Polymorphism of the Human Genome: Markers for Genetic Linkage Analyses in Heritable Diseases of the Skin

Angela M. Christiano and Jouni Uitto

Departments of Dermatology, and Biochemistry and Molecular Biology, Jefferson Medical College, and Section of Molecular Dermatology, Jefferson Institute of Molecular Medicine, Thomas Jefferson University, Philadelphia, Pennsylvania, U.S.A.

The human haploid genome consists of approximately three billion nucleotides (bp) throughout which are dispersed genes encoding upwards of 100,000 unique polypeptides. Genetic diseases can reflect deleterious alterations in as few as one of the three billion bp. Localization of the altered gene and precise identification of the mutation will provide understanding of the molecular basis of such diseases. Recently, several approaches of molecular biology have been developed to precisely map disease genes using naturally occurring variation, or polymorphism, in the DNA as markers. This mini-review summarizes some of the advances in the technology for mapping of disease genes and illustrates the feasibility of the approaches in the study of heritable skin diseases.

POLYMORPHIC NATURE OF THE HUMAN GENOME

DNA polymorphism is a reflection of the plasticity of the human genome, and it is believed that there is variability occurring within every 100–1,000 nucleotides [1]. If we assume that one in every 500 bases is variable, then two unrelated individuals will differ at several million sites within the genome [2]. It has been estimated that approximately 95% of this polymorphism will have little or no impact on the function of genes involved, and therefore only 5% could potentially be deleterious, and manifest as a disease [1,2]. Currently, several thousand DNA polymorphisms have been identified in the human genome [3], yet this is only a small fraction of the variation that potentially exists [1]. Among the DNA polymorphisms that have been identified thus far, about 25% reside within well-characterized genes, whereas the remainder are found in anonymous sequences without known function [1,3].

Variability in the human genome can exist in the form of single nucleotide substitutions (simple polymorphisms), or can include several basepairs resulting in DNA insertions or deletions (complex polymorphisms) [4]. In either case, genetic variation is detectable as inherited patterns of DNA fragments that differ between individuals. The identification of polymorphic markers that can be utilized for genetic linkage analysis has resulted in spectacular success in mapping many human genetic diseases.

SIMPLE POLYMORPHISMS

Single base substitutions have frequently been detected on Southern blot analysis of genomic DNA upon restriction enzyme digestion of DNA followed by hybridization with a DNA probe, which reveals the altered pattern. Variation within the nucleotides of an enzyme-recognition sequence can result in the deletion or creation of a cleavage site. This type of single base substitution is usually recognized as a simple two-allele system in which the fragment sizes are additive and can be attributed to the loss or gain of a restriction site (Fig 1*A*). In addition to Southern analysis, this type of polymorphism can be detected by PCR amplification of the corresponding genomic region, followed by cleavage of the PCR amplimer with the restriction endonuclease (Fig 1*B*).

In many cases, however, single basepair substitutions do not alter the recognition sites of restriction endonucleases. Instead, these variations can be detected by using a variety of techniques of molecular biology, including single-strand conformation polymorphism (SSCP) analysis, denaturing gradient gel electrophoresis (DGGE), RNase A or chemical cleavage at mismatched bases, allele-specific oligomer (ASO) hybridizations, or simply direct comparative microsequence analysis of DNA [1].

The major limitation of single basepair polymorphisms used in linkage studies is that most of them are simply dimorphic, ie, two-allele systems, which allow the assignment of only one type of heterozygous genotype (+/−) to an individual. The heterozygosity of such a system, by virtue of the allele frequencies, can never exceed 50%, and frequently is much lower [5]. The polymorphism information content (PIC) is a function of the heterozygosity of the locus, and is an indication of the usefulness of a particular marker within families subjected to genetic linkage studies. A polymorphism must have a PIC of >0.5 (range 0–1.0) to be considered highly informative. The PIC of simple dimorphic systems can never exceed 0.375, and they are therefore of somewhat limited value in linkage studies, because such RFLPs will often be uninformative in pedigrees in which many individuals are homozygous [6].

