Published online 31 January 2007

Nucleic Acids Research, 2007, Vol. 35, No. 5 ***e34***
doi:10.1093/nar/gkl1152

Allelic imbalance in gene expression as a guide to *cis*-acting regulatory single nucleotide polymorphisms in cancer cells

Lili Milani[1], Manu Gupta[1], Malin Andersen[3], Sumeer Dhar[2], Mårten Fryknäs[2], Anders Isaksson[2], Rolf Larsson[2] and Ann-Christine Syvänen[1,*]

[1]Molecular Medicine, [2]Clinical Pharmacology, Department of Medical Sciences, Uppsala University, Uppsala, Sweden and [3]Department of Biotechnology, AlbaNova University Center, Royal Institute of Technology (KTH), Stockholm, Sweden

Received November 3, 2006; Revised and Accepted December 6, 2006

ABSTRACT

Using the relative expression levels of two SNP alleles of a gene in the same sample is an effective approach for identifying *cis*-acting regulatory SNPs (rSNPs). In the current study, we established a process for systematic screening for *cis*-acting rSNPs using experimental detection of AI as an initial approach. We selected 160 expressed candidate genes that are involved in cancer and anticancer drug resistance for analysis of AI in a panel of cell lines that represent different types of cancers and have been well characterized for their response patterns against anticancer drugs. Of these genes, 60 contained heterozygous SNPs in their coding regions, and 41 of the genes displayed imbalanced expression of the two cSNP alleles. Genes that displayed AI were subjected to bioinformatics-assisted identification of rSNPs that alter the strength of transcription factor binding. rSNPs in 15 genes were subjected to electrophoretic mobility shift assay, and in eight of these genes (*APC*, *BCL2*, *CCND2*, *MLH1*, *PARP1*, *SLIT2*, *YES1*, *XRCC1*) we identified differential protein binding from a nuclear extract between the SNP alleles. The screening process allowed us to zoom in from 160 candidate genes to eight genes that may contain functional rSNPs in their promoter regions.

INTRODUCTION

Single nucleotide polymorphisms (SNPs) in genomic regions that regulate gene expression are major causes of human diversity and may also be important susceptibility factors for complex diseases and traits. Several studies have used linkage or association analysis with microarray-based expression data from lymphoblastoid cell-lines from healthy individuals as the quantitative trait, and have identified putative *cis*- and *trans*-acting genetic variants that regulate the gene expression levels (1–4). So far only a few studies have addressed the relationship between SNPs in regulatory regions of multiple genes and gene expression levels in human diseases in a systematic way. A recent exception is a study, in which the association between SNPs in 200 candidate genes was analyzed against gene expression levels determined using cDNA arrays in breast cancer tumor samples (5). Using novel statistical tools, this study of 50 tumor samples identified both *cis*- and *trans*-acting putative regulatory SNPs (rSNPs).

To use the relative expression levels of two SNP alleles (allelic imbalance (AI)) of a gene in the same sample, instead of the total expression level as the quantitative phenotype is an alternative approach for identifying *cis*-acting rSNPs or haplotypes (6–10). A major advantage of this approach is that the two SNP alleles are measured in the same environment, and serve as internal standards for each other to control other than *cis*-acting genetic factors and environmental factors that may cause differences in the expression levels between samples. AI in expression has proven to be a common phenomenon for human genes. One study detected AI in the expression of 326 out of 602 human genes (54%) by using Affymetrix HuSNP oligonucleotide arrays to study kidney and liver tissues from seven fetuses (11). By analyzing leukocytes from 12 individuals using allele-specific oligonucleotide hybridization arrays (Perlegen Sciences) another study found

*To whom correspondence should be addressed. Tel: +46 18 6112959; Fax: +46 18 553601; Email: Ann-Christine.Syvanen@medsci.uu.se
Present addresses:
Manu Gupta, Cancer Research UK Medical Oncology Laboratory, Barts and the Royal London School of Medicine, Queen Mary College, London, UK.
Sumeer Dhar, Department of Pharmacogenomics and Experimental Therapeutics, UNC School of Pharmacy, Chapel Hill, NC, USA.

AI in 731 out of 1389 informative genes (53%) (12). In addition to allele-specific hybridization on microarrays, a variety of other genotyping methods have been applied to detect AI (6–8,13–15). Due to variation in the sensitivity and specificity of the methods, and the limited number of samples or SNPs included in these studies, the frequency estimates for AI vary largely between studies, from 18 to 60% of the analyzed genes. Imbalanced expression of alleles has also been detected using bioinformatics tools, comparing the allele frequencies of SNPs in expressed sequence tags (ESTs) databases to the allele frequencies in Centre d'Etude du Polymorphisme Humain (CEPH) samples from the Haplotype Mapping project (16). This study estimated that AI occurred for 36% of over 2500 analyzed genes, and AI was experimentally verified for 40 of the genes by sequencing.

In the current study, we established a process for systematic screening for *cis*-acting rSNPs using experimental detection of AI as an initial approach. An approach with similar steps as in our study, but performed in a different order than in our process, was recently described for identifying rSNPs that are associated with osteoarthritis (7). Inspired by a number of studies that have identified putative rSNPs in genes related to cancer (14,17,18) and the response to treatment with anticancer drugs (19–22), we used a panel of cell lines that represent different types of cancers and have been well characterized for their response patterns against anticancer drugs (23) as target cells in our study. For detecting AI in the expression of candidate genes for cancer and anticancer drug response we used our 'in house' developed tag-microarray minisequencing system, which we have previously shown to be accurate and sensitive for quantitative detection of AI (15). Genes that displayed AI were then subjected to bioinformatics-assisted identification of rSNPs that alter the strength of transcription factor binding in their upstream regulatory regions. The putative rSNPs were tested for their protein-binding capacity using electrophoretic mobility shift assays (EMSAs). This process allowed us to zoom in from the 160 originally selected candidate genes to eight genes that might contain rSNPs that affect the transcription levels of the genes.

MATERIALS AND METHODS

Cell lines

A panel of 13 human tumor cell-lines consisting of drug-sensitive parental cell-lines and resistant subtypes was analyzed. Table 1 presents a summary of the cell lines, including their origin, parental cell-lines and the resistant subtypes and the selecting agents used to create resistant subtypes. The cell-line cultures have regularly been monitored and found negative for mycoplasma contamination. The cell lines have been described in detail by Dhar *et al*. (23).

Extraction of DNA, RNA and cDNA synthesis

Genomic DNA was extracted from the 13 cell-lines using the Genelute™ Mammalian Genomic DNA kit (Sigma, St. Louis, MO, USA), and the DNA was stored at –20°C until use. Total RNA was extracted by a standard guanidine isothiocyanate method (TRIZOL® Reagent; Gibco BRL/Invitrogen). The quality of the RNA was verified by running the samples on a 1% agarose gel, and the RNA was quantified by measuring the ultraviolet absorbance at 260 and 280 nm (NanoDrop Technologies). Twenty micrograms of RNA was treated with DNase I to remove genomic DNA using the RNeasy Mini Kit (Qiagen, 74104). Adequate removal of the genomic DNA after DNase I treatment was verified by absence of PCR products from RNA samples using primers for genomic DNA. Five micrograms of purified RNA was reverse transcribed to cDNA using the High-Capacity cDNA archive kit (Applied Biosystems, 4322171).

