Chromosome-wide identification of novel imprinted genes using microarrays and uniparental disomies

Citation for published version:

Digital Object Identifier (DOI):
10.1093/nar/gkl461

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
Nucleic Acids Research

Publisher Rights Statement:
© 2006 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Chromosome-wide identification of novel imprinted genes using microarrays and uniparental disomies

Reiner Schulz, Trevelyan R. Menheniott, Kathryn Woodfine, Andrew J. Wood, Jonathan D. Choi and Rebecca J. Oakey*

King’s College London, School of Medicine at Guy’s, King’s College and St. Thomas’ Hospitals, Department of Medical and Molecular Genetics, 8th Floor Guy’s Tower, London SE1 9RT, UK

Received March 21, 2006; Revised May 26, 2006; Accepted June 15, 2006 GEO accession no GSE4870

ABSTRACT
Genomic imprinting refers to a specialized form of epigenetic gene regulation whereby the expression of a given allele is dictated by parental origin. Defining the extent and distribution of imprinting across genomes will be crucial for understanding the roles played by imprinting in normal mammalian growth and development. Using mice carrying uniparental disomies or duplications, microarray screening and stringent bioinformatics, we have developed the first large-scale tissue-specific screen for imprinted gene detection. We quantify the stringency of our methodology and relate it to previous non-tissue-specific large-scale studies. We report the identification in mouse of four brain-specific novel paternally expressed transcripts and an additional three genes that show maternal expression in the placenta. The regions of conserved linkage in the human genome are associated with the Prader–Willi Syndrome (PWS) and Beckwith–Wiedemann Syndrome (BWS) where imprinting is known to be a contributing factor. We conclude that large-scale systematic analyses of this genre are necessary for the full impact of genomic imprinting on mammalian gene expression and phenotype to be elucidated.

INTRODUCTION
Genomic imprinting refers to a specialized form of epigenetic gene regulation whereby the expression of a given allele is dictated by its maternal versus paternal origin. Imprinting plays a crucial role in mammalian reproduction and imposes an absolute requirement for maternal and paternal genomes for the generation of viable offspring (1,2). Approximately 80 imprinted genes have been identified in mouse (http://www.mgu.har.mrc.ac.uk/research/imprinted/), and around 40 in human (3). The extent of imprinting in mouse or human is not fully known, but has been estimated in mouse to range between 100 and 600 genes (4–6). Misregulation of the imprinted genes has been associated with growth and developmental abnormalities in mice (1,2,4), birth defects (7,8) and neoplasias in humans (9) as well as abnormalities in cloned mammals (10,11).

The rationale for the identification of new imprinted genes is based on further understanding imprinted gene regulation and the genetic components of developmental phenotypes in the mouse and birth defects in humans, such as Prader–Willi Syndrome (PWS), Beckwith–Wiedemann Syndrome (BWS) and cancer. Imprinted genes are frequently associated with asynchronous DNA replication (12) and epigenetic control mechanisms that can extend over large genomic regions. Studying aspects of epigenetic regulation requires examination of the genomic environment, especially in cases of placenta-specific imprinting where evidence suggests that not only DNA methylation but also histone modification is important for imprinted gene expression (13,14). Thus an inventory of monoallelic expression of genes and transcripts in a given region is advantageous.

Imprinted genes are characteristically differentially expressed in mice with a maternally derived uniparental duplication or disomy (matUpDp/UpD) compared to mice where the same UpDp/UpD is of paternal origin (patUpDp/UpD). Ideally, expression of an imprinted gene transcribed from the maternally inherited allele will increase 2-fold in a matUpDp sample compared to wild-type (wt) and will be undetectable in a patUpDp sample, and vice versa. In contrast, the expression of a non-imprinted gene is not expected to differ. Strains of mice carrying reciprocal and Robertsonian translocation chromosomes have been used to produce progeny where both copies of a particular chromosome or chromosomal region have been inherited from only the mother or the father. Here, the reciprocal translocation mouse strain T65H (15) was used to generate progeny with UpDps of Chromosomes (Chrs) 7 and 11 either proximal

*To whom correspondence should be addressed. Tel: 020 7188 3711; Fax: 020 7188 2585; Email: rebecca.oakey@genetics.kcl.ac.uk

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

© 2006 The Author(s).
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
or distal of the translocation breakpoint as described previously (16). These mice provide the basis for both the computational and molecular studies presented here. They are viable until birth, so that differential gene expression profiles over a range of different tissues were investigated. Robertsonian translocation mouse strains that generated UpDs of Chr 18 (17) and Chr 12 (18) were used to gain additional statistical power for the evaluation of this method.

To a limited degree, the effectiveness of this method has already been demonstrated by the identification of a novel brain-specific imprinted gene, Inpp5f_v2 (19). Here, we report on the results of systematically refining our approach and extending its coverage to the whole of Chr 7 and Chr 11. We describe this methodology, quantitatively evaluates its effectiveness and compare it to previous non-tissue-specific whole genome studies, reviewed in (20). This study has identified and validated four novel brain-specific paternally expressed transcripts and three placenta-specific maternally expressed genes.