Manuscript received July 22, 1992; accepted for publication August 3, 1992.

Reprint requests to: Dr. Jouni Uitto, Jefferson Medical College, Department of Dermatology, 233 South 10th Street, Room 450, Philadelphia, PA 19107.

Abbreviations:
ASO: allele specific oligomer
DGGE: denaturing gradient gel electrophoresis
EB: epidermolysis bullosa
EHK: epidermolytic hyperkeratosis
LOD score: logarithm of the odds
PIC: polymorphism information content
RFLP: restriction fragment length polymorphism
SSCP: single-strand conformation polymorphism
VNTR: variable number of tandem repeats

0022-202X/92/$05.00

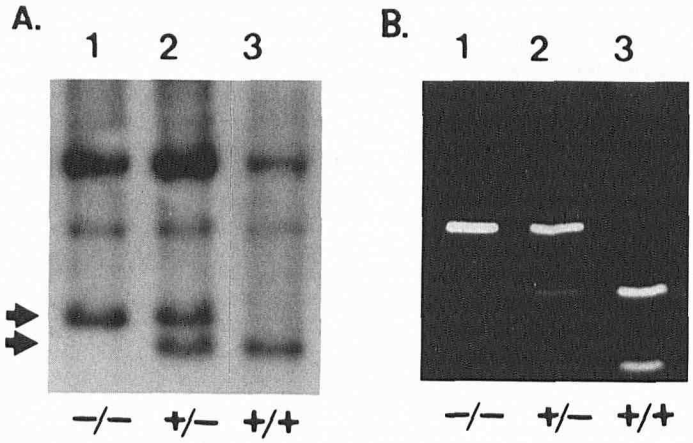

Figure 1. Representative examples of simple polymorphisms detectable by Southern blotting (*A*) or by PCR amplication (*B*). *A*: Southern hybridization of DNA, after digestion with the restriction enzyme *Hind*III, using an anonymous DNA probe (D3F15S2) on chormosome 3 reveals two variable fragments (*arrow*) that allow assignment of the genotype with respect to the presence (+) or absence (−) of the restriction enzyme site. *B*: PCR amplification of a genomic segment containing an *Alu*I in COL7A1. The bands represent PCR amplimers after digestion with the endonuclease *Alu*I. The presence of the restriction site results in generation of two fragments (*lane 3*, +/+ genotype), whereas the absence of the site preserves the original amplimer (*lane 1*, −/−). In a heterozygous individual (*lane 2*, +/−) all three bands are evident.

COMPLEX POLYMORPHISMS

Genetic linkage studies involving human diseases have recently been advanced by the availability of polymorphic markers contained within hypervariable regions of DNA. Complex polymorphisms frequently display multiallelic variation and correspondingly higher heterozygosities and PIC values than single polymorphisms [5]. Insertion or deletion of tracts of DNA gives rise to this type of variation, and the most useful types of RFLP result from a variable number of tandem repeat (VNTR) sequences [7]. The variable region is comprised of tandem repeats of a short sequence (a minisatellite), and the multiallelic polymorphism arises from differences in the number of such repeats. Alleles at VNTR loci can differ by as little as two basepairs or can vary by several hundred nucleotides. Such VNTRs are usually detected by electrophoresis on DNA sequencing gels following amplification of the repeat sequence by PCR.

Originally, the generation of multiallelic polymorphisms at VNTRs was thought to arise from unequal basepairing or "illegitimate" recombination events due to the highly repetitive nature of the loci. However, more recent studies suggest that new alleles are generated by polymerase slippage, sister chromatid exchange, or loopout deletions during meiosis [1,8].

