Supplementary Notes and Figures include:

Supplementary Note 1: Lexico-statistical methods.

Supplementary Note 2: Comparing historical, linguistic and genetic dates.

Supplementary Figure 1: Phylogenetic networks for the high-frequency Caucasus haplogroups.

Haplogroup G2a*-P15(xP16,P303,M406)

Haplogroup G2a3b1-P303

Haplogroup J1* - M267 (xP58)

Haplogroup J2*-M172 (xM12, xM47, xM67)

Haplogroup J2a4b-M67 (xM92)

Haplogroup L3-M357

Haplogroup Q-M242

Haplogroup R1a1a*-M198 (xM458)

Haplogroup R1a1a7-M458

Haplogroup R1b1b2-M269

Supplementary Figure 2. Lexico-statistical reconstruction of the evolution of North Caucasus languages.

Supplementary Figure 3. Genetic reconstruction of the evolution of North Caucasus populations.

Supplementary Figure 4. BATWING tree of the population splits. Numbers on branches indicate time (in years) between corresponding splitting events.

Supplementary Note 1.

LEXICO-STATISTICAL METHODS

The basis

To reconstruct a phylogenetic tree of a set of languages, it is essential to consider what a human language is, and how languages could be classified as the nodes of a phylogenetic tree. A language could be considered as a semiotic system for the transmission of information in a society, existing independently of speakers, changing in the course of time, and consisting of 'signs' and the relationships between them. A language 'sign' is generally a double-faced entity, including both semantics and text. Within a sign, the link between the semantic face and textual face is conventional and caused by tradition of a concrete language. Therefore, the coincidence of the textual faces of the words (morphemes, signs) with the same meanings in two different languages (which is what linguists can observe in reality) needs explanation. When considering large sets of words, random coincidence is statistically unlikely. Therefore, the basic assumption of comparative historical linguistics is that these multiple homogenous coincidences indicate that these pairs of signs are two different reflections of one proto-sign. In other words, comparative historical linguistics considers those coincidences of a set of signs where both semantics and text are matched between two (or more) languages.

Matches between textual faces need not necessarily be exact (literal); most often, two sets of words in two languages could be deduced from the third (hypothetical) proto-form of the words following regular phonetical rules. These proto-signs could be either inherited from the common proto-language or could be obtained from a contact language (loanwords, borrowings). To distinguish between these two cases, we shall examine the sets of rules relating our sets of words (signs) with their protoforms. Generally, it turns out that some phonetical rules are specific to certain thematic groups of words, usually related with some achievements of culture or technology (for example terms of harness, science, or bronze working); these matches are typically loan words. Other phonetic rules are, in principle, universal for the given pair of languages; these matches could be considered as inherited from the common language.

The phonetical changes described above (which do not disrupt the identity of signs) are one type of language change. If we can demonstrate that two language systems contain large identical sets of signs (not related with specific "cultural" areas) which differ only as a result of regular phonetical changes, we have proved the affinity of those languages. However, another type of change is the replacement of signs. Namely, to express the same meaning, a language may use two different words. The basic assumption of the branch of comparative historical linguistics named "glottochronology" is that the process of replacement of words in a given language could be described generally as by random probability, such that an average replacement rate can be calculated.   As replacement is non-probabilistic for the names of cultural or technologic entities, for such a calculation we need words with meanings belonging to a so-called pre-cultural vocabulary. Such words, as a rule, do not get borrowed, as the things designed by them are common to any human culture and the borrowing is unnecessary. Thus, the word lists used in lexicostatistical calculations are compiled from thematic groups within which loan words are quite rare (Swadesh, 1955; Starostin, 1989). These rare loan words are considered an “informational noise” and should be excluded from computations (Starostin, 1989).

The applications

This approach in its initial version was proposed by Swadesh (1955) but has subsequently been criticized. Present-day studies in the field of glottochronology generally take into account the main critical points, such as the possibility of borrowing words (borrowings are now carefully identified and excluded) and the assumption of a linear rate of word substitution (in present-day studies the “relaxed clock” model (Kitchen et al., 2009), or the exponential formula are used instead of the linear formula).

Swadesh word lists were recently successfully used for reconstructing Indo-European (Gray and Atkinson, 2003), Semitic (Kitchen et al., 2009), Austronesian (Greenhill et al., 2010) and other language families. In these studies, most effort was made in the statistical analysis of existing data sets, including mathematical modeling and the use of advanced calculating approaches (like Bayesian Markov chain Monte Carlo methods) similar to those used in evolutionary biology.

Another research group, the Moscow school of glottochronology, has applied Swadesh word lists for classifying Turkic (Dybo, 2006), Paleo-Siberian (Mudrak, 2008), North Caucasian (Nikolaev and Starostin 1995), Altaic (Starostin, 1991) and other language families. In these studies, most effort was put into creating new high-quality data sets, including etymological studies, establishing phonetic rules, and intensive exclusion of borrowings. The analytical methods did not significantly change after Starostin’s (1989) correction of the Swadesh method.

The method

Lexico-statistical analyses, including the evaluation of the taxonomic characteristics of the North Caucasian family and the dating of its splits into different branches, were carried out using the method of glottochronology proposed by Swadesh (1955) and then improved by Starostin (Starostin, 1989; Starostin, 2000; Embleton, 2000).

Among the five postulates of Swadesh, the third was that the percentage of words shared between two languages decreases linearly with the time after splitting. When analyzing Indo-

European and Semitic languages, as well as comparing Old Chinese, Middle Chinese, and Modern Chinese dialects, Starostin found an exponential, rather than linear, dependence to be the best approximation to empirical data (i.e., the rate of changes is maximal after a split, and then gradually decreases, up to the next splitting event). The important feature of Starostin's (and all the Moscow school's) work is that we apply procedures to carefully identify and exclude loan words from our calculations; for this reason, we cannot analyze languages where the etymologies of most words are unknown.

The decay ("erosion", or replacing) of the stable lexical elements from the initial list can be mathematically described by the formula suggested by Starostin (1989):

$$C(t) = C_0 \cdot e^{-\lambda C(t) t^2}$$

where C(t) is the frequency of words from the basic list surviving in a given language at time t, C_0 is the initial basic list, and λ is a coefficient indicating the rate of replacing words from the basic list. For lexical glottochronology using the 100-word list of Swadesh, λ=0.05 (Starostin, 1989).

When dating the split of two languages, this dependence can be written as follows:

$$C_2(t) = C_0 \cdot e^{-2\lambda t^2 \sqrt{C_2(t)}}$$

where $C_2(t)$ is the frequency of words surviving in both descendant languages.

Therefore, one can calculate the time t from the coefficient λ and the frequency of words surviving within a language pair:

$$t^2 = \frac{\ln \dfrac{C_2(t)}{C_0}}{-2\lambda\sqrt{C_2(t)}}$$

Linguistic distances (needed for the Mantel test) were calculated as the reciprocal of linguistic similarities ($1/C_2(t)$).

Lexical versus typological comparisons

The credibility of glottochronology methods based on lexical information was shown for a time depth of approximately 10,000 years (Gell-Mann et al., 2008; Starostin et al., 2003). To extend this limit, it was suggested to use typological (grammatical) data. as well (see, for example, Greenhill et al., 2010).

As suggested above, the concept of historical affinity of languages is directly related to the phenomenon of retaining the identity of a large set of signs in the course of divergence of languages. When some linguists working within comparative historical linguistics say that a language affinity can be proved only by the presence of grammatical coincidences between the studied languages, they mean the following. The signs (e.g., the morphemes) whose meanings are grammatical are borrowed substantially more rarely then lexical morphemes (e.g., word stems). Therefore, if we can demonstrate that a given pair of languages contains some sets of identical (differing only by the results of phonetical changes) grammatical morphemes, then we have proven the affinity of those languages. The sticking point is that the quantity of grammatical morphemes in a given language is substantially less than the quantity of lexical ones, and far from enough for establishing regular phonetical correspondences. In a number of human languages, there are no grammatical morphemes. Thus, the lexical criterion of language affinity is more universal.

