View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Online Research Database In Technology

Technical University of Denmark

A resource of genome-wide single-nucleotide polymorphisms generated by RAD tag sequencing in the critically endangered European eel

Pujolar, J.M.; Jacobsen, M.W.; Frydenberg, J.; Als, Thomas Damm; Larsen, P.F.; Maes, G.E.; Zane, L.; Jian, J.B.; Cheng, L.; Hansen, M.M.

Published in:
Molecular Ecology Resources

Link to article, DOI:
10.1111/1755-0998.12117

Publication date:
2013

Link back to DTU Orbit

Citation (APA):
Pujolar, J. M., Jacobsen, M. W., Frydenberg, J., Als, T. D., Larsen, P. F., Maes, G. E., ... Hansen, M. M. (2013). A resource of genome-wide single-nucleotide polymorphisms generated by RAD tag sequencing in the critically endangered European eel. Molecular Ecology Resources, 13, 706-714. DOI: 10.1111/1755-0998.12117

DTU Library
Technical Information Center of Denmark

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Molecular Ecology Resources
http://dx.doi.org/10.1111/1755-0998.12117 **© 2013 John Wiley & Sons Ltd**

A resource of genome-wide single nucleotide polymorphisms generated by RAD tag sequencing in the critically endangered European eel

J. M. Pujolar[1]*, M. W. Jacobsen[1], J. Frydenberg[1], T.D. Als[2], P. F. Larsen[3], G. E. Maes[4], L. Zane[5], J. B. Jian[6], L. Cheng[7], M. M. Hansen[1]

[1]Department of Bioscience, Aarhus University, Ny Munkegade 114, Bldg. 1540, DK-5 8000 Aarhus C, Denmark
[2]National Institute of Aquatic Resources, Technical University of Denmark, Vejlsøvej 39, DK-8600 Silkeborg, Denmark
[3]Kopenhagen Fur, Langagervej 60, DK-2600 Glostrup, Denmark
4Laboratory of Biodiversity and Evolutionary Genomics, Deberiotstraat 32, University of Leuven (KU Leuven), B-3000 Leuven, Belgium
[5]Department of Biology, University of Padova, Via G. Colombo 3, I-35131 Padova, Italy
[6]BGI-Shenzhen, Main Building, Beishan Industrial Zone, Yantian District, 518083 Shenzhen, China
[7]BGI-Europe, Copenhagen Bio Science Park, Ole Maaløes Vej 3, DK-2200 Copenhagen, Denmark

ABSTRACT
Reduced representation genome sequencing such as RAD (Restriction-site Associated 2 DNA) sequencing is finding increased use to identify and genotype large numbers of 3 single nucleotide polymorphisms (SNPs) in model and non-model species. We 4 generated a unique resource of novel SNP markers for the European eel using the 5 RAD sequencing approach that were simultaneously identified and scored in a 6 genome-wide scan of 30 individuals. Whereas genomic resources are increasingly 7 becoming available for this species, including the recent release of a draft genome, no 8 genome-wide set of SNP markers was available until now. The generated SNPs were 9 widely distributed across the eel genome, aligning to 4,779 different contigs and 10 19,703 different scaffolds. Significant variation was identified, with an average 11 nucleotide diversity of 0.00529 across individuals. Results varied widely across the 12 genome, ranging from 0.00048 to 0.00737 per locus. Based on the average nucleotide 13 diversity across all loci, long-term effective population size was estimated to range 14 between 132,000 and 1,320,000, which is much higher than previous estimates based 15 on microsatellite loci. The generated SNP resource consisting of 82,425 loci and 16 376,918 associated SNPs provides a valuable tool for future population genetics and 17 genomics studies and allows for targeting specific genes and particularly interesting 18 regions of the eel genome.

Keywords: effective population size; population genomics; RAD sequencing; SNP discovery

*Corresponding author: jmartin@biology.au.dk

Article first published online: 9 MAY 2013
Please note that this is an author-produced PostPrint of the final peer-review corrected article accepted for publication. The definitive publisher-authenticated version can be accesses here:
http://dx.doi.org/10.1111/1755-0998.12117 © 2013 John Wiley & Sons Ltd

Introduction

Recent advances in the speed, cost and accuracy of next-generation sequencing technologies are revolutionizing the field of population genetics and facilitating the application of genomic approaches into ecological and evolutionary studies (Allendorf *et al*. 2010; Davey *et al*. 2011). The growing accessibility to high-throughput sequencing methods allows the production of extremely large collections of data and the discovery of genome-wide resources at relatively modest and decreasing costs. Although ecological and evolutionary genomic studies involving the complete sequencing of multiple individuals and/or populations are still costly and have been restricted to few organisms (Jones *et al*. 2012a), genotyping-by-sequencing approaches (i.e. sequencing of a reduced representation of the genome followed by single-nucleotide polymorphism (SNP) discovery) can provide data on hundreds of thousands of SNPs that are to some extent evenly distributed across the genome.

One such genotyping-by-sequencing approach is the use of high-throughput sequencing of Restriction-site Associated DNA tags (RADs) (Miller *et al.* 2007; Baird *et al.* 2008). RAD tags are short fragments of DNA adjacent to each instance of a particular restriction enzyme recognition site. Different RAD tag densities can be achieved by choice of restriction enzyme. By focusing sequencing efforts only on those tags flanking a restriction site in multiplexed individually-barcoded samples, RAD sequencing allows efficient high-density identification of SNPs. Recently, a number of related genotyping-by-sequencing methods have been developed, including double-digest methods that considerably simplify library construction but generally also provide less coverage of the genome as compared to the original RAD method (Elshire *et al.* 2011; Peterson *et al.* 2012; Bruneaux *et al.* 2013). Different types of genotyping-by-sequencing approaches have been successfully used to discover thousands of SNPs in fish (Hohenlohe *et al.* 2010; 2011; Bruneaux *et al.* 2013), mammals (Peterson *et al.* 2012), insects (Emerson *et al.* 2010) and plants (Barchi *et al.* 2011; Scaglione *et al.* 2012). Hence, these methods by themselves allow for dense genome scans, but also identify thousands of markers, subsets of which can subsequently be genotyped in larger numbers of individuals using different genotyping technologies (Helyar *et al.* 2011).

