DNA Sequence, December 2005; 16(6): 462–467

Taylor & Francis
Taylor & Francis Group

SHORT COMMUNICATION

Heterozygous indels as useful tools in the reconstruction of DNA sequences and in the assessment of ploidy level and genomic constitution of hybrid organisms

CARLA SOUSA-SANTOS[1], JOANA I. ROBALO[1], MARIA-JOÃO COLLARES-PEREIRA[2], & VITOR C. ALMADA[1]

[1]*Instituto Superior de Psicologia Aplicada (ISPA), Unidade de Investigação em Eco-Etologia, Rua Jardim do Tabaco 34, 1149-041 Lisboa, Portugal, and* [2]*Universidade de Lisboa, Faculdade de Ciências, Centro de Biologia Ambiental, Campo Grande, 1749-016 Lisboa, Portugal*

(Received 18 May 2005)

Abstract
In this paper we describe a simple approach using double peaks in chromatograms generated as artefacts in the vicinity of heterozygous indels, to identify the specific sequences present in individual strands of a given DNA fragment. This method is useful to assign bases in individuals that are heterozygous at multiple sites. In addition, the relative sizes of the double peaks help to determine the ploidy level and the relative contribution of the parental genomes in hybrids. Our interpretation was confirmed with the analysis of artificial mixtures of DNA of two different species. Results were robust with varying PCR and sequencing conditions. The applicability of this method was demonstrated in hybrids of the *Squalius alburnoides* complex and in heterozygotes of *Chondrostoma oligolepis*. Far from being limited to these fish models and the gene where it was tested (*beta-actin*), this sequence reconstruction methodology is expected to have a broader application.

Keywords: *Double peaks, interspecific hybrids, haplotype reconstruction, heterozygotes, beta-actin gene, cyprinid fish*

Introduction

When an organism is heterozygous for several linked sites in a given DNA fragment that is being sequenced, one major problem is to assign the individual bases in each heterozygous position to each of the parental genomes. This problem is found both in intraspecific studies and especially in interspecific hybrids whose parental sequences often differ in many nucleotide positions.

The presence of a heterozygous indel in a fragment of nuclear DNA generates a disturbance in the sequencing process characterized by a succession of false double peaks. Bhangale et al. (2005) used the double peaks generated to identify indels in a set of 330 human genes. Once identified, the length of the indel was inferred from the pattern of peaks by performing a pairwise alignment of bases corresponding to the two allelic sequences, obviating the need of having a previous knowledge of both of the sequences involved (Bhangale et al. 2005).

In this paper, we explore further potentialities of this approach. First, we show with intra and interspecific data how the two parental haplotypes can be read from the chromatogram if we previously know its characteristic indels. In addition, we evidence how this method can help to determine the ploidy of each hybrid and to access the relative contributions of the parental genomes, constituting a useful alternative tool to quantitative PCR methods.

Methods

To test this method in interspecific hybrids, we used individuals of the Iberian minnow *Squalius alburnoides* complex. Its hybrid origin resulted from interspecific

Correspondence: C. Sousa-Santos, Instituto Superior de Psicologia Aplicada (ISPA), Unidade de Investigação em Eco-Etologia, Rua Jardim do Tabaco, 34, 1149 041 Lisboa, Portugal. Tel: 35 1218811700. Fax: 35 1218860954. E-mail: carla.santos@ispa.pt

ISSN 1042-5179 print/ISSN 1029-2365 online
DOI: 10.1080/10425170500356065

crosses between *S. pyrenaicus* females (PP) and males from an unknown species (AA), generating $2n = 50$, $3n = 75$ and $4n = 100$ hybrid forms (reviewed in Alves et al. 2001) and reconstituted diploid non-hybrids with the nuclear AA genome of the missing paternal ancestor (Alves et al. 2002, Robalo et al. n.d.), which are morphologically distinct from the diploid hybrid form of the complex.

