Gene expression

Maximum likelihood inference of imprinting and allele-specific expression from EST data

Cathal Seoighe*, Victoria Nembaware and Konrad Scheffler

Computational Biology Group, Institute of Infectious Disease and Molecular Medicine, University of Cape Town, South Africa

Received on August 18, 2006; revised on September 26, 2006; accepted on October 6, 2006

Associate Editor: John Quackenbush

ABSTRACT

Motivation: In a diploid organism the proportion of transcripts that are produced from the two parental alleles can differ substantially due, for example to epigenetic modification that causes complete or partial silencing of one parental allele or to cis acting polymorphisms that affect transcriptional regulation. Counts of SNP alleles derived from EST sequences have been used to identify both novel candidates for genomic imprinting as well as examples of genes with allelic differences in expression.

Results: We have developed a set of statistical models in a maximum likelihood framework that can make highly efficient use of public transcript data to identify genes with unequal representation of alternative alleles in cDNA libraries. We modelled both imprinting and allele-specific expression and applied the models to a large dataset of SNPs mapped to EST sequences. Using simulations, matched closely to real data, we demonstrate significantly improved performance over existing methods that have been applied to the same data. We further validated the power of this approach to detect imprinting using a set of known imprinted genes and inferred a set of candidate imprinted genes, several of which are in close proximity to known imprinted genes. We report evidence that there are undiscovered imprinted genes in known imprinting regions. Overall, more than half of the genes for which the most data are available show some evidence of allele-specific expression.

Availability: Software is available from the authors on request.

Contact: cathal@science.uct.ac.za

Supplementary information: http://cbio.uct.ac.za/publication_support/ML_EST

1 INTRODUCTION

Different rates of transcription from the two alleles in a diploid organism can result from epigenetic modification that causes the complete or partial suppression of one allele or from cis acting sequence polymorphisms that affect the rate of transcription. Genomic imprinting, the heritable suppression of one parental allele through epigenetic modification, plays a role in several important heritable human diseases and remains a topic of interest and debate among evolutionary biologists (Wilkins, 2005; Reik and Lewis, 2005; Constancia et al., 1998; Morison and Reeve, 1998). More recently, sequence polymorphisms that affect gene expression, both in cis and in trans, have become recognized as a major source of phenotypic variation and disease (Oleksiaik et al., 2002; Morley et al., 2004; Schadt et al., 2003; Yan et al., 2002; Pastinen and Hudson, 2004). Because of their biological importance and implications for human health, several high-throughput methods have been applied to find novel candidate imprinted genes and genes with allelic differences in expression (allele-specific expression), including methods that have made use of the data available in public transcript databases (Ge et al., 2005; Lin et al., 2005; Yang et al., 2003).

Genomic imprinting is thought to affect a small minority of genes. In human and mouse, genomic imprinting has been confirmed for ~41 and 71 transcriptional units (TUs), respectively (Morison et al., 2005), often occurring in genomic clusters. Although the number of human diseases and phenotypes with parent-of-origin effects is greater than can be accounted for by known imprinted genes (Morison and Reeve, 1998; Robertson, 2005), it has been suggested that only a small number of imprinted genes remain to be discovered (Morison et al., 2005). In many cases genomic imprinting affects just a subset of the tissues and developmental stages in which a gene is expressed. As a result it can be difficult to confirm that a specific gene is never imprinted and the imprinting status of several genes is disputed (Morison et al., 2005).

Luodi et al. (2005) inferred imprinted genes in mouse from genomic sequence features and put the number of imprinted mouse genes at 600. Contrary to the view expressed by Morison et al. (2005), this result strongly suggests that the vast majority of imprinted genes remain undiscovered. A wide range of methods has also been applied to detect allele-specific expression by measuring expression level differences between the two alleles of a polymorphic marker in heterozygous individuals (Knight, 2004; Pastinen and Hudson, 2004; Yan et al., 2002; Backlund, 2004). Two groups have recently used public transcript data for this purpose (Ge et al., 2005; Lin et al., 2005).

In the current work we have applied statistical models, in a maximum likelihood framework, to detect genes for which there is evidence of unequal expression of alternative alleles in a subset of the cDNA libraries in dbEST. The models are sensitive to the absence of apparent heterozygous libraries and imbalance between alternative alleles in libraries in which both alleles are observed.

*To whom correspondence should be addressed

© 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.
We have modelled two alternative causes of unequal representation of alleles: imprinting and allele-specific expression and investigated a set of genes for which the model of genomic imprinting is favoured over the null model. Using simulation and a positive control data set we confirm the power of this method to detect imprinted genes and show that it has far greater power than existing methods that have been applied to the same data. To test whether there is evidence of undiscovered imprinted genes in known imprinted regions we also assessed the evidence in favour of the imprinting model for genes that are located close to known imprinted genes.

