D698–D702 *Nucleic Acids Research, 2009, Vol. 37, Database issue* *Published online 5 November 2008*
doi:10.1093/nar/gkn813

EVOG: a database for evolutionary analysis of overlapping genes

Dae-Soo Kim[1], Chi-Young Cho[2], Jae-Won Huh[3], Heui-Soo Kim[4] and Hwan-Gue Cho[2,*]

[1]Korean BioInformation Center (KOBIC), KRIBB, Daejeon 305-806, [2]School of Computer Science and Engineering, College of Engineering, Pusan National University, Busan 609-735, Korea, [3]National Primate Research Center (NPRC), KRIBB, Ochang, Chungbuk 363-883, Republic of Korea and [4]Division of Biological Sciences, College of Natural Sciences, Pusan National University, Busan 609-735, Korea

Received August 14, 2008; Accepted October 10, 2008

ABSTRACT

Overlapping genes are defined as a pair of genes whose transcripts are overlapped. Recently, many cases of overlapped genes have been investigated in various eukaryotic organisms; however, their origin and transcriptional control mechanism has not yet been clearly determined. In this study, we implemented evolutionary visualizer for overlapping genes (EVOG), a Web-based DB with a novel visualization interface, to investigate the evolutionary relationship between overlapping genes. Using this technique, we collected and analyzed all overlapping genes in human, chimpanzee, orangutan, marmoset, rhesus, cow, dog, mouse, rat, chicken, Xenopus, zebrafish and Drosophila. This integrated database provides a manually curated database that displays the evolutionary features of overlapping genes. The EVOG DB components included a number of overlapping genes (10074 in human, 10 009 in chimpanzee, 67 039 in orangutan, 51 001 in marmoset, 219 in rhesus, 3627 in cow, 209 in dog, 10 700 in mouse, 7987 in rat, 1439 in chicken, 597 in Xenopus, 2457 in zebrafish and 4115 in Drosophila). The EVOG database is very effective and easy to use for the analysis of the evolutionary process of overlapping genes when comparing different species. Therefore, EVOG could potentially be used as the main tool to investigate the evolution of the human genome in relation to disease by comparing the expression profiles of overlapping genes. EVOG is available at http://neobio.cs.pusan.ac.kr/evog/.

INTRODUCTION

Overlapping genes are described as different genes whose genomic regions overlap to some extent. This is frequently observed in viral and prokaryotic genomes as well as in mitochondrial DNA and is believed to be a common strategy for genome organization and gene regulation in bacteria (1). However, there is a growing body of evidence that suggests overlapping genes may regulate key gene expression mechanisms in eukaryotes (2) with genomic imprinting (3), RNA interference, translational regulation (4), transcriptional interference (5) and RNA editing (6). Moreover, using bioinformatic approaches based on expressed sequence tags and full-length cDNA sequences it has been estimated that ~20% of human genes are overlapping genes (7,8). Despite their abundance, the origin and evolution of overlapping genes in eukaryotes remain unclear (9).

Recently, several studies have reported that the occurrence of overlapping genes may have been an advantageous factor for gene expression, regulation and/or a harmful factor for provoking new diseases during evolution (10). In humans, an increasing amount of evidence indicates that overlapping genes were a major factor in the culmination of various human diseases (11). For example, imprinting related SNURF-SNRPN and UBE3A overlapping genes is associated with Prader–Willi and Angelman syndromes (12). Beckwith–Wiedemann syndrome (13), Angelman syndrome (14) and transient neonatal diabetes (15) were also suggested as overlapping gene related diseases. Cross-species comparative analysis on large-scale datasets of nucleotide sequences, genomic structure and gene expression are considered to be an effective approach to enrich our knowledge of the functionally important elements (16). The availability of the complete genome

*To whom correspondence should be addressed. Tel: +82 51 510 2283; Fax: +82 51 515 2208; Email: hgcho@pusan.ac.kr
Correspondence may also be addressed to Heui-Soo Kim. Email: khs307@pusan.ac.kr

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

sequence and the accumulation of millions of expressed sequences have made it possible for large-scale predictions of naturally occurring overlapping genes (3).

