5-13-2015

Development of diagnostic microsatellite markers from whole-genome sequences of Ammodramus sparrows for assessing admixture in a hybrid zone

Adrienne I. Kovach  
*University of New Hampshire, Durham*, Adrienne.Kovach@unh.edu

Jennifer Walsh  
*University of New Hampshire, Durham*

Jordan Ramsdell  
*University of New Hampshire, Durham*

William Kelley Thomas  
*University of New Hampshire, Durham*, Kelley.Thomas@unh.edu

Follow this and additional works at: [https://scholars.unh.edu/nhaes](https://scholars.unh.edu/nhaes)

**Recommended Citation**

This Article is brought to you for free and open access by the Research Institutes, Centers and Programs at University of New Hampshire Scholars' Repository. It has been accepted for inclusion in New Hampshire Agricultural Experiment Station by an authorized administrator of University of New Hampshire Scholars' Repository. For more information, please contact nicole.hentz@unh.edu.
Development of diagnostic microsatellite markers from whole-genome sequences of Ammodramus sparrows for assessing admixture in a hybrid zone

Rights
© 2015 The Authors. Ecology and Evolution published by John Wiley & Sons Ltd.

Comments
This is an article published by Wiley in Ecology and Evolution in 2015, available online: https://dx.doi.org/10.1002/ece3.1514

This article is available at University of New Hampshire Scholars' Repository: https://scholars.unh.edu/nhaes/377
Development of diagnostic microsatellite markers from whole-genome sequences of Ammodramus sparrows for assessing admixture in a hybrid zone

Adrienne I. Kovach¹, Jennifer Walsh¹, Jordan Ramsdell² & W. Kelley Thomas²

¹Department of Natural Resources and the Environment, University of New Hampshire, Durham, New Hampshire 03824
²Department of Molecular, Cellular and Biomedical Sciences, University of New Hampshire, Durham, New Hampshire 03824

Keywords
Admixture, Ammodramus, diagnostic markers, hybridization, next-generation sequencing.

Abstract
Studies of hybridization and introgression and, in particular, the identification of admixed individuals in natural populations benefit from the use of diagnostic genetic markers that reliably differentiate pure species from each other and their hybrid forms. Such diagnostic markers are often infrequent in the genomes of closely related species, and genomewide data facilitate their discovery. We used whole-genome data from Illumina HiSeqS2000 sequencing of two recently diverged (600,000 years) and hybridizing, avian, sister species, the Salt-marsh (Ammodramus caudacutus) and Nelson’s (A. nelsoni) Sparrow, to develop a suite of diagnostic markers for high-resolution identification of pure and admixed individuals. We compared the microsatellite repeat regions identified in the genomes of the two species and selected a subset of 37 loci that differed between the species in repeat number. We screened these loci on 12 pure individuals of each species and report on the 34 that successfully amplified. From these, we developed a panel of the 12 most diagnostic loci, which we evaluated on 96 individuals, including individuals from both allopatric populations and sympatric individuals from the hybrid zone. Using simulations, we evaluated the power of the marker panel for accurate assignments of individuals to their appropriate pure species and hybrid genotypic classes (F1, F2, and backcrosses). The markers proved highly informative for species discrimination and had high accuracy for classifying admixed individuals into their genotypic classes. These markers will aid future investigations of introgressive hybridization in this system and aid conservation efforts aimed at monitoring and preserving pure species. Our approach is transferable to other study systems consisting of closely related and incipient species.

Introduction
Interspecific hybridization is common in nature (Mallet 2005; Abbott et al. 2013), especially among species in early stages of speciation or in secondary contact (Via 2009; Ellegren et al. 2012). Wild hybrids are a mosaic of phenotypes and genotypes, creating challenges for their taxonomic identification and confusion about their conservation status (Stronen and Paquet 2013). Accurate identification of admixed individuals in wild populations aids evolutionary investigations of introgressive hybridization as well as conservation management.

Studies of genetic admixture are most powerful when they use diagnostic species-specific markers, that is markers that are highly differentiated between the two parental species (Moccia et al. 2007; Hohenlohe et al. 2011). Yet, diagnostic markers are infrequent in the genomes of closely related species, and they are rarely found by anonymous marker development approaches. Current sequencing technologies present solutions to the challenges of identifying diagnostic markers, through efficient development of large genomewide panels of SNPs or microsatellite loci. Despite the advent and potential power of large SNP panels, there are many research questions, including those involving genetic hybrid indices, that can be effectively addressed with a carefully selected suite of highly informative microsatellite markers (Guichoux et al. 2011; Wei et al. 2014; Vukosavljev et al. 2015).
Diagnostic Markers for Admixture in *Ammodramus*

A. I. Kovach et al.

2015). High-throughput sequencing greatly enhances de novo microsatellite development and results in the low cost recovery of tens of thousands of repeat-containing sequences (Malaua et al. 2011; Reid et al. 2012). Diagnostic marker development can capitalize on this wealth of repeat sequence data to identify markers in the few genomic regions that are differentiated between closely related species. By generating sequencing data from both species’ genomes, screening markers for repeat differences can be performed in silico, thereby saving tremendously on the laborious process of screening loci in the laboratory. In this study, we developed such a strategy for comparing the repeat sequences generated by whole-genome shotgun sequencing of two hybridizing avian sister species, to identify a suite of diagnostic markers for high-resolution identification of pure and admixed individuals in an avian hybrid zone.

Saltmarsh and Nelson’s sparrows (*Ammodramus caudacutus* and *A. nelsoni*; Fig. 1) belong to a unique group of terrestrial vertebrates that rely primarily or exclusively on tidal marsh habitats (Greenberg et al. 2006a). As such, they are excellent models for studying local environmental adaptation and ecological speciation (Greenberg and Maldonado 2006; Greenberg 2006). They are also species of high conservation priority along the northeastern Atlantic coast of North America, where they breed (U.S. Department of Interior (USDI) 2008). *A. caudacutus* breeds exclusively in coastal marshes from mid-Maine to Virginia, USA (Greenlaw and Woolfenden 2007); it is globally threatened because of its limited range and obligate habitat requirements (IUCN Red List criteria; Birdlife International 2004). *A. nelsoni* has a wider ecological niche, and one of three subspecies (*A. n. subvirgatus*) breeds in tidal marshes, brackish waters, and hay fields from coastal Quebec to northeastern Massachusetts (Greenlaw and Woolfenden 2007; Nocera et al. 2007).

![Figure 1. Pure Saltmarsh Sparrow (Ammodramus caudacutus) on the left and pure Nelson's Sparrow (Ammodramus nelsoni) on the right. The two tidal marsh birds hybridize in an overlap zone along the northeastern Atlantic coast, from northern Massachusetts to southern Maine, USA.](image)

These young species diverged ~600,000 years ago (Rising and Avise 1993), as evidenced by weak genetic divergence (1.2% differentiation at the COI gene and $F_{ST}$ of ~0.15 for neutral microsatellite markers; Shriver et al. 2005; Walsh et al. 2011). They co-occur and hybridize in tidal marshes of the New England coast, where they are now in secondary contact.

Hybrid *A. caudacutus-nelsoni* sparrows are prevalent within the overlap zone and reveal a complex and poorly understood pattern of morphological and genetic introgression (Hodgman et al. 2002; Shriver et al. 2005; Walsh et al. 2011). Currently available microsatellite markers yield low levels of differentiation within and between species (Shriver et al. 2005; Walsh et al. 2012) and lack the resolution to differentiate genotypic classes of admixed individuals (e.g., F1, F2, and backcrossed to each parental species). Difficulties in distinguishing pure and admixed individuals hinder efforts to evaluate productivity and viability of populations in the hybrid zone, as well as to fully evaluate the geographic extent of introgression. Diagnostic markers are germane for addressing these concerns as well as for investigating patterns and mechanisms of introgressive hybridization.

The aim of this study was to use whole-genome sequence data of *A. caudacutus* and *nelsoni* for de novo development of a suite of species-specific diagnostic microsatellite markers with high resolution for identifying pure and hybrid genotypic classes (F1, F2, and backcrosses). To do so, we identified putative diagnostic markers by in silico comparison of repeat sequences in the two genomes, and we screened 37 of them in the laboratory on individuals of both species. We then developed a panel of the 12 most diagnostic markers, which we found through additional screening to be highly suitable for a genetic hybrid index. We evaluated the power of the markers for accurate assignments of simulated individuals to their appropriate hybrid genotypic classes. Our approach and PERL script for identifying diagnostic repeats between two genomes are readily transferrable to other study systems consisting of closely related and incipient species.