Typological (grammatical) similarities could be caused either by relationship (inherited from a common proto-language) or by convergence (when speakers of one language use their habitual grammatical structures when they switch to another language or by chance (because of the restricted quantity of such features). However, linguists have no general methods for distinguishing between these three cases. In particular, the inherited character of some typological feature can only be argued if a set of morphemes expressing this feature is inherited (allowing for phonetical changes), which returns us to the previous case. For this reason, in our work, we prefer to prove a common origin of sets of languages using vocabulary rather than "grammar".

Similarities between linguistic and genetic dating methods

The calculation of linguistic dates is based on counting the differences that have accumulated after a set of languages split. Similarly, genetic dates are calculated from the number of mutations that have accumulated after a haplogroup started to expand. Both approaches depend on a coefficient that converts diversity (genetic or linguistic) into time; both obtain this rate by calibrating their clocks from case studies. This methodological similarity allows a direct comparison of the genetic and linguistic dates for the same events in population history.

From a genetic perspective, when a population splits and a haplotype cluster starts to expand in one of the daughter populations, this cluster will be restricted to only this daughter (assuming that male-mediated migration between the populations is negligible). Therefore, estimation of its age will not date the split itself (i.e., the population divergence) but the cluster expansion event, which could have happened at the time of split or *after* the split, at any time during the evolution of the daughter population.

From a linguistic perspective, innovations present in one language/dialect but absent in another arise after the physical separation of a proto-population of speakers into partly isolated subpopulations. The linguistic dating of the split will thus show the point in time when the language differences became fixed, rather than the beginning of this process. Therefore, the start of population divergence could be some generations before the date shown in language development. An additional lag of linguistic dates might be caused by the time needed for incorporation of the newly introduced words into written sources which are used for lexico-statistical analysis. As a result, both linguistic and genetic dating methods used in our study provide a recent (lower) estimate of the population splitting event.

Supplementary Note 2.

COMPARING HISTORICAL, LINGUISTIC AND GENETIC DATES

Here, we consider two population events in the Caucasus which are documented archaeologically, allowing comparison among historical, linguistic and genetic dates.

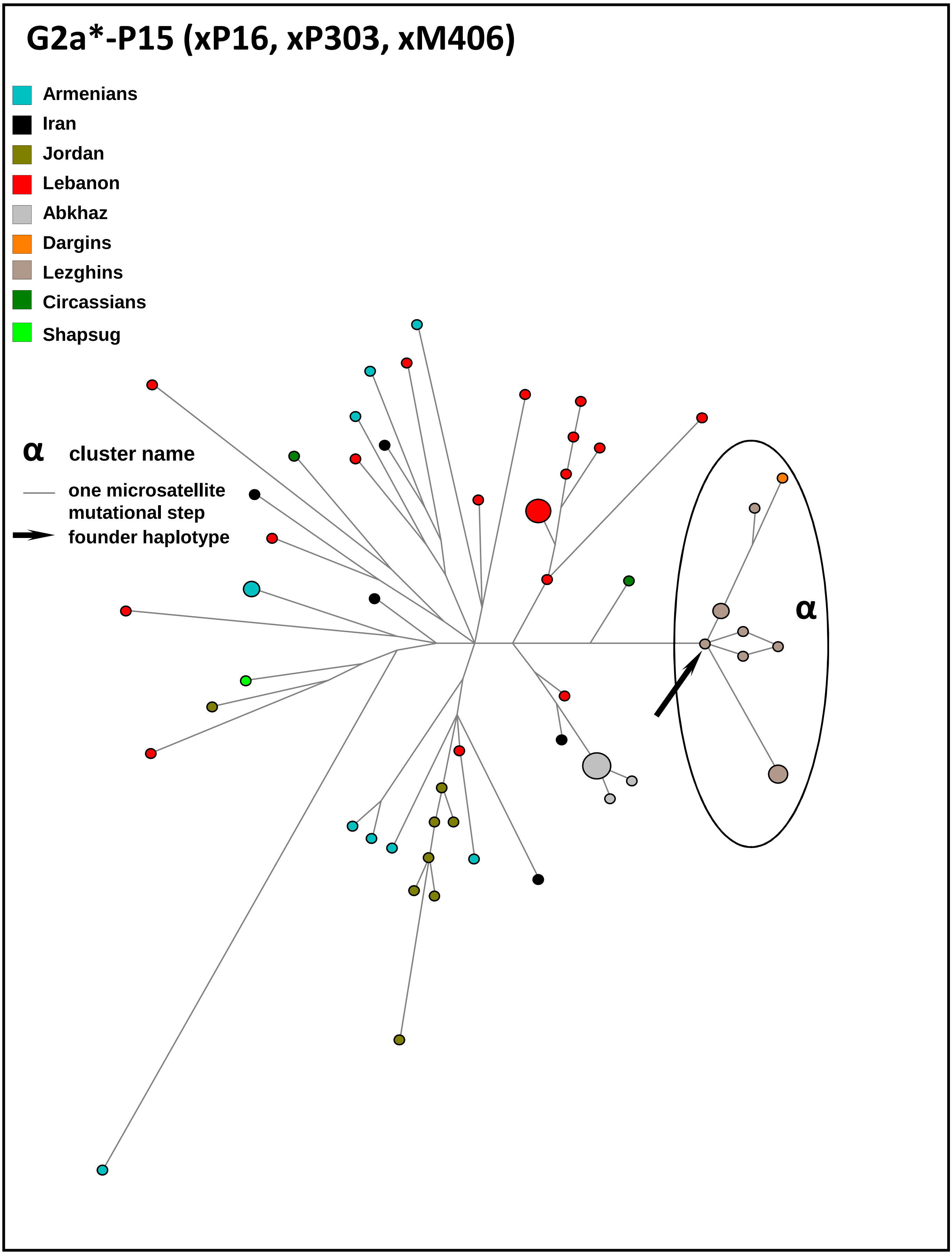
The first case is the split between two Ossetian subpopulations, the Ironians and the Digorians. There are three historical estimates of this split. The first supposes that Ironians and Digorians are descendants of different Sauromatian tribes, dwelling not in the Caucasus but in the nearby steppes. The second opinion traces the Ossetian subpopulations back to the Iranic-speaking Syraks and Aors, who appeared in the Central Caucasus in the 3rd-2nd centuries BC and remained there up to the time of the Huns (3rd-4th centuries AD) in the foothills. The third view postulates that the split into Ironians and Digorians occurred in the Alan era, whose culture reached its maximum spread in the 6th century AD in the Central Caucasus mountains and foothills (Kovalevskaya, 1981). Digorians were first mentioned in written sources in "Armenian Geography" in the 7th century AD (Ageeva, 2000). The linguistic date obtained in our study is the 7th century (Supplementary Figure 2). Overall, different historical estimates are consistent with the splitting of the Ironians and Digorians at times from the 6th century BC to the 8th century AD.

Genetic dates have also been obtained for the clusters G1a1a-P18-α and G1a1a-P18-β. Because each cluster is specific to one of two Ossetian subpopulations (Ironians and Digorians) and both clusters show the same age, their expansion probably occurred soon after the split of the Ossetian proto-population. The genetic diversity ρ=1.43 (averaged between the two clusters) yields a date of 1,400±500 years when using the genealogical mutation rate and 4,100±1400 years when using the evolutionary rate. Thus, the genealogical rate corresponds with historical and linguistic dates, while the estimate based on the evolutionary rate is too old, even considering the recent skew of the confidence intervals. BATWING analysis yields estimates of 1,100 to 5,200 YBP for these clusters.