2 MATERIALS AND METHODS

ESTs and clone library information was downloaded from dbEST (ftp://ftp.ncbi.nih.gov/repository/dbEST/). Human chromosomal sequences, as well as, flat files of pre-computed EST and SNP genomic locations, based on NCBI genome assembly 35 were downloaded from the UCSC genome database [ftp://hgdownload.cse.ucsc.edu/goldenPath/hg17; (Karolchik et al., 2003)]. We discarded ESTs for which <90% of the sequence was mapped to the genome. The EST base at the position corresponding to the SNP was inferred from the EST sequences and genomic coordinates of the SNPs and ESTs. EST sequences derived from cDNA libraries annotated as pooled by tissue donor (Kelso et al., 2003) were removed from the analysis. Known and novel transcripts based on NCBI Genome build 35 (Hubbard et al., 2005) were downloaded from Ensembl (ftp://ftp.ensembl.org/pub/ current_human/data/fasta/). We aligned the transcript sequences to the genome using BLAT (Kent, 2002) with the default parameters and the -fine switch, recommended for mRNA to genome alignment. We removed all transcripts that mapped to more than one location on the genome and discarded any transcripts for which <90% of the total transcript length could be mapped. All Ensembl genes with at least one transcript that mapped multiple times to the genome were removed from our dataset in order to reduce the impact of paralogous sequences on our results. The final dataset consisted only of ESTs and SNPs corresponding to Ensembl transcripts that mapped to unique genomic locations, based on NCBI genome assembly 35 were downloaded from the UCSC genome database [ftp://hgdownload.cse.ucsc.edu/goldenPath/hg17; (Karolchik et al., 2003)]. We discarded ESTs for which <90% of the total transcript length could be mapped.

2.1 Simulations

Simulations were used to compare the performance of alternative methods to recover imprinted genes and genes with allelic-specific expression levels.

We took a sample of 100 SNPs from our dataset satisfying the power criteria described above and designated 50 as imprinted and the remainder as non-imprinted. We constructed 1000 random replicates of this data, matching allele frequency, sequence error rate and EST coverage to the values estimated from the real data. For each simulated cDNA library the genotype for a particular SNP was assigned randomly under the assumption of Hardy–Weinberg equilibrium and estimated population allele frequencies for the SNP. We simulated imprinting with complete and with partial silencing of the imprinted allele. In the case of partial silencing the probability of observing the silenced allele in an imprinted library was sampled from a normal distribution with mean and standard error equal to 0.1 (truncated to the left at zero). For the allele-specific expression simulations, we again simulated replicate datasets consisting of 50 cases and 50 controls. For these cases, the highly-expressed allele was assigned randomly and the probability of observing the low-expression allele was obtained by again sampling from the normal distribution with a mean and standard deviation of 0.1 (truncated at 0). Simulations as well as statistical tests described in Yang et al. (2003) were implemented in Perl.

3 ALGORITHM

For a given SNP, with alleles A and B, that could be mapped to transcript sequences, the data consisted of counts, ai and bi, of the SNP alleles in the cDNA libraries i = 1 . . . N, in which the transcript was found. The likelihood of the observed data, under the null model with no allele-specific expression or imprinting, can be expressed as a function of the frequency of allele A, fa, and the sequencing error rate, ε. We defined ε as the probability that the incorrect allele was read at the SNP position for a given EST and estimated this probability separately for each SNP. The error rate was estimated as half of the proportion of ESTs for which the nucleotide at the polymorphic position corresponded to neither of the SNP alleles, with pseudocounts (one per allele) added to allow for sparse data. In the case of SNPs with more than two alleles this would result in a conservative overestimate of ε. Assuming Hardy–Weinberg equilibrium the likelihood of the data is given by

\[
L(D) = \prod_i \left( f_A^a e^b (1-\epsilon)^b + 2f_A (1-f_A) 0.5^a + b \right.
\]

where the first and last terms correspond to the likelihood for homoyzgous individuals and the middle term corresponds to the likelihood for heterozygous individuals. Note that we are assuming that each library contains data from a single individual: if data from multiple individuals are pooled in a single library, the null model may provide a poor fit because a pool containing homozygous individuals may appear to be heterozygous but with alleles A and B appearing at very different frequencies.

