Order Number 9029953

The gene encoding alpha-galactosidase A and gene rearrangements causing Fabry disease

Kornreich, Ruth, Ph.D.

City University of New York, 1990

U·M·I

300 N. Zeeb Rd.
Ann Arbor, MI 48106

17

THE GENE ENCODING ALPHA-GALACTOSIDASE A AND GENE REARRANGEMENTS CAUSING FABRY DISEASE

by

RUTH KORNREICH

A dissertation submitted to the Graduate Faculty in Biomedical Sciences in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York.

1990

This manuscript has been read and accepted by the Graduate Faculty in Biomedical Sciences in satisfaction of the dissertation requirements for the degree of Doctor of Philosophy.

4/19/90
date

Robert J. Desnick, Ph.D., M.D.
Chairman of the Examining Committee

4/20/90
date

Terry A. Krulwich, Ph.D.
Executive Officer

David F. Bishop, Ph.D.

Edward H. Schuchman, Ph.D.

James G. Wetmur, Ph.D.

Daniel Meruelo, Ph.D.

Supervisory Committee

The City University of New York

ABSTRACT

THE GENE ENCODING ALPHA-GALACTOSIDASE A AND GENE REARRANGEMENTS CAUSING FABRY DISEASE

by

Ruth Komreich

Advisors: David F. Bishop, Ph.D. and Robert J. Desnick, Ph.D., M.D.

Human α-galactosidase A (EC 3.2.1.22) is a lysosomal hydrolase encoded by a single gene localized to the chromosomal region Xq21.22-q22. The deficient activity of this enzyme results in Fabry disease, an X-linked recessive disorder which leads to premature death in affected males. To study the structure of α-galactosidase A and the molecular nature of mutations causing Fabry disease, the full-length cDNA was isolated, sequenced and used to obtain the human chromosomal gene. The 1393 base pair full-length cDNA had a 60 nucleotide 5' untranslated region and encoded a precursor peptide of 429 amino acids including a signal peptide of 31 residues. The ~12 kilobase chromosomal gene was sequenced in its entirety. The gene had seven exons whose sequences were identical to those in the full-length cDNA. All intron-exon splice junctions conformed to the GT/AG consensus sequence. The 5' flanking region of α-galactosidase A contained SP1 and TATA and CCAAT box promoter elements. In addition, sequence motifs corresponding to the AP1, "OCTA" and "core" enhancer elements were identified. Four direct repeats of the "chorion box" enhancer were preceded by an upstream "HTF" island. A unique feature of the noncoding regions of the α-galactosidase A gene was the presence of 12 *Alu* sequences. These repetitive elements accounted for about 30% of the entire gene. The sequence information was used to characterize five naturally occurring α-

galactosidase A partial gene deletions (0.4 to 4.6 kb) and one partial gene duplication (8.1 kb) which cause Fabry disease. The breakpoints of each were determined by either cloning and sequencing the mutant gene from an affected hemizygote or by DNA amplification using the polymerase chain reaction and sequencing the genomic region containing the junction. Although the α-galactosidase A gene contains numerous *Alu* elements, only a 3.2 kilobase deletion was the result of recombination between two *Alu* repeats. The remaining five rearrangements occurred between short direct repeats of 2-6 base pairs at the deletion/duplication termini of which only one copy was retained in the novel junction. Of particular interest was the finding of the tetranucleotide CCAG and the trinucleotide CAG (or their respective complements CTGG and CTG) at the 5' breakpoints in three and four of the five α-galactosidase A gene rearrangements involving short direct repeats, respectively, suggesting a possible functional role for these sequences in the recombinational event.

FOREWARD

Portions of this thesis have been presented in the following publications:

Bishop, D. F., Kornreich, R. and Desnick, R. J.: Structural organization of the human α-galactosidase A gene: Further evidence for the absence of a 3' untranslated region. *Proc. Natl. Acad. Sci.* USA **85**:3903-3907, 1988.

Bishop, D. F., Kornreich, R., Eng, C. E., Ioannou, Y. A., Fitzmaurice, T. F. and Desnick, R. J.: Human α-galactosidase A: Characterization and eukaryotic expression of the full-length cDNA and structural organization of the gene. *In*: Lipid Storage Disorders: Biological and Medical Aspects, Salvayre, R., Douste-Blazy, L. and Gatt, S. Plenum Press, pp. 503-509, 1988.

Kornreich, R., Desnick, R. J. and Bishop, D. F.: Nucleotide sequence of the human α-galactosidase A gene. *Nucleic Acids Res.* **17**:3301-3302, 1989.

Kornreich, R., Bishop, D. F. and Desnick, R. J.: The gene encoding α-galactosidase A and gene rearrangements in Fabry disease. *In*: Transactions of the American Association of Physicians **102**:30-43, 1989.

Kornreich, R., Desnick, R. J. and Bishop, D. F.: α-Galactosidase gene rearrangements causing Fabry disease: Identification of short direct repeats at breakpoints in an *Alu*-rich gene. *J. Biol. Chem.* (accepted, March 1990).

ACKNOWLEDGEMENTS

I wish to express my sincere gratitude to Dr. Robert J. Desnick and Dr. David F. Bishop for the opportunity to complete my doctoral training under their supervision. My warmest thanks are extended to Dr. David Bishop not only for his guidance, patience and many hours spent analyzing data, but also for setting high standards for research which I hope I can emulate. It is with respect that I acknowledge Dr. Desnick for creating a strong academic environment and whose unique wisdom and love of science have truly been an inspiration to me. I am also grateful to the other members of my examination committees, especially Dr. Gregory Grabowski and Dr. James Wetmur, for providing invaluable suggestions and criticism throughout the course of my research.

I am indebted to Dr. Christine Eng, Thomas Fitzmaurice, Yiannis Ioannou, Richard Gotlib, Heidi Giordano and Minxu Lu for their assistance in sequencing the α-galactosidase A chromosomal gene. Special thanks are extended to Yiannis Ioannou for his helpful scientific suggestions and assistantship. I would also like to acknowledge the groundwork for this work that was set by Drs. Harold S. Bernstein and Kenneth Astrin who identified the rearrangements used in this study.

Finally, I wish to thank my family, especially my husband Lewis, for their never ending love and support. Without their patience, encouragement and confidence in me, the completion of this work would not have been possible- you have enabled me to achieve this goal!

ABBREVIATIONS

α-galactosidase A	α-Gal A
base pair(s)	bp
complementary DNA	cDNA
counts per minute	cpm
deoxyribonucleic acid	DNA
hemoglobin	Hb
hour(s)	h
kilobase pairs	kb
micrograms	μg
nanograms	ng
polymerase chain reaction	PCR
restriction fragment length polymorphism	RFLP
ribonucleic acid	RNA

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Page

INTRODUCTION

More than 2000 different human Mendelian disorders are known, however, almost all of these mutations remained elusive to analysis until recombinant DNA technology permitted their characterization. During the past decade, the application of these techniques has led to an understanding of normal gene structure and organization, as well as the delineation of the nature of the specific defects underlying various inherited disorders. Application of this information has permitted improved diagnosis and heterozygote detection. DNA probes for many important human genes are currently being used for these purposes in such diseases as hemophilia A (1) and B (2), phenylketonuria (3), α (4) and β thalessemias (5), sickle cell anemia (6), Duchenne muscular dystrophy (7) and cystic fibrosis (8).

A group of genetic diseases known as lysosomal storage disorders result from inborn errors of metabolism. Each of the approximately 30 known inherited disorders result from the deficient activity of one or more specific lysosomal enzymes. Most of the defective enzymes are glycosidases and result in the accumulation of complex carbohydrates and lipids in lysosomes. In many of these disorders there is significant variability in the age at onset and severity of the disease manifestations (9). Although much progress has been made in the elucidation of the biochemistry and enzymology of these disorders, effective treatments have not been developed. However, recent advances in recombinant DNA technology may lead to future therapeutic methodologies. For example, large amounts of enzyme may be produced in various expression systems for therapeutic evaluation. Enzyme replacement therapy for lysosomal storage diseases may be feasible because preliminary studies demonstrated that low levels of the appropriate exogenous normal enzyme were able to correct the metabolic defect in cultured fibroblasts from patients with these disorders (10-12). Until recently, research was hindered due to the lack of sufficient quantities of lysosomal enzymes needed for these studies.

The overall objective of the following investigations was the characterization of the gene encoding α-galactosidase A. This gene is of interest because it encodes a human lysosomal hydrolase and is a representative of this class of enzymes. In addition, mutations in this gene cause Fabry disease. Specifically, molecular lesions which cause Fabry disease due to gene rearrangements in the α-galactosidase A locus were studied.

Biochemistry of Lysosomal Hydrolases— Lysosomal enzymes, like secretory and membrane proteins, are synthesized on polysomes bound to the endoplasmic reticulum. These proteins share a common signal sequence (prepeptide) of 15-30 predominantly hydrophobic amino acids at the amino-terminus that directs the ribosomes involved in their synthesis to the endoplasmic reticulum and participates in the translocation of the protein across its membrane (13). A complex is formed with an 11S cytoplasmic ribonucleoprotein called the signal recognition particle (SRP) (14). This complex then binds to the SRP receptor, also known as the docking protein, and the nascent polypeptide passes through to the lumen of the rough endoplasmic reticulum by a poorly understood mechanism (13). During their passage through the endoplasmic reticulum, several modifications of the lysosomal enzymes occur including the cleavage of the signal peptide and cotranslational glycosylation. Oligosaccharides are transferred as a group through dolichol carbohydrate intermediates to certain asparagine residues in the polypeptide chain (15). N-linked glycosylation occurs only with asparagine residues in the sequence asparagine-X-(threonine/serine). The oligosaccharides undergo sequential modifications including the removal of the glucose residues yielding high-mannose type oligosaccharide moieties (15).

The newly synthesized proteins are then transported to the Golgi apparatus where the secretory, plasma membrane and lysosomal proteins are eventually targeted to their proper post-Golgi destinations. Additional carbohydrate alterations can occur to form the complex oligosaccharide chains on these enzymes. The majority of lysosomal enzymes are

selectively phosphorylated in the 6-hydroxy position of one or more mannose residues of the N-linked oligosaccharide chains by a two step process. First, N-acetylglucosamine-1-phosphate is transferred to the C-6 position of mannose residues (16); followed by the removal of N-acetylglucosamine by α-N-acetylglucosaminyl phosphodiesterase (17, 18). The resulting mannose-6-phosphate residues serve as signals for receptor-mediated targeting to the lysosome (19). This pathway, however, is not the only route for enzymes to reach the lysosome. This is evidenced by the fact that patients with I cell disease, who are deficient in phosphotransferase and are thus unable to synthesize the mannose-6-phosphate recognition signal, can still target residual amounts of enzyme to lysosomes (20). In addition to the above modifications, some lysosomal enzymes undergo additional proteolytic processing at either the amino and/or the carboxy terminus involving the removal of propeptide sequences (21). The precise cellular locations where the processing occurs is not yet known.

Fabry Disease: α-Galactosidase A Deficiency— Fabry disease is an X-linked recessive inborn error of catabolism which results from the deficient activity of the lysosomal hydrolase, α-galactosidase (α-Gal A, α-D-galactoside galactohydrolase, EC 3.2.1.22) (22); an enzyme that hydrolyzes the terminal α-galactosyl residue of globotriaosylceramide and other substrates with terminal α-galactosyl moieties. In affected males, accumulation of neutral glycosphingolipids, predominantly globotriaosylceramide, occurs in most tissues and fluids of the body with a predilection for lysosomes of the vascular endothelium (23). This produces the clinical manifestations of the disease including angiokeratoma, acroparesthesias, episodic crises of excruciating pain, corneal and lenticular opacities, hypohydrosis, and cardiac and renal dysfunction (24). Death usually occurs in the third to fifth decades of life from renal, cardiac, and/or cerebral complications of vascular disease. Heterozygous females are usually clinically asymptomatic or only mildly symptomatic and live a normal life span (24).

The clinical diagnosis of affected males can be confirmed by the demonstration of deficient α-Gal A activity in various sources such as plasma, leukocytes, tears and cultured cells. Classically affected males have no detectable α-Gal A activity (25-27). Biochemical identification of female carriers of Fabry disease is not always possible. Due to random X-chromosomal inactivation (28), heterozygotes can express levels of enzymatic activity theoretically ranging from zero to normal values. The determination of carrier status has been demonstrated in normal and mutant cell populations with α-Gal A assays of single hair roots (29-31) or individual cloned fibroblasts (32). However, these studies are time-consuming as well as extremely difficult to perform and interpret (33). A more reliable method for heterozygote determination needs to be established. Molecular diagnosis using probes from the α-Gal A gene or closely linked regions will permit more accurate diagnosis (34).

Fabry disease and α-Gal A represent an excellent model system to investigate the effects of various mutations which alter the expression of a lysosomal hydrolase and to study the structure and organization of a human lysosomal enzyme due to the following unique features: 1) The active enzyme is a homodimer (35), therefore, only a single gene needs to be isolated; 2) Of the glycolipid storage disorders, only Fabry disease is transmitted by an X-linked gene; when studying mutations of Fabry hemizygotes, only one allele has to be isolated; 3) Cell lines from over 140 unrelated Fabry families are available (in our laboratories) permitting the characterization of the variety of molecular lesions which cause this lysosomal storage disorder. Furthermore, a detailed knowledge of the genomic structure of the α-Gal A gene and the availability of cloned genomic sequences are critical for the characterization of the molecular defects underlying Fabry disease.

The availability of the fine genomic structure of α-Gal A and other lysosomal hydrolases may provide insight into the structure-function relationships of this class of enzymes. Common structural or functional domains may be revealed. To date, the isolation and characterization of only a limited number of lysosomal genes have been

reported (36-39). In addition, the current application of molecular biological technology to the study of lysosomal enzymes is expected to provide a more thorough understanding of the synthesis, maturation and transport processes of proteins synthesized in the rough endoplasmic reticulum. Such studies may provide fundamental knowledge of the normal physiology of lysosomal enzymes and their genes analogous to the insights provided by the investigation of type II familial hypercholesterolemia and the low density lipoprotein receptor (40).

To facilitate studies on the characterization of the molecular pathology underlying Fabry disease, provide probes for heterozygote detection and express large amounts of enzyme, a cDNA clone encoding α-Gal A (designated λAG18) was isolated from a λgt11 human liver expression library (41). This clone was identified using monospecific polyclonal antibodies and synthetic oligonucleotide mixtures corresponding to the amino-terminus and internal tryptic and cyanogen bromide peptide sequences. The nucleotide sequence of the 1234 bp λAG18 insert (Fig. 1) was determined by dideoxy chain termination sequencing of generated M13 deletion subclones (42). The nucleotide sequence contained an open reading frame of 1226 nt encoding the entire 398 amino acids of the mature form of α-Gal A and the last 5 amino acids of the prepeptide sequence (42). The authenticity of this clone was demonstrated by its colinearity with 86 non-overlapping microsequenced amino acids from the N-terminus and internal peptides, X-chromosomal dosage, gene mapping using mouse-human somatic cell hybrid panels and localization to the region Xq21.33-Xq22 by *in situ* hybridization (43) and RFLP linkage studies (44).

Poly(A)$^+$ RNA transfer hybridization experiments demonstrated the message encoding the entire α-Gal A subunit was 1.45 kb (42). Immunologic studies of human α-Gal A biosynthesis have shown that a 50.5 kilodalton glycosylated precursor undergoes proteolytic processing to a mature 46 kilodalton form (45, 46). Native α-Gal A has a molecular mass of 100 kilodaltons of which 7-15 percent is due to carbohydrate and is composed of two identical subunits (45-47). The remaining sequences in the α-Gal A

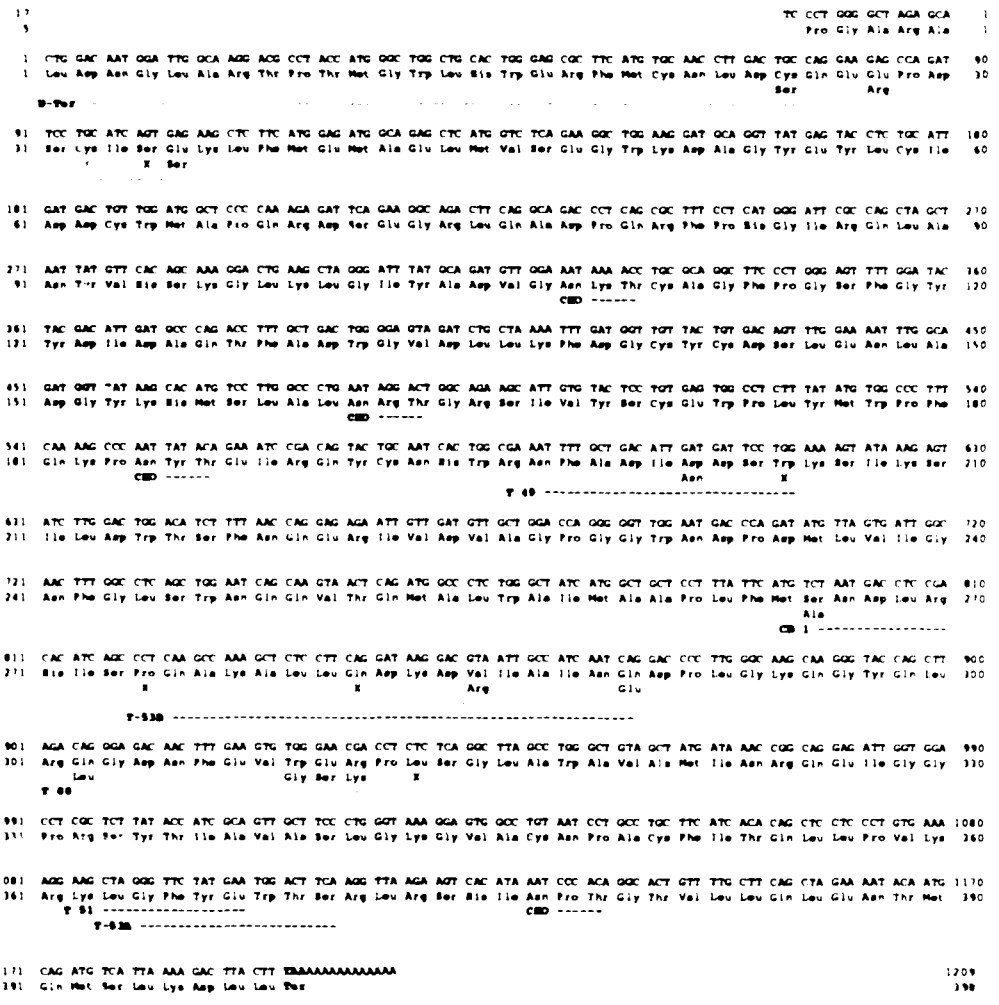


Figure 1. Nucleotide and predicted amino acid sequences of the λAG18 insert encoding the human mature α-Gal A subunit. Amino acid 1 is assigned to the amino-terminal residue. Nucleotides -17 to -1 encode five amino acids of the prepeptide. Bold underlines indicate confirmed amino acid sequence obtained by microsequencing of the N-terminus and tryptic (T) and cyanogen bromide (CB) peptides. Differences between microsequenced and predicted amino acids are shown; X denotes unidentified amino acids. Note that peptide T-53B corrects two errors in the sequence of CB-1. CHO indicates potential sites of N-glycosylation and the 3' termination signals are overlined.

mRNA must encode the prepeptide (signal peptide), 5' untranslated sequences and the poly(A) tail.

Two consensus sequences for cleavage at the polyadenylation site, ATTAAA and AATACA, were present 12 and 28 nt upstream from the stop codon, respectively in the cDNA (42). The former is present in approximately 12 percent of vertebrate messages and the latter occurs only in 2 percent (48). An unusual feature of the α-Gal A mRNA was the absence of a 3' untranslated region between the termination codon and the poly(A) tail. This is unique among non-mitochondrial genes. The only other known mammalian example of a nuclear encoded mRNA lacking a 3' untranslated region is that for murine thymidylate synthase (the human thymidylate synthase transcript did have a 3' untranslated region) (49).

Study of Gene Rearrangements— Southern hybridization analysis of the α-Gal A gene in families with Fabry disease have been conducted to characterize the molecular lesions underlying the enzymatic defect. Several gene rearrangements have been detected in the α-Gal A gene including five partial deletions ranging in size from 0.4 to 4.8 kb and a partial gene duplication of ~8 kb (50). These naturally occurring germ line mutations provide a unique system to study genetic recombination of higher eukaryotes. By characterizing the sequences found at the rearrangement termini, information about the events underlying the recombinational events can be obtained.

Studies that define the breakpoints of gene rearrangements have revealed the frequent presence of repetitive sequences at rearrangement termini, especially in genes enriched with such sequences. Repetitive DNA comprises approximately 20 to 30 percent of human DNA (51). A portion of this DNA consists of long and short sequences that are interspersed among unique regions. The number of different families of interspersed repeats is difficult to determine (52). Renaturation rate studies of denatured human DNA indicate that the number of dispersed sequence families may be few (52). The most

abundant interspersed repetitive sequence family is the *Alu* family. This family comprises 3-6 percent of the human genome (52). Assuming these sequences are randomly distributed, an *Alu* family member should be found every 4 kb (54). This spacing has been found to be variable. Eight *Alu* family members have been found in the ~60 kb β-globin gene cluster, however, some are only 700-800 bp apart (55). Also, the 19 kb human insulin gene has only one *Alu* sequence while the 12 kb c-sis gene has three (55).

Alu family members share a related 300 nt DNA sequence composed of an imperfect tandemly repeated 130 bp sequence flanked by short direct repeats 8-20 bp in length and adenine rich terminal and internal regions (56). The function of these interspersed repeats has been purely speculative. Some investigators believe they are important in the regulation of gene activity, while others believe they perform neither a useful nor harmful function. *Alu* DNA is partially homologous to 7SL RNA which is a component of the signal recognition particle that mediates protein secretion across the endoplasmic reticulum (52). Both are transcribed by RNA polymerase III. For *Alu* sequences, the transcription initiation site is close to the first nucleotide. Initiation has been found to occur only within the first arm even though the two arms are similar in sequence (52). Although *Alu* family members account for the majority of interspersed repeats in humans, members of related families of long repeats, some over 6 kb in length, occur an estimated 100,000 times within the human genome (57). Long interspersed repeat sequences, LINES, are a multigene family (known as the *Kpn*I family in primates) composed of one or more functional genes and a large number of psuedogenes (58). The number of functional genes and when and where they are expressed is still unknown.

Due to their frequency and sequence homology, the interspersed repetitive sequences may serve as hot spots for illegitimate recombination by unequal crossing over events between repeats during meiosis. There may also be an inherent instability in DNA containing repetitive sequences that predisposes these regions to recombination. For example, *Alu* sequences occur at the breakpoints of most of the characterized low density

lipoprotein receptor gene rearrangements which cause familial hypercholesterolemia. This gene is unusually rich in these repeat elements containing ~1 *Alu* per 1.8 kb (21 *Alu* monomers or dimers identified to date in 38 kb, H.H. Hobbs, pers. commun.). More specifically, six of the eight rearrangements characterized at the DNA sequence level had both breakpoints in an *Alu* sequence and a seventh had one breakpoint in an *Alu* repeat (59-66). In four cases, the *Alu* sequences are found in the same orientation (59, 61, 63, 64). It is probable that unequal crossing over events took place during meiosis after *Alu* sequences on different chromatids mispaired. In two other cases, the *Alu* family members were oriented in opposite directions (60, 66). It is thought that recombination occurred between the two repetitive sequences on a single strand of DNA. A stem and loop structure that would bring the ends of the deletion into close proximity may be hypothesized. Following cleavage of the loop, the observed deletions would be produced. There are also several examples of rearrangements in different genes that appear to have arisen from unequal crossing over between similarly oriented *Alu* repeat sequences (67-70). Yet another reported example of a deleterious event resulting from recombination between *Alu* sequences is a chromosome rearrangement producing an XX male (71). In this case described by Rouyer et al. (71), the XX male presumably arose from an abnormal cross-over between regions on the X and Y chromosomes. The breakpoints were mapped within *Alu* repeats and there was no sequence homology between the normal X and Y regions outside the *Alu* elements. This suggests, along with the fact that the recombined *Alu* sequence can be aligned with normal counterparts, that the rearrangement is the product of recombination between the repetitive sequences.

In contrast, *Alu* repetitive elements do not appear to play a significant role in the generation of deletions in the human ß-globin gene cluster which spans ~60 kb of chromosome 11 and has only 8 *Alu* elements (55). Over 20 deletions in the ß-globin gene cluster have been characterized at the DNA sequence level and no rearrangements were due to *Alu-Alu* recombination and only three breakpoints occurred within an *Alu* sequence (for

review, see 72). Since germinal rearrangements in only a few human genes have been intensively investigated, the correlation between the frequency of *Alu* sequences in a given gene and their occurrence at rearrangement termini cannot be precisely determined.

