BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 16 2011, pages 2187–2193
doi:10.1093/bioinformatics/btr346

Genome analysis

Advance Access publication June 27, 2011

Paired-end RAD-seq for *de novo* assembly and marker design without available reference

Eva-Maria Willing[1], Margarete Hoffmann[1], Juliane D. Klein[2], Detlef Weigel[1] and Christine Dreyer[1,*]

[1]Department of Molecular Biology, Max Planck Institute for Developmental Biology and [2]Department of Algorithms in Bioinformatics, University of Tübingen, Sand 14, 72076 Tübingen, Germany

Associate Editor: Alex Bateman

ABSTRACT

Motivation: Next-generation sequencing technologies have facilitated the study of organisms on a genome-wide scale. A recent method called restriction site associated DNA sequencing (RAD-seq) allows to sample sequence information at reduced complexity across a target genome using the Illumina platform. Single-end RAD-seq has proven to provide a large number of informative genetic markers in reference as well as non-reference organisms.

Results: Here, we present a method for *de novo* assembly of paired-end RAD-seq data in order to produce extended contigs flanking a restriction site. We were able to reconstruct one-tenth of the guppy genome represented by 200–500 bp contigs associated to EcoRI recognition sites. In addition, these contigs were used as reference allowing the detection of thousands of new polymorphic markers that are informative for mapping and population genetic studies in the guppy.

Availability: A perl and C++ implementation of the method demonstrated in this article is available under http://guppy.weigelworld.org/weigeldatabases/radMarkers/ as package RApiD.

Contact: christine.dreyer@tuebingen.mpg.de

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on April 18, 2011; revised on May 27, 2011; accepted on June 1, 2011

1 INTRODUCTION

The availability of increasing amounts of DNA sequence information has greatly facilitated studying of many biological questions, especially in the context of genome evolution, natural variation and adaptive processes and association mapping. Next-generation sequencing (NGS) technologies have revolutionized the field of genome research, at first by allowing cheap re-sequencing projects for organisms with an already existing reference genome. Recently, more and more methods have been developed incorporating NGS to analyze also non-reference organisms, taking advantage of improvements such as longer read lengths and paired-end (PE) reads. However, *de novo* assemblies of large genomes from very short reads remain difficult, in spite of recent improvements in assembly algorithms (Gnerre *et al.*, 2011). Yet, for a large number of interesting questions a high number of genetic markers equally distributed over the genome would already be very informative, even without a complete genome sequence. Baird *et al.* (2008) developed a protocol for high-throughput sequencing of restriction site-associated DNA (RAD) tags using the Illumina platform (RAD-seq). It has the advantage that only a reduced representation of the genome is sequenced leading to deep sequence coverage of fragments near a specific type of restriction site. They showed that single end (SE) sequencing of RAD tags could be used for rapid marker development in Stickleback for which a reference genome is available. Since then, SE RAD-seq has become a popular tool in next-generation population genetics (Davey and Blaxter, 2010; Emerson *et al.*, 2010; Hohenlohe *et al.*, 2011; Pfender *et al.*, 2011). In addition, Illumina PE sequencing could extend the sequence information on each side of the restriction sites (Baird *et al.*, 2008; Davey and Blaxter, 2010; Etter *et al.*, 2011). Because each RAD can provide a unique genomic sequence tag that can be characterized without its immediate genomic context, the first reads may be aligned to each other, building subsets that are associated to one restriction site each. As a strategy for obtaining longer sequence tags, we exploited the fact that random mechanical shearing leads to a family of staggered second reads that can be assembled to longer subsets associated to the RE site defined by the first read cluster. This strategy subdivides the assembly problem into a high number of less complex local assemblies. In this study, we analyze PE RAD-seq data from two very diverged guppy populations, namely Quare and Cumana, which have been previously used to generate a genetic linkage map (Tripathi *et al.*, 2009). The guppy (Poecilia retitculata) is an important model organism in ecological genetics, and adaptation to contrasting habitats has been extensively studied in field experiments (Magurran, 2005; Reznick *et al.*, 2001). However, due to the lack of a sufficient number of genetic markers the molecular background is still unknown. We show that our approach can generate *de novo* 283 842 RAD tags that are 200–400 bp long and cover ~10% of the guppy genome. Furthermore, these tags can be used as reference to design thousands of new polymorphic markers useful for population genetic and mapping studies. All tools developed for the analysis can be downloaded from http://guppy.weigelworld.org/weigeldatabases/radMarkers/ as package RApiD.

2 MATERIAL AND METHODS

2.1 Creation and sequencing of the RAD library

The genomic RAD libraries were created as described by Baird *et al.* (2008). Briefly, genomic DNA pooled from six individuals each was digested with EcoRI (NEW ENGLAND BioLabs). Pools represented Cumaná and Quare males and females and technical replicates of Quare males and Cumaná

*To whom correspondence should be addressed.

