Y-Chromosome Analysis Favors the Fatimid Expansion into the Levant as the Conveyor of the Sickle-Cell Gene into Lebanon

by

Simon G. Khoury

Submitted in partial fulllment of the requirements

for the Degree of Master of Science

In Molecular Biology

Under the supervision of Dr. Pierre Zalloua

School of Arts and Sciences

Lebanese American University

January 2010

LEBANESE AMERICAN UNIVERSITY

Thesis approval Form (Annex III)

Student Name: Simon Khoury I.D. #: 200203224

Thesis Title : **Y-Chromosome Analysis Favors the Fatimid expansion as the conveyor of Sickle Cell Gene into Lebanon**

Program : M.S. in Molecular Biology

Division/Dept : Natural Sciences Department

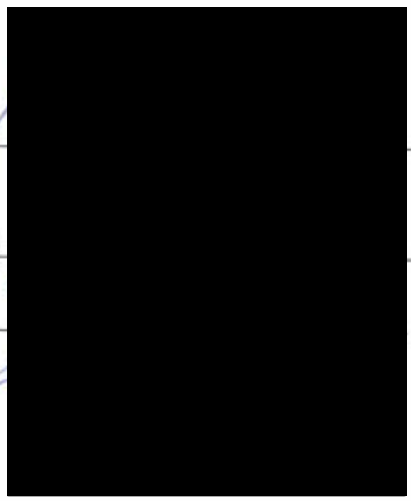
School : **School of Arts and Sciences**

Approved by:

Thesis Advisor: Dr. Pierre Zalloua

Member : Dr. Ralph Abi Habib

Member : Dr. Roy Khalaf



Date: June 2010

I grant to the **LEBANESE AMERICAN UNIVERSITY** the right to use this work, irrespective of any copyright, for the University's own purpose without cost to the University or to its students, agents and employees. I further agree that the University may reproduce and provide single copies of the work, in any format other than in or from microfilms, to the public for the cost of reproduction.

Dedicated to

The sweet memory of my mother, and to my family:

My father, Samer, Sandy and Stephany...

Y-Chromosome Analysis Favors the Fatimid Expansion into the Levant as the Conveyor of the Sickle-Cell Gene into Lebanon

Simon G. Khoury

Abstract

Sickle cell disease (SCD) in Lebanon was historically thought of by researchers as the import of the Arab slave trade, which flourished in the Arabian Peninsula. Previous studies have shown that the disease clusters North and South of Lebanon, and particularly within the Sunni community. We wished to test whether the previous facts could be translated into a correlation with specific Y-chromosomal haplogroups/haplotypes. The Y-chromosomes of 36 anonymous sickle-cell gene carriers and patients were haplotyped for 17 Short Tandem Repeats (STR) loci using an Applied Biosystems AmpFlSTR Yfiler kit. Y-Haplogroup was determined using a custom TaqMan SNP genotyping assay from Applied Biosystems, wherewith the 58 biallelic markers genotyped for each subject defined his Y-Haplogroup. Results indicated a significant association with haplogroup J1, followed by haplogroup E3b. Subjects belonging to the J1 haplogroup had closely-related STR haplotypes, indicating a recent ancestor and a probable founder effect, all of which probably accentuated by high rates of endogamy and large sibship size. We constructed a hypothetical modal haplotype closely representing that of the probable ancestor, and estimated the time of coalescence to be around 800-1200 A.D. Geographical mapping of the modal haplotype corresponded to the North-African Coast, in a declining gradient from East to West. Thus geographical, historical and genetic data suggest that the Fatimid expansion, and not the Arab slave trade, was the migratory event responsible for introducing the SCD into Lebanon. The modal J1 haplotype we constructed could also be used as a fingerprint for the Fatimid expansion into the Levant.

Acknowledgements

First and foremost, I would like to extend praise and thankfulness to the Lord for his fatherly love, relentless watch and unmerited favors, that he has bestowed upon me.

My deepest gratitude goes to my advisor, Dr. Pierre Zalloua, an extraordinary scientist and brilliant researcher whose love for knowledge drove me forward and provided me with a constant challenge to learn and improve. I also thank him for the opportunity he presented me with to delve into new scientific territories and explore fascinating concepts that made me appreciate science and Biology even more.

To my defense committee members, Drs. Roy Khalaf and Ralph Abi-Habib, as well as Drs. Mirvat Sibai, Constantine Daher and Georges Baroudy, I say: You were more family than faculty to me. Your supervision helped me become a better scientist and a better man, and for that I am forever in your debt.

To my lab colleagues, Marc Haber, Sonia Youhanna and Stephanie Saade: You helped in more ways than I can ever thank you for. It was only with your valuable input and pleasant company that this work was made possible.

Finally, all love and thanksgiving go to my family: To my beloved father, to the memory of my mother, and to Samer and Sandy and to my beloved Stephany. You have been there for me every step of the way... And when the going got tough, it was you who got me going.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The first to hypothesize that a molecular abnormality in the hemoglobin molecule might be at the origin of the sickle cell disease was Linus Pauling in 1945. Pauling later confirmed his hypothesis in 1949, as demonstrated by gel electrophoresis and the differential mobility of sickle hemoglobin (Hb S) as opposed to normal adult hemoglobin (Hb A) (Pauling *et al.*, 1949). That same year, the disease was proven to be autosomal recessive by Neel *et al.* (Neel, 1949), and was as such the first human disease to be understood at the molecular level. In 1956, Ingram *et al.* showed that the mutant sickle hemoglobin (Hb S) and the normal adult hemoglobin differed only by a single amino acid substitution, namely a glutamine to valine substitution at the sixth residue of the β-globin polypeptide (6 Glu $\rightarrow$ Val) (Ingram, 1989).

1.1 A Molecular Disease

As mentioned above, the pioneering work conducted by Pauling *et al.* demonstrated that SCD results from a mutation affecting the hemoglobin molecule, and the mode

of inheritance for this disease was proven to be autosomal recessive in the same year.

As such, individuals who are affected with SCD carry two copies of the affected hemoglobin variant (Hb SS, or Hb S homozygotes), and the predominant hemoglobin found in their erythrocytes is the sickle hemoglobin. Heterozygotes (Hb AS) have only one copy of the sickle variant and one copy of the normal (Hb A) hemoglobin, and both hemoglobin types are found in their bloodstream. They are referred to as carrier individuals, or having the "sickle cell trait" as opposed to "sickle cell disease" (Ashley-Koch *et al.*, 2000).

There are also other types of SCD in which individuals are compound heterozygotes. Compound heterozygotes possess one copy of the sickle hemoglobin variant (Hb S) in addition to another β-globin gene variant, such as Hb β-Thalassemia or Hb C (hemoglobin C, incorporating lysine in- stead of glutamic acid in codon 6). Figure 1.1 shows various mutations in the β-globin gene that can lead to Hb S, Hb C and other abnormal variants of hemoglobin.

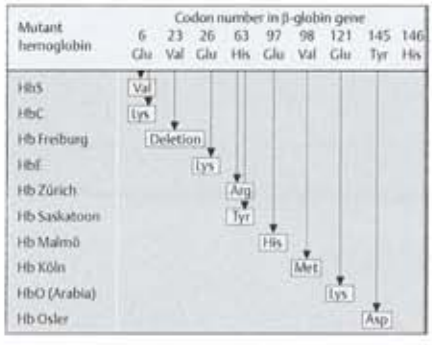


Figure 1.1: Amino acid sequence variations in mutant hemoglobin form. Modified from Weatherall (2001).

1.1.1 β-globin and α-globin gene clusters

Two β-globin sub-units, in addition to two α-globin sub-units make up a normal adult hemoglobin molecule, as opposed to fetal hemoglobin (Hb F) that comprises two β-globin and two γ-globin subunits. The globin sub-units, as well as several other related genes are encoded on two separate gene clusters (See figure 1.2). In man, the β-globin-like genes (α, Gγ, Aγ, δ and β) map to the β-globin gene cluster on the short arm of chromosome 11, region 1, band 5 and sub-band 5 (11p15.5), over a span of 60 kb. The two γ genes, A_γ and G_γ arose via duplication events . The α-globin-like genes on the other hand are located on the short arm of chromosome 16 (Bank, 2005).

1.1.2 The Sickle-Cell Mutation

The molecular genetic basis of SCD is a missense mutation in the second position of codon 6 - codon 7 if start codon was taken into account - of the β-globin gene (on the β-globin gene cluster), as demonstrated decades later by Marotta *et al.* (1977) . It is a transversion of an adenine (A) to a thymine (T), thus changing codon GAG to GTG. This reflects on the amino acid sequence, and a change from a glutamic acid moiety (E) to a valine moiety (V) is observed (E6V). The affected amino acid is on the surface of the molecule. Therefore, a substitution of glutamic acid by the hydrophobic Valine results in sickle-cell hemoglobin (Hb S) that is less soluble than the normal (Hb A) hemoglobin.

The abnormal Hb S can crystallize in the deoxygenated state (via polymer-

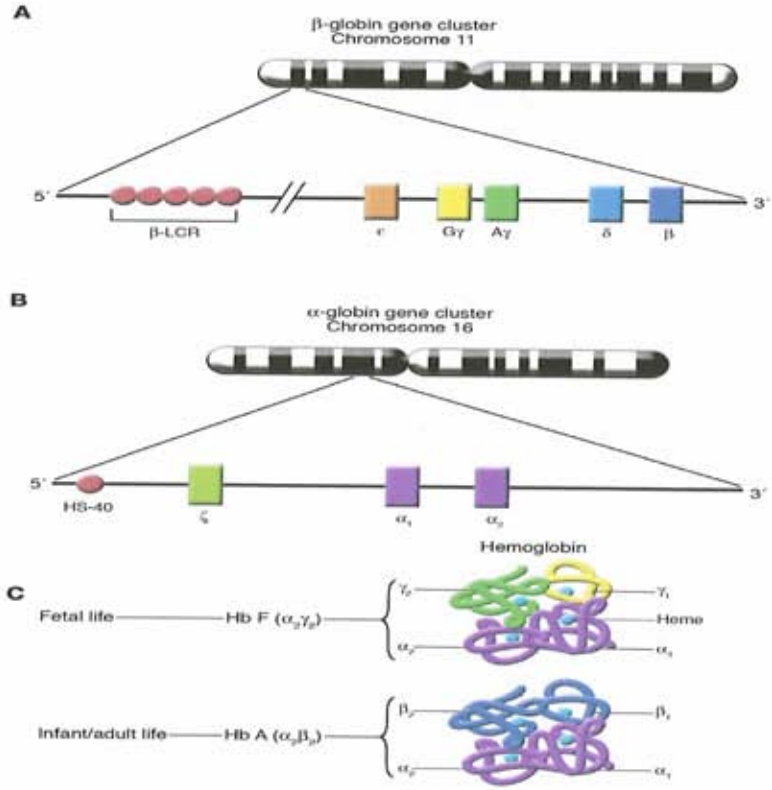


Figure 1.2: α- and β-globin gene clusters

(A) The β-globin gene cluster (harboring genes α, Gγ, Aγ, δ and β) is present on chromosome 11. Interestingly, the cluster genes occur in the same order in which they are expressed during development. β-LCR (β-Locus Control Region) is an upstream regulatory region that is essential for the proper expression of the included genes in the required levels. (B) The α-globin gene cluster is located on chromosome 16, and includes genes α, α_1 and α_2. They also figure within this cluster in the same order in which they are expressed during development. HS-40 is an upstream regulatory region necessary for satisfactory expression levels of the cluster genes. (C) In the fetal stage, Hb F predominates. It includes two α subunits and two β subunits ($\alpha_2\gamma_2$). Hemoglobin switching is the phenomenon defined on the molecular level by the silencing of the γ-globin gene and the activation of the β-globin (adult) gene. As such, Hb F is replaced by Hb A, the predominant type of hemoglobin in adult life. Figure modified from Bank (2005).

ization), adopting the morphology of "small rods". This causes erythrocytes to transmute from the usual biconcave disc shape into an irregularly shaped sickle-like form. In addition to their irregular form, sickled erythrocytes lose their flexibility that allows them to maneuver their way through small blood vessels, and gain a propensity to adhere to the latter. As a result, affected erythrocytes can clog small arteries and capillaries, resulting in local oxygen deficiencies in various organs, with all the implicated pathological sequelae (Ashley-Koch *et al.*, 2000).

1.2 The Sickle Cell Gene: Origin, Malarial Interaction and Gene Flow

1.2.1 Multicentric Origin Of The Sickle Cell Gene

The origin of the sickle cell trait can only be understood in light of β-globin haplotypes. Work on the chromosomal framework surrounding the β-globin gene (β-globin gene cluster, *vide supra*) debuted with the pioneering work of Kan and Dozy (Kan & Dozy, 1978)who described the polymorphism of a restriction endonuclease (*HpaI*) recognition site located about 5 kb 3' to the β-globin gene. Instead of a normal 7.6-kb fragment containing the β-globin gene, 7.0- and 13.0-kb variants were detected and were found in African Americans, Asians, and Caucasians. Such a polymorphism suggests that the Hb S gene was linked to two chromosome 11 "types", *i.e.* one bearing the sequence recognized by *HpaI* and one without it.

This suggested that the sickle cell mutation had occurred in at least two historically distinct and independent events. These limited initial data were interpreted to

mean that one of the chromosomes was characteristic of West Africa and the other of Equatorial and East Africa. Subsequent work by Mears *et al.* further rebutted the unicentric origin, suggesting even more diversity, such as the possibility of three distinct origins of Hb S: In Senegal, West Africa and Bantu-speaking Africa (Mears *et al.*, 1981; Wasi & Bowman, 1983).

Pagnier *et al.* used 11 polymorphic sites in the β-globin gene cluster that could be detected via sequence analysis and restriction fragment length polymorphism (RFLP). Those 11 restriction endonucleases sites are distributed over a region stretching from the 5' proximity of the ϵ-globin gene to 8 kb 3' to the β-globin gene. Pagnier *et al.* assumed those markers to be closely linked to the β-globin gene, and therefore if the mutational event leading to Hb S was fairly recent, then that mutation should be accompanied by a defined haplotype formed by a set of those polymorphic DNA markers, preexisting in the chromosome before the mutational event (Pagnier *et al.*, 1984a). In theory, there should be a wide variety of such haplotypes associated with the β-globin cluster of different subjects and populations; however Pagnier *et al.* revealed that only a specific subset was found to be associated with clusters carrying the sickle mutation. Although the latter haplotypes were commonly occurring, they mapped the highest frequency of each to a specific geographic region: The Benin haplotype found in central west Africa, the Senegal haplotype in the African west coast and the Bantu (Central African republic) haplotype found in Central Africa (*i.e.* Bantu speaking Africa) (Pagnier *et al.*, 1984b; Nagel *et al.*, 1985a).