The advent of next-generation sequencing technologies such as RAD sequencing is driving a shift from microsatellite to SNP genotyping in organisms with and without a

reference genome. The main advantages of SNPs are their high abundance and regular distribution across the genome, low scoring error rates, high reproducibility, a simple mutation model and the ability to concurrently screen neutral variation and regions of the genome under selection (Morin *et al.* 2004). Despite microsatellites typically presenting higher diversity per locus, a panel of several hundred SNPs is likely to be more informative than the 10-20 microsatellite loci used in standard population genetic studies (Helyar *et al.* 2011; Seeb *et al.* 2011), as shown in mapping (Ball *et al.* 2010), parentage (Hauser *et al.* 2011) and stock identification studies (Hess *et al.* 2011). The use of genotyping-by-sequencing methods to identify SNPs has many applications in ecological, evolutionary and population genetic studies. For example, Emerson *et al.* (2010) showed that RAD sequencing can be used to reveal previously unresolved genetic structure and detailed patterns of postglacial phylogeography of a non-model organism, the North American pitch planter mosquito, *Wyeomyia smithii*. Besides the assessment of population structure, genotyping-by-sequencing methods can also be used to detect signatures of selection and local adaptation. Hohenlohe *et al.* (2010) measured genome-wide genetic diversity across marine and freshwater populations of threespine stickleback (*Gasterosteus aculeatus*) using a high-density genome scan of 45,000 SNPs, which identified genomic regions exhibiting signatures of both balancing and directional selection.

Here, we use RAD tag sequencing to generate a resource of genome-wide SNPs in the European eel, *Anguilla anguilla*, a catadromous fish species with a particularly complex life cycle. After spawning in frontal zones of the southern Sargasso Sea, larvae cross the Atlantic Ocean following the Gulf Stream and metamorphose into glass eels upon reaching the Eastern Atlantic. Glass eels complete the migration into continental (freshwater, brackish, coastal) habitats as yellow eels, and after a highly variable feeding period, they metamorphose into silver eels that migrate back to the Sargasso Sea utilizing their high fat reserves, spawn once and die (van den Thillart *et al.* 2009). Remarkably, despite occupying a broad range of habitats from Subarctic environments in Iceland and northern Scandinavia to Subtropical environments in North Africa and the Mediterranean region, the European eel has been demonstrated to be a panmictic species (Als *et al*. 2011), a pattern that has also been revealed in the closely-related American eel *A. rostrata* (Coté *et al*. 2012).

In 2008, the long-term stock decline of the European eel prompted its inclusion in the IUCN (International Union for the Conservation of Nature) Red List of Threatened Species (www.iucnredlist.org), with a current status as "critically endangered". All over Europe, the abundance of all life-stages of eel (glass eel, yellow eel, silver eel) has severely decreased since the mid 1980s. The recruitment of glass eels entering rivers has been exceptionally low over the last five years, with a decline of 99% (continental North Sea) and 95% (rest of Europe) in comparison with the 1960-1979 levels (ICES 2011). Possible causes for the decline include anthropogenic factors such as overfishing, pollution, man-introduced parasites (the swimbladder nematode *Anguillicola crassus*) and diseases (EVEX virus) (van den Thillart *et al.* 2009), as well as climate and ocean current change (Knights 2003; Friedland *et al.* 2007; Bonhommeau *et al.* 2008).

A better understanding of crucial aspects of the biology of the European eel, including genetic diversity, effective population size and possible evolutionary responses to anthropogenic stressors, may promote measures to protect the species. Traditionally these issues have been addressed by using a low number of genetic markers due to the limited genomic resources available for eels. Two new rich sources of data have been recently made available: the first European eel transcriptome database Eeelbase (Coppe *et al.* 2010), which was recently updated to about 45,000 contigs (Pujolar *et al.* 2012); and the first eel draft genome based on Illumina sequencing and a *de novo* assembly (Henkel *et al.* 2012), with the genome size determined to be 1.1 Gbp. The present study reports the generation of genomic RAD tags from a total of 30 glass eels from three separate sampling locations. The RAD tags enabled the discovery of novel candidate SNP markers, thereby providing the first genotyping-by-sequencing data set for a wide-spread, highly fecund marine fish species, and generating a SNP resource that can be used for selecting subsets of markers to be genotyped using medium- or high-throughput platforms.

Material and Methods

RAD tag sequencing

Samples of glass eels were collected at three separate locations: one location in the western Mediterranean, the gulf of Valencia in Spain (39°49'N; 0°24'W), and two

locations in the eastern Atlantic, the Gironde estuary north of Bordeaux in France (45°15′N; 0°69′W) and the Burrishoole river in North-west Ireland (53°53′N; 9°34′W). Although the species is panmictic, sampling of geographically distinct localities accounts for the possibility that spatially and temporally variable selection might occur (Gagnaire *et al*. 2012). Genomic DNA was purified from a total of 30 individuals (10 from each location) using standard phenol-chloroform extraction.

Genomic DNA from each individual was digested with restriction enzyme EcoRI. A preliminary analysis suggested on average one cutting site every 2,346 bp. The digested product was ligated to a modified Illumina P1 adapter containing individual-specific nucleotide barcodes 4-8 bp long for sample tracking. All barcodes differed by at least two nucleotides to minimize sample mis-assignment due to sequencing error. Adapter-ligated fragments were subsequently pooled and sheared to an average size of 500 bp. Sheared DNA was separated by electrophoresis on a 2% agarose gel and fragments in the 350-500 bp size range were isolated using a MinElute Gel Extraction kit (Qiagen). After dsDNA ends were treated with end blunting enzymes and 3´-adenine overhangs were added, a modified Illumina P2 adapter was ligated. Finally, libraries were enriched by PCR amplification and RADs for each individual were sequenced (10 individuals per sequencing lane) on an Illumina Genome Analyzer II by Beijing Genomics Institute (BGI, Hong Kong, China) using paired-end reads.

