Elucidation of Pattern of Variation for the Amylase Locus in Type 1 Diabetes Patients

THESIS

Presented in Partial Fulfillment of the Requirements for the Degree Master of Science in the Graduate School of The Ohio State University

By

Andrea Marie Rutherford, B.S.

Graduate Program in Molecular, Cellular and Developmental Biology

The Ohio State University

2012

Master's Examination Committee:

Dr. Chack Yung Yu, Advisor

Dr. Heithem El-Hodiri

Dr. Robert Hoffman

ABSTRACT

Proper metabolism of food, especially glucose, is vital to an organism's survival. Failure to metabolize polysaccharides and regulate blood glucose levels are hallmarks of diabetes mellitus. Type 1 diabetes (T1D) is a disease where pancreatic beta cells, producers of the insulin that regulates blood glucose levels, are destroyed by the immune system. Most carbohydrates yield glucose, with starch and glycogen being especially rich in glucose. Starch degradation in humans begins with α-amylase, which is responsible for the catalytic hydrolyzation of α1,4-glycosidic bonds between glucose monomers in starch. Humans express two forms of α-amylase that display high tissue specificity. Pancreatic amylase (encoded by *AMY2*) is expressed by the cells of the exocrine pancreas. Salivary amylase (encoded by *AMY1*) is expressed in the acinar and intercalated duct cells of the parotid salivary gland.

Human amylase genes are one of the first reported instances of copy number variation (CNV) among healthy individuals. Six different genes exist in a cluster on chromosome 1 in the human genome – three salivary genes (*AMY1A,*

1B, and *1C*), two pancreatic genes (*AMY2A* and *2B*) and one truncated pseudogene (*AMYP1*). In an effort to explain the diversity of amylase CNV in healthy individuals, Peter Groot developed the following: *AMY2B-AMY2A-(AMY1A-AMY1B-AMYP1)*$_n$*-AMY1C*, with n = 0 – 2 copies in a diploid individual. Evidence from other reports, such as the absence of *AMY2A* in the NCBI Alternative Reference genome, suggests that Groot's model is not an accurate description of variation at this locus. This thesis focuses on further interrogating the amylase locus in T1D individuals to better understand its pattern of variation.

Dot-plot analysis of the amylase locus indicated the presence of four large blocks ranging in size from 64kb to 44kb that share sequence homology. Within these blocks are smaller segments of 7-30kb with sequence similarity. *Pme*I digests revealed segmental duplications of ~20-50kb increments, implicating these blocks in amylase diversification.

Comparisons of *Taq*I, *Pst*I, and *Pvu*II/*Psh*AI digests revealed variable copy number for all amylase genes in individuals. Individuals with short haplotypes provided further insight into the variation and complexity of the locus. Pancreatic *AMY2* genes appear to vary in intensity throughout the population. Another individual was found to be homozygous deficient for *AMY2A*, further highlighting limitations in Groot's model. Our T1D population as a whole showed large copy number variation for all amylase genes.

In conclusion, the human amylase locus undergoes complex variation that most likely includes unequal homologous recombination and alignment of large and smaller segments sharing sequence homology. Further study of healthy individuals and quantitative measures of amylase copy number should provide further insight into a pattern of variation.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor, Dr. Chack Yung Yu, for all the scientific wisdom he imparted, his patience, the guidance he gave me throughout my time with him, and for his efforts in pushing me to be the best I could be. Without his help I would not have succeeded as well nor accomplished as much as I have been able to do in this short time here. I would also like to express my thanks to Bi, Alaaedin, and Yee Ling for teaching me various techniques and making me feel welcome. Special thanks to my committee members, Dr. Robert Hoffman and Dr. Heithem El-Hodiri, for agreeing to listen to my work and critique it. Thanks to all the other members of our center for providing an ear to listen when I needed it and suggestions when I asked. Thanks as well to my friends and family for supporting me and believing in me, even when I didn't necessarily believe in me myself. I especially want to thank Francisco for enduring a lot of late night dinners, weekend visits to the lab, and general stress that comes with dating a scientist. I love you all and couldn't accomplish what I have without your assistance. Many thanks.

VITA

2006……Archbishop McNicholas High School

2010……...B.S. Molecular Genetics, The Ohio State University

2010 to present……..…Graduate Research Assistant

Molecular, Cellular, and Developmental Biology

The Ohio State University

PUBLICATIONS

Rutherford A., Zhou B., Wu Y. L., S.E. K., Germak J., S. B., Hoffman R. P. and C.Y. Y. (2011) Analysis of the Human Amylase Locus Revealed Extensive Copy Number Variation of Salivary and Pancreatic Amylase Genes in Type 1 Diabetes Patients *Pediatric Research* **70**, 429-429. (Abstract)

FIELDS OF STUDY

Major Field: Molecular, Cellular, and Developmental Biology

TABLE OF CONTENTS

Abstract……..……ii

Acknowledgements……..……iv

VITA……...…v

List of Figures.....……..…….....vi

List of Tables.. .....……...……...…….ix

Abbreviations……..…….xi

1. Introduction…….……1

1.1 General Introduction……...……....1

1.2 Structure of Human Salivary and Pancreatic Amylases.....…….....2

1.3 Enzymatic Activity of Human Salivary and Pancreatic Amylase Proteins……..…....9

1.4 Regulation of Amylase Gene Expression…….....12

1.5 History of Copy Number Variation…….....17

1.6 Initial Studies of Amylase Gene Copy Number Variation…….....18

1.7 Copy Number Variation of Human Amylase Gene Cluster…….....23

1.8 Human Amylase Gene Copy Number Variation and Phenotypic Correlations…….....26

1.9 Type 1 Diabetes: Overview…….....……31

1.10 Genetics of Type 1 Diabetes……..……34

1.11 Goals of This Study…….....……...36

2. Elucidation of Pattern of Variation for the Amylase Locus in Type 1 Diabetes Patients…….....……...39

2.1 Introduction……..……...39

2.2 Materials and Methods……..……...42

2.2.1 Study Subject Recruitment…….....42

2.2.2 Genomic DNA samples…….....42

2.2.3 Amylase Specific Probes……...43

2.2.4 Restriction Enzymes……..…...44

2.2.5 *Taq*I, *Pvu*II/*Psh*AI, and *Pst*I Restriction Enzyme Digests of genomic DNA with Southern blot analysis…….....44

2.2.6 *Pme*I Restriction Enzyme Digests and Pulsed Field Gel Electrophoresis.....45

2.2.7 Pressure Blot Transfer, UV Crosslinking, and Hybridization.....46

2.2.8 Bioinformatic Analysis of the Human Amylase Locus.....46

2.2.9 Additional Declarations.....47

2.3 Results.....48

2.3.1 Dot Plot of the Human Amylase Locus.....48

2.3.2 Restriction Map Analysis of the Human Amylase Genes and Probe Development.....49

2.3.3 *Taq*I digest results.....51

2.3.4 *Pvu*II/*Psh*AI digest results.....53

2.3.5 *Pst*I digest results.....53

2.3.6 *Pme*I digest results.....54

2.3.7 Segmental Duplications of the *AMY* Locus.....55

2.3.8 TD80P – Simplest confirmed haplotype.....55

2.3.9 Homozygous for 354kb *Pme*I fragment and elucidation of *Taq*I Fragment H identity.....57

2.3.10 Support for Groot's Model for human amylase pattern of variation.....58

2.3.11 TD72P, TD110P, and copy number variation of *AMY2A*.....59

2.3.12 TD25P and copy number variation *AMY2B*.....60

2.3.13 Pattern of variation for the human amylase locus in T1D.....61

2.4 Discussion.....64

References.....84

LIST OF FIGURES

Fig.1.1 An alignment of human pancreatic amylase (AMY2A) and salivary amylase (AMY1A) protein sequences…….....6

Fig.1.2 Crystal structure of human salivary amylase protein…….7

Fig.1.3 Human pancreatic amylase protein……..8

Fig. 1.4 Subsites along the active site cleft and representation of polysaccharide cleavage……...11

Fig.1.5 The 5’ region of the actively transcribed amylase genes……15

Fig1.6 The human amylase gene cluster……16

Fig1.7 Initial amylase protein patterns observed in humans……...21

Fig.1.8 Human salivary protein patterns obtained from isoelectric focusing……....22

Fig.1.9 Human salivary amylase copy number and phenotype correlation……..30

Fig.2.1 Dot plot analysis of the entire human amylase locus……..72

Fig.2.2 *Taq*I RFLP Southern analyses for copy number of human salivary and pancreatic amylase genes in 10 individuals...……74

Fig.2.3 *Pvu*II/*Psh*AI RFLP Southern analysis for copy number variation of human salivary and pancreatic amylase genes in 13 individuals..……..75

Fig.2.4 *Pst*I RFLP Southern analysis for copy number of human salivary and pancreatic amylase genes……76

Fig.2.5 *Pme*I RFLP pulsed field gel electrophoresis for human amylase locus size in 13 unrelated individuals with Type 1 diabetes……...77

Fig.2.6 *Pme*I RFLP pulsed field gel electrophoresis for human amylase locus size in 13 unrelated individuals with Type 1 diabetes……...78

Fig.2.7 RFLP Southern analyses for copy number of human salivary and pancreatic amylase genes in TD80P……..79

Fig.2.8 RFLP Southern analysis for copy number of human salivary and pancreatic amylase genes in 11 unrelated individuals…….80

Fig.2.9 All RFLP digest results for TD72P/TD110P……..81

Fig.2.10 RFLP Southern analyses for copy number of human salivary and pancreatic amylase genes in TD25P……..82

Fig.2.11 Variations of human salivary and pancreatic amylase genes in the diploid genomes of Type 1 diabetes patients…….83

LIST OF TABLES

Table 2.1 Estimated fragment sizes for each amylase gene depending on which restriction enzyme and probe (AMY300, AMY345, AMY543) is used…...73

ABBREVIATIONS

3’	3 prime
5’	5 prime
AIRE	autoimmune regulator
AMY	amylase
Arg	arginine
Array CGH	array comparative genomic hybridization
Asn	asparagine
Asp	aspartic acid
BAC	bacterial artificial chromosome
BSA	bovine serum albumin
Ca^{2+}	calcium ion
cDNA	complementary DNA
CNV	copy number variation
DNA	deoxyribonucleic acid
EDTA	ethylenediamine tetra-acetic acid
EtBr	ethidium bromide
FISH	fluorescence in situ hybridization
Gln	glutamine
Glu	glutamic acid
Gly	glycine
GWA	genome-wide association
His	histidine
HLA	human leukocyte antigen
IRB	institutional review board
LGT	low gelling temperature
LTR	long terminal repeat
Lys	lysine
Met	methionine
NCBI	National Center for Biotechnology Information
PFGE	pulsed field gel electrophoresis

PMN	ploymorphonuclear neutrophils
PPA	porcine pancreatic amylase
qPCR	quantitative polymerase chain reaction
RBC	red blood cell
RFLP	restriction fragment length polymorphism
Ser	serine
SNP	single nucleotide polymorphism
T1D	Type I diabetes mellitus
TCA cycle	tricarboxylic acid cycle
T_{reg}	regulatory T cells
Trp	tryptophan
Tyr	tyrosine
VNTR	variable number tandem repeats
WBC	white blood cell

CHAPTER 1

INTRODUCTION

1. 1 General Introduction

The ability of an organism to metabolize food is critically important for its survival. In order for cells of an organism to obtain energy from the food it ingests or stores, the food must undergo numerous steps of degradation into its basic building blocks and converted into something cells can use. Ingested items are generally broken down based on their contents. Protein metabolism results in amino acids, fat metabolism yields fatty acids and glycerol, and carbohydrate metabolism yields monosaccharides. Because eukaryotes obtain the most energy from glucose, most cells prefer to convert nutrients into glucose or subsequent derivatives that can then enter glycolysis, the tricarboxylic acid (TCA) cycle, and the electron transport chain. Of all sources of carbohydrates, starch and glycogen yield the most glucose. It is, therefore, extremely important that an organism be able to properly process these complex carbohydrates.

The first step in starch degradation begins with amylase. The importance of this enzyme is highlighted by the fact that every organism, from bacteria (Fleming and Neill, 1927) to higher organisms, has some form of amylase being utilized. Discovered in the 1830s, α-amylase is expressed in humans in two forms that show high tissue specificity– salivary and pancreatic amylase (Kamaryt and Laxova, 1966). Both human salivary and pancreatic amylase are responsible for the catalytic hydrolyzation of the α-(1,4)-glycosidic bonds in starch.

1. 2 Structure of Human Salivary and Pancreatic Amylases

The primary structures of human salivary and pancreatic amylase proteins are highly similar (Figure 1.1). Both proteins consist of 496 amino acids plus a leader peptide of 15 residues. The two isozymes share 97% sequence identities, with only 16 amino acid differences occurring between them, suggesting that their secondary and tertiary structures would also be highly similar (Brayer et al., 1995). Crystallography studies have indicated that both amylases are arranged into three domains – Domains A, B, and C (Figures 1.2 and 1.3) (Brayer et al., 1995; Ramasubbu et al., 1996). Domain A is comprised of amino acid residues 1-99 and 169-404. The major structural feature of this domain is the eight stranded beta barrel that is surrounded by alpha helices. All three catalytic residues – Asp197, Glu233, and Asp300 – reside in this domain, located on the

top of the beta barrel. The active site itself forms a V-shaped cleft that houses, in addition to the scissle bond site, additional subsites where secondary substrate binding takes place (Kandra and Gyemant, 2000; Nahoum et al., 2000; Ramasubbu et al., 1996). These subsites orient the substrate within the active site. Close to the active site is a binding site for a chloride ion, which appears to neutralize the positive charge of R337, lowering the pKa of E233 and, therefore, raising the activity of the enzyme (Numao et al., 2002). The chloride ion may also play a role in forcing E233 to maintain a specific conformation that is conducive to catalysis (Maurus et al., 2005).

Domain B is comprised of amino acid residues 100-168 that form beta structures with a small helix present as well (Brayer et al., 1995). This domain forms a pocket against Domain A, within which rests a calcium ion that is required for all alpha amylases. The calcium ion is bound tightly within its binding site, interacting with at least three water molecules and four residues in the pancreatic protein (Brayer et al., 1995). The calcium ion interacts with four amino acids, three in Domain B (Asp167, Asn100, and Arg158) and one in Domain A (His201). The required presence of Ca^{2+} is noted by the fact that not only does it anchor and hold Domain B near the substrate binding cleft (hence, also stabilizing the integrity of Domain A), but it is also shown to orient His201 in the cleft, which provides the asymmetric environment the human amylase proteins employ to orient the substrate during binding (Ramasubbu et al., 1996).

Domain C, the final domain in both amylase proteins, is comprised of the C-terminal amino acids, residues 405-496. Pancreatic amylase's Domain C forms a compact antiparallel β-barrel type structure that is nestled closely against the side of Domain A opposite Domain B (Brayer et al., 1995). The salivary amylase Domain C is comprised of ten β-strands, eight of which form a flattened Greek-key topology (Ramasubbu et al., 1996). The other two strands are arranged as loops, separate from the Greek-key configuration. The interface between Domains A and C is inundated with hydrophobic residues, peppered with met residues (Ramasubbu et al., 1996). Salivary amylase, which has been shown to under glycosylation post-transcriptionally (Kauffman et al., 1973), has an N-glycosylation motif (Asn412-Gly413-Ser414) located within the Greek key configuration of Domain C (Ramasubbu et al., 1996). Although these residues are identical in both pancreatic and salivary amylase proteins, no report of glycosylation of pancreatic amylase has been published.

One further similarity shared between both proteins is the spontaneous formation of pyrrolid-2-one-5-carboxylic acid, which is formed when a side chain carbon atom of the N-terminal glutamine becomes covalently bonded to the main-chain amino group (Brayer et al., 1995; Ramasubbu et al., 1996). This occurs as the protein is being synthesized and is believed to protect the proteins from degradation by other digestive enzymes in the surrounding environment.

The vast majority of amino acid substitutions between human salivary and pancreatic amylase proteins (13 of the 16 substitutions) occurs within Domain A.

Several of these substitutions are believed to cause the slight difference in activity between salivary and pancreatic amylases. The substitution of isoleucine for leucine at residue 196 (L196I) is implicated because it is close to three other important residues – Arg195, which binds the chloride ion; Asp197, a catalytic residue; and His 201, which binds the calcium ion (Brayer et al., 1995). Five more substitutions occur along a loop containing residues 341-288, possibly causing polarity differences that would affect Km values and changing specificity for inhibitors (Ramasubbu et al., 1996). The residue substitutions at residues 349 (Gln in the pancreatic enzyme to Glu) and 352 (Asn in pancreatic enzyme to a Lys) occur at one of the subsites of the cleft, possibly altering the affinity of salivary amylase for larger substrates (Ramasubbu et al., 1996).

An alignment of human pancreatic amylase (AMY2A) and salivary amylase (AMY1A)

Figure 1.1. An alignment of human pancreatic amylase (AMY2A) and salivary amylase (AMY1A) protein sequences.

Figure 1.2. Crystal structure of Human salivary amylase protein. Stereoview of human salivary amylase. The locations of Ca^{2+} and chloride ions are also shown. Taken from Ramasubbu et al., 1996.

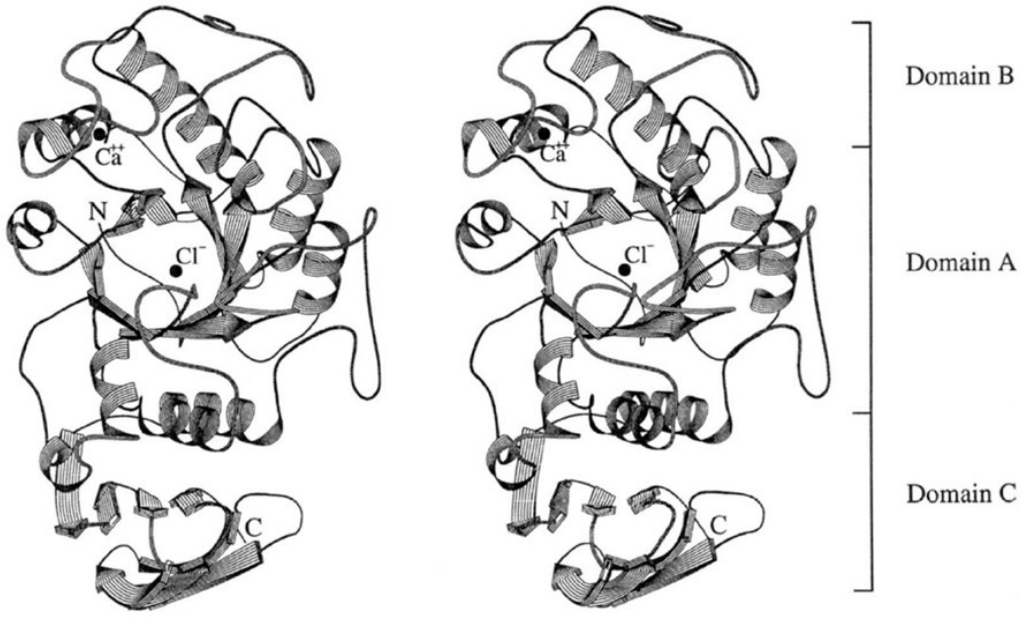


Figure 1.3. Human pancreatic amylase protein. Stereoview of human pancreatic amylase. Overall fold of the enzyme is shown, along with the Ca^{2+} and chloride ions. Domains are indicated. Taken from Brayer et al., 1995.

