The Pennsylvania State University
The Graduate School

STATISTICAL MODELS FOR MAPPING GENES THAT CONTRIBUTE TO SHAPE VARIATION

A Dissertation in
Statistics
by
Guifang Fu

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2012

The dissertation of Guifang Fu was reviewed and approved* by the following:

Rongling Wu
Professor of Public Health Sciences and Statistics
Dissertation Co-Advisor, Co-Chair of Committee

Runze Li
Professor of Statistics and Public Health Sciences
Dissertation Co-Advisor, Co-Chair of Committee

Vernon M Chinchilli
Distinguished Professor of Public Health Sciences and Statistics

Jia Li
Associate Professor of Statistics

Qiang Du
Verne M. Willaman Professor of Mathematics

David Hunter
Professor and Head of the Department of Statistics

*Signatures are on file in the Graduate School.

Abstract

Living things come in all shapes and sizes, from bacteria, plants, and animals to humans. Knowledge about the genetic mechanisms for biological shape has far-reaching implications for a range of spectrum of scientific disciplines including anthropology, agriculture, developmental biology, evolution and biomedicine. Despite the fundamental importance of morphological shape, the difficulty in quantifying the shape and modeling the ultra-high dimension of the image data make the task of genetic mapping on it increasingly difficult.

In this dissertation, we derived several statistical models for mapping specific genes or quantitative trait loci (QTLs) that govern the variation of morphological shape. We are pioneer in the functional shape genetics area and able to detect several significant genes that control the static allometry of the leaf shape traits by incorporating image analysis, statistical model and marker-based linkage disequilibrium (LD) analysis.

After quantifying the morphological shapes numerically through RCC (Radius Centroid Contour) skills, each phenotype, as a datum, is in the form of samples of functional curves or trajectories with high dimension. In the first model, we decreased the dimension by PCA and illustrated the shape variation piece by piece. In the second model, we developed a nonparametric method to model the mean curve by GEE (Generalized Estimating Equation) local polynomial kernel and model the covariance matrix by functional PCA (Principal Component Analysis. Through functional PCA, we characterized the dominant modes of variation around the overall mean trend function and avoided facing directly the extremely huge dimensional covariance matrix. The models are formulated within the mixture framework, in which different types of shape are thought to result from genotypic discrepancies at a QTL. The EM algorithm was implemented to estimate

QTL genotype-specific shapes based on a shape correspondence analysis.

Through incorporating these procedures into the LD based mapping framework, our model led to the detection of several individual significant QTLs responsible for global and local shape variability, addressed many questions in the genetic control of biological shape, and simultaneously estimated QTL allele frequency and marker-QTL linkage disequilibrium. The statistical behavior of the model and its utilization were verified by both real data analysis on the leaf data from China, and computer simulated data.

Table of Contents

List of Figures

List of Tables

Acknowledgments

First of all, I would like to express my deepest appreciation to my advisor, Dr. Rongling Wu, for his wonderful guidance, inspiration, encouragement, patience, and denoted support in so many aspects. Whenever I have problems, Dr. Wu is always there and spare no effort to help. His immense enthusiasm and rich ideas in research has become a good model for me. Especially, he is able to bring very precious projects and data to me. I feel lucky that I have him as my advisor, I pursue a research area that I liked, and I have the opportunity to access the interesting data resources. Second, I would like to deeply thank my co-advisor, Dr. Runze Li, for his wonderful guidance, suggestions, and help. Especially Dr. Li brings the idea of functional data analysis into my project and guides me progress further. Last but not the least, I would also like to deeply express my appreciation to Dr. Vernon M Chinchilli, Dr. Jia Li, and Dr. Qiang Du for their precious time and valuable suggestions in improving my research.

Chapter 1

Introduction

In this dissertation, we tackle the challenge of mapping quantitative trait loci (QTLs) that control the variation of leaf shape traits through shape analysis, statistical model, and marker-based linkage disequilibrium (LD) analysis. Three major advances in life and physical science during the last decades will make it possible to study shape variation and its genetic underpinnings. First, DNA-based molecular markers allow the identification of quantitative trait loci (QTLs) and biochemical pathways that contribute to quantitatively inherited traits such as shape. Second, functional mapping of longitudinal traits such as growth curve, HIV dynamics, programmed cell death, circadian rhythms and pharmacodynamics/pharmacokinetics, constructed by Wu et al. (Wu et al. 2003; 2004a; 2004b; 2004c; 2006; 2007; Wang et al. 2004; Lou et al. 2003) has unearthed high throughput statistical models to locate QTLs that underly quantitative traits. Third, the past two decades have witnessed an increasing interest and development in shape analysis technologies, such as shape acquisition, shape detection, shape representation, shape transforms, shape classification, and shape retrieval (Kendall 1984; Small 1996; Cootes et al. 1995; Belongie et al. 2002; Chang et al. 2002; Kong et al. 2007; Yushkevich et al. 2001; Mcneill 2006; Gower 2004; Stegmann 2002). Although each of these three fields is not new, we are pioneer in integrating them to extend the idea of QTLs to map the genes governing the phenotypic shape trait, which is quite complicated compared to other phenotype traits.

1.1 Motivation

Morphology is one of the most complex physical phenomenon in the world, as all living creature from microbes, bacteria, plants, insects, animals to humans taking an extraordinary diversity of morphological variability. Many biological process, from cellular metabolism, embryonic growth, fruit yield, heartbeat, times of blood circulation, lifespan, to population dynamics, are affected by shape (Wu et al. 2002a). Therefore, knowledge about the genetic mechanisms for biological shape has far-reaching implications for a range spectrum of scientific disciplines including anthropology, agriculture, developmental biology, evolution and biomedicine. For example, it can help us to detect genes that control fruit shape to improve yield and control root shape to improve environment; help us to locate the gene that might cause cancer, as abnormality in organ shape can be related to certain diseases; and help us to address why all organisms persistently attain morphological and anatomical variations in their own respective form, and so on. It has long been known that schizophrenia runs in families. And schizophrenia often caused by the structural changes in the brain. Hence, by locating the genes that control the shape change of the brain might cure the schizophrenia, which is one of the incurable disease in the world.

Despite the fundamental importance of morphological shape, little is known about the detailed genetic mechanisms of shape variation. The motivation of this dissertation is to develop a statistical and computational model for mapping specific QTLs that are responsible for the variation of morphological shape. Historically, genetic mapping has been focused on the numerical phenotypic traits, with an operable dimension, say 1 dimension or at most 30 dimensions. But image data are neither numerical values nor operable dimensions. Generally, image are input in the form of photo, which is saved in the form of huge dimensional matrix. What shape analysis does is to turn photos into numerical values.

This dissertation proposes a new techniques called ***shape mapping*** for linking gene action with key morphometric parameters of a shape within a statistical framework. We will perform both real data and computer simulation to examine the statistical properties of the model. While the models described in this dissertation are belong to the field of statistical genetics, they can be widely applicable to

many scientific areas such as finance, biology, computer vision, pattern recognition, and so on.

1.2 Background and Previous Work

Due to the rapid development of digital and information technologies, image data are more and more easy to be acquired in digital form from varieties of sources. However, there is still a big gap before we know how to take advantage of these visual information. Human's visual system is able to recognize and compare objects by their shapes easily. But, making computer has this visual ability is a hard process. Moreover, development of quantitative methods for describing shapes are even harder. By overcoming these difficulties, shape analysis is a process that recognizes and describes shapes by quantitative nature.

1.2.1 Shape Analysis Background

Shape is the contour or geometrical boundary of an object. Kendall (1984) gives a more accurate definition, by stating that "shape is all the geometrical information that remains when location, scale and rotational effects are filtered out from an object". The procedure that remove location, scale and rotational effects is called alignment or similarity transformation.

The input to shape analysis is a colorful or gray-scale image of a scene containing the objects of interest. In order to manipulate shape variation, there are basically three steps in 2D shape analysis to turn an image into numeric values. First is the **shape representation**. Shape representation establishes a geometric representation of the original shape such as a graph to preserve the important characteristics of the shape. Second is the **shape description**. Shape description generates a set of features from the representation, and these features must be invariant under translation, scale and rotation. Third is the **shape classification**. Shape classification refers to methods for analyzing and comparing shapes.

Shape representation can be roughly divided into two big categories: **region based**, which use all the pixels in an image, and **contour based** which only exploit shape boundary. A very popular contour based representation method is

landmarks, which are a finite set of points on the boundary assigned by either geometrical property such as high curvature or an extremum point, or specific biological meaning to efficiently describe a shape (Cootes et al. 1995; Belongie et al. 2002; Chang et al. 2002). The theory of contour based shape representation has been well established by Kendall (1984) and Small (1996). A very popular region based representation is level set functions, which embed the boundary curve C into the zero level set of a 2D function $z = \phi(x, y)$, and hence represent a shape implicitly (Tsai et al. 2005). The theory of region based shape representation has been well invented by Osher and Sethian (1988).

Each method has its own advantages and disadvantages. So, the choice of shape representation or description depends on the specific problem at hand. Region based method is very time-consuming and inefficient. It need huge space to save a shape, and hard to accomplish any transformation such as scale and rotation. On the contrary, contour based method only need the information of the boundary, and hence obtaining a considerable data reduction without loss of information. Therefore, any numerical implementation can be readily applied. Although contour-based representation is a powerful method that can describe a shape efficiently and accurately, it has several drawbacks. First, is is possible that different images yield different landmark locations and different number of landmark points. To make a good performance of shape analysis, there should one to one correspondence between landmarks of one shape and those of another shape in such a way that the corresponding landmarks are at the same location of the same shape boundaries. Second, landmark method might not be able to deal with the non-convex shapes or high curvature locations. Especially for the shapes whose radii cross the boundary more than once. Third, landmark method has difficulty in handling topological changes, and hard to generalize to 3D shapes. Forth, landmark method might not stable to noise and initial contour placements. But region based method does not have these drawbacks. On the contrary, region based method is able to capture some global properties that might be missed by contour based methods, and can be generalized to 3D case easily.

Different methods work well for different shapes. Hence, it is necessary for us to apply both methods in this dissertation.

1.2.2 Previous Work of Shape Analysis

As is known, shape is one of the most important features. In recent years, significant progress has been made in shape analysis.

1.2.2.1 Contour-based Shape representation

Contour based method is very popular, hence many work has been done. Cootes represent shapes by landmarks and choose points manually to make sure the correspondence problem (Cootes et al. 1995). Abbasi et al. (2000) and Super (2004) use maximum curvature to determine points automatically. Zhang et al. (2003) represent shape by points chosen from the manifold (shape space). i.e. high-dimensional surface, on which different view (transformation, rotation, scale) of the same shape will correspond to a single point. Then they compute geodesic distance to recognize the shape boundary. Bookstein pick the points on the boundary of a shape by some specific biological meaning (Bookstein 1978). Rhodri. proposed a way to automate the choice of landmarks using the minimum description length criterion (Davies et al. 2001; 2002). Although they can avoid the manually procedure, this method is very time consuming. Staib et al. (1992) and Szekely represent boundaries of the shape as a weighted sum of Fourier basis functions and consider the weights as the interest. Golland et al. (1999) represent a shape by a fixed topology skeleton, which is to describe a shape by the width and curvature along the medial axis. Yushkevich et al. (2001) describe shapes using a multiscale medial representation. They use coarse-scale for entire object and fine-scale for part of the shape. Tan et al. (2000) use centroid-radii model to represent shapes. However, it cannot deal with the non-convex shapes. i.e. whenever the radii cross the boundary more than once, the algorithm will fail to represent the boundary correctly. Kong et al. (2007) overcome this shortcoming by combining centroid-raddi model with Haar wavelet transform, and expand the representation to any kind of shape, no matter convex or not. Freeman chain code approximates a curve with a sequence of directional vectors lying on a square grid. But, it is very sensitive to noise. Sun et al. (2006) propose CCCV (chain code coherence vector) and CCDV (Chain code distribution vector) to overcome the shortcomings of Freeman chain code, and hence invariant to translation, rotation and scaling. Dubois et

al. (1986) use autoregressive model to perimetrically express the landmark points equispaced angularly obtained from the boundary, since these points are spatially correlated. Chuang et al. (1996) developed a planar curve descriptor that has a multiscale analysis capability by using the wavelet transform. Yadav et al. (2007) compared three descriptor techniques: FD (Fourier descriptors), GFD (generic Fourier descriptors), and WFD (Wavelet-Fourier descriptors), and conclude that WFD performs best among the three. Greenander et al. (1993) proposed a general stochastic shape model to characterize the random shape variability among objects by using Bayesian formulation.

1.2.2.2 Region-based Shape representation

The level set method represent regions and set of interfaces with a continuous function defined over every pixel of the whole image. So, it is called the region-based method. Using level set function to represent implicitly a boundary of a shape was first proposed by Osher and Sethian (1988). By evolving a higher-dimensional embedding function, they can propagate the boundary points in the 2D plane. As early as 90's, Malladi, Caselles, Kichenassamy, and Deriche started to use level set function to do image segmentation. In recent years, a lot of work have been developed. Tsai et al. (2003) adopt the level set method and propose a shape based approach to do segmentation. After representing the shape boundary by signed distance function of each pixel, Tsai et al. (2005) incorporates the level set method within the framework of EM algorithm to do classification. Samson et al. (2000) also apply level set to represent the shape implicitly, then they assume different classes own different level sets and use optimal partition to do classification by minimizing a unique functional.

Besides above level set method, moment based shape representation is one of the earliest and most popular region based representations. It provides a numerical similarity measurement that is invariant to translation, scale, and rotation. The moment representation describe a shape by defining the gray level image function as a probability density of 2D random variable. Flusser (2000) use geometrical moments, Teague (1980) and Khotanzad (1990) use zernike moments, Mukundan et al. (2001) use Tchebichef moments.

Zhang et al. (2002) proposed a generic Fourier descriptor (GFD) by apply-

ing 2D Fourier transform on a polar raster sampled shape image. GFD is the modified Fourier transform to treat the polar image in polar space as a normal two dimensional rectangular image in Cartesian space. Goshtasby (1985) develop a polar quantization matrix by considering not only the outer geometry but also inner geometry. Lu et al. (1999) suggested a grid based approach that is simple and intuitive.

1.3 Basic Genetics and Previous Work

It is well known that many things including IQ, height, characters, and some disease such as schizophrenia, hypertension, and diabetes are easily inherited from parent to offspring. Hence, locating the genes that regulate all kinds of phenomena has become a very important field nowadays.

1.3.1 Basic Genetics

Genes are pairwise units by which the biological characteristics can transmit unchangeably from parents to offspring. If a pair have similar genes, it is called homozygous. Otherwise, it is called heterozygous. For example genes named A and a, then AA and aa are homozygous, and Aa is heterozygous. These alternative genes are named alleles. It is easy to understand that a single pair of alleles can make three possible genotypes AA, Aa, and aa. Chromosome is the microscopic body that genes located in some specific order. The location that a specific gene lies is called locus. Since both genes and chromosomes come in pairs, we call them loci. The final goal of QTLs mapping is to locate the genes that affecting some phenotypic traits, which is described by quantitative values.

Gamete (ova and spermatozoa) is the reproductive cells by which only one chromosome from each parent passes one gene. Fertilization is a process when a sperm carrying one gene from the father integrating with an ovum carrying one gene from the mother to complete one pair. This makes the new offsprings owning one gene from mother and the other from father in each body cell of them. The number of chromosomes in a gamete is called haploid, and that in a fertilized zygote is called diploid. The Mendel's first law says that if a parent has genotype Aa, then

either A or a has the probability of 0.5 to be passed into the gamete. Assume we cross two individuals with AA and aa, then all zygotes in F1 (first generation) will be Aa. If continue crossing two individuals of F1, then three possible genotypes AA, Aa, aa will be generated in F2 with the ratio of 1:2:1. Backcross is the process that one individual from F1 carrying Aa cross over with one homozygous parent carrying either AA or aa. Mendel's second law states that different pairs of genes segregate independently and do not affect each other. For the genes in different chromosomes, this law might be true. However, for the genes located in the same chromosome, it is high likely that they are correlated to each other. Considering the relation between neighbor genes is what linkage analysis does. As a matter of fact, the segregating rules in real life often violate the Mendel's first and second laws. Hence, Linkage Disequilibrium (LD) analysis is used in this dissertation.

Consider a gene with alleles A (with probability p_0) and a (with probability p_1). From above, we know that the individual in F2 will have three genotypes AA,Aa,aa. Assume their population frequencies are P_2, P_1, and P_0, respectively. Hardy-Weinberg law says that, if the individuals mated with each other randomly, then the Hardy-Weinberg equilibrium

$$P_1^2 = 4P_2P_0 \tag{1.1}$$

is always hold for any generation. See (Wu et al. 2006) for more details.

In later chapters, we will describe in detail the models for LD, and backcross.

1.3.2 Previous Work of QTLs Mapping on Shape

As early as the beginning of this century, Hedrick et al. (1907) and Price et al. (1908) started genetic map on tomato fruit shape. However, the relationship between morphological shape and genes are poorly understand during the past 100 years, since almost all focus measured shape mainly by some simple scale such as length or width (Grandillo et al. 1999). Currence (1934) found that locus *o* in chromosome 2 controls the relative length of tomato fruit got by length/diameter (Zygier et al. 2005). Young et al. (1947) detected gene *f* on chromosome 11 and *lc* on chromosome 2 control the fruit shape by representing the tomato fruit shape by the locule number. Grandillo et al. (1996) reported *fs8.1* that controls

tomato fruit shape by using the ratio of longitudinal diameter (L) and equatorial diameter (D) as phenotypic traits (Ku et al. 2000). Jiang et al. (2000) did a more advanced work by recording 14 measures (lobe numbers, main-lobe length and width, second-lobe length and width, et. al) to describe a leaf shape. Fulton et al. (1997) detected 16 QTLs that control tomato fruit shape by denoting 1 for round tomato shape, and denoting 2 for elongated tomato shape. In his seminal review, Tanksley (2004) summarized some major discoveries of genes for fruit size and shape in tomato. In a long process of domestication, tremendous shape variation has occurred in tomato fruit from almost invariably round (wild or semiwild types) to round, oblate, pear-shaped, torpedo-shaped, and bell pepper-shaped (cultivated types). Some of the QTLs that cause these differences, namely *fw2.2*, *ovate*, and *sun*, have been cloned (Fray et al. 2000; Liu et al. 2001; Xiao et al. 2008).

While these above past work once brought a great breakthrough to this area, they failed to describe allometry accurately and thoroughly, since two objects with totally different shape, say a circle and a diamond, can take exactly the same mass, length, and even surface area. Unlike roughly computing the ratio of size such as width or height, shape variability analysis is the best way to describe allometry meticulously, as it can measure not only size, width, or any above scale but also all kinds of unobservable morphological value. Despite its powerful skills, there is, so far as we know, few literature can be found about mapping genes in morphology using shape analysis. What aggravate the complexity is that the phenotypic trait of each shape is not traditional number but photo or picture.

So far in the literature, there is only one paper (Langlade et al. 2005) that is a little close to our present work. Langlade et al. also used shape analysis skills to map the genes that control allometry of leaf shape. Their work is much accurate than above simply measuring shape by simple scale such as length or width (Whitfield 2001; Enquist et al. 1998; 1999; West et al. 1997; 1999a; 1999b; 2008). However, our work have five main benefits and differences. First, Langlade et al. use interval mapping to locate the genes that control **evolutionary allometry** in leaf shape for 18 different antirrhinum species. Instead, we apply linkage disequilibrium to map the genes that affect **static allometry** in leaf shape for different individual trees from the same poplar species. Second, Langlade et al. roughly connected 19 points for all different leaves to represent their boundaries. But, us-

ing only 19 fixed points for all leaves is not a forceful way to represent any shape boundaries, especially for those with complicated and non-smooth outlines (For example Fig. 2.1. Leaf LS29-2). Instead, we use radius centroid contour (RCC) to describe each leaf shape in a very meticulous way. Third, Langlade et al. put all middle veins horizontal to filter rotation effect. However, The middle veins of different leaves are not always parallel and straight (For example Fig. 2.1. Leaf LS17-1), and hence alignment using middle veins is not a very persuadable and accurate way. Instead, we use procrustes analysis to align leaves systematically and automatically by model. Fourth, Langlade et al. only capture the biggest shape variability by PCA. We not only capture the global shape variability, but also catch the local minor shape variability. Moreover, we also transform back from reduced space to original image domain, and show the different effect of genotype in both image domain and vector domain. Fifth, our work is unique since the poplar trees are planted in different elevation, longitude and latitude of Tibet of China. Hence, it can at least illustrate a genetic property in a unique location of the earth.

1.4 Overview of Chapters

The remainder of this dissertation is organized as follows:

In Chapter 2, we tackle the challenge of mapping quantitative trait loci (QTLs) that control the variation of leaf shape traits through contour-based shape analysis, statistical model, and marker-based linkage disequilibrium (LD) analysis. The model is validated by analyzing a mapping data collected from two different natural populations of poplar, and identifying several QTLs for leaf shape in this species.

In Chapter 3, simulated data is used to test the power of the statistical model. In this experiment, to make it simple, we make many assumptions such as independent pixels and backcross markers, etc. We use region-based shape analysis here because the leaves are not all convex.

In Chapter 4, we developed a nonparametric smoothing method to model the mean curve by GEE (Generalized Estimating Equation) local polynomial kernel and model the covariance matrix by functional PCA (Principal Component Analysis). Through this model, we estimate both the mean and covariance as the

function of spatial angle, characterize the dominant modes of variation around the overall mean trend function, and hence avoid facing directly the extremely huge dimensional covariance matrix.

Chapter 5 concludes with a discussion of the contributions of this dissertation and possible future work.

Chapter 2

Mapping Shape QTLs Using a Radius-Centroid-Contour Model

In this chapter, we will explain in detail the contour based shape analysis, genetic design, and the statistical model. To the end, the accurate and quantitative representation of a shape is produced with aligned Radius-Centroid-Contour (RCC) curves, i.e., a function of radial angle at the centroid. The high dimensionality of the RCC data, crucial for a comprehensive description of the geometric feature of a shape, is reduced by principal component (PC) analysis, and the resulting PC axes are treated as phenotypic traits, allowing specific QTLs for global and local shape variability to be mapped, respectively. The usefulness and utilization of the new model for shape mapping in practice are validated by analyzing the mapping data collected from two natural populations of poplar, and identifying several QTLs for leaf shape in this species. The model provides a powerful tool to compute which genes determine biological shape in plants, animals and humans.

2.1 Introduction

Tremendous variation in morphological shape provides a fuel for the evolution of biological function and the formation of new species that best adapt to a specific environment (Albertson et al. 2005; Klingenberg 2010; Klingenberg et al. 2012). Genes are thought to play an important role in controlling phenotypic variation in shape; according to quantitative genetic analyses in animals, shape may have

a heritability of 0.60 - 0.70 (Klingenberg and Leamy 2001; Monteiro et al. 2002; Klingenberg 2003; Mezey and Houle 2005; Gilchrist and Crisafulli 2006). With the development of genotyping techniques, genetic mapping that dissects phenotypic variation into individual quantitative trait loci (QTLs) (Lander and Botstein 1989) has been used to detect specific QTLs for morphological shape in mice and Drosophila, providing many promising results (Klingenberg et al. 2001; 2004; Leamy et al. 2008; Weber et al. 1999; Mezey et al. 2005). More recently, Fu et al. (2010) developed a binary model for shape mapping based on computer-simulated black and white shape data. Langlade et al. (2005) used 19 representative points for a leaf to map the QTLs that control the allometry of leaf shape and pioneered the integration of shape QTLs with interspecific divergence and evolution.

