Elliott, C.P., Enright, N.J., Allcock, R.J.N., Gardner, M.G., Meglécz, E., Anthony, J. and Krauss, S.L. (2014) Microsatellite markers from the Ion Torrent: a multi-species contrast to 454 shotgun sequencing. Molecular Ecology Resources, 14 (3). pp. 554-568.

Microsatellite markers from the Ion Torrent: a multi-species contrast to 454 shotgun sequencing.

Carole P. Elliott[1,2], Neal J. Enright[1], Richard J.N. Allcock[4], Michael G. Gardner[5], Emese Meglécz[6], Janet Anthony[2,3], Siegfried L. Krauss[2,3]

1. School of Veterinary and Life Sciences, Environmental and Conservation Sciences, Murdoch University, Murdoch, Western Australia 6150, Australia.
2. Botanic Gardens and Parks Authority, Kings Park and Botanic Garden, Fraser Avenue, West Perth, Western Australia 6005, Australia.
3. School of Plant Biology, University of Western Australia, Crawley, Western Australia 6009, Australia.
4. Lotterywest State Biomedical Facility: Genomics, School of Pathology and Laboratory Medicine, University of Western Australia, Crawley, Western Australia 6009, Australia.
5. School of Biological Sciences, Flinders University, Adelaide, South Australia 5001, Australia; and Evolutionary Biology Unit, South Australian Museum, North Terrace Adelaide, South Australia 5000, Australia.
6. Equipe Evolution, Génome et Environnement, Aix-Marseille University CNRS, IRD, UMR 7263 – IMBE, Marseille, France.

Key words: Next generation sequencing, PGM, Ion Torrent, 454, GS-FLX, gymnosperm, angiosperm.

Corresponding Author: Carole Elliott

School of Veterinary and Life Sciences, Environmental and Conservation Sciences, Murdoch University, Murdoch, Western Australia 6150, Australia.
Botanic Gardens and Parks Authority, Kings Park and Botanic Garden, Fraser Avenue, West Perth, Western Australia 6005, Australia.
Fax: +61-8-9480-3641
Email: c.elliott@murdoch.edu.au or carole.elliott@bgpa.wa.gov.au

Running title: Microsatellite discovery with Ion Torrent and 454

Abstract

The development and screening of microsatellite markers has been accelerated by next generation sequencing (NGS) technology and in particular GS-FLX pyro-sequencing (454). More recent platforms such as the PGM semi-conductor sequencer (Ion Torrent) offer potential benefits such as dramatic reductions in cost, but to date have not been well utilised. Here we critically compare the advantages and disadvantages of microsatellite development using PGM semi-conductor sequencing and GS-FLX pyro-sequencing for two gymnosperm (a conifer and a cycad) and one angiosperm species. We show that these NGS platforms differ in the quantity of returned sequence data, unique microsatellite data and primer design opportunities, mostly consistent with the differences in read length. The strength of the PGM lies in the large amount of data generated at a comparatively lower cost and time. The strength of GS-FLX lies in the return of longer average length sequences and therefore greater flexibility in producing markers with variable product length, due to longer flanking regions, which is ideal for capillary multi-plexing. These differences need to be considered when choosing a NGS method for microsatellite discovery. However, the ongoing improvement in read lengths of the NGS platforms will reduce the disadvantage of the current short read lengths, particularly for the PGM platform, allowing greater flexibility in primer design coupled with the power of a larger number of sequences.

Introduction

The explosion of next generation sequencing (NGS) technology has been a substantial catalyst for the progression of research in all fields of molecular genetics (Mardis, 2008; Metzker, 2010). It has expanded research horizons in the areas of evolutionary genetics (e.g. phylogeny), ecological genetics (e.g. gene flow) and gene expression (Egan *et al.*, 2012). This scope has been facilitated by the evolving capacity of NGS to generate increasingly large volumes of data (e.g. millions of reads or Mb of data) cheaply and quickly (Glenn, 2011; Metzker, 2010). Along with these appealing characteristics, another major advantage of NGS technology is that genomic information can be obtained for non-model species, where limited genetic information is currently available (Ekblom, Galindo, 2011). The vast amount of genomic data obtained with NGS technologies has fostered similar advances in the areas of bioinformatics and data management (Shendure, Ji, 2008). The development of new software programmes to handle sequence alignment, sequence assembly, detection of variation, sequence annotation and data analysis has been an essential tool for the effective and accurate application of NGS technologies to answering genetic and biological questions (see review by Zhang *et al.*, 2011 for details).

A key application of next generation sequencing has been in molecular marker development (Bertozzi *et al.*, 2012). Thousands of markers can be produced for screening in a matter of days (Davey *et al.*, 2011; Egan *et al.*, 2012). In particular, the discovery and development of simple sequence repeats (SSRs or microsatellites) is straightforward using NGS technologies (Gardner *et al.*, 2011; Malausa *et al.*, 2011; Zalapa *et al.*, 2012). To date, the two main NGS platforms that have been used for microsatellite development are the GS-FLX or 454 (Roche Diagnostics) and the GAII/HiSeq/Miseq (Illumina) sequencer, with GS-FLX (454) proving to

be the most popular due to the production of longer sequence reads of 400-500 bp (see review of Zalapa *et al.*, 2012).

Next generation sequencing platforms, including GS-FLX (454), Illumina (HiSeq/Miseq) and PGM (Ion Torrent), are based on a process known as sequencing by synthesis (Glenn, 2011). DNA fragments are prepared by ligation of adaptors, amplification and subsequent immobilisation on a surface before detection. The detection technique is fundamentally different between GS-FLX (454), the Illumina (HiSeq/Miseq) and the PGM (Ion Torrent) platforms. The first two use variations on fluorescence imaging (either single or multi-colour) to determine nucleotide incorporation, whereas the PGM is based on semiconductor technology and uses changes in pH (i.e. the release of H^+ ions) to detect nucleotide incorporation during the sequencing process (Egan *et al.*, 2012; Schadt *et al.*, 2010). As such, GS-FLX (454) and Illumina sequencing are considered to be second generation technologies, whereas PGM semiconductor technology is considered to sit between second and, the single molecule approach, of third generation sequencing (Schadt *et al.*, 2010). PGM sequencing has thus far been effective in the shotgun sequencing of microbial genomes (Egan *et al.*, 2012) and other applications requiring relatively low amounts of data (i.e. Hall-Mendelin *et al.*, 2013). The characteristics of the PGM platform, particularly its low cost and short run-time, suggests that it may also be an effective method for microsatellite marker discovery (Egan *et al.*, 2012).

Microsatellite characteristics such as microsatellite composition, including the relative abundance of different motif type (e.g. AT, ACC), motif class (e.g. dinucleotide), motif length (number of repeat units), and microsatellite coverage (i.e. number of bases of microsatellites

per Mb of DNA) have been shown to vary widely among different taxa (Dieringer, Schlotterer, 2003; Katti *et al.*, 2001; Meglécz *et al.*, 2012a). For example, one study using a fingerprint approach found GA and CA dinucleotides to be major components of the conifer genome (Schmidt *et al.*, 2000) in comparison to the high proportion of AT dinucleotides found in angiosperms (Dieringer, Schlotterer, 2003). The most informative and hence sought-after microsatellites for detecting genetic variability among individuals and populations are the highly polymorphic markers (Ellegren, 2004). The generally accepted consensus is that microsatellites with a shorter motif class and higher number of repeats provide the greatest polymorphism rates (Gardner *et al.*, 2011; Zalapa *et al.*, 2012). Discovering this target type of microsatellite has been made easier with NGS platforms and may be advantageous for species with low microsatellite coverage, like gymnosperms (Meglécz *et al.*, 2012a), where microsatellite discovery has been particularly challenging (Elsik, Williams, 2001).

In this study, we used three non-model species (two Gymnosperms: the cycad *Macrozamia riedlei* and the conifer *Podocarpus drouynianus*; one Angiosperm: the euphorb *Ricinocarpos brevis*) to critically compare the costs and benefits of two NGS platforms (Ion PGM and GS-FLX sequencing) for microsatellite discovery. Specifically, we compared the two platforms in their capacity to provide unique microsatellites for which primers could be designed. In addition, we examine the polymorphic variation of several loci from each platform, for two species (the Gymnosperms). We then outline several points to consider when deciding which NGS platform to choose (including cost, primer design, etc).

Method

We extracted DNA from leaf material from one individual of each of the two gymnosperm species, *Macrozamia riedlei* (Cycadaceae; MR) and *Podocarpus drouynianus* (Podocarpaceae; PD), and the angiosperm species, *Ricinocarpos brevis* (Euphorbiaceae; RB). We used the Sinclair *et al.* (2009) extraction method for the two gymnosperm species and a modified version of the Carlson *et al*. (1991) extraction method for the angiosperm. The modifications involved adding a potassium acetate step (0.7M KAc in supernatant, freeze for 15 min, centrifuge at 13 000 *g* for 10 min and transfer supernatant) before the isopropanol step, and adding a sodium chloride step (3M NaCl in dissolved pellet solution, centrifuge at 16 600 *g* for 10 min and discard the supernatant) before ethanol precipitation. Gymnosperm samples were cleaned using Agencourt AMPure XP beads before sequencing (PGM platform only).

NGS was performed using the Ion Personal Genome Machine Sequencer (Life Technologies; also known as PGM or Ion Torrent) and the Genome Sequencer FLX Titanium Instrument (Roche Diagnostics; commonly known as GS-FLX or 454) and these are outlined below. PGM semi-conductor sequencing was performed at the LotteryWest State Biomedical Facility Genomics Node in Perth, Western Australia. Briefly, 100ng of DNA was sheared to approximately 200-300 bp using an S2 sonicator (Covaris, UWA). Barcoded libraries were prepared using an Ion Xpress Fragment Library kit (Life Technologies, USA). Size selection (insert sizes 200-250 bp) was performed by gel excision (E-gel, Invitrogen) and the libraries were assessed and quantified using a Bioanalyser 2100 (Agilent Technologies, USA). Individual libraries were then diluted to 9pM for template preparation using a OneTouch Template 200 kit (Life Technologies, USA) and enriched. Sequencing was performed on a PGM using 520 flows (generating approx 200-250 bp read lengths) on a 316 sequencing chip.

After sequencing, signal processing and base-calling was performed using TorrentSuite 2.2 and library-specific fastq files were also generated.

GS-FLX pyro-sequencing runs were conducted at the Ramaciotti Centre for Gene Function Analysis, in Sydney, New South Wales, Australia. Five µg of each DNA sample were shotgun sequenced on a Titanium GS-FLX (Roche Applied Science, Indianapolis, Indiana, USA), with each species occupying 12.5% of a plate, following Gardner et al. (2011). Sequences for the three species from both the PGM and GS-FLX platforms have been lodged with the Dryad Digital Repository (http://dx.doi.org/10.5061/dryad.vv82q; http://dx.doi.org/10.5061/dryad.jd183, respectively). We compared the two platforms by utilising the same three plant species, the same analysis parameters with the returned sequence data (see below), and obtained the costs and instrument running times for each platform.

Microsatellite discovery and primer design was performed using a slightly modified version of QDD v 2.2 to reduce run time (Meglécz *et al.*, 2010). The default parameters of the programme were used for the screening steps and for primer design, except for the following stringency: GC clamp (PRIMER_GC_CLAMP) set to two and Max self-complementarity (PRIMER_SELF_ANY) was set to six. QDD output was summarised and t-tests were performed on primer design data. Geneious (v 5.6.5) was used to calculate average sequence length and the GC content of sequences. Since genome coverage was very low (0.2-5%; Table 1) assembly of sequences (previous to running QDD) would produce probably meaningless contigs of repetitive elements, therefore raw reads were used in the pipeline. Primer design was performed on unique microsatellite sequences only (UMS; i.e. sequences containing

microsatellites that were singletons) since our preliminary results indicated that consensus sequences obtained from low coverage data may be consensuses of repetitive elements (E. Meglécz, N. Pech, V. Dubut, A. Gilles, P. Hingamp, A. Trilles, R. Grenier & JF. Martin, unpublished data). Furthermore, only primers with stringent design (A, B or C) were taken into account, to concentrate on markers with only one microsatellite in the target region and homopolymers longer than four bases were not allowed in the amplicon. Statistics for the unique microsatellite sequences (UMS) were obtained from analysing and summarising Batch3 output (You *et al.*, 2008), and performing t-tests and pairwise comparisons.

