A survey of genetic and epigenetic variation affecting human gene expression

Tomi Pastinen,1,2 Robert Sladek,1 Scott Gurd,1 Alya’a Sammak,1 Bing Ge,1 Pierre Lepage,1 Karine Lavergne,1 Amelie Villeneuve,1 Tiffany Gaudin,1 Helena Brändström,1 Allon Beck,1 Andrei Verner,1 Jade Kingsley,1 Eef Harmsen,1,2 Damian Labuda,3 Kenneth Morgan,2,4 Marie-Claude Vohl,5 Anna K. Naumova,2,4,6 Daniel Sinnett,3 and Thomas J. Hudson1,2,4

1McGill University and Genome Quebec Innovation Centre H3A IA4; 2Departments of Human Genetics and Medicine, McGill University H3A 1B1; 3Hospital Sainte Justine Research Centre, Department of Pediatrics, University of Montreal H3T 1C5; 4Research Institute of the McGill University Health Centre, Montreal H3G 1A4; 5Lipid Research Centre and Department of Food Science and Nutrition, Laval University, Quebec G1K 7P4; and 6Department of Obstetrics and Gynecology, McGill University Health Centre, Montreal, Quebec H3A 1A1, Canada

Submitted 26 September 2003; accepted in final form 23 October 2003


A COMPENDIUM OF GENETIC VARIANTS affecting gene regulation will be a key resource for the identification of disease-associated polymorphisms and will complement the ever-growing databases of coding polymorphisms (12). Large-scale strategies are needed to build databases of potential regulatory polymorphisms including the identification of single-nucleotide polymorphisms (SNPs) in sequence elements identified by computational-based predictions of promoters and epigenomically conserved sequences as well as large-scale association studies comparing genetic variants with gene expression levels. Most functional studies of these putative regulatory polymorphisms are based on in vitro techniques such as transient transfection assays with allele-specific promoter constructs (29). The results of these systems are often difficult to interpret as the reporter gene constructs are studied outside of their normal chromosomal environment. In contrast, direct assessment of the relative abundance of allelic transcripts allows investigation of allele-specific expression differences in normal chromosomal context as well as elucidation of epigenetic cis-acting mechanisms. In this study, differential expression of allelic transcripts is assessed directly by relative quantitation of intragenic marker alleles in messenger or heteronuclear RNA derived from cells or tissues. This survey used 193 single-nucleotide polymorphisms (SNPs) from 129 genes expressed in lymphoblastoid cell lines to identify 23 genes (18%) with common allele-specific transcripts whose expression deviated from the expected equimolar ratio. A subset of these deviations, or “allelic imbalances,” can be observed in multiple samples derived from reference CEPH (“Centre d’Etude du Polymorphisme Humain”) pedigrees and demonstrate a spectrum of patterns of transmission, including cosegregation of allelic skewing across generations compatible with Mendelian inheritance as well as random monoallelic expression for three genes (IL1A, HTRA2, and FGB). Additional studies for BTN3A2 provide evidence of SNPs and haplotypes in complete linkage disequilibrium with high- and low-expressing transcripts. The pipeline described herein offers tools for efficient identification and characterization of allelic expression allowing identification of regulatory sequence variants as well as epigenetic variation affecting human gene expression.

transcription; polymorphism

METHODS

Samples and RNA preparation. A description of the lymphoblastoid cell lines (LCLs) used in this study is listed in Supplementary Table 2, available at the Physiological Genomics web site.1 The primary set of 63 LCLs, derived from unrelated donors of Caucasian, Asian, Oceanic, and African origin, partially overlaps with a set used to identify SNPs in the coding regions of human genes (4). A second set of LCLs derived from five CEPH pedigrees (1420, 1423, 1424, 1444, and 1416), each composed of four grandparents (only maternal grandparents for family 1416), two parents, and four children, was

1The Supplementary Material for this article (Supplementary Table 1, with detailed data on 129 genes and SNPs assayed; and Supplementary Table 2, listing all COriell cell lines used in this study) is available online at http://physiogenomics.physiology.org/cgi/content/full/00163.2003/DC1.
used in this study. The LCLs were obtained from Coriell (Coriell Institute for Medical Research, Camden, NJ) and grown in RPMI 1640 medium (Invitrogen, Carlsbad, CA) supplemented with penicillin/streptomycin, with 2 mM l-glutamine and 15% heat-inactivated fetal bovine serum (Sigma-Aldrich, St. Louis, MO). The cell growth was monitored using a hemocytometer: when the cultures reached a density of 0.8–1.1 × 10^6 cells/ml, these were pelleted and lysed by resuspension in TRIzol reagent (Invitrogen). Subcutaneous and epicardial adipose tissue samples were obtained from morbidly obese male patients (n = 21) undergoing a biliopancreatic diversion for the treatment of obesity. The tissue samples were homogenized in TRIzol reagent (Invitrogen), and the lysate was preclarified by centrifugation at 4°C prior to RNA isolation. Patients provided written consent for the project. The study was approved by the ethics committees at McGill University Health Centre Research Institute and at Laval University.

cDNA synthesis. RNA was isolated from the lymphoblastoid lines and adipose tissue biopsies using TRIzol reagent according to the manufacturer’s instructions (Invitrogen). In a typical probe reaction, 50-μg aliquots of total RNA were treated with 8 U of DNase I for 40 min at 37°C (Ambion, Austin, TX), extracted with phenol/chloroform (Invitrogen), and precipitated. The resulting RNA was annealed to 1,000 ng random hexamers (Invitrogen), and first-strand cDNA synthesis was performed using SuperScript II reverse transcriptase according to the manufacturer’s instructions (Invitrogen). Adequate removal of the genomic DNA was verified by PCR amplification of the intergenic microsatellite markers D4S2367, D5S816, and D14S597.