1. 3 Enzymatic Activity of Human Salivary and Pancreatic Amylase Proteins

Both human salivary and pancreatic amylase are responsible for the hydrolyzation of the α-(1,4)-glycosidic bonds in starch. This hydrolyzation occurs within the cleft located at the active site in Domain A. Both amylases contain a v-shaped cleft in Domain A where the active site is located (Brayer et al., 1995; Kandra and Gyemant, 2000; Ramasubbu et al., 1996). Along this cleft are a number of subsites where the enzymes can bind their substrate (

Figure). While both enzymes employ a total of six subsites – four glycon and three aglycon sites – only salivary amylase utilizes all six during catalysis (Brayer et al., 2000; Kandra and Gyemant, 2000). The preference of pancreatic amylase to use five of its subsites may contribute to the difference in enzymatic activity noted between the enzymes. The orientation of starch is also aided by a combination of aromatic residues, such as Trp58, Trp59, and Tyr62, and histidine residues, all of which either directly interact with starch within the glycon sites or facilitate this interaction (Ragunath et al., 2008).

Despite having different activities in response to starch (Kaczmarek and Rosenmund, 1977), both amylases employ the same mechanism of double displacement to catalyze the hydrolyzation of starch (Brayer et al., 2000). Following substrate binding, the process begins when Asp197 (Rydberg et al., 2002), aided by acid catalysis from Glu233 or Asp300, attacks the anomeric carbon via nucleophilic displacement. This attack results in a covalently bonded

B-glycosyl enzyme intermediate and a freed glucose monomer of maltose, which is released with the aid of a glycine-rich loop (residues 304-310) (Ramasubbu et al., 2003). With the help of the same Glu233 or Asp300 residue acting this time as base, a water molecule attacks the same anomeric center via nucleophilic displacement. The starch molecule is then free to reorient itself within the enzyme to be re-hydrolyzed once again (Brayer et al., 2000; Numao et al., 2002).

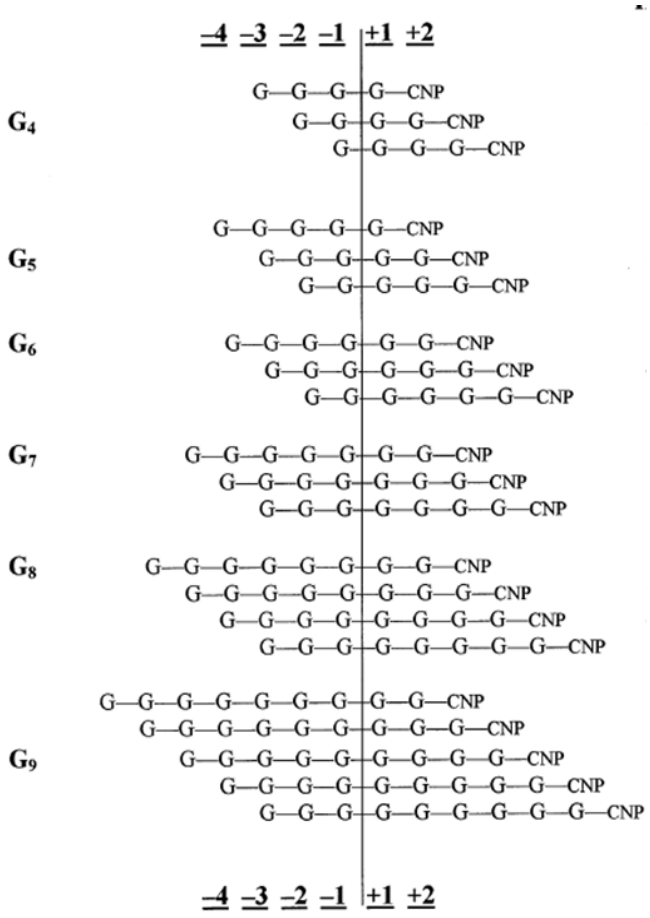


Figure 1.4. Subsites along the active site cleft and representation of polysaccharide cleavage. Subsites (-4 through +2) are represented at the top and bottom. Active site is represented by the vertical line situated between -1 and +1. $G_{4\text{-}9}$ signifies the number of glycosyl residues. Taken from Kandra and Gyemant, 2000. (G, glucose)

1. 4 Regulation of Amylase Gene Expression

A number of studies have shown the expression of amylase is highly tissue specific. Salivary *AMY1* is only expressed in the acinar and intercalated duct cells of the parotid gland (Kraus and Mestecky, 1971). Pancreatic *AMY2* is expressed in the pancreas (Samuelson et al., 1988). Both genes have been shown to be expressed in various cancer cells. Hyperamylasemia associated with tumors usually occurs when the tumors express *AMY1* (Koyama et al., 2001; Seyama et al., 1994). The difference in tissue specificity drove further investigations into the loci of both *AMY1* and *AMY2*. It was quickly shown that both genes are located in the p22-p21 region of chromosome 1 through a number of linkage studies and hybrid cell lines (Hill et al., 1972; Merritt et al., 1973; Tricoli and Shows, 1984). The salivary and pancreatic amylase genes showed that the salivary gene is 10kb long with 11 exons and 10 introns (Nishide et al., 1986). The gene *AMY2A* is 8.4kb long, with 10 exons (Horii et al., 1987). The gene complex is shown in (Figure 1.6). The differences in size is the fact that *AMY1* has an extra exon that is not translated (Horii et al., 1987). A TATA box was also identified for both genes. *AMY2*'s TATA box lies 29-23 bases upstream from its first exon, while that of *AMY1* lies 34-28 base pairs upstream of its untranslated exon (Horii et al., 1987). This divergence between *AMY1* and *AMY2* was not enough to explain the difference in tissue specificity, especially since *AMY2* shared considerable sequence homology at the same location as

AMY1's untranslated exon. More in-depth study of the promoter and intergenic regions would be needed to explain the high tissue specificity.

A ribonuclease protection assay was done to better determine the expression patterns of the amylase genes gave the first insight into the promoter regions of the amylase genes. The promoter region of the amylase genes is depicted in (Figure 1.5). It was determined that the liver expresses very low levels of the *AMY2B* gene product and that this specificity was due to an upstream splice site that includes part of the sequence that corresponds to the untranslated region of *AMY1*'s untranslated exon (Samuelson et al., 1988). Additional, unexpected products from the liver study revealed the presence of a 3' untranslated region of a human γ-actin pseudogene located 5' of all five amylase genes. This γ-actin pseudogene stopped 200bp upstream of the first exon in all amylase genes and continued at least 1.4kb upstream of each gene (Samuelson et al., 1990). While all the amylase genes were flanked by an untranslated region of γ-actin, only *AMY2B* included part of the actual coding sequence of the ancestral gene (Samuelson et al., 1988). Furthermore, this pseudogene lays upstream of *AMY1*'s first exon, comprising the promoter region and untranslated exon (Samuelson et al., 1988). Additional study of the salivary amylase gene showed that the γ-actin pseudogene extended not just into that first exon, but also into the first intron as well (Emi et al., 1988). This γ-actin pseudogene appeared to occur before the retroviral insert and, because it was present in Old World monkeys as well human amylase loci, it is believed to have

been present prior to the split of apes from Old World monkeys (Samuelson et al., 1990).

It was also shown that *AMY1A, AMY1B, AMY1C,* and *AMY2A* exhibited an endogenous retroviral insert (Samuelson et al., 1988). Originally believed to be just a gag-related sequence and long terminal repeat (LTR) located upstream of the genes, the retroviral insert was later shown to not only flank both sides of each amylase salivary gene, but that an *env*-related sequence was also present (Samuelson et al., 1990). This insert was believed to have occurred later in the evolution of the amylase locus, after the original divergence of human salivary and pancreatic amylase genes (Meisler and Ting, 1993). A transgenic mouse study using various portions of the 5' region of *AMY1C* linked to a reporter gene showed that the only part of the sequence upstream of *AMY1* needed to convey parotid tissue specificity was a region spanning 10kb upstream of the gene to 826bp upstream of *AMY1C* (Ting et al., 1992). Further investigation of this region showed that it was comprised entirely of the retroviral insert and the γ-actin pseudogene inserts (Ting et al., 1992). It was speculated that the recent retroviral insert, and the subsequent juxtaposition of the retroviral insert with the γ-actin pseudogene, was what led to a working promoter, since neither insert is actually capable of acting alone to convey specificity. This was supported by the fact that while the salivary genes have complete copies of retroviral-like elements, *AMY2A* only has the retroviral LTR sequence. There is no evidence

of retroviral elements flanking either side of the *AMY2B* gene (Meisler and Ting, 1993).

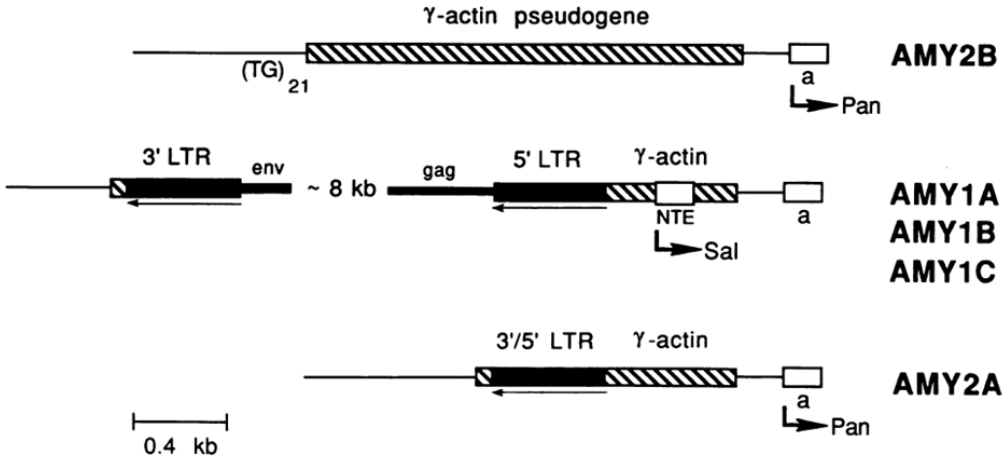


Figure 1.5. The 5' region of the actively transcribed amylase genes. The 5' region of each amylase gene is shown (the pseudogene *AMYP1* is not included). The white boxes indicate exons. The γ–actin pseudogene inserts are represented by striped boxes. Solid black boxes represent the retroviral insert. Arrows below the LTR sequences represent the orientation of the insert. *Pan* and *Sal* indicate start sites for pancreatic and salivary amylase gene transcription, respectively. Taken from Samuelson et al., 1990).

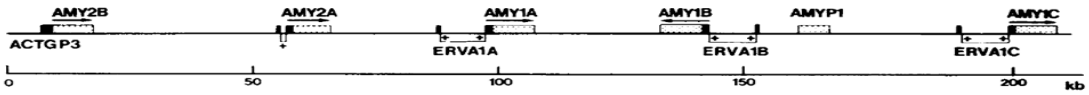


Figure 1.6. The human amylase gene cluster. A schematic representation of the human amylase gene cluster. The arrows indicate orientation of the gene. Black boxes indicate γ-actin pseudogene inserts. ERVA is the retroviral insert. Plus signs indicate retroviral LTRs. Taken from Samuelson et al., 1990.

1. 5 History of Copy Number Variation

Copy number variation is defined as a heritable duplication or deletion event that involves DNA greater than 1kb in size (Freeman et al., 2006). Most of the initial studies noting copy number variation occurred with cytogenetic observations. For example, multiple individuals with various degrees of mental retardation were shown to have chromosomal abnormalities by cytogenic techniques where chromosomal bands were either missing or duplicated (Jacobs et al., 1978). It was also believed to occur in regions where repeat sequences were plentiful (Freeman et al., 2006).

Since the initial reports of the phenomenon, variation in copies of a genomic region, especially genes, were believed to be associated only with genetic disorders or diseases (Ji et al., 2000). More recent studies have demonstrated, however, that this is not always the case, showing that copy number variation can occur in healthy individuals. High levels of sequence identity between at least 20 genes for olfactory receptors supported the idea that most of the receptors are subject to copy number variation (Trask et al., 1998). Investigation into chromosomal band 8p23.1 showed that region, which has been involved in numerous rearrangements, contains antimicrobial β-defensin genes that range in copy from 2 to 12 (Hollox et al., 2003). Studies of the complement

C4 has shown that, in a diploid genome, the gene on chromosome 6 can vary in number from 2 to 8 copies as well (Chung et al., 2002).

Reports of human amylase copy number variation were among the very first to document copy number variation among healthy individuals. However, such common CNV phenomenon attracted little attention in the field of human genetics until the past five years as the advent of comparative genomic hybridization experiments using microarrays reveal that CNV may account for 70% of genetic variations among different human subjects (Girirajan et al., 2011). It is now widely accepted that normal and healthy individuals can exhibit extensive copy number variation, leading to acceptance of the idea that copy number variation is responsible for a greater portion of variation between individuals than SNPs (Perry et al., 2007).

1.6 Initial Studies of Amylase Gene Copy Number Variation

Initial studies of copy number variation at the amylase locus began as studies of mainly salivary amylase protein isozymes. The first report of amylase variation described six different amylase protein patterns, observed in agar electrophoresis, which were inherited in a codominant manner (Kamaryt and Laxova, 1966) (Figure 1.7). Subsequent reports demonstrated a similar phenomenon in *Mus musculus* (Sick and Nielsen, 1964) and *Drosophila melanogaster* saliva (Bahn, 1967).

When the co-dominant alleles theory of an amylase gene in a diploid genome failed to explain some of the more complex patterns obtained from other electrophoresis techniques, two opposing models were proposed to explain why salivary amylase phenotypes are so varied. The two locus model, proposed originally in 1969 (Boettcher and La Lande, 1969) and again in 1971 (Ward et al., 1971), stated that the differences in fraction intensity for individuals with the same pattern, as well as variation in the patterns themselves, indicated that more than one locus was coding for the isozymes.

The one locus-multiple allele model, proposed in 1973, stated that amylase isozymes were not due to multiple loci but rather a mixture of post-translational modifications (such as glycosylation, deamidation, etc.) occurring before and after secretion of the enzyme from the parotid glands (Karn et al., 1973). Another report showed that salivary amylase could exist in glycosylated and non-glycosylated forms (Kauffman et al., 1973). The one locus-multiple allele model was supported by the finding that some fractions, but not all, were lost when amylase was stored in various conditions. One pattern could mimic another observed pattern once it was stored at different temperatures. In addition, there were a number of immunological studies on the salivary and pancreatic isozymes, all of which showed that both shared immunological identity (Karn et al., 1974; Ogita, 1966). Following the introduction of this model the research involving amylase variants centered more on the biochemistry and

enzymatic activity of the salivary and pancreatic isozymes rather than the genetics behind them.

Both models were revisited in 1982, following inheritance studies for a rare, unexplainable amylase variant demonstrated by the proposers of the one-locus-multiple allele model (Pronk et al., 1982). Amylase was extracted from the saliva of individuals from several families and the protein subjected to isoelectric focusing to look for patterns. One of the findings from this study was a “1-2-3 phenotype”, where three fractions appeared in the same individual (Figure). This indicated that in these family members, there were three genes encoding salivary amylase. The observation of varying gene product intensities made in earlier reports was also shown. These variations remained after samples were treated to ensure all possible isozymes formed, demonstrating that the intensities from the initial phenotypes were due entirely to separate gene products. It was speculated that the ancient salivary gene was duplicated and then went through multiple mutations to give rise to three separate genes. From familial studies, it was shown that multiple loci could be inherited as haplotypes. Another study looking at two isolated clones from a human lambda library showed further evidence of duplication at the locus when it was discovered that they both encoded pancreatic amylase genes but differed enough in their sequences to be considered separate genes (Groot et al., 1988).

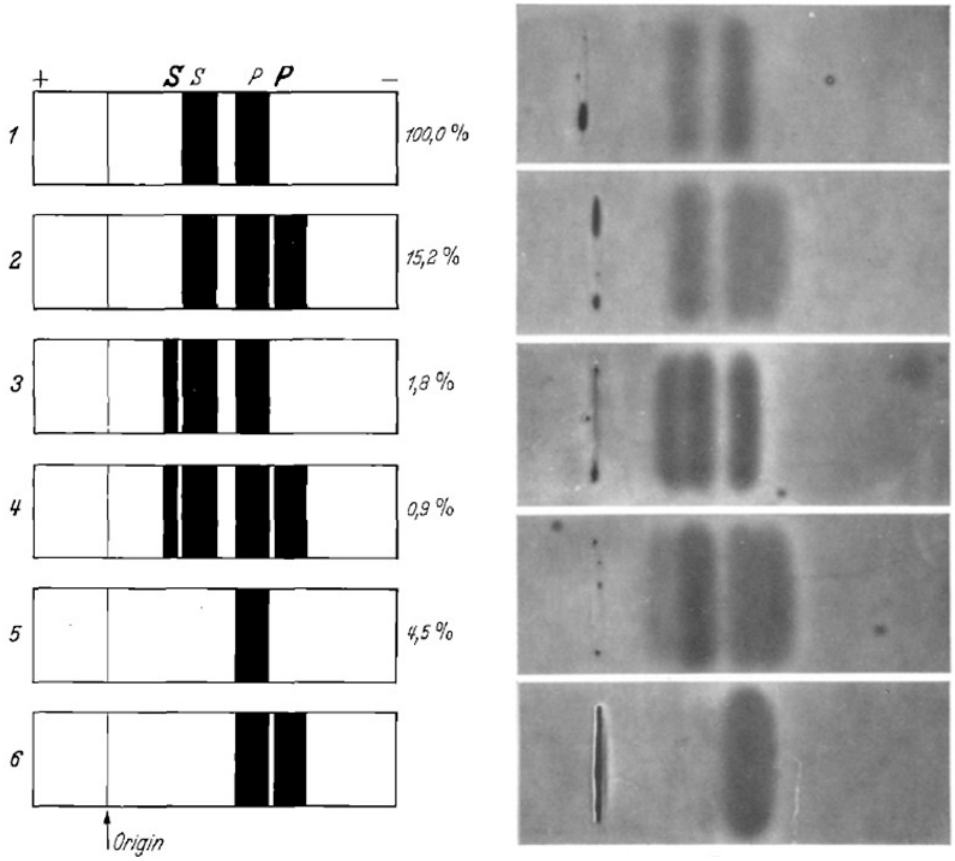


Figure 1.7. Initial amylase protein patterns observed in humans.

Schematic representation (left) and examples of observed amylase protein patterns. S signifies salivary fractions and P signifies pancreatic fractions. Taken from Kamaryt and Laxova, 1966.

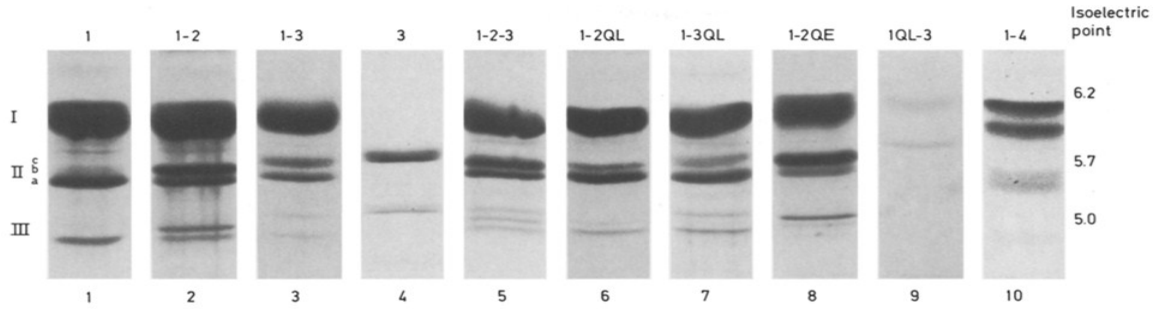


Figure 1.8. Human salivary protein patterns obtained from isoelectricfocusing. Phenotype 1-2-3 presents in individual 5. The three gene products were visible in Fraction 3. Taken from Pronk et al., 1982.

