Population genetic inferences using immune gene SNPs mirror patterns inferred by microsatellites

JEAN P. ELBERS[1,3], RACHEL W. CLOSTIO[2], SABRINA S. TAYLOR[1]

[1]School of Renewable Natural Resources, 227 RNR Bldg., Louisiana State University and AgCenter,
Baton Rouge, Louisiana, 70803, USA

[2]Department of Biology, 300 E. St. Mary Blvd., University of Louisiana at Lafayette,
Lafayette, Louisiana, 70503, USA

Keywords: microsatellites, target enrichment, sequence capture,
next-generation sequencing, immunogenetics, population genomics

[3]Corresponding author: Fax: 225-578-4227, Email: jean.elbers@gmail.com

Running title: Immune gene SNPs mirror microsatellites

Abstract

Advances in genomic resources and next-generation sequencing have allowed researchers to access large panels of genetic markers such as single nucleotide polymorphisms (SNPs). These markers are replacing microsatellites for population genetic analyses, but it is not apparent how many SNPs are needed or how well SNPs correlate with microsatellite-derived estimates of genetic diversity, differentiation, or admixture. We used data from the gopher tortoise, *Gopherus polyphemus*, to compare the utility of SNPs and microsatellites to estimate population genetic parameters. Specifically, we compared inferences from 101 tortoises from 4 populations previously genotyped at 10 microsatellite loci with 18,000 immune gene SNPs from 16 randomly sampled tortoises (4 per population). We found SNPs generally mirrored patterns inferred by microsatellites. Observed and expected heterozygosities, F_{ST} values, and population admixture estimates were correlated between SNPs and microsatellites; however, allelic richness was not. We also found that the number of randomly chosen SNPs required to correlate with microsatellite-derived parameters varied depending on the question asked. In particular, 1,600, 800, or 100 randomly chosen SNPs were needed to correlate with microsatellite-estimates of observed heterozygosity, expected heterozygosity, or F_{ST} values, respectively. Our study illustrates that estimates of population genetic parameters obtained with SNPs generally mirror those obtained with microsatellites. Moreover, the number of SNPs typically obtained from next-generation sequencing far exceeds the number of SNPs needed to obtain parameter estimates similar to those obtained with microsatellites. These findings suggest that results from recent studies using a large panel SNPs will be largely comparable to older studies using microsatellites.

Introduction

Molecular markers vary in their utility and application to population genetic studies, and geneticists use available markers suited to answering questions at hand. Initially, geneticists only had allozymes and used them to infer nucleotide changes underlying differences in protein migration during electrophoresis. Later, variable mitochondrial DNA markers were used because of the availability of conserved primers and the high copy number of mitochondria, but mitochondrial markers mostly provided information on broad-scale genetic patterns (Moritz, 1994). Presently, markers such as microsatellites are commonly used in population genetics because they are neutrally evolving, are spread across genomes, and can elucidate fine-scale spatial genetic patterns (e.g., Clostio *et al.*, 2012).

Genomic resources, hybridization arrays, fluorescent probes, and next-generation sequencing (NGS) have allowed researchers to access other types of genomic markers, and recently large arrays of single nucleotide polymorphisms (SNPs) have become particularly popular in population genetic studies of not only model but also non-model organisms (Allendorf *et al.*, 2010). SNPs are one of the most numerous molecular markers (Gupta *et al.*, 2001), and thousands to millions of them can be examined simultaneously using NGS techniques compared to dozens observed in traditional Sanger sequencing-based approaches. However, it is not apparent how population genetic inferences vary between thousands of NGS derived SNPs and traditional microsatellites markers. Prior research has shown that genetic differentiation and diversity are correlated between 7 di-allelic markers (SNPs and indels (insertions/deletions)) and 14 microsatellites in 21 *Salmo salar* (Atlantic salmon) populations (Ryynanen *et al.*, 2007). Other studies have examined how microsatellite-derived population genetic parameters relate to fluorescent probe-assayed (e.g., Narum *et al.*, 2008), Sanger sequenced (e.g., Coates *et al.*, 2009), or array-assayed (e.g., Glover *et al.*, 2010) SNP parameters, but little research has compared genetic inferences derived from thousands of NGS generated SNPs to inferences from microsatellites. As more and more studies utilize NGS data, a better understanding of this relationship is imperative because many current management and recovery plans currently in effect are based on genetic data from microsatellites.

We recently applied genomic approaches to the threatened *Gopherus polyphemus* (gopher tortoise) to isolate genes involved in immune responses and better understand susceptibility to a chronic and occasionally fatal upper respiratory tract disease (Elbers & Taylor, 2015) caused by pathogens such as the bacteria *Mycoplasma agassizii* (Brown *et al.*, 1999). In addition to being subject to epidemiology studies, *G. polyphemus* populations have been examined genetically to infer population genetic diversity and differentiation and inform management decisions (e.g., Schwartz & Karl, 2005; Ennen *et al.*, 2010; Richter *et al.*, 2011; Clostio *et al.*, 2012).

We use the NGS data leveraged in Elbers & Taylor (2015) and the microsatellite data obtained by Clostio *et al.* (2012) to compare estimates of population genetic diversity, differentiation, and admixture derived from immune gene SNPs and microsatellites using samples from the same populations to better understand how NGS SNP inferences relate to those from microsatellites. We also subsample our SNPs to determine how many are needed to replace a given number of microsatellites for estimating genetic diversity and differentiation. We predict SNP inferences will mostly correlate with microsatellite inferences and that not all of the discovered SNPs will be needed to replace microsatellites for estimating diversity and differentiation.

Methods

Samples

For microsatellite analyses, we used 101 *G. polyphemus* from 4 populations along an east to west gradient (Fig. 1, Table 1) that were previously genotyped at 10 microsatellite loci (Clostio *et al.*, 2012). For SNP analyses, we used a subset of 16 randomly chosen tortoises (4 per population) from the full subset of 101 tortoises.

Target region for sequencing SNPs

We created a target region to capture the immunome (i.e., genes involved in immune response, *sensu amplo* Ortutay & Vihinen (2006)) of *Chrysemys picta bellii* (western painted turtle) using the GO2TR workflow (Elbers & Taylor, 2015). The workflow filtered the *C. p. bellii* 3.0.1 genome assembly (Shaffer *et al.*, 2013) annotated by the NCBI Eukaryotic Genome Annotation Pipeline (annotation release 100) using the gene ontology term "immune response" (i.e., genes that function in the immune system's response to internal or invasive threats). Jean-Marie Rouillard of MYcroarray Inc. (Ann Arbor, MI, USA) generated 120-bp bait sequences with 60-bp overlap to capture our 1.4Mbp target region.

Library preparation and sequence capture

We used biotinylated RNA baits from MYcroarray in an in-solution hybridization experiment to capture the immunomes of 16 *G. polyphemus*. We created Illumina adaptor-ligated libraries using Agilent SureSelect XT2 Reagent Kits for the Illumina MiSeq (Agilent Technologies, Santa Clara, CA, USA), pooled 16 prepared libraries per capture reaction, and used MYcroarray reagents and protocols for sequence capture. We then sequenced post-capture amplification libraries on two Illumina MiSeq sequencer flow cells (i.e., all individuals sequenced twice) using 75-bp paired-end reads.

