New gains with a slimmer genome – An automated approach to improve reference genome assemblies

Mikko Kivikoski[1,†,*], Pasi Rastas[2,†], Ari Löytynoja[2,*], and Juha Merilä[1]

[1]Organismal and Evolutionary Biology Research Programme, Faculty of Biological and Environmental Sciences, FI-00014 University of Helsinki, Finland
[2]Institute of Biotechnology, FI-00014 University of Helsinki, Finland
[†]Contributed equally to the work
[*]Corresponding author

Abstract

The utility of genome-wide sequencing data in biological research depends heavily on the quality of the reference genome. Although the reference genomes have improved, it is evident that the assemblies could still be refined, especially in non-model study organisms. Here, we describe an integrative approach to improve contiguity and haploidy of a reference genome assembly. With two novel features of Lep-Anchor software and a combination of dense linkage maps, overlap detection and bridging long reads we generated an improved assembly of the nine-spined stickleback (*Pungitius pungitius*) reference genome. We were able to remove a significant number of haplotypic contigs, detect more genetic variation and improve the contiguity of the genome, especially that of X chromosome. However, improved scaffolding cannot correct for mosaicism of erroneously assembled contigs, demonstrated by a de novo assembly of a 1.7 Mbp inversion. Qualitatively similar gains were obtained with the genome of three-spined stickleback (*Gasterosteus aculeatus*).

Introduction

Great deal of present-day research in biology is based on genomic data that are processed and analyzed in the context of a linear reference genome. Typical examples of this are whole-genome sequencing studies where sequencing reads are mapped to the reference genome and the characteristics of interest are derived from local dissimilarities and statistics based on the alignments (1, 2). Reliability of those characteristics and the conclusions drawn from them depend not only on the quality of the sequencing data but also on the quality of the reference genome. Assembling and evaluating the quality of reference genomes is not easy (3, 4, 5, 6). The profound problem is that the physical connectivity is lost during sequencing and recovering that in the assembly stage is notoriously difficult. To this end, high-quality linkage maps are valuable and allow inferring the physical order and orientation of the assembled contigs (7, 8, 9).

Although a linear reference genome is ill-suited for describing many structural variations, most genome analysis methods assume the reference genome to contain each genomic region only once. The continuous development of the human reference genome (10, 11) has shown that creating a linear haploid reference genome for a diploid species is a non-trivial task. Reaching this ideal can be especially challenging in organisms where the genetic variation cannot be reduced in controlled inbreeding designs, and most reference genomes are likely based on reference individuals carrying long alternative haplotypes (12, 7, 13). Presence of homologous haplotypes, that is, differing copies of the same genomic region inherited from the two parents, is against the assumptions of the linear reference genome and affects for instance the read mapping. If reads from distinct haplotypes map to different copies of the same region, single nucleotide variants (SNPs) separating the haplotypes cannot be detected and variation is underestimated. This affects various statistics in population genomics, and may lead to wrong conclusions in many different contexts, including estimation of substitution rate (14), inbreeding (15) or population history (16).

Lep-Anchor software (9) can improve assembly and scaffolding of even high-quality reference genomes with joint use of linkage map based genome anchoring and long-read sequencing data. Its performance was tested and demonstrated in the original publication, and here we have a closer look on the actual changes and asses their impact on typical genome analyses. Starting from an existing high-quality contig assembly, original PacBio reads and ultra-dense linkage maps for the nine-spined stickleback (*Pungitius pungitius*), we were able to generate a significantly improved reference genome (ver. 7) using largely automated methods. When evaluating the differences to the published version of the reference genome (ver. 6) (17) we detected haplotypes in three contexts. First, some haplotypes were originally assembled as separate contigs leading to false duplication of a region in the assembly. Second, haplotypes were assembled to the ends of subsequent contigs and occurred as duplicates on both sides of a contig gap. Third, haplotypic regions, exemplified by an inversion in LG19, were assembled as mosaics of the two haplotypes. Using the novel features of Lep-Anchor, we could automatically remove a large proportion of the first two types of haplotypes while the correction of haplotypes of the last category was possible but demanded manual effort. Recognition and removal of haplotypes shortens the nine-spined stickleback reference genome and increases heterozygosity of the reference individual while the contig re-scaffolding enabled the identification of the centromere in all linkage groups. To demonstrate that this approach works for contig assemblies in general, we reassembled the latest published reference genome of the three-spined stickleback (*Gasterosteus aculeatus*) (18) using one new linkage map and publicly available 10X Genomics linked read sequencing data (19).

Results

We used Lep-Anchor software and information from linkage map anchoring, pairwise contig alignment and long-read bridging to reassemble the nine-spined stickleback genome. Linkage map anchoring allowed assigning 274 previously unassigned contigs to the linkage groups (LGs) and pairwise contig alignments revealed 10% of the previous assembly as haplotypes (Fig. 1a). Of the 843 contigs in linkage groups, Lep-Anchor could assess 763 to be scaffolded in correct orientation. Removal of haplotypes and linking of adjacent contigs reduced the number of contig gaps and increased the N50 contig length with more than two-fold (Table 1). With a more accurate representation of the haploid genome, the total length of the reference decreased with 55 Mbp (Table 1). It is noticeable that, with the exception of one linkage map produced here, these improvements were gained with a more efficient use of data generated for the original assembly. We also assembled and incorporated the native mitochondrial genome while additional data from related homozygous individuals were used to reassemble a large inversion in LG19.

The improved assembly brings noticeable gains, and we could now successfully map the centromere associated repeat and unambiguously identify the centromere positions in all linkage groups (previously missing from LG1 and LG16, and incoherent in LG10 and LG14, Fig. 1b; see reference (17)). Removal of haplotypes and other changes in the genome assembly affects read mapping and single nucleotide variant calling. More even read depth and the anticipated mean depth indicate that the reference has become more haploid and contains fewer haplotype copies (Fig. 1c). In comparison to the ver. 6, the heterozygosity of the reference individual increased by 14% (Table 2). As expected, most (78%) of the newly identified SNPs were in regions where haplotype variants were removed from the reference, allowing reads from variant alleles to map to the same copy of the genomic region (Fig. 2, Table 2). New SNPs in other regions were a minority and their allelic depth deviated from the expected value (Supplementary Fig. S1).

Content of the LG12 changed considerably from ver. 6 to ver. 7 as one of the homologous copies in X and Y chromosomes were removed (Fig. 1a). As a result, the sex determining part of LG12 is close to a haploid representation and few regions show zero read depth (Fig. 3a). This also increases the heterozygosity of the male reference individual (Fig. 3a), the newly identified SNPs arising from differences between X and Y chromosomes, while no increase is observed in females (Fig. 3a). Although homologous sequences are represented only once, the sex chromosome is still a mosaic of X and Y chromosomes and females show both homozygous variant and reference alleles (Fig. 3a). An HMM analysis confirmed the mosaicism and indicated the sex chromosome assembly to be 57% of X chromosome (Supplementary Fig. S2). Despite the mosaicism, the reassembly improved the synteny of LG12 with the three-spined stickleback counterparts (Fig. 3b).