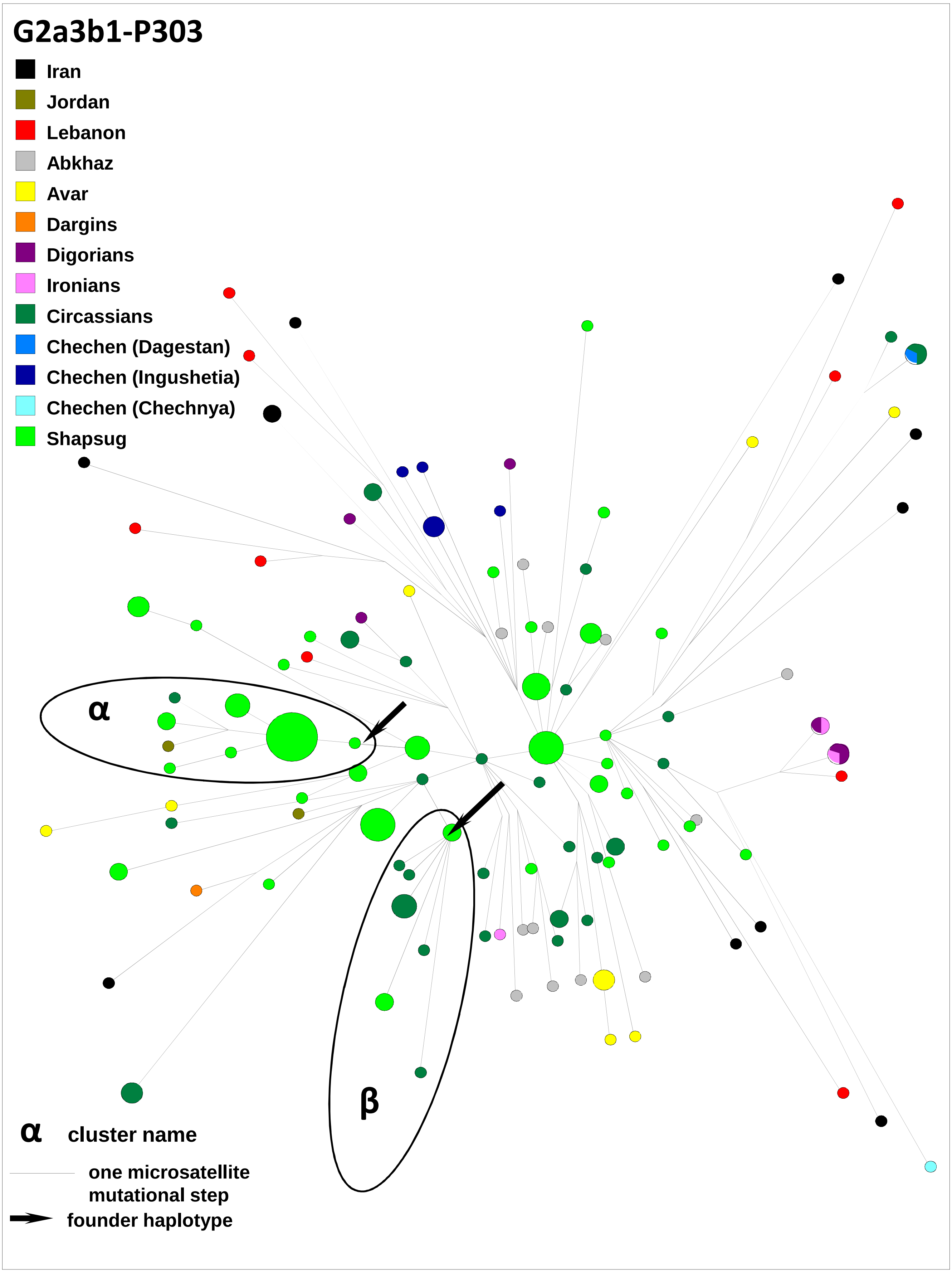
The second example is the split between the Shapsugs and Circassians. Both of these populations originated from the Adyghe community, which formed in the West Caucasus by the 10th century AD (Fedorov, 1982; Bertrozov, 1991) and consisted of many tribes. The ancestors of the Shapsugs and Circassians lived in adjacent areas of the lower Kuban River valley (Ageeva, 2000). In the 11th century, under pressure from incoming Turkic groups, many Adyghes migrated eastward to the Central Caucasus. When the Alans were defeated by the Mongols in this region, the Adyghes migrated further eastwards and gave rise to Kabardian and Circassian populations in the 12th-14th centuries (Fedorov, 1982; Ageeva, 2000). Therefore, historical data place the separation of Shapsug and Circassian populations between the 10th and 14th centuries, while the linguistic date is 14th century (700 BP) (Table 5, Supplementary Figure 2).

Genetic dating of this event is based on cluster P303-α, which is specific to the Shapsugs and absent from the Circassians. The genetic diversity of this cluster is ρ=0.55 (Table 5). A "genealogical" rate provides a date of 500±200 years BP, in agreement with history and linguistics (the cluster seems to emerge in the Shapsug after the ancestors of the Circassians migrated eastward). Using the “evolutionary” rate again gives too old of a date, 1,600±700 BP.

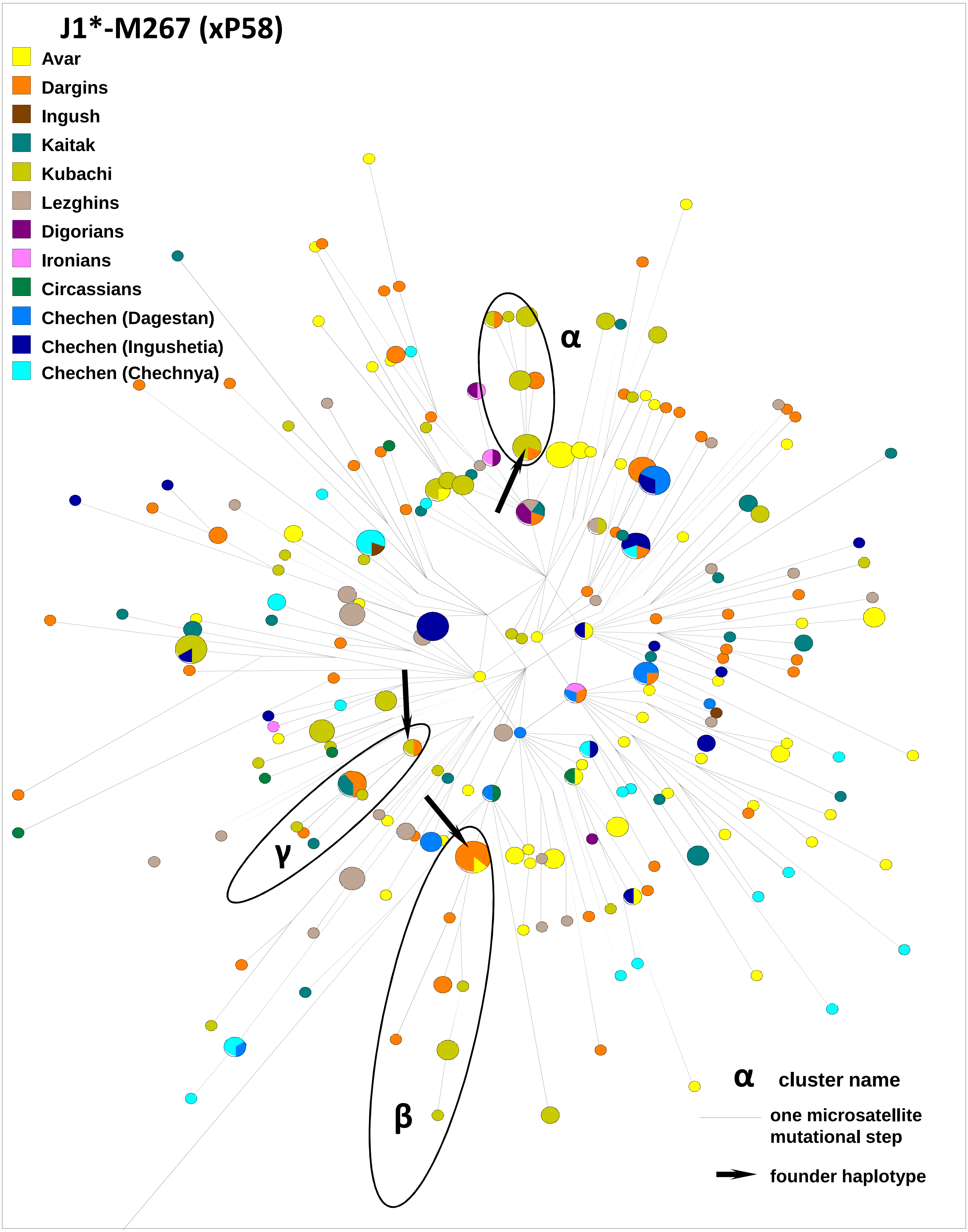
In addition, we compared genetic and linguistic evidence from the Nakh populations (Chechens and Ingushes). Their linguistic affinity is one of the most disputed questions for the North Caucasian languages. Some linguists consider Nakh and Dagestan languages as two branches of the East Caucasian group, while other linguists stress that variation within Dagestan languages is so significant that Nakh languages should be considered as one of many East Caucasian branches or even as part of Avar branch of the Dagestan languages (Alexeev, 1999). The linguistic tree obtained in our study (Supplementary Figure 2) supports the first interpretation. The genetic results (Figure 1) also clearly indicate the peculiarity of the Nakh paternal pool and its deep separation. Unfortunately, the lack of historical evidence from the East Caucasus did not allow us to compare it with our genetic and linguistic results.



