ARTICLES IN ENGLISH

Evaluation of the convergence of sets in STR phylogeny and analysis of the haplogroup R1a1 tree.

I. Rozhanskii

ABSTRACT

An approach has been developed to verify a convergence of Y-chromosome haplotype sets to single ancestors. This is a modification of the previously proposed method relying on the correspondence between the number of base haplotypes and the total number of mutations in the set [Klyosov, 2009]. The convergence parameter of the set is defined as the ratio of time spans to the (most recent) common ancestor (TSCA), calculated by logarithmic (from the number of base haplotypes) and linear (from the total number of mutations) methods, respectively. Parameters were calculated by using independent short fragments of extended (25 markers) haplotypes, and the average value was used for evaluation. This approach is able to employ relatively small number of extended haplotypes in order to estimate the convergence of trees to the single ancestor. The typical lower limit for evaluation is assumed as 20 haplotypes for 2000 years TSCA. The method is illustrated by examples from the phylogeny tree of R1a1 haplogroup.

INTRODUCTION

Anyone who ever calculated time spans to the common ancestors (TSCA) for Y-chromosome haplotypes knows that this is not a simple task. Apart from purely computational problems, it is very difficult in many instances to decide whether the set with the calculated base (modal) haplotype does converge to the single ancestor or there are several of them. This is the issue, because in the latter case the so-called «phantom» ancestor and incorrect TSCA are inevitably deduced. One can discard irrelevant haplotypes and select appropriate branches by analyzing the general structure of the tree, but this is not the general case. The tree itself appears sometimes either like a staircase or like an interwoven web, with heavily blurred boundaries between branches (Fig. 1). It is not easy to decide how to divide or combine closely positioned clusters. Decisions tend to be personally biased, that makes an analysis of large sets of extended haplotypes

increasingly unreliable, especially when the branches of the tree are not known *a priori*. There is a need in some independent method, which could allow us to estimate the means of division of the tree by branches.

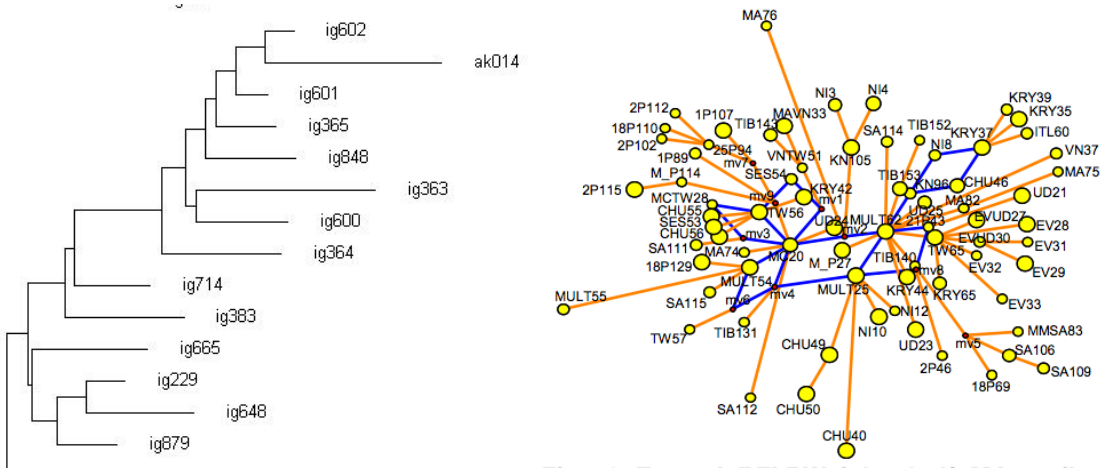

Fig. 1. Examples of trees in PHYLIP (left) and Network (right) software.

