RECOMBINANT DNA TECHNOLOGY AND HUMAN DNA-POLYMORPHISM

K.D. Zang and N. Blin

Institut für Humangenetik der Universität des Saarlandes, Universitätsklinik, 6650 Homburg/Saar, Federal Republic of Germany

Hitherto, only phenotypical markers from blood, other tissues and excretory fluids served in the process of identification of persons from stains and in affilitation analysis. They base on two phenomena:

- numerous human characteristics are monogenically inherited and display polymorphisms at the level of the gene product, i.e. they are not homogeneous;
- these polymorphisms strictly follow Mendelian laws of dominance and recessivity. Generally, they are inherited in a codominant way, in exceptions in a dominant/recessive way (e.g. ABO system).

The occurrence of a variant at a frequency of at least one percent is called polymorphism; in the codominant way of inheritance this corresponds directly to its gene frequency. The empirical value of 1% guarantees a population genetic equilibrium. Variants with a lower frequency are essentially determined by new mutations. This is useful for the identification of persons but may cause substantial problems in paternity testing.

In paternity cases the routinely applied set of about 20-30 systems (ABO, MNSs, P, Rh, K, Fy, Jk, Lu, Xg, Se, acP, AK, ADA, PGM1, GPT, EsD, GLO, HP, Gm, Km, Gc, C3, Bf, Tf, Pi und HLA A, B, C) already provides a cumulative a priori-probability of exclusion of 99.9 % and a plausibility of paternity in the same range according to Essen-Möller.

Disadvantages of these systems result from limited sample size from using quite different techniques, from the uneven distribution of the corresponding genes on the chromosomes and from linkage disequilibiria (e.g. HLA genes) which also have to be taken into account. In general, this set of standard polymorphisms is sufficient for forensic application. If neccessary, it may be extended to about 40 scientifically proven polymorphic systems.

The investigation at the level of the gene product, i.e. the protein, mirrors only vaguely the extent of the polymorphisms at the level of the gene. The diploid human genome is composed of about $2x3x10^9$ nucleotides (base pairs). Considering that three nucleotides code for one amino acid and that an average polypeptide chain shows a length of 150 to 350 amino acids a number of about $3x10^6$ primary gene products can theoretically be calculated. According to realistic estimates, however, only 20.000 to 100.000 genes exist or are expressed in man. This means that we are using less than 5 % of our genetic material translating it into gene products. Out of these, about one third demonstrate polymorphisms.

Advances in Forensic Haemogenetics 2
Edited by W.R. Mayr

Admittedly, this still represents quite a large number. Recombinant gene technology, however, allows to demonstrate all changes in the primary gene sequence, even those which occur in the 95 % of non-transcribed DNA.
Only in the minority of the cases, in particular when investigating genetic defects, we are trying to detect the (pathological) polymorphism of the gene product at the level of the gene itself. In the majority of these cases, anonymous DNA sequences which are dissected from the genome by restriction enzymes (see below) are used in recombinant gene technology. It was found that about each 1/100 to 1/200 base pair (corresponding to 0.5 - 0.1% of nucleotides) is changed by mutations thus generating a special class of "alleles" which are inherited as codominant polymorphisms at the DNA level and are present in a homozygous or heterozygous state depending on the molecular phenotype of the individual. At the level of gene product this allelism can generally not be recognized.

About 70% of our genome is represented by so called single copy genes. 20% are known as middle repetitive sequences (several thousand copies), 10% as highly repetitive sequences (several hundred thousand and more). Excluding several gene families, e. g. the ribosomal genes, leaves the repetitive DNA predominantly existing as short sequences ("minisatellites", see below). These can be found at multiple homologous loci of the corresponding sister chromosomes, for example in the centromeric region in a fashion of tandem repeats. They display an enormous individual variability in the number of copies and therefore present multiple alleles in an electrophoretic pattern specific for individuals. Point mutations and mutations of the number of repeats together result in a tremendous number of about 30 million polymorphisms at the DNA level which obviously seem difficult to relate to our biological existence. It was shown, however, that this high variability accumulated in the course of millions of years, mainly in the non-transcribed, i.e. genetically inactive DNA sequences - this means in sections of the genome which do not present a disadvantage in selection for the carrier. Analysis of these two different kinds of polymorphisms in human DNA presently results in two basically different approaches:
- the RFLP-analysis, i.e. detection and evaluation of the restriction fragment length polymorphisms which mainly represent a Mendelian two or few-alleles system;
- the analysis of the minisatellite sequences i.e. detection and evaluation of hypervariable regions which correspond to a multiallele system and deliver an individual-specific DNA pattern.