1. 7 Copy Number Variation of Human Amylase Gene Cluster

A few years later, three reports came out in quick succession showing definitive proof of copy number variation of the amylase genes. The first used a combination of chromosome walking via cosmid clones to determine the total copy number of salivary and pancreatic amylase genes (Gumucio et al., 1988). A total of seven genes were detected using this method – two pancreatic genes (labeled *AMY2A* and *AMY2B*), three salivary genes (*AMY1A*, *AMY1B*, *AMY1C*), and two pseudogenes (*AMYP1* and *AMYP2*). The salivary genes were reported as identical in coding sequence but differing in position within the cluster. It was also shown that *AMY1B* was oriented in the opposite direction from the other amylase genes. When hybridized to a mixture of salivary and pancreatic amylase cDNAs, both pseudogenes were shown to be truncated and missing the first two exons. The second report on amylase copy number variation used a ribonuclease protection assay to determine the number of functional genes present at the locus (Samuelson et al., 1988). This report confirmed that five genes – two pancreatic and three salivary – were responsible for all amylase protein in the human body. The reported two pseudogenes were not confirmed.

The last report to describe copy number variation of amylase came from studying the individuals with the 1-2-3 phenotype described in the Pronk paper (Groot et al., 1989). Also employing clones isolated from a genomic DNA library, the authors found only six genes, as opposed to seven, at the amylase

locus that could be inherited as a haplotype – two pancreatic, three salivary genes, and only one pseudogene. *AMYP2* was considered a cloning artifact, rather than a true gene. Unlike in previous reports, the layout of the locus was determined to be in the order of *AMY2B-AMY2A-AMY1A-AMY1B-AMYP1-AMY1C*. To date, this is the adopted layout of the AMYLASE locus, as is listed in NCBI's reference of the human genome (http://www.ncbi.nlm.nih.gov/projects/mapview/maps.cgi?taxid=9606&chr=1).

Investigating the amylase locus in other familial individuals showed that while only one copy of *AMY2A*, *AMY2B*, and *AMY1C* is present per haplotype; the rest of the genes can vary from one copy to three in a given haplotype. The shortest haplotype reported consisted of one copy each of *AMY2A*, *AMY2B*, and *AMY1C* while the longest also included two copies of *AMYP1*, four copies of *AMY1A*, and one copy of *AMY1B*. Based on this observation, an equation was given to describe future predicted haplotypes: *AMY2B-AMY2A-*(*AMY1A-AMY1B-AMYP1*)$_n$*-AMY1C* (Groot et al., 1989). A model for the evolution of the modern amylase family was proposed. The initial step was the duplication of an ancient amylase gene. Following this rare event, Groot speculated that further duplication could occur through unequal homologous recombinations between various haplotypes. This model and the equation were further detailed in following papers a few years later (Groot et al., 1991; Groot et al., 1990).

In the earlier report, a combination of sequence analysis and hybridization studies was used to derive a more detailed model of evolution than what was

previously described. Following the initial duplications of the ancestral pancreatic gene (believed to be *AMY2B*) to create the smallest haplotype observed, a series of homologous but unequal recombinations led to the extensive duplications of amylase genes. Deletion of part of a pancreatic gene similar to *AMY2A* led to the truncated *AMYP1*. A combination duplication-inversion event saw the creation of *AMY1B*. These events, therefore, led to the creation of the modern amylase cluster with all genes present in one copy. Finally, additional unequal and homologous crossovers between either *AMY2A/P1* or *AMY1A/B* were hypothesized to lead to further variation at the amylase locus (Groot et al., 1991).

In the last report from Groot's group, a combination of segregation analysis and Southern blot analyses using various restriction enzyme digests were conducted on genomic DNA of both families and unrelated individuals in order to identify more haplotypes than those previously reported. The first of the digests, done with the *Sst*I enzyme, displayed a fragment pattern predicted based on the original equation predictions. There was only one instance of a fragment of unexpected size obtained from this part of the study, which the authors described as a restriction fragment length polymorphism (RFLP) of *AMY2B*. A number of other restriction enzyme digests were done that showed no divergence from expected patterns. Two other digests revealed fragments of unexpected size – a *Hind*III digest and a *Taq*I digest. It was noted with the HindIII digest that the *AMY2A* and *AMY2B* fragments did not always display

equal intensities. It was suggested that *AMY2A* may be partially lost in some haplotypes due to recombination events that create “chimeric genes”. No explanation was given for the changes in *AMY2B* fragment intensity.

In the *Taq*I digests, there was a repeated loss of the fragment designated *AMY1C*. This did not agree with the original equation and model presented, as this gene was predicted to be always present in a haplotype. The authors resolved this by stating that all fragments that were associated with *AMY1C* were also associating with *AMY1B* and *AMY1A*. No further segregation analysis or cloning was done to determine if this was indeed the case.

At the conclusion of their work, the authors determined that their theory regarding the evolution of the modern amylase haplotypes was correct. The body of work done by the Groot group concerning copy number variation at the amylase locus was upheld by data from the Human Genome Sequencing project, which is published on NCBI, and is generally accepted as the model of amylase variation to this day. Other than a few RFLP studies done using restriction enzyme digests and gel electrophoresis (Ishizaki et al., 1985; Tsuchida and Ikemoto, 1989), studies of amylase again moved in the direction of protein studies, particularly protein structure and biochemical assays.

1. 8 Human Amylase Gene Copy Number Variation and Phenotypic Correlations

While the body of work on amylase gene copy number variation focused on the phenomenon itself and the mechanism behind it, nothing was proposed regarding the evolutionary pressures driving it until well after the start of the new millennium. Variation in copy number of genes has been shown in a number of studies to result in varying doses of gene product in individuals (Freeman et al., 2006). Since the primary substrate of amylase is starch, it would follow that the availability of starch would drive the need for a higher concentration of salivary amylase. A study aimed to determine if both hypotheses held true for the amylase locus (Perry et al., 2007).

In this study, populations were studied based on a history of starch consumption. The “high starch” populations studied included two agricultural groups and a hunter-gatherer group whose primary source of nutrition included highly starchy foodstuffs. The “low starch” populations included hunter-gatherer groups that traditionally consumed more protein and vegetables than starch and a pastoralist group. Because these groups were either traditionally high or low starch populations, the variation in copy number of salivary amylase observed could be tracked through the population. Comparison of salivary *AMY* gene copy between all populations via real time qPCR revealed that low starch populations have a lower median copy number than high starch populations. This part of the study also showed that not only was the median copy number for amylase in high

starch populations higher than in the low starch, but that the salivary locus was much more varied, ranging from two to as many as 15 copies in a diploid genome.

A high resolution fiber-FISH experiment both validated the qPCR results and showed that the variation in copy number did not occur in chimpanzee, which was used as a reference. Unlike humans, chimpanzee salivary copy number was consistently two per diploid genome. Initial divergence from chimpanzees at the locus was postulated to have occurred relatively early in human evolution, as tubers rich in starch were believed to be a prominent source of nutrition to hominids pre-*Homo erectus*. The model proposed for the reason for copy number variation at the amylase locus was one of directional or positive selection for higher copy number of salivary amylase genes in at least the high starch populations, thus providing an explanation for the driving force behind CNV at the amylase locus.

While most genes that are subject to CNV show a trend of higher protein production with increased gene copy, it was uncertain how closely amylase genes and proteins followed the same trend. For example, it was shown that salivary amylase production could increase in response to stress (Chatterton et al., 1996). The real time qPCR done by Perry's group, when combined with protein blots, showed that amylase protein levels were positively correlated with the copy number of salivary genes (Perry et al., 2007), indicating the greater copy number of salivary genes did lead to greater amount of salivary protein

(Figure). Another study showed that not only did increased copy number of *AMY1* correlate with a greater amount of protein, but also correlated with an increase in salivary enzyme activity and perceived starch viscosity (Mandel et al., 2010). Since creaminess of food and release of flavor from starch is affected by how efficiently broken down the starch becomes before swallowing, it would stand to reason that an increase in both amylase enzymatic activity and protein level would affect not only metabolism/digestion of starch, but how well an individual enjoys food (Mandel et al., 2010). Amylase gene copy number variation and variations in protein levels, therefore, could possibly play a role in obesity or other metabolic diseases.

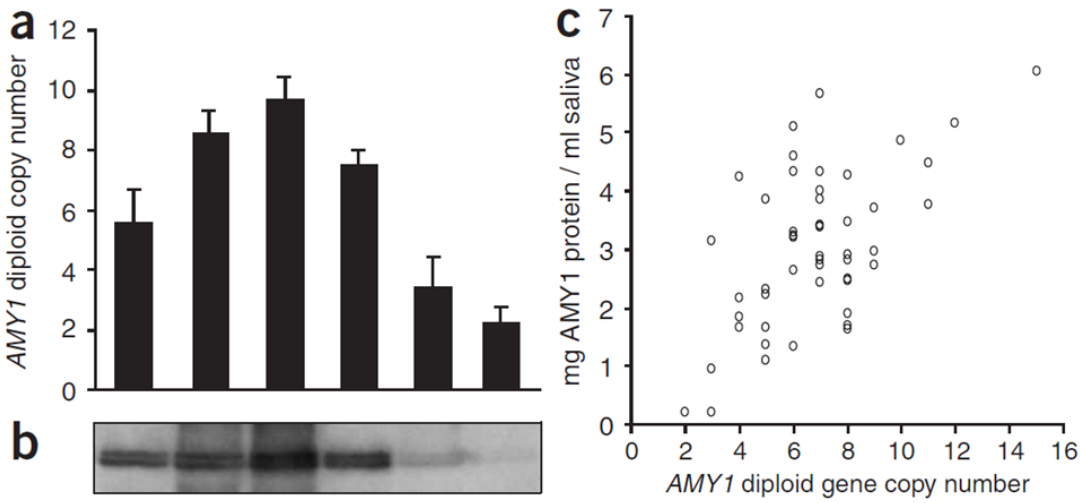


Figure 1.9. Human salivary amylase copy number and phenotype correlation. Diploid *AMY1* copy number for six individuals was determined by qPCR (a). A corresponding protein blot of salivary amylase for these individuals was also done (b). *AMY1* diploid copy number taken in conjunction with the actual concentration of amylase is saliva shows a positive correlation of protein concentration with *AMY1* copy number. Taken from Perry et al., 2007.

1. 9 Type 1 Diabetes: an Overview

Diabetes affects 26. 3 million people in United States with a 5. 3% rise in new cases each year (van Belle et al., 2011). One-tenth of the diabetic population belongs to Type 1 diabetes (T1D). Believed to be an autoimmune disease, T1D is the end result of immune-mediated destruction of the pancreatic β-cells in the endocrine pancreas's islets of Langerhans (Bluestone et al., 2010). T1D is believed to occur in two phases – insulitis, when leukocytes invade the islets, and diabetes, when most of the β-cells are destroyed (Mathis et al., 2001). β-cells are producers of insulin, a hormone that signals liver, adipose, and muscle cells to take up glucose circulating in the blood to be stored as glycogen. Once β-cells are destroyed, an individual with T1D is incapable of regulating blood glucose levels, leading to a variety of acute and chronic complications such as ketoacidosis, blindness, and kidney failure (Bluestone et al., 2010). It is currently believed that T1D is a culmination of genetic susceptibility and exposure to environmental factors that trigger an autoimmune response toward the endocrine pancreas.

There are currently multiple environmental factors that could act as the trigger that sets off autoimmunity and progression to T1D. Because of the seasonal nature of T1D (most patients are diagnosed about the same season each year), and a possible link with viral infections, it is speculated that these

infections may be a major event that can lead to disease onset (van Belle et al., 2011). While no one specific virus strain has been directly shown to be causative, enteroviruses as a group have been implicated by several studies because of their ability to infect β-cells, such as coxsackieviruses (van Belle et al., 2011). Rotaviruses have also been studied due to the fact that their proteins can mimic T-cell epitopes. Maintaining the natural flora of bacteria in the intestine has also been linked to some degree with T1D incidence (van Belle et al., 2011). The "hygiene hypothesis" – the idea that little or no exposure to infectious agents early in life can lead to greater risk of immunological disorders later – has been supported by reports of some viral infections conferring a level of protection against T1D in certain conditions (van Belle et al., 2011). Other, less obvious candidates for an environmental trigger, such as the albumin in cow's milk, gluten in wheat products, and a lack of vitamin D, have all been implicated in increasing risk of T1D development based on several studies of molecular mimicry, T-cell reactivity, and general observation of patients already diagnosed with the disease (van Belle et al., 2011).

Patients usually are admitted when they begin exhibiting symptoms associated with hyperglycemia such as excessive thirst, urination, and hunger (van Belle et al., 2011). Once in the hospital, they are usually then diagnosed based on the presence of hyperglycemia, partial-to-complete loss of C-peptide secretion, complete dependence on an external source of insulin, or a combination of the three (Rowe et al., 2011). At the time of diagnosis, most of

the β-cells (60-90% of total mass) have been destroyed, or at least are non-functional (van Belle et al., 2011). Upon administration of insulin, more than 60% of patients experience a “honeymoon phase”, during which the need for exogenous insulin is low to non-existent (van Belle et al., 2011). While the presence and length of honeymoon phase varies with the age of onset, all patients eventually “relapse”, showing renewed immune-mediated destruction of β-cells and ultimately complete loss of C-peptide secretion (van Belle et al., 2011). Once this occurs, patients are entirely reliant on regulated doses of insulin for the rest of their lives.

While the clinical symptoms of T1D have been well defined, the actual pathogenesis of the disease is not currently well understood, partly due to the fact that the initial stage of insulitis (when leukocytes actually begin invading the islets) can occur for years without symptoms (Mathis et al., 2001). Since the advent of excellent medical treatment, T1D is no longer an acutely fatal disease (Rowe et al., 2011). Despite this, there are a number of suggested models that strive to explain how the exposure to environmental triggers in a certain genetic background can lead to development of the disease. The most widely accepted model, known as the linear beta-cell decline hypothesis, states that an environmental trigger(s) leads to a linear loss of β-cell mass and development of diabetic symptoms (van Belle et al., 2011).

A second model depicts T1D as a relapse-remit disease, where cyclical disequilibrium between T_{regs} and effector T cells ultimately leads to a drop in β-

cell mass (van Belle et al., 2011). The fertile field hypothesis speculates that a viral infection creates a time window during which a number of mechanisms such as molecular mimicry could lead to autoreactive effector T cells. Eventually, the buildup of such T cells could lead to an autoimmune attack of β-cells and ultimately lead to T1D (van Belle et al., 2011). Until better methods of studying the insulitis and pre-insulitis stages become available, it will be difficult to determine which model, or combination of models, depicts the correct path of T1D pathogenesis.

1.10 Genetics of Type 1 Diabetes

A number of genetic studies looking into which genes and their variants cause a genetic predisposition to T1D has resulted in a large number of loci that could play a role in disease susceptibility. Most of the major known contributors to T1D have also been associated with other autoimmune diseases (van Belle et al., 2011). The genes that show the strongest association with T1D are those genes located within the HLA region (Pociot et al., 2010; van Belle et al., 2011). Of these genes, those within class II have been shown to have the greatest effect, with some haplotypes being highly protective and others conferring the highest risk. For example, the presence of the haplotype DRB1*1501-DQA1*0102-DQB1*0602 provides a huge level of protection against T1D (van Belle et al., 2011). The haplotype conveying the highest risk is DR3/4-DQ8

haplotype, which is present in 30-50% of the T1D population (Pociot et al., 2010; van Belle et al., 2011). Several HLA class I alleles have also been shown to carry risk for T1D, including HLA-B*39 and HLA-A*02 (van Belle et al., 2011). Because several protective HLA alleles have been shown to be dominant, it is speculated that the HLA region is involved more with protection against the disease, rather than predisposition for it (Bluestone et al., 2010).

Besides the HLA region, several other candidate genes have been identified whose variants can confer a higher risk for T1D. The insulin gene has VNTR regions that flank it. The class I alleles, which have shorter VNTRs, have been shown to increase the risk of T1D (Pociot et al., 2010; van Belle et al., 2011). This class is also associated with less protein and mRNA expression in the thymus, possibly leading to AIRE's inability to bind to the promoter and reduce tolerance (van Belle et al., 2011). The longer, class III alleles for insulin, on the other hand, convey dominant protection against T1D (Pociot et al., 2010).

A variant of the gene *CTLA-4*, A49G, also has been shown to increase the risk of T1D. *CTLA-4* encodes a receptor that aids in the inhibition of T-cell activation (Pociot et al., 2010). The A49G variant has been shown to lead to less surface expression of CTLA-4 protein levels in vitro, although it's possible that it also lowers mRNA levels (van Belle et al., 2011). Another gene, *PTPN22*, also has a variant (R620W) that increases the risk for T1D. This gene encodes another inhibitor of T-cell activation and the high-risk variant is believed to be a gain-of-function allele (Pociot et al., 2010). *IL2RA*, which encodes CD25

expression on naïve and memory T cells as well as activated monocytes, have noncoding variations that alter gene transcription, leading to increased risk of T1D (Pociot et al., 2010). Besides these candidate genes and the strong HLA haplotypes associations, GWA studies have uncovered even more regions of interest, implicating more than 41 areas within the human genome that could be associated with increased risk of T1D (Pociot et al., 2010). Most of these regions, however, have yet to be investigated.

1.11 Goals of This Study

There is a model that describes the pattern of duplication for the human amylase genes (Groot et al., 1989). There have been various restriction enzyme digests and Southern blot analyses done by this group but the results were varied, with some results supporting the model and other contradicting it. Southern blots with SstI, for example, exhibited patterns that were predicted by the authors according to their model. However, they noticed varying intensities for the *Taq*I fragment that corresponded to *AMY1C* that could not be explained by their model (Groot et al., 1991). There have been no population studies data to detail the pattern of inheritance for the amylase locus either. Therefore, the first goal of this study is to determine patterns of duplication for the amylase genes. A combination of Southern blot analyses using new and similar restriction

enzymes, along with the new technique of pulsed field electrophoresis, should allow us to better distinguish how the amylase genes duplicate.

Other than the obvious initial duplication of the pancreatic gene and the creation of the pseudogene P1, no further investigation into the CNV of *AMY2* has been reported. According to the model proposed by Groot, *AMY2* will not undergo further duplication/deletion events (Groot et al., 1989). An observation was made that *AMY2A* fragment in Southern blots can vary in intensity and the explanation was that, while it may be slightly truncated due to another recombination event, complete loss or duplication of *AMY2A* would not be possible (Groot et al., 1991). *AMY2B* has also not been reported to be subject to further CNV. A second goal for this study, therefore, is to determine first if *AMY2* CNV is possible and, if so, investigate the nature of *AMY2* CNV. This can be done via Southern blots using a number of restriction enzymes and by using a probe that can distinguish between *AMY2* and *AMY1*.