**MATERIALS AND METHODS**

**Tissue sources**

Mouse: the tissue sources for the T65H translocation strain are described in (16). Essentially, newborn mice with maternal and paternal duplications for specific regions of Chrs 7 and 11 were generated using the T65H translocation (21,22). The tissues selected are shown in Table 1 and include 13.5 dpc embryo and placenta, newborn brain, carcass, heart and liver for matUpDp prox 7 versus patUpDp prox 7 (the same samples also generate data for patUpDp prox 11 versus matUpDp prox 11). MatUpDp distal 7 samples were compared to normal sibling samples at 13.5 dpc since the patUpDp distal 7 embryo is not viable at this stage of development. 8.5dpc embryos with maternal versus paternal UpD for Chr 18 were also compared using a Robertsonian translocation. 8.5dpc embryos with maternal versus paternal duplications for specific regions of Chrs 7 and 11 were generated using the T65H translocation strain and liver for matUpDp prox 7 versus patUpDp prox 7 (the same samples also generate data for patUpDp prox 11 versus matUpDp prox 11). MatUpDp distal 7 samples were compared to normal sibling samples at 13.5 dpc since the patUpDp distal 7 embryo is not viable at this stage of development. 8.5dpc embryos with maternal versus paternal UpD for Chr 18 were also compared using a Robertsonian translocation Chr Rb(2.8)2Lub(7.18)9Lub or RB92.82Lub (R) and C57BL/6JEi-Rb(7.18)9Lub (B) strains obtained from the Cytogenetic Models Resource at the Jackson Laboratory (17). Placentae were dissected free from the decidua. However, due to the structure of the placenta with invading maternal vessels, some maternal material is likely to be included in these preparations. 15.5 dpc embryos and placentae from Chr 12 UpD (18) were obtained in collaboration with Anne Ferguson-Smith. Human: for DHCR7 and AMPD3, anonymous placenta DNA and RNA samples were collected in collaboration with Dr B.S. Emanuel (CHOP) in accordance with ethical guidelines.

**Microarray protocols**

Affymetrix Genechips™ U74v2 and 430v2 microarrays were used (Table 1). The U74v2 triple array series represents ~36 000 genes and ESTs. The U74v2 series represents all sequences (~6 000) in the Mouse UniGene database (Build 74) that had been functionally characterized at the time plus EST clusters. The 430v2 dual array series probes ~39 000 genes and ESTs where the 430Av2 array predominantly represents well-characterized genes.

Caesium chloride isolated total RNA from samples in Table 1 were quantified on an Agilent Bioanalyser™ and 5–7 μg was used to prepare biotin-labelled cRNA target essentially as in the Affymetrix™ expression manual (23). Biotinylated cRNA targets were purified using cRNA Cleanup spin columns (Affymetrix™), fragmented in 5x fragmentation buffer (Affymetrix™), and quantified prior to hybridization on an Agilent Bioanalyser™. A total of 0.15 μg of labelled probe was hybridized per GeneChip expression array followed by washing and staining on the Affymetrix™ fluidics station 450. An Affymetrix Scanner 3000 was used to quantitate the signal.

**Gene chip operating software (GCOS) data analysis**

Data from the scanner were analysed using the Affymetrix™ MASv5 or GCOS software. The software computes the signal for each pair of corresponding PM and MM probes in a probe set as $S_i = \log_2(\text{PM}_i) - \log_2(\text{MM}_i)$ where PM and MM are the measured fluorescence levels of the $i$th probes in the PM and MM sets. In cases where $S_i$ would barely exceed the background noise level or be negative, GCOS makes adjustments to ensure a conservative and positive estimate of genespecific hybridization. GCOS arrives at a single signal value (S) for each probe set by calculating the Tukey-biweight of the $S_i$, a weighted outlier-resistant statistic. GCOS reports $2^\Delta$ as the estimate of the absolute expression level of the gene represented by the probe set. GCOS allows a comparative analysis of gene expression in two samples, each used to hybridize a separate microarray. GCOS computes a signal $2^\Delta$-ratio (SLR) for each probe set as a measure of the degree of differential expression between the two samples and a $P$-value as a measure of confidence in any measured difference in expression. In the most common case, the SLR equals the Tukey-biweight over $i$ of $\log_2(\text{PM}_i) - \log_2(\text{MM}_i) - [\log_2(\text{PM}_i) - \log_2(\text{MM}_i)]$; where $\text{PM}_i$, e.g., is the measured fluorescence of the $i$th PM probe in a particular probe set on array A (hybridized with sample A). The $P$-value is based on a Wilcoxon signed rank test applied to the pairs of differences $[\log_2(\text{PM}_i) - \log_2(\text{MM}_i), \log_2(\text{PM}_i) - \log_2(\text{MM}_i)]$ and $[\log_2(\text{PM}_i) - A, \log_2(\text{PM}_i) - B]$, where

Table 1. Overview of the microarray data used in this study

<table>
<thead>
<tr>
<th>UpDp/UpD</th>
<th>Microarray Series</th>
<th>Model</th>
<th>Tissue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prox Chrs 7 and 11</td>
<td>Affymetrix™ U74v2</td>
<td>A</td>
<td>x</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Dist Chrs 7 and 11</td>
<td>Affymetrix™ U74v2</td>
<td>A</td>
<td>x</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Chr 12</td>
<td>Affymetrix™ U74v2</td>
<td>A</td>
<td>x</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Chr 18</td>
<td>430v2</td>
<td>A</td>
<td>x</td>
</tr>
</tbody>
</table>

An ‘x’ indicates the availability of data for a specific combination of UpDp/UpD, microarray and tissue (Pl = placenta, Em = embryo, Cc = carcass, Br = brain, Lv = liver and Ht = heart). Note that for the experiment using a prox Chrs 7 and 11 sample of heart, there was only enough material to hybridize the U74v2 A and B arrays.
A and B are measures of the background noise of the respective array.