Another frequent finding at germinal gene rearrangement termini is the presence short direct repeats of 2-8 nucleotides. The material between the repeat sequences is deleted (or duplicated) and the resulting chromosome contains a single copy of the original repeat (73). Several models have been proposed to define the role of short stretches of homology in the generation of rearrangements including: 1) errors of DNA replication by the "slipped mispairing" of the direct repeats; 2) unequal crossing over between the short regions of homology on different chromatids; 3) intrachromosomal excision of the region between aligned direct repeats releasing a circular DNA fragment, and 4) improper rejoining of DNA after double-stranded breakage events (for review, see 74). Replication errors due to slippage of the single-stranded DNA template appears to be the favored model. "Slipped mispairing" of direct repeats during DNA replication was originally proposed by Streisinger et al. to explain the generation of frameshift mutations (75). As the replicative enzyme complex moves through the region containing the direct repeats, the region will become single-stranded and one of the repeats could base-pair with the complementary downstream repeat sequence. This event would produce a single-stranded loop containing one of the repeats and the intervening sequence between the repeats. This loop could then be removed by DNA repair enzymes or would be removed following another round of replication. If the DNA "slips" after replicating the second repeat, a duplication would be produced after another round of DNA replication. Analyses of gene rearrangements in various systems have suggested that the sequence context surrounding the direct repeats was important in their formation (i.e., 76, 77). For example, in the clustered deletions of the *E. coli lac* I gene, a potential secondary structure intermediate was found between the short direct repeats at the breakpoints which would bring the termini into close proximity, thereby allowing their interaction (77). In addition, seven of

the ten reported *aprt* somatic gene rearrangements in cultured hamster cells had short direct repeats at their breakpoints which clustered in a 40 bp region that contained inverted repeats capable of forming stable secondary structures (76). These structures may act as substrates for enzymes that resolve the structural intermediates. Therefore, as evidenced by studying rearrangement breakpoints in various genes, much can be learned about the nature of DNA which predisposes it to illegitimate recombination.

REFERENCES

1. Wood, W.I., Capon, D.J., Simonsen, C.C., Eaton, D.L., Gitschier, J., Keyt, B., Seeburg, P.H., Smith, D.H., Hollingshead, P., Wion, K.L., Delwart, E., Tuddenham, G.D., Vehar, G.A. and Lawn, R.M. (1984) *Nature* **312**, 330-337

2. Choo, K.H., Gould, K.G. Rees, D.J.G. and Brownlee, G.G. (1982) *Nature* **299**, 178-180

3. Woo, S.L.C., Lidsky, A.S., Guttler, F., Chandra, T. and Robson, K.J.H. (1983) *Nature* **306**, 151-155

4. Orkin, S.H., Old, J., Lazarus, H., Altay, C., Gurgey, A., Weatherall, D.J. and Nathan, D.G. (1979) *Cell* **17**, 33-42

5. Orkin, S.H., Kazazian, H.H., Antonarakis, S.E., Goff, S.C., Boehm, C.D., Sexton, J.P., Waber, P.G. and Giardina, P.J.V. (1982) *Nature* **296**, 627-630

6. Geever, R.F., Wilson, L.B., Nallaseth, F.S., Milner, P.F., Bittner, M. and Wilson, J.T. (1981) *Proc. Natl. Acad. Sci.* USA **78**, 5081-5085

7. Chamberlain, J.S., Gibbs, R.A., Ranier, J.E., Nguyen, P.N. and Caskey, C. (1988) *Nucleic Acid Res.* **16**, 11141-11156

8. Beaudet, A.L., Feldman, G.L., Fernbach, S.D., Buffone, G.J. and O'Brian, W.E. (1989) *Am. J. Hum. Genet.* **44**, 319-326

9. Scriver, C.R., Beaudet, A.L., Sly, W.S., and Valle, D., eds. (1989) *The Metabolic Basis of Inherited Disease* McGraw-Hill, New York, 6th ed.

10. Porter, M.T., Fluharty, A.L. and Kihara, H. (1971) *Science* **172**, 1263-1265

11. Hickman, S. and Neufeld, E.F. (1972) *Biochem. Biophys. Res. Commun.* **49**, 992-996

12. Dawson, G., Matalon, R. and Ki, Y.T. (1973) *Pediat. Res.* **7**, 690-694

13. Walter, P., Gilmore, R. and Blobel, G. (1984) *Cell* **38**, 5-8

14. Walter, P. and Blobel, G. (1982) *Nature* **299**, 691-98

15. Kornfeld, R. and Kornfeld, S. (1985) *Ann. Rev. Biochem.* **54**, 631-634

16. Reitman, M.L. and Kornfeld, S. (1981) *J. Biol. Chem.* **256**, 11977-11980

17. Varki, A. and Kornfeld, S. (1980) *J. Biol. Chem.* **255**, 8398-8401

18. Waheed, A., Hasilik, A. and von Figura, K. (1981) *J. Biol. Chem.* **256**, 5717-5721

19. Sly, W.S. and Fischer, H.D. (1982) *J. Cell. Biochem.* **18**, 67-85

20. Miller, A.L., Kress, B.C., Stein, R., Kinnon, C. and Kern, H. (1981) *J. Biol. Chem.* **256**, 9352-9362

21. Erikson, A.H., Conner, G.E. and Blobel, G. (1981) *J. Biol. Chem.* **256**, 11224-11231

22. Brady, R.O., Gal A.E., Bradley, R.M., Martensson, E., Warshaw, A.L. and Laster, L. (1967) *New Engl. J. Med.* **276**, 1163-1166

23. Johnson, D.L. and Desnick, R.J. (1978) *Biochem. Biophys. Acta* **538**, 195-204

24. Desnick, R.J. and Bishop, D.F. (1989) in *The Metabolic Basis of Inherited Disease* (Scriver, C.R., Beaudet, A.L., Sly, W.S., and Valle, D., eds.) pp.1751-1796, McGraw-Hill, New York, 6th ed.

25. Desnick, R.J., Allen, K.Y., Desnick, S.J., Raman, M.K., Berneohr, R.W. and Krivit, W. (1973) *J. Lab. Clin. Med.* **81**, 157-171

26. Johnson, D.L., Del Monte, M.A., Cotlier, E. and Desnick, R.J. (1975) *Clin. Chim. Acta* **63**, 81-90

27. Mayes, J.S., Scheerer, J.B., Sifers, R.N. and Donaldson, M.L. (1981) *Clin. Chim. Acta* **112**, 247-251

28. Lyon, M. (1961) *Nature* **190**, 372-373

29. Grimm, T., Wienker, T.F. and Ropers, H.H. (1976) *Hum. Genet.* **32**, 329-336

30. Vermorken, A.J.M., van Bennekom, C.A., deBruyn, C.H.M.M. and Oei, T.L. (1980) *Br. J. Derm.* **103**, 101-103

31. Beaudet, A.L. and Caskey, C.T. (1978) *Clin. Genet.* **13**, 251-256

32. Romeo, G. and Migeon, B.R. (1970) *Science* **170**, 180-182

33. Rietra, P.J.G.M., Brouwer-Kelder, E.M., deGroot, W.P. and Tager, J.M. (1976) *J. Mol. Med.* **1**, 237-242

34. Botstein, D., White, R.L., Skolnick, M. and Davis, R.W. (1980) *Am. J. Hum. Genet.* **32**, 314-331

35. Bishop, D.F. and Desnick, R.J. (1981) *J. Biol. Chem.* **256**, 1307-1316

36. Proia, R.L. and Soravia, E. (1987) *J. Biol. Chem.* **262**, 5677-5681

37. Proia, R.L. (1988) *Proc. Natl. Acad. Sci.* USA **85**, 1833-1887

38. Horowitz, M., Wilder, S., Horowitz, Z., Reiner, O., Gelbart, T., and Beutler, E. (1989) *Genomics* **4**, 87-96

39. Geier, C., von Figura, K. and Pohlmann, R. (1989) *Eur. J. Biochem.* **183**, 611-616

40. Brown, M.S. and Goldstein, J.L. (1986) *Science* **232**, 34-47

41. Calhoun, D.H., Bishop, D.F., Bernstein, H.S., Quinn, M., Hantzopoulos, P. and Desnick, R.J. (1985) *Proc. Natl. Acad. Sci.* USA **82**, 7364-7368

42. Bishop, D.F., Calhoun, D.H., Bernstein, H.S., Hantzopoulos, P., Quinn, M. and Desnick, R.J. (1986) *Proc. Natl. Acad. Sci.* USA **83**, 4859-4863

43. Desnick, R.J., Bernstein, H.S., Astrin, K.H., and Bishop, D.F. (1987) *Enzyme* **38**, 54-64

44. Astrin, K.H., Vlasak, I., Snir Lev-ran, Bishop, D.F., and Desnick, R.J. (1988) *Am. J. Hum. Genet.* **43**, A135

45. LeDonne, N.C., Fairly, J.L. and Sweeley, C.C. (1983) *Arch. Biochem. Biophys.* **224**, 186-195

46. Lemansky, P., Bishop, D.F., Desnick, R.J., Hasilik, A. and von Figura, K. (1987) *J. Biol. Chem.* **262**, 2062-2065

47. Kusiak, J.W., Quirk, J.M. and Brady, R.O. (1978) *J. Biol. Chem.* **253**, 184-190

48. Wickens, M. and Stephenson, P. (1984) Science **226**, 1045-1051

49. Jenh, C.H., Deng, T., Li, D., DeWille, J. and Johnson, L.F. (1986) *Proc. Natl. Acad. Sci.* USA **83**, 8482-8486

50. Bernstein, H.S., Bishop, D.F., Astrin, K.H., Kornreich, R., Eng, C.M., Sakuraba, H., and Desnick, R.J. (1989) *J. Clin. Invest.* **83**, 1390-1399

51. Schmid, C.W. and Deininger, P.L. (1975) *Cell* **6**, 345-358

52. Jelinek, W.R. and Schmid, C.W. (1982) *Ann. Rev. Biochem.* **51**, 813-844

53. Houck, C.M., Rinehart, F.P. and Schmid, C.W. (1978) *Biochim. Biophys. Acta* **518**, 37-52

54. Hwu, H.R., Roberts, J.W., Davidson, E.H. and Britten, R.J. (1986) *Proc. Natl. Acad.* Sci. USA **83**, 3875-3879

55. Schmid, C.W. and Jelinek, W.R. (1982) *Science* **216**, 1065-1070

56. Deininger, P.L., Jolly, D.J., Rubin, C.M., Friedmann, T. and Schmid, C.W. (1981) *J. Mol. Biol.* **151**, 17-33

57. Grimaldi, G., Skowronski, J. and Singer, M.F. (1984) *EMBO J.* **3**, 1753-1759

58. Singer, M.F. and Skowronski, J. (1985) *Trends Biochem. Sci.* **10**, 119-122

59. Lehrman, M.A., Goldstein, J.L., Russell, D.W., and Brown, M.S. (1987) *Cell* **48**, 827-835

60. Lehrman, M.A., Schneider, W.J., Sudhof, T.C., Brown, M.S., Goldstein, J.L., and Russell, D.W. (1985) *Science* **227**, 140-146

61. Hobbs, H.H., Brown, M.S., Goldstein, J.L., and Russell, D.W. (1986) *J. Biol. Chem.* **261**, 13114-13120

62. Lehrman, M.A., Russell, D.W., Goldstein, J.L., and Brown, M.S. (1986) *Proc. Natl. Acad. Sci.* USA **83**, 3679-3683

63. Lehrman, M.A., Russell, D.W., Goldstein, J.L., and Brown, M.S. (1987) *J. Biol. Chem.* **262**, 3354-3361

64. Horsthemke, R., Beisiegel, U., Dunning, A., Havinga, J.R., Williamson, R., and Humphries, S. (1897) *Eur. J. Biochem.* **164**, 77-81

65. Aalto-Setala, K., Helve, E., Kovanen, P.T., and Kontula, K. (1989) *J. Clin. Invest.* **84**, 499-505

66. Miyake, Y., Tajima, S., Funahashi, T., and Yamamoto, A. (1989) *J. Biol. Chem.* **264**, 16584-16590

67. Myerowitz, R., and Hogikyan, N.D. (1987) *J. Biol. Chem.* **262**, 15396-15399

68. Markert, M.L., Hutton, J.J., Wiginton, D.A., States, J.C., and Kaufman, R.E. (1988) *J. Clin. Invest.* **81**, 1323-1327

69. Nicholls, R.D., Fischel-Ghodsian, N., and Higgs, D.R. (1987) *Cell* **49**, 369-78

70. Huang, L-S., Ripps, M.E., Korman, S.H., Deckelbaum, R.J., and Breslow, J.L. (1989) *J. Biol. Chem.* **264**, 11394-11400

71. Rouyer, F., Simmler, M.C., Page, D.C. and Weissenbach, J. (1987) *Cell* **51**, 417-425

72. Collins, F.S. and Weissman, S.M. (1984) *Prog. Nucleic Acid Res.* **31**, 315-436

73. Efstratiadis, A., Posakony, J.W., Maniatis, T., Lawn, R.M., O'Connell, C., Spritz, R.A., DeRiel, J.K., Forget, B.G., Weissman, S.M., Slightom, J.L., Blechl, A.E., Smithies, O., Baralle, F.E., Shoulders, C.C., and Proudfoot, N.J. (1980) *Cell* **21**, 653-668

74. Meuth, M. (1989) in *Mobile DNA* (Berg, D.E. and Howe, M.M., eds.) pp.833-860, American Society for Microbiology, Washington, D.C.

75. Streisinger, G., Okada, Y., Emrich, J., Newton, J., Tsugita, A., Terzaghi, E., and Inouye, M. (1966) *Cold Spring Harbor Symp. Quant. Biol.* **31**, 77-84

76. Nalbantoglu, J., Hartley, D., Phear, G., Tear, G., and Meuth, M. *EMBO J.* (1986) **5**, 1199-1204

77. Albertini, A.M., Hofer, M., Calos, M.P. and Miller, J. (1982) *Cell* **29**, 319-328

Chapter One

Structural Organization of the Human α-Galactosidase A Gene

ABSTRACT

For studies of the structure and function of α-galactosidase A and for characterization of the genetic lesions in families with Fabry disease, the full-length cDNA was isolated, sequenced and used to screen human genomic libraries. The 1393 bp full-length cDNA had a 60 nt 5' untranslated region and encoded a precursor peptide of 429 amino acids including a signal peptide of 31 residues. Three overlapping lambda clones spanning 32 kb were isolated which contained the entire ~12 kb chromosomal gene as well as ~9 and ~11 kb of 5' and 3' flanking sequence, respectively. The gene had seven exons. The genomic exonic and full-length cDNA sequences were identical. All intron-exon splice junctions conformed to the GT/AG consensus sequence. The 5' flanking region of this lysosomal housekeeping gene contained SP1 and CCAAT box promoter elements as well as sequences corresponding to the AP1, "OCTA" and "core" enhancer elements. There was an upstream "HTF" island followed by four direct repeats of the "chorion box" enhancer. The unique lack of a 3' untranslated sequence in the α-galactosidase A cDNA was confirmed by sequencing additional cDNA clones and the genomic 3' region.

INTRODUCTION

This chapter describes the cloning and sequencing of a full-length human α-Gal A cDNA and its use for the isolation of the entire chromosomal gene. In addition, the characterization of this lysosomal housekeeping gene's intron/exon organization, 5' regulatory elements and the 3' flanking sequence are discussed. This work was greatly facilitated by the availability of a cDNA clone (λAG18) which encoded the entire mature lysosomal form of human α-Gal A (1, 2). The λAG18 cDNA was isolated from a λgt11 expression library using monospecific polyclonal antibodies and synthetic oligonucleotide probes (1). This cDNA, however, did not contain the entire 5' sequence and encoded only five residues of the α-Gal A signal peptide. It was notable that λAG18 and a subsequently cloned α-Gal A cDNA (2) did not have 3' untranslated regions. In these cDNAs, the poly(A) tract immediately followed the termination codon. This unusual finding was confirmed by sequencing the chromosomal gene. Additionally, to fully understand the molecular nature of α-Gal A genetic lesions causing Fabry disease, the structure of the normal gene was elucidated.

EXPERIMENTAL PROCEDURES

Materials— Restriction endonucleases, the Klenow fragment of DNA polymerase I, M13 cloning vectors and universal sequencing primers were purchased from New England Biolabs; T4 DNA ligase was from International Biotechnologies Inc.; β-cyanoethyl diisopropyl phosphoramidites and ancillary DNA synthesis reagents were obtained from Biosearch; agarose was from FMC Corp.; nitrocellulose filters (type HATF) were purchased from Millipore and Zetabind nylon transfer membranes were from AMF Cuno; [α- and γ-^{32}P] dNTPs (3000 Ci/mmole) and [α-^{35}S] dATP (1000 Ci/mmole) were obtained from Amersham. A human fibroblast cDNA library in the pcD vector (3) was

kindly provided by Dr. H. Okayama, NIH. A human lung λgt11 cDNA library (#HL1004, lot #2007) was obtained from Clontech Laboratories. The 49,XXXXY bacteriophage library (4), designated λ4X, was kindly provided by Dr. W.I. Wood, Genentech, Inc. A 49,XXXXX lymphoblast cell line (GM 6061A) was from the Human Genetic Mutant Cell Repository. The EMBL3 vector, *in vitro* packaging extracts and *E. coli* strain K802 were kindly provided by Dr. J. Gordon, Mount Sinai School of Medicine.

Isolation of Full-Length α-Gal A cDNA Clones— The unamplified pcD human fibroblast cDNA library was plated at a density of 5 x 10^4 colonies per 137 mm HATF filters on χ-broth agar plates (3). After growth for 10 h at 37 °C, replicas were regrown and transferred to chloramphenicol plates for an additional 12 h. Colony hybridization was carried out according to the method of Hanahan and Meselson (5). For use as hybridization probe, the 1.2 kb *Eco*RI insert of pAG18 (from λAG18; 2) was purified by electroelution from a 0.8% agarose gel and nick-translated to a specific activity of 5 x 10^8 cpm/μg. In addition, a λgt11 human lung cDNA library was screened by plaque hybridization (6) with the nick-translated pAG18 cDNA insert. The inserts from the positive clones were subcloned directly into M13 mp18 and mp19 (7). All DNA sequencing reactions were carried out by primer extension in both orientations (8) using either the M13 universal primer or synthetic oligonucleotides (17-mers) constructed to α-Gal A gene sequences with a Biosearch 8700 DNA synthesizer.

Construction and Screening of X-Chromosome-Enriched Genomic Libraries— Genomic DNA was isolated (9) from the 49,XXXXX human lymphoblast line, partially digested with *Mbo* I and fractionated in a 0.9% agarose gel. Purified target DNA (13-22 kb) was ligated to the lambda replacement vector, EMBL3 (10), which had been digested to completion with *Bam*HI and *Eco*RI to prevent religation to the middle stuffer fragment. The ligated DNA was packaged with extracts prepared by the method of Ish-Horowicz and

Burke (11). Approximately 700,000 plaques from the unamplified 49,XXXXX library, designated λ5X, and 2 x 10^6 recombinants from the λ4X library (4) were screened (6) at a density of 10,000 plaques per 150 mm plate. Filters were hybridized with the [^{32}P]-nick-translated insert from pAG18 as described (12).

Characterization of Genomic Clones— Phage DNA was isolated from purified positive plaques (13), digested with various restriction endonucleases, separated by agarose gel electrophoresis, and transferred to nylon membranes (14). *Sac*I- and *Pvu*II-digested DNAs were hybridized with the [^{32}P]nick-translated, α-Gal A M13 deletion subclones (2) in order to identify and orient the location of exonic sequences. Finer mapping was accomplished with double digests of the genomic inserts and isolated *Sac*I restriction fragments. Selected genomic fragments containing exonic sequences and 5' and 3' flanking genomic fragments were subcloned into M13 vectors and sequenced by the dideoxynucleotide chain termination method in both orientations as described above. Additional restriction mapping was performed to position the intron/exon boundaries.

RESULTS AND DISCUSSION

Isolation of a Full-Length cDNA— pAG18 was used to screen the human fibroblast pcD cDNA library (15) and the human lung cDNA library. Following purification, hybridization with 5' and 3' λAG18 M13 subclones (2) and restriction analyses, only one of the 15 positive clones, pcDAG126, was putatively full-length. The 1437 nt *Pst*I/*Bam*HI insert was subcloned into M13 mp18 and mp19 and sequenced (Fig. 1). The pcDAG126 insert was 136 nt longer than that of λAG18 and contained 60 nt of 5' untranslated sequence, the initiation codon and the entire open reading frame which encoded a 31 amino acid signal peptide and the 398 residues of the mature enzyme. The 60 nt 5' untranslated sequence was average in length for such sequences and contained one

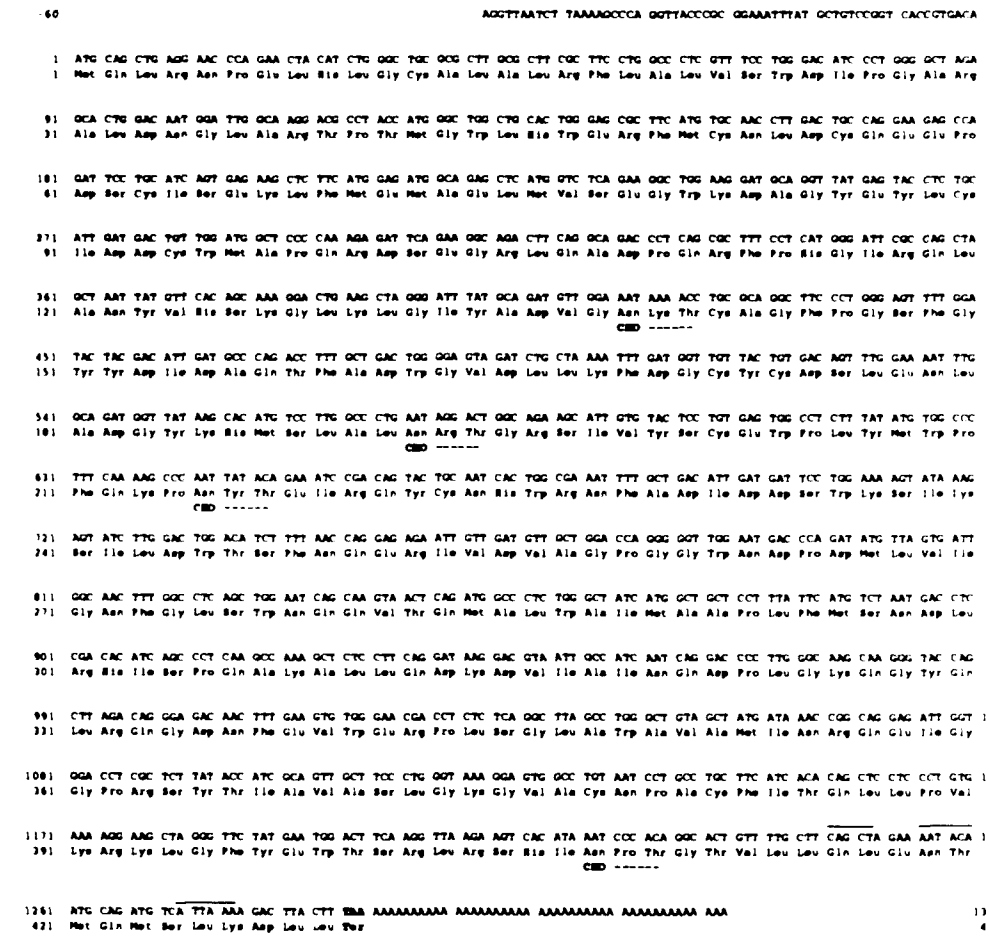


Figure 1. Nucleotide and predicted amino acid sequences of the 1437 nt full-length human α-Gal A cDNA insert of pcDAG126. This clone contained 60 nt of 5' untranslated sequence and an open reading frame which encoded a 31 amino acid signal peptide and 398 residues of the mature enzyme. Nucleotide A of the ATG initiation codon is designated +1. Overlines indicate the poly(A) signals AATACA and ATTAAA and the CAGCT site implicated in U4 small nuclear RNA binding.

out-of-frame ATG at -22 which did not have the conserved purine at -25 or the other consensus nucleotides for initiation codons (16) (Fig. 1). The 31 amino acid signal peptide was consistent with that predicted from maturation studies, indicating that human α-Gal A was synthesized as a ~50 kDa precursor glycoprotein which was proteolytically cleaved to the mature ~46 kDa lysosomal glycoprotein (17). As shown in Figure 2, the predicted signal peptide was typical of such sequences (18-20) and included a basic amino acid in the first 5 residues (Arg-4) followed by a central hydrophobic core of at least 9 residues (Leu-8 to Val-22), an alpha helix breaker such as proline or glycine -4 to -8 from the cleavage site (Pro-27,Gly-28), a more polar C-terminal region (Asp-25 and Arg-30) and the most frequently observed C-terminal sequence, Ala-X-Ala. Identification of the signal peptidase cleavage site using the weight-matrix method of von Heijne (20), unequivocally predicted cleavage after Ala-31 (score = 7.36). This site was consistent with Leu-32 being the amino-terminal residue, as had been established previously by microsequencing the purified enzyme (1). Thus, microsomal cleavage of the signal peptide is the only amino-terminal processing of human α-Gal A. An identical signal peptide sequence for α-Gal A was reported by Tsuji et al. (21), based on sequencing an α-Gal A cDNA from the pcD library (pcD-AG210) which had a 5' untranslated region 37 nt shorter than pcDAG126. The amino-terminal processing of the human α-Gal A prepeptide is similar to that of other human lysosomal enzymes (e.g., 22) and differs from those that contain pre- and pro-segments requiring a second cleavage to form the mature polypeptide (e.g., 23, 24).

