31. Lam DS, Lee WS, Leung YF, et al. TGFbeta-induced factor: a candidate gene for high myopia. *Invest Ophthalmol Vis Sci.* 2003; 44:1012–1015.

32. Sundin OH, Leppert GS, Silva ED, et al. Extreme hyperopia is the result of null mutations in MFRP, which encodes a Frizzled-related protein. *Proc Natl Acad Sci U S A.* 2005;102:9553–9558.

33. Han W, Yap MK, Wang J, Yip SP. Family-based association analysis of hepatocyte growth factor (HGF) gene polymorphisms in high myopia. *Invest Ophthalmol Vis Sci.* 2006;47:2291–2299.

34. Lin HJ, Wan L, Tsai Y, et al. The TGFbeta1 gene codon 10 polymorphism contributes to the genetic predisposition to high myopia. *Mol Vis.* 2006;12:698–703.

35. Wang IJ, Chiang TH, Shih YF, et al. The association of single nucleotide polymorphisms in the 5'-regulatory region of the lumican gene with susceptibility to high myopia in Taiwan. *Mol Vis.* 2006;12:852–857.

36. Inamori Y, Ota M, Inoko H, et al. The COL1A1 gene and high myopia susceptibility in Japanese. *Hum Genet.* 2007;122:151–157.

37. Majava M, Bishop PN, Hagg P, et al. Novel mutations in the small leucine-rich repeat protein/proteoglycan (SLRP) genes in high myopia. *Hum Mutat.* 2007;28:336–344.

38. Mutti DO, Cooper ME, O'Brien S, et al. Candidate gene and locus analysis of myopia. *Mol Vis.* 2007;13:1012–1019.

39. Tsai YY, Chiang CC, Lin HJ, Lin JM, Wan L, Tsai FJ. A PAX6 gene polymorphism is associated with genetic predisposition to extreme myopia. *Eye.* 2008;22:576–581.

40. Paluru PC, Scavello GS, Ganter WR, Young TL. Exclusion of lumican and fibromodulin as candidate genes in MYP3 linked high grade myopia. *Mol Vis.* 2004;10:917–922.

41. Scavello GS, Paluru PC, Ganter WR, Young TL. Sequence variants in the transforming growth beta-induced factor (TGIF) gene are not associated with high myopia. *Invest Ophthalmol Vis Sci.* 2004; 45:2091–2097.

42. Scavello GS, Jr., Paluru PC, Zhou J, White PS, Rappaport EF, Young TL. Genomic structure and organization of the high grade Myopia-2 locus (MYP2) critical region: mutation screening of 9 positional candidate genes. *Mol Vis.* 2005;11:97–110.

43. Hasumi Y, Inoko H, Mano S, et al. Analysis of single nucleotide polymorphisms at 13 loci within the transforming growth factor-induced factor gene shows no association with high myopia in Japanese subjects. *Immunogenetics.* 2006;58:947–953.

44. Liang CL, Hung KS, Tsai YY, Chang W, Wang HS, Juo SH. Systematic assessment of the tagging polymorphisms of the COL1A1 gene for high myopia. *J Hum Genet.* 2007;52:374–377.

45. Simpson CL, Hysi P, Bhattacharya SS, et al. The Roles of PAX6 and SOX2 in Myopia: lessons from the 1958 British Birth Cohort. *Invest Ophthalmol Vis Sci.* 2007;48:4421–4425.

46. Metlapally R, Li YJ, Tran-Viet KN, et al. Common MFRP sequence variants are not associated with moderate to high hyperopia, isolated microphthalmia, and high myopia. *Mol Vis.* 2008;14:387–393.

47. Pertile KK, Schache M, Islam FM, et al. Assessment of TGIF as a candidate gene for myopia. *Invest Ophthalmol Vis Sci.* 2008;49: 49–54.

48. Zayats T, Guggenheim JA, Hammond CJ, Young TL. Comment on 'A PAX6 gene polymorphism is associated with genetic predisposition to extreme myopia'. *Eye.* 2008;22:598–599.

49. McBrien NA, Gentle A. Role of the sclera in the development and pathological complications of myopia. *Prog Retin Eye Res.* 2003; 22:307–338.

50. Rada JA, Shelton S, Norton TT. The sclera and myopia. *Exp Eye Res.* 2006;82:185–200.

51. Keeley FW, Morin JD, Vesely S. Characterization of collagen from normal human sclera. *Exp Eye Res.* 1984;39:533–542.

52. Siegwart JT Jr, Norton TT. The time course of changes in mRNA levels in tree shrew sclera during induced myopia and recovery. *Invest Ophthalmol Vis Sci.* 2002;43:2067–2075.

53. Gentle A, Liu Y, Martin JE, Conti GL, McBrien NA. Collagen gene expression and the altered accumulation of scleral collagen during the development of high myopia. *J Biol Chem.* 2003;278:16587–16594.

54. Paluru P, Ronan SM, Heon E, et al. New locus for autosomal dominant high myopia maps to the long arm of chromosome 17. *Invest Ophthalmol Vis Sci.* 2003;44:1830–1836.

55. Gushima H. Pharma SNP Consortium (PSC). Research on pharmacokinetics related genetic polymorphism among Japanese Population [in Japanese]. *Xenobiotic Metabolism and Disposition.* 2001; 16:340–345.

56. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics.* 2005;21:263–265.

57. Purcell S, Cherny SS, Sham PC. Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. Bioinformatics, 2003;19(1):149–150.

58. Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat Genet.* 2003;33:177–182.

J Hum Genet (2008) 53:74–86
DOI 10.1007/s10038-007-0220-8

ORIGINAL ARTICLE

Evaluation of resequencing on number of tag SNPs of 13 atherosclerosis-related genes in Thai population

Chintana Tocharoentanaphol · Somying Promso · Dianna Zelenika · Tassanee Lowhnoo · Sissades Tongsima · Thanyachai Sura · Wasun Chantratita · Fumihiko Matsuda · Sean Mooney · Anavaj Sakuntabhai

Received: 24 August 2007 / Accepted: 29 October 2007 / Published online: 28 November 2007

Abstract In the candidate gene approach, information about the distribution of single nucleotide polymorphisms (SNPs) is a crucial requirement for choosing efficient markers necessary for a case-control association study. To obtain such information, we discovered SNPs in 13 genes related to atherosclerosis by resequencing exon-flanking regions of 32 healthy Thai individuals. In total, 194 polymorphisms were identified, 184 of them SNPs, four insertions, and the rest deletions. Fifty-nine of the SNPs were characterized as novel polymorphisms, and these accounted for 30% of the identified SNPs. Comparing allele frequency distributions of the Thai population with other Asian populations shows similar patterns. In contrast, a low correlation pattern ($r = 0.521$) was found when comparing with either Caucasian or African populations. However, some rare alleles (rs11574541 and rs10874913) are found in the Thai population but not in other Asian populations. Most of the novel SNPs found were located outside the haplotype blocks generated by known SNPs in the Thai population. Only 5.77% of the novel SNPs lies in these defined haplotype blocks. The selection of haplotype-tagging SNPs shows that 8 of 13 genes benefited from the ethnic-specific genotype information. That is, when at least one novel SNP was present, the tagging SNPs chosen were altered. Functional prediction of 16 nonsynonymous SNPs (nsSNPs) by three different algorithm tools demonstrated that five nsSNPs possibly alter their corresponding protein functions. These results provide necessary information for conducting further genetic association studies involving the Thai population and demonstrate that resequencing of candidate genes provides more complete information for full genetic studies.

Keywords Thai · Atherosclerosis · SNP · Tag SNP · Resequencing · Novel SNP

Introduction

Cardiovascular disease (CVD) has been one of the leading causes of death in the Thai population over the last decade

Chintana Tocharoentanaphol and Somying Promso contributed equally to this work.

C. Tocharoentanaphol · D. Zelenika · T. Lowhnoo · F. Matsuda
Centre National de Génotypage, 2 rue Gaston Crémieux,
CP 5721, 91057 Evry Cedex, France

C. Tocharoentanaphol (✉)
Chulabhorn Cancer Centre, Viphavadee-Rangsit road, Laksi,
Bangkok 10120, Thailand
e-mail: chintaphol@hotmail.com

S. Promso · W. Chantratita
Department of Pathology, Faculty of Medicine Ramathibodi
Hospital, Mahidol University, Bangkok 10400, Thailand

T. Lowhnoo
Research center, Faculty of Medicine Ramathibodi Hospital,
Mahidol University, Bangkok 10400, Thailand

S. Tongsima
National Center for Genetic Engineering and Biotechnology,
Pathumthani, Thailand

T. Sura · A. Sakuntabhai
Department of Medicine, Faculty of Medicine Ramathibodi
Hospital, Mahidol University, Bangkok 10400, Thailand

A. Sakuntabhai
Laboratoire de Génétique de la réponse aux infections chez
l'homme, Institut Pasteur, 75724 Paris Cedex 15, France

S. Promso · S. Mooney
Center for Computational Biology and Bioinformatics,
Department of Medical and Molecular Genetics, Indiana
University School of Medicine, Indianapolis, IN 46202, USA

(Department of Medical Service Thailand 2006a). Atherosclerosis is one of the major factors among cardiovascular disease (Department of Medical Service Thailand 2006b). It is characterized by abnormal fatty deposits and fibrosis of the inner layer of arteries. This progressive disease normally takes years to manifest itself as a partial or complete interruption of blood flow. Depending on the artery affected, atherosclerosis can lead to a life-threatening condition. However, the risk of atherosclerosis is influenced by many factors, such as body mass index (BMI), blood pressure, plasma lipid levels, genetics, and environment. A number of genetic factors have shown significant correlation to processes related to atherosclerosis (Laukkanen and Yla-Herttuala 2002). Currently, candidate-gene-based association studies are the most common approach used in disease-causing gene identification research. Resequencing is a technique for SNP typing and discovery and is useful for fine mapping as well as discovery of unique single nucleotide polymorphisms (SNPs) for ethnic groups in a specific gene. SNP discovery is then an alternative choice for SNP typing prior to starting a case-control association study using a candidate gene approach (Do et al. 2006).

Inflammation plays an essential role in the etiology and progression of atherosclerosis. Simon et al. (2000) presented evidence that inflammation also has a role in vascular repair after mechanical arterial injury. Several genes involved in inflammation processes have been reported, namely, *ITGAM*, *ITGAX*, *ITGB7*, and *SELPLG*. The cell adhesion molecules, such as *ITGAM* (Integrin, alpha M) and *ITGAX* (Integrin, alpha X), are important in the adherence of monocytes to stimulate endothelium as well as in the phagocytosis of complement-coated particles. *ITGB7* (Integrin, beta 7) is also expected to play a role in adhesive interactions of leukocytes by being a receptor for fibronectin, whereas *SELPLG* (Selectin P lingand) is the high-affinity counter-receptor for P-selectin on myeloid cells and stimulates T lymphocytes, where it plays a critical role in the tethering and rolling of these cells to activate platelets or endothelia expressing P-selectin (Schneider et al. 2004).

The *CCL1*–chemokine (C–C motif) ligand 1 and *CCL2*–chemokine (C–C motif) ligand 2 belong to the CXC subfamily of the cytokine group, which is involved in immunoregulatory and inflammatory processes. This cytokine displays chemotactic activity for monocytes and basophils but not for neutrophils or eosinophils (Maglott et al. 2005). Recent studies reported that *CCL2* or *MCP-1* is involved in the pathogenesis of human atherosclerosis and myocardial infarction. This evidence also agrees with the hypothesis that this group is involved in inflammation (McDermott et al. 2005; Tabara et al. 2003).

In addition to the role of an anti-inflammatory cytokine, the role of transforming growth factor β (TGF-β) in atherosclerosis has been the subject of considerable debate for a decade (Singh and Ramji 2006). *TGFB2* (TGF-β 2) and *TGFB3* (TGF-β 3) show significant mRNA expression differences in atherogenic animal models (Tabibiazar et al. 2005; Wang et al. 2003). *TGFBR2* (TGF-β type II receptor) defects have been recently associated with Marfan syndrome (MFS) with a prominent cardioskeletal phenotype (Disabella et al. 2006; Matyas et al. 2006; Mizuguchi et al. 2004).