Read quality control and mapping

We demultiplexed reads for each MiSeq run, allowing for up to one mismatch in the 8-bp barcode using `MiSeq Reporter` software. We used `TRIMMOMATIC` v0.32 (Bolger *et al.*, 2014) default settings for adapter trimming, and for base quality filtering, we trimmed leading and trailing bases with quality scores less than 5 and 15, respectively. We also used sliding window scans to remove the 3' end of reads when average quality dropped below 15, and discarded reads with less than 40 bases. We next merged overlapping paired-ends reads with BBMerge v5.4 from the BBMap suite (https://sourceforge.net/projects/bbmap/) and then combined mateless single reads and merged paired reads for downstream analysis. Paired and single plus merged reads were first mapped separately to the *C. p. bellii* 3.0.3 genome using the BWA-MEM algorithm (Li, 2013) implemented in BWA v0.7.12 (Li & Durbin, 2009), and then less stringently using STAMPY v1.0.23 (Lunter & Goodson, 2011). NCBI remap (http://www.ncbi.nlm.nih.gov/genome/tools/remap) was used to convert our bait intervals from *C. p. bellii* 3.0.1 to *C. p. bellii* 3.0.3 coordinates.

Variant and genotype calling

Mapped reads were then processed using the Genome Analysis Toolkit v3.3.0 (McKenna *et al.*, 2010, GATK), adhering to best practices for exome sequencing and calling variants such as SNPs with GATK's Haplotye Caller and Unified Genotyper. Following variant calling, we used PICARD's v1.128 (http://broadinstitute.github.io/picard/) CalculateHSMetrics to estimate sequencing metrics, and featureCounts (Liao *et al.*, 2014) to estimate the number of genes and exons covered by each sample.

We then filtered variants to remove those with bad validation, low quality, low read depth, or low genotype quality to produce a high quality set of SNPs called by the Unified Genotyper. Next, we called variants from base-recalibrated BAM files using the Haplotype Caller and filtered variants in the same manner as before. We then looked for concordance between the two variant callers and used concordant SNPs for variant quality filtering of the Haplotype Caller's call set. Finally, we used BEAGLE v4.0 r1398 (Browning & Browning, 2007) for genotype imputation on the variant-recalibrated SNP set.

Population genomic analyses

For all population genomic analyses, we analyzed only di-allelic polymorphic SNP loci, as the tri- (n=758) and tetra-allelic (n=7) loci we obtained would influence SNP heterozygosity estimates. We used VCFTOOLS v0.1.12b (Danecek *et al.*, 2011) to recalculate allele frequencies from our Beagle-imputated SNPs and then removed loci with allele frequencies of one. We then pruned SNP loci that were out of Hardy-Weinberg Equilibrium (HWE) or in Linkage Disequilibrium (LD) within each population using VCFTOOLS. We used the p.adjust function in R (R Core Team, 2015) to correct P values for HWE and LD tests using a false discovery rate (Benjamini & Hochberg, 1995) of 0.05. For genetic diversity analyses and all subsequent file conversions, we used PGDSpider v2.0.7.4 (Lischer & Excoffier, 2012) and the R package hierfstat v0.04-10 (Goudet, 2005) to assess observed and expected heterozygosity and allelic richness.

For population genomic differentiation, we estimated FST values with hierfstat. For estimating admixture, we performed principle component analyses (PCA) with `hierfstat`, and we also assessed population admixture using `STRUCTURE` v2.3.4 (Hubisz *et al.*, 2009; Pritchard *et al.*, 2000). We ran `STRUCTURE` with 50,000 burnins and 100,000 replicates using correlated allele frequency and the admixture ancestry models from K=1–5 with 10 replicates per K value. We used `STRUCTURE HARVESTER` web v0.6.94 (Earl & vonHoldt, 2012) to select the best K value and `CLUMPAK` web server (Kopelman *et al.*, 2015) to average data from multiple runs and to visualize population assignments.

Microsatellite analyses

We assessed HWE and LD for 10 microsatellite loci using `ARLEQUIN` v3.5 (Excoffier & Lischer, 2010). Genetic diversity, differentiation, and admixture were estimated in the same manner as SNPs using `hierfstat` and `STRUCTURE`.

Random sampling of SNPs for power analysis

We examined how many SNP loci would be needed to obtain P values $<$ 0.05 for Pearson's r correlation coefficient with our 10 microsatellite loci for heterozygosity and FST values by randomly subsampling our 17,901 SNPs. We did not include allelic richness because SNP and microsatellites were not significantly correlated at 0.05 level. We randomly chose 10, 20, 40, 100, 200, 400, 800, 1,600, 3,200, 6,400, or 13,200 SNPs and calculated the P value of the correlation coefficient 10 times for each sample size of SNP loci for observed and expected heterozygosity and FST.

Results

From two Illumina MiSeq sequencer runs, we obtained 47.5 million reads that passed quality control and were assignable to individuals. Each tortoise had 3 ± 0.7 (mean ± standard deviation) million reads of which 47.9 ± 3.2 % were unique (i.e., were not PCR duplicates), and 98.8 ± 0.1 % of these unique reads could be aligned to our target region. Mean sample coverage over the entire target region was 65.4 ± 13 reads, and each sample had 69.3 ± 3.6 % target bases with coverage greater than 20 reads (Fig. 2, Fig. 3). Only 4.7 % (66.3 Kbp) of the 1.4 Mbp target region had coverage of less than 2 reads. Although our target region contained a total of 632 immune genes and 37,275 exons, only 611 genes and 4,837 exons were represented by usable reads. Each sample had reads for 592.1 ± 4.2 genes and 4,106 ± 98.1 exons (mean ± standard deviation).

There were 17,901 di-allelic polymorphic SNP loci after filtering and imputation. None of these loci were out of HWE or in LD, but the lack of LD is unlikely given the close proximity of loci within the same exon and may have occurred because we had to correct P values to account for thousands of multiple tests. Polymorphic SNPs were present in 491 immune genes (Table S1, Supporting information) and included broad classes such as histocompatibility and Toll-like receptor genes (Table 2).

SNP allelic richness was not posivitely correlated with values derived from microsatellites (Fig. 4A, Pearson's r = 0.411, P = 0.294); however, SNP and microsatellite observed (Fig. 4B, Pearson's r = 0.945, P = 0.028) and expected heterozygosities (Fig. 4C, Pearson's r = 0.976, P = 0.012) were highly correlated. The LA population followed by FL then GA and AL populations had the lowest to highest heterozygosity and allelic richness for SNPs. This suggests lower genetic diversity in the western LA population versus eastern FL, GA, and AL populations based on SNPs, a similar result to that obtained with microsatellites.

Pairwise FST values were also positively correlated for SNP and microsatellite markers (Fig. 4D, Pearson's r = 0.96, P = 0.001). However, LA and AL had the lowest differentiation for SNPs compared to second lowest for microsatellites. This discrepancy was also apparent when comparing PCA results as LA and AL were the closest groups in the very tight clusters of the SNP PCA (Fig. 5A) but not for the looser clusters of microsatellite-derived PCA (Fig. 5B) where AL and GA clusters were closer together.

