Analysis of allelic differential expression in the human genome using allele-specific serial analysis of gene expression tags

Daniel Onofre Vidal, Jorge Estefano S. de Souza, Lilian Campos Pires, Cibele Masotti, Anna Christina Matos Salim, Maria Cristina Ferreira Costa, Pedro Alexandre Favoretto Galante, Sandro José de Souza, and Anamaria Aranha Camargo

Abstract: Recent reports have demonstrated that a significant proportion of human genes display allelic differential expression (ADE). ADE is associated with phenotypic variability and may contribute to complex genetic diseases. Here, we present a computational analysis of ADE using allele-specific serial analysis of gene expression (SAGE) tags representing 1295 human genes. We identified 472 genes for which unequal representation (>3-fold) of allele-specific SAGE tags was observed in at least one SAGE library, suggesting the occurrence of ADE. For 235 out of these 472 genes, the difference in the expression level between both allele-specific SAGE tags was statistically significant ($p < 0.05$). Eleven candidate genes were then subjected to experimental validation and ADE was confirmed for 8 out of these 11 genes. Our results suggest that at least 25% of the human genes display ADE and that allele-specific SAGE tags can be efficiently used for the identification of such genes.

Key words: allelic differential expression, allele-specific SAGE tags, monoallelic expression, SAGE.

Résumé : De récents travaux ont démontré qu'une proportion significative des gènes humains présentent une expression allélique différentielle (ADE). L'ADE est associée à la variabilité phénotypique et pourrait contribuer à des maladies génétiques complexes. Dans ce travail, les auteurs présentent une analyse bioinformatique de l'ADE au moyen d'étiquettes SAGE allèle-spécifiques pour 1295 gènes humains. Les auteurs ont identifié 472 gènes pour lesquels une abondance différentielle (>3 fois) d'étiquettes SAGE allèle-spécifiques a été observée au sein d'au moins une banque d'étiquettes SAGE; ce qui suggère l'occurrence du phénomène d'ADE. Pour 235 de ces 472 gènes, la différence quant au niveau d'expression des deux étiquettes SAGE était statistiquement significative ($p < 0,05$). Onze gènes candidats ont ensuite été assujettis à une validation expérimentale et l'ADE a été confirmée pour 8 de ces 11 gènes. Ces résultats suggèrent qu'au moins 25 % des gènes humains présentent de l'ADE et que les étiquettes SAGE allèle-spécifiques constituent un outil efficace pour identifier de tels gènes.

Mots-clés : expression allélique différentielle, étiquettes SAGE allèle-spécifiques, expression monoallélique, SAGE.

[Traduit par la Rédaction]

Introduction

Until recently, it was generally assumed that in diploid eukaryotic organisms both alleles of each gene were expressed at the same level and that allele-specific differences in expression levels were restricted to imprinted genes, X-chromosome-inactivated genes, and a few autosomal genes. However, recent reports have demonstrated that a significant proportion of non-imprinted autosomal human genes display allelic differential expression (ADE) (Yan et al. 2002; Lo et al. 2003; Ge et al. 2005; Pant et al. 2006; Serre et al. 2008; Palacios et al. 2009) and that allele-specific differences in expression levels are heritable, involving both genetic and epigenetic mechanisms (Yan et al. 2002; Pastinen et al. 2004).

ADE has been associated with phenotypic variability between individuals and may contribute to both Mendelian and complex genetic diseases (Milani et al. 2007; Wilkins et al. 2007; Jordheim et al. 2008; Milani et al. 2009). Because of their implications for human health, several high-throughput methods measuring the relative expression level of different alleles using intragenic polymorphisms have