Many of these shape genetic studies are based on a simple geometric analysis and, thus, do not intend to resolve the inherently complicated structure of a biological shape. For example, simple morphological measures for length, width, height, ratio, and angle do not separate size and shape clearly (Rohlf and Marcus 1993), although these two aspects perform different biological functions. In addition, some more advanced genetic analysis of shape mostly focus on drastic morphological changes, but do not allow a quantitative description of detailed structures of organ morphology, such as leaf margins that can be entire, serrated, or lobed (reviewed in Klingenberg 2010).

As an important approach for shape analysis, geometric morphometrics (GM) has a capacity to quantify each piece of subtle variation that accumulatively contributes to shape (Klingenberg 2010). By analyzing the polar coordinates of anatomical landmarks, shape analysis based on the GM model retains geometric information from digitized data and relates abstract, multivariate results to the physical structure of the original specimens (Adams et al. 2004; Slice 2007). The development of image and digital technologies has greatly facilitated shape recognition and shape registration based on the theory that a shape can be represented by a number of carefully selected and coded image patches extracted from images taken from different view-points (Belongie et al. 2002). The recent years have seen the development of new technologies used to analyze and interpret the molecular, mechanical and dynamic mechanisms that form shape (Nath et al. 2003; Rolland-Lagan et al. 2003; Coen et al. 2004). Coen and colleagues used clonal analysis

techniques to study the dynamic relationship between gene expression pattern and leaf shape (Rolland-Lagan et al. 2005). Liang and Mahadevan (2009) capitalized on a combination of scaling, stability and asymptotic analysis to quantify leaf shape and the conditions that cause different morphologies of leaves. As a first step of our shape gene identification project, here we develop a model for studying the genetic mechanisms of morphological shape by mapping specific quantitative trait loci (QTLs) involved in shape variation. This model integrates existing GM analysis into a framework for QTL mapping through a series of statistical bridges. By measuring radii from the centroid to the contour at regular intervals, we quantify the geometric features of a shape and further use a procrustes analysis to align shapes with different poses, scales and rotations. The high dimension of shape data measured by a Radius-Centroid-Contour (RCC) analysis is reduced by principal component (PC) analysis producing orthogonal PC axes that capture global and local variability, respectively. Based on the PC axes of RCC values, a QTL mapping model is derived and then the QTL effects detected on shape structure are transformed back to image domains in order to intuitively visualize how QTLs affect shape variation. To demonstrate the utility and usefulness of the new model, we used it to analyze a mapping population of a poplar species, leading to the detection of several significant QTLs that govern leaf shape. The new model combines the strengths from genetic mapping and shape analysis, providing a powerful tool for the genome-wide identification of QTLs with varying sizes of genetic effects on shape diversity.

2.2 Contour-Based Shape Analysis

The theory of shape analysis has well been established by Kendall (1984), in which a finite number of landmarks are used to represent a shape of an object. According to Kendall's definition, "shape is all the geometrical information that remains when location, scale and rotational effects are filtered out from an object." Here we integrate this theory into the genetic mapping framework by which to characterize the structural, functional, and developmental features of shape.

To capture the complicated structure of a shape, we used a high dimension of pixels to describe its boundary and detailed inner feature. A vector of representa-

tion for the shape can be denoted as coordinates $(x(s), y(s))(s = 0, 1, \ldots, m-1)$ extracted from a digital image, where m is the number of coordinates, determining the accuracy of shape representation. Below are steps for shape analysis with digital images.

During this shape representation, we divide the procedures into four steps: shape compression to decrease the original resolution to an operable size; shape detection to distinguish the background from the object; shape alignment to minimize the variation caused by location, scale and rotation; and shape description to describe a shape using numerical vectors.

2.2.1 Shape Detection

After decrease the dimension to 150 by 225, we need to recognize the leaf from the background. Since each pixel of color image is saved in RGB (Red, Green, Blue) value, we notice that the object (leaf here) has very high green value and the background in purple color happen to has very high red value. By applying a simple threshold on both R and G value, we successfully convert the color image into a binary image. At each pixel, we use 0 to denote the background (black) and 1 to denote the object (white). Finally we complete the preprocessing procedure through de-noising and removing all isolated segments. Once a black and white image is obtained, the shape boundary can be easily detected. The vector that represents a shape can be denoted as $(x(s), y(s)), s = 0, 1, \ldots, m-1$. Here m determines the accuracy of the shape representation, the larger m is, the more details of the shape information can be kept, and consequently, the shape analysis will be more accurate. We use 360 points to finely describe the boundary.

2.2.2 Shape Alignment

All shapes need to be aligned, in order to minimize variation caused by pose. Shape alignment is a process by which to establish a coordinate reference for all shapes with respect to position, scale and rotation, commonly known as pose. An orthogonal procrustes analysis is used to undertake this alignment (Gower et al. 2004).

To make shape representation invariant to translation, we shift all shapes to

their centroids by

$$(x_1(s), y_1(s)) \quad = \quad (x(s) - x_c, y(s) - y_c), \tag{2.1}$$

where (x_c, y_c) is the centroid of the shape, which is defined as

$$x_c = \frac{1}{m}\sum_{s=0}^{m-1} x(s), y_c = \frac{1}{m}\sum_{s=0}^{m-1} y(s). \tag{2.2}$$

By using the new coordinate system $(x_1(s), y_1(s))$, all shapes have the origin at the centroid and thus eliminate any influence caused by position.

To filter a scale effect, we normalized all shapes by dividing each shape by its Euclidean or Frobenius norm, which produces the normalized shape:

$$(x_2(s), y_2(s)) \quad = \quad \frac{(x_1(s), y_1(s))}{||(x_1(s), y_1(s))||}. \tag{2.3}$$

The last and most complicated step for shape alignment is to remove the rotation effect. The idea behind is to rotate each shape one by one so that they can be close to a reference shape as much as possible. We use the Euclidean or Frobenius norm to measure the distance between two shapes. The smaller the Euclidean norm, the closer they are. In addition, the average of all shapes is used as the reference shape (denoted as $\overline{Z}$). Now, we assume

$$Q = \begin{pmatrix} cos(\theta) & -sin(\theta) \\ sin(\theta) & cos(\theta) \end{pmatrix},$$

is the rotation matrix, by multiplying which on the right hand side of (2.3), the shape get rotated θ angle clockwise. Denote Z as $(x_2(s), y_2(s))$. By definition, we hope to solve Q by minimizing

$$||ZQ - \bar{Z}||.$$

Since

$$||ZQ - \bar{Z}|| \quad = \quad trace(Q^T Z^T ZQ + \bar{Z}^T \bar{Z}) - 2trace(\bar{Z}^T ZQ),$$

$$= \quad trace(Z^T Z + \bar{Z}^T \bar{Z}) - 2trace(\bar{Z}^T ZQ),$$

where the first part does not contain Q so that we only need to maximize the trace of $\bar{Z}^T ZQ$. By singular value decomposition, there exist an orthogonal matrix U and V, and diagonal matrix D such that $\bar{Z}^T Z = UDV^T$. Hence

$$\begin{aligned} trace(\bar{Z}^T ZQ) &= trace(UDV^T Q), \\ &= trace(DV^T QU), \\ &= trace(DH), \\ &= \sum_{i=1}^{p} (d_i h_{ii}). \end{aligned}$$

where $H = V^T QU$ is an orthogonal matrix, d_l is the lth diagonal element of diagonal matrix D, and h_{ll} is the lth diagonal element of H. Therefore, $trace(\bar{Z}^T ZQ)$ is maximized when $H = I$. This is equivalent to $Q = VU^T$.

It can be seen from the above derivation that we should multiply the right-hand side of $(x_2(s), y_2(s))$ by VU^T to rotate a shape to be closed to the average of all shapes. The three steps described above are repeated and iterated until the rotated shapes provide the best fit of differences among all shapes caused by pose. We use $(\tilde{x}(s), \tilde{y}(s)), s = 0, 1, \ldots, m-1$ to denote final coordinates of each shape after alignment.

2.2.3 Shape Representation

As a popular contour based method, we use landmarks for shape representation. Landmarks are a set of points on the boundary assigned by either geometrical property (such as high curvature), or an extremum point, or specific biological meaning (Cootes et al. 1995; Belongie et al. 2002). To make a one to one correspondence between landmarks of one shape and all other shapes, we choose the same angle or the same arc length. We select points on the boundary spaced at equal radial angle $\theta = 2\pi/m$, where m is the number of points. This gives an accurate and robust description of shape. A shape can be described by Radius-Centroid-Contour (RCC) values (Belongie et al. 2002), i.e.,

$$r(s) = (\tilde{x}(s)^2 + \tilde{y}(s)^2)^{1/2}. \tag{2.4}$$

which are used for QTL mapping.

2.3 Statistical Design

2.3.1 Likelihood

A segregating population is prerequisite for mapping trait QTLs. Consider a natural population from which a sample of n individuals is drawn randomly. All these individuals are genotyped for a panel of molecular markers. Meanwhile, the shape of an organ, such as leaf, is measured for each individual by taking a photo of representative leaves. It is likely that a set of QTLs controls shape, forming a total of J genotypes. Although we cannot observe these QTL genotypes directly, they can be inferred from the markers (M) that are linked to the QTLs. For this reason, a basic statistical model for QTL mapping is a mixture model, in which each observation Y is assumed to have arisen from one of the J QTL genotypes, each genotype (j) being modeled from a density function (frequently a normal distribution is assumed). Thus, the likelihood of Y is expressed as

$$L(\omega, \phi, \eta | Y, M) = \prod_{i=1}^{n} \sum_{j=1}^{J} \omega_{j|i} f_j(Y_i | \phi_j, \eta_i), \tag{2.5}$$

where ω is composed of mixture proportions $\omega_{j|i}$ of individual i carrying a QTL genotype j, ϕ_j is the expectation parameter vector specific to a QTL genotype j, and η_i is the variance-covariance parameter common to all genotype groups, and $f_j(Y_i|\phi_j, \eta_i)$ is the probability density function of observations for individual i at QTL genotype j. For a natural population, the mixture proportions ($\omega_{j|i}$) of each QTL genotype j in likelihood (2.5) are described in terms of allele frequencies at the markers and QTLs and their linkage disequilibria (LD) (Wang et al. 2004). The size of LD reflects the degree to which the markers and QTLs are associated.

2.3.2 Dimension Reduction

Since the dimension of the RCC values is still too high to be handled, the dimension reduction skills need to be applied to obtain the MLEs from likelihood (2.5). Many approaches can be used to decompose the original m-dimensional space to a space of reduced dimension. Principal component analysis (PCA) is one of such powerful approaches by removing redundant information through mapping the high dimensional data to the subspace that best accounts for the distribution of the original pattern. Denote n shape data by $R = \{r_1, r_2, \ldots, r_n\}$ in the R^m space, where r_i is the RCC curve of the ith leaf shape with length m. The average of these data is defined by

$$\mu = \frac{1}{n}\sum_{i=1}^{n} r_i,$$

and the MLE of variance can be given by $\Sigma_R = \frac{1}{n}\sum_{i=1}^{n}(r_i - \mu)(r_i - \mu)^T$. Let $X = \{r_1 - \mu, r_2 - \mu, \ldots, r_n - \mu\}$, then we have $\Sigma_R = XX^T$, a $m \times m$ matrix, which is too big to be manipulated practically. The main idea behind PCA is to maximize the variance by finding a certain number of orthogonal axes, called principal components (PCs), that is much fewer than m. Therefore, through PCA, we can use $Y_i = v_k^T X_i^T X_i$, where v_k $(k = 1, \ldots, K)$ is the eigenvector of $X^T X$ in terms of the kth PC, to model the likelihood (2.5). The first K largest PCs are chosen. Next, we will describe a procedure for linkage disequilibrium mapping of QTLs using these PC values (Wang et al. 2004).

2.3.3 Linkage Disequilibrium Mapping

To map QTLs in a natural population, we need to implement linkage disequilibrium as a parameter that link markers with QTLs. For clarity of model description, we assume one QTL controlling a shape which is associated with a marker, with two alleles M (with a probability p) and m (with a probability $1-p$), through a linkage disequilibrium $\mathfrak{D}$. At the shape QTL, there are two alleles A (with a probability q) and a (with a probability $1 - q$) that form three genotypes, expressed as AA (denoted as 1), Aa (denoted as 2), and aa (denoted as 3). The marker and QTL form four haplotypes MA, Ma, mA, and ma, with the frequencies denoted as $p_{11} = pq + \mathfrak{D}, p_{10} = p(1-q) - \mathfrak{D}, p_{01} = (1-p)q - \mathfrak{D}$, and $p_{00} = (1-p)(1-q) + \mathfrak{D}$,

respectively, where $\max(-pq, -(1-p)(1-q)) \leq D \leq \min(p(1-q), (1-p)q)$. The haplotypes from maternal and paternal parents unite randomly to generate nine marker-QTL genotypes.

2.3.4 Modeling Mixture Proportions

The conditional probabilities of a given QTL genotype, conditional upon a marker genotype for individual j, expressed as $\omega_{j|i}$ in the likelihood (2.5), can be calculated (Wang et al. 2004). The observations of three genotypes at the marker are denoted as $n1$ for MM, $n2$ for Mm, and $n3$ for mm.

In a natural population at HWE, the frequency of a joint marker and QTL diplotype can be expressed as the product of the frequencies of the two haplotypes derived from different parents that constitute the diplotype. For the genotypes which are homozygous at one or two loci, the diplotype frequency is the same as the genotype frequency (Table 2.1). The double heterozygote $AaMm$ contains two possible diplotypes $AM|am$ and $Am|aM$, where the haplotypes derived from maternal and paternal parents are separated by the vertical lines. Thus, the total frequency of genotype $AaMm$ is the sum of the frequencies of these two diplotypes. Since unobservable QTL genotypes can be inferred from observed marker genotypes due to the marker-QTL association, we will derive the the conditional probability of a QTL genotype (AA, Aa, aa) given a marker genotype (MM, Mm, mm) using the joint genotype frequencies from Table 2.1.

Table 2.1. Joint genotype frequencies at the marker and QTL

	AA	Aa	aa
MM	p_{11}^2	$2p_{11}p_{10}$	p_{10}^2
Mm	$2p_{11}p_{01}$	$2p_{11}p_{00} + 2p_{10}p_{01}$	$2p_{10}p_{00}$
mm	p_{01}^2	$2p_{01}p_{00}$	p_{00}^2

2.3.5 Parameter Estimation

The parameters that define the likelihood (2.5) are obtained by differentiating the likelihood with respect to each parameter, letting the derivative equal to zero, and then solving the log-likelihood equations. We implemented the EM algorithm

to estimate the parameters. The E step is designed to calculate the posterior probability with which subject i carries QTL genotype j given its marker and phenotypic information, expressed as

$$\Omega_{ij} = \frac{\omega_{j|i} f_j(Y_i)}{\sum_{j'=1}^{3} \omega_{j'|i} f_{j'}(Y_i)}. \tag{2.6}$$

Using the calculated posterior probabilities, the M step is derived to solve the haplotype frequencies expressed as

$$\begin{aligned}
\mu_j &= \frac{\sum_{i=1}^{n}(\Omega_{ij} * Y_i)}{\sum_{i=1}^{n} \Omega_{ij}}, \quad \forall j = 1, 2, 3 \\
\sigma^2 &= \frac{1}{n} \sum_{i=1}^{n} \left[\Omega_{i1}(Y_i - \mu_1)^2 + \Omega_{i2}(Y_i - \mu_2)^2 + \Omega_{i3}(Y_i - \mu_3)^2 \right], \\
\hat{p}_{11} &= \frac{1}{2n} \left[\sum_{i=1}^{n_1} (2\Omega_{i1} + \Omega_{i2}) + \sum_{i=1}^{n_2} (\Omega_{i1} + \theta \Omega_{i2}) \right], \\
\hat{p}_{10} &= \frac{1}{2n} \left[\sum_{i=1}^{n_1} (\Omega_{i2} + 2\Omega_{i3}) + \sum_{i=1}^{n_2} (\Omega_{i3} + (1-\theta) \Omega_{i2}) \right], \\
\hat{p}_{01} &= \frac{1}{2n} \left[\sum_{i=1}^{n_3} (2\Omega_{i1} + \Omega_{i2}) + \sum_{i=1}^{n_2} (\Omega_{i1} + (1-\theta) \Omega_{i2}) \right], \\
\hat{p}_{00} &= \frac{1}{2n} \left[\sum_{i=1}^{n_3} (\Omega_{i2} + 2\Omega_{i1}) + \sum_{i=1}^{n_2} (\Omega_{i3} + \theta \Omega_{i2}) \right],
\end{aligned} \tag{2.7}$$

where $\theta = p_{11}p_{00}/(p_{11}p_{00} + p_{10}p_{01})$. The iteration are repeated between including equation (2.6) and equations (2.7) until the estimates converge to stable values. These stable values are the maximum likelihood estimates (MLEs) of parameters.

2.3.6 Hypothesis Tests

Based on likelihood (2.5), the significance of a shape QTL can be tested by using the following hypotheses:

$$\begin{aligned} H_0: &\quad \mu_{\mathbf{j}} = \mu, \ (j = 1, 2, 3) \\ H_1: &\quad \textit{At least one of the equalities above does not hold.} \end{aligned} \tag{2.8}$$

where the H_0 corresponds to the reduced model, in which the data can be fit by a single shape, and the H_1 corresponds to the full model, in which three QTL genotype-specific shapes exist to fit these data. The log-likelihood ratio (LR) of the full to reduced model is calculated as the test statistics for the above hypotheses. An empirical approach based on permutation tests is used to determine the critical threshold (Churchill and Doerge 1994). The significance level was further corrected for multiple comparisons using Bonferroni's criterion.

After a significant QTL is found to exist, we need to test whether this QTL can be detected by a given marker using the hypotheses:

$$\begin{aligned} H_0: &\quad \mathfrak{D} = \mathbf{0}, \\ H_1: &\quad \mathfrak{D} \neq \mathbf{0}, \end{aligned} \tag{2.9}$$

where the H_0 corresponds to the reduced model, in which the marker and QTL are at the linkage equilibrium, and the H_1 corresponds to the full model, in which there is a linkage disequilibrium between the marker and QTL. The test statistics for this hypothesis is calculated as $\chi^2 = 2nD^2/[p(1-p)q(1-q)]$, which is χ^2-distributed with one degree of freedom. The significance level was corrected for multiple comparisons using Bonferroni's criterion.

2.4 Examples

2.4.1 *Populus szechuanica var. tibetica*

The new model was used to analyze leaf shape data for a mapping population of poplar, *Populus szechuanica var. tibetica*. Belonging to the Tacamahaca section, *P. szechuanica* is naturally distributed throughout the Tibet Plateau, growing in

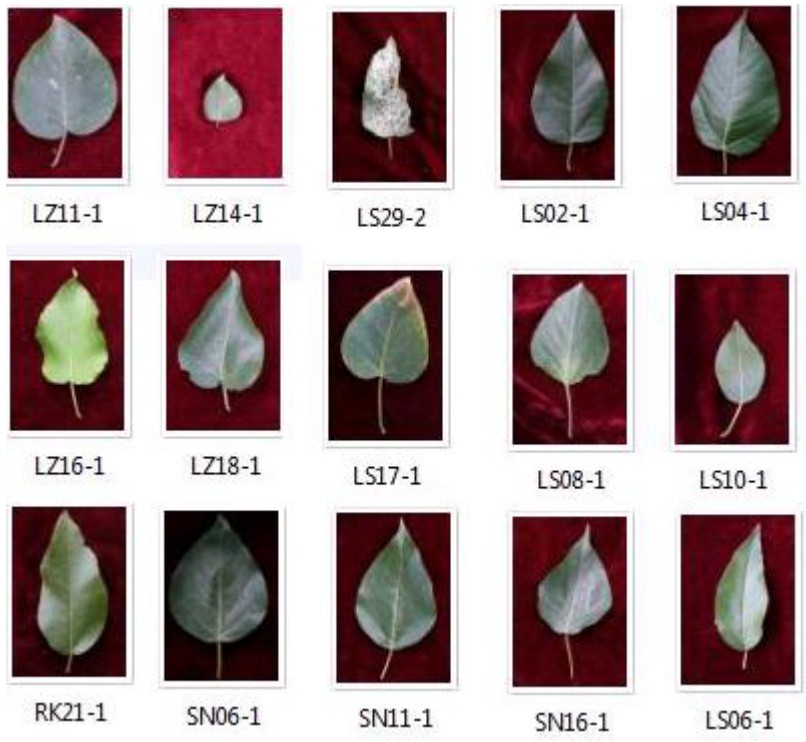

Figure 2.1. A set of original leaf images (with IDs given at the bottoms) chosen from the mapping population for *Populus szechuanica var. tibetica*, showing pronounced variation in leaf shape.

mountains at an altitude of 1100 - 4600 m over a wide range of regions in Gansu, Shaanxi, Sichuan, Xizang, and Yunnan Provinces of China (Hamzeh et al. 2004). Its wide ecological adaptation of this species, along with its pronounced variation in leaf size and shape (Fig. 2.1), makes this species ideal to study the genetic variation of leaf morphology using molecular markers. The overall shape of leaf blade in *Populus szechuanica var. tibetica* varies markedly from broadly ovate to ovate-orbicular to ovate-lanceolate. The bases of leaf blades are rounded, cuneate, or shallowly cordate with glandular dentate margins at the first ciliate. Leaves grow from short branchlets with petioles 2.5 - 8 cm. A precise shape analysis approach is needed to identify and quantify such a diversity of leaf shape.

Langlade et al. (2005) pioneered a numerical analysis for shape variation in leaves by placing 19 key landmarks on the leaf margin and leaf mid vein from digital images. However, joining these 19 points with straight lines can only capture the global feature of a leaf outline. The choice of sparse anatomical landmarks by this

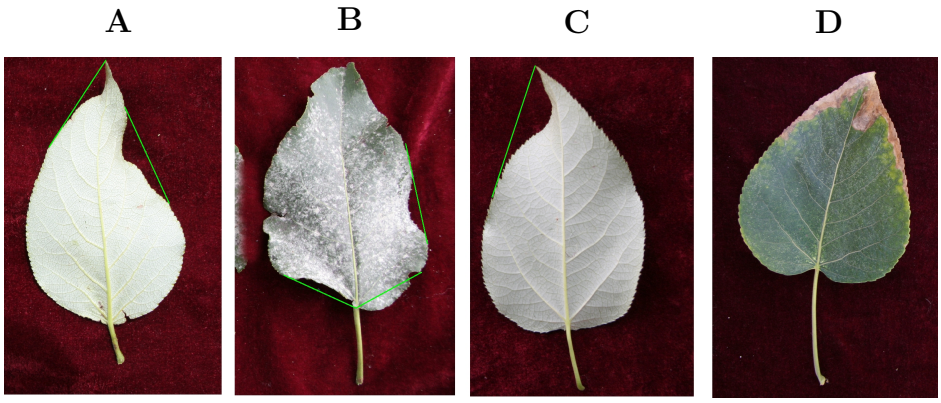

Figure 2.2. Four typical leaf shapes detected from the mapping population. In A, B and C, leaf margins are not always smoothly curved, as shown by green lines, which makes it difficult to determine anatomical landmarks on the leaf outlines using traditional approaches. In D, the mid-vein is crooked, which cannot be used as a reference to align leaf shapes.

approach is extremely difficult when some leaves (see examples in Fig. 2.2 A,B,C) are abruptly curved. This part of leaf shape variation may be linked with some particular ecological function (Kessler and Sinha 2004) and, therefore, should be taken into account. Furthermore, in Langlade et al. (2005), a straight mid-vein was used to align leaf shapes (see their Fig. 2.2). In our example, however, many leaves display a curved mid-vein (Fig. 2.2 D), making it difficult to align shapes using the mid-vein as a reference.