Hypervariable regions of the DNA are thought to be randomly distributed throughout the genome, though some tend to cluster at chromosome telomeres that are thought to be unstable [1]. For example, hypervariable regions such as the CA-dinucleotide repeats are present every 30–60 kb in the human genome, amounting to approximately 50,000–100,000 copies. In addition, a number of mono-, di-, tri-, and tetranucleotide repeats have been demonstrated to be polymorphic [1]. In some cases, repeated trinucleotide sequences located within disease genes undergo expansion to hundreds or even thousands of copies in affected individuals. For example, in the genes for fragile X syndrome and myotonic dystrophy, the length of the amplified trinucleotide repeat corresponds with the severity of the phenotype [9]. The discovery of several examples of heritable unstable DNA sequences within the past year suggests that this phenomenon may be involved in the development of other genetic diseases as well.

The most variable minisatellite regions can have mutation rates as high as 5% over one generation. In general, the most variable VNTRs have a high repeat number and a long repeat sequence. Consequently, the greater the likelihood of polymorphism, the higher the heterozygosity and the PIC value for a given locus. VNTR polymorphisms associated with hypervariable minisatellite regions of the genome can have heterozygosities of 99%, and PIC values approaching 1.0. These polymorphic systems often display a large number of DNA fragments of variable sizes, thereby increasing the efficiency of linkage analyses due to the strong likelihood that two individuals will be heterozygous at the VNTR marker locus [10]. This feature also allows their use in genetic "fingerprinting" and forensic studies [7].

GENETIC LINKAGE ANALYSIS

Many heritable diseases involve extensive heterogeneity at the molecular level. It is not a trivial task to identify the exact molecular defect in a particular family; however, knowledge of the precise mutation is not a prerequisite for diagnosis by molecular analysis, for example in prenatal testing [4]. More specifically, the diagnosis can be inferred from the segregation of a disease phenotype with a defective gene through genetic linkage. The basic premise of genetic linkage is the assumption that a polymorphic DNA marker and a disease gene that are close together on the same chromosome are less likely to be separated by recombinational events than loci that are further apart [8]. Thus, co-segregation of a marker with the disease in every affected member of a family indicates that the two loci are "linked" [10]. The evidence in favor of linkage is expressed as a LOD score (Z), which reflects the strength of the statistical likelihood that a defective allele harbors the mutation causing the disease phenotype [11]. The LOD score (i.e., logarithm of the odds in favor of linkage) is calculated using sophisticated computer programs (such as LIPED), and LOD scores for separate families with the same disease can be summed together [11]. The cumulative LOD score is considered to establish linkage when $Z > 3$, or the odds in favor of linkage are 10^3, or 1000 to 1. Alternatively, linkage between the marker locus and the disease can be excluded when $Z < -2$ [8,11].

The one requirement for successful linkage analysis is the ability to distinguish the abnormal gene from the normal allele and trace its inheritance through affected families [12]. Recognition of two different alleles is accomplished by delineating polymorphisms within or near the gene that can differentiate between the alleles inherited from either parent. In many cases, the DNA polymorphism will not be the actual mutation that causes the disease, but it simply serves to mark the normal and abnormal alleles on the two chromosomes [10]. Linkage analyses have several limitations, including the correct statement of paternity in pedigrees, the informativeness of the parents of affected individuals for the polymorphic markers, the availability of extended families for testing, and, more importantly, the ability to detect polymorphisms using the markers of interest [4]. Nevertheless, polymorphic markers have provided a powerful tool for research of human genetic diseases through linkage analysis, either using the candidate gene approach when the gene is suspected, or by positional cloning when the gene is entirely unknown.