Population admixture inferred using SNPs suggested an optimum number of two clusters with STRUCTURE, the first consisting of AL, GA, and LA; the second with FL by itself (Fig. S1, Supporting information). For microsatellite-inferred admixture, there was an optimum of three clusters: the first with LA; the second with AL and GA; and the third with FL (Fig. S2, Supporting information). PCA analysis produced four clusters for SNPs (one for each population, Fig. 5A) and three clusters for microsatellites (the first with LA; the second with AL and GA; and the third with FL, Fig. 5B).

Random sampling of SNP loci showed that at least 1,600 SNPs were needed to obtain a significant correlation between SNP- and microsatellite-estimated observed heterozygosity (Fig. 6A). Nearly 800 SNPs were needed for expected heterozygosity (Fig. 6B), but only 100 SNPs were needed for SNP- and microsatellite-derived FST values to be correlated (Fig. 6C). Variability decreased as the number of randomly chosen SNPs increased, especially after 100, 40, and 40 SNPs for observed heterozygosity, expected heterozygosity, and FST values respectively (Fig. 6A-6C).

Discussion

Here we sequenced the immunomes of 16 *G. polyphemus* and compared genetic diversity, differentiation, and admixture derived from immunome gene SNPs and values derived from 10 microsatellites from the full set of 101 *G. polyphemus*. We identified nearly 18,000 SNPs among several hundred immune response genes, and observed correlations between estimates of genetic diversity derived from immunome SNPs and microsatellites.

Genetic diversity

Other studies have observed similar and contrasting correlations between SNP versus microsatellite-derived estimates of genetic diversity. For example, previous work using 7 SNPs/indels and 14 microsatellites found that expected heterozygosity and allelic richness are positively correlated between the two types of markers in Atlantic salmon populations (Ryynanen *et al.*, 2007). On the contrary, SNP(n=1–46) and microsatellite (n=10-27) heterozygosities are not correlated for European and North American wolf populations (Vali *et al.*, 2008). Likewise, microsatellite-estimated diversity is different between *Bombus* bumble bee species, but similar when using restriction site-associated DNA sequencing (RADseq) loci (Lozier, 2014), thus diversity estimates from these two markers are not correlated. Further, correlations may or may not exist depending on the diversity of microsatellites as these markers are poly-allelic, compared to SNPs which are typically di-allelic.

Previous work with microsatellites showed that genetic variation was lower in western versus eastern *G. polyphemus* populations (Ennen *et al.*, 2010), and our results support this finding. However, because we only sampled a single western population (Fig. 1), it is not appropriate for us to generalize or label all western populations as genetically depauperate. Ultimately, additional sampling and immunome sequencing from other western *G. polyphemus* populations is warranted.

Although similar, the rank order for allelic richness was not the same for immune gene SNPs and microsatellites. Similar observations have been made by other studies including those comparing SNPs and microsatellites in Atlantic salmon (Ryynanen *et al.*, 2007). Rank order may be skewed between the markers because microsatellites are poly-allelic while SNPs are di-allelic. Differences in rank order may also be caused by the manner in which allelic richness penalizes larger populations. For example, in a preliminary analysis, we incorrectly calculated allelic richness using the wrong population sizes (i.e., we made all populations equal to their original size) and actually found the rank order to be the same for both markers until we noticed the calculation error.

Genetic differentiation

We also observed strong correlations between SNP and microsatellite-derived genetic differentiation, albeit the order of least to most differentiated comparisons varied. The same was observed for SNP- and microsatellite-derived F_{ST} estimates from four populations of western corn rootworms (Coates *et al.*, 2009). The incongruence in rank order may have occurred in both scenarios because of homoplasy issues with microsatellites, where high mutation rates can cause repeat number to revert to a particular allele size, which can then inflate estimates of gene flow (Coates *et al.*, 2009).

Genetic admixture

Population admixture assessments had few inconsistencies between SNPs and microsatellites. Both PCAs suggested four clusters using either marker, but the PCAs varied in which populations were admixed. In particular, LA and AL were closer in the SNP PCA versus GA and FL for the microsatellite PCA. PCAs involving SNPs, expressed sequence tag microsatellites, and anonymous microsatellites among four populations of western corn rootworms found differences among markers in which populations were more closely clustered (Coates *et al.*, 2009). We also observed differences in STRUCTURE admixture results with the optimum number of clusters being 2 for SNPs and 3 for microsatellites. Morin *et al.* (2012) compared 42 SNPs versus 22 microsatellites in bowhead whales and also found the optimum number of clusters is different when using STRUCTURE (optimum number of clusters is 3 for SNPs versus 2 for microsatellites).

Experimental design considerations

So far, we have discussed how population genetic parameters estimated from immune gene SNPs mirror patterns estimated from microsatellite loci, but marker choice also depends on additional considerations such as cost, number of loci, computational issues with NGS generated SNPs, and neutral versus selective processes. First, although sequencing costs are decreasing, NGS techniques can be more expensive than microsatellites on a per sample basis depending on availability of equipment. In particular, the NGS technique used in this paper, in-solution hybridization, requires synthesis of expensive RNA baits/probes, in the order of several thousand dollars (USD). Although tagged microsatellite primers are not trivial in cost, they are far cheaper than biotinylated RNA baits. Further, most genetics labs are not equipped for NGS workflows that require specialized equipment, so lab work must either be outsourced to commercial or non-commercial core facilities.

The number of loci required to adequately address the genetic question at hand is also an important consideration when choosing between SNPs and microsatellites and will vary depending on the question being asked. In general, simulations suggest many more SNPs are needed than microsatellite loci when trying to achieve similar statistical power or parameter estimates. For example, simulations suggest between 60–100 SNP loci are needed for accurate parentage assignment (Anderson & Garza, 2006), and empirical data from sockeye salmon suggest 80 SNPs have higher assignment success and are more accurate for parentage assignment than 11 microsatellites (Hauser *et al.*, 2011). Furthermore, a similar number of SNPs is needed for detecting low levels of divergence (i.e., FST < 0.005) (Morin *et al.*, 2009). Ryynanen *et al.* (2007) had significant correlations between 7 SNPs/indels and 14 microsatellite loci when estimating FST. Our simulation results suggest at least 100 SNP loci are needed for correlating SNP and microsatellite-derived FST. For heterozygosity, our data suggest more than 800 SNP loci are needed to correlate with 10 microsatellite loci, but Ryynanen *et al.* (2007) only needed 7 SNP/indel loci to reach similar correlation levels possibly because they analyzed 21 populations. Acquiring data from a large number of SNPs is not a problem with NGS approaches, rather not all SNP loci are equally informative, and smaller SNP panels may occasionally perform well in comparison to much larger SNP arrays.

Computational issues with NGS are also not trivial, as our own NGS analysis relied on high performance computing resources and required many gigabytes of data storage. This does not include the time or expertise required to write code and scripts to analyze the gigabytes of raw data.

Neutral versus selective processes are also important to consider when deciding between SNPs and microsatellites. Markers such as microsatellites will be neutrally evolving while SNPs could represent both functional and neutral markers and be influenced by both neutral and adaptive processes. NGS generated SNPs are also to be appreciated because the sequences used to identify SNPs can represent functional, coding regions of the genome, which can provide information on the adaptive potential of populations to respond to environmental change (Meyers & Bull, 2002; van Tienderen *et al.*, 2002).

