MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2014) **14**, 953–965 doi: 10.1111/1755-0998.12245

The effects of read length, quality and quantity on microsatellite discovery and primer development: from Illumina to PacBio

NA WEI, JORDAN B. BEMMELS and CHRISTOPHER W. DICK
Department of Ecology and Evolutionary Biology, University of Michigan, 830 North University Avenue, Ann Arbor, MI 48109-1048, USA

Abstract

The advent of next-generation sequencing (NGS) technologies has transformed the way microsatellites are isolated for ecological and evolutionary investigations. Recent attempts to employ NGS for microsatellite discovery have used the 454, Illumina, and Ion Torrent platforms, but other methods including single-molecule real-time DNA sequencing (Pacific Biosciences or PacBio) remain viable alternatives. We outline a workflow from sequence quality control to microsatellite marker validation in three plant species using PacBio circular consensus sequencing (CCS). We then evaluate the performance of PacBio CCS in comparison with other NGS platforms for microsatellite isolation, through simulations that focus on variations in read length, read quantity and sequencing error rate. Although quality control of CCS reads reduced microsatellite yield by around 50%, hundreds of microsatellite loci that are expected to have improved conversion efficiency to functional markers were retrieved for each species. The simulations quantitatively validate the advantages of long reads and emphasize the detrimental effects of sequencing errors on NGS-enabled microsatellite development. In view of the continuing improvement in read length on NGS platforms, sequence quality and the corresponding strategies of quality control will become the primary factors to consider for effective microsatellite isolation. Among current options, PacBio CCS may be optimal for rapid, small-scale microsatellite development due to its flexibility in scaling sequencing effort, while platforms such as Illumina MiSeq will provide cost-efficient solutions for multispecies microsatellite projects.

Keywords: circular consensus sequencing, error trimming simulation, microsatellites, quality control, read length simulation, sequencing error simulation

Received 7 October 2013; revision received 19 February 2014; accepted 24 February 2014

Introduction

Microsatellites, also referred to as simple sequence repeats (SSRs) or short tandem repeats, are repetitive short DNA sequences that are scattered throughout the genomes of prokaryotes and eukaryotes (Morgante *et al.* 2002; Ellegren 2004). These molecular markers have seen extensive use in ecology and evolutionary biology (Provan *et al.* 2001; Schlotterer 2004; Selkoe & Toonen 2006). The dominance of microsatellites as the marker of choice for many applications in molecular ecology is, nevertheless, facing new challenges from large genomic data sets generated by next-generation sequencing (NGS) technologies (Ouborg *et al.* 2010). Yet, due to their hypervariability (Schlotterer 2000; Ellegren 2004), microsatellites remain invaluable for investigations of fine-scaled spatial demographic and genetic processes where individuals of interest are closely related, such as dispersal, parentage inference, pedigree reconstruction, linkage mapping and population structure (Selkoe & Toonen 2006; Guichoux *et al.* 2011; Haasl & Payseur 2011).

Interestingly, the advent of NGS may bolster microsatellite use because acquiring adequate genomic sequences from which microsatellites are retrieved is no longer technically and monetarily difficult. Instead, the bottleneck in microsatellite development is now the laborious and costly process of marker validation. Many researchers have advocated NGS-based microsatellite detection in nonmodel organisms (Abdelkrim *et al.* 2009; Gardner *et al.* 2011; Jennings *et al.* 2011), with the 454 and Illumina platforms dominating such efforts (reviewed in Zalapa *et al.* 2012). More recently, microsatellite detection has employed other NGS platforms, including Ion Torrent PGM (Huey *et al.* 2013; Elliott *et al.*

Correspondence: Na Wei, Fax: +1-734-763-0544;
E-mail: weina@umich.edu

in press), Illumina MiSeq (Nowak *et al.* 2014; McCracken *et al.* in press) and Pacific Biosciences (PacBio) RS (e.g. this study; Grohme *et al.* 2013). All these platforms can deliver hundreds to thousands of microsatellite loci per species, many more than identified using traditional methods (Zane *et al.* 2002), and with substantial reductions in time and capital investment.

The popularity of the 454 platform for microsatellite isolation owes primarily to its long read length sequencing (Zalapa *et al.* 2012). Long read length is advantageous in that it could benefit primer design by providing sufficient flanking regions (Guichoux *et al.* 2011; Zalapa *et al.* 2012). In addition, longer reads are suggested to allow better detection of genomic redundant sequences that contain low-complexity regions unfavourable for microsatellite amplification and interpretation (Elliott *et al.* in press). However, the 454 platform is economically inefficient (i.e. high cost per megabases; Glenn 2013), and involves laborious titration steps required in emulsion PCR to precisely link one DNA template to a single bead (Margulies *et al.* 2005). These aspects of the 454 platform eventually translate into a high total cost for microsatellite isolation. In terms of cost reduction, the most dramatic drop has been seen using the Illumina platform due to its high sequence throughput (Jennings *et al.* 2011; Castoe *et al.* 2012). Although the Illumina platform has much higher sequencing capacity relative to other platforms (Glenn 2013), it produces short reads (single-end up to 150 bp, paired-end up to 300 bp in GAIIx and HiSeq), except for the Illumina MiSeq sequencer, which can generate paired-end reads up to 600 bp (Illumina Incorporation 2013). The Ion Torrent platform represents an intermediate solution regarding the trade-off between (single-end) read length and read quantity (Glenn 2013), as well as the sequencing cost for microsatellite development (Jennings *et al.* 2011; Castoe *et al.* 2012; Elliott *et al.* in press).

Compared with the above NGS platforms, single-molecule real-time sequencing (SMRT; Eid *et al.* 2009) implemented on the PacBio RS system has the longest sequencing capability (Glenn 2013; Pacific Biosciences 2013), which offers potential advantages for microsatellite detection. The PacBio platform differs fundamentally from other platforms in that sequencing is performed on individual molecules without involving DNA amplification (e.g. emulsion PCR on 454 and Ion Torrent; Bridge PCR on Illumina; Glenn 2011), thereby resulting in a more uniform representation of genomic regions (Pacific Biosciences 2013). Although the long read length sequencing of PacBio comes with a high single-pass error rate (~11%; Pacific Biosciences 2013), improved base-calling accuracy is achieved by circular consensus sequencing (CCS), that is, reading through the same circular template DNA fragment multiple times (Travers *et al.* 2010). In addition, the insensitivity to various types of sequence context biases, such as homopolymers, GC-biased DNA regions and highly repetitive sequences (Eid *et al.* 2009; Quail *et al.* 2012; Zhang *et al.* 2012), makes PacBio a compelling alternative sequencing platform in this context. Success in microsatellite marker development using CCS has recently been reported on this platform (Grohme *et al.* 2013; Wainwright *et al.* 2013). However, an in-depth evaluation is not yet available regarding sequence characteristics of CCS and corresponding strategies of quality control for microsatellite development; therefore, providing this evaluation is the first objective of this study.

Independently from specific platforms and organisms, the development of microsatellite markers is in general influenced by read length, read quantity and read quality. Although the commonly agreed-upon benefits of long reads are conceptually straightforward, robust quantitative evidence for this consensus has been lacking. In addition, sequencing errors can undermine the efficiency of converting *in silico* loci into working markers, because unambiguous and unique sequences are crucial to the construction of amplifiable primers. However, most NGS-based microsatellite development work has been carried out in the absence of the inspection and control of sequence quality (for a counterexample, see Fernandez-Silva *et al.* 2013). It remains unclear the extent to which read quality inflicts a measurable effect on microsatellite marker development. Therefore, the second objective of this work is to provide a quantitative investigation of microsatellite development effectiveness in relation to read length, read quality and sequence quality control.

