American Genetic Association

Journal of Heredity, 2017, 693–700
doi:10.1093/jhered/esx058
Brief Communication
Advance Access publication August 3, 2017

OXFORD

Brief Communication

Improvement of the Threespine Stickleback Genome Using a Hi-C-Based Proximity-Guided Assembly

Catherine L. Peichel, Shawn T. Sullivan, Ivan Liachko, and Michael A. White

From the Divisons of Basic Sciences and Human Biology, Fred Hutchinson Cancer Research Center, Seattle, WA (Peichel); Phase Genomics, Inc., Seattle, WA (Sullivan and Liachko); and Department of Genetics, University of Georgia, 120 Green Street, Athens, GA 30602 (White). Catherine L. Peichel is now at the Institute of Ecology and Evolution, University of Bern, 3012 Bern, Switzerland.

Address correspondence to M. A. White at the address above, or e-mail: whitem@uga.edu.

Received December 18, 2016; First decision February 20, 2017; Accepted July 5, 2017.

Corresponding Editor: William Murphy

Abstract

Scaffolding genomes into complete chromosome assemblies remains challenging even with the rapidly increasing sequence coverage generated by current next-generation sequence technologies. Even with scaffolding information, many genome assemblies remain incomplete. The genome of the threespine stickleback (*Gasterosteus aculeatus*), a fish model system in evolutionary genetics and genomics, is not completely assembled despite scaffolding with high-density linkage maps. Here, we first test the ability of a Hi-C based proximity-guided assembly (PGA) to perform a de novo genome assembly from relatively short contigs. Using Hi-C based PGA, we generated complete chromosome assemblies from a distribution of short contigs (20–100 kb). We found that 96.40% of contigs were correctly assigned to linkage groups (LGs), with ordering nearly identical to the previous genome assembly. Using available bacterial artificial chromosome (BAC) end sequences, we provide evidence that some of the few discrepancies between the Hi-C assembly and the existing assembly are due to structural variation between the populations used for the 2 assemblies or errors in the existing assembly. This Hi-C assembly also allowed us to improve the existing assembly, assigning over 60% (13.35 Mb) of the previously unassigned (~21.7 Mb) contigs to LGs. Together, our results highlight the potential of the Hi-C based PGA method to be used in combination with short read data to perform relatively inexpensive de novo genome assemblies. This approach will be particularly useful in organisms in which it is difficult to perform linkage mapping or to obtain high molecular weight DNA required for other scaffolding methods.

Subject area: Genomics and gene mapping
Key words: chromosome conformation capture, de novo genome assembly, *Gasterosteus aculeatus*

Although short-read genome sequencing has become a staple in genetic research, scaffolding complete eukaryotic genomes from fragmented assemblies remains a remarkably difficult task as most modern scaffolding techniques utilize purified high-molecular weight DNA as the source of contiguity information (Das et al. 2010; Teague et al. 2010; Putnam et al. 2016). Purification results in broken DNA

molecules and loss of long-range intra-chromosomal genetic contiguity information typically yielding incomplete scaffolds. One traditional method for retaining chromosome-scale contiguity is to use genetic crosses to establish maps of relative linkage distances between sequences. However, genetic mapping is very laborious, cannot be applied to many organisms, and often falls short of scaffolding all contigs due to a low resolution from a limited number of crossovers (reviewed in Fierst 2015). Chromosome conformation capture techniques such as Hi-C (Lieberman-Aiden et al. 2009) retain ultra-long-range genomic contiguity information through in vivo crosslinking of chromatin and subsequent sequencing of proximal pairs of sequences. The rate of Hi-C interaction decreases rapidly with increasing genomic distance between pairs of loci. Taking advantage of this relationship between intersequence distance and proximity interaction allows the construction of chromosome-scale genome scaffolds (Burton et al. 2013; Kaplan and Dekker 2013; Marie-Nelly et al. 2014; Bickhart et al. 2017; Dudchenko et al. 2017).

Here, we used the Hi-C-based proximity-guided assembly (PGA) method to assemble the genome of the threespine stickleback (*Gasterosteus aculeatus*). This small, teleost fish is a widely used model system in diverse fields including ecology, evolution, behavior, physiology, and toxicology (Wootton 1976; Bell and Foster 1994; Östlund-Nilsson et al. 2007). Sticklebacks are well known for the extensive morphological, behavioral, and physiological variation present in freshwater populations that have evolved since the retreat of the glaciers across the Northern hemisphere in the past 15 000 years (Bell and Foster 1994; Hendry et al. 2013). Recent research has led to the identification of the genetic and genomic basis of this phenotypic diversity, providing new insights into the genetic basis of adaptation (Peichel and Marques 2017). To facilitate research in this model system, a high-quality genome assembly for *G. aculeatus* was generated by Sanger sequencing of plasmid, fosmid, and BAC genomic libraries made from a single female from Bear Paw Lake, AK. Scaffolds were anchored to the 21 known stickleback chromosomes or linkage groups (LG) using genetic linkage mapping. The original genome assembly comprised 400.7 Mb of scaffolds anchored to LGs, with an additional 60.7 Mb of assembled scaffolds not anchored to LGs (Jones et al. 2012). Two further revisions to the genome assembly have used genetic linkage mapping in 3 additional crosses to assign some of these unanchored scaffolds to LGs and to correct errors in the original assembly (Roesti et al. 2013; Glazer et al. 2015). The most recent genome assembly comprises 436.6 Mb, with 26.7 Mb remaining unassigned to LGs (Glazer et al. 2015). Here, we took advantage of the existence of a high-quality assembly for *G. aculeatus* to test the performance of Hi-C in generating a de novo assembly. Additionally, we used Hi-C to further improve the existing *G. aculeatus* genome assembly.