**Materials and Methods**

**Sampling and DNA extraction**

To obtain samples for marker development, we sampled a total of 120 *A. caudacutus* and *nelsoni* individuals from multiple putatively allopatric ($n = 48$ individuals of each species) and sympatric ($n = 24$ individuals) locations along the northeastern coast of the United States, within and north and south of the species’ overlap zone (Table 1). Adult sparrows were captured using 12-m mist nests with size 36-mm mesh. Blood samples (10–20 μL)
Table 1. Sampling locations and sample sizes of *Ammodramus caudacutus* and *A. nelsoni* used in this study. Locations outside of the currently documented overlap zone (considered allopatric populations in this study) are in bold.

<table>
<thead>
<tr>
<th>Sampling Location</th>
<th>Latitude</th>
<th>Longitude</th>
<th>n</th>
<th>Sample use</th>
</tr>
</thead>
</table>
| Lubec, Maine      | 44.822    | –66.991   | 9  
*nelsoni, 1 caudacutus* | Initial screening/marker characterization |
| Columbia Falls, Maine | 44.644 | –67.719   | 9  
*nelsoni* | Initial screening/marker characterization |
| Narraguagus River – Milbridge, Maine | 44.551 | –68.891   | 9  
*nelsoni* | Initial screening/marker characterization |
| Penobscot River – Penobscot, Maine | 44.591 | –68.859   | 1  
*nelsoni* | Whole-genome sequencing |
| Frankfort, Maine  | 44.587    | –68.858   | 12  
*nelsoni* | Initial screening/marker characterization |
| Winterport, Maine | 44.623    | –68.854   | 9  
*nelsoni* | Initial screening/marker characterization |
| Weskeag Marsh – South Thomaston, Maine | 44.077 | –69.142   | 1  
*nelsoni, 1 caudacutus* | Marker characterization |
| Sheepscot River – Newcastle, Maine | 44.065 | –69.597   | 2  
*nelsoni* | Marker characterization |
| Popham Beach – Phippsburg, Maine | 43.739 | –69.806   | 1  
*nelsoni, 1 caudacutus* | Marker characterization |
| Maquoit Bay – Brunswick, Maine | 43.867 | –69.988   | 1  
*nelsoni* | Marker characterization |
| Cousins River – Yarmouth, Maine | 43.811 | –70.156   | 1  
*nelsoni* | Marker characterization |
| Saco River – Saco, Maine – Rachel Carson NWR | 43.492 | –70.391   | 2  
*nelsoni, 1 caudacutus* | Marker characterization |
| Marshall Point – Arundel, Maine – Rachel Carson NWR | 43.381 | –70.433   | 1  
*nelsoni* | Marker characterization |
| Elderidge Marsh – Wells, Maine – Rachel Carson NWR | 43.292 | –70.572   | 1  
*nelsoni, 1 caudacutus* | Marker characterization |
| Kittery Point, Maine – Rachel Carson NWR | 43.087 | –70.664   | 1  
*caudacutus* | Marker characterization |
| Lubberland Creek – Newmarket, New Hampshire | 43.073 | –70.903   | 1  
*nelsoni, 1 caudacutus* | Marker characterization |
| Great Bay NERR |  |  |  |  |
| Chapman’s Landing – Stratham, New Hampshire | 43.041 | –70.924   | 1  
*caudacutus* | Marker characterization |
| Great Bay NERR |  |  |  |  |
| Axcomin Marsh – Rye, New Hampshire | 43.006 | –70.752   | 1  
*nelsoni, 1 caudacutus* | Marker characterization |
| Hampton Beach, New Hampshire | 42.926 | –70.806   | 1  
*caudacutus* | Marker characterization |
| Salisbury, Massachusetts | 42.844 | –70.822   | 1  
*caudacutus* | Marker characterization |
| Plum Island – Newbury, Massachusetts (Parker River NWR) | 42.774 | –70.809   | 4  
*caudacutus* | Initial screening/marker characterization |
| Revere, Massachusetts | 42.436 | –71.011   | 5  
*caudacutus* | Marker characterization |
| Duxbury, Massachusetts | 42.053 | –70.681   | 1  
*caudacutus* | Marker characterization |
| Waquoit Bay – Mashpee, Massachusetts (Waquoit Bay NERR) | 41.555 | –70.506   | 2  
*caudacutus* | Marker characterization |
| Prudence Island, Rhode Island | 41.625 | –71.321   | 9  
*caudacutus* | Initial screening/marker characterization |
| Middletown, Rhode Island – Sachuest Point NWR | 41.488 | –71.249   | 1  
*caudacutus* | Whole-genome sequencing |
| Narragansett, Rhode Island – John H. Chafee NWR | 41.442 | –71.467   | 9  
*caudacutus* | Initial screening/marker characterization |
| Shirley, New York – Wertheim NWR | 40.771 | –72.889   | 3  
*caudacutus* | Initial screening |
| Oceanside, New York – Oceanside Marine Nature Center | 40.622 | –73.624   | 2  
*caudacutus* | Initial screening |
| North Sea, New York – Scallop Pond Preserve | 40.944 | –72.429   | 2  
*caudacutus* | Marker characterization |
| Sag Harbor Bay – Noyack, New York | 41.022 | –72.306   | 3  
*caudacutus* | Marker characterization |
| North Cider Island – Lido Beach, New York | 40.602 | –73.611   | 3  
*caudacutus* | Marker characterization |
| Plum Bank Creek – Old Saybrook, Connecticut | 41.269 | –72.391   | 2  
*caudacutus* | Marker characterization |
| Farm River State Park – East Haven, Connecticut | 41.155 | –72.857   | 2  
*caudacutus* | Marker characterization |
| Milford, Connecticut | 41.218 | –73.035   | 1  
*caudacutus* | Marker characterization |
| Watts Island – Niantic, Connecticut | 41.299 | –72.219   | 1  
*caudacutus* | Marker characterization |
| West River – West Haven, Connecticut | 41.291 | –72.945   | 1  
*caudacutus* | Marker characterization |
were collected from the brachial vein onto Nobuto blood filter strips (Advantec MFS Inc., Dublin CA). For de novo marker development, two additional females, one nelsoni captured from Penobscot, Maine, and one caudacutus captured from Middletown, Rhode Island, were blood-sampled for whole-genome sequencing. These individuals were assumed to be “pure” for each parental species, as they were sampled from locations outside of the currently recognized hybrid zone (Hodgman et al. 2002). DNA was extracted from blood samples using a DNeasy Blood Kit (Qiagen, Valencia, CA).

Genome sequencing and assembly

Illumina TruSeq DNA libraries were generated including electrophoretic, gel-based, manual size selection targeting an average insert size of 300 bp. Whole-genome 100–base pair, paired-end sequencing was performed in two separate lanes on an Illumina HiSeqS2000. This resulted in 213,519,998 and 384,563,744 100-base-pair reads for A. caudacutus and nelsoni genomes, respectively. Total reads and the most variation in the distribution of alleles to screen further as a panel of putatively diagnostic loci. We counted the number of alleles shared between the species across the 24 allopatric individuals (Table 2). We chose 12 loci with the fewest number of shared alleles and the most variation in the distribution of alleles to screen further as a panel of putatively diagnostic loci. These 12 chosen loci were screened in an additional 96 individuals (36 allopatric and 12 sympatric individuals of each species), using dye-labeled primers (HEX, FAM, or NED) in two multiplex PCRs. The 15 µL polymerase chain reactions contained 3 µL of eluted genomic DNA, 0.3 µL of each primer, 0.3 µL of taq DNA polymerase (Promega), 1X PCR buffer (Promega, Madison, WI, USA), 0.2 mmol/L of deoxyribonucleotides, and 1 unit of taq DNA polymerase (Promega). Cycling conditions were as follows: initial denaturation at 94°C for 4 min, followed by 30 cycles of 94°C for 30 sec, 56°C for 45 sec, 72°C for 1 min, and a final extension step at 72°C for 5 min. PCR products were resolved on a 1% agarose gel. Of the 37 primers, 34 consistently amplified the target regions in both species and were used for an initial screening of 24 individuals from eight allopatric marshes (Table 1). PCR was repeated with the addition of 0.04 mmol/L of fluorescently labeled ChromaTide Alexa Fluor 488-5-dUTPs (Invitrogen, Life Technologies, Grand Island, NY, USA) to allow for the visualization of amplified products on an automated DNA sequencer (ABI 3130 genetic analyzer, Applied Biosystems, Foster City, CA).