Overlapping genes have been primarily considered to be important in regulatory gene structure for mRNA degradation and translational repression. In this study, we employed evolutionary approaches to address the functional roles of overlapping genes by comparing the genome organization between humans and different species. We performed a comprehensive comparative genomics analysis on 10 different genomes and found overlapping patterns of overlapping genes during the evolution of the genome. In addition, we developed a semi-automatic system to identify overlapping genes on a massive scale using publicly available sequence databases. We identified features and conservation of overlapping genes, and inferred possible mechanisms responsible for overlap formation. Our system could be a valuable resource for analyzing and comparing overlapping genes between animal genomes that range from human to insects. In addition, Evolutionary visualizer for overlapping genes (EVOG) could be potentially used as the main tool to investigate the evolution of the human genome in relation to disease by comparing the expression profile of overlapping genes.

DATABASE CONSTRUCTION

Dataset

We used the combined data of publicly available mRNA and genome alignment from the University of California, Santa Cruz genome browser database (http://genome.ucsc.edu; March, 2006, release). These alignments were produced by BLAT using mRNA databases. Our DB includes 13 genome datasets (human, chimpanzee, orangutan, marmoset, rhesus, cow, dog, mouse, rat, chicken, Xenopus, zebrafish and drosophila), which were obtained from the UCSC genome database recently updated.

Searching for overlapping genes from transcript sequences

In this study, overlapping genes are defined as the adjacent gene sequences that overlap partially and share one or more nucleotides. Nevertheless, the complete gene structure should include both the transcription regulatory regions at the 5′ upstream end and the termination region at the 3′ downstream end of coding sequences. We systematically analyzed all overlapping genes in the genomes of thirteen species. We limited our analysis to protein-coding genes and we did not consider alternative splicing forms of a gene to be overlapping genes. Because we were especially interested in addressing the evolution of overlapping genes, we required that all genes in our study have strict orthologs between human and other genomes. To obtain high quality genome data, we focused on only mRNA sequences for the genome alignment data from the UCSC genome browser database and did not include ESTs. Therefore, we used the genome to mRNA sequence alignment data calculated in the UCSC Genome Browser database. In the data, we attempted to map the mRNA sequence to the genome sequence. To reduce the workload and improve the mapping quality, we first applied the selected sense orientation reliable transcripts. All imperfect alignments were removed. The transcripts sequences that were aligned to more than one genomic fragment were discarded as suspected chimeras. We searched for overlapped genes from genes that were transcribed on the opposite strands of the same genomic locus. All of the putative overlapped genes were also mapped onto the genome. If the RefSeq mRNA sequences overlapped, only the longest was considered. We searched for overlapping sense/antisense and gene-in-gene pairs based on the coordinates of the RefSeq in the genome sequence. To cover both the transcriptional initiation and termination sites of all the gene structures, we expanded the overlapping regions to allow adjacent upstream or downstream gene regions to partially overlap. Using this procedure, Human, Chimpanzee, Orangutan, Marmoset, Rhesus, Cow, Dog, Mouse, Rat, Chicken, Xenopus, Zebrafish and Drosophila genes were identified as overlapping genes (Table 1).

Table 1. Statistics for EVOG database

Species	Genome assembly	No. of overlapping gene pairs	No. of genes
Human	hg18	10 074	27 062
Chimp	panTro2	10 009	27 306
Orangutan	ponAbe2	67 039	187 740
Marmoset	calJac1	51 001	204 327
Rhesus	rheMac2	219	513
Cow	bosTau4	3627	10 161
Dog	canFam2	209	923
Chicken	galGal3	1439	4326
Mouse	mm9	10 700	46 859
Rat	rn4	7987	21 551
Xenopus	xenTro2	597	8987
Zebrafish	danRer5	2457	30 540
Drosophila	dm3	4115	21 158

Identification of overlapping genes according to pairing region

Overlapping genes were then categorized into seven different types, 3′UTR-to-3′UTR, 5′UTR-to-5′UTR, Intron-to-Exon, 3′UTR-to-5′UTR, 5′UTR-to-3′UTR, Non-exon overlapping and single exon to-UTR overlapping. These classes were determined according to the relative genome location using genomic mapping data from UCSC genome browser. The antisense transcripts from the opposite strand of the same genomic locus were included. Chimeras of overlapping genes were collected from the databases described above by selecting sequences that contained two parts, were each at least 50 bp long, were aligned to different genes and had opposite orientations.

