Journal of Translational Medicine BioMed Central

Methodology

Open Access

A strategy for detection of known and unknown SNP using a minimum number of oligonucleotides applicable in the clinical settings

Ena Wang[1], Sharon Adams[1], Yingdong Zhao[2], Monica Panelli[1], Richard Simon[2], Harvey Klein[1] and Francesco M Marincola*[1]

Address: [1]Immunogenetics Section, Department of Transfusion Medicine, Clinical Center, National Institutes of Health, Bethesda, MD USA and [2]Biometric Research Branch, Division of Cancer Treatment and Diagnosis, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA

Email: Ena Wang - EWang@mail.cc.nih.gov; Sharon Adams - SAdams@mail.cc.nih.gov; Yingdong Zhao - zhaoy@ctep.nci.nih.gov; Monica Panelli - MPanelli@mail.cc.nih.gov; Richard Simon - rsimon@mail.nih.gov; Harvey Klein - HKlein@mail.cc.nih.gov; Francesco M Marincola* - FMarincola@cc.nih.gov

* Corresponding author

Published: 20 August 2003

Journal of Translational Medicine 2003, **1**:4

Received: 02 July 2003
Accepted: 20 August 2003

This article is available from: http://www.translational-medicine.com/content/1/1/4

Abstract

Detection of unknown single nucleotide polymorphism (SNP) relies on large scale sequencing expeditions of genomic fragments or complex high-throughput chip technology. We describe a simplified strategy for fluorimetric detection of known and unknown SNP by proportional hybridization to oligonucleotide arrays based on optimization of the established principle of signal loss or gain that requires a drastically reduced number of matched or mismatched probes. The array consists of two sets of 18-mer oligonucleotide probes. One set includes overlapping oligos with 4-nucleotide tiling representing an arbitrarily selected "consensus" sequence (consensus-oligos), the other includes oligos specific for known SNP within the same genomic region (variant-oligos). Fluorescence-labeled DNA amplified from a homozygous source identical to the consensus represents the reference target and is co-hybridized with a differentially-labeled test sample. Lack of hybridization of the test sample to consensus- with simultaneous hybridization to variant-oligos designates a known allele. Lack of hybridization to consensus- and variant-oligos indicates a new allele. Detection of unknown variants in heterozygous samples depends upon fluorimetric analysis of signal intensity based on the principle that homozygous samples generate twice the amount of signal. This method can identify unknown SNP in heterozygous conditions with a sensitivity of 82% and specificity of 90%. This strategy should dramatically increase the efficiency of SNP detection throughout the human genome and will decrease the cost and complexity of applying genomic wide analysis in the context of clinical trials.

Background

The human genome project provides the first reference sequence of all human chromosomes with the remaining challenge of characterizing the frequency of deviations from this reference among individuals [1]. It is estimated that 1.42 million SNP are distributed throughout the human genome and about 60,000 SNP fall within exons [2]. Detection of SNP due to genetic variation in a given population (polymorphisms) or epigenetic changes throughout life (mutations) is important since it often has

functional implications. In fact, 25 % of the known non-synonymous SNP could affect the function of the correspondent gene product [3–6]. Yet, it is still unclear whether the prevalence of common diseases can be truly attributed, at least in part, to SNP because of the incomplete information available about SNP prevalence throughout the genome. The completion of the human genome project could not provide comprehensive knowledge about sequence variations because sequences are based on information derived from randomly chosen individuals [1,2] and there are only few examples of systematic searches for genetic variants within a specific genomic region [7]. However, in the context of clinical research a large number of individuals may need to be screened when investigating associations between genetic variation and disease susceptibility. In such an endeavor, a tool capable of efficiently identifying known and flagging unknown SNP could dramatically increase the efficiency of the study of human pathology through direct application of genome-derived information [8].

Known SNP can be readily detected using oligo-array-based techniques [9–12] or comparable high-throughput systems [13]-[14,15]. Detection of unknown SNP, however, is not as readily achievable because most current methodologies are based on the utilization of probes encompassing only known variant sequences [16,17]. Thus, identification of unknown SNP has relied on high-throughput sequencing which is burdened by high cost and demanding requirements for sample preparation. To improve the efficiency of SNP detection, high-density oligonucleotide arrays have been proposed that cover all possible sequence permutations of the genomic region of interest [7,9,18,19]. These arrays are characterized by extreme accuracy not only for detecting but also in providing definitive sequence information about SNP [9]. However, for each genomic region a complex SNP array needs to be assembled as for the 4L (length of nucleotide) oligomer probes [9]. These arrays are composed of oligomer probes that query sequential positions in the genome spanning the length of the probe each one overlapping the previous one of one base. For each position a set of four oligos is prepared identical except at a single, generally central, position systematically substituted with each of the four nucleotides. Thus for a given genomic region a number of oligos equal to the number of bp investigated times 4 is spotted to the array. To query a 16,569-base pairs (bp) sequence 66,276 probes were necessary [9]. Similarly others have investigated BRCA1 and ATM genes using 96,600 and >90,000 oligonucleotides for genomic regions encompassing 3,450 and 9,170 bp respectively [10,19]. Although this approach could potentially cover the full genome, it might not be justified for genomic areas with no polymorphism [9]. In addition, preparation of these arrays would be disproportionate for genomic areas with very low density of SNP. In those cases it would be preferable to obtain more information about the location of highly polymorphic sites prior to the design of 4L tiled arrays or other comprehensive high-throughput systems. Finally, this approach would not be justifiable in situation where SNP occur extremely rarely in a given population. In those cases a tool that could identify the rare individuals carrying SNP could indicate few instances where routine sequencing of a limited genomic region could be more appropriate than the preparation of complex high-density arrays. Thus, a simplified screening tool that could discriminate conserved from polymorphic genomic regions or identify rare individuals carrying unusual SNP could dramatically restrict the use of high-throughput sequencing or guide the production of high-density 4L tiled arrays.

Results

We describe here a strategy that utilizes the well-established principle of loss or gain of hybridization signal [9,19]. for the screening of genomic regions that requires ~250 overlapping oligos to cover a 1 kb consensus sequence (consensus oligos) or 4,142 rather than 66,276 oligos to cover a 16.569 bp genomic region as for the previous example [9]. Thus, the proposed strategy should be considered a screening tool applicable for the investigation of unexplored areas of the human genome prior to extensive sequencing expeditions or the construction of high-density arrays. In addition, in situations where allelic variation is already revealed, the array can be complemented by a number of oligos equal to the number of known SNP within that region. This number, therefore, is proportional to the degree of pre-documented polymorphism of a given genomic fragment. This strategy proposes a fluorimetric detection of SNP by proportional hybridization to oligonucleotide arrays. The reference sample, from a homozygous cell line identical to the consensus (*a,a*), and test sample are amplified by PCR followed by *in vitro* transcription to generate single stranded RNA. Array data generated from hybridization of fluorescence-labeled reference (i.e. Cy3, green) and test (i.e. Cy5, red) cDNA sample to consensus and additional oligos, representing known SNP (variant oligos), is compared and represented as natural log of the fluorescence intensity ratio ($_{Log}$Ratio). In diploid organisms, four type of combinations can occur: I) Homozygosity identical to the consensus (*a,a*); II) Homozygosity different from the consensus (*b,b*); III) heterozygosity containing one allele identical and one different from the consensus (*a,b*); IV) heterozygosity with both alleles different from the consensus (*b,c*). Although this conceptually applies to whole genes, in loci containing more than one polymorphic site, this distinction applies to regions investigated by individual oligos; while, for the whole gene various combinations can simultaneously occur. Thus, in this paper we will