Genotyping and microsatellite characterization

To test the 37 diagnostic loci for amplification, we chose two individuals of each species. Polymerase chain reactions were prepared in 12.5 µL reactions and contained 2 µL of eluted genomic DNA, 0.4 µmol/L of each primer, 2.5 mmol/L MgCl2, 5X PCR buffer (Promega, Madison, WI, USA), 0.2 mmol/L of deoxyribonucleotides, and 1 unit of Taq DNA polymerase (Promega). Cycling conditions were as follows: initial denaturation at 94°C for 4 min, followed by 30 cycles of 94°C for 30 sec, 56°C–63°C for 45 sec, 72°C for 1 min, and a final extension step at 72°C for 5 min. PCR products were resolved on a 1% agarose gel. Of the 37 primers, 34 consistently amplified the target regions in both species and were used for an initial screening of 24 individuals from eight allopatric marshes (Table 1). PCR was repeated with the addition of 0.04 mmol/L of fluorescently labeled ChromaTide Alexa Fluor 488-5-dUTPs (Invitrogen, Life Technologies, Grand Island, NY, USA) to allow for the visualization of amplified products on an automated DNA sequencer (ABI 3130 genetic analyzer, Applied Biosystems, Foster City, CA).

To characterize the diagnostic potential of these 34 loci, we counted the number of alleles shared between the species across the 24 allopatric individuals (Table 2). We chose 12 loci with the fewest number of shared alleles and the most variation in the distribution of alleles to screen further as a panel of putatively diagnostic loci. These 12 chosen loci were screened in an additional 96 individuals (36 allopatric and 12 sympatric individuals of each species), using dye-labeled primers (HEX, FAM, or NED) in two multiplex PCRs. The 15 µL polymerase chain reactions contained 3 µL of eluted genomic DNA,
Table 2. Description of 34 putatively diagnostic markers screened in 12 *Ammodramus caudacutus* and 12 *A. nelsoni* individuals from allopatric populations. For each locus, columns contain the repeat motif, size in base pairs of the amplified fragments, primer sequences and annealing temperature, number of individuals genotyped (*n*), number of unique alleles (NA), observed (*H*<sub>o</sub>) and expected (*H*<sub>e</sub>) heterozygosities, number of alleles shared between the two species, the chromosome location and annotation of the amplified sequence in the Zebra Finch genome, and the GenBank accession number. Loci in bold are the 12 diagnostic markers selected after initial screening.

<table>
<thead>
<tr>
<th>Locus</th>
<th>Repeat motif</th>
<th>Size Range (bp)</th>
<th>Primer sequences (5′–3′)</th>
<th>TA (°C)</th>
<th>n</th>
<th>NA</th>
<th># of Shared Alleles</th>
<th>Chromosome number (Zebra Finch)</th>
<th>Zebra Finch Annotation</th>
<th>GenBank Accession Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ammo001</td>
<td>ACTC</td>
<td>138–158</td>
<td>F: CTTTCATCCATCCCTGTGCT</td>
<td>63</td>
<td>12</td>
<td>6</td>
<td>0.318</td>
<td>0.374</td>
<td>0</td>
<td>13</td>
</tr>
<tr>
<td>Ammo002</td>
<td>AAT</td>
<td>194–242</td>
<td>F: GGTGTTGCAAGCCAGGGATTAT</td>
<td>60</td>
<td>23</td>
<td>9</td>
<td>0.436</td>
<td>0.684</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Ammo003</td>
<td>GAT</td>
<td>139–157</td>
<td>F: TGTGGAAAGAAACAAAGCCAAAT</td>
<td>60</td>
<td>23</td>
<td>4</td>
<td>0.261</td>
<td>0.403</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Ammo005</td>
<td>AAAT</td>
<td>184–224</td>
<td>F: TGCCATTCCTCAGAAGCACCA</td>
<td>60</td>
<td>23</td>
<td>9</td>
<td>0.523</td>
<td>0.601</td>
<td>NA</td>
<td>2</td>
</tr>
<tr>
<td>Ammo006</td>
<td>ACAT</td>
<td>228–264</td>
<td>F: TCCAGCCCCTTTTGGAGAG</td>
<td>60</td>
<td>23</td>
<td>9</td>
<td>0.648</td>
<td>0.686</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Ammo008</td>
<td>GAT</td>
<td>238–250</td>
<td>F: AAGGCAGATGTCTCCTCCAC</td>
<td>60</td>
<td>24</td>
<td>3</td>
<td>0.208</td>
<td>0.309</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Ammo009</td>
<td>AAAC</td>
<td>242–270</td>
<td>F: TGGGTGACTAAGGTGGTCAG</td>
<td>56</td>
<td>22</td>
<td>7</td>
<td>0.545</td>
<td>0.636</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Ammo010</td>
<td>AAAC</td>
<td>230–254</td>
<td>F: AGCCCTATCGAGGTGAAAGA</td>
<td>60</td>
<td>22</td>
<td>7</td>
<td>0.275</td>
<td>0.597</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Ammo011</td>
<td>AAT</td>
<td>206–236</td>
<td>F: GGCACTCTGAGCATCAAT</td>
<td>60</td>
<td>23</td>
<td>9</td>
<td>0.424</td>
<td>0.503</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Ammo012</td>
<td>GTT</td>
<td>176–191</td>
<td>F: TTTGAAACTCTCTTCAATGG</td>
<td>56</td>
<td>23</td>
<td>3</td>
<td>0.000</td>
<td>0.243</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Ammo013</td>
<td>ACCT</td>
<td>252–272</td>
<td>F: GAGCAATGACAGGAGGAAC</td>
<td>60</td>
<td>23</td>
<td>6</td>
<td>0.254</td>
<td>0.466</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Ammo014</td>
<td>AAT</td>
<td>178–193</td>
<td>F: GAGAAACTCTTCAGCTC</td>
<td>60</td>
<td>20</td>
<td>4</td>
<td>0.400</td>
<td>0.350</td>
<td>2</td>
<td>1A</td>
</tr>
<tr>
<td>Ammo015</td>
<td>AGG</td>
<td>242–254</td>
<td>F: TCACCAAAAGGAGGTTT</td>
<td>60</td>
<td>22</td>
<td>4</td>
<td>0.083</td>
<td>0.179</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Ammo016</td>
<td>AAT</td>
<td>244–259</td>
<td>F: GCCAAACATGCTGGAGAAT</td>
<td>60</td>
<td>22</td>
<td>4</td>
<td>0.087</td>
<td>0.344</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Ammo017</td>
<td>AATC</td>
<td>112–116</td>
<td>F: GCTCTGGACTGCTGCAAAAT</td>
<td>60</td>
<td>11</td>
<td>2</td>
<td>0.045</td>
<td>0.118</td>
<td>0</td>
<td>1A</td>
</tr>
<tr>
<td>Ammo018</td>
<td>ATT</td>
<td>200–236</td>
<td>F: GGCAGAAGGCTGTGTA</td>
<td>60</td>
<td>24</td>
<td>8</td>
<td>0.542</td>
<td>0.569</td>
<td>2</td>
<td>No match</td>
</tr>
<tr>
<td>Ammo019</td>
<td>AT</td>
<td>150–171</td>
<td>F: CCTGCAAGAATGAGAGA</td>
<td>60</td>
<td>22</td>
<td>4</td>
<td>0.182</td>
<td>0.432</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Ammo020</td>
<td>AAT</td>
<td>265–292</td>
<td>F: TGTTGCTGAAAGAGT</td>
<td>60</td>
<td>23</td>
<td>9</td>
<td>0.867</td>
<td>0.739</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>Locus</td>
<td>Repeat motif</td>
<td>Size Range (bp)</td>
<td>Primer sequences (5′–3′)</td>
<td>TA (°C)</td>
<td>n</td>
<td>NA</td>
<td>H₀</td>
<td>Hₑ</td>
<td># of Shared Alleles</td>
<td>Chromosome number</td>
</tr>
<tr>
<td>---------</td>
<td>--------------</td>
<td>-----------------</td>
<td>--------------------------</td>
<td>--------</td>
<td>---</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>---------------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>Ammo021</td>
<td>CTGT</td>
<td>146–166</td>
<td>F: GGGTGCCACAGTCACATTTT</td>
<td>63</td>
<td>24</td>
<td>6</td>
<td>0.917</td>
<td>0.660</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>Ammo022</td>
<td>CCT</td>
<td>239–263</td>
<td>F: TGAGAGCTCTGAGCCCTTG</td>
<td>60</td>
<td>23</td>
<td>7</td>
<td>0.133</td>
<td>0.526</td>
<td>2</td>
<td>No match</td>
</tr>
<tr>
<td>Ammo023</td>
<td>ATT</td>
<td>213–261</td>
<td>F: GGAACAGAGAGGCTCCACCT</td>
<td>60</td>
<td>22</td>
<td>11</td>
<td>0.500</td>
<td>0.657</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Ammo024</td>
<td>AAAT</td>
<td>268–284</td>
<td>F: TTTCAAGGCTGCTGAGC</td>
<td>60</td>
<td>22</td>
<td>4</td>
<td>0.364</td>
<td>0.390</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Ammo025</td>
<td>ATT</td>
<td>183–225</td>
<td>F: GCCCTTCTTTCCTCCAGT</td>
<td>60</td>
<td>22</td>
<td>12</td>
<td>0.592</td>
<td>0.780</td>
<td>4</td>
<td>No match</td>
</tr>
<tr>
<td>Ammo027</td>
<td>AAAG</td>
<td>190–232</td>
<td>F: AAAAGGAAAGCTCAGTGA</td>
<td>63</td>
<td>22</td>
<td>9</td>
<td>0.592</td>
<td>0.713</td>
<td>3</td>
<td>SET and MYND domain-containing protein 3</td>
</tr>
<tr>
<td>Ammo028</td>
<td>ATCC</td>
<td>228–260</td>
<td>F: GCACTGCTCTCTACCTT</td>
<td>60</td>
<td>24</td>
<td>9</td>
<td>0.750</td>
<td>0.793</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>Ammo029</td>
<td>AATG</td>
<td>116–148</td>
<td>F: TGAAACAAGGATTTTGAAAGA</td>
<td>63</td>
<td>20</td>
<td>6</td>
<td>0.450</td>
<td>0.675</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Ammo030</td>
<td>ATT</td>
<td>243–270</td>
<td>F: GCCATGAGCTCTCCAAT</td>
<td>60</td>
<td>20</td>
<td>7</td>
<td>0.438</td>
<td>0.414</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Ammo031</td>
<td>ACT</td>
<td>244–259</td>
<td>F: AAAAGTTAACCGCTAGCC</td>
<td>60</td>
<td>23</td>
<td>4</td>
<td>0.174</td>
<td>0.349</td>
<td>2</td>
<td>Z</td>
</tr>
<tr>
<td>Ammo032</td>
<td>GAT</td>
<td>137–158</td>
<td>F: AAAACCCTAGGGGAAGGCA</td>
<td>63</td>
<td>23</td>
<td>6</td>
<td>0.871</td>
<td>0.637</td>
<td>2</td>
<td>Z</td>
</tr>
<tr>
<td>Ammo033</td>
<td>ATCC</td>
<td>262–278</td>
<td>F: TACCAGAAAGATGCCACACAA</td>
<td>60</td>
<td>22</td>
<td>5</td>
<td>0.545</td>
<td>0.630</td>
<td>4</td>
<td>No match</td>
</tr>
<tr>
<td>Ammo034</td>
<td>AAT</td>
<td>129–165</td>
<td>F: AGGGAAGATCTCGAGCCCTCG</td>
<td>60</td>
<td>21</td>
<td>9</td>
<td>0.864</td>
<td>0.765</td>
<td>4</td>
<td>No match</td>
</tr>
<tr>
<td>Ammo035</td>
<td>ATCT</td>
<td>161–229</td>
<td>F: ACACCGCAAGCTCCAGCTC</td>
<td>60</td>
<td>24</td>
<td>11</td>
<td>0.590</td>
<td>0.743</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>Ammo036</td>
<td>CTT</td>
<td>189–210</td>
<td>F: TCAAGAGCCCTGGCGTTCC</td>
<td>60</td>
<td>20</td>
<td>5</td>
<td>0.527</td>
<td>0.528</td>
<td>2</td>
<td>1A</td>
</tr>
<tr>
<td>Ammo037</td>
<td>ATCT</td>
<td>268–312</td>
<td>F: CATGCTGCTCTGCGACCTCTCA</td>
<td>63</td>
<td>23</td>
<td>10</td>
<td>0.739</td>
<td>0.646</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

