MetaImprint: an information repository of mammalian imprinted genes

Yanjun Wei1, Jianzhong Su1, Hongbo Liu2, Jie Lv2, Fang Wang1, Haidan Yan2, Yanhua Wen1, Hui Liu2, Qiong Wu2,* and Yan Zhang1,*

ABSTRACT
Genomic imprinting is a complex genetic and epigenetic phenomenon that plays important roles in mammalian development and diseases. Mammalian imprinted genes have been identified widely by experimental strategies or predicted using computational methods. Systematic information for these genes would be necessary for the identification of novel imprinted genes and the analysis of their regulatory mechanisms and functions. Here, a well-designed information repository, MetaImprint (http://bioinfo.hrbmu.edu.cn/MetaImprint), is presented, which focuses on the collection of information concerning mammalian imprinted genes. The current version of MetaImprint incorporates 539 imprinted genes, including 255 experimentally confirmed genes, and their detailed research courses from eight mammalian species. MetaImprint also hosts genome-wide genetic and epigenetic information of imprinted genes, including imprinting control regions, single nucleotide polymorphisms, non-coding RNAs, DNA methylation and histone modifications. Information related to human diseases and functional annotation was also integrated into MetaImprint. To facilitate data extraction, MetaImprint supports multiple search options, such as by gene ID and disease name. Moreover, a configurable Imprinted Gene Browser was developed to visualize the information on imprinted genes in a genomic context. In addition, an Epigenetic Changes Analysis Tool is provided for online analysis of DNA methylation and histone modification differences of imprinted genes among multiple tissues and cell types. MetaImprint provides a comprehensive information repository of imprinted genes, allowing researchers to investigate systematically the genetic and epigenetic regulatory mechanisms of imprinted genes and their functions in development and diseases.

KEY WORDS: Imprinted genes, Database, Genetic and epigenetic information, Imprinting disorder

INTRODUCTION
Genomic imprinting, a complex epigenetic phenomenon whereby gene expression occurs in a parent-of-origin specific manner in various cell or tissue types, has an essential role in normal growth and development (Reik and Walter, 2001). Fetal development is promoted generally by paternally expressed genes (Wu et al., 2006), whereas maternally expressed genes have a tendency to restrict fetal growth (Abu-Amero et al., 2008). Imprinting is found predominantly in placental mammals, and may participate in the balance of parental resource allocation to the offspring (Ishida and Moore, 2013). It has been reported that imprinting is related to the evolution of the placenta in mammals (Fang et al., 2012). Defects in imprinting have been associated with congenital diseases, infertility, molar pregnancy and defective assisted reproductive cells (Tomizawa and Sasaki, 2012). Loss of imprinting (LOI) induced by methylation disruption at the imprinted loci may lead to serious imprinting disorders (e.g. Beckwith–Wiedemann syndrome) (Bliik et al., 2001), and cancers (e.g. Wilms’ tumor) (Rancourt et al., 2013; Wu et al., 2008).

Extensive studies have attempted to identify the imprinted genes in various cell and tissue types of mammals, and several experimental strategies have been used to identify imprinted genes, based on DNA methylation status analysis between two alleles, gene-targeting experiments and whole-transcriptome RNA-sequencing analysis (Babak et al., 2008; Barlow et al., 1991; Luedi et al., 2007; Wang et al., 2008). In addition, computational methods have been used to predict new imprinted genes, based on DNA sequence features, CpG islands, repetitive elements, miRNA clusters and epigenetic modifications in the human and mouse genomes (Brideau et al., 2010; Luedi et al., 2005). The mechanisms that establish and maintain imprinted genes have been found using different perspectives, such as DNA sequence features (Khatib et al., 2007), non-coding RNAs (ncRNAs) (Royo and Cavaille, 2008) and epigenetic modifications (Barlow, 2011; Koerner and Barlow, 2010).