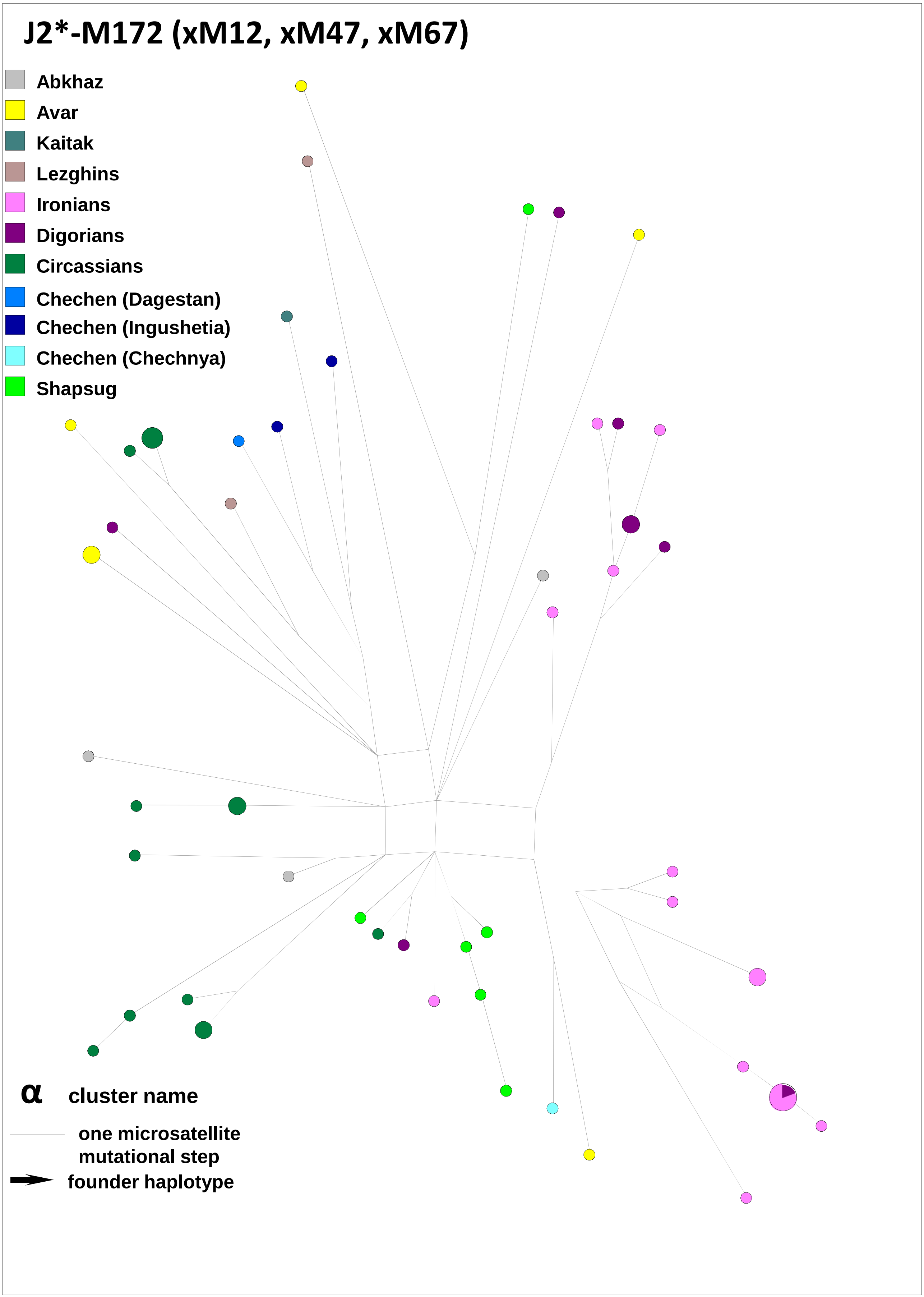
Supplementary Figure 1. Phylogenetic networks for the high-frequency Caucasus haplogroups. Haplogroup G2a*-P15(xP16,P303,M406).



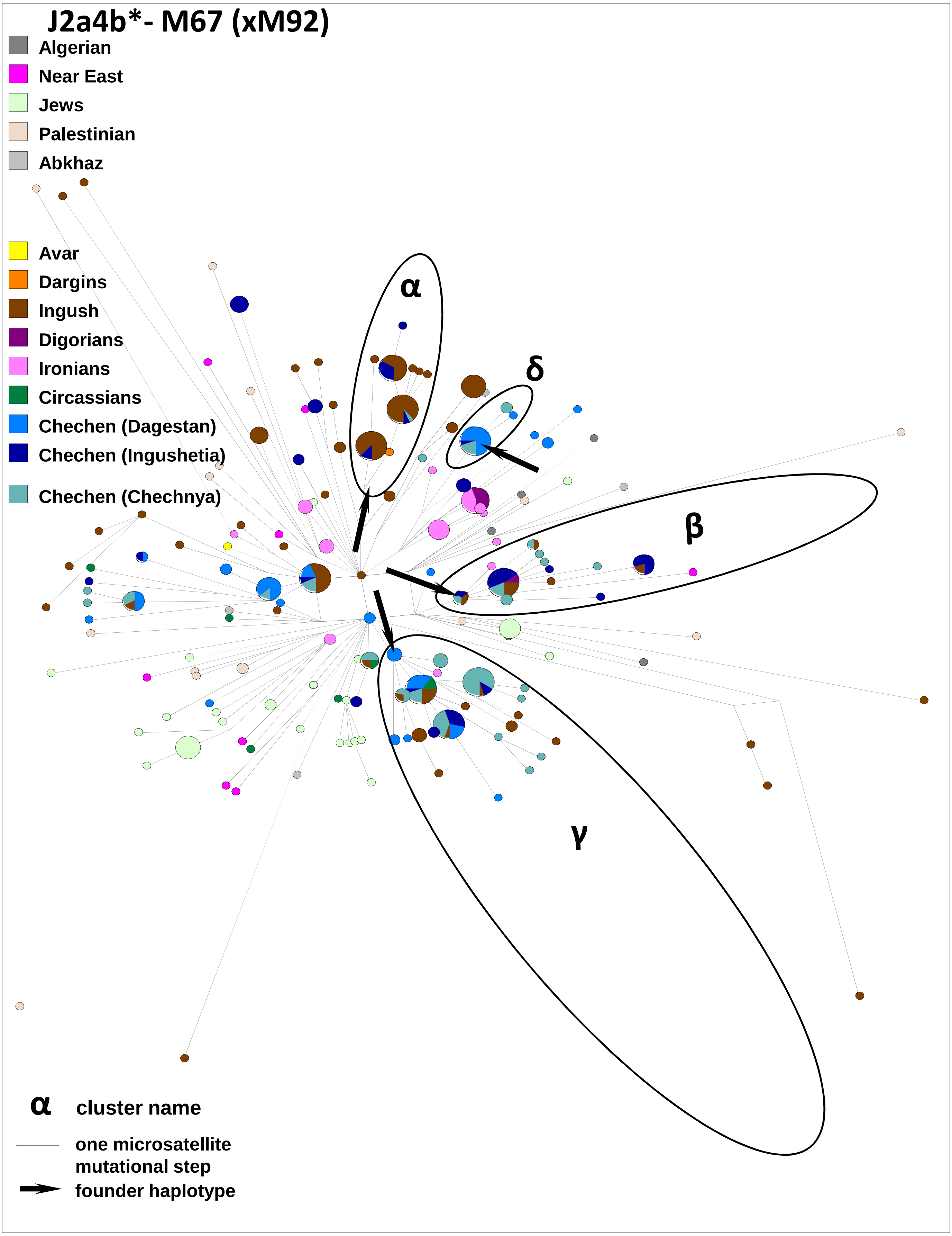
Supplementary Figure 1 (continued). Phylogenetic networks for the high-frequency Caucasus haplogroups. Haplogroup G2a3b1-P303