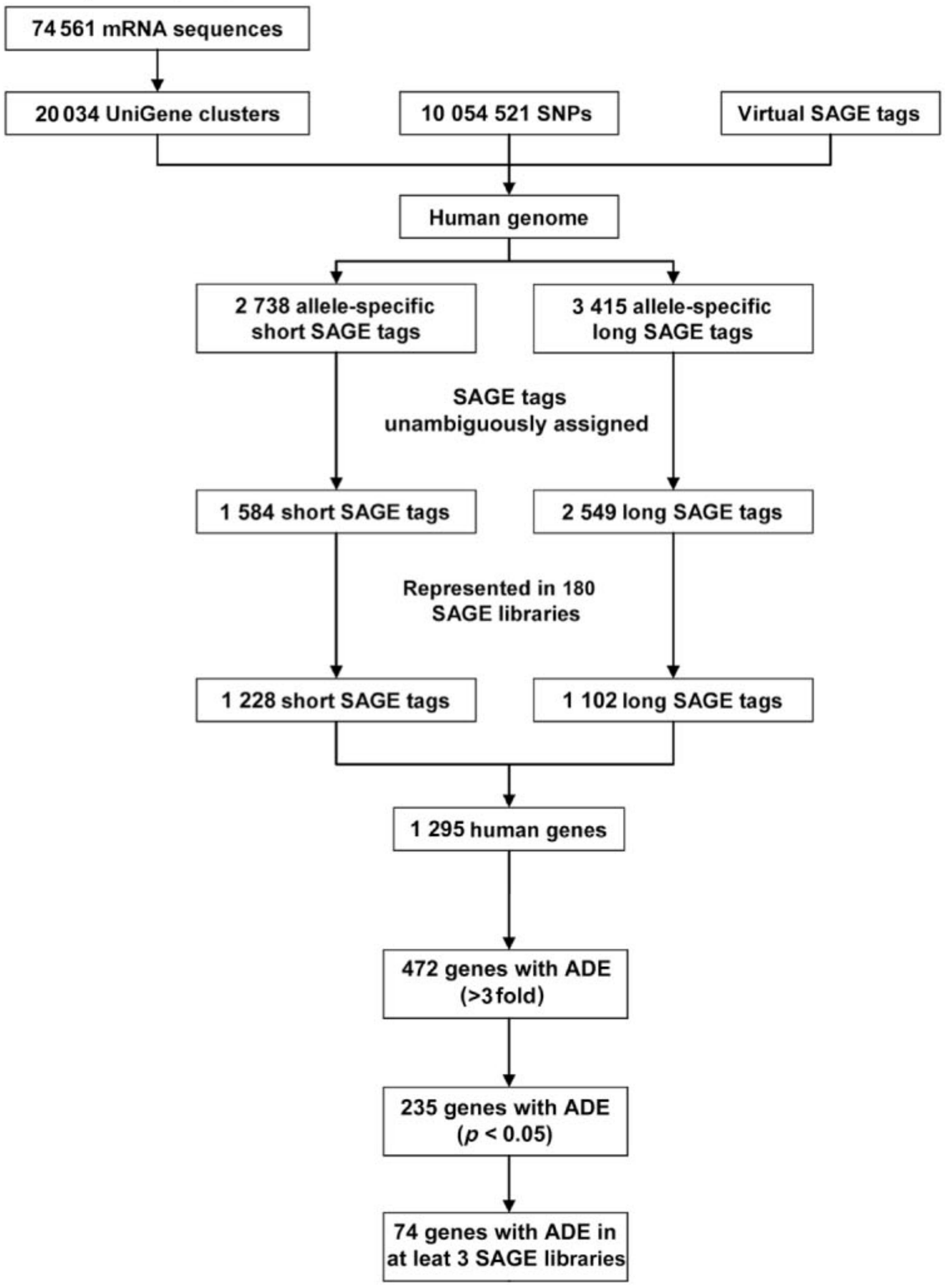
Received 7 July 2010. Accepted 19 October 2010. Published on the NRC Research Press Web site at genome.nrc.ca on 25 January 2011.

Corresponding Editor: A. Naumova.

D.O. Vidal,[1] J.E.S. de Souza,[1] L.C. Pires, C. Masotti, A.C.M. Salim, M.C.F. Costa, P.A.F. Galante, S.J. de Souza, and A.A. Camargo.[2] Ludwig Institute for Cancer Research, São Paulo Branch, Rua João Julião, 245 – 1°Andar, 01323-903, São Paulo, SP, Brazil.

[1]These authors contributed equally to this study.
[2]Corresponding author (e-mail: anamaria@compbio.ludwig.org.br).

Fig. 1. Representation of the computational approach used for the identification of allele-specific serial analysis of gene expression (SAGE) tags and allelic differential expression (ADE).



been applied to identify genes displaying ADE. Together these studies demonstrate that approximately 20%–65% of the human genes display ADE (Yan et al. 2002; Lo et al. 2003; Ge et al. 2005; Pant et al. 2006; Serre et al. 2008; Palacios et al. 2009).

Serial analysis of gene expression (SAGE) is a powerful technique for genome-wide analysis of gene expression that is capable of measuring expression levels without a priori knowledge of the transcript sequence. In the SAGE technique, a short sequence tag with a variable length (10 or 17 nucleotides) adjacent to the 3′ most *Nla*III restriction site is extracted from each transcript (Velculescu et al. 1995). The extracted tags are then concatenated for high-throughput sequencing and tag counts are used to measure the relative abundance of their corresponding transcripts.

We have previously analyzed the impact of single nucleotide polymorphisms (SNPs) on the generation of allele-specific SAGE tags (Silva et al. 2004). The identification of allele-specific SAGE tags was achieved through the construction of a reference database in which the analysis of mRNA sequences from UniGene was combined with information available from the National Center for Biotechnology Information (NCBI) SNP database (Sherry et al. 2001) and SAGE Genie (Boon et al. 2002). Allele-specific SAGE tags were identified by analyzing the presence of SNPs within the original SAGE tag sequence or SNPs cre-

Table 1. Allele-specific SAGE expression data for a gene displaying ADE.

Gene name:	*NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 4, 9kDa (NDUFA4)*				
UniGene:	Hs.50098				
Original tag:	(long) TTGGAGATCTCTATTGT				
Alternative tag:	(long) TTGGAGATCTCTATTGC				
SNP ID:	rs1804855				
SAGE library No.	Origin	Tissue	Original tag frequency[a]	Alternative tag frequency[a]	*p* value (binomial)
675	T	Mammary gland	24	0	0.000000*
673	**T**	**Mammary gland**	**28**	**4**	**0.000010***
963	T	Lung	16	0	0.000015*
1566	**N**	**White blood cells**	**18**	**1**	**0.000038***
655	N	Vascular	11	0	0.000488*
645	**T**	**Mammary gland**	**13**	**1**	**0.000916***
1564	**N**	**White blood cells**	**13**	**1**	**0.000916***
653	T	Colon	9	0	0.001953*
1563	N	White blood cells	9	0	0.001953*
657	T	Mammary gland	8	0	0.003906*
1567	N	White blood cells	8	0	0.003906*
723	**T**	**Mammary gland**	**9**	**1**	**0.010742***
703	T	Mammary gland	6	0	0.015625*
683	T	Mammary gland	5	0	0.031250*
649	T	Mammary gland	4	0	0.062500
1565	N	White blood cells	4	0	0.062500
1645	**T**	**Brain**	**3**	**1**	**0.312500**

Note: Presence of both allele-specific SAGE tags in SAGE libraries is in bold. *, SAGE libraries displaying significant allele-differential expression ($p < 0.05$); N, normal; T, tumoral.

