3-19-2004

The genetic characterization of populations comprising the Austronesian language family

Rachel Anne Chow
Florida International University

FLORIDA INTERNATIONAL UNIVERSITY

Miami, Florida

THE GENETIC CHARACTERIZATION OF POPULATIONS COMPRISING THE

AUSTRONESIAN LANGUAGE FAMILY

A thesis submitted in partial fulfillment of the

requirements for the degree of

MASTER OF SCIENCE

in

FORENSIC SCIENCE

by

Rachel Anne Chow

2004

To: Dean R. Bruce Dunlap
   College of Arts and Sciences

This thesis, written by Rachel Anne Chow, and entitled The Genetic Characterization of Populations Comprising the Austronesian Language Family, having been approved in respect to style and intellectual content, is referred to you for judgment.

We have read this thesis and recommend that it be approved.

George Duncan

Stephen Winkle

Rene J. Herrera, Major Professor

Date of Defense: March 19, 2004

The thesis of Rachel Anne Chow is approved.

Dean R. Bruce Dunlap
College of Arts and Sciences

Dean Douglas Wartzok
University Graduate School

Florida International University, 2004

DEDICATION

I dedicate this thesis to my family and friends, especially my parents. Mom and Dad without your support and encouragement during the past two years, this never would have been possible.

ACKNOWLEDGMENTS

ABSTRACT OF THE THESIS

THE GENETIC CHARACTERIZATION OF POPULATIONS COMPRISING THE

AUSTRONESIAN LANGUAGE FAMILY

by

Rachel Anne Chow

Florida International University, 2004

Miami, Florida

Professor Rene J. Herrera, Major Professor

Ascertaining the genetic relationships between Austronesian populations is pivotal to understanding their dispersal throughout the islands of the Pacific and Indian Oceans. The Austronesian expansion dates to approximately 6,000 years ago and from the linguistic and archeological evidence, the origin of this dispersal appears to be Taiwan. In this study, six polymorphic point mutation loci were studied in Taiwanese aborigines and compared with 32 other populations. The genetic relationships were characterized by maximum likelihood analysis, principal component maps, centroid gene flow plots, expected heterozygosities, power of discrimination values and pair wise G-tests. Following these analyses, it was apparent that genetic similarities existed between the Atayal and the Chinese, whereas the Ami displayed similarities with the Native Americans. Thus, the Atayal have little or no affinity for the Ami and other Austronesian populations. The large genetic differences between the two groups most likely arise from genetic isolation, and/or small population sizes.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Population Genetics and Austronesian History

Population genetics is integral to forensic DNA analysis and is also important in tracing the movements of peoples worldwide throughout time. This interdisciplinary study will focus on characterizing the genetic makeup of Austronesian groups, particularly those in Taiwan, Madagascar, Java and Bali. Genetics, particularly allelic and genotypic frequencies of a population can suggest the path of human dispersals of the past. Yet before population genetics, the clues to human dispersals existed mainly in the form of archaeological and linguistic data. Similarities between cultures and languages have been used to recreate the history of man's movement throughout the world. Linguistic data is essential, since divergence of language groups occurs in a manner similar to the speciation of isolated biological populations (Holden, 2002). Information of this type is important, and will be combined with the genetic character of a population to further elucidate the path of dispersal. Population geneticists hope that by tracing the history of a gene that they can trace the history of a population. Allelic frequencies generated from neutral loci have been the method used to compare populations since they generate the data used to estimate gene flow and population subdivision (Cann, 2001).

The colonization of the islands of the Pacific and Indian Oceans approximately 6000 years ago was the most recent dispersal of a post-agricultural population (Cann, 2001). There are many theories surrounding the peopling of these islands, and no single

1

theory enjoys full approval. Both archeological and linguistic data indicate that Austronesian speaking populations were responsible for this expansion. Currently, there are 1,200 Austronesian languages and 200 million native speakers spanning approximately two thirds (26,000 km) the circumference of the globe (Bellwood, 1991; Diamond, 2000). Austronesian languages are found as far west as Madagascar, off the coast of Mozambique in Africa, to the reaches of Easter Island, in the remote East Pacific, and include the languages of the islands of Southeast Asia, such as those of the Taiwanese aborigines (Kirch, 2000). The discontinuous pattern of island habitation led to social and linguistic isolation which resulted in a unique and autonomous cultural development. Thus, in turn, the art, linguistics, physical anthropology, and genetics of the present day aboriginal populations can be used to illuminate the Austronesian diaspora (Sewerin et al., 2002), a past which is extremely complicated to reconstruct, since no preserved Austronesian writings are dated earlier than 670 AD, near the end of the language family's expansion (Diamond, 2000). From the compiled archeological evidence, Jared Diamond synthesized "the Express Train to Polynesia" theory on the Austronesian expansion. Diamond theorized, that people spread from southern China, to Taiwan, the Philippines, Indonesia and finally to the coast of New Guinea and Polynesia (Diamond, 1988). Yet, there is still some debate over whether the actual homeland is in Taiwan or southern China. In either scenario, Taiwan is extremely important as a gateway to the Austronesian expansion.

Taiwan, also known as Formosa from colonial times, has been home to Austronesian tribal groups since before 4000 B.C. (Ruhlen, 1994). Furthermore, Taiwan

is the location of the greatest Austronesian linguistic variety (with 9 of the 10 subgroups present) and thus, according to linguistic theory, is believed to be near the origin the language family (Sewerin et al, 2002). Archeological evidence revealed that the Austronesian people practiced horticulture, domesticated animals, fished in both inshore and offshore waters and crafted red-slipped earthenware pottery. They were also expert canoe builders skilled in navigation, as evidenced by their spread throughout the islands (Kirch, 2000). The oldest archeological sites were found on Taiwan and were known as the Ta-p`en-k`eng culture, the earliest ceramic phase (Kirch 1997, 2000). The ceramic cultural tradition then diffused from Taiwan through the Philippines and into the equatorial islands of Southeast Asia. The phylogeny of Austronesian languages derived from previous linguistic studies parallels the pattern of island settlement supported by this archeological data, with Taiwan as the projected homeland (Gray and Jordan, 2000). The increasing use of genetic data is beginning to substantiate the more traditional lines of investigation and provide additional information on this major human dispersal.

There are nine tribal groups that constitute the aboriginal population of Taiwan: Ami, Atayal, Bunun, Paiwan, Puyama, Rukai, Saisiya, Tsou, and Yami (Chu, 1997). They comprise approximately 2% of Taiwan's total population, which currently numbers 21 million (Chu, 1997). The remainder of the population consists mainly of the Han Chinese, who have had a long history of conflict with the Austronesian aborigines and thus, were instrumental in the displacement of these groups since ancient times (Kao, 1958) to the rugged inland mountains and the eastern coastline. It is interesting to note that eight of the nine tribes have distinct languages along with different material cultures

and social organizations (Chai, 1967). The Atayal is Taiwan's second largest tribe after the Ami. Currently, there are approximately 90,000 Atayal tribal members. This Austronesian aboriginal group resides in a large area in northern Taiwan and is the only tribe known for facial tattooing, a practiced outlawed since the Japanese occupation.

Previous studies of these populations have provided a limited understanding of the genetic origins of each tribe and the genetic relationship between the tribes and other worldwide populations (Sewerin et al., 2002). Sewerin's investigation (2002), which focused on the genetic polymorphisms in low-density lipoprotein receptor (LDLR), glycophorin A (GYPA), hemoglobin G gammaglobin (HBGG), D7S8, group specific component (GC) and HLA-DQA1 loci of the Ami tribe, suggested that the Ami displayed a higher affinity to the Native American populations than to East Asian groups. The study also indicated that the Ami were genetically distinct from all other populations. The results of studies such as Sewerin's have generated a series of questions involving, the intertribal relationships and the relationship between the aborigines and other Austronesian speaking populations.

There are few genetics studies involving the aborigines of Formosa. Those that have been done, concern diseases that are prevalent with the population, but few actually attempt to determine the origin of these populations. Sewerin's study, as discussed above is a prime example of such a study. Previous studies examined the blood types and their frequencies among the tribes (Ikemoto et al., 1942; Huang, 1964; Nakajima et al., 1971). More recently, mitochondrial DNA (mtDNA) had been used to analyze these aborigines

of Formosa as well as the Austronesian influence in Melanesians and Polynesians (Melton et al., 1998; Merriwether et al, 1999; Lum et al., 2000). Other recent studies utilized polymorphisms such as those of the ABO blood type gene and examined the Austronesian influence in Melanesian populations (Ohashi et al., 2004). The makeup of this dispersal is still not fully characterized and the analysis is hindered by the difficulties that collecting samples from these populations present.

In the present study, we are interested in elucidating the relationship between the Atayal and the Ami, along with those of the Atayal and Ami to other Austronesian populations from Bali, Java and Madagascar. Two additional Austronesian groups (Chamorro and Guam Filipino) were also included (Budowle et al., 1998). The six polymorphic point mutation loci commonly known as LDLR, GYPA, HBGG, D7S8, GC and HLA-DQA1 were examined, to provide information on these relationships. The HLA-DQA1 and Polymarker (PM) loci have been utilized extensively in genetic studies and no detectable deviations from linkage or Hardy Weinberg equilibrium were reported for any of the six loci (Budowle et al., 1995). The HLA-DQA1 and polymarker loci have also been utilized extensively in forensic studies and the database of population frequencies is immense. These databases will be utilized for comparisons with the populations to be studied. HLA-DQA1 is a single point mutation locus, while the Polymarker loci consist of five point mutation loci: LDLR, GYPA, HBGG, D7S8, and GC. Genotypic and allelic frequencies from these loci were produced for the populations from Java, Bali, Madagascar and the Atayal tribe of Taiwan. Genotypic and allelic

frequency data from 29 populations were taken from the literature to determine the relationship between the Atayal and other worldwide populations.

Forensic History and Background

Genetic markers have been used by forensic laboratories for over 90 years to identify the origin of stains that have been taken from biological stains (Saferstein, 2002). The Polymarker/HLA-DQA1 (PM/DQA1) kit was originally released in 1990 with only the HLA-DQA1 locus. This kit was the first to use PCR in the forensic arena and utilized radiolabeled allele-specific oligonucleotide probes instead of the colorimetric system used today (Saferstein, 2002). The advent of PCR revolutionized forensic science, since it can be used to analyze samples containing extremely small amounts of DNA, along with degraded samples. The second kit, the PM/DQA1, was released in 1994 and quickly became widely used in forensic laboratories. The kit was extremely efficient and simple because all loci could be typed from one PCR reaction.

The PM/DQA1 kit is based on single nucleotide polymorphisms, used to create sequence specific probes. Polymorphisms are DNA variations that exist in the form of different alleles and make up only a small portion of the genome, since 99.7% of DNA is the same between people (Butler, 2001). There are two types of polymorphisms, point mutations (sequenced based) or length mutations (Saferstein, 2002). The point mutations are utilized in the PM/DQA1 kit, while as a side note, length based mutations are exploited in Short Tandem Repeat (STR) typing, the golden standard of DNA typing today. Sequences for the most common alleles are used to create probes that are

6

immobilized on nylon strips. The PCR products (amplicon) are 239-240bp in length and are hybridized to the probes (Saferstein, 2002). Only single stranded DNA will hybridize, so the DNA is heated to denature it, also, the sequence must be complementary to the probe, and a difference of even one base can cause the amplicon not to hybridize (Rudin and Inman, 2002). The results of this study are suitable for forensic databasing, but the PM/DQA1 databases are becoming obsolete following the discontinuation of the kit in 2003. That point aside, the information is still extremely important since these populations are so poorly characterized, due to the difficulty of collection.

