AMERICAN JOURNAL OF PHYSICAL ANTHROPOLOGY 143:488–492 (2010)

Technical Note: A Method for Assignment of the Weight of Characters

M. Muzzio,[1]* J.C. Muzzio,[2] C.M. Bravi,[1] and G. Bailliet[1]

[1]*Laboratorio de Genética Molecular Poblacional, Instituto Multidisciplinario de Biología Celular (IMBICE), CICPBA, CCT La Plata-CONICET, Argentina*
[2]*Facultad de Ciencias Astronómicas y Geofísicas de la UNLP, Instituto de Astrofísica de La Plata (IALP), CCT La Plata-CONICET and UNLP, Argentina*

KEY WORDS character weight; mathematical scale; median-joining networks

ABSTRACT The weight of characters is a crucial step in different population analyses. We propose a new formula to facilitate this while establishing a scale that follows the criteria of the probability of change in each character. This method is described for drawing of median-joining networks, yet it could also be used for other methods in which the weight of the characters is required. Am J Phys Anthropol 143:488–492, 2010.

Median-joining networks (Bandelt et al., 1999) are widely used in molecular anthropology and population genetics, since they allow rebuilding phylongenies at an intraspecies level, while using nonrecombining population data and without solving "ties" between trees. The most used software system for drawing them is NETWORK (fluxus-engineering.com) that allows the weight of characters: the less likely events should be given a higher weight, considering that when they occur it is significant, while the most probable events should have a lower weight assigned. This step is crucial, because it allows giving differential importance to each marker when building the network. This way, the more plausible connections are showed, which makes the interpretation of the resulting network simpler.

In the present article, we propose a new system for weight assignment originally designed for Y-chromosome short tandem repeats (Y-STR) data that could also be employed for other types of traits, since it is based on the probability of a change of state. In the case of Y-STR data, the most used system was proposed by Qamar et al. (2002), which is employed by many researchers (Bolnick et al., 2006; King et al., 2007; Malhi et al., 2008). It consists of a fixed scheme of asymmetrical intervals, based on the variance of each STR, which gives weights from 1 to 5. These authors comment that their resulting networks contained a high amount of high-dimensional cubes (these occur when it is not possible to draw the network in a clear way and complicate its interpretation), so they also needed to use the reduced median algorithm before drawing the median-joining network.

Also, there are a few other complications that were not addressed by these authors: 1) Basing the weight on the STR variation requires having estimated it beforehand. 2) This variation could be reduced by drift processes, so we could find, in a given population—at least theoretically—a low variance to a STR that is highly polymorphic in others as a result of a recent bottleneck. 3) Even if the STR variances were estimated employing independent data bases (such as Y-Chromosome Haplotype Reference Database (YHRD), among others), when the single-step mutational model is considered, more specifically its characteristic of presenting a maximum and minimum of possible allele length, the variance of highly recurrent mutations could be underestimated. 4) It is a very restricted weight range that goes from 1 to 5, unpractical in cases where many different STRs are employed.

We propose the following formula for the assignment of weight per character:

$$W(x) = M(\text{mutational rate}_y/\text{mutational rate}_x)^a,$$

where $W(x)$ is the weight of a given STR or marker, M is the maximum weight value that the researcher decides to employ, mutational rate_y is the lowest mutational rate in the analyzed STR set, mutational rate_x is the mutational rate of the STR for which the researcher wishes to estimate the weight, and "a" is a value that enables the researcher to adapt the resulting scale to a desired range. To estimate it:

$$\text{Lowest weight value} = M(\text{mutational rate}_y/\text{mutational rate}_z)^a,$$

where mutational rate_z is the highest mutational rate in the analyzed STR set, which is conveniently put in a logarithmic form, linear in a, which is the unknown, and simplifies its computation:

Additional Supporting Information may be found in the online version of this article.