Over the past few years, several databases have been established to store information related to imprinted genes, including MouseBook (Blake et al., 2010), WAMIDEX (Schulz et al., 2008), the Otago Catalogue of Imprinted Genes (Morison and Reeve, 1998) and Genieprint (Bischoff et al., 2009). Both MouseBook and WAMIDEX mainly focus on mouse imprinted genes. Although the Otago Catalogue of Imprinted Genes and Genieprint cover diverse species, the records for the imprinted gene are limited, such that even the genomic loci information is limited or even lacking. In addition, as reference genes in public nucleotide databases are updated, a portion of the imprinted gene records in these databases have become outdated and cannot be positioned in the recent versions of the genomes. In addition, the emergence of high-throughput sequencing technologies in recent years has made it possible to perform genome-wide annotations of imprinted genes, such as cell/tissue specificity, genetic and epigenetic information, as well as related diseases. Therefore, a specifically designed database, MetaImprint, was constructed by manually mining the literature related to imprinted genes and incorporating meta-information from the high-throughput genetic and epigenetic data. The current version of MetaImprint covers imprinted genes from eight species (human, mouse, sheep, rat, dog, rabbit, cow and pig).

Compared with the pre-existing databases, MetaImprint has four advantages. (1) The allele or cell/tissue specificity and detailed
research processes of imprinted genes are shown in the detailed pages as a result of literature mining. (2) MetaImprint provides the integrative annotation of genetic and epigenetic information for each imprinted gene at any given chromosome location, such as imprint control regions (ICRs), single nucleotide polymorphisms (SNPs), repetitive elements, ncRNAs, DNA methylation and histone modifications. In particular, a visual engine, Imprinted Gene Browser, was developed to show a landscape view of the imprinted genes in the context of existing genetic and epigenetic annotations. (3) The detailed relationships between imprinted genes and human diseases are also incorporated. (4) MetaImprint provides two analysis tools to identify epigenetic differences and for functional enrichment analysis of imprinted genes among cell/tissue types. In brief, MetaImprint is a comprehensive and cross-species database for the systematic study of imprinted genes in terms of genetics and epigenetics, as well as diseases and evolution.

RESULTS

Database content and construction

The current release of MetaImprint was designed to provide imprinted gene-related information regarding the imprinting states, epigenetic modifications, and related diseases. By mining the literature manually, we collected 845 imprinted gene-related studies from 1991 to date. From these studies, 539 mammalian imprinted genes from eight mammalian species were obtained, including 317 human genes, 146 mouse genes, three sheep genes, seven rat genes, one dog gene, one rabbit gene, 35 cow genes and 29 pig genes (Table 1). Among these imprinted genes, 255 imprinted genes have been validated experimentally and the remainder were predicated by computational methods from the human and mouse genomes.

Based on the collected information, we investigated the research trends of imprinted genes over time. This revealed that the number of studies related to imprinting increased from 1900, especially in the years since 2007 (Fig. 1A). In most of the studies, allele-specific expression of genes was used to discover imprinted genes by experimental and computational methods. For example, Igf2 was identified as the first imprinted gene in mouse in 1991, and H19 in human in 1992. The time spectrum for the discovery of imprinted genes revealed that there was a spurt in the identification of novel imprinted genes around 2008 (Fig. 1B). However, the number of newly discovered imprinted genes decreased year by year from 2010 whereas the number of studies related to imprinting genes increased. Although the identification of novel imprinted genes is useful, researchers have been paying more attention to the regulatory mechanisms and functions of imprinted genes. Thus, the regulatory and functional information of imprinted genes would be necessary for current analysis.

DNA methylation and histone modification information may be useful to uncover epigenetic regulatory effects on genome imprinting. Recently, increasing numbers of studies related to epigenetic modifications (e.g. DNA methylation) have revealed that epigenetic modifications are regulators of allele-specific expression of imprinted genes (Court et al., 2014; Fang et al., 2012; Shoemaker et al., 2010). In addition, some studies focused on the imprinting mechanisms in the development and related functions of imprinted genes (Chang et al., 2013). Studies on SNPs, which are ideal markers for identifying imprinted genes, revealed the roles of abnormal imprinting in complex traits, including complex diseases (Ishida and Moore, 2013; Lawson et al., 2013). It has been reported that imprinted genes contain tandem repeat arrays in their CpG islands more frequently (Hutter et al., 2006). Thus, we obtained and normalized the genomic and epigenomic data from public data resources including the University of California Santa Cruz (UCSC) Genome Bioinformatics Site (Karolchik et al., 2014), the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) database (Barrett et al., 2013) and the NCBI Sequence Read Archive (SRA) (Wheeler et al., 2007).