1Did not amplify in A. caudacutus.
2Out of Hardy Weinberg Equilibrium (HWE) in A. nelsoni at P = 0.05.
3Out of HWE in A. nelsoni with Bonferroni correction.
4Out of HWE in both species at P = 0.05.
5Out of HWE in A. caudacutus at P = 0.05.
6Out of HWE in A. caudacutus with Bonferonni correction.
0.1–0.3 μmol/L of each dye-labeled primer, 2.0 mmol/L MgCl₂, 5X PCR buffer (Promega), 0.1 mmol/L of deoxyribonucleotides, and 1 unit of Taq DNA polymerase (Promega). We used the same cycling conditions described above with a 60°C annealing temperature for all loci. Amplified products were again electrophoresed on an ABI 3130 automated DNA sequencer, and individual genotypes were scored manually using PEAKSCANNER software (ABI).

For the 12 diagnostic loci, the number of private alleles, allele frequencies, and estimates of expected and observed heterozygosities were calculated for allopatric individuals in GENALEX, version 6.41 (Peakall and Smouse 2006). The proportion of shared alleles was calculated for each locus as the number of alleles shared between allopatric *A. nelsoni* and *A. caudacutus* divided by the total number of alleles. The frequency of the most common allele in each species was calculated in GENALEX. We performed selection tests for the 12 loci using an *F*ₜₛₜ outlier approach ( Beaumont and Nichols 1996) in LOSITAN (Antao et al. 2008). Tests of Hardy–Weinberg equilibrium and linkage equilibrium were conducted in GENEPOL, version 4.2 (Raymond and Rouset 1995). Significance was assessed with a Bonferroni correction for multiple tests. Locus-specific *F*ₜₛₜ values were also calculated for all pairwise combinations of allopatric and sympatric sparrows in GENEPOL. We used a Bayesian clustering method implemented in the program STRUCTURE v. 2.3.4 ( Pritchard et al. 2000) to assess how membership proportions differed between allopatric and sympatric populations of both species. We ran ten replications with *K* = 2, using the admixture model with correlated allele frequencies and a 100,000 burn-in followed by 100,000 iterations.

**Power assessment of diagnostic marker panel**

We assessed the power of the panel of 12 diagnostic markers by evaluating the accuracy of each locus in assigning known individuals to hybrid classes. We simulated 100 genotypes for each of six genotypic classes (pure *A. nelsoni*, pure *A. caudacutus*, backcrossed *A. nelsoni*, backcrossed *A. caudacutus*, F₁ hybrids, and F₂ hybrids) using the program HYBRIDLAB 1.0 (Nielsen et al. 2006). Simulated individuals were analyzed using the program NEWHYBRIDS 1.1 BETA (Anderson and Thompson 2002); we used the *z* and *s* option to identify the 36 pure individuals of each species as known reference individuals. We ran NEWHYBRIDS using the default options with 200,000 sweeps and a 200,000 burn-in. We calculated mean posterior probabilities of the individuals assigned to each category and the percentage of correctly assigned individuals. Individuals were considered correctly assigned when their true category was the category with the highest posterior probability.

**Results**

**Marker development and characterization**

Sizes of the repeat regions for the 34 markers ranged from 112 to 284 bp, and loci were variably polymorphic with 2–12 alleles (Table 2; see Appendix A2 for allele frequency data). Mean observed and expected heterozygosities ranged from 0.133 to 0.917. Eighteen loci showed significant deviations from Hardy–Weinberg in one or both species at *P* < 0.05, and 6 loci showed deviations in one or both species after Bonferroni correction (*P* < 0.0007; Table 2). These deviations are not unexpected and most likely result from a Wahlund effect (Wahlund 1928), given that samples for each species were collected from a diversity of geographic locations, potentially comprised of distinct populations. The number of shared alleles between species ranged from 0 to 6 across the 34 loci. Across the panel of 12 diagnostic loci, no pairs showed significant deviations from linkage equilibrium. Two loci (*Ammd0012* and *Ammd0015*) were candidates for positive selection.

