INDIVIDUAL VARIATION IN PROTEIN-CODING SEQUENCES OF HUMAN GENOME

SHAMIL SUNYAEV, JENS HANKE, DAVID BRETT, ATAKAN AYDIN, INGA ZASTROW, WARREN LATHE, PEER BORK, and JENS REICH

Max-Delbrück-Centrum of Molecular Medicine, Berlin-Buch (Germany) and European Molecular Biology Laboratory, Heidelberg (Germany)

I. Introduction

Everyday experience shows us that all human beings are unique in their external appearance and can be safely distinguished if only sufficient independent physical traits are compared. From so-called identical twins we learn that a considerable part of this phenotype is inherited.

One of the principal discoveries of modern experimental biology was the importance of proteins for cellular and organismal life. "Life is the form of existence of protein bodies" stated Friedrich Engels, the co-founder of Marxism, in his "Dialectics of Nature," which appeared at about the same time as Darwin's writings. This was a bold speculation, which at the time only a political philosopher might make, but it was soon confirmed by biochemical facts. With proteins as the basis of cellular physiology and biochemistry, the problem arose whether they exist as uniform entities or whether as individual variants such as does any other biological trait. Slight variation between species was detected when the first proteins, such as hemoglobin or collagen, were systematically studied.

Copyright © 2000 by Academic Press.
All rights of reproduction in any form reserved.

0065-3233/00 $30.00

Soon individual variation was also detected, mainly in human proteins. In particular, immunology demonstrated the intraspecies individuality of proteins. The first systematically studied object was the system of blood group antigens, beginning with the ABO system detected by Landsteiner and Wiener in 1901. The blood groups were demonstrated (1913) to be stable individual properties of cellular antigens, and their individuality was heritable, strictly following mendelian rules. In later years several hundred blood groups were detected, and the topic developed into a discipline of its own (Race and Sanger, 1975). Blood group antigens became the model of human biochemical individuality and of inheritance of traits. Family as well as population genetics in humans rested on blood group antigen diversity as a main methodical tool, as the application to paternity conflicts and to ethnological history testifies.

Clinical genetics contributed knowledge on individual genetics from a quite different field, usually based on the study of rare diseases that were found to segregate in certain families. It was surmised that absence or deleterious variation of proteins is responsible for the majority of diseases whose transmission follows mendelian principles. Pauling coined the term "molecular disease" for sickle cell anemia caused by a biochemically demonstrable variant of hemoglobin (Pauling *et al.*, 1949). Later on Ingram (1956) showed, by way of peptide fingerprint analysis, the exchange of one amino acid in sickle cell hemoglobin. This was the first case of an individual variant on the level of primary molecular structure of a protein.

In the 1950s, new methods of protein separation were developed that enabled the systematic study of molecular variation in many more human proteins. Starch gel electrophoresis allowed the separation of closely related protein variants by differences in charge and molecular size. Smithies (1955) detected the amazing polymorphism of haptoglobin. In later years the method was extended to the study of allozymes (enzyme polymorphisms).

Electrophoresis remained for several decades the most powerful method for demonstrating polymorphism in human proteins. The striking biochemical individuality of human proteins emerged slowly. Harris (1966) demonstrated the existence of many polymorphisms in 3 of 10 enzymes studied in detail. He found an average heterozygosity (in his case the mean fraction of electrophoretically visible allele differences) of 10% in a sample of ethnic Europeans. Later on he extended the set of enzymes and confirmed the range of frequency values.

Harris was also among the first to present the distribution of enzyme polymorphisms in different human populations (Harris and Hopkinson, 1972). Their cumulative mean value was 28% polymorphism and 6.7% heterozygosity.

Electrophoresis as a method of study of molecular differences has its limitations. It can demonstrate variation of the primary structure only if it changes migration in the electric field, which is the case only for a fraction (about one third) of the conceivable variants; the others are electrophoretically silent. On the other hand, different amino acid replacements may cause the same electrophoretic variant and thus be indistinguishable. Furthermore, the whole method was applicable only to selected types of proteins and could not yield a genetically satisfactory overview.

With the advent of protein sequencing, also in the 1950s, attempts were made to study protein variation directly on the primary structure. However, the method was very expensive and time-consuming and could not be applied to population genetics. It remained confined to evolutionary study of differences between species (applied to molecular phylogenetics) and to the demonstration of sequence mutation in important heritable diseases.

The advent of various techniques of DNA analysis (restriction fragment length polymorphism [RFLP] analysis, gene cloning and sequencing, hybridization with polymerase chain reaction (PCR) generated probes) in the 1970s led to an explosion of studies on individual variation in coding and noncoding sequences of the human genome. Cooper *et al.* (1985), for instance, used the RFLP technique to show a nucleotide diversity (i.e., on the nucleotide level the fraction of differences between two alleles randomly selected from the population) of 0.0037 in a bulk DNA segment of the human genome; Nei (1975) arrived at a much lower value (of 0.0004) in the coding regions of the genome (comprising only 5% of the human genome). Li and Sadler (1991) collected data from 49 fully sequenced genes (75 kb of coding and perigenic regions) and found "low nucleotide diversity in man," namely only less than 1 polymorphism per 1000 sites, in contrast to certain strains of *Drosophila* whose genome was shown to contain about 10-fold more variants.

This chapter describes recent developments in the study of the individual variation in the coding part of the human genome. It is now possible to estimate genetic parameters on the molecular scale. This field is still in its infancy, although the necessary large-scale array technology is expected to mature soon and will be applied to genomic sequences of large population samples. The studies so far published are still restricted to certain segments of the genome (where interesting genes are located). Larger noncoding regions have not been studied. Meanwhile our group, as well as others, concentrated on database analysis of expressed sequence tags (ESTs) as a source of preliminary statistical information of the amount and distribution of individual variation along the whole human genome. As the most recent and presumably most efficient method of study is the indirect one, namely establishing genomic nucleotide variation of DNA or of cDNA derived from expressed genes rather than directly of protein sequence, it is important to put variation in coding and noncoding genomic parts into comparative perspective.

II. Polymorphism versus Mutation—Neutral Drift versus Selection

If in a DNA or protein sequence a single position is variant, this may be called a polymorphism or mutation. The difference is set by convention: a mutation is rare, and a polymorphism is common. By implication, a mutation may be deleterious, or in rare cases it may be advantageous for the reproductive fitness of its carriers. A widespread polymorphism, on the other hand, is unlikely to affect the fitness to such an extent. A frequency value of 1% of the polymorphic allele is usually taken as a threshold between mutation and polymorphism (Kimura, 1983; Li, 1997). A polymorphic site is called biallelic if two variants segregate in the population, and multiallelic if there are more than two variants, which is a rare event in human sequences.

There is a long-standing controversy whether mutation or selection is the major driving force of molecular evolution. ''Neutralists'' say that its rate is mainly determined by an excess of mutation, and the main changes in time are explainable by a selectionally neutral drift of genes in the space of possible sequence variation. ''Selectionists'' say that there is always enough variation (i.e., mutation) in a population, and decision comes from environmental change and selection against unfit or in favor of fit variants. Proponents of the neutralist theory are Kimura (1983) and Li (1997), whereas the neodarwinian selectionist argument has been stressed by others (Gillespie, 1994). The truth is perhaps that both factors play a role, and which of them dominates depends on the particular problem.

A polymorphism can be a risk factor for an individual's health or life expectancy without impairing its reproduction. This appears to be the case for many common human diseases, which manifest themselves only after the generative period, so that evolutionary selection against such a trait cannot be operative.

III. Population Dynamics of Sequence Variation

A polymorphism that occurs at considerable frequency in a population is likely to be very old (i.e., must have originated many generations ago).

Population genetics has produced a detailed theory of the relationship between frequency, drift, and selective value of genetic variation (Nei, 1975; Hartl and Clark, 1989; Li, 1997). The problem with its straightforward application is that the main parameters, such as history of population size, fixation rate, and selective pressure, are difficult to establish.

The age of a common variant, i.e., the number of generations during which it existed, is in the numeric range of the effective population size N_e, provided that the mutation rate did not vary dramatically and that the selective value is nearly neutral. A rare variant, on the other hand, may be rare because reduced reproductive fitness has prevented its spreading over the population. In this case it is on its way to extinction or to equilibrium with new mutations, but it may also be a neutral polymorphism on the rise, i.e., a young one; a polymorphism that is just disappearing because of random genetic drift; or linked to a site that is functionally favorable and is therefore "hitchhiking" with that site under heavy selective pressure (see discussion by Chakravarti, 1998, and Clark et al., 1998).