Expression profiles. Microarray analysis was performed using 10 μg of total RNA hybridized to Affymetrix HG-U133A GeneChips (Affymetrix, Santa Clara, CA). Detailed protocols for the probe synthesis and hybridization reactions as well as the posthybridization washing and staining have been previously described (21). The hybridized arrays were scanned and raw data extracted using the MicroArray Analysis Suite 5.0 (Affymetrix). All expression profiles were scaled to a mean signal intensity of 1,000 units prior to analysis. To assess clonality of the lymphoblastoid lines, expression levels of IgL were scaled to a mean signal intensity of 1,000 units prior to analysis. expression separately for the two probe pairs, these independent measurements showed high concordance (r = 0.99); thus we averaged the values. Expression profiles obtained from peripheral blood mononuclear cells as well as malignant B-cell lines (OCI-Ly1, OCI-Ly7, OCI-Ly8, and SU-DHL4) were used as reference standards. Allelic imbalance assays and genotyping. Publicly available intragenic SNPs (http://www.ncbi.nlm.nih.gov/SNP/, http://pga.mbh.washington.edu/) were identified and mapped to the genomic sequence (http://genome.ucsc.edu/cgi-bin/hgGateway). PCR primer design was carried using Primer 3.0 (http://www.genome.wi.mit.edu/cgi-bin/primer/primer3.cgi). PCR testing for genomic DNA and cDNA samples (including a RT− control) was carried out using the same conditions for all fragments, and the amplicon size was verified by agarose gel electrophoresis. A subset of genes were studied using primers for both mRNA and heteronuclear or unprocessed RNA (hnRNA) (n = 6), i.e., using RT-PCR designs with one or both primers located in introns of target genes: the use of intronic SNPs showed consistency of allelic imbalance calls between intronic SNPs and coding SNPs (eSNPs); subsequently intronic assays were designed for six genes which had no common eSNPs. Optimization of the PCR performance was carried out for failed PCR fragments. Independent replicate PCR reactions were carried out for all cDNA samples in the screening process. All primer sequences and PCR conditions can be accessed at http://www.genome.mcgill.ca/Regulotory/PCR. Single base extension (SBE) detection primers of minimum of 19 bp length and melting temperature above 55°C were designed in both orientations for each SNP using PrimerPremier 5.0 (Premier Biosoft International, Palo Alto, CA). The PCR reactions were treated with exonuclease I and shrimp alkaline phosphatase, applying protocols suggested by the manufacturer rescaled for reduced reaction volumes (Acyctylomycin Kits; PerkinElmer, Wellesley, MA). Fluorescent polarization-single base extension (FP-SBE) reactions were performed in both orientations for each duplicate PCR reaction; a total of eight replicate assays were studied for each cDNA sample. Following the FP reaction, the extension products were incubated with single-stranded binding protein (USB, Cleveland, OH); reading buffer was added and the plates were read using an Analyst HT reader (Molecular Devices, Sunnyvale, CA) as described previously (11). The FP reaction conditions as well as the SBE primer sequences can be viewed at http://www.genome.mcgill.ca/Regulotory/FP. Genotype calling in genomic DNA samples was carried out using a clustering algorithm (kindly provided by Dr. A. D. Long, Univ. of California at Irvine), which groups the data into four categories (homozygous AA, heterozygous AB, homozygous BB, and failed). To eliminate failed RT-PCRs, we applied the FP values of the successful genomic DNA data points (i.e., belonging to clusters AA, AB, or BB) to set the signal intensity threshold. In the samples above threshold, a ratio of rhodamine-110 (R110)-labeled allele (A) vs. the total signal [carboxytetramethylrhodamine (Tamra) allele (B) + of R110-labeled allele (A)] was calculated, and these A/(A + B) ratios were used in all subsequent data analysis. The averages of the A/(A + B) ratios over the three genotype clusters were the basis for heterozygote allele ratio estimation: if the heterozygous data point (AB) had A/(A + B) ratio below the average AB cluster ratio, then the relative distance of this data point between the average ratios for AB and BB clusters provided the estimate of the allelic deviation; conversely, if the AB ratio for the data point was above the average AB cluster ratio, then the relative distance of this data point between the average ratios for AB and AA was used for allele ratio estimation. The correspondence of the AB cluster averages in RNA (RT-PCR) and DNA samples was high (Pearson correlation coefficient r = 0.97). Thus in cases where common allelic imbalances affected the average AB cluster ratio in RNA (evident by discrepancy between average AB ratios in the genomic DNA and RNA samples) we applied the average AB ratio from genomic DNA for allele ratio estimation (applied to BTNSA2, IGFI, NFATC4, KLK1, HTRA2, FEN1, CX3CR1, IL1A, PDCD1, ERCCCI, FCER1G, FGB, and ADRB2). The use of the relative distance of A/(A + B) ratios between AB and BB or AB and AA for allele ratio estimation assumes that the allele ratios between the heterozygote and either homozygote clusters show linear correlation; this slightly underestimates the true deviations, as demonstrated by an analysis of standard curves generated for 30 genes (Fig. 2B). Detailed allele ratio data for samples exceeding our threshold (>60:40 or <40:60 allele ratio) are given in Supplementary Table 1.

