The advantages of next generation sequencing against the enrichment technique in isolating biomarkers from the termite genome.

by

G. Veera Singham, Ahmad Sofiman Othman, Edward Vargo, Warren Booth and Chow-Yang Lee

Urban Entomology Laboratory, Vector Control Research Unit, School of Biological Sciences, Universiti Sains Malaysia, 11800 Penang, Malaysia

Abstract

The relative advantage of two different molecular techniques in isolating microsatellite markers from the genome of a subterranean termite, *Macrotermes gilvus* (Hagen) was investigated. The cost, efficacy, and success rate of the traditional enrichment protocol was compared against the high-throughput next generation sequencing method. Here we show that the next generation sequencing technology was found to be a rapid, cost-efficient method for isolating novel microsatellite markers from the termite genome. Hundreds of sequences with suitable flanking sites for primer design were detected and novel polymorphic microsatellites were successfully characterized for downstream applications in population genetics study of *M. gilvus*.

Keywords: Microsatellites, 454 pyrosequencing, next generation sequencing, termite, *Macrotermes* sp.

Introduction

Microsatellite DNA, widely distributed in the genome of prokaryotes and eukaryotes (Field and Willis 1998, Schlotterer 2000), has been repeatedly used in studies of genetic diversity, population genetics, genome mapping, and phylogeography (Sunnucks 2000, Zhang and Hewitt 2003). High levels of polymorphism, reproducibility and co-dominance make microsatellites one of the best choices of molecular marker for fine scale studies of the population genetic structure of the Asian subterranean termite, *Macrotermes gilvus* (Hagen). Although the application of microsatellites is simple and robust, their identification and development display substantial challenges. Typically, there are two ways of acquiring microsatellites: constructing a genomic library and developing microsatellites *de novo* or to search published literature for existing microsatellite primers for the target species. Traditionally, microsatellite loci have been isolated from partial genomic libraries by screening several thousand

clones via colony hybridization with repeat-motif containing probes (the enrichment technique). Whereas being successful for taxa with an abundance of microsatellites, this approach remains inefficient for species with low microsatellite frequencies such as invertebrates including termites. Next generation sequencing (NGS) is revolutionizing biomarker discovery through simplification of library preparation, template preparation, and sequencing chemistries (Margulies et al. 2005, Malausa et al. 2011). It has been increasingly popular in supplanting the traditional enrichment procedures as the preferred method for developing microsatellites (e.g. Abdelkrim et al. 2009, Castoe et al. 2010). As such, this paper is dedicated in highlighting the various advantages of next generation sequencing in isolating microsatellite markers from *M. gilvus* in comparison to the enrichment technique.

Materials and Methods

Sample preparation and DNA extraction. Worker termites were collected from 22 colonies in suburban habitats within the state of Sarawak, East Malaysia and stored in absolute ethanol. Total genomic DNA was extracted after crushing the whole termite worker in liquid nitrogen by following the CTB Tissue Extraction Kit protocol (Intron, Seongnam-Si, Gyeonggi-do, Korea).

Enrichment technique. DNA from a pooled sample of five workers was digested with *Rsal* and followed by ligation with M1uI annealed adaptor. The ligation was enriched with microsatellite repeats by hybridizing on Hybond N+ membrane with bound oligonucleotides comprising of di-, tri- and tetra-nucleotide motifs (GATA, AAAT, GATG, GACA, AAG, CAA, AAT, CA, GT, and CT). Fragments with SSR repeats (enriched fragments) were then incorporated into a cloning vector pGEM-T Easy vector (Promega, Madison, Wisconsin, USA) before being transformed into JM109 *E. coli* competent cells (Promega) to obtain a clone library. Blue-white screening was performed to select positive clones (i.e. colonies with vector- incorporated insert). 95 positive clones were selected and sequenced.

Next generation sequencing. ~ 2 μg of genomic DNA extracted from a single wholetermite and used in a 1/16th of a PicoTiter plate for sequencing on a Roche 454 GS FLX sequencer with titanium chemistry at the Genome Sequencing and Analysis Core Facility, Duke University (Durham, NC). Sample preparation and analytical processing were performed according to manufacturer's protocol for the Titanium series as detailed in Margulies et al. (2005).

Microsatellite detection and primer design. Reads containing microsatellite repeats were identified using the software Msatcommander v. 0.8.2 (Faircloth 2008). The search criteria were set to a minimum of six repeats of di- to hexa-nucleotides with perfect repeat motifs. Primer3 v. 0.4.0 was subsequently

used to design a subset of primer pairs that were optimized in a standard three step PCR with an initial denaturation step at 94 °C for 2min, followed by 35 cycles of three step PCR ; the denaturation step at 94 °C for 30 sec, the annealing step was optimized on a gradient of temperatures based on the melting point of the designed primers for 30 sec and the extension step at 72 °C for 1 min followed by a post cycle extension at 72 °C for 10 min. PCR products were run on 6 % polyacrylamide gel electrophoresis and genotyping was performed using Alpha Imager software.

Results and Discussion

Sequence analysis from the 95 positive clones resulted in 16 sequences comprising relatively long SSR motifs. 12 sets of locus specific primers were successfully designed out of the 16 sequences after elimination of overlapping sequence. Five of the 12 developed primer sets successfully PCR amplified with a single distinct band observed on 2% agarose gel electrophoresis. The rest of the primer sets were either non-amplifying or amplified with multiple unspecific bands. The success rate of finding SSR motifs in the sequence data was 16.8 % for the enrichment technique. Subsequent characterization of five loci resulted in characterization as monomorphic when tested against termites from different colonies except one locus with revealed a low level of polymorphism.