Isolation and Restriction Mapping of Genomic Clones— Bacteriophage libraries enriched for human X-chromosomal DNA were screened with the nick-translated pcDAG126 and pAG18 inserts. Three overlapping clones containing the entire α-Gal A cDNA sequence were isolated, purified and subjected to extensive restriction mapping and Southern hybridization with the radiolabeled λAG18 M13 deletion subclones (2) to locate fragments containing exonic sequences (Fig. 3). Clone λ-W2, isolated from the λ4X

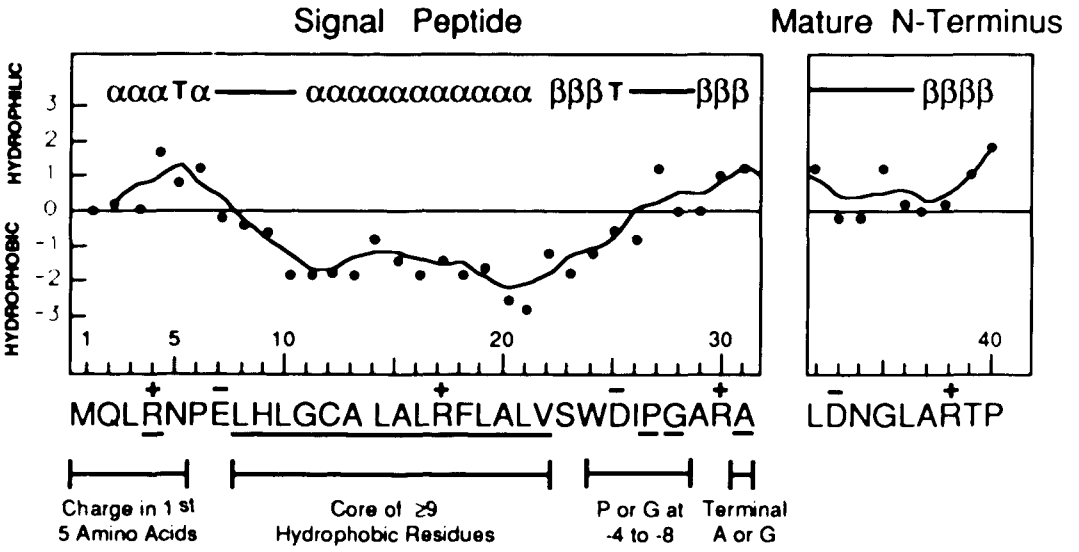


Figure 2. Secondary structure of the α-Gal A signal peptide. The hydrophobicity profiles and the predicted secondary structure for the 31 amino acid residue signal peptide and the first 9 residues of the mature N-terminus are shown. Amino acids are represented by their one letter code beginning with the initiation methionine at +1. Consensus sequences required for signal peptides are indicated and the corresponding residues in the α-Gal A signal peptide are underlined. Charged residues are indicated (+ or -). Deduced α-helical regions (α), β-pleated sheets (β), random coils (--) and turns (T) are indicated.

library, was ~12 kb and contained the 5' end of the gene (exon 1) and an additional 9 kb of 5' flanking sequences. Two overlapping clones, λ-B5 and λ-B18, were isolated from the unamplified λ5X human lymphoblast library. λ-B5 was ~14 kb and contained the entire coding sequence. λ-B18 was ~15 kb and contained 3' coding sequences as well as about 11 kb of 3' flanking sequence. These three clones spanned 32 kb of the X-chromosome which included the entire α-Gal A gene (~12 kb) and about 20 kb of flanking sequences.

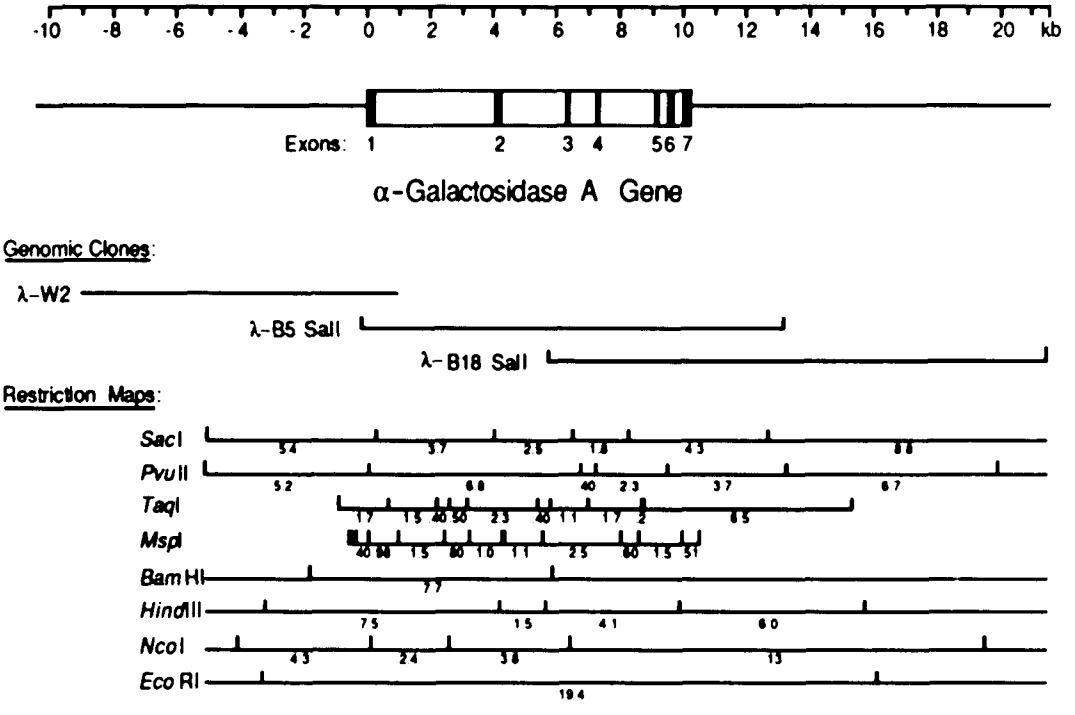


Figure 3. Organization of the human α-Gal A gene. Exons 1-7 are denoted by solid vertical boxes, introns by open areas. The three overlapping genomic clones (λ-W2, λ-B5, λ-B18) used to characterize the gene and their complete restriction maps are indicated.

The Intron/Exon Boundaries— Sequencing of selected α-Gal A restriction fragments revealed the presence of seven exons in the gene (Fig. 3). There were no discrepancies between the exonic sequence in the genomic clones and that of pcDAG126 or

pAG18 (2). The exons ranged in length from 92 to 291 nt (Table I) while the introns ranged from 200 nt (intron 5) to 3.7 kb (intron 1). The first exon (254 nt) contained the 60 nt 5' untranslated sequence as well as the sequence encoding the signal peptide and the first 33 residues of the mature enzyme. The intron/exon boundary sequences for all seven exons are shown in Table I. All intron/exon splice junctions followed the "GT/AG" rule (25) and were consistent with the consensus sequences for splice junctions of RNA polymerase II-transcribed genes (26). Putative lariat branch points were identified between -23 and -28 nt from the splice junction for all 6 introns by similarity to the less well conserved consensus sequence (C/T)N(C/T)T(A/G)A(A/C/T) (27, 28). All three codon phases were observed at the intron/exon junctions in the α-Gal A gene (Table I). Phase 0 junctions (between codons) occurred four times, while phase I and II junctions each occurred once.

To date, the only other reported isolated and characterized human lysosomal genes are those for the β-hexosamimidase α-chain (29), β-hexosamimidase β-chain (30), glucocerebrosidase (31) and acid phosphatase (32). The α-Gal A gene, was smaller in size (~12 kb with 7 exons) than the β-hexosamimidase α-chain gene which had 14 exons spanning ~35 kb (29) and the β-hexosamimidase β-chain gene which extended over 40 kb and was also split among 14 exons (30). α-Gal A was larger in size than the glucocerebrosidase gene which had 11 exons spanning ~7 kb of DNA (31) and the acid phosphatase gene which was ~9 kb and had 11 exons (32). In contrast to many secretory and membrane proteins (33), the signal sequences (prepeptides) were not contained on separate exons in these five lysosomal genes. However, intron 1 interrupts the coding sequence of the β-hexosamimidase α-chain gene at the proteolytic processing site of the propeptide indicating a possible functional domain (29).

Regulatory Elements— Transcriptional activity of eukaryotic genes is mediated by different combinations of promoter and enhancer elements. The promoter is required for

Table I. Nucleotide sequence of the intron/exon boundaries in the human α-Gal A gene.

Exon number and (size)	cDNA position of exon	5' Splice donor	Intron number and (size)	3' Splice acceptor	Codon phase
(nt)			(kb)	5' cctgAGGTTAATC	
1 (254)	-60-194	TGCATCAG gtatca	1 (3.8)	tgaattgtaatgattattgaaatttctctttcag TGAGAAGC	II
2 (175)	195-369	CTAATTAT gtgagt	2 (2.0)	acaatggtgactcttttcctccctctcatttcag GTTCACAG	0
3 (178)	370-547	GGCAGATG gtaatg	3 (1.5)	tttcccttattttacccattgttttctcatacag GTTATAAG	1
4 (92)	548-639	TTCAAAAG gtgaga	4 (1.2)	aaagtaaacagaagagtcatatctgttttcacag CCCAATTA	0
5 (162)	640-801	CAGATATG gtaaaa	5 (0.2)	tctcttgtttgaattatttcattctttttctcag TTAGTGAT	0
6 (198)	802-999	TTAGACAG gtaaat	6 (0.4)	gttgctaagcaaccacactttcttggttttcag GGAGACAA	0
7 (291)	1000-1290	TACTTTAAaatgt 3'			

Consensus sequences: donor: $^{A}_{C}AG\ gt^{a}_{g}agt$ lariat: $^{c}_{t}n^{c}_{t}t^{a}_{g}a^{c}_{t}$ acceptor: $^{cccccccccc}_{ttttttttttt}ncag$

Exon sequences are in upper case letters; intron sequences in lower case. Introns 5 and 6 were completely sequenced and the sizes of introns 1 through 4 were determined by restriction enzyme analyses. Donor and acceptor consensus sequences as well as those for the lariat branch point are indicated. The putative lariat branch points are underlined in the 3' splice acceptor sequences. Codon phase 0 intron/exon junctions occur between codons, codon phases I and II interrupt codons after the first and second nucleotides, respectively.

accurate and efficient initiation of transcription, while enhancers increase the rate of transcription from promoters (34). A typical promoter region directing transcription by RNA polymerase II is usually about 100 nucleotides in length and has an A/T-rich region, the TATA box, that lies between 20 and 30 bases upstream from the initiation of transcription site (CAP site) (35). More distal elements that are predominantly involved in controlling the frequency of initiation are found in many promoters. CAAT box elements are often located in this region and are typically present 70-80 nucleotides upstream from the CAP site (36). α-Gal A belongs to a class of genes known as "housekeeping" genes, which are expressed at relatively low levels in most cell types. The promoters of several housekeeping genes have been characterized and unusual structural features were uncovered (37-40). Sequencing the promoters for adenosine deaminase (37), 3-hydroxy-3-methyl glutaryl coenzyme A reductase (38), hypoxanthine phosphoribosyl transferase (39) and 3-phosphoglycerate kinase (40) revealed the lack of consensus sequences for TATA and CAAT boxes characteristic for eukaryotic promoters. A notable feature of the promoters was the presence of GGGCGG repeat sequences which are binding sites for the transcription factor SP1, a DNA binding protein that has been shown to enhance transcription by RNA polymerase II 10- to 50- fold (41).

The 5' flanking regions of the housekeeping genes analyzed are all highly rich in G+C nucleotides and may be related to the "methylation-free" or "HTF (*Hpa* II tiny fragments) islands" reported by Bird et al. (42, 43). These islands of 500 to greater than 2000 base pairs contain at least 50% G+C and the number of CpG dinucleotides approximately equals GpCs. The CpG clusters in this region are non-methylated and this has been shown to be correlated to the transcriptional activity of the genes; i.e., the transcription of genes with "HTF islands" is inhibited when the island is methylated.

Several possible regulatory elements were identified in the α-Gal A 5' flanking region (Fig. 4). The TAATAA sequence was 86 nt from the initiation codon (all distances are from the initiation codon to the most 3' nt of the indicated element). Canonical CCAAT

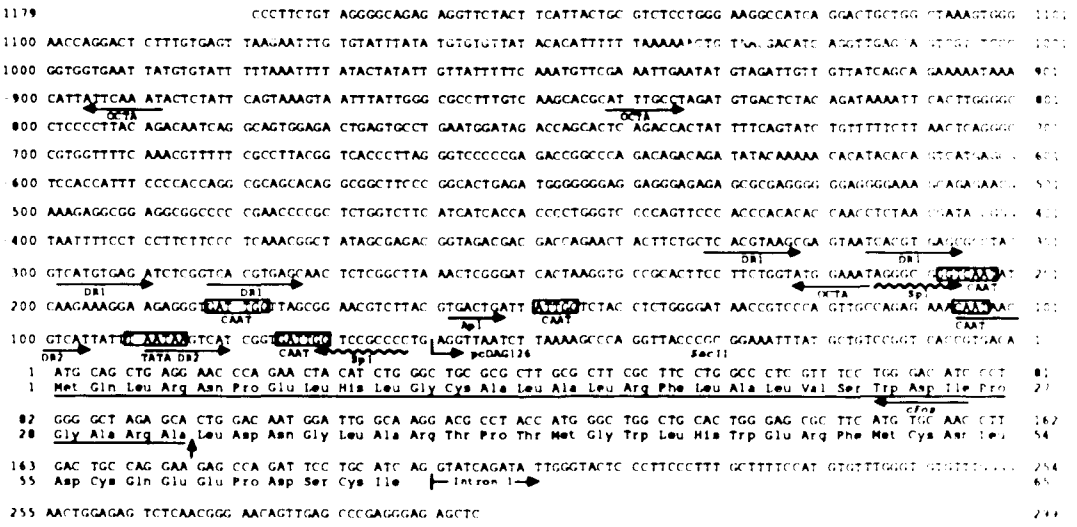


Figure 4. The 5' nucleotide sequence of the α-Gal A gene and its 5' flanking region. The 1478 nt sequence was derived from restriction fragments of genomic clones λ-W2 and λ-B5 (see Fig. 3). The sequence includes 1179 nt of 5' flanking DNA, 192 nt of coding sequence and 107 nt of intron 1. The sequence was determined in both orientations. The sequence from -60 to +192 was identical to that in the full-length cDNA clone, pcDAG126. The signal peptidase cleavage site is indicated by an upward arrow. Nucleotide A of the initiation codon, ATG, is designated +1. Sequences resembling TATA and CCAAT motifs are boxed. SP1 binding sites (SP1) are underlined with a wavy arrow; "OCTA" enhancer elements (OCTA), an AP-1 enhancer sequence (AP1) and direct repeats (DR1, DR2) are underlined with straight arrows. A GC rich region ranges from -660 to +1. See text for details.

box sequences (5' GGYCAATCT 3'; 36) were located at -71, -146 and -178 nt from the initiation codon on the antisense strand, while degenerate forms, CAAT and GGTCAAT, occurred at positions -104 and -203 nt, respectively. The GC box consensus sequence for the promoter-specific transcription factor Sp1 (41), occurred in the reverse orientation at -63 nt and in the forward orientation at -207 nt. For comparison, the β-hexosaminidase α-chain gene had a 5' region -253 to -283 nt from the initiation codon (the cap site is yet to be determined) which contained a TATA-like box (TTATTTA) and a CCAAT box (CCATC) with an intervening GC-rich region (29). The human glucocerebrosidase gene also had TATA-like and CAAT-like boxes in its 5' upstream region, but lacked binding sites for the SP1 transcription factor (31), while the acid phosphatase gene contained two GC boxes but lacked TATA and CAAT sequence motifs.

Several potential enhancer binding sites also were present in the α-Gal A gene. The conserved recognition motif (TGACTCA) of the AP1 enhancer-binding protein (44) was identified at -153 nt. The immunoglobulin "OCTA" enhancer element (ATTTGCAT; 45) was present with one mismatch at -835 nt and, in the opposite orientation, at -889 nt. A perfect match with the reverse complement of the *c-fos* enhancer element GATGTCC (46), occurred at +70 to +78. Interestingly, the "chorion box" enhancer [TCA(T/C)GT(G/A)AG(C/A); (47, 48)] was found in four direct repeats, each separated by ~6 nt at -274, -290, -307 and -323 nt. Future evaluation of the importance of these putative promoter/enhancer elements in the regulation of α–Gal A gene expression will require footprinting experiments and functional assays with deletion/mutation mapping (49).

The α–Gal A gene contained a methylation-free or "HTF island" (42) in the region from -260 to -660 nt. This 5' region had a G+C content of 59% (typically >50%) and a CpG/GpC dinucleotide ratio of 1.5 (typically CpG/GpC > 1.0). Extension of the region to +1 nt, in order to include a *Sac*II site which is indicative of "HTF islands" (43), slightly reduced the G+C content to 54% and the dinucleotide ratio to 1.4. Since "HTF islands"

have been implicated in maintaining inactivation of X-linked genes (50, 51), analysis of the methylation patterns in this region from active vs inactive X-chromosomes will complement previous studies of X-reactivation of α-Gal A expression in mouse-human somatic cell hybrids (51).

The 3' Flanking Region— A most unusual feature of the λAG18 cDNA insert was the absence of a 3' untranslated sequence; the polyadenylation signal sequence was in the coding region 12 nt from the termination codon which was followed by the poly(A) tract (2). This finding is unique among human nuclear-encoded mRNAs. In mammals, the only other example of an mRNA lacking a 3' untranslated region is the mRNA for murine thymidylate synthase (53). It is of interest that the human thymidylate synthase transcript did have a 3' untranslated region (53). To further investigate the possibility that α-Gal A transcripts lack a 3' untranslated region, additional cDNA clones from different human cDNA libraries were isolated and their 3' sequences were determined. The possibility that the α-Gal A transcript contained a 3' poly(A) sequence which could serve as a site for oligo(dT) binding in the construction of cDNA libraries was ruled out by sequencing the 3' genomic region. As shown in Table II, five cDNA clones from three different libraries lacked 3' untranslated regions. In two other clones, (pcDAG7 and pcDAG41), the TAA termination codons were followed by the short 3' untranslated sequences, AATGTTT and AATGTC, respectively, while a previously reported cDNA (21) had the sequence AATGTT (Table II). Thus, while the majority of the isolated α-Gal A cDNAs lack a 3' untranslated sequence, alternative cleavage and polyadenylation can result in a short 6 or 7 nt untranslated sequence, maintaining the termination codon. It is possible that these alternative, short 3' regions represent microheterogeneity of the primary α-Gal A cleavage reaction (54). Additionally, the second upstream (-28 nt) polyadenylation signal, AATACA, may be involved in the variant cleavages.

Table II. 3' Sequences of α-Gal A cDNAs and Genomic Clone λ-B18.

Clone	3' Sequence
cDNA	
λAG18	....TAA . . . . . $(A)_{12}$
λHLAG4	....TAA . . . . . $(A)_{122}$
pcDAG8	....TAA . . . . . $(A)_{33}$
pCDAG69	....TAA . . . . . $(A)_{92}$
pcDAG126	....TAA . . . . . $(A)_{43}$
pCDAG41	....TAA AATGTC. . . . . $(A)_{57}$
pcDAG7	....TAA AATGTTT . . . . $(A)_{91}$
Genomic	
λ-B18	....TAA AATGTTTATTTTATTGCCAACT
	ACTACTTCCTGTCCACCTTTTTCTCC....
Downstream consensus:	YTGTTYY TTTTTTTT

The α-Gal A cDNA and genomic clones were isolated and sequenced as described in the text. The 51 nt genomic 3' sequence, beginning with the termination codon(TAA) and ending with CTCC, is aligned with the sequence in pcDAG7. The "GT-box" and "T-rich" element homologies each cccur twice in the genomic sequence as indicated by underlines.

Since the murine thymidylate synthase gene did not have a 3' untranslated sequence, it was proposed that cleavage and polyadenylation was facilitated by the formation of a stem and loop structure in which the polyadenylation signal completely base-paired with downstream sequences (53). When the human α-Gal A genomic sequence was similarly folded (BIOFLD program, BIONET), the polyadenylation signal was not completely base-paired with downstream signals. In contrast to the thymidylate gene, the 3' flanking region of the α-Gal A gene contained sequences with similarities to the consensus downstream elements, YGTGTTYY ("GT-box") and T_n ("T-rich"), recently shown to be involved in polyadenylation (55-57). Two pairs of these downstream elements occur in the α–Gal A gene first at +16 and +22 nt and again at +43 and +52 nt

from the polyadenylation signal (Table II). It is possible that the alternative cleavage sites may be due to the presence of these additional, more 3', elements. Whatever the mechanism responsible for these alternative 3' sequences, it is notable that the α-Gal A transcript does not require a 3' untranslated region for stability or translation.

REFERENCES

1. Calhoun, D.H., Bishop, D.F., Bernstein, H.S., Quinn, M.,Hantzopoulos, P.and Desnick, R.J. (1985) *Proc. Natl. Acad. Sci. USA* **82**, 7364-7368

2. Bishop, D.F., Calhoun, D.H., Bernstein, H.S., Hantzopoulos, P., Quinn, M. and Desnick, R.J. (1986) *Proc. Natl. Acad. Sci. USA* **83**, 4859-4863

3. Okayama, H. and Berg, P. (1983) *Mol. Cell. Biol.* **3**, 280-289

4. Wood, W.I., Capon, D.J., Simonsen, C.C., Eaton, D.L., Gitschier, J., Keyt, B., Seeburg, P.H., Smith, D.H., Hollingshead, P., Wion, K.L., Delwart, E., Tuddenham, G.D., Vehar, G.A. and Lawn, R.M. (1984) *Nature* **312**, 330-337

5. Hanahan, D. and Meselson, M. (1983) *Methods Enzymol.* **100**, 333-342

6. Benton, W.D. and Davis, R.W. (1977) *Science* **196**, 180-182

7. Messing, J. (1983) *Methods Enzymol.* **101**, 20-78

8. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 5463-5467.

9. Aldridge, J., Kunkel, L., Bruns,G., Tantravahi, U., Lalande, M., Brewster, T., Moreau, E., Wilson, M., Bromley, W., Roderick, T. and Latt, S. (1984) *Am. J. Hum. Genet.* **36**, 546-564.

10. Frischauf, A.M., Lehrach, H., Poustaka, A. and Murray, N. (1983) *J. Mol. Biol.* **170**, 827-842

11. Ish-Horowicz, D. and Burke, J.F. (1981) *Nuc. Acids. Res.* **9**, 2989-2998

12. Maniatis, T., Fritsch, E.F. and Sambrook, J. (1982) Molecular Cloning: A Laboratory Manual (Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.)

