A Snafu in SNP Analysis in a Member of a Multigene Family

BioTechniques 30:726-732 (April 2001)

Single nucleotide polymorphisms (SNPs), the most common variant in the human genome (3), have become of strong interest in efforts to develop a third-generation human genome map useful for the efficient identification of genes that influence complex traits (8). This interest is based on characteristics that include high incidence (about 1 in every 500 bp) (12) and low mutability (2). These characteristics make SNPs suitable for association studies of mutations and complex traits (8). SNPs are also useful for population genetic studies, as methods have been developed that allow rapid estimation of allele frequency at each locus (9). In addition, the identification of SNPs unlike that of other genetic markers is more amenable to automation, and, thus, a rapid genome-wide scan for causative mutations in genes that influence complex traits is feasible. Because of this utility, several reports have recently described large-scale programs of SNP discovery in diverse genes in humans (2,6,15) and birds (14). There are three primary methods of SNP discovery: shotgun sequencing of cloned PCR products, oligonucleotide arrays, and resequencing of PCR products. Of these three, the resequencing of PCR products remains the most straightforward and least technically demanding. However, the resequencing of PCR products may not be an appropriate method for SNP detection within genes that are members of multigene families. These genes make up a significant fraction of the coding sequences of the genome (4). Because members of the same family share significant DNA sequence homology (7), oligonucleotides designed to amplify fragments from one member may amplify a similar product from another member. One can thus hypothesize that some SNPs detected by scanning a member of a multigene family may be

This work was partially supported by the Deutsche Forschungsgemeinschaft grant no. INK-16A1-1. Address correspondence to Dr. Wolfgang Höhne, Institut für Biochemie (Charité), Humbolt-Universität zu Berlin, Monbijoustr. 2, 10117 Berlin, Germany. e-mail: wolfgang.hoehne@charite.de

Received 26 June 2000; accepted 11 December 2000.

Matthias Paschke, Grit Zahn1, Axel Warsinke, and Wolfgang Höhne1

Universität Potsdam
Postdam
1Humbolt-Universität zu Berlin
Berlin, Germany

Figure 1. Consed view of the variant nucleotides at position 376 (vertical arrow) in the aligned genomic DNA sequences. This SNP, VPlagSNP-005, corresponds also to position 3769 within L21913, the target sequence. The three heterozygotes within this alignment and at the variant position are tagged dark gray. The mismatch homozygotes are color coded as black for A and red for both G and C genotypes, respectively.
false as a result of the heterogeneity of the amplified products. Here, we evaluated the effect of two nucleotide differences between the sequences of two members of the chicken (Gallus gallus) proteoglycan core protein multigene family on the validity of SNP detection in one of the members.

Proteoglycans are macromolecules that are widely distributed in extracellular matrices of all tissues in which they function in diverse developmental processes including maintaining cartilage integrity during skeletal development (10) and promoting neurite outgrowth during brain development (1). In the chicken, three members of the proteoglycan core protein gene family have been described, including brain chondroitin sulfate proteoglycan (GenBank accession no. U78555), white leghorn aggrecan mRNA sequence (GenBank accession no. M88101), and the chicken cartilage proteoglycan core protein (GenBank accession no. M88101), and the chicken cartilage proteoglycan core protein (GenBank accession no. L21913). Though the coding sequences for U78555 and L21913 produce distinct core proteins that function in the brain and cartilage, respectively, it is yet unclear if they are the products of a single gene (11). On the other hand, while M88101 and L21913 share a sequence identity of about 99.98% within a 5552-bp overlap region, the two differ by only 2 bp that correspond to positions 3769 and 3931 in L21913. To test whether the differences at these two sites may influence the validity of SNPs detected by resequencing, we amplified a PCR product of about 600 bp using oligonucleotides specific for L21913. The sequences of the oligonucleotides used are 5'-AGAAGCAGAAGCAAAAGAA-3' (forward) and 5'-GGCCTTGTCACAACTTCTAC-3' (reverse). The binding sites of the forward and reverse primers within U78555, L21913, and M88101 were 3621 and 4230, 3382 and 3981, and 3538 and 4137, respectively. Based on these binding sites, the expected product size from each targeted region, if each sequence represents a unique gene, is 609 bp from U78555 and 599 bp from L21913 and M88101. The sequence of the forward and reverse primers showed 100% homology to U78555, L21913, and M88101 reference sequences.