Primers were subsequently tested for polymorphism using genomic DNA of *Macrozamia riedlei* (Cycadaceae; MR) and *Podocarpus drouynianus* (Podocarpaceae; PD; Western Australia), which were extracted from 10-20mg of milled, freeze-dried leaf material, using the protocol outlined in Botieux *et al*. (method 7; 1999) with the following modifications. The extraction buffer contained 100mM Tris-HCl pH 8.0, 50mM EDTA pH 8.0, 1.25% SDS and 1.25% PVP. Samples were mixed and incubated overnight at 50°C. After the addition of ammonium acetate (6M), samples were incubated at 4°C for 30 min and centrifuged for 30 min (16 600 *g*). After the addition of isopropanol (equal volume), samples were put in the freezer for 30 min and then centrifuged for 30 min. The supernatant was discarded; the pellet was washed in 70% ethanol and then centrifuged for 20 min. After the ethanol was discarded, the pellet was air-dried and resuspended in water (or TE buffer) overnight. The samples were centrifuged for 30 min and the supernatant transferred to new tubes.

We trialled primers sourced from both platforms and for both species, all primers were initially tested on 7 individuals for two populations (40km apart) from jarrah (*Eucalyptus*

199 *marginata*) forest east of Perth, Western Australia (MR: 35 microsatellites sourced from PGM
200 and 10 microsatellites from GS-FLX; PD: 18 primer pairs sourced from PGM and 17 sourced
201 from GS-FLX). Amplification via two PCR programmes and screening on 2.0% agarose gels
202 showed six MR loci (four PGM and two GS-FLX) and three PD loci (two PGM and one GS-
203 FLX) failed to amplify. We selected 23 (15 PGM and 8 GS-FLX sourced) MR loci and 12
204 (six PGM and six GS-FLX sourced) PD loci that were investigated for variation by
205 genotyping 18-22 individuals from four populations for each species. A further four
206 *Macrozamia* loci (PGM sourced) produced uninterpretable banding patterns of the expected
207 size.
208
209 Microsatellite loci were amplified in a 10µL reaction volume containing 20ng of DNA, 0.5x
210 PCR buffer (Fisher Biotec), 0.1µM labeled forward primer (WellRED oligos, Sigma Aldrich),
211 0.1µM reverse primer, 0.5U *Taq* polymerase (Fisher Biotec), variable $MgCl_2$ concentrations
212 and the addition of bovine serum albumin at 2% (Table 2a and 2b). Amplifications were
213 completed in a Veriti thermocycler (Applied Biosystems) with one of the following
214 conditions: 1) 96°C for 2 min, 40 cycles at 95°C for 30 s, variable annealing (see Table 2a and
215 2b) for 1 min and 72°C for 30 s followed by 72°C for 5 min and a 4°C holding step or, 2)
216 96°C for 2 min, 20 cycles at 95°C for 30 s, variable annealing (see Table 2a and 2b) for 30s
217 decreasing by 0.5°C per cycle, 72°C for 30 s and 30 cycles at 95°C for 30 s, variable
218 annealing (see Table 2a and 2b) for 30 s, 72°C for 30 s, followed by 72°C for 5 min and a 4°C
219 holding step. For sequence electrophoresis, 2µL of each loci was added to 30µL of loading
220 mix (29.65µL sample loading solution and 0.35µL DNA size standard kit – 400; Beckman
221 Coulter) and separated on CEQ8800 (Beckman Coulter). Fragment sizes and analysis was
222 conducted using Genetic Analysis System software version 9.0.25 (Beckman Coulter).

For each locus we calculated the number and range of alleles, observed and expected heterozygosity and polymorphic information content using CERVUS (Kalinowski *et al.*, 2007). In addition, we checked for deviation from Hardy-Weinberg Equilibrium (HWE) and linkage disequilibrium for all pairs of loci using GENEPOP 3.4 (Raymond, Rousset, 1995). *P* values from HWE and linkage disequilibrium tests were adjusted for multiple tests of significance using the method of Benjamini and Hochberg (1995). We used MICROCHECKER 2.2.3 (Van Oosterhout *et al.*, 2004) to check each locus for evidence of null alleles, scoring error due to stuttering, and large allele drop out (Table 2a and 2b).

We conducted a simulation experiment to investigate the effect that different size distributions and coverage may have on the discovery of microsatellites and their primer design. We randomly sampled the human genome (version GRCh37.p5) to generate data sets that represented the different read length distributions of the PGM and GS-FLX platforms, at several levels of genome coverage (0.01x, 0.04x, 0.07x, 0.1x).

In order to detect platform specific differences, independent of read length distribution and read number, we made reduced data sets from the real sequence data of both platforms, where each reduced data set contained the same number of reads as the real GS-FLX data set and followed the same read length distribution as the PGM data set. To create these reduced data sets, first we randomly chose reads from the PGM data to equal the number of reads in the corresponding GS-FLX data set, then we truncated the reads of the GS-FLX data so they had the same read length distribution as the PGM data set. All simulated or reduced data sets were

analysed by QDD in the same manner as described above for microsatellite discovery and primer design.

Results

Returned sequence data

The differences between the results from the two NGS platforms and the two phyla used to assess these sequencing methods are summarised in Table 1. Individual PGM sequencing runs returned four to fifteen times more reads and data (in Mbases) per species than the GS-FLX (t=7.7 d.f. 3 P=0.004; t=4.3 d.f. 4 P=0.006, respectively). However, the average read length of PGM sequences was 59-62% shorter than GS-FLX sequences (Table 1; Fig. 1; t=-17.3 d.f. 2 P=0.003). The GC content of these sequences was no different between the platforms or the phyla (Table 1; t=-0.8 d.f. 4 P=0.47). The proportion of sequences that contained microsatellites was very low (0.5 – 5.0%) and not significantly different between the platforms (Table 1; t=-2.5 d.f. 2 P=0.13). The distribution of read lengths of these microsatellite-containing sequences clearly showed inter-species consistency within a platform, but substantial difference between the two platforms themselves, with the majority being 150-250 bp long for PGM sequences and 350-500 bp for the GS-FLX platform (Supplementary Material 1). Among microsatellite containing sequences, PGM runs had a significantly higher proportion of unique reads (singletons) than the GS-FLX runs (41-58% PGM; 21-29% GS-FLX; Table 1; t=4.3 d.f. 3 P=0.01), but the proportion of consensus sequences was not significantly different between platforms (4-6% PGM; 4-13% GS-FLX; Table 1; t=-1.3 d.f. 2 P=0.15). Accordingly, the proportion of sequences not used for primer design (redundant, potentially repetitive, low complexity) was significantly lower in PGM than GS-FLX sequences (36-55% for PGM; 59-70% for GS-FLX; Table 1; t=-3.3 d.f. 3

P=0.02). These results were consistent within each species. This species consistency between PGM and GS-FLX platforms was also reflected in the higher proportion of microsatellites contained in angiosperm sequences compared to gymnosperm sequences (Table 1).

There were no detectable differences between the two NGS platforms or between the two phyla (gymnosperm and angiosperm) in the most common microsatellite motifs found (i.e. AT; AAT and AAG motifs were the most common; Supplementary Material 2). However, there was a comparative difference in terms of the proportional distribution of certain motif types within each nucleotide class. For example, PGM sequencing returned a higher proportion of AG dinucleotide and both ATC and AAG trinucleotide microsatellites than GS-FLX sequencing, whereas, GS-FLX sequencing returned a higher proportion of AT dinucleotide and AAT trinucleotide microsatellites than PGM sequencing (Supplementary Material 2). Both PGM and GS-FLX platforms were consistent in terms of the differences in the proportional distribution of motif types among species, as the two gymnosperms had a higher proportion of AC dinucleotides and AAC and ATC trinucleotide microsatellites than the angiosperm, which had a higher proportion of the AT and AG dinucleotides and AGC trinucleotide microsatellites, for both platforms (Supplementary Material 2).

Microsatellites of unique sequences

A high number of microsatellite containing sequences could be used for primer design for both platforms. When considering the number of microsatellite containing singleton sequences (unique microsatellite sequences) with a successful stringent primer design, their percentage to the total number of UMSs was higher in GS-FLX than PGM sequences (14-

21% GS-FLX; 8-12% PGM; Table 1; *t*=-2.9 d.f. 3 *P*=0.03). This result was consistent with the expectation that longer reads provide a greater probability for successful primer design.

Considering the unique microsatellite sequences (UMS, i.e. sequences contained microsatellites and were singletons) a higher proportion contained more than one microsatellite from the GS-FLX platform (Table 1; *t*=-4.9 d.f. 4 *P*=0.01), a result entirely consistent with the significantly longer read length of UMS produced by this platform (Fig. 1; *t*=-11.3 d.f. 2 *P*=0.01). The dominance of the shorter, motif classes (i.e. dinucleotides were the most abundant) was not different between NGS platforms (*t*=0.69 d.f. 3 *P*=0.54), however, there was a significantly higher proportion of tetranucleotides returned in GS-FLX UMS compared to PGM UMS (Fig. 2; *t*=-5.3 d.f. 4 *P*=0.01). The average number of repeat units differed between the two platforms for the dinucleotide motif class, with GS-FLX sourced microsatellites longer by an average of 2.3-2.6 repeat units than PGM sourced microsatellites, across all three species (Fig. 3a; *t*=-4.3 d.f. 2 *P*=0.05). The other motif classes were no different in length between platforms, except for the GS-FLX sourced trinucleotides of the angiosperm species (RB), which was 2.6 repeat units on average longer than the other species and for the comparable data from this angiosperm on the other platform (Fig. 3a). There was no difference between platforms in the number of motif types (e.g. AT or AG) within each nucleotide class (e.g. dinucleotide; Table 1). In addition, when comparing between the two platforms for each individual species, the average difference between the proportion of each nucleotide class (Fig. 2) and the average number of repeat units of each nucleotide class (Fig. 3a) were not significantly different than zero (nucleotide class: all species, *t*>0.001 d.f. 4 *P*=1; repeat unit length: MR - *t*=-0.28 d.f. 4 *P*=0.79; PD - *t*=-1.0 d.f. 4 *P*=0.38; RB - *t*=-1.81 d.f. 4