Table 1. Sequence information and read counts for each 12 bp MID

MID	Sequence	Sample	Million reads
1	ATGTGTCGCCAA	6 Quare males[a]	4.6
2	TCTGAGCGTACA	6 Quare males[a]	3.4
3	GATCTGAAGCTC	6 Quare females	0.015
4	CGACGATACTTG	6 Cumaná males	5.1
5	CTAGATGCTGAC	6 Cumaná females[a]	4.4
6	GACACCGTATGT	6 Cumaná females[a]	5.4

[a]Technical replicates.

females were included (Table 1). Illumina P1 adaptors including a unique 12 bp multiplex identifier (MID) preceding the EcoRI site were added by ligation. All MIDs differed by at least seven bases and were therefore tolerant to up to three errors. After ligation of the P1 adaptors containing the different MIDs, the six DNA samples were pooled in proportionate amounts before shearing (Covaris) and addition of the P2 adaptor. A single library with an insert size range of 200–400 bp was prepared and sequenced from both ends with 100 bp read lengths in one lane of an Illumina GAIIX sequencer (Fig. 1A). Sequence reads can be downloaded from http://guppy.weigelworld.org/weigeldatabases/radMarkers/.

2.2 *De novo* assembly of RAD tags

For quality control, all first reads were checked for presence of the partial 5 bp EcoRI motif (AATTC) following the 12 bp MID. Second, all reads containing uncalled nucleotides were removed from the dataset. After removal of the MID and restriction site sequence, the remaining first reads were grouped into pools representing the same RAD tag, using vmatch (www.vmatch.de) (Fig. 1B). We allowed a maximal hamming distance of three within the same cluster. After clustering the first reads, the second reads could be sorted into groups accordingly that were assembled separately (Fig. 1B). Ideally, the local assembly of second reads in a cluster results in one contig indicating that they indeed originate from the same RAD tag. Mixed clusters of first reads could be caused by RE sites in repetitive regions. Such clusters might be resolved if the second reads were in a region outside the repeat and could be assembled into unique sequences. In such cases, the assembly of the second reads resulted in more than one contig and allowed resolution of the mixed first reads accordingly. Every single cluster for each tag could have a different set of optimal assembly parameters, because of different repeat content and number of reads per cluster. For example, if the coverage of a tag is low, a smaller overlap length should be used for the assembly. We used the assembler LOCAS (http://ab.inf.uni-tuebingen.de/software/locas/) that uses an Overlap-Layout-Consensus approach to keep track of the overlaps among reads and is especially developed for low coverage data. LOCASopt is a wrapper that calls the assembler LOCAS with a different set of parameters in order to assemble the reads in a cluster several times under different conditions. Parameters that can be optimized are overlap length, percent of mismatches allowed in overlap and seed size (see LOCAS Manual). LOCASopt keeps track of all the assemblies in order to choose the optimal one. We defined the optimal assembly as the one resulting in the smallest number of contigs and incorporating the largest fraction of available reads in a cluster. In order to test if optimizing each local assembly leads to better results, we assembled our data once with LOCASopt iterating over a large set of different parameter combinations, namely overlap = 21, 23, … 67, *k*-mer = 13, 15, 17 and mismatch rate = 0.05, 0.07, 0.09. Additionally, we assembled the same set of clusters a second time with the parameters fixed at the values mostly used in the previous assembly (see Section 3).

After assembling, the second read contigs are joined with the consensus of the corresponding first reads (Fig. 1B). In order to generate a high-quality reference, we performed an additional quality control by mapping back all read pairs to the assembled tags and calling the majority consensus (see Section 2.3) for each tag requiring a minimal quality of 20 (corresponding to a 0.01% chance that a base was wrongly called) and a minimal coverage of two per base. After that, uncalled nucleotides at the ends of the second read contigs were removed. If there were uncalled nucleotides in the middle of a second read contig, the contig was split up at these positions and the longest resulting substring remained as representative of this contig. Depending on the insert size of the library, the first read consensus and second read contig can be overlapping or non-overlapping (Fig. 1B). Therefore, we checked for an overlap between the two parts requiring a minimal overlap length of 10 bp and a maximal mismatch rate of 5%.

2.3 Consensus and SNP calling

After generating a comprehensive high-quality reference, reads were sorted according to their MIDs and separately mapped back to the reference (Fig. 1C). We used genomemapper (Schneeberger *et al.*, 2009) to map the reads back to the reference allowing up to five mismatches and no gaps. A mapped read pair has to pass several quality controls to be considered for consensus or SNP calling (Fig. 1B). Both reads in a pair have to map to the same contig in the right direction, with the start of the first read at the first position in a tag. A pair is only considered if at least one member uniquely maps to one contig in the reference. Furthermore, read pair clones are removed in order to prevent false positive SNP calls that were caused by errors occurring during the amplification of the library. A read pair was considered to be a clone, if the second read maps to the same position in a reference tag as another second read in a previous pair. After mapping the reads back, the consensus base for each position in the reference was called by determining the major base at that position in the reads that could be mapped back. We used only bases with a minimal quality of 20 for consensus calling. Each consensus base got a quality value that was the average over the quality values of the bases used for consensus calling. If a position in the assembled reference was not covered during the consensus calling, it was marked with a 'N' as uncalled nucleotide.