13. Blattner, F.R., Williams, B.G., Blechl, A.E., Denniston-Thompson, K., Faber, H.E., Furlong, L.-A., Grunwald, D.J., Kiefer, D.O., Moore, D.D., Sheldon, E.L. and Smithies, O. (1977) *Science* **196**, 161-169

14. Southern, E. (1975) *J. Mol. Biol.* **98**, 503-517

15. Okayama, H. and Berg, P. (1982) *Mol. Cell. Biol.* **2**, 161-170

16. Kozak, M. (1984) *Nucleic Acids Res.* **12**, 857-872

17. Lemansky, P., Bishop, D.F., Desnick, R.J., Hasilik, A. and von Figura, K. (1987) *J. Biol. Chem.* **262**, 2062-2065

18. Perlman, D. and Halvorson, H. O. (1983) *J.Mol. Biol.* **167**, 391-409

19. Watson, M.E.E. (1984) *Nucleic Acids Res.* **13**, 5145-5164

20. von Heijne, G. (1986) *Nucleic Acids Res.* **14**, 4683-4690

21. Tsuji, S., Martin, B.M., Kaslow, D.C., Migeon, B.R., Choudary, P.V., Stubblefield, B.K., Mayor, J.A., Murray, G.J., Barranger, J.A. and Ginns, E.I. (1987) *Eur. J. Biochem.* **165**, 275-280

22. Sorge, J., West., C., Westwood, B. and Beutler, E. (1985) *Proc. Natl. Acad. Sci. USA* **82**, 7289-7293

23. Faust, P.L., Kornfeld, S. and Chirgwin, J.M. (1985) *Proc. Natl. Acad. Sci. USA* **82**, 4910-4914

24. Myerowitz, R., Piekarz, R., Neufeld, E.F., Shows, T.B. and Suzuki, K. (1985) *Proc. Natl. Acad. Sci. USA* **82**, 7830-7834

25. Breathnach, R. and Chambon, P. (1981) *Ann. Rev. Biochem.* **50**, 349-383

26. Mount, S.M. (1982) *Nucleic Acids Res.* **10**, 459-472

27. Ruskin, B., Krainer, A.R., Maniatis, T. and Green, M.R. (1984) *Cell* **38**, 317-331

28. Keller, E.B. and Noon, W.A. (1985) *Nucleic Acids Res.* **13**, 4971-4981

29. Proia, R.L. and Soravia, E. (1987) *J. Biol. Chem.* **262**, 5677-5681

30. Proia, R.L. (1988) *Proc. Natl. Acad. Sci. USA.* **85**, 1833-1837

31. Horowitz, M., Wilder, S., Horowitz, Z., Reiner, O., Gelbart, T. and Beutler, E. (1989) *Genomics* **4**, 87-96

32. Geier, C., von Figura, K. and Pohlmann, R. (1989) *Eur. J. Biochem.* **183**, 611-616

33. Gilbert, W. (1985) *Science* 228, 823-824

34. Maniatis, T., Goodbourn, S. and Fischer, J.A. (1987) *Science* **236**, 1237-1245

35. Cordon, J., Wasylyk, B., Buchwalder, A., Sassone-Corsi, P., Kedinger, C. and Chambon, P. (1980) *Science* **209**, 1406-1414

36. Benoist, C. and Chambon, P. (1981) *Nature* **290**, 304-310

37. Valerio, D., Duyvesteyn, M.G.C., Bekker, B.M.M., Weeda, G., Berkvens, T.M., van der Voorn, van Ormondt, H. and van der Eb, A.J. (1985) *EMBO* **4**, 437-443

38. Reynolds, G.A., Basu, S.K., Osborne, T.F., Chin, D.J., Gil, G., Goldstein, J.L. and Luskey, K.L. (1984) *Cell* **38**, 275-285

39. Melton, D.W., Konecki, D.S., Brennand, J. and Caskey, T.C. (1984) *Proc. Natl. Acad. Sci.* USA **81**, 2147-2151

40. Singer-Sam, J., Keith, D.H., Tani, K., Simmer, R.L., Shively, L., Lindsay, S., Yoshida, A. and Riggs, A.D. (1984) *Gene* **32**, 409-417

41. Dynan, W.S. and Tjian,R. (1985) *Nature* **316**, 774-778

42. Bird, A.P. (1986) *Nature* **321**, 209-213

43. Lindsay, S. and Bird, A.P. (1987) *Nature* **327**, 336-338

44. Lee, W., Mitchell, P. and Tjian, R. (1987) *Cell* **49**, 741-752

45. Lenardo, M., Pierce, J.W. and Baltimore, D. (1987) *Science* **236**, 1573-1577

46. Prywes, R. and Roeder, R.G. (1986) *Cell* **47**, 777-784

47. Spoerel, N., Nguyen, H.T. and Kafatos, F.C. (1986) *J. Mol. Biol.* **190**, 23-35

48. Spradling, A.C., deCicco,D., Wakimoto, B.T., Levine, J.F., Kalfayan, L.J. and Cooley, L. (1987) *EMBO J.* **6**, 1045-1053

49. Gorman, C.M., Moffat, L.F. and Howard, B.H. (1982) *Mol. Cell. Biol.* **2**, 1044-1051

50. Yang, T.P. and Casky, T.C. (1987) *Mol. Cell. Biol.* **7**, 2994-2998

51. Keith, D.H. Singer-Sam, J. and Riggs, A.D. (1986) *Mol. Cell. Biol.* **6**, 4122-4125

52. Mohandas,T., Sparkes, R.S., Bishop, D.F. Desnick, R.J. and Shapiro, L.J. (1984) *Am. J. Hum. Genet.* **36**, 916-925

53. Jenh, C.-H., Deng, T., Li, D., DeWille, J. and Johnson, L.F. (1986) *Proc. Natl. Acad. Sci. USA* **83**, 8482-8486

54. Birnstiel, M.L., Busslinger, M. and Strub, K. (1985) *Cell* **41**, 349-359

55. McDevitt, M.A., Hart, R.P., Wong, W.W. and Nevins, J.R. (1986) *EMBO J.* **5**, 2907-2913

56. Gil, A. and Proudfoot, N.J. (1987) *Cell* **49**, 399-406

57. McLauchlan, J., Gaffney, D., Whitton, J.L. and Clements, J.B. (1985) *Nucleic Acids Res.* **13**, 1347-1368

Chapter Two

Determination of the Complete α-Galactosidase A Genomic Sequence

INTRODUCTION

The isolation and characterization of full-length cDNA and genomic clones encoding α-Gal A have provided the necessary tools for the molecular analysis of Fabry disease and for the expression of large amounts of the enzyme for structural studies and therapeutic trials of enzyme replacement. However, in order to precisely characterize α-Gal A gene rearrangements and to provide probes for the identification of intronic polymorphic sites, the complete sequence of the α-Gal A gene was determined. This information will provide a basis for the understanding of the molecular mechanisms underlying α-Gal A gene rearrangements. Interestingly, *Alu* sequences comprised about 30 percent of the α-Gal A gene. These 300 nucleotide middle repetitive elements are members of the largest dispersed repeat family representing ~3-6 percent of the human genome (1) or an average of 1 *Alu* every 4 kb (2). During evolution, *Alu* sequences presumably were inserted randomly into the human genome by retroposition of a series of related genes derived from an ancestral 7SL sequence (2, 3). *Alu* family members have been classified into evolutionarily related groups based on the extent of their divergence from consensus sequences at certain nucleotide positions (3-6). In addition, divergence of the 7SL integrated product occurred at hypervariable CpG dinucleotides (24 in the consensus sequence). According to the classification of Jurka and Smith (4), the oldest subfamily has been designated the *Alu*-J subfamily and is estimated to have arisen more than 60 million years ago. The more recent subfamily, *Alu*-S has been further split into three branches, *Alu*-a, -b and -c. In the α-Gal A gene, all 12 *Alu* repeats were classified in the more recent S subfamily, nine in the *Alu*-a, two in the *Alu*-b and one in the *Alu*-c branches. Whatever the origin the of *Alu* sequences may be, the unusually high frequency of these repeats in the α-Gal A gene predicts that this gene may be particularly prone to genetic recombination by unequal crossing over events.

EXPERIMENTAL PROCEDURES

Genomic clones described in Chapter 1 which spanned approximately 30 kb of the human X-chromosome were used to sequence the ~12 kb α-Gal A gene in its entirety in both orientations. The majority of the sequence was determined from unidirectional exonuclease III (Exo III) deletion subclones of the 3.7 kb, 2.5 kb and 1.8 kb *Sac*I restriction fragments (see Figure 3, Chapter 1) in pUC19 generated with the "Erase-a-Base System" (Promega). Five micrograms of each plasmid were digested with two restriction enzymes that cut uniquely in the pUC19 polylinker. The enzymes were chosen such that a 5' overhang was produced adjacent to the insert and a 3' overhang next to the priming site, thus protecting the vector. The doubly digested DNAs were treated with 360 units of Exo III at 35 °C. Approximately 200 ng aliquots were removed every 30 seconds and placed on ice. The samples were then treated with 2 units of S1 nuclease for 30 minutes at room temperature, followed by heat inactivation at 70 °C for ten minutes. The large fragment of DNA polymerase was used to fill in the ends and then the plasmids were circularized with T4 DNA ligase. The deleted plasmids were used to transform JM109 cells rendered competent by incubation with 0.1 M $MgCl_2$ and 0.1 M $CaCl_2$. A rapid screening of recombinants from each time point was performed by lysing the bacterial colonies directly in an SDS-agarose gel and separating the plasmid contents by gel electrophoresis (7). Subclones that contained a nested set of deletions were used for sequencing with the M13/pUC universal primer. Plasmid template DNA was isolated by the rapid boiling method of Wang et al. (8). Double-stranded sequencing of plasmid DNAs was accomplished using the GemSeq Klenow System (Promega) and Sequenase (United States Biochemical Corp.). Other genomic fragments were subcloned into M13 vectors (9); the single-stranded M13 DNA templates were sequenced using the dideoxy chain termination method (10) with universal M13 primers or oligomers constructed to α-Gal A sequences

synthesized on an Applied Biosystems Model 380B DNA Synthesizer. Computer-assisted alignment of the nucleotide sequences was performed using the Microgenie program (Beckman).

RESULTS AND DISCUSSION

The complete nucleotide sequence for the gene encoding human α-Gal A which includes 1179 nt of 5' flanking sequence and 1169 nt of 3' flanking sequence is shown in Figure 1 (pp. 50-57). The 12,436 bp sequence was determined on both strands of DNA in their entirety. Nucleotide A of the initiation codon, ATG, is 1180. Figure 2 is a schematic diagram of the α-Gal A gene indicating the positions of the seven exons, six introns and repeat elements. Of note, there were 12 *Alu* repetitive elements (11) distributed throughout the intervening sequences and the 3' flanking region. Four *Alu* repeats were located in intron 1 (*Alu* elements 1 and 2 were arranged in a tandem array), three in intron 2, one in intron 3, one in intron 4 and three in the 3' flanking region (*Alu* elements 10 and 11 were

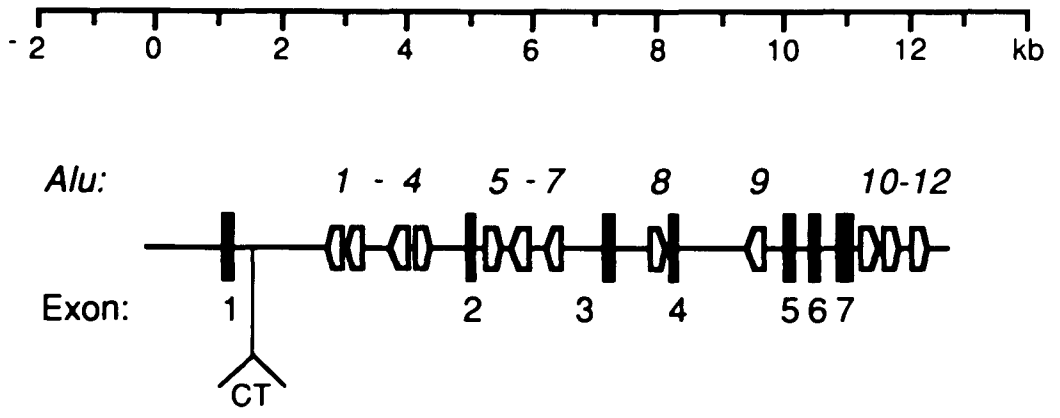


Figure 2. Schematic diagram of the α-Gal A gene. Exons 1-7 are shown as solid rectangles. The positions and orientations of the 12 *Alu* repeats are indicated. CT designates the position of the 84 bp CT repeat in intron 1. The nucleotide coordinates for the α-Gal A gene (top) are as for the entire 12,436 bp sequence (Fig. 1).

arranged in a tandem array). The finding of 12 *Alu* sequences in the 12 kb α-Gal A gene was well above the theoretical average of one *Alu* every 4 kb (2), but is not unprecedented. For example, the human β-tubulin gene has 10 *Alu* sequences clustered in a single 4.8 kb intron (12) and the ~33 kb human tissue plasminogen activator gene has 28 copies distributed throughout the 13 introns and the 5' flanking region of the gene (13).

The properties of the 12 *Alu* sequences are summarized in Table 1. *Alu* repeats 1, 2, 3, 6, 7 and 9 were in the reverse orientation, whereas repeats 4, 5, 8, 10, 11 and 12 were in the sense orientation. When categorized into "subfamilies" according to the classification of Jurka and Smith (4), all 12 had sequences homologous to the more modern "S subfamily"; nine were further subdivided into the older "a branch", two were assigned to the more recent "b branch" and one to the intermediate "c branch". Figure 3 is a comparison the the *Alu* elements in the α-Gal A gene to the consensus sequence. When each of these repeats was aligned with the consensus sequence, the percent divergence (based on the number of mismatches due to base substitutions, insertions and/or deletions) ranged from 9 to 15 percent (Table I and Fig. 3). Of the 24 CpG dinucleotides in the consensus sequence, the number in each of the 12 *Alu* repeats that had undergone spontaneous deamination of 5-methylcytosine to thymidine (14) ranged from 8 to 20 (Table I and Fig. 3). Analogously, comparison of the *Alu* sequences to each other revealed a divergence of 12 to 21 percent (data not shown) suggesting that each *Alu* resulted from a separate retroposition event. Moreover, all 12 *Alu* sequences were flanked by direct repeats of 4 to 18 bp (indicated by underlines in Figure 3), possibly indicating a staggered breakage of the host chromosome occurred during insertion of the mobile element (15). As previously observed with other direct repeats flanking *Alu* sequences, the 5' end of many of the α-Gal A direct repeats were especially A-rich and these regions have been proposed to be involved in the integration process of the repetitive elements (16).

In addition to the *Alu* elements, a CT repeat sequence of 84 bp was found in intron 1 (nucleotides 1603-1686). This repeat was predicted to form particularly stable secondary

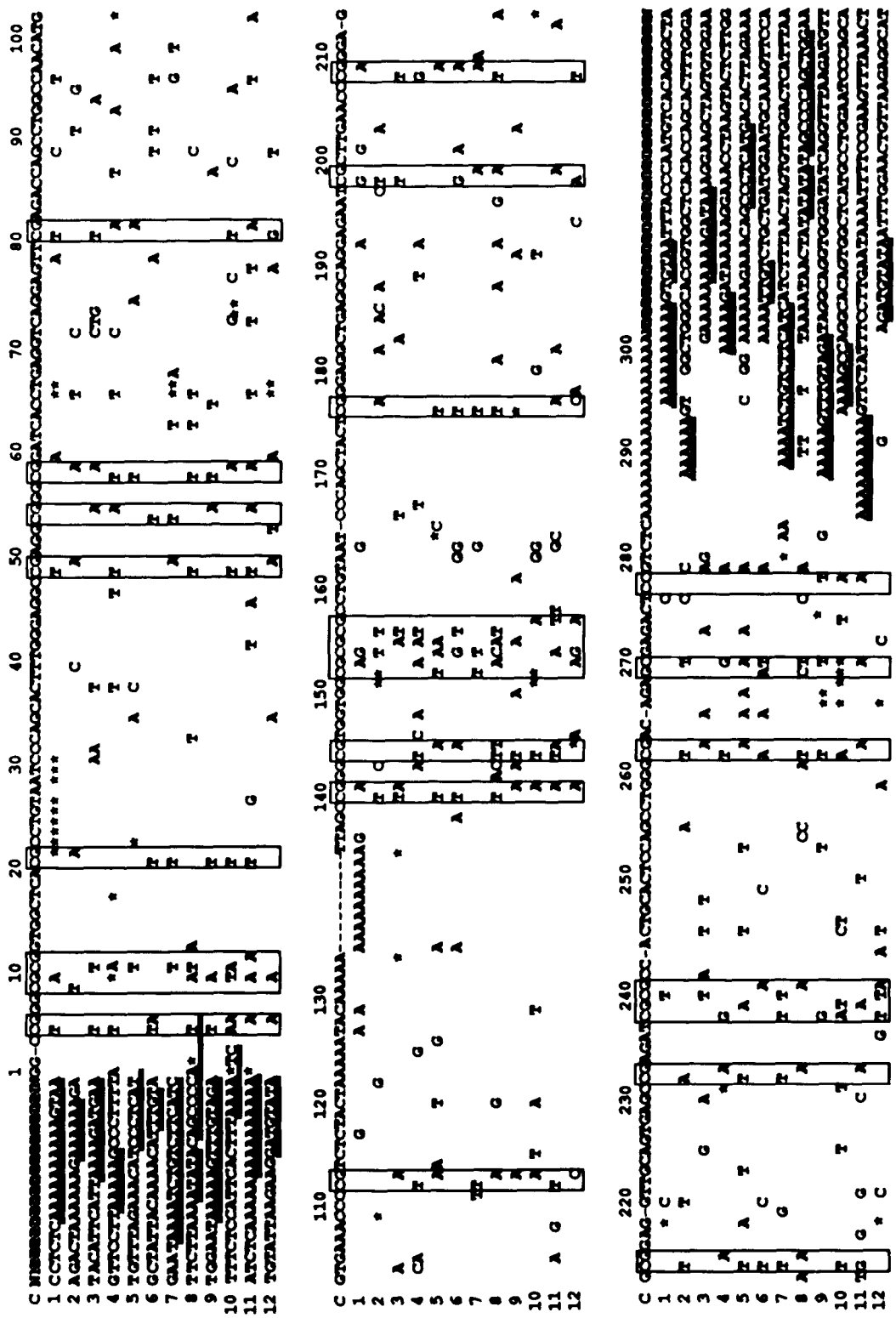
structures with a GA rich region in the 5' flanking sequence (nucleotides 626-692) when analyzed for dyad symmetry by the SEQ program (17). Other small imperfect direct repeats and inverted repeats were found throughout the gene, but the significance of these is unknown.

Table I. Properties of the Alu sequences in the human α-galactosidase A gene.

Alu Repeat	Nucleotide Coordinates (bp)	Sense	Subfamily Branch	% Homology With *Alu* Consensus Sequence	% Mutated CpGs
1	2851-3150	-	B	90.0	33.3
2	3155-3433	-	A	88.7	50.0
3	3828-4125	-	A	88.3	53.3
4	4214-4512	+	A	87.0	62.5
5	5351-5644	+	A	89.2	62.5
6	5759-6060	-	B	90.3	50.0
7	6306-6595	-	A	91.4	53.3
8	7945-8229	+	A	85.0	83.3
9	9508-9793	-	A	91.3	50.0
10	11330-11620	+	A	86.5	53.3
11	11621-11912	+	A	85.3	79.2
12	12101-12398	+	C	87.0	50.0

Coordinates correspond to the α-Gal A 12,436 bp sequence. The A of the ATG initiation codon is nucleotide 1180. *Alu* elements were classified according to the nomenclature of Jurka and Smith (4). For homology comparisons, gaps have been counted as single events even if they involved more than one base pair.

Figure 3. Sequence comparison of Alu repetitive elements within the human α-Gal A gene. The 12 α-Gal A *Alu* family members (1-12 shown on the sense strand) are aligned with the *Alu* consensus sequence (C) taken from the literature (4). Open spaces denote homology; deleted sequences are indicated by asterisks and inserted nucleotides are shown beneath the dashes placed in the consensus sequence. Direct repeats flanking the *Alu* sequences are underlined.



REFERENCES

1. Jelinek, W.R. and Schmid, C.W. (1982) *Ann. Rev. Biochem.* **51**, 813-844

2. Hwu, H.R., Roberts, J.W., Davidson, E.H. and Britten, R.J. (1986) *Proc. Natl. Acad.* Sci. USA **83**, 3875-3879

3. Deininger, P.L. and Slagel, V.K. (1988) *Mol. Cell. Biol.* **8**, 4566-4569

4. Jurka, J. and Smith, T. (1988) *Proc. Natl. Acad. Sci.* USA **85**, 4775-4778

5. Britten, R.J., Baron, W.F., Stout, D.B., and Davidson, E.H. (1988) *Proc. Natl. Acad. Sci.* USA **85**, 4770-4774

6. Willard, C., Nguyen, H.T., and Schmid, C.W. (1987) *J. Mol. Evol.* **26**, 180-186

7. Sekar, V. (1987) *BioTechniques* **5**, 11-13

8. Wang, L-M., Weber, D.K., Johnson, T., and Sakaguchi, A.Y. (1988) *Biotechniques* **6**, 839-843

9. Messing, J. (1983) *Methods Enzymol.* **101**, 20-78

10. Sanger, F., Nicklen, S., and Coulson, A.R. (1977) *Proc. Natl. Acad. Sci.* USA **74**, 5463-5476

11. Deininger, P.L., Jolly, D.J., Rubin, C.M., Friedmann, T. and Schmid, C.W. (1981) *J. Mol. Biol.* **151**, 17-33

12. Lee, M.G-S, Loomis, C. and Cowan, N.J. (1984) *Nucleic Acids Res.* **12**, 5824-5836

13. Friezner Degen, S.J., Rajput, B. and Reich, E. (1986) *J. Biol. Chem.*. **261**, 6972-6985

14. Coulondre, C., Miller, J.H., Farabaugh, P.J., and Gilbert, W. (1978) *Nature* **274**, 775-780

15. Weiner, A.M., Deininger, P.L., and Efstratiadis, A. (1986) *Ann. Rev. Biochem.* **55**, 631-661

16. Daniels, G.R. and Deininger, P.L. (1985) *Nucleic Acids Res.* **13**, 8939-8954

17. Brutlag, D.L., Clayton, J., Friedland, P., and Kedes, L.H. (1982) *Nucleic Acids Res.* **10**, 279-294

Figure 1. Complete nucleotide sequence of the human α-Gal A gene. The 12,436 bp sequence derived from α-Gal A genomic clones was determined on both strands of DNA in their entirety. Exonic sequences are in uppercase; intronic sequences are in lowercase and *Alu* sequences are in italics. A CT-rich region in intron 1 is underlined.