Many tests exist for ascertaining whether NGS generated sequences or SNPs are putatively under selection (reviewed in Vitti *et al.*, 2013). In particular, SNP allele frequencies can vary among populations possibly due to selection acting on SNPs in one but not the other populations, and these different allele frequencies can influence genetic differentiation that can be approximated with Wright's fixation index (FST). Several methods exist for detecting outlier SNPs that are putatively under selection, and these so-called outlier FST tests have been reviewed in Narum & Hess (2011).

Conclusion

As more and more population genetic studies are publishing NGS generated SNPs as opposed to microsatellites, it would be useful to identify patterns between microsatellites and NGS derived SNPs and to appreciate the additional functional information commonly provided by SNPs. One apparent pattern is that high variation observed at microsatellites will likely translate into high SNP-estimates of genetic diversity (Ryynanen *et al.*, 2007) and vice versa. Further, genetic diversity estimated by allelic richness between microsatellites and SNPs may be a less stable metric than diversity estimated by observed and/or expected heterozygosity because large populations are penalized by allelic richness and more alleles are present in microsatellites than SNPs. This does not mean allelic richness should be ignored especially for conservation purposes because some traits including disease resistance are associated with particular alleles (e.g., Langefors *et al.*, 2001), which is not accounted for by heterozygosity. Another important pattern likely to be observed between microsatellites and SNP studies is presence/absence of genetic structure, with any potential inconsistencies resulting from different evolutionary forces acting on the markers. The addition of adaptive processes acting on SNPs can then result in similar but disparate structure patterns between the two marker types. Finally, given the consistencies found between the two markers types here, we don't think it is necessary for researchers to replace older microsatellite data with NGS data as microsatellite-based management plans are probably still relevant.

Acknowledgements

We thank LSU and the LSU AgCenter for financial and logistical support. The Lucius Gilbert Foundation provided support for sequencing and for J.P.E. We are grateful to Richard Carmouche of Pennington Biomedical Research Center's Genomic Core Facility for performing next-generation sequencing laboratory work. This project/work used Genomics core facilities that are supported in part by COBRE (NIH 8 P20 GM103528) and NORC (NIH 2P30DK072476) center grants from the National Institutes of Health. We used LSU High Performance Computing resources to analyze next-generation sequencing data.

References

Allendorf FW, Hohenlohe PA, Luikart G (2010) Genomics and the future of conservation genetics. *Nature Reviews Genetics*, **11**, 697–709.

Anderson EC, Garza JC (2006) The power of single-nucleotide polymorphisms for large-scale parentage inference. *Genetics*, **172**, 2567–2582.

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**, 289–300.

Bolger AM, Lohse M, Usadel B (2014) TRIMMOMATIC: a flexible trimmer for Illumina sequence data. *Bioinformatics*, p. btu170.

Brown MB, McLaughlin GS, Klein PA, *et al.* (1999) Upper respiratory tract disease in the gopher tortoise is caused by Mycoplasma agassizii. *Journal of Clinical Microbiology*, **37**, 2262–2269.

Browning SR, Browning BL (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American Journal of Human Genetics*, **81**, 1084–97.

Clostio RW, Martinez AM, LeBlanc KE, Anthony NM (2012) Population genetic structure of a threatened tortoise across the south-eastern United States: implications for conservation management. *Animal Conservation*, **15**, 613–625.

Coates BS, Sumerford DV, Miller NJ, *et al.* (2009) Comparative performance of single nucleotide polymorphism and microsatellite markers for population genetic analysis. *Journal of Heredity*, **100**, 556–564.

Danecek P, Auton A, Abecasis G, *et al.* (2011) The Variant Call Format and VCFtools. *Bioinformatics*, **27**, 2156–8.

Earl DA, vonHoldt BM (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources*, **4**, 359–361.

Elbers JP, Taylor SS (2015) GO2TR: a gene ontology-based workflow to generate target regions for target enrichment experiments. *Conservation Genetics Resources*, **7**, 851–857.

Ennen JR, Kreiser BR, Qualls CP (2010) Low genetic diversity in several gopher tortoise (Gopherus polyphemus) populations in the Desoto National Forest, Mississippi. *Herpetologica*, **66**, 31–38.

Excoffier L, Lischer HE (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources*, **10**, 564–7.

Glover KA, Hansen MM, Lien S, Als TD, Hoyheim B, Skaala O (2010) A comparison of SNP and STR loci for delineating population structure and performing individual genetic assignment. *BMC Genetics*, **11**, 1–12.

Goudet J (2005) HIERFSTAT, a package for R to compute and test hierarchical F-statistics. *Molecular Ecology Notes*, **5**, 184–186.

Gupta P, Roy J, Prasad M (2001) Single nucleotide polymorphisms SNPs: a new paradigm in molecular marker technology and DNA polymorphism detection with emphasis on their use in plants. *Current Science*, **80**, 524–535.

Hauser L, Baird M, Hilborn RAY, Seeb LW, Seeb JE (2011) An empirical comparison of SNPs and microsatellites for parentage and kinship assignment in a wild sockeye salmon (Oncorhynchus nerka) population. *Molecular Ecology Resources*, **11**, 150–161.

Hubisz MJ, Falush D, Stephens M, Pritchard JK (2009) Inferring weak population structure with the assistance of sample group information. *Molecular Ecology Resources*, **9**, 1322–32.

Kopelman NM, Mayzel J, Jakobsson M, Rosenberg NA, Mayrose I (2015) CLUMPAK: a program for identifying clustering modes and packaging population structure inferences across k. *Molecular Ecology Resources*, **15**, 1179–91.

Langefors A, Lohm J, Grahn M, Andersen O, von Schantz T (2001) Association between major histocompatibility complex class iib alleles and resistance to Aeromonas salmonicida in atlantic salmon. *Proceedings of the Royal Society of London B*, **268**, 479–485.

Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*, **1303.3997**.

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–60.

Liao Y, Smyth GK, Shi W (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, **30**, 923–30.

Lischer H, Excoffier L (2012) PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics*, **28**, 298–299.

Lozier JD (2014) Revisiting comparisons of genetic diversity in stable and declining species: assessing genome-wide polymorphism in North American bumble bees using RAD sequencing. *Molecular Ecology*, **23**, 788–801.

Lunter G, Goodson M (2011) STAMPY: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Research*, **21**, 936–9.

McKenna A, Hanna M, Banks E, *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, **20**, 1297–1303.

Meyers LA, Bull JJ (2002) Fighting change with change: adaptive variation in an uncertain world. *Trends in Ecology and Evolution*, **17**, 551–557.

Morin PA, Archer FI, Pease VL, *et al.* (2012) An empirical comparison of SNPs and microsatellites for population structure, assignment, and demographic analyses of bowhead whale populations. *Endangered Species Research*, **19**, 129–147.

Morin PA, Martien KK, Taylor BL (2009) Assessing statistical power of SNPs for population structure and conservation studies. *Molecular Ecology Resources*, **9**, 66–73.

Moritz C (1994) Applications of mitochondrial DNA analysis in conservation: a critical review. *Molecular Ecology*, **3**, 401–411.

Narum SR, Banks M, Beacham TD, *et al.* (2008) Differentiating salmon populations at broad and fine geographical scales with microsatellites and single nucleotide polymorphisms. *Molecular Ecology*, **17**, 3464–3477.

