

An EST-based genome scan using 454 sequencing in the marine snail *Littorina saxatilis*

J. GALINDO*, J. W. GRAHAME† & R. K. BUTLIN*

*Animal and Plant Sciences, The University of Sheffield, Western Bank, Sheffield, UK

†Institute of Integrative and Comparative Biology, The University of Leeds, Leeds, UK

Keywords:

local adaptation;
outlier analysis;
population genomics;
transcriptome scan;
454.

Abstract

Genome scans have been used in the studies of ecological speciation to find genomic regions ('outlier loci') showing reduced gene flow between divergent populations/species. High-throughput sequencing ('454') offers new opportunities in this field via transcriptome sequencing. Divergent ecotypes of the marine gastropod *Littorina saxatilis* represent a good example of incipient ecological speciation. We performed a 454-based genome scan between H and M ecotypes of *L. saxatilis* from the British Isles using cDNA of pooled individuals. Allele frequencies were calculated for 2454 single nucleotide polymorphisms (SNPs), within 572 contigs, and 7% of loci were detected as outliers. Functional annotation of the contigs containing outlier SNPs showed that they included shell matrix and muscle proteins (lithostathine, mucin, titin), proteins involved in energetic metabolism (arginine kinase, NADH dehydrogenase) and reverse transcriptases. Follow-up investigations into these proteins and unannotated outliers will be a promising route in the study of ecological speciation in *L. saxatilis*.

Introduction

Interest in the role of ecological divergence in speciation dates back to Darwin and the modern synthesis (reviewed in Sobel *et al.*, 2010). The term 'ecological speciation' is now used for the evolution of barriers to gene flow because of ecologically based divergent selection (Schluter, 2000, 2001). Populations adapted to different habitats are expected to progress towards complete reproductive isolation through pleiotropic effects of the adaptive traits themselves or through the acquisition of additional barriers such as habitat or mate choice (Nosil & Rundle, 2009 and references therein). Unravelling the mechanisms by which divergent selection can lead to reproductive barriers and discovering genes responsible for the evolution of those barriers are central questions in the study of ecological speciation (Schluter, 2009). During the early stages of ecological speciation with gene flow, genomically localized barriers because of divergent selection promote variation in the level of genetic differentiation across the genome (Nosil *et al.*, 2009; Via, 2009; Wu, 2001; Butlin, 2010). Regions involved in local adaptation will show high genetic

differentiation (i.e. reduced gene flow), whereas the rest of the genome will remain weakly differentiated because of recent divergence and/or ongoing gene flow. This heterogeneous genomic differentiation is expected to evolve towards broader genomic divergence along with the evolution of additional reproductive barriers. For this reason, pairs of populations in the initial stages of the speciation process are particularly suitable for the detection and further study of genomic regions involved in adaptation and isolation.

The population genomics approach (Luikart *et al.*, 2003) can be used to identify the regions of reduced gene flow between divergent populations, because it can separate locus-specific effects such as selection from genomewide effects of drift and gene flow. So-called outlier loci can be detected via a variety of approaches (reviewed in Storz, 2005). Population genomics has been extensively used in the study of ecological speciation for the detection of outlier loci (Nosil *et al.*, 2009 and references therein). AFLP-based genome scans were performed in many cases, but an important drawback of these genome scans is the laborious work necessary to link the outlier loci with a gene or function because, in general, these anonymous markers occur in noncoding regions (e.g. Wood *et al.*, 2008). One shortcut is to perform a scan based on markers in expressed sequence tags (ESTs) (Bonin, 2008).

Correspondence: Juan Galindo, Animal and Plant Sciences, The University of Sheffield, Western Bank, S10 2TN Sheffield, UK.
Tel.: +44 (0)114 2220112; fax: +44 (0)114 2220002;
e-mail: j.galindo@shef.ac.uk

Obtaining an EST database by traditional approaches was time-consuming and expensive and so was restricted to a few organisms (Wheat, 2010; Weber *et al.*, 2007). With the advent of the new generation of sequencing technologies (see Metzker, 2010 for a review), genetic/genomic resources for nonmodel species have become far more accessible and transcriptome sequencing is becoming one of the most important applications of next-generation sequencing in evolutionary biology (Ellegren, 2008; Hudson, 2008). For example, '454' pyrosequencing (Margulies *et al.*, 2005) allows for rapid transcriptome sequencing at a reasonable cost in any species, including those undergoing ecological speciation (Elmer *et al.*, 2010; Renaut *et al.*, 2010; Schwarz *et al.*, 2009). Single nucleotide polymorphism (SNP) discovery can be performed if pools of cDNA from multiple individuals are used (e.g. Novaes *et al.*, 2008; Vera *et al.*, 2008). For species undergoing ecological speciation, SNP frequency differences between divergent populations can be estimated for hundreds of loci and interesting candidates for adaptation can be found (e.g. Renaut *et al.*, 2010; Schwarz *et al.*, 2009). A more sophisticated approach would use simulations to perform outlier detection in 454-based genome scans. Here, we investigate the practicalities of this approach.

Our study organism, the marine gastropod *Littorina saxatilis*, has captured the attention of evolutionary biologists because of its high level of polymorphism over small spatial scales (reviewed in Johannesson, 2003). It has direct development (lacking a pelagic larval stage) with restricted movement of the adults, and this underlies local adaptation and ecotype formation (Rolán-Alvarez, 2007). Divergent ecotype pairs have been studied in detail on various different shores of the North Atlantic (western Sweden, northeast England and northwest Spain), and these ecotypes may represent parallel cases of nonallopatric ecological speciation (Panova *et al.*, 2006; Quesada *et al.*, 2007; Rolán-Alvarez *et al.*, 2004) (but see Johannesson *et al.*, 2010; Butlin *et al.*, 2008; Sadedin *et al.*, 2009). Within a single shore, divergent ecotypes are adapted to contrasting habitats with different exposure regimes and divergent natural selection has been shown to be responsible for the maintenance of the polymorphism (Grahame *et al.*, 2006; Johannesson, 2003; Rolán-Alvarez, 2007). Despite reproductive barriers such as assortative mating and habitat choice (reviewed in Rolán-Alvarez, 2007), there is ongoing gene flow between the ecotypes, average neutral genetic differentiation is low (Galindo *et al.*, 2009; Panova *et al.*, 2006; Wilding *et al.*, 2001) and AFLP genome scans showed that only a small proportion of the genome had greater than neutral differentiation (Galindo *et al.*, 2009; Wilding *et al.*, 2001).

In this study, we used 454 technology to perform an EST-based genome scan by sequencing cDNA pools of H and M ecotypes of *L. saxatilis*. SNP discovery, allele frequency estimation, outlier analysis and functional annotation of outlier loci were carried out. We have

adapted a simulation approach to outlier detection, used in previous genome scans, for use with 454 sequence data. This study represents only one step in the long process of understanding adaptation and ecological diversification in *Littorina's* ecotypes, but these initial stages are much facilitated by the availability of next generation sequencing technologies.

Materials and methods

H and M ecotypes of *L. saxatilis*

On the northeast coast of England, two ecotypes of *L. saxatilis* inhabit different levels of the shore (Hull *et al.*, 1996; Wilding *et al.*, 2001). In the upper-shore, on cliff walls, there is a small ecotype with thin shell and wide aperture (H ecotype, High-shore). This habitat is dominated by strong wave action, and in these conditions, a wide aperture and correspondingly large foot appear to be adaptations to resisting dislodgement (Grahame & Mill, 1986). In the mid-shore boulder fields, there is a large and robust ecotype (M ecotype, Mid-shore). Boulder movement and crab predation select for a thicker and larger shell (Raffaelli, 1978). Partial assortative mating between ecotypes has been observed in laboratory conditions (Hull, 1998; Pickles & Grahame, 1999). Neutral genetic divergence between the ecotypes is low ($F_{ST} \sim 0.03$; Wilding *et al.*, 2001; Grahame *et al.*, 2006), but an AFLP genome scan suggested that 5% of loci were influenced by selection (Wilding *et al.*, 2001).

