ELSEVIER

Gene 234 (1999) 177–186

GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES

www.elsevier.com/locate/gene

Review

The essence of SNPs

Anthony J. Brookes *

Department of Genetics and Pathology, Biomedical Center, Uppsala University, 751 23 Uppsala, Sweden

Received 26 February 1999; received in revised form 18 May 1999; accepted 26 May 1999

Abstract

Single nucleotide polymorphisms (SNPs) are an abundant form of genome variation, distinguished from rare variations by a requirement for the least abundant allele to have a frequency of 1% or more. A wide range of genetics disciplines stand to benefit greatly from the study and use of SNPs. The recent surge of interest in SNPs stems from, and continues to depend upon, the merging and coincident maturation of several research areas, i.e. (i) large-scale genome analysis and related technologies, (ii) bioinformatics and computing, (iii) genetic analysis of simple and complex disease states, and (iv) global human population genetics. These fields will now be propelled forward, often into uncharted territories, by ongoing discovery efforts that promise to yield hundreds of thousands of human SNPs in the next few years. Major questions are now being asked, experimentally, theoretically and ethically, about the most effective ways to unlock the full potential of the upcoming SNP revolution. © 1999 Published by Elsevier Science B.V. All rights reserved.

Keywords: Association studies; Haplotype; Linkage disequilibrium; Polymorphism

1. Introduction

The Human Genome Project (HGP) is progressing rapidly, with over one million partial cDNA sequences and approximately 10% of a 'reference' genomic sequence now in public databases. With this advance has come an appreciation of the need to also study naturally occurring sequence variations, i.e. to understand human DNA polymorphism, about 90% of which is single nucleotide polymorphism (SNP) (Collins et al., 1998). Significant efforts towards large-scale characterisation of human SNPs have been initiated in the last year or so, a somewhat late stage given that almost two decades ago the original incarnation of SNPs [as restriction fragment length polymorphisms (RFLPs)] clearly indicated the existence of widespread subtle genome variation. Now, the renewed and extensive interest in genome polymorphism signifies a development in human genetics research that will have a major impact upon population genetics, drug development, forensics, cancer and genetic disease research. One consequence of all this activity is that the acronym 'SNP' (pronounced 'S' 'N' 'P' or 'SNiP') has appeared in many diverse articles and reviews, leading many to ponder "what are SNPs and why all the fuss?". This review is an attempt to answer these questions.

We can start with a working definition — SNPs are single base pair positions in genomic DNA at which different sequence alternatives (alleles) exist in normal individuals in some population(s), wherein the least frequent allele has an abundance of 1% or greater. Thus, single base insertion/deletion variants (indels) would not formally be considered to be SNPs. In principle, SNPs could be bi-, tri-, or tetra-allelic polymorphisms. However, in humans, tri-allelic and tetra-allelic SNPs are rare almost to the point of non-existence, and so SNPs are sometimes simply referred to as bi-allelic markers (or di-allelic to be etymologically correct). This is somewhat misleading because SNPs are only a subset of all possible bi-allelic polymorphisms (e.g. indels, multiple base variations).

In practice, the term SNP is typically used more loosely than required by the above definition. Single base variants in cDNAs (cSNPs) are usually classed as SNPs since most of these will reflect underlying genomic

* Present address: Center for Genomics Research, Karolinska Institute, Doktorsringen 2P, 171 77 Stockholm, Sweden.
Tel.: +46-18-4714151; Fax: +46-18-526849.
E-mail address: tony.brookes@medgen.uu.se (A. Brookes)

PII: S0378-1119(99)00219-X

DNA variants. This, however, ignores the possibility that they may be the result of RNA editing. Genomic DNA indels involving single or multiple bases are commonly discovered in SNP search efforts and so can become deposited in SNP lists and databases. In a similar way, such data-sets also contain SNP variants of less than 1% allele frequency. Complications with the above definition also exist. Specifically, some people might not want to consider disease predisposing single base variants to be SNPs — but the above definition would encompass such things as recessively acting, low penetrance dominant, quantitative trait loci, or risk associated alleles, since all of these will occur in some normal (non-diseased) individuals. Also the 'some population' component of the definition is limited by practical challenges of attaining and surveying representative global population samples. Consequently, claims of non-polymorphic sequences should always be accompanied by statements of the actual populations and the numbers of chromosomes tested. Overall, it is therefore apparent that the term 'SNP' is being widely and imprecisely used as a catch-all label for many different types of subtle sequence variation. To maintain clarity within this review, I shall restrict myself to the SNP definition given above. I shall also use the term polymorphism consistently and correctly to refer to the set of alleles at a locus, rather than to any one allele alone.

2. SNP basics

Bi-allelic SNPs comprise four distinct types. Using the abbreviation X⇔Y (X_1⇔Y_1) to represent allelic nucleotides X and Y of an SNP on one DNA strand, with their base pairing nucleotides X_1 and Y_1 of the second strand shown in parentheses, then the four SNP alternatives include one transition C⇔T (G⇔A) and three transversions C⇔A (G⇔T), C⇔G (G⇔C), and T⇔A (A⇔T). This four-way classification is valid if one considers each DNA strand to be equivalent, so C⇔T (G⇔A) is an identical 'mirror image' or sequence complement of G⇔A (C⇔T). However, in certain situations, such as the analysis of DNA replication or transcription, the two DNA strands must be distinguished. In these cases, two of the four SNP types [C⇔T (G⇔A) and C⇔A (G⇔T)] must each be separated into two SNP subtypes, yielding a total of six fully distinct alternatives. The frequencies of the four basic SNP types in the human are not equal, with most SNPs (about 2/3) involving the C⇔T (G⇔A) variety, while the other three types occur at similar levels to each other to comprise the remainder. The higher level of C⇔T (G⇔A) SNPs is probably partly related to 5-methylcytosine deamination reactions that are known to occur frequently, particularly at CpG dinucleotides (Holliday and Grigg, 1993). For this reason, just as with the known spectrum of sequence variants that underlie disease (Krawczak et al., 1998), the abundance of C⇔T (G⇔A) SNPs in particular will perhaps be higher in the gene and C+G rich isochores, though this is yet to be confirmed.