In order to be able to analyse the hybrid nuclear genomes, a total of 31 individuals of the parental species were analysed: 11 *S. pyrenaicus*, nine *S. carolitertii* and 11 diploid nonhybrid *S. alburnoides* (GenBank: AY943863–AY943896). Samples of *S. carolitertii* were used since in the river basins where *S. pyrenaicus* is absent, although the mtDNA found in *S. alburnoides* fish is also *S. pyrenaicus*-like, the complex seems to be maintained by crosses with males of *S. carolitertii* (CC) and by diploid hybrid males (CA) (Cunha et al. 2004; Pala and Coelho 2005). Samples from 19 hybrids of *S. alburnoides* (eight diploids, 10 triploids and one tetraploid) were used. The ploidy of the hybrids was previously determined by flow cytometry using fresh fin clips, following an adaptation of the method proposed by Lamatsch et al. (2000) (Collares-Pereira 1985).

To illustrate the applicability of the method in intraspecific studies we sampled 13 individuals of *Chondrostoma oligolepis* (formerly known as *Chondrostoma macrolepidotum*) another Iberian minnow with $2n = 50$ (Collares-Pereira 1985).

Total genomic DNA was extracted from fin clips preserved in ethanol by an SDS/proteinase-k based protocol (adapted from Sambrook et al. 1989). A total of 927 bp of the *beta-actin* gene was amplified using the primers For-5′-ATGGATGATGAAATTGCCGC-3′ and Rev-5′AGGATCTTCATGAGGTAGTC-3′ (Robalo et al. n.d.). The amplification process was conducted as follows: 35 cycles of [94°C(30 s), 55°C(40 s) and 72°C(1 min 30 s)]. Amplification and sequencing of DNA from six diploids and six triploids was repeated using different PCR conditions: 35 cycles of [94°C(30 s), 42°C(40 s), 72°C(1 min 30 s)]. The amplified fragment is homologous to a region of the *beta-actin* gene of *Cyprinus carpio* (GenBank: M24113), between the positions 1622 and 2550, including introns B and C and three exons. Each sample was sequenced in both directions with the same primers used for PCR. Sequences were aligned with BioEdit® v.5.0.6.

Ploidy assessment methodology

To test the hypothesis that unbalanced proportions of parental genomes can be detected in non-quantitative PCR products, we produced six groups of artificial hybrids simulating hybrid forms of the *S. alburnoides* complex. Each group included PA, PAA and PPA forms made with mixtures of re-suspended DNA from the same genome donors (previously sequenced and all differentiated by specific point mutations): 9 μl of each DNA suspension from the A- and P-genome donors to produce PA hybrids; 6 μl of DNA suspension from the P-genome donor and 12 μl from the A-genome donor for PAA hybrids; and the reversed quantities for PPA hybrids.

The contribution of each parental complement to the hybrid genome was quantified by the "measuring method": measuring both overlapping peaks in each position (using ImageTool 2.0 UTHSCSA® with a screen resolution of 1024 × 768 pixels) and calculating the ratio "height of peak from P/(height of peak from P + height of peak from A)" (P/P + A ratio). The provenience of the higher overlapping peak was also registered to calculate the percentage of P-peaks that were greater than A-peaks ("P-count")—"count method".

Different DNA concentrations caused by variations in the extraction procedure could be responsible for excesses of one of the genomes. The value of the P/A ratio for each position measured in an artificial diploid was used as a "correction coefficient" for the peak heights measured in the chromatograms of the artificial triploids made with the same genome donors. These corrected P-peak values were then used to calculate the P/P + A ratio in each position. In natural hybrids the P/A correction was unnecessary.

Results

There was no clear distinction between *S. pyrenaicus* and *S. carolitertii* for the analysed gene segment (one to four mutations between pairs of haplotypes) thus they were here designated as "*S. pyrenaicus*/*S. carolitertii*" (PP genome). The chromatograms of the parental species showed single peaks, except for one to four single nucleotide polymorphisms for some fish.

All hybrids sequenced showed double peaks in segments that varied between 544 and 667 bp (58.7 and 72.0% of the amplified fragment, respectively), involving the bases we expected to find if we overlapped the parental genomes. We assumed that analysing about a hundred double peaks should provide a sufficient number of distinct points to allow an adequate statistical analysis. Thus, a segment of 176 bp was randomly selected in the double peaks region, containing 118 positions with overlapping peaks (the remaining were single peaks as a result of the addition of equal bases—see Figure 1).