A study has been published showing that when serum from patients who presented both autoimmune pancreatitis and fulminant type 1 diabetes was incubated with a human pancreatic cDNA library, autoantibodies against AMY2A were detectable in 88% of cases (Endo et al., 2009). This study indicates that the human amylase genes, and their products, may play a possible unknown role in development of T1D. However, there have been no other studies reporting a connection between the two. The positive correlation between amylase CNV and both enzymatic activity and protein concentration in saliva also shows that

individuals can vary in their baseline expression of amylase protein (Mandel et al., 2010). Because no studies have been done to characterize amylase CNV in T1D patients, it is necessary to obtain initial data on amylase CNV patterns in this population. Thus, a third goal of this study is to determine amylase CNV in a T1D patient population. This can be done by identifying individuals with simple haplotypes and determining the genes present in those haplotypes. Determining CNV of *AMY1* and *AMY2* genes in the same patient population will also provide initial data with which more studies can be done to characterize correlations between amylase CNV and T1D.

CHAPTER 2

ELUCIDATION OF PATTERN OF VARIATION FOR THE AMYLASE LOCUS IN TYPE 1 DIABETES PATIENTS

2.1 Introduction

Copy number variation for the human amylase locus has been extensively reported, but very little work has been done to determine the exact pattern of variation. Previous studies have focused exclusively on human salivary amylase genes (*AMY1*), with no contemplation given to CNV of human pancreatic *AMY2.*

The first model to explain amylase variation was proposed by Peter Groot. According to his model, the amylase locus would see segmental duplication of specific amylase genes according to the equation *AMY2B-AMY2A-(AMY1A-AMY1B-AMYP1)$_n$-AMY1C* (Groot et al., 1989). A single individual homozygous for the simple, small haplotype (*AMY2B-AMY2A-AMY1C*)$_1$ was the basis for the model. According to this theory, only human salivary *AMY1* would differ in copy

number. None of the pancreatic amylase genes would be capable of undergoing copy number variation, nor would the final salivary amylase gene *AMY1C*. The method for variation associated with this model is one of unequal homologous crossovers of simple amylase haplotypes (such as the simple haplotype on which the model is based) that lead to greater, more complex amylase loci.

Although Groot's AMY-locus model is consistent with Reference sequence from data of the Human Genome Project, there is a building body of evidence suggesting that the pattern of variations for *AMY1* and *AMY2* is far more complex. Some of the evidence against the model was presented in subsequent reports from Groot. Various restriction enzyme digests of genomic DNA from unrelated individuals were conducted by Groot's group (Groot et al., 1991). Some of the digests, such as those using restriction enzymes *Bgl*II or *Sst*I, gave support to the model. Some restriction enzyme digests, however, gave rise to results that did not fit the model as Groot described. In the *Hind*III digests, 6. 3 and 5. 8 kb fragments representing *AMY2A* and *AMY2B*, respectively, varied in intensity between individuals. Variance in intensity for fragments representing these two genes were considered to be the product of partial deletion of these genes due to unequal crossover (Groot et al., 1991).

Reports from other researchers also suggest that Groot's model is not entirely accurate. Perry's study of causes for human amylase CNV showed variation of copy number of salivary genes that exceeds that suggested by Groot (Perry et al., 2007). Another report details inserts of a γ-actin pseudogene and

retroviral elements found in the intergenic regions of the human amylase locus (Samuelson et al., 1988). The presence of similar inserts coupled with highly similar genes suggest that it could be possible to see recombination between any of the amylase genes, including those originally postulated not to vary in copy number.

A report was published in 2004 that showed results for an investigation into common CNVs in healthy individuals (Iafrate et al., 2004). Using array CGH and confirming with both FISH and qPCR results, the authors sought to determine what large scale variation could be found in the human genome. The most common CNV described by the group came from the amylase locus. Unlike what Groot had suggested, this report found that the region where most variance occurred was that encompassing amylase genes *AMY2A* and *AMY1A* (Iafrate et al., 2004). The group also found that the variation in this region occurred at a larger scale than Groot predicted.

The strongest evidence against Groot's model comes from the alternative reference sequence present in NCBI's database, which showed an AMY haplotype consisting of three AMY genes: *AMY2B-AMYP1-AMY1C*. The presence of an Alternative Reference Sequence for the *AMY* locus without *AMY2A* shows that Groot's model is no longer accurate. However, since no further work has been done on the amylase locus to detail the pattern of variation, no current accurate model exists. In this study, we interrogate copy number of individual amylase genes and variation in size for the amylase locus

as a whole for a large cohort of recently diagnosed type I diabetes patients. Our goals were to determine the variation present in a T1D population to obtain evidence to propose a more accurate model for the pattern of variation at the human amylase locus.

2.2 Materials and Methods

2.2.1 Study Subject Recruitment

Informed consent was obtained from T1D patients and healthy subjects according to IRB-approved protocol. All T1D patients were recruited from the Endocrinology Clinic at Nationwide Children's Hospital in Columbus, OH.

2.2.2 Genomic DNA samples

Peripheral whole blood samples in EDTA-tubes were obtained from blood donors by venipuncture. To isolate DNA for Southern blot analysis, blood cells were lysed using QIAGENE's RBC lysis solution and PureGene's cell lysis solution. Samples were stored in 37^{o}C for a half hour. Once sample had cooled to room temperature, proteins were precipitated out using PureGene's protein precipitation solution. Samples were centrifuged for 10 min at 2000G. Ethanol precipitation is used to extract genomic DNA from supernatant. DNA was re-suspended in TE buffer and stored at 4^{o}C.

White blood cells (WBCs) and ploymorphonuclear neutrophils (PMNs) were separated from whole blood using histopaque layering. Samples were centrifuged for 30 min at 700G. WBCs were isolated and encased in 1% molten low gelling temperature (LGT) agarose plugs to use for pulsed field gel electrophoresis (PFGE). Plugs were placed at 4^{o}C for a half hour to harden.

Plugs were then incubated at 50^{o}C overnight with 85 υL proteinase K and NDS ($K_4[ON(SO_3)_2]_{2)}$ solution (pH = 9.0). All plugs were stored in NDS at 4^{o}C.

2.2.3 Amylase Specific Probes

All probes for DNA hybridization were designed and generated by PCR using the cDNA of the human amylase genes as template. Details of PCR primers are as follows.

Forward primer for AMY300: 5' – CGATGGCGCCAAATAAGGAACATGG -3'.

Reverse primer for AMY300: 5' – GATTCAGCATGAATTGCAATAAATGG -3'.

Forward primer for AMY345: 5' – CTGGAAAGGACACTGACAACTTCAAAGC -3'.

Reverse primer for AMY345: 5' – CCAACATTGTTACATCTAGTCACCATG -3'.

Forward primer for AMY543: 5' – GATATTGCTCTTGAATGTGAGGG -3'.

Reverse primer for AMY543: 5' – GCCACATGTGCTTGGAAGCATC -3'.

Template for AMY300 and AMY345 was *AMY1A* cDNA. *AMY2B* cDNA was used as the template for AMY543 development. PCR reactions were carried out using Epicentre's FailSafe Enzyme mix and premixes. 100 υg of DNA template and 150 υg of each primer was used for PCR reactions. Following completion of PCR reactions, DNA was obtained using phenol/chloroform extraction followed by alcohol precipitation. DNA was then run in a 1.0% LGT agarose gel at 54V for

4-5 hours to ensure proper amplification occurred during PCR reaction. Fragments were then excised from LGT gel and purified using QIAquick gel extraction kit according to company instructions. Probes were stored in TE buffer at 4^{o}C.

2.2.4 Restriction Enzymes

All restriction enzymes and enzyme buffers were purchased from New England Biolabs (NEB). Restriction digests were conducted using the suggested conditions from NEB. NEB3 buffer was used with the *Pst*I digest. All other digests used NEB4 buffer. All digest cocktails included a restriction enzyme, its suggested buffer, and 30x BSA.

2.2.5 *Taq*I, *Pvu*II/*Psh*AI, and *Pst*I Restriction Enzyme Digests of genomic DNA with Southern blot analysis

Approximately 6 μg of genomic DNA was used in each digest. *Taq*I digests were conducted at 65^{o}C. *Psh*AI/*Pvu*II digests were conducted first at 25^{o}C followed by an increase in temperature to 37^{o}C. *Pst*I digests were conducted at a constant temperature of 37^{o}C. All three digests were carried out overnight. For Southern blot analysis, 0.7% agarose gels were made with 1xTBE and 25 μL of ethidium bromide (EtBr, 0.05%). Gels were cast, cooled and loaded

with digested DNA samples. Gels were run at a constant voltage until DNA had run full length of the gel. Following gel run, gel was exposed to UV light and a picture was taken.

2.2.6 *Pme*I Restriction Enzyme Digests and Pulsed Field Gel Electrophoresis

Plugs were removed from NDS and washed several times, first with TE buffer followed by NEB4 buffer. For *Pme*I digests, a master mix of enzyme, digest buffer, and 30x BSA was added. Plugs were immersed in master mix. *Pme*I digests were conducted at a constant temperature of 37^{o}C for at least 3.5 hours. Plugs were then placed at 4^{o}C for at least 30 min to re-harden the plug. Pulsed field certified agarose was purchased from BioRad and used to make gels. 1.0% agarose gels were made with 0.5xTBE. Gels were cast, cooled, and loaded with digest plugs. All pulsed field gels were run using BioRad's CHEF Mapper XA system. Programs were set to resolve 70-700kb of genomic DNA fragments. Conditions are as follows:

Switch time: 9.65 sec to 1 min 8.65 sec

Gradient: 6.0 V/cm

Angle: 120^{o}

Temp: 14^{o}C

Ramp: linear

Run time: 44 hours

Following their run, gels were submerged in 0.5xTBE with 100 μL of 0.05% EtBr for 20 min. Gels were then exposed to UV light before a picture was taken.

2.2.7 Pressure Blot Transfer, UV Crosslinking, and Hybridization

All gels were first washed in 0. 5M NaCl, 0. 2M HCl buffer for 25 min to denature double-stranded genomic DNA. Second wash was 1.5M NaCl, 0.5M NaOH buffer for 30 min to neutralize acidic buffer. Third buffer was 1.5M NaCl, 0.5M TRIS-HCl buffer for 30 min. Final wash was 10xSSC for 10 min. A nylon membrane was labeled and soaked in 10xSSC with gel and two pieces of Whatman chromatography paper. DNA was transferred from gels to nylon membrane using pressure blotting. Transfer was conducted at a pressure between 70 and 80 mmHg for at least 6 hours for genomic DNA fragment between 0.5kb to 20kb resolved by regular agarose gel electrophoresis, at least 24 hours for DNA fragments between 20 kb to 1000 kb resolved by PFGE. Membranes were then UV cross-linked twice. Prior to hybridization, membranes were washed in 0.1% SDS-SET ($CH_3(CH_2)_{11}OSO_3Na$) buffer at a constant 42^{o}C for a half hour. Hybridization was conducted using one of the prepared

radioactive ^{32}P probes (AMY300, AMY345, or AMY543). To view results, X-ray film was exposed to hybridized membranes at a constant temperature of -80^{o}C in a cassette with intensifying screens for 2-14 days.

2.2.8 Bioinformatic Analysis of the Human Amylase Locus

Dot plots were performed using the dottup program from Mobyle @Pasteur (EMBOSS 6.3.1). For comparison of the entire amylase locus, the FASTA sequence of the reference human amylase locus as published on NCBI was used (Build 37.3, Chromosome 1: 104,096,322-104,303,310). Word-length for dot-plot was set at 25 nucleotides. To compare individual amylase genes and cDNA, NCBI's BLAST program was used. Amylase gene and cDNA sequences were obtained from NCBI's reference genome.

2.2.9 Additional Declarations

I would grateful to Dr. Yee Ling Wu, who was responsible for pilot studies on the human amylase locus, preparation of *AMY1* and *AMY2* plasmid clones, and generation of AMY-specific probes. I am also grateful to her help in recruitment of T1D patients and processing of blood samples. I would like to thank senior research associate Bi Zhou for her help in processing blood samples for genomic DNA, for performing all regular Southern blot analyses, and

for conducting all hybridizations. I am grateful for Zhenyu Yang's help in recruitment of T1D patients and preparation of genomic DNA. A special thanks goes to Dr. Chack Yung Yu for aiding me in the bioinformatic analysis of the *AMY* locus and coordination of this project. I am grateful for the efforts of Dr. Suzanne Kingery, Dr. Robert Hoffman, Dr. Sasigarn Bowden, Dr. John Germak, and Dr. William Zipf for their efforts in diagnosing T1D patients and recruiting them to our study. Finally, I am eternally grateful and indebted to the patients and their family members that agreed to participate in our study.

2.3 Results

2.3.1 Dot plot of the human amylase locus

A dot plot of the entire human amylase locus shows a high degree of shared similarity (Figure 2.1A). This similarity is not just observed between individual genes, but also intergenic regions as well. Pancreatic *AMY2B* deviates the most from the other genes. It is not only considerably longer than the other amylase genes, but also shows noticeable breaks in shared sequence that shares no sequence similarity with the rest of the locus. This break corresponds to a 1kb region within *AMY2B* that runs from midway through the first intron up to the second intron. This sequence, unique to *AMY2B* for the amylase locus, is a β-actin enteric pseudogene insert within intron 2.

Several blocks of similar sequence exist within the human amylase locus (Figure 2.1, panel A). The first, measuring 64kb in size, encompasses only *AMY2B*. Within this block are smaller stretches of similar sequence varying in size from 7 kb to 29kb. The 29kb segment, located 3' to *AMY2B*, is homologous to the 5' region of both *AMY1A* and *AMY1C*. The second block measures 50kb in size and encompasses the genes *AMY2A* and *AMY1A*. As with the *AMY2B* block, there exist several smaller segments of similar sequence present within the block. The largest is 26kb, followed by 18kb, 10kb, and 7kb segments. The 26kb segment, present within this block immediately 3' to *AMY2A*, is homologous to a region 5' of *AMY2A*.

AMY1B is in the opposition orientation within the locus. In order to determine the sequence similarity for this gene and surrounding intergenic region, we compared sequences of the amylase locus in the forward orientation with the locus in the reverse (or opposite) orientation (Figure 2.1B). When we did this, we found a 44kb block that encompassed *AMY1B* and corresponding extragenic regions. Again, smaller segments ranging from 7 kb to 30 kb exist within the larger block. The largest segments of 30kb correspond to shared sequence with the other amylase salivary genes and extragenic regions.

The final block, encompassing genes *AMYP1* and *AMY1C*, measures slightly smaller than the block for *AMY2A-AMY1A*. Measuring 48kb in size, the *AMYP1-AMY1C* block is identical to that for the *AMY2A-AMY1A* block, including the presence of the smaller sequences of similarity.

2.3.2 Restriction map analysis of the human amylase genes and probe development

To investigate the pattern of variation for the blocks of segmental duplication in amylase locus, we examined restriction maps for all amylase genes to identify diagnostic markers. The human Amylase locus spans from nucleotide104,097,322 that corresponds to the 5' end of *AMY2B*, to nucleotide 104301310 that corresponds to the 3' end of *AMY1C*. It was determined that no *Pme*I cut sites were present within the amylase locus itself. The proximal 5'

*Pme*I site is 118 kb from *AMY2B*; the distal 3' *Pme*I is 33 kb from *AMY1C.* Thus *Pme*I restriction fragment length polymorphism an ideal assay to investigate the physical size variations of haplotypes for the entire *AMY* locus . The sizes of *Pme*I fragments would correlate with the total number of genes present at the locus for each haplotype. In order to detect the fragments, we designed a 300 bp *AMY* probe cDNA, labeled AMY300. AMY300 encompasses a common region at the 3' end of all human amylase genes.

While *Pme*I would provide information on haplotype sizes, it would not be informative in deciphering for the composition of individual amylase genes present in the locus. We, therefore, developed a number of assays that rely on different restriction fragment lengths and their relative dosages (i.e., their band intensities on X-ray films) to distinguish between amylase genes and their relative copy number. The digests and expected fragment lengths are shown in (Table 2.1). With AMY300 as a probe, a *Taq*I digest would provide the most information on individual gene copy number as it would generate fragments unique to each gene. The results of a *Taq*I digest could be corroborated from a second assay that utilizes a double digest of *Pvu*II and *Psh*AI. With these two assays, we would be capable of differentiating the copy number for all amylase genes with the exception of *AMY2A* and *AMYP1*, which were indistinguishable.

AMY2A and *AMYP1* share 99% sequence similarity, with *AMYP1* missing a 2kb sequence at the 5' end and therefore not functional. We, therefore, designed two new probes that would allow for separation of these two genes.

The first, AMY345, is a 345bp probe that extends from the first exon of *AMY2A* to its second exon. Because it hybridizes to part of the region lacking in *AMYP1*, any assay using this probe would exclude *AMYP1* from analysis. The second probe, designated AMY543, extends from exon 1 of *AMY2A* to exon 4. AMY543 is a capable of partially hybridizing to *AMYP1*, ensuring its detection in assays.

Using a combination of a double digest with AMY345 probe would allow us to distinguish between *AMY2A* and *AMYP1* but the size of the fragment generated for *AMY2B* would ensure its exclusion from the assay. It therefore became necessary to develop one more digest that would allow us to see all genes. We chose to use a *Pst*I digest, which would not interfere with probe hybridization and would create a small enough fragment representing *AMY2B* to ensure its detection.

2.3.3 *Taq*I digest results

The *Taq*I RFLPs yielded insights into the copy number of specific amylase genes. However, we have also observed a number of unexpected genotypes. This resulted from novel RFLPs present in the population that were not predicted based on the restriction map of NCBI's amylase locus.

The two most common genotypes were the expected five fragment pattern, designated ABCDE for simplicity (Figure 2.2, ER68P and ER70F) and a three fragment pattern, designated BEH (Figure 2.2, ER71M). In both, the

fragment specific to *AMY2A/AMYP1* (fragment B) and *AMY2B* (fragment E) are present. The H fragment, which was not predicted by restriction map analysis, measures 3.0kb in size. In both patterns, fragments (including fragment H) vary in intensity. This indicates that different individuals vary in the composition and the copy number of individual amylase genes. The ABCDE and BEH patterns account for almost half of all individuals. A combination of both patterns (ABCDEH) has also been observed fairly regularly.

A myriad of variations on ABCDE and BEH are observed in subjects. Variations include ABEH, BCEH, BCDEH, ABDEH, etc. In all variants of the main phenotypes, individual fragments vary in intensity. The fragments for pancreatic amylase (B and E) are also present in all variations on the main phenotypes.

Another unpredicted RFLP appeared fairly often in the subject population. Measuring 4.2 kb in size, this fragment (designated fragment F) was not identified to associate with a specific AMY gene. When its presence was observed, the fragment for *AMY2B*, fragment E, was less intense. Examples of phenotypes with fragment F include ABCDEF, ABEFH, BEFH, etc. This suggests that the F-fragment could be a polymorphic variant of *AMY2B*.

One last novel RFLP was observed in our subject population. Designated fragment G, it measured 3.7kb in size. Due its low prevalence in the population,

we have not determined which amylase gene is represented by this fragment at this time.

2.3.4 *Pvu*II/*Psh*AI digest results

Unlike as in the *Taq*I phenotypes, no novel RFLPs were observed in any double digest results (Figure 2.3). All three expected fragments were present in all individuals studied. The fragments representing *AMY1A/AMY1C* (6.5 kb) and *AMY2A/AMYP1/AMY1B* (5.8 kb) varied greatly in intensity among individuals. The fragment representing *AMY2B* also varied in intensity, though not to the extent observed in the other two fragments.