Allelic imbalance pipeline. Genes were selected based on their potential role in immunity and inflammation, metabolic disorders, and cancer. Assays were designed for 239 genes: an average of 2 SNPs were included for each locus (495 SNPs in total). The intragenic SNPs were obtained from public databases. The assays were first tested for positive RT-PCR signals with a appropriately sized amplified fragment using RNA derived from LCLs. Genomic DNA samples were genotyped to validate the SNPs as well as to confirm good technical performance of the FP-SBE assay. Poor or absent RT-PCR amplification was seen for 22% of SNPs (n = 107), whereas the validation of SNPs failed in 37% of cases (n = 185). Of these, 80% were rejected due to their rarity or absence (monomorphic in the tested cell lines) and 20% due to poor technical performance of the FP assay. The remainder of the failed SNPs (n = 7) did not amplify from genomic DNA. As pilot studies demonstrated that genes with low expression levels were associated with high failure rates in RT-PCR, we modified...
the pipeline to exclude genes whose expression was not reliably detected using expression microarray studies of the LCLs.

Validation studies in adipose tissues and CEPH pedigrees. BTN3A2 showing evidence of differential allelic expression in LCLs was assayed in paired genomic DNA and adipose tissue-derivied RNA samples (n = 21); the assays were carried both in mRNA as well as in hnRNA in epiploic and subcutaneous adipose tissue yielding same results. In addition, ~50% of all SNPs were assayed in LCLs (n = 48) from five 3-generation CEPH-families for identification of potential transmission of allelic imbalances. Mendelian transmission of allelic imbalance in this study is defined as full cosegregation of alleles and/or haplotypes with the allelic imbalance phenotype (including the direction of allelic deviation). Seventeen of the 23 genes with common allelic imbalances in the primary screening had partial or complete family data available. In addition to the nine genes reported in the RESULTS, five genes had either no allelic imbalance (ADPRTL and MICA) in the five pedigrees, or one instance that was only present in a grandparent (KLK1, FCER1G, and IL19). Three genes (CD44, CAT, and NFATC4) have insufficient informative occurrences and allelic imbalances to allow meaningful interpretations. The unrelated screening panel included the nine grandmothers from the CEPH families, which were cultured independently providing a general control for consistency of the allele ratio measurements in independent samples. Additional experiments for FGB and HTR2A using independently cultured cells from CEPH families 1420 and 1444 were carried out to further confirm that the observed effect was not caused by sample contamination, as these genes are expressed at very low levels in LCLs and only a small subset of heterozygotes can be amplified by RT-PCR.

Sequencing. Ten genes with common allelic imbalances were validated by sequencing in independent RNA preparations. Sequencing was done in parallel for amplified genomic DNA and cDNA samples using the BigDye Terminator Cycle Sequencing v3.0 kit (Applied Biosystems, Foster City, CA), applying conditions sug-

Fig. 1. Pipeline for screening allele-specific expression. A: overview of allelic imbalance pipeline. Data and examples for the screening phase are shown in B and in Figs. 2 and 3A and in Table 1. Validation and characterization data is presented in Figs. 3B through 6 and in Table 1. B: parallel quantitative genotyping in paired genomic DNA and RNA (cDNA) samples. Raw data for two genes, PLCG2 (rs1143688) and KLK1 (rs1054713), are shown in the top left and top right of B, respectively (negative controls not shown). The x-axis and y-axis of the scatter plots show fluorescence polarization counts for R110 (A allele) and Tamra (B allele), respectively. Each assay is replicated with four independent data points for each genomic DNA sample (black squares) and eight independent data points for each RNA sample (red circles). The genomic DNA data points form three genotype clusters (AA, AB, and BB), and the data of the heterozygous (AB) samples is further examined in corresponding RNA samples. There is minimal dispersion of the heterozygous RNA data points for PLCG2 gene (top left), whereas the heterozygote cluster in KLK1 shows deviation toward the A as well as the B allele (top right). The magnitude of the deviation can be estimated based on the average A/(A+B) ratios in the three clusters, and consistency of estimated deviation can be investigated by observing the independent replicate data points for each sample. Bottom of B illustrates the estimated A:B allele ratios (y-axis) of independent replicates for KLK1 gene in eight samples (x-axis) extracted from the data presented in the scatter plot in top right of B. The largest possible deviation is ±50%, which would indicate monoallelic expression of either allele (see also Fig. 3A). The genomic DNA and RNA data points for each sample are represented by gray and red bars, respectively. The consistency of the independent replicates indicates that the dispersion of the RNA heterozygote cluster seen in the scatter plot for KLK1 gene (top right) is not random; rather RNA samples consistently show one or the other allele overexpressed. Three of the RNA samples (LY49, LY23, and LY34) shown at the bottom have allelic imbalance according to our criteria: averaged estimated deviation exceeding 40:60 or 60:40 ratio, which are consistent over three or more independent replicates (see Fig. 2 for additional details).
gested by the manufacturer. After purification by ethanol precipitation, the samples were resuspended in Hi-Di formamide and run on the 3700 DNA Analyzer (Applied Biosystems). Sequencing for BTN3A2 SNP discovery followed the same protocols and was performed using primers without sequence homology to other BTN3 genes. BTN3A2 SNP flanking and primer sequences are available at http://www.genome.mcgill.ca/regulatory/sequencing.

