BMC
Genomics

RESEARCH ARTICLE   **Open Access**

Genomic divergence between nine- and three-spined sticklebacks

Baocheng Guo[1*], Frédéric JJ Chain[2], Erich Bornberg-Bauer[3], Erica H Leder[4] and Juha Merilä[1]

Abstract

Background: Comparative genomics approaches help to shed light on evolutionary processes that shape differentiation between lineages. The nine-spined stickleback (*Pungitius pungitius*) is a closely related species of the ecological 'supermodel' three-spined stickleback (*Gasterosteus aculeatus*). It is an emerging model system for evolutionary biology research but has garnered less attention and lacks extensive genomic resources. To expand on these resources and aid the study of sticklebacks in a phylogenetic framework, we characterized nine-spined stickleback transcriptomes from brain and liver using deep sequencing.

Results: We obtained nearly eight thousand assembled transcripts, of which 3,091 were assigned as putative one-to-one orthologs to genes found in the three-spined stickleback. These sequences were used for evaluating overall differentiation and substitution rates between nine- and three-spined sticklebacks, and to identify genes that are putatively evolving under positive selection. The synonymous substitution rate was estimated to be 7.1×10^{-9} per site per year between the two species, and a total of 165 genes showed patterns of adaptive evolution in one or both species. A few nine-spined stickleback contigs lacked an obvious ortholog in three-spined sticklebacks but were found to match genes in other fish species, suggesting several gene losses within 13 million years since the divergence of the two stickleback species. We identified 47 SNPs in 25 different genes that differentiate pond and marine ecotypes. We also identified 468 microsatellites that could be further developed as genetic markers in nine-spined sticklebacks.

Conclusion: With deep sequencing of nine-spined stickleback cDNA libraries, our study provides a significant increase in the number of gene sequences and microsatellite markers for this species, and identifies a number of genes showing patterns of adaptive evolution between nine- and three-spined sticklebacks. We also report several candidate genes that might be involved in differential adaptation between marine and freshwater nine-spined sticklebacks. This study provides a valuable resource for future studies aiming to identify candidate genes underlying ecological adaptation in this and other stickleback species.

Keywords: *Pungitus pungitius*, *Gasterosteus aculeatus*, Comparative genomics, Transcriptome, Substitution rate, Adaptive evolution

Background

The rapid advances in sequencing technologies have facilitated the development of comparative genomics – an important approach in contemporary evolutionary biology research [1,2]. The stickleback fishes (Gasterosteidae) provide an excellent model system for such comparative studies. The three-spined stickleback (*Gasterosteus aculeatus*) has become a vertebrate

'supermodel' allowing a combination of studies at molecular, developmental, phenotypic, and population genetic levels to explore factors and processes relevant for adaptive evolution in ecologically relevant contexts [3,4]. The three-spined stickleback is a small teleost populating diverse ecosystems across a wide geographic distribution in the northern hemisphere and occurs in marine, brackish, and freshwater habitats. Populations that have colonized freshwater habitats after the retreat of Pleistocene ice sheets have evolved remarkable morphological and behavioral diversity as compared to marine populations [5,6]. For example, they have repeatedly

* Correspondence: baochengguo@gmail.com
[1]Ecological Genetics Research Unit, Department of Biosciences, University of Helsinki, Helsinki, Finland
Full list of author information is available at the end of the article

evolved changes in body shape, skeletal armor, trophic apparati, pigmentation, osmoregulatory functions, life history, and behavior [5]. The genetic architecture for several of these phenotypic adaptations has been – or is being – deciphered [7-15]. Interestingly, the parallel evolution (similar phenotypes evolving independently in different populations derived from a common ancestor) of armor loss, pelvic reduction, and pigmentation has been found to result from parallel genetic changes in similar genes [8,9,11,14]. However, relatively little is known about the genetics of these or other traits in other stickleback species (but see: [16-20]).

The nine-spined stickleback (*Pungitius pungitius*) is an emerging model for evolutionary biology research [21] and has diverged from the three-spined stickleback around 13 million years ago (Mya) [22], but the two species are ecologically – and to some degree also phenotypically – very similar [23]. Phylogeographic and population genetic analyses of the nine-spined stickleback demonstrate that their colonization and adaptation to freshwater habitats from marine environments has occurred independently multiple times [24-26]. Meanwhile, freshwater nine-spined sticklebacks have also evolved – repeatedly and independently – similar morphological [26-28], behavioral [29,30], neurological [31-36], and physiological [37,38] phenotypes in different localities. Notably, similar adaptive traits also have been evolved in parallel between nine- and three-spined sticklebacks [6]. For example, both marine nine- and three-spined sticklebacks have a complete pelvis, but several different freshwater populations in both species have undergone a genetically based reduction - or even total loss - of the pelvic girdle and associated spines [16-18]. However, it is still uncertain whether or not the genetic underpinnings of the pelvic reduction in nine- and three-spined sticklebacks are the same. For instance, Shapiro et al. [16] first suggested that changes of *Pitx1* expression might contribute to pelvic reduction in both species, but later discovered that the major loci controlling for pelvic development were completely different between the two species. This suggests that the pelvic reduction in these species is an example of genetic convergence [17] (but see [18]). Hence, nine- and three-spined sticklebacks offer a powerful opportunity to study whether or not similar phenotypic changes across species are associated with the same genes or genetic mechanisms.

A genome-wide comparative study can help us to better understand how selection has shaped divergence and illuminate the genetic basis of parallel evolution in nine- and three-spined sticklebacks. It can also reveal the extent of genome-wide differentiation across protein-coding and non-coding regions and the prevalence of species-specific genes that may influence the evolu-tionary trajectory of divergent species. However, compared to the three-spined stickleback with abundant genomic resources [3,39,40], genomic resources for the nine-spined stickleback are still largely lacking (but see: [17,41,42]). For example, development of microsatellite markers for study of nine-spined stickleback currently is based on the three-spined genome, but cross-species utility of microsatellite primers is limited due to low amplification success [43]. Fortunately, the recent explosion of affordable Next Generation Sequencing (NGS) technology provides evolutionary and ecological researchers a great opportunity to conduct genome-wide studies of non-model organisms with limited genetic and genomic resources [44-46]. For instance, transcriptome, a collection of expressed sequences, represents a sample of the spatiotemporally expressed genome that can be used for comparative genomic studies at an interspecific level, as well as genetic diversity analyses at an intraspecific level. Here, we used deep sequencing (454 GS FLX) to characterize partial brain and liver transcriptomic libraries of nine-spined sticklebacks from marine and freshwater populations exhibiting a high degree of morphological and genetic divergence [26-38,41]. With the resulting transcripts, we (1) characterized the sequence divergence between the two closely related stickleback species, (2) investigated rates of molecular evolution for patterns consistent with positive selection, and (3) evaluated sequence differentiation between marine and freshwater nine-spined sticklebacks.

Results

Sequencing and assembly
We obtained a total 337,630 high quality reads with mean length of 250 bp from 454 sequencing of four cDNA libraries from nine-spined sticklebacks (Additional file 1: Figure S1). Contig assembly of the reads were combined from the four cDNA libraries into one "nine-spined stickleback transcriptome" containing 7,932 contigs ≥ 100 bp (median = 403 bp, Additional file 1: Figure S2) with an average coverage depth of 38 reads (Additional file 2: Table S1).