Table 1. Summary of cell lines analyzed

Parental CL	Resistant CL	Origin	Selecting agent
8226/S	8226/Dox	Myeloma	Doxorubicin
	8226/LR5	Myeloma	Melphalan
CCRF-CEM	CEM/VM-1	T-cell leukemia	Teniposide
NCI-H69	H69AR	Small cell lung cancer	Doxorubicin
U937-GTB	U937/VCR	Histiocytic lymphoma	Vincristine
	GTB/CHS	Histiocytic lymphoma	Cynoguanidine
HELA	–	Cervical cancer	
HTERT	–	Normal epithelial retina	
ACHN	–	Renal adenocarcinoma	

Gene expression profiling

The expression levels of 7400 genes in 13 of the parental and drug-resistant cell-lines had been previously determined using mRNA expression microarrays with cDNA probes (24). Twelve of the cell lines were selected to represent all cancer types in expression profiling on Sentrix® Genome-Wide Expression BeadChips (Illumina Inc., San Diego, CA, USA). Biotinylated cRNA was prepared from 500 ng of RNA, using the TotalPrep™ RNA Labeling Kit (Ambion). The *in vitro* transcription product was purified and labeled with Cy3-labeled streptavidin, followed by overnight hybridization of 1.5 µg of the labeled product to the BeadChips. The following day, the slides were washed and scanned using a Bead Station GX 500 Array Reader (Illumina Inc., San Diego, CA, USA). The image data files were analyzed using the BeadStudio software (Illumina Inc., San Diego, CA, USA), where the 'rank invariant' normalization model was applied, as recommended by the manufacturer. The limit of detection was set at 98% confidence.

PCR

Primers for PCR and minisequencing primers with 20-nucleotide tag sequences were designed using the Primer3 (http://frodo.wi.mit.edu/cgi-bin/primer3/primer3.cgi) and Autoprimer (http://www.autoprimer.com) (Beckman Coulter) softwares. The primers were obtained from Integrated DNA Technologies (IDT Inc., Coralville,

IA, USA). The fragments comprising the SNPs were amplified by PCR from genomic DNA in multiplex reactions with 6–12 amplicons per reaction, using 10 ng of DNA, 0.1 mM dNTPs, 1 U Smart-Taq hot DNA polymerase (Naxo Ltd., Tartu, Estonia), 4 mM $MgCl_2$ and 0.2 μM of primers in a final volume of 30 μl. PCR from cDNA was performed in individual reactions using 1/20 of the cDNA products, 0.1 mM dNTPs, 0.5 U Smart-Taq hot DNA polymerase, 1.5 mM $MgCl_2$ and 0.2 μM of primers in a final volume of 30 μl. The PCR conditions were initial activation of the enzyme at 95°C for 10 min followed by 40 cycles of 95°C for 1 min, 55°C for 30 s and 72°C for 1 min in a Thermal Cycler PTC225 (MJ Research, Watertown, MA, USA). The amplified cDNA fragments were pooled and concentrated to 40 μl using Microcon® YM-30 Centrifugal Filter Devices (Millipore Corporation, Bedford, MA, USA).

Preparation of microarrays

Oligonucleotides that were complementary to the tag sequences on the minisequencing primers were immobilized covalently on CodeLink™ Activated Slides (GE Healthcare, Uppsala, Sweden) by the mediation of a NH_2-group at their 3′-end as described earlier (25). Each oligonucleotide was applied as duplicate spots to the slides at a concentration of 25 μM in 150 mM sodium phosphate pH 8.5 using a ProSys 5510A instrument (Cartesian Technologies Inc., Irvine. CA, USA) equipped with four Stealth Micro Spotting pins (SMP3B, TeleChem International Inc., Sunnyvale, CA, USA). The oligonucleotides were spotted in an 'array-of-arrays' configuration, which facilitates analysis of 80 individual samples in parallel on each microscope slide. In each 'subarray,' a fluorophore-labeled oligonucleotide was included as a control for the immobilization process. After printing, the slides were incubated in a humid chamber for at least 24 h, followed by treatment with ethanolamine. The slides were stored desiccated in the dark until use.

Tag-microarray minisequencing

Excess of PCR primers and dNTPs was removed by treatment of the PCR mixtures with 5 U of Exonuclease I and 1 U of shrimp alkaline phosphatase (USB Corporation, Cleveland, OH, USA). Multiplex cyclic minisequencing primer extension reactions were performed in the presence of 80 tagged primers in both DNA polarities at 10 nM concentration, 0.1 μM Texas Red-ddATP, Tamra-ddCTP, R110-ddGTP and 0.15 μM Cy5-ddUTP (Perkin Elmer Life Sciences, Boston, MA, USA) and 0.065 U of KlenThermase™ DNA polymerase (GeneCraft, Germany), as described earlier (26). Alternatively, reagents from the SNPstream® genotyping system (Beckman Coulter, Fullerton, California, USA) were used for the cyclic minisequencing reaction. A reference oligonucleotide that is complementary to a synthetic template to mimic a four-allelic SNP was added to the minisequencing reaction to monitor the difference in incorporation efficiency of the four nucleotides by the DNA polymerase. The reaction conditions were initial activation of the enzyme at 96°C for 5 min followed by 33 cycles of 95 and 55°C for 20 s each. The extension products were allowed to anneal to the immobilized complementary tag oligonucleotides at 42°C for 1–2 h followed by washing of the slide with 2 × SSC and 0.1% SDS twice for 5 min at 42°C and twice with 0.2 × SSC for 1 min at room temperature. Five replicates of DNA and cDNA from the same cell-line were analyzed in parallel.

Signal detection and data analysis

Fluorescence was measured from the microarrays using a ScanArray® Express instrument (Perkin Elmer Life Sciences, Boston, MA, USA) with the excitation lasers Blue Argon 488 nm (R110 and fluorescin), Green HeNe 543.8 nm (Tamra), Yellow HeNe 594 nm (Texas Red) and Red HeNe 632.8 nm (Cy5) with the laser power set to 88% and the photomultiplier tube gain adjusted to obtain equal signal intensities from the reaction control for all fluorophores. The fluorescence signals were quantified using the QuantArray® analysis 3.1 software (Perkin Elmer Life Sciences, Boston, MA, USA).