The typical frequency with which one observes single base differences in genomic DNA from two equivalent chromosomes is of the order of 1/1000 bp (Li and Sadler, 1991; Wang et al., 1998; Lai et al., 1998; Nickerson et al., 1998; Harding et al., 1997; Taillon-Miller et al., 1998). Many, but not all of these, will be polymorphisms for which the least abundant allele is present at or above 1% frequency in the tested population, the level required for designation as an SNP. Alleles of lower frequency will be examples from a sea of 'rare variants' in the population, each of which will be represented by only a small number of (or individual) chromosomes. The rate of nucleotide difference between two randomly chosen chromosomes is an index termed nucleotide diversity (Nei and Li, 1979). Simplistically speaking, the 1/1000 figure means that there is an average 0.1% chance of any base being heterozygous in an individual. Of course, by screening more individuals (more chromosomes), more base differences can be found, but the nucleotide diversity index remains unchanged. Within coding exons the nucleotide diversity is some four-fold lower, with about half resulting in non-synonymous codon changes (Li and Sadler, 1991; Nickerson et al., 1998). Genome-wide there are region specific differences in SNP density that are at least as great as 100-fold. For example, nucleotide diversity in some regions is way below 0.1% (Nachman et al., 1998), whilst some peculiar non-coding HLA regions show nucleotide diversity levels of 5–10% (Guillaudeux et al., 1998; Horton et al., 1998). Overall then, these numbers add up to several million single base differences between any two individuals and something like 100 000 amino acid differences between their proteomes. When this is compared to the only 10-fold greater degree of difference that exists between human and chimpanzee genomes, the enormous functional relevance of so many SNPs becomes strikingly apparent.

To understand present day SNPs more fully, it helps to consider them in an evolutionary context relative to the time of divergence of humans and chimpanzees (~5 million years ago) (Kumar and Hedges, 1998; Takahata, 1995), and the time at which modern humans are believed to have spread globally from a common ancestral population in Africa (0.1–0.2 million years ago) (Stoneking, 1997; Hammer and Zegura, 1996). Although genomic DNA sequence variations are created continuously at a rate of some 100 new single base changes per individual (Kondrashov, 1995; Crow, 1995) (typically to then be eliminated by drift or remain for a period as rare variations), most present day human SNPs (i.e. with 1% minimum allele frequencies) originated long after speciation but before the emergence of different populations (Mountain et al., 1992). Also, one needs to consider the low rate (about 10^{-8} changes per nucleotide per generation) and essentially random nature of base changing events (Crow, 1995; Li et al., 1996), which together make single-base alleles very stable. Thus, few of the SNP alleles that were present when humans emerged from Africa will have yet become fixed (reached 0% or 100%). The consequence of all this is that human SNPs are generally not shared with our primate cousins, but most ($\sim$85%) are common to all human populations (with differing allele frequencies) with only 15% or so being population 'private' (Barbujani et al., 1997). Hence it is often stated that the majority of human genome variance is represented within rather than between populations.

3. SNP discovery and scoring

Significant efforts towards large-scale SNP discovery have now begun, in what started as something of a hectic race between industry and academia. Both camps appreciate the functional importance and practical utility of SNPs, and whilst the former is keen to secure intellectual property protection on them, the latter would generally like them to be available to all as a research tool. With so many SNPs out there to be gathered and no real indication as to which will be the most useful (with the possible exceptions of cSNPs and promoter region SNPs), one wonders why there has been so much frantic competition. Fortunately, there are now some encouraging signs of mutually beneficial partnerships aimed at jointly discovered and shared SNP data-sets, a move that logically couples the upstream funding and discovery potential of industry to the downstream broad-range functional research activities of academia.

As of April 1999, there were some 7000 SNPs in the public domain, about half of which were cSNPs. But this number is increasing rapidly. Some large-scale discovery endeavors set for completion within the next two to three years include: (i) a US National Institutes of Health funded program (http://www.nhgri.nih.gov/About_NHGRI/Der/variat.htm; Marshall, 1997) expected to yield over 50 000 SNPs (cDNA and genomic); (ii) a private effort by Genset towards 60 000 genomic SNPs, for public release once intellectually protected (Marshall, 1997); (iii) some 30 000 SNPs from overlapping genomic clones sequenced by academia (Taillon-Miller et al., 1998), plus many-fold more once private shotgun sequence data becomes available (Venter et al., 1998); (iv) in silico extraction of cSNPs from multiply redundant cDNA sequences in the dbEST division of Genbank and other similar databases (perhaps 10 000 cSNPs) (Gu et al., 1998; Picoult-Newberg et al., 1999; Buetow et al., 1999); (v) a joint academic/drug industry venture called The SNP Consortium (TSC) committing $45 million over two years to find 300 000 SNPs and map half of these (Marshall, 1999); and (vi) proprietary efforts by various companies that together might yield many hundred thousand genomic and cDNA SNPs. In all, it is therefore reasonable to expect many hundred thousand SNPs to enter the public domain in just a few years, with still more residing in private databases. However, one major missing component in all of this is a generally agreed, high utility, and readily accessible series of global population sample DNAs within which allele frequencies can be determined. A first step towards such a resource is being made by the National Human Genome Research Institute and others (http://www.nhgri.nih.gov:80/Grant_info/Funding/RFA/discover_polymorphisms.html; Collins et al., 1998), but there are some serious concerns regarding the practical value of these samples which, for ethical reasons, are being stripped of all population descriptors.

Ready access to the rapidly growing mass of SNP data is a prerequisite to its effective utilisation. To this end, public SNP databases are being constructed. There are, of course, a range of generic and locus specific disease mutation databases that carry some SNP data (given the uncertain distinction between disease predisposing single base variants and SNP alleles), and a number of individual discovery efforts offer access to their own data via dedicated Web pages (http://www-genome.wi.mit.edu/SNP/human/index.html; http://www.ibc.wustl.edu/SNP/; http://www.chlc.org/cgap/nature_genetics_snps.html). However, to date there exist only two major public databases that attempt to provide a comprehensive summary of human SNPs. One is dbSNP (http://www.ncbi.nlm.nih.gov/SNP/), an archival database designed to provide full details of discovered genomic and cDNA SNPs from any species, including methods of assay and discovery, and flanking sequence PCR conditions. The second is the Human Genic Bi-Allelic Sequences (HGBASE) database (Sarkar et al., 1998; http://hgbase.interactiva.de/) which focuses upon the relationship between SNPs and gene function. HGBASE therefore provides details of human gene related (promoter, exonic and intronic) SNPs, and will in time provide details of facile scoring assays and information on gene expression and disease relationships. It is anticipated that within a decade most non-synonymous human cSNPs will be compiled within HGBASE. The databases dbSNP and HGBASE are both freely available to the public, and will mutually exchange SNP data at regular intervals to fully benefit the research community.

