Title: L1b-M317 Haplogroup shows Mount Lebanon Maronites as a persistently non-Emigrating population.

Daniel E. Platt¹, Hovig Artinian², Francis Mouzaya², Wissam Khalil³, Francois G. Kamar², David Comas⁴, Lisa Matisoo-Smith⁵, Francesc Calafell⁴, Nassim Nicolas Taleb^{6*}, Pierre Zalloua^{2,7,*}.

¹Bioinformatics and Pattern Discovery, IBM T. J. Watson Research Centre, Yorktown Hgts, NY 10598, USA

²The Lebanese American University, School of Medicine, Beirut 1102 2801, Lebanon

³Department of Arts and Archeology, Lebanese University, Lebanon

⁴Department of Anatomy, University of Otago, Dunedin, New Zealand

⁵Institut de Biologia Evolutiva (CSIC-UPF), Barcelona, Spain

⁶Risk Engineering, School of Engineering, New York University, USA

⁷Harvard T.H. Chan School of Public Health, Boston, MA 02215, USA.

*Corresponding author: Pierre Zalloua (pierre.zalloua@lau.edu.lb) or Nassim Nicolas Taleb (nnt1@nyu.edu)

Abstract

The various communities in Lebanon genetically differentiated prior to the establishment of today's religions. While evolving primarily within Lebanon, these communities have been influenced by local isolation as evident by their Y-haplogroup distributions. Additional genetic analyses of the Mount Lebanon and surrounding communities may provide details related to the origin and the migration time and routes of this mountainous community and how it evolved within the region. We used Y chromosome and whole genome autosomal data from Mount Lebanon and other surrounding communities to trace the migratory pattern that may have led to the genetic isolation and autosomal clustering in the mountainous regions of Lebanon. We observed individual clustering indicating that the Y chromosome L1b Levantine STR cluster branches each were around 5,000 years old, and shared a common ancestor 7,300 years ago that had immigrated to the Levant from the Caucasus. Both descriptive statistics and admixture modelling suggest that North Lebanese Mountain Maronites bear a distinctive ancestral Fertile Crescent genetic component far above every other population in the region. The qpAdm admixture model results are very consistent with the L1b results identifying an admixture event into the North Lebanese Mountain populations that originated in more northern populations, which was a much more recent event than the early basal expansion suggested by qpGraph. We believe that the Levantine L1b group split from the Caucasus around 7,300 years ago and migrated to the Levant. Differential religious adaption by populations from the North Lebanese Mountains are clearly aligned with the L1b haplotype STR haplogroup clusters, indicating pre-existing and persistent cultural barriers marked by the transmission of L1b lineages. Our findings highlight the value of uniparental haplogroups and STR haplotypes in probing historical events among these populations.

Introduction

Many of the early archaeologists and researchers who undertook archaeological and historical studies on Mount Lebanon believed that it was either not occupied, or very sparsely populated up until the conquest Arab-Muslim in the 7th century CE^{1; 2}. Currently, there are no written historical documents describing the populations that may have occupied Mount Lebanon during the Iron or Bronze Ages or before. Apart from the name Amurru which, in the first millennium BCE, was used to designate a geographical area in the north of hinterland modern-day Lebanon, and its inhabitants (Amorites), there is no other reference to the inhabitants of the Lebanese mountains and / or their territory in the Iron Age^{3; 4}.

In the first millennium AD, Roman historians, such as Strabo, Pliny and Flavius Josephus referred to the inhabitants of Mount Lebanon as Itureans⁵⁻⁷ describing them as mountain bandits occupying a rugged terrain that was impossible to cultivate and living in shelters underground and in caves along with their livestock.

Many of the twentieth century archaeologists and geographers who investigated the question of the inhabitants of the mountains of Lebanon and the nature of its human occupation also concluded that it was an unpopulated region during Antiquity and that would have been visited only by loggers and hunters^{1; 8}.

Today, active archaeological research is under way investigating the question of the occupation of mountainous and rural areas of Lebanon, attempting to provide some answers. Recent archaeological findings of some of the isolated sites in Mount Lebanon show evidence of human occupation since at least the Bronze age (3,200 BCE). Evidence of agglomerations and fortifications from the Bronze and Iron Age are very common in Mount Lebanon⁹. Further, a considerable number of inhabited sites and sanctuaries dating to the Roman (64 BCE-391/2 CE)

and the Proto-Byzantine (391/2 – 635 CE) periods dot the Lebanese mountains from north to south¹⁰. These sites, covering a large chronological range (from the Bronze Age to the early Arab Muslim conquest), depict an intense use of the mountainous territories as agricultural, trade and cult centers.

Today Mount Lebanon is mostly inhabited by the Maronite and Druze communities. The Maronites represent the oldest and most dominant Christian branch within Lebanon and has been occupying Mount Lebanon since the late 6th century CE¹¹⁻¹³. The Maronites draw their name from the Syrian Monastery that was built to honor the pious Marun or Maro, who lived, according to Maronite church historians, in the early 6th century on the banks of the Orontes, between Emesa and Apamea^{13; 14}.

We have previously shown, based on Y-chromosome and mitochondrial DNA analyses, that religious communities genetically differentiated prior to the establishment of religions¹⁵⁻¹⁷. We also showed that these communities, while evolving primarily within Lebanon, have been influenced by local isolation as evident by their Y-haplogroups distributions¹⁸⁻²⁰. The over-representation of haplogroup L1b-M317 among the Maronites of Mount Lebanon and the near total absence of this haplogroup from the rest of the Lebanese and surrounding populations irrespective of their religious backgrounds calls for further analyses to answer what factors created such strong genetic stratification.

Additional genetic analyses of the Mount Lebanon and surrounding communities focusing primarily on haplogroup L1b-M317 may provide details related to the origin and the migration time and routes of this mountainous community and how it evolved within the region.