*Correspondence to: Marina Muzzio, IMBICE, 526 e/10 y 11, 1900 La Plata, Argentina, Postal Box C.C. 403.
E-mail: marinamuzzio@yahoo.com.ar

Received 13 November 2009; accepted 3 May 2010

DOI 10.1002/ajpa.21366
Published online 18 August 2010 in Wiley Online Library (wileyonlinelibrary.com).

$$\log(\text{lowest weight value}) = \log(M) + a\{\log(\text{mutational rate}_y) - \log(\text{mutational rate}_z)\},$$

So that

$$a = \{\log(\text{lowest weight value}) - \log(M)\} / \{\log(\text{mutational rate}_y) - \log(\text{mutational rate}_z)\}$$

For instance, if we decide to assign a minimum value of 3 and a maximum of 10, for the minimum haplotype whose lowest mutational rate is DNA Y-Chromosome Segment (DYS) 392 (of 0.45×10^{-3}) and highest is DYS 389 II (of 3.43×10^{-3}) (www.yhrd.org):

$$3 = 10(0.45 \times 10^{-3}/3.43 \times 10^{-3})^a$$

$$a = 0.593.$$

MATERIALS AND METHODS

To assess the quality of our scheme compared with the one of Qamar et al. (2002) or a uniform weight assignment, we employed published data of samples belonging to the E haplogroup from a Lebanese population (Zalloua et al., 2008). We chose the seven STRs from the minimum haplotype (Kayser et al., 1997): DYS 19, DYS 389 I, DYS 389 II, DYS 390, DYS 391, DYS 392, DYS 393, and used these data to calculate and draw three median-joining networks, employing the software NETWORK (fluxus-engineering.com), one per each weight assignment system. In the case of a network built with our weight assignment proposal, the mutation rates were obtained from the international database YHRD (www.yhrd.org), choosing a maximum of 10 and a minimum of 3, resulting in a weight value of 3 for DYS 389 II and DYS 391; 4 for DYS 19, DYS 389 I, and DYS 390; 7 for DYS 393; and 10 for DYS 392.

We estimated the amount of median vectors and maximum length to the node per network, and presence or absence of high dimension cubes, as indicators of parsimony (the less nodes and high-dimension cubes, the more parsimonious the network).

To analyze whether our weighting scheme reflected the data ancestry and did not force the results, we generated a series of haplotypes with a FORTRAN code created to build haplotype simulations. This code generates haplotypes starting on a given ancestor, with a given number of bearers, and for a given amount of generations, where different mutational rates can be specified for each locus. This code employs the RANLUX routine (James, 1994; Lüscher, 1994) for the random num-

TABLE 1. Simulated haplotypes employed in Figure 4

Lineage	Haplotype label	DYS19	DYS389I	DYS389II	DYS390	DYS391	DYS392	DYS393
A	A1	13	13	30	24	10	14	13
	A2	14	13	30	24	10	14	13
	A3	14	13	30	23	10	14	13
	A4	14	13	30	23	10	14	12
	A5	14	12	30	23	10	14	12
	A6	14	12	31	23	10	14	12
	A7	13	12	31	23	10	14	12
	A8	13	13	31	23	10	14	12
	A9	12	13	31	23	10	14	12
	A10	12	13	31	23	11	14	12
B	B1	13	13	30	24	10	14	13
	B2	13	13	30	24	9	14	13
	B3	14	13	30	24	9	14	13
	B4	14	13	30	24	10	14	13
C	C1	13	13	30	24	10	14	13
	C2	13	13	30	25	10	14	13
	C3	13	13	30	25	11	14	13
	C4	13	13	30	25	12	14	13
	C5	14	13	30	25	12	14	13
	C6	14	13	30	25	13	14	13
	C7	13	13	30	25	13	14	13
	C8	13	13	30	24	13	14	13
	C9	13	13	29	24	13	14	13
	C10	13	14	29	24	13	14	13
D	D1	13	13	30	24	10	14	13
	D2	13	13	29	24	10	14	13
	D3	13	13	30	24	10	14	13
	D4	12	13	30	24	10	14	13
	D5	12	13	30	25	10	14	13
	D6	12	13	31	25	10	14	13
	D7	12	13	31	26	10	14	13
	D8	12	13	31	26	10	13	13
	D9	12	14	31	26	10	13	13
E	E1	13	13	30	24	10	14	13
	E2	13	13	31	24	10	14	13
	E3	13	13	32	24	10	14	13
	E4	12	13	32	24	10	14	13
	E5	12	12	32	24	10	14	13
	E6	12	12	32	24	11	14	13
	E7	12	13	32	24	11	14	13