A modest sequencing volume of the next generation sequencing on the 454 platform generated 34,122 reads with 1,212 reads containing microsatellites with di-, tri, tetra-, penta- and hexa- nucleotides perfect repeat motifs. Subsequently, LI-COR 4300 DNA analyzer and fragment analysis were used to test for polymorphism among the 37 randomly selected primer pairs. Fifteen loci were found to be polymorphic averaging between 2 - 4 alleles per locus. Observed and expected heterozygosities ranged between 0.09 – 0.73 and 0.09 – 0.54, respectively. Four multiplex groups were designed for future population genetic studies. Successful development of these novel molecular markers will provide new a perspective for understanding the population dynamics and also to elucidate the phylogeographic structure of fungus-growing termites in this region. Table 1 summarizes the relative advantage of next generation sequencing over enrichment method in isolating microsatellites from *M. gilvus*.

Table 1: Development of microsatellite markers using two different isolation methods.

Criteria	Enrichment technique	Next generation sequencing
Library preparation and isolation time	~ 2 – 4 weeks	~ 1 – 2 weeks (supplier dependent)
Costs (in USD)	Up to $ 2000 (for a single plate screening including library set up)	~ $ 1500 (1/ 16th of picotiter plate)
Rate of isolating repeat containing sequences	16.8 % (16 out of 95)	3.6 % (1,212 out of 34,122)
Sequence with successful primer design	12 primer sets	207 primer sets (a subset of 37 primers selected for screening)
Positive PCR amplification	5 primer sets	34 primer sets
Polymorphic loci	1	15
Estimated cost for enrichment technique to match results of NGS	15 plate screening (approx. $ 15,000)	-

Based on the results, it is clear that NGS application surpasses the traditional enrichment technique in the development of polymorphic microsatellite loci from *M. gilvus*. We chose the 454 sequencing platform as it provided output reads with longer read lengths (~400bp) compared to other NGS platforms such as Illumina (~100bp). Longer reads are important so that the repeat motifs have sufficient flanking sites for primer design. A major drawback of the enrichment technique is that the cost for the Sanger sequencing of the selected clones is relatively expensive and thousands of clones are required for identifying functional microsatellite repeats, particularly for species with a low abundance of microsatellite repeats in their genome, like *M. gilvus*. Clearly, the success rate of isolating functional microsatellites relies heavily on a coverage-dependent manner. At similar costs, NGS outputs ~13Mbp of sequence coverage compared to the enrichment method with ~60 Kbp provides about 200 fold deeper coverage. The NGS method is thus more successful case as it outputs greater coverage of the genome with limited input at a minimum cost. Moreover, the enrichment method is highly laborious and time consuming in library preparation and includes cloning which can be bypassed in the NGS method. Consequently, while the enrichment method fell short of finding the number of functional microsatellites needed for downstream population genetics applications, 15 novel polymorphic microsatellite markers were successfully characterized using the NGS approach.

Conclusion

In conclusion, the NGS technology offers advantages for microsatellite isolation over the traditional enrichment method from the termite genome as: 1) there is no need to decide *a priori* which motif types to enrich, 2) problems in attaining DNA fragments of manageable size (using restriction enzymes) can be overcome, 3) over-representation of transposable elements are avoided, 4) there is no need to edit sequence chromatograms and finally 5) development of microsatellites via NGS is quicker, cost effective, and resulted in high-throughput data.

References

Abdelkrim, J., B. C. Robertson, J. A. L. Stanton, and N. J. Gemmel. 2009. Fast, cost-effective development of species-specific microsatellite markers by genomic sequencing. *BioTechniques* **46**, 185-192.

Castoe, T.A., A. W. Poole, W. J. Wu, et al. 2010. Rapid identification of thousands of copperhead snake (*Agkistrodon contortrix*) microsatellite loci from modest amounts of 454 shotgun genome sequence. *Mol. Ecol. Resour.* **10**, 341-347.

Faircloth, B. C. 2008. Msatcommander: detection of microsatellite repeat arrays and automated, locus-specific primer design. *Mol. Ecol. Resour.* **8**, 92-94.

Field, D., and C. Wills. 1998. Abundant microsatellite polymorphism in *Saccharomyces cerevisiae*, and the different distributions of microsatellites in eight prokaryotes and *S. cerevisiae*, result from strong mutation pressures and a variety of selective forces. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 1647-1652.

Malausa, T., A. Gilles, E. Meglecz, H. Blanquart, S. Duthoy, et al. 2011. High-throughput microsatellite isolation through 454 GS-FLX Titanium pyrosequencing of enriched DNA libraries. *Mol. Ecol. Resour.* **11**, 638-644.

Marguilies, M., M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376-380.

Schlotterer, C. 2000. Evolutionary dynamics of microsatellite DNA. *Chromosoma* **109** 365-371.

Sunnucks, P. 2000. Efficient genetic markers for population biology. *Trends Ecol. Evol.* **15**, 199-203.

Zhang, D. X., and G. M. Hewitt. 2003. Nuclear DNA analyses in genetic studies of population: practice, problems, and prospects. *Mol. Ecol.* **12**, 563-584.