In this study, we selected populations that may have contributed genetic admixture to the Lebanese population in general and to Mount Lebanon in particular. Throughout ancient and modern history, Lebanon has been the site for multiple events that were associated with major human mobility. Large scale population movement have been occurring in and out of Lebanon since the rule of the Babylonians and Assyrians, and continued through the Roman and Ottoman rules. These population movements must have certainly led to noticeable gene flow. While the most recent genetic admixture events are likely to be more readily detectable and have indeed contributed genetically their impact alone is not significant enough to explain the observed genetic diversity and its patterns observed in the modern population that occupied Lebanon today, especially the populations of Mount Lebanon. We used Y chromosome data from Mount Lebanon and other surrounding communities, including populations from Arabia, Armenia, Caucasus, Greece, Italy, Levant, Mesopotamia, and Turkey, as well as autosomal data primarily from communities in Mt Lebanon and from Beirut, to trace the migratory pattern that may have led to this pattern of L1b distribution and the autosomal clustering in the mountainous regions of Lebanon. We also sought to investigate the reasons for the isolation of haplogroup L1b-M317 as well as the autosomal clustering within the Maronite community of Mount Lebanon and whether this haplogroup can be associated with the first inhabitants of the mountain that later became Maronites.

Material and Methods

Populations and samples

The new samples collected for this study were selected based on analysis of their historic, ethnic, and geographical context and were actively collected in the field. A total of 2,944 samples were genotyped for Y SNPs and markers in our laboratory and analyzed; these included Lebanese (1,403), Iranians (416), Syrians (552), Cypriotes (166), Kuwait (42), and Palestinians (365). In addition, comparative data from Turkey (452 samples), Armenia (1,010 samples), Iran (104 samples), Kuwait (117 samples), Libya (175 samples) and the Caucasus (1,525) were obtained from published sources^{18; 19; 21-26} making a total of 6,327 samples (Supplementary Table 1).

Samples gathered for autosomal analysis were primarily collected from Beirut communities and include 43 samples from the Ashrafieh Greek Orthodox community, 35 from Shia community members, and 48 Sunni members. 53 Mount Lebanon Maronite samples were drawn from the Northern Mount Lebanon. Samples were also collected from Armenia (61), Cyprus (16), Iran (148), Iraq (50), Jordan (43), Syria (71), Turkey (39), Ukraine (3) (Supplementary Table 2).

The study participants recruited by our team had three generations of paternal ancestry and detailed information on their geographical origin and religious group or religious affiliation. They all signed an informed consent form approved by the IRB of the Lebanese American University.

For our STR analyses and in order to have a balanced representation of the L1b Haplogroup among the various regions we included, in addition to our samples (2,537), 128 L1b samples from key comparative regions (Caucasus, Turkey, Iran, Italy and Greece) from published work²⁷ and online available data from Family Tree DNA (Supplementary Table 4).

Finally, for the L1b analyses the study samples were divided into regions (10 in total) according to their geographic distribution (Supplementary Table 5). The North Lebanese Mountain Maronites (NLMM) represent the Maronite community in the Northern Lebanese Mountains.

Genotyping

DNA was extracted from buccal swab or blood samples. DNA from buccal swabs was extracted using phase lock gel (Qiagen, USA) and phenol and from blood using phenol/chloroform/isoamyl alcohol and subsequently precipitated in ethanol. SNP genotyping was done with the Applied Biosystems 7900HT Fast Real-Time PCR System with a set of custom Y-chromosomal binary marker assays (Thermo Fisher Scientific). The Y-chromosomal STR loci were amplified using the Y-filer™ PCR Amplification kit (Thermo Fisher Scientific) and analyzed on an Applied Biosystems 3130xl Genetic Analyzer. L1b-M317 was assigned by its terminal M317 SNP that classically defines L1b in the 2011 ISOGG tree

(https://isogg.org/tree/2011/ISOGG_HapgrpL11.html).

For the autosomal genotypes, samples were processed on the Infinium Omni Express-24 v1.2 Bead Chip, or the Human 1M-duo v3.0 array. Published ancient DNA data was used representing samples²⁸ from Anatolia_ChL (1), Anatolia_N (24), Armenia_ChL (5), Armenia_EBA (3), Armenia_MLBA (9), Iran_ChL (5), Iran_LN (1), Iran_N (5), Levant_BA (3), Levant_N (13), Natufian (6), Steppe_EMBA (28), Steppe_Eneolithic (3), Steppe_MLBA (22), details are provided in Supplementary Table 6. Other samples included Mbuti.DG (3)²⁹, and 1000 Genome Project samples GBR (91), CEU (99), FIN (99), GIH (103), and IBS (107)³⁰.

Analyses

NETWORK 5 (Fluxus-Engineering: <http://www.fluxus-engineering.com/>) was applied to determine a reduced median network of L1b STR haplotypes by using a reduction threshold of 1³¹. These networks were constructed with the 13 Y-STRs common to all samples DYS19, DYS389i, DYS389b, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, DYS439, DYS448, Y-GATA-H4, DYS456 (Supplementary Table 5).

Arlequin 3.5³² was applied to compute R_{ST} distances between populations. Euclidean distances between individual STR genotypes were clustered via UPGMA and plotted via heatmap in R 3.5.0³³ (<https://www.r-project.org/>). BATWING³⁴ was employed to compute TMRCAs for the main branches (which, on the assumption that they contain undiscovered unique event polymorphisms, we termed UEPs) identified in the NETWORK analysis, and the L1b haplogroup, run both on L1b samples alone, as well as together with all haplogroups in our dataset. Mutation rates were derived from multiple estimates, with Zhivotovsky's evolutionary and three pedigree based⁴²⁻⁴⁵ tabulations. Generation times were assumed to be 25 years for the evolutionary rates, and 30 years for the others.

