Am. J. Hum. Genet. 65:183–191, 1999

Sequence Diversity in 36 Candidate Genes for Cardiovascular Disorders

François Cambien,[1,2] Odette Poirier,[1,2] Viviane Nicaud,[1] Stefan-Martin Herrmann,[1,2] Christine Mallet,[2] Sylvain Ricard,[2] Isabelle Behague,[2] Vincent Hallet,[2] Hervé Blanc,[2] Valérie Loukaci,[2] Joëlle Thillet,[3] Alun Evans,[4] Jean-Bernard Ruidavets,[5] Dominique Arveiler,[6] Gérald Luc,[7] and Laurence Tiret[1]

[1]INSERM U525 and [2]INSERM SC7 and [3]INSERM U321, Paris; [4]Belfast MONICA Project, Department of Epidemiology and Public Health, the Queen's University of Belfast, Belfast; [5]Projet MONICA Haute-Garonne, Toulouse; [6]Projet MONICA Bas-Rhin, Strasbourg; and [7]SERLIA-INSERM U325, Lille

Summary

Two strategies involving whole-genome association studies have been proposed for the identification of genes involved in complex diseases. The first one seeks to characterize all common variants of human genes and to test their association with disease. The second one seeks to develop dense maps of single-nucleotide polymorphisms (SNPs) and to detect susceptibility genes through linkage disequilibrium. We performed a molecular screening of the coding and/or flanking regions of 36 candidate genes for cardiovascular diseases. All polymorphisms identified by this screening were further genotyped in 750 subjects of European descent. In the whole set of genes, the lengths explored spanned 53.8 kb in the 5′ regions, 68.4 kb in exonic regions, and 13 kb in the 3′ regions. The strength of linkage disequilibrium within candidate regions suggests that genomewide maps of SNPs might be efficient ways to identify new disease-susceptibility genes, provided that the maps are sufficiently dense. However, the relatively large number of polymorphisms within coding and regulatory regions of candidate genes raises the possibility that several of them might be functional and that the pattern of genotype-phenotype association might be more complex than initially envisaged, as actually has been observed in some well-characterized genes. These results argue in favor of both genomewide association studies and detailed studies of the overall sequence variation of candidate genes, as complementary approaches.

Received November 25, 1998; accepted for publication May 4, 1999; electronically published June 3, 1999.

Address for correspondence and reprints: Dr. Laurence Tiret, INSERM U525, 17 rue du Fer à Moulin, 75005 Paris, France. E-mail: tiret@idf.inserm.fr

0002-9297/99/6501-0024$02.00

Introduction

In recent years, there has been considerable effort to understand the contribution of genetic factors to human disease. Although success has been undeniable in the discovery of genes underlying Mendelian disorders, the identification of genes involved in common diseases appears to be far more complicated than initially anticipated (Collins et al. 1997). It has been progressively realized that linkage studies, the strategy successfully applied to Mendelian disorders, has low power to detect genes with modest effects, such as those involved in complex diseases, and it is now often proposed that association studies might be more efficient for identification of susceptibility genes underlying common disorders (Lander 1996; Risch and Merikangas 1996; Cambien et al. 1997; Collins et al. 1997).

A current limitation of association studies is that they are restricted to already known candidate genes. Major recent developments, however, have led to the consideration of whole-genome association studies. Two different strategies have been proposed (Collins et al. 1997). The first one is linked to the completion of the Human Genome Project, aimed at producing the complete nucleotide sequence of the ~100,000 human genes. One of the goals proposed for the postgenome phase is the systematic identification of all common variants in human genes (Lander 1996). By cataloging all common variants, it would be possible to test directly the association of each of them with disease.

The second strategy involves the use of very dense maps of single-nucleotide polymorphisms (SNPs) evenly distributed over the whole genome so that every gene would be covered or narrowly flanked by one or a few SNPs, allowing indirect detection of its effects through linkage disequilibrium between SNPs and a functional variant of the gene (Risch and Merikangas 1996; Collins et al. 1997).

There is limited information about the natural sequence diversity in human genes that could help to identify the most efficient strategy. The choice of a strategy

is dependent on a number of critical elements, including the number of functional polymorphisms within a gene, the combination of their effects, their allele frequencies, and the extent of linkage disequilibrium among them. Examples of some well-characterized genes—such as the angiotensin I–converting enzyme (*ACE*) (Villard et al. 1996), apolipoprotein (*apo*) E (Davignon et al. 1988; Lambert et al. 1998), or cholesteryl ester transfer protein (*CETP*) (authors' unpublished data) genes—indicate that the presence of several functional polymorphisms within a gene is likely and that the pattern of genotype-phenotype association can be quite complex, owing, in particular, to considerable linkage disequilibrium between polymorphisms.

We have recently undertaken a program of exploration of candidate genes for cardiovascular diseases, and, as a first step of this ongoing program we have performed an extensive molecular screening of the coding and/or flanking regions of 36 genes. All polymorphisms identified by this screening were further genotyped in 750 European population-based control subjects of the ECTIM (Etude Cas-Témoin sur l'Infarctus du Myocarde) study (Parra et al. 1992). The statistics derived from the study of these 36 genes provide new insights into the type and amount of DNA-sequence variation that might be expected in human genes.