A fourth African haplotype, the Cameroon haplotype, was discovered later, and

is found exclusively in the Eton people of The Cameroons (Lapouméroulie *et al.*, 1992) indicating a fourth independent origin of the Hb S gene in Africa. It seems that little to no expansion occurred beyond this original ethnic group.

Studies were conducted later on variable repeats of the ATTTT motifs located about 1.5 Kb 5' of the β- globin gene and the AT repeats (followed by T runs of different size) located about 0.5 kb 5' of the same gene. DNA sequencing in this region suggested that these repeats were polymorphic. Moreover, ATTTT motifs repeated either four or five times with $(AT)^{x}T^{y}$probes having X = 7 or 11 and Y = 7 or 3. Those results were complemented by sequence data from the –1080 bp stretch located 5' to the cap site of the β-globin gene. It became apparent that the combination of these polymorphic areas was unique for each haplotype, supporting the previous conclusions, and revealing a novel (fifth) haplotype (i.e. origin), known henceforth as the Arab-Indian haplotype. It is found today among the tribesmen of India and among Arabs living in the eastern oases of Saudi Arabia and in Oman and at varying frequencies in countries of the Arabian Peninsula.(Bandyopadhyay *et al.*, 1999; Das & Talukder, 2001; Rahimi *et al.*, 2006)

Indeed, sickle cell disease occurs in India mainly in tribal populations, in what can be viewed as genetic "pockets" isolated from the rest of mainstream society. Since it is improbable that an influx of sickle cell gene from the outside of India is capable of explaining the rates of heterozygosity that can amount to 35% in certain Indian tribes, the most likely scenario is that the mutation correlated with the fifth haplotype occurred originally in India, and that gene flow carried this trait from India to the Middle-East and the neighboring Arab countries (and other

non-Arab countries) by means of migrations, the Great Silk Road and commercial exchange.(Bandyopadhyay *et al.*, 1999; Das & Talukder, 2001; Rahimi *et al.*, 2006)

Other than their anthropological importance, sickle-cell haplotypes have clinical significance. For instance, the Senegal and the Arab-Indian haplotypes are associated with a milder clinical course, stemming from the fact that they are correlated with higher levels of Fetal Hemoglobin production (Hb F) (Green *et al.*, 1993).

1.2.2 Interaction With Malaria: The Heterozygous Advantage

Malaria has probably exerted the greatest selective pressure on the human genome in recent history. This is mostly due to the substantial rates of mortality inflicted by malaria, claiming between one to three million victims yearly, most of whom are from Sub-Saharan Africa(Kwiatkowski, 2005), where 90% of malaria-caused deaths occur.

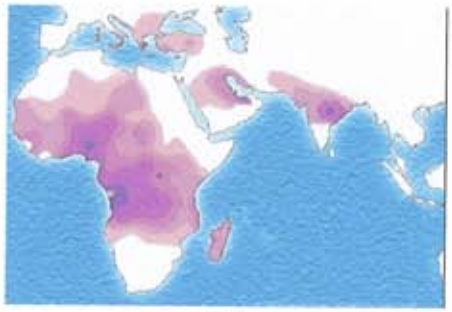
Malaria is caused by protozoan parasites of the genus *Plasmodium*. The gravest malarial infections are caused by *Plasmodium falciparum*. The vector for malaria is the female *Anopheles* mosquito.

Individuals heterozygous for the sickle cell mutation are a classic example of heterozygous advantage. Whereas individuals homozygous for the mutation (Hb SS) have full sickle-cell disease and rarely live beyond adolescence in traditional societies, heterozygous individuals (Hb AS) on the other hand are at distinct advantage in societies where malaria is endemic, where the frequency of the sickle cell gene could reach figures higher than 10% (Allison, 1964; Aidoo *et al.*, 2002). sickle cell trait

(HbAS) has been shown to confer strong protection against *Plasmodium falciparum* malaria in numerous studies conducted over the course of more than 50 years. However, the protective mechanisms at work remain incompletely understood (Aidoo *et al.*, 2002; Williams *et al.*, 2005b). The sickling of infected Hb AS red blood cells (RBCs) may be resulting in their premature destruction by the spleen before the daughter parasites emerge. Part of this protective effect could also be attributed to the altered biochemical/physical characteristics of Hb AS erythrocytes, and their effect on the malarial parasites. Indeed, the invasion, growth, and development of *Plasmodium falciparum* parasites are all reduced in such cells *in vitro*. Recent observations suggest that the mechanism at hand might also involve an immune component (Williams *et al.*, 2005a).

Whereas Hb AA (normal) individuals are at risk of succumbing to severe infections in endemic malarial areas, and Hb SS (sicklers) are prone to morbid and mortal sequelae, heterozygotes (Hb AS) produce only a limited amount of sickled erythrocytes, not enough to be symptomatic, but enough to impart resistance to malaria. It ensues that the heterozygotes have a higher fitness than either of the homozygotes. This is known as heterozygote advantage, and it explains the high frequencies of the sickle gene in areas where malaria is (or was) endemic (see Figure 1.3).

(a)



(b)

Figure 1.3: Geographical distribution of malaria and sickle-cell trait.As such, (a) Historical distribution of malaria. Note that malaria is no longer endemic in Europe, however it hasn't been eradicated on the southern borders of the former USSR and Turkey. Outbreaks were registered in the 1970's and 1980's in Azerbaijan, Turkey, and Tadjikistan. (b) Geographical distribution of sickle-cell trait, ranging from high frequency (dark fill) to low frequency (light fill).

1.3 Studies Of Gene Flow

β-globin and α-globin gene clusters have been exploited for several purposes in linkage studies and studies of gene flow. Below are some notable examples.

1.3.1 β-globin and α-globin Gene Cluster Haplotypes in Biology, Medicine and Anthropology

1. Anthropological correlations: β-globin and α-globin Gene Cluster Haplotypes have been used to determine the African and Indian origins of the sickle cell gene, as discussed hereinabove; They have been also used to give biological justification to the linguistic evidence of the Bantu expansion in Africa, as well as defining and estimating the likelihood of an ancestral home of the tribesmen of India (Muralitharan *et al.*, 2003).

2. Explaining clinical diversity among SC (Sickle Cell) patients: Clinical evidence suggests that linkage of the β^s gene to the Arab-Indian and Senegal haplotypes is associated with higher levels of production of fetal hemoglobin (Hb F) in the Hb SS state and therefore with a better clinical and hematological profile. In contrast, the Bantu haplotype has the worst clinical course (Nagel *et al.*, 1985b; Green *et al.*, 1993).

1.3.2 Slave Trade-Based Gene Flow of Sickle Cell Gene to America

Many Africans brought to America as slaves carried the sickle cell gene, and the genetic profile of current African-Americans reflects this fact. For instance, out of more than 20 haplotypes associated with the β^s-globin gene in Jamaica (which was populated by Africans enslaved from different parts of Africa), the three major haplotypes (See section1.2.1) constitute more than 95% of the cases (Wainscoat *et al.*, 1983; Antonarakis *et al.*, 1984). This further denotes that the three geographically distinct haplotypes were the major ones linked with β^s in Africa, whereas the rest of the haplotypes were "fresh" linkages established in Jamaica by novel mutations, gene conversion or crossing-over events (Muralitharan *et al.*, 2003).

1.3.3 Expansion in the Middle-East via the Sassanian Empire

The Arab-Indian haplotype has been found to be in linkage with the sickle gene in several countries of the Arabian Peninsula. This further corroborates the hypothesis that this African-independent mutation arose on the margins of the Indus valley, most probably within the Harappa culture 4000-5000 years ago, in present-day Pakistan, where the development of agriculture could have been the drive behind the emergence and exacerbation of malaria, thus contributing to the significant expansion of the gene frequency. The presence of the gene in the modern-day Arabian Peninsula countries such as eastern Saudi-Arabia, Kuwait, Bahrain, northern Oman (Kulozik *et al.*, 1986; Adekile, 1997; el Kalla & Baysal, 1998; Daar *et al.*, 2000; Bashwari *et al.*, 2001; Muralitharan *et al.*, 2003; Teebi & Teebi, 2005)and in Iran

(Rahimi *et al.*, 2003; Rahimi *et al.*, 2006), as well as India (Pagnier *et al.*, 1984a; Kulozik *et al.*, 1986) and Afghanistan (Daar *et al.*, 2000) suggests that the expansion of the Sassanian Empire (200-600 A.D.) could have been the carrier that brought the gene to Iran, the Arabian Peninsula and parts of the Middle-East (See Figure 1.4). In fact, the Sassanian Empire was bordered from the east by modern-day Pakistan and Afghanistan, whereas the southern limits extended to the northern shores of the Arabian Peninsula, including Kuwait, Bahrain, Qatar, United Arab Emirates and Sultanate Oman. In all of the latter countries, the sickle cell gene and the Arab-Indian haplotype were found (Rahimi *et al.*, 2003).

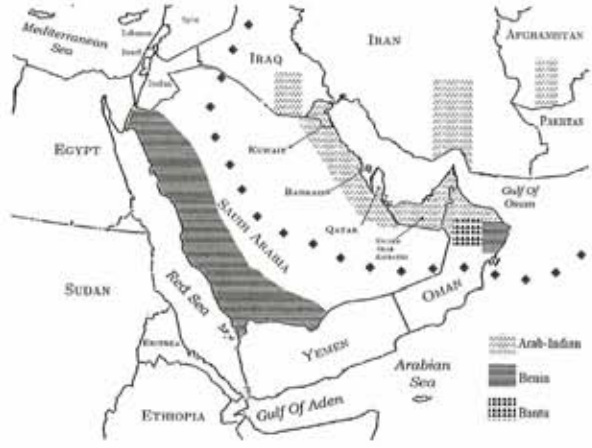


Figure 1.4: Distribution of haplotypes in different Middle-East regions. The diamond-dashed line represents the boundaries of the Sassanian empire during its peak. Figure constructed using data from Daar *et al.* (2000) and Teebi *et al.* (2005)

1.4 Sickle Gene in the Arab World

Several factors contribute to the genetic diversity of Arab populations. From major demographic events in antiquity (*e.g.* Roman, Greek, Sassanian and Phoenician expansions amongst others) to the migrations of Semitic tribes radiating from the

Arabian Peninsula, the Islamic conquests in the 7th century A.D., the Fatimid expansion and the Crusades in the 11th entury as well as the Mameluk and Ottoman occupation from the 13th to the 20th century.

Worthy of notice is the high prevalence of hemoglobinopathies in the Arabian Peninsula, including Hb S, Hb C and α- and β-thalassemia (Teebi & Farag, 1997; Daar *et al.*, 2000). More generally, autosomal recessive disorders are present in high frequencies, often much higher than in the western populations. Further, significant differences are encountered in abnormal gene frequencies between different Arab countries and also within different regions of the same country (Teebi & Farag, 1997; Teebi & Teebi, 2005).

1.4.1 Factors Affecting Hemoglobinopathies and Sickle Gene Frequencies the Arab World

Consanguinity Consanguineous marriages are a common feat in Arab communities and date back to biblical times (Kamal *et al.*, 1967). Recent literature abounds with reports of high rates of inbreeding in most Arab societies. In Lebanese Muslim communities for instance, the rate of consanguinity is about 25-30% with a high prevalence of first-cousin marriages (Khlat, 1988; Inati *et al.*, 2003; Inati *et al.*, 2007). An early study conducted on 84 Lebanese thalassemics reported that 49% of the patients were begotten by first-cousin marriages (Basson, 1979). More recently, a study by Inati *et al.* (2007) on 387 Lebanese SCD and β-thalassemia patients reported that nearly all patients were Muslim and that 56% were the offspring of consanguineous marriages, mostly between first degree relatives. Moreover, two-

thirds belonged to the same religious sect (Muslim Sunnis). This is consistent with the propensity of Arabs in general and Muslims in particular to pick their spouse from within the same religion and often from the same sect (Inati *et al.*, 2007).

Malaria Malaria was endemic in several Arab countries (including certain coastal areas of Lebanon before its eradication in the 1950's) (Teebi & Farag, 1997; El-Hazmi *et al.*, 1995).

Large Progeny Size This is a rather common feature of Arab families in general. In Saudi Arabia for instance, the average is 6-7 children/family (Teebi & Teebi, 2005). A resulting disadvantage is the large number of affected offspring in families with mutant or defective genes.

Conquests and Slave Trade The Arab slave trade refers to the practice of slavery in Arab countries. Slave trade was sanctioned in the religion of Islam, wherewith Muslim conquerors had the right to enslave their non-Muslim war-captives. Contrary to common beliefs, slaves were not limited to a certain ethnicity or color: In the early days, slaves were mostly Arabs and Berbers. Later, during the 8th and 9th century and under the rule system of the Caliphate, most slaves were Eastern Europeans (Slavic), Mediterranean, Persians, Turks, Caucasians, Scandinavian and Central Asian, in addition to those of African origin. It is only in the 18th and 19th century, however, that East Africa became the primary supply of slaves (Lewis & Lewis, 1990). It is forseeable that the massive numbers of slaves imported into various Muslim and Arab nations would had a discernable fingerprint on the genetic/genotypic profile of autochtonous populations, and would have definitely

altered the picture of hemoglobinopathies. This can be ascribed to the endemicity of malaria both in Africa and in Turkic/Caucasus regions where the majority of slaves were imported. On the other hand, the Islamic expansion that occurred in the 7th century served to shuttle previously localized genetic traits to the rest of Arab countries, as well as certain European and African countries.

1.4.2 Demography of the Sickle Gene in Arabia

The presence of Hb S in almost all Arab countries was confirmed by numerous studies and reports. Again, substantial differences in frequencies occur within and between different countries. HbS trait frequency ranges from less than 1% in Jordan and Lebanon, to 20% or more in Bahrain or eastern Saudi Arabia (Teebi & Teebi, 2005). Table 1.1 shows upper and lower range estimates of the frequency of Hb S gene in different Arab countries. A major difference is readily noticeable between Mediterranean Arab countries with frequencies that are much lower than then rest of the Gulf Arab countries. This can be attributed to lower rates of consanguineous marriages, less involvement in the practice of slavery, less influx of sickle gene from India and Persia and, in general, fewer foci of malarial endemicity.