SNP detection and scoring methods are many and various (Landegren et al., 1998). A thorough account of these is beyond the scope of this review, but a few general words might be useful. The two recurrent themes in the various assay designs are elegant simplicity (as typified by Dynamic Allele Specific Hybridisation) (Howell et al., 1999) and advanced technology (e.g. micro-fabricated hybridisation arrays) (Hacia et al., 1998), two things that need not be mutually exclusive. Currently, most procedures involve target sequence PCR amplification, a costly and time-consuming burden that limits possibilities for scale-up and automation. This is equally true for the much acclaimed miniature hybridisation array (DNA-chip) concept. Removal of the PCR step would be highly desirable, and this may be possible with some of the newest assay concepts (Nilsson et al., 1997; Lizardi et al., 1998; Lyamichev et al., 1999). Genomic level scoring of cSNPs presents another major hurdle, and one that is perhaps not widely recognised. The key problem is the possible existence (usually unknown) of processed pseudogenes that are highly similar in sequence to the target gene, but devoid of introns. Since exon–intron structures are presently unknown for many target genes, it can be impossible to reliably design assays that will not also, or even solely, interrogate the processed pseudogene(s). Personal experience suggests this may be an obstacle for as many as 20% of all cSNP assays. Furthermore, even knowing the detailed structure of a target gene does not necessarily enable one to design around the problem, since assays for SNPs within large exons may be impossible to construct in such a way as to exploit intronic (true gene specific) sequences. Finally, most current scoring assays involve allele discrimination via, or secondary to, matched and mismatched base pair detection at the SNP locus. The most stable mismatched base pair in such assays is G:T, which is almost as stable as its A:T counterpart (Ikuta et al., 1987). Ironically, it is precisely this mismatch that one needs to distinguish in order to score the most abundant of the four SNP types, [C⇔T (G⇔A)].

4. Population genetics and linkage disequilibrium

Population genetics is the study of the genetic composition and inter-relationships between populations. The major research tool it uses is DNA polymorphism. Unfortunately, population genetics and human molecular genetics have in some ways been running along parallel research paths, with much population genetics effort over the last few decades being directed towards non-human organisms. With the new SNP era, these fields are beginning to interact far more closely. Population genetics researchers will be able to exploit appropriate sets of SNP markers which, due to their abundance, stability and ease of scoring, will allow them to undertake far more detailed and rapid human genome studies than were previously possible. And the great wealth of accumulated population genetics understanding can be incorporated into human molecular genetics studies in order to best exploit SNPs for effective analysis of genotype–phenotype relationships.

To appreciate the role of SNPs in population genetics one must be familiar with the concept of linkage disequilibrium (LD), the principles of which are shown in Fig. 1. Most SNPs in modern humans probably arose by single base modifying events that took place within single DNA molecules (chromosomes) a long time ago. A single newly created allele, at its time of origin, would have been surrounded by a series of alleles at other polymorphic loci. Thus, at the instance of creation, a unique grouping of alleles (a haplotype) was established. As (and if) the surrounding chromosomal region became replicated in the next few generations, the haplotype would probably remain intact. In this situation, complete LD would be said to exist between the new allele and each of the nearby polymorphisms — meaning that the new allele would be 100% predictive of the alleles present at these nearby polymorphic sites. Thus, the existence of LD enables an allele of one polymorphic marker to be used as a surrogate for a specific allele of another.

Unfortunately, LD is not stable over long time periods. With successive generations the level of LD between two markers will typically decrease. This is due to meiotic recombination events (that may be non-randomly distributed with enrichment at 'hotspots') which, by exchanging polymorphism carrying portions of sister chromosomes, will tend to shuffle alleles at different loci along the DNA. More closely positioned loci are less prone to this effect as recombination events are simply less likely to occur between them. Similarly, gene conversion events may also change the pattern of LD. Through these effects, LD will be lost with time. Other forces, however, can act to create or preserve LD. Specifically, random drift of haplotype frequencies may occur, thereby increasing LD. This is most likely to occur in smaller populations of stable size (Slatkin, 1994; Laan and Pääbo, 1997). Also, the action of natural selection against or for certain sequences would concomitantly drive alleles of adjacent loci (that were in prior LD) to much higher or lower frequencies, thus raising the total LD in that particular genomic region (Terwilliger et al., 1998).

LD is thus a complex phenomenon, and one which is of great interest to population geneticists. Its regional distribution will reflect not only the biological processes mentioned above, but also population specific demographic history, such as bottlenecks, admixture, inbreeding, migration, immigration, and assortative mating (Terwilliger et al., 1998). Analysis of all this in humans was previously confounded by the lack of a suitably dense series of readily scorable polymorphic markers that would enable comparison of chromosome portions with sufficient resolution. Now, the expectation is that this can be remedied by the use of SNPs. These very stable and abundant markers, including both global and