Material and Methods

The ECTIM Population-Based Samples

The ECTIM study (Parra et al. 1992) is a case-control study of myocardial infarction conducted in four regions covered by WHO MONICA (monitoring in cardiovascular disease) registers: Belfast (Northern Ireland), Lille (northern France), Strasbourg (eastern France), and Toulouse (southern France). In each region, a population-based control group was composed of men of age 25–64 years (mean age 53.3 ± 8.5 years) randomly sampled from the population covered by the register. To increase ethnic homogeneity, the subjects had to be residents of the region where they were recruited, their parents had to have been born in this same region, and their four grandparents had to have been born in Europe. Informed consent was obtained from all participants.

Screening of the Genes and Genotyping of Polymorphisms

Genomic DNA was prepared from white blood cells by phenol extraction. The molecular screening of the genes was performed by comparing 40 chromosomes from 20 unrelated patients, each of whom had suffered a myocardial infarction. DNA-sequence variations were identified by PCR/SSCP (Orita et al. 1989). From the known (published or identified) sequences of genes, overlapping fragments were amplified enzymatically to cover the entire coding sequence and the flanking regions. The mean size of PCR fragments used for the initial SSCP was 254 bp (SE 67; range 87–478). Most (95%) of the fragments were <350 bp in size.

Each amplification was performed by use of 250 ng of DNA in a total volume of 50 μl, containing 10 mmol Tris-HCl (pH 9)/liter, 50 mmol KCl/liter, 1.5 mmol $MgCl_2$/liter 0.1% Triton X-100, 0.2 mg BSA/ml, 200 μmol each dNTPs/liter, 25 pmol of each amplimer/liter, and 0.2 U *Taq* polymerase. For the SSCP analysis, 0.12 μCi of α[-^{32}P]-dCTP were added to the 50-μl mix. Amplification products were then diluted twofold in a solution containing 95% formamide, 10 mmol EDTA/liter, 0.05% bromophenol blue, and 0.05% xylene cyanol. After denaturation at 95°C for 5 min, the samples were placed on ice, and 4 μl were loaded onto nondenaturing 8% acrylamide gels (acrylamide:bisacrylamide ratio 39:1). Electrophoresis was performed, at room temperature, with a cooling fan and 40 W constant power for 6 h on gels containing 0% and 7.5% glycerol. Thereafter the gels were dried and autoradiographed overnight at −80°C, with an intensifying screen.

DNA from subjects presenting a different SSCP pattern of migration was reamplified by PCR with unlabeled primers. PCR products were then purified and sequenced by the Sanger method (Sanger et al. 1997) in 25 cycles of PCR with γ[-^{32}P]-dATP end-labeled primer, by a direct sequencing kit (AmpliCycle; Perkin-Elmer, Roche Molecular Systems). The method of detection by PCR/SSCP is not 100% sensitive but >90% of already known polymorphisms were actually detected, which provides an estimate of the sensitivity of the technique used.

Polymorphisms were then genotyped in the 750 population-based controls of the ECTIM study, by allele-specific oligonucleotides (ASOs) (Saiki 1986). After enzymatic amplification of the fragment encompassing the polymorphism, 1/5 of the PCR product was denatured in 150 μl of 0.5M NaOH and 1.5 M NaCl with 10 μl of 0.05% bromophenol blue solution and was blotted onto nylon membranes (Hybond N+; Amersham). Each allele was detected after preincubation of the membranes, for 2 h with 100 pmol of unlabeled oligonucleotide probe specific for the other allele, followed by incubation for 2 h with 20 pmol of the labelled probe specific for the allele. The melting temperature (T_m) used for hybridization was calculated by adding 4°C for each C or G and 2°C for each A or T and then subtracting 5°C from the total. The membranes were washed twice at room temperature in 1 × SSC for 5 min, followed by 5 min in 0.5 × SSC at T_m − 3°C.

All information needed for genotyping polymorphisms (PCR primers, probes, and temperatures) can be obtained at our Internet site, Canvas (see Electronic-

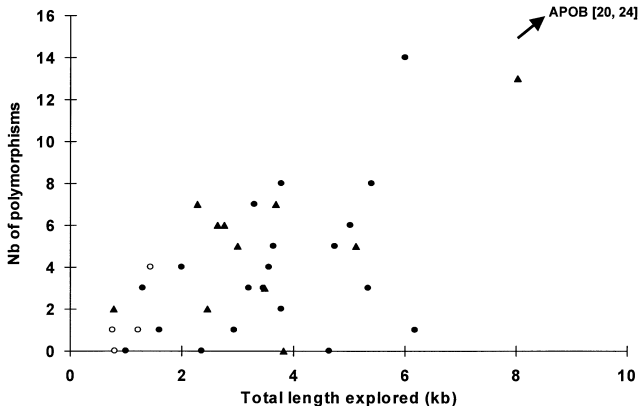

Figure 1 Plot of number of detected polymorphisms versus length of scanned regions. Blackened circles denote genes in which 5′, exonic, and 3′ regions were scanned ($n = 21$); triangles denote genes in which 5′ and exonic regions were scanned ($n = 11$); unblackened circles denote genes in which only the 5′ region was scanned ($n = 4$).

