Wei Han · Shea Ping Yip · Jing Wang
Maurice K. H. Yap

Using denaturing HPLC for SNP discovery and genotyping, and establishing the linkage disequilibrium pattern for the all-*trans*-retinol dehydrogenase (*RDH8*) gene

Received: 24 September 2003 / Accepted: 14 October 2003 / Published online: 6 December 2003
© The Japan Society of Human Genetics and Springer-Verlag 2003

Abstract All-*trans*-retinol dehydrogenase (RDH8) is a visual cycle enzyme that reduces all-*trans*-retinal to all-*trans*-retinol. As part of an on-going effort to map genes involved in complex eye diseases, myopia in particular, using association studies, single nucleotide polymorphisms (SNPs) were identified and linkage disequilibrium (LD) pattern was established within and around the *RDH8* gene. We used denaturing high-performance liquid chromatography (DHPLC) to screen SNPs in four DNA pools each consisting of DNA from five individuals and genotyped the identified SNPs in 150 Chinese subjects from Hong Kong. Fifteen SNPs were identified: seven were common with the minor allele frequency >0.05 and ten were novel. Common SNPs were included in LD and haplotype analysis using the ASSOCIATE and EH programmes. Four SNPs in the 3' region exhibited significant LD and formed a haplotype block, while three common SNPs in the 5' region did not exhibit useful LD. The LD pattern around the *RDH8* gene suggested that one SNP from the 3' region and two to three SNPs from the 5' region were needed in association studies involving *RDH8*. Our results demonstrated the efficiency of DHPLC in screening SNPs when coupled with DNA pooling strategy and in genotyping SNPs.

Keywords Single nucleotide polymorphism (SNP) · All-*trans*-retinol dehydrogenase (*RDH8*) · Linkage disequilibrium (LD) · Denaturing high-performance liquid chromatography (DHPLC) · Association studies · Chinese population

Introduction

All-*trans*-retinol dehydrogenase (RDH8) is the enzyme that catalyses the reduction of all-*trans*-retinal to all-*trans*-retinol in the presence of NADPH (Rattner et al. 2000). It is a member of the short chain dehydrogenase/reductase family and is located in the outer segments of photoreceptors (hence also known as photoreceptor retinol dehydrogenase). It is very important in the visual cycle and begins the rhodopsin regeneration pathway by reducing all-*trans*-retinal, the product of bleached and hydrolysed rhodopsin (Rando 2001). This is a rate-limiting step in the visual cycle (Saari et al. 1998). The human *RDH8* gene consists of six exons spanning about 9 kilobasepairs (kb), and maps to chromosome 19p13 (Rattner et al. 2000).

Myopia (or short-sightedness) is a complex eye trait with both genetic and environmental risk factors contributing to its genesis (Saw et al. 1996). It is much more common in Orientals than in Caucasians (Saw et al. 1996; Wong et al. 2000). In Hong Kong, the prevalence peaks at age 19–39 at 70% (Goh and Lam 1994). In parallel with our existing efforts to identify myopia susceptibility genes, we identify single nucleotide polymorphisms (SNPs) in myopia candidate genes, which can then be tested by association studies. Association studies are powerful in detecting genes with small effects in complex traits (Lander and Schork 1994; Johnson and Todd 2000). Retinoic acid is an important signalling molecule in the eye and has been shown to be a potential mediator between refractive error and compensatory eye growth (Seko et al. 1998; Mertz and Wallman 2000). In other words, genes encoding enzymes, binding proteins,

W. Han · J. Wang
Department of Ophthalmology, The First Affiliated Hospital,
Medical College, Zhejiang University, Hangzhou, China

W. Han · M. K. H. Yap
Department of Optometry and Radiography,
The Hong Kong Polytechnic University,
Hung Hom, Kowloon, Hong Kong, SAR, China

S. P. Yip (✉)
Biomedical Science Section, School of Nursing,
The Hong Kong Polytechnic University,
Hung Hom, Kowloon, Hong Kong, SAR, China
E-mail: hsspyip@polyu.edu.hk
Tel.: +852-2766-7906
Fax: +852-2364-9663

or receptors involved in the metabolism of retinoic acid are potential candidate genes. Here we report the identification of SNPs for the *RDH8* gene and its pattern of linkage disequilibrium (LD) in a Chinese population.

This study employed denaturing high-performance liquid chromatography (DHPLC) to screen SNPs in DNA pools. In DHPLC, products amplified by polymerase chain reaction (PCR) are fractionated in a special column under partially denaturing conditions by ion-pair reverse-phase liquid chromatography (Xiao and Oefner 2001). Heteroduplexes formed from different alleles have melting properties different from perfectly matched homoduplexes and can be differentially eluted from the latter. Potential SNPs were then confirmed and characterised by direct DNA sequencing. DNA samples were pooled before PCR amplification to increase the throughput efficiency of SNP detection by DHPLC (Wolford et al. 2000).

Materials and methods

Subjects

Blood samples were obtained with written consent from 20 university students and used in the screening stage of the study. Anonymous blood samples were also collected from 150 unrelated blood donors from the Hong Kong Red Cross Blood Transfusion Service and used to establish SNP allele frequencies for the local Chinese population. DNA was extracted with a modified salting-out method (Miller et al. 1988). The study was approved by the University's Human Subject Ethics Subcommittee.