Comparative analysis of overlapping genes

A comparative analysis of overlapping genes between species was conducted by performing a comprehensive comparative genomics analysis across 13 genomes and

identifying genes that were overlapped. An interesting phenomenon that has been observed is the gain and loss of overlapping states. To address this question, we also examined the overlapping states of orthologous genes in 13 other genomes, human, chimpanzee, orangutan, marmoset, rhesus, cow, dog, mouse, rat, chicken, Xenopus, zebrafish and drosophila. The evolutionary relationships between the overlapping genes in the 13 analyzed species were assessed by extracting all overlapping genes conserved between each pair of species. To analyze the evolutionary impact of overlapping genes among human, chimpanzee, orangutan, marmoset, rhesus, cow, dog, mouse, rat, chicken, Xenopus, zebrafish and drosophila, we compared the human genome with other genomes. Selected species were chosen based on wide-range cross-species comparisons with human, in addition to compiling relatively complete datasets of genomic sequences and abundant transcript sequences.

Constructing phylogenetic trees from overlapping genes

Our multi species dataset enabled us to identify how many overlapping genes were conserved between human and different species. The distance matrix was calculated using the number of overlapping genes that existed between the two-genome pair. For this analysis we used a metric function to compute the similarity of gene proximity. In the following, let $g_{n,m}$ denote a gene whose name is n on a species S_m.

From this we aimed to estimate the difference between two different gene pairs, $\{g_{a,x}, g_{b,x}\}$ and $\{g_{a,y}, g_{b,y}\}$ over two different species (chromosome) S_x, S_y, where $g_{a,x}$ and $g_{b,x}$ are orthologous to $g_{a,y}$ and $g_{b,y}$, respectively. In the following analysis g_x will denote the generic gene identity. Therefore, g_x is orthologous to all $g_{x,W}$, and S_W represents all species. To demonstrate the utility of this analytical method, we initially considered only two different genes. However, computing the similarity between multiple genes from two different species can be easily done by extending the following procedure. Let $begin(p,A)$ denote the starting position (in terms of base pair) of gene $g_{p,A}$ on species S_A. In a similar way $end(p,A)$ denotes the ending position of gene $g_{p,B}$ on species S_B. And $|g_{a,x}|$ is the total length of a gene such that $|g_{a,x}| = |begin(a,x) - end(a,x)|$. (Figure 1) Let $\mathrm{sim}(S_x, S_y \mid g_a, g_b)$ be the similarity of the configuration of two genes g_a, g_b on S_x compared with g_a, g_b on S_y. Note that our measure is not symmetric;

$$\mathrm{sim}(S_x, S_y|g_a, g_b) \neq \mathrm{sim}(S_y, S_x|g_a, g_b)$$

The formal definition of $\mathrm{sim}(S_x, S_y \mid g_a, g_b)$ is as follows.

$$\mathrm{sim}(S_x, S_y|g_a, g_b) = \frac{\text{Maximal Common interval}[(g_{a,x}, g_{b,x}), (g_{a,y}, g_{b,y})]}{|g_{a,y}| + |g_{b,y}|}$$

The Maximal Common interval between $(g_{a,x}, g_{b,x})$ and $(g_{a,y}, g_{b,y})$ can be maximized by moving $(g_{a,y}, g_{b,y})$ over S_x. If $g_{a,x}$, $g_{b,x}$ is completely identical to $g_{a,y}, g_{b,y}$, respectively then $\mathrm{sim}(S_x, S_y \mid g_a, g_b) = 1$. In Figure 2, S_y was slightly aligned by moving it right in order to maximize the common (overlapping intervals). Common_a and

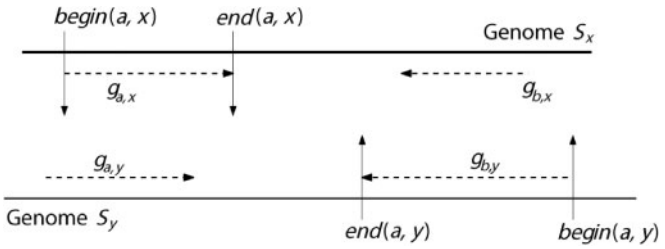

Figure 1. Computing the similarity of gene proximity between two different species.