Sample collection and RNA extraction

Samples were collected in December 2007 from two shores on the northeast coast of England, Thornwick Bay (0°07'W, 54°08'N) and Old Peak (0°29'W, 54°24'N). These sites have been used in previous studies (Wood *et al.*, 2008; Wilding *et al.*, 2001; Grahame *et al.*, 2006). We collected H and M ecotypes in their characteristic habitats. After sampling, snails were maintained alive in the laboratory in artificial sea water for 48 h to reduce the gut contents. Then, the individuals were sexed and 15 females from each ecotype and sampling site were transferred into RNAlater (Ambion, Austin, TX, USA). Males were not included because they cannot be reliably distinguished from the sister species *Littorina arcana* (Grahame & Mill, 1989; Reid, 1996). Total RNA was extracted from whole bodies (including the embryos from the brood pouch) using TRIzol[®] reagent (Invitrogen, Carlsbad, CA, USA) according to the manufacturer's protocol. RNA quality and quantity was measured using the RNA 6000 Nano Chip Kit with an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA). Low-quality samples, including those with evidence of DNA contamination, were discarded. All the samples were standardized to 200 ng μL^{-1} , and equal volumes of the samples were combined into two pools

each representing 30 individuals (H and M pools). Both pools were DNase-treated with Turbo DNA-free (Ambion) and purified with RNeasy Mini Kit (QIAGEN, Valencia, CA, USA) following the manufacturer's instructions; RNA quality and quantity was determined again at the end of the process.

cDNA synthesis, normalization and 454 sequencing

The two total RNA pools (H and M) were used for double-strand cDNA synthesis using SMART technology (Zhu *et al.*, 2001; BD Biosciences Clontech), and then the cDNA was purified using the QIAquick PCR Purification Kit (QIAGEN). Full-length cDNA was normalized using the duplex-specific nuclease (DSN; Shagin *et al.*, 2002) normalization method (Zhulidov *et al.*, 2004). cDNA synthesis and normalization were performed by Evrogen (<http://www.evrogen.com>, Moscow, Russia). Normalization decreases the prevalence of highly abundant transcripts which should lead to an increase in the number of different transcripts represented in the sequence reads and consequently the number of SNPs available for analysis. However, normalization may also influence allele frequencies: if alleles are sufficiently different to reduce duplex formation between sequences derived from different alleles, normalization will tend to reduce the frequency of the more common allele. If this occurs, it will always tend to reduce apparent allele frequency differences between pools and so make the subsequent outlier analysis conservative. For this reason, we consider the gain in coverage from normalization to outweigh the risk of biasing allele frequencies, in the context of a transcriptome scan.

Each normalized cDNA pool was sequenced in one-half of a PicoTiterPlate™ using the Genome Sequencer FLX System (454 Life Sciences, Branford, CT, USA) (NCBI Sequence Read Archive accession SRA 020871.5) at the NERC-funded Biomolecular Analysis Facility at the University of Liverpool (<http://www.nbaf.nerc.ac.uk/nbaf-liverpool>).

Assembly, SNP discovery and genotyping

Prior to assembly, we used the program SEQMAN NGEN™ v1.2 (DNA*®, Madison, WI, USA) to perform adaptor, SMART primer and poly-A tail trimming and also quality filtering (threshold quality score = 20). The sequence reads obtained from each of the libraries/ecotypes were combined in the assembly, but always keeping track of their origin (H and M ecotypes). Two different software packages were used to perform *de novo* assemblies, SEQMAN NGEN™ v1.2 (DNA*®) and GS *de novo* Assembler v.2.0.00.20 ('Newbler'; 454 Life Sciences), because of the differences observed in their performance (see Results). Then, we used the program GS Reference Mapper (454 Life Sciences) for SNP detection in each one of the assemblies (NGen and Newbler). Assembly and mapping

parameters are shown in Table S1. As described in the Single Nucleotide Polymorphism Database (dbSNP, <http://www.ncbi.nlm.nih.gov/projects/SNP/>), we included single-base substitutions, single-base or small-scale multi-base deletions/insertions, small-scale multi-base substitutions and tandem repeat variation in our analysis. For each SNP detected with GS Reference Mapper, we estimated the allele frequencies for each ecotype (H and M) and removed from the analysis those SNPs with < 4 reads for each ecotype. Then, we calculated the logarithm of odds (LOD) score (Sokal & Rohlf, 1994), with continuity correction ($LOD = \ln(((p_H + 0.5)/(q_H + 0.5))/((p_M + 0.5)/(q_M + 0.5)))$ where p_i, q_i are counts of reads for the two alleles for ecotype i), as a measure of differentiation between the ecotypes for each SNP.

Differential representation of sequence reads between ecotypes

We determined the number of reads from each ecotype pool (H and M) within the assembly for each contig. Despite normalization, large differences in these counts may reflect, in some cases, real biological processes rather than methodological artefacts. In this sense, cases of differential gene expression, alternative splicing or copy number variation might produce differences in the number of reads per contig. However, in the absence of a clear expectation for the effect of normalization, we cannot apply formal tests for significant differences in numbers of reads and observed differences must be interpreted with caution. To detect transcripts with the greatest differences in representation, we applied an arbitrary cut-off of at least 15 reads in total and then we looked for departures from an expectation based on the total number of reads for each ecotype (Newbler: 52.9% M, 47.1% H; NGen: 53.8% M, 46.2% H) using a χ^2 test with sequential Bonferroni correction for the number of contigs tested. Contigs meeting our criteria were included for annotation. These contigs represent a start point for further studies trying to determine the origin of differential representation between the ecotype pools.

Outlier detection

We used a modified version of the program WINKLES (Wilding *et al.*, 2001), for the analysis of our SNP data set. The new version, WINKLES454 (available upon request), simulates a two-island symmetrical model at migration, mutation and drift equilibrium. Following Wilding *et al.* (2001), effective population size was set to 1000 and mutation rate to 0.0001. These parameters do not have major effects on genetic differentiation under the conditions of high gene flow (Beaumont & Nichols, 1996). The program simulated 5000 bi-allelic loci. An initial stage of panmixia was followed by 10 000 generations of reduced gene flow ($F_{ST} \sim 0.03$; Wilding *et al.*, 2001), then a sample of alleles equivalent to that used in the cDNA

pool (i.e. 60 alleles) was extracted from each population (ecotype) and for each locus. To simulate the 454 sequencing process, these alleles were re-sampled, with replacement (because multiple reads may be present in the 454 data from a single allele), to give the number of reads observed for each SNP and each ecotype in our data set. LOD values were then calculated for the simulated loci. This process was repeated in ten successive cycles separated by 500 generations to avoid allele frequency correlations between samples. The whole simulation was repeated 100 times to generate a total of 1000 simulated LOD values for each locus, based on the same numbers of reads as the observed data. Then, probabilities were calculated by comparing simulated and empirical |LOD| values. Finally, *P*-values were adjusted for multiple testing using the sequential goodness of fit test (SGoF) correction (Carvajal-Rodríguez *et al.*, 2009, SGoF program available at <http://webs.uvigo.es/acraaj/SGoF.htm>). The SGoF adjustment controls for the familywise error rate. It is particularly suited to analyses, like ours, where a large number of tests have been conducted; *P*-values are known imprecisely and departures from the null hypothesis are not expected to be large (see Carvajal-Rodríguez *et al.*, 2009 for a comparison with other multiple test corrections). All SNPs with |LOD| values greater than 95% of simulated values were considered potential outliers, and contigs containing these SNPs were annotated where possible.

For any one locus, representation of alleles in the ecotype-specific pool may be influenced by variation in the total amount of RNA contributed by each individual (despite attempts to keep contributions equal) or by differences in expression level among individuals. This source of variance is not accounted for by the WINKLES454 simulations. Therefore, we have conducted a separate set of simulations to assess its likely impact. We simulated population allele frequencies from 0.1 to 0.5, drew a sample of individuals (10 or 60) assuming Hardy–Weinberg equilibrium within ecotype, formed an RNA pool with a log-normal distribution of contributions from the sampled individuals ($\ln(x)$ where x is normally distributed with mean zero and SD varied from 0 to 2) and then sampled 10 or 60 reads from the pool. This process was repeated 1000 times for each parameter combination, and the means and standard deviations of the output allele frequencies were calculated. The simulation was implemented in GENSTAT v.10 (VSN International, Hemel Hempstead, UK).

