International Congress Series 1288 (2006) 201–203

www.ics-elsevier.com

Making the most of Y-STR haplotypes: The HapYDive

C. Alves [a,*], L. Gusmão [a], J. Meirinhos [a], A. Amorim [a,b]

[a] *Instituto de Patologia e Imunologia Molecular da Universidade do Porto (IPATIMUP), Porto, Portugal*
[b] *Faculdade de Ciências da Universidade do Porto, Porto, Portugal*

Abstract. The HapYDive is the latest version of a software devised to evaluate the increase of haplotype diversity (HD) by the addition of any combination of Y-STRs to a fixed number of markers. Created in Excel format (available at www.ipatimup.pt/app/), it is not only a software for Y-STR HD calculation but, more importantly, it allows the determination of which combination of Y-STRs is the most informative in a certain population sample. With the HapYDive it is possible to analyse any set of Y markers up to a maximum of 20, with a minimum number of 4 markers fixed for calculations. Results on the application of this program to different population samples and sizes and with a certain combination of Y-STR markers are presented and discussed, together with its usefulness mainly to the forensic community.

Keywords: HapYDive; Y-STR combination; Haplotype diversity

1. Introduction

Since the informative power of a Y-STR marker can only be recognised in a haplotype context, a software was devised to evaluate the increase of haplotype diversity (HD) by the addition of any combination of markers to a fixed number of markers. The first version of the program [1] was quite limited and not very user-friendly. The HapYDive is the latest version, created in Excel format (available at www.ipatimup.pt/app/). It not only calculates Y-STR HD but, more importantly, it allows the determination of which combination of Y-STRs is the most informative in a certain population sample. With the HapYDive it is possible to analyse any set of Y markers up to a maximum of 20, with a minimum number

* Corresponding author. Tel.: +351 22 5570700; fax: +351 22 5570799.
E-mail address: calves@ipatimup.pt (C. Alves).

0531-5131/
doi:10.1016/j.ics.2005.09.034

of 4 markers fixed for calculations. The program's features may not only be useful to commercial companies when choosing sets of Y-STRs for multiplex kits, but also for smaller forensic labs with lower fundings which are only now beginning to work with these markers. For anthropological and evolutionary studies, it can also provide the best markers for population differentiation.

Here we show some results when applying the HapYDive to different population samples and discuss the effect of population size.

2. Material and methods

The program for the study on the informativeness of all possible Y-STR combinations was created for Microsoft in Excel file format. It addresses the following points: it can be used on populations from any region or size; it combines all possible results for any Y-STRs; in a given sample it determines the number of different haplotypes and calculates HD for each set of Y-STRs. It can be obtained from the web page www.ipatimup.pt/app/.

The HapYDive was applied on non-related male individuals previously typed for 17 Y-STRs (DYS19, DYS385, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, DYS439, DYS460, DYS461, GATA A10, DYS635 and GATA H4) sampled from different origins (Portugal, $N=657$ [2], and Mozambique, $N=112$ [1]).

3. Results and discussion

The results shown for all samples consider a specific set of Y-STRs that are fixed, namely the currently used in the "YHRD—Y Chromosome Haplotype Reference Database" (http://www.yhrd.org), comprising the extended haplotype or SWGDAM Core Set (minimal haplotype markers, 9 loci, plus DYS438 and DYS439). Therefore, the HapYDive analyses that we are reporting concern only the behaviour of all the other Y-STRs on HD increments (DYS434, DYS437, DYS460, DYS461, GATA A10, DYS635 and GATA H4). The question is which set and what number of these other available Y-STRs will increase more rapidly the HD value?

As shown before [1], the choice of markers depends on population origin. Table 1 shows the HapYDive results obtained from Portugal ($N=657$) and Mozambique ($N=112$). In the Portuguese sample, all the other Y-STRs contribute to a certain degree to an increment of HD, but DYS460 contributes the most and DYS635 the least. However, for Mozambique the order is different and there is one marker, DYS437, that does not contribute in any way to a higher HD.

Table 1

(a) HapYDive data for Portugal ($N=657$)		(b) HapYDive data for Mozambique ($N=112$)	
Y-STR sets	HD	Y-STR sets	HD
11 Y-STRs	0.99771	11 Y-STRs	0.99212
11 Y-STRs+DYS460	0.99866	11 Y-STRs+GATA A10	0.99437
12 Y-STRs+GATA H4	0.99915	12 Y-STRs+DYS460	0.99582
13 Y-STRs+GATA A10	0.99936	13 Y-STRs+DYS635	0.99646
14 Y-STRs+DYS461	0.99950	14 Y-STRs+DYS461	0.99695
15 Y-STRs+DYS437	0.99960	15 Y-STRs+GATA H4	0.99727
16 Y-STRs+DYS635	0.99966	16 Y-STRs+DYS437	0.99727

Table 2
Average HapYDive data obtained from sub-sample groups created randomly from the original Portuguese sample of 657 individuals

Y-STR order	Group 1 (N=219)	Group 2 (N=438)	Group 3 (N=525)	Total (N=657)
1st	GATA H4	DYS460	DYS460 or GATA H4	DYS460
2nd	DYS460	GATA H4	GATA H4 or DYS460	GATA H4
3rd	GATA A10	GATA A10	GATA A10	GATA A10
4th	DYS437	DYS461	DYS461	DYS461
5th	DYS461 or DYS635	DYS635	DYS635 or DYS437	DYS437
6th	DYS635 or DYS461	DYS437	DYS437 or DYS635	DYS635

What about sample size? If we compare the HapYDive results obtained with this sample from Portugal (N=657) with the previous smaller sample in [1] (N=208), the picture is slightly different, since the order of choice does not coincide. However, it is clear that in both sample sizes the best three Y-STRs to be added are DYS460, GATA H4 and GATA A10, whilst the remaining three are less informative, although in different orders.

Randomly re-shuffling the 657 Portuguese individuals into smaller sub-populations and applying the HapYDive was our next trial. First, a division into six random sub-samples consisting of 219 individuals (Group 1) was tested, then three random sub-samples of 438 individuals (Group 2), and finally five random sub-samples of 525 individuals (Group 3). The results obtained for each sub-sample of these three groups varied considerably from each other. However, if for each group an average of the results is taken, it becomes clear again the consistency observed before: the set DYS460, GATA H4 and GATA A10 is the most informative, whilst the set DYS461, DYS437 and DYS635 is less, although the order of choice varies inside both sets (Table 2).

It seems perceptible that the informativeness of Y-STRs is sample size dependent. Indeed, when using a large number of Y-STRs, it is expected to find a high number of unique haplotypes, whose estimated frequencies tend to present a high standard deviation. In a small sample size, it is also predictable that a large number of haplotypes will not be represented. As demonstrated in this work, large samples are needed to discriminate markers with similar impact on haplotype diversity. Nevertheless, it is also clear that even for small samples (N~200) it is possible to select groups of markers with more or less informativeness in a certain haplotype context. Since very distinct Y-STR orders of preference were obtained from each sub-sample within each group, we consider the sub-sampling strategy important in order to get a more reliable picture of which markers are really making the most of the Y-STR haplotypes you may have.

Acknowledgements

Financial support was granted by Fundação para a Ciência e Tecnologia, Programa Operacional Ciência, Tecnologia e Inovação (POCTI).

References

[1] C. Alves, et al., Evaluating the informative power of Y-STRs: a comparative study using European and new African haplotype data, Forensic Sci. Int. 134 (2–3) (2003) 126–133.
[2] S. Beleza, et al. Micro-phylogeographic and demographic history of Portuguese male lineages. Ann. Hum. Genet. (in press).