DNA Research, 2018, 25(5), 535–545
doi: 10.1093/dnares/dsy023
Advance Access Publication Date: 17 July 2018
Full Paper

Full Paper

Design and evaluation of a sequence capture system for genome-wide SNP genotyping in highly heterozygous plant genomes: a case study with a keystone Neotropical hardwood tree genome

Orzenil Bonfim Silva-Junior[1,2,*], Dario Grattapaglia[1,2], Evandro Novaes[3], and Rosane G. Collevatti[4]

[1]EMBRAPA Recursos Genéticos e Biotecnologia, EPqB, Brasília, DF 70770-910, Brazil, [2]Programa de Ciências Genômicas e Biotecnologia, Universidade Católica de Brasília, SGAN 916 Modulo B, Brasilia, DF 70790-160, Brazil, [3]Departamento de Biologia, Universidade Federal de Lavras, Lavras, MG 37200-000, Brazil, and [4]Laboratório de Genética & Biodiversidade, Instituto de Ciências Biológicas, Universidade Federal de Goiás, Goiânia, GO 74001-970, Brazil

*To whom correspondence should be addressed. Tel. +55 61 3448 4792. Fax. +55 61 3340 3624.
Email: orzenil.silva@embrapa.br

Edited by Prof. Kazuhiro Sato

Received 22 December 2017; Editorial decision 20 June 2018; Accepted 22 June 2018

Abstract

Targeted sequence capture coupled to high-throughput sequencing has become a powerful method for the study of genome-wide sequence variation. Following our recent development of a genome assembly for the Pink Ipê tree (*Handroanthus impetiginosus*), a widely distributed Neotropical timber species, we now report the development of a set of 24,751 capture probes for single-nucleotide polymorphisms (SNPs) characterization and genotyping across 18,216 distinct loci, sampling more than 10 Mbp of the species genome. This system identifies nearly 200,000 SNPs located inside or in close proximity to almost 14,000 annotated protein-coding genes, generating quality genotypic data in populations spanning wide geographic distances across the species native range. To provide recommendations for future developments of similar systems for highly heterozygous plant genomes we investigated issues such as probe design, sequencing coverage and bioinformatics, including the evaluation of the capture efficiency and a reassessment of the technical reproducibility of the assay for SNPs recall and genotyping precision. Our results highlight the value of a detailed probe screening on a preliminary genome assembly to produce reliable data for downstream genetic studies. This work should inspire and assist the development of similar genomic resources for other orphan crops and forest trees with highly heterozygous genomes.

Key words: target enrichment, sequence capture, SNPs, Handroanthus

1. Introduction

The remarkable advances in whole-genome sequencing (WGS) and genome-wide genotyping technologies have provided unprecedented opportunities to gain detailed insights into the overall patterns of genetic variation underpinning the basis of adaptive evolution.[1–4] Such advances took place first in species for which reference genome sequences and high-throughput single-nucleotide polymorphism (SNP) genotyping were readily available, including humans[5] and model plant species.[6,7] However, little advance has taken place for the vast majority of plant species besides the mainstream crops,[8] some minor ones[9] and the main plantation forest trees.[10]

The dropping costs and increasing data yields of next-generation sequencing (NGS) technologies have fostered the development of several sequencing-based genotyping methods that allow for simultaneous discovery and genotyping of very large numbers of markers.[11] Among these methods, targeted enrichment or sequence capture[12] has been increasingly used in a large array of organisms.[13] Early reports published in humans[14,15] and in plants such as maize,[16] eucalypts,[17] poplars[18] and pines[19] showed the potential of the capture systems to assay sequence variants across the genomes. Sequence capture can significantly reduce costs in comparison to WGS because only specific loci of interest are sequenced at high coverage, increasing confidence of true variant detection. The practical advantages and increasing accessibility of target sequence capture methods have been reviewed in the context of evolutionary and ecological genomics, predicting a rapid expansion of this approach.[20] For example, targeted enrichment was recently used to increase phylogenetic resolution within the Neotropical tree genus Inga (Leguminosae),[21] generating alignments of over 0.3 Mbp of coding sequences revealing nearly 5,000 informative sites. Although providing large volumes of phylogenetically informative sequence data, analyses indicated that to further enhance understanding of the evolution of Inga, for instance by increasing ability to discriminate among more parameter-rich models and more recent divergence histories, a much larger set of sites, covering several thousand loci, would be needed.

Handroanthus impetiginosus (Mart. ex DC.) Mattos (syn. *Tabebuia impetiginosa*, Bignoniaceae) commonly known as Pink Ipê or Lapacho tree, is a Neotropical hardwood widely distributed in seasonally dry tropical forests (SDTFs) of South America and Mesoamerica.[22,23] Our previous studies with *H. impetiginosus* populations in Central Brazil have revealed incomplete lineage sorting and a highly structured genetic diversity,[23] representing an excellent system to better understand evolutionary processes that result in correlation between ecological and genetic differentiation in tropical biomes. We are now interested in disentangle the uncertainty in relating phylogenetic inference of ancestral geography and genealogical histories of individual loci in *H. impetiginosus*. Attempts to reconstruct species phylogeny using gene sequences has shown to be affected by the number of loci used to estimate the phylogeny and the number of individuals sampled per species.[24] The availability of a high-throughput genotyping system at large numbers of loci in a reasonable number of individuals can provide, despite the presence of retention and sorting of ancestral polymorphism, clues on the occurrence of past contacts that may have facilitated evolutionary divergence of modern population sub-groups. From a practical standpoint, an improved knowledge of the evolutionary history and population genomics of this highly exploited species may also aid the precise identification of geographical origin of timber trade in forensic applications to combat the significant illegal trading pressure[25] and help define management units for conservation planning.[26]

We have recently generated a preliminary genome assembly and annotation of the nuclear genome sequence of a single individual of *H. impetiginosus*.[27] Using these resources as reference, in this work we designed and carried out an evaluation of a set of 30,795 sequence capture probes for the identification and genotyping of nearly 200,000 SNPs located inside *circa* of 14,000 annotated protein-coding genes or in close proximity to them to support multiple sequence-based genetic analyses in the species.