As a pilot study of shape mapping, we selected 107 trees randomly from a natural population of *Populus szechuanica* var. *tibetica* and from each tree three representative leaves were sampled to take photos. The sampled trees were genotyped for 29 microsatellite markers to be used to detect leaf shape QTLs. By reading 600×900 pixels from a leaf digital image, we obtained three matrices for red, green and black colors that discern the object and background (Fig. 2.3 A), from which binary smaller matrix was generated to capture the leaf shape by recording its contour (Fig. 2.3 B). Using the procedure for shape alignment described in METHODS, we obtained a vector of 360 coordinates $(\tilde{x}(s), \tilde{y}(s))(s = 0, 1, \ldots, 359)$ to represent leaf shape. It turns out that 360 points can well describe the leaf boundary (Fig. 2.3 C,D). Fig. 2.4 is the diagrammatic representation of several key steps (A, B, C, and D) described in Fig. 2.3. The 360 representative points shown in a vector (Fig. 2.3 D) can be actually expressed as a Radius-Centroid-

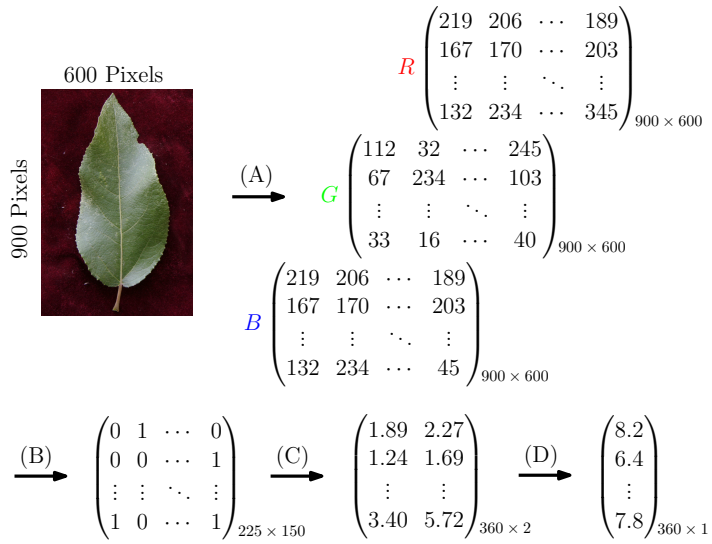

Figure 2.3. The procedure of extracting leaf-shape information from a leaf image. In A, the leaf is read by 900×600 pixels based on different colors, red (R), green (G) and black (B) for the object and background. In B, the leaf outline is read as a 1/0 binary variable with a dimension-reduced matrix. In C, the Cartesian coordinates of points on the leaf outline are calculated. In D, all coordinates in C are expressed as single Radius-Centroid-Contour (RCC) values.

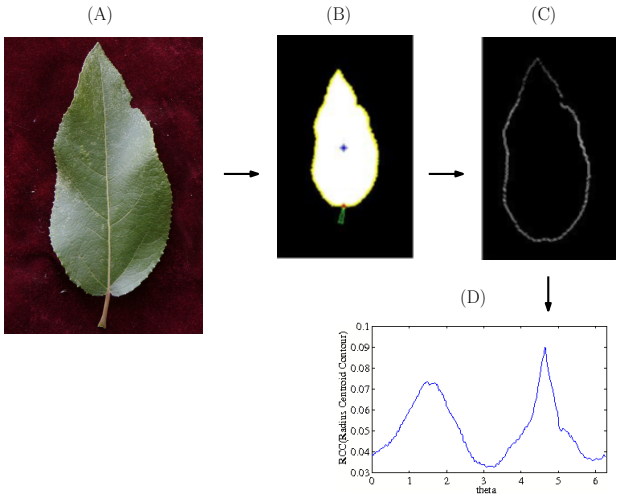

Figure 2.4. Diagrammatic representation of the extracting procedure described in Fig. 2.3 A - D in this figure correspond to those in Fig. 2.3, respectively. The vector of RCC values in Fig. 2.3D is expressed as a curve which is a function of radial angle (θ) (see the text).

Contour (RCC) curve (Fig. 2.4 D).

Leaf shape shows considerable variation caused by scale, rotation and translation (Fig. 2.5 A). Through alignment (see Subsection 2.2.2), all this has been filtered out from the objects (Fig. 2.5 B). A high dimension of leaf shape data described by RCC values, i.e., the coordinates along the leaf boundaries, is reduced using PCA. It was found from PCA that six orthogonal axes, termed PCs, could explain 88.1% of the variation among the samples, which, ordered according to the percentages of variance they explained, are PC1, 47.3%, PC2, 23.2%, PC3, 6.7%, PC4, 5.1%, PC5, 3.5%, and PC6, 2.3%. These PCs can describe each leaf shape by capturing different aspects of leaf shape variability including global and local.

To map the QTLs that affect leaf shape, the PC values were associated with 29 microsatellite markers. Table 2.2 tabulates the names of significant markers, their allele frequencies, the allele frequencies of the QTLs detected by these markers, and marker-QTL linkage disequilibria. PC1, PC3, PC4 and PC5 were each found to exhibit significant associations with three markers, whereas PC6 is associated with one marker. Some markers may be associated with different types of PC axes, suggesting that the same QTLs have a pleiotropic effect on different features of a leaf shape. For example, marker $GCPM\_1063$ is significantly associated with PC1 ($P = 1.01 \times 10^{-10}$), PC3 ($P = 1.88 \times 10^{-8}$), PC4 ($P = 1.14 \times 10^{-7}$) and PC5 ($P = 3.55 \times 10^{-7}$). It is possible that the same QTL causes the association of

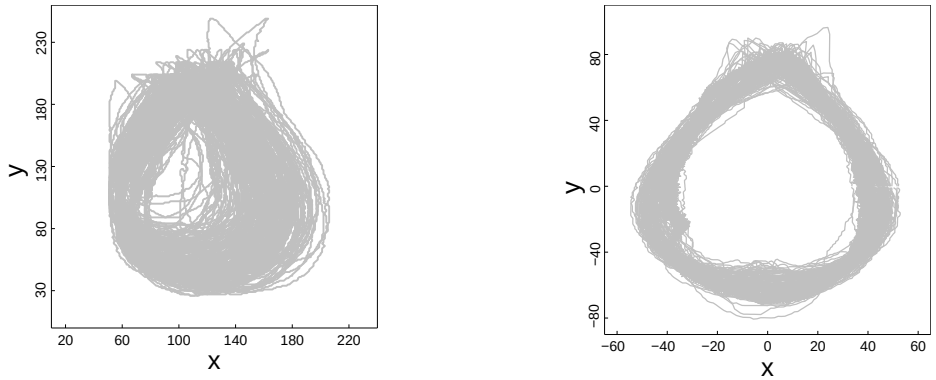

Figure 2.5. Linking 360 coordinates on the leaf outlines for leaves of all sampled trees from the mapping population. In A, raw leaf shapes, showing variation in scale, position and orientation. In B, this variation is removed from the objects through shape alignment.

$GCPM\_1063$ with these four PC axes because the QTLs detected for all the four PC axes have a similar allele frequency (0.49 - 0.51) and linkage disequilibrium (0.09 - 0.12).

Table 2.2. Detection of Leaf Shape QTLs by the Linkage Disequilibrium Analysis of Microsatellite Markers in a Natural Population of *Populus szechuanica var. tibetica*

PC(%Explained)	Microsatellite Marker	Effect p-value	$\hat{q}$	$\hat{p}$	$\hat{D}$	LD p-value
PC1(47.3%)	$GCPM\_1063$	1.01×10^{-10}	0.74	0.49	0.12	5.77×10^{-15}
	$GCPM\_1026-1$	2.42×10^{-10}	0.31	0.43	-0.12	2.42×10^{-13}
	$GCPM\_1093-1$	1.93×10^{-05}	0.43	0.47	-0.05	1.57×10^{-3}
PC3(6.7%)	$GCPM\_1063$	1.88×10^{-08}	0.81	0.51	0.09	2.94×10^{-13}
	$GCPM\_1064-1$	1.68×10^{-05}	0.44	0.43	-0.05	8.20×10^{-6}
	$GCPM\_1-1$	3.42×10^{-04}	0.73	0.56	0.08	1.72×10^{-11}
PC4(5.1%)	$GCPM\_1063$	1.14×10^{-07}	0.77	0.51	0.13	0
	$GCPM\_1026-1$	9.56×10^{-07}	0.60	0.23	0.07	4.64×10^{-13}
	$GCPM\_1034-1$	8.54×10^{-04}	0.45	0.14	-0.06	3.49×10^{-7}
PC5(3.5%)	$GCPM\_1063$	3.55×10^{-07}	0.79	0.51	0.10	0
	$GCPM\_1064-1$	1.63×10^{-04}	0.72	0.43	0.11	0
	$GCPM\_1025-1$	4.95×10^{-04}	0.73	0.44	0.11	0
PC6(2.3%)	$GCPM\_1053-1$	2.11×10^{-04}	0.26	0.49	-0.12	0

In Table 2.2, p is the allele frequency of a marker, q is the allele frequency of a QTL detected by the marker, and $\mathfrak{D}$ is the linkage disequilibrium (LD) between the marker and QTL. The effects of QTLs are tested by hypothesis (2.8), and the LD between markers and QTLs tested by hypothesis (2.9).

In general, the QTLs detected by PC1 control overall leaf shape variation, whereas the QTLs detected by the other PCs are responsible for local leaf variation. Generally speaking, the QTL detected by marker $GCPM\_1063$ alters leaf shape from lanceolate (AA) to ovate-orbicular (Aa) to ovate (aa) through PC1 (Fig. 2.6 A), whereas this QTL determines the detailed structure of broadly-ovate leaf shape, e.g., different degrees of deltoidness at leaf base among the three genotypes (Fig. 2.6 B). Fig. 2.7 illustrates the fitness of PC1 curves (A) and PC3 curves (B) to the RCC curves of all poplar trees, respectively, for three genotypes, AA, Aa and aa, at the QTL detected by marker $GCPM\_1063$. Difference in leaf shape explained by PC1 and PC3 curves of the same QTL genotype is diagrammed in Fig. 2.8 where such a difference is found to be genotype-specific.

The linkage disequilibria of markers with the QTLs are highly significant ($P =$

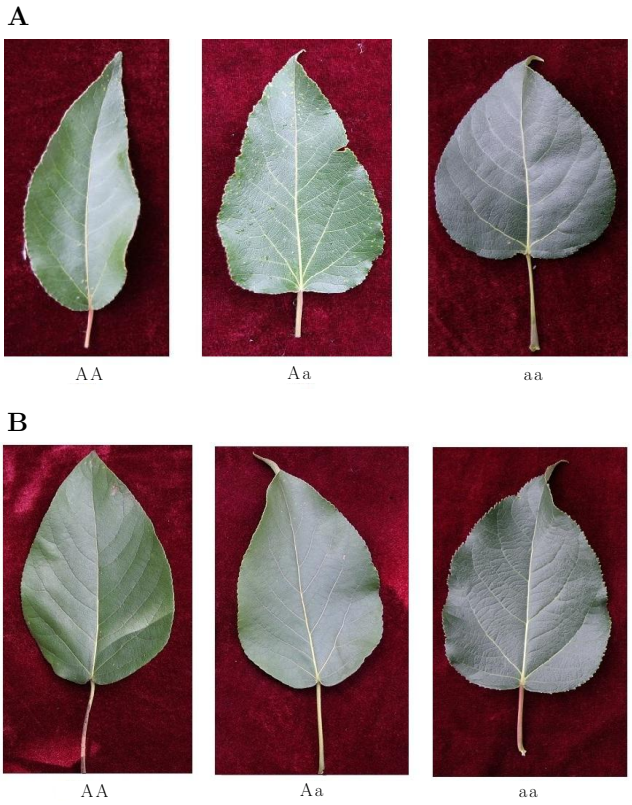

Figure 2.6. Three representative leaf shapes of *Populus szechuanica var. tibetica* corresponding to three different genotypes, AA, Aa and aa, at the QTL detected by marker $GCPM\_1063$. PC1 defines overall leaf shape (A), whereas PC3 defines local shape variability (B). In B, three genotypes all have broadly ovate leaf shape, but genotypes AA and Aa are more deltoid than genotype aa at leaf base.

1.57×10^{-3}), suggesting that these QTLs can possibly map to a narrow genomic region. Of the two other PC1 QTLs that control overall leaf shape in a similar manner, but with a lesser extent, one detected by marker $GCPM\_1026-1$ displays a larger effect on shape variation and is also closer to the QTL than one detected by marker $GCPM\_1093-1$ (Table 2.2). The QTLs associated with the other PCs tend to affect the local variation of leaf shape at various positions of leaves. Although it is subtle, such local variation may be tightly linked with gradient changes of some environmental factors. Thus, ecological functions of "local" QTLs deserve further investigations.

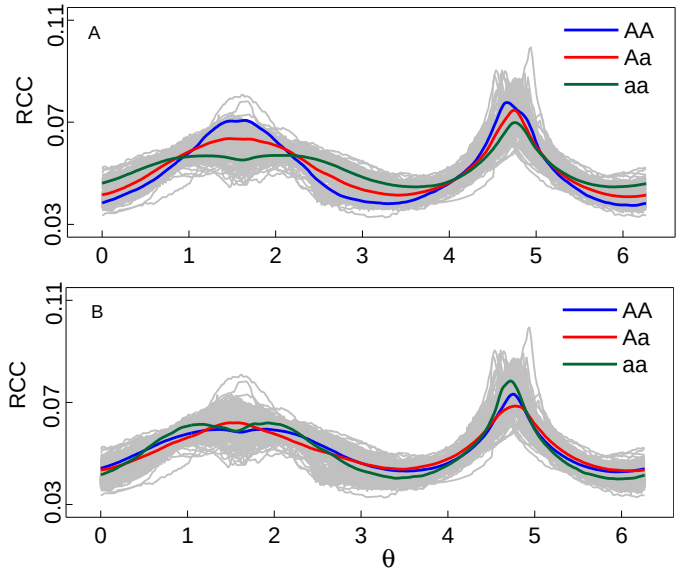

Figure 2.7. RCC curves of leaf shape as a function of radial angle θ at the centroid, explained by the PC1 curve (A) and PC3 curves (B) for the three genotypes, AA, Aa and aa, at the QTL detected by marker $GCPM\_1063$.

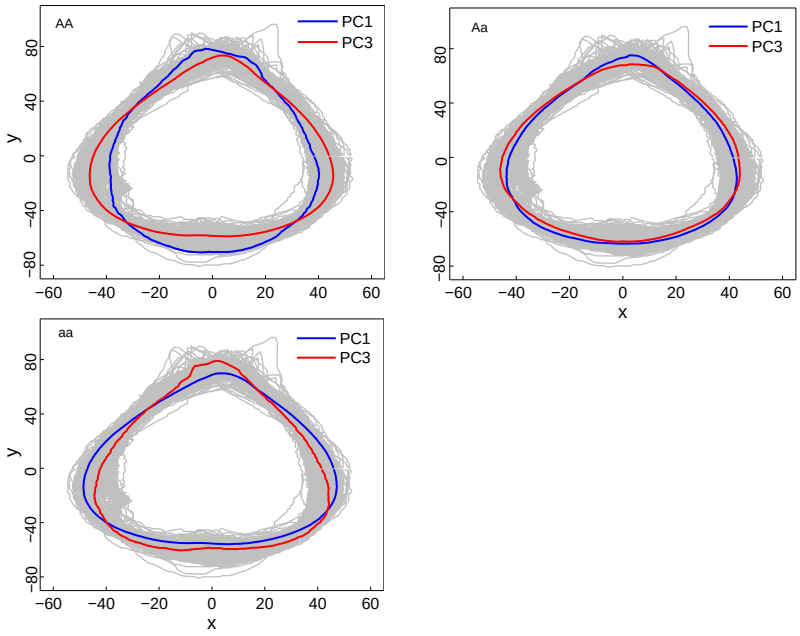

Figure 2.8. The pleiotropic control of the same QTL on different features of leaf shape specified by PC1 and PC2. The difference of leaf shape defined by PC1 (blue) and PC3 axes (red) for the same genotype, AA, Aa or aa, is shown.

2.4.2 *Populus euphratica oliv.*

Populus euphratica oliv., as one of the oldest members of poplar family, is a very special species to be researched because of its vast plasticity ability in dramatically different environment, temperature, light, and humidity, and so on. It is distributed from Mediterranean Woodlands and Shrublands, Semi-steppe shrublands, Shrub-steppes, to Deserts, and even prosperous in the Takla makan Desert. The leaves of *Populus euphratica oliv.* are polymorphic, i.e. have quite different shapes among different trees or among different branches of the same tree (Fig. 2.9. $T036-1, T036-6, T036-7, T036-11$ are four examples from the same tree). The leaf blade shapes of *Populus euphratica oliv.* vary significantly from lanceolate, elliptic, oblong, to rhomboid. Moreover, the most strikingly difference is the irregular dentation on the boundary of the leaves (Fig. 2.9).

In order to detect the significant QTLs that contribute to the variation of the *Populus euphratica oliv.* leaf shapes, we randomly selected 471 trees from a natural

Figure 2.9. A set of original leaf images (with IDs given at the bottoms) chosen from the mapping population for *Populus euphratica oliv.*, showing pronounced variation in leaf shape.

population of *Populus euphratica oliv.*, and took the photos of 25 leaves randomly from each tree. Therefore, there were altogether 11775 leaves under experiment, which has a huge sample size. Each leaf was saved in a separate colorful image with the resolution of 350×440. In Matlab, each colorful image was saved in the form of three dimensional matrix to record the RGB (Red, Green, and Blue) values that are enough to describe the original photo. Among these 471 trees, there were 421 progenies genotyped using 104 molecular markers. Because the different leaves from the same tree have quite big variations in shape and it is unreasonable to assume that the leaves from the same tree are independent, we computed the average of all 25 leaves as the phenotype for each progeny.

Since we use similar Linkage Disequilibrium genetic design and the morphological shape of *Populus euphratica oliv.* is not much different from that of *Populus szechuanica var. tibetica*, the methodologies for these two data are similar, except for some tiny differences. In the following, we only address the parts that are different.

In Subsection 2.4.1, we used the Radius-Centroid-Contour method to represent the leaf boundary by recording one point per θ. It worked well for many 2-dimensional shapes. However, this method with fixed θ has two problems. One problem is that it can not describe the non-convex shapes properly. For example, the result will be bad if there are more than one point when connecting the centroid to the contour. Another problem is that it will loss the information between two neighbor θs if the outline is not smooth. In the case of the leaf shape of *P. euphratica oliv.*, there exists a lot of irregular and sharp dentation, which can not be represented accurately by the fixed θ method any more. Therefore, a more accurate shape representation skill is necessary to capture every tiny and sharp corner of the dentation that is one of the most striking variation of the *Populus euphratica oliv.* leaves.

Kong et al. (Kong et al. 2007) proposed a new shape descriptor based on directional radii vector and the Haar wavelet transform. Because the boundary of the shape was closed, we were able to choose any two consecutive points on the boundary, one as the starting point and the other as the end point. If all the points between the starting point and the end point on the boundary were recorded one by one, there was no any information loss. Still used $(x(s_i), y(s_i)),\ (s_i = 1, \ldots, M_i)$

to denote the x and y coordinates of all the boundary points for individual shape i. But in the new method, the coordinates were subject to the individual. i.e. for each different shape i, the coordinates $x(s_i)$ and $y(s_i)$, and the number of points M_i all changed as i changed. In addition, the points on the boundary were visited by location or index rather than θ. By the new shape representation method, we obtained a curve or a long vector, with different length for each different shape. To make data analysis and shape alignment feasible, we must standardize the length of different vectors by interpolation. Let C to denotes the common length of all shapes, which is much larger than $\theta = 360$. Once the length of the curves were uniformed, the shape alignment formula described in Section 2.2.2 and the shape representation formula (2.4) were still applicable.

Since C (=910) was a big number which made the computation heavy and improved the complexity of statistical modeling, we used Haar wavelet transform (Kong et al. 2007; Zhao et al. 2008) to capture the important information. The goal of wavelet-transform is to take advantage of redundancy in the original information and obtain a good reconstruction upon decompression by dividing the original signal into two sequences of wavelet coefficients of equal length. Mathematically, Haar wavelet transformation is to represent the original signal as a superposition of a set of basis functions formed by a sequence of rescaled "square-shaped" functions. These basis functions are obtained from a single prototype wavelet called the mother wavelet $\psi(t)$, by dilations or scaling and translations. The Haar wavelet's mother wavelet function can be described as

$$\psi(t) = \begin{cases} 1, & 0 \leq \mathrm{t} < 1/2; \\ -1, & 1/2 \leq \mathrm{t} < 1; \\ 0, & \mathrm{ow}. \end{cases}$$

By Implementing the Haar wavelet transform, we compressed the original directional radii vectors into half of its original length. Then, we used exactly the same methods as Subsection 2.4.1 for the remaining parts. It turned out that the first eleven PCs, explain total 97.86% of the variation among the samples from different aspect, are in detail, PC1, 42.89%, PC2, 32.77%, PC3, 10.68%, PC4, 4.58%, PC5, 2.50%, PC6, 1.57%, PC7, 0.98%, PC8, 0.86%, PC9, 0.40%, PC10, 0.33%, and PC11, 0.30%. To map the QTLs that affect leaf shape, these PC values were

associated with 104 molecular markers. Then we performed two hypothesis tests. As mentioned in Section 2.3.6, a significant QTL that contributes to the shape variation must reject both of the two hypothesis tests. Considering there are 104 markers, we use Bonferroni correction to control the family rate. Then, the critical value for the hypothesis test (2.8) with the df value 2 is 15.28, for hypothesis test (2.9) with the df value 1 is 12.19. Among these eleven PCs, PC 4, PC8, PC10, and PC11 failed to reject both the two hypothesis tests. For the other seven significant ones, in Table 2.3, we summarized their allele frequencies, the allele frequencies of the QTLs detected by these markers, and marker-QTL linkage disequilibria.