RFLPs

Using bacterial restriction endonucleases high molecular weight human DNA can be dissected into numerous fragments. These enzymes are distinguished by the fact that they do not cut the DNA unspecifically resulting in ever shorter oligonucleotides in the course of their action but that they only act in very specific recognition sequences or at a defined distance from this recognition sites. These recognition sites are composed of 4 to 8 base pairs. Presently, more than 200 of such enzymes catalyzing the specific section of double-stranded DNA are known. Their biological significance obviously lies in protecting bacteria from

foreign DNA, for example from penetrating bacteriophages. The application of these enzymes to human genomic DNA results in reproducible fragments of defined length which can be separated electrophoretically (for example using agarose gels) according to their molecular weight and thus to their length.

Mutations may lead to deletions, insertions or exchange of one or several nucleotides. These events can generate a new restriction site or the previously present site may disappear. This may change the size of the respective DNA fragments resulting in a different electrophoretic mobility of the bands. If this mutation presents no disadvantage for the carrier (e.g. by inactivating an important gene) it can be passed on to the next generation via germ line and can be demonstrated as a codominant polymorphism at the DNA level. As mentioned above, such restriction fragment lenght polymorphisms are rarely detected in active genes themselves. In close neighbourhood to a particular gene or in non-expressed intron sequences of a gene, however, they offer useful information for detection of patients or predisposed individuals in the area of clinical genetics by means of linkage analysis. For such an indirect analysis of the phenotype of all monogenic characteristics about 1.600 precisely localized polymorphic markers spread evenly throughout the human genome would suffice (Botstein et al., 1980). This is a number which will be reached shortly. For forensic means, the location of the markers is irrelevant and the number could be much smaller. In analogy to analysis of polymorphic proteins it is the number and relative frequency of alleles which is decisive.

DNA probes

If the complete genomic DNA of an individual were used in a restriction enzyme assay hundreds of thousands of fragments would be distributed across the gel according to their length and disguise a clear pattern, even more so a change of such a pattern. To reach this goal it is necessary to apply DNA probes. Such probes are presented by defined DNA sequences, even defined genes under certain conditions, which are amplified by cloning and radioactively labeled. Specific or anonymous DNA sequences of the human genome can be incorporated into a naturally occuring or a artificially constructed vector molecule (e.g. plasmid or bacteriophage). Dissection of the genomic DNA and of the vector molecule using the same restriction enzyme provides identical DNA termini and facilitates the incorporation of a human DNA sequence into the vector and the recircularization of the recombinant. This recombinant can be manipulated into host bacteria and thus, multiplied ad libitum (Figure 1).

This process of multiplying human DNA sequences into large copy numbers is called molecular cloning. Following the separation of the insert from the vector and tagging the DNA by a radioactive label (e.g. by "nick-translation") such DNA sequences may be used as a genetic probe in order to visualize particular genomic fragments by means of autoradiography (for technical details, see Maniatis et al., 1982). This procedure favourably applies such probes which are present in the human genome only once (single copy sequences), therefore, a simple and clear banding pattern can be obtained, as it is known from analyzing protein polymorphisms.

The technical steps are the following:

After electrophoretic separation the DNA fragments in the gel are denatured by alkali resulting in single-stranded DNA. Then, they are transferred from the gel to a solid support membrane (nitrocellulose or nylon). These steps constitute the classical "Southern blot" procedure. By incubating this membrane with a radioactively labeled probe a molecular hybrization occurs precisely at that position where the probe detects a homologous, complementary DNA sequence.