Autosomal Analyses

We applied EIGENSOFT³⁵ for principal component analysis, ADMIXTURE³⁶, qpGraph³⁷, and qpAdm³⁸ to these sets (Supplementary Tables 2 & 6). Inbreeding coefficients were computed first using PLINK³⁹ ver 1.090b4.5 (25 Jul 2017), computing $F_{HOM} = (E_{HOM} - O_{HOM}) / (L - E_{HOM})$, where L is the number of genotyped autosomal SNPs, E_{HOM} is the expected homozygous genotypes, and O_{HOM} is the observed count of homozygous autosomal genotypes reported by PLINK. Second, inbreeding coefficients based on runs of homozygosity were computed from PLINK ver 1.90b6.10 (17 Jun 2019), computing $F_{ROH} = L_{ROH} / L_{AUTO}$ where L_{ROH} is the total length of ROH covered by the genotyped SNPs estimated by PLINK, and L_{AUTO}

is the length of the autosomes covered by the genotyped SNPs, which was 2,780,358 (caveat: while centromeres represent very distinctive and unsampled regions that are not typically sampled or characterized in ROH, statistics regarding sizes are not characterized for the same reasons; they were not excluded in the estimates of L_{AUTO}). PLINK parameters for ROH were --homozyg-kb=300, --homozyg-snp=50, --maf=0.05. Results were displayed using violin plots⁴⁰ available in the ggplot collection⁴¹ in R³³.

Results

Haplogroup distribution and frequencies.

Frequencies of haplogroup L1b were obtained from 10 different populations (Supplementary Table 1). The L1b haplogroup frequencies were the highest in Lebanon reaching 3.06%, followed by Turkey (2.88%) and Armenia (1.09%) (Figure 1, Table 1). Libya, Syria and the Caucasus had the lowest frequencies, at 0.57%, 0.54%, and 0.52% respectively. Within Lebanon (Supplementary Table 3b), the L1b haplogroup was significantly more frequent in the Maronite community of Northern Mount Lebanon (NLMM) than any other community in Lebanon or any other group tested in the region (χ^2 tests, $p < 0.05$), consistent with previous results by our group¹⁵. The L1b haplogroup was significantly more present in the Maronites (6.4%) compared to the general Lebanese population (3%). Out of the 43 L1b haplogroups found in Lebanon, 34 of them were Maronites from the North Lebanese Mountains.

Network, Clustering, and MDS Analysis

We assessed the distribution of the L1b haplogroup through STR network analysis by observing the geographic distribution of the L1b STR haplotypes. NETWORK results are shown in Figure 2. This NETWORK has three heavily localized branches that shared a common root from the L1b network in general. To these three branches, virtual UEP (unique event polymorphism) labels UEP2, UEP3, and UEP4 were assigned. UEP2 was comprised almost entirely of NLMM, but includes two Italians, while UEP3 and UEP4 were comprised of individuals broadly distributed over all of Lebanon.

The NETWORK computation was cross-checked by application of UPGMA clustering based on Euclidean distances between individual sample STRs, and plotted on a heat map, shown in Figure 3. Sidebar colors correspond to the NETWORK population colormap described above. The three virtual UEP groups identified by NETWORK are largely confirmed by the Euclidean distance UPGMA clustering. The UPGMA clusters agree with NETWORK in also indicating these clusters split relatively early in the history of L1b, though they are all primarily found in Lebanon. One interesting difference is that one of the Italians that reduce median network analysis placed in UEP2 was not associated with UEP2 haplotypes by UPGMA clustering by Euclidian distances. Turkish and Greek L1b haplotypes also tend to be marked by a distinct division among L1b STR genotypes compared to the other branches, also reflected in the NETWORK structure. The Arlequin R_{ST} population differences were subject to UPGMA (Figure 4) and MDS clustering (Figure 5). The UPGMA R_{ST} clustering separates the population groups identified in the Euclidean distance UPGMA clustering between individuals in the heatmap. Individuals tend to break into 3 major sub-clusters, with some mixing between populations. The R_{ST} statistics reflect the similarity among individuals in these branches, accounting for admixture affinities. The Greek/Caucasus/Turkish cluster emerges clearly in both clusters. The R_{ST} grouping of populations

included non-NLMM Lebanon, Syria as the “Levant.” The MDS analysis in 2D showed relatively poor STRESS (0.178) and R² (0.406) suggesting that the data are not well represented in a simple 2D plot. Nevertheless, the samples do reflect the organization of the UPGMA clustering, except that MDS appears to reflect the Reduced Median network for placing the Egypt sample over the UPGMA R_{ST} clustering.

Time and place of origin of L1b branches

We used BATWING to investigate timing of isolating or group splits while taking into consideration admixture events and immigration. In estimating times of split, BATWING will assign times based on dates of last admixture events, even when limited gene flow is involved (by immigration or admixture), thus it may significantly reduce split time estimates. For that reason, this report focused on TMRCA times of UEPs to mark expansion events.

BATWING computations were performed to estimate the ages of UEP2, UEP3, and UEP4, their combination, and L1b. UEP1 is the union of 2, 3, and 4, and represents a cluster of identifiably Lebanese L1b haplotypes. Table 2a displays estimates based only on L1b samples. The TMRCA for the whole measures the age of L1b. Table 2 b applied BATWING to the general set of all haplogroups, with UEP5 marking L1b in this analysis. BATWING estimates for the virtual UEPs are each around 4.6-5 ky. They shared a common ancestor around 7.3ky. We computed the age of L1b in two ways. First, using only L1b haplotypes, BATWING identified an overall TMRCA of around 11ky old. When combined with other haplogroups for a more representative population, UEP5 was assigned to mark L1b chromosomes (M317). This showed a TMRCA of around 10.6ky old. UPGMA on individual Euclidean distances does not suggest an oldest region,

though the NETWORK reduced median plot suggests a center of radiation with the Caucasus near the center. The Caucasus haplotypes are distributed over a majority of the branches.