PCR amplification

Twenty-four primer pairs (Table 1) were designed with the software Oligo (version 6.57; Molecular Biology Insights, Cascade, USA) to amplify the six *RDH8* exons and their immediate flanking regions, and non-coding sequences about 2.5 kb upstream of the start codon and 2.0 kb downstream of the stop codon. NT_011295 was used as the reference sequence. The size of the PCR fragments ranged between 220 and 439 basepairs (bp). In the screening stage, four DNA pools were constructed, with each consisting of equal amounts from five individuals. If at least one heterozygous sample was present in a pool, then the relative frequency of the less frequent allele in the pool would be at least 10% and the DHPLC elution profile would show multiple peaks (Wolford et al. 2000). For any pool showing multiple peaks, its constituent samples were amplified and analysed individually by DHPLC to identify homozygous and heterozygous samples. In the genotyping stage, 150 samples were individually amplified before DHPLC analysis.

The 50- μ l reaction volume contained 1 $\times$ Gold Buffer (15 mM Tris-HCl, 50 mM KCl, pH 8.0), 0.2 mM of each dNTP, 0.3 μ M of each primer, 1U AmpliTaq Gold DNA polymerase (Applied Biosystems, Foster City, USA) and 100 ng genomic DNA (pooled or individual). The MgCl₂ concentration was 1.5 mM for the amplification of all fragments except four, namely, Rdh831, Rdh833, Rdh836 and Rdh837, for which it was 2.5 mM. Touch-down PCR was used to avoid excessive optimisation of the reaction conditions. PCR amplification consisted of initial denaturation for 5 min at 95°C, eight touch-down cycles, 35 main cycles and final extension for 7 min at 72°C. Both touch-down and main cycles consisted of 30 s at 95°C, 30 s at the annealing temperature and 45 s at 72°C. The annealing temperature of the main cycles for each primer pair is shown in Table 1, and the initial annealing temperature for the touch-down cycles was 7°C above this with 1°C reduction for each successive touch-down cycle.

DHPLC analysis

In the screening stage, PCR products were denatured for 5 min at 95°C and then cooled to 25°C at a rate of 1°C/min to allow for heteroduplex formation. Denatured products were analysed with the WAVE DNA Fragment Analysis System (Transgenomic) using a DNasep column kept at a certain temperature in an oven. Next, 5 μ l aliquots of denatured products were automatically injected into the column and eluted with a linear acetonitrile (ACN) gradient in 0.1 M triethylammonium acetate (TEAA) buffer, pH 7.4, at a constant flow rate of 0.9 ml/min. The gradient was composed of a range of proportions (percent) of buffer B increased at a rate of 2%/min and created by mixing buffers A and B, where buffer A was 0.1 M TEAA with 0.025% ACN and buffer B 25% ACN in 0.1 M TEAA. The gradient was calculated automatically by the software WAVEMAKER (version 4.1, Transgenomic) on the basis of the DNA sequence and hence was specific to each PCR product (Table 1). The column was cleaned with 75% ACN after each analytical run. The eluate was monitored by an ultraviolet light detector at 260 nm. The WAVEMAKER software also predicted the melting domains of input DNA fragment sequence. Column temperatures for screening of each fragment were selected on the basis of its melting domains whereby the domains were 75–90% in helical conformation at the given temperatures. Consequently, the PCR fragments were each analysed at three to seven column temperatures (Table 1).

For each identified and characterised SNP, 150 Chinese samples were genotyped to establish the allele frequencies. The samples were first analysed individually without mixing with a reference sample and using the same DHPLC conditions as above, but only at a single column temperature at which the multiple-peak elution profile was most distinctive (Table 1). Known homozygous and heterozygous samples were run in parallel as controls. Samples showing a single peak (homozygotes) were then mixed with a reference homozygous sample in equal volumes and re-analysed again using the same conditions. With this second analysis, the sample had the same genotype as the control homozygote if a single peak was observed, and was homozygous for the other allele if multiple peaks were seen. Similarly, samples showing multiple peaks (heterozygotes) at the first analysis were also mixed with a reference heterozygous sample and re-analysed. Identical multiple peaks as in the first analysis confirmed that the sample had the same heterozygous genotype as the control heterozygote. A distorted multiple-peak elution profile indicated the presence of additional SNP in the fragment and would be further investigated.

DNA Sequencing

The DNA sequences of representative samples showing distinct elution patterns were determined by direct sequencing. After amplification as described above, the PCR products were purified with shrimp alkaline phosphatase (Amersham Biosciences, Piscataway, USA) and exonuclease I (New England Biolabs, Beverly, USA). Cycle sequencing was performed on the purified products using the Big Dye Terminator Cycle Sequencing Ready Reaction Kit (Version 2.0 or 1.1, Applied Biosystems). Forward and/or reverse primers were used as sequencing primers as appropriate. The sequencing products were purified by ethanol precipitation and analysed in ABI PRISM 310 Genetic Analyzer (Applied Biosystems) according to the manufacturer's instructions.