Database Information), except for the MSR1 gene, for which the gene sequence was confidentially communicated to us. The gene abbreviations used in this paper are taken from OMIM (see Electronic-Database Information).

Statistical Analysis

All population-related statistics were estimated in the 750 population-based controls from the ECTIM study. Since, for the vast majority of polymorphisms, allele frequencies and pairwise linkage-disequilibrium coefficients did not statistically differ between populations, subjects from the four populations were pooled for analysis. The statistics computed were (i) the lengths of the regions explored in each gene; (ii) the number of polymorphic sites found; (iii) the population allele frequencies of polymorphisms; (iv) the heterozygosity (h) estimated at each single varying site; (v) the nucleotide diversity (π), defined as the expected number of differences, per nucleotide site, between a random pair of chromosomes drawn from the population, a number that is equivalent to heterozygosity at the nucleotide level, averaged across all nucleotide sites, including monomorphic sites; (vi) the overall heterozygosity (H) yielded by the whole set of polymorphisms identified in a gene, which is the proportion of heterozygotes per gene locus when all polymorphisms across the gene are combined into haplotypes, as computed from haplotype frequencies estimated by the MYRIAD program (MacLean and Morton 1985); and (vii) the pairwise linkage-disequilibrium coefficients (D') between diallelic polymorphisms, estimated by log-linear model analysis (Tiret et al. 1991). The extent of disequilibrium was expressed as the ratio of the unstandardized coefficient to its maximal/minimal value. For each gene, the mean linkage disequilibrium was calculated by averaging the $k(k-1)/2$ pairwise disequilibrium coefficients (k = no. of diallelic polymorphisms within a gene). An association between two diallelic polymorphisms was considered as complete or nearly complete if <10% of subjects were off the diagonal of the 3 × 3 contingency table of genotypes observed in the 750 controls.

Results

Sequence Diversity in the 36 Genes

The investigated genes were selected for their possible involvement in common cardiovascular disorders such as coronary heart disease or hypertension. They code for growth factors, for cytokines, for neuromediators, and for molecules involved in lipid metabolism, calcium metabolism, vascular and cardiac trophicity, thrombosis, and adhesion (table 1). The screening of the genes was performed by comparing 40 chromosomes from 20 unrelated patients with myocardial infarction.

The gene sequences were divided into three regions: 5′ flanking (upstream from the transcription start), exonic (including untranslated exons), and 3′ flanking. In the whole set of genes, the total lengths explored spanned 53.8 kb in the 5′ flanking regions, 68.4 kb in the exonic regions, and 13 kb in the 3′ flanking regions. The mean length explored in a gene was 3.8 kb, with a large range of variation: from 750 bp in the FGG gene to nearly 20 kb in the APOB gene. In 4 of the 36 genes (LPA, FGA, FGG, and VDR), only the 5′ region was explored in a systematic way.

The total number of identified polymorphic sites was 164 (mean 4.6 [range 0–24]). The largest number of polymorphisms ($n = 24$) was in the APOB gene (Poirier et al. 1996). In 5 of the 36 genes (FGA, ADRB1, CSF1, ECE, and HGF), no polymorphic site was found in the regions explored (table 1). As expected, the number of polymorphisms in a gene was positively correlated with the length of the scanned sequences (Spearman correlation coefficient [r] = 0.48; $P = .003$) (fig. 1).

Allele frequencies of the polymorphisms were estimated in 750 subjects from the ECTIM study. Sixteen of the polymorphisms were rare variants (frequency <.01), and 10 were multiallelic polymorphisms, the remainder being common diallelic polymorphisms (SNPs or insertion/deletion variations). For polymorphisms whose minor-allele frequency was in the range .01–.10, the observed number was corrected for their probability of detection, by comparison of 40 chromosomes. Polymorphisms with a higher allele frequency had a probability of almost 1 of being detected by this approach. For rare variants (frequency <.01), no attempt to correct their number was made, because accurate correction