For the purpose of assessing the applicability of PacBio CCS in microsatellite isolation, we outline the process of (i) performing quality control (QC) on CCS reads, (ii) identifying microsatellite loci from post-QC CCS reads and (iii) validating microsatellite markers for three plant species for which no prior genomic information was available. For the second objective of quantifying how sequence characteristics limit microsatellite development, (iv) we conduct read length simulations to test whether increases in sequence length are associated with improvements in primer design success, microsatellite throughput and genomic redundancy detection; (v) we use sequencing error simulations to examine whether and how read quality affects microsatellite amplification; and (vi) we perform error trimming simulations to validate the need for sequence quality control in microsatellite development. Then, we use the findings from these simulations to guide the performance evaluation of PacBio CCS in comparison with other NGS platforms and highlight some key considerations for NGS use in microsatellite isolation.

Materials and methods

DNA sources and PacBio library preparation

We collected leaf tissues of three nonmodel tropical tree species from the 50-ha Forest Dynamics Plot (FDP) on Barro Colorado Island, Panama: *Alchornea costaricensis* Pax & K. Hoffm. (Euphorbiaceae), *Cecropia insignis* Liebm. (Cecropiaceae) and *Triplaris cumingiana* Fisch. & C.A. Mey. ex C.A. Mey. (Polygonaceae). Genomic DNA was isolated from freeze-dried leaves using DNeasy Plant Mini Kit (Qiagen, Valencia, CA, USA). DNA quality was checked using NanoDrop 2000 (Thermo Scientific, Wilmington, DE, USA), and dsDNA concentration was measured using Qubit® 2.0 Fluorometer (Invitrogen, Carlsbad, CA, USA). Double-stranded DNA of at least 30 ng/μL in a 50-μL volume from one tree of each species was sent to the DNA Sequencing Core Laboratory at the University of Michigan for PacBio 500-bp DNA library preparation and CCS.

In brief, genomic DNA was first sheared to fragments averaging 500 bp in length, and quantified with 2200 TapeStation using DNA 1k Tape (Agilent, Santa Clara, CA, USA). Sheared dsDNA was end repaired and ligated with hairpin adapters that contain a sequencing primer binding site to form the SMRTbell™ structure (i.e. two 55-nt single-stranded hairpin loops plus a dsDNA fragment). Unsuccessful ligation products were removed afterwards by exonuclease (ExoIII and ExoVII). Postligation products were quantified a second time with 2200 TapeStation, showing a mean fragment size of 363 bp for *A. costaricensis*, 487 bp for *C. insignis* and 445 bp for *T. cumingiana*. Then, the SMRTbell™ templates were annealed with sequencing primers and bound to biotinylated phi29 DNA polymerase mounted at the base of individual reaction chambers in SMRT cells. Nucleotide incorporation in a SMRT cell was monitored using 2 × 45-min collection mode. Four SMRT cells were run for each species on a PacBio RS sequencer using C2 chemistry. Fragments inserted between adapters of ≥3 × sequencing depths (including the sense and antisense strand) were retained for generating highly accurate adapter-free consensus sequences from CCS (referred to as CCS reads).

Quality control of CCS reads

Species-specific ccs.fastq files from four SMRT cells were combined to fetch CCS reads and the corresponding Phred +33 quality scores. The mean quality score of a CCS read was typically higher than 30 (median = 64, *A. costaricensis*; 62, *C. insignis*; 60, *T. cumingiana*; solid lines, Fig. S1), suggesting that trimming sequences based on average read quality would be ineffective (also see simulation results below). Thus, we removed terminal low-quality portions of each CCS read using a sliding-window approach implemented in MOTHUR v1.29.2 (Schloss *et al.* 2009). The window size was set to 10 bases, moving one base per step. The minimum window-wide mean quality score was set to 30, equivalent to an error-tolerance rate of 0.1%. If a window below this threshold was encountered, the CCS read was truncated from the last base in the window until the end of the read. We also filtered sequences according to homopolymer length. Some CCS reads contained homopolymers of 30–40 bases long, but more than 75% of CCS reads had homopolymers of ≤8 bases (Fig. S2). To retain adequate sequence numbers and eliminate long homopolymers, we omitted from further analyses CCS reads bearing a homopolymer longer than eight bases. Comparisons of pre-QC and post-QC base quality were visualized using the QRQC package (Buffalo 2012) in R v2.15.0 (R Development Core Team 2012).

Microsatellite identification and primer design

Circular consensus sequencing reads that passed the preceding quality control (referred to as trimmed CCS reads) were used to retrieve microsatellite loci. Perl pipelines in QDD v2.1 (Meglécz *et al.* 2010) were employed to automate the process of detecting microsatellites and designing primers. An initial purging step removed reads either too short (<80 bp) for successful primer design or holding microsatellite motifs of less than five repeats. The resulting microsatellite-containing sequences were screened for genomic redundancy (i.e. low-complexity regions and interspersed repeats) and sequencing redundancy (i.e. multiple copies of the same sequence) based on sequence similarities using BLAST v2.2.25 (Altschul *et al.* 1990) all-against-all pairwise alignments, in which microsatellites were soft masked. Once significant BLAST hits were discerned, the sequences with flanking region similarity less than 95%, probably resulting from genomic redundancy, were eliminated; those of ≥95% flanking region similarity were realigned by CLUSTALW2 (Larkin *et al.* 2007) to generate consensus sequences. The resulting nonredundant microsatellite-containing reads (i.e. singletons and unique consensus sequences) were used to locate appropriate priming regions using Primer3 (Rozen & Skaletsky 2000). Stringent primer-designing criteria (A + B design; definition *sensu* QDD program) were utilized as follows: (i) the absence of tandem repeats in priming regions and no homopolymers more than three bases; (ii) no multiple microsatellites in the target region; (iii) primer size between 18 and 22 bases; (iv) PCR product of 100–500 bp; (v) optimal GC content of 50% (range 40–60%); (vi) 57–63 °C melting temperature with a

maximum intrapair difference of 5 °C; (vii) maximum self-complementarity score of 3; and (viii) presence of one GC clamp.

Microsatellite marker validation

We prioritized the test array of microsatellite markers as follows: for tri- to hexanucleotide repeat motifs, the number of repeat units is >7; for a dinucleotide repeat motif, there are at least seven or eight repeats depending on species; no compound repeat motif is allowed. In total, we synthesized 59 primer pairs for *A. costaricensis*, 69 for *C. insignis* and 62 for *T. cumingiana*. For each species, we first screened the markers in three individuals. If more than one allele was present at the focal locus, we then assessed marker polymorphism in nine more individuals collected from the same population in the 50-ha FDP. We defined successful amplification as consistently resulting in easily interpretable allelic patterns, and polymorphism as possessing at least two alleles. We used a fluorescently labelled M13 primer coupled with M13-tagged microsatellite primers in individual PCRs as detailed previously (Wei *et al.* 2013). PCRs were performed using a touchdown protocol of an initial denaturation at 94 °C for 4 min; 28 cycles of 94 °C for 30 s, 59 °C (a decrement of 0.2 °C per cycle) for 40 s and 72 °C for 60 s; 10 cycles of 94 °C for 30 s, 53 °C for 40 s and 72 °C for 60 s; and a final extension at 72 °C for 10 min. Amplicons were sized in ABI 3730 DNA Analyzer (Applied Biosystems, Carlsbad, CA, USA), and scored using GENEMARKER version 1.7 (Softgenetics, State College, PA, USA).

Read length simulations

In the simulations of read length effect on microsatellite detection, reads of uniform length at each of 100, 150, 200, 250, 300, 350, 400, 500, 600, 700, 800, 1000 and 1200 bp were drawn at random from *Populus trichocarpa* chromosomes 1–19 (v3.0, DOE-JGI, http://www.phytozome.net/poplar; Tuskan *et al.* 2006) using 0.1× coverage (~40 Mb). This equal genome coverage ensures that the ability to locate microsatellites, eliminate genomic redundancy and design suitable primers for individual loci depends only on how long the reads are, rather than how much sequencing effort was exerted in individual simulations. We conducted platform-independent read length simulations by allowing no sequencing errors in reads using GRINDER v0.5.3 (Angly *et al.* 2012). In addition, we incorporated sequencing errors into read length simulations using PBSIM v1.0.3 (Ono *et al.* 2013) with built-in PacBio CCS error profiles (substitutions/insertions/deletions ratio of 6:21:73; read accuracy of 98 ± 2%). Both error-free and error-embedded simulated reads were used directly (i.e. no quality control) to detect microsatellites and design primers, following the above-described procedure except using a relaxed primer GC content of 30–70% (also for the following simulations).