ccctctgta ggggcagaga ggttctactt cattactgcg tctcctggga aggccatcag 60

gactgctggc taaagtggga accaggactc tttgtgagtt aagaatttgt gtatttatat 120

gtgtgttata cacatttttt aaaaaactgt aacgacatca ggttgagcag tcgtctccgg 180

gtggtgaatt atgtgtattt ttaaatttta tactatattg ttatttttca aatgttcgaa 240

attgaatatg tagattgttg ttatcagcag aaaaataaac attattcaaa tactctattc 300

agtaaagtaa tttattgggc gcctttgtca agcacgcatt tgcctagatg tgactctaca 360

gataaaattc acttggggcc tccccttaca gacaatcagg cagtggagac tgagtgcctg 420

aatggataga ccagcactca gaccactatt ttcagtatct gtttttctta actcagggcc 480

gtggttttca aacgtttttc gccttacggt cacccttagg gtcccccgag accggcccag 540

acagacagat atacaaaaac acatacacag tcatgagcgt ccaccatttc cccaccaggc 600

gcagcacagg cggcttcccg gcactgagat gggggggagg agggagagag cgcgaggggg 660

gagggggaaag cagagaacga aagaggcgga ggcggccccc gaaccccgct ctggtcttca 720

tcatcaccac ccctgggtcc ccagttccca cccacacacc aacctctaac gataccgggt 780

aattttcctc cttcttccct caaacggcta tagcgagacg gtagacgacg accagaacta 840

cttctgctca cgtaagcgag taatcacgtg agcgcctacg tcatgtgaga tctcggtcac 900

gtgagcaact ctcggcttaa actcgggatc actaaggtgc cgcacttcct tctggtatgg 960

aaatagggcg ggtcaatatc aagaaaggaa gagggtgatt ggttagcgga acgtcttacg 1020

tgactgatta ttggtctacc tctgggggata accgtcccag ttgccagaga aacaataacg 1080

tcattattta ataagtcatc ggtgattggt ccgcccctgA GGTTAATCTT AAAAGCCCAG 1140

GTTACCCGCG GAAATTTATG CTGTCCGGTC ACCGTGACAA TGCAGCTGAG GAACCCAGAA 1200

CTACATCTGG GCTGCGCGCT TGCGCTTCGC TTCCTGGCCC TCGTTTCCTG GGACATCCCT 1260

GGGGCTAGAG CACTGGACAA TGGATTGGCA AGGACGCCTA CCATGGGCTG GCTGCACTGG 1320

GAGCGCTTCA TGTGCAACCT TGACTGCCAG GAAGAGCCAG ATTCCTGCAT CAGgtatcag 1380

atattgggta ctcccttccc tttgcttttc catgtgtttg ggtgtgtttg gggaactgga 1440

gagtctcaac gggaacagtt gagcccgagg gagagctccc ccacccgact ctgctgctgc 1500

ttttttatcc ccagcaaact gtcccgaatc aggactagcc ctaaactttc tctgtgtgac 1560

ctttcctggg atgggagtcc ggccagcggc ccctgtttct ttctctctct ctctctctct 1620

cgttctcgtt ctctttctct tttctcttgtt tcctctctct ttctctctct ccctgccccg 1680

ttctctcttt tcactgctcc ttgcagagca gggccacccc ataggcagtg tgcccaaagt 1740

agccctgccc ggttctattc agacccttct tgtgaacttc tgctcttcct ctgccgggtg 1800

ctaaccgtta gaacatctag ggtgggtagg aggaatgggg aactaagatt cgtgccattt 1860

tttctccttt tggggtcgtg gatttctcgg cagtatctcg agggagttag agagaccata 1920

aggtcgctga gatctctccc acctcgccca tgagcgtggc atcaggctgg aaggttgaca 1980

tggaggaact ttatacattt acacctttgc gtgagggttg aggctggatt agataggtat 2040

tgaacatatc tgaccctcac aatccttatc tgtaaattgg gattacaacc ttttaatttc 2100

agggagctga caaaaaaaat ctgaaaaata gttcttatct cacacaggtg agttttcaag 2160

gagataacct atttaaagta catagcacag cgcttgacca ttcaactgcg cttacagagc 2220

aaatgttcaa tgggaaaatg aatgtaaatc tacaaatctg aatgaatatg tgtatttttc 2280

tggagaggagg atatttacct ttcttcaaat tctcaaaggg ctctgtgatt taaaaaaggt 2340

taggaatcac tgatagatgt tggtaaaagg tggcagtcac agtacatttc tgtgtccata 2400

agttattcct atgaatatct ttatagataa agtcaggatg ttggtcagac atcacagaag 2460

aaattggcct tgtaagtttc atgtgaccct gtggtacagt atgtgtggca attttgccca 2520

tcacggattt ttttttattg gtatttgcat ctgattataa aactaatgca tgatcattgc 2580

aaaaaatgta gataaagaag agcaaaatga aaataaagat ttccccccac cgttccacca 2640

cccagaaata atcatggttt aaatgttaat atacaacctt acaattgttt tctatataaa 2700

tgaaaacata gatttcttta tttcattatt ttccataaaa aatggatcat gtttatgtca 2760

tgtttggcta atggcaagac cctggcaccc agtctgggct caaattctgc ctcattgtta 2820

cttagccctg tgacattggg taaattacac tttttttttt tttttttttt tgagacgggg 2880

tctcgctctg tcgcccaggc tggagtgcag tggcacgatc tcggctcact gcaagtccgc 2940

ctcctgggtt cacgccattc ttctgcctca gcctcccgag tagctgggac tacaggcgcc 3000

tgccaccacg cctggctctt tttttttttt tttttttttt tagtacagac ggggtttcac 3060

catgttagcc agggtggtct caatctcctg acctcgtgat tcgcccgcct cagcctccca 3120

aagtgctggt gtgagccacc gtgcccagcc ttacttttta ttttgagagg gggtctcact 3180

ctgtcaccca ggttggagtg cagtggcgcg atctctgctc agtgcaaact ccacctcccg 3240

ggtttaagca gttctcctgt cgtagtctcc tgagtagctg ggattacagg cacaccacca 3300

cggccagcta atttttgtat tttcagtaga gacgggtttc accatgttgc ccaagctggt 3360

ctcgaactcc tggcctcaag tgatctgccc gccttggcct cccagagtgc tgggattaca 3420

ggtgtgagcc accgcacccg gcctcttttt tctttttag tctatcatac cttgcaaata 3480

cagtggttct tcctatgtgt tggttttgat atttatgtaa tcaaacacat cagttttcc 3540

tttctgattt ctgactttgg ggtcatgctg agaaagtcct ttcctacctg aagataatac 3600

agtatatacg tttcttacta gtatttttgt ggatttttaa aatatttaaa tctttagtcc 3660

atctgaactt gttcttctat cagaaatgcc acatttaata aataataagt cccatggtat 3720

cagatggctg gaaggacctc tttcgaaact ttgtttaatt ccattaatct gtgtattctt 3780

attctaatgc taatagttcc acactagctt cctttatctt ttttttcttt tttttttttt 3840

ttttgagctg gagtttcgct cttgttgccc aggctggagt acaatgtcac gatctcggtt 3900

caccgcaacc tccgcctccc aggttcaagc aattctcctg cctcatcctc gcgagtagct 3960

ggaattacag gcatgcgcca ccacgcctag ctatttttgta tttttagtag agatggggtt 4020

tctccatgtt ggtcaggctg gtctcaaact cccagcctca ggtgatctgc ctgcctcggc 4080

ctcccaaaat gctgttatta caggcgtgag ccaccacgcc cagccttcat cttttaatga 4140

atgtacatgt atgtaatctt ttaggtgaac tttttgtaat gttgtgccaa gttccttaaa 4200

aagccctttt ggaagctggg caggtggcca cgcctgtaat cccagcattt tgggaggtctg 4260

aggcaggtgg atcacttgag gccaggagtt caagactagc ctagccaaaa tgcaaaaccc 4320

tgtctctact aaagatacaa aaattagccg gatgcgatgg cacatgcctg taatctcagc 4380

tactcgggag gctgaggtag aagaatcgct tgaaccgggg aggcagaggt tgcagtgagc 4440

aagatggcgc cactgcactc cagcctgggt gacagaggga gactccatct caaaaaaaaa 4500

aaaaaaaaaa aagataaaaa ggaaacctaa gtactcttgg gctttgttaa ggattttgtt 4560

aaatatacaa aggattgcag ggaaaattaa cttatttta atattgagta tgcttatcca 4620

agagcaaaat aatatttctc catttattca aatcatttag gagcatcata gttttaacat 4680

atgggccttg cacgtatctt aaatttatct ctaggcattt taggttgttc agttgttgtt 4740

gtgaatggga tctttttctc caaataggat tattgttgat atctgttgat tatgttaact 4800

ttgtagtttc tgactttact gaactgtctt cttagatcta atactctttt caatttcatc 4860

atatatttct cattcctatt ttgtttgggg tttttagggc gggaatatta acgggataag 4920

agagacaaaa gaaaatctgg aaaaacaatt cattttacct tacattgctt gtgattacta 4980

ccacactatt actgggttgg aaaaaattgt gaaatcccaa ggtgcctaat aaatgggagg 5040

tacctaagtg ttcatttaat gaattgtaat gattattgga atttctcttt cagTGAGAAG 5100

CTCTTCATGG AGATGGCAGA GCTCATGGTC TCAGAAGGCT GGAAGGATGC AGGTTATGAG 5160

TACCTCTGCA TTGATGACTG TTGGATGGCT CCCCAAAGAG ATTCAGAAGG CAGACTTCAG 5220

GCAGACCCTC AGCGCTTTCC TCATGGGATT CGCCAGCTAG CTAATTATgt gagtttatag 5280

ataatgttct tgttcattca gaggactgta agcacttctg tacagaagct tgtttagaaa 5340

cagccctcat ggccgggcgt ggtggctcac gctgtaatcc caacactttg ggaggccgag 5400

gcgggtggat cacctgaggt caagagttca agaccagcct ggccaacatg gtgaaacccc 5460

aactctatta aaagtacaaa aaattagctg ggcatggtgg tgaacgcctg taaccccagc 5520

tacttgggag gctgaggcag gagaatcgct tgaacccagg aggtggaagt ttcagtgagc 5580

tgagatcacg ccattgcact ctagcctggg caacaaaaga gaaactccat ctcaaaaaaa 5640

aaaacaagga aaaaaagaaa cagccctcat gacacttaga aagtagaata gctggctgtt 5700

atctgaacat tgaattgtaa ggcttatcag gtggactttg cattccatca gcagacaatt 5760

tttttttttt tttttttttg agatggagtc tcattctgtc tcccaggctg gagggcagtg 5820

gtgcgatctc ggctcactgc aagctccacc tcctgggttc atgccattct cctgcctcag 5880

cctcccaagt agctgggacc acaggcaccc gccaccatgc ccagttaatt ttttgtattt 5940

ttagtagaga cggggtttca ccatgttagc caagatggtc tcgatctcct gacctcgtga 6000

tccgcccacc tcggcctccc aaagtgctgg gattacaggc atgagccacc gcgcctagcc 6060

tacaaatgtt ttgtaatagc tcttgaggcc catcttggag ttctcctttt gctaaaacca 6120

ctgaactctc taggaggaaa aaggaacttg gttcttgaca tatgtgtgca tgtatttcca 6180

tataacattt aggaagctat tgcaatggta ctataaacta gaattttaga agatagaagg 6240

aaaatattct ggagatcatt gaagagaaat ggagtccaac actagttaaa gatgatgaag 6300

acagattttt ttttttgacg gagtctcgct ctgtcgccca ggctggagtg cagtggcaca 6360

atctcagctc actgcaaccc tccacctctt gggttcaagt gattctcctg cctcagcctc 6420

ccaagtagct gggactacag gcgcacacca ccacgcccgg ctaatttttg tatttttagt 6480

agagacaagg tttcaccata ttggccaggc tggtctcgaa ctcctgacct tgtaatccgc 6540

ccaccttggc ctcccaaagt gctgggatta caggcatgag ccaccacgcc cggccgatga 6600

agacagattt tattcagtac taccacagta gaggaaagag ccaagttcaa ttccaaatac 6660

aacaaagaca ggtggagatt tatagccaat gagcagattg agggggtcag tggatggaat 6720

atttaagaag acatcaaggg tagggagctt cttgctaaag cttcatgtac ttaaacaaga 6780

agggtggggg atgagggaaa ttgatcagat atcaatggtg gcagtattga cttagcagga 6840

ttcttgctaa gaggtcttgc taggacagac ataggaagcc aaggtggagg tctagtcgaa 6900

aagaaggctc atcagagaag tctaactaaa gtttggtcaa gaagagtctt tgtcaaggta 6960

aatctatcat ttccctcaaa aggtaatttt caggatccca tcaggaagat tagcatggct 7020

gctagctttc tcctcagttc tgggctatag ctcacatgcc tagtttgaac tagctcagca 7080

gaactggggg atttattctt tgtcttccaa caaactcatc tggatgattt tgggggtttg 7140

tggggaaaag cccccaatac ctggtgaagt aaccttgtct cttcccccag cctggaatgg 7200

ttctctcttt ctgctacctc acgattgtgc ttctacaatg gtgactcttt tcctccctct 7260

catttcagGT TCACAGCAAA GGACTGAAGC TAGGGATTTA TGCAGATGTT GGAAATAAAA 7320

CCTGCGCAGG CTTCCCTGGG AGTTTTGGAT ACTACGACAT TGATGCCCAG ACCTTTGCTG 7380

ACTGGGGAGT AGATCTGCTA AAATTTGATG GTTGTTACTG TGACAGTTTG GAAAATTTGG 7440

CAGATGgtaa tgtttcattc cagagattta gccacaaagg aaagaacttt gaggccatgg 7500

tagctgagcc aaagaaccaa tcttcagaat tttaaatacc ctgtcacaat actggaaata 7560

attattctcc atgtgccaga gctcccatct cttctctttc agttcattaa ttaattaatt 7620

aattcatgta aaatccatgc atacctaacc atagctaata ttgtgcactt ataattcaag 7680

agggctctaa gagttaatta gtaattgtaa ctctctataa catcatttag gggagtccag 7740

gttgtcaatc ggtcacagag aaagaagcat cttcattcct gcctttcctc aatatacaca 7800

ccatctctgc actacttcct cagaacaatc ccagcagtct gggaggtact ttacacaatt 7860

taagcacaga gcaactgcct gtccctgctg ctagtttaaa catgaacctt ccaggtagcc 7920

tcttcttaaa atatacagcc ccagctgggc atgatggctc atgcctgtaa tcctagcact 7980

ttgggaggct gaggcgggtg gattacttga ggtcaggagt tcgagaccac cctggccaac 8040

atggtgaaac cccatctcta gtaaaaatac aaaaattagc tgactttggt ggcacatgcc 8100

tgtaatccca gctacttggg aagctgagac agaagagtca cttgaacctg ggaaacagag 8160

gttgcagtga gccaagatcg caccactgca ctccaccctg gatgacagac tgaaccccat 8220

ctcaaaaaat taaaataaaa taaaataaaa taactatata tatagcccca gctggaaatt 8280

catttctttc ccttatttta cccattgttt tctcatacag GTTATAAGCA CATGTCCTTG 8340

GCCCTGAATA GGACTGGCAG AAGCATTGTG TACTCCTGTG AGTGGCCTCT TTATATGTGG 8400

CCCTTTCAAA AGgtgagata gtgagcccag aatccaatag aactgtactg atagatagaa 8460

cttgacaaca aaggaaacca aggtctcctt caaagtccaa cgttacttac tatcatccta 8520

ccatctctcc caggttccaa ccacttctca ccatccccac tgctgtaatt atagcctaag 8580

ctaccatcac ctggaaagtc atccttgtgt cttccccttt atttcaccat tcatgtcctg 8640

tctatcaaca gtccttccac cagtatctct aaaatatctc ctgaatcagc ccacttcctt 8700

ccatcttcac tacatgcacc ctggccttcc aagctactat cggctctcaa ccagactgct 8760

gggaccacct gatctctctg cttccactct gtctcaaccc ccatctattt tccaagcagc 8820

actagagtta tcatattaaa atgtaaatat cagttttttt tttaaagaaa aaaaccctga 8880

gacttaacag agttataaaa aatataaatg tcatcatcag ttccctgctt aaaaccctta 8940

actcgcttcc aattgcactt ggaatgaaac caaactgcac tgatccagcc cttgcctgcc 9000

tccccaaagt ccaaggggtc atggctcttt ccctggctac actggttttc tttctgtccc 9060

tcaacactgc aagcctattg ctgccccagg gcctttacac ttgcttttt tctgcctaga 9120

acagttcttc cccaaagatt tttaaagggc cgggctcctt aacattgaag tcgcagacca 9180

aacgccacat atgcagacag ttcttctcta actactttaa aatagccctc tgtccattca 9240

ttcttcatca cattaacctg tttaattttc ttctcagagc tccacactat ttggaagtat 9300

ttgttgactt gttaccatgt ctccccacta gagtgtaagt ttcatgaggg cagggacctt 9360

gtctgacttt gactgtatct ctcgcatatg gttaagtgtt aaatagttat ttatggaatg 9420

aatccctatt attccctcat tatctctgca aaatagtctt ttttctcaac atcttaaacc 9480

tgatatccca cctgcctatc tacaaacttt ttttttgcga cagagtctca ctgtcaccca 9540

ggctagagtg cagtggcgcc atctcggctc actgcaacct ccgcctcccg ggtttaagcg 9600

attctcttgc ctcagcctcc cagtagctgg gattataggc gtgcgctacc acatctggct 9660

aatttttgta ttttagtag agatggtttc accatgttgg ccaggcttgt ctcgaactcc 9720

tgacctcaga tgatccacct gcctcggcct cccaaagtgc tgggattaca ggcatgagcc 9780

accgtgccca gcctctacaa acttttatt ccattaacaa actatatgct gggatttaag 9840

ttttcttaat acttgatgga gtcctatgta attttcgagc ttttaatttt actaagacca 9900

ttttagttct gattatagaa gtaaattaac tttaagggat ttcaagttat atggcctact 9960

tctgaagcaa acttcttaca gtgaaaattc attataaggg tttagacctc cttatggaga 10020

cgttcaatct gtaaactcaa gagaaggcta caagtgcctc ctttaaactg ttttcatctc 10080

acaaggatgt tagtagaaag taaacagaag agtcatatct gttttcacag CCCAATTATA 10140

CAGAAATCCG ACAGTACTGC AATCACTGGC GAAATTTTGC TGACATTGAT GATTCCTGGA 10200

AAAGTATAAA GAGTATCTTG GACTGGACAT CTTTTAACCA GGAGAGAATT GTTGATGTTG 10260

CTGGACCAGG GGGTTGGAAT GACCCAGATA TGgtaaaaac ttgagccctc cttgttcaag 10320

accctgcggt aggcttgttt cctattttga cattcaaggt aaatacaggt aaagttcctg 10380

ggaggaggct ttatgtgaga gtacttagag caggatgctg tggaaagtgg tttctccata 10440

tgggtcatct aggtaacttt aagaatgttt cctcctctct tgtttgaatt atttcattct 10500

ttttctcagT TAGTGATTGG CAACTTTGGC CTCAGCTGGA ATCAGCAAGT AACTCAGATG 10560

GCCCTCTGGG CTATCATGGC TGCTCCTTTA TTCATGTCTA ATGACCTCCG ACACATCAGC 10620

CCTCAAGCCA AAGCTCTCCT TCAGGATAAG GACGTAATTG CCATCAATCA GGACCCCTTG 10680

GGCAAGCAAG GGTACCAGCT TAGACAGgta aataagagta tatattttaa gatggcttta 10740

tatacccaat accaactttg tcttgggcct aaatctattt ttttcccttg ctcttgatgt 10800

tactatcagt aataaagctt cttgctagaa acattacttt atttccaaaa taatgctaca 10860

ggatcatttt aatttttcct acaagtgctt gatagttctg acattaagaa tgaatgccaa 10920

actaacaggg ccacttatca ctagttgcta agcaaccaca ctttcttggt ttttcagGGA 10980

GACAACTTTG AAGTGTGGGA ACGACCTCTC TCAGGCTTAG CCTGGGCTGT AGCTATGATA 11040

AACCGGCAGG AGATTGGTGG ACCTCGCTCT TATACCATCG CAGTTGCTTC CCTGGGTAAA 11100

GGAGTGGCCT GTAATCCTGC CTGCTTCATC ACACAGCTCC TCCCTGTGAA AAGGAAGCTA 11160

GGGTTCTATG AATGGACTTC AAGGTTAAGA AGTCACATAA ATCCCACAGG CACTGTTTTG 11220

CTTCAGCTAG AAAATACAAT GCAGATGTCA TTAAaagact tactttaaaa tgtttatttt 11280

attgccaact actacttcct gtccaccttt ttctccattc actttaaaag ctcaaggcta 11340

ggtggctcat gcctgtaatc ccagcacttt gggaggctga ggcgggcaga tcacctgagg 11400

tcgggacttt gagacccgcc tggacaacat ggtgaaaccc catttctaat aaaaatataa 11460

aaattagcca ggtgtggtgg cgcacctgtg gtcccagcta ctctgggggc tgaggcatga 11520

gaatcgcttg aacccgggag tggaggttgc attgagctga gatcatgcca cctcactcca 11580

gcctgggcaa caaagattcc atctcaaaaa aaaaaaaaaa gccaggcaca gtggctcatg 11640

cctggaatcc cagcactttt ggaagctgag gcaggcagat cacttgaggt taggatttca 11700

agaccagcct ggctaacata gtaaagcccc gtctctacta aaaatacaaa aattagccag 11760

gtatggtggc gagcttctgt agccccagct actcaggaga ctgaggcagg agaatcactt 11820

gaacccggga agtgggggag tgcagtgacc caagatcacg ccactgcatt ccagcctggg 11880

caacagagca agactccatc tcaaaaaaaa aagttctatt tccttgaata aaattttccg 11940

aagtttaaac tttaggaata aaactattaa acccgtattt actcatccag atacccaccc 12000

cccttgttga gattctctcc caattatcaa aatgtgtagc atatttaact accaagagct 12060

aaacatcatt aagactgaaa tgtattaaga aggatgtata ggccaggcac ggtgtctcac 12120

gcctgtaatc ccaacacttt gggaggccaa gtcgggcgga tcacgaggtc aggagatgga 12180

gaccatcctg gccaacatgg tgaaaccccc tctctactaa aaatacaaaa attagccagg 12240

caggtggcag gcacctgtaa tcccagctac tccagaggct gaggcaggac aatcacttga 12300

acctgggagg cagaggctgc agtgagctga ggttgtacca attgcactcc agcctaggta 12360

acgagcaaca ctccatctca aaaaaagaaa aaaaaaaaga tgtataattt ggaactgtta 12420

agaggcattt taaaga

Chapter Three

α-Galactosidase A Gene Rearrangements

ABSTRACT

Six α-galactosidase A gene rearrangements that cause Fabry disease were investigated to assess the role of *Alu* repetitive elements and short direct and/or inverted repeats in the generation of these germinal mutations. The breakpoints of five partial gene deletions and one partial gene duplication were determined by either cloning and sequencing the mutant gene from an affected hemizygote, or by PCR amplifying and sequencing the genomic region containing the novel junction. Although the α-galactosidase A gene contains 12 *Alu* repetitive elements (representing ~30% of the 12 kb gene or ~1 *Alu*/1.0 kb), only one deletion resulted from an *Alu-Alu* recombination. The remaining five rearrangements involved illegitimate recombinational events between short direct repeats of 2 to 6 bp at the deletion or duplication breakpoints. Of these rearrangements, one had a 3' short direct repeat within an *Alu* element, while another was unusual having two deletions of 1.7 kb and 14 bp separated by an 151 bp inverted sequence. These findings suggested that slipped mispairing or intrachromosomal exchanges involving short direct repeats were responsible for the generation of most of these gene rearrangements. There were no inverted repeat sequences or alternating purine-pyrimidine regions which may have predisposed the gene to these rearrangements. Intriguingly, the tetranucleotide CCAG and the trinucleotide CAG (or their respective complements, CTGG and CTG) occurred within or adjacent to the direct repeats at the 5' breakpoints in three and four of the five α-galactosidase A gene rearrangements, respectively, suggesting a possible functional role in these illegitimate recombinational events. These studies indicate that short direct repeats are important in the formation of gene rearrangements, even in human genes like α-galactosidase A that are rich in *Alu* repetitive elements.

INTRODUCTION

In man, studies of the illegitimate recombinational events which cause gene rearrangements have focused on the molecular characterization of 1) somatic rearrangements in the immunoglobulin loci in B- or T- cell leukemias and lymphomas, and 2) germinal rearrangements in various genes which cause inherited diseases (for review, see 1). Most of the germinal gene rearrangements described to date have been deletions, however, several duplications and a complex deletion-inversion have been reported (e.g., 2-4). Insights into the origin of these rearrangements have been derived from the molecular analysis of their novel junctions. To date, the breakpoint junctions of about 45 human germinal gene rearrangements have been determined. Over half of these rearrangements are in the α- and β-globin gene clusters (4-17) and eight are in the LDL-receptor gene (2, 18-24). The remainder include genes in which only one or two germinal rearrangements have been characterized at the molecular level (e.g., 25, 26). These studies have revealed the occurrence of breakpoint clusters or "hot spot regions" in specific genes (6, 27), the frequent involvement of the *Alu* family of short interspersed repetitive elements at breakpoint junctions (especially the *Alu* RNA polymerase III promoter region, 2, 21), and the presence of short direct repeats of 2 to 7 nucleotides at rearrangement termini (e.g., 5, 26).

Although only a few human genes have been investigated intensively, a notion has evolved that germinal rearrangements in genes enriched in *Alu* repetitive sequences have a propensity for breakpoints in these elements, whereas genes in which *Alu* repeats are less frequent tend to have rearrangement breakpoints involving short direct or inverted repeats. For example, in the *Alu*-rich LDL receptor gene (21 *Alu* monomers or dimers identified to date in 38 kb or ~1 *Alu*/1.8 kb, H.H. Hobbs, pers. commun.), six of the eight rearrangements had both breakpoints in *Alu* repetitive elements, while a seventh had a

single breakpoint in an *Alu* repeat (2, 18-24). Of the four analyzed deletions in the ~40 kb α-globin gene cluster (which contains 16 *Alu* repeats or ~1 *Alu*/2.5 kb), one involved an *Alu-Alu* recombination, while each of the others had a breakpoint in an *Alu* element (6). In contrast, of the 20 sequenced rearrangements in the 60 kb β-globin gene cluster (which has 8 *Alu* elements or ~1 *Alu*/7.5 kb), there were no rearrangements due to *Alu-Alu* recombination and only three deletions (or ~15%) had a breakpoint in an *Alu* sequence (4, 5, 7-17). Based on these findings, it might be expected that rearrangements in genes enriched with *Alu* sequences would frequently result from illegitimate recombinational events involving these repetitive elements. However, further understanding of the nature and frequency of the events that cause germinal rearrangements in man requires the analysis of rearrangement junctions in additional genes.

The isolation and characterization of the full-length cDNA and entire genomic sequence encoding α-Gal A has facilitated the identification and analysis of gene rearrangements causing Fabry disease. Notably, the genomic sequence contains 12 *Alu* repetitive elements dispersed among its six introns and 3' flanking region, representing about 30% of the gene or ~1 *Alu*/1.0 kb. Previous studies of 130 unrelated Fabry families revealed six different germinal rearrangements of the α-Gal A gene (28). Of these, five were partial deletions which ranged in size from 0.4 to at least 4.8 kb, while the sixth was a partial duplication of 8.1 kb. None of these gene rearrangements expressed a functional mRNA and no α-Gal A activity was detected in affected males (28). Interestingly, four of the five deletions had a breakpoint that mapped within intron 2, a region that contained three *Alu* elements. In this chapter, the precise characterization of these six germinal rearrangements which were analyzed by cloning and sequencing the mutant gene from an affected hemizygote or by polymerase chain reaction (PCR) amplification and sequencing of the genomic region containing the breakpoint junction is described. Notably, only one rearrangement was the result of an *Alu-Alu* recombinational event, even though the gene

was enriched with these repeat elements. Instead, five of the rearrangements had short direct repeats flanking the deleted or duplicated sequences. Among these, certain tri- and tetra- nucleotides occurred more frequently than expected within or immediately adjacent to the 5' breakpoints, but not in or near the 3' breakpoints.

EXPERIMENTAL PROCEDURES

Materials— Restriction endonucleases, the Klenow fragment of DNA polymerase I, T4 polynucleotide kinase, M13 cloning vectors and universal sequencing primers were purchased from New England Biolabs. T4 DNA ligase and pUC vectors were from International Biotechnologies Inc. SP6 and T7 promoter primers and pGEM4 DNA were obtained from Promega. *Taq* polymerase was purchased from Cetus. DNA synthesis reagents and β-cyanoethyl-diisopropylphosphoramidites were from Applied Biosystems and radioisotopes were from Amersham. The λEMBL3 vector and the Gigapack *in vitro* packaging system were obtained from Stratagene. The GemSeq Klenow System was from Promega and Sequenase was purchased from United States Biochemical Corp.