Scaffolding with Lep-Anchor had a minor impact on the variant allele frequencies in the LG19 inversion region (Fig. 4a). The reason for this is that the original contigs were mosaics of the two alleles and an improved ordering of contigs does not correct for their internal errors. The newly assembled contigs and the scaffolded alleles for the LG19 block revealed that there, indeed, are two segregating inversion haplotypes in the study population, and that the reference individual (see methods) is heterozygous (Fig. 4a). As expected, the variant allele frequencies across the newly assembled haplotypes are either zero or twice as high as with the original mosaic assembly for individuals homozygous for the two alleles (Fig. 4a). Although the mosaicism had a large impact on variant allele frequencies, its effect on SNP frequency was small. There are more SNPs according to ver. 7 but most of them are found due to haplotype removal and few of them are in the inversion region. With the HMM and data from the four females, we found four observable large regions

that indicate fine-scale mosaic of two diverged haplotypes (Fig. 4b). All these four regions were identified in both genome versions which suggests that the corresponding contigs are erroneously assembled as in LG19.

In the reassembly of the three-spined stickleback genome, we identified 1,831 haplotype contigs, most of them unassigned in the previous assembly. On the other hand, we were able to add 176 previously unassigned contigs to the linkage groups. The ungapped length of the 21 linkage groups, representing the 21 chromosomes, decreased from 426 Mbp to 423 Mbp and the ungapped length of the unassigned contigs decreased from 21 Mbp to 13 Mbp. N50 of the original and our new genome are 83,717 and 87,370 bp, respectively (Table 3). With the new reference we found 0.62% more autosomal SNPs in a sample from Paxton Lake, Canada, than were found using the original assembly. As in the nine-spined stickleback, most of the newly identified SNPs (52%) were in regions where haplotype variant was removed from the reference. Whereas the median sequencing depth of the sample was 15x for both genome versions, the depth for the identified haplotype regions was 9x and 15x in the published and new assembly, respectively, indicating successful haplotype removal.

Discussion

Reconstructing a linear reference genome is a challenging, yet an instrumental task. Interpretation of genomic data is often made with the assumption that the reference genome is a complete haploid representation of the actual genome. The errors in the genome directly affect the conclusions drawn, and for instance, missing SNPs influence the site frequency spectrum that is essential in demographic analyses (20). More directly, presence of haplotype copies in a reference genome can make a highly diverged region seem exceptionally conserved and can thus seriously mislead variation-based functional analyses. Given the severe consequences of the errors, efforts to improve reference genomes are needed, and here we have described an approach to make reference genomes more haploid and more contiguous using the Lep-Anchor software (9).

Faced with the dilemma of correctly separating duplicated genome regions while simultaneously collapsing and merging haplotypic differences into a haploid sequence, all assembly programs are poised to make errors. The magnitude of these errors depends on the heterozygosity of the reference individual and on the type input data, long reads spanning more distant sites and thus capable of creating longer haplotype blocks, while the direction of the bias to either too long or too short genome depends on the algorithm. While the three-spined stickleback genome is based on relatively old data and is established over years of refinement, the nine-spined stickleback genome is an example of a modern reference genome built using the best practices. We demonstrated our method's potential by showing how the latter, an already very high quality reference genome, could be greatly improved by more efficient use of the original sequencing and mapping data (Fig.1, Table 1). Improvements were based on linking, reassembly and improved scaffolding of the contigs with joint use of linkage map anchoring and long read sequencing data, as well as characterization and removal of alternative haplotypes. The improvements on the three-spined stickleback genome were more modest but we could still both add new contigs into the linkage groups and remove haplotype copies (Table 3, Supplementary Fig. S3), resulting in an 0.62% increase in number of segregating sites in a sample from the Paxton Lake benthic population (Table 2). While the three-spined stickleback analyses demonstrate that Lep-Anchor can improve even highly polished assemblies, they also illustrate how various data types, for example contigs from the 10X Genomics platform, can be incorporated in genome refinement.

In the nine-spined stickleback, most of the removed haplotypes were among the unassigned contigs and only one contig was moved between two linkage groups (Fig. 1a), underlining the high quality of the original

scaffold. Although we were not able to place all contigs in the linkage groups, we were able to divide them in putative classes based on the read depth and their repeat content, those with high repeat content (either centromere or other) forming the largest groups of unassigned contigs. Although repetitive regions are difficult to assemble and scaffold using the type of data available, we were able to improve the centromeric regions (Fig. 1b) and our approach can be useful for repetitive regions more generally. Some unassigned contigs had low or even zero read depth (Fig. 1a), but as we did not detect any obvious contamination when aligning them to the NCBI database, those were retained in the reference genome.

Removal of haplotypes lead to identification of ca. 14% more autosomal SNPs in the nine-spined stickleback reference individual (Table 2). Finding more SNPs *per se* is not evidence for better assembly, but together with more uniform sequencing depth (Fig. 1c) and strong evidence of successfully identified haplotypes (Fig. 2), it shows that genetic variability can be underestimated if the reference genome contains haplotypes. The minority of newly identified SNPs that were not within haplotype regions (22% of the novel SNPs) may have emerged because of short similarities between contigs that were not classified as haplotypes. They may also be related to changes in mapping of the read pairs in regions where haplotype copies have been removed or contig orientation or order has changed. Nonetheless, the evidence for some of those SNPs is questionable as their allelic depth deviates from the expected (i.e. 0.5; Supplementary Fig. S1), and one may want to filter them from downstream analyses.

Nine-spined stickleback LG12 was formed by fusion of chromosomal segments that correspond to chromosomes 7 and 12 of the three-spined stickleback (21) (Fig. 3b). This rearrangement has occurred after the split of the three-spined and the nine-spined sticklebacks 17 million years ago (22) but the exact timing is unclear (21). While 15 Mbp in one end of LG12 behaves like an autosomal chromosome, the 17 Mbp (25 Mbp in ver. 6) in the other end forms the sex-determination region and behaves like a sex chromosome (Fig. 3a). While parts of the sex-determination region seem very similar, other parts have differentiated significantly, and assembling complete X and Y chromosomes based on a single male reference individual is extremely challenging. Although our HMM analysis indicated that the LG12 assembled here is only 57% of X (Supplementary Fig. S2), we are confident that the sequence content of the current version is close to haploid presentation of X and the error is mainly in the SNP polarization. This is supported by the improved synteny with the three-spined stickleback genome (Fig. 3b) but especially by the more uniform read depth and more constant nucleotide diversity across the whole LG12 in females (Fig. 3a). The original sequencing data for the nine-spined stickleback reference are slightly outdated by modern standards, and we did not attempt to scaffold both X and Y copies of LG12. Fully separating the two should be relatively straightforward by obtaining long-read or linked-read data for both sexes with the latest sequencing technology.

Without genotype phasing, a haploid reference genome is a mosaic of maternal and paternal haplotypes and the reference alleles are drawn randomly. If parental haplotypes are clearly different, they are assembled as separate copies and appear as duplicates in the contig assembly; if the differences are punctuated by local similarities, the haploid consensus may alternate between the two parental haplotypes. It is evident that if the underlying contigs are erroneously assembled, their re-ordering cannot make the reference perfect. In the nine-spined stickleback, the inversion in LG19 and the sex chromosome region LG12 demonstrate how diverged haplotypes complicate the assembly of a haploid reference genome. On the other hand, the characterization of the inversion haplotypes provides an example of how TrioBinning (23) can be utilized without a trio and long-read sequencing data combined with population level whole-genome sequencing data allow assembling the segregating haplotypes. We acknowledge that our HMM for identifying regions of diverged haplotypes provides only indicative results (Fig. 4b) but it does suggest that haplotypes can be

fairly common in the nine-spined stickleback which is in line with findings regarding other fish (7) and humans (24). We also anticipate that highly concentrated alternation between two homozygous genotypes is a usable statistic for exploration and more sophisticated detection methods based on that could be devised. Identification of such regions requires the studied individual to be heterozygous and therefore all regions were not supported by all individuals. Having a single continuous haplotype, such as the inversion in LG19, in the reference genome correctly phases the alternative alleles (Fig. 4a) and allows studying the differences between the haplotypes. However, representation of potential structural differences is difficult and it is evident that methodological work to incorporate multiple haplotypes in a reference genome, e.g. using variation graph data structures, is urgently needed (25).