The search for polymorphic sites was done in a similar way as the consensus calling. A given site was considered polymorphic if the polymorphism occurred in at least a certain number of reads and if the site had a minimal coverage above threshold. In order to call a homozygous SNP, all reads must contain the same nucleotide that must be different from the reference. As in the consensus calling, only bases were considered that reached a certain quality threshold. The quality of a SNP is the average of the qualities of the single bases at the SNP position.

3 RESULTS AND DISCUSSION

3.1 PE sequencing

In order to generate a dense set of RAD markers, we chose the restrictions enzyme EcoRI, which recognizes the palindromic 6 bp sequence G′AATT,C. The guppy genome size is nearly 1 Gb as estimated by flow cytometry (M.Schartl, personal communication). Based on sequenced BAC ends from a genomic library of the Cumaná guppy, we predicted the guppy genome to be relatively AT rich (60%), close to the AT content of the EcoRI recognition site. For simplicity, we will assume that EcoRI sites occur close to the expected frequency of 1/4096 bp, and that we have therefore an expected number of 500 000 RAD tags. To test the sequencing depth required as well as reproducibility of the results, we pooled six independently digested bulks of DNA from six individuals each, representing males and females of two different populations and technical replicates (Table 1). PE sequencing with 101 bp read length of this pool on a single lane of an Illumina flowcell resulted in 23.4 million read pairs, of which 97% (22.6 million) contained the correct restriction site pattern (AATTC) at the beginning of the first read

