

Organisation of the human genome and our tools for identifying disease genes

P.E. Slagboom*, I. Meulenbelt

Section Molecular Epidemiology, Sylvius Laboratory, Leiden University Medical Centre, PO Box 9503, 2300 RA Leiden, The Netherlands

Abstract

Determination of the sequence of the human genome has been a major undertaking. It provided powerful tools to explore the genetic component in complex diseases. To fully understand the genetic pathways contributing to complex disease traits, we must not only reveal the genomic locus of all genes involved, but also delineate the functionally relevant allelic variation in such genes and understand the patterns of gene expression leading up to the actual disease trait. Insight in the genetic contribution to clinical endpoints of complex disease and their biological risk factors, therefore, requires an understanding of both the structure and the biology of the genome. This paper constitutes a tutorial overview of the organisation of the human genome, the tools it provides for molecular genetic studies, and the genomic background of the current strategies for gene identification. © 2002 Published by Elsevier Science B.V.

Keywords: DNA; Genetic variation; Gene expression; Genetic markers; Linkage; Positional cloning

1. Introduction

This chapter attempts to provide a background for an understanding of molecular genetic studies aimed at the dissection of the genetic component in human disease. Without trying to be complete we will discuss what kind of landscape the genome (all genetic information in an organism) presents to the geneticist that tries to find genes with a major contribution to disease. What do genes look like? What DNA sequences inside genes or surrounding genes may harbour mutations that cause

* Corresponding author. Tel.: +31-71-527-1945; fax: +31-71-527-1985
E-mail address: p.slagboom@lumc.nl (P.E. Slagboom).

disease? What genetic markers in the genome are tools to localise disease genes? The landscape of the genome is just as much the result of evolutionary history as the Grand Canyon. The human genome is subject to continuous variation, generated in the DNA sequence organisation of both our somatic cells and our germ cells (sperm and egg cells and their progenitor cells). Sequence variation in germ cells is transmitted to offspring. Germ line variations that add benefit to an individual's reproductive capacities may become more widely distributed or even become fixed in the population during evolution; other genetic variations disappear from the population within only a few generations. Neutral variations may go either way depending on chance (genetic drift).

During evolution the genome has expanded. A reflection of this evolutionary process is observed throughout the genome. The most simple organisms, single cells without a nucleus (prokaryotes) such as bacteria, have a small circular genome consisting mainly of genes. Other single cell organisms with a nucleus (eukaryotes), such as yeast, have a larger genome in which 20% consists of genes (20 000), and multicellular eukaryotes, such as humans, have a genome 200 times larger than the yeast genome consisting of 2% genes ($\pm 30\,000$). Only a small part of the additional DNA sequences created during evolution actually contains DNA coding for gene products (polypeptides). Many genes in higher organisms resemble those in lower organisms and the difference in gene number does not seem large enough to explain the huge differences in complexity. A biological explanation for the increased organismal complexity generated during evolution may not so much be in the increased number of genes, but in the fine tuned regulation of gene expression: how and when do genes become expressed. Genome studies into the biology of complex traits, therefore, require an integrated understanding of the sequence variations in human genes and their corresponding genes in (experimental) animals as well as the factors that influence their expression and the functional pathways to which they contribute (see [Lewin, 1997](#); [Strachan and Read, 1999](#) for excellent textbooks on these topics).

2. From DNA to polypeptide, the central dogma

Deoxyribonucleic acid (DNA) is in most organisms the primary material containing the information for all hereditary traits of the organism. DNA molecules are polymers consisting of two strands each made up of sequences of four types of nitrogenous bases (Guanine, Adenosine, Thymine, Cytosine) arranged like beads on a long string. The two strands are intertwined to form a double stranded helix that closely resembles a spiral staircase. The steps of this staircase are formed by the bases that pair to each other in a fixed order (cytosine(C) opposite to guanine (G) and adenosine (A) opposite to thymine (T)). Large stretches of the double stranded DNA helix, together with binding proteins, form a chromosome. The linear sequence of the four bases contains the code for proteins, the functional endpoints of the DNA template. This is comparable with the way a linear sequence of characters of the alphabet contains the words in written language. The genetic information of the cell

is processed from DNA through another polymer, ribonucleic acid (RNA), into polypeptides that form the basic components of all proteins (Fig. 1).

In most somatic cells of the body the human genome contains 3000 million basepairs arranged in 22 pairs of chromosomes (autosomes) and a pair of X chromosomes or X and Y chromosome (sex chromosomes). For all chromosome numbers (1–22) and the sex chromosomes, offspring obtains a chromosome from each parent resulting in 22 pairs of autosomes and a pair of sex chromosomes. This chromosomal DNA is permanently situated in the nucleus of the cell. The number and shape of the chromosomes differs between different eukaryote species. Except for the nucleus, DNA also resides in the mitochondrion as circular molecules. In contrast to somatic cells, human gametes (sperm and egg cells) have only one copy of each of the 22 autosomes and one sex chromosome. Somatic cells are diploid because they contain two copies of the genome ($2n$), gametes are haploid (n). The cell division process of somatic cells is called mitosis. The genome of the somatic cell during mitosis is first replicated ($4n$) and subsequently distributed over two daughter cells ($2n$). Gametes are formed from diploid precursor (germ) cells in another cell division process that is called meiosis. Meiosis of one germ cell involves a cell division round generating two haploid daughter cells (gametes) followed by one round of DNA replication and a second cell division round generating 4 gametes. During meiosis, homologous chromosomes, that is the maternal and paternal chromosomes of chromosome number 1, 2, 3, . . . , etc. take position adjacent to each other. Before the chromosome pairs are subsequently separated to undergo reduction to the haploid state, genetic information becomes exchanged between the maternal and paternal chromosome (Fig. 2). This exchange process is called recombination. Meiosis results in the production of gametes (haploid: one copy of the complete genome) that contain for each chromosome either the original maternal chromosome, the original paternal chromosome, or new (recombinant) chromosomes as a result of recombination. During each meiosis, recombination occurs at least twice per chromosome. Gene localisation (mapping) techniques depend on recombinant children.