Autosomal Analyses

We applied eigenstrat for PCA, ADMIXTURE, qpGraph, and qpAdm to samples in supplementary Table 2 & 6 as well as a selected 1000 genome data sets with Mbuti samples. PCA results, computed with least-squares projections, are shown in Figure 6. Figure 6a highlights the samples sharing the most information with the principal component. The PCA shows the North Lebanese Mountain Maronite samples strongly localized in the upper-left quadrant, Iranian and Iraqi samples in the upper-right quadrant, Jordanian, Syrian samples with some Iraqis in the lower quadrant. In the middle top, Armenian, just above ancient Steppes groups and Ukraine, mix into Iraqi samples, then Turkish, and Iranian samples moving from center to right. Moving from top center towards the left yields Cyprus and other Lebanese Christians, and the tail of the North Lebanese Mountain Maronite samples. Mutual information, augmented by the sign, between two normal factors A and B , shown in Figure 6b, was computed as $I_{AB} = -\text{sgn}(\rho_{AB}) \ln \sqrt{1 - \rho_{AB}^2}$ (Supplementary Information). In this case, we consider shared information between principal component as A and individual scaled and centered genotypes as B . The information scaling in Figure 6b suggests that the principal components are dominated by inter-regional genetic variations. PC1 is primarily Iranian. PC2 carries Syrian, Jordanian, Lebanese Shiia and Sunni. Turkey, Steppes, and Armenia project to the origin suggesting little shared information between their variation and PCs 1 and 2. Most of the

NLMM samples share information with both PC1 and PC2, but are anti-correlated with the others, while a scattering of Iranian samples share information with PC2.

ADMIXTURE at K=5, shown in Figure 7 identified basal populations among Natufians, Steppe_EMBA and Anatolia_N. North Lebanese Mountain Maronites appears to be dominated by an ancient component, colored pale blue, that is widespread across the Fertile Crescent populations. Another component heavily represented among Iranian samples, colored green, is also widespread throughout the more modern Fertile Crescent populations, is represented in Armenia in most modern DNA samples. Anatolia_N is marked by an ancestral population by yellow, and shows no Iranian genetic contributions, while Turkey shows significant admixture. Lastly, Natufians are identified as an ancient population marked by dark blue, which shows up in low contributions to modern Lebanese populations, North Lebanese Mountain Maronites least of all, though Syria and Jordan show contributions.

Since North Lebanese Mountain Maronites showed such a distinctive position in the PCA and such a high ancestral population in the ADMIXTURE analysis, we sought to understand whether this was a highly isolated population, or could be explained by modern admixture. Two very simple candidate hypotheses were constructed with qpGraph, one with North Lebanese Mountain Maronites deriving primarily from Africa (sharing a common ancestor with Mbutti.DG), with modern admixtures, and the other graph representing an admixture from other neighboring ancient populations. qpGraph tests yielded a z-score of 96.3 for an admixture of nearby populations producing North Lebanese Mountain Maronites, while an African rooted placement yielded a z-score of 18.98. This seems to support the North Lebanese Mountain Maronites isolation on the PCA plot, as well as an ancestral status, however its best fit deriving Anatolia_N as a strong admixture between North Mount Lebanon Maronites and a proto-Anatolia is

contradictory to ADMIXTURE. Further, while the z-score is much better for an ancestral African root than modern, the regression shows highly significant deviation from expected sampling variation. To test whether such a strong ancestral isolation is supported taking advantage of information from other immigration events that qpAdm affords and which qpGraph does not use, qpAdm was used to test North Lebanese Mountain Maronites as an admixture of Levant_N, Iran_N, and Armenia. qpAdm was also used to test LNMM as an admixture of Levant_N, Iran_N, and Armenia_ChL. These were tested using GBR as the (right-hand) base and FIN, CEU, GIH, and IBS as terminal migration samples (right-hand samples). The Armenia_ChL model scored a p-value of 0.0290, while the modern Armenia model scored 0.3510, indicating that North Lebanese Mountain Maronites can be described by admixtures of recent populations, and such cannot be excluded. The admixture coefficients for this model were 0.773 ± 0.034 for Armenia, 0.216 ± 0.023 for Levant_N, and 0.011 ± 0.013 for Iran_N, indicating Iranian genetic contributions are not significant, or were adequately explained by modern Armenian contributions.

Homozygosity analysis was performed to see if the isolation of NLMM was distinct from that of other groups within Lebanon that we explored. Figure 8 indicates distinct homozygosity inbreeding coefficients associated with each population. Both the NLMM and OLC are relatively isolated communities, though with different histories. The Sunni and Shia populations may have more regional immigrations, but consanguineous autozygosity is also present in the population. Both the inbreeding coefficients and runs of homozygosity present similar features to each other for all populations, except for the Shia, which carry some samples with low inbreeding coefficient estimates, not reflected in the ROH estimates, and all show individuals with very high homozygosity in all four populations. The modal inbreeding coefficients for

NLMM are largest across all groups, indicating higher typical inbreeding, though the range of variations is strongly overlapping across populations. The Christian individuals do not contain the tail of very strongly homozygous samples suggestive of consanguineous autozygosity as the Muslim samples do.

Discussion

The NETWORK, UPGMA individual clustering and BATWING analyses indicate that the L1b Levantine STR cluster branches carry diversity that took around 5ky to generate, and shared a common ancestor 7,300 years ago that had immigrated to the Levant from the Caucasus.

Migration from the Caucasus was likely triggered by the amelioration of the climate and induced migratory events in search for better habitats throughout the Northern Levant. There was a major climatic event around 8,200 years ago that caused severe drought throughout the Levant and most of Europe and southwest Asia⁴⁶. After that event multiple expansion events took place across Southwest Asia, reaching ultimately as far as the earliest Indus Valley Harappans⁴⁷.

In the Levant, these migrants must have arrived in the region, splitting into different groups, such as the Northern Lebanese Mountains population, between 5,000-4,600 years ago. All archeological evidence recovered to date from various locations in Mount Lebanon provide dates that are consistent with our DNA findings, all suggesting assimilation into community life in the Lebanese Mountains around the early Bronze age (5,000 years ago)^{9; 10}.