By integrating genetic and epigenetic information with imprinted genes and other information, we constructed the MetaImprint database based on three major software components: a MySQL relational database, an Apache Tomcat web server and Java-based backstage processing programs. To guarantee the stability and high performance of the database, a flexible query engine was developed using a Java web application framework (Apache Struts2) and a persistence framework (iBATIS), which automates the mapping between MySQL databases and objects in Java. We also developed two pragmatic online tools to analyze imprinted genes, one for functional enrichment of imprinted genes and the other to analyze the differences in epigenetic modification among multiple tissues or cells. Asynchronous JavaScript and XML (AJAX) and JavaServer Pages (JSP) were used to build the ImprintedGeneBrowser for visualization of information related to imprinted genes, such as imprinting states, imprinting control region, allelic-specific methylation, and imprinting-related diseases.

Table 1. MetaImprint data content and statistics (1 January 2014)

<table>
<thead>
<tr>
<th>Data content</th>
<th>Data statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Imprinted gene</td>
<td>Human Mouse Dog Cow Sheep Pig Rabbit Rat</td>
</tr>
<tr>
<td>Total</td>
<td>317 146 1 35 3 29 1 7</td>
</tr>
<tr>
<td>Experimentally validated (I)</td>
<td>65 132 1 27 3 20 1 6</td>
</tr>
<tr>
<td>Predicted (P)</td>
<td>200 1 0 0 0 0 0 1</td>
</tr>
<tr>
<td>Loss of imprint (LOI)</td>
<td>21 4 0 0 0 0 0 0</td>
</tr>
<tr>
<td>Not imprinted (NI)</td>
<td>31 7 0 8 0 9 0 0</td>
</tr>
<tr>
<td>Genome information</td>
<td></td>
</tr>
<tr>
<td>RefGene</td>
<td>hg19 mm9 canFam3 bosTau7 oviAri1 susScr3 oryCun2 rn4</td>
</tr>
<tr>
<td>RefGenome</td>
<td>23,850 23,427 1327 13,584 710 3950 1301 16,849</td>
</tr>
<tr>
<td>DNA methylation</td>
<td>19 10 0 0 0 0 0 0</td>
</tr>
<tr>
<td>Histone modification</td>
<td>21 18 0 0 0 0 0 0</td>
</tr>
<tr>
<td>Allele-specific methylation</td>
<td>5 0 0 0 0 0 0 0</td>
</tr>
<tr>
<td>Disease</td>
<td>344 83 0 18 0 0 0 0</td>
</tr>
<tr>
<td>Other integrated data</td>
<td>Repeat SNP</td>
</tr>
<tr>
<td>10 10 10 10 10 10 10 10</td>
<td></td>
</tr>
<tr>
<td>11 1 0 0 0 0 0 1</td>
<td></td>
</tr>
</tbody>
</table>
Database usage and access

Overview of the usage of the MetaImprint database

MetaImprint is available at http://bioinfo.hrbmu.edu.cn/MetaImprint. The database supports four basic operations: search, browser, analysis and download. As shown in Fig. 1C, users can obtain the imprinted genes and related annotations by exercising user-specified options. The query results are listed as a table with links to the Imprinted Gene Browser for visualization, and links to download for further analysis. Two analysis tools were developed to identify epigenetic differences of imprinted genes among different tissue/cell types and for functional enrichment analysis of imprinted gene sets. All of the datasets hosted in the MetaImprint database can be downloaded to local servers for further analysis. Novel imprinted genes and their related information can be directly uploaded into the database using the ‘submit’ option to keep MetaImprint up to date.