[a]Tag frequencies correspond to absolute count tags in each SAGE library.

ating or disrupting *Nla*III sites used for SAGE library construction (Silva et al. 2004).

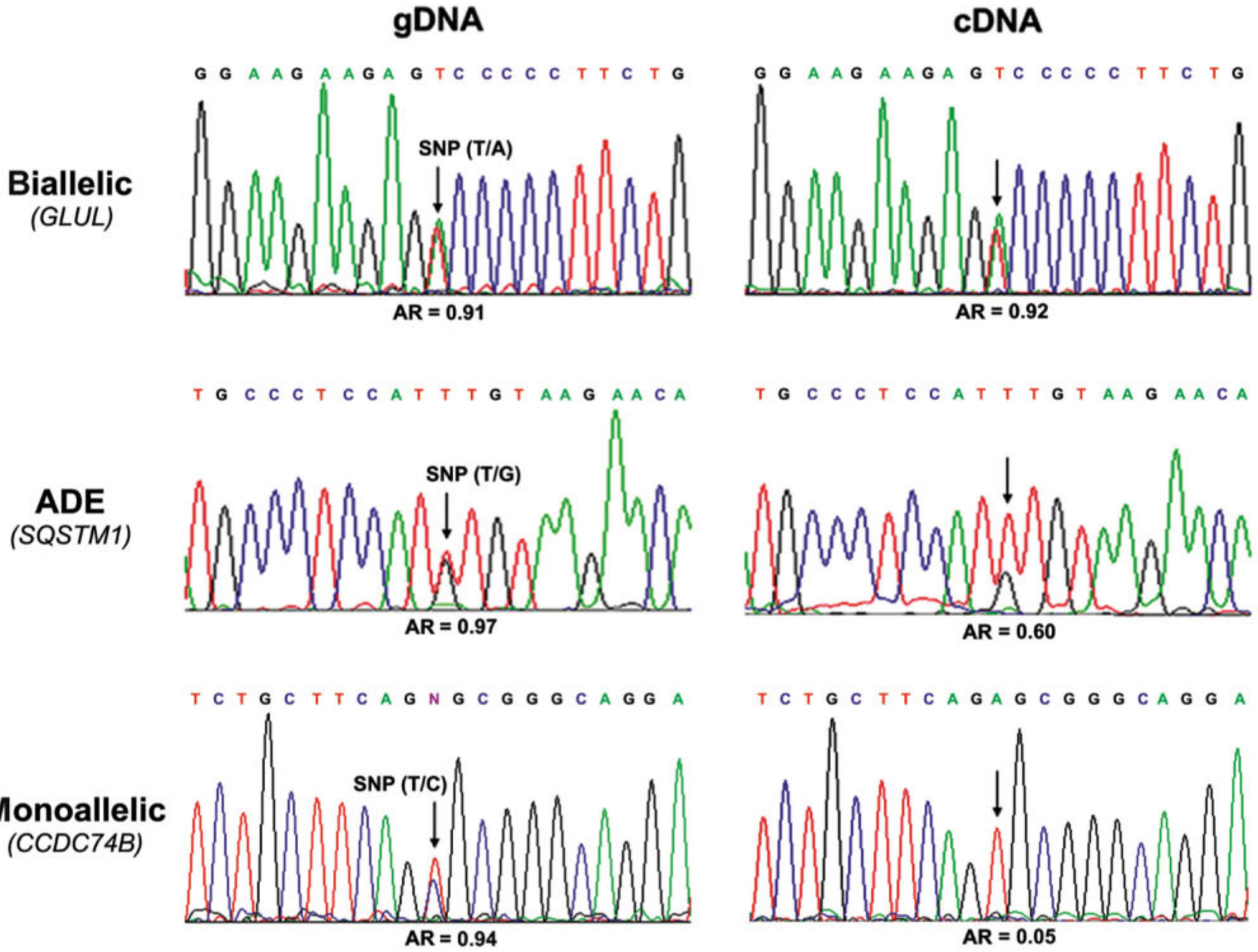
In the present work, we developed a computational method to identify genes that display ADE using our database of allele-specific SAGE tags and publicly available SAGE expression data. We first identified SAGE libraries in which both allele-specific SAGE tags were present and measured allelic variation in gene expression based on the SAGE tag counts for each allele. We identified 472 genes for which allele-specific SAGE tags were concomitantly expressed with a frequency difference >3-fold in at least one SAGE library, suggesting the occurrence of ADE. For 235 out of these 472 genes, the difference in the expression level between both allele-specific SAGE tags was statistically significant ($p < 0.05$). Complementary DNA (cDNA) sequencing of heterozygotes was then used to validate a subset of genes displaying ADE. Our results suggest that at least 25% of the human genes display ADE and that allele-specific SAGE tags can be efficiently used for the identification of such genes.

