MOLECULAR ECOLOGY

Molecular Ecology (2012) doi: 10.1111/j.1365-294X.2012.05680.x

Genome-wide patterns of standing genetic variation in a marine population of three-spined sticklebacks

PHILINE G. D. FEULNER,*[1] FRÉDÉRIC J. J. CHAIN,†[1] MAHESH PANCHAL,†[1] CHRISTOPHE EIZAGUIRRE,†‡ MARTIN KALBE,† TOBIAS L. LENZ,†¶ MARVIN MUNDRY,* IRENE E. SAMONTE,† MONIKA STOLL,§ MANFRED MILINSKI,†[2] THORSTEN B. H. REUSCH‡[2] and ERICH BORNBERG-BAUER*[2]

*Institute for Evolution and Biodiversity, Evolutionary Bioinformatics, Westfaelische Wilhelms University, Huefferstr. 1, 48149 Muenster, Germany, †Max Planck Institute for Evolutionary Biology, Department of Evolutionary Ecology, August-Thienemann-Str. 2, 24306 Ploen, Germany, ‡Helmholtz Center for Ocean Research (GEOMAR), Evolutionary Ecology of Marine Fishes, Duesternbrooker Weg 20, 24105 Kiel, Germany, §Genetic Epidemiology of Vascular Disorders, Leibniz-Institute for Arteriosclerosis Research, Domagkstr. 3, 48149 Muenster, Germany

Abstract

Since the end of the Pleistocene, the three-spined stickleback (*Gasterosteus aculeatus*) has repeatedly colonized and adapted to various freshwater habitats probably originating from ancestral marine populations. Standing genetic variation and the underlying genomic architecture both have been speculated to contribute to recent adaptive radiations of sticklebacks. Here, we expand on the current genomic resources of this fish by providing extensive genome-wide variation data from six individuals from a marine (North Sea) stickleback population. Using next-generation sequencing and a combination of paired-end and mate-pair libraries, we detected a wide size range of genetic variation. Among the six individuals, we found more than 7% of the genome is polymorphic, consisting of 2 599 111 SNPs, 233 464 indels and structural variation (SV) (>50 bp) such as 1054 copy-number variable regions (deletions and duplications) and 48 inversions. Many of these polymorphisms affect gene and coding sequences. Based on SNP diversity, we determined outlier regions concordant with signatures expected under adaptive evolution. As some of these outliers overlap with pronounced regions of copy-number variation, we propose the consideration of such SV when analysing SNP data from re-sequencing approaches. We further discuss the value of this resource on genome-wide variation for further investigation upon the relative contribution of standing variation on the parallel evolution of sticklebacks and the importance of the genomic architecture in adaptive radiation.

Keywords: copy-number variation, *Gasterosteus aculeatus*, inversions, population genome evolution, standing genetic variation, structural variation

Received 7 March 2012; revision received 4 May 2012; accepted 15 May 2012

Introduction

Making the link between phenotype and causal mutations has enabled scientists to track adaptive changes in natural populations demonstrating evolution in action (Barrett *et al*. 2008; Gratten *et al*. 2008). With the promises brought by next-generation sequencing, genomic tools have revolutionized the capacity to identify alleles underpinning adaptation (Ellegren & Sheldon 2008; Stapley *et al*. 2010). But what is the origin and form of the genetic variation underlying phenotypic traits on which natural selection operates? What are the relative contributions of newly arisen mutations and standing genetic variation in adaptive evolution? In contrast to selection immediately acting on newly arisen variation,

Correspondence: Philine G. D. Feulner, Fax: +49 251 83 24668; E-mail: p.feulner@uni-muenster.de

[1] These authors contributed equally to this work.

[2] Shared senior authorship.

¶Present address: Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, 77 Ave. Louis Pasteur, Boston, MA 02115, USA.

adaptation from standing genetic variation implies that there are neutral and slightly deleterious variations maintained in a population and that those become beneficial upon a change of selection regime (Barrett & Schluter 2008). In a broader sense, standing variation also includes variation introduced by gene flow from another population. As standing genetic variation has persisted in the population over time, it therefore cannot be highly detrimental and can occur at higher frequency than newly arisen mutations (Barrett & Schluter 2008). Thus, standing variation can contribute to fast adaptation after a sudden change of environment (Kitano *et al.* 2008; Eizaguirre *et al.* 2012). Simulations demonstrate that fixation probabilities are increased for alleles stemming from standing genetic variation (Orr & Betancourt 2001), especially for alleles experiencing relatively weak selection (Hermisson & Pennings 2005). Experimental evolution has provided evidence that adaptation from standing variation is more repeatable than evolution resulting from novel mutations (Teotonio *et al.* 2009). However, the genetic basis for instances of parallel adaptation has not always been traced back to the same mutation in the same gene, but sometimes to different mutations in the same gene, to mutations in different genes within the same biochemical pathway or to mutations of genes in different pathways with a similar phenotypic effect (Manceau *et al.* 2010; Elmer & Meyer 2011). The adaptive course is modulated by the type and quantity of the available genetic variation (Hermisson & Pennings 2005). Therefore, a comprehensive description of variation in a natural population facilitates our understanding of the contribution of standing variation in the context of ecological adaptation and parallel evolution—but such studies describing genome-wide variation in populations remain rare.

While standing genetic variation can contribute to the emergence of rapid adaptations, variation exists in numerous shapes and sizes. Genetic differences can be categorized by size into single-nucleotide polymorphisms (SNPs), small insertion and deletions (commonly referred to as indels; ≤50 bp) and larger structural changes (SV, structural variation; >50 bp). Combined, these shape the genome architecture. SV can further be differentiated into 'balanced' changes, like inversions and translocations, and 'unbalanced' changes altering the genomic content, like large duplications and deletions (CNV, copy-number variation). The pervasiveness and impact of these different types of variation in natural populations are of great interest to yield a better understanding of genome evolution and adaptation. Recent advancements in technology have made whole-genome inspections much more accessible, especially the investigation of SV (Korbel *et al.* 2007; Medvedev *et al.* 2009; Alkan *et al.* 2011). This has led to the emergence of dense variation maps (covering a major proportion of the genome) summarizing a large size spectrum of variations for many eukaryotic organisms, especially those of economic relevance such as rice (Xu *et al.* 2012), maize (Lai *et al.* 2010), soybean (Lam *et al.* 2010), chicken (Rubin *et al.* 2010) and cattle (Zhan *et al.* 2011). There is also increasing information on SV of model species like *Arapidopsis thaliana* (DeBolt 2010), *Drosophila melanogaster* (Cridland & Thornton 2010), zebrafish (Brown *et al.* 2012), mice (Quinlan *et al.* 2010) and humans (Handsaker *et al.* 2011; Mills *et al.* 2011a,b). Additionally, evidence for the adaptive potential of CNV is growing and supported by findings in diverse taxa from flies (Emerson *et al.* 2008) to fish (Chen *et al.* 2008) to apes (Gazave *et al.* 2011). There is also reason to believe that balanced SV could sometimes be related to adaptive success. For example, inversions have been linked to adaptation to different soil types in the monkeyflower (Lowry & Willis 2010) and correlated with a change in diapause in *Rhagoletis* flies (Feder *et al.* 2003). The importance of inversions for the speciation process, including theoretical models and empirical evidence, has been reviewed by Faria & Navarro (2010). However, despite the evolutionary impact of SV, only a few studies provide a comprehensive characterization of genetic variation covering a broad size spectrum in a natural population.

The three-spined stickleback (*Gasterosteus aculeatus*) has been called a supermodel in evolution as it combines supreme knowledge of its life history traits with the availability of many genomic resources (Gibson 2005). Sticklebacks occur in various types of aquatic habitats throughout the northern hemisphere, having repeatedly adapted to novel environments (Bell & Foster 1994). Populations demonstrate a huge diversity of phenotypes differing in size, shape, behaviour and mating preference (Bell & Foster 1994; McKinnon & Rundle 2002; Gibson 2005). From an evolutionary perspective, these recently diverged populations that have undergone multiple independent adaptations serve as replicates of evolution (Rundle & Nosil 2005) and offer promising opportunities to gain further insights into adaptation and speciation. The adaptive radiation of sticklebacks has been extensively studied (Schluter 1996a,b), and this process might be facilitated by standing genetic variation of a largely intermixed marine source population (Bell & Foster 1994). Following from there, Schluter & Conte (2009) developed a hypothesis explaining the maintenance of standing genetic variation within the marine populations by a metapopulation scenario in which alleles with adaptive potential are 'transported' between populations through gene flow. This hypothesis is supported by observations of the same alleles repeatedly rising in frequency in freshwater populations with parallel selection regimes (Colosimo *et al.* 2005; Barrett *et al.* 2008; Jones *et al.* 2012a). One striking example is the identification of the causal mutation for the adaptive reduction in bony amour in freshwater sticklebacks, a SNP in the EDA locus that is maintained at low frequency in the marine population (Colosimo *et al.* 2005). However, parallel evolution of the same phenotypic changes can also be driven by SV such as deletions in the Pitx1 enhancer causing pelvic reduction (Chan *et al.* 2010). In contrast, this case discloses independent parallel evolution (different mutations in the same genomic region), rather than recurrent sweeps of the same allele (same mutation in the same gene). A more comprehensive account of variation in the entire genome is needed to evaluate the relative impact of different origins (newly arisen vs. standing) and different kinds and sizes (SNP, indel or SV) of variation. Genome scans utilizing various markers [microsatellites, restriction-site associated DNA tags (RAD), SNPs] display signatures of selection around the locations of previously identified adaptive genes and reveal further candidate regions suggestively shaped by selection (Mäkinen *et al.* 2008; Hohenlohe *et al.* 2010; DeFaveri *et al.* 2011; Deagle *et al.* 2012; Jones *et al.* 2012a,b). These studies provided an initial insight into the proportion of the stickleback genome that is variable between different populations and potentially shaped by adaptation. Here, we expand on these studies by providing a genome-wide account of the vast array of genetic variation in marine sticklebacks using whole-genome re-sequencing, which has enabled us to evaluate the contribution of large SV in genome evolution.

