INFORMATION TO USERS

While the most advanced technology has been used to photograph and reproduce this manuscript, the quality of the reproduction is heavily dependent upon the quality of the material submitted. For example:

- Manuscript pages may have indistinct print. In such cases, the best available copy has been filmed.
- Manuscripts may not always be complete. In such cases, a note will indicate that it is not possible to obtain missing pages.
- Copyrighted material may have been removed from the manuscript. In such cases, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, and charts) are photographed by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each oversize page is also filmed as one exposure and is available, for an additional charge, as a standard 35mm slide or as a 17"x 23" black and white photographic print.

Most photographs reproduce acceptably on positive microfilm or microfiche but lack the clarity on xerographic copies made from the microfilm. For an additional charge, 35mm slides of 6"x 9" black and white photographic prints are available for any photographs or illustrations that cannot be reproduced satisfactorily by xerography.

Quinn, Merrigene

FABRY DISEASE: ISOLATION, CLONING, AND SEQUENCE ANALYSES OF COMPLEMENTARY-DNA AND GENOMIC CLONES ENCODING ALPHA-GALACTOSIDASE A

City University of New York PH.D. 1987

PLEASE NOTE:

In all cases this material has been filmed in the best possible way from the available copy. Problems encountered with this document have been identified here with a check mark __✓__.

1. Glossy photographs or pages __✓__
2. Colored illustrations, paper or print __✓__
3. Photographs with dark background __✓__
4. Illustrations are poor copy _____
5. Pages with black marks, not original copy _____
6. Print shows through as there is text on both sides of page _____
7. Indistinct, broken or small print on several pages __✓__
8. Print exceeds margin requirements _____
9. Tightly bound copy with print lost in spine _____
10. Computer printout pages with indistinct print _____
11. Page(s) _____ lacking when material received, and not available from school or author.
12. Page(s) _____ seem to be missing in numbering only as text follows.
13. Two pages numbered _____. Text follows.
14. Curling and wrinkled pages _____
15. Dissertation contains pages with print at a slant, filmed as received __✓__
16. Other _____

_____

_____

FABRY DISEASE: ISOLATION, CLONING, AND SEQUENCE ANALYSES OF cDNA AND GENOMIC CLONES ENCODING ALPHA-GALACTOSIDASE A

By

Merrigene Quinn

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.

1987

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.

10 Apr 87
Date

David H. Calhoun, Ph.D.
Chairman of Examining Committee

10 April 87
Date

Terry A. Krulwich, Ph.D.
Executive Officer

Beatriz Pogo, M.D.

Catherine L. Squires, Ph.D.

Jerome L. Schulman, M.D.

James G. Wetmur, Ph.D.

Supervisory Committee

The City University of New York

Abstract

FABRY DISEASE: ISOLATION, CLONING, AND SEQUENCE ANALYSES OF cDNA AND GENOMIC CLONES ENCODING ALPHA-GALACTOSIDASE A

By

Merrigene Quinn

Adviser: David H. Calhoun, Associate Professor

Fabry disease, an X-linked error of glycosphingolipid metabolism, results from the deficient activity of the lysosomal hydrolase, alpha-galactosidase A. To investigate the nature of the molecular mutations involved with Fabry disease and to characterize the structure, organization, and expression of the alpha-galactosidase A locus, this study reports the isolation of cDNA and genomic clones specific for human alpha-galactosidase A. In addition, since this is the first genomic clone described to date for a lysosomal enzyme, it establishes a reference for future analyses of the molecular events that mediate the expression of lysosomal hydrolases. The human alpha-galactosidase A cDNA clone (designated λAG18) was isolated from a λgt11 expression library using the antibody detection method. Phage λAG18 contains an EcoRI cDNA insert of 1226 nucleotides with an open reading frame encoding 398 amino acids and a protein with a predicted

molecular weight of 45,346. This clone also includes sequences encoding the last five amino acids of the signal peptide. Two polyadenylation signals, AATACA and ATTAAA, are located 28 and 11 nucleotides, respectively, before the TAA stop codon. Phage λAG18 lacks a 3'- untranslated region since the poly(A) sequence immediately follows the TAA codon. Four possible N-glycosylation sites were identified within the propeptide sequence. Hybridization of HeLa cell mRNA with nick translated cDNA insert revealed a single band of approximately 1.45 kilobases. A genomic clone specific for human alpha-galactosidase A, was isolated from a lymphoblast 4X Charon 30 library. Nucleotide sequence analysis of a 1243 nucleotide genomic fragment includes sequences for the promoter, complete signal peptide, first exon, and 104 nucleotides of the first intron. Direct and inverted repeat elements of 10, 11, 16, 19, and 22 nucleotides flank the promoter site. A GA-rich repeat element of approximately 60 nucleotides, found to be homologous to similar elements in several species, is located upstream of the promoter. A GGGCGG site specific for the DNA binding protein Sp1 is located next to a CAAT box, and the inverted Sp1 binding site, CCGCCC, is located next to the TATA box. The sequence immediately flanking the ATG initiation codon of the human alpha-galactosidase A gene was shown to be highly homologous to sequences flanking the ATG inititiaon codons of four of the other five human lysosomal hydrolases for which sequence information is available, but not for any of the other 133 human signal peptides examided. Information is presented which indicates conversion to the mature enzyme occurs by cleavage of a carboxy-terminal propeptide fragment.

ACKNOWLEDGEMENTS

It is with my sincere gratitude that I would like to thank Dr. David H. Calhoun for the opportunity to complete my Ph.D. thesis in his laboratory. Through his patience, assistance, and guidance I have gained valuable insight and experience as a scientist. I would also like to thank Dr. Robert J. Desnick and Dr. David F. Bishop for their introduction to the the study of Fabry disease.

To Lisa Traub (Szeto), I would like to express my appreciation for her assistance with various laboratory procedures and her friendly lunchroom conversations. To Jim and Betty Cox, your Mid-West hospitality and friendliness was a welcome encounter. I would also like to thank Lavanya Rajachandran for her assistance in the sequencing project. I would like to say grazie to Vicenzo Fidanza (Enzo) for helping me to improve my conversational Italian and Neopolitan sign language. To Petros Hantzopoulos (the Greek), I would like to sincerely thank you for your computer, artistic, and photographic direction, and your most helpful exchange of scientific data. And to my officemate Audrey Manheimer-Lory, a thank you only begins to express my appreciation for the shared experiences and comraderie that developed during our five years as graduate students. I would also like to thank Scott, Eric, and Scott, and in that order, for their patience and support during this time.

Mostly, I would like to thank my mother for her encouragement, support, and love, which made it all possible. And lastly, to my father for his faith and instillation of the belief that achievement is an end and perserverance, dedication, and ambition are the means.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

INTRODUCTION

At the turn of the twentieth century, Sir Archibald Garrod began his studies of pathologic or disease states in relation to the body's normal physiologic processes. Through his studies of alcaptonuria, albinism, cystinuria, and pentosuria, Garrod proposed the hypothesis that an enzyme blocking a single metabolic reaction was either reduced or missing completely in patients with these disorders (Stanbury et. al., 1983). His analyses of familial distribution and knowledge of Mendelian genetics (Mendel, 1901), enabled Garrod to postulate that these disorders were due to recessive genes. Subsequent investigations proved Garrod's hypotheses to be correct and because of his substantial contributions to this field, he has become known as the father of inborn errors of metabolism. A group of genetic diseases known as lysosomal storage disorders result from inborn errors of metabolism. There are at least 30 lysosomal storage disorders and each results from an inherited trait which affects the intralysosomal enzymes (Dingle et al., 1984). The first clinical description of a lysosomal storage disorder came in 1881 by Warren Tay, a British opthalmologist, when he described a cherry red spot surrounded by a golden halo in the eye of an infant. Tay hypothesized that this clinical manifestation was due to an inherited factor. In 1897 a neurologist at Mt. Sinai Hospital in New York, B. Sachs, reported several additional cases of the cherry red spot in the eye of his patients and he also noted that these patients had extensive central nervous system damage (Sloan and Fredrickson, 1972). This disease has

been given the eponym of Tay-Sachs disease. Within the next two decades inherited lysosomal storage diseases were described and have now been categorized into three groups according to the chemical nature of the accumulated substrate: i) sphingolipidoses (e.g. Gaucher disease), ii) mucopolysaccharidoses (e.g. Hurler disease), and iii) glycoproteinoses (e.g. I cell disease) (Stanbury et al., 1983). Although each disorder is individually rare, taken collectively, their clinical incidence is significant.

Another lysosomal storage disorder known as Fabry disease was discovered independently in 1898 by Anderson in England and Fabry in Germany. Both described patients with angiokeratoma corporis diffusum. Thus, the disease might more appropriately be called Fabry-Anderson disease, however, the name Fabry disease has been retained for classification purposes. Clinically, the term angiokeratoma corporis diffusum arose from the characteristic skin lesions seen on hemizygous males. Additional clinical symptoms reported were corneal opacities, crises of fever, peripheral edema, and burning sensation in the extremeties (Desnick and Sweeley, 1983).

In 1947, Pompen et al. performed autopsies on two men known to have Fabry disease in order to examine the histopathology of the disorder. They made the significant observation of the presence of vacuoles in the tunica media of abnormal arteries throughout the body. This prompted them to propose that a generalized storage disease was the key to the disorder. The lipid nature of the storage material was established in 1950 (Scriba), and chemical analyses have since confirmed Scriba's studies. Sweeley and Klionsky (1963) were the first to classify Fabry disease as a sphingolipidosis due to their isolation and characterization of two neutral glycosphingolipids: i) galactosyl-

galactosylglucosyl ceramide (Gal-Gal-Glc-Cer) and ii) digalactosyl ceramide (Gal-Gal-Cer) from the autopsied kidney of a Fabry hemizygote. The observation that Fabry disease was due to the absence of trihexosyl ceramide galactosyl hydrolase activity was demonstrated by Brady et al. (1967).

Wallace (1958) and Colley et al. (1958) were the first to confirm Fabry disease in a woman. This woman's son had died due to Fabry disease and her subsequent autopsy material revealed vacuolated glomerular epithelial cells. Subsequently, Wise et al. (1962) and Burda and Winder (1967) demonstrated very slight clinical features of Fabry disease in carrier heterozygous females. The X-linked transmission of the disorder was confirmed by Opitz and Stiles (1963) and Opitz (1964) through studies of at least 21 carrier females and affected males. Literature reports of over 45 documented heterozygotes show that the females are generally less severely affected with corneal dystrophy as the most frequent and often singular manifestation (Desnick and Sweeley, 1983). Prognosis is better for the heterozygote than for the hemizygote. Although life expectancy is longer in affected females, most become more symptomatic with increasing age and generally die of the disease (Burda and Winder, 1967).

Of the glycolipid storage disorders, only in Fabry disease is the deficient enzyme transmitted by an X-linked structural gene. Opitz et al. (1965) first studied the relative position of the Fabry locus on the X chromosome. The initial data of Opitz et al. (1965) and then Johnston et al. (1969), indicated a linkage between alpha-galactosidase A and the Xga blood group antigens. However, subsequent data derived from somatic cell hybridization assigned the alpha-galactosidase A structural gene to the long arm of X (Xq) (de la

Chapelle and Miller, 1979), while other data placed the Xg blood groups on the short arm of X (Xp). Further linkage data from two other group negates the previous evidence for linkage between alpha-galactosidase A and Xg (Ropers et al., 1977; Johnston and Sanger, 1981). Specifically, the strucural gene for alpha-galactosidase A has been localized to a narrow region of Xq (Xq21-Xq22) (Fox et al., 1984) by somatic cell hybridization techniques.

Lysosomal enzymes break down down macromolecules and and their end products may be eliminated from the cell or used as anabolic buiding blocks (Dingle et al., 1984). Failure of one of these hydrolytic enzymes can result in the accumulation of their substrates or their abnormal metabolites within specific cell types of the body. Resultant accumulation can cause a variety of clinical sequelae ranging from dermal lesions to mild retardation to death. The lysosomal storage disorders provide examples of genetic heterogeneity associated with similar or indistinguishable clinical manifestations. Indistinguishable phenotypes can be caused by, i) mutations of different genes which affect the metabolism of the same end product or by ii) mutations at different alleles which affect the nature of a gene product (genotype). An example of the first situation is seen in mucopolysaccharidosis type III where there are four forms of the disease, Sanfilippo A, B, C, and D. Clinically, all result in accumulation of heparin sulfate in the tissues, however genetically, the activity of four separate enzymes is affected (McKusick et al., 1978). Tay-Sachs disease is an example of the second situation since the alpha and beta subunits of hexosaminidase A are encoded on chromosome 15 and 5, respectively (O'Brien, 1983).

The control, regulation, expression, biosynthesis, and trafficking

of lysosomal enzymes is not completely understood. The study of patients with various forms of lysosomal storage disorders has helped to unravel these complex biologic processes. Lysosomal enzymes are synthesized on membrane-bound polysomes on the rough endoplasmic reticulum along with secretory proteins and membrane proteins. All of these proteins are thought to contain an amino terminal signal peptide which participates in translocation of the protein across the membrane of the endoplasmic reticulum (Walter et. al., 1984). To date, two components have been purified and shown to be necessary for translocation. The first is the signal recognition particle (SRP), an 11S cytoplasmic ribonucleoprotein, which recognizes the signal sequence of nascent polypeptides (Erickson et al., 1981; Erickson et al., 1983; Rosenfeld et al., 1982). The second component of the translocation event is the SRP receptor, also known as the docking protein. The SRP receptor is an integral membrane protein which associates with the SRP and allows targeting of the ribosomes and protein to the membrane of the rough endoplasmic reticulum (Walter et al., 1984). The steps following the targeting process are poorly understood, however, somehow the permeability barrier of the membrane is altered and the nascent polypeptide passes through to the lumen.

During the concomitant translocation and translation several modifications occur, including cleavage of the the signal peptide from the nascent lysosomal polypeptide and addition of high mannose core oligosaccharides. Specifically, three glucose, nine mannose, and 2 N-acetylglucosamine residues are added en bloc by a lipid carrier to selected asparagine residues along the lysosomal polypeptide (Kornfeld and Kornfeld, 1985). Processing of asparagine-linked oligosaccharides begins by the removal of three glucose residues and one mannose

residue. Within the lumen of the rough endoplasmic reticulum, the lysosomal enzymes are in the same pool as the secretory proteins since they share a similar mechanism of transport.

From the lumen the proteins are transported to the Golgi apparatus for posttranslational modifications and sorting to their proper destinations (e.g. lysosome, secretory granule, or plasa membrane). The site(s) and mechanism(s) by which lysosomal enzymes are sorted from the pool of secretory proteins is not known. However, it is known that within the Golgi complex, one important acquisition for lysosomal enzymes is the addition of mannose-6-phosphate (M-6-P) by a two step process. First, N-acetylglucosamine-1-phosphotransferase catalyzes the transfer of N-acetylglucosamine-1-phosphate to a mannose residue on the oligosaccharide (Hasilik et al., 1981; Reitman and Kornfeld, 1981). Subsequently, N-acetylglucosamine is removed by alpha-N-acetyl-glucosaminyl phosphodiesterase and the mature recognition marker, M-6-P, is exposed (Varki and Kornfeld, 1980; Waheed et al., 1981; Varki and Kornfeld, 1981). These residues are recognized by M-6-P receptors located in the plasma membrane of the golgi apparatus and the receptor-ligand complex travels to vesicles destined to become primary lysosomes via an undetermined pathway (Sly and Fischer, 1982). Within the acidic milieu of prelysosomal compartments the receptor-ligand complex dissociates and the M-6-P receptor is free to recycle (Gonzalez-Noriega et al., 1980). Recently it has been shown that there are actually two types M-6-P receptors based on their requirement for divalent cations. The new receptor has been named cation-dependent M-6-P receptor and the other cation-independent M-6-P receptor (Hoflack and Kornfeld, 1985). This finding raises interesting questions. Do lysosomal enzymes exhibit random or selective binding to these

different receptors and if so, what determines their selectivity? And do these different receptors serve to traffick different acid hydrolases to separate types of lysosomes with specialized functions?

Lang et al. (1984) and Kornfeld et al. (1985) demonstrated that the selective recognition marker of acid hydrolases by N-acetylglucosaminyl phosphotransferase is contained within the protein rather than the oligosaccharide portion of lysosomal enzymes. Specifically, using an in vitro assay, deglycosylated enzymes (endonuclease H-treated) acted as both specific and competitive inhibitors of phosphorylation of the intact lysosomal enzymes. In addition, they showed that the intact protein is required for inhibition since large proteolytic fragments and denatured proteins did not possess the information neccessary to inhibit phosphorylation. Of the available cDNA sequences for human lysosomal enzymes, (Faust et al., 1985; Fukushima et al., 1985; Sorge et al., 1985; Guise et al., 1985; Myerowitz et al., 1985; O'Dowd et al., 1985; Fong et al., 1986; Oshima et al., 1987), no significant homologies were found at the nucleotide or amino acid level. This suggests that there may be secondary or tertiary protein structure in common which allows specific recognition of acid hydrolases by N-acetylglucosaminyl phosphotransferase.

Although M-6-P receptor mediated delivery to lysosomes is an an important pathway for channeling enzymes, it is known that this is not the only route to the lysosome and there must be a pathway(s) independent of this receptor. The evidence comes from the study of patients with I cell disease (mucolipidosis II) and the genetically related disease, pseudo-Hurler polydystrophy (mucolipidosis III), who are unable to synthesize the M-6-P recognition marker due to an impaired N-acetylglucosaminyl phosphotransferase (Hasilik et al.,1981;

Reitman et al., 1981; and Varki et al., 1981). Thus, their lysosomal enzymes are not suitable ligands to the M-6-P receptor. For fibroblasts and certain other cells, the lysosomal enzymes are secreted into the external environment rather than delivered to the lysosomes. Interestingly, Kupffer cells, leukocytes, and hepatocytes contain nearly normal levels of enzymes within their lysosomes (Owada and Neufeld, 1982; Waheed et al., 1982). It also has been shown by Gabel et al. (1983) that a number of cells without the M-6-P receptor (e.g. P388D, J774, and L cells) have high levels of lysosomal acid hydrolases. Clearly these cells have a M-6-P independent mechanism for trafficking enzymes to the lysosomes, however, this mechanism is unknown at this time.

Besides cleavage of the signal peptide in the endoplasmic reticulum and oligosaccharide processing, lysosomal enzymes undergo additional proteolytic processing. Lysosomal enzymes are synthesized as prepropeptides (Hasilik et al., 1981); the prepropeptide includes the amino terminal signal peptide which is removed first followed by removal of the pro piece after the enzyme has entered the prelysosomal compartment (Gieselmann et al., 1983). Propeptide processing has been shown to occur at the amino and/or carboxy terminus (Erickson et al., 1981; Erickson and Blobel, 1983). It has been suggested that removal of the pro piece of cathepsin D activates the zymogen and contains the activity of the enzyme prior to its sorting and delivery to the lysosome (Erickson et al., 1981). Zymogen activation by removal of the pro piece does not appear to pertain to all lysosomal enzymes. Kornfeld (1986) speculates that the presence of the pro piece of the enzyme might be necessary for proper folding of the peptide for sorting via the M-6-P receptor independent pathway. His speculation is based

on the evidence that the pro piece is not necessary for the addition of M-6-P residues, since purified mature protein can be phosphorylated by the phosphotransferase (Lang et al., 1984; Kornfeld et al., 1985). Once the enzyme has been delivered to the lysosome this pro piece becomes dispensable and is removed. An alternative postulation suggests that removal of the pro piece of the peptide stabilizes the enzyme in the acidic milieu of the lysosomal compartment (Kornfeld, 1986). It is possible that removal of the pro piece may serve a dual role in some lysosomal enzymes. Specifically, removal of the pro piece may activate and stabilize the enzyme in the newly found acidic environment of the lysosome.

During the past twenty years, significant observations were made in the understanding of the molecular pathologies of lysosomal storage disorders and other inherited metabolic diseases. This information gave way to the development of specific and facile tests to accurately diagnose and identify specific enzymatic defects in more than 200 of the over 400 characterized, recessively inherited, inborn errors of metabolism (McKusick, 1978). These techniques and in vitro somatic cell hybridization assays have been utilized to assign structural genes for these enzymes to specific human chromosomes (Evans et al., 1979).

Research involving patients with inborn errors of metabolism has led to a better understanding of the cell biology, physiology, and biochemistry of these disorders. However, an obvious goal is to provide specific therapy for patients suffering from these devastating diseases. Because efficient treatment is not yet available for patients with many inherited metabolic deficiencies, investigators devised studies to identify specific treatments to alter the disease course. Early efforts were divided into two major themes: i) decrease

the progressively accumulated substrate through dietary restrictions or ii) replace the deficient gene product through intravenous injections. The first approach has proved to be successful, for example, with phenylketonuria (Hoskins et al., 1980). The second approach was prompted by preliminary in vitro studies for lysosomal storage disorders which revealed that the exogenously added, appropriate enzyme could correct the metabolic defect in cultured fibroblasts from patients with lysosomal storage disorders (Porter et al., 1971; Hickman and Neufeld, 1972).

Clinical trials were designed to administer and target the normal human enzyme in patients with lysosomal enzyme deficiencies (Tagler et al., 1974; Brady et al., 1973; Desnick and Grabowski, 1981). These trials encountered several obstacles which decreased the desired effect of this approach. One major hinderance was the fact that many lysosomal storage disorders are characterized by storage of lipids and mucopolysaccharides within the neurons, and hence, the administered enzyme must cross the blood-brain barrier. Also, the intravenously administered enzyme was rapidly cleared from the plasma and uptaken primarily by the hepatic system. Finally, a major stumbling block was the lack of sufficient amounts of exogenous preparations of purified enzyme for assessment of long term clinical effectiveness.

For Fabry disease, clinical efforts were designed to reduce the circulating substrate, thereby, reducing vascular accumulation. The first clinical infusion of alpha-galactosidase A was reported in 1970 (Mapes et al.). These and additional trials (Brady et al., 1973) showed transient depletion of substrate in the circulation following purified alpha-glactosidase A infusion. In a later study (Desnick et al., 1979), affected males (hemizygotes) were given several injections

of alpha-galactosidase A purified spleen or plasma at 100,000 nmole/hr per injection. In brief, the results were the following: i) the more highly sialyated plasma enzyme remained in the circulation seven times longer than the splenic form, ii) both enzymatic forms were detected in biopsied liver 30 min post-injection iii) systemically, a 25-fold greater reduction of plasma substrate by the plasma enzyme was reported compared to the depletion by the splenic enzyme, iv) during a span of 120 days, no detectable immune response was generated to either form of the enzyme, and v) stored substrate was mobilized into the circulation by the plasma form of the enzyme, but not by the enzyme purified from spleen.

Several lines of evidence support the application of recombinant DNA technology for the treatment of Fabry disease. First of all, the absence of detectable immunologic reactions and the clinical effectiveness of replacement therapy in preliminary human trials has been demonstrated as previously noted. In addition to the positive in vivo results, Dawson et al. (1973) reported that exogenous alpha-galactosidase A added in vitro to cultured fibroblasts from unrelated hemizygotes with Fabry disease corrected the biochemical defect. Also, Fabry disease has no central nervous system involvement (Desnick and Sweeley, 1983). In this context, the identification of a 42 year old male variant whom has low alpha-galactosidase A activity (3-10% of the normal activity) and is completely asymptomatic (Bishop et al., 1981), has provided an additional stimulus to the efforts designed to treat Fabry disease. The report of this variant provides evidence to the hypothesis that low levels of administered alpha-galactosidase A may be effective. Since alpha-galactosidase A is a homodimer (Bishop and Desnick, 1981), only a single gene is required for cloning and eventual

expression in microbial systems.

In addition to clinical treatment, basic investigations of structural and functional characterizations of lysosomal enzymes and in vitro studies also have been delayed due to the lack of sufficient quantities of homogeneous enzyme. However, due to recent advances in microbial expression systems and the cloning of several lysosomal hydrolase cDNAs, a new book is opened on the application of enzyme replacement therapy for the treatment of lysosomal storage disorders. Recombinant DNA technology can be employed to construct plasmids that direct the production in microorganisms of large amounts of enzyme that can be used for research leading to therapeutic applications. Indeed, expression and purification of large amounts of active gene products are foreseeable. Besides therapeutic applications, molecular studies and genetic manipulations can be utilized to determine gene structure, regulation, and transcriptional/translational processing. Nucleotide sequence analyses of the normal and disease states will determine the molecular defects associated with inherited metabolic diseases.

Recent experiments indicate the feasibility of retrovirus-mediated gene transfer as an approach to gene therapy for lysosomal storage diseases and other inherited metabolic disorders. Gaucher disease, a glycolipid storage disorder, results from the deficient activity of the lysosomal enzyme glucocerebrosidase (Brady and Barranger, 1983). Sorge et al. (1987) cloned the human glucocoerebrosidase cDNA into a retrovirus vector and infected mouse fibroblasts with this construct. Infected mouse fibroblasts expressed active human glucocerebrosidase. The retrovirus vector was then rescued from mouse fibroblasts by a helper virus and used to transform cultured fibroblasts and lymphoblasts from Gaucher patients. Sorge et al. (1987) demonstrated

complete correction of the enzymatic defect in these transformed cells. In a similar experiment by Palmer et al. (1987), a retrovirus vector containing a human adenosine deaminase cDNA, was used to infect human fibroblasts from a patient with adenosine deaminase deficiency. The infected cells produced 12-fold more adenosine deaminase than uninfected cells and were able to metabolize adenosine and deoxyadenosine, substrates that accumulate in the plasma of affected individuals (Kredich and Hershfield, 1983). Another hereditary disorder, alpha-antitrypsin deficiency, is characterized by reduced serum levels of alpha-antitrypsin resulting in destruction of the lower respiratory tract by neutrophil elastase (Gadek and Crystal, 1983). Garver et al. (1987) used a retrovirus vector containing alpha-antitrypsin cDNA to produce glycosylated, active human alpha-antitrypsin in mouse fibroblasts. Their hope is to produce sufficient quantities of the enzyme for therapeutic applications.

In summary, the availability of cDNA and genomic clones for alpha-galactosidase A would facilitate studies of the molecular basis of the disease, provide specific probes for heterozygote identification, and permit expression of large amounts of the enzyme for further structural analyses and for therapeutic trials of enzyme replacement. The first part of the study reported here focuses on a group effort directed towards the isolation of a cDNA encoding human alpha-galactosidase A. Antibodies prepared from purified alpha-galactosidase A protein and oligonucleotides synthesized from knowledge of protein sequence were utilized to screen cDNA libraries. Partial nucleotide sequences of a positive cDNA clone was used to verify clone authenticity. The complete nucleotide sequence of a cDNA clone specific for human alpha-galactosidase A was determined and the

predicted protein sequence was obtained. Consensus poly(A) addition sites, N-glycosylation sites, restriction endonuclease sites, and secondary structure of the mature protein were identified.

The second half of this study reports my specific efforts that led to the isolation of genomic clones specific for human alpha-galactosidase A. In fact, these are the only genomic clones reported to date for any mammalian lysosomal enzyme. Nucleotide data from the alpha-galactosidase A cDNA clone was utilized to synthesize an oligonucleotide corresponding to the amino terminal amino acid sequence for screening genomic libraries. Nucleotide sequence analysis of a 1243 nucleotide genomic segment revealed the presence of the 5'-flanking promoter sequences, transcription initiation sites, signal peptide sequence, and first exon-intron splice junction. Analyses of these regions should give insight into the mechansisms that regulate lysosomal enzymes and intracellular processing. In addition, these studies should lay the foundation for future analyses of the molecular mutations involved in Fabry disease and the expression of the alpha-galactosidase A locus in normal individuals and in Fabry patients.