In the situation of equal genome coverage, there exists a balance between read quantity and read length to maintain the total sequence bases, that is, libraries of longer reads contain fewer reads (Tables S1 and S2). To relax the equal genome coverage assumption, we further used equal read quantity in each simulation. To do so, microsatellite parameters (e.g. microsatellite-containing reads, microsatellite loci; Tables S1 and S2) were converted to relative estimates by dividing by the corresponding total read numbers in individual simulations and then we multiplied these relative estimates by the same read quantity of 160 000.

Sequencing error simulations

Sequencing errors of substitutions and indels (insertions and deletions) were introduced to reads of uniformly 350 bp simulated from the reference genome of *P. trichocarpa* at 0.1× coverage. Taking into account potential effects of sequencing error types on simulated results, we considered both substitution-biased (substitutions/indels ratio of 90:10) and indel-biased (substitutions/indels ratio of 10:90) sequencing errors. In terms of sequencing error rate distribution, we assumed that sequencing errors occurred either uniformly or linearly from the 5′ end to 3′ end of each read. With uniformly distributed sequencing errors, reads were simulated with an error rate of 0, 0.01%, 0.1%, 0.5%, 1%, 2%, 3% and 5%. With linearly distributed errors, the error rate doubled from the 5′ end to 3′ end: 0, 0.01–0.02%, 0.1–0.2%, 0.2–0.4%, 0.5–1%, 1–2% and 2–4%. To check the extent to which sequencing errors impair the amplification of microsatellite markers, designed primer pairs (A + B design) from the simulated error-containing reads were aligned back to the reference genome of *P. trichocarpa* using iPCRess implemented in EXONERATE v2.2 (Slater & Birney 2005). Successful *in silico* locus amplification was defined conservatively as having unique and perfect (i.e. zero mismatch) alignment between the reference genome and the forward and reverse primer.

Error trimming simulations

To investigate whether sequence quality control is essential for microsatellite development, we compared the rate of *in silico* locus amplification between simulated sequence libraries that were treated with different quality control criteria. We simulated reads of uniform length of 350 bp from the *P. trichocarpa* genome (0.1× coverage) using PBSIM, following the observed read length

distribution and quality profiles of *T. cumingiana* CCS reads in this study. Then, two types of quality control were used to filter the simulated sequences. The first QC method was based on mean read quality, requiring a minimum average read quality score of 30, as well as no reads containing homopolymers longer than eight bases. The second QC method was based on the sliding-window approach as described above, including the control of homopolymers. Microsatellite detection and primer design were conducted on both post-QC reads and raw reads. Designed microsatellite primers (A + B type) were tested *in silico* for locus amplification.

Results

Sequencing capacity of PacBio 500-bp CCS

Pacific Biosciences CCS of 500-bp genomic DNA inserts returned on average 161 000 CCS reads using four SMRT cells (Table 1). Among species, the number of CCS reads varied (one-way ANOVA, $F_{2,9} = 7.273$, $P < 0.05$; Table 1); *A. costaricensis* yielded fewer CCS reads ($n = 105\,881$; Table 1) relative to *C. insignis* (198 989; Holm's adjusted $P < 0.05$ for pairwise *t*-tests) and *T. cumingiana* (178 122; Holm's adjusted $P < 0.05$), whereas the difference between the latter two was negligible (Holm's adjusted $P = 0.436$). At a per-SMRT-cell scale, the number of CCS reads ranged from 19 801 to 57 046 (mean = 40 249; Fig. S3), species identity notwithstanding. The frequency distribution of CCS read lengths revealed a wide size range (11–1391 bp, *A. costaricensis*; 9–1751 bp, *C. insignis*; 12–1917 bp, *T. cumingiana*; Fig. 1a), but on average, only 0.01% (0.03%, *A. costaricensis*; 0.003%, *C. insignis*; 0.001%, *T. cumingiana*) of CCS reads were shorter than 80 bp, the minimum read length required in the QDD program for microsatellite detection.

Quality control of CCS reads

Mean sequence quality of CCS reads was greatly augmented after QC (*A. costaricensis*, one-sided Wilcoxon rank sum test, $W = 4.75 \times 10^9$, Holm's adjusted $P < 0.001$; *C. insignis*, $W = 1.62 \times 10^{10}$, adjusted $P < 0.001$; *T. cumingiana*, $W = 1.62 \times 10^{10}$, adjusted $P < 0.001$). The minimum mean quality score of post-QC CCS reads (20, *A. costaricensis*; 25, *C. insignis*; 23, *T. cumingiana*; dotted lines, Fig. S1) was nearly double that of raw CCS reads (14, 14, and 13, respectively; solid lines, Fig. S1). CCS read quality was negatively correlated on a log-log scale with read length before QC (Fig. S4), but was positively correlated after QC (Fig. S5).

In addition, QC improved base accuracy, as the percentiles and the mean of base quality scores were increased (Fig. 2). Although base accuracy declined along the length of a CCS read both prior to QC (*A. costaricensis*, $F_{1,1389} = 1.40 \times 10^4$, $P < 0.001$, adjusted $R^2 = 0.910$; *C. insignis*, $F_{1,1749} = 2.24 \times 10^3$, $P < 0.001$, adjusted $R^2 = 0.561$; *T. cumingiana*, $F_{1,1915} = 3.11 \times 10^4$, $P < 0.001$, adjusted $R^2 = 0.942$) and after QC (*A. costaricensis*, $F_{1,879} = 383.6$, $P < 0.001$, adjusted $R^2 = 0.303$; *C. insignis*, $F_{1,1042} = 491.5$, $P < 0.001$, adjusted $R^2 = 0.320$; *T. cumingiana*, $F_{1,979} = 773.6$, $P < 0.001$, adjusted $R^2 = 0.441$), the fitted slope of post-QC base quality with read base position was significantly smaller than the

Table 1 Sequencing capacity and microsatellite throughput of 500-bp genomic shotgun circular consensus sequencing using four SMRT cells per species.

	Alchornea costaricensis (Euphorbiaceae)	*Cecropia insignis* (Cecropiaceae)	*Triplaris cumingiana* (Polygonaceae)
Sequence megabases (Mb)	31.4	70.2	65.2
Number of reads			
CCS reads	105 881	198 989	178 122
Trimmed CCS reads	95 265	177 161	157 026
Average read length (bp)			
CCS reads	297	353	366
Trimmed CCS reads	201	225	222
Microsatellite detection			
Nonredundant SSR-containing CCS reads	5433	6212	5793
Nonredundant SSR-containing trimmed CCS reads	3146	3072	3001
SSR loci (≥5 repeats; primer design A + B)	390	512	795

CCS, circular consensus sequencing; CCS reads, adapter-free consensus sequences generated by CCS; trimmed CCS reads, CCS reads passing quality control.