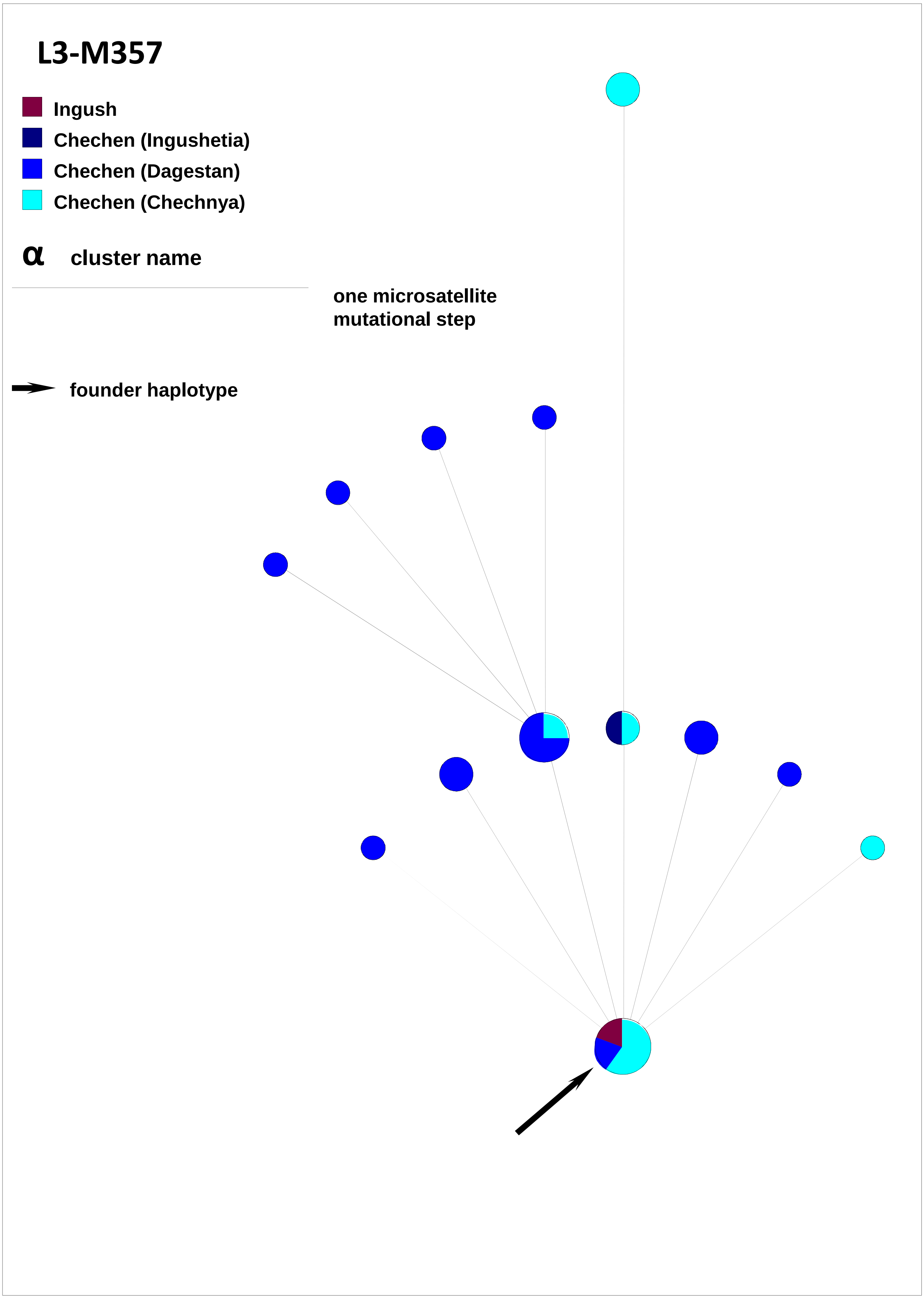
Supplementary Figure 1 (continued). Phylogenetic networks for the high-frequency Caucasus haplogroups. Haplogroup J1* - M267 (xP58)



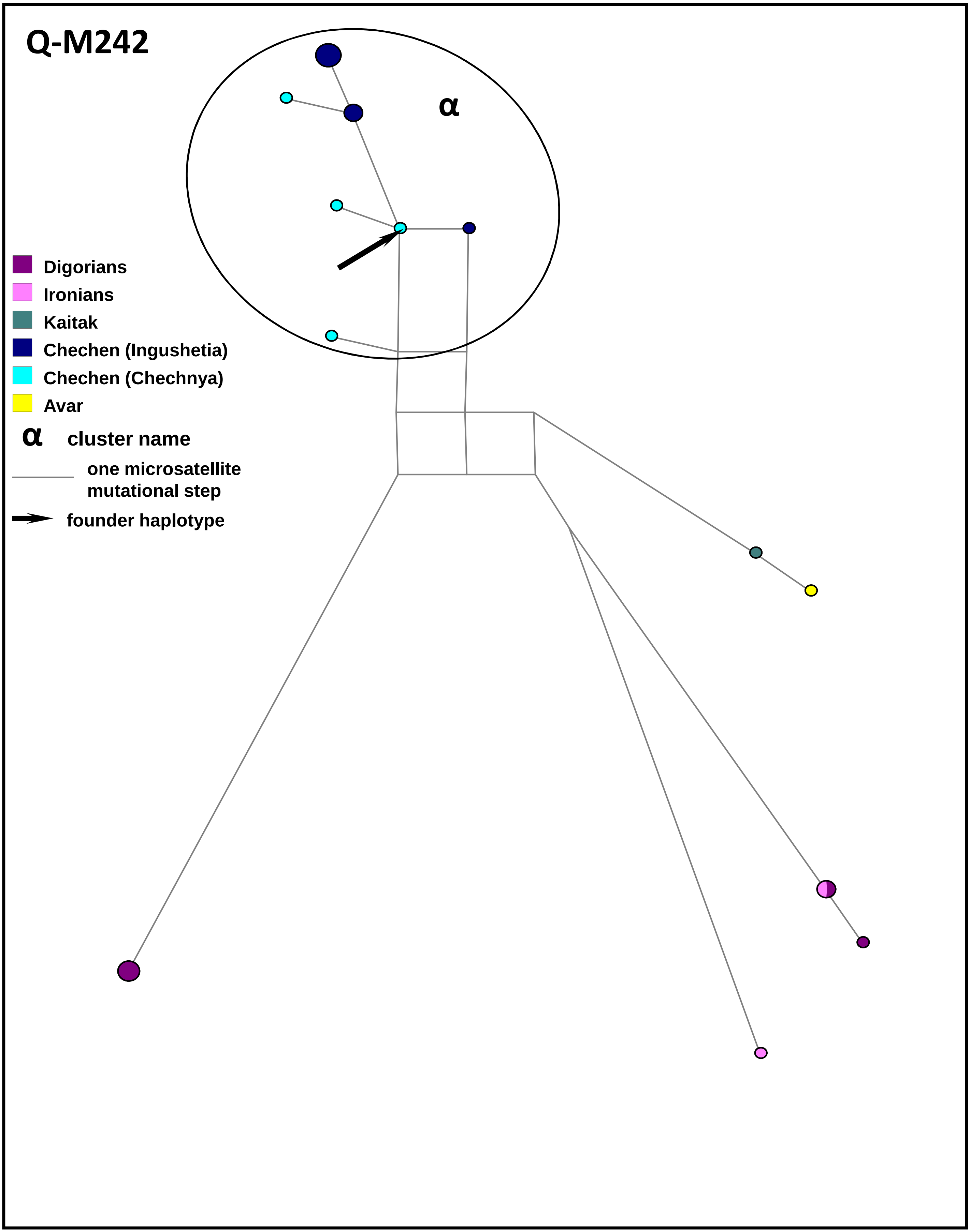
Supplementary Figure 1 (continued). Phylogenetic networks for the high-frequency Caucasus haplogroups. Haplogroup J2*-M172 (xM12, xM47, xM67)



