Int J Legal Med (2008) 122:435–440
DOI 10.1007/s00414-008-0233-7

TECHNICAL NOTE

The SNP*for*ID browser: an online tool for query and display of frequency data from the SNP*for*ID project

Jorge Amigo · Christopher Phillips · Maviky Lareu · Ángel Carracedo

Received: 7 May 2007 / Accepted: 13 March 2008 / Published online: 20 May 2008

Abstract The SNP*for*ID browser is a web-based tool for the query and visualization of the SNP allele frequency data generated by the SNP*for*ID consortium (http://www.snpforid.org/). From this project, validated panels of single nucleotide polymorphisms (SNPs) for a variety of forensic applications have been generated with the browser concentrating on the single-tube identification SNP set comprising 52 markers. A web interface allows the visitor to review the allele frequencies of the studied markers from all the available populations used by SNP*for*ID to validate global SNP variability. The interface has been designed to offer the useful facility of combining populations into appropriate geographic groups for visual comparison of populations individually or amongst user-defined groupings and with equivalent HapMap data.

Keywords SNP · SNP*for*ID · Online databases · Forensic allele frequency databases · HapMap

Accessibility: web access to this tool is granted at http://spsmart.cesga.es/snpforid.php

Electronic supplementary material The online version of this article (doi:10.1007/s00414-008-0233-7) contains supplementary material, which is available to authorized users.

J. Amigo (✉) · C. Phillips · Á. Carracedo
Spanish National Genotyping Center (CeGen) and Genomic Medicine Group, CIBERER,
University of Santiago de Compostela,
Santiago de Compostela, Spain
e-mail: jamigo@usc.es

C. Phillips
e-mail: chrisp@usc.es

M. Lareu · Á. Carracedo
Insitute of Legal Medicine, Genomic Medicine Group,
University of Santiago de Compostela,
Santiago de Compostela, Spain

Introduction

The SNP*for*ID consortium was set up in 2003 to develop single nucleotide polymorphisms (SNP) loci for use in human identification analysis: principally focused on forensic analysis but encompassing relationship testing (e.g., paternity analysis, confirmation of pedigree, etc.), enhanced prediction of geographic origin and medical sample identification. The main requirement from novel forensic marker sets, hitherto lacking in short tandem repeat loci (STRs), is the ability to successfully genotype highly degraded DNA without dropout: the differential loss of loci or alleles caused by PCR fragment sizes above ~125 bp or resulting from large differences in repeat number within a locus. For this reason, SNP*for*ID prioritized SNP sets that could be genotyped from amplified fragments generally below 100 bp and in multiplexes sufficiently large to provide equivalent, or better, discrimination power to the widely used 16-STR kits. A core 52 SNP multiplex has been developed for forensic analysis comprising loci primarily targeted from the p-arm and q-arm of each autosome [13]. This has been supplemented with SNP sets that allow the prediction of the geographic origin of a sample [9], enhanced characterization of the Y chromosome [3] and typing of haplotype-informative coding region SNPs in the mitochondrial genome [2].

An important aspect of the work of the consortium has been the promotion of an open source ethos for reporting the technical aspects of the SNP typing assays developed and the scientific findings together with the provision of tools to analyze SNP genotype data. The SNP*for*ID browser falls into the third category—an online tool that permits any researcher with genotyping data for the 52 SNPs in the forensic marker set to obtain allele frequency estimates from populations relevant to their own analyzes. The data is

presented in such a way that it is easy to collect and, when required, to combine allele frequency estimates from several populations into groups that better represent continental groups or geographic regions. HapMap allele frequency estimates from the four phase I study populations can also be listed to assist comparisons with the SNP*for*ID populations and as a benchmark for assessing the reliability of the estimates for each locus.