As for haplotypes, it seems that the Arab world can be divided into two categories:

- In the eastern part of the Arabian Peninsula (See Figure 1.4) , including eastern Saudi Arabia, the Arab-Indian haplotype seems to be dominant. This can be attributed to several reasons, foremost amongst which are the aforementioned demographic dynamics with the Harappa culture and the Sassanian expansion,

Table 1.1: Estimates of Hb S frequency in different Arab populations*

Country	HbS %
Algeria	0.83-3.5
Bahrain	~1-20
Jordan	0.5-1
Egypt	<1-22.7
Lebanon	0.34
Libya	0.44-6.31
Saudi Arabia	<1-21.3
Sudan	1.52-10
Syria	<1
Iraq	5.25
Tunisia	6.0
United Arab Emirates	1.9-2.4
Oman	5.3-6.2
Yemen	0.95

*** Numbers are best estimates available at the time of writing of this manuscript. Modified from Teebi & Farag (1997) and Daar *et al.* (2000) and Teebi & Teebi (2005).**

as well as geography. Other explanations can refer to the more benign clinical course of SCD in this haplotype as opposed to the others, and the recency of events that could have brought the Benin or Bantu haplotypes, *i.e.* The Arab slave trade, which is predated by the interaction with the Persian/Indian culture and by a large margin.

- In the Mediterranean Arab countries, as well as North African Arab countries and the eastern province of Saudi Arabia , the Benin haplotype is the prevalent one. Proximity to the seminal foci of the mutation and trans-Saharan trade routes can explain this fact in North African Arab countries such as Morocco, Tunisia, Algeria and Egypt. As for Mediterranean Arab countries such as Lebanon, Syria and Jordan, several processes could have been conducive to the current state of affairs. In 1992, Oner *et al.* remarked that nearly all

Hb SS patients in the Mediterranean basin carried the Benin haplotype. He concluded that this chromosome carrying the sickle mutation was introduced from central West Africa via Algeria, Tunisia and Egypt. According to Adekile (1997), the desertification of substantial areas of Africa circa 4000-3000 B.C. drove many migrations out of Africa and into many destinations, one of which was the Western Arabian Coast, which was later occupied by the Semites, along with the Benin sickle chromosome. However, since this area was not markedly endemic with malaria, it is unlikely that this migration could be solely responsible for the frequencies observed today, since genetic drift and lack of positive selection would favor the dilution or even the extinction of this gene in this area . One might argue in favor of an Arabian-Peninsula origin of the sickle gene in the west Arabian Coast, driven by the Islamic expansion in the 7th century. However, that would imply that the Arab-Indian haplotype would be the predominant one in the Jordan, Syria and Lebanon, since it is the most frequent in the Arabian Gulf. Again this is negated by the prevalence of Benin haplotype in West Arabia. For an explanation to be valid, it has to answer for the prevalence of the Benin haplotype, and be of enough recency and demographic impact to explain the current frequencies observed. Recent historical events in the region might provide some clues for the answer. For instance, the recent import of great numbers of North African slaves by the Franks, Venetians, Mameluks and Ottomans between the 10th and the 18th centuries (AKSOY, 1961; Boussiou *et al.*, 1991; Adekile, 1992) might have resulted in foci of β^S-Benin carriers in those empires, only to merge later with the

indigenous populations of those Mediterranean countries, expanding through endogamy and positive malarial selection (See Figure 1.3). All the latter civilizations had a well documented and extended presence in the Mediterranean Arab countries that stretched for centuries, and could have therefore played carrier to the sickle gene into western Arabia. Other researchers have sought explanation in the practice of slavery in the Arab and Muslim world, as explained hereinabove.

1.4.3 Demography of the Sickle Gene in Lebanon

In Lebanon, the overall incidence of sickle cell gene was estimated at 0.34%. However, this figure is easily surpassed in southern-Lebanese villages where a figure of 24% is encountered ((Teebi & Farag, 1997; Inati *et al.*, 2007).

In an extensive, nationwide study, Inati *et al.* reported SCD to be clustered in two geographical foci, namely the southern and northern coastal areas (Inati *et al.*, 2007). Nearly all of the 387 SCD patients in the study were Muslims (66.5% Sunnis and 25% Shiite). Only two patients were Christians, and one of those was of Turkish descent. 56.3% of the patients were the progeny of consanguineous marriages. A previous study by the same author investigated β-globin haplotypes in the chromosomes of 50 Lebanese SCD patients and found Benin to be the dominant one (73% of chromosomes), followed by Bantu (15%), Arab-Indian (10%) and Senegal (2%).

This prevalence of SCD among Lebanese Muslims (and Sunnis in particular) can be ascribed to several factors: Firstly, consanguineous marriages are preferable and abound within the Muslim communities. This results in elevated frequencies of

individuals carrying or affected by genetic anomalies (refer to section 1.4.1). Second, the Muslim populations in Lebanon were historically bound to coastal areas (as opposed to Maronites, the major Christian sect in Lebanon, who historically took garrison in the Lebanese mountain chains). As such, malaria exerted a much bigger selective role on Muslim communities. And third, the larger average sibship size in Muslim communities might have played a role in shaping the sickle gene frequency and protecting it against genetic dilution.

1.5 Y-Chromosome Polymorphisms as Phylogeographical and Historical Investigative Tools

The ever growing body of information available on Y-chromosome polymorphisms, as well as the fact that it is passed on patrilineally, and that most of this chromosome is a non-recombining region (NRY) whose diversity is more or less solely generated by random mutational events makes it ideal for phylogenetic/evolutionary and historical or phylogeographical studies.

A Y-haplogroup is a haplotype defined by a set of biallelic markers (which are unique event polymorphisms or UEPs. *e.g.* Single-nucleotide polymorphisms or SNPs). Therefore, a Y-haplogroup originates when a new binary-marker mutation occurs (Jobling & Tyler-Smith, 2003). On the other hand, a different kind of Y-chromosome haplotypes can be defined by relying on microsatellites *i.e.* short tandem repeats (STRs). Since Y-haplogroups are defined by UEPs, they are more stable over time (exhibit lower mutation rates), and therefore are used as deep ances-

tral markers, whereas the higher mutation rates of STRs makes STR-haplotypes a more dynamic entity, and therefore endows it with a higher discriminatory capacity.

The Y Chromosome Consortium (YCC) is a consortium entrusted with coordinating the influx of information available from different research groups involved with this field, as well as creating and mainting a unified nomenclature system for Y-haplogroups, while accommodating for new expansions or haplogroups when discovered.

The YCC nomenclature system has defined 18 main haplogroups, named A through R (capital letters). Each of these major haplogroups (also called clades) can encompass subgroups (or subclades) numbered numerically *e.g.* E haplogroup has 3 subgroups: E1, E2 and E3 each defined by their own marker. There is also subgroup E*, defined as any subclade belonging to E but lacking the markers defining E1, E2 or E3. Sub-subgroups are labeled with lower case characters, such as E3a or E3b (Jobling & Tyler-Smith, 2003).I

In 1992, Lutz Roewer described the first polymorphic Y-chromosome marker, now known as the STR locus DYS19 (Butler, 2003). In 1997, the European forensic community selected a core set of Y-STR markers. This "minimal haplotype" included the following STR loci: DYS19, DYS389I/II, DYS390, DYS391, DYS392, DYS393, and DYS385 a/b with YCAII a/b as an optional marker (See Table 1.2. A large proportion of data available to date were generated with these loci. In early 2003, the U.S. Scientific Working Group on DNA Analysis Methods (SWGDAM) selected a core set of markers that included the 9 markers in the minimal haplotype, in addition to markers DYS438 and DYS439. Several commercially released

Y-STR profiling kits included this core set. As of today, commercial kits are available that can generate a profile for 17 or more STR loci in a single, multiplex PCR reaction, such as the PowerPlex-Y kit (Promega Corporation, Madison, WI) and the Yfiler kit (Applied Biosystems, Foster City, CA). PowerPlex Y examines 12 Y-STRs: DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, DYS439, and DYS385 a/b. Y-filer types 17 Y-STRs: DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392.

In addition, online electronic databases house thousands of Y-STR profiles from various populations, and are freely searchable. This renders the study of Y-chromosome polymorphisms an invaluable tool in the field.

Table 1.2: Date of discovery of Y-STR markers since 1992. Certain markers exist in several copies (multi-copy markers) and are indicated with "a/b" designation. From Butler (1998)

Year	Number of Available Markers (with multicopy)	Markers
1992	1	DYS19
1994	5 (8)	YCAI a/b, YCAII a/b, YCAIII a/b (DYS413), DXYS156
1996	11 (14)	DYS389I/II, DYS390, DYS391, DYS392, DYS393
1996	14 (17)	DYF371, DYS425, DYS426
1997	16 (19)	DYS288, DYS388
1998	17 (21)	DYS385 a/b
1999	22 (26)	A7.1 (DYS460), A7.2 (DYS461), A10, C4, H4
2000	28 (32)	DYS434, DYS435, DYS436, DYS437, DYS438, DYS439
2001	30 (34)	DYS441, DYS442
2002	33 (37)	DYS443, DYS444, DYS445
2002	34 (38)	DYS462
2002	48 (56)	DYS446, DYS447, DYS448, DYS449, DYS450, DYS452, DYS453, DYS454, DYS455, DYS456, DYS458, DYS459 a/b, DYS463, DYS464 a/b/c/d
2002	177	DYS468-DYS596 (+129)
2003	227	DYS597-DYS645 (+50)

1.6 Thesis Objectives

Since SCD in Lebanon was shown to be prevalent in a specific religious group characterized by high rates of consanguinity, and since the disease clustered in two well-defined geographical enclaves, we were interested in investigating whether his demographic association with SCD could be translated on the molecular level into an association/correlation between sickle-carrying chromosomes 11 and specific chromo-

some Y lineages. Such an association would be very dispersed or even non-existent in communities where endogamy, large sibship size and frequent consanguineous marriages are not a common feature.

In more detail, we wanted to investigate whether the sickle chromosomes 11 would be associated with certain Y-chromosomal haplogroups, or Y-STR haplotypes). If such an association was indeed present, we would then apply described molecular descriptive and comparative techniques to try to determine the date, origin and modality through which the sickle cell gene was introduced into the Lebanese population.

Chapter 2

Materials and Methods

2.1 Patients

DNA samples from 36 anonymized SCD patients were used for our dataset. This dataset comprised 23 SCD patients (β^s homozygotes), 9 Sickle-trait carriers (heterozygotes) and 4 sickle-beta thalassemia (compound hterozygotes). The presence of Hb S was confirmed via acid electrophoresis. Characterization of the polymorphisms of β-globin gene was performed by polymerase chain reaction (PCR) amplification of the 5' end of the β-globin gene followed by digestion with *Bsu36I* restriction endonuclease and DNA sequencing. In total, 59 sickle chromosomes were available for analysis.

2.2 Genotyping

2.2.1 Y-Haplogroup Determination

Y-SNP genotyping for haplogroup determination was performed using a custom TaqMan SNP genotyping assay from Applied Biosystems (Applied Biosystems, Foster City, CA, USA). Figure 2.1 illustrates the principle behind the TaqMan Real-Time PCR assay for allele calling. TaqMan probes are oligonucleotide probes having a fluorophore dye (*e.g.* tetrachlorofluorescin or FAM) on the 5' end, and a quencher attached at the 3' end. The role of the quencher is to "quell" the fluorescence emitted by the fluorophore when the latter is excited by the light-cycler's light source. This quenching occurs via Fluorescence Resonance Energy Transfer or FRET (Bustin, 2000; Bustin, 2005), as long as the reporter (fluorophore) and the quencher are within a threshold distance from each other (i.e. tethered to the ends of the probe). TaqMan probes are designed such that they anneal within a DNA region amplified by a specific set of primers. Usually the probed area spans a polymorphism (such as an SNP). Two different probes can be designed to match both alleles (ancestral and mutated) at that site, with each probe carrying a different fluor. As the TaqMan polymerase extends the primer and synthesizes the new strand, the 5' to 3' exonuclease activity of the polymerase degrades any probe annealed to the template. This releases the fluorophore, thus relieving the quenching effect and allowing fluorescence of the fluorophore since the two are no longer in close proximity (Bustin, 2000; Bustin, 2005). Fluorescence detected in the real-time PCR thermal cycler is directly proportional to the amount of fluorophore released and the amount of DNA template present in the PCR. The fluorescence spectrum at the end of amplification

allows to determine the type of allele present in the template. If the template is diploid, this also allows to determine homozygosity or heterozgosity.

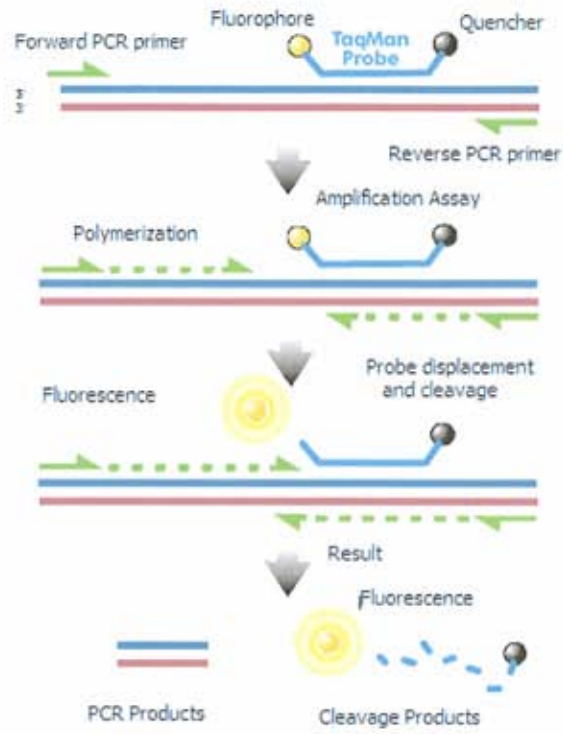


Figure 2.1: TaqMan assay principle. Two probes, each complementary to a different allele can be used simultaneously for allelic calling (not shown).

To determine the Y-haplogroup, (UEPs) were genotyped for each sample. This set defines all the Y-haplogroups encounered in the Lebanese population. Figure 2.2 shows the biallelic markers used as well as the Lebanese Y-haplogroups.

