Proc. Assoc. Advmt. Anim. Breed. Genet. Vol 14

TIPS: A PROCESS FOR RAPID FINE MAPPING OF QTLS USING ESTS AND COMPARATIVE MAPPING

J.C. McEwan[1], K.A. Paterson[2], A. Zadissa[2], T. Van Stijn[2], C. Diez-Tascon[2], A.M. Crawford[2]

[1]AgResearch, Invermay Research Station, PB 50034, Mosgiel, New Zealand
[2]AgResearch Molecular Biology Unit, Biochemistry Department, University of Otago, PO Box 56, Dunedin, New Zealand

SUMMARY

Targeted **I**ntronic **P**olymorphic **S**equence identification (**TIPS**) is a procedure, which allows rapid identification of novel SNPs in ruminant species, targeted to an area of interest. Normally, this region would be a QTL and the markers would aid in its fine mapping. An advantage of TIPS is that it can be used for cross species SNP identification within expressed genes, and therefore allows rapid development of Type 1 markers across species. Current results suggest that the process is both rapid and cost effective relative to previous alternatives. The process is dependent on the availability of draft human genomic sequence, and ESTs, the latter derived either from the species of interest or from a closely related genus. Results to date suggest that for an average 10 Mbp human region, around 50% of unique genes will have a bovine contig. Of the primers developed using bovine sequence, approximately 72% work in ovine DNA, and SNPs have been be identified in 64% of cases. This results in an overall success rate of 46% per primer pair designed. Approximately 9 SNPs per 10Mbp can be identified with the bovine ESTs available in February 2001. The costs per targeted SNP marker developed are likely to be less than NZ$1000. This is approximately an order of magnitude cheaper than previous alternatives.

Keywords: SNP, DNA, ruminant, introns, comparative mapping.

INTRODUCTION

In recent years there has been an upsurge of reports where QTL affecting production traits have been detected in ruminant livestock species. Many more have been unreported because of commercial considerations. While there have been a few reports of subsequent successful identification of the genetic polymorphisms underlying these traits (e.g. Kambadur *et al*. 1997; Galloway *et al*. 2000; Wilson *et al*. 2001), the present methods are expensive in time and personel. The key problems have been the lack of sufficient polymorphisms in farmed ruminant livestock in a given region of interest, and also the lack of detailed comparative maps of type 1 markers. For example the latest (3rd generation) sheep map contains 1061 markers of which only 120 are type 1 markers (Maddox *et al*. 2001). Previous work (Zadissa 2000) has described the level of homology required for successful automated design of PCR primers and suggested that while orthology comparisons could be made *in silico* between humans and ruminants, successful PCR primer design required ruminant sequence. Fortunately, the number of public and private bovine ESTs has increased rapidly and as at February 2001 there were more than 160,000 sequences in Genbank (http://www.ncbi.nlm.nih.gov/). Single nucleotide polymorphisms (SNPs) are the most common DNA variant, but unfortunately few SNPs are conserved across species (Hacia *et al*. 1999). However, it is possible to detect SNPs across closely related species using ESTs, because of the high level of conservation of the length and position of

intron/exon boundaries in mammals (Batzoglou *et al*. 2000). This enables primers to be designed in the conserved exon sequence bordering an intron with a high chance of amplification success in the related species. Repeated sequencing of the intronic region in different individuals then identifies polymorphisms in the species of interest. The TIPS process described subsequently provides a cost effective solution to the various constraints imposed by targeted SNP discovery.

METHOD

The TIPS process involves 4 separate steps and only the middle two are described in detail in this paper. All of the results attempt to identify targeted SNPs in sheep using draft human genomic sequence and bovine ESTs.

The first step involves the identification of the likely comparative region in the human genome to the QTL identified using manual methods (O'Brien *et al*. 1999), although automated methods are under development by ourselves (Zadissa *et al*. 2001) and others (Rebiez and Lewin 2000). Normally, this identifies a region spanning 10-20 Mbp of DNA.