3.1 Model of genomic imprinting

We used a mixture model and two additional parameters to model genomic imprinting, which may be dependent on the context of gene expression and which may involve partial rather than complete silencing of one of the parental alleles. For a given gene the mixture weight, pi, represents the proportion of cDNA libraries in which the gene is imprinted and the parameter q represents the probability of observing the suppressed allele in an imprinted cDNA library. The likelihood of the data under the imprinting model is

\[
L(D) = \prod_i \left( p_i \left( f_A^a e^b (1-\epsilon)^b + f_A (1-f_A) q^a (1-q)^b \right.ight.
\]

\[
+ q^b (1-q)^a + (1-f_A)^a e^b (1-\epsilon)^b + (1-p_i) \left( f_A^e b (1-\epsilon)^a \right) + 0.5^a + b \right)
\]

\[
+ 2f_A (1-f_A) 0.5^a + b \left. (1-f_A)^a e^b (1-\epsilon)^b \right) \right).
\]

3.2 Model of allele-specific expression

We considered a cis regulatory variant which is closely linked to the observed SNP allele and which may affect expression level in a subset of expression contexts. In the absence of recombination between the SNP allele and the regulatory variant we would expect either no effect in a given cDNA library or overrepresentation of one allele, with the overrepresented allele consistent across libraries. Using the mixture weight pq to represent the proportion of cDNA libraries in which there is an allelic difference in gene expression
and a new parameter $\gamma$ to represent the probability of observing allele A in an affected cDNA library. The likelihood of the data is

$$L(D) = \prod p_i \left( \frac{\gamma}{C_0} \theta_C^{1-\gamma} (1-\gamma)^{1-\theta_C} \right)^{p_C} + \frac{2\gamma^2 - \gamma}{C_1} \theta_C^{1-\gamma} (1-\gamma)^{1-\theta_C}$$

$$\left( (1-\gamma)^{1-\theta_C} (1-\gamma)^{1-\theta_C} + \frac{2\gamma^2 - \gamma}{C_1} \theta_C^{1-\gamma} (1-\gamma)^{1-\theta_C} \right).$$

3.3 Model of loss of heterozygosity in cancer

Both the imprinting and the allele-specific expression models can be adapted to model loss of heterozygosity associated with cancer. Loss of heterozygosity involving consistent loss of the same allele, such as would be expected for examples with cancers resulting from heterozygosity for a defective tumour suppressor gene, can be modelled by adapting the allele-specific expression model. We compare the likelihood of a model in which the parameter $p_C$ is estimated separately for libraries annotated as cancer or normal by eVOC (Kelso et al., 2003) to the likelihood of a model in which a single $p_C$ is estimated for all libraries. Similarly the imprinting model can be adapted to model loss of heterozygosity that does not involve consistent loss of the same allele, such as would be expected for example in the case of a haploinsufficient tumour suppressor gene.

3.4 Optimization

Maximum likelihood parameter estimates were obtained by optimizing the likelihood using Powells method (Press et al., 1992), implemented in the Perl programming language. Raw data and software are available from the authors on request.

4 RESULTS AND DISCUSSION

Excluding genes on the X and Y chromosomes, we found 81 SNPs out of a total of 52,027 tested, for which the imprinting model provided a better fit to the data than the null model at the 1% significance level (a complete list is provided as Supplementary Data online). The power to detect an improved fit to the data depends on the number of ESTs that map to the SNP and on the allele frequencies: if there are two few ESTs or if one allele is too rarely observed, the data do not provide a chance of identifying imprinting or allele-specific expression. We therefore restricted the analysis to SNPs that were spanned by at least 50 ESTs in the database and for which the expected number of ESTs derived from the less frequent allele was at least five. This resulted in 76 SNPs for which the fit of the imprinting model was significantly better than the null model from a total of 3969 SNPs tested (corresponding to 1898 different genes). These power criteria succeeded in identifying a much smaller set of SNPs to test while losing very few cases in which the null model could be rejected. The SNPs that favoured the imprinting model were derived from 60 different genes. The genes in which the imprinting model was favoured at the 1% significance level in at least one SNP, ranked in order of the largest difference in log likelihood between the imprinting and null models, are shown in Table 1. We have also provided a more complete version of the data corresponding to these 60 genes as Supplementary Table 2. For each candidate gene this includes a list of libraries in which transcripts of the gene were found as well as eVOC (Kelso et al., 2003) anatomical annotations for each of the cDNA libraries. The table also includes SNP IDs and the number of transcripts of each allele of the SNP observed in each of the cDNA libraries.