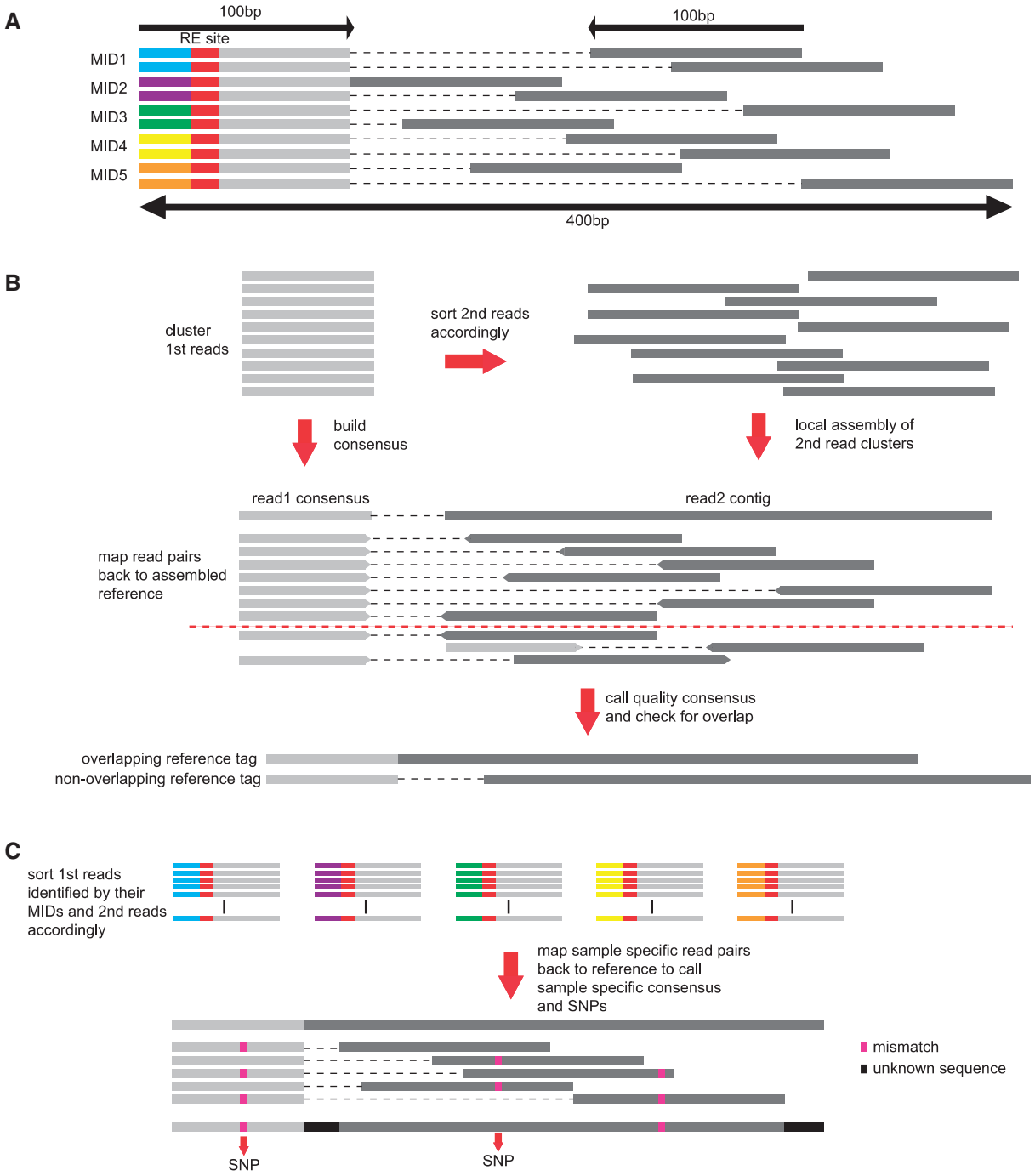

Fig. 1. (**A**) RAD-seq output. Fragments are sheared randomly. By PE sequencing of fragments between 200 and 400 bp, the obtained second reads are staggered and cover a range of 100–300 bp, whereas the first reads contain the MID and the restriction site and therefore start always at the same genomic position. (**B**) After removing the MID and the restriction site, the first reads can be aligned to each other (clustering). According to these clusters, the second reads can also be sorted and assembled separately to a contig, which is then linked to the first read majority consensus. The tag pairs then serve as a reference to which all reads are mapped back and a majority consensus is called using only high-quality bases from read pairs that mapped to the assembly fulfilling certain constrains (read pairs below red dashed line are discarded). After this step, all remaining tag pairs are checked whether they overlap. (**C**) All reads can be sorted according to their MIDs, and sample-specific consensus sequences and SNPs can be called by mapping the sorted reads back to the reference.

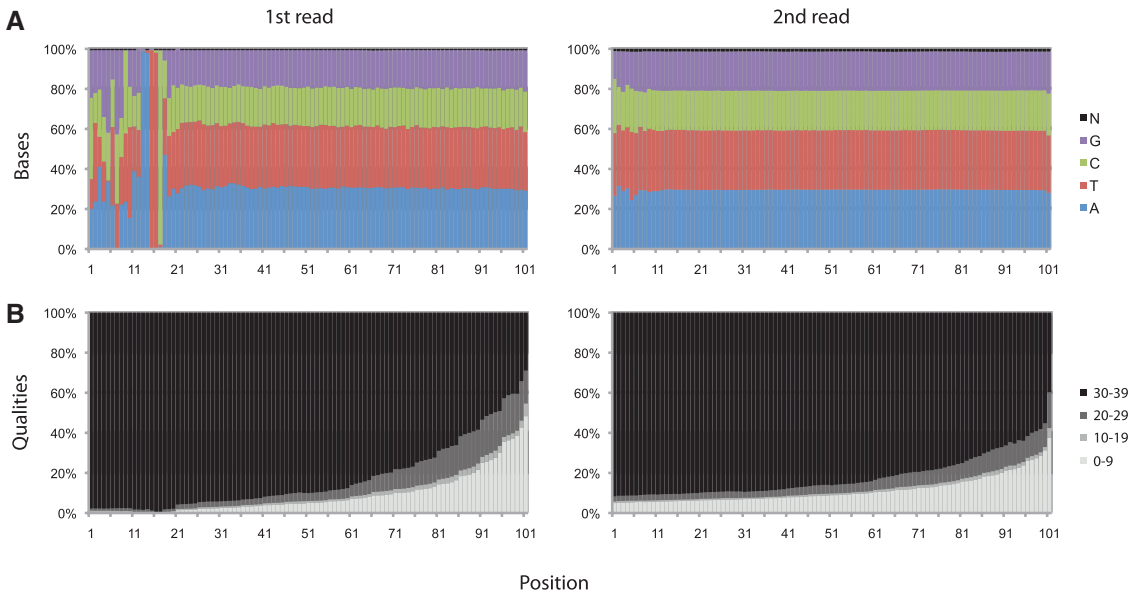

Fig. 2. Base distribution and quality scores along the reads. (**A**) The number of bases per position in each read was determined. For read1, positions 1–12 contain the sample specific MID and at position 13–17 is the restriction site (clearly seen in the base counts at these position). Read2 is completely composed of genomic bases. The base distribution at positions containing solely genomic regions reflects the expected distribution of ∼60% AT. (**B**) Quality values were counted along the reads. As expected, quality values decrease to the end of the reads.

and no uncalled positions. Consequently, assuming 500 000 tags, each tag should be covered by ∼46 read pairs on average. Figure 2 shows the base and quality score counts per site in each read. The base distribution over the first 17 bp in the first read nicely depicts the MIDs and the restriction site. However, at the first position after the restriction site G is significantly underrepresented, possibly as a consequence of genomic CG methylation inhibiting EcoRI. Yet after position 18, the distribution converges on the expected values (60% AT, 40% GC), which is seen over the entire second read. As expected, quality values of both reads decrease over their length (Bansal *et al.*, 2010), with a slightly faster decline in the first read, possibly caused by the unequal base distribution in the first 17 bp. For consensus and SNP calling, the reads were sorted according to sample specific MIDs (Table 1). Differences between read counts for the different samples deviated less than a factor of 1.6 from each other, with one exception. This is within the range previously encountered when sequencing multiplexed samples (Craig *et al.*, 2008). We obtained only ∼15 000 reads encoded with MID3, suggesting technical failure (Craig *et al.*, 2008).