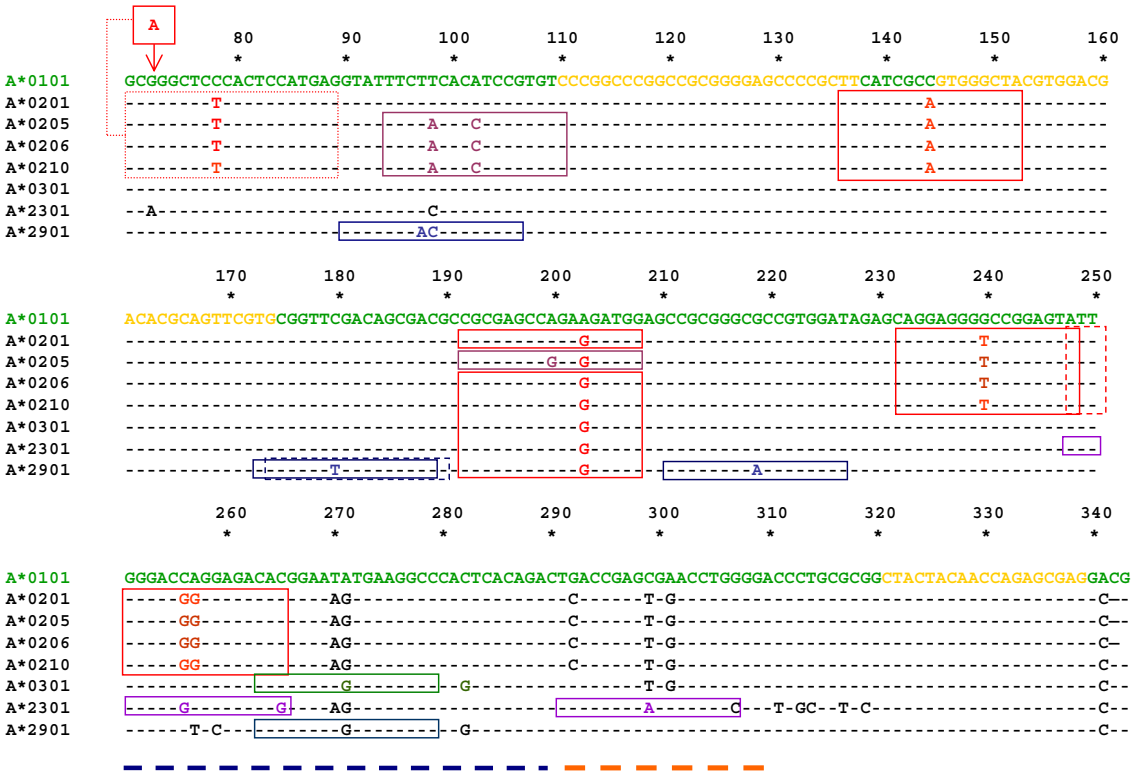


Figure 1

HLA-A Locus, Exon II nucleotide sequence alignment of alleles relevant to the discussion of this manuscript. The consensus sequence is shown in green in areas containing polymorphic sites and in yellow at conserved areas (*k* oligos). Genomic regions covered by variant oligos are indicated by boxes with colored borders (red: HLA-A*02; violet: sub-types of HLA-A*02 different from A*0201; green: HLA-A*0301, purple: HLA-A*2301; blue: HLA-A*29) corresponding to the sequences shown in Figure 2. Dashed borders indicate overlapping oligos. Dotted borders indicate variant oligos 1-SP-A-02a and b that contain an additional G > A switch in a peripheral region to test the influence of this mismatch in various conditions. In black are SNP in areas for which no variant alleles were designed ("unknown SNP"). The dashed line represents the areas where most "unknown SNP" are present for HLA-A*02 and A*29 (blue) or only for HLA-A*02 (orange).

refer to the various combinations by specifying *ad hoc* whether we are referring to the whole gene or a specific region of the same gene.

To illustrate this strategy, we selected a region of the human genome spanning exon 2 of the Human Leukocyte Antigen (HLA)-A locus, which contains multiple polymorphic sites (Figure 1). This gave us the opportunity to identify different homo- heterozygous combinations using genomic DNA from previously classified HLA-typed Epstein-Barr virus-transformed B lymphoblastoid (EBV) cell lines obtained from the European Collection of Animal Cell Cultures. The HLA type of all cell lines was verified using sequence-based typing methods [20]. According to conventional display, the HLA-A*0101 sequence was chosen as the consensus http://www.anthonynolan.com/HIG/data.html. A set of 75 18-mer oligos with 4-nucleotide tiling were selected to cover the HLA-A*0101 allele from position 71 to 343. In addition, several variant oligos relevant to the HLA phenotype of the EBV lines tested were designed (Figure 2) with the SNP in the centermost position to enhance the power of discrimination caused by single nucleotide mismatch. Other known SNP close to the 3' region were not covered by variant oligos to test the capability to detect unknown variants ("unknown SNP" displayed in black in Figure 1).

Name of oligo	Position	5' Sequence	SNV		SNV	3' sequence
1-SP-A-02 a*	70-87	70-CC	G->A	GGCTC	C->T	CACTCCATG-87
2-SP-A-02 b*	71-88	71-C	G->A	GGCTC	C->T	CACTCCATGA-88
3-SP-A-2901	90-106	90-GTATTTC	TT->AC			CACATCCG-106
4-SP-A-0205,0206	93-111	93-TTTCT	T->A	CAC	A->C	TCCGTGTC-111
5-SP-A-02,30	136-153	136-TTCATCGC	C->A			GTGGGCTAC -153
6-SP-A-2901 b	172-189	172-GTGCGGTT	C->T			GACAGCGAC-189
7-SP-A-2901 a	173-190	173-TGCGGTT	C->T			GACAGCGACG-190
8-SP-A-0205	191-208	191-CcGCGAGCC	A->G	GA	A->G	ATGG-208
9-SP-A-02,A03,A23, A29	191-206	191-CCGCGAGCCAGA	A->G			GATGG-208
10-SP-A-29	211-229	211-CCGCGGGC	G->A			CCGTGGATA-229
11-SP-A-02 b	232-249	232-CAGGAGGG	G->T			CCGGAGTAT-249
12-SP-A-23 a	248-265	248-ATTGGGAC	C->G			AGGAGACAG-265
13-SP-A-02	248-265	248-ATTGGGAC	CA->GG			GGAGACAC-265
14-SP-A-03, A29	263-280	263-CACGGAAT	C->G			TGAAGGCCC-280
15-SP-A-23 b	291-308	291-TGACCGAG	C->A			GAACCTGCG-308

Figure 2
Variant oligos used in this study. Oligos are color coded to correspond to the regions described in Figure 1. *1-SP-A-02a and 1-SP-A-02b contain one HLA-A*0201-specific mismatch and one additional mismatch (G > A in position 72).

In case of *a,a* homozygosity, similar fluorescence intensity is expected in both channels with a theoretical Cy5/Cy3 fluorescence intensity ratio = 1 ($_{Log}$Ratio = 0). This can be experimentally tested by arbitrarily selecting genomic fragments within the investigated region that, based on available information, are most likely conserved in every potential test sample (displayed in yellow in Figure 1) [2]. In this region, reference and test samples can be predicted to be *a,a* homozygous. Consistent deflection from 0 of $_{Log}$Ratios in these oligos denotes biases of labeled target or reference. Thus, the average of the $_{Log}$Ratio for these oligos is used as a normalization factor, constant (*k*), to correct the bias of both channels in the rest of the data set. The unlikely occurrence of SNP within the constant region could still be detected since in such cases, the $_{Log}$Ratio of one oligo will diverge from the rest of the constant region oligos.

After normalization, individual oligos are investigated. Homozygosity (*a,a*) is characterized by $_{Log}$Ratio ~0, digitally displayed as yellow in the array image (Figure 3, row I). Conversely, in the context of homozygosity different from the consensus (*b,b*), specific hybridization of the test sample to the variant oligo results in strong red (Cy5) signal (Va_1 in Figure 3, row II) while the consensus oligo spanning the same region reveals strong green signal caused by hybridization of the reference sample only (Figure 3, row II). Non-specific low affinity hybridization to an irrelevant variant oligo spanning the same regions is similar for both reference and test sample resulting in fluorescence intensity ratios close to 1 ($_{Log}$Ratio ~0) (Va_2 in Figure 3, row II).