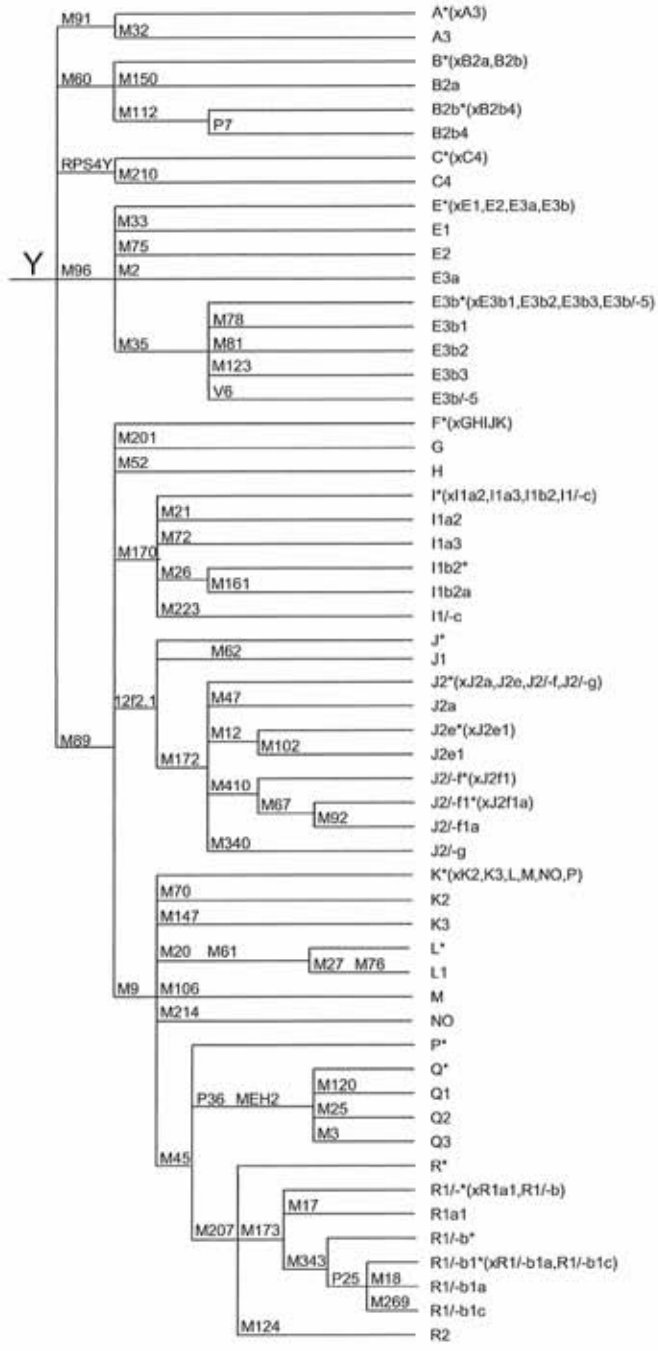


Figure 2.2: Y-Haplogroups identified in the Lebanese population. Biallelic markers' designation is shown above clade line, whereas haplogroup names are shown at the end of each clade line. based on the 2003 YCC tree with departures (Jobling & Tyler-Smith, 2003).

Genotyping was performed using a custom TaqMan SNP Genotyping Assay, using a 7900HT Real-Time PCR light-cycler (both from Applied Biosystems).

Each machine run accomodated 96 PCR reactions (25 μl reaction volume), using a MicroAmp™ Fast Optical 96-Well Reaction Plate (ABI, P/N 4366932). For each PCR reaction, we used:

- 12.5 μl TaqMan Genotyping Master Mix (as PCR mastermix, ABI P/N 4371355)
- 1.25 μl Custom TaqMan SNP Genotyping Assays (primers and biallelic markers detection probes, VIC and FAM fluorophores, ABI P/N 4351379)
- 11.25 μl of genomic DNA in PCR-grade water (10 ng).

The following thermal cycling conditions were used, according to manuacturer's instructions:

Step	**Temp (°C)**	**Duration**	**Cycles**
AmpliTaq Gold Enzyme Activation	95	10 min	Hold
Denature	95	15 sec	40
Anneal / Extend	60	1 min	

Pre- and Post-amplification fluorescence readings were taken, and used for Allele calling on SDS ver2.3 software.

2.2.2 Y-STR Haplotype Determination

Y-STR haplotypes were determined using the the AmpFlSTR Yfiler PCR Amplification Kit (Applied Biosystems). This is a short tandem repeat (STR) multiplex assay that amplifies 17 Y-STR loci in a single PCR reaction using fluorescent primers.

Those fluorescently-labeled PCR fragments can then be separated via capillary electrophoresis, and haplotypes can be assigned after inferring the repeat number of electrophoretic fragments.

The Yfiler kit amplifies the following loci:

- European minimal haplotype: DYS19, DYS385a/b, DyS389I/II, DYS390, DyS391, DyS392, DYS393.
- Scientific Working Group-DNA Analysis Methods (SWGDAM)-recommended Y-STR panel: Consists of the European minimal haplotype plus DY5438 and DY5439.
- Additional highly polymorphic loci: DYS437, DYS448, DYS456, DYS458, DYS635 and Y GATA H4.

Since DYS389I is embedded in DYS389II, we substracted it from the latter, to obtain DYS389b, whose value was used with the remaining allele values in the calculation of genetic distance between different samples.

Multiplex PCR Amplification The PCR reactions were run on a GeneAmp PCR System 9700 thermal cycler (Applied Biosystems). Reaction volume was 25 μl total, as follows:

Component	Volume in μl
AmpFlSTR Yfiler PCR Reaction Mix	9.2
AmpFlSTR Yfiler Kit Primer Set	5.0
AmpliTaq Gold DNA Polymerase	0.8
Sample or Control DNA (0.1 ng/μl) in PCR-Grade Water	10.0

Control DNA 007 (Bundled with YFiler kit) was used as positive control provided with the Yfiler kit, to assess the efficiency of the amplification step and STR genotyping.

Batches of 96 PCR reactions were run using MicroAmp Optical 96-Well Reaction Plates (Applied Biosystems). The following thermal cycling conditions were used, according to manuacturer's instructions:

	30 Cycles				
Enzyme Activation	**Denature**	**Anneal**	**Extend**	**Final Extension**	**Final Hold**
HOLD	CYCLE			HOLD	HOLD
95 °C	94 °C	61 °C	72 °C	60 °C	4 °C
11 min	1 min	1 min	1 min	80 min	∞

Electrophoresis and Fragment Analysis Electrophoresis was performed on an ABI Prism 3130*xl* Genetic Analyzer capillary electrophoresis workstation. Following is a panel of standards used for PCR product base pair sizing and genotyping (all bundled with the YFiler kit) :

- GeneScan-500 LIZ Size Standard: For assessing base pair sizing results and use as an internal lane standard. GeneScan-500 LIZ Size Standard is designed to size DNA fragments in the 35-500 bp range.

- Yfiler Kit Allelic Ladder: To accurately characterize the alleles amplified by the AmpFlSTR Yfiler kit multiplex PCR assay. It contains the majority of alleles reported for the 17 STR loci to aid in allele calling.

Each batch run on the 3130xl Genetic Analyzer consisted of 96 samples (inlcuding 1 Control DNA 007 sample used as a positive control) that were run using a 96-well optical plate after denaturation of double-stranded genomic DNA.

The plate loading mix is as follows for each reaction:

Reagent	**Volume per reaction (μl)**
Hi-Di Formamide	8.5
GeneScan 500 LIZ Size Std	0.5
PCR product (or	1
Total	10.0

The palte was placed into an Place plate into an BI 9700 Thermalcycler with a 96-well head with compression pads. Denaturing was accomplished using the following cycle parameters:

Temperature	**Time**
95 °C	5 min
4 °C	∞

Next, the optical plates with denatured PCR products were loaded into the Genetic Analyzer. Fragment analysis was performed using the bundled GeneMaper v4.0 software. Haplotype assignement for each sample was conducted with reference

to the allelic ladder provided.

2.3 Statistical Analysis

χ^2 test and the Fisher test (with one-tailed and two-tailed *p-value*) were performed using the SPSS software version 14.0 (Inc, 2001). a Pairwise genetic distances matrix for STR loci between samples was constructed using the software GENALEX version 6.2 (Peakall & Smouse, 2006). Principal Component Analysis (PCA) was performed using GENALEX with the genetic distance matrix generated earlier as input. Analysis Of Molecular Variance AMOVA (Excoffier *et al.*, 1992) was performed using the package ARLEQUIN version 3.3 (Excoffier *et al.*, 2005). Input for ARLEQUIN was the STR haplotypes constructed with 17 loci. AMOVA was also performed; ϕ_{ST} as an estimator of the mean genetic distance between the two populations was computed after linearization according to Slatkin's transformation (Slatkin, 1995). To test for significance, 10000 permutations were performed.

To calculate time to most common recent ancestor or TMRCA, we used the Walsh method ((Walsh, 2001)), using his constructed TMRCA calculator

(http://nitro.biosci.arizona.edu/ftDNA/TMRCA.html).

The Y chromosome haplotype reference database (YHRD) online search tool (Willuweit *et al.*, 2007) was used whenever the goeographical distribution of a certain haplotype was needed.

Chapter 3

Results

Table 3.1 shows the Y-haplogroup of the genotyped subjects as well as the β-globin mutation(s) carried. Out a total of 36 patients total, 23 were SCD patients (β^s homozygotes), 9 were sickle-trait carriers (heterozygotes) and 4 sickle/β-thalassemia compound heterozygotes. Two compound heterozygotes carried the codon 8 (CD8) β-thalassemia mutation, while 2 others carried the inversion at codon 110 β-thalassemia mutation (IVSI-110). 14 patients carried the J1 haplogroup (38.9%), 8 were E3b (19.44%), 6 carried the J2 haplogroup (16.65%), 4 were R1b (11.11%), 3 were K2 (8.33%) and 1 patient carried the N haplogroup (2.76%).

The total number of β^s-chromosomes per each haplogroup is illustrated in table 3.2.

Since we had estimates of the prevalence of each of the relevant Y-haplogroups in the Lebanese population (Zalloua *et al.*, 2008b), we compared the actual distribution of the β^s chromosomes over the different haplogroups with the expected distribution of 59 chromosomes (if they were to be sampled at random from the Lebanese population) (Table 3.3; figure 3.2). χ^2 test was first used for significance, and returned

Table 3.1: Y-Haplogroups and sickle state per patient.

Sample	Y-Haplogroup	Mutation
4H89	E3b	IVSI-110/β^S *
9W14	E3b	IVSI-110/β^S *
11P110	E3b	β^S/-
12S42	E3b	β^S/-
1D121	E3b	β^S/-
3D54	E3b	β^S/-
7S167	E3b	β^S/-
5S59	J1	β^S/-
14S116	J1	β^S / β^S
18P110	J1	β^S / β^S
19P110	J1	β^S / β^S
1H89	J1	β^S / β^S
1S144	J1	β^S / β^S
2D17	J1	β^S / β^S
4P110	J1	β^S / β^S
4S156	J1	β^S / β^S
6P175	J1	β^S / β^S
6S156	J1	β^S / β^S
7P110	J1	β^S / β^S
8S167	J1	β^S / β^S
8S59	J1	β^S / β^S
1S59	J2	β^S / β^S
5S89	J2	β^S / β^S
1AK140	K2	β^S / β^S
3S144	K2	β^S / β^S
6P110	N	β^S / β^S
2AK147	R1b	β^S / β^S
4B119	R1b	β^S / β^S
7S107	R1b	β^S / β^S
9S167	R1b	β^S / β^S
15S188	J2	β^S/-
4S42	J2	β^S/-
2S144	K2	β^S/-
2S59	J2	CD8/β^S *
5D121	J2	CD8/β^S *
2P110	E3b	β^S / β^S

*Indicates a compound β-Thalassemia/β^S heterozygote.

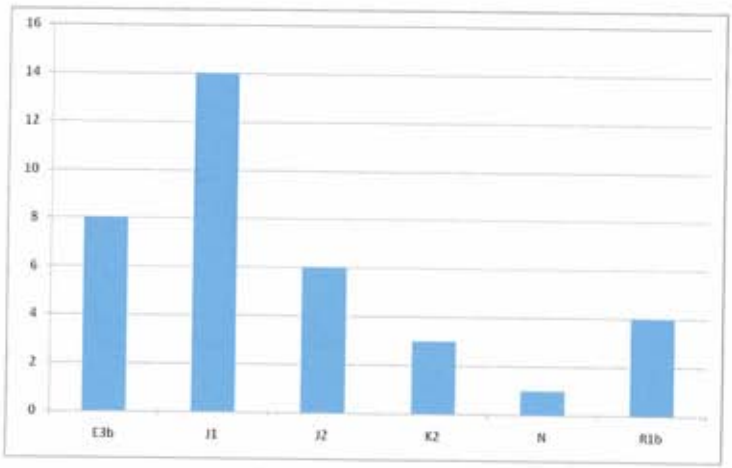


Figure 3.1: Patients + carriers per haplogroup

Table 3.2: Number of βs Chromosomes per haplogroup.