This study used next-generation sequencing data to describe the genome-wide variation and genomic architecture in a natural marine stickleback population. Whole-genome sequencing has numerous advantages such as the ability of (i) covering the entire genome for a comprehensive characterization of diversity, (ii) circumventing the polymorphism ascertainment bias of SNP arrays, (iii) identifying new rare variants and also (iv) detecting numerous types of SV (Zhan *et al.* 2011). To some extent, our findings of genomic variation impart the standing genetic variation in a European marine stickleback population, the putative founding source for the colonization of many surrounding freshwater systems. The genetic variation herein provides us with a foundation for identifying variation contributing to adaptation in an ecological context. Moreover, our study further contributes to develop a holistic view of the genomic architecture of a natural population. This may illuminate potential mechanisms for rapid adaptation in diverse environments (Hohenlohe *et al.* 2012).

Methods

Data collection

Thirty fish were caught in September 2010 from Lemvig in Limfjord, Denmark, in the North Sea (56°36′9.19″N, 8°18′1.94″E). Muscle tissue from six randomly sampled individuals (three females and three males) was used for DNA extraction using a Qiagen DNA Midi Kit following the manufacturer's protocol for high molecular weight DNA. Sequencing was outsourced (GATC Biotech AG) and performed using Illumina technology following Illumina protocols, producing for each individual a paired-end library and a mate-pair library. The paired-end library read length was 96 bp with an average insert size of 140 bp, and the mate-pair library read length was 36–51 bp with an average insert size of 3 kb. The total amount of raw data for all six fish together was 58 Gb (for details about the amount of data per individual, see Table S1, Supporting information).

Data processing

Raw data were first analysed by applying the software FastQC v0.10.0 (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/) to find the quality score encoding and view data quality statistics. We then used SeqPrep (https://github.com/jstjohn/SeqPrep) to remove adapter sequence from the reads. To prepare the data for the genotyping process, SeqPrep was used to merge overlapping reads in the paired-end libraries. Mate-pair data were reverse-complimented. To prepare the data for SV detection, each paired-end library read was trimmed to a maximum of 50 bp to use the paired-end signal in the data, because most reads overlap with the second read in the pair. Reads were mapped to version 65 of the ensembl three-spined stickleback genome (Broad Institute 2007; Flicek *et al.* 2011; Jones *et al.* 2012b). Merged reads were aligned using the bwase module in bwa v0.5.9-r16 (Li *et al.* 2008), while paired reads for the paired-end library, mate-pair library and trimmed paired-end library were aligned using the bwape module. For the mate-pair data, read pairs that aligned facing away from each other and had an insert size <1 kb were labelled as single end reads by changing the flag in the sam format output. All reads were processed using the picard toolkit v1.53 (http://picard.sourceforge.net/) by applying fixmateinformation.jar, then cleansam.jar and finally sortsam.jar to sort the data by coordinate. SAMtools v0.1.18 (Li *et al.* 2009) was used to index the resulting bam format files. Data from different lanes belonging to the same individual were combined using mergesamfiles.jar in the picard toolkit.

Each combined data set was locally realigned using the GATK toolkit v1.4 (McKenna *et al.* 2010), followed by removing duplicate reads with MARKDUPLICATES.JAR in PICARD, and base quality score recalibration using the GATK toolkit (DePristo *et al.* 2011).

SNPs and indels

Following the above filtering steps, the resulting bam files were utilized for calling genotypes, specifically SNPs and small indels using GATK (DePristo *et al.* 2011). For variant recalibration with the GATK toolkit, SNPs were also called using SAMTOOLS (Li *et al.* 2009), and concordant variants between the two callers were used. The remaining variant sites were phased and imputed using BEAGLE v3.1 (Browning & Browning 2009) and annotated for impact of variant effects using SNPEFF v2.0.2 (Cingolani 2012) based on gene annotations from version 65 of ENSEMBL (Flicek *et al.* 2011). For further analyses, all variations that overlapped with unknown bases and masked regions based on the ENSEMBL repeat-masked genome file (Flicek *et al.* 2011) were excluded. Our final SNP set also excluded all SNP calls within 10 bp of indels. For functional analyses, only the highest-impact effect annotated for each variant was used. As we are interested in polymorphisms within the sampled marine population, variation with respect to the reference sequence was not analysed. This has the additional advantage of reducing the impact of recurrent mis-mapping due to genomic characteristics, such as repetitive and duplicated sequences, or due to variation limited to the reference genome. VCFTOOLS (Danecek *et al.* 2011) and custom scripts were used to acquire frequency properties of the variant genotypes.

Copy-number variation regions (CNVRs)

Copy-number variations (large duplications and deletions) were assessed through read-depth analysis using CNVNATOR (Abyzov *et al.* 2011). For each individual, postfiltered mate-pair reads and paired-end reads (merged but not trimmed—see above Methods) were combined and processed in CNVNATOR with a bin size of 500 bp along each linkage group (LG). CNV calls that overlapped (more than 50% of their length) with masked regions were excluded from CNV analyses. Remaining calls were compared between individuals using MERGEBED and ANNOTATEBED of BEDTOOLS (Quinlan & Hall 2010), and only CNV polymorphic among the marine individuals were identified as CNVRs and analysed for reasons already stated above. Although not analysed, the identification of read-depth differences compared to the reference genome that exists in all individuals ('fixed' deletions and duplications) helped in detecting regions that may be prone to mapping over- and under-representation. This actually becomes a crucial component when interpreting the results from the genome scan analysis based on SNPs.

Deletions and inversions

In addition to using read-depth to call CNVRs, paired-end and split-read methods were also performed to detect SV. Deletions, inversions as well as small indels were called utilizing the paired-end mapping of the trimmed reads (see above) and the reads of the mate-pair library, separately. For both libraries, read pairs that mapped uniquely to the reference were extracted. The paired-end approach uses information on the relative distance (span) and orientation of read pairs. For this, we used BREAKDANCERMAX-1.1r112 with default parameters except for (i) increasing the cut-of to seven units of standard deviation, (ii) setting the minimum number of reads required to establish a connection to four, (iii) setting the maximum threshold of sequence coverage to 50 and (iv) utilizing the long insert-size option for the mate-pair libraries (Chen *et al.* 2009). The split-read analysis (PINDEL 0.2.2 with default parameters; Ye *et al.* 2009) uses information on continuous breaks in the alignment due to split-read mappings. Both approaches were performed on libraries of all six individuals together. This way we have the most power (from sequence and physical coverage) to detect SV while still being able to trace back reads to their respective individual library. BREAKDANCER variation calls overlapping by more than 50% of their breakpoint range were merged into a nonredundant set and filtered by a confidence score >60 for deletions and >90 for inversions. PINDEL calls were filtered to remove variants supported by <4 reads. Utilizing INTERSECTBED (v2.14.2) of BEDTOOLS (Quinlan & Hall 2010), calls of different libraries and approaches were intersected to establish a set of nonredundant calls of deletions and inversions. Calls present in all individuals or overlapping (more than 50%) with masked regions of the genome were excluded. Again, this ensures that our set of SV contains only polymorphisms of the marine population, and reduces false positives due to characteristics unique to the reference assembly. The deletions called with BREAKDANCER and PINDEL are a particular subset of CNV that involve the loss of either 1 or 2 copies in some individuals.

General analysis

As mentioned above, we excluded from further analyses variation that overlapped with masked regions (7.6% of the autosomes); therefore, all numbers and percentages reported are based on the reference genome after masking. We only report variation on the 20 annotated autosomes (LGs I–XVIII, XX and XXI). This comprises 84% of the reference genome and excludes the mitochondrial genome, the sex chromosome and unassembled scaffolds. We designed random primer pairs across the genome and performed Sanger sequencing to validate genotypes called using our Illumina data. In total, we confirmed 59 SNPs, 10 indels and 17 554 invariant sites with 99% accuracy.