Materials and methods

Identification of allele-specific SAGE tags

Identification of allele-specific SAGE tags was carried out as previously described (Silva et al. 2004) except that updated versions of UniGene and NCBI SNP database were used in the present analysis. A total of 74 560 mRNA sequences containing a poly(A) tail and corresponding to 20 034 human genes according to UniGene (build #198) were mapped onto the publicly available human genome sequence (build #35.1) using BLAT (http://genome.ucsc.edu/) and Sim4 (Florea et al. 1998). Spurious and multiple alignments were eliminated by using an additional set of alignment criteria. These included a minimum identity of 93% and coverage (percentage of sequence length aligned) of >55%. Sequences mapping to more than one location on the genome were given a score for alignment quality. A higher score was associated with a higher identity and coverage. Only the alignments with the highest scores were kept in the database. Poly(A)-containing mRNA sequences were then scanned for the presence of *Nla*III restriction sites and virtual SAGE tags downstream of the 3′ most restriction site were extracted and denominated original tags. A total of 10 054 521 SNPs from the NCBI SNP database (build #124) were also mapped to the human genome sequence. Mapping was achieved through the alignment of sequences flanking the SNPs according to the NCBI criteria for SNP mapping. We have restricted our analysis to SNPs that mapped only once to the human genome sequence. A MySQL database (available at http://www.mysql.com/) was loaded with mapping information for all mRNAs, SNPs, and original SAGE tags that shared an overlap in genomic coordinates. Alternative allele-specific SAGE tags were then identified by crossing mapping information stored in the relational database and by analyzing the presence of SNPs within the original SAGE tag sequence or SNPs affecting (creating or disrupting) the restriction enzyme site used for SAGE library construction. The complete list of allele-specific SAGE tags and related information are available upon request. Alternative allele-specific SAGE tags were then compared with a list of experimentally obtained tags and were used to measure ADE using SAGE expression data available in public databases.

Fig. 2. Experimental validation of allelic differential expression (ADE) using gDNA and cDNA direct sequencing. Representative sequencing results for genes with samples presenting bialleic expression (*GLUL*), ADE (*SQSTM1*), and monoallelic expression (*CCDC74B*). Sequencing traces from gDNA and cDNA samples are represented in the left and right panels, respectively. SNPs are indicated by the arrows; AR means adjusted peak ratio.



Analysis of ADE using SAGE data

SAGE expression data was obtained from SAGE Genie (http://cgap.nci.nih.gov/SAGE). SAGE library descriptions were manually curated to exclude libraries that were made with pooled RNA from different individuals or libraries that were derived from a same individual. ADE was measured using the frequency of each allele-specific SAGE tag in 180 SAGE libraries (163 short SAGE libraries and 17 long SAGE libraries), assuming that curated libraries represent different single individuals. A complete list of curated SAGE libraries is available as supplementary data (Table S1).[3] We then searched for genes displaying ADE and selected from our database genes that showed at least 3-fold difference in the frequencies of both allele-specific tags in at least one SAGE library. Differential expression of both allele-specific tags in each SAGE library was further confirmed using a binomial test. X-linked genes were removed from ADE analysis, as their observed allelic bias could be influenced by tissue clonality.