Expression analysis

Two assays were used; the first used cDNA prepared from maternal and paternal UpDp and wt RNA samples in RT–PCR assays. Briefly, 4 μg of total RNA was reverse transcribed using Superscript III (Invitrogen™) reverse transcriptase (manufacturers’ instructions) and PCR-amplified using ABgene™ Reddymix and gene-specific primers (Supplementary Table S4). The second method used inter-subspecies hybrid RNA as described in (19). Essentially, single nucleotide polymorphisms (SNPs) between *Mus m. musculus* [C57Bl6J (B6)] and *Mus m. castaneus* (CAST) strains were identified and assayed by DNA sequencing (primers are listed in Supplementary Table S4). RT–PCR followed by sequencing identified the expressing allele(s) for each gene tested (Figures 2 and 3). Sequencing was performed on an ABI 3730 DNA analyser using ABI BigDye™ reagents (manufacturer’s protocols). Fetal or parental human DNAs were sequenced for expressed SNPs, which were assayed as described above.

Gene prediction programs

Gene predictions tools used include Ensembl (24–26), SGP (27), Geneid (28), Genscan (29), Superfamily/SCOP (30), Twinscan (31) and Augustus (32).

RESULTS

The ‘one chromosome at a time’ approach excludes most false positives

Each of the microarrays used here measures the expression of thousands of distinct transcripts, all of which can potentially change in expression levels between the respective matUpDp/UpD and patUpDp/UpDs. However, most transcripts can be ruled out based on their location within the genome that places them outside the selected chromosomal duplication or disomy.

Four chromosome anomalies were investigated (Table 1). Filtering by genomic position reduced the number of potential false positives by between 85 and 94% (Table 2). For example, of the 45 037 distinct non-control probe sets, representing over 39 000 transcripts, on the Affymetrix™ 430v2 arrays, only 6595 (14.6%) generate a match to proximal Chrs 7 or 11 when their target sequences are aligned with the genome (March 2005 NCBI build 34) using BLAT (33). The remaining 38 442 probe sets do not pass the henceforth called ‘map filter’ for this UpDp and thus can be excluded from the search. On average, about a tenth of the transcripts represented on the microarrays mapped to the respective region of UpDp/UpD in our experiments and could legitimately change in expression between the matUpDp/UpD and pat UpDp/UpD samples as a direct result of being imprinted. This reduction of potential false positives by an order of magnitude compared to the analogous whole genome approach (parthenogenote versus androgenote), is compounded by the more limited downstream effects of UpDp/UpDs.

Gains in sensitivity by combining differential expression measures

Arrays hybridized with the matUpDp/UpD sample were compared to arrays hybridized with the patUpDp/UpD sample using the gene operating software (GCOS) that provides two measures of differential expression (34). The SLR of a probe set captures the fold change in expression of the represented transcript between the two samples, i.e., in simplified terms, $\text{SLR} = \log_{2}(\text{expression in matUpDp/UpD / expression in patUpDp/UpD})$. In addition, GCOS supplies a $P$-value that expresses the confidence in and direction of any measured expression difference, that is, $P = 0.5$ if there is no evidence for a change in expression, and $P$ is the closer to 0 or 1 the more statistical evidence there is for an increase or decrease in expression, respectively.

$SLR$ and $P$-value are closely correlated but non-redundant measures. Frequently, a probe set’s absolute $SLR$ (|$SLR$|) is large while $0.5 - |P - 0.5| \approx 0.5$, meaning that while a large fold change is measured, there is little statistical evidence for differential expression. The converse happens similarly frequently. Which of the two measures more truthfully represents the actual degree of differential expression is not immediately clear. Previously, $SLR$ and $P$-value have been shown to give qualitatively different results (35). Here, we systematically evaluated the performance of $SLR$, $P$-value and combinations of both, namely, $P$-value-weighted $SLRs$ (xSLR), in the context of imprinted gene detection. Specifically, $xSLR = 0$ if $0.5 - |P - 0.5| \geq x$, and otherwise, $xSLR = [1 - 1/x (0.5 - |P - 0.5|)] SLR$. Within the interval

Table 2. For each combination of microarray series, UpDp/UpD and tissue (Pl = placenta, Em = embryo, Cc = carcass, Br = brain, Lv = liver and Ht = heart) used in this study, the table shows the number of distinct non-control probe sets detecting differential expression that map to within (normal script) versus to only outside (italic script) the UpDp/UpD

<table>
<thead>
<tr>
<th>UpDp/UpD: Array series</th>
<th>Prox Chrs 7 and 11 Number of distinct non-control probe sets</th>
<th>Dist Chrs 7 and 11 Chrs 12</th>
<th>Chr 18</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number of distinct non-control probe sets</td>
<td>Mapping to within (normal) versus to only outside (italic) the UpDp/UpD</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>Total</td>
<td>luSLR &gt; 6 Pl</td>
<td>Pl Em</td>
</tr>
<tr>
<td>430</td>
<td>45037</td>
<td>6595</td>
<td>222</td>
</tr>
<tr>
<td>U74</td>
<td>36701</td>
<td>3679</td>
<td>161</td>
</tr>
</tbody>
</table>

The table also provides the total number of non-control probe sets per microarray series and, for each UpDp/UpD in turn, the total number of probe sets that map to within versus to only outside the UpDp/UpD.
<x> linearly changes from SLR to zero where it remains at all lower levels of statistical confidence. We explored the whole range of possible values for <x>, i.e. \( 0 < x < 0.5 \), and found \( x = 0.01 \) (uSLR) to perform best, closely followed by \( x = 0.05 \) (vSLR). The simple linear weighing of SLR over the whole range of \( P \)-values, \( (x = 0.5; pSLR) \) was not nearly as effective.

