Benchmark

Molecular Haplotyping by PyrosequencingTM

BioTechniques 33:1104-1108 (November 2002)

Jacob Odeberg, Kristina Holmberg, Per Eriksson¹, and Mathias Uhlén

Royal Institute of Technology and ¹Karolinska Institutet, Stockholm, Sweden

Several million SNPs have been identified and are available in the public databases. For most, no effect on gene function or expression is known. "Functional polymorphisms" refers to SNPs to which a putative effect on gene function or expression is attributed and implies that they are "non-silent" variations located in regions that are coding, regulatory, splice sites, or untranslated regions that affect mRNA stability and translation. In association studies (12), a polymorphism state that is correlated with the occurrence of disease suggests that the etiological agent is either in linkage with the analyzed polymorphism site or is identical with the polymorphism analyzed. It is likely that haplotypes, which are the specific combinations of nucleotides located on the same chromosomal molecule (allele), will be more informative on the complex relationship between phenotypes (disease) and DNA variation than any SNP. The disease association of a specific allele may be dependent on *cis* effects involving functional variants at other loci within the gene, and an association may not be found unless a sufficient marker for the haplotype on which a *cis* effect arises is typed. Regardless of possible *cis* effects, when an analyzed SNP has no effect on gene function but

becomes a marker for a yet uncharacterized functional genetic variation within the same genomic region, if the functional variant(s) is of more recent origin, then the polymorphism state at the analyzed (non-functional) marker polymorphism may be shared between the allele harboring the functional variant(s) and one that does not. In such a case, genotype analysis of the marker polymorphism could fail to detect a significant association with disease existing for one of the underlying alleles.

Based on studies of genomic variance in extended chromosomal regions and of the entire chromosome 21 (2,11), a picture of discrete haplotype blocks spanning up to 100 kb has emerged. Remarkably, the diversity is very limited, with a few haplotypes (two to four) accounting for more than 90% of the chromosomes/alleles present in a population sample (2,11). This implies that even if a large number of polymorphic sites exists in such haplotype blocks, then the corresponding haplotypes can be identified by using a small number of haplotype "tags" (6), which do not necessarily have to be spread out over a haplotype block. Thus, these haplotype tags can be a small subset of SNPs that uniquely distinguishes the different common haplotypes in each block and allows for exhaustive testing for whether common variation within a longer genomic region is associated with disease, regardless of whether it is the functional variant(s) that are being genotyped or not. Current methods for analyzing haplotypes include statistical estimation from population genotype data or from family genotype data (3,13,15). Direct

molecular genetic typing methods such as allele-specific PCR, with match/mismatch 3'-end primers for two SNPs located some distance apart, have been described (10), but the robustness of this approach is potentially liable to be sequence dependent. For example, GT or CA mismatches are clearly not refractory to extension in PCR (5,8).

We demonstrate here how the PyrosequencingTM technique (Pyrosequencing AB, Uppsala, Sweden) (14) is applicable to molecular haplotyping, demonstrated for the Cathepsin S (CTSS) and Matrix metalloproteinase-7 (MMP7) genes. Inherent to the sequencing-by-synthesis principle, on which the Pyrosequencing technique is based, is the possibility to design a nucleotide addition order with the extension over a polymorphic site becoming out of phase on the two different alleles, so the extension on one strand is lagging behind when passing over and continuing beyond the polymorphism. This results in distinctive raw data profiles that will distinguish between genotypes without ambiguity. By keeping the extension out-of-phase up to and beyond a second polymorphism, the raw data profile obtained at this site will be dependent on which nucleotide variant was present at the first polymorphic position on the same allele/chromosomal molecule. To illustrate the principle of out-of-phase extension, a haplotype marker in the CTSS gene, consisting of two polymorphic sites located 3 bp apart in the proximal promoter, is shown in Figure 1, where arrows indicate the position of the 3'-ends of the extended strands on the alleles after the fifth and eleventh nucleotide

DRUG DISCOVERY

AND GENOMIC TECHNOLOGIES

addition. Of the four theoretically possible alleles, only three are found in the Scandinavian population sample analyzed (320 individuals), with allele frequencies of 33%, 8%, and 59%, respectively. In the second example, a haplotype marker of two polymorphic sites in the MMP7 gene promoter located 28 bp apart at -181(A/G) and -153(C/T) are shown (7). The 184-bp

Figure 1. Raw data genotype profiles for the CTSS haplotype unit. The polymorphisms analyzed are positioned at nucleotides 902 and 899 with reference to GenBank® accession no. AH0035692. The sequence is displayed and analyzed in the antisense direction. The 902 polymorphism was previously reported by Cao et al. (1) as -25(G/A), while the 899 (-28) polymorphism is new. Out of the four theoretical haplotypes, three are found in the population analyzed (alleles 1–3). The position on the three alleles of the 3'-end of the extended strands after the fifth (I) and eleventh (II) nucleotide dispensations are marked by arrows, and the corresponding position in the raw data output is shown (dotted lines). The raw data corresponding to the three haplotypes are shown by the homozygotes [1/1], [2/2], and [3/3]. The genotype profiles of heterozygotes [1/2], [1/3], and [2/3] are the result from overlaying the corresponding homozygote profiles.