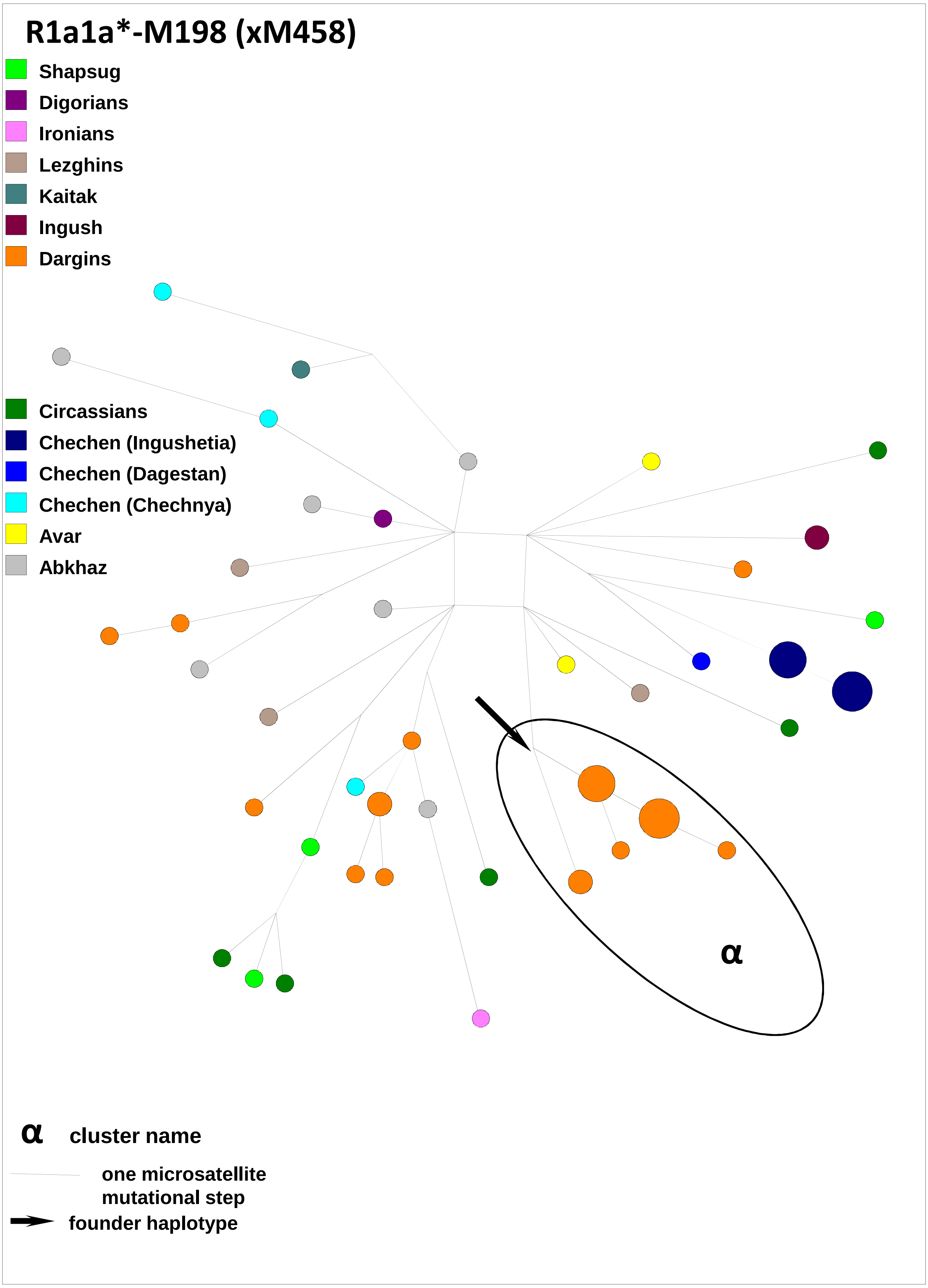
Supplementary Figure 1 (continued). Phylogenetic networks for the high-frequency Caucasus haplogroups. Haplogroup J2a4b-M67 (xM92)



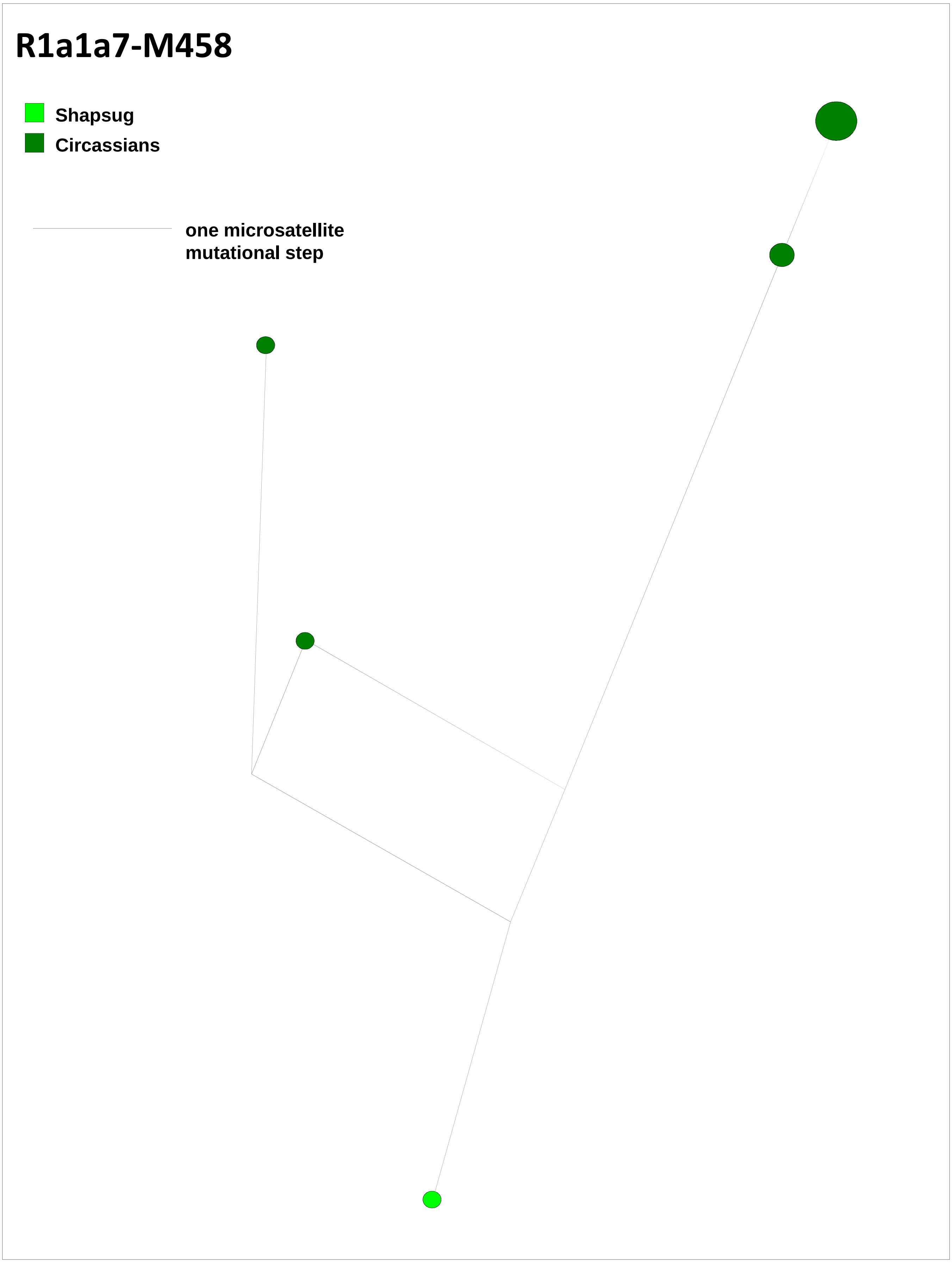
Supplementary Figure 1 (continued). Phylogenetic networks for the high-frequency Caucasus haplogroups. Haplogroup L3-M357