MATERIALS AND METHODS

Materials. Beta-cyanoethyl diisopropylphosphoramidites and heat sealable Kapak pouches (8 inches X 9.5 inches) were purchased from Fisher. Reverse transcriptase was obtained from Life Sciences. The M13 universal primer (17-mer; 5'-GTAAAACGACGGCCAGT-3') and the following enzymes were purchased from New England Biolabs: phage T4 ligase, phage T4 polynucleotide kinase, the Klenow fragment of E. coli DNA polymerase I, and restriction endonucleases AccI, BamHI, EcoRI, HindIII, PstI, RsaI, SacI, and TaqI. Plasmids pUC8 and pUC9 were obtained from Bethesda Research Laboratories. Calf intestinal phosphatase, [α-^{35}S]dNTPs (1200 Ci/mmole), [α-^{32}P]dATP (800 Ci/mmole), and [γ-^{32}P]dATP (3000 Ci/mmole) were purchased from New England Nuclear. The rapid deletion subcloning kit and synthetic oligonucleotide RD22 (5'-CGACGGCCAGTGAATTCCCCCC-3') were purchased from International Biotechnologies, Inc. DEAE cellulose membranes (NA-45), Zetabind nylon membranes, colony/plaque screen nylon membranes, and nitrocellulose membranes (137 mm) were from Schleicher and Schuell, ABF Cuno, New England Nuclear, and Millipore, respectively. Other chemicals and reagents were obtained from several vendors.

Bacteria, Bacteriophages, and Recombinant Plasmids. Bacteria strains, various libraries, and recombinant plasmids utilized during the course of this study are described in Table 1.

Table 1. Bacteria, bacteriophages, and recombinant plasmids.

Strain	Description	Reference
BACTERIA		
C600	F⁻,thi-1,thr-1,leuB6,lacY1, supE44,tonA21,λ⁻	(Appleyard, 1954)
HB101	F⁻,hsdS20,(r^-_B,m^-_B)recA13, ara-14,proA2,lacY1,galK2, rpsL20(Sm^R),Xyl-5,Mtl-1, supE44, λ⁻	(Boyer and Rouilland-Dussoix, 1969)
JM103	Δ(lac pro),thi,strA,supE,endA, sbcB,hsdR⁻,F'traD36,proAB, lacIq,lacZAM15	(Messing et al., 1981)
X1776	F⁻,tonA53,dapD8,minA1,glnr44, (supE42),Δ(gal-uvrB)40,λ⁻, minB2,rfb-2,gyrA25,thyA142,oms-2, metC65,oms-1,(tte-1),Δ(bioH-asd)29 cycB2,cycA1,hsdR2	(Curtiss et al., 1977).
Y1088	lacV169,supE,supF,hsdR⁻,hsdM⁺, metB,trpR,tonA21,proC::Tn5(pMC9)	(Young and Davis, 1983a).
BACTERIOPHAGES		
Charon 30	b1007,KH54,nin5,dup1(sbh2-3)	(Rimm et al., 1980)
λgt7	lac5(b522,nin5)	
λgt11	lac5,cI857,nin5,S100	(Young and Davis, 1983b)
λgtWES.λB	wam403,sam1100,sam100,cI857, Requires host bacteria with SupF	(Leder et al., 1977)
λ1059	hλs,bam1o,b189(int29,nin144, cI857,pac129)Δ(int-cIII) KH54	(Karn et al., 1980)
λAG18	λgt11 with EcoRI cDNA insert of alpha-galactosidase A	(Calhoun et al., 1985)
λAG1,λAG2, λAG3	Lysogenic form of λAG18 (in Y1088; TN1614a)	This work
λMQ1, λMQ2	Independent isolates containing genomic alpha-galactosidase A inserts in Charon 30	This work
λC1, λC2	Independent isolates from λ1059 genomic library	This work

Table 1. (cont.)

Vector	Description	Reference
PLASMIDS		
pBR322	plasmid vector	(Bolivar et al., 1977; Sutcliffe, 1978)
pAG18	pBR322 with alpha-galactosidase A cDNA 1.2 kb EcoRI insert	(Calhoun et al., 1985)
pUC8, pUC9	plasmid vectors	(Vierra and Messing, 1982)
pUC9-18-5	pUC9 with EcoRI 1.2 kb alpha-galactosidase A cDNA insert (in HB101; TN1615a)	This work
pUC9-18-3	same as pUC9-18-5 with insert in opposite orientation (in HB101; TN1616a)	This work
pUC8-18-19	pUC8 with EcoRI 1.2 kb alpha-galactosidase A cDNA insert (in C600; TN1619a)	This work
pUC8-18-20	same as pUC8-18-6 with insert in opposite orientation (in C600; TN1620a)	This work
pcDX	plasmid vector containing cDNA inserts	(Okayama and Berg, 1983)
pcD1-21	pcD clones isolated as positive to oligonucleotide probes	This work
pSPRI	plasmid vector	(Krystal et al., 1986)
pMQ1, pMQ2	pSPRI containing SacI insert from λMQ1 and λMQ2, respectively	This work

Table 1. (cont.)

Vector	Description	Reference
M13 VECTORS[b]		
M13mp18, M13mp11	cloning vectors	(Messing, 1983)
M25	M13mp18 with 1.2 kb EcoRI insert of alpha-galactosidase A cDNA (same sense as message[c])	(Bishop et al., 1986)
M25.21,.22, and .28	Deletion derivatives of M25 for sequence analysis; Fig. 12.	This work
M27	M13mp18 with 1.2 kb EcoRI insert of alpha-galactosidase A (anti-message sense[c])	(Bishop et al., 1986)
M27.18,.15, .30,and .5	Deletion derivatives of M27 for sequence analysis; Fig. 12.	This work
mMQU	M13mp11 with 1.265 kb TaqI genomic insert of alpha-galactosidase A (same sense as message[c])	This work
mMQL	M13mp11 with 1.265 kb TaqI genomic insert of alpha-galactosidase A (anti-message sense[c])	This work
mMQZ	mMQU minus the 22 nucleotide polylinker region of pSPRI	This work

[a] TN refers to laboratory number of E. coli host strain which carries indicated plasmid or bacteriophage.
[b] M13 vectors and derivatives were grown in JM103.
[c] The same sense as message means that the bacteriophage M13 packages the strand of the human cDNA with the same polarity as mRNA, or its anti-message complement.

Purification and Amino Acid Sequence Analyses of Alpha-Galactosidase A. Human lung alpha-galactosidase A, purified in Dr. D.F. Bishop's laboratory according to the method of Bishop and Desnick (1981) with slight modifications (Calhoun et al., 1985), was used for amino acid sequence and composition analyses. Two independent preparations of 95% pure alpha-galactosidase A were sequenced in Dr. Leroy Hood's laboratory at The California Institute of Technology by extrapolation of each amino acid concentration to its zero-time value from 24-, 48-, and 72-hour (hr) hydrolyses in 6 M HCl at 110°C. Performic acid oxidation was used for analysis of cysteine as cysteic acid and for methionine as the sulfone (Hirs, 1967). Amino acid concentrations were determined in a Durrum model D-500 analyzer. The tryptophan concentration was obtained from the ratio of its absorbance to that of tyrosine by the spectro- photometric method of Edelhoch (1967). Homogeneous alpha-galactosidase A was digested with trypsin treated with tosylphenylalanine chloromethyl ketone (Allen, 1981) or cleaved with cyanogen bromide (Gross, 1967), and the peptides were isolated by reversed-phase HPLC (Browne et al., 1982). The amino acid sequences of the amino-terminal, tryptic, and cyanogen bromide peptides were determined by automated gas-phase microsequencing and HPLC identification of the phenylthiohydantoin derivatives of the amino acids as described by Hunkapiller and Hood (1983).

Synthesis of Deoxyoligonucleotide Mixtures. Within the first 37 amino acids of the alpha-galactosidase A sequence and within an internal tryptic peptide, several regions of low codon redundancy were identified and utilized to synthesize oligonucleotide mixtures corresponding to predicted mRNA sequence. Oligonucleotide mixtures

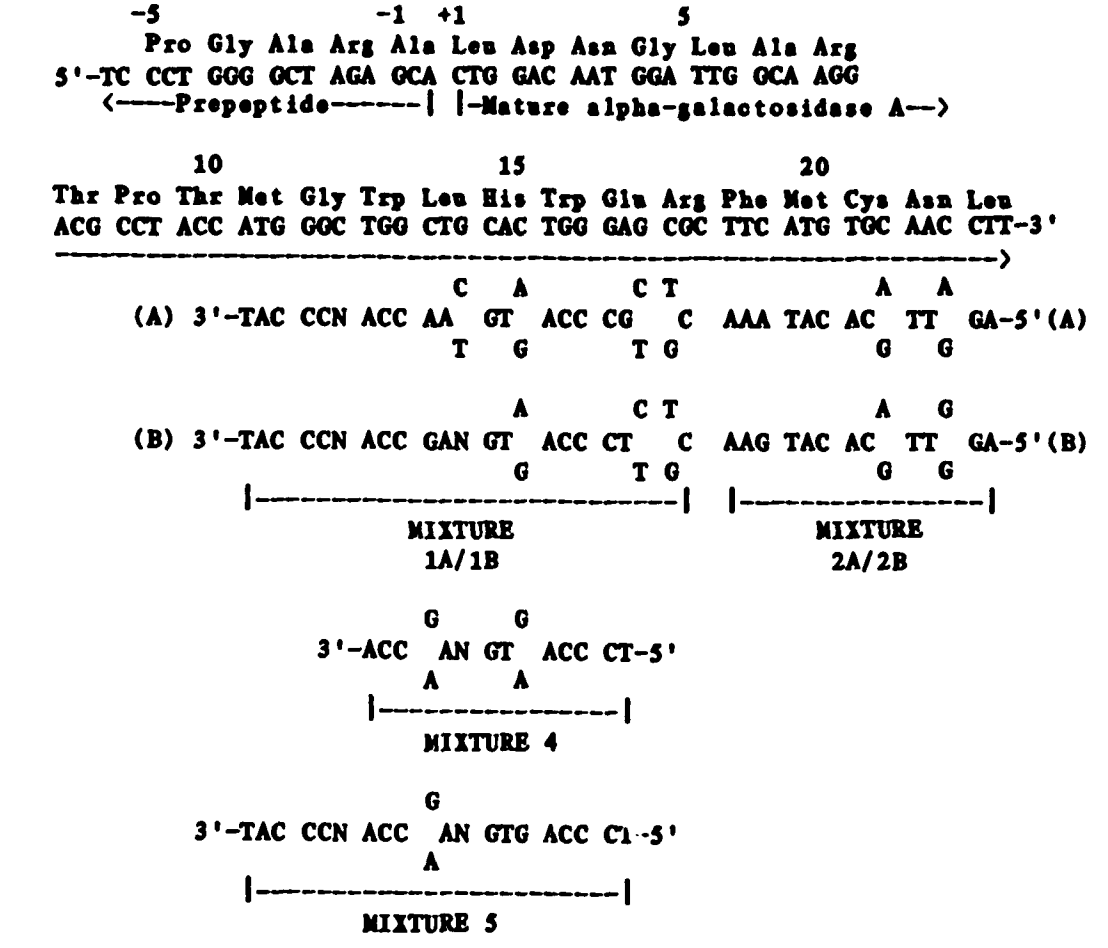
2A/2B (Fig. 1; the designation 2A/2B refers to a combination of oligonucleotide mixtures 2A and 2B) and oligonucleotide mixture 4 were purchased from Applied Biosystems and New England Biolabs, respectively. Oligonucleotide mixture 5 was constructed for us by Dr. Peter Model, Rockefeller University. The other oligonucleotide mixtures were synthesized in the Microbiology department on a Sam One synthesizer (Biosearch), using phosphotriester or, later, using phosphoramidite chemistries with Beta-cyanoethyl diisopropylphosphoramidites. Oligonucleotide mixtures were purified by gel electrophoresis on 20% polyacrylamide gels containing 8 M urea and DNA was localized by UV shadow-casting.

1. Oligonucleotide mixture 1A/1B. Oligonucleotide mixtures 1A and 1B are 23-mers (mixtures of 64 and 128 oligonucleotide species, respectively) synthesized to correspond to amino acids 11 through 18. Oligonucleotides 1A and 1B differed in that they were specific for leucine condons UUP (P= purine) and CUN (N= any nucleotide), respectively.

2. Oligonucleotide mixture 2A/2B. Oligonucleotide mixures 2A and 2B, are each composed of four different 14-mers spanning amino acids 19 through 23. The complexity of these mixtures was reduced by selecting G for the first nucleotide codon for leucine, based on the frequency (94%) of its human codon usage (Grantham et al., 1981).

3. Oligonucleotide mixture 3. Corresponding to an internal tryptic peptide sequence, oligonucleotide mixture 3 is composed of 96 different 17-mers.

Panel A.



Panel B.

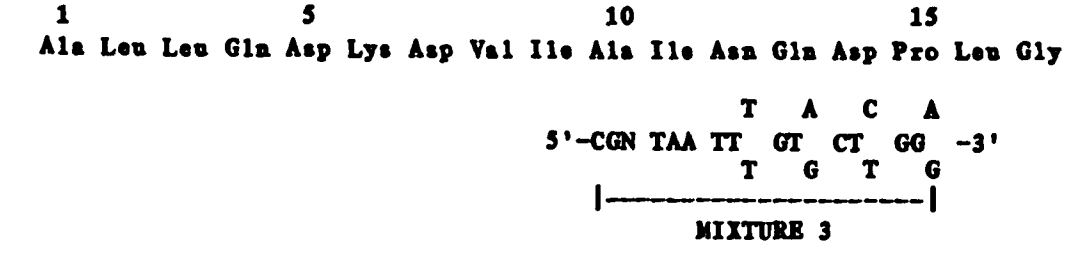


Fig. 1. Synthetic oligonucleotide mixtures. Panel A shows oligonucleotide mixtures synthesized from the amino acid sequence of amino terminal fragment of the alpha-galactosidase A protein. Panel B shows an oligonucleotide mixture synthsized from the amino acid sequence of an internal tryptic fragment. Mixtures were 5'-labelled with [γ^{32}P]dATP by T4 polynucleotide kinase and utilized as hybridization probes to screen cDNA and genomic libraries for clones with complementary sequences.

4. Oligonucleotide mixture 4. This oligonucleotide mixture of 16 different 14-mers spans sequences encoding amino acids 13 to 17 of alpha-galactosidase A.

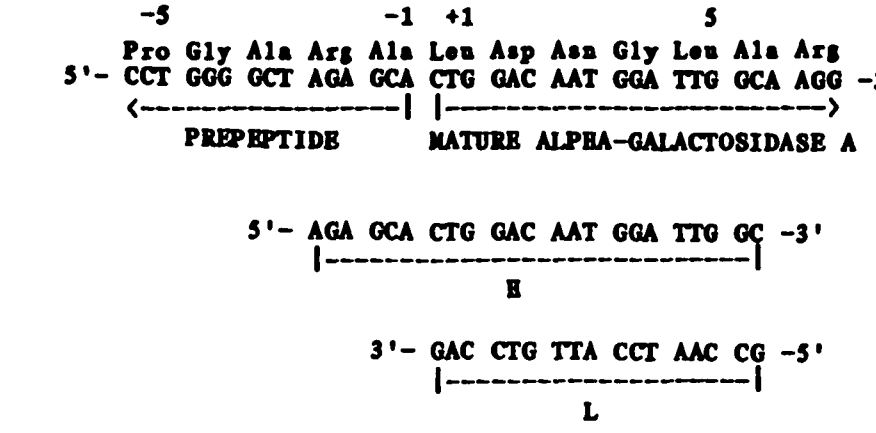
5. Oligonucleotide mixture 5. This mixture was designed to be complementary to sequences encoding alpha-galactosidase A amino acids 11 to 17 and consists of 32 different 20-mers.

Synthesis of Unique Deoxyoligonucleotides. Unique oligonucleotides, as opposed to the oligonucleotide mixtures described above, were based on alpha-galactosidase A cDNA nucleotide sequence or from genomic nucleotide sequence (Fig. 2). These were also synthesized in the Microbiolgy department as described above. Unique oligonucleotides were distinguished from mixtures by letter designations compared to numerical designations used for oligonucleotide mixtures.

1. Oligonucleotide H. This unique 23-mer, synthesized by Petros Hantzopoulos, spans nucleotides encoding the prepropeptide cleavage site. Specifically, it spans the last six nucleotides encoding the signal sequence and the first 17 nucleotides of the propeptide sequence and is the same polarity as message RNA.

2. Oligonucleotide L. Oligonucleotide L was synthesized to read sequence towards the signal peptide initiation codon and promoter region of the alpha-galactosidase A gene. It spans the first 17 nucleotides encoding the propeptide and is complementary to the message RNA.

A.



B.

Fig. 2. Unique synthetic oligonucleotides. These were designed from sequence data of the cDNA clone (probe H) for screening genomic libraries or from genomic clone sequence data (L, M, and N) for sequencing internal regions of the genomic clone. Numbers above oligonucleotides M and N (B) represent the corresponding nucleotide of the alpha-galactosidase A genomic clone (Fig. 25).

3. Oligonucleotide M. Oligonucleotide M was designed to bridge a gap in the DNA sequence not covered by deletion mutants. This unique 15-mer (anti-message sense) spans nucleotides 344 to 358 of the genomic DNA sequence.

4. Oligonucleotide N. This 16-mer (same sense as message) was constructed to cover a gap in the genomic DNA sequence from nucleotides 680 to 696.

Labelling 5'- Ends of Synthetic Oligonucleotides with T4 Polynucleotide Kinase. Prior to use, 50 pmole aliquots of an oligonucleotide were dissolved in 10 mM triethylammonium bicarbonate, lyophilized, and stored at -76°C in 1.5 ml siliconized microfuge tubes. The reaction was carried out in these tubes with the addition of 18 µl distilled water, 5 µl of 10X kination buffer (0.5 M Tris-HCl, pH 7.6 containing 100 mM $MgCl_2$; 50 mM dithiothreitol; 1 mM spermidine and 1 mM EDTA), 2 µl of T4 polynucleotide kinase (10 units/µl) and 25 µl of [γ-^{32}P]dATP. The incubation was performed at 37°C for 30 min. (Maxam and Gilbert, 1980). The incubation mixture was passed over a spun column using Sephadex G-50 (superfine) as described by Maniatis et al. (1982) to remove free ATP.

Selection of Hybridization and Wash Temperatures. For all oligonucleotide mixtures, the melting temperature (Tm) was calculated as described by Wallace (1981): where T= (2°C)(number A/T base pairs) + (4°C)(number G/C base pairs). Since the oligonucleotides are composed of different species, an average Tm was calculated and hybridizations

and washes were done at 5°C to 10°C below the average Tm. For all of these experiments, optimal temperatures were empirically determined by sequential wash experiments of 5°C increments in temperature . After each increase in temperature, the filters were exposed to X-ray film and the autoradiographs were examined for comparison of signal to noise ratios. The calculated Tm ranges for the various oligonucleotide mixtures are: oligonucleotide mixture 1A/1B (60°C to 69°C), oligonucleotide mixture 2A/2B (34°C to 40°C), oligonucleotide mixture 3 (46°C to 53°C), oligonucleotide mixture 4 (40°C to 46°C), and oligonucleotide mixture 5 (57°C to 62°C).

Plating the Okayama and Berg cDNA Library for Screening . The colony hybridization method of Hanahan and Meselson (1980) was used to screen the human fibroblast cDNA library of Okayama and Berg (1983) for clones containing sequences complementary to the synthetic oligonucleotide mixtures. Agar plates (150 mm) containing ampicillin (25 μg/ml) were prepared, and nitrocellulose filters were spread with 0.2 ml of cells to yield 50,000 to 100,000 colonies per plate. This library of 10^6 transformants in E. coli strain χ-1776 (Okayama and Berg, 1983) was screened on 10 plates total for each screening. When colonies reached 0.1 mm in diameter, replica filters were prepared from each master filter. All filters were returned to plates and incubated until colonies were 0.5 mm in diameter. At this point, the master plates were stored at 4°C and served as colony sources. The replica filters were transferred to agar plates containing chloramphenicol (10 μg/ml) for an additional 12 hr to amplify plasmid copy number. The bacteria were lysed in situ by treatment with sodium dodecyl sulfate and NaOH. The DNA was then fixed to the filters by baking in a vacuum oven for

2 hr at 80°C.

Hybridization of cDNA Filters with [γ-^{32}P]-labelled Oligonucleotides. Filters were placed in heat sealable Kapak pouches (8 inches x 9.5 inches), five filters maximum per pouch, with 10 ml of hybridization solution containing 6X SSPE (0.9 M NaCl; 0.06 M sodium phosphate, pH 7.4; 6 mM EDTA), 5X Denhardt's solution (0.1% Ficoll, 0.1% polyvinlypyrrolidone, 0.1% bovine serum albumin, 0.45 M NaCl and 0.045 M sodium citrate, pH 7.0) and 0.5% sodium dodecyl sulfate. The pouches were then sealed and incubated for 2 hr in a 68°C water bath followed by a 2 hr incubation at the appropriate hybridization temperature. One corner of the pouch was cut and the prehybridization solution squeezed out. The pouch was then refilled with 8 ml of hybridization solution containing 300,000 to 10,000,000 cpm/ml of [γ-^{32}P]-labelled oligonucleotide mixture and resealed. Pouches were returned to a shaking water bath at the appropriate temperature and incubated for 16-18 hr.

Processing Filters Following Hybridization. The nitrocellulose filters were washed at the appropriate hybrization temperature using three changes of 1500 ml of 6X SSC (0.9 M NaCl, 0.09 M sodium citrate). Each of the three washes was done in plastic containers placed in a shaking water bath (60 revolutions per minute) for 1 hr. After air drying, autoradiography of the filters was accomplished by exposure to Kodak XAR film with one or two Cronex Lightning-plus intensifying screens at -70°C for various times (several hours to a few days) until the desired exposure was achieved.

Isolation and Purification of cDNA Colonies. Positive colonies, as determined by hybridization experiments were resuspended in broth, and spread on filters at a density of approximately 500 colonies per filter. The screening procedure was repeated until well isolated, hybridization positive colonies were obtained. Plasmid DNA was amplified and recovered as described by Curtiss et al. (1977).

Screening λgt7, λgtWES.λB, and λ1059 Genomic Libraries. Three human X chromosome phage libraries were screened with [γ-^{32}P]-labelled synthetic oligonucleotide mixtures using the in situ plaque hybridization technique of Benton and Davis (1977). The libraries were as follows: i) A λgt7 phage library obtained from Dr. M. Siniscalco and Dr. P. Szabo of the Sloan Kettering Laboratories (Siniscalco et al., 1982), containing 80,000-100,000 different 7-14 kilobases (kb) EcoRI inserts from a mouse (A9/HRBc2)-human hybrid line, ii) a λgtWES.λB library obtained from Dr. R. Williamson of St. Mary's Hospital Medical School, London, (Davies et al., 1981), containing 50,000 recombinants that was prepared from EcoRI digestion of flow sorted human X chromosomes, and iii) a λ1059 library obtained from Dr. Howard J. Cooke of the University of Edinborough, Scotland, (unpublished) containing BamHI digested flow sorted X chromosome DNA with 70,000 individual recombinants. The libraries were plated initially at a density of 20,000 to 30,000 plaque forming units, using 10 plates (150 mm) per library. Plaques were transferred to nitrocellulose filters, denatured in NaOH, neutralized, baked and hybridized as described for colony hybridization above.

Isolation and Purification of λ1059 and λgt7 Genomic Clones. Positive phage, determined by hybridization signals, were isolated using the large end of a sterile pasteur pipette, resuspended in broth, titered, and replated at a density of 500 to 1000 plaque forming units per plate. These plates were rescreened as described above and this purification procedure was repeated until well isolated, hybridization positive plaques were obtained.

Determination of Insert Size from cDNA and Genomic Clones. Plasmid DNA was isolated from positive cDNA clones (Okayama and Berg library) by the alakaline rapid extraction method of Birnboim and Doly (1979) and phage DNA was isolated by the small lysate method of Leder et al., (1977). Plasmid DNA was digested with BamHI restriction enzyme endonuclease to release the cDNA insert. The cDNA insert sizes were estimated by electrophoresis of digested DNA on 1.0% agarose gels along with molecular weight markers of the appropriate size range. Similarly, digestion of the λ clones with either EcoRI (λgt7 clones) or BamHI (λ1059 clones) released the left and right arms from the genomic insert. Size estimates for genomic clones were determined by electrophoresis of digested DNA on 0.7% agarose gels containing the appropriate molecular weight markers (i.e. HindIII, EcoRI, and PstI digested wild type λ DNA).

Analyses of cDNA and Genomic Clones by Southern Hybridization. DNA from positive cDNA and genomic clones was digested with several restriction endonucleases to generate insert fragments of varying lengths. DNA was electrophoresed on 0.7%, 1.0%, or 1.7% agarose gels, stained with ethidium bromide (0.5 μg/ml) to localize bands, and

transferred to Zetabind nylon membrane using 0.025 M sodium phosphate buffer, pH 6.5, according to the method of Southern (1975). The membranes were prewashed in 0.1X SSC (0.015 M NaCl, 0.0015 M sodium citrate) and 0.1% sodium dodecyl sulfate (500 ml per filter) at 68°C for 1 hr. Filters were transferred to heat sealable Kapak pouches and prehybridization, hybridization, and wash conditions were carried out as described above for screening cDNA libraries.

Primer Extension Using Poly(A) mRNA Templates. The 5'- ends of oligonucleotide mixtures 2A/2B and 4 were labelled with [γ-^{32}P]dATP using T4 polynucleotide kinase as described above. The cDNA was synthesized in 50 mM Tris-HCl, pH 8.3, containing 1-2 nmole [γ-^{32}P]-labelled oligonucleotide primer, 130 µg/ml poly(A)$^+$ mRNA, 10 mM dithiothreitol, 50 mM NaCl, 60 µg of bovine serum albumin/ml, 600 µM each of the four unlabelled deoxynucleotide triphosphates and 400 units of reverse transcriptase/ml (Noyes et al., 1979; Sood et al., 1981).

Analysis of mRNA Primer Extension Products. An equal volume (usually 5 µl) of loading buffer (8 M urea, 0.05% xylene cyanol, and 0.05% bromophenol blue) was added to the synthesized cDNA products. The mixture was heated at 90°C for 2 min and layered on a 12.5% (wt/vol) polyacrylamide gel containing 7 M urea. Electrophoresis was performed in 50 mM Tris-borate buffer, pH 8.3 and 1 mM EDTA as described by Noyes et al. (1979) and Sood et al. (1981). Autoradiography was accomplished by exposure of the gel to Kodax XAR film at -70°C for various times (e.g. 3 hr to 2 days). Individual cDNA products were sliced from the gel and sequenced according to the technique of Maxam and Gilbert (1980) by Lisa Traub and Harold

Bernstein. DNA sequence was compared to known amino acid sequence data.

DNA Sequence Analysis of Okayama and Berg cDNA Clones. The cDNA clones which hybridized to more than one of the synthetic oligonucleotide mixtures or those which hybridized with great intensity (above the average Tm), were subjected to DNA sequence analysis by the enzymatic method of Sanger et al. (1977) or the chemical method of Maxam and Gilbert (1980).

Antibody Screening of the λgt11 cDNA Library. Homogeneous alpha-galactosidase A protein (Bishop and Desnick, 1981) was inoculated into rabbits to produce rabbit anti-human alpha-galactosidase A antibodies which were absorbed and titered as described by Calhoun et al. (1985). The human liver cDNA library was provided by T. Chandra and S.L.C. Woo, Baylor College of Medicine. This library, which contains approximately 1.4 X 10^7 independent clones, was plated and screened in Dr. D.F. Bishop's laboraory, as described (Young and Davis, 1983a; de Wet et al., 1984; Young and Davis, 1983b).

Characterization of λgt11 cDNA Clones. Antibody-positive cDNA clones were subjected to antibody competition studies to demonstrate binding specificity and to Southern blot analyses to identify insert fragments that hybridized to synthetic oligonucleotides (Calhoun et al., 1985). One positive clone, designated λAG18 with an EcoRI insert of 1.2 kb, was subcloned to pBR322 by Petros Hantzopoulos and designated pAG18 (Calhoun et al., 1985; Table 1).