METHOD

The cornerstone assumption in computational methods of DNA genealogy states that mutations in the non-coding regions of Y chromosome are strictly stochastic. As a consequence, the following equations can be derived [Klyosov, 2009]:

$$M = \mu N_0 t \qquad (1)$$

$$N/N_0 = e^{-\mu t} \qquad (2)$$

with M being the total number of mutation from the base haplotype during time interval t in the set, N_0 corresponding to total number of haplotypes in the set, N defining the number of non-mutated (≡ base) haplotypes in the set, and μ being the average mutation rate per haplotype.

Equations (1) and (2) yield the apparent relation:

$$\ln(N_0/N) \sim M/N_0 \qquad (3)$$

It turns to equality under the following conditions:

- reasonably large set is considered,
- there is one, and only one common ancestor, and
- multi-step mutations are absent.

Since multi-step mutations are rather rare and tend to be statistically insignificant, one can omit condition (c), and relation (3) can be applied to verify the condition (b). It can be re-written for convenience as:

$$\nu = \ln(N_0/N)/(M/N_0) \qquad (4)$$

with ν being defined as the parameter of convergence of the set to one common ancestor. The closer its value to unity and the lesser it deviates from this upon addition or withdrawal of presumably related haplotypes, the higher is the probability to consider the particular set of haplotypes as descending from the single common ancestor. Neither of factors affecting TSCA values (reverse mutations, asymmetry, size of haplotypes, differences in individual mutation rates per marker) should affect the parameter ν, because the product μt remains the same both in (1) and (2).

This verification method can work efficiently, if condition (a) is fulfilled. That is, the set of haplotypes should be large enough. The critical issue is how to offer statistically significant number of base haplotypes, because their fraction falls exponentially as a function of time and length of the haplotype (see Eq. 2). For example, relatively young (28 generations by documental genealogy) Donald Clan retains now 25 % of its base haplotypes in 25 marker format, i.e. 21 from 84 [Klyosov, 2009], but its fraction will drop to mere 0.7 % after 100 generations from the ancestor. It is unlikely to find any base haplotypes in the set of the same size. If we limit our count by shorter 12 marker standard, we would expect 9.3 % of base haplotypes after 100 generation (7-8 from 84), that is enough for evaluation. However, the deviation of the number of base haplotypes just by one from the expected value (i.e., 6 or 9 from 84) would result in 15 % error in determination of the parameter ν. It is too rough to be of practical use. This error can be reduced in case of much larger sets, but it not always possible to collect enough haplotypes.

Where can we take more data, if the number of samples is limited? There is a simple solution – pick them up from the «cuts» of extended haplotypes, which have remained after employing their 1st panels (12 markers) in calculations. Since equations (1) and (2) are universal, they should be valid for any sequence of markers, not necessary standard. It is enough to obtain only 3 characteristics of the set - N_0, N and M, with tedious work on calibrating mutation rates being unnecessary for the scheduled task. In case of 25 marker haplotypes of FTDNA standard kit, this additional set can be compiled from haplotypes of the 2nd panel, considered as independent. Its base haplotype, the number of mutations and the convergence parameter can be calculated exactly by the same means as for the standard 12-marker panel. In an ideal case, parameters ν for both sets should coincide, because they belong to the same samples. In practice, they

differ, but their average should be closer to the true value, provided there is the single ancestor with corresponding base haplotype.

The number and composition of such «cut» haplotypes can be arbitrary. It appeared to be convenient to use 3, rather than 2 independent sets of markers, taken from the standard 25 marker panel.

Set 1: DYS 393, DYS 391, DYS 388, DYS 389-2, DYS 458, DYS 459b, DYS 437, DYS 464c, DYS 464d;

Set 2: DYS 390, DYS 385a, DYS 385b, DYS 426, DYS 389-1, DYS 392, DYS 447, DYS 464b;

Set 3: DYS 19, DYS 439, DYS 459a, DYS 455, DYS 454, DYS 448, DYS 449, DYS 464a.

The calculation of the convergence parameters for all sets was carried out by MS Excel. Simultaneous calculations of TSCA were performed by the linear method, corrected by reverse mutations [Klyosov, 2009].

