Molecular evolution of the non-coding eosinophil granule ontogeny transcript

Dominic Rose1* and Peter F. Stadler2,3,4,5,6,7

1 Bioinformatics Group, Department of Computer Science, University of Freiburg, Freiburg, Germany
2 Bioinformatics Group, Department of Computer Science, Interdisciplinary Center for Bioinformatics, University of Leipzig, Leipzig, Germany
3 Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany
4 Fraunhofer Institut für Zelltherapie und Immunologie, Leipzig, Germany
5 Department of Theoretical Chemistry, University of Vienna, Wien, Austria
6 Center for non-coding RNA in Technology and Health, University of Copenhagen, Frederiksberg, Denmark
7 Santa Fe Institute, Santa Fe, NM, USA

Eukaryotic genomes are pervasively transcribed. A large fraction of the transcriptional output consists of long, mRNA-like, non-protein-coding transcripts (mlncRNAs). The evolutionary history of mlncRNAs is still largely uncharted territory. In this contribution, we explore in detail the evolutionary traces of the eosinophil granule ontogeny transcript (EGOT), an experimentally confirmed representative of an abundant class of totally intronic non-coding transcripts (TINs). EGOT is located antisense to an intron of the ITPR1 gene. We computationally identify putative EGOT orthologs in the genomes of 32 different organisms human and mouse.

INTRODUCTION

Large surveys of transcriptomes, such as ENCODE (ENCODE Project Consortium et al., 2007) and FANTOM (Maeda et al., 2006), demonstrated that eukaryotic genomes are pervasively transcribed (Jacquier, 2009). Long, mRNA-like, non-protein-coding transcripts (mlncRNAs) are an important component of this transcriptional output, often arising from regions unlinked to annotated protein-coding genes (Khalil et al., 2009). Apart from a few exceptions, the detailed function of these transcripts, however, still remains in the dark. The cases that are reasonably well understood, on the other hand, implicate mlncRNAs as key molecules orchestrating essential cellular processes, including gene expression, transcriptional, and post-transcriptional regulation, chromatin-remodeling, differentiation and development (Mercer et al., 2009).

As a group, mlncRNAs show evidence of stabilizing selection (Ponjavic et al., 2007; Marques and Ponting, 2009). Although the evidence for wide-spread evolutionary constraints on the sequence evolution of ncRNAs is the most direct evidence that at least a large fraction of them is in fact functional, we know very little about the evolutionary history of individual transcripts. In contrast to protein-coding genes or short structured ncRNAs, for which comprehensive evolutionary information is available in databases like Pfam (Finn et al., 2010) or Rfam (Gardner et al., 2011), there is no comparable resource for long ncRNAs. The lncRNA database (Amaral et al., 2011) is a first pioneering step in this direction, predominately compiling non-coding transcripts from the model organisms human and mouse.

To-date, only a few detailed case studies are available. Chodroff et al. (2010) recently considered the conservation of a few brain-specific mlncRNAs, reporting weak sequence conservation and major changes in gene structure across amniotes. Even more detailed descriptions of mlncRNA evolution zooming in on the sequences are available only for a few “famous” transcripts. Xist, an eutherian-specific regulatory long ncRNA that plays a central role in inactivation of one female X chromosome by recruiting chromatin-remodeling complexes, reviewed, e.g., by Arthold et al. (2011), is the only long ncRNAs whose evolutionary origin is understood in detail. It arose after the divergence of marsupials and placental mammals from the protein-coding Lnx3 gene upon incorporation of additional, repeat-derived exons (Duret et al., 2006; Elisaphenko et al., 2008; Kolesnikov and Elisafenko,
Wagner et al. (2007) demonstrated that EGOT is not associated dominantly unspliced transcripts (Engelhardt and Stadler, 2011), with ribosomes and thus most likely functions as a mlncRNA. EGO-B transcript, which because of these properties is classified between human and chicken. That was already recognized by Wagner et al. (2007) to be conserved at sequence level. The intron, however, contains a sequence element about 190 nt into the intron. Both transcripts are polyadenylated about 1.4 kb are exonic. In contrast, EGO-A remains unspliced, reaching 2.4 kb. Its primary transcript covers about 2.4 kb, of which about 1.4 kb are exonic. In contrast, EGO-A remains unspliced, reaching about 190 nt into the intron. Both transcripts are polyadenylated (Wagner et al., 2007). Overall, EGOT is quite poorly conserved at sequence level. The intron, however, contains a sequence element that was already recognized by Wagner et al. (2007) to be conserved between human and chicken.

Here, we report on an in-depths computational analysis of EGOT, focusing in particular on the spliced and polyadenylated EGO-B transcript, which because of these properties is classified as a mlncRNA.