Since human genomes are diploid, polymorphisms can occur in combinations and, therefore, detection of SNP

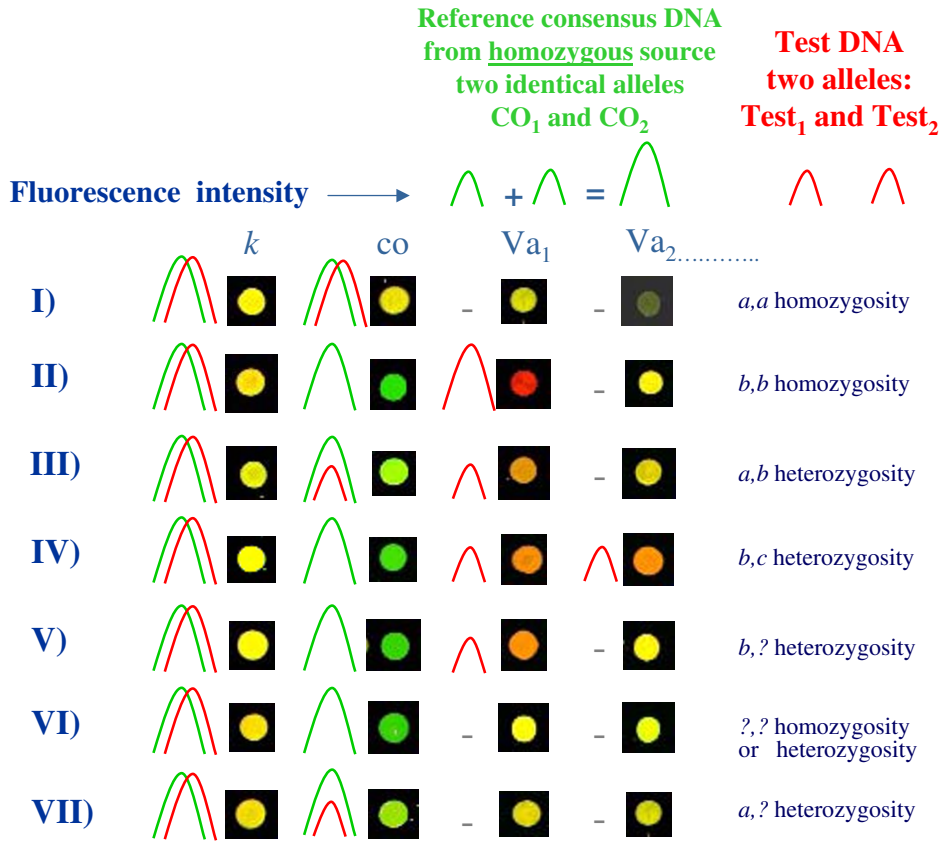


Figure 3
Principle of fluorimetric detection of SNP by proportional hybridization to oligonucleotide arrays in homozygous and heterozygous conditions. Portrait of proportional hybridization (of differentially labeled reference (Cy3) and test (Cy5) sample to overlapping oligos encompassing an arbitrarily chosen "consensus" sequence (consensus oligos = CO) or variant-specific oligos (Va_1 and $Va_{2....}$, see Figures 1 and 2). The homozygous reference sample consists of two alleles identical to the consensus sequence (HLA-A*0101). Differentially fluorescence-labeled reference and test samples are co-hybridized to an array slide spotted with the 18-nucleotide overlapping consensus oligos and variant oligos. Four consensus oligos representing a "conserved" region (*k*) were used to normalize the data set (see text). Reference sample consistently hybridizes to CO and never to Va oligos. Hybridization of test sample will determine variability in ratio of fluorescence intensity as portrayed by the digital images from a GenePix scanner. Possible combinations are: **I**) *a,a* type homozygosity (row I); **II**) *b,b* homozygosity (row II); **III**) *a,b* heterozygosity with one known allele (row III) and one unknown allele (row VII); **IV**) *b,c* heterozygosity with two known alleles (row IV), one known and one unknown allele (row V) and two new alleles (row VI). Question marks represent the hypothetical discovery of new allele(s).

should be possible in the context of heterozygosity. This discrimination can be achieved through fluorimetric assessment of the hybridization pattern based on the general principle that homozygous samples (two identical alleles) will generate twice the amount of signal for a given sequence than a heterozygous sample (Figure 3, rows III-VII). In the context of heterozygosity, a single allele hybridization to variant oligos generates a weaker signal than in the homozygous condition resulting in a lower $_{Log}$Ratio (specific hybridization over background). (Figure 3, row III-V Va_1 and row IV Va_2). Competitive hybridization to consensus oligos will lead to two patterns. In *a,b* heterozygosity (portrayed as lighter green by the digital image) at least one allele of the test sample will hybridize to the consensus resulting in $_{Log}$Ratio depression of lesser magnitude (lighter green) than in *b,b*