Reconstruction of the parental sequences in a hybrid

In the presence of one heterozygous indel in a fragment of nuclear DNA, the sequencer starts to read two bases in the same position, a situation that generates a pattern of overlapping peaks in the chromatogram. In these regions of double peaks, the

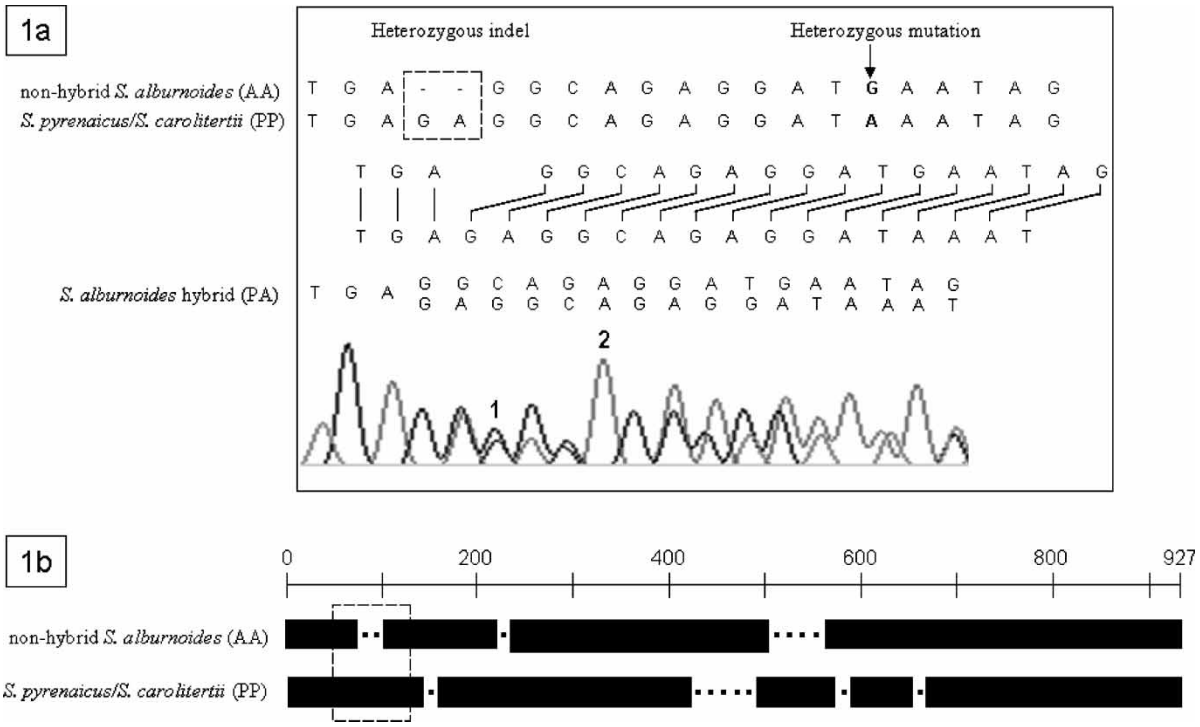

Figure 1. (1a) Demonstration of the disturbance generated by the first heterozygous indel in the sequencing process of a *S. alburnoides* individual with P and A complements. When the bases in both alleles are the same. The chromatogram shows a single higher peak reflecting the presence of double quantities of the same base, which means that the sequences can be read even in the homologous regions flanked by overlapping peaks. 1—double peak; 2—single peak resulting from the addition of equal bases. (1b) Schematic representation of the characteristic indels of the parental species in the analysed fragment of 927 bp (deletions are represented by black dots). Indel size ranged from one to five bases (mean = 1.88 ± 1.68). The broken line box indicates the localization of the indel represented in Figure 1a.

bases are out of phase as many positions as the number of bases of the indel (Figure 1a), a condition that will be maintained until a second indel of opposite direction counterbalances the first, which could be many dozens or hundreds of bases downstream from the initial indel. Thus, starting at the position where the first heterozygous indel occurs, it is possible to read the complements involved in the formation of a hybrid genome.