Methods

Tissue Collection and Hi-C Sequencing

The liver of a single, lab-reared adult male from the Paxton Lake benthic population (Texada Island, British Columbia) was dissected and flash frozen in liquid nitrogen. Tissue processing, chromatin isolation, library preparation, and sequencing were all performed by Phase Genomics (Seattle, WA). A total of 176 461 081 read-pairs were sequenced.

PGA Scaffolding

Scaffolding was conducted in 2 phases. First, to scaffold the entire *G. aculeatus* revised genome assembly (Glazer et al. 2015), the genome was divided into 8342 contiguous contigs of varied length, excluding contigs that were not previously assigned to LGs and the mitochrondria sequence. Gaps present within the revised genome assembly were not removed before dividing it into contigs; the resulting contigs, therefore, included gaps ranging from 0.00% to 100.00% of their length, with a median of 0.31% and mean of 2.69% of the length of a contig being composed of gaps. Contig length followed a normal distribution that ranged from 20 to 100 kb (median contig size: 52 339 bp; standard deviation 18 261 bp). Paired-end reads were aligned to the contigs, only retaining reads that aligned uniquely. Contigs were scaffolded using PGA with an adapted version of the Lachesis method (Burton et al. 2013) by Phase Genomics. The known number of *G. aculeatus* chromosomes (21) was used as a starting input parameter during the scaffolding process (Ross and Peichel 2008). The final set of Lachesis parameters were selected from randomized parameter sweeps from over 60 000 scaffolding iterations. Parameters were varied within the following bounds: CLUSTER_N between 1 and 46; CLUSTER_MIN_RE_SITES between 1 and 4988; CLUSTER_MAX_LINK_DENSITY between 0.0008 and 28.9; CLUSTER_NONINFORMATIVE_RATIO between 1.0004 and 24.3; ORDER_MIN_RES_IN_TRUNK between 1 and 5131; and ORDER_MIN_RES_IN_SHREDS between 1 and 6489. The 4 best sets of candidate parameters were identified among these sweeps that best reflected the expected patterns of the Hi-C data and the likelihood of the resulting scaffolds having generated the observed Hi-C data. The patterns examined were intra-cluster link density (the ratio of Hi-C linkage contained within scaffolds as opposed to between them), ordering enrichment (the concentration of observed Hi-C link density between contigs near each other as compared to a null hypothesis of uniform Hi-C link density), and orientation quality score (the differential log-likelihood of the chosen orientation of a contig having resulted in the observed Hi-C data as compared to alternatives) (for additional detail see Burton et al. 2013; Bickhart et al. 2017). The final set of parameters that generated the largest scaffolds were CLUSTER_N = 21, CLUSTER_MIN_RE_SITES = 141, CLUSTER_MAX_LINK_DENSITY = 1.593, CLUSTER_NONINFORMATIVE_RATIO = 5.163, ORDER_MIN_N_RES_IN_TRUNK = 69, ORDER_MIN_N_RES_IN_SHREDS = 18.

The second phase of scaffolding used PGA to assign the contigs that were previously not assigned to LGs to gaps in the reference genome. The reference assembly (excluding the mitochondria sequence) was split into contigs at gaps and Ns were removed, and all contigs (13 435 contigs previously assigned to LGs and 3499 contigs not previously assigned to LGs) were assembled with PGA. Previously unassigned contigs that were placed in the PGA assembly were divided into 3 groups based on the level of certainty in their placement. If the contigs assembled before (contig A) and after (contig B) the previously unassigned contig were sequential (i.e., occurred in the expected order relative to the reference assembly), the previously unassigned contig was considered an accurate placement and was inserted in the gap between contig A and contig B. If contig A and contig B were from the same LG, but were not sequential, the previously unassigned contig could not be accurately placed in the gap and was instead assigned to the LG within a narrowed range of possible locations. If contig A and contig B were from different LGs or if contig A or contig B were missing (i.e., the previously unassigned contig only had linkage information on one end), the previously unassigned contig could not be placed and was not considered further (216 total unplaced contigs).

BAC End Alignments

Sequenced BAC ends from the CHORI-215 BAC library made from 2 Paxton Lake benthic males (Kingsley et al. 2004; Kingsley and Peichel 2007) were aligned to the unmasked *G. aculeatus* revised genome assembly (Glazer et al. 2015) using the BLAST-like alignment tool (BLAT) (Kent 2002) (67 979 total paired BAC ends). Alignments were only retained if at least 90% of the sequenced BAC end aligned to the genome. If a BAC end aligned to multiple locations in the genome, the highest scoring alignment was kept according to the formula: alignment matches + alignment matches that are part of repeats – mismatches in alignment – number of gap openings in the query sequence – number of gap openings in the target sequence. Alignments were discarded if there were multiple alignments tied for the same highest alignment score. In addition, alignments were only considered if both BAC ends aligned to the same LG because we were only focused on identifying putative intrachromosomal rearrangements.