**MATERIALS AND METHODS**

Based on the human EGO-B transcript (acc. no. NR_004428.1), orthologs have been retrieved from the UCSC multiz and the Ensembl EPO alignments but were also manually collected by iterative blat/blast searches against genomes publicly available at the UCSC Genome Browser and the Ensembl database, covering the evolutionary range from human to insects. Finally, a multiple sequence alignment was generated using MUSCLE (Edgar, 2004). Beyond reasonable sequence conservation, we applied additional criteria to collect the putative EGO-B orthologs, i.e., the syntenic conservation of flanking genes or an intact exon/intron gene structure with two conserved splice sites. In order to search for potential homologs outside the eutheria, we first identified the region homologous to the conserved element in the intron of EGO-B, extracted the complete ITPR1 intron plus some flanking sequence and used clustalw to construct separate pairwise alignments of each of the two EGOT exons with the genomic DNA sequence. RNA secondary structures were analyzed using the Vienna RNA package (Hofacker et al., 1994) and RNAz (Washietl et al., 2005). The significance of RNAz-predicted structures was analyzed by a control screen consisting of randomized sequence alignments generated by rnazRandomizeAln.pl, which is part of the RNAz package. This script columnwisely shuffles each sequence alignments such that local alignment characteristics and conservation patterns are preserved while the correlation between columns is destroyed. The UCSC Genome Browser was used for visualization of the EGOT locus. Stabilizing selection was quantified using phastCons (Siepel et al., 2005).

**RESULTS**

**GENE PHYLOGENY**

We identified putative EGO-B orthologs in the genomes of 32 different amniotes, see Table 1; Figure 1. Based on the conservation of DNA sequence, gene structure, splice sites, and synteny, we found 25 strong candidate orthologs in primates, rodents, ungulates, carnivores, afrotherians, and xenarthrans. However, seven of the 32 putative orthologs have to be considered as weak. Their exons exhibit additional insertions, no convincing splice sites, or are extremely diverged in sequence from the members of the strong ortholog set. We could not identify EGO-B in all placental mammals: no homolog was found in pika, alpaca, microbat, and hedgehog genomes. We suspect that this is due to the low coverage and incomplete assembly of these genomes and hence constitutes an artifact rather than true gene loss. No indication for the existence of paralogs of the EGOT locus was found.

Trying to resolve distant homologies, we have also compiled EGO-B candidates for opossum, platypus, and chicken. This search was restricted to the ITPR1 locus to increase sensitivity. The putative ortholog in the opossum genome is most likely a true positive: it shows several compositional and syntenic features conserved throughout the eutherian orthologs, such as comparable exon/intron lengths, putatively functional splice sites, as well as the highly conserved intronic element discussed in detail below. Although the sequence of both exons is highly diverged, and hence the alignment of the opossum candidate to the eutherian sequences is rather poor, we hypothesize that EGOT most likely dates back before the divergence of eutheria and marsupials. In contrast, the candidates in platypus and sauropsids are not well supported.

**SEQUENCE CONSERVATION**

The two known human EGO-B exons exhibit average phastCons (Siepel et al., 2005) scores close to zero (~0.04) among mammals (as well as vertebrates) suggesting a remarkably low level of sequence conservation, see Figure 2. In contrast, the two ITPR1 exons flanking EGO-B have phastCons scores of 0.87 and 0.96,
Table 1 | Approximate genomic locations of EGO-B orthologs.

| Species                  | Assembly | Chr.  | 5’ EGO-B    | 3’ EGO-B    | Strand | Size (nt) |
|--------------------------|----------|-------|-------------|-------------|--------|-----------|
| Homo sapiens             | hg19     | chr2  | 4790878     | 4793274     | –       | 2397      |
| Macaca mulatta           | rheMac2  | chr2  | 5627601     | 56278426    | +       | 2410      |
| Mus musculus             | mm9      | chr6  | 108404678   | 108407558   | –       | 2881      |
| Bos taurus               | bosTau4  | chr22 | 22291950    | 22294301    | +       | 2352      |
| Equus caballus           | equCab2  | chr16 | 11378820    | 11381084    | +       | 2265      |
| Felis catus              | felCat4  | A2    | 55998823    | 56001116    | –       | 2294      |
| Canis familiaris         | canFam2  | chr20 | 15833826    | 15836114    | +       | 2289      |
| Choleopus hoffmanni      | choHof1  |       | 145093      | 147373      | +       | 2281      |
| Monodelphis domestica    | monDom5  | chr6  | 236476850   | 236479951   | –       | 3102      |

The coordinates refer to the unspliced genomic regions of EGO-B. Recall that some entries are based on draft assemblies (GeneScaffolds) and the respective coordinates are thus preliminary. A full list of all 32 orthologs is available as Table A1 in Appendix.

respectively. At first glance, this observation conflicts with the initial findings of Wagner et al. (2007) who reported a high level of sequence conservation, which is present nearly exclusively in a highly conserved element (HCE) inside the intron of EGO-B, however. We used phastCons to quantify stabilizing selection. PhastCons uses a hidden Markov model to estimate the probability that each nucleotide of a multiple alignment belongs to a conserved element. Despite differences in detail, the alignment method has surprisingly little influence on the estimates. The average phastCons-score is about 0.09 for the 5’-exon and 0.02 for the 3’-exon, see Figure A1 in Appendix. In fact, major parts of both exons have no measurable conservation signal.

