Parallel gene analysis with allele-specific padlock probes and tag microarrays

Johan Banér, Anders Isaksson, Erik Waldenström¹, Jonas Jarvius, Ulf Landegren* and Mats Nilsson

The Beijer Laboratory, Department of Genetics and Pathology, Rudbeck Laboratory, SE-751 85 Uppsala, Sweden and ¹Department of Medical Sciences, Uppsala University Hospital, SE-751 85 Uppsala, Sweden

Received June 6, 2003; Revised and Accepted July 8, 2003

ABSTRACT

Parallel, highly specific analysis methods are required to take advantage of the extensive information about DNA sequence variation and of expressed sequences. We present a scalable laboratory technique suitable to analyze numerous target sequences in multiplexed assays. Sets of padlock probes were applied to analyze single nucleotide variation directly in total genomic DNA or cDNA for parallel genotyping or gene expression analysis. All reacted probes were then co-amplified and identified by hybridization to a standard tag oligonucleotide array. The technique was illustrated by analyzing normal and pathogenic variation within the Wilson disease-related ATP7B gene, both at the level of DNA and RNA, using allele-specific padlock probes.

INTRODUCTION

Over the last few years millions of single nucleotide polymorphisms (SNPs) have been identified that represent a large proportion of common genetic variation among the humans. It is an important objective of future genetic studies to correlate these molecular genetic variations with human phenotypic variation in health and disease. It remains an analytical challenge, however, to comprehensively access information about the genetic constitution of individual patients. A great number of genotyping methods have been developed to analyze large sets of SNPs. Most methods are based on analysis of individual markers, typically using PCR to amplify the segments harboring the variable nucleotide positions (1). In the interest of greater throughput, a number of approaches have recently been established for parallel analyses of large sets of markers in individual reactions. Amplification by PCR is inherently difficult to perform in a highly multiplexed format, but multiple individual or modestly multiplexed PCR reactions can be pooled and analyzed in parallel on dense resequencing microarrays, revealing thousands of genotypes per hybridization in a serial–parallel process (2,3). A parallel–parallel process was used to genotype and analyze the expression of up to 1000 markers by performing multiplexed target-dependent ligation of pairs of oligonucleotides, followed by amplification and sorting by hybridization to fiber optic bead arrays with DNA tag sequences (4,5).

We have recently presented an approach for parallel gene analysis using large pools of sequence-tagged padlock probes. These were added to individual DNA samples and subjected to gap-fill ligation followed by exonuclease treatment to remove both unreacted and any cross-reactive products that might arise through intermolecular ligation. Next, the circularized probes were molecularly inverted by linearization at a position remote from the site of ligation and amplified by PCR. Finally, the probes were analyzed on commercially available tag sequence arrays (6). Using this strategy individual DNA samples were analyzed for more than 1500 SNPs in parallel, and more recently the technique has been used to analyze sets of 13 000 markers (T. Willis, ParAllele BioScience, personal communication). The approach exploits some useful properties of padlock probes, oligonucleotides that circularize in target sequence-dependent ligation reactions (7). The requirement for recognition of adjacent target sequences by the two ends of these linear probes provide sufficient specificity to analyze SNPs in total genomic DNA without prior target amplification (6,8–10). Moreover, as shown by Hardenbol et al. (6), large numbers of padlock probes can be combined in single reactions since amplification of unreacted and cross-reacted probes can be counteracted by exonuclease treatment.

In this paper we report an alternative protocol for padlock-based analyses of sequence variants in genomic DNA or transcribed sequences. Sets of allele-specific padlock probes were combined in a single ligation reaction, followed by exonuclease removal of unreacted and cross-reacted probes. The remaining circularized probes were amplified and the products identified in an array-of-arrays, prepared using a standard printing robot. Twenty-six allele-specific padlock probes were used to genotype 75 individuals at 100% call rate and accuracy, and with sufficient signal strength for further scale-up. The method was also used to evaluate allele-specific gene expression by comparing genomic DNA samples from liver necropsies with cDNA prepared from the same samples. We discuss differences between the present approach and that used in Hardenbol et al. (6).