Two examples serve to illustrate the potential use of a frequency browser tool and highlight the flexibility of a combinational approach to reviewing SNP allele frequency data. In the first hypothetical example, a forensic laboratory in a nonurban region of northern Canada might wish to interpret a SNP profile by obtaining the frequency for the genotypes in both European and Inuit populations. Although Inuit data is available from the SNP*for*ID browser, Canadian European population data is lacking but could be adequately substituted with the combined European data readily obtained from the frequency page, allowing the investigator to report to court two appropriate cumulative frequency estimates for comparison. In the second, real case, a challenging paternity analysis of closely related individuals in Galicia (NW Spain) required SNP typing as a supplement to STR analysis. The investigator used the browser to compare and contrast Galician population estimates with various combinations of European populations and to obtain relevant frequency data permitting the assessment of the degree of local variation compared to European-wide patterns of variability for the 52 SNPs. In interpreting paternity analysis data involving related individuals, it is particularly important to gauge the degree of variability in the family investigated, the local population, and continent-wide to properly assess the significance of the genotypes and paternity indices obtained.

Data curation

Although it is easy to provide an open access website accessing the full set of population validation genotypes available, more power is provided by constructing a web tool that can read directly from a database of combinable data. Designing and programming a suitable search web tool with emphasis on visualization of allele frequencies became our main priority. From the start, it was decided that access to individual genotypes or sample profiles would not normally be required by forensic users and that data from multiple centers that can be combined or compared provides more flexibility. As such, each genotype is not particularly important as a single entity, but is considered as a whole when the allele frequency estimates are calculated from the query of joined databases. This does not preclude the possibility of SNP*for*ID centers genotyping standardized control samples such as those from the Coriell cell repositories (http://ccr.coriell.org/ccr/) or the positive control DNA supplied with standard forensic STR typing kits, then listing such profiles for each of the SNP sets developed by the consortium. Furthermore, the complete dataset of 52 SNP genotype profiles from all the populations listed (outside of HapMap) that underlie the allele frequency estimates are available as a flat text file download for each selected population, allowing user-defined analyses such as tests for independence or intrapopulation and interpopulation F_{st}.

The SNP*for*ID project represents the sum of efforts from six laboratories spread across Europe, so all the genotyping data generated required curation before being combined. A simple format database was created to form the basis for joining all available data and to allow for future developments that can also work from the same data. Data was indexed by sample and contained information of contributing laboratory, gender, population of origin (ascertained from the donors' declaration of their immediate ancestry), and 52 SNP genotypes. The curation process that checks data quality encompasses scrutiny of GeneMapper ID or Genotyper output from SNaPshot genotyping submissions made outside the SNP*for*ID laboratories plus assessment of Hardy–Weinberg equilibrium using chi-squared analysis together with F_{st} measurements comparing new populations with those of the same group.

A minor logistical problem in the initiation of the database was the collation and standardization of the data into a single repository. The binary nature of autosomal SNP data makes this process much easier than with multiple allele and haploid polymorphic loci utilized elsewhere in forensic science population databases like the Y Chromosome Haplotype Reference Database (YHRD, [10]) and Mitochondrial DNA Control Region Database (EMPOP, [7]). Therefore, all bases were inverted when necessary (e.g., CT base calls converted to AG) to match those listed in the Santa Cruz genome browser summary of dbSNP reference SNP data. Heterozygote genotypes were alphabetized and the locus listing order was, by previous convention, p-arm SNaPshot electrophoretic mobility (Auto1 SNPs) then q-arm (Auto2). Because all contributing laboratories used SNaPshot for the validation of populations, base standardization anticipates future submissions to the database from alternative genotyping platforms. To check genotyping quality, chi-squared analysis was made of the observed and expected genotype ratios in all populations having sufficient numbers of samples, although this had been previously performed on similar data to study interlaboratory concordance [13]. In addition, SNP*for*ID allele frequency estimates for African, European, and East Asian population groups were compared to those from the equivalent population panels of HapMap (termed Yoruba

from Ibadan, Nigeria [YRI]; CEPH Utah residents with European ancestry [CEU], and ASN, respectively, with ASN representing a panel of Chinese from Beijing [CHB] and Japanese from Tokyo [JPT] populations combined [1]).