Overlap Between Misorderings in the PGA Assembly and BAC Ends That Aligned Discordantly

BAC ends that aligned in a forward/reverse orientation and were separated by over 250 kb in the genome (the average insert size of the library is 148 kb) were considered putative deletions in the Paxton Lake benthic population or insertions in the reference assembly. BAC ends that aligned in a forward/forward or reverse/reverse orientation in the genome were considered putative inversions. A contig was considered misordered in the PGA reassembly of the *G. aculeatus* genome if either of the neighboring contigs in the scaffold was located further than 250 kb away from the coordinates in the reference genome assembly (Glazer et al. 2015). This method defined the end breakpoints of misordered regions within the scaffolds. Overlap was scored if the position of a discordantly aligned BAC end fell within the contig that was misordered in the PGA assembly. Permutations were conducted for each LG separately and for the total genome to test for significance. Random subsets of BAC ends were drawn from each LG equal to the number of discordant BAC end alignments. Overlap was scored between the misordered PGA contigs and the random subsets of BAC ends. The *P*-value reflects how often the same number of overlaps is recovered among a set of 10 000 random permutations.

BioNano Scaffolding

High molecular weight DNA was isolated from the blood of a single adult male from the Paxton Lake benthic population (Texada Island, British Columbia) following protocols outlined in (Kingsley et al. 2004). This male was not the same individual used for Hi-C, nor for creation of the BAC libraries. Irys optical mapping (BioNano Genomics, San Diego, CA) was performed at Kansas State University. DNA was nicked with the BspQI restriction enzyme, which cuts at a frequency of 15.8 sites per 100 kb across the *G. aculeatus* genome, which is around the ideal cutting frequency of 10–15 sites/100 kb for optical mapping with the BioNano Irys System (Shelton et al. 2015). DNA was labeled with fluorescent nucleotides and repaired according to BioNano protocols. DNA was imaged on the BioNano Irys System using 2 IrysChips. BioNano molecules were filtered to only include segments that were at least 150 kb and contained at least 8 labels. The *P*-value threshold for the BioNano assembler was set to a minimum of 2.2×10^{-9}. Molecule stretch was adjusted using AssembleIrysCluster.pl (v. 1.6.1) (Shelton et al. 2015). To assess the effectiveness of BioNano optical maps in rescaffolding the *G. aculeatus* genome, the revised genome assembly was split into contiguous 100 kb contigs (the minimum recommended size for BioNano assembly). The split genome was digested with BspQI into *in silico* CMAP files using fa2cmap_multi.pl (BioNano) and iteratively scaffolded with the BioNano optical maps using sewing_machine.pl (v. 1.0.6) (Shelton et al. 2015). Two different filtering options were used: the default filters (--f_con 20, --f_algn 40, --s_con 15, --s_algn 90) and relaxed filters set at half of the default thresholds (--f_con 10, --f_algn 20, --s_con 7.5, --s_algn 45). For both sets of filters, the default alignment parameters were used (-FP 0.8, -FN 0.08 -sf 0.20 -sd 0.10).

Results

Hi-C/PGA Rescaffolding of the *G. Aculeatus* Genome Assembly

To investigate how well Hi-C-based PGA (provided by Phase Genomics) can assemble a genome composed of small contigs, we split the revised *G. aculeatus* genome assembly (Glazer et al. 2015) into contigs of varied length, ranging from 20 to 100 kb and used a PGA to rescaffold the contigs together. PGA reconstructed a highly accurate genome assembly, with 8042 of the 8342 (96.40% of contigs; 97.15% of the genome length) contigs correctly assigned to 1 of the 21 LGs in the *G. aculeatus* genome during clustering (1 contig was incorrectly assigned to an alternate LG, 2 contigs were not aligned within the cluster, and 297 contigs were not assigned to a LG) (Figure 1). Among the LGs, most of the contigs (7404 contigs, or 92.06% of the 8042 contigs correctly assigned to LGs) had an ordering identical to the revised *G. aculeatus* reference assembly (Figure 2; Supplementary Figure S1). The 638 contigs (34.33 Mb of the 436.6 Mb total genome length) that were ordered incorrectly

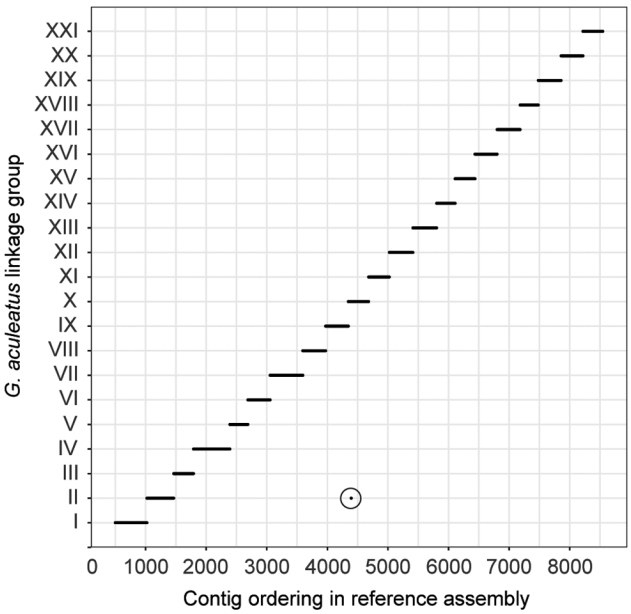

Figure 1. PGA clusters all LGs of the *Gasterosteus aculeatus* genome. The revised *G. aculeatus* genome assembly (Glazer et al. 2015) was divided into contiguous contigs of varying length (20–100 kb) and assembled using PGA. The *G. aculeatus* revised reference assembly contig order is preserved along the *x* axis. PGA clustering was largely congruent with the revised *G. aculeatus* reference genome, as shown by each LG assembled as a contiguous segment. One contig was assigned to a different LG in the PGA clustering than in the reference assembly (circled).