Statistical analysis

Allele frequencies were calculated using the simple gene counting method. Testing for Hardy-Weinberg equilibrium was performed using chi-square test. LD pattern and haplotype frequencies were also established for common SNPs with minor allele frequency >0.05 in the Chinese population under study. Allelic association and hence LD between a pair of SNPs were determined with the

Table 1 Polymerase chain reaction (PCR) fragments, primer sequences and annealing temperatures (Ta), and conditions of denaturing high-performance liquid chromatography (DHPLC) analysis

Fragment (size)	Primer Sequence (5' > 3')		Reverse primer	Ta (°C)	DHPLC analysis	
	Forward primer				Column temperature (°C) ^a	% buffer B
Rdh858 (423 bp) ^b	CAG TCG GGG CTT TTT GGA G		TGG CTT GAA CGC AGG AGT TG	57	53, [59], 60, 61	57 → 66
Rdh857 (390 bp)	GTT TGG AGA CAG GGT CTT GA		AGC TGA GGG ACT GAG GAT AA	55	55, 59, 60, 61, 62	56 → 65
Rdh856 (389 bp) ^b	CAC CCC GAA CTC TCA CTC A		CCA GCT CAG ACT CCC TCA TT	57	58, 59, [60]	56 → 65
Rdh855 (405 bp) ^b	CCA ACT CAG GTC TCC AGT G		CTG CGA CAG GGA CAG ATT C	57	58, 59, 60, [61], 62	57 → 66
Rdh854 (392 bp)	CTG AGT TTG TGC CAC TGC GA		ACT GCT TGC GTC TCC CTC TT	57	60, 61, 62, 63	56 → 65
Rdh853 (406 bp)	TAT TCT GAG CCC TGA ACT GC		AGA GAG TGA GAG AAG GAG ACA	55	59, 60, 61, 62	57 → 66
Rdh852 (388 bp)	CCC TCT ATG AAA CAC TTC TGG A		GGT TCA ATT TCG GCT CAC TAC	57	57, 58, 59, 60	56 → 65
Rdh851 (439 bp) ^{b, c}	GAG GCA GGA GAA TCA TTT TAC C		TCC GCA CCT ACT ATA GAC ATC A	57	51, 55, 56, 57, [58]	57 → 66
Rdh8e11 (225 bp)	GCC TGA GTG ACA GAG CAA G		CTA ACC CCC TCC CAC AGA T	56	57, 58, 59, 60, 63	51 → 60
Rdh8e12 (301 bp) ^b	CTG TTC AGG GTC AGT CCA GAA		ATG CTG CCT CCT AGT GCC A	56	61, 62, [63], 64	54 → 63
Rdh8e2 (223 bp)	GTC TCT GCT TTA CTC TCT GC		GAT GGG GCT TAG TGA ATC AG	55	61, 63, 64, 65	51 → 60
Rdh8e3 (275 bp)	GCT AGA AGA GTG AGG GAC TTT A		ATC AGA CAC TGG TGG GAT TTC	55	56, 57, 60, 62, 63	53 → 62
Rdh8e4 (220 bp)	CCT CCC AGA CCC TCA AAG A		GGC TTC ATC TCT CCT CTC C	56	60, 61, 62, 63	51 → 60
Rdh8e5 (273 bp) ^b	GCA TCA GAC TTA CAC TAC CCA		ATA ACC CCG CAA CAT ACC TG	55	61, 62, [63]	53 → 62
Rdh8e61 (349 bp) ^{b, c}	AAG ACG TGG CTC AGG TAT GTT		CTA GAT GTG GAA TGA AGA GGT C	55	60, 64, [65]	55 → 64
Rdh8e62 (338 bp) ^b	ACT ATC CCC ATC CCT GAA CAA		AGC ATG AAG TCA CGA GGT TGA A	55	57, [58], 60, 61	55 → 64
Rdh8e63 (370 bp)	TGG AGG AAG CGG CAT TGT TAT		GGC TTG CTT GCT TGC TTT CT	55	52, 53, 55, 57, 59, 60	56 → 65
Rdh831 (425 bp)	CGT GAA AAG CCA GTT GAA CTA C		GCC AGA CTC CAT CTC AAA GA	59	52, 53, 54, 57, 59, 60	57 → 66
Rdh832 (325 bp)	CTG GGC AAC AGA ATG AGA CTT		TGG ACA GAT CAC AAG GTC AG	56	54, 55, 57	55 → 64
Rdh833 (381 bp)	GTT AGC CAG CAT GGT CTT GAT		GCC TTA GGT TCA TTC ACA CTG	59	56, 57, 58, 60	56 → 65
Rdh834 (324 bp)	ATG GAT AGC GGT GAT GGT TG		GAT GGT CTC AAA CTC CTG ACT T	57	55, 56, 57, 59, 60, 61, 62	55 → 64
Rdh835 (379 bp)	CTA TCA ATT CTC CTG CGT CAC		GAT GGT CTC AAA CTC CTG ACT T	59	55, 56, 57	56 → 65
Rdh836 (401 bp) ^{b, c}	GCT TAA TAT AGA CCA GGC ACA G		CCG AGG AAA CAA CTA AAG TTC AC	59	52, 54, 57, [59], 60	57 → 66
Rdh837 (285 bp)	GCC CTA AAA GCA GCA TAT ATC		GAC TAA CAC GGT GAA ACC CT	59	53, 54, 56	54 → 63

^a Selection of column temperatures was based on the melting domains present in the PCR fragments (see text for details). The column temperature recommended by the software WAVEMAKER is shown in *italics*, while the temperature at which a single nucleotide polymorphism (SNP) was identified is placed within square brackets