S. pyrenaicus/*S. carolitertii* and nuclear nonhybrid *S. alburnoides* differed by a total of seven indels. The reconstruction of the genomes of the hybrids was possible because each parental species had a characteristic number and location of indels (Figure 1b). It was also favoured by the existence of several point mutations characteristic of each parental species (in homozygous condition) that marked out highly conserved regions and made possible to ascribe unambiguously each peak in a double peaks region to the correct parental genome.

Information on the ploidy and hybrid genome constitution

Since the height of a peak in a chromatogram reflects approximately the amount of a specific base in that position, one can expect that different genome constitutions exhibit chromatograms with different peak heights. Exception is made when there is a suppression of signal at a given position because during the sequencing process the reading of a specific base may be affected by the constitution of the preceding one (for example, it is frequent to observe weak G's after C's or A's—see Hills et al. 1997 for more information on peak patterns). However, this "effect of the adjacent base" affects all the samples in the same way, as demonstrated by a strong correlation of the heights of the peaks between samples (for six samples, Spearman-R ranged between 0.88 and 0.97, for 70 analysed nucleotide positions of a homozygous segment).

After the reconstruction of the genomes involved, it was possible to assess the ploidy level and to determine the hybrid genome constitution. In artificial hybrids we obtained P/P + A mean values of 0.29 ± 0.08 for PAA and 0.70 ± 0.07 for PPA hybrids. For all of the six groups the P/P + A values of the PAA and PPA triploids did not overlap with those of the diploids (forced to be 0.50). All differences were highly significant even after Bonferroni correction for multiple comparisons ($N = 118$, p values = 4.5×10^{-20}–2.4×10^{-16}, Wilcoxon tests). The "count method" was also applied and the resulting P-count was significantly different: 0.71% ± 0.84 for PAA and 96.32% ± 5.76

Table I. Results from the measuring and count methods applied to the six groups of artificial hybrids. Mean values and standard deviations for P/P + A ratios (both raw and uncorrected values) and P-counts calculated from the analysis of 118 sites are presented for each group. Total means and standard deviations of the P/P + A ratios and P-counts for the three types of artificial hybrids are also presented.

Group	Measuring method						Count method	
	Raw P/P+A ratio			Corrected P/P+A ratio			P-count (%)	
	PA	PAA	PPA	cPA	cPAA	cPPA	PAA	PPA
1	0.45 ± 0.13	0.33 ± 0.11	0.63 ± 0.13	0.50	0.37 ± 0.06	0.69 ± 0.06	1.74	99.12
2	0.30 ± 0.18	0.15 ± 0.09	0.48 ± 0.21	0.50	0.29 ± 0.07	0.70 ± 0.11	0.85	97.46
3	0.31 ± 0.12	0.18 ± 0.08	0.58 ± 0.17	0.50	0.32 ± 0.04	0.77 ± 0.08	0.00	99.15
4	0.47 ± 0.12	0.34 ± 0.10	0.65 ± 0.10	0.50	0.36 ± 0.03	0.68 ± 0.02	0.00	100.00
5	0.27 ± 0.20	0.05 ± 0.04	0.58 ± 0.27	0.50	0.16 ± 0.10	0.79 ± 0.13	1.69	97.46
6	0.35 ± 0.17	0.16 ± 0.11	0.42 ± 0.21	0.50	0.26 ± 0.06	0.59 ± 0.10	0.00	84.75
total mean	0.36	0.20	0.56	0.50	0.29	0.70	0.71	96.32
total sd	0.08	0.11	0.09	0.00	0.08	0.07	0.84	5.76

for PPA hybrids ($p = 0.0036$, $N_1 = N_2 = 6$, Mann-Whitney test). The data for each group of artificial hybrids are summarized in Table I.