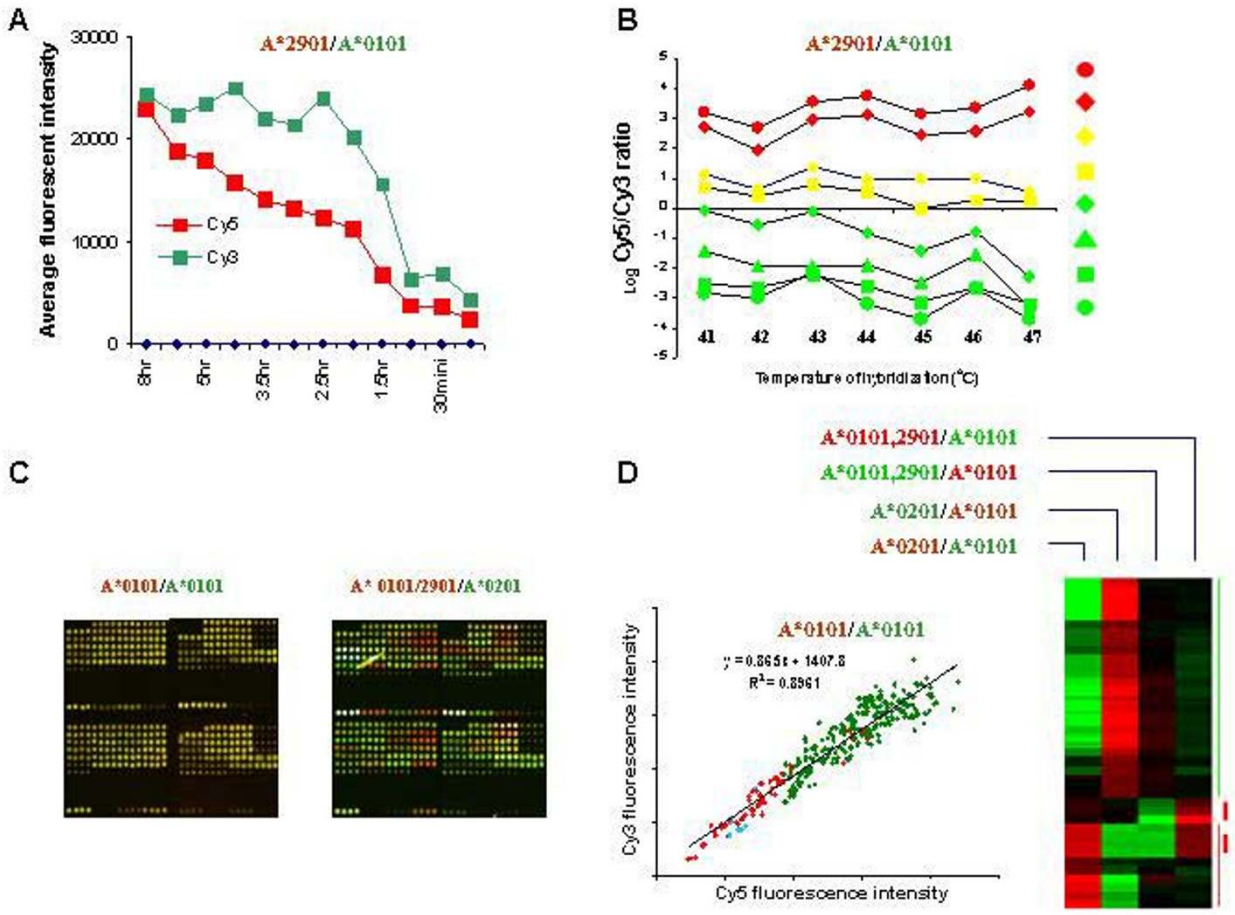


Figure 4

Effect of hybridization time (**a**) and temperature (**b**) in conditions of *b,b* type homozygosity (HLA-A*2091 labeled with Cy5 *vs* HLA-A*0101 labeled with Cy3). **b**) $_{Log}$Ratio for Va (HLA-A*2901-specific) oligos (red), *k*-oligos (yellow) and CO-oligos. **c**) Test for labeling bias and reproducibility in conditions of *a,a* homozygosity and *b,c* heterozygosity: gene pix image. **d**) Scatter plot of Cy3 and Cy5 fluorescence intensity in conditions of *a,a* homozygosity. In red are portrayed Va oligos; in green *k* and consensus oligos. A cluster of consensus oligos with poor hybridization kinetics due to high CG content (average of 75%) is portrayed in light blue. Reproducibility was tested in conditions of *b,b* homozygosity (HLA-A*0201 *vs* HLA-A*0101) and *a,b* heterozygosity (HLA-A*0101, 2901 *vs* HLA-A*0101) by reciprocally labeling test and reference targets in duplicate experiments as shown by the clusterogram. In spite of variable hybridization kinetics, differential hybridization could be observed consistently among Va and corresponding consensus oligos. Only HLA-A*0201-specific and relevant homologous consensus oligos (solid red and green bar respectively) or HLA-A*2901-specific oligos (dashed red line) are shown.

homozygosity where none of the test samples hybridizes to the consensus (Figure 3, row III and VII). In *b,c* heterozygosity none of the alleles hybridizes to consensus oligos and, therefore, a situation similar to *b,b* homozygosity occurs in the consensus oligos with strongly depressed $_{Log}$Ratios (Figure 3, row IV-VI).

Hybridization conditions were optimized at variable incubation times (Figure 4a) and temperatures (Figure 4b). Temperatures of 47°C for at least 4 hours yielded the best hybridization results that were adopted thereafter. Target density had no major repercussion on fluorescence intensity ratio beyond 80 ng (data not show). Labeling bias was assessed in conditions of *a,a* homozygosity by co-hybridizing samples from an HLA-A*0101 source dif-

ferentially labeled with Cy5 and Cy3 (Figure 4c). Good correlation of Cy5 with Cy3 signal intensity was observed based on individual oligo although the kinetics of hybridization varied remarkably among different oligos Figure 4d. In general, the hybridization efficiency correlated to the theoretical expectation that the variant oligos (red in the scatter plot) should have low affinity for HLA-A*0101 targets. However, exceptional variant oligos with strong hybridization affinity to HLA-A*0101 and, conversely, a cluster of consensus oligos (average CG content = 75%, range 67 to 94%) with poor hybridization kinetics (blue in the scatter plot) were noted. Strong hybridization of the reference sample to a variant oligo could be associated with GC rich oligo or a T → A or T → C replacement, while decreased hybridization of reference sample to consensus oligo could be caused by self-annealing of GC rich sequences. These, fluctuations in strength of hybridization had, however, little impact on the analysis, since differential hybridization could still be easily detected when samples from different genotypes were tested (clusterogram in Figure 4d). For instance, reciprocal labeling in the context of *b,b* homozygosity or *b,c* heterozygosity demonstrated perfect symmetry in hybridization to variant and consensus oligos.

Data could be ordered sequentially into genotypic profiles according to sequence position. The *a,a* homozygosity phenotype was characterized by $_{Log}$Ratios with minimal deviation from 0 (Figure 5a,5F). Conversely, *b,b* homozygosity yielded extremely high $_{Log}$Ratios in variant oligos and strongly depressed $_{Log}$Ratio in the corresponding consensus oligos (Figure 5b). Noteworthy is the pattern of proportional hybridization to the variant oligo *p* (*p* = oligo 13-SP-A-02 at position 248–265) that encompasses a double nucleotide variant from CA → GG in position 256–257. The oligo containing this double mismatch further reduced the hybridization affinity to the reference sample, proportionally enhancing the affinity of the test sample leading to higher $_{Log}$Ratio for oligo *p* in Figure 5b. The "unknown" SNP (from position 250 to 310) was revealed by absence of variant-specific hybridization but highly depressed $_{Log}$Ratios of the corresponding consensus region oligos (orange asterisks and dashed line in Figure 5b). This phenomenon may exemplify the ability of this method to detect allelic variants in the absence of variant-specific oligos. Even in conditions of *a,b* heterozygosity (Figure 5c), "unknown" SNP could be suspected due to the relatively reduced $_{Log}$Ratios (0.5 ~1) in variant oligos and more characteristically less depressed $_{Log}$Ratios in corresponding consensus oligos compared with *b,b* homozygosity (Figure 5b). Noteworthy is the exceptional behavior of the variant oligo *p* that contains a double nucleotide variant (*p* in Figure 5c) which leads to higher $_{Log}$Ratios even in the context of *a,b* heterozygosity. However, the corresponding consensus oligos are not affected by the double nucleotide variant and, therefore, display a typical *a,b* heterozygosity pattern distinguishable from b,b homozygosity since one allele in the test sample is identical to the reference.

Heterozygosity with both alleles different from the consensus (*b,c*) represents the most complex situation (Figure 5d). In this case, heterozygous samples such as A*0201, 2901, $_{Log}$Ratios behave as in *b,b* homozygosity at this specific region since both alleles hybridize to the same variant oligo and neither to the correspondent consensus oligos. The discriminatory power of this phenomenon can be better illustrated by comparing the hybridization pattern of the same oligo in conditions of *b,b* homozygosity (HLA-A*2901 homozygous) and *a,b* heterozygosity (HLA-A*0101,2901 heterozygous) (*x* in Figure 6). In this case, differential hybridization pattern is marked by doubling and halving of $_{Log}$Ratios in variant and consensus oligos respectively (see also Table in Figure 6). Thus, variable types of heterozygosity in the test sample may affect differently the proportional hybridization to the correspondent consensus oligos. An extreme example occurs when one allele of the test sample hybridizes to the variant oligo (14-SP-A*29; position 263–280) while an unknown SNP is present in the other allele within the same region (HLA-A*0201, *y* in Figure 5d). This exemplifies a case of *b,c* heterozygosity in which one SNP is unknown as theorized in row V, Figure 3. In this case, a reduced $_{Log}$Ratio in the variant oligo associated with extremely depressed $_{Log}$Ratios in the corresponding consensus oligo are observed because both test alleles are different from the consensus. The power of discrimination *a,b* and *b,c* heterozygosity in regions with "unknown " SNP is underlined by the dashed line in Figure 5d. The blue line highlights polymorphisms occurring in both alleles (A*0201,2901). Thus, the corresponding consensus oligos display strongly depressed $_{Log}$Ratios comparable to *b,b* type homozygosity. In the following region, the dashed orange line indicates polymorphisms occurring only in one allele (A*0201 and not 2901) resulting in less depressed $_{Log}$Ratios in the corresponding consensus oligos as in the case of a/b heterozygosity.