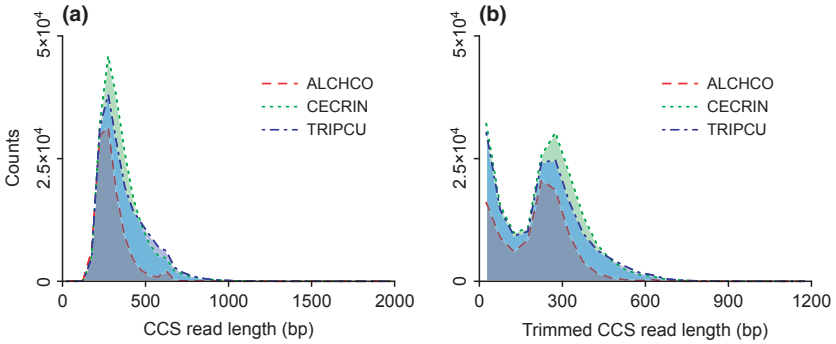

Fig. 1 Frequency distribution of CCS read lengths generated by 500-bp genomic shotgun circular consensus sequencing. CCS, circular consensus sequencing; CCS reads, adapter-free consensus sequences generated by CCS; trimmed CCS reads, CCS reads passing quality control. ALCHCO, *Alchornea costaricensis*; CECRIN, *Cecropia insignis*; TRIPCU, *Triplaris cumingiana*.

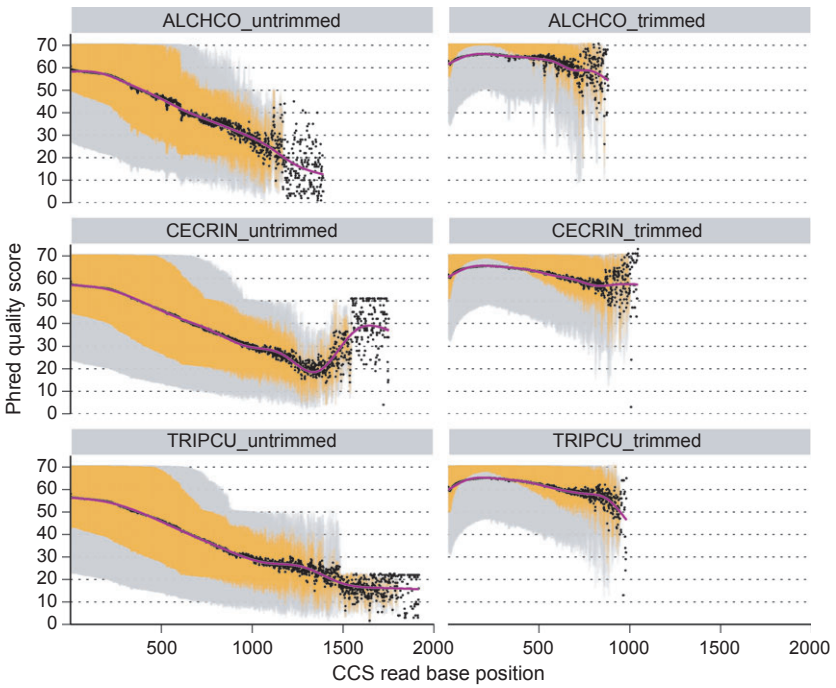

Fig. 2 Base quality scores of CCS reads before (untrimmed) and after (trimmed) quality control. Outer whiskers (grey regions) represent the 10th to the 90th percentile of position quality scores; inner whiskers (orange regions) represent the 25th to the 75th percentile; dots are the mean quality score at each base position; lines are fitted generalized additive model (GAM) smooth lines. ALCHCO, *Alchornea costaricensis*; CECRIN, *Cecropia insignis*; TRIPCU, *Triplaris cumingiana*. CCS, circular consensus sequencing.

pre-QC fitted slope (*A. costaricensis*, slope $\beta_{\text{post-QC}} = -0.010$, $\beta_{\text{pre-QC}} = -0.036$, one-sided Welch's *t*-test, $t = 1333.8$, d.f. = 1264, $P < 0.001$; *C. insignis*, $\beta_{\text{post-QC}} = -0.010$, $\beta_{\text{pre-QC}} = -0.018$, $t = 508.1$, d.f. = 1958, $P < 0.001$; *T. cumingiana*, $\beta_{\text{post-QC}} = -0.013$, $\beta_{\text{pre-QC}} = -0.025$, $t = 821.8$, d.f. = 1078, $P < 0.001$).

A positive correlation was found between homopolymer length and square root-transformed raw CCS read length (Fig. S6). But homopolymer lengths were appreciably reduced after QC (*A. costaricensis*, Wilcoxon rank sum test, $W = 6.50 \times 10^9$, Holm's adjusted $P < 0.001$; *C. insignis*, $W = 2.28 \times 10^{10}$, adjusted $P < 0.001$; *T. cumingiana*, $W = 1.86 \times 10^{10}$, adjusted $P < 0.001$). Furthermore, we found no effect of QC on position GC content of CCS reads; pre-QC sequence position GC content averaged between 36.6% and 39.8%, and post-QC between 38.2% and 38.7% (Fig. S7).

Quality control filtered out approximately 10% of CCS reads (Table 1). Remaining CCS reads were significantly shortened (*A. costaricensis*, Wilcoxon rank sum test, $W = 7.28 \times 10^9$, Holm's adjusted $P < 0.001$; *C. insignis*, $W = 2.60 \times 10^{10}$, adjusted $P < 0.001$; *T. cumingiana*, $W = 2.10 \times 10^{10}$, adjusted $P < 0.001$; Table 1), with a mean read length reduction by 32–39%. Post-QC CCS reads were bimodally distributed (Fig. 1b), in which on average, 76% were longer than 80 bp.

Microsatellite detection

Without quality control, approximately 5400 to 6200 nonredundant microsatellite-containing sequences were retrieved in individual species (Table 1), corresponding to 3.1–5.1% of raw CCS reads. With quality control, the nonredundant microsatellite-containing sequences

decreased to around 3000 (Table 1), accounting for 1.7–3.3% of raw CCS reads. Selected from the nonredundant microsatellite-containing trimmed CCS reads, microsatellite loci (A + B design) varied from 390 in *A. costaricensis* to 512 in *C. insignis* and 795 in *T. cumingiana* (Table 1). These loci accounted for 12.4–26.5% of the nonredundant microsatellite-containing trimmed CCS reads, and 0.3–0.5% of raw CCS reads. With respect to repeat motifs, dinucleotide motifs were most abundant (69–77% of all repeat motifs), followed by trinucleotide motifs (21–26%; Fig. 3). Other repeat motifs collectively constituted <5%.

We also checked the extent of microsatellite throughput reduction resulting from QC, by comparing the number of microsatellite loci retrieved from trimmed CCS reads with those retrieved from raw CCS reads. The ratio of post-QC microsatellite loci to pre-QC microsatellite loci (pre-QC $n = 663$, *A. costaricensis*; 1024, *C. insignis*; 1534, *T. cumingiana*) averaged 54% (range 52–59%).

Microsatellite marker validation

For *A. costaricensis*, 59 microsatellite markers were inspected for locus amplification and polymorphism, of which 62.7% ($n = 37$) were amplifiable and 42.4% ($n = 25$) were polymorphic. Likewise, the amplification success in *C. insignis* reached 73.9% (51 of 69); polymorphic loci accounted for 39.1% (27 of 69). In *T. cumingiana*, the rate of locus amplification and polymorphism was 62.9% (39 of 62) and 45.2% (28 of 62), respectively. On average, irrespective of species identity, 66.8% of the screened microsatellite markers were successfully amplified, whereas 42.1% exhibited polymorphism. Further details about the informativeness of species-specific microsatellite markers will be provided elsewhere.