**Resolution and power of the diagnostic marker panel**

The proportion of shared alleles between allopatric *A. caudacutus* and *A. nelsoni* at the 12 diagnostic loci ranged from 0.11 to 0.95, with the frequency of most common alleles as high as 1.0 in *A. caudacutus* and 0.984 in *A. nelsoni* (Table 3). The number of private alleles ranged from 1 to 12 among allopatric populations. Locus-specific *F*ₜₛₜ values between allopatric *A. nelsoni* and *A. caudacutus* ranged from 0.21 to 0.81 with a global *F*ₜₛₜ of 0.46 (Table 4). Differentiation between sympatric *A. nelsoni* and allopatric *A. caudacutus* was similar to that of the two allopatric populations; however, differentiation between allopatric *A. nelsoni* and sympatric *A. caudacutus* and between sympatric populations of each species was slightly lower (Fig. 2). *F*ₜₛₜ values for within-species comparisons were much lower (0.004 to 0.027 overall; Table 4; Fig. 2). STRUCTURE Q values (proportion of the genome attributed to the parental species, with 1 being pure *caudacutus* and 0 pure *nelsoni*) for allopatric individuals were above or below the pure species cutoffs of 0.9 and 0.1, respectively. Introgression was apparent in sympatric individuals, however, with slightly lower Q values, including some above/below the pure species cutoffs, especially for sympatric *A. caudacutus* (Fig. 3).
Based on the 12 diagnostic microsatellite markers, NEWHYBRIDS assigned 92% of all the simulated individuals to their true category. Assignment accuracies varied for the categories, ranging from 76% (F2) to 100% (pure A. caudacutus), with posterior probabilities for the correctly assigned categories ranged from 0.75 (F2) to 0.991 (A. caudacutus; Table 5). Pure individuals had the highest percentage of correct assignments with 98% (A. nelsoni) and 100% (A. caudacutus) of individuals correctly assigned with posterior probabilities of 0.947 and 0.991, respectively. F1 individuals were also assigned with high accuracy (97% and posterior probability of 0.936). F2 individuals were the most difficult to assign, especially with respect to distinguishing them from backcrossed individuals, with 76% of individuals correctly assigned with a mean posterior probability of 0.75. For backcrossed individuals, 91% were assigned at nearly identical mean posterior probabilities of 0.844 and 0.843. The majority of misassignments were between backcrossed and F2 individuals. There were no instances where backcrossed A. nelsoni were assigned as backcrossed A. caudacutus and vice versa.

**Discussion**

While current sequencing technologies afford the potential for generating tens of thousands of genomewide markers for population genomics research (Davey et al. 2011), not all research and conservation applications will require genomewide data (Allendorf et al. 2010). For such applications, including research questions focused on discerning processes for closely related individuals – such as dispersal, kinship, population structure, and admixture – an informative panel of microsatellite markers will remain a valuable tool (Guichoux et al. 2011; Wei et al. 2014; Vukosavljev et al. 2015). In our case study of an avian hybrid zone, we highlight the utility of a carefully selected, high-resolution panel of microsatellite markers for discriminating genotypic classes of pure and admixed individuals. Our strategy for diagnostic marker discovery via in silico screening for microsatellite repeat differences in two species’ genomes eliminates the laborious process of manually screening markers in the laboratory. As such, this efficient and highly effective approach should prove useful for other studies requiring diagnostic microsatellite markers for closely related species.

From whole-genome sequence data, we identified 34 polymorphic and diagnostic or partially diagnostic microsatellite markers that amplified in both A. caudacutus and A. nelsoni. We developed a panel of the 12 loci with the fewest shared alleles between species. All markers in this diagnostic panel amplified consistently using the same routine PCR conditions, making them highly conducive
for multiplexing. We demonstrated the power of these loci for the identification of pure and admixed individuals in this avian hybrid zone.

After screening the 12 diagnostic markers on 96 sparrows from allopatric and sympatric sites, we found them to be highly informative for species discrimination. This panel of loci had high resolution for classifying pure and admixed individuals into their genotypic classes. The markers were most powerful for distinguishing among F1, backcrossed, and parental groups (with 91–100% accuracy), while F2s were difficult to distinguish from either F1 or backcrossed groups (76% accuracy). For all loci, the most common allele differed between the two species; this allele was typically rare (<0.05%) in the other species. Only one locus, Ammo030, showed a fixed allele in either species—with a single allele in A. caudacutus and three private alleles in A. nelsoni. An additional four markers had an allele with >90% frequency in one of the two species. While most of the markers exhibited a relatively large portion of shared alleles between species (0.11–0.98), allele frequency distributions differed strongly between the species, and all loci had at least one private allele. Locus-specific $F_{ST}$s indicated strong divergence (ranging from 0.2144 to 0.819, with overall $F_{ST} = 0.4667$)

Table 4. Locus-specific and overall $F_{ST}$ values for all pairwise comparisons of allopatric and sympatric Ammodramus caudacutus and A. nelsoni using the panel of 12 diagnostic microsatellite markers developed in this study.

<table>
<thead>
<tr>
<th>Locus</th>
<th>Allopatric nelsoni/Allopatric caudacutus</th>
<th>Allopatric nelsoni/Sympatric caudacutus</th>
<th>Allopatric caudacutus/Sympatric nelsoni</th>
<th>Sympatric nelsoni/Sympatric caudacutus</th>
<th>Allopatric nelsoni/Sympatric nelsoni</th>
<th>Allopatric caudacutus/Sympatric caudacutus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ammo001</td>
<td>0.3033</td>
<td>0.2592</td>
<td>0.4037</td>
<td>0.3707</td>
<td>-0.0703</td>
<td>-0.0027</td>
</tr>
<tr>
<td>Ammo003</td>
<td>0.4019</td>
<td>0.3946</td>
<td>0.3635</td>
<td>0.3477</td>
<td>-0.0312</td>
<td>-0.024</td>
</tr>
<tr>
<td>Ammo006</td>
<td>0.281</td>
<td>0.2237</td>
<td>0.2524</td>
<td>0.1813</td>
<td>-0.0234</td>
<td>-0.0182</td>
</tr>
<tr>
<td>Ammo008</td>
<td>0.6245</td>
<td>0.5498</td>
<td>0.6262</td>
<td>0.5204</td>
<td>0.0078</td>
<td>0.0326</td>
</tr>
<tr>
<td>Ammo012</td>
<td>0.819</td>
<td>0.7123</td>
<td>0.6915</td>
<td>0.4454</td>
<td>0.0632</td>
<td>0.1041</td>
</tr>
<tr>
<td>Ammo015</td>
<td>0.8073</td>
<td>0.849</td>
<td>0.7748</td>
<td>0.8202</td>
<td>-0.0387</td>
<td>0.0111</td>
</tr>
<tr>
<td>Ammo016</td>
<td>0.5586</td>
<td>0.6217</td>
<td>0.4856</td>
<td>0.5076</td>
<td>-0.0183</td>
<td>0.0202</td>
</tr>
<tr>
<td>Ammo017</td>
<td>0.3968</td>
<td>0.4448</td>
<td>0.2555</td>
<td>0.2814</td>
<td>0.0056</td>
<td>-0.0213</td>
</tr>
<tr>
<td>Ammo023</td>
<td>0.2629</td>
<td>0.2185</td>
<td>0.1842</td>
<td>0.1371</td>
<td>0.0169</td>
<td>-0.0135</td>
</tr>
<tr>
<td>Ammo027</td>
<td>0.2144</td>
<td>0.2421</td>
<td>0.3734</td>
<td>0.489</td>
<td>0.1711</td>
<td>-0.0027</td>
</tr>
<tr>
<td>Ammo030</td>
<td>0.5198</td>
<td>0.3685</td>
<td>0.694</td>
<td>0.4561</td>
<td>0.0006</td>
<td>0.064</td>
</tr>
<tr>
<td>Ammo036</td>
<td>0.4352</td>
<td>0.465</td>
<td>0.4172</td>
<td>0.4614</td>
<td>-0.0241</td>
<td>0.0166</td>
</tr>
<tr>
<td>Overall</td>
<td>0.4667</td>
<td>0.4282</td>
<td>0.4567</td>
<td>0.4137</td>
<td>0.0272</td>
<td>0.004</td>
</tr>
</tbody>
</table>

Figure 2. Distribution of $F_{ST}$ values across the 12 diagnostic markers for pairwise comparisons of allopatric and sympatric Ammodramus caudacutus and A. nelsoni.
between allopatric populations of each species. In comparison, anonymous neutral loci yielded a between-species $F_{ST}$ of 0.15 (Shriver et al. 2005).