Nucleotide diversity (π) and Tajima's *D* were calculated with LIBSEQUENCE (Thornton 2003) in 100-kb nonoverlapping sliding windows across the genome. To find potential regions evolving under positive or balancing selection, we identified outlier windows for π and Tajima's *D* (in the top and bottom 1% quantiles) of each respective LG following an empirical outlier approach (Akey *et al.* 2010; Kolaczkowski *et al.* 2011; Cheng *et al.* 2012). Imperfect mapping and unbalanced SV can affect SNP calling and thus genome scan inferences. We therefore excluded outlier windows that contained 50% or more nongenotyped sites caused by missing data, masked sites or deletions in our samples compared to the reference genome. Importantly, we also excluded all outliers that overlapped with fixed duplications compared to the reference sequence; these regions have potentially multiple loci mapping to the same reference position affecting heterozygosity estimates.

Variants overlapping with gene annotations from version 65 of ENSEMBL were identified using INTERSECTBED (v2.14.2) of BEDTOOLS (Quinlan & Hall 2010). The variation effect software SNPEFF (Cingolani 2012) was used to annotate the impact of the variation calls (e.g. synonymous or nonsynonymous). To determine enrichment of gene ontology (GO) terms among regions and gene lists, TOPGO was used with a universe of autosomal genes, and significance was determined using FDR-adjusted *P*-values to help correct for multiple testing (Alexa *et al.* 2006). Representations of genomic variation were visualized using CIRCOS (Krzywinski *et al.* 2009) and R (R Development Core Team 2011).

Results

We performed whole-genome sequencing of six marine three-spined stickleback individuals using a small-insert paired-end library in conjunction with a large-insert mate-pair library per individual. Of the 58 Gb of raw sequence data, 3.4–4.7 Gb per individual remained after filtering for paired-end libraries and 1.3–3.5 Gb per individual for mate-pair libraries (for details, see Table S2, Supporting information). After mapping to the autosomes, this resulted in an average of 10x depth of coverage per individual. Average insert sizes ranged from 130 to 150 bp for paired-end libraries and 2.7 to 3.0 kb for mate-pair libraries. We used the ENSEMBL stickleback genome as a reference for our assembly but only report variation exclusive to our samples on the 20 autosomal LGs. This comprised 345 424 403 sites (98% of the 20 LGs) with at least one genotyped individual (94% of sites have genotypes for all individuals). Overall, we found that more than 25.6 Mb of the assembled genome (>7%) are affected by some type of genetic variation (Table 1). These polymorphism calls include SNPs, indels, CNVRs and inversions (Table 2). Below, we highlight the pervasiveness of these variants, as well as the parts of the genome that displayed the highest and lowest amounts of nucleotide diversity.

Over 2.5 Mb of autosomal sites was affected by SNPs (Table 1). The average nucleotide diversity, π, was 0.0025 (ranging from 0.0003 to 0.0066), and the average Tajima's *D* was −0.06 (ranging from −1.55 to 1.82). Using a genome scan with 100-kb windows, we detected a total of 103 outlier (π and/or Tajima's *D*) regions (Fig. 1). The combined outlier regions overlapped with 415 protein-coding genes (Table S3, Supporting information). Some large SV intersected with outliers: a total of four inversions encompassed four different outlier regions (three on LG XIII and one on LG XV), and 96 duplication and deletion regions overlapped with 57 outliers. It is likely that the latter unbalanced SV causes an increase or decrease in observed heterozygosity; therefore, we present our outlier analyses mindful of CNVRs.

Genomic regions that have high nucleotide diversity as well as more intermediate-frequency polymorphisms than expected are potentially evolving under balancing selection. Furthermore, these positions harbour genetic diversity that can be subsequently partitioned during colonization of new environments. Two outlier windows, one on LG V and one on LG XV, are consistent with molecular patterns of balancing selection as they are outliers for both the top 1% ('high outliers') of π and Tajima's *D* (Fig. 1A—'High'). The two windows contain 15 genes with various functions (protein glycosylation, protein binding, ion transport, hydrolase activity, translational initiation, protein phosphorylation, transferase activity, kinase activity). The window on LG XV has one deletion, and the window on LG V has one CNVR containing both duplication and deletion, possibly contributing to the elevated observed diversity (Fig. 1B). Additionally, a total of 150 genes lie in the high π outliers, which show an enrichment of genes involved in protein ubiquitination and ligase activity, even after removing 44 genes in CNVRs. We found no enrichment of GO terms in 108 genes within the high outliers of Tajima's *D* after accommodating for CNVRs.

Table 1 Summary of polymorphism calls found on the 20 autosomes of six marine stickleback genomes

	Genotypes	SNPs (GATK)	Indels (GATK)	CNVRs (CNVNATOR)	Deletions (BD)	Deletions (PINDEL)	Inversions (BD)	Inversions (PINDEL)
Total	345 424 403	2 599 111	233 464	1054	1075	28	29	19
bp size	345 424 403	2 599 111	534 969	10 646 500	4 372 633	763 067	10 889 318	59 323
Mean size		1	2.29	10 101	4068	2725	375 500	3122
Median size		1	2	5000	310	2055	15 670	2273
Rare		0.39	0.27	0.63	0.26	0.00	0.03	0.00
In genes	0.44	0.43	0.43	0.45	0.39	0.43	0.72	0.58
In CDS	0.08	0.04	0.006	0.34	0.07	0.18	0.72	0.32
Genes affected	0.99	0.96	0.76	0.04	0.04	0.00	0.03	0.00
CDS affected	0.96	0.88	0.07	0.03	0.03	0.00	0.03	0.00

The metrics reported are the total count (total), the total length in base pairs (bp size), the mean length in bp (mean size), the median length in bp (median size), the proportion of calls that are rare variants (rare, which represent variants that differ in only one individual), the proportion of calls in genes (in genes) and in protein-coding regions of genes (in CDS), the proportion of all genes that intersect with the calls (genes affected) and the proportion of genes in which their protein-coding regions intersect with the calls (CDS affected). The genotypes column represents the total amount of genotyped positions after genome masking, as well as the proportion of genotyped sites that are in genes (in genes) and in protein-coding regions (in CDS), and the proportion of genes (genes affected) and CDS (CDS affected) that contain genotyped sites. The rest of the columns represent different types of genetic variation: single-nucleotide polymorphisms (SNPs), small insertions and deletions (Indels), copy-number variation regions (CNVRs), deletions and inversions. The methods used for detection were GATK, CNVNATOR, BREAKDANCER (BD) and PINDEL.

Table 2 Distribution of genetic variation by autosomal linkage group (LG)

LG	Length [bp]	SNPs	Indels	CNVRs	Deletions	Inversions
Group I	26 006 010	183 580	16 490	68	77	5
Group II	21 641 308	155 217	14 204	54	77	2
Group III	15 639 605	117 364	10 412	36	44	3
Group IV	30 011 177	209 401	18 532	129	74	6
Group V	11 246 971	81 654	7197	42	35	1
Group VI	15 852 356	122 512	10 969	33	59	2
Group VII	25 726 737	180 171	16 099	112	64	0
Group VIII	17 701 717	126 437	11 839	58	31	5
Group IX	18 672 053	147 123	13 284	72	60	2
Group X	14 296 404	108 869	10 063	49	57	0
Group XI	15 348 408	120 290	10 479	43	59	0
Group XII	16 987 265	120 811	10 836	45	54	1
Group XIII	18 746 755	134 482	12 004	49	71	8
Group XIV	14 073 442	116 744	10 829	22	37	4
Group XV	15 125 368	114 338	10 494	35	74	1
Group XVI	16 932 203	127 911	10 943	66	59	3
Group XVII	13 546 504	108 956	10 075	28	31	3
Group XVIII	14 875 199	107 985	9937	42	53	1
Group XX	18 186 289	135 915	11 903	51	58	0
Group XXI	10 862 270	79 351	6875	20	28	1

The lengths of LG in bp are reported after removing masked regions. We report the total number of SNPs, indels, copy-number variation regions (CNVRs), deletions and inversions.