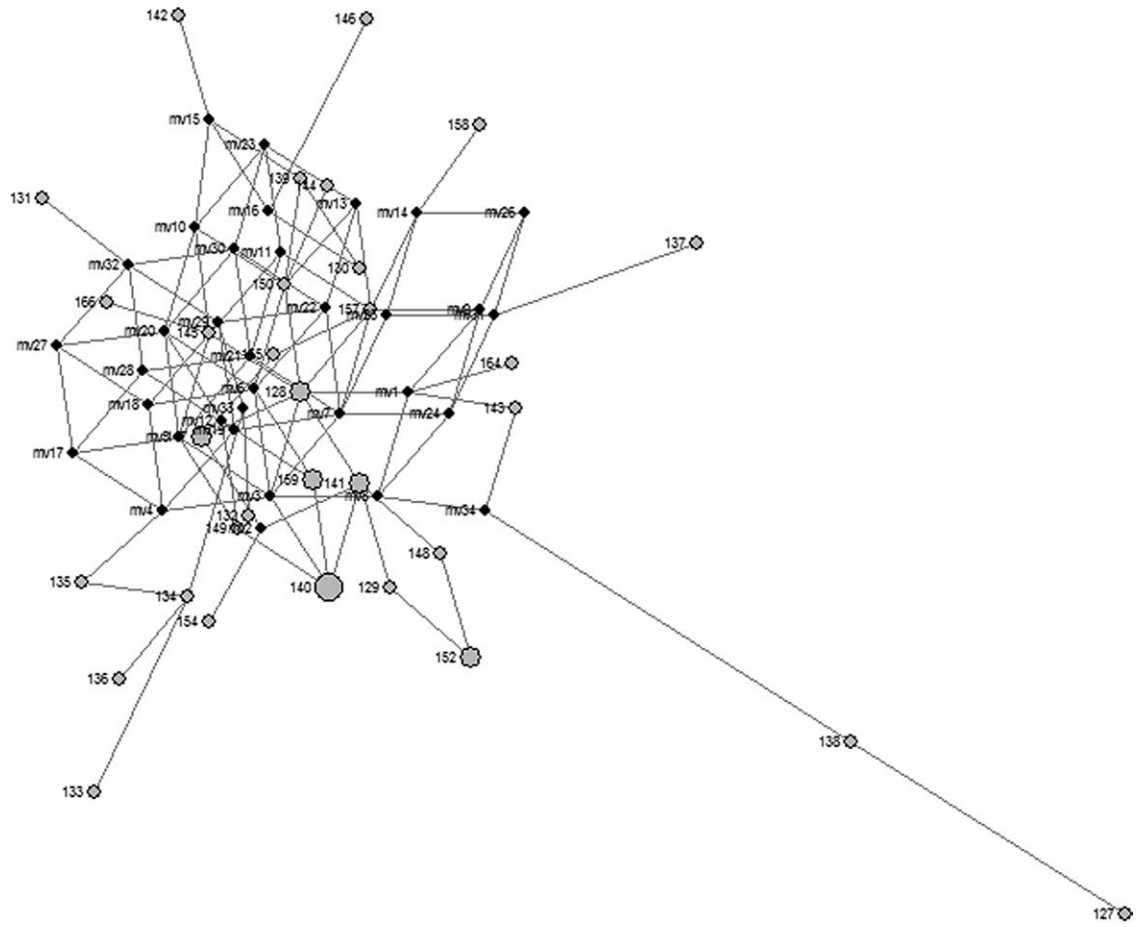

Fig. 1. Microsatellite network drawn with uniform weights. Black dots show the median vectors and gray dots the samples. Nodes are proportional to frequencies, and branches are proportional to mutational steps.

ber generation, and it is available at http://sites.google.com/site/haplosim/.

From the generated haplotypes, we randomly picked four sets of five lineages (each lineage in these cases includes all haplotypes from the founder to the final haplotype in the 70th generation), summarized in Table 1 and Table S1, and calculated networks with our weight proposal, to analyze whether the resulting graphics depicted correctly the relations of kinship or if our weighting scheme forced the data. Each lineage in these cases include all haplotypes from the founder to the final haplotype in the 70th generation, labeled with a letter (corresponding to the lineage) and a number (corresponding to the order of changes, i.e., "1" belongs to the founder, and it increases with each mutational step).

RESULTS

The resulting networks are represented by Figures 1–3. In Figure 1, we present the network calculated by assigning a weight of 10 for each STR, as the reader can see, a high amount of median vectors (in total 34), high-dimension cubes, and a maximum length of 10 mutational steps.

Figure 2 shows the network estimated with the weight assignment that follows the method of Qamar et al. (2002). There are eight median vectors, high-dimension shapes, and a maximum length to the node of 10 mutational steps.

Figure 3 presents the network computed and drawn with our proposal. There are a total of six median vectors, no high-dimension cubes, and a maximum length to the node of eight mutational steps. In all three networks, the ancestral haplotype remains the same, yet their topology is quite different.