*To whom correspondence should be addressed. Tel: +46 18 471 4907; Fax: +46 18 471 4808; Email: ulf.landegren@genpat.uu.se
Table 1. Oligonucleotide sequences used for parallel gene analyses (5’→3’)

<table>
<thead>
<tr>
<th>Padlock probe</th>
<th>5’-End segment</th>
<th>Common sequence</th>
<th>Allele tag</th>
<th>Locus tag</th>
<th>3’-End segment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1216(G/T)</td>
<td>gggatatattaaaaagctgtgc</td>
<td>CTCGACGGTTAGACAGCTGTAgT</td>
<td>(A1/A2)</td>
<td>GTCCTGATCGATGTCTGCACT</td>
<td>ttctctggtctatcgaagc/ta</td>
</tr>
<tr>
<td>1366(C/G)</td>
<td>agatgagctgtacacAgtag</td>
<td>CTCGACCGTACAGCACTGTAtt</td>
<td>(A1/A2)</td>
<td>TCITGCTTGTCTGACCGAT</td>
<td>ggacagccagatgtaaaggg/ta</td>
</tr>
<tr>
<td>1877(C/G)</td>
<td>caatctctctgatcataaa</td>
<td>CTCGACCGTATAGACAGCTGTAtt</td>
<td>(A1/A2)</td>
<td>GCATGATGATCAGTGGCAT</td>
<td>tttctgagtctatcgaagc/ta</td>
</tr>
<tr>
<td>2128(A/G)</td>
<td>gaggcttaagctgagaacgaA</td>
<td>CTCGACCGTACTAGCACTGTAtt</td>
<td>(A1/A2)</td>
<td>TAGACGATGGTACTGCTATG</td>
<td>Cagctggtctatcgaagc/ta</td>
</tr>
<tr>
<td>2495(G/A)</td>
<td>tagatgagatccgcctgacTca</td>
<td>CTCGACCGTACTAGCACTGTAtt</td>
<td>(A1/A2)</td>
<td>GACGGAGATTTGCGCCACAT</td>
<td>actatcctcgacctgcttg/ta</td>
</tr>
<tr>
<td>2855(G/A)</td>
<td>tctgaaaccaccaaactgA</td>
<td>CTCGACCGTACTAGCACTGTAtt</td>
<td>(A1/A2)</td>
<td>ATGTTCGTCAAGATGTCGAT</td>
<td>Aatctggtctatcgaagc/ta</td>
</tr>
<tr>
<td>2930(C/T)</td>
<td>gtaggtgcttgtgaacgaA</td>
<td>CTCGACCGTACTAGCACTGTAtt</td>
<td>(A1/A2)</td>
<td>GCITCCTAAAGACTGATGCTG</td>
<td>gcagctggtctatcgaagc/ta</td>
</tr>
<tr>
<td>2973(G/A)</td>
<td>gttggccgctgcctgacg</td>
<td>CTCGACCGTACTAGCACTGTAtt</td>
<td>(A1/A2)</td>
<td>AACGAGGATGCAAGCTGTA</td>
<td>gggacacagctgcttg/ta</td>
</tr>
<tr>
<td>3207(C/A)</td>
<td>tgttcactgctggcctccgca</td>
<td>CTCGACCGTACTAGCACTGTAtt</td>
<td>(A1/A2)</td>
<td>AGCGCGCAGTAGAGCTGTA</td>
<td>cctcagctggtctatcgaagc/ta</td>
</tr>
<tr>
<td>3317(A/T)</td>
<td>cttgctcaaccatcaatcaagec</td>
<td>CTCGACCGTACTAGCACTGTAtt</td>
<td>(A1/A2)</td>
<td>TACATGATGATCAGTGGCAT</td>
<td>gggatcctggtctatcgaagc/ta</td>
</tr>
<tr>
<td>3419(T/C)</td>
<td>ctgcactcatttcaaaagggc</td>
<td>CTCGACCGTACTAGCACTGTAtt</td>
<td>(A1/A2)</td>
<td>ATATCGGCTACACAGTGTTG</td>
<td>ctatcctcgacctgcttg/ta</td>
</tr>
<tr>
<td>3547(C/T)</td>
<td>ttcaggtttcctgacgac</td>
<td>CTCGACCGTACTAGCACTGTAtt</td>
<td>(A1/A2)</td>
<td>AGCGCGGATGCAAGCTGTA</td>
<td>gggatcctggtctatcgaagc/ta</td>
</tr>
<tr>
<td>3955(T/C)</td>
<td>gacgctctctgtaagggA</td>
<td>CTCGACCGTACTAGCACTGTAtt</td>
<td>(A1/A2)</td>
<td>TCCGCTGTGATGATGATGGA</td>
<td>gggatcctggtctatcgaagc/ta</td>
</tr>
</tbody>
</table>

