

1
2
3 **1 Improvement of the threespine stickleback genome using a Hi-C-based Proximity-**
4 **2 Guided Assembly**
5

6
7 Catherine L. Peichel^{1*}, Shawn T. Sullivan², Ivan Liachko³, Michael A. White^{4†}
8

9 ¹*Divisions of Basic Sciences and Human Biology, Fred Hutchinson Cancer Research Center,*
10 *Seattle, Washington 98109, USA, catherine.peichel@jee.unibe.ch;* ²*Phase Genomics, Inc.,*
11 *4000 Mason Road, Suite 225, Seattle, Washington 98185, USA, shawn@phasegenomics.com;*
12 ³*Phase Genomics, Inc., 4000 Mason Road, Suite 225, Seattle, Washington 98185, USA,*
13 *ivan@phasegenomics.com;* ⁴*Department of Genetics, University of Georgia, Athens, Georgia*
14 *30602, USA, whitem@uga.edu*
15

16
17 †Author for correspondence:
18 Michael A. White
19 Department of Genetics
20 University of Georgia
21 120 Green St.
22 Athens, GA 30602
23 706-542-2464
24 whitem@uga.edu
25
26
27

28
29 *Current address: Institute of Ecology and Evolution, University of Bern, 3012 Bern, Switzerland
30

31
32 Running title: Hi-C genome assembly in threespine stickleback
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 28 **Abstract**
4 29

5 30 Scaffolding genomes into complete chromosome assemblies remains challenging even with the
6 31 rapidly increasing sequence coverage generated by current next-generation sequence
7 32 technologies. Even with scaffolding information, many genome assemblies remain incomplete.
8 33 The genome of the threespine stickleback (*Gasterosteus aculeatus*), a fish model system in
9 34 evolutionary genetics and genomics, is not completely assembled despite scaffolding with high-
10 35 density linkage maps. Here, we first test the ability of a Hi-C based Proximity-Guided Assembly
11 36 to perform a *de novo* genome assembly from relatively short contigs. Using Hi-C based
12 37 Proximity-Guided Assembly, we generated complete chromosome assemblies from a
13 38 distribution of short contigs (20-100 kb). We found that 96.40% of contigs were correctly
14 39 assigned to linkage groups, with ordering nearly identical to the previous genome assembly.
15 40 Using available bacterial artificial chromosome (BAC) end sequences, we provide evidence that
16 41 some of the few discrepancies between the Hi-C assembly and the existing assembly are due to
17 42 structural variation between the populations used for the two assemblies or errors in the existing
18 43 assembly. This Hi-C assembly also allowed us to improve the existing assembly, assigning over
19 44 60% (13.35 Mb) of the previously unassigned (~21.7 Mb) contigs to linkage groups. Together,
20 45 our results highlight the potential of the Hi-C based Proximity-Guided Assembly method to be
21 46 used in combination with short read data to perform relatively inexpensive *de novo* genome
22 47 assemblies. This approach will be particularly useful in organisms in which it is difficult to
23 48 perform linkage mapping or to obtain high molecular weight DNA required for other scaffolding
24 49 methods.
25 50

26 51 Keywords: *de novo* genome assembly, chromosome conformation capture, *Gasterosteus*
27 52 *aculeatus*
28 53

54 Introduction

55 While short-read genome sequencing has become a staple in genetic research, scaffolding
56 complete eukaryotic genomes from fragmented assemblies remains a remarkably difficult task
57 as most modern scaffolding techniques utilize purified high-molecular weight DNA as the source
58 of contiguity information (Putnam et al. 2016; Teague et al. 2010; Das et al. 2010). Purification
59 results in broken DNA molecules and loss of long-range intra-chromosomal genetic contiguity
60 information typically yielding incomplete scaffolds. One traditional method for retaining
61 chromosome-scale contiguity is to use genetic crosses to establish maps of relative linkage
62 distances between sequences. However, genetic mapping is very laborious, cannot be applied
63 to many organisms, and often falls short of scaffolding all contigs due to low resolution from a
64 limited number of crossovers (reviewed in Fierst 2015). Chromosome conformation capture
65 techniques such as Hi-C (Lieberman-Aiden et al. 2009) retain ultra-long-range genomic
66 contiguity information through *in vivo* crosslinking of chromatin and subsequent sequencing of
67 proximal pairs of sequences. The rate of Hi-C interaction decreases rapidly with increasing
68 genomic distance between pairs of loci. Taking advantage of this relationship between inter-
69 sequence distance and proximity interaction allows the construction of chromosome-scale
70 genome scaffolds (Burton et al. 2013; Marie-Nelly et al. 2014; Kaplan and Dekker 2013;
71 Bickhart et al. 2017; Dudchenko et al. 2017).

72 Here, we used the Hi-C-based Proximity-Guided Assembly (PGA) method to assemble
73 the genome of the threespine stickleback (*Gasterosteus aculeatus*). This small, teleost fish is a
74 widely used model system in diverse fields, including ecology, evolution, behavior, physiology
75 and toxicology (Wootton 1976; Bell and Foster 1994; Östlund-Nilsson et al. 2007). Sticklebacks
76 are well known for the extensive morphological, behavioral and physiological variation present
77 in freshwater populations that have evolved since the retreat of the glaciers across the Northern
78 hemisphere in the past 15,000 years (Bell and Foster 1994; Hendry et al. 2013). Recent
79 research has led to identification of the genetic and genomic basis of this phenotypic diversity,

1
2
3 80 providing new insights into the genetic basis of adaptation (Peichel and Marques 2017). To
4
5 81 facilitate research in this model system, a high quality genome assembly for *G. aculeatus* was
6
7 82 generated by Sanger sequencing of plasmid, fosmid, and bacterial artificial chromosome (BAC)
8
9 83 genomic libraries made from a single female from Bear Paw Lake, Alaska. Scaffolds were
10
11 84 anchored to the 21 known stickleback chromosomes or linkage groups (LG) using genetic
12
13 85 linkage mapping. The original genome assembly comprised 400.7 Mb of scaffolds anchored to
14
15 86 linkage groups, with an additional 60.7 Mb of assembled scaffolds not anchored to linkage
16
17 87 groups (Jones et al. 2012). Two further revisions to the genome assembly have used genetic
18
19 88 linkage mapping in three additional crosses to assign some of these unanchored scaffolds to
20
21 89 linkage groups and to correct errors in the original assembly (Roesti et al. 2013; Glazer et al.
22
23 90 2015). The most recent genome assembly comprises 436.6 Mb, with 26.7 Mb remaining
24
25 91 unassigned to linkage groups (Glazer et al. 2015). Here, we took advantage of the existence of
26
27 92 a high quality assembly for *G. aculeatus* to test the performance of Hi-C in generating a *de novo*
28
29 93 assembly. Additionally, we used Hi-C to further improve the existing *G. aculeatus* genome
30
31 94 assembly.
32
33
34
35
36
37