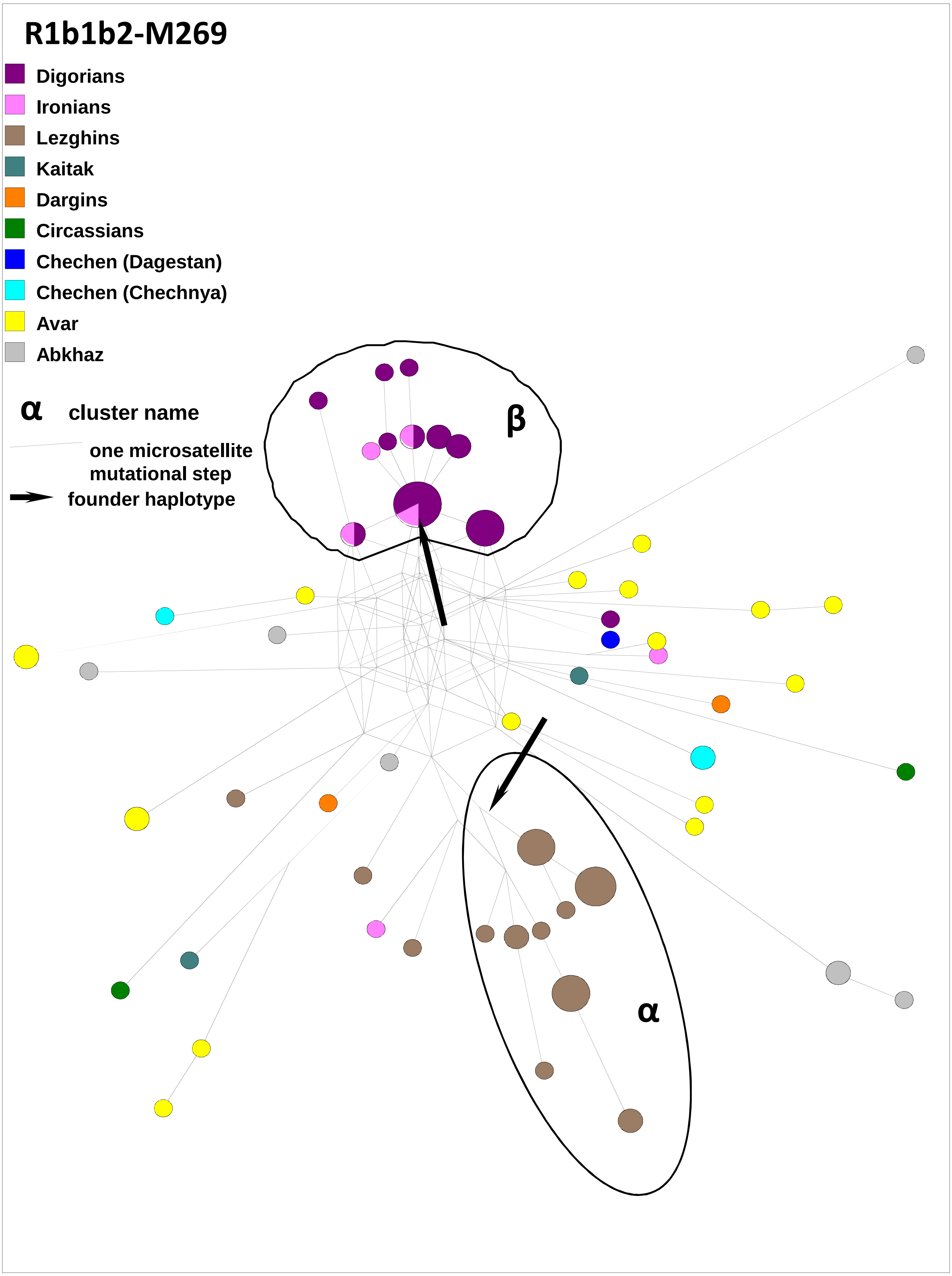
Supplementary Figure 1 (continued). Phylogenetic networks for the high-frequency Caucasus haplogroups. Haplogroup Q-M242.



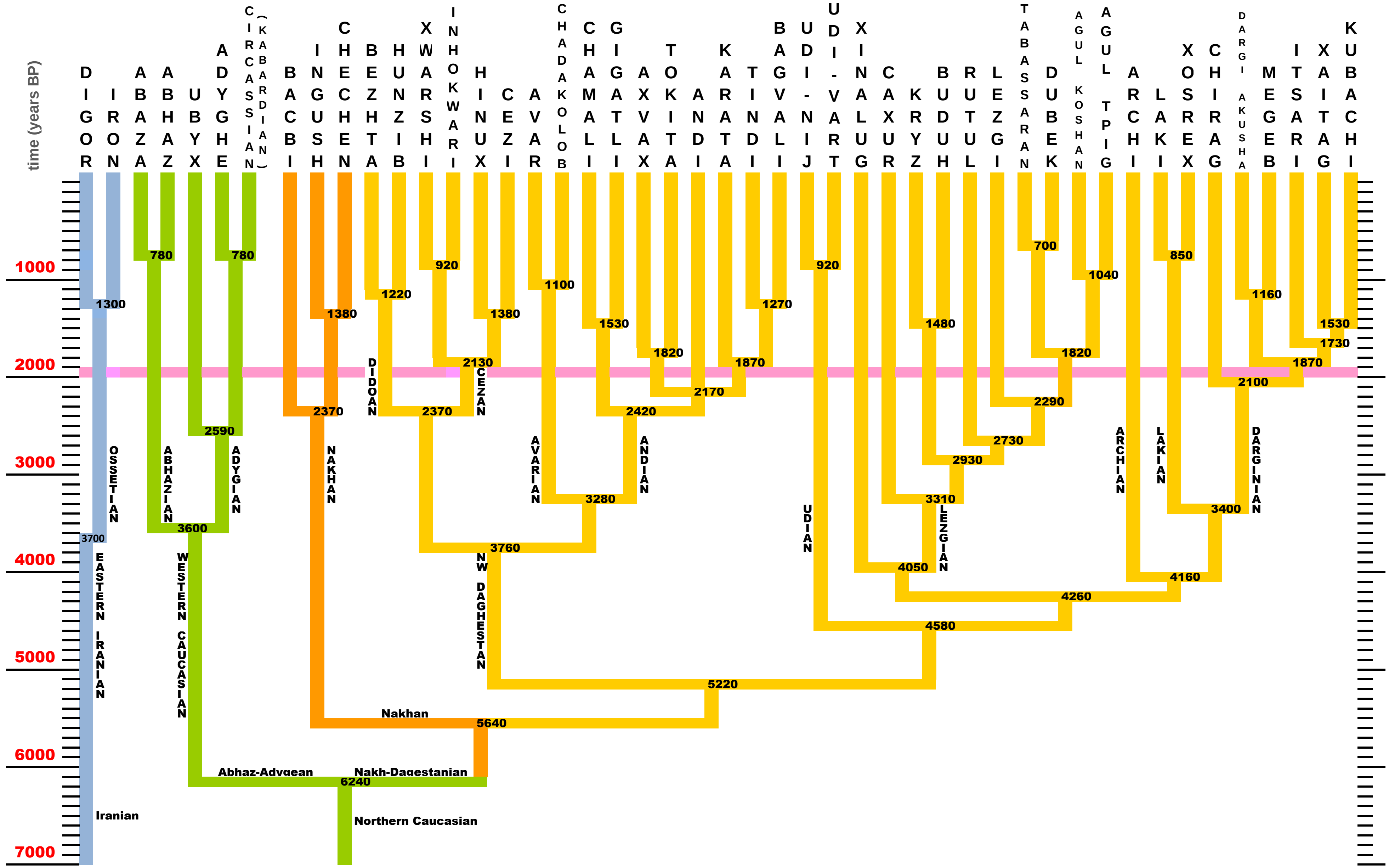
Supplementary Figure 1 (continued). Phylogenetic networks for the high-frequency Caucasus haplogroups. Haplogroup R1a1a*-M198 (xM458)