The networks calculated with the simulated haplotype data from Table 1 are represented in Figure 4. Further networks calculated with simulated data are presented as supplementary materials (Table S1 and Figs. S1–S3). Each haplotype is labeled with a letter, representing the lineage, and the number that starts in 1 (the founder) and grows higher as it advances to the final haplotype. Clearly, these figures show that our weight scheme maintains the ancestor–descent relations.

DISCUSSION

Our proposal permits the researcher to obtain more parsimonious networks than those computed with other weight schemes, without employing pre- or postprocessing, since the resulting network contained less median vectors, a shorter maximum distance to the node, and no high-dimensional cubes or figures. The networks built with simulated data reflected the correct ancestry relations of the haplotypes employed, suggesting that our proposal provides the more parsimonious results without driving the data. In those cases where the haplotype order goes a step back, it is due to back mutations, as

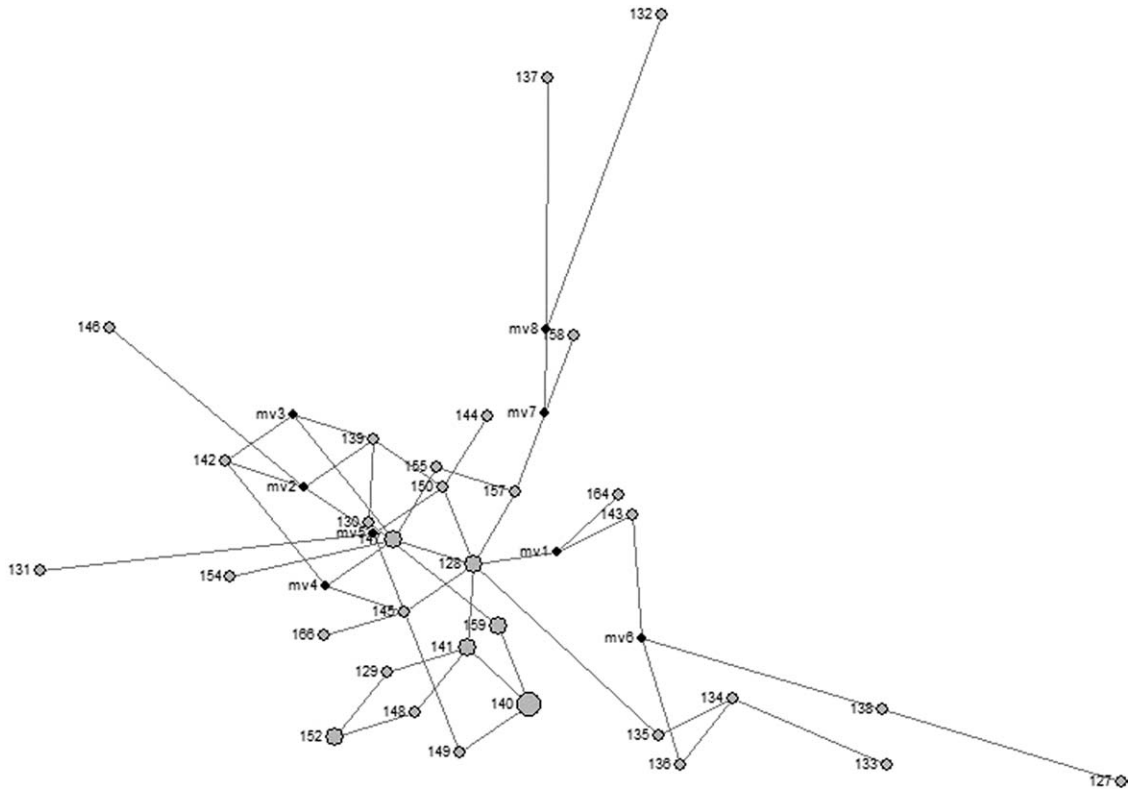

Fig. 2. Microsatellite network drawn with the weights established following the method proposed by Qamar et al. (2002). Black dots show the median vectors and gray dots the samples. Nodes are proportional to frequencies, and branches are proportional to mutational steps.