3.2 Clustering and *de novo* assembly

All-against-all alignment of reads resulted in 451 981 first-read clusters with ∼48 reads on average (range 2–66 393). For assembly, we considered only 297 147 (65.7%) clusters within a certain coverage range (5–184), in order to avoid highly repetitive regions. These clusters had an average size of 63 reads and included 18.9 million (81%) of the reads.

The second reads belonging to each first-read cluster were sorted and assembled separately to obtain a second-read contig for each cluster (Fig. 1B). If the assembly of a cluster resulted in more than one contig or if not all the reads were used in the assembly, the first reads were sorted anew, according to the assembled contigs. We performed the assembly twice, once iterating over different parameter settings and once fixing the parameters at the values mostly used in the optimized assembly (overlap = 21, *k*-mer = 13, mismatch rate = 0.05). The assembly with fixed parameters resulted in 503 748 contigs with an average length of 286 bp, representing 291 149 clusters and incorporating 76.6% of the reads. On average, 28 read pairs contributed to one RAD tag (Table 2). In the optimized assembly, 291 159 clusters were assembled resulting in 334 215 second-read contigs with an average length of 349 bp using 76.8% of the input reads. On average, 43 read pairs contributed to one RAD tag, which is close to the 46 read pairs expected per tag (Table 2). Figure 3 shows that after optimizing the assembly the increase in the number of longer contigs was marginal, but most of the very short contigs may have been merged with longer contigs by choosing a different set of parameters. This notion is supported by the fact that significantly less clusters result in more than one contig in the optimized assembly (8.7% compared with 31.0%, Table 2). Consequently, optimizing the set of parameters for each local assembly led to less, but on average longer second read contigs with a higher number of reads used per contig. We therefore used these contigs for all the following analyses.

3.3 Quality control

Following the strategy detailed in Section 2, we found 283 842 contigs fulfilling the quality requirements, corresponding to ∼57% of the tags expected. This is comparable to the number of EcoRI RAD tags found in stickleback (Baird *et al.*, 2008), where short (36 bp) reads were aligned to a reference genome. Of the assembled guppy RAD tags, 51.4% were overlapping with their corresponding first read consensus, over a length of 29 bp and with an average

Table 2. Results of the optimized versus the not optimized assembly

	Fixed	Optimized
Clusters resulting in an assembly	291 149	291 159
Contigs	503 748	334 215
Custers resulting in >1 contig (%)	31.0	8.7
Reads used	75.6	76.8
Average contig length (bp)	286	349
Sum of all contigs (Mb)	143.8	116.6

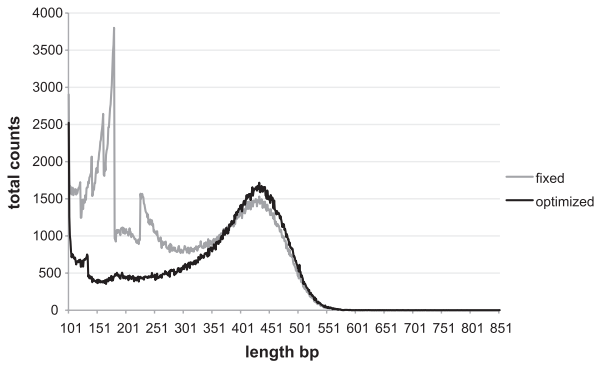

Fig. 3. Length distribution of assembled contigs. The assembly of the second read clusters was performed twice. Once with parameters fixed at certain values and once trying a large set of different assembly parameter combinations to find the optimal one for each cluster. The optimal assembly was defined as the one resulting in the least number of contigs, but incorporating the largest fraction of reads.

mismatch rate of 0.002%. For these tags, we obtained on average 417 bp continuous sequence corresponding to 60 Mb in total. The second read contigs of the non-overlapping tags were on average 259 bp long. Taking the sequence information together from all overlapping and non-overlapping tags, we obtained 108.2 Mb of sequence, corresponding to about one-tenth of the guppy genome, close to the expectation.

In order to assess the quality of the *de novo* assembled reference, we predicted RAD markers of ≥150 bp length by digesting 6165 sequenced Cumaná BAC ends with EcoRI *in silico*. These were used as queries in a Blastn search against our high-quality reference. Of 1112 predicted RAD markers, 862 (77.5%) matched (≤1e-100) our assembled reference (Supplementary Table S1). Of these 862 hits, 798 (92.6%) covered >90% of the query or the subject sequence and included the restriction site at one end (Supplementary Table S1). These results show that our strategy led to a high-quality reference of *de novo* assembled RAD markers that can be further used for sample-specific consensus and SNP calling.