Genomic Cloning— High molecular weight DNA was isolated from cultured skin fibroblasts of an affected hemizygous male from Fabry Family A and from cultured lymphoblast lines of affected hemizygotes from Families B and J (29). The DNA was partially digested with *Mbo*I, size fractionated in a 0.9% agarose gel and 13-22 kb fragments were ligated into the λEMBL3 replacement vector (30) which had been digested to completion with *Bam*HI and *Eco*RI, and then packaged with the Gigapack Plus extracts. The unamplified libraries were screened in *E. coli* strain K802 using the nick-translated full-length α-Gal A cDNA insert pcDAG126 after transfer to nitrocellulose filters (31). Recombinant clones containing the rearranged alleles were obtained from the respective library and were mapped by restriction analyses. Following electrophoresis in agarose,

fragments containing the deletion junctions were isolated and subcloned into M13 or pGEM vectors for sequencing.

Amplification of deletion/duplication junctions— DNA sequences that flank the rearrangement junctions in Fabry Families D, E and F were amplified by the PCR technique (32). Sense and antisense primers containing restriction endonuclease cleavage sites and an additional 6 to 8 nt 5' to the restriction sites (to facilitate enzymatic cleavage) were synthesized on an Applied Biosystems Model 380B DNA synthesizer. For Fabry Family E, a 433 bp region containing the deletion junction was amplified. The sense primer (5'-TCTTGCTAAAGCTTCATGTACTTAA-3') corresponded to 25 nt of intron 2 (nt 6750-6774) and contained a unique *Hind*III restriction site. All numbered coordinates refer to nucleotide positions in the normal α-Gal A gene (Fig. 1, Chapter 2). The antisense primer (5'-ACGTACGTGAGCTCTGGCACATGGA-3') was complementary to 17 nt of intron 3 (nt 7568-7584) and contained a *Sac*I restriction site. For Family F, amplification of a 793 bp region containing the recombination junction was accomplished using a sense primer (5'-ACTACTGAGCTCTTGAGGCCCATC-3') constructed to 17 nt of intron 2 (nt 6078-6094) with an additional 7 nt that established a *Sac*I site, and an antisense primer (5'-ACTACTGAATTCGTTTAAAGGAGGCAC-3') complementary to intron 4 (nt 10054-10070) which also contained an *Eco*RI site. Using DNA isolated from a Family D hemizygote, a 1.4 kb region surrounding the duplication junction was amplified using a sense 30-mer (5'-ACTACTGAGCTCTGGGTCATCTAGGTAACT-3') constructed to 18 nt of intron 5 (nt 10441-10458) and containing a *Sac*I recognition sequence, and an antisense 30-mer (5'-ACTACTGAATTCATTAGAATAAGAATACAC-3') which was complementary to 19 nt of intron 1 (nt 3771-3789) and had an *Eco*RI restriction site.

Genomic DNA from affected hemizygotes was amplified by a modification of the method of Saiki et al. (32). Briefly, the amplification mixtures contained 1 μg of genomic DNA, 50 mM KCl, 10 mM Tris-HCl, pH 8.3, 1.5 mM $MgCl_2$, 0.01% gelatin, 200 μM

each of dATP, TTP, dGTP, dCTP and 1.0 μM each of sense and antisense primers in a reaction volume of 100 μl. Each reaction mixture was incubated initially at 94 °C for 7 min followed by the addition of 2.5 U of *Taq* polymerase. Thirty amplification cycles of denaturation at 94 °C for 1 min, nonstringent annealing at 37 °C for 2 min and extension at 60 °C for 5 min were performed with the Perkin-Elmer/Cetus Thermal Cycler. After amplification, the authenticity of each PCR product was verified by Southern hybridization to 17-mers synthesized to genomic sequences internal to the amplification primers. The products were then cleaved with the appropriate restriction endonucleases and subcloned into M13 mp18 and mp19 vectors (33).

DNA sequencing— The single-stranded M13 DNA templates containing the rearrangement junctions were sequenced in both orientations by the dideoxy chain termination method of Sanger (34) using universal M13 primers or synthetic oligomers constructed to α-Gal A sequences. Sequences were determined from multiple clones to control for the possible misincorporation of nucleotides by *Taq* polymerase. Plasmid template DNA was isolated by the rapid boiling method of Wang et al. (35). Double-stranded sequencing of plasmid DNAs in both orientations was accomplished using the GemSeq Klenow System or Sequenase according to the manufacturers' instructions. Nucleotide sequences were aligned using an IBM AT computer with the Microgenie program (Beckman).

RESULTS

Characterization of α-Gal A gene rearrangements— The structure of the normal α-Gal A gene and the positions of the five previously identified partial gene deletions and the partial gene duplication are shown in Figure 1. Since the locations of the 5' or 3' breakpoints in Families A, B and J were not identified by restriction analyses using the full-

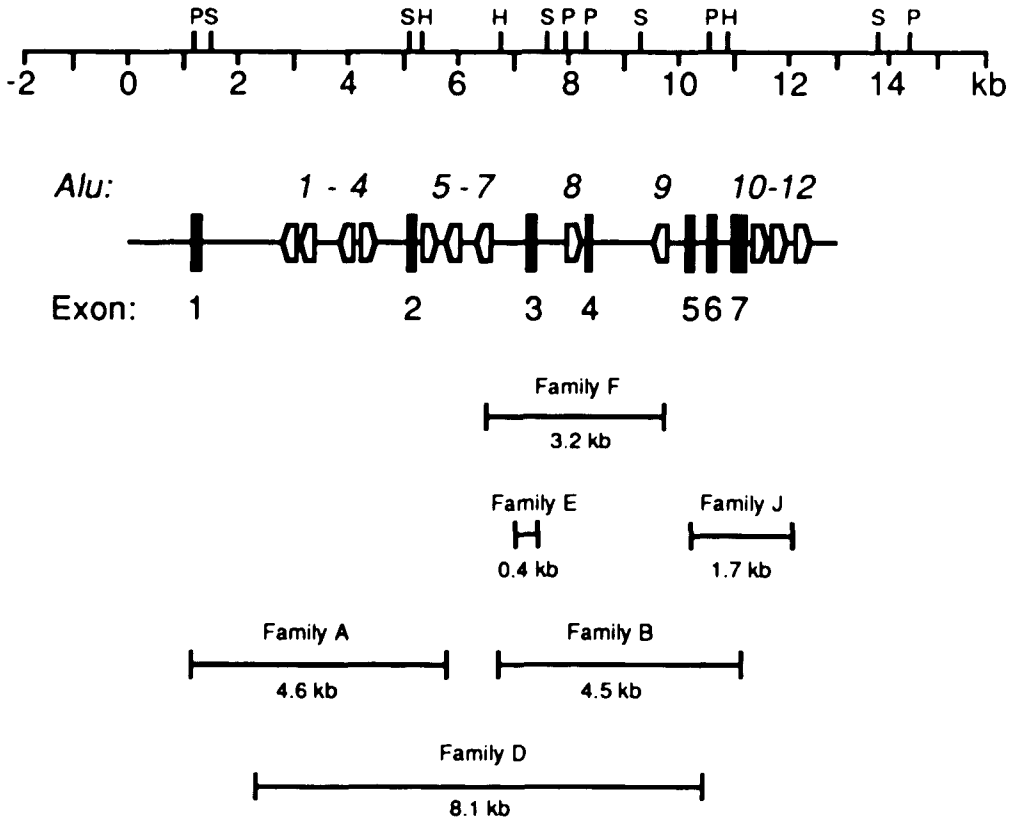


Figure 1. The structure of the α-Gal A gene and the locations of six intragenic rearrangements causing Fabry disease. The α-Gal A nucleotide sequence coordinates as numbered for the entire 12,436 bp sequence (Chapter 2, Fig. 1) and *Hind*III (H), *Pvu*II (P), *Sac*I (S) restriction maps are shown. Exons 1 through 7 are represented as solid rectangles and the positions and orientations of the 12 *Alu* repeat elements are indicated by pentagonal figures. The locations and lengths of the deletions in Fabry Families A, B, E, F and J and the duplicated region in Fabry Family D are shown.

length cDNA as a probe, the mutant alleles were isolated from genomic libraries constructed in λEMBL3 with DNA obtained from cultured cells of an affected hemizygote from each family. In Fabry Families D, E and F, sequences flanking the deletion or duplication junctions were PCR-amplified from genomic DNA of affected hemizygotes. Sequencing of the novel breakpoint junctions revealed that all six α-Gal A gene rearrangements were intragenic (Figs. 1-8). Interestingly, the rearrangement in Family J was complex involving two deletions and an inversion (Fig. 8). Four of the six gene rearrangements had one breakpoint in intron 2 which contains three *Alu* repetitive elements within a 1.3 kb region. However, only two breakpoints occurred in the intron 2 repetitive elements. In fact, of the 14 rearrangement breakpoints, only three occurred within *Alu* repetitive elements. Fabry Family F had both breakpoints within *Alu* repetitive elements, whereas Family A had one breakpoint within an *Alu* repeat. The other rearrangements had short direct repeats at their breakpoint termini. Computer-assisted analysis of the entire α-Gal A genomic sequence for dyad symmetry by the SEQ program (36) predicted that stable secondary structures could be formed by a CT-rich sequence of 84 bp in intron 1 (nt 1603-1686) pairing with a GA-rich region in the 5' flanking sequence (nt 626-692) and by *Alu* repeats in opposite orientations. However, there were no stable secondary structures (i.e., inverted repeats or dyads) or alternating purine-pyrimidine stretches (37) at or near the breakpoints that could have facilitated the formation of these rearrangements.

Deletion involving Alu-Alu recombination— The breakpoints of the 3.2 kb partial deletion in Fabry Family F were determined by sequencing a PCR-amplified region of genomic DNA which spanned from the first third of intron 2 into intron 4 (nt 6078-10070 in the normal gene) (Figs. 1-3). Sequence analysis of the ~800 bp PCR product revealed that the 5' breakpoint was in the left arm of *Alu* 7 in intron 2, and the 3' breakpoint was in the left arm of *Alu* 9 in intron 4 (Fig. 2, 3). Both *Alu* 7 and *Alu* 9 were in the antisense orientation, and both breakpoints were in identical 38 bp regions of their

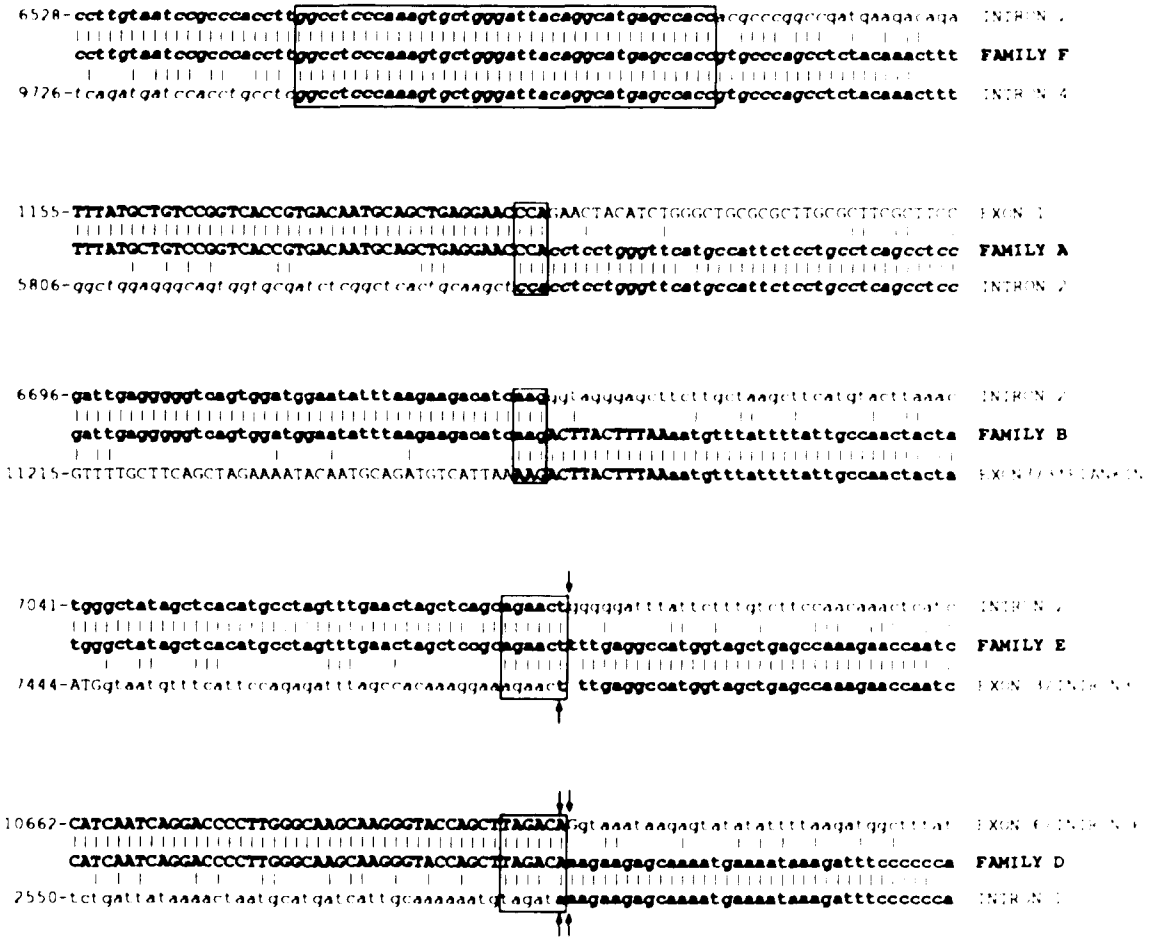


Figure 2. Nucleotide sequences across the breakpoint junctions in four partial deletions and the partial duplication of the α-Gal A gene. Exonic sequences are in uppercase, intronic sequences are in lowercase, and *Alu* repetitive sequences are italicized. For each rearrangement (middle sequence), the corresponding normal 5' and 3' sequences are shown above and below, respectively. The nucleotides in the rearranged alleles are in bold. Vertical lines indicate positions at which the sequences are identical. The regions of homology at the crossover site in Fabry Family F and the short direct repeats present at the rearrangement termini in Fabry Families A, B, D and E are boxed. The precise 5' and 3' breakpoints in Families D and E are indicated by arrows.

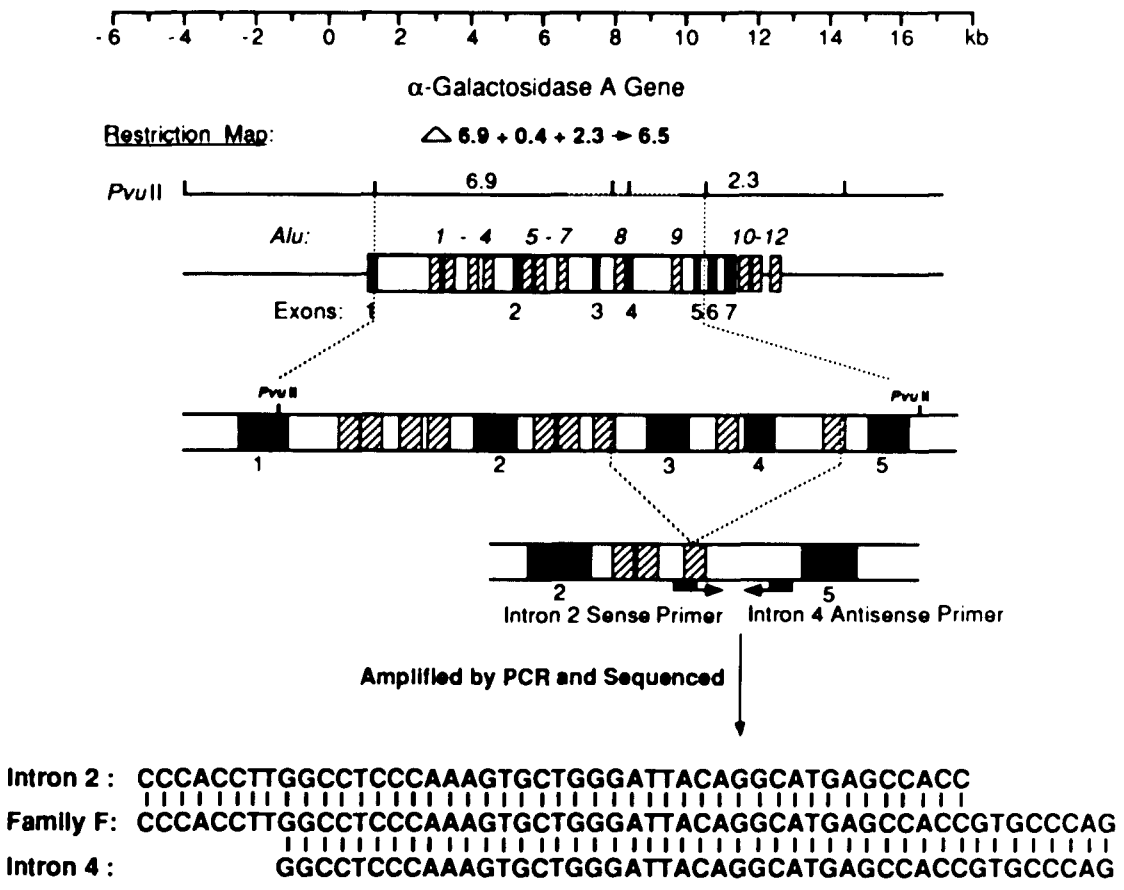


Figure 3. Analysis of the deletion in Fabry Family F. The 3.2 kb partial gene deletion was identified by Southern hybridization studies and the breakpoints were mapped to introns 2 and 4 using several restriction endonucleases (28). The restriction map for *PVU*II is shown (top). The deletion breakpoints indicated in the middle panel by broken lines were determined by sequencing genomic DNA from an affected hemizygote that was PCR-amplified using sense and antisense primers constructed to intron 2 and intron 4 sequences, respectively. The deletion involved recombination between two similarly oriented *Alu* repeat elements, one in intron 2 and the other in intron 4. The mutant allele and the DNA sequences in the *Alu* repeats at the breakpoints are shown below. Note the 38 bp of identity shared between the two *Alu* sequences at the site of recombination.

respective *Alu* RNA polymerase III promoter (38). The mutant allele contained a complete novel *Alu* element that presumably arose from an unequal cross-over in the 38 bp common region of the two *Alu* repeats.

Recombination involving short regions of homology— The mutant allele in Fabry Family A was isolated by screening a library of $1x10^6$ recombinants constructed with genomic DNA from an affected hemizygote. The novel 2.1 kb *Pvu*II fragment, identified previously by Southern hybridization analysis (28), was subcloned into pGEM4 and sequenced. The 5' breakpoint was in exon 1 whereas the 3' breakpoint was in the right arm of *Alu* 6 in intron 2 (Figs. 1, 2, 4). This partial deletion eliminated 4651 bp including 179 bp of exon 1, all of intron 1 and exon 2, and 577 bp of intron 2. The breakpoint junction was localized to the triplet, CCA, which was found at both the 5' and 3' deletion breakpoints (i.e., nt 1195-1197 in exon 1 and nt 5846-5848 in *Alu* 6), with only one triplet repeat being retained in the novel junction. The precise breakpoints of this recombinational event occurred immediately 5', 3' or within these trinucleotide direct repeats.

The breakpoints of the 4.5 kb deletion in Fabry Family B were identified by isolating the mutant allele from a genomic library of 4×10^6 recombinants constructed with DNA from an affected hemizygote (Figs. 1, 5). The 2.5 kb *Sac*I fragment containing the deletion junction was subcloned into pGEM4 and sequenced with primers corresponding to selected regions of intron 2. As shown in Figures 2 and 5, the 5' breakpoint was in intron 2 and the 3' breakpoint was in exon 7. The 4519 bp deletion eliminated 533 bp of intron 2, all of exons and introns 3 to 6 and 277 bp of exon 7. The direct repeat, AAG, was present at both deletion termini; however, only one AAG repeat was retained in the novel junction. The precise breakpoints occurred immediately 5', 3' or within these direct repeats.

The partial α-Gal A gene deletion in Fabry Family E was localized by restriction mapping to a region of about 400 bp which included all of exon 3 (Figs. 1, 6). The

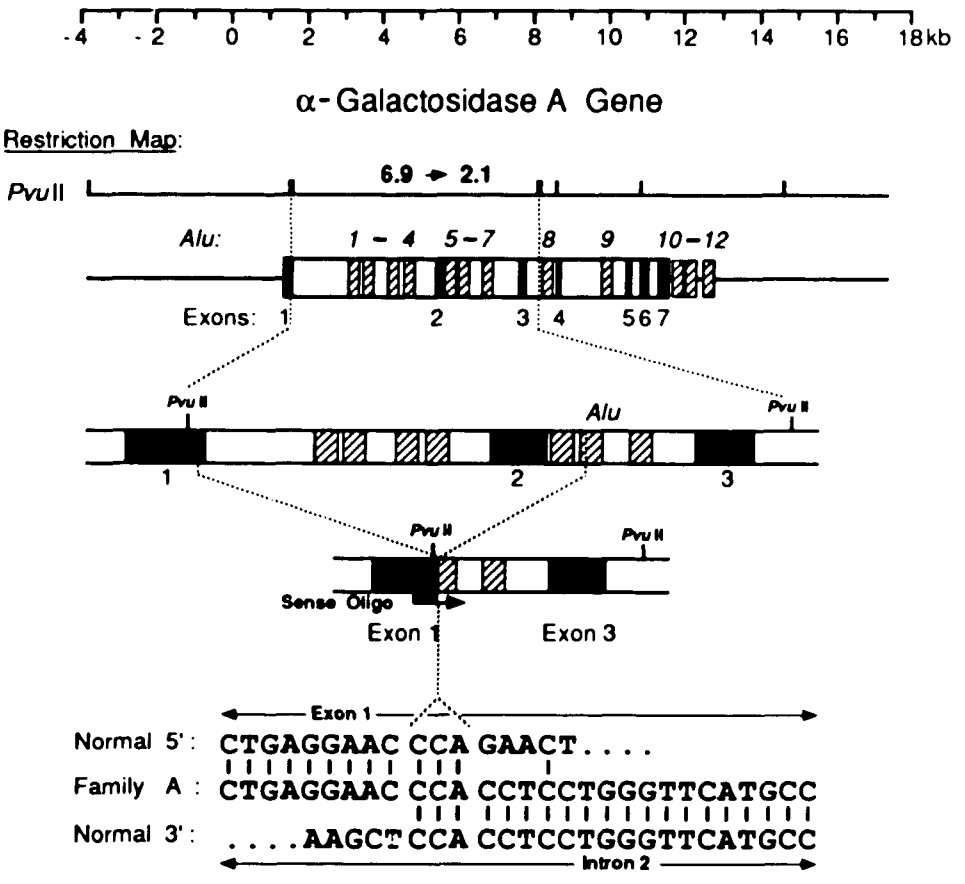


Figure 4. Characterization of the deletion in Fabry Family A. Restriction mapping of genomic DNA from affected hemizygotes of this family revealed the absence of the normal α-Gal A 6.9 kb *Pvu*II fragment and the presence of a new 2.1 kb fragment, indicating a deletion of about 4.8 kb (28). The deletion breakpoints were localized to exon 1 and intron 2 as indicated by the dashed lines in the middle panel. To determine the precise breakpoint junction, the mutant allele was isolated as described in the "Methods". A putative positive clone was isolated and characterized by restriction analysis. The novel 2.1 kb *Pvu*II fragment was subcloned into pGEM4 and the breakpoint junction was sequenced. The sequence across the deletion junction and the corresponding normal sequences are shown (bottom). Note the CCA trinucleotide present at the deletion termini.