38 96 **Methods**

39 97 *Tissue collection and Hi-C sequencing*

40 98 The liver of a single, lab-reared adult male from the Paxton Lake benthic population (Texada
41
42 99 Island, British Columbia) was dissected and flash frozen in liquid nitrogen. Tissue processing,
43
44 100 chromatin isolation, library preparation, and sequencing were all performed by Phase Genomics
45
46 101 (Seattle, WA). A total of 176,461,081 read-pairs were sequenced.
47
48
49
50

51 102 52 103 *PGA scaffolding*

53 104 Scaffolding was conducted in two phases. First, to scaffold the entire *G. aculeatus* revised
54
55 105 genome assembly (Glazer et al. 2015), the genome was divided into 8,342 contiguous contigs
56
57
58
59
60

1
2
3 106 of varied length, excluding contigs that were not previously assigned to linkage groups and the
4
5 107 mitochondria sequence. Gaps present within the revised genome assembly were not removed
6
7 108 prior to dividing it into contigs; the resulting contigs therefore included gaps ranging from 0.00%
8
9 109 to 100.00% of their length, with a median of 0.31% and mean of 2.69% of the length of a contig
10
11 110 being composed of gaps. Contig length followed a normal distribution that ranged from 20 kb to
12
13 111 100 kb (median contig size: 52,339 bp; standard deviation 18,261 bp). Paired-end reads were
14
15 112 aligned to the contigs, only retaining reads that aligned uniquely. Contigs were scaffolded using
16
17 113 PGA with an adapted version of the Lachesis method (Burton et al. 2013) by Phase Genomics.
18
19 114 The known number of *G. aculeatus* chromosomes (21) was used as a starting input parameter
20
21 115 during the scaffolding process (Ross and Peichel 2008). The final set of Lachesis parameters
22
23 116 were selected from randomized parameter sweeps from over 60,000 scaffolding iterations.
24
25 117 Parameters were varied within the following bounds: CLUSTER_N between 1 and 46;
26
27 118 CLUSTER_MIN_RE_SITES between 1 and 4,988; CLUSTER_MAX_LINK_DENSITY between
28
29 119 0.0008 and 28.9; CLUSTER_NONINFORMATIVE_RATIO between 1.0004 and 24.3;
30
31 120 ORDER_MIN_RES_IN_TRUNK between 1 and 5131; and ORDER_MIN_RES_IN_SHREDS
32
33 121 between 1 and 6489. The four best sets of candidate parameters were identified among these
34
35 122 sweeps that best reflected the expected patterns of the Hi-C data and the likelihood of the
36
37 123 resulting scaffolds having generated the observed Hi-C data. The patterns examined were intra-
38
39 124 cluster link density (the ratio of Hi-C linkage contained within scaffolds as opposed to between
40
41 125 them), ordering enrichment (the concentration of observed Hi-C link density between contigs
42
43 126 near each other as compared to a null hypothesis of uniform Hi-C link density), and orientation
44
45 127 quality score (the differential log-likelihood of the chosen orientation of a contig having resulted
46
47 128 in the observed Hi-C data as compared to alternatives) (for additional detail see Burton et al.
48
49 129 2013; Bickhart et al. 2017). The final set of parameters that generated the largest scaffolds were
50
51 130 CLUSTER_N=21, CLUSTER_MIN_RE_SITES=141, CLUSTER_MAX_LINK_DENSITY=1.593,
52
53
54
55
56
57
58
59
60

1
2
3 131 CLUSTER_NONINFORMATIVE_RATIO=5.163, ORDER_MIN_N_RES_IN_TRUNK=69,
4
5 132 ORDER_MIN_N_RES_IN_SHREDS=18.
6

7 133 The second phase of scaffolding used PGA to assign the contigs that were previously
8
9 134 not assigned to linkage groups to gaps in the reference genome. The reference assembly
10
11 135 (excluding the mitochondria sequence) was split into contigs at gaps and Ns were removed, and
12
13 136 all contigs (13,435 contigs previously assigned to linkage groups and 3,499 contigs not
14
15 137 previously assigned to linkage groups) were assembled with PGA. Previously unassigned
16
17 138 contigs that were placed in the PGA assembly were divided into three groups, based on the
18
19 139 level of certainty in their placement. If the contigs assembled before (contig A) and after (contig
20
21 140 B) the previously unassigned contig were sequential (i.e. occurred in the expected order relative
22
23 141 to the reference assembly), the previously unassigned contig was considered an accurate
24
25 142 placement and was inserted in the gap between contig A and contig B. If contig A and contig B
26
27 143 were from the same linkage group, but were not sequential, the previously unassigned contig
28
29 144 could not be accurately placed in the gap and was instead assigned to the linkage group within
30
31 145 a narrowed range of possible locations. If contig A and contig B were from different linkage
32
33 146 groups or if contig A or contig B were missing (i.e. the previously unassigned contig only had
34
35 147 linkage information on one end), the previously unassigned contig could not be placed and was
36
37 148 not considered further (216 total unplaced contigs).
38
39
40
41
42
43

44 150 *Bacterial Artificial Chromosome (BAC)-end alignments*

45
46 151 Sequenced BAC-ends from the CHORI-215 BAC library made from two Paxton Lake benthic
47
48 152 males (Kingsley et al. 2004; Kingsley and Peichel 2007) were aligned to the unmasked *G.*
49
50 153 *aculeatus* revised genome assembly (Glazer et al. 2015) using the BLAST-like alignment tool
51
52 154 (BLAT) (Kent 2002) (67,979 total paired BAC ends). Alignments were only retained if at least
53
54 155 90% of the sequenced BAC-end aligned to the genome. If a BAC-end aligned to multiple
55
56 156 locations in the genome, the highest scoring alignment was kept according to the formula:
57
58
59
60

1
2
3 157 alignment matches + alignment matches that are part of repeats – mismatches in alignment –
4
5 158 number of gap openings in the query sequence – number of gap openings in the target
6
7
8 159 sequence. Alignments were discarded if there were multiple alignments tied for the same
9
10 160 highest alignment score. In addition, alignments were only considered if both BAC-ends aligned
11
12 161 to the same linkage group because we were only focused on identifying putative
13
14 162 intrachromosomal rearrangements.
15