GENE STRUCTURES

RefSeq annotated human exons are on average 307 nt long (Pruitt et al., 2009). In contrast, exons of human pseudogenes are substantially longer. For example, the exons of the Yale pseudogene annotation have average lengths of 482 nt (Zhang et al., 2003). This difference can be explained by a lack of selective constraints to preserve the gene structure of pseudogenes. Among others, retrotransposition may lead to the incorporation of repeats and other artifact sequences. We used the two EGO-B exons as anchors for a local alignment approach to collect orthologs. Thus, the loss or inclusion of additional sequence elements at orthologous EGO-B loci can easily be measured. The lengths of orthologous EGO-B genes vary between 1.9 and 3.2 kb, given that we neglect the 9 kb long Procavia capensis or the 12 kb long Ornithorhynchus anatinus loci because of assembly issues. However, the average gene size (2.4 kb) of all collected orthologs is in perfect agreement with the initially reported 2.4 kb of EGO-B in human (Wagner et al., 2007). In particular, the sizes of the EGO-B 5’-exon, the intron, as well the 3’-exon fit fairly well to the human reference transcript for the majority of orthologs, see Figure 3. The deeply conserved gene structure supports our set of EGO-B candidates and suggests selective constraints acting on EGOT to preserve the spliced isoform.

SPlice SITE CONSERVATION

The presence of evolutionarily conserved splice sites would further support our set of putative EGO-B orthologs and is usually indicative of a functionally relevant transcript. The majority of the 32 transcript candidates shows canonical splice site sequences at positions homologous to the known splice sites in human: 56% (18/32) have both a standard GT donor and an AG acceptor (59% (19/32) have a GT donor, 88% (28/32) an AG acceptor). Furthermore, we classified the EGO-B splice sites using MaxEntScan (Yeo and Burge, 2004), a maximum entropy modeling approach that discriminates real from false splice sites. As depicted in Figure 4A, 50% (16/32) of all donors and 94% (30/32) of all acceptors yield positive MaxEntScan scores implying that the sequence motifs of these sites are in agreement with known splice sites and therefore likely functional. Scoring the potential splice sites with a novel log-odds scoring scheme that evaluates substitution patterns of vertebrate splice sites and their ancestral sequences along a phylogenetic tree (Rose et al., 2011) yields −16.48 for EGO-B donors and 14.67 for EGO-B acceptors. Again, positive scores are indicative of functional splice sites. The evolutionary traces of substitutions at EGO-B acceptors are summarized in Figure 4B. Interestingly, we observed twice as many (24) substitution events typical for real acceptors compared to 12 atypical substitution events. However, there is a highly conserved TATA box-like motif at the EGO-B donor (see Figure 4C), which might explain the low donor scores as the consequence of an additional selective constraint. Even in human, the MaxEntScan donor score is only half of the corresponding acceptor signal. In summary, our results suggest intact splice sites for at least half (16/32) but likely even more of the analyzed species.

SYNTENIC CONSERVATION

Wagner et al. (2007) have previously reported that EGO-B is transcribed antisense to an intron of the ITPR1 gene inositol triphosphate receptor type 1. However, ITPR1 is strictly syntenically linked to SUMF1 and BHLHE40 throughout vertebrates. The ancestral gene order of the ITPR1 locus seems to be SETMAR(+), SUMF1(−), ITPR1(+), BHLHE40(+), ARL8B(+), since this arrangement is present in basically all species in which we have detected EGO-B. Figure 2 (top) gives a compact overview of the gene synteny in human. The fact that synteny is intact and deeply conserved among a variety of vertebrate species supports our collection of EGO-B orthologs. The ITPR1 gene is conserved throughout vertebrates and the HCE in the intron of eutharian EGO-B is detectable throughout amniotes, with a plausible
candidate also visible in *Xenopus*. Nevertheless, no convincing EGO-B orthologs were found outside placental mammals and marsupials.

**PROMOTERS**

Not much is known about the transcriptional regulation of EGOT. ENCODE data suggest four possible promoter regions for EGOT, see Figure A3 in Appendix. On the one hand, digital DNaseI hypersensitivity clusters obtained via tiling array experiments (Sabo et al., 2006) indicate three possible promoter regions upstream of EGOT. On the other hand, ChipSeq histone marks (Ernst et al., 2011) suggest an internal promoter located at the 5′-exon of EGO-B. However, the putative promoter regions are only moderately conserved at the sequence level. Among the four candidates, the external one, which is directly located upstream of EGO-B, exhibits the highest sequence conservation, better phastCons scores (0.21) than EGO-B, and can be traced back until zebrafish.

**MYSTERIOUS HIGHLY CONSERVED ELEMENTS**

The EGOT locus contains three elements of unknown function that are highly conserved at the sequence level, see Figure 5. Two of these HCEs flank EGOT and another is located within the intron of EGO-B. As suggested above, the upstream HCE may function as a promoter. Using Q-RT-PCR, Wagner et al. (2007) already confirmed abundant expression at the intronic HCE. Next, there is transcriptional evidence from EST data (FN099218) derived from 454 deep sequencing of primary human breast cancer (Guffanti et al., 2009) and an RNA-seq library of healthy breast tissue (Wang et al., 2008). In the recent release of the Rfam database (10.1, June 2011) the intronic HCE is already listed as EGOT (RF01958; Gardner et al., 2011). However, it is still not satisfactorily resolved whether these HCEs are part of novel EGOT isoforms, belong to independent, yet undiscovered, transcripts, or other functionally relevant regions.