Annotation

Contigs containing potential outlier SNPs ($P < 0.05$; without multiple test correction) and contigs with differential representation of sequence reads between ecotypes were used for Blastn and Blastx searches and GO (gene ontology; The Gene Ontology Consortium, 2000) annotation. The Netblast 2.2.12 client ([\[www.ncbi.nlm.nih.gov/BLAST/download.shtml\]\(http://www.ncbi.nlm.nih.gov/BLAST/download.shtml\)\) was used to perform Blastn searches \(E-value cut-off, \$1E-5\$ \) against the nonredundant \(nr\) nucleotide database at NCBI. We also performed more restricted searches within the Phylum Mollusca. An initial screen revealed some sequences that were not derived from *L. saxatilis* \('xenobiotics'\). The program Blast2GO \(Conesa *et al.*, 2005\) was used for Blastx searches \(E-value cut-off, \$1E-5\$ \) against the NCBI nr protein database and Swiss-Prot database and to extract the GO terms associated with the Blast hits. The annotation parameters were pre-E-value-Hit Filter \(\$10^{-6}\$ \), annotation cut-off \(55\) and GO weight \(5\). Directed acyclic graphs were also generated in Blast2GO using default parameters, and GO assignment \(level 2\) results for the three principal GO categories \(Biological Process, Molecular Function and Cellular Component; see The Gene Ontology Consortium\) were obtained. Contigs containing nonoutlier SNPs \(\$P > 0.05\$; 'neutral'\) were used to compare the GO categories against the categories obtained with outliers and contigs with differential representation.](http://</p>
</div>
<div data-bbox=)

Results

Assembly

A summary of the results of one GS-FLX 454 sequencing run on normalized cDNA pools of H and M ecotypes of *L. saxatilis* is presented in Table S2. We obtained 58.2 Mb (52.9% M, 47.1% H) of sequence with an average read length of 195 bp. After primer trimming and quality filtering, the total number of reads was reduced to 262 817 (88% of the initial reads) and the average length was 171 bp (see Table S2 for details). Two independent *de novo* assemblies were performed (Newbler and NGen), a summary of the assemblies' statistics is presented in Table S3. Singleton reads were excluded from further analysis. Newbler assembled 65% of the reads into 11 298 contigs and NGen 78% into 33 359 contigs. Average contig length was 343 bp for Newbler and 309 for NGen, and average numbers of reads per contig were 15 and 6, respectively. The NGen assembly had a much greater proportion of contigs with low number of reads/contig (Fig. S1). This indicates that in our dataset, and with the parameters we used, NGen generates a more stringent assembly than Newbler.

Differential representation of sequence reads between ecotypes

The number of contigs with differential representation between H and M ecotypes that met our criteria (see Materials and methods), using a nominal cut-off at $P < 0.05$ after sequential Bonferroni correction, was 393 (3.5%) for the Newbler assembly and 335 (1%) for the NGen assembly. In Newbler, 208 contigs showed over-expression in the M ecotype and 185 contigs in the H

ecotype. In NGen, the numbers of contigs were 160 and 175, respectively. When the nominal cut-off was reduced to $P < 0.01$, the number of contigs was 309 for Newbler and 238 for NGen.

SNP analysis

The total numbers of SNPs detected were 11 305 (3126 contigs, 3.6 SNP/contig on average) and 7755 (3598 contigs, 2.2 SNP/contig) for Newbler and NGen assemblies, respectively (Table 1). Transitions were the most common SNP (see Table 1). The average total depths of coverage per SNP were 8.3 for Newbler and 4.9 for NGen. The numbers of reads per SNP were similar, on average, for the two ecotypes.

Outlier detection

Those contigs that had Blastn matches (E-value $< 1E-5$) with non-Eumetazoan sequences (xenobiotics, mainly gut bacterial sequences), or with *Littorina* ribosomal sequences, were removed from the analysis. We also removed those SNPs with < 4 reads for either of the ecotypes. The final number of SNPs analysed was 2454 (572 contigs) for Newbler but many fewer for NGen (510 from 197 contigs) because of the lower average coverage (Table 1). We performed Fisher's exact tests to provide an initial, assumption-free comparison of allele frequency differences between ecotypes, 186 SNPs were significant

($P < 0.05$) for Newbler (122.7 expected) and 41 for NGen (25.5 expected). This result suggests that highly differentiated SNP cannot be explained by chance alone. However, the null hypothesis tested allows for neither population history nor the lack of independence between reads. Therefore, for outlier locus detection, we used a simulation approach. We calculated the LOD value for each SNP and determined significance by comparison of absolute LOD scores with simulated distributions (Fig. 1). We found approximately 12% of SNPs to be outliers ($P < 0.05$) and approximately 7% to be significant after multiple test correction (SGoF, experiment-wide $P < 0.05$). These proportions were similar between assemblies, but the number of contigs with outliers differed between assemblies (Table 2). For the outlier contigs, the average length and depth per contig were close to 700 bp and 15 reads. In Fig. S2, we show an example of the aligned sequences for a contig containing one outlier SNP.

Our analysis of the impact of variation in the contributions of individuals to the sequencing pools showed that it makes only a small contribution to the uncertainty in allele frequency estimation, compared to the effects of sampling individuals and reads (Fig. 2). This is true even when the coefficient of variation in RNA amount is as high as 3. The reason for the small effect is that variation in RNA contribution is not correlated with genotype.

We were also interested to examine the distribution of the number of outlier loci within individual contigs

Table 1 Summary of the results for the SNP detection performed with gs Reference Mapper for Newbler and NGen assemblies. SNPs were classified into four categories, and the number of SNPs, number of reads and average depth of coverage (number of reads) are shown for each category. Total counts and average values are also presented. '% M' represents the percentage of reads that belong to the *Littorina saxatilis* M ecotype. These results are shown for the initial SNP discovery (Total) and after removing those SNPs with less than four reads for each ecotype and those SNPs with Blastn matches (E-value $< 1E-5$) with *Littorina* rRNA or non-Eumetazoan sequences using the nr database at NCBI (Filtered).

	Transition	Transversion	Indel	Multiple base substitution	Total
Newbler					
Total					
<i>n</i> SNPs (%)	4515 (39.9)	2877 (25.4)	2241 (19.8)	1672 (14.9)	11305
<i>n</i> reads (% M)	39360 (49.5)	24489 (49.8)	16002 (48.6)	13758 (49.6)	93609 (49.5)
Average depth	8.7	8.5	7.1	8.2	8.3
Filtered					
<i>n</i> SNPs (%)	1055 (43.0)	685 (27.9)	347 (14.1)	367 (15.0)	2454
<i>n</i> reads (% M)	15686 (50.1)	10276 (50.7)	5025 (50.1)	5430 (49.7)	36417 (50.1)
Average depth	14.9	15.0	14.5	14.8	14.8
NGen					
Total					
<i>n</i> SNPs (%)	3343 (43.1)	2106 (27.2)	1420 (18.3)	886 (11.4)	7755
<i>n</i> reads (% M)	17693 (50.5)	11242 (49.5)	5575 (50.1)	3947 (48.7)	38457 (49.9)
Average depth	5.3	5.3	3.9	4.5	4.9
Filtered					
<i>n</i> SNPs (%)	264 (51.8)	171 (33.5)	34 (6.7)	41 (8.0)	510
<i>n</i> reads (% M)	3414 (53.1)	2272 (52.5)	477 (51.8)	534 (46.3)	6697 (52.3)
Average depth	12.9	13.3	14.0	13.0	13.3

SNP, single nucleotide polymorphism; nr, nonredundant.

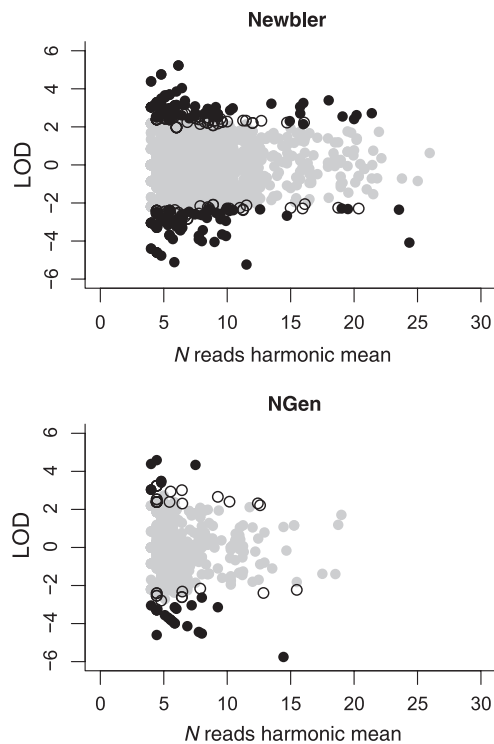


Fig. 1 Single nucleotide polymorphism (SNP) divergence between the M and H ecotypes of *Littorina saxatilis* for Newbler (2454 SNPs) and NGen (510 SNPs) assemblies. Logarithm of odds values are plotted against the harmonic mean of the number of reads per ecotype. The results of the outlier detection analysis are also included: grey closed circles are nonoutlier SNPs, open circles are outliers ($P < 0.05$), black closed circles are outliers after SGoF correction.

(Fig. 3a, Newbler assembly) and the relationship between the numbers of outliers ($P < 0.05$, without correction) and the total number of SNPs within each contig (Fig. 3b). It is clear that contigs containing multiple outlier SNPs are more common than expected from the overall frequency of outliers. Contigs containing a high proportion of outlier SNPs are considered good candidates for local adaptation and so for further study.