Querying imprinted gene records

MetaImprint provides a flexible search engine based on a MySQL backend. The starting point of MetaImprint is the search page, which provides three entries including the Imprinted gene search, Disease search and Imprinted Gene Browser. Each search entry provides several query parameters, such as species, gene name/ID, genomic location, chromosome band, expressed parental allele and disease type. The more query parameters used to query imprinted genes, the more precise the query results are. Quick search results are shown in Fig. 2A. The links on these genes would bring users to the detailed information for the corresponding imprinted genes.

As shown in Fig. 2B, the imprinting states of a gene are divided into four states (see Materials and Methods). Moreover, the related Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway, gene ontology (GO) terms and disease information for the imprinted genes are also provided in MetaImprint. The research courses of imprinted genes include gene names, genome locations, EnsembleID, imprinting states, allele-specific expression, tissue-specific types, measurement methods and relevant publications shown in chronological order. As shown in Fig. 2B, the summary information for the human imprinted gene IGF2 consists of six kinds of information. (1) Its genome location is chr11: 2150346-2160204 (hg19). (2) Its Ensembl (space) ID is ENSG00000167244.
(3) Its imprinting states are imprinted (I) according to 46 related publications, including 39 records supporting maternal-origin-specific expression, only seven supporting paternal-origin-specific expression and 22 supporting biallelic expression. (4) The KEGG and GO terms related to IGF2. (5) The studies of IGF2 as an imprinted gene are from 1994 in human. (6) IGF2 is related to some diseases, including medulloblastomas, hepatocarcinomas, neural tube defects and Silver Russell syndrome (SRS). All the query results can be viewed by the Imprinted Gene Browser and downloaded to a local computer for further analysis.

**Imprinted gene browser**

A user-friendly and configurable genome browser, Imprinted Gene Browser, was developed to view the query results for imprinted genes intuitively. Imprinted Gene Browser is linked to the MySQL backend to display the genetic and epigenetic features of imprinted genes, including the upstream and downstream genetic information, repetitive elements, SNPs, ncRNAs, ICRs, CpG islands, DNA methylation and histone modifications. As shown in Fig. 2C, the Imprinted Gene Browser provides options to zoom through the predefined genomic regions or gene ID/symbols. Users can view a comprehensive landscape of each imprinted gene along with the epigenetic and genetic features in the Imprinted Gene Browser. Users can also review the imprinted genes that are located in specific chromosome regions, such as a given chromosome band.

**Analysis tools**

MetaImprint also provides two analysis tools: the Epigenetic Changes Analysis tool and the Functional Enrichment Analysis tool (Fig. 2D). As is well known, DNA methylation mediates imprinted gene expression by passing an epigenetic state across generations and differentially marking specific regulatory regions on maternal and paternal alleles. The dynamics of epigenetic modification may be indispensable for genomic imprinting, which is a feature of mammalian development (Liu et al., 2013). Therefore, DNA methylation and histone modification information may be useful to uncover the mechanisms by which epigenetic modification mediates the allele-specific expression of imprinted genes. The Epigenetic
Changes Analysis tool is based on analysis of variance (ANOVA) to identify the epigenetic differences of the average histone modification signals or the average DNA methylation levels of imprinted genes among different cells or tissues. The results list the information of RefSeq genes, repetitive elements, SNPs and ncRNAs. The epigenetic features and $P$-values of ANOVA can be adjusted by the users when necessary. Moreover, the Functional Enrichment Analysis tool supported by the Database for Annotation, Visualization and Integrated Discovery (DAVID) application programming interface (API) (Dennis et al., 2003) is provided for users to identify the enriched biological functions for any given lists of imprinted genes. And users can also submit a list of gene (NCBI gene ID/gene symbol) to identify enriched biological function based on KEGG and GO.