THE CANDIDATE GENE APPROACH

Linkage analysis using the candidate gene approach requires some pre-existing knowledge of the basic biochemical defect in the pathoetiology of the disease of interest. For example, insight into the abnormal protein may come in the form of ultrastructurally abnor-

Table I. Representative Examples of Heritable Diseases Affecting the Skin That Have Successfully Been Mapped by Genetic Linkage Analyses

Disease	Map Location	Candidate Gene(s)
Epidermolysis bullosa		
Simplex	12q	Keratin 5 (KRT5)
Simplex	17q	Keratin 14 (KRT14)
Dominant dystrophic	3p21	Type VII collagen (COL7A1)
Recessive dystrophic	3p21	Type VII collagen (COL7A1)
Epidermolytic hyperkeratosis	12q	Keratin 1 (KRT1)
X-linked ichthyosis	Xp	Steroid sulfatase (SST)
Marfan syndrome	15q11	Fibrillin-15 (FIB1)
Congenital contractural arachnodactyly	5q23	Fibrillin-5 (FIB2)

mal tissues in which the disease manifests, from biochemical understanding of the normal protein product, or through immunologic evidence for a defective protein by the use of antibodies [13]. The candidate gene approach, or "functional cloning," takes advantage of this pre-existing knowledge of the protein to facilitate cloning of the corresponding gene. If the normal protein and partial amino acid sequences or antibodies against the protein are available, these can serve as powerful tools to clone the candidate gene. In functional cloning, linkage mapping is performed only after successful cloning of the candidate gene. Once the disease gene has been cloned and its chromosomal location established, polymorphisms within or near the gene can be used as markers for the defective gene in kindreds with the disease. This approach has proved successful in the study of many genetic diseases, including the Marfan syndrome, retinitis pigmentosa, and different forms of epidermolysis bullosa, among others [13,14].

POSITIONAL CLONING

Although the candidate gene approach has been successful in some disorders, there are many genetic diseases for which the prerequisite biochemical insight is simply not available. In these instances, an alternative approach, known as "positional cloning," can be undertaken. In positional cloning, mapping the gene is the critical first step, and function of the protein is determined only after it has been cloned. Positional cloning begins with the identification of large pedigrees in which the disease is segregating in a Mendelian fashion. Such kindreds are studied by linkage analysis using a large number, up to several hundred, polymorphic markers evenly distributed throughout the human genome, until close linkage is established between one or more markers and the disease phenotype [13]. Using fine-mapping techniques, the candidate interval can usually be narrowed to a one centimorgan (cM) sub-region of a chromosome (1 cM contains approximately 10^6 bp). The availability of chromosomal aberrations or gross rearrangements in the genome can greatly expedite the identification of the candidate gene. However, in the absence of cytogenetic rearrangements, the significant effort in positional cloning resides in identifying each transcript in the candidate interval, and searching for mutations in each gene in patients with the disease. The pattern of tissue-specific expression of the transcripts can be used to prioritize the cloning of candidate genes, but the ultimate proof will involve the sequencing of mutations in affected individuals. Despite the effort involved, the application of positional cloning has been successful in Duchenne muscular dystrophy, cystic fibrosis, and neurofibromatosis, among other diseases [13].

SUCCESSES IN MAPPING OF GENODERMATOSES

Remarkable progress has recently been made in understanding the molecular basis of several heritable diseases of the skin (Table I). In each case, the candidate gene approach was used to establish linkage between the disease and the suspected defective gene. For example, in the simplex forms of epidermolysis bullosa (EB), ultrastructural