Primers and hybridization probes

Common primer | TCTACGGTCTAGCCTGAG
Allele 1 specific primer and A1 | CGGAGATGCTGATCAATGCTG
Allele 2 specific primer and A2 | AGCGCGGATGCAAGCTGTA
Hybridization control oligonucleotide | F/TA – ttctctggtctatcgaagc/ta

Padlock probes were designed to analyze sequence variations within the human ATP7B gene (GenBank accession no. NT_024524). Names of the probe pairs are given as the base pair numbers for the variable nucleotide positions, according to Petrukhin et al. (19), followed by the nucleotide variation in parentheses; allele 1 first, allele 2 second. For each padlock probe pair, the target-complementary sequences are in lower case and bold nucleotides indicate the segments where the common primer hybridizes and normal upper case indicates the locus tag sequences. Tag codes were selected from the GeneFlex™ Tag Array sequence collection (Affymetrix, Santa Clara, CA). The tag sequences specific for alleles 1 and 2 (A1 and A2) at any locus are located between the common and locus tag sequences. The A1 and A2 PCR primers correspond to alleles 1 and 2, respectively. Italic upper case indicates the segments where the common primer hybridizes and normal upper case indicates the locus tag sequences. Tag codes were selected from the GeneFlex™ Tag Array sequence collection (Affymetrix, Santa Clara, CA). The tag sequences specific for alleles 1 and 2 (A1 and A2) at any locus are located between the common and locus tag sequences. The A1 and A2 PCR primers correspond to alleles 1 and 2, respectively, with a hybridization control oligonucleotide, were labeled with FITC or TAMRA. Spotting oligonucleotides for each locus and for the hybridization control are complementary to individual locus tag sequences and to the hybridization control oligonucleotide, respectively, with an additional 3'-TC)-T-NH2 tail.

MATERIALS AND METHODS

DNA and RNA samples

DNA samples were from patients with Wilson’s disease, cared for at Uppsala University hospital (11), and from anonymous blood donors. Liver necropsies were kindly provided by Liver necropsies were kindly provided by Dr Fredrik Pontén. DNA and RNA were extracted using a Trizol protocol (Gibco BRL). Second strand cDNA was prepared by incubating 0.1 μg/μl total RNA in 50 μl of 50 mM Tris–HCl pH 8.3, 3 mM MgCl2, 75 mM KCl, 10 mM DTT, 1 μM random hexamer primers (Gibco BRL), 0.8 U/μl HRP ribonuclease inhibitor (Amersham Bioscience) and 3 U/μl M-MLV reverse transcriptase (USB) for 30 s at 50°C, followed by 1 h at 37°C and finally 80°C for 10 min. An aliquot of 5 μl of second strand reaction mix, containing 50 mM Tris–HCl pH 7.5, 10 mM MgCl2, 1 mM DTT, 0.25 μg BSA, 5 μM random primers and 5 U Klenow fragment (Amersham Bioscience) was added to 20 μl of the first strand reaction and incubated at 37°C for 1 h, followed by 65°C for 15 min. The reactions were finally diluted 10-fold in water.