Wagner et al. (2007) considered the intronic HCE to be independent of EGOT, since it, contrary to EGO-B, was not inducible with IL-5. This assumption is further backed by our bioinformatic analyses predicting a putative novel exon with conserved splice sites at the intronic HCE (Rose et al., 2011), see Figure 5. The putative exon cannot be part of another EGOT isoform, since it is in opposite reading direction. Spliced short reads from the ENCODE Caltech RNA-seq track (Mortazavi et al., 2008) verify the predicted splice site and reveal that the predicted exon is part of a novel ITPR1 isoform.

Moreover, the consensus sequence of the TATA box-like motif at the EGO-B donor (see Figure 4) is TAATA. This element might act as a promoter for an individually transcribed element. It has previously been shown that the TAATA motif can enhance transcription, i.e., it is part of the promoter of the human glucocorticoid receptor gene (Govindan et al., 1991).

**EXPERIMENTAL EVIDENCE FOR TRANSCRIPTION**

In addition to sequence homology, EST data are typically used to determine the approximate evolutionary extent of a long ncRNA. There are several cDNAs available experimentally confirming EGO-A and EGO-B, see Figure 5. Analyzing the UniGene EST profiles reveals approximate gene expression patterns. EGOT has been detected in various adult human body sites, predominantly adipose tissue, bone marrow, and kidney. Beyond healthy cell lines it is also expressed in various tumor tissues, such as liposarcoma or breast cancer.

However, the available EST data for EGOT mainly derives from human tissues, cDNAs from other species are rare. Beyond human cDNAs, there are only ESTs from *Macaca fascicularis*.

---

**FIGURE 1 | Overview of eutherian EGO-B orthologs.** Species in which a (nearly) complete EGO-B gene was found are highlighted. EGO-B candidates are labeled with question marks. The figure is based on the Ensembl species tree.
The figure illustrates the difficulties of obtaining orthologs of long ncRNAs due to a lack of sequence conservation. Black horizontal arrows highlight our manually curated EGO-B orthologs of human and horse. Blat searches are basically not sensitive enough, since they only recover fragments of the actual EGO-B exons. In case of horse, for example, even the RefSeq track insufficiently lists only the partial gene structure. Next, sequence conservation, or actually the probability of the EGO-B locus to be under negative selection, is close to zero according to the phastCons program. Nevertheless, we were able to compile a set of at least 25 strong candidates of EGO-B orthologs.

The gene structure, in particular the exon/intron lengths of the identified EGO-B transcripts are well conserved among orthologs. Horizontal lines mark the lengths of the human reference. In summary, the gene structure of the majority of species fits fairly well to the human reference. Notable exceptions are hyrax and platypus due to incomplete genome assemblies. As expected, especially rodents exhibit additional insertions compared to other higher vertebrates.
We found that EGOT is highly structured. Using RNAz (Washietl et al., 2005), genome-wide RNAz-based studies have estimated their false discovery rates (FDR) at ∼20–60% (Missal et al., 2005, 2006; Rose et al., 2011). Both sequences strongly support the expression of the EGO-B 3′-exon, but do not provide a complete proof, since they are unspliced and their reading direction is not known. However, many of the human ESTs can successfully be mapped to several non-human EGOT orthologs recovering the established human gene structure.

Non-coding RNA profiling by high throughput sequencing of nuclear RNA in bone marrow-derived macrophages (De Santa et al., 2010) reveals extragenic Pol-II transcription sites at the mouse EGOT ortholog. As depicted in Figure A5 in Appendix, deep sequencing confirms transcription of the intronic HCE and parts of the 3′-end of the mouse EGOT ortholog. Although the data do not validate the full mouse ortholog, their experiments are still in line with our results. On the one hand, the two independently transcribed regions at the intronic HCE support our hypothesis that the HCE consists of two independent domains, a non-coding and a protein-coding one. Next, since it was previously postulated that EGOT may act via siRNAs to repress its targets MBP and EDN (Wagner et al., 2007), the signals at the 3′-end on the other hand might indeed indicate small RNAs that are hosted by EGOT. In summary, the experimental data of (De Santa et al., 2010) from mouse tally well with what is known from human EGOT.

SECONDARY STRUCTURES

We found that EGOT is highly structured. Using RNAz (Washietl et al., 2005), we identified five regions that exhibit thermodynamically stable and evolutionary conserved secondary structure motifs, see Figure 6. EGO-A contains a distinctive secondary structure at its 3′-end, which therefore might act as a termination signal. Remarkably, one of the EGO-B elements is located at the splice junction and thus can only be formed by the mature (spliced) transcript. In total, 43% (635/1462 nt) of the mature EGO-B transcript exhibit such prominent secondary structure motifs. In line with EGOT, the intronic HCE also shows RNAz-predicted signatures of preserved secondary structures. Figure A4 in Appendix depicts the predicted minimum free energy structures for several species and illustrates their evolutionary conservation in more detail. As expected, a sequence/structure-based clustering using LocARNA (Will et al., 2007) of the corresponding orthologs nearly perfectly recovers the six structural groups.