Table 1. Genes for which the null model was rejected in favour of the imprinting model at the 1% significance level for at least one SNP

<table>
<thead>
<tr>
<th>Gene</th>
<th>ΔL</th>
<th>p</th>
<th>C Band</th>
<th>Gene</th>
<th>ΔL</th>
<th>p</th>
<th>C Band</th>
<th>Gene</th>
<th>ΔL</th>
<th>p</th>
<th>C Band</th>
</tr>
</thead>
<tbody>
<tr>
<td>TFP12</td>
<td>29.6</td>
<td>1E-13</td>
<td>7</td>
<td>q21.3</td>
<td>LIPA</td>
<td>8.4</td>
<td>2E-04</td>
<td>10 q23.31</td>
<td>PRG2</td>
<td>5.3</td>
<td>5E-03</td>
</tr>
<tr>
<td>GNAS</td>
<td>20.6</td>
<td>1E-09</td>
<td>20</td>
<td>q13.32</td>
<td>PPGB</td>
<td>8.3</td>
<td>2E-04</td>
<td>20 q13.12</td>
<td>C17orf49</td>
<td>5.3</td>
<td>5E-03</td>
</tr>
<tr>
<td>C20orf114</td>
<td>19.2</td>
<td>5E-09</td>
<td>20</td>
<td>q11.21</td>
<td>CGA</td>
<td>8.1</td>
<td>3E-04</td>
<td>6 q15</td>
<td>TMEM66</td>
<td>5.3</td>
<td>5E-03</td>
</tr>
<tr>
<td>MBP</td>
<td>19.1</td>
<td>5E-09</td>
<td>18</td>
<td>q23</td>
<td>COL3A1</td>
<td>7.9</td>
<td>4E-04</td>
<td>2 q32.2</td>
<td>CGI-69</td>
<td>5.3</td>
<td>5E-03</td>
</tr>
<tr>
<td>ALB</td>
<td>16.5</td>
<td>7E-08</td>
<td>4</td>
<td>q13.3</td>
<td>SLCO4A1</td>
<td>7.6</td>
<td>5E-04</td>
<td>2 q32.2</td>
<td>COASY</td>
<td>5.2</td>
<td>6E-03</td>
</tr>
<tr>
<td>IGFBP1</td>
<td>14.8</td>
<td>4E-07</td>
<td>7</td>
<td>p13</td>
<td>CHIT1</td>
<td>7.1</td>
<td>8E-04</td>
<td>1 q32.1</td>
<td>MGCI3170</td>
<td>5.2</td>
<td>6E-03</td>
</tr>
<tr>
<td>H19</td>
<td>14.5</td>
<td>5E-07</td>
<td>11</td>
<td>p15.5</td>
<td>SNURF</td>
<td>7.1</td>
<td>8E-04</td>
<td>15 q11.2</td>
<td>CH3L1</td>
<td>5.1</td>
<td>6E-03</td>
</tr>
<tr>
<td>CRYAA</td>
<td>13.1</td>
<td>2E-06</td>
<td>21</td>
<td>q22.3</td>
<td>CRYBA1</td>
<td>6.9</td>
<td>1E-03</td>
<td>17 q11.2</td>
<td>ADAR</td>
<td>5.0</td>
<td>7E-03</td>
</tr>
<tr>
<td>SPARC1L1</td>
<td>11.5</td>
<td>1E-05</td>
<td>4</td>
<td>q22.1</td>
<td>KISS1</td>
<td>6.5</td>
<td>2E-03</td>
<td>1 q32.1</td>
<td>TF</td>
<td>4.9</td>
<td>7E-03</td>
</tr>
<tr>
<td>STPTC</td>
<td>11.4</td>
<td>1E-05</td>
<td>8</td>
<td>p21.3</td>
<td>PPT1</td>
<td>6.3</td>
<td>2E-03</td>
<td>1 p34.2</td>
<td>ILK</td>
<td>4.9</td>
<td>7E-03</td>
</tr>
<tr>
<td>RHOB</td>
<td>11.4</td>
<td>1E-05</td>
<td>2</td>
<td>p24.1</td>
<td>ANXA1</td>
<td>6.2</td>
<td>2E-03</td>
<td>9 q21.13</td>
<td>ITGB2</td>
<td>4.9</td>
<td>7E-03</td>
</tr>
<tr>
<td>FAM107A</td>
<td>11.3</td>
<td>1E-05</td>
<td>3</td>
<td>p14.2</td>
<td>SLCO2A1</td>
<td>6.0</td>
<td>2E-03</td>
<td>1 p34.2</td>
<td>RAFTLIN</td>
<td>4.9</td>
<td>7E-03</td>
</tr>
<tr>
<td>CRYBA4</td>
<td>10.6</td>
<td>2E-05</td>
<td>22</td>
<td>q12.1</td>
<td>COMT</td>
<td>5.9</td>
<td>3E-03</td>
<td>22 q11.21</td>
<td>OLFM1</td>
<td>4.9</td>
<td>7E-03</td>
</tr>
<tr>
<td>CLU</td>
<td>10.1</td>
<td>4E-05</td>
<td>8</td>
<td>p21.1</td>
<td>SF1</td>
<td>5.9</td>
<td>3E-03</td>
<td>11 q13.1</td>
<td>FMOD</td>
<td>4.8</td>
<td>8E-03</td>
</tr>
<tr>
<td>SILV</td>
<td>10.0</td>
<td>5E-05</td>
<td>12</td>
<td>q13.2</td>
<td>PHF11</td>
<td>5.9</td>
<td>3E-03</td>
<td>13 q14.2</td>
<td>MEGF8</td>
<td>4.8</td>
<td>8E-03</td>
</tr>
<tr>
<td>DLK1</td>
<td>9.6</td>
<td>7E-05</td>
<td>14</td>
<td>q32.2</td>
<td>MRPS34</td>
<td>5.7</td>
<td>3E-03</td>
<td>16 p13.3</td>
<td>TGFBI</td>
<td>4.8</td>
<td>8E-03</td>
</tr>
<tr>
<td>BSG</td>
<td>9.4</td>
<td>8E-05</td>
<td>19</td>
<td>p13.3</td>
<td>C11orf43</td>
<td>5.7</td>
<td>3E-03</td>
<td>11 p15.5</td>
<td>FGB</td>
<td>4.7</td>
<td>9E-03</td>
</tr>
<tr>
<td>SERPIN1A</td>
<td>9.2</td>
<td>1E-04</td>
<td>14</td>
<td>q32.13</td>
<td>SERPIN2</td>
<td>5.4</td>
<td>5E-03</td>
<td>18 q21.3</td>
<td>CTSD</td>
<td>4.7</td>
<td>9E-03</td>
</tr>
<tr>
<td>CTSS</td>
<td>9.2</td>
<td>1E-04</td>
<td>1</td>
<td>q21.2</td>
<td>COL1A2</td>
<td>5.4</td>
<td>5E-03</td>
<td>7 q21.3</td>
<td>ERCC5</td>
<td>4.7</td>
<td>9E-03</td>
</tr>
<tr>
<td>INS</td>
<td>9.1</td>
<td>1E-04</td>
<td>11</td>
<td>p15.5</td>
<td>ACAT2</td>
<td>5.4</td>
<td>5E-03</td>
<td>6 q25.3</td>
<td>QN81843</td>
<td>4.7</td>
<td>9E-03</td>
</tr>
</tbody>
</table>