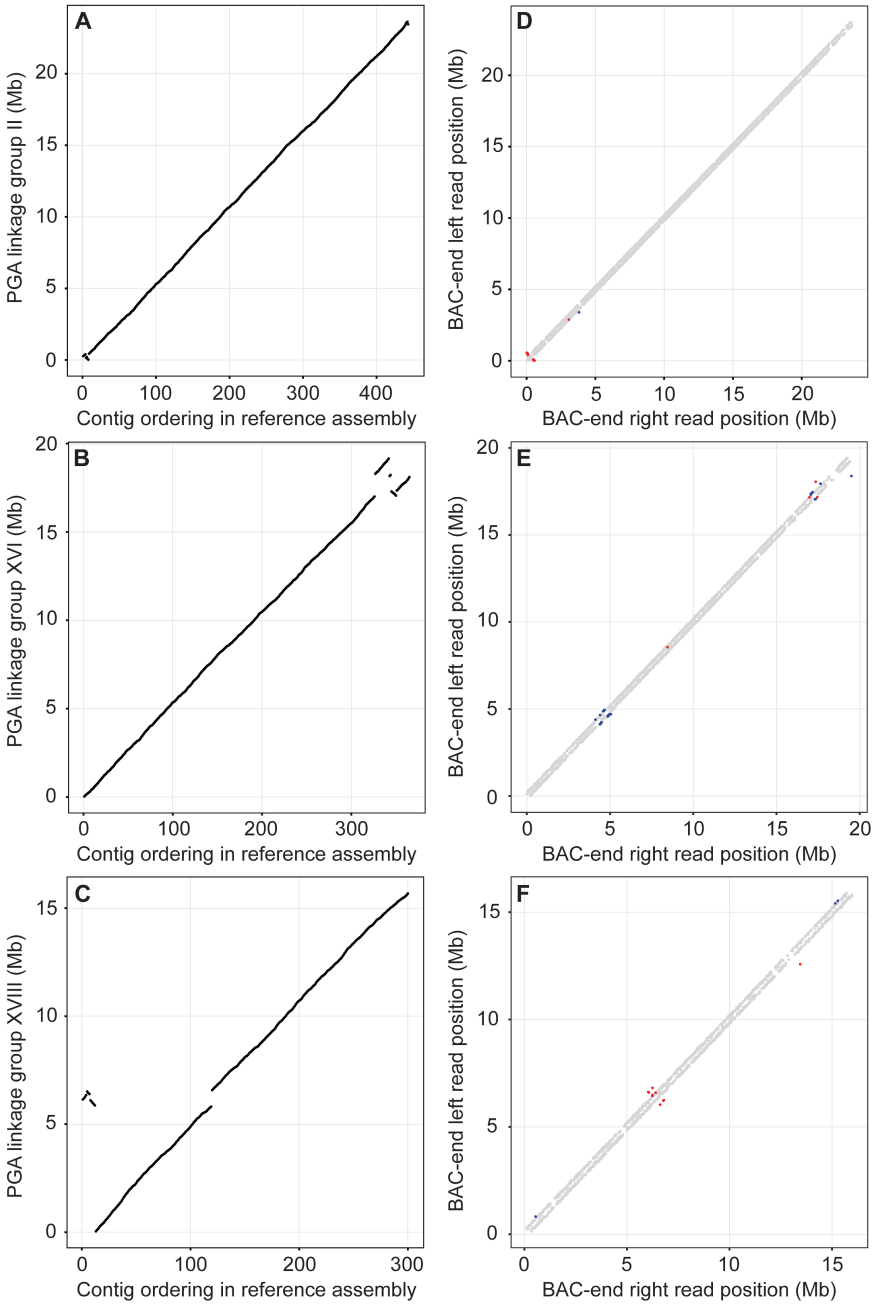

Figure 2. Misorderings within LGs are consistent with structural variation between populations of *Gasterosteus aculeatus*. Misorderings in the PGA scaffolding within LGs (**A**–**C**) and alignment of BAC mate-pair sequences (**D**–**F**) relative to the *G. aculeatus* revised reference assembly (Glazer et al. 2015) are shown for 3 LGs that had significant overlap between the 2 data sets: II (A, D), XVI (B, E), and XVIII (C, F). Alignment of BAC end sequences from the same Paxton Lake benthic population used for the PGA scaffolding reveal structural variation around the breakpoints of the misordered PGA scaffolds. Gray points are left and right concordantly aligned mate-pairs from the BAC library and fall slightly off either side of the 1:1 diagonal, reflecting the 148 kb average insert size of the library (Kingsley et al. 2004). Discordant read pairs are highlighted in color. Blue points indicate mate pairs that align in a forward/reverse orientation, reflecting a putative deletion relative to the reference genome assembly. Red points indicate mate pairs that align in a forward/forward or reverse/reverse orientation indicating a putative inversion relative to the reference genome assembly. The remaining LGs are shown in Supplementary Figure S1.

within LGs may represent errors in the PGA scaffolding, assembly errors in the reference genome, or could reflect structural variation between the Paxton Lake (British Columbia) benthic population used for the PGA scaffolding and the Bear Paw Lake (Alaska) population used for the reference assembly.