CHAPTER II

METHODOLOGY

Populations Studied

Forty unrelated individuals from the Atayal tribe of Taiwan were collected. Also studied were 24 individuals from Java, 34 individuals from Bali and 69 from Madagascar. Java and Bali are both islands within the borders of Indonesia, whereas Madagascar is located approximately 250 miles off the south-east coast of Africa and approximately 6,000 miles from the island of Taiwan, yet the individuals from Madagascar are known to speak Malagasy, an Austronesian language. Figure 1 illustrates the geographical locations of the populations studied. These populations were compared with the data from 29 other worldwide populations found within the literature. Individuals were identified as Atayal members by tracing back biographical information for at least two generations. Table 1 lists the populations and their location along with their reference.

Data Collection Technique

Blood Collection and DNA Purification

The samples were collected as whole blood in EDTA Vacutainer tubes in strict adherence to the guidelines set forth by Florida International University's Institutional Review Board. Cells were lysed and leukocyte nuclei were isolated by centrifugation, followed by digestion of the cells with proteinase K, as previously described (Antunez de Mayolo et al., 2002). The total genomic DNA was then isolated by a standard phenol

FIGURE 1. Geographical Map of Populations

1 Ami
2 Atayal
3 Bunun
4 Paiwan
5 Puyama
6 Rukai
7 Saisiat
8 Tsou
9 Yami
10 Non-Aborigine

Madagascar

Java

Bali

9
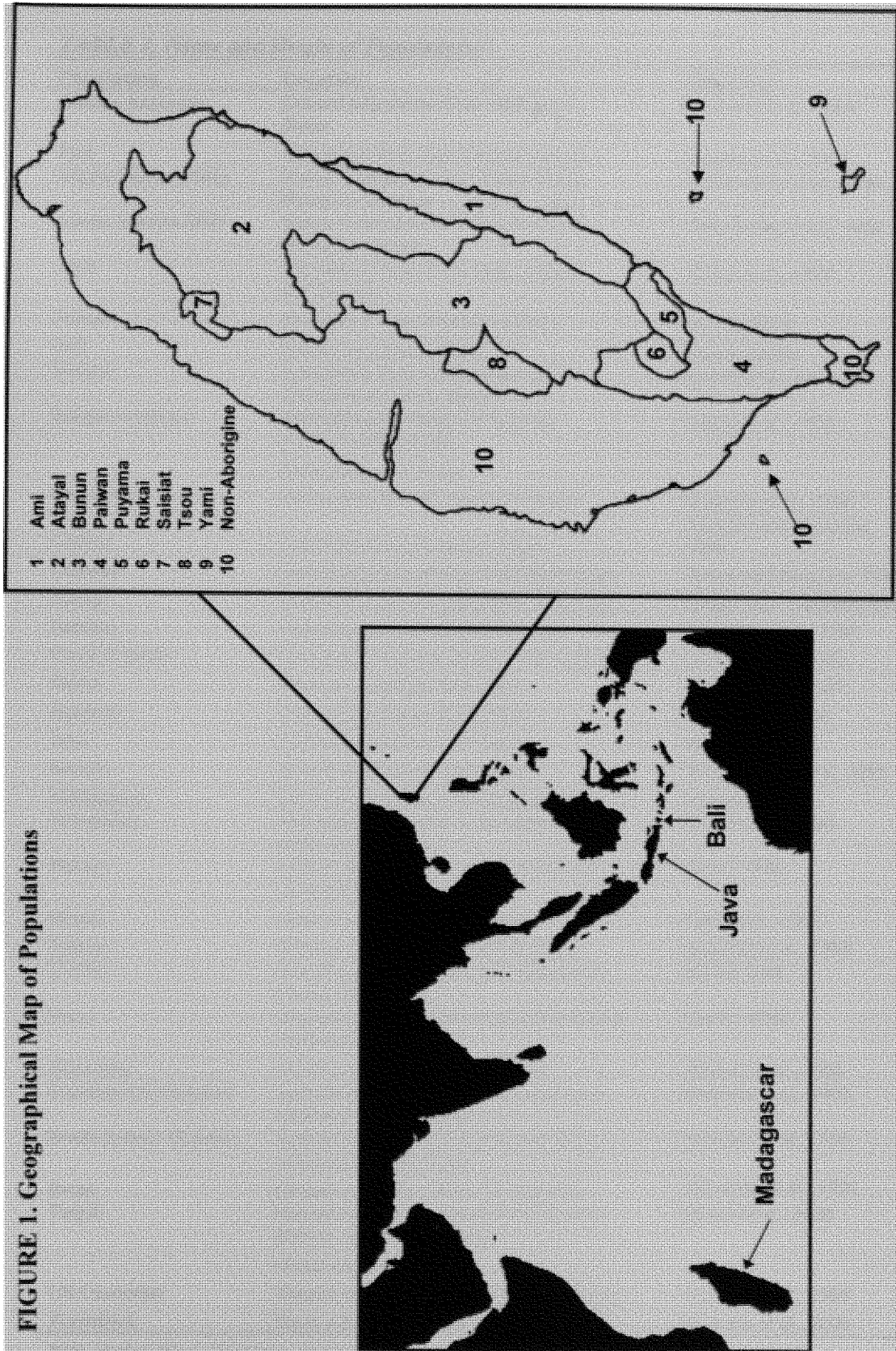
TABLE 1. Name and Origin of Populations*

Population	Location	Reference
African American	General population from the United States	Budowle et al., 1995
Ami	Aborigines from East Central Taiwan	Sewerin et al., 2002
Alaska Eskimo (BH)	Native Americans from Bethel-Wade Hampton, Alaska	Walkinshaw et al., 1996
Alaska Eskimo (NS)	Native Americans from North Slope Borough, Alaska	Walkinshaw et al., 1996
Arab	Populations from Gaza Strip, Judea, and Samaria	Hayes et al., 1995
Atayal	Aborigines from Northern Taiwan	Present Study
Bali	General population from Bali	Present Study
Basque	Individuals from the Goiherri Valley, Gipuzkoa province, Spain	Brown et al., 2000
Basque residents	Heterogeneous sample of Spanish people residing in the Basque Country, Spain	Garcia et al., 1994
Chamorro	Aborigines from the Island of Guam	Budowle et al., 1998
Chilean	Population from Santiago, Chile	Jorquera and Budowle, 1998
Chinese	General population from mainland China	Huang and Budowle, 1995
Elazig	Individuals from Elazig region of Turkey	Dulger and Tokdemir, 2000
German	General population from Germany	Schneider et al., 1991
Greek Cypriot	Individuals of Greek decent from Cyprus	Cariolou et al, 1998
Italian	General population from Italy	Spinella et al., 1997
Japanese	General population from Japan	Fildes et al., 2000
Java	General population from Java	Present Study
Korean	General population from Korea	Woo and Budowle, 1995
Madagascar	General population from Madagascar	Present Study
Mongolian	Individuals from Ulan Ude region of northwest Mongolia	Brown et al., 2000
Navajo	Nadene-speaking tribe from Southwest United States	Scholl, 1996
Omani	General population from Oman	Tahir et al., 1997
Filipino	Filipino population residing in Guam	Budowle et al., 1998
Pueblo	Amerind-speaking tribe from Southwest United States	Scholl, 1996
Sioux	Amerind-speaking tribe from north-central United States	Scholl, 1996
Slovenian	General population from Slovenia	Drobnic et al., 2000
Southeastern Hispanic	Spanish-speaking individuals residing in southeast United States	Budowle et al., 1995
Southwestern Hispanic	Spanish-speaking individuals residing in southwest United States	Budowle et al., 1995
Swiss	General population from Switzerland	Schneider et al., 1991
Tadjik	Residents from the regions of Dushanbe, Kulyab, Kurgan-Tube and Hudzhent in Tajikistan	Brown et al., 2000
US Caucasian	General population from the United States	Budowle et al., 1995
Zimbabwe	Black African population from Zimbabwe	Wolfarth et al., 2000

*BH, Bethel-Wade Hampton, Alaska; NS, North Slope Borough, Alaska

10

chloroform extraction and ethanol precipitation (Luis et al., 2003). All samples were kept in -80^{0}C storage when not in use.

DNA Amplification and Polymarker-DQA1 Genotyping

The extracted DNA was amplified by polymerase chain reaction (PCR) using the AmpliType® PM-DQA1 PCR Amplification and Typing Kit (Perkin Elmer Corp, Norwalk, CT) using the conditions specified by the manufacturer. PCR was performed using the Perkin-Elmer 480 thermal cycler. Following amplification, samples were screened for successful amplification by electrophoresis in a 1X TAE 2% agarose gel, followed by ethidium bromide staining and UV visualization. Figure 2 is an example of a successful amplification. Within Figure 2, lanes 1, 2, 3 and 5 in the gel are samples that have been successfully amplified. Lane 6 is the positive control that was used and Lane 4 is the molecular weight marker (HAE III digested with phi-X174. A negative control was also utilized, containing only the reaction mix with no DNA to ensure that the reactants were not responsible for the positive results. The following loci were then genotyped for each sample, HLA-DQA1, LDLR, GYPA, HBGG, D7S8, and GC according to the directions from the manufacturer. The chromosomal locations of the loci are: HLA-DQA1, 6p21.3; LDLR, 19p13.1-13.3; GYPA, 4q28-31; HBGG, 11p15.5; D7S8, 7q22-31.1; and GC, 4q11-13 (Budowle et al., 1995). Typing of these samples involves reverse dot blot methods with allele specific oligonucleotide probes bound to strips that allow the typing of multiple loci at one time. The samples are visualized colorimetrically; a blue color will develop through an enzymatic reaction. When the DNA is amplified, a biological tag, biotin, attached to the primer is used (Rudin and

11

Inman, 2002). The amplicon is now tagged with biotin, and following DNA hybridization, the excess DNA is washed, so only those amplicons that hybridize remain. Streptavidin protein and the enzyme horseradish peroxidase (HRP) are chemically linked and added, the streptavidin, which has a strong affinity for biotin, binds tightly to it (Rudin and Inman, 2002). Tetramethylbenzidine (TMB) is added and HRP produces a blue color in the presence of hydrogen peroxide, simply, the hybridized DNA will turn blue. Figure 3 is a depiction of PM-DQA1 strips. All alleles can be typed from a single PCR reaction.