Occurrence of more than one SNP within an oligo could complicate the data analysis. For instance, in the case described in Figure 5d, two regions with a double mismatch are observed which are characterized by $_{Log}$Ratios disproportionately high for the HLA-A*0201/2901 heterozygous state (*p* and *q* depict oligos 13-SP-A-02 and 3-SP-A-2901 respectively, specific for the two alleles). In conditions of HLA-A*0201/2901 heterozygosity, consensus oligos corresponding to polymorphism *p* representing A*0201, strongly hybridized to the reference sample resulting in a *b,b* homozygosity hybridization pattern. In

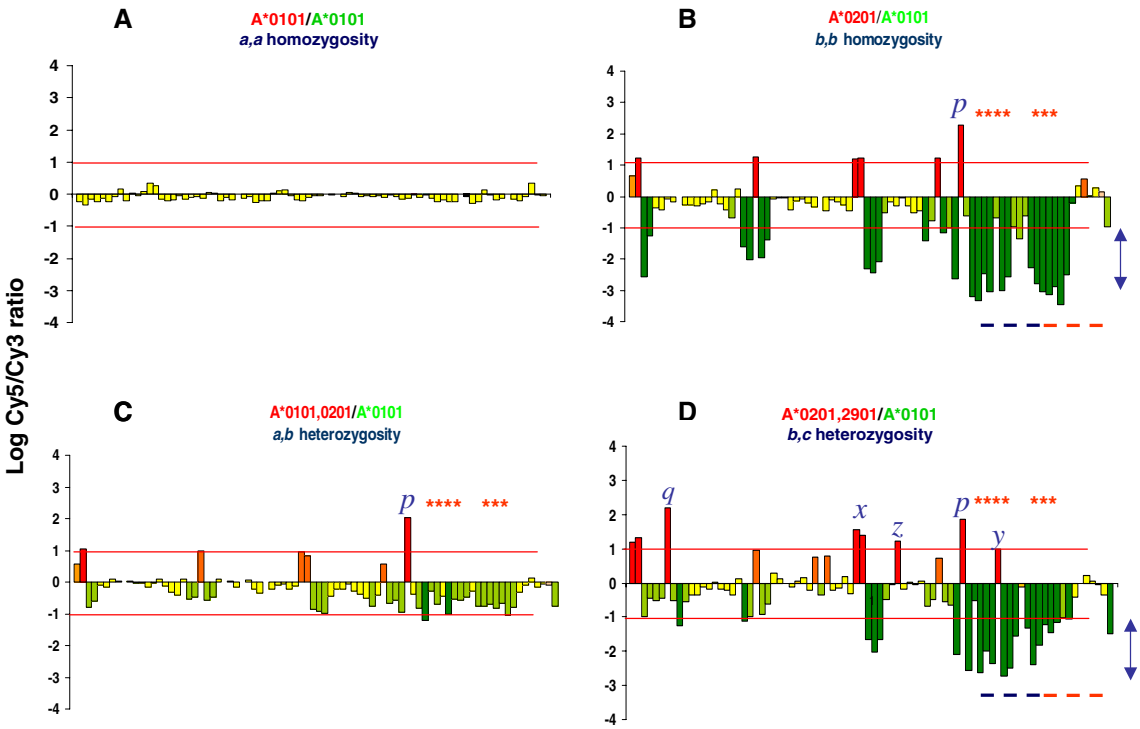
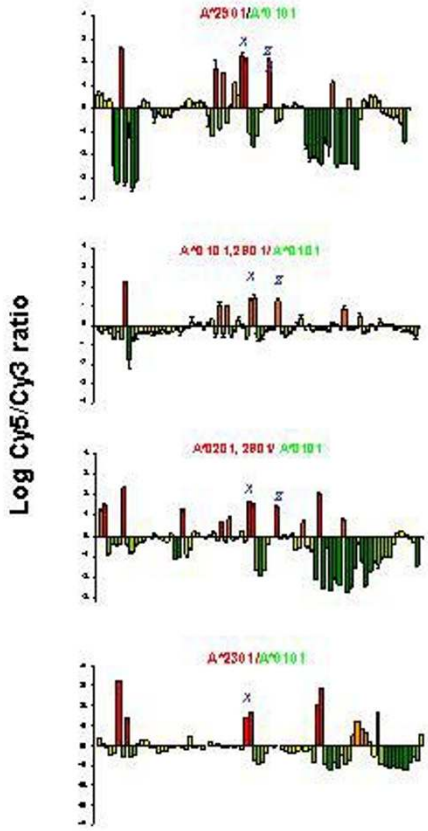


Figure 5
Profiling by proportional hybridization of HLA-A exon 2. Four patterns are described: *a,a* homozygosity (**A**); *b,b* homozygosity (**B**); *a,b* heterozygosity (**C**) and *b,c* heterozygosity (**D**). Each bar represents the $_{Log}$Ratio for individual oligos sequentially positioned in 3' to 5' direction. Variant oligos are nested between homologous consensus oligos and appear as bright ($_{Log}$Ratio >1) or light red ($_{Log}$Ratio > 0.5 and ≤ 1). $_{Log}$Ratio for any oligo between ≤ 0.5 or ≥ -0.5 are shown in yellow. Consensus oligos in various conditions of mismatch with test DNA are shown as light ($_{Log}$Ratio < -0.5 and ≤ -1) or dark ($_{Log}$Ratio < -1) green. The orange lines delimit $_{Log}$Ratio between 1 and -1. The letter *p* points to a variant oligo with a double nucleotide variant (CA → GG at 256–257). Another double mismatch-containing oligo is pointed out by *q* (CA → AC at 99–100). The letter *x* points to a variant oligo (9-SP-A-02) spanning a region of *b,c* type heterozygosity (HLA-A*0201 and A*2901 differ A → G from the consensus). The letter *y* points to a situation where a variant oligo (14-SP-A-03, A29 at 263–280) encompasses also a "unknown" SNP present in the other allele (HLA-A*0201). *z* shows a variant oligo (10-SP-A-29 at 211–229) whose increased $_{Log}$Ratio is not associated with decreased $_{Log}$Ratio in the homologous consensus oligo. The orange asterisks and the dashed lines show a genomic region in which "unknown" HLA-A*0201 polymorphisms are associated (blue) or not associated (orange) with HLA-A*2901 polymorphic sites.

the case of polymorphism *q* representing A*2901 variant, the same A*0201 region is identical to the consensus and, therefore, *a,b* heterozygosity pattern is observed.

An unexplainable finding was observed where specific hybridization to a variant oligo is not associated with depressed $_{Log}$Ratios of the correspondent consensus oligo (*z* in Figure 5d). A purposeful mismatch, a C → A switch, may have less repercussions on the hybridization pattern of this oligo. Indeed, even in homozygous conditions (*z* in Figure 6) the $_{Log}$Ratios are only minimally depressed in this case. In these relatively rare occurrences (only case in our study) an "unknown" polymorphism would not be detected.



$_{Log}$Cy5/Cy3 ratio (average ± SDev of 3 consecutive experiments)

	Allele-specific oligos		Consensus oligos	
	Homo	Hetero	Homo	Hetero
Name of oligo	Ave $_{Log}$ Cy5 / Cy3 ± SDev			
1-SP-A-02a*	0.4 ± .13	-0.1 ± .12	0.3 ± .09	-0.3 ± .14
2-SP-A-02b*	0.5 ± .15	-0.3 ± .20	0.3 ± .09	-0.3 ± .14
3-SP-A-2901	2.6 ± .08	2.0 ± .12	-3.2 ± .11	-0.8 ± .12
4-SP-A-0205,0206	-0.6 ± .63	-0.8 ± .06	-3.4 ± .13	-0.5 ± .04
5-SP-A-02,30	-0.1 ± .00	-0.4 ± .22	0.3 ± .09	-0.4 ± .11
6-SP-A-2901 b	1.7 ± .38	0.6 ± .01	0.9 ± .01	0.6 ± .17
7-SP-A-2901 a	1.5 ± .02	0.7 ± .06	0.9 ± .01	0.6 ± .17
8-SP-A-0205	2.3 ± .10	1.3 ± .11	-1.6 ± .05	-0.7 ± .10
9-SP-A-02,A03,A23, A29	2.1 ± .14	1.1 ± .13	-1.6 ± .05	-0.7 ± .10
10-SP-A-29	2.1 ± .04	1.3 ± .14	-0.5 ± .08	-0.5 ± .17
11-SP-A-02 b	0.2 ± .04	-0.2 ± .22	0.1 ± .06	-0.4 ± .12
12-SP-A-23 a	-2.1 ± .14	-0.4 ± .22	-2.1 ± .09	-0.6 ± .04
13-SP-A-02	-2.2 ± .12	-0.4 ± .07	-2.1 ± .09	-0.7 ± .11
14-SP-A-03, A29	1.2 ± .13	-0.6 ± .18	-2.4 ± .07	-0.6 ± .05
15-SP-A-23 b	0.1 ± .13	0.4 ± .19	-0.1 ± .25	-0.4 ± .11

•1-SP-A-02a and 1-SP-A-02b contain one HLA-A*0201-specific mismatch and one Additional mismatch in position 72 as a Pre-fabricated "error". n.d. = not done

Figure 6
Genomic "fingerprinting" using proportional hybridization. Description of individual genotype combinations. Four allelic combinations are described here using graphics similar to the ones described in Figure 5. The data for each allelic combination represent the average $_{Log}$Ratio ± SDev of three separate tests. In the accompanying Table, the $_{Log}$Ratio in variant oligos and corresponding homologous consensus oligo are shown for *b,b* homozygosity (HLA-A*2901 vs reference HLA-A*0101) and *a,b* heterozygosity (HLA-A*0101, A*0201 vs reference HLA-A*0101) that are portrayed by the top two bar graphs.