We explored whether the incorrect orderings within LGs were due to errors in PGA scaffolding or were due to structural variation between populations using the BAC end sequences available from a BAC library made from 2 Paxton Lake benthic males (Kingsley et al. 2004; Kingsley and Peichel 2007). We aligned the BAC end sequences to the revised *G. aculeatus* genome assembly (Glazer et al. 2015) and scanned for discordant mate-pair alignments (i.e., mate-pairs that aligned in the same orientation or that aligned across genomic regions larger than 250 kb, which is larger than the 148 kb average insert size of the library). Such discordant alignments would indicate large structural differences between the Paxton Lake benthic

population and the Bear Paw Lake reference assembly population, or errors in the Bear Paw reference genome assembly (Supplementary Table S1). In many LGs, we found that discordantly aligned BAC mate pairs were significantly more often associated with the contig at either end of the discordant orderings in the PGA scaffolding (Figure 2; Supplementary Figure S1; Table 1). This indicates that at least some of the PGA scaffold misorderings may reflect true insertions, deletions, or inversions between populations of *G. aculeatus*, and/or errors in the reference assembly. Full sequencing of these BAC clones will allow the breakpoints to be fine-mapped beyond the resolution offered by these analyses.

Placement of Unassembled *G. Aculeatus* Contigs

In the most recent *G. aculeatus* genome assembly, 26.7 Mb of sequence (including Ns) still remained unassigned to LGs (Glazer et al. 2015). Here, we used the PGA scaffolding data to assign these contigs to LGs. The LGs of the *G. aculeatus* genome were split into their underlying contigs (13 435 previously assigned contigs, total length after removing Ns: 424.9 Mb, median contig length: 10 661 bp, N50 = 87 544 bp) and reassembled using PGA along with the unassigned contigs (3499 previously unassigned contigs, total length after removing Ns: 21.7 Mb, median contig length: 3076 bp, N50 = 10 954 bp). Accuracy was slightly reduced during the clustering step when including the previously unassigned contigs, compared with the assembly that was generated by splitting only the sequence assigned to LGs into 50 kb bins (Figure 1). In this second assembly, 11 927 contigs from the assigned portion of the genome were correctly clustered by LG (88.78% of 13 435), 1381 assigned contigs did not cluster at all with LGs (10.28% of 13 435), and 127 assigned contigs were clustered incorrectly to a LG (0.94% of 13 435) (Figure 3). Of the previously unassigned contigs, 2015 (57.59% of 3499) clustered with LGs. During the ordering step, 1604 of the 2015 were scaffolded by PGA (45.84% of the 3499 previously unassigned contig count). However, 216 of these 1604 contigs could not be assigned to a single LG and were not considered further. The remaining contigs were split into 2 groups based upon the confidence of their placement within a LG (see Methods). One hundred twenty-five contigs from the previously unassigned contigs were unambiguously placed in gaps between sequential contigs in the revised genome assembly. This resulted in an additional 1.1 Mb of sequence (5.1% of the total previously unassigned length) scaffolded into the *G. aculeatus* genome assembly. The remaining 1263 previously unassigned contigs (12.25 Mb, 56.4% of the total unassembled length) were mapped to regions of LGs (median range: 832.8 kb; max range: 33.6 Mb; min range: 8958 bp), but could not be assigned to specific gaps in the genome assembly (Table 2).

To refine the chromosomal regions of the contigs not placed into specific gaps, we used long-distance mate pair information from the Paxton Lake benthic BAC end sequences (Kingsley et al. 2004; Kingsley and Peichel 2007) to identify connections with contigs within the LG. We identified BACs where one end of a BAC insert aligned to an unscaffolded contig, while the other end aligned to a contig within the LG assigned by the PGA scaffolding. Identifying such linkage associations allowed us to narrow the location of many unscaffolded contigs to approximately 148 kb, the average insert size of the BAC library. Of the 1263 previously unassigned contigs assigned to LGs by PGA scaffolding, 229 had alignments with BAC ends. Of these contigs, 195 (2.9 Mb, 23.6% of the sequence length) had BAC end alignments that matched the PGA LG associations (85.2%), confirming that the PGA scaffolding can accurately localize segments of the genome that are challenging to assemble through traditional methods. A revised genome assembly (Gac-HiC) is provided as supplemental data, with 125 new contigs placed into gaps in the assembled genome and 1263 of the previously unassigned contigs narrowed to LGs (available from Dryad Digital Repository).