Implementation

The web tool has been written in PHP and HTML, and it acts as an interface to the underlying database, an example of which is shown in Fig. 1. It was designed to go beyond text queries, and so certain graphical aids were developed to address this need. The first query point is a browsable world map allowing the visitor to locate each studied population and obtain frequency data with a single click. We used our own customized version of the DIY Map [5], a clickable zooming map written in Flash and configurable through an XML file providing the ability to not only spot the population locations and their population groups, but also to implement simple queries activated directly through clicks.

The graphical system of the data summary returned from the query provides visitors with a flexible and intuitive approach to the scrutiny of allele frequencies from single populations and in comparison to combinations of populations, enhanced to allow comparison of results using two different queries in parallel. This search system established itself as the main core of the application because all the possible queries that visitors were predicted to run had to be included together with the ability to preempt incorporation of future submissions of new populations or SNP sets to the database. As a result, the database is dynamically updated at the point in time each query is made, so the search page contains all current available data once it has been checked, curated, and incorporated. The same real-time updating process applies to the HapMap frequency data that is included in the data summary when available (48 out of the 52 SNPs have now been characterized by HapMap). In summary, the SNP data obtained from a query will always provide the most current frequency estimates for each SNP*for*ID and equivalent HapMap population: updated in real-time at the moment the query is made.

In keeping with the clean, easily interpreted pie chart summaries of SNP variability used successfully in the HapMap genome browser [14], we have mirrored the same approach in the pie charts used to visualize frequencies for each SNP*for*ID population or their combination, although actual allele frequencies are also listed as numeric values alongside the pie charts in the search return page. Charts display blue segments denoting the reference allele and red segments denoting the alternative allele with frequencies charted from 0.01 to 0.99. It is important to note two elements of the HapMap pie chart approach: (1) the reference allele segment is positioned counterintuitively on the left side of the zero point, i.e., from −3.6° (0.01 frequency) to −356° (0.99) and (2) triallelic SNPs that are now also in the browser as part of the ancestry-informative SNP sets from SNP*for*ID [8, 9] and were not included in the 1.1 million phase I SNPs characterized by HapMap. Therefore, the convention we propose to adopt for triallelic SNPs is to add a green segment for the third allele, denoting the least frequent allele observed in Africans and so likely to be the most recently derived substitution at the SNP.

Results

Depending on the options chosen for a search, the pie charts plotted in the query return page represent allele frequency estimates calculated from single populations or

L	15019	L-15019	M	Nigeria	AFRICA	CT	CT	AA	TT
L	16410	L-16410	M	Nigeria	AFRICA	CT	CC	AT	AA
L	16417	L-16417	M	Nigeria	AFRICA	TT	CT	AA	TT
L	16744	L-16744	M	Nigeria	AFRICA	CT	CC	AA	AA
L	16314	L-16314	M	Britain	EUROPE	TT	CC	AT	AT
L	16402	L-16402	M	Britain	EUROPE	CC	CT	TT	AT
L	16854	L-16854	M	Britain	EUROPE	CT	CC	AT	AT
L	16873	L-16873	M	Britain	EUROPE	TT	CC	TT	TT
L	17347	L-17347	M	Britain	EUROPE	CT	CT	AT	AA
L	17677	L-17677	M	Britain	EUROPE	CC	TT	AT	AT
L	17863	L-17863	M	Britain	EUROPE	CT	CC	AT	AT
L	18086	L-18086	M	Ireland	EUROPE	CT	CT	AT	AT
L	18123	L-18123	M	Ireland	EUROPE	TT	CC	AA	AA
L	18126	L-18126	M	Ireland	EUROPE	TT	CT	AA	AT
L	18147	L-18147	M	Ireland	EUROPE	CT	CC	TT	AT
L	18150	L-18150	M	Ireland	EUROPE	CC	CC	AA	TT
L	18244	L-18244	M	Ireland	EUROPE	TT	CT	AT	AT
L	18250	L-18250	M	Ireland	EUROPE	CT	CC	AT	AT
L	18274	L-18274	M	Ireland	EUROPE	CC	CT	AT	AT
L	18277	L-18277	M	Ireland	EUROPE	TT	TT	AA	AT
L	18280	L-18280	M	Ireland	EUROPE	CT	CT	AT	AT
L	18322	L-18322	M	Ireland	EUROPE	TT	CC	AT	AA
L	18999	L-18999	M	Britain	EUROPE	CT	TT	AA	TT
L	19021	L-19021	M	Ireland	EUROPE	CT	TT	AT	AA
L	19073	L-19073	M	Ireland	EUROPE	CC	CT	AA	AT

