Molecular pathology of single gene disorders

D J WEATHERALL

From the Medical Research Council Molecular Haematology Unit, Nuffield Department of Clinical Medicine, John Radcliffe Hospital, Oxford

SUMMARY Recent studies using recombinant DNA technology have led to an understanding of the basic molecular pathology of single gene disorders. Furthermore, methods are being developed for finding genes for conditions, whose underlying biochemistry is still not understood, or which may contribute to polygenic systems that underlie common diseases. As well as providing new approaches to carrier detection, prenatal diagnosis, and treatment of single gene disorders, these advances promise to provide important information about the pathophysiology of many common polygenic diseases.

Considering that it is only about seven years since the first human genes were cloned and sequenced, a remarkable amount of progress has been made in unravelling the molecular pathology of single gene disorders. We already probably have a good idea of the repertoire of molecular defects that underlie most of them, and a start has been made in trying to relate these lesions to associated clinical phenotypes. These advances have important practical implications for carrier detection and prenatal diagnosis of genetic diseases, and in the long term may enable us to start to understand the molecular basis for common polygenic conditions such as heart disease, diabetes, and the major psychoses. Thus the new techniques which have led to these advances are likely to have broad application in diagnostic pathology in the future.

In this short review I shall summarise what is known about the structure of normal human genes, describe some of the different types of defects which give rise to abnormal gene function and how they may have arisen, and describe a few examples of how it is possible to relate abnormal gene structure and function to the associated clinical picture. There are several accounts of recombinant DNA technology for non-specialised readers.1 2

Normal gene structure and function

Fig 1 summarises a typical human gene and the mechanisms by which its messenger RNA product is processed and translated. With a few exceptions, all mammalian genes examined so far are broken up into coding regions (exons) and non-coding regions called introns. The number of introns varies widely, ranging from only two in the case of the globin genes to about 50 in the gene for the α chain of collagen. Although the function of the introns is still far from clear, it has been observed that, by and large, they separate regions of genes that code for different functional domains of proteins. It has been suggested, therefore, that their presence offers an evolutionary advantage. The cutting out of intervening sequences supposedly facilitates the juxtaposition and hence joint expression of DNA sequences, which may previously have been widely separated throughout the genome and subsequently brought together by various recombination events. In this way the presence of introns would increase the speed at which selection for functionally useful fusion products might be produced. Certainly the sequence of some genes—that for the low density lipoprotein receptor, for example—looks like an evolutionary patchwork of diverse sequences with homology to a variety of other genes.

There are highly conserved sequences at the junctions between coding regions and intervening sequences. In all mammalian genes the dinucleotides GT and AG are found at the 5' and 3' ends of the introns. It is also apparent from comparative studies and analysis of the various human mutations, which will be described later, that there are other sequences at both the 5' and 3' ends of the introns near their junctions with exons that are also critical for the normal splicing of messenger RNA.
Fig 1 Typical mammalian gene and steps entailed in its transcription and translation. Exons are shown in black and introns (intervening sequence, IVS) unshaded. Regions of gene which code for untranslated portions of messenger RNA are indicated as NC (non-coding regions). Position of 5' regulatory boxes is indicated.

There are other highly conserved sequences in most mammalian genes. At the 5' flanking region there are three "boxes" of homology: the CACCC box, located between -87 and -95 (87-95 nucleotides upstream from the initiation codon); the CCAAT box, located between -72 and -77; and the TATA box located between -26 and -30 relative to the start site of transcription. There is good experimental evidence that these boxes represent cis acting sequences, which are required for the accurate and efficient initiation of transcription—that is, they are the major promoter sequences for structural genes. Such a sequence can be defined as a region on a DNA molecule to which an RNA polymerase binds and initiates gene transcription. At the 3' non-coding region of all mammalian genes there is a sequence AATAAA that is thought to be a signal sequence for the processing and polyadenylation of the 3' end of messenger RNA transcripts. The initiation codon of messenger RNA is invariably AUG while any one of three codons, UAA, UAG, or UGA, can act as termination signals.

When a gene is transcribed, the primary transcript contains both introns and exons; while in the nucleus a considerable amount of modification and processing occurs. Introns seem to be removed as a single piece after which the exons splice together. The splicing of pre-messenger RNA entails endonucleolytic cleavage and ligation of intron-exon junctions with the formation of a complex lariat-like structure. The splicing machinery must be capable of aligning and holding together the intermediates of pre-messenger RNA during the splicing process; it is now clear that several nuclear proteins have a role in this highly complex reaction. After the introns have been removed and it has been polyadenylated messenger RNA moves into the cytoplasm where it acts as a template for protein synthesis, the major steps of which are summarised in fig 1.