RAD data analysis and SNP identification

Sequence reads from the Illumina runs were sorted according to their unique barcode tag. Sequences were quality-filtered using the FASTX-Toolkit (http://hannonlab.cshl.edu/fastx-toolkit) and reads with ambiguous barcodes and of poor quality were removed from the analysis. A minimum Phred score of 10 (equivalent to 90% probability of being correct) per nucleotide position was chosen, meaning that reads were dropped if a single nucleotide position had a score lower than 10. This is the Phred score generally used in SNP discovery studies (Ellison *et al*. 2011; Scaglione *et al*. 2012; Van Bers *et al*. 2012; Wagner *et al*. 2012). Final read length was trimmed to 75 nucleotides, following a preliminary analysis that showed a substantial increase in the number of SNPs at the tails of the sequences (from position 76 onwards), suggestive of sequencing errors (Figure 1). For subsequent analyses, only the first (left) paired-read was used. The DNA fragments created by RAD tag

library preparation have a restriction site at one end and are randomly sheared at the other end, which results in each instance of a restriction site sequence being sampled many times by the first reads and the genomic DNA sequence in the nearby region being randomly sampled at a lower coverage by the second paired-end reads (Etter *et al*. 2011), which are therefore less suitable for calling SNPs.

Sequence reads were aligned to the European eel genome draft (www.eelgenome.com) using the un-gapped aligner Bowtie version 0.12.8 (Langmead *et al*. 2009). A maximum of two mismatches between the individual reads and the genome were allowed and alignments were suppressed for a particular read when more than one reportable alignment existed, thereby decreasing the risk of paralogous sequences in the data.

The reference-aligned data were then used to assemble the RAD sequences into loci and identify alleles using the ref_map.pl pipeline in Stacks version 0.9995 (Catchen *et al*. 2011). First, exactly-matching sequences are aligned together into stacks, which are in turn merged to form putative loci. At each locus, nucleotide positions are examined and SNPs are called using a maximum likelihood framework. Second, a catalog is created of all possible loci and alleles. Third, each individual is matched against the catalog. A minimum stack depth of 10 reads was used, which is the number of exactly matching reads that must be found to create a stack in an individual. Finally, the program Populations in Stacks was used to process all the SNP data across individuals. The minimum number of individuals to process a locus was set to 66.7% of the individuals sequenced.

Genome-wide measures of genetic diversity, including observed (H_o) and expected (H_e) heterozygosities and nucleotide diversity (π), were calculated at each nucleotide site for all individuals as described in Hohenlohe *et al*. (2010). Using the average nucleotide diversity across all loci, long-term effective population size (N_e) was estimated using $\pi = 4*N_e*\mu$ (Tajima 1983), where μ is the mutation rate per site per generation. SNPs have relatively low mutation rates ($1x10^{-8}$ - $1x10^{-9}$ per generation; Brumfield *et al*. 2003) in comparison with other markers such as microsatellites that have mutation rates per generation of the order of 10^{-4}.

Finally, batch BLAST similarity searches were conducted locally for all loci in the catalog using BLAST+ (NCBI). All sequences were blasted against the predicted complete transcripts from either scaffolds or unscaffolded contigs in the European eel

genome database (www.eelgenome.com). BLASTN searches were conducted using default parameters. Alignments with an e-value < 0.001 were considered significant. In case of multiple hits, best match was kept. Different annotation similarity cut-off values (60%, 80%, 90%) were considered.

Results

Sequencing of the RAD libraries generated an average of 8.67 million reads of 90 bp per individual, prior to any quality filtering. The number of reads ranged from 5.33 to 13.03 million reads per individual. After quality filtering, on average 6.94 million (80.2%) sequences per individual were retained and 1.73 million (19.8%) sequences were eliminated. Retained sequences presented a mean quality score of 38.61, a median of 39.41 and a GC content of 40.6% (Table 1).

Out of the retained sequences, an average of 4.89 million (70.41%) aligned to the European eel draft genome, 1.75 million (25.17%) were not aligned and 306,969 (4.42%) sequences were discarded due to alternative alignments (more than one reportable alignment existed) (Table 1).

Aligned sequences were assembled into an average of 489,870 stacks per individual and subsequently into a set of 328,812 loci. Using a minimum coverage of 10 reads per individual, an average of 202,923 (61.5%) loci were retained. Average coverage was 22.52 ± 2.18 read per locus. A total of 125,890 (38.5%) loci were discarded per individual due to insufficient depth of coverage (Table 1). The ratio between observed and expected loci (based on the number of EcoRI cutting sites) was 65.1% when using a minimum stack depth of 10 reads per locus and 88.0% when using a minimum stack depth of 1 read per locus.

A catalog of 422,634 loci was constructed using all 30 individuals. After a final filtering step focused on loci genotyped in >20 out of the 30 individuals, a total of 142,509 loci were retained for SNP discovery. Out of these, 13,220 (9.27%) loci were monomorphic, 8,770 (6.14%) loci showed more than 2 alleles per individual (and were consequently eliminated from further analyses) and 120,539 (84.58%) were polymorphic, producing a total of 530,030 candidate SNP markers.

Average number of SNPs per locus was 3.96, ranging between 1 and 22 (Figure 2). Only 14.70% of the loci presented one single SNP, with 2 SNPs being the most frequent (17.61%). SNPs were evenly distributed across nucleotide positions in the

sequence reads and no apparent increase of SNPs toward the end of the reads was observed. About two thirds of the SNPs proved to be transitions in our dataset, with an observed transition:transversion ratio of 1.6:1 (Figure 3).

In order to support the validity of the large number of SNPs detected, data was re-analyzed using different parameters in the analysis. Firstly, we tested the effect of the alpha value used for the chi-square significance level when SNP calling. Similar results were obtained when using the default alpha of 0.05 (530,030 SNPs) or when using a more stringent alpha of 0.001 (527,352 SNPs), with a difference of less than 1%. Secondly, we tested the effect of quality filtering using different Phred scores. Using a more conservative Phred score of 20, a large number of SNPs was still detected (461,380 SNPs). The use of different Phred scores had no apparent effect on the total number of loci (422,634 using a Phred score of 10; 407,401 using a Phred score of 20), number of loci with more than two alleles (6.1% using a Phred score of 10; 5.8% using a Phred score of 20), average number of SNPs per locus (3.96 using a Phred score of 10; 3.89 using a Phred score of 20) and maximum number of SNPs per locus (over 20 in both cases). Thirdly, we re-analyzed all data using also the second (right) paired-end for alignment (but not for SNP calling), which makes the process more conservative. By comparing the results obtained when using the left paired-end only and when using both left and right paired-ends for alignment, we can determine if those loci presenting high numbers of SNPs are the consequence of poor alignment. Using both paired-ends, loci with high number of SNPs were still detected, up to 21 SNPs per loci, and the average number of SNPs per loci was 3.64, similar to the value found when using only the left paired-end (3.96). The fact that loci with over 20 SNPs were found independently of quality filtering, SNP calling or alignment procedure suggests that the method used for SNP discovery is accurate.