Functional annotation
A BLASTX search returned 3,347 (42.2%) nine-spined stickleback contigs with significant hits to three-spined stickleback genes. This proportion of contigs with BLAST hits is similar to previous transcriptome studies [47-49], in which contigs without significant hits may consist of untranslated transcripts, chimeras or assembly artifacts. Blast2Go with the Gene Ontology (GO) annotations database was used for further annotation and 2,071 contigs have one or more GO terms (Additional file 1: Figure S3). We additionally found that 104 contigs had no significant BLASTX hit with protein sequences

from the three-spined stickleback but had significant hits with protein sequences in at least one of the other seven fish genomes available from Ensembl. By using BLASTN and BLAT searches, we confirmed that 15 of the 104 contigs had no hits in the current three-spined stickleback genome (Additional file 2: Table S2). Because these contigs correspond to genes in other teleost genomes, this suggests that the orthologous sequences of these contigs have probably been lost in the three-spined stickleback rather than gained in nine-spined sticklebacks.

Sequence comparison between nine- and three-spined sticklebacks

We found that 3,091 out of the 3,347 nine-spined stickleback contigs (92.4%) had a pairwise $K_s \leq 0.5$ compared to their three-spined stickleback orthologs (Figure 1), and these had an average length of 690 bp (Additional file 1: Figure S4). We restricted all further analyses to these 3,091 contigs, or "unigenes", in an attempt to curtail the effects of erroneously called orthologs with large K_s values. The corresponding genes are more or less evenly distributed across the three-spined stickleback genome with 2.3% to 7.1% of genes on each chromosome, and the gene number per chromosome is positively correlated with chromosome size ($r_s = 0.84$, $P < 1.2 \times 10^{-6}$, in Additional file 1: Figure S5). Given the conserved genomic synteny between the two species [35], these observations suggest that the unigenes are a relatively unbiased sample of nine-spined stickleback genes in terms of genomic distribution.

We used three methods to detect positive selection on genes in sticklebacks. We first calculated the pairwise
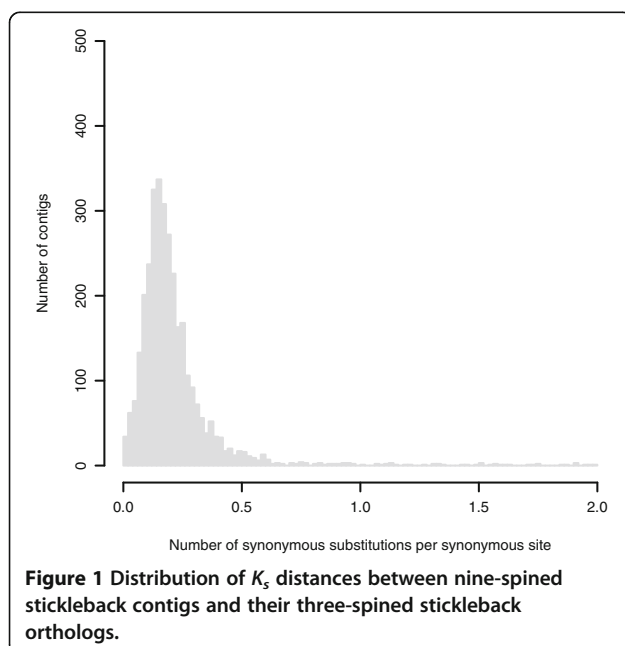
substitution rates K_s, K_a, and K_a/K_s between the nine-spined stickleback unigenes and their putative orthologs in the three-spined stickleback (Figure 2). Genes are generally under strong purifying selection (low K_a/K_s values), with a mean pairwise K_s value was 0.1841 ± 0.0017 (mean ± SD). A total of 194 (6.3%) orthologous pairs had a K_a/K_s ratio between 0.5 and 1 (points above the grey line in Figure 2), and 74 (2.4%) had a K_a/K_s ratio > 1 (points above the black line in Figure 2). The latter 74 unigenes are distributed across 16 chromosomes (Additional file 3: Table S3).

We also performed the branch-site test with medaka as an outgroup to detect positive selection operating on sites along each stickleback lineage. The branch-site test revealed a total of 33 unigenes ($p < 0.05$, eight after multiple test correction with q-value < 0.05) that are putatively evolving under positive selection in the nine-spined stickleback lineage and 39 unigenes (seven after multiple test correction) in the three-spined stickleback lineage (Additional file 3: Table S4). We also found 82 unigenes (37 after multiple test correction) with sites evolving under positive selection in the ancestral lineage before the split between nine- and three-spined sticklebacks (Additional file 3: Table S4).

A third method was used for inferring positive selection by utilizing nine-spined stickleback SNPs. We analyzed the patterns of selection among genes with the MK test and the direction of selection (DoS). We found 48 unigenes that departed from neutrality (Chi-square test with Yates correction, df = 1, $p < 0.05$), 18 of which show a signature of positive selection (Additional file 3:

Figure 1 Distribution of K_s distances between nine-spined stickleback contigs and their three-spined stickleback orthologs.

Figure 2 Distribution of pairwise K_a and K_s distances between nine-spined stickleback unigenes and their three-spined stickleback orthologs. Genes with K_a/K_s ratio >1 fall above the black line while those with K_a/K_s ratio between 0.5 and 1 fall between the gray and black lines.
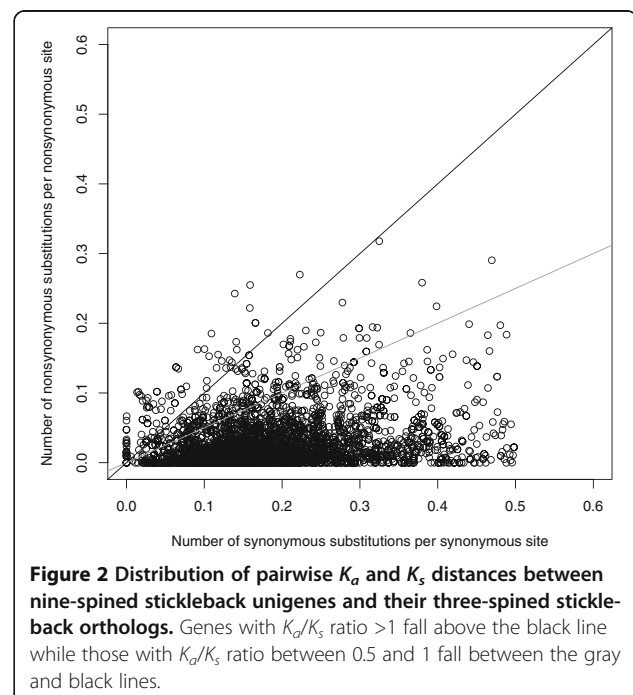
Table S5). However, none of these signatures remained statistically significant after correction for multiple tests.

It is noteworthy that positive selection on seven genes was detected by at least two of the three methods mentioned above. For example, two genes with a pairwise K_a/K_s ratio ≥ 1 that are involved in lipid transport are also detected using the branch-site test, of which one gene (apolipoprotein B - ENSGACG00000009637) is consistent with positive selection in the nine-spined stickleback lineage and the other gene (a vitellogenin gene - ENSGACG00000009711) is consistent with positive selection in the three-spined stickleback lineage. Other overlaps from methods of detecting positive selection include a gene (adenylate cyclase 6 - ENSGACG 00000008575) detected by the MK test and the branch-site test in the nine-spined stickleback lineage, and four genes (complement factor H-related 3 - ENSGACG000 00001733, fetuin B - ENSGACG00000005690, HECT domain containing E3 ubiquitin protein ligase 1 - ENSG ACG00000012853 and an uncharacterized gene - ENS GACG00000007507) detected by both pairwise K_a/K_s and the MK test. Combining all three tests, we found a total of 165 genes with patterns of adaptive evolution in either the nine- or three-spined stickleback, or both. These genes are distributed rather evenly across all of the three-spined stickleback chromosomes except XIV (Additional file 1: Figure S5). We found 126 of these 176 genes with associated GO annotations spanning a broad range of functions (Figure 3). We found that nine GO terms (*viz.* translational elongation, macromolecule

biosynthesis, protein biosynthesis, translation, cellular biosynthesis, physiological process, macromolecule metabolism, biosynthesis, and energy derivation by oxidation of organic compounds) were significantly overrepresented among these 165 genes by comparing to all three-spined stickleback genes, which suggested that these 165 genes have been subject to adaptive evolution (family-wise error rate, $P < 0.05$; Additional file 3: Table S6).