The way in which mammalian genes are regulated is not yet fully understood. DNA does not exist in cells in the form of a naked strand but is associated with various proteins to form chromatin. It is now apparent that the physical state of chromatin can vary in regions where DNA is being actively transcribed. Thus active genes are packaged in a changed form of chromatin that shows increased sensitivity to...
Molecular pathology of single gene disorders

**Fig 2** Human globin gene clusters.

digestion by nucleases such as DNase I. In general, genes that are being actively described are hypomethylated, compared with inactive genes. There is increasing evidence that other sequences play a part in the regulation of mammalian genes with respect to their expression in different tissues—for example, a number of so called enhancer elements with this property have been defined for different human gene families.

Many human genes are found in families at particular chromosomal locations. Among the best studied are the globin genes (fig 2) and the immunoglobulin “super gene” family. The α-like genes of globin are found on chromosome 16 and the β-like genes on chromosome 11; each cluster is arranged such that the genes are in the order in which they are expressed during development. Within these gene clusters there are a number of inactive pseudogenes which may be evolutionary remnants of once active loci. Scattered among and within the structural genes there are inherited base variations which may either produce new sites or remove pre-existing sites of cleavage by restriction enzymes, so giving rise to so called restriction fragment length polymorphisms (RFLPs). In addition, in some gene clusters so called hyper-variable regions (HVRs) have been found—that is, lengths of DNA which vary considerably in length and which are highly polymorphic in this respect. These regions, which usually consist of simple repeat units or satellite DNA, also constitute a valuable series of polymorphic markers.

Most human gene clusters examined to date contain other families of repetitive DNA sequences. For example, the Alu I family, so called because of the presence of a recognition site for this restriction endonuclease in the centre of the repeat sequence, constitutes a family of repeats of about 300 nucleotides which occur some 300,000 times within the human genome. These units have a high level of homology, and although they have no known function, at least in some cases there is evidence that they are transcribed. The general reader is referred to Lewin for an excellent account of gene structure.4

**Human molecular pathology**

Most of what is known about the molecular pathology of single gene disorders has been derived from a study of the genetic abnormalities of haemoglobin production, particularly the abnormal haemoglobins and the thalassaemias. Indeed, work in this field carried out over the past five years has probably given us a very good idea of the complete repertoire of mutations that can affect human genes. As knowledge about the molecular defects in other single gene disorders is amassing it is apparent that most of the abnormalities are similar to those which have been observed in different forms of thalassaemia. The table lists some of the different types of human gene mutations, together with the disorders in which they have been found. The molecular aspects of the globin gene disorders are reviewed by Weatherall and Wainscoat5 and Bunn and Forget,6 and other single gene disorders by Cooper and Schmidtke,7 Gusella,8 Orkin,9 and Davies and Robson.10

**SINGLE BASE SUBSTITUTIONS LEADING TO STRUCTURAL CHANGE IN A PROTEIN**

There are numerous examples of inherited structural changes in proteins. By far the best studied are the
haemoglobin variants. Adult haemoglobin consists of a pair of α chains comprising 141 amino acid residues and a pair of β chains with 146 residues. As each residue is coded by three bases, theoretically, there are a total of 2583 single base substitutions that are possible for these two genes. Of these, 1690 would result in an amino acid replacement, and about one third of these would cause a change in charge, allowing separation of the haemoglobin variant by electrophoresis. Remarkably, over 400 variants have been identified which, in most cases, can be explained by a single base substitution in the corresponding triplet codon of the globin gene. A few variants have amino acid replacements at two different sites on the same chain, three of which entail β chains with the β6 Glu→Val substitution that produces sickle haemoglobin. These may represent second mutations in genes which already contained the sickle mutation or may have resulted from crossing over between two variant β chain genes.

Usually, single amino acid substitutions have no effect on the overall length of a peptide chain. There are exceptions, however; again we have to look to haemoglobin for examples. Elongated globin chains can be produced in various ways. There is a family of chain termination mutations which give rise to elongated α globin chains. These result from single amino acid substitutions in the chain terminating codon UAA—for example, in haemoglobin Constant Spring there is a change from UAA to CAA; the latter is the codeword for glutamine. Thus instead of the chain terminating in the usual way glutamine is inserted and then 3' messenger RNA sequences which are not normally utilised are translated for another 90 bases until another in-phase stop codon is reached. Thus haemoglobin Constant Spring has α chains containing 30 additional residues at their C terminal end. Other single base substitutions in the termination codon can produce similarly elongated α chains but with different substitutions at position 142.

Elongated products can also be produced by frameshift mutations; the additional residues are found at the C terminal ends of the globin chains. There are also several examples of elongation of globin chains at the N terminal end. These usually result from replacement of the N terminal valine by methionine, and this residue is preceded by an additional methionine. This interesting change is the result of a change in the normal processing of newly synthesised globin. Methionine is the first residue to be incorporated but during translation of the nascent peptide the N terminal methionine is normally cleaved, leaving valine as the N terminal residue of the α and β chains. The replacement of valine for methionine probably inhibits the peptidase that normally cleaves the NH terminal methionine, leading to the synthesis of chains with the N terminal residues Met-Met. Interestingly, the substitution of a proline for a histidine residue at position 2 in the β chain also interferes with...
Molecular pathology of single gene disorders

the removal of the methionine residue on nascent globin chains.

