



HAL
open science

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

Nicolas Bierne, Arnaud Tanguy, Matthieu Faure, Jean-Baptiste Faure, Elise David, Isabelle Boutet, E. Boon, Nolwenn Quéré, Sophie Plouviez, P. Kemppainen, et al.

► To cite this version:

Nicolas Bierne, Arnaud Tanguy, Matthieu Faure, Jean-Baptiste Faure, Elise David, et al.. Mark-recapture cloning: a straightforward and cost-effective cloning method for population genetics of single-copy nuclear DNA sequences in diploids. *Molecular Ecology Notes*, 2007, 7 (4), pp.562-566. 10.1111/j.1471-8286.2007.01685.x . hal-00670005

HAL Id: hal-00670005

<https://hal.univ-brest.fr/hal-00670005v1>

Submitted on 18 Oct 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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. Tanguy², M. Faure¹, B. Faure^{2,3}, E. David⁴, I. Boutet⁴, E. Boon¹, N. Quere², S. Plouviez^{1,2}, P. Kempainen⁵, D. Jollivet², D. Moraga⁴, P. Boudry³, P. David⁶

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

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

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

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

15 PCR products were quantified on agarose gel stained with ethidium bromide then mixed
16 together in such a way as to equalise concentration of each PCR product. Pools of PCR
17 products were purified with the QIAquick PCR purification kit or the QIAEX II Purification
18 Kit (Qiagen, Crawley, UK), and cloned with the pGEM-T Vector System (Promega, WI,
19 USA) according to manufacturer's recommendations. Positive clones were screened for the
20 presence of appropriate-sized inserts by PCR amplifications then sent to the Genoscope
21 platform (Evry, France; <http://www.genoscope.cns.fr/>) where plasmid extraction and
22 sequencing with vector-specific primers SP6 (5'-TATTTAGGTGACACTATAG-3') and T7
23 (5'-TAATACGACTCACTATAGGG-3') were performed.

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

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

4 Stochastic processes during PCR, ligation, transformation and bacterial growth can
5 sometimes generate an overrepresentation of a few sequences at the end of the experiment. To
6 circumvent this drift effect, we choose to pool an appreciable number of individuals (usually
7 48 which corresponds to half a PCR plate). Our aim was not to capture every individual of the
8 initial sample. The average number of sequences obtained and number of individuals captured
9 in each experiment are given in supplementary Table 1. The rate of capture (number of
10 individuals captured / initial sample size) increased with the capture effort (number of
11 sequences / initial sample size) but was on average slightly lower than the expectation based
12 on a uniform distribution (Figure 1). The rate of multiple capture which provides more
13 reliable data (see below) increased linearly with the capture effort (slope = 0.2) for the range
14 of capture effort investigated in this study (Figure 2).

15 In the course of the development of the protocol, we encountered a number of technical
16 artefacts. First, a number of tags were partially or totally deleted during the cloning process.
17 Tag deletion led to an average rate of unassigned sequences of ~7%, but this rate was highly
18 variable depending on the locus studied (supplementary Table 1). We suspect that the
19 sequence upstream of the primer in the matrix DNA may have an impact because a high rate
20 of deletion has been observed for a primer immediately designed after a poly-T repetition
21 (25%). However, other primers sometimes reached as a high rate of deletion without any
22 visible distinctiveness at the DNA primary structure. Unassigned sequences should not
23 inevitably be removed from the data analysis (see Kronforst *et al.* 2006) but the consequences
24 of their use need to be considered. Second, the impact of classical technical artefacts usually
25 encountered in this kind of protocol –i.e. mutation during PCR, cloning and sequencing, is not

1 easy to appreciate with our technique. We expect an individual to have a maximum number of
2 two different sequences (*i.e.* alleles) and when two sequences are observed, the divergence
3 should be in accordance with the global diversity observed. A small proportion of individuals
4 captured several times displayed more than two alleles (~8%). However, in such cases
5 differences were only due to the presence of a single artefactual mutation in one sequence.
6 We also observed individuals with two alleles, of which one was sequenced only once,
7 differing by a single nucleotide, while the average pairwise difference in the whole sample
8 was much greater. Thirdly and most problematically, we observed in a few cases multiple
9 captured individuals for which more than two alleles presented such a divergence that
10 sequence misassignment to this individual was the only valid explanation. Misassignment can
11 occur owing to a mutation in a tag (during PCR, ligation or bacterial replication) or *in vitro*
12 recombination. Indeed, in some instances one of the sequences retrieved was in good
13 agreement with an event of recombination between divergent alleles present in our sample.