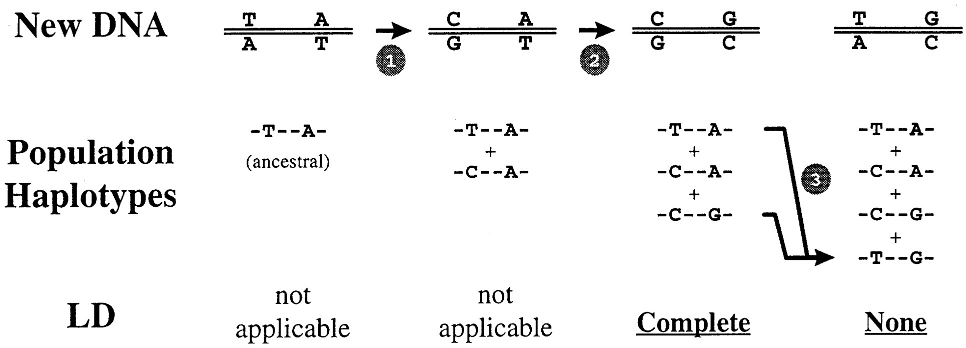

Fig. 1. Creation and loss of linkage disequilibrium. An ancestral chromosomal fragment is shown giving details for two separate base pair positions. Unless both positions are polymorphic the concept of LD does not apply. Event 1 involves a change [T⇔C (A⇔G)] of the first base pair to create one SNP and a new (second) haplotype that is assumed to increase to an appreciable frequency in the population. Then, event 2 changes [A⇔G (T⇔C)] the other base pair to create the second SNP and yet another (third) haplotype that also increases to an appreciable frequency. The choice of which haplotype to show modified by the second event is arbitrary, but rarely will both be altered. At this stage complete LD exists between the two SNP positions — thus, a 'T' at the first position is always accompanied by an 'A' at the second position, and a 'G' at the second position likewise by a 'C' at the first. Event 3 represents subsequent meiotic recombination (and/or gene conversion) activity through which the fourth possible haplotype is formed and LD is lost. Many factors affect the degree to which event 3 goes to completion.

population specific examples, are poised to facilitate a rapid advance of this important field.

5. Complex phenotypes and genome variation

The myriad of human phenotype variations one might wish to study are likely to be caused by genetic and non-genetic (environmental) factors, as well as by an interplay between the two and even a sprinkling of chance events. Clearly, many clinical phenotypes do seem to have a considerable genetic component. The underlying genetic factors of relevance will be encoded in the spectrum of genomic variation that is primarily SNPs. Thus, risks of major common diseases such as cancer, cardiovascular disease, mental illness, autoimmune states, and diabetes, are expected to be heavily influenced by the patterns of SNPs one possesses in certain key susceptibility genes yet to be identified. The same reasoning can be applied to gene based inter-individual variations in drug responses, a research area termed pharmacogenomics that is of great interest to the pharmaceutical industry. In many cases, genetic epidemiology data from twin, adoption, and family studies strongly support the above ideas with high numerical indices (heritability values) of the degree of total genetic contribution to disease causation in present day environments. For example, 74% of the risk of suffering Late Onset Alzheimer's Disease is estimated to be genetic (Gatz et al., 1997), and the heritability figures for Attention Deficit Hyperactivity Disorder and Autism are even higher (Folstein and Rutter, 1977; Stevenson, 1992).

For practical purposes, a somewhat artificial distinction can be made between sequence variants that strongly predispose to disease [e.g. Cystic Fibrosis gene defects (Rommens et al., 1989; Kerem et al., 1989)], and SNP alleles. Whereas the former can be considered to *cause* disease (with variable penetrance), the latter are imagined to merely *modify risk*. Clearly, this is really only a matter of degree, and the truth is that the two alternatives are opposite ends of a spectrum. But the point here is that single disease related SNP alleles alone are neither necessary nor sufficient to cause illness. Instead, it is probably the combined effect of a collection of SNP alleles in sets of key genes, plus environmental factors, that together determine whether an individual suffers some disease. Hence the term 'complex disease' is often used to describe these scenarios. The level of this complexity could potentially be enormous. For example, the number of interacting key genes could be a few, a few tens, or even a few hundred (oligogenic to polygenic). There could be many different predisposing risk alleles in these genes (allelic heterogeneity). Different or overlapping sets of genes could be important in different affected individuals (locus heterogeneity). Interactions between the genes could be additive, synergistic, or epistatic. Resulting pathologies could be quantitative rather than all-or-nothing traits, with thresholds determining clinical manifestation. And layered upon all this is the effect of the environment, and interactions therewith.

Attempts to understand the genetic basis of disease have previously tackled simpler single gene disorders. Here, the strategy of positional cloning (Collins, 1992), anchored typically upon initial linkage findings (meiotic mapping) to localise the mutant locus, has been very successful. In complex disorders there are sometimes rarer and usually more severe versions of the illness that are due to single gene defects. These provide a great way into the genetic etiology of the disease via traditional linkage analysis, for example in Alzheimer's Disease

(Goate et al., 1991; Sherrington et al., 1995; Levy-Lahad et al., 1995; Rogaev et al., 1995). However, when attempts have been made to use linkage mapping on the more complex disease forms, practical problems such as late onset of illness, lack of sufficient large informative families, or uncertain diagnoses, can severely impede progress. Worst still, for some common diseases it may be impossible to successfully apply linkage analysis due to the existence of too much multi-locus etiology, i.e. no single locus contributing enough of the disease causation in affected individuals for it to stand out in any human meiotic mapping study of practical scale. For these reasons, many laboratories are now directing considerable attention, and appropriately cautious certitude, towards an alternative strategy known as association analysis.

6. SNP based association studies

If a factor contributes an increased risk for disease occurrence, then that factor should be found at higher frequency in individuals with that disease compared to non-diseased controls, i.e. associated with the phenotype. A non-genetic example would be smoking which is associated with lung cancer (Vial, 1986), and a good genetic example would be the $\epsilon4$ allele of the apolipoprotein E gene (*APOE*4) which is associated with Alzheimer's Disease (Strittmatter and Roses, 1995). In common diseases that are due to high frequency, low risk (<10-fold) alleles, an association signal from a disease locus allele can be far greater than any produced by a linkage analysis (Greenberg, 1993; Hodge, 1994; Risch and Merikangas, 1996). For example, being homozygous compared to being a non-carrier for the $\epsilon4$ allele of the *APOE* gene is genetically associated with an approximately 10-fold increased risk for AD (Strittmatter and Roses, 1995), but family linkage studies with this locus do not produce convincing signals (Liu et al., 1996). The process of performing an association study involves simply determining the frequency of a test factor (e.g. an SNP allele) in many patients and age and race matched controls. The validity of this test will then depend critically upon an appropriate selection of these patients and controls. Population case-control studies are somewhat vulnerable to inappropriate patient-to-control matching (population stratification), and family based alternatives have been suggested. These include the Haplotype Relative Risk method (Khoury, 1994) (employing non-transmitted parental alleles as controls), and the Transmission Disequilibrium Test (Spielman et al., 1993) (comparing allele transmission rates from a heterozygous parent).

