Molecular Ecology Notes
July 2007, Volume 7 Issue 4, Pages 562 - 566
http://dx.doi.org/10.1111/j.1471-8286.2007.01685.x

Mark–recapture cloning: a straightforward and cost-effective cloning method for population genetics of single-copy nuclear DNA sequences in diploids

N. Bierne[1, *], A. Tanguy[2], M. Faure[1], B. Faure[2, 3], E. David[4], I. Boutet[4], E. Boon[1], N. Quere[2], S. Plouviez[1, 2], P. Kemppainen[5], D. Jollivet[2], D. Moraga[4], P. Boudry[3], P. David[6]

[1] Génome Populations Interactions Adaptation, UMR 5171, Université Montpellier II - IFREMER - CNRS, Station Méditerranéenne de l'Environnement Littoral, 34200 Sète, France.
[2] Equipe Evolution & Génétique des Populations Marines, UMR 7144, UPMC - CNRS, Station biologique de Roscoff, BP. 74, Place Georges Teissier, 29682 Roscoff, France.
[3] Laboratoire de Génétique et Pathologie, IFREMER, 17390 La Tremblade, France.
[4] Laboratoire des sciences de l'environnement marin, UMR CNRS 6539, Institut Universitaire Européen de la Mer, Université de Bretagne Occidentale 29280 Plouzané, France.
[5] Department of Marine Ecology, Tjärnö Marine Biological Laboratory, 45296 Strömstad, Sweden.
[6] Centre d'Ecologie Fonctionnelle et Evolutive - CNRS, 34293 Montpellier cedex 5, France.

*: Corresponding author : N. Bierne, email address : n-bierne@univ-montp2.fr

Abstract:

We describe a simple protocol to reduce the number of cloning reactions of nuclear DNA sequences in population genetic studies of diploid organisms. Cloning is a necessary step to obtain correct haplotypes in such organisms, and, while traditional methods are efficient at cloning together many genes of a single individual, population geneticists rather need to clone the same locus in many individuals. Our method consists of marking individual sequences during the polymerase chain reaction (PCR) using 5'-tailed primers with small polynucleotide tags. PCR products are mixed together before the cloning reaction and clones are sequenced with universal plasmid primers. The individual from which a sequence comes from is identified by the tag sequences upstream of each initial primer. We called our protocol mark–recapture (MR) cloning. We present results from 57 experiments of MR cloning conducted in four distinct laboratories using nuclear loci of various lengths in different invertebrate species. Rate of capture (proportion of individuals for which one or more sequences were retrieved) and multiple capture (proportion of individuals for which two or more sequences were retrieved) empirically obtained are described. We estimated that MR cloning allowed reducing costs by up to 70% when compared to conventional individual-based cloning. However, we recommend to adjust the mark:recapture ratio in order to obtain multiple sequences from the same individual and circumvent inherent technical artefacts of PCR, cloning and sequencing. We argue that MR cloning is a valid and reliable high-throughput method, providing the number of sequences exceeds the number of individuals initially amplified.

Keywords: High throughput allele recognition, Sequence, DNA polymorphism, Population Genetics

The analysis of gene genealogies by increasingly powerful methods (Balding *et al.* 2001; Slatkin & Veuille 2002; Zhang & Hewitt, 2003) and the development of methods to quantify adaptation at the molecular level (Yang & Bielawski 2000; Fay & Wu 2001) make DNA sequence a major tool in population genetics. Although the literature abounds in studies of mitochondrial DNA (mtDNA) and concertedly evolving multiple-copy ribosomal DNA (rDNA) loci, the analysis of single-copy nuclear DNA sequences remains surprisingly infrequent and limited to model organisms (Zhang & Hewitt 2003). The lack of reference sequences in non-model organisms does not explain everything (Zhang & Hewitt 2003). Another technical challenge is the difficulty to identify alleles in heterozygous state in outcrossing diploid organisms. Heterozygous individuals at a given locus have two different alleles that should ideally be sequenced independently. Even though alternative methods exist and are continuously explored (Zhang & Hewitt 2003), cloning of PCR products often remains an essential step. While PCR and sequencing have become universally-used low-cost techniques, individual cloning still remains time consuming and expensive. As a consequence, molecular ecologists endeavour to avoid the cloning procedure when possible, restricting the analysis of DNA sequences to mtDNA, rDNA or sex chromosomes in the hemizygous sex when available, or losing the benefit of genealogical information by typing SNPs, even when nucleotide diversity is high. When individual cloning is performed, the cost increases proportionally to sample size, setting a strong limit to the latter.