Biological samples

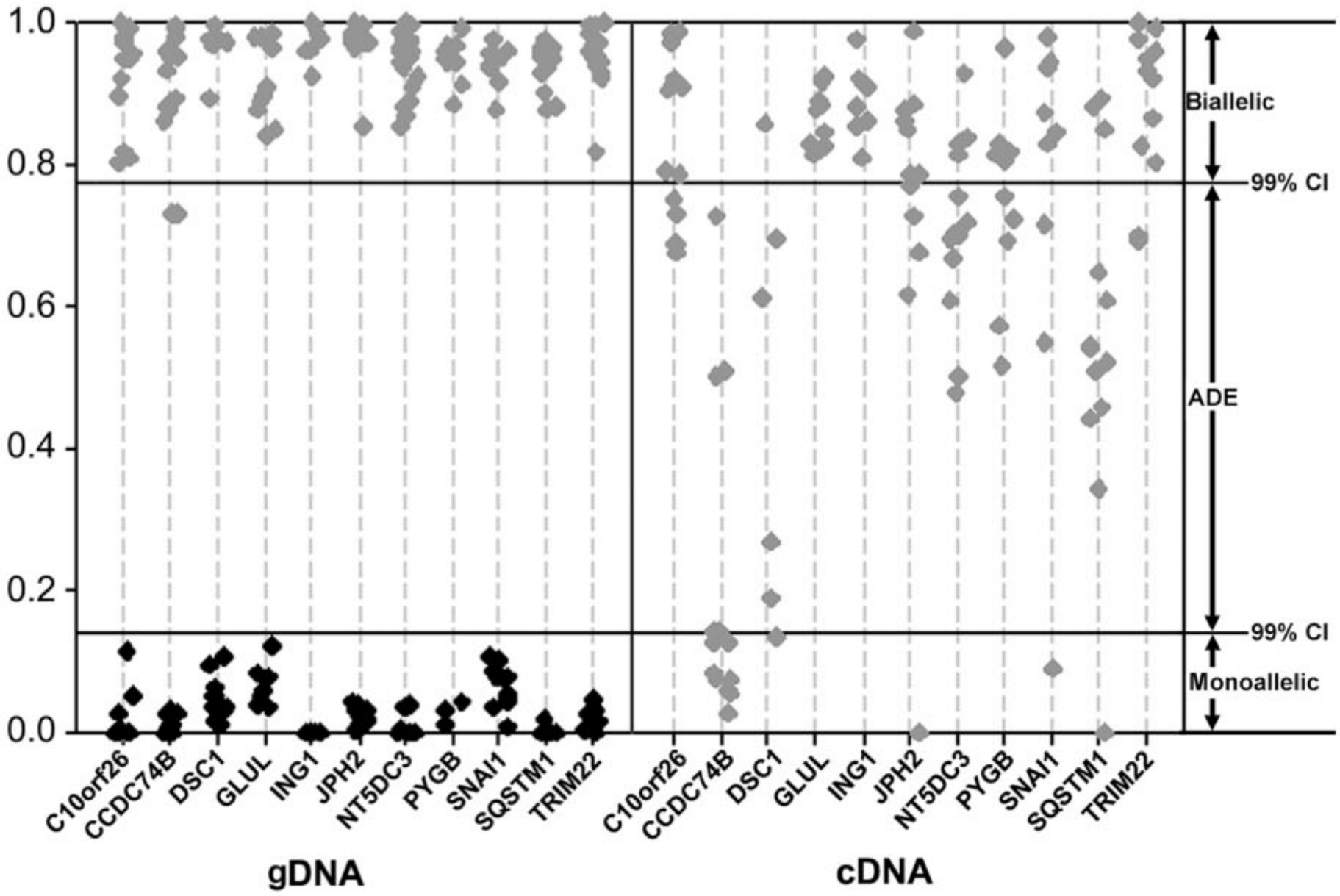
Blood samples were obtained from the blood bank of Hospital A.C. Camargo, São Paulo, Brazil and were processed immediately after collection. Samples were collected after informed consent and the study was approved by the institution's ethic committee.

DNA extraction and genotyping

Genomic DNA (gDNA) was isolated by digestion with proteinase K (Invitrogen, Carlsbad, California), followed by phenol–chloroform extractions. PCR primers used for genotyping (Supplementary data, Table S2) were designed flanking the SNP region associated with the allele-specific SAGE tag and avoiding other known SNPs within the primer sequences. Sequences corresponding to the universal sequencing primer M13 (5′-GTAAAACGACGGCCAGT-3′) were appended to the forward or reverse primers used for genotyping to apply the same sequencing conditions to all PCR fragments. PCR was carried out in a final volume of 25 μL,

[3] Supplementary data for this article are available on the journal Web site (http://genome.nrc.ca).

Fig. 3. Allelic differential expression (ADE) analysis in blood samples. Graphic representation of adjusted peak ratios (AR) for gDNA (left) and cDNA (right) samples from heterozygotes of the 11 genes assayed for ADE. Upper and lower black lines represent the 99% confidence intervals (CI) for biallelic and monoallelic expression calculated using the AR from heterozygotes and homozygotes, respectively. Grey and black squares represent the AR from heterozygotes and homozygotes, respectively.



containing 50 ng of gDNA, 1× *Taq* platinum DNA polymerase buffer (Invitrogen), 1.5 mmol/L $MgCl_2$, 0.2 mmol/L dNTPs, 0.4 μmol/L of each primer, and 1 U of *Taq* platinum DNA polymerase (Invitrogen). PCR conditions were 95 °C for 2 min, followed by 35 cycles at 95 °C for 35 s, 60 °C for 35 s, and 72 °C for 40 s. Reactions were kept at 72 °C for 6 min after the last cycle. The amplified products were treated with 10 U of exonuclease and 1 U of shrimp alkaline phosphatase (USB, Cleveland, Ohio) according to manufacturer's instructions and used for direct sequencing using BigDye Terminator (Applied Biosystems, Foster City, California) and an ABI3130 sequencer (Applied Biosystems). Sequence traces were manually inspected to identify heterozygotes.

RNA extraction and RT-PCR

RNA was isolated using Trizol (Invitrogen) according to the manufacturer's instructions. The RNA quality was verified on agarose gels and 50 μg of total RNA were treated with 8 U of DNase I (Ambion, Austin, Texas) for 40 min. at 37 °C, extracted with phenol–chloroform, and reprecipitated. The presence of residual gDNA contamination after DNase treatment was detected using PCR amplification of intronic sequences of the *hMLH-1* gene (forward, 5′-TGGTGTCTCTAGTTCTGG-3′; reverse 5′-CATTGTTGTAGTAGCTCTGC-3′). RNA samples with positive *hMLH-1* amplification were subjected to another round of DNase treatment.

cDNA synthesis was performed using 1 μg DNA-free RNA, oligo(dT) primer, and Superscript II reverse transcriptase (Invitrogen) according to manufacturer's instruction. Primers used for RT-PCR were the same used for gDNA amplification and genotyping. RT-PCRs were carried out in a final volume of 25 μL, containing 1 μL of cDNA, 1× *Taq* platinum DNA polymerase buffer (Invitrogen), 1.5 mmol/L $MgCl_2$, 0.2 mmol/L dNTPs, 0.4 μmol/L of each primer, and 1 U of *Taq* platinum DNA polymerase (Invitrogen). PCR conditions were the same used for genotyping. RT-PCR products were purified and sequenced as described above.