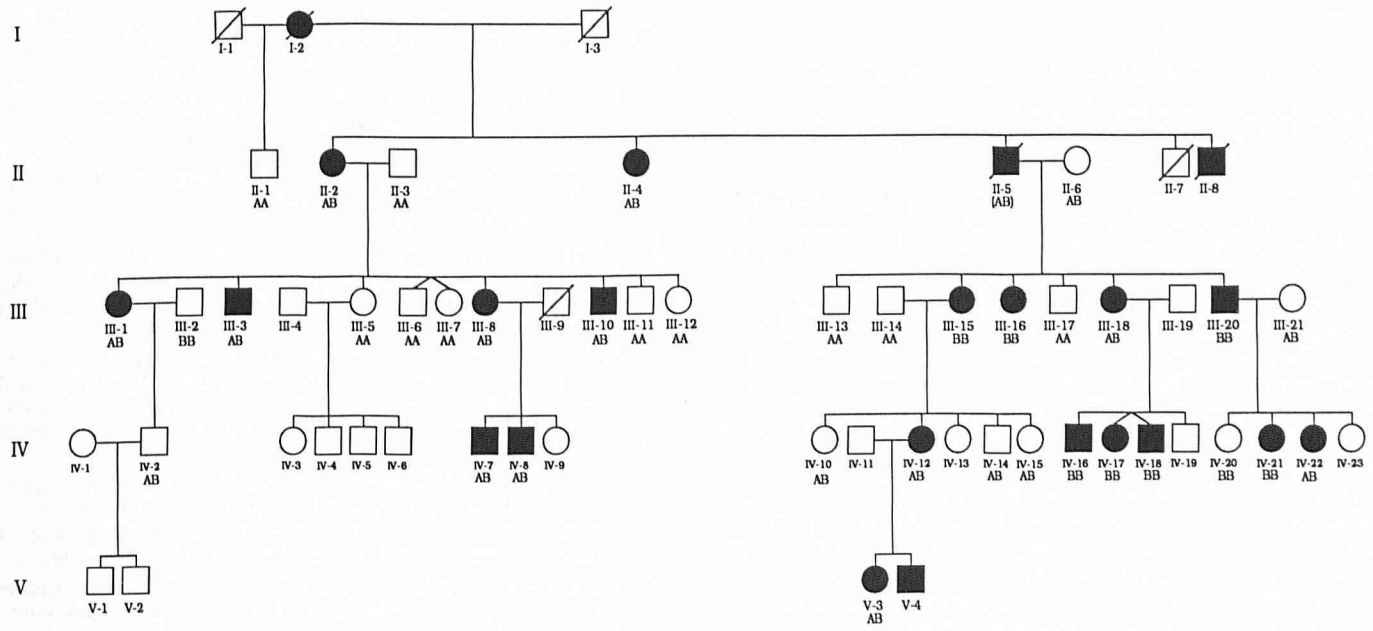

Figure 2. Co-segregation of type VII collagen alleles with the clinical phenotypes in a Finnish pedigree with dominant dystrophic EB with 20 affected living individuals in four generations. The genotype with respect to a *Pvu*II RFLP in COL7A1 is indicated under each individual symbol. In this family, each affected individual has inherited the B allele. The maximum LOD score (Z) is 6.6 with recombination fraction (θ) 0.0, indicating strong linkage. (Reprinted with permission from [26].)

observations initially revealed clumping of intracellular tonofilaments and blister formation in the region of the basal keratinocytes. For this reason, the keratins expressed in the basal keratinocytes (K5 and K14) were suspected as candidate genes. Linkage studies performed in several EB simplex families using polymorphic markers revealed close linkage of EB simplex with the keratin gene clusters either on chromosome 12 or 17, near the genes for K5 and K14, respectively [15,16]. Subsequently, mutations in K14 have been shown to result in poor keratin filament formation, leading to fragility of the basal keratinocytes [17]. In a similar approach, ultrastructural evidence suggested defects in the suprabasilar keratins in the skin of patients with epidermolytic hyperkeratosis (EHK) [18]. The candidate genes in EHK would then be the suprabasilar keratins, K1 and K10. Again, using polymorphic markers flanking the keratin gene clusters on chromosomes 12 and 17, linkage has been demonstrated between EHK and the keratin 1 gene on chromosome 12q [19,20].

In the dystrophic forms of EB, abnormalities in the anchoring fibrils of the basement membrane zone had clearly been demonstrated. In the dominant forms of dystrophic EB (DDEB), electron microscopy had shown a reduced number and/or morphologic alterations in the anchoring fibrils, and in the more severe, recessive forms of dystrophic EB (RDEB), the anchoring fibrils were virtually non-existent. Because type VII collagen is the major structural component of anchoring fibrils, it was logical to postulate that the type VII collagen gene, COL7A1, was the candidate gene in the dystrophic forms of EB [14].