16 163
17
18 164 *Overlap between misorderings in the PGA assembly and BAC-ends that aligned discordantly*
19
20 165 BAC-ends that aligned in a forward/reverse orientation and were separated by over 250 kb in
21
22 166 the genome (the average insert size of the library is 148 kb) were considered putative deletions
23
24 167 in the Paxton Lake benthic population or insertions in the reference assembly. BAC-ends that
25
26 168 aligned in a forward/forward or reverse/reverse orientation in the genome were considered
27
28 169 putative inversions. A contig was considered misordered in the PGA reassembly of the *G.*
29
30 170 *aculeatus* genome if either of the neighboring contigs in the scaffold was located further than
31
32 171 250 kb away from the coordinates in the reference genome assembly (Glazer et al. 2015). This
33
34 172 method defined the end breakpoints of misordered regions within the scaffolds. Overlap was
35
36 173 scored if the position of a discordantly aligned BAC-end fell within the contig that was
37
38 174 misordered in the PGA assembly. Permutations were conducted for each linkage group
39
40 175 separately and for the total genome to test for significance. Random subsets of BAC ends were
41
42 176 drawn from each linkage group equal to the number of discordant BAC-end alignments. Overlap
43
44 177 was scored between the misordered PGA contigs and the random subsets of BAC ends. The p-
45
46 178 value reflects how often the same number of overlaps is recovered among a set of 10,000
47
48 179 random permutations.
49
50

51
52 180
53
54
55 181 *BioNano scaffolding*
56
57
58
59
60

1
2
3 182 High molecular weight DNA was isolated from the blood of a single adult male from the Paxton
4
5 183 Lake benthic population (Texada Island, British Columbia) following protocols outlined in
6
7 184 (Kingsley et al. 2004). This male was not the same individual used for Hi-C, nor for creation of
8
9 185 the BAC libraries. Irys optical mapping (BioNano Genomics, San Diego, CA) was performed at
10
11 186 Kansas State University. DNA was nicked with the BspQI restriction enzyme, which cuts at a
12
13 187 frequency of 15.8 sites per 100 kb across the *G. aculeatus* genome, which is around the ideal
14
15 188 cutting frequency of 10-15 sites / 100 kb for optical mapping with the BioNano Irys System
16
17 189 (Shelton et al. 2015). DNA was labeled with fluorescent nucleotides and repaired according to
18
19 190 BioNano protocols. DNA was imaged on the BioNano Irys System using two IrysChips. BioNano
20
21 191 molecules were filtered to only include segments that were at least 150 kb and contained at
22
23 192 least eight labels. The P-value threshold for the BioNano assembler was set to a minimum of
24
25 193 2.2×10^{-9} . Molecule stretch was adjusted using AssembleIrysCluster.pl (v. 1.6.1) (Shelton et al.
26
27 194 2015). To assess the effectiveness of BioNano optical maps in resc scaffolding the *G. aculeatus*
28
29 195 genome, the revised genome assembly was split into contiguous 100 kb contigs (the minimum
30
31 196 recommended size for BioNano assembly). The split genome was digested with BspQI into *in*
32
33 197 *silico* CMAP files using fa2cmap_multi.pl (BioNano) and iteratively scaffolded with the BioNano
34
35 198 optical maps using sewing_machine.pl (v. 1.0.6) (Shelton et al. 2015). Two different filtering
36
37 199 options were used: the default filters (--f_con 20, --f_algn 40, --s_con 15, --s_algn 90) and
38
39 200 relaxed filters set at half of the default thresholds (--f_con 10, --f_algn 20, --s_con 7.5, --s_algn
40
41 201 45). For both sets of filters, the default alignment parameters were used (-FP 0.8, -FN 0.08 -sf
42
43 202 0.20 -sd 0.10).
44
45
46
47
48
49
50

204 Results

205 *Hi-C/PGA resc scaffolding of the G. aculeatus genome assembly*

206 To investigate how well Hi-C-based PGA (provided by Phase Genomics, Seattle, WA) can
207 assemble a genome composed of small contigs, we split the revised *G. aculeatus* genome

1
2
3 208 assembly (Glazer et al. 2015) into contigs of varied length, ranging from 20 kb to 100 kb and
4
5 209 used a proximity-guided assembly to rescaffold the contigs together. PGA reconstructed a
6
7 210 highly accurate genome assembly, with 8,042 of the 8,342 (96.40% of contigs; 97.15% of the
8
9 211 genome length) of the contigs were correctly assigned to one of the 21 linkage groups in the *G.*
10
11 212 *aculeatus* genome during clustering (1 contig was incorrectly assigned to an alternate linkage
12
13 213 group, 2 contigs were not aligned within the cluster, and 297 contigs were not assigned to a
14
15 214 linkage group) (Figure 1). Among the linkage groups, most of the contigs (7,404 contigs, or
16
17 215 92.06% of the 8,042 contigs correctly assigned to linkage groups) had an ordering identical to
18
19 216 the revised *G. aculeatus* reference assembly (Figure 2; Figure S1). The 638 contigs (34.33 Mb
20
21 217 of the 436.6 Mb total genome length) that were ordered incorrectly within linkage groups may
22
23 218 represent errors in the PGA scaffolding, assembly errors in the reference genome, or could
24
25 219 reflect structural variation between the Paxton Lake (British Columbia) benthic population used
26
27 220 for the PGA scaffolding and the Bear Paw Lake (Alaska) population used for the reference
28
29 221 assembly.
30
31
32

33 222 We explored whether the incorrect orderings within linkage groups were due to errors in
34
35 223 PGA scaffolding or structural variation using the BAC-end sequences available from a BAC
36
37 224 library made from two Paxton Lake benthic males (Kingsley et al. 2004; Kingsley and Peichel
38
39 225 2007). We aligned the BAC-end sequences to the revised *G. aculeatus* genome assembly
40
41 226 (Glazer et al. 2015) and scanned for discordant mate-pair alignments (i.e. mate-pairs that
42
43 227 aligned in the same orientation or that aligned across genomic regions larger than 250 kb, which
44
45 228 is larger than the 148 kb average insert size of the library). Such discordant alignments would
46
47 229 indicate large structural differences between the Paxton Lake benthic population and the Bear
48
49 230 Paw Lake reference assembly population, or errors in the Bear Paw reference genome
50
51 231 assembly (Table S1). In many linkage groups, we found that discordantly aligned BAC mate
52
53 232 pairs were significantly more often associated with the contig at either end of the discordant
54
55 233 orderings in the PGA scaffolding (Figure 2; Figure S1; Table 1). This indicates that at least
56
57
58
59
60

1
2
3 234 some of the PGA scaffold misorderings may reflect true insertions, deletions, or inversions
4
5 235 between populations of *G. aculeatus*, and/or errors in the reference assembly. Full sequencing
6
7 236 of these BAC clones will allow the breakpoints to be fine-mapped beyond the resolution offered
8
9 237 by these analyses.