SNPs were widely distributed across the genome and were found in a total of 4,779 different contigs and a total of 19,703 different scaffolds. When loci sequences were compared to the European eel genome using BLASTN, a significant similarity was found for 10,376 (6.8%) loci. Monomorphic loci showed a higher association with transcripts from either scaffolds or contigs in the eel genome (10.1%) than polymorphic loci (6.4%). Few loci were annotated, 0.2% using a cut off of 90, 0.3% using a cut off of 80 and 3.3% using a more relaxed cut off of 60% similarity. Annotations were higher in monomorphic loci (0.2% using a cut off of 90, 1.1% using

a cut off of 80 and 5.7% using a cut off of 60) than in polymorphic loci (0.1% using a cut off of 90, 0.7% using a cut off of 80 and 3.1% using a cut off of 60).

Finally, genome-wide measures of genetic diversity were calculated from the SNP data. A sequence length of 70 nucleotides was considered, since the first 5 nucleotides constitute the recognizing sequence motif for the restriction endonuclease. Substantial variation was identified, with average nucleotide diversity (π) equal to 0.00529 ± 0.00110 across all 30 individuals included in the study. Results varied widely across loci, ranging from 0.00048 to 0.00737. Average observed and expected heterozygosity were 0.00468 and 0.00518, respectively.

Using the average nucleotide diversity across all loci, long-term effective population size (N_e) was estimated using Tajima's (1983) formula $\pi = 4*N_e*\mu$, where μ is the mutation rate per site per generation. N_e was estimated to range between 132,000 (using a mutation rate of $1x10^{-8}$ per site per year) and 1,320,000 (using a mutation rate of $1x10^{-9}$ per site per year).

As a final step, we generated a SNP resource available as an Excel spreadsheet (Table S1), including sequences of RAD tags, identified SNPs and their position in the European eel draft genome. For the resource, we excluded those loci in which all SNPs were singletons. In total the resource includes 82,425 loci in which at least one SNP was present in a minimum of two individuals. For these loci, apparent singleton SNPs are also reported since their presence may be relevant for primer design and for assessing if the SNPs are found in particularly variable genomic regions. The total number of SNPs in the resource is 376,918.

Discussion

Large Scale SNP identification

We report the discovery of a large number of SNPs in the European eel genome using the RAD sequencing approach. After excluding those loci in which all SNPs were singletons, we generated a large resource consisting of 82,426 loci and 376,918 associated SNPs. While the amount of genomic resources available for this species are rapidly increasing, with the recent release of a draft genome, no genome-wide set of SNP markers was available until now. The generation of such a large panel of novel SNPs represents a major step in terms of genomic resources available for this species (Table S1). In this sense, only 49 microsatellite markers have been developed to date

in the European eel, including a panel of 12 dinucleotide microsatellites identified from enriched libraries (Wielgoss *et al*. 2008) and a larger set of 28 expressed sequence tag (EST)-linked microsatellite loci (Pujolar *et al*. 2008). Additionally, 232 proteins, 177 ESTs and the complete mitochondrial genome are available in GenBank. The low number of markers available has somehow constrained genetic studies during the last two decades, and most studies have been conducted using <20 (or even <10) microsatellite loci. While classic population and conservation studies based on a few markers provide a "snapshot" of the variation in the genome, the panel of novel SNPs presented here will facilitate the development of population genomics studies on the European eel. Obviously, such studies can proceed using RAD sequencing for more samples, or they can make use of the generated SNP resource (Table S1) for selecting subsets of markers for genotyping in high numbers of individuals. The latter would be particularly advantageous when focusing on specific genes or parts of the genome or when analyzing degraded samples, such as DNA extracted from historical samples of otoliths or other hard parts (Nielsen & Hansen 2008), for which RAD sequencing and related methods are not suitable (Davey *et al.* 2011).

The feasibility of genome-scan approaches has been illustrated by several recent studies in a variety of organisms, including eukaryotes (Ellison *et al*. 2011), plants (Namroud *et al*. 2008; Turner *et al*. 2010), invertebrates (Turner *et al*. 2005, 2008) and fishes (Hohenlohe *et al*. 2010; Willing *et al*. 2010; Jones *et al*. 2012b). Genome-scan approaches such as SNP discovery using genotyping-by-sequencing can also provide a better understanding of adaptive evolution by means of identifying genes associated with ecologically important traits. Candidate genes and genomic regions can be identified using an F_{ST} outlier approach by detecting loci showing increased or decreased differentiation across populations compared to neutral expectations, suggestive of directional or purifying natural selection. Specifically for a panmictic species like the European eel, SNP based genome scans could be used to test within-cohort selection resulting from geographically varying environmental conditions encountered by glass eels across different regions of Europe and North Africa. In the case of the American eel, the recent study of Gagnaire *et al*. (2012) identified SNPs under possible temperature-related selection, with 13 loci showing correlations between allele frequencies and environmental variables across the entire species range. Moreover, introduced pathogens and parasites may have contributed to the

recent decline of the European eel (van den Thillart *et al.* 2009). Retrospective monitoring of SNPs associated with immune system related genes could be conducted based on contemporary and historical samples (e.g. archived otoliths) (Hansen *et al*. 2012), which would allow for testing for possible adaptive responses to pathogens and parasites in the species.