The SNP genotypes were assigned using the SNPsnapper software v3.0.0.191 (http://www.bioinfo.helsinki.fi/SNPSnapper/) based on scatter plots with the logarithm of the sum of both fluorescence signals ($S_{Allele1} + S_{Allele2}$) on the vertical axis and the fluorescence signal fractions [$S_{Allele2}/(S_{Allele1} + S_{Allele2})$] on the horizontal axis. The genotypes together with the signal intensities of the incorporated nucleotides were exported to Microsoft® Excel. The data was handled and interpreted using the Microsoft® Excel and Access programs. AI was determined by calculating the fluorescence signal ratio between the two alleles ($S_{Allele1}/S_{Allele2}$) in cDNA and DNA for each heterozygous SNP. The signal ratio from cDNA was divided by the corresponding ratio in DNA to obtain a measure for AI. In this calculation, the mean signal intensity of duplicate spots in one sub-array was considered as one replicate assay, and five replicate assays were performed for each SNP. A two-tailed student's *t*-test was used to calculate the significance of the difference in the allelic ratios for the SNPs in genomic DNA and cDNA (Figure 1).

Analysis of regulatory regions affected by SNPs

The bioinformatics tool Regulatory Analysis of Variation in Enhancers (RAVEN) (M. Andersen, B. Lenhard *et al.*, in preparation) was used for the identification of potential rSNPs in the promoter regions of the genes with imbalanced expression. RAVEN (http://mordor.cgb.ki.se/CONSNP/) combines position weight matrices for transcription factor binding sites (TFBSs) from the manually curated Jaspar database (27,28) with phylogenetic footprinting to increase the likelihood of identifying functional variants. The RAVEN interface enables automatic analysis of SNPs from dbSNP as well as uploading of additional SNPs. Based on the application of position-specific weight matrices, RAVEN gives a score that ranges from 1 to 15 for binding sites of 6–14 nucleotides in length

that contain the two SNP alleles. Putative rSNPs with MAF >0.05 were selected for further genotyping by applying a minimum SNP-caused score difference over 2 between the high- and low-scoring SNP alleles in the TFBS profile and a conservation cut-off above 70% between human and mouse, based on the phylogenetic footprinting.

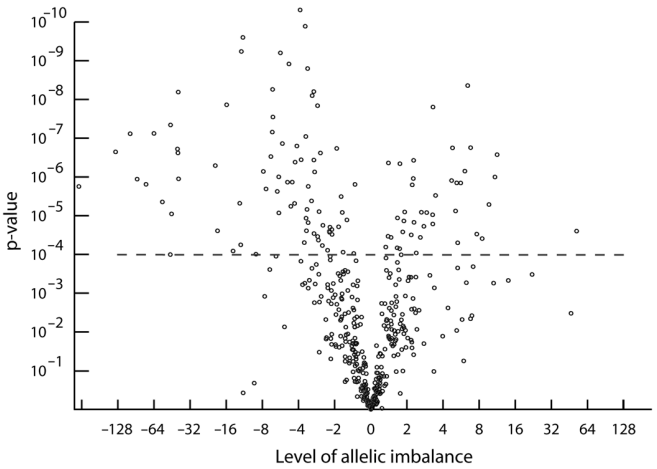

Figure 1. Volcano plot of the AI data from 105 heterozygous cSNPs in 13 cell lines. AI for each SNP was determined by calculating the fluorescence signal ratio between the two alleles ($S_{Allele1}/S_{Allele2}$) in RNA (cDNA) and genomic DNA for each heterozygous SNP. The level of AI obtained by dividing the signal ratio in RNA by the corresponding ratio in DNA is plotted on the horizontal axis. The *P*-value for the difference between allelic ratios in RNA and DNA based on five replicate assays is plotted on the vertical axis. Spots above the horizontal dashed line represent the SNPs showing AI at a *P*-value < 0.0001 that were selected for further analysis.

Electrophoretic mobility shift assays (EMSA)

Complementary double-stranded 5′ biotinylated as well as unmodified 30 bp oligonucleotides, containing the predicted TFBS, were designed for each allele of putative rSNPs (Table 2). The oligonucleotides were obtained from Integrated DNA Technologies (IDT Inc., Coralville, IA, USA). The complementary oligonucleotides were allowed to anneal in 10 mM Tris–HCl, pH 7.5, 50 mM NaCl, 1 mM EDTA to generate double-stranded probes for the EMSA reaction. Twenty femtomoles of the labeled double-stranded probes were incubated with 5 µg of HELA or Jurkat nuclear extracts (Active Motif, Carlsbad, CA, USA) in a freshly made binding buffer containing 12 mM HEPES pH 7.4, 5 mM $MgCl_2$, 60 mM KCl, 1% glycerol, 0.05% NP-40, 50 µg/µl BSA, 1 mM DTT, 0.5 mM EDTA with 50 ng/µl of poly(dI-dC) · poly(dI-dC) (Amersham Biosciences, Piscataway, NJ, USA) and Halt™ Protease Inhibitor Cocktail (Pierce Biotechnology, Rockford, IL, USA) in a final volume of 20 µl. Three reactions were prepared for each double-stranded oligonucleotide (see Figure 2). The mixtures were incubated at room temperature for 20 min, and analyzed using electrophoresis on 6% polyacrylamide gels (Bio-Rad Laboratories, Hercules, CA, USA). The gels were run for 1.5 h at 100 V, followed by transfer to Hybond-N+ nylon membranes (Buckinghamshire, England) in 0.5 × TBE for 1 h at 550 mA, using a Criterion Blotter (Bio-Rad Laboratories, Hercules, CA, USA). The LightShift Chemiluminescent EMSA kit (Pierce Biotechnology, Rockford, IL, USA) was used to visualize the biotinylated oligonucleotide signals on the membranes and a ChemiDoc XRS system (Bio-Rad Laboratories, Hercules, CA, USA). The EMSA experiments were performed twice with reproducible results.

Table 2. Result from validation of the transcription factor binding sites predicted by RAVEN by electrophoretic mobility shift assays

Gene[a]	SNP [b]	EMSA probes (one strand)[c]	Transcription factors[d]	Confirmed by EMSA[e]
APC	rs2439591	GAAATCCATTACACA**G**AATAAGGCAGACA GAAATCCATTACACA**A**AATAAGGCAGACA	AGL3, E4BP4, HLF, SOX17, SQUA	+
BCL2	rs1944423	TTCATAAACTTGGAGA**A**A**T**ATTTATATTGA TTCATAAACTTGGAGAA**C**ATTTATATTGA	Athb-1, HFH-1, HFH-2, HFH-3, HNF-3beta, MEF2, SOX17	–
CCND2	rs3812821	ACCAGAACAACGTCCCTT**G**TGCCCCCCCC ACCAGAACAACGTCCCTT**C**TGCCCCCCCC	SOX17	–
MLH1	rs3172297	ATTTAAGACTA**T**ATGAATC**A**GAATTTTAA ATTTAAGACTA**C**ATGAATCAGAATTTTAA	CF2-II	+
PARP1	rs1317170	CTCGATGGGGT**G**CAT**G**ACATACACAGGATA CTCGATGGGGTGCAT**A**ACATACACAGGATA	CREB, bZIP910	+
SLIT2	rs564041	ACCTAAAATCTCTGCA**A**TAT**T**CTCATTAA ACCTAAAATCTCTGCAATAT**C**CTCATTAA	SOX17	+
XRCC1	rs12608635	CGGCGGCGGGGAGCAGGTG**C**CA**C**GGCCAAA CGGCGGCGGGGAGCAGGTGCCA**T**GGCCAAA	Chop-cEBP, bZIP911	+
YES1	rs7233932	GGAGCGCTCCGATT**G**TGCCCCTCT**T**GCCTT GGAGCGCTCCGATT**C**TGCCCCTCTGCCTT	SOX17, Sox-5	+