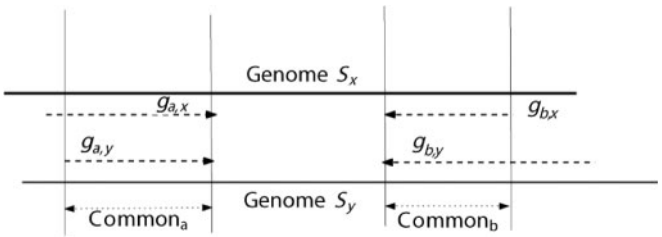

Figure 2. Computing the maximal common interval. S_y was slightly moved to the right in order to maximize the common intervals.

Common_b intervals denote the overlapping intervals between g_a and g_b over S_x and S_y. This yields the following result in Figure 2.

$$\mathrm{sim}(S_x, S_y|g_a, g_b) = \frac{|\mathrm{Common}_a| + |\mathrm{Common}_b|}{|g_{a,y}| + |g_{b,y}|}$$

We should also consider the direction of the gene (upstream or downstream) when computing sim( ). Thus, those above computations are valid only if the direction of the matched gene is consistent, since matching an upstream gene to a downstream gene is not reasonable.

By exploiting this sim( ) calculation, we can construct a phylogenetic tree in terms of the proximity information on multiple genes. Due to this a typical method, e.g. Nearest-neighborhood Joining, can be easily applied.

USER INTERFACE

The EVOG database is publicly accessible at http://neobio.cs.pusan.ac.kr/evog/. Before using EVOG, every user has to confirm that JRE (JRE 1.6 or newest) is installed on their local computer. There are various ways for users to access the data stored in the EVOG database. The database can be browsed by selecting a specific genome and gene name from the main page. The web interface allows users to access the database content via three different search options. First, users can search the genes of interest by using the HUGO symbol name. In addition, one can use this route to get gene sequences and detailed gene information from the NCBI data bank (Figure 3A). Second, users can search overlapping genes by clicking one of the genomes listed on the main page (Figure 3C). The genome browser of EVOG will then show annotation features of all the overlapped genes when a particular organism on the multiple genome menu is clicked (Figure 3A). To investigate evolutionary