The second step involves the identification of orthologous bovine ESTs that map within this region. The sequence for this section of human genomic DNA and the full and partial cDNA sequences mapped to this region, are then downloaded in the form of FASTA files. Commonly we use http://genome.cse.ucsc.edu/index.html in order to undertake this step. Genomic sequence is downloaded directly. The html source for the region is then parsed by computer for cDNA Genbank identifiers. The relevant cDNA sequence is then extracted using batch entrez (http://www.ncbi.nlm.nih.gov/entrez). The full and partial human cDNA sequences are then masked for repetitive bovine and human sequences and compared to a file of bovine ESTs and cDNAs using a stand-alone BLAST (Altschul *et al*. 1997) with a threshold expectation < 1e-20. Matching bovine ESTs are then extracted from the bovine EST file and masked for repetitive bovine sequences. Experience has shown masking is necessary to avoid subsequent EST assembly problems and eliminate the possibility of accidentally designing primers within repetitive elements. In particular there are a surprising number of retrotransposon elements or their deriviatives transcribed in the 3' UTRs of many ruminant genes. Subsequently, these bovine ESTs are assembled using CAP (Haung 1992) modified so that assembled ESTs require greater than 40bp overlap and greater than 96% homology. The contig assembly process has several benefits: first it eliminates much of the redundancy from the EST data, secondly the consensus sequence is longer and more accurate than the original ESTs. Finally, it allows identification of closely related and paralogous genes, chimeric ESTs and alternatively spliced variants. The full and partial human cDNA sequences are then compared using BLAST against a file containing: draft human sequence of the selected region, the full and partial human cDNA sequences and the tentative bovine consensus sequences using the following options expectation < 1e-20, -m 4, –F “m D”. Length of putative introns are then calculated from the output. The user can from this single output file identify the human cDNA, any matching human cDNAs (outputs for these cDNAs are subsequently ignored), the intron/exon boundaries in human genomic DNA and their length and the matching bovine contig. Close examination of the contigs identifies chimeric, alternative spliced and paralogous sequences which normally have a lower homology than the orthologous sequence. Simple cutting and pasting and minor editing of selected bovine exon sequence surrounding the selected intron/exon boundaries allows for automated

PCR primer design of multiple target sequences in batch mode using PRIMER (http://www-genome.wi.mit.edu/genome_software/). Introns selected for amplification are normally longer than 500 bp and less than 1500 bp in length. An additional BLAST is undertaken to check that the cDNA selected maps uniquely to a single position in the human genome. Cumulated statistics to date show the success at each step of this procedure (Table 1).

The third step is to sequence genomic DNA amplified by the selected PCR primers from the target species. Initially the amplified sequence length is determined and compared to the expected length and the presence of any multiple bands is noted. The single pass DNA sequence obtained from using one or both primers is also compared with the consensus sequence. If all results suggest the correct region has been amplified and is unique, then three to six individuals, usually the sires used in the QTL experiment or a mapping flock are single pass sequenced. Sequence comparison, confirmed via examination of ABI files identifies putative SNPs, indels, and occasionally potential microsatellite repeats. Generally, the presence of one or both homozygotes and a heterozygote is sufficient evidence to proceed to the next step, although doubtful cases can be confirmed by repeated sequencing, sequencing further individuals or sequencing in the opposite direction. The success rate at each of these steps has been tabulated in Table 1. The final step, is to map the SNP in an appropriate flock. In our laboratory this is likely to require primer redesign and use of a Mass Spectrometry based technique (Leushner and Chiu 2000)

RESULTS AND DISCUSSION

The numbers tabulated in table 1 suggest that for the unique human cDNAs in a region approximately 50% have a matching bovine contig and of these 82% have a suitable intron available for primer design. This normally results in some 20 or more potential primers available within a 10 Mbp region. Of the primers around 72% amplify the expected ovine product. Of those amplifying the expected product around 64% have one or more identifiable SNPs. Thus the overall success rate in ovine is 46% per primer designed from bovine sequence or around 9 SNPs per 10 Mbp. While it is unlikely that the success rate will improve markedly, the number of SNPs identified in a region will increase as more human cDNAs are mapped and more ovine and bovine ESTs become available. As all the SNPs identified by this technique are linked to type 1 markers in humans, mapping these markers in the target species also greatly improves the comparative map in the region of interest. Of course successful primers can also be used to probe various Cosmid, BAC and YAC libraries, which in turn can be probed by a variety of methods to identify more SNPs, or to act as anchors for physical mapping.