Among the outliers in the bottom 1% (‘low outliers’) of π and Tajima’s *D*, we found 7 windows that are low outlier for both statistics, each one on a different LG (Fig. 1A—‘Low’). These are regions with low nucleotide diversity and more low-frequency polymorphisms than expected, suggestive of directional selection. They include 26 genes with various functions (protein binding, protein phosphorylation, histone ubiquitination and acetylation, transferase activity, regulation of transcription, translation, signal transduction, calcium ion

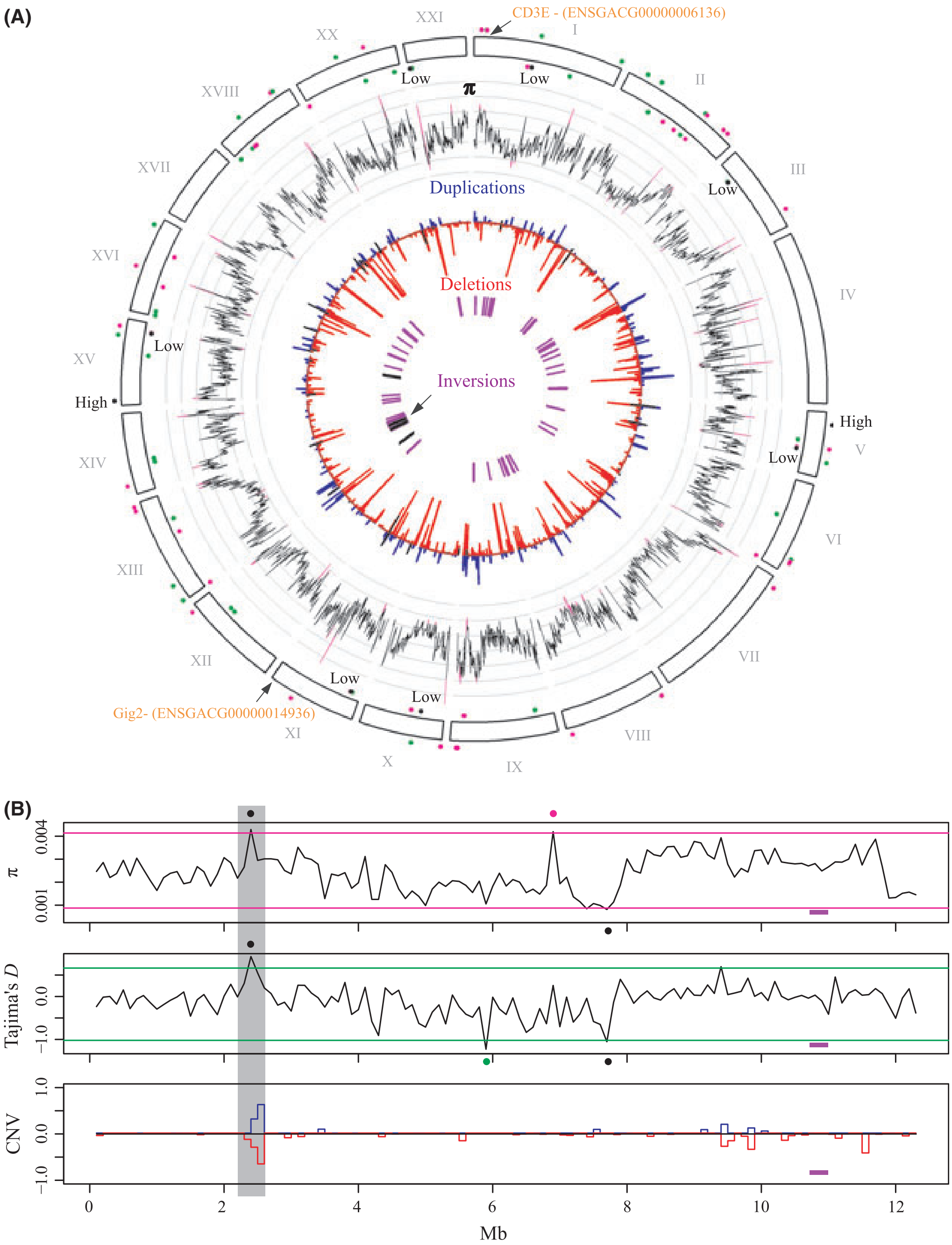

Fig. 1 (A) Distribution of genomic variation in *Gasterosteus aculeatus* across autosomes represented in circular tracks. Linkage groups (LGs) are labelled in grey Roman numerals and represented as white blocks in a circle. The dots surrounding the blocks represent genome scan outliers based on SNPs (after correcting for read-depth coverage) of nucleotide diversity π (pink), Tajima's *D* (green) and the intersection of both (black). The top 1% outlier quantiles (high π, high Tajima's *D*) are on the exterior track of the circle, and the bottom 1% (low π, low Tajima's *D*) are on the interior track. The histogram in black represents π in 100-kb sliding windows with higher values peaking outwards from the centre, and the outliers (before correcting for read-depth coverage) are in pink. The proportional length of CNVRs in 100-kb windows is shown with duplications (blue) facing outwards and deletions (red) facing inwards, whereas inversions are represented on the innermost track (purple). SV overlapping with outlier regions are in black, and some highlights from the main text are labelled. (B) Distribution of genomic variation across LG V. Nucleotide diversity π, Tajima's *D* and copy-number variation (CNV) proportions are represented in 100-kb windows. Horizontal lines indicate 1% and 99% quantiles, and outliers are indicated with dots (π pink, Tajima's *D* green, both black). Duplications are in blue, deletions are in red, and the position of an inversion is marked in purple. Grey highlighting represents a high outlier (π and Tajima's *D*) overlapping with pronounced CNV. CNVRs, Copy-number variation regions; SV, structural variation.

binding, ion transport, B-cell differentiation, ribokinase activity and polyadenylation), and one outlier on LG III contains no genes. There are four small deletions and curiously one 23-kb duplication (despite the low observed diversity) overlapping these low outliers. However, the 23-kb duplicated region does not overlap any gene and only consists of 102 SNPs, which is suggestive of a relatively recent duplication. After correcting for CNVRs, we found no enrichment of GO terms in 90 and 98 genes within the low outliers of π and Tajima's *D*, respectively.

As a complement to genome scans, whole-genome sequencing provides the ability to investigate specific genes in more molecular detail, either inside or outside the outlier regions. Whereas most SNPs are intergenic (58%), many SNPs affect protein-coding regions (2.4% synonymous sites, 1.8% nonsynonymous sites) with an overall nonsynonymous/synonymous SNP ratio of 1:1.37. Genes in low outliers have an average nonsynonymous/synonymous SNP ratio of 1:1.65 compared with 1.47:1 for genes in high outliers, consistent with the corresponding suggested selection regimes. Of all the annotated autosomal protein-coding genes, 88% have at least one SNP and 68% have at least one nonsynonymous SNP within coding regions. The top 1% of genes with the highest proportion of protein-coding nonsynonymous SNPs (per CDS length) is enriched with genes involved in protein ubiquitination and ligase activity. The two genes with the highest proportion of nonsynonymous SNPs that are not within CNVRs (Fig. 1A) have genotype calls for all six individuals for 100% and 89% of their lengths, respectively. The first is a short gene (ENSGACG00000014936) similar to interferon-inducible protein Gig2, a gene involved in pathogen response in fish (Baerwald *et al.* 2008; Jiang *et al.* 2009). Among our samples, this gene contains three synonymous and 19 nonsynonymous SNPs in its 456-bp exon. The second is CD3E (ENSGACG00000006136), a T-cell regulating gene important in antigen recognition. Among our samples, this gene has 61 SNPs (four synonymous and 21 nonsynonymous) and two intronic indels, and it is also found in one of the high outliers of Tajima's *D*. These genes are both functionally involved in immune response and have the molecular signature indicative of balancing selection.

In addition to SNPs, the prevalence of indels and some types of larger SV was assessed. The GATK framework called 233 464 indels (Table 1). Proportionally, indels occur as frequently as SNPs in genes but very infrequently in coding regions (0.6%), giving a SNP/indel ratio of 72:1 in coding regions vs. 11:1 overall. Based on the ENSEMBL gene models, some indels are predicted as frameshift mutations but the length of indels in coding regions is more likely to be a multiple of 3 (the length of codons) compared to indels in the rest of the genome (Fig. 2).

Using CNVNATOR, a total of 4573 deletions and duplications were called among our six individuals, resulting in 1054 unique autosomal CNVRs (Table 1). More deletions (76%) were found than duplications, and in some cases (6%), a deletion in one individual co-occurred with a duplication event in another individual. Genes in duplication regions are enriched with electron carrier and hydrolase activity. In those CNVRs with co-occurring deletions and duplications, we found a total of 58 genes that are enriched with glycosylation activity. Overall, CNVRs are enriched with G-protein-coupled receptor (GPCR) genes and major histocompatibility complex (MHC) genes.

A total of 60 972 SNPs and 3058 indels lie within CNVRs. Many of these smaller variants within CNVRs could actually represent divergent loci mapping to the same position of the reference genome. One indication of this comes from SNPs that are heterozygous in all individuals, of which we found 10–20 times greater proportions in CNVRs. Furthermore, CNVRs also harbour a very high proportion of nonsynonymous SNPs, containing a nonsynonymous/synonymous SNP ratio of 1.88:1 compared to 1:1.37 for the rest of the genome.