High SNP density points to large effective population size in the European eel

One interesting result in our study is the high density of SNPs identified, with an average of 3.96 SNPs per locus and a maximum of 22. Sequencing errors, mostly found in the last nucleotide positions of the sequence reads, can mistakenly be identified as SNPs. If a substantial number of predicted SNPs in the dataset are the result of sequencing errors, an increase in the amount of SNPs toward the tails of the reads is expected. This was apparent in a pre-analysis of sequences trimmed to 80 bp showing a 20% over-representation of SNPs in positions 76-80. The fact that SNPs were equally distributed over the reads after trimming all sequences to 75 bp, indicates that the majority of our SNPS are not the result of sequencing errors and that our large scale SNP identification approach is valid. Additionally, we calculated the transition:transversion ratio of the SNPs in our dataset. If polymorphisms were introduced at random, a transition (A<->G or C<->T) to transversion (A<->C, A<->T, C<->G, G<->T) rate of 1:2 would be expected. The SNPs in our dataset showed a transition:transversion ratio of 1.6:1, which suggests a very small influence of sequencing error on SNP calling. Similar transition:transversion ratios have been reported in the eggplant (1.65:1; Barchi *et al*. 2011) and the great tit (1.7:1; Van Bers *et al*. 2010). The fact that the number of SNPs found per locus did not change when applying more conservative quality filtering, SNP calling significance level and alignment procedures further supports the validity of the SNPs.

One explanation to the substantial polymorphism detected might be that the vast majority of our data is intergenic and intronic, as suggested by the low number of loci annotated using BLAST analysis (0.2-5.7% depending on the sequence similarity criterion used). In comparison, 2% of validated SNPs generated by deep sequencing of a reduced representation library were associated with rainbow trout transcripts (Castaño-Sanchez *et al*. 2009), and similar values have been found in humans (2%) and chimpanzee (1%) SNPs (Hodgkinson *et al*. 2009). Despite the high occurrence of

SNPs in our study, the presence of a large number of singletons and alleles in low frequency resulted in only a moderately high nucleotide diversity (π =0.00529). π also allows to estimate the long-term effective population size (N_e) using $\pi = 4*N_e*\mu$ in a model in which sites evolve neutrally. Nevertheless, it should be noted that the model assumes an idealized population with random mating and constant size, which might not be necessarily met in the case of the European eel. The estimated N_e ranged between 132,000 and 1,320,000 individuals, depending on the mutation rate used, a much larger value than those previously reported in the literature. Using seven microsatellite loci, Wirth and Bernatchez (2003) estimated a long-term N_e of 4,410-5,388 individuals inferred by the coalescent-based geneological method in MSVAR (Storz & Beaumont 2002). Using a larger panel of 22 EST-derived microsatellite loci, Pujolar *et al*. (2011) estimated a long-term N_e of 5,444-10,474 individuals inferred by a different Bayesian genealogy sampler (LAMARC; Kuhner 2006), which was consistent with the estimated values of short-term N_e of 5,031 (2,986-12,810) inferred by the comparison of allele frequencies across samples. The differences across studies, with a higher long-term N_e estimated in our study, can be due to the number and nature of microsatellite loci. In particular, MSVAR and LAMARC assume a simplistic stepwise mutation rate, whereas mutational properties at microsatellite loci are in reality more complex (Di Rienzo *et al*. 1994). The estimation of a relatively high effective population size is not surprising given that the species consists of one single large panmictic unit (Als *et al.* 2011). Nevertheless, it might be seen to contrast with the low abundance of recruitment and landings of yellow and silver eels occurring all over Europe (ICES 2011). However, it should be noted that this is a historical N_e estimate, whereas a short-term N_e estimate would be required to detect the recent declines (Waples 2005).

Collectively, the generation of a resource of 82,425 loci and 376,918 associated SNPs provides a valuable tool for future population genetics and genomics studies in the European eel and allows for targeting particularly interesting regions of the eel genome. All RAD tag sequences and associated SNPs are presented in a spreadsheet along with their map position in the draft eel genome (Table S1). Such resources were until recently only available for model organisms, whereas European eel must definitely be considered a non-model organism. Crucial aspects of its life cycle are still

unresolved and attempts to artificially propagate the species have so far proven unsuccessful (Tomkiewicz 2012). Hence, the generated eel SNP resource provides a clear illustration of the advances in next-generation sequencing and its potentials for overcoming the gap between model and non-model species.

Acknowledgements

We thank Annie Brandstrup for technical assistance, Russel Poole, Javier Lobon and Eric Feunteun for samples and Julian Catchen for advice on data analysis using Stacks. We also thank the subject editor and three anonymous referees for comments on a previous version of the paper. We acknowledge funding from the Danish Council for Independent Research, Natural Sciences (grant 09-072120).

Author's contribution

MMH and PFL conceived and designed the project. JMP, MMH and MWJ conducted all bioinformatics analyses. JBJ and LC were involved in data generation. JMP wrote the manuscript with contributions from MMH, MWJ, LZ, JF, TDA, PFL and GEM.

Data Accessibility Statement

Sequence reads have been deposited in the NCBI Sequence Read Archive (Accession number SRP020485).

References

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

Als TD, Hansen MM, Maes GE *et al*. (2011) All roads lead to home: panmixia of European eel in the Sargasso Sea. *Molecular Ecology*, **20**, 1333-1346.

Baird NA, Etter PD, Atwood TS *et al*. (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One*, **3**, e3376.

Ball AD, Stapley J, Dawson SA, Birkhead TR, Burke T, Slate J (2010) A comparison of SNPs and microsatellites as linkage mapping markers: lessons from the zebra finch (*Taeniopygia guttata*). *BMC Genomics*, **11**, 218.

Barchi L, Lanteri S, Acquadro A, Vale G, Toppino L, Rotino GL (2011) Identification of SNP and SSR markers in eggplant using RAD tag sequencing. *BMC Genomics*, **12**, 304.

Bonhommeau S, Chassot E, Rivot E (2008) Fluctuations in European eel (*Anguilla anguilla*) recruitment resulting from environmental changes in the Sargasso Sea. *Fisheries Oceanography*, **17**, 32-44.

Brumfield RT, Beerli P, Nickerson DA, Edwards SV (2003) The utility of single nucleotide polymorphisms in inferences of population history. *Trends in Ecology and Evolution*, **18**, 249-256.