X-inactivation assays. Clonality of the LCLs was assessed using a PCR-based androgen receptor methylation assay (1) with modifications (20). LCL DNA from seven female donors from CEPH families 1423 and 1444 was studied. For these same donors X-inactivation phenotypes from peripheral blood lymphocytes were established earlier (19). Six of the LCL samples showed nonrandom X-inactivation ratios (from 0.75 to 0.99) that differed considerably from the X-inactivation ratios in noncultured lymphocytes of the same females.

RESULTS

To investigate allele-specific cis-acting regulatory mechanisms, we devised a pipeline to rapidly screen genes for allelic imbalance (Fig. 1A). This strategy was applied to a primary screening cohort composed of LCLs from 63 unrelated individuals to study genes with potential roles in human complex disease phenotypes. None of the target genes had prior in vivo evidence of cis-acting polymorphisms in LCLs, other than three imprinted genes (MEST, PEG-10, and ATP10C) included as positive controls (17, 18, 25). We optimized the single base extension with fluorescence polarization detection assay (FP-SBE) (11) to provide quantitative detection of allele-specific transcripts (Fig. 1B). The allelic imbalance screening approach relies on quantitative genotyping of heterozygous individuals for intragenic SNPs in RNA transcripts and comparing the observed allele ratios in corresponding genomic DNA samples, which are assumed to represent 50:50 allele ratios. Based on allelic transcript measurements performed using independent RNA samples, we selected a calculated allele ratio of 40:60 as the cutoff for determining the presence of a significant allelic imbalance (Fig. 2A). We observed a minor amount of nonlinearity of standardized allele ratios (determined by DNA dilution studies) when compared with our estimates of allele ratios obtained using RT-PCR templates (see METHODS for details), such that the allele ratio estimates correspond on average to slightly larger true deviations. Thus the calculated 40:60 cutoff corresponds to an average 37:63 allelic deviation representing a 1.7-fold difference between the relative transcript levels (Fig. 2B). Examples of the screening data for 13 heterozygote samples for three genes are shown in Fig. 3A; no allelic imbalance was observed for PLCG2, whereas five and six samples show allelic imbalance for IGF1 and BTN3A2, respectively.