Read length simulations

Read length simulations examined the relationship between read length and microsatellite isolation effectiveness, regarding (1) the likelihood of finding shotgun reads that carry microsatellites, (2) the ability to detect genomic redundancy from microsatellite-containing reads, (3) the success of designing primers and (4) the amount of putative microsatellite loci. First, the percentage of microsatellite-containing reads, relative to total simulated reads, increased in proportion to read length (Pearson correlation coefficient $r = 0.998$ for both error-free and error-bearing reads; Tables S1 and S2). For instance, a twofold increase in read length, such as from 200 bp to 400 bp, resulted in a nearly twofold increase in the proportion of reads containing microsatellites (from 5.1% to 10.1%, error-free reads; from 4.8% to 9.2%, error-bearing reads). Second, with respect to genomic redundancy detection under equal genome coverage, the proportion of grouped sequences that have low levels of similarity (<95%) and thus are not regarded as from the same locus (Meglécz *et al.* 2010), increased by two orders of magnitude with read length from 0.14% at 100 bp to 24.2% at 1200 bp, relative to all microsatellite-containing reads; multihit sequences that contain interspersed repetitive regions increased from 0% at 100 bp to 0.7% at 300 bp, and to 15.6% at 1200 bp, in the situation of no sequencing errors (Fig. 4c). A similar magnitude of increase in detectable genomic redundancy was observed when sequencing errors were considered (from 0.1% to 25.1%, grouped sequences; from 0% to 12.9%, multihit sequences; Fig. 4c). Third, the rate of primer design (A + B type) for nonredundant microsatellite-containing reads increased 50-fold from 100 bp (0.3%, error-free reads; 0.4%, error-bearing reads) to 400 bp (18.4%, error-free reads; 17.2%, error-bearing reads), and

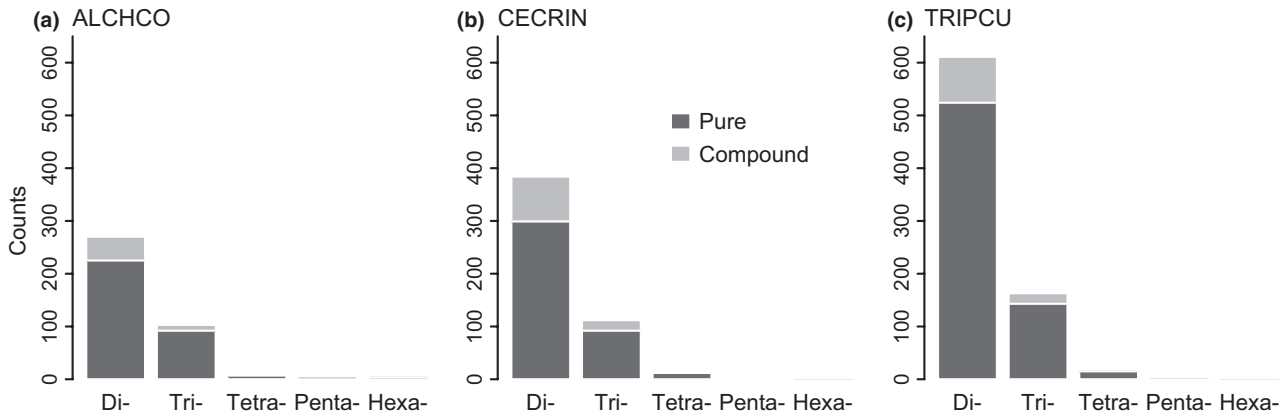

Fig. 3 Motif length-specific microsatellite loci identified from quality-controlled CCS reads. See text for microsatellite searching and primer design criteria. ALCHCO, *Alchornea costaricensis*; CECRIN, *Cecropia insignis*; TRIPCU, *Triplaris cumingiana*; CCS, circular consensus sequencing.

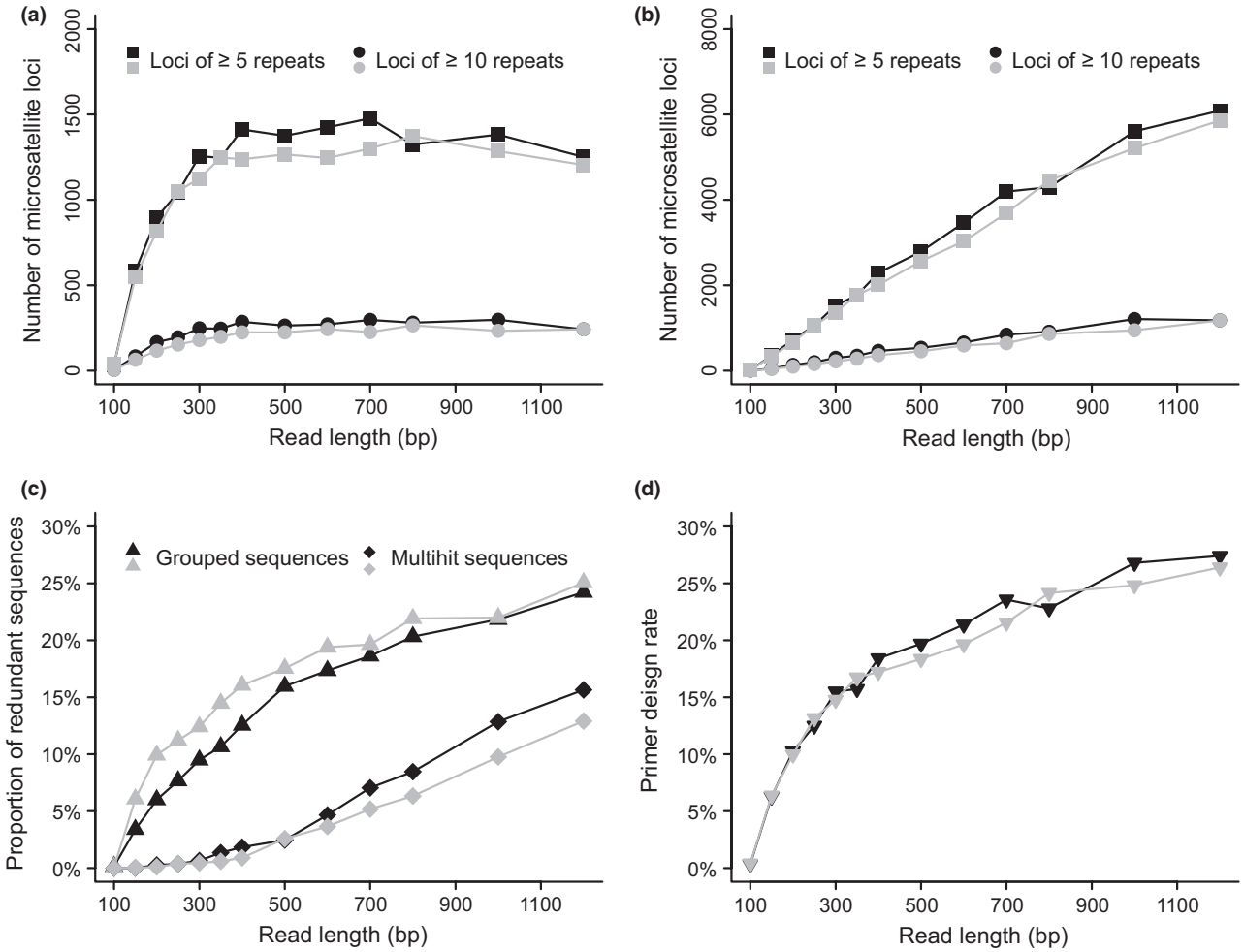

Fig. 4 The effects of read length on microsatellite yield (a–b), genomic redundancy detection (c) and primer design success rate (d). (a–b) Microsatellite loci (primer design A + B, see text) were retrieved from simulated sequences assuming (a) equal genome coverage of 0.1× and (b) equal read numbers of 160 000 for each read length size from the *Populus trichocarpa* genome. (c) The proportions of grouped and multihit sequences are relative to total microsatellite-containing reads. (d) Primers (A + B type) were designed for nonredundant microsatellite-containing reads. Black symbols indicate the absence of sequencing errors in simulations; grey symbols indicate the inclusion of sequencing errors (read accuracy of 98 ± 2%).

