Molecular Ecology Notes (2007) 7, 400–403 doi: 10.1111/j.1471-8286.2007.01697.x

TECHNICAL ARTICLE

A pooling approach to detect signatures of selective sweeps in genome scans using microsatellites

MEIKE THOMAS,* FRIEDRICH MÖLLER,† THOMAS WIEHE* and DIETHARD TAUTZ*
**Department of Genetics, University of Cologne, Zuelpicher Str. 47, 50674 Cologne, Germany, †Freie Universität Berlin and Berlin Center for Genome Based Bioinformatics (BCB), Arnimallee 22, 14195 Berlin, Germany*

Abstract

We have evaluated a pooling approach that can reduce the number of polymerase chain reactions in a screen for selective sweeps by more than an order of magnitude. We show that the complex peak pattern that results from pooling of all samples from a given population is a faithful reflection of the composite pattern of the individual alleles, although with an under-representation of the larger alleles. Candidate loci for selective sweeps can be identified by visual inspection of the pool patterns. We have also implemented a software tool, which can find suitable microsatellite loci in the vicinity of annotated genes.

Keywords: microsatellites, PCR, pooling, selective sweeps

Received 29 June 2006; revision accepted 11 December 2006

Identification of selective sweeps in natural populations has been proposed as an approach to identify genes involved in local adaptations (Schlötterer 2003). Positive selection can lead to the fixation of a favourable mutation in a population. This leads also to a loss of variability in the flanking region, due to hitch hiking (Maynard Smith & Haigh 1974). Because of the high density of microsatellite loci in eukaryotic genomes it should be possible to trace recent local selective sweeps by systematically scanning for population-specific loss of variability at individual loci (Schlötterer 2003). However, since higher eukaryotic genomes may contain about 40 000 selectable loci and the detection of polymorphic variants requires testing of multiple individuals (at least 20) for several populations, a complete genome scan would require millions of genotypes to be determined. We present here a pooling strategy that allows to reduce the number of genotyping reactions significantly and which has been shown to be possible in principle (Pacek *et al*. 1993). The basic idea is that DNA samples from all individuals of a given population can be pooled in equal amounts and only one polymerase chain reaction (PCR) is performed with primers flanking a microsatellite locus. This will result in a complex pattern of peaks in those cases where the locus is polymorphic, but in a relatively simple peak pattern in cases where polymorphism was lost, that is in the regions that are candidates for selective sweeps. Such loci can then be re-typed from individuals to confirm the sweep signature and to calculate exact allele frequencies to be used in appropriate statistics. In the following, we describe this approach for a genome screen based on approximately 1000 microsatellite loci in five different populations of the house mouse (*Mus musculus*).

Correspondence: Diethard Tautz, Fax: +49 221470 5975; E-mail: tautz@uni-koeln.de

Sample collection schemes and DNA extraction are described in Ihle *et al*. (2006). DNA of single individuals was normalized to 10 ng/μL and combined in population specific pools, each consisting of 40 samples. Primers were purchased in 96-well plate format (primer sequences see Table S1, Supplementary material) from Sigma Aldrich Co. Forward primers were labelled with FAM dye at the 5′ end. PCR was performed using a multiplex kit (QIAGEN Cat. No. 206143). All reactions were carried out in 10-μL volumes using 30 ng of pooled DNA template and following the protocol of the supplier of the kit. Amplification was started with a 15′ melting step at 95 °C, followed by 40 cycles of 30″ melting at 94 °C, standard annealing temperature for all primers at 60 °C for 1′30″, and elongation at 72 °C for 1′, followed by a final elongation step at 72 °C for 30′. PCR products were diluted 1:20 in water and 1 μL of this dilution was added to 0.1 μL GeneScan 500 LIZ Size Standard (Applied Biosystems, Part Number 4322682) in 10 μL HiDi formamide (Applied Biosystems, Part Number 4311320). Fragments were run on a 48-capillary 3730 DNA sequencer.

