Rapid Microsatellite Isolation from a Butterfly by *De Novo* Transcriptome Sequencing: Performance and a Comparison with AFLP-Derived Distances

Alexander S. Mikheyev[1]*, **Tanya Vo**[1], **Brian Wee**[2], **Michael C. Singer**[3], **Camille Parmesan**[3]

1 Okinawa Institute of Science and Technology, Onna, Japan, 2 National Ecological Observatory Network, Incorporated, Washington, District of Columbia, United States of America, 3 University of Texas, Section of Integrative Biology, Austin, Texas, United States of America

Abstract

Background: The isolation of microsatellite markers remains laborious and expensive. For some taxa, such as Lepidoptera, development of microsatellite markers has been particularly difficult, as many markers appear to be located in repetitive DNA and have nearly identical flanking regions. We attempted to circumvent this problem by bioinformatic mining of microsatellite sequences from a *de novo*-sequenced transcriptome of a butterfly (*Euphydryas editha*).

Principal Findings: By searching the assembled sequence data for perfect microsatellite repeats we found 10 polymorphic loci. Although, like many expressed sequence tag-derived microsatellites, our markers show strong deviations from Hardy-Weinberg equilibrium in many populations, and, in some cases, a high incidence of null alleles, we show that they nonetheless provide measures of population differentiation consistent with those obtained by amplified fragment length polymorphism analysis. Estimates of pairwise population differentiation between 23 populations were concordant between microsatellite-derived data and AFLP analysis of the same samples (r = 0.71, p<0.00001, 425 individuals from 23 populations).

Significance: De novo transcriptional sequencing appears to be a rapid and cost-effective tool for developing microsatellite markers for difficult genomes.

Introduction

Many types of genetic analysis take advantage of microsatellite markers, which are highly polymorphic loci of simple sequence repeats located through the genome. For example, microsatellite analysis is useful in studies of paternity, population structure and history, as well as to make conservation decisions for the management of endangered species [1,2].

Given the broad-scale utility of these markers, a large number of approaches have been developed for their isolation from genomic DNA [3]. These approaches typically involve some form of microsatellite enrichment, followed by time consuming and costly brute force sequencing. Aside for the labor and cost associated with traditional approaches, the microsatellite enrichment step sometimes fails. For example, for reasons not fully understood, isolation of microsatellites from Lepidopteran genomes is extremely difficult [4–6]. This problem is not confined to Lepidoptera, affecting bivalve mollusks [7], mosquitoes [8], mites [9], ticks [8], nematodes [10,11] and birds [12,13].

The increase in publicly available EST data for many species has made bioinformatic isolation of microsatellite markers increasingly commonplace (*e.g.*, [14–17]). However, microsatellites isolated from EST libraries differ from those typically found in regions of the genome unassociated with genes. Gene-associated microsatellites are physically linked to particular alleles of a gene, and may hitchhike if the gene is under selection. Microsatellite variation in untranslated regions of transcribed DNA may affect the rates of gene expression or translation, and thus may be under selection. Indeed, EST-derived microsatellites almost universally show strong deviations from Hardy-Weinberg equilibrium. However, the relatively few studies that compare the performance of EST-derived microsatellites with that of other genotyping techniques have generally found comparable results [18–22]. Here we used the Roche 454 Titanium platform for transcriptional sequencing of Edith's checkerspot butterfly (*Euphydryas editha*), in order to rapidly isolate polymorphic microsatellite loci for a conservation genetics study. We then compared the estimates of population differentiation and biogeographic structure obtained by this approach with those from AFLP genotyping of the same set of populations [23].

Materials and Methods

Microsatellite identification

RNA was extracted from a larva, a pupa and an adult *E. editha*. RNA extraction, normalized library preparation, sequencing and

assembly using the Roche Newbler assembler was performed by the University of Illinois W.M. Keck Center for Comparative and Functional Genomics using protocols and reagents supplied by Roche. The assembled data were then queried for the presence of microsatellites using a simple python script using all possible sequences combinations of di-, tri- and tetra-nucleotide repeats, with at least eight perfect repeats. Primers for microsatellite-containing sequences were designed using Primer3 [24] and tested for amplification and polymorphism.