The chromosome and band to which the genes map are shown. The table also includes SNP IDs and the number of cDNA libraries. The table shows data for the SNP for which the difference in log likelihood between the imprinting and null models (ΔL) is largest, and a P-value obtained by comparing this log likelihood difference to a chi-squared distribution with two degrees of freedom. Known imprinted genes are in bold font.
datasets, each consisting of 50 genes simulated under imprinting (with complete silencing of the imprinted allele) and 50 non-imprinted genes, we recovered on average 67% of the imprinted genes before the first false-positive discovery (Fig. 1). Yang et al. (2003) presented two alternative methods to detect novel candidate imprinted genes from dbEST. Both methods use only ESTs with trace information and high Phred scores rather than modelling sequence error. In the first method, Yang et al. (2003) compared the expected frequency of heterozygous libraries based on inferred frequencies of individual alleles to the expected number of heterozygous libraries, given the EST data mapping to a specific SNP. The difference between these quantities was summarized as a Z-statistic, which was used to calculate a P-value. Few, if any, of the P-values obtained by this approach are statistically significant and instead Yang et al. (2003) rank the logarithm of the P-values and consider the top-ranking examples as candidates for imprinting. The way in which Bayes rule was applied in this method is questionable (the use of flat priors for the alternative genotypes underlying the data in individual cDNA libraries ignores the information about the allele frequencies contained in the collection of cDNA libraries) and it is possible to construct trivial data sets in which the expected value of their Z-statistic is not zero (e.g., a dataset that consists of just one EST in a single cDNA library). However, in simulated data, we found that the method used by Yang et al. (2003) did have some power to detect imprinting (Fig. 1a), although, for any given false positive rate, the power was much lower than the method presented here.