^b SNPs were identified in these nine PCR fragments

^c For these three PCR fragments, the SNPs were each identified at a column temperature that was different from that recommended by the software WAVEMAKER. These column temperatures are shown in boldface for easy recognition

expectation-maximisation algorithm (Excoffier and Slatkin 1995) using the ASSOCIATE programme (Ott 1985), as described by Yip et al. (2003). Briefly, the output of ASSOCIATE allowed the determination of the Lewontin's LD parameter (D) and its maximum value ($D_{\max}$), and hence the calculation of the standardised Lewontin's LD parameter ($D' = D/D_{\max}$). ASSOCIATE also gave the p value for the statistical significance of allelic association and hence LD. In order to obtain an overall significance level (α') of 0.05, Bonferroni procedure was applied to correct for multiple testing (Weir 1996). There were 21 pairwise LD values. Thus, the significance level (α) for each individual testing was obtained from $\alpha = 1 - (1 - \alpha')^{1/21}$ and found to be 0.0024. In addition, another commonly used LD measure (r^2) was also calculated; $r^2 = D^2/[P_A P_B (1 - P_A) (1 - P_B)]$ where the P_A and P_B are the frequencies of allele A and allele B at two different loci. Finally, frequencies of haplotypes across all seven common SNP sites were estimated with the EH programme (Terwilliger and Ott 1994).

Results

Twenty-four PCR fragments (Table 1) were amplified to cover the 5' flanking region, six exons and their immediate flanking intron sequences, and the 3' flanking region of the *RDH8* locus, which amounted to a total genomic distance of about 7 kb. DHPLC analysis found 15 SNPs in nine of these fragments. Most fragments harboured one or two SNPs, but one fragment (Rdh858) contained three SNPs (Tables 1 and 2). It is interesting to note that five of these 15 SNPs were identified not during the screening stage but during the genotyping stage when 150 samples were genotyped by DHPLC

(Table 2). This was not unexpected in view of the much larger sample size for genotyping purpose. On the other hand, two PCR fragments showing potential heteroduplex peaks on initial DHPLC analysis of the DNA pools were finally confirmed not to contain any SNP on repeated DHPLC analysis and direct DNA sequencing of individual samples of the positive pools. This gave a false positive rate of 18% for DHPLC analysis for screening sequence variations.

Characteristics of identified SNPs

Table 2 shows the details of the SNPs identified in this study. Of the 15 identified, 12 were transitions (seven A/G and five C/T transitions), two transversions, and one was a two-base insertion/deletion. Only three were located in coding regions, while the rest were in non-coding regions with nine found upstream of the start codon and three downstream of the stop codon. One coding SNP was a synonymous transition (8117C > T or RDH8E61) in exon 6. The other two coding SNPs were adjacent to each other in the same codon (202) located in exon 5 (7826T > C and 7827G > A, or RDH8E5a and RDH8E5b, respectively). Among the 150 samples genotyped, only three of the four possible haplotypes were found for these two adjacent SNPs: ATG (methionine), ACG (threonine) and ACA (threonine). The ACA haplotype was found only in four individuals

Table 2 Details of the single nucleotide polymorphisms (SNPs) identified in the all-*trans*-retinol dehydrogenase (*RDH8*) gene in Hong Kong Chinese population ($n = 150$)

SNP ^a	Designation ^a	Location	Reference SNP ID no. (NCBI)	Flanking sequences (5' > 3') and alleles [major/minor] ^b	Minor allele frequency
-2881G > C	RDH858a	5' flanking region	Novel	TTGTGCTGAG [G/C] GCTGATTGGC	0.0033
-2720C > G	RDH858b	5' flanking region	Novel	CAAAGGAGAC [C/G] TTAACCTCAT	0.0033
-2710T > C ^c	RDH858c	5' flanking region	Novel	CTTAACCTCA [T/C] TCAAGACTTC	0.0033
-2076G > A	RDH856	5' flanking region	Novel	CCTAAGTCAC [G/A] TGTCCTGGTC	0.0033
-1799A > G ^c	RDH855a	5' flanking region	Novel	TCAGCTGGCC [A/G] ACATGGCGAA	0.0067
-1715G > A	RDH855b	5' flanking region	Novel	CTACTTGGGA [G/A] GCTGAGGTGG	0.1633
-472C > T	RDH851	5' flanking region	rs2233789	GCCAAGAGTT [C/T] GAGACCAATC	0.2667
-130G > A ^c	RDH8E12a	Exon 1, 5'UTR ^d	Novel	AGGGTCTCCG [G/A] GGGAGGGCAG	0.0033
-126A > G	RDH8E12b	Exon 1, 5'UTR	Novel	TCTCCGGGGG [A/G] GGGCAGAGGC	0.0533
7826T > C^c	RDH8E5a	Exon 5	rs1644731	CAGGTTTCTA [C/T] GGCTGAGTTC	0.4367
7827G > A ^{c, e}	RDH8E5b	Exon 5	Novel	AGGTTTCTAT [G/A] GCTGAGTTCC	0.0133
8117C > T^f	RDH8E61	Exon 6	rs747574	TGACCACGCT [C/T] AAAACCGTGG	0.4033
8344C > T ^c	RDH8E62a	Exon 6, 3'UTR	Novel	TTCATTATT [C/T] ATTCTGCAAA	0.0067
8566-8567delGA	RDH8E62b	Exon 6, 3'UTR	rs3217240	CTTGAAGGAA [-/GA] GACAGACTCT	0.4067
10254G > A	RDH836	3' flanking region	rs1644727	AGTGGCTCAC [G/A] TCTGTAATTC	0.4000