Many studies support that lipid plays a major role in the formation of atheromatous plaque. High density lipoprotein (HDL) is negatively correlated with risk of CVD (Asztalos 2004). Genes involved in lipid transport, such as *APOA1* (apolipoprotein A-I), *SCARB1* (scavenger receptor class B, member1), and *LIPG* (lipase, endothelial) may also be predisposed to the disease. *APOA1* is the major protein component of HDL in plasma. Previous studies have reported that SNPs, including -75[G/A] in the *APOA1* promoter, are significantly associated with HDL cholesterol (HDL-C) variability (Brown et al. 2006; Ruano et al. 2006). *SCARB1* is an HDL receptor that mediates selective cholesterol uptake from HDL to cells. McCarthy et al. (2003) examined polymorphisms in the HDL receptor gene *SCARB1* in 371 Caucasian patients suffering from coronary artery disease to determine their association with plasma lipids. They found an association between a combination of genotypes to the disease; however, this association was found only in women. One of their findings was that the genetic variants in *SCARB1* may be an important determinant of abnormal lipoproteins in women, which confers particular susceptibility on coronary artery disease (McCarthy et al. 2003). *LIPG*, a new member of the triglyceride lipase family, is expressed in endothelial cells (Jaye et al. 1999) and may have indirect atherogenic actions in vivo through its effect on circulating HDL-C and directly mediates cellular lipoprotein uptake (Fuki et al. 2003; Ishida et al. 2004).

Nowadays, many studies testing common SNPs associated with higher risk of developing common disease was done by using available SNP databases. When commercial microarray genotyping chips are applied to genetic studies, most SNPs in the databases come from the common SNPs in the Caucasian population exclusively. It is not yet clear whether public SNP databases are adequate for studies on other ethnic groups, such as the Thai population.

In this study, these 13 atherosclerosis-candidate genes were resequenced focusing on exons flanking regions in the Thai population. Using 32 unrelated healthy Thai DNA samples, this study established common genetic variation in these regions. We compared many parameters [allele frequency, ethnic differences, linkage disequilibrium (LD) block, and tags SNPs] needed for association study by using two data sets: (1) our resequencing data, which include all novel SNPs found in our study, and (2) only SNPs available in the public SNP databases and excluding novel SNPs.

Materials and methods

DNA sample

The 32 healthy Thais were recruited in this study and were blindly picked from the 64 healthy individuals conducted under the Thailand SNP Discovery Project. To capture the genetic makeup of Thais, all recruited individuals' families must have resided in Thailand for more than two generations. To avoid late-onset unhealthy samples, the individuals must be at least 50 years of age with no medical history of chronic diseases such as hypertension, diabetes mellitus, and cancer. Informed consent was obtained from each participant.

Resequencing analysis for polymorphism identification

Total genomic DNA from these 32 unrelated Thai individuals was isolated from peripheral leukocytes or Epstein–Barr virus (EBV)-transformed B lymphocytes by using the standard phenol–chloroform extraction method. The same amount of the two DNA samples was pooled together. SNPs were analyzed by direct sequencing of the pooled DNA samples. Before using, the pooled DNA was tested with a known SNP assay to test feasibility of using pooled DNA. For all genes analyzed, polymerase chain reaction (PCR) primers were newly designed by using Primers3 software (Rozen and Skaletsky 2000). The PCR targets included all exons based on GenBank; NT_033899 (*APOA1*), NT_010799 (*CCL1*), (*CCL2*), NT_010393 (*ITGAM and ITGAX*), NT_029419 (*ITGB7*), *NT*_010966 (*LIPG*), NT_009755 (*SCARB1*), NT_019546 (*SELPLG*), NT_021877 (*TGFB2*), *NT*_026437 (*TGFB3*), NT_032977 (*TGFBR2*), and NT_022517 (*TGFBR3*). Each exon was amplified separately and sequenced in both directions. PCR conditions and primer oligo sequences are available on the ThaiSNP database Web site (http://thaisnp2.biotec.or.th). Sequencing was performed using BigDye (Applied Biosystems) on an ABI 3730 DNA sequencer, according to the manufacturer's instruction. Sequence comparison, SNP discovery, and allele frequency were determined using Genalys program version 3.326a (Takahashi et al. 2002) and verified with visual inspection by two independent individuals. Before analysis, each chromatogram was trimmed to remove low-quality sequence.

Data and statistical analysis

Analysis of the resequencing of 16 pooled DNAs for each gene was performed using the Pool2 Package (Hoh et al. 2003). Based on the estimated frequencies, the Pool2 statistically calculates individual genotypes. The haplotypes were then estimated by using pairwise LD, statistics and the haplotype blocks were defined using confidence intervals (Gabriel et al. 2002) through the Haploview software package (Barrett et al. 2005). The ethnic group correlation was determined by comparing minor allele frequency (MAF) to dbSNP and HapMap data;—Caucasian, African, Japanese, and Chinese—using Pearson's moment correlation and Fisher's exact test function of the R-package software (R Development Core Team 2006). Statistical significance was determined at a two-sided value of $P < 0.001$ with 500,000 replications using the Monte Carlo test.

Tag SNPs of each gene were selected by using the aggressive tagging algorithm (de Bakker et al. 2005) within the Haploview software package. The calculation was applied to each gene in two categories: ThaiSNP data excluding novel SNPs, and ThaiSNP data including novel SNPs. All parameters were used according to the program's default, for which log of odds (LOD) threshold for multimarker tests was set as 3.0 and the default r^2 threshold was used (0.8). Differences in the number of tag SNPs in both conditions were compared using the tagging efficiency parameter. This parameter was calculated from the percentage of the ratio between the differences of the number of tagging SNPs found in the same gene with novel SNPs and without novel SNPS. Tag SNP efficiency = $(N_{\mathrm{RT}} - N_{\mathrm{DT}})/(N_{\mathrm{R}} - N_{\mathrm{D}})100$, where N_{RT} is the total number of tag SNPs selected from resequencing data, N_{DT} is the total number of tag SNPs selected from the SNP data verified by the National Center for Biotechnology Information (NCBI), N_{R} is the total number of SNPs from the resequencing process, and N_{D} is the total number of SNPs validated by the NCBI. $N_{\mathrm{RT}} - N_{\mathrm{DT}}$ is the total number of tag SNPs additionally discovered after resequencing, whereas $N_{\mathrm{R}} - N_{\mathrm{D}}$ is the total number of new SNPs additionally discovered after resequencing. When there is no new SNPs, $N_{\mathrm{R}} - N_{\mathrm{D}} = 0$, which means that no additional SNPs were found from resequencing and there is no additional tag SNPs. The higher percentage will represent the impact of novel SNPs to effect tag SNP selection, whereas zero means those novel SNPs were not found and were not gaining information.

Assessment of nonsynonymous SNPs

All SNPs identified in this study were mapped onto the human genome using NCBI Refseq (genome_build 36.1) and dbSNP (build126). The nonsynonymous SNPs (nsSNPs) were analyzed using SIFT (Ng and Henikoff 2002), PolyPhen (Ramensky et al. 2002), and SNPs3D (Yue et al. 2006) for assessing the potential impact of amino acid substitution on protein function. Those three tools were performed using their respective Web sites. SIFT (http://blocks.fhcrc.org/sift/SIFT.html) and PolyPhen (http://genetics.bwh.harvard.edu/pph/index.html) were run using the services on their Web sites with the default parameters. SNPs3D (http://www.snps3d.org/) scores were determined by the sequence-based NCBI scores provided on the Web site.

Results

In the identification of novel polymorphisms in the Thai population, the resequenced regions spanned a total of 453,533 bp and covered around 1 kb 5′ upstream from the starting codon and around 1 kb after the 3′ stop codon. We identified 194 polymorphisms from 16 pooled DNA samples located on the 13 atherosclerosis candidate genes, which comprised 184 SNPs, four insertions, and six deletions (Table 1). One hundred and thirty (67%) SNPs had common allele frequencies (MAF greater than 0.05), and the rest (64; 33%) were characterized as rare alleles (MAF less than 0.05). Figure 1 illustrates all genetic variations found in this study, separated by gene. The novel polymorphisms are indicated by an asterisk. Fifty-nine polymorphisms (30%) were discovered by mapping their positions to the RefSeq (http://www.ncbi.nlm.nih.gov/RefSeq/) sequences mapped to the human genome in RefSeq database build 36.1 and comparing with the location of SNPs reported in dbSNP database build 126 (Table 2). Of those, 16 SNPs (27%) were located in coding regions, whereas 43 SNPs (73%) mapped to intronic regions. Among the novel SNPs, 11 had MAF greater than 0.05 and were located in intronic regions, whereas 15 coding SNPs had allele frequencies less than 0.05.

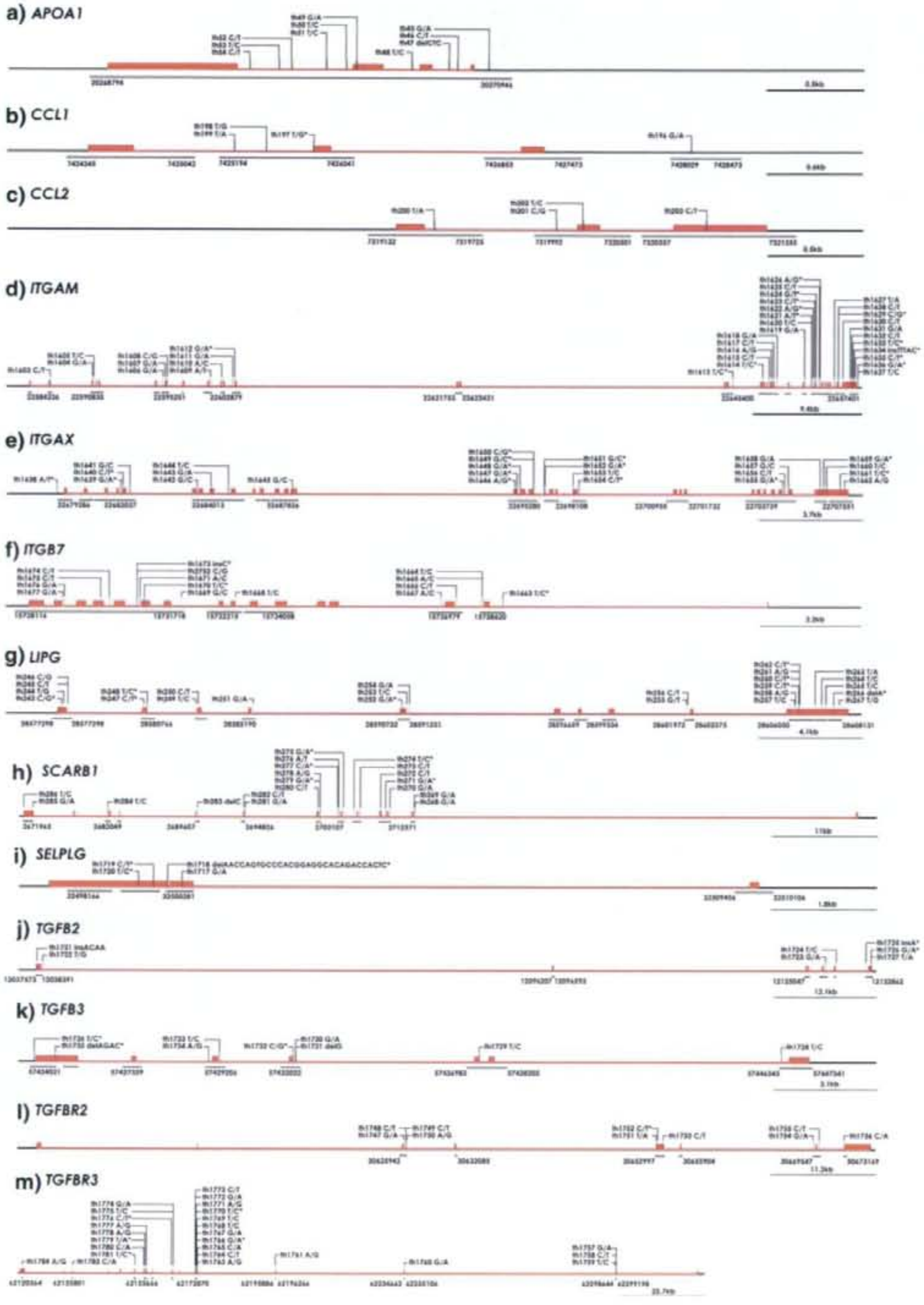
Ethnic differences in genetic-variation allele frequency