[a]Gene symbol according to the HUGO gene nomenclature committee http://www.gene.ucl.ac.uk/nomenclature/
[b]The SNPs rs8073706, rs907187, rs8176077, rs5016499, rs7655084, rs2717701 and rs3810378 in the respective *ABCC3, PARP1, BRCA1, DCTD, SLIT2, TNFRSF12A* and *XRCC1* genes were not confirmed by EMSA.
[c]EMSA probe containing the SNP, the top probe contains the SNP allele that is predicted to give stronger transcription factor binding.
[d]Transcription factors predicted by RAVEN to bind to the probes.
[e]The probes for the SNP alleles giving a stronger signal in EMSA that matched the predictions by RAVEN.

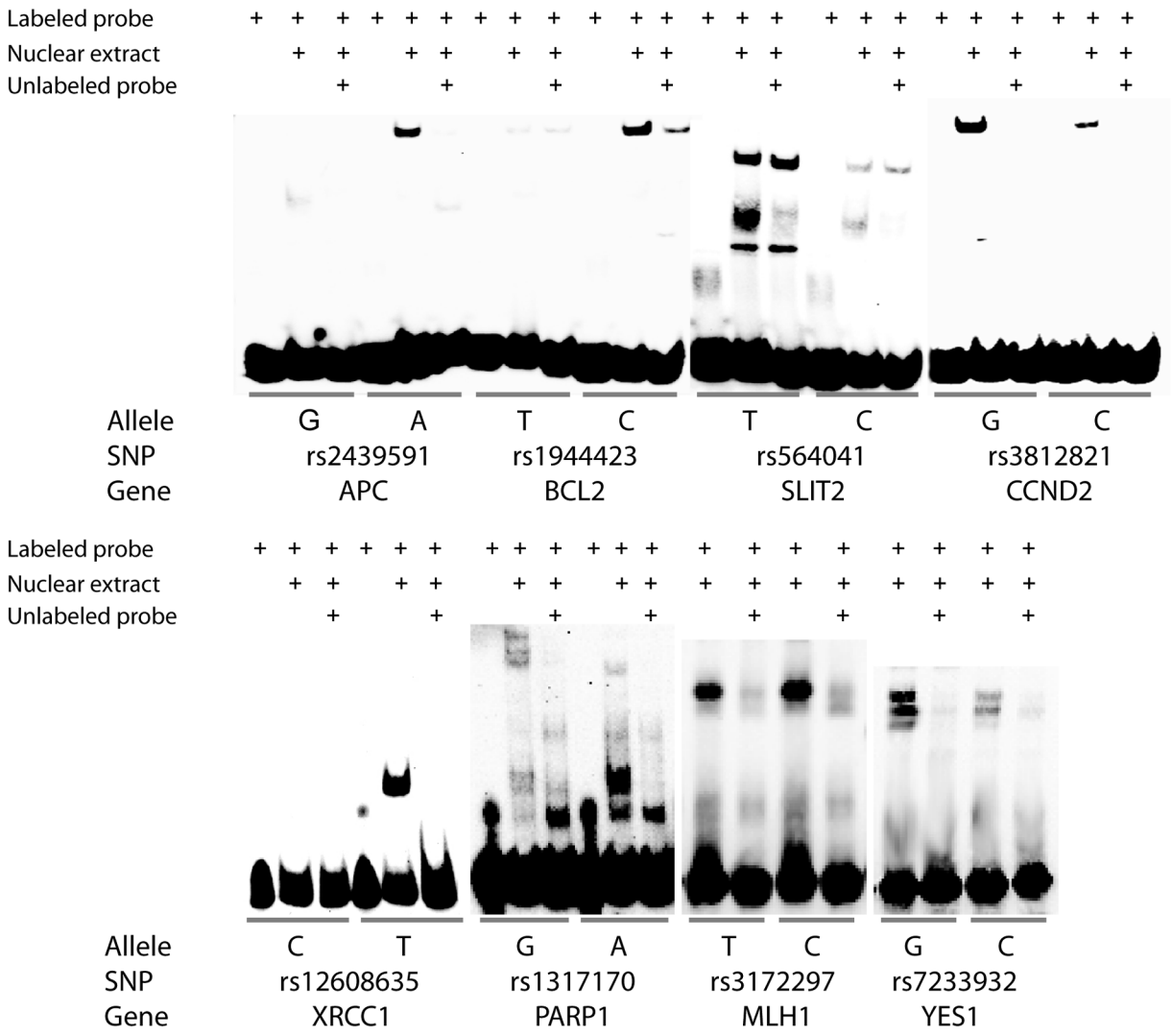

Figure 2. Electrophoretic mobility shift assay images for the SNP alleles of the *APC*, *BCL2*, *SLIT2*, *CCND2*, *XRCC1*, *PARP1*, *MLH1* and *YES1* genes. Three lanes are shown for each SNP allele. From left to right these are: a control reaction with labeled probe only, a reaction containing both labeled probe and nuclear extract and a reaction where an unlabeled probe is added in excess as a competitor, in addition to the labeled probe and nuclear extract. For the *MLH1* and *YES1* genes, the two lanes are shown: a reaction with labeled probe and nuclear extract and a reaction where the unlabeled competitor probe is added. The sequences of the allele-specific EMSA probes are given in Table 2.

RESULTS AND DISCUSSION

Selection of candidate genes and coding SNPs

A panel of 13 human tumor cell-lines that includes drug-sensitive parental cell-lines and their corresponding resistant subtypes was analyzed to detect AI in the expression of candidate genes involved in cancer progression and response to anticancer drugs (Table 1). These cell lines have previously been well characterized for their response patterns against 66 different anticancer drugs (23,24). Initially, we selected a panel of 210 candidate genes for our study. The panel included oncogenes and tumor suppressor genes selected from the literature and genes relevant for the pathways of nine anticancer drugs (irinotecan, 5-fluorouracil, platinum, taxanes, methotrexate, topotecan, gemcitabine, cyclophosphamide and doxorubicin) according to the Pharmacogenetics and Pharmacogenomics knowledge base website (http://pharmacogenetics.wustl.edu/). Based on expression data for 7400 human genes using cDNA microarrays (24) and expression profiling using bead arrays with probes for 46,000 transcripts (Illumina Inc.) (Milani *et al.*, unpublished results), 160 of the 210 genes appeared to be expressed in at least one of the cell lines (see Supplementary Table 1). By searching the dbSNP and Ensembl databases, we identified 237 SNPs with minor allele frequencies above 10% in the coding region of the expressed candidate genes.