FIGURE 2. Successful Amplification of PM and HLA-DQA1 Loci
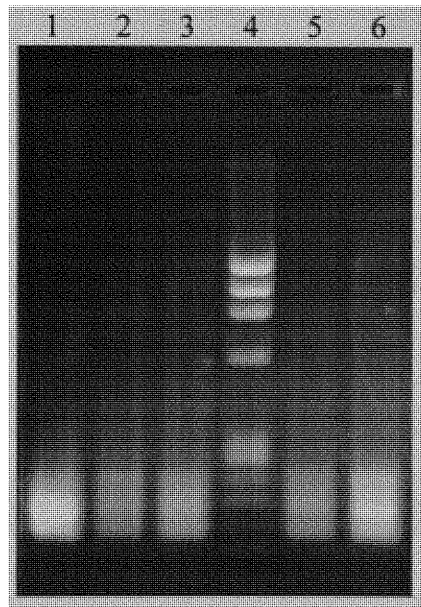
FIGURE 3. Polymarker and DQA1 Typing Strips
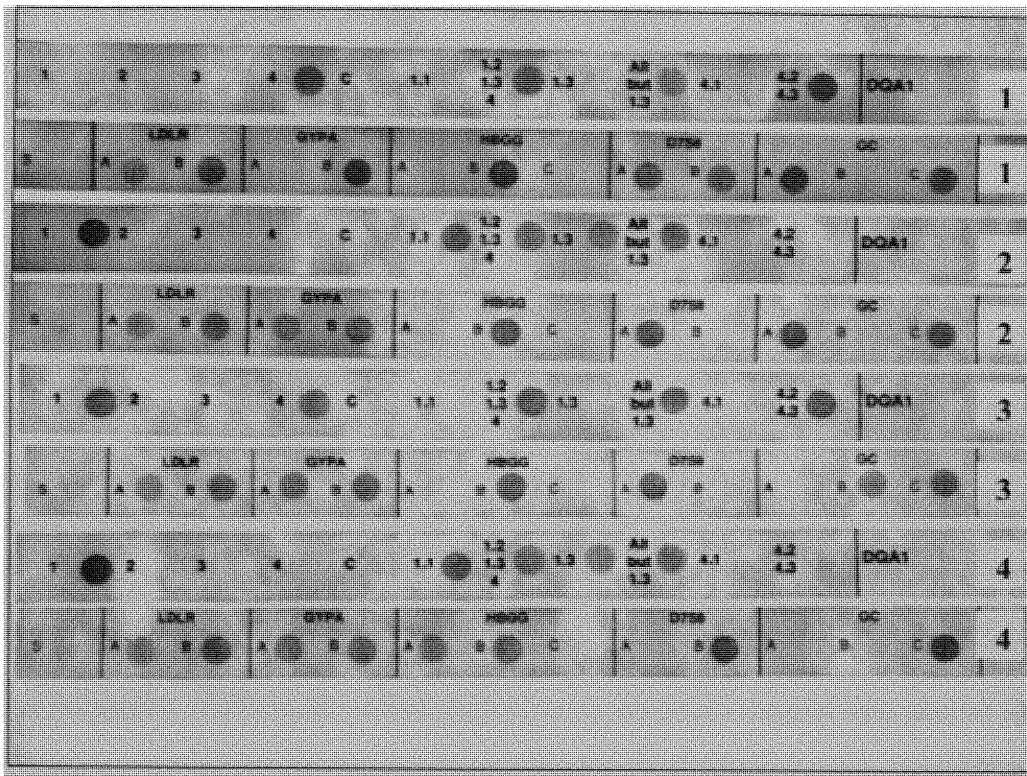
The strips in Figure 3 represent four individuals from the population of Java. The typing is achieved visually, by comparing the intensity of the dots that appear. To type the DQA1 strip, the C dot (read to the right of a designation) is used as a control. Any alleles that have more intense color than the C dot are considered positive. So for example, the type of individual 1 is 4.2/4.3, 4.2/4.3 or in other words, they are homozygous for the 4.2/4.3 allele. The PM strip is typed in a similar manner, with the S dot as a control instead of the C dot. The types for the strips in Figure 3 for all individuals are displayed below in Table 2.

13

Table 2. Genotypes of Individuals from Figure 3.

Individual	LDLR	GYPA	HBGG	D7S8	GC	DQA1
1	AB	BB	BB	AB	AC	4.2/4.3, 4.2/4.3
2	AB	AB	BB	AA	AC	1.1, 1.3
3	AB	AB	BB	AA	BC	1.2, 4.2/4.3
4	AB	AB	AB	BB	CC	1.1, 1.3

Statistical Analyses

For all loci, genotypic and allelic frequencies were calculated using the gene counting method (Li, 1976). To ascertain the phylogenetic relationships between populations, Maximum likelihood trees (ML) based on the allelic frequency distributions of the loci were generated by the PHYLIP 3.52c program (Felsenstein, 1993). Bootstrap consensus phylogenies (1000 replications) were produced by SEQBOOT and CONTML options programs of PHYLIP. The bootstrap is a statistical method that can be used to place confidence levels on phylogenies (Felsenstein, 1985). The CONTML and CONSENSE programs determined the best-fit tree using the allele frequencies.

Principal component (PC) analyses were then performed using the Numerical Taxonomy and Multivariate Analysis System (NTSys) Principal Component program (Rohlf, 2002), to summarize the information concerning relationships among the populations. Centroid analysis was conducted to examine the relative gene flow experienced by a population and can be used to compare relative gene flow among several populations along with the effective population size (Sewerin et al, 2002). The centroid analysis ascertains the level of gene flow by following the procedures of

Harpending and Ward (1982), and assumes an island model of population structure and expects a single linear relationship between the heterozygosity of a population and the genetic distance of the population from the centroid. The centroid is defined as the overall mean allelic frequency of the populations. The theory is particularly useful in detecting and analyzing outliers. The theory assumes that every population is exchanging migrants at the same rate, both among other populations in the sample as well as, with outside populations. If a population is receiving gene flow from elsewhere at a higher-than-average rate, then the heterozygosity will be higher than predicted. Those populations would plot above the linear regression line. Conversely, if the population has remained isolated, the heterozygosity will be lower than expected due to less-than-average gene flow and would segregate below the linear regression line (Sewerin et al., 2002). The BIOSYS II program was used to create expected heterozygosities and to detect any disagreement from Hardy-Weinberg equilibrium, using the Fisher exact test and chi-square deviation test. Power of discrimination values were also determined for all six loci independently and cumulatively. The power of discrimination is the chance that individuals will have a different set of combined genotypes. This serves as the basis for forensic DNA typing. G-tests were completed, with each 2 x 2 contingency table determining the statistical significance, indicating whether the populations are homogeneous with each other (Lewontin and Felsenstein, 1965).

CHAPTER III

RESULTS

Genetic composition of Atayal, Java, Bali and Madagascar

Table 2 displays the allelic and genotypic frequencies for the LDLR, GYPA, HBGG, D7S8 and GC loci in all four populations. Allelic and genotypic frequencies for the HLA-DQA1 locus are shown in Table 3. The expected and observed heterozygosities for all six loci in the four populations are presented in Table 4. The observed heterozygosities ranged from 10% (HBGG) to 83% (HLA-DQA1), both exhibited within the Atayal population. Except for two instances, all loci were in Hardy-Weinberg equilibrium (as seen in Table 5). One of the exceptions was GYPA in the population from Java, which following a Bonferroni correction (0.008), was also in Hardy-Weinberg equilibrium. The Bonferroni correction is used to reduce the number of falsely significant results in a statistical analysis. The second exception was for the Fisher test of all populations in the GYPA locus. At 0.007, it was slightly below the Bonferroni correction (0.008). Power of discrimination (PD) values are presented in Table 6. The largest range in PD existed in the Atayal population, with 0.176 in the HBGG locus and 0.913 in the HLA-DQA1 locus. Overall the combined PM-DQA1 power of discrimination was between 0.998 and 0.999 for all four populations. The six loci G-test proved that all differences were statistically significant for all 32 populations that were compared with the Atayal (P < 0.0001) as can be seen in Table 7.

TABLE 3. Atayal Sample Genotypes

Sample ID	LDLR	GYPA	HBGG	D7S8	GC	DQA1
5654	AB	BB	BB	AB	AB	1.1, 3
5663	BB	AB	BB	BB	BB	1.3, 4.1
5061	BB	AB	BB	BB	BB	1.1, 4.2/4.3
5071	BB	AB	BB	AB	AB	1.3, 4.1
5081	BB	AB	BB	AA	AC	1.3, 1.3
5094	BB	AA	BB	AA	AC	3, 4.1
5064	BB	BB	BB	AA	BB	1.1, 1.3
5095	BB	AB	BB	AB	BB	1.1, 4.1
5069	BB	BB	BB	AB	AB	1.1, 3
5072	BB	AB	AB	AA	AB	1.2, 1.3
5091	BB	AB	AB	AB	AB	1.1, 3
5092	AB	BB	BB	AA	BB	1.1, 3
5098	BB	AA	BB	AB	AB	3, 3
5104	BB	AB	BB	BB	BB	1.2, 3
5550	AB	AA	BB	AA	CC	3, 3
5557	BB	AA	BB	AB	BB	3, 4.1
5565	BB	AA	BB	AB	AB	1.3, 4.1
5571	BB	AB	BB	BB	BB	1.1, 4.2/4.3
5576	BB	AA	BB	AA	BC	1.3, 1.2
5582	BB	AA	BB	AA	BC	1.3, 4.2/4.3
5583	AB	AB	BB	AA	CC	1.2, 4.1
5649	BB	AB	BB	AB	BB	1.2, 4.1
5585	BB	AA	BB	AA	AB	4.1, 4.2/4.3
5648	BB	AA	BB	AA	BC	1.1, 1.1
5586	BB	AA	BB	AB	AB	1.1, 1.3
5597	AB	AA	AB	AB	AB	1.3, 3
5611	BB	BB	AB	AA	AC	1.1, 4.1
5619	BB	AA	BB	BB	BC	1.3, 4.1
5642	BB	AA	BB	AB	CC	1.1, 1.1
5621	BB	BB	BB	AB	BB	1.1, 3
5623	BB	AB	BB	AB	AB	1.1, 1.2
5645	BB	AB	BB	AB	AB	3, 4.2/4.3
5634	BB	AA	BB	BB	AB	1.1, 1.2
5639	BB	AB	BBG	AA	BC	1.1, 4.1
5641	AB	AA	BB	AB	BB	1.1, 4.2/4.3
5103	BB	AA	BB	AB	AB	1.1, 1.2
5096	BB	AB	BB	AA	BB	1.1, 1.1
5609	BB	AA	BB	AA	AA	1.2, 3
5655	BB	AA	BB	AA	BB	1.1, 4.1
5659	AB	BB	BB	AB	AB	1.1, 3

TABLE 4. Java Sample Genotypes

Sample ID	LDLR	GYPA	HBGG	D7S8	GC	DQA1
75	AB	BB	BB	AB	AC	4.2/4.3,4.2/4.3
76	BB	AB	BB	AB	BC	1.3, 4.2/4.3
77	AA	AB	BB	BB	AC	4.2/4.3,4.2/4.3
78	BB	AB	AB	AB	BC	1.2, 4.2/4.3
79	BB	BB	AB	AA	BB	1.1, 4.2/4.3
80	BB	AB	AB	AB	BB	1.1, 1.2
81	AB	AB	BB	AA	AC	1.1, 1.3
82	AB	AB	BB	AA	AC	1.2, 4.2/4.3
83	AB	AB	AB	BB	CC	1.1, 1.3
84	BB	AB	AB	AA	CC	1.2, 3
85	AB	AA	BB	AB	AC	1.1, 4.2/4.3
86	BB	AB	BB	AB	CC	2, 4.2/4.3
87	AB	AA	BB	AB	AB	1.1, 4.2/4.3
88	AB	AB	BB	BB	CC	1.2, 4.2/4.3
89	AB	AB	BB	BB	AB	4.1, 4.2/4.3
90	BB	AA	AB	BB	AB	4.2/4.3,4.2/4.3
91	BB	AB	AA	BB	AB	4.2/4.3,4.2/4.3
92	BB	AB	BB	BB	BB	4.2/4.3,4.2/4.3
93	BB	AB	AB	AB	AC	1.1, 3
94	AB	AB	BB	AA	AC	2, 3
95	AA	AA	AB	AB	BC	4.2/4.3,4.2/4.3
96	AB	AB	BB	AB	BC	1.1, 4.2/4.3
97	AB	AA	BB	AB	AB	1.3, 4.2/4.3
98	BB	AB	BB	AB	CC	3, 4.2/4.3