Here we describe a simple protocol that allows the cloning of PCR products of several individuals from a population sample at once, leading to a less time- and resource consuming cloning procedure. Our method is based on the observation that cloning can separate single alleles from several individuals as well as it does within a single individual. A simple solution to reduce the number of cloning reactions would therefore be to pool the PCR products of several individuals before cloning and to sequence many clones (e.g. Kronforst *et al.* 2006).

However, with such a procedure it is no longer possible to know the individual from which an allele sequence comes from. To solve this problem, PCR products need to be individually marked. The method we found consists of marking individual sequences during the PCR using slightly different primer pairs for each individual. To this aim, every primer is 5'-tailed with a small poly-nucleotide tag. Tags do not match the matrix DNA sequence in the initial stages of the PCR and does not perturb the reaction. The method is essentially similar to the M13 tailing technique (Oetting *et al.* 1995) although the tail is much smaller. PCR products of similar quantities are mixed together and cloned with standard protocols. Clones are then sequenced with universal plasmid primers flanking the insert. The small poly-nucleotide tags upstream of primers are therefore sequenced and allow identifying the individual from which the sequence comes from. Using the combination of the forward and reverse primers, it is not necessary to use different primer pairs for each PCR-amplified individual. For instance, we usually used eight different tags for the forward primers and six for the reverse primers, yielding 48 unique combinations by which sequences can be recognised.

PCR products were quantified on agarose gel stained with ethidium bromide then mixed together in such a way as to equalise concentration of each PCR product. Pools of PCR products were purified with the QIAquick PCR purification kit or the QIAEX II Purification Kit (Qiagen, Crawley, UK), and cloned with the pGEM-T Vector System (Promega, WI, USA) according to manufacturer's recommendations. Positive clones were screened for the presence of appropriate-sized inserts by PCR amplifications then sent to the Genoscope platform (Evry, France; http://www.genoscope.cns.fr/) where plasmid extraction and sequencing with vector-specific primers SP6 (5'-TATTTAGGTGACACTATAG-3') and T7 (5'-TAATACGACTCACTATAGGG-3') were performed.

The method has been tested in four distinct laboratories accounting for 57 experiments of MR-cloning using various species of marine invertebrates and genes (supplementary Table

1). We present observed rates of capture (i.e. the proportion of individuals for which one or more sequences was obtained), technical artefacts we have encountered and recommendations to accommodate artefacts in the lab or during statistical analysis.

Stochastic processes during PCR, ligation, transformation and bacterial growth can sometimes generate an overrepresentation of a few sequences at the end of the experiment. To circumvent this drift effect, we choose to pool an appreciable number of individuals (usually 48 which corresponds to half a PCR plate). Our aim was not to capture every individual of the initial sample. The average number of sequences obtained and number of individuals captured in each experiment are given in supplementary Table 1. The rate or capture (number of individuals captured / initial sample size) increased with the capture effort (number of sequences / initial sample size) but was on average slightly lower than the expectation based on a uniform distribution (Figure 1). The rate of multiple capture which provides more reliable data (see below) increased linearly with the capture effort (slope = 0.2) for the range of capture effort investigated in this study (Figure 2).