Oligonucleotides

All padlock probes were evaluated using Oligo 6.6 software (Molecular Biology Insights Inc.) to avoid secondary structure that might interfere with probe function. The 104 at long probes were synthesized, chemically 5’-phosphorylated and gel purified by Thermo Electron (Ulm, Germany). Tag oligonucleotide sequences were selected from the GeneFlex™ Tag Array collection (Affymetrix, Santa Clara, CA) that contains sequence information for 2000 oligonucleotides with minimal tendency for cross-hybridization. Padlock probe sequences and amplification primers are presented in Table 1.

Ligation, exonuclease treatment and amplification

Genomic DNA was reduced in size by digestion using Dral and SspI (New England Biolabs) for 2 h in the recommended buffers. Aliquots of 10 μl of ligation reactions, containing 50 ng digested DNA or 2 μl of cDNA in 20 mM Tris–HCl pH 9.0, 100 mM KCl, 10 mM MgCl2, 1 mM EDTA, 1 mM DTT, 0.1% Triton X-100, 400 μM Tth ligase or Ampligase (Epigenec) and 400 pM each padlock probe were placed in a thermal cycler at 95°C for 5 min, then cycled 20 times between 95°C for 2 min and 55°C for 20 min, followed by 95°C for 2 min. Homozygous genotypes that were not present among our samples were instead represented by synthetic 40mer oligonucleotide targets added at 45 zmol to separate ligation reactions. After ligation, 10 μl of exonuclease mix was added to the reactions for a final concentration of 67 mM Tris–HCl pH 9.0, 50 mM KCl, 6.7 mM MgCl2, 0.05 μg/μl BSA, 0.35 U/μl exonuclease I (New England Biolabs) and 0.35 U/μl exonuclease III (Amersham Biosciences). Samples were incubated at 37°C for 2 h, followed by 95°C for 10 min. For PCR amplification, 6 μl of exonuclease reaction was combined with 24 μl of PCR mix, bringing the concentration to 30 mM Tris–HCl pH 8.8, 18 mM KCl, 3.0 mM MgCl2, 0.08% Triton X-100, 200 μM dNTP, 400 nM each of the three PCR primers and 0.1 U/μl HotStart Pfu polymerase (Stratagene). Reactions were placed in a thermal cycler at 95°C for 2 min, and then cycled 26 times between 98°C for 30 s and 55°C for 2 s. The competitor experiment (Fig. 3) was
performed by mixing 45 zmol of circularized probes specific for the 1216 locus with the indicated molar excess of circularized probes for the remaining 12 loci. The probes were circularized using an excess of synthetic target oligonucleotides and then diluted to the desired concentrations. The probes were amplified and hybridized as described.

**Oligonucleotide arrays**

Amino-modified tag oligonucleotides were dissolved at 10 μM concentration in 150 mM sodium carbonate buffer pH 9 with 0.08% SDS and printed in triplicate on 3D-Link slides (Motorola) as 12 subarrays using a four pin GMS 417 Microarrayer (Genetic Microsystems Inc.). Slides were incubated in a 75% humidity chamber at room temperature for 2 h and then transferred to a phosphate-buffered saline bath containing 6.9 mM NaBH₄ and 23% ethanol for 5 min, washed in dH₂O, dried in a stream of pressurized air and finally stored under denaturing conditions (7).

Therefore, in the following experiments all nearby ends of target DNA sequences under denaturing conditions (7). The probes were amplified using target-complementary arms of these probes extended same samples. Four loci were omitted from the liver analysis since the target-complementary arms of these probes extended over exon–intron junctions. A control reaction in which reverse transcriptase was omitted gave no signals, demonstrating that no contaminating genomic DNA was detected (data not shown). We observed no evidence of bias of allelic expression or splicing in the liver samples, which would have resulted in altered allele ratios at the two heterozygous loci (data not shown). We observed no evidence of bias of allelic expression or splicing in the liver samples, which would have resulted in altered allele ratios at the two heterozygous loci (data not shown). We observed no evidence of bias of allelic expression or splicing in the liver samples, which would have resulted in altered allele ratios at the two heterozygous loci (data not shown). We observed no evidence of bias of allelic expression or splicing in the liver samples, which would have resulted in altered allele ratios at the two heterozygous loci (data not shown).