2.3.5 *Pst*I digest results

Like the *Pvu*II/*Psh*AI digests, the *Pst*I digests exhibited novel RFLPs that existed in our subject population (Figure 2.4). All expected fragments appeared in all individuals, with the exception of the fragment for *AMYP1* and, in one instance, *AMY2A*. As with *Taq*I and *Pvu*II/*Psh*AI double digests, the fragments vary in intensity between individuals and the fragment specific to *AMY2B* is always present.

There are two novel *Pst*I RFLPs that are prevalent in our subject population, regardless of which probe is used (AMY543 or AMY345). The first is

a 20kb fragment, present for SK3P as seen in Figure 2.4's top panel. When AMY543 is used, the 20kb is always present when the 2.1kb *AMYP1* fragment is also present. When AMY345 is used, the 20kb is not always present. The 20kb fragment is never present in *Pst*I-AMY345 for an individual if it isn't present as well in the *Pst*I-AMY543 digest.

A second novel *Pst*I RFLP, as seen in ER41M's genotype, is also fairly prevalent. If the fragment is present in *Pst*I-AMY543 digest for an individual, it presents in the *Pst*I-AMY345 digest (and vice versa). This fragment, measuring 7.4kb in size, is not associated with any other fragment. Its presence does not result in lessening intensity for another fragment, or in another fragment's absence. It does vary in intensity when present, indicating that it is associated with the amylase locus, rather than nonspecific hybridization. However, we have not determined which amylase gene(s) is being represented here.

2.3.6 *Pme*I digest results

Because there are no *Pme*I cleavage sites within the amylase locus, all individuals either displayed a two fragment or one fragment pattern. An individual with two fragments, such as MS639 in Figure 2.5, is carrying two haplotypes of different sizes while individuals like TD86P with one fragment is carrying two haplotypes of the same size. The NCBI reference sequence for the amylase locus lists one of each gene (*AMY2B-AMY2A-AMY1A-AMY1B-AMYP1-*

AMY1C). The predicted size of the *Pme*I fragment for such locus with one copy of all amylase genes is 354kb. In our population of 260 T1D patients, the 354kb fragment has an allele frequency of 0.46. However, *Pme*I fragment sizes ranging from 260kb to 740kb were detectable in our study population (Figure 2.5).

2.3.7 Segmental duplications of the *AMY* locus

Despite extensive variation in *Pme*I fragment sizes, we found a noticeable pattern when samples were organized in order of increasing fragment size (Figure 2.6). Fragments appear to increase in size mainly in ~20kb and ~30kb increments, although minor increases of 10kb appear to exist as well. Major segmental increases may correlate with the 44-50kb blocks of sequence similarity observed during dot plot analysis. At this time, resolution for pulsed field prevents us from definitively ascertaining which blocks are being duplicated during this segmental duplication. The smallest fragment observed in our population measures 260kb in size. The next largest fragment commonly seen measured 280 kb in size. The 354kb fragment estimated from NCBI's reference genome contains just one copy of each amylase gene and has all three blocks representing *AMY2A/AMY1A, AMY1B,* and *AMYP1/AMY1C* present. The next fragment observed measures 410kb. The human amylase locus continues to undergo segmental duplication, reaching fragment sizes of 740 kb .

2.3.8 TD80P – Simplest confirmed haplotype

All digests for TD80P are presented in Figure 2.7. As shown in panel B, TD80P presented just one *Pme*I fragment measuring 280kb in size. Having one long-monomodular fragment indicated the presence of two haplotypes of equal size that contained very few amylase genes. *Taq*I digest (panel D) indicates that TD80P has a haplotype pattern of ABE. ABE pattern indicates the presence of *AMY1C, AMY2A/AMYP1,* and *AMY2B.* Each fragment is present in equal intensities, suggesting that each fragment contains the same number of human amylase genes. *Taq*I digest results were supported by the double digest (panel E). TD80P had the expected three fragment pattern for the double digest and, again, all fragments were of equal intensity. At the very least, TD80P would be expected to have one copy of *AMY1C* and *AMY2B* per haplotype (two copies each for the diploid individual), but it is unclear from these two digests whether TD80P contains copies of *AMY2A, AMYP1,* or a combination of the two. To determine this, we looked to the *Pst*I digest using AMY543 for detection. Results from *Pst*I are displayed in panels C. It shows three fragments only – 8.2 kb representing *AMY1*, 15.2 kb representing *AMY2B* and 5.4 kb representing *AMY2A.* No fragment for *AMYP1* was detected. The novel 20 kb for *AMYP1* was also not present. Since *AMYP1* was never detected by *Pst*I digest, it is unlikely that TD80P has a copy of *AMYP1* present in his/her genome.

From various digests, it would appear that TD80P has only three amylase genes present in his/her genome – *AMY2B, AMY2A,* and *AMY1C.* From NCBI's reference genome, we know that having one copy of all amylase genes would create a *Pme*I fragment of 354 kb. Since TD80P has a *Pme*I fragment of less than 354kb, it would be unlikely that this individual would have more than one copy of each gene. Based on this fact, we postulate that TD80P has two haplotypes containing one copy each of the three amylase genes. TD80P, therefore, would have two haplotypes that have the following layout: (*AMY2B-AMY2A-AMY1C*)$_1$.

2.3.9 Homozygous for 354kb *Pme*I fragment and elucidation of *Taq*I Fragment H identity

From NCBI's reference genome, we were able to predict that having one copy of each human amylase gene results in a 354kb *Pme*I fragment. Using this information, we have been able to elucidate the identities of several novel RFLPs within *Taq*I digests and *Pst*I digests.

*Taq*I fragment patterns for individuals homozygous for the 354kb *Pme*I fragment are overwhelmingly ABCDE. In individuals with 354kb and ABCDE, fragment B is slightly more intense than the others, indicating the presence of both *AMY2A* and *AMYP1*. Double digest results for these individuals corroborates with *Taq*I results. The 5.8kb fragment in the double digest was

slightly more intense than the 6.5kb fragment. *Pst*I digests further supported these results. The fragment representing all *AMY1* genes was more intense than those for *AMY2A, AMYP1,* and *AMY2B*. In *Pst*I digests, those fragments for all other genes appeared to be of the same intensity, showing that there were equal copies of all genes present.

There are also a number of individuals in the subpopulation of homozygous 354kb that present with fragment H as part of their *Taq*I phenotype. For example, some individuals with homozygous 354kb display a BEH phenotype. Fragment B and E are known to represent *AMY2A/AMYP1* and *AMY2B*, respectively, indicating that fragment H must represent all *AMY1* genes present. Individuals with TaqI phenotypes that differ from ABCDE and BEH (such as ABEH) show changes in intensity of fragment H that corresponds to the number of *AMY1* genes present.

2.3.10 Support for Groot's model for human amylase pattern of variation

Groot et al. proposed that the human amylase locus would vary according to the model *AMY2B-AMY2A-(AMY1A-AMY1B-AMYP1)*$_n$*-AMY1C* (Groot et al., 1990). Within our study population, many subjects displayed phenotypes that support Groot's model (Figure 2.9, TD55P). *Taq*I fragments that varied the most in intensity within the population are fragments B, C, D, and H (when present). Individuals with intense B fragments usually showed a more intense fragment for

AMYP1 than for *AMY2A* in *Pst*I-AMY543 digests. This indicated that increased *Taq*I B fragment intensity was due to increased copies of *AMYP1*, rather than increased copy of *AMY2A.* In our entire study population, we never saw a change in *Taq*I *AMY1C* fragment intensity unless fragment H was also present, indicating that *AMY1C* may play a role in anchoring the locus. When looking at individuals with very large *Pme*I fragments, we tended to see a large increase in copies of salivary *AMY1A* and *AMY1B*, rather than increases in copies of *AMY1C*. We also saw a trend of increased copies of *AMYP1* as well.

2.3.11 TD72P, TD110P, and copy number variation of *AMY2A*

TD72P/TD110P (referred to as TD72P hereafter) presented with one *Pme*I fragment of 280kb in size, indicating that this individual had two haplotypes of equal size (Figure 2.10E). Unlike TD80P's ABE pattern, TD72P's *Taq*I phenotype was ABDEH (Figure 2.10A). Despite having *Pme*I fragments of equal size to TD80P, TD72P's haplotypes were very different from TD80P. Presenting with fragment D meant that at least one copy of *AMY1A* is present. The presence of fragment A suggested at least one copy of *AMY1C* as well. Since fragment H can represent any or all salivary *AMY1* genes, the presence of this fragment was not overly informative. *Taq*I fragment B can represent either *AMYP1* or *AMY2A*. Further digests needed to be performed in order to determine the haplotype of this individual.

Double digest results for TD72P did not show fragments of equal intensity like in the case of TD80P (Figure 2.10B). TD80P displayed fragments of equal intensity. TD72P's 6.5kb fragment was much more intense than the others, indicating that this individual has a greater number of salivary amylase genes (*AMY1A* and *AMY1C*) than pancreatic genes. Based on *Taq*I and double digest results, it is impossible to say with certainty whether TD72P has a haplotype containing *AMY1B*. We also could not determine from these digests alone whether this individual has *AMY2A, AMYP1* or both genes present. To determine which of the two genes are present, we consulted the *Pst*I digest results.

*Pst*I digests was especially informative for TD72P (Figure 2.10C-D). *AMYP1* is present in this individual, as indicated by the presence of both 17.0kb and 2.1kb fragments. As expected from other digests' results, TD72P presented with the 8.2kb fragment representing all salivary *AMY1* genes and the 15.2 kb fragment representing *AMY2B*. What was extraordinary about TD72P was the fact that this individual lacked the fragment representing *AMY2A*. Expressing no fragment would indicate that TD72P has no copies of *AMY2A* present within their genome, making this individual the first in our study to be homozygous deficient for this pancreatic gene. Based on all digests together, we speculate that TD72P has two copies of the following haplotype: $(AMY2B\text{-}AMY1A\text{-}AMP1\text{-}AMY1C)_1$.

Although we saw no other individuals homozygous deficient for *AMY2A*, we did see individuals that appear to vary in the copy number of this pancreatic

gene. At present, evidence suggests that individuals can vary in *AMY2A* copy number from 0-4 copies.

2.3.12 TD25P and copy number variation of *AMY2B*

Unlike as in the case of *AMY2A*, no individuals in our study presented with a homozygous deficiency of pancreatic *AMY2B*. However, we did have evidence supporting the possibility of this gene also varying in copy number (Figure 2.11). An example of an individual with possible CNV of pancreatic *AMY2B* is TD25P. As shown previously, the amount of DNA loaded for each individual is comparable but we saw an intense 4.9kb *Taq*I fragment corresponding to *AMY2B*. Double digest results for this individual showed the 3.6kb fragment for *AMY2B* was much more intense than those for other individuals with the same amount of DNA. Both digests suggested that TD25P contains higher copy number for *AMY2B* than the expected 2 copies per individual. Within the population, *AMY2B* is estimated to vary in copy from 1-4 copies per individual.

2.3.13 Pattern of Variation for the Human Amylase Locus in T1D

Copy number of *AMY2A*, *AMY2B* and total copy number of human salivary *AMY1* and pancreatic *AMY2* (excludes *AMYP1*) genes were estimated for 344 individuals diagnosed with T1D. Estimation of gene copy was

determined by comparing *Taq*I, *Pvu*II/*Psh*AI, *Pst*I-AMY345, and PstI-AMY543 results for each sample. The results are presented in Figure 2.12.

From *Pme*I digests, we determined that the expected 354kb fragment containing 6 amylase genes was present at an allele frequency of 0.46. Individuals homozygous for this haplotype would be expected to have a total of 12 genes – 6 salivary *AMY1* genes, 2 copies of pancreatic *AMY2B*, 2 copies of pancreatic *AMY2A*, and 2 copies of pseudogene *AMYP1*. In all four categories (total *AMY1*, *AMY2A*, *AMY2B*, and total *AMY2*), expected copy number had the highest frequency. Having six copies of *AMY1* genes occurred at a frequency of 0.27. 86% of the population had the expected 2 copies of *AMY2B*, while 2 copies of *AMY2A* occurred at a frequency 0.58. Slightly more than half the population (0.55) had the expected 4 copies of *AMY2* genes. However, we found variation for all amylase genes from expectations.

Copy number of salivary *AMY1* genes varied the most within the population. For simplicity, individuals having greater than 10 salivary genes were grouped together. Individuals having less than the expected 6 salivary genes were also grouped together for simplicity of study. Number of *AMY1* genes ranged from as little as 2 copies to as many as 17 copies within a diploid genome. Having 8 copies of *AMY1* genes within a diploid genome was the median for the population. 6 copies of *AMY1* occurred at the highest frequency (0.27) with those having greater than 10 copies in their genome having the

second highest frequency (0.24). This indicates a great degree of variation within the population.

Total number of *AMY2* genes also varied, although not to the extent seen in the AMY1 genes. Although most of the population contained only 4 copies of *AMY2* in their respective genomes, as many as 10 copies of *AMY2* were seen in individuals. As was the case with *AMY1*, the smallest copy number seen for total *AMY2* was 2 copies in a diploid genome. The median for the population was 4 copies of *AMY2*. Having 4 copies of *AMY2* also was the most prevalent in the population, with 55% of individuals having 4 copies in their genome. The next most common number of copies seen in the population was having 5 copies of *AMY2* present (0.18). Surprisingly, the population trended toward having more copies of *AMY2* genes. Of the remaining individuals that did not have 4 copies of *AMY2*, 85% had greater than 4, rather than fewer copies.

Copy number of *AMY2B* deviated the least out of the four categories studied. The vast majority of the population only had two copies of *AMY2B* present in their diploid genomes, with 2 copies occurring at a frequency of 0.86 within the group. We did see a range of 1-4 copies of *AMY2B* in the population. In no instance did we see an individual that was homozygous deficient for *AMY2B*. The next most frequent genotype seen was 3 copies of *AMY2B*. As was the case with total copy number of *AMY2*, the population trended toward more copies than the expected 2 copies of *AMY2B*, rather than having less.

Of the pancreatic *AMY2* genes, *AMY2A* showed the most variation, with copy number ranging from 0 to 6 copies in diploid individuals. For simplicity of study, those with greater copy number than 4 were grouped together. Again, the majority of our population had the expected 2 copies of *AMY2A*, with this genotype occurring within 58% of the group. The median was also 2 copies of *AMY2A*. The second most common genotype, having 3 copies of AMY2A, occurred at a frequency of 0.19. Of the remaining individuals that deviated from 2 copies of *AMY2A*, 84% had more than 2 copies present in their genomes. This indicated that the population trends toward having more copies of *AMY2A*.

2.4 Discussion

There is a large body of work on human amylase CNV but very little has been done to elucidate the exact pattern of variation and copy number of specific amylase genes. It, therefore, became crucial to investigate the pattern of variation at the locus. Restriction enzyme digests and use of a larger population has provided us with a clearer picture of amylase locus diversification.

Dot plot analysis and study of the locus as a whole revealed a region that contains large stretches of sequence that share a high degree of similarity. These stretches span not just amylase genes but extragenic regions as well. This is not unexpected, since high degree of sequence similarity between amylase genes themselves, as well as retroviral and γ-actin pseudogene inserts 5' to all amylase genes have been previously reported (Samuelson et al., 1988). Study of the amylase region revealed large blocks of 44-64kb in size that share sequence similarity. Present within these blocks are smaller segments ranging in sizes of 7kb to 26kb that also share similarity. The large blocks contain either 2 genes (like the block spanning *AMY2A* and *AMY1A*) or just one gene (such as the large, 64kb block that includes *AMY2B,* or the 44kb block that contains *AMY1B*). Some of the smaller segments span genes, while some only span intragenic regions.

The block containing *AMY1B* and surrounding extragenic regions is situated in opposite orientation from the rest of the locus. Nevertheless, this

region shares high sequence similarity with other regions in the amylase cluster. It is important to note, however, that while highly similar to the rest of the locus, this 44kb block is not identical to the blocks containing *AMY2A-AMY1A* and *AMYP1-AMY1C*. It is also important to note that this block is identical to the region spanning 3' of *AMYP1* to 3' of *AMY1C*, including all of *AMY1C* gene. This indicates that this block probably evolved as a duplication specifically of *AMY1C* and surrounding regions. Whether inversion of *AMY1B* block occurred during this duplication or afterwards is not immediately clear.

While the 64kb block that includes *AMY2B* contains large spans of sequence that share similarity with other regions in the amylase locus, it is unique in being the only block that is not 100% identical in its entirety with other regions of the amylase locus. Two large breaks occur within this block, one within *AMY2B* itself and another that occurs 3' of *AMY2B*. Break 1, located within *AMY2B*, was determined to be a β-actin enteric pseudogene insert. We have yet to determine the identity of break 2. Because this block lacks sequence homology with the rest of the locus, it is improbable that the block as a whole would be duplicated. During our analysis, we never saw a segmental duplication that equaled 64kb in size.

The blocks containing two genes, a 50kb block spanning *AMY2A-AMY1A* and a 48kb block spanning *AMYP1-AMY1C*, are identical in sequence. It is probable that the 48kb block originated initially from a duplication of the 50kb block. Loss of 2kb at the start of *AMYP1*, which resulted in a truncated

pseudogene, led to the smaller block. Since a portion of the 5' region of *AMY1B* block shares similarity with both *AMY2A* and *AMYP1*, it is possible that inversion of *AMY1B*'s block resulted in truncation of *AMYP1*. Due to shared similarity between *AMY2A* and *AMYP1*'s blocks, we would expect to see duplications of 48-50 kb that result in gains of copy number for either *AMY2A-AMY1A* or *AMYP1-AMY1C*. Indeed, we did see segmental duplications that differed in size of about 50kb. In individuals with such duplications, we did see increased copy of corresponding genes.

Residing within all larger blocks, are smaller segments of shared sequence homology that ranged from 30kb (present only in *AMY1B* block) to 7kb. Having such a high degree of similarity between large and smaller spans of sequence has several implications for human amylase cluster. Unlike in other instances of copy number variation, where whole genes are gained or lost, the amylase locus has the potential for partial loss or gain of genes, as well as loss/gain of intergenic regions without loss of genes. This is seen in our *Pme*I digests and even between individuals. Both of our case studies, TD80P and TD72P, share a homozygous 280kb *Pme*I fragment, yet their actual genotypes differ. TD80P's genotype is (*AMY2B-AMY2A-AMY1C*), while TD72P's genotype is (*AMY2B-AMY1A-AMYP1-AMY1C*). Such variation for smaller *Pme*I haplotypes leads to the question of whether a "simple haplotype" for the amylase locus exists. For those with fragments larger than 354kb, variation between individuals with the same *Pme*I fragment size prevents us from quantitatively defining copy

number of amylase genes. More quantitative measures, such as real-time qPCR or sequencing the region, will be necessary to definitively define the total copy number of human salivary and pancreatic amylase genes.

*Taq*I, *Pvu*II/*Psh*AI, and both *Pst*I digests allowed us to determine copy number for individual genes. Results from these digests revealed not just copy number of amylase genes, but also revealed a number of RFLPs that exist fairly commonly in our population. The H fragment, present in 62% of the *Taq*I digest results, represents not just one change to the locus but three. All salivary amylase *AMY1* genes can condense to this 3.1kb fragment. Its presence within a genotype renders our *Taq*I less informative, as at this time we have yet to determine which *AMY1* genes are being represented by this fragment. Even if an expected *Taq*I fragment representing one of the salivary genes were missing in a genotype with fragment H, we are incapable of determining if one or more genes are being represented by fragment H without another form of verification. Since salivary amylase genes are almost identical in sequence, a real time qPCR may not be sufficient to answer this question. Creating a BAC library of clones followed by sequencing of the amylase region would allow us to ascertain the total number of *AMY1* genes present and how many individual *AMY1* genes exist for an individual.