316 *P*=0.14), which also supported the observed consistency between the platforms within each
317 species.
318
319 *Simulations*
320 By using the reduced PGM and GS-FLX sequence data sets, we tested if there are platform
321 specific differences, apart from the read length and number of reads. Microsatellite content
322 and the proportion of UMS with successful primer design were not different between the
323 platforms (Table 3). Interestingly, the proportion of singletons remained significantly lower
324 (34-50% GS-FLX; 76-86% PGM; Table 3; *t*=6.9 d.f. 3 *P*=0.003) and the consensus sequences
325 were significantly higher with the reduced GS-FLX data set than for the reduced PGM data
326 set (7-17% versus 0.5-2%; Table 3; *t*=-3.7 d.f. 2 *P*=0.03). With low coverage data, few
327 sequences are expected to cover the same locus, thus the high proportion of GS-FLX
328 consensus sequences are likely to indicate a bias towards sequencing interspersed repetitive
329 regions in the GS-FLX data set. However, a similar proportion of microsatellite containing
330 reads between the reduced data set of the two platforms suggests, that there is no important
331 platform specific bias towards sequencing microsatellite containing regions.
332
333 In order to investigate the effect of read length on the number of markers with stringent
334 primers we randomly sampled the human genome at four different levels of data coverage.
335 For all coverage levels two samples were generated, one following the PGM data set read
336 length distribution (named HG_PGM samples) and the other the GS-FLX data set read length
337 distribution (named HG_GS-FLX samples). In this way, the only difference between samples
338 of the same coverage was the read length distribution and consequently the number of reads,
339 and they were free from all potential platform induced bias (Table 4). Independent of the level

of coverage, 83% of the HG_PGM sequences were longer than 80bp (limit used in QDD pipeline), while in the HG_GS-FLX samples, it was 95%. The percentage of reads containing microsatellites was 3-4% for HG_PGM sequences and 7% for HG_GS-FLX sequences, independently of coverage level. As a result, for each coverage level a very similar number of raw microsatellite containing sequences were obtained (Table 4). The QDD pipeline eliminates redundancy and reads that are less likely to suit for PCR amplification. The major groups of microsatellite containing sequences are the singletons (unique sequences with no similarity to other sequences), consensus sequences (highly similar reads pooled into a consensus), grouped sequences (significant but below limit or partial similarity to other sequences), nohit_css sequences (Cryptically Simple Sequence; low complexity or repetitive region covers most of the read) and multihit_css (Cryptically Simple Sequence; probable repeated region within sequence). These later three categories together with the redundant reads pooled into consensus sequences, form a pool of sequences that were not used for primer design. The percentage of singletons decreased with increasing coverage and it was significantly higher for HG_PGM samples (78-86%) than HG_GS-FLX samples (61-71%; t=-5.7 d.f. 6 P=0.001). As expected, the percentage of consensus sequences increased with coverage and was not significantly different between the pair of samples (0.3-3.1%; t=-0.4 d.f. 6 P=0.70). In addition, the percentage of sequences not used for primer design (i.e. redundant, grouped, css) increased with coverage and was significantly lower in HG_PGM (13-19% compared to GS-FLX with 26-40%; Table 4; coverage: t=-7.4 d.f. 6 P>0.001). For grouped, nohit_css and multihit_css sequences, there was a clear read length effect and there was no difference between coverage levels (Table 4). For example, 5-6% of the HG_PGM samples and 1% of the GS-FLX samples were nohit_css. The difference between read length distributions is expected, since shorter reads have higher chance to be covered by low

complexity sequences. Grouped and multihit_css sequences were significantly more frequent in HG_GS-FLX samples than HG_PGM samples (grouped: 16.4-18.7% and 7.5-9.2%, respectively, t=-14.6 d.f. 6 P>0.001; multihit_css: 9.7-11.6% and 0-0.1%, respectively, t=-27.4 d.f. 3 P>0.001). The percentage of UMS for which stringent primers were designed, was significantly higher for the HG_GS-FLX samples, regardless of the coverage level (t=-125 d.f. 6 P>0.001).

Consequences for microsatellite primer design

PGM sequencing produced many more sequences, and a higher number of UMS than GS-FLX sequencing, regardless of phyla. As a consequence, primers could be designed for a larger number of microsatellite sequences (Table 1; t=2.7 d.f. 3 P=0.04). As dinucleotide microsatellites were the most common, they were also the dominant microsatellite for which primers could be designed (Table 1). The average repeat length of PGM sourced dinucleotide microsatellites, for which primers could be designed (design levels A-G in QDD), was 0.5 repeat units shorter than those sourced from GS-FLX sequencing, and this was consistent among species (Fig 3b; t=-11.7 d.f. 4 P<0.001). The trinucleotide class of microsatellites followed a similar pattern to the dinucleotides (0.4 repeat units shorter with PGM sequencing; t=-1.3 d.f. 2 P=0.32), except for the GS-FLX sourced angiosperm which was 1.5 repeat units longer that the others. There was no clear pattern amongst the other motif classes (Table 1). The average product length from primers designed based on GS-FLX sequences (design levels A-C as recommended by QDD) was 35-47 bases longer than PGM (Fig 1; t=-12.6 d.f. 3 P=0.001).

Polymorphism of designed primers

388 Of the 19 MR microsatellite loci, 17 were polymorphic (Table 2a) with the total number of
389 alleles ranging from two to fourteen. Four loci (two PGM and two GS-FLX sourced) showed
390 significant deviation from Hardy-Weinberg equilibrium (P<0.05) due to heterozygote
391 deficiency (Table 2a). Locus MR2 showed evidence of null alleles at the target site (Table
392 2a). None of the other loci showed evidence for large allele drop out, or evidence of scoring
393 error due to stuttering. No loci showed evidence of linkage disequilibrium.
394
395 All of the twelve PD microsatellite loci were polymorphic with the total number of alleles
396 ranging from three to twelve (Table 2b). Eight loci (three PGM and five GS-FLX sourced)
397 showed significant deviation from Hardy-Weinberg equilibrium (P<0.05) due to heterozygote
398 deficiency, and showed evidence of null alleles (Table 2b). None of the other loci showed
399 evidence for large allele drop out, or evidence of scoring error due to stuttering. One pair of
400 loci showed evidence of linkage disequilibrium (PD15 and PD31). The causes of these
401 departures require further investigation.
402
403 For both species, there was no significant difference in the number of alleles, observed
404 heterozygosity, expected heterozygosity or polymorphic information content between primers
405 sourced from the two platforms. These results indicate that similar levels of polymorphism
406 were produced for the microsatellite markers sourced from either platform, regardless of
407 species.
408

409 Discussion

410 It is now well established that NGS has the capacity to accelerate molecular marker
411 development by providing more sequence material to be screened for markers such as
412 microsatellites. GS-FLX pyro-sequencing was the first NGS platform suitable for
413 microsatellite discovery and marker development (Castoe *et al.*, 2010; Gardner *et al.*, 2011;
414 Malausa *et al.*, 2011; Santana, 2009), recently followed by the Illumina sequencing (Castoe *et*
415 *al.*, 2012). Our study illustrates that PGM semi-conductor sequencing is also a suitable
416 platform for microsatellite discovery. The time and cost estimates of NGS services, as
417 indicated by Glenn (2013) and described in Egan *et al*. (2012), are both congruent with our
418 results that indicate PGM was comparatively less expensive per run, required less DNA for a
419 sequence run, and was faster for sequence preparation (ligation etc.) and instrument running
420 times than the GS-FLX platform (Table 1).
421
422 In terms of the sequence output obtained, the major difference between the two platforms was
423 the number of sequences and their length, as PGM sequencing produced a larger number of
424 sequences than GS-FLX, albeit with a shorter average repeat length. Interestingly, the PGM
425 sequence output contained a higher proportion of sequences that were unique within species
426 (41-58%), in comparison to GS-FLX sequencing (21-29%). This is unexpected, since due to a
427 higher coverage with PGM data (1-5%) than GS-FLX data (<1%), a smaller number of
428 singletons were expected for this data set, as demonstrated by the human genome simulations
429 (Table 4). However it should be acknowledged, that the reduction of redundancy in the QDD
430 pipeline used in this study, does not only mean making consensus sequences from highly
431 similar reads, but also eliminating potentially problematic reads, that could come from

repetitive regions. Such detection, depends on the read length and genome coverage, and thus affects the two data sets differently.

Our simulations from the human genome using PGM and GS-FLX sequence size distributions allowed us to study the effect of read length distribution, regardless of coverage, and potential platform specific biases. For equal coverage, despite the higher number of reads of the PGM distribution data, the number of sufficiently long sequences with microsatellites was similar for the two size distributions. Among these sequences, a lower percentage of unique sequences in GS-FLX distribution data was mainly due to the high proportion of grouped and multihit sequences. Grouped sequences show either partial similarity to other sequences, or their similarity is not high enough to suppose that they are reads of the same locus. Although there is no experimental proof that primers designed for them cannot amplify a unique PCR product, it is unwise to use them if there are a sufficiently high number of unique sequences. As expected, our simulations showed that the chance of detecting partial similarity between sequences increases with read length. Multihit sequences are detected by the presence of more than one local alignment between two sequences, suggesting that they contain a sequence region repeated within a read. Therefore, unlike unique sequences, they are less likely to provide good support for PCR amplifications. As for grouped reads, their detection also depends on read length and therefore, is easier for GS-FLX than for PGM sequences. Thus, the lower percentage of unique sequences following the GS-FLX read length distribution data set is not necessarily a disadvantage, since overall they may amplify better than unique sequences from PGM distribution, where the detection of the potential problems, such as partial similarities between sequences or intra-sequence repetitions, is more difficult. Furthermore, due to longer read lengths, a higher percentage of the UMS of the GS-FLX

456 distribution were suitable for stringent primer design. In summary, for equal coverage levels
457 of the human genome, the GS-FLX size distribution data set produced a higher number of
458 potential markers, with an overall PCR success rate likely to be higher. Therefore, a trade-off
459 between the two platforms must be considered: due to its high throughput, the PGM platform
460 returns many more UMS at comparably less financial cost, but the GS-FLX platform returns
461 longer sequence lengths that proved better for microsatellite selection because there is a
462 higher chance to contain microsatellites with enough flanking regions, and longer read length
463 allowing the detection of potentially repetitive regions. However, these results are based on
464 simulated data to compare the effect of coverage and read length, and do not take into account
465 other potential platform specific biases.
466
467 To test whether the high proportion of unique sequences in our experimental PGM data set
468 was due only to the shorter read length, or was also a result of platform specificities other than
469 coverage and read length, we compared the proportion of unique and consensus sequences in
470 the reduced data sets originating from both platforms, but containing the same number of
471 sequences and following the length distribution of the PGM data. A higher proportion of
472 consensus sequences and unused sequences in the reduced GS-FLX data set indicated that
473 there is a potential bias in this sequencing platform. Although 454 shotgun sequencing
474 provides a biologically meaningful sample of the genome, it is not a perfect representation of
475 the genome (Meglécz *et al.*, 2012b). It has been shown that the GS-FLX platform may
476 sequence the same molecule more than once if a DNA fragment attaches to neighbouring
477 empty beads during the emulsion PCR (Gomez-Alvarez *et al.*, 2009). Although no major bias
478 has been detected in an earlier study (Meglécz *et al.*, 2012b), a further possible explanation of

a relatively high proportion of consensus sequences of GS-FLX system is that this method might be slightly biased toward sequences of repetitive regions.

Shorter read lengths can influence a number of processes involved in microsatellite marker development. First, shorter read lengths reduce the likelihood of discovering longer repeat length microsatellites and this is undesirable as higher polymorphism is generally associated with longer microsatellites (Gardner *et al.*, 2011). PGM UMS were on average 150 bp shorter in length, resulting in an average decrease in dinucleotide motif length of 2.3 to 2.6 repeat units (1-3 repeat units for the subset of microsatellites for which primers were designed). However, no such decrease in the number of repeat units was detected among the other motif classes, potentially due to a reduced number of these larger classes not facilitating a comparison. The reduction in repeat number of dinucleotide microsatellites suggests there could be a potential shortfall in the average polymorphism of dinucleotide markers produced from PGM sequencing. The extent of this shortfall would be strongly dependent on the species in question and its mutational mechanisms (Ellegren, 2000). Second, shorter sequence read lengths reduce the potential to develop primers of suitable specificity as flanking regions are too short or not conducive for primer design (Zalapa *et al.*, 2012). Despite shorter unique microsatellite sequence read lengths from PGM sequences, we were able to successfully design primers for a high proportion of microsatellites, thus providing a larger pool of potential markers to choose from. However, the longer average fragment length of GS-FLX microsatellites (24-28% longer) provides greater flexibility in capillary multi-plexing (i.e. the simultaneous fragment separation of multiple markers) with greater numbers of markers with non-overlapping sizes than for PGM sequencing. The current disadvantages in shorter sequence read lengths will largely be overcome with improved upgrades in the technology of