Ultimately, studies into disease genetics are trying to determine precisely which genomic sequence variants alter function and so have a 'pathogenic effect'. To find examples of such variants that are relevant to many individuals one could develop assays for all human SNPs and score these in large sets of patients plus matched non-diseased controls for all the complex phenotypes one wished to understand. However, present day costs and logistical problems aside, there are two reasons why this experiment might be problematic. First, the architectural complexity of a complex disease may, in many cases, be so elaborate (entail too many different, interacting and weak risk factors) that no genuine strong signals would exist to be detected. This concern is actually quite a disturbing 'big unknown' that may truly apply in some situations (Terwilliger and Weiss, 1998). However, the relatively small number of non-synonymous cSNPs that exist per gene (Wang et al., 1998; Lai et al., 1998; Nickerson et al., 1998; Harding et al., 1997; http://hgbase.interactiva.de/), compared to the wide spectrum of observed disease predisposing alleles for some disease loci (Krawczak et al., 1998), probably indicates that allelic heterogeneity may not be a limiting factor in so many cases. Indeed, one might reasonably expect that, contrary to the situation for Mendelian type disease variants, the weakly deleterious and perhaps late age effects of presumed cSNP alleles with pathogenic influence would allow some to be tolerated by natural selection and thus drift to high frequencies — and it is precisely these variants that we can hope to detect by association analysis. The second problem with comprehensive SNP association analysis would be that since each marker investigated would be essentially an independent test, screening millions of markers would lead to thousands of confounding chance (false) associations at any reasonable significance threshold, obscuring any real signals. This problem, however, can be minimised by thoughtful experimental design (see Fig. 2).

One way to perform association studies more effectively is to limit, by careful pre-selection, which SNPs are tested for pathogenic effect. The basis of the selection might be to focus upon biologically defined candidate genes, genes suggested by differential display experiments, or positional candidates from prior linkage investigations. Employing SNPs that are more likely to have functional consequences, such as non-synonymous cSNPs and promoter variants, is obviously sensible. For example, in our research (neurodegenerative disorders) we are using association analysis to investigate 250 component genes from four candidate pathways, and our reasonable goal is to extend this to all human cSNPs as these are discovered in future years. Studies of this type worldwide are now yielding many statistically positive associations. To determine which of these represent true or false signals will require numerous independent replication studies and a pro-active willingness by journals to also publish negative association data. Requiring multiple replications of positive signals will then become an effective way to sequentially filter out the many false

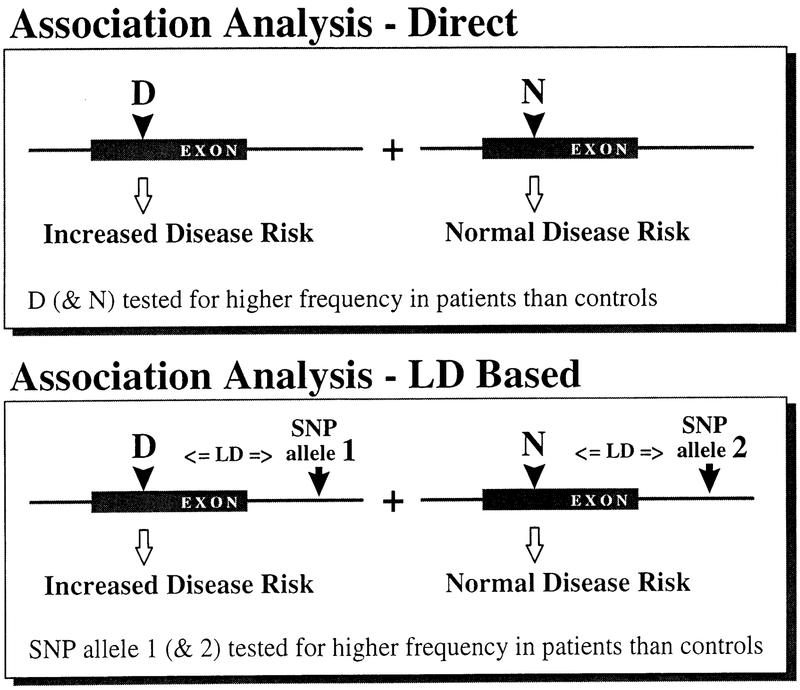
Association Analysis - Direct
D
N
EXON
+
EXON
Increased Disease Risk
Normal Disease Risk
D (& N) tested for higher frequency in patients than controls
Association Analysis - LD Based
D
<= LD =>
SNP allele 1
N
<= LD =>
SNP allele 2
EXON
+
EXON
Increased Disease Risk
Normal Disease Risk
SNP allele 1 (& 2) tested for higher frequency in patients than controls

Fig. 2. Strategies for association analysis. A pathogenic allele (D) and its non-pathogenic counterpart (N) are shown in a disease predisposing gene. An exonic variant is given by way of example, but disease predisposing changes in control and other regions could also be imagined. In direct association analysis, one tests polymorphisms that are candidates for the pathogenic sequence itself, e.g. cSNPs that involve non-synonymous codon changes. In LD based association analysis, one tests random SNP alleles distributed over a large genomic region and relies upon the existence of sufficient LD between one or more of the tested markers and the pathogenic sequence. Logically, positive signals from direct association analysis could equally be due to pathogenic sequences in LD with, but distinct from, the tested variant.

associations that we will have to accept when using any reasonable (i.e. not excessively stringent) significance level threshold.