Cloning and chromosomal mapping of COL7A1 to 3p21 facilitated the identification of several intragenic and flanking markers for this locus. A *Pvu*II RFLP in COL7A1 was first applied to a large Finnish family with DDEB (Fig 2). Co-segregation of the RFLP and the phenotype was observed in this family, demonstrating strong genetic linkage ($Z = 6.6$) [21]. Subsequently, linkage has been shown in several additional DDEB families, as well as in 19 kindreds with RDEB [22,23]. The high LOD scores in many families, together with the defective anchoring fibrils in affected individuals, strongly suggest that COL7A1 harbors the mutations resulting in the dystrophic forms of EB. These approaches are clearly
applicable to other heritable skin diseases, including pseudoxanthoma elasticum, cutis laxa, junctional forms of EB, and others [14,24].

IMPLICATIONS OF LINKAGE ANALYSIS

The mapping of the disease genes and the precise identification of the individual mutations in the genodermatoses have far-reaching consequences for affected patients.

The first important contribution of gene mapping is related to prenatal diagnosis of the disease in families at risk. Identification of linked RFLPs or elucidation of the precise mutation, together with the sensitivity of PCR-based detection systems, will allow prenatal diagnosis from DNA obtained from chorionic villus biopsies or periumbilical vein blood samples during the early weeks of pregnancy. These approaches, which are already being used for the prenatal diagnosis of RDEB, will eventually obliterate the use of invasive procedures, such as fetal skin biopsy, which can be performed only relatively late during gestation and carry substantial risk for the fetus.

Secondly, with the advent of gene therapy, the phenotype resulting from mutations in DNA sequences can be corrected or reversed by introduction of vectors expressing the normal gene into transplantable skin cells, such as keratinocytes. Even more directly, a novel technology to introduce DNA into the cells in situ using a biolistic microprojectile accelerator is currently being developed. Due to its accessibility and the ability to propagate dermal fibroblasts and epidermal keratinocytes in culture, the skin is an ideal target organ for these state-of-the-art gene-transfer technologies [25,26].

We thank Debbie Pawlicki for expert secretarial assistance.

This work was supported in part by NIH grants P01-AR38923 and T32-AR07561. Dr. Christiano is a Fellow of the Dermatology Foundation.