The Lebanese Northern Mountains with their hilly topography constituted a suitable habitat for the Levantine travelers, providing many shelters, lush and verdant spaces with good precipitation. The inhabitants of Northern Lebanon have therefore been present since at least the early Bronze age as attested by archeological findings. Interestingly, the modern genetic record

indicates complete affiliation with the Maronite faith, without the conversions to Islam observed in the regions outside of the northern Lebanese mountains.

The L1b haplogroup showed different frequencies and haplotype differentiation among regional and religious grouping in Lebanon. L1b was almost exclusively present in the Maronites of Mount Lebanon and is nearly absent in any of the other religious groups tested. This haplogroup is also nearly absent from most of the Levant and the rest of Southwest Asia while it is present in Turkey and the Caucasus. Two important observations can be stated about the haplotype differentiation of this haplogroup in the Levant. First, there is substantial geographical structure seen in the Levant. The Network analyses identified STR branches that are pretty strongly localized to Mount Lebanon highlands (L1b NLMM). Second, this L1b NLMM is exclusively Maronite and it is present within a more diverse, but geographically restricted Levantine branch (L1b Levant). These observations indicate that the L1b NLMM marked by UEP2 has remained relatively isolated with very little emigration over its estimated 5,000 year history. Otherwise we would have observed UEP2 members in other regions. The Euclidean distance-based UPGMA heat map clusters support the NETWORK results showing distinctive clusters of STRs strongly connected to the northern mountains of Lebanon. UEP3 appears to be much more broadly distributed in Lebanon with some random sample of its STR haplotypes appearing in the NLMM, it would appear this branch evolved in the Levant, with relatively recent immigration into the NLMM. Both UEP2 and UEP3 have very similar ages, with very slightly older UEP3 dates than for UEP2. UEP4 is another non-NLMM Lebanese cluster that shares a similar age with UEP2 and 3.

One observation is that the NLMM community was not completely isolated. Other UEPs besides UEP2 that appear to be rooted elsewhere in Lebanon include individuals now living in

the NLMM community, so immigration *into* the region occurred. Further, it is likely some of the inhabitants left genetic traces as far as Italy, suggesting that NLMM inhabitants participated in Lebanese emigrations over history. However, the numbers are quite low, perhaps due to the small population size in the region. Such low numbers would reduce the likelihood that STR haplotypes carried by NLMM emigrants would fix in the population rather than being lost to drift.

Y haplogroups show religious affinities in Lebanon¹⁶, with L1b being notably Maronite. The NLMM UEP2 cluster offers no evidence that its members participated in the expansion of the Maronite faith into regions outside of the Northern Lebanese mountains since its haplotypes are not found outside of the mountains, though as noted above, given the relatively small NLMM population, the emigrant lineages from NLMM impacting the spread of the Maronite Church could have failed to fix in the population due to drift. Meanwhile, the other Lebanese UEP3 and UEP4 groups are largely Maronite with a few Muslim members, though they appear not to represent NLMM rooted lineages, and split from UEP2 much earlier than the establishment of the Maronite Church. The haplogroup affinities and expansion associations further emphasize the isolation of the NLMM UEP2 group, which reflects the general very long-term resistance to successful emigration of member lineages of that population in general.

While the two Sicilian individuals among Lebanese UEPs may have also been the descendants of Phoenician sailors there is another more plausible explanation. In 1584 the Maronite College of Rome was founded to form a cultural bridge between the Christians of the Levant and Rome⁴⁸. It happened also that this College was established by a Maronite Patriarch from the mountains of North Lebanon. The College remained active for 228 years and was instrumental in instituting similar cultural and practices on both sides of the Mediterranean that are still observed today.

The College staff of clerical and lay people provided education to several hundreds of disciples from various age groups, not limited to Lebanon alone but also included Maronites from neighboring Syria and Cyprus. Most of these disciples would arrive to Italy from North Lebanon through Sicily. In fact, Sicily constituted a major hub of interaction between the Maronite community and Europe throughout most of the 16th and 17th Centuries. Among the names of the disciples of the Maronite College a substantial number came from the northern Lebanese mountains and it is not unlikely that the two Sicilian Levantine L1bs came to Sicily as part of the cultural exchange activities between Rome and the Christians of the Levant.

PCA, ADMIXTURE, Figure 7, and qpGraph suggests that North Lebanese Mountain Maronites bear a distinctive ancestral Fertile Crescent genetic component far above every other population in the region. However, qpAdm explains its composition with much stronger statistical agreement with its model than qpGraph offered, but restricts these to expansions listed in the “right-hand” population list, marking admixtures represented among those expansions. qpGraph, PCA, and ADMIXTURE all reflect the ancient population component preserved in the LNMM, while qpAdm appears to pick up more recent admixture events. Given our selection of populations, local populations that ADMIXTURE picked up will tend to be projected out by the F4 scores. However, the qpAdm admixture model results are also very consistent with the L1b results identifying an admixture event into the North Lebanese Mountain populations that originated in more northern populations, which was a much more recent event than the early basal expansion suggested by qpGraph.

In general, all populations within Lebanon present similar pictures of homozygosity, with NLMM’s modal homozygosity scores being largest among groups, yet the ranges of all groups

tend to overlap. Muslim populations appear to present some autozygosity. This suggests that all the populations are isolated, though with distinctive features specific to each.

The PCA also showed affinity between the “Other Christian Lebanese” (primarily Orthodox Catholic community members long present in Beirut) and Cyprus. The other communities showed population similarities with other regions typical of what was observed in prior Y-chromosome studies of Lebanon^{16; 18}.