Pairs of padlock probes were designed to recognize the two alleles at individual loci. The segment between the target-complementary ends of the probes included sequence elements required for amplification, along with a unique locus-specific tag sequence. Pools of all probes were added to individual samples along with a thermostable DNA ligase. After ligation the reactions were treated with a mix of exonucleases to remove probes that had not been circularized or that might have undergone intermolecular ligation (Fig. 1A). Next, reacted probes were amplified by PCR with one common primer and two allele-specific fluorescence-labeled ones that allowed amplification products representing different alleles to be distinguished. Finally, the products were sorted by hybridization on a tag sequence microarray (Fig. 1B). Genotypes at individual loci were revealed by analyzing the allele-specific fluorescence signals at the corresponding microarray positions.

A subset of 29 samples were also genotyped by Sanger sequencing (11) or minisequencing, and results for all these 377 genotypes were in complete agreement with those obtained using the parallel padlock probe approach. The distinction between alleles at individual loci was not detectably influenced by the presence of probes for the other 12 loci, attesting to a lack of interference between probes (data not shown).

The same probe set was also used to analyze cDNA preparations from two liver necropsies. Thereby, expression data of transcribed ATP7B alleles could be related to genotyping data recorded using a DNA preparation from the same samples. Four loci were omitted from the liver analysis since the target-complementary arms of these probes extended over exon–intron junctions. A control reaction in which reverse transcriptase was omitted gave no signals, demonstrating that no contaminating genomic DNA was detected (data not shown). We observed no evidence of bias of allelic expression or splicing in the liver samples, which would have resulted in altered allele ratios at the two heterozygous loci (Fig. 2B). The results from the remaining homozygous loci were in agreement with their corresponding genomic DNA samples (data not shown).

To investigate how the signal strength from individual features of the arrays depended on the level of multiplexing, a dilution series was prepared where a constant amount of circularized tracer probes for one locus was mixed with increasing amounts of circularized competitor probes specific for the remaining 12 loci. Signals from the three possible
genotypes of the tracer probes remained clearly distinct in the presence of a 3000-fold molar excess of competitor probes (Fig. 3), indicating that a far greater multiplicity of probes could be used in the present experimental set-up.

DISCUSSION

We present a strategy for parallel analyses of large sets of DNA or RNA sequence variants using padlock probes to detect and distinguish gene sequences in total genomic DNA or cDNA. The approach represents a scalable alternative to the highly multiplexed method we recently described (6), differing in a few aspects as described below.

In the method used by Hardenbol et al. (6), circularization of probes was affected in combined polymerization and ligation reactions, i.e. separate gap fill reactions were performed for each nucleotide triphosphate (14). These procedures require only one probe per polymorphism, eliminating probe pair imbalances and reducing probe cost, but four reactions must be performed per individual typed. In the present strategy a single ligation reaction was used to determine all genotypes in a sample which ensured that identical reaction conditions were used, but twice as many probes had to be prepared.

The combination of allele-specific incorporation of nucleotides followed by ligation in the previous study provides two levels of allele distinction. First, the correct nucleotide must be incorporated, and any misincorporated nucleotide would still hamper the ligation reaction. However, ligation reactions by themselves are known to distinguish single nucleotide variants by factors of thousands (15,16). Accordingly, the allele-specific padlock probe ligation reactions used in the present study also yield robust discrimination of alleles.

We confirm in this study that topologically linked padlock probes are poor substrates for polymerization reactions, but PCR is less affected than the previously studied isothermal rolling circle replication reaction (13). Here we have resolved the topological inhibition by fragmenting the genomic DNA.