In the course of the development of the protocol, we encountered a number of technical artefacts. First, a number of tags were partially or totally deleted during the cloning process. Tag deletion led to an average rate of unassigned sequences of ~7%, but this rate was highly variable depending on the locus studied (supplementary Table 1). We suspect that the sequence upstream of the primer in the matrix DNA may have an impact because a high rate of deletion has been observed for a primer immediately designed after a poly-T repetition (25%). However, other primers sometimes reached as a high rate of deletion without any visible distinctiveness at the DNA primary structure. Unassigned sequences should not inevitably be removed from the data analysis (see Kronforst *et al.* 2006) but the consequences of their use need to be considered. Second, the impact of classical technical artefacts usually encountered in this kind of protocol –i.e. mutation during PCR, cloning and sequencing, is not easy to appreciate with our technique. We expect an individual to have a maximum number of two different sequences (*i.e.* alleles) and when two sequences are observed, the divergence should be in accordance with the global diversity observed. A small proportion of individuals captured several times displayed more than two alleles (~8%). However, in such cases differences were only due to the presence of a single artefactual mutation in one sequence. We also observed individuals with two alleles, of which one was sequenced only once, differing by a single nucleotide, while the average pairwise difference in the whole sample was much greater. Thirdly and most problematically, we observed in a few cases multiple captured individuals for which more than two alleles presented such a divergence that sequence misassignment to this individual was the only valid explanation. Misassignment can occur owing to a mutation in a tag (during PCR, ligation or bacterial replication) or *in vitro* recombination. Indeed, in some instances one of the sequences retrieved was in good agreement with an event of recombination between divergent alleles present in our sample.

We found no satisfactory solution for tag deletion. Initial experiments were conducted with two-nucleotide tags which was enough to create our 14 primers. Tag length was sometimes increased in successive experiments with no significant impact on this problem. We observed a strong variation in the rate of tag deletion according to the locus analysed (supplementary Table 1). We therefore suspect an effect of the primer sequence (hairpin or duplex effect) or the sequence upstream of the primer, although we were unable to find convincing evidence for such an effect.

The problem of artefactual mutations could be circumvented by restricting genetic data analysis to alleles captured several times. However, the rate of artefactual mutations was always low. One can then compare the results obtained with reliable alleles (for which several sequences were captured) and results obtained with the whole dataset. Because artefactual mutations should mainly create singletons (mutations observed in a single sequence of the dataset) an interesting parameter to evaluate in this respect is the proportion of singleton mutations. One can also choose the dataset required depending on the analysis conducted. For instance, any sequences can be used in most analyses of molecular evolution that compare the relative rate of evolution between different categories of mutations within the same sequence (synonymous, non-synonymous, non-coding, indels). The McDonald-Kreitman test (McDonald & Kreitman 1991a) falls in this category of analysis (McDonald & Kreitman 1991b). In addition, singletons can sometimes be removed from the data in some analyses of molecular evolution (e.g. Bierne & Eyre-Walker 2004; Andolfatto 2005). Here, attention might be called to the fact that such a technical artefact is an ubiquitous problem not restrained to the MR-cloning protocol (Zhang & Hewitt 2003).

Misassignment (tag mutation or *in vitro* recombination) could have been a serious problem if the rate was high. When nucleotide diversity is low, misassignment can easily be confounded with standard artefactual mutations. Luckily, marine invertebrates usually exhibit very high nucleotide diversities (π often > 0.01, Table 1). We were able not only to detect misassignment, but also to estimate its rate. The rate of misassignment turned out to be low (<2%, supplementary Table 1). The occurrence of *in vitro* recombination is known to occur at a non-negligible rate during PCR (Meyerhans *et al.* 1990) or cloning (Tang & Unnasch 1995). Such chimeric DNA products are well-known in surveys of bacterial 16S rRNA genes (Kopczynski *et al.* 1994). However, this artefact is not easily detected when nucleotide diversity is low. We argue that *in vitro* recombination is not a more serious bias in MR-cloning than in standard protocols but is detected in multiple captures (recombination during PCR) or because of tags rearrangement (recombination during cloning). As for artefactual mutations, the problem can be solved by restricting genetic data analysis to alleles captured several times.