Table 1

Lengths Explored and Number of Polymorphisms Found in 36 Genes

Gene	Abbreviation	Chromosome	Length Explored [No. of Polymorphisms Found][a]							
			5′ Regions		Exonic Regions		3′ Regions		Total	
α_{2A} Adrenergic receptor	ADRA2A	10q	2,041	[3]	1,350	[1]	173	[0]	3,564	[4]
Aldosterone synthase	CYP11B2	8q	660	[2]	2,097	[4]	0	[…]	2,757	[6]
Angiotensin-II type 1 receptor	AGTR1	3q	2,558	[8]	2,606	[6]	836	[0]	6,000	[14]
Apolipoprotein (a)	LPA	6q	1,441	[4]	0	[…]	0	[…]	1,441	[4]
Apolipoprotein B	APOB	2p	3,656	[4]	15,803	[19]	300	[1]	19,759	[24]
β_1 adrenergic receptor	ADRB1	10q	3,097	[0]	1,444	[0]	93	[0]	4,634	[0]
Cholesteryl ester transfer protein	CETP	16q	805	[2]	1,590	[3]	911	[2]	3,306	[7]
Colony stimulating factor 1	CSF1	5q	748	[0]	225	[0]	34	[0]	1,007	[0]
Early growth response protein 1	EGR1	5q	918	[1]	1,632	[0]	1,228	[1]	3,778	[2]
Endothelin 1	EDN1	6p	2,955	[3]	639	[2]	48	[0]	3,642	[5]
Endothelin converting enzyme	ECE	1p	219	[0]	3,600	[0]	0	[…]	3,819	[0]
Endothelin-A receptor	EDNRA	4	858	[0]	4,122	[5]	38	[1]	5,018	[6]
Endothelin-B receptor	EDNRB	13q	1,001	[1]	4,307	[2]	36	[0]	5,344	[3]
Fibrinogen α	FGA	4q	800	[0]	0	[…]	0	[…]	800	[0]
Fibrinogen β	FGB	4q	1,500	[5]	1,395	[2]	2,500	[1]	5,395	[8]
Fibrinogen γ	FGG	4q	750	[1]	0	[…]	0	[…]	750	[1]
Granulocyte-macrophage colony–stimulating factor	CSF2	5q	663	[0]	435	[1]	503	[0]	1,601	[1]
Hepatocyte growth factor	HGF	7q	86	[0]	2,190	[0]	75	[0]	2,351	[0]
Insulin-like growth factor-1	IGF1	12q	1,630	[2]	927	[1]	648	[0]	3,205	[3]
Intercellular circulating adhesion molecule-1	ICAM1	19p	1,392	[1]	1,599	[4]	0	[…]	2,991	[5]
Interleukin 1-α	IL1A	2q	1,437	[1]	2,021	[2]	0	[…]	3,458	[3]
Interleukin 6	IL6	7p	1,158	[3]	609	[0]	237	[1]	2,004	[4]
Macrophage-scavenger receptor	MSR1	8p	2,345	[2]	117	[0]	0	[…]	2,462	[2]
Matrix Gla protein	MGP	12p	3,200	[6]	311	[2]	269	[0]	3,780	[8]
Microsomal triglyceride transfer protein	MTP	4q	732	[2]	52	[0]	0	[…]	784	[2]
NO synthase endothelial	NOS3	7q	1,021	[3]	1,614	[3]	0	[…]	2635	[6]
NO synthase inducible	NOS2A	17q	1,090	[3]	191	[0]	10	[0]	1,291	[3]
P-selectin	SELP	1q	4,866	[5]	3,152	[8]	0	[…]	8,018	[13]
Platelet-derived growth factor A	PDGFA	7p	1,787	[0]	645	[1]	2,305	[4]	4,737	[5]
Platelet-derived growth factor A-receptor A	PDGFRA	4q	2,122	[6]	160	[1]	0	[…]	2,282	[7]
Scavenger receptor	CD36	7q	273	[0]	2,641	[1]	25	[0]	2,939	[1]
Thrombin receptor	TR	5q	225	[1]	1,278	[0]	1,970	[2]	3,473	[3]
Transforming growth factor β1	TGFB1	19q	1,138	[3]	2,533	[4]	0	[…]	3,671	[7]
Tumor necrosis factor α	TNF	6p	1,053	[5]	4,066	[0]	0	[…]	5,119	[5]
Vascular-cell adhesion molecule-1	VCAM1	1p	2,355	[1]	3,064	[0]	766	[0]	6,185	[1]
Vitamin D receptor	VDR	12q	1,228	[1]	0	[…]	0	[…]	1,228	[1]
Total			53,808	[79]	68,415	[72]	13,005	[13]	135,228	[164]

[a] An ellipsis (…) indicates that the no. of polymorphisms was not determined.

would be impossible. As a consequence, these variants were not included in the estimation of sequence diversity. These variants, although of interest from a population-history aspect, are unlikely to contribute substantially to the genetic components of common diseases (Cambien et al. 1997).

After correction for the probability of detection, the expected number of polymorphisms, excluding rare variants, was 164 ($n = 79$ in 5′ regions, $n = 71$ in exonic regions and $n = 14$ in 3′ regions), coincidentally equaling the total number of observed polymorphisms. The number of polymorphisms decreased as their allele frequency increased in the population (fig. 2). Nearly one-third of polymorphisms had a frequency <0.1.

The mean ± SE h estimated across all polymorphic sites (when rare variants were excluded) was .301 ± .014. The mean ± SE π estimated from the pooled genes was .00037 ± .00005, which means that two randomly chosen sequences from the population are expected to differ approximately every 2,700 bp. The mean ± SE π was 0.00044 ± 0.00008 in the 5′ and 3′ regions (1 variable site/2,300 bp) and .00030 ± .00007 in the exonic regions (1 variable site/3,400 bp). The lower sequence diversity observed in exonic regions versus flanking regions of the 36 genes presumably reflects a more stringent selective pressure in coding versus noncoding sequences.

