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/weigel_databases/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 represented by 200–500 bp contigs associated to EcoRI restriction site. We were able to reconstruct one-tenth of the guppy genome 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 Cumaná, which have been previously used to generate a genetic linkage map (Tripathi et al., 2009). The guppy (Poecilia reticulata) 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/weigel_databases/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...
Table 1. Sequence information and read counts for each 12 bp MID

| MID         | Sample   | Million reads |
|-------------|----------|---------------|
| ATGTGTCGCCAA | 6 Quare males* | 4.6 |
| TCTGAGGTGACA | 6 Quare males* | 3.4 |
| GATCTGAGCTCTC | 6 Quare females | 0.015 |
| CGAGCAGACTGTG | 6 Cumana males | 5.1 |
| GTATCTGCACGAC | 6 Cumana males* | 4.4 |
| GACACGGTATGTC | 6 Cumana females* | 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, dsDNA 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.weighworld.de/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 (AATT,C). The guppy genome size is nearly 1 Gb as estimated by flow cytometry (M. Scharl, personal communication). Based on sequenced BAC ends from a genomic library of the Cumana guppy, we predicted the guppy genome to be relatively AT rich (60%), close to the AT content of the EcoRI recognition pattern (AATTTC). The guppy genome size is nearly 1 Gb as estimated by flow cytometry (M. Scharl, personal communication).

In order to generate a dense set of RAD markers, we chose the restrictions enzyme EcoRI, which recognizes the palindromic 6 bp sequence GAATTTC. The guppy genome size is nearly 1 Gb as estimated by flow cytometry (M. Scharl, personal communication).

In order to generate a dense set of RAD markers, we chose the restrictions enzyme EcoRI, which recognizes the palindromic 6 bp sequence GAATTTC. The guppy genome size is nearly 1 Gb as estimated by flow cytometry (M. Scharl, personal communication).

In order to generate a dense set of RAD markers, we chose the restrictions enzyme EcoRI, which recognizes the palindromic 6 bp sequence GAATTTC. The guppy genome size is nearly 1 Gb as estimated by flow cytometry (M. Scharl, personal communication).

In order to generate a dense set of RAD markers, we chose the restrictions enzyme EcoRI, which recognizes the palindromic 6 bp sequence GAATTTC. The guppy genome size is nearly 1 Gb as estimated by flow cytometry (M. Scharl, personal communication).

In order to generate a dense set of RAD markers, we chose the restrictions enzyme EcoRI, which recognizes the palindromic 6 bp sequence GAATTTC. The guppy genome size is nearly 1 Gb as estimated by flow cytometry (M. Scharl, personal communication).
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 gappy RAD tags, 51.4% were overlapping with their corresponding first read consensus, over a length of 29 bp and with an average
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 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 heterogamy (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 markers 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) contrasts, corresponding to ~149 female-specific tags and ~477 male-specific tags. Because 40% of these
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 sixfold. 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 Cumaná consensus as reference and for the Cumaná 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 Cumaná 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; Malik 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 Cumaná 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 homonymous 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.

Bass,W. 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.

Guerrero,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 three-spined 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,L. 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.

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

Mallozzi,N. and Jones,S.F. (2010) High-quality SNP calling using Illumina data at shallow coverage. Bioinformatics, 26, 1025–1035.

Pflueger,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.

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

Schnellerbergs,K. et al. (2009) Simultaneous alignment of short reads against multiple genomes. Genome Biol., 10, R56.

Tropu,N. et al. (2009) Genetic linkage map of the guppy, Poecilia reticulata, and quantitative trait loci analyses of male size and color 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.