To determine ethnic differences in SNP allele frequencies, we searched reported SNPs from the HapMap database (http://www.hapmap.org) and dbSNP database (http://www.ncbi.nlm.nih.gov/SNP/) and compared our data with those from other ethnic groups. Differences in allele frequency distribution could be observed among various ethnic groups (Fig. 2). The allele frequencies found in Thai populations were similar to other Asian populations, particularly the combined Japanese and Chinese (JPT.CHB) data (correlation coefficient $r = 0.765$), but appeared as a low correlation to Caucasian (correlation coefficient $r = 0.521$) and African (correlation coefficient $r = 0.206$) populations. The significantly different markers among these ethnic groups are presented in Table 3. Two SNPs, rs4264407, and rs11165378, appeared monomorphic in the Thai population, whereas they were polymorphic in Caucasian and African populations. Also, rs7188189 appeared monomorphic in the Thai (data not shown) and African American populations, but it was polymorphic in European American populations (The Innate Immunity PGA 2000). In contrast, rs11574541 located on exon #13 of *ITGB7* and rs10874913 located on intron of *TGFBR3* were only found in the Thai population with an MAF of 0.03. There were some polymorphisms unique to only the Asian population, i.e., rs11057824, rs11057825, rs10482823, and rs2306888, whereas rs9936831 and rs11574635 were

Table 1 Summary of genetic variation in the 13 cardiovascular related genes

Gene	All genetic variations	SNPs	Ins/del	Novel genetic variation	SynSNPs	nsSNPs	UTR	Total base pairs sequenced (kb)	Frequency (bp/1SNP)
APOA1	10	9	1	2	0	1	2	2.2	215
CCL1	4	4	0	1	0	1	0	2.7	678
CCL2	4	4	0	0	1	0	1	1.9	476
ITGAM	35	34	1	13	1	4	5	13.0	371
ITGAX	25	25	0	14	4	3	5	11.8	473
ITGB7	15	14	1	3	2	0	3	6.3	417
LIPG	25	24	1	7	1	2	13	7.0	280
SCARB1	19	18	1	6	3	0	1	7.0	370
SELPLG	4	3	1	3	1	2	0	2.5	634
TGFB2	7	5	2	2	0	0	3	4.0	569
TGFB3	9	7	2	3	0	0	1	5.2	580
TGFBR2	10	10	0	1	1	1	0	3.9	391
TGFBR3	27	27	0	4	4	2	2	6.8	251
TOTAL	194	184	10	59	18	16	36	74.4	439

Ins/del Insertion/deletion variation, *UTR* untranslated region, *synSNP* synonymous single nucleotide polymorphism, *nsSNP* nonsynonymous single nucleotide polymorphism


a) APOA1
b) CCL1
c) CCL2
d) ITGAM
e) ITGAX
f) ITGB7
g) LIPG
h) SCARB1
i) SELPLG
j) TGFB2
k) TGFB3
l) TGFBR2
m) TGFBR3

◀ **Fig. 1** Location of single nucleotide polymorphisms (SNPs) in the *APOA1* (**a**), *CCL1* (**b**), *CCL2* (**c**), *ITGAM* (**d**), *ITGAX* (**e**), *ITGB7* (**f**), *LIPG* (**g**), *SCARB1* (**h**), *SELPLG* (**i**), *TGFB2* (**j**), *TGFB3* (**k**), *TGFBR2* (**l**), and *TGFBR3* (**m**) genes, indicated by *vertical lines*. Exons are indicated by a *solid rectangle*. The regions that have been sequenced are indicated by a *horizontal line*. The polymorphism numbers are accession numbers from the ThaiSNP database (correspondence to dbSNP rsIDs is given in Tables 2 and 3). The novel polymorphisms are indicated by an *asterisk*. The genomic sequences used for alignment are NT_033899 (*APOA1*), NT_010799 (*CCL1*), (*CCL2*), NT_010393 (*ITGAM and ITGAX*), NT_029419 (*ITGB7*), *NT*_010966 (*LIPG*), NT_009755 (*SCARB1*), NT_019546 (*SELPLG*), NT_021877 (*TGFB2*), *NT*_026437 (*TGFB3*), NT_032977 (*TGFBR2*), and NT_022517 (*TGFBR3*)

polymorphic in Thai, Caucasian, and African but not in Chinese and Japanese.

Linkage disequilibrium (LD) analysis and haplotype-block definition

LD statistics (D' or r^2) for the individual genotypes were calculated using the confidence intervals algorithm (Gabriel et al. 2002) implemented in the Haploview program for defining a haplotype block. To evaluate the effect the novel SNPs found in this study on the definition of haplotype blocks, we redefined the LD block in the Thai population using Thai SNPs excluding the novel SNPs (defined as known ThaiSNPs) and compared them with those from the combined Chinese–Japanese population. The populations were combined because the Japanese and Chinese populations were recently shown to be insignificantly different (The International HapMap 2005). Figure 3 shows haplotype-block definitions for the *ITGB7* gene using HapMap data, Thai population data with known SNPs, and Thai population data with novel SNPs. Both combined Chinese–Japanese and Thai population data had been defined with one block, with small differences in SNP members in the block. By introducing two novel SNPs from the Thai population to the SNP set and recalculation of the haplotype blocks, both novel SNPs appeared outside the original haplotype blocks. The haplotype block was then calculated for the rest of the 13 genes (data not shown); from this, 5.77% of novel SNPs were located within the LD block defined by the known Thai SNPs reported in dbSNP.

Tag SNP efficiency

Tag SNP efficiency was calculated by the number of tag SNPs from resequencing data and the number of tag SNPs from data verified by the NCBI. We identified that eight of 13 genes achieved 100% tag SNPs among discovered novel SNPs from the resequencing process, which means all SNPs newly discovered from the resequencing process were tagging SNPs (Table 4). In contrast, only one gene, *CCL2*, did not benefit from resequencing, because no novel SNPs were found. *ITGAM*, *TGFBR3*, *ITGAX*, and *LIPG* each had novel tag SNPs identified when the newly discovered SNPs were included. Our defined parameter, tag SNP efficiency (see "Materials and methods") of *ITGAM*, *TGFBR3*, *ITGAX*, and *LIPG* were 75.00%, 66.76%, 64.29%, and 50.00%, respectively. This suggested that a high percentage of novel SNPs define new tags. However, the overall tag SNP efficiency for all genes was 81.23%.

Functional polymorphism assessment

To assess the impact of amino acid substitutions on protein activity, we analyzed 16 nsSNPs, including eight novel nsSNPs, using three nsSNP functional prediction tools that utilize different algorithms: SIFT, PolyPhen, and SNPs3D (Table 5). Eleven nsSNPs were concordantly predicted to be intolerant and seven predicted to be neutral with these three functional prediction tools, whereas the remaining five were predicted to have a mixture of neutral and damaging activity. Allele frequencies were also used to classify the potential effects of nsSNPs. Interestingly, eight out of 16 nsSNPs were common SNPs (MAF > 5%). SNP rs2230429 located on *ITGAX* had the highest MAF observed (0.5) and was predicted to be damaging by all these tools.

Discussion

Cardiovascular disease is a complex disease that is influenced by many factors, including genetics. Candidate-gene-based association studies are the most common approach used in disease-causing gene identification research. Choosing markers for association approaches is based on extensive information on the distribution of SNPs across the genome. To obtain such information, 13 candidate genes, which had been associated to atherosclerosis, were resequenced in exon-flanking regions in the Thai population. We decreased all the possible known errors by using only high-quality chromatograms for the analysis. The sequencing data was obtained from both strands with difference primers. More than 80% of SNPs found were associated with both strands. We identified 59 novel polymorphisms (30%) by comparing them with dbSNP build 126. The percentage of novel polymorphisms found in this study was quite similar to the other SNP discovery studies (Michiels et al. 2007), but the allele frequencies were observed to have a high number of rare alleles; as many as 45 (76%) of those novel SNPs were rare alleles.

Table 2 Summary of 61 novel genetic variants identified in the 13 cardiovascular-related genes

Gene (contig ID)	ThaiSNP ID	Position in contig	Allele	Frequency	AA change	Type	Gene (contig ID)	ThaiSNP ID	Position in contig	Allele	Frequency	AA change	Type
APOA1	th49	20270152	G:A	0.969:0.031	A/T	nsSNP	ITGB7	th1673	15730710	_:C	0.984:0.016	–	Intron
(NT_033899)	th47	20270625	CTC:_	0.984:0.016	–	UTR	(NT_029419)	th1670	15730869	T:C	0.969:0.031	N/N	synSNP
CCL1	th197	7425784	T:G	0.969:0.031	M/R	nsSNP		th1663	15738568	T:C	0.955:0.046	–	Locus
(NT_010799)													
ITGAM	th1612	22602325	G:A	0.984:0.016	–	Intron	SELPLG	th1720	32499354	T:C	0.922:0.078	Y/H	nsSNP
(NT_010393)	th1613	22646387	T:C	0.984:0.016	–	Intron	(NT_019546)	th1719	32499604	C:T	0.984:0.016	T/T	synSNP
	th1614	22648985	T:C	0.922:0.078	–	Intron		th1718	32499840	AACCAGTGCCCACGGAGGCACAGACCACTC:_	0.712:0.289	–	Intron
	th1621	22653328	A:T	0.659:0.341	–	Intron	SCARB1	th281	2694784	G:A	0.984:0.016	–	Intron
	th1622	22653447	A:G	0.563:0.438	–	Intron	(NT_009755)	th279	2702586	G:A	0.938:0.063	–	Intron
	th1623	22653599	C:T	0.969:0.031	–	Intron		th277	2704352	C:A	0.938:0.063	–	Intron
	th1624	22653688	G:T	0.984:0.016	M/I	nsSNP		th275	2704961	G:A	0.955:0.046	–	Intron
	th1626	22654062	A:G	0.977:0.023	–	Intron		th274	2706013	T:C	0.865:0.135	–	Intron
	th1629	22655695	C:G	0.984:0.016	–	Intron		th271	2709346	G:A	0.938:0.063	–	Intron
	th1633	22656933	T:C	0.984:0.016	–	UTR	LIPG	th243	28577675	C:G	0.969:0.031	–	UTR
	th1634	22657040	_:TTTAC	0.953:0.047	–	UTR	(NT_010966)	th245	28577928	C:T	0.641:0.359	–	Intron
	th1635	22657150	C:T	0.969:0.031	–	UTR		th246	28577932	C:G	0.641:0.359	–	Intron
	th1636	22657362	G:A	0.923:0.077	–	Locus		th247	28580851	C:T	0.981:0.019	R/C	nsSNP
ITGAX	th1638	22679349	A:T	0.984:0.016	–	Locus		th248	28580992	T:C	0.981:0.019	–	Intron
(NT_010393)	th1639	22681584	G:A	0.984:0.016	–	Intron		th252	28590899	G:A	0.967:0.033	P/P	synSNP
	th1640	22681658	C:T	0.984:0.016	C/C	synSNP		th259	28606442	C:T	0.984:0.016	–	UTR
	th1646	22695534	A:G	0.984:0.016	A/A	synSNP		th260	28606473	C:T	0.984:0.016	–	UTR
	th1647	22695594	G:A	0.984:0.016	R/R	synSNP		th266	28607810	A:_	0.983:0.017	–	UTR
	th1648	22695702	G:A	0.984:0.016	–	Intron	TGFB2	th1725	12077374	_:A	0.859:0.141	–	Intron
	th1649	22696175	G:C	0.983:0.017	L/L	synSNP	(NT_021877)	th1726	12077946	G:A	0.969:0.031	–	UTR
	th1650	22696182	C:G	0.983:0.017	P/A	nsSNP	TGFB3	th1736	57424151	T:C	0.984:0.016	–	Locus
	th1651	22696570	G:C	0.984:0.016	–	Intron	(NT_026437)	th1735	57424782	AGAC:_	0.984:0.016	–	UTR
	th1652	22696601	G:A	0.817:0.183	–	Intron		th1732	57431803	C:G	0.984:0.016	–	Intron
	th1654	22697835	C:T	0.933:0.067	–	Intron	TGFBR3	th1781	46001605	T:C	0.984:0.016	Y/Y	synSNP
	th1655	22705064	G:A	0.981:0.019	–	Intron	(NT_032977)	th1779	46004452	T:A	0.982:0.018	–	Intron
	th1659	22706509	G:A	0.95:0.05	–	UTR		th1776	46012430	C:T	0.969:0.031	–	Intron
	th1661	22707107	T:C	0.953:0.047	–	UTR		th1770	46019870	T:C	0.984:0.016	F/L	nsSNP
							TGFBR2	th1752	30653256	C:T	0.969:0.031	R/W	nsSNP
							(NT_022517)						

UTR untranslated region, *synSNP* synonymous single nucleotide polymorphism, *nsSNP* nonsynonymous single nucleotide polymorphism