Scaffolding with BioNano Irys Optical Maps

We also used the BioNano Irys system (San Diego, CA) to generate optical maps of the Paxton Lake benthic population genome

Table 1. Misorderings in the PGA scaffolding often overlap with BAC ends that align discordantly to the reference genome

LG	PGA misorderings (*N*)	Discordant BAC ends (*N*)	Overlap (*N*)	*P* value
I	42	66	6 (14%)	0.123
II	4	17	3 (75%)	<0.001
III	4	8	0 (0%)	—
IV	10	41	0 (0%)	—
V	6	15	1 (17%)	0.124
VI	2	8	1 (50%)	0.015
VII	18	40	2 (11%)	0.231
VIII	6	13	0 (0%)	—
IX	15	22	1 (7%)	0.215
X	25	55	6 (24%)	0.038
XI	12	22	1 (8%)	0.302
XII	22	33	3 (14%)	0.088
XIII	0	46	0 (0%)	—
XIV	7	4	1 (14%)	0.029
XV	2	2	0 (0%)	—
XVI	8	34	2 (25%)	0.024
XVII	8	8	1 (13%)	0.147
XVIII	6	15	2 (33%)	0.031
XIX	19	61	1 (5%)	0.231
XX	6	8	1 (17%)	0.062
XXI	40	25	7 (18%)	<0.001
Total	262	543	39 (15%)	<0.001

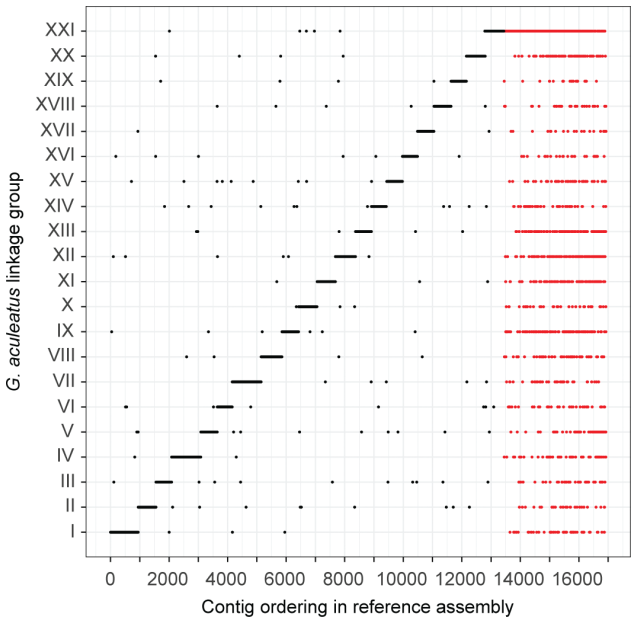

Figure 3. PGA clusters a large proportion of previously unassigned contigs to LGs. The revised *Gasterosteus aculeatus* genome assembly (Glazer et al. 2015) was split into contigs at gaps and clustered with PGA along with 21.7 Mb of contigs previously unassigned to LGs in the *G. aculeatus* genome. The previously unassigned contigs (red vertical band) are distributed across LGs by the PGA clustering. PGA is less accurate clustering the *G. aculeatus* genome when these previously unassigned contigs are included, shown by an increased number of contigs being incorrectly assigned to different LGs (127 incorrectly assigned contigs, 0.94% of the previously assembled contig count).

Table 2. Distribution of contigs scaffolded by PGA that were previously unassigned in the *Gasterosteus aculeatus* genome

LG	Contigs placed into gaps (*N*)	Contigs placed into gaps (total length, bp)	Contigs narrowed to region (*N*)	Contigs narrowed to region (total length, bp)
I	10	93 296	45	433 601
II	6	48 648	11	135 728
III	1	8144	23	216 619
IV	4	92 707	54	455 301
V	7	34 105	20	108 253
VI	1	8060	20	191 189
VII	6	26 377	39	487 407
VIII	8	68 459	46	475 569
IX	31	291 837	109	909 866
X	1	1635	34	411 527
XI	7	41 888	42	333 362
XII	6	46 023	124	965 250
XIII	6	49 046	81	719 633
XIV	1	8315	51	527 141
XV	5	64 865	24	180 933
XVI	7	45 332	17	145 214
XVII	0	—	15	111 395
XVIII	0	—	30	398 982
XIX	4	33 136	10	66 498
XX	4	24 291	39	370 734
XXI	10	123 249	429	4 607 221
Total	125	1 109 413	1263	12 251 423

to verify the PGA scaffolding and to help refine location estimates of the previously unassigned contigs. The BioNano optical map was composed of 615 total contigs (N50 = 1.35 Mb) with a total map length of 569.7 Mb. We split the *G. aculeatus* revised genome assembly (Glazer et al. 2015) into 4377 consecutive 100 kb bins (the minimum recommended contig length for BioNano assemblies) for rescaffolding with the BioNano optical map contigs. Using the default filtering parameters and alignment thresholds, the automated scaffolding pipeline (Shelton et al. 2015) was unable to join many contigs into scaffolds. The N50 remained at 100 kb, assembling 150 of the 4377 contigs into 52 scaffolds. To improve scaffolding, we reduced the filtering thresholds of the scaffolding software by half. This increased the N50 of the assembly from 100 to 796 kb, incorporating 2293 of the 4377 contigs into 460 scaffolds; however, this also increased the number of misassembled scaffolds. Seventy-eight of the 460 scaffolds (19.0%) contained contigs from more than one LG. In addition to scaffolding, we aimed to use the BioNano optical maps to narrow the range estimates of the 1263 previously unassigned contigs within their PGA assigned LGs, but this was not possible because only 2 of the 1263 previously unassigned contigs were over the 100 kb minimum length required for scaffolding with BioNano optical maps.