RNAz is a window based approach. To demonstrate that all six structured regions found at the EGOT locus can indeed be attributed to constraints on EGOT orthologs, we set up a control screen consisting of shuffled alignment windows. The standard screen consisted of 351 input alignment windows, which partially overlap, not only because EGO-B and EGO-A already overlap, but also because several window sizes and various step-widths were tested. Overall, 45 of 351 windows were classified as structured RNA in the standard screen. However, only a single window was classified as structured in the control screen. This significant enrichment of structured windows in real versus control screen supports the significance of these RNAz predictions. We note that genome-wide RNAz-based studies have estimated their false discovery rates (FDR) at ∼20–60% (Missal et al., 2005, 2006; Rose et al., 2007, 2008). Here, we consider only a small locus with a highly significant signal for conserved structure.

Next, we applied LocARNA-P (Will et al., submitted), a novel approach estimating the precise boundaries of non-coding RNAs. It combines sequential and structural reliability information to a profile that depicts constrained and therefore likely functional regions. As illustrated in Figure 6, RNAz considers only a sub-region of the HCE to be structured. It is at least partially confirmed by EST data. However, the LocARNA-P reliability profile reveals additional signals of viable secondary structures next to the RNAz hit and suggests a larger non-coding gene. In summary, the HCE is not only conserved at the sequence level, it also harbors distinct secondary structures possibly associated with relevant biological functions.

We propose that the intronic HCE has ambiguous functions (at least dual), since we could show that it contains both protein-coding domains as well as non-coding elements. Most strikingly,
The figure illustrates RNAi-predicted secondary structures of EGOT and the adjacent intronic HCE. Both loci exhibit signals of thermodynamically stable and evolutionary conserved secondary structures. We provide the approximate genomic position and the predicted minimum free energy structure (RNAfold) for each RNAi hit. Computing the potential gene boundaries of a putative regulatory element located at the HCE using LocARNA-P reveals additional signals of conserved secondary structures. Peaks in the reliability profile are indicative of constrained and therefore most likely functionally relevant regions. Structural contributions to the reliability profile are depicted in dark gray, sequential in light gray.

Although both the genomic location in an intron of ITPR1 and the gene structure (i.e., both splice sites) is conserved at least throughout the placental mammals, the putative transcript is quite poorly conserved at the sequence level. In contrast to protein-coding genes and short, structured ncRNAs, this is a rather common feature of long ncRNAs in general (Marques and Ponting, 2009; Chodroff et al., 2010). Hence EGOT appears to be a rather typical representative of the mRNA-like ncRNAs.

Superimposed on the overall low level of sequence conservation, the EGOT locus contains also highly conserved regions. In particular, we have characterized the intronic HCE and untangled the LocARNA-P-derived reliability profile apparently visualizes this dual character of the HCE. The sharp decrease of reliability signal clearly separates the patterns of putative non-coding RNAs in form of conserved secondary structures from the novel protein-coding ITPR1 exon.

**DISCUSSION**

We have traced here the evolutionary history of EGOT, one of the first totally intronic long ncRNA that has been studied in detail. The spliced isoform, EGO-B, may be present throughout the placental mammals, and most likely dates back even further.
its complex nature. The 3′ part of the HCE can be recognized as an undescribed exon of ITPR1. Thus, it might even be that EGOT expression affects the (alternative) splicing of ITPR1 as it is known from the Saf/Fas locus (Yan et al., 2005). Its 5′ side shows evidence for expression unrelated to both ITPR1 and EGOT, exhibits a well conserved secondary structure element and features a conserved TATA-like element potentially acting as a promoter. Our results could furthermore be used to extend and refine the EGOT Rfam entry (RF01958), which at the moment just covers the intrinsic HCE but not the actual EGOT transcript.

In order to assess EGOT orthologs computationally, we have analyzed apparent indexes of conservation like synteny or the presence of functional splice sites at the EGOT locus. Although we have collected computational indication for a deep evolutionary conservation of EGOT, it is still theoretically possible that some of our signals might not be due to the putative EGO-B transcript orthologs, but to other yet unidentified functional elements in the region.

Surprisingly, a large part of EGO-B is folded into evolutionary conserved secondary structures. This sets it apart from the few other well-studied long ncRNAs. HOTAIR, for instance, has been reported to contain functional secondary structure elements whose evolutionary conservation appears to be weak (Tsai et al., 2010; He et al., 2011; Schorderet and Duboule, 2011) and requires further analysis. MALAT-1, on the other hand, exhibits only a few small conserved structured elements despite its overall high level of sequence conservation (Stadler, 2010). Furthermore, Marques and Ponting (2009) reported a moderate enrichment of conserved structural elements in some but not all types of long ncRNAs. This calls for a more systematic analysis of RNA secondary structures in long ncRNAs. The difference in structure content suggests, in particular, that this could be an important means of distinguishing functional classes of long ncRNAs.

The overall low level of sequence conservation is a serious obstacle for comparative genomics approaches. It limits first the sensitivity of homology search and then the accuracy of multiple sequence alignments. The large size of the molecules and the often complex and variable exon/intron structures, on the other hand, makes it extremely tedious to resort to manual improvements of alignments, in particular since currently available alignment editors are unable to accommodate complex annotation data. Recently developed tools (Rose et al., 2011) for the systematic assessment of splice site conservation were instrumental both in recognizing the additional exon in the HCE and in providing computational evidence for the conservation of the EGO-B orthologs.