Genomic DNA samples from a total of 10 unrelated chickens from a randombred population were used as templates in the PCR. The PCR was carried out at an annealing temperature of 56°C using cycling and reaction conditions as reported elsewhere (14). Amplified products were processed for sequencing as previously described (13). Sequences were analyzed using a combination of Phred, Phrap, and Consed for assembly, alignment, and editing, as described by Gordon et al. (5).

Table 1. Nucleotide Variants Identified in L21913 and the Corresponding Alleles and Positions in M88101

<table>
<thead>
<tr>
<th>ID^a</th>
<th>Variant^b</th>
<th>Allele/Position^c</th>
</tr>
</thead>
<tbody>
<tr>
<td>VPlagcSNP-001</td>
<td>C/A</td>
<td>L21913: A3427, M88101: A3570</td>
</tr>
<tr>
<td>VPlagcSNP-002</td>
<td>A/G</td>
<td>L21913: G3475, M88101: G3618</td>
</tr>
<tr>
<td>VPlagcSNP-003</td>
<td>A/G</td>
<td>L21913: G3492, M88101: G3635</td>
</tr>
<tr>
<td>VPlagcSNP-004</td>
<td>A/G</td>
<td>L21913: G3539, M88101: G3682</td>
</tr>
<tr>
<td>VPlagcSNP-005</td>
<td>A/C/G</td>
<td>L21913: C3769, M88101: C3912</td>
</tr>
<tr>
<td>VPlagcSNP-006</td>
<td>A/G</td>
<td>L21913: C3824, M88101: C3967</td>
</tr>
<tr>
<td>VPlagcSNP-007</td>
<td>T/C</td>
<td>L21913: C3931, M88101: T4074</td>
</tr>
</tbody>
</table>

^a Represents the identity of each SNP as described in our database at ftp: genome.apsc.vt.edu
^b Validation of each variant was based on the incidence of the variant nucleotide in a second PCR product and in the aligned sequences of the reverse primer.
^c The nucleotide and position within the GenBank sequences L21913 and M88101.
^d An anomalous variant position: three alleles were observed and validated at this locus.
^e Nucleotide positions at which L21913 and M88101 differ.

Alignment of GenBank sequences corresponding to the regions amplified by each primer pair showed that U78555 and L21913 share a 100% sequence identity, while L21913 and M88101 have a two-base difference as indicated above. Within the region scanned, a total of seven candidate SNPs were detected including VPlagcSNP-005 and -007, positions at which L21913 and M88101 differ (Table 1).

Figure 2. Consed view of the variant nucleotides at position 188 (vertical arrow) in the aligned DNA sequences from selected clones of individual amplified products. The G, T, and C genotypes (from T7 sequencing) in this alignment are complimentary to the G, A, and C observed in the alignment of the genomic DNA sequences observed in Figure 1 for VPlagcSNP-005. As in Figure 1, the nucleotides at this position suggest an anomaly that required further investigation.
At the VPIagcSNP-005 locus, three alleles, G, A, and C were detected (Figure 1 and Table 1). Like other SNPs identified in this study, the variation at this locus involving the three nucleotides was validated by sequencing a second PCR product as well as reverse primer sequencing. This anomalous position in our alignment corresponds to 3769 in L21913 and 3912 in M88101, a location at which the two sequences differ naturally. At this position, the L21913 allele is a G, while that for M88101 is a C. With the three alleles, six possible genotypes were expected and observed: GG, CC, AA, GC, CA, and GA. To determine if this anomaly was a result of heterogeneous PCR products arising from simultaneous amplification of L21913- and M88101-based products, representative fragments of all six genotypes were cloned. After purification of the amplified products and checking that each product consisted of a single band, the TA Cloning Kit was used for cloning according to the manufacturer’s recommendation (Invitrogen, Carlsbad, CA, USA). Ten clones from each genotype were sequenced and analyzed for nucleotide variants as previously described (14). An example of a Consed alignment from selected clones representing GG, AA, and CC genotypes is shown in Figure 2. The sequences, produced by T7 primer (with SP6 primer-based sequencing for confirmation), confirm the anomaly of the three-nucleotide variants observed at this locus in the alignment of the genomic DNA sequences (Figure 1). The alignment shown in Figure 3 provided the confirmation that genotypes that included a C were due to the presence of heterogeneous fragments in the PCR products. While heterozygous genotypes without a C were true to form in the alignments of the sequences of the clones, those with a C were all homozygous, with either a C or the alternate allele. For example, the sequences of the four clones from individual 103 were either a GG or CC and not, as was expected, a GC, the genotype of the original amplified product (Figure 3). All C-containing genotypes also had a T genotype at VPIagcSNP-007, supporting the M88101 origin of C-containing sequences. It should be noted, however, that the 100% sequence homology of the expected amplified region in U78555 and L21913 does not exclude U78555 as the fragment within which the SNPs were detected.