The lack of microsatellite loci with fixed differences between the two species is notable, given our whole-genome approach. By comparing all of the microsatellite repeats identified from whole-genome shotgun sequencing, we only found 79 loci to differ in repeat numbers between the two genomes (and of these, only 42 met our criteria of differing by four or more repeats). Our resulting panel of 12 diagnostic loci therefore likely represents the largest microsatellite differences across the genomes of these two species. The overall similarity in microsatellite repeats between the two species exemplifies their close relationship as recently diverged sister species and suggests that high genetic similarity is characteristic not only at the mitochondrial level (Rising and Avise 1993; Walsh et al. 2011), but also potentially on a genomewide level.

This finding gives insight into genomic similarity of hybridizing taxa and highlights the challenges of identifying diagnostic markers for recently diverged species, as well as the utility of whole-genome sequencing in high-resolution marker development.

The elevated divergence of the diagnostic panel relative to neutral loci previously used in this system (Shriver et al. 2005; Walsh et al. 2012) suggests that these loci may be under selection (Storz 2005; Strasburg et al. 2012). Selection tests identified two of the markers to be under selection in this dataset; however, further research with more targeted sampling schemes may identify additional selected loci. This is supported by the fact that 10 of the 34 (6 of 12 diagnostic) loci aligned with an annotated protein-coding region of the Zebra Finch genome (Table 2). These markers may be associated with a region of the genome with a functional role that diverges between the two species; accordingly, these may be
important portions of the genome with respect to speciation. As allelic changes induced by natural selection occur faster than those due to neutral processes (Nei 1987), high-resolution gene-associated markers are more powerful than neutral markers for applications that require assigning individuals to distinct population or species groupings (Nielsen et al. 2009, 2012).

The pattern of between-species divergence that we found using the 12 diagnostic markers in this study was fairly consistent across sympatric and allopatric populations. While $F_{ST}$s were highest for allopatric comparisons of the species, they were only slightly lower for comparisons that included sympatric populations, suggesting divergence at these loci is maintained in the face of interspecific gene flow in the hybrid zone (Walsh et al. 2011, 2015). These markers therefore appear to be associated with gene regions that do not introgress freely between the two species. For within-species comparisons, the $F_{ST}$s are slightly lower within A. caudacutus than within A. nelsoni, supporting the hypothesis that introgression is biased in the direction of the A. caudacutus genome (Shriver et al. 2005; Walsh et al. 2011).

The low within-species divergence we found in this study is an expected outcome, especially for diagnostic markers. $F_{ST}$ among A. caudacutus in this study is similar to that previously reported by Walsh et al. (2012) using anonymous neutral loci. Despite high levels of gene flow, Walsh et al. (2012) found evidence for fine-scale population structure within A. caudacutus. The sampling scheme in the current study, however, was not designed for evaluating within-species population structure, as pooling across many geographically separate sympatric or allopatric marshes likely masks some of the underlying population differentiation. Nonetheless, the higher within-species $F_{ST}$ found in A. nelsoni compared to A. caudacutus suggests that a finer scale population genetic structure may be characteristic of the former. More pronounced population structure in A. nelsoni relative to A. caudacutus is consistent with differences in the species’ distributions and demography – nelsoni typically occur in smaller numbers in small marshes that tend to be more spatially disconnected than the larger, more continuous coastal marshes typically occupied by caudacutus (J. Walsh and A. Kovach, pers. obs.). These are the first population genetic data collected on A. nelsoni; future research with a more robust sampling scheme is warranted to characterize population genetic structure in this species. The preliminary data in this study suggest that these markers will be useful for such within-species population comparisons.

In conclusion, our comparative, whole-genome approach has proven useful for identifying high-resolution diagnostic markers in sister species with high genetic similarity. This approach is superior to anonymous marker development, not only because it enables pinpointing species-specific differences, but also because it links the markers to large contigs that can be mapped to genomic regions. The markers identified in this study will aid future research that requires distinguishing pure and admixed individuals in the A. caudacutus – nelsoni hybrid zone, as doing so from morphology alone is unreliable (Walsh et al. 2011, 2015). A hybrid index based on 12 diagnostic microsatellite markers provides an inexpensive and simple genetic assay. This diagnostic assay for hybrid identification will prove valuable in efforts that seek to track shifts in species distributions, which is of particular relevance to the conservation of threatened A. caudacutus populations (Shriver et al. 2005; Walsh et al. 2011). The diagnostic marker panel will also be useful for studies of evolutionary ecology, such as providing insight into the rates and direction of introgression and estimates of the width and center of the hybrid zone (Barton and Gale 1993). Our marker development approach is easily transferable to other studies, and we provide our PERL script for comparing repeat sequences of two genomes as an appendix.

Acknowledgments

Funding was provided by the United States Fish & Wildlife Service, Region 5, Division of Natural Resources, National Wildlife Refuge System, and the New Hampshire Agricultural Experiment Station. We thank C. Elphick, T. Hodgman, E. King, O. Lane, G. Mittlehauser, N. Pau, K. O’Brien, B. Olsen, K. Ruskin, and G. Shriver for logistical support and sample collection assistance. The Maine Department of Inland Fisheries and Wildlife, The Nature Conservancy, and Waquoit Bay National Estuarine Research Reserve generously allowed sampling on protected marshes. We thank J. Anderson for technical assistance with fragment analysis and K. Okamoto for genome assembly help. T. Kristensen, B. Benvenuti, K. Papanastassiou, and C. Mills and two anonymous reviewers provided helpful comments on a previous draft of this manuscript. This is Scientific Contribution Number 2601 of the New Hampshire Agricultural Experiment Station. This work was supported by the USDA National Institute of Food and Agriculture McIntire-Stennis Project Number 225575.

Conflict of Interest

None declared.

Data Accessibility

Sequences of the 34 microsatellite loci are deposited in GenBank (Accession numbers in Table 2). Raw Illumina
reads are deposited in the Short Read Archive (SRS893469 and SRS897222). Our PERL script for screen-
ing repeat sequences across the two genomes is provided as Appendix A1, and allele frequencies for the panel of 12 diagnostic loci are in Appendix A2. Genotypes of the 96 individuals, simulated genotypes used to evaluate the marker panel, and the assembled N50 contigs of A. nelsoni and A. caudapectus are deposited in Dryad (http://dx.doi.org/10.5061/dryad.f8m84).

References


Appendix A1: Perl script for screening diagnostic microsatellite sequences between two genomes

This script takes the microsatellite loci identified from one dataset with MSATCOMMANDER and screens the repeat sequences against a fasta file of sequences (in this case from a second genome) to pull out those loci common to both datasets that differ in repeat number between the two genomes. The resulting output contains information on the number of repeats for each locus in each dataset. This output can then be filtered in a spreadsheet to thresholds according to the study design (e.g., loci that differ by >4 repeats).


#!/usr/bin/perl

#Title: Microsat Scanner
#Version: 1.5
#Author: Jordan Ramsdell
#Use: Allows the user to import microsat pattern data from MSATCOMMANDER

# And compare these patterns to data found in a fasta file.

#Exports results to csv format.

#Documentation: This script uses the output from Msat commander to take the
#flanking nucleotides from either side of the microsatellites. This is then used
#to find a potential match in the target fasta file, under
#the assumption that
#the flanking sequences are identical.
#The reverse-complement of the flanks are also searched, in case the
#microsatellite is found on the 3’ to 5’ end.

#USAGE: perl MicroSatScan.pl -msat <input.msat> -target <input.fasta> -flank_length <int, default 20> -out <output.csv>

#Options:
#-msat <input.msat> Name of the msat commander file to use with this program.
#-source <input.fasta> Name of the fasta file that msat was run on, used to extract flanking sequences.
#-target <input.fasta> Name of the target fasta file to scan through with flanking sequences.
#-flank_length <integer> Specifies the number of flanking nucleotides used in the regex search.
#-out <output.csv> Name of the comma delimited file to create after finishing scan

The output is in a comma-separated format with the following columns:

####[Clone]: The contigs from the original fasta file that msat commander was run on.
####[Startbp]: The starting location of the microsatellite within the source contig.
####[Repeat]: The number of times the microsatellite repeats within the source contig.
####[Endbp]: The ending location of the microsatellite within the source contig.
####[Type]: The type of repeat
####[Comments]: Whether or not this repeat was found on the forward or reverse strand in the source.
####[MatchContig]: The target contig where the nucleotide repeat matched (based on flanking regions).
####[MatchRepeats]: How many times the microsatellite repeats within the matching contig.
####[Difference]: The difference in the amount of times the microsatellite repeats in the source contig compared to the target contig.
####[MatchSeq]: Displays the nucleotide sequence found between the flanking regions within the matching target contig.
####[MatchStart]: The starting location of the microsatellite within the matching target contig
####[MatchEnd]: The ending location of the microsatellite within the matching target contig
####[MatchLength]: The total length of the matching sequence within the target contig
####[Type]: Type 0 indicates that there was a change in expected length, likely due to the insertion or deletion of a single bp.