A total of 368 nine-spined stickleback unigenes contained partially sequenced UTRs ≥ 50 bp. The average K2P distance of these UTRs and their three-spined stickleback orthologous sequence was 0.0709 ± 0.0020 (median = 0.0688), whereas the average K2P distance of the coding regions for these same genes was 0.0513 ± 0.0017 (median = 0.0436). The average pairwise K_s for the 368 unigenes was 0.1746 ± 0.0047 (median = 0.1637) and is close to that of the all 3,091 unigenes (mean = 0.1841 ± 0.0017; median = 0.1687), which suggests no bias of the 368 unigenes with UTR information, at least with respect to K_s. The divergence of UTRs was significantly higher as compared to the divergence in corresponding coding regions (Wilcoxon signed rank test, $P < 2.2 \times 10^{-16}$) but significantly lower than that of K_s (Wilcoxon signed rank test, $P < 2.2 \times 10^{-16}$), which suggests that UTRs have evolved under lower selective pressures than coding regions, albeit not neutrally (assuming that synonymous sites evolve close to neutrally). Based on the divergence estimates above and the species divergence time of 13 Mya [22], we calculated the substitution rate as 2.0×10^{-9} per site per year in coding regions
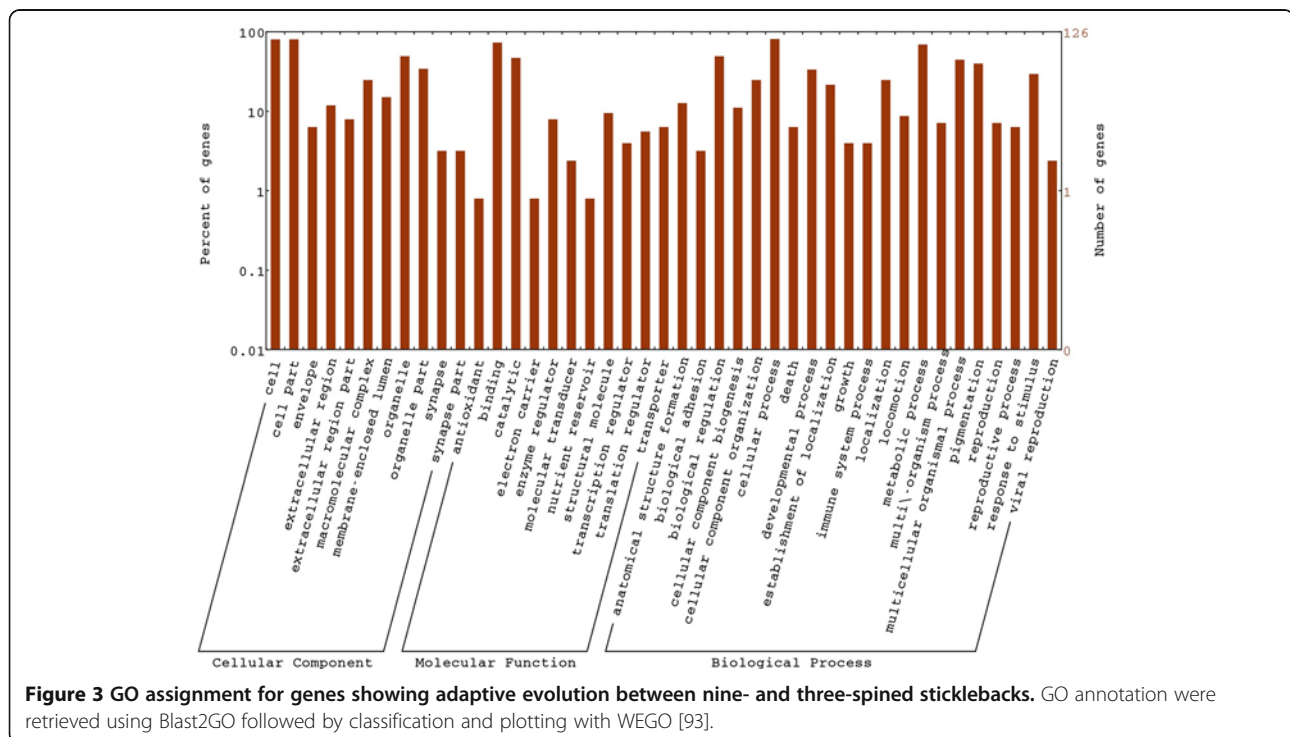

Figure 3 GO assignment for genes showing adaptive evolution between nine- and three-spined sticklebacks. GO annotation were retrieved using Blast2GO followed by classification and plotting with WEGO [93].

(including both synonymous and nonsynonymous sites) and 2.7×10^{-9} per site per year in UTRs between nine- and three-spined sticklebacks.

Divergence between marine and freshwater nine-spined sticklebacks

We found 1,814 SNPs (0.044% of evaluated genic sites) among 718 unigenes in the sampled nine-spined sticklebacks (934 unique SNPs in the marine sample, 642 unique SNPs in the pond sample, and 238 SNPs shared in common). Many of the SNPs (567) are predicted to be nonsynonymous changes, while the remaining are either synonymous (665) or in UTRs (582) (Additional file 3: Table S7). We found 47 SNPs in 28 unigenes (spanning 25 different genes) that lead to 'fixed' genotypes between the two ecotypes, including 17 homozygous differences. These divergent SNPs occur in both tissue types and as such are not tissue-specific differences but most probably reflect general genetic differences between the ecotypes (at least from the sampled individuals). Of the fixed homozygous differences, five are nonsynonymous SNPs, ten are synonymous SNPs and two are SNPs found in UTRs (Additional file 3: Table S8).

Discovery of microsatellite markers

Microsatellites are important genetic markers for non-model organisms and have been widely used for studies of nine-spined sticklebacks [18,19,24,41,43]. We analyzed the nine-spined stickleback unigenes to identify microsatellite markers. We obtained 468 SSRs in 358 unigenes (Additional file 3: Table S9). In terms of abundance, dinucleotide repeats were most abundant (178, 38.0%) followed by trinucleotide repeats (148, 31.6%), mononucleotide repeats (139, 29.7%), 1 tetranucleotide repeat, and 2 hexanucleotide repeats. Of the 468 SSRs, 428 are perfect and 40 are compound. AC/GT (124 out of 178, 69.7%) was the most abundant dinucleotide repeat motif and AGG/CTT (58 out of 148, 39.2) was the most abundant trinucleotide repeat motif.