Plasmid pAG18 Containes the Authentic Alpha-Galactosidase A cDNA Insert. The method of McGraw (1984) was used to obtain the amino-terminal coding cDNA sequence by primer extension using synthetic oligonucleotide mixture 2B (Fig. 1). Oligonucleotides 2A and 2B were [γ-^{32}P]-labelled as described earlier and annealed separately to heat denatured pAG18 plasmid DNA by placing the mixtures in a 100°C water bath for 3 min and then plunging into ice for 5 min. DNA sequence was obtained by the addition of the Klenow fragment of DNA polymerase I and the appropriate dideoxynucleotides with subsequent electrophoresis on a 20% polyacrylamide gel containing 8 M urea. Electrophoresis was done at 70 Watts (constant power) for 3 hr. The gel was exposed to Kodak XAR film with two Cronex Lightning-plus intensifying screens at -70°C for 18, 36, and 144 hr. The DNA sequence was compared to the known amino terminal amino acid sequence of the alpha-galactosidase A protein for verification of identity. The DNA sequence in this region of the cDNA of pAG18 was confirmed (Petros Hantzopoulos, Ph.D. thesis) using the method of Maxam and Gilbert (1980).

Preparation of λAG18 Lysogen of E. coli strain Y1088. Approximately 7 X 10^8 plaque forming units of λAG18 were added to 100 µl of an overnight culture of E. coli strain Y1088. After adsorption at 37°C for 15 min, the mixture was transferred to a 125 ml Ehrlenmeyer flask containing 20 ml NZ-CYM medium (Maniatis et al., 1982). The culture was incubated at 30°C for five hours then dilutions of 10^{-4} through 10^{-7} were prepared and 100 µl of each dilution was plated on a YT plate (Maniatis et al., 1982) using a sterile glass rod to disperse the liquid. Plates were incubated overnight at 30°C. Well isolated single colonies were transferred to two separate YT plates with sterile

toothpicks in a grid fashion (50 total, each colony streaked on both plates in same numbered grid for identification). One plate was incubated at 42°C and one plate at 30°C. Colonies growing at 30°C but not at 42°C were purified three times at 30°C, with replicas tested at 42°C after each purification round.

Subcloning the 1.2 kb EcoRI Fragment Encoding Alpha-Galactosidase A. In a 10 µl reaction, 1 µg of pUC9 was digested with 10 units of EcoRI in high salt assay buffer (Maniatis et al., 1982) containing 100 mM NaCl, 50 mM Tris-HCl, pH 7.5, 10 mM $MgCl_2$, and 1 mM dithiothreitol). The reaction mixture was incubated at 37°C for 2.5 hr and one third of the reaction was analyzed on a 0.7% agarose gel at 150 volts (V) for 1 hr to test for linearization of the plasmid. To inhibit self ligation, the plasmid DNA was dephosphorylated with calf intestinal phosphatase as described by Maniatis et al. (1982). In a 20 µl reaction, EcoRI digested pUC9 (28 ng) was combined with EcoRI digested λAG18 (500 ng) with 1 unit of ligase in ligase buffer (50 mM Tris-HCl, pH 7.5, 10 mM $MgCl_2$, 20 mM dithiothreitol, 1 mM ATP, and 50 µg of bovine serum albumin/ml). The reaction was incubated overnight at 16°C and transformed into E. coli strain HB101 made competent as described by Hanahan (1983). Plasmid DNA was extracted from transformed colonies (Birnboim and Doly, 1977), digested with EcoRI, and electrophoresed on a 0.7% agarose gel at 150 V for 1 hr to test for presence of the 1.2 kb alpha-galactosidase A cDNA insert. Insert orientation analysis was performed by digestion of plasmid DNA with RsaI followed by agarose gel electrophoresis and comparison of DNA bands (clones were designated pUC9-18-19 and pUC9-18-20; Table 1). The pUC9 recombinant clones with the 1.2 kb alpha-galactosidase A cDNA insert in both orientations were

identified. The 1.2 kb alpha-galactosidase A EcoRI fragment was cloned into pUC8 as described for cloning into pUC9, however, recombinant plasmids were transformed into E. coli strain C600 (clones were designated pUC8-18-5 and pUC8-18-3; Table 1).

Construction of M13mp18 Deletion Derivatives. The 1.2 kb EcoRI fragment of plasmid pAG18 was subcloned into M13mp18 (Messing, 1983) in both orientations and deletion subclones were made according to the method of Dale et al. (1985) in Dr. D.F. Bishop's laboratory.

Nucleotide Sequence Analysis of M13mp18 Deletion Subclones. The deletion subclones (M25.21, M25.22, M25.28, M27.15, M27.18, M27.30, and M27.5; Table 1) were sequenced by the method of Sanger et al. (1977) using the M13 universal primer (17-mer).

Computer Analyses of the Alpha-Galactosidase A cDNA Insert. Overlapping nucleotide sequences were aligned on a microcomputer using the MicroGenie program (Beckman). Protein structural analyses and DNA and amino acid homology searches of the National Institutes of Health GenBank and the National Biomedical Research Foundation protein data base were performed in February, 1986, using IFIND/ALIGN programs on the Bionet Network (Intelligenetics).

Screening the Charon 30 Genomic Library with a Unique 5'-Oligonucleotide Probe. A Charon 30 4X human genomic library, karyotyped 49, XXXXY, was obtained from William Wood (Wood et al., 1984). The library was plated at a density of 2 X 10^4 phage per 150 mm petri dish (15 dishes total) and screened at 55°C with [γ-^{32}P]-labelled

oligonucleotide H (Fig. 2). Phage DNA was transferred to nylon membranes by a procedure adapted from the method of Benton and Davis (1977). A nylon filter was placed on the agar for two to three min and orientaion marks were established by stabbing a needle through the filter and agar in an asymmetric pattern. The filter was then transferred to a glass dish containing 200 ml of 0.5 M NaOH. After two min the filter was transferred to a second glass dish containing 200 ml of 1.0 M Tris-HCl, pH 7.5. The filter was neutralized for two min and then placed on blotting paper to dry at room temperature. Hybridization and wash procedures were done as described above. Two strongly hybridizing phage, designated λMQ1 and λMQ2 (Table 1), were identified and plaque purified three times.

Characterization of Charon 30 Genomic Clones. Phage DNA, isolated by the liquid culture method (Maniatis et al., 1982), was analyzed by endonuclease digestion with various restriction enzymes and by Southern blot hybridization analyses. DNA was electrophoresed on 0.7% agarose gels, stained with ethidium bromide (0.5 μg/ml) to localize DNA bands, and transferred to Zetabind nylon membranes using 0.025 M sodium phosphate buffer, pH 6.5. These blots were screened with end-labelled probe H at 55°C, probe 2B at 32°C, or nick translated (Rigby et al., 1977) DNA from pAG18 at 68°C.

Subcloning a Genomic Fragment from Phage λMQ1 to Plasmid pSPRI. Digestion of phages λMQ1 and λMQ2 with SacI generated DNA fragments of approximately 5.3 kb that strongly hybridized to probe H, probe 2B, and nick translated pAG18. This SacI fragment of phage λMQ1 was subcloned into the SacI site of plasmid pSPRI (Krystal et al., 1986), provided by

Mark Krystal. Recombinant clones were detected by electrophoresis of undigested plasmid DNA on 0.7% agarose gels a 30 V for 18 hr. Undigested DNA samples migrating slower than undigested pSPRI vector DNA were analyzed for inserts using restriction endonucleases. Digestion of plasmid DNA with SacI followed by Southern blot hybridization analyses using oligonucleotide H, 2B, or nick translated plasmid pAG18 as described earlier, identified two pSPRI clones (pMQ1 and pMQ2; Table 1) containing the 5.3 kb SacI fragment. Southern blot hybridization analysis of TaqI digested plasmid DNA from pMQ1 and pMQ2, using oligonucleotide H as a hybridization probe, identified a strongly hybridizing subfragment of approximately 1.2 kb.

Subcloning a 1.2 kb TaqI Genomic Fragment into M13mp11. Approximately 50 µg of DNA from plasmid pMQ1, digested with TaqI, was loaded into a preparative well (4 cm wide) and electrophoresed on 1.7% agarose gel at 150 V for 2.5 hr. A portion of the lane (approxiamtely 10%) was excised from the gel, stained with ethidium bromide, and photographed with a ruler along the edge. The 1.2 kb TaqI fragment was identified and localized on the ruler. A DEAE cellulose membrane was inserted 2 cm in front of the unstained portion of the gel containing the 1.2 kb TaqI fragment. Electrophoresis was continued at 200 V for 10 min. DNA was eluted from the membrane as described recommended by the manufacturer. Specifically, the DEAE cellulose membrane was placed in a 1.5 ml microcentrifuge tube with enough high salt buffer (0.15M NaCl; 0.1 mM EDTA; 20 mM Tris, pH 8) to cover the membrane. The tube was spun for 5 seconds in a microcentrifuge to submerge the membrane, and incubated at 68°C for 20 min with occasional swirling. The buffer was removed with a pasteur pipette and the membrane washed with an

additional 100 µl of high salt buffer. DNA was precipitated with 2.5 volumes of ethanol (20 min at -70°C) and precipitated with 0.3 M sodium acetate to remove residual NaCl. The purified TaqI fragment of 1.2 kb was subcloned into the AccI site of M13mp11 (Messing, 1983) in both orientations.

Analysis of M13mp11 Clones for the Charon 30 Genomic Insert. M13mp11 derivatives were analyzed for the presence of the 1.2 kb TaqI genomic fragment from plasmid pMQ1. The replicative form of M13mp11 clones were digested with TaqI and electrophoresed on 0.7% agarose gels. The orientations of the TaqI fragments of randomly chosen clones, was determined by the dideoxy sequencing method (Sanger et al., 1977). Clones were designated mMQU (packages the strand with the same sense as message) and mMQL (packages the opposite strand, anti-message sense). The polylinker regions of plasmid pSPRI and M13mp11 were removed from clone mMQU to eliminate the potential formation of a stem and loop hairpin structure by the single stranded DNA. The replicative form of DNA from clone mMQU was digested with SacI, ligated with T4 ligase, and transformed into E.coli strain JM103. DNA from M13mp11 clones were analyzed for the presence of BamHI and SmaI restriction enzyme sites, which are only present in the polylinker regions of pSPRI and M13mp11. Clones which had lost the polylinker regions were designated mMQZ (Table 1).

Nucleotide Sequencing of the Charon 30 Genomic Fragment. Deletion clones of M13mp11 derivatives, mMQU and mMQL, were generated by the method of Dale et al. (1985). Deletion clones were sized by electrophoresis of single stranded DNA on 0.7% agarose gels at 30 V for

18 hr and compared to DNA from mMQU (undeleted 1.2 kb insert) and M13mp11 (no insert). Seventeen deletion clones were sequenced by the method of Sanger et al. (1977) using [α-^{35}S]dNTPs and the M13 universal primer (17-mer) and the RD22 oligonucleotide. Five synthetic oligonucleotide primers (Figs. 1, 2 and 24) were utilized to complete the sequence not obtained from deletion mutants or to confirm sequence already obtained.

Computer Analyses of the Charon 30 Genomic Fragment. Overlapping nucleotide sequences were aligned using the MicroGenie program (Beckman). DNA and amino acid homology searches were initially done using FASTP and FASTN (Lipman and Pearson, 1985) with the Protein Identification Resource Protein Library (version 7) and the GenBank nucleotide sequence data bank (version 40), respectively. The data base search was repeated using MicroGenie with the September, 1986 releases of the GenBank and National Biomedical Research Foundation data banks. Signal peptide protein structural analysis and hydrophilicity matrix analyses were performed using the MicroGenie program. The region of homology flanking the ATG initiation codon of human lysosomal genes was detected using MicroGenie. A compilation of signal peptide sequences (Watson, 1984) was used as a guide to obtain the corresponding nucleotide sequences of 42 human signal peptides. Subsequently, the Quest routine of Bionet was used to generate a file containing 133 human genes with complete nucleotide sequences of the signal peptides, and these were examined for homology to the human alpha-galactosidase A signal peptide.

RESULTS

Amino Acid Sequence of Alpha-Galactosidase A. Treatment of the purified enzyme with N-glycanase, reduced the molecular weight of the deglycosylated, monomeric enzyme from approximately 45,000 to approximately 41,800 (Fig. 3). The amino acid composition analyses of two preparations of human alpha-galactosidase A were determined (Table 2). Based on an estimated subunit molecular weight of 41,800, it was calculated that the alpha-galactosidase A subunit contains 370 amino acids. Microsequencing of the mature enzyme provided amino-terminal sequence of 23 residues (Fig. 1). In addition, two cyanogen bromide and five tryptic peptides were subsequently sequenced, providing amino acid sequence data for a total of 101 residues (approximately 27% of the mature enzyme). Synthetic oligonucleotides were constructed based on the amino acid sequence data in regions of low codon redundancy (Fig. 1).

Screening the Okayama and Berg cDNA Library with Oligonucleotides. The first oligonucleotide mixture synthesized was 2A and 2B (2A/2B), and each were comprised of four different 14-mers with a Tm range of 34°C to 40°C. Hybridization conditions were selected 5°C to 10°C below the lowest Tm, and the stringency was increased during screening and purification to identify clones that most likely had a perfect match to one member of the oligonucleotide mixture. These conditions would be expected to identify the authentic alpha-galactosidase A cDNA clone, as well as cDNA clones, that, by chance, hybridized to one or more members

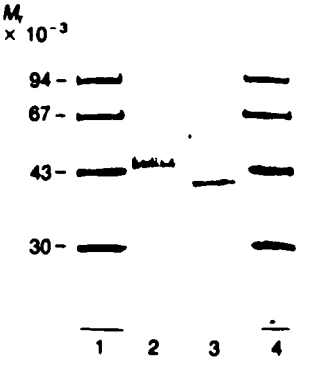


Fig. 3. N-glycanase digestion of human lung alpha-galactosidase A. Lanes 1 and 4, 1.0 µg each of the standards: phosphorylase b, 94,000; bovine serum albumin, 67,000; ovalbumin, 43,000; and carbonic anhydrase, 30,000. Lane 2, 0.40 µg of purified alpha-galactosidase A. Lane 3, 0.34 ug deglycosylated alpha-galactosidase A. The deglycosylated enzyme migrated with an apparent molecular weight to 41,800 (from Calhoun et. al., 1985).

Table 2. Amino acid composition of alpha-galactosidase A.
Moles residues per mole subunit

Residue	Prep 1	Prep 2	Average integral no.	Predicted mature protein
Asx	41.2	45.7	44	46
Thr	16.1	13.4	15	11
Ser	24.2	21.7	23	21
Glx	38.9	35.8	37	37
Pro	19.1	18.4	19	18
Gly	28.9	29.8	29	30
Ala	25.8	27.2	27	28
Val	19.2	13.7	16	15
Met	7.4	8.7	8	13
Ile	15.3	18.8	17	20
Leu	36.3	37.7	37	34
Tyr	15.9	14.3	15	15
Phe	14.9	14.3	15	15
His	7.0	9.0	8	6
Lys	16.1	15.1	16	16
Arg	18.7	15.5	17	17
Cys	9.1	9.6	9	12
Trp	14.9	13.4	14	16

Amino acid composition analysis predicts a subunit molecular weight of 41,800 with an estimated 370 amino acid residues per alpha-galactosidase A monomer. This is compared to the amino acid composition of the predicted mature protein (based on genomic nucleotide sequence and carboxy-terminal processing at amino acid 371 (Arg) of the propeptide, see text). The molecular weight of the predicted mature protein is 42,307 and contains 370 amino acids residues.

of the oligonucleotide mixture. Our rationale was to select clones with a second oligonucleotide mixture (i.e. 4) to eliminate spurious clones from further analyses. Subsequently, other synthetic oligonucleotide mixtures were constructed to various segments of the peptide, since there was uncertainty to the identity of several amino acid residues. This uncertainty was due to the limited quantities of purified enzyme available for microsequencing. The human fibroblast cDNA library of Okayama and Berg (1983) was screened for clones with sequences complementary to oligonucleotide mixture 2A/2B (Fig. 4). Upon colony purification (Fig. 5) and rescreening, 20 transformants, named pcD1-20, were identified. Restriction enzyme analyses of plasmid DNA from clones pCD1-20 digested with SalI (Fig. 6), PstI, BamHI, HaeIII, and Sau3AI indicated that the cDNA inserts ranged from 100 to 2500 base pairs.

Southern Blot Hybridization Analyses of plasmids pcD1-20. Southern blot hybridization with [γ-^{32}P]-labelled oligonucleotide mixture 2A/2B to DNA from plasmids pcD1-20, digested separately with Sau3AI or HaeIII, revealed that these clones did not share common restriction fragments. Hybridization of these filters with [γ-^{32}P]-labelled oligonucleotide mixture 4 revealed binding to DNA from clone pcD1, but not to DNA from clones pcD2-pcD20. This result was confirmed by duplicate Southern blot analyses of DNA from clone pcD1 digested with HaeIII or Sau3AI hybridized to [γ-^{32}P]-labelled oligonucleotide mixtures 2A/2B or 4 (Fig. 7). This experiment indicated that oligonucleotide 2A/2B and 4 hybridized to the same HaeIII and Sau3AI fragments.

Fig. 4. Screening of the Okayama and Berg (1983) cDNA library. Panels A and B are autoradiographs obtained after exposure to filters washed at 25°C or 28°C, respectively. Arrows indicate colonies which were chosen for purification. The arrows in panel C show some of the single colonies selected for further analyses. Panel D shows hybridization with oligonucleotide 2A/2B at 25°C (H. Bernstein, unpublished) to a λgt7 genomic library (Siniscalco et al., 1982).

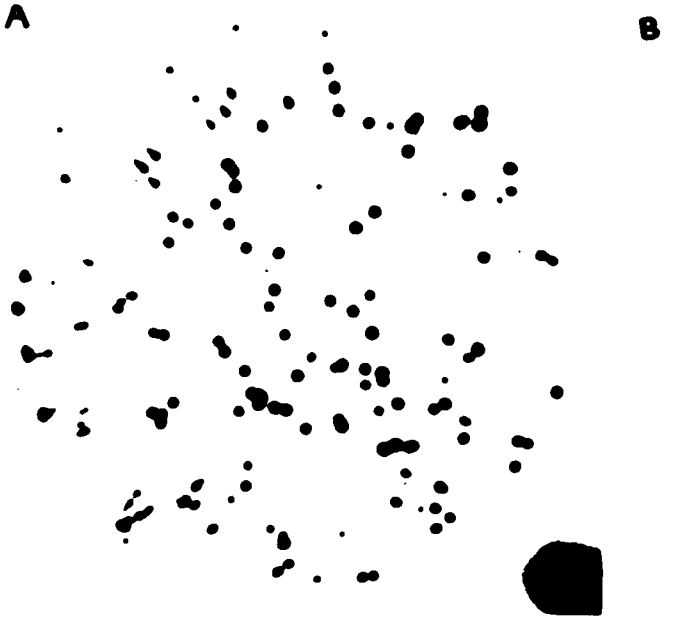


Fig. 5. Purification of clone pcD1. Panel A shows an autoradiograph after hybridization with [γ-32P]-labelled oligonucleotide mixture 2A/2B (Fig. 1) at 28oC. Panel B shows an autoradiograph of a parallel experiment to clone pcD19, a purified negative control. Filters were exposed to Kodak XAR film at -70oC for 18 h. Culture was considered pure when every colony on the filter hybridized to the oligonucleotide. Arrow idicates a well isolated single colony selected for subsequent experiments.

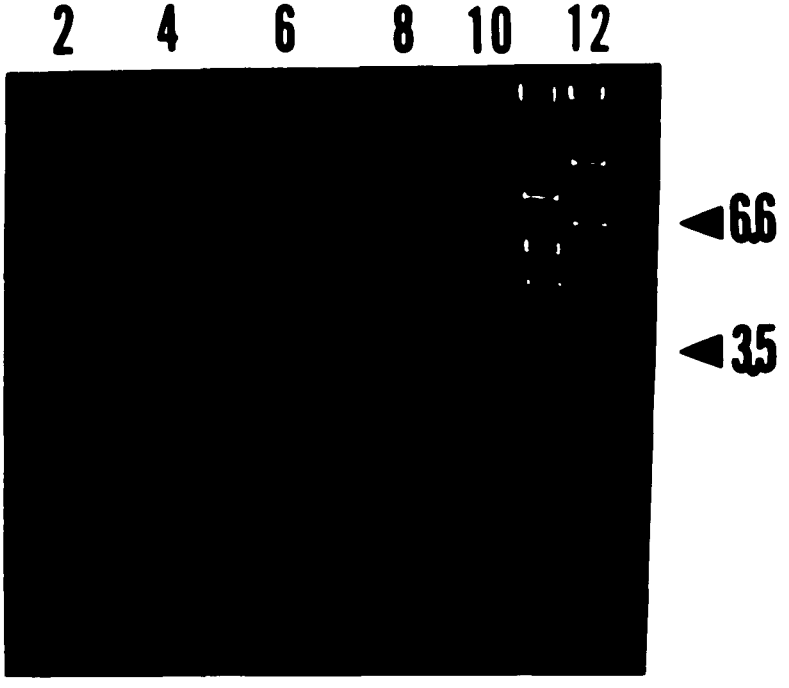


Fig. 6. SalI restriction enzyme analysis of plasmids pcD1-10. Lanes 1-10; pcD1-10 DNA digested with SalI. Lanes 11-13; molecular weight markers of lambda DNA digested with EcoRI, HindIII, and PstI, respectively. Arrows indicate (in kb) size of corresponding DNA fragment.

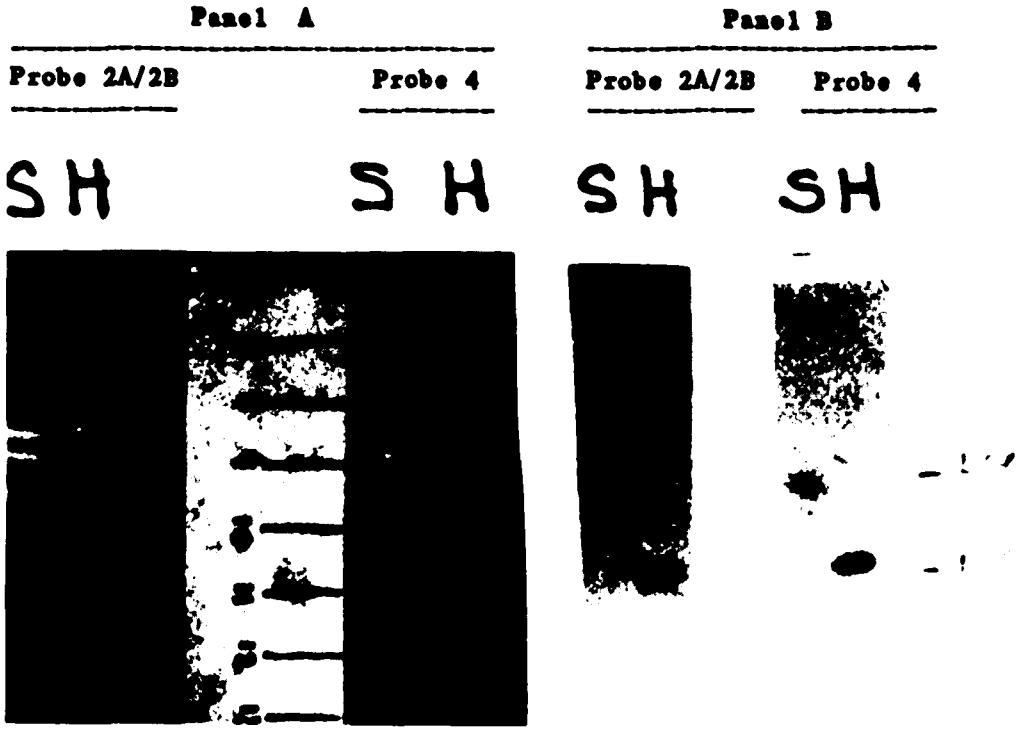


Fig. 7. Southern blot analyses of pcD1 digested with Sau3AI (S) or HaeIII (H). Panel A shows parallel lanes of a 1.7% agarose gel stained with ethidium bromid before Southern blot analysis. Panel B shows the autoradiographs after transfer and hybridization to probe 2A/2B at 25oC or probe 4 at 28oC. The Sau3AI restriction enzyme digest of pcD1 identified a band of approximately 400 nucleotides and the HaeIII restriction enzyme digest identified a band of approximately 150 nucleotides that bound both probes.

Nucleotide Sequence of the cDNA from plasmid pcD1. Additional Southern blot hybridization analysis of BamHI digested pcD1 plasmid DNA revealed a 411 base pair fragment which showed hybridization to oligonucleotide mixtures 2A/2B and 4. This 411 base pair fragment was cloned into M13mp7 and sequenced by the method of Sanger et al. (1977) to determine if it encoded the known amino terminal amino acid sequence. It was found that the BamHI fragment from clone pcD1 had a 13 of 14 base pair match to one member of oligonucleotide mixture 2A/2B and a 12 of 14 base pair match to one member of oligonucleotide mixture 4 (Fig. 8). Using the empirical formula of Wallace (1981), minus 10°C for an internal mismatch, the estimated Tm for plasmid pcD1 with the most homologous member of oligonucleotide mixtures 2A/2B and 4 were 24°C and 30°C, respectively. This result illustrates an inherent limitation of the mixed oligonucleotide approach for screening libraries, since it is impossible to distinguish between a perfect match with an A-T rich sequence, compared to a partial match to a G-C rich sequence. We interpreted these results to indicate that an authentic full length alpha-galactosidase A cDNA clone was not detected by oligonucleotide mixture 2A/2B during the screening procedure and the library was rescreened.

Rescreening the Okayama and Berg cDNA Library at Increased Stringency. The cDNA library of Okayama and Berg (1983) was rescreened at the increased stringency of 31°C with [γ-^{32}P]-labelled oligonucleotide mixtures 2A/2B. One transformant, designated pcD21, was identified that hybridized to probe 2A/2B at 34°C following colony

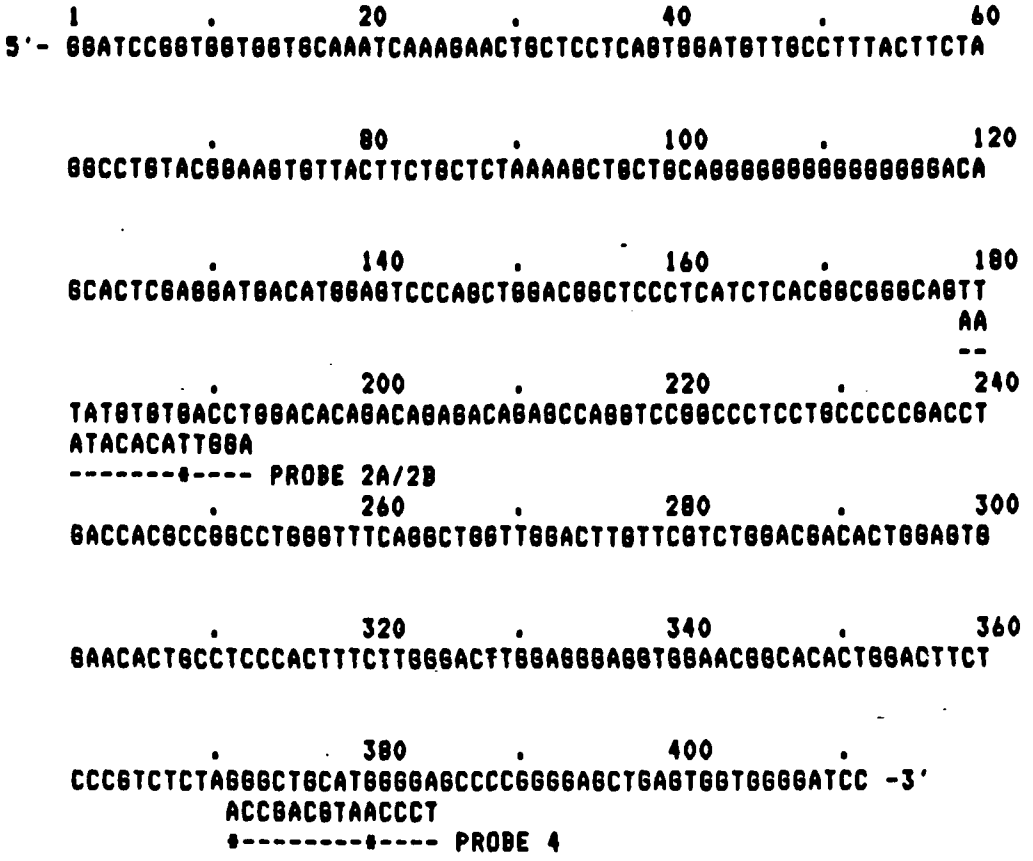


Fig. 8. Nucleotide sequence of the cDNA of plasmid pcD1. Duplicate Southern blot analysis revealed that a 411 nucleotide BamHI fragment bound oligonucleotide mixture 2A/2B at 25°C and oligonucleotide mixture 4 at 28°C (Fig. 7). This fragment was cloned into M13mp7 and the complete nucleotide sequence is shown. Complementary sequences were identified in the nucleotide sequence of pcD1 with a 13 of 14 base pair match to one member of oligonucleotide mixure 2A/2B and a 12 of 14 base pair match to one member of oligonucleotide mixture 4. The dashed (—) line indicates a match and the asterick (*) indicates a mismatch in the sequence.

purification and rehybridization. Southern blot hybridization analysis of clone pcD21 digested with several restriction enzymes, revealed a 220 nucleotide HinfI-HaeIII fragment which hybridized to probe 2A/2B. Subsequent nucleotide sequence of this 220 base pair fragment by the method of Maxam and Gilbert (1980), indicated a perfect match to one member of oligonucleotide mixture 2A/2B (Fig. 9), with a calculated Tm of 38°C. However, the flanking nucleotides of clone pcD21 did not correspond to the known amino acid codons for human alpha-galactosidase A. This result illustrates another limitation of the mixed oligonucleotide approach for screening libraries, specifically, undesired clones with a perfect match to a member of the oligonucleotide mixture can be selected. Although these first two clones selected from the Okayama and Berg library were not authentic, the results did provide evidence that the techniques used were adequate to identify the desired clone if it were present and well isolated. In addition, plasmids pcD1 and pcD21 provided internal controls for subsequent screenings using oligonucleotide 2A/2B.