For each measure, we first estimated from our microarray data, the true positive rate (TPR) and false positive rate (FPR) for a range of differential expression thresholds. The TPR was estimated from a total of 43 known imprinted genes that map to within the investigated UpDp/UpDs and for which there are representative probe sets. For each tissue, we counted (a) the number of genes within the respective UpDp/UpD that the literature has reported to be imprinted in the tissue and for which tissue-matched array data were available and (b) the number of such genes for which in addition, at least one representative probe set detected an increase (for maternally expressed genes) or decrease (for paternally expressed genes) in expression that met or exceeded the differential expression threshold. The TPR was computed as the ratio between the two sums over all tissues of counts (b) and (a). The FPR is the fraction of probe sets mapping to the investigated UpDp/UpDs that do not represent known imprinted genes but nevertheless met or exceeded the differential expression threshold. Plotting the FPR versus the TPR over the range of decision thresholds yields a receiver operating characteristics (ROC) curve shown for SLR, \( P \)-value, uSLR, vSLR and pSLR (Figure 1).

A low FPR is paramount to achieve a good cost to discovery ratio in the molecular validation. Even with the map filter, an FPR of 5% still corresponds to, in the proximal Chrs 7 and 11 region, roughly 300 distinct probe sets on the 430 array that will be flagged as representing differentially expressed sequences, but are unlikely to represent imprinted transcripts. We therefore chose 5% as the upper limit on the FPR. Up to this limit, uSLR consistently performed best, i.e. gave the highest TPR compared to the other measures that we considered. Consequently, the ROC curve for uSLR dominates the other curves in Figure 1 for an FPR < 5%. Specifically, the differential expression threshold of \( |uSLR| > 0.6 \) delivered a TPR = 65.8% and an FPR = 4.9% so that henceforth, ‘differentially expressed’ will imply \( |uSLR| > 0.6 \).

The map filter saturates candidates with imprinted genes

To prioritize the molecular validation experiments, we ranked the probe sets passing the map filter by degree of differential expression, i.e. in descending order of \( |uSLR| \). On average, 19% (29%, 38%, 46% and 52%) of all genes known to be imprinted in the respective tissue and mapping to within the respective duplication were ranked among the top 10 (20, 50, 100 and 200). Compared to the analogous ranking of all probe sets, i.e. without the prior application of the

\[
0 \leq 0.5 - |P - 0.5| \leq x,
\]

xSLR linearly changes from SLR to zero where it remains at all lower levels of statistical confidence. We explored the whole range of possible values for x, i.e. \( 0 < x \leq 0.5 \), and found \( x = 0.01 \) (uSLR) to perform best, closely followed by \( x = 0.05 \) (vSLR). The simple linear weighing of SLR over the whole range of \( P \)-values, \( (x = 0.5; pSLR) \) was not nearly as effective.

For each measure, we first estimated from our microarray data, the true positive rate (TPR) and false positive rate (FPR) for a range of differential expression thresholds. The TPR was estimated from a total of 43 known imprinted genes that map to within the investigated UpDp/UpDs and for which there are representative probe sets. For each tissue, we counted (a) the number of genes within the respective UpDp/UpD that the literature has reported to be imprinted in the tissue and for which tissue-matched array data were available and (b) the number of such genes for which in addition, at least one representative probe set detected an increase (for maternally expressed genes) or decrease (for paternally expressed genes) in expression that met or exceeded the differential expression threshold. The TPR was computed as the ratio between the two sums over all tissues of counts (b) and (a). The FPR is the fraction of probe sets mapping to the investigated UpDp/UpDs that do not represent known imprinted genes but nevertheless met or exceeded the differential expression threshold. Plotting the FPR versus the TPR over the range of decision thresholds yields a receiver operating characteristics (ROC) curve shown for SLR, \( P \)-value, uSLR, vSLR and pSLR (Figure 1).

A low FPR is paramount to achieve a good cost to discovery ratio in the molecular validation. Even with the map filter, an FPR of 5% still corresponds to, in the proximal Chrs 7 and 11 region, roughly 300 distinct probe sets on the 430 array that will be flagged as representing differentially expressed sequences, but are unlikely to represent imprinted transcripts. We therefore chose 5% as the upper limit on the FPR. Up to this limit, uSLR consistently performed best, i.e. gave the highest TPR compared to the other measures that we considered. Consequently, the ROC curve for uSLR dominates the other curves in Figure 1 for an FPR < 5%. Specifically, the differential expression threshold of \( |uSLR| > 0.6 \) delivered a TPR = 65.8% and an FPR = 4.9% so that henceforth, ‘differentially expressed’ will imply \( |uSLR| > 0.6 \).

The map filter saturates candidates with imprinted genes

To prioritize the molecular validation experiments, we ranked the probe sets passing the map filter by degree of differential expression, i.e. in descending order of \( |uSLR| \). On average, 19% (29%, 38%, 46% and 52%) of all genes known to be imprinted in the respective tissue and mapping to within the respective duplication were ranked among the top 10 (20, 50, 100 and 200). Compared to the analogous ranking of all probe sets, i.e. without the prior application of the
map filter, this corresponds to an average increase in the number of known imprinted genes in the top 10 (20, 50, 100 and 200) ranks by 85% (71%, 57%, 61% and 48%). Table 3 provides the absolute values. This illustrates how the map filter increases the saturation of the top ranks with known imprinted genes and, by extrapolation, with likely truly imprinted candidate genes.