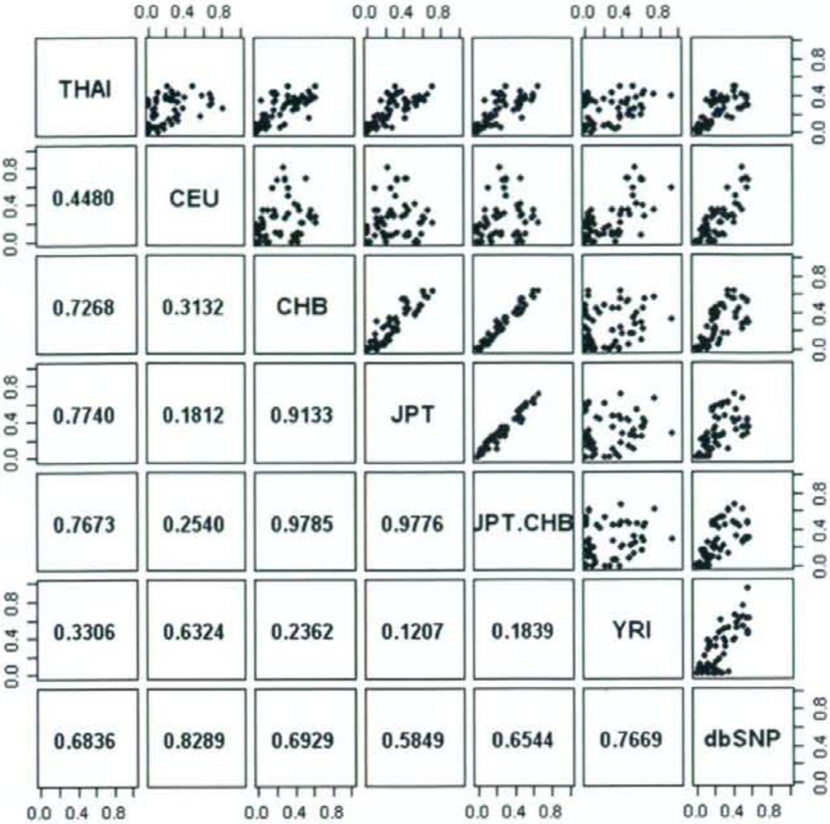


Fig. 2 Scatter plots of pairwise comparison of allele frequency distribution between Thai, Caucasian, Chinese, Japanese, Chinese + Japanese, African, and Single Nucleotide Polymorphism Database (dbSNP). Pearson's product moment correlation of allele frequency between the two populations is presented in the *lower diagonal matrix*

Table 3 Ethnic comparison of single nucleotide polymorphism (SNP) allele frequencies in the 13 cardiovascular-related genes

Gene	SNP(rs no.)	Type	Allele	Thai[a]	Hapmap[b]				Ethnic difference[c]
					Caucasian	Han Chinese	Japanese	African	
CCL1	rs2282691	Intron	A:T	0.375:0.625	0.404:0.596	0.405:0.595	0.523:0.477	0.551:0.449	0.176
CCL2	rs4586	synSNP	T:C	0.583:0.417	0.667:0.333	0.422:0.578	0.352:0.648	0.258:0.742	0.409
ITGAM	rs1143678	nsSNP	C:T	0.933:0.067	0.867:0.133	0.989:0.011	1.000:0.000	0.8:0.2	0.2
	rs3815801	Intron	A:G	0.672:0.328	0.667:0.333	0.822:0.178	0.739:0.261	0.608:0.392	0.214
	rs4077810	Intron	C:T	0.568:0.432	0.737:0.263	0.841:0.159	0.775:0.225	0.966:0.034	0.398
	rs4597342	UTR	C:T	0.607:0.393	0.724:0.276	0.8:0.2	0.705:0.295	0.967:0.033	0.36
	rs7184677	Intron	G:A	0.938:0.062	0.892:0.108	0.989:0.011	1.000:0.000	0.642:0.358	0.358
	rs7206295	Intron	C:T	0.65:0.35	0.717:0.283	0.8:0.2	0.705:0.295	0.967:0.033	0.317
	rs9936831	Intron	A:T	0.95:0.05	0.9:0.1	1.000:0.000	1.000:0.000	0.636:0.364	0.364
ITGAX	rs11150620	Intron	G:C	0.654:0.346	0.75:0.25	0.42:0.58	0.386:0.614	0.975:0.025	0.589
	rs1140195	UTR	A:G	0.609:0.391	0.737:0.263	0.367:0.633	0.398:0.602	0.975:0.025	0.608
	rs11574635	Intron	G:C	0.917:0.083	0.828:0.172	1.000:0.000	1.000:0.000	0.737:0.263	0.263
	rs2929	UTR	G:A	0.75:0.25	0.712:0.288	0.822:0.178	0.744:0.256	0.517:0.483	0.305
	rs4264407	Intron	G:C	1.000:0.000	0.892:0.108	1.000:0.000	1.000:0.000	0.933:0.067	0.108
ITGB7	rs11170465	Intron	G:A	0.839:0.161	0.949:0.051	0.92:0.08	0.802:0.198	0.975:0.025	0.173
	rs11170466	Intron	G:A	0.839:0.161	0.942:0.058	0.898:0.102	0.778:0.222	0.975:0.025	0.197
	rs11574541	synSNP	C:T	0.969:0.031	1.000:0.000	1.000:0.000	1.000:0.000	1.000:0.000	0.031

Table 3 continued

Gene	SNP(rs no.)	Type	Allele	Thai[a]	Hapmap[b]				Ethnic difference[c]
					Caucasian	Han Chinese	Japanese	African	
	rs2272299	Intron	G:A	0.85:0.15	N/A	0.9:0.1	0.761:0.239	0.884:0.116	0.139
	rs2272300	Intron	T:G	0.812:0.188	0.942:0.058	0.9:0.1	0.761:0.239	0.405:0.595	0.537
	rs2272301	Intron	C:G	0.95:0.05	0.851:0.149	0.942:0.058	0.872:0.128	0.982:0.018	0.131
	rs3817537	Intron	G:C	0.984:0.016	1.000:0.000	0.978:0.022	1.000:0.000	N/A	0.022
	rs3825084	Intron	A:C	0.906:0.094	0.821:0.179	0.94:0.06	0.805:0.195	0.949:0.051	0.144
LIPG	rs2000812	Intron	T:C	0.633:0.367	0.8:0.2	0.522:0.478	0.42:0.58	1.000:0.000	0.58
	rs2000813	nsSNP	C:T	0.667:0.333	0.692:0.308	0.656:0.344	0.761:0.239	0.966:0.034	0.31
	rs2276269	Intron	T:C	0.6:0.4	0.417:0.583	0.667:0.333	0.738:0.262	0.067:0.933	0.671
	rs3744843	UTR	A:G	0.922:0.078	N/A	0.744:0.256	0.726:0.274	0.667:0.333	0.255
	rs3786247	UTR	A:C	0.732:0.268	0.931:0.069	0.557:0.443	0.545:0.455	0.642:0.358	0.386
	rs3786248	UTR	A:G	0.938:0.062	0.933:0.067	0.756:0.244	0.727:0.273	1.000:0.000	0.273
	rs3819166	Intron	G:A	0.6:0.4	0.808:0.192	0.511:0.489	0.432:0.568	1.000:0.000	0.568
	rs3826577	UTR	A:T	0.938:0.062	0.933:0.067	0.756:0.244	0.727:0.273	1.000:0.000	0.273
	rs6507931	Intron	C:T	0.833:0.167	0.425:0.575	0.833:0.167	0.909:0.091	0.492:0.508	0.484
	rs9958734	UTR	T:C	0.75:0.25	0.946:0.054	0.605:0.395	0.583:0.417	0.839:0.161	0.363
SCARB1	rs11057824	Intron	C:T	0.667:0.333	1.000:0.000	0.622:0.378	0.486:0.514	1.000:0.000	0.514
	rs11057825	Intron	C:T	0.646:0.354	1.000:0.000	0.567:0.433	0.444:0.556	1.000:0.000	0.556
	rs3825140	UTR	C:T	0.783:0.217	1.000:0.000	0.622:0.378	0.524:0.476	N/A	0.476
	rs4765615	Intron	C:T	0.5:0.5	0.509:0.491	0.655:0.345	0.697:0.303	0.491:0.509	0.206
	rs5889	synSNP	C:T	0.708:0.292	0.992:0.008	0.58:0.42	0.464:0.536	1.000:0.000	0.536
	rs5892	synSNP	C:T	0.906:0.094	1.000:0.000	0.977:0.023	0.965:0.035	0.892:0.108	0.108
SELPLG	rs2228315	nsSNP	G:A	0.808:0.192	0.908:0.092	0.756:0.244	0.83:0.17	0.617:0.383	0.291
TGFB2	rs900	UTR	A:T	0.281:0.719	N/A	N/A	N/A	0.667:0.333	0.386
	rs10482823	Intron	T:C	0.984:0.016	1.000:0.000	0.989:0.011	0.989:0.011	1.000:0.000	0.016
	rs6684205	Intron	A:G	0.266:0.734	0.808:0.192	0.278:0.722	0.227:0.773	0.525:0.475	0.581
TGFB3	rs3917147	Intron	T:C	0.969:0.031	1.000:0.000	0.911:0.089	0.966:0.034	0.667:0.333	0.333
	rs3917187	Intron	G:A	0.55:0.45	0.731:0.269	0.44:0.56	0.616:0.384	0.357:0.643	0.374
	rs3917200	Intron	T:C	0.95:0.05	0.944:0.056	0.966:0.034	0.911:0.089	0.708:0.292	0.258
	rs3917201	Intron	A:G	0.839:0.161	0.708:0.292	0.444:0.556	0.58:0.42	0.625:0.375	0.395
TGFBR2	rs1155705	Intron	A:G	0.297:0.703	0.683:0.317	0.3:0.7	0.33:0.67	0.608:0.392	0.386
	rs1155708	Intron	G:A	0.297:0.703	0.675:0.325	0.3:0.7	0.33:0.67	0.608:0.392	0.378
	rs2276767	Intron	C:A	0.875:0.125	0.667:0.333	0.889:0.111	0.911:0.089	0.942:0.058	0.275
	rs2276768	Intron	C:T	0.567:0.433	0.898:0.102	0.8:0.2	0.67:0.33	0.707:0.293	0.331
	rs9843942	Intron	G:A	0.734:0.266	0.616:0.384	0.557:0.443	0.583:0.417	0.382:0.618	0.352
TGFBR3	rs10874913	Intron	C:T	0.969:0.031	1.000:0.000	1.000:0.000	1.000:0.000	N/A	0.031
	rs11165376	Intron	A:G	0.683:0.317	0.31:0.69	0.477:0.523	0.573:0.427	0.563:0.438	0.373
	rs11165377	Intron	C:T	0.906:0.094	0.742:0.258	0.689:0.311	0.9:0.1	0.92:0.08	0.231
	rs11165378	Intron	T:C	1.000:0.000	1.000:0.000	1.000:0.000	1.000:0.000	0.942:0.058	0.058
	rs12069176	Intron	A:G	0.367:0.633	0.686:0.314	0.533:0.467	0.411:0.589	0.467:0.533	0.319
	rs1805109	UTR	G:A	0.661:0.339	0.915:0.085	N/A	N/A	0.936:0.064	0.275
	rs1805110	nsSNP	C:T	0.661:0.339	0.907:0.093	0.551:0.449	0.557:0.443	0.877:0.123	0.356
	rs1805117	UTR	A:G	0.859:0.141	0.78:0.22	0.936:0.064	0.841:0.159	0.956:0.044	0.176
	rs2279455	Intron	T:C	0.589:0.411	0.333:0.667	0.778:0.222	0.67:0.33	0.381:0.619	0.445
	rs2296621	Intron	C:A	0.659:0.341	0.75:0.25	0.956:0.044	0.841:0.159	0.911:0.089	0.297
	rs2306886	Intron	G:A	0.733:0.267	N/A	0.644:0.356	0.789:0.211	N/A	0.145
	rs2306887	Intron	C:T	0.714:0.286	N/A	0.676:0.324	0.667:0.333	N/A	0.047
	rs2306888	synSNP	T:C	0.933:0.067	1.000:0.000	0.878:0.122	0.795:0.205	1.000:0.000	0.205