The effective population size N_e (defined as the number of people contributing chromosomes to the next generation) is lower than, but in the order of magnitude of, the true size N of a breeding population. If N fluctuates with time, then the average effective population size is an intermediate value (the harmonic means) between all values that it had during the evolution, but such a value is closer to the value of the "bottleneck" of the population history. Also geographic or other causes of isolation of subpopulations tend to modify the effective population size. As all such circumstances are difficult to retrace after many generations, the estimation of the effective population size is a difficult undertaking. Bipedal tool-using hominids have occupied much of the Old World for approximately 1 million years. Modern human population appears to have originated from a small group (of a few thousand, at most) about 100,000 years (about 5000 generations) ago. It underwent several narrow environmental bottlenecks since, so 1000 to 5000 generations might be the age of most segregating sites on the human genome. Our population size remained at a low value (perhaps as low as 10^4, see also Bergstrom *et al.*, 1998) for a long time and exploded at an exponential rate only about 50,000 years (about 1000 generations) ago (Harpending *et al.*, 1998). For species older than *Homo* such an estimate is even more difficult to obtain.

The usual situation at a molecular site (such as a sequence position) is monomorphism (only one nucleotide present in the population) or biallelism. Multiallelism is rare. This fact can be explained by population dynamics over a large number of generations. When a variant is deleterious or very favorable, it will soon die out or replace all others, respectively. If it is neutral, or nearly so, then the theory predicts that a newly appearing variant in a population of N individuals has a chance of $\frac{1}{2}N$ of becoming fixed (replacing the competitor) or $1\frac{1}{2}N$ that it will disappear again by random drift (Kimura, 1962). Fixation of a neutral allele, if it occurs, takes roughly N generations (Kimura and Otha, 1969), which can be longer than the existence of that species, whereas disappearance or fixation due to selection, if it occurs, is complete after a number of generations proportional to log N, i.e., rather soon. This holds when the mean number of new mutations at sites per generation is small. The evolutionary story of a site depends then on its selective value. If it is advantageous or disadvantageous, long periods of monomorphism of a site are interrupted by short periods where rare variants appear and disappear or become fixed very quickly. Neutral sites, on the other hand, display slow changes of frequency together with long fixation or elimination time. Thus common polymorphisms are liable to be neutral or nearly neutral variants.

The mutation rate μ of the nucleotide (or amino acid) at a sequence site is related to the popular notion of a "molecular clock" (Zuckerkandl and Pauling, 1965), because it determines after which time the clock ticks and a new mutation arises because of a copying error during meiosis. Whether this clock ticks uniformly is a topic of prolonged debate (summarized in Li, 1997). The question is usually treated by comparing sequence difference at (supposedly) neutral sites with evolutionary distance between species.

Typical values of μ range between 10^{-8} and 10^{-9} per individual site per generation. The actual value seems to vary for different types of replacements, in different gene regions, in different genes and different populations, and between species. Under the neutral hypothesis it is possible to equate mutation rate (number of mutations appearing every generation) and evolutionary rate (number of substitutions reaching fixation per generation), independent of the population size. If selection is operative at a site, the evolutionary rate will explicitly depend, apart from the mutation rate, on effective population size and selective advantage (Kimura, 1983).

In an extant population the heterozygosity at a given site may be measured. Under the neutral hypothesis and assuming that the mutation rate is sufficiently low, one may calculate the product $4N_e\mu$, which is in the numerical range of the expected heterozygosity. A typical nucleotide heterozygosity is in the range of 5/10,000, which implies that if μ is 10^{-9} or (10^{-8}) then N_e is 500,000 (or 50,000, respectively).

IV. Species Difference versus Intraspecies Variation

Molecular phylogeny is a discipline that studies species differences between DNA or protein sequences. Its basic tenet is that during evolution, the sequences have drifted apart by mutation and selection as well as by random drift and fixation of variants in certain positions. The earlier two species separated the more differences became fixed. Phylogenetic trees are constructed on the basis of mutual differences of protein and/or DNA sequence. Comparison of intraspecies variation with between-species variation may in the future yield information on the neutralist/selectionist alternative. McDonald and Kreitman (1991) devised an interesting test against neutrality that compared the ratio of silent/replacement mutation of a given locus within a species with the same ratio between two related species. Under the neutral theory this should be equal (corrected for sample size), but in fact it is not (see Li, 1997, and Hudson, 1993, for a discussion).

V. Studies on Single-Nucleotide Polymorphism (SNPs)

Large-scale identification of single nucleotide polymorphism is one of the major goals of the genome research in the next years. This will furnish rich information about individual genetic diversity in humans. A further obvious application is the usage of mapped SNPs as genetic markers of susceptibility for diseases or of variation in drug response. Also SNPs that in themselves affect the phenotype are important. Such variations may occur in noncoding regulatory regions or splice sites of the DNA, but perhaps more frequently in the primary structure of proteins. If variation of primary structure is involved and its common polymorphisms are known, study of direct association between gene polymorphism and traits may be more powerful than the indirect demonstration of the responsible locus by linkage mapping to marker loci [compare the relevant discussion by Risch and Merikangas (1996) and Collins *et al.* (1997)].

Chip-based technology for finding SNPs in the human genome are emerging (Wang *et al.*, 1998; Winzeler *et al.*, 1998; Hacia, 1999; Ramsay, 1998) and will enable the study of polymorphism in large populations. It requires knowledge of polymorphic sequence tagged sites (STSs), allowing the synthesis of specific oligonucleotides on the chip. To find cSNPs, one must screen large sets of overlapping oligonucleotides over the whole coding segment of many individual chromosomes. So far this systematic identification of SNPs has been confined to limited gene regions of special interest (Harding *et al.*, 1997; Nickerson *et al.*, 1998). Another study (Wang *et al.*, 1998) covered a broad part of the genome, but only with a small population sample. Winzeler *et al.* (1998) screened the whole (but relatively small) yeast genome, whereas Chee *et al.* (1996) studied the complete variation of the complete (and again relatively small) mitochondrial genome of humans on several dozen individuals.

Wang *et al.* (1998) studied SNPs in the human genome. They examined 2.3 megabases of human genomic DNA by gel-based and DNA chip methods and identified more than 3000 SNPs and established the genomic location of more than 2000 among them to provide a powerful mapping tool for genetic studies of traits important in medical applications. They produced DNA chips that allow simultaneous genotyping of 500 SNPs. Part of their study dealt with the frequency of SNPs in gene regions of the genome, by genotyping individual variation at sequence-tagged sites obtained from 3'ESTs (two-thirds) and from random genomic sequence (one third). Their estimate of average heterozygosity was 4.5 in 10000 sites. A difference between this value in 3'ESTs and random ESTs was interpreted to be consistent with more selective constrain in coding than in noncoding regions. The authors did not specifically address the frequency of cSNPs in the narrower sense.

Nickerson *et al.* (1998) studied the individual genomic variability of a contiguous stretch (9.7 kb) of the human lipoprotein lipase gene by sequencing 142 chromosomes from individuals from three very different ethnic origins; 90% of this region was noncoding (introns) and 10% was coding, which is fairly typical for human genes. They found 79 single base substitutions and 9 insertion/deletion variants; 81 variable sites were found in the noncoding region of 8736 nucleotides (nt). There were 7 variants in the coding region of 998 sites; 4 of some were missense-variants on the protein sequence level. More than half the variants were found in more than 10% of the individuals examined. The nucleotide sequence diversity, defined as expected heterozygosity averaged over all sites, was 1 per 500 nt overall (approximately equal over all ethnicities), but was fairly different in different parts of the gene. It was highest in regions of genomic repeats (1 per 312), followed by noncoding regions (1 per 476) and coding regions (1 per 2000 nt). There was 1 segregating variable site per 142 nt in coding and 1 per 108 nt in noncoding regions. Obviously cSNPs are more carefully selected against and tend to occur predominantly at low frequency. This explains why the ratio of segregating sites in coding versus noncoding parts may be about unity, but the frequency-dependent score of heterozygosity is not.

A series of papers with first systematic studies of SNPs in selected genes have been published recently (Cargill *et al.*, 1999; Halushka *et al.*, 1999, Hacia *et al.*, 1999). In all studies high-density variant detector arrays (Chee *et al.*, 1996; Lipshutz *et al.*, 1999) were used to detect SNPs in PCR products. In the study denaturing high-performance liquid chromatography was used for detection of variants in PCR products. In all cases the authors confirmed candidates obtained by these screening methods by direct sequencing of the relevant genomic regions.

