Computation method to identify differential allelic gene expression and novel imprinted genes

Howard H. Yang, Ying Hu, Michael Edmonson, Kenneth Buetow and Maxwell P. Lee*

Laboratory of Population Genetics, National Cancer Institute, Gaithersburg, MD 20877, USA

Received on August 23, 2002; revised on November 13, 2002; accepted on November 17, 2002

ABSTRACT

Motivation: Genomic imprinting plays an important role in both normal development and diseases. Abnormal imprinting is strongly associated with several human diseases including cancers. Most of the imprinted genes were discovered in the neighborhood of the known imprinted genes. This approach is difficult to extend to analyze the whole genome. We have decided to take a computational approach to systematically search the whole genome for the presence of mono-allelic expressed genes and imprinted genes in human genome.

Results: A computational method was developed to identify novel imprinted or mono-allelic genes. Individuals represented in human cDNA libraries were genotyped using Bayesian statistics, and differential expression of polymorphic alleles was identified. A significant reduction in the number of libraries that expressed both alleles, measured by $Z$-statistics, is a strong indicator for an imprinted or a mono-allelic gene.

Availability: The data sets are available at http://leelab.nci.nih.gov/leelab/jsp/IGDM/IGDM.html

Contact: leemax@mail.nih.gov.

Supplementary information: http://leelab.nci.nih.gov/leelab/jsp/IGDM/IGDM.html

INTRODUCTION

Genomic imprinting is characterized by parent-of-origin-dependent expression of genes resulting in allele-specific gene expression (Feinberg, 2000; Surani, 2001). Abnormal imprinting is strongly associated with several human diseases including Prader–Willi syndrome (PWS; Nicholls and Knepper, 2001), Beckwith–Wiedemann syndrome (BWS; Mannens et al., 1994) and some types of cancer (Ogawa et al., 1993; Rainier et al., 1993). Most imprinted genes have been identified by genotyping and allele-specific expression analysis targeted to candidate genes. About 50 imprinted genes have been identified in human and mouse genomes. Mono-allelic expression also plays an important role in allelic exclusion, which ensures that mono-specific antibody is produced from B cells, and in X chromosome inactivation, which ensures that gene expression is equal from male and female.

The current release of human Unigene (build #153) contains 3 569 546 EST clones that are in 101 602 clusters. 14 758 of these Unigene clusters contain at least 33 EST clones. Genes with multiple ESTs can be used to deduce information about gene expression and quantitative analysis of gene expression such as the Digital Differential Display has also been achieved by the Cancer Genome Anatomy Project (Strausberg et al., 2000). Computational methods have been used to identify single nucleotide polymorphisms (SNPs) in redundant EST clones (Buetow et al., 1999; Irizarry et al., 2000; Marth et al., 1999).

METHODS

SNP data set

The SNP data set for modeling allele-specific gene expression is extracted from the SNP database described previously in (Buetow et al., 1999). Unigene is known to contain EST clones with variable quality. We have taken several steps to ensure that high quality of EST clones are used in our data set. The EST clones and SNPs must meet the following three criteria to be included in our data set. (1) Phred quality score of an EST clone is equal or greater than 20; (2) SNP score is equal or greater than 0.99 (Buetow et al., 1999); and (3) SNPs are mapped to Locuslink. Any transcript that has not been assembled into Unigene is not considered here. This may include imprinted anti-sense transcripts, which are usually very long and lack introns. This data set consists of 112 812 records for 19 312 unique SNPs. Table 1 is a tiny portion of this data set.

This table consists of three blocks of EST clone records for three SNPs with SNP ids 604639, 605408, and 608434. The SNP 604639 is heterozygous in the two cDNA libraries whereas the genotypes of the SNP 604639 in the
The SNPs 604639 and 608434 were the two examples that were selected by our method. The SNP 608434 is in the known imprinted gene IGF2. The clone count distribution of the SNPs in the normal tissue data set is shown in Figure 1.

