Molecular Evolution of the non-coding Eosinophil Granule Ontogeny Transcript EGOT
– Supplement –

Dominic Rose [a,h] and Peter F. Stadler [b,c,d,e,f,g]

[a] *Bioinformatics Group, Department of Computer Science, University of Freiburg, Georges-Köhler-Allee 106, D-79110 Freiburg, Germany.*

[b] *Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany.*

[c] *Max Planck Institute for Mathematics in the Sciences, Inselstrasse 22, D-04103 Leipzig, Germany*

[d] *Fraunhofer Institut für Zelltherapie und Immunologie – IZI Perlickstraße 1, D-04103 Leipzig, Germany*

[e] *Department of Theoretical Chemistry, University of Vienna, Währingerstraße 17, A-1090 Wien, Austria*

[f] *Center for non-coding RNA in Technology and Health, University of Copenhagen, Grønnegårdsvej 3, DK-1870 Frederiksberg C, Denmark*

[g] *Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA*

[h] *Corresponding author:* `dominic@bioinf.uni-freiburg.de`

Table 1
Approximate genomic locations of EGO-B orthologs. 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.

Species	Assembly	Chr.	5' EGO-B	3' EGO-B	±	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
Callithrix jacchus	calJac3	chr15	56602911	56605332	-	2422
Tarsius syrichta	tarSyr1	GeneScaffold_4896	162358	164326	-	1969
Microcebus murinus	micMur1	-	-	-	-	-
Otolemur garnettii	BUSHBABY1	GeneScaffold_2768	553545	555837	-	2293
Tupaia belangeri	tupBel1	scaffold_127316	516	2657	+	2142
Mus musculus	mm9	chr6	108404678	108407558	-	2881
Rattus norvegicus	rn4	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_1935	210524	212948	-	2425
Bos taurus	bosTau4	chr22	22291950	22294301	+	2352
Equus caballus	equCab2	chr16	11378820	11381084	+	2265
Felis catus	felCat4	A2	55998823	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
Choloepus_hoffmanni	choHof1	GeneScaffold_4676	145093	147373	+	2281
Monodelphis domestica	monDom5	chr6	236476850	236479951	-	3102
Ornithorhynchus anatinus	ornAna1	X1	44628568	44640839	-	12271
Gallus gallus	galGal3	chr12	19134732	19138187	-	3455

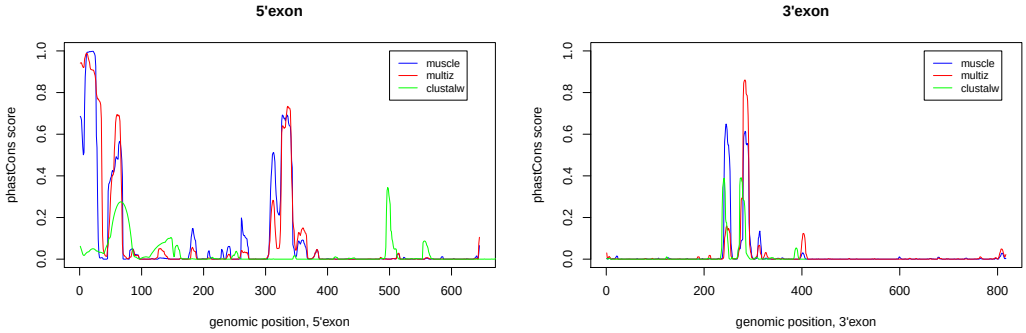

Fig. 1. **Sequence conservation of EGO-B as indicated by the `phastCons` program.** We have computed the **`phastCons`** scores for three different alignment approaches using the vertebrate model available at the UCSC Genome Browser: (1) our own **`muscle`** alignment, (2) the pre-computed 46-way vertebrate **`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.

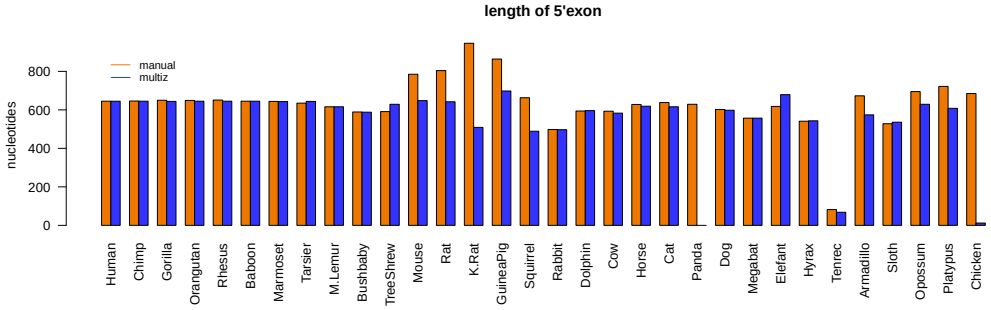

Fig. 2. **Differences in aligned sequence data between our approach and pre-computed UCSC `multiz` alignments.** We plot the sequence portion of the 5'exon of EGO-B for different species and alignment approaches. There is a large overlap but also substantial differences between the two alignment approaches. 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.