Table 3 continued

Gene	SNP(rs no.)	Type	Allele	Thai[a]	Hapmap[b]				Ethnic difference[c]
					Caucasian	Han Chinese	Japanese	African	
	rs284176	Intron	G:A	0.667:0.333	0.664:0.336	0.544:0.456	0.545:0.455	0.703:0.297	0.159
	rs3738441	Intron	C:T	0.5:0.5	0.792:0.208	0.367:0.633	0.3:0.7	0.619:0.381	0.492
	rs6696224	Intron	A:G	0.633:0.367	0.917:0.083	0.522:0.478	0.411:0.589	0.792:0.208	0.506
	rs6699304	Intron	C:T	0.969:0.031	0.825:0.175	1.000:0.000	0.9:0.1	0.9:0.1	0.175
	rs7524066	Intron	G:T	0.812:0.188	0.664:0.336	0.944:0.056	0.878:0.122	0.542:0.458	0.402

[a] Allele frequency of the Thai population was determined by direct sequencing of DNA-pooled two samples selected from 32 unrelated Thai

[b] Allele frequencies were determined from data obtained by searching the Hapmap database (http://www.hapmap.org). SNP of some genes are not included in this table because of lacked of information in the Hapmap database

[c] Ethnic differences in allele frequency were calculated by subtracting the lowest allele frequency of the minor allele from the highest allele frequency of the minor allele among the ethnic groups for each SNP site

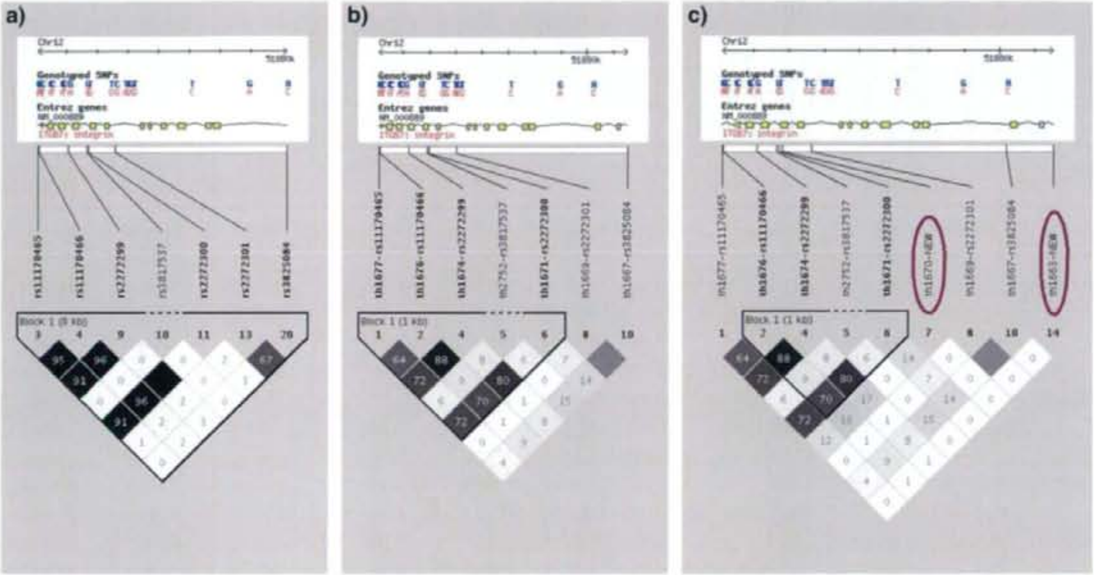


Fig. 3 The haplotype-block definition of the *ITGB7* gene comparing Japanese and Chinese HapMap data (**a**), Thai population without novel single nucleotide polymorphism (SNP) data (**b**), and Thai population with novel SNPs (**c**) using confidence intervals (Gabriel et al. 2002) in the Haploview software. SNP locations linked to the physical map on the chromosome are shown on the *white rectangle*. The novel SNPs are marked by *red ovals*

However, these novel SNPs are still important to consider for high-resolution association study design.

The ethnic differences between these SNPs could be responsible for differences in gene regulation and differences in the prevalence of diseases among these ethnic groups. Because of this, allele frequencies in the Thai population were compared with Chinese, Japanese, Caucasian, African, and average allele frequencies in dbSNP. Not surprisingly, the results showed that the allele frequency distribution of the Thai population was more correlated to other Asian populations, Chinese and Japanese, than to Caucasian and African populations. Correlation coefficients were similar to other recent studies (Cha et al. 2004; Kim et al. 2005; Mahasirimongkol et al. 2006). When compared with a similar study performed on the Korean population (Kim et al. 2005), the allele frequency of Korean populations was very similar to that of the Japanese population (correlation coefficient $r = 0.907$), whereas it had very different patterns of allele frequency compared with Caucasian (correlation coefficient $r = 0.359$) or African (correlation coefficient $r = 0.156$) populations. When focusing only on the correlation

Table 4 Number of tag single nucleotide polymorphisms (SNPs) selected from resequencing data, number of tag SNPs selected from data verified by the National Center for Biotechnology Information (NCBI) and percentage of tag SNPs efficiency

Gene	No. of tag SNPs from resequencing data (N_{RT})	No. of total SNPs from resequencing process (N_R)	No. of tag SNPs from data verified by NCBI (N_{DT})	No. of total SNPs validated by NCBI (N_D)	Percentage of tag SNP efficiency (%)
APOA1	8	8	7	7	100.00
CCL1	4	4	3	3	100.00
CCL2	4	4	4	4	0.00
ITGAM	20	31	11	19	75.00
ITGAX	18	23	9	9	64.29
ITGB7	9	14	7	12	100.00
LIPG	18	24	15	18	50.00
SCARB1	17	17	11	11	100.00
SELPLG	3	3	1	1	100.00
TGFB2	4	5	3	4	100.00
TGFB3	6	6	4	4	100.00
TGFBR2	7	10	6	9	100.00
TGFBR3	19	26	17	23	66.67
TOTAL	137	175	98	124	81.23

Table 5 Assessment of the 16 nonsynonymous single nucleotide polymorphisms (SNPs) in the 13 cardiovascular-related genes using SIFT, PolyPhen, and SNP3D

Gene (protein ID)	ThaiSNP ID	NCBI SNP ID	Frequency	AA variant	SIFT score*	SIFT prediction	PolyPhen prediction	SNPs3D prediction
APOA1 (NP_000030)	th49	New	G[0.969]/A[0.031]	A61T	1.00	Tolerant	Benign	Neutral
CCL1 (NP_002972)	th197	New	T[0.969]/G[0.031]	I63R	0.00	Intol	PRB	NA
ITGAM (NP_000623)	th1604	rs1143679	G[0.969]/A[0.031]	R77H	0.47	Tolerant	Benign	Neutral
	th1617	rs7201448	C[0.933]/T[0.067]	A858V	0.19	Tolerant	Benign	Neutral
	th1624	New	G[0.984]/T[0.016]	M951I	0.38	Tolerant	Benign	Neutral
	th1630	rs1143678	C[0.933]/T[0.067]	P1146S	0.49	Tolerant	Benign	Neutral
ITGAX (NP_000878)	th1643	rs12928508	G[0.923]/A[0.077]	A251T	1.00	Tolerant	Benign	Neutral
	th1645	rs2230429	C[0.5]/G[0.5]	P517R	0.00	Intol	PRB	Damaging
	th1650	New	C[0.983]/G[0.017]	P720A	0.00	Intol	PRB	Damaging
SELPLG (NP_002997)	th1720	New	T[0.922]/C[0.078]	Y297H	0.25	Tolerant	Benign	Neutral
	th1717	rs2228315	G[0.808]/A[0.192]	M62I	0.00	Intol	Benign	Neutral
LIPG (NP_006024)	th247	New	C[0.981]/T[0.019]	R54C	0.00	Intol	PRB	Damaging
	th250	rs2000813	C[0.667]/T[0.333]	T111I	0.00	Intol	Benign	Neutral
TGFBR2 (NP_003233)	th1752	New	C[0.969]/T[0.031]	R193W	0.00	Intol	PRB	Neutral
TGFBR3 (NP_003234)	th1770	New	T[0.984]/C[0.016]	F142L	0.71	Tolerant	POS	NA
	th1758	rs1805110	C[0.661]/T[0.339]	S15F	0.03	Intol	Benign	NA

NCBI National Center for Biotechnology Information, *AA* amino acid, *Intol* Intolerant, *POS* possibly damaging, *PRB* probably damaging, *NA* no data or SNP analysis available

* TI scores ≤0.05 are predicted to be Intolerant, whereas TI scores >0.05 are tolerant variants

coefficient (*r*) among the Asian population, the Thai population and combined Chinese–Japanese frequencies had a higher correlation than between the Thai population and Chinese or Japanese populations. These results support the findings of the international HapMap consortium that the Chinese and Japanese populations are insignificantly different (The International HapMap 2005). Consistent with their population histories, the admixture event between Thai and Chinese is believed to have occurred quite some time ago. Although the allele frequency distribution of the Thai population was similar to the other Asian populations, there are some rare allele markers (rs11574541 and rs10874913) found only in the Thai population but not in other Asian populations.

For LD-block comparison, most novel SNPs were located out of the block defined by known SNPs. Some of them were formed a new LD block, but there were no LD-block formations of mostly novel SNPs. The number of blocks might reflect population age, suggesting the greater LD block is the older population. We found a lower number of blocks than in other Asian populations, in which the referred age of the Thai population was younger than other Asian populations. Only 5.77% of the novel SNPs appear in the defined haplotype block. The LD blocks might be affected by the lower allele frequencies observed for the novel SNPs found in our study. These data indicate that using published SNP data alone would not have adequate coverage of the target region associated with disease. SNP discovery approaches can help identify causative SNPs, which were not reported in the public SNP databases.

When the number of tag SNPs of these two groups was compared, the average tag SNP efficiency (see "Materials and methods") was 81.23. Eight of 13 genes showed 100% tag SNP efficiency, that is, each of the newly discovered SNPs were also defined as tag SNPs. Therefore, additional SNP discovery is needed to assemble a map for use in the Thai population. The tag SNP results showed concordance to the LD-block results. Consequently, our data agree with the study of Carlson et al. (2003), who suggested that the study of populations other than European Americans required additional SNP discovery before conclusions can be drawn as to the adequacy of dbSNP for each population.

Additionally, SNP location is another factor that influences the selection of prominent SNPs for further association studies. Using three commonly known algorithms for protein function damaging SNP prediction, several novel SNPs located in coding regions were found to likely affect the alteration of protein function. This hypothesis should be investigated further using experimental functional assays to determine their corresponding effects.

In summary, resequencing is a powerful method to discover novel SNPs and SNPs that are specific to certain ethnic groups. This approach could also reveal the distribution of SNPs along interesting genes, which will be useful for future association studies. Consequently, this SNP discovery project provided sufficient information for marker selection used in case-control association studies utilizing candidate gene approaches.

Acknowledgments We thank Justin Dantzer for designing figures used in this article and Lang Li and Jeesun Jung for helpful advice. CT received the postdoctoral fellowship from the National Center for Genetic Engineering and Biotechnology, Thailand, under the Thailand SNP Discovery Project. SP is funded by Thailand Center for Excellence in Life Sciences (TCELS) under the Postdoctoral Scholarships Program. SDM is funded by NIH K22LM009135.