^a The SNPs are named in accordance with the recommended nomenclature systems (Nebert 2002; den Dunnen and Antonarakis 2000; Antonarakis and the Nomenclature Working Group 1998), while alternative designations (based on the names of the fragments) are also given for the sake of easy discussion in this article. NT_011295 is used as the reference sequence. Note the SNPs shown in boldface are included in linkage disequilibrium (LD) and haplotype analysis

^b The 5' and 3' flanking regions, each 10 bases long, are shown before and after the SNP, respectively, with the alleles of the SNP within square brackets. The major allele refers to the more frequent allele and the minor allele the less frequent allele

^c These five SNPs were identified only during the stage of genotyping population samples

^d UTR = untranslated region

^e These two SNPs are located at positions 2 and 3 of the same codon (202). If heterozygous, most individuals are heterozygous only for 7826T > C and the amino acid is changed from methionine (ATG) to threonine (ACG). Only four individuals in our sample set are heterozygous for both sites and were shown by single strand conformation polymorphism analysis (data not shown) to have the haplotypes of ATG (methionine) and ACA (threonine)

^f 8117C > T causes the synonymous transition L267L

heterozygous for both SNPs, who, in other words, carried the two-site genotype *ATG/ACA*. This was determined by single-strand conformation polymorphism analysis (data not shown).

Ten of the 15 SNPs were novel and not previously reported. Seven SNPs (two novel and five previously documented in NCBI with reference SNP ID numbers) were common with the minor allele frequency >0.05. The observed genotype frequencies were in Hardy-Weinberg proportions for all SNPs.

Three common SNPs (RDH851, RDH8E61 and RDH836) were of particular interest, because they might have been missed if DHPLC analysis had been performed only at the single column temperatures recommended by the software WAVEMAKER (Tables 1 and 2). They showed distinct heteroduplex peaks at temperatures 2, 1 and 5°C, respectively, above those recommended. These findings testify the usefulness of our approach to selecting multiple column temperatures on the basis of the melting domains, whereby the domains are 75–90% in helical conformation at the given temperatures.

Analysis of LD and haplotypes

LD pattern and haplotype frequencies were established for the seven common SNPs. Table 3 shows the pairwise LD measures (both $|D'|$ and r^2) and the corresponding statistical significance for allelic association. Strong to complete LD was observed across four SNPs at the 3' end of the *RDH8* locus, namely, RDH8E5a, RDH8E61, RDH8E62b and RDH836. The pairwise LD measures were >0.75 for $|D'|$ and >0.42 for r^2 , with all p values for allelic association <0.0001. While this LD block was obvious and certain, LD among other SNP pairs was either questionable (with inconsistent LD measures) or absent. In particular, the $|D'|$ values showed significant ($p < 0.0024$) LD between RDH851 on one hand and three other SNPs (RDH855b, RDH8E5a and RDH836) on the other hand. However, the r^2 values were all <0.09 for these three SNP pairs. Similar situations occurred for three other SNP pairs (RDH855b-RDH8E61, RDH851-RDH8E12b, and RDH8E12b-

RDH836) that had $|D'| \geq 0.39$, but $r^2 < 0.05$ with p values significant at the 0.05 level but not at the Bonferroni-adjusted level (0.0024).

Table 4 shows the seven-site haplotypes together with the corresponding frequencies estimated by the EH programme. Although there are theoretically a total of 128 possible haplotypes for seven bi-allelic SNPs, six major seven-site haplotypes gave a total cumulative frequency of ~0.75. The cumulative frequency added up to ~0.95 with ten additional seven-site haplotypes.

Discussion

We are interested in mapping susceptibility genes predisposing humans to complex eye diseases including myopia. One major mapping approach is association studies based on either unrelated cases and controls from the same population (population based) or families with at least one affected sib (family based). Both types of association studies require identification and characterisation of suitable SNPs within and around potential candidate genes followed by testing with appropriate sample sets. Discovery and characterisation of SNPs in candidate genes form part of our concerted mapping effort based on association studies. This study reports the SNPs in the *RDH8* locus and their LD pattern. The gene product is one of the proteins involved in the metabolism of retinoic acid, which may play a role as a mediator between refractive error and compensatory eye growth (Seko et al. 1998; Mertz and Wallman 2000).