Summary statistics of probe sets representing differentially expressed sequences

Table 2 gives a complete account, broken down by UpDp/UpD, tissue and microarray series, of how many probe sets whose target sequences map to within versus to exclusively outside the UpDp/UpD detected differential expression. On average, 4.9% (5.3%) of the probe sets mapping to within (outside) the UpDp/UpD detected differential expression. The differences underlying these averages are not statistically significant ($P = .083$; Wilcoxon signed rank test), which suggests that proportionately and in terms of differential expression, a UpDp/UpD affects the region of duplication approximately as much as the rest of the genome.

Differential gene expression profiles across tissues

For the UpDp of proximal Chrs 7 and 11, we conducted separate microarray experiments using placenta, embryo, carcass, brain, liver and heart samples (Table 1). Thus, we created tissue-specific differential expression profiles. Only limited or no tissue-specific data were available for the distal UpDp of Chrs 7 and 11 and the UpDs of Chrs 12 (embryo and placenta) and 18 (embryo).

Profiles of 43 known imprinted genes on Chrs 7, 11, 18 and 12 are compared with the literature-reported imprinting status in Supplementary Table S1. Profiles of the 59 genes on Chrs 7 and 11 that were not known to be imprinted and for which we conducted imprinting validation experiments are listed in Supplementary Table S2, where the results of the validation assays appear side-by-side with the microarray data. The profiles of all transcripts on proximal Chrs 7 and 11 (935 transcripts represented by 1705 probe sets) that were differentially expressed in at least one of the tissues are shown in Supplementary Table S3.

Confirmation of seven differentially expressed candidates

Allele-specific assays using UpDp/UpD RNAs have been used extensively to determine parent of origin specific gene expression (36), but could be argued not to be independent of the microarray assay since the starting material is the same (37). Interspecies hybrids have been used to assay for parent of origin specific expression (38–40) but loss of imprinting of some genes has been detected in hybrid crosses (41). So, where possible we have used more than one method. Moreover, we have tested known imprinted genes and have in every case confirmed imprinting status.

A set of 59 candidate genes with an unknown imprinting status were selected and initially screened in an RT–PCR approach using matUpDp versus patUpDp RNA templates (where available). Candidates showing evidence of parental origin specific expression in this assay on the proximal regions of mouse Chrs 7 were subjected to further validation using RT–PCR combined with a SNP allele-specific assay in mouse inter-sub-specific hybrids.

On proximal Chr 11 in whole embryo or isolated brain, eight candidate imprintected transcripts were tested for monoallelic expression by RT–PCR in mat versus patUpDp material (Supplementary Table S2), along with two control known imprinted genes, U2af1-rs1 and Grb10. The control genes showed imprint expression and six out of eight non-control genes were biallelic. Pnpt1 was expressed in paternally duplicated samples only but was not present in wt, Gabral showed some maternal bias, but this difference was not considered robust enough to warrant further study. Lsm11 is located within band B1.1 of Chr 11 containing the T65H translocation breakpoint thus may have been duplicated in the distal rather than the proximal experiment. However, the distal microarray data did not suggest differential expression of Lsm11.

On proximal Chr 7, 25 transcripts were tested for monoallelic expression in brain, 17 in whole embryo, 5 in placenta and 2 in heart (Supplementary Table S2). By UpDp assay, 10 transcripts showed allele bias in brain and were tested further by SNP analysis, which confirmed that four of these (BB077283, BM117114, AK080843 and AV328498) were paternally expressed (Figure 2). BB182944 and BB312372 were also validated as paternally expressed candidates showing evidence of parental origin specific gene expression (38–40) but loss of imprinting of some genes has been detected in hybrid crosses (41). So, where possible we have used more than one method. Moreover, we have tested known imprinted genes and have in every case confirmed imprinting status. On the proximal experiment. However, the distal microarray data did not suggest differential expression of Lsm11.

On proximal Chr 7, 25 transcripts were tested for monoallelic expression in brain, 17 in whole embryo, 5 in placenta and 2 in heart (Supplementary Table S2). By UpDp assay, 10 transcripts showed allele bias in brain and were tested further by SNP analysis, which confirmed that four of these (BB077283, BM117114, AK080843 and AV328498) were paternally expressed (Figure 2). BB182944 and BB312372 were also validated as paternally expressed but were subsequently found to correspond to the imprinted Pec2 and Pec3 transcripts (42) and were excluded. Of the remainder, Ampd3 was paternally expressed in placenta. The placenta UpDp assay showed a maternal bias with some paternal expression (data not shown) rather than the exclusive maternal expression seen with the interspecies SNP assay (Figure 3). BB264453 was robustly differentially expressed
by RT–PCR in mat versus pat UpDp brain, but did not contain a SNP between B6 and cast and so was not considered further. The location of AI114950 within band F4 of Chr 7 makes it analogous to Lsm11 and the distal microarray data did not suggest differential expression of AI114950. The remaining tested transcripts were biallelic by SNP assay. AK080843 and AV328498 are transcribed in the antisense orientation relative to Ube3A and could be part of the large imprinted antisense transcript LNCAT, (43,44) but because the LNCAT cDNA is not clearly defined or publicly available, this is uncertain.

Ten transcripts were selected from distal mouse Chr 7 (Supplementary Table S2). Of these, Th and Dhcr7 were maternally expressed in placenta (Figure 3) but were biallelic in embryo. These transcripts were not as robustly monoallelic as the novel brain transcripts in the proximal region, although the allele preferences exchanged with reversed parental transmission (Figure 3).