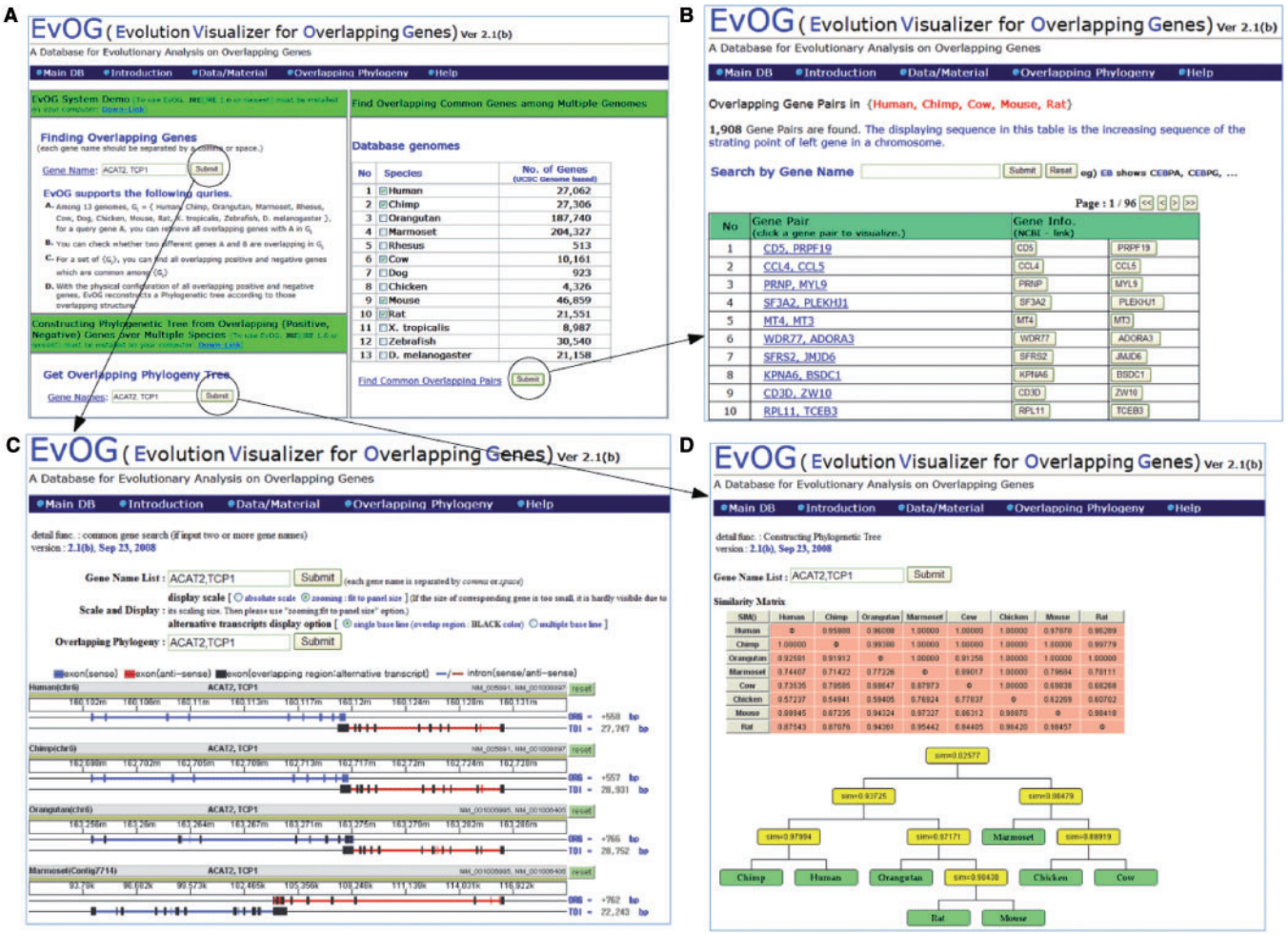

Figure 3. (**A**) Overlapping genes web retrieval interface. The web interface allows users to access the database contents via three different searching options. This is useful for finding organisms with overlapping genes specified by the users. (**B**) The results page from the EvOG database. The results page is very effective and easy to use for the comparative analysis and investigation of the evolutionary process of overlapping genes. (**C**) The users can search interesting overlapping genes by click genome listed on the main page. (**D**) The distance matrix was calculated by the number of overlapping genes between each two-genome pair.

aspects of overlapping genes, we implemented EVOG with a novel visualization interface. The web-based genome browser was implemented using JavaServer Faces (JSF) technology, which has the advantage of constructing a clearly defined architecture by separating application logic and presentation. In addition, the EVOG database supports a visualization interface that shows the comparative configuration of overlapping genes across multiple species along the whole chromosome scale. As shown in Figure 3B, we show one overlapping gene pairs (AUP1, HTRA2), which appears commonly in human, chimpanzee, cow and mouse. It displays the distance measured between overlapping genes as a phylogenetic tree (Figure 3D), enabling users to infer the evolutionary history of the overlapping genes at a glance. In the picture, sense transcripts are in red and anti-sense transcripts are in blue. The small thick segment denotes the exon and the thin line denotes the intron. Users can freely zoom in/out of a specified region to more closely investigate the overlapping regions of gene pairs. EVOG database supports two different viewing scales, the absolute scale and the zoomed scale, both of which are fit to the panel size. The absolute scale is a fixed viewing scale, which includes the overlapping genes that appear in multiple genomes. Zooming makes the viewing scale fit the actual size. Users can enlarge the viewing interval by selecting the scale lines with the mouse. Moreover, the EVOG database incorporates multiple genome and tree visualization tools to facilitate online visualization of the data.

SUMMARY AND FUTURE DIRECTIONS

The EVOG database is an integrated database for overlapping genes that includes bioinformatics analysis data. EVOG supports all overlapping genes that are common in a subset of Human, Chimpanzee, Orangutan, Marmoset, Rhesus, Cow, Dog, Mouse, Rat, Chicken, Xenopus, Zebrafish and Drosophila. In addition, it provides a manually curated database that displays the evolutionary

features of overlapping genes. The EVOG database is very effective and easy to use for the comparative analysis and investigation of evolutionary processes of overlapping genes. Furthermore, the EvOG database supports a visualization interface that shows the comparative configuration of overlapping genes across multiple species along the whole chromosome scale. The database is constantly being supplemented with new genome data from a range of other available sources. In the near future, the EVOG database will include the whole genomes of more than 20 species for comparison of commonly overlapping genes. Therefore, EVOG could potentially be used as the main tool to investigate the evolution of the human genome in relation to disease by comparing the expression profiles of overlapping genes.