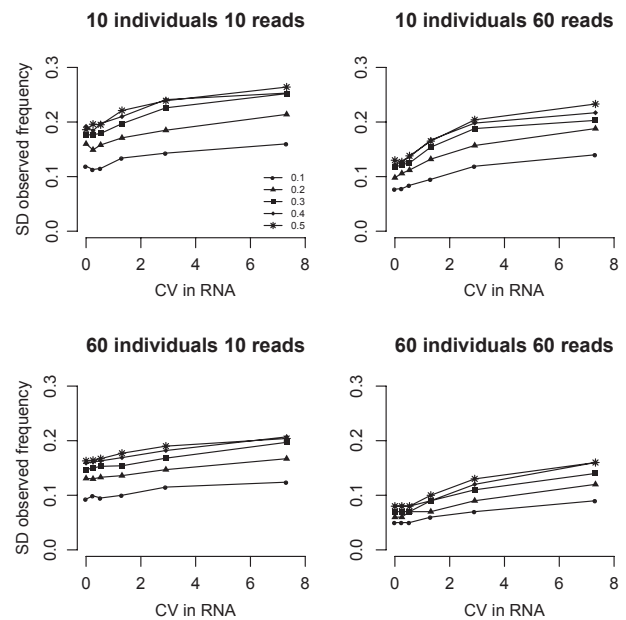


Fig. 2 Standard deviation in the single nucleotide polymorphism allele frequency inferred from simulated 454 reads in relation to the variation in RNA contribution to a sequencing pool. The lines on each plot are for different true allele frequencies. 'CV in RNA' is the coefficient of variation in RNA amount among the pooled individuals.

Annotation

A total of 187 contigs from the Newbler assembly and 47 for NGen contained at least one potential outlier SNP ($P < 0.05$, without correction) and were annotated using Blast2GO. Contigs with differential representation above our cut-off were also annotated: 393 and 335 contigs for the two assemblies, respectively. Table S4 shows the GO annotation (level 2), against Swiss-Prot, for the predominant GO categories. None of the comparisons between 'outlier', 'differential representation' and 'neutral' contigs showed a significant difference in GO categories.

The number of Blastx hits (E-value $< 1E-5$) in the nr protein and Swiss-Prot searches was 30 and 6, respec-

Table 2 Results of the outlier detection analysis using WINKLES454 for Newbler and NGen assemblies. Two different levels of significance were used: $P < 0.05$ and SGoF_{0.05} correction. We present the number of outlier SNPs and outlier contigs, total number of bases for outlier contigs, average length and number of reads per contig. Blastx search results against nr protein (NCBI) and Swiss-Prot databases for the outlier contigs are also shown (note that the majority of contigs with Swiss-Prot hits also have hits in the nr database).

Outlier	N SNPs (%)	N contigs (%)	N bases	Average length (bp)	Average N reads	Blastx hit nr protein	Blastx hit Swiss-Prot
Newbler							
$P < 0.05$	312 (12.7)	187 (32.7)	196145	629	14.9	29	24
SGoF _{0.05}	179 (7.3)	119 (20.8)	113028	631	15.0	20	18
NGen							
$P < 0.05$	65 (12.7)	47 (23.8)	48778	750	12.9	6	3
SGoF _{0.05}	31 (6.1)	19 (9.6)	22586	753	12.8	3	1

SNP, single nucleotide polymorphism; SGoF, sequential goodness of fit; nr, nonredundant.

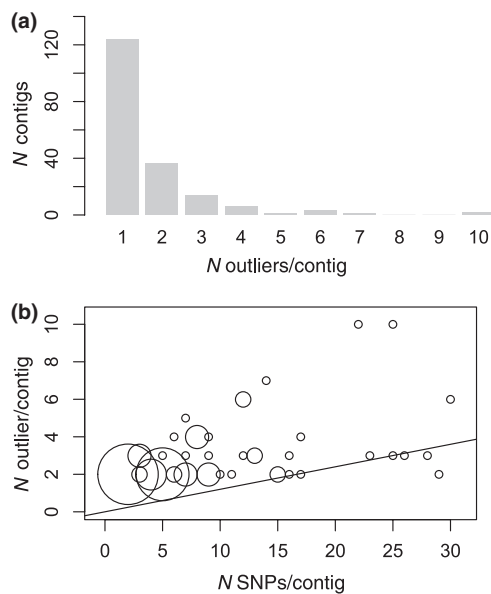


Fig. 3 (a) Histogram showing the number of outliers per contig for the Newbler assembly. Contigs with zero outlier single nucleotide polymorphism (SNP) (2142 contigs) are omitted. (b) The number of outlier SNPs per contig in relation to the total number of SNPs detected per contig in the analysis (Newbler). Contigs with zero or one outlier are omitted. Circle diameter is proportional to the number of outliers. The line shows the expectation of 12% of outliers, based on their overall frequency.

tively (Table 2). Results where the ‘sequence description’ was ‘hypothetical protein’ were not included. Seventeen of the potential outlier loci with Blastx hits were significant outliers after SGoF correction for the Newbler assembly and 2 for NGen (Table 3). Annotated contigs containing outliers that were significant after multiple test correction contained more than one potential outlier in 5 of 17 cases for Newbler and one of two in NGen (see Table 3). From the Newbler assembly, two additional contigs, annotated as NADH-dehydrogenase subunit 4 and selenoprotein precursor (Table 3), contained more than one individually significant outlier but none of these was significant after multiple test correction. These loci may also be worthy of follow-up studies together with the SGoF outliers. In the NGen assembly, this analysis did not find any additional candidates.

Blastx results for the contigs showing differential representation of reads between the ecotypes are presented in Table S5 for Newbler and Table S6 for NGen. The numbers of significant hits were 39 and 24, respectively. Again, ‘hypothetical proteins’ were omitted. Tables S5 and S6 also show the number of reads per contig for each ecotype (H and M). No contig was identified by both the outlier SNP and expression analyses (partly because the outlier analysis required at least four reads from each ecotype). However, contigs with Blast matches to titin and mucin were included in

both annotations and could be from the same loci (see Discussion).

Xenobiotics

In the Newbler assembly, we found 15 different bacterial rRNA matches (7.9 SNPs/contig and 22 reads/SNP on average). In NGen, we found two contigs matching bacterial rRNA. Some of these bacterial contigs showed marked SNP frequency differences (high |LOD| values) between the ecotypes (Table S7). We also found two contigs with marked differences in the number of reads between ecotypes and with interesting matches: an uncultured bacterium in the class Bacteroidetes (Contig-10401, Acc. No. EF123624, E-value = $1E-47$) with 34 reads in the H ecotype and none for the M and a ciliate protozoan *Cyclidium glaucoma* (Contig-10717, Acc. No. DQ442839, E-value = 0) with 28 reads in H and none in M.

Repetitive elements

Blastn searches (nr database, NCBI) also resulted in Blastn matches (E-value < $1E-5$) against *L. saxatilis* sequences from four BAC clones sequenced by Wood *et al.* (2008) (Table S8). We found 21 contigs containing outlier SNPs (15 significant after SGoF correction) that matched the BAC clones. Most of these contigs had matches in more than one BAC (Table S8); mean match length was 215 bp. These regions therefore appear to be repetitive elements that are widespread across the genome but may be differentiated between ecotypes.

Discussion

Ecological speciation and genome scans

The population genomics approach has been widely used in recent years as a strategy in the study of the genetic basis of local adaptation and ecological speciation (Nosil *et al.*, 2009; Storz, 2005). ‘Next-generation’ sequencing technologies can potentially increase the power of genome scans, ultimately through re-sequencing of whole genomes. Turner *et al.* (2010) showed how genes involved in local adaptation in *Arabidopsis lyrata* can be detected in this way when a reference genome is available. However, one of the advantages of genome scans is that they can be applied to organisms with well-characterised ecology that lack genomic resources. In such cases, EST-based genome scans are more informative, because they focus on coding regions (Bonin, 2008). Recent studies have shown how 454 pyrosequencing of the transcriptomes of divergent populations can be used to detect candidate genes for adaptation (*Coregonus* spp., Renaud *et al.*, 2010; *Rhagoletis pomonella*, Schwarz *et al.*, 2009). Neither study used simulations to determine the expected distribution of SNP frequency

Table 3 Blastx results for outlier loci ($P < 0.05$) for Newbler and NGen assemblies. Those remaining significant after multiple test correction (SGoF) are indicated. Number of SNPs detected for each contig and number of outlier SNPs (uncorrected $P < 0.05$) per contig are also included. Matches to the nr protein (NCBI) and Swiss-Prot databases are indicated, with accession numbers (Acc. No.) for the best hit and their E-values.