Bruneaux M, Johnston SE, Herczeg G *et al.* (2013) Molecular evolutionary and population genomic analysis of the nine-spined stickleback using a modified restriction-site-associated DNA tag approach. *Molecular Ecology*, **22**, 565-582.

Catchen JM, Amores A, Hohenlohe PA, Cresko WA, Postlethwait JH (2011) Stacks: building and genotyping loci de novo from short-read sequences. *G3: Genes, Genomes, Genetics*, **1**, 171-182.

Castaño-Sanchez C, Smith TPL, Wiedmann RP *et al*. (2009) Single nucleotide polymorphism discovery in rainbow trout by deep sequencing of a reduced representation library. *BMC Genomics*, **10**, 559.

Coppe A, Pujolar JM, Maes GE *et al*. (2010) Sequencing, *de novo* annotation and analysis of the first *Anguilla anguilla* transcriptome: EeelBase opens new perspectives for the study of the critically endangered European eel. *BMC Genomics*, **11**, 635.

Côté C, Gagnaire PA, Bourret V, Verrault G, Castonguay M, Bernatchez L (2012) Population genetics of the American eel (*Anguilla rostrata*): F_{ST}=0 and North Atlantic Oscillation effects on demographic fluctuations of a panmictic species. *Molecular Ecology*, in press.

Davey JW, Hohenlohe PA, Etter PD *et al*. (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, **12**, 499-510.

Di Rienzo A, Peterson AC, Garza JC *et al*. (1994) Mutational processes of simple-sequence repeat loci in human populations. *Proceedings of the National Academy of Sciences of the United States of America*, **91**, 3166-3170.

Ellison CE, Hall C, Kowbel D *et al.* (2011) Population genomics and local adaptation in wild isolates of a model microbial organism. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 2831-2837.

Elshire RJ, Glaubitz JC, Sun Q *et al.* (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One*, **6**, e19379.

Emerson KJ, Merz CR, Catchen JM *et al*. (2010) Resolving postglacial phylogeography using high-throughput sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 16196-16200.

Etter PD, Preston JL, Bassham S, Cresko WA, Johnson EA (2011) Local *de novo* assembly of RAD paired-end contigs using short sequencing reads. *PLoS One*, **6**, e18561.

Friedland KD, Miller MI, Knights B (2007) Oceanic changes in the Sargasso Sea and declines in recruitment of the European eel. *ICES Journal of Marine Science*, **64**, 519-530.

Gagnaire PA, Normandeau E, Côté C *et al*. (2012) The genetic consequences of spatially varying selection in the panmictic American eel (*Anguilla rostrata*). Genetics, **190**, 725-736.

Hansen MM, Olivieri I, Waller DM *et al.* (2012) Monitoring adaptive genetic responses to environmental change. *Molecular Ecology*, **21**, 1311-1329.

Hauser L, Baird M, Hilborn R, 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.

Helyar SJ, Hemmer-Jansen J, Bekkevold D *et al.* (2011) Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges. *Molecular Ecology Resources*, **11**, 123-136.

Henkel CV, Burgerhout E, Danielle L *et al.* (2012) Primitive duplicate hox clusters in the European eel's genome. *PLoS One*, **7**, e32231.

Hess J, Matala AP, Narum SR (2011) Comparison of SNPs and microsatellites for fine-scale application of genetic stock identification of chinook salmon in the Columbia River Basin. *Molecular Ecology Resources*, **11**, 137-149.

Hohenlohe PA, Basshan S, Etter PD, Stiffler N, Johnson EA, Cresko WA (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genetics*, **6**, e1000862.

Hohenlohe PA, Amish SJ, Catchen JM, Allendorf FW, Luikart G (2011) Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westlope cutthroat trout. *Molecular Ecology Resources*, **11**, 117-122.

Hodgkinson A, Ladoukakis E, Eyre-Walker A (2009) Cryptic variation in the human mutation rate. *PLoS Biology*, **7**, e10000027.

ICES (2011) Report of the Joint EIFAAC/ICES Working Group on Eels (WGEEL), 5-9 September 2011, Lisbon, Portugal. ICES CM 2011/ACOM:18. International Council for the Exploration of the Seas, Copenhagen, Denmark.

Jones FC, Grabherr MG, Chan YF *et al*. (2012a) The genomics basis of adaptive evolution in threespine sticklebacks. *Nature*, **484**, 55-61.

Jones FC, Chan YF, Schmutz M *et al*. (2012b) A genome-wide SNP genotyping array reveals patterns of local and repeated species-pair divergence in sticklebacks. *Current Biology*, **22**, 83-90.

Knights B (2003) A review of the possible impacts of long-term oceanic and climatic changes and fishing mortality on recruitment of anguillid eels of the Northern hemisphere. *Science of the Total Environment*, **310**, 237-244.

Kuhner MK (2006) LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters. *Bioinformatics*, **22**, 768-770.

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, **10**, R25.

Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA (2007) Rapid and cost-effective polymorphism identification and genotyping using restriction-site associated DNA (RAD) markers. *Genome Research*, **17**, 240-248.

Morin PA, Luikart G, Wayne RK (2004) SNPs in ecology, evolution and conservation. *Trends in Ecology and Evolution*, **19**, 208-216.

Namroud MC, Beaulieu J, Juge N, Larouche J, Bousquet J (2008) Scanning the genome for gene single nucleotide polymorphisms involved in adaptive population differentiation in white spruce. *Molecular Ecology*, **17**, 3599-3613.

Nielsen EE, Hansen MM (2008) Waking the dead: the value of population genetic analyses of historical samples. *Fish and Fisheries*, **9**, 450-461.

Peterson BK, Weber JN, Kay EH *et al.* (2012) Double Digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS One*, **7**, e37135.

Pujolar JM, Maes GE, Van Houdt JK, Zane L (2008) Isolation and characterization of expressed sequence tag-linked microsatellite loci for the European eel, *Anguilla anguilla*. *Molecular Ecology Resources*, **9**, 233-235.

Pujolar JM, Bevacqua D, Cappocioni F, Ciccotti E, De Leo GA, Zane L (2011) No apparent genetic bottleneck in the demographically declining European eel using molecular genetics and forward-time simulations. *Conservation Genetics*, **12**, 813-825.