In the first two studies (Cargill, Halushka) known genes were investigated, which are thought to be relevant for polygenic cardiovascular (including blood pressure regulatory), endocrinological and neuropsychiatric disease traits. A total of 181 genes (about 380 kb of genomic sequence) were screened for variants in their coding and adjacent 3′ and 5′UTR regions. The data came from approximately 260 independent alleles; donors were from different African, European, and Asian sources. The authors of this companion publication reported a total of about 1500 SNPs, of which 780 (about one-half) were in the coding region of the respective genes, and half of these were synonymous and one-half nonsynonymous cSNPs. As only about 40% of all possible variants were synonymous, an equal number of SNPs means that they are about two fold to three fold more frequent in synonymous sites.

The nucleotide diversity in silent positions may be calculated for a given sample size and sequence length and is about 8 to 10 per 10,000 sites. Estimates of θ and π (diversity and heterozygosity) were close to each other, as suggested by the neutral theory assuming constant population size (10,000 individuals) under the infinite sites model of population genetics (Li, 1997).

The situation was different for replacement variants. Here the number of segregating sites, as well as the average heterozygosity, was much smaller than in silent positions and in noncoding positions. This was explained as selection effect. Silent and replacement polymorphisms are believed to occur at the same basic rate, but only about one third of the latter are "accepted" by selection owing to reproductive fitness. This conclusion was further corroborated by the more subtle argument that estimates of the mutation parameter θ were greater if based on the number of segregating sites than on heterozygosity. This points to slightly deleterious alleles in the replacement fraction (Tajima's test, cf. Li, 1997, p. 248).

The distribution of cSNPs in different genes is not homogenous. There were genes without any cSNPs and others with up to 30 cSNPs. This pattern was also observed in *Drosophila* (Moriyama and Powell, 1996).

Both articles included comparative studies on ape genomes in an attempt to reveal information on the evolutionary age of polymorphisms. If at a biallelic site one allele is present in the related species, it thought to be the ancestral allele. In most cases the more frequent allele was also the ancestral one, although there were exceptions with the less frequent allele being that of the chimpanzee (this was ascribed to drift or negative selection pressure). The nucleotide diversity between chimpanzee and humans was confirmed to be in the percentage range, i.e., about one order of magnitude higher than the intraspecies variation in humans. Also the third article of the cooperative group using *Affymetrix arrays* (Hacia *et al.,* 1999) dealt with ancestral variants among a set of SNPs. In their set a considerable fraction of alleles were dominant in humans but not of common ancestry with the other primates. Most alleles newly acquired by humans involved the highly mutable CpG dinucleotide in genomic sequences.

Polymorphisms were much more frequent in individuals whose ancestors came from Africa rather than from other continents. This finding is in keeping with previous studies (e.g., Zietkiewicz *et al.,* 1997). It supports the idea that non-African populations originated from Africa and on emigration in prehistoric time were subject to a "population bottleneck."

The articles contain cautious extrapolations to the whole genome. It is predicted that 75,000 genes contain about 1 million SNPs, 500,000 of them in noncoding regions, 250,000 silent, and 250,000 missense cSNPs. Because more than 40,000 genes have already been "hit" by ESTs, one may expect a large number of such polymorphisms to appear in EST collections if their volume continues to increase at the present rate. About 80% of all genes may be expected to be polymorphic at the protein level, with an average heterozygosity of 17% between sequences, which is higher than the classic studies (such as Harris, 1966; and Harris and Hopkinson, 1972) with physicochemical methods in the pregenomic area.

VI. ESTs as Data Source

A. *Studying Individual Sequence Variation*

Expressed sequence tags are short sequence segments (usually up to 500 nt long, but rarely they may be longer by a factor of 2 to 3) obtained by reverse transcription into cDNA clones from mRNA preparations of a cell or tissue in a specified functional or developmental stage. They are produced by automatic procedures and released by their producers (after a certain time lag) into public databases. At present EST collections (in particular from human origin) grow much faster than any other genomic sequence information. The main application of EST analysis is gene expression. As a by-product they may be evaluated for the study of individual variation in the expressed part of the human genome.

Mining for SNPs in EST databases requires only computer resources and does not incur experimental cost (as do the various techniques of large-scale DNA chip analysis).

An ideal EST meets the following criteria:

It is a short contiguous reverse-transcribed segment excised from a spliced mRNA. It should contain either the 5'untranslated region (5'UTR) and/or spliced exonic sequence, and/or 3'untranslated region (3'UTR).

The ensemble of ESTs in the available databases should cover all genes of the genome and all parts of each gene. At present, there are about 1.3 million human ESTs covering about 75,000 human mRNAs. Thus one mRNA is hit on the average by >10 ESTs, but one EST can cover only a fraction of mRNA sites (about 300 nt per about 2000 sites).

It contains neither intergenic material away from the coding region nor intronic sequences.

Its abundance is approximately proportional to the equilibrium between synthesis and hydrolysis of mRNA.

To avoid heavy overrepresentation of mRNA species typical for the respective tissue (such as globin in red blood cells), normalization procedures reduce the redundant population.

Alignment of autologous EST stretches from different donors reflect individual genomic variation in the coding region (missense and silent), and/or the adjacent expressed regulatory parts (e.g., promoter region, terminator region).

It displays part of the correct amino acid sequence of the gene product when read in the correct complementarity and reading frame.

It reveals splice variants.

In practice, the EST collection does not live up to these ideal demands:

It can cover only a fraction of the expressed part of the genome, because some genes are read off at a very low level or not at all.

Coverage of expressed information is far from uniform. Figure 1 shows a typical example of a gene whose mRNA sequence is known. The coverage is skewed toward the 3'UTR. Only about 30% of all mRNA sites are at the present time covered by more than one EST library (see results later). This reduces the chance of finding many of the existing SNPs. As a result any large-scale *in silico* analysis of polymorphic variations will be biased toward the tail region of strongly expressed genes.

There is some error in the sequences (Ewing *et al.*, 1998), which is no problem for the usual whole-sequence-based approach to expression analysis, but a drawback when individual sites are studied. In particular, the automatic base calling by a computer may increase the error.

```
                                   400       799      1199      1598      1998      2397      2796      3196      3595
SequenceName            1 ----<-----|---------|---------|---------|---------|---------|----->---|---------|---------|---------|  3994
     3882166 (   1)      1 ------------------------------------------------------------------------------------------------       3799
     2008103 (   3)    357          ----------                                                                                     730
     1626046 (   2)    494             ----------                                                                                  850
     1358718 (  -8)    508              ---------                                                                                  856
     1441765 (  -1)    511              -------                                                                                    769
     1965614 (   0)    531              -------------                                                                              986
     1439436 (  -8)   1129                             --------                                                                   1411
     1440515 (  -4)   1151                              ------                                                                    1358
     4900686 (  -2)   1292                                 ------------------                                                     1971
     1978776 (   1)   1517                                       ----------                                                       1870
     1637382 ( -10)   1619                                         -----------                                                    1981
     1260651 (   1)   1660                                          -----                                                         1812
     2015904 (   0)   1800                                              --------                                                  2077
      831553 (  15)   1940                                                 ------------                                           2372
     2356787 (   0)   1951                                                  -------+--                                            2317
     2357528 (   0)   1951                                                  -------+--                                            2317
     1183638 (  -1)   2130                                                      -----------                                       2504
      767580 (  18)   2178                                                       -----------                                      2546
     2013469 (   0)   2227                                                         --------                                       2526
      672938 (   0)   2269                                                          --------                                      2548
     1201054 (   8)   2345                                                            ----------                                  2730
     1099916 (   8)   2355                                                            -----------                                 2759
     2787361 (  -1)   2520                                                                ----------                              2888
     2355879 (   0)   2523                                                                ----------------                        3130
      844136 (   0)   2633                                                                   -------------                        3124
     4564720 (   0)   2759                                                                      ----------                        3130
     4085761 (   0)   2766                                                                      ----------                        3134
     2567972 (   1)   2779                                                                      -+--+-----                        3129
      839442 (   0)   2870                                                                         --+----                        3122
      841813 (   2)   2888                                                                         -+---------                    3281
     1689649 (   0)   2901                                                                          ----------                    3296
     2184366 (   0)   3017                                                                            -----+--------              3514
     1270542 ( -20)   3049                                                                             ----+----------            3604
     2457176 (   0)   3239                                                                                  +-------------        3769
     2616416 (   1)   3261                                                                                   -------              3504
     4686943 (  -2)   3283                                                                                   --------------       3799
     2947719 (  -5)   3327                                                                                    ------              3520
     1202843 (   1)   3332                                                                                    ------------        3767
     4511078 (   0)   3333                                                                                    ---+--------        3769
     4153004 (   0)   3334                                                                                    ---+---------       3793
     2255752 (   0)   3407                                                                                      --+-------        3769
     3181454 (   0)   3407                                                                                      --+-------        3769
      856511 (  18)   3408                                                                                      -----------       3795
     1137708 (   2)   3409                                                                                      -------------     3899
     4898083 (   0)   3411                                                                                      ----------        3758
     2540735 (   0)   3418                                                                                      -----------       3793
     2589445 (   0)   3418                                                                                      ----------        3769
```

There is also a small error (said to be about 1/10,000;) owing to reverse transcription and synthesis involved in the generation of cDNA clones from mRNA.