TABLE 5. Bali Sample Genotypes

Sample ID	LDLR	GYPA	HBGG	D7S8	GC	DQA1
1	AB	AB	BB	AB	CC	1.1, 4.1
2	AB	AB	AB	AB	BB	1.1, 4.2/4.3
3	AB	AA	BB	AB	BB	1.2, 4.2/4.3
4	BB	AA	AB	AB	BC	1.1, 1.1
5	AA	AB	AB	AB	BC	1.1, 1.2
6	AB	AA	AB	BB	BB	1.2, 1.3
7	AB	BB	AB	AB	BB	1.1, 4.2/4.3
8	AA	AB	BB	AA	BC	1.1, 4.2/4.3
9	BB	AA	BB	AB	BB	1.2, 4.2/4.3
10	AB	AA	AB	AB	BC	1.1, 2
11	BB	BB	BB	AB	BB	1.1, 4.2/4.3
12	AA	AB	AB	AB	CC	1.2, 3
13	BB	AA	BB	BB	BC	4.2/4.3, 4.2/4.3
14	BB	AA	AB	BB	BC	1.1, 4.1
15	AB	AB	BB	AB	BC	2, 4.2/4.3
16	BB	AB	AB	AB	BB	1.2, 4.2/4.3
17	BB	AA	BB	AB	BB	1.1, 1.1
18	AB	AB	BB	AA	BC	1.1, 1.2
19	AB	AB	BB	AB	BC	1.1, 1.2
20	BB	AB	BB	AA	BC	1.1, 1.1
21	BB	AA	BB	AA	BC	1.1, 1.1
22	AA	AB	BB	AB	BB	1.2, 4.2/4.3
23	BB	BB	BB	AB	BB	1.2, 1.2
24	AB	AB	BB	BB	BB	1.2, 4.1
25	AB	AA	BB	AB	CC	1.1, 4.2/4.3
26	BB	AA	BB	AB	BB	1.1, 4.2/4.3
27	BB	AA	AB	AA	CC	1.1, 1.1
28	AB	AB	BB	AA	BC	1.1, 1.2
29	BB	AB	AB	AA	BC	1.1, 4.2/4.3
30	BB	AB	BB	AA	BB	1.2, 4.2/4.3
31	BB	AB	BB	BB	BC	1.1, 1.2
32	BB	AA	BB	BB	CC	1.1, 1.2
33	AB	AA	BB	AB	BB	1.1, 1.2
34	BB	AA	AA	BB	AC	4.2/4.3, 4.2/4.3

TABLE 6. Madagascar Sample Genotypes

Sample ID	LDLR	GYPA	HBGG	D7S8	GC	DQA1
1	AB	AB	AA	AA	AC	1.2, 3
2	BB	AA	BB	AB	BC	1.1, 4.2/4.3
4	AB	AA	AB	AA	AB	3, 4.2/4.3
5	AB	AB	AA	AB	BB	4.1, 4.2/4.3
6	BB	BB	CC	AA	BB	4.1, 4.2/4.3
9	BB	BB	AC	AB	BB	1.2, 1.3
10	AB	AB	AA	AB	BC	4.1, 4.2/4.3
11	BB	AA	BB	AB	AB	1.2 4.2/4.3
12	BB	AB	CC	AB	BC	3, 4.1
15	BB	BB	BB	AB	BB	1.2, 1.2
18	AB	AA	BC	AB	BB	1.1, 4.1
19	BB	AA	AC	AB	BB	1.1, 4.2/4.3
22	AB	BB	AB	AB	AC	1.1, 4.1
25	AB	AA	BC	AA	AB	1.1, 4.1
27	BB	AA	BC	BB	AB	4.1, 4.2/4.3
30	AB	AB	BB	AB	BB	1.1, 4.1
31	AB	AA	AC	AB	AB	4.2/4.3, 4.2/4.3
32	BB	AA	AA	AB	BB	1.1, 4.2/4.3
33	AB	BB	BC	AA	BB	1.3, 4.2/4.3
36	AB	AA	CC	AA	BB	1.1, 4.2/4.3
37	AB	BB	CC	AB	BC	1.3, 4.1
38	BB	BB	BC	AB	BB	1.2, 4.2/4.3
40	BB	AB	CC	AB	AB	1.3, 4.1
41	BB	BB	CC	AA	AB	1.3, 2
42	BB	BB	CC	BB	BB	1.2, 1.3
43	BB	BB	AA	AA	BB	1.1, 1.2
44	AB	AB	BC	AB	BC	2, 2
45	AB	AA	BC	BB	BC	4.2/4.3, 4.2/4.3
46	BB	AB	AA	AA	AB	4.1, 4.1
47	AB	AA	BC	AB	BB	1.1, 4.1
48	AB	AA	CC	AA	BB	1.2, 4.2/4.3
49	BB	BB	CC	AB	AB	1.2, 4.2/4.3
50	AB	BB	BB	AB	BB	4.2/4.3, 4.2/4.3
51	BB	AB	BB	AB	AB	1.1, 3
53	BB	BB	BC	AB	BB	1.3, 1.2
55	BB	AA	BC	AA	BB	1.2, 2
57	BB	AB	AA	BB	AB	3, 4.2/4.3
58	AB	AA	AB	AB	BC	1.2, 3
59	AB	BB	BC	BB	BB	4.2/4.3, 4.2/4.3
60	BB	AB	AC	AA	BB	3, 4.2/4.3
61	AB	AB	AB	AB	BC	3, 4.1
62	BB	AB	AB	AA	AB	4.2/4.3, 4.2/4.3
64	BB	AA	AA	AB	BB	1.1, 1.2
65	BB	AA	BB	AA	CC	4.1, 4.1
69	AB	AA	BC	AA	BC	2, 4.2/4.3
70	AB	BB	AC	BB	CC	1.1, 4.1

20

Table 6. Continued

71	AB	AA	AC	AB	BB	1.1, 3
72	BB	AA	AA	BB	BB	1.1, 4.2/4.3
73	AB	AA	AB	AB	AB	1.2, 4.2/4.3
74	BB	AB	AC	AA	AB	1.3, 4.2/4.3
75	AB	AA	AA	AA	BB	1.2, 4.2/4.3
76	BB	AA	AC	AB	BC	1.3, 4.1
77	AB	AB	AB	AB	BB	2, 4.2/4.3
78	AB	AA	AC	AA	BB	1.2, 1.2
79	BB	AA	BB	AA	AB	1.2, 3
83	BB	AB	AB	BB	AB	1.2, 3
85	BB	AB	AB	AA	BB	3, 4.1
86	AB	AB	CC	AB	AB	1.1, 1.2
87	BB	AA	BC	BB	BB	4.1, 4.2/4.3
88	BB	BB	BC	AA	BB	1.1, 1.2
89	AA	AB	AA	AA	AB	4.1, 4.2/4.3
90	BB	AA	BB	AA	BB	1.1, 1.2
91	BB	AA	AB	AA	BB	1.2, 4.2/4.3
92	BB	AA	AB	AB	BB	1.2, 3
93	AB	AB	BB	AB	BB	4.1, 4.1
94	BB	AB	AC	AB	BB	1.1, 3
96	BB	AB	AC	AA	BB	4.1, 4.1
97	BB	AB	AC	AA	BB	1.1, 4.2/4.3
99	BB	AB	AC	BB	BB	1.1, 4.1

TABLE 7. Genotypic and allelic frequencies for Polymarker loci

Locus	LDLR	GYPA	HBGG	D7S8	GC
ATAYAL (N = 40)*					
Genotype					
AA	0.000	0.450	0.000	0.400	0.025
AB	0.175	0.375	0.100	0.450	0.375
BB	0.825	0.175	0.900	0.150	0.325
BC	NA	NA	0.000	NA	0.125
CC	NA	NA	0.000	NA	0.075
AC	NA	NA	0.000	NA	0.075
Allele					
A	0.087	0.637	0.050	0.625	0.250
B	0.913	0.363	0.950	0.375	0.575
C	NA	NA	0.000	NA	0.175

Table 7. Continued

JAVA (N = 24)*

Genotype					
AA	0.083	0.208	0.042	0.208	0.000
AB	0.458	0.708	0.333	0.500	0.208
BB	0.459	0.084	0.625	0.292	0.125
BC	NA	NA	0.000	NA	0.167
CC	NA	NA	0.000	NA	0.208
AC	NA	NA	0.000	NA	0.292
Allele					
A	0.313	0.563	0.208	0.458	0.250
B	0.688	0.438	0.792	0.542	0.313
C	NA	NA	0.000	NA	0.438

BALI (N = 34)*

Genotype					
AA	0.147	0.441	0.029	0.235	0.000
AB	0.353	0.471	0.324	0.559	0.000
BB	0.500	0.088	0.647	0.206	0.412
BC	NA	NA	0.000	NA	0.412
CC	NA	NA	0.000	NA	0.147
AC	NA	NA	0.000	NA	0.029
Allele					
A	0.309	0.676	0.191	0.515	0.015
B	0.691	0.324	0.809	0.485	0.618
C	NA	NA	0.000	NA	0.376

MADAGASCAR (N = 69)*

Genotype					
AA	0.015	0.420	0.159	0.377	0.000
AB	0.420	0.348	0.159	0.478	0.261
BB	0.565	0.232	0.188	0.145	0.536
BC	NA	NA	0.145	NA	0.145
CC	NA	NA	0.204	NA	0.029
AC	NA	NA	0.145	NA	0.029
Allele					
A	0.225	0.594	0.333	0.616	0.145
B	0.775	0.406	0.326	0.384	0.739
C	NA	NA	0.341	NA	0.116

*NA, not applicable because allele or genotypes do not exist at specific loci;
N is the number of individuals studied.