RESULTS AND DISCUSSION

Prior to discussion of practical examples, one can note that the proposed method allows not only estimate the probability of the convergence of sets to single ancestors. It can be also useful in analyzing the character of deviations from uniform convergence if they are observed. Let's consider possible cases, as shown in Fig. 2.

Case A. If the set is statistically significant, uniform, and it converges to the single ancestor, it leads, by definition, to $v = 1$. This case can be imagined as a tree with its branches fitting the circle. The size of the trunk corresponds to the number of base haplotypes, whereas the length of branches represents the total number of mutations.

Case B. If the set contains haplotypes both from the main tree and from some remote unrelated branch, the total number of mutations appears to be overestimated compared with the number of base haplotypes. Accordingly, $v < 1$, and the corresponding double tree fits the oblate ellipse. This is rather frequent case in STR phylogeny, which happens if generous young branches "pull" the base haplotype to themselves. In many instances, such branches can be visibly recognized upon drawing the tree.

Case C. Branches of the uniform tree are overlapped with those of the nearby tree, which is not directly related to the former. The set looks rather compact by eye, but it contains less base haplotypes than expected. It results in $\nu > 1$, and the whole «composite» tree fits the prolate ellipse. This case appears frequently, when sets are composed by some artificial criteria, disregarding «natural» mutations. The particular example is sorting haplotypes by the fixed values of selected markers, which is popular among beginners. It is very difficult to recognize unrelated haplotypes in such sets. Sometimes it is more convenient to recalculate the whole tree, rather than to unravel the knot.

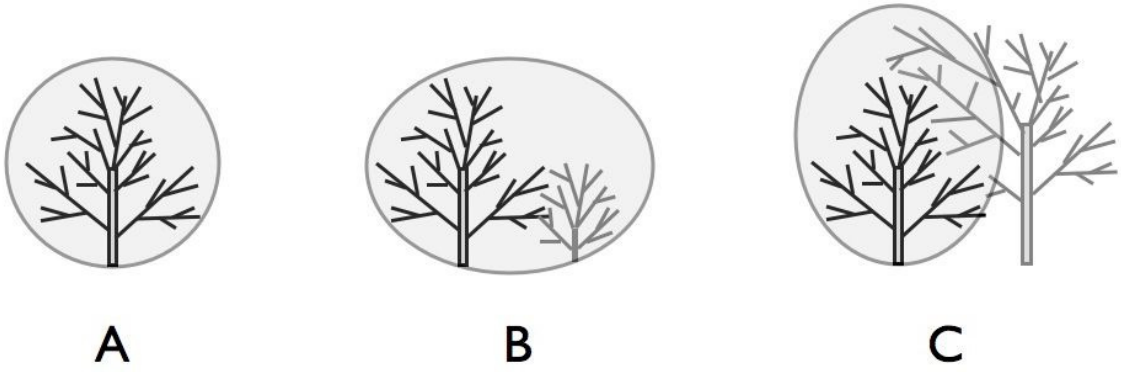

Fig. 2. Typical cases of the superposition of trees.

Even if $\nu = 1$, it is still important to confirm, that this is not accidental coincidence. If the set is large enough, one can apply rather strict test. The set is divided randomly by two parts, and convergence criteria and base haplotypes are calculated separately for both halves. If differences in their values are negligible, the whole branch can be considered as homogeneous.

If this test cannot be performed because of the limited size of the set, one can judge about stability of the tree indirectly, by comparing the base haplotype resulted from the optimization of ν and the modal one. If they differ markedly, it might be (but not always) a sign of the superposition.

Finally, there is a useful hint when statistically significant sets of 67 marker haplotypes are considered, with TSCA and errors being calculated separately for 25- and 67 marker panels. The more these values are overlapped, the higher is the probability for the set to be descending from the single ancestor.

The present method has been developed in the course of the analysis of haplogroup R1a1, for which any information on its SNP phylogeny was nearly absent to the moment of beginning of that work. Base haplotypes of branches and their geographical distribution were published [Rozhanskii and Klyosov,

2009; Rozhanskii and Klyosov, 2010], while TSCA have been recalculated considering the growth of the database in time. Only average values of ν were used in the analysis, without standart errors. The latters characterize mostly the mutation patterns in particular branches, rather than bear any statistical meaning.