Discussion
The nine-spined stickleback transcriptome

In recent years, the use of comparative genomic approaches in a phylogenetic framework has shed much light on a variety of fundamental evolutionary questions, such as adaptive evolution [3,50,51], genetic variation [52-55], and speciation [56-58]. Development of genomic resources is the first step towards such biological questions. Using 454 pyrosequencing, we have contributed to the improvement of genomic resources for nine-spined sticklebacks. We provide over three thousand transcript sequences that correspond to an orthologous gene in the three-spined stickleback, and report

hundreds of genic microsatellites that can be used as markers in future experiments with nine-spined sticklebacks. The data provided here significantly increase the number of available gene sequences for nine-spined sticklebacks since there are currently fewer than 1,000 sequence entries in the National Center for Biotechnology Information. Given its status as an emerging model for evolutionary biology research [21], this transcriptomic data will be of interest to researchers investigating the evolution of nine-spined sticklebacks, for example by using the identified SNPs or microsatellite markers for population genetics studies. It also allows for more refined inferences concerning stickleback and teleost evolution in a phylogenetic framework by providing orthologs of closely related fish species. Thus, apart from contributing a large number of new gene sequences to the research domain, the results of this study represent the first reported nine-spined stickleback transcriptomic resource, and as such, provide a starting point for intra- and inter-specific genomic comparisons in sticklebacks.

Sequence divergence between nine- and three-spined sticklebacks

The nine-spined stickleback transcriptome characterized in this study allowed us to survey sequence divergence between two closely related species – nine- and three-spined sticklebacks. Because the two species diverged 13 Mya [22], we anticipated that the genetic differences would be considerable despite the highly ecological, phenotypic, and genetic similarities between the species [23,43]. The rate of sequence substitution is of central importance to understand mechanisms underlying molecular evolution. Rates of nonsynonymous and synonymous substitutions are good indicators of selective pressures at the sequence level of protein-coding genes [59,60]. Synonymous sites usually evolve neutrally and can provide insights on the background rate of sequence evolution [59,60], thus we used the K_s values of protein-coding genes to estimate neutral substitution rates in sticklebacks. The average substitution rate was estimated to be 7.1×10^{-9} per synonymous site per year between nine- and three-spined sticklebacks (Additional file 1: Figure S6) when calibrated to the divergence time of 13 Mya. This rate is faster than previously published genome-wide substitution rate estimates available across mammals (2.2×10^{-9} per synonymous site per year; [61]), but is nearer the substitution rate of teleosts (1.25×10^{-6} in cichlids; [46]) as the rates of molecular evolution in fish are known to be fast compared to other vertebrates [62]. Additionally, the unigenes we identified may be enriched with highly expressed genes that are more easily detected in transcriptomic sequencing, and thus the estimated substitution rate might be an underestimation because highly expressed protein coding genes usually

evolve slowly [63]. Nevertheless, this estimated substitution rate should be a useful yardstick for research in teleost molecular evolution in general, and particularly for those studies on stickleback phylogeny and molecular clock dating.

Identifying genes that show evidence of positive selection can help us in understanding whether closely related species occupying similar ecological niches share genetic attributes involved in adaptation. The K_a/K_s ratio (= 1: neutral evolution; > 1: positive selection; < 1: purifying selection) is often used for diagnosing the extent and direction of selection on sequence evolution [59,60]. Using three analyses based on nonsynonymous and synonymous substitutions, a total of 165 genes show indications of positive selection in one or both species of sticklebacks. These 165 genes have significantly smaller pairwise K_s (Wilcoxon–Mann–Whitney test, $P = 1.4 \times 10^{-7}$) but significantly larger pairwise K_a (Wilcoxon–Mann–Whitney test, $P = 5.7 \times 10^{-4}$) compared to the other analyzed genes (Additional file 1: Figure S7). Despite a broad range of GO annotations that these genes are involved with, we found that they showed enrichment in several functional categories. Such genes may be of particular interest for further studies aiming to investigate their detailed functions, as well as possible associations with ecological differences between stickleback species.

In addition to coding sequence changes, regulatory sequence changes may play an important role in repeated adaptive evolution of freshwater three-spined sticklebacks [3]. In general, UTRs, especially 3′-UTRs, are found to evolve neutrally among very closely related taxa [46]. However, we found that UTRs between nine- and three-spined sticklebacks are under stronger purifying selection as compared to synonymous sites, but under more relaxed selection as compared to coding regions (both synonymous and nonsynonymous sites). These findings suggest that some UTRs may be important in shaping stickleback evolution [3].

Gene gains and losses are important processes contributing to evolutionary innovation and differentiation [64,65], perhaps especially so in teleosts because of the teleost-specific whole genome duplication event [66]. The comparison between stickleback orthologs revealed that some genes are likely to have been lost in the three-spined stickleback, as they exist both in nine-spined sticklebacks and other model fish genomes. It is also possible that these genes are missing from the current three-spined stickleback genome assembly, or that the genes have evolved so rapidly that they no longer resemble the same gene in other fishes. Of the genes that might have been lost in three-spined sticklebacks, nine have associated GO terms related to binding (protein and iron), cell migration, and membrane component.

However, a more complete grasp of the number of genes differentially lost and retained between nine- and three-spined sticklebacks can only be answered with a complete nine-spined stickleback genome. Nevertheless, our results suggest that as in the case of other vertebrates [65,67], stickleback divergence is also accompanied with gene losses.

However, we are aware that our results largely depend on the initial dataset for which we can make comparisons between genes. Because we used a subset of all genes in the genome, we cannot capture the entire list of variation and genes that are evolving under positive selection. In fact, our dataset may further be biased towards slowly evolving genes under stronger purifying selection if we are capturing mainly highly expressed genes, and those with low K_s values. Nevertheless, our results should provide a useful first step towards unraveling the genetics underlying divergence between nine- and three-spined sticklebacks. Taken together, our analyses of substitution rates, positive selection and gene loss suggest that there are considerable genetic differences between these two ecologically and phenotypically similar species.

Genetic divergence between marine and freshwater nine-spined sticklebacks

Much research has been directed towards investigating genome-wide divergence between marine and freshwater three-spined sticklebacks and many genes associated with their divergence have been identified [3,68]. Genetic differentiation between marine and freshwater nine-spined sticklebacks also has been described in studies utilizing microsatellites [41] and restriction-site-associated DNA sequencing [42]. For example, Shikano et al. [41] found several functionally- and physiologically-important genes that had experienced divergent selection between different habitats, and Bruneaux et al. [42] showed that genomic regions enriched for genes having functions related to immunity, chemical stimulus response, lipid metabolism, and signaling pathways had experienced positive selection. However, in-depth genome-wide studies of genetic differentiation between marine and freshwater nine-spined sticklebacks have been lacking. Here, we probed the genome-wide genetic differentiation between marine and freshwater nine-spined sticklebacks to understand whether similar or different genetic changes underlying divergence between freshwater and marine populations exist in the two stickleback species. We found 25 genes with 'fixed' genotypes between marine and freshwater nine-spined sticklebacks (Additional file 3: Table S8), and these represent candidates for ecotypic differentiation in nine-spined sticklebacks. Interestingly, one of these genes, the *enolase 1a* (ENSGACG00000007396) gene has also been

found to be associated with the divergence of marine and freshwater three-spined sticklebacks [3]. *ATPases* are another group of interesting genes that have been associated with the marine and freshwater divergence in sticklebacks. We found that the *ATP5B* and *ATP6v1ba* genes have SNPs differentiating marine and freshwater nine-spined sticklebacks, and similar evidence is available from *ATP6V1Aa* [41] in nine-spined and *ATP6V0A1* and *ATP6V0E1* in three-spined sticklebacks [3]. Furthermore, a transferrin gene (ENSGACG00000013533) with a putative function in iron ion transport may be of particular interest for understanding adaptive population divergence of marine and freshwater nine-spined sticklebacks, since ion concentration is one of the notable environmental differences demarcating marine and freshwater habitats. Hence, *enolase 1a*, *ATP5B*, *ATP6v1ba* and transferrin provide promising candidates for further investigations focused on understanding the molecular mechanisms of differentiation and adaptation between marine and freshwater stickleback populations. Further studies screening more populations and individuals are needed to evaluate the robustness of these results, as well as to understand how often adaptive divergence between marine and freshwater populations of different stickleback taxa is occurring through evolution in the same or in different genes or genetic elements.