Haploid reference genomes based on a single individual, such as the one here, represent only one version of the species' genome which may cause reference bias and thus affect various downstream analyses and the conclusions drawn from them (25, 26). Although a linear reference does not represent the full species diversity, they are widely used and provide a starting point for analysis of genomic variation between individuals and populations. In the future, pan-genome representations and graph-based algorithms will likely change the way reference genomes are represented and analyzed (25, 11). Since linear genomes are still widely used, their improvements are relevant (11) and our work demonstrates that significant enhancements can be obtained with efficient use of the existing data. Moreover, characterization of haplotypes is instrumental in more inclusive genome representations, increasing the relevance of our approach.

Materials and methods

Nine-spined stickleback assembly

The starting point for this reference was the genomic DNA sequence data from Varadharajan et al. (17) In short, sequence data on one male individual from Pyöreälampi pond (66°15'N, 29°26'E; Finland) were obtained with PacBio and Illumina sequencing (PacBio RS II, 63 and 23 SMRT cells with P6v1/C4 and P6v2/C4 chemistry, respectively; Illumina HiSeq 2000 rapid run 2 X 250 nt; see ref. (17) for details). To reassemble the genome, the contig assembly reported by Varadharajan et al. was used with two refinements: a large inversion in LG19 was characterized and the region was reassembled (see below), and the mitochondrial genome was reassembled (see below). The resulting contig-assembly was anchored using Lep-Anchor (9) following the standard pipeline (https://sourceforge.net/p/lep-anchor/wiki/Home) with default parameters (exception: minQuality=1 for Map2Bed to assign more contigs into chromosomes). For this, we (1) utilized three original linkage maps (17) and reconstructed one new ultra-high density linkage map concordant with the existing maps (see below); (2) generated contig-contig alignments by running the two first steps of HaploMerger2 (27); and (3) incorporated the raw PacBio reads by aligning them to the contig assembly with minimap2 (28).

The mitochondrial genome was misassembled in the ver. 6 contig assembly and ver. 6 genome included an earlier published *P. pungitius* mitogenome (29). Here, the mitochondrial genome was assembled from the short-read Illumina data of the reference individual. Reads that mapped to the published mitochondrial genome (29) were extracted, converted to fastq format and assembled with the program MEGAHIT (ver. 1.2.9) (30) using the default parameters. The only assembled contig longer than 1000 base pairs (bp) was 16720 bp long and was confirmed as the mitochondrial genome by aligning it to published stickleback mitochondrial genomes (31, 32, 29).

Nine-spined stickleback linkage map

The new linkage map was based on crosses of wild-caught marine nine-spined sticklebacks from Helsinki, Finland 60°13'N, 25°11'E). Adult nine-spined sticklebacks were collected randomly from an outbred population in April 2016 and were verified to be largely unrelated (average relatedness = 0.17). 99 F_1-generation families were generated at the University of Helsinki fish facility through artificial fertilizations using methods described in Rastas et al. (33). Half-sib families were formed by mating one female to two different males. Once the eggs had hatched and larvae started feeding, the families were thinned to 25 offspring per family and moved to two large aquaria so that about half of each family were placed in each unit. The larvae were mass-reared in these aquaria, and their family identity was later identified from the genotype data.

The parental fish were whole-genome sequenced (WGS) with an Illumina Hiseq 4000 and other Hiseq platforms (BGI Hong Kong) at 5–10x sequencing coverage. The offspring were sequenced/genotyped by Diversity Arrays Technology, Pty Ltd, Australia, using their DarTseq technology. The fastq files (from DArTseq and WGS) were mapped to the contig assembly using BWA-MEM (34) and SAMtools (ver. 1.9) (35). The genotype likelihoods were called using Lep-MAP3 (36) pipeline. The linkage mapping and the pedigree construction were conducted using Lep-MAP3. Since the offspring were mass-reared, we assigned the offspring to parents with the aid of genotype data using the IBD module in Lep-MAP3 and by calculating the Mendelian genotypic error rate for each combination of parents and offspring. Including all combinations with error rate $< 10\%$ (average error rate 5.5%) the dataset contained 938 offspring in 87 families and 130 parents. The offspring sex was obtained by comparing the sequencing coverage on X and Y specific regions using SAMtools depth (see Supplementary Fig. S4). The parental genotypes were called using ParentCall2 module utilizing half-sib families (halfsibs=1) and calling also sex markers (XLimit=2). After running Filtering2 on default parameters, 50568 markers were retained and SeparateChromosomes2 module with LOD score limit 40 (lodLimit) clustered these markers into 21 linkage groups (LGs), matching the chromosome number of the species (33, 17). JoinSingles2All with lodLimit=20 and lodDifference=4 assigned 22753 markers to the LGs and each LG was then ordered with OrderMarkers2, producing the first maps. After genome anchoring with the novel maps and maps from Varadharajan et al. (17), OrderMarkers2 was run one more time, separately for each LG, with physically ordered markers (parameter evaluateOrder). This yielded the final map from which the crossover locations were acquired.

Contig classification via sequencing depth

In ver. 7, 1644 of the total 2487 contigs were not assigned to any of the 21 linkage groups (Table 1). To classify these contigs, we analysed the depth (coverage) of the Illumina and PacBio sequencing data for the reference individual and for a pool of four female individuals from the Pyöreälampi pond (Illumina only, see 'Characterization and assembly of the LG19 inversion' for the sequencing details). The Illumina and PacBio data (subreads) were mapped to the reference genomes (contig assembly, ver. 6 and ver. 7) using BWA-MEM and minimap2 (28), respectively, and SAMtools (ver. 1.9). The raw sequencing depth was computed with SAMtools depth.

The coverage analysis was carried out using Lep-Anchor's novel modules CoverageAnalyser and CoverageHMM. The documentation and the source code of these modules can be found at https://sourceforge.net/projects/lep-anchor/. Using CoverageAnalyser and a simple mixture model, sequencing depth histogram was classified to (about) zero, half, normal or high: Half and normal depths were modelled using two normal distributions and the zero and high depth as a zeta distribution (coverage + 1 $\sim$ Zeta, the same distribution was used for both,

zero and high). Then CoverageHMM and a four-state hidden Markov model (HMM) were used to classify each genomic position to four states: zero, half, normal or high. The emission probabilities of the HMM were taken from the mixture model (CoverageAnalyser) and maximum likelihood transition probabilities along the physical (contig) coordinates were learned using the Baum-Welch expectation-maximization algorithm (37).

Repetitive regions in the assembly were identified with RepeatMasker software (ver. open-4.0.5; Smit *et al.* 2013–2015 http://www.repeatmasker.org) by using the species specific repeat libraries by Varadharajan *et al.* (38). Contigs with repeat content greater than 20% were classified as repetitive contigs (Fig. 1a). Connection to centromeres was investigated by aligning the centromere-associated repeat sequence characterized by Varadharajan et al. (17) against each of the unassigned contigs with blastn (BLAST+ applications version 2.2.31+) (39, 40). All contigs with at least one hit with e-value lower than 10^{-5} were classified as putative centromeric contigs (Fig. 1a). Results were visualized with R (ver. 3.4.4; R Core Team 2018 https://www.R-project.org/) using packages ggplot2 (ver. 3.0.0) (41) and ggforce (ver. 0.3.1; Pedersen 2019 https://CRAN.R-project.org/package=ggforce).