References

Asztalos BF (2004) High-density lipoprotein metabolism and progression of atherosclerosis: new insights from the HDL Atherosclerosis Treatment Study. Curr Opin Cardiol 19:385–391

Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21:263–265

Brown CM, Rea TJ, Hamon SC, Hixson JE, Boerwinkle E, Clark AG, Sing CF (2006) The contribution of individual and pairwise combinations of SNPs in the APOA1 and APOC3 genes to interindividual HDL-C variability. J Mol Med V84:561–572

Carlson CS, Eberle MA, Rieder MJ, Smith JD, Kruglyak L, Nickerson DA (2003) Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans. Nat Genet 33:518–521

Cha PC, Yamada R, Sekine A, Nakamura Y, Koh CL (2004) Inference from the relationships between linkage disequilibrium and allele frequency distributions of 240 candidate SNPs in 109 drug-related genes in four Asian populations. J Hum Genet 49:558–572

de Bakker PI, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, Altshuler D (2005) Efficiency and power in genetic association studies. Nat Genet 37:1217–1223

Department of Medical Service Thailand G (2006a) Statistical report. (http://www.dms.moph.go.th/statreport/index.html)

Department of Medical Service Thailand G (2006b) Statistical report of cardiovascular disease in Thai population. (http://www.dms.moph.go.th/statreport/2547/table0647.htm)

Disabella E, Grasso M, Marziliano N, Ansaldi S, Lucchelli C, Porcu E, Tagliani M, Pilotto A, Diegoli M, Lanzarini L, Malattia C, Pelliccia A, Ficcadenti A, Gabrielli O, Arbustini E (2006) Two novel and one known mutation of the TGFBR2 gene in Marfan syndrome not associated with FBN1 gene defects. Eur J Hum Genet 14:34–38

Do H, Vasilescu A, Diop G, Hirtzig T, Coulonges C, Labib T, Heath SC, Spadoni JL, Therwath A, Lathrop M, Matsuda F, Zagury JF (2006) Associations of the IL2Ralpha, IL4Ralpha, IL10Ralpha, and IFN (gamma) R1 cytokine receptor genes with AIDS progression in a French AIDS cohort. Immunogenetics 58:89–98

Fuki IV, Blanchard N, Jin W, Marchadier DH, Millar JS, Glick JM, Rader DJ (2003) Endogenously produced endothelial lipase enhances binding and cellular processing of plasma lipoproteins via heparan sulfate proteoglycan-mediated pathway. J Biol Chem 278:34331–34338

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D (2002) The structure of haplotype blocks in the human genome. Science 296:2225–2229

Hoh J, Matsuda F, Peng X, Markovic D, Lathrop MG, Ott J (2003) SNP haplotype tagging from DNA pools of two individuals. BMC Bioinform 4:14

Ishida T, Choi SY, Kundu RK, Spin J, Yamashita T, Hirata K, Kojima Y, Yokoyama M, Cooper AD, Quertermous T (2004) Endothelial lipase modulates susceptibility to atherosclerosis in apolipoprotein-E-deficient mice. J Biol Chem 279:45085–45092

Jaye M, Lynch KJ, Krawiec J, Marchadier D, Maugeais C, Doan K, South V, Amin D, Perrone M, Rader DJ (1999) A novel endothelial-derived lipase that modulates HDL metabolism. Nat Genet 21:424–428

Kim JY, Moon SM, Ryu HJ, Kim JJ, Kim HT, Park C, Kimm K, Oh B, Lee JK (2005) Identification of regulatory polymorphisms in the TNF-TNF receptor superfamily. Immunogenetics 57:297–303

Laukkanen J, Yla-Herttuala S (2002) Genes involved in atherosclerosis. Exp Nephrol 10:150–163

Maglott D, Ostell J, Pruitt KD, Tatusova T (2005) Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res 33:D54–D58

Mahasirimongkol S, Chantratita W, Promso S, Pasomsab E, Jinawath N, Jongjaroenprasert W, Lulitanond V, Krittayapoositpot P, Tongsima S, Sawanpanyalert P, Kamatani N, Nakamura Y, Sura T (2006) Similarity of the allele frequency and linkage disequilibrium pattern of single nucleotide polymorphisms in drug-related gene loci between Thai and northern East Asian populations: implications for tagging SNP selection in Thais. J Hum Genet 51:896–904

Matyas G, Arnold E, Carrel T, Baumgartner D, Boileau C, Berger W, Steinmann B (2006) Identification and in silico analyses of novel TGFBR1 and TGFBR2 mutations in Marfan syndrome-related disorders. Hum Mutat 27:760–769

McCarthy JJ, Lehner T, Reeves C, Moliterno DJ, Newby LK, Rogers WJ, Topol EJ (2003) Association of genetic variants in the HDL receptor, SR-B1, with abnormal lipids in women with coronary artery disease. J Med Genet 40:453–458

McDermott DH, Yang Q, Kathiresan S, Cupples LA, Massaro JM, Keaney JF Jr, Larson MG, Vasan RS, Hirschhorn JN, O'Donnell CJ, Murphy PM, Benjamin EJ (2005) CCL2 polymorphisms are associated with serum monocyte chemoattractant Protein-1 levels and myocardial infarction in the Framingham Heart Study. Circulation 112:1113–1120

Michiels S, Danoy P, Dessen P, Bera A, Boulet T, Bouchardy C, Lathrop M, Sarasin A, Benhamou S (2007) Polymorphism discovery in 62 DNA repair genes and haplotype-associations with risks for lung, and head and neck cancers. Carcinogenesis (in press)

Mizuguchi T, Collod-Beroud G, Akiyama T, Abifadel M, Harada N, Morisaki T, Allard D, Varret M, Claustres M, Morisaki H, Ihara M, Kinoshita A, Yoshiura K, Junien C, Kajii T, Jondeau G, Ohta T, Kishino T, Furukawa Y, Nakamura Y, Niikawa N, Boileau C, Matsumoto N (2004) Heterozygous TGFBR2 mutations in Marfan syndrome. Nat Genet 36:855–860

Ng PC, Henikoff S (2002) Accounting for human polymorphisms predicted to affect protein function. Genome Res 12:436–446

R Development Core Team G (2006) R: a language and environment for statistical computing

Ramensky V, Bork P, Sunyaev S (2002) Human non-synonymous SNPs: server and survey. Nucleic Acids Res 30:3894–900

Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol 132:365–386

Ruano G, Seip RL, Windemuth A, Zollner S, Tsongalis GJ, Ordovas J, Otvos J, Bilbie C, Miles M, Zoeller R, Visich P, Gordon P, Angelopoulos TJ, Pescatello L, Moyna N, Thompson PD (2006) Apolipoprotein A1 genotype affects the change in high density lipoprotein cholesterol subfractions with exercise training. Atherosclerosis 185:65–69

Schneider M, Tognolli M, Bairoch A (2004) The Swiss-Prot protein knowledgebase and ExPASy: providing the plant community with high quality proteomic data and tools. Plant Physiol Biochem 42:1013–1021

Simon DI, Chen Z, Seifert P, Edelman ER, Ballantyne CM, Rogers C (2000) Decreased neointimal formation in Mac-1−/− mice reveals a role for inflammation in vascular repair after angioplasty. J Clin Invest 105:293–300

Singh NN, Ramji DP (2006) The role of transforming growth factor-beta in atherosclerosis. Cytokine Growth Factor Rev 17:487–499

Tabara Y, Kohara K, Yamamoto Y, Igase M, Nakura J, Kondo I, Miki T (2003) Polymorphism of the monocyte chemoattractant Protein (MCP-1) gene is associated with the plasma level of MCP-1 but not with carotid intima-media thickness. Hypertens Res 26:677–683

Tabibiazar R, Wagner RA, Ashley EA, King JY, Ferrara R, Spin JM, Sanan DA, Narasimhan B, Tibshirani R, Tsao PS, Efron B, Quertermous T (2005) Signature patterns of gene expression in mouse atherosclerosis and their correlation to human coronary disease. Physiol Genom 22:213–226

Takahashi M, Matsuda F, Margetic N, Lathrop M (2002) Automated identification of single nucleotide polymorphisms from sequencing data. Proc IEEE Comput Soc Bioinform Conf 1:87–93

The Innate Immunity PGA P (2000) IIPGA genetic data: ITGAM allele frequencies. (https://innateimmunity.net/IIPGA2/PGAs/InnateImmunity/ITGAM/allele_freqs?flavor=masked)

The International HapMap C (2005) A haplotype map of the human genome. Nature 437:1299–1320

Wang X, Le Roy I, Nicodeme E, Li R, Wagner R, Petros C, Churchill GA, Harris S, Darvasi A, Kirilovsky J, Roubertoux PL, Paigen B (2003) Using advanced intercross lines for high-resolution mapping of HDL cholesterol quantitative trait loci. Genome Res 13:1654–1664

Yue P, Melamud E, Moult J (2006) SNPs3D: candidate gene and SNP selection for association studies. BMC Bioinform 7:166

CLINICAL STUDY

Genetic investigation of four meiotic genes in women with premature ovarian failure

Béatrice Mandon-Pépin, Philippe Touraine[1], Frédérique Kuttenn[1], Céline Derbois[2], Agnes Rouxel[1], Fumihiko Matsuda[2], Alain Nicolas[3], Corinne Cotinot and Marc Fellous[4]

INRA, UMR 1198, ENVA, CNRS, FRE 2857, Biologie du Développement et Reproduction, Jouy-en-Josas F-78350, France, [1]APHP, Department of Endocrinology and Reproductive Medicine, GH Pitié Salpêtrière, 47-83 Bd de l'Hôpital, 75651 Paris Cedex 13, France, [2]Centre National de Génotypage, 2 rue Gaston Crémieux, 91057 Evry, Cedex, France, [3]CNRS, UMR144, Génétique Moléculaire de la Recombinaison, Institut Curie, 26 rue d'Ulm, 75248 Paris Cedex 05, France and [4]INSERM, U709, Genomics and Epigenetics of Placentary Pathology, Hôpital Cochin, 123 Bld de Port Royal, 75014 Paris, France

(Correspondence should be addressed to B Mandon-Pépin; Email: beatrice.pepin@jouy.inra.fr)

Abstract

Objective: The goal of this study was to determine whether mutations of meiotic genes, such as disrupted meiotic cDNA (*DMC1*), MutS homolog (*MSH4*), *MSH5*, and *S. cerevisiae* homolog (*SPO11*), were associated with premature ovarian failure (POF).

Design: Case–control study.

Methods: Blood sampling, karyotype, hormonal dosage, ultrasound, and ovarian biopsy were carried out on most patients. However, the main outcome measure was the sequencing of genomic DNA from peripheral blood samples of 41 women with POF and 36 fertile women (controls).

Results: A single heterozygous missense mutation, substitution of a cytosine residue with thymidine in exon 2 of *MSH5*, was found in two Caucasian women in whom POF developed at 18 and 36 years of age. This mutation resulted in replacement of a non-polar amino acid (proline) with a polar amino acid (serine) at position 29 (P29S). Neither 36 control women nor 39 other patients with POF possessed this genetic perturbation. Another POF patient of African origin showed a homozygous nucleotide change in the tenth of *DMC1* gene that led to an alteration of the amino acid composition of the protein (M200V).

Conclusions: The symptoms of infertility observed in the *DMC1* homozygote mutation carrier and in both patients with a heterozygous substitution in exon 2 of the *MSH5* gene provide indirect evidence of the role of genes involved in meiotic recombination in the regulation of ovarian function. *MSH5* and *DMC1* mutations may be one explanation for POF, albeit uncommon.

European Journal of Endocrinology **158** 107–115

Introduction

Premature ovarian failure (POF; OMIM no. 311360) is a cause of female infertility due to the loss of normal ovarian function in women before the age of 40 years (1). The condition is defined by the absence or cessation of normal menses for at least 6 months (primary or secondary amenorrhea), menopausal level of follicle-stimulating hormone (FSH) >40 mIU/ml, hypoestrogenism and infertility (2, 3). POF affects 1 and 0.1% of women by 40 and 30 years of age respectively. POF is not uncommon considering the incidence rate of 1–2% of women during their reproductive life.

Several mechanisms may be involved in POF pathogenesis such as viral or autoimmune inflammatory disease, environmental toxics, and radiation or chemotherapy, but the genetic contribution is a significant etiological component. However, the disorder can occur on a familial basis, and there is evidence for a genetic mechanism in at least some cases. Deletions and translocations involving three regions of the X chromosome (Xq13-22, Xq26-28, and Xp11.2-22.1) have been associated with POF (4–9). Several genes located on this chromosome (i.e., bone morphogenetic protein-15 *BMP15*, kit ligand *KITLG*) have been sequenced in cohorts of POF patients, and heterozygous variants were detected but their frequency remained rare and did not appear to be a common cause of POF (10–13).

Candidate gene approaches have revealed few mutations in the gonadotropins and their receptors (14, 15) except noteworthy missense variant Ala189 Val of the *FSH* receptor gene which was strongly associated with POF in the Finnish population but rare in other world populations (16–19). POF can also be associated in familial syndromes such as type 1 blepharophimosis, ptosis, and epicanthus inversus syndrome (BPES; OMIM no. 110100) (20). Several *FOXL2* gene mutations have been reported in the type 1 BPES and nonsyndromic POF cases but are uncommon in diverse

DOI: 10.1530/EJE-07-0400
Online version via www.eje-online.org

populations (21, 22). Recently, attention has been focused on members of the transforming growth factor-β (TGF-β) superfamily synthesized by the oocyte, growth differentiation factor-9, and BMP15. These studies identified several heterozygous variants that are significantly more prevalent among women with POF but they are not a major cause of ovarian insufficiency (13, 23–25). Mutations in autosomal genes (galactose-1-phosphate uridylyltransferase, *GALT1*; transforming growth factor beta receptor, *TGFBR3*; inhibin alpha, *INHa*; forkhead box E1, *FOXE1*; and *β-glycan*) have also been related to POF (23, 24, 26–30). Nevertheless, in most cases, the etiopathology of the disease remains unknown.