Type of Polymorphism

In accordance with previous observations (Li and Sadler 1991; Kwok et al. 1996; Nickerson et al. 1998), the vast majority of varying sites (85%) were single-base substitutions. Although transversions are theoretically twice as frequent as transitions, transitional changes actually occurred more often than transversional changes, as already reported (Vogel and Kopun 1977; Li et al. 1984; Collins and Jukes 1994). In the 36 genes, the transition rate exceeded the transversion rate by a ratio of 2.3:1. The relative number of transitions/transversions and the type of substitution were not significantly different among regions of the genes. The frequencies of the different substitutions were A/G (39.3%), C/T (30.0%), G/C (12.9%), G/T (7.9%), A/C (7.1%), and A/T (2.9%). The higher prevalence of A/G substitutions is in accordance with previous observations in functional genes (Gojobori et al. 1982). Among the 164 polymorphisms identified, 13 (7.9%) were insertion/deletion variations and 10 (6.1%) were repeat polymorphisms. All insertion/deletion variations (with one exception, which affected the coding sequence of the signal peptide of the APOB gene) and all repeat polymorphisms were found in untranslated sequences (5′ and 3′ flanking regions or untranslated exons).

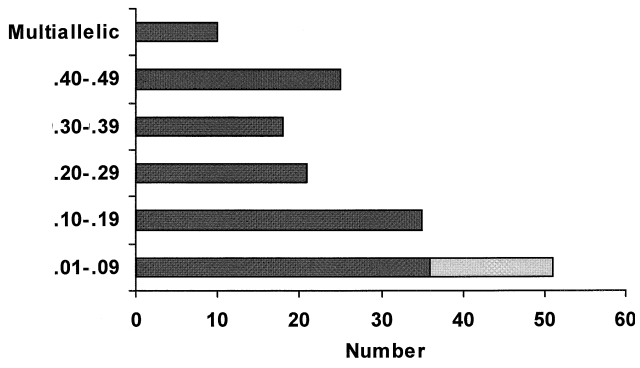

Figure 2 Distribution of polymorphisms according to allele frequency estimated in 750 subjects. The light gray shading indicates the frequency corrected for the probability of detection by using 40 chromosomes. Rare variants (frequency <.01) are not represented because of the lack of precision of correction.

Seventy-two polymorphic sites were identified in exonic regions, among which 12 (17%) occurred in untranslated sequences. Among the 60 remaining polymorphisms, 40 (67%) were nonsynonymous (i.e. they led to an amino acid change). The proportion of nonsynonymous mutations was not significantly different from the 71% proportion expected by chance if all codons were assumed to be equally frequent in the genome and if the probability of substitution was the same for all pairs of substitutions (Nei 1987). All substitutions occurring at the first nucleotide position ($n = 18$) were nonsynonymous, as compared with a proportion of 95% expected under random substitution. At the second nucleotide position, all substitutions were nonsynonymous by definition. At the third nucleotide position, 20 (87%) of the 23 substitutions were synonymous, a proportion not significantly different from the 72% expected by chance. This result suggests, at most, only a weak selective pressure acting on common polymorphisms affecting coding regions of candidate genes for complex diseases. This might be because, unlike rare genetic defects involved in monogenic diseases, common polymorphisms that influence human health and disease susceptibility have no dramatically deleterious effect. Moreover, in most cases their detrimental effect would appear relatively late in life.

Linkage Disequilibrium and H

The mean pairwise linkage disequilibrium could be estimated in 22 of the 36 genes. Among the 14 remaining genes, 5 were not polymorphic (FGA, ADRB1, CSF1, ECE, and HGF), 5 had only one polymorphism (CD36, FGG, CSF2, VCAM1, and VDR), and in the remaining 4 genes (IGF1, NOS2A, MSR1, and PDGFA), the polymorphisms involved were too rare (<.01) or multiallelic (repeat polymorphisms). In the 22 genes, the mean

± SE pairwise disequilibrium was .86 ± .03 (fig. 3). In 18 of the 22 genes, the mean linkage disequilibrium was >0.80, reflecting a strong nonrandom association among polymorphisms within candidate regions. There were, however, a few exceptions, in which polymorphisms within a gene exhibited little association. As an example, in the AGTR1 gene, a group of polymorphisms located in the 5′ flanking region was in weak linkage disequilibrium with a group of polymorphisms located in the exonic and 3′ flanking regions, probably because the two groups are separated by one large intron (Poirier et al. 1998).

The mean ± SE of *H* conveyed by the whole set of polymorphisms within a gene was .48 ± .05, and reached .57 ± .04 after exclusion of the nonpolymorphic genes. There was a strong negative correlation between the mean pairwise linkage disequilibrium and *H* ($r = -.72$, $P < .001$) (fig. 3). This correlation reflects that the stronger the nonrandom association between polymorphisms, the lower the information added by each polymorphism to the set of the others. The high *H* observed in several genes (>0.7 in 10 of the 36 genes) has potential implications for linkage studies, since it may be more efficient to combine several highly informative SNPs within a candidate region than to use a microsatellite.

Complete or Nearly Complete Association Between Polymorphisms

Complete or nearly complete association between polymorphisms was frequently observed (i.e., genotypes at different sites were almost completely concordant and generated only two main haplotypes). Such a pattern of nearly complete association could extend over several polymorphic sites within a gene. For example, in the APOB gene (Poirier et al. 1996), we observed three pairs of polymorphisms in nearly complete association, as well as one triplet (fig. 4). An extreme example was found in the FGB gene, where six polymorphisms were in nearly complete association at both ends of the gene and generated two main haplotypes (Behague et al. 1996).