genomic fragment of the CTSS gene (AH0035692) was amplified with primers (5'-biotin-ACCTCATGTGAC-AAGTTCCAATTTC-3') and (5'-GACCAAATGGGAGAAAAAGAACAAG-3'). The 165-bp genomic fragment of the MMP7 gene (U25346) was amplified using primers (5'-AGTCTACAGAACTTTGAAAGTATGTG-3') and (5'-biotin-CTATGAGAGCAGTCATT-TGACTTTG-3'). One nanogram of genomic DNA was used in 50-μL reactions. Sample preparations and annealing of the CTSS sequencing primer (5'-GATAGAACCAGCAGTTGCTC-3') or the MMP7 sequencing primer (5'-CTAAACGAGGAAGTATTACATC-3') were performed on MBS automated workstations (MBS, Stockholm, Sweden) using streptavidin-coated beads, following the manufacturers instructions. Pyrosequencing was performed, with the addition of 1 μg single-stranded protein (SSB) per reaction. Nucleotide dispensation orders were [GCACTGACTGAGTAGAGTC] and [ATAGTGATCAGCAGCAGCACATGTACTGTCTATGTATGTCTCAGT-CT], respectively.

Both CTSS and MMP7 have attracted interest in relation to the progression and stability of atherosclerotic plaques in coronary arteries. The C allele of CTSS promoter polymorphism 902C/T

[also referred to as -25G/A (1)] was previously reported with a frequency of 54% (1). The (G/A) polymorphism at position 899 is a previously unknown polymorphism we found when typing the C902T polymorphism. As the previously known C allele of the C/T polymorphism is here found resolve into two alleles [902C-899G] and [902C-899A], with allele frequencies of 33% and 8% (summing up to 41%), haplotype marker analysis of this region could yield a better resolution in an association study. The difference in allele frequency between the Scandinavian population here analyzed and that previously reported for the Caucasian and Inuit populations analyzed by Cao et al. (1) can reflect its population history. The short distance between the two polymorphisms located just upstream of the transcription start site could have potential implications for transcriptional activity that includes a *cis* interaction between the two variants. The combination of the two polymorphisms in haplotype 2 (Figure 1) creates a putative AML-1 transcription factor binding site (TGTGGG in the sense direction) not shared with haplotype 1 or 3 (9), but whether this has any biological significance requires further work. For the MMP7 gene promoter haplotype marker, an interaction was previously

Figure 2. Raw data profiles for the MMP7 haplotype marker analysis. The -153 and -181 polymorphism in the promoter region (7) is here displayed and analyzed in the antisense direction. Six genotypes were found in the population analyzed, which represents three different alleles. Pyrosequencing profiles for the three haplotypes identified are represented by the homozygotes [1:1], [3:3], and [4:4]. Arrows mark peaks from which the state of the second polymorphism is determined. The state of the first polymorphism is identified by the distinct pattern obtained for the first 20 nucleotide additions [4:4]. Genotype profiles for heterozygotes result from overlaying the corresponding haplotype profiles.

indicated (7). In vitro functional analysis of the transcriptional activity of the four possible theoretical alleles demonstrated a synergistic positive effect for the combination of the -181G and -153T variants, suggesting a *cis* effect on the recruitment and binding of transcription factors (7). In principle, haplotype markers involving three or four polymorphic sites located in proximity could be analyzed by this approach. The number of possible genotype pyrogram profiles would be equal to:

$$n = X \\ \sum (X-n) \\ n = 0$$

where X is the number of haplotype marker combinations present in a population. The nucleotide addition order would be designed to continue the out-of-phase-extension beyond the second or third polymorphism, and from the point of having at least 2 or 3 informative peaks for each allele. The strategy would be to establish a gap between the two extending strands when passing over the first polymorphism that is 3–6 bp when reaching the second polymorphism, where the two strands become resolved into underlying haplotypes. This will give a necessary flexibility when designing the nucleotide addition order at this position so as to obtain gaps of 2–3 bp between different extending alleles when continuing up to the third position, where further resolution will occur, and so on. Large gaps between extending strands/alleles should possibly be avoided, as the total number of nucleotide additions relative to the length of the target sequence may become unnecessary high. The limitation of Pyrosequencing's applicability to molecular haplotyping lies in the read length of the Pyrosequencing technique, restricting it to proximal haplotype tag SNPs. Recent work (4) has shown that this can be extended up to 160 bp de novo sequence [stepping through unknown sequence with a (CAGT)_n dispensation order] by optimization of the enzyme and reagent composition, and further work may increase this. As the sequence and the variation within the target region in molecular haplotyping is known beforehand, the nucleotide dispensation order is optimized both with respect to the intervening sequence and to obtain

adequate out-of-phase extension profiles that distinguish between alleles. Nucleotide additions for which no extension on either allele will occur are eliminated, and, consequently, also part of the eventual effect of the accumulated successive nucleotide dispensations that can reduce data quality. Therefore, Pyrosequencing-based analysis of haplotype markers should be applicable to typing of haplotype tag SNPs separated by distances in parity with the upper limit of de novo sequencing.