Incompletely spliced primary messenger, as well as unprocessed genomic material, may be present as impurities in a mRNA preparation and may obscure the alignment of autologous ESTs necessary for finding variants.

SNP candidates derived from ESTs refer to one allele of the donor person, so the zygosity of the carrier may remain obscure.

Some EST preparations come from pooled material rather than from one person, which may make statistical calculation dubious.

Several libraries are from one person, which also confuses statistical calculation.

An error source is that a variation in an alignment of EST sequences may not come from the same gene but rather from a highly similar paralogous copy elsewhere on the genome or from a pseudogene ("dead" gene: transcribed, processed, but not translated into protein). This necessitates restriction to high sequence identity as a criterion of inclusion of an EST into the aligned cluster. This does not fully rule out the paralalogy error and nevertheless risks to exclude some true variants that do not pass such a strict threshold.

Despite these problems EST databases are a valuable source of large-scale analysis of human variation. They will become even more valuable as the data continue to grow at the present rate. An algorithm for computer-aided SNP mining should contain filters to eliminate the potential sequence errors. Such filters can be based on the probabilistic analysis of sequence features. It can also take into account that multiple occurrences of a variant are more trustworthy, and it may furthermore focus on improving the quality of base-calling if the fluorescent traces are available for closer srcutiny.

FIG. 1. Example of EST coverage of the coding region of a gene. Shown in schematic form is the coding and perigenic sequence obtained from *Homo sapiens* mRNA for KIAA0723 protein (GI code 38982166). It served as master sequence to which all ESTs with nucleotide identity >99% were aligned. The first line after the scale symbolizes the complete messenger (3994 nt) from its 5′UTR (left) to the 3′UTR. The symbols < and > symbolize the end of the coding region. The cluster of ESTs is curtailed: on the bottom right follow 179 more ESTs, all in the 3′EST region. It is seen that the EST cluster covers nearly the whole gene with random ESTs. The +−symbol shows 13 positions where the alignment contains a SNP candidate (relative to the mRNA master sequence).

B. *Methodics of EST-Driven Generation and Evaluation of cSNP Candidates*

ESTs were obtained from dbEST (Boguski *et al.*, 1993), a regularly expanding database as division of GenBank (Benson *et al.*, 1999) that contains sequence data and other information on "single-pass" cDNA sequences and/or expressed sequence tags from a number of organisms including homo sapiens. A brief account on the temporal development of that collection is given by Boguski (1995).

mRNA could be obtained from GenBank or EMBL entries identified by the appropriate annotation ("mRNA," "complete cDNA"). GenBank is the genetic sequence database maintained at the NCBI of National Institutes of Health in Bethesda, MD. EMBL is a sequence database maintained in EBI outstation of EMBL in Hinxton. There were approximately 3 billion nt in 4 million sequence records in these databases as of June 1999. About 9000 mRNA or cDNA entries may be used as a fully sequenced master template for studies of variation.

Several groups focused on hunting SNPs from assembled EST clusters such as collected by UNIGENE database (Schuler, 1996, 1997; UNIGENE, 1999; for such applications see Taillon-Miller *et al.*, 1998; Buetow *et al.*, 1999; Picoult-Newberg *et al.*, 1999). The last two groups used the Phred base calling program together with the Phrap sequence assembling tool (Ewing *et al.*, 1998; Ewing and Green, 1998; Green, 1998; Gordon *et al.*, 1998). This procedure yields a quality score for each base as called, which expresses its statistical trustworthiness on a logarithmic scale (e.g., Phred value >20 is already a reliable call, whereas values below 20 are increasingly doubtful).

The approach of Buetow *et al.*, (1999) includes two additional features. With the help of the PHYLIP package, they exclude possible paralogous genes via analysis of the phylogeny of the cluster. The EST set so purified is analyzed by the DEMIGLACE tool, which extracts protential sites of variation from the multiple alignment and applies several filters. A SNP candidate is rejected if its neighbors have low Phred quality, if the fluorescence peak is too small compared with standard peaks of that base, and if there is a double peak or one read from one DNA strand only. All remaining candidates are scored by Phred quality values converted to bayesian posterior probability.

Picoult-Newberg *et al.* (1999) also applied after Phred/Phrap a set of filters to avoid sequencing errors. They excluded candidates if there were indels or further mismatches nearby in the sequence. They neglected variants suggested in the first 100 EST positions since this region is known to have a high error rate. They also discarded variants seen only once in the EST collection.

We performed a benchmark analysis on a set of genes for which the full mRNA and/or the pertinent protein sequence was documented in the literature. Instead of clustering ESTs without any template, we aligned them by a BLAST search (Altshul *et al.*, 1997) to this set of master mRNA sequences and looked in these alignments for variant letters. We applied a set of filters as follows:

Only subalignments of length >100 nucleotide above 99% sequence identity and with >15 exact matches at both ends were considered. This is a hard criterion for excluding paralogs and other unreliable candidates.

Positions were excluded when there were closely located further mismatches (window sizes applied: 33 and 3 nucleotide).

Sequence patterns were excluded that were liable to cause gel compressions (Yamakawa and Ohara, 1997), or homopolymer stretches, which often lead to base miscalls.

ESTs that aligned to >1 mRNA of the panel were excluded.

A significant improvement of the prediction reliability is achieved by considering only variants that occur more than once. The price to be paid is a strong sampling bias toward frequent variants.

About 60% of the data collections offer the pertinent EST chromatograms. In these cases we applied a filter based on Phred quality.

C. *SNPs Identified by EST Data Mining*

Buetow *et al.* (1999) reported >3000 candidates with a score >0.99 from the set of >8000 UNIGENE clusters. A subset of nearly 200 candidates was subjected to a direct validation in a pooled preparation from 10 individuals (20 chromosomes). More than 80% of these candidates were indeed present as variation in this pool.

Picoult-Newberg *et al.* (1999) analyzed >21,000 5′ESTs and >19,000 3′ESTs. More than 6000 candidates were localized, but only 850 passed the filters applied. They inspected the fluorescence traces of 100 randomly selected specimens. A total of 88 verified candidates were then validated as common variants by sequencing from a panel of individuals; 55 out of 88 sites were confirmed to be polymorphic. In four cases all samples appeared to be heterozygous, which points to sampling from more than one gene of a multigene family (they did not pursue possible paralogy).

Our own data may summarized as follows. To benchmark the mining our method and also to provide access to information of medical interest, study focused on a subset of 500 human genes, called disease-associated genes, because experimental data on genetic variation and trait association are available from the OMIM database of mendelian traits (McKusick, 1999). All ESTs were aligned to the mRNA sequences of these genes. To test doubtful candidates a less strict sequence identity threshold of 95% was applied at the amino acid level for inclusion into the alignment. All other methodical details were as described previously. We selected 100 predicted nonsynonymous SNP candidates from this alignment and subjected them to direct resequencing of the cDNA clone. In 61 cases, we also obtained the fluorescence traces. Thus we could evaluate the reliability of the Phred scores as predictors of nucleotide variants. Table I shows the comparison and leads to the conclusion that variants with Phred value >20 are fairly confident candidates of a true SNP. The results of this benchmark test allow cross-validation of the fraction of known polymorphisms (mentioned in OMIM or SWISS-PROT protein sequence collection, Bairoch and Apweiler, 1999) found, as well as the percentage of false-positive and of false-negative results.