TABLE 8. HLA-DQA1 genotypic and allelic frequencies*

ATAYAL (N = 40)

Genotype	Frequency	Allele	Frequency
1.1, 1.1	0.075	1.1	0.312
1.1, 1.2	0.075	1.2	0.112
1.1, 1.3	0.050	1.3	0.163
1.1, 3	0.150	3	0.175
1.1, 4.1	0.100	4.1	0.163
1.1, 4.2/4.3	0.075	4.2/4.3	0.075
1.2, 1.3	0.050		
1.2, 3	0.050		
1.2, 4.1	0.050		
1.3, 1.3	0.025		
1.3, 3	0.025		
1.3, 4.1	0.100		
1.3, 4.2/4.3	0.025		
3, 3	0.050		
3, 4.1	0.050		
3, 4.2/4.3	0.025		
4.1, 4.2/4.3	0.025		

BALI (N = 34)

Genotype	Frequency	Allele	Frequency
1.1, 1.1	0.147	1.1	0.397
1.1, 1.2	0.207	1.2	0.250
1.1, 2	0.029	1.3	0.015
1.1, 4.1	0.059	2	0.029
1.1, 4.2/4.3	0.207	3	0.015
1.2, 1.2	0.029	4.1	0.044
1.2, 1.3	0.029	4.2/4.3	0.250
1.2, 3	0.029		
1.2, 4.1	0.029		
1.2, 4.2/4.3	0.147		
2, 4.2/4.3	0.029		
4.2/4.3, 4.2/4.3	0.059		

JAVA (N = 24)

Genotype	Frequency	Allele	Frequency
1.1, 1.2	0.042	1.1	0.167
1.1, 1.3	0.083	1.2	0.104
1.1, 3	0.042	1.3	0.083
1.1, 4.2/4.3	0.166	2	0.042
1.2, 3	0.042	3	0.083
1.2, 4.2/4.3	0.124	4.1	0.021
1.3, 4.2/4.3	0.083	4.2/4.3	0.500
2, 3	0.042		
2, 4.2/4.3	0.042		
3, 4.2/4.3	0.042		
4.1, 4.2/4.3	0.042		

Table 8. Continued

Genotype	Frequency	Allele	Frequency
4.2/4.3, 4.2/4.3	0.250		
MADAGASCAR (N = 69)			
Genotype	Frequency	Allele	Frequency
1.1, 1.2	0.072	1.1	0.152
1.1, 3	0.044	1.2	0.181
1.1, 4.1	0.101	1.3	0.065
1.1, 4.2/4.3	0.087	2	0.043
1.2, 1.2	0.029	3	0.102
1.2, 1.3	0.044	4.1	0.196
1.2, 2	0.014	4.2/4.3	0.261
1.2, 3	0.072		
1.2, 4.2/4.3	0.101		
1.3, 2	0.014		
1.3, 4.1	0.044		
1.3, 4.2/4.3	0.029		
2, 2	0.014		
2, 4.2/4.3	0.029		
3, 4.1	0.044		
3, 4.2/4.3	0.044		
4.1, 4.1	0.059		
4.1, 4.2/4.3	0.087		
4.2/4.3, 4.2/4.3	0.072		

* If genotype or allele is not listed, then frequency is zero.

N is the number of individuals studied.

TABLE 9. Expected and observed heterozygosities for all six loci, all populations

	LDLR	GYPA	HBGG	D7S8	GC	DQA1
Java						
Expected	0.439	0.503	0.337	0.507	0.662	0.689
Observed	0.458	0.708	0.333	0.500	0.667	0.708
Atayal						
Expected	0.162	0.468	0.096	0.475	0.584	0.788
Observed	0.175	0.375	0.100	0.450	0.575	0.825
Bali						
Expected	0.444	0.444	0.314	0.507	0.490	0.702
Observed	0.353	0.471	0.324	0.559	0.441	0.765
Madagascar						
Expected	0.351	0.494	0.671	0.477	0.422	0.724
Observed	0.420	0.362	0.551	0.478	0.435	0.739

TABLE 10. Hardy-Weinberg equilibrium tests for all six loci

	LDLR	GYPA	HBGG	D7S8	GC	DQA1
Chi-Square						
Java	0.823	0.040	0.957	0.944	0.298	0.763
Atayal	0.577	0.202	0.774	0.739	0.109	0.853
Bali	0.222	0.724	0.853	0.545	0.568	0.933
Madagascar	0.096	0.111	0.176	0.976	0.279	0.107
Fisher	0.044	0.007	0.217	0.016	0.080	0.041
All populations						

TABLE 11. Power of discrimination values for all six loci

Population	LDLR	GYPA	HBGG	D7S8	GC	DQA1	Combined
Java	0.581	0.619	0.496	0.623	0.802	0.877	0.999
Atayal	0.280	0.604	0.176	0.608	0.755	0.913	0.998
Bali	0.581	0.588	0.475	0.625	0.618	0.868	0.998
Madagascar	0.515	0.616	0.363	0.610	0.625	0.946	0.999

TABLE 12. G-Test Results For Atayal

Population	G-Test Score	P-Value
African American	245.3556	0.0000 ± 0.0000
Alaskan Eskimo (BH)	207.1993	0.0000 ± 0.0000
Alaskan Eskimo (NS)	138.7254	0.0000 ± 0.0000
Ami	99.3873	0.0000 ± 0.0000
Arab	148.2016	0.0000 ± 0.0000
Bali	76.3896	0.0000 ± 0.0000
Basque	224.3466	0.0000 ± 0.0000
Basque Residents	246.4193	0.0000 ± 0.0000
Chamorro	65.0545	0.0000 ± 0.0000
Chilean	224.6996	$0.0000 + 0.0000$
Chinese	53.8396	0.0000 ± 0.0000
Elazig	123.4093	0.0000 ± 0.0000
German	199.0623	0.0000 ± 0.0000
Greek Cypriot	193.5212	0.0000 ± 0.0000
Italian	241.6420	0.0000 ± 0.0000
Japanese	74.0617	0.0000 ± 0.0000
Java	50.7302	0.0000 ± 0.0000
Korean	52.7988	0.0000 ± 0.0000
Madagascar	139.0750	0.0000 ± 0.0000
Ulan Ude	83.1581	0.0000 ± 0.0000
Navajo	195.5626	0.0000 ± 0.0000
Omani	216.2193	0.0000 ± 0.0000
Pueblo	217.9319	0.0000 ± 0.0000
Filipino	72.7306	0.0000 ± 0.0000
Sioux	217.6512	0.0000 ± 0.0000
Slovenian	255.2212	0.0000 ± 0.0000
SE Hispanic	150.3322	0.0000 ± 0.0000
SW Hispanic	175.5403	0.0000 ± 0.0000
Swiss	175.3553	0.0000 ± 0.0000
Tadjik	125.1079	0.0000 ± 0.0000
US Caucasian	201.7074	0.0000 ± 0.0000
Zimbabwe	210.2916	0.0000 ± 0.0000

<u>Maximum Likelihood Trees</u>

Figure 2 displays the Maximum Likelihood phylogeny and bootstrap values. In the resulting dendrogram, three major clusters were observed: 1) African American and Arabs, 2) Caucasians and 3) East Asians and Native Americans. The third cluster of East Asians and Native Americans bifurcates into two subclusters. The Atayal group segregates with the Chinese, whereas the Ami tribe clusters with the Native Americans. It should also be noted that the population from Madagascar segregates with the African/African descent populations. Bootstrap values that are above 50% (or 500) are regarded as significant phylogenetic relationships. All bootstrap values, except for two were above 50%. The two exceptions were slightly below 50%, at 46% and 40%.

The second Maximum Likelihood tree, contains the major Austronesian populations of the current study (Figure 5), had several bootstrap values well below 50% (500). Adding more populations accounts for the lower bootstrap values. There are three major clades in the tree; 1) East Asians/Native American groups, 2) Caucasian, and 3) African/African descent. The Madagascar and Arab populations both cluster with the African/African descent groups. The topology of the two trees is similar, but the addition of Bali and Java, split the East Asian/Native American clade into three subclades. The Ami continue to cluster with the Native Americans and the Atayal with the Han Chinese. The third monophyletic grouping consists of the Bali and Java populations. Another significant difference between the first and second ML tree, is that in second, the population from Madagascar segregates further from the Austronesian populations.
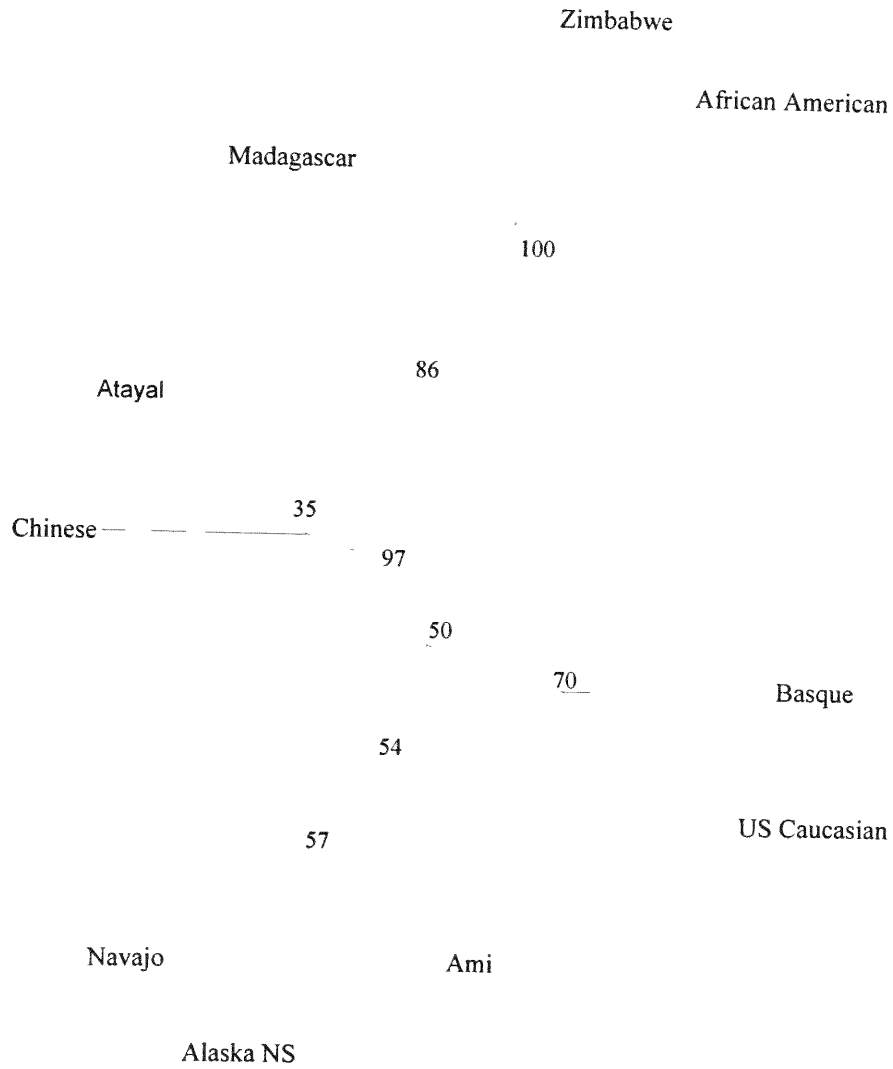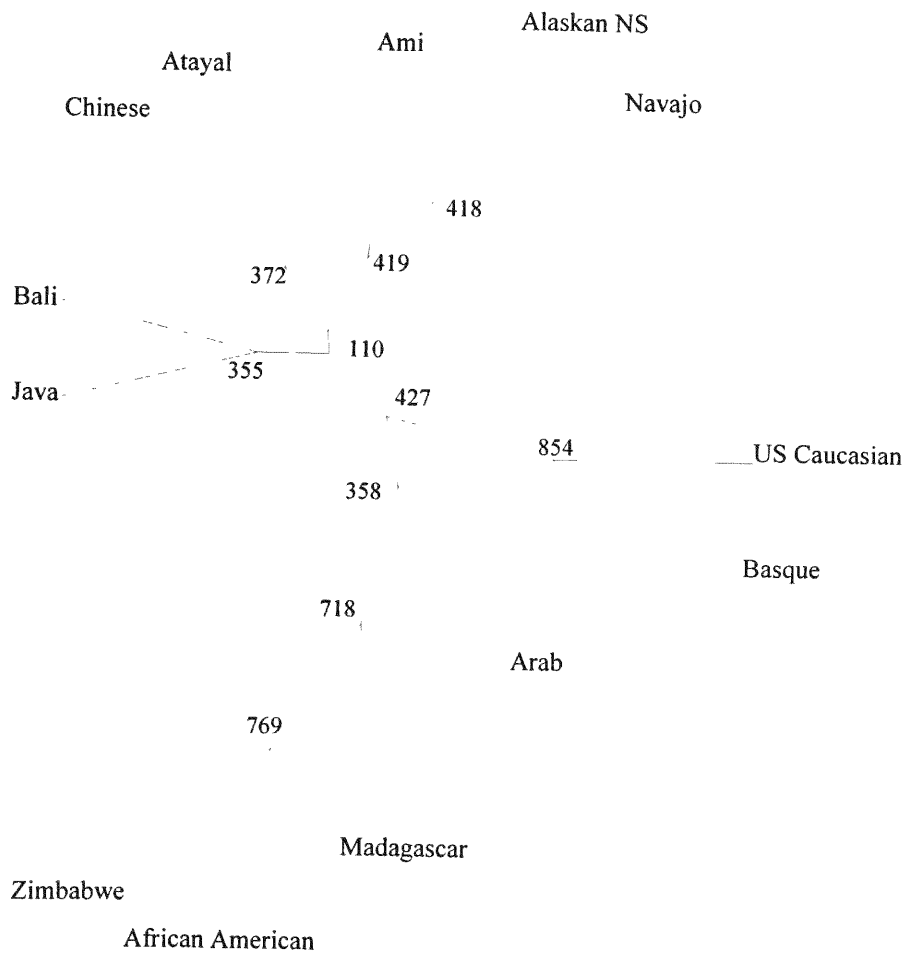
FIGURE 4. Maximum Likelihood Tree (Ten Populations)