Conclusion

We now believe that the Levantine L1b group split from the Caucasus around 7,300 years ago and migrated to the Levant, with different branches inhabiting coastal and inland hills areas. The L1b immigration marks one of possibly several layers of immigrations into an otherwise genetically highly isolated population. Migration from the Caucasus was likely triggered by the amelioration of the climate and induced migratory events in search for better habitats throughout the Northern Levant. Differential religious expansions into populations into NLMM vs. the rest of Lebanon are clearly seen in the L1b haplotype STR haplogroup clusters, indicating cultural barriers marked by the transmission of L1b lineages, highlighting the value of uniparental haplogroups and STR haplotypes in probing historical events among these populations.

Figures Legends

Figure 1. Map showing the geographical location and the frequency of the L1b M20 samples included in this study. The sizes of the circles correspond to the frequencies of the respective populations.

Figure 2. Reduced median network of L1b M20 showing STR haplotype distribution among populations; area is proportional to haplotype frequency, and color indicates populations. Connecting lines represent putative phylogenetic relationships between haplotypes. The three heavily localized branches that share a common root from the L1b network are assigned on the figure as UEP2, UEP3, and UEP4. Color code: Arabia - purple, Caucasus - brown, Egypt - yellow, Greece - blue, Italy - pink, Levant - dark green, Libya - light blue, NLMM - light green, Mesopotamia - orange, and Turkey - red.

Figure 3. Heatmap showing Euclidean distances between samples based on STRs clustered by UPGMA. Color code: Arabia - purple, Caucasus - brown, Egypt - yellow, Greece - blue, Italy - pink, Levant - dark green, Libya - light blue, NLMM - light green, Mesopotamia - orange, and Turkey - red.

Figure 4. UPGMA clustering using Arlequin RST distances between populations based on L1b STRs. GRE = Greece, CAU = Caucasus, TUR = Turkey, MES = Mesopotamia, ARA = Arabia, EGY = Egypt, NLMM = North Lebanon Mountain Maronites, LEV = Levant, ITL = Italy, LIB = Libya.

Figure 5. MDS plot of Rst distances between populations derived from L1b Y-STR data in Supplementary Table S3. GRE = Greece, CAU = Caucasus (Armenia, Azerbaijan, Georgia, Kurdistan, Russian Federation), TUR = Turkey, MES = Mesopotamia, ARA = Arabian Peninsula (Kuwait, Qatar, Saudi Arabia, UAE), EGY = Egypt, NLMM = North Lebanon Mountain Maronites, LEV = Levant (Non-LNM Lebanon, Syria, Cyprus), ITL = Italy, LIB = Libya.

Figure 6. PCA computed and displayed for the full list of populations indicated in Supplementary Table 2&6. a. PCA components 1 and 2 for recent and ancient regional populations, highlighting differentiation of NLMM from all other regional populations. b. Mutual information scaling for principal components indicating shared bitwise genotypic information between samples and principal components.

Figure 7. ADMIXTURE plot for admixtures computed on the full population list with populations plotted listed in Supplementary Table 2&6. ADMIXTURE ancestral components computed for regional populations with K=4 and 5 shows that NLMM contains the strongest regional representation of an ancestral component found distributed among Levantine and Fertile Crescent populations.

Figure 8. Plots of homozygosity computed a by coancestry, and b runs of homozygosity between strands per sample, showing a frequency chart as a fiddle diagram, and median and 95% quantiles of the fixation coefficients.

Tables:

Region	Nb. L1b (M20) Hg	Total Population	Frequency	Source
Lebanon	43	1403	3.06%	Present Study; Zalloua, 2009 ¹⁸
Turkey	13	452	2.88%	Cinnioğlu, 2004 ²¹
Armenia	11	1010	1.09%	Hovhannisyan, 2014 ²² ; Weale, 2001 ²³
Palestine	3	365	0.82%	Zalloua, 2008 ¹⁹
Iran	4	520	0.77%	Pereira, 2009 ²⁴ ; Zalloua, 2009 ¹⁸
Kuwait	1	159	0.63%	Zalloua, 2009 ¹⁸ ; Rebai, 2015a ²⁵
Cyprus	1	166	0.60%	Zalloua, 2008 ¹⁹ ; Zalloua, 2009 ¹⁸
Libya	1	175	0.57%	Rebai, 2015b ²⁶
Syria	3	552	0.54%	Zalloua, 2008 ¹⁹ ; Zalloua, 2009 ¹⁸
Caucasus	8	1525	0.52%	Balanovsky, 2011 ⁴⁹

Table1. Haplogroup L1b M20 frequencies in Lebanon and surrounding regions

Mutation model	L1b TMRCA	UEP1 TMRCA	UEP2 TMRCA	UEP3 TMRCA	UEP4 TMRCA
Zhivotovsky	36 (21 – 71)	27 (12 – 57)	16 (7.9 – 33)	17 (8.4 – 35)	17 (8.4 – 35)
Burgarella ImMR	11.5 (8.0 – 19.2)	9.0 (4.8 – 16.5)	5.5 (3.0 – 9.7)	5.7 (3.2 – 9.9)	5.8 (3.2 – 9.8)
Shi OMR	10.9 (7.5 – 19.2)	8.2 (4.3 – 16.9)	5.0 (2.8 – 9.4)	5.2 (2.9 – 9.5)	5.4 (3.0 – 9.9)
Burgarella OMR	11.9 (8.2 – 20.2)	9.1 (4.7 – 17.7)	5.6 (3.1 – 10.0)	5.8 (3.2 – 10.2)	5.9 (3.3 – 10.6)

Table 2a: STR derived time estimates based on L1b samples. “OMR” = “observed mutation rates” by both Shi et al⁴³ and Burgarella et al⁴², “ImMR” = “logistic model mutation rates” by Burgarella et al – each with 30 year generation times, and Zhivotovsky “EMR” = “evolutionary mutation rates” estimated by Zhivotovsky⁴⁵ with 25 yr generation times. A recalibrated evolutionary mutation rate “rEMR” by Shi et al differed from the Zhivotovsky EMR in an STR not included in these analyses, so was identical to the Zhivotovsky EMR for our purposes. Shi et al also presented summary parameters for the other references.