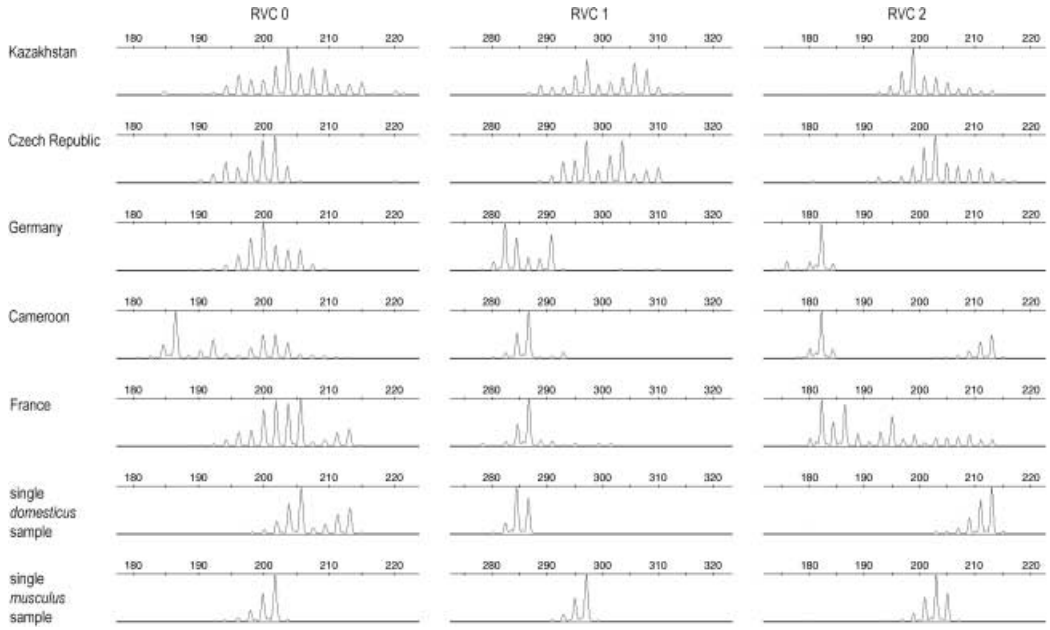

Fig. 1 Examples of three output files from the pooling approach each representing one of the three different reduced variability classes (RVC). The loci were amplified from all five mouse populations plus a single animal from each subspecies. The classification of the loci in RVC1 and 2 are based on a comparison of the German and the French population. In the example for RVC1 a weak reduction in variability is present in France, but also in Cameroon. The locus classified in RVC2, exhibits almost no variability in Germany compared to a variable French population. With respect to the Cameroon population, it would have been classified as RVC1.

The fragment patterns from the pools were displayed with the GENEMAPPER software version 3.5 (Applied Bioscience, Part Number 4346647) and then analysed by eye. All output files were inspected by pairwise comparison between populations. Candidate loci were defined as those showing a rather simple pattern of peaks in one population but a complex one in the others. Figure 1 shows an example from three loci typed in the five populations, two belonging to the subspecies *Mus musculus musculus* (sampled in Kazakhstan and in the Czech Republic) and three to *Mus musculus domesticus* (sampled in Germany, France and Cameroon) (Ihle *et al*. 2006). Because each locus shows characteristic slippage patterns, we have also always included a single individual from each subspecies for comparison.

To pick candidate loci for selective sweeps, we scored the trace patterns obtained from the sequencer for signs of population specific reduction in variability. We used a ranking scheme of 'Reduced Variability Classes' (RVC) 0 to 2, where RVC 0 contains all loci that show no obvious difference in variability between the populations (clearly no candidate loci), RVC 1 represents possible cases of reduction in variability (potential candidate loci) and RVC 2 represents clear cases (clear candidate loci) (Fig. 1). This classification was performed by eye, since pool patterns represent a composite of allele peaks plus stutter bands from 80 chromosomes, and additional 'noise peaks' due to unspecific amplification. Thus, the estimation of reduction in variability has to be based on the unique pattern of each locus. We have tried automatic scoring methods, such as counting peak numbers, or assessing the distribution of peak areas, but found that this cannot cope satisfactorily with the complexity of the pattern encountered. In contrast, the scoring by eye allows to adjust each case in direct comparison with the amplification patterns from the two individuals, which provides information on the locus specific slippage patterns and unspecific peaks. Independent scoring by two persons yielded very similar classifications.