Detection of allelic imbalance

Next, we genotyped the cSNPs by multiplex tag-microarray minisequencing in genomic DNA from the cell lines and found that 79 of the candidate genes contained coding SNPs (cSNPs) that were heterozygous in at least one of the cell lines. These heterozygous cSNPs

were then genotyped in five replicate reactions in both genomic DNA and RNA (cDNA) from the relevant cell lines. Genotyping of the RNA samples was successful for 105 cSNPs in 60 genes. For 19 genes, genotyping assays that were successful for genomic DNA failed in cDNA, presumably due to the low expression level of these genes. AI between the expressed alleles was initially observed by aberrant clustering of the genotype data from RNA compared to data from DNA in scatter plots. To obtain a quantitative measure for the observed AI, the fluorescence signal (S) ratio between the two alleles ($S_{Allele1}/S_{Allele2}$) in RNA was divided by the corresponding signal ratio in DNA for each SNP. A student's *t*-test was then used to assess the significance of the difference between allelic ratios in DNA and RNA based on five replicate measurements. The 'volcano plot' in Figure 1 displays the AI data from all 105 cSNPs and 13 cell lines with the magnitude of AI plotted on the horizontal axis and the *P*-values for the differences in signal ratios for the detected AI on the vertical axis. The complete data underlying Figure 1 is provided in Supplementary Table 2. Using a conservative *P*-value of 0.0001 as significance threshold we detected AI in the expression of 41 of the genes (Table 3). Figure 3 summarizes the recovery of genes at the different stages of our screening process.

Despite the conservative approach for defining AI, the relative number of genes that displayed AI in our study (68%) was higher than that previously observed by others based on screening with allele-specific hybridization microarrays (11,12). The reason for this difference could be the high sensitivity of detecting minority alleles using minisequencing primer extension, which we have previously shown to be 1–5%, depending on the sequence context of the SNP (15,29). Alternatively it is possible that cancer-related genes in cancer cells are more frequently expressed in an allele-specific manner than randomly selected genes in lymphoblastoid cell-lines that have been analyzed for AI in other studies.

As can be seen in Supplementary Table 2 the level of AI that we measured in our study varied largely, from 1.3-fold (44% of the minor allele expressed) to over 40-fold (2.4% of the minor allele expressed). For a subset of 15 genes we observed apparent monoallelic expression in at least one of the cell lines, based on an allelic ratio in RNA that was indistinguishable from a homozygous genotype. Extreme AI, or monoallelic expression could, in addition to a strong *cis*-acting regulatory effect, be due to lack of transcription of one allele because of methylation of the promoter region as a consequence of imprinting. In accordance with this notion, we have detected methylation in the CpG islands in the 5′ region of the *ERBB2* gene in the CCRF-CEM and CEM-VM1 cells that showed monoallelic expression in the current study, but not in HELA cells that displayed equal expression of the *ERBB2* alleles (Milani *et al*., unpublished results).

Eleven of the genes that displayed AI in our study contained more than one heterozygous cSNP (Supplementary Table 2). For example, SNPs rs12917 and rs1803965 in exon 3 of *MGMT* yielded 2.3-fold and 1.7-fold AI, respectively, in the CCRF-CEM cell line, and no AI in any of the other cell lines. Three SNPs in different exons of *PARP1* (rs1136410, rs1805404, rs3219061) all yielded 2.3–2.9 fold AI in the HTERT cell line. This data supports that our system yields reproducible results. On the other hand, SNP rs602990 in exon 20 of *VAV2* displayed 12-fold AI, while both alleles of SNP rs509590 in the 3′ UTR of *VAV2* were expressed at equal levels. These apparently discordant results could be caused by differential expression of alternatively spliced transcripts, where the exon containing one of the SNPs has been removed from one of the splice variants. Hence, measurement of AI using SNPs distributed over different exons could be used for relative quantification of alternatively spliced transcripts, as an alternative approach to assays based on detection of exon-specific nucleotides only (29,30). AI could thus be used as a guide to SNPs that regulate alternative splicing, analogously to the process for identifying rSNPs that affect the expression of the entire transcript.

Bioinformatics-assisted identification of SNPs that cause allelic imbalance

Next, we attempted to identify SNPs in the 5′-regulatory regions of the 41 genes that displayed AI. For this purpose, we used the RAVEN application. RAVEN reports evolutionary conserved regions based on the human and mouse genome sequences and scans the sequences for the presence of potential TFBSs that are affected by SNPs. We used RAVEN to scan 3–5 kb of the 5′-regulatory regions of the 41 genes that were found to display AI, and selected about 100 putative rSNPs in the genes based on this analysis. The putative rSNPs identified using RAVEN were subsequently genotyped in all the cell lines (data not shown). The 15 rSNPs that were heterozygous in the same samples as the originally genotyped cSNPs in the corresponding genes were selected for further analysis.

Functional analysis of rSNPs

Fifteen of the rSNPs predicted by RAVEN and that appeared to be in linkage disequilibrium (LD) with the initially genotyped cSNP were analyzed for their capacity to bind transcription factors or other proteins from a nuclear cell extract by EMSAs (31,32). Allelic pairs of eight of these SNPs that are located in the promoter regions of the *APC*, *BCL2*, *SLIT2*, *CCND2*, *XRCC1*, *PARP1*, *MLH1* and *YES1* genes displayed a reproducible signal intensity difference in a product with altered mobility in EMSA. Protein binding to only one of the SNP alleles can be seen for the *APC*, *BCL2* and *XRCC1* genes, while for the *SLIT2*, *CCND2*, *PARP1*, *MLH1* and *YES1* genes a difference in the amount of protein bound is seen (Figure 2). For as many as six of the SNPs, the allele that showed stronger protein binding had been predicted by RAVEN to have a stronger binding affinity for a transcription factor. The transcription factors predicted to bind to the binding sites containing the rSNPs are listed in Table 2. No protein binding or allele-specific differences in binding were detectable using EMSA for the remaining seven SNPs that are located in the A*BCC3*, *PARP1*, *BRCA1*, *DCTD*, *SLIT2*, *TNFRSF12A*