Mutation model	L1b TMRCA UEP5	UEP1 TMRCA	UEP2 TMRCA	UEP3 TMRCA	UEP4 TMRCA
Zhivotovsky	20 (13 – 28)	15 (6.9 – 23)	9.6 (4.9 – 17)	10 (5.1 – 17)	9.9 (5.0 – 18)
Burgarella ImMR	10.6 (7.4 – 14.6)	6.4 (4.3 – 10.2)	4.3 (2.6 – 6.8)	4.3 (2.6 – 6.8)	4.5 (2.8 – 7.2)
Shi OMR	10.8 (7.4 – 15.1)	7.3 (4.5 – 10.7)	4.6 (2.8 – 7.8)	4.8 (3.1 – 7.9)	4.6 (2.8 – 7.5)
Burgarella OMR	9.8 (7.5 – 13.8)	7.5 (4.8 – 10.5)	5.1 (3.1 – 8.2)	5.2 (3.2 – 7.9)	5.2 (3.2 – 8.5)

Table 2b. STR derived time estimates based on all haplogroups. UEP5 marks all the L1b’s among all the samples, and it is used to compute L1b’s TMRCA. “OMR” = “observed mutation rates” by both Shi et al⁴³ and Burgarella et al⁴², “ImMR” = “logistic model mutation rates” by Burgarella et al – each with 30 year generation times, and Zhivotovsky “EMR” = “evolutionary mutation rates” estimated by Zhivotovsky⁴⁵ with 25 yr generation times. A recalibrated evolutionary mutation rate “rEMR” by Shi et al differed from the Zhivotovsky EMR in an STR not included in these analyses, so was identical to the Zhivotovsky EMR for our purposes. Shi et al also presented summary parameters for the other references.

References

1. Planhol, X.d. (1968). Les Fondements géographiques de l'histoire de l'Islam.(Paris: Flammarion).
2. Vaumas, t.d. (1955). Le Liban : montagne libanaise, Bekaa, Anti-Liban, Hermon, Haute Galilée libanaise : Étude de géographie physique.(Paris: Firmin-Didot).
3. Clay, A.T. (2013). The empire of the Amorites.(Miami: HardPress Publishing).
4. Starcky, J. (1982). La flèche de Zakarba'Al roi d'Amurru. pp 179-186.
5. Josephus, F., and Henten, J.W.v. (2014). Judean antiquities 15 : translation and commentary.
6. Pline, I.A., and LittrÉ, m. (2016). Histoire naturelle.(Paris: Les Belles Lettres).
7. Strabon, and Jones, H.L. (1966). The geography of Strabo. 7, 7.(London; Cambridge, Mass.: W. Heinemann ; Harvard University press).
8. Planhol, X.d. (1993). Les nations du prophète : manuel géographique de politique.(Paris: Fayard).
9. Khalil, W. (2012). Routes et fortifications dans le Chouf libanais Dossiers d'Archéologie 22 22, pp. 70-75.
10. Aliquot, J. (2009). La vie religieuse au Liban sous l'Empire romain.(Beyrouth: Institut Français du Proche-Orient).
11. Hitti, P.K. (1967). Lebanon in history : from the earliest times to the present.(London ; MacMillan; New York, St. Martin's).
12. Lammens, H. (1902). Tasrih al-absar fi ma yahtawi Lubnan min al-athar.([Beirut]: [publisher not identified]).
13. Moosa, M. (2005). The Maronites in history.(Piscataway, NJ: Gorgias Press).
14. Henley, A.D.M. (2008). Politics of a Church at War: Maronite Catholicism in the Lebanese Civil War. Mediterranean Politics 13, 353-369.
15. Haber, M., Platt, D.E., Badro, D.A., Xue, Y., El-Sibai, M., Bonab, M.A., Youhanna, S.C., Saade, S., Soria-Hernanz, D.F., Royyuru, A., et al. (2011). Influences of history, geography, and religion on genetic structure: the Maronites in Lebanon. Eur J Hum Genet 19, 334-340.
16. Zalloua, P.A., Xue, Y., Khalife, J., Makhoul, N., Debiane, L., Platt, D.E., Royyuru, A.K., Herrera, R.J., Hernanz, D.F., Blue-Smith, J., et al. (2008). Y-chromosomal diversity in Lebanon is structured by recent historical events. Am J Hum Genet 82, 873-882.
17. Haber, M., Gauguier, D., Youhanna, S., Patterson, N., Moorjani, P., BotiguÉ, L.R., Platt, D.E., Matisoo-Smith, E., Soria-Hernanz, D.F., Wells, R.S., et al. (2013). Genome-Wide Diversity in the Levant Reveals Recent Structuring by Culture. PLoS Genet PLoS Genetics 9, e1003316.
18. El-Sibai, M., Platt, D.E., Haber, M., Xue, Y., Youhanna, S.C., Wells, R.S., Izaabel, H., Sanyoura, M.F., Harmanani, H., Bonab, M.A., et al. (2009). Geographical structure of the Y-chromosomal genetic landscape of the Levant: a coastal-inland contrast. Ann Hum Genet 73, 568-581.
19. Zalloua, P.A., Platt, D.E., El Sibai, M., Khalife, J., Makhoul, N., Haber, M., Xue, Y., Izaabel, H., Bosch, E., Adams, S.M., et al. (2008). Identifying genetic traces of historical expansions: Phoenician footprints in the Mediterranean. Am J Hum Genet 83, 633-642.
20. Badro, D., A., Bouchra, D., Haber, M., Youhanna, S., C. , Salloum, A., Ghassibe-Sabbagh, M., Johnsrud, B., Khazen, G., Matisoo-Smith, E., Soria-Hernanz, D., F. , et al. (2013). Y-Chromosome and mtDNA Genetics Reveal Significant Contrasts in Affinities of Modern Middle Eastern Populations with European and African Populations. PLOS ONE 8.
21. Cinnioglu, C., King, R., Kivisild, T., Kalfoglu, E., Atasoy, S., Cavalleri, G.L., Lillie, A.S., Roseman, C.C., Lin, A.A., Prince, K., et al. (2004). Excavating Y-chromosome haplotype strata in Anatolia. Hum Genet 114, 127-148.