10 238

14 239 *Placement of unassembled G. aculeatus contigs*

16 240 In the most recent *G. aculeatus* genome assembly, 26.7 Mb of sequence (including Ns) still
17
18 241 remained unassigned to linkage groups (Glazer et al. 2015). Here, we used the PGA scaffolding
19
20 242 data to assign these contigs to linkage groups. The linkage groups of the *G. aculeatus* genome
21
22 243 were split into their underlying contigs (13,435 previously assigned contigs, total length after
23
24 244 removing Ns: 424.9 Mb, median contig length: 10,661 bp, N50=87,544 bp) and reassembled
25
26 245 using PGA along with the unassigned contigs (3,499 previously unassigned contigs, total length
27
28 246 after removing Ns: 21.7 Mb, median contig length: 3,076 bp, N50=10,954 bp). Accuracy was
29
30 247 slightly reduced during the clustering step when including the previously unassigned contigs,
31
32 248 compared to the assembly that was generated by splitting only the sequence assigned to
33
34 249 linkage groups into 50 kb bins (Figure 1). In this second assembly, 11,927 contigs from the
35
36 250 assigned portion of the genome were correctly clustered by linkage group (88.78% of 13,435),
37
38 251 1,381 assigned contigs did not cluster at all with linkage groups (10.28% of 13,435), and 127
39
40 252 assigned contigs were clustered incorrectly to a linkage group (0.94% of 13,435) (Figure 3). Of
41
42 253 the previously unassigned contigs, 2,015 (57.59% of 3,499) clustered with linkage groups.
43
44 254 During the ordering step, 1,604 of the 2,015 were scaffolded by PGA (45.84% of the 3,499
45
46 255 previously unassigned contig count). However, 216 of these 1,604 contigs could not be
47
48 256 assigned to a single linkage group and were not considered further. The remaining contigs were
49
50 257 split into two groups based upon the confidence of their placement within a linkage group (see
51
52 258 methods). 125 contigs from the previously unassigned contigs were unambiguously placed in
53
54 259 gaps between sequential contigs in the revised genome assembly. This resulted in an additional
55
56
57
58
59
60

1
2
3 260 1.1 Mb of sequence (5.1% of the total previously unassigned length) scaffolded into the *G.*
4
5 261 *aculeatus* genome assembly. The remaining 1,263 previously unassigned contigs (12.25 Mb,
6
7 262 56.4% of the total unassembled length) were mapped to regions of linkage groups (median
8
9 263 range: 832.8 kb; max range: 33.6 Mb; min range: 8,958 bp), but could not be assigned to
10
11 264 specific gaps in the genome assembly (Table 2).

12
13
14 265 To refine the chromosomal regions of the contigs not placed into specific gaps, we used
15
16 266 long-distance mate pair information from the Paxton Lake benthic BAC-end sequences
17
18 267 (Kingsley et al. 2004; Kingsley and Peichel 2007) to identify connections with contigs within the
19
20 268 linkage group. We identified BACs where one end of a BAC insert aligned to an unscaffolded
21
22 269 contig, while the other end aligned to a contig within the linkage group assigned by the PGA
23
24 270 scaffolding. Identifying such linkage associations allowed us to narrow the location of many
25
26 271 unscaffolded contigs to approximately 148 kb, the average insert size of the BAC library. Of the
27
28 272 1,263 previously unassigned contigs assigned to linkage groups by PGA scaffolding, 229 had
29
30 273 alignments with BAC-ends. Of these contigs, 195 (2.9 Mb, 23.6% of the sequence length) had
31
32 274 BAC-end alignments that matched the PGA linkage group associations (85.2%), confirming that
33
34 275 the PGA scaffolding can accurately localize segments of the genome that are challenging to
35
36 276 assemble through traditional methods. A revised genome assembly (Gac-HiC) is provided as
37
38 277 supplemental data, with 125 new contigs placed into gaps in the assembled genome and 1,263
39
40 278 of the previously unassigned contigs narrowed to linkage groups (available from Dryad Digital
41
42 279 Repository).

43
44
45
46 280

47 48 281 *Scaffolding with BioNano Irys optical maps*

49
50 282 We also used the BioNano Irys system (San Diego, CA) to generate optical maps of the Paxton
51
52 283 Lake benthic population genome in order to verify the PGA scaffolding and to help refine
53
54 284 location estimates of the previously unassigned contigs. The BioNano optical map was
55
56 285 composed of 615 total contigs (N50=1.35 Mb) with a total map length of 569.7 Mb. We split the
57
58
59
60

1
2
3 286 *G. aculeatus* revised genome assembly (Glazer et al. 2015) into 4,377 consecutive 100 kb bins
4
5 287 (the minimum recommended contig length for BioNano assemblies) for rescaffolding with the
6
7 288 BioNano optical map contigs. Using the default filtering parameters and alignment thresholds,
8
9 289 the automated scaffolding pipeline (Shelton et al. 2015) was unable to join many contigs into
10
11 290 scaffolds. The N50 remained at 100 kb, assembling 150 of the 4,377 contigs into 52 scaffolds.
12
13 291 To improve scaffolding, we reduced the filtering thresholds of the scaffolding software by half.
14
15 292 This increased the N50 of the assembly from 100 kb to 796 kb, incorporating 2,293 of the 4,377
16
17 293 contigs into 460 scaffolds; however, this also increased the number of misassembled scaffolds.
18
19 294 78 of the 460 scaffolds (19.0%) contained contigs from more than one linkage group. In addition
20
21 295 to scaffolding, we aimed to use the BioNano optical maps to narrow the range estimates of the
22
23 296 1,263 previously unassigned contigs within their PGA assigned linkage groups, but this was not
24
25 297 possible because only two of the 1,263 previously unassigned contigs were over the 100 kb
26
27 298 minimum length required for scaffolding with BioNano optical maps.
28
29
30
31
32

33 300 **Discussion**

34
35 301 Hi-C-based PGA was able to accurately re-scaffold the *G. aculeatus* revised genome assembly
36
37 302 from a set of small contigs into full linkage groups with internal ordering that closely matched the
38
39 303 reference genome. Our results indicate PGA is highly effective at scaffolding relatively short
40
41 304 contigs together into a contiguous assembly. Illumina short-read sequences are widely used to
42
43 305 construct *de novo* genome assemblies in non-model organisms (reviewed in Ekblom and Wolf
44
45 306 2014). But, short sequencing reads typically cannot span highly repetitive segments of genomes
46
47 307 (Gordon et al. 2016; Treangen and Salzberg 2012). This limits the length of contigs that can be
48
49 308 built from short-read technologies alone, often with contig N50 sizes of 10-50 kb (Ekblom and
50
51 309 Wolf 2014). To assemble contigs into larger scaffolds, many genome assemblies incorporate
52
53 310 long range information from a variety of sources, including BAC and fosmid libraries (Myers et
54
55 311 al. 2000; Salzberg et al. 2012), jump libraries (Salzberg et al. 2012; Nagarajan and Pop 2013),
56
57
58
59
60