Contig	M test	N SNPs	N outlier SNPs	nr	Swiss-prot	Sequence description	Length (bp)	Acc. No.	E-value
Newbler									
Contig-03815	SGoF	1	1	X	X	Acyl-CoA-binding protein	352	P82934	1.72E-24
Contig-03328		1	1	X	X	Alpha actinin	820	XP 972324	4.80E-115
Contig-03312	SGoF	2	2	X	X	Arginine kinase	410	P51544	2.82E-35
Contig-10406	SGoF	1	1	X	X	CD109	774	XP 419879	7.25E-22
Contig-11152	SGoF	2	1	X	X	Chymotrypsin inhibitor	631	XP 002517007	1.09E-10
Contig-10074	SGoF	7	1	X	X	Dermatopontin	1355	ACH48240	1.05E-34
Contig-00793	SGoF	5	2	X	X	Endonuclease reverse transcriptase	910	XP 799307	1.21E-40
Contig-00110	SGoF	2	1	X	X	Fibrillin	953	Q61554	1.65E-07
Contig-11253	SGoF	7	2	X	X	Lithostathine	810	ABO26661	1.87E-16
Contig-09772	SGoF	9	4	X	X	Myosin heavy chain	794	AAD13782	5.60E-99
Contig-00829	SGoF	6	1	X	X	NADH dehydrogenase subunit 1	947	CAA10597	5.20E-129
Contig-02751		15	2	X	X	NADH dehydrogenase subunit 4	1535	YP 492550	2.60E-117
Contig-03280		5	1	X	X	NADH dehydrogenase subunit 4	894	XP 001990826	3.02E-12
Contig-10350	SGoF	4	2	X	X	NADH dehydrogenase subunit 6	207	CAM58062	7.37E-24
Contig-00262		7	1	X		Pol-like protein	1259	BAC82624	3.82E-07
Contig-00850		2	1	X		Retinitis pigmentosa GTPase regulator	959	NP 850032	1.12E-07
Contig-00890		4	1	X		Reverse transcriptase-like protein	503	XP 001190798	2.49E-22
Contig-10674	SGoF	1	1	X	X	Reverse transcriptase-like protein	381	XP 002160419	1.09E-19
Contig-00024		2	1	X	X	Ribosomal protein L39	372	XP 002400257	2.94E-20
Contig-10433		11	2	X	X	Selenoprotein precursor	1337	A8YXY3	3.66E-33
Contig-01851	SGoF	1	1	X	X	Tectorin alpha precursor	963	Q9YH85	1.16E-16
Contig-11254	SGoF	1	1		X	Titin	201	Q8WZ42	6.88E-13
Contig-02371	SGoF	1	1	X	X	Transmembrane protein 59 precursor	645	ACI33199	1.04E-27
Contig-01496	SGoF	4	1	X	X	Ubiquinol-cytochrome c reductase	791	P00130	1.27E-12
Contig-01407	SGoF	2	1	X		Zinc finger protein	475	XP 001201671	1.52E-13
NGen									
Contig-15375		5	1	X	X	Arginine kinase	517	BAB41095	8.99E-21
Contig-27612	SGoF	8	7	X	X	Mucin	364	XP 002225487	5.03E-09
Contig-02103		5	1	X	X	NADH dehydrogenase subunit 4	1633	YP 492550	2.40E-124
Contig-26142		2	1	X		Pol-like protein	996	EER05122	1.62E-09
Contig-27784	SGoF	2	1	X		Reverse transcriptase-like protein	393	XP 002160419	8.50E-20

SNP, single nucleotide polymorphism; nr, nonredundant.

differences, which we consider important because of both population genetic effects and the sampling process during sequencing.

Our study was based on approximately 60 Mb of sequence (compared to the approximately 500 Mb that can be obtained with the recent GS-FLX Titanium Series), but we were able to perform a *de novo* assembly and SNP detection, and this procedure provided a large number of potentially informative markers. The two assemblers performed quite differently with the settings we used, the more stringent NGen assembly providing around three times more contigs but with fewer reads per contig than for Newbler. For the genome scan analysis, we were primarily interested in contigs with high coverage and the Newbler assembly was preferable in this respect. Although the lower stringency increases the risk of including reads from paralogous loci, there is no obvious reason why this should create false-positive outliers. Despite lower coverage per SNP on average, the

NGen assembly did allow the identification of some additional outliers. We expect this sensitivity to differences between assembly packages and settings to be lower for data sets with more and longer reads.

Indexing of cDNA from individuals before 454 sequencing provides an alternative to pooling but it significantly increases costs and restricts the number of individuals that can be analysed. Where the primary question of interest is to detect frequency differences between populations, pooling of large samples of individuals may continue to be the preferred strategy (see Futschik & Schlötterer, in press). Our simulations (Fig. 2) show that variation in RNA amounts among individuals is not a serious problem but indicate that the number of reads per SNP should, ideally, be higher than in our current data set. This will be easily achieved with the current sequencing technology.

Normalization is another potential source of variation in SNP frequencies. However, as argued earlier, it is

unlikely to generate marked frequency differences between pools. It has the advantage of evening out the representation of different transcripts and so increasing the number with sufficient coverage for analysis but the disadvantage of removing information about variation in levels of expression among transcripts, within pools. However, transcripts with large expression differences between pools may still be differentially represented in the 454 reads. Therefore, we have identified those transcripts with the greatest differences in representation as targets for future study but do not make any further inferences about them.

SNPs detected using 454 sequencing of pooled cDNA should be validated by direct genotyping from genomic DNA of individuals. Sequencing errors may generate false SNPs, and allele frequencies may not be estimated accurately from pools, as illustrated by comparisons between read frequencies and direct genotyping (e.g. Van Tassel *et al.*, 2008). In our study, sequencing errors may contribute to the total number of SNPs detected but are very unlikely to generate outliers, with strong frequency differences between pools. Our test for outliers incorporates the major sources of variation in SNP frequencies, although clearly not all sources. Nevertheless, the outliers detected will include false positives (and also exclude many loci that are, in fact, influenced by selection) for these experimental reasons as well as the limitations of the demographic model discussed in the following paragraphs. Therefore, we emphasize that the objective of this 454 genome scan approach, as with any such methodology, is to identify candidate loci for further study rather than to provide strong evidence for selection influencing individual loci.

Divergent ecotypes of *L. saxatilis* provide an example of local adaptation and partial reproductive isolation (see Johannesson *et al.*, 2010 for a discussion of repeated evolution of ecotypes, Rolán-Alvarez, 2007 for a review, Sadedin *et al.*, 2009 for a model). Previous genome scans with AFLPs have detected a small proportion of markers, 2–5% outlier loci, with greater than neutral genetic differentiation (Galindo *et al.*, 2009; Wilding *et al.*, 2001). Our EST-based genome scan showed a similar pattern, with a small proportion of the markers showing greater than neutral differentiation (outlier loci), around 7% when applying a multiple test correction. A recent review by Nosil *et al.* (2009) on divergent selection and the extent of genomic divergence during population differentiation and/or speciation points out that the percentage of the genome apparently affected by divergent selection (% of outlier loci) lies in the range of 5–10% in most of the studies reviewed (Nosil *et al.*, 2009 and references therein). However, they remark that the results of different genome scans should be compared with caution because of variation among the analyses (number of populations, type of molecular markers, methodology and level of significance). Population structure is also an important variable to take into

account (Excoffier *et al.*, 2009). It is clear that false positives (and false negatives) are a feature of all genome scan studies (Hermisson, 2009). Therefore, the percentage of outliers detected in a genome scan is not the most important outcome, and individual outliers should be treated as candidates, in need of further investigation into the effects of natural selection. In studies of ecological speciation, a genome scan is one of the first steps towards the genetics of divergent adaptation, rather than an end in itself.

Outlier loci

Outlier loci may be the direct targets of selection, they may be regions tightly linked to selected loci or they may be false positives (Nosil *et al.*, 2009). An advantage of using ESTs is that a proportion of loci can be functionally annotated, and this may reveal loci likely to be associated with traits under selection. However, the annotation step may be problematic for 454 data from nonmodel species. In this study, we were able to annotate only 16% of the contigs with outlier SNPs (30 of 187 contigs, Newbler). This is for several reasons: contigs are often short, they may contain primarily untranslated regions, the Mollusca are not well represented in sequence databases and the molluscan sequences that are available are not themselves well annotated.

A potential confounding factor in using cDNA is the possibility of allele-specific expression (e.g. Pant *et al.*, 2006). As pointed out by Renaut *et al.* (2010), if different alleles are preferentially expressed in the populations being compared, false outliers may be generated. This would actually represent an interesting form of divergence and should be considered in follow-up studies of EST outliers.