Table 3. Allelic imbalance levels in 13 cell lines

Gene[a]	SNP	Alleles[b]	8226/S	8226/Dox	8226/LR5	CCRF-CEM	CEM/VM-1	NCI-H69	H69AR	U937-GTB	U937/VCR	GTB/CHS	HELA	HTERT	ACHN
ABCB1	rs3842	A/G	–	–	–	−1,4	−3,1	–	–	–	−4,6	–	–	−5,7	–
ABCC3	rs4148416	G/A	–	–	–	–	–	–	–	–	–	–	2,8	–	–
APC	rs2229992	T/C	nd	nd	nd	nd	nd	1.7	nd	nd	nd	nd	–	nd	–
ATF5	rs283525	G/A	−1,7	–	nd	–	–	−7,5	–	nd	nd	nd	nd	–	–
ATF5	rs8647	G/A	–	–	–	nd	nd	−2,9	−2,7	–	–	–	–	–	nd
ATF5	rs8667	T/C	–	–	–	nd	8,5	nd	nd	–	–	–	–	–	nd
BAK1	rs210135	A/T	nd	–	nd	nd	−2,1	nd	nd	nd	nd	nd	–	–	–
BCL2	rs1801018	A/G	–	–	–	−5,8	−3,4	–	–	–	–	–	–	–	–
BCL2	rs4987843	A/G	–	–	–	−3,8	−2,1	nd	nd	–	–	–	–	−2,2	–
BDH1	rs1050119	T/C	–	–	–	−3,4	−4,8	–	–	nd	−2,8	−2,2	–	–	–
BLM	rs1063147	G/A	–	–	–	–	–	–	–	nd	nd	2,4	–	–	–
CBR1	rs20572	G/A	–	–	–	–	−6,8	–	–	–	–	–	–	–	–
CBR1	rs9024	C/T	–	–	–	–	−46,6	–	–	–	–	–	–	–	–
CCND2	rs1049606	T/C	–	–	–	6,1	10,8	nd	3,5	9,7	nd	nd	–	–	–
CCND2	rs3217926	G/A	–	–	nd	–	–	–	–	6,4	nd	nd	–	−11,6	–
CDH6	rs2302904	T/C	–	–	–	–	–	nd	–	–	–	–	–	−2,6	–
ERBB2	rs1058808	C/G	–	–	–	−40,8	−40,2	–	–	–	–	–	nd	–	–
ERBB2	rs1801200	A/G	–	–	–	5,3	7,6	–	–	–	–	–	nd	−3,3	–
ERBB2	rs2230698	A/G	nd	nd	−1,9	–	–	–	–	–	–	–	–	–	–
ERBB2IP	rs36303	G/A	–	–	–	nd	nd	−1,8	nd	–	–	–	–	–	–
ERBB2IP	rs706679	T/C	nd	nd	nd	nd	nd	−5,5	nd	–	–	–	nd	–	–
ERCC2	rs13181	A/C	–	–	–	–	–	–	–	–	–	–	−3,0	–	–
FANCA	rs2239359	T/C	–	–	–	nd	nd	–	–	−133,7	nd	nd	–	–	–
FDXR	rs690514	G/A	–	–	–	–	–	–	–	–	–	–	–	11,3	–
HMGA2	rs8756	T/G	–	–	–	–	–	–	–	–	–	–	–	51,9	–
HMGB1	rs590050	G/A	–	–	–	−2,5	−2,2	−6,6	−4,3	−7,9	−14,0	−9,0	–	−12,0	−6,6
MAP4K2	rs2071313	T/C	–	–	–	–	–	nd	2,2	1,5	nd	nd	–	2,2	–
MAPT	rs1052594	C/G	–	–	–	–	–	nd	−3,1	nd	nd	nd	–	nd	–
MAPT	rs9468	C/T	–	–	–	–	–	nd	nd	nd	−3,3	nd	–	nd	–
MCM7	rs2070215	T/C	–	–	–	nd	−1,9	–	–	nd	nd	nd	–	–	–
MGMT	rs12917	G/A	–	–	–	2,3	nd	–	–	–	–	–	nd	–	–
MGMT	rs1803965	G/A	–	–	–	1,7	nd	–	–	–	–	–	nd	–	–
MLH1	rs1799977	G/A	–	–	–	−40,0	−40,4	–	–	−2,3	nd	nd	–	–	–
MSH6	rs1800935	T/C	–	–	–	–	3,3	–	–	–	–	–	–	–	nd
NF2	rs1008515	T/C	2,3	nd	−5,8	–	–	–	–	nd	nd	3,3	nd	–	–
NME4	rs14293	G/A	1,4	−4,1	−19,7	–	–	nd	nd	–	–	–	−1,6	−1,4	–
PARP1	rs1136410	G/A	–	–	–	–	–	–	–	–	–	–	–	2,6	–
PARP1	rs1805404	G/A	–	–	–	–	–	–	–	–	–	–	–	2,9	–
PARP1	rs3219061	T/C	–	–	–	–	–	–	–	–	–	–	–	2,3	–
PCNA	rs3626	C/G	–	–	–	–	–	–	–	–	–	–	–	−64,1	–
PMS2	rs1059060	G/A	−2,8	–	nd	nd	nd	nd	nd	nd	nd	nd	nd	nd	nd
RET	rs1800858	G/A	–	–	–	–	–	–	–	−3,5	nd	5,1	–	1,7	–
SLIT2	rs7655084	T/G	–	–	–	–	10.6	–	–	–	–	–	–	nd	–
TERT	rs2736098	G/A	nd	–	nd	–	–	−12,4	nd	–	–	–	–	−270,5	–
TK1	rs1065769	C/T	nd	nd	−3,5	–	–	nd	nd	–	–	–	–	nd	–
TK1	rs1071664	T/C	nd	–	nd	–	–	−4,3	nd	–	–	–	–	nd	–
TK1	rs1143696	G/A	1,9	nd	nd	–	–	–	–	–	–	–	–	–	–
TNFRSF12A	rs13209	T/C	nd	4,8	nd	nd	nd	−1,7	nd	–	–	–	–	–	nd
TSHZ1	rs3744908	T/C	1,7	1,3	nd	nd	nd	6,8	nd	–	–	–	1,9	nd	–
TSHZ1	rs3809997	T/C	nd	nd	nd	nd	nd	5,6	nd	–	–	–	nd	nd	–

(continued)

Table 3. Continued

Gene[a]	SNP	Alleles[b]	8226/S	8226/Dox	8226/LR5	CCRF-CEM	CEM/VM-1	NCI-H69	H69AR	U937-GTB	U937/VCR	GTB/CHS	HELA	HTERT	ACHN
UMPS	rs1139538	A/G	5,2	–	nd	–	–	–	–	nd	nd	nd	–	–	–
VAV2	rs509590	T/C	nd	nd	nd	nd	2,6	–	–	nd	nd	nd	–	–	–
VAV2	rs602990	G/A	nd	nd	−12,1	–	–	nd	nd	–	–	–	–	–	–
WT1	rs16754	T/C	–	–	–	−74,6	−54,5	–	–	–	–	–	–	−101,1	–
XPC	rs2470352	T/A	−3,0	−3,6	nd	–	–	–	–	–	−6,0	−3,9	–	–	–
XRCC1	rs25487	A/G	nd	1,8	–	–	–	nd	nd	nd	nd	nd	–	nd	–
YES1	rs1060922	T/C	–	–	–	nd	nd	−18,9	nd	–	–	–	–	–	–

[a]Gene symbol according to the HUGO gene nomenclature committee http://www.gene.ucl.ac.uk/nomenclature/
[b]SNP alleles, the first nucleotide is referred to as Allele 1 and the second as Allele 2.
nd, heterozygous sample where no AI was detected; - (hyphen), homozygous (or failed) samples.
Overexpression of Allele 1 gives positive values, overexpression of Allele 2 gives negative values.