Narum SR, Hess JE (2011) Comparison of FST outlier tests for SNP loci under selection. *Molecular Ecology Resources*, **11**, 184–194.

Ortutay C, Vihinen M (2006) Immunome: a reference set of genes and proteins for systems biology of the human immune system. *Cellular Immunology*, **244**, 87–89.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

R Core Team (2015) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Richter SC, Jackson JA, Hinderliter M, Epperson DM, Theodorakis CW, Adams SM (2011) Conservation genetics of the largest cluster of federally threatened gopher tortoise (Gopherus polyphemus) colonies with implications for species management. *Herpetologica*, **67**, 406–419.

Ryynanen HJ, Tonteri A, Vasemagi A, Primmer CR (2007) A comparison of biallelic markers and microsatellites for the estimation of population and conservation genetic parameters in Atlantic salmon (Salmo salar). *Journal of Heredity*, **98**, 692–704.

Schwartz TS, Karl SA (2005) Population and conservation genetics of the gopher tortoise (Gopherus polyphemus). *Conservation Genetics*, **6**, 917–928.

Shaffer HB, Minx P, Warren DE, *et al.* (2013) The western painted turtle genome, a model for the evolution of extreme physiological adaptations in a slowly evolving lineage. *Genome Biology*, **14**, R28.

van Tienderen PH, de Haan AA, van der Linden CG, Vosman B (2002) Biodiversity assessment using markers for ecologically important traits. *Trends in Ecology and Evolution*, **17**, 577–582.

Vali U, Einarsson A, Waits L, Ellegren H (2008) To what extent do microsatellite markers reflect genome-wide genetic diversity in natural populations? *Molecular Ecology*, **17**, 3808–3817.

Vitti JJ, Grossman SR, Sabeti PC (2013) Detecting natural selection in genomic data. *Annual Review of Genetics*, **47**, 97–120.

Data Accessibility

Raw sequencing data are available from the Sequence Read Archive (accession: SRP061247). BAM and VCF files are available from Dryad repository (doi:). Detailed analytical methods and scripts to create Tables and Figures are available from `https://github.com/jelber2/immunome_2014`.

Author Contributions

J.P.E. designed the study and performed SNP analyses. R.W.C. performed microsatellite analyses. J.P.E. and S.S.T. wrote the paper.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table S1 All genes with di-allelic, polymorphic SNPs.

Fig. S1 `STRUCTURE` plot for 17,901 immune gene SNPs with optimum number of clusters $K = 2$ determined by `STRUCTURE HARVESTER`.

Fig. S2 `STRUCTURE` plot for 10 microsatellites with optimum number of clusters $K = 3$ determined by `STRUCTURE HARVESTER`.

Tables and Figures

Table 1 *Gopherus polyphemus* sample descriptions.

Site (Abbreviation)	N	Latitude	Longitude
Florida Gas Pipeline, Washington Parish, LA (LA)	36	30.78	-90.00
Solon Dixon, Andalusia, AL (AL)	20	31.16	-86.70
Jones Ecological Research Center, GA (GA)	26	31.23	-84.47
Private Site, Nassau County, FL (FL)	19	30.59	-81.56

Table 2 Histocompatibility and Toll-like Receptor Loci with di-allelic, polymorphic SNPs.

Histocompatibility Loci
CD74 molecule, major histocompatibility complex, class II invariant chain
Class I histocompatibility antigen, F10 alpha chain-like
Class II histocompatibility antigen, M alpha chain
Class II, major histocompatibility complex, transactivator
DLA class II histocompatibility antigen, DR-1 beta chain-like
H-2 class II histocompatibility antigen, A-R alpha chain-like
H-2 class II histocompatibility antigen, E-S beta chain-like
HLA class II histocompatibility antigen, DP alpha 1 chain-like
HLA class II histocompatibility antigen, DR alpha chain-like
HLA class II histocompatibility antigen, DR beta 5 chain-like
HLA class II histocompatibility antigen, DRB1-15 beta chain-like
Major histocompatibility complex class I-related gene protein-like
Rano class II histocompatibility antigen, A beta chain-like
Toll-like Receptor Loci
Toll-like Receptor 13
Toll-like Receptor 2
Toll-like Receptor 7
Toll-like Receptor 8
Toll-like Receptor adaptor molecule 1
Toll-like Receptor adaptor molecule 2