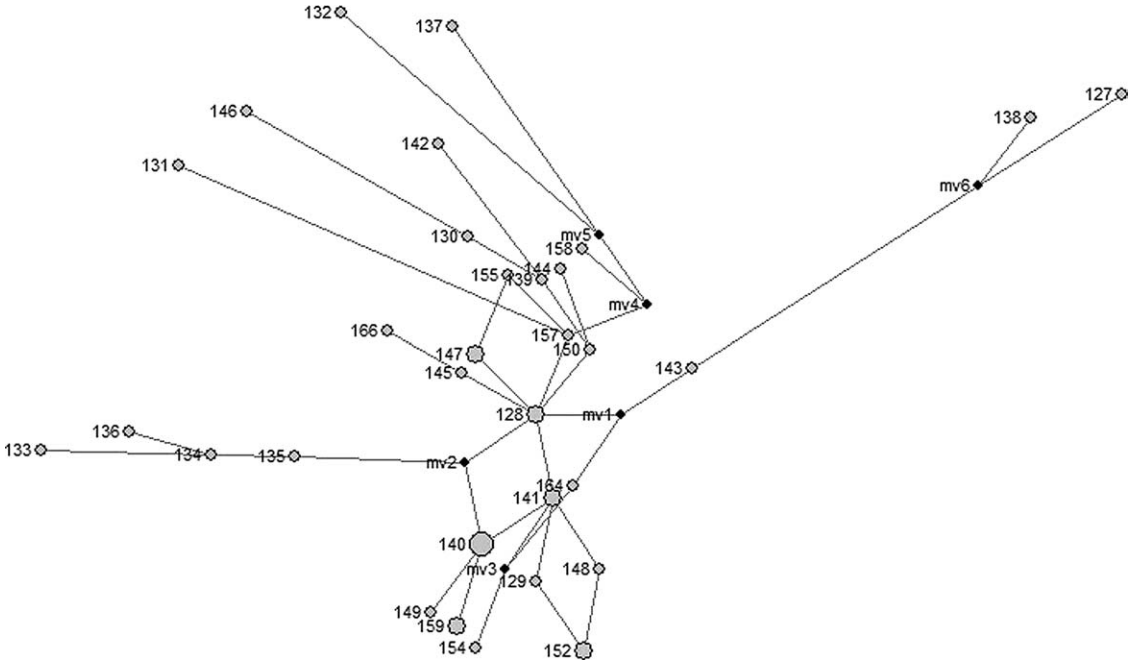

Fig. 3. Microsatellite network drawn with the weights established following the formula proposed in the present article. Black dots show the median vectors and gray dots the samples. Nodes are proportional to frequencies, and branches are proportional to mutational steps.

can be easily seen when analyzing each lineage individually by comparing the haplotypes on Tables 1 and S1. Note that all graphics (Figs. 1–4, Figs. S1–S3) were copied exactly from those produced by the software, without altering the topology in any way that could simplify or modify them.

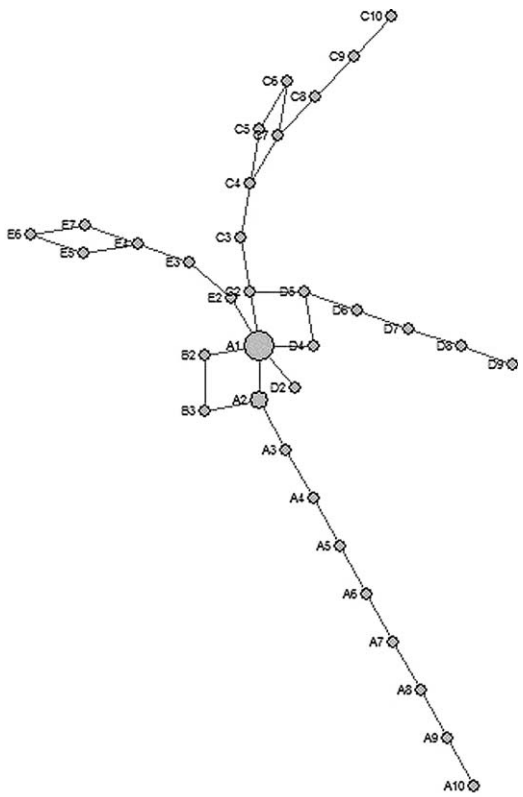

Fig. 4. Microsatellite network resulting from the first set of five lineages from the simulated data.

The proposed formula considers the mutational rates of the STRs, instead of their variance. Accordingly, those markers with a high mutational rate receive low weights, while those with low rates are assigned a high weight, since they are less probable events than the former. A mutation-based weighting scheme was originally proposed by Forster et al. (2000); however, their suggestion only consisted in establishing two weight classes, one with high mutation rates and another with lower rates, weighted by a factor of 2:1. Our formula allows a fan of weights that gives a finer resolution according the differences in mutation rates.