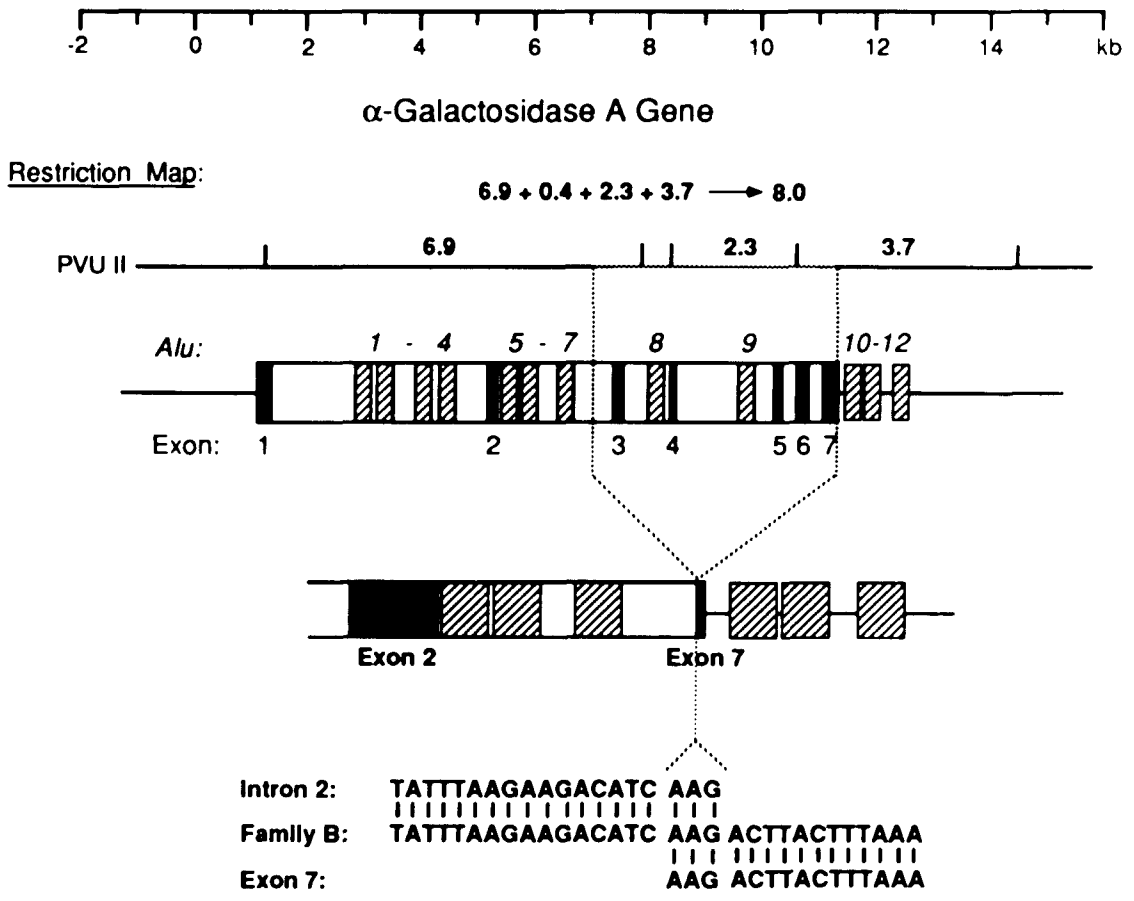


Figure 5. Fabry Family B Rearrangement. The 5' breakpoint in Fabry Family B was mapped to intron 2 by Southern hybridization analysis, but the 3' breakpoint could not be determined using α-Gal A cDNA as a probe (28). The precise breakpoints were identified by sequencing a clone obtained from a Family B genomic library. As shown by vertical dashed lines on the gene map, the 5' and 3' breakpoints were localized to intron 2 and exon 7, respectively. The rearranged allele is depicted below followed by the sequences at the deletion junction.

deletion breakpoints were determined by PCR amplification and sequencing of the genomic region flanking the deletion (nt 6750-7584). The partial deletion was 402 bp and included 183 bp of intron 2, all of exon 3 and 41 bp of intron 3. The hexanucleotide direct repeat, AGAACT, was found at both the 5' and 3' deletion breakpoints (Figs. 2, 6). However, only the 5' repeat was retained in the novel junction; apparently the pentanucleotide, AGAAC, from the 3' direct repeat had been deleted by the recombinational event. The precise breakpoints occurred immediately after the T in the 5' repeat and before the T in the 3' repeat to account for the presence of the TTT sequence immediately following the direct repeat in the novel junction.

In Family D, Southern hybridization analysis identified an 8.1 kb partial α-Gal A gene duplication based on the presence of unique 4.0 kb *Sac* I and 5.5 kb *Pvu*II fragments in addition to all of the normal restriction fragments (28). The partial duplication appeared to have resulted from a recombinational event between regions in introns 1 and 6 (Figs. 1, 7). The regions flanking the putative duplication joint were PCR-amplified with primers corresponding to sequences in introns 1 and 5, resulting in a 1.4 kb PCR product. As shown in Figures 2 and 7, the sequence 5' to the recombinational joint was identical to that of exon 6 up to the breakpoint at position 10706; the sequence 3' to the breakpoint was identical to that of intron 1 beginning at position 2595. Thus, an 8,112 bp region of the α-Gal A gene was duplicated including 2,500 bp of intron 1, all of exons and introns 2 through 5, and 197 bp of exon 6. At the duplication breakpoints there were short regions of homology, TAGACA and TAGATA, in exon 6 and intron 1, respectively. Since the duplication junction retained the sequence TAGACA, the precise breakpoint had to occur after the fifth or sixth base of the hexanucleotide repeats.

Double deletion/inversion rearrangement— To determine the precise breakpoints of the partial gene deletion in Fabry Family J, the mutant allele was isolated by screening a library of 3.6×10^5 recombinants constructed with genomic DNA from the affected

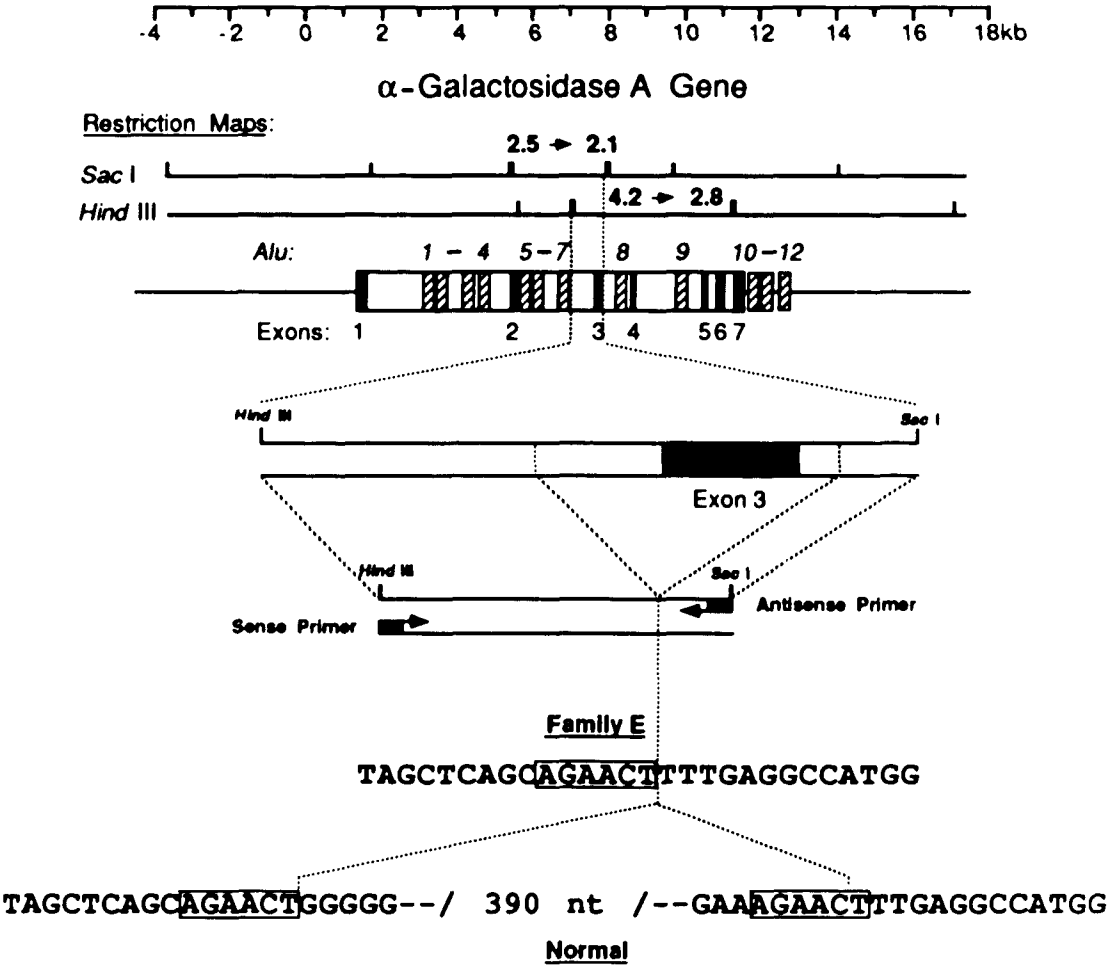


Figure 6. Characterization of the deletion in Fabry Family E. Restriction mapping of genomic DNA from an affected hemizygote revealed the absence of the normal α-Gal A 2.5 kb *Sac*I fragment and the presence of a new 2.1 kb fragment indicating a deletion of 400 bp. In addition, the absence of the 4.2 kb *Hind*III fragment and the presence of a new 2.8 kb fragment localized the deletion to the region between the *Hind*III and *Sac*I sites flanking exon 3 (28) indicated by the dashed lines. This region was amplified from genomic DNA of an affected hemizygote using sense and antisense oligonucleotide primers; the sense primer was constructed to intron 2 sequences and contained the unique *Hind*III restriction site and the antisense oligo was complementary intron 3 and contained the *Sac*I restriction site. Following 30 cycles of PCR amplification, the product was digested with *Sac*I and *Hind*III and subcloned into M13 vectors for sequencing. The sequences from a Family E hemizygote and a normal individual are shown. A region of 402 nt was deleted. As denoted by boxes, the hexanucleotide direct repeat, AGAACT was present at the 5' and 3' breakpoints. Note the rearrangement in this family eliminated the 397 nt between the hexanucleotide direct repeats as well as 5 nt from the second repeat.

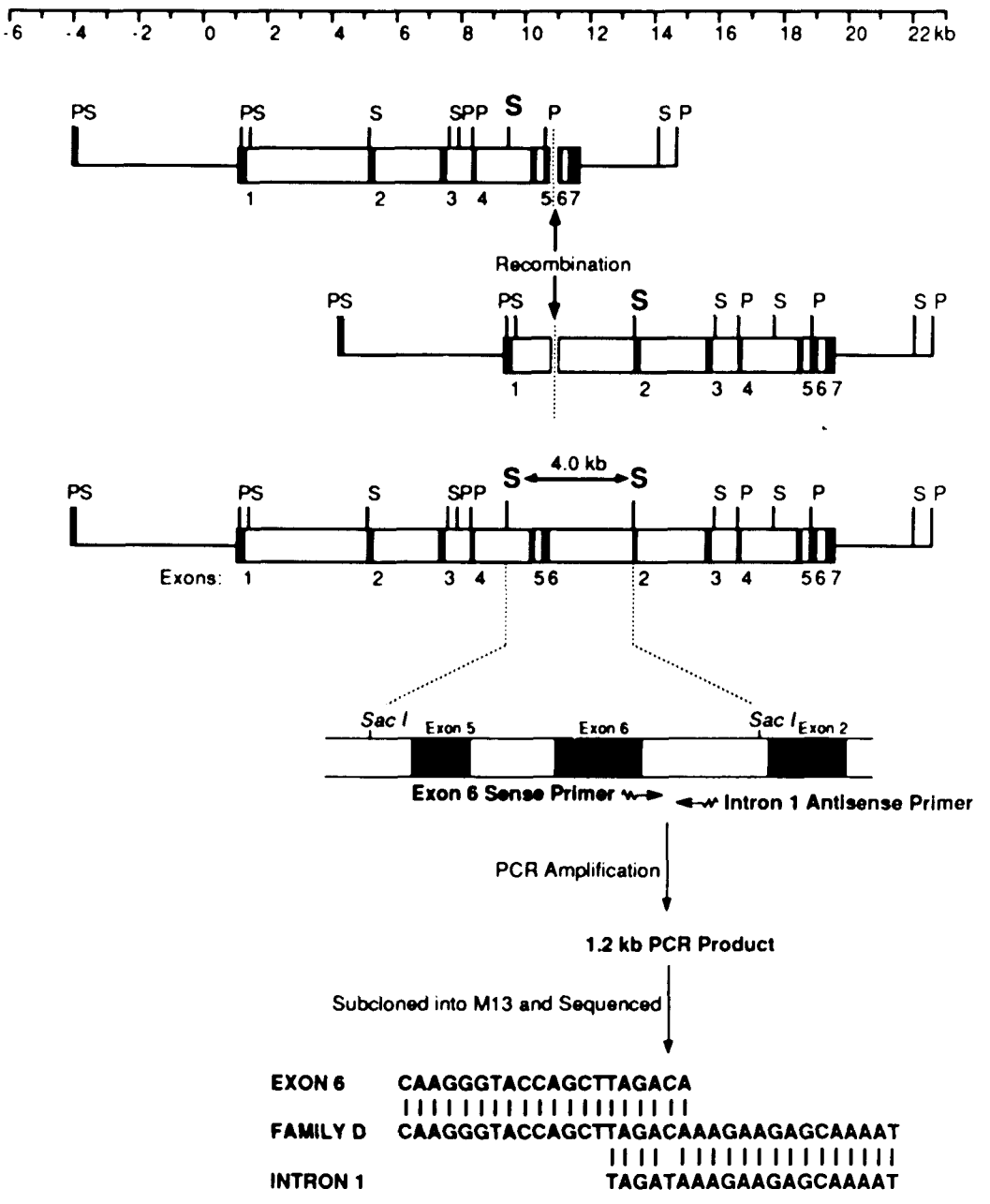
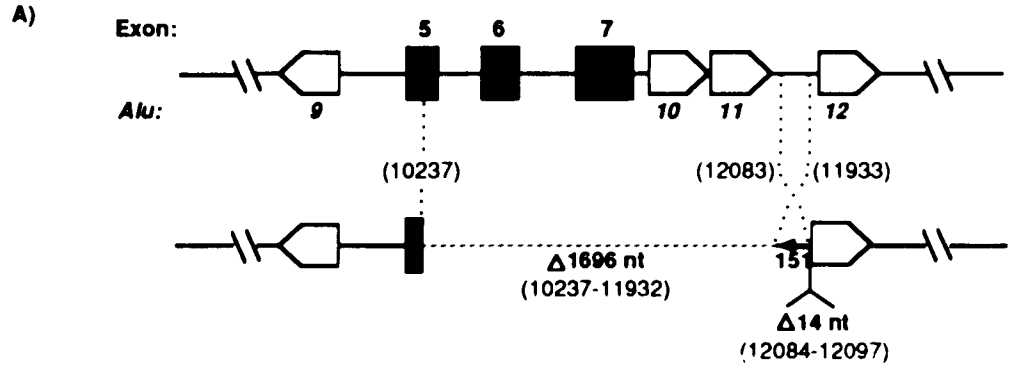


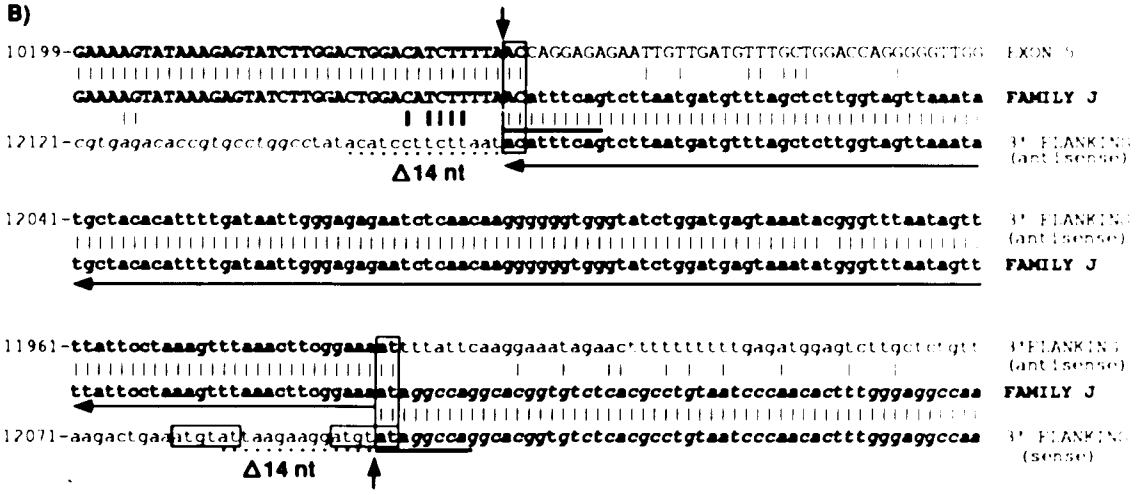
Figure 7. Characterization of the partial gene duplication in Fabry Family D. The presence of the unique 4.0 kb *Sac*I fragment and densitometric analysis of the fragments from the Family D hemizygotes and heterozygotes mapped the breakpoints to introns 6 and 1 (28). The recombinational event is shown diagrammatically. The region amplified by PCR and the information obtained from sequencing the PCR product is shown. Note, the imperfect hexanucleotide repeats at the duplication termini.

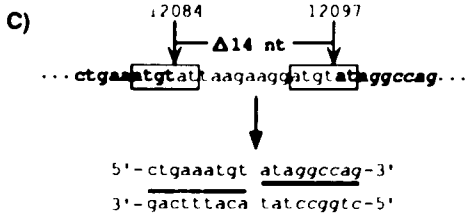
hemizygote. The fragment containing the deletion junction was subcloned into M13 vectors and sequenced. As depicted in Figure 8A, the rearrangement was complex and involved two deleted regions: 1) a 1696 bp deletion (nt 10237-11932) eliminating 56 bp of exon 5, all of exon 6, intron 6, exon 7 and 664 bp of the 3' flanking region, and 2) a deletion of 14 bp (nt 12084-12097) which was 3 bp upstream from *Alu* 12 (12101-12398). Notably, the 151 nt sequence (nt 11933-12083) between the two deleted regions was inverted in the rearranged allele (Fig. 8A and 8B). The dinucleotide direct repeat AC was present at both breakpoints of the 1.7 kb deletion (Fig. 8B). In addition, the hexanucleotide direct repeat ATGTAT occurred at both breakpoints of the 14 bp deletion (Fig. 8C). In the novel junctions of both deletions, only one copy of the respective repeat was retained. Interestingly, the inverted 151 bp sequence that separated the two deletions contained the AC direct repeat and the AT dinucleotide of the ATGTAT direct repeat at its 5' and 3' ends, respectively.

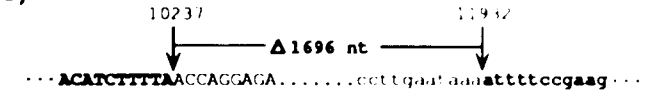
PCR amplification and sequencing of the genomic region from the affected hemizygote which contained the deleted-inverted sequence confirmed the above findings and eliminated the possibility that the mutant allele was further altered by library construction or subcloning procedures. In addition, efforts were directed to determine if the inverted sequence and/or the 14 bp deletion was present in genomic DNA from lymphoblasts of the maternal grandfather (shown by RFLP analysis to be the source of the new mutation; K. Astrin and R. Desnick, unpublished results), since his α-Gal A gene appeared normal as assessed by Southern hybridization analyses (28). Genomic DNA isolated from the maternal grandfather was PCR-amplified using a sense primer from the 3' flanking region (nt 11286-11302) and primers in both orientations within the inversion (nt 12056-12078). A PCR product was obtained only with the antisense primer complementary to the normal 3' flanking sequences, ruling out the presence of the inversion. In addition, the maternal's grandfather's genomic DNA was amplified with the

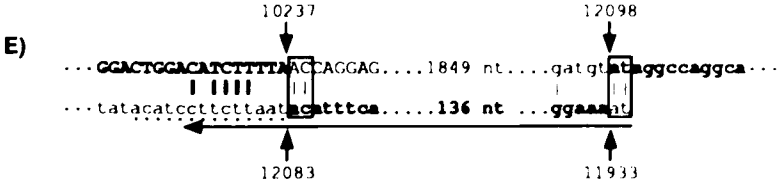
Figure 8. The complex α-Gal A gene rearrangement in Fabry Family J. (A) Schematic of the normal gene with exons 5 through 7 and *Alu* repeat elements 9 to 12 indicated (above) and the Family J rearrangement showing the locations of the 1696 and 14 bp deletions, and the 151 bp inversion (below). The nucleotides are numbered according to the genomic sequence (Chapter 2, Fig. 1) and vertical dotted lines indicate the presumed breakpoints. (B) The nucleotide sequence of the Family J rearrangement with the corresponding normal 5' and 3' sequences shown above and below, respectively. The 151 bp inversion is shown by the long arrows below the sequence, and the nucleotides of the 14 bp deletion (Δ 14 nt) on both sense and antisense strands are indicated by dots below the deleted nucleotides. Direct repeats are boxed and the regions surrounding the 14 bp deletion are shown by bold underlines. Other symbols are as in Fig. 2. (C) Nucleotides surrounding the 14 bp deletion with boxed hexanucleotide direct repeats, and the breakpoints indicated by short arrows. The resulting deleted sequence on both strands is shown below. (D) Nucleotides surrounding the 1696 bp deletion and deduced breakpoints. (E) Alignment of normal sequences (top) with the sequences from the inverted 3' flanking region (below). Note the presence of dinucleotide direct repeats at both ends of the inversion as well as the region of homology between the 5' sequence and the nucleotides excised from the 14 bp deletion (bold vertical lines). It is possible that these excised nucleotides were looped out by misalignment of the hexanucleotide direct repeats (Fig. 3C) and facilitated the juxtaposition of the regions involved in this complex rearrangement.


A)
Exon:
5
6
7
Alu:
9
10
11
12
(10237)
(12083)
(11933)
Δ1696 nt
(10237-11932)
Δ14 nt
(12084-12097)


B)
EXON 5
FAMILY J
3' FLANKING (antisense)
Δ14 nt
3' FLANKING (antisense)
FAMILY J
3'FLANKING (antisense)
FAMILY J
3' FLANKING (sense)
Δ14 nt


C)
12084
12097
Δ14 nt
5'-ctgaaatgt ataggccag-3'
3'-gactttaca tatccggtc-5'


D)
10237
Δ1696 nt


E)
10237
12098
136 nt
12083
11933

above sense primer and an antisense primer that included the 14 bp deletion at the 3' end (5'-GCCTATACATCCTTCTTAAT-3'). These studies did not reveal the presence of the inversion or the 14 bp deletion (data not shown).

DISCUSSION

Recent molecular studies have identified illegitimate *Alu-Alu* recombinational events as important causes of germinal gene rearrangements in a variety of human inherited disorders (6, 25, 27, 39, 40). Of the ten sequenced *Alu-Alu* rearrangements, the breakpoints occurred primarily in the internal RNA polymerase III promoter region in the left arm of the *Alu* sequence, indicating that this region was predisposed to breakage possibly by a conformational change occurring during transcription (21). The fact that the human α-Gal A gene is particularly enriched for *Alu* repetitive elements (12 *Alu* sequences or ~30% of the 12 kb gene) suggested that unequal cross-over events between *Alu* repeats might be the major type of molecular lesion causing Fabry disease. However, only 5% of the 130 unrelated Fabry families analyzed had gene rearrangements (28), a frequency similar to those in other X-linked diseases involving genes of 10 to 186 kb (41-44). Initial restriction mapping of these six α-Gal A gene rearrangements revealed that four of the five deletions had a breakpoint in intron 2 which contained multiple *Alu* sequences (28). Knowledge of the entire α-Gal A genomic sequence permitted the characterization of the precise breakpoints in all six rearrangements. Each breakpoint in the intron 2 cluster occurred at a different site. Moreover, the *Alu* sequences in intron 2 were involved in only two breakpoint junctions, and only the deletion in Family F resulted from an *Alu-Alu* recombinational event with breakpoints in the polymerase III promoter regions of *Alu* 7 and *Alu* 9 (Figs. 1-3). This rearrangement may have been facilitated by the fact that there were 38 identical nucleotides in the breakpoint regions, a high level of homology even among the

Alu family of repeats (45). It is also possible that the alignment of *Alu* elements in intron 2 with oppositely-oriented *Alu* sequences in this or in other regions could form structural intermediates which are prone to recombinational events.

In contrast, five of the six gene rearrangements causing Fabry disease involved short direct repeats of 2-6 bp (Figs. 2-8). The occurrence of short direct and/or inverted repeats at the breakpoints of gene rearrangements has been well-documented in both prokaryotic and eukaryotic systems (e.g., 5, 46-50). These studies indicated that the misalignment of short direct repeats during replication or recombination was an important mechanism responsible for the generation of gene rearrangements. In addition, short inverted repeats may be involved in the alignment and stabilization of structural intermediates in the deletion process (e.g., 50, 51). In man, short direct repeats of 2 to 7 bp have occurred at the breakpoints in about 40% (17 of 42) of reported germinal gene rearrangements (2-26, 39, 40, 52-56). In addition, short direct repeats have been found in three out of six human gene rearrangements in which one breakpoint occurred in an *Alu* sequence (6, 8, 11, 20), as was the case for the rearrangement in Fabry Family A.