3.4 Sample-specific consensus calling

After assembly and quality control, we sorted the reads according to their MIDs and mapped each batch on the reference in order to call sample-specific consensus sequences. Baird *et al.* (2008) used the presence or absence of a tag to identify it as polymorphic. The absence of a RAD tag in one sample is probably most often caused by a polymorphism in the associated restriction site. However, random

Table 3. Pairwise comparison of missing RAD tags between the five sample pools using different coverage thresholds

MID	Minimum coverage	1	2	4	5	6
	1×	218 946	2.88	19.49	19.89	19.35
1	6×	174 735	0.26	16.81	17.07	16.75
	10×	137 679	0.04	15.11	15.31	15.06
	1×	1.89	216 720	19.53	19.93	19.37
2	6×	0.06	153 463	16.02	16.28	15.98
	10×	0.01	103 495	14.03	14.26	14.06
	1×	25.11	25.91	235 368	3.54	2.86
4	6×	20.70	21.17	185 295	0.73	0.60
	10×	18.13	18.46	148 954	0.34	0.32
	1×	25.17	25.97	3.15	234 404	1.80
5	6×	20.14	20.57	0.39	178 107	0.05
	10×	17.25	17.53	0.10	137 078	0.01
	1×	25.30	26.07	3.28	2.62	236 381
6	6×	20.97	21.47	0.50	0.15	190 100
	10×	18.27	18.60	0.12	0.02	155 298

Diagonal contains the total number of tags in the sample with the required coverage. Remaining entries give the percentage of RAD tags that can be found in sample *i* (rows) but are missing in sample *j* (columns).

sampling in the sequencing process can cause false positives. Therefore, Baird *et al.* (2008) scored only such markers as absent that were represented by at least eight reads in one sample and by none in the other sample. We tested whether this strategy also works with *de novo* RAD tags by comparing the intersections between the different samples using different coverage cutoffs (1×, 6×, 10×) to assign a marker as polymorphic. The technical replicates provided the opportunity to estimate the false positive rate at the different coverage cutoffs. Table 3 shows how many tags we found per sample at different coverage cutoffs (diagonal) and the percentage of markers that could not be found in the intersection between the different samples and would therefore be scored as polymorphic. The false positive rate declines from >1% with a minimum coverage of 1× to <0.3% and 0.04% with minimum coverage of 6× and 10×, respectively. We see from Table 3 that the percentage of absent markers between the samples from the two different populations is much higher (>14% at all coverage thresholds) than the highest false positive rate. We infer that a significant number of polymorphic markers is caused by sequence variation that changes restriction enzyme sites. At 10× coverage, <0.3% of these markers are false positives.

In guppies, sex is genetically determined and sex-linked inheritance and sex chromosome evolution are topics of general interest in this species (Lindholm and Breden, 2002). Sex is determined by male heterogamety (XY), but the master sex determining locus, which appears to be located at the distal end of the Y chromosome, has not yet been precisely mapped due to a lack of markers (Tripathi *et al.*, 2009). We inspected our *de novo* assembled RAD tags for sex-specific markers. At 10× coverage, there were at least 2.5-fold more markers polymorphic (0.1%/0.12% and 0.34%/0.32%, Table 3) in the Cumaná female/male (4 compared to 5 and 6, Table 3) contrasts, compared with the Cumaná female/female (0.02 and 0.01%, 5 and 6, Table 3) or Quare male/male (0.04 and 0.01%, 1 and 2, Table 3) constrasts, corresponding to ∼149 female-specific tags and ∼477 male-specific tags. Because 40% of these

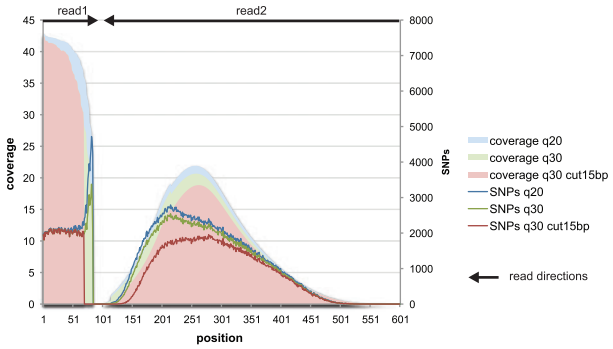

Fig. 4. Distribution of polymorphic sites in assembled tags. All read pairs were mapped back to the tag pairs in order to calculate the total coverage and the total number of polymorphisms per site as a function of the quality cutoff and read length used.

markers are expected to be false positives at 10× coverage, a higher coverage threshold should be used.

3.5 Distribution and fidelity of polymorphic sites

The distribution of polymorphic sites along the assembled RAD tags was analyzed by mapping all reads back to the assembled reference. A site was regarded as polymorphic if the polymorphism was covered by at least two reads and the coverage was at least six-fold. SNPs were called with quality thresholds of either 20 or 30. Figure 4 shows that the coverage decreases significantly toward the end of the first read, with declining quality scores, as is typical for the end of the reads (Bansal *et al.*, 2010). Over the first 69 bp, SNPs are found with equal frequency at each position in the first read, but the number of SNPs significantly increases to the end of the first read even when using a quality threshold of 30. However, this might not only be caused by decreasing quality values at the end of the reads, but also might be due to more misalignments at the end of the reads. When we do not use the last 15 bp of each mapped read for SNP calling, we reduce the number of SNPs mainly at the proximal end of the second read part of the tag (red curve in Fig. 4).