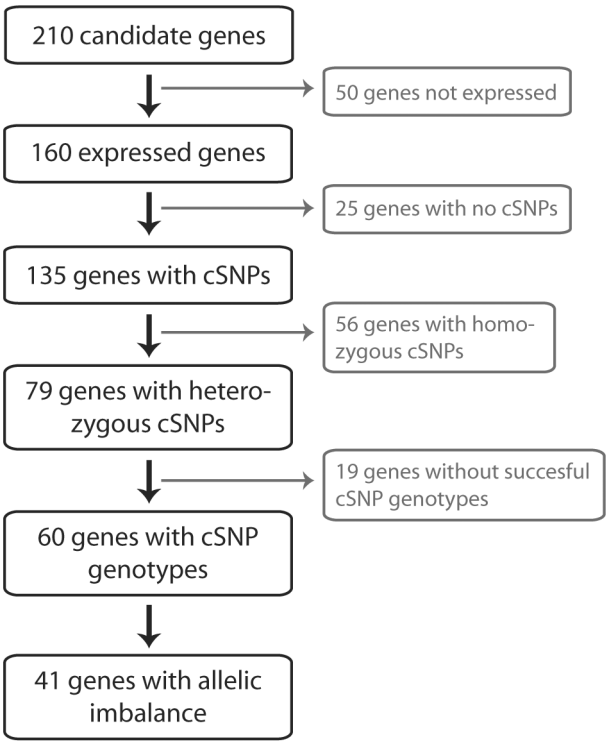

Figure 3. Recovery of genes and SNPs at the different stages of our process for screening for allelic imbalance.

and *XRCC1* genes. RAVEN appears to be an excellent tool for enriching functional SNPs in TFBSs, although fine-tuning of the parameters of RAVEN as well as raising the cut-off for detecting AI could increase its success rate.

The genes in which we identified rSNPs were mainly genes involved in cancer progression. *BCL2* (the B-cell CLL/lymphoma 2 gene), has been reported to be over-expressed in different leukemias and to be involved in leukemogenesis (33) and the expression of *TNFRSF12A* (the tumor necrosis factor receptor superfamily, member 12A gene) is important in cells undergoing apoptosis (34). The Yamaguchi sarcoma viral oncogene homolog 1 (*YES1*) gene has been shown to be differentially expressed in colon cancer cells treated with histone deacetylase inhibitors (35). *ABCC3*, which encodes the multi-drug resistance-associated protein 3, has been reported to be involved in resistance to doxorubicin (36). The identified rSNPs in these genes should be further validated by functional assays. The genes and rSNPs identified in our study are promising candidates for genetic association studies with samples from patient cohorts with relevant types of cancer or drug response patterns.

CONCLUSION

By the screening process presented here, we detected AI in the expression levels of 41 out of 160 candidate genes that were expressed in cancer cells, and applied AI as a guide to putative rSNPs in these genes. Using bioinformatics tools that predict TFBSs, we selected SNPs in the 5′-regulatory

regions of genes for which AI was detected. We identified rSNPs that had a suggestive allele-specific effect, which was shown experimentally by EMSA for eight genes. We conclude that a screening process, such as the one established in our study, that combines allele-specific gene expression analysis with powerful bioinformatics tools offers a shortcut for the detection of potential *cis*-acting regulators of gene expression. The process allows a substantial reduction of the number of candidate rSNPs to be subjected to labor-intensive genetic association or functional studies.

ACKNOWLEDGEMENTS

The study was supported by grants from the Swedish Cancer Foundation and the Swedish Research Council for Science and Technology (to ACS), Swedish Research Council for Medicine (to ACS and MG) and Selander Foundation (to ACS). We thank Raul Figueroa for producing the tag microarrays, and Kristo Käärmann and Mats Jonsson for assistance with data analyses. Funding to pay the Open Access publication charge was provided by Swedish Research Council.

Conflict of interest statement. None declared.