The mechanisms proposed to account for the generation of gene rearrangements involving short direct repeats include: 1) errors of DNA replication by the "slipped mispairing" of direct repeats; 2) unequal crossing over between the short regions of homology on different chromatids; 3) intrachromosomal excision of the region between aligned direct repeats releasing a circular DNA fragment, and 4) improper rejoining of DNA after double-stranded breakage events (for review, see 1). Among these possibilities, attention has focused on the "slipped mispairing" mechanism as an important cause of germinal deletions and duplications since the short regions of homology may not be sufficient to promote the misaligning of chromosomes for unequal crossing over during meiosis (5). As illustrated in Figure 9, slipped mispairing can occur during replication when the DNA becomes single-stranded (5). Although single-stranded regions as long as

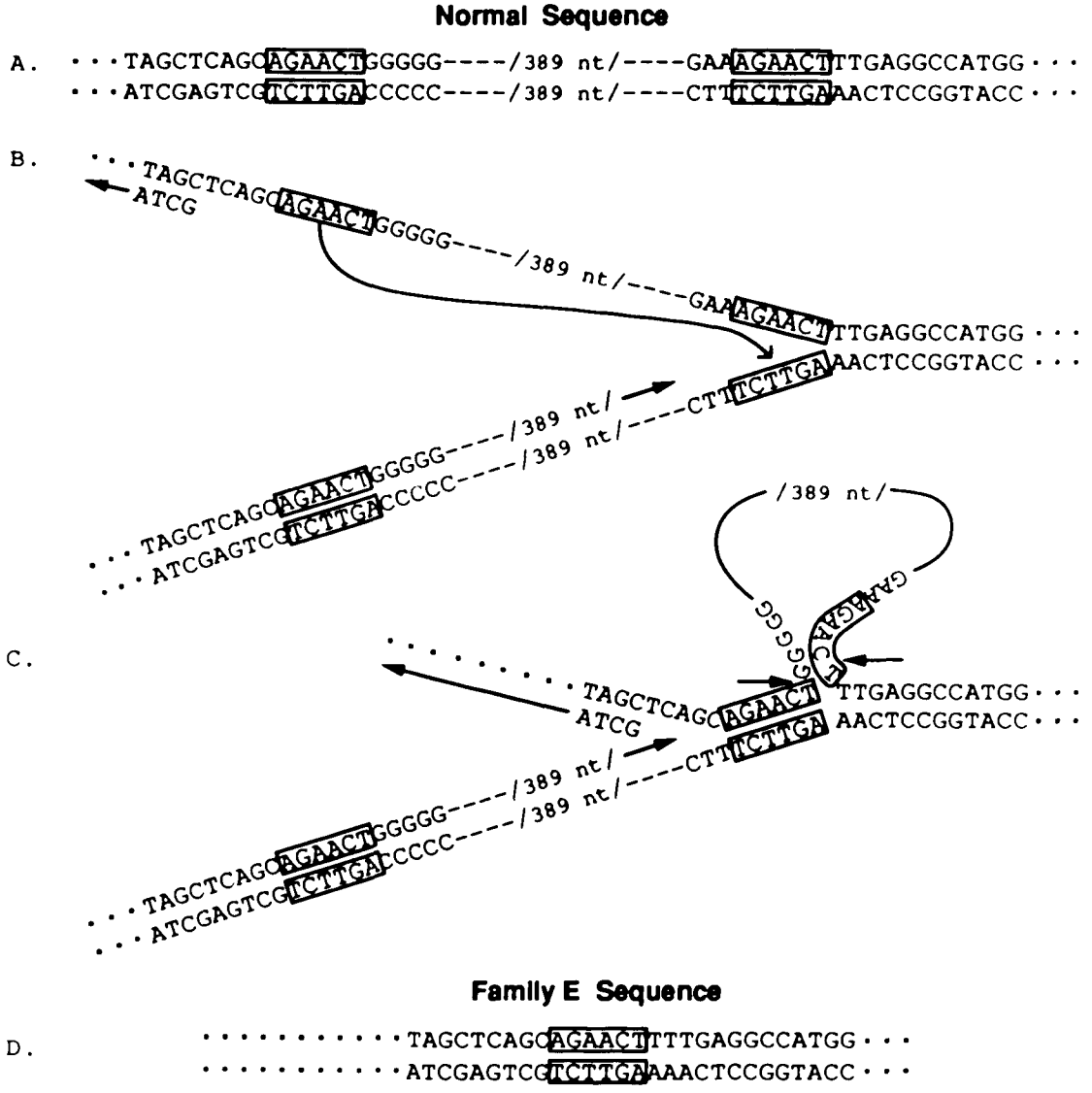


Figure 9. Schematic representation of the slipped mispairing mechanism for the generation of the deletion in Fabry Family E (after 5). (A) Normal sequence with the hexanucleotide direct repeats indicated in boxes. (B) During replication, the two single-stranded repeat elements can mispair leading to the formation of a loop structure containing one of the repeats and the intervening sequence between the repeats. (C) To generate a deletion, the loop can be excised by DNA repair enzymes or would be eliminated after another round of replication. (D) Resulting Family E sequence with only a single repeat element retained adjacent to the deleted 402 bp sequence.

100 kb have been reported in human cells undergoing replication (57), mispairing of direct repeats may occur between single-stranded stretches of DNA separated by a double-stranded area as proposed by Brunier et al. (58). Analogously, large duplications also could result from the slipped mispairing of direct repeats in single-stranded regions between stretches of double stranded DNA, if the replication machinery was displaced to an upstream DNA sequence (58). This may be the mechanism responsible for the formation of the 8 kb duplication in Fabry Family D since imperfect hexanucleotide direct repeats flank the duplication breakpoints. Additionally, chromatin structure may bring distant termini close together. For example, Vanin et al. (8) proposed that four large deletions in the human β-globin gene cluster arose from the loss of integral numbers of chromatin loops anchored to the nuclear matrix during DNA replication. Alternatively, it is possible that the duplication in Fabry Family D, in particular, as well as the deletions in Families A, B, and/or E, may have resulted from unequal crossing-over between the direct repeats on different chromatids during meiosis.

In contrast to the above α-Gal A gene rearrangements which could be explained by simple illegitimate recombinational mechanisms, the Family J rearrangement involving two deletions separated by an inversion presumably resulted from a more complex event(s) (Fig. 8A). Analysis of the Family J sequence provides insights into the possible origin of this rearrangement. As shown in Figures 8B and 8C, the hexanucleotide direct repeat ATGTAT occurred at both the 5' and 3' breakpoints of the 14 bp deletion. The precise breakpoint in these repeats (ATGT - AT) was deduced from the Family J sequence in which the 3' breakpoint of the 14 bp deletion also was the 3' breakpoint of the 151 bp inversion. Also note that the 5' breakpoint of the 14 bp deletion (in the antisense orientation) was joined to the exon 5 breakpoint (Figs. 8B and 8C). Thus, it is likely that both the 14 bp deletion and the inversion resulted from a single breakage event. In contrast, examination of the region surrounding the 1696 bp deletion did not reveal direct

or inverted repeats which could have been aligned in the formation of the larger deletion (Fig. 8D). However, it is possible that the sequences surrounding the larger deletion were in close proximity to the sequences involved in the 14 bp deletion and inversion. These breakpoint regions may have been adjacent due to chromatin structure or the formation of stable secondary structures involving the oppositely-oriented *Alu* 9 and *Alu* 12 sequences (Fig. 8A). It is conceivable that the homology (5 of 6 nt) between the sequences in exon 5 (immediately upstream from the 5' breakpoint of the 1696 bp deletion) and those in the 14 bp deletion (which may have been looped out by the two flanking hexanucleotide direct repeats) (Figs. 8B and 8E), also could have been involved in their juxtaposition. It is notable that the inverted 151 bp sequence was just slightly longer than the 146 bp required to encircle a nucleosome (59), thereby protecting and/or positioning the inverted sequence as well as permitting the short direct repeats (AC and AT) at the respective 5' and 3' termini of the inverted segment (Fig. 8E) to facilitate the alignment of the recombinational event. Thus, a single, albeit complex, recombinational event could have resulted from the abnormal juxapositioning of these sequences during meiosis. In support of this hypothesis is the fact that the Family J rearrangement could be traced to a germinal mutation in the maternal grandfather. Alternatively, it is conceivable that this complex rearrangement was the result of two independent events, each involving two double-stranded breakages and reunions in the maternal grandfather's spermatogonia. The only other reported germinal inversion-deletion rearrangement occurred in the human β-globin gene cluster in which two deletions of 0.8 and 7.5 kb were separated by a 15.5 kb inverted segment (4). Although it was not possible to deduce the origin of this deletion-inversion rearrangement which caused a $^{A}\gamma\delta\beta$ thalassemia, it was suggested that chromatin structure, alignment of oppositely-oriented *Alu* or *Kpn*I repetitive elements, and/or short direct repeats at the breakpoints of the larger deletion may have been involved. Thus, the similarity of these two deletion-inversion rearrangements suggests that common underlying recombinational events may be responsible for their formation.

Comparison of the sequences surrounding the breakpoints in the five α-Gal A gene rearrangements which involved short direct repeats revealed that the trinucleotide CAG, or its complement CTG, occurred in or immediately adjacent to the direct repeat at the 5' breakpoint in four, or 80%, of these rearrangements (Table I). In contrast, the CAG/CTG trinucleotides were not present at any of the α-Gal A 3' breakpoints. The CAG and CTG sequences were the seventh and fifth most frequent trinucleotides in the α-Gal A gene, occurring on the average every 47 and 45 nt, respectively. Based on these frequencies, the expected occurrence of either trinucleotide in or adjacent to the 5' or 3' short direct repeat would be 36% in the five rearrangements. It was also noted that the sequence CCA, or its complement TGG, occurred at three of five α-Gal A 5' breakpoints and at one of the 3' breakpoints. The finding that the CCA/TGG and CAG/CTG sequences overlapped in three of the five α-Gal A 5' breakpoints suggested that the tetranucleotide CCAG (or CTGG), which was present at three of the 5' breakpoints and at none of the 3' breakpoints, might be important in the origin of these rearrangements. Analysis of the α-Gal A gene revealed that the sequences CCAG/CTGG were the twentieth and twelfth most frequent tetranucleotides in the α-Gal A gene, occurring on the average every 162 and 134 nt, respectively. Based on these frequencies, it would be expected that either tetranucleotide would occur in or next to the short direct repeats at the 5' or 3' breakpoints in about 10% of the five rearrangements. Thus, the tetranucleotide and CAG/CTG trinucleotide sequences were present more frequently than expected at the 5' breakpoints of the α-Gal A gene rearrangements.

Analysis of all mammalian germinal gene rearrangements with short direct repeats including α-Gal A (Table I) revealed that the trinucleotides CAG or CTG occurred in or immediately next to the 5' and 3' breakpoints in 52% and 19% of rearrangements, respectively (4-6, 8, 11, 13, 17, 26, 54). The tetranucleotides CCAG/CTGG occurred at 33% and 14% of 5' and 3' breakpoints, respectively. It is tempting to hypothesize that the

Table I. Breakpoint sequences in mammalian germinal gene rearrangements involving short direct repeats.

Rearrangement	5' Breakpoint	3' Breakpoint	Reference
Hb Leiden	TC**CT GA** GGAG	AGGA **GA** AGTC	5
Hb Lyon	GGGC **AA** GGTG	GGTG **AA** CGTG	5
Hb Freiburg	TGAA **GT** TGGT	GTTG **GT** GGTG	5
Hb Niteroi	GGTT **CTTTG** AGTC	AGTC **CTTTG** GGGA	5
Hb Gun Hill	AGTG **AGCTGCA** CTGT	GACA **AGCTGCA** CGTG	5
Hb Tochigi	TATG **GG** CAAC	CTAA **GG** TGAA	5
Hb St. Antoine	TGAT **GGC** CTGG	GC**CT GGC** TCAC	5
Hb Coventry	TAAT **GCC**C TGGC	C**CTG GCCC** ACAA	5
γδβ-Thal 1	TC**CC AG** CACT	GAAA **AG** TCTG	8
Hb BK	GGTA **TCT GGAG**	AATT **TCT** ATTA	4
Indian HPFH	CGCG **CCACT GCAC**	ATCC **CCACT** ATAT	11
Dutch β^0 Thal	AACC **AAATTT** GCAC	GAGA **AAATTT** TTGC	13
Turkish β Thal	GTCT **ACCC** TTGG	TTGG **ACCC AGAG**	17
$-(\alpha)^{20.5}$	CCTA **GGC** AACA	TAAG **GGC** CACG	6
RB 1	AGCT **TTTATAC** TTGA	TGAA **TTTAAAC** ATAA	26
Pro-α2(I)	TTTC **TTTC** TAAG	GTGG **TTTC** CCTG	61
Fabry Family A	GAAC **CCA GAAC**	AGCT **CCA** CCTC	
Fabry Family B	CATC **AAG** GGTA	TTAA **AAG** ACTT	
Fabry Family D	AGCT **TAGACA GGTA**	AATG **TAGATA** AAGA	
Fabry Family E	CAG**C AGAACT GGGG**	GGAA **AGAACT** TTGA	
Fabry Family J	TTTA **AC CAGG**	TAAT **AC** ATTT	
	GAAA **AT** TTTA	ATGT **AT** AGGC	

For each gene rearrangement, the 5' and 3' breakpoint regions are shown. Spaces have been inserted between the direct repeats (in bold) and the four adjacent 5' and 3' nucleotides which may be involved in the recombinational event. CTGG/CCAG tetranucleotides and/or CTG/CAG trinucleotides are indicated by underlines.

CAG/CTG trinucleotide (or possibly the CAGG/CTGG tetranucleotide) sequences may be recognition sites for proteins involved in replication or recombination, and that their frequent presence in the α-Gal A gene at 5' breakpoints may predispose sequences involving short direct repeats to illegitimate recombinational events. The recent identification of a tetranucleotide from the mouse immunoglobulin switch signal region at four of seven breakpoints in the hamster *aprt* gene also supports the potential role for short nucleotide sequences in somatic illegitimate recombinational events (1, 50, 60). Furthermore, this concept is consistent with the fact that short sequences of three or four nucleotides are recognition sites for proteins involved in DNA interactions such as the recognition site for topoisomerase I (61).

In summary, six gene rearrangements causing Fabry disease have been characterized by sequencing their novel junctions. These included the second deletion-inversion rearrangement and the third duplication described to date in man. All six rearrangements were intragenic, suggesting the possibility that flanking regions may encode essential functions such that larger deletions involving these genes might be lethal. Despite the fact that ~30% of the α-Gal A gene is comprised of *Alu* sequences, only one deletion resulted from recombination between two *Alu* repeats. The other rearrangements involved short direct repeats, further documenting the importance of small regions of homology in the generation of gene rearrangements causing this and other inherited diseases. Thus, the occurrence of short direct repeats and the paucity of *Alu* sequences in rearrangements of the α-Gal A gene suggests that the type and relative frequency of illegitimate recombinational events cannot be predicted, even in genes highly enriched with *Alu* repetitive sequences.

REFERENCES

1. Meuth, M. (1989) in *Mobile DNA* (Berg, D.E. and Howe, M.M., eds.) pp.833-860, American Society for Microbiology, Washington, D.C.

2. Lehrman, M.A., Goldstein, J.L., Russell, D.W., and Brown, M.S. (1987) *Cell* **48**, 827-835

3. Gitschier, J. (1988) *Am. J. Hum. Genet.* **43**, 274-279

4. Jennings, M.W., Jones, R.W., Wood, W.G., and Weatherall, D.J. (1985) *Nucleic Acids Res.* **13**, 2897-2906

5. Efstratiadis, A., Posakony, J.W., Maniatis, T., Lawn, R.M., O'Connell, C., Spritz, R.A., DeRiel, J.K., Forget, B.G., Weissman, S.M., Slightom, J.L., Blechl, A.E., Smithies, O., Baralle, F.E., Shoulders, C.C., and Proudfoot, N.J. (1980) *Cell* **21**, 653-668

6. Nicholls, R.D., Fischel-Ghodsian, N., and Higgs, D.R. (1987) *Cell* **49**, 369-378

7. Spritz, R.A. and Orkin, S.H. (1982) *Nucleic Acids Res.* **10**, 8025-8029

8. Vanin, E.F., Henthorn, P.S., Kioussis, D., Grosveld, F., and Smithies, O. (1983) *Cell* **35**, 701-709

9. Nicholls, R.D., Higgs, D.R., Clegg, J.B., and Weatherall, D.J. (1985) *Blood* **65**, 1434-1438

10. Mager, D.L., Henthorn, P.S., and Smithies, O. (1985) *Nucleic Acids Res.* **13**, 6559-6575

11. Henthorn, P.S., Mager, D.L., Huisman, T.H.J., and Smithies, O. (1986) *Proc. Natl. Acad. Sci.* USA **83**, 5194-5198

12. Popovich, B.W., Rosenblatt, D.S., Kendall, A.G., and Nishoika, Y. (1986) *Am. J. Hum. Genet.* **39**, 797-810

13. Gilman, J.G. (1987) *Br. J. Haem.* **67**, 369-372

14. Anand, R., Boehm, C.D., Kazazian, H.H., and Vanin, E.F. (1988) *Blood* **72**, 636-641

15. Kulozik, A.E., Yarwood, N., and Jones, R.W. (1988) *Blood* **71**, 457-462

16. Spiegelberg, R., Aulehla-Scholz, C., Erlich, H., and Horst, J. (1989) *Blood* **73**, 1695-1698

17. Schnee, J., Griese, E.-U., Eigel, A., and Horst, J. (1989) *Blood* **73**, 2224-2225

18. Lehrman, M.A., Schneider, W.J., Sudhof, T.C., Brown, M.S., Goldstein, J.L., and Russell, D.W. (1985) *Science* **227**, 140-146

19. Hobbs, H.H., Brown, M.S., Goldstein, J.L., and Russell, D.W. (1986) *J. Biol. Chem.* **261**, 13114-13120

20. Lehrman, M.A., Russell, D.W., Goldstein, J.L., and Brown, M.S. (1986) *Proc. Natl. Acad. Sci.* USA **83**, 3679-3683

21. Lehrman, M.A., Russell, D.W., Goldstein, J.L., and Brown, M.S. (1987) *J. Biol. Chem.* **262**, 3354-3361

22. Horsthemke, R., Beisiegel, U., Dunning, A., Havinga, J.R., Williamson, R., and Humphries, S. (1897) *Eur. J. Biochem.* **164**, 77-81

23. Aalto-Setala, K., Helve, E., Kovanen, P.T., and Kontula, K. (1989) *J. Clin. Invest.* **84**, 499-505

24. Miyake, Y., Tajima, S., Funahashi, T., and Yamamoto, A. (1989) *J. Biol. Chem.* **264**, 16584-16590

25. Myerowitz, R., and Hogikyan, N.D. (1987) *J. Biol. Chem.* **262**, 15396-15399

26. Canning, S., and Dryja, T.P. (1989) *Proc. Natl. Acad. Sci.*, USA **86**, 5044-5048

27. Langlois, S., Kastelein, J.J.P., and Hayden, M.R. (1988) *Am. J. Hum. Genet.* **43**, 60-68

28. Bernstein, H.S., Bishop, D.F., Astrin, K.H., Kornreich, R., Eng, C.M., Sakuraba, H., and Desnick, R.J. (1989) *J. Clin. Invest.* **83**, 1390-1399

29. Aldridge, J., Kunkel. L., Bruns, G., Tantravani, U., Lalande, M., Brewster, T., Moreau, E., Wilson, M., Bromley, W., Roderick, T., and Latt, S.A. (1984) *Am. J. Hum. Genet.* **36**, 546-564

30. Frischauf, A., Lehrach, H., Poustka, A., and Murray, N. (1983) *J. Mol. Biol.* **170**, 827-842

31. Benton, W.D., and Davis, R.W. (1977) *Science* **196**, 180-182

32. Saiki, R.K., Scharf, S., Faloona, F., Mullis, K.B., Horn, G.T., Erlich, H.A., and Arnheim, N. (1985) *Science* **230**, 1350-1354

33. Messing, J. (1983) *Methods Enzymol.* **101**, 20-78

34. Sanger, F., Nicklen, S., and Coulson, A.R. (1977) *Proc. Natl. Acad. Sci.* USA **74**, 5463-5476

35. Wang, L-M., Weber, D.K., Johnson, T., and Sakaguchi, A.Y. (1988) *Biotechniques* **6**, 839-843

36. Brutlag, D.L., Clayton, J., Friedland, P., and Kedes, L.H. (1982) *Nucleic Acids Res.* **10**, 279-294

37. Freund, A-E., Bichara, M., and Fuchs, R.P.P. (1989) *Proc. Natl. Acad. Sci.* USA **86**, 7465-7469

38. Paolella, G., Lucero, M.A., Murphy, M.H. and Baralle., F.E. (1983) *EMBO J.* **2**, 691-696

39. Markert, M.L., Hutton, J.J., Wiginton, D.A., States, J.C., and Kaufman, R.E. (1988) *J. Clin. Invest.* **81**, 1323-1327

40. Huang, L-S., Ripps, M.E., Korman, S.H., Deckelbaum, R.J., and Breslow, J.L. (1989) *J. Biol. Chem.* **264**, 11394-11400

41. Rozen, R., Fox, J., Fenton, W.A., Horwich, A.L., and Rosenberg, L.E. (1985) *Nature* **313**, 815-817

42. Youssoufian, H., Antonarakis, S.E., Aronis, S., Tsiftis,G., Phillips, D.G., and Kazazian, Jr., H.H. (1987) *Proc. Natl. Acad. Sci.* USA **84**, 3772-3776

43. Yang, T.P., Patel, P.I., Chinault, A.C., Stout, J.T., Jackson, L.G., Hildebrand, B.M., and Caskey, C.T. (1984) *Nature* **310**, 412-414

44. Matthews, R.J., Anson, D.S., Peake, I.R., and Bloom, A.L. (1987) *J. Clin. Invest.* **79**, 746-753

45. Deininger, P.L., Jolly, D.J., Rubin, C.M., Friedmann, T., and Schmid, C.W. (1981) *J. Mol. Biol.* **151**, 17-33

46. Farabaugh, P.J. and Miller, J.H. (1978) *J. Mol. Biol.* **126**, 847-863

47. Hogan, A. and Faust, E.A. (1984) *Mol. Cell Biol.* **4**, 2239-2242

48. Swebilius-Singer, B. and Westyle, J. (1988) *J. Mol. Biol.* **202**, 233-243

49. de Zamaroczy, M., Faugeron-Fonty, G, and Bernardi, G. (1983) *Gene* **21**, 193-202

50. Nalbantoglu, J., Hartley, D., Phear, G., Tear, G., and Meuth, M. *EMBO J.* (1986) **5**, 1199-1204

51. Albertini, A.M., Hofer, M., Calos, M.P. and Miller, J. (1982) *Cell* **29**, 319-328

52. Barsh, G.S., Roush, C.L., Bonadio, J., Byers, P.H., and Gelinas, R.E. (1985) *Proc. Natl. Acad. Sci.* USA **82**, 2870-2874

53. Chu, M-L, Gargiulo, V., Williams, C.J., and Ramirez, F. (1985) *J. Biol. Chem.* **260**, 691-694

54. Kuivaniemi, H., Sabol, C., Tromp, G., Sippola-Thiele, M. and Prockop, D.J. (1988) *J. Biol. Chem.* **263**, 11407-11413

55. Shapiro, L.J., Yen, P., Pomerantz, D., Martin, E., Rolewic, L., and Mohandas, T. (1989) *Proc. Natl. Acad. Sci.* USA **86**, 8477-8481

56. Lee, B., Vissing, H., Ramirez, F, Rogers, D., and Rimoin, D. (1989) *Science* **244**, 978-980

57. Bjursell, G., Gussander, E., and Lindahl, T. (1979) *Nature* **280**, 420-423

58. Brunier, D., Michel, B., and Ehrlich, S.D. (1988) *Cell* **52**, 883-892

59. McGhee, J.D. and Felsenfeld, G. (1980) *Ann. Rev. Biochem.* **49**, 1115-1156

60. Nalbantoglu, J., Phear, G., and Meuth, M. (1987) *Mol. Cell. Biol.* **7**, 1445-1449

61. Bullock, P., Champoux, J.J., and Botchan, M. (1985) *Science* **230**, 954-958

CONCLUSIONS

The mechanisms underlying gene rearrangements in higher eukaryotes has long been of interest to geneticists and molecular biologists. One approach to the study of genetic recombination in man has been the characterization of naturally occurring germ line mutations at the DNA sequence level. By comparing the mutant sequences to their normal counterparts, insights into the molecular mechanisms underlying their formation may be gained. However, extensive sequence data is available for only a limited number inherited disorders. As more information becomes available, the molecular factors predisposing chromosomal rearrangements may be elucidated.

To investigate the nature of the gene rearrangements causing Fabry disease, the breakpoints in the α-Gal A gene of five naturally occurring partial gene deletions and one partial gene duplication were determined. In order to characterize the precise location and nature of the breakpoints in these gene rearrangements, the entire α-Gal A chromosomal gene was sequenced. The first two chapters of this thesis discuss the isolation and characterization of the α-Gal A chromosomal gene and full-length cDNA. The third chapter describes the use of these tools in the analysis of α-Gal A gene rearrangements.

Knowledge of the sequence and genomic organization of the α-Gal A gene permitted characterization of the molecular nature of mutations causing Fabry disease. Application of this information lead to improved diagnosis in some Fabry families. Although the identification of a specific lesion in a family allows the precise diagnosis of affected hemizygotes and heterozygotes in that family, it is not feasible to determine the molecular defect in each family. An indirect approach to molecular diagnosis is the use of restriction length polymorphisms (RFLPs) within or closely linked to the disease locus. This powerful method has been greatly simplified by *in vitro* DNA amplification using the polymerase chain reaction. Genetic diagnoses can be achieved rapidly, with a minimal amount of sample and without the use of radioactivity. The nucleotide sequence

information described in this study will serve as a basis for the construction of specific oligonucleotide primers to use for amplification of sequences within or closely linked to the α-Gal A gene. Identification of several frequent α-Gal A RFLPs should make the molecular diagnosis of Fabry disease available to most families. In addition, the investigations detailed in this thesis have provided a foundation for the future study of such important processes as the regulation of X-linked housekeeping genes, X-chromosomal inactivation and the biosynthesis and processing of lysosomal enzymes.