503 both NGS platforms that include longer sequence read lengths of 400 bp to 1kb (Life
504 Technologies Corporation 2012; Roche Diagnostics Corporation 2012).
505
506 There were several differences in microsatellite characteristics between the two platforms,
507 with the PGM sourced sequences returning a proportionally different motif type (e.g. AG), a
508 smaller proportion of tetranucleotides (motif class), and shorter dinucleotide microsatellites
509 (i.e. lower number of repeat units) than the GS-FLX sourced sequences. Tetranucleotide
510 microsatellites are reported as having greater usability than other nucleotide classes (Gardner
511 *et al.*, 2011) and GS-FLX returned proportionally more of this nucleotide class, indicating a
512 potential advantage in targeting microsatellites that are more likely to be polymorphic with
513 this platform. However, the number of tetranucleotide microsatellites for which primers could
514 be designed was not significantly different between the platforms, suggesting that successful
515 primer design may restrict this potential nucleotide advantage of the GS-FLX platform.
516 However, when considering the levels of polymorphism of the markers developed and tested
517 on the two gymnosperm species from the two platforms there was no significant difference.
518 This indicated that obtaining successful microsatellite markers and their level of
519 polymorphism was not dependent on the platform from which the microsatellite sequences
520 were discovered.
521
522 Interestingly, the two NGS platforms consistently showed the same patterns, in terms of the
523 number of microsatellites and their characteristics, even when considering the different
524 genomes of gymnosperms and angiosperms. For example, both platforms showed AT to be
525 the most common motif type, which is congruent with other studies on plants (Dieringer,
526 Schlotterer, 2003; Katti *et al.*, 2001; Sonah *et al.*, 2011), and both showed gymnosperms to

have a higher proportion of AC dinucleotides (11-29% more) than the angiosperm, which also supports the evidence for this dinucleotide being a major component of their genome in comparison to angiosperms (Schmidt *et al.*, 2000). In addition, sequences from both platforms indicated that the angiosperm genome contained a greater proportion and diversity (i.e. a higher number of different motif types in all classes, except the dinucleotide class) of microsatellites than the gymnosperm genomes. Evidence suggests that the frequency of microsatellites is positively correlated with the abundance of single or low copy DNA in angiosperms (Morgante *et al.*, 2002). Single or low copy DNA has been shown to make up only 25-28% of two gymnosperm genomes (Elsik, Williams, 2000; Kurdi-Haidar *et al.*, 1983) and it is possible that this was the reason for the smaller number of microsatellites discovered in the two gymnosperm species here. As such, developing microsatellite libraries by other methods, such as library enrichment and searching within cDNA libraries, has been difficult and often restrictive in the quantity of microsatellites discovered (Elsik, Williams, 2001). Here we show that the PGM and GS-FLX platforms demonstrate great capacity to generate large microsatellite libraries for gymnosperms. Considering the large number of economically important gymnosperms, like the conifers, both NGS platforms could prove valuable in providing additional, suitable molecular markers in a relatively short time frame (Schmidt *et al.*, 2000).

For this study, we chose the QDD pipeline (Meglécz *et al.*, 2010) since the 200-250 bp read length of the PGM platform allows a meaningful use of this method, and comparison to our existing GS-FLX data. The QDD pipeline uses an all against all BLAST and conservatively eliminates sequences with high proportion of low complexity regions and intra-read repetitions of blocks of sequences. QDD also makes consensus sequences from highly similar

reads and eliminates reads with partial similarities, but it is not designed to replace dedicated genome assemblers. The dominant alternative approach is the PAL_finder pipeline (Castoe et al. 2012) for Illumina reads, which uses the number of exact matches of primers against the whole dataset to evaluate redundancy and/or repetitiveness of a region. It is an elegant approach that is able to treat Gigabases of the short sequences produced by Illumina, however, it cannot account for natural variation between alleles, copies of repetitive elements or sequencing errors, when counting only exact matches. QDD is thus a more conservative approach, which can be valuable in reducing wet-lab cost while setting up markers, but it is slower, and does not work well with short reads and very large datasets produced by Illumina. Both bioinformatics pipelines are based on relatively low coverage data (Castoe *et al.*, 2012; Meglécz *et al.*, 2010). PAL_finder's most stringent selection is based on primer pairs that both appear only once in the dataset, thus selects loci that have not been sequenced more than once. Thus, Gigabases of sequence data for a genome of 100-200 Mb, would result in a high coverage, and would give few unique loci for primer design with either of the two pipelines. In this case, the most efficient way of analysing the data would be to assemble them first and use contigs for microsatellite research and marker development. On the other hand, species with large genomes and low microsatellite density, such as the two species in this study, would definitely benefit from large datasets, even at the expense of obtaining shorter read length. A meaningful comparison of the QDD and the PAL_finder pipelines, as well as comparisons of the behaviour of the data produced by different platforms (including Illumina), would be very desirable. However, these comparisons would be complex and beyond the scope of the current paper, since the principle of redundancy elimination of the two pipelines are quite different, and they are both influenced by genome coverage and read length issues.

Conclusions

Our multi-species comparison between PGM (Ion Torrent) and GS-FLX (454) platforms showed that a trade-off needs to be considered when choosing a NGS platform for microsatellite discovery, as the smaller sequence size of the PGM platform resulted in shorter dinucleotide microsatellites (0.5 repeat units shorter in length), but returned a significantly larger number of markers to screen at a comparatively lower cost and shorter run time than the GS-FLX sourced markers. We also show that both NGS platforms showed no species bias in microsatellite characteristics, which were congruent with the available evidence on the specific genome content of the gymnosperm and angiosperm phyla, or the level of marker polymorphism, suggesting that limited consideration needs to be given to these issues when choosing a NGS platform. The ongoing improvement to the sequence quality and capacity of these NGS platforms will alter the differences observed between the PGM and GS-FLX platforms, including the increased read lengths of sequences, which will minimise the disadvantage of shorter microsatellite motifs and limited flanking regions for primer design, particularly for sequences sourced from the PGM semi-conductor platform.

Acknowledgements

This research was supported by an Australian Research Council Discovery Project grant (DP110101480 to NJE and SLK). We would like to thank Nina Kresoje, Vanessa Atkinson and Alison Fitch for library preparation and sequencing.

596 References

597 Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and
598 powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series*
599 *B (Statistical Methodology)* **57**, 289-300.
600 Bertozzi T, Sanders KL, Sistrom MJ, Gardner MG (2012) Anonymous nuclear loci in non-
601 model organisms: making the most of high-throughput genome surveys.
602 *Bioinformatics* **28**, 1807-1810.
603 Boiteux LS, Fonseca MEN, Simon PW (1999) Effects of plant tissue and DNA purification
604 method on randomly amplified polymorphic DNA-based genetic fingerprinting
605 analysis in carrot. *Journal of the American Society for Horticultural Science* **124**, 32-
606 38.
607 Carlson JE, Tulsieram LK, Glaubitz JC, *et al.* (1991) Segregation of random amplified DNA
608 markers in F1 progeny of conifers. *Theoretical and Applied Genetics* **83**, 194-200.
609 Castoe TA, Poole AW, de Koning APJ, *et al.* (2012) Rapid microsatellite identification from
610 Illumina paired-end genomic sequencing in two birds and a snake. *PLoS ONE* **7**,
611 e30953.
612 Castoe TA, Poole AW, Gu W, *et al.* (2010) Rapid identification of thousands of copperhead
613 snake (*Agkistrodon contortrix*) microsatellite loci from modest amounts of 454
614 shotgun genome sequence. *Molecular Ecology Resources* **10**, 341-347.
615 Davey JW, Hohenlohe PA, Etter PD, *et al.* (2011) Genome-wide genetic marker discovery
616 and genotyping using next-generation sequencing. *Nature Reviews Genetics* **12**, 499-
617 510.

618 Dieringer D, Schlotterer C (2003) Two distinct modes of microsatellite mutation processes:
619 Evidence from the complete genomic sequences of nine species. *Genome Res.* **13**,
620 2242-2251.
621 Egan AN, Schlueter J, Spooner DM (2012) Applications of next-generation sequencing in
622 plant biology. *American Journal of Botany* **99**, 175-185.
623 Ekblom R, Galindo J (2011) Applications of next generation sequencing in molecular ecology
624 of non-model organisms. *Heredity* **107**, 1-15.
625 Ellegren H (2000) Microsatellite mutations in the germline: implications for evolutionary
626 inference. *Trends in Genetics* **16**, 551-558.
627 Ellegren H (2004) Microsatellites: simple sequences with complex evolution. *Nature Reviews*
628 *Genetics* **5**, 435-445.
629 Elsik CG, Williams CG (2000) Retroelements contribute to the excess low-copy-number
630 DNA in pine. *Molecular and General Genetics MGG* **264**, 47-55.
631 Elsik CG, Williams CG (2001) Low-copy microsatellite recovery from a conifer genome.
632 *Theoretical and Applied Genetics* **103**, 1189-1195.
633 Gardner MG, Fitch AJ, Bertozzi T, Lowe AJ (2011) Rise of the machines – recommendations
634 for ecologists when using next generation sequencing for microsatellite development.
635 *Molecular Ecology Resources* **11**, 1093-1101.
636 Geneious version 5.6.5 created by Biomatters. Available from http://www.geneious.com/
637 Glenn TC (2011) Field guide to next-generation DNA sequencers. *Molecular Ecology*
638 *Resources* **11**, 759-769.
639 Glenn, TC. (2013) 2013 NGA Field Guide – Table 2b: Costs/run, cost/MB, minimum costs.
640 The Molecular Ecologist http://www.molecularecologist.com/next-gen-fieldguide-
641 2013.

642 Gomez-Alvarez V, Teal TK, Schmidt TM (2009) Systematic artifacts in metagenomes from
643 complex microbial communities. *ISME J* **3**, 1314-1317.
644 Hall-Mendelin S, Allcock R, Kresoje N, van den Hurk AF, Warrilow D (2013) Detection of
645 arboviruses and other micro-organisms in experimentally infected mosquitoes using
646 massively parallel sequencing. *PLoS ONE* **8**, e58026.
647 Kalinowski ST, Taper ML, Marshall TC (2007) Revising how the computer program cervus
648 accommodates genotyping error increases success in paternity assignment. *Molecular*
649 *Ecology* **16**, 1099-1106.
650 Katti MV, Ranjekar PK, Gupta VS (2001) Differential distribution of simple sequence repeats
651 in eukaryotic genome sequences. *Molecular Biology and Evolution* **18**, 1161-1167.
652 Kurdi-Haidar B, Shalhoub V, Dib-Hajj S, Deeb S (1983) DNA sequence organization in the
653 genome of *Cycas revoluta*. *Chromosoma* **88**, 319-327.
654 Life Technologies Corporation (2012)
655 http://find.lifetechnologies.com/sequencing/ion400base/pr. USA.
656 Malausa T, Gilles A, MeglÉCz E, *et al.* (2011) High-throughput microsatellite isolation
657 through 454 GS-FLX Titanium pyrosequencing of enriched DNA libraries. *Molecular*
658 *Ecology Resources* **11**, 638-644.
659 Mardis ER (2008) Next-generation DNA sequencing methods. *Annual Review of Genomics*
660 *and Human Genetics* **9**, 387-402.
661 Meglécz E, Costedoat C, Dubut V, *et al.* (2010) QDD: a user-friendly program to select
662 microsatellite markers and design primers from large sequencing projects.
663 *Bioinformatics* **26**, 403-404.