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

21 The problem of artefactual mutations could be circumvented by restricting genetic data
22 analysis to alleles captured several times. However, the rate of artefactual mutations was
23 always low. One can then compare the results obtained with reliable alleles (for which several
24 sequences were captured) and results obtained with the whole dataset. Because artefactual
25 mutations should mainly create singletons (mutations observed in a single sequence of the

1 dataset) an interesting parameter to evaluate in this respect is the proportion of singleton
2 mutations. One can also choose the dataset required depending on the analysis conducted. For
3 instance, any sequences can be used in most analyses of molecular evolution that compare the
4 relative rate of evolution between different categories of mutations within the same sequence
5 (synonymous, non-synonymous, non-coding, indels). The McDonald-Kreitman test
6 (McDonald & Kreitman 1991a) falls in this category of analysis (McDonald & Kreitman
7 1991b). In addition, singletons can sometimes be removed from the data in some analyses of
8 molecular evolution (e.g. Bierne & Eyre-Walker 2004; Andolfatto 2005). Here, attention
9 might be called to the fact that such a technical artefact is an ubiquitous problem not
10 restrained to the MR-cloning protocol (Zhang & Hewitt 2003).

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

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

1 instance when it allows sampling the rare lineage that has survived a bottleneck or a selective
2 sweep or that has introgressed through a barrier to gene flow.

3 We would conclude that MR-cloning is a valid and reliable high-throughput method.

4 From the experience we gained with MR-cloning, we would recommend to use an appreciable
5 effort of capture (say 2-3) in order to obtain multiple sequences from the same individual (see
6 Figure 2) and circumvent inherent technical artefacts of PCR, cloning and sequencing.

7 However, the level of precision required depends on the nucleotide diversity observed and the
8 data analysis one wants to conduct. MR-cloning offers an opportunity to appreciate the
9 consequences of technical artefacts by comparing more or less stringent datasets (e.g. raw
10 datasets to datasets restricted to sequences obtained more than once).

11

12 **Acknowledgements**

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

1 **References**

- 2 Andolfatto P (2005) Adaptive evolution of non-coding DNA in *Drosophila*. *Nature*, **437**,
3 1149-1152.
- 4 Balding DJ, Bishop M, Cannings C (2001) Handbook of statistical genetics, p. 847. John
5 Wiley & Sons, Ltd., Chichester.
- 6 Bierne N, Eyre-Walker A (2004) The genomic rate of adaptive amino-acid substitution in
7 *Drosophila*. *Molecular Biology and Evolution*, **21**, 1350-1360.
- 8 Fay JC, Wu CI (2001) The neutral theory in the genomic era. *Current Opinion in Genetics
9 and Development*, **11**, 642-646.
- 10 Felsenstein J (2005) Accuracy of coalescent likelihood estimates: do we need more sites,
11 more sequences, or more loci? *Molecular Biology and Evolution*, **23**, 691-700.
- 12 Kopczynski ED, Bateson MM, Ward DM (1994) Recognition of chimeric small-subunit
13 ribosomal DNAs composed of genes from uncultivated microorganisms. *Applied and
14 Environmental Microbiology*, **60**, 746-48.
- 15 Kronforst MR, Young LG, Blume LM, Gilbert LE (2006) Multilocus analyses of admixture
16 and introgression among hybridizing *Heliconius* butterflies. *Evolution*, **60**, 1254-68.
- 17 McDonald JH, Kreitman M (1991a) Adaptive evolution at the *Adh* locus in *Drosophila*.
18 *Nature*, **351**, 652-654.
- 19 McDonald JH, Kreitman M (1991b) Neutral mutation hypothesis test. *Nature*, **354**, 116.
- 20 Meyerhans A, Vartanian JP, Wain-Hobson S (1990) DNA recombination during PCR.
21 *Nucleic Acids Research*, **18**, 1687-91.
- 22 Oetting WS, Lee HK, Flanders DJ, et al. (1995) Linkage analysis with multiplexed short
23 tandem repeat polymorphisms using infrared fluorescence and M13 tailed primers.
24 *Genomics*, **30**, 450-458.
- 25 Sambrook J, Fritsch EF, Maniatis T (1989) Molecular cloning: a laboratory manual (2nd ed.)
26 Cold Spring Harbor Laboratory, New York.
- 27 Slatkin M, Veuille M (2002) Modern developments in theoretical population genetics, the
28 legacy of Gustave Malecot. Oxford University Press, Oxford.
- 29 Tang J, Unnasch TR (1995) Discriminating PCR artifacts using directed heteroduplex
30 analysis (DHDA). *Biotechniques*, **19**, 902-905.
- 31 Yang Z, Bielawski JP (2000) Statistical methods for detecting molecular adaptation. *Trends in
32 Ecology and Evolution*, **15**, 496-503.
- 33 Zhang DX, Hewitt GM (2003) Nuclear DNA analyses in genetic studies of populations:
34 practice, problems and prospects. *Molecular Ecology*, **12**, 563-584.