Figure 2. Identification of novel paternally expressed transcripts on mouse proximal Chr 7. (A) The 2 Mb PWS/AS orthologous region with positions of the novel paternally expressed transcripts indicated*. Genes are shown as open boxes with the relative transcriptional orientations defined by arrows. Blue, red or no colouring define paternal, maternal or biallelic expression respectively. (B) RT–PCR analysis of candidates identified on the proximal Chr 7 brain array in cDNA derived from patDp prox 7, matDp prox 7 and wild-type brain tissue. Controls for paternal (Snrpn), maternal (Grb10) and biallelic (Igf1r) expression are shown. The molecular weight marker is a 100 bp DNA ladder where the bright band corresponds to 500 bp. Samples treated with and without reverse transcriptase are indicated as + or −RT. ESTs AK080843, BB077283, BM117114 and AV328498 were paternally expressed. (C) Allele-specific RT–PCR analysis of proximal Chr 7 candidates. Newborn brain tissues with expressed SNPs were obtained from reciprocal crosses between M.m.musculus (B6) and M.m.castaneus (CAST) animals. cDNA fragments containing the SNPs were recovered by RT–PCR and direct sequencing to determine allele-specific expression.
Verified imprinted ESTs: gene predictions and genomic context

Gene predictors were applied to the genomic regions containing the ESTs for which validation by RT–PCR and SNP analysis had confirmed imprinting (Figure 2). The genomic position of each EST was uniquely identifiable. In brain and embryo, the paternally expressed AK080843 and AV328498 map to a region between 6.7 and 13 kb centromeric of Snrpn (Figure 2A). Based on the genomic sequence of this region, no prediction program provided evidence for a separate gene or an extended transcript of Snrpn whose coding region aligns with the ESTs. AV328498 is the 3′-read of the full insert sequence AK078094 that partially overlaps AK080843. Both AK078094 and AK080843 partially overlap BC070450, a transcript extending much further 3′. The distinct splicing patterns of these transcripts suggest that they might be alternative transcripts of the same transcriptional unit. Whether these transcripts are extensions of Snrpn is not completely clear but based on the ESTs in the region (including AK080843 and AV328498, AK078094), Ensembl predicts a gene (ENSMUSG00000016158) that is distinct from Snrpn at this location. To further ascertain whether transcripts AV328498 and AK080843 could splice...
onto Snrpn, a Snrpn forward primer was combined with the AV328498 and AK080843 reverse primers (Supplementary Table S4, SNP assay primers) to amplify brain cDNA by RT–PCR. ABGene thermoprime plus Taq polymerase was used to amplify large products of up to 12 kb. No products were seen from any primer combination (data not shown). This provides empirical data to support these transcripts being independent of Snrpn, but they are limited in assaying the absence of a product. Similarly, no genes were predicted that coincide with BB077283 (corresponding 5’-read: BB625859) or BM117114 (no 5’-read available, but identical to BQ555876 with corresponding 5’-read BQ555877). These ESTs map telomeric of Snrpn and centromeric of Ndn (Figure 2A) and are located distant (1.8 Mb; BM117114 and 1.3 Mb; BB077283) from Snrpn and hence are not likely to be part of the Snrpn transcript.

Human orthologues of mouse distal Chromosome 7 genes are not maternally expressed in placenta

We examined the expression status of the human orthologues DHCR7 and AMPD3 in placenta using a combined RT–PCR and SNP assay essentially as performed in the mouse tissues. Both genes were biallelically expressed in human term placenta (Figure 4). The imprinting status of TH was not addressed, since informative polymorphisms could not be found in our sample sets. These genes may however be imprinted in other non-tested human tissues since tissue-specific imprinting need not be the same between species (45).

Limited overlap with previous non-tissue-specific whole genome imprinting studies

A 2-fold change in expression was used as the cut-off for differential expression in RIKEN’s FANTOM2 imprinting screen (46), which identified 2110 imprinting candidate transcripts. A large fraction of these transcripts were represented by probe sets on the Affymetrix™ arrays (430v2: 79%, U74av2: 65%), and between 17 (350) and 41% (870) of them were represented by probe sets mapping to one of the investigated UpDp/UpDs. Given that the above transcripts constitute imprinting candidates, we expected a large fraction of them to be represented by one or more differentially expressed probe sets in our experiments. However, at the most, this fraction was 45% (UpD of Chr 18), and across all UpDp/UpDs, tissues and microarray series’, the average was 20.4%. The overlap between the RIKEN imprinting candidates and differentially expressed probe sets in our experiments increased when a less stringent differential expression threshold was used, permitting a higher FPR. For example, using |sSLR| = 0.5 with a TPR of 74% and an FPR of 21.1% (Figure 1), the overlap increases to 66.6% in the best case (UpD of Chr 18), and to 50.3% on average. This suggests that the |uSLR| ≥ 0.6 is more stringent than the 2-fold change in expression threshold used in the RIKEN study.

Recent work that used bioinformatic methods predicted 600 out of 23 788 annotated (Ensembl) autosomal mouse genes to be imprinted based on their similarity to sequence features surrounding 44 known imprinted genes, and their dissimilarity to ~500 assumed non-imprinted genes (6). The TPR and FPR of the analysis, determined by cross-validation, was 100 and 7%, respectively. Using the same method, the authors also predicted allele preference, that cross-validation showed to be 97.7% accurate. We observed the largest overlap with our study on proximal Chr 7 and 11 to which 27 of the 600 predicted imprinted genes mapped and were represented by probe sets on the Affymetrix arrays. For 10 of these genes there was at least one probe set that detected differential expression in at least one of the investigated tissues, and in three cases, the predicted allele preference consistently agreed with the direction of change detected by the probe sets (Supplementary Table S5).