22. Hovhannisyan, A., Khachatryan, Z., Haber, M., Hrechdakian, P., Karafet, T., Zalloua, P., and Yepiskoposyan, L. (2014). Different waves and directions of Neolithic migrations in the Armenian Highland. *Investig Genet* 5, 15.
23. Weale, M.E., Yepiskoposyan, L., Jager, R.F., Hovhannisyan, N., Khudoyan, A., Burbage-Hall, O., Bradman, N., and Thomas, M.G. (2001). Armenian Y chromosome haplotypes reveal strong regional structure within a single ethno-national group. *Hum Genet* 109, 659-674.
24. Alshamali, F., Pereira, L., Budowle, B., Poloni, E.S., and Currat, M. (2009). Local population structure in Arabian Peninsula revealed by Y-STR diversity. *Hum Hered* 68, 45-54.
25. Triki-Fendri, S., Sanchez-Diz, P., Rey-Gonzalez, D., Alfadhli, S., Ayadi, I., Ben Marzoug, R., Carracedo, A., and Rebai, A. (2016). Genetic structure of the Kuwaiti population revealed by paternal lineages. *Am J Hum Biol* 28, 203-212.
26. Triki-Fendri, S., Sanchez-Diz, P., Rey-Gonzalez, D., Ayadi, I., Carracedo, A., and Rebai, A. (2015). Paternal lineages in Libya inferred from Y-chromosome haplogroups. *Am J Phys Anthropol* 157, 242-251.
27. Boattini, A., Martinez-Cruz, B., Sarno, S., Harmant, C., Useli, A., Sanz, P., Yang-Yao, D., Manry, J., Ciani, G., Luiselli, D., et al. (2013). Uniparental markers in Italy reveal a sex-biased genetic structure and different historical strata. *PLoS One* 8, e65441.
28. Lazaridis, I., Nadel, D., Rollefson, G., Merrett, D.C., Rohland, N., Mallick, S., Fernandes, D., Novak, M., Gamarra, B., Sirak, K., et al. (2016). Genomic insights into the origin of farming in the ancient Near East. *Nature* 536, 419-424.
29. Swapan, M., Heng, L., Mark, L., Iain, M., Melissa, G., Fernando, R., Mengyao, Z., and Ninj, C. (2016). The Simons Genome Diversity Project: 300 Genomes from 142 Diverse Populations. *Nature* 7624.
30. Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., Abecasis, G.R., et al. (2015). A global reference for human genetic variation. *Nature* 526, 68-74.
31. Bandelt, H.J., Forster, P., and Rohl, A. (1999). Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol* 16, 37-48.
32. Excoffier, L., and Lischer, H.E. (2010). Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol Ecol Resour* 10, 564-567.
33. R Core Team. (2019). R: A Language and Environment for Statistical Computing.(Vienna, Austria: R Foundation for Statistical Computing).
34. Wilson, I.J., Weale, M.E., and Balding, D.J. (2003). Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 166 155-188.
35. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38, 904-909.
36. Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* 19, 1655-1664.
37. Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., and Reich, D. (2012). Ancient admixture in human history. *Genetics* 192, 1065-1093.
38. Haak, W., Lazaridis, I., Patterson, N., Rohland, N., Mallick, S., Llamas, B., Brandt, G., Nordenfelt, S., Harney, E., Stewardson, K., et al. (2015). Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* 522, 207-211.
39. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81, 559-575.

40. Hintze, J.L., and Nelson, R.D. (1998). A Box Plot-Density Trace Synergism. *The American Statistician* 52, 181-184.
41. Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. (New York: Springer-Verlag).
42. Burgarella, C., and Navascues, M. (2011). Mutation rate estimates for 110 Y-chromosome STRs combining population and father-son pair data. *Eur J Hum Genet* 19, 70-75.
43. Shi, W., Ayub, Q., Vermeulen, M., Shao, R.G., Zuniga, S., van der Gaag, K., de Knijff, P., Kayser, M., Xue, Y., and Tyler-Smith, C. (2010). A worldwide survey of human male demographic history based on Y-SNP and Y-STR data from the HGDP-CEPH populations. *Mol Biol Evol* 27, 385-393.
44. Wei, W., Ayub, Q., Xue, Y., and Tyler-Smith, C. (2013). A comparison of Y-chromosomal lineage dating using either resequencing or Y-SNP plus Y-STR genotyping. *Forensic Sci Int Genet* 7, 568-572.
45. Zhivotovsky, L.A., Underhill, P.A., Cinnioglu, C., Kayser, M., Morar, B., Kivisild, T., Scozzari, R., Cruciani, F., Destro-Bisol, G., Spedini, G., et al. (2004). The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time. *Am J Hum Genet* 74, 50-61.
46. Alley, R.B., Mayewski, P.A., Sowers, T., Stuiver, M., Taylor, K.C., and Clark, P.U. (1997). Holocene climatic instability: A prominent, widespread event 8200 yr ago. *GEOLOGY -BOULDER-* 25, 483-486.
47. Sharma, R.S. (2007). *Harappan Culture: Bronze Age Urbanization in the Indus Valley*.
48. Khairallah, S. (1998). The Maronite College in Rome. *THEOLOGICAL REVIEW -BEIRUT-* 19, 105-124.
49. Balanovsky, O., Dibirova, K., Dybo, A., Mudrak, O., Frolova, S., Pocheshkhova, E., Haber, M., Platt, D., Schurr, T., Haak, W., et al. (2011). Parallel evolution of genes and languages in the Caucasus region. *Mol Biol Evol* 28, 2905-2920.