1
2
3 312 optical mapping (Shelton et al. 2015; Zhang et al. 2012; Dong et al. 2013), genetic linkage maps
4
5 313 (Fierst 2015), and single-molecule real-time sequencing (Gordon et al. 2016; Shi et al. 2016;
6
7 314 Bickhart et al. 2017). However, application of these technologies can often be limited by cost,
8
9 315 the ability to perform crosses, and the availability of material. For example, optical mapping and
10
11 316 BAC library construction require isolation of high molecular weight DNA (Shelton et al. 2015;
12
13 317 Teague et al. 2010; Kingsley et al. 2004), which is not possible in many situations. Some
14
15 318 technologies, like BioNano lrys optical mapping, also require a minimum contig length (Shelton
16
17 319 et al. 2015) for scaffolding that is not typically achievable with short-read Illumina sequencing
18
19 320 alone. For example, even with 100 kb contigs, we were not able to re-scaffold the *G. aculeatus*
20
21 321 reference genome to the completeness observed with PGA scaffolding. Our results therefore
22
23 322 offer a promising example of constructing a nearly-complete genome assembly *de novo* using
24
25 323 only short-read technologies paired with PGA scaffolding.
26
27
28

29 324 Several unmapped contigs from the *G. aculeatus* reference assembly were either placed
30
31 325 into specific gaps or localized to regions within linkage groups in the re-scaffolded genome.
32
33 326 Among linkage groups, there was an overabundance of previously unassigned contigs that were
34
35 327 assigned to LG XXI. A similar excess of previously unassigned contigs was placed on LG XXI in
36
37 328 the revised reference assembly (Glazer et al. 2015). Among linkage groups in the Glazer et al.
38
39 329 (2015) assembly, LG XXI had the greatest relative increase in length (1.48 fold increase in
40
41 330 length versus an average 1.09 fold increase across the remainder of the linkage groups).
42
43 331 Combined, our Hi-C reference assembly (Gac-HiC) and the revised reference assembly from
44
45 332 high-density linkage maps (Glazer et al. 2015) indicate LG XXI was the least complete linkage
46
47 333 group in the original reference assembly (Jones et al. 2012).
48
49
50

51 334 Within linkage groups, we identified several discordant orderings between the Hi-C
52
53 335 assembly and the *G. aculeatus* reference genome. Although some of these are likely due to
54
55 336 errors in the PGA scaffolding (from errors in the assembly algorithm or regional differences in
56
57 337 chromatin interactions), many of the misorderings in the Hi-C assembly matched discordant
58
59
60

1
2
3 338 mate-pair alignments in a BAC library from the same population of *G. aculeatus* used for the
4
5 339 PGA scaffolding, suggesting structural variation among populations. Three population-specific
6
7 340 inversions on LG I, LG XI, and LG XXI have previously been identified in sticklebacks (Jones et
8
9 341 al. 2012). These inversions were not identified by the PGA scaffolding or by aligning the BAC-
10
11 342 ends to the reference genome. However, these inversions are polymorphisms present between
12
13 343 freshwater and marine populations and would not be expected in our comparison between two
14
15 344 freshwater populations (Paxton Lake benthic and Bear Paw Lake). Future work will focus on
16
17 345 identifying the nature of the discordant orderings in the PGA assembly, and whether they reflect
18
19 346 errors in the reference assembly or structural polymorphisms between the Paxton Lake benthic
20
21 347 and Bear Paw Lake populations. The results presented here suggest that PGA scaffolding may
22
23 348 be a useful method to identify errors in reference assemblies or structural variation across
24
25 349 genomes.
26
27
28
29
30

350

351 **Funding**

352 This work was supported by an Evolutionary, Ecological, or Conservation Genomics Research
353 Award from the American Genetic Association to M.A.W.; the Office of the Vice President of
354 Research at the University of Georgia to M.A.W.; the National Institutes of Health (R01
355 GM116853 to C.L.P.); and the Fred Hutchinson Cancer Research Center Division of Basic
356 Sciences to C.L.P.
357

357

358 **Acknowledgements**

359 We thank Chris Amemiya for isolating high molecular weight DNA for use in optical mapping,
360 and Susan Brown and the Kansas State University Bioinformatics Center for performing the
361 optical mapping. All procedures were approved by the Fred Hutchinson Cancer Research
362 Center Institutional Animal Care and Use Committee (protocol 1575).
363

363

1
2
3 364 **Data Availability**
4

5 365 We have deposited the primary data underlying these analyses as follows:

6
7 366 -Hi-C sequences are deposited in the NCBI SRA database: SRP081031

8
9 367 -BioNano XMAP optical map files: Dryad

10
11 368 -Revised genome assembly: Dryad

12
13 369

14
15
16 370 **Disclosure Declaration**

17
18 371 STS and IL are employees of Phase Genomics. MAW and CLP declare no competing interests.

19
20 372

21
22
23 373 **References**

24 374

25 375 Bell MA, Foster SA. 1994. *The evolutionary biology of the threespine stickleback*. Oxford
26 376 University Press, Oxford, U.K.

27
28 377 Bickhart DM, Rosen BD, Koren S, Sayre BL, Hastie AR, Chan S, Lee J, Lam ET, Liachko I,
29 378 Sullivan ST, et al. 2017. Single-molecule sequencing and chromatin conformation capture
30 379 enable de novo reference assembly of the domestic goat genome. *Nat Genet* **49**: 643–650.

31
32 380 Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, Shendure J. 2013. Chromosome-scale
33 381 scaffolding of de novo genome assemblies based on chromatin interactions. *Nat Biotechnol*
34 382 **31**: 1119–1125.

35
36 383 Das SK, Austin MD, Akana MC, Deshpande P, Cao H, Xiao M. 2010. Single molecule linear
37 384 analysis of DNA in nano-channel labeled with sequence specific fluorescent probes. *Nucleic*
38 385 *Acids Res* **38**: e177–e177.

39
40 386 Dong Y, Xie M, Jiang Y, Xiao N, Du X, Zhang W, Tosser-Klopp G, Wang J, Yang S, Liang J, et
41 387 al. 2013. Sequencing and automated whole-genome optical mapping of the genome of a
42 388 domestic goat (*Capra hircus*). *Nat Biotechnol* **31**: 135–141.

43
44 389 Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, Shamim MS, Machol I,
45 390 Lander ES, Aiden AP, et al. 2017. De novo assembly of the *Aedes aegypti* genome using
46 391 Hi-C yields chromosome-length scaffolds. *Science* **356**: 92–95.

47
48 392 Ekblom R, Wolf JBW. 2014. A field guide to whole-genome sequencing, assembly and
49 393 annotation. *Evol Appl* **7**: 1026–1042.

50
51 394 Fierst JL. 2015. Using linkage maps to correct and scaffold de novo genome assemblies:
52 395 methods, challenges, and computational tools. *Front Genet* **6**: 220.