Allelic transcript ratios were successfully measured in 129 genes (including the 3 control genes) using 193 SNPs. The data set generated using the primary screening panel has 3,192 informative heterozygous occurrences and 8,346 uninformative homozygous genotypes. Each of the imprinted genes included as controls demonstrated strong to complete allelic imbalance in all heterozygous samples. Evidence of replicable deviations exceeding the 40:60 threshold in two or more LCL samples was observed for 23 of the 126 genes tested (Table 1). Five of these genes are informative for two or more cSNPs and show concordant allelic imbalances in assays performed using different amplicons. Furthermore, we carried out parallel allelic imbalance studies using primers to assay unspliced transcripts (hnRNA) for BTN3A2 (Fig. 3B, left), KL (Fig. 3B, right), and PDCD1 (data not shown) revealing concordant allelic imbalances between mRNA- and hnRNA-specific assays. Sequence validation performed in independent RNA
Fig. 3. Screening and validation of relative expression of allelic transcripts. A: three genes screened by the FP-SBE-based system for relative allelic transcript level quantitation are illustrated with 13 representative heterozygotes (x-axis) shown for each gene (black bars). The A:B allele ratios (y-axis) are averaged over replicate measurements (see Fig. 1B). The correspondence of the A:B ratios to actual alleles at each single-nucleotide polymorphism (SNP) can be read on the right. The threshold for calling significant allelic deviation is shown by a dashed red line. Left: the relative expression of alleles in \textit{PLCG2} gene (rs1143688) are equal for all informative samples tested. Middle: \textit{IGF1} (rs6220) shows allelic deviations exceeding the threshold in multiple samples; the numbers to the right of the bars correspond to the samples shown in B. Right: strong overexpression of A allele for 6 of the 13 heterozygotes shown is evident for \textit{BTN3A2} gene (rs1985732). Among all the informative heterozygous lymphoblastoid cell line (LCL) samples, 60% showed no allelic imbalance and 40% showed unidirectional allelic imbalance of similar magnitude. B: heteronuclear or unprocessed RNA (hnRNA) assays were validated both using intronic RT-PCR primers in amplification of a coding SNP as shown on the left for \textit{BTN3A2} and using intronic SNPs as shown on the right for \textit{KL}. In both graphs, each pair of white and black bars represents hnRNA-specific and mRNA-specific allele ratios, respectively. The y-axis represents the estimated deviation from the 50:50 allele ratio, and hnRNA-specific allele ratios (white bars) are arbitrarily defined as positive. The dotted red line corresponds to our threshold for significant allelic deviation (i.e., >60:40 or <40:60 ratio). Both hnRNA and mRNA results were results were available for 15 \textit{BTN3A2} heterozygotes showing a high correlation (r = 0.98) for allele ratios, and all allelic imbalance calls are concordant (left). For \textit{KL} (right), results from 17 heterozygotes informative for three pairs of independent intronic (white bars) and exonic (black bars) SNPs (the three SNP pairs are indicated above the three groups of bars) are shown. The samples are derived from our family panel; thus the phase of the independent SNPs for each of the 17 pairs shown is unequivocal. Good correlation of allele ratio estimates in the independent intronic-exonic SNP pairs (r = 0.92) was observed: in 16/17 (94%) heterozygotes the allelic imbalance call is concordant between the SNPs (indicated by a red arrow). C: validation by sequencing in independent RNA preparations was carried out for a subset of genes showing evidence of allelic expression variation in the screening. For each sample, the genomic DNA sequencing (top) was carried out in parallel with the cDNA sequencing (bottom). Results of validation for three samples in the \textit{IGF1} gene (rs6220) are shown, confirming the allelic imbalance screening results shown in the middle of A.
either allele was seen in multiple informative heterozygotes for regulatory polymorphism. In contrast, monoallelic expression of allelic imbalance with at least fourfold overexpression of one transcriptional allelic imbalance.

of a high LD block (8) in world populations of origins similar included in Table 1), these distances exceed the average length downstream of the transcription start site (22 kb for genes geneity of one or more.

Bidirectional allelic imbalance could result from allelic heterozygosity of one or more cis-acting regulatory polymorphisms present in coding, intronic and/or other noncoding regulatory sequences (9), or from DNA methylation, histone acetylation, and other epigenetic factors (13, 16). Alternatively, the SNPs used in the allelic imbalance screen could be detecting a single polymorphism that is responsible for the differential expression of allelic transcripts, but that does not show unidirectional allelic imbalance when it is not in complete linkage disequilibrium (LD). The SNPs used in this study are on average 34 kb downstream of the transcription start site (22 kb for genes included in Table 1), these distances exceed the average length of a high LD block (8) in world populations of origins similar to those studied in this survey. Allele-specific splice variants or unequal allele-specific mRNA decay rates could also mimic transcriptional allelic imbalance.

BTN3A2 is an example of a gene displaying unidirectional allelic imbalance with at least fourfold overexpression of one allele, suggesting that the marker SNP is in LD with a regulatory polymorphism. In contrast, monoallelic expression of either allele was seen in multiple informative heterozygotes for ILIA and HTR2A and in all informative samples of FGB.

In an attempt to classify the allelic imbalance based on transmission patterns, allelic imbalances were studied for nine genes in three-generation CEPH pedigrees. Of these, BTN3A2 showed strong unidirectional allelic imbalance segregating in a Mendelian fashion in two families (Fig. 4A). One additional family (1420) had a single grandparent carrying allelic imbalance, which was not transmitted. Four additional genes (IGF1, CX3CR1, PDCD1, and PLAUR) also demonstrated transmission compatible with Mendelian segregation; however, inconclusive phase assignment precluded unambiguous determination of the mode of inheritance. Notably, three genes with predominantly monoallelic expression patterns (ILIA, HTR2A, and FGB) showed inheritance patterns incompatible with Mendelian inheritance and classic parent-of-origin effects (as shown for PEG-10, Fig. 4B). Two sample pedigrees show transmission patterns of monoallelic expression for ILIA that is discordant for the expressed allele in two pairs of siblings that inherited haplotypes identical-by-descent (Fig. 4C). Finally, KL demonstrated moderate allelic deviations for multiple independent SNPs in all five pedigrees; however, we observed two pairs of siblings with complete sharing of alleles identical-by-descent in which one of the siblings shows the allelic imbalance observed in his parent while the other sibling shows no skewing in allelic ratios. The transmission pattern of KL is compatible with incomplete penetrance of allelic imbalance.