Conclusions

With the massively parallel pryrosequencing of nine-spined stickleback cDNA libraries, we identified over three thousand unique gene transcripts and hundreds of genic microsatellites. Using these transcripts, we calculated sequence substitution rates in coding regions, in UTRs, and across synonymous sites between nine- and three-spined sticklebacks. We identified over a hundred genes with molecular patterns of positive selection in one or both stickleback lineages and found several candidate genes that might be involved in differential adaptation between marine and freshwater nine-spined sticklebacks. Both the same and different genes were found to associate with marine and freshwater divergence across stickleback taxa. Apart from these specific findings, the study brings about significant amount of new resources (*viz.* gene sequences, microsatellites, and SNPs) to the reach of the research community interested in fish and stickleback genomics in particular.

Methods

This study did not involve human subjects, and our experimental protocol was approved by the ethics committee of National Animal Experiment Board, Finland (permission numbers: ESLH-STSTH223A and STH037A).

Fish sampling, RNA isolation, and cDNA library construction

We sampled two male and two female nine-spined sticklebacks from the Baltic Sea (Helsinki, Finland; 60°12′N, 25°11′E), and one male and one female from an isolated freshwater pond (Rytilampi, Finland; 66°23′N, 19°19′E). We chose to sequence the brain and liver transcriptomes to access a large number of diverse transcripts, as these are highly complex organs with complex transcriptomes. Total RNA was extracted from brain and liver tissues using TRIzol reagent (Invitrogen, Carlsbad, CA, USA) according to the manufacturer's protocol. We constructed four cDNA libraries (marine brain, marine liver, freshwater brain, and freshwater liver) with the Super-Script® Double-Stranded cDNA Synthesis Kit (Invitrogen, cat. no. 11917-010), according to the manufacturer's protocol. Equimolar amounts of the total RNA from each of the two males and two females from marine population were pooled for construction of the marine brain library, but only one male and one female were used for the marine liver library. Likewise, RNA from one male and one female were used for the freshwater brain and liver libraries.

Transcriptome sequencing and assembly

We barcoded the four cDNA libraries and sequenced them in a half plate of GS FLX (Roche 454) Standard Chemistry run by DNA Sequencing and Genomics Laboratory, Institute of Biotechnology, University of Helsinki at Helsinki, Finland. Sequences have been deposited in the NCBI Short Read Archive (Accession no. SRR846896, SRR846899, SRR846900, and SRR846901). We trimmed adaptors and low quality reads using custom Perl scripts. We then assembled the cleaned reads using v2.5.3 of the GS *De Novo* Assembler [69] into contigs representing four transcriptomic libraries, one brain and one liver library from each population. We obtained three additional transcriptomic libraries by pooling the contigs from the four cDNA libraries into a marine transcriptome, a freshwater transcriptome and the all-encompassing 'nine-spined stickleback transcriptome'. These transcriptomic libraries were assembled from reads using a minimum overlap length of 40 bp and a minimum overlap identity of 98%. Detailed information of the transcriptome assemblies are listed in Additional file 2: Table S1.

Gene annotation

Our annotations focused on the 'nine-spined stickleback transcriptome' that was assembled with all reads combined from the four cDNA libraries. We only included assembly contigs with a minimum length of 100 bp for further analyses and used two comprehensive methods to annotate the remaining contigs. We first assigned

their putative identities using BLASTX [70] searches against protein datasets of the three-spined stickleback reference (a freshwater individual [3]) from the Ensembl database [71] (release-68) with an E-value cutoff of 1×10^{-10}, and paired the contigs with their top BLAST hit. The resulting gene pairs are herein referred to as orthologs. Importantly, because of varying transcript lengths and alternative transcription, different nine-spined stickleback contigs can map to different regions or to alternative transcripts of the same three-spined stickleback gene. To identify genes that are possibly lost (or missing) from the three-spined stickleback genome, we used contigs without hits against three-spined stickleback proteins as queries in BLASTX searches against protein datasets of the other model fishes *Danio rerio*, *Gadus morhua*, *Oreochromis niloticus*, *Oryzias latipes*, *Takifugu rubripes*, and *Tetraodon nigroviridis* from the Ensembl database release-68 and *Xiphophorus maculatus* from the Ensembl database release-70. We then used those contigs with hits in other model fish as queries in BLASTN and BLAT [72] searches against the three-spined stickleback genome to validate that these putative genes are lost (or missing) from the three-spined stickleback genome.

We assigned putative functions for each selected nine-spined stickleback contig using version 2.5.0 of Blast2GO [73], which performs a BLASTX search against the non-redundant database from NCBI with default parameters. We obtained annotated accession numbers and Gene Ontology (GO) [74] numbers from NCBI QBLAST [70] based on an E-value of 1×10^{-10} and a high-scoring segment pair cut-off greater than 33. We conducted the annotation procedure with the following parameters: a pre-E-value-Hit-Filter of 10^{-6}, a pro-Similarity-Hit-Filter of 15, an annotation cut-off of 55, and a GO weight of 5. GO term enrichment test was conducted using GOSSIP [75].

To obtain putative protein-coding and amino acid sequences, we employed GeneWise2 [76] to deduce the open reading frame for each contig sequence using its corresponding best-match protein in the three-spined stickleback as a guide. The putative untranslated region (UTR) of each contig was obtained based on the results of the ORF prediction and further assessed by alignment with UTRs of their corresponding putative orthologs using MUSCLE [77] with default settings to avoid including assembly artifacts.

Substitution rate estimation

We aligned the amino acid sequences of each pair of orthologs from nine- and three-spined sticklebacks using MUSCLE [77] with default settings and manually inspected for possible alignment artifacts. We performed DNA sequence alignments from the resulting protein

alignments using a custom Perl script. The number of nonsynonymous substitutions per nonsynonymous site (K_a) and synonymous substitutions per synonymous site (K_s) between each orthologous pair was computed using a maximum-likelihood method [78] with the YN00 program implemented in PAML version 4.4 [79]. Only nine-spined stickleback contigs with $K_s \leq 0.5$ compared to their three-spined stickleback orthologs were selected for further analyses (e.g., GO annotation and SNP calling) and are referred to as unigenes. In addition, if different nine-spined stickleback contigs aligned to the same three-spined gene, nine-spined stickleback contig with smallest K_s to the three-spined gene was kept; in case two contigs aligned to the same three-spined gene with equal K_s values, we randomly kept one of them for further analysis. As mentioned previously, several nine-spined stickleback contigs or unigenes can correspond to different regions or transcripts of the same three-spined stickleback gene.

We estimated the overall substitution rate between the nine- and three-spined stickleback genomes based on the divergence between unigenes and their orthologs (coding region and UTRs at least 50 bp long) while considering a divergence time around 13 Mya [22]. Distances of coding regions and UTRs were calculated separately using Kimura's two parameter (K2P) model [80] in EMBOSS [81].

We performed the branch-site test [82,83] with the codeml program in PAML [79] to detect positive selection operating on sites in the nine- and three-spined stickleback lineages. For this test, we used the corresponding 1-to-1 ortholog in *O. latipes* (determined from Ensembl) as an outgroup. We were able to perform this test for 2,458 unigenes. We calculated the P values based on the Chi-square critical values of 3.84 (5%) as recommended in PAML [79]. Multiple test correction was performed using the qvalue package in R [84] with default settings to correct for the false discovery rate (FDR).