53
54 396 Glazer AM, Killingbeck EE, Mitros T, Rokhsar DS, Miller CT. 2015. Genome Assembly
55 397 Improvement and Mapping Convergent Evolved Skeletal Traits in Sticklebacks with

- 1
2
3 398 Genotyping-by-Sequencing. *G3* **5**: 1463–1472.
- 4
5 399 Gordon D, Huddleston J, Chaisson MJP, Hill CM, Kronenberg ZN, Munson KM, Malig M, Raja
6 400 A, Fiddes I, Hillier LW, et al. 2016. Long-read sequence assembly of the gorilla genome.
7 401 *Science* **352**: aae0344.
- 8
9 402 Hendry AP, Peichel CL, Matthews B. 2013. Stickleback research: the now and the next. ...
10 403 *Ecology Research* **15**: 111–141.
- 11
12 404 Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, Swofford R, Pirun M, Zody
13 405 MC, White S, et al. 2012. The genomic basis of adaptive evolution in threespine
14 406 sticklebacks. *Nature* **484**: 55–61.
- 15
16 407 Kaplan N, Dekker J. 2013. High-throughput genome scaffolding from in vivo DNA interaction
17 408 frequency. *Nat Biotechnol* **31**: 1143–1147.
- 18
19 409 Kent WJ. 2002. BLAT--the BLAST-like alignment tool. *Genome Res* **12**: 656–664.
- 20
21 410 Kingsley DM, Peichel CL. 2007. The molecular genetics of evolutionary change in sticklebacks.
22 411 In *Biology of the Three-Spined Sticklebacks* (eds. S. Östlund-Nilsson, I. Mayer, and F.
23 412 Huntingford), pp. 44–81, Boca Raton.
- 24
25 413 Kingsley DM, Zhu B, Osoegawa K, De Jong PJ, Schein J, Marra M, Peichel CL, Amemiya C,
26 414 Schluter D, Balabhadra S, et al. 2004. New genomic tools for molecular studies of
27 415 evolutionary change in threespine sticklebacks. *Behaviour* **141**: 1331–1344.
- 28
29 416 Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragozy T, Telling A, Amit I,
30 417 Lajoie BR, Sabo PJ, Dorschner MO, et al. 2009. Comprehensive mapping of long-range
31 418 interactions reveals folding principles of the human genome. *Science* **326**: 289–293.
- 32
33 419 Marie-Nelly H, Marbouty M, Cournac A, Flot J-F, Liti G, Parodi DP, Syan S, Guillén N, Margeot
34 420 A, Zimmer C, et al. 2014. High-quality genome (re)assembly using chromosomal contact
35 421 data. *Nat Commun* **5**: 5695.
- 36
37 422 Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, Kravitz SA, Mobarry CM,
38 423 Reinert KH, Remington KA, et al. 2000. A whole-genome assembly of *Drosophila*. *Science*
39 424 **287**: 2196–2204.
- 40
41 425 Nagarajan N, Pop M. 2013. Sequence assembly demystified. *Nat Rev Genet* **14**: 157–167.
- 42
43 426 Östlund-Nilsson S, Mayer I, Huntingford F. 2007. *Biology of the three-spined stickleback*. CRC
44 427 Press, Boca Raton, FL.
- 45
46 428 Peichel CL, Marques DA. 2017. The genetic and molecular architecture of phenotypic diversity
47 429 in sticklebacks. *Philosophical Transactions of the Royal Society B: Biological Sciences* **372**:
48 430 20150486.
- 49
50 431 Putnam NH, O'Connell BL, Stites JC, Rice BJ, Blanchette M, Calef R, Troll CJ, Fields A, Hartley
51 432 PD, Sugnet CW, et al. 2016. Chromosome-scale shotgun assembly using an in vitro
52 433 method for long-range linkage. *Genome Res* **26**: 342–350.
- 53
54
55
56
57
58
59
60

- 1
2
3 434 Roesti M, Moser D, Berner D. 2013. Recombination in the threespine stickleback genome-
4 435 patterns and consequences. *Mol Ecol* **22**: 3014–3027.
- 6 436 Ross JA, Peichel CL. 2008. Molecular cytogenetic evidence of rearrangements on the Y
7 437 chromosome of the threespine stickleback fish. *Genetics* **179**: 2173–2182.
- 9 438 Salzberg SL, Phillippy AM, Zimin A, Puiu D, Magoc T, Koren S, Treangen TJ, Schatz MC,
10 439 Delcher AL, Roberts M, et al. 2012. GAGE: A critical evaluation of genome assemblies and
11 440 assembly algorithms. *Genome Res* **22**: 557–567.
- 13 441 Shelton JM, Coleman MC, Herndon N, Lu N, Lam ET, Anantharaman T, Sheth P, Brown SJ.
14 442 2015. Tools and pipelines for BioNano data: molecule assembly pipeline and FASTA super
15 443 scaffolding tool. *BMC Genomics* **16**: 734.
- 17 444 Shi L, Guo Y, Dong C, Huddleston J, Yang H, Han X, Fu A, Li Q, Li N, Gong S, et al. 2016.
18 445 Long-read sequencing and de novo assembly of a Chinese genome. *Nat Commun* **7**:
19 446 12065.
- 21 447 Teague B, Waterman MS, Goldstein S, Potamouisis K, Zhou S, Reslewic S, Sarkar D, Valouev
22 448 A, Churas C, Kidd JM, et al. 2010. High-resolution human genome structure by single-
23 449 molecule analysis. *Proc Natl Acad Sci USA* **107**: 10848–10853.
- 25 450 Treangen TJ, Salzberg SL. 2012. Repetitive DNA and next-generation sequencing:
26 451 computational challenges and solutions. *Nat Rev Genet* **13**: 36–46.
- 28 452 Wootton RJ. 1976. *The Biology of Sticklebacks*. Academic Press, U.K.
- 30 453 Zhang Q, Chen W, Sun L, Zhao F, Huang B, Yang W, Tao Y, Wang J, Yuan Z, Fan G, et al.
31 454 2012. The genome of *Prunus mume*. *Nat Commun* **3**: 1318.