664 Meglécz E, Nève G, Biffin E, Gardner MG (2012a) Breakdown of phylogenetic signal: A
665 survey of microsatellite densities in 454 shotgun sequences from 154 non model
666 eukaryote species. *PLoS ONE* **7**, e40861.
667 Meglécz E, Pech N, Gilles A, Martin J-F, Gardner MG (2012b) A shot in the genome: how
668 accurately do shotgun 454 sequences represent a genome? *BMC Research Notes* **5**,
669 259.
670 Metzker ML (2010) Sequencing technologies - the next generation. *Nat Rev Genet* **11**, 31-46.
671 Morgante M, Hanafey M, Powell W (2002) Microsatellites are preferentially associated with
672 nonrepeatitive DNA in plant genomes. *Nature Genetics* **30**, 194-200.
673 Raymond M, Rousset F (1995) GENEPOP (Version 1.2): Population Genetics Software for
674 Exact Tests and Ecumenicism. *Journal of Heredity* **86**, 248-249.
675 Roche Diagnostics Corporation (2012) http://454.com/products/gs-flx-system/. USA.
676 Santana QC (2009) Microsatellite discovery by deep sequencing of enriched genomic
677 libraries. *BioTechniques* **46**, 217.
678 Schadt EE, Turner S, Kasarskis A (2010) A window into third-generation sequencing. *Human*
679 *Molecular Genetics* **19**, R227-R240.
680 Schmidt A, Doudrick RL, Heslop-Harrison JS, Schmidt T (2000) The contribution of short
681 repeats of low sequence complexity to large conifer genomes. *Theoretical and Applied*
682 *Genetics* **101**, 7-14.
683 Shendure J, Ji H (2008) Next-generation DNA sequencing. *Nat Biotech* **26**, 1135-1145.
684 Sinclair EA, Anthony J, Coupland GT, *et al.* (2009) Characterisation of polymorphic
685 microsatellite markers in the widespread Australian seagrass, *Posidonia australis*
686 Hook. f. (Posidoniaceae), with cross-amplification in the sympatric *P. sinuosa*.
687 *Conservation Genetics Resources* **1**, 273-276.

688 Sonah H, Deshmukh RK, Sharma A, *et al.* (2011) Genome-wide distribution and organization
689 of microsatellites in plants: An insight into marker development in *Brachypodium*.
690 *PLoS ONE* **6**, e21298.
691 Van Oosterhout C, Hutchinson WF, Wills DPM, Shipley P (2004) Micro-checker: software
692 for identifying and correcting genotyping errors in microsatellite data. *Molecular*
693 *Ecology Notes* **4**, 535-538.
694 You F, Huo N, Gu Y, *et al.* (2008) BatchPrimer3: A high throughput web application for PCR
695 and sequencing primer design. *BMC Bioinformatics* **9**, 253.
696 Zalapa JE, Cuevas H, Zhu H, *et al.* (2012) Using next-generation sequencing approaches to
697 isolate simple sequence repeat (SSR) loci in the plant sciences. *American Journal of*
698 *Botany* **99**, 193-208.
699 Zhang J, Chiodini R, Badr A, Zhang G (2011) The impact of next-generation sequencing on
700 genomics. *Journal of Genetics and Genomics* **38**, 95-109.
701
702

Data Accessibility:

DNA sequences: DRYAD entry for PGM doi.org/10.5061/dryad.vv82q; GS-FLX doi.org/10.5061/dryad.jd183.

Supporting Information Online:

Supplementary Material 1: Comparison between next generation sequencing platforms (a) PGM semi-conducting and (b) GS-FLX (454) pyro-sequencing in the proportional distribution of unique microsatellite sequence lengths, amongst the three taxa (MR – *Macrozamia riedlei*, PD – *Podocarpus drouynianus*, RB – *Ricinocarpos brevis*).

Supplementary Material 2: Comparison between next generation sequencing platforms (PGM semi-conducting and GS-FLX pyro-sequencing) in the proportion of different motif types (e.g. AT) within each motif class (di- and tri-nucleotide), amongst the three taxa (MR – *Macrozamia riedlei*, PD – *Podocarpus drouynianus*, RB – *Ricinocarpos brevis*). Sequences of >80 bp were used. We adopted minimal names for motifs with circular permutation and reverse complementary sequences grouped together (e.g. ATG is for ATG/CAT, TGA/TCA, GAT/ATC). Tetranucleotide, pentanucleotide and hexnucleotide classes are not graphically presented due to the small numbers within each type.

Supplementary Material 3: Simulation script

Figure legends

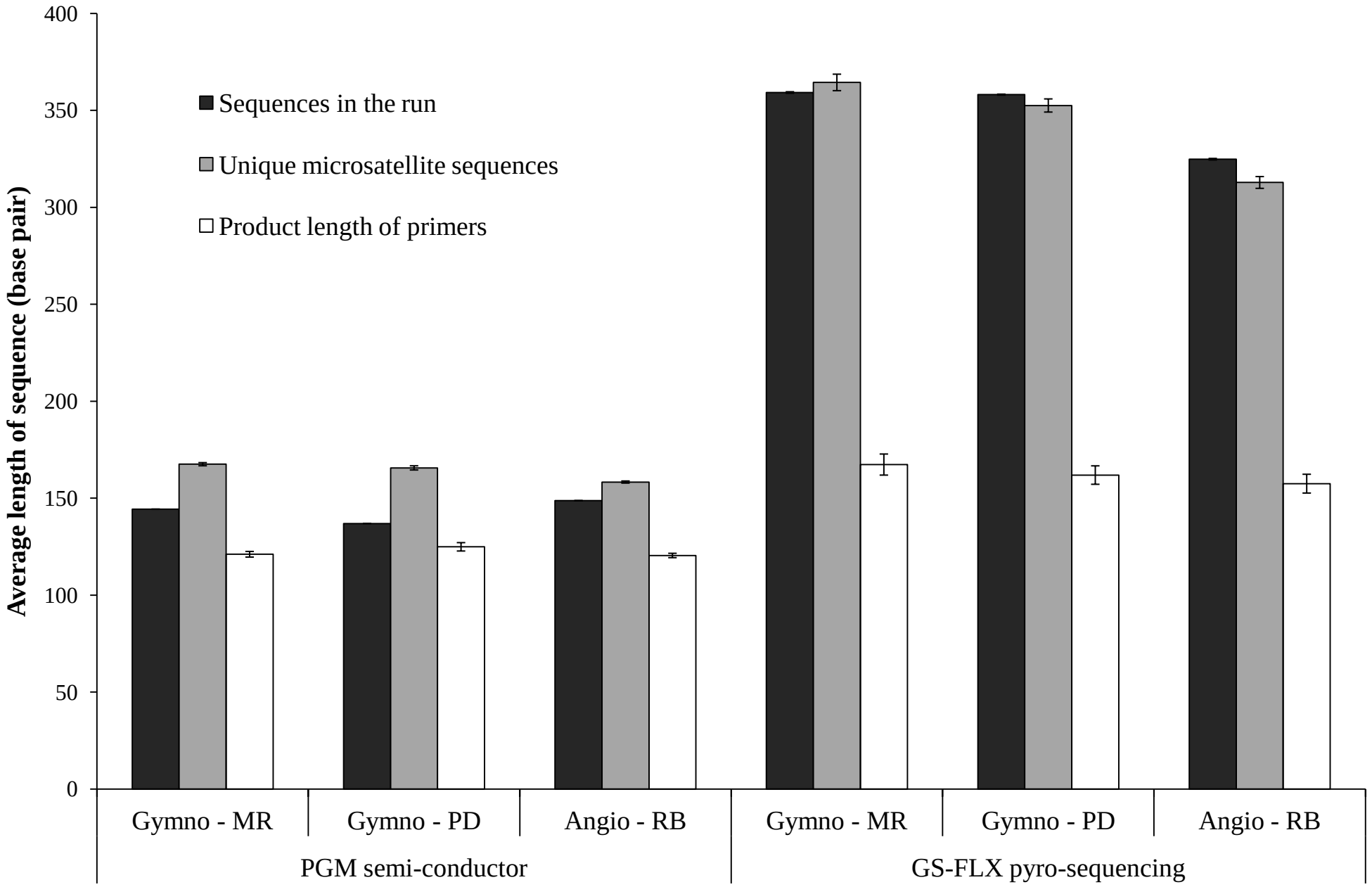
Fig. 1 Comparison between PGM (Ion Torrent) and GS-FLX (454) pyro-sequencing platforms in the average length of 1) sequences in the run; 2) unique microsatellite sequences (UMS) and; 3) the product produced from primers designed in QDD with stringent parameter settings (A+B+C designs only). The two gymnosperms (MR – *Macrozamia riedlei*, PD – *Podocarpus drouynianus*) and one angiosperm species (RB – *Ricinocarpos brevis*) are shown.

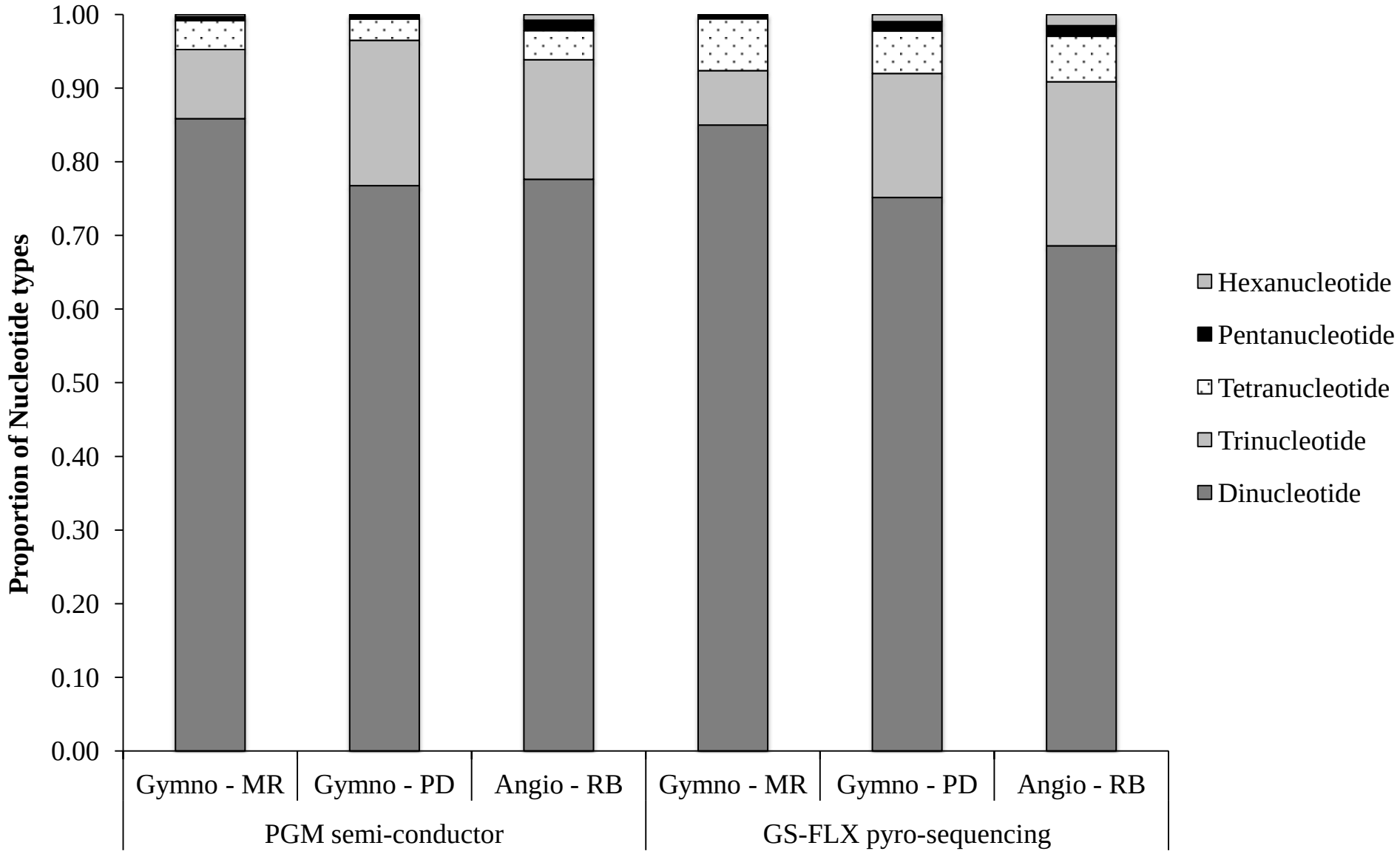
Fig. 2 Comparison of PGM (Ion Torrent) and GS-FLX (454) pyro-sequencing platforms in the proportion of unique microsatellite sequences that contain dinucleotide, trinucleotide, tetranucleotide, pentanucleotide and hexanucleotide microsatellite motifs, for the two gymnosperms (MR – *Macrozamia riedlei*, PD – *Podocarpus drouynianus*) and one angiosperm species (RB – *Ricinocarpos brevis*).