Table 2.3. Detection of Leaf Shape QTLs by the Linkage Disequilibrium Analysis of Molecular Markers in a Natural Population of *Populus euphratica oliv.*

PC(%Explained)	Microsatellite Marker	Effect p-value	$\hat{q}$	$\hat{p}$	$\hat{D}$	LD p-value
PC1(42.89%)	Marker1	1.45×10^{-12}	0.43	0.17	-0.07	0
	Marker28	5.70×10^{-10}	0.42	0.42	-0.05	3.13×10^{-09}
	Marker56	1.93×10^{-09}	0.46	0.17	-0.04	4.99×10^{-08}
	Marker54	4.63×10^{-09}	0.64	0.60	0.06	5.73×10^{-11}
	Marker13	9.45×10^{-08}	0.57	0.34	0.04	4.34×10^{-08}
PC2(32.77%)	Marker7	2.00×10^{-15}	0.41	0.59	-0.04	1.90×10^{-06}
	Marker96	2.22×10^{-15}	0.46	0.24	-0.04	3.63×10^{-07}
	Marker10	6.93×10^{-14}	0.44	0.35	-0.05	3.84×10^{-08}
	Marker17	1.42×10^{-13}	0.43	0.243	-0.05	2.13×10^{-10}
	Marker95	9.47×10^{-11}	0.55	0.16	0.04	2.24×10^{-09}
PC3(10.68%)	Marker36	8.22×10^{-12}	0.88	0.86	0.06	0
	Marker90	1.13×10^{-11}	0.61	0.29	0.07	0
	Marker70	1.82×10^{-11}	0.61	0.28	0.07	0
	Marker76	1.79×10^{-10}	0.60	0.37	0.06	1.21×10^{-11}
	Marker54	1.01×10^{-09}	0.69	0.60	0.07	0
	Marker35	1.27×10^{-09}	0.62	0.24	0.08	0
PC5(%)	Marker10	0	0.60	0.35	0.05	1.07×10^{-08}
	Marker86	0	0.57	0.24	0.03	2.41×10^{-06}
	Marker69	0	0.64	0.34	0.08	0
	Marker13	0	0.58	0.34	0.04	3.71×10^{-07}
PC6(%)	Marker67	3.57×10^{-11}	0.30	0.43	-0.12	0
PC7(%)	Marker89	1.36×10^{-4}	0.44	0.22	-0.05	7.07×10^{-11}
PC9(%)	Marker55	1.79×10^{-4}	0.83	0.66	0.11	0

In Table 2.3, p is the allele frequency of a marker, q is the allele frequency of a QTL detected by the marker, and $\mathfrak{D}$ is the linkage disequilibrium (LD) between

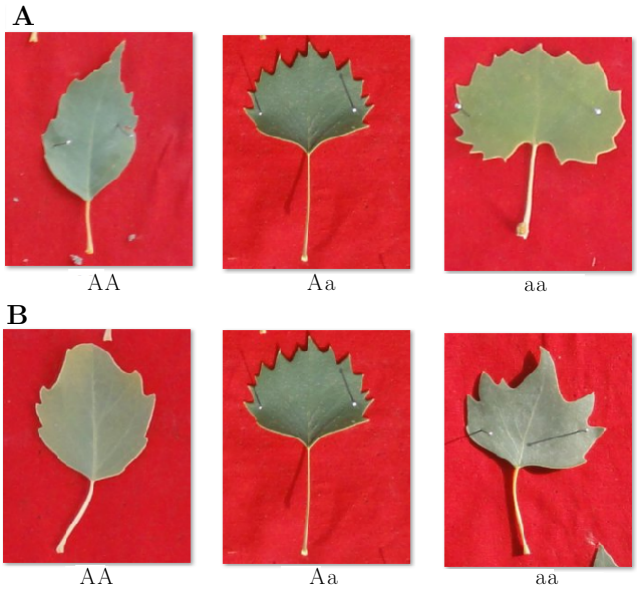

Figure 2.10. Three representative leaves of *Populus euphratica oliv.* corresponding to three different genotypes, AA, Aa and aa, at the QTL detected by marker1 for PC1 (A) and marker7 for PC2 (B). In A, the genotypes corresponding to PC1 show quite big variation. AA has lanceolate, Aa has oblong, and aa has rhomboid leaf shape. In B, three genotypes all have broadly ovate leaf shape, but the dentation patterns are different among three different genotypes associated with PC2.

the marker and QTL. The effects of QTLs are tested by hypothesis (2.8), and the LD between markers and QTLs tested by hypothesis (2.9).

Since the QTLs detected by PC1 and PC2 control the first two biggest overall leaf shape variation (almost half of the total variation), we illustrated the effect of PC1 and PC2 in Fig. 2.10. Generally speaking, the QTL detected by marker1 alters leaf shape from lanceolate (AA) to oblong (Aa) to rhomboid (aa) through PC1 (Fig. 2.10 A). And the QTL detected by marker 7 determines the detailed structure of the dentation. (Fig. 2.10 B). Fig. 2.11 illustrates the fitness of PC1 curves (Top panel) and PC2 curves (Bottom panel) to the DRV (Directional Raddi Vectors) of all trees, respectively, for three genotypes, AA, Aa and aa, at the QTL detected by marker1 and by marker7, respectively. Difference in leaf shape explained by PC1 and PC2 leaves of the same QTL genotype is diagrammed in Fig. 2.12 where such a difference is found to be genotype-specific.

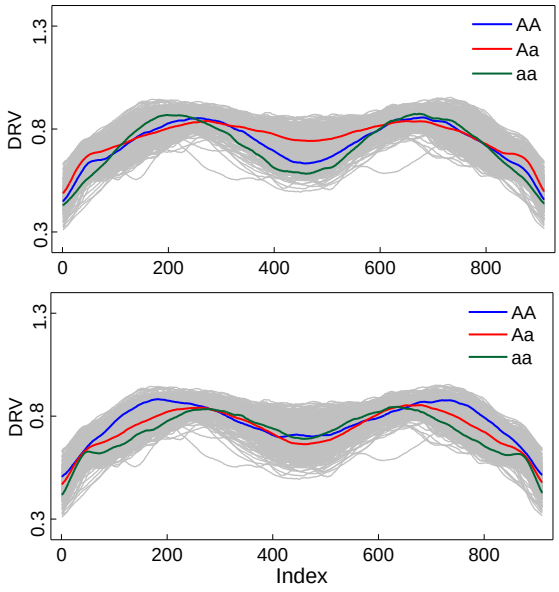

Figure 2.11. Directional Radii Vectors of leaf shape as a function of index explained by the PC1 curve (Top) and PC3 curves (Bottom) for the three genotypes, AA, Aa and aa, at the QTL detected by marker1 and marker 7, respectively.

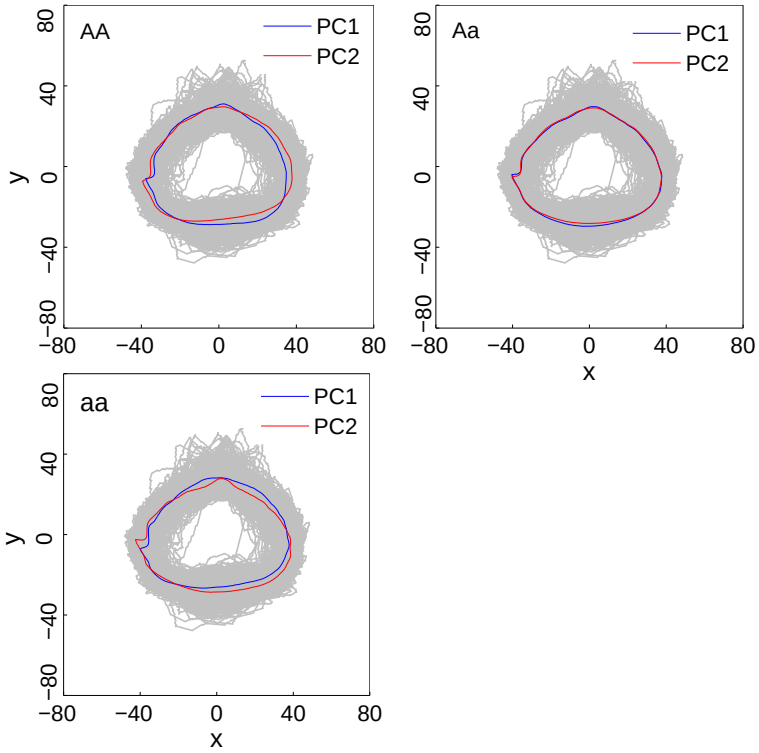

Figure 2.12. The control of the same QTL on different features of leaf shape specified by PC1. The difference of leaf shape defined by PC1 for the same genotype, AA, Aa or aa, is shown from image domain.

Chapter 3

Mapping Shape QTLs Using Level Set Method

In this Chapter, we derived a statistical model for mapping specific genes or quantitative trait loci (QTLs) that control morphological shape. After the region based shape analysis skills, we obtained n level set functions $Y = \{Y_1, Y_2, \cdots, Y_n\}$ corresponding to n aligned shapes. Each Y_i is a L by L matrix. The model was formulated within the mixture framework, in which different types of shape are thought to result from genotypic discrepancies at a QTL. The EM algorithm was implemented to estimate QTL genotype-specific shapes based on a shape correspondence analysis. Computer simulation was used to investigate the statistical property of the model. By identifying specific QTLs for morphological shape, the model developed will help to ask, disseminate and address many major integrative biological and genetic questions and challenges in the genetic control of biological shape and function.

3.1 Introduction

Morphological shape is one of the most conspicuous aspects of an organism's phenotype and provides an intricate link between biological structure and function in changing environments (Ricklefs et al. 1994; Reich 2001). For this reason, comparing the anatomical and shape feature of organisms has been a central element of biology for centuries. Nowadays, attempts have been made to unlock the genetic

secrets behind phenotypic differentiation in developmental shape (Tanksley 2004), understand the origin and pattern of shape variation from a developmental perspective (Klingenberg 2001; Klingenberg et al. 2001), and predict the adaptation of morphological shapes in a range of environmental conditions (Tsukaya 2005).

Three major advances in life and physical science during the last decades will make it possible to study shape variation and its biological underpinnings. First, DNA-based molecular markers allow the identification of quantitative trait loci (QTLs) and biochemical pathways that contribute to quantitatively inherited traits such as shape. In his seminal review, Tanksley (Tanksley 2004) summarized some major discoveries of genes for fruit size and shape in tomato. In a long process of domestication, tremendous shape variation has occurred in tomato fruit from almost invariably round (wild or semiwild types) to round, oblate, pear-shaped, torpedo-shaped, and bell pepper-shaped (cultivated types). Some of the QTLs that cause these differences, namely fw2.2, ovate, and sun, have been cloned (Frary et al. 2000; Liu et al. 2002; Xiao et al. 2008).

Second, digital technologies through computerized analysis and processing procedures can obtain a comprehensive representation of the involved objects, capable not only of representing most of the original information, but also of emphasizing their less redundant portions (Bookstein 1978; Monteiro et al. 2002; Adams et al. 2004; Bernal 2007; Stegmann et al. 2002; Basri et al. 1998). Third, statistical and computational technologies have well been developed for analyzing high-dimensional, large-scale, high-throughput data of high complexity (Dempster et al. 1977; Tsai et al. 2005). With the development of missing data analysis, Lander and Botstein (1989) have been able to pioneer an approach for dissecting complex quantitative traits into individual QTLs using genetic linkage maps constructed with molecular markers. There has been a vast wealth of literature in the development of QTL mapping models (Zeng 1994; Jansen et al. 1994; Xu et al. 1995; Lynch et al. 1998; Broman et al. 2002; Zou et al. 2004; Yi et al. 2005).

The motivation of this study is to develop a statistical and computational model for mapping specific QTLs that are responsible for differences in morphological shape. Historically, genetic mapping has been focused on the genetic control of a trait at a static point, ignoring the dynamic behavior and spatial properties of the trait. Now, by integrating the developmental principle of trait growth, a

new genetic mapping approach, called functional mapping (Ma et al. 2002; Wu et al. 2003; Wu et al. 2006), can be used to study the dynamic control of genes in time course. The central idea of functional mapping is to connect the genetic control of a developmental trait at different time points through robust mathematical and statistical equations. Complementary to functional mapping, the model developed for shape mapping in this study links gene action with key morphometric parameters of a shape within a statistical framework. We will perform computer simulation to examine the statistical properties of the model.

3.2 Region-Based Shape Analysis

For the region based shape representation method, we divide the procedures into two steps: shape alignment to minimize the variation caused by location, scale and rotation; and shape description to describe a shape using numerical matrices.

3.2.1 Shape Alignment

According to the definition of Kendall (Dryden et al. 1998), "shape is all the geometrical information that remains when location, scale and rotational effects are filtered out from an object". Assume that each backcross progeny is measured for the leaf shape as shown in Fig. 3.1. For a given shape, I^i $(i = 1, ..., n)$, described by a black and white image, it is gridded as an $L \times L$ matrix, where L is the number of pixels in the row and column. At each point in the matrix, we use 0 to denote the background (black) and 1 to denote the leaf (including an arbitrary shape of it) (white). The 1/0 value of the matrix is assumed to follow a Bernoulli distribution. All these n shapes, $T = \{I^1, I^2, \cdots, I^n\}$, need to be aligned, in order to minimize the interference caused by pose variations. This can be carried out by establishing a coordinate reference with respect to position, scale and rotation, commonly known as pose to which all shapes are aligned (Bookstein 1978; Adams et al. 2004; Stegmann et al. 2002).

Denote the pose parameter for each shape I^i by $p^i = [a, b, h, \theta]^T$ where a and b correspond to x and y translations, h is the scaling parameter, and θ corresponds to rotation. The transformed image of I^i, based on the pose parameter p^i, is denoted

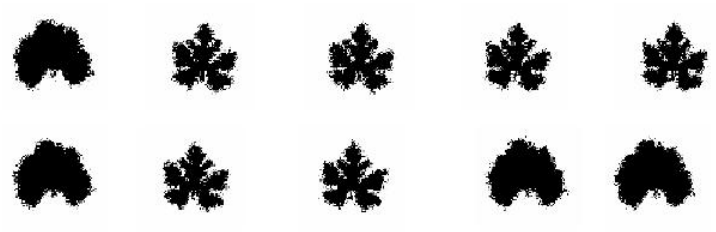

Figure 3.1. The Diagram of twelve leaf shapes from the backcross population. Five of them are wild *Cucurbita argyrosperma sororia* and seven of them are cultivated *cucurbita argyrosperma*.

by $\tilde{I}^i$, defined as

$$\tilde{I}^i(\tilde{x}, \tilde{y}) = I^i(x, y),$$

where

$$\begin{pmatrix} \tilde{x} \\ \tilde{y} \\ 1 \end{pmatrix} = T[p] \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & a \\ 0 & 1 & b \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} h & 0 & 0 \\ 0 & h & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} cos(\theta) & -sin(\theta) & 0 \\ sin(\theta) & cos(\theta) & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix},$$

which yields

$$\begin{cases} \tilde{x} &= a + hxcos(\theta) - hysin(\theta), \\ \tilde{y} &= b + hycos(\theta) + hxsin(\theta). \end{cases} \tag{3.1}$$

The translation matrix $T[p]$ is the product of three matrices: a translation matrix $M(a, b)$, a scaling matrix $H(h)$, and an in-plane rotation matrix $R(\theta)$. The transformation matrix $T[p]$ maps the coordinates $(x, y) \in R^2$ into coordinates $(\tilde{x}, \tilde{y}) \in R^2$, where $x, y = 1, \cdots, L$.

An effective strategy to jointly align the n binary images is to use a gradient descent to minimize the following energy function:

$$E = \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \left\{ \frac{\int\int_{\Omega} (\tilde{I}^i - \tilde{I}^j)^2 dA}{\int\int_{\Omega} (\tilde{I}^i + \tilde{I}^j)^2 dA} \right\}, \tag{3.2}$$

where Ω denotes the image domain. Minimizing the energy function (3.2) is equivalent to simultaneously minimizing the difference between any pair of binary images in the training database. What we would like to estimate is the pose parameter p^i

for each I^i.

The derivative respective to p^i of equation (3.2) is

$$\begin{aligned}\nabla_{p^i} E =& 2\Sigma_{j=1,j\neq i}^{n} \left\{ \frac{2\int\int_{\Omega}(\tilde{I}^i - \tilde{I}^j)\nabla_{p^i}\tilde{I}^i dA}{\int\int_{\Omega}(\tilde{I}^i + \tilde{I}^j)^2 dA} \right. \\ & \left. - \frac{2\int\int_{\Omega}(\tilde{I}^i - \tilde{I}^j)^2 dA \int\int_{\Omega}(\tilde{I}^i + \tilde{I}^j)\nabla_{p^i}\tilde{I}^i dA}{(\int\int_{\Omega}(\tilde{I}^i + \tilde{I}^j)^2 dA)^2} \right\},\end{aligned} \tag{3.3}$$

where $\nabla_{p^i}\tilde{I}^i = \left[\frac{\partial \tilde{I}^i}{\partial a}, \frac{\partial \tilde{I}^i}{\partial b}, \frac{\partial \tilde{I}^i}{\partial h}, \frac{\partial \tilde{I}^i}{\partial \theta}\right]^T$.

By a chain rule and equation (3.1), we get

$$\begin{aligned}\frac{\partial \tilde{I}^i}{\partial a} &= \frac{\partial \tilde{I}^i}{\partial \tilde{x}} = \frac{\partial I^i}{\partial x}, \\ \frac{\partial \tilde{I}^i}{\partial b} &= \frac{\partial \tilde{I}^i}{\partial \tilde{y}} = \frac{\partial I^i}{\partial y}, \\ \frac{\partial \tilde{I}^i}{\partial h} &= \frac{\partial I^i}{\partial x}(xcos(\theta) - ysin(\theta)) + \frac{\partial I^i}{\partial y}(ycos(\theta) + xsin(\theta)), \\ \frac{\partial \tilde{I}^i}{\partial \theta} &= \frac{\partial I^i}{\partial x}(-hxsin(\theta) - hycos(\theta)) + \frac{\partial I^i}{\partial y}(-hysin(\theta) + hxcos(\theta)).\end{aligned}$$

Hence, we can obtain the value of $\nabla_{p^i} E$ as long as p^i and $\tilde{I}^i$ are given in each iterative step. The steepest gradient algorithm is then used to minimize E in (3.2) and get the pose parameter p^i for each shape I^i. All the training shapes after the alignment procedure described above are obtained (Fig. 3.2).

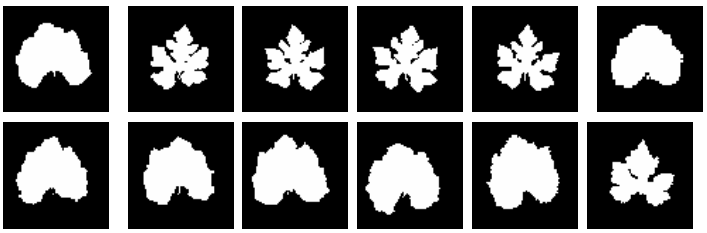

Figure 3.2. Leaf shapes after alignment for leaf shapes shown in Fig. 3.1.

3.2.2 Shape description

After all the training shapes are aligned, a shape representation scheme needs to be chosen for $T = \{\tilde{I}^1, \tilde{I}^2, \cdots, \tilde{I}^n\}$., i.e., the transformed images, which now become continuous variables. We will use the implicity representation of boundary by level set method, which is an Eulerian approach (Osher et al. 1988).

A closed curve can divides the image domain into three parts: The region inside the curve I, the region outside the curve I^c, and the boundary C. The boundary of a shape is a curve satisfying some specific function. Osher and Sethian define a smooth function $\phi(x, y)$ such that the set where $\phi(x, y) = 0$ corresponding to the boundary C. If ϕ has the following property, then it is said to be a level set function:

$$\begin{aligned} \phi(x,y) < 0, \quad if\ (x,y) \in I \\ \phi(x,y) > 0, \quad if\ (x,y) \in I^c \\ \phi(x,y) = 0, \quad if\ (x,y) \in C \end{aligned}$$

By this definition, we are able to embed the boundary curve C into the 0 level set of a 2D function $z = \phi(x, y)$, and hence represent a shape implicitly. For any given boundary, you might find several level set functions, but once the level set function is chosen, the boundary will be uniquely determined. Signed distance function $|\phi(x)| = min_{x_I \in C}\ d(x, X_I)$ is a traditional way to be served as a shape descriptor to represent the contours of the shape. As you can see, it satisfies the definition of level set function. Each contour is embedded as the zero level set of a signed distance function with negative distances assigned to the inside and positive distances assigned to the outside. This technique yields n level set functions $Y = \{Y_1, Y_2, \cdots, Y_n\}$ corresponding to above n aligned training shapes.

By now, we finish all the steps of shape analysis and are able to substitute Y_i to the statistical model in the following to map the genes that control the shape variability for the simulated leaf shapes.

3.3 Statistical Design

3.3.1 Genetic Design

We assume a backcross design although the model can be modified to accommodate any other mapping designs. Consider a backcross progeny population of size n, founded with two inbred lines that are sharply contrasting in leaf shape. Because of gene segregation, there is a range of variation in leaf shape among the backcross progeny. Such shape variation is illustrated in Fig. 3.1 by using leaf morphology in cucurbit plants (Schlichting et al. 1998). To map the shape trait, the mapping population is typed for a panel of molecular markers from which a genetic linkage map covering the genome is constructed. The statistical approach for linkage analysis and map construction is reviewed in Wu et al. (Wu et al. 2007). Assume that there are some specific QTLs responsible for the biological shape. The approach being developed aims to detect and map such QTLs by capitalizing on knowledge about shape analysis and biological principles behind shape formation and variation.

3.3.2 Statistical Model

From the standpoint of QTL mapping, we treat $Y = \{Y_1, Y_2, \cdots, Y_n\}$ as the multiple phenotypic traits of n individuals. For a progeny i ($i = 1, 2, \cdots, n$), we have

$$Y_i = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1L} \\ y_{21} & y_{22} & \cdots & y_{2L} \\ \vdots & \vdots & \ddots & \vdots \\ y_{L1} & y_{L2} & \cdots & y_{LL} \end{pmatrix}. \tag{3.4}$$

Thus, each individual has a total of $m = L^2$ dimensions.

For the backcross progeny population, there are always two different genotypes at each locus. The genotypes at a shape QTL, expressed as QQ (denoted as 1) and Qq (denoted as 2), cannot be observed directly but can be inferred from the markers that are linked to the QTL. For this reason, the basic statistical model for QTL mapping is based on a mixture model, in which each observation Y is assumed

to have arisen from one of the two groups of QTL genotypes, each group being modeled from a density function (frequently a normal distribution is assumed). Thus, the population density function of Y is

$$f(Y|\omega, \phi, \eta) = \sum_{j=1}^{2} \omega_{j|i} f_j(Y_i|\mu_j, \eta), \tag{3.5}$$

where ω represents the mixture proportions $(\omega_{1|i}, \omega_{2|i})$, which are constrained to be nonnegative and sum to unity, μ_j is the expectation of different QTL genotypes $j = 1, 2$, and η is the variance-covariance parameter common to all genotype groups, and $f_j(Y_i|\mu_j, \eta)$ is the probability density function for QTL genotype j. After images are transformed, Y_i can be assumed to follow a multivariable normal distribution, i.e.,

$$f_j(Y_i) = \frac{1}{(2\pi)^{m/2}|\Sigma|^{1/2}} \exp[-(Y_i - \mu_j)^T \Sigma^{-1}(Y_i - \mu_j)/2], \tag{3.6}$$

with the expectation matrix of each QTL genotype expressed as

$$\mu_j = \begin{pmatrix} \mu_{11}^j & \mu_{12}^j & \cdots & \mu_{1L}^j \\ \mu_{21}^j & \mu_{22}^j & \cdots & \mu_{2L}^j \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{L1}^j & \mu_{L2}^j & \cdots & \mu_{LL}^j \end{pmatrix}, \quad \text{for } j = 1, 2, \tag{3.7}$$

and $(m \times m)$ residual variance-covariance matrix of the variables Σ.

In order to simplify the problem, we use the most natural sampling strategy to utilize the $L \times L$ rectangular grid of the training shapes to generate $m = L \times L$ lexicographically ordered samples (where the columns of the matrix grid are sequentially stacked on top of one other to form one large row). Also, we assume that all the components of the same random vector are independent and the variance of the different random vectors are common among the different progenies. now, from equation (3.5), we get the likelihood function as

$$L(y) = \prod_{i=1}^{n} f(Y_i|\omega, \mu, \eta)$$

$$= \prod_{i=1}^{n}\sum_{j=1}^{2}\omega_{j|i}f_j(Y_i|\mu_j,\eta)$$
$$= \prod_{i=1}^{n}\sum_{j=1}^{2}\omega_{j|i}\frac{1}{(2\pi)^{m/2}|\Sigma|^{1/2}}\exp[-(Y_i-\mu_j)^T\Sigma^{-1}(Y_i-\mu_j)/2], \quad (3.8)$$

where the mean matrix (μ_j) of QTL genotype j is modeled by parameter μ_j, and covariance matrix (Σ) is a diagonal matrix with common values in the diagonals.

