EICO (Expression-based Imprint Candidate Organizer): finding disease-related imprinted genes

Itoshi Nikaido1,2,3, Chika Saito3, Akiko Wakamoto3, Yasuhiro Tomaru3, Takahiro Arakawa3, Yoshihide Hayashizaki1,3 and Yasushi Okazaki2,3,*

1Division of Genomic Information Resource Exploration, Science of Biological Supramolecular Systems, Yokohama City University, Graduate School of Integrated Science, Yokohama, Kanagawa 230-0045, Japan, 2Division of Functional Genomics and Systems Medicine, Research Center for Genomic Medicine, Saitama Medical School, Saitama 350-1241, Japan and 3Laboratory for Genome Exploration Research Group, RIKEN Genomic Sciences Center (GSC), RIKEN Yokohama Institute, Yokohama, Kanagawa 230-0045, Japan

Received August 13, 2003; Revised and Accepted October 7, 2003

ABSTRACT
We have developed an integrated database that is specialized for the study of imprinted disease genes. The database contains novel candidate imprinted genes identified by the RIKEN full-length mouse cDNA microarray study, information on validated single nucleotide polymorphisms (SNPs) to confirm imprinting using reciprocal mouse crosses and the predicted physical position of imprinting-related disease loci in the mouse and human genomes. It has two user-friendly search interfaces: the SNP-central view (MuSCAT: MoUse SNP CATalog) and the candidate gene-central view (CITE: Candidate Imprinted Transcripts by Expression). The database, EICO (Expression-based Imprint Candidate Organizer), can be accessed via the World Wide Web (http://fantom2.gsc.riken.jp/EICODB/) and the DAS client software. These data and interfaces facilitate understanding of the mechanism of imprinting in mammalian inherited traits.

INTRODUCTION
Genomic imprinting results in the expression of individual genes from only one of two parental chromosomes and affects growth and behavior after birth in mammals (1). Aberrant imprinting can lead to various diseases due to an effective doubling of gene dosage. Conversely, genetic diseases display complex inheritance patterns, through the male or female line, when the affected gene falls within a maternally or paternally imprinted locus. Identification of the network of imprinted genes will provide insight into the molecular mechanisms that underlie imprinting-related phenotypes and diseases. To date ~60 imprinted mouse genes have been identified using various methods (http://www.mgu.har.mrc.ac.uk/imprinting/all_impmaps.html). Genomic imprinting involves promoter methylation and/or natural antisense transcripts (NATs) of imprinted or neighboring genes (2); however, the details are unclear. Imprinting clearly cannot be predicted from genomic sequencing and annotation alone (1). We have established an efficient method of screening for candidate imprinted transcripts, and target genes by comparing mRNA expression profiles between parthenogenotes and androgenotes using RIKEN cDNA microarrays (3,4). Although our screening method is very efficient, a fraction (32%) of the identified candidate genes proved to be non-imprinted (3). These non-imprinted genes could be regulated by imprinted genes. To confirm the imprinted status of candidate transcripts, we performed reciprocal crosses with Mus musculus molossinus (MSM), a Japanese wild mouse strain, and analyzed the resulting transcripts for polymorphisms that distinguish paternal from maternal loci. Since MSM is phylogenetically 1 million years apart from common laboratory mouse strains, it exhibits frequent genetic polymorphisms with laboratory mice. To this end, we searched for polymorphisms in the 3¢-end of the transcripts between MSM and C57BL/6J mouse lines and the results were assembled into the EICO. In this paper, we report the construction and implementation of the EICO (http://fantom2.gsc.riken.jp/EICODB/), which efficiently stores and retrieves three kinds of data: (i) candidate imprinted transcripts from microarray analysis, (ii) single nucleotide polymorphisms (SNPs) between the 3¢-end sequences of the RIKEN full-length cDNAs from C57BL/6J and MSM mice, and (iii) imprinting-related disease loci extracted from OMIM (5). The relationship between disease loci and novel imprinted mRNAs identifies new candidates that may be involved causally in imprinting-related human genetic diseases.

DATABASE STRUCTURE AND CONTENTS
The EICO contains 2850 SNPs between C57BL/6J and MSM found in 1281 RIKEN mouse full-length cDNA clones and 2101 candidate imprinted genes derived from microarray experiment data (Table 1). Of the 2101 candidate imprinted genes, 1403 showed maternal expression and 698 showed paternal expression. There were 243 candidate imprinted genes with reciprocal polymorphisms in the 3¢-end sequences of the RIKEN full-length cDNAs that may be involved in human genetic diseases.

