International Congress Series 1261 (2004) 18–20

ELSEVIER

www.ics-elsevier.com

Selecting single nucleotide polymorphisms for forensic applications

C. Phillips*, M. Lareu, J. Sanchez, M. Brion, B. Sobrino, N. Morling, P. Schneider, D. Syndercombe Court, A. Carracedo

The SNP for ID Consortium, Spain

Abstract. The selection criteria that can be used for choosing a core set of discriminatory SNPs from large scale databases for forensic use are reviewed.

Keywords: SNP; Linkage disequilibrium; LD blocks; Forensic analysis

1. Introduction

With nearly 6 million SNPs in numerous online databases available to choose and develop, it is increasingly important to formulate appropriate criteria for selecting the most useful and robust loci for forensic analysis. We aim to develop a core panel of SNPs for the optimum discrimination of individuals, in addition to developing specialized SNP sets. Studies of binary polymorphisms suggest that between 50 and 80 loci are required to match the discrimination levels of 16 STRs [1]. This makes linkage a critical factor in SNP selection, as several markers will be needed on each chromosome. Furthermore, SNPs rarely have optimum allele frequencies and often show marked differences in frequency distributions between populations, making polymorphic content an important additional characteristic. Lastly, flanking sequence quality needs to be considered in the context of the kind of genotyping assay to be used when selecting candidates.

2. Materials and methods

2.1. Number of candidate loci selected and databases used

We decided to genotype selected markers using SNaPshot minisequencing since this was a simple and reliable technique that most forensic labs can readily perform. SNP typing assays have varying conversion rates (i.e. successful incorporation of loci into the

* Corresponding author. Institute of Legal Medicine, University of Santiago de Compostela, Calle San Fransisco s/n, 15705 Santiago de Compostela, Spain. Tel.: +34-981582327; fax: +34-981580336.

E-mail address: chrisp@usc.es (C. Phillips).

0531-5131/
doi:10.1016/j.ics.2003.12.001

assay), averaging between 70% and 80%. Some candidates prove less polymorphic than supposed or difficult to amplify in multiplexes, reducing the conversion rate to 50–60%. Primer design constraints caused by the small amplicon sizes required (60–120 bp) and complex interactions as the primer pool expands further reduce the number. So it was decided to select an initial pool of 400 candidates, taken from the telomeric region of the p-arm and q-arm of each autosome. Four different online SNP databases were used to select candidate markers:

(a) Celera Discovery Systems database (CDS) comprises 4,802,233 SNPs discovered from the private human genome mapping initiative plus the best validated SNPs from public databases (www.celeradiscoverysystem.com).
(b) Applied Biosystems Assays-on-Demand (ABI) database comprises 146,636 SNPs with the best allele information currently available (derived from 46 Africans and Europeans) (http://www.appliedbiosystems.com/products/assays).
(c) NCBI dbSNP database comprises 2,243,761 unique SNPs discovered from centres contributing to the public HGMP (http://www.ncbi.nlm.nih.gov/entrez/).
(d) The SNP Consortium (TSC) database comprises 1,255,326 SNPs taken from dbSNP but with higher levels of validation from three population groups (http://snp.cshl.org/).

2.2. Selection criteria: validation, polymorphic content, flanking sequence and linkage

Validation status is important to ensure each SNP is both real (i.e. not due to a sequencing error) and unique (i.e. mapping once in the genome). Growing numbers of SNPs are validated by genotyping and this provides the allele frequencies necessary to properly assess locus variability. The polymorphic content of a marker is often measured in forensic analysis by the discrimination index (Dp). The Dp values for SNPs change very little between minimum allele frequencies of 0.5–0.3 (dropping from 62.5% to 58%), so this range was used for selecting SNPs with sufficient variability. Polymorphism levels also need to be sufficiently high in at least two of the three population groups quoted in the TSC and ABI databases—many SNPs exhibit high variability in only one population and so are not applicable—these comprised about 10% of the, otherwise good, candidate SNPs found. Sequence features that are of importance in candidate selection include repeat motifs, % GC, polybases, secondary structures and SNP clustering. Good GC balance and an absence of tandem repeats around the SNP are essential to ensure good primer design. Potential secondary structures were not actively detected during selection but 7 of the 46 first SNP candidates failed the primer design process due to this problem. Examples of problem sequence encountered are shown in Fig. 1. SNP clustering is a characteristic that

Repeats	TCAAAATAATAATAATAATAATDATAATAATAATAATAATAAT	rs379091
%GC	TGGTTGGGGGGTGTGTGTGGGBGGGGGGACGGTGTGGAGGGC	rs1574185
	TAATTTTTTTTCTTTTCTTTTHTTTTTTTNTGCAAAAAGATGTCT	rs727936
Polybases	AAGAATTCAAGATGGTATTTWAAAAAAAAACCTCATATCTT	rs113492

Fig. 1. Sequence characteristics of several SNPs that gave assay design problems.

often interferes with primer design: about 15–20% of otherwise good markers were rejected because of additional variable positions in close proximity to the SNP of interest.

Establishing if linkage disequilibrium (LD) is shown by SNPs is important in order to use a cumulative frequency for a multiple SNP profile. Current studies examining LD distribution report that chromosomal blocks exist where markers are in almost complete LD across relatively large distances. Such blocks are not consistent with the established model of genetic distance measured by a uniform recombination rate. Recent analysis of whole chromosomes (6, 21 and 22) using high density SNP maps [2] reveals LD, in certain cases, extending as far as 300 kb between two positions and varying in extent between different populations. However, the average size of these regions, termed LD or haplotype blocks, was only 26 kb in Caucasians (with just 30% of LD blocks longer than 25 kb) and 18 kb in Africans (with just 22% of LD blocks longer than 25 kb). Clearly, an average block size of $\sim$ 20–25 kb means the risk of LD between SNP and gene is extremely low over distances greater than 100 kb making this a realistic window to use around known genes. In addition, LD between two SNP positions though highly variable is also unlikely to regularly extend over greater distances. Association analysis of two SNPs, 1 Mb apart, on chromosome 10 will form part of our initial validation process. Once SNP analysis is established, it is likely that forensic testing will often make use of both STRs and SNPs in parallel. The STR markers in routine forensic use are all mapped and equivalent distances can be used around these loci to avoid linkage issues.

3. Results and discussion

The first set of 23 markers selected using the criteria outlined above have now been successfully combined into a single multiplexed PCR and minisequencing reaction. This assay is currently being used to validate allele frequencies in several populations and to test for robustness and sensitivity in routine forensic analyses. It appears that none of the criteria used and outlined here will need major revision when further sets of loci are selected.

References

[1] P. Gill, Int. J. Leg. Med. 114 (2001) 204–210.
[2] F. De La Vega, et al., (poster 85), Proceedings of the 8th Int. Human Genome Meeting, 2003.