DISCUSSION

Microarray measurements versus allele-specific assays

In principle, microarrays are an ideal tool for the large-scale detection of imprinted genes in UpDp/UpD material because imprinted genes are expected to exhibit an extreme expression differential between the matUpDp/UpD and patUpDp/UpD samples. Empirically however, the array measurements will lead to both false negatives and false positives due to, among other factors, the sharing of probe sets between multiple alternative transcripts, cross-hybridization and downstream effects of the chromosome anomaly. Most downstream targets can be excluded by map position.
Table 4. Contingency table summarizing Supplementary Table S2 by contrasting the microarray-derived mode of expression (either M = maternal; uSLR \(>\) .6, P = paternal; uSLR \(<\) -.6 or B = biallelic) with the mode determined by the UpDp/UpD PCR and/or SNP sequencing assays

<table>
<thead>
<tr>
<th>Array-based expression mode: Tissue</th>
<th>Mode of expression based on microarray measurement</th>
<th>Paternal</th>
<th>Biallelic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Maternal</td>
<td>P</td>
<td>B</td>
</tr>
<tr>
<td>Pl</td>
<td>14</td>
<td>2</td>
<td>Ampd3</td>
</tr>
<tr>
<td>Em</td>
<td>28</td>
<td>9</td>
<td>An2</td>
</tr>
<tr>
<td>Br</td>
<td>30</td>
<td>1</td>
<td>Gabral*</td>
</tr>
<tr>
<td>Ht</td>
<td>2</td>
<td>1</td>
<td>Shhgl3</td>
</tr>
</tbody>
</table>

For each combination of microarray-derived and assay-derived expression mode, the number of distinct genes that exhibited this combination is provided. Empty entries correspond to zero occurrences. The gene names are given except for genes that showed a biallelic mode of expression in both the microarray and assay experiments. New imprinted genes/transcripts are in bold. *indicates no SNP between Bb and Cast.

However some false positives are retained, and unless a gene has been verified by allele-specific assay, its imprinting status remains uncertain. Thus, the fine-tuning of the true and especially the FPR in microarray data analysis is of paramount importance in a screen, such as this and has been a focus of this study.

Probe sets representing differentially expressed sequences (uSLR \(>\) .6) were present for 40 of the 59 genes that were tested for imprinting using the UpDp and/or the SNP assay. Nineteen transcripts were assayed based on past analyses using outdated Affymetrix™ software and annotation (MAS4). For individual tissues, Table 4 shows the total number of transcripts for individual tissues that were tested using an allele-specific assay and compares the results with the microarray measurements. The control gene, H19, showed robust maternal expression in the SNP assay (Figure 3). However, contrary to expectation, on examining the differential expression on the embryo and placenta distal matUpDp versus wt microarrays, H19 had similar expression levels in both samples (Supplementary Table S1). In contrast, Igf2, was expressed in wt and vastly reduced in matUpDp RNA as would be predicted (Supplementary Table S1). The probe sequences for H19 were examined but there was no evidence for cross-hybridization in the Affymetrix™ annotation. Since H19 is a very highly expressed transcript in the cell, the fluorescent signal may have saturated, but this was not the case because GCOS excludes saturated readings from the comparative analysis, and for H19, all probes were included. Hence this false negative result may be an artifactual problem, but could also be due to a biological modulation of H19 RNA levels in the matUpDp samples by an unknown mechanism.

More than 66% of the 40 genes with microarray evidence for differential expression were found to be biallelic in the UpDp and/or SNP assays (Table 4). This is in stark contrast to the FPR of 5% that we estimated for our threshold for differential expression (uSLR \(>\) .6; Figure 1). However, this estimate is with respect to differential expression detected by a single probe set. Each of the tested 59 genes is on average represented by four probe sets that were each used for measurements in at least two different tissues. Assuming independence, the probability of one of these probe sets detecting differential expression in one of the tissues by chance is roughly one-third. However, only 9 of the 31 genes detected as differentially expressed by their probe sets subsequently exhibited an allele preference in the UpDp assay. For a per-gene and cross-tissue FPR of 1/3 the expected number is 20. The UpDp assay is not strand-specific, so an antisense transcript could mask an allele bias in the UpDp assay, but this is unlikely to explain the ~50% shortfall in genes with an allele bias. The probe sets that apparently mistakenly indicate differential expression do not obviously share any characteristics that distinguish them from ‘correct’ probe sets. So, a large fraction of the inconsistencies between the microarray and the UpDp assay results remain unexplained.

Limited overlap with previous whole genome imprinting studies

Previous studies report that the genome-wide identification of imprinted genes via differential expression between parthenogenotes and androgenotes likely suffers from a high FPR...
due to the extreme and different downstream effects of partheno- and androgenesis (39,47). This is likely to explain our limited overlap with the results of the RIKEN FANTOM2 imprinting study (46), especially since relaxing our threshold for differential expression increased the overlap significantly.

The small overlap between our results and the genes predicted to be imprinted in (6) has several possible explanations. The predictor may have misclassified imprinted genes for which the training set was not representative, e.g. imprinted genes with a distinct regulatory mechanism that would have made the sequence features appear atypical for an imprinted gene. For a given gene prediction, it is still unknown in which tissue(s) and developmental stages the imprinting occurs. Our microarray measurements may not have covered the relevant tissues or developmental stages.

A general difficulty in establishing a closer correspondence between the classifier and our microarray-based approach is the classifiers’ limited applicability to characterized genes, while the microarrays contain a large number of EST-complementary probe sets.