The application of the "count method" to natural hybrids showed that diploids had higher peaks attributable to P-genome (55.77% ± 6.88; range 47.01%–65.25%), which were significantly different from the 8.13% ± 0.95 (range 6.84%–9.40%) calculated for the PAA triploids ($p = 0.0005$, $N_1 = 8$, $N_2 = 9$, Mann-Whitney test). Among the triploids we found a morphologically distinct individual that had extremely high values of P-count (95.69%), suggesting a PPA genome. The P-count for the single tetraploid (1.69%) was lower than the smallest value for PAA triploids (6.84%), suggesting a PAAA constitution. Sections of chromatograms of the hybrids are shown in Figure 2.

The "measuring method" also returned significant differences: mean values of P/P + A for PA and PAA hybrids were, respectively, 0.53 ± 0.01 (range 0.51–0.55) and 0.34 ± 0.01 (range 0.33–0.36) ($p = 0.0005$, $N_1 = 8$, $N_2 = 9$, Mann-Whitney test). Values of P/P + A ratio for the morphologically distinct triploid (0.70) and for the tetraploid (0.23) also differed markedly from the remaining fish, corroborating the genomic constitutions suggested by the "count method".

The comparison between the measures taken for the same individuals in two different PCR and sequencing runs demonstrated a repeatable distinction between diploids and triploids. In the second PCR and sequencing of the same material, the mean P/P + A values were still significantly different between diploids and triploids: 0.58 ± 0.03 and 0.39 ± 0.01, respectively ($p = 0.0039$, $N_1 = N_2 = 6$, Mann-Whitney test). The results showed a high correlation between the P/P + A values obtained for the two chromatograms from the same individual for diploids and triploids (Spearman-R ranged between 0.66 and 0.77 and between 0.76 and 0.83, respectively).

A comparison of the results obtained when lower numbers of double peaks are analysed showed that all the differences between ploidies were recovered and the values showed only slight deviations from those obtained with the entire data series (Figure 3). Indeed, even groups of 20 overlapping peaks provided estimates that were sufficiently accurate to assign the

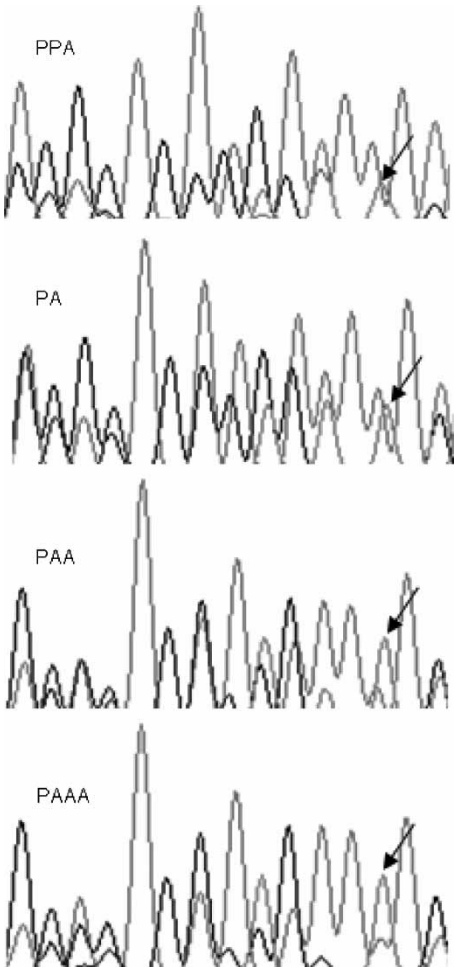

Figure 2. Sections of the chromatograms of *S. alburnoides* natural hybrids where different dosages of A and P genomes are evident. The arrow indicates a base attributed to the non-hybrid *S. alburnoides* progenitor (A genome) that increases its relative height from PPA to PA to PAA to PAAA.