SNP screening based on DNA pooling and DHPLC

DHPLC analysis at a partially denaturing temperature separates homoduplex and heteroduplex DNA fragments very efficiently if the injected sample contains more than one allele. Conventionally, a test sample is mixed with a reference sample of wild-type homozygous genotype before injection to allow detection of sequence variations (SNPs or pathological mutations). If the test sample is heterozygous, then the variant allele frequency in the mixed sample is 25% (one variant allele in a pool

Table 3 Pairwise linkage disequilibrium (LD) measures for common single nucleotide polymorphisms (SNPs) of the all-*trans*-retinol dehydrogenase (*RDH8*) locus and the corresponding statistical significance for allelic association^a

	RDH855b	RDH851	RDH8E12b	RDH8E5a	RDH8E61	RDH8E62b	RDH836
RDH855b		1.00 (0.07)	0.64 (0.00)	0.40 (0.02)	0.39 (0.05)	0.26 (0.01)	0.31 (0.01)
RDH851	0.0003		1.00 (0.02)	0.41 (0.08)	0.31 (0.02)	0.17 (0.01)	0.40 (0.09)
RDH8E12b	0.4736	0.0136		0.00 (0.00)	0.26 (0.01)	0.19 (0.00)	0.77 (0.02)
RDH8E5a	0.0829	0.0010	0.9904		1.00 (0.52)	0.88 (0.68)	0.94 (0.76)
RDH8E61	0.0220	0.0778	0.3962	<0.0001		1.00 (0.46)	0.97 (0.43)
RDH8E62b	0.3151	0.1379	0.4816	<0.0001	<0.0001		0.76 (0.56)
RDH836	0.1966	0.0002	0.0199	<0.0001	<0.0001	<0.0001	

^a Above the diagonal are the values of $|D'|$ and r^2 (in parenthesis), and below the diagonal are the p values for allelic association (from the ASSOCIATE programme). Note that the Bonferroni-adjusted significance level for each individual test is 0.0024 (see "Materials and methods" for details)

Table 4 Frequencies of haplotypes across seven common single nucleotide polymorphisms (SNPs) of the All-*trans*-retinol dehydrogenase (*RDH8*) locus^a

SNPs							Haplotype frequency	Cumulative frequency
RDH855b	RDH851	RDH8E12b	RDH8E5a	RDH8E61	RDH8E62b	RDH836		
1 ^b	1 ^b	1 ^b	2 ^d	2 ^d	2 ^d	2 ^d	0.2065	0.2065
1 ^b	1 ^b	1 ^b	1 ^e	1 ^e	1 ^e	1 ^e	0.1569	0.3634
1 ^c	2 ^c	1 ^c	1 ^e	1 ^e	1 ^e	1 ^e	0.1322	0.4956
2	1	1	2 ^d	2 ^d	2 ^d	2 ^d	0.0995	0.5951
1 ^b	1 ^b	1 ^b	2	1	2	2	0.0923	0.6874
1 ^c	2 ^c	1 ^c	2 ^d	2 ^d	2 ^d	2 ^d	0.0664	0.7538
1 ^c	2 ^c	1 ^c	1	1	2	1	0.0308	0.7846
2	1	1	1 ^e	1 ^e	1 ^e	1 ^e	0.0273	0.8119
1 ^b	1 ^b	1 ^b	1	1	1	2	0.0243	0.8362
1 ^c	2 ^c	1 ^c	2	1	2	2	0.0240	0.8602
1	1	2	1	1	1	2	0.0169	0.8771
1 ^b	1 ^b	1 ^b	1	1	2	1	0.0164	0.8935
2	1	1	2	1	2	2	0.0162	0.9097
1 ^b	1 ^b	1 ^b	2	1	1	1	0.0125	0.9222
1 ^b	1 ^b	1 ^b	1	1	2	2	0.0109	0.9331
1	1	2	2 ^d	2 ^d	2 ^d	2 ^d	0.0102	0.9433
G	C	A	C	C	—	G	1 (= major allele)	
A	T	G	T	T	GA	A	2 (= minor allele)	

^a For the sake of clarity, the major allele of each SNP is denoted as 1 and the minor allele as 2 (see the bottom of the table). The haplotypes are listed in the order of decreasing frequency

^{b, c} The two most common haplotypes for the three SNPs at the 5' end
^{d, e} The two most common haplotypes for the 4 SNPs at the 3' end
 Note that only haplotypes with frequency > 0.01 are shown

of four alleles). It was shown previously that DHPLC analysis could detect one variant allele (5%) in a pool of 20 alleles (Wolford et al. 2000). Consequently, we increased the screening throughput five-fold by using DNA pools composed of samples from five individuals instead of mixing every sample with a reference. In the DNA pool, every constituent sample forms a reference for all other samples in the pool. If at least one of the five constituent samples is heterozygous for an SNP, then the variant allele frequency is at least 10% (one variant allele in a pool of ten alleles). This can easily be detected by DHPLC. In addition, our pooling strategy still had a wide safety margin. DNA pooling has proved very efficient in SNP discovery when coupled with DHPLC analysis.

The same study also reported that fluorescent cycle sequencing could detect variants in the same pools if the frequency of the less frequent allele was at least 10% (Wolford et al. 2000). However, we found this approach very unreliable, and heterozygous peaks very often could not be detected in the sequencing traces for the DNA pools showing heteroduplex signals in DHPLC. As a result, we adopted the approach that individual samples of the positive pools were re-amplified and re-analysed by DHPLC before individual representative samples were sequenced to characterise the SNPs.

The present study highlighted the inadequacy of the WAVEMAKER software in predicting the column temperature to be used for DHPLC analysis of a particular PCR fragment (Table 1). There are recommendations to conduct the analysis also at temperatures 2°C below and above the predicted temperature (Colosimo et al. 2002). Even with this recommendation, the SNP RDH836 (or 10254G > A) would still be missed, since it

was detected at a temperature 5°C above that recommended by WAVEMAKER. This justifies the selection of multiple temperatures at which the melting domains are 75–90% in helical conformation.