RNA is the more mobile polymer that is able to transfer the genetic information content of DNA from the nucleus to the cytoplasm of the cell (Fig. 1). DNA and RNA are slightly different molecules. The four types of bases in double stranded DNA are A, C, G and T. RNA is a single stranded molecule consisting of A, C, G and uracyl (U), meaning that only three out of four bases are the same as in DNA. The process in which sequence information of DNA is copied into RNA is called transcription. DNA strands have an orientation, one end being marked by a chemical group and called the 5' end and the other being called the 3' end. The two strands in a DNA duplex are positioned in opposite orientation. During transcription, the DNA strands separate at a site determined by specific a DNA sequence to which transcription proteins bind (promoter area, see below). Subsequently RNA is copied, starting again at a specific sequence (INR, see below), from the 3' to 5' DNA template strand (RNA has, therefore, a 5'–3' orientation). After the RNA molecule has been transcribed from one strand of a gene it is processed in a number of steps (see below), and travels from the nucleus to the cytoplasm. In the cytoplasm the information enclosed in RNA polymers is translated into polypeptides. Analogous

to DNA and RNA, the polypeptides are polymers of a linear sequence of units, the amino acids of which 20 different types exist. Three bases in the DNA sequence are copied into RNA sequence to form a codon that is translated into one amino acid. During translation, amino acids are tied together like beads on a string to form a polypeptide that is subjected to posttranslational modifications (i.e. cleavage of peptide fragments, cross linking of polypeptides). The unit of genetic information that is expressed into a processed polypeptide is defined as a gene. Various polypeptides, often originating from different genes, can merge to fold into a single protein. Proteins, often in complex interaction with other proteins, enzymatically regulate cellular functions, constitute the cellular structural components such as cell membranes, function as storage proteins, or act as receptors for signal transduction, hormones and transcription factors.

3. What does a human gene look like: coding and non-coding DNA

The genes of vertebrate species contain pieces of sequence information that become translated into chains of amino acids (exons, the coding sequence) interrupted by pieces of non-coding sequence information (introns). Upstream (5') of the first exon is an untranslated region (5' UTR) and downstream (3') of the last exon is an untranslated region (3' UTR). When DNA is transcribed into the primary RNA transcript in the nucleus, the full sequence information (exons and introns) is copied. This primary RNA molecule is subsequently processed by enzymes that remove the intronic sequences and fuse the exons (Fig. 3). This process is called RNA splicing and depends on specific sequences (GT and AG) at the exon/intron boundaries (splice junctions) in the gene. A string of fused exon sequences is generated that forms the coding template for translation into a polypeptide. Additional to the splicing process, the 5' end of the RNA polymer is blocked by a chemical group (CAP at the 5' end) and a number of adenines (poly(A) tail) to the other (3') end. Now a messenger RNA (mRNA) molecule has been created that travels to the cytoplasm and becomes translated into a polypeptide. The CAP site and poly(A) signal, among others, assist in binding and stabilising the mRNA to the cytoplasmic particle at which translation occurs, the ribosome. The polypeptide in

Fig. 1. Schematic representation of the central dogma. A magnified gene sequence on one of the 46 chromosomes in the nucleus at the bottom, (1) is transcribed, starting at a start site at the left end of the gene sequence at the 3: into a single stranded RNA molecule (2). The information content in RNA is then translated in the cytoplasm into a polypeptide (3) where each three bases in RNA (codon) form the information for one amino acid to be sequentially bound in the polypeptide. Subsequently, the polypeptide folds into a protein (4) with other polypeptides.

Fig. 2. Recombination during meiosis. (A) During specific phases in cell replication, a chromosome consists of two sister chromatids (two copies of the chromosome). A single cross over (grey cross) occurs between homologous maternal and paternal chromatids generating two recombinant and two non-recombinant chromatids. (B) The gametes produced as a result of recombination carry (for chromosomes 1, 2, ..., 22, X, Y) either the maternal, paternal or a recombinant chromosome.

turn will also be processed (post-translational modification) to form, together with other polypeptides, the functional endpoint of this process: the protein.

As stated earlier, as a result of the variation produced during evolution the size of the genome (and our genes) has expanded compared with more simple organisms. Only 2% of the DNA sequence information in the human genome is translated into polypeptide sequence (Venter et al., 2001; Lander et al., 2001). The human genome contains approximately 30 000 genes and the average amount of coding sequence in a gene is 2000 base pairs. The size of human genes varies from hundreds of bases to several megabases especially due to the large intronic sequences. The human Dystrophin gene is 2, 4 Mb (2 400 000 bases) including 79 exons. The non-coding part of the genome (98%) is not 'junk' DNA. It contains numerous signals that are necessary for the regulation of gene expression. All somatic cells in an organism carry the same set of genes. Yet different cell types or cells in different developmental and differentiation stages show large differences in cell function and responses to internal and external stimuli. This is entirely brought about by different patterns of

Fig. 3. Gene transcription in vertebrate species. At the bottom a gene is depicted consisting of exonic and intronic sequences (1). The region in front of the first exon contains sequences that regulate gene transcription (symbolised by the signpost, see Fig. 4). The gene is transcribed into a primary RNA transcript (2) and this RNA molecule is further processed by splicing, capping, poly A tailing into an mRNA molecule (3) that will enter the cytoplasm for translation.

Fig. 4. Elements involved in vertebrate gene transcription. Transcription starts at INR, GT/AG are the recognition sequences for splicing (splice sites), AAUAA the recognition site for polyadenylation (poly A signal). TATA and BRE boxes belong to the core promoter, GC and CCAAT boxes are enhancers in the non core promoter area. More distal elements are RE (responsive to hormones for example) and insulators. Enhancers and silencers may be present kilobases from the promoter or very close to the gene (in intronic sequences, for example).