2. Materials and methods

2.1. Biological material and sequencing

Total genomic DNA samples from six unrelated trees of *H. impetiginosus* were extracted using the Qiagen DNeasy Plant Min kit (Qiagen, DK) and pooled in equimolar amounts to be whole-genome sequenced. These six trees used in the pooled WGS were part of the 24 trees later used for the sequence capture system. They were sampled at the rate of one per each population group in an attempt to sample a representative sequence diversity of the species for designing probes. One barcoded DNA library was built on the pooled sample and paired-end sequencing of this library (2 × 150 nt) was performed in half lane of an Illumina HiSeq 2500 instrument (Illumina, CA, USA) at the High-Throughput Sequencing and Genotyping Center of the University of Illinois Urbana-Champaign, USA. DNA samples of 24 *H. impetiginosus* individual haplotypes sampled from six subpopulations spanning a wide geographic range of the species (Supplementary Table S1) were used for the development and evaluation of the sequence capture system.

2.2. Sequence-capture probe design

Two different sets of probe sequences were designed. The first one targeted transcript sequences of protein-coding genes predicted from the genome assembly of *H. impetiginosus* UFG-1.[27] These probes were designed by RAPiD Genomics LLC, Gainesville, FL, USA, from a total of 17,424 transcript sequences, which were tiled with 26,526 120-mer DNA sequence probes. The second set, designed by our group, sampled SNPs identified in pooled low coverage WGS data. We used nearly 20 Gbp of data (70,049,993 pairs of reads) for the SNP calling process. From the empirically determined haploid genome size of 557 Mbp/1C for *H. impetiginosus*,[27] this read set corresponded to nearly 3× of sequencing coverage in the pool of 12 chromosomes. Based on the 23,508 variants derived from the GATK[28] and the 9,079 from the Cortex[29] analysis, the genome assembly of the species was used as a template for the design of 120-mer probes surrounding the SNP vicinity. A total of 4,269 120-mer DNA probe sequences were extracted from the template genome for subsequent use. The 4,269 WGS derived probes together with the 26,526 probes from transcript data, composed the capture system with a total of 30,795 probes. Additional details on the design of these probes are described in Supplementary File S1.

2.3. Sequence-capture data generation

Rapid Genomics LLC (Gainesville, FL, USA) Capture-Seq method provided target DNA enrichment across this custom probe set on the 24 individuals using an 8-plex capture pool.[19] The products of multiplexed reactions were sequenced in a single lane of an Illumina HiSeq2000 instrument, 1 × 101 bp.

2.4. Analyses of sequence capture data and definition of target loci

Sequencing reads were trimmed for adapters and aligned to a genome sequence assembly of *H. impetiginosus*[27] with BWA[30] version 0.5.9 using BWA-backtrack algorithm, default settings. These steps were performed in a grid computing cluster software system using a pipeline available from the International Cassava Genetic Map Consortium.[31] Preliminary analysis of variants was performed by joint-sample SNV calling and genotyping of the 24 individuals using GATK-GenotypeGVCFs method on gVCF files produced by HaplotypeCaller on each sample (options: *-mbq 10 -mmq 10 -hets 0.01 -ERC GVCF -ploidy 2*). For the resulting VCF file, the raw SNPs were analyzed. Coordinates of the 30,795 sequence probes on the genome were reputable initially as 'on-target' regions, while all the complementary intervals on the genome assembly were considered 'off-target'. Following the preliminary analysis of on-target and off-target variants, sequence coordinates for the probes passing preliminary quality control were adjusted to include 200 bp in each direction. All book-ended ('touching') entries were then merged into a single interval using, respectively, *slopBed* and *mergeBed* utilities from the BedTools v2.25 package.[32] A set of non-overlapping intervals thereafter called *loci* was defined. This criterion was set to reduce the negative impact of unintended sequence capture while allowing for accurate variant detection in genomic regions contiguous to the region of interest thus extending the utility of the targeted sequencing to novel genomic sequences that would be more divergent from the probe sequence. It also allowed the identification of faulty or improper probe sequences in our initial design. Based on this criterion, the coordinates of the 30,795 probe sequences have determined 23,232 distinct target loci across the genome assembly. These encompass 19,962 loci determined by the coordinates of probe sequences from transcripts and 3,270 loci for the low-coverage WGS. For the assessment of the capture efficiency, only the defined loci were considered as callable regions for on-target SNP discovery and further genotyping, while calls outside these regions were discarded. Sequencing coverage and capture efficiency for loci determined by probe coordinates from transcripts and low-coverage WGS were computed per sample. Coverage was quoted as the ratio of the aligned read depth, which denotes the number of quality reads after alignment to the genome assembly, to the number of sequenced loci in the corresponding sample. Capture efficiency was estimated as the ratio of the number of loci for which at least one quality SNP was detected to the total number of loci. Finally, following GATK best practices and other recommendations to improve variant calling in NGS data,[33] we devised a procedure to establish useful quality dataset for further refinement of the SNP calls in the defined target loci as described in the Supplementary File S1.

2.5. Performance of the initially designed probe sequences for target enrichment and capture

Bioinformatics filtering steps were developed to obtain quality SNP data across the defined 23,232 loci. The successive application of these filters led us to define call sets that encompassed a collection of variant calls across loci and probe sequences with increasing stringency, which were named STANDARD, GQ20+VQSR and GQ20+VQSR+MM80 (see Results for definitions). To detect improper designed or faulty probes for reliable SNP detection, capture efficiencies across these call sets were inspected by counting the number of probe sequences across target loci for which at least one quality SNP was detected. A success rate for probes was obtained by dividing the number of successful probes by their totals in the initial design, i.e. 30,795, whether the source of design were transcripts (26,526 probes) or low-coverage WGS (4,269 probes). Conversely, a failure rate was defined as (1 − success rate). The Success/Failure rate was also stratified according the probe location in the genome assembly for predicted gene model features or intergenic region.

2.6. Reproducibility of the capture assay

Reproducibility was assessed in terms of SNP recall rate and precision between technical replicates conducted in the same laboratory. Reproducibility was assessed for a partial set of 14,135 sequence capture probes selected out of the 30,795 probes (23,232 loci). The coordinates of these 14,135 probe sequences fell within 11,026 loci on the genome assembly. Nearly 50% (5,978 out of 11,026) of these loci were randomly selected in the well-curated GQ20+VQSR set of probes. It included 6,843 probe sequences from transcripts and 1,313 probes from low-coverage WGS. The remaining 5,048 loci (3,023 probe sequences from transcripts and 2,956 from low-coverage WGS) were taken randomly from the remaining set of loci. For an adequate reproducibility assay, the same level of DNA pooling for the capture assay, i.e. 8 samples per pool, and sequencing effort, i.e. 24 barcoded samples in one lane of a HiSeq2000 instrument, were kept between replicates. The sequence data in the first assay (called HIMP-1 thereafter) and the second partial replicate (called HIMP-2 thereafter) were processed using the analytical protocol developed in this study, including variant recalibration using a database of likelihoods for markers in the HIMP-1 study. VCF files for the two assays were combined and evaluated as described in Supplementary File S1.