**Computation**

This study presents a computational method, which can identify differential allelic expressed genes and candidate imprinted genes from gene expression data in the EST database. Imprinted genes and mono-allelic genes differentially express one allele. To model allele-specific gene expression, it is assumed that each cDNA library represents an individual and all libraries (or the sum of available libraries) constitute a population. If both A and B alleles of an SNP in a gene X are represented in an EST library, the individual is heterozygous at the SNP in the gene X. If only allele A is represented in the EST library (Fig. 2, D1), the genotype for that individual could be either AA or AB. The probability that the individual is AA or AB can be inferred using Bayes’ rule (Weir, 1996). We consider the following three kinds of allele observations D for an SNP from a library:

- **D1** the allele A appeared \( n \) times in the library, or
- **D2** the alleles A and B appeared \( n1 \) and \( n2 \) times respectively in the library, or
- **D3** the allele B appeared \( n \) times in the library.

---

**Table 1.** A tiny portion of the SNP data set to show three examples of SNPs in the data set

<table>
<thead>
<tr>
<th>SNP</th>
<th>Lib</th>
<th>Base</th>
<th>EST Count</th>
<th>Gene</th>
<th>SNP sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>604639</td>
<td>1042</td>
<td>A.C</td>
<td>5</td>
<td>NUDE1</td>
<td>GTTTATCATAMGCWGCGCCT</td>
</tr>
<tr>
<td>604639</td>
<td>1494</td>
<td>A</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>604639</td>
<td>1728</td>
<td>C</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>604639</td>
<td>1730</td>
<td>A</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>604639</td>
<td>424</td>
<td>A</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>604639</td>
<td>589</td>
<td>A.C</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>604639</td>
<td>595</td>
<td>A</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>604639</td>
<td>628</td>
<td>C</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>605408</td>
<td>1730</td>
<td>G</td>
<td>1</td>
<td>KIAA0737</td>
<td>TGACCTCCCGRGGGCTCCAAC</td>
</tr>
<tr>
<td>605408</td>
<td>262</td>
<td>G</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>605408</td>
<td>278</td>
<td>A</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>605408</td>
<td>298</td>
<td>A</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>605408</td>
<td>367</td>
<td>G</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>605408</td>
<td>415</td>
<td>A</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>605408</td>
<td>452</td>
<td>G</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>605408</td>
<td>486</td>
<td>G</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>608434</td>
<td>250</td>
<td>G</td>
<td>3</td>
<td>IGF2</td>
<td>GCAATCGGAAKTGAGCAAAAC</td>
</tr>
<tr>
<td>608434</td>
<td>252</td>
<td>G</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>608434</td>
<td>262</td>
<td>G</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>608434</td>
<td>271</td>
<td>T</td>
<td>2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

605408 and 608434 are the two examples of SNPs selected from the top 1%, whereas 604639 is the example of SNP that was not selected. Lib denotes library id; Base indicates nucleotides of the SNP in the library; EST count refers to the number of EST clones in the library. Gene symbols are provided. Ten bases flanking each site of the SNPs are also provided.

Fig. 1. This figure shows the average clone count per library distribution of the SNPs in the normal tissue data set. The SNPs are sorted by the maximum clone counts per library of those SNPs. Each dot represents an average clone count per library in logarithm scale for one SNP. There are 3651 SNPs with library sizes equal or greater than 10 and 2767 SNPs with clone counts equal or greater than 20.

other are unknown. The genotypes of the SNP 605408 and the SNP 608434 are also unknown. However, their genotypes can be inferred as described in **Computation**.
The expected heterozygote frequency based on the genotype frequencies, $Q_A$ and $Q_B$ represent the population allele frequencies. The allele frequency is calculated as $Q_A = P_{AA} + 0.5P_{AB}$ and $Q_B = 1 - Q_A$. The expected heterozygote frequency based on the Hardy-Weinberg equilibrium distribution is calculated as $Q_{AB} = 2Q_A Q_B$. $P_{AB}$ tends to be lower than $Q_{AB}$ for imprinted genes and genes displaying mono-allelic expression. This behavior can be analyzed using $Z$-statistics described in Figure 2. Bayes’ inference of genotypings procedures. The computational results from Bayes’ inference are used in computing $Z$-statistics.

RESULTS

Normal tissue libraries were mined for SNPs producing a data set of 19 312 SNPs in 112 812 records. The difference between $P_{AB}$ and $Q_{AB}$ was calculated for each SNP using $Z$-statistics. The probability of differential allele-specific expression is indicated by the $p$-value for each SNP. The $p$-values are shown in Figure 3 in negative logarithm scale.