In the total of 36 genes, we observed 10 combinations of two polymorphisms (in the APOB, CETP, EDN1, EDNRA, IL1A, IL6, and TGFB1 genes), 3 combinations of three polymorphisms (in the APOB, NOS3, and SELP genes), 1 combination of four polymorphisms (in the CYP11B2 gene), and 3 combinations of six polymorphisms (in the AGTR1, FGB, and PDGFRA genes) in nearly complete association. Such associations among polymorphisms over long distances suggest the existence of ancestral haplotypes and raise important questions about both the population-historic processes that generated these haplotypes and the potential functionality of the haplotypes.

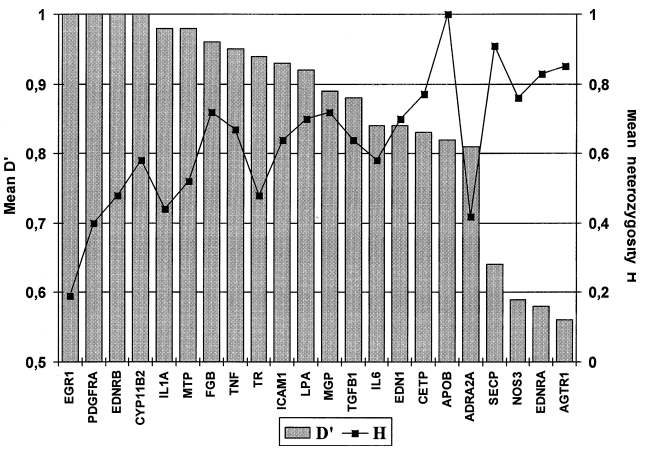

Figure 3 Mean *D*′ between diallelic polymorphisms and overall heterozygosity (*H*), conveyed by the whole set of polymorphisms within a gene, estimated in 750 subjects.

Discussion

The present study, based on a systematic screening of 36 candidate genes for cardiovascular diseases, provides an estimate of sequence diversity in human genes from European populations. Because rare variants were not included in the calculation of nucleotide diversity, the figures reported here represent a lower bound of the actual DNA-sequence diversity in the explored genes. However, it should be noted that the contribution of rare variants to nucleotide diversity is rather small. For example, 9 variants, each with a frequency of .01, would have roughly the same contribution as one single polymorphism with an allele frequency of .10. The nucleotide diversity estimated from the present data was of the same order of magnitude as that obtained from an analysis comparing published human DNA sequence data from 49 different loci (Li and Sadler 1991). By contrast, a much higher nucleotide diversity (.002) was recently reported across a region of the lipoprotein lipase gene, in a study comparing sequence data from 71 individuals from three populations (Nickerson et al. 1998). The nucleotide diversity was very similar in the three populations, despite their different demographic histories (African Americans, Europeans, and European Americans). Several reasons might explain the difference observed for the nucleotide diversity estimated in the present study. First, 90% of the sequence screened in the lipoprotein lipase gene was located in introns, in which, as expected, the sequence diversity was found to be fourfold higher than that in coding regions. Second, most of the varying sites in the lipoprotein lipase gene were rare variants found in one or two individuals, whereas, in our estimation of nucleotide diversity, we did not include such variants. Third, our study of 36 genes indicates

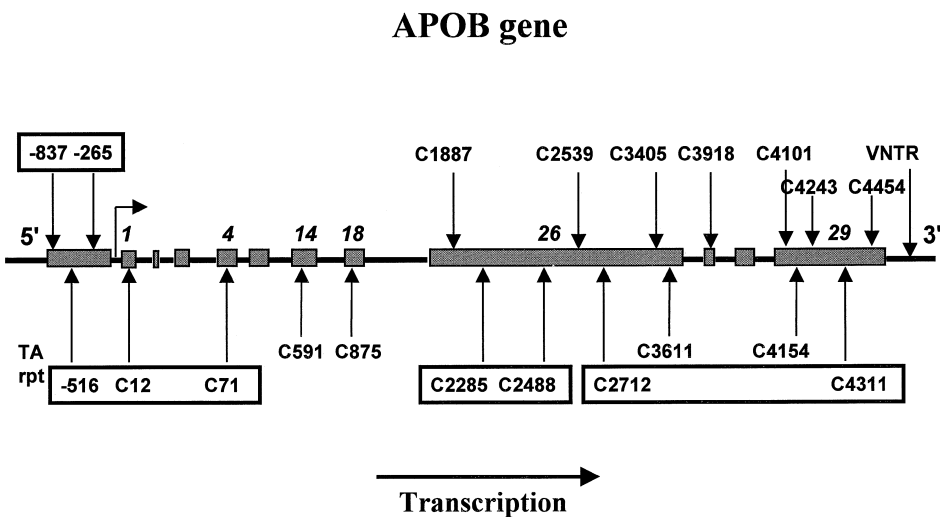

Figure 4 Examples of nearly complete associations between groups of polymorphisms in APOB gene. The codon affected is identified by a "C" prefix to the codon number. Exons are shown as gray boxes, and their numerical designations are in italics (some exons are skipped, and the drawing is not to scale). The numbering of the variants in the 5′ flanking region is from the transcription start site. Polymorphisms enclosed in boxes are in almost complete association.

that there is a variability, in the sequence diversity, among genes. For example, when we selected the 10 genes in which the scanned sequences were >4 kb (in order to have more reliable estimates), the nucleotide diversity varied from 0 (ADRB1 gene) to .006 (PDGFA gene).