Type 1 (what we’re probably more interested in) means that this sequence differs only in the number of microsatellites it has.

```perl
use Getopt::Long;
use Pod::Usage;
use strict();
use warnings();

my $fasta;
my $source;
my $target;
my $debug;
my $input;
my $output;

GetOptions("msat=s" => $msat_file,
"source=s" => $source_file,
"target=s" => $target_file,
"flank_length=i" => $flank_number,
"out=s" => $output_name)
or die(pod2usage(1));

#Open Msat file
{
  open(MSAT, $msat_file) || die "Couldn't open -msat file: $msat_file";
  $msat = <MSAT>
  close MSAT;
}

print "Successfully imported MSAT data.\nImporting Source FASTA file.\n";

#Open Source Fasta File
{
  local $/ = undef;
  open(SOUR, $source_file) || die "Couldn't open -source file: $source_file";
  $source = <SOUR>
  close SOUR;
}

print "Successfully imported Source data.\nImporting Target FASTA file.\n";

#Open Target Fasta File
{
  local $/ = undef;
  open(TARG, $target_file) || die "Couldn't open -target file: $target_file";
  $target = <TARG>
  close TARG;
}

print "Successfully imported Target data.\nFiltering data.\n";
```
# Clean up data and prepare header
$target =~ s/\n/\r/; $msat =~ s/\r\n\r/\n/g;
my @linesplit = split("\r", $msat);
Soutput = shift (@linesplit);
Soutput =~ ".MatchContig,MatchRepeats,Difference, MatchSequence,MatchStart,MatchEnd,MatchLength, Type\r";
print "Data filtering complete. Beginning search.\n\n";
# Begin searching for matches in target sequence
while ($#linesplit != -1)
{
    my $element = shift @linesplit;
    my $reference = $element;
    my $contig = shift @split;
    # Find Contig index match in source
    my $contig_index = index ($source, $contig."\n");
    if ($contig_index != -1) & & ($split[0] ne "No repeats found" or "\n")
    {
        my $stopt = index ($source,">", (contig_index + 1));
        # Extract the sequence (if it's the last contig, we just extract to the end)
        my $sequence = substr ($source, contig_index + length ($contig));            my $startbp = shift @split;
        my $repeat = shift @split;
        my $endbp = shift @split;
        $sequence =~ s/\n/\r/g;
        my $extract = substr ($sequence, $startbp - 1, ($endbp - $startbp));
        my $flank_left = substr ($sequence, $startbp - 1 - $flank_number,
                                $flank_number);
        my $flank_right = substr ($sequence, $endbp - 1, $flank_number);
        # Extract repeat nucleotides and number of times they occur.
        $repeat =~ m/(?:[\ATCGN])/g;
        my $repeat_source = \$1;
        my $times_source = \$2;
        # Now do a regex lookup on the target fasta file
        my $regex = "(?:\$flank_left)([\ATCGN]\{1,200\})(?: \$flank_right)";
        print "Searching Contig:".$contig."nPattern:".$repeat."n";
        my $reverse = 0;
        # Is this match different? If so, send to output. Otherwise, throw out match.
        if (((($target =~ m/(:\$flank_left)/g) ||
                 (($target =~ m/(:\$flank_right)/g) & & ( ++$reverse)) ))
        {
            my $snippet = substr($target, (pos($target) - 1000), 2000);
            my $repeat_start = \$1;
            my $repeat_end = \$2;
            my $repeat_start2 = reverse($repeat_start);
            my $repeat_end2 = reverse($repeat_end);
            my $repeat_right = reverse($repeat_right);
            my $repeat_left = reverse($repeat_left);
            if (((($target =~ m/(:\$flank_left)/g) ||
                     (($target =~ m/(:\$flank_right)/g) & & ( ++$reverse)) ))
            {
                my $matched_repeat = \$1;
                my $reverse_repeat = reverse ($matched_repeat);
                $repeat_repeat =~ tr/ATCG/TAGC/;
                if (((($matched_repeat eq $extract) ||
                           ($matched_repeat ne $extract) & & ($reverse_repeat eq $extract)) )
                {
                    print "Located difference in pattern: \n---------------------\n";
                    # Something's different in this sequence.
                    $output = "$contig","$startbp","$repeat","$endbp","$shift (@split)","$shift(@split);
                    my $pos = pos($target);
                    print "Interior Sequence:".$matched_repeat."\n";
                    # Figure out which contig this is on.
                    my $contig_start = rindex($target, ">");
                    my $contig_end = pos($target);
                    print "CONTIG START: $pos\n\n";
                    pos($target) = $contig_start;
                    $target =~ m/[:\ATCGN]/g;
                    my $contig_end = pos($target);
                    # Extract name of the contig that had the match in the target fasta file
                    my $contig_name = substr($target, ($contig_start + 1),
                                              $contig_end - $contig_start - 2);
                    $output =~ ".$contig_name;\nRepeat Start: $repeat_start;\nRepeat End: $repeat_end;\nRepeat Start End: $repeat_start2;\nRepeat End: $repeat_end2;\nRepeat Right: $repeat_right;\nRepeat Left: $repeat_left;
                }
            }
        }
    }
}
#Figure out where the repeat differs.
my $times_target = 0;
if ($reverse == 1) {
    $repeat_source = reverse($repeat_source);
    $repeat_source =~ tr/ATCG/TAGC/;
}

$repeat_source += while $matched_repeat =~ /(\$repeat_source)/g;
$output .= ",$times_target;

$repeat_dif = $times_target - $times_source;
$output .= ",$repeat_dif," . length($matched_repeat);

#Did a bp change, or was a whole repeat inserted/deleted?
if (length($matched_repeat) != (length($repeat_source) * $times_target)) {
    $output .= ",0\n";
} else {
    $output .= ",1\n";
}

print "Repeated: $times_target times.\nDifference from Original: $repeat_dif.\n--------------------\n";

else {
    $output .= join("", @reference)."No Match\n";
}

print "\nRemaining Lookups: $#linesplit.\n\n\n\n\nSearch complete.\nExporting data to: $output_name;\nopen FILE, ">$output_name or die $!
print FILE $output;
close FILE;

=head1 NAME
=sample - Using GetOpt::Long and Pod::Usage
=head1 SYNSOPSIS
=perl MicroSatScan.pl -msat <input.msat> -source <input.fasta> -target <input.fasta> -flank_length <int, default 20 > -out <output.csv>

Options:
=msat <input.msat> Name of the msat commander file to use with this program.
=source <input.fasta> Name of the fasta file that msat was run on, used to extract flanking sequences.
=target <input.fasta> Name of the target fasta file to scan through with flanking sequences.
=flank_length <integer> Specifies the number of flanking nucleotides used in the regex search. Default is 20.
=out <output.csv> Name of the comma delimited file to create after finishing scan.
=cut

Appendix A2: Allele frequencies for the 12 diagnostic loci screened in 36 Ammodramus caudacutus and A. nelsoni sampled from outside of the known overlap zone. Diagnostic alleles (found only in one species) are indicated in bold, and the most common allele in each species is indicated with an asterisk.