80-fold to 1200 bp (27.4% and 26.4% in error-free and error-bearing reads, respectively; Fig. 4d). Lastly, with respect to microsatellite throughput, the number of microsatellite loci (≥5 repeats and ≥10 repeats; A + B design) responded positively to read length until approximately 400 bp, after which relationships became nearly asymptotic, for both error-free and error-bearing reads under equal genome coverage (Fig. 4a). But with equal read quantity rather than equal genome coverage, the measures of microsatellite yield increased monotonically with read length (e.g. in error-free reads, loci of ≥5 repeats and A + B design, linear regression slope $\beta = 5.76$, $F_{1,11} = 583.7$, $P < 0.001$, adjusted $R^2 = 0.980$; loci of ≥10 repeats and A + B design, $\beta = 1.18$, $F_{1,11} = 364.1$, $P < 0.001$, adjusted $R^2 = 0.968$; Fig. 4b).

Sequencing error simulations

Reductions in microsatellite amplification success were associated with increased sequencing error rate, irrespective of sequencing error type and error rate distribution (Fig. 5). When error rate was 0 (i.e. no sequencing errors), 93.6% of microsatellite loci (primer design A + B) recovered from simulated shotgun sequences amplified *in silico*. Compared against this baseline, reductions in locus amplification became detectable when error rate increased to 0.1% given uniformly distributed base accuracy (indel bias, amplification rate = 89.4%, one-sided proportion test, $\chi^2 = 13.80$, d.f. = 1, Holm's adjusted $P < 0.001$; substitution bias, 88.9%, $\chi^2 = 16.96$, d.f. = 1, adjusted $P < 0.001$), and to

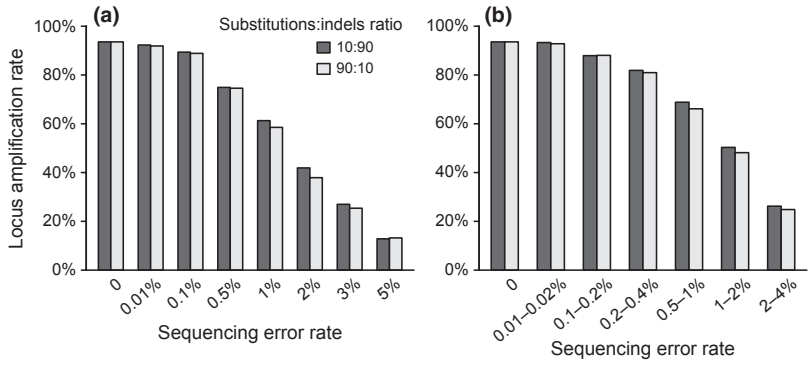

Fig. 5 Simulations of microsatellite amplification rate in relation to sequencing errors. Sequencing errors of substitutions and indels (insertions and deletions) were introduced to simulated reads of 350 bp (a) uniformly or (b) linearly increasing from the 5′ end to 3′ end from the *Populus trichocarpa* genome with 0.1× coverage. Substitution- and indel-dominated sequencing errors are represented by light bars and dark bars, respectively.

0.1–0.2% given linearly distributed base accuracy (indel bias, 87.9%, $\chi^2 = 23.54$, d.f. = 1, adjusted $P < 0.001$; substitution bias, 88.1%, $\chi^2 = 22.94$, d.f. = 1, adjusted $P < 0.001$). For reads of uniform 1% per-base error rate, approximately 60% of microsatellite loci were amplifiable (61.3%, indel bias; 58.5%, substitution bias) and 40% (41.9%, indel bias; 37.9%, substitution bias) for reads of 2% per-base error rate (Fig. 5a). With a linearly distributed sequencing error rate, approximately 68% of microsatellite loci amplified *in silico* at 0.5–1% error rate and 49% at 1–2% error rate (Fig. 5b).

Error trimming simulations

The *in silico* experiment of the effect of error trimming on microsatellite amplification was conducted on simulated reads that closely mimicked the characteristics of observed CCS reads (of *T. cumingiana*) in this study. Quality control based on the sliding-window method reduced the test array of microsatellite loci (primer design A + B) by 57% from 1259 to 539, whereas QC based on mean read quality reduced microsatellite loci by 23% to 970. The amplification rate of microsatellite loci with sliding-window-based QC was 60.7% (Fig. 6), significantly higher than that without QC (53.9%; one-sided proportion test, $\chi^2 = 6.838$, d.f. = 1, Holm's adjusted $P < 0.05$) and that with mean read-quality-based QC (55.3%; $\chi^2 = 3.924$, d.f. = 1, adjusted $P < 0.05$). When comparing the locus amplification rate of simulated reads (60.7%) with that of observed reads in *T. cumingiana* (62.9%) based on the same method of QC, no significant difference was detected ($\chi^2 = 0.042$, d.f. = 1, $P = 0.419$).

Discussion

By elaborating the workflow from sequence quality control to marker validation, we demonstrate the effectiveness of shotgun genome CCS for isolating microsatellites in nonmodel plant species. On average, approximately

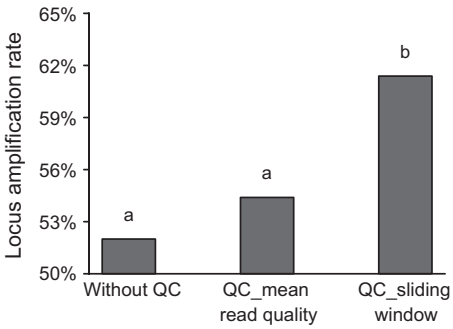

Fig. 6 The effect of quality control on microsatellite locus amplification. *In silico* microsatellite amplification rate is significantly higher with quality control based on a sliding-window approach (QC_sliding window) relative to that without QC, as well as that with QC based on mean read quality score (QC_mean read quality).

160 000 CCS reads were acquired using four SMRT cells for each species. Quality control reduced microsatellite throughput by *ca*. 50%, but several hundred microsatellite loci were still obtained per species. These loci are also expected to have higher amplification success than the total pool of microsatellite loci before quality control, as indicated by the error trimming simulations. The initial marker screening revealed that two-thirds of the loci consistently resulted in easily interpretable amplicons, and two-fifths of the loci were polymorphic. Here, we discuss the performance of PacBio CCS in comparison with other NGS platforms for microsatellite isolation, in the context of read length, read quality and sequence quality control.

Our read length simulations substantiate the postulated importance of read length in microsatellite development. Long reads increase the likelihood that a sequence contains microsatellites (Tables S1 and S2) by searching through more bases in a genomic location. They also increase the probability that a sequence contains intact microsatellites with sufficient flanking regions (Fig. 4d), because a microsatellite is less likely to be located in proximity to either of the two ends in a

longer read (Abdelkrim *et al.* 2009). A primer-design success rate of up to 33% was empirically predicted for reads averaging 200 bp (reviewed in Guichoux *et al.* 2011), which coincides with that predicted by simulations here, that is, *ca*. 25% primer design rate for nonredundant microsatellite-containing reads under relaxed criteria (primer design A–G; data not shown). Furthermore, our read length simulations provide platform-independent evidence of improved genomic redundancy detection with increases in read length under equal genome coverage (Fig. 4c). This finding compliments previous inferences by Elliott *et al.* (in press) based on the comparison between Ion Torrent- and 454-specific read lengths.

Notwithstanding the aforementioned benefits of longer reads, increases in read length do not result in a continuing increase in microsatellite yield, when total sequencing effort is held constant. The relationship between the number of microsatellite loci and read length predicts a threshold read length of approximately 400 bp (Fig. 4a; based on visual inspection). Above the threshold, read length is not a limiting factor for microsatellite throughput: any potential gains in the number of microsatellite loci, due to increased probability of reads containing microsatellites and increased primer design success, are offset by the losses resulting from decreased read numbers and an increased portion of unused redundant reads (e.g. grouped and multihit sequences). Nonetheless, microsatellite loci recovered from longer reads (e.g. 1000 bp) may have a higher chance of successful amplification than loci from reads of 400 bp, because of more effective genomic redundancy removal. An additional comparison of *in silico* microsatellite amplification indeed revealed a small yet significant increase by *ca*. 3% when read length was increased from 400 to 1000 bp (data not shown).