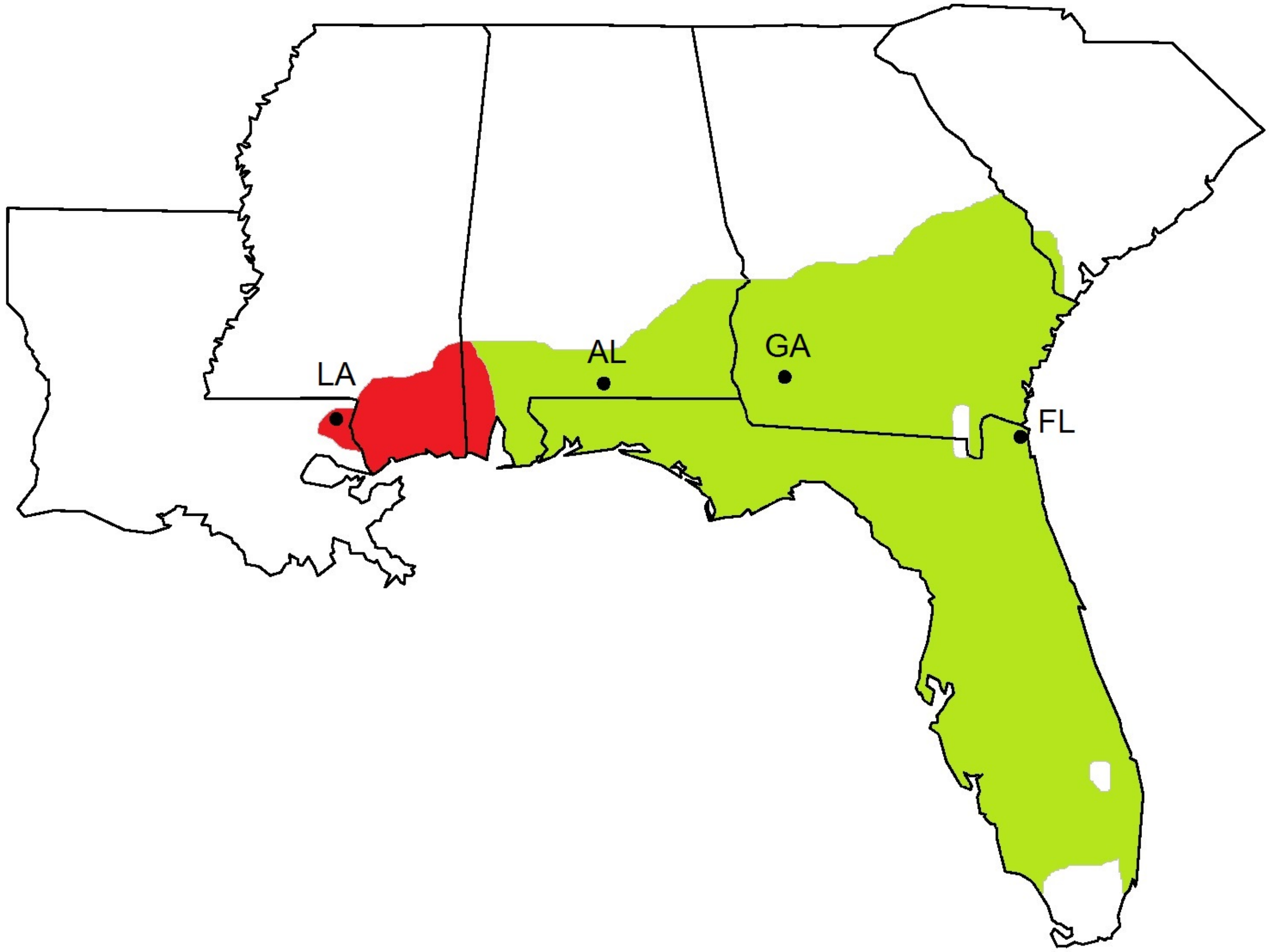

Fig. 1 *Gopherus polyphemus* range map and sampling sites. Range of western *G. polyphemus* populations darkly shaded on the left with eastern populations lightly shaded on the right.

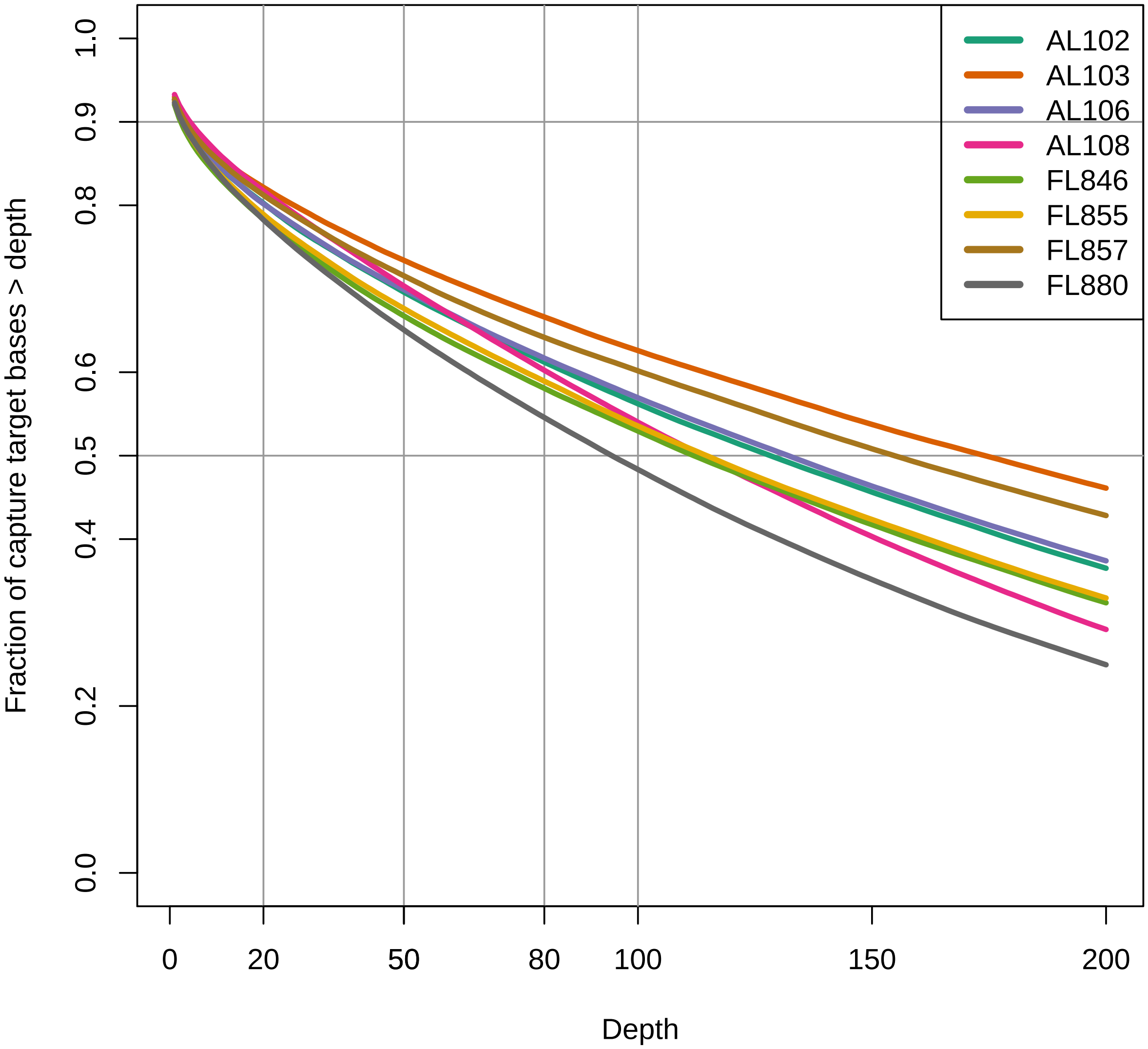

Fig. 2 Coverage plots for first eight samples showing number of sequencing reads at or above specified proportions. A value at 100 Depth and 0.5 fraction means 50 percent of bases were at or above 100X coverage.