Zimbabwe

African American

Madagascar

100

Atayal

86

Chinese —— —— —— 35

97

50

70 Basque

54

US Caucasian

57

Navajo   Ami

Alaska NS

FIGURE 5. Maximum Likelihood Tree (Thirteen Populations)

Atayal

Ami

Alaskan NS

Chinese

Navajo

418

372

419

Bali

110

355

Java

427

854

US Caucasian

358

Basque

718

Arab

769

Madagascar

Zimbabwe

African American

<u>Principal Component Results</u>

Figure 6 displays the PC analysis. PC 1 (the x-axis) and PC 2 (the y-axis) represents 28% and 25% of the variability, respectively. The Caucasian populations cluster together in the right center region of the map. The Southwest Hispanics lie at the outer edge of the Caucasian cluster towards the Native American cluster. PC 2 separates the Native American and East Asian populations. The Native American group appears in the upper central area on the map and includes the two Alaskan Eskimo populations. The East Asians (Ulan Ude, Chinese, Chamorro, Filipino, and Korean) cluster loosely together at the lower center, with the populations from Java and Japan on the periphery at opposite ends of the East Asian cluster along PC 2. Within Figure 3, the populations of Bali and Java segregate apart from each other, with Bali plotting closer to the Atayal and Ami. Java on the other hand, does not plot close to any group, but may be considered on the far fringes of the East Asian group. Interestingly, also included within the East Asian cluster is the population from Madagascar. There appear to be four outliers on the PC map: the Ami population in the left central edge, the Atayal in the lower left quadrant, the population from Bali, which lies in lower left quadrant and the African Americans in the lower right quadrant.

The second principal component map, Figure 7, compares only the seven Austronesian populations included in the study; Ami, Atayal, Bali, Chamorro, Filipino, Java and Madagascar. It is obvious that only three, Chamorro, Filipino, and Java, cluster together in the upper right quadrant of the map. The other four populations exist as outliers. Java and Bali are separated, as are the Ami and the Atayal. Within the plot,

PC1 accounts for 38% of the variability and PC2 for 23%. PC 1 separates the Ami and Atayal from the remaining populations, while PC2 separates the Atayal and Madagascar (lower half of plot) from the others.

Centroid analysis of Atayal

Figure 7 depicts the centroid analysis of the Atayal, Bali, Java and Madagascar along with 29 other populations. The Caucasian populations are positioned above the linear regression close to the centroid, whereas the East Asian populations fall directly below. The Native American populations are located farther beneath regression line and the apparent outliers, Ami and Atayal are the most distant groups away from the regression line. Two other possible outliers, the African Americans and the population from Madagascar are the furthest distance above the regression line.

FIGURE 6. Principle Component Map of All Populations
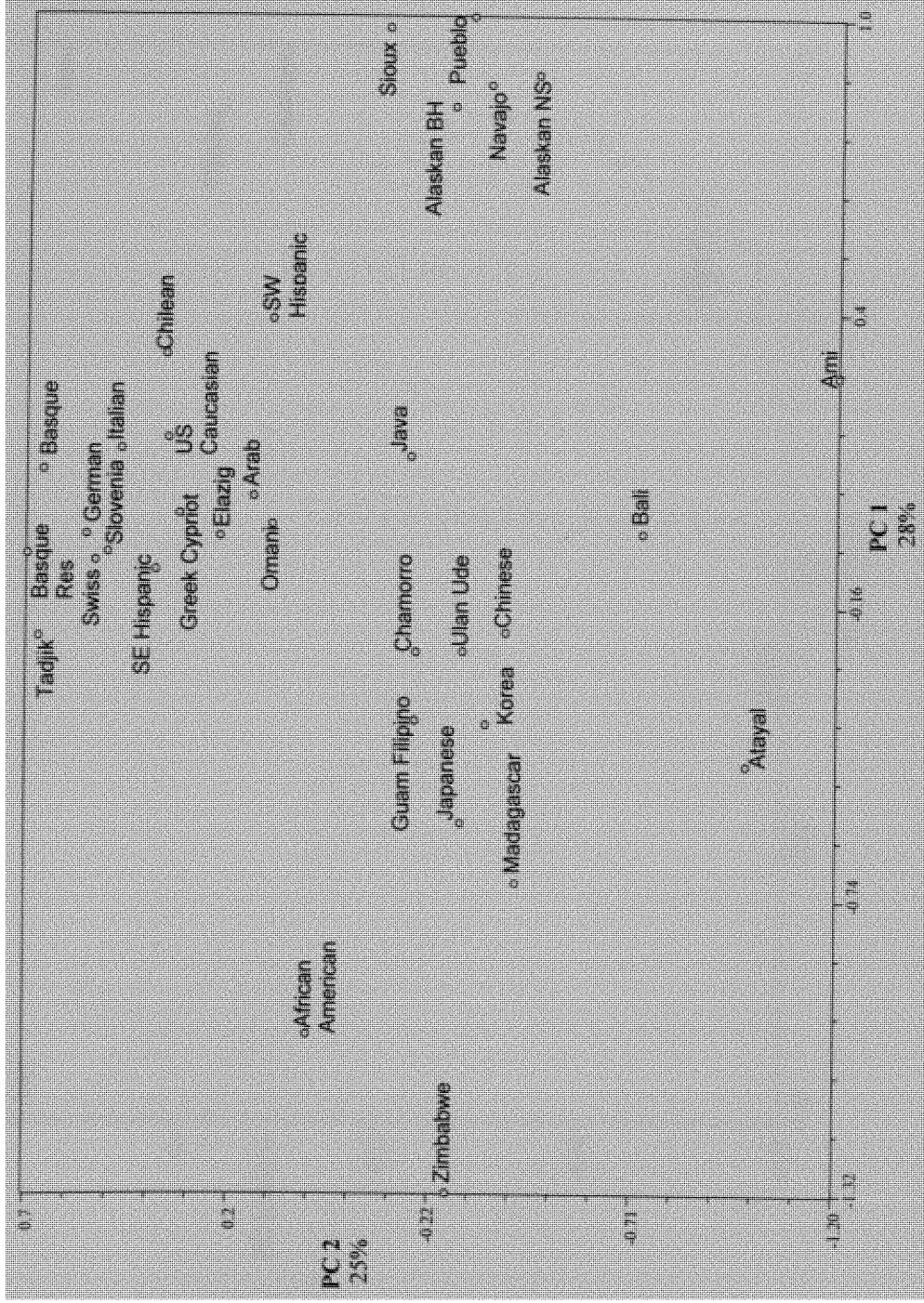
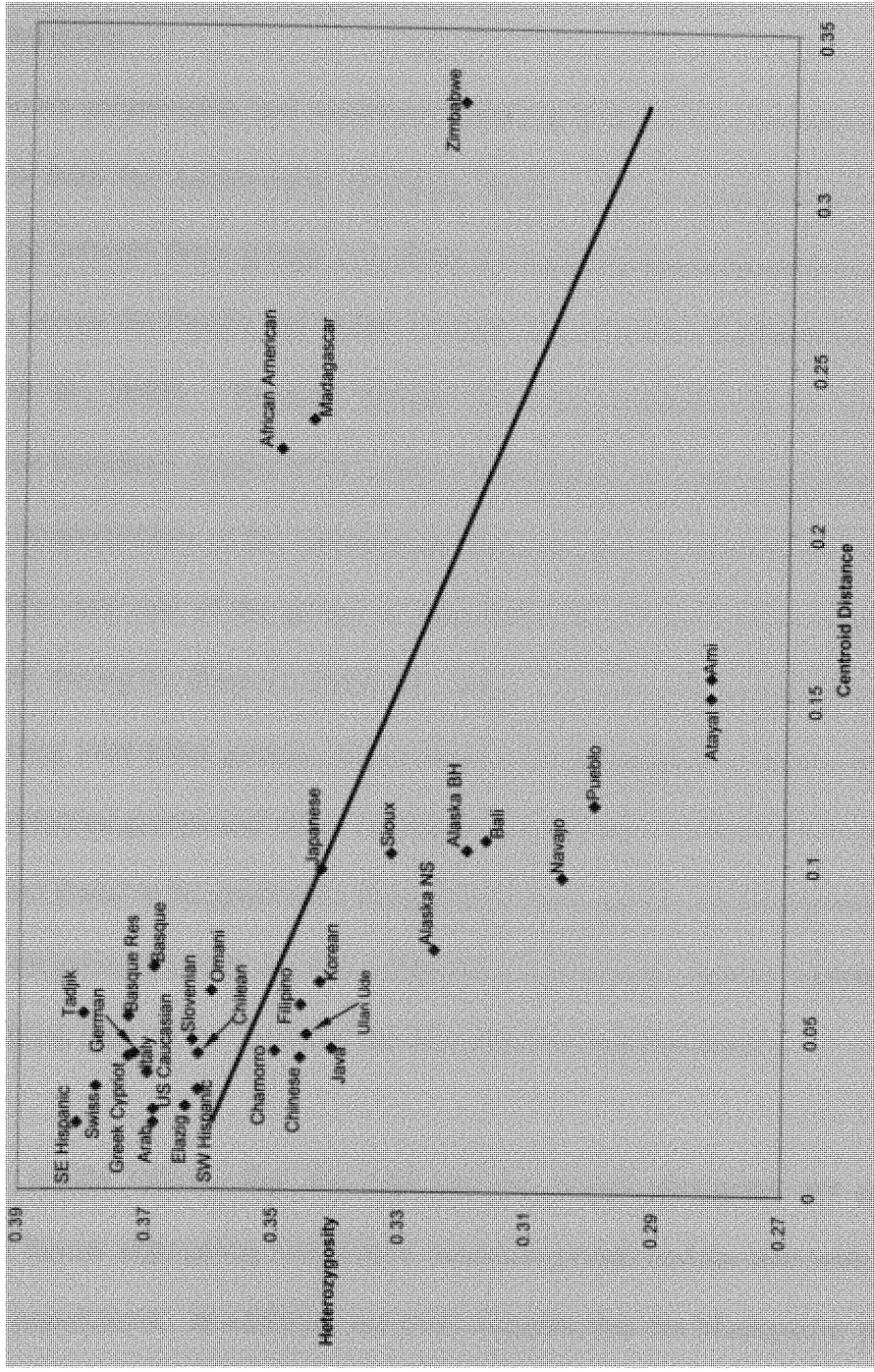# FIGURE 7. Principle Component Map of 7 Austronesian Populations