These test results encouraged us to do a SNP search in all presently available mRNA sequences. The EMBL database contains approximately 9000 fully sequenced mRNAs. Table II describes the overall coverage of mRNA nucleotide sites by at least one, and the coverage by more than one, EST. Figure 2 depicts the tail of the distribution of high coverage in more detail. Only 32% of all sites are covered by more than one EST, and this percentage is skewed toward perigenic 3′regions. A small

TABLE I

Validating of SNP Candidates: Phred Value vs. Resequencing[a]

Threshold	SNPs confirmed by resequencing	SNPs rejected by resequencing
Total number of SNPs among them	38	23
Phred > 40	15	0
Phred > 30	22	0
Phred > 20	28	1
Phred > 15	30	2

[a] A total of 61 predicted nonsynonymous SNPs were selected for which the fluorescence traces and the perinent Phred quality values were available. Resequencing of the clones confirmed 38 of the 61 SNPs, whereas 23 were rejected as sequencing errors in the database.

It is seen that candidates with high phred values are always confirmed, but a number of true SNPs will be missed (false negative). Lowering the Phred value threshold will increase selectivity, but at the cost of including more and more false positives. A value >20 is an appropriate selection threshold: candidates with higher values are reliable (only 1 false-positive in 29), but only 28/38 true polymorphisms will be found. Lists of candidates with lower values contain nearly all polymorphisms, but also 3/23 false-positive or more.

TABLE II
EST Coverage of mRNA Sites[a]

Site	(Sub) total	Covered by ≥1 EST	Covered by >1 EST
All site classes	19.4 (100%)	9.7 (50% of total)	6.3 (32% of total)
5'UTR sites	1.3 (7%)	0.4 (31%)	0.2 (15%)
Coding sites	12.7 (65%)	5.6 (44%)	3.2 (25%)
3'UTR sites	5.5 (28%)	3.7 (68%)	2.9 (52%)

[a] Counted by million nt sites, rounded values. The percent values refer to the first line, and those on the first line to the total number.

percentage (but still tens of thousands) of mRNA sites were represented by between 10 and 87 different libraries.

Table III shows that about 83,000 mismatches (SNP candidates) were identified, but only 9,228 were present in more than one library. About half of all these pass our algorithmic filters. Traces were available for

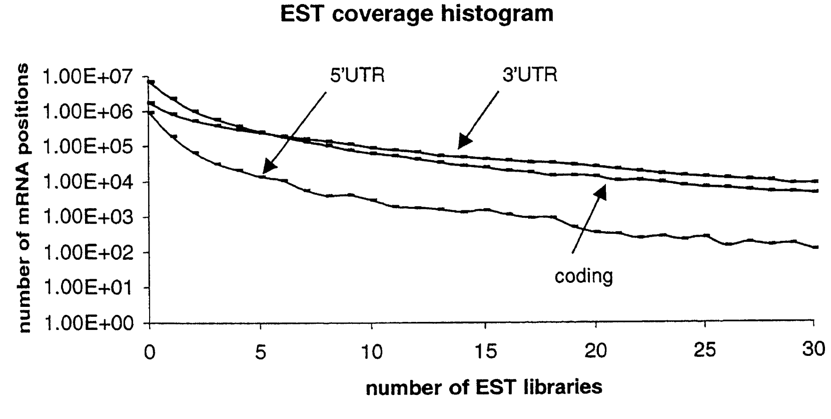

FIG. 2. Histogram of the number of distinct EST library reports of mRNA sites. For all 19 million positions of the mRNA collection, classified according to genomic function (see Table II), the number of positions that was covered by the specified number of libraries (= independent individuals) was counted. Example: About 10,000 (1.00E + .04) 5'UTR sites were reported by exactly 6 EST libraries. Only values up to reports of 30 libraries are displayed. Positions covered by more than 30 libraries (i.e., very frequently expressed genes) amount to 1.5% of the 3'UTR sites and much less than 1% of the 5'UTR and coding regions. The highest library coverage found was 87 libraries reporting a set of coding sites. The 3'UTRs contribute most of the EST coverage in the tail region (of many ESTs per site), although only 28% of the total number of positions is in this region. It is also seen that in the tail region of the distribution 5'UTR and coding regions are covered in proportion of the available sites (about 1:10 distance on the log scale, see values in Table II).

TABLE III
SNP Candidates Found and Their Predicted Accuracy[a]

Candidates	Number of SNPs candidates	
	Found in ≥1 library	Found in >1 library
Raw candidates:	82,673	9,228
passing all filters:	45,254 (55%)	5,041 (55%)
traces available for:	40,836 (49%)	4,228 (53%)
Phred value > 20:	9,231 (23%)	2,611 (53%)

[a] ESTs were aligned against the set of about 9,000 human mRNA sequences. Candidate cSNPs were extracted and subjected to various algorithmic quality filters, as described in the text.

It is seen that the Phred >20 filter confirms about 25% of all candidates. This confirmation rate is much higher (53%) if more than one library reports the variant.

On the other hand, the algorithmic filters without access to fluorescence traces confirm 55%.

50% of the candidates, and the algorithmic filter based on Phred >20 confirmed only one-fourth of the candidates represented only once, but 53% of those represented more than once.

The 9231 candidates with Phred value >20 were further characterized as to their regional position in the gene (see Table IV). These absolute numbers may be converted to estimates of mutation parameters as applied in population genetics (Table V).

D. *Estimating Population Parameters*

Our EST studies cover about 9000 mRNAs. About 6.3 million positions were aligned to more than one EST. There were about 9200 reliably

TABLE IV
Number of Candidate SNPs with Phred Values > 20

Untranslated		Coding			
5106		4125			
3′UTR	5′UTR	Synonymous		Nonsynonymous	
4891	215	1680		2445	
		4 fold degenerate sites	2 fold degenerate sites	Nondegenerate sites	2 fold degenerate sites
		955	645	2116	319

TABLE V

Population Estimates of Genomic Variation in Different Human Samples[a]

Site	EST data[b]		Cargill data[c]		Halushka data[d]	
	θ	π	θ	π	θ ("Europeans")	θ ("Africans")
Noncoding[e]	0.00062	0.00058	0.00053	0.00052	0.00054	0.00068
Coding[f]	0.00050	0.00046	0.00054	0.00050	0.00045	0.00063
(Silent)[g]	0.00094	0.00090	0.00100	0.00110	0.00090	0.00129
(Replacing)[h]	0.00036	0.00032	0.00036	0.00028	0.00031	0.00042

[a] θ, Population estimate of the number of segregating per total sites (normalized to infinite sample); π, estimate of the nucleotide diversity per nt site (heterozygosity). Both estimates have the same expected mean value (see Li, 1997). If significantly $\theta > \pi$: suggests presence of nonneutral sites. It is seen that the different samples gave similar estimates. "Africans" have somewhat higher estimates of diversity than "Europeans." The θ estimate tends to be somewhat higher than the π estimate.

[b] Own SNP data as found by EST analysis (see Table IV).

[c] See Cargill *et al.* (1999).

[d] See Halushka *et al.* (1999); Europeans, calculated from sample of Americans of European origin; Africans, calculated from sample of Americans of African origin.

[e] Refers to SNPs at perigenic sites of the 3′UTR and 5′UTR region.

[f] SNPs at sites coding for protein sequence.

[g] cSNPs at coding sites which do not replace the amino acid as coded for.

[h] cSNPs at coding sites which replace the amino acid as coded for.

reported occurrences of SNPs. As the range of different genes probed is greater than in the previous studies by other authors (which focused on certain stretches of the genome), it is interesting to compare the population genetic parameters estimated from such data.

Table V contains estimates of θ (normalized number of segregating sites) and of π (average heterozygosity) from the EST data. The values are in the numerical range of the estimates by Cargill *et al.* (1999). As expected, variation is somewhat higher in silent mutation sites than in nonsynonymous sites (0.9/1000 vs. 0.4/1000). Surprisingly, the variation is less than intuitively expected in noncoding regions (0.6/1000), which suggests that evolutionary selection does not accept polymorphisms as readily as in the synonymous region of the coding segments. Cargill data show a similar trend.

The general tendency that values based on number of variant sites are higher than those based on heterozygosity/nucleotide diversity points to the possible presence of more rare variants under selective pressure, although this effect needs further statistical corroboration before becoming a hard fact.