SNP calling

To determine single nucleotide polymorphisms (SNPs) among sampled nine-spined stickleback individuals, we mapped all of the cleaned reads from each of the four cDNA libraries separately to the nine-spined stickleback unigenes using BWA-SW in BWA 0.6.1 [85] with default settings. SNPs from each of the four mappings were called using samtools 0.1.18 [86] with mpileup -I to disable indel calling as insufficient flushing during 454 sequencing usually leads to indel events around homopolymers [87]. Only bases with a Phred quality score of at least 20 were considered for the SNP calling. Combined with the three-spined stickleback ortholog, SNPs were used for performing the McDonald-Kreitman (MK) test of neutral evolution [88] using libsequence [89]. The

MK test is used for evaluating the ratio of polymorphism (intraspecies differences) and divergence (interspecies differences) at nonsynonymous and synonymous sites. Under neutrality, the ratio of polymorphism and divergence at these site classes is equal. We calculated an unbiased estimator of the direction of selection (DoS) developed by Stoletzki & Eyre-Walker [90], which is a modification of the neutrality index (NI) [91] by calculating the difference between the proportion of divergent and polymorphic nonsynonymous substitutions. Whereas DoS is zero under neutrality, positive selection driving an excess of nonsynonymous divergence between species would render DoS positive, and purifying selection reflected by an excess of nonsynonymous polymorphisms within species would decrease DoS below zero. Statistical significance in the departure from neutrality for each gene was determined by the Chi-square test with Yates correction as implemented in libsequence [89].

Microsatellite identification

We used a microsatellite identification program – MISA [92] to identify microsatellite motifs in our nine-spined unigenes. We searched for all types of Simple Sequence Repeats (SSRs) from mononucleotide to hexanucleotides using the following parameters: at least 10 repeats for mono-, 6 repeats for di- and 5 repeats for tri-, tetra-, penta- and hexanucleotide for simple repeats. We identified both perfect (SSRs containing a single repeat motif) and compound (SSRs composed of two or more motifs separated by < 100 bp) SSRs.

Availability of supporting data

The data sets supporting the results of this article are available in the in the National Center for Biotechnology Information Short Read Archive repository (Accession no. SRR846896, SRR846899, SRR846900, and SRR846901), and included within the article and its additional files.

Additional files

Additional file 1: Figure S1. Read length distribution; **Figure S2.** Contig length distribution; **Figure S3.** Unigene GO annotation plotting with WEGO; **Figure S4.** Unigene length distribution; **Figure S5.** Chromosomal distribution of unigenes along the three-spined stickleback genome (grey) and genes with molecular patterns of adaptive evolution between nine- and three-spined sticklebacks (black); **Figure S6.** Gene-specific substitution rates per synonymous site per year between nine- and three-spined sticklebacks; **Figure S7.** Boxplot of pairwise K_a and K_s between nine- and three-spined sticklebacks of the 165 genes showing evidence for adaptive evolution and other genes.

Additional file 2: Table S1. Assembly summary of nine-spined stickleback transcriptomic libraries; **Table S2.** Genes putatively lost in three-spined sticklebacks.

Additional file 3: Table S3. Results from pairwise Ka/Ks calculation for detection of positive selection between nine- (9SS) and three-spined

(3SS) sticklebacks; **Table S4.** Results from a branch-site test for positive selection; **Table S5.** Direction of Selection (DoS) from the MK test; **Table S6.** GO term enrichment test of the genes with adaptive evolution between nine- and three-spined sticklebacks using GOSSIP; **Table S7.** Genes with SNPs between ecotypes; **Table S8.** Fixed genotype differences between ecotypes; **Table S9.** SSRs in unigenes of nine-spined stickleback.

Author details
[1]Ecological Genetics Research Unit, Department of Biosciences, University of Helsinki, Helsinki, Finland. [2]Max Planck Institute for Evolutionary Biology, Department of Evolutionary Ecology, Plön, Germany. [3]Institute for Evolution and Biodiversity, Evolutionary Bioinformatics, Westfalian Wilhelms University of Münster, Münster, Germany. [4]Division of Genetics and Physiology, Department of Biology, University of Turku, Turku, Finland.

References
1. Hardison RC: **Comparative genomics.** *PloS Biol* 2003, **1**:e58.
2. Stapley J, Reger J, Feulner PG, Smadja C, Galindo J, Ekblom R, Bennison C, Ball AD, Beckerman AP, Slate J: **Adaptation genomics: the next generation.** *Trends Ecol Evol* 2010, **25**:705–712.
3. Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, Swofford R, Pirun M, Zody MC, White S, *et al*: **The genomic basis of adaptive evolution in threespine sticklebacks.** *Nature* 2012, **484**:55–61.
4. Gibson G: **The synthesis and evolution of a supermodel.** *Science* 2005, **307**:1890–1891.
5. McKinnon JS, Rundle HD: **Speciation in nature: the threespine stickleback model systems.** *Trends Ecol Evol* 2002, **17**:480–488.
6. Bell MA, Foster SA: *The evolutionary biology of the threespine stickleback.* London: Oxford University Press; 1995.
7. Chan YF, Marks ME, Jones FC, Villarreal G, Shapiro MD, Brady SD, Southwick AM, Absher DM, Grimwood J, Schmutz J, *et al*: **Adaptive Evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 Enhancer.** *Science* 2010, **327**:302–305.
8. Miller CT, Beleza S, Pollen AA, Schluter D, Kittles RA, Shriver MD, Kingsley DM: **cis-regulatory changes in kit ligand expression and parallel evolution of pigmentation in sticklebacks and humans.** *Cell* 2007, **131**:1179–1189.
9. Colosimo PF, Hosemann KE, Balabhadra S, Villarreal G, Dickson M, Grimwood J, Schmutz J, Myers RM, Schluter D, Kingsley DM: **Widespread parallel evolution in sticklebacks by repeated fixation of ectodysplasin alleles.** *Science* 2005, **307**:1928–1933.
10. Peichel CL, Ross JA, Matson CK, Dickson M, Grimwood J, Schmutz J, Myers RM, Mori S, Schluter D, Kingsley DM: **The master sex-determination locus in threespine sticklebacks is on a nascent Y chromosome.** *Curr Biol* 2004, **14**:1416–1424.
11. Colosimo PF, Peichel CL, Nereng K, Blackman BK, Shapiro MD, Schluter D, Kingsley DM: **The genetic architecture of parallel armor plate reduction in threespine sticklebacks.** *PloS Biol* 2004, **2**:635–641.
12. Shapiro MD, Marks ME, Peichel CL, Blackman BK, Nereng KS, Jonsson B, Schluter D, Kingsley DM: **Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks.** *Nature* 2004, **428**:717–723.

13. Peichel CL, Nereng KS, Ohgi KA, Cole BLE, Colosimo PF, Buerkle CA, Schluter D, Kingsley DM: **The genetic architecture of divergence between threespine stickleback species.** *Nature* 2001, **414**:901–905.

14. Cresko WA, Amores A, Wilson C, Murphy J, Currey M, Phillips P, Bell MA, Kimmel CB, Postlethwait JH: **Parallel genetic basis for repeated evolution of armor loss in Alaskan threespine stickleback populations.** *Proc Natl Acad Sci USA* 2004, **101**:6050–6055.

15. Kimmel CB, Ullmann B, Walker C, Wilson C, Currey M, Phillips PC, Bell MA, Postlethwait JH, Cresko WA: **Evolution and development of facial bone morphology in threespine sticklebacks.** *Proc Natl Acad Sci USA* 2005, **102**:5791–5796.