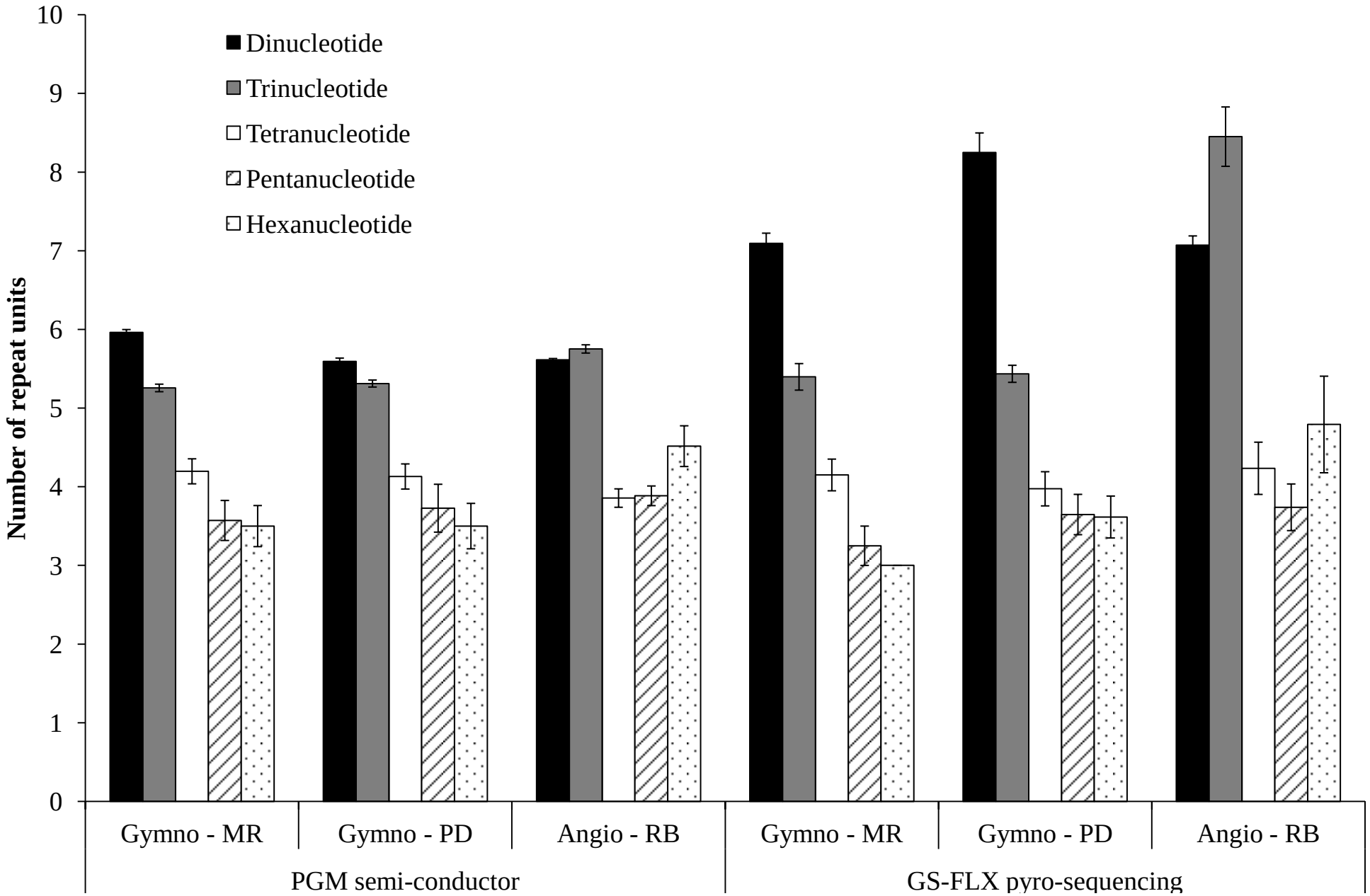
Fig. 3 Comparison of PGM (Ion Torrent) and GS-FLX (454) pyro-sequencing methods in the average number of repeat units for (a) all unique microsatellite sequences (UMS) and (b) the unique microsatellite sequences for which primers could be designed (stringent designs A-G), in each motif class (dinucleotide, trinucleotide, tetranucleotide, pentanucleotide and hexanucleotide microsatellite motifs) for the two gymnosperms (MR – *Macrozamia riedlei*, PD – *Podocarpus drouynianus*) and one angiosperm species (RB – *Ricinocarpos brevis*).

Author Contributions

C.P.E was responsible for conception and design of the study, analysed data, and wrote and finalised the paper. N.J.E and S.L.K were responsible for conception and design of the study. R.J.N.A and M.G.G were responsible for generating the sequences and providing expert advice. E.M. was responsible for generating simulation data sets and providing expert advice. J.A was responsible for preparing high quality DNA samples. S.L.K and J.A collected samples. All authors critically revised the article.


Average length of sequence (base pair)
400
350
300
250
200
150
100
50
0
Sequences in the run
Unique microsatellite sequences
Product length of primers
Gymno - MR
Gymno - PD
Angio - RB
Gymno - MR
Gymno - PD
Angio - RB
PGM semi-conductor
GS-FLX pyro-sequencing


Proportion of Nucleotide types
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
Gymno - MR
Gymno - PD
Angio - RB
Gymno - MR
Gymno - PD
Angio - RB
PGM semi-conductor
GS-FLX pyro-sequencing
Hexanucleotide
Pentanucleotide
Tetranucleotide
Trinucleotide
Dinucleotide


Number of repeat units
10
9
8
7
6
5
4
3
2
1
0
Dinucleotide
Trinucleotide
Tetranucleotide
Pentanucleotide
Hexanucleotide
Gymno - MR
Gymno - PD
Angio - RB
Gymno - MR
Gymno - PD
Angio - RB
PGM semi-conductor
GS-FLX pyro-sequencing

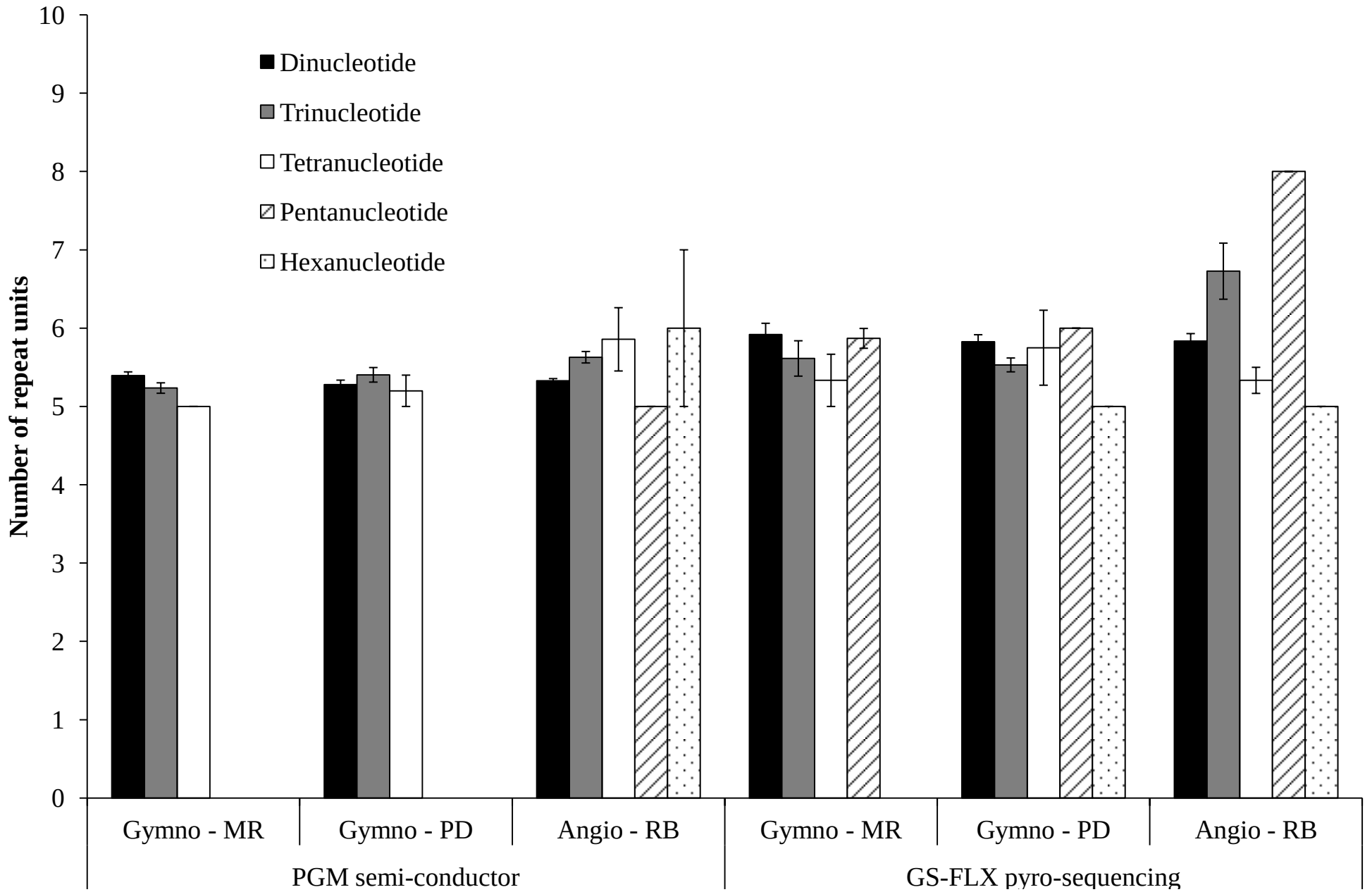

Number of repeat units
10
9
8
7
6
5
4
3
2
1
0
Dinucleotide
Trinucleotide
Tetranucleotide
Pentanucleotide
Hexanucleotide
Gymno - MR
Gymno - PD
Angio - RB
Gymno - MR
Gymno - PD
Angio - RB
PGM semi-conductor
GS-FLX pyro-sequencing

Table 1: Summary of the comparison between PGM semi-conducting (Ion torrent) and GS-FLX (454) pyro-sequencing platforms in the 1) Type of sequence data returned, 2) Proportion of microsatellite containing unique and consensus sequences; 3) Number and proportion of microsatellite motif types of the unique microsatellite sequences (UMS: unique (singleton) microsatellite containing sequence) and; 4) Number and proportion of sequences with stringent primers designed (A+B+C only), for the two gymnosperms (MR – *Macrozamia riedlei*, PD – *Podocarpus drouynianus*) and one angiosperm species (RB – *Ricinocarpos brevis*).