Fig. 5. Gene localisation by linkage analysis. Depicted are genetic markers on a chromosome pair of mother, father and two of their children. Suppose three loci (1, 2, 3) are linked (close together) on this chromosome. Locus 1 carries a mutation or variant in the mother (dot at Locus 1) associated with a phenotype (a disease, a high trait score, an increased level of a biochemical parameter, etc.). Locus 2 and 3 are polymorphic loci. Locus 1 is an unknown locus (that needs to be localised) but its resulting phenotype can be established in all individuals. The mother and child 1 have the deviant phenotype ('affected'), the father and child 2 do not have this phenotype. At locus 2 the mother is homozygous. This polymorphic locus is not informative which is demonstrated when typing locus 2 in the children (both have allele A from the mother). Since the affected child and the non-affected child both have allele A, one could not tell whether the known genetic marker (locus 2) is close to the gene causing the deviant phenotype, or not. By typing locus 3, for which the mother is heterozygous (informative), it can be detected that child 1 inherited one chromosome (allele A) and the other child inherited the other chromosome (allele a) from the mother. Throughout the whole family, one can follow locus 3 and one can observe that when a child inherits allele A at locus 3, this child carries also the deviant phenotype ('affected status'), and when allele a at locus 3 is transmitted, the child does not express the deviant phenotype. If the association of disease to allele A at locus 3 in the pedigree is statistically significant, it means that locus 3 (of which the position in the genome is known) is likely to be genetically linked (physically close) to the gene one tried to localise. Note that the exact genotype at the marker locus 3 is irrelevant, in another pedigree it could well be that the disease is linked to a paternal allele a at locus 3. For a marker locus it only matters that it can be shown to co-segregate with the disease in a pedigree. In genetic studies the terms IBS and IBD are used. Although both children carried allele A at locus 2 they are IBS, genotyping locus 3 shows they are not IBD for this chromosome (they did not inherit the same A allele at locus 2 from their mother).

expression of the same set of genes. Only some genes, for example those involved in protein synthesis machinery, are expressed in almost all cells (housekeeping genes).

Signals involved in the regulation of gene expression include specific DNA sequences that form recognition sites (*cis*-acting elements) for proteins (*trans*-acting factors) to bind to DNA, and by binding influence gene expression (Fig. 4). Transcription of DNA into RNA starts after binding of the basal transcription apparatus (RNA polymerase complex and general transcription factors) at the transcription start site (Initiator sequence, INR). The promoter area (about 200 bp at the 5' side of the INR, which is called: –200 'upstream' of INR) contains the main signals mediating the binding of the transcription apparatus and the separation of the two DNA strands to start transcription. *Cis*-acting elements (present on the same DNA molecule as the gene they regulate) of the core promoter are the TATAbox (nucleotide –25) and BRE box. In the non core promoter area (–50 to –200) GC rich sequences are present (GC boxes) and the CCAAT box. These *cis*-acting elements are called enhancers that modulate basal transcription of the core promoter. More distal promoter elements may contain other enhancers, silencers (elements involved in reducing transcription levels) and insulators (elements involved in blocking large DNA segments against all transcription). The distal promoter sequences may harbour (within approximately 1 kb from the initiator sequence) responsive elements (RE) modulating transcription of the gene in response to glucocorticoids, steroids and the second messenger cAMP. Enhancers and silencers are short sequence elements that may be located tens of kilobases upstream from the transcription start site or may be present in intronic sequences within the gene. These elements can afford to be positioned at a distance from the promoter because the DNA strand between such enhancer signals and the promoter signal usually loops out, bringing element and promoter together.

Gene expression is, in addition to the presence of above described DNA sequences, also influenced by DNA methylation of C bases (cytosines). 3% of cytosines in human DNA is methylated mostly in CpG dinucleotide sequences. Methylation of the cytosine at a specific CpG dinucleotide may repress gene expression in some tissues, whereas, the absence of methylation of such a site in other tissues correspond with the expression of the gene. Unmethylated CpG islands (Bird, 1986) occur in the promoter regions of all housekeeping genes and about 40% of tissue specific genes. DNA methylation is believed to be involved in imprinting, a mechanism that leads to the specific transcription of one of the two alleles of a gene locus: the paternally or maternally inherited allele. For a number of genes it has been reported that this imprinting leads to the selective transcription of one of the alleles depending on the parental origin of the chromosome. In diseases such as Beckwith–Wiedeman syndrome the disease gene is only expressed when inherited from the mother, the mutated allele inherited from the father is imprinted and transcriptionally silent.

In summary, regulation of gene expression may act on the level of transcription and processing: presence of *trans*-acting factors binding to *Cis*-acting elements, hormone or transcription factors binding to RE in promoter areas, the use of alternative promoters, alternative splicing, alternative polyadenylation and RNA

editing (introduction of RNA sequence changes). In addition, regulation of gene expression occurs at the level of translation: posttranslational cleavage, polypeptide stability etc. (generating on average three different functional proteins per human gene, Banks et al., 2000) and by epigenetic control (DNA methylation, chromatin structure). It is not yet clear which of all these modifications contribute most to organismal complexity. Similar levels of alternative splicing were observed in different species making it unlikely that this mechanism is a major source of the increased complexity (Brett et al., 2002).

Given the importance of regulated gene expression, it is not difficult to imagine why, in biological psychology, gene-environment and gene-behaviour interactions are thought to play an important role. Environment (stress) and behaviour (smoking, drinking, exercise) influence serum hormone levels, the balance of endo/paracrine cell signals, and/or the extracellular micro-environment. Serum hormones can influence the transcription of sets of genes by signal transduction in a multistep process. Signal transduction may start by binding of hormones to cell surface receptors eventually leading to induction of expression of sets of genes by binding of *trans*-acting factors to RE surrounding such inducible genes in the nucleus. In normal cell signalling, endo- or paracrine signals can lead to the temporary induction of gene expression. Finally, extracellular concentrations of ions, oxygen radicals, nutrient molecules, temperature and shock can directly affect the expression of inducible genes. By influencing gene expression, all these ‘environmental’ effects will lay bare or amplify the effect of existing genetic DNA sequence variation, and induce disease in carriers of disease gene variants.

4. Repeated DNA sequences

Apart from the DNA sequences with a regulatory function or a coding function, 40% of the non-coding part of the human genome contains structures of repeated DNA sequences. These may be sequence motives numerously repeated in a dispersed way throughout the genome. The most expanded example is the Alu repeat, a 300 bp long element of which the human genome carries a million copies (10% of the genome) (Lander et al., 2001). A part of these copies in the human genome have gained a function in the regulation of gene expression, these repeats have a relatively high GC content and are preferentially located at transcriptionally active regions in the genome (Deininger and Batzer, 1999). Other sequence motives are repeated adjacent to each other in head to tail (tandem) or inverse orientation. An example of tandemly arranged DNA sequences are microsatellites, arrays of repeat units of 1–4 basepairs that form relatively small size motifs ((CAG)_n, or (CA)_n). These tandem repeat arrays may expand or contract in size at a given location in somatic cells and gametes (increase or decrease their copynumber *n*). In general, the presence of tandem repeat arrays at a chromosomal location increases the chance for genetic variation to occur.