The comparison of original genome-wide alignments and manually curated alignments of the EGOT locus demonstrates several drawbacks of pre-computed alignments (see also Figure A2 in Appendix). Pre-computed genome-wide alignments require substantial post-processing. Separated into alignment blocks, reference-based alignments often contain only partial sequences for some species since the orthologous sequence is not included in some alignment blocks, while on the other hand insertions not included in the reference are not represented at all. A third type of artifact consists in misaligned sequences that violate synteny. Of course, all these issues in principle also pertain to protein-coding regions. High levels of sequence conservation of coding regions and comparably little variability of intron/exon structure in coding regions, however, makes coding regions the most high-quality parts of genome-wide alignments. In-depth case studies such as the present one are thus instrumental in determining the types of problems that need to be considered in constructing analysis pipelines that deal with long non-coding RNAs at genome-wide levels.

Eosinophil granule ontogeny transcript has previously been proposed to affect myeloid development by regulating eosinophil gene expression in human. Eosinophils are generally responsible for an immune response to multicellular parasites and certain infections, not only in human, but in all vertebrates. Therefore, it would be conclusive that EGOT is also present in vertebrates fulfilling similar regulatory roles as in human. In turn, the functional assessment of a putative human-specific EGOT gene bears also great potential for evolutionary as well as clinical bioinformatics. However, further experimental evidence validating the expression of the proposed EGOT orthologs is required to ultimately assess the depth of evolutionary conservation of EGO-B.

ACKNOWLEDGMENTS

We are thankful to Sebastian Will and co-authors for permission to use LocARNA-P prior to its publication. We gratefully acknowledge the contributions of Manja Marz and Annegret Wilde.

REFERENCES

Amaral, P., Clark, M., Gascoigne, D., Dinger, M., and Mattick, J. (2011). IncRNAdb: a reference database for long noncoding RNAs. Nucleic Acids Res. 39, D146–D151.

Arthold, S., Kurowski, A., and Wutz, A. (2011). Mechanistic insights into chromosome-wide silencing in X inactivation. Hum. Genet. 130, 295–305.

Chodroff, R., Goodstadt, L., Sirey, T., Oliver, P., Davies, K., Green, E., Molnár, Z., and Ponting, C. (2010). Long noncoding RNA genes: conservation of sequence and brain expression among diverse amniotes. Genome Biol. 11, R72.

De Santa, F., Barozzi, L., Mietton, F., Ghisletti, S., Polletti, S., Tusi, B., Muller, H., Ragoussis, J., Wei, C., and Natoli, G. (2010). A large fraction of extragenic RNA pol II transcription sites overlap enhancers. PLoS Biol. 8, e1000384. doi:10.1371/journal.pbio.1000384

Duret, L., Chureau, C., Samain, S., Weissenbach, J., and Avner, P. (2006). The Xist RNA gene evolved in eutherians by pseudogenization of a protein-coding gene. Science 312, 1653–1655.

Edgar, R. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797.

Elisaphenko, E. A., Kolesnikov, N. N., Shevchenko, A. I., Rogaizin, I. B., Nesterova, T. B., Brockdorff, N., and Zakian, S. M. (2008). A dual origin of the Xist gene from a protein-coding gene and a set of transposable elements. PLoS ONE 3, e2521. doi:10.1371/journal.pone.0002521

ENCODE Project Consortium, Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigó, R., Gingeras, T. R., Margulies, E. H., Weng, Z., Snyder, M., Dermitzakis, E. T., Thurman, R. E., Kuehn, M. S., Taylor, C. M., Neph, S., Koch, C. M., Asthana, S., Maltotra, A., Adzhubei, I., Greenbaum, J. A., Andrews, R. M., Flicek, P., Boyle, P. J., Cao, H., Carter, N. P., Clelland, G. K., Davis, S., Day, N., Dhami, P., Dillon, S. C., Dorschner, M. O., Fiegler, H., Giresi, P. G., Goldy, J., Hawrylycz, M., Haydock, A., Humbert, R., James, K. D., Johnson, B. E., Johnson, E. M., Frum, T. T., Rosenzweig, E. R., Karnani, N., Lee, K., Lefebvre, G. C., Navas, P. A., Neri, F., Parker, S. C., Sabo, P. J., Sandstrom, R., Shafer, A., Vetrie, D., Weaver, M., Wilcox, S., Yu, M., Collins, F. S., Dekker, J., Lieb, J. D.,
Tullius, T. D., Crawford, G. E., Sun—yaev, S., Noble, W. S., Dunham, I., Denoeud, F., Reymond, A., Kapra—nov, P., Rozowsky, J. S., Zheng, D., Castelo, R., Frankish, A., Harrow, J., Ghosh, S., Sandelin, A., Hofacker, I. L., Baertsch, R., Keefe, D., Dike, S., Cheng, J., Hirsch, H. A., Sekinger, E. A., Lagarde, J., Abril, J. F., Sha—hab, A., Flammi, C., Fried, C., Hack—ermüller, J., Hertel, J., Lindemeyer, M., Missal, K., Tanzer, A., Washietl, S., Korbel, J. Emannuelsson, O., Pedersen, J. S., Holroyd, N., Taylor, R., Swift, B., Swarbreck, D., Matthews, N., Dickson, M. C., Thomas, D. J., Weirauch, M. T., Gilbert, J., Drenkow, J., Bell, I., Zhao, X., Srinivasan, K. G., Sung, W. K., Osi, H. S., Chiu, K. P., Fois—Hartman, S., Urban, A. E., Kraus, P. R., King, D. C., Ameur, A., Enroth, D., Zhang, Z. D., Barrera, L., Stuart, Miller, W., Sidow, A., Trinklein, N. D., Children's Hospital Oakland Research Center, Broad Institute, Chil—baylor College of Medicine Human Genome Sequencing Center, Washington University Genome Sequencing Center, Broad Institute, Children's Hospital Oakland Research Institute, Batzoglou, S., Goldman, N., Hardison, R. C., Hausler, D., Miller, W., Sidow, A., Trinklein, N. D., Zhang, Z. D., Barrera, L., Stuart, R., King, D. C., Ameur, A., Enroth, S., Bieda, M. C., Kim, J., Bihinge, A. A., Jiang, N., Liu, J., Yao, F., Gougle, J., Frith, M., Enstroem, P. G., Lenhard, B., Attaralaya, R. N., Bat—slow, S., Beisel, K. W., Bult, C. J., Chen, F. C., Forrest, A. R., Furuno, M., Hill, D., Itoh, M., Kamonori—Katayama, M., Katayama, S., Kato, M., Kawashima, T., Quackenbush, J., Ravasi, T., Ring, B. Z., Shibata, K., Sugira, K., Takahana, Y., Teasdale, R. D., Wells, C. A., Zhu, Y., Kai, C., Kawai, J., Hume, D. A., Carninci, P., and Hayashizaki, Y. (2006). Transcript annotation in FANTOM3: mouse gene catalog based on physical cDNAs. PLoS Genet. 2, e62. doi:10.1371/journal.pgen.0020062