Screening of λgt7 and λ1059 Genomic Libraries. The mouse-human X chromosome λgt7 phage library (Sinescalco et al., 1982) was screened using oligonucleotide mixture 2A/2B at 28°C, and 10 positive clones were identified and plaque purified. Restriction enzyme analysis of these phage indicated that the insert sizes ranged from 6.8 kb to 9.6 kb. Duplicate Southern hybridization analyses of these 10 EcoRI digested λgt7 phage to probes 2A/2B and 4, indicated that only one phage showed positive hybridization signals to both probes. However, at 31°C, the hybridization signals to both oligonucleotides became

A.

```
                                  15                  20
alpha-
galactosidase A   TrpLeuHisTrpGluArgPheMetCysAsnLeuAspCysGlnGlnGluPro
                   X  X  X  X  X  X  ||||||||||||||| X ||| X  X  X  X
pcD21             AlaSerAspGlyArgAsnPheMetCysAsnLeuSerCysThrThrLeuPhe
```

B.

```
pcD21             GGCTAGTGACGGTAGGAATTTCATGTGTAACCTAAGAACAACTCTCTTCA
                  CCGATCACTGCCATCCTTAAAGTACACATTGGATTCTTGTTGAGAGAAGT
                   ••••••••••••••••
```

Tm= 38°C

Fig. 9. Nucleotide sequence of the cDNA of plasmid pcD21. Shown is a portion of the nucleotide sequence (B) flanking the probe 2B complementary region (•••) of plasmid pcD21. The calculated Tm for this perfect match is 38 C. Beneath the amino terminal amino acid sequence of alpha- galactosidase A is shown the deduced amino acid sequence from the nucleotide sequence of pcD21 (A). A match in the amino acid sequence is indicated by vertical bars (|||) and a mismatch by an X between the two residues.

faint for this clone. Also, when phage DNA from this clone was tested in Southern blot analysis using oligonucleotide 1A/1B (Fig. 1) at 51°C, no hybridization to this probe was detected even though the most A-T rich member (Tm range 57°C to 62°C) of the oligonucleotide mixture would bind at this temperature. Therefore, the clone was not studied further.

The flow-sorted, human X chromosome λ1059 library (H.J. Cooke, unpublished) was screened for phage containing sequences complementary to [γ-^{32}P]-labelled oligonucleotide mixture 1A/1B at 51°C, and two positive phage (λC1 and λC2) were identified (Fig. 10). These phage did not hybridize to oligonucleotides 2A/2B or 4 with subsequent Southern blot analysis. For this reason, clones λC1 and λC2 were not characterized further.

Oligonucleotide Primer Extension of mRNA. Oligonucleotide mixtures 2A and 2B were used to separately to prime reverse transcriptase in the presence of mRNA from HeLa and KNH3 cells. Three major extension products were detected using mixture 2B, but none were seen with mixture 2A tested in parallel. The nucleotide sequence of the three major cDNA products indicated that they did not correspond to the human alpha-galactosidase A mRNA. Similarly, the use of [γ-^{32}P]-labelled oligonucleotide mixture 4 as a primer for HeLa or KNH3 mRNA yielded three major cDNA products along with several less intense bands. The nucleotide sequence analysis of these major cDNA products indicated that they were not synthesized from the human alpha-galactosidase A mRNA, since the predicted amino acid sequence differed from that obtained from the purified enzyme. Primer extension

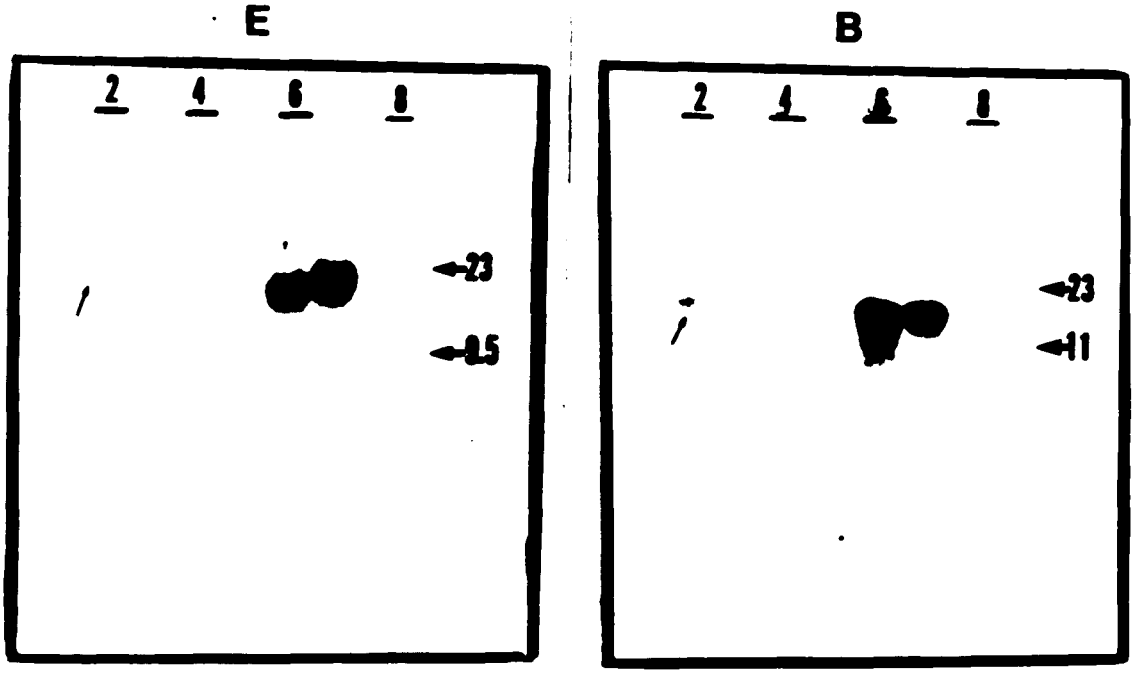


Fig. 10. Southern blot analysis of phage isolated from the λ1059 genomic library (Cooke, unpublished). The left and right panels show phage DNA digested with EcoRI (E) or BamHI (B), respectively. Both filters were hybridized to oligonucleotide mixture 1A/1B at 51°C. Lanes 2-7 contain phage DNA, and lanes 8-9 contain molecular weight markers of λ DNA digested with HindIII and PstI, respectively. Arrows indicate size (in kb) of DNA band migrating at that position. Two strongly hybridizing phage (lanes 6-7) were identified and designated λC1 and λC2. One weakly hybridizing band was seen in both blots (see arrow), however, this was considered too faint at 51°C for further analysis.

using oligonucleotide mixture 5 (32 different 17-mers) resulted in a complex mixture of cDNA products that were not suitable for DNA sequence analysis.

Isolation and Characterization of Phage λgt11 cDNA clones. A λgt11 cDNA expression library derived from human liver and containing 1.4×10^7 independent clones, was screened using the antibody detection method (Calhoun et al., 1985). Four antibody-positive clones were isolated (designated λAG2, λAG14, λAG15, and λAG18). Of the four clones, only λAG18 showed positive hybridization in Southern blot analyses to oligonucleotide mixtures 1A/1B, 2A/2B, and 3. The size estimate of the EcoRI insert of λAG18 was 1250 base pairs, sufficient to encode the entire mature enzyme subunit (370 amino acids).

Initial Nucleotide Sequence Analyses of Plasmid pAG18. The EcoRI cDNA insert of clone λAG18 was subcloned to plasmid pBR322, and DNA from this clone, designated plasmid pAG18, was used directly as a template for nucleotide sequence analysis. When [γ-^{32}P]-labelled oligonucleotide mixure 2B was used to prime the Klenow fragment of DNA polymerase I using heat denatured plasmid pAG18 DNA as a template and the dideoxy sequencing protocol of Sanger et al. (1977), a readable sequence was obtained. However, no discrete extension products were detected when oligonucleotide mixture 2A was used as a primer in parallel reactions. The DNA sequence obtained using oligonucleotide mixture 2B as a primer predicted the known amino terminal amino acid sequence for alpha-galactosidase A (Fig. 11). This nucleotide region of clone pAG18 was confirmed (Petros Hantzopoulos, Ph.D. thesis) using the chemical method of Maxam and Gilbert (1980). This identified λAG18

as an authentic human alpha-galactosidase A cDNA clone, and justified an effort to obtain the nucleotide sequence of the complete 1250 nucleotide EcoRI cDNA segment.

Nucleotide Sequence Analysis of the cDNA from Phage λAG18. Fig. 12 shows the strategy for sequencing the cDNA from phage λAG18 using a set of deletion subclones derived from the full length insert cloned to M13mp18 in both orientations. Fig. 13 shows an autoradiograph obtained as a part of the sequencing project, which was a concerted effort (P. Hantzopoulos, H. Bernstein, and M Quinn). The sequence was confirmed in its entirety on both strands from the deletion derivatives (Figs. 12, 13), with the exception of a short segment (nucleotides 939-1004) on the message strand that was sequenced by primer extension using a synthetic oligonucleotide (17-mer). Other regions, including the 3'- and 5'- ends, were confimed by primer extension of pAG18 and by Maxam and Gilbert analysis as described (Bishop et al., 1986).

The complete 1226 nucleotide sequence of the alpha-galactosidase A cDNA obtained from phage λAG18 and its deduced amino acid sequence are shown in Fig. 14. This sequence contains an open reading frame from positions -15 to 1194 followed by a TAA termination codon. A poly(A)$^+$ tail of 12 nucleotides immediately follows the termination codon. Two hexanucleotide poly(A)$^+$ addition signals, AATACA and ATTAAA, are located 28 and 11 nucleotides prior to the TAA stop codon, respectively. In addition, the putative U4 small nuclear ribonucleoprotein recognition sequence, CAGCT, which may be involved polyadenylation (Berget, 1984), is present 38 nucleotides upstream of the stop codon. The absence of a 3'- untranslated region was also

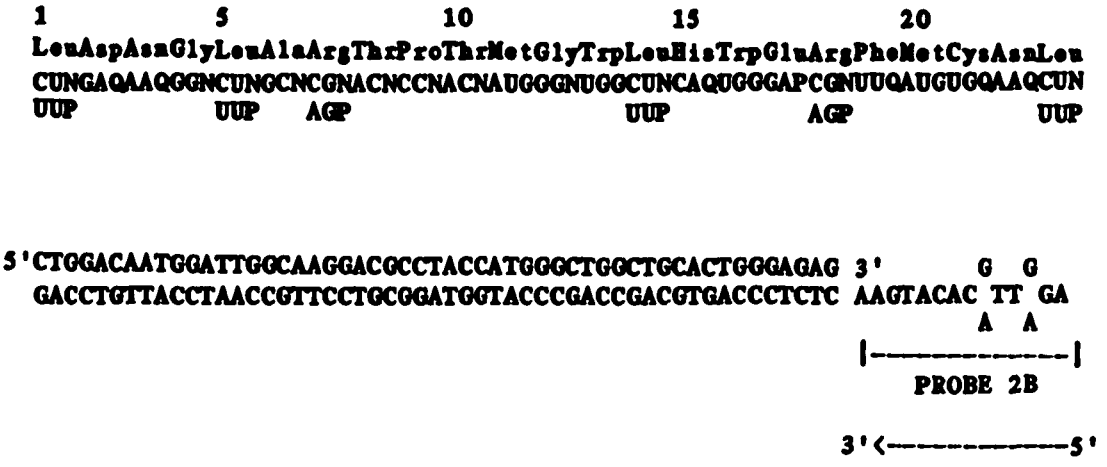


Fig. 11. Nucleotide sequence of the 5'-end of the cDNA of plasmid pAG18. The top line shows the first 23 amino acids of the amino terminal alpha-galactosidase A protein. Beneath the amino acid sequence, the possible nucleotide codons are shown for comparison to the nucleotide data (Q= C or T; N= any base; P =A or G). Oligonucleotide mixture 2B (4 different 14-mer), corresponding to amino acid 19 to 23, was end-labelled with [γ-^{32}P]dATP and used to prime heat denature DNA from plasmid pAG18 in the presence of the klenow fragment of DNA polymerase I and dideoxy nucleotides (see Methods). The nucleotide sequence read from the sequencing gel is shown (arrow designates the direction of the sequence reaction when priming with probe 2B). The nucleotide sequence was colinear with the amino acid sequence.

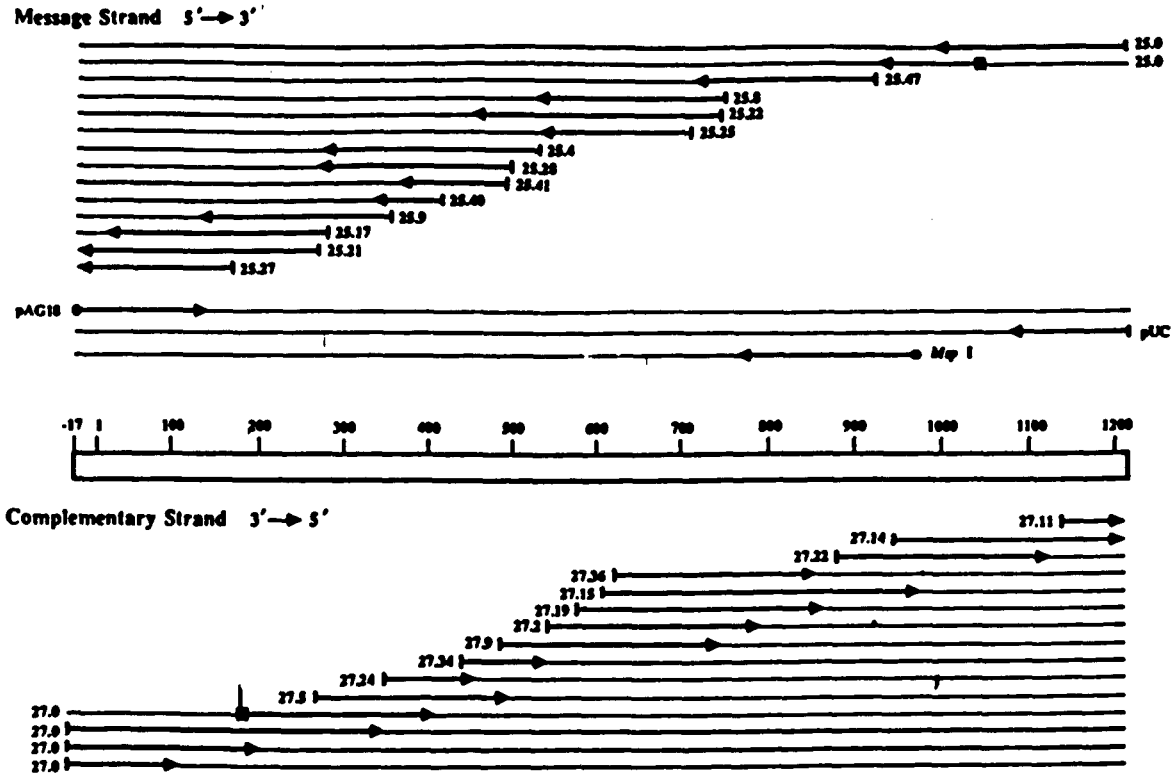


Fig. 12. Strategy for sequencing the alpha-galactosidase A cDNA. M13mp18 clones M25 and M27 package the entire message and complementary strands, respectively, from λAG18 cDNA EcoRI insert. Full length clones were used to generate deletion clones (e.g. 25.21 and 27.18) and the entire length of each deletion clone is shown; heavy arrows indicate the extent and direction of the sequence determined per clone. Arrows originating from vertical bars designate sequence data obtained by the enzymatic method of Sanger et al. (1977) using the M13 universal primers and sequence data obtained from synthetic internal oligonucleotides are shown with solid squares. Arrows originating from closed circles indicate sequence obtained by the chemical method of Maxam and Gilbert (1980). Figure taken from Bishop et al., 1986.

Fig. 13. Autoradiograph of a sequencing gel of clone M25.28. The sequence was obtained by the enzymatic method of Sanger et al. (1977) using the M13 universal primer (17-mer). The sequence reads from bottom to top, 5'- AAACTGTCACAGTAACAACCATCAAATTTTAGCAGATCTA CTCCCCAGTCAGCAAAGGTCTGGGCATCAATGTCGTAGTATCCAAAACTCCCAGGGAGCCTGCGCAGGTT TTTATTTCC -3'.

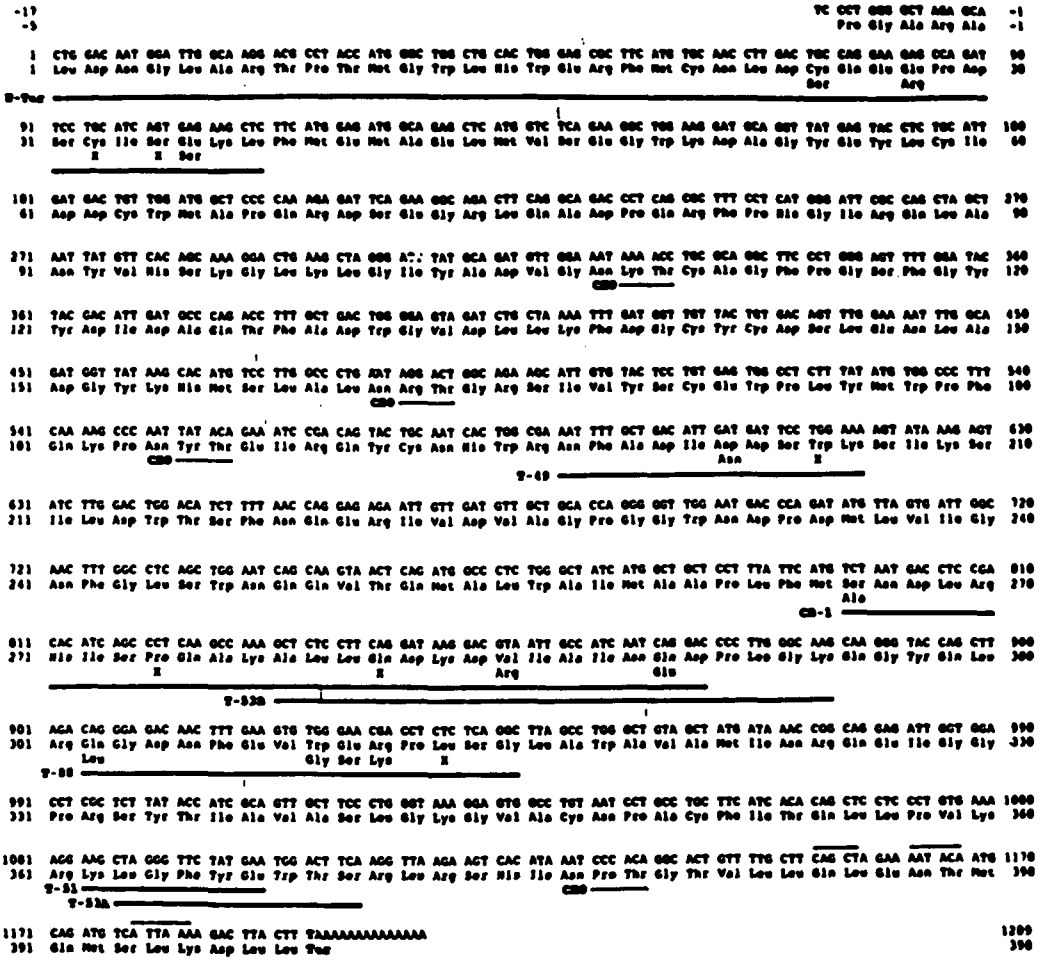


Fig. 14. Combined nucleotide and predicted amino acid sequences of the λAG18 cDNA sequence encoding the alpha-galactosidase A propeptide. Amino acid 1 is assigned to the amino terminal residue determined from microsequencing of the mature protein. Nucleotides -17 to -1 encode the five carboxy terminal amino acids of the signal peptide. Bold underlines indicate confirmed amino acid sequence from tryptic (T) and cyanogen bromide (CB) peptides. Discrepancies between the predicted amino acids and the microsequenced are shown; X denotes unidentified amino acids. CHO denotes potential N-glycosylation sites. Overlines indicate the poly(A) signals AATACA and ATTAAA, and the CAGCT site postulated in U4 small nuclear RNA binding. Figure taken from Bishop et al., 1986.

seen in an independently isolated alpha-galactosidase A cDNA clone from human lung (Bishop et al., 1986).

Predicted Amino Acid Sequence of Alpha-Galactosidase A. The cDNA includes an open reading frame of 1194 nucleotides that encodes 398 amino acids corresponding to a predicted molecular weight of 45,346. Nucleotide positions 1-3 were assigned to the amino-terminal leucine residue of the microsequenced mature enzyme. The nucleotide sequence agrees with 86 of 100 amino acid residues determined by microsequencing of five tryptic peptides, one cyanogen bromide peptide, and the amino-terminal sequence of the homogeneous protein. Minor differences were observed between the amino acid sequences and those predicted from the cDNA sequence (shown in Fig. 14), presumably, due to the limited quantities of protein available for microsequencing. Also shown in Fig. 14 are the four possible glycosylation sites (Asn-Xxx-Ser/Thr) for asparagine-linked oligosaccharides at asparagine residues.

Protein Structural Analysis of Alpha-Galactosidase A. The protein structural analysis of the alpha-galactosidase A peptide (398 amino acids) is shown in Fig. 15. Alpha helix regions of 10 or more residues are located at positions 31-50, 146-155, 253-267, 275-284, and 385-398. The longest beta sheet regions found are three stretches of 7 residues located at positions 234-240, 334-340, and 351-357. Possible N-glycosylation sites are located at beta turns as expected for surface localization.

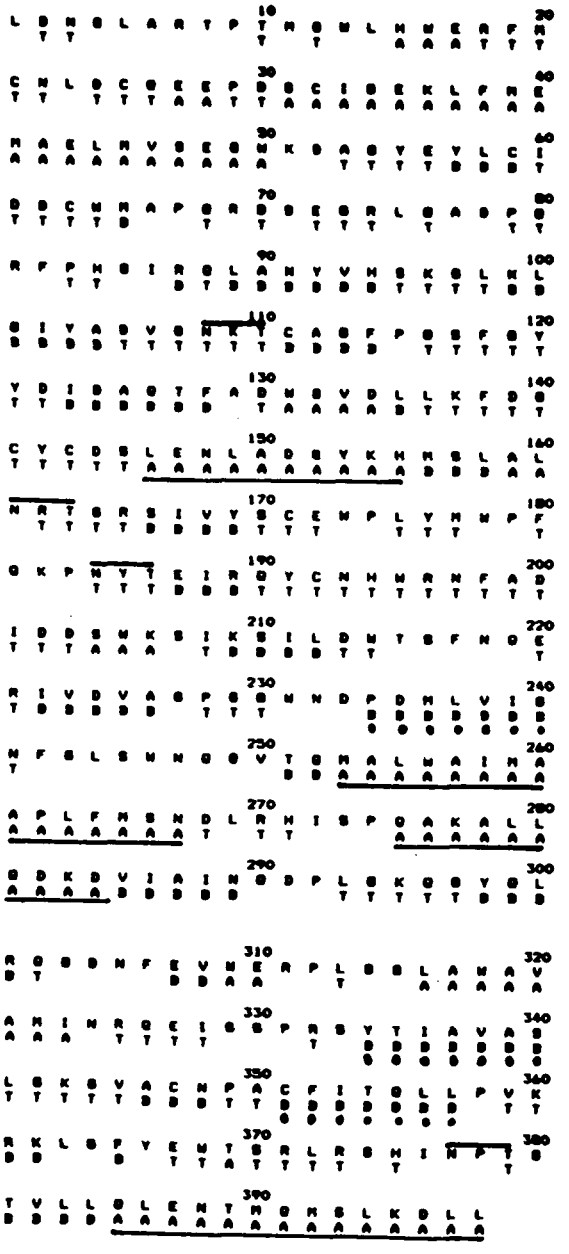


Fig. 15. Protein structural analysis of human alpha-galactosidase A propeptide of 398 amino acids. Stretches of 10 or more alpha helices (A) are underlined. The three longest beta sheets (B) are dotted and the four possible N-glycosylation sites occuring at beta turns (T) are overlined.

Restriction Enzyme Analysis of the Alpha-Galactosidase A cDNA. The restriction enzyme analysis of of the cDNA from phage λAg18 encoding alpha-galactosidase A, generated by the MicroGenie program, is shown in Fig. 16. The entire sequence is shown with the enzyme site listed beneath the numbered nucleotide sequence. The numbering system is the same as that used in Fig. 14.

Isolation of Genomic Clones from a Charon 30 Library. A Charon 30 library derived from human genomic 4X lymphoblastic DNA (Wood et al., 1984) was screened using a synthetic oligonucleotide probe (probe H; Fig. 2) which has a sequence derived from the cDNA of phage λAG18. Probe H corresponds to the last six nucleotides encoding the signal peptide and the first seventeen nucleotides coding for the mature enzyme. Since the λAG18 cDNA clone, and other clones subsequently isolated, did not encode the complete signal peptide, an immediate priority was to derive this information from a genomic clone. In addition, the promoter and other regulatory regions of interest would most likely be present in genomic clones isolated with probe H.

Approximately 1.5 X 10^5 plaque forming units of the Charon 30 library were screened and two positive clones, designated λMQ1 and λMQ2, were identified and plaque purified three times (Fig. 17). DNA from these two phage specifically hybridized in Southern blot analyses to probe H, probe 2B (corresponding to amino acids 19 to 23 of the mature enzyme; Fig. 1), and to nick translated plasmid pAG18.