Fig. 1 Example snapshot from the joined SNP*for*ID database. Entry columns denote, from *left* to *right*, originating center; center sample ID; SNP*for*ID sample identifier; gender, population of origin; population group; and genotypes (A01–A54 in the same order of SNPs as search page top to bottom, allowing a direct transposition from a curated Excel file to the database)

their combinations as a single column for the search option plus multiple columns for up to four user-defined comparisons. Five population groups are summarized in a set of pie charts using the grouping of populations outlined in the search page listings that Fig. 2 shows. This grouping is based on a previous study of global variability that found a close match between geographic distribution of populations and genetic clustering using STRUCTURE to arrange populations into groups based on patterns of variability [11]. Using the same clustering algorithm for the 52 SNPs and 9 of the validation populations in the browser gave a broadly similar grouping within the confines of a much smaller range of loci and study populations (Fig. 3, K=4 in [13]). The separate listing of the South Asian population sample to those from Europe is a potentially contentious arrangement because populations from this region tended to cluster with other Eurasian populations from Europe, North Africa, and the Middle East in the Rosenberg study; however, the browser allows this population sample to be included with the six European populations or analyzed separately so the added flexibility provided is worth retaining, particularly as additional South Asian populations are likely to be sampled and submitted to the browser. This last point also illustrates the potential of a combine-and-compare approach in studying differences between populations because the pie charts provide an intuitive system for visualizing the contrasting allele frequency distributions found in some of the SNPs in the 52-SNP set. Such SNPs comprise about 10% of the full set and were chosen deliberately to provide indicators of geographic origin in the same way STR data can be used for this purpose [6]. Therefore, it seems likely that the use in the near future of dedicated sets of ancestry-informative SNP sets including those of SNP*for*ID [9] will also benefit from the system of allele frequency visualization adopted for this browser.

To statistically assess the goodness of fit of allele frequency estimates from SNP*for*ID and HapMap, an r^2 analysis was performed on appropriate population groupings matched to the HapMap study panels described above. ESM Fig. 1 presents an analysis of allele frequency estimate correlation between SNP*for*ID and HapMap genotyping for 48 of 52 SNPs analyzed in common. Goodness of fit between the paired datasets was assessed using r^2 analysis of appropriate SNP*for*ID study population groupings matched to the HapMap study panels: (a) European (EUR vs CEU), (b) East Asian (ASN vs combined CHB/JPT), and (c) African (AFR vs YRI). The listed r^2 values indicate good correlation of SNP*for*ID and HapMap frequency estimates for all loci and each pair of population groups.

As an illustration of the standard display features of the browser, a dataset of samples from Spain and Mozambique is illustrated in ESM Fig. 2 because both populations represent a data subset that can be readily compared to their continental-based population groups of Europeans and Africans, respectively. ESM Fig. 2 illustrates a complete query result for NW Spain and Mozambique with summary population-group pie charts showing allele frequency data for each SNP and the equivalent HapMap estimates when present. The SNP*for*ID population-group pie charts are designed to match the order of HapMap charts: EUR/CEU (SNP*for*ID European/HapMap CEPH European from Utah of northern and western European ancestry), ASN/CHB+JPT combined (SNP*for*ID East Asian/combined Chinese from