In the ovary, primordial germ cells enter into meiosis from week 9 post-conception, oocytes pass through leptotene, zygotene, and pachytene stages before arresting in the last stage of meiotic prophase I, the diplotene, or dictyate stage at about the time of birth. It is widely accepted, although recently debated, that in mammals a female is born with a fixed number of oocytes within the ovaries (31, 32). The fertile lifespan of a female depends on the size of the oocyte pool at birth and the rapidity of the oocyte pool depletion. The phenotype of ovaries in null mutant mice for several meiotic genes could be strikingly similar to clinical observations found in human infertility and POF. In female mice lacking the *Dmc1* gene, normal oogenesis was aborted in embryos, and germ cells disappeared in the adult ovary (33, 34). The ovaries of *Msh5*−/− female mice are normal in size at birth, but degenerate progressively to become rudimentary, concomitant with the decline in oocyte numbers from day 3 pp until adulthood (35). The aim of this study was to screen a cohort of 41 clinically well-characterized patients who present unexplained infertility (normal XX karyotype, women with POF) for mutations in four meiotic genes. For this purpose, the exons of these four genes (*DMC1*, *SPO11*, *MSH4*, and *MSH5*) were sequenced and compared with the human corresponding gene to evaluate the impact of meiotic prophase arrest in 46 XX females with ovarian disorders.

Materials and methods

Patients and control population

Patients ($n=41$) were mainly ($n=35$) recruited from the reproductive endocrine unit of Pitie-Salpetriere Hospital, Paris, France. The diagnostic criteria for POF include at least 6 months of amenorrhea before the age of 40 years, with high serum FSH levels (>40 IU/l). In two cases, however, patients were included without fulfilling these criteria. The first one had an FSH level of 38 mUI/l but with a familial history of POF. The second patient had clinical symptoms suggesting Turner's syndrome (short size, bradymetacarpia, and multiple nevi) but with a normal karyotype. However, hypoestrogenism was associated with mild increase of FSH level (18 mUI/l). Since a mutation of one of the studied meiotic genes has been identified in this patient, we considered it necessary to still maintain the patient in our cohort. Karyotyping with a high-resolution GTG banding was carried out for all the patients. This study was approved by the institutional review board of the hospitals, and all participants gave their written informed consent. The control population provided by the Centre National of Genotypage (CNG) included 36 Caucasian women having at least one child and no history of infertility. A second group of control population originating from Senegal ($n=32$) was also tested for the tenth exon of *DMC1* gene.

DNA extraction and PCR

Genomic DNA was isolated from peripheral blood samples using the standard phenol–chloroform procedure. The *DMC1*, *MSH4*, *MSH5*, and *SPO11* genes are composed of 14, 20, 25, and 13 exons respectively. The sequencing project was performed at the CNG (Evry). All primers were designed using the software Primer 3 (36) (http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi). The first PCR amplification, using intronic oligonucleotide primers flanking the exons, was performed in a 15 μl volume containing 25 ng genomic DNA, 2.5 pM of each primer, and 0.75 U Taq polymerase (ExTaq, Takara, Cambrex, MD, USA). After an initial denaturation step at 94 °C for 5 min, 34 cycles of amplification were performed consisting of 5 s at 98 °C, 30 s at 60 °C, a 30 s elongation step at 72 °C, and one 10 min terminal elongation step. Primer sequences of *DMC1* and *MSH5* genes are given in Tables 1 and 2.

DNA sequencing and in silico analysis

All the PCR products containing the exons and flanking regions of each gene were purified using BioGel P-100 (Bio-Rad laboratories). To 2 μl sense or antisense sequencing primer (1.5 μM) and 3 μl Bigdye terminator mix (Applied Biosystems, Foster City, CA, USA), 1 μl purified PCR products was added. The amplification consisted of an initial 5 min denaturation step at 96 °C, 25 cycles of 10 s of denaturation at 96 °C, and a 4 min annealing/extension step at 60 °C. The purified reaction products (G50 Sephadex spin column, Boehringer Mannheim) were sequenced on an ABI PRISM 3700 DNA Analyzer (Applied Biosystems).

Both strands from all patients and controls were sequenced for the entire coding region and the exon/intron boundaries. Alignment and single nucleotide polymorphism (SNP) analysis were performed with Genalys software developed by the CNG (37).

The sequence of each variant was confirmed by a new round of PCR amplification and sequencing. The potential deleterious effect of the amino acid

Table 1 Sequences of PCR and sequencing primers for the human *DMC1* gene.

DMC1 gene	PCR primers	Sequencing primers
Exon 1	5′-TCCCAGGTTCAAGCGAT-3′ 5′-GCCATACCAGCTGTTAAG-3′	5′-TCAGGCATCTGTGTGCATGT-3′ 5′-GTAGCTAACAGGGAAGGAAC-3′
Exon 2	5′-TGAAATGAAATCAGAGGCCC-3′ 5′-GAAAAGCCTGTTGGTGGAAA-3′	5′-CAGCCCTTTCAATGTTGGTG-3′ 5′-TCAAAGGCTGGATTTCTGCC-3′
Exon 3	5′-TTTCCACCAACAGGCTTTTC-3′ 5′-ACCTGGAAGTTACTGCCCCT-3′	5′-CATTCTTGGGAAATCAGGGC-3′ 5′-CCAGGTCTCTTAATCCCTAC-3′
Exon 4	5′-CACTGTTGCATGTTTGACCC-3′ 5′-CTTGCTCCTCCAAGCAGTCT-3′	5′-TTGAACCTAGAAAGGGCAGC-3′ 5′-AATCTTGCTCCTCCAAGCAG-3′
Exon 5	5′-CAGCCAAGAATTGCTGTTCA-3′ 5′-GTGAAACCCCGTCTCCACTA-3′	5′-GGCATGCTATTTGTTCAGCC-3′ 5′-GCGAGACTCCGTCTCAAAAA-3′
Exon 6	5′-TGTAATCCCAGCTACTCAGG-3′ 5′-TGTAATCCCAGCTACTCAGG-3′	No forward sequencing primer 5′-TCAGGCACATAGTAGATGTTTG-3′
Exon 7	5′-GCAACAGCAGATTCCATGTG-3′ 5′-TTACCCAAACAGGTTCTGCC-3′	5′-CAACTATGCTGGCAGAATAC-3′ 5′-CCATATGAAGAAGTGAAACC-3′
Exon 8	5′-TGCAGGTGCACTTAGTTTGC-3′ 5′-CTTGAAGCCAGGAGTTGGAG-3′	5′-TGGTTGCTAGCATCCTCTAG-3′ 5′-TCTGCCTTAGCATGTATACC-3′
Exons 9+10	5′-GTAGCATTTGGTATACATGC-3′ 5′-AAGAGTTGTAAAGCCGGG-3′	5′-TATTTTGCCTGGCTCCCAAG-3′ 5′-CGCTGCCTCCTGACATTATA-3′
Exon 11	5′-ACTTTGCAGAGAAGCTTGG-3′ 5′-GCGCCCAGTAATAAAGTG-3′	5′-AGCCCGGCTTTACAACTCTT-3′ 5′-CGGAGTAGCTGAGATTACAG-3′
Exon 12	5′-GAGGTTGCAGTGAGTGAGAT-3′ 5′-GTTAGGGAAAGGTTCCCTGA-3′	5′-ACACAGCTAGACTCCATCTC-3′ No reverse sequencing primer
Exon 13	5′-CCTGTTTCCAAGTTTGGAGT-3′ 5′-GCCCAGCCCTGGAATTTT-3′	5′-GGCACATAATGCCTGTGACA-3′ 5′-CCAGCCCTGGAATTTTCATG-3′
Exon 14	5′-CCTGTTTCCAAGTTTGGAGT-3′ 5′-GCCCAGCCCTGGAATTTT-3′	5′-GTTGTTGGGAAAGGAGTACG-3′ 5′-AAGCACATGCCACTGCACTT-3′

changes was determined using PolyPhen software (http://tux.embl-heidelberg.de/ramensky/index.shtml). The multiple protein sequence alignment was realized with BioEdit and ClustalW (http://www.mbio.ncsu.edu/BioEdit/bioedit.html).

Results

Sequence analysis

The analysis of the coding sequence of *DMC1* revealed a homozygous substitution in the tenth exon of one case (patient A), g.33551A>G (with respect to the sequence AY520538 in Genbank; Fig. 1). This leads to the change in amino acid M200V. The patient A was of African origin (from Senegal, Sarakholé ethnic group). In order to determine the frequency of this genetic perturbation in a control population originating from Senegal, 32 additional DNA samples provided by Pasteur Institute (Dakar) were tested for exon 10. All individuals originated from the same geographic region and ethnic group (Sarakholé) as the patient's family.

Two DNA controls presented a heterozygous substitution at the same position. This variant frequency (3%) was comparable with those previously described in Genbank database. This variant is predicted to be probably damaging by the Polyphen program prediction (PSIC score difference=2.053). The familial analysis revealed that both parents and one sister are carriers of the same mutation with heterozygous status (Fig. 2A).

We have also detected, in the second exon of *MSH5*, a heterozygous transition g.2547C>T (with respect to the sequence AY943816 in Genbank) that altered codon 29 of the protein resulting in a proline-to-serine change (P29S). This mutation leads to the change from a medium size hydrophobic amino acid (P) to a small polar amino acid (S), and this variant is predicted to be possibly damaging by the Polyphen program prediction (PSIC score difference=1.800).

This variant was present in two POF patients (patients B and C). It was not found in any control ($n=36$). The sequencing of one patient's family (patient B) revealed the presence of the variant in the DNA of the father and the young sister (Fig. 2B).

The sequencing of *MSH4* and *SPO11* genes revealed no intragenic mutation. We detected only several SNPs present in similar frequency in patients and controls (data not shown).

Patient's phenotype

The mean patient age was 26.5 (15–39) years. The patients presented with the following clinical patterns: primary amenorrhea with absence of or interrupted puberty ($n=6$) and secondary amenorrhea with normal puberty ($n=23$). Eleven patients had a familial history of POF. Mean FSH level was 73.2 mUI/l (18–141).

DMC1-M200V

Patient A was a 28-year-old African woman. Puberty occurred normally when she was 15, with regular

Table 2 Sequences of PCR and sequencing primers for the human MutS homolog 5 (*MSH5*) gene.

MSH5 gene	PCR primers	Sequencing primers
Exon 1	Sense: 5′-ATGTCCCAGTAGGGGTGT-3′	5′-AATCAGCGTCCAGACTCTTC-3′
	Antisense: 5′-TGTGGACACAGGAGGTGA-3′	5′-AGATTGTGGGAAACTCCACG-3′
Exon 2	Sense: 5′-ATGAGGGTGGGGCGC-3′	5′-CCTCTGTGAATCGTTGCTTC-3′
	Antisense: 5′-TAGGCATCATCACCCCCA-3′	5′-GGCTCCCAACCCTCTTTTAT-3′
Exon 3	Sense: 5′-AGATTGCTCCACTGCACTTC-3′	5′-CTAAATGGGGGTGATGATGC-3′
	Antisense: 5′-GGTTGAGTCAGGAGAATTGC-3′	5′-GAGGAATTCATGGTTCCATC-3′
Exons 4+5	Sense: 5′-GAATCTGCCATCACGCCT-3′	5′-GAGGGCTATGGGTTTTCTCT-3′
	Antisense: 5′-CTGAGGCAGTGCCTTTTG-3′	5′-GGAACAGGGAGTTAGGCTAA-3′
Exons 6–8	Sense: 5′-ACTGCCTCAGTGACCCTT-3′	5′-TACAAGACCGTTCCCTTTGC-3′
		5′-AGCCCCCAGGAGATTTAAGA-3′
	Antisense: 5′-CCCCTTCCCTTTCCTTCA-3′	5′-CCACAACTCCACTTCCTTTG-3′
		5′-AGCATGCCTCCACCTCTTTA-3′
Exon 9	Sense: 5′-GTAATCCCAGCCACTCAGGA-3′	5′-AAAGACGTGATCTCAGGAGG-3′
	Antisense: 5′-ACAAGGTCTCCCAAAGTCCC-3′	5′-GGAGCCAATTGCTTTTCTGG-3′
Exon 10	Sense: 5′-CCTGTGAGTGTCCATCCCTT-3′	5′-AGCTTCCTCAACAACCAGCA-3′
	Antisense: 5′-AATCCAAGGTTCATGGCTTG-3′	5′-GAAATGCAGTTAGCCAGTGC-3′
Exons 11+12	Sense: 5′-CCTCAGAGTGAGCTGCAGTG-3′	5′-GTAACTTGTAGTACCCCCAC-3′
	Antisense: 5′-GTGTTGAAACTGCATGGTGG-3′	5′-GGCCTTTACCTGGACTTTTG-3′
Exons 13+14	Sense: 5′-TCTGTCTTCCTTCCTAGACTG-3′	5′-CTGTGATCTTCCCTACTGGT-3′
	Antisense: 5′-GACCACCTGCCAAGGATG-3′	5′-TGCCAAGGATGGTACTCCAT-3′
Exons 15–18	Sense: 5′-CGCAGTGATGGAGTACCAT-3′	5′-AGGGCAGGAGACTCACTTTT-3′
		5′-AAGTCCACAGCTTTGAACCC-3′
	Antisense: 5′-TTGGGCCCCTCATGTCTA-3′	5′-CATCACTCACCTTACAGAGG-3′
		5′-CTCATGTCTATTCCTCCACC-3′
Exons 19–21	Sense: 5′-TAGACATGAGGGGCCCAA-3′	5′-CTGGGGGTTCATCTATCTTG-3′
		5′-TCCTGTTTCACCCTGTCCAT-3′
		5′-TGCGTTACGGGCTTCCAATA-3′
	Antisense: 5′-CATATGCCCCTCTGCACT-3′	5′-CTCACTGTCATGCTCCTTCA-3′
Exons 22–25	Sense: 5′-GCTGTGTGGGCAGAAAAGAA-3′	5′-AATGCTAACCTCTGCCCTCT-3′
		5′-CTCCCACCTTCTTGCTTGTT-3′
	Antisense: 5′-TACTGAGGCAGGGCAGGT-3′	5′-GGTGGTTGCACATTTGGATC-3′
		5′-CCTGCTCTGTGTTTTGGATC-3′