2.7. Genomic variant characterization

We combined the sequence data in HIMP-1 and HIMP-2 capture assays and used our pipeline to compile a catalog of SNPs across the genome assembly. SnpEff[34] version 4.3t was used to annotate variants in the genome assembly based on their targeted regions and predicted coding effects. As a first glimpse of the ability of this sequence capture system to carry out population genomics studies of *H. impetiginosus*, we used the SNPs data to look at the site frequency spectrum (SFS) and some population genetics features. To account for the fact that the sequence capture system provides data in 'blocks', with SNPs likely in complete linkage disequilibrium (LD), we used Plink v1.90b2h[35] to greedily prune SNPs. A window size of variant count was used such that no SNP pairs remained with r^2 greater than a threshold for tight LD (–indep-pairwise 50 5 0.5), and only included SNPs for which missing data was ≤20% (–geno 0.2). We used the data of the LD-pruned SNPs to obtain estimates of population heterozygosity, the ratio of transitions and transversion (Ts/Tv) and the number of SNP with heterozygous or non-reference homozygous genotypes (het/non-ref-hom).

3. Results

To facilitate following the sequence of methods and results obtained along the development and evaluation of the sequence capture system, a comprehensive flowchart is presented (Fig. 1). The overall process involved five steps, starting with the sequence probe design, all the way to the definition of a set of loci (non-overlapping intervals), recognition of improper or faulty probes, and genome annotation of the sampled SNPs. Each step is detailed in its components

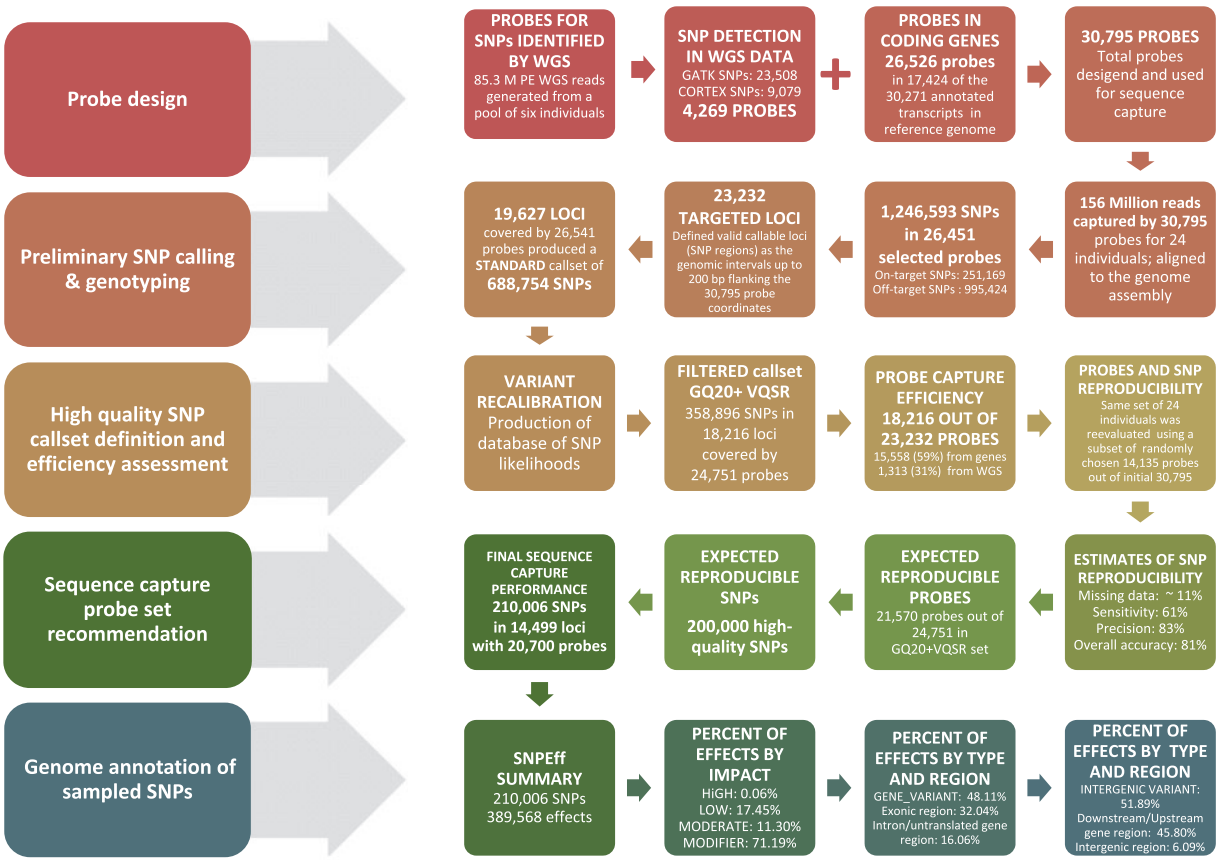

Figure 1. Flowchart of the sequence of steps and corresponding input and output results in terms of sequence data, probes and SNPs obtained along the development and evaluation of the sequence capture system for *Handroanthus impetiginosus*.

such that one may follow the numbers of input sequences, probes or SNPs, and the resulting output numbers achieved.

3.1. Sequence capture data

A total of 156,896, 801 single-end reads were generated from sequencing of DNA fragments captured by the 30,795 probes in the sample of 24 individuals of *H. impetiginosus* (mean: 6,537,367; sd: 2,865,856) (Supplementary Table S2). Alignment of the reads to the genome assembly was determined for each sample. We observed a varying fraction of unmapped reads which ranged from 10% to nearly 40% for some individuals. For mapped reads, mappability within targeted regions was further evaluated using HCMappingQualityFilter in GATK HaplotypeCaller (-mmq 10, default value). Overall good read mappability was achieved, with only 3–4% of the reads being discarded prior to variant calling.