1 Figure legends

2

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

7

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

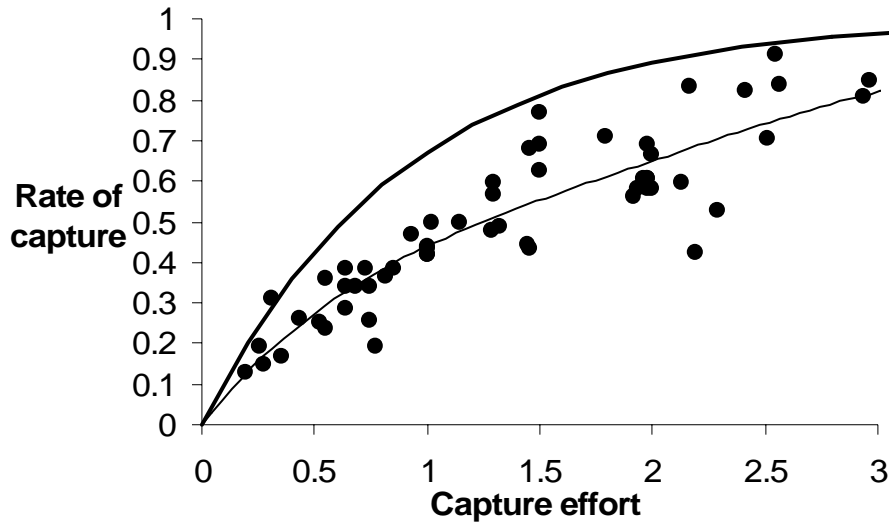
11

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

22

1 Figure 1

2



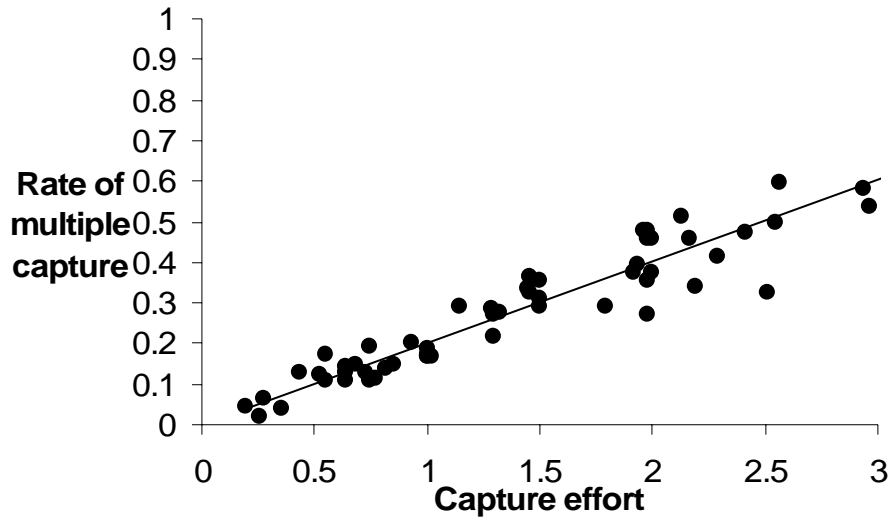
3

4

5

1 Figure 2

2



3

4

5

Figure 3