Tandem arrays of DNA sequence are prone to deletion/insertion of repeat units. Such tandem repeat arrays expand or contract as a result of homologous

recombination. When homologous chromosomes pair during meiosis, genetic information can be exchanged between the chromosome pairs. The homology between tandem repeat arrays on pairing chromosomes may lead to erroneous pairing and after recombination, expansion of the tandem repeat array on one of the chromosomes and contraction on the other is the result. Alternatively the length of tandem repeat arrays at a locus may be affected during DNA replication by a mechanism called slippage replication. Expansion in some tandem repeat loci may disturb normal gene functioning and be a cause of disease. Most of the repetitive DNA sequence motifs, however, are neutrally present in non-coding DNA and create an endless source of neutral genetic variation between individuals. As outlined below, by their variable length, tandem repeat arrays are important tools to localise genes.

5. Genetic variation and disease

The study of families carrying heritable diseases (McKusick, 1997) has not only provided immense insights into the function of many genes but has illustrated what type of genetic variation occurs in the genome giving rise to disease. In fact, the complete spectrum of genetic variation that is observed in disease is also observed during evolution. Large chromosomal aberrations were the earliest mutations detected because many of these can be observed by cytogenetic analysis of chromosomes under the microscope. Molecular genetic analysis has since revealed smaller sources of genetic variation (Cooper and Krawczak, 1993) such as single base substitutions, the most frequent of all variations. Deletions or insertions of one or small numbers of basepairs or of larger DNA segments occur, be it more frequent in non-coding than in coding DNA. Both loss and gain of function mutations occur in genes. Changed gene functioning can be the result of mutations occurring in coding DNA, in 3' or 5' UTRs, at splice junctions (sequences that direct the splicing of RNA), in proximal or more distal promoter sequences, enhancers, silencers, insulators, and CpG sites. For mutations causing disease, the severity of the clinical manifestations (age of onset, systemic or local disease, extreme level) frequently depends on the type of mutations that occur at a gene locus (variable expression). In coding DNA, mutations in the first or second base of a codon have usually more consequences than mutations in the third base. Mutations may cause a codon change into a stopcodon (abolishing translation), or a frameshift (disturbing the complete polypeptide sequence following the codon change). The effect of codon changes due to mutations depend further on whether a functional domain of the protein becomes mutated, for example a hormone binding domain, a DNA binding domain, a domain necessary for the cleavage of a protein etc.

Methylated cytosines deaminate spontaneously to produce thymine. Therefore, CpG sites of which the C is methylated present hot spots prone to mutate into TpG. Logically the vertebrate genome has an underrepresentation of CpG dinucleotides (since most of these will have mutated to TpG). Yet CpG islands occur in the promoter region of many genes. It is an interesting finding that a part of DNA

sequences that in evolution have gained functionality in the regulation of gene expression simultaneously form a threat because they represent hot spots of mutation. Examples are CpG dinucleotides and structures of repetitive sequences (tandemly arranged, inverted or dispersed) at the 5' or 3' gene areas or in introns. Quite a number of diseases result from the presence of repetitive DNA. In neurofibromatosis type 1 (NF1) an Alu sequence (interspersed repeat) was inserted in intron 5, in some breastcancer families an Alu is inserted in an exon (number 22) of the BRCA2 gene. In both cases this resulted in the formation of a shortened polypeptide by exon skipping.

Especially a number of neurological disorders (Fragile-X syndrome, myotonic dystrophy, Huntingtons disease) originate in some families from the expansion (above a threshold copynumber) of trinucleotide repeats. For example, in spinocerebellar ataxia a (CAG)₃₀ array can become a deleterious (CAG)₁₀₀ array (Koob et al., 1999). Diseases caused by triplet repeat expansion in 5' UTR or 3' UTR or intronic gene sequences, show a lower age of onset and increased severity in successive generations (anticipation). The mechanism by which repeat arrays expand in disease may also apply to the numerous gene families consisting of partly homologous copies of genes.

When studying the genetics of disease, one may distinguish mild (relatively symptomless) forms of the disease from severe clinical manifestations. Severe disease is more likely due to mutations affecting coding DNA. Milder disease forms may be expected to be due to sequence changes at *cis*-elements, for example, which may alter the regulation of gene expression. Whether carriers of a mild gene variant express the disease or trait (penetrance) may depend on environmental factors. The phenotypic effect of a mutation in a RE of a gene might become apparent only under environmentally induced circumstances where a *trans*-acting factor (hormone) is expressed that cannot bind optimally to the mutated RE.

6. Genetic variation as a tool to identify disease genes

Most of DNA sequence variations are generated during DNA replication and DNA repair. DNA sequence changes with a functional effect on gene function are designated as mutations, variations with no effect (or very small effects) are usually designated as polymorphisms. The different sequence variants at a locus are called alleles. Since a copy of each gene is present on each of the homologous chromosomes (one paternal, one maternal), each human individual has two alleles for every locus (gene, polymorphism, or any other DNA sequence). Suppose that subject 1 has at nucleotide 20 of a gene an A (AT base-pair) on the maternal chromosome and a C (CG base-pair) on the paternal chromosome. Subject 2 may have an A on both maternal and paternal chromosomes. Subject 1 is then called to be heterozygous for the polymorphism at nucleotide 20 (carries alleles A and C; genotype AC), whereas, subject 2 is called to be homozygous for this polymorphism (carries allele A on both paternal and maternal chromosomes; genotype AA). When the heterozygous subject 1 transmits the A allele to two of his children, the children are identical by descent

(IBD) for the A allele (both children inherited the same A allele from a single parent, subject 1). For the children of a homozygous parent, like subject 2, this is slightly more complicated. If two children of this subject inherit the same A allele they are IBD at this locus, but if one child inherits one A allele (the grand-paternal A allele) and the other child inherits the second A allele (the grand-maternal A allele) from subject 2, they are not IBD. Since both children have an A allele in the genotype, they are called to be identical by state (IBS) for this locus. In linkage analysis (explained in Fig. 5) variation in the paternal and maternal contribution to the various genomes of the offspring is used to localise disease genes or, in general, genes that influence a quantitative trait.