After applying our stringent filtering we detected a total of 1102 deletions and 48 inversions using BREAKDANCER and PINDEL (Table 1). The complementary approaches of BREAKDANCER and PINDEL detected a single common deletion, and <5% of the deletions (52) overlapped with deletion calls from CNVNATOR, which is based on read-depth differences. The advantage of our approach of using different insert-size libraries is demonstrated by the size range comparison of SV (inversions and deletions) called by each library; although the majority of large SV was detected using paired-end libraries, the long insert-size mate-pair library substantially

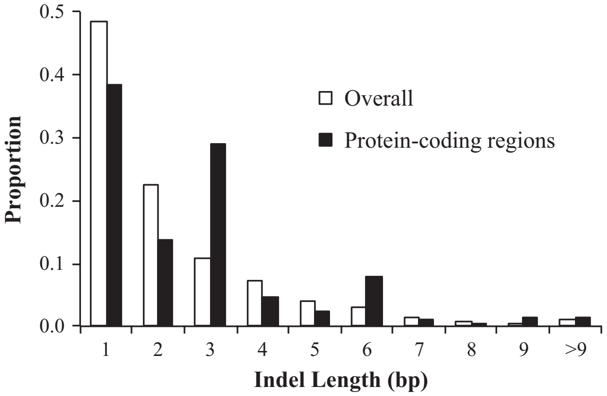

Fig. 2 Indel length distribution. The proportion of indels found overall (white) and only in protein-coding regions (black), binned by indel length.

extended the size spectrum of our SV calls (Fig. 3). Inversions spanned 105 759 SNPs and 9526 indels, and more than half of the called inversions (27 of 48) encompassed protein-coding sequences. Of the 20 autosomes, 16 contained at least one inversion (Table 2). One region along LG XIII appears to be prone to SV, encompassing seven inversions and many CNVRs across <10 Mb (Fig. 1A—'Inversions'). These inversions overlap with 347 genes with no detectable enrichment of GO terms. A summary of the SV detected in this study can be accessed in Table S4 (Supporting information).

Discussion

Adaptive evolution requires genetic change; however, insights into the relative contribution of different types of genetic variation are rare in natural populations. Three-spined sticklebacks are an excellent ecological vertebrate model to study genetic variation in an adaptive context because within a short period of time, these fish have repeatedly colonized various environments (Bell & Foster 1994; Reusch *et al.* 2001). Numerous recurrent genetic and phenotypic patterns have been observed in parallel stickleback systems, and the maintenance of population divergence appears to be driven mainly by ecology (Gow *et al.* 2007; Berner *et al.* 2009; Eizaguirre *et al.* 2009; Kaeuffer *et al.* 2012). This lends support to the notion that standing genetic variation maintained in marine stickleback populations functions as a resource for freshwater colonizations and has probably enabled the rapid diversification and recurring adaptation observed in nature. Therefore, we sought to generate an in-depth evaluation of genome-wide variation in a marine stickleback population. Our focus was on characterizing the breadth of polymorphisms present in a marine population and not on detecting regions under selection. In follow-up studies involving additional individuals, we will build upon this catalogue of variation to further investigate the adaptive relevance of the different types of variation (SNPs, indels and SV) and the generality of outlier patterns in this population and others.

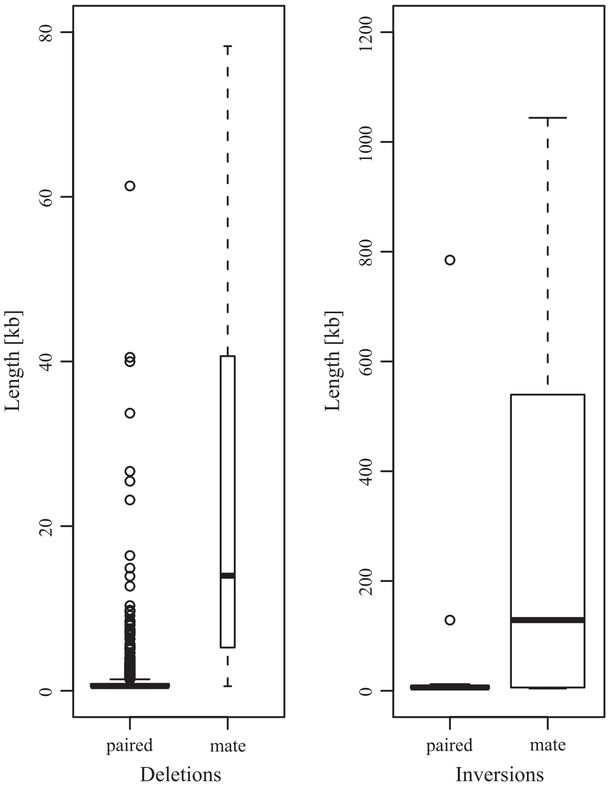

Fig. 3 Length distribution of structural variation (SV) called from two different insert-size libraries: paired-end libraries (insert ~140 bp; paired) and mate-pair libraries (insert ~3 kb; mate). The boxplots show the median (horizontal line), the 25th and 75th percentiles (bottom and top of box), twice the standard deviation (dashed lines) and some outliers (circles; extreme outliers are cut-off for presentations purposes; maximal values are 753 kb for deletions and 3254 kb for inversions). SV calls are much longer if based on mate-pair libraries. The width of the boxplot reflects the relative number of deletions called by library (1077 for paired compared to 35 for mate) and inversions called by library (26 for paired compared to 23 for mate).

The study presented here used next-generation sequencing to detect and characterize numerous forms of genetic variation in a marine population (six individuals) of three-spined sticklebacks. The advantages of whole-genome re-sequencing data include the ability to detect novel SNPs and assess the pervasiveness of a broad size range of variation incorporating indels and larger SV. Generally, population genomic scans are very well suited to demonstrate overall patterns of diversity, while alone they are usually not sufficient to make the final causal connection between genotype and phenotype (Stinchcombe & Hoekstra 2008). Here, we examined the relative abundance and distribution of different types of variation along the stickleback autosomes. This helps assess the contribution of different parts of the genomic architecture in the face of recurrent and rapid adaptations observed in sticklebacks. To some extent, the polymorphism found in our study population represents the standing genetic variation available as a source to fuel and promote the adaptation to diverse surrounding habitats. Overall, we found that 0.75% of the analysed autosomes were affected by SNPs, 0.15% by indels, 4.2% by CNVRs, and another 3.2% were encompassed in inversions. The majority of genetic variation calls were not restricted to one individual (Table 1), suggesting that most variants are not rare. Most genes contained polymorphisms within their protein-coding regions, including 88% with SNPs, 7%

with indels, 6% within CNVRs (including both duplications and deletions) and 4% within inversions. We show that each of these different types of variants contributes substantially to genetic diversity, shaping the overall evolutionary trajectory of the genome.

Single-nucleotide polymorphisms were the most abundant form of detected variation from which we estimated an average π of 0.0025. These results are very similar to previous findings with more individuals (20 per population); based on a RAD approach evaluating <1% of reference genome sites, Hohenlohe *et al.* (2010) estimated that the variation in two marine populations ranged from 0.81% to 1.12%, while the overall average π was approximately 0.002–0.003.

Utilizing our dense marker set, evaluating more than 75% of reference genome sites, we have identified autosomal outlier regions of nucleotide diversity (π) and Tajima's *D* (Table S3, Supporting information). We detected two regions with high π and Tajima's *D* suggestive of balancing selection and seven regions with low π and Tajima's *D* suggestive of directional selection. Our empirical approach attempts to correct for demographic effects and variation in mutation and recombination rates by comparison with LG-specific backgrounds. However, we acknowledge that we cannot entirely rule out other factors as an explanation for the observed outlier patterns. The sample size might prevent inferring the forces shaping variation patterns (demography, drift or selection) but we identified candidate regions with molecular patterns consistent with adaptive evolution. Conversely, we might have overlooked small effect and soft sweep signatures. Comparative approaches contrasting diversity and divergence between populations are a more powerful approach to exploit candidate regions shaped by selection (Mäkinen *et al.* 2008; Hohenlohe *et al.* 2010; DeFaveri *et al.* 2011; Deagle *et al.* 2012; Jones *et al.* 2012a). The difference in approach also explains why a gene such as EDA, involved in adaptation to freshwater habitats, is not highlighted in our analysis of standing genomic variation within a marine population. However, due to the unbiased re-sequencing approach, our findings can be integrated into future studies examining the rich diversity of stickleback populations. In addition to SNP analysis, our sequencing approach integrates and overlaps further types of variation (indels, SV, including CNV), their abundance and potential functional relevance.

Small insertions and deletions (indels) are scattered across genomes and impact exonic regions as has been evinced in humans (Mullaney *et al.* 2010; Mills *et al.* 2011a), cattle (Zhan *et al.* 2011), silkworm (Xia *et al.* 2009) and *Arabidopsis* (Cao *et al.* 2011). We have identified thousands of small indels, some of which cause protein-coding frameshifts and others which occur in parts of genes that might interrupt regulatory function. In contrast to SNPs, there are relatively few indels in coding regions of genes and when they do occur, there is a preference for trinucleotide indels. These observations mirror findings in other species (Mullaney *et al.* 2010; Zhan *et al.* 2011) and are not surprising if there is selection against indels that disrupt protein sequences, such as frameshift mutations within exons. Finding this expected pattern also solicits confidence in our set of indels. Nevertheless, we cannot exclude that some frameshift mutations are due to incorrect gene models or that some indels are sequencing errors. Assuming that this has minimal impact on our overall numbers, we demonstrate the preponderance of multinucleotide polymorphisms contributing to standing genetic variation in a natural population, but that this is levelled out by natural selection in protein-coding regions.