Shortened gene products can also be produced, usually by non-homologous crossing over between chromosomes, with the deletion of a varying number of bases. There is one haemoglobin variant with a shortened globin chain that results from a nonsense mutation.

GENE DELETIONS AND VARIATION IN GENE NUMBER

There are many examples of partial or complete deletions of genes as the basis for inherited diseases (table). Furthermore, we are starting to gain insights about how such deletional events may have occurred. One field that has been particularly productive in this respect is the analysis of the α thalassaemias.

Fig 2 shows that there are two closely linked α globin genes on chromosome 16. In the α⁺ thalassaemias there is a deletion entailing this chromosome which leaves a single functional α gene. In many forms of α⁺ thalassaemia both α globin genes are lost. The most likely mechanism for the production of a chromosome with a single α globin gene is non-homologous crossing over between the two α globin gene loci after mispairing of homologous chromosomes during meiosis. Duplicated loci like the α genes have arisen by a reduplication event which is mirrored by regions of homology in the flanking regions of the particular genes involved. In the case of the α globin genes these regions are designated X, Y, and Z. In fact, several different crossovers have occurred within these homology boxes, resulting in different types of α⁺ thalassaemia. If the crossover theory is correct the reciprocal product of the crossover event, a chromosome carrying three α globin gene loci, should be observed. In fact, such cases have been found in every human population that has been observed to date. Similar mechanisms almost certainly play a part in the generation of variable numbers of γ globin genes on chromosome 11; individuals with one, three, or even four γ genes on chromosome 11 have been found.

There are other mechanisms for the production of gene deletions. Recent work on both the α⁺ thalassaemia gene deletions and on long deletions which involve the β globin gene cluster and give rise to the phenotype of hereditary persistence of fetal haemoglobin (HPFH) have shown that some of the breakpoints of the deletions entail sequences with many characteristics of the Alu I repeat sequences, as described earlier. Similarly, there is an example of a deletion involving the gene for the low density lipoprotein receptor, in which Alu-Alu recombination has occurred. Because of their high degree of homology the Alu repetitive sequences may serve as “hotspots” for recombination. On the other hand, at least some of the long deletions which produce HPFH and α⁺ thalassaemias seem to be through non-related sequences—that is, they are examples of so called illegitimate recombination.

Another interesting feature of deletions of the α and β globin gene clusters is that in many cases they are of similar length, although at different points along the genome.11 A rather novel mechanism has been proposed to explain these observations: the deletions are generated by the loss of chromatin loops at different stages of DNA replication as chromatin moves through specific attachment sites on the nuclear matrix. After breakage the two ends of the DNA become reunited with the loss of a loop; recent evidence in favour of this observation has been obtained from the study of an α⁺ thalassaemia deletion, in which the gap across the deletion seems to have been filled in by a DNA sequence that is normally found at least 34 kilobases upstream from the site of the deletion. This could only have happened if the deletion entailed a large loop of DNA which brought the “filler” sequence into the appropriate place and orientation.12

FUSION GENES

Another interesting and important result of abnormal chromosomal crossing over is the production of fusion genes which code for hybrid proteins. The first and best studied example is haemoglobin Lepore, which has normal α chains combined with non-α chains that have the N terminal amino acid sequence of δ chains and the C terminal sequence of β chains. This variant seems to have arisen through non-homologous crossing over between part of the δ locus on one chromosome and part of the β locus on the complementary chromosome. Of course, such an event should give rise to two abnormal chromosomes, one with a Lepore gene and the other with its opposite counterpart, an anti-Lepore gene. In fact, both these arrangements have been found in a number of patients. A similar mechanism seems to have played a part in the production of one of the sialo-glycoproteins of the red cell surface.

Another elegant example of the generation of variation in gene number and the production of fusion genes is provided by recent studies on the molecular genetics of colour vision and colour blindness.13 Human colour vision is based on three light sensitive pigments. The genes for the red and green pigments show 96% identity and lie in a tandem array on the X chromosome on which there is a single red pigment gene and variable numbers of green pigment genes; the blue pigment gene shows less homology and is on
an autosome. Many of the different forms of red-green colour blindness seem to have resulted from unequal crossing over between the red and green pigment genes with the production of a variety of different fusion genes.

**INVERSIONS**

There is one example of a gene inversion in man. This has been found in several patients with the phenotype of δβ thalassaemia; the inversion occurs in a region of DNA between the δ and γ globin genes; and there is also a small deletion at each end of the inversion. A model has been proposed whereby an inversion of this type is generated by interactions between two chromosomal loops.