Centromere annotation

Centromere positions were identified by aligning the centromere-associated repeat sequence against each linkage group (see above; Supplementary Fig. S5). Blast alignments with low mapping quality (E-value $\geq 10^{-5}$) were discarded and remaining hits for each linkage group were assigned to three groups with k-mean clustering according to their position. Clusters with less than 10% of the total number of hits were considered outliers and discarded. Centromeric region was defined to span the region between the most distant hits. Clustering was conducted and the results were visualized with R.

Synteny analysis

The quality of ver. 6 and ver. 7 were assessed by comparing their synteny with the three-spined stickleback genome (18). Based on the previously reported large-scale synteny to the three-spined stickleback genome (17); see also (42, 33), the homologous linkage groups of nine- and three-spined sticklebacks were aligned with minimap2 software. Previous studies (21, 33) have shown that LG12 is a fusion chromosome, and it was aligned against the three-spined stickleback linkage groups 7 (1–14 Mbp) and 12. Alignment fragments with less than 5000 matching base pairs were discarded and the syntenies of the two assemblies with the three-spined stickleback genome were compared by counting the number of changes in orientation of consecutive fragments (Fig. 3b).

Content of the sex chromosome LG12

Based on the sequencing coverage of the female and male individuals, the sex-determination part (1–25 Mpb) of the ver. 6 LG12 appeared to contain contigs derived from X and Y chromosomes. We aimed to make LG12 haploid and purely X, and to identify significantly different Y-origin haplotypes. For that, X/Y-related contigs were removed as highly similar haplotypes, or labelled and retained as differentiated X/Y-related haplotypes (Table 1). To investigate the composition of the new assembly of LG12 we used GATK4 to joint-call variable sites in a pool of the reference individual and four females, and defined a HMM based on the frequency of homozygous reference and variant alleles in females. We assumed that females are homozygous for the reference allele in regions representing X and homozygous for the variant allele in regions representing Y. The emitted statistic was $\frac{[0/0]}{[0/0]+[1/1]+1} \times 100$, where [0/0] and [1/1] are the number of loci where an individual is

homozygous for reference or variant allele, respectively. The statistic was calculated in 50 kb windows and rounded to the closest integer. Low and high values of the statistic indicate X and Y chromosomes, respectively, whereas values of around 50 indicate fine-scale mosaicism of X and Y. The analysis was carried out in the sex-determination region of the ver. 7 LG12 (1–16.9 Mbp) with R package HMM (ver. 1.0; Scientific Software Development, Himmelmann 2010 https://CRAN.R-project.org/package=HMM).

Variant analysis for assembly comparisons

To compare the nine-spined ver. 6 and ver. 7 references, we called autosomal SNPs of the reference individual (FIN-PYO-0). Reads were mapped to both references using BWA-MEM and variants were called with bcftools mpileup. To ensure that SNPs were called reliably, we applied several criteria for pruning: SNPs in close proximity of repetitive regions, within 20 bp of an indel, of low quality (< 20) or with low (< 30) or high (> 70) depth were discarded. Finally, only SNPs in regions of good mappability were included. These regions were determined using approach of Li (http://lh3lh3.users.sourceforge.net/snpable.shtml) and converted to bed format using script by Schiffels (https://github.com/stschiff/msmc-tools). SNPs found using ver. 6 were grouped in three categories: found in autosomal linkage groups of ver. 7, locus removed from autosomal linkage groups of ver. 7, or not called with ver. 7. SNPs called using ver. 7 were grouped similarly but there were two additional groups for SNPs in regions where haplotype copy was removed (Table2).

Characterization and assembly of the LG19 inversion

Mapping of sequencing data on the ver. 6 reference genome revealed a block of high nucleotide diversity approximately from 5.8–7.5 Mbp in LG19 (Fig. 4a). Based on the Marey map (Supplementary Fig. S6), this appeared to be caused by two inversion haplotypes. To study this region in more detail, the original PacBio data for the reference individual were combined with resequencing data of individuals from the Pyöreälampi pond and nearby Kirkasvetinen pond (66°26'N, 29°08'E). Sequencing of additional individuals was performed on Illumina HiSeq 2000/4000 and on Nextseq 500 (150bp paired-end reads) to 10-20x coverage (BGI, Hong Kong and Institute of Biotechnology, Finland). Reads were mapped to the ver. 6 reference genome using BWA-MEM, SAMtools and GATK3 (ver. 3.8.0, commands RealignerTargetCreator, IndelRealigner) (43). Variants were called with GATK4 (ver. 4.0.1.2; commands HaplotypeCaller, CombineGVCFs, GenotypeGVCFs) (44, 45) as a part of a larger sample set.

Haplotypic contigs associated with the region were identified using pair-end resequencing data. Using program Plink (ver. v1.90b6.9) (46) and variant data from Pyöreälampi pond and the nearby Kirkasvetinen pond, heterozygosities and identity-by-state distances across the inversion region were computed and the latter was used to infer a neighbor-joining tree. Based on the clustering of low-heterozygosity samples, 14 and 13 individuals were identified to be homozygotes for the different alleles, A and B, respectively. Utilizing resequencing data for these individuals read-pairs were scanned and those with one read mapping to the assembled inversion region and the other read to an unplaced contig were identified. 30 unplaced contigs were identified using a threshold of at least 10 inward edges.

PacBio data was aligned to the original ver. 6 assembly, reads mapping to the inversion region or to selected unplaced contigs were identified, and the original full length PacBio reads were extracted. Reads longer than 5 kb in length were kept and the data, PacBio-5k, were assembled with Falcon Unzip (ver. 0.4.0) (13) and its default parameters. Using Trio Binning (prerelease version) (23) unique k-mers among the a and b alleles in Illumina data were searched and the PacBio-5k reads were binned into allele-specific subsets.

These subsets, PacBio-5k-A and PacBio-5k-B, were assembled separately with Canu (ver. 1.6) (47) and its default parameters, and the resulting contigs were polished with Pilon (ver. 1.22) (48) using the allele-specific Illumina data.

The Falcon Unzip graph was visualized with Bandage (ver. 0.8.1) (49), the Canu contigs for each haplotype were aligned separately against it, and the contigs were ordered manually based on their alignment (Supplementary Fig. S7). To confirm that the two segregating haplotypes originate from an inversion, they were aligned to each other and to the reference genome with minimap2. Alternative versions of the region were created by inserting the newly assembled alles into the reference sequence.

Individuals homozygous for A and B alleles were mapped to different versions of LG19 with BWA-MEM and SAMtools. Variants were called with bcftools mpileup (ver. 1.9) (50) and single nucleotide variants with quality score ≥ 5 were retained. Frequencies of sites with homozygous and heterozygous variant alleles were calculated in 100 kb windows with Bedtools software (ver. 2.27.1) (51).

To identify other potential inversion haplotypes in the genome, we defined a HMM. We anticipated that a dense mosaic of haplotypes in the reference genome results in variation between homozygous reference and variant alleles in an individual homozygous for one haplotype. Therefore, the emitted statistic was defined as $-10\log_{10}(\frac{([0/0]-[1/1])^2+1}{([0/0]+[1/1])^2+1})$, where [0/0] and [1/1] are the number of loci where an individual is homozygous for reference or variant allele, respectively. The statistic was estimated in 50kb windows and rounded to the closest integer; values above 40 were truncated to 40. Small values (e.g. high proportion of both homozygous genotypes) indicated inversion region. The HMM was applied to the four female individuals and all 21 linkage groups.