Figure 1. Parallel padlock probe analysis of single nucleotide variation. (A) Padlock probes included two target-complementary sequences at both the 5’ and 3’ ends (grey) and the allele-specific nucleotide was positioned at the 3’-end. The target-specific ends were connected by the following sequence elements, 20 nt each, in 5’→3’ order: one sequence common for all probes used for amplification (black), either of two allele-specific tags (green or red) and, finally, a tag sequence unique for each locus (blue). Only probe arms correctly hybridized to a target sequence (black line) are efficiently joined by a ligase. After ligation, unreacted or dimerized probes are degraded by exonuclease treatment, preserving circularized ones. (B) Circularized probes at any locus serve as templates in a PCR with one primer specific for the common sequence (black) and two allele-specific primers, labeled with either FITC for allele 1 (green primer, marked F) or TAMRA for allele 2 (red primer, marked T). PCR products are then denatured and applied to a DNA microarray with complementary tag sequences, guiding the labeled strands to their corresponding positions. An overlaid scanning image is shown from an analysis of one individual (red color, TAMRA; green color, FITC; a combination of TAMRA and FITC appears as yellow). The spot position key identifies the investigated loci. Numbers represent locus nomenclature according to Petrukhin et al. (19).

Figure 2. (Opposite) Results from parallel gene analysis. (A) Seventy-five individuals were genotyped at 13 loci. Signal intensities from FITC- and TAMRA-labeled PCR products were plotted on logarithmic x- and y-axes, respectively. Locus identities are shown in the lower left corners. Individuals with confirmed genotypes that were homozygous for allele 1 at any locus are identified by green squares, homozygotes for allele 2 by red squares and heterozygotes by yellow squares. Blue diamonds represent individuals that have not been investigated by other genotyping techniques. Oligonucleotide targets representing all homozygous genotypes were also assayed, and included in the graphs if any of the homozygous genotypes were not represented in our set of samples (triangles). (B) Parallel genotyping of cDNA samples (open squares and triangles) from two liver necropsies and the corresponding genomic samples (filled squares and triangles). The results from two loci are plotted with the results from 75 genomic samples (A) shown as shaded diamonds to indicate the genotype clusters.
before padlock probe hybridization and ligation, so that circularized probes can slip off the target strands during the denaturation step that precedes the polymerization reaction (7). An alternative approach was taken in Hardenbol et al. (6), where intact genomic DNA was used but probes were released from their target sequences by cleaving the probes at uracil
residues after circularization, molecularly inverting the probes before amplification (6). In the present study the ligation reactions were repeated cyclically to increase circularization yield (7), something that may be less efficient using the uracil approach.

In the present study, amplification products from each individual were hybridized to individual tag sequence subarrays on a microscope slide prepared in-house. This array-of-arrays design permits analysis of many samples on each slide (12,17). Intermediary scalable levels of multiplexing can therefore be performed at moderate expense compared to standard procedures by a commercial vendor. We estimate that ligation efficiency of the different full-length probes in this study ranged from 10 to 50%. The performance of the assay can be expected to be further improved with higher probe quality. It will be of importance to develop improved means to design and also to produce large sets of probes at high purity. This work is currently in progress in our group (J. Stenberg et al., unpublished results).

In conclusion, pools of padlock probes, each molecularly tagged to reflect its locus and allele specificity, are promising tools to interrogate minute DNA or RNA samples in extensive association studies or to investigate the genetic basis of responses to drug treatment and for highly precise transcript analyses. A general set of amplification primers and arrays can be used for different sets of padlock probes. The reaction kinetics of these unimolecular dual recognition probes allows lower probe concentrations to be used, reducing risks of amplification artifacts and providing for low cost, high throughput analyses of large sets of gene sequences.

ACKNOWLEDGEMENTS

Drs F. Barany and F. Pontén kindly provided Th ligase and liver samples, respectively. The work was supported by the Borgström, Beijer and Wallenberg Foundations and by Polysaccharide Research AB (Uppsala), the Research Council of Sweden for Natural Science and Research Council of Sweden for Medicine, the Swedish Cancer Fund and by a long term EMBO fellowship to A.I.

REFERENCES