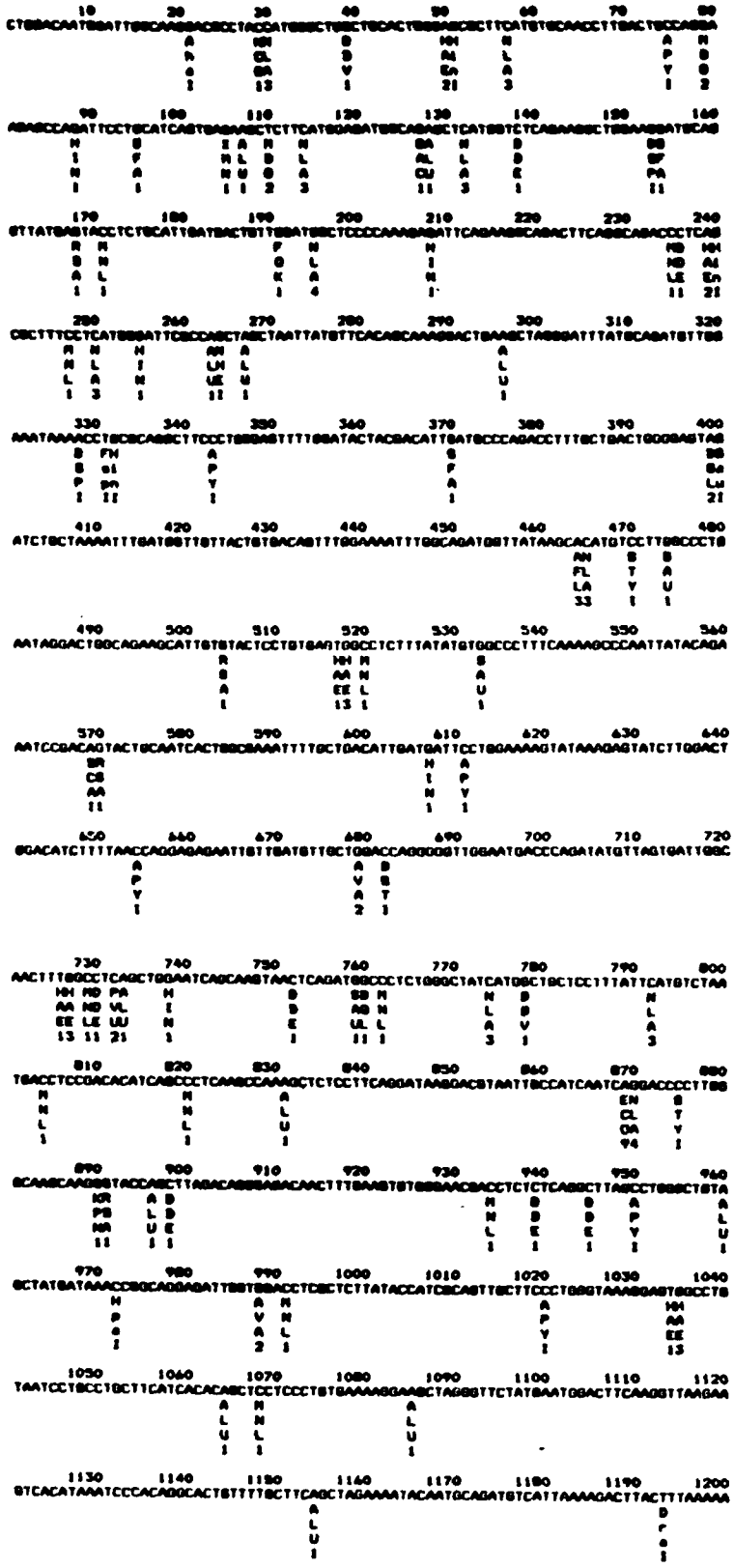


Fig. 16. Restriction enzyme sites of alpha-galactosidase A cDNA. Several restriction enzymes which recognize six bases are indicated beneath the nucleotide sequence. Nucleotide numbers are the same as used in Fig. 14.

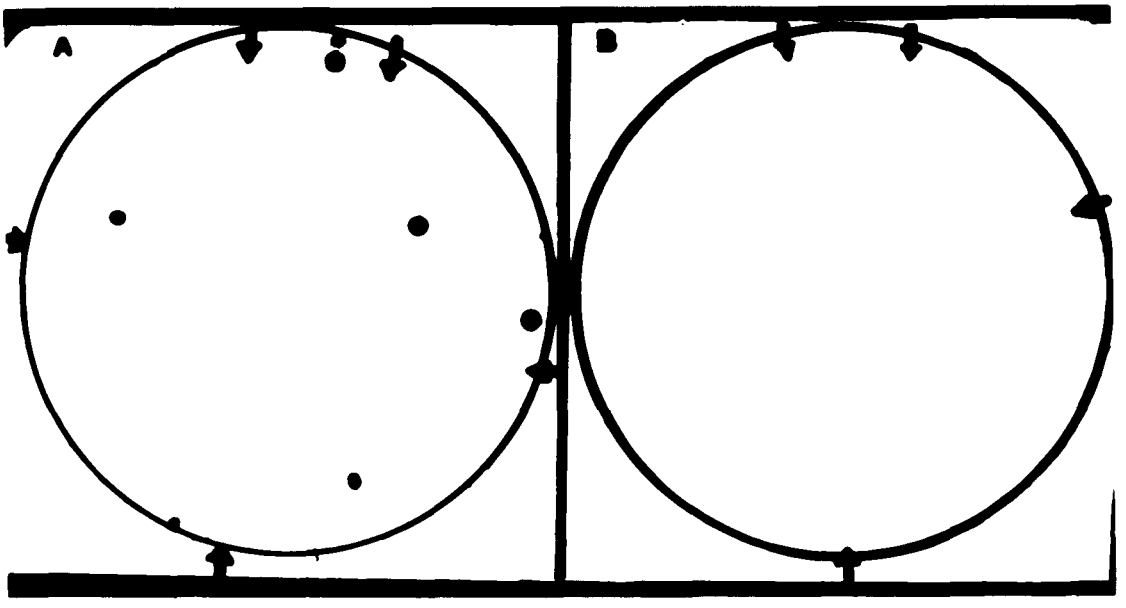


Fig. 17. Detection of alpha-galactosidase A genomic clones. λMQ1 was identified with end-labelled probe H, a 23-base long probe (5'-AGAGCACTGGACAATGGATTGGC-3') based on the last six nucleotides encoding the signal sequence and the first seventeen nucleotides encoding the mature enzyme (Fig. 14). Panel A depicts the hybridization of [γ-^{32}P]-labelled probe H to plaques of λMQ1 at the third cycle of plaque purificatin. Panel B shows a purified negative control processed in parallel that contained approximately 100 plaques. Each purification consisted of selection of a well-isolated positive plaque determined by probe H hybridization followed by replating at a density of less than 50 plaque forming units per plate. Arrows indicate filter orientation marks. Exposure times for filters A and B with one intensifying screen were 3.5 and 17 hr, respectively.

Southern Blot Analysis and Subcloning Strategy for clones λMQ1 and λMQ2. The steps used to generate alpha-galactosidase A genomic nucleotide sequence from M13mp11 subclones are indicated in Fig. 18. A single SacI fragment derived from genomic clones λMQ1 and λMQ2 specifically hybridized in Southern blots to probe H, probe 2B, and nick translated pAG18. Fig. 19 shows restriction enzyme digestion and Southern blot analysis of these clones. A SacI fragment of approximately 5.3 kb from λMQ1 (Fig. 19, lane 8) was subcloned to plasmid pSPRI generating plasmids pMQ1 and pMQ2 (Fig. 19, lanes 6 and 7, respectively). Digestion of λMQ1 DNA with TaqI revealed a genomic fragment of approximately 1.8 kb (Fig. 19, lane 3). Digestion of plasmids pMQ1 and pMQ2 with TaqI revealed the presence of a 1.2 kb fragment that bound oligonucleotide H (Fig. 19, lanes 1 and 2). This results from one internal TaqI site within the genomic alpha-galactosidase A 5.3 kb SacI fragment and a TaqI site present in the polylinker region of pSPRI (22 nucleotides from a SacI site in the genomic fragment). This 1.2 kb TaqI fragment from clone pMQ1 was then subcloned to the AccI site of M13mp11. M13 derivatives were analyzed for inserts by digestion of phage DNA with TaqI and electrophoresis on agarose gels (Fig. 20). Clones with the insert in both orientations were obtained and designated mMQU (packages the strand with the same sense as message RNA) and mMQL (packages the anti-message RNA sense strand).

Deletion Subcloning and Sequencing Strategy of a Genomic Fragment from Clone λMQ1. The method of Dale et al. (1985) was used to generate seventeen deletion subclones for nucleotide sequence analysis.

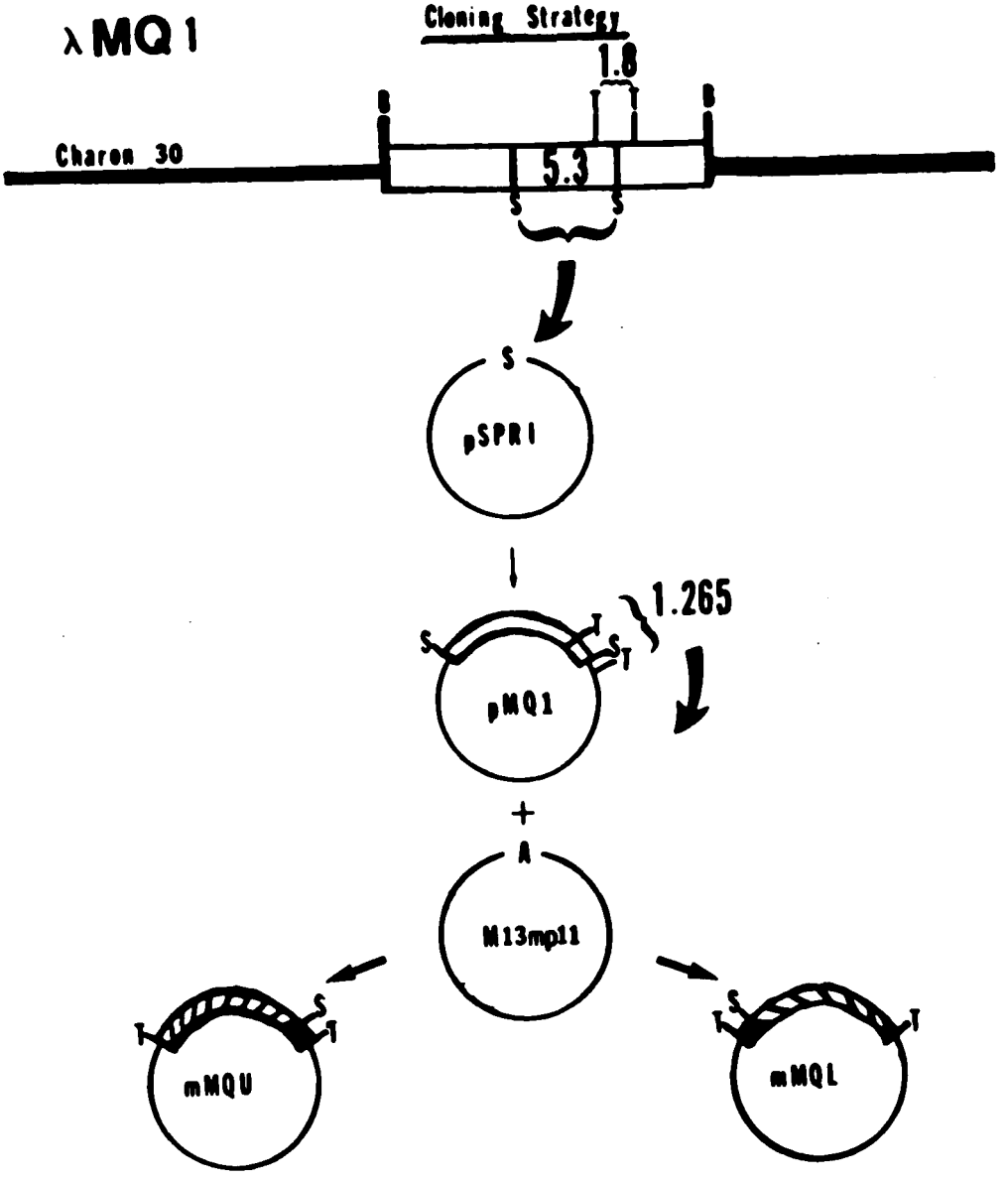


Fig. 18. Cloning strategy for sequencing the genomic fragment from phage λMQ1. A 5.3 kb SacI fragment from λMQ1 was subcloned to the SacI site of plasmid pSPRI, and named pMQ1. Clones were identified by SacI digestion of plasmid DNA and by Southern blot analysis (Fig. 19). A TaqI band of 1.265 kb from pMQ1 was obtained by electroelution of the DNA fragment to DEAE cellulose membrane, and was subcloned to the AccI site of M13mp11. Clones with the insert in both orientations were obtained and designated mMQU (packages the strand with the same sense as message) and mMQL (packages the anti-message strand). The solid box of mMQU and mMQL represents the 22 base pair polylinker region of pSPRI which was removed from clone mMQU (designated mMQZ; Table 1) for sequencing purposes (Figs. 22 and 23).

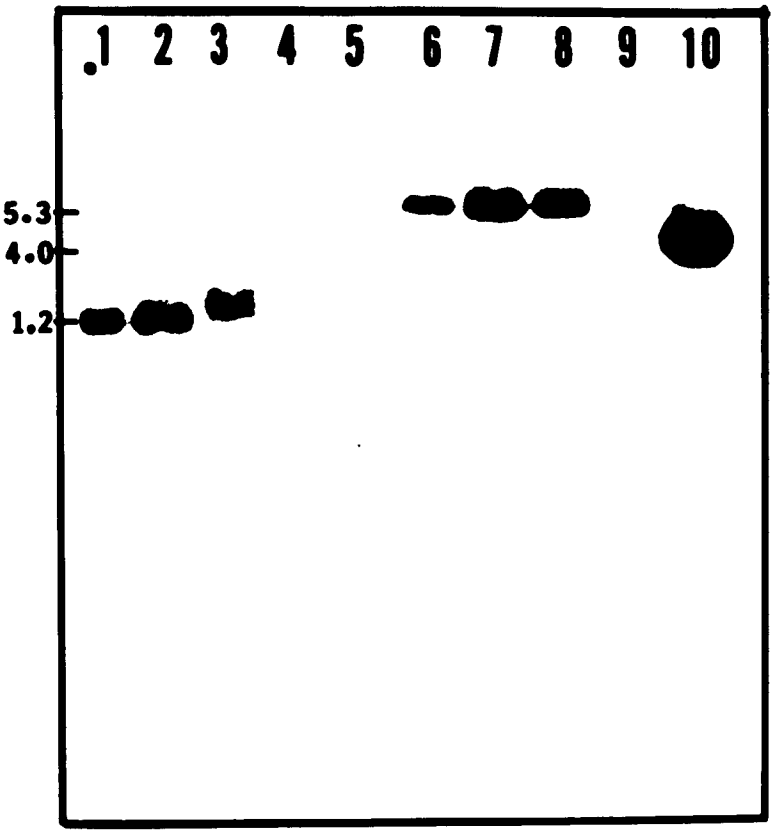


Fig. 19. Southern blot analysis of alpha-galactosidase A genomic clones. DNA was electrophoresed on a 0.7% agarose gel and transferred to Zetabind nylon membrane as described in Methods. Subclones were hybridized to probe H (Fig. 2) at 55°C. Lanes 1 and 2; SacI and TaqI double digest of plasmids pMQ1 and pMQ2, respectively. Lane 3; TaqI digest of pMQ1. Lane 4; TaqI digest of plasmid pSPRI. Lane 5; molecular weight markers of λ DNA digested with HindIII. Lanes 6 and 7; SacI digest of λMQ1 and λMQ2, respectively. Lane 8; SacI digest of pMQ1. Lane 9; SacI digest of λgt7 and lane 10; PstI digest of pUC9-18-5 (Table 1) containing alpha-galactosidase A cDNA insert as a hybridization positive control. Tic marks on the left indicate DNA sizes (in kb) of positive restriction fragments.

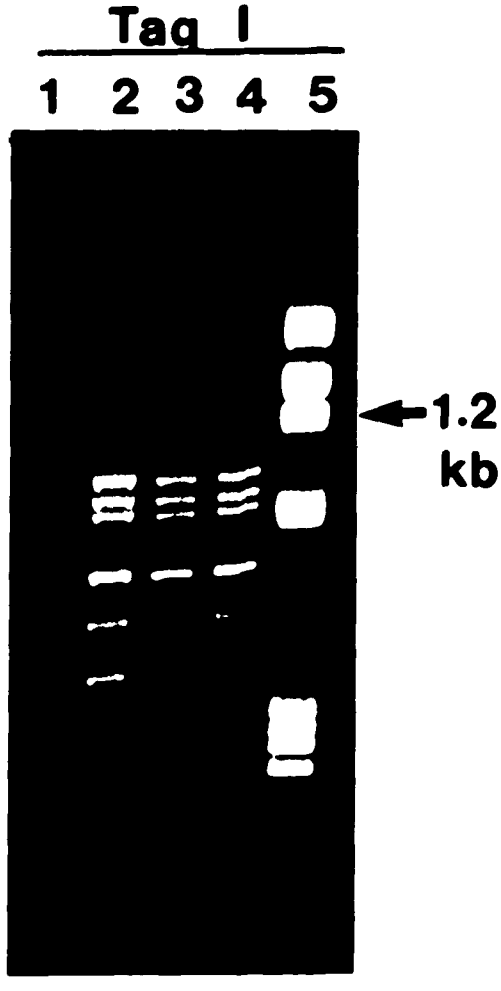


Fig. 20. TaqI restriction enzyme analysis of M13mp11 derivatives. M13mp11 subclones were analyzed for the presence of the 1.2 kb genomic fragment from plasmid pMQ1 by digestion of DNA with TaqI followed by electrophoresis on 0.7% agarose gels. Lanes 1-4; TaqI digestion of M13 subclones. Lane 5; TaqI digestion of plasmid pMQ1. The arrow indicates the 1.2 kb genomic fragment from plasmid pMQ1 which hybridizes to probe H (Fig. 18). The 1.2 kb TaqI insert can be seen in lane 1, and subsequent sequence analysis revealed that this clone (designated mMQL; Table 1), packages the strand with complementary sequence to the message.

Preliminary size estimates of deletion clones were determined by electrophoresis of single stranded phage DNA on agarose gels (Fig. 21). Five synthetic oligonucleotides (Figs. 1, 2) were designed to generate sequence in regions not covered by deletion subclones. The overall sequencing strategy is depicted in Fig. 22.

Interestingly, no sequence information was obtained from the undeleted mMQU clone using the M13 universal primer (17-mer). However, sequence data was obtained from mMQU using probe H as an internal primer. Perusal of the polylinker nucleotide sequence from pSPRI and M13mp11 identified an inverted repeat. Single stranded DNA from clone mMQU could potentially form a stem and loop structure with a calculated $\Delta G = -71.4$ (Fig. 23). To remove the polylinker regions from pSPRI and M13mp11, the replicative form of clone mMQU was digested with SacI, ligated with T4 DNA ligase, and transformed into strain JM103 (generating clone mMQZ; Fig. 24). It is postulated that the Klenow fragment of DNA polymerase I was able to move freely along the single stranded DNA of clone mMQZ without obstruction from a secondary structure due to the removal of the potential stem and loop region.

Nucleotide Sequence Analysis of Alpha-Galactosidase A Genomic Fragment. The seventeen deletion subclones previously described (Fig. 22) were sequenced by the Sanger method (1977) using the M13 (17-mer) and RD22 universal primers. The complete sequence of 1243 nucleotides (Fig. 25) was sequence in its entirety from both strands with the exception of the first 95 nucleotides of the message sense strand. The genomic sequence from position 1021 to 1138 was identical to the

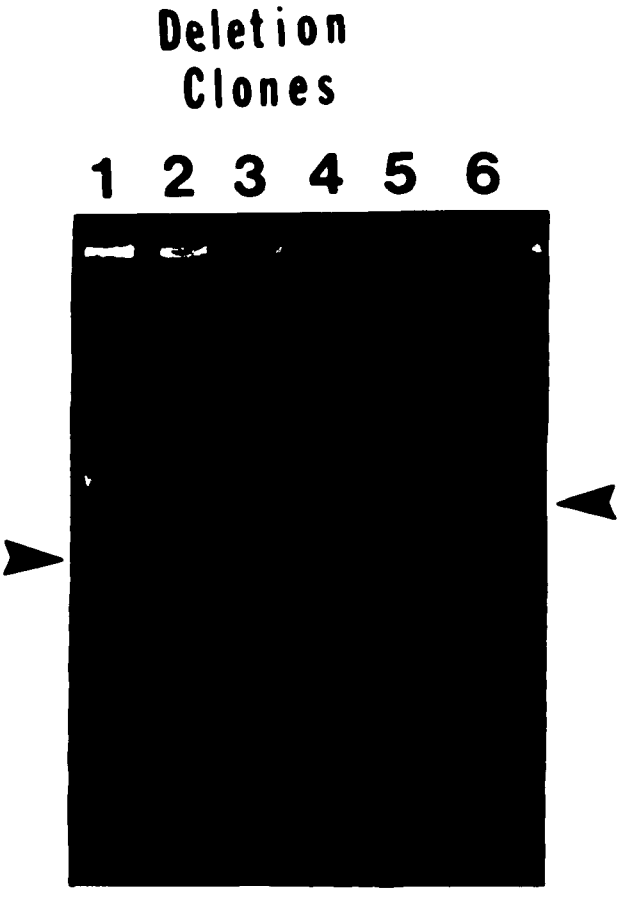


Fig. 21. Size analysis of single stranded DNA from deletion derivatives of clone mMQL. Each lane contains two bands, corresponding to the circular (upper) and linear (lower) forms of the single stranded viral DNA. Lane 2 contains M13mp11 with no insert and lane 6 contains mMQL with no deletion. Lanes 1, 3, 4, and 5 contain deletion clones that subsequently proved to have deletions of 132, 908, 886, and 278 nucleotides, respectively (determined by sequence analysis). The arrow on the right indicates the fastest migrating linear band of M13mp11 and the arrow on the left indicates the slowest migrating linear band of mMQL.

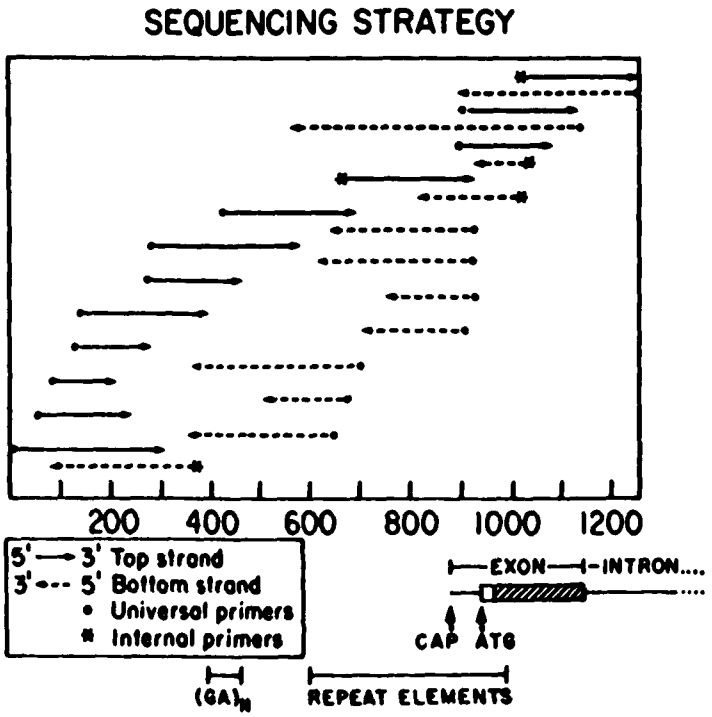


Fig. 22. Strategy for sequencing the alpha-galactosidase A genomic segment. The diagram shows the strategy for sequencing the 1265 nucleotide TaqI fragment from phage λMQ1. This fragment was subcloned into M13mp11 in both orientations for sequence analysis using the method of Sanger et al. (1977). Deletion clones with the end points indicated by the closed circles were generated using the method of Dale et al. (1985). In order to fill gaps in the sequence obtained from deletion clones, specific oligonucleotides were synthesized to prime the undeleted M13mp11 clones at sites indicated by astericks. The ruler indicates position by nucleotide (see Fig. 25). Some relevant features are indicated in the lower part of the figure, including the positions of the GA-rich sequence, the repeat elements flanking the promoter, the CAP site, the ATG initiation codon, the signal peptide (open box), the amino terminal amino acid sequence (hatched box), and the first exon-intron junction. The 22 base pairs extending from position 1243 to the TaqI site at position 1265, which are derived from the polylinker of the pSPRI vector, are not shown in Figs. 22 and 23.

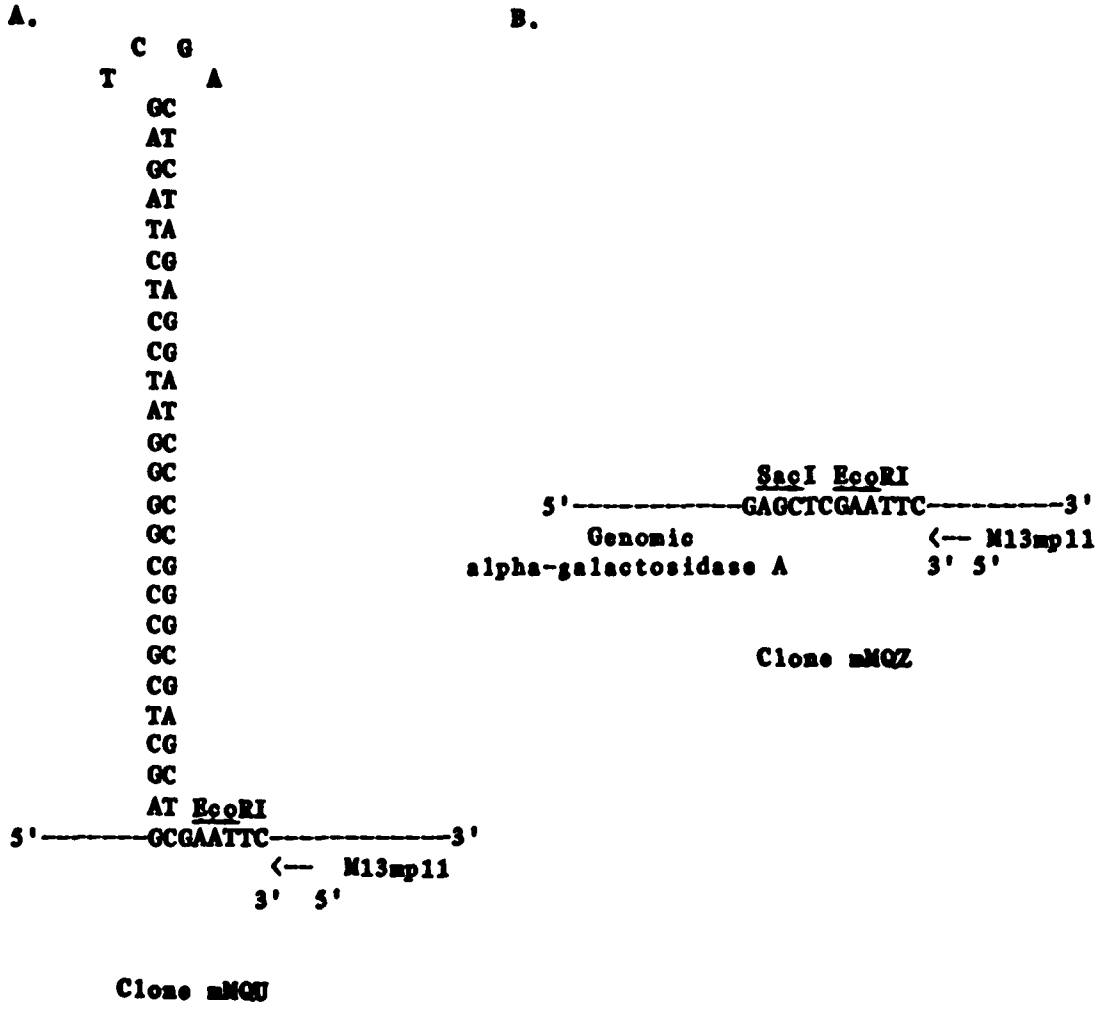


Fig. 23. Potential stem and loop structure of single stranded DNA from clone mMQU. The order of the inverted restriction enzyme sites from plasmid pSPRI and M13mp11 is: SacI-SmaI-BamHI-AccI-XbaI-AccI-BamHI-SmaI-SacI. The 25 base pair region from the polylinker regions of pSPRI and M13mp11 contain 17 G-C pairs with a calculated $\Delta G = -71.4$ (A). The arrow indicates the binding region of the M13 universal primer (17-mer) and the direction, 5'- to 3'-, of the dideoxy sequence reaction. SacI restriction enzyme digestion followed by ligation with T4 DNA ligase removed this inverted repeat region (B; clone mMQZ). Clones were screened for the presence of SacI and EcoRI restriction enzyme sites and the loss of SmaI and BamHI restriction enzyme sites (see Fig. 24).