The demonstration of heritable allele-specific expression for BTN3A2 gene allowed us to correlate genetic variants at the BTN3A2 locus with gene expression levels. The initial exonic

### Table 1. Genes with common allelic imbalances

<table>
<thead>
<tr>
<th>Accession No.</th>
<th>Gene</th>
<th>Chr</th>
<th>No. of SNPs</th>
<th>No. of Informative Unrelated Hets</th>
<th>No. of Samples with AI</th>
<th>AI Allele Frequency in Unrelateds</th>
<th>Average AI Magnitude</th>
<th>Direction of Deviation</th>
<th>Independent SNPs</th>
<th>Independent Sequence Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>NM_006437</td>
<td>ADPRT1</td>
<td>13q11</td>
<td>2</td>
<td>32</td>
<td>4</td>
<td>6%</td>
<td>65.35</td>
<td>+/−</td>
<td>NA</td>
<td>ND</td>
</tr>
<tr>
<td>NM_000024</td>
<td>ADRB2</td>
<td>5q31-q32</td>
<td>1</td>
<td>10</td>
<td>6</td>
<td>30%</td>
<td>30.70</td>
<td>+/−</td>
<td>NA</td>
<td>+</td>
</tr>
<tr>
<td>NM_001641</td>
<td>APEX</td>
<td>14q11.2-q12</td>
<td>1</td>
<td>28</td>
<td>4</td>
<td>7%</td>
<td>30.70</td>
<td>+/−</td>
<td>NA</td>
<td>ND</td>
</tr>
<tr>
<td>NM_007047</td>
<td>BTN3A2</td>
<td>6p22.2</td>
<td>2</td>
<td>26</td>
<td>13</td>
<td>25%</td>
<td>90.10</td>
<td>major</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>NM_001752</td>
<td>CAT</td>
<td>11p13</td>
<td>2</td>
<td>27</td>
<td>3</td>
<td>6%</td>
<td>70.30</td>
<td>+/−</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>NM_000648</td>
<td>CCR2</td>
<td>3p21</td>
<td>1</td>
<td>9</td>
<td>3</td>
<td>18%</td>
<td>75.25</td>
<td>+/−</td>
<td>NA</td>
<td>+</td>
</tr>
<tr>
<td>NM_001337</td>
<td>C3XCR1</td>
<td>3p21.3</td>
<td>2</td>
<td>14</td>
<td>7</td>
<td>25%</td>
<td>70.30</td>
<td>+/−</td>
<td>+</td>
<td>ND</td>
</tr>
<tr>
<td>AL532642</td>
<td>CD44</td>
<td>11p13</td>
<td>2</td>
<td>44</td>
<td>6</td>
<td>7%</td>
<td>65.35</td>
<td>+/−</td>
<td>+</td>
<td>ND</td>
</tr>
<tr>
<td>NM_001983</td>
<td>ERCC1</td>
<td>10q13.2-q13.3</td>
<td>1</td>
<td>18</td>
<td>7</td>
<td>19%</td>
<td>65.35</td>
<td>+/−</td>
<td>NA</td>
<td>ND</td>
</tr>
<tr>
<td>NM_004106</td>
<td>FER1G</td>
<td>1q23</td>
<td>2</td>
<td>26</td>
<td>9</td>
<td>17%</td>
<td>70.30</td>
<td>+/−</td>
<td>NA</td>
<td>+</td>
</tr>
<tr>
<td>NM_004111</td>
<td>FEN1</td>
<td>11q12</td>
<td>1</td>
<td>11</td>
<td>9</td>
<td>41%</td>
<td>70.30</td>
<td>+/−</td>
<td>NA</td>
<td>ND</td>
</tr>
<tr>
<td>NM_005141</td>
<td>FGB</td>
<td>4q28</td>
<td>2</td>
<td>6</td>
<td>6</td>
<td>50%</td>
<td>100.0</td>
<td>+/−</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>NM_000621</td>
<td>HTR2A</td>
<td>13q14-q21</td>
<td>1</td>
<td>5</td>
<td>5</td>
<td>50%</td>
<td>100.0</td>
<td>+/−</td>
<td>NA</td>
<td>+</td>
</tr>
<tr>
<td>NM_000618</td>
<td>IGF1</td>
<td>12q22-q23</td>
<td>1</td>
<td>29</td>
<td>13</td>
<td>22%</td>
<td>70.30</td>
<td>+/−</td>
<td>NA</td>
<td>+</td>
</tr>
<tr>
<td>NM_013371</td>
<td>IL19</td>
<td>3q22.2</td>
<td>2</td>
<td>31</td>
<td>5</td>
<td>8%</td>
<td>65.35</td>
<td>+/−</td>
<td>+</td>
<td>ND</td>
</tr>
<tr>
<td>NM_005755</td>
<td>ILIA</td>
<td>2q14</td>
<td>1</td>
<td>22</td>
<td>12</td>
<td>27%</td>
<td>85.15</td>
<td>+/−</td>
<td>NA</td>
<td>+</td>
</tr>
<tr>
<td>NM_004795</td>
<td>KL</td>
<td>13q12</td>
<td>7</td>
<td>53</td>
<td>8</td>
<td>8%</td>
<td>70.30</td>
<td>+/−</td>
<td>+</td>
<td>ND</td>
</tr>
<tr>
<td>NM_002257</td>
<td>KLK1</td>
<td>1q13.3</td>
<td>1</td>
<td>31</td>
<td>8</td>
<td>13%</td>
<td>65.35</td>
<td>+/−</td>
<td>NA</td>
<td>ND</td>
</tr>
<tr>
<td>NM_000247</td>
<td>MICA</td>
<td>6p21.3</td>
<td>1</td>
<td>31</td>
<td>4</td>
<td>6%</td>
<td>65.35</td>
<td>minor</td>
<td>NA</td>
<td>+</td>
</tr>
<tr>
<td>NM_004554</td>
<td>NFATC4</td>
<td>14q11.2</td>
<td>1</td>
<td>10</td>
<td>6</td>
<td>30%</td>
<td>70.30</td>
<td>major</td>
<td>NA</td>
<td>ND</td>
</tr>
<tr>
<td>AF363458</td>
<td>PDCD1</td>
<td>2q37.3</td>
<td>2</td>
<td>37</td>
<td>12</td>
<td>16%</td>
<td>65.35</td>
<td>+/−</td>
<td>+</td>
<td>ND</td>
</tr>
<tr>
<td>NM_002659</td>
<td>PLAUR</td>
<td>19q13</td>
<td>2</td>
<td>22</td>
<td>5</td>
<td>11%</td>
<td>70.30</td>
<td>+/−</td>
<td>NA</td>
<td>+</td>
</tr>
<tr>
<td>NM_003006</td>
<td>SELPLG</td>
<td>12q24</td>
<td>1</td>
<td>18</td>
<td>4</td>
<td>11%</td>
<td>60.40</td>
<td>minor</td>
<td>NA</td>
<td>ND</td>
</tr>
</tbody>
</table>