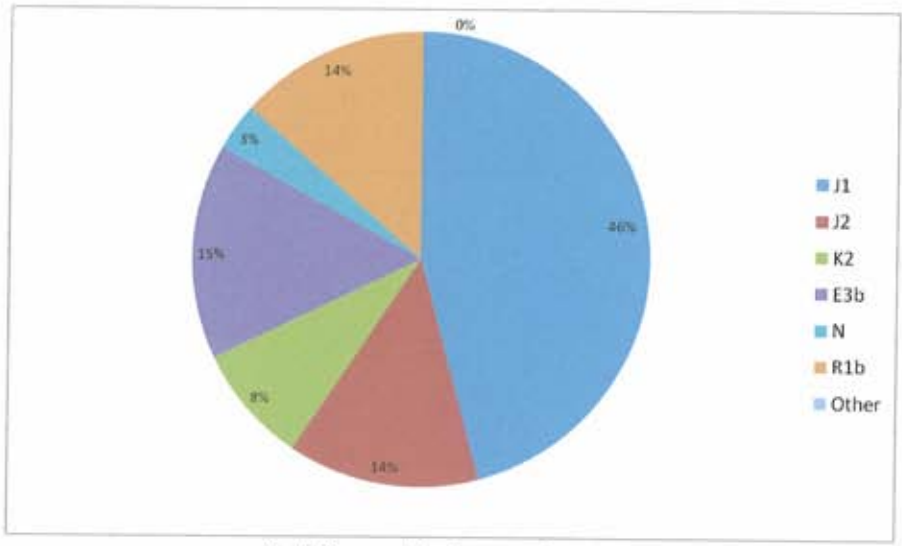
Y-Haplogroup	No. βs Chromosomes	Frequency βs
J1	27	45.76 %
J2	8	13.56 %
K2	5	8.47 %
E3b	9	15.25 %
N	2	3.39 %
R1b	8	13.56 %

Table 3.3: Number of βs Chromosomes per haplogroup. $p < 0.001$

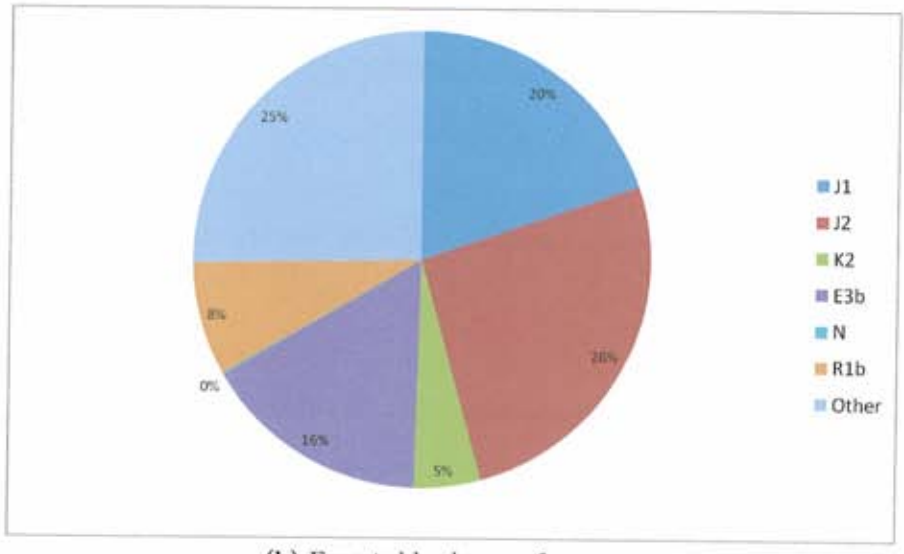
Y-Haplogroup	No. of βs Chromosomes	Freq. of βs Chromosomes	Expected No. (Normal population)	Expected Freq (Normal populati
J1	27	45.76 %	11.6	20 %
J2	8	13.56 %	15.34	26 %
K2	5	8.47 %	2.95	5 %
E3b	9	15.25 %	9.44	16 %
N	2	3.39 %	0.064	0 %
R1b	8	13.56 %	4.72	8 %
Other	-	-	14.886	25 %
Total	**59**	**100**	**59**	**100**

a *P-value* < 0.0001. This was foreseeable, since there is an obvious deviation from the expected frequencies, especially when examining the observed-to-expected ratio of chromosomes belonging to the J1 Y-haplogroup (figure 3.2).

Given the presence of several observations (haplogroups) at low numbers (<5), we performed a Fisher exact test, again to test for significance. Fisher exact test is better suited for tables with one or several sparsely populated cells. We were most interested in testing the significance of the association of β^s chromosomes with the J1 haplogroup, since it represented the most prominent deviation from expected



(a) Observed haplogroup frequency



(b) Expected haplogroup frequency

Figure 3.2: Observed versus expected haplogroup frequencies for β^8 chromosomes

Table 3.4: Fisher exact test for significance of association with J1 haplogroup with β^s.

	observed	expected
J1	27	12
Non-J1	32	47

p-value (one-tailed) = *0.002*
p-value (two-tailed) = *0.005*

results. For this purpose, we opted for a dichotomous Fisher exact test, which is a 2x2 matrix. All observations were pooled adequately to fit under two column categories: [observed] and [expected], whereas row categories were defined as [J1] and [non J1] . Expected frequency was rounded off to the nearest integer as required for this statistical test. Again, the null-hypothesis was clearly rejectable, with a one-tailed *p-value* of 0.002 and a two-tailed *p-value* of 0.005, both less than 0.05 (refer to table 3.4). STR haplotypes for the 17 Loci typed are shown in table **??**. Upon closer inspection, the STR haplotypes of patients carrying the J1 haplogroup showed little divergence.

To formally investigate this observation, it was first necessary to determine whether this Y-STR haplotype clustering was a feature that distinguished the sickle J1 subjects exclusively, or if it was explainable by the amount of variation inherent in the Lebanese J1 subjects taken as a whole. To quantitate this proposal, we sought to test our J1 sicklers set against a control set of 81 Lebanese individuals carrying the J1 haplogroup.

Pairwise Genetic distances were calculated for all samples (14 J1 sicklers) and controls (81 Lebanese J1) using the R_{ST}-like sum of squared differences between loci. A potential caveat here is the possibility of having sicklers amongst the control

Table 3.5: 17 loci Y-STR haplotypes versus haplogroup and sickle-mutation. Haplotype defined by Alleles (repeat numbers) at each locus (DYS numbers)

Sample	Haplogroup	Mutation	DYS 389I	DYS 389II	DYS 390	DYS 456	DYS 19	DYS 385a	DYS 385b	DYS 458	DYS 437	DYS 438	DYS 448	GATA_ H4	DYS 391	DYS 392	DYS 393	DYS 439	DYS 635
4H89	E3b	IVSI-110/βs	13	30	23	15	13	17.1	18	16	14	10	20	12	10	11	13	12	20
9W14	E3b	IVSI-110/βs	13	29	24	16	13	13	15	17	14	10	20	12	9	11	13	10	22
11P110	E3b	βs/-	13	29	24	16	13	13	14	18	14	10	20	12	9	11	13	10	21
12S42	E3b	βs/-	13	30	24	16	14	18	19	16	14	10	19	11	11	11	13	12	20
1D121	E3b	βs/-	13	30	23	15	13	17	18	16	14	10	20	12	10	11	13	12	20
3D54	E3b	βs/-	12	29	21	16	13	15	16	18	14	10	21	12	9	11	12	11	20
7S167	E3b	βs/-	14	31	24	15	14	18	18	16	14	10	20	12	10	11	13	13	20
2P110	E3b	βs/βs	12	29	21	17	13	15	17	18	14	10	21	12	9	11	12	11	20
5S59	J1	βs/-	13	30	23	14	14	13	19	19.2	14	9	20	11	11	11	12	11	21
14S116	J1	βs/βs	13	30	23	14	15	13	19	19.2	14	10	20	11	11	11	12	11	21
18P110	J1	βs/βs	14	31	23	14	14	13	18	17.2	14	10	20	11	11	11	12	11	21
19P110	J1	βs/βs	13	29	23	14	14	13	18	20	14	10	20	11	9	11	12	11	21
1H89	J1	βs/βs	13	31	23	14	14	13	20	18.2	14	10	20	11	11	11	12	11	21
1S144	J1	βs/βs	13	29	23	15	14	13	18	19.2	14	10	20	11	11	11	12	12	21
2D17	J1	βs/βs	13	29	23	16	13	12	19	19.2	14	10	21	11	10	11	12	11	21
4P110	J1	βs/βs	13	29	23	14	14	13	18	18.2	14	10	20	11	9	11	12	11	21
4S156	J1	βs/βs	13	31	23	13	14	13	19	18.2	14	10	20	11	11	11	12	11	21
6P175	J1	βs/βs	13	30	23	14	14	13	19	19.2	14	9	20	11	11	11	12	11	21
6S156	J1	βs/βs	14	31	23	16	14	12	18	19.2	14	10	21	10	10	11	12	11	22
7P110	J1	βs/βs	13	29	23	16	13	12	19	19.2	14	10	21	11	10	11	12	11	22
8S167	J1	βs/βs	13	30	23	14	14	13	19	18.2	14	10	20	11	10	11	12	11	21
8S59	J1	βs/βs	13	30	23	15	15	13	18	18.2	14	10	20	11	11	11	12	11	21
15S188	J2	βs/-	14	31	25	14	15	15	18	19	15	9	19	11	10	11	12	11	21
4S42	J2	βs/-	13	30	22	15	14	13	15	16	15	9	21	11	10	11	13	10	21
2S59	J2	βs/CD8	14	29	22	16	14	13	20	15	15	9	21	11	9	11	12	11	21
5D121	J2	βs/CD8	13	29	23	16	15	13	16	14	14	9	21	12	10	11	14	12	22
1S59	J2	βs/βs	13	29	25	15	14	18	18	19.2	14	10	19	11	10	11	13	12	21
5S89	J2	βs/βs	14	30	24	15	14	13	16	14	15	9	20	11	10	11	12	12	21
2S144	K2	βs/-	14	31	23	15	13	15	15	16	15	9	19	12	10	13	13	11	20
1AK140	K2	βs/βs	14	31	23	15	13	15	15	16	15	9	19	12	10	13	13	11	20
3S144	K2	βs/βs	14	31	23	15	13	15	19	16	15	9	19	12	10	13	13	11	20
6P110	N	βs/βs	14	30	23	15	14	11	13	18	14	11	19	11	11	16	14	10	22
2AK147	R1b	βs/βs	14	30	24	14	14	11	15	17	15	12	19	13	11	13	13	12	23
4B119	R1b	βs/βs	14	30	24	15	15	13	17.1	16	14	12	19	11	10	13	13	13	23
7S107	R1b	βs/βs	14	30	24	15	15	13	17	16	14	12	19	11	10	13	13	13	23
9S167	R1b	βs/βs	14	30	24	15	15	13	16	16	14	12	19	12	11	13	13	12	23

J1 population, which would inevitably tend to decrease genetic difference amongst the two sub-populations. The calculation was performed assuming a Step-Wise Mutation Model (SMM) instead of the infinite-allele model. In the SMM, it is assumed that every mutation event can only vary the mutated allele by increasing or decreasing the copy number by exactly one. Therefore, if two individuals had a difference of two copy numbers at a given locus, this is scored as 2 mutational events.

To represent the 81x14 genetic distance matrix obtained in human-interpretable form, principal component analysis (PCA) was performed. PCA is a mathematical, multi-dimensional reduction technique that can be used - without any need for prior modeling - to reduce dimensions of variations in a k-dimensional space (such as one defined by a pairwise genetic distance matrix generated earlier) into two-dimensional major axes of variation that can be used to detect any inherent stratification encoded by the data.

AMOVA, or analysis of molecular variance was also performed. ϕ_{ST}, an F_{ST} analogue, is conventionally defined as the ratio of the estimated variance component due to differences among P populations (σ_a^2) , divided by the estimated total variance (σ^2): $\phi_{ST} = \sigma_a^2/\sigma^2$, with the total variance defined as the sum of within- and among-populations variance $\sigma^2 = \sigma_a^2 + \sigma_w^2$. This estimator of the mean genetic distance between the two populations was computed after linearization according to Slatkin's transformation (Slatkin, 1995). To test for significance, 10000 permutations were performed, to obtain a *P-value* representing the probability of getting a similar or more extreme distribution via chance alone.

A PCA plot showing the two major axes of variation for the J1 sicklers and J1 control dataset is shown in figure 3.3.

As was partly discernable from Y-STR haplotypes, the PCA plot showed a clustered distribution of the J1 Sickle dataset within the larger space defined by Lebanese J1 carriers we used as controls. This was indicative of an inbreeding/consanguineous deme; this fact could be further supported by the otherwise unexplainable full homozygosity within this group (All J1 were SCD patients, not carrieres). This could also be indicative of a potential and recent founder effect.

Figure 3.4 shows a summary of AMOVA performed on the previous dataset. 6% of the total variance was due to differences among the two sub-populations (*p-value=0.002)*, which is significant taking into consideration that both test populations shared the same haplogroup.

To test whether clustering of the sickle J1 dataset could be further emphasized, we decided to repeat the PCA analysis using fewer loci, since a distance matrix based on 17 loci could be powerfully discriminating, even for somewhat tightly clustering groups (figure 3.5). The loci used were DYS19, DYS389I, DYS389b, DYS390, DYS391, DYS392, DYS393 and DYS385a/b.

The clustering of J1 sicklers became more obvious as the J1 control samples seemed to "spread" more using this PCA scheme, although one sickle sample seemed to be an outlier and possibly not descendant from the conjectured patrilineal founder of this group.

These observations warranted further investigation of the possible origin of the J1 sickler founder. For this purpose, a putative "modal" or ancestral haplotype was

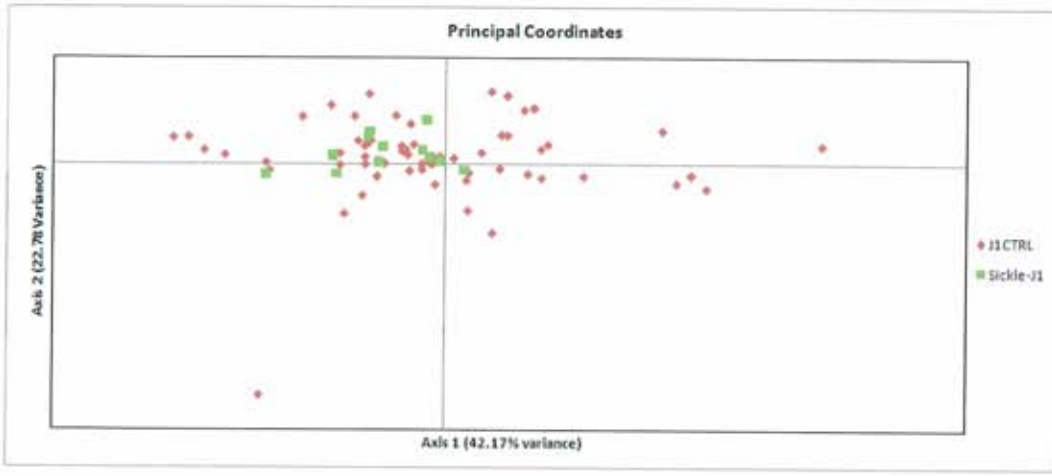
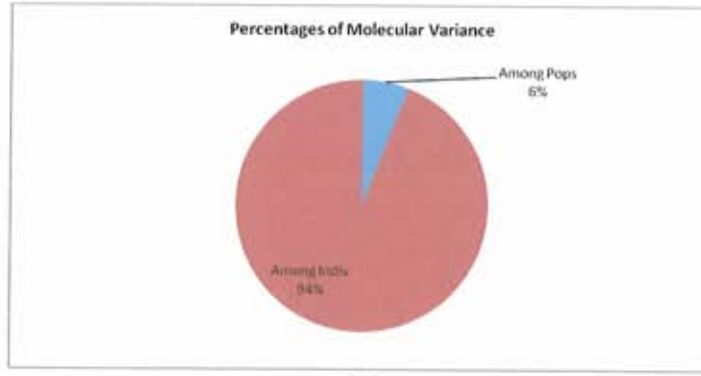


Figure 3.3: PCA plot showing axes of variation in a J1 sickle versus J1 control dataset. Plot input data was genetic distance between the two sets using all Y-STR loci. Amount of variance explained by each axis is also noted. Lebanese J1 Control dataset in red. Sickle J1 dataset in green.