In our study population, we found one individual that was homozygous deficient for AMY2A. Rather than having a copy of *AMY2A*, this individual's digest results suggest s/he has one copy of *AMYP1* and one copy of *AMY1A*

instead. The absence of an *AMY2A* fragment in both *Taq*I and *Pst*I results indicate that the entire gene has been lost, rather than just partial loss or detection of *AMYP1* instead. TD72P's genotype, along with that from the Alternative Reference Sequence of the NCBI Human Genome Database , suggest that loss of *AMY2A*, while perhaps not as common as loss of salivary *AMY1* genes, is still a common occurrence. Loss of *AMY2A* could be explained as alignment of the short 30kb segment 5' of *AMY2A* on one chromosome with the corresponding region 3' of the gene on a second. Subsequent unequal recombination would result in two copies of *AMY2A* in one haplotype and complete loss of *AMY2A* in the other. TD72P most likely inherited two haplotypes deficient for *AMY2A*, rather than obtaining this deficiency through novel recombination. Obtaining genomic DNA and genotyping of TD72P's parents would elucidate how this individual came to be homozygous deficient.

Although we found no individuals homozygous deficient for *AMY2B*, we did find a number of individuals that appeared to have *AMY2B* CNV. Because the block containing *AMY2B* does not share 100% sequence similarity with any other block in the amylase locus, it is unlikely that the whole block is capable of duplicating. A β-actin pseudogene insert within *AMY2B* itself discourages complete duplication of the gene. However, the presence of smaller segments that share sequence homology with other segments in the cluster provide opportunity for at least partial duplication of *AMY2B*. *AMY2B's* region 3' to b-actin insert shares sequence homology with *AMY2A*'s 3' intragenic region, *AMY1A*'s 5'

extragenic region, all of *AMYP1*, and *AMY1C*'s 5' extragenic region. Alignment of any of these sequences on one chromosome with corresponding *AMY2B* sequence on another chromosome could lead to a recombination event that leads to only partial loss of *AMY2B*. To confirm this, sequencing the locus of this individual would prove useful.

An RFLP that results in a 4.2kb fragment in *Taq*I digests occurs at a rate of only 5% in our T1D population. Fragment F's presence in a *Taq*I genotype normally results in a less intense 4.9kb fragment corresponding with *AMY2B* copy number, indicating that the gene represented by fragment F is *AMY2B*. To determine fragment F's identity, the fragment can be excised directly from a gel and incorporated into a TA cloning vector for sequencing.

Surprisingly, we found no individual that appeared homozygous deficient for *AMY1C*, despite finding numerous individuals that appeared to have variable copy number of *AMY1B*. For individuals with varying copy number of *AMY1B*, the most probable explanation behind its duplication would be alignment of 44kb block with *AMY1C* and its corresponding regions 5' and 3' of the gene and subsequent recombination. Doing so would lead to complete loss of *AMY1C* in one of the resulting haplotypes. Since loss of *AMY1C* is not overtly noted, it is possible that *AMY1B* is aligning with *AMY1A* instead. Because of current assay limitations, it is possible that individuals with a *Taq*I pattern of BEH could be homozygous for *AMY1C* but at this time, we are unable to determine this for sure. If, in fact, *AMY1C* is still present in these individuals, it is unclear why

AMY1C is never deleted within this locus. Perhaps *AMY1C* plays the role of anchor for the entire amylase locus or is vital in some other way. More investigation into a possible mechanism behind *AMY1C*'s preservation within the amylase locus is needed.

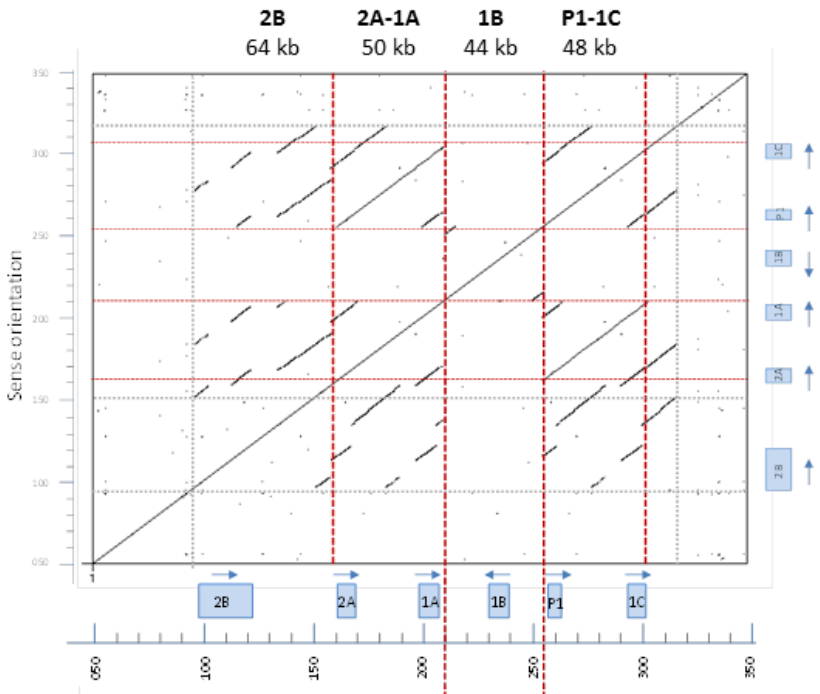
One of the goals of this study was to characterize and present the copy number of human amylase genes in our T1D population. As expected, the vast majority of individuals displayed the same copy of *AMY1* and *AMY2* genes as reported in NCBI's reference – 6 salivary *AMY1* genes, 4 pancreatic AMY2 genes, and 2 copies of pseudogene *AMYP1*. Salivary *AMY1* copy number varied the most, with the number of individuals having more than 10 copies or less than 6 copies making up the vast majority of individuals deviating from 6 copies per diploid genome. *AMY2* copy number showed a more directional trend. Most individuals that deviated from expected copy number had increased copy number, rather than decreased copy number of *AMY2*. This is also true when pancreatic amylase genes are separated to view copy number of *AMY2B* and *AMY2A*. Given the limitations of our study, these results should be confirmed with by real-time qPCR.

It is not immediately clear why we should see a trend of increased copy number of pancreatic *AMY2* genes in our T1D population. There are a number of further studies that must be done to better understand this trend. Usually, more copies of a gene would equate to higher levels of protein made. Is this true for pancreatic amylase genes? Pancreatic amylase levels can be determined

from extracted blood of individuals. Western blots coupled with Bradford assays would allow us to equate protein concentrations with *AMY2* copy number. Also, is this trend specific to T1D or does the same apply for individuals without T1D? At this present time, we are gathering information about copy number of amylase genes in a population of Caucasian and African American individuals without T1D.

Groot presented a model for how diversification at amylase locus originated (Groot et al., 1990). This model was based on unequal homologous recombination between salivary *AMY1* genes and between *AMY2A* and *AMYP1*, giving rise to greater number of *AMY1* and *AMYP1* genes. However, such a high degree of sequence similarity throughout the locus indicates a much more complex picture than Groot's simple model suggests. We believe the model should be adjusted to explain how *AMY2A* and *AMY2B* is also undergoing partial and complete duplication. While it is probable that variation at the amylase locus is occurring via unequal homologous recombination events, we believe these events are occurring between more genes than just between *AMY2A* and *AMYP1*. We have presented evidence that suggests it is possible for multiple matchups to occur, including genes with intergenic regions. While more work needs to be done to further characterize and investigate this locus, it would appear that the pattern of variation is exceedingly complex and complicated.

A.



B.

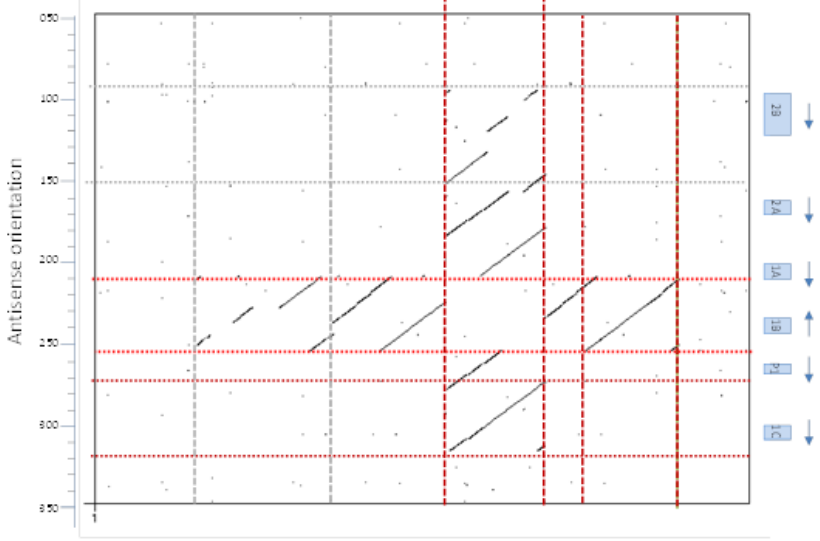


Figure 2.1 Dot plot analysis of the entire human amylase locus as listed in NCBI's reference sequence. A. Human amylase locus in the forward orientation compared to itself in same orientation. B. Human amylase locus in correct orientation (X-axis) compared with itself in opposite orientation.

Table 2.1 Estimated fragment sizes for each amylase gene depending on which restriction enzyme and probe (AMY300, AMY345, AMY543) is used.

	TaqI	PvuII/PshAI		PstI	
gene	AMY300	AMY300	AMY345	AMY345	AMY543
AMY1A	5.3	6.5	3.3	8.2	8.2
AMY1B	5.7	5.8	3.3	8.2	8.2
AMY1C	7.8	6.5	3.3	8.2	8.2
AMY2A	7.1	5.8	6	5.4	5.4
AMY2B	4.9	3.6	21.4	15.2	15.2
AMYP1	7.1	5.8	X	X	2.1

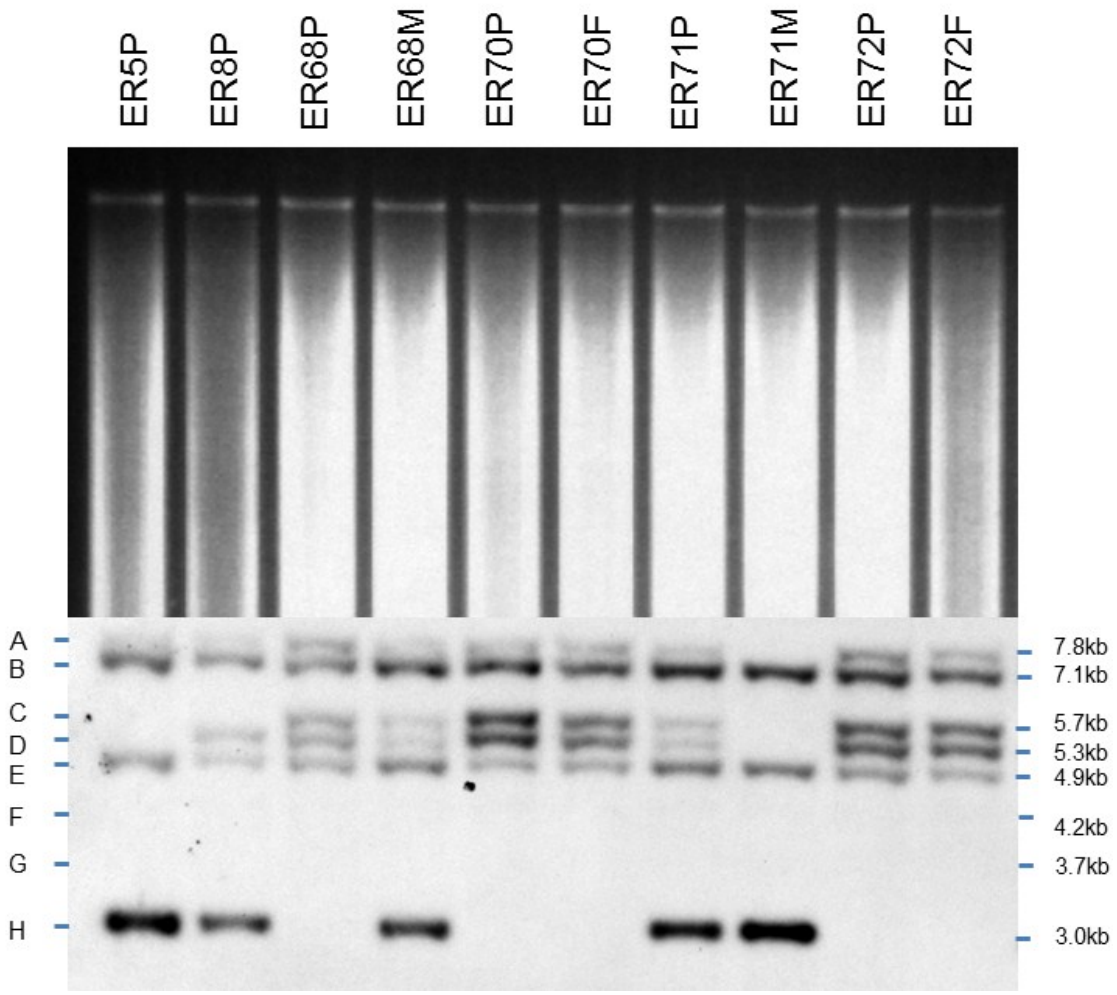


Figure 2.2 *Taq*I RFLP Southern analyses for copy number of human salivary and pancreatic amylase genes in 10 individuals. Top panel shows DNA of each sample after gel had been soaked in EtBr. Bottom panel shows *Taq*I results following a week long exposure.

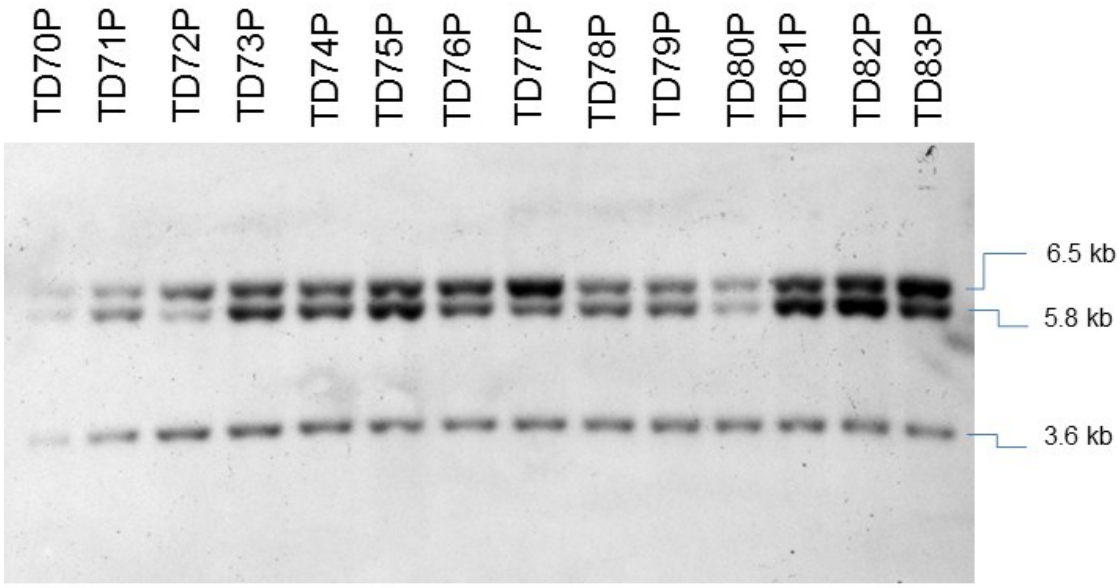


Figure 2.3 *Pvu*II/*Psh*AI RFLP Southern analysis for copy number variation of human salivary and pancreatic amylase genes in 13 individuals.

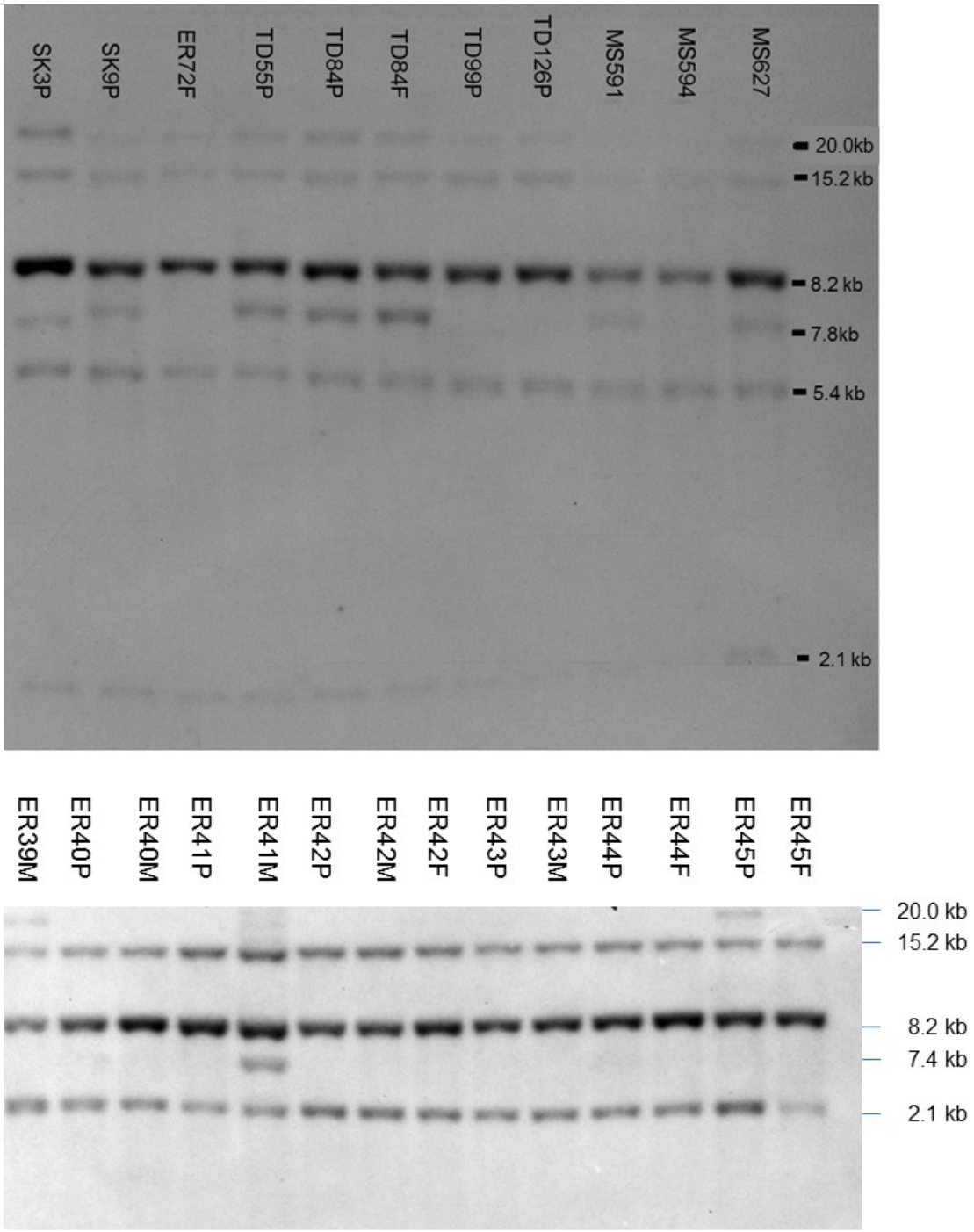


Figure 2.4 *Pst*I RFLP Southern analysis for copy number of human salivary and pancreatic amylase genes in 26 unrelated individuals. Probes used were AMY-543 (top) and AMY345 (bottom).