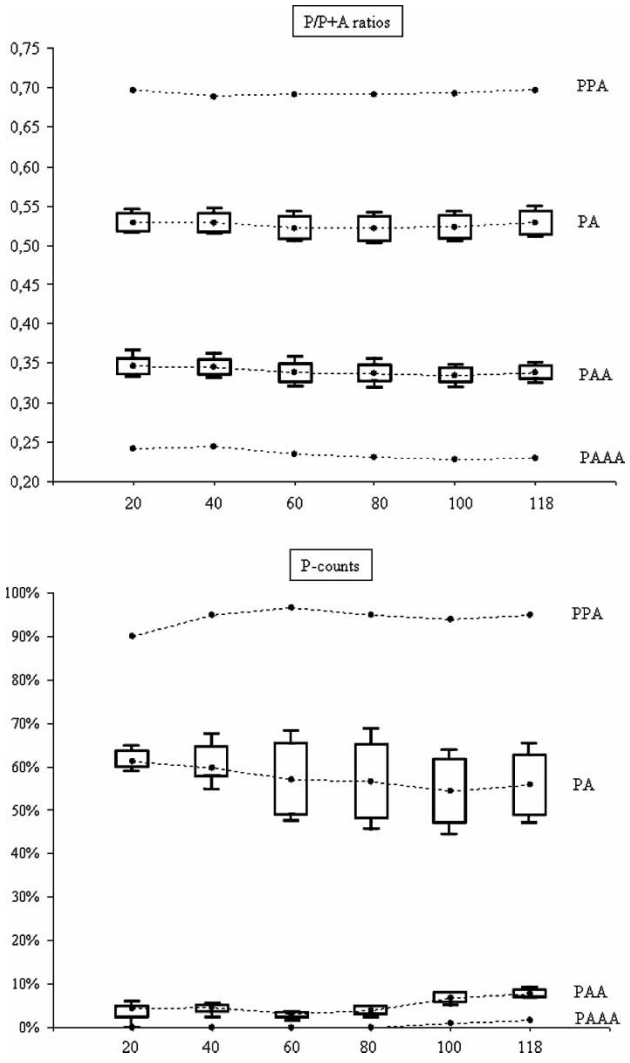

Figure 3. Plot of means, standard deviations, minimums and maximums of the P/P + A ratios and P-counts in PA diploids ($N = 8$), PAA triploids ($N = 9$), PPA triploid ($N = 1$) and PAAA tetraploid ($N = 1$), as a function of the number of double peaks analysed (20, 40, 60, 80, 100 and 118).

proportion of each genome in different ploidy groups. This means that even disturbances that are much shorter than the ones used in this study may recover basically the same ploidy information.

Application to intraespecific heterozygotes

Concerning intraspecific heterozygotes, four of the *C. oligolepis* samples sequenced revealed one heterozygous indel of seven bases in the same fragment of the *beta-actin* gene, generating a double peaks region of 823 bp. The reconstructed parental haplotypes were recovered in the remaining individuals (GenBank: AY943897–AY943905): the strand with the deletion was present in homozygosity in seven individuals (CO1 genome), and the other strand was present in two individuals homozygous for the insertion (CO2 genome). Measures of the peaks in the chromatograms of the heterozygotes yielded a CO1/CO1 + CO2 ratio of 0.58 ± 0.03, a value that is comparable to that obtained for the PA diploids of *S. alburnoides* from the same PCR and sequencing run ($p = 1.000$, $N_1 = 6$, $N_2 = 4$, Mann-Whitney test).

Discussion

The results demonstrated that the pattern of double peaks generated by heterozygous indels proved to be useful: (i) to reconstruct the parental sequences involved in large DNA segments of individuals that are heterozygotes for several linked sites and of interspecific hybrids; (ii) to determine the ploidy of the hybrids and (iii) to identify the relative contributions of the parental sequences in non-diploid hybrids. The method was repeatable in different PCR and sequencing conditions and the interpretation of ploidy and genomic proportions was always consistent. To control for possible variations we recommend that: in all PCR and sequencing runs at least one diploid hybrid should be included to serve as a standard control to that run; and results should be confirmed with forward and reverse sequencing.

One may ask which of the two methods described in this paper is the more accurate. The "count method" is much less time consuming and in the present study discriminated all the genomic constitutions that had been identified with the measuring method. However, we suggest that the "measuring method" should be preferred over the "count method" since it presents the great advantage of controlling for varying PCR artefacts and PCR conditions and to obviate the effects of neighbouring bases on the height of each peak. In addition, the "measuring method" is more reliable and powerful than the "count method" to discriminate different hybrid forms in which on of the parental complements is predominant—for instance, when we want to discriminate between a PAA and a PAAA individual.