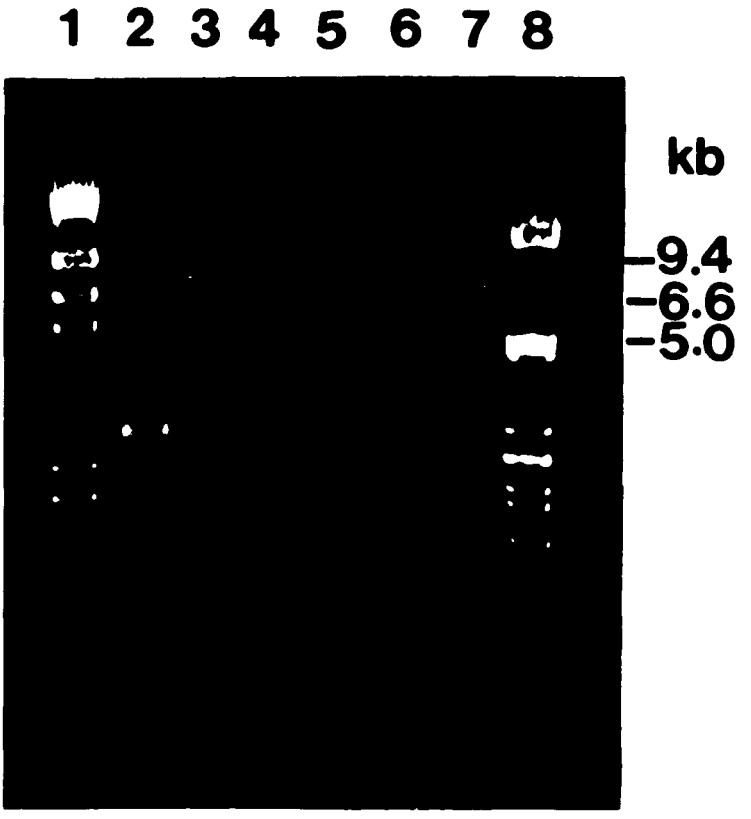


Fig. 24. Analysis of clone mMQU derivatives for the presence of the polylinker regions derived from pSPRI and M13mp11. DNA was electrophoresed on a 0.7% agarose gel at 100 V for 4 hr and stained with ethidium bromide. Lanes 1 and 8; molecular weight markers of λ DNA digested with HindIII and PstI, respectively. Lanes 2-4; mMQU DNA undigested, BamHI and SacI digested, respectively. Lanes 5-7; mMQZ DNA undigested, BamHI and SacI digested, respectively. Clone mMQZ has lost the inverted polylinker regions of pSPRI and M13mp11, thus, the BamHI restriction enzyme is lost and no digestion occurs. However, the SacI restriction enzyme site is retained and the clone is linearized after digestion with SacI.

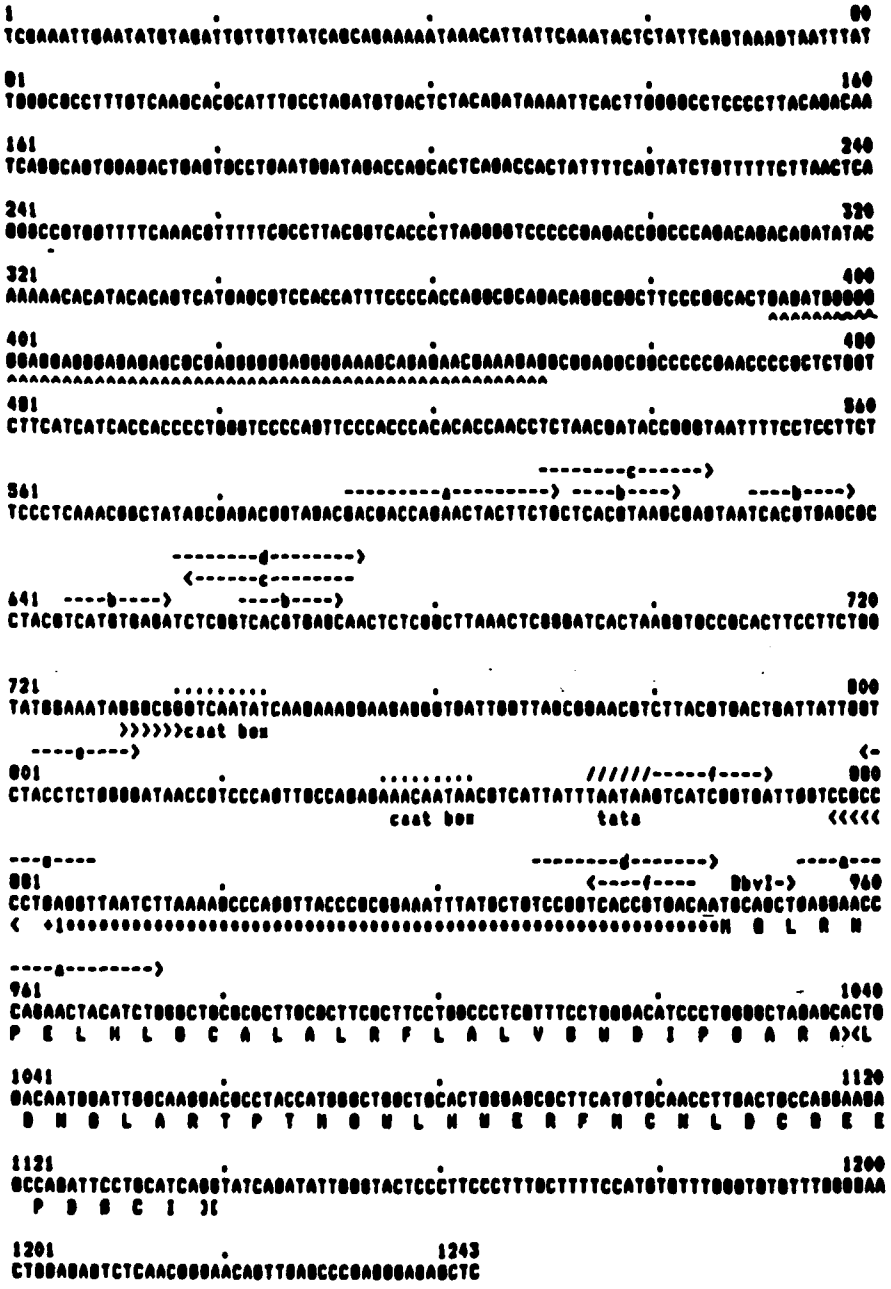


Fig. 25. Genomic alpha-galactosidase A nucleotide sequence. The 1243 nucleotide fragment from phage lambda-MQ1 includes the 5'-flanking sequences, complete signal peptide, first exon, and 104 bases of the first intron. The BbvI element flanking the ATG initiation codon is compared to other similar sequences in Table 6. Other sequences present are annotated as follows: GA-rich element ( ), CAAT boxes (...), TATA box (////), mRNA start site (+1***), signal peptide cleavage site ()<), exon-intron boundary (][), inverted and direct repeat elements (—> <—, respectively), and inverted transcription factor Sp1 recognition sites (>>> and <<<).

sequence that we previously obtained for the cDNA clone. The sequence extending from position 1139 to 1243 represents the first intron of the alpha-galactosidase A gene, since it has no similarity to the cDNA sequence, and the sequence, CAGGTATCA at positions 1136 to 1144 is a good match to the CAGGTAAGT consensus exon donor sequence (Mount and Steitz, 1983). This first exon-intron splice junction is compared to the consensus sequence in Table 3.

Signal Peptide of 31 Amino Acids for Alpha-Galactosidase A. The open reading frame at the amino-terminal end of the alpha-galactosidase A coding region extends to an initiation codon 31 amino acids from the first amino acid of the mature enzyme. The sequence CCACCATG has been identified (Kozak, 1984) as a consensus sequence for eukaryotic initiation sites (bases underlined are those most highly conserved), and the TGACAATG sequence at position 940 to 947 (Fig. 25) exhibits homology with that observed among the 211 sequences previously tabulated. The hydrophilicity matrix analysis of the signal peptide is shown in Fig. 26. The structure possesses the three components identified for signal sequences (Von Heijne, 1986), which are a charged amino acid within the first five residues (Arg at position 4), a hydrophobic central core (amino acids 6 to 26), and a more polar carboxy-terminal region (amino acids 27 to 31). Furthermore, the predicted signal peptide conforms to the "(-3,-1) rule" of Von Heijne (1986) which specifies that the residue at -1 (Ala) must be small (Ala, Ser, Gly, Cys, or Thr) and the residue at -3 (Ala) must not be aromatic (Phe, His, Tyr, Trp), charged (Asp, Glu, Lys, Arg), or large and polar (Asn, Gln). Also, as seen with other signal sequence cleavage sites

Table 3. First exon-intron splice junction of alpha-galactosidase A.

NUCLEOTIDE FREQUENCY	83	64	73	100	100	91	68	84	63
DONOR CONSENSUS	C/A	A	G	/ G	T	A/G	A	G	T
ALPHA-GALACTOSIDASE A DONOR	C	A	G	/ G	T	A	T	C	A

Donor consensus splice site sequence was tabulated from 139 donor sequences. Nucleotide frequency indicates percentage of nucleotide occurance by position (Mount and Steitz, 1983). The first exon donor sequence of human alpha-galactosidase A is shown beneath the consensus sequence. Note that the first six nucleotides are invariant from the consensus model.

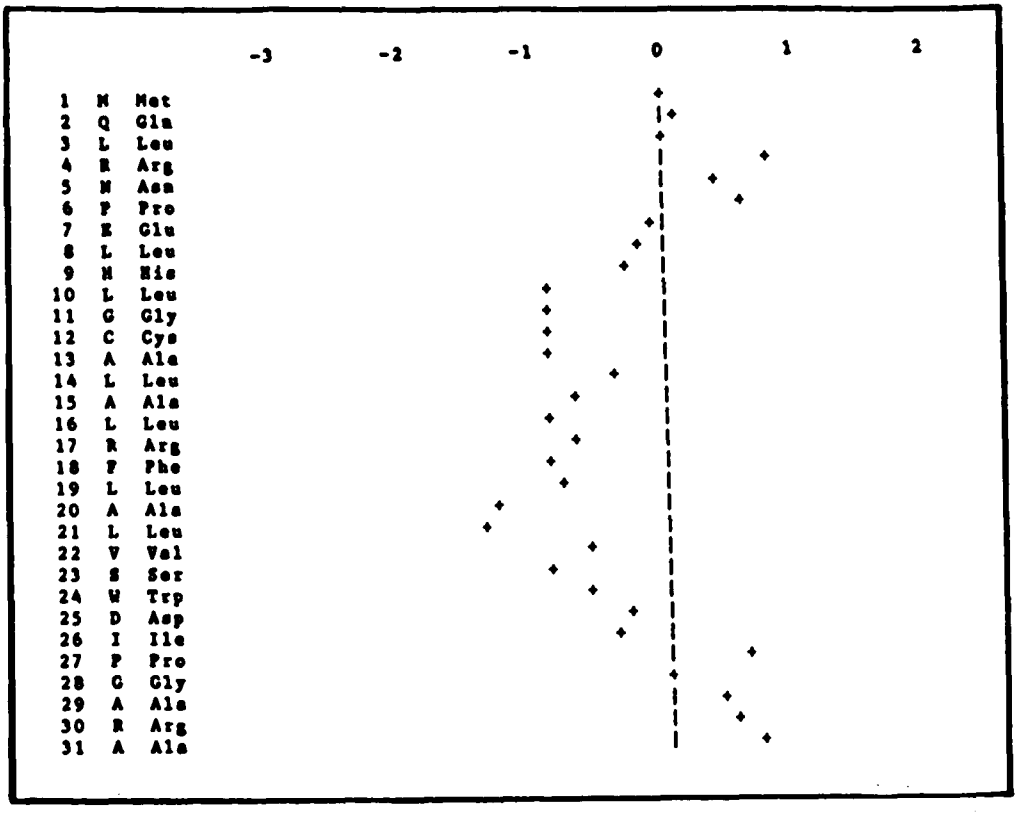


Fig. 26. Hydrophilicity profile of the 31 amino acid signal peptide of alpha-galactosidase A. Note the presence of Arg, a charged amino acid within the first four residues, the hydrophobic core spanning from Pro at residue 6 to Ile at residue 26, and the alpha helix breaking residue, Pro at position 27 of the peptide. Computer analysis of the protein structure predicts an alpha helix of 13 residues within the hydrophobic core from amino acid 8 to amino acid 20.

(Von Heijne, 1986), Pro is absent from positions -3 to +1 (Ala-Arg-Ala-Leu) but is present at residue -5 near the boundary of the hydrophobic core.

The Promoter and 5'- Flanking Sequences of Alpha-Galactosidase A. A consensus TATA box is present (TAATAA) at position -31 relative to a putative +1 transcription initiation site (the A residue at position 885 in Fig. 25). This A residue at position 885 is a candidate for the first base in the transcript since (i) most eukaryotic transcripts begin with A or G, with A being more common (Mount and Steitz, 1983), (ii) this is the only A residue located in the region 26 to 31 nucleotides from the TATA box, which is the distance most commonly observed between the TATA box and the initiation site, (iii) this start site should generate spliced mRNA of 1349 nucleotides which, assuming approximately 100 nucleotides in a poly(A)$^+$ tract, is consistent with the size of 1.45 kb that we estimated for mRNA in Northern blots hybridized to the nick translated cDNA clone (Fig. 27), and (iv) it is consistent with the size of the cDNA generated when a 131 base pair HgiAI restriction fragment derived from the 5' -end of our cDNA clone is used to prime reverse transcriptase using HeLa cell poly(A)0+ mRNA as a template (Petros Hantzopoulos, Ph.D. thesis).

Sequences homologous to the cannonical CAAT box (GGQCAATCT; Q is C or T) are seen at position 736 (GGTCAATAT), which corresponds to position -148 relative to the proposed transcription initiation start site and at position 835 (AAACAATAA), which is at position -50 relative to the initiation start site. The CAAT box is commonly present in eukaryotic promoters recognized by RNA polymerase II, and it is most

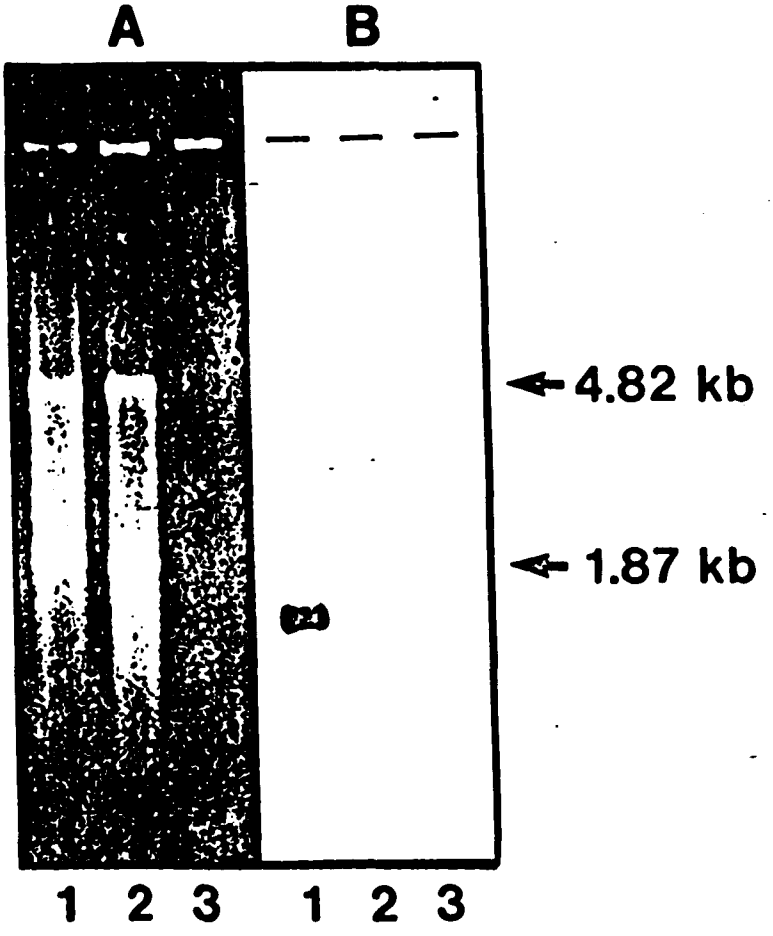


Fig. 27. Northern blot analysis of alpha-galactosidase mRNA. HeLa cell poly(A)+ RNA before (lane 1) or after (lane 2) polysome immunoabsorption with anti-alpha-galactosidase A antibodies. Panel A shows the picture of the agarose gel stained with ethidium bromide before transfer and panel B shows an autoradiograph of the Northern blot after hybridization to nick translated pAG18 cDNA insert. Based on the migration of the 18S and 28S human ribosomal subunits (lane 3), the alpha-galactosidase A mRNA was estimated to be 1.45 kb. (Figure taken from Bishop et al., 1986).

often found at position -70 to -80 relative to the transcription initiation site (Benoist et al., 1980; Corden et al., 1980; Efstratiadis et al., 1980).

Computer Analyses of Alpha-Galactosidase A Genomic Fragment. The European Molecular Biology Library and GenBank Library were searched for homologies to the alpha-galactosidase A genomic fragment shown in Fig. 25. A GA-rich segment of approximately 60 nucleotides that extends from position 391 to 450 (Fig. 25) is highly homologous to a sequence found in eleven other genes (Table 4) from various organisms, including human, chimpanzee, rat, mouse, frog, insect, and sea urchin. No other significant nucleic acid homologies to sequences in the data bases were detected. A search of the National Biomedical Resource Foundation amino acid sequence libraries revealed no significant homologies to the 31 amino acids of the signal peptide for human alpha-galactosidase A.

Several direct and inverted repeats are located between positions 590 and 980 (Fig. 25; Table 5). There are four tandem repeats of a 10 nucleotide element located upstream of the CAAT box (position 613 to 671), and it may be significant that these are regularly spaced, being separated by 6, 7, and 6 nucleotides, respectively. The significance of these and other repeated sequences, if any, is unknown, and the presence of lysosomal enzyme specific sequences cannot be determined since no other genomic clones for this class of enzymes have been characterized.

Table 4. The homologous GA-rich element upstream of the alpha-galactosidase A promoter.

	HOMOLOGOUS SEQUENCE	POSITION
1.	GAGATGGGGGGGAGGAGGGAGAGAGCGCGAGGGGGGAGGGGAAAGCAGAGAACGAAAGAGGCGGAG	391-456
2.	....TAG..A.....AA.A..AA...A....AGG..*..*...A...:..A.A....	2966-2911
3.	.....AAA.....	776-754
4.	....G.....A.A.*.....A.A..G..*.....GA..G....	1771-1721
5.	...:..C.....C...:..G....*.....	292-252
6.	....CA.....CT....CT.....	405-431
7.	...GC...A.A...:..A.....A....A.....	55-15
8.	.C.....G.....:....C...T.....TCG.G..*.....	570-523
9.	...GG...C.....A.....C.C.....	2531-2496
10.	....*.A.A.A....G.....A.A.....A.*.G..*.....	267-220
11.	....*.A.A.A..:...A.....A.A.....G.....	21-62
12.	.A.T..:.....CT....A*...T...:...A.GG..AG..A...AG..A..	43-96

For each gene, the sequence and nucleotide position of the homologous region is indicated. An asterick (*) indicates a gap and the colon (:) indicates a match with one base loopout. The genes are: 1) alpha-galactosidase A (this work), 2) chimpanzee Alu type DNA (Maeda et al., 1983), 3) mouse Ig kappa (Van Ness et al., 1981), 4) mouse glandular kallikrein (Mason et al., 1983), 5) mouse alpha-2 type I collagen (Schmidt et al., 1984), 6) rat cardiac alpha-myosin heavy chain (Mahdavi et al., 1984), 7) rat 18S rRNA (Cassidy et al., 1982), 8) rat 28S rRNA (Chan et al., 1983), 9) frog 28 S rRNA (Ware et al., 1983), 10) Drosophilia virilis simple repeat sequence (Tautz and Renz, 1984), 11) sea urchin histone H4 (Hentschel, 1982), and and 12) human histone H4 (Sierra et al., 1983). The numbering system is used to indicate the positions of these repeats is that used by the GenBank data base. The sequence position is indicated in ascending order (6, 11, and 12) or descending (2-5, 7-10) manner to indicate its location in the top or bottom strands, respectively.

Table 5. Direct and inverted repeat sequences in the promoter region.

Element	Orientation	Sequence	Match
A	Direct	GACGA**CCAGAACTACTTCTG GAGGAACCCAGAACTACATCTG	18/22
B	Direct	TCACGTAAGC TCACGTGAGC	9/10
		TCACGTGAGC TCACGTGAGC	10/10
		TCACGTGAGC TCATGTGAGA	8/10
C	Inverted	TGCTCACGTAAGCGAG ACGAGTGCACTGGCTC	14/16
D	Direct	TCTCGGTCAC*GTGAGCAA TC*CGGTCACCGTGA*CAA	16/19
E	Inverted	ACCTCTGGGG TGGAGTCCCC	9/10
F	Inverted	GTCATCGGTGA CAGT*GCCACT	10/11

The repeat elements labelled A-F are aligned to permit maximum homology. The asterick (*) indicates a mismatch or gap. The locations of the repeat elements are indicated in Figure 25.

The BbvI Consensus Sequence of Alpha-Galactosidase A. The nucleotide and amino acid sequences corresponding to the signal peptide of alpha-galactosidase A was compared to those known for the other human lysosomal enzymes: cathepsin B (Chan et al., 1986), cathepsin D (Faust et al., 1985), beta-glucocerebrosidase (Tsuji et al., 1986), beta-hexosaminidase alpha-subunit (Myerowitz et al., 1985) and beta-glucuronidase (Oshima et al., 1987). Although no extensive regions of homology were apparent at the amino acid or nucleic acid level, several patterns of homology, including a consensus sequence GCAGCQ (Q is C or T), that overlaps the ATG initiation codon of the signal peptide or follows within 4 nucleotides for five of these enzymes (Table 6) were noted. This consensus sequence (which fortuitously includes a GCAGC recognition sequence for the restriction enzyme BbvI) overlaps the ATG initiation codon for both alpha-galactosidase A and cathepsin D, and is separated from the ATG by two or four nucleotides for cathepsin B and beta-glucocerebrosidase, respectively. The homology also extends to beta-hexosaminidase alpha-subunit but this requires a loop out of two nucleotides (Table 6).

Watson (1984) prepared a compilation of the amino acid sequences of signal peptides, including 42 human genes with complete signal peptides. The corresponding nucleotides sequences for these genes were obtained and none contained the BbvI consensus sequence flanking the ATG codon. In order to expand the search, all of the human signal peptides present in the GenBank data base (December, 1986) were examined using the Quest routine of Bionet. Among the 133

Table 6. Homologies among the nucleotides encoding the signal peptides of lysosomal enzymes.

GENE	NUCLEOTIDE SEQUENCE
1. alpha-galactosidase A	atgCAGCTGAGG]]]]]]]]
2. cathepsin D	atgCAGCCCTCC]]]]]]]]
3. cathepsin B	atgTGGCAGCTCTGG \/\/\/\/\/\/
4. beta-glucocerebrosidase	atgGCTGGCAGCCTCAC \/\/\/\/\/\/
5. alpha subunit beta-hexosaminidase	atgACAAGCTAG]]]]]]]]

The atg initiation codon is indicated in lower case, and the symbols (-,], and \/) indicate consensus sequences shared by the various members of this group. The BbvI consensus sequence (GCAGCQ; Q is C or T) is common to all five genes.

non-lysosomal human signal peptides, none contained the BbvI consensus sequence within four nucleotides of the ATG codon. Six of these human sequences contain the BbvI consensus sequence, but these are located 13 to 37 nucleotides distal to the ATG codon.

The BbvI sequence homology noted among five of the six human lysosomal sequences may not extend to other mammals or may exist for only a subset of lysosomal enzymes, since the rat preputial beta-glucuronidase signal peptide lacks this sequence (Nishimura et al., 1986). Among the 132 non-human mammalian signal peptides examided, only one contains the BbvI repeat immediately flanking (separated by four nucleotides) the ATG codon. This signal peptide is encoded in one of the seven rearranged V_H genes from a BALB/c hybridoma elicited by immunization with proteins coupled with the hapten, (4-hydroxy-3-nitrophenyl) acetyl (Loh et al., 1983).

Location of the Propeptide of Alpha-Galactosidase A. We observed that the deglycosylated mature alpha-galactosidase A protein migrates with an apparent molecular weight of approximately 41,800 (Fig. 3). Based on this subunit molecular weight and the amino acid composition analysis, it was estimated that there are approximately 370 amino acid residues in the polypeptide chain. A predicted molecular weight of 45,346 was based on the nucleotide sequence of the cDNA clone λAG18 (Fig. 14), not including the signal peptide. This cDNA clone included the complete 3'-coding region, but only part of the 5'-coding region upstream from the first amino acid of the mature form of the enzyme. The results presented here complete the sequence at the 5'-end, and indicate the presence of the signal (prepeptide) but not a propeptide at the amino terminus. The most carboxy-terminal tryptic

peptide for which amino acid sequence is available spans amino acids 363 to 370 (Fig. 14), while the cDNA sequence indicates an open reading frame extending to amino acid 398. This most carboxy-terminal region presumably constitutes the propeptide that is cleaved from the proenzyme form. If proenzyme processing took place at basic amino acids, as occurs for the human cathepsin B propeptide (Chan et al., 1986), then potential protease cleavage sites at Arg residues at positions 371 and 373 would release polypeptides with 28 amino acids (3,223 daltons) or 26 amino acids (2,954 daltons), respectively. This agrees well with the 3,546 daltons predicted for the propeptide based on the difference in the observed molecular weight of the mature enzyme (41,800), and the coding capacity predicted by the cDNA clone (45,346 daltons) after removal of the signal peptide. In addition, Table 2 shows good agreement in comparison of the predicted amino acid composition of the predicted mature enzyme (if cleaved at amino acid 371) to the amino acid composition analysis performed on purified human alpha-galactosidae A and based on a subunit molecular weight of 41,800.

DISCUSSION

Our initial efforts to isolate a cDNA clone for human alpha-galactosidase A were based on the descriptions of two techniques for isolating specific DNA sequences and on our limited knowledge of the protein sequence. In 1979 Noyes et al. developed a technique and obtained the partial nucleotide sequence of hog gastrin mRNA using an oligonucleotide to prime poly(A)+ mRNA. Following the work of Noyes et al. (1979), several groups used this general approach to isolate several genes (e.g. rabbit beta-globin; Wallace et al., 1981, and rat relaxin; Hudson et al., 1981). The use of synthetic oligonucleotide mixtures that represent all possible codon combinations of a portion of amino acid sequence as hybridization probes to screen for cloned DNA sequences, was pioneered by Suggs et al. (1981). Other groups (e.g. Woods et al., 1982; Breslow et al., 1982) used this methodology to successfully isolate genes for the major histocompatibility complex factor B and apolipoprotein A-I, respectively. When Young and Davis (1983) described a λgt11 expression system that permitted immunologic detection of cloned cDNA sequences, we pursued this screening technique as a complementary approach. To date, these methodologies have proved to be powerful techniques in the isolation of specific genes or proteins.