Candidate loci from RVC2 were then amplified from individuals to compare the pool patterns with those determined from single typing. Figure 2 shows three such comparisons. It is evident that the allele frequencies determined from the individual typings reflect rather well the patterns seen in the pools. However, in several cases we found that larger alleles tend to be under-represented in the pools, when compared to the individual typings. Comparison of the patterns in Fig. 2(e, f) shows that the large alleles in the German population appear only as minor peaks in the pool, but with significant frequencies in the individual typings. This effect has therefore to be kept in mind when evaluating the pool patterns. However, among approximately 40 loci screened in this way, we found no case where an allele appeared with a significant frequency in the individual typings, which was not also at least present as a small peak in the pool pattern. Thus, the pool patterns are fairly reliable indicators of the allele spectrum found in a population.

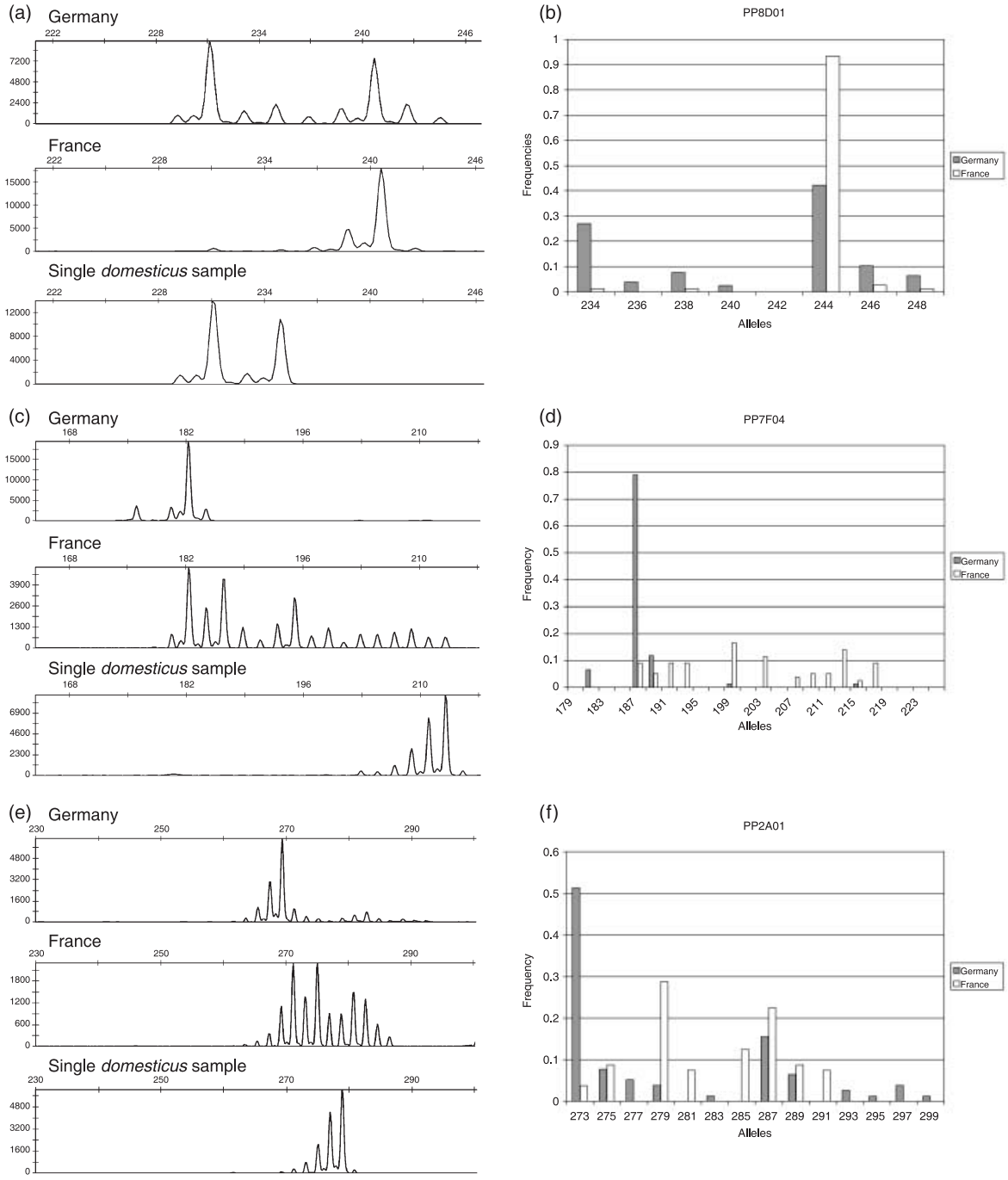

Fig. 2 Pattern of the pooled samples from three candidate sweep loci (a, c, e) and the corresponding allele frequencies estimated form genotyping single individuals for the respective loci (b, d, f).