(a)

Summary AMOVA Table		
Source	Est. Var.	%
Among Pops	0.430	$\phi_{ST} = 6\%$
Among Indiv	6.690	94%
Within Indiv	0.001	0%
Total	7.120	100%

P(rand>=data via 10000 permutations) = 0.002

(b)

Figure 3.4: AMOVA summary for the two datasets at 17 Y-STR loci. (a) Chart showing the apportionment of variance among individuals and between the two sub-populations. (b) Table summarizing variance, ϕ_{ST} and significance derived from 10000 permutations.

constructed using the "median" (*i.e.* the most occurring) allele at each loci. In an event of a tie, the larger allele was used. However, the outlying sickle-J1 sample detected via PCA was excluded from this process. Figure 3.6 shows the adopted modal haplotype, as well as the difference in mutational steps at each locus from sample haplotypes.

Next, the YHRD database (www.yhrd.org) was queried using this defined modal haplotype. Again, we decided to use a subset of the total 17 loci for this purpose, since the YHRD database housed almost twice as much 11-loci haplotypes as 17-loci haplotypes (46244 versus 17384 haplotypes, respectively), and since we wanted to query as many populations as possible. We included the same subset of 11 loci used for the PCA analysis shown in figure 3.5.

YHRD was queried for the modal haplotype. Figure 3.7 shows the geographical distribution and frequency of matching haplotypes. A clear geographical cline was easily discernable along the North-African coast, with matches in Italy. Unsurprisingly, the matches found in Italy were of Tunisian origin.

Finally, time to most recent common ancestor (TMRCA) was estimated for all pairs of the Sickle-J1 dataset, as well as between each sample and the modal haplotype. The Walsh method was applied as described in section 2.3. We selected results within a confidence interval of 50%. Generation time was assumed to be 25 years (figure 3.8).

A TMRCA estimation for a pair of haplotypes is fairly approximative, and the statistical confidence intervals for such estimations are very wide, especially because high confidence in the underlying mutation model for the markers is lacking. How-

ever, we only sought to determine the historical time-frame in which the ancestral haplotype was extant/introduced to the population and for this purpose the reliance on TMRCA is warranted.

Having said that, analyses of the TMRCA table could indicate that 8 out of 13 of the samples diverged from a common ancestor some 1000 year ago, whereas one sample showed an estimate of 600 years. Hence we can safely bracket our analysis to a time frame between 800 A.D and 1200 A.D.

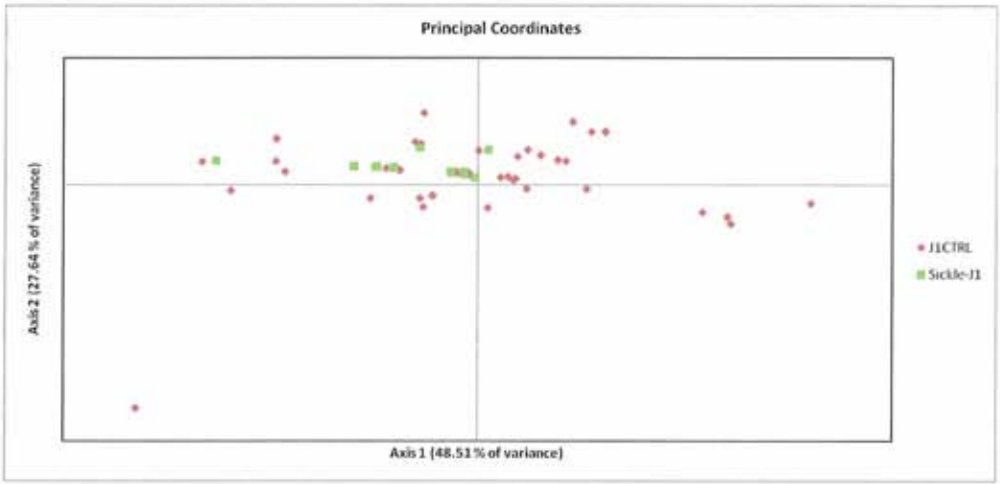
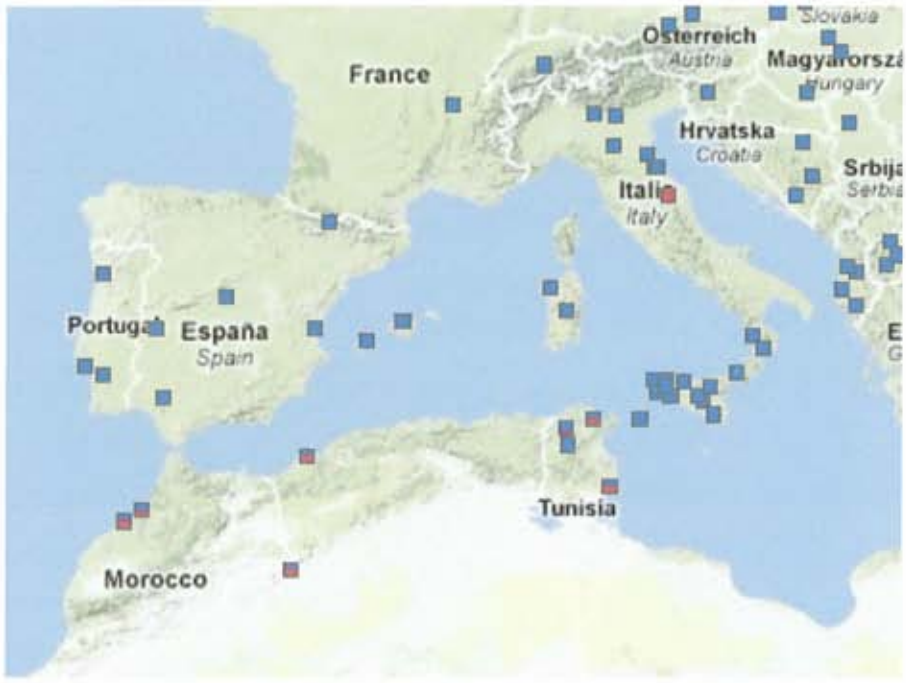


Figure 3.5: PCA plot showing axes of variation in a J1 sickle versus J1 control dataset. Plot input data was genetic distance between the two sets using a reduced (11) subset of the 17 Y-STR loci. Amount of variance explained by each axis is also noted. Lebanese J1 Control dataset in red. Sickle J1 dataset in green.

	ID	DYS19	DYS389I	DYS389II	DYS390	DYS391	DYS392	DYS393	DYS385	DYS438	DYS439
1	modal	14	13	30	23	11	11	12	13.19	10	11
2	J1Sickle	15	13	30	23	11	11	12	13.19	10	11
3	J1Sickle	14	14	31	23	11	11	12	13.18	10	11
4	J1Sickle	14	13	29	23	9	11	12	13.18	10	11
5	J1Sickle	14	13	31	23	11	11	12	13.20	10	11
6	J1Sickle	14	13	29	23	11	11	12	13.18	10	12
7	J1Sickle	13	13	29	23	10	11	12	12.19	10	11
8	J1Sickle	14	13	29	23	9	11	12	13.18	10	11
9	J1Sickle	14	13	31	23	11	11	12	13.19	10	11
10	J1Sickle	14	13	30	23	11	11	12	13.19	9	11
11	J1Sickle	14	14	31	23	10	11	12	12.18	10	11
12	J1Sickle	13	13	29	23	10	11	12	12.19	10	11
13	J1Sickle	14	13	30	23	10	11	12	13.19	10	11
14	J1Sickle	14	13	30	23	11	11	12	13.19	9	11
15	J1Sickle	15	13	30	23	11	11	12	13.18	10	11
Distance from reference:						Zero	One	Two	Three+		

Figure 3.6: Modal haplotype defined from 11 loci. Deviations (mutational steps) at each locus are highlighted.



(a)

n of N	Geoposition [Population]	Metapopulation	Continent
12 of 61	Marche, Italy [Tunisian]	Afroeurasian - Semitic	Europe
5 of 166	Casablanca, Morocco [Arab]	Afroeurasian - Semitic	Africa
5 of 155	Sfax, Tunisia [Tunisian]	Afroeurasian - Semitic	Africa
4 of 102	Oran, Algeria [Arab]	Afroeurasian - Semitic	Africa
4 of 61	Marche, Italy [Moroccan]	Afroeurasian - Semitic	Europe
3 of 126	Kuwait [Kuwaiti]	Afroeurasian - Semitic	Asia
3 of 52	Figuig, Morocco [Berber]	Afroeurasian - Berber	Africa
3 of 68	Rabat, Morocco [Sahraouis]	Afroeurasian - Semitic	Africa
2 of 69	Rabat, Morocco [Berber]	Afroeurasian - Berber	Africa
2 of 54	Tunis, Tunisia [Tunisian]	Afroeurasian - Semitic	Africa
2 of 132	Tunisia [Andalusian Arab]	Afroeurasian - Semitic	Africa
2 of 114	Jordan [Arab-Qahtanit]	Afroeurasian - Semitic	Asia

(b)

Figure 3.7: YHRD query: Geographic/frequency distribution of modal-haplotype matches. Higher red-fill ratio of squares indicates higher number of matches. Blue squares indicate no matches found.

Time to Most Recent Common Ancestor (Years)															
	ID	modal	1J1Sickle	2J1Sickle	3J1Sickle	4J1Sickle	5J1Sickle	6J1Sickle	7J1Sickle	8J1Sickle	9J1Sickle	10J1Sickle	12J1Sickle	13J1Sickle	14J1Sickle
1	modal	15	1000	600	1000	1000	1450	3650	1000	1450	1000	3000	1000	1000	1000
2	1J1Sickle	1000	15	1450	1925	1450	2425	3000	1925	1450	1000	4375	1000	1000	1000
3	2J1Sickle	600	1450	15	1450	1450	1925	4375	1450	1925	1450	2425	1450	1450	1450
4	3J1Sickle	1000	1925	1450	15	1450	1450	3000	250	1925	1925	3650	1450	1925	1925
5	4J1Sickle	1000	1450	1450	1450	15	1925	3650	1450	1000	1450	4375	1450	1450	1925
6	5J1Sickle	1450	2425	1925	1450	1925	15	3650	1450	1925	2425	4375	2425	2425	1450
7	6J1Sickle	3650	3000	4375	3000	3650	3650	15	3000	3000	3650	2425	2425	3650	3650
8	7J1Sickle	1000	1925	1450	250	1450	1450	3000	15	1925	1925	3650	1450	1925	1925
9	8J1Sickle	1450	1450	1925	1925	1000	1925	3000	1925	15	1450	4375	1450	1450	1925
10	9J1Sickle	1000	1000	1450	1925	1450	2425	3650	1925	1450	15	4375	1000	250	1925
11	10J1Sickle	3000	4375	2425	3650	4375	4375	2425	3650	4375	4375	15	3000	4375	3650
12	12J1Sickle	1000	1000	1450	1450	1450	2425	2425	1450	1450	1000	3000	15	1000	1925
13	13J1Sickle	1000	1000	1450	1925	1450	2425	3650	1925	1450	250	4375	1000	15	1925
14	14J1Sickle	1000	1000	1450	1925	1925	1450	3650	1925	1925	1925	3650	1925	1925	15
0-225 Years		250-475 Years			500-725 Years			750-975 Years							

Figure 3.8: TMRCA between dataset- and modal- haplotype pairs. Corresponding mutation rates were used for different loci. The infinite-allele model was used, and a generation time of 25 years was assumed. Probability is 50% that the TMRCA is no longer than indicated.

Chapter 4

Discussion

The distribution of haplogroups within the sickle dataset differed significantly from that of the overall Lebanese population, and this was especially salient when considering the discrepancy in the representation of the J1 haplogroup. Care must be taken not to strictly interpret this as the proportion of haplogroups within the seminal conveyors of the sickle gene into the Lebanese population. What can be inferred however, and as further corroborated by the PCA analysis, is that we have stumbled upon a probable founder effect, in a deme apparently characterized by elevated rates of inbreeding and/or consanguinity, and carrying the J1 haplogroup. In fact, 13 of the 22 SCD homozygotes in this study were J1 haplogroup carriers, in further support of this observation.

This finding could be valuable for studying the origin/time of introduction of the sickle gene in Lebanon. First, it meant that a modal/ancestral haplotype could be constructed easily. Second, it also meant that this modal haplotype could have been one of the seminal disseminators of the sickle gene into present-day Lebanon. In other terms, it is very likely that a non-indigenous individual, carrying both the sickle

mutation and this ancestral Y-STR haplotype, initiated a founder effect for the sickle gene within part of the Lebanese populace. Recency as well as high consanguinity rates, paired with large sibship size could have protected the association between the mutation on chromosome 11 and the Y-Haplogroup J1 against the genetic "dilution" that is bound to occur with every generation.

As such, we queried the YHRD database for the constructed 11-loci modal haplotype. The highest number of matches came from Italy, and specifically in samples from Tunisian descent/ethnicity. The rest of the matching haplotypes were distributed along the North-African coast (namely Morocco, Tunisia and Algeria).

We can assume that the Lebanese descendants of this ancestral haplotype diverged from it some 1000 years ago. One can sensibly assume that this correlates to a historic event that took place around 800-1200 A.D. When combined, these observations, as well as the other findings we presented earlier, allow for an evidence-based hypothesis of the origin of the sickle gene in a large section of the Lebanese population.

The Genetic evidence we gathered, as well as the suspected timeframe led us to propose that the expansion of the Fatimid dynasty from North-Africa into the Levant was the demographic event mainly responsible of the introduction of SCD into Lebanon, as opposed to the traditional hypothesis referring to the practice of slave trade in the Arabian Peninsula as the event that brought SCD to Lebanon through migration.