Therefore, the method proposed here for verification of SNPs within genes that belong to multiple families requires a difference of at least 1 bp, which must be a non-variant site. While several approaches are available for SNP detection, the report described here seems to suggest that the validity of some oligonucleotide-based surveys may be affected if the targeted genes belong to multigene families.

REFERENCES

6. Halushka, M.K., J.B. Fan, K. Bentley, L.
Benchmarks


Comments by Eric Wong on the manuscript as well as on interpretation of the data are gratefully acknowledged. The NHGRI and USDA-SCD programs provided support for our work. Address correspondence to Dr. Edward J. Smith, 3130 Litton Reaves Hall, Virginia Tech, Blacksburg, VA 24061, USA. e-mail: esmith@vt.edu

Received 29 August 2000; accepted 2 January 2001.

Edward J. Smith, Gliceria E. Smith, and Li Shi
Virginia Tech
Blacksburg, VA, USA

Small Bones from Dried Mammal Museum Specimens as a Reliable Source of DNA

BioTechniques 30:732-736 (April 2001)

Although many museums now routinely archive animal tissue for molecular systematic studies, the vast majority of museum specimens predate the advent of molecular techniques and consist of dried skins, skeletons, and formalin-fixed, ethanol-preserved tissues. The difficulties of obtaining usable DNA from such specimens are well known (1,3,4,6,8), yet such specimens form a potentially invaluable source of data, especially for populations of animals that are extinct or locally extirpated. Engstrom et al. (3) pointed to the fact that bone may be a better source of DNA, particularly to obtain long fragments of DNA. However, the amount of material suggested by the authors from voucher specimens for PCR may be on the range of gram-quantities. During our studies of mitochondrial variation of fruit bats of the genus Sturnira, we were unable to amplify cytochrome b from standard phenol-chloroform extractions of museum specimens. Consequently, we refined extraction protocols using only milligram-quantities that have routinely yielded amplifiable DNA from up to 66-year-old museum skins, bones, and formalin-fixed, ethanol-preserved tissues from bats of the genus Sturnira. Our techniques are novel improvements of existing DNeasy® Tissue Extraction protocols (Qiagen, Valencia, CA, USA), apply to very tiny bones, and could potentially be used with other mammals and small vertebrates. These modifications are (i) repeated washing and rehydration of tissues; (ii) prolonged (up to 72 h) digestion and repeated additions of proteinase K; (iii) checking and adjusting the pH of the extracted DNA solution before adsorption on the Qiagen column; (iv) reducing the elution volume; and (v) amplification of short overlapping segments utilizing internal primers.

Samples consisted of museum specimens of Sturnira, frugivorous bats averaging about 20 g. Samples from dried specimens included bones (ribs 5–9 mm in length or 2–3 mm of the diaphysis of a wing bone) or 3 × 2 mm of dry skin with hair. Formalin-fixed, ethanol-preserved tissue samples included 3 × 2 mm pieces of liver or skin with muscle or 1–2 ribs. Liquid-preserved tissue was chopped into small (<0.5 mm) pieces; dry bone was crushed within a folded piece of a 10 × 10 cm weighing paper (Fisher Scientific, Pittsburgh, PA, USA) with a pair of needle-nose pliers. Chopped or crushed samples were placed in 1.5-mL tubes and washed 3–5 times in 250 μL PBS (9) for 10 min at 55°C with occasional vortex mixing. The sample was spun briefly,