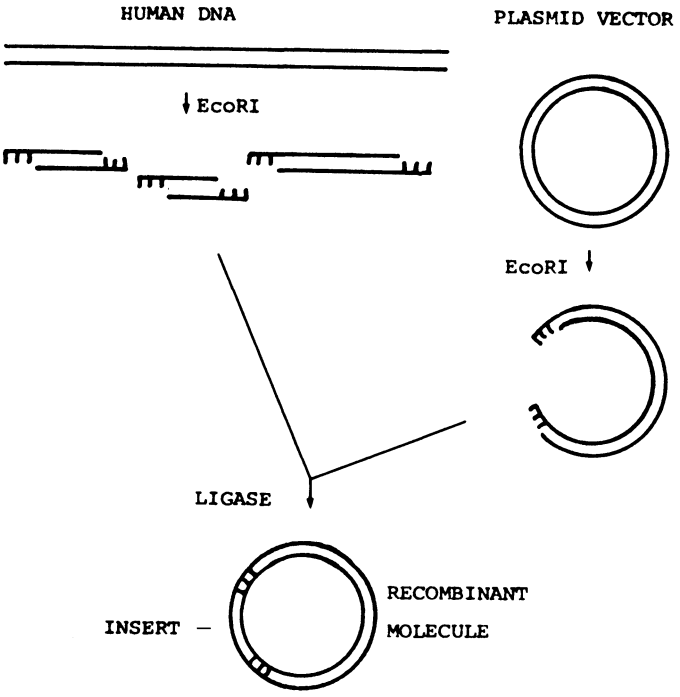

Figure 1. Double-stranded linear and circular DNA is dissected by using the EcoRI restriction endonuclease and generating "sticky ends". Then, the fragments are linked using ligase to form a recombinant DNA.

Finally, the specific bands and the resulting pattern are visualized by autoradiography (Figure 2). Since the filter can be washed after the autoradiographic step and successively incubated with different probes it is possible to search for several polymorphisms using the same gel. In the meantime, non-radioactive techniques have been developed which also show a high resolution. Under optimal conditions 20 micrograms of high molecular weight DNA can be isolated from one milliliter of blood. About 5 micrograms of DNA are necessary for gel electrophoresis; thus, DNA from one ml of blood can be studied using four different restrictions enzymes and four electrophoretic separations which then can be hybridized with a panel of probes and will result in a quadruplicate of information.

Minisatellites

Analysis of minisatellites is aimed at the enormous polymorphism of the middle and high repetetive DNA sequences. In contrast to the RFLP analysis we are confronted here with quantitative and not with qualitative polymorphism. The biological significance of the repetive sequences is presently not understood. It is speculated that they are somehow connected with gene protection ("body-guard-hypothesis"), stabilization of chromatin, gene regulation, and possibly spontaneous DNA recombination.

For forensic analysis it is decisive that these minisatellite sequences are scattered in tandem repeats throughout the entire genome and that the copy number varies substantially in different individuals. Futhermore it is important, that these clusters are

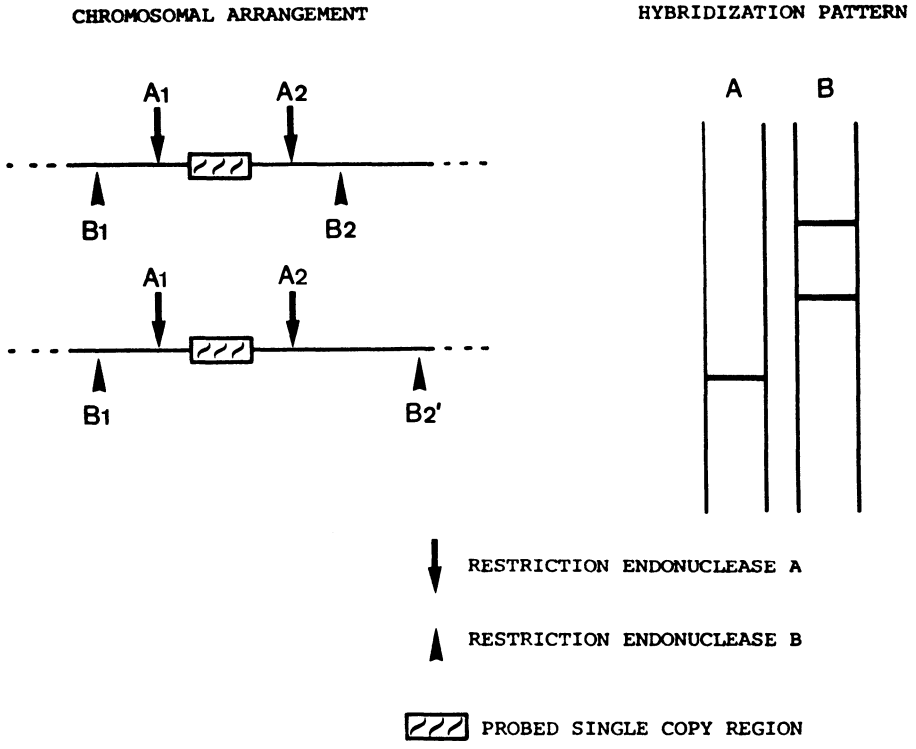

Figure 2. Schematic representation of two homologous chromosomes restricted by two endonucleases, A and B, respectively, yielding two corresponding fragments of identical size for A. B generates two larger fragmens of different size due to a deviating restriction site on one of the chromosomes (B2 and B2', respectively)

localized at corresponding loci of the homologous chromosomes. Thus, they present a form of alleles which do not, however, differ in their structure but in their length i.e. in the number of short core sequences.