Finally, we would need to estimate the time/money saved with MR-cloning over standard protocols for a comparable amount of data collected. The time saved seems obvious to us as a cloning reaction is far more time-consuming than a sequencing reaction; especially when accounting for the recent technical progress made in the automatisation of sequencing. In addition, sequencing platforms have flourished and the sequencing step is increasingly outsourced to these platforms. Estimating the money saved is more difficult because costs and lab facilities can vary widely among laboratories and countries. First, we used our estimated costs of primers, PCR, PCR product purification, cloning and sequencing reactions to evaluate the cost of a MR-cloning. Then, using our empirical rate of capture (quartic regression in figure 1) we estimated the cost of obtaining the same final number of sequences with standard individual-based cloning protocols. However, the estimate we made is an underestimation because we neglected our salaries in the calculation. To take costs of manual work into account, we used in a second estimate prices given by a private company (information one can easily get on the web). The financial gain of a MR-cloning protocol primarily depends on the ratio of the cost of a cloning reaction to the cost of a sequencing reaction which turned out to be 5 in our case but was estimated to be 15 from the costs provided by private companies. The relative cost of MR-cloning to standard protocols of individual cloning is presented in Figure 3 as a function of the sample size for population genetics analysis. As expected, the bigger is the final sample size, the more is the saving of money provided by MR-cloning. MR-cloning was estimated reducing costs by up to 70% when compared to conventional individual-based cloning (Figure 3). We do not claim that MR-cloning would be so cost-effective in every lab. In addition, one may not plan to obtain a big sample size simply to save money while the genetic information sought could emerge in a small sample size (e.g. Felsenstein 2005). However, we would argue that big sample sizes can often be highly valuable for population genetics inference in non-equilibrium populations for instance when it allows sampling the rare lineage that has survived a bottleneck or a selective sweep or that has introgressed through a barrier to gene flow.

We would conclude that MR-cloning is a valid and reliable high-throughput method. From the experience we gained with MR-cloning, we would recommend to use an appreciable effort of capture (say 2-3) in order to obtain multiple sequences from the same individual (see Figure 2) and circumvent inherent technical artefacts of PCR, cloning and sequencing. However, the level of precision required depends on the nucleotide diversity observed and the data analysis one wants to conduct. MR-cloning offers an opportunity to appreciate the consequences of technical artefacts by comparing more or less stringent datasets (e.g. raw datasets to datasets restricted to sequences obtained more than once).

Acknowledgements

The authors are very indebted to the National Sequencing Centre (Genoscope at Evry) for providing access to their sequencing platform. They thank François Bonhomme, Grant Pogson and Tobias Lindskog for helpful discussions. This research was funded by the national programme GIS "Génomique Marine", by the Fish & Shellfish node of the NoE "Marine Genomics Europe" and by La Région Bretagne. This is publication GPIA 2006-XXX.

References

Andolfatto P (2005) Adaptive evolution of non-coding DNA in Drosophila. *Nature*, **437**, 1149-1152.

Balding DJ, Bishop M, Cannings C (2001) Handbook of statistical genetics, p. 847. John Wiley & Sons, Ltd., Chichester.

Bierne N, Eyre-Walker A (2004) The genomic rate of adaptive amino-acid substitution in Drosophila. *Molecular Biology and Evolution*, **21**, 1350-1360.

Fay JC, Wu CI (2001) The neutral theory in the genomic era. *Current Opinion in Genetics and Development*, **11**, 642-646.

Felsenstein J (2005) Accuracy of coalescent likelihood estimates: do we need more sites, more sequences, or more loci? *Molecular Biology and Evolution*, **23**, 691-700.