Pujolar JM, Marino IAM, Milan M *et al*. (2012) Surviving in a toxic world: transcriptomics and gene expression profiling in response to environmental pollution in the critically endangered European eel. *BMC Genomics*, **13**, 507.

Scaglione S, Acquadro A, Portis E, Tirone M, Knapp SJ, Lanteri S (2012) RAD tag sequencing as a source of SNP markers in *Cynara cardunculus*. *BMC Genomics*, **13**, 3.

Seeb JE, Carvalho GR, Hauser L, Niash K, Roberts S, Seeb LW (2011) Single-nucleotide polymorphism (SNP) discovery and application of SNP genotyping in nonmodel organisms. *Molecular Ecology Resources*, **11**, 1-8.

Storz, JF, Beaumont MA (2002) Testing for genetic evidence of population expansion and contraction: An empirical analysis of microsatellite DNA variation using a hierarchical Bayesian model. *Evolution*, **56**, 154-166.

Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations. *Genetics*, **105**, 437-460.

Tomkiewicz J (2012) Reproduction of European eel in aquaculture (REEL): consolidation and new production methods. DTU Aqua Report No 249-2012. National Institute of Aquatic Resources, Technical University of Denmark, 47p.

Turner TL, Bourne EC, Von Wettberg EJ, Hu TT, Nuzhdin SV (2010) Population resequencing reveals local adaptation in *Arabidopsis lyrata* to serpentine soils. *Nature Genetics*, **42**, 260-263.

Turner TL, Han MW, Nuzhdin SV (2005) Genomic islands of speciation in *Anopheles gambiae*. *PLoS Biology*, **3**, e285.

Turner TL, Levine MT, Eckert M L, Begun DJ (2008) Genomic analysis of adaptive differentiation in *Drosophila melanogaster*. *Genetics*, **179**, 455-473.

Van Bers NEM, Van Oers K, Kerstens HHD (2010) Genome-wide SNP detection in the great tit *Parus major* using high throughtput sequencing. *Molecular Ecology*, **19**, 89-99.

van den Thillart G, Rankin JC, Dufour S (2009) Spawning migration of the European eel: reproduction index, a useful tool for conservation management. Springer, Dordecht, The Netherlands.

Wagner CE, Keller I, Wittwer S *et al*. (2012) Genome-wide RAD sequence data provide unprecedented resolution in species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. *Molecular Ecology*, **22**, 787-798.

Waples RS (2005) Genetic estimates of contemporary effective population size: to what time periods do the estimates apply? *Molecular Ecology*, **14**, 3335-3352.

Wielgoss S, Wirth T, Meyer A (2008) Isolation and characterization of 12 dinocleotide microsatellites in the European eel, *Anguilla anguilla*, and tests for amplification in other species of eels. *Molecular Ecology Resources*, **8**, 1382-1385.

Willing EM, Bentzen P, van Oosterhout C *et al*. (2010) Genome-wide single nucleotide polymorphisms reveal population history and adaptive divergence in wild guppies. *Molecular Ecology*, **19**, 968-984.

Wirth T, Bernatchez (2003) Decline of North Atlantic eels: a fatal synergy? *Proceedings of the National Academy of Sciences of the United States of America*, **270**, 681-688.

Supporting Information

Table S1. Spreadsheet encompassing all RAD tag sequences and associated SNPs, along with their position in the draft eel genome.

Figure 1. Number of SNPs per nucleotide position (1-80). There is an apparent increase in number of SNPs in the last 5 nucleotides (76-80), suggestive of sequencing errors, which were consequently removed from the analyses.

Figure 2. Distribution of the number of SNPs per loci.

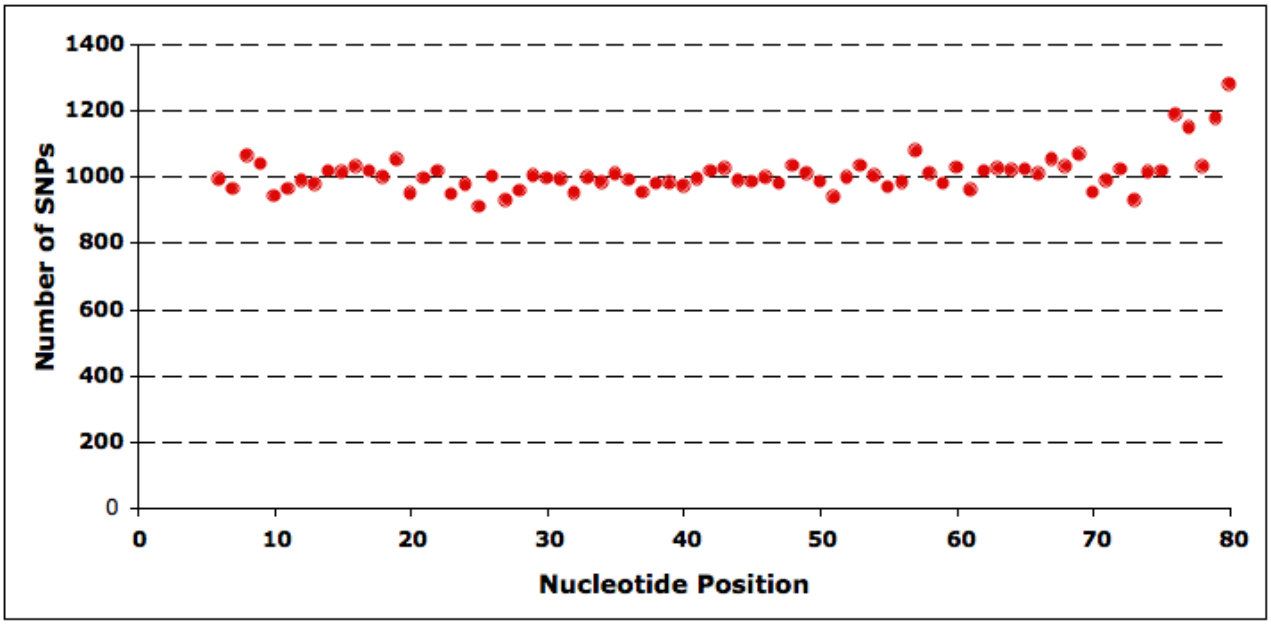
Figure 3. Transitions and transversions occurring within a set of 551,429 European eel SNPs.