3.2. Preliminary analyses of SNPs and definition of target loci

A total of 251,169 SNPs were located in the genomic intervals of the 30,795 probe coordinates ('on-target variants'). We identified a set of targeted regions that consistently produced data in the sampled individuals. For three samples the sequencing coverage was <40× (Supplementary Table S2). With the exclusion of these samples, a total of 26,451 probe sequences successfully captured on-target SNPs. In our evaluation of off-target variants (see Supplementary File S1), we found that additional 134,126 quality SNPs were spread across the genome assembly in regions up to 25 kbp from the closest targeted location (Fig. 2). These distances were unevenly distributed and most of them (75% of the total off-target SNPs) were only 200 bp upstream or downstream to the targeted location on the genome assembly. Sequence coordinates for these 26,451 probes passing the preliminary analyses were adjusted then to include 200 bp in each direction (see Materials and Methods). A set of 19,627 capturable *loci* was thus defined (average length: 607 bp; median length: 520 bp) and none of them had matches to annotated coordinates to repeats in the genome assembly of *H. impetiginosus*. Consequently, 4,344 probe sequences (~14% of 30,795) were considered improper in the design of the capture system.

3.3. Improvement procedure for SNP calling and genotyping refinement

We applied a bioinformatics protocol for SNP discovery and progressive filtering of low-confidence variants across the 26,451 120-mer sequences and the 19,627 targeted loci defined across this probe set. Overall, the numbers of aligned reads were 63,868,629 (~41% of the total reads) and 107,275,224 (~60% of the total reads) across probe and loci coordinates, respectively. The average sequencing coverage and standard deviation for the samples across the loci was 136x ± 75x (9x – 304x). Sequencing coverage and capture efficiency for loci determined by probe coordinates from transcripts and low-coverage WGS were estimated per sample (Table 1). Considering the samples with capture efficiency greater than 60%, the coverage was equal or greater than 70x. This coverage is enough

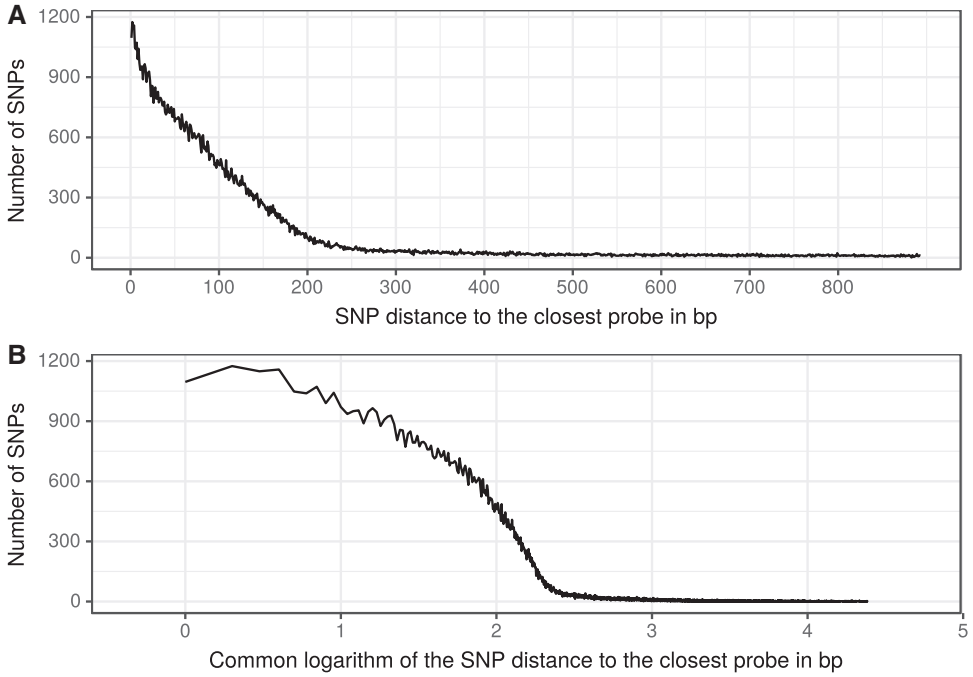

Figure 2. Distribution of off-target SNPs distances to the closest probe sequence coordinate. (A) The narrow spectrum of distances indicates that most of the identified SNPs are located up to 200 bp upstream or downstream to a targeted location in the genome. (B) The wide spectrum of distances shows that off-target variants were found spread across the genome in regions up to 25 kbp from the closest targeted probe location.

Table 1. Summary of per sample coverage and capture efficiency over the 23,232 target loci in the genome assembly of *H. impetiginosus*

Sample	Coverage[a]	Number of sequenced loci out of the 23,232[b]	Efficiency over 19,962 loci determined by probes from transcripts[c(a)]	Efficiency over 3,270 loci determined by probes from low-coverage WGS[c(b)]	Efficiency over all 23,232 loci[c(c)]
POS-15-1	304	19,627	72%	63%	71%
POT-05-1	265	18,951	70%	66%	69%
CAR-05-1	210	19,077	70%	67%	69%
POS-10-1	211	18,671	68%	63%	68%
SEC-23-1	217	18,944	70%	63%	69%
SEC-08-1	245	18,784	69%	65%	68%
SEC-05-1	146	18,524	68%	63%	68%
CAR-02-1	100	18,047	66%	63%	66%
SUM-13-1	132	17,909	68%	47%	65%
SUM-12-1	114	17,797	68%	46%	65%
MOC-08-1	112	17,877	70%	37%	65%
MOC-11-1	136	17,788	69%	44%	65%
MOC-02-1	148	17,765	68%	41%	64%
MOC-09-1	129	17,547	68%	39%	64%
MOC-12-1	133	17,625	68%	42%	64%
MOC-13-1	125	17,705	68%	43%	64%
MOC-07-1	118	17,524	67%	41%	63%
MOC-03-1	112	17,436	67%	41%	63%
SUM-10-1	72	17,443	66%	48%	63%
MOC-06-1	94	17,226	66%	42%	63%
MOC-10-1	70	16,986	65%	43%	62%
MOC-05-1	25	16,652	62%	45%	60%
SEC-09-1	9	16,229	58%	65%	59%
MOC-04-1	28	15,942	61%	44%	58%

[a]Coverage is the ratio of the aligned read depth, which denotes the number of quality reads after alignment to the genome assembly, to the number of sequenced loci in the corresponding sample.

[b]Denotes the number of target loci in the initial design of the capture system for which at least one quality read was detected after alignment to the genome assembly.