16. Shapiro MD, Bell MA, Kingsley DM: **Parallel genetic origins of pelvic reduction in vertebrates.** *Proc Natl Acad Sci USA* 2006, **103**:13753–13758.

17. Shapiro MD, Summers BR, Balabhadra S, Aldenhoven JT, Miller AL, Cunningham CB, Bell MA, Kingsley DM: **The genetic architecture of skeletal convergence and sex determination in ninespine sticklebacks.** *Curr Biol* 2009, **19**:1156–1156.

18. Shikano T, Laine VN, Herczeg G, Vilkki J, Merilä J: **Genetic architecture of parallel pelvic reduction in ninespine sticklebacks.** *G3- Genes Genom Genet* 2013, **3**:1833–1842.

19. Laine VN, Shikano T, Herczeg G, Vilkki J, Merilä J: **Quantitative trait loci for growth and body size in the nine-spined stickleback *Pungitius pungitius* L.** *Mol Ecol.* in press.

20. Laine VN, Shikano T, Herczeg TG JV, Merilä J: **QTL analysis of behavioral traits in nine-spined sticklebacks (*Pungitius pungitius*).** *Behav Genet.* in press.

21. Merilä J: **Nine-spined stickleback (*Pungitius pungitius*): an emerging model for evolutionary biology research.** *Ann N Y Acad Sci* 2013, **1289**:18–35.

22. Bell MA, Stewart JD, Park PJ: **The world's oldest fossil threespine stickleback fish.** *Copeia* 2009, **2009**:256–265.

23. Wootton RJ: *The biology of the sticklebacks.* New York: Academic; 1976.

24. Shikano T, Shimada Y, Herczeg G, Merilä J: **History vs. habitat type: explaining the genetic structure of European nine-spined stickleback (*Pungitius pungitius*) populations.** *Mol Ecol* 2010, **19**:1147–1161.

25. Teacher AGF, Shikano T, Karjalainen ME, Merilä J: **Phylogeography and genetic structuring of European nine-spined sticklebacks (*Pungitius pungitius*)-mitochondrial DNA evidence.** *PLoS One* 2011, **6**:e19476.

26. Aldenhoven JT, Miller MA, Corneli PS, Shapiro MD: **Phylogeography of ninespine sticklebacks (*Pungitius pungitius*) in North America: glacial refugia and the origins of adaptive traits.** *Mol Ecol* 2010, **19**:4061–4076.

27. Herczeg G, Gonda A, Merilä J: **Evolution of gigantism in nine-spined sticklebacks.** *Evolution* 2009, **63**:3190–3200.

28. Herczeg G, Gonda A, Merilä J: **Morphological divergence of North-European nine-spined sticklebacks (*Pungitius pungitius*): signatures of parallel evolution.** *Biol J Linnean Soc* 2010, **101**:403–416.

29. Herczeg G, Gonda A, Merilä J: **Predation mediated population divergence in complex behaviour of nine-spined stickleback (*Pungitius pungitius*).** *J Evol Biol* 2009, **22**:544–552.

30. Herczeg G, Gonda A, Merilä J: **The social cost of shoaling covaries with predation risk in nine-spined stickleback, *Pungitius pungitius*, populations.** *Anim Behav* 2009, **77**:575–580.

31. Gonda A, Herczeg G, Merilä J: **Habitat-dependent and -independent plastic responses to social environment in the nine-spined stickleback (*Pungitius pungitius*) brain.** *Proc R Soc B-Biol Sci* 2009, **276**:2085–2092.

32. Gonda A, Herczeg G, Merilä J: **Adaptive brain size divergence in nine-spined sticklebacks (*Pungitius pungitius*)?** *J Evol Biol* 2009, **22**:1721–1726.

33. Gonda A, Välimäki K, Herczeg G, Merilä J: **Brain development and predation: plastic responses depend on evolutionary history.** *Biol Lett* 2012, **8**:249–252.

34. Trokovic N, Herczeg G, Ab Ghani NI, Shikano T, Merilä J: **High levels of fluctuating asymmetry in isolated stickleback populations.** *BMC Evol Biol* 2012, **12**:115.

35. Trokovic N, Herczeg G, Scott McCairns R, Izza Ab Ghani N, Merilä J: **Intraspecific divergence in the lateral line system in the nine-spined stickleback (*Pungitius pungitius*).** *J Evol Biol* 2011, **24**:1546–1558.

36. Gonda A, Herczeg G, Merilä J: **Population variation in brain size of nine-spined sticklebacks (Pungitius pungitius)–local adaptation or environmentally induced variation?** *BMC Evol Biol* 2011, **11**:75.

37. Shimada Y, Shikano T, Kuparinen A, Gonda A, Leinonen T, Merilä J: **Quantitative genetics of body size and timing of maturation in two nine-spined stickleback (*Pungitius pungitius*) populations.** *PLoS One* 2011, **6**:e28859.

38. Waser W, Sahoo TP, Herczeg G, Merilä J, Nikinmaa M: **Physiological differentiation among nine-spined stickleback populations: effects of copper exposure.** *Aquat Toxicol* 2010, **98**:188–195.

39. Feulner PG, Chain FJ, Panchal M, Eizaguirre C, Kalbe M, Lenz TL, Mundry M, Samonte IE, Stoll M, Milinski M, *et al*: **Genome-wide patterns of standing genetic variation in a marine population of three-spined sticklebacks.** *Mol Ecol* 2013, **22**:635–649.

40. Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA: **Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags.** *PLoS Genet* 2010, **6**:e1000862.

41. Shikano T, Ramadevi J, Merilä J: **Identification of local- and habitat-dependent selection: scanning functionally important genes in nine-spined sticklebacks (*Pungitius pungitius*).** *Mol Biol Evol* 2010, **27**:2775–2789.

42. Bruneaux M, Johnston SE, Herczeg G, Merilä J, Primmer CR, Vasemagi A: **Molecular evolutionary and population genomic analysis of the nine-spined stickleback using a modified restriction-site-associated DNA tag approach.** *Mol Ecol* 2013, **22**:565–582.

43. Shikano T, Ramadevi J, Shimada Y, Merilä J: **Utility of sequenced genomes for microsatellite marker development in non-model organisms: a case study of functionally important genes in nine-spined sticklebacks (*Pungitius pungitius*).** *BMC Genomics* 2010, **11**:334.

44. Andres JA, Larson EL, Bogdanowicz SM, Harrison RG: **Patterns of transcriptome divergence in the male accessory gland of two closely related species of field crickets.** *Genetics* 2013, **193**:501–513.

45. Barreto FS, Moy GW, Burton RS: **Interpopulation patterns of divergence and selection across the transcriptome of the copepod Tigriopus californicus.** *Mol Ecol* 2011, **20**:560–572.

46. Elmer KR, Fan S, Gunter HM, Jones JC, Boekhoff S, Kuraku S, Meyer A: **Rapid evolution and selection inferred from the transcriptomes of sympatric crater lake cichlid fishes.** *Mol Ecol* 2010, **19**(Suppl 1):197–211.

47. Meyer E, Aglyamova GV, Wang S, Buchanan-Carter J, Abrego D, Colbourne JK, Willis BL, Matz MV: **Sequencing and de novo analysis of a coral larval transcriptome using 454 GSFlx.** *BMC Genomics* 2009, **10**:219.

48. Parchman TL, Geist KS, Grahnen JA, Benkman CW, Buerkle CA: **Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery.** *BMC Genomics* 2010, **11**:180.

49. Vera JC, Wheat CW, Fescemyer HW, Frilander MJ, Crawford DL, Hanski I, Marden JH: **Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing.** *Mol Ecol* 2008, **17**:1636–1647.