We found that about 7% of the genome varies in structural organization within our samples, encompassing over 10% of genes. This supports previous observations in numerous organisms that SV is a prevalent phenomenon and has the potential to substantially differentiate genomes (Feuk *et al.* 2006; Redon *et al.* 2006; Korbel *et al.* 2007; Conrad *et al.* 2010; Gazave *et al.* 2011; Ventura *et al.* 2011; Yalcin *et al.* 2011). It has been shown that the parallel evolution of the same adaptive phenotype can be caused by SV, such as multiple variants of Pitx1 deletion in sticklebacks. Inherent properties of the genome architecture may influence the amount of SV observed in certain regions such as the high prevalence of deletion mutations in Pitx1 (Chan *et al.* 2010). Unfortunately, such regions can be difficult to access as they are hard to sequence and to assemble (e.g. Pitx1 has not yet been assembled to a LG, while there is evidence that it should be positioned on LG VII; see Chan *et al.* 2010 for details). Mapping issues might at least partially contribute to the rather high false discovery rates of many of the SV detection approaches, especially for low coverage data (Mills *et al.* 2011b). An evaluation of linkage disequilibrium patterns further pointed out that the genomic architecture potentially plays an essential role for divergence and adaptation in the stickleback radiation (Hohenlohe *et al.* 2012). There is also evidence for the impact of SV in other fish such as zebrafish (Brown *et al.* 2012) and salmonids (Miller *et al.* 2012).

Genome evolution is a dynamic process and in constant size-flux, as duplications and deletions can occur even more frequently than point mutations (Lipinski *et al.* 2011). Unbalanced SVs may lead to strong and potentially adaptive phenotypic changes such as altering gene expression or regulation, especially when CNVRs involve gene duplication (Perry *et al.* 2007; Colbourne *et al.* 2011; Zhou *et al.* 2011). CNVRs have associations with numerous human diseases (Zhang *et al.* 2009; Almal & Padh 2012) and are enriched with immune genes in various organisms such as primates (Fortna *et al.* 2004) and cattle (Hou *et al.* 2011; Stothard *et al.* 2011; Bickhart *et al.* 2012). Immune-related genes are often evolving under balancing selection (van Oosterhout 2009), and this could contribute to inhibiting copy-number fixation of these genes (Gokcumen *et al.* 2011). In this study, we found an enrichment of G-Protein-Coupled receptor (GPCR) genes and major histocompatibility complex (MHC) genes in CNVRs. These are both large gene families with functions that involve interacting with extracellular molecules, and as such have adaptive potential (Yokoyama *et al.* 1989; Apanius *et al.* 1997). It has been suggested that MHC genes evolve in a birth-and-death process of frequent gene duplication, pseudogenization and deletion (Nei *et al.* 1997) and field data from three-spined sticklebacks support this view (Lenz *et al.* 2009; Eizaguirre *et al.* 2011). Our data are consistent with this 'accordion model of evolution' (Klein *et al.* 1993) and represent further evidence at the genomic level for an elevated rate of copy-number variation of the MHC loci.

In addition to being a significant contributor to genetic variation, unbalanced SV can also be used in conjunction with other types of polymorphisms to gain greater perspective on the root of molecular signatures. Specifically, overall SNP detection accuracy is affected by CNVRs (Zhan *et al.* 2011), for example with the occurrence of heterozygous SNPs. This raises the concern of attributing phenotypic effects to SNPs in the absence of the recognition of CNV. As an example, we find an enrichment of GPCR genes with high numbers of heterozygous SNPs in their coding regions, suggesting that they may be evolving under strong balancing selection. On the other hand, given that these genes mostly lie in CNVRs, they might instead have multiple copies that map to the same reference position where copy-number and nucleotide sequences of each locus may be evolving neutrally. A visualization of this problem is given for the highlighted region on LG V, where a high outlier consistent with balancing selection overlaps with a pronounced CNVR (Fig. 1B). Consequently, we cannot exclude that multiple copies mapping to one reference region actually cause the increased SNP variation detected in this region. As such, the functional impacts of SNPs and indels inside and near unbalanced SV should be evaluated with caution (Zhan *et al.* 2011). This highlights the importance of interpreting SNP data within its broader genomic context.

Our re-sequencing approach additionally allowed the use of paired-end and split-read information to detect balanced SV such as inversions. Most inversions in our study are large and overlap with numerous genes (Table 1). Because the current methods give rather imprecise breakpoint ranges (predicted start and end position of SV), it is difficult to assess whether gene sequences are interrupted by inversions. However, inversions can impact phenotypic change without affecting coding sequences, for example, by altering expression (Harewood *et al.* 2012). Furthermore, divergence may proceed by reducing recombination in inverted regions (Rieseberg 2001; Navarro & Barton 2003; Kirkpatrick & Barton 2006; Hoffmann & Rieseberg 2008; Faria & Navarro 2010). Therefore, inversions could function as islands of divergence promoting the speciation process. As sticklebacks are one of the prime examples of ecological speciation, with multiple recently diverging populations (Schluter 1996a,b; McKinnon & Rundle 2002; Hendry *et al.* 2009), our data set of inversion calls may provide a starting point to further investigate the role of inversions during population divergence. In this context, inversions fixed between populations would be highly informative. Methods improving the certainty of SV localization and particularly pinpointing breakpoints are general challenges that lie ahead. In particular, the precise determination of SV locations and SV genotypes is essential in calculating allele frequencies and in turn will allow the utilization of SV in a population genetic framework (Conrad & Hurles 2007). This will help to further quantify the adaptive relevance of SV and contrast it to the relative contribution of SNPs.

Our analysis of whole stickleback genomes provides an account of numerous forms of standing genetic variation in a European marine population. Advantages of performing whole-genome sequencing include the ability of characterizing novel genetic variants, which may be particularly of interest when studying populations in their natural ecological context, as well as large structural changes in the organization of the genome. Along with SNPs, we investigated small insertions and deletions as well as larger SV, namely duplications, deletions and inversions. We detected substantial variation despite sampling only six individuals. Some of these variants affect large portions of the genome including entire stretches of exons and genes. Much of this variation is structural, which has arguably been undervalued as a major contributor to population genomic differentiation despite its potential to contribute to phenotypic variation and speciation. We additionally stress that unbalanced SV such as CNVRs can, if neglected, affect inferences made by solely concentrating on genetic variation signals from SNPs. The genetic variation we have described in a natural population brings us closer to understanding the impact of variation (novel or standing) and the influence of genomic architecture on adaptation processes. As freshwater populations are derived from marine populations, this catalogue of variation may be composed of the genetic diversity that has been utilized and promoted the adaptation to various environmental conditions. A next step is to investigate the variability of these sites and regions in other three-spined stickleback populations. To this end, we have provided a data set that can be used to better understand the generality of adaptation from standing genetic variation as well as the role of the genomic architecture during stickleback radiation.

Acknowledgements

We thank Jan Korbel, Antoine Branca, Andrew Moore and Sarah Behrens for their helpful suggestions regarding technical details and statistical analysis. We thank Anna Niehues, Steffen Klasberg, Anja Hasselmeyer and Sybille Liedtke for technical assistance. We also extend our gratitude to Derk Wachsmuth for technical support in computing. Thanks to three anonymous reviewers for advice on improving the manuscript. PGDF is supported by ERG (proposal nr 270891), and MMu was funded by the DFG SPP 1399 (BO-2544/7-1). The project has been mainly financed by a grant from the innovation fund of the Max-Planck-Society to MMi.

References

Abyzov A, Urban AE, Snyder M, Gerstein M (2011) CNVnator: an approach to discover, genotype and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Research*, **21**, 974–984.

Akey JM, Ruhe AL, Akey DT *et al.* (2010) Tracking footprints of artificial selection in the dog genome. *Proceedings of the National Academy of Sciences of the USA*, **107**, 1160–1165.

Alexa A, Rahnenführer J, Lengauer T (2006) Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*, **22**, 1600–1607.

Alkan C, Coe BP, Eichler EE (2011) Genome structural variation discovery and genotyping. *Nature Reviews Genetics*, **12**, 363–376.

Almal SH, Padh H (2012) Implications of gene copy-number variation in health and diseases. *Journal of Human Genetics*, **57**, 6–13.

Apanius V, Penn D, Slev PR, Ruff LR, Potts WK (1997) The nature of selection on the major histocompatibility complex. *Critical Reviews in Immunology*, **17**, 179–224.