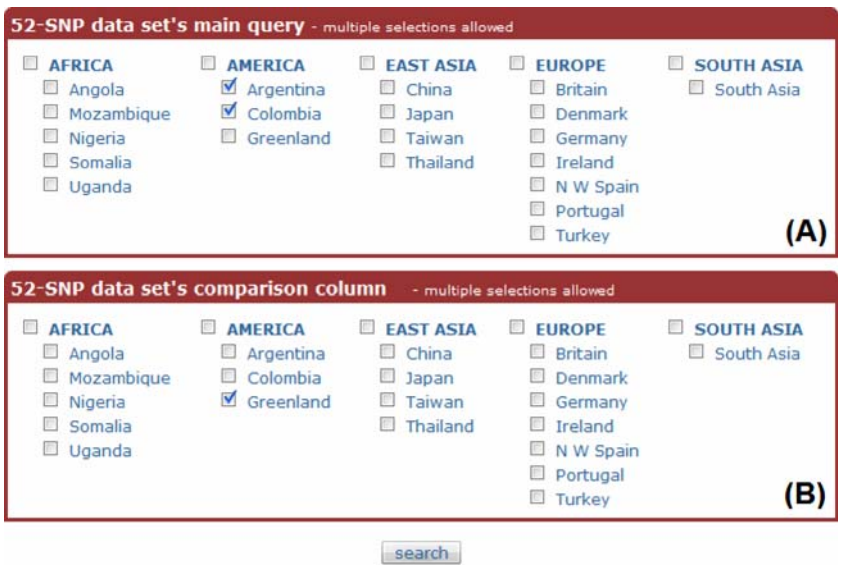

Fig. 2 **a** Search options available in the search page. Offset upper row tick-boxes allow combination of the listed populations of each region to create a full panel or population group. **b** Comparison options available in the search page. In each case, combinations can be tailored by the user to more closely match geographic distribution; in the example, ticking Argentina and Colombia in the search populations query and Greenland alone in the compare populations query permits comparison of North and South American population groups

Beijing and Japanese from Tokyo), AFR/YRI (SNP*for*ID African/Yoruba of Ibadan, Nigeria), plus SAS=SNP*for*ID South Asian and AME=SNP*for*ID American.

It is important to note that although all database profiles are complete, the sample number ranges from 7 (Japan) to 156 (Denmark) and clearly certain small population samples require interpretation with caution or exclusion altogether. The population data is structured in columns and the SNP data is structured in rows for all collated pie chart sets and corresponding full-frequency figures. These allele frequencies are shown numerically in columns under their corresponding genotyped base to four decimal places, and the pie charts are drawn to 1% allele frequency precision. A column of hyperlinks to dbSNP provides a convenient system for obtaining additional data for the individual SNP locus if required. The complete dataset of 52 SNP genotype profiles from all the populations listed (outside of HapMap) that underlie the allele frequency estimates are available as a flat text file download for each selected population, allowing user-defined analyses such as tests for independence or intrapopulation and interpopulation F_{st}.

Finally, at the time of writing, the website registered an average of 150 visits per month. The browser has been available to the public since December 2005 and has benefited in particular from links placed in the STRbase forensic marker information portal run by the National Institute of Standards and Technology (NIST, [12]) and the SNP*for*ID homepage (http://www.snpforid.org/).

Discussion

The SNP*for*ID browser represents a simple but highly effective visualization method to query and display the genotype data of the SNP*for*ID project. The format of the pie chart graphics also helps the researcher to quickly review the data, and the comparison with HapMap data as an external resource adds an appropriate system for confirming the precision of the allele frequency estimates given with both datasets being updated in real-time immediately before the display of the query results. This browser has been designed to be a web tool that can be rapidly accessed by the forensic practitioner requiring instant allele frequency data retrieval for a specific population plus a comparison at the same time with samples of global variability and is directly available at http://spsmart.cesga.es/snpforid.php.