This systematic survey of the sequence variation of a set of candidate genes for cardiovascular diseases provides a number of elements that might be relevant to the strategy of association studies. First, they demonstrate the presence of a relatively large number of common polymorphisms within coding and regulatory regions of candidate genes. This raises the possibility that several of these polymorphisms might be functional. Second, they reveal a strong pattern of nonrandom association among polymorphic sites within genes, complete linkage disequilibrium between two polymorphisms being a common feature. Third, nearly complete association among several polymorphisms within a gene is frequently present. Apart from the redundancy of information yielded by completely concordant polymorphisms, this observation raises the possibility of functional haplotypic combinations.

The strong linkage disequilibrium observed within candidate regions suggests that genomewide maps of SNPs might be efficient tools for identification of new genes involved in complex diseases. Indeed, one can expect from the results of the present study that an anonymous SNP, if it occurs within a gene or in its immediate vicinity, has a high probability of being in linkage disequilibrium with functional mutations of the gene. However, genomewide association studies imply that maps of SNPs should be sufficiently dense, since linkage disequilibrium rapidly decreases with distance across the genome. This strategy might be more efficient in genetically isolated populations, in which linkage disequilibrium extends over larger distances. Moreover, it is important to realize that this strategy does not allow the exclusion of regions of the genome, since a given SNP could be in linkage disequilibrium with several functional mutations that have opposite effects on the phenotype, which might obscure association.

If a whole-genome association study may be suitable for identification of new disease-susceptibility genes, understanding the underlying biological mechanisms requires a gene-specific approach and a study of the overall sequence variation of candidate genes. Indeed, the pattern of genotype-phenotype association might be far more complex than envisaged in early association studies that assumed a single functional variant within a gene. There are several examples in which this assumption proved to be too simplistic. One such example is that of the ACE gene, in which an insertion/deletion polymorphism in intron 16 was shown to be strongly associated with plasma ACE levels (Tiret et al. 1992). An extensive molecular screening of the regulatory and coding sequences of the ACE gene further revealed that there were probably two functional polymorphisms in strong linkage disequilibrium with each other, acting additively on ACE levels (Villard et al. 1996). Another example is provided by the CETP gene, in which a *Taq*I polymorphism was shown to be associated with both CETP mass and HDL-cholesterol, in interaction with alcohol consumption (Fumeron et al. 1995). A systematic molecular

screening of the CETP gene further suggested that there were at least three functional polymorphisms influencing CETP mass and HDL-cholesterol levels, through distinct mechanisms (authors' unpublished data). As a final example, we may consider the well-known APOE polymorphism, which is characterized by the presence of three common alleles generated by two polymorphisms affecting codons 112 and 158 (Davignon et al. 1988). A third polymorphism, located in the promoter of the gene and in linkage disequilibrium with the two others, has recently been shown to affect apoE expression (Lambert et al. 1998). These examples suggest that the overall polymorphism of a gene should be investigated, not just a few markers. They also highlight the importance of measuring relevant intermediate phenotypes, as a way to progress toward the identification of functional polymorphisms, which is the ultimate goal of genetic studies.

If it is assumed that the mean number of common polymorphisms within regulatory and coding sequences is ~4/gene, then 400,000 polymorphisms would have to be identified to characterize the entire variability of human genes. Such identification may soon be within the scope of the new technologies developed for investigation of DNA-sequence variation. One of the major goals of the postgenome phase (Lander 1996) will be to characterize this variability. Whether intronic sequences will have to be investigated integrally or only partly is an open question. As part of our work on 36 genes, a number of polymorphisms were identified in the regions of introns flanking the exons (data not reported). Obviously, some of these polymorphisms may have a functional impact.

Our own objective is to focus on the variability of candidate genes for cardiovascular disorders and to investigate the role of the detected polymorphisms in a number of association studies (see our Internet site, Canvas, for a list of available studies). For the genes already explored, our Internet site provides a detailed description of polymorphisms, allele frequencies, pairwise linkage-disequilibrium coefficients, haplotype frequencies, and conditions of PCR/SSCP and genotyping.

In conclusion, the present study on 36 candidate genes argues in favor of both genomewide association studies and detailed studies of the overall sequence variation of candidate genes, as complementary approaches. For the gene-specific approach, the availability of intermediate phenotypes and clinical endpoints relevant from a genetic perspective will be a crucial requirement. Phenotype, rather than genotype, may rapidly become the major challenge of genetic research of complex diseases.