[c]Capture efficiency was computed at sample level by the ratio between the number of loci for which at least one quality SNP was detected using the improvement procedure for SNP calling and genotyping refinement and the total number of loci determined by probes in the corresponding design: 19,962 loci from transcripts of protein-coding genes (c(a)); 3,270 loci from low-coverage WGS (c(b)); total of 23,232 loci (c(c)).

Table 2. Summary of the numbers of loci, size in base pairs (bp), numbers of probes, SNPs, gene models and intergenic regions retained following each filtering step in the variant analysis using the GATK framework

SNP call set	Number of loci	Target loci			Number of probes	Number of SNPs	Number of gene models
		Total size (bp)	Mean (bp)	Median (bp)			
STANDARD	19,627	11,913,787	607	520	26,451	688,754	11,024
GQ20+VQSR	18,216	10,983,469	603	520	24,771	352,879	10,400
GQ20+VQSR+MM80	11,748	7,392,636	630	520	16,901	83,476	7,991

See Results for the definitions of the SNP call sets.

to produce accurate joint genotype calls for common variants across the samples (minimum allele frequency, MAF > 5%) using per-sample read depth as low as 3 (70x0.05) at the variant site. The call set across these target loci resulted in 688,754 SNPs. It was named 'STANDARD' and its main features are shown in Table 2.

Additional filtering based on SNP quality score recalibration using a custom database of likelihoods and genotyping refinement steps (see Supplementary File S1) resulted in a set of 352,879 SNPs (51.2% of the variants in the STANDARD call set). Coordinates of these SNPs were within 18,216 targeted loci (≈11 Mbp), encompassing the locations of 24,771 probe sequences (Supplementary File S2). This call set was named 'GQ20+VQSR' (Table 2). With this improvement procedure, an additional set of 1,680 probe sequences were deemed faulty for SNP detection. Together with the preliminary analysis, this additional step to detect faulty probes resulted in the exclusion of 6,024 (4,344 + 1,680) probes. It represents ∼20% of the probes in the initial design. Based on our assessment of genetic diversity from the sequence data, using average θ (the population mutation rate) of 0.009 (Supplementary Table S7), and sample size ($N = 21$ diploid individuals), we found that the number of quality SNPs closely matches the expectation of ∼367,000 polymorphic sites using the Watterson estimator[36] (=18,216 × 520 × 0,009 × 4.3), where 4.3 is the partial sum of the harmonic series with $(2*N - 1)$ terms. Finally, we applied a threshold of less than 20% of missing data for any site on the former 'GQ20+VQSR' call set. Considering the 21 individuals, the number of retained polymorphic SNPs dropped from 352,879 to 83,476, within 11,748 loci (∼65% of 18,216). This call set resulting from all three filters applied was named 'GQ20+VQSR+MM80' (Table 2).

3.4. Efficiency of the capture system

To assess the efficiency of the system to detect genomic variants, we considered the initial number of 30,795 probe sequences and 23,232 distinct target loci. The number of loci targeted by probes in which we successful detected quality SNPs ranged from 11,748 (50.6% of 23,232) to 18,216 (78.4% of 23,232) across quality detected variants, whether we accounted for polymorphism (MAF > 5%) and low-missing data at the genotype level (<20%) or only polymorphism requirement, respectively. Table 3 summarizes the capture efficiencies for probes across the progressively filtered call sets whether the source of design were transcripts or low-coverage WGS and considering the location of the probes in the genome assembly.

3.5. Technical reproducibility and the performance of the sequence capture system for genotyping

Sequence data generated in the partially replicated assay, named HIMP-2, produced 105,665,340 reads in total. Considering the sequencing intervals defined by the 14,135 probe sequences and 11,026 loci in this second assay (see Materials and Methods), 52,717,622 reads (∼50% of the total reads) and 75,394,548 reads (∼71% of the total reads) were aligned across on-target regions, respectively. After SNP detection using the improvement procedure for variant analysis, the VCF files contained records of 129,423 SNPs in HIMP-1 and a total of 175,361 in HIMP-2. The overall sensitivity—the fraction of 'true called SNPs' to 'true SNPs'—was 61% while the precision—the fraction of 'true called SNPs' to 'called SNPs'—was 83%. False positive (FP) rate, calculated as the FP/(FP + TN) ratio (TN is the true negative rate), was 8% while the overall accuracy was 81%. Contingency table values were summarized (Supplementary Table S3). The true positive (TP) set of SNPs consisted of 93,441 variants with high confidence to the genotype calls between replicates (see Supplementary File S1 and Table S4).

For the 14,135 probe sequences and 11,026 loci (47.5% of the 23,232 total targeted loci) evaluated across replicates, the overall performance of the capture system was 84% (8,315 out of 9,866) for the transcript-based probes and only 12% (524 out of 4,269) for the WGS-designed probes. Taking into account only the well-curated set of probes (GQ20+VQSR), the efficiency between replicates for transcript-based probes was 6,651 out of 6,843 (97%) and 430 out of 1,313 (33%) for probes from low-coverage WGS. Applying these results to the complete set of 24,751 probe sequences in the GQ20+VQSR set, we infer that *circa of* 21,570 (20,941 × 97% + 3,810 × 33%) probes constitute a highly efficient capture-based genotyping system for the species. Regarding the SNP call and genotyping confidence and considering the overall accuracy we estimated that the GQ20+VQSR probe set can sample nearly 204,000 true SNPs [93,441 × (23,232/11,026) × (21,570/24,751) × (1 + (1 − 81%))] across 18,216 loci (11.1 Mbp) in the genome assembly. Sample-to-sample variation in terms of read counts indicates that this outcome is influenced by the quality of the input DNA and sources of technical variation not evaluated here.