**NONSENSE MUTATIONS**

Several point mutations have been described, which cause scrambling of the genetic code and hence make it impossible for the translation of a normal gene product. Again most of these examples occur in thalassaemia, although similar lesions have been observed as the basis for disorders such as haemophilia, Christmas disease, and several other single gene conditions. The first of these to be identified was a substitution of codon 17 of the β globin messenger RNA, AAG→UAG, which changes a lysine codon to a premature termination codon. Another premature termination codon of this type is commonly found in patients with β thalassaemia—in this case a change in codon 39, CAG→UAG; CAG codes for glutamine in the normal β globin chain. Clearly, if there is a premature chain termination codon in the middle of a structural gene, translation will cease prematurely with the production of a shortened and physiologically useless peptide chain.

Another way in which the genetic code can be scrambled is by the generation of a so called frame-shift mutation, the basis of at least seven different forms of β thalassaemia and some cases of haemophilia and Christmas disease. As proteins are encoded by a triplet code the loss or insertion of one, two, or four nucleotides in the coding region of a gene will throw the reading frame out of sequence. As a result, a completely anomalous amino acid sequence will be added to a normally initiated globin chain. Sometimes the changed base sequence generates a new termination codon leading to premature termination of translation of the abnormal messenger RNA. Occasionally the messenger RNA may be translatable, and in this case there is a complete change of sequence from normal after the site of the frame-shift mutation. As mentioned earlier, at least one form of human haemoglobin variant with an elongated β chain results from a frame-shift mutation; in this case the normal stop codon is rendered out of sequence and therefore the scrambled messenger RNA is translated until another stop codon is produced so leading to an elongated translation product.

**DEFECTIVE PROCESSING OF MESSENGER RNA**

As mentioned earlier the primary transcript has to be processed by the removal of introns, joining together of exons, and by polyadenylation. Work on thalassaemia has provided a wealth of examples of molecular pathology involving these complex processes.

We have already discussed how normal splicing of messenger RNA depends on having GT and AG dinucleotides at the 5' and 3' intron-exon junctions. There are several examples of forms of β thalassaemia in which a single base substitution in one of these critical sites completely abolishes β globin chain production; no normal messenger RNA is produced. These findings underline the critical importance of these sequences for normal splicing.

There are, however, much more subtle abnormalities of messenger RNA processing due to point mutations (table). Single base substitutions within introns may result in preferential alternative splicing of the precursor β messenger RNA molecules at the site of the mutation—for example, a common form of β thalassaemia which occurs in the Mediterranean population results from a single nucleotide substitution, G→A, at position 110 of the first intervening sequence of the β globin gene. This change produces an AG sequence which happens to be preceded by a stretch of pyrimidines and so forms a good 3' acceptor consensus sequence. Thus about 80% of the processed messenger RNA is the result of splicing into this site rather than the normal 3' IVS 1 AG. The messenger RNA produced from the abnormal splic-
Molecular pathology of single gene disorders

Fig 4 Some point mutations at intron-exon junctions of first intron of human β globin gene. Different thalassaemia phenotypes are shown.

Several other forms of thalassaemia have been described which result from the production of alternative splicing sites within introns: there are two varieties which result from single base changes at positions 5 or 6 at the 5' end of the first intervening sequence (fig 4). Splicing occurs both at the normal site and at the new sites generated by these base changes. The effect of these point mutations is remarkably subtle. Some mutations at position 5 cause a severe defect in β chain production, while those at position 6 are associated with an extremely mild phenotype. Several other types of thalassaemia have been described in which point mutations in the second intervening sequence cause similar alternate splicing and hence abnormal β globin messenger RNA molecules.

Perhaps even more remarkable is the fact that mutations have been found in exons of the globin genes that seem to activate cryptic splice sites. One of these is particularly interesting because it is also associated with the production of a structural haemoglobin variant, haemoglobin E. This variant has a lysine for glutamic acid substitution at position 26. This results from a codon change GAG→AAG (fig 5). The latter seems to activate a “cryptic” splice site which competes with the normal 5' splice site and hence leads to a reduced output of β globin chains. This may be why haemoglobin E is associated with a β thalassaemia phenotype.

Several other single gene disorders are now known to result from splicing defects. These include one form of phenylketonuria and several varieties of haemophilia and Christmas disease. Recent studies suggest that one form of factor IX deficiency results from a deletion of an entire exon of the factor IX gene. Despite this an abnormal gene product is produced, presumably by linking the remaining exons together during messenger RNA processing. The functional importance of this unexpected finding is discussed below.

Finally, polyadenylation site mutations may also interfere with the normal processing of messenger

Fig 5 Activation of cryptic splice site in exon 1 as basis for some genetic disorders of β chain production. 