E. Alternative Splice Forms

The EST data bases contain a wealth of extractable information about gene structure, expression, and gene family members. ESTs can be used to identify expressed paralog gene members within the same family and/or ortholog genes expressed in other species. When the tissue type is reported, a simple expression profile can also be generated. Processed pseudogenes (lacking introns) are also identifiable from within the EST database. ESTs also represent a valuable source of structural information within a gene. Alternative splicing occurs within genes and provides a mechanism by which a specific cell or tissue type can generate a variant protein product by changing the sequence of exons normally expressed. In practice, the splicing mechanism is able to choose alternative donor and acceptor sites in the DNA sequence from which to splice out introns. This alternative splicing leads to the gain of an additional exon or the loss of an exon or part of an exon. These inframe alternative splice forms evidently lead to a change in expressed peptide sequence and can radically alter a protein's function and or location (e.g., see Klamt *et al.*, 1998; Qi and Byers, 1998). ESTs have been derived from a wide variety of tissue types including normal tissues, diseased tissues, and immortalized cell lines. There is also a wide degree of time points represented ranging from 2-week-old embryos to old age (75 years old). This inherent variability within the EST databases can give rise to a number of alternative splice forms of a gene occurring as single hit or multiple EST hits to a gene.

F. Method of Investigation

A nucleotide or peptide sequence is searched against the EST database using the BLAST (see BLAST server; Karlin and Altschul, 1990 and 1993; Altschul *et al.*, 1997) searching programs.

The stringency of the search and the length of sequence matched are set to a value where only ESTs representing the gene are matched and related structural domains from family members or more distant proteins are mostly avoided. In practice a typical expect value parameter is $E = 1e-30$ when comparing proteins against translations of the EST database (TBLASTN). The older ungapped BLAST program with the $X = 1$ parameter prevents the program from continuing an alignment where there are a number of differences between EST and query. To identify an EST with an alternative splice form, a difference in length of match between the query gene sequence and the EST is sought. Multiple sequence alignments of ESTs and the query gene sequence can be made to clarify the picture. Small artifactual differences can occur between the query sequence and EST (typically 5 to 10 base pairs or smaller) and need to be excluded. Once a large difference is noted, the full sequence of the EST must then be experimentally verified. This should exclude sequencing mistakes or incorrect annotation of the EST. A large number of original sequencing traces of ESTs are provided by Washington University (http://genome.wustl.edu/gsc/gschmpg.html). These can be used to check for quality of sequence data in the target area of the EST. The next step and probably the most important is to test the existence of the possible alternative splice form in the cell type the EST was derived from by independent means. The laboratory researcher can test for the presence of the alternative splice form either by northern blot analysis (Ozon *et al.*, 1998) of the mRNA when the difference between the two forms is large enough or by RT-PCR (reverse transcription polymerase chain reaction) from specific tissue or cell type mRNA. Direct PCR from cDNA libraries or panels can also used (Sadoulet-Puccio *et al.*, 1996). Once verified a panel of different tissue mRNAs can be tested to give an expression profile of the alternative splice form. Further proof can be obtained with antibodies against the protein to test whether the alternative splice form was translated into protein (Sadoulet-Puccio *et al.*, 1996). For an example of a novel alternative splice form discovered within an EST see Fig. 3.

G. *Alternative Splicing Within Disease-Associated Genes*

An interesting question arising from the production of alternative splice forms is that of disease association. Are alternative splice forms of a gene associated with the development of a specific disease type? Another possibility is that a specific splice form might present as a strong risk factor in the development of more complex disease types such as heart disease or diabetes. A number of such examples have been reported (see Table VI for examples). These range from the drastic reduction of a specific alternative splice form leading to a distinct form of disease (WT1gene/Frasier syndrome: Klamt *et al.*, 1998; Menkes gene/occipital horn syndrome: Qi and Byers, 1998) to specific alternative splice forms exclusively expressed or overexpressed in diseased tissue (G-protein β_3 subunit/hypertension: Siffert *et al.*, 1998; presenilin gene/Alzheimer's disease: Sato *et al.*, 1999; CD44 gene/esophageal carcinomas: Koyama *et al.*, 1999).

The discovery of new alternative splice forms of genes associated with disease has the exciting potential to lead to new rapid PCR-based diagnostic markers. The ability to extract such alternative splice forms together

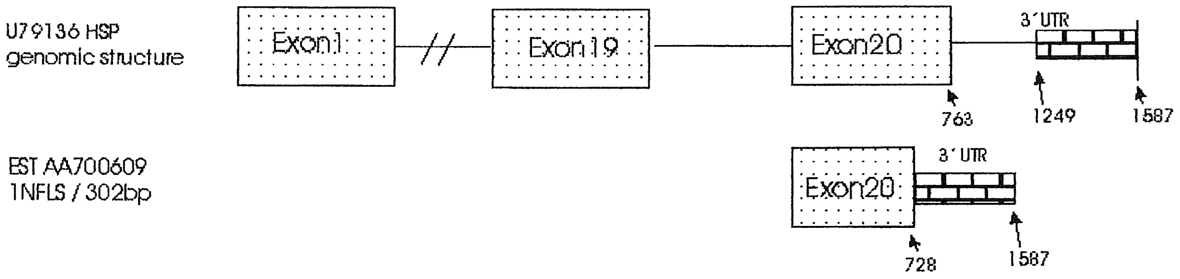

FIG. 3. Predicted alternative splicing in HPS (Hermansky-Pudlak Syndrome). The HPS polypeptide is a novel transmembrane protein that is likely to be a component of multiple cytoplasmic organelles and is apparently crucial for their normal development and function. We demonstrated an putative alternate transcript of the HPS gene. By RT-PCR, two transcripts were found in normal human brain, kidney, liver, lung, pancreas, and placenta. The short transcript (AA700609, length 302 bp) shows a deletion of 521 bp in the 3'UTR region of the HPS gene.

TABLE VI
Splice Variants and Their Associated Phenotype

Example	Form of alternative splice	Associated phenotype	Refs.
G-protein β_3	Alternative exon in the 5′UTR	Hypertension	Siffert *et al.* (1998)
MDM-2	Loss of p53 binding domain	Cancer Tissue	Sigalas *et al.* (1996)
Presenilin-2 (PS-2)	Stress induced loss of exon 5	Sporadic Alzheimer's disease	Sato *et al.* (1999)
CD44	Overexpression of exon 9 form	Metastatic adenocarcinoma	Koyama *et al.* (1999)
ATP7A	Loss of Golgi localization motif	Occipital horn syndrome	Qi and Byers (1998)
WT1 Wilms' tumor	Loss of +KS form	Frasier syndrome	Klamt *et al.* (1998)

with as yet unknown new disease-associated genes from the EST databases has made private EST collections a valuable commercial resource.

H. *Estimate of Rate of Alternative Splicing*

Taking sample sets of complete mRNAs or protein sequences and searching for possible alternative splice products can be semiautomated. The possible alternative splice forms found within ESTs are checked for mistakes arising from matches to paralogous sequences and possible pseudogenes. Additional filters are applied to the raw data to exclude repeat regions and contamination with vector sequence (Smit and Green, 1997). Given that ESTs are derived from a wide variety of human tissues and individuals, the number of possible alternative splice forms extracted from an EST database can be argued to give a reasonable estimate of the general level of alternative splicing occurring in human genes. In a recent study of this type, a sample of 475 proteins annotated in the SWISS-PROT database (Bairoch and Apweiler, 1999) as disease associated were searched against the EST database for the presence of possible alternative splice forms (Hanke *et al.*, 1999). After filtering the data to remove possible premature mRNAs or pseudogenes, 204 candidate sites were predicted from 162 of the proteins in the set. A final calculation of 34% of the proteins studied had a candidate alternative splice site. This initial study was extended to cover 8503 full-length mRNAs and confirmed the figure in the first study with an initial value of approximately 30% (work in progress). Table VII outlines the number and type of alternative splice forms detected within ESTs matching the collection of proteins.

Of these possible alternative splice forms, 70% were found to be exon skipping events, 30% had additional inserted sequence. The coverage of matching ESTs in the set of 475 proteins was approximately 50% of all positions only, and the average report was from about two different tissues per position. As a result of this low coverage, it could be argued that 30% is an underestimate of the true value. Interestingly, both these percentages are considerably higher than the previous estimates (≈5%: Sharp, 1994; Wolfsberg and Landsman, 1997). To what degree this represents reality in terms of alternative protein forms finally expressed at any one time in a given tissue type remains to be verified experimentally. In many cases different alternative splices forms coexist at a given ratio within the same cell. Whether or not the existence of a particular alternative splice form represents a functional protein is also open to question. It is quite possible that cells could tolerate quite high levels of incorrect alternative splicing if the half-lives of the mRNAs or peptides produced were relatively short and/or if the variants do not impair function.