menstrual cycles up to 21 years. A secondary amenorrhea occurred definitely since then. She was referred to our department when she was 28. POF was confirmed with a high FSH level (91 mUI/l); estradiol and inhibin B levels were low (<10 pg/ml and 15 ng/ml respectively). No anti-ovarian antibodies were found positive. Pelvic ultrasonography showed a small uterus (50 mm in its maximal length) and small-sized ovaries; no follicle was observed. An ovarian biopsy was performed confirming the follicular depletion. Familial study identified one sister with a long-standing history of infertility with eight spontaneous abortions.

MSH5-P29S

The first heterozygous patient (patient B) was a Caucasian woman who was 18 years old when she was referred to our department. She first had menstruations when she was 14 with a normal puberty. However, an oligomenorrhea and a secondary amenorrhea appeared progressively. Clinically, she was short (1.46 m) and associated with an obesity (BMI: 32); had rough face, bradymetacarpia, and a mild intellectual deficiency. Hormonal evaluation identified a mild increase of FSH level (18 mUI/l); inhibin B was low (12 pg/ml). Ten days of progestin treatment induced vaginal bleeding. Ultrasonography identified two ovaries that are small in size with multiple follicles depicted. Turner's syndrome was suggested but high-resolution karyotype was found normal and repeated twice. Since the syndrome appeared uncommon, an ovarian biopsy was performed, identifying multiple primary follicles; a secondary follicle was also observed.

The second heterozygous patient (patient C) was also a Caucasian woman who was 36 years old when she was referred to our department. She had her first menstruations, associated with a normal pubertal development, when she was 13. She had oligomenorrhea between 13 and 18 years of age and then used oral contraceptive pills until she was 30. She became pregnant 6 months later and gave birth to a normal boy. She had menstruations following this but a secondary amenorrhea occurred when she was 32. Hormonal results confirmed the existence of POF with a high FSH level (71 mUI/l). Pelvic ultrasonography identified two ovaries small in size without follicles. An ovarian biopsy was performed, depicting small streak gonads with a complete follicular depletion. Table 3 showed major characteristics of these three patients with mutation in meiotic gene.

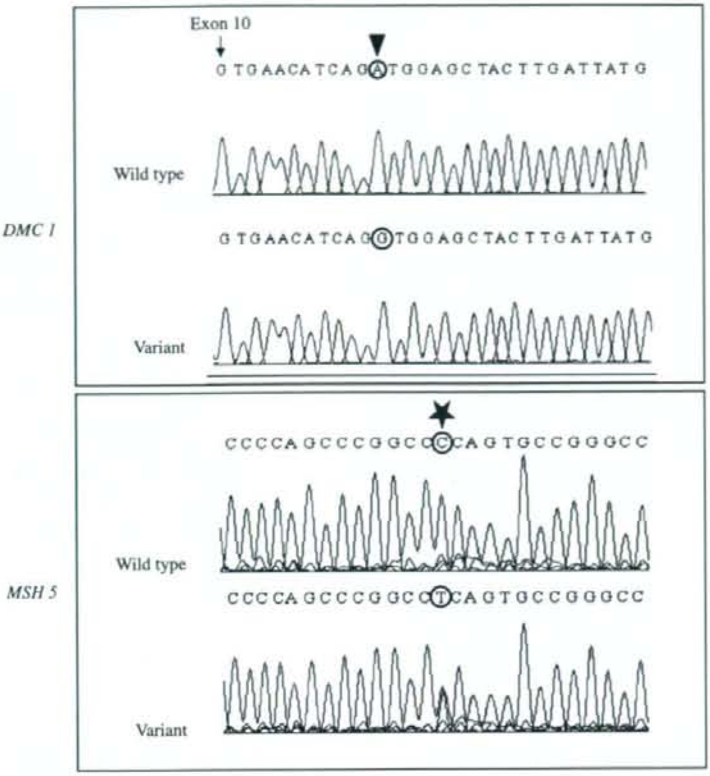


Figure 1 Analysis of *DMC1* and *MSH5* coding sequences. Electropherogram showing (above) the sequence of exon 10 *DMC1* variant in comparison with the sequence of wild type and (below) the sequence of a part of the exon 2 *MSH5* variant in comparison with the sequence of wild type. The arrowhead indicates nucleotide 12 of exon 10 with the A>G homozygous substitution. The star indicates nucleotide 97 of exon 2 with the C>T heterozygous substitution.

Discussion

The POF syndrome is a very heterogeneous clinical disorder probably due to the complex genetic networks controlling human oogenesis and folliculogenesis. It is often associated with small pedigrees that make it difficult to perform genetic linkage analysis to identify responsible genes. An alternative approach is to test candidate genes on the basis of existing knowledge of ovarian physiology.

Several meiotic genes known in yeast have been isolated in mammals, including *Dmc1*, *Msh4*, *Msh5*, and *Spo11* genes (38–41).

Dmc1 is important for meiotic recombination in many organisms; for example, mice with targeted mutations of the *Dmc1* gene are sterile and show hallmarks of poorly repaired DNA double-strand breaks. At birth, the mutant ovaries formed in *Dmc1* −/− mice contained a high proportion of oocytes whose nuclear features were characteristic of leptonema or zygonema, in contrast to the littermates, in which the majority of oocytes had progressed to the pachytene stage. Histological analysis showed that the adult ovaries from *Dmc1* −/− deficient mice contained no follicle at any developmental stage (33, 34). These results indicate that while germ cells are indeed formed in *Dmc1* −/− ovaries, the meiotic progression is blocked leading to progressive death of oocytes and subsequent complete depletion in the ovary by adulthood. The description of our clinical case is perfectly compatible with the animal model. Patient A had a normal gonadal function during a few years, which disappeared when she was 21. Since then, ovarian description, either by ultrasonography or histology, showed a complete absence of follicular reserve and/or maturation.

Msh5 is a member of a family of proteins known to be involved in DNA mismatch repair (42). *Msh5* −/− mice are viable but sterile (35). Meiosis in these mice is affected due to the disruption of chromosome pairing in prophase I. The ovaries of *Msh5* −/− females are normal in size at birth, but degenerate progressively to become rudimentary (35). The phenotype of *Msh5* −/− females differs from *Dmc1* −/− mice, in that the few oocytes remaining at 4–5 day pp in *Msh5* −/− ovaries are normal in appearance and have formed follicles (43). In contrast, ovaries from *Dmc1* knockout females were very small and contained no follicle at any developmental stages. A less severe

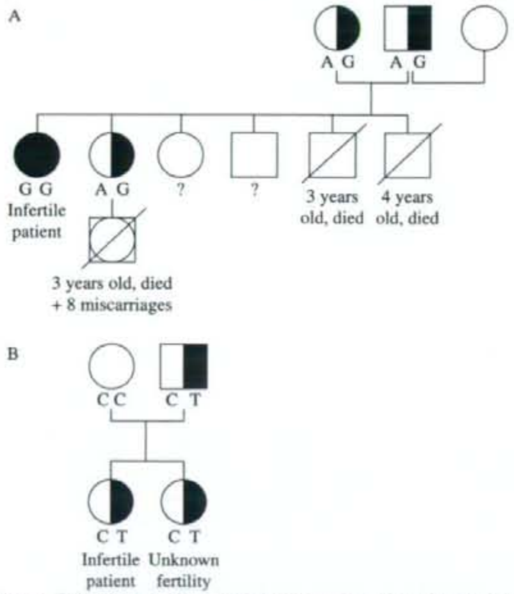


Figure 2 (A) Pedigrees of the infertile patient's family (patient A) with *DMC1* homozygous mutation. Circles, females; squares, males; and slashes through symbols, deceased. (B) Pedigrees of the infertile patient's family (patient B) with *MSH5* variant. Circles, females and squares, males.

phenotype for *Msh5* versus *Dmc1* mutation has also been observed in yeast and worms (44, 45).

This type of phenotype is in accordance with those observed in POF patients with a progressive loss of activity of the ovary leading to gonads reduced in size without germ cells; the oocytes having failed to progress to the dictyate stage *in utero* and subsequently degraded. Nevertheless, the MSH5 protein of the POF patients is probably not completely defective and phenotypes could be less severe than those observed in null mice. However, in our patients, only heterozygous *MSH5* mutation was described. The interspecific genotype difference (heterozygous in humans and homozygous in mice) could be explained by a more dosage-sensitive system in humans.

In both cases with *MSH5* alteration, gonadal function appeared normal with a progressive involution. The most surprising data concern the youngest woman who presented with syndromic features. However, similarities in the ovarian phenotype in female *Msh5−/−* mice and Turner's syndrome patients have been reported (35). It is also probable that this young woman will present with a complete POF in the near future. Indeed, the variability observed in clinical phenotypes (complete or partial infertility) could result from the age that the patients consult with clinicians. The sequencing of *MSH5* gene in the family of this patient revealed that her 20-year-old young sister was also a carrier of the same variant. Until now, she had normal menses but she could be considered as a carrier of a genetic predisposition to develop POF in the future.

The resulting P29S alteration within MSH5 protein is located within the N-terminal region and it is conceivable that this amino acid change could directly impact the integrity of the protein interaction between MSH5 and MSH4. Amino acid sequence analysis revealed that the MSH5 N-terminal region contains a contiguous (Px)5 dipeptide repeat flanked by two PxxP motifs (46). This dipeptide repeat is disrupted in the *MSH5* P29S variant. Moreover, this same mutation has previously been described in genomic DNA of patients with ovarian carcinoma (47). To address the effect of P29S alteration on the interaction between MSH4 and MSH5, a quantitative two-hybrid analysis has been performed. This alteration causes moderate but significant reduction between both proteins and could affect the formation of MSH4–MSH5 heterocomplex (47). The functionality of both proteins in meiotic homologous recombination process probably needs a precise interaction between them and any deviation from this precise coordination will be expected to affect the accuracy of DNA recombination. It is noteworthy that this alteration was found in two patient populations with ovarian pathology; the previous with ovarian cancer and the present with POF. These two pathologies could affect the capacity of DNA repair leading to either a progressive loss of germ cells or cancer formation. For this reason, it will be crucial to follow the degeneration of the ovary from our two patients on a long-term period to prevent an eventual ovarian carcinoma.

In summary, one homozygous missense mutation in *DMC1* gene and one heterozygous in *MSH5* were described here, in 3 of 41 POF patients. The *DMC1* M200V mutation seems to generate a deleterious effect only in homozygous states since the mother and the

Table 3 Major characteristics of the patients with mutation in meiotic genes.

Patient	Age (years)	Origin	FSH levels (mUI/l)	Follicle	Mutated gene
A	28	African	91	None	*DMC1:* g.33551A>G Homozygous mutation
B	18	Caucasian	18	Multiple primary, 1 secondary	*MHS5:* g.2547C>T Heterozygous mutation
C	36	Caucasian	71	None	*MHS5:* g.2547C>T Heterozygous mutation