To statistically assessed the ability of this method to discriminate differences between *a,a* homozygosity from other combinations in conditions of unknown SNP, we analyzed the behavior of consensus oligos only in various experimental conditions. For each 18-mer probe, starting from 3rd base and ending at 16th base, if the test target contained at least one single nucleotide different for the consensus sequence HLA-A*0101, we considered this specific region SNP(+); otherwise it was SNP(-). Various combinations were tested as shown in Figure 7. In all permutations tested (given appropriate sample size) the set of 4-nucleotide tiling consensus oligo alone demonstrated a high SNP discriminatory power. In particular, even the theoretically most problematic combination (*a,b* vs *a,a*), least divergent from the consensus, could be easily significantly discriminated in all permutations tested.

We then tested the sensitivity and specificity of this strategy by examining the relationship between LogRatio and the actual SNP condition via ROC (Relative Operating Characteristics) analysis [21]. The ROC curves in Fig. 5 were composed of 3 different groupings. As it could be expected, optimal threshold for detection of polymorphism related sensitivity and specificity varied according to the allelic combination. When the test sample is most different from the consensus reference as for *b,b* an *b,c*

***t*-test comparison of Log_2Ratio differences among various allele combinations**

Allelic Combination	#samples	Allelic Combination	#samples	df	p-value
(*b,b*)		(*a,a*)			
A0201	2	A0101, A0101	4	4	0.00001
A0301	1	A0101, A0101	4	3	0.0008
A2301	1	A0101, A0101	4	3	0.0002
A2901	8	A0101, A0101	4	10	0.0003
(*a,b)*		(*a,a*)			
A0101,0201	1	A0101,A0101	4	3	0.0015
A0101,0301	1	A0101,A0101	4	3	0.0025
A0101,2301	2	A0101,A0101	4	4	0.0015
A0101,2901	5	A0101,A0101	4	7	0.0008
(b,b)		(*a,b*)			
A0201	2	A0101,0201	1	1	0.0819
A2301	1	A0101,2301	2	1	0.0554
A2901	8	A0101,2901	5	11	0.0003

Figure 7
Degree of discrimination among different HLA alleles using proportional hybridization to oligonucleotide arrays. Two sample *t*-tests were used to differences between two means. P-values were calculated for each (*a,b*) vs.(*a,a*); (*a,b*) vs. (*b,b*); and (*b,b*) vs. (*a,a*)

higher accuracy with sensitivity 82% and specificity 96% is observed (red line). The worst accuracy was noted when test and reference samples were closest as in the case of *a,b* heterozygosity (sensitivity 82% and specificity of 82%) (blue line). The most informative analysis was, however, provided using data from all the possible combinations (green line in Figure 8) since in most common experimental condition the relationship between test and consensus sample is not known and, therefore, all possible allelic combinations should be expected. In that case an optimal threshold of $_{Log}$Ratio ≤ -0.62 yielded sensitivity at 82% and specificity at 89%. Thus, this strategy may identify 4 out of 5 SNP with 90% accuracy with the highest chance of discriminating false positive result when an *a,b* heterozygous sample is tested.

Discussion

In this study, we describe a simplified strategy for the detection of known and unknown SNP that, because of the decreased cost, could be applied to large clinical studies. A flow chart describing the algorithm for the identification of SNP using this method is shown in Figure 9. Although this strategy was conceived to identify SNP in genes for which little is known about their polymorphism(s), it could also be used for routine typing of known alleles. This could be achieved by preparing allele-specific profiles and matching test results with known permutations collected in databases compiling repeated hybridization profiles as template to eliminate possible false interpretations. Annotation of the test sample genotype will be reported after comparison with established templates. The HLA alleles and corresponding oligos used

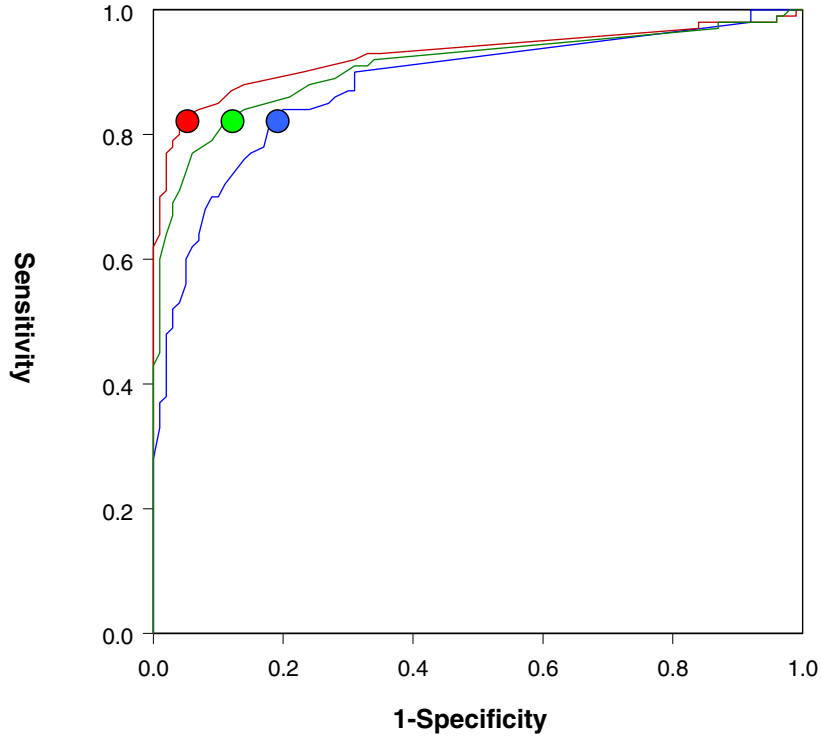


Figure 8
Sensitivity and specificity analysis of the strategy. 18-mer probes were assigned to SNP (+) if the sequence of the test channel probe of a specific spot contained at least one single nucleotide different from the sequence A*0101. The other probes were considered SNP(-). ROC (Relative Operating Characteristic) analysis [21] of curves for 3 different allelic combinations. The red line represents a ROC curve generated from 448 spots in 7 *b,b* homozygosity and *b,c* heterozygosity experiments with 269 SNP(-) and 179 SNP(+) spots. The blue line represents ROC curves generated from 265 spots in *a,b* heterozygosity experiments with 169 SNP(-) and 87 SNP(+) spots. The green line represents a ROC curve from 704 spots in all experiments with a total of 438 SNP (-) and 266 SNP (+) spots. The dots on the curves stand for optimal threshold in each group. Threshold for $_{Log}$Ratio were based on ROC analysis to balance sensitivity and specificity. For the group with *b,b* and *b,c* heterozygosity (red line) experiments the optimal threshold was a $_{Log}$Ratio ≤ -0.91. The sensitivity using this threshold was 82% and specificity 96%. The optimal threshold for the *a,b* heterozygosity (blue line) was $_{Log}$Ratio ≤ -0.43 with a sensitivity of 82% but a specificity drop to 82%. For the analysis combined all the data (green line) the optimal threshold was $_{Log}$Ratio ≤ -0.62 with a sensitivity at 82% and specificity at 89%. This last ROC exemplifies the most common experimental condition where the relationship between test and reference sample is not known and, therefore, all possible allelic combinations should be expected.

here were selected to exemplify various combinations. Various permutations of homozygosity and heterozygosity can be observed that produce complex hybridization patterns highly specific for a particular phenotype. In these highly polymorphic conditions, each haplotype combination maintains a highly reproducible profile characterized by minimal variance. This allows the creation of "genotypic masks" within narrow ranges of variation to "fingerprint" known haplotype permutations for high-throughput typing of highly polymorphic genes.