Table. Examples of branches of R1a1 trees.

Branch	N_0	ν	TSCA (years)	
			25 markers	67 markers
Kyrgyz	79	1.01	875±120	900±110*)
Ashkenazi	67	0.97	1125±150	1100±140
Northern Eurasian	53	1.00	1875±240	1950±220
Northern Carpathian	31	1.01	2300±320	1975±250
Western Slavic	65	1.00	2800±330	2250±240
Central Eurasian-1	56	1.00	3625±420	3725±420
Young Scandinavian	138	1.06	2050±230	1975±210
same, parent sub-branch	99	0.99	1850±220	2050±220
Central European	110	0.96	3525±380	2775±290
same, sub-branch 1	43	1.02	2125±280	2275±250
same, sub-branch 2 (recLOH)	67	0.98	2425±290	2475±260
Cluster «K» (polish FTDNA project)	64	1.06	3525±400	3275±370
Cluster «K» borderline (polish FTDNA project)	36	0.82	4425±540	3575±390

*) – calculated for haplotypes in SMGF format (43 markers)

.

The first 6 branches show all signs of convergence: their parameters ν are close to 1.00, they are stable upon random dissection of sets, and their TSCA match each other when calculated by different sets of markers. This method works well in the wide time scale, both for Kyrgyz and Central European branches, in spite of their 4 times difference in TCSA. These branches correspond to the case "A" (Fig. 2).

Young Scandinavian branch is an example of the case "C" (Fig. 2). Its convergence parameter deviates significantly from 1.00, while the base haplotype of the branch appears to be rather unstable, since it switches between several optima upon random dissection of the set. These are visible signs of superposition. Indeed, the tree of this branch is not entirely homogeneous, because it contains a younger compact sub-branch (so-called "Scottish" cluster), which adds some more mutations or even can "usurp" the base haplotype if outnumbers the rest of the set. When this younger sub-branch was withdrawn,

the remaining ("parent") branch immediately gave good convergence to the single ancestor.

The next example is the Central European branch, which seems to fit the case «B» (Fig. 2), with a superposition of parent and daughter sub-branches. However, this is another, non-standard case. More than half of the haplotypes (67 from 110) bear a complex mutation known as "recombinational loss of heterozygosity" (recLOH). One can find more information about this mutation elsewhere (http://freepages.genealogy.rootsweb.ancestry.com/~langolier/krahn.pdf). In relevance to the present subject, it is important that it takes place in pair markers of palindrome regions of DNA and its probability is less than for the most of STR markers, but more than for SNP. The specific mark of recLOH in Central European branch is "doubling" of alleles in the quadruple marker DYS464a-d. For example, the present author (as well as dozens of his neighbours by this branch) bears alleles DYS464a-f **12-12-15-15**-15-16 instead of characteristic for R1a1 sequence **12-15**-15-16. Formally, these are 4 consecutive mutations. In fact, this is a single, albeit rare event. Apparently, it is necessary to make correction to recLOH in order to obtain undistorted TSCA and base haplotypes.

This branch was divided by two parts, and parameters of both subsets were calculated separately. Both sub-branches showed good convergence parameters and TSCA, close to each other. Counting recLOH as a single mutation, one can calculate the genetic distance between two base haplotypes as 4 in 67-marker format. It puts the common ancestor of both sub-branches at 2725±400 years before present, which matches closely TSCA obtained for the same branch using another dataset [Underhill et al., 2009; Klyosov and Rozhanskii, 2009], that is 2550±360 ybp. Therefore, Central European branch shows patterns of the case «A», representing the homogeneous tree with the single common ancestor. Probably, recLOH mutation occurred at the very beginning of the history of this branch, and both sub-branches developed in the same fashion, producing virtually indistinguishable haplotypes. At the best of author's knowledge, it is one of very few examples when recLOH defines a genealogical line of such temporal, demographical and geographical scale.