On average one in 250–1000 bases in the nuclear human genome persist in the population as different sequence alleles. The most frequent of these are single base differences between alleles (Single Nucleotide Polymorphism's, SNP) (Wang et al., 1998). Again, except for specific genetic disorders, each individual in a population inherits a maternal and paternal allele at all chromosomal loci. In the population as a whole, many alleles of a locus may persist. This number is especially high for repeat loci. Many microsatellite repeat loci persist in 20 different length alleles in the population (for example (CAG)₁₀, (CAG)₁₄, (CAG)₂₀, etc.). If many of these alleles occur at reasonable frequencies in the population, most individuals will be heterozygous at such a locus. This means that the maternal and paternal chromosomes of such heterozygous individuals can be distinguished by this polymorphic locus. The maternal chromosome may have the (CAG)₁₀ allele at that repeat locus and the paternal chromosome may carry the (CAG)₂₀ allele. In many genetic approaches to disease, one aims to follow the transmittance of different (maternal and paternal) chromosomes in pedigrees to search for the chromosome that is repeatedly co-transmitted with disease traits. It is, thus, essential that paternal and maternal chromosomes can be distinguished in each meiosis. Polymorphic loci

Fig. 6. Searching informative recombinants. The mosaic revealed by a number of recombinants can show which minimal region in a chromosome is sufficient to generate the trait. Suppose the gene causing the trait to be positive is present on the paternal chromosome (disease mutation depicted by a dot). The homologous wild type gene (causing the trait to be negative) is present on the maternal chromosome. Recombinants 1, 2, 3 would show that the gene must be present on the left side of the chromosome the minimal chromosome region being that of the trait positive recombinant 1. Only when recombinant 4, positive for the trait, is investigated by genetic markers, the gene can be further localised to the area between the arrows.

Fig. 7. Association of a neutral SNP with a causal disease gene variant. (1) A SNP polymorphism is generated in the DNA sequence flanking a gene (G to T substitution). Two alleles (and two haplotypes: A–G and A–T) now reside in the population. (2) If a causal disease mutation in the gene (represented by a dot) occurs on the chromosome with the T allele in the first affected person (the disease 'founder') shortly after the T allele was generated, significant association can be detected many generations later if mutation and T allele become transmitted together. The T allele will be over represented in affected (cases) over unaffected (control) subjects. (3) If, on the other hand, the T allele becomes widely distributed before a causal disease mutation occurs in the gene, no association will be detected in the population many generations later. The frequency of T alleles among unaffected subjects will simply be too high.

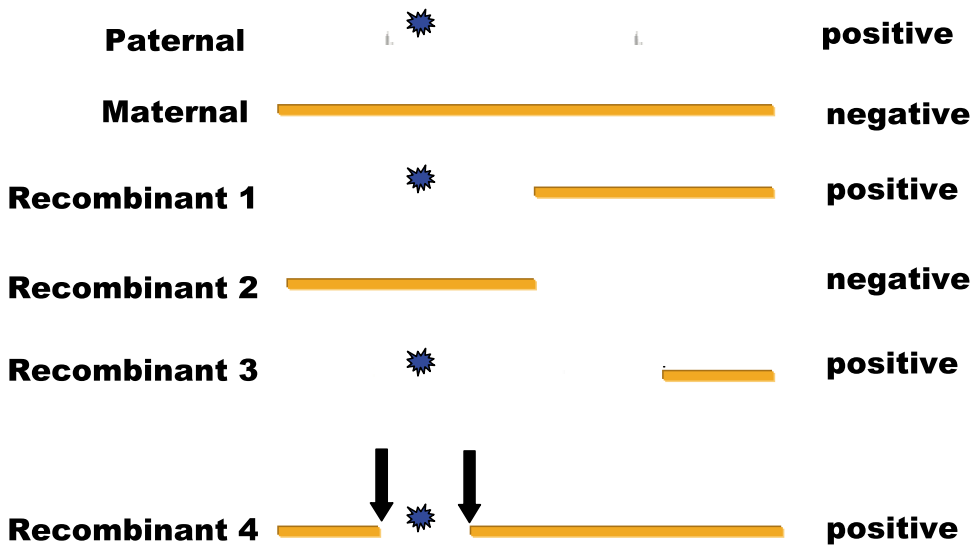


Fig. 6

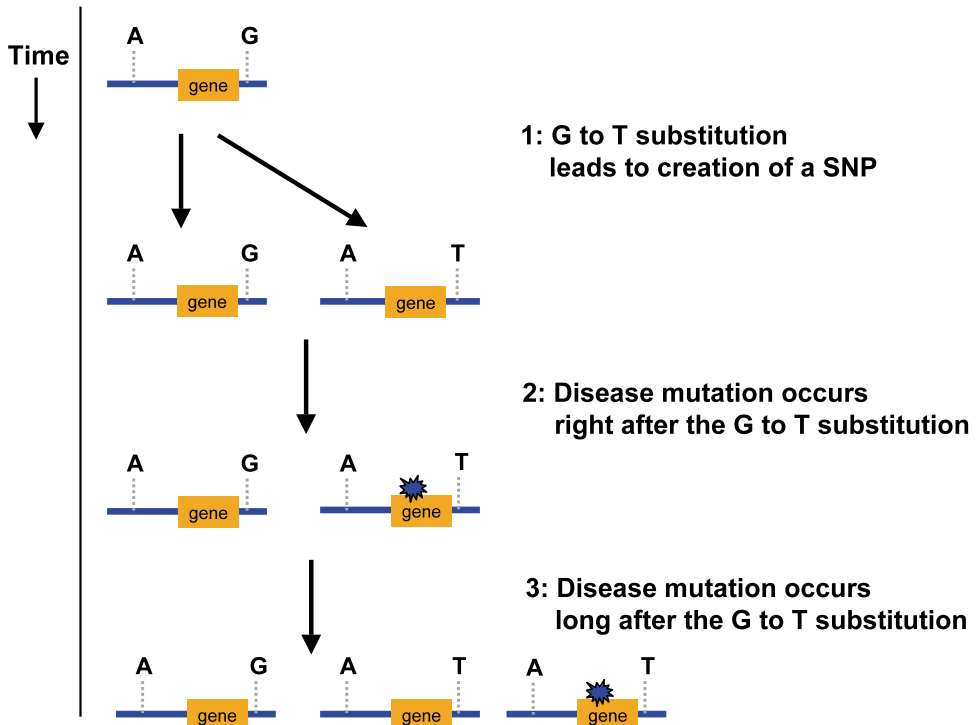


Fig. 7

that have so many different alleles in the population that most individuals are heterozygous are very useful tools as genetic markers in genetic studies.