G→A change, which is responsible for amino acid substitution in haemoglobin E, activates cryptic splice site that is responsible for reduced rate of production of haemoglobin E and associated thalassaemia phenotype. Haemoglobin Knossos has similar phenotype with same molecular basis. T→A substitution produces no amino acid difference but results in reduced rate of β globin chain production and, again, phenotype of β thalassaemia.
RNA—for example, the single base change AATAAA→AATAAC, which is found in the α globin genes of patients with α thalassaemia in the Middle East and Mediterranean region, seems to prevent almost entirely the production of α globin chains from the α2 globin gene. Instead of the normal cutting and polyadenylation of the messenger RNA precursor a long molecule is produced which does not appear in the cell cytoplasm. There may be a small amount of polyadenylated messenger RNA produced but the overall effect is to inactivate almost entirely the α2 globin gene. It is also possible that this mutation may in some way interfere with the termination of α2 globin gene transcription. A similar polyadenylation site mutation has been observed in the β globin gene as the basis for one type of β thalassaemia.

**INITIATION CODON MUTATIONS**

Several mutations have been observed in patients with α thalassaemia, entailing either the initiation codon itself or the sequences which immediately precede it. As would be expected these cause a complete absence of normal α gene product.

**PROMOTOR BOX MUTATIONS**

Several forms of β thalassaemia have been described in which point mutations have been found upstream from the β globin gene (fig 6), either within or adjacent to the promoter boxes described earlier. These mutations are associated with a variable reduction in output from the adjacent locus. Their existence underlines the importance of these highly conserved regions of DNA and confirms their likely promotor function.

**Structure function relations**

It is now possible at least to make a start in trying to understand how the molecular pathology of human genes is reflected in differences in clinical phenotypes.

In this section I shall summarise a few examples of recent successes in this important aspect of human molecular genetics.

**HAE MOGLOBIN**

Structure-function relations have been studied in detail for the many varieties of structural haemoglobin mutations. Amino acid substitutions at critical sites may change oxygen binding properties of the molecule in a variety of ways, including modification of the interaction between subunits. This occurs as part of the allosteric changes involved in producing a sigmoid oxygen dissociation curve, by modifying interactions with regulatory molecules such as 2,3 diphosphoglycerate, or by interfering with critical residues within the haem pocket, and leads to the production of permanent met-haemoglobinæmia. A variety of clinical phenotypes are associated with these mutations, including genetic polycythaemia and forms of congenital cyanosis. Chronic haemolytic anaemia may result from point mutations which lead to changed molecular configurations such as occur in the sickling and haemoglobin C disorders, or by a variety of different mutations which produce molecular instability.

**POST-TRANS LATIONAL MODIFICATION; INSULIN AND SIGNAL SEQUENCES FOR FACTOR IX AND GLOBIN PROCESSING**

As mentioned earlier some proteins undergo a considerable amount of post-translational modification, and it is becoming apparent that point mutations may interfere with this process. Insulin consists of two dissimilar peptide chains, A and B, linked by two disulphide bonds. Unlike many other proteins which consist of structurally distinct subunits, however, insulin is under the control of a single gene locus and chains A and B are derived from a one chain precursor, pro-insulin. Proinsulin is converted to insulin by the
enzymatic removal of a segment that connects the amino end of the A chain to the carboxyl end of the B chain, called the C peptide. Familial hyperproinsulinaemia results from mutations at the cleavage sites connecting the A chain to the C peptide. There have been a few reports of structurally variant insulins that are functionally defective: the replacement of phenylalanine by leucine at position 24 in the insulin gene is associated with a diabetic phenotype.

Another recently described defect in posttranslational modification of a protein is a mutation responsible for one form of Christmas disease. Factor IX is synthesised as a precursor and might be expected to be proteolytically cleaved in at least two positions during maturation to remove a prepeptide and a propeptide region. One form of Christmas disease results from a single amino acid substitution at position —4 in the propeptide region: an arginine is replaced by a glutamine. This change results in the expression of a stable, longer protein with 18 additional amino acids of the N terminal propeptide region still attached. This is an important finding because it suggests that during the normal maturation of factor IX a signal peptidase cleaves the peptide bond between amino acid residues —18 and —19, generating an unstable pro-factor IX intermediate. Further proteolytic processing to the mature factor IX molecule must depend on the arginine residue at position 4.

Interestingly, the arginine at —4 is not unique to the factor IX precursor but is also found in factor X and prothrombin and in many other sequences processed by site specific trypsin-like enzymes—C3, C4, and C5 of the complement system and tissue type plasminogen activator.

As mentioned earlier, haemoglobin variants have been found in which there are elongated globin chains entailing residues at the N terminal end. These offer some particularly interesting insights into posttranslational modification of proteins. The amino terminal methionine residue, the translation product of the AUG initiation codon, is present only transiently in the nascent peptide chains of most proteins. One variant that has been analysed in detail, haemoglobin Long Island, shows a methionine residue at the end of the β globin chain; the second residue is proline instead of histidine. It seems likely that the latter base substitution results either in a structural or charge difference in the nascent β globin chain, which interferes with a methionine aminopeptidase mechanism, or causes a change in the secondary structure of the messenger RNA of sufficient magnitude to impair the removal of the amino-terminal methionine residue.