Microsatellite yield behaves as a threshold response to read length in the context of equal sequencing depth. However, NGS platforms differ in sequencing throughput; therefore, the estimation of platform-dependent microsatellite yield needs to consider both read length and read numbers without the constraint of equal genome coverage or equal read quantity. Despite amassed NGS-based microsatellite data sets, cross-platform comparisons of the number of microsatellite loci, based on these empirical investigations, are complicated by heterogeneities in genomic microsatellite frequency and genome size among taxa (Toth *et al.* 2000; Morgante *et al.* 2002; Ellegren 2004), and in microsatellite searching and primer design criteria. Therefore, we base this assessment on our simulations, taking into account platform-dependent read numbers and mean read length (Table 2). Specifically, we multiplied platform-specific read numbers by the estimated proportion of reads containing microsatellite loci (primer design A + B; Table S1) for the corresponding mean read length of individual NGS platforms. Despite the use of uniform read length rather than platform-dependent read length distributions, our simulations provide good predictions of microsatellite throughput, as evidenced by the concordance between simulated and observed number of microsatellite loci on PacBio and Illumina MiSeq (Table 2). The discrepancy between simulations and empirical findings on Ion Torrent and 454 may result primarily from low microsatellite density in targeted organism genomes (Elliott *et al.* in press), as noted by the authors. In general, PacBio and Ion Torrent produce a comparable number of microsatellite loci relative to 454 but with a *ca*. 50% reduction

Table 2 Cross-platform comparisons of next-generation sequencing (NGS) use in microsatellite development. Platform-specific information of read length, read quantity and observed microsatellite loci are from this study on PacBio, Elliott *et al.* (in press) on Ion Torrent and 454, and Nowak *et al.* (2014) on Illumina MiSeq. Predicted microsatellite loci are calculated based on read length simulations according to platform-specific mean read length and read quantity, assuming no sequencing errors

NGS platform	PacBio CCS	Ion Torrent PGM	454 GS-FLX	Illumina MiSeq
Sequencing unit	4 SMRT cells	1 '316' chip	1/8 PTP	1 PE 250 bp
Average read length (bp)	350	150	350	400*
Number of reads	160 000	1 000 000	150 000	6 300 000
Sequencing cost†($)	~1000	~1000	~2000	~1400
Predicted SSR loci of ≥5 repeats (primer design A + B)	1769	2213	1658	90 194
Observed SSR loci of ≥ 5 repeats	1645‡	413	165	81 886
Predicted SSR loci of ≥10 repeats (primer design A + B)	349	319	327	18 269

PTP, PicoTiterPlate; PE, paired-end; SMRT, single-molecule real-time sequencing.

*Predicted mean contig length of MiSeq paired-end 250 bp sequencing (observed contig lengths were not reported in the original data).

†Sequencing cost including library preparation from Glenn (2013).

‡Observed SSR loci of ≥5 repeats (primer design A + B; default parameters used in the QDD program) retrieved from CCS reads without quality control.

in cost, but Illumina MiSeq generates approximately 30 times more microsatellite loci than PacBio and Ion Torrent at the same total cost (Table 2). Nevertheless, all the NGS platforms are able to deliver thousands of microsatellite loci, far more than a project could practically screen and genotype.

As *in silico* locus acquisition is no longer the bottleneck for microsatellite development, the efficiency of converting these loci into functional makers is of equivalent, if not greater, importance relative to the initial sequencing step. This consideration is particularly salient in light of the distinct effects that sequencing errors have on microsatellite yield and microsatellite amplification. Microsatellite yield is little affected by the presence of sequencing errors, as error-bearing (read accuracy of 98 ± 2%) and error-free reads of the same read length were able to retrieve a similar number of microsatellite loci (Fig. 4a). However, the amplification rate of these microsatellite loci plummets when sequencing errors are present (Fig. 5). In practice, all NGS platforms produce sequencing errors but to varying degrees, such as 1.07% reported for 454 GS-FLX Titanium (Gilles *et al.* 2011), 1.71% for Ion Torrent PGM and 0.80% for Illumina MiSeq (Quail *et al.* 2012). Given these error rates, approximately 50–68% of microsatellite loci are predicted to be able to amplify unique and interpretable PCR products, according to our simulations; this estimate is consistent with empirical findings (Castoe *et al.* 2012; Fernandez-Silva *et al.* 2013; Wei *et al.* 2013).

One important implication of the adverse effects of sequencing errors on microsatellite amplification is that quality control is essential for the consideration of cost- and labour-effective marker validation. In this study, significant improvement in microsatellite amplification was achieved by quality control using a sliding-window approach (Fig. 6). Although this finding is based on PacBio CCS-dependent error trimming simulations, it can also apply to sequences from other NGS platforms, as base-calling accuracy in general declines with base position (Gilles *et al.* 2011; Loman *et al.* 2012). One question regarding quality control concerns the possible negative effect of shortened read lengths after error trimming. By comparing the *in silico* amplification rate of microsatellite loci retrieved from error-bearing reads of 1000 bp with that of loci from error-free reads of 200 bp (Fig. 4a), we found the amplification rate was twofold higher when errors were absent, despite a fivefold decrease in read length (data not shown). This finding suggests that the importance of read quality outweighs that of read length in terms of obtaining successfully amplifiable microsatellite loci. The methods of quality control may vary between platforms that emphasize long read length (e.g. 454 and PacBio) and those that emphasize high throughput (e.g. Illumina MiSeq and Ion Torrent). The sliding-window-based quality control described in this study can be used for 454 and PacBio, because this approach shortens read lengths but is unlikely to result in a substantial reduction in sequence quantity. Meanwhile, more stringent quality control, such as removing reads of per-base quality score below 30, can be afforded for Illumina MiSeq or Ion Torrent because of high sequence throughput.

Conclusion

This study provides a quantitative demonstration of microsatellite development in relation to sequence attributes, based on which the performance of PacBio CCS is evaluated in comparison with other NGS platforms. PacBio CCS is suitable for fast, small-scale microsatellite development due to its flexibility in scaling sequencing effort, in terms of the number of SMRT cells utilized per project. A single SMRT cell can potentially deliver enough functional microsatellite markers (Grohme *et al.* 2013; Wainwright *et al.* 2013), at a sequencing cost of ~$200 (not including library preparation). On the other hand, Illumina MiSeq paired-end sequencing can be particularly cost-efficient when greater sequencing effort is required, such as for multispecies microsatellite projects, as well as for organisms that have low genomic microsatellite density. In light of the continuing advances in sequence length on all the platforms, read length may not be the primary concern for NGS use in microsatellite isolation. Instead, sequencing accuracy and the corresponding strategies of quality control are essential for time- and cost-effective microsatellite isolation.

Acknowledgements

We are grateful to two anonymous reviewers and the editor Travis Glenn for providing invaluable feedback on the manuscript. We thank the Smithsonian Tropical Research Institute and Center for Tropical Forest Science for facilitating fieldwork on Barro Colorado Island. This work was supported by the Rackham Graduate Student Research Grant to N.W. and a grant to C.W.D. from the College of Literature, Science and the Arts at the University of Michigan. N.W. was supported by a Barbour Scholarship from the University of Michigan, and J.B.B. was supported by an NSF Graduate Research Fellowship.

References

Abdelkrim J, Robertson BC, Stanton JAL, Gemmell NJ (2009) Fast, cost-effective development of species-specific microsatellite markers by genomic sequencing. *BioTechniques*, **46**, 185–191.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.

Angly FE, Willner D, Rohwer F, Hugenholtz P, Tyson GW (2012) Grinder: a versatile amplicon and shotgun sequence simulator. *Nucleic Acids Research*, **40**, e94.

Buffalo V (2012) *qrqc: Quick Read Quality Control*. R package version 1.10.0. http://github.com/vsbuffalo/qrqc.

Castoe TA, Poole AW, de Koning APJ *et al.* (2012) Rapid microsatellite identification from Illumina paired-end genomic sequencing in two birds and a snake. *PLoS One*, **7**, e30953.