Allele-specific expression analysis using high sensitivity sequencing

Allele expression levels were determined using the PeakPicker software (Ge et al. 2005) specifically developed to quantify the relative amount of two alleles by measuring the peak intensity of the two polymorphic bases from sequence traces. Subsets of at least five informative heterozygotes for each SNP associated with the allele-specific SAGE tag were initially identified, and their gDNA was amplified in identical conditions to establish the sequence peak intensity ratio between polymorphic bases that would correspond to a 50:50 representation of both alleles. Because peak heights vary depending on sample, base type, and their position within the sequence, the PeakPicker software carries out a normalization step in which the SNP allele peak intensity is compared with the intensity of reference peaks in the flanking sequence. We limited our analysis to sequence traces in which 80% of the bases within a 21 base window flanking the SNP presented a Phred quality score of >15.

A text file with SNP allele peak intensity ratios normalized by the reference peaks is generated as an output. Peak intensity ratios were calculated for gDNA and cDNA samples from all informative heterozygotes. Ratios with values above 1 were transformed to 1/ratio to set all ratios in a 0–1 scale and then adjusted to the mean of the peak intensity ratios from DNA samples. The adjusted peak intensity ratios (AR) of DNA samples from heterozygotes were used to estimate the methodological variability and establish a 99% confidence interval (CI) for equal representation of both alleles. The 99% CI was calculated assuming that normalized peak height ratios of DNA samples are normally distributed according to the Anderson–Darling test. A similar approach was used to identify samples displaying monoallelic expression. However, in this case, normalized ratios of homozygote DNA samples were used to establish a 99% CI.

Results

ADE analysis using allele-specific SAGE expression data

We have previously demonstrated that allele-specific SAGE tags can be generated owing to the presence of SNPs creating or disrupting *Nla*III sites used for SAGE library construction or within the original tag sequence (Supplementary data, Fig. S1). The identification of allele-specific SAGE tags was achieved through the construction of a reference database in which the analysis of mRNA sequences from UniGene was combined with information available from the NCBI SNP database and SAGE Genie (Silva et al. 2004).

After updating our reference database, a total of 2738 allele-specific short SAGE tags and 3415 allele-specific long SAGE tags, representing 2892 known human genes, were identified (Fig. 1). Alternative allele-specific SAGE tags generated by SNPs that disrupted the *Nla*III restriction site (640 short SAGE tags and 640 long SAGE tags) were not used for ADE analysis, as allele-specific SAGE tags located upstream of the original SAGE tag could also be associated with other biological phenomena like alternative polyadenylation or alternative splicing. It should be noted, however, that this is a conservative approach that might lead to an underestimation of candidate genes displaying ADE.

Of the remaining allele-specific SAGE tags, 1584 short SAGE tags (10 bp) and 2549 long SAGE tags (17 bp) could be unambiguously assigned to a known human transcript. Of those unambiguous tags, 1228 allele-specific short SAGE tags and 1102 allele-specific long SAGE tags were represented in SAGE libraries derived from single individuals (Fig. 1). These tags were used for allele-specific expression analysis of 1295 genes for which both allele-specific SAGE tags were represented in curated SAGE libraries. The complete list of genes under analysis is available as supplementary data (Table S3). Allele-specific expression was measured using the frequency of each allele-specific SAGE tag in these libraries.

We then searched for genes displaying ADE and selected from our database 472 genes that showed at least 3-fold difference in the frequencies of both allele-specific tags in at least one SAGE library. For 235 out of these 472 genes, the difference in the expression level between both allele-specific SAGE tags was statistically significant ($p < 0.05$), and for 74 of these genes, differential expression was observed in more than three SAGE libraries presenting both allele-specific SAGE tags. The complete list of genes displaying ADE and their allele-specific SAGE expression data is available as supplementary data (Table S4).

A representative example of a gene displaying ADE is presented in Table 1. The *NDUFA4* gene (Hs.50098) presents two allele-specific SAGE tags that differ by the presence of a SNP (rs1804855) within the tag sequence. *NDUFA4* expression was observed in 17 SAGE libraries derived mainly from mammary gland and white blood cells. The presence of both allele-specific SAGE tags was detected in six out of these 17 libraries and differential expression between the two alleles ($p < 0.05$) was observed in five of these six libraries.

Allele-specific gene expression analysis using high sensitivity sequencing

Twenty candidates from the list of 472 genes displaying ADE were randomly selected for allele-specific gene expression analysis in blood samples. For 11 of these 20 candidates, we were able to identify at least five heterozygotes to carry out the analysis. These 11 genes were represented on average in 66 different SAGE libraries (min. 22 – max. 146 libraries) and the presence of both tags in a same SAGE library was detected on average in six libraries (min. 2 – max. 11). ADE for these 11 genes was detected on average in two different libraries (min. 1 – max. 4).