*To whom correspondence should be addressed at Division of Functional Genomics and Systems Medicine, Research Center for Genomic Medicine, 1397-1 Yamane, Hidaka City, Saitama 350-1241, Japan. Tel: +81 429 85 7319; Fax: +81 429 85 7329; Email: okazaki@saitama-med.ac.jp
The EICO contains 2850 SNPs, 2101 candidate imprinted genes, and 529 candidate imprinted genes within predicted imprinting-related disease loci.

### Table 1. Contents of the EICO

<table>
<thead>
<tr>
<th>Contents</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNPs between MSM and C57BL/6J (1)</td>
<td>2850 (1281 genes)</td>
</tr>
<tr>
<td>Candidate imprinted genes (2)</td>
<td>Paternal: 698; maternal: 1403</td>
</tr>
<tr>
<td>Genes overlapped between 1 and 2 (3)</td>
<td>243</td>
</tr>
<tr>
<td>Candidate imprinted genes on predicted imprinting-related disease loci on human genome (4)</td>
<td>529 genes; 65 diseases; 109 loci</td>
</tr>
<tr>
<td>Genes overlapped between 3 and 4</td>
<td>114 genes</td>
</tr>
</tbody>
</table>

The EICO has 2850 SNPs, 2101 candidate imprinted genes, and 529 candidate imprinted genes within predicted imprinting-related disease loci.
position of a candidate gene is within an imprinted-related disease locus, (iii) whether the genomic position of a candidate gene is close to a known imprinting cluster and (iv) whether the candidate gene is non-coding RNA (ncRNA) (14). This information can be accessed by the color bar code on the web interface (Fig. 1). The EICO includes 159 NATs, 56 ncRNA and 39 genes mapped to known imprinted cluster loci. Finally, the EICO can be queried with elements such as RIKEN clone ID, RIKEN Rarray ID, FANTOM Annotation, nucleic acid and amino acid sequence using SSAHA (15) and BLAST (16). These data and interfaces in the EICO will serve as a major resource for understanding the mechanism of imprinting in mammalian inherited traits.

IMPLEMENTATION

The EICO is currently implemented using MySQL, an open source relational database management system (http://www.mysql.com/), on Konkura MNU/Linux. MuSCAT. The CITE interface systems are based on an Apache web server (http://www.apache.org/) and CGI programs written in Perl (http://www.cpan.org/) and the object-oriented scripting language Ruby (http://www.ruby-lang.org/). To make hyperlinks to other databases interactive, the EICO uses the DAS protocol using Lightweight Distributed Annotation System (LDAS) (http://www.biodas.org/servers/).

DATA AVAILABILITY AND CITING THE EICO

All users can interactively access all candidate imprinted genes, SNPs and candidate imprinted genes mapped to predicted imprinting-related disease loci via the world wide web at the following URL: http://fantom2.gsc.riKEN.jp/EICO/DB/. The MuSCAT and CITE searching systems can be accessed at http://fantom2.gsc.riKEN.jp/EICO/SNP/, http://fantom2.gsc.riKEN.jp/EICO/Imprinting/. The sequence similarity search interfaces for the EICO are http://fantom2.gsc.riKEN.jp/EICO/ssaha/ and http://fantom2.gsc.riKEN.jp/EICO/BLAST/. The server for the DAS for the EICO services is at http://fantom2.gsc.riKEN.jp/EICO/DAS. Please refer to this article and Nikaido et al. (4) when citing the EICO.

FUTURE DIRECTIONS

Novel imprinted candidate genes in the EICO will be increased by progressive accumulation of RIKEN full-length cDNA microarray data. The information of validated candidate imprinted genes will be reflected in the EICO when the data are updated. The EICO will import public mouse SNPs within confirmed candidate imprinted genes.

ACKNOWLEDGEMENTS

We thank the following: Yosuke Mizuno, Hidemasa Bono, Shiro Fukuda, Takeya Kasukawa, Ken Yagi, Naoko Tominaga, Yuki Tsujimura, Tomohiro Kono, Yukiko Yamazaki, Toshihiko Shiroishi and Kazuo Moriwaki for technical assistance and discussion. We would like to acknowledge David Hume of the University of Queensland and Elva Diaz of the University of California at Davis for helpful discussion and English editing. This study was also supported by the Special Coordination Fund for the Promotion of Science and Technology; a fund entrusted by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) to Y.O. and by a Research Grant for the RIKEN Genome Exploration Research Project from the Ministry of Education, Culture, Sports, Science and Technology of the Japanese Government to Y.H.

REFERENCES