Baerwald M, Welsh A, Hedrick R, May B (2008) Discovery of genes implicated in whirling disease infection and resistance in rainbow trout using genome-wide expression profiling. *BMC Genomics*, **9**, 37.

Barrett RDH, Schluter D (2008) Adaptation from standing genetic variation. *Trends in Ecology & Evolution*, **23**, 38–44.

Barrett RDH, Rogers SM, Schluter D (2008) Natural selection on a major armor gene in threespine stickleback. *Science*, **322**, 255–257.

Bell MA, Foster SA (1994) *The Evolutionary Biology of the Threespine Stickleback*. Oxford University Press, Oxford.

Berner D, Grandchamp A-C, Hendry AP (2009) Variable progress towards ecological speciation in parapatry: stickleback across eight lake-stream transitions. *Evolution*, **63**, 1740–1753.

Bickhart DM, Hou Y, Schroeder SG *et al.* (2012) Copy number variation of individual cattle genomes using next-generation sequencing. *Genome Research*, **22**, 778–790.

Broad Institute (2007) Stickleback draft genome assembly: gasAcu1.0. (http://www.broadinstitute.org/models/stickleback).

Brown KH, Dobrinski KP, Lee AS *et al.* (2012) Extensive genetic diversity and substructuring among zebrafish strains revealed through copy number variant analysis. *Proceedings of the National Academy of Sciences of the USA*, **109**, 529–534.

Browning BL, Browning SR (2009) A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *American Journal of Human Genetics*, **84**, 210–223.

Cao J, Schneeberger K, Ossowski S *et al.* (2011) Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nature Genetics*, **43**, 956–963.

Chan YF, Marks ME, Jones FC *et al.* (2010) Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. *Science*, **327**, 302–305.

Chen Z, Cheng C-HC, Zhang J *et al.* (2008) Transcriptomic and genomic evolution under constant cold in Antarctic notothenioid fish. *Proceedings of the National Academy of Sciences of the USA*, **105**, 12944–12949.

Chen K, Wallis JW, McLellan MD *et al.* (2009) BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature Methods*, **6**, 677–681.

Cheng C, White BJ, Kamdem C *et al.* (2012) Ecological genomics of *Anopheles gambiae* along a latitudinal cline in Cameroon: a population resequencing approach. *Genetics*, **190**, 1417–1432.

Cingolani P (2012) snpEff: variant effect prediction. http://snpeff.sourceforge.net.

Colbourne JK, Pfrender ME, Gilbert D *et al.* (2011) The ecoresponsive genome of *Daphnia pulex*. *Science*, **331**, 555–561.

Colosimo PF, Hosemann KE, Balabhadra S *et al.* (2005) Widespread parallel evolution in sticklebacks by repeated fixation of Ectodysplasin alleles. *Science*, **307**, 1928–1933.

Conrad DF, Hurles ME (2007) The population genetics of structural variation. *Nature Genetics*, **39**, S30–S36.

Conrad DF, Pinto D, Redon R *et al.* (2010) Origins and functional impact of copy number variation in the human genome. *Nature*, **464**, 704.

Cridland JM, Thornton KR (2010) Validation of rearrangement break points identified by paired-end sequencing in natural populations of *Drosophila melanogaster*. *Genome Biology and Evolution*, **2010**, 83–101.

Danecek P, Auton A, Abecasis G *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.

Deagle BE, Jones FC, Chan YF *et al.* (2012) Population genomics of parallel phenotypic evolution in stickleback across stream—lake ecological transitions. *Proceedings of the Royal Society. Biological Sciences Series B*, **279**, 1277–1286.

DeBolt S (2010) Copy number variation shapes genome diversity in *Arabidopsis* over immediate family generational scales. *Genome Biology and Evolution*, **2**, 441–453.

DeFaveri J, Shikano T, Shimada Y, Goto A, Merilä J (2011) Global analysis of genes involved in freshwater adaptation in threespine sticklebacks (*Gasterosteus aculeatus*). *Evolution*, **65**, 1800–1807.

DePristo MA, Banks E, Poplin R *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, **43**, 491–498.

Eizaguirre C, Lenz TL, Traulsen A, Milinski M (2009) Speciation accelerated and stabilized by pleiotropic major histocompatibility complex immunogenes. *Ecology Letters*, **12**, 5–12.

Eizaguirre C, Lenz T, Sommerfeld R *et al.* (2011) Parasite diversity, patterns of MHC II variation and olfactory based mate choice in diverging three-spined stickleback ecotypes. *Evolutionary Ecology*, **25**, 605–622.

Eizaguirre C, Lenz TL, Kalbe M, Milinski M (2012) Rapid and adaptive evolution of MHC genes under parasite selection in experimental vertebrate populations. *Nature Communications*, **3**, 621.

Ellegren H, Sheldon BC (2008) Genetic basis of fitness differences in natural populations. *Nature*, **452**, 169–175.

Elmer KR, Meyer A (2011) Adaptation in the age of ecological genomics: insights from parallelism and convergence. *Trends in Ecology & Evolution*, **26**, 298–306.

Emerson JJ, Cardoso-Moreira M, Borevitz JO, Long M (2008) Natural selection shapes genome-wide patterns of copy-number polymorphism in *Drosophila melanogaster*. *Science*, **320**, 1629–1631.

Faria R, Navarro A (2010) Chromosomal speciation revisited: rearranging theory with pieces of evidence. *Trends in Ecology & Evolution*, **25**, 660–669.

Feder JL, Roethele JB, Filchak K, Niedbalski J, Romero-Severson J (2003) Evidence for inversion polymorphism related to sympatric host race formation in the apple maggot fly, *Rhagoletis pomonella*. *Genetics*, **163**, 939–953.

Feuk L, Carson AR, Scherer SW (2006) Structural variation in the human genome. *Nature Reviews Genetics*, **7**, 85–97.

Flicek P, Amode MR, Barrell D *et al.* (2011) Ensembl 2011. *Nucleic Acids Research*, **39**, D800–D806.

Fortna A, Kim Y, MacLaren E *et al.* (2004) Lineage-specific gene duplication and loss in human and great ape evolution. *PLoS Biology*, **2**, e207.

Gazave E, Darré F, Morcillo-Suarez C *et al.* (2011) Copy number variation analysis in the great apes reveals species-specific patterns of structural variation. *Genome Research*, **21**, 1626–1639.

Gibson G (2005) EVOLUTION: the synthesis and evolution of a supermodel. *Science*, **307**, 1890–1891.

Gokcumen O, Babb P, Iskow R *et al.* (2011) Refinement of primate copy number variation hotspots identifies candidate genomic regions evolving under positive selection. *Genome Biology*, **12**, R52.

Gow JL, Peichel CL, Taylor EB (2007) Ecological selection against hybrids in natural populations of sympatric threespine sticklebacks. *Journal of Evolutionary Biology*, **20**, 2173–2180.

Gratten J, Wilson AJ, McRae AF *et al.* (2008) A localized negative genetic correlation constrains microevolution of coat color in wild sheep. *Science*, **319**, 318–320.

Handsaker RE, Korn JM, Nemesh J, McCarroll SA (2011) Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nature Genetics*, **43**, 269–276.

Harewood L, Chaignat E, Reymond A (2012) Structural variation and its effect on expression. *Methods in Molecular Biology*, **838**, 173–186.

Hendry AP, Bolnick DI, Berner D, Peichel CL (2009) Along the speciation continuum in sticklebacks. *Journal of Fish Biology*, **75**, 2000–2036.

Hermisson J, Pennings PS (2005) Soft sweeps. *Genetics*, **169**, 2335–2352.

Hoffmann AA, Rieseberg LH (2008) Revisiting the impact of inversions in evolution: from population genetic markers to drivers of adaptive shifts and speciation? *Annual Review of Ecology Evolution and Systematics*, **39**, 21–42.

Hohenlohe PA, Bassham S, Etter PD *et al.* (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genetics*, **6**, e1000862.

Hohenlohe PA, Bassham S, Currey M, Cresko WA (2012) Extensive linkage disequilibrium and parallel adaptive divergence across threespine stickleback genomes. *Philosophical Transactions of the Royal Society of London. Series B*, **367**, 395–408.

Hou Y, Liu G, Bickhart D *et al.* (2011) Genomic characteristics of cattle copy number variations. *BMC Genomics*, **12**, 127.

Jiang J, Zhang Y-B, Li S *et al.* (2009) Expression regulation and functional characterization of a novel interferon inducible gene Gig2 and its promoter. *Molecular Immunology*, **46**, 3131–3140.

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

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

Kaeuffer R, Peichel CL, Bolnick DI, Hendry AP (2012) Parallel and nonparallel aspects of ecological, phenotypic, and genetic divergence across replicate population pairs of lake and stream stickleback. *Evolution*, **66**, 402–418.

Kirkpatrick M, Barton N (2006) Chromosome inversions, local adaptation and speciation. *Genetics*, **173**, 419–434.