Microsatellite amplification and polymorphism testing

Microsatellite loci were tested for amplification and polymorphism in 10 µl PCR mixes containing 1 ng genomic DNA, 10 mg BSA, 10 pmol primers, 6.7 nmol of ChromaTide® Rhodamine Green™-5-dUTP (Molecular Probes, presently discontinued) and 5 µl AmpliTaq Gold® PCR Master Mix (Applied Biosystems). The temperature cycling conditions were as follows: 7 min at 94°C, then 35 cycles of 10 sec at 94°C, 1.5 minutes at 60°C and 2 minutes at 68°C. The reaction was terminated with a final incubation of 30 minutes at 72°C. 1 µl of each reaction was then analyzed using an ABI3100 DNA sequencer. For genotyping each well had 0.1 µl LIZ labeled GeneScan 500 size standard (Applied Biosystems) and enough deionized formamide for a total volume of 10 µl. Alleles were scored using GeneMarker.

Quality control

Deviations from Hardy-Weinberg equilibrium were assessed using GenAlEx [25]. Many individuals in the present study were previously genotyped by Wee [26] using AFLP markers. Thus, we were able to assess concordance between results of the two studies by comparing Fst matrixes generated by the two techniques. We computed Fst distances for 23 populations (425 individuals) used in Arlequin (v.3) [27], and compared them to the Fst matrix from Wee [26] using a Mantel test with 10,000 bootstrap replicates. We also screened an additional 406 individuals from 48 more populations for polymorphism analysis (Table S1).

Results

After quality filtering, the 454 run generated 864,056 reads, totaling 245,064,986 bases, which were assembled into 14,244 contigs with a threshold of 200 bp overlap and 95% identity. 49,937 singleton reads remained unassembled and were not included in the subsequent analysis, although if needed, they may be used for microsatellite mining. The assembled contigs contained 92 microsatellite loci, 72 of which were selected for

microsatellite development. Of these, 36 loci amplified successfully and appeared polymorphic (see Table S2). Following the initial screening performed of eight individuals, we developed four multiplex PCR cocktails containing a total 10 polymorphic loci for large-scale genotyping (Table S1). Sequences for the other loci are available from the authors upon request. The reaction conditions were as above, but without fluorescently labeled dUTPs in reactions 1 and 2, and with primer concentrations as noted in Table S2. The 10 loci are deposited in GenBank under accession numbers GU997598-GU997607.

The markers show significant deviations from Hardy-Weinberg equilibrium in the many of the populations (Figure 1). The difference between observed and expected heterozygosities was positively correlated with the number of failed amplification for each locus, suggesting that null alleles may in part be responsible for driving this difference ($r_s = 0.81$, n = 10, p = 0.0042). However, estimates of pairwise population differentiation were concordant between microsatellite-derived data and an earlier AFLP analysis of the same samples by Wee [26] (r = 0.71, n = 23, p<0.00001).

Raw microsatellite data generated in this study have been deposited in the Dryad database (www.datadryad.org) under accession number 1540.

Discussion

Microsatellite isolation from lepidopteran genomes has been difficult, possibly because microsatellite loci appear to be rare, and may have very similar flanking regions [6], which makes the design of primers problematic. We hypothesized that microsatellite loci isolated from non-translated transcripts may be less likely to exist as duplicate copies, and thus be more amenable to marker development. This has made microsatellite isolation relatively straightforward in our case. Given the decrease in next-generation sequencing costs, transcriptional re-sequencing will be a faster and cheaper way to isolate microsatellites, compared with traditional enrichment techniques. We were able to complete microsatellite development and screening in about three months of part time work by a single technician. Our actual cost of library construction and sequencing, was about US$15,000, is comparable to that charged by private companies for microsatellite enrichment [3]. Since then, the actual cost of library construction and next generation sequencing has dropped by at least 50%, and is decreasing further.

In this and several other studies, microsatellites derived from transcribed sequence data significantly depart from Hardy-Weinberg equilibrium (Figure 1) [14–17]. This could be due to selection on polymorphisms in untranslated gene regions where

Figure 1. Hardy-Weinberg equilibrium statistics. Significant deviations from Hardy-Weinberg (chi-squared test, p<0.05) are indicated in dark grey. Loci monomorphic in that population are shown in light grey. Every population is represented by a column, with each row corresponding to a microsatellite locus. The order of the populations is the same as in Table S1 (alphabetical).
doi:10.1371/journal.pone.0011212.g001

Table 1. Primers used for large-scale genotyping.