The aforementioned techniques were used in parallel since all have inherent advantages and disadvantages. First, all screening procedures can detect false positive clones due to cross-reactive antigenic sites or identical nucleotide sequences. Second, the clone may not be

detected, for example, if the library does not contain a full length cDNA sequence and only amino teminal amino acid sequence is known. A clone will be missed with immunologic screening technique if the clone of interest does not express the protein. This will occur if the cDNA is out of the proper frame for translation or if the insert is cloned in the opposite orientation. The antibody detection method can only be used to screen cDNA libraries, whereas, the oligonucleotide technique can be used to screen both genomic and cDNA libraries. However, a problem can arise with the oligonucleotide approach if the amino acid data is uncertain, which is not unusual for proteins that are difficult to purify in large quantities. For this reason, we made more than one oligonucleotide mixture spanning separate amino acid sequences. both the oligonucleotide and immunologic screening procedures employ high density screening which enables one to screen a large number of recombinant clones in one screening. The use of oligonucleotides as hybridization probes has the advantage of obtaining greater specificity than the mRNA priming technique for the production of cDNA probes. Under the appropriate hybridization conditions, formation of a base pair mismatch is not detected, whereas base mismatches are tolerated at the 5'- end of the primer by reverse transcriptase. An advantage of the antibody detection method is that a relatively small number of specific antibodies are produced to the protein. However, with oligonucleotides, only one member of a large mixture is specific for the clone due to amino acid codon redundancy.

Besides the inherent disadvantages to the screening methodologies, the isolation of human lysosomal hydrolase genes has been hampered because they constitute a very small percent of the mRNA pool. Individual lysosomal enzymes make up less than 0.1% of the total

protein in most human cells (Myerowitz and Neufeld, 1981). For example, within human lung, alpha-galactosidase A constitutes approximately 0.002% of total cellular protein. In addition, with the exception of the steroid induced rat beta-glucuronidase, there is no known induction method which can significantly increase the synthesis of lysosomal enzymes. In spite of these hurdles, several investigators have been successful in isolating cDNA clones encoding human lysosomal hydrolases. The cDNA obtained for alpha-fucosidase was considered to be one of the rarest clones isolated to date; alpha-fucosidase mRNA is estimated to be 0.002% and its protein concentration is estimated at 0.01% of liver protein (Grantham et al., 1981). Although a major challenge, the isolation of these rare mRNAs will obviously aid elucidation of structure, function, trafficking, and expression of the gene products. Moreover, significant insight into the molecular lesions of lysosomal storage diseases will be advanced.

Our efforts to screen cDNA and genomic libraries with mixed oligonucleotide probes identified several positive hybridizing clones, however, none proved to be a bona fide clone specific for alpha-galactosidase A. When the oligonucleotide probe was limited to a mixture of eight species (i.e. probe 2A/2B; Fig. 1), over 30 positive clones which specifically hybridized to this probe were isolated from cDNA and genomic libraries. Of these 30 clones, only one demonstrated positive hybridization to a second oligonucleotide (pcD1). Nucleotide sequence analysis of a 411 base pair fragment from clone pcD1, which bound both oligonucleotides, revealed that the surrounding nucleotides did not match the known amino acid codons for the alpha-galactosidase A gene product. Additional sequence analysis of another clone, pcD21, isolated under increased stringency conditions, revealed that this

clone contained complementary sequence with a perfect match to oligonucleotide 2B, but the surrounding nucleotides did not match the known alpha-galactosidase A sequence. This result demonstrates an obvious limitation to the use of oligonucleotide mixtures.

When oligonucleotide mixtures were utilized to prime mRNA in the presence of reverse transcriptase, Nucleotide sequence analyses of the major cDNA band products identified more false positives. That is to say, false in that they did not correspond to the known amino acid codons for alpha-glactosidase A gene. However, upon long exposures of the X-ray film, several weak cDNA band products were evident, and possibly one of these minor cDNA products was copied from the alpha-galactosidase A mRNA. However, after the isolation of a cDNA clone specific for alpha-galactosidase A, we obtained a unique product using a unique 131 nucleotide restriction fragment from this clone to prime mRNA.

We isolated a cDNA clone specific for human alpha-galactosidase A by the immunologic detection method. Our success was presumably due to the availability of highly specific antiserum and a large library. A cDNA clone was isolated from a human liver cDNA library constructed in a λgt11 expression system when approximately 1.4 X 107 plaques were screened with monospecific polyclonal antibodies. Four antibody-positive clones were identified initially; of the four, only one (λAG18) demonstrated both antibody specificity in competition assays with purified enzyme and specific hybridization to synthetic oligonucleotide mixtures. In this situation, the oligonucleotide mixtures were used to confirm the identity of the clone as opposed to primary selection of the clone. The authenticity of clone λAG18 was verified by nucleotide sequence analysis of the 5'- end of the cDNA

insert and exact correspondence between the predicted and known amino-terminal amino acid sequence was shown.

Nucleotide sequence analysis of the complete alpha-galactosidase A cDNA revealed an open reading frame of 1226 nucleotides. The amino-terminal codon for the mature enzyme was located by agreement between the predicted amino acid sequence from the nucleotide sequence and the known amino-terminal amino acid sequence. Translation of the open reading frame showed agreement with 86 of 100 non-overlapping amino acid residues. This is an excellent correlation considering that there was substantial uncertainty concerning the accuracy of the amino acid sequence data due to the limited quantity of available enzyme. From the amino-terminal codon of the mature enzyme, the alpha-galactosidase A insert encoded information for 398 amino acids. Also, the last five amino acids of a precursor form were encoded in this clone (Fig. 14). At that time, it was not clear whether these five amino acids were the carboxy terminus of the signal peptide (prepeptide) or the carboxy terminus of the pro piece.

Two consensus polyadenylation cleavage sites, ATTAAA and AATACA, were identified 12 and 28 nucleotides, respectively, upstream from the stop codon in the alpha-galactosidase A cDNA clone. Approximately 12% of vertebrate messages contain the sequence ATTAAA, and AATACA only occurs in 2% of the messages (Wickens and Stephenson, 1984). The pentanucleotide sequence, CACUG, associated with the U4 small ribonuclear binding protein, was located at the 3'- end of the alpha-galactosidase A cDNA. Manipulations (i.e. deletions and nucleotide mutations) of the cDNA clone will enable studies to ascertain the essential features necessary for polyadenylation.

The predicted amino acid sequence identified four possible

N-glycosylation sites (Fig. 14). This is consistent with the observation that the purified mature enzyme from human tissues and plasma contains one or more asparagine-linked oligosaccharide chains (Desnick and Sweeley, 1983; LeDonne et al., 1983). The four possible N-glycosylation sites were all located in beta-turns within hydrophilic regions of the peptide (Fig. 15) as would be predicted for surface localization and exposure.

Previous to this clone, the absence of a 3'- untranslated region between the UAA stop codon and the poly(A)$^+$ tail has not been reported for nuclear coded transcripts. However, the absence of a 3'- untranslated region has been described for mitochondrial genes (Anderson et al., 1981). A search of the data base identified 3'- untranslated regions as short as 10 and 15 nucleotides for human factor X (Leytus et al., 1984) and for the beta-subunit of human chorionic gonadotropin (Fiddes and Goodman, 1980), respectively. The lines of evidence which support this unique feature are: i) the 3'- region of the clone was sequenced independently four times from both strands, ii) the most carboxy-terminal tryptic peptide, T-53A (Fig. 14), confirms the reading frame at the 3'- end, iii) the consensus polyadenylation signals and U4 small ribonuclear binding site were identified as described above, and iv) an alpha-galactosidase A cDNA clone isolated from an independent library contained the identical 3'- sequence as the original clone (Bishop et al., 1986). After the identification of our clone, Jenh et al. (1986) identified a rat thymidylate synthase messenger which similarly lacks a 3'- untranslated region.

Computer aided homology searches of amino acid and nucleotide data bases did not reveal any significant homology to other characterized genes. A computer search of the data available for other lysosomal

hydrolases revealed that only alpha-fucosidase (nucleotides 31 to 178) showed limited homology (52%) to nucleotides 34 to 168 of the alpha-galactosidase A cDNA. The lack of homology at the amino acid level between the sequenced lysosomal enzymes and the knowledge that the denatured enzyme does not contain the necessary recognition signal for phosphorylation and hence, transport to the lysosome (Lang et al., 1984), supports the hypothesis that secondary or tertiary amino acid structures define lysosomal determinants necessary in the M-6-P recognition pathway.

After the nucleotide sequence for the cDNA clone was determined, a unique oligonucleotide (23-mer) was synthesized to screen genomic libraries to obtain the complete nucleotide sequence encoding the preproenzyme. This region was of interest to us since the cDNA clone was truncated at the 5'- end, and we wished to identify the sequence of the promoter region and any associated regulatory signals. An obvious advantage of a unique oligonucleotide probe compared to a nick translated cDNA to screen genomic libraries is that the identification process is specifically narrowed to clones containing that particular region. Depending on the size and number of introns, a gene may span several kilobases of DNA, and a nick translated cDNA will select all clones containing exons. Also, the signal to noise ratio was more favorable with a long (e.g. greater than 17 nucleotides) unique oligonucleotide compared to that of nick translated probes.

Two positive clones, λMQ1 and λMQ2, with inserts of approximately 15 and 17 kb, respectively, were isolated from the initial screening of a 4X Charon 30 genomic library. Clone identity was confirmed when Southern blot analysis of SacI digested phage DNA detected a 5.3 kb fragment from both clones that bound specifically to oligonucleotides H

and 2B and to nick translated alpha-galactosidase A cDNA insert.

The nucleotide sequence information for the region that extends approximately 880 nucleotides upstream of the proposed site of transcription was obtained, and this region would be expected to include the entire promoter. The alpha-galactosidase A gene is recognized and transcribed by RNA polymerase II (reviewed by Roeder, 1976). Many RNA polymerase II type genes have been sequenced and the 5'- ends of the mRNAs they encode have been located. There is a clear hierarchy of mRNA CAP site preferences with A>G>>U>C (Baker and Ziff, 1981). Two short regions of control have been defined. The first, recognized by Goldberg and Hogness, is referred to as the TATA box and is generally found at position -26 to -31 (+1 is the transcriptional start site). A TATA box is located at position -31 relative to the proposed transcription initiation site for alpha-galactosidae A (Fig. 25). The second region of control is located at the -70 to -80 region and is known as the CAAT box (Benoist et al., 1980; Corden et al., 1980; Efstratiadis et al., 1981). Two CAAT boxes are located in the promoter region of the alpha-galactosidase A gene at positions -148 and -50 relative to the putative start site of transcription.

The ratio of one lysosomal enzyme to another remains relatively constant, although the absolute levels of the enzymes may vary widely, as seen during organ development and during the changes in the number of lysosomes in a cell (Paigen, 1979). This indicates the presence of mechanisms to coordinately regulate the synthesis of this class of enzymes. The presence of TATA and CAAT boxes in the alpha-galactosidase A gene promoter implies the presence of specific controls at the level of transcription, since these features are absent

from constitutively expressed mammalian housekeeping genes, such as human hypoxanthine phosphoribosyltransferase (Kim et al., 1986).

The alpha-galactosidase A genomic clone contains the sequence encoding the complete 31 amino acid signal peptide that was absent in our cDNA isolates. It has been postulated that the amino-terminal signal sequences of lysosomal proteins are recognized by an 11S ribonucleoprotein (signal recognition particle; Walter et al., 1984) in a manner similar to some secretory and integral membrane proteins. All three protein classes are thought to be processed by a similar but not necessarily identical pathway from their site of synthesis at the rough endoplasmic reticulum through the Golgi apparatus. At different stages in this process the proteins in transit may be covalently modified by glycosylation, oligosaccharide modifications, or partial proteolysis. Evidence has been presented which indicates that routing from the Golgi apparatus to lysosomes requires a M-6-P residue on the asparagine-linked oligosaccharide side chain (Kornfeld, 1986). As discussed earlier, Lang et al. (1984) have shown that for cathepsin D, the determinant recognized by the phosphorylating enzyme is present on the native, but not the denatured polypeptide. Newly synthesized lysosomal enzymes bind with high affinity to the M-6-P receptor and are translocated to the lysosomes. However, there are two different receptor proteins and at least one or more M-6-P independent pathways to the lysosome (Kornfeld, 1986). We compared the alpha-galactosidase A signal peptide to other lysosomal proteins and other proteins synthesized on the rough endoplasmic reticulum to test for similarities. In this context, the BbvI consensus sequence present at the nucleotide level in five of the six human lysosomal enzymes (Table 6) is unexpected, and does not suggest a simple hypothesis with regard

to potential biological significance at the level of transcription, translation, or enzyme sorting. Furthermore, the presence of the EbvI consensus sequence does not correlate with the dependence upon the M-6-P receptor pathway, since beta-glucocerebrosidase does not use this pathway (Tager, 1985).

The significance of the GA-rich element located upstream from the alpha-galactosidase A promoter is unknown. The search of the nucleotide data base revealed homologous sequences in 11 other genes (Table 4). The GA-rich sequence is located 5'- to the CAP site of several of these genes (human histone H4, mouse alpha-2 type I collagen, mouse glandular kallikrein, and rat cardiac alpha-myosin heavy chain). It is present as a spacer between histone genes in sea urchins. The sequence occurs 3'- to the chimpanzee Alu type DNA and within the coding region for the rat 28S rRNA. The GA-rich sequence is located in the 5'- flanking regions of the rat 18S rRNA where it encodes TC-rich regions postulated to direct processing and cleavage of pre-ribosomal RNA. This sequence was also identified in genomic clones of Drosophila virilis isolated by screening a genomic library with mRNA labelled in vitro. These and other simple sequences are present, as judged by hybridization, in the genomes of all eukaryotes examined (e.g. human, frog, sea urchin, wheat, and yeast). Tautz and Renz (1984) proposed that these are non-functional transcripts encoded by sequences arising from slippage reactions in regions of the genome that are not strongly selected.

To the best of our knowledge, there have been no reports of the isolation, cloning, and sequencing of a genomic clone specific for a mammalian lysosomal hydrolase. The mechanisms that regulate the levels of expression of alpha-galactosidase A and other lysosomal enzymes are

not well understood. The characterization and discussions reported in this study for the 5'- flanking promoter regions, signal peptide, propeptide, and putative carboxy-terminal processing of the propeptide to the mature enzyme for human alpha-galactosidase A, lay a foundation for future studies. This information will help to elucidate the understanding of the expression of the alpha-galactosidase A locus in normal individuals as well as in Fabry patients, and for the analyses of factors that control regulation of lysosomal enzyme synthesis.

At the molecular level, restriction enzyme information will help elucidate restriction fragment length polymorphisms (RFLP) of unrelated Fabry patients DNA in Southern blot analyses. A computer-generated restriction endonuclease map is presented in Fig. 28 and enzyme sites are shown beneath the numbered nucleotide sequence of the genomic clone. The nucleotide sequence information described in this study in concert with future studies to complete the genomic sequence will define the organization of this gene on the long arm of the X chromosome (Fig. 29). Also, the nucleotide sequence will serve as a basis for the isolation of specific probes for heterozygote identification and for comparison of the normal molecular state to that of the various Fabry patients. The availability of both genomic and cDNA clones for alpha-galactosidase A and other lysosomal hydrolases will further delineate the nature of lysosomal enzyme biosynthesis and trafficking to the lysosomes. Moreover, this information will expedite the microbial and mammalian expression and purification of active human alpha-galactosidase A for in vitro and in vivo trials of enzyme replacement therapy.

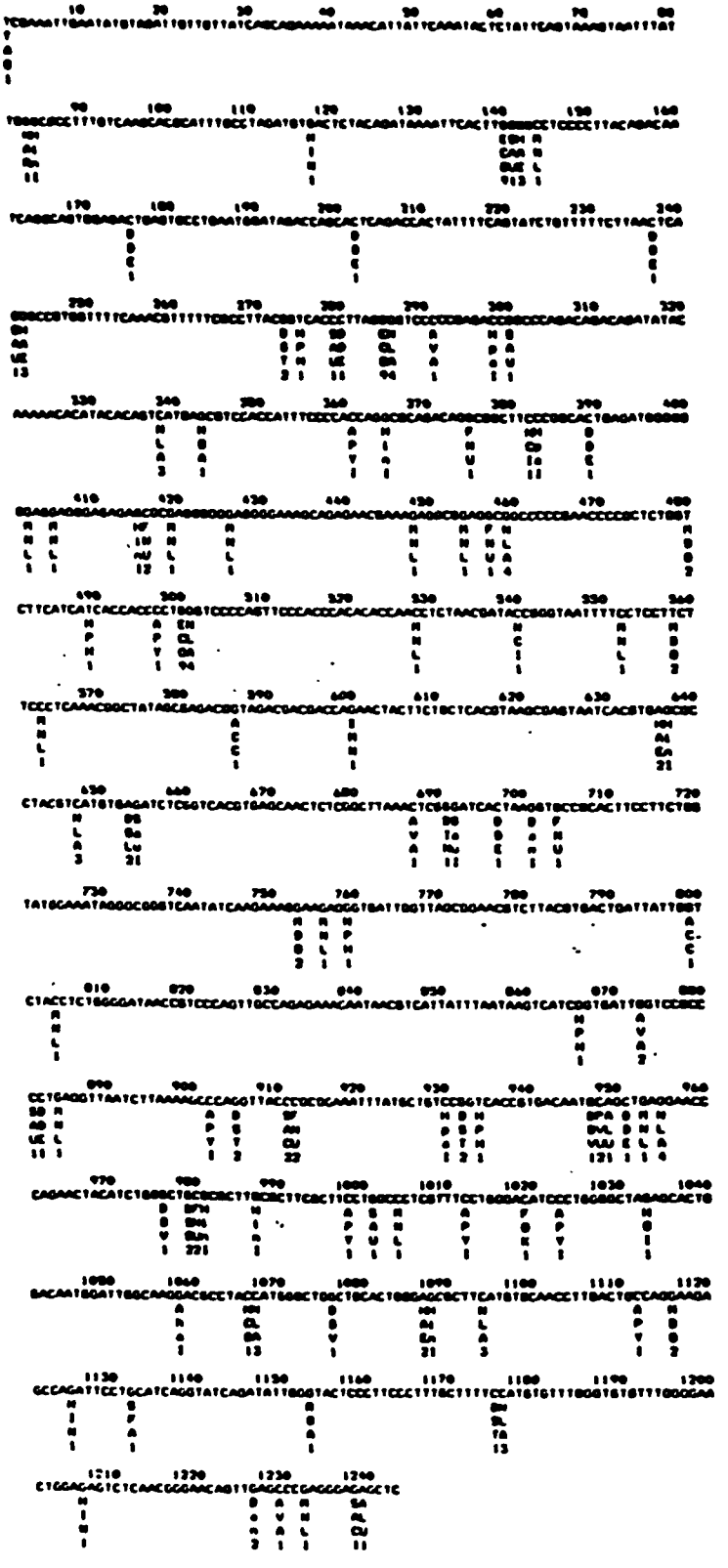


Fig. 28. Genomic alpha-galactosidase A restriction endonuclease sites. The restriction enzyme sites found within the 1243 nucleotide genomic fragment are shown beneath the nucleotide where they occur. Nucleotide numbering is the same as used in Fig. 25.

Fig. 29. Linear depiction of the alpha-galactosidase A genomic fragment. The 1243 nucleotide TaqI (T) to SacI (S) fragment located on the long arm of the X chromosome (Xq) is shown. The GA-rich region spanning nucleotides 391 to 456 (Table 4) is indicated by the solid box. The TATA box (open box) is located at position -31 before the putative transcription initiation site (see text). The slashed lines following the TATA box indicate the 60 nucleotides of 5' untranslated sequence. The open box of exon one encodes 64 amino acids including the 31 amino acid signal peptide and the first 33 amino acids of the propeptide. The intron following exon one to the SacI site contains 104 nucleotides. The following restriction enzyme sites are shown: (H) HaeIII at nucleotide 243; (A) AvaI at nucleotides 293, 689, and 1231; (C) AccI at nucleotides 498 and 801; and (R) RsaI at nucleotide 1155.

BIBLIOGRAPHY

Allen, G. 1981. In: Sequencing of Proteins and Peptides. Elsevier, New York. pp. 53-54.

Anderson, S., Bankier, A.T., Barrell, B.G., de Bruijn, M.H.L., Coulson, A.R., Drouin, J., Eperon, I.C., Nierlich, D.P., Roe, B.A., Sanger, F., Schreier, P.H., Smith, A.J.H., Staden, R., and Young, I.G. 1981. Sequence and organization of the human mitochondrial genome. Nature 290: 457-464.

Anderson, W. 1898. A case of angiokeratoma. Brit. J. Derm. 10: 113-118.

Appleyard, R.K. 1954. Segregation of new lysogenic types during growth of a doubly lysogenic strain derived from E. coli K12. Genetics 39: 440-449.

Baker, C.C. and Ziff, E.B. 1981. Promoters and heterogeneous 5' termini of messenger RNAs of adenovirus serotype 2. J. Mol. Biol. 149: 189-221.

Bankaitis, V.A., Johnson, L.M., and Emr, S.D. 1986. Isolation of yeast mutants defective in protein targeting to the vacuole. Proc. Natl. Acad. Sci. USA. 83: 9075-9079.

Benoist, C., O'Hare, K., Breathnach, R., and Chambon, P. 1980. The ovalbumin gene-sequence of putative control regions. Nucl. Acids Res. 8: 127-142.

Benton, W.D. and Davis, R.W. 1977. Screening λgt recombinant clones by hybridization in situ. Science 196: 180-182.

Berget, S.M. 1984. Are U4 small nuclear riboproteins involved in polyadenylation? Nature 309: 179-182.

Birnboim, H.C. and Doly, J. 1979. A rapid alkaline extraction procedure for screening recombinant plasmid DNA. Nucl. Acids Res. 7: 1513-1523.

Bishop, D.F. and Desnick, R.J. 1981. Affinity purification of alpha-galactosidase A from human spleen, placenta, and plasma with elimination of pyrogen contamination. J. Biol. Chem. 256: 1307-1316.

Bishop, D.F., Grabowski, G.A., and Desnick, R.J. 1981. Fabry disease: an asymptomatic hemizygote with significant residual alpha-galactosidase A activity. Am. J. Hum. Genet. 33: 71A.

Bishop, D.F., Calhoun, D.H., Bernstein, H.S., Hantzopoulos, P., Quinn, M., and Desnick, R.J. 1986. Human alpha-galactosidase A: nucleotide sequence of a cDNA clone encoding the mature enzyme. Proc. Natl. Acad. Sci. USA. 83: 4859-4863.

Bolivar, F., Rodriguez, R.L.,Greene, P.J., Betlach, M.C.,Heynecker, H.L., Boyer, H.W., Crosa, J.H., and Falkow, S. 1977. Construction and characterization of new cloning vehicles. II. A multipurpose cloning system. Gene 2:95-113.

Boyer, H.W. and Roulland-Dussoix, D. 1969. A complete analysis of the restriction and modification of DNA in E. coli. J. Mol. Biol. 41: 459-472.

Brady, R.O. and Barranger, J.A. 1983. Glucosylceramide lipidosis: Gaucher's disease. In: The Inherited Basis of Metabolic Diseases. Stanbury, J.B., Wyngaarden, J.B., Fredrickson, D.S., Goldstein, J.L., and Brown, M.S. (eds.) McGraw-Hill, New York. pp.842-846.

Brady, R.O., Gal, A.E., Bradley, R.M., Martensson, E., Warshaw, A.L., and Laster, L. 1967. Enzymatic defect in Fabry's disease: ceramide trihexosidase deficiency. New Eng. J. Med. 276: 1163-1167.

Brady, R.O., Tallman, J.F., Johnson, W.G., Gal, A.E., Leahy, W.R., Quirk, J.M., and Dekaban, A.S. 1973. Replacement therapy for inherited enzyme deficiency. Use of purified ceramidetrihexosidase in Fabry's disease. N. Engl. J. Med. 289: 9-14.

Browne, C.A., Bennent, H.P.J., and Solomon, S. 1982. The isolation of peptides by high-performance liquid chromotography using predicted elution positions. Anal. Biochem. 124: 201-208.

Burda, C.D., and Winder, P.R. 1967. Angiokeratoma corporis diffusum universale (Fabry's disease) in female subjects. Amer. J. Med. 42: 293-299.

Calhoun, D.H., Bishop, D.F., Bernstein, H.S., Quinn, M., Hantzopoulos, P., and Desnick, R.J. 1985. Fabry disease: isolation of a cDNA clone encoding human alpha-galactosidas A. Proc. Natl. Acad. Sci. USA. 82: 7364-7368.

Cassidy, B.G., Subrahmanyan, C.S., and Rothblum, L.I. 1982. The nucleotide sequence of rat 18S rDNA and adjoining spacer. Biochem. Biophys. Res. Comm. 107: 1571-1576.

Chan, S.J., San Segundo, B., McCormick, M.B., and Steiner, D.F. 1986. Nucleotide and predicted amino acid sequences of clones human and mouse prepro-cathepsin B cDNAs. Proc. Natl. Acad. Sci. USA. 83: 7721-7725.

Chan, Y-L., Olvera, J., and Wool, I.G. 1983. The structure of rat 28S ribosomal ribonucleic acid inferred from the sequence of nucleotides in a gene. Nucl. Acids Res. 11: 7819-7831.

Colley, J.R., Miller, D.L., Hutt, M.S.R., Wallace, H.J., and de Wardener, H.E. 1958. The renal lesion in angiokeratoma corporis diffusum. Brit. J. Med. 1: 1266-1275.

Corden, J., Wasylyk, B., Buchwalder, A., Sassone-Corsi, P., Kedinger, C., and Chambon, P. 1980. Promoter sequences of eukaryotic protein-coding genes. Science 209: 1406-1414.

Curtiss, R., III, Inoue, M., Pereira, D., Hsu, J.C., Alexander, L., and Rock, L. 1977. Construction in use of safer bacterial host strains for recombinant DNA research. In: Molecular Cloning of Recombinant DNA. Scott, W.A. and Werner, R.A. (eds.) Academic Press, New York. 13: 99-114.

Dale, R.M.K., Mc Clure, B.A., and Houchins, J.P. 1985. A rapid single-stranded cloning strategy for producing a sequential series of overlapping clones for use in DNA seqencing: application to sequencing the corn mitochondrial 18S rDNA. Plasmid 13: 31-40.

Davies, K.E., Young, B.D., Elles, R.G., Hill, M.E., and Williamson, R. 1981. Cloning of a representative genomic library of the human X chromosome after sorting by flow ctometry. Nature 293: 374-376.

Dawson, G., Matalon, R., and Li, Y.T. 1973. Correction of the enzymatic defect in cultured fibroblasts from patients with Fabry's disease: treatment with purified alpha-galactosidase from ficin., Pediatr. Res. 7: 684- 690.

de la Chapelle, A. and Miller, O.J. 1979. In: Human Gene Mapping 5. pp. 47-48.

Desnick, R.J. (ed.). 1980. In: Enzyme Therapy in Genetic Diseases: 2. Alan R. Liss Inc., New York.

Desnick, R.J., Dean, K.J., Grabowski, G.A., Bishop, D.F., and Sweeley, C.C. 1979. Enzyme therapy in Fabry disease: differential in vivo plasma clearance and metabolic effectiveness of plasma and splenic alpha-galactosidase A isozymes. Proc. Natl. Acad. Sci. USA. 76: 5326-5330.

Desnick, R.J., Dean, K.J., Grabowski, G.A., Bishop, D.F., and Sweeley, C.C. 1980. Enyme therapy XVII: metabolic and immunologic evaluation of alpha-galactosidase A replacement in Fabry disease. In: Enzyme Therapy in Genetic Diseases. 2. Desnick, R.J. (ed.) Alan R. Liss, Inc., New York. pp. 393-413.