33
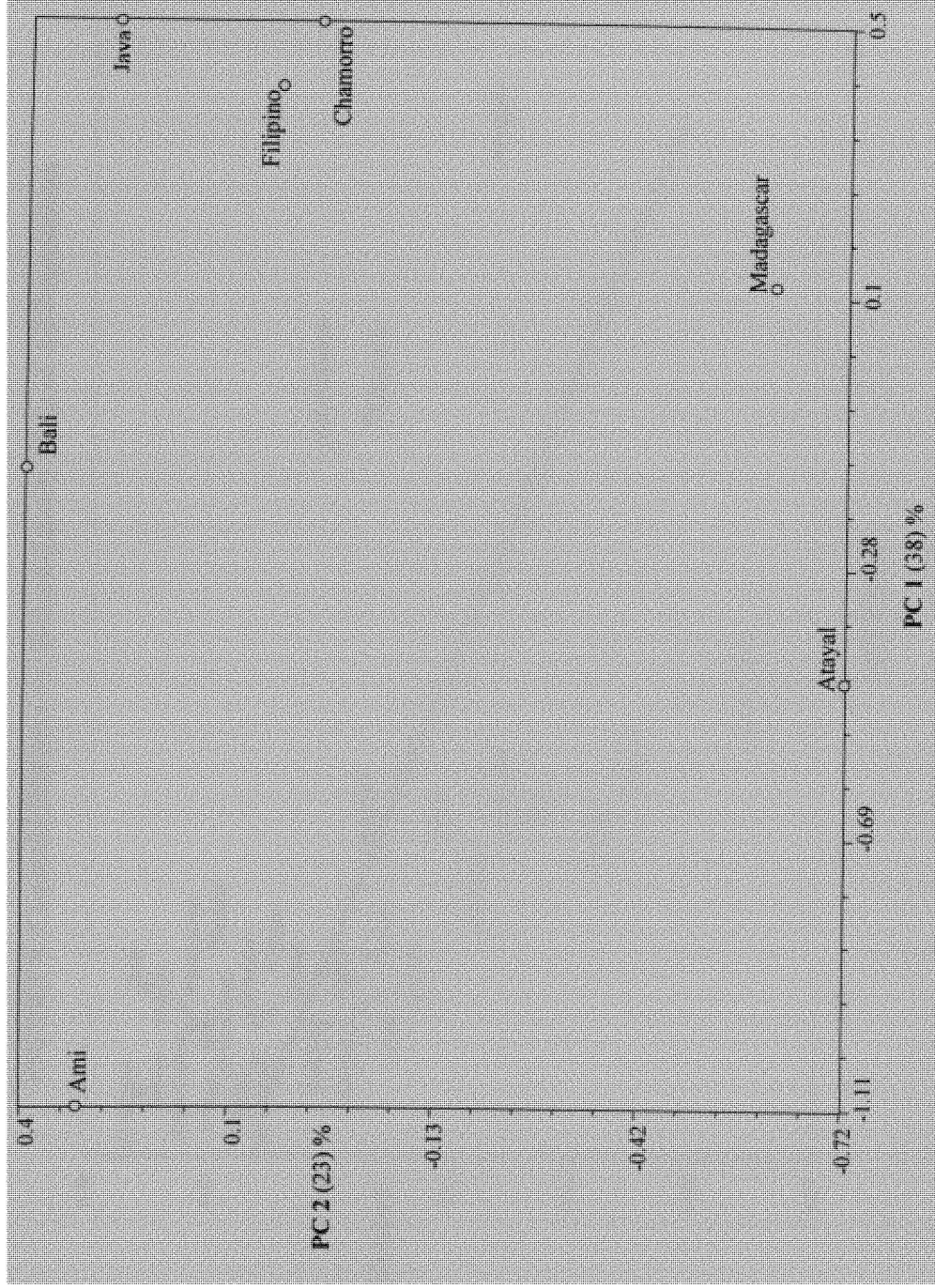
FIGURE 8. Centroid Distance vs. Heterozygosity Map of All Populations

34

CHAPTER IV

DISCUSSION

Research Objectives

The six polymorphic loci commonly known as PM/DQA1 were examined in four populations from the Austronesian language family. The investigation of the genetic relationships between these groups and others may help illustrate the path of past human migrations across the Pacific and Indian Oceans, known as the Austronesian expansion (Sewerin et al., 2002). Because Taiwan has been hypothesized as the potential Austronesian homeland, studies of the country's aborigines are extremely important. It is likely that diversification of the Austronesian language family occurred on Taiwan, and that, possibly, only one of the groups is responsible for the successive colonization of other islands, during the diaspora (Diamond, 2000). This idea emphasizes the importance of determining inter-tribal genetic relationships and relationships between the tribes to other Austronesian groups and other worldwide populations. In Sewerin's study (2002) the Ami were found to be genetically distinct from other groups around the world. In the current investigation the Ami are also compared to the Atayal, another Taiwanese tribal group, shedding light on the history of the Austronesian language family.

Our results indicate that the Ami and the Atayal aborigines are genetically distinct, the former segregating with Native American groups and the later with mainland East Asian populations. In addition, our work suggests that the Madagascar population exhibits phenetic affinities to African and East Asian groups. Also of interest are the

genetic differences between the Bali and Java populations, in spite of their geographic proximity, and the closer genetic relationship of the former to the Austronesian tribes examined. This information suggests that the Ami and the Atayal may have had independent origins in the Asian mainland and that the Balinese may represent a more ancient group of Austronesian ancestry as compared to a more admixed recent Java population. The present genetic evidence reaffirms linguistic data substantiating Austronesian influence in Madagascar and the westward reaches of the diaspora.

Implications of the Maximum Likelihood Trees

Figure 2 depicts the first ML tree, with four clusters. One clade contains populations from Africa (Zimbabwe and Madagascar) or of African descent (African American). In another cluster, the Native Americans segregate with the Ami, while the Atayal group with the mainland Han Chinese in a third clade. Caucasians are found together in the fourth cluster. The only Austronesian groups represented are the two Taiwanese aboriginal tribes and Madagascar. It is not surprising that the Zimbabwe, African American and Madagascar groups cluster together and away from the Caucasian and Asian/Native American groups. The population from Madagascar segregates with the Zimbabwe and African American groups as expected since the island of Madagascar lies just off the east coast of Africa in close proximity (about 500 miles) from Zimbabwe. As previous studies have shown, the Ami show a higher affinity to the Native Americans than to the East Asians (Sewerin et al., 2002). The Atayal tribe, on the other hand, clusters with the Han Chinese, which supports the connection between the Austronesian language family and Southeast Asia. Previous studies, using 13 classical loci, indicates

that the Toroko, a branch of the Atayal, have higher affinities to the Philippines and Thailand, than to the populations of south China and Vietnam (Chen et al., 1985). It is interesting to note that the two Taiwanese aboriginal groups segregate into different clades. There are two possible explanations for this. One is that their differences may reflect separate origins from diverse populations that arrived in successive waves to inhabit Formosa and the second is that these variations may simply be due to the subsequent geographical partitioning and genetic differentiation of tribes from a common origin (Melton el al., 1998).

The second ML tree, with all four populations of interest (Atayal, Bali, Java and Madagascar) supports the ideas that were suggested by the first ML tree. The Atayal continued to show differences from the Ami and grouped with the Han Chinese. This supports the idea that the tribes may be the result of separate origins. In contrast, the population from Madagascar showed less of Austronesian influence in the second ML, and did not support the findings on the first ML tree. Bali and Java both exhibited Austronesian influences, but also displayed that differences, existed between the two. This was surprising, since the two are geographically close. The PC plots described in the following section, elaborate on the relationships suggested between these populations.

The relationship exhibited by the groups from Africa and the East Asian/Native American populations underscore two important issues. First, is the segregation of the Madagascar groups away from the Zimbabwe and the African Americans in the African clade and its proximity to the Atayal and Han Chinese. It is significant that the

Zimbabwe, an East African population, is genetically closer to African Americans, an admixed group of West African descent, than to the sample from Madagascar, just off the coast of Africa. It is possible that the Madagascar population's genetic affinity to the Atayal and the continental East Asians (Hans) may be the result of the Austronesian migration into the island approximately 3,200 years ago (Ruhlen, 1994). The Ami aborigines, on the other hand, segregate distant from the Madagascar group in a cluster with two Native American groups. The extensive genetic distance between the Ami and the Madagascar population argue against a close relationship between the two. These results point to the Atayal and not the Ami aborigines as a stronger candidate for the Formosan source population responsible for the westward Austronesian dispersal. As previously reported, the clustering of the Ami with the Native Americans may be the result of a common ancestral source population for both in mainland East Asia (Sewerin et al., 2002). The Navajo and the Alaskan Eskimos signify recent arrivals (approximately within the last 2,000 years) to the New World representing distinct language groups, Nadene and Eskaleut, respectively. The segregation of these African and East Asian/Native American groups in the ML tree mirror the genetic affinities reflected in the following PC analyses.

Relationships Suggested by PC Plot and Centroid Analysis

The principal component plot, a two dimensional illustration of the allelic variability between the populations is depicted in Figure 3. The first and second principal components account for 53% of the total variation. Within the plot five groups are evident: 1) Caucasian, 2) East Asian, 3) Native American, (4) African or African descent

and (5) a scattered set that includes the Ami, Atayal and Bali. The second principal component separates the Caucasians from the East Asian and Native American clusters. Also along PC 2, the scattered Ami/Atayal/Bali grouping segregates away from the East Asian cluster. The Native American groups clearly separate from the East Asians. The position of the Southwest Hispanic group which lies on the fringe of the Caucasian cluster and towards the Native Americans reflects the admixed nature of this population, which is known to have Amerind and European contributions. The large Euclidean distance between the Ami and the Atayal belies the fact that both tribal groups are in close geographical proximity. It is possible that these two aboriginal groups have different ancestral populations in the Asian mainland. It is also likely that their migration to the island of Formosa occurred at different times from unique regions. Alternatively, extreme genetic drift and/or genetic isolation may be at least partially responsible for their phenetic differences. Another interesting observation is that the population from Madagascar clusters away from the African American and Zimbabwe, while grouping within the East Asian cluster at the fringes in the direction of the Africans and groups of African descent. As with the ML analysis, the position of the Madagascar group may reflect the Austronesian contribution to the island population. Of the seven Austronesian populations in the plot, only three (Madagascar, Chamorro, and Guam Filipino) cluster together. The three form part of the East Asian group. The other four (Ami, Atayal, Bali, and Java) segregate loosely along PC 2. It is surprising that Bali and Java plot distantly from each other in spite of their geographic proximity. It is obvious that these populations are genetically unique and distinct, even though they are geographically close and belong to the Austronesian language family. Compared to Bali, the island of Java is

much larger and more culturally diverse. Subsequent to the Austronesian dispersal, Java has been invaded by Hindus, Buddhists and Muslims. Culturally, Bali is predominately Hindu. Greater genetic flow from different groups and/or effective population size in Java may be an explanation for the observed differences between these two islands. This position of Java above and Bali below the linear regression in the centroid analysis (Figure 4) corroborate this possibility. Alternatively, these populations may have undergone extreme drift and/or genetic isolation subsequent to migration into the two islands.