PCR #	Locus	Primers sequence	Primer amount (pmol)	Label	Repeat Motif	Range (bp)	Allele Count	H_o	H_e	Percent missing
1	euphy 2	tgatgataacgagcgggaag	0.5	5'TAM	CAG	144–191	20	0.42	0.72	0.60%
		cggtaccgctacgtgactact								
	euphy 3	gctgtaatttggtaaggggttg	0.5	5' HEX	ATC	121–171	18	0.52	0.83	0.84%
		tacgttcagtgatggacatgc								
	euphy 21	acgcaaggtgctccacttat	0.5	5' HEX	CAA	220–239	9	0.18	0.24	1.32%
		ttgctacgctaacagcatcg								
	euphy 69	ctcctccgcaccaacaagta	1	5' FAM	GTT	72–103	13	0.17	0.39	3.59%
		aaacgtctacgttagaaggtatgt								
2	euphy 14	tgactgaacacacggacgat	0.5	5' TAM	TACA	99–170	32	0.15	0.68	14.0%
		tccatcatgctttaagtgagga								
	euphy 61	aaagcgtgcttacattacatgg	0.5	5' TAM	AC	186–246	42	0.44	0.87	12.9%
		tcccgtttaacataatctgtgg								
3	euphy 35	atagaaataaacatgcggccata	10	dUTP	TG	267–335	56	0.33	0.96	13.1%
		cagatgtacaagaggctgcctta								
	euphy 50	atgcgatttcatgccacata	10	dUTP	CA, A	135–176	28	0.22	0.85	22.5%
		ccatcctgacatgtgaaacg								
4	euphy 37	tgcaagacttgaaatatggttatca	10	dUTP	C, CA	130–182	21	0.41	0.80	2.28%
		gtccattggaaggatcagga								
	euphy 47	cacgtgagcattccagtttg	10	dUTP	AT	172–335	34	0.44	0.87	5.99%
		tcggcgtaacggtttaaatg								

Summary statistics are based on a survey of 835 individuals from 72 populations (Table 1). Even and odd numbered reactions were pooled and analyzed together in the same sequencer run. The percentages of missing were significantly different among the PCR mixes, being significantly higher in reactions 2 and 3 ($F_{3,6} = 15.4$, $p = 0.0038$).
doi:10.1371/journal.pone.0011212.t001

these microsatellites typically reside, or to non-neutral dynamics of the genes to which they are physically linked. In our study, percent reaction failure explained most of the variance in the differences between observed and expected heterozygosities (Table 1). Therefore, at least in our case, Hardy-Weinberg disequilibrium may be partially due to insufficient optimization of PCR conditions and allele dropout. Whether or not higher levels of null alleles are common in EST-derived microsatellites is not clear, since these data are not routinely reported with such studies. We strongly recommend further optimization of the reaction conditions for the loci presented here, especially since the manufacture of fluorescent dUTPs used in this study has been discontinued.

In principle, deviations from Hardy-Weinberg can create substantial biases [28], limiting the utility of such markers. The extent to which these issues may affect analysis with EST-derived microsatellites is presently unclear, but should be carefully investigated by future studies. Ideally, studies isolating microsatellites from ESTs should verify their performance by comparing results with another genotyping method, as we have done with AFLPs. Likewise, it would be useful to present an analysis of null allele presence.

Supporting Information

Table S1 Sample sizes and locations of the populations used for polymorphism screening.
Found at: doi:10.1371/journal.pone.0011212.s001 (0.10 MB DOC)

Table S2 This file lists all the primers tested in the study, and the results of polymorphism testing based on a small sample of 8 individuals. Loci used for further analysis are highlighted in gray.
Found at: doi:10.1371/journal.pone.0011212.s002 (0.12 MB DOC)

Acknowledgments

Author Contributions

Conceived and designed the experiments: ASM MCS CP. Performed the experiments: ASM TV. Analyzed the data: ASM. Contributed reagents/materials/analysis tools: BW MCS CP. Wrote the paper: ASM.

References

1. Wright J, Bentzen P (1994) Microsatellites: genetic markers for the future. Reviews in Fish Biology and Fisheries 4: 384–388.
2. Varshney R, Graner A, Sorrells M (2005) Genic microsatellite markers in plants: features and applications. TRENDS in Biotechnology 23: 48–55.
3. Zane L, Bargelloni L, Patarnello T (2002) Strategies for microsatellite isolation: a review. Molecular Ecology 11: 1–16.
4. Nève G, Meglécz E (2000) Microsatellite frequencies in different taxa. Trends in Ecology & Evolution 15: 376–377.
5. Meglécz E, Petenian F, Danchin E, D'Acier A, Rasplus J, et al. (2004) High similarity between flanking regions of different microsatellites detected within each of two species of Lepidoptera: Parnassius apollo and Euphydryas aurinia. Molecular Ecology 13: 1693–1700.

6. Zhang D (2004) Lepidopteran microsatellite DNA: redundant but promising. Trends in Ecology & Evolution 19: 507–509.

7. Cruz F, Pérez M, Presa P (2005) Distribution and abundance of microsatellites in the genome of bivalves. Gene 346: 241–247.