Contig classification, variant analysis, synteny comparisons and other downstream analyses of the genome assembly, were executed using Anduril 2 workflow platform (52).

Three-spined stickleback assembly and linkage map

We also tested the performance of Lep-Anchor with the three-spined stickleback genome assembly (18). First, a linkage map was constructed with Lep-MAP3 based on the data set of 517 F_1-offspring from 60 families (30 males, each crossed with two females) described by Pritchard et al. (53). The parents were wild caught from the Baltic Sea and artificially crossed in laboratory facilities (see refs. (54, 53)) for more details. The linkage map reconstruction was similar to that of the nine-spined stickleback with two differences: the pedigree was obtained from Pritchard et al. (53) and, in SeparateChromosomes2, lodLimit was set to 25 to obtain 21 linkage groups.

The original scaffolded genome was partitioned into (about 16,000) underlying contigs by cutting it at long runs of N's. We then made an artificial map containing one marker per contig, listing contigs in the scaffold order within each of the 21 linkage groups. To allow deviations from the contig order of Peichel et al. (18), the marker for the i:th contig was given a map interval of [i, i+9]. Finally, an artificial alignment file (paf format) was constructed with alignments for each adjacent contig in the scaffolds. As for the nine-spined stickleback, we then run the Lep-Anchor pipeline using the linkage map produced with Lep-MAP3 and the artificial map and alignment files. To further improve the result, we incorporated a scaffold level 10X Genomics genome assembly (Boot Lake population, Vancouver Island, Canada) (19) into the input data. The assembly was aligned with minimap2 to the three-spined stickleback contigs and the resulting alignment was included as two copies to Lep-Anchor to increase its weight in the optimisation score.

To demonstrate the effect of haplotype removal, we called SNPs for a male three-spined stickleback from Paxton Lake benthic population (49°42'28"N, 124°31'30"W), Canada (55). The Illumina whole-genome

sequencing data (HiSeq 2000) for the sample SRR5626529 were downloaded from European Nucleotide Archive (ENA). Sequencing reads were mapped with BWA-MEM to the published three-spined stickleback genome and to the genome assembled here. SNPs were called with bcftools mpileup as in the nine-spined stickleback (see above). As the mean sequencing coverage of the sample as 15x, only SNPs with depth between 7 and 23 were retained.

Acknowledgements

We thank Chris Eberlein, Sami Karja, Takahito Shikano, and everyone else who helped in catching and rearing the fish utilized in this work. Thanks are also due to those who helped in DNA extractions, Laura Hänninen, Kirsi Kähkönen, Miinastiina Issakainen and Sami Karja in particular. We thank Antoine Fraimout for help with relatedness analysis. Our research was supported by LUOVA doctoral programme of the University of Helsinki (funding to MK), Academy of Finland (# 129662, 134728 and 218343 to JM; # 1322681 to AL), and grant from Helsinki Institute for Life Sciences (HiLife; to JM). We also wish to acknowledge CSC – IT Center for Science, Finland, for computational resources.

Author contributions

J.M. directed and organized the data collection. P.R. generated linkage maps and assembled reference genomes. M.K., A.L. and P.R. carried out the data analysis and interpretation of the results. M.K. was responsible for structuring and compiling the manuscript. The manuscript was written and edited by all authors.

References

1. Korneliussen, T. S., Albrechtsen, A. & Nielsen, R. ANGSD: analysis of next generation sequencing data. *BMC Bioinformatics* **15,** 356 (2014).
2. Schraiber, J. G. & Akey, J. M. Methods and models for unravelling human evolutionary history. *Nat. Rev. Genet.* **16,** 727–740 (2015).
3. Church, D. M. *et al.* Modernizing reference genome assemblies. *PLoS Biol.* **9,** e1001091 (2011).
4. Baker, M. *De novo* genome assembly: What every biologist should know. *Nat. Methods* **9,** 333–337 (2012).
5. Meltz Steinberg, K., Schneider, V. A., Alkan, C., Montague, M. J., Warren, W. C., Church, D. M. & Wilson, R. K. Building and Improving Reference Genome Assemblies. *P. IEEE* **105,** 422–435 (2017).
6. Rice, E. S. & Green, R. E. New approaches for genome assembly and scaffolding. *Annu. Rev. Anim. Biosci.* **7,** 17–40 (2019).
7. Stemple, D. L. So, you want to sequence a genome... *Genome Biol.* **14,** 128 (2013).
8. Pengelly, R. J. & Collins, A. Linkage disequilibrium maps to guide contig ordering for genome assembly. *Bioinformatics* **35,** 541–545 (2019).
9. Rastas, P. Lep-Anchor: automated construction of linkage map anchored haploid genomes. *Bioinformatics* **36,** 2359–2364 (2020).

10. Schneider, V. A. *et al.* Evaluation of GRCh38 and *de novo* haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.* **27,** 849–864 (2017).

11. Sherman, R. M. & Salzberg, S. L. Pan-genomics in the human genome era. *Nat. Rev. Genet.* **21,** 243–254 (2020).

12. Howe, K. *et al.* The zebrafish reference genome sequence and its relationship to the human genome. *Nature* **496,** 498–503 (2013).

13. Chin, C.-S. *et al.* Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* **13,** 1050–1054 (2016).

14. Kong, A. *et al.* Rate of *de novo* mutations and the importance of father's age to disease risk. *Nature* **488,** 471–475 (2012).

15. Ceballos, F. C., Joshi, P. K., Clark, D. W., Ramsay, M. & Wilson, J. F. Runs of homozygosity: windows into population history and trait architecture. *Nat. Rev. Genet.* **19,** 220–234 (2018).

16. Roux, C., Fraïsse, C., Romiguier, J., Anciaux, Y., Galtier, N. & Bierne, N. Shedding light on the grey zone of speciation along a continuum of genomic divergence. *PLoS Biol.* **14,** 1–22 (2016).

17. Varadharajan, S. *et al.* A high-quality assembly of the nine-spined stickleback (*Pungitius pungitius*) Genome. *Genome Biol. Evol.* **11,** 3291–3308 (2019).

18. Peichel, C. L., Sullivan, S. T., Liachko, I. & White, M. A. Improvement of the threespine stickleback genome using a Hi-C-based proximity-guided assembly. *J. Hered.* **108,** 693–700 (2017).

19. Berner, D. *et al. De Novo* Sequencing, assembly, and annotation of four threespine stickleback genomes based on microfluidic partitioned DNA libraries. *Genes* **10,** 426 (2019).

20. Han, E., Sinsheimer, J. S. & Novembre, J. Characterizing bias in population genetic inferences from low-coverage sequencing data. *Mol. Biol. Evol.* **31,** 723–735 (2014).

21. Shikano, T., Laine, V. N., Herczeg, G., Vilkki, J. & Merilä, J. Genetic architecture of parallel pelvic reduction in ninespine sticklebacks. *G3: Genes Genom. Genet.* **3,** 1833–1842 (2013).

22. Guo, B., Fang, B., Shikano, T., Momigliano, P., Wang, C., Kravchenko, A. & Merilä, J. A phylogenomic perspective on diversity, hybridization and evolutionary affinities in the stickleback genus *Pungitius*. *Mol. Ecol.* **28,** 4046–4064 (2019).

23. Koren, S. *et al. De novo* assembly of haplotype-resolved genomes with trio binning. *Nat. Biotechnol.* **36,** 1174–1182 (2018).

24. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526,** 75–81 (2015).

25. Paten, B., Novak, A. M., Eizenga, J. M. & Garrison, E. Genome graphs and the evolution of genome inference. *Genome Res.* **27,** 665–676 (2017).

26. Ballouz, S., Dobin, A. & Gillis, J. A. Is it time to change the reference genome? *Genome Biol.* **20,** 1–9 (2019).

27. Huang, S., Kang, M. & Xu, A. HaploMerger2: rebuilding both haploid sub-assemblies from high-heterozygosity diploid genome assembly. *Bioinformatics* **33,** 2577–2579 (2017).

28. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34,** 3094–3100 (2018).

29. Guo, B., Toli, E.-A. & Merilä, J. Complete mitochondrial genome of the nine-spined stickleback *Pungitius pungitius* (Gasterosteiformes, Gasterosteidae). *Mitochondr. DNA Part B* **1,** 72–73 (2016).

30. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31,** 1674–1676 (2015).

31. Kawahara, R., Miya, M., Mabuchi, K., Near, T. J. & Nishida, M. Stickleback phylogenies resolved: evidence from mitochondrial genomes and 11 nuclear genes. *Mol. Phylogenet. Evol.* **50,** 401–404 (2009).

32. Hwang, D.-S., Song, H. B. & Lee, J.-S. Complete mitochondrial genome of the Amur stickleback *Pungitius kaibarae* (Gasterosteiformes, Gasterosteidae). *Mitochondr. DNA* **23,** 313–314 (2012).

33. Rastas, P., Calboli, F. C. F., Guo, B., Shikano, T. & Merilä, J. Construction of ultradense linkage maps with Lep-MAP2: stickleback F_2 recombinant crosses as an example. *Genome Biol. Evol.* **8,** 78–93 (2016).

34. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at http://arxiv.org/abs/1303.3997, 1–3 (2013).

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

36. Rastas, P. Lep-MAP3: robust linkage mapping even for low-coverage whole genome sequencing data. *Bioinformatics* **33,** 3726–3732 (2017).

37. Baum, L. E. *et al.* An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. *Inequalities* **3,** 1–8 (1972).

38. Varadharajan, S. Ninespine_RepeatLibrary.fasta. figshare. `https://figshare.com/articles/dataset/Ninespine_RepeatLibrary_fasta/10565792` (2019).

39. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).

40. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. & Madden, T. L. BLAST+: architecture and applications. *BMC Bioinformatics* **10,** 421 (2009).

41. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag New York, 2016).

42. Guo, B., Chain, F. J. J., Bornberg-Bauer, E., Leder, E. H. & Merilä, J. Genomic divergence between nine- and three-spined sticklebacks. *BMC Genomics* **14,** 756 (2013).

43. McKenna, A. *et al.* The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–1303 (2010).

44. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43,** 491–498 (2011).

45. Poplin, R. *et al.* Scaling accurate genetic variant discovery to tens of thousands of samples. Preprint at https://doi.org/10.1101/201178 (2017).

46. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).

47. Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H. & Phillippy, A. M. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27,** 722–736 (2017).

48. Walker, B. J. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* **9,** e112963 (2014).

49. Wick, R. R., Schultz, M. B., Zobel, J. & Holt, K. E. Bandage: interactive visualization of *de novo* genome assemblies. *Bioinformatics* **31,** 3350–3352 (2015).

50. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27,** 2987–2993 (2011).

51. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).

52. Cervera, A. *et al.* Anduril 2: upgraded large-scale data integration framework. *Bioinformatics* **35,** 3815–3817 (2019).

53. Pritchard, V. L., Viitaniemi, H. M., McCairns, R. J. S., Merilä, J., Nikinmaa, M., Primmer, C. R. & Leder, E. H. Regulatory architecture of gene expression variation in the threespine stickleback *Gasterosteus aculeatus*. *G3: Genes Genom. Genet.* **7,** 165–178 (2017).

54. Leder, E. H. *et al.* The evolution and adaptive potential of transcriptional variation in sticklebacks—signatures of selection and widespread heritability. *Mol. Biol. Evol.* **32,** 674–689 (2014).

55. Samuk, K., Owens, G. L., Delmore, K. E., Miller, S. E., Rennison, D. J. & Schluter, D. Gene flow and selection interact to promote adaptive divergence in regions of low recombination. *Mol. Ecol.* **26,** 4378–4390 (2017).

Tables

Table 1: Summary of the differences between the two nine-spined stickleback genome assemblies

Feature	ver. 6	ver. 7	%Change
N50 contig size (bp)	1,202,809	2,794,615	232.34
Total length of the assembly (bp)[1)]	521,233,387	466,582,808	–10.48
Total length of the 21 linkage groups (bp)	444,482,085	439,721,235	–1.07
LG12 length (bp)	40,899,740	33,585,825	–17.88
Contigs in linkage groups (contig chains)[2)]	686 (not assessed)	843 (362)	22.89 (NA)
Contigs in LG12	244	150	–38.52
Contigs not assigned to linkage groups (length)	4,616 (76,734,720 bp)	1,644 (27,251,636 bp)	–64.38 (–64.49)
Contigs not in linkage groups of other assembly (LG12)	117 (109)	274 (15)	
Contigs with known orientation	Not assessed in ver. 6	763 (427,086,963 bp)	

[1)] Includes 21 linkage groups with gaps, unassigned contigs and mitochondrial sequence.
[2)] Contig chain refers to group of contigs joined without gap in ver. 7. In ver. 6, all contigs had a gap in between.

Table 2: Number of autosomal SNPs detected by mapping short read data against the published and the new assembly of the stickleback genomes. ”Other” indicates SNPs that were identified in regions with no changes or removed haplotype pair. The reference individual and a Paxton Lake individual were used for the nine-spined and three-spined stickleback, respectively.

	SNPs in different assemblies						
	Both	Published		New			
Species		Not in New	Other	Haplot. removed	Haplot. trimmed	Not in Published	Other
Nine-spined	23,576	11	88	2,110	514	248	500
Three-spined	2,278,066	3,672	1,792	6,912	3,217	5,160	4,342

Table 3: Summary of the differences between the two three-spined stickleback genome assemblies.

Feature	Peichel *et al.* 2017	New assembly	%Change
N50 contig size (bp)	83,717	87,370	4.4
Ungapped length of the assembly (bp)	446,619,272	435,522,568	–2.50
Ungapped length of the 21 linkage groups (bp)	426,017,013	422,866,090	–0.74
Ungapped length of the unassigned contigs (bp)	20,602,259	12,656,478	–39.57
No. contigs in linkage groups	13,460	12,946	–3.82
No. contig chains in linkage groups	Not assigned	12,723	
Contigs not in linkage groups of other assembly	60	176	
No. unassigned contigs	3,373	2,056	39.05
No. removed contigs	Not applicable	1,831	
Length of removed contigs (bp)	Not applicable	8,341,160	