In the case of the ecotypes of *L. saxatilis*, we expected, *a priori*, that the outlier loci should be involved or linked to genes involved in local adaptation and assortative mating (Butlin *et al.*, 2008; Johannesson *et al.*, 2010). Shell size represents an important variable in local adaptation (wave exposure and crab predation) and mate choice, but also shell shape and the foot muscle that attaches the snail to the substrate play an important role in adaptation to withstand wave exposure (see Rolán-Alvarez, 2007 for a review). Thus, attention should be directed towards genes related to shell formation, muscle physiology and energy metabolism. Because the females used in our study carried developing embryos in their brood pouches and we used all tissues in our RNA extractions, there is the potential for genes involved in all of these processes to be included in our data. Here, we briefly discuss outliers that stand out as promising targets for future investigation on the basis of their annotation.

Skeletal matrix proteins

Shells in *L. saxatilis* are composed of calcite and aragonite (Taylor & Reid, 1990), with an associated organic matrix

that is thought to be involved in shell formation and so influences the properties of the shell (e.g. size, shape) (Gunthorpe *et al.*, 1990). Some of the SGoF outliers showed matches with skeletal matrix proteins including lithostathine, mucin and dermatopontin. Mucin was also detected in the differential representation analysis (see Tables S5 and S6). Lithostathine is a C-type lectin-like protein, which plays an important role in calcium carbonate biomineralization in a wide variety of organisms (Matsubara *et al.*, 2008 and references therein). Mucins are heavily glycosylated proteins, and there is evidence that they have a role in molluscan shell calcification (Marin *et al.*, 2000). The fact that mucin also stands out in the differential representation analysis could be because members of the mucin gene family are differentially expressed between the ecotypes. Dermatopontin is considered a major component of the shell matrix proteins in molluscs (Marxen *et al.*, 2003).

Skeletal muscle proteins

The muscular foot size of *L. saxatilis* is associated with the level of wave exposure (Grahame & Mill, 1986). Contigs with matches to myosin and titin were SGoF outliers, and titin was also differentially represented in the two pools. Twitchin is a titin-like protein in molluscan smooth muscle, and it is involved in the 'catch' contraction in molluscs, a unique energy-saving contraction (reviewed in Funabara *et al.*, 2005). Molluscan catch muscle can maintain tension for a long time with little energy consumption, and twitchin interacts with myosin in this contraction (Funabara *et al.*, 2001).

Energetic metabolism

Energetic metabolism is likely to vary with environmental stressors (e.g. temperature, anoxia, wave action) that differ between tidal levels. Outlier loci (SGoF) matched genes involved in energetic metabolism, arginine kinase (ARK) and NADH dehydrogenase (NADH-dh) and were detected as outliers in both assemblies (Newbler and NGen). ARK regulates the availability of ATP in cells involved in metabolic work (Morrison, 1973). ARK has shown evidence for local adaptation in previous studies of *Littorina* spp.: *L. fabalis* (Johannesson & Mikhailova, 2004; Tatarenkov & Johannesson, 1994), *L. obtusata* (Schmidt *et al.*, 2007) and *L. saxatilis* (Martínez-Fernández *et al.*, 2008). The hypothesis in these studies is that certain ARK alleles provide a faster supply of ATP for contracting the foot muscle in wave-exposed habitats. NADH dehydrogenase subunits are encoded by the mitochondrial genome, and they are involved in the respiratory chain. Natural selection may affect genes within the mtDNA (Ballard & Whitlock, 2004 and references therein). It is also possible that these contigs are false positives caused by transcribed nuclear pseudogenes derived from mtDNA (Bensasson *et al.*, 2001). However, previous studies of

littorinid mtDNA have not detected nuclear copies (Wilding *et al.*, 2000).

Reverse transcriptases

In eukaryotes, reverse transcriptases are responsible for the movement of transposable elements (TEs). TEs account for a major fraction of eukaryotic genomes (Feschotte *et al.*, 2002 and references therein), and they are able to modify gene expression and promote genome evolution (reviewed in Gogvadze & Buzdin, 2009). As the genome of *Littorina* is large (estimates range from 0.81 to 1.47 pg, $0.79\text{--}1.35 \times 10^9$ bp, <http://www.genomesize.com/index.php>), it is likely to contain many TEs and it is not surprising that reverse transcriptase transcripts occur in our cDNA pools. They may occur as outliers because different gene copies are combined within contigs.

Repetitive elements

Repetitive elements such as TEs have already been described in *L. saxatilis* (see Wood *et al.*, 2008). In a previous study (Wilding *et al.*, 2001), an AFLP genome scan discovered outlier loci between H and M ecotypes of *L. saxatilis*, two of which (E10 and E12) were associated with putative TE insertions (Wood *et al.*, 2008). At the same time, BAC clones containing these outlier loci and another two BAC clones containing neutral AFLP loci (A30 and A37) were sequenced and found to contain additional partial copies of these inserted sequences (Wood *et al.*, 2008). In our study, 21 contigs with outlier SNPs matched sequences within these BACs, in most cases matching sequences in several BACs at the same time. This result suggests that these repetitive elements must be very widespread across the *Littorina* genome because the sequenced BACs represent such a small proportion of the total. It is difficult to interpret their 'outlier' status because the expectation against which SNP frequency divergence was compared is not appropriate for repetitive elements. Nevertheless, it is intriguing that repetitive elements appear to differ between ecotypes. In some cases, these elements might play a role in gene regulation, and they could represent a mechanism for adaptation to changing environments through genetic novelty (reviewed in Gogvadze & Buzdin, 2009). However, the number of studies of TEs remains small relative to their ubiquity and abundance in eukaryote genomes, and more studies are needed to address their role in adaptation. Next generation sequencing will play an important role in this respect.

Xenobiotics

We found several contigs within our data set that have a bacterial origin, which was not unexpected because we used whole snails, including the digestive gland. For example, Vera *et al.* (2008) also found microbial

sequences in their butterfly data set, including an intracellular parasite. In our case, these microbial sequences are interesting because they reflect characteristics of the environment. *Littorina saxatilis* feeds on the epilithic biofilm of diatoms, cyanobacteria and bacteria (Norton *et al.*, 1990). Microgradients of light, moisture, temperature and wave exposure are common in the rocky intertidal habitats, and this determines the distribution, diversity and abundance of algae and invertebrates (Menge & Branch, 2001). Thus, we might expect the resources available to differ between ecotypes. Some of the contigs identified showed marked differences either in SNP frequency or in abundance of reads between the ecotypes, and most of these had Blastn matches with cyanobacteria (see Table S7). Our results suggest that the microbial biofilm may differ between tidal levels, and this provides another, previously unrecognized, habitat axis that may contribute to the divergent selection that operates on *Littorina* populations, as well as on other marine gastropods.

Sequences matching the bacterivorous ciliate *C. glaucoma* were also identified within the contigs with reads present only in the H ecotype. Ciliates can be observed inside of the brood pouch of female *L. saxatilis*, in contact with the developing embryos (personal observation), and are likely to be this species. We do not know whether they have negative effects on the reproductive success or what proportion of each ecotype is infested. These two questions should be addressed in future studies.

Conclusions

This study shows how high-throughput sequencing can help in performing EST-based genome scans for the study of local adaptation and ecological speciation, without the need for extensive genomic resources in the study species. The outlier loci detected could be followed up through SNP genotyping assays, development of EPIC (exon-primed intron-crossing) markers or INDEL (insertion deletion) markers. These studies will test the results obtained in this study and will help to unravel the history and ecological diversification of other *L. saxatilis* populations, ecotypes or other littorinid species.

Acknowledgments

We thank from the staff of the NERC Biomolecular Analysis Facility in Liverpool, especially Margaret Hughes, for carrying out the sequencing and Kevin Ashelford for assistance with initial data analysis. Jon Slate and Robert Ekblom were involved in many discussions about next-generation sequencing analysis. Carole Smadja, Jon Slate and anonymous referees provided useful comments on earlier versions of this manuscript. This work was funded by grants from the European Commission (FP6-042815) and NERC (NE/G018375/1).