We, therefore, describe a potentially powerful and efficient strategy for high-throughput screening of genes for which little is known about their polymorphism. This

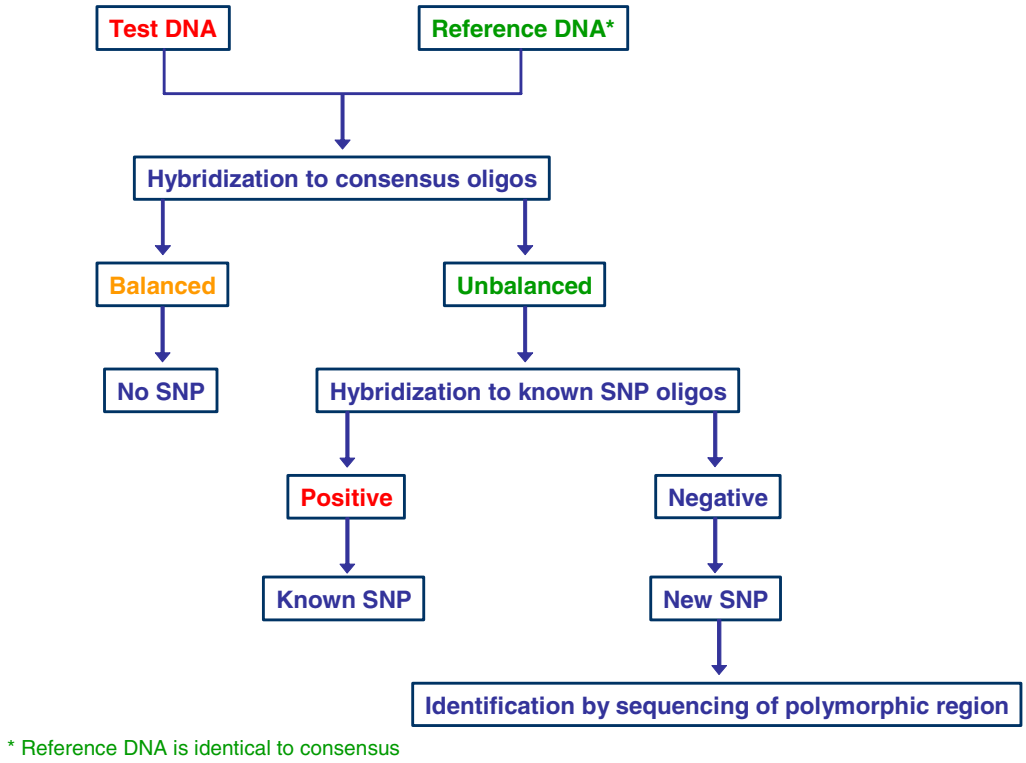


Figure 9
Flow chart summarizing the proposed algorithm to screen for known or novel SNP using proportional hybridization to oligonucleotide arrays.

strategy could also be used to identify mutations in disease processes or for typing known allelic variants of well-characterized genes such as HLA. This, however, would require specialized design of numerous oligos encompassing known variants and supportive software for efficient data interpretation. Various scenarios are best exemplified by using as a model a highly polymorphic region of the human genome such as exon 2 of the HLA-A locus. The simplest case would be the investigation of a gene characterized by minimal or no polymorphism. In this case, screening of samples from different ethnic groups would yield a pattern described as *a,a* type homozygosity, similar to the one shown in Figure 8a. Another possibility would be the investigation of a gene with few but relatively common polymorphism(s). In this case the occurrence of *b,b* type homozygosity would be common as depicted in Figure 8b. In the same situation, *a,b* type heterozygosity would also frequently occur as shown in Figure 8c. Finally, a most complex scenario, likely to occur only for genes characterized by high polymorphic prevalence is portrayed in Figure 8d. In this case, *b,c* type heterozygosity should occur frequently as it might be expected for the HLA loci. However, SNP occur in the human genome on average every 600 – 2,000 bases [1,2]. Therefore, most genes are characterized by a relatively narrow range of polymorphism that would allow a relatively simple design of oligo-array chips and interpretation of results. Independently of the genomic region investigated, this strategy can identify unknown variants through observation of disproportionably depressed $_{Log}$Ratios in consensus oligos.

Conclusions

The SNP detection system described her may provide a great improvement in the ability to screen different genes for the frequency and location of polymorphic sites, which ca be confirmed by site directed sequencing limited to the region of interest. Thus, the best application of this strategy stems from the clinical need to rapidly segregate genes characterized by presence or lack of polymorphisms in their coding or regulatory regions that may affect clinical behaviors. A good example of such application is the screening of cytokines, chemokines and their receptors whose polymorphism(s) have been associated to individual predisposition to immune pathology, survival of transplanted organs and predisposition to cancer [16,22–25].

Material and methods

Oligo nucleotide probe design

The HLA-A locus exon 2 region from position 73–346, allele A*01011, was used for the design of consensus oligos (Figure 1). Allele-specific (variant) oligos were designed based on single or double nucleotide variants according to alignment to the arbitrarily selected consensus sequence (HLA-A*0201). Variant and consensus oligos consisted of 18-mers with a 5' amino-modifier having a six-carbon spacer for immobilization (Operon Technologies, Inc, Alameda, CA). The polymorphic site in the variant oligos was designed in the centermost position. Melting temperatures of oligo probes were maintained as close as possible to a range of 56–60°C. 350–400 pmol/**μ**l oligo probes synthesized in 96-well format were dried and re-suspended in 3 × SSC for printing.

Oligonucleotide array printing and post process of slides

Probes were spotted onto a 3D-link activated slide for covalent immobilization (Motor roller) using OmniGrid robotic printer (Genemachine) with four printing pins picking up 0.25 **μ**l of probe solution and depositing 0.6 nl per spot (TeleChem International, Inc.). Each spot was quadruplicated to minimize printing bias and test reproducibility. Spot diameter was 90–100 **μ**m, spaced at 250 **μ**m to prepare 4 × 16 × 6 spot/arrays. After printing, slides were kept in a sealed humidifier chamber at room temperature overnight and blocked with 50 mM ethanolamine, 0.1 M Tris, pH9 and 0.1% SDS at 50°C for 15 minutes followed with rinsing in water twice, washing with 4 × SSC/0.1% SDS 50°C for 45 minutes, rinsing with water briefly and centrifuging at 800 rpm for 3 minutes with microplate carriers. Arrays were then stored in a desiccator until use.

Test and consensus reference samples

Genomic DNA was isolated from EBV transformed B cell line. 12 heterozygous or homozygous samples were tested for oligo nucleotide array hybridization and confirmed by sequence-based typing using the ABI Prism 3700-96 Analyzer.

Preparation of target nucleic acids

In order to generate single strand DNA, PCR products from Exon 1 to Exon 5 of the HLA-A locus were amplified with an attachment of T7 promoter to the 5' end using 5' T7-EX1A-6 primer 5'AAACGACGGCCAGTGAATACGACTCACTATAGGCGC CAGACGCCGAGGATGGCC3') and three 3' primers (3' EX 5-A 993-1 CAT TGC TGG CCT GGT TCT CC; 3' EX 5-A 993-2 CAT TGC TGG CCT GGT TCT CTT; 3' EX 5-A 993-3 CAT TGC TGG CCT AGT TCT CTT). PCR reaction was mixed with 25 **μ**l of HotStart PCR reagents (Qiagen, CA), 5 ng-0.5 **μ**g of genomic DNA, 5 **μ**l of 15 **μ**M 5'primer, 5 **μ**l of 15 **μ**M 3'primer mix and H_2O to a 50 **μ**l final volume. The reaction was cycled at 95°C for 10 minutes, 96°C for 35 seconds, 65°C for 45 seconds, 72°C for 3 minutes, 4 cycle; 96°C for 30 seconds, 60°C for 40 seconds, 72°C for 3 minutes, 19 cycles and 96°C for 30 seconds, 55°C for 40 seconds, 72°C for 2 minutes, 9 cycles. One **μ**l of PCR product from each sample was analyzed using a Bioanalyzer on DNA7500 chip (Agilent Biotechnology). Approximately 2,000 bp amplicons from each sample were amplified. The PCR products were then precipitated with EtOH at room temperature and re-suspended in DEPC-treated H_2O at 0.1 **μ**g/**μ**l concentration. *In vitro* transcription (IVT) was performed using an Ambion T7 Megascript Kit (Cat. #1334). For each sample, the following reaction mixture was made: 4 **μ**l of each 75 mM NTP (A, G, C and UTP), 4 **μ**l reaction buffer, 4 **μ**l enzyme mix (RNase inhibitor and T7 phage polymerase) and 1 **μ**g purified PCR product in 16 **μ**l DEPC-treated H_2O. The reactions were then incubated at 37°C for six hours to permit transcription. Amplified RNA was then purified using TRIzol reagent according to manufacture instruction (GibcoBRL) and re-suspended in 40 **μ**l of DEPC water. RNA concentrations were estimated by using a Bioanalyzer on RNA 6000 chip (Agilent Biotechnology).