Discussion

Hi-C-based PGA was able to accurately re-scaffold the *G. aculeatus* revised genome assembly from a set of small contigs into full LGs with internal ordering that closely matched the reference genome. Our results indicate PGA is highly effective at scaffolding relatively short contigs together into a contiguous assembly. Illumina short-read sequences are widely used to construct de novo genome assemblies in nonmodel organisms (reviewed in Ekblom and Wolf 2014). But, short sequencing reads typically cannot span highly repetitive segments of genomes (Treangen and Salzberg 2011; Gordon et al. 2016). This limits the length of contigs that can be built from short-read technologies alone, often with contig N50 sizes of 10–50 kb (Ekblom and Wolf 2014). To assemble contigs into larger scaffolds, many genome assemblies incorporate long range information from a variety of sources, including BAC and fosmid libraries (Myers et al. 2000; Salzberg et al. 2012), jump libraries (Salzberg et al. 2012; Nagarajan and Pop 2013), optical mapping (Shelton et al. 2015; Zhang et al. 2012; Dong et al. 2013), genetic linkage maps (Fierst 2015), and single-molecule real-time sequencing (Gordon et al. 2016; Shi et al. 2016; Bickhart et al. 2017). However, application of these technologies can often be limited by cost, the ability to perform crosses, and the availability of material. For example, optical mapping and BAC library construction require isolation of high molecular weight DNA (Kingsley et al. 2004; Teague et al. 2010; Shelton et al. 2015), which is not possible in many situations. Some technologies, like BioNano Irys optical mapping, also require a minimum contig length (Shelton et al. 2015) for scaffolding that is not typically achievable with short-read Illumina sequencing alone. For example, even with 100 kb contigs, we were not able to re-scaffold the *G. aculeatus* reference genome to the completeness observed with PGA scaffolding. Our results, therefore, offer a promising example of constructing a nearly complete genome assembly de novo using only short-read technologies paired with PGA scaffolding.

Several unmapped contigs from the *G. aculeatus* reference assembly were either placed into specific gaps or localized to regions within LGs in the re-scaffolded genome. Among LGs, there was an overabundance of previously unassigned contigs that were assigned to LG XXI. A similar excess of previously unassigned contigs was placed on LG XXI in the revised reference assembly (Glazer et al. 2015). Among LGs in the Glazer et al. (2015) assembly, LG XXI had the greatest relative increase in length (1.48-fold increase in length versus an average 1.09-fold increase across the remainder of the LGs). Combined, our Hi-C reference assembly (Gac-HiC) and the revised reference assembly from high-density linkage maps (Glazer et al. 2015) indicate LG XXI was the least complete LG in the original reference assembly (Jones et al. 2012).

Within LGs, we identified several discordant orderings between the Hi-C assembly and the *G. aculeatus* reference genome. Although some of these are likely due to errors in the PGA scaffolding (from errors in the assembly algorithm or regional differences in chromatin interactions), many of the misorderings in the Hi-C assembly matched discordant mate-pair alignments in a BAC library from the same population of *G. aculeatus* used for the PGA scaffolding, suggesting structural variation among populations. Three population-specific inversions on LG I, LG XI, and LG XXI have previously been identified in sticklebacks (Jones et al. 2012). These inversions were not identified by the PGA scaffolding or by aligning the BAC ends to the reference genome. However, these inversions are polymorphisms present between freshwater and marine populations and would not be expected in our comparison between 2 freshwater populations (Paxton Lake benthic and Bear Paw Lake). Future work will focus on identifying the nature of the discordant orderings in the PGA assembly, and whether they reflect errors in the reference assembly or structural polymorphisms between the Paxton Lake benthic and Bear Paw Lake populations. The results presented here suggest that PGA scaffolding may be a useful method to identify errors in reference assemblies or structural variation across genomes.

Supplementary Material

Supplementary data are found at *Journal of Heredity* online.

Funding

This work was supported by an Evolutionary, Ecological, or Conservation Genomics Research Award from the American Genetic Association to M.A.W.; the Office of the Vice President of Research at the University of Georgia to M.A.W.; the National Institutes of Health (R01 GM116853 to C.L.P.); and the Fred Hutchinson Cancer Research Center Division of Basic Sciences to C.L.P.

Acknowledgments

We thank Chris Amemiya for isolating high molecular weight DNA for use in optical mapping, and Susan Brown and the Kansas State University Bioinformatics Center for performing the optical mapping. All procedures were approved by the Fred Hutchinson Cancer Research Center Institutional Animal Care and Use Committee (protocol 1575).

Conflict of Interest

S.T.S. and I.L. are employees of Phase Genomics. M.A.W. and C.L.P. declare no competing interests.

Data Availability

We have deposited the primary data underlying these analyses as follows: Hi-C sequences are deposited in the NCBI SRA database: SRP081031. BioNano XMAP optical map files and revised genome assembly are available at Dryad (doi: 10.5061/dryad.h7h32).

References

Bell MA, Foster SA. 1994. *The evolutionary biology of the threespine stickleback*. Oxford, UK: Oxford University Press.

Bickhart DM, Rosen BD, Koren S, Sayre BL, Hastie AR, Chan S, Lee J, Lam ET, Liachko I, Sullivan ST, *et al*. 2017. Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nat Genet*. 49:643–650.

Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, Shendure J. 2013. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat Biotechnol*. 31:1119–1125.

Das SK, Austin MD, Akana MC, Deshpande P, Cao H, Xiao M. 2010. Single molecule linear analysis of DNA in nano-channel labeled with sequence specific fluorescent probes. *Nucleic Acids Res*. 38:e177.