Desnick, R.J. and Grabowski, G.A. 1981. Advances in the treatment of inherited metabolic diseases. In: Advances in Human Genetics. Harris, H. and Hirschhorn, K. (eds.) 12: 281-370.

Desnick, R.J. and Sweeley, C.C. 1983. Fabry's disease: glycosphingolipidoses. In: The Metabolic Basis of Inherited Disease. Stanbury, J.B., Wyngaarden, J.B., Fredrickson, D.S., Goldstein, J.L., and Brown, M.S. (eds.) McGraw-Hill, New York. pp. 906-944.

Desnick, R.J., Thorpe, S.R., and Fiddler, M.B. 1976. Toward enzyme therapy. Physiol. Rev. 56: 57-99.

de Wet, J.R., Fukushima, H., Dewji, N.N., Wilcox, E., O'Brien, J.S., and Helsinki, D.R. 1984. Chromogenic immunodetection of human serum albumin and alpha-L-fucosidase clones in a human hepatoma cDNA expression library. DNA 3: 437-447.

Dingle, J.T., Dean, R.T., and Sly, W.S. (eds.). 1984. In: Lysosomes in Biology and Pathology. Elsevier, New York.

Dyan, W.S. and Tjian, R. 1983. The promoter-specific transcription factor Sp1 binds to the upstream sequences in the SV40 early promoter. Cell 35: 79-87.

Edelhoch, H. 1967. Spectroscopic determination of tryptophan and tyrosine in proteins. Biochem. 6: 1948-1954.

Efstratiadis, A., Posakony, J., Maniatis, T., Lawn, R., O'Connell, C., Spritz, R., de Riel, J., Blechl, A., Smithies, O., Baralle, F., Shoulders, C., and Proudfoot, N. 1980. The structure and evolution of the human beta-globin gene family. Cell 21: 653-668.

Erickson, A.H., Conner, G.E., and Blobel, G. 1981. Biosynthesis of a lysosomal enzyme. J. Biol. Chem. 256: 11224-11231.

Erickson, A.H. and Blobel, G. 1983. Carboxyl-terminal proteolytic processing during biosynthesis of lysosomal enzymes beta-glucuronidase and cathepsin D. Biochem. 22: 5201-5205.

Erickson, A.H., Walter, P., and Blobel, G. 1983. Translocation of a lysosomal enzyme across the microsomal membrane requires signal recognition particle. Biochem. Biophys. Res. Comm. 115: 275-280.

Evans, J.H.M., Hamerton, J.L., Klinger, H.P., and McKusick, V.A. 1979. Human Gene Mapping 5, Cyto. Cell. Genet. p. 25.

Fabry, J. 1898. Ein beitragzur kenntnis der purpura haemorrhagica nodularis (Purpura papulosa hemorrhagica hebrae). Arch. Derm. Syph. 43: 187-196.

Faust, P.L., Kornfeld, S., and Chirgwin, J.M. 1985. Cloning and sequence analysis of a cDNA for human cathepsin D. Biochem. 82: 4910-4914.

Fiddes, J.C. and Goodman, H.M. 1980. Isolation, cloning, and sequence analysis of the cDNA for the alpha-subunit of human chorionic gonadotropin. Nature 281: 351-356.

Fong, D., Calhoun, D., Hsieh, W-T., Lee, B., and Wells, R.D. 1986. Isolation of a cDNA clone for the human lysosomal proteinase cathepsin B. Proc. Natl. Acad. Sci. USA. 83: 2909-2913.

Fukishima, H., de Wet, J.R., and O'Brien, J.S. 1984. Molecular cloning of a cDNA for human alpha-L-fucosidase. Proc. Natl. Acad. Sci. USA. 82: 1262-1265.

Gabel, C.A., Goldberg, D.E., and Kornfeld, S. 1983. Identification and characterization of cells deficient in mannose 6-phosphate receptor: evidence for an alternate pathway for lysosomal enzyme targeting. Proc. Natl. Acad. Sci. USA. 80: 775-779.

Gadek, J.E. and Crystal, R.G. 1983. Alpha$_1$-antitrypsin deficiency. In: The Metabolic Basis of Inherited Diseases. Stanbury, J.B., Wyngaarden, J.B., Fredrickson, D.S., Goldstein, J.L., and Brown, M.S. (eds.) McGraw-Hill, New York. pp.1450-1467.

Garver, R.I., Chytil, A., Karlsson, S., Felts, G.A., Brantly, M.L., Courtney, M, Kantoff, P.W., Nienhuis, A.W., Anderson W.F., and Crystal, R.G. 1987. Production of glycosylated physiologically "normal" human alpha-antitrypsin by mouse fibroblasts modified by insertion of a human alpha-antitrypsin cDNA using a retroviral vector. Proc. Natl. Acad. Sci. USA. 84: 1050-1054.

Gieselmann, V., Pohlmann, R., Hasilik, A., and von Figura, K. 1983. Biosynthesis and transport of cathepsin D in cultured fibroblasts. J. Cell Biol. 97: 1-5.

Ginns, E.I., Choudary, P.V., Martin, B.M., Winfield, S., Stubblefield, B., Mayor, J, Merkle-Lehman, D., Murray, G.J., Bowers, L.A. and Barranger, J.A. 1984. Isolation of cDNA clones for human beta-glucocerebrosidase using λgt11 expression system. Biochem. Biophys. Res. Comm. 123: 574-580.

Gonzalez-Noeiega, A., Grubb, J.H., Fallad, V., and Sly, W.S. 1980. Chloroquine inhibits lysosomal enzyme pinocytosis and enhances secretion by impairing receptor recycling. J. Cell Biol. 8: 839-852.

Grantham, R., Gautier, C., Gouy, M., Jacobzone, M., and Mercier, R. 1981. Codon catalog usage is a genome strategy for gene expressivity. Nucl. Acids Res. 9: r43-r74.

Gross, E. 1967. The cyanogen bromide reaction. Methods Enzymol. 11: 238-255.

Guise, K.S., Korneluk, R.G., Waye, J., Lamhonwah, A., Quan, F., Palmer, R., Granschow, R.E., Sly, W.S., and Gravel, R.A. 1985. Isolation and expression in Escherichia coli of a cDNA clone encoding human beta-glucuronidase. Gene 34: 105-110.

Hanahan, D. 1983. Studies on transformation of Escherichia coli with plasmids. J. Mol. Biol. 166: 557-580.

Hanahan, D. and Meselson, M. 1980. Plasmid screening at high colony density. Gene 10: 63-67.

Hasilik, A., Voss, B., and von Figura, K. 1981. Transport and processing of lysosomal enzymes by smooth muscle cells and endothelial cells. Exp. Cell Res. 133: 23-30.

Basilik, A., Waheed, A., and von Figura, K. 1981. Enzymatic phosphorylation of lysosomal enzymes in the presence of UDP-N-acetylglucosamine. Absence of activity in I-cell fibroblasts. Biochem. Biophys. Res. Comm. 98: 761-767.

Heinrich, G., Kronenberg, H.M., Potts, J.T., and Habener, J.F. 1984. Gene encoding parathyroid hormone. J. Biol. Chem. 259: 3320-3329.

Hentschel, C.C. 1982. Homocopolymer sequences in the spacer of a sea urchin histone gene repeat are sensitive to S1. Nature 295: 714-716.

Hickman, S. and Neufeld, E.F. 1972. A hypothesis for I-cell disease: Defective hydrolases that do not enter the lysosomes. Biochem. Biophys. Res. Comm. 49: 992-999.

Hirs, C.H.W. 1967. Determination of cysteine as cysteic acid. Methods Enzymol. 11: 59-62.

Hoflack, B. and Kornfeld, S. 1985. Lysosomal enzyme binding to mouse P388D1 macrophage membranes lacking the 215 Kd mannose 6-phosphate receptor: Evidence for the existence of a second mannose 6-phosphate receptor. Proc. Natl. Acad. Sci. USA. 80: 4428-4432.

Hoskins, J.A., Jack, G., Peiris, R.J.D., Starr, D.J.T., Wase, H.E., Wright, E.C., and Stern, J. 1980. Enzymatic control of phenylalanine intake in phenylketonuria. Lancet. 1: 392.

Hunkapiller, M.W. and Hood, L.E. 1983. Protein sequence analysis: automated microsequencing. Science 219: 650-659.

Jenh, C-H., Deng, T., Li, D., DeWille, J., and Johnson, L.F. 1986. Mouse thmidylate synthase messenger RNA lacks a 3' untranslated region. Proc. Natl. Acad. Sci. USA. 83: 8482-8486.

Johnson, D.A. and Desnick, R.J. 1978. Molecular pathology of Fabry's disease. Physical and kinetic properities of alpha-galactosidase A in cultured human endothelial cells. Biochem. Biophys. Acta 538: 195-204.

Johnston, A.W., Frost, P., Spaeth, G.L., and Renwick, J.H. 1969. Linkage relationships of the angiokeratoma (Fabry) locus. Ann. Hum. Genet. London. 32: 369-378.

Karn, J., Brenner, S., Barnett, L., and Cesareni, G. 1980. Novel bacteriophage λ cloning vector. Proc. Natl. Acad. Sci. USA. 77: 5172-5176.

Kim, S.H., Moores, J.C., David, D., Respess, J.G., Jolly, D.J., and Friedmann, T. 1986. The organization of the human HPRT gene. Nucl. Acids Res. 14: 3103-3118.

Kornfeld, R. and Kornfeld, S. 1985. Assembly of asparagine-linked oligosaccharides. Annu. Rev. Biochem. 54: 631-664.

Kornfeld, S., Lang, L., and Hoflack, B., 1985. Lysosomal enzyme targeting. Fed. Proc. 44(3):ix.

Kredich, N.M. and Hershfield, M.S. 1983. Immunodeficiency diseases caused by adenosine deaminase deficiency and purine nucleoside phosphorylase deficiency. In: The Metabolic Basis of Inherited Diseases. Stanbury, J.B., Wyngaarden, J.B., Fredrickson, D.S., Goldstein, J.L., and Brown, M.S. (eds.) McGraw-Hill, New York. pp. 1157-1183.

Kozak, M. 1984. Compilation and analysis of sequences upstream from the translational start site in eukaryotic mRNAs. Nucl. Acids Res. 12: 857-872.

Krystal, M., Li, R., Lyles, D., Pavlakis, G., and Palese, P. 1986. Expression of the three influenza virus polymerase proteins in a single cell allows growth complementation of viral mutants. Proc. Natl. Acad. Sci. USA. 83: 2709-2713.

Lang, L., Reitman, M., Tang, J., Roberts, R.M., and Kornfeld, S. 1984. Lysosomal enzyme phosphorylation. Recognition of a protein-dependent determinant allows specific phosphorylation of oligosaccharides present on lysosomal enzymes. J. Biol. Chem. 259: 14663-14671.

Leder, P., Tiemeir, D. and Enquist, L. 1977. EK2 derivatives of bacteriophage λ useful in the cloning of DNA from higher organisms: the λgtWES system. Science 196: 175-177.

LeDonne, N.C., Fairley, J.L., and Sweeley, C.C. 1983. Biosynthesis of alpha-galactosidase A in cultured Chang liver cells. Arch. Biochem. Biophys. 224: 186-195.

Leytus, S.P., Chung, D.W., Kisiel, W., Kurachi, S., and Davie, E.W. 1984. Characterization of a cDNA coding for human factor X. Proc. Natl. Acad. Sci. USA. 81:3699-3702.

Lipman, D.J. and Pearson, W.R. 1985. Rapid protein sequence similarity searches. Science 227: 1435-1441.

Loh, D.Y., Bothwell, A.L.M., White-Scharf, M.E., Imanishi-Kari, T., and Baltimore, D. 1983. Molecular basis of a mouse strain-specific anti-hapten response. Cell 33: 85-93.

Maeda, N., Bliska, J.B., and Smithies, O. 1983. Recombination and balanced chromosome polymorphism suggested by DNA sequence 5' to the human gamma-globin gene. Proc. Natl. Acad. Sci. USA. 80: 5012-5016.

Mahdavi, V., Chambers, A.P., and Nadal-Ginard, B. 1984. Cardiac alpha- and beta-myosin heavy chain genes are organized in tandem. Proc. Natl. Acad. Sci. USA. 81: 2626-2630.

Mapes, C.A., Anderson, R.L., Desnick, R.J., Krivit, W., and Sweeley, C.C. 1970. Enzyme replacement as a possible therapy for Fabry's disease. Science 169: 987-989.

Maniatis, T., Fritsch, E.F., and Sambrook, J. 1982. Rapid, small-scale isolation of bacteriophage λ DNA by liquid culture. In: Molecular Cloning: A Laboratory Manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York. p. 373.

Mason, A.J., Evans, B.A., Cox, D.R., Shine, J., and Richards, R.I. 1983. Structure of mouse kallikrein gene family suggests a role in specific processing of biologically active peptides. Nature 303: 300-307.

Maxam, A.M. and Gilbert, W. 1980. Sequencing by chemical modification. Methods Enzymol. 65: 499-560.

McGraw, R.A., III. 1984. Dideoxy DNA sequencing with end labelled oligo-nucleotide primer. Anal. Biochem. 143: 298-303.

McKusick, V.A. 1978. In: Mendelian Inheritance in Man, 5th ed., Johns Hopkins Press, Baltimore.

McKusick, V.A., Neufeld, E.F., and Kelley, T.E. 1978. The mucopolysaccharide storage diseases. In: The Metabolic Basis of Inherited Disease. Stanbury, J.B., Wyngaarden, J.B., and Fredrickson, D.S. (eds.) McGraw-Hill, New York. pp. 1282-1307.

Mendel, G. 1901. Versuche uber Pflanzenhybriden. Engelmann, Leipzig.

Messing, J. 1983. New M13 vectors for cloning. Methods Enzymol. 101: 20-78.

Messing, J. and Vierra, J. 1982. A new pair of M13 vectors for selecting either DNA strand of double digest restriction fragment. Gen. 19: 269-276.

Mount, S.M. and Steitz, J.A. 1983. Signals for the splicing of eukaryotic messenger RNA transcripts. In: Methods of DNA and RNA Sequencing. Weissman, S.M. (ed.) Praeger Publishers, New York. pp. 399-424.

Myerowitz, R. and Neufeld, E.F. 1981. Maturation of L-iduronidase in cultured human fibroblasts. J. Biol. Chem. 256: 3044-3048.

Myerowitz, R., Piekarz, R., Neufeld, E.F., Shows, T.B., and Suzuki, K. 1985. Human beta-hexosaminidase alpha chain: coding sequence and homology with the beta chain. Proc. Natl. Acad. Sci. USA. 82: 7830-7834.

Nishimura, Y., Rosenfeld, M.G., Kreibich, G., Gubler, U., Sabatini, D.D., Adesnik, M., and Andy, R. 1986. Nucleotide sequence of rat preputial gland beta-glucuronidase cDNA and in vitro insertion of its encoded polypeptide into microsomal membranes. Proc. Natl. Acad. Sci. USA. 83: 7292-7296.

Noyes, B.E., Mevarech, M., Stein, R., and Agarwal, K.L. 1979. Detection and partial sequence analysis of gastrin mRNA by using an oligonucleotide probe. Proc. Natl. Acad. Sci. USA. 76: 1770-1774.

O'Brien, J.S. 1983. The gangliosidoses. In: The Metabolic Basis of Inherited Diseases. Stanbury, J.B., Wyngaarden, J.B., Fredrickson, D.S., Goldstein, J.L., and Brown, M.S. (eds.) McGraw-Hill, New York.

O'Dowd, B.F., Quan, F., Willard, H.F., Lamhonwah, A-M., Korneluk, R. G., Lowden, A., Gravel, R.A., and Mahuran, D.J. 1985. Isolation of cDNA clones coding for the beta-subunit of human beta-hexosaminidase. Proc. Natl. Acad. Sci. USA. 82: 1184-1188.

Okayama, H. and Berg, P. 1983. A cDNA cloning vector that permits expression of cDNA inserts in mammalian cells. Mol. Cell. Biol. 3: 280-289.

Opitz, J. 1964. Angiokeratoma corporis diffusum. Arch. Derm. Chicago. 90: 330-336.

Opitz, J.M., Stiles, F.C., Wise, D., von Gemmingen, G., Race, R.R., Sander, R., Cross, E.G., and de Groot, W.P. 1965. The genetics of angiokeratoma corporis diffusum (Fabry's disease) and its linkage with Xg(a) locus. Amer. J. Hum. Genet. 17: 325-334.

Oshima, A., Kyle, J.W., Miller, R.D., Hoffmann, J.W., Powell, P.P., Grubb, J.H., Sly, W.S., Tropak, M., Guise, K.S., and Gravel, R.A. 1987. Cloning, sequencing, and expression of a cDNA for human beta-glucuronidase. Proc. Natl. Acad. Sci. USA. 84: 685-689.

Owada, M. and Neufeld, E. F. 1982. Is there a mechanism for introducing acid hydrolases into liver lysosomes that is independent of mannose-6-phosphate recognition. Evidence from I-cell disease. Biochem. Biophys. Res. Comm. 105: 814-820.

Paigen, K. 1979. Acid hydrolases as models of genetic control. Ann. Rev. Genet. 13: 417-466.

Palmer, T.D., Hock, R.A., Osborne, W.R.A., and Miller, A.D. 1987. Efficient retrovirus-mediated transfer and expression of a human adenosine deaminase gene in diploid skin fibroblasts from an adenosine deaminase-deficient human. Proc. Natl. Acad. Sci. USA. 84: 1055-1059.

Perlman, D. and Halvorson, H.O. 1983. A putative signal peptidase recognition site and sequence in eukaryotic and prokaryotic signal peptides. J. Mol. Biol. 167: 391-409.

Pompen, A.W.M., Ruiter, M., and Wyers, H.J.G. 1947. Angiokeratoma corporis diffusum universale (Fabry) as a sign of unknown internal disease: two autopsy reports. Acta. Med. Scand. 128: 234-240.

Porter, M.T., Fluharty, A.L., and Kiharce, H. 1971. Correction of abnormal cerebroside sulfate metabolism in cultured metachromatic leukodystrophy fibroblasts. Science 172: 1263-1265.

Reitman, M.L., and Kornfeld, S. 1981. UDP-N-acetylglucosamine: glycoprotein N-acetylglucosamine-1-phosphotransferase. J. Biol. Chem. 256: 4275-4281.

Reitman, M.L., Varki, A., and Kornfeld, S. 1981. Fibroblasts from patients with I-cell disease and pseudo-Hurler polydystrophy are deficient in uridine 5'-diphosphate-N-acetylglucosamine: glycoprotein N-acetylglucosaminylphosphotransferase activity. J. Clin. Invest. 67: 1574-1579.

Rigby, P.W.J., Dieckmann, M., Rhodes, C., and Berg, P. 1977. Labeling deoxyribonucleic acid to high specific activity in vitro by nick translation with DNA polymerase I. J. Mol. Biol. 113: 237-251.

Rimm, D.L., Horness, D., Kucera, J., and Blattner, F.R. 1980. Construction of coliphage λ Charon vectors with BamHI cloning sites. Gene 12: 301-309.

Roeder, R.G. 1976. Eukaryotic nuclear RNA polymerases. In: RNA Polymerases. Losick, R. and Chamberlin, M. (eds.) Cold Spring Harbor Press, Cold Spring Harbor, New York.

Rosenfeld, M G., Kreibich, G., Popov, D., Kato, K., and Sabatini, D. D. 1982. Biosynthesis of lysosomal hydrolases: Their synthesis in bound polysomes and the role of co- and post-translational processing in determining their subcellular distribution. J. Cell Biol. 93: 135-143.

Sanger, F., Nicklen, S., and Coulson, A.R. 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA. 74: 5463-5467.

San Segundo, B., Chan, S.J., and Steiner, D.F. 1985. Identification of cDNA clones encoding a precursor of rat liver cathepsin B. Proc. Natl. Acad. Sci. USA. 82: 2320-2324.

Schmidt, A., Yamada, Y., and de Crombrugghe, B. 1984. DNA sequence comparison of the regulatory signals at the 5' end of the mouse and chick alpha 2 type I collagen genes. J. Biol. Chem. 259: 7411-7415.

Scriba, K. 1950. Zur pathogenese des angiokeratoma corporis diffusum Fabry mit cardio-vasorenalem symptomenkomplex. Verh. Deutsch. Ges. Path. 34: 221-228.

Sierra, F., Stein G., and Stein, J. 1983. Structure and in vitro transcription of a human H4 histone gene. Nucl. Acids Res. 11: 7069-7086.

Siniscalco, M., Szabo, P., Fillippi., G., and Rinaldi, A. 1982. Combination of old and new strategies for the molecular mapping of the human X-chromosome. In: Progress in Clinical and Biological Research. Alan R. Liss, Inc., New York. 103A: 103-124.

Sloan, H.R. and Fredrickson, D.S. 1972. G^{M2} gangliosidoses: Tay-Sachs disease. In: The Metabolic Basis of Inherited Disease. Stanbury, J.B., Wyngaarden, J.B., and Fredrickson, D.S. (eds.) McGraw-Hill, New York. pp. 615-662.

Sly, W.S. and Fischer, H.D. 1982. The phosphomannosyl recognition system for intracellular and intercellular transport of lysosomal enzymes. J. Cell Biochem. 18: 67-85.

Sogawa, K., Gotoh, O., Kawajiri, K., and Fuji-Kuriyama. 1984. Distinct organization of methylcholanthrene- and phenobarbital-inducible cytochrome p-450 genes in the rat. Proc. Natl. Acad. Sci. USA. 81: 5066-5070.

Sood, A.K., Pereira, D., and Weissman, S.M. 1981. Isolation and partial nucleotide sequence of a cDNA clone for human histocompatibility antigen HLA-B by use of an oligodeoxynucleotide primer. Proc. Natl. Acad. Sci. USA. 78: 616-620.

Sorge, J., Kuhl, W., West, C., and Beutler, E. Complete correction of the enzymatic defect of type I Gaucher disease fibroblasts by retroviral-mediated gene transfer. Proc. Natl. Acad. Sci. USA. 84:906-909.

Sorge, J., West, C., Westwood, B., and Beutler, E. 1985. Molecular cloning and nucleotide sequence of human glucocerebrocidase cDNA. Proc. Natl. Acad. Sci. USA. 82: 7289-7293.

Southern, E.M. 1975. Detection of specific sequences among DNA fragments separated by gel electrophoresis. J. Mol. Biol. 98: 503-517.

Stanbury, J.B., Wyngaarden, J.B., Fredrickson, D.S., Goldstein, J.L., and Brown, M.S. (eds.). 1983. Inborn errors of metabolism in the 1980's. In: The Metabolic Basis of Inherited Diseases. McGraw-Hill, New York. pp. 3-38.

Stiles, F.D., and Opitz, J.M. 1963. Diffuse angiokeratosis (Fabry's disease) in children (abstract). Meeting of the Midwest Society for Pediatric Research, Chicago, November.

Sutcliffe, J.G. 1978. pBR322 restriction map marked from the DNA sequence. Accurate DNA size markers up to 4361 nucleotide pairs long. Nucl. Acids Res. 5: 2721-2728.

Sweeley, C.C., and Klionsky, B. 1963. Fabry's disease: classification as a sphingolipidosis and partial characterization of a novel glycolipid. J. Biol. Chem. 238: 3148-3152.

Sweeley, C.C., Klionsky, B., Krivit, W., and Desnick, R.J. 1983. Fabry's disease: gylcosphingolipid lipidosis. In: Metabolic Basis of Inherited Diseases. Stanbury, J.B., Wyngaarden, J.B., Fredrickson, D.S., Goldstein, J.L., and Brown, M.S. (eds.) McGraw-Hill, New York, 5th ed. pp. 663-687.

Tager, J.M. 1985. Biosynthesis and deficiency of lysosomal enzymes. TIBS. 10: 324-326.

Tagler, J.M., Hoogwinkel, G.J.M., and Daems, W.Th., (eds.). 1974. In: Enzyme Therapy in Lysosomal Storage Diseases. North-Holland/American Elsevier, Amsterdam. p. 308.

Tautz, D. and Renz, M. 1984. Simple DNA sequences of Drosophila virilis isolated by screening with RNA. J. Mol. Biol. 172: 229-235.

Van Ness, B.G., Weigert, M., Coleclough, C., Mather, E.L., Kelley, D.E., and Perry, R.P. 1981. Transcription of the unrearranged mouse Ck locus: sequence of the initiation region and comparison of activity with a rearranged Vk-Ck gene. Cell 37: 593-602.

Varki, A., and Kornfeld, S. 1980. Identification of a rat liver alpha-N-acetylglucosaminyl phosphodiesterase capable of removing "blocking" alpha-N-acetylglucosamine residues from phosphorylated high mannose oligosaccharides of lysosomal enzymes. J. Biol. Chem. 255: 8398-8401.

Varki, A., and Kornfeld, S. 1981. Purification and characterization of rat liver alpha-N-acetylglucosaminyl phosphodiesterase. J. Biol. Chem. 256: 9937-9943.

Varki, A.P., Reitman, M.L., and Kornfeld, S. 1981. Identification of a variant of mucolipidosis III (pseudo-Hurler polydystrophy): A catalytically active N-acetylglucosaminylphosophotransferase that fails to phosphorylate lysosomal enzymes. Proc. Natl. Acad. Sci. USA. 78: 7773-7777.

Vierra, J. and Messing, J. 1982. The pUC plasmids: an M13mp7 derived system for insertion, mutagenesis and sequencing with universal primers. Gene 19: 259-268.

von Heijne, G. 1983. Patterns of amino acids near signal-sequence cleavage sites. Eur. J. Biochem. 133: 17-21.

von Heijne, G. 1986. A new method for predicting signal sequence cleavage sites. Nucl. Acids Res. 14: 4683-4690.

Waheed, A., Hasilik, A., and von Figura, K. 1981. Processing of the phosphorylated recognition marker in lysosomal enzymes. J. Biol. Chem. 256: 5717-5721.

Waheed, A., Hasilik, A., and von Figura, K. 1982. UDP-N-acetylglucosamine: lysosomal enzyme precursor N-acetylglucosamine-1-phosphotransferase. J. Biol. Chem. 257: 12322-12331.

Wallace, H.J. 1958. Angiokeratoma corporis diffusum. Brit. J. Derm. 70: 354-356.

Wallace, R.B., Johnson, N.J., Hirose, T., Miyake, M., Kawashima, E.H., and Itakura, K. 1981. The use of synthetic oligonucleotides as hybridization probes. II. Hybridization of oligonucleotides of mixed sequence to rabbit beta-globin DNA. Nucl. Acids Res. 9: 879-894.

Walter, P., Gilmore, R., and Blobel, G. 1984. Protein translocation across the endoplasmic reticulum. Cell 38: 5-8.

Ware, V.C., Tague, B.W., Clark, C.G., Gourse, R.L., Brand, R.C., and Gerbi, S.A. 1983. Sequence analysis of 28S ribosomal DNA from the amphibian Xenopus laevis. Nucl. Acids Res. 11: 7795-7817.

Watson, M.E.E. 1984. Compilation of published signal sequences. Nucl. Acids Res. 12: 5145-5164.

Wickens, M. and Stephenson, P. 1984. Role of the conserved AAUAAA sequence: four AAUAAA point mutants prevent messenger RNA 3' end formation. Science 226: 1045-1051.

Wise, D., Wallace, H.J., and Jellinck, E.H. 1962. Angiokeratoma corporis diffusum: a clinical study of eight affected families. Quart. J. Med. 42: 177-185.

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, E.G.D., Vehar, G.A. and Lawn, R.M. 1984. Expressin of active human factor VIII from recombinant DNA clones. Nature 312: 330-337.

Young, R.A. and Davis, R.W. 1983a. Yeast RNA polymerase II genes: isolation with antibody probes. Science 222: 778-782.

Young, R.A. and Davis, R.W. 1983b. Efficient isolation of genes using antibody probes. Proc. Natl. Acad. Sci. USA. 80: 1194-1198.