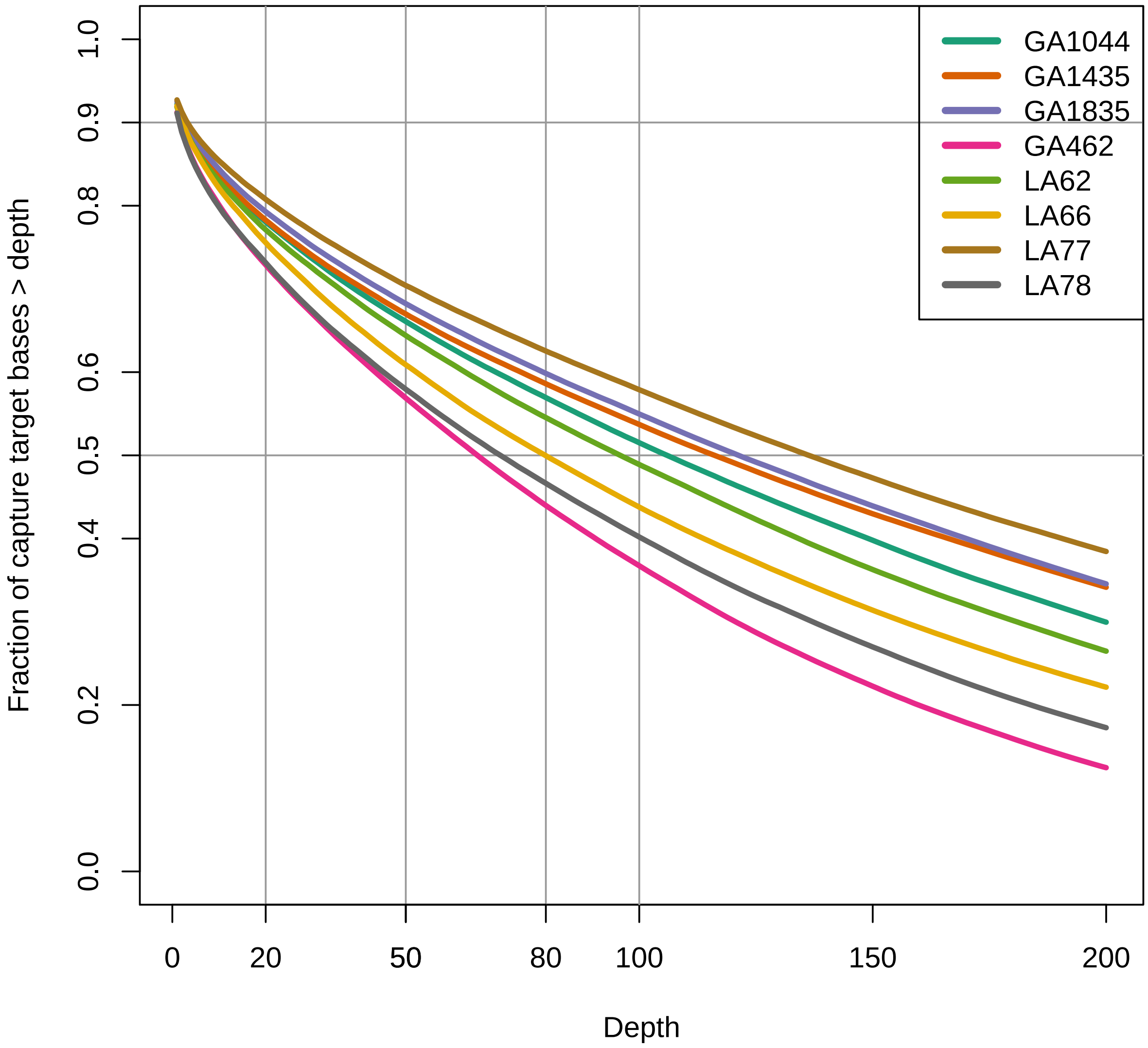

Fig. 3 Coverage plots for last eight samples showing number of sequencing reads at or above specified proportions.

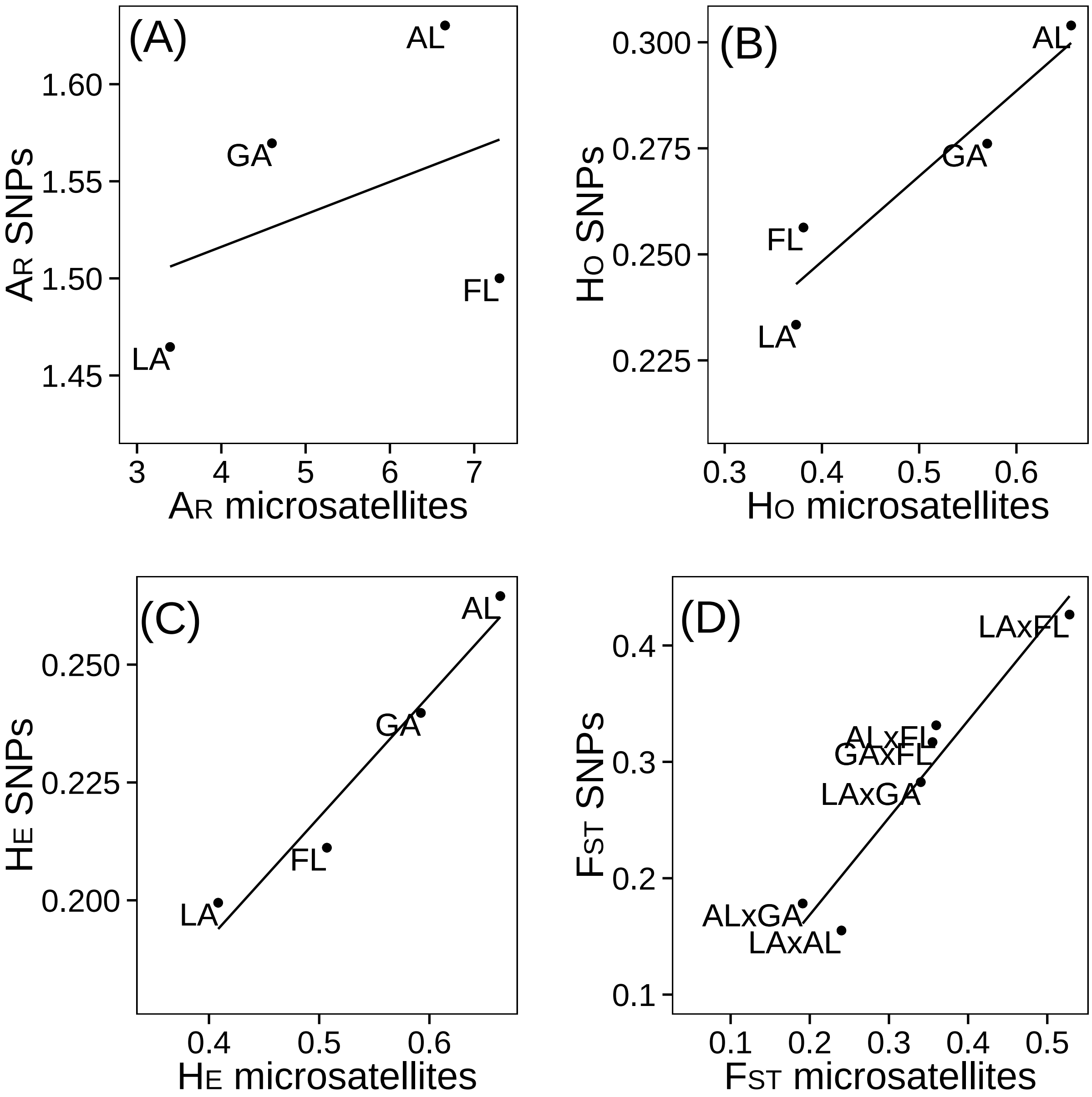

Fig. 4 (A) Allelic richness, (B) observed heterozygosity, (C) expected heterozygosity, and (D) FST comparison between 10 microsatellites and 17,901 immune gene SNPs. AR for allelic richness, HO for observed heterozygosity, HE for expected heterozygosity.