8. Fagerberg AJ, Fulton RE, Black WC, IV (2001) Microsatellite loci are not abundant in all arthropod genomes: analyses in the hard tick, *Ixodes scapularis* and the yellow fever mosquito, *Aedes aegypti*. Insect Molecular Biology 10: 225–236.

9. Navajas MJ, Thistlewood HM, Lagnel J, Hughes C (1998) Microsatellite sequences are under-represented in two mite genomes. Insect Molecular Biology 7: 249–256.

10. Grillo V, Jackson F, Gilleard J (2006) Characterisation of *Teladorsagia circumcincta* microsatellites and their development as population genetic markers. Molecular and Biochemical Parasitology 148: 181–189.

11. Johnson P, Webster L, Adam A, Buckland R, Dawson D, et al. (2006) Abundant variation in microsatellites of the parasitic nematode *Trichostrongylus tenuis* and linkage to a tandem repeat. Molecular and Biochemical Parasitology 148: 210–218.

12. Primmer C, Raudsepp T, Chowdhary B, Møller A, Ellegren H (1997) Low frequency of microsatellites in the avian genome. Genome Research 7: 471.

13. Neff B, Gross M (2001) Microsatellite evolution in vertebrates: inference from AC dinucleotide repeats. Evolution 9: 1717–1733.

14. Qiu X, Liu S, Wang X, Meng X (2009) Eight SSR loci from oyster *Crassostrea gigas* EST database and cross-species amplification in *C. plicatula*. Conservation Genetics 10: 1013–1015.

15. Sharma R, Bhardwaj P, Negi R, Mohapatra T, Ahuja P (2009) Identification, characterization and utilization of unigene derived microsatellite markers in tea (*Camellia sinensis*, L). BMC Plant Biology 9: 53.

16. Wang S, Zhang L, Matz M (2009) Microsatellite characterization and marker development from public EST and WGS databases in the reef-building coral *Acropora millepora* (Cnidaria, Anthozoa, Scleractinia). Journal of Heredity 100: 329–337.

17. Yang J, Yang J, Li H, Zhao Y, Yang S (2009) Isolation and characterization of 15 microsatellite markers from wild tea plant (*Camellia taliensis*) using FIASCO method. Conservation Genetics 10: 1621–1623.

18. Coulibaly I, Gharbi K, Danzmann RG, Yao J, Rexroad CE (2005) Characterization and comparison of microsatellites derived from repeat-enriched libraries and expressed sequence tags. Animal Genetics 36: 309–315.

19. Garoia F, Guarniero I, Grifoni D, Marzola S, Tinti F (2007) Comparative analysis of AFLPs and SSRs efficiency in resolving population genetic structure of Mediterranean *Solea vulgaris*. Molecular Ecology 16: 1377–1388.

20. Scariot V, De Keyser E, Handa T, Deriek J (2007) Comparative study of the discriminating capacity and effectiveness of AFLP, STMS and EST markers in assessing genetic relationships among evergreen azaleas. Plant Breeding 126: 207–212.

21. Varshney R, Chabane K, Hendre P, Aggarwal R, Graner A (2007) Comparative assessment of EST-SSR, EST-SNP and AFLP markers for evaluation of genetic diversity and conservation of genetic resources using wild, cultivated and elite barleys. Plant Science 173: 638–649.

22. Woodhead M, Russell J, Squirrell J, Hollingsworth P, Mackenzie K, et al. (2005) Comparative analysis of population genetic structure in *Athyrium distentifolium* (Pteridophyta) using AFLPs and SSRs from anonymous and transcribed gene regions. Molecular Ecology 14: 1681–1695.

23. Vos P, Hogers R, Bleeker M, Reijans M, Van de Lee T, et al. (1995) AFLP: a new technique for DNA fingerprinting. Nucleic Acids Research 23: 4407–4414.

24. Rozen S, Skaletsky H (1999) Primer3 on the WWW for General Users and for Biologist Programmers. Bioinformatics Methods and Protocols. pp 365–386.

25. Peakall R, Smouse P (2006) GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. Molecular Ecology Notes 6: 288–295.

26. Wee P-P (2004) Effects of geographic distance, landscape features and host association on genetic differentiation of checkerspot butterflies. Ph.D. Dissertation, Austin: University of Texas.

27. Excoffier L, Laval G, Schneider S (2005) Arlequin (version 3.0): an integrated software package for population genetics data analysis. Evolutionary Bioinformatics Online 1: 47.

28. Chapuis M, Estoup A (2007) Microsatellite null alleles and estimation of population differentiation. Molecular Biology and Evolution 24: 621–631.