References

- Ballard, J.W.O. & Whitlock, M.C. 2004. The incomplete natural history of mitochondria. *Mol. Ecol.* **13**: 729–744.
- Beaumont, M.A. & Nichols, R.A. 1996. Evaluating loci for use in the genetic analysis of population structure. *Proc. R. Soc. Lond. B Biol. Sci.* **263**: 1619–1626.
- Bensasson, D., Zhang, D.X., Hartl, D.L. & Hewitt, G.M. 2001. Mitochondrial pseudogenes: evolution's misplaced witnesses. *Trends Ecol. Evol.* **16**: 314–321.
- Bonin, A. 2008. Population genomics: a new generation of genome scans to bridge the gap with functional genomics. *Mol. Ecol.* **17**: 3583–3584.
- Butlin, R.K. 2010. Population genomics and speciation. *Genetica* **138**: 409–418.
- Butlin, R.K., Galindo, J. & Grahame, J.W. 2008. Sympatric, parapatric or allopatric: the most important way to classify speciation? *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **363**: 2997–3007.
- Carvajal-Rodríguez, A., de Uña-Alvarez, J. & Rolán-Alvarez, E. 2009. A new multitest correction (SGoF) that increases its statistical power when increasing the number of tests. *BMC Bioinformatics* **10**: 209.
- Conesa, A., Götz, S., García-Gómez, J.M., Terol, J., Talon, M. & Robles, M. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**: 3674–3676.
- Ellegren, H. 2008. Sequencing goes 454 and takes large-scale genomics into the wild. *Mol. Ecol.* **17**: 1629–1631.
- Elmer, K.R., Fan, S., Gunter, H.M., Jones, J.C., Boekhoff, S., Kuraku, S. & Meyer, A. 2010. Rapid evolution and selection inferred from the transcriptomes of sympatric crater lake cichlid fishes. *Mol. Ecol.* **19**: 197–211.
- Excoffier, L., Hofer, T. & Foll, M. 2009. Detecting loci under selection in a hierarchically structured population. *Heredity* **103**: 285–298.
- Feschotte, C., Jiang, N. & Wessler, S.R. 2002. Plant transposable elements: where genetics meets genomics. *Nat. Rev. Gen.* **3**: 329–341.
- Funabara, D., Kinoshita, S., Watabe, S., Siegman, M.J., Butler, T.M. & Hartshorne, D.J. 2001. Phosphorylation of molluscan twitchin by the cAMP-dependent protein kinase. *Biochemistry* **40**: 2087–2095.
- Funabara, D., Kanoh, S., Siegman, M.J., Butler, T.M., Hartshorne, D.J. & Watabe, S. 2005. Twitchin as a regulator of catch contraction in molluscan smooth muscle. *J. Muscle Res. Cell Motil.* **26**: 455–460.
- Futschik, A. & Schlötterer, C. in press. Massively parallel sequencing of pooled DNA samples – the next generation of molecular markers. *Genetics*, genetics.110.114397.
- Galindo, J., Morán, P. & Rolán-Alvarez, E. 2009. Comparing geographical genetic differentiation between candidate and noncandidate loci for adaptation strengthens support for parallel ecological divergence in the marine snail *Littorina saxatilis*. *Mol. Ecol.* **18**: 919–930.
- Gogvadze, E. & Buzdin, A. 2009. Retroelements and their impact on genome evolution and functioning. *Cell. Mol. Life Sci.* **66**: 3727–3742.
- Grahame, J. & Mill, P.J. 1986. Relative size of the foot of two species of *Littorina* on a rocky shore in Wales. *J. Zool. (Lond.)* **208**: 229–236.

- Grahame, J. & Mill, P.J. 1989. Shell shape variation in *Littorina saxatilis* and *Littorina arcana*, a case of character displacement? *J. Mar. Biolog. Assoc. UK* **69**: 837–855.
- Grahame, J.W., Wilding, C.S. & Butlin, R.K. 2006. Adaptation to a steep environmental gradient and an associated barrier to gene exchange in *Littorina saxatilis*. *Evolution* **60**: 268–278.
- Gunthorpe, M.E., Sikes, C.S. & Wheeler, A.P. 1990. Promotion and inhibition of calcium-carbonate crystallization invitro by matrix protein from blue-crab exoskeleton. *Biol. Bull.* **179**: 191–200.
- Hermisson, J. 2009. Who believes in whole-genome scans for selection[quest]. *Heredity* **103**: 283–284.
- Hudson, M.E. 2008. Sequencing breakthroughs for genomic ecology and evolutionary biology. *Mol. Ecol. Resour.* **8**: 3–17.
- Hull, S.L. 1998. Assortative mating between two morphs of *Littorina saxatilis* on a shore in Yorkshire. *Hydrobiologia* **378**: 79–88.
- Hull, S.L., Grahame, J. & Mill, P.J. 1996. Morphological divergence and evidence for reproductive isolation in *Littorina saxatilis* (Oliv) in northeast England. *J. Molluscan Stud.* **62**: 89–99.
- Johannesson, K. 2003. Evolution in *Littorina*: ecology matters. *J. Sea Res.* **49**: 107–117.
- Johannesson, K. & Mikhailova, N. 2004. Habitat-related genetic substructuring in a marine snail (*Littorina fabalis*) involving a tight link between an allozyme and a DNA locus. *Biol. J. Linn. Soc.* **81**: 301–306.
- Johannesson, K., Panova, M., Kempainen, P., Rolán-Alvarez, E. & Butlin, R.K. 2010. Repeated evolution of reproductive isolation in a marine snail – unveiling mechanisms of speciation. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **365**: 1735–1747.
- Luikart, G., England, P.R., Tallmon, D., Jordan, S. & Taberlet, P. 2003. The power and promise of population genomics: from genotyping to genome typing. *Nat. Rev. Gen.* **4**: 981–994.
- Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z.T., Dewell, S.B., Du, L., Fierro, J.M., Gomes, X.V., Godwin, B.C., He, W., Helgesen, S., Ho, C.H., Irzyk, G.P., Jando, S.C., Alenquer, M.L.I., Jarvie, T.P., Jirage, K.B., Kim, J.B., Knight, J.R., Lanza, J.R., Leamon, J.H., Lefkowitz, S.M., Lei, M., Li, J., Lohman, K.L., Lu, H., Makhijani, V.B., McDade, K.E., McKenna, M.P., Myers, E.W., Nickerson, E., Nobile, J.R., Plant, R., Puc, B.P., Ronan, M.T., Roth, G.T., Sarkis, G.J., Simons, J.F., Simpson, J.W., Srinivasan, M., Tartaro, K.R., Tomasz, A., Vogt, K.A., Volkmer, G.A., Wang, S.H., Wang, Y., Weiner, M.P., Yu, P.G., Begley, R.F. & Rothberg, J.M. 2005. Genome sequencing in microfabricated high-density picoliter reactors. *Nature* **437**: 376–380.
- Marin, F., Corstjens, P., de Gaulejac, B., Vrind-De Jong, E.D. & Westbroek, P. 2000. Mucins and molluscan calcification – molecular characterization of mucoperlin, a novel mucin-like protein from the nacreous shell layer of the fan mussel *Pinna nobilis* (Bivalvia, Pteriomorpha). *J. Biol. Chem.* **275**: 20667–20675.
- Martínez-Fernández, M., Rodríguez-Piñeiro, A.M., Oliveira, E., de la Cadena, M.P. & Rolán-Alvarez, E. 2008. Proteomic Comparison between Two Marine Snail Ecotypes Reveals Details about the biochemistry of adaptation. *J. Proteome Res.* **7**: 4926–4934.
- Marxen, J.C., Becker, W., Finke, D., Hasse, B. & Eppele, M. 2003. Early mineralization in *Biomphalaria glabrata*: microscopic and structural results. *J. Molluscan Stud.* **69**: 113–121.
- Matsubara, H., Hayashi, T., Ogawa, T., Muramoto, K., Jimbo, M. & Kamiya, H. 2008. Modulating effect of acorn barnacle C-type lectins on the crystallization of calcium carbonate. *Fish. Sci.* **74**: 418–424.
- Menge, B.A. & Branch, G.M. 2001. Rocky intertidal communities. In: *Marine Community Ecology* (M.D. Bertness, S.D. Gaines & M.E. Hay, eds), pp. 221–251. Sinauer Associates Inc., Sunderland, MA.
- Metzker, M.L. 2010. Sequencing technologies – the next generation. *Nat. Rev. Gen.* **11**: 31–46.
- Morrison, J.F. 1973. Arginine kinase and other invertebrate guanidino kinases. In: *The Enzymes* (P.D. Boyer, ed.), pp. 457–486. Academic Press, New York.
- Norton, T.A., Hawkins, S.J., Manley, N.L., Williams, G.A. & Watson, D.C. 1990. Scraping a living: a review of littorinid grazing. *Hydrobiologia* **193**: 117–138.
- Nosil, P. & Rundle, H.D. 2009. Ecological speciation. In: *The Princeton Guide to Ecology* (S.A. Levin, ed.), pp. 134–142. Princeton University Press, Princeton, NJ.
- Nosil, P., Funk, D.J. & Ortiz-Barrientos, D. 2009. Divergent selection and heterogeneous genomic divergence. *Mol. Ecol.* **18**: 375–402.
- Novaes, E., Drost, D.R., Farmerie, W.G., Pappas, G.J., Grattapaglia, D., Sederoff, R.R. & Kirst, M. 2008. High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. *BMC Genomics* **9**: 312.
- Panova, M., Hollander, J. & Johannesson, K. 2006. Site-specific genetic divergence in parallel hybrid zones suggests nonallopatric evolution of reproductive barriers. *Mol. Ecol.* **15**: 4021–4031.
- Pant, P.V.K., Tao, H., Beilharz, E.J., Ballinger, D.G., Cox, D.R. & Frazer, K.A. 2006. Analysis of allelic differential expression in human white blood cells. *Genome Res.* **16**: 331–339.
- Pickles, A. & Grahame, J. 1999. Mate choice in divergent morphs of *Littorina saxatilis* (Oliv): speciation in action? *Anim. Behav.* **58**: 181–184.
- Quesada, H., Posada, D., Caballero, A., Morán, P. & Rolán-Alvarez, E. 2007. Phylogenetic evidence for multiple sympatric ecological diversification in a marine snail. *Evolution* **61**: 1600–1612.
- Raffaelli, D.G. 1978. The relationship between shell injuries, shell thickness and habitat characteristics of the intertidal snail *Littorina rudis* Maton. *J. Molluscan Stud.* **44**: 166–170.
- Reid, D.G. 1996. *Systematics and Evolution of Littorina*. The Ray Society, Dorchester.
- Renaut, S., Nolte, A.W. & Bernatchez, L. 2010. Mining transcriptome sequences towards identifying adaptive single nucleotide polymorphisms in lake whitefish species pairs (*Coregonus* spp. Salmonidae). *Mol. Ecol.* **19**: 115–131.
- Rolán-Alvarez, E. 2007. Sympatric speciation as a by-product of ecological adaptation in the galician *Littorina saxatilis* hybrid zone. *J. Molluscan Stud.* **73**: 1–10.
- Rolán-Alvarez, E., Carballo, M., Galindo, J., Moran, P., Fernandez, B., Caballero, A., Cruz, R., Boulding, E.G. & Johannesson, K. 2004. Nonallopatric and parallel origin of local reproductive barriers between two snail ecotypes. *Mol. Ecol.* **13**: 3415–3424.
- Sadedin, S., Hollander, J., Panova, M., Johannesson, K. & Gavrillets, S. 2009. Case studies and mathematical models of ecological speciation. 3: ecotype formation in a Swedish snail. *Mol. Ecol.* **18**: 4006–4023.
- Schluter, D. 2000. *The Ecology of Adaptive Radiation*. Oxford University Press, Oxford.
- Schluter, D. 2001. Ecology and the origin of species. *Trends Ecol. Evol.* **16**: 372–380.