VII. Conclusions

Studying variation in human genomic sequences may serve two general purposes: to characterize genetic population structure and its history and to elucidate the genotype/phenotype relationship in individuals or families. Both aspects are strongly interdependent, and it is only with the advent of new methods of individual sequencing on a mass scale that they become technically feasible.

Polymorphism in the coding part of the genome as well as in the regulatory perigenic regions may be a major factor explaining individual variation of the phenotypes associated with common diseases and with pharmacogenetically determined traits (Housman and Ledley, 1998). For common diseases, the current hypothesis is that relatively common

TABLE VII
Type Classification of Splice Variants Found in Expressed Genes

Starting sample	Exon skipping	Inserted DNA
475 proteins	282	45
8503 mRNAs	4893	1932

allele variants in a number of loci contribute to the causation and/or the susceptibility for relatively common traits that mark the widespread multifactorial diseases such as arteriosclerosis, diabetes, hypertension, and others. Chakravarti (1998, 1999) and Collins *et al.* (1997), among others, have outlined this "common variant–common disease" hypothesis whose confirmation or refutation is a major goal of individual human genomics. Single nucleotide polymorphisms are one of the possible factors in that complex network, and they well serve at two fronts: They either participate directly (if they are phenotype-modifying variants) in the pathogenesis, or serve as markers of genomic localization and suggest, if linked to a trait, the possible location of a contributing factor with which the SNP is in "linkage disequilibrium" (i.e., closely located and not yet reshuffled by meiotic recombination). The interpretation of the complex multifactorial network's behavior requires detailed understanding of the population structure in which the alleles arose and segregate, as well as its history. Linkage disequilibrium, for instance, is a transient phenomenon that disappears after a number of generations. Its use as direct mapping tool depends on the prehistory of the marker. Association, on the other hand, of a combination of genetic variants of neutral or slightly nonneutral character, may severely mislead the interpretation, if there is admixture in the control population. Interpretation of the disease-causing effect of variation must take into account that modern civilization tends to suppress the "purifying" effect of natural selection because it deals with factors that often have no influence on the reproductive fitness of its carriers. Myopia, once perhaps a seriously disadvantageous trait, is no longer under heavy selective pressure and may now behave like a selectively neutral trait. Similar arguments apply to organic diseases linked to nonstandard lipid metabolism: In the savanna intensive fat ingestion and effective assimilation may be favorable, but not in the zoo.

The present state of knowledge indicates that protein sequences are subject to genetic variation, in the range of a few SNPs per thousand sites (being silent or sense-changing on the protein level). The extent of influence of those individual variants on the physiology and pathology of the organism is to be elucidated in future.

REFERENCES

Altschul, S., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST anmd PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25,** 3389–3402.

Bairoch, A., and Apweiler, R. (1999). The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. *Nucl. Acids. Res.* **27,** 49–54.

Benson, D. A., Boguski, M. S., Lipman, D. J., Ostell, J., Ouellette, B. F., Rapp, B. A., and Wheeler, D. L. (1999). Genbank. *Nucl. Acids Res.* **27,** 12–17.

Bergstrom, T. F., Josefsson, A., Ehrlich, H. A., and Gyllensten, U. (1998). Recent origin of HLA-DRB 1 alleles and implications for human evolution. *Nat. Genet.* **18,** 237–242.

BLAST server (1999): http://www.ncbi.nlm.nih.gov/blast.

Boguski, M. S. (1995). The turning point of the history of human ESTs in genome research. *Trends Biochem. Sci.* **20,** 295–296.

Boguski, M. S., Lowe, T. M. J., and Tolstoshev, C. M. (1993) dbEST—data base for expressed sequence tags. *Nat. Genet.* **4,** 332–333.

Buetow, K. H., Edmonson, M. N., and Cassidy, A. B. (1999). Reliable identification of large numbers of candidate SNPs from public EST data. *Nat. Genet.* **21,** 323–325.

Cargill, M., Altschuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., Lane, C. R., Lim, E. P., Kalyanaraman, N., Nemesh, J., Ziaugra, L., Friedland, L., Rolfe, A., Warrington, J., Lipshutz, R., Daley, G. O., and Lander, E. S. (1999). Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat. Genet.* **22,** 231–238.

Chakravarti, A. (1998). It's raining SNPs, hallelujah? *Nat. Genet.* **19,** 216–217.

Chakravarti, A. (1999). Population genetics—making sense out of sequence. *Nat. Genet.* **21,** 56–60.

Chee, M., Yang, R., Hubbell, E., Berno, A., Huang, X. C., Stern, D., Winkler, J., Lockhart, D. J., Morris, M. S., and Fodor, S. P. A. (1996). Accessing genetic information with high-density DNA arrays. *Science* **274,** 610–614.

Clark, A. G., Weiss, K. M., Nickerson, D. A., Taylor, S. L., Buchanan, A., Stengard, J., Salomaa, V., Virtianen, E., Perola, M., Boerwinkle, E., and Sing, C. F. (1998). Haplotype structure and population genetic inferences from Nucleotide-sequence variation in human lipoprotein lipase. *Am. J. Hum. Genet.* **63,** 595–612.

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

Cooper, D. N., Smith, B. A., Cooke, H., Niemann, S., and Schmidtke, J. (1985). An estimate of unique sequence heterozygosity in the human genome. *Hum. Genet.* **69,** 201–205.

Ewing, B., Hillier, L., Wendl, M. C., and Green, P. (1998). Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. *Genome Res.* **8,** 175–185.

Ewing, B., and Green, P. (1998). Base-calling of automated sequencer traces using Phred. II. Error probabilities. *Genome Res.* **8,** 186–194.

Gillespie, J. H. (1994). Alternatives to the neutral theory. *In* "Non-Neutral Evolution: Theories and Molecular Data."(B. Golding, ed.) pp. 1–17. Chapman & Hall, New York.

Gordon, D., Abajian, C., and Green, P. (1998). Consed: a graphical tool for sequence finishing. *Genome Res.* **8,** 195–202.

Green, P. (1998). Phrap, sequence alignment and contig assembly program. http://genome.washington.edu.

Hacia, J. G. (1999). Resequencing and mutational analysis using oligonucleotide arrays. *Nat. Genet.* **21,** 42–47.

Hacia, J. G., Fan, J. B., Ryder, O., Jin, L., Edgemon, K., Ghandour, G., Mayer, R. A., Sun, B., Hsie, L., Robbins, C. M., Brody, L. C., Wang, D., Lander, E. S., Lipshutz, R., Fodor, S. P. A., and Collins, F. S. (1999). Determination of ancestral alleles for human single-nucleotide polymorphisms using high-density oligonucleotide arrays. *Nat. Genet.* **22,** 164–167.

Halushka, M. K., Fan, J. B., Bentley, K., Hsie, L., Shen, N., Weder, A., Cooper, R., Lipshutz, R., and Chakravarti, A., (1999). Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nat. Genet.* **22,** 239–247.

Hanke, J., Brett, D., Zastrow, I., Aydin, A., Delbrück, S., Lehmann, G., Luft, F., Reich, J., and Bork, P. (1999). *Trends Genet.* **15,** 389–390.

Harding, R. M., Fullerton, S. M., Griffiths, R. C., Bond, J., Cox, M. J., Schneider, J. A., Moulin, D. S., and Clegg, J. B. (1997) Archaic African and Asian lineages in the genetic ancestry of modern humans. *Am. J. Hum. Genet.* **60,** 772–789.

Harpending, H. C., Batzer, M. A., Gurvens, M., Jorde, L. B., Rogers, A. R., and Sherry, S. T. (1998). Genetic traces of ancient demography. *Proc. Natl. Acad. Sci. U.S.A.* **95,** 1961–1967.

Harris, H. (1966). Enzyme polymorphisms in man. *Proc. Roy. Soc. London* [*B*] **164,** 298–310.

Harris, H., and Hopkinson, D. A. (1972). Average heterozygosity per locus in man: an estimate on the incidence of enzyme polymorphisms. *Ann. Hum. Genet.* **36,** 9–20.

Hartl. D. L., and Clark, A. G. (1989). "Principles of Population Genetics." Sinauer Ass., Sunderland, MA.