Kitano J, Bolnick DI, Beauchamp DA *et al.* (2008) Reverse evolution of armor plates in the threespine stickleback. *Current Biology*, **18**, 769–774.

Klein J, Ono H, Klein D, O'hUigan C (1993) The accordion model of MHC evolution. *Progress in Immunology*, **8**, 137–143.

Kolaczkowski B, Kern AD, Holloway AK, Begun DJ (2011) Genomic differentiation between temperate and tropical Australian populations of *Drosophila melanogaster*. *Genetics*, **187**, 245–260.

Korbel JO, Urban AE, Affourtit JP *et al.* (2007) Paired-end mapping reveals extensive structural variation in the human genome. *Science*, **318**, 420–426.

Krzywinski M, Schein J, Birol I *et al.* (2009) Circos: an information aesthetic for comparative genomics. *Genome Research*, **19**, 1639–1645.

Lai J, Li R, Xu X *et al.* (2010) Genome-wide patterns of genetic variation among elite maize inbred lines. *Nature Genetics*, **42**, 1027–1030.

Lam H-M, Xu X, Liu X *et al.* (2010) Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nature Genetics*, **42**, 1053–1059.

Lenz TL, Eizaguirre C, Becker S, Reusch TBH (2009) RSCA genotyping of MHC for high-throughput evolutionary studies in the model organism three-spined stickleback *Gasterosteus aculeatus*. *BMC Evolutionary Biology*, **9**, 57.

Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, **18**, 1851–1858.

Li H, Handsaker B, Wysoker A *et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Lipinski KJ, Farslow JC, Fitzpatrick KA *et al.* (2011) High spontaneous rate of gene duplication in *Caenorhabditis elegans*. *Current Biology*, **21**, 306–310.

Lowry DB, Willis JH (2010) A widespread chromosomal inversion polymorphism contributes to a major life-history transition, local adaptation, and reproductive isolation. *PLoS Biology*, **8**, e1000500.

Mäkinen HS, Cano JM, Merilä J (2008) Identifying footprints of directional and balancing selection in marine and freshwater three-spined stickleback *Gasterosteus aculeatus* populations. *Molecular Ecology*, **17**, 3565–3582.

Manceau M, Domingues VS, Linnen CR, Rosenblum EB, Hoekstra HE (2010) Convergence in pigmentation at multiple levels: mutations, genes and function. *Philosophical Transactions of the Royal Society of London. Series B*, **365**, 2439–2450.

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

McKinnon JS, Rundle HD (2002) Speciation in nature: the threespine stickleback model systems. *Trends in Ecology & Evolution*, **17**, 480–488.

Medvedev P, Stanciu M, Brudno M (2009) Computational methods for discovering structural variation with next-generation sequencing. *Nature Methods*, **6**, S13–S20.

Miller MR, Brunelli JP, Wheeler PA *et al.* (2012) A conserved haplotype controls parallel adaptation in geographically distant salmonid populations. *Molecular Ecology*, **21**, 237–249.

Mills RE, Pittard WS, Mullaney JM *et al.* (2011a) Natural genetic variation caused by small insertions and deletions in the human genome. *Genome Research*, **21**, 830–839.

Mills RE, Walter K, Stewart C *et al.* (2011b) Mapping copy number variation by population-scale genome sequencing. *Nature*, **470**, 59–65.

Mullaney JM, Mills RE, Pittard WS, Devine SE (2010) Small insertions and deletions (INDELs) in human genomes. *Human Molecular Genetics*, **19**, R131–R136.

Navarro A, Barton NH (2003) Accumulating postzygotic isolation genes in parapatry: a new twist on chromosomal speciation. *Evolution*, **57**, 447–459.

Nei M, Gu X, Sitnikova T (1997) Evolution by the birth-and-death process in multigene families of the vertebrate immune system. *Proceedings of the National Academy of Sciences of the USA*, **94**, 7799–7806.

van Oosterhout C (2009) A new theory of MHC evolution: beyond selection on the immune genes. *Proceedings of the Royal Society. Biological Sciences Series B*, **276**, 657–665.

Orr HA, Betancourt AJ (2001) Haldane's sieve and adaptation from the standing genetic variation. *Genetics*, **157**, 875–884.

Perry GH, Dominy NJ, Claw KG *et al.* (2007) Diet and the evolution of human amylase gene copy number variation. *Nature Genetics*, **39**, 1256–1260.

Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.

Quinlan AR, Clark RA, Sokolova S *et al.* (2010) Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. *Genome Research*, **20**, 623–635.

R Development Core Team (2011) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Redon R, Ishikawa S, Fitch KR *et al.* (2006) Global variation in copy number in the human genome. *Nature*, **444**, 444–454.

Reusch TBH, Wegner KM, Kalbe M (2001) Rapid genetic divergence in postglacial populations of threespine stickleback (*Gasterosteus aculeatus*): the role of habitat type, drainage and geographical proximity. *Molecular Ecology*, **10**, 2435–2445.

Rieseberg LH (2001) Chromosomal rearrangements and speciation. *Trends in Ecology & Evolution*, **16**, 351–358.

Rubin C-J, Zody MC, Eriksson J *et al.* (2010) Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature*, **464**, 587.

Rundle HD, Nosil P (2005) Ecological speciation. *Ecology Letters*, **8**, 336–352.

Schluter D (1996a) Ecological causes of adaptive radiation. *The American Naturalist*, **148**, S40–S64.

Schluter D (1996b) Ecological speciation in postglacial fishes. *Philosophical Transactions of the Royal Society of London. Series B*, **351**, 807–814.

Schluter D, Conte GL (2009) Genetics and ecological speciation. *Proceedings of the National Academy of Sciences of the USA*, **106**, 9955–9962.

Stapley J, Reger J, Feulner PGD *et al.* (2010) Adaptation genomics: the next generation. *Trends in Ecology & Evolution*, **25**, 705–712.

Stinchcombe JR, Hoekstra HE (2008) Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. *Heredity*, **100**, 158–170.

Stothard P, Choi J-W, Basu U *et al.* (2011) Whole genome resequencing of black Angus and Holstein cattle for SNP and CNV discovery. *BMC Genomics*, **12**, 559.

Teotonio H, Chelo IM, Bradic M, Rose MR, Long AD (2009) Experimental evolution reveals natural selection on standing genetic variation. *Nature Genetics*, **41**, 251–257.

Thornton K (2003) Libsequence: a C++ class library for evolutionary genetic analysis. *Bioinformatics*, **19**, 2325–2327.

Ventura M, Catacchio CR, Alkan C *et al.* (2011) Gorilla genome structural variation reveals evolutionary parallelisms with chimpanzee. *Genome Research*, **21**, 1640–1649.

Xia Q, Guo Y, Zhang Z *et al.* (2009) Complete resequencing of 40 genomes reveals domestication events and genes in silkworm (*Bombyx*). *Science*, **326**, 433–436.

Xu X, Liu X, Ge S *et al.* (2012) Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nature Biotechnology*, **30**, 105–111.

Yalcin B, Wong K, Agam A *et al.* (2011) Sequence-based characterization of structural variation in the mouse genome. *Nature*, **477**, 326–329.

Ye K, Schulz MH, Long Q, Apweiler R, Ning Z (2009) Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, **25**, 2865–2871.

Yokoyama S, Isenberg KE, Wright AF (1989) Adaptive evolution of G-protein coupled receptor genes. *Molecular Biology and Evolution*, **6**, 342–353.

Zhan B, Fadista J, Thomsen B *et al.* (2011) Global assessment of genomic variation in cattle by genome resequencing and high-throughput genotyping. *BMC Genomics*, **12**, 557.

Zhang F, Gu W, Hurles ME, Lupski JR (2009) Copy number variation in human health, disease, and evolution. *Annual Review of Genomics and Human Genetics*, **10**, 451–481.

Zhou J, Lemos B, Dopman EB, Hartl DL (2011) Copy-number variation: the balance between gene dosage and expression in *Drosophila melanogaster*. *Genome Biology and Evolution*, **3**, 1014–1024.

P.G.D.F. and F.J.J.C wrote the manuscript. P.G.D.F, F.J.J.C and M.P. contributed to the design of the study and analysed the data. C.E. performed the sampling. C.E., M.K., T.L.L., M.Mu. and I.S.P. contributed to the design of the study and the analysis of the data. M.S., M.Mi., T.B.H.R. and E.B.B. designed the study and contributed to the analysis of the data.

Data accessibility

Data have been deposited in the EBI Sequence Read Archive (SRA) with accession no. ERP001339.

Supporting information

Additional supporting information may be found in the online version of this article.

Table S1 List of quality estimates of sequencing.

Table S2 List of assembly quality estimates.

Table S3 Location of outliers on autosomal linkage groups, with start position (zero-based) of 100 kb windows, the number of genotyped sites (after masking) within the window, and the the top (99) or bottom (01) 1% windows designated as outliers for either π or Tajima's *D*, or both.

Table S4 List of large SV and the method by which they are called.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.