Interestingly, DHPLC failed to detect at all selected temperatures an SNP found by direct sequencing in one particular sample. This SNP was located in a GC-rich domain and was analysed at a temperature best for this domain. At this particular column temperature, all other domains melted completely and the elution peak appeared very early in the chromatogram. On the other hand, cycle sequencing failed to confirm convincingly in heterozygous samples an SNP detected by DHPLC. This particular SNP could only be confirmed by cycle sequencing of the two types of homozygous samples.

SNP genotyping by DHPLC

In genotyping SNPs for population samples with DHPLC analysis, all samples were first analysed without mixing with a reference sample. This allowed the detection of homozygotes (of either type) and heterozygotes. A second analysis required mixing with a reference sample (either a homozygous or a heterozygous sample, depending on the initial results) allowed accurate genotyping of the SNPs concerned and simultaneous detection of any additional novel SNPs present in the same PCR fragment. A similar, but not identical, DHPLC strategy was also recently reported for genotyping purpose (Schaeffeler et al. 2001). In our study, the second analysis additionally required the heterozygous sample to be mixed with a reference heterozygous sample and then re-analysed. This extra step would allow

additional SNPs within the same fragment to be identified if this additional SNP did not reveal itself in the first analysis. This approach has proved very successful in our hands.

SNP genotyping based on direct DHPLC analysis requires two injections per sample. But this approach does offer many advantages. There is no need to re-design primers, re-optimize PCR, and use other new reagents (like restriction enzymes). Sample injection is automated and throughput is high. Reproducibility is very high. Furthermore, this approach allows identification of new SNPs within the same fragment (Table 2), because DHPLC is used as both a screening and a diagnostic method in this setting.

Analysis of LD and haplotypes

D' and r^2 are two most commonly used LD measures (Ardlie et al. 2002; Weiss and Clark 2002). In some ways, r^2 is complementary to D' but is increasingly recommended as the measure of choice for quantifying and comparing LD in the context of mapping genes by association studies. In addition, r^2 is less affected by sample size and allele frequency. Both measures were estimated for seven common SNPs identified in the *RDH8* locus in this study (Table 3). As expected, the r^2 values were lower than the $|D'|$ values. The $|D'|$ measure reached its maximum value of 1 for some SNP pairs, but the r^2 measure was at most 0.76 (for the SNP pair RDH8E5a-RDH836) in this study. The case of $|D'| = 1$ (complete LD) occurs if, and only if, two SNPs have not been separated by recombination (Ardlie et al. 2002). In this case, at most, three out of the four possible two-site haplotypes are observed in the sample. On the other hand, the case of $r^2 = 1$ (perfect LD) occurs if, and only if, two SNPs have not been separated by recombination and, in addition, have the same allele frequency (Ardlie et al. 2002). In this case, exactly two out of the four possible two-site haplotypes are observed in the sample. From the practical viewpoint of mapping genes using association studies, an r^2 value of at least 1/3 is the recommended minimum, because this requires an increase of sample size by less than three-fold in order to compensate for the weak LD between an SNP marker and a susceptibility locus.

Both LD measures strongly supports the notion that the four common SNPs (RDH8E5a, RDH8E61, RDH8E62b and RDH836) at the 3' end of the *RDH8* locus forms a LD block with $|D'| > 0.75$ and $r^2 > 0.42$ (Table 3). This conclusion is also supported by haplotype analysis for these four SNPs, which spans a genomic distance of about 2.5 kb. For the sake of clarity and easy discussion, we denote the major allele of each SNP as 1 and the minor allele as 2. As estimated by EH on the basis of the observed frequencies of genotypes in the Chinese population under study, the *observed* four-site haplotype frequency across these four SNPs (in the order listed above) was 0.3164 for 1-1-1-1, 0.3826 for 2-2-2-2 and 0.1325 for 2-1-2-2, with a total observed

frequency of 0.8315 (Table 4). The *expected* haplotype frequencies could be obtained by multiplying the corresponding allele frequencies and were found to be 0.1197, 0.0287 and 0.0424, respectively, with an expected total frequency of only 0.1908. The large differences between the observed and the expected haplotype frequencies testify the strong LD among these four SNPs, which can be regarded as forming a haplotype block.

That three common SNPs (RDH855b, RDH851 and RDH8E12b) at the 5' end of the locus does not exhibit significant useful LD is also supported by a similar analysis of the haplotype frequencies. As estimated by EH, the observed three-site haplotype frequency was 0.5198 for 1-1-1, 0.2534 for 1-2-1 and 0.1430 for 2-1-1, with a total observed haplotype frequency of 0.9162 (Table 4). The expected haplotype frequencies were 0.5808, 0.2113 and 0.1134, respectively, with an expected total frequency of 0.9055; all were not much different from the observed frequencies. In other words, these three SNPs were in linkage equilibrium, which is not expected on the basis of short genomic distance (~ 1.6 kb) spanning by these SNPs.

In other words, LD pattern established by pairwise LD measures is supported by multiple-site haplotype analysis. Our results indicate that two sets of markers are required for association studies involving the *RDH8* gene. One SNP is required in the 3' haplotype block. Of the four SNPs in this haplotype block, RDH8E5a is best suited for this purpose because it involves a non-synonymous change (Met/Thr) and might have functional consequences. Since strong and useful LD does not exist among the common SNPs at the 5' region of the gene, two to three SNPs are needed.