An alternative method contained in the same paper and applied in a subsequent paper (Lin et al., 2005) to the problem of identifying allele-specific expression from EST data relies on observing both alleles in heterozygous individuals. It therefore cannot be used to detect cases of complete imprinting or allele-specific expression (where one allele is never observed). In this method the statistical significance of the difference in allele frequencies within the cDNA library is evaluated using the binomial distribution. We simulated datasets with partial suppression of one allele to compare the performance of this method and the previous method to the likelihood method. For most of the false-positive rate range the likelihood method has far greater power to recover the imprinted genes than either of the previously published methods (Fig. 1b). We also found that the likelihood method had improved power to detect cases of allele-specific expression than the binomial method (Fig. 1c).

To validate the power of the imprinting model to highlight candidate imprinted genes, we obtained a set of 34 known imprinted TUs from the imprinted gene catalogue ([Morison et al., 2001]; http://igc.otago.ac.nz] that could be mapped to Ensembl genes. Of the imprinted genes, six had SNPs that met the power criteria described above and five of the six were among the genes for which the imprinting model was favoured over the null model at the 1% significance level (P-values from the likelihood ratio test with two additional parameters ranged from 0.003 to 8 × 10^{-10}). The remaining known imprinted gene (IGF2) showed a non-significant improvement in likelihood for the imprinting model over the null model (P = 0.1), and a significant improvement in likelihood for the allele-specific expression model over the null model (P = 0.007). The enrichment for imprinted genes among the genes for which the null model was rejected in favour of the imprinting model for at least one SNP was highly statistically significant (P = 1 × 10^{-7} from two-sided Fisher’s exact test).

---

**Fig. 1.** Receiver operator characteristic (ROC) curves, showing the power of the methods that have been used to detect allelic differences in expression from data in dbEST as a function of false positive rate. The likelihood method presented here (triangles) outperforms the Z-test method (squares) and the binomial test method (circles). The curves were constructed from 1000 simulations, each consisting of 50 imprinted genes and 50 non-imprinted genes. Simulations were carried out separately for (a) imprinting with complete suppression of the imprinted allele, (b) imprinting with partial suppression of the imprinted allele and (c) allele-specific gene expression. Note that the binomial test is not applicable to the case of imprinting with complete suppression of the imprinted allele and the Z-test has not been applied to allele-specific expression.

We used simulation to evaluate the accuracy and power of the statistical models we present and to compare to previously published methods of detecting imprinting and allele-specific expression using data in the cDNA libraries of dbEST. In 1000 replicate
Several of the genes in Table 1, not previously identified as imprinted, are promising candidates. TFP12 (tissue factor pathway inhibitor 2), the top candidate from the table \( (P = 1 \times 10^{-13}) \), is located within a region on chromosome 7 that contains several known imprinted genes (Fig. 2a). The TFP12 gene has a 44 bp antisense overlap with an alternative isoform of GNGT1, which shows evidence of maternal expression (Okita et al., 2003). Another of the candidate imprinted genes in Table 1, COL1A2 (collagen, type 1, alpha 2; \( P = 0.005 \)), is also located in this region, 150 kb from a known imprinted gene (SGCE). COL1A2 was considered a candidate for genomic imprinting in mouse on the basis of differential expression between parthenogenote and androgenote mouse embryos but found to be biallelically expressed in normal mouse embryos (Mizuno et al., 2002). Our data suggest that this gene should be reconsidered as a possible candidate for imprinting in human and that a greater proportion of the genes in this region may be imprinted than previously thought.

Five genes from Table 1 map to 11p15, an imprinted region involved in Beckwith Wiedemann syndrome (Morison and Reeve, 1998). Two of these are the well-known imprinted genes INS and H19. One of the candidate genes from Table 1, putative insulin-like growth factor II associated protein, C11orf43, maps to the 162 kb region between INS and H19, just 3 kb from the imprinted insulin-like growth factor 2 (IGF2; Fig. 2b). A second candidate, CTSD, located 230 kb from H19, is among a set of candidate imprinted genes in mouse (Luedi et al., 2005). Although biallelic expression of CTSD has previously been reported (Rachmilewitz et al., 1993), this may be due to limited tissues in which imprinting of this gene has been tested (Luedi et al., 2005). The integrin-linked kinase gene (ILK) from Table 1, also on chromosome 11, is <350 kb from a putatively imprinted zinc finger gene (ZNF215) in 11p15.4. Given the limited proportion of genes for which sufficient data were available to test for imprinting using this approach, the presence of three novel candidate imprinted genes suggests that the number of genes in this region that are imprinted in at least some individuals and under at least some conditions, is underestimated.