During the past 10 years numerous laboratory methods have been developed to detect polymorphisms. To be useful as genetic markers, the rarest variant of the polymorphisms must occur with a frequency greater than 0.01 in the human population. Modern methodologies depend on specific amplification of known polymorphisms in the genomic DNA of patients, control subjects, family members etc., to be used as markers to detect unknown disease polymorphisms. The DNA sequence information of the human genome that has been produced as a result of the Human Genome Project (HGP) enables the amplification of almost any locus in the genome with high specificity. After amplification (Polymerase Chain Reaction, PCR) of such a locus, typically a technical step follows (electrophoresis, sequencing or hybridisation step) to visualise the allelic variants at the polymorphic marker locus within the amplified fragment. Polymorphic marker loci can be found in public databases at least at every 1000 basepairs across all chromosomes. One can scan a chromosome by amplifying tens of DNA fragments across the chromosome and measure polymorphic marker loci in such fragments in order to distinguish the paternal from maternal chromosomes and to detect at which chromosomal regions recombination occurred between paternal and maternal chromosomes.

7. Genetic marker maps

The mapping of genes depends on the use of a genetic marker map, a set of polymorphic DNA markers with a known relative position in the genome. If two marker loci are positioned on the same chromosome not too far from each other, alleles of these marker loci will be transmitted together from one generation to the next. If the marker loci are far apart, recombination between the marker loci will prevent alleles from being transmitted together. Starting with the known position of one marker locus, it can be tested whether another marker is genetically close (linked) by counting the number of recombinations that occurred between the known and unknown marker loci in a given number of meioses. If the marker loci are linked (because physically close), alleles from both loci segregate together in a pedigree. By studying the segregation of microsatellite repeat loci in families for example, the genetic distance of thousands of such loci could be calculated.

As an example, let a subject be heterozygous for locus 1 (genotype Aa) and locus 2 (genotype Bb) on a specific chromosome. Suppose allele A and B are present together on the maternal chromosome as haplotype AB (a haplotype is a set of alleles at different loci on the same chromosome) and the two other alleles form haplotype ab on the paternal chromosome. If recombination occurs once between locus 1 and 2 during meiosis, four types of gametes will be generated: haplotypes corresponding to the two original chromosomes AB, ab and two new haplotypes Ab and aB. Children of this subject carrying haplotypes AB and ab are non-recombinant and children carrying haplotypes Ab and aB are recombinant. The proportion of recombinant children as a result of this meiosis is the recombination

fraction (θ). The larger the distance between the two loci, the higher the chance of recombination. If two loci are at 1 Centimorgan (cM) genetic distance, these are separated by recombination once in 100 meioses. On average the physical distance between the loci is then 880 kb. This is only a rough estimation of physical distance since some chromosomal regions undergo more recombination than others. Two loci at 1 cM genetic distance in a region of frequent recombination (hot spot) are in fact present at a physical distance much smaller than 880 kb and vice versa.

8. Positional cloning: finding a disease gene by defining its position

In 1991 the HGP was started to obtain the complete nucleotide sequence of the human genome. As one of the first steps in this huge biological project, the location of numerous microsatellite repeat loci was established to such extent that one marker per cM was identified and localised (Weissenbach et al., 1992; Murray et al., 1994; Weber and Wong, 1993). This genetic marker map has greatly facilitated the localisation of disease loci. Microsatellite loci are used for this purpose because these loci have many different alleles in the population and, as a consequence, many subjects in a study will be heterozygous and are, therefore, likely to produce informative meioses at such loci.

To find disease genes we again use linkage analysis, as for mapping marker loci but this time to test proximity of the unknown disease gene to the known position of marker loci on the genetic marker map. How can we test for linkage between the disease gene and a marker locus if we have not yet identified that gene? Although we cannot observe polymorphisms in the disease gene we can measure the presence or absence of the disease phenotype. Thus, one can estimate the genetic distance between a marker locus of known position and a disease locus (of unknown position) by observing the segregation of the marker locus in a pedigree together with the disease status and then counting through how many meioses an allele or a haplotype becomes transmitted together with the disease status (linkage analysis, see Fig. 5). Using the genetic marker maps as a tool, genome wide scans can be performed to localise the most likely position of a disease locus (Lander and Schork, 1994). In a genome wide scan one tests in a pedigree whether linkage can be observed of a set of adjacent marker loci with the disease status, using about 400 marker loci of known position, evenly distributed over the genome. As a result one may have excluded 98% of the genome and designated areas of 0–20 cM as the most likely location of the disease gene. In such regions (of up to 100 genes) the disease gene has to be identified. In family studies, a first step to get closer to a gene is to investigate the area of linkage in all available family members in search of informative recombinant offspring (Fig. 6). Once the minimal area of linkage is established, mutation analysis may be performed at likely candidate genes in such areas.

This is done firstly by demonstrating mutations in any of the genes in a region of linkage, secondly by showing an association of the mutated allele and the disease, and thirdly by showing that the mutation is causal to the disease in cell systems, transgenic mice, knock out mice, etc. As many genes in a region of positive linkage

will carry DNA sequence variations that may be neutral to the disease, identification of a gene on the basis of linkage is usually not an easy accomplishment. The approach in which the disease gene is first localised and subsequently identified is called positional cloning. Thus far, positional cloning has mainly been successful in the identification of Mendelian traits caused by a single locus such as the Huntington disease gene (localised at chromosome 4 in 1983, cloned in 1993). Many efforts are in progress to localise genes contributing to complex and common diseases such as diabetes type II (Davies et al., 1994), osteoporosis, schizophrenia and other heritable traits such as quantitative scales of anxious depression. In such complex late onset diseases and traits it is usually not possible to collect families with multiple affected and non affected family members, since the parents of patients are rarely alive and the children of the patients are too young to establish disease status. If only a single generation of subjects can be accessed, the optimal approach is a linkage study in large populations of sibling pairs (and additional siblings). In a sib-pair linkage study, it is calculated for any marker locus whether sibling pairs concordant for the trait also have a tendency to share alleles IBD above the expected sharing by chance. Alternatively, marker loci at which sibling pairs discordant (different) for the trait share alleles IBD below the expected sharing by chance, may also prove linkage of the marker locus to the trait. Identity By Descent (IBD) is estimated from highly polymorphic markers, for which it is unlikely that the same variant occurs in offspring through chance rather than through shared lineage. Estimation of IBD is greatly improved if parental DNA or DNA of additional siblings can be obtained.