- Schluter, D. 2009. Evidence for ecological speciation and its alternative. *Science* **323**: 737–741.
- Schmidt, P.S., Phifer-Rixey, M., Taylor, G.M. & Christner, J. 2007. Genetic heterogeneity among intertidal habitats in the flat periwinkle, *Littorina obtusata*. *Mol. Ecol.* **16**: 2393–2404.
- Schwarz, D., Robertson, H.M., Feder, J.L., Varala, K., Hudson, M.E., Ragland, G.J., Hahn, D.A. & Berlocher, S.H. 2009. Sympatric ecological speciation meets pyrosequencing: sampling the transcriptome of the apple maggot *Rhagoletis pomonella*. *BMC Genomics* **10**: 633.
- Shagin, D.A., Rebrikov, D.V., Kozhemyako, V.B., Altshuler, I.M., Shcheglov, A.S., Zhulidov, P.A., Bogdanova, E.A., Staroverov, D.B., Rasskazov, V.A. & Lukyanov, S. 2002. A novel method for SNP detection using a new duplex-specific nuclease from crab hepatopancreas. *Genome Res.* **12**: 1935–1942.
- Sobel, J.M., Chen, G.F., Watt, L.R. & Schemske, D.W. 2010. The biology of speciation. *Evolution* **64**: 295–315.
- Sokal, R.R. & Rohlf, F.J. (1994) *Biometry*. W H Freeman & Co., New York, p. 764.
- Storz, J.F. 2005. Using genome scans of DNA polymorphism to infer adaptive population divergence. *Mol. Ecol.* **14**: 671–688.
- Tatarenkov, A. & Johannesson, K. 1994. Habitat related allozyme variation on a microgeographic scale in the marine snail *Littorina mariae* (Prosobranchia: Littorinacea). *Biol. J. Linn. Soc.* **53**: 105–125.
- Taylor, J.D. & Reid, D.G. 1990. Shell microstructure and mineralogy of the Littorinidae: ecological and evolutionary significance. *Hydrobiologia* **193**: 199–215.
- The Gene Ontology Consortium 2000. Gene Ontology: tool for the unification of biology. *Nature Genetics* **25**: 25–29.
- Turner, T.L., Bourne, E.C., Von Wettberg, E.J., Hu, T.T. & Nuzhdin, S.V. 2010. Population resequencing reveals local adaptation of *Arabidopsis lyrata* to serpentine soils. *Nat. Genet.* **42**: 260–263.
- Van Tassell, C.P., Smith, T.P.L., Matukumalli, L.K., Taylor, J.F., Schnabel, R.D., Lawley, C.T., Haudenschild, C.D., Moore, S.S., Warren, W.C. & Sonstegard, T.S. 2008. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nat Methods* **5**: 247–252.
- Vera, J.C., Wheat, C.W., Fescemyer, H.W., Frilander, M.J., Crawford, D.L., Hanski, I. & Marden, J.H. 2008. Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Mol. Ecol.* **17**: 1636–1647.
- Via, S. 2009. Natural selection in action during speciation. *Proc. Natl Acad. Sci. USA* **106**: 9939–9946.
- Weber, A.P.M., Weber, K.L., Carr, K., Wilkerson, C. & Ohlrogge, J.B. 2007. Sampling the arabidopsis transcriptome with massively parallel pyrosequencing. *Plant Physiol.* **144**: 32–42.
- Wheat, C.W. 2010. Rapidly developing functional genomics in ecological model systems via 454 transcriptome sequencing. *Genetica* **138**: 433–451.
- Wilding, C.S., Grahame, J. & Mill, P.J. 2000. Mitochondrial DNA *CoI* haplotype variation in sibling species of rough periwinkles. *Heredity* **85**: 62–74.
- Wilding, C.S., Butlin, R.K. & Grahame, J. 2001. Differential gene exchange between parapatric morphs of *Littorina saxatilis* detected using AFLP markers. *J. Evol. Biol.* **14**: 611–619.
- Wood, H.M., Grahame, J.W., Humphray, S., Rogers, J. & Butlin, R.K. 2008. Sequence differentiation in regions identified by a genome scan for local adaptation. *Mol. Ecol.* **17**: 3123–3135.
- Wu, C.I. 2001. The genic view of the process of speciation. *J. Evol. Biol.* **14**: 851–865.
- Zhu, Y.Y., Machleder, E.M., Chenchik, A., Li, R. & Siebert, P.D. 2001. Reverse transcriptase template switching: a SMART (TM) approach for full-length cDNA library construction. *BioTechniques* **30**: 892–897.
- Zhulidov, P.A., Bogdanova, E.A., Shcheglov, A.S., Vagner, L.L., Khaspekov, G.L., Kozhemyako, V.B., Matz, M.V., Meleshkevitch, E., Moroz, L.L., Lukyanov, S.A. & Shagin, D.A. 2004. Simple cDNA normalization using kamchatka crab duplex-specific nuclease. *Nucleic Acids Res.* **32**: e37.

Supporting information

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

Figure S1 Histograms representing the length of the contigs and the coverage (number of reads per contig) of the assemblies performed with Newbler and NGen.

Figure S2 Partial assembly of an outlier contig (Contig-00865, unannotated).

Table S1 Parameters used for *de novo* assembly with programs Newbler (gs *de novo* Assembler v.2.0.00.20; 454 Life Sciences) and NGen (SEQMAN NGEN™ v1.2, DNA*®) and mapping assembly with gs Reference Mapper (454 Life Sciences).

Table S2 Summary of the gs FLX 454 sequencing run performed on cDNA pools of M and H ecotypes of *Littorina saxatilis*.

Table S3 Summary of the results of the assemblies performed using Newbler and NGen.

Table S4 Results of the functional annotation performed with Blast2GO.

Table S5 Blastx results for contigs with apparently differential expression between M and H ecotypes for the Newbler assembly.

Table S6 Blastx results for contigs with apparently differential expression between M and H ecotypes for the NGen assembly.

Table S7 Summary of the highly differentiated contigs between H and M ecotypes of *L. saxatilis* obtained in the xenobiotics analysis (Blastn search against nr nucleotide database in NCBI).

Table S8 Contigs with outlier loci ($P < 0.05$) that hit BAC clones (E10, E12, A30 and A37) from *Littorina saxatilis* (see Discussion) after Blastn searches.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organized for online delivery but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.

Received 5 May 2010; revised 30 June 2010; accepted 3 July 2010