Figure 4 also illustrates that the second read contigs have their maximal coverage around position ∼270 bp and that the coverage decreases as expected toward both ends of the contigs. Furthermore, the likelihood to detect a SNP at a certain position in the second read part of a tag is positively correlated to the coverage. However, on average above a coverage of ∼15 fold SNP detection does not seem to increase further, suggesting that such coverage is sufficient to detect the majority of alleles.

To determine the number of SNPs that could be confirmed in the intersection of technical replicates, we analyzed each sample separately. Based on the observations described above, we performed the sample-specific SNP calling disregarding the last 15 bp of each read and considering only those positions in the reference having a coverage at least equal to a certain cutoff in all samples. For the Quare male replicates, we used the Cumana consensus as reference and for the Cumana female replicates the Quare consensus as reference, in order to compare high fidelity rates for heterozygous as well as homozygous SNPs. At 6× coverage, 84% of the heterozygous SNPs within the Quare replicates, and 86% of heterozygous SNPs within the Cumana replicates, could be found in the intersection. At 10× coverage, these numbers increased as expected slightly to 89 and 90%, respectively. In order to determine whether this applies to both parts of a tag, we examined the intersections of the first and second read part separately. We found that 87–91% of the SNPs detected in the first read lie in the intersection between the technical replicates, but only 78–80% of the heterozygous SNPs in the second read. Apart from the higher coverage in the first read, this could also be partly due to position-dependent systematic errors in the base calling that are equally likely in each sample. Since the first reads in a RAD tag are completely overlapping, position-dependent systematic errors can lead to false positive heterozygous SNPs that are shared among different multiplexed samples. The position-dependent effect cannot occur in the second part of a tag because we did not consider read pair clones. However, homozygous SNPs differ from the heterozygous SNPs in their fidelity. At 6× coverage, we find >97.1% of the SNPs in the intersection of the technical replicates, and this increases only to >97.7% at 10× coverage. Moreover, the intersections between the first and the second part differ by <1%. This indicates that the detection of homozygous SNPs between populations is highly reproducible with this method. Nevertheless, our approach also allows the detection of a high number of high fidelity heterozygous SNPs within populations at a specificity rate of >78% at comparatively low coverage. We have scored polymorphic sites using a newly developed approach, because the first read and error models developed for SNP calling in whole-genome sequencing data do not apply to RAD-seq data. As the first read of a specific tag starts at an invariant position, a SNP within the first read will always be at the same position. This is severely punished by some error models used for SNP calling, because sequencing errors at the same site are correlated (Li *et al.*, 2008). In addition, we do not expect a large number of insertions and deletions causing misalignments, because the reference is assembled with the reads that are also used for SNP calling. Moreover, repetitive sequences are removed by removing large first read clusters. These properties make the alignment problem fairly easy and eliminate the main sources of false positive SNPs in genomic data (Li *et al.*, 2008; Malhis and Jones, 2010). While our approach supports the use of other SNP calling algorithms using the assembled consensus tags as reference, we would advise to filter the mapping file used as input, following the criteria for informative reads defined in this study (Sections 2 and 2.3).

3.6 Polymorphic markers within and between Quare and Cumaná populations

To determine the number of polymorphic markers within and between Quare and Cumaná specimens, we pooled the technical replicates to increase the coverage. At a minimum coverage of 6×, we found that 28.9% of the assembled 283 842 RAD markers are polymorphic between the two populations due to a polymorphism affecting the enzyme recognition site.

Including only those positions in the reference with at least 6× coverage in each population sample, we scored 302 693 polymorphic sites, of which 148 770 (49.1%) were homozygous SNPs differentiating the two populations, and 153 923 (50.9) sites contained SNPs that were heterozygous in at least one population. We found 116 861 (41.2%) tags containing at least

one polymorphism of which 81 405 (28.7%) contained at least one homozygous SNP and 73 199 (25.9%) at least one heterozygous SNP, indicating that some tags can be either scored for a homozygous or heterozygous SNP.

Genetic studies on wild populations addressing questions about population structure and adaptation to different habitats are of great interest in guppies. Using the complete set of 302 693 SNPs, we estimated two important population parameters namely expected heterozygosity (H_e) (Excoffier, 2007) within each population and genetic differentiation measured by F_{ST} (Reich *et al.*, 2009) between the two population samples. We found $H_e = 0.078$ and $H_e = 0.138$ within Quare and Cumaná samples, respectively. These values are similar to a previous study using genome-wide SNP markers for population structure analysis (Willing *et al.*, 2010). However, the estimated F_{ST} of 0.71 is somewhat lower, perhaps due to the less biased choice of markers compared with the previous work, which used markers designed for mapping crosses, with fixed SNPs between the two populations being preferred over segregating ones, inflating the estimation of genetic differentiation between the two populations (Willing *et al.*, 2010). Consequently, our approach cannot only be used to identify SNPs for generating a high-density genetic map, but it will also produce a high number of unbiased informative SNPs that are ideal for population genetic analyses.