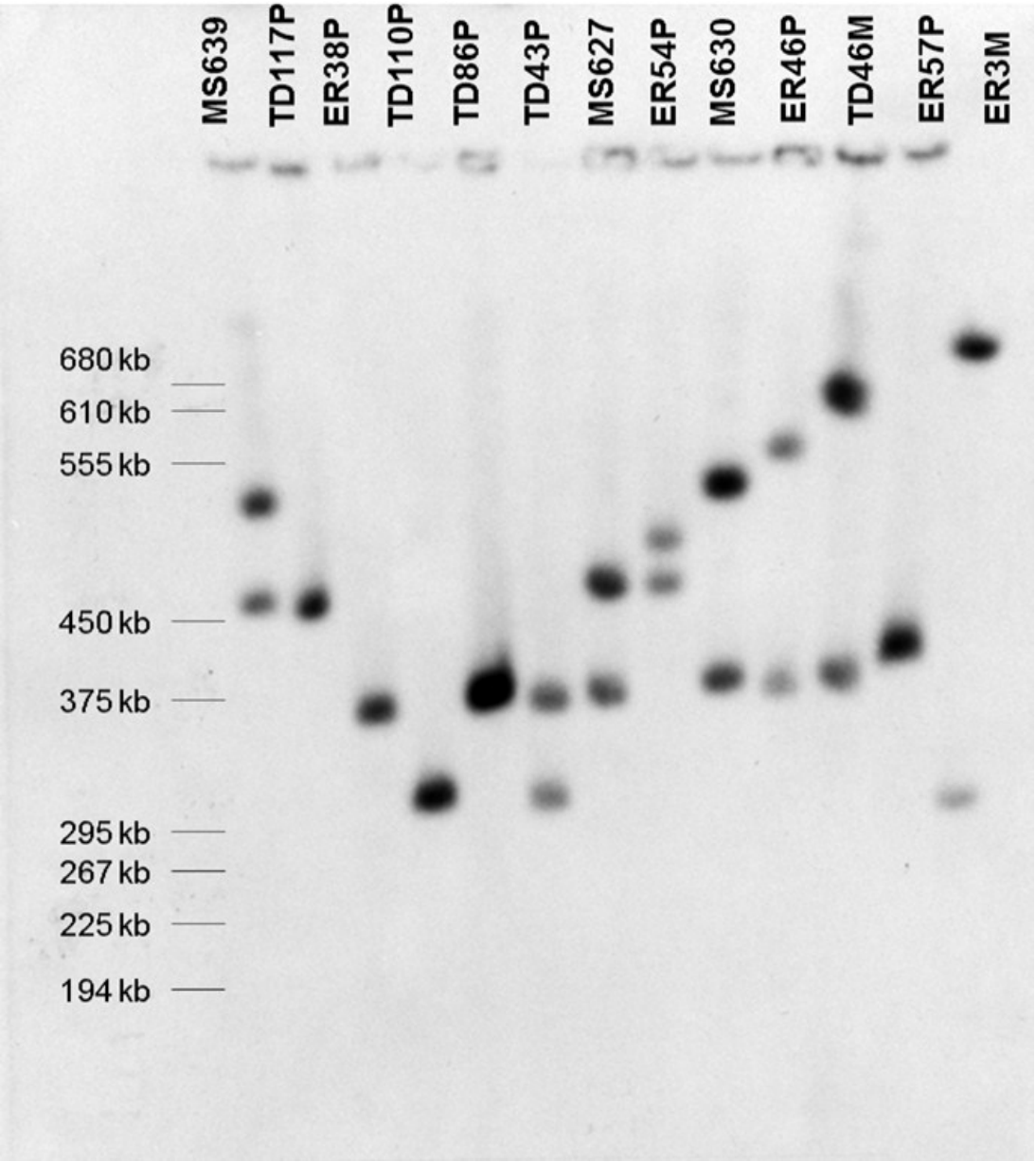


Figure 2.5 *Pme*I RFLP pulsed field gel electrophoresis for human amylase locus size in 13 unrelated individuals with Type 1 diabetes.

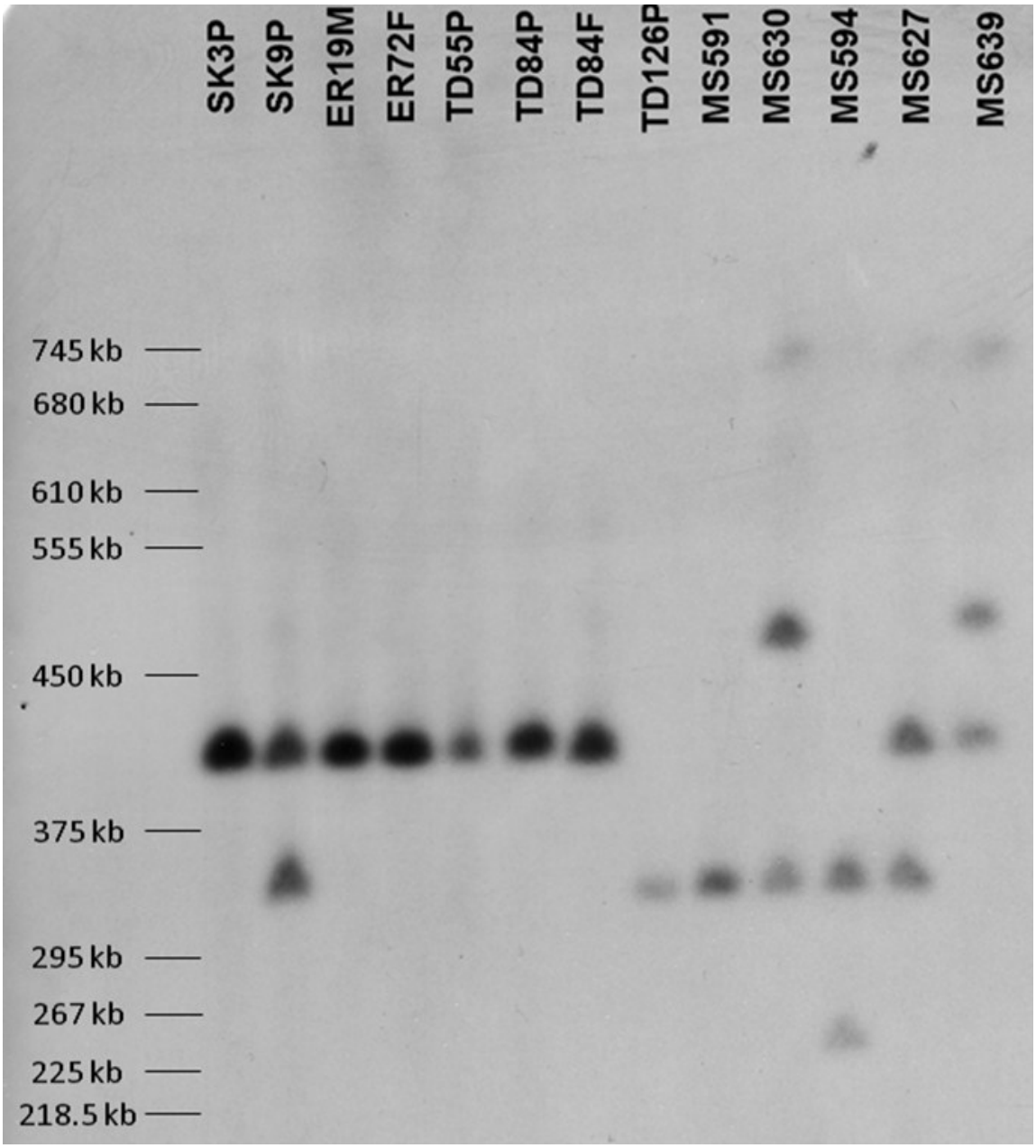


Figure 2.6 *Pme*I RFLP pulsed field gel electrophoresis for human amylase locus size in 13 unrelated individuals with Type 1 diabetes.

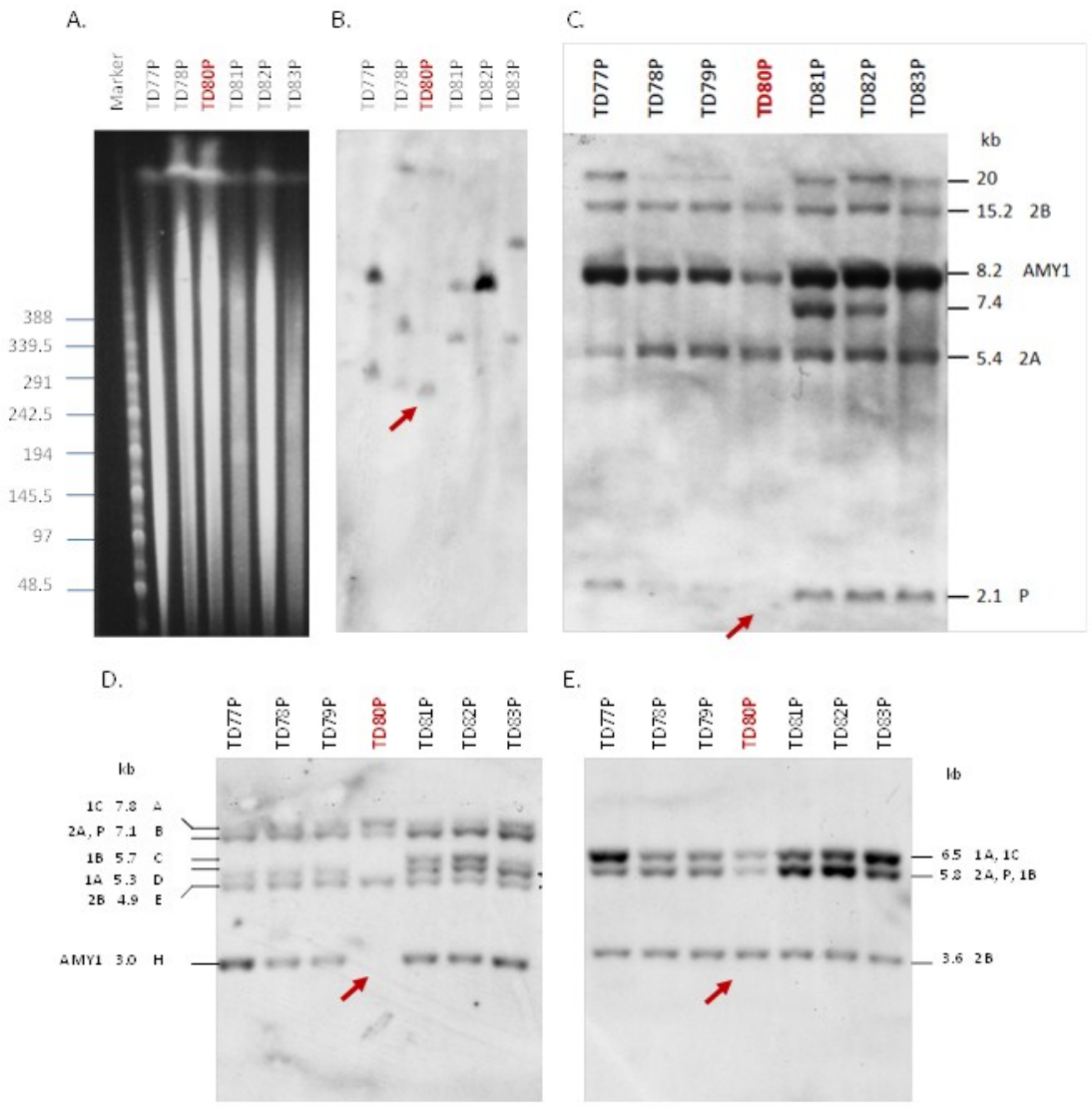


Figure 2.7 RFLP Southern analyses for copy number of human salivary and pancreatic amylase genes in TD80P. **A**. DNA for individuals following a gel run, made visible by soaking gel in EtBr. **B**. *Pme*I RFLP pulsed field gel electrophoresis results. **C.** *Pst*I-AMY543 Southern analysis results. **D.** *Taq*I Southern analysis results. **E.** *Pvu*II/*Psh*AI Southern analysis results.

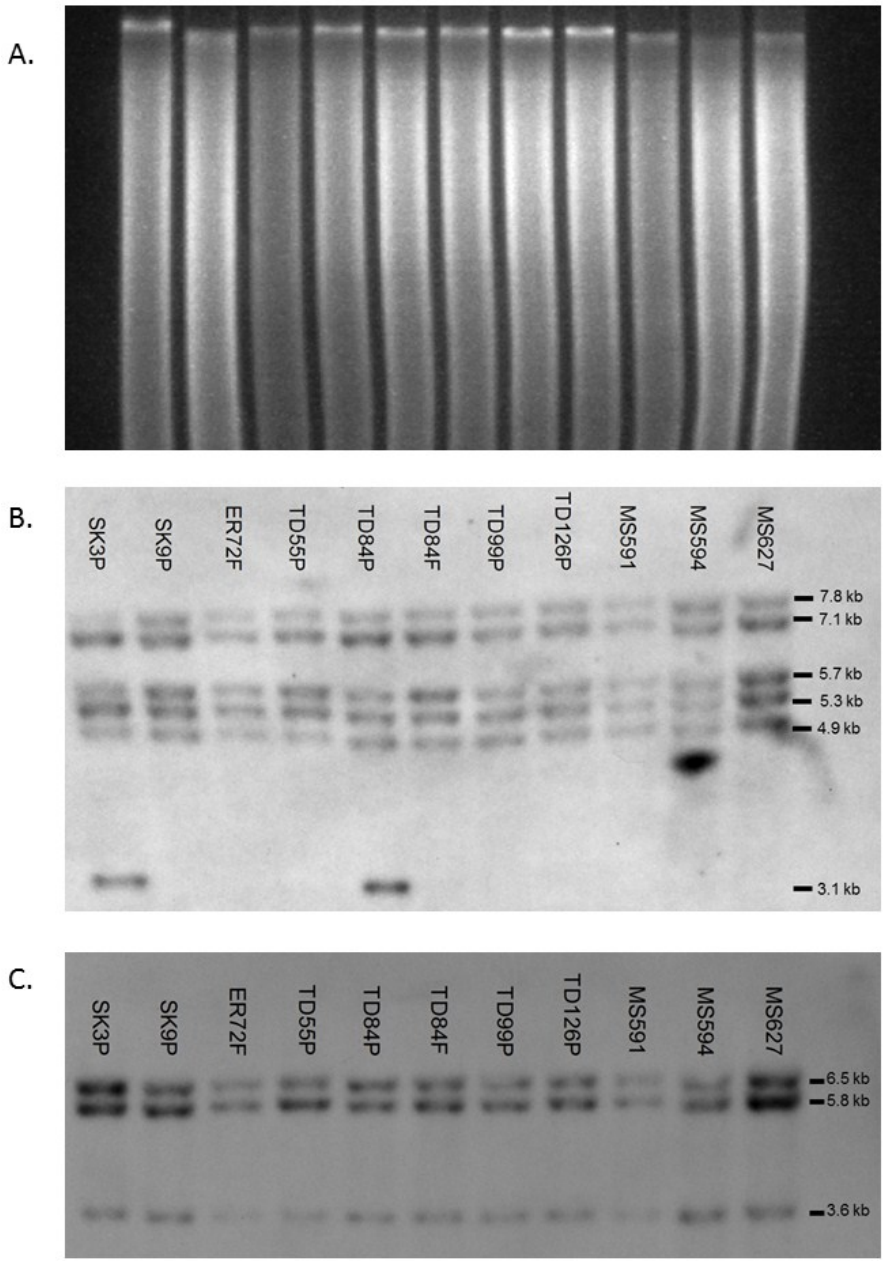


Figure 2.8 RFLP Southern analysis for copy number of human salivary and pancreatic amylase genes in 11 unrelated individuals. **A.** DNA smears for individuals after soaking gel in EtBr. **B.** *Taq*I RFLP Southern analysis results. **C.** *Pvu*II/*Psh*AI RFLP Southern analysis results.

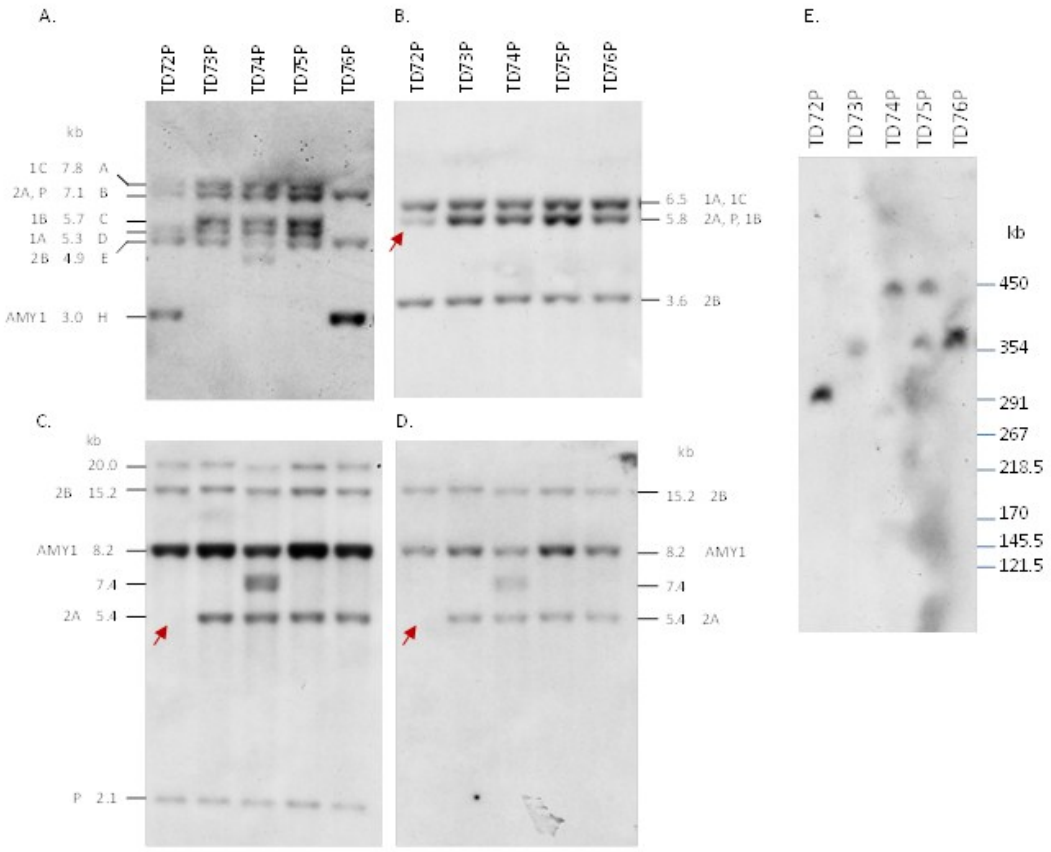


Figure 2.9 All RFLP digest results for TD72P/TD110P. **A.** *Taq*I RFLP Southern analysis results. **B.** *Pvu*II/*Psh*AI RFLP Southern analysis results. **C**. *Pst*I-AMY543 RFLP Southern analysis. **D.** *Pst*I-AMY345 RFLP Southern analysis results. **E.** *Pme*I pulsed field gel electrophoresis results.