Experimental validation was carried out by direct sequencing of gDNA and cDNA fragments from heterozygous individuals, followed by comparison between peak intensity ratios corresponding to the polymorphic bases in the sequencing traces from gDNA and cDNA samples (Fig. 2). The peak intensity ratios from DNA samples of heterozygotes were used to estimate the methodological variability and to establish a 99% CI for equal representation of both alleles (50:50 ratio). If normalized peak height ratios in cDNA samples showed significant deviation beyond the 99% CI established from gDNA data, the sample was classified as displaying ADE. A similar approach was used to identify samples displaying monoallelic expression. However, in this case, normalized ratios of homozygote DNA samples were used to establish the 99% CI and cDNA samples showing ratios within the 99% CI were considered as displaying monoallelic expression.

Of the 11 genes subjected to allele-specific gene expression analysis, 8 presented ADE in at least 20% of the heterozygotes (Fig. 3). Moreover, ADE was detected in more than 75% of the heterozygotes analyzed for the *SQSTM1*, *CCDC74B*, and *DSC1* genes. Interestingly, monoallelic expression was also detected in some heterozygotes for *CCDC74B*, *SQSTM1*, and *DSC1* genes (Fig. 3). Together, these results demonstrate that allele-specific SAGE tags can be efficiently used to measure allele-specific gene expression to identify genes displaying ADE.

Discussion

ADE has been observed in several species and can result

from *cis*-acting sequence polymorphisms that affect the rate of transcription or from epigenetic modifications that cause complete or partial suppression of one allele (Pastinen et al. 2004). Expression level differences between alleles are directly associated with phenotypic variability and have been linked to the development of common genetic disorders in humans (Milani et al. 2007, 2009; Wilkins et al. 2007; Jordheim et al. 2008).

Several genome-wide computational and experimental studies have been carried out to identify genes displaying ADE (Yan et al. 2002; Lo et al. 2003; Ge et al. 2005; Pant et al. 2006; Serre et al. 2008; Palacios et al. 2009). Collectively, these studies revealed that approximately 20%–60% of all non-imprinted autosomal human genes display ADE. However, it has already been shown that ADE can be context specific with regard to cell type and developmental stage and that for some genes even small differences in expression level between alleles are physiologically relevant (Wilkins et al. 2007). In this context, the development of tools for identification of genes displaying ADE and, most importantly, for measuring differences in allele expression levels will allow us to dissect the genetic and epigenetic mechanisms underlying ADE and will contribute to a better understanding of phenotypic variability in humans.

In this study, we developed a computational method to identify genes displaying ADE and to measure allele-specific expression using publicly available SAGE data. Unambiguous allele-specific SAGE tags were identified for 1295 (6.5%) human genes, and allele-specific gene expression was measured by determining the frequency of allele-specific SAGE tags in 180 SAGE libraries derived from different single individuals. The applicability of our approach was experimentally confirmed by cDNA sequencing of heterozygotes. Eight out of 11 genes (73%) selected for experimental validation displayed ADE in at least 20% of the heterozygotes. If we consider that the 1295 genes under analysis is a representative subset of all human genes and taking into account our experimental validation efficiency (73%), our analysis suggests that at least 25% of all human genes display ADE and confirm previous estimates made with other computational and experimental methods (Lo et al. 2003; Pant et al. 2006; Palacios et al. 2009).

Interestingly, we noticed that for 442 genes out of the 1295 genes under analysis, both allele-specific SAGE tags were never concomitantly observed in any SAGE library, a finding that is compatible with genes displaying monoallelic expression. However, this number might include a significant number of false positives, because of the number of SAGE libraries in which expression of these genes was detected, as well as a low frequency of one of the alleles in the population, would directly influence the probability of detecting the simultaneous expression of both allele-specific SAGE tags.

One major limitation of our approach is the yet relatively small number of SAGE libraries available in public databases and the low expression level of a significant fraction of genes under analysis, which are represented by a small number of SAGE tags. These issues might lead to an underestimation of the number of genes displaying ADE and to an overestimation of genes displaying monoallelic expression. Another important limitation is the yet relatively small number of human genes with allele-specific SAGE tags, which in turn is related to limited information on the distribution of SNPs in the human genome. All these limitations will, nevertheless, be surpassed in the near future by the disseminated use of next generation sequencers for the generation of expression data and for identification of new SNPs.

The SAGE protocol has been recently adapted to be used in combination with next generation sequencing platforms such as Illumina GA and SOLiD. These adapted protocols allow for the highest level of expression profile sensitivity and quantification with minimal sequencing requirements, and they will remain the choice protocols for gene expression quantification, especially when a large number of samples is to be analyzed. Moreover, although more recent techniques such as RNA-Seq (Wang et al. 2009) will allow broader gene coverage and improve SNP detection, the application of this type of data for ADE studies is not straightforward. First, SNP calling in next generation sequencing data is still complex and one of the major advantages of our approach is that SNPs under analysis are experimentally confirmed in an independent fashion by sensitivity to restriction enzyme digestion. Second, events of alternative splicing (which are less common in the 3′ end of the transcripts) will directly interfere in allele representation, increasing the complexity of the analysis. Finally, because sequence coverage is not equally distributed along the transcript, the use of RNA-Seq data will introduce further complexity to the analysis if more than one SNP per transcript is considered.

In conclusion, we have demonstrated that allele-specific SAGE tags can be efficiently used to measure allele-specific gene expression and that SAGE data provide a valuable resource for studying phenotypic variation and complex diseases. To our knowledge, this is the first time that SAGE data is used for allele-specific expression analysis. We anticipate that the applicability of our approach will certainly increase in the near future with the disseminated use of next generation sequencing technologies.