3.3.3 Parameter estimation

To obtain the maximum likelihood estimates (MLEs) of parameters in likelihood (3.8), we implement a standard EM algorithm. In the E step, we compute the posterior probability with which a backcross individual carries a QTL genotype j using

$$\Omega_{ij} = \frac{\omega_j f_j(Y_i|\mu_j,\eta)}{\sum_{l=1}^{2}\omega_l f_l(Y_i|\mu_l,\eta)}. \quad (3.9)$$

In the M step, we estimate the parameters using

$$\mu_{jk} = \frac{\sum_{i=1}^{n}\Omega_{ij}y_{ik}}{\sum_{i=1}^{n}\Omega_{ij}}, \quad (3.10)$$

for $j = 1, 2$ and $k = 1, 2, \cdots, m$.

The EM steps are iterated between equations (3.9) and (3.10) until the estimates converge to stable values. It should be pointed out that the data set for shape analysis is highly sparse and high-dimensional. For example, if a shape is described by (75×75) pixels, i.e., $L = 75$, then we will have $m = 75^2 = 5625$, and an $(n \times 5625)$ matrix for the phenotypic observations. Several approaches will be developed to model the structure of the variance-covariance matrix. One of the simplest approaches is to use $\sigma = \frac{1}{2}\sqrt{2L^2}$. This choice is large enough to assure that various levels of differences lie well within a Gaussian distribution.

3.3.4 Hypothesis tests

A hypothesis about the existence of a significant QTL that controls a morphological shape can be tested by calculating the log-likelihood ratio under the hypotheses:

$$H_0 : \mu_1 = \mu_2 \quad vs \quad H_1 : \mu_1 \neq \mu_2 \tag{3.11}$$

As like an usual mapping approach, shape mapping has a problem of uncertain distribution for the log-likelihood test statistic. However, an empirical approach based on permutation tests, which does not rely on the distribution of log-likelihood ratios, can be used to determine the threshold for claiming the existence of a significant QTL.

3.4 Simulation Design and Experimental Results

Cucurbit (Cucurbita argyrosperm) plants display tremendous variation in shape between cultivars and wild types (Schlichting et al. 1998). By mimicking leaf morphologies of this species, we performed simulation studies to examine the statistical behavior of our shape mapping model. A backcross population of 200 progeny was simulated for a linkage group with 11 equally spaced markers. A QTL that determines leaf shape is hypothesized on the third marker interval. The phenotypic values of the shape were simulated with a (75 * 75) dimension by $Y_i = \xi_j \mu_1 + (1 - \xi_j)\mu_2 + e_i$, where μ_j is the mean shape matrix for QTL genotype j (j = 1, 2), ξ_j is the indicator variable defined as 1 and 0 if progeny i carries QTL genotype QQ (1) and qq (2), respectively, and e_i follows a multivariate normal distribution with mean vector zero and covariance matrix Σ. To simplify computing, we assumed that Σ is an identity matrix. We designed two simulation schemes to test our shape mapping algorithm.

The first scheme assumes that there exists a "big" QTL which triggers a tremendous effect on the difference in leaf shape of cucurbit plants between their cultivars and wild types. This QTL has two different genotypes, one, QQ, corresponding to the wild type shape (right) and the second, Qq, to the domesticated shape (left) (Fig. 3.3A). The QTL genotypes are determined by the conditional probability of

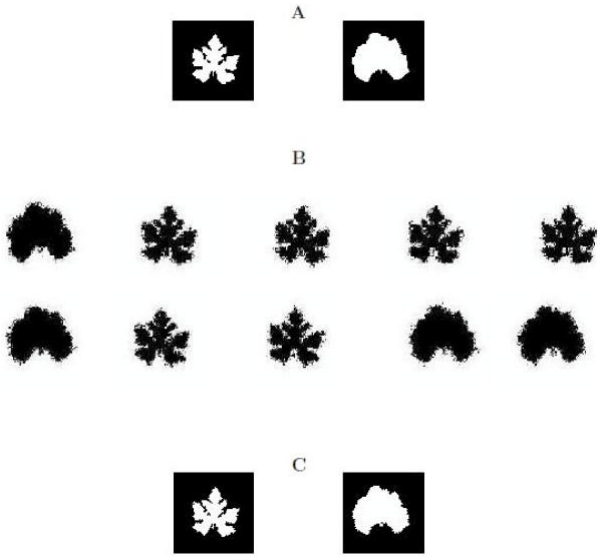

Figure 3.3. The first simulation scheme: A "big" QTL controls differences in leaf shape between wild types and cultivars for *cucurbit* plants. A: Two given QTL genotypes, QQ for the wild type (left) and Qq for the cultivar (right); B: Part of the simulated backcross progeny; C: Two estimated QTL genotypes, QQ for the wild type (left) and Qq for the cultivar (right).

a QTL genotype, conditional upon the genotypes of the two markers that flank the QTL (Wu et al. 2006). Part of the 200 progeny simulated with two assumed QTL genotypes were given in Fig. 3.3B, in which some leaf shape looks more like the wild type, some more like the domesticated type, and the other is in between. The model described above was used to analyze the simulated data. The log-likelihood ratio test statistic calculated under hypotheses (3.11) is greater than the critical threshold for testing the existence of a QTL obtained from permutation tests, suggesting that two genotype-specific shapes for QQ and Qq were detected and identified. Fig. 3.3B also illustrates the shapes of two detected QTL genotypes from the simulated data. As shown, the estimated shapes are similar to the true shapes for the two backcross QTL genotypes, suggesting that our model has great power to identify the QTL that control morphological shape.

The second scheme simulated two QTLs that determine the differences of leaf shape among wild-type plants and domesticated plants, respectively. Compared to the "big" QTL assumed in the first scheme, these two QTLs are "small" because their two genotypes correspond to slightly different leaf shapes. Fig. 3.4 and Fig.

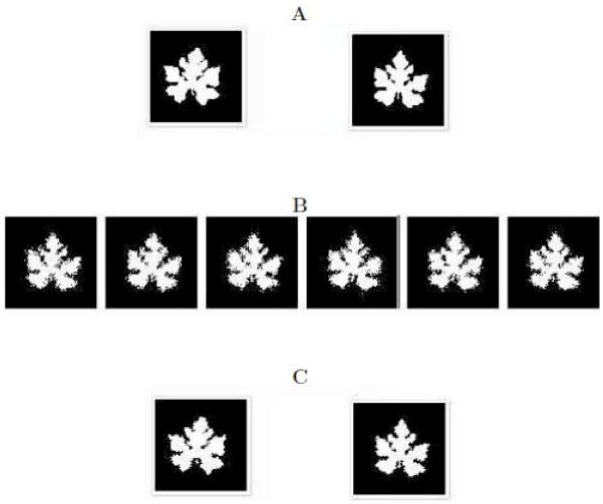

Figure 3.4. The second simulation scheme: A "small" QTL controls differences in leaf shape among different plants from wild types of cucurbit plants. A: Two given QTL genotypes, QQ for the wild type (left) and Qq for the cultivar (right); B: Part of the simulated back-cross progeny; C: Two estimated QTL genotypes, QQ for the wild type (left) and Qq for the cultivar (right).

3.5 provide the results about shape mapping for wild-type plants and domesticated plants, respectively. In the upper panel (A) of each figure, two original QTL genotypes are assumed, from which 200 backcross progeny were simulated with a range of leaf shape. The middle panel (B) gives part of the backcross. In the bottom panel (C), two genotypes were estimated using our algorithm. It can be seen that the model can well detect a QTL even if it has a small effect on morphological shape.

To show the fitness of our model, we put the estimated QTL genotypes on the simulated backcross population for the first (A) and second (B and C) simulation scheme (Fig. 3.6) on the image domain. The leaf shape of two QTL genotypes in each case well covers the simulated leaf shape, showing a good fitness of the mapping model.

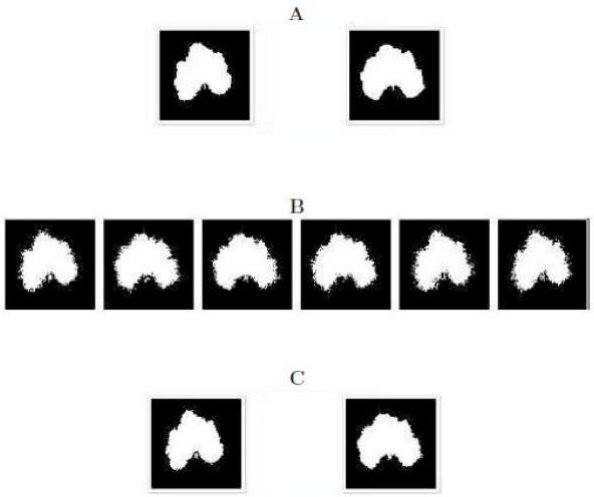

Figure 3.5. The second simulation scheme: A "small" QTL controls differences in leaf shape among different plants from cultivars of cucurbit plants. A: Two given QTL genotypes, QQ for the wild type (left) and Qq for the cultivar (right); B: Part of the simulated backcross progeny; C: Two estimated QTL genotypes, QQ for the wild type (left) and Qq for the cultivar (right).

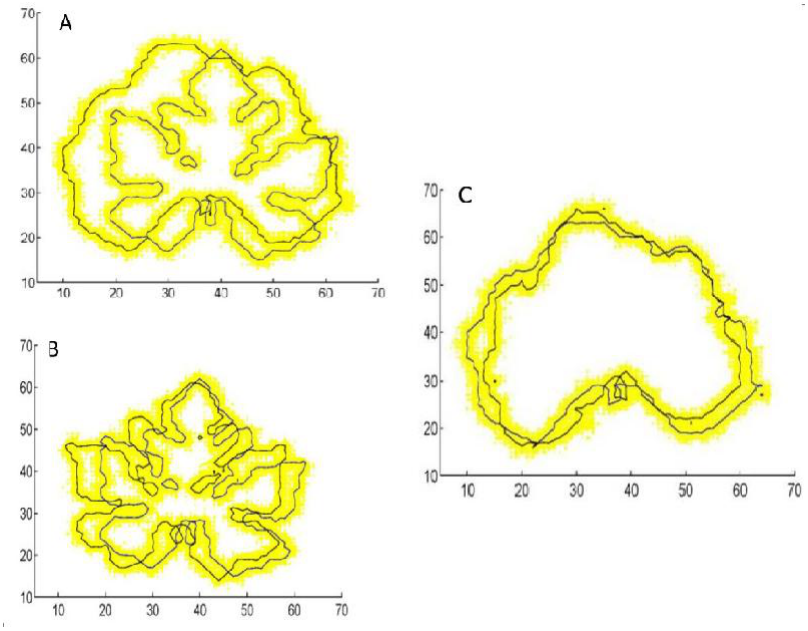

Figure 3.6. The fitness of estimated QTL genotypes to simulated leaf shape in a backcross. A: A "big" QTL for the shape difference between wild types and cultivars of *cucurbit* plants. B: A "small" QTL for the shape difference between different wild types. C: A "small" QTL for the shape difference between different cultivars.

Chapter 4

Functional QTL Mapping for Ultra High Dimensional Biological Shape Curves

After quantifying the morphological shapes numerically through RCC (Radius Centroid Contour) skills, each phenotype, as a datum, is in the form of samples of functional curves or trajectories with high dimension. In this section, we developed a nonparametric smoothing method to model the mean curve by GEE (Generalized Estimating Equation) local polynomial kernel and model the covariance matrix by functional PCA (Principal Component Analysis). Through this model, we estimate both the mean and covariance as the function of spatial angle, characterize the dominant modes of variation around the overall mean trend function, and hence avoid facing directly the extremely huge dimensional covariance matrix.

4.1 Introduction

Despite the fundamental importance of morphological shape, the difficulty in quantifying the image photo and modeling the ultra-high dimension of the shape data make the task of genetic mapping on it increasingly difficult.

In this Chapter, we provide novel insights into the genetic mechanisms that

control the structure of the morphological shapes, by developing a framework under which the semi- and non-parametric statistical models, functional data analysis skills, shape analysis skills, and functional QTL mapping schemes can be integrated together to achieve our purpose. In this model, both the mean and variance are modeled as functions of the spatial angle, utilizing the semi- and non-parametric kernel regression and functional PCA, respectively. Functional PCA describes observed random trajectories in terms of a number of functional principal component scores, and eigenfunctions have been interpreted as the modes of variation of the curves (Yao et al. 2005a; 2003; 2006; Muller et al. 2006). The proposed model is data-adaptive and does not require pre-specified functional forms and it automatically detects characteristic patterns.

By a Nature Review Genetics paper (Klingenberg 2012), quantifying the shape by landmarks is a quite new skill. After the mathematical definition of shape (Dryden et al. 1998) and the manual landmark points (Cootes et al. 1995) first proposed in 1995, the landmark has become a popular contour based shape representation method (Renaud et al. 2010; Langlade et al. 2005; Abbasi et al. 2000; Super 2004; Zhang et al. 2003; Rhodri et al. 2001; 2002; Tan et al. 2000; Kong et al. 2007). In current literature, multivariate analysis models have been mainly used to model the multidimensional landmark points. For example, principal component analysis used for examining the main patterns of variation in the data (Langlade et al. 2005; Drake et al. 2010), multivariate regression used for analyzing allometry or evolutionary change in shape over time (Drake et al. 2008; Monteiro 1999). However, the shape representation data is more meaningful to be treated as curves or trajectories rather than un-constructed vectors. Although the landmarks are recorded discretely, a continuous curve or function lies behind these data and the spatial dynamics is a major factor for shape variation. After quantifying the morphological shapes numerically through RCC (Radius Centroid Contour) skills (Tan et al. 2000; Kong et al. 2007), each phenotype, as a datum, is in the form of samples of functional curves with respective to the spatial angle rather than scalars or vectors. By considering the curves as the function of spatial angle, FDA (Functional Data Analysis) models the trends as a smooth dynamic function of spatial angel and is able to handle infinite-dimensional or/and irregular sparse curves. Therefore, FDA is more appropriate than simply modeling the

random vectors like the multivariate analysis does (Muller 2005). FDA is also a quite new area and the first book on FDA was published in 1997 (Ramsay et al. 1997), hence FDA has hardly been used in image analysis area (Epifanio et al. 2011). Currently in the literature, there are only two articles implementing FDA to the shape analysis area. But they had nothing related to genetic shape mapping. They just applied the "fda" library installed in MATLAB or R to the shape data and had no specified statistical models either. All in all, to our best knowledge, our work stated in this paper is the first article that related to functional genetic mapping on shape analysis area.

As early as 1946, karhunen (Karhunen 1946) founded the theoretical ideas about the stochastic process in Hilbert space. Grenander (Grenander 1950) expanded the Karhunen Loeve Theorem to the functional data and proposed the first functional regression theory. Rao (Rao 1958) applied functional PCA (principal components analysis) to the growth curves. Ramsay (Ramsay et al. 1997; 2002) gave an detailed introduction on the functional PCA and its applications. The discussions of Ramsay make the functional regression models popular. Hall (Hall et al. 2006) summarized two different approaches to model the functional data. If the measurements are recorded on a sense grid of time points, then the data are typically termed as one curve per subject and the nonparametric approaches are employed. On the other hand, if the measurements are recorded sparsely and irregularly varying among subjects, then the time will be added into the model as a random variables and GEE (generalized estimating equations) are applied. The importance of smoothing in the estimation of functional PCA has been emphasized several times (Yao et al. 2005a; Rice et al. 1991). Lin (Lin et al. 2000) ignored the correlation structure and provided theoretical evidence in support of the semi-parametric GEE independence model for longitudinal data.

In our application, we use the AIC (Akaike Information Criterion) to choose the tuning parameters, such as the number of principal components, and the bandwidth (Yao et al. 2005a). Yao mentioned that AIC is more efficient than the leave-one-curve-out Cross Validation, as far as the computation cost is concerned.

4.2 Statistical Models

In Chapter 2, through the shape analysis skills, each leaf shape was uniquely represented by a RCC curve, denoted as $Y_{i1}, \ldots, Y_{iN_i}$. Here $ij,\ i = 1, \ldots, n,\ j = 1 \ldots, N_i$ denoting the jth spatial angle on the ith subject. Since the measurements might vary among different subjects, for instance, irregular or sparse data, we need to treat spatial angle as a random variable, denoted as $t_{ij},\ i = 1, \ldots, n,\ j = 1, \ldots, N_i$, lying in the compact interval $\mathcal{T}$. Then for each subject i, the observations show up in pairs (t_{ij}, Y_{ij}).

In order to figure out the possible QTLs, forming a total of three genotypes, that control the shape curve variations, we take advantage of the association between the observable markers (M) and the latent but unobservable QTLs. The basic statistical model for QTL mapping is a mixture model, in which each observation (t_{ij}, Y_{ij}) is assumed to have arisen from one of the three QTL genotypes, each genotype $(c = 1, 2, 3)$ being modeled from a density function (the Gaussian Process is assumed). Therefore, considering the QTL effect associated with the marker M, we have the model

$$Y_{ij} = \sum_{c=1}^{3} \xi_{ic} X_{ic}(t_{ij}) + \varepsilon_{ij}, \quad i = 1, \ldots, n, j = 1, \ldots, N_i, \tag{4.1}$$

where ξ_{ic} is an indicator variable describing a possible QTL genotype c for subject i ($c = 1$ for AA, 2 for Aa, and 3 for aa) which is defined as 1 if a particular genotype is observed and 0 otherwise. $X_{ic}(t)$ is a smooth random trajectory of an underlying stochastic process in $L_2(\mathcal{T})$ for a particular genotype, $c = 1, \ldots, 3$; and ε_{ij} is the experimental error and assumed to be independent and identically distributed as $N(0, \sigma^2)$. We also assume that t_{ij} are independent and identically distributed and $X_i's$, $t_{ij}'s$ and $\varepsilon_{ij}'s$ are totally independent on each other.

For a fixed genotype c, the mean effect of $X_{ic}(t)$ is $\mu_c(t)$ and covariance function is $G_c(s,t) = cov\{X_c(s), X_c(t)\},\ s \in \mathcal{T},\ t \in \mathcal{T}$. Here $\mu_c(t)$ is interpreted as the genotypic value of the QTL for the phenotypes with genotype c. Throughout this paper, it is assumed that $\mu_c(t)$ is a smooth function of t, and $G_c(s,t)$ is a positive definite and bivariate smooth function of s and t, for $s, t \in \mathcal{T}$. Then, the path of $X_c(t)$ is also a smooth function. The "smooth" refers to twice continuously differ-

entiable. The idea of above model (4.1) is that the observed data are decomposed into a smooth process that is sampled on a discrete dense grid and additive noise.

Based on model (4.1), the likelihood of Y_{ij} can be expressed as

$$L(\mathbf{\Delta}) = \prod_{i=1}^{n} \left[\sum_{c=0}^{2} \pi_{c|i} \; f_c\{Y_{ij}|\mu_c(t_{ij}), \Sigma_c(t_{ij}, t_{il})\} \right], \tag{4.2}$$

where $\mathbf{\Delta}$ denotes all the unknown parameters specifying the likelihood, $\pi_{c|i}$ is the conditional prorability for the individual i to carry a QTL genotype c, and $f_c\{Y_{ij}|\mu_c(t_{ij}), \Sigma_c(t_{ij}, t_{il})\}$ is the probability density function of the observation Y_{ij} at QTL genotype c, which is assumed to be the Gaussian Process with mean function $\mu_c(t)$ and covariance function $G_c(s,t)$, $s \in \mathcal{T}$, $t \in \mathcal{T}$. For a natural population, the mixture proportions ($\pi_{c|i}$) of each QTL genotype c in likelihood (4.2) are described in terms of allele frequencies at the markers and QTLs and their linkage disequilibria (LD) (Wang et al. 2004; Wu et al. 2007). The size of LD reflects the degree to which the markers and QTLs are associated.

4.2.1 Semi-parametric Independent Model

Lin (Lin et al., 2000) proposed a semi-parametric model for longitudinal data using local polynomial kernel. They stated that the estimator is efficient if ignoring within subject correlation. In that case, independence is assumed and the covariance matrix will be an identity matrix.

Following Lin's idea, for a fixed c, the covariance matrix $G_c(s,t)$ will be

$$G_c(s,t) = \begin{cases} G(t,t) + \sigma^2, & s = t \\ 0, & s \neq t \end{cases} \tag{4.3}$$

Then, substituting this covariance structure (4.3) into model (4.1), we have

$$Y_{ij} = \sum_{c=1}^{3} \xi_{ic}[\mu_c(t_{ij}) + \varepsilon_c^*(t_{ij})], \quad i = 1, \ldots, n, j = 1, \ldots, N_i, \tag{4.4}$$

where $\varepsilon_c^*(t_{ij})$ are independent with mean 0, and variance $\sigma_c^{*2}(t_{ij})$. (4.3) implies that $\sigma_c^{*2}(t_{ij}) = G_c(t,t) + \sigma^2$ if $i = j$. And if $i \neq j$, $\sigma_c^{*2}(t_{ij}) = 0$. Therefore, in this independent model, for fixed QTL genotype c, the density function in likelihood function (4.2) will be equivalent to

$$f_c\{Y_{ij}|\mu_c(t_{ij}), G_c(s_{ij},t_{ij}) + \sigma^2 \cdot I\} = f_c\{Y_{ij}|\mu_c(t_{ij}), \sigma_c^{*2}(t_{ij})\} \sim N(\mu_c(t), \sigma_c^{*2}(t)).$$

4.2.2 Nonparametric Functional PCA Model

If counting the correlation structure, based on the above definition, we have $cov\{X_c(s), X_c(t)\} = G_c(s,t),\ s \neq t,\ s \in \mathcal{T},\ t \in \mathcal{T}$ and $cov\{X_c(t), X_c(t)\} = G_c(t,t) + \sigma^2,\ t \in \mathcal{T}$. The main idea of functional PCA is to interpret $G_c(s,t)$ as the kernel of a liner mapping on the space $L_2(\mathcal{T})$ of square-integrable functions on $\mathcal{T}$, mapping $f \in L_2(\mathcal{T})$ to $A_G f \in L_2(\mathcal{T})$ defined by (Hall et al., 2006)

$$(A_G f)(t) = \int_{\mathcal{T}} f(s) G_c(s,t) ds.$$

An eigenfunction v of the operator A_G is a solution of the equation $(A_G v)(t) = \lambda v(t)$, with eigenvalue λ. For the fixed c, we assume that the operators A_G have a sequence of smooth orthonormal eigenfunctions v_{qc} satisfying $\int_{\mathcal{T}} v_{kc}(t) v_{qc}(t) dt = \delta_{kq}$ (here δ_{kq} is the Kronecker symbol), with ordered eigenvalues $\lambda_{1c} \geq \lambda_{2c} \geq \ldots \geq 0$. By Mercer's Theorem (Indritz 1963), applying a spectral decomposition on the function G_c, Hilbert-Schmidt kernel, yields

$$G_c(s,t) = \sum_{q=1}^{\infty} \lambda_{qc} v_{qc}(s) v_{qc}(t). \tag{4.5}$$

Since the eigenfunctions v_{qc}'s form a complete orthonormal sequence on $L_2(\mathcal{T})$, by the generalized Fourier expansion ($Karhunen - Loeve$ Theorem (Karhunen 1946) or functional principal component expansion) of the stochastic process X_{ic}, we have

$$X_{ic}(t) = \mu_c(t) + \sum_{q=1}^{\infty} \zeta_{iqc} v_{qc}(t), \tag{4.6}$$

where the sum is defined in the sense of L_2 convergence, with uniform convergence, and

$$\zeta_{qc} =< X_c - \mu_c, v_{qc} >= \int_{\mathcal{T}} (X_c(t) - \mu_c(t))v_{qc}(t)dt \tag{4.7}$$

are uncorrelated random variables with $E(\zeta_{qc}) = 0$, and $var(\zeta_{qc}) = \lambda_{qc}$, subject to the L_2 convergence, i.e. $\Sigma_q \lambda_{qc} < \infty$. ζ_{qc} are frequently referred to as the qth functional principal component score or the qth dominant modes of variation effect.