Table 1. Statistics describing the distribution of different properties of RAD sequences after each step of filtering (FASTX-Toolkit), alignment to the eel draft genome (Bowtie) and assemblage into loci (Ref_map.pl).

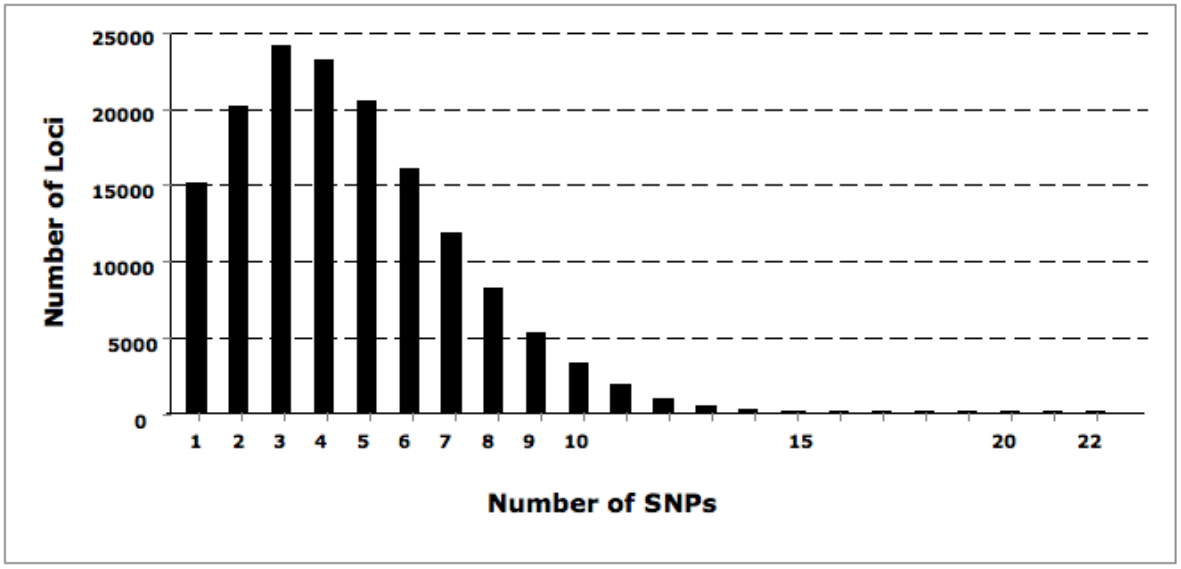
FASTX-Toolkit										
Raw reads	**Filtered reads**	**% Eliminated**	**Mean Q**	**Q1**	**Med**	**Q3**	**%A**	**%C**	**%G**	**%T**
8670526	6942282	19.8	38.6	38	39.4	40	29.8	20.5	20.1	29.7

Bowtie						
Reads	**Aligned**	**% Aligned**	**Non-aligned**	**% Non-aligned**	**Discarded**	**% Discarded**
6942282	4886517	70.4	1749063	25.2	306969	4.4

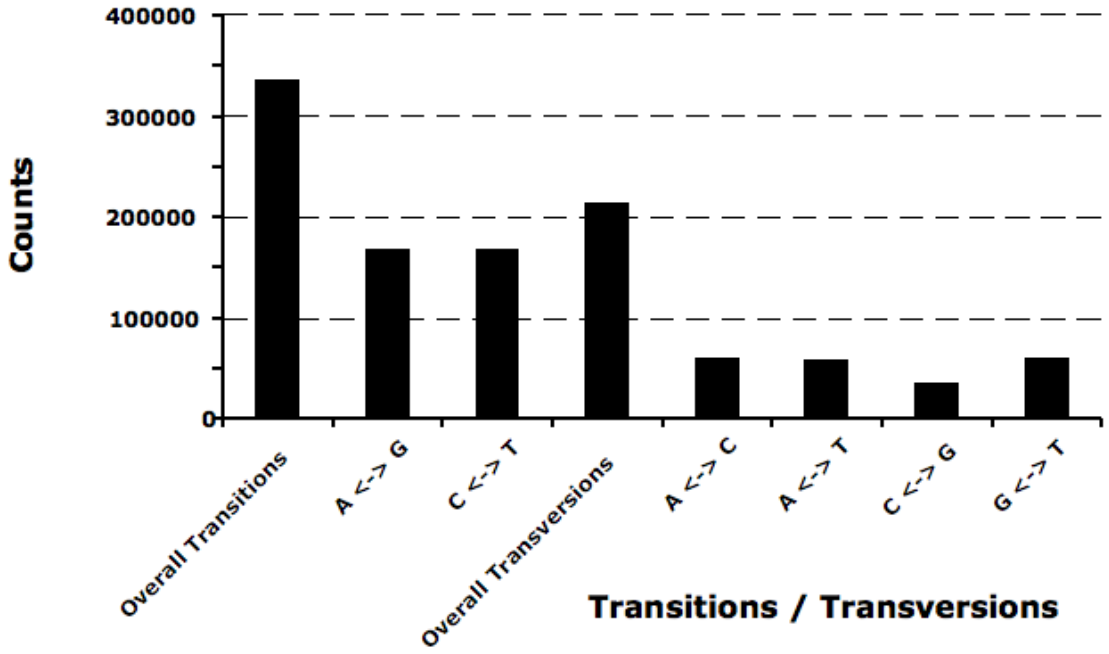
Ref_map.pl						
Reads	**Stacks**	**Loci**	**Loci used**	**% Loci used**	**Loci discarded**	**% Loci discarded**
4886517	489870	328812	202923	61.5	125889	38.5


1400
1200
1000
800
600
400
200
0
Number of SNPs
0
10
20
30
40
50
60
70
80
Nucleotide Position

1


25000
20000
15000
10000
5000
0
Number of Loci
1
2
3
4
5
6
7
8
9
10
15
20
22
Number of SNPs

1


400000
300000
200000
100000
0
Counts
Overall Transitions
A <-> G
C <-> T
Overall Transversions
A <-> C
A <-> T
C <-> G
G <-> T
Transitions / Transversions

1