In proteins that are secreted or membrane-bound methionine constitutes the amino-terminal residue of a peptide of about 20 residues in length is apparently essential for both protein secretion and its incorporation into membranes. This “signal sequence”, which is hydrophobic, varies in length and sequence (except that methionine is always present at the amino-terminal end) and it is cleaved by a membrane bound enzyme in secreted proteins. In fact, it has been found that the signal sequence is only one of the constituents of an 11S protein termed the “signal recognition protein”. As haemoglobin is a cytoplasmic non-secreted protein it remains to be established whether the preserved amino-terminal methionine present in these mutant haemoglobins has any effect on the processing or cellular compartmentalisation of this mutant.

**RECEPTOR FUNCTION: MUTATIONS OF THE LOW DENSITY LIPOPROTEIN RECEPTOR**

One of the most elegant emerging stories of structure-function relations at the molecular level is the elucidation of the mutations that involve different regions of the low density lipoprotein receptor gene, and which lead to the disruption of the normal control of cholesterol metabolism. The low density lipoprotein receptor is a cell surface glycoprotein. It is synthesised in the rough endoplasmic reticulum as a precursor, after which it travels to the Golgi complex and hence to the cell surface, where it is capable of binding two proteins, apo B which is the sole protein of low density lipoprotein, and apo E. These receptors undergo a quite remarkable recycling process. They appear on the cell surface in coated pits and within a few minutes of their formation the pits invaginate to form endocytic vesicles. Multiple vesicles fuse to create larger sacs called endosomes. When the pH of the endosome falls below 6.5 the low density lipoprotein dissociates from the receptor. The receptor then returns to the surface. Each low density lipoprotein receptor makes one round trip about every 10 minutes in a continuous fashion, whether or not it is occupied by low density lipoprotein.

The gene for the low density lipoprotein receptor has been isolated and sequenced and most of its protein structure has been worked out. Patients homozygous for familial hypercholesterolaemia show a variety of different mutations which, based on the way in which they change receptor function, can be separated into four classes. In the first, no receptors are synthesised. In one case this is because there is a large deletion of the low density lipoprotein receptor gene. A second class of mutations result in a reduced rate of transportation from the endoplasmic reticulum to the Golgi apparatus; the receptors do not appear on the surface of the cell but remain in the endoplasmic reticulum until they are degraded. The molecular defect has not yet been determined. A third class of mutations is characterised by normal receptor synthesis but failure
to bind low density lipoprotein. It is believed that these mutations entail amino acid substitutions, deletions, or duplications in a cystein rich low density lipoprotein binding domain. Finally, there is a class of mutations in which the receptors reach the cell surface and bind low density lipoprotein but fail to cluster in coated pits. Three different molecular changes have been defined as the basis for this abnormality, all of which entail the participation of the cytoplasmic tail of the receptor that protrudes into the cell cytoplasm. In one case a tryptophan codon has been converted to a nonsense codon. In another there is a single amino acid substitution, tyrosine for cysteine, again in the middle of the cytoplasmic tail domain. These remarkable studies not only underline the extraordinary molecular heterogeneity of what seem to be similar genetic disorders but also provide considerable insights into the physiology of receptor function.

PHENOTYPIC VARIABILITY: \(\beta\) THALASSAEMIA INTERMEDIA\(^5\)

Analysis of the molecular basis of the thalassaemias also provides some important information about the way in which phenotypic variation can occur. Most \(\beta\) thalassaemia homozygotes have a severe transfusion dependent disease. Some, however, have a milder condition called \(\beta\) thalassaemia intermedia. In many of the severe forms of \(\beta\) thalassaemia no \(\beta\) globin chains are synthesised. The molecular basis of these disorders is heterogeneous and may result from deletions, nonsense mutations, frame-shift mutations, or point mutations at the intron/exon junctions. There are, however, milder forms of \(\beta\) thalassaemia in which \(\beta\) chains are produced but at a reduced rate. Some of these conditions result from point mutations in the promotor regions of the \(\beta\) globin genes. Many of them, however, are caused by single base changes in consensuses regions at the intron/exon junctions or from the activation of cryptic splice sites, all of which provide the opportunities for alternative sites of splicing of messenger RNA. Depending on the degree to which the normal, compared with the abnormal, splice site is used, a whole series of phenotypes of varying severity are produced.