9. Genetic association studies, linkage disequilibrium: testing candidate genes

Linkage studies are usually complemented by genetic association studies in patient populations and population based cohorts. In this approach, the genotype of markers within and surrounding specific candidate disease genes is investigated in groups of unrelated cases (e.g. patients) and controls (e.g. healthy subjects). Candidate genes may be studied by this approach based on their putative protein function, because they result from positional cloning, or because they are homologous with animal genes proven relevant in animal models of the disease. Genetic association studies are being performed for a wide variety of disease traits such as cardiovascular disease, rheumatoid arthritis, schizophrenia (Sklar et al., 2001; Mc Ginnis et al., 2001). HLA genes (chromosome 6) have frequently been associated with disease. Interestingly, paternally inherited HLA alleles have been associated with women's preferred choice of male odour (Jacob et al., 2002).

A DNA sequence variant that influences the expression of a gene is denoted 'functional variant'. The role of a candidate gene in a disease trait can be investigated by testing for association between the trait and a functional variant of the gene. For most of the human genes, however, the functional variants residing in the population are not known yet. The contribution of a candidate gene to a trait is then studied by testing for association between the trait and various markers in the gene (bi-allelic markers, most often SNPs) in search for the polymorphism that was created in an

evolutionary time scale shortly before or after the mutation occurred in the ancestor of the case (Fig. 7).

If an allele at a marker locus or a haplotype (the combined alleles of several loci) occurs at significantly higher frequency in the case group than in the control group, the locus is associated with the disease. This association is biologically meaningful if it is due to a causal mutation in the gene that is co-inherited with the marker allele/haplotype. As explained in Fig. 7, such meaningful association can only be found if the cases inherited the mutation from a common but distant ancestor.

If marker allele and disease allele are close to each other, they may co-segregate through many generations. These alleles are said to be in linkage disequilibrium (LD), they co-occur at frequencies higher than predicted on the basis of their individual allele frequencies. Most microsatellite loci are less suitable for genetic association studies than SNPs because they mutate to other length alleles at a relatively high rate, thereby, disturbing the LD between a specific marker allele and the disease allele. In both linkage and association studies one is searching for chromosomal segments detected by genetic markers that are common among cases (e.g. affected subjects) and not common among control individuals. Note that in association study one is testing for shared chromosome segments across many more meioses (on a much longer evolutionary time scale) than in most linkage studies. A disease allele shared by descendants (patients) from a remote common ancestor will only be detected by a very close marker locus of which an allele was coincidentally generated in the same evolutionary time frame as the disease locus (as explained in Fig. 7). Genome scans using association, therefore, require huge numbers of loci to be analysed (hundreds of thousands). Our knowledge of the patterns of LD in human populations is growing but remains far from complete. It has been suggested that LD blocks of up to 50 kb stable LD exist, interrupted by hotspots of variation (and disrupted LD). If such a pattern were true across the genome, a scan in association studies might require less (1 or 2 SNPs per block) markers (Goldstein, 2002). Others do not observe LD blocks and have demonstrated that presence of LD between loci is highly unpredictable but only slightly dependent on distance. In general, only sparse replication of findings from linkage and association studies is found in the current literature (Altmuller et al., 2001).

10. High throughput technology in future studies

The sequencing of the complete human genome and localisation of the position of the approximately 30 000 human genes was a huge task, much of which has now been accomplished. This explosion of genomic information created the tools for research on the function of these genes (functional genomics), i.e. a further understanding of transcription, translation and appearance as metabolites under various circumstances such as stress, inflammation, disease, etc. A recent technological development that will further contribute to the functional studies is the ability to simultaneously analyse tens of thousands of genes by DNA expression arrays. Although promising, high throughput technologies produce many false positive

findings because of the many tests that are performed. The follow-up to all such findings for verification in functional studies is usually performed at much lower speed if it can be done at all. Therefore, the new high throughput technology should go hand in hand with developments in the fields of biocomputation and statistical methodology to extract the relevant information from huge data sets. Such a holistic approach to the genome may prove the best route to identify the complex pathways leading to disease. Once it precipitates in testable ‘systems biological’ hypotheses, it likely requires the collection of many intermediate traits (endophenotypes) in human populations and the creation of animal models and experimental cell systems. When a functional mutation that increases disease susceptibility is finally identified, its effects on the endophenotypes and various clinical endpoints can be tested. Subsequently other variations in the same gene and other genes in the same pathway can be analysed in parallel. If identification of disease gene variants and disease specific gene expression patterns is successful, it may become feasible to use this genomic information to monitor and predict the progression of disease and response to therapy.

Acknowledgements

We would like to thank Professor Dr J.C.N. de Geus for helpful discussions and his contribution to this manuscript and Dr B.T. Heijmans for his contribution to this manuscript.

Appendix A: Glossary

Allele	one of the alternative forms of a gene at a specific chromosomal location (locus). At the autosomal loci each individual possesses two alleles, one inherited from the father, one inherited from the mother
Autosome	any chromosome other than a sex chromosome. Humans have 22 pairs of autosomes
<i>Cis</i> -acting element	DNA sequence that regulates expression of a gene present on the same DNA molecule or chromosome as the element
centiMorgan	a unit of genetic distance equivalent to a 1% probability of recombination during meiosis
Codon	a nucleotide triplet that specifies an amino acid or a translation stop signal
Diploid	the status of a cell that has two copies of the genome and a chromosome number of 2n. In humans, the diploid number is 46
Exon	the part of the DNA sequence in a gene that is transcribed into a spliced RNA transcript