Traditionally, the "Arab slave trade" was thought of as being the conveying agent of the sickle gene into the Arabian Peninsula through black African slaves, and thus

indirectly, by Arab migration, into Lebanon and the Levant. However, there are several indices precluding this practice from being a plausible explanation of the expansion of this gene in Lebanon. The sturdiest molecular counter-argument can be obtained by comparison of R1b haplogroup frequencies. R1b haplogroup is highly characteristic of western Europeans. In Italy for instance, it is prevalent at about 40% (Capelli *et al.*, 2007). In Lebanon, this haplogroup is present at differential frequencies between non-Muslims (11%) and Muslims (4.7%) (Zalloua *et al.*, 2008b). Our sickle dataset comprised R1b haplogroups at 14%, very different from the expected 4.7% assuming that most of our dataset are Muslims. As such, explaining R1b frequencies in our dataset in terms of Arab slave trade and Peninsulan Arabs settling in Lebanon seems like a very remote possibility, knowing that the R1b haplotype is only present at 2% in the Arabian Peninsula. From a historical and sociological perspective, the ratio of male-to-female African slaves brought to Arabia was 1 to 3 (Lewis & Lewis, 1990; Ewald, 1992)(Lewis & Lewis, 1990; Ewald, 1992). And while female slaves were destined for concubinage and domestic service/menial work, the majority of male slaves were castrated (eunuchs) and were destined to work as harem guards or at the service of mosques. Add to that the substantial mortality rates amongst slaves, and the condemnation/quasi-prohibition of marriages between slaves and free women (Lewis & Lewis, 1990), to conjecture that genetic admixture with slaves was not phenomenal. Moreover, the use of captured Africans as slaves was a relatively late practice in the slavery timeline, peaking only in the 18th and 19th century. Within the time-frame we are considering, slaves in Arabia came from sundry origins, but they were mostly Slavic Eastern Europeans ("Saqaliba"),

people from surrounding Mediterranean areas, Persians, Turks and peoples from the Caucasus mountain regions (such as Georgia, Armenia and Circassia) (Lewis & Lewis, 1990). It was only later, towards the 18th and 19th century, that slaves were increasingly Africans, captured mostly from East Africa. In addition, most of the globin haplotypes in the Arabian Peninsula are Arab-Indian, and not Benin or Bantu. The only exceptions to this are eastern Saudi Arabia (Benin haplotype, directly involved in African slave trade across the Red Sea), and Yemen and Oman (coastal enclaves of Benin and Bantu haplotypes; Benin is due to genetic influx from the eastern parts of the Arabian peninsula, whereas the presence of Bantu can be linked to close contacts with Mombasa and Zanzibar, modern day Tanzania through the Indian Ocean slave trade route (Daar *et al.*, 2000)). Combining all of the former facts vindicates our line of thought: If influx from the Arabian Peninsula was the major source of sickle gene in Lebanon, then the major haplotype globin haplotype would be the Arab-Indian, and not the Benin as is the case. Or at the least, we would have found matches for our modal sickle-J1 haplotype scattered around the Arabian Peninsula.

Our hypothesis based on The Fatimid expansion could provide a better explanation given the genetic and historic data presented so far. The Fatimid rule, or Fatimid caliphate, debuted in modern-day Tunisia and Algeria. The foundation of the dynasty occured in 909 By Abdullah al-Mahdi Billah. Soon after, the Fatimids were in control of most central Maghreb, which comprises modern-day Tunisia, Libya, Algeria and Morocco *i.e.* The North-African coast. However, in the late 900s, the Fatimids conquered Egypt, and in 969 they changed their capital from

Tunisia to Egypt, where they founded the city of "al-Qahira", better known as Cairo (Steindorff & Seele, 1957; Holt *et al.*, 1977; Ewald, 1992). Cairo became the newly-found capital of the Fatimid Caliph, the ruling Elite and the army. From Egypt, the Fatimids exerted an expansionist policy, conquering surrounding regions, until their dominion stretched from Tunisia to Syria, and even crossing into Sicily. At its zenith, the Fatimid empire centered in Egypt encompassed North Africa, Sicily, Palestine, Lebanon, Syria, the Red Sea coast of Africa, Yemen and the Hejaz, as shown in figure 4.1 (Steindorff & Seele, 1957; Beeson, 1969; Ewald, 1992; Wintle, 2003).

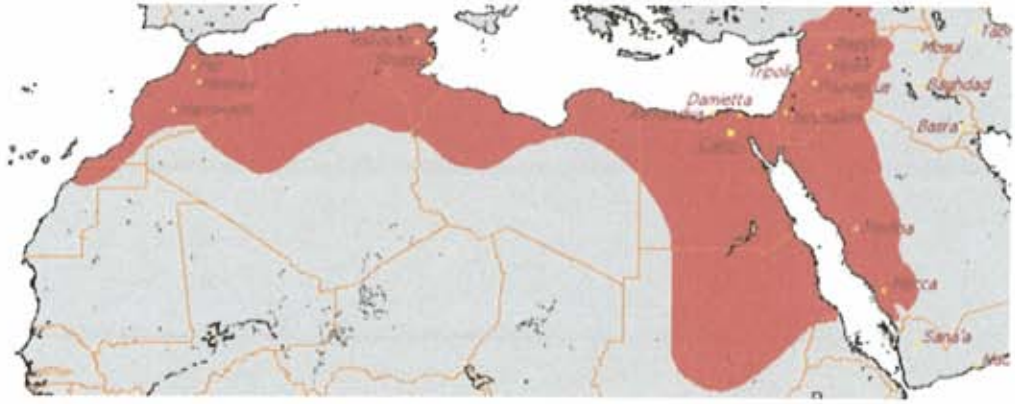


Figure 4.1: The Fatimid empire (909-1171 A.D.)

In addition to the fact that Lebanon came under the Fatimid rule, the Fatimid army comprised mainly North-African Berbers (and other African ethnicities, to a lesser extent). This is in total accordance with the geographical mapping of our modal haplotype. But to validate this proposal, it was imperative to establish a genetic link with Egypt, the capital of the Fatimids for more than 200 years. To achieve this, we decided to expand our query with even more haplotypes, by further reducing our modal Y-STR haplotype from 11 to 7 core loci. Performing another

search in YHRD yielded the expected result of finding additional haplotype matches in Egypt; suitably, matches with a lower frequency were also present in Syria (figure 4.2). These additional haplotype matches were not detected when we used 11-loci haplotypes.

Figure 4.2: Distribution of modal-haplotype matches for 7 core loci.

With this, one can confidently postulate that admixture occured between sickle-carrying Berber/North-African members of the invading Fatimid and certain Lebanese/Syrian communities/enclaves. This founder-effect-like process, coupled with high rates of consanguinity/inbreeding within the recipient population resulted in the clustered distribution of the sickle cell gene we are studying.

Other than the distribution of Y-STR haplotype matches, genetic evidence is clear in terms of haplogroup representation in the source (north-African) and target (Lebanese SCD) population (refer to Table 4.1). Other data useful for this comparison is a recent study of Y-haplogroup frequencies in Algeria showing that E3b was at more than 45%, whereas J1 was observed at 22.5% and R1b was present at 10.8% (Robino *et al.*, 2008). Similar frequencies are oberved in the remaining neighboring

Table 4.1: Dominant Haplogroup Frequencies in North Africa versus our SCD dataset

Population		N	Reference	ExE3	E3	F*	G	JxJ2	J2	K2	R*	R1+R2
	Algeria (Arabs)	35	Arredi et al., 2004		57	11		23	5.7			
	Algeria (Berbers)	20	Arredi et al., 2004		55	10		15			15.0	
North-African	Egypt	147	Luis et al., 2004	1.0	38		9	20	12	8.0		8.0
	Libya	20	Shen et al., 2004		30		10	10	40			10.0
	Morocco	20	Shen et al., 2004		20		30	10	20	10.0		10.0
	Tunisia	148	Arredi et al., 2004	0.6	50	4.7		33	3.4	0.7	6.1	0.7
Lebanese SCD (Patient Frequency)					19.44			38.9	16.65	8.33		11.11

North African countries.

The concordance in terms of haplogroup frequencies also suggests heavy genetic contribution to the Y-chromosome pool of Lebanese SCD patients from an invading force comprising peoples of the north African coast, from Morocco to Egypt, with Tunisia and Egypt probably being the top contributors. The over-representation of J1 in our dataset might be due to several reasons. A founder effect coupled with high consanguinity rates and large sibship size of the target population could provide explanation, although we cannot rule out the fact that our data was obtained from a clinical referral centre, where symptomatic homozygotes are more likely to show up. Indeed, most homozygotes were J1, whereas the highest number of heterozygotes were E3b in our dataset.

The fact that R1b individuals in our dataset seem to originate from north Africa as opposed to western Europe where R1b dominates is quite interesting. In fact, when we compared our R1b STR-haplotypes to Italian databases, we could not find a

n	19	389i	389ii	390	391	392	393	385
1	15	14	30	23	10	13	13	13,17

Figure 4.3: A 1-step neighbor of our R1b consensus haplotype in Tunisia

single match and the genetic distance was striking. However, a one-step neighbor to a consensus R1b-haplotype constructed from our dataset found a match in Tunisia as shown in Figure 4.3. It is therefore obvious that this is an African subclade (R1b1) of the R1b haplogroup.

It should be noted that during the conquest of the Levant, the Fatimids employed Turkic slaves/mercenaries to leverage against the trained Turkic archers of the rival Abbasids in Syria. This might be the root of the Anatolian CD8 thalassemia mutation detected in our dataset, as well as the IVSI-110 thalassemia mutations found in our compound heterozygotes, knowing that the latter mutation is the most common thalassemia mutation in Turkey (Keser *et al.*, 2004). Finally, this fact could explain the presence of the N haplogroup in our sickle dataset (3%) as opposed to being quasi-absent in the total Lebanese population. Figure 4.4shows haplogroup frequencies in the North African coast as well as in Europe, including Turkey.

The geographical clustering of SCD north and south of Lebanon provides further support for our dual hypothesis. The coast of Northern and Southern Lebanon

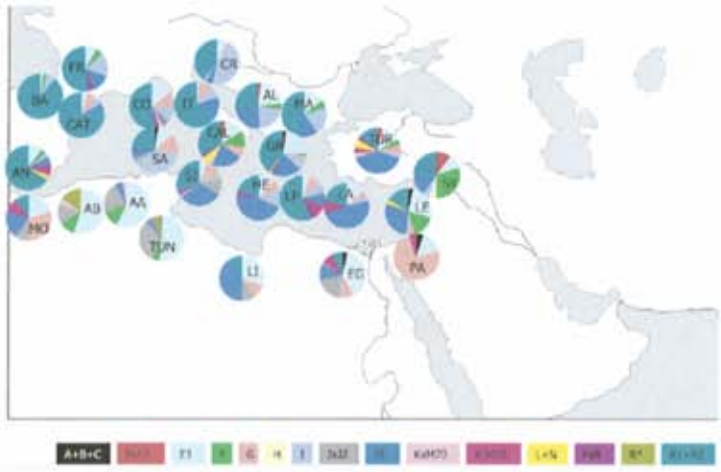


Figure 4.4: Haplogroup Frequencies in north Africa and Europe. Note the presence of the N haplogroup in Turkey. MO:Morocco, AB/AA:Algeria, TUN:Tunisia, LI:Libya, EG:EGYPT, TUR:Turkey.

are regions where the invading armies of Fatimid (and concomitant Crusader) invaders laid siege, built forts, and inevitably interacted demographically with the local populace; On a remarkable historic coincidence, The demarcation line between the invading crusaders and the Fatimids occurred in Lebanon, and in the south in particular, where the Fatimid ruler was based in the city of Sidon:

> *"...This time they clung to the narrow coastal plains that placed the towering mountains of Lebanon between the invaders and the powerful amirate of Damascus. The ruler of Tripoli supposedly agreed that if the Crusaders defeated the Fatimids, he would convert to Christianity and hold his lands under Crusader suzerainty [...] After crossed the Dog river north of Beirut, the Crusaders entered Fatimid territory, where several local governors supplied the intruders with money, food and guides in return for no damage to the surrounding agricultural area. But the Fatimid governor of Sidon refused to cooperate and his garrison attacked*

the Crusader host when it looted local villages. The towns further south generally followed the example of Beirut and by the time the Crusaders reached Acre they seem to have learned a lot about local religious and political rivalries..." (Nicolle & Hook, 2003).

With this, the historical picture gets clear: A Crusader invasion from the North, Fatimid stronghold in the South, and both factions keeping to the coastal areas: The Crusaders, to avoid confrontation with the emirate of Damascus, employing the Lebanese mountainous hinterland tot heir advantage; the fatmids, to avoid skirmishes with the Maronites, whose natural bastion of Mount Lebanon made gave them a keen military edge. We cannot tell for certain whether the current coastal distribution of sickle cell pockets in Lebanon was due to this cautious attitude of the invaders towards non-coastal areas or to malarial endemicity on certain coastal spots. However, we think that selection exerted by malaria couldn't have been so dramatic for the time period spanning the 10th-11th century hitherto, and thus we tend to favor the former hypothesis.

Chapter 5

Conclusion

When studying anthropological or historical events through molecular and genetic indices, few tools are as powerful as the study of Y-chromosome polymorphisms, notably with the great strides the field has taken in the past few years. In particular, the study of Y-polymorphisms proves more than handy when investigating past events in regions such as the Levant or the Middle-East, *i.e.* crossroads and melting-pots of different civilizations whose conquests, expansions, trade routes and interactions with neighbors eventually lead to an inextricable amalgamation of overlaying genetic and molecular signatures.

However, the fact that the markers studied on the Y-chromosome are on a non-recombining chromosomal framework, in addition to the patrilineal mode of inheritance allows the investigation of the aforementioned events. Another inherent advantage in using Y-molecular markers is that those anthropological events themselves were mostly mediated by males. In other terms, the Y-chromosome's patrilineal mode of inheritance is well suited for studying events such as conquests, invasions and pioneering/settling, within a patrilocal setting.

Recently, the genetic imprint of the phoenicians in the mediterranean was elucidated using the same methodology (Zalloua *et al.*, 2008a). It was a proof-of-concept that a systematic investigation using Y-chromosomal polymorphisms, with proper historical contextualization, can de-obfuscate even the most tangled of historical/anthropological events, much like the phoenician expansion in the Mediterranean.

With this paradigm in mind, we wished to investigate whether we could determine the dynamics through which sickle cell disease / gene was introduced to Lebanon and the neighboring region. Different hypotheses were forwarded previously in a bid to tackle this question.