Combining the above equations (4.1) and (4.6), we have

$$Y_{ij} = \sum_{c=1}^{3} \xi_{ic}[\mu_c(t_{ij}) + \sum_{q=1}^{\infty} \zeta_{iqc} v_{qc}(t_{ij})] + \varepsilon_{ij}, \quad i = 1, \ldots, n, j = 1, \ldots, N_i. \tag{4.8}$$

4.2.3 Parameter Estimate

Using the observation data set $\mathcal{D} = \{(t_{ij}, Y_{ij}), 1 \leq j \leq N_i, 1 \leq i \leq n\}$, we will propose the estimating procedures for all unknown parameters $\hat{\mu}(t), \hat{G}_c(s,t)$ and $\hat{\sigma}^2$ stated in model (4.1). Applying the smoothing procedures, we are able to obtain consistent estimates for μ_c and G_c.

If the estimator of $\hat{\mu}(t)$ is obtained, then we can compute a rough estimate of covariance from all observed pairs of data points for the same subject, (t_{ij}, Y_{ij}), and (t_{il}, Y_{il}) by

$$\bar{G}_{ijlc} = (Y_{ij} - \hat{\mu}_c(t_{ij}))(Y_{il} - \hat{\mu}_c(t_{il})).$$

The local linear smoother estimate $\hat{G}_c(s,t)$ for $G_c(s,t)$ is obtained by minimizing (Yao et al. 2006; Muller et al. 2006)

$$\sum_{i=1}^{n} \pi_{c|i} \sum_{1 \leq j \neq l \leq N_i} K_2(\frac{t_{ij} - s}{h_G}, \frac{t_{il} - t}{h_G})\{\bar{G}_{ijlc} - \beta_0 - \beta_{11}(s - t_{ij}) - \beta_{12}(t - t_{il})\}^2, \tag{4.9}$$

with respect to $\beta = (\beta_0, \beta_{11}, \beta_{12})$. Here $K_2(\cdot, \cdot)$ is the bivariate nonnegative compactly supported kernel function used as weights for locally weighted least squares smoothing in two dimensions. As a valid kernel function, K_2 is symmetric with zero mean and finite variance and $\| K_2 \|^2 = \int \int K_2^2(s,t)dsdt < \infty$. h_G is the bandwidth corresponding to the kernel function K_2. The minimization with respect to (4.9) yields $\hat{G}_c(s,t) = \hat{\beta}_0(s,t)$ (Muller et al. 2006), which can be solved in a close

form

$$\hat{G}_c(s,t) = \frac{\sum_{i=1}^{n} \pi_{c|i} \sum_{1\le j\ne l\le N_i} \bar{G}_{ijlc} K_2(\frac{t_{ij}-s}{h_G}, \frac{t_{il}-t}{h_G})}{\sum_{i=1}^{n} \pi_{c|i} \sum_{1\le j\ne l\le N_i} K_2(\frac{t_{ij}-s}{h_G}, \frac{t_{il}-t}{h_G})} \tag{4.10}$$

Once the smoothed covariance function $\hat{G}_c(s,t)$ is computed from (4.10), it is then discretized on a suitable finite grid and represented as a covariance matrix.

The estimate of eigenfunctions are obtained by the corresponding spectral decomposition on $\hat{G}_c(s,t)$ (Rice et al. 1991). To be more specific, $\hat{\lambda}_{qc}$ are eigenvalues of $\hat{G}_c$, given by

$$\int_{\mathcal{T}} \hat{G}_c(s,t)\hat{v}_{qc}(s)ds = \hat{\lambda}_{qc}\hat{v}_{qc}(t).$$

And $\hat{v}_{qc}$ are the eigenfunctions corresponding to $\hat{\lambda}_{qc}$, satisfying $\int_{\mathcal{T}} \hat{v}_{qc}^2(t)dt = 1$ and $\int_{\mathcal{T}} \hat{v}_{pc}\hat{v}_{qc}(t)dt = 0$ if $p \ne q$. Here $\hat{G}_c$ also agrees an empirical version of the expansion (4.5)

$$\hat{G}_c(s,t) = \sum_{q=1}^{\infty} I(\hat{\lambda}_{qc} > 0)\hat{\lambda}_{qc}\hat{v}_{qc}(s)\hat{v}_{qc}(t) \tag{4.11}$$

Here the I is an identity function. The positive definiteness of the estimated covariance matrix $\hat{G}_c(s,t)$ is not always guaranteed and might be a problem in practical applications. Yao proposed a trick to avoid this (Yao et al. 2003; Muller 2005). Once $\hat{\lambda}_{qc}$ and $\hat{v}_{qc}$ are obtained, we should check whether or not $\hat{\lambda}_{qc} > 0$. If some $\hat{\lambda}_{qc}$ is negative, then we drop this negative eigenvalue and its corresponding eigenfunction, and reconstitutes the estimate from the remaining eigenvalue and eigenfunction estimates.

After $\hat{v}_{qc}$ and $\hat{\lambda}_{qc}$ are got, the fitting of individual trajectories requires estimation of functional principal component scores. By the discretization on the equation (4.7), plugging $\hat{\mu}_c$ and $\hat{v}_{qc}$ into a Riemann sum approximation of the integral, we have

$$\hat{\zeta}_{iqc} = \Sigma_{j=1}^{N_i}(Y_{ij} - \hat{\mu}(t_{ij}))\hat{v}_{qc}(t_{ij})(t_{ij} - t_{i,j-1}) \tag{4.12}$$

setting $t_{i0} = 0$ (Muller 2005). In our motivating example, the spatial angle are recorded densely spaced so approximation this sum in formula (4.12) to the integral is reasonable. If the data is noisy, sparse or irregular, another approach called PACE (Principal Analysis through Conditional Expectation) can be referred (Yao et al. 2005a).

After above complicated procedures, we are able to get estimators related to the covariance structure. Now, it is time to estimate the mean. The idea is that we will estimate the mean from residuals after removing the covariance part. Define

$$Y_{ij}^{*} = Y_{ij} - \sum_{c=1}^{3} \xi_{ic} \sum_{q=1}^{\infty} \hat{\zeta}_{iqc} \hat{v}_{qc}(t_{ij}), \quad i = 1, \ldots, n, j = 1, \ldots, N_i.$$

Then from model (4.8), it is easy to understand that

$$Y_{ij}^{*} = \sum_{c=1}^{3} \xi_{ic} \mu_c(t_{ij}) + \varepsilon_{ij}, \quad i = 1, \ldots, n, j = 1, \ldots, N_i. \tag{4.13}$$

Model (4.4) and (4.13) are quite similar in that they are both Mixture Gaussian Process with independent covariance matrix. Therefore, we can apply the EM algorithm here if the dimension is feasible. For the ultra-high dimension case in our motivating example, we need to simultaneously apply nonparametric kernel smoothing.

Let $\mathcal{W} = \{\omega_1, \ldots, \omega_m\} \in \mathcal{T}$ be m knots at which the mean and variance functions can be estimated. For any $\omega_l \in \mathcal{W}$, we approximate $\mu_c(t_{ij})$ by $\mu_c(\omega_l)$, and $\sigma_c^{*2}(t_{ij)}$ by $\sigma_c^{*2}(\omega_l)$ for t_{ij} located within bandwidth h_u neighborhood of ω_i. Then the corresponding local logikelihood function of model (4.13) will be

$$\sum_{i=1}^{n} \log \left[\sum_{c=1}^{3} \pi_{c|i} \prod_{j=1}^{N_i} f_c\{Y_{ij}^{*} | \mu_c(t_{ij}), \sigma^2\} \right] K_1(\frac{t_{ij} - \omega_l}{h_u}), \tag{4.14}$$

Here $K_1(\cdot)$ is a nonnegative univariate compactly supported kernel function that is used as weights for local polynomial smoothing in one dimension satisfying the basic requirement for a kernel function. The log-likelihood function of model (4.4) will be very similar if changing Y_{ij}^{*} by Y_{ij} and σ^2 by $\sigma_c^{*2}(t_{ij})$. After a derivation, we can apply the EM algorithm based on $\mathcal{W} = \{\omega_1, \ldots, \omega_m\} \in \mathcal{T}$. The E step is designed to calculate the posterior probability with which subject i has QTL genotype c given its marker and phenotypic information, expressed as

$$\Pi_{c|i} = \frac{\pi_{c|i} \prod_{j=1}^{N_i} f_c\{Y_{ij}^{*} | \mu_c(t_{ij}), \sigma^2\}}{\sum_{k=1}^{3} \pi_{k|i} \prod_{j=1}^{N_i} f_k\{Y_{ij} | \mu_k(t_{ij}), \sigma^2\}}$$

Using the calculated posterior probabilities, the M step is derived to solve the haplotype frequencies expressed as (Wang et al. 2004)

$$
\begin{aligned}
\mu_c(\omega_l) &= \frac{\sum_{i=1}^{n} \Pi_{c|i} \sum_{j=1}^{N_i} Y_{ij}^* K_1(\frac{t_{ij}-\omega_l}{h_u})}{\sum_{i=1}^{n} \Pi_{c|i} \sum_{j=1}^{N_i} K_1(t_{ij}-\omega_l)}, \\
\hat{\sigma}^2 &= \frac{1}{\sum_{i=1}^{n} N_i} \sum_{i=1}^{n} \sum_{c=1}^{3} \Pi_{c|i} \sum_{j=1}^{N_i} [Y_{ij}^* - \mu_c(t_{ij})]^2, \\
\sigma_c^{*2}(\omega_l) &= \frac{\sum_{i=1}^{n} \Pi_{c|i} \sum_{j=1}^{N_i} [Y_{ij} - \mu_c(\omega_l)]^2 K_1(\frac{t_{ij}-\omega_l}{h_u})}{\sum_{i=1}^{n} \Pi_{c|i} \sum_{j=1}^{N_i} K_1(t_{ij}-\omega_l)}, \\
\hat{p}_{11} &= \frac{1}{2N} \left[\sum_{i=1}^{N_1} (2\Pi_{i1} + \Pi_{i2}) + \sum_{i=1}^{N_2} (\Pi_{i1} + \theta \Pi_{i2}) \right], \\
\hat{p}_{10} &= \frac{1}{2N} \left[\sum_{i=1}^{N_1} (\Pi_{i2} + 2\Pi_{i3}) + \sum_{i=1}^{N_2} (\Pi_{i3} + (1-\theta) \Pi_{i2}) \right], \\
\hat{p}_{01} &= \frac{1}{2N} \left[\sum_{i=1}^{N_3} (2\Pi_{i1} + \Pi_{i2}) + \sum_{i=1}^{N_2} (\Pi_{i1} + (1-\theta) \Pi_{i2}) \right], \\
\hat{p}_{00} &= \frac{1}{2N} \left[\sum_{i=1}^{N_3} (\Pi_{i2} + 2\Pi_{i1}) + \sum_{i=1}^{N_2} (\Pi_{i3} + \theta \Pi_{i2}) \right],
\end{aligned}
\tag{4.15}
$$

where $\theta = p_{11}p_{00}/(p_{11}p_{00} + p_{10}p_{01})$.

The last thing need to mention is the selection of the tuning parameter. As we known, the number of eigenfunctions used to approximate the infinite-dimensional longitudinal process and the degree of smoothness are simultaneously determining the performance of the model. And, in practical implementation, the degree of smoothness is determined by the number and location of knots, and the size of the bandwidth. One-curve-leave-out CV (Cross Validation) is a very popular method to select the tuning parameters. Without loss much efficiency, we use AIC instead. Yao (Yao et al. 2005a) compared two methods and pointed out that AIC is computationally more efficient but the results are similar to those obtained by cross

validation. Then the optimal $\hat{K}$ is chosen by minimizing

$$AIC(\hat{K}) \propto \sum_{i=1}^{n}\sum_{c=1}^{3}\pi_{c|i}\{-\frac{1}{2}(Y_i-\hat{\mu}_{ic}-\sum_{q=1}^{\hat{K}}\hat{\zeta}_{iqc}\hat{v}_{iqc})^T\Sigma_i^{-1}(Y_i-\hat{\mu}_{ic}-\sum_{q=1}^{\hat{K}}\hat{\zeta}_{iqc}\hat{v}_{iqc})\}+3*\hat{K}, \tag{4.16}$$

here $Y_i=(Y_{i1},\ldots,Y_{iN_i})^T$, $\hat{\mu}_{ic}=(\hat{\mu}_c(t_{i1}),\ldots,\hat{\mu}_c(t_{in_i}))^T$, $\Sigma_i=diag(\hat{\sigma}^2,\ldots,\hat{\sigma}^2)$, and $\hat{v}_{iqc}=(\hat{v}_{qc}(t_{i1}),\ldots,\hat{v}_{qc}(t_{in_i}))$.

Finally, we summarize the estimating procedures in the following:

1) From the marker information, give an initial guess of $\hat{p},\hat{q},\hat{D}$, and $\hat{\pi}_{c|i}$, through which give a rough initial estimate $\hat{\mu}_c(t)$ and initial $\hat{\xi}_{ic}$. Then, compute the rough estimate of the covariance matrix $\bar{G}_{ijlc}$.

2) Estimate the smooth covariance surface $\hat{G}_c(s,t)$ by two-dimensional local liner smoothing formula (4.9,4.10).

3) Compute the eigenfunctions $\hat{v}_{qc}(t)$ and eigenvalues $\hat{\lambda}_{qc}$ of $\hat{G}_c(s,t)$. Only keep the terms with positive eigenvalues.

4) Use formula (4.12) to compute the functional principal component scores $\hat{\zeta}_{iqc}$

5) Compute $\hat{Y}^*_{ij}$.

6) Apply EM algorithm (4.15) to update $\hat{\mu}_c(t)$, $\hat{\xi}_{ic}$, $\hat{\sigma}^2$, $\hat{p}_{11},\hat{p}_{10},\hat{p}_{01},\hat{p}_{00},\hat{p},\hat{q},\hat{D}$, and $\hat{\pi}_{c|i}$.

Then back to step 1). Repeated until convergence.

4.2.4 Hypothesis Tests

From the model structure description in the beginning of this section, it is easy to understand that significant QTL effects show evidence in the significant differentiation among three genotype curves, which are described by the three smooth mean functions $(\mu_1(t),\mu_2(t),\mu_3(t))$ in model (4.2). Therefore, the significance of a shape QTL can be tested by using the following hypotheses:

$$\begin{aligned} H_0 &: \mu_j(t)\equiv\mu(t),(j=1,2,3) \\ H_1 &: \textit{At least one of the equalities above does not hold}, \end{aligned}$$

where the H_0 corresponds to the reduced model, in which the data can be fit by a single function, and the H_1 corresponds to the full model, in which there exist

three QTL genotype-specific functions to fit these data. The test statistics for the above hypotheses is calculated as the log-likelihood ratio (LR) of the reduced to the full model. The degree of freedom is $2 * N_i$ (=720 in our case). An empirical approach based on permutation tests is used to determine the critical threshold (Churchill et al. 1994).

In our genetic design, we locate the true but unobservable genes by its association with the marker. Therefore, the linkage disequilibrium is our basic assumption. After a significant QTL is found to exist, it is necessary for us to test whether or not this QTL exists a significant linkage disequilibrium with the given marker using the hypotheses:

$$H_0 : D = 0 \;\; vs. \;\; H_1 : D \neq 0,$$

where the H_0 corresponds to the reduced model, in which the marker and QTL are at the linkage equilibrium (i.e. independence), and the H_1 corresponds to the full model, in which there is a linkage disequilibrium between the marker and QTL. The test statistics for this hypothesis is $\chi^2 = 2nD^2/(p(1-p)q(1-q))$, which follows a χ^2 distribution with one degree of freedom (Wu et al. 2007).

4.3 Numerical Implementation

The objects of our analysis is to identify significant QTLs that regulate the variation of leaf shapes in a natural population of poplar species, *Populus szechuanica var. tibetica*, distributed throughout the Tibet Plateau. Applying the advanced new model in Chapter 4 to the data set described in Chapter 2.4.1 and using the same genetic design in Chapter 2.3.3 and 2.3.4, in the following, we will describe the implementation result.

Now, apply both the semi-parametric independent model (4.4) and the non-parametric functional PCA model (4.8), we get all smooth estimates on the mean function $\mu_c(t)$, variance function $\sigma_c^{*2}(t)$, and the eigenfunctions extracting the dominant modes of variation $v_{qc}(t)$, for three different genotypes. The optimal number of principal components chosen by AIC is 6. We also roughly chose the bandwidth to be 0.08, and chose 50 knots equally spaced to make sure that there are enough

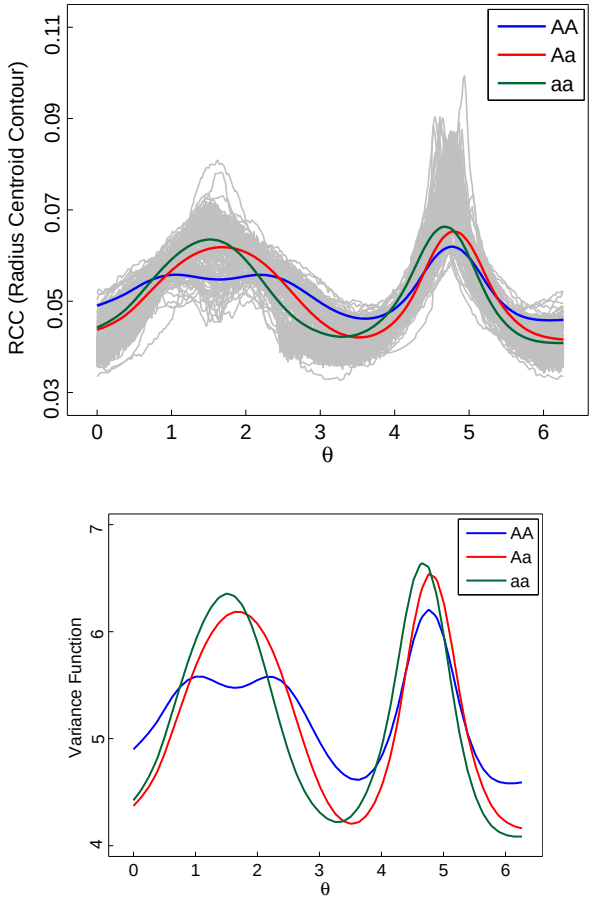

Figure 4.1. The smooth estimates of the mean function $\mu_c(t)$ (Top Panel) and smooth estimates of the variance function $\sigma_c^{*2}(t)$ (Bottom Panel) of the RCC curves for three genotypes obtained from semi-parametric independent model (4.4)

neighborhood points within each smoothing window.

In Fig. 4.1, we illustrate the smoothing estimates of $\hat{\mu}_c(t)$, and $\hat{\sigma}_c^{*2}$ obtained from model (4.4). The mean has an approximate periodic trend because we measured the observation circularly from 0 to 2π clockwise. But it is not exactly periodic and the two peaks are not symmetric because the leaf shape is not a round circle. For all the three genotypes, there are two peaks, with the first one located near $t = \frac{\pi}{2}$ and the second peak located near $t = \frac{3\pi}{2}$. The variance is clearly nonstationary, with high variability near $t = \frac{\pi}{2}$ and $t = \frac{3\pi}{2}$, which corresponding to the tip and bottom area of the leaf blades. Therefore, we claim that the most significant effect of genes is found to narrowing the leaf shape from the tip and bottom parts. In Fig. 4.2, we illustrate the smoothing estimates of $\hat{\mu}_c(t)$ obtained

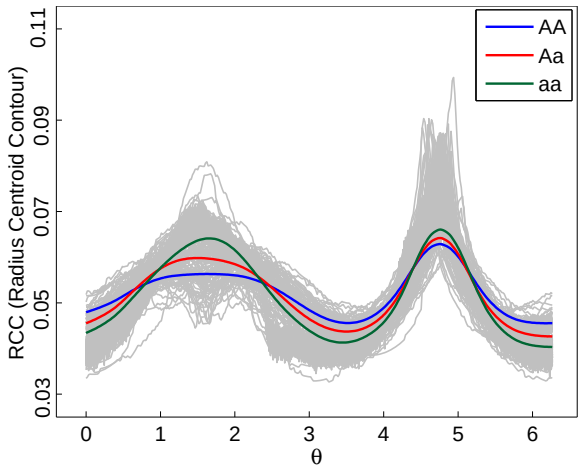

Figure 4.2. The smooth estimates $\mu_c(t)$ of the RCC curves for three genotypes obtained from nonparametric functional PCA model (4.8).

from model (4.8). The two estimated mean functions $\hat{\mu}_c(t)$ got from two different models are very similar, except some small tiny details. From both results, it seems that the genotype AA (in blue color) control leaves with more round shape and short tips.

Next, consider the eigenfunction decomposition of the smooth estimated covariance function $\hat{G}_c(s,t)$. Six eigenfunctions $v_{qc},\ q = 1,\ldots,6$ shown in Fig. 4.3 are used to approximate the infinite dimensional process. The first two eigenfunctions have similar trends as the mean function. We also notice that the first eigenfunction $\hat{v}_{1c}$ have negative values along intervals $(0, \frac{\pi}{4})$, $(\frac{3\pi}{4}, \frac{5\pi}{4})$, and $(\frac{7\pi}{4}, 2\pi)$, for all $c = 1, 2, 3$. It means that a RCC curve Y_i with a positive (or negative) functional principal component score $\hat{\zeta}_{i1c}$ along the direction of $\hat{v}_{1c}$ tends to have smaller (or larger) values in these intervals than the overall population average. In addition, the results agrees with the estimated variance function in Fig. 4.1. For the genotype $c = 1$, these eigenfunctions account for $\%63.66, \%14.36, \%11.45, \%4.30, \%2.67$, and $\%1.82$ (altogether explains $\%98$) of the total variation, respectively. For the genotype $c = 2$, these eigenfunctions account for $\%68.30, \%11.85, \%8.02, \%5.58, \%3.20$, and $\%1.50$ (altogether explains $\%98$) of the total variation, respectively. For the genotype $c = 3$, these eigenfunctions account for $\%43.86, \%28.97, \%9.10, \%7.63$, $\%3.78$, and $\%3.45$ (altogether explains $\%96$) of the total variation, respectively. It

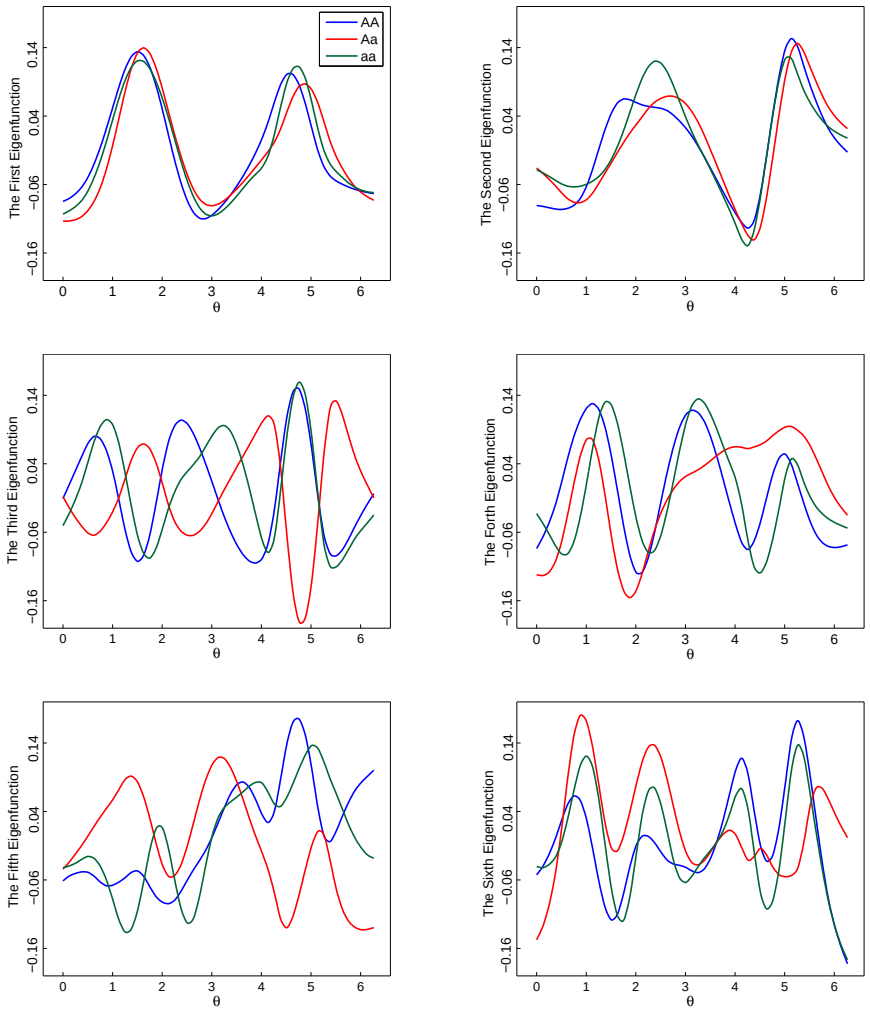

Figure 4.3. Smooth estimates of the first six eigenfunctions of RCC curves for three genotypes obtained from model (4.8).

seems that different genotype has different partition on the total variations. The first eigenfunction of genotype $c = 3$ (in color green) does not contribute as much as that of genotype $c = 1$ (in color blue) or $c = 2$ (in color red). And the second eigenfunction of genotype $c = 3$ has almost double contribution as that of $c = 1$ and $c = 2$.