Kopczynski ED, Bateson MM, Ward DM (1994) Recognition of chimeric small-subunit ribosomal DNAs composed of genes from uncultivated microorganisms. *Applied and Environmental Microbiology*, **60**, 746-48.

Kronforst MR, Young LG, Blume LM, Gilbert LE (2006) Multilocus analyses of admixture and introgression among hybridizing *Heliconius* butterflies. *Evolution*, **60**, 1254-68.

McDonald JH, Kreitman M (1991a) Adaptive evolution at the Adh locus in Drosophila. *Nature*, **351**, 652-654.

McDonald JH, Kreitman M (1991b) Neutral mutation hypothesis test. *Nature*, **354**, 116.

Meyerhans A, Vartanian JP, Wain-Hobson S (1990) DNA recombination during PCR. *Nucleic Acids Research*, **18**, 1687-91.

Oetting WS, Lee HK, Flanders DJ, et al. (1995) Linkage analysis with multiplexed short tandem repeat polymophisms using infrared fluorescence and M13 tailed primers. *Genomics*, **30**, 450-458.

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

Slatkin M, Veuille M (2002) Modern developments in theoritical population genetics, the legacy of Gustave Malecot. Oxford University Press, Oxford.

Tang J, Unnasch TR (1995) Discriminating PCR artifacts using directed heteroduplex analysis (DHDA). *Biotechniques*, **19**, 902-905.

Yang Z, Bielawski JP (2000) Statistical methods for detecting molecular adaptation. *Trends in Ecology and Evolution*, **15**, 496-503.

Zhang DX, Hewitt GM (2003) Nuclear DNA analyses in genetic studies of populations: practice, problems and prospects. *Molecular Ecology*, **12**, 563-584.

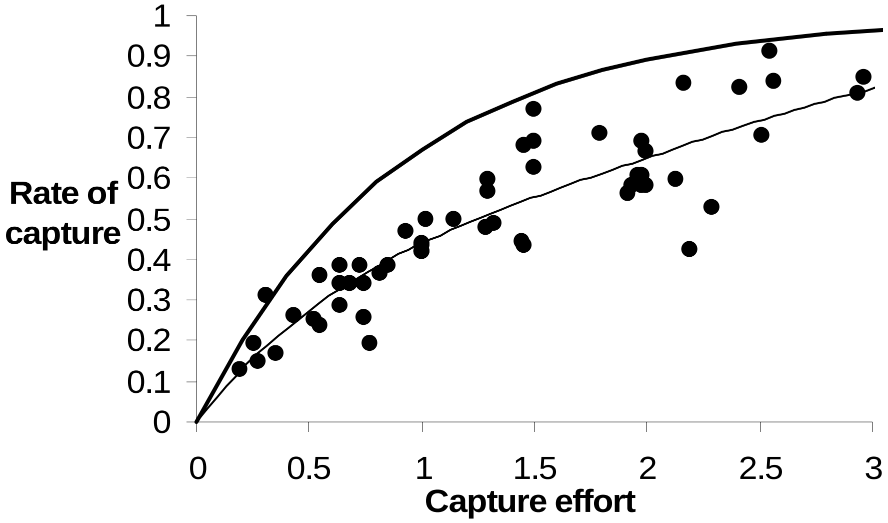
Figure legends

Figure 1: Rate of capture (number of individuals captured / initial sample size) as a function of the capture effort (number of sequences / initial sample size). The thick line is the expectation based on a uniform distribution and the thin line is a quartic polynomial regression on the data.

Figure 2: Rate of multiple captures (number of individuals captured more than once / initial sample size) as a function of the capture effort (number of sequences / initial sample size). The line is a linear regression on the data (slope = 0.2).