Databases can fall into the trap of becoming static and out-of-date entities if they are not updated regularly. We have avoided this problem by recalculating allele frequency results at the moment a query has been submitted and by retrieving the current HapMap data at the same time. As well as ensuring all data displayed is the most current available, the dynamic system of data management we have adopted makes it easier to incorporate new data and to welcome submissions via e-mail from the worldwide forensic community (see contact information on the title page). This may represent a more efficient way to disseminate allele frequency data from an extending range of global populations than the conventional system of journal publication of allele frequency data. However, such an approach brings with it the problems of quality management more easily addressed in the curation of online haplotype loci databases mentioned previously (YHRD and EMPOP) where phylogenetic methods can be applied to check for typing errors. For this reason, we have decided to require scrutiny of raw genotyping data generated by contributing laboratories outside the SNP*for*ID consortium. We now include the ancestry-informative SNPs developed by SNP*for*ID [9] that supplement the identification SNP set. Ancestry-informative SNPs in particular benefit from the broadest range of shared population data because they show higher overall variability between populations. One favorable feature of autosomal SNP data in general is that relatively small population samples provide reliable allele frequency estimates. Therefore, submitting data to a shared database for SNPs of forensic interest should not represent a prohibitive amount of effort from those interested in validating these loci for forensic applications in their own laboratories.

Finally, we intend to allow for the possibility of linking allele frequency data to individual genotype profiles from widely used standard control samples such as the CEPH–HGDP panel of population samples [4] or the Coriell cell repositories control sample set. This would offer the simplest system for providing control profiles to help researchers that are establishing genotyping assays for the SNP*for*ID loci in their laboratories for the first time.

Acknowledgements The authors wish to thank Albert Vernon Smith and Lalitha Krishnan of the HapMap Project for their guidance in helping us link the browser to the HapMap SNP dataset, and Antonio Salas for his help with the genotyping quality assessment. We also would like to thank the Centro de Supercomputación de Galicia (CESGA) for their web hosting service and technical support. Funding from Xunta de Galicia: PGIDTIT06PXIB228195PR and Ministerio de Educación y Ciencia: proyecto BIO2006-06178 given to ML partially supported this work.

References

1. Altshuler D, Brooks LD, Chakravarti A, Collins FS, Daly MJ, Donnelly P, The International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437:1299–1320
2. Brandstatter A, Salas A, Niederstatter H, Gassner C, Carracedo A, Parson W (2006) Dissection of mitochondrial superhaplogroup H using coding region SNPs. Electrophoresis 27:2541–2550

3. Brion M, Sanchez JJ, Balogh K et al (2005) Introduction of an single nucleotide polymorphism-based "major Y-chromosome haplogroup typing kit" suitable for predicting the geographical origin of male lineages. Electrophoresis 26:4411–4420
4. Cann HM, de Toma C, Cazes L et al (2002) A human genome diversity cell line panel. Science 296:261–262
5. Emerson J (2006) DIY Map: a clickable and zoomable map written in Flash. Available at http://www.backspace.com/mapapp/
6. Lowe AL, Urquhart A, Foreman LA, Evett IW (2001) Inferring ethnic origin by means of an STR profile. Forensic Sci Int 119:17–22
7. Parson W, Brandstatter A, Alonso A et al (2004) The EDNAP mitochondrial DNA population database (EMPOP) collaborative exercises: organisation results and perspectives. Forensic Sci Int 139:215–226
8. Phillips C, Lareu V, Salas A, Carracedo A (2004) Non binary single-nucleotide polymorphism markers. In: Doutremepuich C, Morling N (eds) Progress in forensic genetics, 10. Elsevier, Amsterdam, pp 30–32
9. Phillips C, Salas A, Sanchez JJ et al (2007) Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. Forensic Sci Int Genetics 1:233–235
10. Roewer L, Krawczak M, Willuweit S et al (2001) Online reference database of European Y–chromosomal short tandem repeat (STR) haplotypes. Forensic Sci Int 118:106–113
11. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW (2002) Genetic structure of human populations. Science 298:2381–2385
12. Ruitberg CM, Reeder DJ, Butler JM (2001) STRBase: a short tandem repeat DNA database for the human identity testing community. Nucleic Acids Res 29:320–322
13. Sanchez JJ, Phillips C, Borsting C et al (2006) A multiplex assay with 52 single nucleotide polymorphisms for human identification. Electrophoresis 27:1713–1724
14. Thorisson GA, Smith AV, Krishnan L, Stein LD (2005) The international HapMap project web site. Genome Res 15:1592–1593