The results from the individual typings can then be directly used for statistical analysis, for example Schlötterer's lnRV statistic (Schlötterer 2002), provided a reference set of randomly chosen loci is available. Using the reference set described in Ihle *et al*. (2006), we found that the first two loci in Fig. 2 show highly significant signatures of selective sweeps, while the third locus is not significant. On average, we found that about 20% of the loci classified as RVC2 and then typed individually remained as highly significant outliers after stringent statistical testing (Thomas *et al.* in preparation).

Given that the pooling approach allows an efficient screening of large numbers of loci, it is also important to have an automatic routine for selecting the microsatellite loci. Because sweep regions may be relatively small (Beisswanger

et al. 2006) it is advisable to use microsatellite loci from the vicinity of genes for a systematic screen. To find such loci, we have written a program that identifies all dinucleotide repeats in the vicinity of annotated genes and lists the results in a comprehensive table. The program reads a GenBank formatted DNA-sequence file and generates a hash table of all words of length four. There are 12 words which represent possible seeds for dinucleotide repeats (dinucleotides AA, CC, GG, TT are ignored) and which are analysed further. For each seed, the program then searches for the minimal left and maximal right boundaries at which the dinucleotide repeat terminates. It stores a 'hit' if the length of the repeat exceeds a user-defined threshold. Subsequently, the positions of the hits are compared to the position of the closest annotated coding sequence (CDS). Only hits within a predefined range of the closest CDS are returned and written into a table. There are two output tables. The first output table lists for each returned hit its position, its dinucleotide type, the index of the closest upstream and downstream CDSs, the repeat sequence itself and upstream and downstream flanking regions around the repeat. The second table contains all CDS annotations from the GenBank file and which can be referred to by the CDS index provided in the first table. This program was used to select the approximately 1000 loci used in this study. Almost all of them amplified successfully, only 5.4% gave no result, or were inclusive. The software is available under http://justus.genetik.uni-koeln.de:8200/software.

By applying the pooling strategy we were able to reduce the number of required PCRs from approximately 200 000 to 7000 for a genome screen in five populations, each represented by 40 individuals. In particular, the time consuming part in microsatellite genotyping, namely the allele calling from the raw data, was reduced to a minimum, because the pre-selection of candidate loci for signatures of positive selection is based on a simple visual inspection of pool patterns. With our approach, it seems therefore feasible to eventually screen all annotated genes of the mouse genome for signatures of selective sweeps and also to apply this approach to other cases where fully sequenced genomes are available.

Acknowledgements

We thank Sonja Ihle and Susanne Kipp for providing the DNA samples used in this study. The work was funded by the Volkswagen Stiftung.

References

Beisswanger S, Stephan W, De Lorenzo D (2006) Evidence for a selective sweep in the wapl region of *Drosophila melanogaster*. *Genetics*, **172**, 265–274.

Ihle S, Ravaoarimanana I, Thomas M, Tautz D (2006) Tracing signatures of selective sweeps in natural populations of the house mouse. *Molecular Biology and Evolution*, **23** (4), 790–797.

Maynard SJ, Haigh J (1974) The hitch-hiking effect of a favourable gene. *Genetical Research*, **23**, 23–35.

Pacek P, Sajantila A, Syvänen A-C (1993) Determination of allele frequencies at loci with length polymorphism by quantitative analysis of DNA amplified from pooled samples. *PCR Methods and Applications*, **2**, 313–317.

Schlötterer C (2002) A microsatellite-based multilocus screen for the identification of local selective sweeps. *Genetics*, **160**, 753–763.

Schlötterer C (2003) Hitch hiking mapping: functional genomics from the population genetics perspective. *Trends in Genetics*, **19**, 32–38.

Supplementary material

The following supplementary material is available for this article:

Table S1 List of 960 loci (with their primer sequences) to which we applied our pooling technique when genotyping them in a large genome screen for selective sweeps. These microsatellite loci were identified applying the program described above.

This material is available as part of the online article from:
http://www.blackwell-synergy.com/doi/abs/10.1111/j.1471-8286.2007.01697.x
(This link will take you to the article abstract).

Please note: Blackwell Publishing are not responsible for the content or functionality of any supplementary materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.