Eid J, Fehr A, Gray J *et al.* (2009) Real-time DNA sequencing from single polymerase molecules. *Science*, **323**, 133–138.

Ellegren H (2004) Microsatellites: simple sequences with complex evolution. *Nature Reviews Genetics*, **5**, 435–445.

Elliott CP, Enright NJ, Allcock RJN *et al.* (in press) Microsatellite markers from the Ion Torrent: a multi-species contrast to 454 shotgun sequencing. *Molecular Ecology Resources*. doi: 10.1111/1755-0998.12192.

Fernandez-Silva I, Whitney J, Wainwright B *et al.* (2013) Microsatellites for next-generation ecologists: a post-sequencing bioinformatics pipeline. *PLoS One*, **8**, e55990.

Gardner MG, Fitch AJ, Bertozzi T, Lowe AJ (2011) Rise of the machines – recommendations for ecologists when using next generation sequencing for microsatellite development. *Molecular Ecology Resources*, **11**, 1093–1101.

Gilles A, Meglecz E, Pech N *et al.* (2011) Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. *BMC Genomics*, **12**, 245.

Glenn TC (2011) Field guide to next-generation DNA sequencers. *Molecular Ecology Resources*, **11**, 759–769.

Glenn TC (2013) 2013 NGS Field Guide: Overview. http://www.molecularecologist.com/next-gen-fieldguide-2013/.

Grohme MA, Soler RF, Wink M, Frohme M (2013) Microsatellite marker discovery using single molecule real-time circular consensus sequencing on the Pacific Biosciences RS. *BioTechniques*, **55**, 253–256.

Guichoux E, Lagache L, Wagner S *et al.* (2011) Current trends in microsatellite genotyping. *Molecular Ecology Resources*, **11**, 591–611.

Haasl RJ, Payseur BA (2011) Multi-locus inference of population structure: a comparison between single nucleotide polymorphisms and microsatellites. *Heredity*, **106**, 158–171.

Huey J, Real K, Mather P *et al.* (2013) Isolation and characterization of 21 polymorphic microsatellite loci in the iconic Australian lungfish, *Neoceratodus forsteri*, using the Ion Torrent next-generation sequencing platform. *Conservation Genetics Resources*, **5**, 737–740.

Illumina Incorporation (2013) Illumina systems. http://www.illumina.com/systems.ilmn.

Jennings TN, Knaus BJ, Mullins TD, Haig SM, Cronn RC (2011) Multiplexed microsatellite recovery using massively parallel sequencing. *Molecular Ecology Resources*, **11**, 1060–1067.

Larkin MA, Blackshields G, Brown NP *et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947–2948.

Loman NJ, Misra RV, Dallman TJ *et al.* (2012) Performance comparison of benchtop high-throughput sequencing platforms. *Nature Biotechnology*, **30**, 434–439.

Margulies M, Egholm M, Altman WE *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.

McCracken G, Wilson K, Paterson I *et al.* (in press) Development of 17 novel microsatellite markers for the longnose sucker (*Catostomus catostomus*) and successful cross-specific amplification of 14 previously developed markers from congeneric species. *Conservation Genetics Resources*. doi: 10.1007/s12686-12013-10086-12683.

Meglécz E, Costedoat C, Dubut V *et al.* (2010) QDD: a user-friendly program to select microsatellite markers and design primers from large sequencing projects. *Bioinformatics*, **26**, 403–404.

Morgante M, Hanafey M, Powell W (2002) Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. *Nature Genetics*, **30**, 194–200.

Nowak C, Zuther S, Leontyev SV, Geismar J (2014) Rapid development of microsatellite markers for the critically endangered Saiga (*Saiga tatarica*) using Illumina® Miseq next generation sequencing technology. *Conservation Genetics Resources*, **6**, 159–162.

Ono Y, Asai K, Hamada M (2013) PBSIM: PacBio reads simulator – toward accurate genome assembly. *Bioinformatics*, **29**, 119–121.

Ouborg NJ, Pertoldi C, Loeschcke V, Bijlsma R, Hedrick PW (2010) Conservation genetics in transition to conservation genomics. *Trends in Genetics*, **26**, 177–187.

Pacific Biosciences (2013) SMRT Technology. http://www.pacificbiosciences.com/products/smrt-technology/smrt-sequencing-advantage/.

Provan J, Powell W, Hollingsworth PM (2001) Chloroplast microsatellites: new tools for studies in plant ecology and evolution. *Trends in Ecology & Evolution*, **16**, 142–147.

Quail MA, Smith M, Coupland P *et al.* (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, **13**, 341.

R Development Core Team (2012) *R: A Language and Environment for Statistical Computing, Version 2.15.0*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods in Molecular Biology*, **132**, 365–386.

Schloss PD, Westcott SL, Ryabin T *et al.* (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, **75**, 7537–7541.

Schlotterer C (2000) Evolutionary dynamics of microsatellite DNA. *Chromosoma*, **109**, 365–371.

Schlotterer C (2004) The evolution of molecular markers – just a matter of fashion? *Nature Reviews Genetics*, **5**, 63–69.

Selkoe KA, Toonen RJ (2006) Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers. *Ecology Letters*, **9**, 615–629.

Slater GS, Birney E (2005) Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, **6**, 31.

Toth G, Gaspari Z, Jurka J (2000) Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Research*, **10**, 967–981.

Travers KJ, Chin CS, Rank DR, Eid JS, Turner SW (2010) A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Research*, **38**, e159.

Tuskan GA, DiFazio S, Jansson S *et al.* (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science*, **313**, 1596–1604.

Wainwright B, Arlyza I, Karl S (2013) Isolation and characterization of twenty-one polymorphic microsatellite loci for *Polycarpa aurata* using third generation sequencing. *Conservation Genetics Resources*, **5**, 671–673.

Wei N, Dick CW, Lowe AJ, Gardner MG (2013) Polymorphic microsatellite loci for *Virola sebifera* (Myristicaceae) derived from shotgun 454 pyrosequencing. *Applications in Plant Sciences*, **1**, 1200295.

Zalapa JE, Cuevas H, Zhu HY *et al.* (2012) Using next-generation sequencing approaches to isolate simple sequence repeat (SSR) loci in the plant sciences. *American Journal of Botany*, **99**, 193–208.

Zane L, Bargelloni L, Patarnello T (2002) Strategies for microsatellite isolation: a review. *Molecular Ecology*, **11**, 1–16.

Zhang XJ, Davenport KW, Gu W *et al.* (2012) Improving genome assemblies by sequencing PCR products with PacBio. *BioTechniques*, **53**, 61–62.

N.W. initiated the project, conducted the experiment and analyses, and wrote the manuscript. J.B.B. and C.W.D. discussed the results and provided helpful comments on earlier drafts of the manuscript.

Data Accessibility

DNA sequences: NCBI SRA accession SRP030127 for CCS fastq files; GenBank accessions KF680324–KF680450 for working microsatellites.

Simulated data: DRYAD entry doi:10.5061/dryad.r20t0

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Fig. S1 Frequency distribution of mean quality score of individual CCS reads before (untrimmed) and after quality control (trimmed).

Fig. S2 Homopolymer lengths of CCS reads.

Fig. S3 The number of CCS reads generated by a single SMRT cell.

Fig. S4 Negative correlation between sequence quality and sequence length in raw CCS reads.

Fig. S5 Positive correlation between sequence quality and sequence length in post-QC CCS reads.

Fig. S6 Positive correlation between homopolymer length and raw CCS read length.

Fig. S7 Base position GC content of CCS reads before (black) and after quality control (grey).

Table S1 Simulations of microsatellite detection effectiveness in relation to read length when sequencing errors were not introduced.

Table S2 Simulations of microsatellite detection effectiveness in relation to read length when PacBio CCS error profiles (Ono *et al*. 2013) were used.