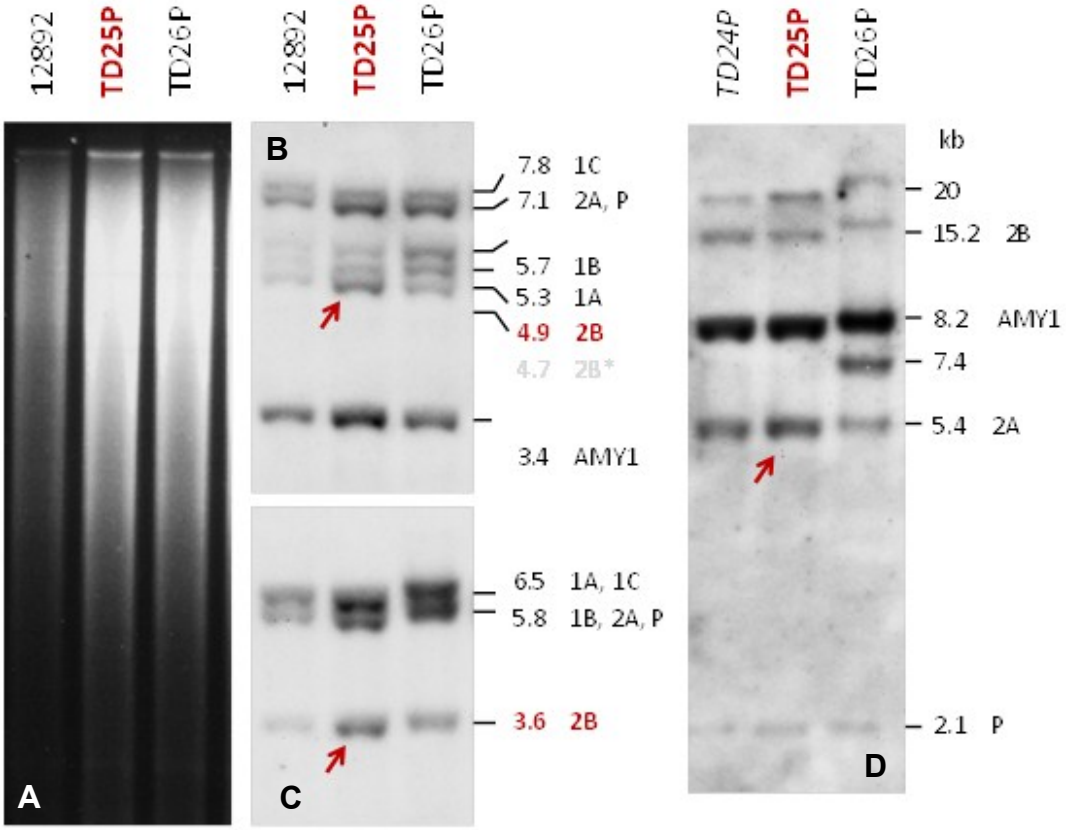


Figure 2.10 RFLP Southern analyses for copy number of human salivary and pancreatic amylase genes in TD25P. **A.** DNA smear for TD25P and other individuals after soaking gel in EtBr. **B.** *Taq*I RFLP digest results. **C.** *Pvu*II/*Psh*AI RFLP digest results. **D.** PstI-AMY543 RFLP Southern result.

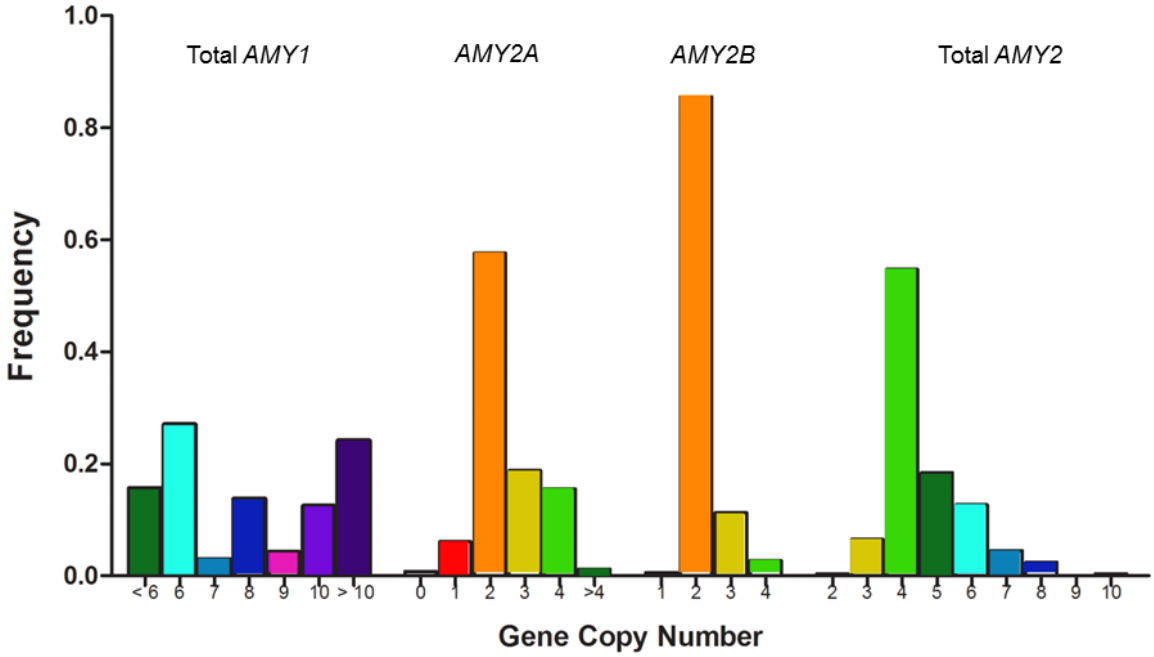


Figure 2.11 Variations of human salivary and pancreatic amylase genes in the diploid genomes of Type 1 diabetes patients.

REFERENCES

Bahn E. (1967) Crossing over in the chromosomal region determining amylase isozymes in Drosophila melanogaster. *Hereditas* **58**, 1-12.

Bluestone J. A., Herold K. and Eisenbarth G. (2010) Genetics, pathogenesis and clinical interventions in type 1 diabetes. *Nature* **464**, 1293-300.

Boettcher B. and La Lande F. A. (1969) Electrophoresis of human saliva and identification of inherited variants of amylase isozymes. *Aust J Exp Biol Med Sci* **47**, 97-103.

Brayer G. D., Luo Y. and Withers S. G. (1995) The structure of human pancreatic alpha-amylase at 1.8 A resolution and comparisons with related enzymes. *Protein Sci* **4**, 1730-42.

Brayer G. D., Sidhu G., Maurus R., Rydberg E. H., Braun C., Wang Y., Nguyen N. T., Overall C. M. and Withers S. G. (2000) Subsite mapping of the human pancreatic alpha-amylase active site through structural, kinetic, and mutagenesis techniques. *Biochemistry* **39**, 4778-91.

Chatterton R. T., Jr., Vogelsong K. M., Lu Y. C., Ellman A. B. and Hudgens G. A. (1996) Salivary alpha-amylase as a measure of endogenous adrenergic activity. *Clin Physiol* **16**, 433-48.

Chung E. K., Yang Y., Rennebohm R. M., Lokki M. L., Higgins G. C., Jones K. N., Zhou B., Blanchong C. A. and Yu C. Y. (2002) Genetic sophistication of human complement components C4A and C4B and RP-C4-CYP21-TNX (RCCX) modules in the major histocompatibility complex. *Am J Hum Genet* **71**, 823-37.

Emi M., Horii A., Tomita N., Nishide T., Ogawa M., Mori T. and Matsubara K. (1988) Overlapping two genes in human DNA: a salivary amylase gene overlaps with a gamma-actin pseudogene that carries an integrated human endogenous retroviral DNA. *Gene* **62**, 229-35.

Endo T., Takizawa S., Tanaka S., Takahashi M., Fujii H., Kamisawa T. and Kobayashi T. (2009) Amylase alpha-2A autoantibodies: novel marker of autoimmune pancreatitis and fulminant type 1 diabetes. *Diabetes* **58**, 732-7.

Fleming W. L. and Neill J. M. (1927) Studies on Bacterial Enzymes : V. The Carbohydrases and Lipase of the Welch Bacillus. *J Exp Med* **45**, 947-59.
Freeman J. L., Perry G. H., Feuk L., Redon R., McCarroll S. A., Altshuler D. M., Aburatani H., Jones K. W., Tyler-Smith C., Hurles M. E., Carter N. P., Scherer S. W. and Lee C. (2006) Copy number variation: new insights in genome diversity. *Genome Res* **16**, 949-61.
Groot P. C., Bleeker M. J., Pronk J. C., Arwert F., Mager W. H., Planta R. J., Eriksson A. W. and Frants R. R. (1988) Human pancreatic amylase is encoded by two different genes. *Nucleic Acids Res* **16**, 4724.
Groot P. C., Bleeker M. J., Pronk J. C., Arwert F., Mager W. H., Planta R. J., Eriksson A. W. and Frants R. R. (1989) The human alpha-amylase multigene family consists of haplotypes with variable numbers of genes. *Genomics* **5**, 29-42.
Groot P. C., Mager W. H. and Frants R. R. (1991) Interpretation of polymorphic DNA patterns in the human alpha-amylase multigene family. *Genomics* **10**, 779-85.
Groot P. C., Mager W. H., Henriquez N. V., Pronk J. C., Arwert F., Planta R. J., Eriksson A. W. and Frants R. R. (1990) Evolution of the human alpha-amylase multigene family through unequal, homologous, and inter- and intrachromosomal crossovers. *Genomics* **8**, 97-105.
Gumucio D. L., Wiebauer K., Caldwell R. M., Samuelson L. C. and Meisler M. H. (1988) Concerted evolution of human amylase genes. *Mol Cell Biol* **8**, 1197-205.
Hill C. J., Rowe S. I. and Lovrien E. W. (1972) Probable genetic linkage between human serum amylase (Amy 2) and Duffy blood group. *Nature* **235**, 162-3.
Hollox E. J., Armour J. A. and Barber J. C. (2003) Extensive normal copy number variation of a beta-defensin antimicrobial-gene cluster. *Am J Hum Genet* **73**, 591-600.
Horii A., Emi M., Tomita N., Nishide T., Ogawa M., Mori T. and Matsubara K. (1987) Primary structure of human pancreatic alpha-amylase gene: its comparison with human salivary alpha-amylase gene. *Gene* **60**, 57-64.
Iafrate A. J., Feuk L., Rivera M. N., Listewnik M. L., Donahoe P. K., Qi Y., Scherer S. W. and Lee C. (2004) Detection of large-scale variation in the human genome. *Nat Genet* **36**, 949-51.
Ishizaki K., Noda A., Ikenaga M., Ida K., Omoto K., Nakamura Y. and Matsubara K. (1985) Restriction fragment length polymorphism detected by human salivary amylase cDNA. *Hum Genet* **71**, 261-2.
Jacobs P. A., Matsuura J. S., Mayer M. and Newlands I. M. (1978) A cytogenetic survey of an institution for the mentally retarded: I. Chromosome abnormalities. *Clin Genet* **13**, 37-60.
Ji Y., Eichler E. E., Schwartz S. and Nicholls R. D. (2000) Structure of chromosomal duplicons and their role in mediating human genomic disorders. *Genome Res* **10**, 597-610.

Kaczmarek M. J. and Rosenmund H. (1977) The action of human pancreatic and salivary isoamylases on starch and glycogen. *Clin Chim Acta* **79**, 69-73.
Kamaryt J. and Laxova R. (1966) Amylase heterogeneity in man. *Humangenetik* **3**, 41-5.
Kandra L. and Gyemant G. (2000) Examination of the active sites of human salivary alpha-amylase (HSA). *Carbohydr Res* **329**, 579-85.
Karn R. C., Rosenblum B. B., Ward J. C., Merritt A. D. and Shulkin J. D. (1974) Immunological relationships and post-translational modification of human salivary amylase (Amy) and pancreatic amylase (Amy) isozymes. *Biochem Genet* **12**, 485-99.
Karn R. C., Shulkin J. D., Merritt A. D. and Newell R. C. (1973) Evidence for post-transcriptional modification of human salivary amylase (amyl) isozymes. *Biochem Genet* **10**, 341-50.
Kauffman D. L., Watanabe S., Evans J. R. and Keller P. J. (1973) The existence of glycosylated and non-glycosylated forms of human submandibular amylase. *Arch Oral Biol* **18**, 1105-11.
Koyama I., Komine S., Iino N., Hokari S., Igarashi S., Alpers D. H. and Komoda T. (2001) alpha-Amylase expressed in human liver is encoded by the AMY-2B gene identified in tumorous tissues. *Clin Chim Acta* **309**, 73-83.
Kraus F. W. and Mestecky J. (1971) Immunohistochemical localization of amylase, lysozyme and immunoglobulins in the human parotid gland. *Arch Oral Biol* **16**, 781-9.
Mandel A. L., Peyrot des Gachons C., Plank K. L., Alarcon S. and Breslin P. A. (2010) Individual differences in AMY1 gene copy number, salivary alpha-amylase levels, and the perception of oral starch. *PLoS One* **5**, e13352.
Mathis D., Vence L. and Benoist C. (2001) beta-Cell death during progression to diabetes. *Nature* **414**, 792-8.
Maurus R., Begum A., Kuo H. H., Racaza A., Numao S., Andersen C., Tams J. W., Vind J., Overall C. M., Withers S. G. and Brayer G. D. (2005) Structural and mechanistic studies of chloride induced activation of human pancreatic alpha-amylase. *Protein Sci* **14**, 743-55.
Meisler M. H. and Ting C. N. (1993) The remarkable evolutionary history of the human amylase genes. *Crit Rev Oral Biol Med* **4**, 503-9.
Merritt A. D., Lovrien E. W., Rivas M. L. and Conneally P. M. (1973) Human amylase loci: genetic linkage with the Duffy blood group locus and assignment to linkage group I. *Am J Hum Genet* **25**, 523-38.
Nahoum V., Roux G., Anton V., Rouge P., Puigserver A., Bischoff H., Henrissat B. and Payan F. (2000) Crystal structures of human pancreatic alpha-amylase in complex with carbohydrate and proteinaceous inhibitors. *Biochem J* **346 Pt 1**, 201-8.
Nishide T., Nakamura Y., Emi M., Yamamoto T., Ogawa M., Mori T. and Matsubara K. (1986) Primary structure of human salivary alpha-amylase gene. *Gene* **41**, 299-304.

Numao S., Maurus R., Sidhu G., Wang Y., Overall C. M., Brayer G. D. and Withers S. G. (2002) Probing the role of the chloride ion in the mechanism of human pancreatic alpha-amylase. *Biochemistry* **41**, 215-25.

Ogita S. (1966) Genetico-biochemical studies on the salivary and pancreatic amylase isozymes in human. *Med J Osaka Univ* **16**, 271-86.

Perry G. H., Dominy N. J., Claw K. G., Lee A. S., Fiegler H., Redon R., Werner J., Villanea F. A., Mountain J. L., Misra R., Carter N. P., Lee C. and Stone A. C. (2007) Diet and the evolution of human amylase gene copy number variation. *Nat Genet* **39**, 1256-60.

Pociot F., Akolkar B., Concannon P., Erlich H. A., Julier C., Morahan G., Nierras C. R., Todd J. A., Rich S. S. and Nerup J. (2010) Genetics of type 1 diabetes: what's next? *Diabetes* **59**, 1561-71.

Pronk J. C., Frants R. R., Jansen W., Eriksson A. W. and Tonino G. J. (1982) Evidence of duplication of the human salivary amylase gene. *Hum Genet* **60**, 32-5.

Ragunath C., Manuel S. G., Venkataraman V., Sait H. B., Kasinathan C. and Ramasubbu N. (2008) Probing the role of aromatic residues at the secondary saccharide-binding sites of human salivary alpha-amylase in substrate hydrolysis and bacterial binding. *J Mol Biol* **384**, 1232-48.

Ramasubbu N., Paloth V., Luo Y., Brayer G. D. and Levine M. J. (1996) Structure of human salivary alpha-amylase at 1.6 A resolution: implications for its role in the oral cavity. *Acta Crystallogr D Biol Crystallogr* **52**, 435-46.

Ramasubbu N., Ragunath C. and Mishra P. J. (2003) Probing the role of a mobile loop in substrate binding and enzyme activity of human salivary amylase. *J Mol Biol* **325**, 1061-76.

Rowe P. A., Campbell-Thompson M. L., Schatz D. A. and Atkinson M. A. (2011) The pancreas in human type 1 diabetes. *Semin Immunopathol* **33**, 29-43.

Rydberg E. H., Li C., Maurus R., Overall C. M., Brayer G. D. and Withers S. G. (2002) Mechanistic analyses of catalysis in human pancreatic alpha-amylase: detailed kinetic and structural studies of mutants of three conserved carboxylic acids. *Biochemistry* **41**, 4492-502.

Samuelson L. C., Wiebauer K., Gumucio D. L. and Meisler M. H. (1988) Expression of the human amylase genes: recent origin of a salivary amylase promoter from an actin pseudogene. *Nucleic Acids Res* **16**, 8261-76.

Samuelson L. C., Wiebauer K., Snow C. M. and Meisler M. H. (1990) Retroviral and pseudogene insertion sites reveal the lineage of human salivary and pancreatic amylase genes from a single gene during primate evolution. *Mol Cell Biol* **10**, 2513-20.

Seyama K., Nukiwa T., Takahashi K., Takahashi H. and Kira S. (1994) Amylase mRNA transcripts in normal tissues and neoplasms: the implication of different expressions of amylase isogenes. *J Cancer Res Clin Oncol* **120**, 213-20.

Sick K. and Nielsen J. T. (1964) Genetics of Amylase Isozymes in the Mouse. *Hereditas* **51**, 291-296.

Ting C. N., Rosenberg M. P., Snow C. M., Samuelson L. C. and Meisler M. H. (1992) Endogenous retroviral sequences are required for tissue-specific expression of a human salivary amylase gene. *Genes Dev* **6**, 1457-65.

Trask B. J., Friedman C., Martin-Gallardo A., Rowen L., Akinbami C., Blankenship J., Collins C., Giorgi D., Iadonato S., Johnson F., Kuo W. L., Massa H., Morrish T., Naylor S., Nguyen O. T., Rouquier S., Smith T., Wong D. J., Youngblom J. and van den Engh G. (1998) Members of the olfactory receptor gene family are contained in large blocks of DNA duplicated polymorphically near the ends of human chromosomes. *Hum Mol Genet* **7**, 13-26.

Tricoli J. V. and Shows T. B. (1984) Regional assignment of human amylase (AMY) to p22----p21 of chromosome 1. *Somat Cell Mol Genet* **10**, 205-10.

Tsuchida S. and Ikemoto S. (1989) Restriction fragment length polymorphism of human salivary alpha-amylase in a Japanese population. *Hum Hered* **39**, 179-81.

van Belle T. L., Coppieters K. T. and von Herrath M. G. (2011) Type 1 diabetes: etiology, immunology, and therapeutic strategies. *Physiol Rev* **91**, 79-118.

Ward J. C., Merritt A. D. and Bixler D. (1971) Human salivary amylase: genetics of electrophoretic variants. *Am J Hum Genet* **23**, 403-9.