Chapter 5

Discussion and Future Work

This chapter reviews and discusses the contribution of this dissertation in Section 5.1. Then, followed by a discussion in Section 5.2 of future work, including unsolved goals for future research, and possible new areas of application of the presented methods.

5.1 Summary of Contributions

Knowledge about the genetic mechanisms for shape variation has far-reaching implications for a range spectrum of scientific disciplines (Ricklefs and Miles 1994; Klingenberg 2010). Comparing the anatomical and shape feature of organisms has been a central element of biology for centuries (Bookstein 1978; Klingenberg and Leamy 2001; Monteiro et al. 2002; Adams et al. 2004). For example, as one of the most conspicuous aspects of a plant's phenotype, leaf shape has been used to provide an intricate link between biological structure and function in changing environments (Tsukaya 2005). With an increasing interest in studying shape genetics (Weber et al. 1999; Langlade et al. 2005; Mezey et al. 2005; Leamy et al. 2008), we have now developed a computational model for mapping specific quantitative trait loci (QTLs) that contribute to shape variation by using leaf shape as an example of demonstration.

In Chapter 2, we present a new statistical model for mapping shape QTLs in a segregating population. We did a very accurate shape analysis by using Radius Centroid Contour to represent a shape, and using the General Procrustes analysis

to align the shapes to minimize any variation caused by translation, scale, and rotation. The new model embeds shape analysis within a mixture model framework in which different types of morphological shape are defined for individual genotypes at a QTL. The advantage of shape mapping lies in its capacity to quantify subtle differences in any corner of a morphological shape and detect specific QTLs that contribute to these differences. Results from both real data and simulation studies suggest that the model has reasonably high power to detect a QTL that control shape difference. Even with a modest sample size, the model is able to discern the effect of a QTL not only in global shape variability (such as elongating, narrowing), but also in local shape variability (such as tail leaning). We also computing the ratio of length of petiole over the length of blade in order to compare the traditional method with our shape analysis method to illustrates the improvements. It turns out that shape analysis is able to locate much more QTLs than traditional simple method. Finally, we find that latitude is strongly significant (under significant level 0.1) in effecting the QTL that control the global shape variability.

Unlike traditional morphological data that concern single measurements of an object, such as size or weight, shape data that capture the proportions and relative positions of various parts of the object are viewed as a photograph (Klingenberg, 2010). We incorporate statistical models for extracting shape information from photographs into a mixture-model framework for QTL mapping. Different aspects of a shape are specified by orthogonal principal components (PCs). Statistical parameters that define genotype-specific differences in shape-related PCs are estimated by implementing the EM algorithm. This so-called shape mapping model enables geneticists to examine the control patterns of specific QTLs on the origin, properties, and functions of leaf shape.

Our model is, to some extent, similar to the approaches for shape-QTL mapping by Langlade et al. (2005) and Klingenberg (2003; 2010) in terms of the use of PCA to reduce data dimension. However, our model is distinct from the latter two types of shape modeling. First, rather than using a limited number of sparse anatomical landmarks, i.e., those points, assigned by an expert, that corresponds between objects of study in a way meaningful in the disciplinary context, our model detects and capitalizes on mathematical landmarks that are located on an object according to its specific mathematical or geometrical property. This shape variation can

be well described by mathematical landmarks. Second, our model expresses a series of coordinates taken on an object as a radius-centroid-contour (RCC) curve (i.e., a function of radial angle at the centroid). Thus, more powerful statistical approaches, such as longitudinal data analysis of RCC, can be incorporated into a QTL mapping framework, enhancing the biological relevance of shape mapping.

To demonstrate its application, shape mapping was used to map QTLs for leaf shape with the data collected from a natural population of *Populus szechuanica var. tibetica*. This poplar species is naturally distributed in the mountains at an altitude of 1100-4600 m in the southwestern China (Hamzeh and Dayanandan 2004), providing an ideal model system to study the genetics of leaf morphology and its relationship with ecological adaptations. Interestingly, we detected a number of shape QTLs associated with microsatellite markers by shape mapping. From the PCA of shape data extracted from leaf images, six major PCs were detected to together explain 88.1% of the variation among leaf shapes. By mapping these PCs, we identified the QTLs that leaf shape from various morphological aspects. Of these QTLs, those obtained through the major PC that accounts for almost a half of the variation determine the overall or global shape variation of leaves, whereas those through the other minor PCs control the local shape variation. It is worthwhile to further investigate specific QTLs that determine the ecological relationships of leaf shape and environmental factors by sampling more poplar trees from different populations.

Different from Langlade et al.'s (2005) work, shape mapping focuses on mapping leaf shape by separating it from leaf size through uniformly scaling leaf images. Although this helps to clarify the genetic control of leaf shape in its own right, the biological functions of leaf size and shape may be inherently linked (Wu et al. 1997). Our model can be readily extended to perform simultaneous mapping of leaf shape and leaf size within a unifying framework, allowing the pleiotropic test of QTL effects on these two leaf traits. Also, given its critical role in trait control (Wang et al. 2011), epistasis between different QTLs should be modeled and tested by implementing multi-QTL genotypes into the mixture likelihood. With the availability of data collected for large-scale and complex problems in genetic, ecological and physiological research, our shape mapping model described will provide a powerful analytical tool to effectively and efficiently test and build

hypotheses, and extract useful information for scientific inferences and prediction.

In Chapter 3, we present a new statistical model for mapping shape QTLs in a segregating population. The new model embeds shape analysis within a mixture model framework in which different types of morphological shape are defined for individual genotypes at a QTL. The model was solved using a traditional shape correspondence analysis approach and EM algorithm. The advantage of shape mapping lies in its capacity to quantify subtle differences in any corner of a morphological shape and detect specific QTLs that contribute to these differences. Results from simulation studies suggest that the model has reasonably high power to detect a QTL that control shape difference. Even with a modest sample size, the model is able to discern the effect of a QTL with a small effect on morphological shape. The model can be easily extended to model epistatic interactions on morphological shape by including more components in the mixture model.

When specific genes that control morphological shape and physiological function are identified, we are in an excellent position to address fundamental questions related to growth, development, adaptation, domestication, and human health. In the past decades, the increasing availability of DNA-based markers has inspired our hope to map genes or quantitative trait loci (QTLs) for complex phenotypes (Zeng 1994; Jansen et al. 1994; Xu et al. 1995; Lynch et al. 1998; Broman et al. 2002; Zou et al. 2004; Yi et al. 2005). However, only several studies have been alert to map so-called shape genes; a few successful examples are the positional cloning of genes for fruit shape in tomato (Tansley 2004; Frary et al. 2000; Liu et al. 2002; Xiao et al. 2008). These successes result from the fact that a major mutation occurs to determine shape difference. For many quantitatively inherited shape traits, genetic mapping will provide a powerful tool for characterizing QTLs affecting morphological shape. Klingenberg and colleagues (Klingenberg et al. 2001; Klingenberg 2001) have developed quantitative genetic theory to estimate the heritability of shape by integrating geometric shape analysis. This theory was used to map specific QTLs for morphometric shapes in the mouse (Leamy et al. 2008; Klingenberg et al. 2004). Airey et al. (Airey et al. 2006) used Procrustes superimposition to study shape differences in the cortical area map of inbred mice.

The model will be needed to be modified for integrating developmental events and their consequences into ontogenetic trajectories of shape. Modern biological

studies display an increasing interest in understanding shape variation in ontogenetic processes that bring about differentiation at an adult stage (Vioarsdottir et al. 2002; Quillevere et al. 2002). In a longitudinal study of radiographs of the Denver Growth Study, Bulygina et al. (Bulygina et al. 2006) investigated the morphological development of individual differences in the anterior neurocranium, face, and basicranium. The modified model can map the QTLs that cause variation in shape developmental trajectories. In biology, a cell or organ fulfill certain biological functions through its shape. Shape is thought to govern the extent and pattern of energy, matter and signal transduction through the surface and inner structure of the biological object. For this reason, an understanding of biological curvature and texture has received a surge of interest in structural biology. The new model can be extended to map the QTLs that determine a three-dimensional (3D) shape and texture of a biological object. Vision technologies have been developed to estimate the 3 D shape of an object from 2 D image data without information about its texture (albedo), its pose and the illumination environment (Romdhani et al., 2005; 2006). These technologies include a 3 D morphable model (3DMM) that represents the 3 D shapes and textures as a linear combination of shapes and textures principal components, a stochastic Newton optimization algorithm that is the 3DMM to a single facial image, thereby estimating the 3 D shape, the texture and the imaging conditions, and a multi-features fitting algorithm that uses not only the pixel intensity but also other image cues such as the edges and the specular highlights. Statistical models can be developed to map QTLs that control the 3 D shape and texture of a biological object with image data. A series of hypothesis tests about the genetic control of topological features (such as stepness and ridgeness) and texture of a shape will be formulated.

In Chapter 4, a mixture functional principal component analysis model for high dimensional functional data (cab be sparse or irregular spaced) is proposed. The functional data are modeled as samples of smooth random trajectories which are observed under additive noise. The noise include white noise caused by experimental error and a smooth random trajectories of variance extended from the concept of a variance function used in non- and semi-parametric regression analysis (Muller et al. 2006). After quantifying the shapes by RCC curves, we use local polynomial kernel smoothing to model the mean function, with the covariance structure

modeled by a set of orthogonal eigenfunctions and random coefficients referred as functional principal component scores (Yao et al. 2003; 2005a; 2006). In addition to the general benefits of other functional data analysis models, this model can handle the mixture gaussian process and integrate with QTL mapping. Since eigenfunctions have been interpreted as the modes of variation of the functional data, we are able to detect significant genes that regulate the variation of the shape.

5.2 Future Work

5.2.1 QTL Mapping on the Growth of Shape

Growth is a physiological process that each organism transforms essential nutrients into living protoplasm. Every physiological mechanisms such as body height, body weight, organs, hormonal, nutritional and so on, underly a growth process. From molecular biological aspect, every growth procedure consists of two steps: hypertrophy (increase in cell size), and hyperplasia(increase in cell number). During growth, morphological or physiological characteristics can be expressed as a function of time t from the embryogenesis to maturity. Fig. 5.1 gives an example of the growth of human body during fetal and postnatal stages. As you can notice, the proportion of head and limb to the whole body changes a lot during growth. The head takes almost 1/2 of the human body for a newborn infant. However, the head is only 1/8 of the whole body for an adult. The proportion of the legs to the whole body for an adult is as twice as that of the new born kids. All human being has the similar growth pattern. However, different individual might growth in a different speed, different form and proportion of the body shape. Gene is an important in determining this difference. Fig. 5.2 illustrates the developmental changes from very young leaves to very old leaves (Wu et al. 2003). As we can seen from Fig. 5.2, leaf shape has different growth changing style from human. Cultivated leaf shape and wild leaf shape also has different growth pattern.

By the research in quantitative genetic, growth is controlled by the particular variants of the genes. Unraveling the genetic control of growth is critical for human to understand the origin of life and ultimately control growth rule to the trajecto-

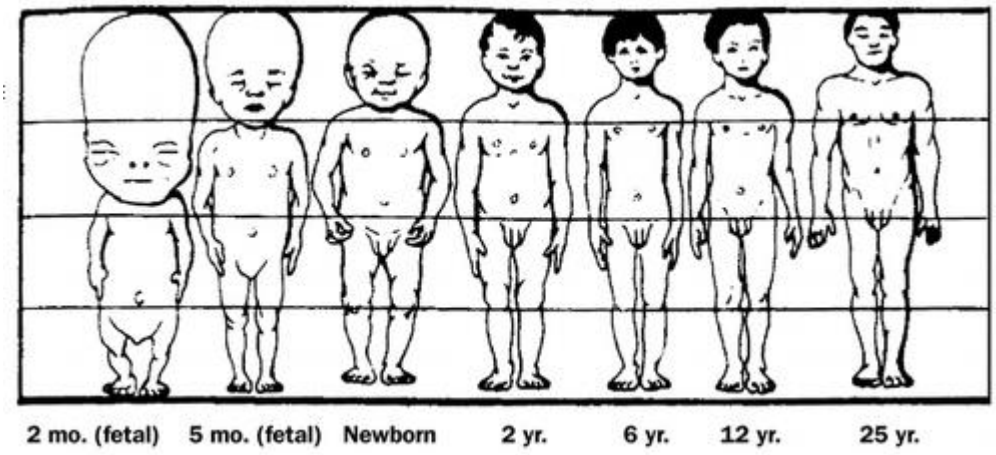

Figure 5.1. Changes in human body proportion from the second fetal month to adulthood.

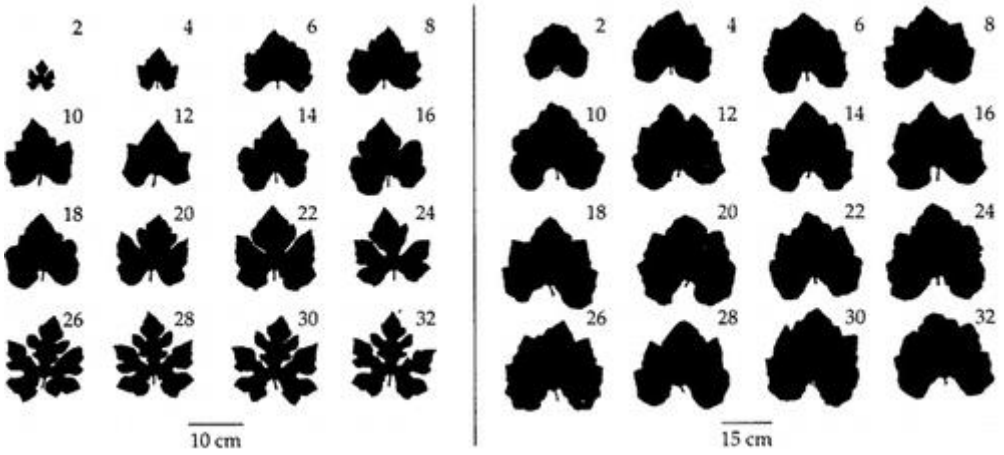

Figure 5.2. Developmental differences in leaf shape between wild *Cucurbita argyrosperma sororia* (left) and cultivated *Cucurbita argyrosperma argyrosperma* (right).

ries beneficial to human. However, growth is a very complex dynamic phenotype and hence figuring out the relationships between the growth and underlying genetic construction is difficult and challenging. Modern biological studies display an increasing interest in understanding shape variation in ontogenetic processes that bring about differentiation at an adult stage (Vioarsdottir et al. 2002; Quillevere et al. 2002). Nowadays, attempts have been made to unlock the genetic secrets behind phenotypic differentiation in developmental shape, understand the origin

and pattern of shape variation from a developmental perspective (Klingenberg et al. 2001; Klingenberg 2001), and predict the adaptation of morphological shapes in a range of environmental conditions (Tsukaya 2005). In a longitudinal study of radiographs of the Denver Growth Study, Bulygina et al. (Bulygina et al. 2006) investigated the morphological development of individual differences in the anterior neurocranium, face, and basicranium. The modified model can map the QTLs that cause variation in shape developmental trajectories.

From statistical aspect, longitudinal data is the repeated observations of the same individuals over a period of time or an interval. As we known, there is correlation exists for the same individual among different time period. Unlike traditional simple longitudinal data, the ontogenetic shape data is special longitudinal data in the form of a huge matrix rather than a single number at each time stage. Hence, the super high dimension for each time stage and the extremely complex covariance matrix among different time stage will make it extremely challenging to develop a statistical model for the shape variation in ontogenetic processes.

Our model in previous chapters will need to be modified for integrating developmental events and their consequences into ontogenetic trajectories of shape. This is our next work.

5.2.2 QTL Mapping on the 3D Morphological Shape

Due to the rapid development of photography, image data can be saved in 3D photo. Since we live in a 3D world, 3D image data can describe a shape even more accurately. For example, 2D image can never reflect the texture, skin, or movement of the human faces.

Vision technologies have been developed to estimate the 3D shape of an object from 2D image data without information about its texture (albedo), its pose and the illumination environment (Quillevere et al. 2002). These technologies include a 3D morphable model (3DMM) that represents the 3D shapes and textures as a linear combination of shapes and textures principal components, a stochastic Newton optimization algorithm that applying the 3DMM to a single facial image, thereby estimating the 3D shape, the texture and the imaging conditions, and a multi-features fitting algorithm that uses not only the pixel intensity but also other

image cues such as the edges and the specular highlights. Statistical models can be developed to map QTLs that control the 3D shape and texture of a biological object with image data. Then, our model can also be extended to map the QTLs that determine a three-dimensional (3D) shape and texture of a biological object. A series of hypothesis tests about the genetic control of topological features (such as stepness and ridgeness) and texture of a shape will be formulated.

Bibliography

[1] Abbasi, S., Mokhtarian, F., and Kittler, J. (2000). Enhancing CSS-Based Shape Retrieval for Objects with Shallow Concavities. *Image and Vision Computing*, **18**, 199 – 211.

[2] Adams, D.C., Rohlf, F.J., and Slice, D.E. (2004). Geometric morphometrics: ten years of progress following the "revolution". *Ital J Zool*, **71**, 5 – 16.

[3] Airey, D.C., Wu, F., Guan, M., and Collins, C.E. (2006). Geometric Morphometrics Defines Shape Differences in the Cortical Area Map of C57BL/6J and DBA/2J Inbred Mice. *BMC Neuroscience*, **63**, 1471 – 2202.

[4] Albertson, R.C., Streelman, J.T., Kocher, T.D., and Yelick, P.C. (2005). Integration and evolution of the cichlid mandible: the molecular basis of alternate feeding strategies. *Proc Natl Acad Sci USA*, **102**, 16287 – 16292.

[5] Basri, R., Costa, L., Geiger, D., and Jacobs, D. (1998). Determining the similarity of de-formable shapes. *Vision Res*, **38**, 2365 – 2385.

[6] Belongie, S., Malik, J., and Puzicha, J. (2002). Shape Matching and Object Recognition Using Shape Contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **24**, 509 – 522.

[7] Bernal, B. (2007). Size and shape analysis of human molars: Comparing traditional and geometric morphometric techniques. *J Comp Hum Biol*, **58**, 279 – 296.

[8] Bookstein, F.L. (1978). The Measurement of Bilogical Shape and Shape Change. *Springer-Verlag: New York*.

[9] Broman, K.W., and Speed, T.P. (2002). A model selection approach for the identification of quantitative trait loci in experimental crosses (with discussion). *J Roy Stat Soc Ser B*, **64**, 641 – 656.

[10] Bulygina, E., Mitteroecker, P., and Aiello, L. (2006). Ontogeny of facial dimorphism and patterns of individual development within one human population. *Am J Phys Anthrop*, **131**, 432 – 443.

[11] Chang, C.C., Hwang, S.M., and Buehrer, A. (2002). A Shape Recognition Scheme Based on Relative Distances of Feature Points from the Centroid. *Pattern Recognition*, **24**, 1053 – 1063.

[12] Chauang, G., and Kuo, C. (1996). Wavelet Descriptor of Planar Curves: Theory and Applications. *IEEE Transactions on Image Processing*, **5**, 56 – 70.

[13] Churchill G.A., and Doerge R.W. (1994). Empirical Threshold Values for Quantitative Trait Mapping. *Genetics*, **138**, 963 – 971.

[14] Coen, E., Rolland-Lagan, A.G., Matthews, M., Bangharn, J.A., and Prusinkiewicz, P. (2004). The genetics of geometry. *Proc Natl Acad Sci USA*, **101**, 4728 – 4735.

[15] Cootes, T.F., Taylor, C.J., Cooper, D.H., and Graham, J. (1995). Active Shape Models–Their Training and Application. *Computer Vision and Image Understanding*, **61**, 38 – 59.

[16] Currence, T.M. (1934). Genes in the First Chromosome of Tomato as Related to Time of Fruit Ripening. *American Naturalist*, 68 – 73.

[17] Davies, R.H., Cootes, T.F., and Taylor, C.J. (2001). A Minimum Description Length Approach to Statistical Shape Modeling. *In Proceedings of Information Processing in Medical Imaging*, **199**, 50 – 63.

[18] Davies, R.H., Twining, C.J., Cootes, T.F., Waterton, J.C., and Taylor, C.J. (2002). A Minimum Description Length Approach to Statistical Shape Modeling. *IEEE Transactions on Medical Imaging*, **21**, 525 – 537.

[19] Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J Roy Stat Soc Ser B*, **39**, 1 – 38.

[20] Drake, A.G., and Klingenberg, C.P. (2008). The Pace of Morphological Change: Historical Transformation of Skull Shape in St. Bernard Dogs. *Proceedings of the Royal Society B: Biological Sciences*, **275**, 71 – 76.

[21] Drake, A.G., and Klingenberg, C.P. (2010). Large-Scale Diversification of Skull Shape in Domestic Dogs: Dsparity and Modularity. *American Naturalist*, **175**, 289 – 301.

[22] Dryden, I.L., and Mardia, K.V. (1998). Statistical Shape Analysis. *Wiley Chichester*.

[23] Dubois, S.R., and Glanz, F.H. (1986). An Autoregressive Model Approach to Two-Dimensional Shape Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **8**, 55 – 65.

[24] Enquist, B.J., Brown, J.H., and West, G.B. (1998). Allometric Scaling of Plant Energetics and Population Density. *Nature*, **395**, 163 – 165.

[25] Enquist, B.J., West, G.B., Charnov, E.L., and Brown, J.H. (1999). Allometric Scaling of Production and Life-History Variation in Vascular Plants. *Nature*, **4015**, 907 – 911.

[26] Epifanio, I., and Campos, N.V. (2011). Functional Data Analysis in Shape Analysis. *Computational Statistics and Data Analysis*, **55**, 2758 – 2773.

[27] Fletcher, S.D., and Geyer, C.J. (1999). The Genetic Analysis of Age-Dependent Traits: Modeling the Character Process. *Genetics*, **153**, 825 – 835.