Figure 3: Estimated cost of MR-cloning protocols relative to standard protocols of individual cloning as a function of the sample size for population genetics analysis. The empirically estimated rate of capture of figure 1 was used for a gradient of initial sample size (the number of individuals PCR-amplified with tagged primers in the MR-cloning), and a gradient of capture effort (number of sequences performed / initial sample size). Numbers closed to curves indicate the initial sample sizes for MR-cloning. Each curve is generated with efforts of capture ranging from one to three. The upper series of curves are estimates that neglect salary costs (based on the prices we get for molecular biology kits and products) and the lower series of curves are estimates that include salary costs (based on prices practiced by private companies for a complete outsource of the experiment).

Figure 1



Rate of capture
Capture effort
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.5
1
1.5
2
2.5
3

Figure 2

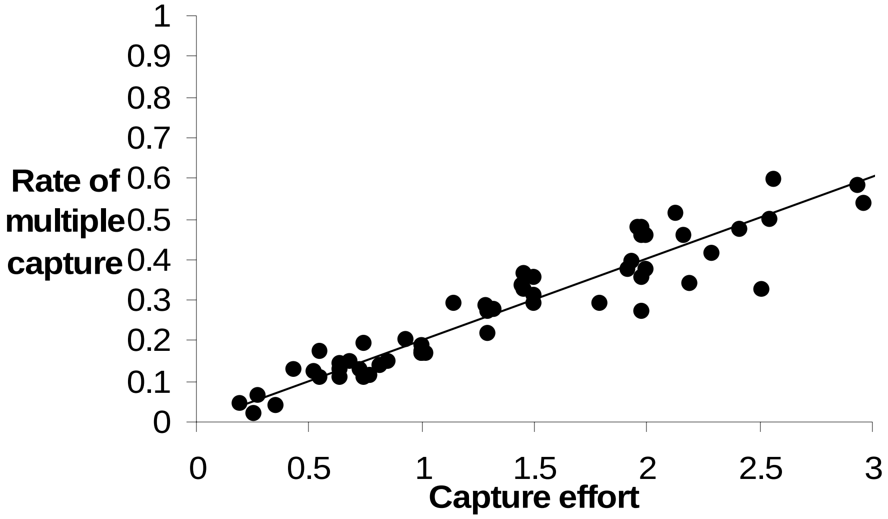
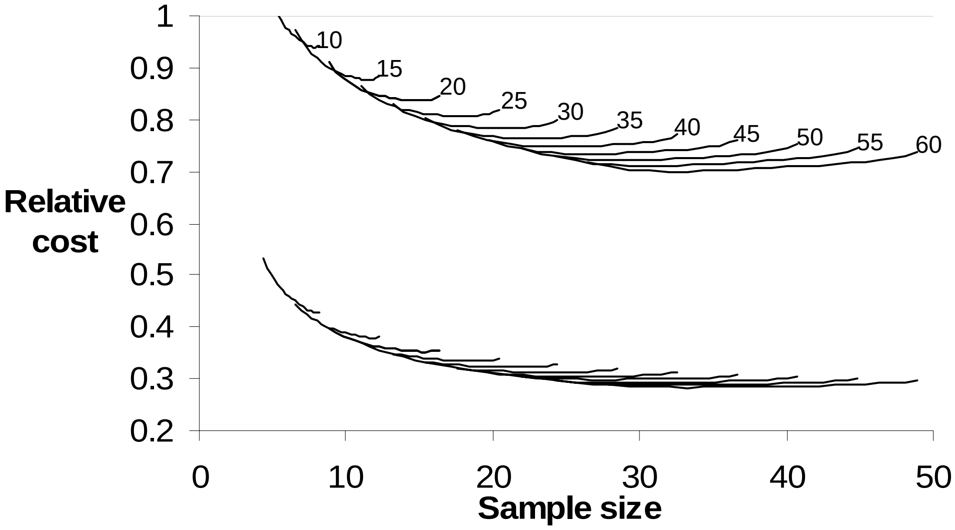

Rate of multiple capture
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.5
1
1.5
2
2.5
3
Capture effort

Figure 3


Relative cost
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0
10
20
30
40
50
Sample size
10
15
20
25
30
35
40
45
50
55
60