3.6. Genomic variant annotation and functional effect prediction

For the 18,216 genomic regions considered reliable analysis targets of this capture system, we produced an annotation of the genetic variants to improve the characterization of the recently reported genome assembly of *H. impetiginosus* and allow for further genetic research and candidate genes studies in this species. Sequence data in HIMP-1 and HIMP-2 capture assays were combined and processed in our pipeline (see Data and scripts availability) to compile a list of high quality SNP sites across the genome assembly. Given the high sequencing coverage acquired following the combined data sets of the two replicates, the number of quality SNPs found was 210,006, which agrees with our previous expectation of 204,000 SNPs based

Table 3. Performance of the initially designed probe sequences for target enrichment and capture

Source	Filter	Genic region						Intergenic region		Total	
		Exon		Exon + flanking		Intron					
		#of successes	#of failures	#of successes	#of failures	#of successes	#of failures	#of successes	#of failures	#of successes	#of failures
Probes from transcripts	STANDARD	3,062 (88%)	405 (12%)	16,208 (90%)	1,870 (10%)	1,163 (83%)	230 (17%)	2,070 (58%)	1,518 (42%)	22,503 (85%)	4,023 (15%)
	GQ20+VQSR	2,872 (83%)	595 (17%)	15,197 (84%)	2,881 (16%)	1,051 (75%)	342 (25%)	1,821 (51%)	1,767 (49%)	20,941 (79%)	5,585 (21%)
	GQ20+VQSR+MM80	2,447 (71%)	1,020 (29%)	11,529 (64%)	6,549 (36%)	511 (37%)	882 (63%)	1,101 (31%)	2,487 (69%)	15,588 (59%)	10,938 (41%)
Probes from low-coverage WGS	STANDARD	102 (90%)	11 (10%)	1,813 (94%)	106 (6%)	214 (94%)	14 (6%)	1,819 (91%)	190 (9%)	3,948 (92%)	321 (8%)
	GQ20+VQSR	102 (90%)	11 (10%)	1,799 (94%)	120 (6%)	209 (92%)	19 (8%)	1,700 (85%)	309 (15%)	3,810 (89%)	459 (11%)
	GQ20+VQSR+MM80	69 (61%)	44 (39%)	966 (50%)	953 (50%)	71 (31%)	157 (69%)	207 (10%)	1,802 (90%)	1,313 (31%)	2,956 (69%)

Capture efficiencies across filtered call sets (see Results for the definitions of the filter criteria) were inspected by counting the number of loci and probe sequences for which at least one quality SNP was detected. A success rate was obtained by dividing the number of successful probes by their totals in the initial design, i.e. 30,795 probe sequences, whether the source of design were mRNA (26,526 probes) or low-coverage WGS (4,269 probes). Conversely, a failure rate was defined as (1 − success rate). The Success/Failure rate was also stratified according the probe location in the genome assembly for predicted gene model features or intergenic region.

on the assessment of the sequence capture efficiency. The average sample call rate was 92% (82–99%) and the average SNP call rate was also 92% (62–100%). The 210,006 high-quality candidate SNPs were found in 14,499 genomic regions of 3,973 scaffolds in the assembly. These scaffolds encompass ∼290 Mbp of sequence in the 557 Mbp/1C genome of the species. SnpEff annotated 389,568 effects with the largest numbers of effects related to SNPs located in downstream or upstream gene regions (45.80%), exons (32.04%), introns and untranslated gene regions (16.06%) and intergenic regions far from genes (6.09%).

4. Discussion

We have described the development and evaluation of an initial collection of 30,795 probe sequences for SNP discovery and genotyping in the highly heterozygous plant genome *Handroanthus impetiginosus*. These probes were used as input to a commercially available system for custom target DNA enrichment followed by Illumina sequencing in a sample of 24 trees from six population groups spanning a wide geographic range of the species. With the aid of a genome assembly of a single tree, we performed extensive evaluation of the capture system to provide recommendations for future developments of similar plant genomes. It resulted in a final set of 24,751 capture probes covering 18,216 distinct loci, sampling more than 10 Mbp of the species genome. Variant analysis and SNP discovery allowed characterization of 210,006 SNPs and genotyping with overall level of missing data low for each site (<20%) across samples. We investigated issues such as probe design, sequencing coverage and bioinformatics protocols including the evaluation of the capture efficiency and a reassessment of the technical reproducibility of the assay for SNPs recall and genotyping precision.

The methods described here for target enrichment share some common aspects to other genome complexity reduction approaches such as genotype by sequencing (GBS) and RAD sequencing and its variations as they also are intended to provide large numbers of SNPs at a lower cost per sample.[37,38] Several additional challenges are faced, however, by these alternative genome complexity reduction methods for direct SNP genotyping. The low sequencing coverage associated to the variable efficiency of DNA digestion results in inconsistent sampling of loci, seriously affecting reproducibility across experiments that tend to exacerbate in heterozygous genomes,[39] ultimately biasing diversity measures.[40–42] The final number of robust SNPs assayed in large samples is typically only a small fraction of the initial set, defeating the alleged cost advantage.

A direct comparison of our sequence capture system with alternative genome complexity reduction methods for direct SNP genotyping is evidently beyond the scope of our research. Studies approaching this issue, however, pointed out that sequence capture consistently provided more reliable and portable data across experiments than RAD for shallow systematics, such as within species.[43] Such observations have, indeed, lead to the recent development of combined protocols that ultimately benefit from the lower cost of RAD and the robustness of targeted capture.[44]

It should be said, however, that laboratory methods for capture systems involve decisions which also have the potential to significantly impact the capture efficiency outcome.[45] Finally, while whole-genome re-sequencing analysis has also been trending for natural population studies, it is still largely cost-prohibitive for the vast

majority of species that are investigated on very low budgets even those with relatively small genomes. Because we were aware of all these issues regarding the challenges of reduced complexity-based methods, we devised several strategies in our analytical pipeline to assess the quality of the sequence data generated in the assay, including a partial replication of the experiment.

4.1 Quality of sequence data and challenges in bioinformatics

The alignment of the sequencing reads obtained for each sample to the genome assembly is an important challenge. Accurate aligner algorithms such as BWA implement mismatch penalty scheme in which sequencing error rate is less than 10% (~6%, for BWA). We observed a varying fraction of unmapped reads which ranged from 10% to nearly 40% for some individuals. These results indicated that genetic variants that distinguish the individuals' genomes from the reference sequence may have not only caused reads to be misaligned but also led to a variable proportion of unmapped reads, resulting in a biased accounting of sequencing coverage across the genetically diverse populations of *H. impetiginosus*. As already noted in the literature, the inability to map reads is most likely due to structural rearrangements or insertions in the query genomes, or deletions in the reference.[46] For mapped reads, however, mappability within targeted regions was further evaluated using HCMappingQualityFilter in GATK HaplotypeCaller (-mmq 10, default value). This filter is applied as a prelude to variant calling to ensure that only reads that are likely to be informative will be used as evidence of possible variation. Overall good read mappability was achieved, with only 3–4% of the reads being discarded prior to variant calling.