Genes with at least two samples with averaged allele ratios below 40:60 or above 60:40 ratio, which is consistent over 3 or more independent replicates. AI, allelic imbalance. The “Magnitude” is an average over all heterozygotes exceeding the threshold ratio. For Direction of Deviation, “+/−” indicates bidirectional allelic imbalances where both alleles of the marker SNP(s) are seen overexpressed in different informative samples, in case of unidirectional allelic imbalances; “minor” and “major” correspond to relative overexpression of the minor or major allele (Supplementary Table 1), respectively. For Independent SNPs “+” indicates that independent SNPs provided corroborating evidence (phase known, deviation to the same direction in all SNPs), and “NA” indicates that no independent SNP data available for those samples demonstrating significant deviation. For Independent Sequence Validation, “+” indicates that sequencing in independent RNA preparations confirmed allelic imbalance, and “ND” indicates not done.

preparations for 10 genes was entirely consistent with the allelic imbalance data obtained using the FP-SBE assay (Fig. 3C). In addition, 10 genes had replicable deviations exceeding the threshold in a single LCL sample (Supplementary Table 1).

Most of the observed allelic imbalances are bidirectional, meaning that either allele is overexpressed in different LCLs. Bidirectional allelic imbalance could result from allelic heterogeneity of one or more cis-acting regulatory polymorphisms present in coding, intronic and/or other noncoding regulatory sequences (9), or from DNA methylation, histone acetylation, and other epigenetic factors (13, 16). Alternatively, the SNPs used in the allelic imbalance screen could be detecting a single polymorphism that is responsible for the differential expression of allelic transcripts, but that does not show unidirectional allelic imbalance when it is not in complete linkage disequilibrium (LD). The SNPs used in this study are on average 34 kb downstream of the transcription start site (22 kb for genes included in Table 1), these distances exceed the average length of a high LD block (8) in world populations of origins similar to those studied in this survey. Allele-specific splice variants or unequal allele-specific mRNA decay rates could also mimic transcriptional allelic imbalance.
SNP used to detect BTN3A2 allelic imbalance (marked +10491 in Fig. 5A) was located in the 3′-untranslated region (3′-UTR) region of the gene; only 40% of heterozygous samples for marker +10491 showed allelic imbalance. We subsequently discovered nine SNPs that span 15 kb encompassing the proximal promoter as well as the first exon and intron (Fig. 5A) and occur on a single haplotype, as determined by unambiguous phase information provided by the CEPH pedigree analysis. The putative “regulatory” haplotype demonstrates 100% correlation with the presence of a low-expressing BTN3A2 transcript in the three CEPH pedigrees carrying the haplotype as well as in 13 unrelated samples carrying haplotype-specific SNPs. RT-PCR with an intronic primer was also used to evaluate allelic imbalance in nonspliced hnRNA. The allele ratios corresponding to each haplotype are shown below the haplotype label; if the estimated allele ratios deviated less than our threshold value of 40:60 or 60:40, then a 50:50 ratio is assigned to the sample. Black squares (male) and circles (female) correspond to allele ratios exceeding our threshold, white symbols signify informative samples with equal transcript ratios, and gray symbols correspond to uninformative samples (homozygous for the marker SNP) with transcript ratios denoted “NA.” Unknown phases are labeled by question marks. A: Mendelian transmission of BTN3A2 allelic imbalance measured using SNPs rs1985732 and +10433 (see Fig. 5A) in CEPH families 1424 and 1444, which demonstrated informative transmissions of the allelic expression phenotype. Five additional SNPs were genotyped to unambiguously assign phase (data not shown). The haplotype showing relative underexpression in informative heterozygotes was identical in the two unrelated families as well as in all unrelated individuals carrying allelic imbalance (Fig. 5). B: transmission of PEG-10 monoallelic expression patterns in CEPH families 1416 and 1424 for SNP rs13073, consistent with paternal imprinting. C: discordant monoallelic expression in IL1A alleles in siblings from CEPH families 1444 and 1423 that inherited both haplotypes identical-by-descent. Six additional SNPs were genotyped at IL1A locus to assign phase (data not shown).
underlying *cis*-acting mechanism for *IL1A* and possibly for *FGB* and *HTR2A* allelic imbalances is random monoallelic expression detected due to clonality of the LCLs.