Fifty of 19 312 SNPs in the data set are in known imprinted genes. The validity of the computational method was tested by determining if SNPs in imprinted genes had small $p$-values, i.e. within the top 1% (194 out of 19 312) of SNPs ordered according to increasing $p$-value. Four SNPs in imprinted genes were in the top 1% of the data set: three in IGF2 and one in PEG3. This finding is highly significant ($p$-value = 0.0016 in one-sided Fisher’s exact test). Interestingly, when ESTs in tumor tissue libraries were used to populate the data set, only one of these four SNPs was in the top 1% of differentially expressed genes. This is consistent with the hypothesis that loss of imprinting (LOI) occurs during tumorigenesis. Bayes’ rule
was used to infer the individual genotype frequencies. As a comparison, we consider the following non-Bayesian rule in the inference:

\[
P_{AA|D1} = 1, \quad P_{AB|D1} = P_{BB|D1} = 0, \quad \text{and} \quad P_{AB|D2} = 1, \quad P_{AA|D2} = P_{BB|D2} = 0, \quad \text{and} \quad P_{BB|D3} = 1, \quad P_{AA|D3} = P_{AB|D3} = 0.
\]

When we replaced the Bayes’ rule by the non-Bayesian rule, the SNPs in known imprinted genes had higher \( p \)-values and were not in the top 1% of SNPs in the data set. Thus, the Bayes’ rule performs much better than non-Bayesian rule.

An alternative method for identifying imprinted genes was also developed. In this case, allele-specific gene expression is analyzed in libraries from heterozygotes. This approach identified 165 SNPs with differential allele-specific expression (\( p \)-value < 0.05, Binomial test) and two of them were in known imprinted genes (\( p \)-value = 0.0681 in one-sided Fisher’s exact test). Thus, this alternative method performs less well than the former method, although it may seem more intuitive.

To examine tissue specific imprinting, we divided the data set into 22 tissue specific subsets and searched for SNPs in the imprinted genes. However, we can not reliably compute \( Z \)-statistics, since the SNP counts, the EST counts, and the EST counts per library in these tissue-divided subsets are small. Nevertheless, if we set the thresholds of 7000 for the SNP counts, 11 000 for the EST clone counts, and 1.8 for the EST clone counts per library for the specified SNP, we have three tissues with all three indices greater than the corresponding thresholds. These are kidney, heart, and placenta. From placenta, we found two SNPs corresponding to the two known imprinted genes PLAGL1 and IGF2 in the top 1% of selected SNPs. The one-side Fisher exact test shows that the two SNPs found are significant with \( p \)-value 0.026. One SNP in the known imprinted gene, p57KIP2, is selected from heart, albeit with \( p \)-value 0.148. None was found from kidney. We also examine the allelic expression of genes on the X chromosome from a subset of 16 libraries from males. Four out of 181 SNPs were selected in the top 1% with the \( p \)-value 0.108.

**DISCUSSION**

The computational method described here identifies genes with differential allele-specific patterns of expression based on the EST data. However, there are several possible reasons that the expected and observed heterozygote frequency could differ: (1) the SNP may be in an imprinted gene; (2) the SNP may be in a mono-allelically expressed gene; and (3) the SNP is from a population that is not in Hardy–Weinberg equilibrium. The performance of our computational method for imprinted genes appears pretty good with the \( p \)-value 0.0016 using the pooled libraries. The performance of the computation method for imprinted genes using tissue-specific libraries and for genes on the X chromosome is marginal. This is most likely due to smaller library size and lower EST clone numbers in these tissue-specific data sets. The performance will improve with more EST clones in Unigene. High-throughput assays are currently being developed to systematically validate the 194 SNPs identified as candidate differential allelic expressed genes and imprinted genes. Our initial validation experiment demonstrated that two out of 18 genes selected from the top 1% showed mono-allelic gene expression in fetal kidney and fetal liver using MALDI-TOF. Thus, our computational result and experimental validation demonstrates the potential utility of this computational method in identifying differential allelic gene expression and novel imprinted genes.

**ACKNOWLEDGEMENTS**

We thank anonymous reviewers for their suggestions to improve the paper.

**REFERENCES**