[28] Fletcher, S.D., and Jaffrezic, F. (2002). Generalized Character Process Models: Estimating the Genetic Basis of Traits That Cannot Be Observed and That Change with Age or Environmental Conditions. *Biometrics*, **58**, 157 – 162.

[29] Flusser, J. (2000). On the Independence of Rotation Moment Invariants. *Pattern Recognition*, **33**, 1405 – 1410.

[30] Fray, A., Nesbitt, T.C., Grandillo, S., Knaap, E., Cong, B., Liu, J., Meller, J., Elber R., Alpert K.B., and Tansley, S.D. (2000). fw2.2 A Quantitative Trait Locus Key to the Evolution of Tomato Fruit Size. *Science*, **289**, 85 – 88.

[31] Fu, G., Berg, A., Das, K., Li, J., Li, R., and Wu, R. (2010). A Modeling Framework to Compute Which Genes Control Leaf Shape. *Theoretical Biology and Medical Modelling*,**7**, 1 – 28.

[32] Fulton, T.M., Beck-Bunn, T., Emmatty, D., Eshed, Y., Lopez, J., Petiard, V., Uhlig, J., Zamir, D., and Tanksley, S.D. (1997). QTL Analysis of an Advanced Backcross of Lycopersicon Peruvianum to the Cultivated Tomato and Comparisons with QTLs Found in Other Wild Species. *Theoretical and Applied Genetics*, **95**, 881 – 894.

[33] Garnier, E., and Laurent, G. (1994). Leaf anatomy, specific mass and water content in congeneric annual and perennial grass species. *New Phytol*, **128**, 725 – 736.

[34] Gilchrist, A.S., and Crisafulli, D. (2006). Using variation in wing shape to distinguish between wild and mass-reared individuals of Queensland fruit fly, Bactrocera tryoni. *Entom Exp App*, **119**, 175 – 178.

[35] Gower, J.C., and Dijksterhuis, G.B. (2004). Procrustes Problems. Oxford University Press, NY, 2004.

[36] Golland, P., Grimson, W.E.L., and Kikinis, R. (1999). Statistical Shape Analysis Using Fixed Topology Skeletons: Corpus Callosum Study, *International Conference on Information Processing in Medical Imaging*, **1613**, 382 – 387.

[37] Goshtasby, A. (1985). Description and Discrimination of Planar Shapes Using Shape Matrices. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **7**, 738 – 743.

[38] Gower, J.C., and Dijksterhuis, G.B. (2004). Procrustes Problems. *Oxford: University Press.*

[39] Grandillo, S., Ku, H.M., and Tanksley, S.D. (1996). Characterization of fs8.1, A Major QTL Influencing Fruit Shape in Tomato. *Molecular Breeding*, **2**, 251 – 260.

[40] Grandillo, S., Ku, H.M., and Tanksley, S.D. (1999). Identifying Loci Responsible for Naturalwh Variation in Fruit Size and Shape in Tomato. *Theoretical and Applied Genetics*, **99**, 978 – 987.

[41] Grenander, U. (1950). Stochastic Processes and Statistical Inference. *Arkiv for Matematik*, 195 – 277.

[42] Grenander, U., and Keenan, D.M. (1993). Towards Automated Image Understanding. *Journal of Applied Statistics*, **16**, 207 – 221.

[43] Hall, B.P., Muller, H.G., and Wang, J.L. (2006). Properties of Principal Component Methods for Functional and Longitudinal Data Analysis. *Annals of Statistics*, **34**, 1493 – 1517.

[44] Hamzeh, M., and Dayanandan, S. (2004). Phylogeny of Populus (Salicaceae) based on nucleotide sequences of chloroplast TRNT-TRNF region and nuclear rDNA. *Am J Bot*, **91**, 1398 – 1408.

[45] Hedrick, U.P., and Brooth, N.O. (1907). Mendelian Characters in Tomatoes. *American Society for Horticultural Science*, **5**, 19 – 24.

[46] Indritz, J. (1963). Methods in Analysis. *Macmillan, New York*.

[47] Jansen R.C., and Stam P. (1994). High Resolution Mapping of Quantitative Traits into Multiple Loci Via Interval Mapping. *Genetics*, **136**, 1447 – 1455.

[48] Jiang, C., Wright, R.J., Woo, S.S., Delmonte, T.A., and Paterson, A.H. (2000). QTL Analysis of Leaf Morphology in Tetraploid Gossypium (Cotton). *Theoretical and Applied Genetics*, **100**, 409 – 418.

[49] Karhunen, K. (1946). Zur Spektraltheorie Stochastischer Prozesse. *Annales Academiae Scientiarum Fennicae*, **37**.

[50] Kendall, D.G. (1984). Shape Manifolds, Procrustean Metrics, and Complex Projective Spaces. *Bulletin of the London Mathematical Society*, **16**, 81 – 121.

[51] Kessler, S., and Sinha, N. (2004). Shaping up: the genetic control of leaf shape. *Curr Opin Plant Biol*, **7**, 65 – 72.

[52] Khotanzad, A. (1990). Invariant Image Recognition by Zernike Moments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **12**, 489 – 497.

[53] Klingenberg, C.P., and Leamy, L.J. (2001). Quantitative Genetics of Geometric Shape in the Mouse Mandible. *Evolution*, **55**, 2342 – 2352.

[54] Klingenberg, C.P., Leamy, L.J., Routman E.J., and Cheverud, J.M. (2001). Genetic architecture of mandible shape in mice: effects of quantitative trait loci analyzed by geometric morphometrics. *Genetics*, **157**, 785 – 802.

[55] Klingenberg, C.P. (2003). Quantitative Genetics of Geometric Shape: Heritability and the Pitfalls of the Univariate Approach. *Evolution*, **57**, 191 – 195.

[56] Klingenberg, C.P., Leamy, L.J., and Cheverud, J.M. (2004). Integration and Modularity of Quantitative Trait Locus Effects on Geometric Shape in the Mouse Mandible. *Genetics*, **166**, 1909 – 1921.

[57] Klingenberg, C.P. (2010). Evolution and Development of Shape: Integrating Quantitative Approaches. *Nature Reviews Genetics*, **11**, 623 – 635.

[58] Klingenberg, C.P., Duttke, S., Whelan, S., and Kim, M. (2012). Developmental plasticity, morphological variation and evolvability: a multilevel analysis of morphometric integration in the shape of compound leaves. *J Evol Biol*, **25**, 115 – 129.

[59] Kong, X.D., Luo, Q.S., Zeng, G.H., and Lee, M.H. (2007). A New Shape Descriptor Based on Centroid-radii Model and Wavelet Transform. *Optics communications*, **273**, 362 – 366.

[60] Ku, H.M., Grandillo, S., and Tanksley, S.D. (2000). Fs8.1, A Major QTL, Sets the Pattern of Tomato Carpel Shape Well Before Anthesis. *Theoretical and Applied Genetics*, **101**, 873 – 878.

[61] Lander, E.S., and Botstein, D. (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, **121**, 185 – 199.

[62] Langlade, N.B., Feng, X.Z., Dransfield, T., Copsey, L., Hanna, A.I., Thebaud, C., Bangham, A., Hudson, A., and Coen, E. (2005). Evolution Through Genetically Controlled Allometry Space. *Proceedings of the National Academy of Sciences*, **102**, 10221 – 10226.

[63] Leamy, L.J., Klingenberg, C.P., Sherratt, E., Wolf, J.B., and Cheverud, J.M. (2008). A Search for Quantitative Trait Loci Exhibiting Imprinting Effects on Mouse Mandible Size and Shape. *Heredity*, **101**, 518 – 526.

[64] Liang, H., and Mahadevan, L. (2009). The shape of a long leaf. *Proceedings of the National Academy of Sciences*, **106**, 22049 – 22054.

[65] Lin, X., and Carroll, R. (2000). Nonparametric Function Estimation for Clustered Data when the Predictor is Measured Without/With Error. *Journal of American Statistical Association*, **95**, 520 – 534.

[66] Liu J., Van E.J., Cong, B., and Tanksley, S.D. (2001). A New Class of Regulatory Genes Underlying the Cause of Pear-Shaped Tomato Fruit. *Proceedings of the National Academy of Sciences*, **99**, 13302 – 13306.

[67] Lou, X., Casella, G., Little, R.C., Yang, M., and Wu, R. (2003). A Haplotype-Based Algorithm for Multilocus Linkage Disequilibrium Mapping of Quantitative Trait Loci with Epistasis. *Genetics*, **163**, 1533 – 1548.

[68] Lu, G., and Sajjanhar, A. (1999). Region Based Shape Representation and Similarity Measure Suitable for Content Based Image Retrieval. *Multimedia Systems*, **7**, 165 – 174.

[69] Lynch, M., and Walsh, B. (1998). Genetics and Analysis of Quantitative Traits. *Sinauer Associates, Sunderland, MA*.

[70] Ma, C., Casella, G., and Wu, R. (2002). Functional mapping of quantitative trait loci under-lying the character process: A theoretical framework. *Genetics*, **161**, 1751 – 1762.

[71] Mcneill, G., and Vijayakumar, S. (2006). A Probabilistic Approach to Robust Shape Matching. *In Proceedings of International Conference on Image Processing*, 937 – 940.

[72] Mezey, J.G., and Houle, D. (2005). The dimensionality of genetic variation for wing shape in Drosophila melanogaster. *Evolution*, **59**, 1027 – 1038.

[73] Mezey, J.G., Houle, D., and Nuzhdin, S.V. (2005). Naturally segregating quantitative trait loci affecting wing shape of Drosophila melanogaster. *Genetics*, **169**, 2101 – 2113.

[74] Monteiro, L.R., Diniz-Filho, J.A., Dos Reis, S.F., and Araujo, E.D. (2002). Geometric estimates of heritability in biological shape. *Evolution*, **56**, 563 – 572.

[75] Monteiro, L.R. (1999). Multivariate Regression Models and Geometric Morphometrics: the Search for Causal factors in the Analysis of Shape. *Systems Biology*, **48**, 192 – 199.

[76] Mukundan, R., Ong, S.H., and Lee, P.A. (2001). Image Analysis by Tchebichef Moments. *IEEE Transactions on Image Processing*, **10**, 1357 – 1364.

[77] Muller H.G. (2005). Functional Modelling and Classification of Longitudinal Data. *Board of the Foundation of the Scandinavian J. Statistics*, **32**, 223 – 240.

[78] Muller, H.G., Stadtmuller, U., and Yao, F. (2006). Functional Variance Process. *Journal of the American Statistical Association*, **101**, 1007 – 1018.

[79] Nath, U., Crawford, B.C.W., Carpenter, R., and Coen, E. (2003). Genetic control of surface curvature. *Science*, **299**, 1404 – 1407.

[80] Osher, S.J., and Sethian, J. A. (1988). Fronts Propagation with Curvature Dependent Speed: Algorithms Based on Hamilton Jacobi Formulations. *Journal of Computational Physics*, **79**, 12 – 49.

[81] Price H.C., and Drinkard A.W. (1908). Inheritance in Tomato Hybrids. *Virginia Agricultural Experiment Station*, **177**, 17 – 53.

[82] Quillevere, F., Debat, V., and Aurray, J.C. (2002) Ontogenetic and Evolutionary Patterns of Shape Dierentiation During the Initial Diversication of Paleocene Acarininids (Planktonic Foraminifera). *Paleobiology*, **28**, 435 – 448.

[83] Ramsay, J.O., and Silverman, B.W. (1997). Functional Data Analysis. *New York: Springer-Verlag.*

[84] Ramsay, J.O., and Silverman, B.W. (2002). Applied Functional Data Analysis. *New York: Springer-Verlag.*

[85] Rao, C.R. (1958). Some Statistical Methods for the Comparison of Growth Curves. *Biometrics*, **14**, 1 – 17.

[86] Reich, P.B. (2001). Body Size, Geometry, Longevity and Metabolism: Do Plant Leaves Behave Like a Animal Bodies? *Trends in Ecology and Evolution* **16**, 674 – 680.

[87] Renaud, S., Auffray, J.C., and De La Porte, S. (2010). Epigenetic Effects on the Mouse Mandible: Common Features and Discrepancies in Remodeling Due to Muscular Dystrophy and Response to Food Consistency. *BMC Evolutionary Biology*, **10**, 28.

[88] Rice, J., and Silverman, B. (1991). Estimating the Mean and Covariance Structure Nonparametrically when the Data are Curves. *Journal of the Royal Statistical Society*, **B 53**, 233 – 243.

[89] Ricklefs, R.E., and Miles, D.B. (1994). Ecological and evolutionary inferences from morphology: an ecological perspective. *Ecological morphology. University of Chicago Press, Chicago.* 13 – 14.

[90] Rohlf, F.J., and Marcus, L.F. (1993). A revolution in morphometrics. *Trends Ecol Evol*, **8**, 129 – 132.

[91] Rolland-Lagan, A.G., Bangham, J.A., and Coen, E. (2003). Growth Dynamics underlying petal shape and asymmetry. *Nature*, **422**, 161 – 163.

[92] Rolland-Lagan, A.G., Coen, E., Impey, S.J., and Bangham, J.A. (2005). A computational method for inferring growth parameters and shape changes during development based on clonal analysis. *J Theor Biol*, **232**, 157 – 177.

[93] Romdhani, S., and Vetter, T. (2005). Estimating 3 D shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. *IEEE Computer Soc Conf Computer Vision Pattern Recog*, **2**, 986 – 993.

[94] Romdhani, S., Ho, J., and Kriegman, D.J. (2006). Face recognition using 3-D models: Pose and illumination. *Proc IEEE*, **94**, 1977 – 1999.

[95] Samson, C., Feraud, L.B., Aubert, G., and Zerubia, J. (2000). A Level Set Model for Image Classification. *International Journal of Computer Vision*, **40**, 187 – 197.

[96] Schlichting, C.D., and Pigliucci, M. (1998). Phenotypic Evolution: A Norm Reaction Perspective Sinauer Associates. *Sunderland*.

[97] Slice, D.E. (2007). Geometric morphometrics. *Annu Rev Anthropol*, **36**, 261 – 281.

[98] Small, C.G. (1996). The Statistical Theory of Shape. *Springer-Verlag, New York*.

[99] Staib, L.H., and Duncan, J.S. (1992). Boundary Finding with Parametrically Deformable Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **11**, 1061 – 1075.

[100] Stegmann, M.B., and Gomez, D.D. (2002). A Brief Introduction to Statistical Shape Analysis. *Lecture Notes of Image Analysis and Computer Graphics Informatical Modeling, Technical University of Denmark*.

[101] Sun, J., and Wu, X. (2006). Chain Code Distribution-Based Image Retrieval. *Proceedings of International Conference on Intelligent Information Hiding and Multimedia*, **273**, 139 – 142.

[102] Super, B.J. (2004). Fast Correspondence-Based System for Shape Retrieval. *Pattern Recognition Letters*, **25**, 217 – 225.

[103] Tan, K.L., Ooi, B.C., and Thiang, L.F. (2000). Indexing Shapes in Image Databases Using the Centroid-Radii Model. *Data and Knowledge Enginerring*, **32**, 271 – 289.

[104] Tanksley S.D. (2004). The Genetic, Developmental, and Molecular Bases of Fruit Size and Shape Variation in Tomato. *Plant cell*, **16**, 181 – 189.

[105] Tasi, A., Yezzi, A., Wells, J.W., Tempany, C., Tucker, D., Fan, A., Grimson, W.E., and Willsky, A. (2003). A Shape Based Approach to the Segmentation of Medical Imagery Using Level Sets. *IEEE Transactions on Medical Imaging*, **22**, 137 – 154.

[106] Teague, M. (1980). Image Analysis Via the General Theory of Moments. *Journal of the Optical Society of America*, **70**, 920 – 930.

[107] Tsai A., Wells W., Warfield S., and Willsky A. (2005). An EM Algorithm for Shape Classification Based on Level Sets. *Medical Image Analysis*, **9**, 491 – 502.

[108] Tsukaya H. (2005). Leaf Shape: Genetic Controls and Environmental Factors. *International Journal of Developmental Biology*, **49**, 547 – 555.

[109] Vioarsdottir, U.S., Higgins, P., and Stringer, C. (2002) A Geometric Morphometric Study of Regional Differences in the Ontogeny of the Modern Human Facial Skeleton. *Genetics*, **201**, 211 – 229.

[110] Wang, Z., and Wu, R. (2004). A Statistical Model for High-Resolution Mapping of Quantitative Trait Loci Determining HIV-1 Dynamics. *Statistics in Medicine*, **23**, 3033 – 3051.

[111] Wang, Z., Liu, T., Lin, Z., Hegarty, J., Koltun, W.A., and Wu, R. (2010). A general model for multilocus epistatic interactions in case-control studies. *PLoS ONE*, **5(8)**, e11384.

[112] West, G.B., Brown, J.H., and Enquist, B.J. (1997). A General Model for the Origin of Allometric Scaling Laws in Biology. *Science*, **276**, 122 – 126.

[113] West, G.B., Brown, J.H., and Enquist, B.J. (1999a). The Fourth Dimension of Life: Fractal Geometry and Allometric Scaling of Organisms. *Science*, **284**, 1677 – 1679.

[114] West, G.B., Brown, J.H., and Enquist, B.J. (1999b). A General Model for the Structure and Allometry of Plant Vascular Systems. *Nature*, **400**, 664 – 667.

[115] West, G.B., and Brown, J.H. (2008). The Origin of Allometric Scaling Laws in Biology from Genomes to Ecosystems: Towards a Quantitative Unifying Theory of Biological Structure and Organization. *The Journal of Experimental Biology*, **208**, 1575 – 1592.

[116] Weber, K., Eisman, R., Morey, L., Patty, A., Sparks, J., Tausek, M., and Zeng, Z. (1999). An analysis of polygenes affecting wing shape on chromosome 3 in Drosophila melanogaster. *Genetics*, **153**, 773 – 786.

[117] Whitfield, J. (2001). All Creatures Great and Small. *Nature*, **413**, 342 – 344.

[118] Wu, R.L., Bradshaw, H.D., and Stettler, R.F. (1997). Molecular genetics of growth and development in Populus. V. Mapping quantitative trait loci affecting leaf variation. *Am J Bot*, **84**, 143 – 153.

[119] Wu, R., Ma, C., Little, R.C. and Casella, G. (2002a). A Statistical Model for the Genetic Origin of Allometric Scaling Laws in Biology. *Theoretical Biology*, **219**, 121 – 135.

[120] Wu, R., Ma, C., Lou, Y., and Casella, G. (2003). Molecular dissection of allometry, ontogeny and plasticity: A genomic view of developmental biology. *BioScience*, **53**, 1041 – 1047.

[121] Wu, R., Ma, C., Zhao, W., and Casella, G. (2003). Functional Mapping of Quantitative Trait Loci Underlying Growth Rates: A Parametric Model. *Physiological Genomics*, **14**, 241 – 249.

[122] Wu, R., Ma, C., Lin, M., and Casella, G. (2004a). A General Framework for Analyzing the Genetic Architecture of Developmental Characteristics. *Genetics*, **166**, 1541 – 1551.

[123] Wu, R., Ma, C., Lin, M., Wang, Z., and Casella, G. (2004b). Functional Mapping of Quantitative Trait Loci Underlying Growth Trajectories Using a Transform-Both-Sides Logistic Model. *Biometrics*, **60**, 729 – 738.

[124] Wu, R., Wang, Z., Zhao, W., and Cheverud, J.M. (2004c). A Mechanistic Model for Genetic Machinery of Ontogenetic Growth. *Genetics*, **168**, 2383 – 2394.

[125] Wu, R., and Lin, M. (2006). Functional Mapping–How to Map and Study the Genetic Architecture of Dynamic Complex Traits. *Nature Reviews Genetics*, **7**, 229 – 237.

[126] Wu, R., Ma, C., and Casella, G. (2007). Statistical Genetics of Quantitative Traits: Linkage, Maps and QTL. *Springer Science and Business Media, LLC.*

[127] Xiao, H., Jiang, N., Schaffner, E., Stockinger, E.J., and Van, K. (2008). A Retrotransposonmediated Gene Duplication Underlies Morphological Variation in Tomato Fruit. *Science*, **319**, 1527 – 1530.

[128] Xu, S., and Atchley, W. (1995). A Random Model Approach to Interval Mapping of Quantitative Trait Loci. *Genetics*, **136**, 1189 – 1197.

[129] Yao, F., Muller, H.G., Cifford, A.J., Dueker, S.R., Follett, J., Lin, Y., Buchholz, B.A., and Vogel, J.S. (2003). Shrinkage Estimation for Functioal Principal Component Scores, with Application to the Population Kinetics of Plasma Folate. *Biometrics*, **59**, 676 – 685.

[130] Yao, F., Muller, H.G., and Wang, H. (2005a). Functional Data Analysis for Sparse Longitudinal Data. *Journal of American Statistical Association*, **100**, 577 – 590.

[131] Yao, F., and Lee, T.C.M. (2006). Penalized Spline Models for Functional Principal Component Analysis. *Journal of Royal Statistical Society B*, **68**, 3 – 25.

[132] Yadav, R.B., Nishchal, N.K., Gupta, A.K., and Rastogi, V.K. (2007). Retrieval and Classification of Shape Based Objects Using Fourier, Generic Fourier and Wavelet Fourier Descriptors Technique: A Comparative Study. *Optics and Lasers in Engineering*, **45**, 695 – 708.

[133] Yi, N., Yandell, B.S., Churchill, G.A., Allison, D.B., Eisen, E.J., and Pomp, D. (2005). Bayesian model selection for genome-wide epistatic quantitative trait loci analysis. *Genetics*, **170**, 1333 – 1344.

[134] Yushkevich, P., Pizer, S.M., Joshi, S., and Marron, J.S. (2001). Intuitive, Localized Analysis of Shape Variability. *Intelligent Platform Management Interface*, **2082**, 402 – 408.

[135] Young, P.A., and MacArthur, J.W. (1947). Horticultural Characters of Tomatoes. *Texas Agricultural Experiment Station*, **698**, 3 – 61.

[136] Zeng, Z. (1994). Precision Mapping of Quantitative Trait Loci. *Genetics*, **1367**, 1457 – 1468.

[137] Zhao, W., and Wu, R. (2008). Wavelet-Based Nonparametric Functional Mapping of Longitudinal Curves. *Journal of the American Statistical Association*, **103**, 714 – 720.

[138] Zhang, D., and Lu, G. (2002). Shape Based Image Retrieval Using Generic Fourier Descriptor. *Signal Processing: Image Communication*, **17**, 825 – 848.

[139] Zhang, J., Zhang, X., Krim, H., and Walter, G. (2003). Object Representation and Recognition in Shape Spaces. *Pattern Recognition Letters*, **36**, 1143 – 1154.

[140] Zou, F., Fine, J.P., Hu, J., and Lin, D. (2004). An efficient resampling method for assessing genome-wide statistical significance in mapping quantitative trait loci. *Genetics*, **168**, 2307 – 2316.

[141] Zygier, S., Chaim, A.B., Efrati, A., Kaluzky, G., Borovsky, Y., and Paran, I. (2005). QTLs Mapping for Fruit Size and Shape in Chromosomes 2 and 4 in Pepper and A Comparison of the Pepper QTL Map with That of Tomato. *Theoretical and Applied Genetics*, **111**, 437 – 445.

Vita

Guifang Fu

I have completed my Ph.D. in Statistics from the Pennsylvania State University in 2012. Prior to Penn State, I obtained a master degree in mathematics from the Department of Mathematics at the University of Florida in 2008. I will join the Department of Mathematics & Statistics at Utah State University as an assistant professor with Tenure Track from fall 2012.

My research interests include Statistical Genetics, Functional Data Analysis or Longitudinal Data Analysis, High-dimensional Data Mining, Statistical Analysis of Images and Shapes, Computational and Mathematical Biology, Ordinary Differential Equations.