Next generation sequencing platform	**PGM semi-conductor**			**GS-FLX pyro-sequencing**		
Minimum Unit Cost $US (% of a run; Glenn, 2013)	~$1000 (100% of a 316 chip)			~$2000 (12% of a plate)		
Quantity of DNA used per run	0.1 ug			5 ug		
Preparation time (ligation and library preparation)	8-9 hours per run			14-16 hours per run		
Instrument running time	~5 hours per run			~12 hours per run		
SPECIES TESTED	Gymno MR	Gymno PD	Angio RB	Gymno MR	Gymno PD	Angio RB
Returned sequence data						
Reads per run*	1 085 313	916 991	1 323 518	101 787	222 517	86 905
Mbases per run*	156.6	125.6	193.7	36.6	79.7	28.2
Approximate genome coverage (proportion)	0.01	0.03	0.05	0.002	0.008	0.008
Average read length (bases)*	144.3	136.9	148.7	359.2	358.1	324.8
GC content of reads (proportion)	0.37	0.34	0.36	0.38	0.35	0.37
Proportion of reads that contained microsatellites	0.005	0.004	0.014	0.021	0.024	0.050
All microsatellite containing reads						
Proportion of unique reads*	0.58	0.55	0.41	0.28	0.21	0.29
Proportion of consensus sequences	0.06	0.05	0.04	0.13	0.09	0.04
Proportion of unused reads (redundant, low complexity, potentially repetitive regions)*	0.36	0.40	0.55	0.59	0.70	0.67
Unique microsatellite data (i.e. singletons only; UMS)						
Number of UMS	3396	2257	7287	616	1107	1264
Proportion of sequences with more than one microsatellite (range 2-12)*	0.11	0.06	0.09	0.21	0.16	0.20

Number of dinucleotide motif types (of a possible 4)	4	3	4	3	4	4
Number of trinucleotide motif types (of a possible 10)	9	8	10	10	10	10
Number of tetranucleotide motif types (total of 29 found)	20	15	22	17	16	17
Number of pentanucleotide motif types (total of 57 found)	9	9	39	2	11	14
Number of hexanucleotide motif types (total of 78 found)	11	4	51	1	12	13
Data on primer design on UMS (stringent design including A+B+C only)						
Number of UMS for which primers could be designed (stringent A+B+C)	401	232	607	134	184	176
Proportion of UMS for which primers could be designed (stringent A+B+C)*	0.12	0.10	0.08	0.21	0.17	0.14
Number of dinucleotide microsatellites with primers designed*	346	154	461	112	127	136
Number of trinucleotide microsatellites with primers designed*	54	79	131	21	54	38
Number of tetranucleotide microsatellites with primers designed	1	4	5	1	2	2
Number of pentanucleotide microsatellites with primers designed	0	0	7	0	1	0
Number of hexanucleotide microsatellites with primers designed	0	0	3	0	0	0

*$P<0.05$

Table 2a. Characterisation of polymorphic loci. NGS platform, locus name and Dryad sequence number, primer sequences, repeat motif, and diversity characteristics of nineteen microsatellite loci from *Macrozamia riedlei*. # indicates PCR cycling conditions described in the methods; Ta is the annealing temperature; ^ indicates the addition of bovine serum albumin (2%) to the PCR reaction; Na indicates number of alleles; *Ho* and *He* indicate observed and expected heterozygosity respectively; [a] indicates locus not in Hardy-Weinberg equilibrium ($P<0.05$), and likely presence of null alleles as determined by Microchecker; [b] indicates locus not in Hardy-Weinberg equilibrium ($P<0.05$); PIC indicates polymorphic information content.

Platform	Locus Seq. No. in Dryad		Primer sequence (5'-3')	Repeat motif	Allele size range	PCR #	Ta (°C)	$MgCl_2$ (mM)	Na	*Ho*	*He*	PIC
PGM	MR2	F:	GTGGGTCGTTATTGTGAGGG	AG_7	119-224	1	66	1.5^	4	0.179[a]	0.324	0.303
	MRGO2:1687:2466	R:	CCAGTGTAATGCACCTAAGGC									
PGM	MR4	F:	AGTGGCAACACCCAAGTCG	AC_7	104	1	65	2	1			
	MRGO2:2612:1824	R:	CGTGGAGACAAGCTTTCAGG									
PGM	MR6	F:	GCCCTGAGTTACCTATGCCC	AG_8	142-148	1	65	2	3	0.405	0.518	0.429
	MRGO2:1752:2233	R:	GCTAAGTGGGTCTAAACATGGC									
PGM	MR11	F:	TCTGCTTTCGACTTCTAGTTTATGC	AC_{10}	101-107	1	66	2	4	0.608	0.532	0.442
	MRGO2:694:2202	R:	AGGTCACAGAAATCATATGCG									
PGM	MR17	F:	GGAGATACAGTTCCCAGACAAGG	AG_{11}	160-173	1	52	2	9	0.855	0.738	0.746
	MRGO2:1601:1559	R:	CTGCCTCCATATACCCTTCC									
PGM	MR19	F:	AGAGAGGGCTCAACACCC	AG_8	100	1	69	3^	1			
	MRGO2:1161:1964	R:	CATAAACACTTGGAATTCTCTCTTGC									
PGM	MR22	F:	GGCACCATTTCCACATACC	$(AC)_{8(}AT)_3$	159-162	1	65	2	4	0.494[b]	0.645	0.588
	MRGO2:2632:1703	R:	TGATCAAGAGGATCCACAGC									
PGM	MR23	F:	GTTGAACCATCATTCTTAAGTCTTC	AAT_7	91-121	1	57	3^	9	0.696	0.614	0.578
	MRGO2:1113:2089	R:	TGCTTGCCTTACGATGATCC									
PGM	MR29	F:	GCTCCTCCTTCAATATGGTTAG	AAG_7	120-123	1	55	2	2	0.101	0.097	0.092
	MRGO2:451:484	R:	ATGCCAATGACCCACTTAGC									
PGM	MR30	F:	AGAAATCACCTAATGGCCAAG	AC_9	103-112	1	64.5	2	6	0.430	0.397	0.372
	MRGO2:999:2437	R:	ATGAAACGAGGATAGAAACCC									
PGM	MR33	F:	GTGCACCATGTGTCGATTTC	AG_8	146-175	1	65	2	4	0.354	0.322	0.277
	MRGO2:1439:1767	R:	GAAGTCAACCTCATAAAGACAGTG									
GS-FLX	MR37	F:	GCAAAGCTAAATGCACGAGG	AC_{14}	169-181	2	69(60)	2	8	0.812[b]	0.821	0.780
	HAT4JNG02GK0GF	R:	CCGCTTCAGACCTTATCCG									
GS-FLX	MR38	F:	AAATACAAGCCAAAGAATCATCC	AC_8	143-174	1	58	3	7	0.405	0.438	0.381
	HAT4JNG01C7LHJ	R:	TGGAAAGTCGATCTCTAGGGC									
GS-FLX	MR39	F:	AGCATCTACATCCAACACATCC	AT_9	145-155	2	70(60)	2^	7	0.618	0.613	0.580

	HAT4JNG02F8ROM	R:	GTATGGCAACCACTTCAGGG									
GS-FLX	MR40	F:	GAGGGTCATATGTGTAGAACTCCC	AC_8	168-174	1	65	3	4	0.532	0.473	0.381
	HAT4JNG01A7XAG	R:	GCAAAGGGACATAAAGATTTGG									
GS-FLX	MR41	F:	CTCTCATGCTTGCTCTATGGG	AG_{12}	202-207	2	69(57)	2	3	0.899[b]	0.511	0.391
	HAT4JNG02G2JPA	R:	GTCAGTGCCAGACCATAACG									
GS-FLX	MR42	F:	TGTAGTATACCTAGAACAAACTTTCCG	AG_{12}	243-274	1	61	2	14	0.610	0.796	0.775
	HAT4JNG02ISZM4	R:	AGAGGAACAAGGCACCAAGC									
GS-FLX	MR43	F:	TTACTTTCCCTTCCTTGAAGC	AAT_{11}	363-396	1	65	3	7	0.506	0.556	0.499
	HAT4JNG01EODTW	R:	CCATGAAGCCTTGGATTGG									
GS-FLX	MR44	F:	TGGAGTCACACAGGATAGATGC	AC_{14}	101-107	1	57	2	4	0.405	0.453	0.409
	HAT4JNG01CVICN	R:	TCTACAACCTGCAATGCTGC									

Table 2b. Characterisation of polymorphic loci. NGS platform, locus name and Dryad sequence number, primer sequences, repeat motif, and diversity characteristics of twelve microsatellite loci from *Podocarpus drouynianus*. # indicates PCR cycling conditions described in the methods; Ta is the annealing temperature; ^ indicates the addition of bovine serum albumin (2%) to the PCR reaction; Na indicates number of alleles; *Ho* and *He* indicate observed and expected heterozygosity respectively; [a] indicates locus not in Hardy-Weinberg equilibrium ($P<0.05$), and likely presence of null alleles as determined by Microchecker; [b] indicates locus not in Hardy-Weinberg equilibrium ($P<0.05$); PIC indicates polymorphic information content.

Platform	Locus Seq. No. in Dryad		Primer sequence (5'-3')	Repeat motif	Allele size range	PCR #	Ta (°C)	$MgCl_2$ (mM)	Na	*Ho*	*He*	PIC
PGM	PD3	F:	TTGATCGAAGAAGTGAAACCC	AC_8	128-133	2	75(63)	3	5	0.577	0.496	0.414
	MRGO2:1534:2376	R:	ACAGAGCCCATTTACCCACC									
PGM	PD5	F:	CCACATTGAAGTAGGCTCCG	AAG_7	146-160	2	72(60)	3	7	0.474	0.455	0.428
	MRGO2:2487:2197	R:	ATATCATCAGGTGCAAACGC									
PGM	PD6	F:	GAAGAAAGCAACCCGATCC	ACT_8	143-180	1	65	3	11	0.346[a]	0.788	0.753
	MRGO2:1570:1798	R:	GGCTTGTTGGTGAAGATTGC									
PGM	PD7	F:	TGAGCCTGCTCCATAGTTAGC	ACT_9	159-185	2	69(60)	3^	8	0.410[a]	0.715	0.674
	MRGO2:1680:2036	R:	ACCTCTCCTCCAGTGCTTGC									
PGM	PD10	F:	ATTGGTGCTATACCTACTGATGC	ACC_8	109-142	2	69(60)	3	7	0.359	0.493	0.451
	MRGO2:1703:725	R:	GGAAGAGAGTAGTGATTGATGGG									
PGM	PD15	F:	GTAAATCGGAGAAGGGAGGC	$AACG_6$	99-120	1	57	3	6	0.000[a]	0.717	0.669
	MRGO2:1330:2427	R:	CCAATTCCATTAAAGAAAGGGC									
GS-FLX	PD19	F:	TTTATGGCTCAACCACACCC	AG_{10}	96-117	2	75(63)	2	8	0.282[a]	0.746	0.695
	HAT4JNG02IITKO	R:	ACATGGAGAGCAACACCAGC									
GS-FLX	PD23	F:	AGAGTTGGATTTACCCATCTAGG	AT_{13}	96-129	2	69(60)	2	12	0.375[a]	0.813	0.785
	HAT4JNG01EYA2D	R:	GATTATTATGCCATGAACATACTCC									
GS-FLX	PD28	F:	GCTAAATCAAGCACTGGGAGC	AT_7	250-272	1	65	3	3	0.351	0.359	0.302
	HAT4JNG02GO352	R:	CAACATATGCCCTCTGTCCC									
GS-FLX	PD31	F:	CAGCATTCTCACTTCTGATACCC	$AAGC_7$	140-161	1	65	3	7	0.013[a]	0.695	0.643
	HAT4JNG02HLTQR	R:	ATAGGGTCAGTAGCGGGACG									
GS-FLX	PD32	F:	TGTAAGACTTCGAGCAATGCC	AT_8	177-194	1	65	3	7	0.289[a]	0.648	0.595
	HAT4JNG02HXJI4	R:	CGCACATTGAAACTATGATTTATGC									
GS-FLX	PD34	F:	TGTCAAACAATTCATGGACCC	AT_{10}	205-227	2	72(60)	3	11	0.360[a]	0.787	0.757
	HAT4JNG01CWFBP	R:	GCAACGGACAAACAGACG									

Table 3: Simulations of the reduced data set from the original PGM and GS-FLX platform data sets, for each species.