Molecular studies of the \(\beta\) thalassaemias have also shown how the interaction of one or more genetic variants at other loci can modify clinical phenotypes. Patients with \(\beta\) thalassaemia have imbalanced globin chain synthesis, and the severity of the disorder depends on the degree of excess \(\alpha\) chain production; excess \(\alpha\) chains precipitate in red cell precursors and cause ineffective erythropoiesis. Thus patients who are \(\beta\) thalassaemia homozygotes, who also inherit \(\alpha\) thalassaemia that reduces the amount of excess \(\alpha\) chains, are phenotypically milder than those with \(\beta\) thalassaemia alone. Similarly, patients with \(\beta\) thalassaemia who inherit genetic determinants that allow persistent fetal haemoglobin production that allow persistent fetal haemoglobin production that allow persistent fetal haemoglobin production that allow persistent fetal haemoglobin production that allow persistent fetal haemoglobin production that allow persistent fetal haemoglobin production that allow persistent fetal haemoglobin production that allow persistent fetal haemoglobin production that allow persistent fetal haemoglobin production that allow persistent fetal haemoglobin production that allow persistent fetal haemoglobin production that allow persistent fetal haemoglobin production. Thus it is becoming apparent that when we know what kinds of genetic interactions to look for at the molecular level a rational explanation for the remarkable variation in expression of mutant genes within members of the same family may be found.

ANTIBODY PRODUCTION IN HAEMOPHILIA AND CHRISTMAS DISEASE\(^1^8\)\(^1^9\)

Recent studies of patients with haemophilia and Christmas disease are also starting to turn up some interesting genotype phenotype associations. Individuals with extensive deletions of the factor IX gene show no factor IX antigen and, in many cases, produce factor IX inhibitors. Recently a patient has been described with Christmas disease in whom there was detectable antigen but whose DNA showed a 2.8 kb deletion that had removed an entire exon and surrounding sequences. Surprisingly, however, this defective gene is transcribed and translated, presumably due to some novel type of splicing event. This patient did not produce any factor IX inhibitors. Thus it seems that even with large deletions some kind of immunologically recognisable gene product can be produced. These observations underline the subtlety of the relations between the molecular defects which underline single gene disorders and their clinical phenotypes.

THE NEOPLASTIC PHENOTYPE\(^2^0\)

Although the story is as yet incomplete, another very important example of structure-function relations seems to be emerging from studies of the rare childhood cancers, which seem to be inherited in a recessive fashion. The best examples are retinoblastoma and Wilms' tumour. The genetic determinants with a role in the production of these tumours are found on chromosomes 13 and 11. In each case individuals may be carriers for a recessive gene on these chromosomes which, in the presence of a normal allele, causes no clinical disability. If the normal allele is lost by some form of genetic rearrangement, however, the unopposed action of the mutant locus produces a malignant phenotype. Restriction enzyme analyses of malignant cells compared with those of normal somatic cells have shown that rearrangements, which give rise to inactivation of the normal allele, may occur in a variety of ways. Other examples of this phenomenon include widespread embryonal tumours (Beckwith-Weidemann syndrome), some bladder cancers, and acoustic neuroma and meningioma. So far, the action of these recessive "tumour genes" have not been worked out. Undoubtedly, the elucidation of their function and of
Molecular pathology of single gene disorders

the effects of these mutations will provide an important chapter in the evolving story of the molecular basis of neoplastic transformation.

It is already apparent that point mutations in cellular oncogenes can sometimes change their function, such that they are capable of producing malignant transformation in appropriate cell lines. Furthermore, studies of disorders like chronic myeloid leukaemia and Burkitt’s lymphoma, in which specific chromosome translocations move oncogenes into other regions of the genome and hence cause their activation, are also providing important information about the steps in neoplastic transformation. The Philadelphia chromosome, which is found in chronic myeloid leukaemia, results from a reciprocal exchange between chromosomes 9 and 22. There is relocation of a portion of the cellular oncogene c-abl and fusion of it with a newly identified locus called bcr (breakpoint cluster region). The genetic fusion creates a chimeric protein that includes the functional domain of the c-abl gene product whose enzymatic activity is more stable than that of the normal gene product.

Although a variety of changes in oncogene structure, location, number, and activities have been found in association with cancers in man, it is still not clear how these findings relate to the generation of a neoplastic phenotype, nor, indeed, whether changed oncogene function is a primary or secondary event. Of the 40 or so cellular oncogenes that have been described, only four major classes of associated biochemical properties have been defined: protein phosphorylation; metabolic regulation of proteins that bind GPT; control of gene expression by influencing biogenesis of mRNA; and the participation in replication in DNA. Genetic changes in oncogene function might cause these cellular regulatory genes to malfunction by causing constitutive activity and a surfeit of an otherwise normal gene product, or by changing the manner by which a protein acts, such as varying the substrate specificity of a protein kinase or the specificity of a transcription factor.

The proliferation of cells is controlled by an elaborate circuitry that stretches from the surface of the cell to the nucleus. Bishop suggested that the products of cellular oncogenes may represent some of the junction boxes in the circuitry, polypeptide hormones that act on the surface of the cell, receptors for these hormones, proteins that carry signals from receptors, and nuclear functions may all interact to orchestrate the genetic response to afferent commands. It is suggested that oncogenes may act as short circuits at these junction boxes, but all this is highly speculative at the moment, and it is still not possible to relate disordered oncogene activity to the primary event which renders a particular cell line neoplastic.