ACKNOWLEDGEMENTS

All authors would like to express their great thanks to editor and one reviewer for improving the stability of our DB.

FUNDING

Korea Science and Engineering Foundation grant funded by the Korea government (MOST) (No. R01-2007-000-20035-0).

Conflict of interest statement. None declared.

REFERENCES

1. Wagner,E.G. and Simons,R.W. (1994) Antisense RNA control in bacteria, phages, and plasmids. *Annu. Rev. Microbiol.*, **48**, 713–742.
2. Peters,N.T., Rohrbach,J.A., Zalewski,B.A., Byrkett,C.M., Casari,G. and Vaughn,J.C. (2003) RNA editing and regulation of *Drosophila* 4f-rnp expression by sas-10 antisense readthrough mRNA transcripts. *RNA*, **9**, 698–710.
3. Lavorgna,G., Dahary,D., Lehner,B., Sorek,R., Sanderson,C.M. and Casari,G. (2004) In search of antisense. *Trends Biochem. Sci.*, **29**, 88–94.
4. Yelin,R., Dahary,D., Sorek,R., Levanon,E.Y., Goldstein,O., Shoshan,A., Diber,A., Biton,S., Tamir,Y., Khosravi,R. *et al.* (2003) Widespread occurrence of antisense transcription in the human genome. *Nat. Biotechnol.*, **21**, 379–386.
5. Brantl,S. (2002) Antisense-RNA regulation and RNA interference. *Biochim. Biophys. Acta*, **1575**, 15–25.
6. Dahary,D., Elroy-Stein,O and Sorek,R. (2005) Naturally occurring antisense: transcriptional leakage or real overlap? *Genome Res.*, **15**, 364–368.
7. Williams,B.A., Slamovits,C.H., Patron,N.J., Fast,N.M. and Keeling,P.J. (2005) A high frequency of overlapping gene expression in compacted eukaryotic genomes. *Proc. Natl Acad. Sci. USA*, **102**, 10936–10941.
8. Rougeulle,C. and Heard,E. (2002) Antisense RNA in imprinting: Spreading silence through Air. *Trends Genet.*, **18**, 434–437.
9. Zhang,Y, Liu,X.S., Liu,Q.R. and Wei,L. (2006) Genome-wide in silico identification and analysis of cis natural antisense transcripts (cis-NATs) in ten species. *Nucleic Acids Res.*, **34**, 3465–3475.
10. Prescott,E.M. and Proudfoot,N.J. (2002) Transcriptional collision between convergent genes in budding yeast. *Proc. Natl Acad. Sci. USA*, **99**, 8796–8801.
11. Kumar,M. and Carmichael,G.G. (1998) Antisense RNA: Function and fate of duplex RNA in cells of higher eukaryotes. *Microbiol. Mol. Biol. Rev.*, **62**, 1415–1434.
12. Runte,M. *et al.* (2001) The IC-SNURF-SNRPN transcript serves as a host for multiple small nucleolar RNA species and as an antisense RNA for UBE3A. *Hum. Mol. Genet.*, **10**, 2687–2700.
13. Lee,M.P., DeBaun,M.R., Mitsuya,K., Galonek,H.L., Brandenburg,S., Oshimura,M. and Feinberg,A.P. (1999) Loss of imprinting of a paternally expressed transcript, with antisense orientation to KvLQT1, occurs frequently in Beckwith–Wiedemann syndrome and is independent of insulin-like growth factor II imprinting. *Proc. Natl Acad. Sci. USA*, **96**, 5203–5208.
14. Buiting,K., Lich,C., Cottrell,S., Barnicoat,A. and Horsthemke,B. (1999) A 5-kb imprinting center deletion in a family with Angelman syndrome reduces the shortest region of deletion overlap to 880 bp. *Hum. Genet.*, **105**, 665–666.
15. Arima,T., Drewell,R.A., Arney,K.L., Inoue,J., Makita,Y., Hata,A., Oshimura,M., Wake,N. and Surani,M.A. (2001) A conserved imprinting control region at the HYMAI/ZAC domain is implicated in transient neonatal diabetes mellitus. *Hum. Mol. Genet.*, **10**, 1475–1483.
16. Boffelli,D., Nobrega,M.A. and Rubin,E.M. (2004) Comparative genomics at the vertebrate extremes. *Nat. Rev. Genet.*, **5**, 456–465.