4 CONCLUSIONS

In this article, we have demonstrated a method for the *de novo* assembly and analysis of PE RAD-seq data. We were able to assemble ∼10% of the guppy genome represented by 283 842 RAD tags of which ∼50% were overlapping. This ratio could be significantly increased either by reducing the insert size of the library or by sequencing with longer read length. About 29% of the tags were polymorphic between the Quare and Cumana populations due to a disruption in the EcoRI recognition site and about 41% of the tags contained at least one SNP site. Estimated population parameters using these SNPs are similar to those previously reported, further confirming the veracity of our approach. We found that 81 405 of the tags contain homozygous SNP between Cumaná and Quare populations. These would be potentially useful in generating a dense genetic map that would greatly aid a whole genome assembly. Furthermore, the PE RAD-seq contigs could be used as artificial long reads in a whole genome assembly, to overcome the problems of assembling an entire genome from short reads only. Moreover, one could use different restriction enzymes to generate an overlapping set of RAD-seq contigs. By counting the restriction sites of 10 additional six-cutter enzymes in our assembled data (unpublished data of EMW), we saw that 167 848 tags contain at least 1 of 10 other restriction enzyme sites analyzed. Similar sequence complexity reduction approaches for aiding genome assemblies have been advocated before [e.g. Hyten *et al.* (2010)].

ACKNOWLEDGEMENTS

We thank Verena Kottler for critical proof reading and helpful suggestions and discussions on earlier versions of the manuscript. Furthermore, we would like to thank Stefan Henz for taking care of the guppy RAD tools project site.

Funding: Gottfried Wilhelm Leibniz Award of the Deutsche Forschungsgemeinschaft (to D.W.); Max Planck Society.

Conflict of Interest: none declared.

REFERENCES

Baird,N.A. *et al.* (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE*, **3**, e3376.

Bansal,V. *et al.* (2010) Accurate detection and genotyping of SNPs utilizing population sequencing data. *Genome Res.*, **20**, 537–545.

Craig,D.W. *et al.* (2008) Identification of genetic variants using bar-coded multiplexed sequencing. *Nat. Methods*, **5**, 887–893.

Davey,J.L. and Blaxter,M.W. (2010) RADSeq: next-generation population genetics. *Brief. Funct. Genomics*, **9**, 416–423.

Emerson,K.J. *et al.* (2010) Resolving postglacial phylogeography using high-throughput sequencing. *Proc. Natl Acad. Sci. USA*, **107**, 16196–16200.

Etter,P.D. *et al.* (2011) Local de novo assembly of RAD paired-end contigs using short sequencing reads. *PLoS ONE*, **6**, e18561.

Excoffier,L. (2007) Analysis of population subdivision. In Balding,D. *et al.* (eds) *Handbook of Statistical Genetics*. John Wiley & Sons, Ltd, West Sussex, p. 982.

Gnerre,S. *et al.* (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl Acad. Sci. USA*, **108**, 1513–1518.

Hohenlohe,P.A. *et al.* (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genet.*, **6**, e1000862.

Hohenlohe,P.A. *et al.* (2011) Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout. *Mol. Ecol. Resour.*, **11** (Suppl. 1), 117–122.

Hyten,D.L. *et al.* (2010) High-throughput SNP discovery through deep resequencing of a reduced representation library to anchor and orient scaffolds in the soybean whole genome sequence. *BMC Genomics*, **11**, 38.

Li,H. *et al.* (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.*, **18**, 1851–1858.

Lindholm,A. and Breden,F. (2002) Sex chromosomes and sexual selection in poeciliid fishes. *Am. Nat.*, **160** (Suppl. 6), S214–S224.

Magurran, A. (2005) *Evolutionary Ecology: The Trinidadian Guppy*. Oxford University Press, Oxford.

Malhis,N. and Jones,S.J. (2010) High quality SNP calling using Illumina data at shallow coverage. *Bioinformatics*, **26**, 1029–1035.

Pfender,W.F. *et al.* (2011) Mapping with RAD (restriction-site associated DNA) markers to rapidly identify QTL for stem rust resistance in Lolium perenne. *Theor. Appl. Genet.*, **122**, 1467–1480.

Reich,D. *et al.* (2009) Reconstructing Indian population history. *Nature*, **461**, 489–494.

Reznick,D. *et al.* (2001) Life-history evolution in guppies. VII. The comparative ecology of high- and low-predation environments. *Am. Nat.*, **157**, 126–140.

Schneeberger,K. *et al.* (2009) Simultaneous alignment of short reads against multiple genomes. *Genome Biol.*, **10**, R98.

Tripathi,N. *et al.* (2009) Genetic linkage map of the guppy, Poecilia reticulata, and quantitative trait loci analysis of male size and colour variation. *Proc. Biol. Sci.*, **276**, 2195–2208.

Willing,E.M. *et al.* (2010) Genome-wide single nucleotide polymorphisms reveal population history and adaptive divergence in wild guppies. *Mol. Ecol.*, **19**, 968–984.