**DISCUSSION**

This survey demonstrates that allelic imbalance assays provide an efficient means to screen for the presence of genetic and epigenetic factors that alter gene expression. The technique can be applied to most genes using either exonic SNPs or intronic SNPs present in hnRNA, and the detected allelic imbalances offer a resource for subsequent characterization of the *cis*-acting regulatory mechanisms affecting these genes. Our study detected common allelic imbalances in 18% (23/126) of genes tested. We note that only 6% (4/69) of mouse genes have been reported to show allelic imbalances (5), although this may be underestimated due to the limited number of mouse strains analyzed. Allelic imbalances in a survey of 13 human genes (33) studied in LCLs showed results similar to this report, if we count the 3/13 (23%) genes where allelic imbalances were seen in more than one unrelated sample. Recently, allelic expression studies have been extended to human cadaver brain tissue as well as fetal liver and kidney samples (3, 16), with both studies reporting 25–50% of genes demonstrating allelic expression differences. Interestingly, a GeneChip-based (HuSNP, Affymetrix) survey of fetal liver and kidney samples from seven unrelated fetuses demonstrated very frequent and strong allelic expression differences; over 20% of all informative heterozygotes showed greater than twofold differences in allelic expression, and >25% of the genes studied demonstrated greater than fourfold differences in allelic ratio.
allelic ratios. The latter study thus reports larger allelic imbalances than our study and to the three other studies carried out by primer extension methods in adult tissues or cell lines (3, 5, 33), all of which report <10% prevalence of allelic expression in informative heterozygotes and only rare occurrences of >4-fold expression differences. It is unclear whether the observed qualitative and quantitative differences of allelic expressions are due to the methodology used or tissues studied. Extension of our allelic imbalance survey to other tissues as demonstrated for BT3N2A2 in adipose tissues is warranted to further explore the questions of tissue specificity and varying prevalence of allelic expression suggested by earlier studies (5, 16).

Given the high level of interest for genes such as KL in aging and osteoporosis (2, 22), PDCD1 in systemic lupus erythematosus (26), KLK1 in end-stage-renal disease (34) and IGF1 in glucose tolerance (31), it is important to develop approaches to characterize allelic imbalances and identify polymorphisms and haplotypes that are markers for high or low levels of transcript expression. Polymorphism discovery and analysis at the BT3N2A2 locus allowed for the identification of a “regulatory” haplotype containing SNPs in complete LD with the presence of allelic imbalance and high correlation with BT3N2A2 expression levels. Increased sample size and marker informativity should allow similar correlations in other genes.

In addition, we propose that the allelic imbalance pipeline can be used to screen the genome for genes that manifest random monoallelic expression. Cataloging genes exhibiting random monoallelic expression may provide clues regarding the pathogenesis of gene dosage-sensitive disorders, such as cancer, in which disturbed epigenetic control in imprinted genes is well documented (7).

We plan to expand our allelic imbalance screening from the moderate number of genes focused on LCLs, to genome-wide coverage in multiple tissues. Emerging high-throughput genotyping technologies (24) may provide the necessary tools to achieve this. Genome-wide allelic imbalance discovery coupled with elucidation of the underlying mechanisms as outlined in this study may provide a comprehensive view of cis-acting transcriptional regulation as well as functional markers for human genes that are commonly modulated by regulatory polymorphisms.

ACKNOWLEDGMENTS

We thank Anthony Long for providing the genotype-calling algorithm used with the FP-SBE genotyping platform. We thank Kazuhiko Nakabayashi for suggesting the control genes. We acknowledge Patrick Beaulieu, Simon Drouin, Helene Belanger, Sylvie Langlois, Sarah Bourgoin, Corine Zotti, and Marie-Christine Theberge for technical assistance. We thank Picard Marceau and Denis Richard for the adipose tissue samples. We thank Annette Hollman for providing clonal LCL lines.

GRANTS

Research funds were provided by Genome Quebec and Genome Canada. T. Pastinen was partially supported by grants from the Academy of Finland, Finnish Heart Foundation, and the Maud Kuistila Foundation. M.-C. Vohl is supported by Fonds de la Recherche en Santé du Québec. A. K. Naumova is a Canadian Institutes of Health Research (CIHR) New Investigator. T. J. Hudson is supported by a Clinician-Scientist Award in Translational Research by the Burroughs Wellcome Fund and an Investigator Award from the Canadian Institutes of Health Research.

REFERENCES