An alternative way to improve association study design is to try to exploit linkage disequilibrium around pathogenic alleles. In principle, when strong LD is present between an SNP marker and an unknown (typically nearby) pathogenic allele, then both may show a similar association with the disease. One can exploit this by using a series of random SNPs (or other polymorphic sequences) to scan the vicinity of each marker for association signals that would indicate some DNA sequence nearby was having a pathogenic effect. In fact, this inference (that some other sequence nearby was pathogenic) should logically be applied to any and all positive associations, even those based upon non-synonymous cSNPs or similar. But in LD based association studies, the explicit idea is to employ a panel of random polymorphic markers and depend upon there being sufficient LD between those located near the target pathogenic locus and the pathogenic variant itself. This can be very effective in isolated and recently expanded populations, where founder risk alleles (in founder haplotypes) may have become abundant (Houwen et al., 1994; Friedman et al., 1995; Laan and Pääbo, 1998). But for other populations, to depend upon LD is a bold thing to do, since current evidence indicates that LD tends to be far from predictable, evenly distributed or strong at anything over a few thousand base pairs (Clark et al., 1998; Kidd et al., 1998 ; Tishkoff et al., 1996; Tishkoff et al., 1998; Harding et al., 1997; Laan and Pääbo, 1997). Nevertheless, the idea of using up to 100 000 or so well mapped SNP markers (still merely an average of one per gene!) from around the genome for comprehensive LD based association studies is being considered (Lai et al., 1998; Risch and Merikangas, 1996). For this to possibly work, very carefully selected test populations with homogeneous disease etiology and high intrinsic LD will need to be employed (Terwilliger et al., 1998; Terwilliger and Weiss, 1998), and evolutionary relationships between haplotypes considered (Sing et al., 1992; Templeton, 1996). More reasonably, the above scanning principle may be applied to home in on disease genes that are initially localised to chromosomal regions by family linkage analysis. Several reports show that this approach has very real potential, at least in some populations (Kerem et al., 1989; Puffenberger et al., 1994; Hastbacka et al., 1994; Lehesjoki et al., 1993; Kestila et al., 1994; Sulisalo et al., 1994; Jorde et al., 1994).

In the long term, the real potential of association studies may depend principally upon the nature of LD in modern human populations. Limited detailed analysis of small chromosomal regions so far indicates that the very different private histories of individual SNPs causes LD to vary greatly for different marker pairs in any one physically linked group (Harding et al., 1997; Clark et al., 1998). Indications are that small, old, stable populations have more intrinsic LD than recently expanded populations and so the former might be the preferred study group of the two (Terwilliger et al., 1998). However, due to founder effects, the latter will often contain high frequency representations of what would normally be low frequency disease alleles (and haplotypes) in other populations. The optimal choice would therefore come down to the 'big unknown' of the true level of complexity of complex diseases. What is already clear and will hopefully be taken on board by all in the field, is that: (i) test populations from which cases and controls are taken should be as homogenous, well defined and characterised as possible; (ii) studies that can combine linkage and association analyses are by far the most powerful; (iii) negative associations with any one or a few markers cannot be taken to exclude nearby sequences as risk loci; and (iv) the vicinity of the strongest of a series of positively associated markers does not necessarily define or even localise the pathogenic sequence.

7. Conclusions

An SNP revolution has begun which promises to challenge and stimulate DNA technologists, population geneticists, and molecular genetics researchers alike, and should bring them closer together than ever before. The field is new and important, with the consequence that much money is being spent with some very different ideas about what are the best initial experiments to perform. Industry is a major player, but joining forces with academia could be the most effective way to reach their goals, as well as to avoid the wasting of effort on proprietary and secret but false associations. Major activities today concern SNP discovery. It is perhaps a matter for concern that in most cases the SNPs being discovered comprise random higher frequency (old) alleles with global representations. These will tend to have low levels of LD with surrounding sequences and so may not be ideal for LD based association studies. Designing ways to collect younger, population specific SNPs might be more productive, and indeed, population geneticists might also be better served by placing more emphasis upon such SNPs. Assembling many thousands of SNPs has to be a current priority, but assembling a large white elephant (if that is what the current efforts amount to) might not be the best way to start. Careful thought and continued flexibility in discovery programs are therefore essential, and certainly an increased emphasis upon cSNPs and promoter SNPs would be welcomed. The characterisation of human SNPs and their role in phenotypic determination represents a truly milestone project, for while human beings are clearly much more than just 'bags of DNA', perhaps on the individual level we are little more than 'sacks of SNPs'.

Acknowledgements

Thanks are given to J.D. Terwilliger for his critical reading of this manuscript. Input from members of our research group is recognised and appreciated. Generous financial support, provided by the Swedish Medical Research Council, Professor U. Pettersson and the Beijer Foundation, for research activities bringing the author into this field are gratefully acknowledged.

References

Barbujani, G., Magagni, A., Minch, E., Cavalli-Sforza, L.L., 1997. An apportionment of human DNA diversity. Proc. Natl. Acad. Sci. USA 94, 4516–4519.

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

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

Collins, F.S., 1992. Positional cloning: let's not call it reverse anymore. Nature Genet. 1, 3–6.

Collins, F.S., Brooks, L.D., Chakravarti, A., 1998. A DNA polymorphism discovery resource for research on human genetic variation. Genome Res. 8, 1229–1231.

Crow, J.F., 1995. Spontaneous mutation as a risk factor. Exp. Clin. Immunogenet. 12, 121–128.

Folstein, S., Rutter, M., 1977. Genetic influences and infantile autism. Nature 265, 726–728.

Friedman, T.B., Liang, Y., Weber, J.L., Hinnant, J.T., Barber, T.D., Winata, S., Arhya, I.N., Asher Jr., J.H., 1995. A gene for congenital, recessive deafness DFNB3 maps to the pericentromeric region of chromosome 17. Nature Genet. 9, 86–91.

Gatz, M., Pedersen, N.L., Berg, S., Johansson, B., Johansson, K., Mortimer, J.A., Posner, S.F., Viitanen, M., Winblad, B., Ahlbom, A., 1997. Heritability for Alzheimer's disease: the study of dementia in Swedish twins. J. Gerontol. 52, M117–M125.

Goate, A., Chartier-Harlin, M.C., Mullan, M., Brown, J., Crawford, F., Fidani, L., Giuffra, L., Haynes, A., Irving, N., James, L., et al., 1991. Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer's disease. Nature 349, 704–706.

Greenberg, D.A., 1993. Linkage analysis of 'necessary' disease loci versus 'susceptibility' loci. Am. J. Human Genet. 52, 135–143.

Gu, Z., Hillier, L., Kwok, P.Y., 1998. Single nucleotide polymorphism hunting in cyberspace. Human Mutat. 12, 221–225.

Guillaudeux, T., Janer, M., Wong, G.K., Spies, T., Geraghty, D.E., 1998. The complete genomic sequence of 424 015 bp at the centromeric end of the HLA class I region: gene content and polymorphism. Proc. Natl. Acad. Sci. USA 95, 9494–9499.

Hacia, J.G., Brody, L.C., Collins, F.S., 1998. Applications of DNA chips for genomic analysis. Mol. Psychiatry 3, 483–492.