35 455

37 456

457 **Figure Legends**

458

459 **Figure 1.** Proximity-Guided Assembly (PGA) clusters all linkage groups of the *G. aculeatus*
460 genome. The revised *G. aculeatus* genome assembly (Glazer et al. 2015) was divided into
461 contiguous contigs of varying length (20-100 kb) and assembled using PGA. The *G. aculeatus*
462 revised reference assembly contig order is preserved along the X-axis. PGA clustering was
463 largely congruent with the revised *G. aculeatus* reference genome, as shown by each linkage
464 group (LG) assembled as a contiguous segment. One contig was assigned to a different linkage
465 group in the PGA clustering than in the reference assembly (circled).
466

467 **Figure 2.** Misorderings within linkage groups are consistent with structural variation between
468 populations of *G. aculeatus*. Misorderings in the Proximity-Guided Assembly (PGA) scaffolding
469 within linkage groups (A-C) and alignment of bacterial artificial chromosome (BAC) mate-pair
470 sequences (D-F) relative to the *G. aculeatus* revised reference assembly (Glazer et al. 2015)
471 are shown for three linkage groups that had significant overlap between the two data sets: II
472 (A,D), XVI (B,E), and XVIII (C,F). Alignment of BAC-end sequences from the same Paxton Lake
473 benthic population used for the PGA scaffolding reveal structural variation around the
474 breakpoints of the misordered PGA scaffolds. Gray points are left and right concordantly aligned
475 mate-pairs from the BAC library and fall slightly off either side of the 1:1 diagonal, reflecting the
476 148 kb average insert size of the library (Kingsley et al. 2004). Discordant read pairs are
477 highlighted in color. Blue points indicate mate pairs that align in a forward/reverse orientation,
478 reflecting a putative deletion relative to the reference genome assembly. Red points indicate
479 mate pairs that align in a forward/forward or reverse/reverse orientation indicating a putative
480 inversion relative to the reference genome assembly. The remaining linkage groups are shown
481 in Figure S1.
482

483 **Figure 3.** Proximity-Guided Assembly (PGA) clusters a large proportion of previously
484 unassigned contigs to linkage groups. The revised *G. aculeatus* genome assembly (Glazer et al.
485 2015) was split into contigs at gaps and clustered with PGA along with 21.7 Mb of contigs
486 previously unassigned to linkage groups in the *G. aculeatus* genome. The previously
487 unassigned contigs (red vertical band) are distributed across linkage groups (LG) by the PGA
488 clustering. PGA is less accurate clustering the *G. aculeatus* genome when these previously
489 unassigned contigs are included, shown by an increased number of contigs being incorrectly
490 assigned to different linkage groups (127 incorrectly assigned contigs, 0.94% of the previously
491 assembled contig count).
492

493

Table 1. Misorderings in the Proximity-Guided Assembly (PGA) scaffolding often overlap with bacterial artificial chromosome (BAC)-ends that align discordantly to the reference genome.

Linkage group	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

497

498

499 **Table 2.** Distribution of contigs scaffolded by Proximity-Guided Assembly (PGA) that were
 500 previously unassigned in the *G. aculeatus* genome.
 501

Linkage group	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	8,144	23	216,619
IV	4	92,707	54	455,301
V	7	34,105	20	108,253
VI	1	8,060	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	1,635	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	8,315	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	1,263	12,251,423

502

503

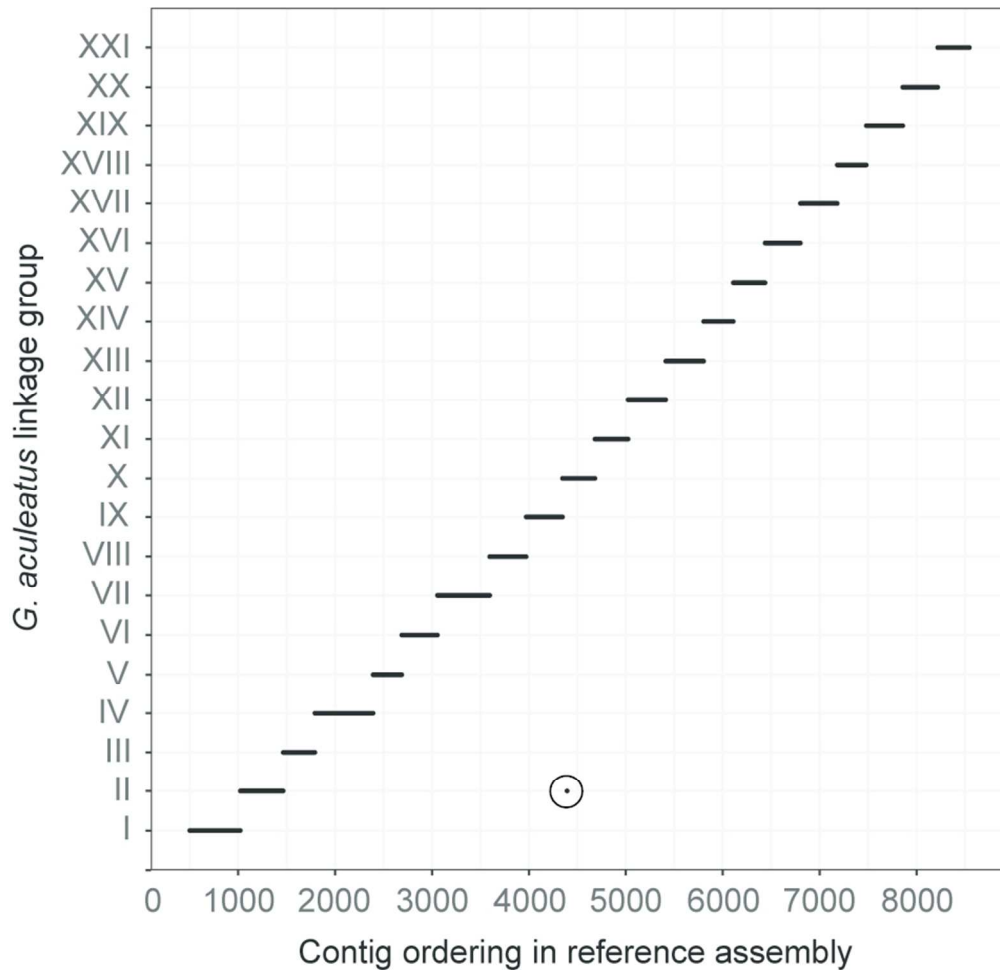


Figure 1. Proximity-Guided Assembly (PGA) clusters all linkage groups of the *G. aculeatus* genome. The revised *G. aculeatus* genome assembly (Glazer et al. 2015) was divided into contiguous contigs of varying length (20-100 kb) and assembled using PGA. The *G. aculeatus* revised reference assembly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

161x156mm (150 x 150 DPI)