Next generation sequencing platform	**PGM semi-conductor**			**GS-FLX pyro-sequencing**		
	Gymno MR	Gymno PD	Angio RB	Gymno MR	Gymno PD	Angio RB
Reduced sequence data set	(Same number of reads as GS-FLX)			(Truncated sequence length like PGM)		
Proportion of run that contained microsatellites	0.005	0.004	0.014	0.007	0.006	0.018
Proportion of microsatellites from unique sequences*	0.86	0.76	0.78	0.39	0.34	0.50
Proportion of microsatellites from consensus sequences*	0.02	0.02	0.005	0.17	0.13	0.07
Proportion of unused reads (redundant, low complexity, potentially repetitive regions)*	0.12	0.22	0.21	0.44	0.53	0.43
Data on primer design (stringent design including A+B+C only)						
Number of UMS for which primers were designed (stringent A+B+C)	54	84	101	37	50	77
Proportion of UMS for which primers were designed (stringent A+B+C)	0.12	0.11	0.11	0.14	0.11	0.10

*$P>0.01$

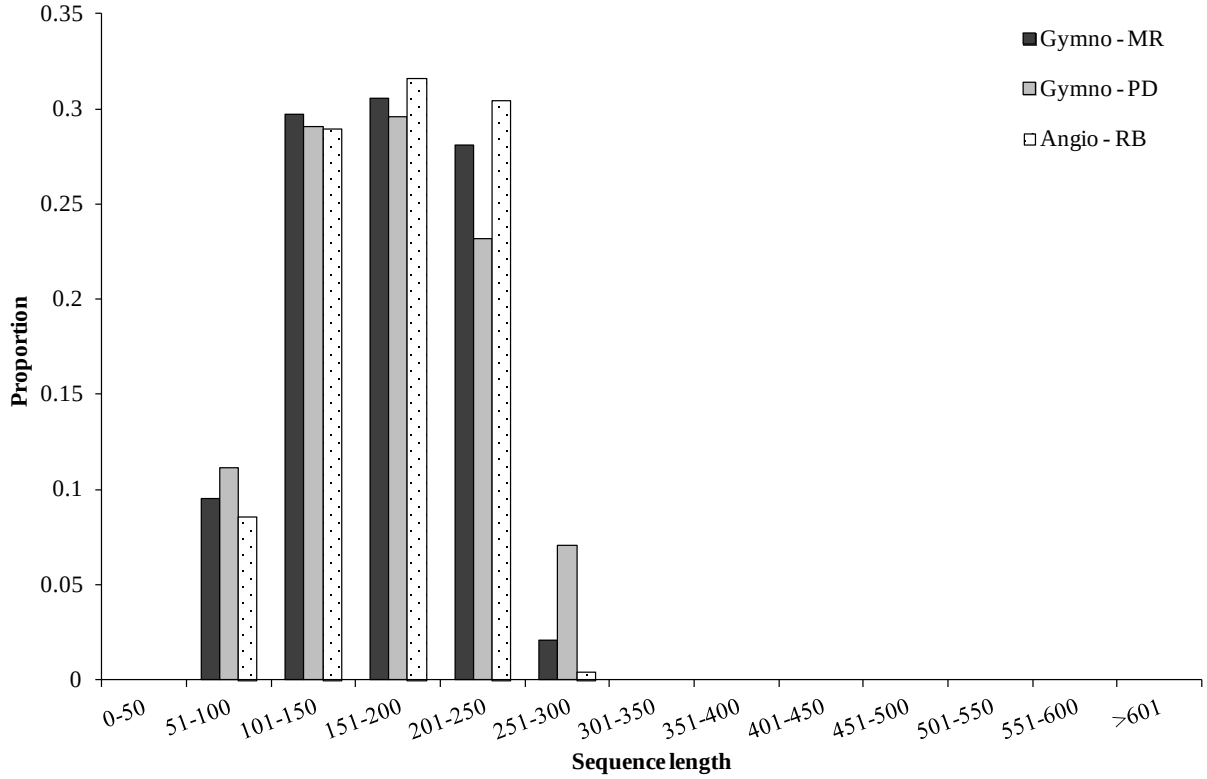
Table 4: Simulations of the PGM and GS-FLX size distributions from the human genome (HG), at four difference levels of coverage.

Next generation sequencing platform	**HG_PGM samples**				**HG_GS-FLX samples**			
Simulated sequence data								
Coverage level	0.01x	0.04x	0.07x	1x	0.01x	0.04x	0.07x	1x
Reads per run	192475	769904	1347331	1924762	79899	319609	559312	799019
Proportion of run containing microsatellites	0.03	0.03	0.03	0.03	0.07	0.07	0.07	0.07
All microsatellite containing reads	5768	22177	38813	55789	5499	22178	39070	55873
Proportion of unique reads*	0.86	0.83	0.80	0.78	0.71	0.66	0.64	0.61
Proportion of consensus sequences	0.00	0.01	0.02	0.03	0.01	0.01	0.02	0.03
Proportion of unused reads (redundant, low complexity, potentially repetitive regions)*	0.13	0.16	0.18	0.19	0.28	0.32	0.34	0.36
Proportion of grouped sequences*	0.08	0.09	0.09	0.09	0.16	0.18	0.18	0.19
Proportion of nohit_css sequences*	0.05	0.05	0.06	0.06	0.01	0.01	0.01	0.01
Proportion of multi_css sequences*	0.0002	0.0005	0.0006	0.0011	0.10	0.11	0.12	0.11
Microsatellite primer design								
Number of UMS for which primers were designed (stringent A+B+C)	690	2545	4299	6123	1056	3942	6751	9119
Proportion of UMS for which primers were designed (stringent A+B+C)	0.14	0.14	0.14	0.14	0.27	0.27	0.27	0.27

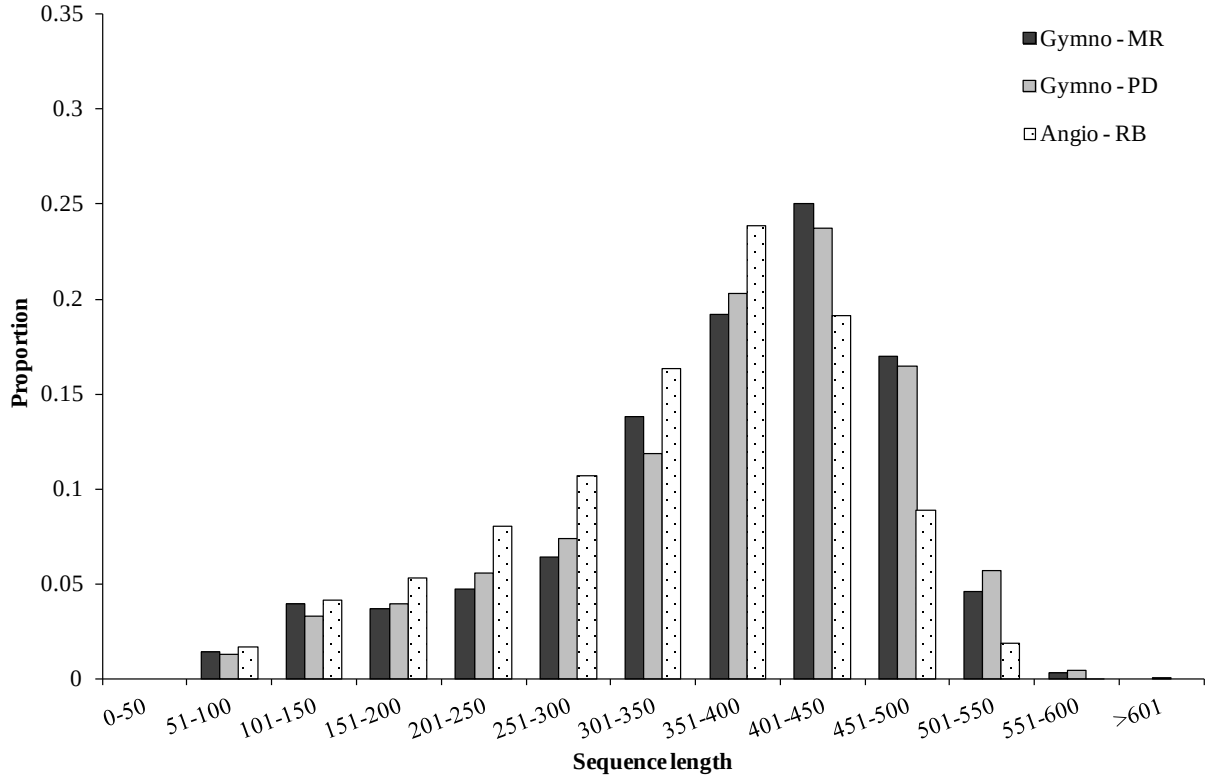
*P>0.01

Supplementary Material 1: Comparison between next generation sequencing platforms (a) PGM semi-conducting and (b) GS-FLX (454) pyro-sequencing in the proportional distribution of unique microsatellite sequence lengths, amongst the three taxa (MR – *Macrozamia riedlei*, PD – *Podocarpus drouynianus*, RB – *Ricinocarpos brevis*).

(a)

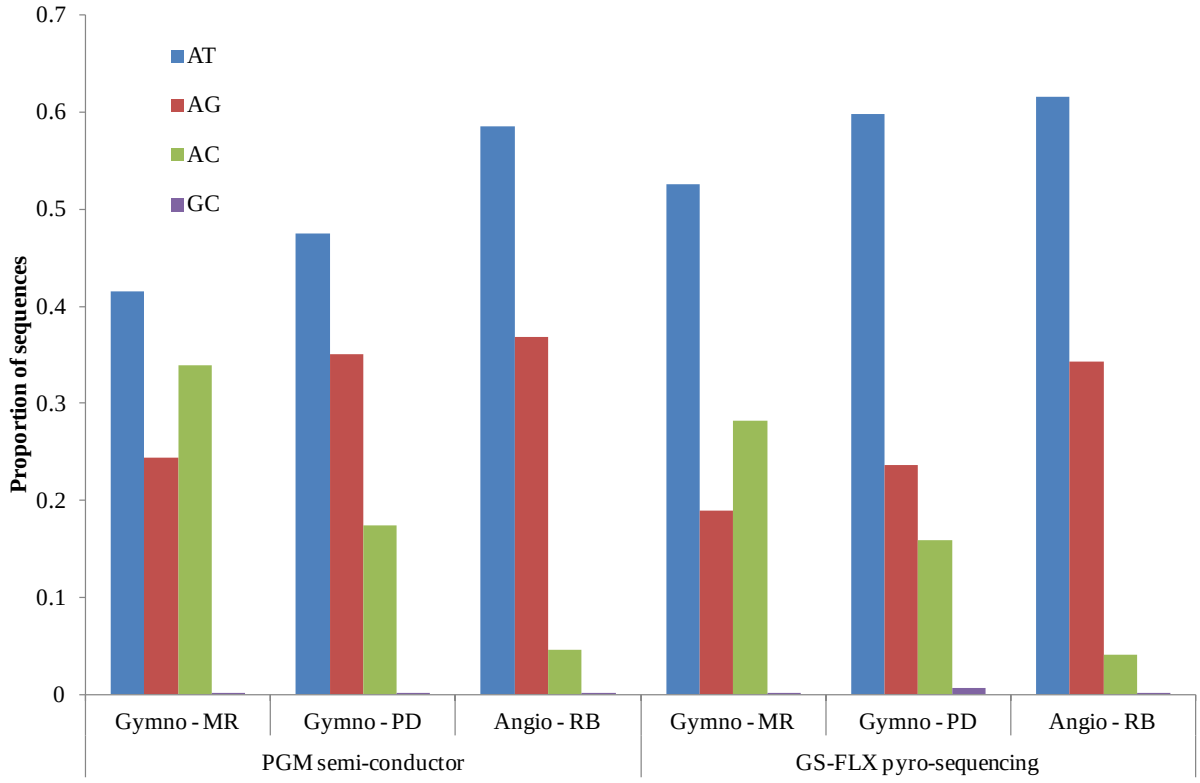


(b)



Supplementary Material 2: Comparison between next generation sequencing platforms (PGM semi-conducting and GS-FLX pyro-sequencing) in the proportion of different motif types (e.g. AT) within each motif class (di- and tri-nucleotide), amongst the three taxa (MR – *Macrozamia riedlei*, PD – *Podocarpus drouynianus*, RB – *Ricinocarpos brevis*). Sequences of >80 bp were used. We adopted minimal names for motifs with circular permutation and reverse complementary sequences grouped together (e.g. ATG is for ATG/CAT, TGA/TCA, GAT/ATC). Tetranucleotide, pentanucleotide and hexnucleotide classes are not graphically presented due to the small numbers within each type.

DINUCLEOTIDE



TRINUCLEOTIDE