The term "satellite DNA" was established because these repetitive sequences deviate in specific density from the bulk of genomic DNA by forming a separate layer in centrifugation tubes during DNA preparations in equilibrium density centrifugation.

Analysis of polymorphisms of this satellite DNA is directly connected with investigations performed by Jeffreys and his colleagues (1985, 1986). The practical steps are as follows: The DNA probe is formed by multiples of core sequences of the particular repetitive sequence (about 15 base pairs, Table 1). The genomic DNA is cut using a restriction enzyme which uses a short recognition sequence of high frequency (for example Hinf I) resulting in numerous fragments. The recognition sequence for this enzyme, however, has to be localized outside of the repeat in the flanking regions and not within the tandem repeats themselves. Otherwise two linked systems of banding patterns would be generated.

Table 1. DNA sequences of repeat units used for minisatellite probes

M13	$(GAGGGTGGXGGXTCT)_n$	core=15bp	(Vassart et al., 1987)
33.15	$(AGAGGTGGGCAGGTGG)_{29}$	core=16bp	(Jefferys et al., 1986)
33.6	$(AGGGCTGGAGG)_{54}$	core=11bp	(Jeffreys et al., 1986)

Given the conditions mentioned above, a set of fragments of varying sizes is produced. They migrate differently and form different bands which, however, can be detected using the same DNA probe. Jeffreys estimates 3 to 29 repeats as a variation range for his system of one core sequence. This would correspond to about 25 electrophoretically distinguishable alleles. When the individual number of repetitive sequences is evenly distributed over a population it would mean an average frequency of alleles of p = 0,04 and lead to the fact that practically all alleles are present in a heterozygous form ($h=1-25x(0.04)^2=0,96$). Consequently, in 96 % of the individuals two bands per locus would be recognized and located at different positions of the gel. For this system, Jeffreys coined the name "fingerprints" describing a pattern of multiple bands as shown in several publications. This complex pattern is caused by an additional polygenic situation. These repetetive sequences which can be demonstrated using one particular DNA probe occur at multiple loci in the human genome, their number corresponding to about half of the visible bands. According to the results obtained by Jeffreys one can expect 60 loci leading to a pattern of about 120 bands in each individual. For forensic purposes, this pattern is limited mainly to the fragments of higher molecular weight in the upper part of the gel containing an increased number of repeats.

In contrast to the RFLP analysis, using a set of several restriction enzymes would not yield new information (if we disregard possible polymorphisms in the flanking regions) but only a modified banding pattern. The information, however, can be

enormously increased when, according to Jeffreys, the repetitive core sequence used in the DNA probe is varied to some extent (e.g. probe 33.15 = 29 x 16 base pairs; probe 33.6 = 54 x 11 base pairs; see Table I). It is as yet unknown why this minor modification of the DNA sequence yields completely different patterns.

The pattern found in offspring can always be deduced from the paternal and maternal bands in a ratio of about 50:50, thus proving a strict codominant Mendelian way of inheritance of these patterns. Extensive pedigree analysis showed that some bands are simultaneously inherited indicating linkage of corresponding loci. Alternatively, the tandem repeats could be occasionally separated by short foreign sequences which then would contain a recognition site for a specific restriction enzyme. This situation explains the occurrence of "pseudo-haplotypes". The broad range of variation between individuals of the repetitive sequences at a particular locus is obviously explained by illegitimate uneven crossing over in the meiosis or by incorrect sister chromatid exchange of somatic chromosomes. The large number of alleles suggests a high mutation frequency. This fact is without relevance for the identification of persons, could, however, lead to misinterpretations in affiliation analysis. As generally known, the mutation frequency for point mutations is at about 10^{-4} to 10^{-6} per generation per locus.(This holds true also for the evaluation of RFLPs). It is estimated at about 10^{-3} for the polymorphisms of minisatellites basing on data by Jeffreys. If approximately 100 bands were analyzed in a child that would result in an a priori probability of 0.1% to find a new band, not present in the patterns of either mother or father. The probability for monozygotic twins, on the contrary, to differ by one band should be significantly lower since their separation takes place after meiosis. Presently, a precise estimation is impossible since the frequency of a somatic recombination between such repetitive sequences is unknown as yet.

The minisatellite analysis most likely will gain significant importance for forensic purposes since using a single probe will allow to evaluate a multigene family with numerous alleles. Thus, using a single test one can obtain a pattern which is specific for individuals. By now, however, the procedure is protected by patents. Even those laboratories possessing the Jeffreys' probes can solely use them for scientific purposes.