Available algorithms for read mapping of divergent reads to a single reference, say up to 10–15%, have been developed such as Stampy.[47] Recently developed bioinformatics tools aim to integrate multiple genomes of the same or closely related species into a single space-efficient representation of the graph structure and align read data to this complex structure.[48–51] Unfortunately, these tools are not fully integrated in confident tools for SNP calling and genotyping such as the GATK-HaplotypeCaller framework which performs better with aligner tool less tolerable to divergent reads (up to 3–6%) such as BWA.[52]

Sequencing coverage contributes significantly to the total physical coverage of the targeted region and to an accuracy of sequence characterization of that region. At the single sample level, usually 150–200x of read depth is necessary to obtain over than 99.9% total coverage and accuracy across targets.[53] However, a depth of coverage as low as 1–2x has been found sufficient for accurate allele calls[46] to find common SNPs in a multi-sample sequencing design for a single species or closely related species. In a diploid organism, beginning with a 2x depth of coverage and dividing by the expected allele frequency to be observed (say 5%), the minimum depth necessary would be 40x. In one of our experiments, three of the 24 samples produced sequencing coverage of less than 40x possibly due to poor DNA quality or issues during library preparation.

We identified a set of targeted regions containing SNPs that consistently produced data in the sampled individuals with a capture efficiency of 60% or superior at appropriate coverage (>70x). However, we also recognized in our pipeline that some unintended regions with high homology to the target sequence had been captured. In fact, the number of quality calls in the non-targeted regions was ~35% of the total calls (134,126 out of 385,295) as assessed in the preliminary analysis of variation. This indicates that the protocol used for target enrichment and sequence capture can generate a considerable number of reads that maps outside the targeted regions, which limits the degree of sample multiplexing per NGS unit run and also represents a challenge to the read alignment process.[53] Nevertheless, while off-target read mapping does represent a challenge to SNP callers it can also be viewed as an opportunity to extend genotyping beyond the designed probe sequences.

These off-target variants can be due to many different factors depending on the capture platform, such as the quantity of input DNA, the degree of the target DNA entanglement with other strands during isolation for library preparation, and the ability of the probes to capture DNA segments that are much larger than their desired length based on the probe sequence. These unintended segments may contain repetitive sequences that can be present in many copies resulting in self-priming events during any subsequent amplification step. We speculate that the fraction of calls in off-target regions in our study was most likely caused by the relatively simplified capture and multiplexing system used by the custom multiplexed CaptureSeq method used for DNA enrichment and indexing sequencing libraries.[19] Designed to be a fast, low-cost and high throughput method, it does not include fragment size selection and the use of blocking oligos to prevent the enrichment of non-specific targets in the custom enrichment, which are steps commonly used in extended but significantly more expensive protocols. Noteworthy, our first capture assay resulted in 40% of total reads mapping on-target related to the probe sequence coordinates. A partially replicated assay resulted in 50% of total reads mapping on-target. However, with an appropriate definition of target loci, which include 200 bp in each direction from the probe sequence coordinates, the amount of on-target read relative to the total generated reads can reach 60–70%. Given the low cost protocol and a customized probe set for a significantly more heterozygous genome, it is satisfactorily competitive with more refined commercial capture probe assays in humans which report average of on-target reads within 68.4% ± 7.8% (s.d.) such as the NimbleGen SeqCap EZ (Roche, Basel, Switzerland).[54]

4.2. The challenge of designing a benchmark strategy to assesses the performance outcome of the custom probe set

We observed that the bioinformatics analysis of the sequence data is another important challenge. Caution is needed to adequately remove probe sequences improperly designed or unreliable for SNP detection. Typically, the design of overlapping sequences tiled across the entire transcript sequences was much satisfactory. Sequences spanning across exon boundaries cannot work well, but they will have neighboring probes which will still give quality data. However, any short exons (below the probe sequence length) may not be recoverable unless they can be 'padded' with intronic sequence. The computational prediction of gene models from a preliminary genome assembly provides the adequate resource to target these regions. Furthermore, because the experimental conditions of the genotyping assay cannot be directly predicted, we recommend an evaluation across replicates to assess the overall accuracy in the performance outcome of the custom probe set. In our work, with a careful bioinformatics protocol for variant analysis we could recognize 4,344 probes that were deemed faulty or improper. Additionally, we inferred from a partial replication that another 4,881 probes contributed negatively to the measures of performance of a binary classification test of success/failure to detect the true genotype of samples.

In terms of the number of distinct sampled loci, both methods used for probe design led to confident SNP calls in regions flanking the exons in transcripts indicating that the whole transcript can be targeted. Intergenic regions will produce reliable detection of true variants, but can introduce penalties in terms of missing data when the sampled individuals are very diverse. Limited read alignment to a single reference genome due to high rates and/or different patterns of sequence polymorphism between individuals can lead to biased accounting of sequencing coverage around the intergenic, UTRs and introns regions, which can result in inflation of missing data at shared sites when genotyping. Noteworthy, we observe that target capture systems can benefit of alternative approaches for probe capture design that directly use transcript sequences, instead of limiting the design to longer exons as is often the choice.

Based on these assessments, we estimated that the efficiency of our custom capture system to reveal high-confidence SNP calls for reliable genotyping in populations of *H. impetiginosus* from which the samples were drawn is 70% (21,570 successful capture probe sequences out of a total of 30, 795 designed probes). To get higher efficiencies, an interested user can simply use the recommended subset of probes in the VQSR+GQ20 set (Supplementary File S2). In this case, under the expectation of the performance evaluation, the percentage of probe regions that may be produce aligned read depth greater than 40x will be ~87.1% (21,570 probe regions in 24,751). This is within the same order of magnitude, for instance, of the performance of a sequence capture system targeting the gene space of another highly heterozygous forest trees. In *Populus trichocarpa* and *Pinus taeda*, 86.8% and 79.1% of the probe regions were on average sequenced at a depth ≥10x.[18,19]