The applications of MetaImprint in the systematic analysis of imprinted genes

Based on the integrated information of imprinted genes in MetaImprint, we studied the functions of imprinted genes. As an example, 317 human imprinted genes in MetaImprint were analyzed for functional enrichment using the Functional Enrichment Analysis tool. This revealed that these imprinted genes are significantly enriched in biological processes, including embryonic morphogenesis, embryonic organ development, pattern specification processes, etc. (Fig. 3A), indicating the important roles of these imprinted genes in development. In addition, numerous genetic diseases are associated with imprinting defects, including Beckwith–Wiedemann syndrome (BWS) and multiple cancers (Ishida and Moore, 2013; Surani et al., 1990). It has been reported that imprinting is related to the evolution of the placenta in mammals and defects in genome imprinting have been associated with several diseases. From the disease information of imprinted genes in MetaImprint, we found that 99 of the 317 human imprinted genes analyzed are associated with disorders (Fig. 3B). For example, imprinted genes IGF2 and H19 are related to 38 diseases. We also found that 84 diseases are induced by at least one imprinted gene (Fig. 3C). It should be noted that 24 imprinted genes are associated with Wilms’ tumor. Furthermore, we built an imprinting disorder network and disorder-related gene network to investigate the potential relationships between imprinted genes and various diseases (Fig. 3D).

This analysis showed that imprinted genes participate in many kinds of cancer, including Wilms’ tumor and bladder cancer, which is consistent with previous findings of loss of imprinting in a large variety of human tumors (Jelinic and Shaw, 2007). Many imprinted genes tend to participate jointly in one disease, indicating their co-regulation in human disorders.

Furthermore, we used the Epigenetic Changes Analysis tool to analyze the methylation difference of human imprinted genes between liver and hepatic carcinoma. Among 317 human imprinted genes, 253 genes with available data were analyzed, and 97 were identified as differentially methylated genes (DMGenes) (Fig. 4A). Based on the disease information in MetaImprint, 11 of these
DMGenes are related to hepatic carcinoma (Fig. 4B). For example, a previous study reported the loss of imprinting in IGF2, which encodes a member of the insulin family of polypeptide growth factors involved in development and growth, influencing liver tumorigenesis (Haddad and Held, 1997). It has been reported that CDKN1C is abnormally expressed in human hepatocarcinomas, but without mutations (Bonilla et al., 1998; Schwienbacher et al., 2000), indicating the potential roles of methylation difference in regulating the differential expression of this gene between the liver and hepatic carcinoma (Fig. 4D). The other DMGenes may also be associated with the development of human hepatic carcinoma and further analysis of these genes is necessary.

DISCUSSION

Genomic imprinting is a common epigenetic phenomenon in mammals. Allele-specific expression and DNA methylation are two well-known characteristics of genomic imprinting involved in development and complex diseases. For promoting the application and understanding of imprinted genes, we developed a comprehensive cross-species database, MetaImprint, which is focused on the systematic study of imprinted genes in the meta-view, including genetics, epigenetics and functional genomes. The latest release of MetaImprint incorporates 539 imprinted genes from eight species, and related genome-wide genetic and epigenetic information, including ICRs, SNPs, repetitive elements, non-coding RNAs, DNA methylation and histone modifications, as well as human imprinting disorders.

The erasing, building and maintaining of gene imprinting are important in the normal development of embryos. Tissue and cell specificity of imprinted genes have been investigated extensively (Yamasaki-Ishizaki et al., 2007). Appropriate expression of imprinted genes is important for normal development. The state transformation of imprinted genes is associated with development, tumors, metabolic diseases and congenital diseases. For example, Xin et al. found that dynamic expression of imprinted genes is associated with maternally controlled nutrient allocation during maize endosperm development (Xin et al., 2013). However, abnormal imprinting may cause various human disorders including Wilms’ tumor and other cancers (Chao and D’Amore, 2008). The disease information also confirmed the potential roles of imprinted genes in human diseases. Based on the information for imprinted genes in this database, we provided an example of its application for studying regulatory mechanisms and potential roles in development and diseases. MetaImprint may be a useful analytical tool to facilitate the systemic mining of related features of imprinted genes. The in-depth study of genomic imprinting facilitated by MetaImprint will be valuable for understanding disease mechanisms, such as carcinogenesis.