Acknowledgments

We thank Christiane Souriau for DNA extraction. The recruitment in the ECTIM study was supported by grants from the Squibb Laboratory, the British Heart Foundation, INSERM, and the Institut Pasteur de Lille. The genetic program was partly supported by an agreement between INSERM and the Merck Sharpe and Dohme Chibret Company. S.-M.H. is supported by Deutsche Forschungsgemeinschaft grant HE 2852/1-1.

Electronic-Database Information

Accession numbers and URLs for data in this article are as follows:

Canvas, http://ifr69.vjf.inserm.fr/~canvas/
Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim

References

Behague I, Poirier O, Nicaud V, Evans A, Arveiler D, Luc G, Cambou J, et al (1996) β fibrinogen gene polymorphisms are associated with plasma fibrinogen and coronary artery disease in patients with myocardial infarction: the ECTIM study. Circulation 93:440–449

Cambien F, Poirier O, Mallet C, Tiret L (1997) Coronary heart disease and genetics: an epidemiologist's view. Mol Med Today 3:197–202

Collins D, Jukes T (1994) Rates of transition and transversion in coding sequences since the human-rodent divergence. Genomics 20:386–396

Collins F, Guyer M, Chakravarti A (1997) Variations on a theme: cataloging human DNA sequence variation. Science 278:1580–1581

Davignon J, Gregg R, Sing C (1988) Apolipoprotein E polymorphism and atherosclerosis. Arteriosclerosis 8:1–21

Fumeron F, Betoulle D, Luc G, Behague I, Ricard S, Poirier O, Jemaa R, et al (1995) Alcohol intake modulates the effect of a polymorphism of the cholesteryl ester transfer protein gene on plasma high density lipoprotein and the risk of myocardial infarction. J Clin Invest 96:1664–1671

Gojobori T, Ishii K, Nei M (1982) Estimation of average number of nucleotide substitutions when the rate of substitution varies with nucleotide. J Mol Evol 18:414–423

Kwok P, Deng Q, Zakeri H, Taylor S, Nickerson D (1996) Increasing the information content of STS-based genome maps: identifying polymorphisms in mapped STSs. Genomics 31:123–126

Lambert J, Pasquier F, Cottel D, Frigard B, Amouyel P, Chartier-Harlin M (1998) A new polymorphism in the APOE promoter associated with risk of developing Alzheimer's disease. Hum Mol Genet 7:533–540

Lander ES (1996) The new genomics: global views of biology. Science 274:536–539

Li W, Sadler L (1991) Low nucleotide diversity in man. Genetics 129:513–523

Li W, Wu C, Luo C (1984) Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. J Mol Evol 21:58–71

MacLean CJ, Morton NE (1985) Estimation of MYRIAD haplotype frequencies. Genet Epidemiol 2:263–272

Nei M (1987) Molecular evolutionary genetics. Columbia University Press, New York

Nickerson D, Taylor S, Weiss K, Clark A, Hutchinson R, Sten-

gard J, Salomaa V, et al (1998) DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. Nature Genet 19:233–240

Orita M, Iwahana H, Kanazawa H, Hayashi K, Sekiya T (1989) Detection of polymorphisms of human DNA by gel electrophoresis as single-strand conformation polymorphism. Proc Natl Acad Sci USA 86:2766–2770

Parra H, Arveiler D, Evans A, Cambou J, Amouyel P, Bingham A, McMaster D, et al (1992) A case-control study of lipoprotein particles in two populations at contrasting risk for coronary heart disease: the ECTIM study. Arterioscler Thromb 12:701–707

Poirier O, Georges J, Ricard S, Arveiler D, Ruidavets J, Luc G, Evans A, et al (1998) New polymorphisms of the angiotensin II type 1 receptor gene and their associations with myocardial infarction and blood pressure: the ECTIM study. J Hypertens 16:1433–1447

Poirier O, Ricard S, Behague I, Souriau C, Evans A, Arveiler D, Marquès-Vidal P, et al (1996) Detection of new variants in the apolipoprotein B (apoB) gene by PCR-SSCP. Hum Mutat 8:282–285

Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273:1516–1517

Saiki RK, Bugawan TL, Horn GT, Mullis KB, Horn HA (1986) Analysis of enzymatically amplified beta-globin and HLA-DQA alpha DNA with allele-specific oligonucleotide probes. Nature 324:163–166

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74: 5463–5467

Tiret L, Amouyel P, Rakotovao R, Cambien F, Ducimetière P (1991) Testing for association between disease and linked marker loci: a log-linear model analysis. Am J Hum Genet 48:926–934

Tiret L, Rigat B, Visvikis S, Breda C, Corvol P, Cambien F, Soubrier F (1992) Evidence from combined segregation and linkage analysis that a variant of the angiotensin I–converting enzyme (ACE) gene controls plasma ACE levels. Am J Hum Genet 51:197–205

Villard E, Tiret L, Visvikis S, Rakotovao R, Cambien F, Soubrier F (1996) Identification of new polymorphisms of the angiotensin I–converting enzyme (ACE) gene, and study of their relationship to plasma ACE levels by two-QTL segregation-linkage analysis. Am J Hum Genet 58:1268–1278

Vogel F, Kopun M (1977) Higher frequencies of transitions among point mutations. J Mol Evol 9:159–180