Dong Y, Xie M, Jiang Y, Xiao N, Du X, Zhang W, Tosser-Klopp G, Wang J, Yang S, Liang J, *et al*. 2013. Sequencing and automated whole-genome optical mapping of the genome of a domestic goat (*Capra hircus*). *Nat Biotechnol*. 31:135–141.

Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, Shamim MS, Machol I, Lander ES, Aiden AP, *et al*. 2017. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science*. 356:92–95.

Ekblom R, Wolf JB. 2014. A field guide to whole-genome sequencing, assembly and annotation. *Evol Appl*. 7:1026–1042.

Fierst JL. 2015. Using linkage maps to correct and scaffold de novo genome assemblies: methods, challenges, and computational tools. *Front Genet*. 6:220.

Glazer AM, Killingbeck EE, Mitros T, Rokhsar DS, Miller CT. 2015. Genome assembly improvement and mapping convergently evolved skeletal traits in sticklebacks with genotyping-by-sequencing. *G3 (Bethesda)*. 5:1463–1472.

Gordon D, Huddleston J, Chaisson MJ, Hill CM, Kronenberg ZN, Munson KM, Malig M, Raja A, Fiddes I, Hillier LW, *et al*. 2016. Long-read sequence assembly of the gorilla genome. *Science*. 352:aae0344.

Hendry AP, Peichel CL, Matthews B, Boughman JW, Nosil P. 2013. Stickleback research: the now and the next. *Ecol Res*. 15:111–141.

Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, Swofford R, Pirun M, Zody MC, White S, *et al*.; Broad Institute Genome Sequencing Platform & Whole Genome Assembly Team. 2012. The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*. 484:55–61.

Kaplan N, Dekker J. 2013. High-throughput genome scaffolding from in vivo DNA interaction frequency. *Nat Biotechnol*. 31:1143–1147.

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

Kingsley DM, Peichel CL. 2007. The molecular genetics of evolutionary change in sticklebacks. In: Östlund-Nilsson S, Mayer I, Huntingford F, editors. *Biology of the three-spined sticklebacks*. Boca Raton (FL): CRC Press. p. 44–81.

Kingsley DM, Zhu B, Osoegawa K, De Jong PJ, Schein J, Marra M, Peichel CL, Amemiya C, Schluter D, Balabhadra S, *et al*. 2004. New genomic tools for molecular studies of evolutionary change in threespine sticklebacks. *Behaviour*. 141:1331–1344.

Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, *et al*. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*. 326:289–293.

Marie-Nelly H, Marbouty M, Cournac A, Flot JF, Liti G, Parodi DP, Syan S, Guillén N, Margeot A, Zimmer C, *et al*. 2014. High-quality genome (re) assembly using chromosomal contact data. *Nat Commun*. 5:5695.

Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, Kravitz SA, Mobarry CM, Reinert KH, Remington KA, *et al*. 2000. A whole-genome assembly of Drosophila. *Science*. 287:2196–2204.

Nagarajan N, Pop M. 2013. Sequence assembly demystified. *Nat Rev Genet*. 14:157–167.

Östlund-Nilsson S, Mayer I, Huntingford F. 2007. *Biology of the three-spined stickleback*. Boca Raton (FL): CRC Press.

Peichel CL, Marques DA. 2017. The genetic and molecular architecture of phenotypic diversity in sticklebacks. *Philos Trans R Soc Lond B Biol Sci*. 372:20150486.

Putnam NH, O'Connell BL, Stites JC, Rice BJ, Blanchette M, Calef R, Troll CJ, Fields A, Hartley PD, Sugnet CW, *et al*. 2016. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res*. 26:342–350.

Roesti M, Moser D, Berner D. 2013. Recombination in the threespine stickleback genome–patterns and consequences. *Mol Ecol*. 22:3014–3027.

Ross JA, Peichel CL. 2008. Molecular cytogenetic evidence of rearrangements on the Y chromosome of the threespine stickleback fish. *Genetics*. 179:2173–2182.

Salzberg SL, Phillippy AM, Zimin A, Puiu D, Magoc T, Koren S, Treangen TJ, Schatz MC, Delcher AL, Roberts M, *et al*. 2012. GAGE: a critical evaluation of genome assemblies and assembly algorithms. *Genome Res*. 22:557–567.

Shelton JM, Coleman MC, Herndon N, Lu N, Lam ET, Anantharaman T, Sheth P, Brown SJ. 2015. Tools and pipelines for BioNano data: molecule assembly pipeline and FASTA super scaffolding tool. *BMC Genomics*. 16:734.

Shi L, Guo Y, Dong C, Huddleston J, Yang H, Han X, Fu A, Li Q, Li N, Gong S, *et al*. 2016. Long-read sequencing and de novo assembly of a Chinese genome. *Nat Commun*. 7:12065.

Teague B, Waterman MS, Goldstein S, Potamousis K, Zhou S, Reslewic S, Sarkar D, Valouev A, Churas C, Kidd JM, *et al*. 2010. High-resolution human genome structure by single-molecule analysis. *Proc Natl Acad Sci USA*. 107:10848–10853.

Treangen TJ, Salzberg SL. 2011. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet*. 13:36–46.

Wootton RJ. 1976. *The biology of sticklebacks*. London (UK): Academic Press.

Zhang Q, Chen W, Sun L, Zhao F, Huang B, Yang W, Tao Y, Wang J, Yuan Z, Fan G, *et al*. 2012. The genome of *Prunus mume*. *Nat Commun*. 3:1318.