Hammer, M.F., Zegura, S.L., 1996. The role of the Y chromosome in human evolutionary studies. Evol. Anthropol. 5, 116–134.

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

Hastbacka, J., de la Chapelle, A., Mahtani, M.M., Clines, G., Reeve-Daly, M.P., Daly, M., Hamilton, B.A., Kusumi, K., Trivedi, B., Weaver, A., et al., 1994. The diastrophic dysplasia gene encodes a novel sulfate transporter: positional cloning by fine-structure linkage disequilibrium mapping. Cell 78, 1073–1087.

Hodge, S.E., 1994. What association analysis can and cannot tell us about the genetics of complex disease. Am. J. Med. Genet. 54, 318–323.

Holliday, R., Grigg, G.W., 1993. DNA methylation and mutation. Mutat. Res. 285, 61–67.

Horton, R., Niblett, D., Milne, S., Palmer, S., Tubby, B., Trowsdale, J., Beck, S., 1998. Large-scale sequence comparisons reveal unusually high levels of variation in the HLA-DQB1 locus in the class II region of the human MHC. J. Mol. Biol. 282, 71–97.

Houwen, R.H., Baharloo, S., Blankenship, K., Raeymaekers, P., Juyn, J., Sandkuijl, L.A., Freimer, N.B., 1994. Genome screening by searching for shared segments: mapping a gene for benign recurrent intrahepatic cholestasis. Nature Genet. 8, 380–386.

Howell, W.M., Jobs, M., Gyllensten, U., Brookes, A.J., 1999. Dynamic allele-specific hybridization. A new method for scoring single nucleotide polymorphisms. Nature Biotechnol. 17, 87–88.

Ikuta, S., Takagi, K., Wallace, R.B., Itakura, K., 1987. Dissociation kinetics of 19 base paired oligonucleotide–DNA duplexes containing different single mismatched base pairs. Nucleic Acids Res. 15, 797–811.

Jorde, L.B., Watkins, W.S., Carlson, M., Groden, J., Albertsen, H., Thliveris, A., Leppert, M., 1994. Linkage disequilibrium predicts physical distance in the adenomatous polyposis coli region. Am. J. Human Genet. 54, 884–898.

Kerem, B., Rommens, J.M., Buchanan, J.A., Markiewicz, D., Cox, T.K., Chakravarti, A., Buchwald, M., Tsui, L.C., 1989. Identification of the cystic fibrosis gene: genetic analysis. Science 245, 1073–1080.

Kestila, M., Mannikko, M., Holmberg, C., Gyapay, G., Weissenbach, J., Savolainen, E.R., Peltonen, L., Tryggvason, K., 1994. Congenital nephrotic syndrome of the Finnish type maps to the long arm of chromosome 19. Am. J. Human Genet. 54, 757–764.

Khoury, M.J., 1994. Case-parental control method in the search for disease-susceptibility genes. Am. J. Human Genet. 55, 414–415.

Kidd, K.K., Morar, B., Castiglione, C.M., Zhao, H., Pakstis, A.J., Speed, W.C., Bonne-Tamir, B., Lu, R.-B., Goldman, D., Lee, C., Nam, Y.S., Grandy, D.K., Jenkins, T., Kidd, J.R., 1998. A global survey of haplotype frequencies and linkage disequilibrium at the DRD2 locus. Human Genet. 103, 211–227.

Kondrashov, A.S., 1995. Contamination of the genome by very slightly deleterious mutations: why have we not died 100 times over? J. Theor. Biol. 175, 583–594.

Krawczak, M., Ball, E.V., Cooper, D.N., 1998. Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes. Am. J. Human Genet. 63, 474–488.

Kumar, S., Hedges, B., 1998. A molecular timescale for vertebrate evolution. Nature 392, 917–919.

Laan, M., Pääbo, S., 1997. Demographic history and linkage disequilibrium in human populations. Nature Genet. 17, 435–438.

Laan, M., Pääbo, S., 1998. Mapping genes by drift-generated linkage disequilibrium. Am. J. Human Genet. 63, 654–656.

Lai, E., Riley, J., Purvis, I., Roses, A., 1998. A 4 Mb high-density single nucleotide polymorphism-based map around human APOE. Genomics 54, 31–38.

Landegren, U., Nilsson, M., Kwok, P.Y., 1998. Reading bits of genetic information: methods for single-nucleotide polymorphism analysis. Genome Res. 8, 769–776.

Lehesjoki, A.E., Koskiniemi, M., Norio, R., Tirrito, S., Sistonen, P., Lander, E., de la Chapelle, A., 1993. Localization of the EPM1 gene for progressive myoclonus epilepsy on chromosome 21: linkage disequilibrium allows high resolution mapping. Human Mol. Genet. 2, 1229–1234.

Levy-Lahad, E., Wasco, W., Poorkaj, P., Romano, D.M., Oshima, J., Pettingell, W.H., Yu, C.E., Jondro, P.D., Schmidt, S.D., Wang, K., et al., 1995. Candidate gene for the chromosome 1 familial Alzheimer's disease locus. Science 269, 973–977.

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

Li, W.H., Ellsworth, D.L., Krushkal, J., Chang, B.H., Hewett-Emmett, D., 1996. Rates of nucleotide substitution in primates and rodents and the generation-time effect hypothesis. Mol. Phylogenet. Evol. 5, 182–187.

Liu, L., Forsell, C., Lilius, L., Axelman, K., Corder, E.H., Lannfelt, L., 1996. Allelic association but only weak evidence for linkage to the apolipoprotein E locus in late-onset Swedish Alzheimer families. Am. J. Med. Genet. 67, 306–311.

Lizardi, P.M., Huang, X., Zhu, Z., Bray-Ward, P., Thomas, D.C., Ward, D.C., 1998. Mutation detection and single-molecule counting using isothermal rolling-circle amplification. Nature Genet. 19, 225–232.

Lyamichev, V., Mast, A.L., Hall, J.G., Prudent, J.R., Kaiser, M.W., Takova, T., Kwiatkowski, R.W., Sander, T.J., de Arruda, M., Arco, D.A., Neri, B.P., Brow, M.A.D., 1999. Polymorphism identification and quantitative detection of genomic DNA by invasive cleavage of oligonucleotide probes. Nature Biotechnol. 17, 292–296.