Early since 1992, Oner *et al.* conjectured that the sickle gene was introduced to various Mediterranean countries from central West Africa via Algeria, Tunisia, and Egypt. He based his speculation on the exclusive occurence of the Benin haplotype among sickle-gene carriers in those countries. We found his proposition to be valid 18 years later

From the establishment of Carthage by the phoenicians, to the Roman and Byzantine hegemony over the Levant, in addition to the Arab Islamic conquests, Fatimid followed by Mameluk expansion, the Crusades, as well as migrations from neighboring Arab countries... Lebanon proves once more to be a true crossroad and a genetic playground for a spectrum of civilizations, even when it comes to a genetic disease such as SCD.

The work presented here seeks to explain the mode of entry of the sickle cell gene into the modern Lebanese population, and tries to explain its geographical and

demographic pattern. However, whilst doing that, we inadvertently came across a haplotype that can be thought of as a footprint of the expansion and influence of the Fatimid expansion from North Africa, and can be used in later studies attempting to characterize this expansion.

Bibliography

Adekile, A. D. (1992). Anthropology of the beta S gene-flow from West Africa to north Africa, the Mediterranean, and southern Europe. *Hemoglobin*, *16*(1-2), 105–121.

Adekile, A. D. (1997). Historical and Anthropological Correlates of Beta-S Haplotypes and Alfa- and Beta-Thalassemia alleles in the Arabian Peninsula. *Hemoglobin*, *21*(3), 281.

Aidoo, M., Terlouw, D. J., Kolczak, M. S., McElroy, P. D., Kuile, F. O., Kariuki, S., et al. (2002). Protective effects of the sickle cell gene against malaria morbidity and mortality. *The Lancet*, *359*(9314), 1311–1312.

Aksoy, M. (1961, May). Hemoglobin s in Eti-Turks and the Allewits in Lebanon. *Blood*, *17*, 657–659.

Allison, A. C. (1964). *Polymorphism and natural selection in human populations* (Vol. 29). New York, NY: Cold Spring Harbor Laboratory Press.

Antonarakis, S. E., Boehm, C. D., Serjeant, G. R., Theisen, C. E., Dover, G. J., & Kazazian, H. H. (1984). Origin of the Beta S-globin gene in blacks: The contribution of recurrent mutation or gene conversion or both. *Proc Natl Acad Sci USA*, *81*(3), 853–856.

Ashley-Koch, A., Yang, Q., & Olney, R. S. (2000). *Sickle hemoglobin (Hb S) allele and sickle cell disease: A huge review* (Vol. 151). Oxford: Oxford University Press.

Bandyopadhyay, A., Bandyopadhyay, S., Chowdhury, M. D., & Dasgupta, U. B. (1999, July). Major beta-globin gene mutations in eastern India and their associated haplotypes. *Human Heredity*, *49*(4), 232–235.

Bank, A. (2005, June). Understanding globin regulation in beta-thalassemia: It's as simple as alpha, beta, gamma, delta. *The Journal of Clinical Investigation*, *115*(6), 1470–1473.

Bashwari, L. A., Mandil, A. M., Bahnassy, A. A., Al-Shamsi, M. A., & Bukhari, H. A. (2001, February). Epidemiological profile of malaria in a university hospital in the eastern region of Saudi Arabia. *Saudi Medical Journal*, *22*(2), 133–138. (PMID: 11299407)

Basson, P. M. (1979). Genetic disease and culture patterns in Lebanon. *J Biosoc Sci*, *11*(2), 201–207.

Beeson, I. (1969, October). Cairo, a millennial. *Saudi Aramco World*, 24, 26–30.

Boussiou, M., Loukopoulos, D., Christakis, J., & Fessas, P. (1991). The origin of the sickle mutation in Greece: Evidence from beta S globin gene cluster polymorphisms. *Hemoglobin*, *15*(6), 459–467.

Bustin, S. A. (2000, October). Absolute quantification of mRNA using real-time reverse transcription polymerase chain reaction assays. *Journal of Molecular Endocrinology*, *25*(2), 169–193.

Bustin, S. A. (2005). Real-Time PCR. *Encyclopedia of diagnostic genomics and*

proteomics, *11*, 17–25.

Butler, J. M. (2003). Recent developments in y-short tandem repeat and y-single nucleotide polymorphism analysis. *Analysis*, *15*, 91.

Capelli, C., Brisighelli, F., Scarnicci, F., Arredi, B., Caglia, A., Vetrugno, G., et al. (2007). Y chromosome genetic variation in the Italian peninsula is clinal and supports an admixture model for the Mesolithic-Neolithic encounter. *Mol Phylogenet Evol*, *44*(1), 228–239.

Daar, S., Hussain, H. M., Gravell, D., Nagel, R. L., & Krishnamoorthy, R. (2000, May). Genetic epidemiology of HbS in Oman: Multicentric origin for the betaS gene. *American Journal of Hematology*, *64*(1), 39–46.

Das, S. K., & Talukder, G. (2001). A review on the origin and spread of deleterious mutants of the beta-globin gene in Indian populations. *Homo: Internationale Zeitschrift Fur Die Vergleichende Forschung Am Menschen*, *52*(2), 93–109.

El-Hazmi, M. A., Al-Swailem, A. R., Warsy, A. S., Al-Swailem, A. M., Sulaimani, R., & Al-Meshari, A. A. (1995). Consanguinity among the Saudi Arabian population. *British Medical Journal*, *32*(8), 623–626.

El-Kalla, S., & Baysal, E. (1998, June). Genotype-phenotype correlation of sickle cell disease in the United Arab Emirates. *Pediatric Hematology and Oncology*, *15*(3), 237–242. (PMID: 9615321)

Ewald, J. J. (1992). Slavery in Africa and the slave trades from Africa. *The American Historical Review*, *97*(2), 465–485.

Excoffier, L., Laval, G., & Schneider, S. (2005). Arlequin (version 3.0): An integrated software package for population genetics data analysis. *Evolutionary*

Bioinformatics Online, *1*, 47.

Excoffier, L., Smouse, P. E., & Quattro, J. M. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data. *Genetics*, *131*(2), 479–491.

Green, N. S., Fabry, M. E., Kaptue-Noche, L., & Nagel, R. L. (1993, Oct). Senegal haplotype is associated with higher HbF than Benin and Cameroon haplotypes in African children with sickle cell anemia. *Am J Hematol*, *44*(2), 145–146.

Holt, P. M., Lambton, A. K. S., & Lewis, B. (1977). *The Cambridge history of Islam.* Cambridge: Cambridge University Press.

Inati, A., Jradi, O., Tarabay, H., Moallem, H., Rachkidi, Y., Accaoui, R. E., et al. (2007, December). Sickle cell disease: The Lebanese experience. *International Journal of Laboratory Hematology*, *29*(6), 399–408.

Inati, A., Taher, A., W, W. B. A., Koussa, S., Kaspar, H., Shbaklo, H., et al. (2003). Beta-Globin gene cluster haplotypes and HbF levels are not the only modulators of sickle cell disease in Lebanon. *European Journal Of Haematology*, *70*(2), 79–83.

Ingram, V. M. (1989). A case of sickle-cell anaemia: A commentary on abnormal human haemoglobins . *Biochim Biophys Acta*, *1000*, 147–150.

Jobling, M. A., & Tyler-Smith, C. (2003). The human Y chromosome: An evolutionary marker comes of age. *Nature Reviews Genetics*, *4*, 599.

Kamal, I., Gabr, M., Mohyeldin, O., & Talaat, M. (1967). Frequency of Glucose 6-phosphate dehydrogenase deficiency in Egyptian children. *Acta Genet Stat Med*, *17*(4), 321–327.

Kan, Y. W., & Dozy, A. M. (1978). Polymorphism of DNA sequence adjacent to human ß-globin structural gene: Relationship to sickle mutation. *Proceedings of the National Academy of Sciences*, *75*(11), 5631–5635.

Keser, I., Sanlioglu, A. D., Manguoglu, E., Kayisli, O. G., Nal, N., Sargin, F., et al. (2004). Molecular analysis of beta-thalassemia and sickle cell anemia in Antalya. *Acta Haematologica*, *111*(4), 205–10.

Khlat, M. (1988). Consanguineous marriage and reproduction in Beirut, Lebanon. *American Journal of Human Genetics*, *43*(2), 188.

Kulozik, A. E., Wainscoat, J. S., Serjeant, G. R., Kar, B. C., Al-Awamy, B., Essan, G. J., et al. (1986, August). Geographical survey of beta s-globin gene haplotypes: Evidence for an independent asian origin of the sickle-cell mutation. *American Journal of Human Genetics*, *39*(2), 239–244. (PMID: 3752087)

Kwiatkowski, D. P. (2005, Aug). How malaria has affected the human genome and what human genetics can teach us about malaria. *Am J Hum Genet*, *77*(2), 171–192.

Lapouméroulie, C., Dunda, O., Ducrocq, R., Trabuchet, G., Mony-Lobé, M., Bodo, J. M., et al. (1992, May). A novel sickle cell mutation of yet another origin in Africa: The Cameroon type. *Hum Genet*, *89*(3), 333–337.

Lewis, B., & Lewis, B. (1990). *Race and slavery in the middle east: A historical enquiry.* New York, NY: Oxford University Press.

Marotta, C. A., Forget, B. G., Cohne-Solal, M., Wilson, J. T., & Weissman, S. M. (1977, July). Human beta-globin messenger RNA.: Nucleotide sequences derived from complementary RNA. *The Journal of Biological Chemistry*,

252(14), 5019–5031. (PMID: 873928)

Mears, J. G., Lachman, H. M., Cabannes, R., Amegnizin, K. P., Labie, D., & Nagel, R. L. (1981). Sickle gene: Its origin and diffusion from West Africa. *Journal of Clinical Investigation*, *68*(3), 606.

Muralitharan, S., Krishnamoorthy, R., & Nagel, R. L. (2003). Beta-globin-like gene cluster haplotypes in hemoglobinopathies. *Methods in Molecular Medicine*, *82*, 195–211.

Nagel, R. L., Fabry, M. E., Pagnier, J., Zohoun, I., Wajcman, H., Baudin, V., et al. (1985). Hematologically and genetically distinct forms of sickle cell anemia in Africa. The Senegal type and the Benin type. *N Engl J Med*, *312*(14), 880–884.

Neel, J. V. (1949). The inheritance of sickle cell anemia. *Science*, *110*(2846), 64–66.

Nicolle, D., & Hook, C. (2003). *The first crusade, 1096-99: Conquest of the Holy Land.* Botley, England: Osprey Publishing.

Pagnier, J., Mears, J. G., Dunda-Belkhodja, O., Schaefer-Rego, K. E., Beldjord, C., Nagel, R. L., et al. (1984, March). Evidence for the multicentric origin of the sickle cell hemoglobin gene in Africa. *Proceedings of the National Academy of Sciences of the United States of America*, *81*(6), 1771–1773.

Pauling, L., Itano, H. A., Singer, S. J., & Wells, I. C. (1949). Sickle cell anemia, a molecular disease. *Science*, *110*(2865), 543–548.

Peakall, R. O. D., & Smouse, P. E. (2006). GENALEX 6: genetic analysis in excel, population genetic software for teaching and research. *Molecular Ecology Notes*, *6*(1), 288–295.

Rahimi, Z., Karimi, M., Haghshenass, M., & Merat, A. (2003, November). Beta-globin gene cluster haplotypes in sickle cell patients from southwest Iran. *American Journal of Hematology*, *74*(3), 156–60.

Rahimi, Z., Merat, A., Gerard, N., Krishnamoorthy, R., & LNagel, R. (2006, December). Implications of the genetic epidemiology of globin haplotypes linked to the sickle cell gene in southern iran. *Human Biology*, *78*(6), 719.

Robino, C., Crobu, F., Gaetano, C. D., Bekada, A., Benhamamouch, S., Cerutti, N., et al. (2008, May). Analysis of Y-chromosomal SNP haplogroups and STR haplotypes in an Algerian population sample. *Int J Legal Med*, *122*(3), 251–255.

Slatkin, M. (1995). A measure of population subdivision based on microsatellite allele frequencies. *Genetics*, *139*(1), 457–462.

Steindorff, G., & Seele, K. C. (1957). *When Egypt ruled the East.* Chicago: University of Chicago Press.

Teebi, A., & Teebi, S. (2005). Genetic diversity among the Arabs. *Public Health Genomics*, *8*(1), 21–26.

Teebi, A. S., & Farag, T. I. (1997). *Genetic disorders among Arab populations.* New York, NY: Oxford University Press.

Wainscoat, J. S., Bell, J. I., Thein, S. L., Higgs, D. R., Sarjeant, G. R., Peto, T. E., et al. (1983). Multiple origins of the sickle mutation: Evidence from beta s globin gene cluster polymorphisms. *Mol Biol Med*, *1*(2), 191–197.

Walsh, B. (2001). Estimating the time to the most recent common ancestor for the y chromosome or mitochondrial DNA for a pair of individuals. *Genetics*,

158(2), 897–912.

Wasi, P., & Bowman, J. E. (1983). *Distribution and evolution of hemoglobin and globin loci.* New York, NY: Elsevier Science Publishing.

Williams, T. N., Mwangi, T. W., Roberts, D. J., Alexander, N. D., Weatherall, D. J., Wambua, S., et al. (2005, May). An immune basis for malaria protection by the sickle cell trait. *PLoS Med*, *2*(5), e128.

Williams, T. N., Mwangi, T. W., Wambua, S., Alexander, N. D., Kortok, M., Snow, R. W., et al. (2005, July). Sickle cell trait and the risk of plasmodium falciparum malaria and other childhood diseases. *The Journal of Infectious Diseases*, *192*(1), 178–186.

Willuweit, S., Roewer, L., & Group, I. F. Y. C. U. (2007, Jun). Y chromosome haplotype reference database (YHRD): Update. *Forensic Sci Int Genet*, *1*(2), 83–87.

Wintle, J. (2003). *History of Islam.* London: Rough Guides Ltd.

Zalloua, P. A., Platt, D. E., Sibai, M. E., Khalife, J., Makhoul, N., Haber, M., et al. (2008). Identifying genetic traces of historical expansions: Phoenician footprints in the Mediterranean. *The American Journal of Human Genetics*, *83*(5), 633–642.

Zalloua, P. A., Xue, Y., Khalife, J., Makhoul, N., Debiane, L., Platt, D. E., et al. (2008, April). Y-Chromosomal diversity in Lebanon is structured by recent historical events. *American Journal of Human Genetics*, *82*(4), 873–882.