Maternally expressed genes in placenta

Two of the three genes found to be maternally expressed in the placenta, Dhcr7 and Th are located within or in close association with the BWS orthologous region on mouse distal Chr 7. The third, Ampd3, is not associated with the cluster and maps ∼33 Mb centromeric of H19 in apparent isolation from any other imprinted gene. The majority of genes contained within the respective human and mouse BWS regions are highly conserved, both in structural organization and, with few exceptions, in imprinting status (48,49). We could not address the imprinting status of TH in humans due to the absence of informative polymorphisms in the available tissues. DHCR7 and AMPD3 were biallelic in placenta, a finding consistent with their respective locations on 11q13 and 11p15.4, neither region being associated with imprinting (Figure 4).

The boundaries of the cluster have been defined by the maternally expressed H19 and Osbpl5 genes (48–50). The finding of preferential maternal expression at Dhcr7, which is located outside these arbitrary boundaries could extend this region of imprinting. Current evidence supports regional imprinting control in the BWS cluster by two imprinting centres (IC’s), the H19 differentially methylated domain (DMD) (51,52), and the Kcnq1 (Kv) DMR1 (40,53). Paternal inheritance of a KvDMR1 deletion leads to loss of imprinting of several maternally expressed imprinted genes with disruption of the repressive paternally expressed Kcnq1ot1 antisense transcript thought to be responsible for this effect. Though not formally addressed here, it is conceivable that the maternal expression of Th and Dhcr7, both of which are located adjacent to KvDMR1 could be regulated by this region, or by the Kcnq1ot1 transcript in cis. Analysis of both genes within the context of a KvDMR1 deletion allele could address this expectation.

One of the limitations of examining monoallelic expression in placenta compared to other organs is the presence of invading maternal blood vessels. The allele-specific assay in interspecies hybrids were performed on placenta with the deciduum removed but the precise fetal:maternal contribution is not known. Some maternal expression could be derived from maternal contamination. The maternal and paternal UPD placentae would not be susceptible to this issue since it applies equally to these tissues harvested for the arrays. Thus the arrays provide independent evidence for differential expression. Maternal allele preference observed in these genes is largely consistent with the directionality of imprinting of other known placental imprinted genes, an exception provided by the Igf2 P0 transcript, expressed exclusively from the paternal allele in labyrinthine trophoblast (54,55).

While Ampd3 shows almost exclusive expression from the maternal allele, for Dhcr7 and Th, transcription was also evident (albeit at a much lower level) from the paternal allele, which indicates that maternal expression of these genes may not be absolute. On the other hand we note the caveat of alternative (non-imprinted) transcripts or imprinting in a cell lineage-dependent manner that potentially complicates this interpretation. However it has not escaped our notice that similarly ‘incomplete’ imprinting has been observed for other placenta-specific genes, Nap1l4, Pbla2 and Osbpl5, located immediately centromeric to Dhcr7 (48,49). Incomplete silencing of the paternal allele in these examples, it has been argued, reflects their relatively distant separation from the KvDMR1 compared with other genes, Kcnq1, Kcnq1ot1 and Cdkn1c, that are more closely associated with this element and robustly imprinted (49), providing a possible mechanistic explanation for these observations.

Imprinted EST transcripts identified in the PWS/Angelman Syndrome region

PWS is thought to arise as a consequence of the loss in expression of several paternally expressed imprinted genes on human 15q11–13. Studies of transgenic models have not revealed an obvious candidate, though the Ndn gene, when disrupted in mice, causes failure to thrive, a frequent observation in PWS (56). The characterization of additional imprinted transcripts within this region could therefore contribute to the further genetic dissection of PWS. Four novel imprinted transcripts AK080843, BM117114, BB077283 and AV328498 were identified in the PWS/AS orthologous region on proximal Chr 7. Bioinformatic analysis did not reveal evidence that these transcripts were contained within larger transcription units or possessed a capacity for protein coding. Significantly, human orthologues could not be found for two transcripts (BM117114 and BB077283), suggesting that they do not have a role in PWS. For AK080843, a 100 bp sequence shares 96% identity in the human genome although extensive RT–PCR analysis failed to detect transcripts from this region (T. R. Menheniott, unpublished data).

Unifying hypotheses to explain imprinting disorders will require a comprehensive mapping of genes in the pertinent critical regions. Methods permitting the global detection of imprinted genes across multiple developmental lineages are likely to shed light upon the role of imprinting processes in such disorders. Indeed a prominent conclusion of this study is that the total number of imprinted genes is likely to exceed the number of currently known imprinted genes and the incidence of tissue-specific imprinting will be significant.

Downloaded from http://nar.oxfordjournals.org/ at Edinburgh University on August 18, 2014
ACKNOWLEDGEMENTS

The authors thank Dr A. C. Ferguson-Smith for use of data on the UpD12 microarrays, C. V. Beechey for the T65H mouse translocation samples and L. A. Underkoffler and J. N. Collins for experimental assistance with some of the microarrays. The authors thank Dr B. S. Emanuel for human DNA/RNA samples. This work was supported by The Wellcome Trust (R.J.O. and T.R.M.), the BBSRC (R.J.O. and R.S.), EMBO (R.S.), The Guy’s and St Thomas’ Charity (R.J.O. and K.W.), The Generation Trust (A.J.W.) and Public Health Service Grant number GM58759 from the National Institutes of Health (R.J.O.). Funding to pay the Open Access publication charges for this article was provided by The Wellcome Trust.

Conflict of interest statement. None declared.

REFERENCES