Figures

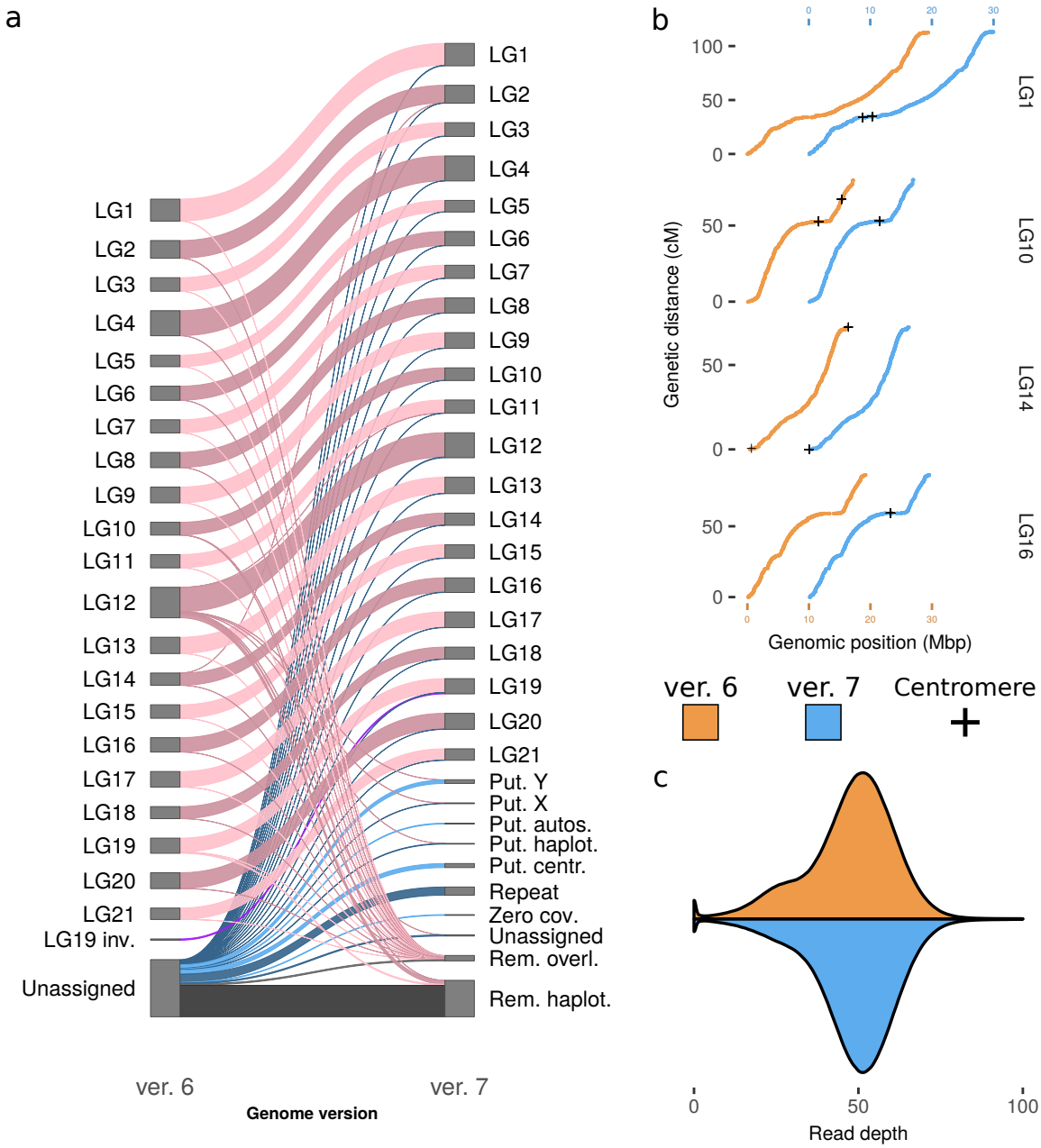

Figure 1: Fig. 1 Summary of the changes between ver. 6 and ver. 7 of the nine-spined stickleback reference genome. (a) Diagonal lines indicate changes in contig placement between different linkage groups (LGs) with band widths proportional to the length of the contigs with the corresponding change. Unassigned contigs in ver. 7 were grouped into putative classes according to their sequencing coverage and repeat content (see methods). (b) Changes in centromere annotation: centromeres were earlier not found in LG1 and LG16 and the annotation became more precise in LG10 and LG14. (c) Coverage of reference individual Illumina data is closer to expected (50x) in ver. 7 and fewer regions show 0 or half (25x) depth

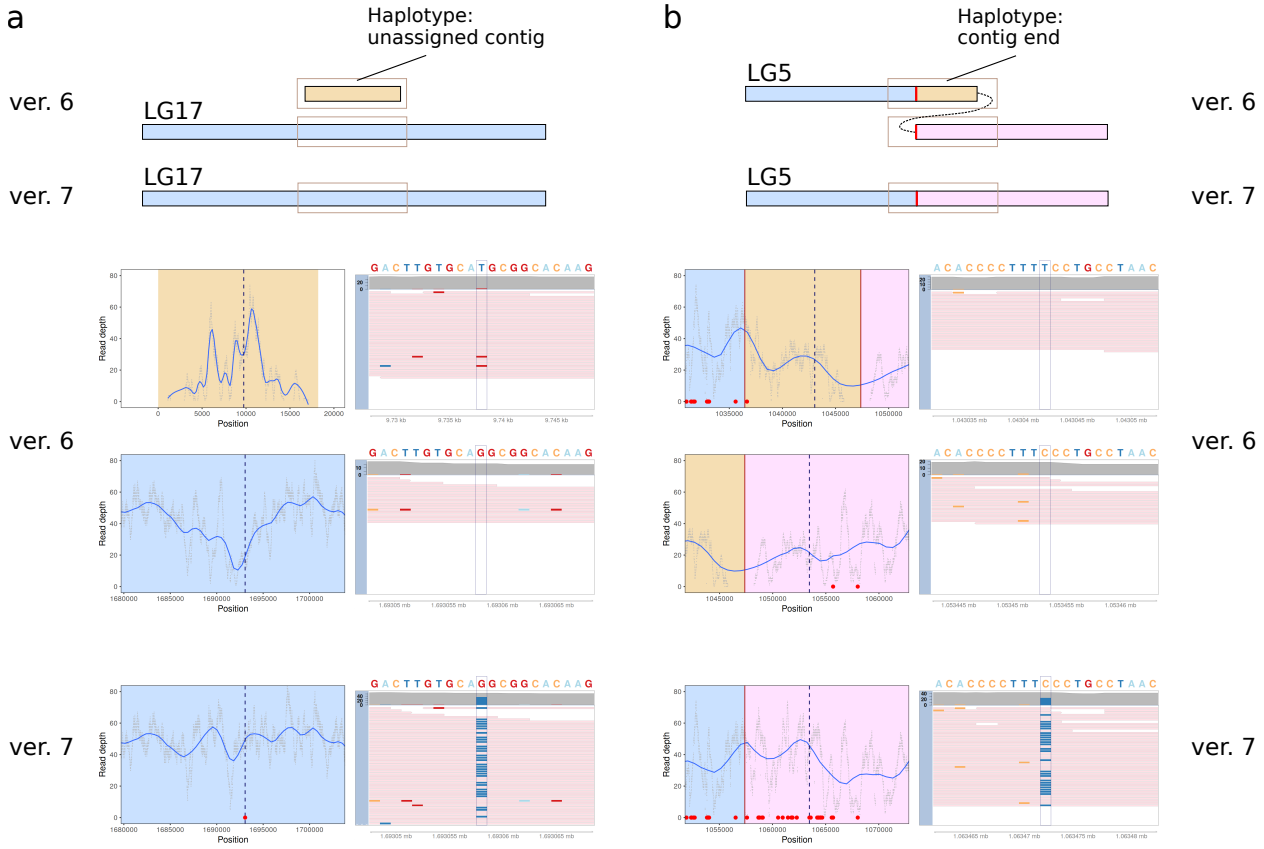

Figure 2: Examples of haplotype removal and identification of new SNPs. A schematic illustration of regions in the two assemblies is shown on top and the data for the highlighted areas (boxes) in the panels below. On the left, blue curves show the smoothed read depth and the dashed lines indicate a SNP position, boxed in the right panel. On the right, the reference sequence is shown on top and the pink bars indicate mapped reads, mismatches shown with matching colors. (a) A short, unassigned contig (orange) was identified as a haplotype within a contig (blue) in LG17. After removal of the short contig (ver. 7, bottom), the read depth is more even and a new SNP (red dot) is identified. (b) A region (orange) was duplicated in the ends of neighboring contigs (blue, pink) in LG5. After its removal (ver. 7, bottom; cut site in red), the read depth is more even and several new SNPs are identified.