Finally, the last two branches have been taken not from our review article [Rozhanskii and Klyosov, 2009], but from the website of the Polish FTDNA project. Data of R1a1 are represented there as clusters, which are composed according to the published method [Gwozdz, 2009]. These data give us the opportunity to examine how the present method competes with the other algorithms of the evaluation of genealogical lines (clusters, by somewhat cautious definition used in the project). A list of 64 haplotypes, assigned by the project administrator to the cluster «K», was treated by the same way as previously considered branches. This set shows close TSCA for 25 and 67

markers panels, while its base haplotype matches very closely that of the Western Eurasian branch (in fact, this is one of the most represented genealogical lines in Poland and surrounding countries). However, the convergence parameter of this set (v = 1.06) suggests, that some unrelated haplotypes can be still present there. It is very difficult to verify this suggestion and to recognize "intruders" by relying entirely on the method used in that project. It is an example of the case «C», with the aforementioned consequences.

Another set of 36 haplotypes has been defined in the project as belonging to the "cluster K borderline". Although such definition is rather obscure, let's consider this set as a separate genealogical line and treat it similarly. It becomes apparent (see the Table above), that this is nothing but a superposition of loosely bound fragments with «phantom» ancestor. Assignment of haplotypes in this set is unreliable and needs refinement by some other methods.

CONCLUSIONS

Examples listed above do not mean that the present method is able to displace existing algorithms used in STR phylogeny. It cannot produce a tree, but it is valuable as a tool for really **independent** verification of various approaches, which should yield genealogical lines descending from single ancestors. Since this approach deals with arrays of data, it is not very informative in positioning individual haplotypes on the tree. By the same reason, evaluations of smaller sets (typically, less than 20 25-marker haplotypes) are less reliable. However, this method is considered as complementary to other techniques of DNA genealogy, providing additional opportunities in research.

ACKNOWLEDGEMENTS. The author expresses his gratitude to A. Klyosov for the encouragement in developing this method and for fruitful discussions. Special thanks to P. Shvarev, who first recognized and characterized branches of R1a1. The very subject of this work would be impossible without his contribution.

References

Gwozdz, P. (2009) Y-STR Mountains in Haplospace, Part I: Methods. J. Genetic Geneal (ISSN 1557-3796), **5**, No. 2, 137 - 159.

Klyosov, A.A. (2009) DNA Genealogy, Mutation Rates, and Some Historical Evidence Written in Y-Chromosome, Part I: Basic Principles and the Method. J. Genetic Geneal (ISSN 1557-3796), **5**, No. 2, 186 - 216.

Klyosov, A., and Rozhanskii, I. (2009). Subclade R1a1a7-M458 – populations, geography, history. Proc. Russian Academy of DNA Genealogy (ISSN 1942-7484), **2**, No. 7, 1200 – 1216 (in Russian).

Rozhanskii, I., and Klyosov, A. (2009). Haplogroup R1a1: haplotypes, genealogical lines, history, geograpgy. Proc. Russian Academy of DNA Genealogy (ISSN 1942-7484), **2**, No. 6, 974-1099 (in Russian).

Rozhanskii, I., and Klyosov, A. (2010). Migrations from Southern Siberia and Central Asia to the Northern Europe from the viewpoint of DNA genealogy. Proc. Russian Academy of DNA Genealogy (ISSN 1942-7484), **3**, No. 1, 66-77 (in Russian).

Underhill, P.A., Myres, N.M., Rootsi, S., Metspalu, M., Zhivotovsky, L.A., King, R.J. et al (2009) Separating the post-Glacial coancestry of European and Asian Y chromosomes within haplogroup R1a. Eur. J. Human. Genet., advance online publication, 4 November 2009, doi: 10.1038/ejhg.2009.194

Web resources

Description of mutations in Y chromosome
http://freepages.genealogy.rootsweb.ancestry.com/~langolier/krahn.pdf

Polish FTDNA project
http://www.familytreedna.com/public/polish/default.aspx?section=yresults