The appearance of insulin-like growth factor binding protein 1 (IGFBP1) in Table 1 is also of interest. This protein is involved in regulation of prenatal growth (Jones and Clemmons, 1995) and is encoded within 5 Mb of the imprinted gene GRB10 on chromosome 7, but previous experimental results reported that IGFBP1 is not imprinted (Wakeling et al., 2000). However the experimental test of imprinting of IGFBP1 involved just one fetal tissue source (fetal liver) and the authors point to the possibility that IGFBP1 might be imprinted in other tissues (Wakeling et al., 2000). Closer examination of our data reveals that the improved fit of the imprinting model over the null model for the SNP shown in Table 1 is primarily due to data from a single placental library in which there is a substantial difference in the proportion of two alleles present for one SNP. There is also a library derived from pregnant uterus in which there is a significant difference between the proportions of the
two alleles for a second SNP (rs7454; see Supplementary data). Interestingly, IGFBP3, which is located adjacent to IGFBP1 on human chromosome 7, also shows evidence of imprinting at the 5% significance level ($P = 0.02$).

Since imprinted genes frequently occur in genomic clusters, we plotted the sliding average of the difference in the log likelihoods between the null and imprinting models ($\Delta L$) using windows of 10 consecutive genes satisfying our power criteria. This highlights several, though not all, of the well-known clusters of imprinted regions (e.g. imprinted regions on chromosomes 7 and 11 shown in Fig. 3). To test whether there is evidence of additional imprinted genes in known imprinted regions we compared $\Delta L$ between genes that are located close to known imprinted genes and the remaining genes, considering only genes that satisfied our power criteria. There were 32 genes in our dataset within 1 Mb of a known imprinted gene, which were not themselves known to be imprinted. The mean value of $\Delta L$ for these genes was 2.5, compared to 1.3 for the remaining genes ($P = 0.003$, from 100000 random samples). Genes in close proximity to known imprinted genes are the most likely to have had their imprinting status tested experimentally (Morison et al., 2005). However our results suggest that there are additional imprinted genes in these regions that remain to be discovered. Genes in Table 1, such as COLIA2, TFPI2, CTSD and C11orf43 (Fig. 2), are good candidates for genomic imprinting.

There were 108 genes for which the allele-specific expression model fitted the data better than the null model for at least one SNP at the 1% significance level (Supplementary Table 3). This included 58 of the 60 genes for which the imprinting model fitted the data significantly better than the null model as well as an additional 50 genes for which only the allele-specific model was favoured over the null model. Because the imprinting and allele-specific models are not nested it was not possible to use the likelihood ratio test to distinguish between them. For most of the genes in Table 1 the difference in log likelihood between the two models is small, suggesting that the data available in dbEST cannot distinguish imprinted genes from genes with allelic differences in expression level.

The overlap between the genes that favour the imprinting model over the null model and the genes that favour the allele-specific expression model over the null model is not surprising, because the effects modelled in both cases are similar. The only difference between these two models is whether or not the overrepresented allele is consistent across cDNA libraries: the imprinting model allows different alleles to be overrepresented in different libraries, while the allele-specific expression model favours situations in which the same allele tends to be overrepresented in different libraries. In many cases, the data do not provide enough information to distinguish between these two possibilities. For a given gene, rejection of the null model in favour of either of the alternative models is evidence of unequal expression of the two alleles of the gene in at least a subset of the cDNA libraries in which the gene is found.

Considering the data set as a whole we find evidence that the null model of exactly equal expression of alternative alleles is unlikely to apply for most of the genes analysed. The proportion of a set of independent hypothesis tests for which the null hypothesis is true can be estimated from a vector of $P$-values using software packages such as QVALUE (Storey and Tibshirani, 2003). Application to the vector of $P$-values from the allele-specific expression model reveals that the null hypothesis is likely to be strictly true for only a minority of genes (~20%). This can be visualized by comparing the distribution of $P$-values we get from applying the likelihood ratio test to data generated under the null hypothesis of equal expression of alternative alleles, to the $P$-values we obtain from the real data (Fig. 4). The $P$-value distribution is strongly skewed to the left for the real data and close to the expected uniform distribution for the data simulated under the null hypothesis. This result is not affected by applying a more stringent criterion for determining pooled libraries (for the less stringent method we removed all libraries annotated as pooled by donor in eVOC and for the more stringent one we removed all libraries unless specifically annotated as not pooled by donor). In the more stringent case the proportion of genes conforming to the null hypothesis estimated by QVALUE increases slightly to 24%.