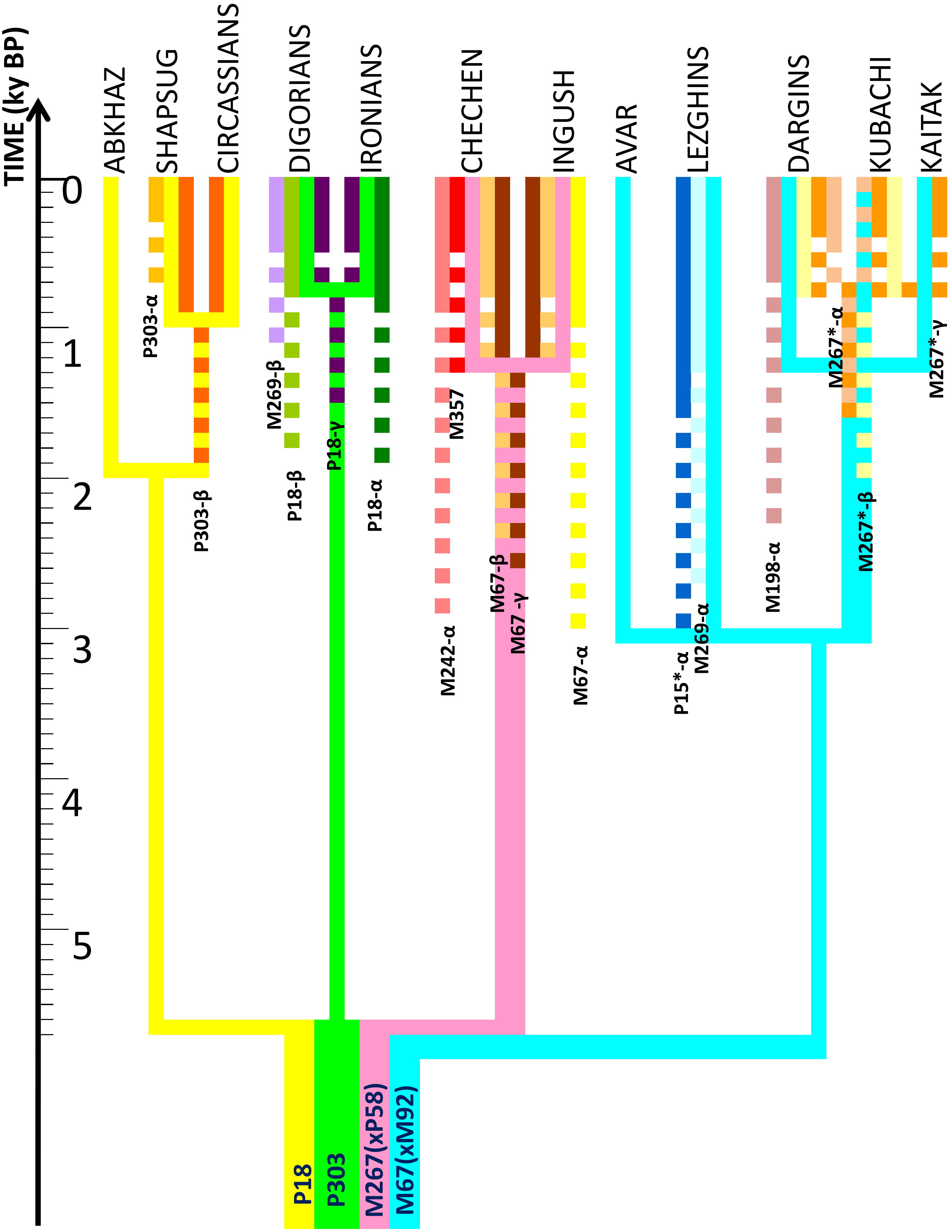
Supplementary Figure 1 (continued). Phylogenetic networks for the high-frequency Caucasus haplogroups. Haplogroup R1a1a7-M458



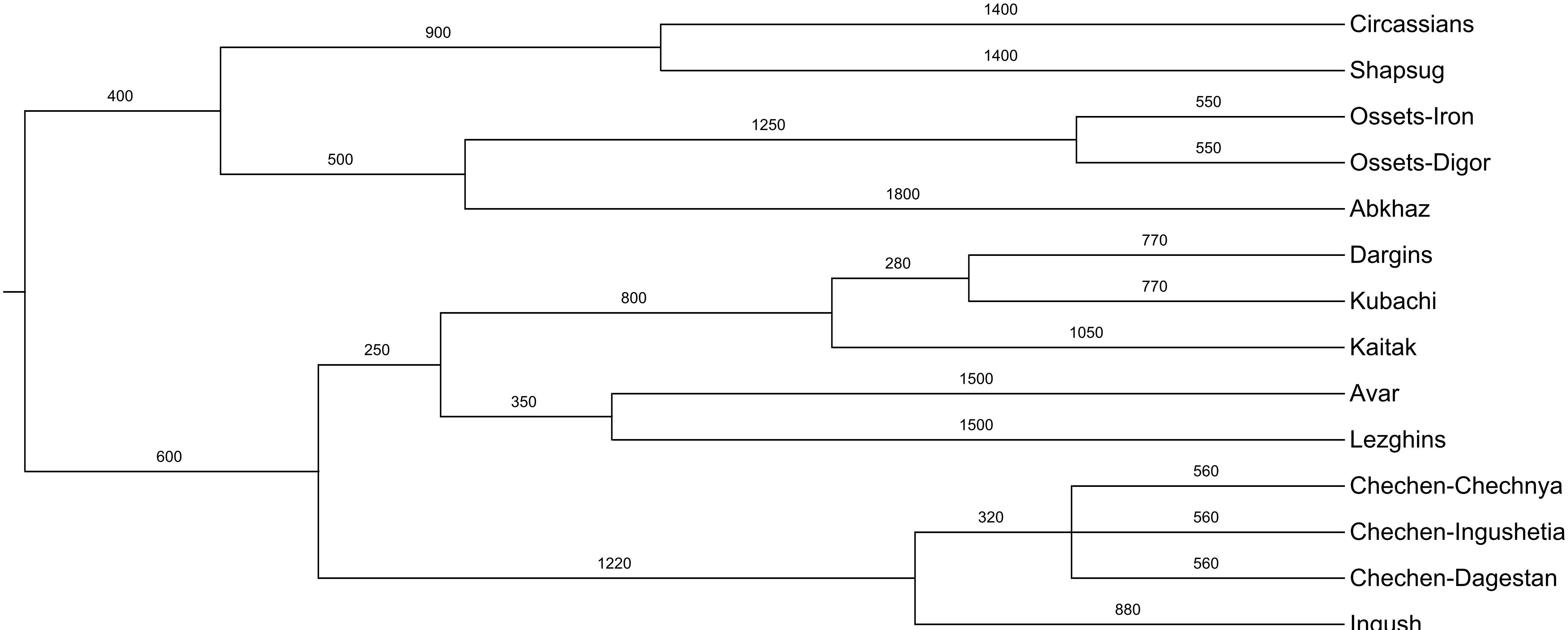
Supplementary Figure 1 (continued). Phylogenetic networks for the high-frequency Caucasus haplogroups. Haplogroup R1b1b2-M269



Supplementary Figure 2. Lexico-statistical reconstruction of the evolution of North Caucasus languages.



Supplementary Figure 3. Genetic reconstruction of the evolution of North Caucasus populations.



Supplementary Figure 4. BATWING tree of the population splits. Numbers on branches indicate time (in years) between corresponding splitting events.