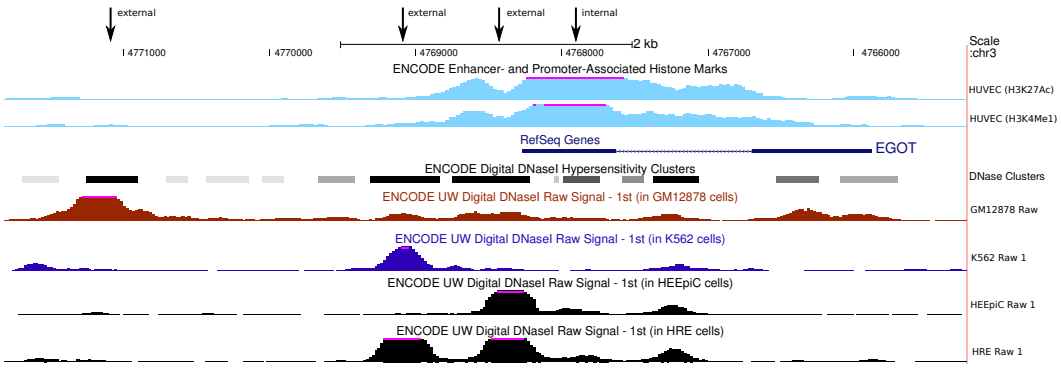

Fig. 3. **EGOT promoter regions.** ENCODE data suggest four possible promoter regions for EGOT (marked by the four arrows). Digital DNase1 hypersensitivity clusters indicate three promoter candidate regions upstream of EGOT. On the other hand, histone marks suggest an internal promoter at the 5'exon of EGO-B. The figure contains only exemplary cell lines. However, the depicted signal peaks, especially the the DNase1 hypersensitivity peaks, are consistently present in numerous cell lines.

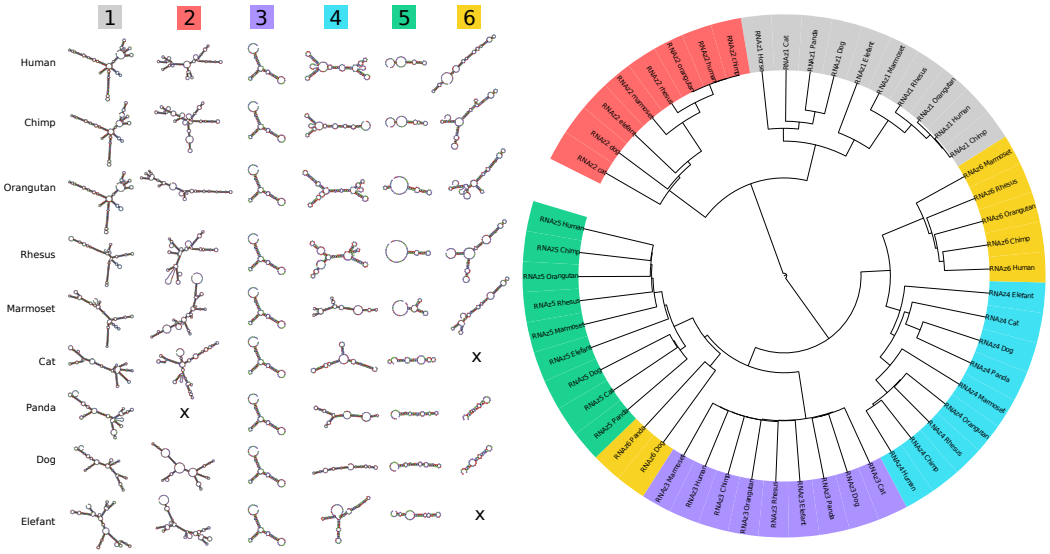

Fig. 4. **Traces of evolutionary conserved secondary structures.** The minimum free energy structures of the six `RNAz`-predicted regions are at least partially conserved throughout higher eukaryotes. A sequence/structure-based clustering using `LocARNA` (Will et al., 2007) visualizes the similarities between the predicted structures in more detail. As expected, the structures nearly perfectly cluster into the six groups.

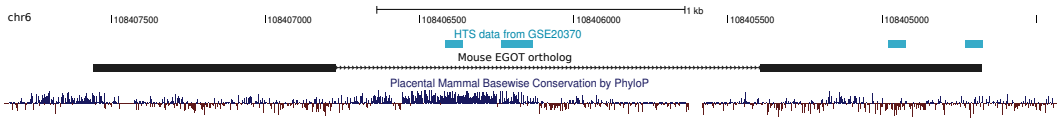

Fig. 5. **Non-coding RNA profiling by high throughput sequencing reveals extragenic Pol-II transcription sites at the mouse EGOT ortholog.** The deep sequencing data from (De Santa et al., 2010) available in the Gene Expression Omnibus (under GEO accession number GSE20370) confirm transcription of the 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 in deed indicate small RNAs that are hosted by EGOT.

References

De Santa, F., Barozzi, I., 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.

Wagner, L., Christensen, C., Dunn, D., Spangrude, G., Georgelas, A., Kelley, L., Esplin, M., Weiss, R., and Gleich, G., 2007. EGO, a novel, noncoding RNA gene, regulates eosinophil granule protein transcript expression. *Blood* **109**: 5191–8.

Will, S., Reiche, K., Hofacker, I., Stadler, P., and Backofen, R., 2007. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comput Biol* **3**: e65. doi:10.1371/journal.pcbi.0030065.