The second PC plot, Figure 7, compares all Austronesian populations in the study. The separation of Bali and Java support the observations in the first PC, of the existence of genetic differences between the two. Only one cluster is evident within this plot, including the populations of Chamorro, Filipino, and Java. The genetic differences between the Atayal and Ami were also supported in this plot. Madagascar segregated closer to the Atayal, lending further support to the relationship between the two. On the other hand, the closer relationship between Bali and the Atayal than Java and the Atayal, was not evident, as within the first PC analysis. Overall, it is obvious that the Ami, Atayal, Bali and Madagascar have little affinity for each other when analyzed using only Austronesian populations.

Figure 4 represents the centroid analysis, which gives an indication of the relative amount of gene flow experienced by populations and/or effective population size. Populations that plot above the regression line could possibly be larger in size or receive

more gene flow than those below the regression line. The African American and Madagascar groups are outliers and the farthest above the regression line. This supports the fact that the African American populations are highly admixed. Madagascar may also have a similar amount of gene flow from diverse sources. It is likely that the Austronesian groups that made it to Madagascar admixed with the populations originally from mainland Africa. The position of the Zimbabwe group above the linear regression most likely is the result of the greater diversity and heterozygosity of sub-Saharan African population (Cavalli-Sforza et al., 1994). Within the plot, all Caucasian populations fall directly above the regression line, while East Asians fall directly below the line. The Native Americans lie directly below the Asian populations as outliers. The Bali population is positioned near the Native American groups which are known for low intra-population diversity but high heterogeneity among groups (Sewerin et al., 2002). The Ami and Atayal lie the furthest below the line, in close proximity. This indicates that while these two groups are genetically different, they both experienced little gene flow and have remained isolated. Population size may have also had an influence, as both populations are extremely small and limited in diversity and heterozygosity. This genetic data supports the idea that the aboriginal tribes of Taiwan have remained isolated from the Han Chinese population and from each other. The fact that almost all tribes have different material cultures and social organizations substantiates this conclusion (Chai, 1967).

Summary

In conclusion, the Atayal and the Ami aboriginal groups are genetically unique, as indicated by the PC plot and ML tree. These results are possibly an indication that the Ami and Atayal may have had different ancestral source populations originating in mainland Asia. The centroid analysis argues for a small effective population size along with limited amount of gene flow for both the Ami and the Atayal. Genetic drift due to founder effect and/or isolation may also be responsible for the differences between the two aboriginal groups. The Madagascar population shows relationships with the African/African descent groups as well as the Atayal, which may be due to an Austronesian influence on the island some 3,200 years ago. Madagascar's segregation with the East Asians in the PC plot is further support of this; it does not segregate with Zimbabwe, which is geographically closer. Madagascar displayed a genetic affinity for the Atayal, while maintaining a large genetic distance with the Ami. This gives support to the suggestion that the Atayal may be a better choice as the source population for the Austronesian expansion. Although the islands of Bali and Java are geographically close, genetic differences were observed between the two. A possible explanation for the differences between the two populations may be that Bali receives less gene flow than Java. Bali and Java have had dissimilar recent migration histories. While Bali is primarily Hindu, Java represents a more diverse population including Hindu, Muslims and Buddhists. Together, these relationships support previous archeological and linguistic data collected and supports a westward expansion of Austronesian populations, originating from or near Formosa reaching to the island of Madagascar. In addition, the

study also provides insight into the origin of the Formosan aborigines, which may be the result of different source populations colonizing the island.

LIST OF REFERENCES

Antunez de Mayolo G, Antunez de Mayolo A, Antunez P, et al. 2002. Phylogenetics of worldwide human populations as determined by polymorphic Alu insertions. Electrophoresis 23: 3346-3356.

Batzer MA, Deininger PL. 1991. A human-specific subfamily of *Alu* sequences. Genomics 9:481-487.

Bellwood P. 1991. The Austronesian dispersal and the origin of the languages. Scientific American 265:88-93.

Brown RJ, Rowold D, Tahir M, Barna C, Duncan G, Herrera RJ. 2000. Distribution of the HLA-DQA1 and polymarker alleles in the Basque population of Spain. Forensic Sci Int 108:145-151.

Budowle B, Lindsey JA, Decon JA, Koons BW, Giusti AM, Comay CT. 1995. Variation and population studies of the loci LDLR, GYPA, HBGG, D7S8, and Gc (PM loci) and HLA-DQα using a multiplex amplification and typing procedure. J Forensic Sci 40:45-54.

Budowle B, Baechtel FS, Fejeran R. 1998. Polymarker, HLA-DQA1, and D1S80 allele frequency data in Chamorro and Filipino populations from Guam. J Forensic Sci 43:1195-1198.

Butler J. 2001. Forensic DNA Typing. California: Academic Press, Elsevier Science.

Cann RL. 2001. Genetic clues to dispersal in human populations: retracing the past from the present. Science 291:1742-1748.

Cariolou MA, Manoli P, Christophorou M, Bashiardes E, Karagrigoriou A, Budowle B. 1998. Greek Cypriot allele and genotype frequencies for Amplitype® PM-DQA1 and D1S80 loci. J Forensic Sci 43:661-664.

Cavalli-Sforza LL, Menozzi P, Piazza A. 1994. The history and geography of human genes. Princeton, Princeton University Press.

Chai CK. 1967. Taiwan aborigines: a genetic study of tribal variations. Cambridge: Harvard University Press.

Chen KH, Cann H, Chen TC, Van West B, Cavalli-Sforza L. 1985. Genetic markers of an aboriginal Taiwanese population. Am J Phys Anthropol 66:327-337.

Chu HL. 1997. An introduction to the indigenous culture of Taiwan. Taipei: Charity Printing Industrial Co. p 6-30.

Diamond JM. 1988. Express train to Polynesia. Nature 336:307-308.

Diamond JM. 2000. Taiwan's gift to the world. Nature 403:709-710.

Drobnic K, Regent A, Budowle B. 2000. The Slovenia population data on the PCR-based loci HLA-DQA1, LDLR, GYPA, HBGG, D7S8, GC, and D1S80. J Forensic Sci 45:689-691.

Dulger HE, Tokdemir M. 2000. HLA-DQA1 and polymarker locus allele frequencies for Elazig/Turkey. J Forensic Sci 45:942.

Felsenstein J. 1985. Confidence limits on phylogenies: an approach to using the bootstrap. Evolution 39:783-791.

Felsenstein J. 1993. PHYLIP (Phylogeny Inference Package). Seattle.

Fildes N, Reynolds R, Ehrlich HA. 2002. HLA-DQA1 and polymarker procedure book. Roche Molecular Systems. Foster City, CA.

Garcia O, Martin P, Albarran C, Alonso A. 1994. Allele frequencies of HLA-DQA1, LDLR, GYPA, HBGG, D7S8, and Gc in the resident population of the Basque country; in Mangin P, Ludes B (eds): Acta Medicinae Legalis. Berlin, Heidelberg and New York. pp 81-82.

Gray RD, Jordan FM. 2000. Language trees support the express train sequence of Austronesian expansion. Nature 405:1052-1055.

Harpending HC, Ward RH. 1982. Chemical systematics and human population. Biochemical aspects of evolutionary biology. Chicago: University of Chicago Press.

Hayes JM, Budowle B, Freund M. 1995. Arab population data on the PCR-based loci: HLA-DQA1, LDLR, GYPA, HBGG, Gc and D1S80. J Forensic Sci 40:888-892.

Holden CJ. 2002. Bantu language trees reflect the spread of farming across sub-Saharan Africa: a maximum parsimony analysis. Proc R. Soc Lond B. 269:793-799.

Huang MC. 1964. Studies in the distribution of Rh blood types among various racial tribes in Formosa. Nippon Koigaku Zasshi [Jpn J Leg Med]. 18: 135-142.

Huang NE, Budowle B. 1995. Chinese population data on the PCR-based loci HLA-DQ alpha, low density lipoprotein receptor, glycophorin A, hemoglobin $^G\gamma$, D7S8 and group specific component. Human Hered 45: 34-40.

Ikemoto S, Ming CT, Haruyama N, Furumata T. 1942. Blood group frequencies in the Ami tribe (Formosa). Proc Jpn Acad: 173-177.

Jorquera H, Budowle B. 1998. Chilean population data on ten PCR-based loci. J Forensic Sci 43:171-173.

Kao TY. 1958. Taiwan historical events. Taipei: Cheng Chung Publishing.

Kirch PV. 1997. The Lapita peoples: ancestors of the oceanic world. Massachusetts: Blackwell Publishers Inc.

Kirch PV. 2000. On the road of the winds: an archeological history of the Pacific Islands before European contact. California: University of California Press.

Lewontin RC, Felsenstein J. 1965. The robustness of homogeneity test in 2 X N tables. Biometrics 21:19-33.

Li CC. 1976. First course in population genetics. Pacific Grove, CA: Boxwood Press.

Luis JR, Terreros MC, Martinez L, Rojas D, Herrera RJ. 2003. Two Problematic human polymorphic Alu insertions. Electrophoresis 24: 2290-2294.

Lum JK, Cann RL. 2000. mtDNA lineage analysis: Origins and migrations of Micronesians and Polynesians. Am J Phys Anthropol 113: 151-168.

Melton T, Clifford S, Martinson J, Batzer M, Stoneking M. 1998. Genetic evidence for the Proto-Austronesian homeland in Asia: mtDNA and nuclear DNA variation in Taiwanese aboriginal tribes. Am J Hum Genet 63:1807-1823.

Merriwether DA, Friedlaender JS, Mediavilla J, Mgone C, Gentz F, Ferrell RE. 1999. Mitochondrial DNA variation is an indicator of Austronesian influence in Island Melanesia. Am J Phys Anthropol 110: 243-270.

Nakajima H, Ohkura K. 1971. The distribution of several serological and biological traits in East Asia, volume 3, the distribution of gamma globulin (Gm [1], Gm [2], Gm [5], and Ivn [1]) and Gc groups in Taiwan and Ryukyu. Hum Hered 21: 362-370.

Ohashi J, Naka I, Ohtsuka R, et al. 2004. Molecular polymorphism of ABO blood group gene in Austronesian and non-Austronesian populations in Oceania. Tissue Antigens 63: 355-361.

Rangel-Villalobos H, Rivas F, Torres-Rodriguez M, Jolama-Cruz AR, Gallegos-Arreola MP, Lopez-Satow J, Cantu JM, Figuera LE. 1999. Allele frequency distributions of 6 Amp-FLPS (D1S80, APO B, VWA, TH01, CSF1PO and HPRTB) in a Mexican population. Forensic Sci Int 105:125-129.

Rohlf FJ. 2002. NTSys pc (Numerical Taxonomy System). Exeter Publishing Ltd. Setauket.