<table>
<thead>
<tr>
<th>Locus</th>
<th>Allele/n</th>
<th>Allopatric SALS</th>
<th>Allopatric NESP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ammo001</td>
<td>118</td>
<td>0.556</td>
<td>0.042</td>
</tr>
<tr>
<td></td>
<td>122</td>
<td>0.389</td>
<td>0.042</td>
</tr>
<tr>
<td></td>
<td>126</td>
<td>0.056</td>
<td>0.028</td>
</tr>
<tr>
<td></td>
<td>134</td>
<td>0.000</td>
<td>0.056</td>
</tr>
<tr>
<td></td>
<td>138</td>
<td>0.000</td>
<td>0.361</td>
</tr>
<tr>
<td></td>
<td>142</td>
<td>0.000</td>
<td>0.250</td>
</tr>
<tr>
<td></td>
<td>146</td>
<td>0.000</td>
<td>0.097</td>
</tr>
<tr>
<td></td>
<td>150</td>
<td>0.000</td>
<td>0.097</td>
</tr>
<tr>
<td></td>
<td>154</td>
<td>0.000</td>
<td>0.028</td>
</tr>
<tr>
<td></td>
<td>232</td>
<td>0.000</td>
<td>0.014</td>
</tr>
<tr>
<td></td>
<td>236</td>
<td>0.014</td>
<td>0.402*</td>
</tr>
<tr>
<td></td>
<td>240</td>
<td>0.000</td>
<td>0.083</td>
</tr>
<tr>
<td></td>
<td>244</td>
<td>0.000</td>
<td>0.403*</td>
</tr>
<tr>
<td></td>
<td>248</td>
<td>0.000</td>
<td>0.056</td>
</tr>
<tr>
<td></td>
<td>252</td>
<td>0.278</td>
<td>0.028</td>
</tr>
<tr>
<td></td>
<td>256</td>
<td>0.250</td>
<td>0.014</td>
</tr>
<tr>
<td></td>
<td>260</td>
<td>0.347*</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>264</td>
<td>0.111</td>
<td>0.000</td>
</tr>
<tr>
<td>Ammo006</td>
<td>112</td>
<td>0.014</td>
<td>0.083</td>
</tr>
<tr>
<td></td>
<td>116</td>
<td>0.056</td>
<td>0.736*</td>
</tr>
<tr>
<td></td>
<td>120</td>
<td>0.028</td>
<td>0.014</td>
</tr>
<tr>
<td></td>
<td>124</td>
<td>0.583*</td>
<td>0.153</td>
</tr>
<tr>
<td></td>
<td>128</td>
<td>0.222</td>
<td>0.014</td>
</tr>
<tr>
<td></td>
<td>132</td>
<td>0.056</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>136</td>
<td>0.042</td>
<td>0.000</td>
</tr>
<tr>
<td>Ammo008</td>
<td>238</td>
<td>0.014</td>
<td>0.333</td>
</tr>
<tr>
<td></td>
<td>244</td>
<td>0.014</td>
<td>0.569*</td>
</tr>
<tr>
<td></td>
<td>247</td>
<td>0.028</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>250</td>
<td>0.944*</td>
<td>0.097</td>
</tr>
<tr>
<td>Ammo027</td>
<td>188</td>
<td>0.097</td>
<td>0.500*</td>
</tr>
<tr>
<td></td>
<td>192</td>
<td>0.000</td>
<td>0.042</td>
</tr>
<tr>
<td></td>
<td>196</td>
<td>0.000</td>
<td>0.292</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>0.000</td>
<td>0.125</td>
</tr>
<tr>
<td></td>
<td>208</td>
<td>0.028</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>212</td>
<td>0.306*</td>
<td>0.014</td>
</tr>
<tr>
<td></td>
<td>216</td>
<td>0.139</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>220</td>
<td>0.153</td>
<td>0.014</td>
</tr>
<tr>
<td></td>
<td>224</td>
<td>0.181</td>
<td>0.014</td>
</tr>
<tr>
<td></td>
<td>228</td>
<td>0.097</td>
<td>0.000</td>
</tr>
<tr>
<td>Ammo015</td>
<td>241</td>
<td>0.000</td>
<td>0.972*</td>
</tr>
<tr>
<td></td>
<td>244</td>
<td>0.014</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>247</td>
<td>0.069</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>250</td>
<td>0.083</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>253</td>
<td>0.819*</td>
<td>0.028</td>
</tr>
<tr>
<td></td>
<td>256</td>
<td>0.014</td>
<td>0.000</td>
</tr>
<tr>
<td>Ammo003</td>
<td>139</td>
<td>0.329</td>
<td>0.014</td>
</tr>
<tr>
<td></td>
<td>142</td>
<td>0.014</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>151</td>
<td>0.471*</td>
<td>0.056</td>
</tr>
<tr>
<td></td>
<td>154</td>
<td>0.186</td>
<td>0.806*</td>
</tr>
<tr>
<td></td>
<td>157</td>
<td>0.000</td>
<td>0.125</td>
</tr>
<tr>
<td>Locus</td>
<td>Allele/n</td>
<td>Allopatric SALS</td>
<td>Allopatric NESP</td>
</tr>
<tr>
<td>----------</td>
<td>----------</td>
<td>-----------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>Ammo023</td>
<td>211</td>
<td>0.100</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>214</td>
<td>0.571*</td>
<td>0.014</td>
</tr>
<tr>
<td></td>
<td>217</td>
<td>0.114</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>220</td>
<td>0.143</td>
<td>0.014</td>
</tr>
<tr>
<td></td>
<td>223</td>
<td>0.014</td>
<td>0.403*</td>
</tr>
<tr>
<td></td>
<td>226</td>
<td>0.057</td>
<td>0.083</td>
</tr>
<tr>
<td></td>
<td>229</td>
<td>0.000</td>
<td>0.014</td>
</tr>
<tr>
<td></td>
<td>232</td>
<td>0.000</td>
<td>0.125</td>
</tr>
<tr>
<td></td>
<td>235</td>
<td>0.000</td>
<td>0.028</td>
</tr>
<tr>
<td></td>
<td>238</td>
<td>0.000</td>
<td>0.056</td>
</tr>
<tr>
<td></td>
<td>241</td>
<td>0.000</td>
<td>0.014</td>
</tr>
<tr>
<td></td>
<td>244</td>
<td>0.000</td>
<td>0.042</td>
</tr>
<tr>
<td></td>
<td>247</td>
<td>0.000</td>
<td>0.028</td>
</tr>
<tr>
<td></td>
<td>250</td>
<td>0.000</td>
<td>0.042</td>
</tr>
<tr>
<td></td>
<td>253</td>
<td>0.000</td>
<td>0.042</td>
</tr>
<tr>
<td></td>
<td>256</td>
<td>0.000</td>
<td>0.097</td>
</tr>
<tr>
<td>Ammo030</td>
<td>243</td>
<td>1.000*</td>
<td>0.181</td>
</tr>
<tr>
<td></td>
<td>249</td>
<td>0.000</td>
<td>0.125</td>
</tr>
<tr>
<td></td>
<td>264</td>
<td>0.000</td>
<td>0.319*</td>
</tr>
<tr>
<td></td>
<td>267</td>
<td>0.000</td>
<td>0.278</td>
</tr>
<tr>
<td></td>
<td>270</td>
<td>0.000</td>
<td>0.014</td>
</tr>
<tr>
<td></td>
<td>273</td>
<td>0.000</td>
<td>0.056</td>
</tr>
<tr>
<td></td>
<td>279</td>
<td>0.000</td>
<td>0.028</td>
</tr>
<tr>
<td>Ammo036</td>
<td>191</td>
<td>0.000</td>
<td>0.764*</td>
</tr>
<tr>
<td></td>
<td>194</td>
<td>0.443*</td>
<td>0.028</td>
</tr>
<tr>
<td></td>
<td>197</td>
<td>0.029</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>0.129</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>203</td>
<td>0.086</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>206</td>
<td>0.229</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>209</td>
<td>0.071</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>212</td>
<td>0.014</td>
<td>0.194</td>
</tr>
<tr>
<td></td>
<td>215</td>
<td>0.000</td>
<td>0.014</td>
</tr>
<tr>
<td>Ammo016</td>
<td>245</td>
<td>0.029</td>
<td>0.944*</td>
</tr>
<tr>
<td></td>
<td>251</td>
<td>0.100</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>254</td>
<td>0.300</td>
<td>0.014</td>
</tr>
<tr>
<td></td>
<td>257</td>
<td>0.386*</td>
<td>0.028</td>
</tr>
<tr>
<td></td>
<td>260</td>
<td>0.157</td>
<td>0.014</td>
</tr>
<tr>
<td></td>
<td>263</td>
<td>0.029</td>
<td>0.000</td>
</tr>
<tr>
<td>Ammo012</td>
<td>177</td>
<td>0.000</td>
<td>0.984</td>
</tr>
<tr>
<td></td>
<td>189</td>
<td>0.803</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>192</td>
<td>0.197</td>
<td>0.016</td>
</tr>
</tbody>
</table>