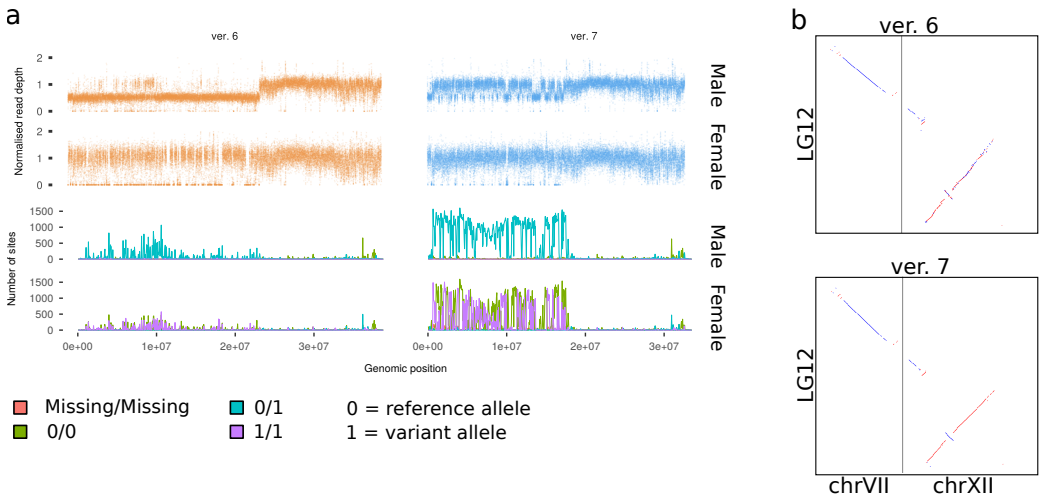

Figure 3: Improvements in LG12 sex chromosome. (a) Normalised read depth of the male reference individual (50x coverage) and a female (10x, FIN-PYO-20) is closer to the expected (one) in ver. 7 (right) and fewer regions show zero depth. Ver. 7 has more segregating sites and especially sites where the reference individual is heterozygous (turquoise) and the female is homozygous (green, purple). Number of sites is calculated in 100 kb windows. (b) The synteny of the nine-spined stickleback LG12 with the three-spined stickleback genome (x axis) is more contiguous in ver. 7, and there are fewer changes in contig order. Red and blue colors indicate forward and reverse alignments, respectively.

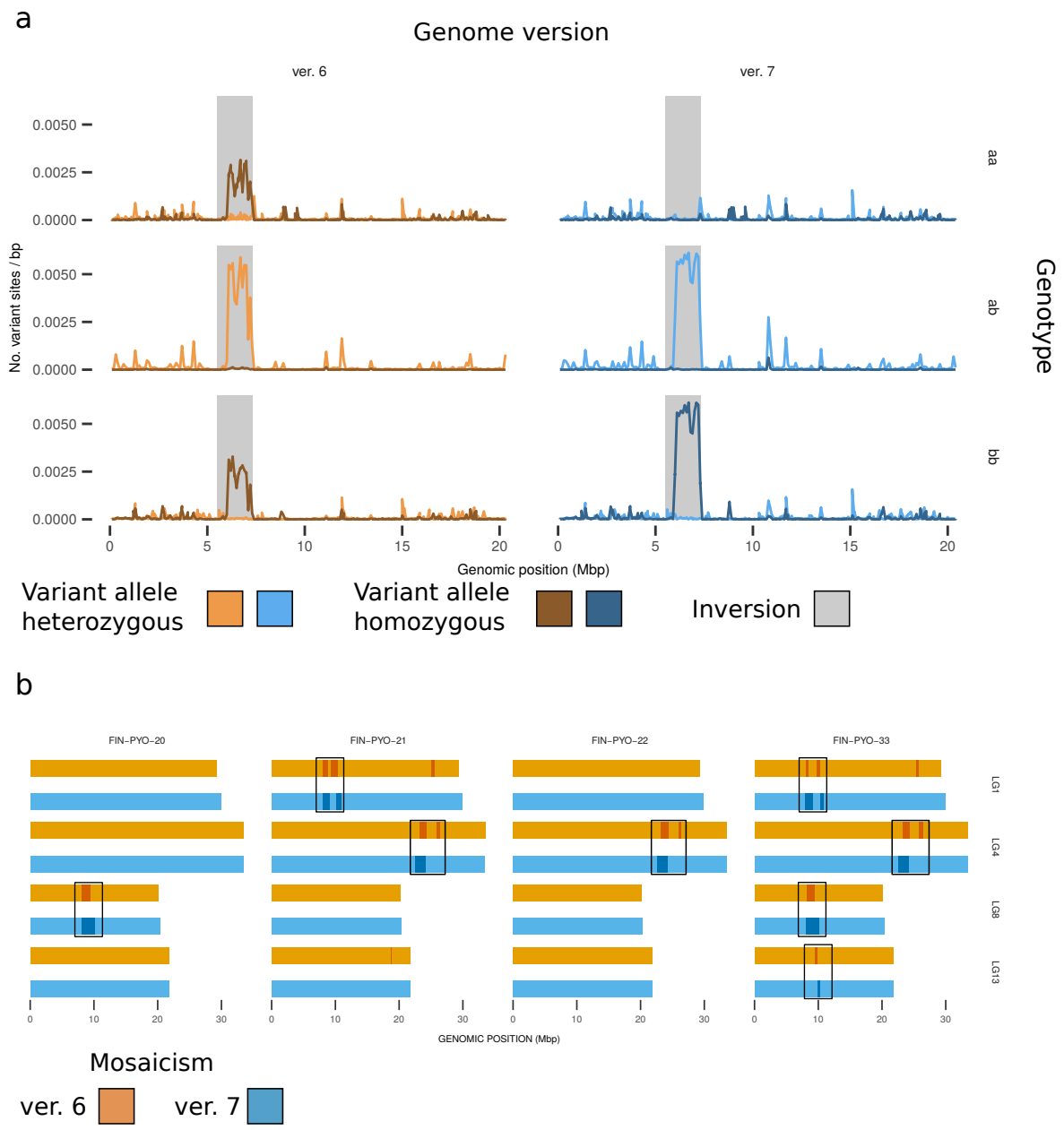

Figure 4: Examples of finescale mosaicism in the nine-spined stickleback reference geome. (a) When mapped to ver. 6, individuals homozygote for the LG19 inversion haplotypes (aa and bb; top and bottom) show high frequency of variant alleles in the inversion region. In ver. 7 with the reassembled inversion haplotype, individuals homozygous for the assembled haplotype (top) have no variant alleles whereas those homozygous for the alternative allele (bottom) show two-fold frequency across the region. (b) Using a HMM, four candidates of finescale mosaicism (dark bands) similar to the LG19 inversion were identified. Regions detected at homologous regions in both genome versions are highlighted.