Gamete	in humans, the gametes are the mature egg and sperm cells. Gametes are haploid (chromosome number n), having a single copy of the genome
Gene	the complete unit of DNA sequence that is transcribed and translated into a polypeptide, including the DNA sequences that regulate transcription
Genome	the entire complement of genetic material of a cell. The HGP defines the human genome as a single haploid set of nuclear chromosomes, plus the mitochondrial genome
Genotype	the genetic makeup of a cell or organism. Genotype can be contrasted with the phenotype (the detectable expression of cellular/organismal functions). Genotype can also refer to the two alleles of a locus in a diploid genome
Germ line	gametes and all precursor cells that give rise to them
Haplotype	a series of alleles found at linked loci on a single chromosome
Haploid	a cell (or organism) having only one copy of the genome and chromosome number n
Heterozygous	the condition in which two different alleles make up the genotype of a locus in a diploid cell. At that locus the homologous chromosomes carry a different DNA sequence variant
Homologue	one of a pair of chromosomes that contains equivalent genetic information. In gametes, homologues pair with one another during meiosis
Homozygous	the condition in which two identical alleles make up the genotype of a locus in a diploid cell. At that locus the homologous chromosomes carry the same DNA sequence variant
Intron	the non-coding part of the DNA sequence which separates neighbouring exons in a gene. Introns are transcribed in the primary RNA transcript but are then removed by RNA splicing
Linkage	the situation in which loci are located at such short distance on a chromosome that they become jointly transmitted through meiosis with a very small probability of recombination to occur between the loci
Locus	a specific position on a chromosome (of a gene, a polymorphism, etc)
Meiosis	a specialised process of cell division that reduces the chromosome number to the single (haploid) complement (n). Meiosis takes place in precursor cells of gametes. Meiosis produces four haploid daughter cells from one diploid precursor cell, involving two rounds of cell division and one round of DNA replication
Microsatellites	a type of DNA polymorphism that is characterised by varying numbers of tandemly arranged repeats of a short DNA sequence (usually 2–4 base pairs)

Mitosis	the process of cell division that ensures that each daughter cell receives an exact copy of the diploid (2n) complement of chromosomes
Mutation	a change in the DNA sequence organisation (such as a substitution, deletion or insertion of a base-pair or DNA sequence fragment). If the mutation occurs in somatic cells, the results affect only the individual bearing those cells; if the mutation occurs in the germ line, the change can be transmitted to offspring
PCR	polymerase chain reaction. A laboratory technique that permits the in vitro production of large amounts of a specific DNA sequence from a very small amount of sample DNA
Phenotype	the externally or internally detectable characteristics of an organism, including behaviour, that represent the influences of environmental and genetic information (genotype)
Polymorphic	any DNA sequence that has more than one variant at significant frequencies in the population
Promoter	a combination of short sequence elements (at the 5' end of the gene) to which the proteins of the transcription apparatus bind to initiate transcription of a gene
Somatic cell	in humans, somatic cells are all cells except cells of the germ line in ovaries and testes that will undergo meiosis
<i>Trans</i> -acting factor	protein that control the expression of all allelic variants of a gene (in contrast to <i>cis</i> -acting DNA elements that only affect expression of the allele they are part of)
Transcription	the process through which an RNA molecule is synthesised by complementary base pairing using DNA as a template. For example, mRNA is the mature product of transcription
Translation	the process through which a polypeptide is synthesised by sequential binding of amino acids using mRNA as a template

References

- Altmuller, J., Palm, L.J., Fischer, G., et al., 2001. Genome wide scans of complex human disease: true linkage is hard to find. *Am. J. Hum. Genet.* 69 (5), 936–950.
- Banks, R.E., et al., 2000. *Lancet* 18, 1749–1956.
- Bird, A., 1986. CpG islands and the function of DNA methylation. *Nature* 321, 209–213.
- Brett, D., Pospisil, H., Valcarcel, J., reich, J., Berk, P., 2002. Alternatice splincings and genome complexity. *Nat. Genet.* 30, 29–30.
- Cooper, D.N., Krawczak, M., 1993. *Human Gene Mutation*. BIOS Scientific Publishers, Oxford.
- Davies, J.L., Kawaguchi, Y., Bennet, S.T., et al., 1994. A genome-wide search for human type 1 diabetes susceptibility genes. *Nature* 371, 130–136.
- Deininger, P.L., Batzer, M.A., 1999. Alu repeats and human disease. *Mol. Genet. Metab.* 67, 183–193.
- Goldstein, D.B., 2002. Islands of linkage disequilibrium. *Nature Genet.* 29, 109–111.

- Jacob, S., Mc Clintock, M.K., Zolano, B., Ober, C., 2002. Paternally inherited HLA alleles are associated with woman's choice of male odor. *Nat. Genet.* 30, 175–179.
- Koob, M.D., Moseley, M.L., Schut, L.J., et al., 1999. Untranslated CTG expansion causes a novel form of spinocerebellar ataxia (SCA8). *Nature Genet.* 21, 379–384.
- Lander, E.S., Schork, N.J., 1994. Genetic dissection of complex traits. *Science* 265, 2037–2048.
- Lander, E.S., Linton, L.M., Birren, B., et al., 2001. The International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860–921.
- Lewin, B., 1997. *Genes VI*. Oxford University Press, Oxford.
- Mc Ginnis, R.E., et al., 2001. Failure to confirm NOTCH4 association with schizophrenia in a large population based sample from Scotland. *Nat. Genet.* 28, 128–129.
- McKusick, V.A., 1997. *Mendelian Inheritance in Man*. Baltimore, 12th ed. (The print version of the OMIM database). Johns Hopkins University Press.
- Murray, J.C., Buetow, K.H., Weber, J.L., et al., 1994. A comprehensive human linkage map with centiMorgan density. *Science* 265, 2049–2054.
- Sklar, P., et al., 2001. Association analysis of NOTCH4 loci in schizophrenic using family and population-based controls. *Nature Genet.* 28, 126–128.
- Strachan, T., Read, A.P., 1999. *Human Molecular Genetics*, vol. 2. BIOS Scientific Publishers, Oxford.
- Venter, J.C., Adams, M.D., Myers, E.W., et al., 2001. The sequence of the human genome. *Science* 291, 1304–1351.
- Wang, D.G., Fan, J.B., Siao, C.J., et al., 1998. Large-scale identification, mapping and genotyping of single nucleotide polymorphisms in the human genome. *Science* 280, 1077–1082.
- Weber, J.L., Wong, C., 1993. Mutation of human short tandem repeats. *Hum. Mol. Genet.* 2, 1123–1128.
- Weissenbach, J., Gyapay, G., Dib, C., Vignal, A., Morissette, J., Milasseau, P., Vaysseix, G., Lanthrop, M., 1992. A second generation linkage map of the human genome. *Nature* 359, 794–801.