Target labeling and hybridization

Fluorescence-labeled single strand cDNA was generated by reverse transcription (RT) and used for hybridization. In the RT reaction, 4 **μ**l of first strand buffer, 1 **μ**l random hexmer (8 **μ**g/**μ**l; Boehringer Mannheim), 2 **μ**l 10 × low TdNTP (5 mM A, C and GTP, 2 mM dTTP), 2 **μ**l 1 mM CydUTP (Cy3 for reference sample or Cy5 for test sample unless otherwise specified), 2 **μ**l 0.1 M DTT, 1 **μ**l Rnasin and 1.2 **μ**g amplified RNA in 8 **μ**l DEPC H_2O were mixed and heated for five minutes at 65°C. This was followed by addition of 1 **μ**l Superscript II (Life technology,), 40 minutes of incubation at 42°C, another 1 ul of Superscript II and 50 more minutes continued incubation at 42°C. Reactions were stopped by addition of 2.5 **μ**l 500 mM EDTA, 5 **μ**l 1 M NaOH and heated to 65°C for 15 minute

to hydrolyze the RNA. Tris buffer (12.5 **μ**l of 1 M) was added immediately to neutralize the pH, and the volume risen to 70 **μ**l by adding 35 **μ**l of 1 × TE. Target solution was then applied to Bio-6 column according to the manufacturer's instructions. The flow through mixed with 200 **μ**l 1 × TE was concentrated to 20–40 **μ**l using Microcon YM-30 column (Millipore) and further concentrated to 8 ul using speed-vacuum.

Cy3- and Cy5-labeled probes were combined (1:1 ratio) and 5 **μ**l 20 × SSC, 0.5 **μ**l 10% SDS and 0.5 **μ**l of 4 mg / ml salmon sperm DNA were added to the probe for hybridization. The samples were then heated for two minutes at 99°C. Prepared probe mixture was applied to an array slide with cover slid and hybridized at 47°C for different amounts of time as described in the text. Slides were washed with 4 × SSC, 2 × SSC with 0.1 % SDS, 1 × SSC, 0.2 × SSC and 0.05 × SSC sequentially for one minute each step and dried by centrifugation at 800 rpm for 3 minutes. The slides were then scanned for fluorescent signal using a GenePix 4000B scanner and the results analyzed using GenePix Pro3 software (Axon Instruments, Inc.).

References

1. Wang DG, Fan J-B, Siao C-J, Berno A, Young P and Sapolsky R *et al.*: **Large-scale identification, mapping and genotyping of single-nucleotide polymorphisms in the human genome.** *Science* 1998, **280:**1077-1082.
2. The International SNP Map Working Group: **A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms.** *Nature* 2001, **409:**928-933.
3. Cooper DN, Ball EV and Krawczak M: **The human gene mutation database.** *Nucleic Acids Res* 1998, **26:**285-287.
4. Ng PC and Henikoff S: **Accounting for human polymorphisms predicted to affect protein function.** *Genome Res* 2002, **12:**436-446.
5. Collins FS, Brooks LD and Chakravarti A: **A DNA polymorphism discovery resource for research on human genetic variation.** *Genome Res* 1998, **8:**1229-1231.
6. Schafer AL and Hawkins JR: **DNA variation and the future of human genetics.** *Nature Biotech* 1998, **16:**33-39.
7. Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM and Hacker CR *et al.*: **Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21.** *Science* 2001, **294:**1719-1723.
8. Kwok PY: **Genetic association by whole-genome analysis.** *Science* 2001, **294:**2669-1670.
9. Chee M, Yang R, Hubbell E, Berno A, Xiaohua C and Stern D *et al.*: **Accessing genetic information with high-density DNA arrays.** *Science* 1996, **274:**610-614.
10. Hacia JG, Brody LC, Chee MS, Fodor SP and Collins FS: **Detection of heterozygous mutations in BRCA1 using high density oligonucleotide arrays and two-colour fluorescence analysis [see comments].** *Nat Genet* 1996, **14:**441-447.
11. Saiki RK, Walsh PS, Levenson CH and Erlich HA: **Genetic analysis of amplified DNA with immobilized sequence-specific oligonucleotide probes.** *Proc Natl Acad Sci U S A* 1989, **86:**6230-6234.
12. Lockhart DJ, Dong H, Byrne MC, Folliette MT, Gallo MV and Chee MS *et al.*: **Expression monitoring of hybridization to high-density oligonucleotide arrays.** *Nature Biotechnol* 1996, **14:**1675-1680.
13. Chen J, Iannone MA, Li M-S, Taylor JD, Rivers P and Nelsen AJ *et al.*: **A microsphere-based assay for multiplex single nucleotide polymorphism analysis using single base chain extension.** *Genome Res* 2000, **10:**549-557.
14. Tong AK and Ju J: **Single nucleotide polymorphism detection by combinatorial fluorescence energy transfer tags and biotinylated dideoxynucleotides.** *Nucleic Acids Res* 2002, **30:**e19.
15. Kwok PY: **High-throughput genotyping assay approaches.** *Pharmacogenomics* 2000, **1:**95-100.
16. Turner D, Choudhury F, Reynard M, Railton D and Navarrete C: **Typing of multiple single nucleotide polymorphisms in cytokine and receptor genes using SNaPshot.** *Hum Immunol* 2002, **63:**508-513.
17. Guo Z, Gatterman MS, Hood L, Hansen JA and Petersdorf EW: **Oligonucleotide arrays for high-throughput SNPs detection in the MHC class I genes: HLA-B as a model system.** *Genome Res* 2002, **12:**447-457.
18. Hacia JG: **Resequencing and mutational analysis using oligonucleotide microarrays.** *Nature Genetics* 1999, **21:**42-47.
19. Hacia JG, Sun B, Hunt N, Edgemon K, Mosbrook D and Robbins C *et al.*: **Strategies for mutational analysis of the large multiexon ATM gene using high-density oligonucleotide arrays.** *Genome Res* 1998, **8:**1245-1258.
20. Adams SD, Barracchini KC, Simonis TB, Stroncek D and Marincola FM: **High throughput HLA sequence-based typing utilizing the ABI prism 3700 analyzer.** *Tumori* 2001, **87:**s41-s44.
21. Swets JA: **Measuring the accuracy of diagnostic systems.** *Science* 1988, **240:**1285-1293.
22. Keen LJ: **The extent and analysis of cytokine and cytokine receptor gene polymorphism.** *Transpl Immunol* 2002, **10:**143-146.
23. McCarron SL, Edwards S, Evans PR, Gibbs R, Dearnaley DP and Dowe A *et al.*: **Influence of cytokine gene polymorphism on the development of prostate cancer.** *Cancer Res* 2002, **62:**3369-3372.
24. Howell WM, Turner SJ, Bateman AC and Theaker JM: **IL-10 promoter polymorphisms influence tumour development in cutaneous malignant melanoma.** *Genes Immun* 2001, **2:**25-31.
25. Howell WM, Bateman AC, Turner SJ, Collins A and Theaker JM: **Influence of vascular endothelial growth factor single nucleotide polymorphisms on tumour development in cutaneous malignant melanoma.** *Genes Immun* 2002, **3:**229-232.