REFERENCES

1. Cheung,V.G., Spielman,R.S., Ewens,K.G., Weber,T.M., Morley,M. and Burdick,J.T. (2005) Mapping determinants of human gene expression by regional and genome-wide association. *Nature*, **437**, 1365–1369.
2. Monks,S.A., Leonardson,A., Zhu,H., Cundiff,P., Pietrusiak,P., Edwards,S., Phillips,J.W., Sachs,A. and Schadt,E.E. (2004) Genetic inheritance of gene expression in human cell lines. *Am. J. Hum. Genet.*, **75**, 1094–1105.
3. Morley,M., Molony,C.M., Weber,T.M., Devlin,J.L., Ewens,K.G., Spielman,R.S. and Cheung,V.G. (2004) Genetic analysis of genome-wide variation in human gene expression. *Nature*, **430**, 743–747.
4. Stranger,B.E., Forrest,M.S., Clark,A.G., Minichiello,M.J., Deutsch,S., Lyle,R., Hunt,S., Kahl,B., Antonarakis,S.E. *et al.* (2005) Genome-wide associations of gene expression variation in humans. *PLoS Genet.*, **1**, e78.
5. Kristensen,V.N., Edvardsen,H., Tsalenko,A., Nordgard,S.H., Sorlie,T., Sharan,R., Vailaya,A., Ben-Dor,A., Lonning,P.E. *et al.* (2006) Genetic variation in putative regulatory loci controlling gene expression in breast cancer. *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 7735–7740.
6. Bray,N.J., Buckland,P.R., Owen,M.J. and O'Donovan,M.C. (2003) Cis-acting variation in the expression of a high proportion of genes in human brain. *Hum. Genet.*, **113**, 149–153.
7. Mahr,S., Burmester,G.R., Hilke,D., Gobel,U., Grutzkau,A., Haupl,T., Hauschild,M., Koczan,D., Krenn,V. *et al.* (2006) Cis- and trans-acting gene regulation is associated with osteoarthritis. *Am. J. Hum. Genet.*, **78**, 793–803.
8. Pastinen,T., Sladek,R., Gurd,S., Sammak,A., Ge,B., Lepage,P., Lavergne,K., Villeneuve,A., Gaudin,T. *et al.* (2003) A survey of genetic and epigenetic variation affecting human gene expression. *Physiol. Genomics*, **16**, 184–193.
9. Pastinen,T., Ge,B., Gurd,S., Gaudin,T., Dore,C., Lemire,M., Lepage,P., Harmsen,E. and Hudson,T.J. (2005) Mapping common regulatory variants to human haplotypes. *Hum. Mol. Genet.*, **14**, 3963–3971.
10. Tao,H., Cox,D.R. and Frazer,K.A. (2006) Allele-specific KRT1 expression is a complex trait. *PLoS Genet.*, **2**, e93.
11. 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**, 1855–1862.
12. 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**, 331–339.
13. Ding,C., Maier,E., Roscher,A.A., Braun,A. and Cantor,C.R. (2004) Simultaneous quantitative and allele-specific expression analysis with real competitive PCR. *BMC Genet.*, **5**, 8.
14. Heighway,J., Bowers,N.L., Smith,S., Betticher,D.C. and Koref,M.F. (2005) The use of allelic expression differences to ascertain functional polymorphisms acting in cis: analysis of MMP1 transcripts in normal lung tissue. *Ann. Hum. Genet.*, **69**, 127–133.
15. Liljedahl,U., Fredriksson,M., Dahlgren,A. and Syvanen,A.C. (2004) Detecting imbalanced expression of SNP alleles by minisequencing on microarrays. *BMC Biotechnol.*, **4**, 24.
16. Ge,B., Gurd,S., Gaudin,T., Dore,C., Lepage,P., Harmsen,E., Hudson,T.J. and Pastinen,T. (2005) Survey of allelic expression using EST mining. *Genome. Res.*, **15**, 1584–1591.
17. Kanamori,Y., Matsushima,M., Minaguchi,T., Kobayashi,K., Sagae,S., Kudo,R., Terakawa,N. and Nakamura,Y. (1999) Correlation between expression of the matrix metalloproteinase-1 gene in ovarian cancers and an insertion/deletion polymorphism in its promoter region. *Cancer Res.*, **59**, 4225–4227.
18. Zhu,Y., Spitz,M.R., Lei,L., Mills,G.B. and Wu,X. (2001) A single nucleotide polymorphism in the matrix metalloproteinase-1 promoter enhances lung cancer susceptibility. *Cancer Res.*, **61**, 7825–7829.
19. Wang,L., Nguyen,T.V., McLaughlin,R.W., Sikkink,L.A., Ramirez-Alvarado,M. and Weinshilboum,R.M. (2005) Human thiopurine S-methyltransferase pharmacogenetics: variant allozyme misfolding and aggresome formation. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 9394–9399.
20. Pullarkat,S.T., Stoehlmacher,J., Ghaderi,V., Xiong,Y.P., Ingles,S.A., Sherrod,A., Warren,R., Tsao-Wei,D., Groshen,S. and Lenz,H.J. (2001) Thymidylate synthase gene polymorphism determines response and toxicity of 5-FU chemotherapy. *Pharmacogenomics J.*, **1**, 65–70.
21. Lee,W., Lockhart,A.C., Kim,R.B. and Rothenberg,M.L. (2005) Cancer pharmacogenomics: powerful tools in cancer chemotherapy and drug development. *Oncologist*, **10**, 104–111.
22. Hirota,T., Ieiri,I., Takane,H., Maegawa,S., Hosokawa,M., Kobayashi,K., Chiba,K., Nanba,E., Oshimura,M. *et al.* (2004) Allelic expression imbalance of the human CYP3A4 gene and individual phenotypic status. *Hum. Mol. Genet.*, **13**, 2959–2969.
23. Dhar,S., Nygren,P., Csoka,K., Botling,J., Nilsson,K. and Larsson,R. (1996) Anti-cancer drug characterisation using a human cell line panel representing defined types of drug resistance. *Br. J. Cancer*, **74**, 888–896.
24. Rickardson,L., Fryknas,M., Dhar,S., Lovborg,H., Gullbo,J., Rydaker,M., Nygren,P., Gustafsson,M.G., Larsson,R. and Isaksson,A. (2005) Identification of molecular mechanisms for cellular drug resistance by combining drug activity and gene expression profiles. *Br. J. Cancer*, **93**, 483–492.
25. Lindroos,K., Liljedahl,U., Raitio,M. and Syvanen,A.C. (2001) Minisequencing on oligonucleotide microarrays: comparison of immobilisation chemistries. *Nucleic Acids Res.*, **29**, E69.
26. Lovmar,L., Fredriksson,M., Liljedahl,U., Sigurdsson,S. and Syvanen,A.C. (2003) Quantitative evaluation by minisequencing and microarrays reveals accurate multiplexed SNP genotyping of whole genome amplified DNA. *Nucleic Acids Res.*, **31**, e129.
27. Sandelin,A., Alkema,W., Engstrom,P., Wasserman,W.W. and Lenhard,B. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, **32**, D91–D94.
28. Vlieghe,D., Sandelin,A., De Bleser,P.J., Vleminckx,K., Wasserman,W.W., van Roy,F. and Lenhard,B. (2006) A new generation of JASPAR, the open-access repository for transcription factor binding site profiles. *Nucleic Acids Res.*, **34**, D95–D97.
29. McCullough,R.M., Cantor,C.R. and Ding,C. (2005) High-throughput alternative splicing quantification by primer extension and matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Nucleic Acids Res.*, **33**, e99.
30. Milani,L., Fredriksson,M. and Syvanen,A.C. (2006) Detection of alternatively spliced transcripts in leukemia cell lines by minisequencing on microarrays. *Clin. Chem.*, **52**, 202–211.

31. Fried,M.G. and Crothers,D.M. (1983) CAP and RNA polymerase interactions with the lac promoter: binding stoichiometry and long range effects. *Nucleic. Acids Res.*, **11**, 141–158.
32. Fried,M. and Crothers,D.M. (1981) Equilibria and kinetics of lac repressor-operator interactions by polyacrylamide gel electrophoresis. *Nucleic. Acids Res.*, **9**, 6505–6525.
33. Wojcik,I., Szybka,M., Golanska,E., Rieske,P., Blonski,J.Z., Robak,T. and Bartkowiak,J. (2005) Abnormalities of the P53, MDM2, BCL2 and BAX genes in acute leukemias. *Neoplasma*, **52**, 318–324.
34. Kokkinakis,D.M., Brickner,A.G., Kirkwood,J.M., Liu,X., Goldwasser,J.E., Kastrama,A., Sander,C., Bocangel,D. and Chada,S. (2006) Mitotic arrest, apoptosis, and sensitization to chemotherapy of melanomas by methionine deprivation stress. *Mol. Cancer Res.*, **4**, 575–589.
35. Hirsch,C.L., Smith-Windsor,E.L. and Bonham,K. (2006) Src family kinase members have a common response to histone deacetylase inhibitors in human colon cancer cells. *Int. J. Cancer*, **118**, 547–554.
36. Liu,Y., Peng,H. and Zhang,J.T. (2005) Expression profiling of ABC transporters in a drug-resistant breast cancer cell line using AmpArray. *Mol. Pharmacol.*, **68**, 430–438.