The Jeffreys probes are very interesting indeed. However, the phenomenon that already slight modifications of the core sequences result in a quite different but also reproducible fingerprint pattern encouraged many groups to work with artificial short nucleotide sequences (e.g. Epplen and coworkers, 1986) or with other natural probes. For example, Vassart et al. have demonstrated that highly specific fingerprint patterns in human genomic DNA can also be obtained by using a different, relatively simple procedure. The E. coli bacteriophage M 13, a very popular vector for cloning a whole range of various DNA sequences, contains a short sequence of 15 nucleotides (Table 1). These display no homology to the core sequences from other probes but also recognize hypervariable regions in the human genome and produce a pattern of about 40 bands. Preliminary experiments from our labo-ratory some month ago testing a family and using the whole M 13 phage DNA as probe

revealed already more than 15 bands (Figure 3). It remains to be seen whether other bacteriophages also contain such repeats which most likely have been conserved during evolution.

In the moment, the conventional procedures of gene product analysis cannot be renounced in forensic medicine. Frequencies of phenotypes are usually well known for different populations as well as linkage disequlibria for many systems. The courts are familiar with the nomenclature and the evidences. A disadvantage is found in the limited number of reproducible polymorphisms and in the relatively broad spectrum of different methods necessary for the investigations.

In paternity cases probably less than one percent of cases cannot be solved with conventional techniques. In criminological cases the percentage will be higher. So, in the long range, it will become inevitable - perhaps also more elegant - to also include the analyses of DNA polymorphisms. Their number is practically unlimited and they can all be detected using the same methodology. Presently, this system is restricted by the fact that only a few DNA probes are commercially available and that their cloning is limited to laboratories which are equipped with the necessary security facilities for gene technology. Like in the beginning of HLA analysis this only will constitute a transient inconvenience. In the backgroundt we can see already the commercial market offering a lot of kits to unexperienced investigations.

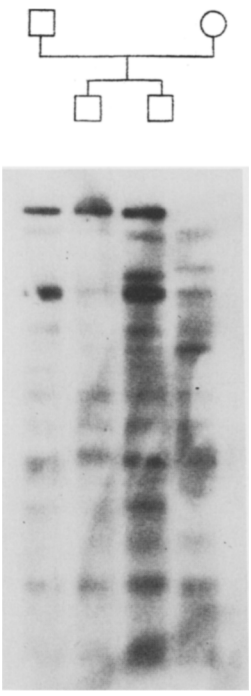

Figure 3. Autoradiogram and schematic representation of a M13-mini-satellite analysis in a family

Detection of restriction fragment length polymorphisms can be compared to that of protein polymorphisms in so far as in each case only one locus with two or several alleles is analyzed. Therefore, in affiliation analysis cumulative probabilities have to be calculated. Investigation of minisatellites, on the contrary, in each case leads to plain yes/no decisions, i.e. identification/exclusion

basing on its extreme polymorphism in a single analysis. Because of psychological reasons, it will surely become extremely difficult to base a forensic investigation on one single test system of such potency; possibly, it will not be advisable since a single, however improbable mistake will be of such decisive significance for the further fate of an individual. In cases where conventional systems (and we include RFLP analysis as well) cannot lead to a final decision minisatellite analysis will gain substantial importance whether used in the version of Jeffreys' probes or other simple probes available to all of us.

LITERATURE

Ali S, Müller CR, Epplen JT (1986) DNA finger printing by oligonucleotide probes specific for simple repeats. Hum. Genet. 74:239-243

Botstein D, White RL, Skolnik M, Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphism. Am. J. Hum. Genet. 32:314-331

Human Gene Mapping #8 (1985) Cytogenet. Cell Genet. 40:1-823

Jeffreys AJ, Wilson V, Thein SL, Weatherall DJ, Ponder BA (1986) DNA "fingerprints" and segregation analysis of multiple markers in human genome. Am. J. Hum. Genet. 39:11-24

Jeffreys AJ, Wilson V, Thein SL (1985) Hypervariable 'minisatellite' regions in human DNA. Nature 314:67-73

Maniatis T, Fritsch EF, Sambrook J (1982) Molecular cloning: a laboratory manual, Cold Spring Harbor Laboratory

Newmark P (1987) DNA fingerprinting at a price at ICI's UK laboratory. Nature 327:548

Vassart G, Georges M, Monsieur R, Brocas H, Lequarre AS, Christophe D (1987) A sequence in M13 phage detects hypervariable minisatellites in human and animal DNA. Science 235:683-684