Marshall, E., 1997. 'Playing chicken' over gene markers. Science 278, 2046–2048.

Marshall, E., 1999. Drug firms to create public database of genetic mutations. Science 284, 406–407.

Mountain, J.L., Lin, A.A., Bowcock, A.M., Cavalli-Sforza, L.L., 1992. Evolution of modern humans: evidence from nuclear DNA polymorphisms. Philos. Trans. Roy. Soc. London B: Biol. Sci. 337, 159–165.

Nachman, M.W., Bauer, V.L., Crowell, S.L., Aquadro, C.F., 1998. DNA variability and recombination rates at X-linked loci in humans. Genetics 150, 1133–1141.

Nei, M., Li, W.H., 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. USA 76, 5269–5273.

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

Nilsson, M., Krejci, K., Koch, J., Kwiatkowski, M., Gustavsson, P., Landegren, U., 1997. Padlock probes reveal single-nucleotide differences, parent of origin and in situ distribution of centromeric sequences in human chromosomes 13 and 21. Nature Genet. 16, 252–255.

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

Puffenberger, E.G., Hosoda, K., Washington, S.S., Nakao, K., de Wit, D., Yanagisawa, M., Chakravart, A., 1994. A missense mutation of the endothelin-B receptor gene in multigenic Hirschsprung's disease. Cell 79, 1257–1266.

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

Rogaev, E.I., Sherrington, R., Rogaeva, E.A., Levesque, G., Ikeda, M., Liang, Y., Chi, H., Lin, C., Holman, K., Tsuda, T., et al., 1995. Familial Alzheimer's disease in kindreds with missense mutations in a gene on chromosome 1 related to the Alzheimer's disease type 3 gene. Nature 376, 775–778.

Rommens, J.M., Iannuzzi, M.C., Kerem, B., Drumm, M.L., Melmer, G., Dean, M., Rozmahel, R., Cole, J.L., Kennedy, D., Hidaka, N., et al., 1989. Identification of the cystic fibrosis gene: chromosome walking and jumping. Science 245, 1059–1065.

Sarkar, C., Ortigao, F.R., Gyllensten, U., Brookes, A.J., 1998. Human genetic bi-allelic sequences (HGBASE), a database of intra-genic polymorphisms. Mem. Inst. Oswaldo Cruz 93, 693–694.

Sherrington, R., Rogaev, E.I., Liang, Y., Rogaeva, E.A., Levesque, G., Ikeda, M., Chi, H., Lin, C., Li, G., Holman, K., et al., 1995. Cloning of a gene bearing missense mutations in early-onset familial Alzheimer's disease. Nature 375, 754–760.

Sing, C.F., Haviland, M.B., Zerba, K.E., Templeton, A.R, 1992. Application of cladistics to the analysis of genotype–phenotype relationships. Eur. J. Epidemiol. 8, 3–9.

Slatkin, M., 1994. Linkage disequilibrium in stable and growing populations. Genetics 137, 331–336.

Spielman, R.S., McGinnis, R.E., Ewens, W.J., 1993. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am. J. Human Genet. 52, 506–516.

Stevenson, J., 1992. Evidence for a genetic etiology in hyperactivity in children. Behav. Genet. 22, 337–344.

Stoneking, M., 1997. Recent African origin of human mitochondrial DNA: review of the current status of the hypothesis. In: Donnely, P., Tavare, S. (Eds.), Progress in Population Genetics and Human Evolution. Springer, New York, pp. 1–13.

Strittmatter, W.J., Roses, A.D., 1995. Apolipoprotein E and Alzheimer Disease. Proc. Natl. Acad. Sci. USA 92, 4725–4727.

Sulisalo, T., Klockars, J., Makitie, O., Francomano, C.A., de la Chapelle, A., Kaitila, I., Sistonen, P., 1994. High-resolution linkage-disequilibrium mapping of the cartilage-hair hypoplasia gene. Am. J. Human Genet. 55, 937–945.

Taillon-Miller, P., Gu, Z., Li, Q., Hillier, L., Kwok, P.Y., 1998. Overlapping genomic sequences: A treasure trove of single-nucleotide polymorphisms. Genome Res. 8, 748–754.

Takahata, N., 1995. A genetic perspective on the origin and history of humans. Annu. Rev. Ecol. Syst. 26, 342–372.

Templeton, A.R., 1996. Cladistic approaches to identifying determinants of variability in multifactorial phenotypes and the evolutionary significance of variation in the human genome. Ciba Found. Symp. 197, 259–277.

Terwilliger, J.D., Weiss, K.M, 1998. Linkage disequilibrium mapping of complex disease: fantasy or reality? Curr. Opin. Biotechnol. 9, 578–594.

Terwilliger, J.D., Zollner, S., Laan, M., Pääbo, S., 1998. Mapping genes through the use of linkage disequilibrium generated by genetic drift: 'Drift mapping' in small populations with no demographic expansion. Human Hered. 48, 138–154.

Tishkoff, S.A., Dietzsch, E., Speed, W., Pakstis, A.J., Kidd, J.R., Cheung, K., Bonne-Tamir, B., Santachiara-Benerecetti, A.S., Moral, P., Krings, M., 1996. Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. Science 271, 1380–1387.

Tishkoff, S.A., Goldman, A., Calafell, F., Speed, W.C., Deinard, A.S., Bonne-Tamir, B., Kidd, J.R., Pakstis, A.J., Jenkins, T., Kidd, K.K., 1998. A global haplotype analysis of the myotonic dystrophy locus: implications for the evolution of modern humans and for the origin of myotonic dystrophy mutations. Am. J. Human Genet. 62, 1389–1402.

Venter, J.C., Adams, M.D., Sutton, G.G., Kerlavage, A.R., Smith, H.O., Hunkapiller, M., 1998. Shotgun sequencing of the human genome. Science 280, 1540–1542.

Vial, W.C., 1986. Cigarette smoking and lung disease. Am. J. Med. Sci. 291, 130–142.

Wang, D.G., Fan, J.B., Siao, C.J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J., Kruglyak, L., Stein, L., Hsie, L., Topaloglou, T., Hubbell, 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., Lander, E.S., 1998. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science 280, 1077–1082.