Table 1. Success rates of various steps in the TIPS process

	N	%
N contigs/unique gene in target region	75/150	50
N introns/contig	121/75	161
N introns 500-1500bp/contig	62/75	82
N PCRs successful	26/36	72
N SNPs identified	16/25	64

Smith *et al*. (2001) have developed a similar methodology using unassembled and untargeted bovine ESTs, and report 80% success in intron amplification per primer pair in bovine, with an SNP detected in 72% of those amplified. Our method differs appreciably in that we target our region, whereas they randomly devise primers over the whole genome. They match their bovine EST sequences against human genomic DNA, whereas we compare against cDNAs and only indirectly against human genomic DNA, as this improves intron/exon boundary detection and matching bovine sequence in low homology regions. Finally, by necessity we use primers designed from bovine sequence to amplify ovine sequence and as expected primer amplification success rates are lower. In conclusion, use of the TIPS process to identify SNPs has been described, and the success rates of the various component steps is reported.

REFERENCES

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) *Nucleic Acids Res*. **25**: 3389.

Batzoglou, S., Pachter, L., Mesirov, J.P., Berger, B. and Lander, E.S. (2000) *Genome Res*. **10**: 950.

Galloway, S.M., McNatty, K.P., Cambridge, L.M., Laitinen, M.P., Juengel, J.L., Jokiranta, T.S., McLaren, R.J., Luiro, K., Dodds, K.G., Montgomery, G.W., Beattie, A.E., Davis, G.H., and Ritvos, O. (2000) *Nat. Genet*. **25**: 279.

Hacia, J.G., Fan, J.B., Ryder, O., Jin, L., Edgemon, K., Ghandour, G., Mayer, R.A., Sun, B., Hsie, L., Robbins, C.M., Brody, L.C., Wang, D., Lander, E.S., Lipshutz, R., Fodor, S.P. and Collins, F.S. (1999) *Nat. Genet*. **22**: 164.

Huang, X. (1992) *Genomics* **14**: 18.

Kambadur, R., Sharma, M., Smith, T.P. and Bass, J.J. (1997) *Genome Res*. **7**: 910.

Leushner, J., Chiu, N.H. (2000) *Mol. Diagn*. **5**: 341.

Maddox, J.F., Davies, K.P., Crawford, A.M., Hulme, D.J., Vaiman, D., Cribiu, E.P., Freking, B.A., Beh, K.J., Cockett, N.E., Kang, N. *et al*. (2001) *Genome Res*. (In Press).

O'Brien, S.J., Menotti-Raymond, M., Murphy, W.J., Nash, W.G., Wienberg, J., Stanyon, R., Copeland, N.G., Jenkins, N.A., Womack, J.E. and Marshall Graves, J.A. (1999) *Science* **286**: 458

Rebeiz, M., and Lewin, H.A. (2000) *Anim. Biotechnol*. **11**: 75

Smith, T.P.L., Grosse, W.M., Stone, R.T., Bennett, G.L., Casas, E., and Keele, J.W. (2001) PAG IX (abstract)

Wilson, T., Wu, X., Juengel, J.L., Ross, I.K., Lumsden, J.M., Lord, E.A., Dodds, K.G., Walling, G.A., McEwan, J.C., O'Connell, A.R, McNatty, K.P. and Montgomery, G.W. (2001) *Biol. Reprod*. **64**: 1225.

Zadissa, A. (2000) MSc thesis, Uppsala University, Sweden

Zadissa, A., Dodds, K.G., McEwan, J.C. (2001) *Proc. Assoc. Advmt. Anim. Breed. Genet*. **14**: 99.