INHERITED DISORDERS OF COLLAGEN

Considerable progress has been made in defining the genes that regulate the structure of human collagen. Clearly there must be considerable constraints on the structure of the fibrillar collagen genes, and it follows that serious mutations must have been eliminated in the population. There are a variety of single gene disorders associated with defects in connective tissue, however, and it is quite possible that inherited connective tissue fragility might contribute to the inheritance of or susceptibility to polygenic conditions such as osteoporosis, osteoarthritis, idiopathic scoliosis and the floppy mitral valve syndrome.

Progress has been made in some of these areas, work which has considerable importance for structure-function relations. Several mutations of the collagen genes have been found in patients with osteogenesis imperfecta. In general, there are two classes of mutation of the structural loci for collagen that have different phenotypic effects. The first is mutations which lead to the expression of a mutant α chain that is incorporated into the collagen molecule. The second class is mutations that cause exclusion of the abnormal gene product from the molecule. It turns out that the former class of mutants is more severe because of the effect they have on the configuration of collagen. At least two well documented cases of a lethal form of osteogenesis imperfecta are produced by mutations of this type. These early studies on the differential effects of collagen mutations indicate that further studies of genetic variability of collagen will be of particular value in understanding structure function relations.

DEVELOPMENTAL MUTATIONS

One of the central questions in human biology is how genes are switched on and off at specific times during human development. The globin genes offer a particularly good example of this phenomenon as human haemoglobin changes its structure between embryonic, fetal, and adult life. There is a group of conditions with the general title hereditary persistence of fetal haemoglobin, in which there is a genetically determined defect in the normal switch from fetal to adult haemoglobin production. A family of point mutations has been found in the region -140 to -202 in the γ globin genes that are associated with the persistent production of the γ genes of fetal haemoglobin into adult life. These observations raise the intriguing possibility that there are critical regulatory regions of DNA, which may have a role in interactions with specific proteins that play a part in the suppression of fetal globin genes during adult life. This seems to be a particularly promising area for further study of the developmental regulation of the globin gene families.
Conclusions

Studies of the molecular basis for single gene disorders are providing a remarkable insight into the repertoire of lesions at the DNA level which are responsible for single gene disorders. As information of this type accumulates we can start to build up a picture of how these molecular lesions are reflected in the clinical phenotypes of patients with these conditions. In particular, we can start to identify some of the factors in the remarkable variation in the clinical picture that are associated with what is apparently the same type of genetic defect. As well as being of considerable clinical value in terms of counselling and prenatal identification of serious genetic diseases, this new information is providing considerable insights into the way in which messenger RNA is processed and into the structure-function relations of proteins such as enzymes and cell surface receptors. Now that it is possible to construct mutant proteins by the new techniques of protein engineering, it should be possible to analyse these findings in more detail. This field is also starting to yield tantalising clues about the general nature of neoplastic transformation.

Our ultimate goal will be to understand in detail how individual genes are regulated and expressed in specific tissues at particular developmental stages. It is also hoped that it will be possible to start using the same types of techniques as have been used to study single gene disorders to analyse the much more complicated question of polygenic disease and hence to understand the genetic factors which contribute to common conditions such as vascular disease, diabetes, and some of the major psychoses.

As far as single gene disorders go the ultimate objective is to reach a stage at which we understand enough about their regulatory sequences that we are in a position to replace defective genes. Until that time is reached, however, it is quite apparent that recombinant DNA technology will assume an increasingly important role in diagnostic pathology, certainly in the analysis of single gene disorders, in the further understanding of neoplastic transformation, and, of course, in the wider diagnostic aspects of infectious, degenerative, and malignant diseases, which it has not been possible to cover in this short review.

References


7 Cooper DN, Schmidtle J. Diagnosis of genetic disease using recombinant DNA. Human Genetics 1986;73:1–11.


16 Prchal JT, Cashman DP, Kan YW. Hemoglobin Long Island is caused by a single mutation (adenine to cytosine) resulting in a failure to cleave amino-terminal methionine. Proc Natl Acad Sci USA 1986;83:24–7.


Molecular pathology of single gene disorders.

D J Weatherall

doi: 10.1136/jcp.40.9.959

Updated information and services can be found at: [http://jcp.bmj.com/content/40/9/959](http://jcp.bmj.com/content/40/9/959)

**Email alerting service**

Receive free email alerts when new articles cite this article. Sign up in the box at the top right corner of the online article.

**Notes**

To request permissions go to: [http://group.bmj.com/group/rights-licensing/permissions](http://group.bmj.com/group/rights-licensing/permissions)

To order reprints go to: [http://journals.bmj.com/cgi/reprintform](http://journals.bmj.com/cgi/reprintform)

To subscribe to BMJ go to: [http://group.bmj.com/subscribe/](http://group.bmj.com/subscribe/)