Anyway, when reliable mutational rates are not available, but STR variation is (which could be the case of newly found STR markers), the formula could be modified considering the last one instead of the former:

$$W(x) = M(\text{variation}_y/\text{variation}_x)^a$$

Also, the informativeness for assignment could be employed instead of variation, and probably with better results, when there are no reliable mutational rates. This ratio is estimated by following the method proposed by Rosenberg et al. (2003), with the software INFOCALC (Rosenberg et al., 2003). Of course, in this case, the formula would be modified accordingly replacing the lowest mutational rate with the highest informative value, and the highest mutational rate with the lowest mutational value.

Another advantage is that it allows the researcher to establish the range of values desired, through the estimation of a. This facilitates the analysis of multiple STRs simultaneously, since it allows using a larger range when the quantity of markers is high. In this article, we chose the minimum weight of 3 and a maximum of 10, because we employed the STRs from the minimum haplotype; nonetheless, the NETWORK (fluxus-engineering.com) software enables the selection of a minimum weight of 0 and a maximum of 100, so the possible range is quite broad.

The formula does not allow a minimum weight of 0 value, but that is not a problem because 0 weights are useless by definition. So if one has a variable with 0 weight value, all one has to do is to exclude it from the analysis and to take the variable with the second lowest weight as that with the minimum weight.

The flexibility of this formula allows its application on other markers and features for whose analysis a weight scheme is required. Simply, the researcher substitutes the mutational rates for a probability of a change of state for the characters of interest in the formula. Thus, an easy to calculate, yet mathematically precise, scale is available for different fields in anthropology and population genetics beyond molecular anthropology.

ACKNOWLEDGMENTS

The authors thank the reviewer of this article for his/her suggestions.

LITERATURE CITED

Bandelt HJ, Forster P, Röhl A. 1999. Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol 16:37–48.

Bolnick DA, Bolnick DI, Smith DG. 2006. Asymmetric male and female genetic histories among Native Americans from eastern North America. Mol Biol Evol 23:2161–2174.

Forster P, Röhl A, Lünnemann P, Brinkmann C, Zerjal T, Tyler-Smith C, Brinkmann B. 2000. A short tandem repeat-based phylogeny for the human Y chromosome. Am J Hum Genet 67:182–196.

James F. 1994. RANLUX: a Fortran implementation of the high-quality pseudorandom number generator of Lüscher. Comput Phys Commun 79:111–114.

Kayser M, Caglià A, Corach D, Fretwell M, Gehrig C, Graziosi G, Heidorn F, Herrmann S, Herzog B, Hidding M, Honda K, Jobling M, Krawczak M, Leim K, Meuser S, Meyer E, Oesterreich W, Pandya A, Parson W, Penacino G, Perez-Lezaun A, Piccinini A, Prinz M, Schmitt C, Schneider PM, Szibor R, Teifel-Greding J, Weichhold G, de Knijff P, Roewer L. 1997. Evaluation of Y-chromosomal STRs: a multicenter study. Int J Legal Med 110:125–133.

King TE, Parkin EJ, Swinfield G, Cruciani F, Scozzari R, Rosa A, Lim SK, Xue Y, Tyler-Smith C, Jobling MA. 2007. Africans in Yorkshire? The deepest-rooting clade of the Y phylogeny within an English genealogy. Eur J Hum Genet 15:288–293.

Lüscher M. 1994. A portable high-quality random number generator for lattice field theory simulations. Comput Phys Commun 79:100–110.

Malhi RS, González-Oliver A, Schroeder KB, Kemp BM, Greenberg JA, Dobrowski SZ, Smith DG, Resendez BM, Karafet T, Hammer M, Zegura S, Brovko T. 2008. Distribution of Y chromosome among Native North Americans: a study of Athapaskan population history. Am J Phys Anthropol 137:412–424.

Qamar R, Ayub Q, Mahyuddin A, Helgason A, Mazhar K, Mansoor A, Zerjal T, Tyler-Smith C, Mehdi SQ. 2002. Y-chromosomal DNA variation in Pakistan. Am J Hum Genet 70:1107–1124.

Rosenberg NA, Li LM, Ward R, Pritchard JK. 2003. Informativeness of genetic markers for inference of ancestry. Am J Hum Genet 73:1402–1422.

Zalloua PA, Xue Y, Khalife J. 2008. Y-chromosomal diversity in Lebanon is structured by recent historical events. Am J Hum Genet 82:873–882.