When compared with allozyme electrophoresis, our method avoids killing the specimens and problems of regulation of gene expression. Although crude when compared with quantitative PCR procedures, it is sufficiently precise and inexpensive to be considered a useful tool in the study of interspecific hybrids and in population genetics.

DNA segments that harbour several and closely located indels provide ideal material for the application of this method as they provide several reference points that allow recovery of the specific sequences starting from both directions and minimize the risk of error caused by recombination. The process of sequence reconstruction is also favoured by the existence of several characteristic point mutations fixed for each parental species. These

mutations mark out highly conserved regions and make it possible to ascribe unambiguously each peak in a double peaks region to the correct parental genome. Although advantageous, these conditions are not essential and their absence would not necessarily make the method impracticable. In fact, as referred by Bhangale et al. (2005), since the length of the indel is inferred from the pattern of double peaks, the process does not require the presence of homozygotes for both of the alleles in the surveyed sample, a situation that widens its applications.

As flow cytometry methodology can only determine the ploidy level of the samples and not their exact genome composition, mainly when parental species have similar DNA contents, our approach also constitutes a valuable complementary tool for the analysis of hybrids with either balanced or non balanced parental genomes.

Acknowledgements

We thank the help of G. Lemos, Sousa-Santos family, T. Bento and A. Levy. We are also grateful to A. Gomes-Ferreira, M. Gromicho and L. Moreira da Costa for technical assistance with flow cytometry. DGF provided authorizations for field work. Study funded by the FCT Pluriannual Program (UI&D 331/94 and UI&D 329/94) (FEDER participation). C. Sousa-Santos was supported by a PhD grant from FCT (SFRH/BD/8320/2002).

References

Alves MJ, Coelho MM, Collares-Pereira MJ. 2001. Evolution in action through hybridization and polyploidy in an Iberian freshwater fish: A genetic review. Genetica 111:375–385.

Alves MJ, Collares-Pereira MJ, Dowling TE, Coelho MM. 2002. The genetics of maintenance of an all-male lineage in the *Squalius alburnoides* complex. J Fish Biol 60:649–662.

Bhangale TR, Rieder MJ, Livingston RJ, Nickerson DA. 2005. Comprehensive identification and characterization of diallelic insertion-deletion polymorphisms in 330 human candidate genes. Hum Mol Genet 14:59–69.

Collares-Pereira MJ. 1985. Cytotaxonomic studies in Iberian Cyprinids. II. Karyology of *Anaecypris hispanica* (Steindachner, 1866), *Chondrostoma lemmingi* (Steindachner, 1866), *Rutilus arcasi* (Steindachner, 1866) and *R. macrolepidotus* (Steindachner, 1866). Cytologia 50:879–890.

Cunha C, Coelho MM, Carmona JA, Doadrio I. 2004. Phylogeographical insights into the origins of the *Squalius alburnoides* complex via multiple hybridization events. Mol Ecol 13:2807–2817.

1997. ABRF'97: Techniques at the genome/proteome interface—DNA sequence analysis, Available from http://www.biotech.iastate.edu/facilities/DSSF/ABRF/default.html via the INTERNET. (Accessed 2005 August 31)

Lamatsch DK, Steinlein C, Schmid M, Schartl M. 2000. Noninvasive determination of genome size and ploidy level in fishes by flow cytometry: Detection of triploid *Poecilia formosa*. Cytometry 39:91–95.

Pala I, Coelho MM. 2005. Contrasting views over a hybrid complex: Between speciation and evolutionary "dead-end". Gene 347:283–294.

Robalo J, Sousa-Santos C, Levy A, Almada V. n.d. Molecular insights on the taxonomic position of the paternal ancestor of the *Squalius alburnoides* hybridogenetic complex. Mol Phylog Evol.

Sambrook J, Fritsch EF, Maniatis T. 1989. Molecular cloning: A laboratory manual. 2nd ed., New York: Cold Spring Harbor Laboratory Press.