Currently, international concerted efforts are being made to produce a haplotype map for the human genome, which serves to facilitate LD mapping for complex human traits or diseases through association studies. This study reports the SNPs within and around the *RDH8* gene and the corresponding LD pattern. This serves to provide the groundwork for association studies mapping genes involved in complex eye diseases.

Acknowledgements The SNP project for myopia candidate genes was supported by an Internal Competitive Research Grant (A-PC13), The Hong Kong Polytechnic University (PolyU). Wei Han was supported by a PolyU competitive grant (G-T538) for Joint Supervision of Ph.D. students with Zhejiang University. The Wave DNA Fragment Analysis System was purchased with a PolyU Big Equipment Grant (G.53.27.9027), and Genetic Analyzer 310 with funds from three sources: a departmental fund (School of Nursing, PolyU); Faculty Dean's Reserve (1.53.27.8721; Faculty of Health and Social Sciences, PolyU); and an Area for Strategic Development Fund for Myopia Research (A362, Department of Optometry and Radiography, PolyU).

References

- Antonarakis SE, The Nomenclature Working Group (1998) Recommendations for a nomenclature system for human gene mutations. *Hum Mutat* 11:1-3

- Ardlie KG, Kruglyak L, Seielstad M (2002) Patterns of linkage disequilibrium in the human genome. *Nat Rev Genet* 3:299–309
- Colosimo A, Guida V, De Luca A, Cappabianca M P, Bianco I, Palka G, Dallapiccola B (2002) Reliability of DHPLC in mutational screening of β -globin (HBB) alleles. *Hum Mutat* 19:287–295
- den Dunnen JT, Antonarakis SE (2000) Mutation nomenclature extensions and suggestions to describe complex mutations: a review. *Hum Mutat* 15:7–12
- Excoffier L, Slatkin M (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol* 12:921–927
- Goh WSH, Lam CSY (1994) Changes in refractive trends and optical components of Hong Kong Chinese aged 19–39 years. *Ophthalmic Physiol Opt* 14:378–382
- Johnson GCL, Todd JA (2000) Strategies in complex disease mapping. *Curr Opin Genet Dev* 10:330–334
- Lander ES, Schork NJ (1994) Genetic dissection of complex traits. *Science* 265:2037–2048
- Mertz JR, Wallman J (2000) Choroidal retinoid acid synthesis: a possible mediator between refractive error and compensatory eye growth. *Exp Eye Res* 70:519–527
- Miller SA, Dykes DD, Polesky HF (1988) A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res* 16:1215
- Nebert DW (2002) Proposal for an allele nomenclature system based on the evolutionary divergence of haplotypes. *Hum Mutat* 20:463–472
- Ott J (1985) A Chi-square test to distinguish allelic association from other causes of phenotypic association between two loci. *Genet Epidemiol* 2:79–84
- Rando RR (2001) The biochemistry of the visual cycle. *Chem Rev* 101:1881–1896
- Rattner A, Smallwood PM, Nathans J (2000) Identification and characterization of all-trans-retinol dehydrogenase from photoreceptor outer segments, the visual cycle enzyme that reduces all-trans-retinal to all-trans-retinol. *J Biol Chem* 275:11034–11043
- Saari JC, Garwin GG, Van Hooser JP, Palczewski K (1998) Reduction of all-trans-retinal limits regeneration of visual pigment in mice. *Vision Research* 38:1325–1333
- Saw SM, Katz J, Schein OD, Chew SJ, Chan TK (1996) Epidemiology of myopia. *Epidemiol Rev* 18:175–187
- Schaeffeler E, Lang T, Zanger UM, Eichelbaum M, Schwab M (2001). High-throughput genotyping of thiopurine S-methyltransferase by denaturing HPLC. *Clin Chem* 47:548–555
- Seko Y, Shimizu M, Tokoro T. (1998) Retinoic acid increases in the retina of the chick with form deprivation myopia. *Ophthalmic Res* 30:361–367
- Terwilliger JD, Ott J (1994) Handbook of human genetic linkage. Johns Hopkins University Press, Baltimore
- Weir BS (1996) Genetic data analysis II: methods for discrete population genetic data. Sinauer Associates, Sunderland, Massachusetts
- Weiss KM, Clark AG (2002) Linkage disequilibrium and the mapping of complex human traits. *Trends Genet* 18:19–24
- Wolford JK, Blunt D, Ballecer C, Prochazka M. (2000) High-throughput SNP detection by using DNA pooling and denaturing high performance liquid chromatography(DHPLC). *Hum Genet* 107:483–487
- Wong TY, Foster PJ, Hee J, Ng TP, Tielsch JM, Chew SJ, Johnson GJ, Seah SK (2000) Prevalence and risk factors for refractive errors in adult Chinese in Singapore. *Invest Ophthalmol Vis Sci* 41:2486–2494
- Xiao W, Oefner PJ (2001) Denaturing high-performance liquid chromatography: a review. *Hum Mutat* 17:439–474
- Yip SP, Leung KH, Lin CK (2003) Extent and distribution of linkage disequilibrium around the SLC11A1 locus. *Genes Immun* 4:212–221