In conclusion, the possibility of using Pyrosequencing for molecular haplotyping of large genomic haplotype blocks promises to reduce the genotyping work in genetic epidemiological studies for detecting associations of genome blocks with disease.

ACKNOWLEDGMENTS

This study was supported by funds from the Magn. Bergvalls Foundation, the Medical Research Council (12660), and the Foundation for Strategic Research (SSF). We thank Annica Åhberg and Lars Svennersten for expert assistance with the genotyping work.

REFERENCES

1. Cao, H. and R.A. Hegele. 2000. Human cathepsin S gene (CTSS) promoter -25G/A polymorphism. *J. Hum. Genet.* 45:94-95.
2. Daly, M.J., J.D. Rioux, S.F. Schaffner, T.J. Hudson, and E.S. Lander. 2001. High-resolution haplotype structure in the human genome. *Nat. Genet.* 29:229-232.
3. Excoffier, L. and M. Slatkin. 1995. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol. Biol. Evol.* 12:921-927.
4. Gharizadeh, B., T. Nordstrom, A. Ahmadian, M. Ronaghi, and P. Nyren. 2002. Long-read pyrosequencing using pure 2'-deoxyadenosine-5'-O'-(1-thiotriphosphate) Spisomer. *Anal. Biochem.* 301:82-90.
5. Huang, M.M., N. Arnheim, and M.F. Goodman. 1992. Extension of base mispairs by *Taq* DNA polymerase: implications for single nucleotide discrimination in PCR. *Nucleic Acids Res.* 20:4567-4573.
6. Johnson, G.C., L. Esposito, B.J. Barratt, A.N. Smith, J. Heward, G. Di Genova, H. Ueda, et al. 2001. Haplotype tagging for the identification of common disease genes. *Nat. Genet.* 29:233-237.
7. Jormsjo, S., C. Whatling, D.H. Walter, A.M. Zeiher, A. Hamsten, and P. Eriksson. 2001. Allele-specific regulation of matrix metalloproteinase-7 promoter activity is associated with coronary artery luminal dimensions among hypercholesterolemic patients. *Arterioscler. Thromb. Vasc. Biol.* 21:1834-1839.
8. Kwok, S., D.E. Kellogg, N. McKinney, D. Spasic, L. Goda, C. Levenson, and J.J. Sninsky. 1990. Effects of primer-template mismatches on the polymerase chain reaction: human immunodeficiency virus type 1 model studies. *Nucleic Acids Res.* 18:999-1005.
9. Meyers, S., J.R. Downing, and S.W. Hiebert. 1993. Identification of AML-1 and the (8;21) translocation protein (AML-1/ETO) as sequence-specific DNA-binding proteins: the runt homology domain is required for DNA binding and protein-protein interactions. *Mol. Cell Biol.* 13:6336-6345.
10. Michalatos-Beloin, S., S.A. Tishkoff, K.L. Bentley, K.K. Kidd, and G. Ruano. 1996. Molecular haplotyping of genetic markers 10 kb apart by allele-specific long-range PCR. *Nucleic Acids Res.* 24:4841-4843.
11. Patil, N., A.J. Berno, D.A. Hinds, W.A. Barrett, J.M. Doshi, C.R. Hacker, C.R. Kautzer, D.H. Lee, et al. 2001. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* 294:1719-1723.
12. Risch, N. and K. Merikangas. 1996. The future of genetic studies of complex human diseases. *Science* 273:1516-1517.
13. Rohde, K. and R. Fuerst. 2001. Haplotyping and estimation of haplotype frequencies for closely linked biallelic multilocus genetic phenotypes including nuclear family information. *Hum. Mutat.* 17:289-295.
14. Ronaghi, M., M. Uhlen, and P. Nyren. 1998. A sequencing method based on real-time pyrophosphate. *Science* 281:363.
15. Stephens, M., N.J. Smith, and P. Donnelly. 2001. A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.* 68:978-989.

Received 17 May 2002; accepted 25 July 2002.

Address correspondence to:

Dr. Jacob Odeberg
Department of Biotechnology
KTH, Royal Institute of Technology
AlbaNova University Center
106 91, Stockholm, Sweden
e-mail: jacob@biotech.kth.se

For reprints of this or
any other article, contact
Reprints@BioTechniques.com