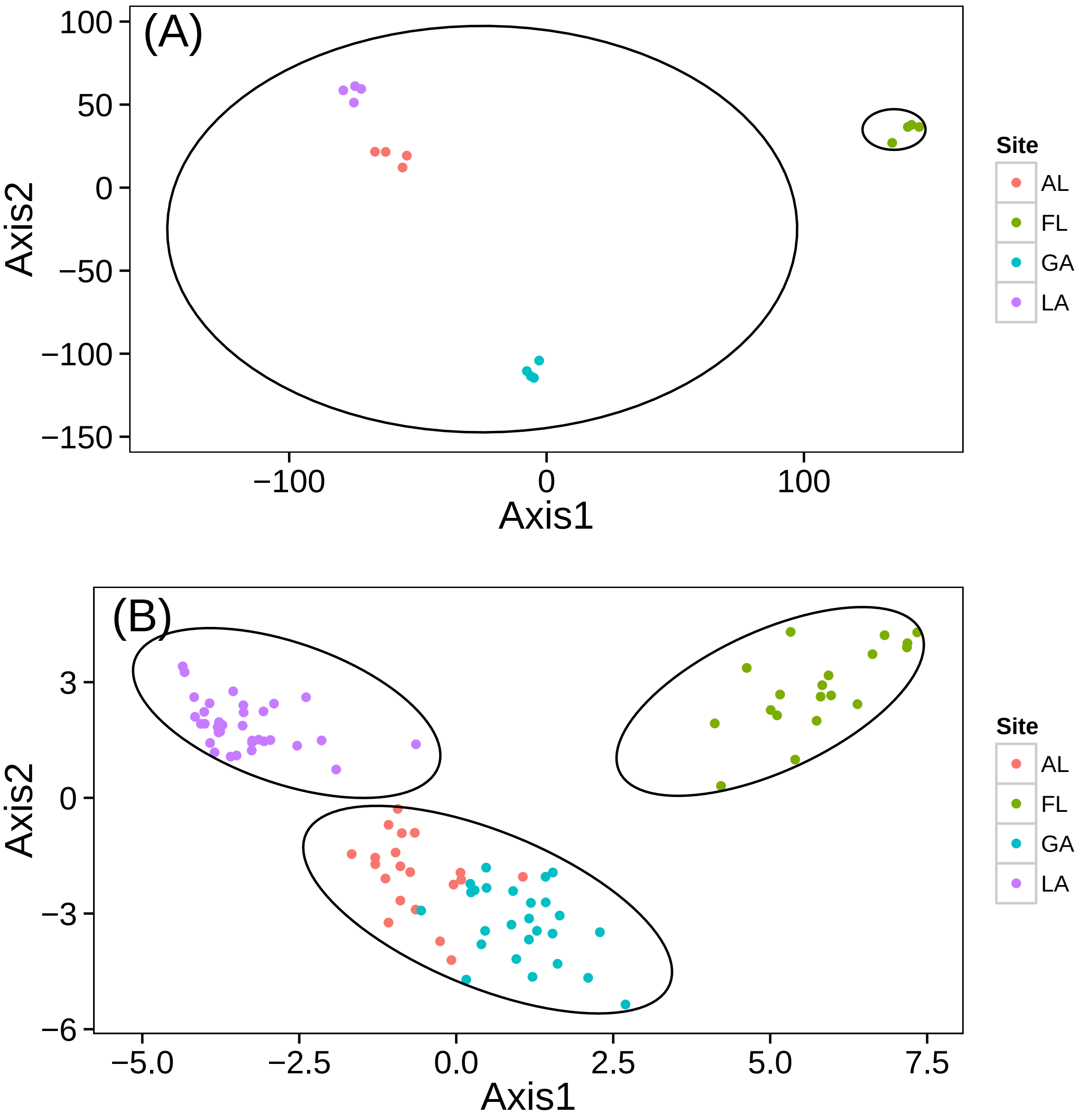

Fig. 5 Principle component analysis for (A) 17,901 immune gene SNPs and (B) 10 microsatellites. Circles indicate optimum clusters indentified using STRUCTURE and STRUCTURE HARVESTER.

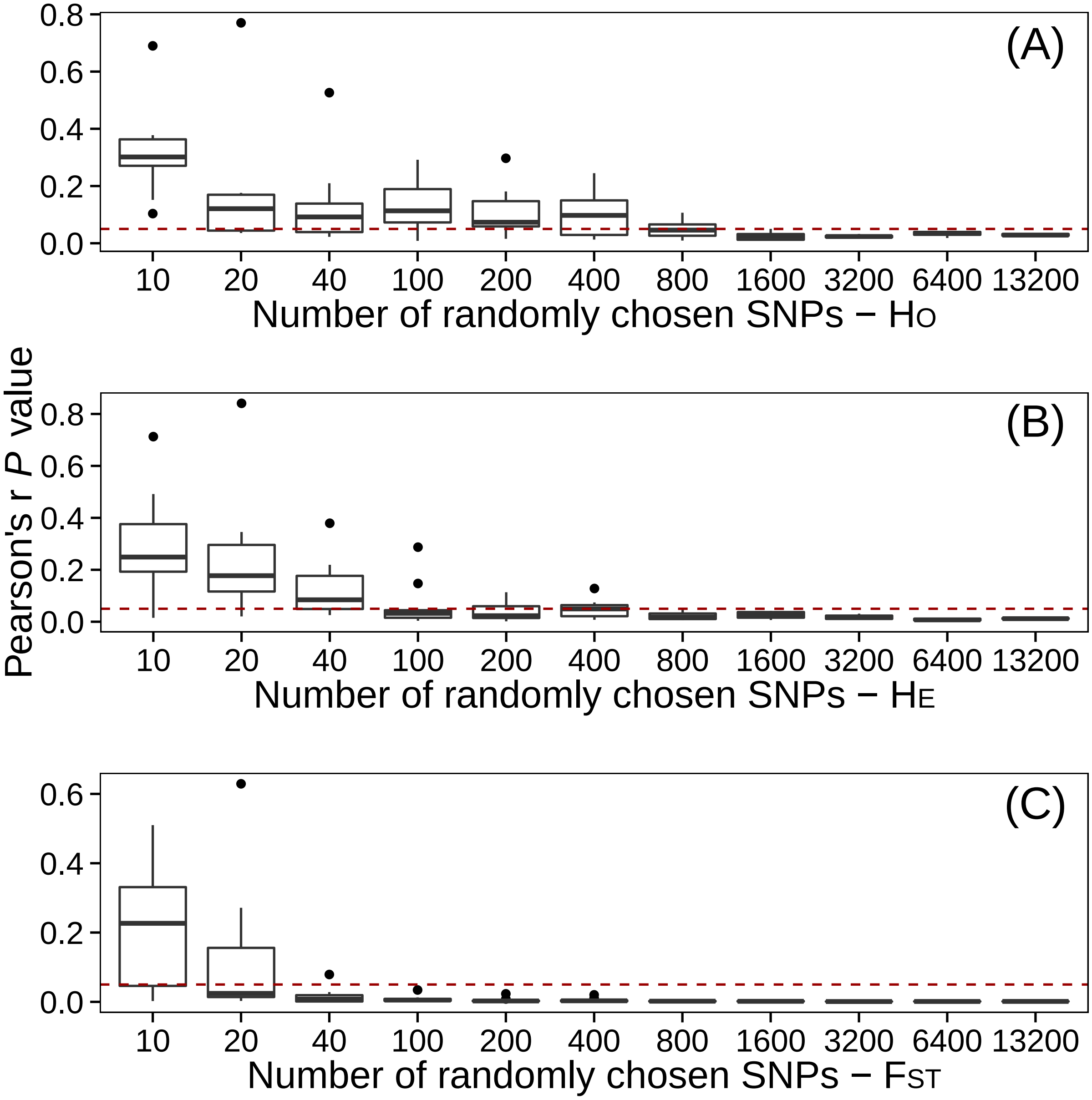

Fig. 6 Power analysis showing how many randomly sampled SNP loci are needed in comparison to 10 microsatellite loci for Pearon's r correlation coefficient to be significant at 0.05 level (dotted line) for (A) observed heterozygosity, (B) expected heterozygosity, and (C) FST. There were 10 simulations for each size class of SNPs. HO for observed heterozygosity, HE for expected heterozygosity.