Marques, A., and Ponting, C. (2009). Catalogues of mammalian long non—coding RNAs: modest conservation and incompleteness. Genome Biol. 10, R124.

McGuire, K., and Glass, E. (2005). The expanding role of microarrays in the investigation of macrophage responses to pathogens. Vet. Immunol. Immunopathol. 105, 259—275.

Misseral, K., Rose, D., and Stadler, P. (2005). Long non—coding RNAs: insights into functions. Nat. Rev. Genet. 10, 155—159.

Missal, K., Rose, D., and Stadler, P. (2005). Non—coding RNAs in ciona intestinalis. Bioinformatics 21(Suppl. 2), i77—i78.

Missal, K., Zhu, X., Rose, D., Deng, W., Skogerboe, G., Chen, R., and Stadler, P. (2006). Prediction of structured non—coding RNAs in the genomes of the nematodes caenorhabditis elegans and caenorhabditis briggsae. J. Exp. Zool. B Mol. Dev. Evol. 306, 379—392.

Mortzavvi, A., Williams, B., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and expression analyses of human intronic noncoding RNAs reveal tissue—specific patterns and enrichment in genes related to regulation of transcription. Genome Biol. 8, R43.

Osada, N., Hashimoto, K., Kameoka, Y., Hirata, M., Tanuma, R., Uno, Y., Inoue, I., Hida, M., Suzuki, Y., Sugano, S., Terao, K., Kusuda, J., and Takahashi, I. (2008). Large—scale analysis of macaca fascicularis transcripts and inference of genetic divergence between M. fascicularis. October 2011 | Volume 2 | Article 69 | 9

www.frontiersin.org
## APPENDIX

Table A1 | Approximate genomic locations of EGO-B orthologs.