One advantage of our method over some previous methods of inferring imprinting and allele-specific expression from dbEST (Yang et al., 2003; Lin et al., 2005) is that it has the capacity to integrate different signatures of these phenomena and to perform a combined inference based on data in many cDNA libraries. The imprinting model is sensitive to the absence of heterozygous libraries, as well as, imbalances between alleles in heterozygous libraries. The allele-specific expression model combines evidence of imbalance between alleles across many cDNA libraries and this model tends to be favoured over the imprinting model when the overexpressed allele is consistent between libraries. Including a parameter for the relatively high EST sequencing error is another advantage of the statistical models that we have...
applied. This parameter, which is estimated from counts of nucleotides at the polymorphic position that do not correspond to either SNP allele, prevents inference of allelic imbalance for homozygote libraries with some sequence errors. Some previous studies have used the alternative strategy of omitting ESTs with Phred scores below some cut-off (Yang et al., 2003; Lin et al., 2005), however this restricts the data to ESTs with trace information. It also fails to resolve the problem, because base calls with high Phred scores still retain a finite probability of error and even a very small number of errors can give a very high probability of generating false-positive results. For example, a highly expressed gene for which there are many ESTs present in a single homozygous library will be interpreted as providing strong evidence of allele-specific expression if only one or two of the ESTs have a sequence error in favour of the alternative allele at the SNP position.

It is possible to compare the set of SNPs for which the allele-specific expression model is favoured over the null model in this study (Supplementary Table 3) to sets of candidates reported in other studies using alternative methods (Ge et al., 2005; Lo et al., 2003; Lin et al., 2005). However, because in each of the other studies to which we compared our results the number of SNPs reported to be associated with allele-specific expression of the corresponding gene is relatively small, the overlap between our results and previously published sets, as well as the overlaps of the previously published sets with each of the others are all relatively small [e.g. in some of the most extensive previous studies, Ge et al. (2005) report 117 SNPs with allelic differences in gene expression and Lo et al. (2003) report 326, but only three SNPs are shared between these studies]. The most meaningful comparison is between the results of Lin et al. (2005) and the results presented here because both studies make use of essentially the same data. Where Lin et al. (2005) present 35 different genes with evidence of allele-specific expression our candidate list consisted of 108 different genes (at a P-value of 0.01). Six genes are common to both candidate lists. Closer examination of the data for some of the cases where we do not find allele-specific expression reveals that sequence error is an equally plausible explanation of the result.

Allele-specific expression is likely to be much more common than genomic imprinting. Human individuals may be heterozygous for functionally significant polymorphisms in the promoter regions of as many as 40% of genes (Rockman and Wray, 2002). Consistent with this, a very high proportion of human genes show evidence of allele-specific expression in at least some individuals (Lo et al., 2003). We found that genes for which the most data were available tended to favour the allele-specific expression model over the null model. For example, for 108 genes with at least 200 ESTs the null model was rejected in favour of the allele-specific model at the 5% level in 55% of cases, for at least one SNP mapping to the gene. This apparently very high prevalence of allele-specific expression, which is consistent with at least one previous report (Lo et al., 2003), has important implications for our understanding of the biology and molecular evolution of diploid organisms.

Unequal representation of alleles in cDNA libraries could also result from loss of heterozygosity in cancer tissues. This is particularly significant because the majority of cDNA libraries in the public domain are likely to be derived from tumours rather than normal tissues (Baranova et al., 2001). Several of the genes in Table 1 are known oncogenes or tumour suppressors [e.g. KISS1 (Lee et al., 1996), CTSD (Iacobuzio-Donahue et al., 2004), SLC2A1 (Smith, 1999) and RHOB (Huang and Prendergast, 2006)]. For some of these genes genomic deletions or aberrant methylation of one allele in neoplastic tissues could explain rejection of the null model and Table 1 may contain novel examples of cancer associated genes that have not previously been annotated as such. We modified the models for allele-specific expression and imprinting to estimate parameters separately for cancer and normal tissues (Materials and Methods). Genes showing greater evidence of unequal expression of alternative alleles in cDNA libraries from cancer tissues compared to normal tissues are listed in Supplementary Tables 4 and 5. As we would predict, for the majority of cases the parameter representing the probability of unequal expression in a library was higher for cancer than for normal libraries (60 cases from a total of 91; $P = 0.002$). However, binary classification of cDNA libraries as cancer or normal is likely to be too coarse to identify loss of heterozygosity that may be restricted to specific cancer types and further, more detailed analysis is required to determine whether allelic expression effects that differ between cDNA libraries with different annotations can be used to identify novel disease associated genes or mutations.

ACKNOWLEDGEMENTS

This work was funded by the South African National Bioinformatics Network. VN is a recipient of the UCT International Student Scholarship. We are grateful to Chris Gehring and Win Hide for comments on the manuscript. Funding to pay the Open Access publication charges for this article was provided by the South African National Bioinformatics Network.

Conflict of Interest: none declared.

REFERENCES