Acknowledgements

This work was supported by the Ludwig Institute for Cancer Research; Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq); and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e Fundação de Amparo a Pesquisa do Estado de São Paulo (FAPESP).

References

Boon, K., Osorio, E.C., Greenhut, S.F., Schaefer, C.F., Shoemaker, J., Polyak, K., et al. 2002. An anatomy of normal and malignant gene expression. Proc. Natl. Acad. Sci. U.S.A. **99**(17): 11287–11292. doi:10.1073/pnas.152324199. PMID:12119410.

Florea, L., Hartzell, G., Zhang, Z., Rubin, G.M., and Miller, W.A. 1998. A computer program for aligning a cDNA sequence with a genomic DNA sequence. Genome Res. **8**(9): 967–974. doi:10.1101/gr.8.9.967 . PMID:9750195.

Ge, B., Gurd, S., Gaudin, T., Dore, C., Lepage, P., Harmsen, E., et al. 2005. Survey of allelic expression using EST mining. Genome Res. **15**(11): 1584–1591. doi:10.1101/gr.4023805. PMID:16251468.

Jordheim, L.P., Nguyen-Dumont, T., Thomas, X., Dumontet, C.,

and Tavtigian, S.V. 2008. Differential allelic expression in leukoblast from patients with acute myeloid leukemia suggests genetic regulation of *CDA*, *DCK*, *NT5C2*, *NT5C3*, and *TP53*. Drug Metab. Dispos. **36**(12): 2419–2423. doi:10.1124/dmd.108.023184. PMID:18775979.

Lo, H.S., Wang, Z., Hu, Y., Yang, H.H., Gere, S., Buetow, K.H., and Lee, M.P. 2003. Allelic variation in gene expression is common in the human genome. Genome Res. **13**(8): 1855–1862. doi:10.1101/gr.1006603. PMID:12902379.

Milani, L., Gupta, M., Andersen, M., Dhar, S., Fryknäs, M., Isaksson, A., et al. 2007. Allelic imbalance in gene expression as a guide to *cis*-acting regulatory single nucleotide polymorphisms in cancer cells. Nucleic Acids Res. **35**(5): e34. doi:10.1093/nar/gkl1152. PMID:17267408.

Milani, L., Lundmark, A., Nordlund, J., Kiialainen, A., Flaegstad, T., Jonmundsson, G., et al. 2009. Allele-specific gene expression patterns in primary leukemic cells reveal regulation of gene expression by CpG site methylation. Genome Res. **19**(1): 1–11. doi:10.1101/gr.083931.108. PMID:18997001.

Palacios, R., Gazave, E., Goñi, J., Piedrafita, G., Fernando, O., Navarro, A., et al. 2009. Allele-specific gene expression is widespread across the genome and biological processes. PLoS One, **4**(1): e4150. doi:10.1371/journal.pone.0004150. PMID:19127300.

Pant, P.V., Tao, H., Beilharz, E.J., Ballinger, D.G., Cox, D.R., and Frazer, K.A. 2006. Analysis of allelic differential expression in human white blood cells. Genome Res. **16**(3): 331–339. doi:10.1101/gr.4559106. PMID:16467561.

Pastinen, T., Sladek, R., Gurd, S., Sammak, A., Ge, B., Lepage, P., et al. 2004. A survey of genetic and epigenetic variation affecting human gene expression. Physiol. Genomics, **16**(2): 184–193. doi:10.1152/physiolgenomics.00163.2003. PMID:14583597.

Serre, D., Gurd, S., Ge, B., Sladek, R., Sinnett, D., Harmsen, E., et al. 2008. Differential allelic expression in the human genome: a robust approach to identify genetic and epigenetic *cis*-acting mechanisms regulating gene expression. PLoS Genet. **4**(2): e1000006. doi:10.1371/journal.pgen.1000006. PMID:18454203.

Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. 2001. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. **29**(1): 308–311. [PMID: 11125122.] doi:10.1093/nar/29.1.308. PMID:11125122.

Silva, A.P., De Souza, J.E., Galante, P.A., Riggins, G.J., De Souza, S.J., and Camargo, A.A. 2004. The impact of SNPs on the interpretation of SAGE and MPSS experimental data. Nucleic Acids Res. **32**(20): 6104–6110. doi:10.1093/nar/gkh937. PMID: 15562001.

Velculescu, V.E., Zhang, L., Vogelstein, B., and Kinzler, K.W. 1995. Serial analysis of gene expression. Science, **270**(5235): 484–487. [PMID: 7570003.] doi:10.1126/science.270.5235.484. PMID:7570003.

Wang, Z., Gerstein, M., and Snyder, M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. **10**(1): 57–63. doi:10.1038/nrg2484. PMID:19015660.

Wilkins, J.M., Southam, L., Price, A.J., Mustafa, Z., Carr, A., and Loughlin, J. 2007. Extreme context specificity in differential allelic expression. Hum. Mol. Genet. **16**(5): 537–546. doi:10.1093/hmg/ddl488. PMID:17220169.

Yan, H., Yuan, W., Velculescu, V.E., Vogelstein, B., and Kinzler, K.W. 2002. Allelic variation in human gene expression. Science, **297**(5584): 1143. doi:10.1126/science.1072545. PMID:12183620.