| Species                              | Assembly     | Chr.      | 5' EGO-B | 3' EGO-B | Strand (±) | Size (nt) |
|--------------------------------------|--------------|-----------|----------|----------|------------|-----------|
| Homo sapiens                         | hg19         | chr3      | 4790878  | 4793274  | –          | 2397      |
| Pan troglodytes                      | panTro2      | chr3      | 4878075  | 4880473  | –          | 2399      |
| Gorilla gorilla                      | gorGor3      | chr3      | 4902164  | 4904567  | –          | 2404      |
| Pongo pygmaeus                       | ponAbe2      | chr3      | 65374639 | 65377039 | –          | 2401      |
| Macaca mulatta                       | rheMac2      | chr2      | 56276017 | 56278426 | +          | 2410      |
| Papio hamadryas                      | Pham 1.0     | Contig1259 Contig623173 | 26345 | 28758 | + | 2413 |
| Callicebus jacchus                   | caIac3       | chr15     | 56602911 | 56605332 | –          | 2422      |
| Tarsius syrichta                     | tarSyr1      | GeneScaffold 4896 | 162358 | 164326 | – | 1969 |
| Microcebus murinus                   | micMur1      | –         | –        | –        | –          | –         |
| Otolemur garnetti                    | BUSHBABY1    | GeneScaffold 2768 | 553545 | 555837 | – | 2293 |
| Tupai belangeri                      | tupBel1      | Scaffold 127316 | 516 | 2657 | + | 2142 |
| Mus musculus                         | mm9          | chr6      | 108404678 | 108407558 | –          | 2881      |
| Rattus norvegicus                    | rm4          | chr4      | 143936406 | 143939404 | –          | 2999      |
| Dipodomys ordii                      | dipOrd1      | GeneScaffold 6600 | 155406 | 158566 | – | 3160 |
| Cavia porcellus                      | cavPor3      | Scaffold 16 | 32052549 | 32055063 | –          | 2514      |
| Spermophilus tridecemlineatus        | SQUIRREL     | GeneScaffold 3331 | 244508 | 246869 | – | 2361 |
| Oryctolagus cuniculus                | oryCun2      | GL018703  | 3454423  | 3456525  | –          | 2102      |
| Tursiops truncatus                   | turTru1      | GeneScaffold 1936 | 210524 | 212948 | – | 2425 |
| Bos taurus                           | bosTaur4     | chr22     | 22291950 | 22294301 | +          | 2352      |
| Equus caballus                       | equCab2      | chr16     | 11378820 | 11381084 | +          | 2265      |
| Felis catus                          | felCat4      | A2        | 56998823 | 56001116 | –          | 2294      |
| Ailuropoda melanoleuca               | ailMel1      | GL192717.1 | 421344 | 423702 | – | 2359 |
| Canis familiaris                     | canFam2      | chr20     | 15833826 | 15836114 | +          | 2289      |
| Pteropus vampyrus                    | pteVam1      | GeneScaffold 2203 | 226110 | 228253 | – | 2143 |
| Loxodonta africana                   | loxAfr3      | chr12     | 42811986 | 42814765 | –          | 2780      |
| Procavia capensis                    | proCap1      | GeneScaffold 4371 | 187873 | 197725 | + | 9853 |
| Echinops telfairi                    | TENREC       | GeneScaffold 5028 | 354037 | 357269 | + | 3233 |
| Dasypus novemcinctus                 | dasNov2      | GeneScaffold 4264 | 285144 | 287278 | – | 2135 |
| Choleopus hoffmanni                  | choHof1      | GeneScaffold 4676 | 145093 | 147373 | + | 2281 |
| Monodelphis domestica                | monDom5      | chr6      | 236476850 | 236479961 | – | 3102 |
| Ornithorhynchus anatinus             | ornAna1      | X1        | 44628568 | 44640839 | – | 12271 |
| Gallus gallus                        | galGal3      | chr12     | 19134732 | 19138187 | – | 3455 |

The coordinates refer to the unspliced genomic regions of EGO-B. Recall that some entries are based on draft assemblies (GeneScaffolds). These genomes contain the EGO-B gene but the respective coordinates are preliminary. In case of assembly problems (e.g., the gene is covered by different scaffolds), no genomic coordinates are given.
FIGURE A1 | Sequence conservation of EGO-B as indicated by the phastCons program. We have computed the phastCons scores for three different alignment approaches using the vertebrate model available at the UCSC Genome Browser: (1) our own muscle alignment, (2) the pre-computed 46-way vertebrate multiz alignment from the UCSC browser, and (3) a clustalw alignment based on our set of orthologs. The peaks are fairly similar, only clustalw slightly differs due to a higher alignment error rate resulting from the lack of consistency transformation or similar alignment refinement/improvement steps. Applied phastCons parameters: transitions 0.01, 0.01; rho 0.4.

FIGURE A2 | Differences in aligned sequence data between our approach and pre-computed UCSC multiz alignments. We plot the sequence portion of the 5′-exon of EGO-B for different species and alignment approaches. There is a large overlap but also substantial differences between the two alignment approaches. An obvious drawback of pre-computed alignments is missing data. For example, the UCSC alignments do not contain the ortholog of panda because the assembly was not ready at the time the alignments were generated. More strikingly, non-human orthologs are only partially included in the UCSC alignments, since it is a reference (human) based approach. Regions that would cause larger gaps in human, such as mouse or rat which exhibit various exonic insertions, are only partially included in the final alignment. However, partial sequences are crucial for any subsequent analysis relying on valid alignments, i.e., RNA secondary structure prediction.
ENCODE data suggest four possible promoter regions for EGOT (marked by the four arrows). Digital DNase1 hypersensitivity clusters indicate three promoter candidate regions upstream of EGOT. On the other hand, histone marks suggest an internal promoter at the 5′-exon of EGO-B. The figure contains only exemplary cell lines. However, the depicted signal peaks, especially the DNase1 hypersensitivity peaks, are consistently present in numerous cell lines.

The minimum free energy structures of the six RNAz-predicted regions are at least partially conserved throughout higher eukaryotes. A sequence/structure-based clustering using LocARNA (Will et al., 2007) visualizes the similarities between the predicted structures in more detail. As expected, the structures nearly perfectly cluster into the six groups.
FIGURE A5 | Non-coding RNA profiling by high throughput sequencing reveals extragenic Pol-II transcription sites at the mouse EGOT ortholog. The deep sequencing data from (De Santa et al., 2010) available in the Gene Expression Omnibus (under GEO accession number GSE20370) confirm transcription of the intronic highly conserved element (HCE) and parts of the 3′-end of the mouse EGOT ortholog. Although the data do not validate the full mouse ortholog, we benefit twice from the depicted transcribed regions. On the one hand the two independently transcribed regions at the intronic HCE support our findings that the HCE consists of two independent non-coding as well as protein-coding domains. Next, since it was previously postulated that EGOT may act via siRNAs to repress its targets MBP and EDN (Wagner et al., 2007), the signals at the 3′-end on the other hand might indeed indicate small RNAs that are hosted by EGOT.