50. Axelsson E, Ratnakumar A, Arendt ML, Maqbool K, Webster MT, Perloski M, Liberg O, Arnemo JM, Hedhammar A, Lindblad-Toh K: **The genomic signature of dog domestication reveals adaptation to a starch-rich diet.** *Nature* 2013, **495**:360–364.

51. Zhao S, Zheng P, Dong S, Zhan X, Wu Q, Guo X, Hu Y, He W, Zhang S, Fan W, *et al*: **Whole-genome sequencing of giant pandas provides insights into demographic history and local adaptation.** *Nat Genet* 2013, **45**:67–71.

52. Guo B, Zou M, Gan X, He S: **Genome size evolution in pufferfish: an insight from BAC clone-based *Diodon holocanthus* genome sequencing.** *BMC Genomics* 2010, **11**:396.

53. Guo B, Gan X, He S: ***Hox* genes of the Japanese eel *Anguilla japonica* and *Hox* cluster evolution in teleosts.** *J Exp Zool Part B* 2010, **314**:135–147.

54. Zou M, Guo B, He S: **The roles and evolutionary patterns of intronless genes in deuterostomes.** *Comp Funct Genomics* 2011, **2011**:680673.

55. Guo B, Zou M, Wagner A: **Pervasive indels and their evolutionary dynamics after the fish-specific genome duplication.** *Mol Biol Evol* 2012, **29**:3005–3022.

56. Guo B, Tong C, He S: ***Sox* genes evolution in closely related young tetraploid cyprinid fishes and their diploid relative.** *Gene* 2009, **439**:102–112.

57. Kulathinal RJ, Stevison LS, Noor MA: **The genomics of speciation in *Drosophila*: diversity, divergence, and introgression estimated using low-coverage genome sequencing.** *Plos Genet* 2009, **5**:e1000550.

58. Noor MA, Feder JL: **Speciation genetics: evolving approaches.** *Nat Rev Genet* 2006, **7**:851–861.

59. Yang Z, Bielawski JP: **Statistical methods for detecting molecular adaptation.** *Trends Ecol Evol* 2000, **15**:496–503.

60. Hurst LD: **The *Ka/Ks* ratio: diagnosing the form of sequence evolution.** *Trends Genet* 2002, **18**:486.

61. Kumar S, Subramanian S: **Mutation rates in mammalian genomes.** *Proc Natl Acad Sci USA* 2002, **99**:803–808.

62. Ravi V, Venkatesh B: **Rapidly evolving fish genomes and teleost diversity.** *Curr Opin Genet Dev* 2008, **18**:544–550.

63. Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH: **Why highly expressed proteins evolve slowly.** *Proc Natl Acad Sci USA* 2005, **102**:14338–14343.

64. Ding Y, Zhou Q, Wang W: **Origins of new genes and evolution of their novel functions.** *Annu Rev Ecol Evol S* 2012, **43**:345–363.

65. Hahn MW, Demuth JP, Han SG: **Accelerated rate of gene gain and loss in primates.** *Genetics* 2007, **177**:1941–1949.

66. Guo B, Wagner A, He S: **Duplicated gene evolution following whole-genome duplication in teleost Fish.** In *Gene Duplication.* Edited by Friedberg F. Rijeka, Croatia: InTech; 2011:27–36.

67. Blomme T, Vandepoele K, De Bodt S, Simillion C, Maere S, Van de Peer Y: **The gain and loss of genes during 600 million years of vertebrate evolution.** *Genome Biol* 2006, **7**:R43.

68. Jones FC, Chan YF, Schmutz J, Grimwood J, Brady SD, Southwick AM, Absher DM, Myers RM, Reimchen TE, Deagle BE, *et al*: **A genome-wide SNP genotyping array reveals patterns of global and repeated species-pair divergence in sticklebacks.** *Curr Biol* 2012, **22**:83–90.

69. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen ZT, *et al*: **Genome sequencing in microfabricated high-density picolitre reactors.** *Nature* 2005, **437**:376–380.

70. Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25**:3389–3402.

71. Hubbard T, Andrews D, Caccamo M, Cameron G, Chen Y, Clamp M, Clarke L, Coates G, Cox T, Cunningham F, *et al*: **Ensembl 2005.** *Nucleic Acids Res* 2005, **33**:D447–D453.

72. Kent WJ: **BLAT–the BLAST-like alignment tool.** *Genome Res* 2002, **12**:656–664.

73. Gotz S, Garcia-Gomez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, Robles M, Talon M, Dopazo J, Conesa A: **High-throughput functional annotation and data mining with the Blast2GO suite.** *Nucleic Acids Res* 2008, **36**:3420–3435.

74. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al*: **Gene Ontology: tool for the unification of biology.** *Nat Genet* 2000, **25**:25–29.

75. Bluthgen N, Brand K, Cajavec B, Swat M, Herzel H, Beule D: **Biological profiling of gene groups utilizing Gene Ontology.** *Genome Inform* 2005, **16**:106–115.

76. Birney E, Clamp M, Durbin R: **GeneWise and genomewise.** *Genome Res* 2004, **14**:988–995.

77. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic Acids Res* 2004, **32**:1792–1797.

78. Yang ZH, Nielsen R: **Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models.** *Mol Biol Evol* 2000, **17**:32–43.

79. Yang ZH: **PAML 4: Phylogenetic analysis by maximum likelihood.** *Mol Biol Evol* 2007, **24**:1586–1591.

80. Kimura M: **A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide-sequences.** *J Mol Evol* 1980, **16**:111–120.

81. Rice P, Longden I, Bleasby A: **EMBOSS: the european molecular biology open software suite.** *Trends Genet* 2000, **16**:276–277.

82. Yang ZH, Nielsen R: **Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages.** *Mol Biol Evol* 2002, **19**:908–917.

83. Zhang JZ, Nielsen R, Yang ZH: **Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level.** *Mol Biol Evol* 2005, **22**:2472–2479.

84. Storey JD, Tibshirani R: **Statistical significance for genomewide studies.** *P Natl Acad Sci USA* 2003, **100**:9440–9445.

85. Li H, Durbin R: **Fast and accurate long-read alignment with burrows-wheeler transform.** *Bioinformatics* 2010, **26**:589–595.

86. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Proc GPD: **The sequence alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25**:2078–2079.

87. Lynch M, Sung W, Morris K, Coffey N, Landry CR, Dopman EB, Dickinson WJ, Okamoto K, Kulkarni S, Hartl DL, *et al*: **A genome-wide view of the spectrum of spontaneous mutations in yeast.** *Proc Natl Acad Sci USA* 2008, **105**:9272–9277.

88. Mcdonald JH, Kreitman M: **Adaptive protein evolution at the *Adh* locus in *Drosophila*.** *Nature* 1991, **351**:652–654.

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

90. Stoletzki N, Eyre-Walker A: **Estimation of the neutrality index.** *Mol Biol Evol* 2011, **28**:63–70.

91. Rand DM, Kann LM: **Excess amino acid polymorphism in mitochondrial DNA: contrasts among genes from drosophila, mice, and humans.** *Mol Biol Evol* 1996, **13**:735–748.

92. Thiel T, Michalek W, Varshney RK, Graner A: **Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.).** *Theor Appl Genet* 2003, **106**:411–422.

93. Ye J, Fang L, Zheng H, Zhang Y, Chen J, Zhang Z, Wang J, Li S, Li R, Bolund L, *et al*: **WEGO: a web tool for plotting GO annotations.** *Nucleic Acids Res* 2006, **34**:W293–297.