Recently, high-throughput sequencing data, including single-cell transcriptomes and MeDIP-based methylomes, have been generated from a genome-wide perspective to study gene imprinting of different tissue/cell/disease types. The re-evaluation of imprinted genes in mammals using the cumulative research over the last few decades should reveal hidden clues to the genomic homology of chromosome imprinting regions. Going forward, many novel imprinted genes from a variety of species and the information related to the imprinted genes should accumulate much faster, which highlights the importance of a comprehensive database such as MetaImprint. The allele-specific expression of noncoding RNA, for instance AK003491 (Zeng et al., 2013), has also been included in MetaImprint. The database is a useful information repository for hosting further imprinted genes. As new studies of imprinting are reported, MetaImprint will be updated regularly. In future, more information on imprinted genes in different mammals will be integrated into the database, which should help the systematic analysis of the dynamics of gene imprinting during...
mammalian development and the potential roles of abnormal imprinting in diseases.

MATERIALS AND METHODS

The collection of imprinted genes

To mine the imprinted genes from the literature, we searched the PubMed database using the keywords ‘imprint OR imprinted OR imprinting’ and obtained 18,696 publications. We then scanned each publication and removed 11,163 that were not gene-related studies. From the remaining 7533 publications, we manually excavated imprinted genes according to the description of the imprinting state in each study. The imprinting state of each imprinted gene was divided into four classes: (1) imprinted states (I): the description of the gene’s imprinting states was mentioned as imprinted; (2) not imprinted states (N): the gene was not detected as imprinted by the authors; (3) loss of imprint (LOI): the imprinting states were lost or abnormal in multiple disorders, such as Prader–Willi syndrome (PWS) and Angelman syndrome (AS); and (4) imprinted states by prediction (P): the imprinting states of the genes were predicted by computational methods.

The genetic and epigenetic annotation information

MetaImprint incorporated the genomic and epigenomic information related to the imprinted genes from reference literatures and public databases. The genetic annotation features (imprinted gene names, genomic location, repetitive elements and chromosome bands data) were downloaded from the UCSC Table Browser (Karolchik et al., 2004) and NCBI’s Entrez gene (Maglott et al., 2011). The CpG islands of mammalian genomes were identified by CpG_MI (Su et al., 2010) and downloaded from the UCSC Table Browser. The SNP data were downloaded from The International HapMap Consortium (2003) and the UCSC Table Browser. The ncRNA data were obtained from NONCODE (Liu et al., 2005). More than 30 ICRs for human and mouse were collected manually from the literature. KEGG IDs (Kanehisa, 2002) and GO terms (Ashburner et al., 2000) were obtained from their web servers or the UCSC Table Browser and preprocessed using a custom Perl program. The high-throughput epigenetic data for DNA methylation and histone modifications in multiple cell types of human and mouse genomes were primarily collected from NCBI Epigenomics (Fingerman et al., 2011), DiseaseMeth (Lv et al., 2012), Human Histone Modification Database (HHMD) (Zhang et al., 2010), DevMouse (Liu et al., 2011) and GEO (Barrett and Edgar, 2006). All these genetic and epigenetic data are available to download from the ftp server of MetaImprint.

Information related to imprinting disorders

Literature mining was performed to meticulously collect the associations between imprinted genes and approximately 70 human diseases (supplementary material Table S1). These records could be valuable references to analyze the potential function of disorganized imprinting in diverse diseases.

Construction of imprinted genes and human disease network

Using the information regarding the relationships between human imprinted genes and diseases in MetaImprint, we used Cytoscape (Shannon et al., 2003) to build an imprinted genes and human disease network. In the human disease network, the nodes are diseases and two diseases are linked if they share at least one gene. In the imprinted genes network, the nodes are imprinted genes, and two genes are linked if they participate in the same disease.

Acknowledgements

We acknowledge Shanshan Jia and Chunlong Zhang from our laboratory, and Jiang Zhu from Harbin Medical University Daqing Campus for their comments.

Competing interests

The authors declare no competing financial interests.

Author contributions

Y.Z. and Q.W. developed the concepts. Y. Wei built the database. J.S. and Hongbo Liu performed the data analysis. J.L., F.W., H.Y., Y. Wen and Hui Liu collected the information on imprinted genes. Y. Wei, J.S. and Hongbo Liu prepared and edited the manuscript.
RESEARCH ARTICLE