Housman, D., and Ledley, F. D. (1998). Why pharmacogenetics? Why now? *Nat. Biotechnol.* **16,** 492–493.

Hudson, R. R. (1993). Levels of DNA polymorphism and divergence yield important insights into evolutionary processes. *Proc. Natl. Acad. Sci. U.S.A.* **90,** 7425–7426.

Ingram, V. M. (1956). A specific chemical difference between the globins of normal human and sickle cell anaemia haemoglobin. *Nature* **178,** 792–794.

Karlin, S., and Altschul, S. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci. U.S.A.* **87,** 2264–2268.

Karlin, S., and Altschul, S. (1993). Applications and statistics for multiple high-scoring schemes in molecular sequences. *Proc. Natl. Acad. Sci. U.S.A.* **90,** 5873–5877.

Kimura, M. (1962). On the probability of fixation of mutant genes in populations. *Genetics* **47,** 713–719.

Kimura, M. (1983). "The Neutral Theory of Molecular Evolution." Cambridge University Press, Cambridge, UK.

Kimura, M., and Otha, T. (1969). Average number of generations until fixation of a mutant gene in a finite population. *Genetics* **61,** 763–771.

Klamt, B., Koziell, A., Poulat, F., Wieacker, P., Scambler, P., Berta, P., and Gessler, M. (1998). Fraier syndrome is caused by defective alternative splicing of WT1 leading to an altered ratio of WT1 +/− KTS splice isoforms. *Hum. Mol. Genet.* **4,** 709–714.

Koyama, S., Maruyama, T., and Adachi, S. (1999). Expression of epidermal growth factor receptor and CD44 splicing variant sharing exons 6 and 9 on gastric and esophageal carcinomas: a two flow-cytometric analysis. *J. Cancer Res. Clin. Oncol.* **125,** 47–54.

Li, W. H. and Sadler, L. A. (1991). Low nucleotide diversity in man. *Genetics* **129,** 513–523.

Li, W. H. (1997). "Molecular Evolution." Sinauer Ass., Sunderland, MA.

Lipshutz, R. J., Fodor, S. P., Gingeras, T. R., and Lockhart, D. J. (1999) High density synthetic oligonucleotide arrays. *Nat. Genet.* **21,** 20–24.

McDonald, J., and Kreitman, M. (1991). Adaptive protein evolution at adh locus in *Drosophila. Nature* **351,** 652–654.

McKusick, V. (1999). Available as "NCBI-distributed Online Mendelian Inheritance in Man", http://www.ncbi.nlm.nih.gov/omim.

Moriyama, E. N., and Powell, J. R. (1996). Intraspecific nuclear DNA variation in *Drosophila. Mol. Biol. Evol.* **13,** 261–277.

Nei, M. (1975). "Molecular Population Genetics and Evolution." North Holland, Amsterdam.

Nickerson, D. A., Taylor, S. L., Weiss, K. M., Clark, A. G., Hutchinson, R. G., Stengard, J., Salomaa, V., Vartiainen, E., Boerwinkle, E., and Sing, C. F. (1998). DNA sequence diversity in a 9.7 kb region of the human lipoprotein lipase gene. *Nat. Genet.* **19,** 233–240.

Ozon, S., Byk, T., and Sobel, A. (1998). SCLIP: anovel SCG10-like protein of the stathmin family expressed in the nervous system. *J. Neurochem.* **70,** 2386–2396.

Pauling, L., Itano, H. A., Singer, S. J., and Wells, I. C. (1949). Sickle cell anemia: a molecular disease. *Science* **110,** 543–546.

Picoult-Newberg, L., Ideker, T. E., Pohl, M. G., Taylor, S. L., Donaldson, M. A., Nickerson, D. A., and Boyce-Jacino, M. (1999). Mining SNPs from EST databases. *Genome Res.* **9,** 167–174.

Qi, M., Byers, P. H. (1998). Constitutive skipping of alternatively spliced exon 10 in the ATP7A gene abolishes Golgi localization of the the Menkes protein and produces the occipital horn syndrome. *Hum. Mol. Genet.* **7,** 465–469.

Race, R. R., and Sanger, R. (1975). "Blood groups in man," 6th ed. Blackwell, Oxford, UK.

Ramsay, G. (1998). DNA chips: state of the art. *Nat. Biotechnol.* **16,** 40–44.

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

Sachse, G., Brockmöller, J., Bauer, S., and Roots, I. (1997). Cytochrome P450 2D6 variants in a Caucasian population: allele frequencies and phenotypic consequences. *Am. J. Hum. Genet.* **60,** 284–295.

Sadoulet-Puccio, H. M., Khurana, T. S., Cohen, J. B., and Kundel, L. M. (1996). Cloning and characterization of the human homologue of a dystrophin related phosphoprotein found at the Torpedo electric organ post-synaptic membrane. *Hum. Mol. Genet.* **4,** 44489–44496.

Sato, N., Hori, O., Yamaguchi, A., Lambert, J. C., Chartier-Harlin, M. C., Robinson, P. A., Delacourte, A., Schmidt, A. M., Furuyama, T., Tohyama, M., and Takagi, T. (1999). A novel presenilin-2 splice variant in the human Alzheimer's disease brain tissue. *J. Neurochem.* **72,** 2498–2505.

Schuler, G. D. (1996). A gene map of the human gene. *Science* **274,** 540–546.

Schuler, G. D. (1997). Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *J. Mol. Med.* **75,** 694–698.

Sharp, P. A. (1994). Split genes and RNA splicing. *Cell* **77,** 805–815.

Siffert, W., Rosskopf, D., Siffert, G., Busch, S., Moritz, A., Erbel, R., Sharma, A. M., Ritz, E., Wichmann, H. E., Jakobs, K. H., and Horsthemke, B. (1998). Association of a human G-protein beta3 subunit variant with hypertension. *Nat. Genet.* **18,** 45–48.

Sigalas, I., Calvert, A. H., Anderson, J. J., Neal, D. E., and Lunec, J. (1996). Alternatively spliced mdm2 transcript with loss of p53 binding domain sequences: transforming ability and frequent detection in human cancer. *Nat. Med.* **8,** 912–917.

Smit, A. F. A., and Green, P. (1997). Repeat Masker: http//ftp.genome.washington.edu/RM/RepeatMasker.html.

Smithies, O. (1955). Zone electrophoresis in starch gels: group variations in the serum proteins of normal human adults. *Biochem. J.* **61,** 629–641.

Taillon-Miller, P., Gu, Y., Li, Q., Hillier, L., and Kwok, P. Y. (1998). Overlapping genomic sequences: a treasure trove of single-nucleotide polymorphisms. *Genome Res.* **8,** 748–784.

UNIGENE web server (1999): http://www.ncbi.nlm.nih.gov/UniGene/index.html.

Wang, D. G., Fan, J. B., Siao, C. J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spence, J., Kruglyak, L., Stein, L., Hsie, L., Topaloglou, T., Hubell, E., Robinson, E., Mittmann, M., Morris, M. S., Shen, N., Kilburn, D., Rioux, J., Nusbaum, C., Rozen, S., Hudson, T. J., Lipshutz, R., Chee, M., and Lander,

E. S. (1998). Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280,** 1077–1082.

Winzeler, E. A., Richards, D. R., Conway, A. R., Goldstein, A. L., Kalman, S., McCullough, M. J., McCusker, J. H., Stevens, D. A., Wodlicka L., Lockhart, D. J., and Davis, R. W. (1998). Direct allelic vaiation scanning of the yeast genome. *Science* **281,** 1994–1197.

Wolfsberg, T. G., and Landsman, D. (1997). A comparison of expressed sequence tags (ESTs) two human genomic sequences. *Nucl. Acids Res.* **25,** 1626–1632.

Yamakawa, H., and Ohara, O. (1997). A DNA cycle sequencing reaction that minimizes compressions on automated fluorescent sequencers. *Nucleic Acid Res.* **25(6),** 1311–1312.

Zietkiewicz, E., Yotova, V., Jarnik, M., Korab-Laskowska, M., Kidd, KK., Modiano, D., Scozzari, R., Stoneking, M., Tishkoff, S., Batzer, M., and Labuda, D. (1997). Nuclear DNA diversity in world-wide distributed human populations. *Gene* **205,** 161–171.

ZuckerKandl, E., and Pauling, L. (1965). Evolutionary divergence and convergence in proteins, *In* "Evolving Genes and Proteins." (V. Bryson, and H. J. Vogel, eds.) pp. 189–225. Academic Press, New York.