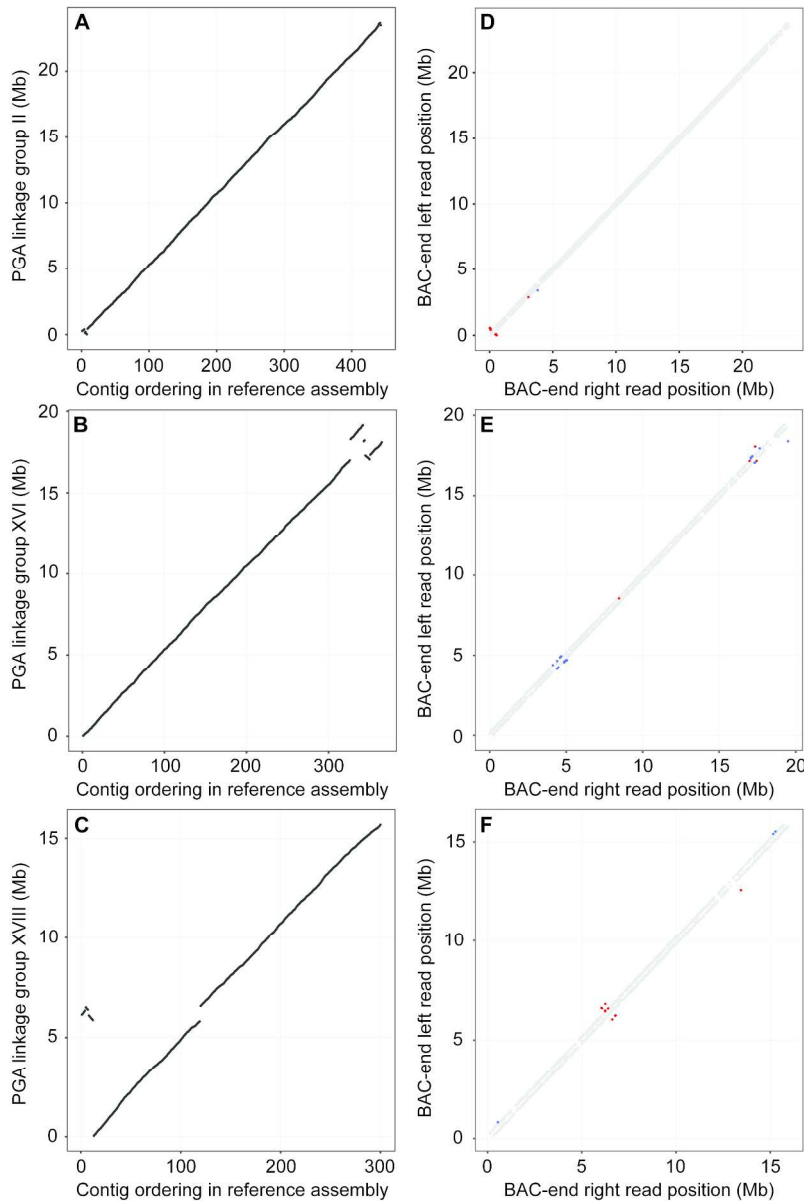


Figure 2. Misorderings within linkage groups are consistent with structural variation between populations of *G. aculeatus*. Misorderings in the Proximity-Guided Assembly (PGA) scaffolding within linkage groups (A-C) and alignment of bacterial artificial chromosome (BAC) mate-pair sequences (D-F) relative to the *G. aculeatus* revised reference assembly (Glazer et al. 2015) are shown for three linkage groups that had significant overlap between the two 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 linkage groups are shown in Figure S1.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

182x274mm (300 x 300 DPI)

For Peer Review

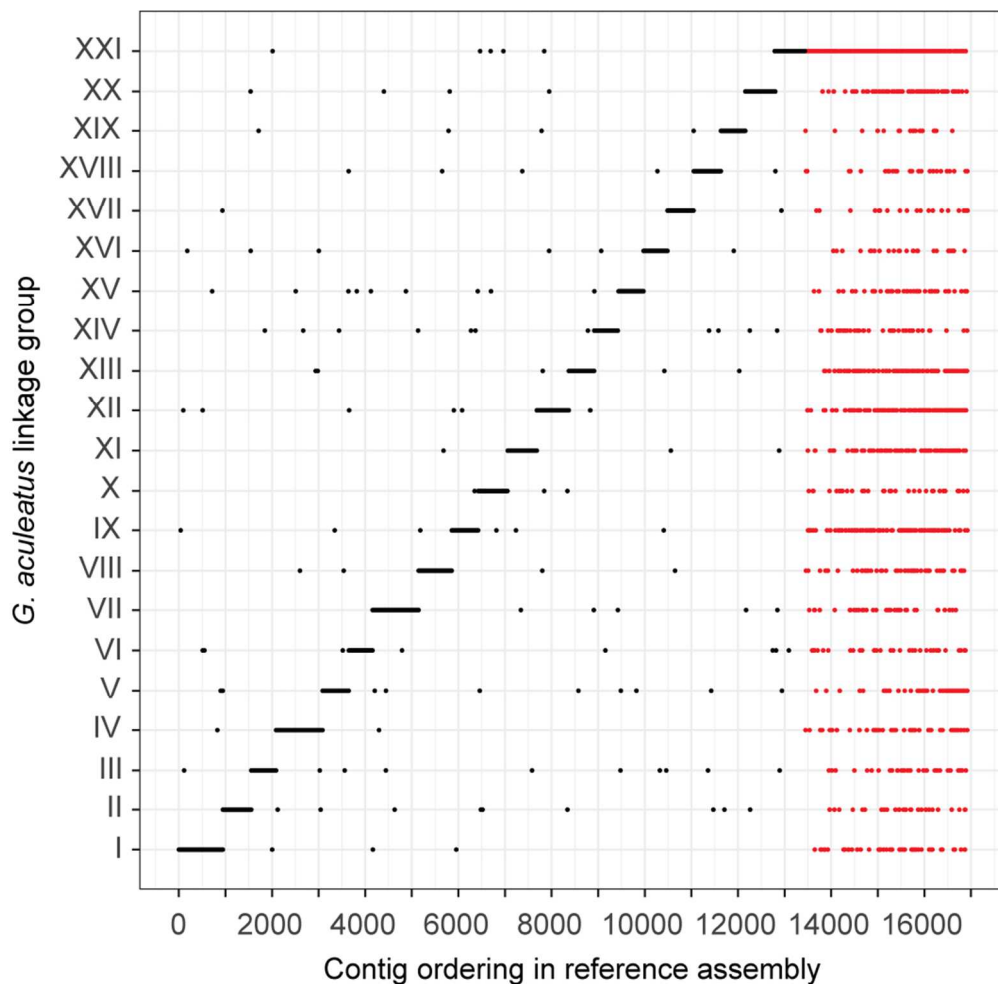


Figure 3. Proximity-Guided Assembly (PGA) clusters a large proportion of previously unassigned contigs to linkage groups. The revised *G. aculeatus* genome assembly (Glazer et al. 2015) was split into contigs at gaps and clustered with PGA along with 21.7 Mb of contigs previously unassigned to linkage groups in the *G. aculeatus* genome. The previously unassigned contigs (red vertical band) are distributed across linkage groups (LG) 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 linkage groups (127 incorrectly assigned contigs, 0.94% of the previously assembled contig count).

179x175mm (150 x 150 DPI)