REFERENCES

1. Bowcock A, Cavalli-Sforza L: The study of variation in the human genome. Genomics 11:491–498, 1991
2. Kimura M: DNA and the neutral theory. Phil Trans R Soc Lond 312:343–354, 1986
3. Williamson R, Bowcock A, Kidd K, Pearson P, Schmidtke J, Chan HS, Chipperfield M, Cooper DN, Hewitt J, Lewitter F, Maidak B, Quitt M, Ricciuti F, Track R: Report of the DNA committee and catalogues of cloned and mapped genes and DNA polymorphisms. Cytogenet Cell Genet 55:457–778, 1990
4. Antonarakis SE: Diagnosis of genetic disorders at the DNA level. New Engl J Med 320:153–163, 1989
5. Jeffreys AJ, Wilson V, Thein SL: Hypervariable 'minisatellite' regions in human DNA. Nature 314:67–73, 1985
6. Botstein D, White RL, Skolnick M, Davis RW: Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet 32:314–331, 1980
7. Nakamura Y, Leppert M, O'Connell P, Wolff R, Holm T, Culver M, Martin C, Fujimoto E, Hoff M, Kumlin E, White R: Variable number of tandem repeat (VNTR) markers for human gene mapping. Science 235:1616–1622, 1987
8. Wolff RK, Nakamura Y, White R: Molecular characterization of a spontaneously generated new allele at a VNTR locus: no exchange of flanking DNA sequence. Genomics 3:347–351, 1988
9. Richards RI, Sutherland GR: Heritable unstable DNA sequences. Nature Genetics 1:7–9, 1992
10. Watkins PC: Restriction fragment length polymorphism (RFLP): applications in human chromosome mapping and genetic disease research. Biotechniques 6:310–320, 1988
11. Ott J: Estimation of the recombination fraction in human pedigrees: efficient computation of the likelihood for human linkage studies. Am J Hum Genet 26:588–597, 1974
12. Skolnick MH, White R: Strategies for detecting and characterizing restriction fragment length polymorphisms (RFLPs). Cytogenet Cell Genet 32:58–67, 1982
13. Collins FS: Positional cloning: let's not call it reverse anymore. Nature Genetics 1:3–6, 1992
14. Uitto J, Bauer EA, Moshell A: Symposium on epidermolysis bullosa: molecular biology and pathology of the cutaneous basement membrane zone. J Invest Dermatol 98:391–395, 1992
15. Ryynänen M, Knowlton RG, Uitto J: Mapping of epidermolysis bullosa simplex mutation to chromosome 12. Am J Human Genet 49:978–984, 1991
16. Bonifas JM, Rothman A-L, Epstein E: Epidermolysis bullosa simplex: evidence in two families for keratin gene abnormalities. Science 254:1202–1205, 1991
17. Coulombe PA, Hutton ME, Letai A, Hebert A, Paller AS, Fuchs E: Point mutations in human keratin 14 genes of epidermolysis bullosa simplex patients: genetic and junctional analyses. Cell 66:1301–1311, 1991
18. Ishida-Yamamoto A, McGrath JA, Judge MR, Leigh IM, Lane EB, Eady RAJ: Selective involvement of keratins K1 and K10 in the cytoskeletal abnormality of epidermolytic hyperkeratosis (bullous congenital ichthyosiform erythroderma). J Invest Dermatol 99:19–26, 1992
19. Compton JG, DiGiovanna JJ, Santucci SK, Kearns KS, Amos CI, Abangan DL, Korge BP, McBride OW, Steinert PM, Bale SJ: Linkage of epidermolytic hyperkeratosis to the type II keratin gene cluster on chromosome 12q. Nature Genetics 1:301–305, 1992
20. Pulkkinen L, Christiano AM, Knowlton RG, Uitto J: Epidermolytic hyperkeratosis: genetic linkage to the type II keratin gene cluster on human chromosome 12q (submitted)
21. Ryynänen M, Knowlton RG, Parents MG, Chung LC, Chu ML, Uitto J: Human type VII collagen: genetic linkage of the gene (COL7A1) on chromosome 3 to dominant dystrophic epidermolysis bullosa. Am J Human Genet 49:797–803, 1991

22. Ryynänen M, Ryynänen J, Sollberg S, Iozzo RV, Knowlton RG, Uitto J: Genetic linkage of type VII collagen (COL7A1) to dominant dystrophic epidermolysis bullosa in families with abnormal anchoring fibrils. J Clin Invest 89:974–980, 1992

23. Hovnanian A, Duquesnoy P, Blanchet-Bardon C, Knowlton RG, Anselem S, Lathrop M, Dubertret L, Uitto J, Goossens M: Genetic linkage of recessive epidermolysis bullosa to the type VII collagen gene. J Clin Invest 90:1033–1037, 1992

24. Christiano AM, Lebwohl MG, Boyd CD, Uitto J: Workshop on pseudoxanthoma elasticum: molecular biology and pathology of the elastic fibers. J Invest Dermatol 99:660–663, 1992 *(this issue)*

25. Jensen PKA, Bolund L: Tissue culture of human epidermal keratinocytes: a differentiating model system for gene testing and somatic gene therapy. J Cell Sci 100:255–259, 1991

26. Uitto J, Christiano AM: Molecular genetics of the cutaneous basement membrane zone. Perspectives on epidermolysis bullosa and other blistering skin diseases. J Clin Invest 90:687–692, 1992