4.3. Missing data at SNPs detected in the capture system

A common issue in studies attempting to use capture probes on divergent genomes is the amount of missing data at any shared site in a diverse sample of individuals. In plants, for instance, a probe set developed based on transcript sequences from loblolly pine (*Pinus taeda*) was used for DNA enrichment across targeted sequences in 48 individuals of whitebark pine (*Pinus albicaulis*), resulting in the identification of 528,873 SNPs in 7,849 transcripts.[55] In that study, data on genotypes at SNPs was found useful only for 12,390 segregating sites (2% out of 528,873 identified SNPs) in 4,452 transcripts (56.7% out of 7,849). In our study, missing data for any site was concentrated in particular SNPs and populations, suggesting a significant underlying population structure likely leading to an important impact on the sharing of SNP sites. In fact, the SNP data clearly recovered patterns of structure in the populations (see details in Supplementary File S1 and Fig. S1B and C). Besides structure in the data, we could also observe that the patterns of polymorphisms in the sampled individuals differ from a single reference genome at different rates across intergenic, transcribed or protein-coding regions. It is an expected outcome since population-level whole-genome resequencing has revealed that diversity levels in large outbred population—assessed by different measures of nucleotide diversity—within stretches of DNA transcribed into a mRNA molecule is reduced in comparison to intergenic DNA. A recent study in outbred bird species,[56] for instance, reported that diversity is related to selective constraint such that in comparison with intergenic DNA, diversity at fourfold degenerate sites was reduced to 85%, 3′ UTRs to 82%, 5′ UTRs to 70% and non-degenerate sites to 12%. Other studies including outbred tree species such as poplars[57] have shown that, among various genomic contexts, the levels of nucleotide diversity were highest at intergenic sites, followed by fourfold synonymous sites, 3′ UTRs, 5′ UTRs, and introns and were the lowest at 0-fold degenerate sites.

4.4. Requirements and recommendations

Given the per-sample sequencing coverage (i.e. the ratio between the number of quality reads after alignment to a genome and the number of loci) greater than 40x, capture efficiencies >60% are expected and genotype call rates at SNPs may be ≥92%. Our results suggest that the successful use of this sequence capture system would encompass the generation of >3.9 million of 100 bp single-end reads per-sample across the 18,216 targeted loci defined by location of the recommended 24,771 probe sequences in the call set 'GQ20+VQSR'. This lower-bound value was adjusted considering the observed ratio between on-target and off-target read alignments (35%, in our data). Adjustment has also been performed due to the limited read alignment to the reference assembly (60–90%, in our data), because of structural rearrangements or insertions in the query genome, or deletions in the reference assembly due to divergent evolutionary history of the population to which the sample belongs. The bioinformatics protocol for data analysis is standard for analysis of variation using Illumina's NGS technology and reference-based methods of SNP identification and involves the steps summarized in Fig. 1 and detailed in the Supplementary File S1. It includes the target loci definition, the creation of a database of variant resources and the likelihoods that these variants exist and the recalibration for all evaluated variants.

4.5. Prospects of the sequence capture system for population genetic studies

We used the genotypic data on the cataloged 210,006 SNP from the two genotyping assays using our custom probe set to look at some population features as a preamble to population genomics studies in *H. impetiginosus*. From a purely observational standpoint, due to the small number of sampled individuals, the shape of the SFS distribution (Supplementary Fig. S1A) may suggest a pattern consistent with incomplete lineage sorting and/or mutation bias among the sampled sub-groups. These preliminary results are consistent with previous phylogenetic analyses showing evidences of incomplete lineage sorting in *Handroanthus* species across populations in Central Brazil but without recovering the shared ancestral haplotypes between them.[23,58] If there is an underlying structure in the data (see Supplementary File S1 and Fig. S1B and C) and if genes did not coalesce at the same time that the sub-groups delimitation occurred in a recent past, differences in the allele frequency distribution can be observed on the entire SFS in comparison with the expected distribution under mutation-drift equilibrium. Thus, the small number of generations that have passed until sub-group delimitation may have contributed in maintaining larger than expected SNPs at intermediate to high frequency. Based on these observations, we expect that this capture system along with available tools for phase and haplotype identification should provide the necessary framework to provide more definitive evidences in support of these findings.

In conclusion, we have reported a targeted SNP resource for *Handroanthus impetiginosus*, a highly valued tree, with a notable ecological keystone status in SDTFs of South America and Mesoamerica. The successful use of this 24,751 capture probes system requires the generation of ~4 million of 100 bp single-end reads per-sample to provide robust genotyping across 18,216 distinct loci

sampling more than 10 Mbp of the species genome. This system identifies nearly 200,000 SNPs located inside or in close proximity to almost 14,000 annotated protein-coding genes, generating quality genotypic data in populations spanning wide geographic distances across the species native range. Prospects of population genetic studies indicate that this platform can be used to estimate population genetics parameters and carry on investigations of the interplay of ecology and evolution at the genome-wide scale. We emphasize though that the preliminary population-level analyses presented here with a limited sample of individuals was only meant to provide a first exploratory assessment on the evolution of genome sequences in *H. impetiginosus*, and by no means we imply that these analyses are representative but rather observational. To that end, we are currently using this capture system to characterize polymorphisms in a follow-up study with a larger set of individuals sampled in 13 demographic groups across the Brazilian territory. Going beyond the significance of these results for the species, this study paves the way for the development of similar genomic resources for other Neotropical forest trees of equivalent or higher relevance. This in turn will open exceptional prospects to advance the understanding of the evolutionary history, species distribution and population demography of the still largely neglected forest trees of the mega diverse tropical biomes.

5. Data and scripts availability

Sequences and corresponding protein-coding gene annotations for the genome assembly of *H. impetiginosus* have been deposited into GenBank, accession NKXS00000000.1. Raw sequencing reads produced in the Capture-Seq experiments HIMP-1 and HIMP-2 are available into SRA under accession numbers SRR6369264-SRR6369287 and SRR6369569-SRR6369592, respectively. BioSample descriptions and all available data are accessible from the BioProject PRJNA324125. Scripts used to process the sequence data and to generate the SNP calls and genotypes across the validated loci in the species genome are available from https://github.com/biozzyn/handroanthus-variant-analysis.

Acknowledgements

This work was supported by competitive grants from CNPq to R.G.C. (project no. 471366/2007-2 and Rede Cerrado CNPq/PPBio project no. 457406/2012-7), to E.N. (CNPq Proc. 476709/2012-1) and to D.G. (PRONEX FAP-DF Project Grant "NEXTREE" 193.000.570/2009). R.G.C. and D.G. have been supported by productivity grants from CNPq, which we gratefully acknowledge. O.B.S. has been supported by an EMBRAPA doctoral fellowship.

Accession numbers

GenBank NKXS00000000.1, SRA SRR6369264-SRR6369287, SRA SRR6369569-SRR6369592

Conflict of interest

None declared.

Supplementary data

Supplementary data are available at *DNARES* online.

References

1. Stapley, J., Reger, J., Feulner, P. G. D., et al. 2010, Adaptation genomics: the next generation, *Trends Ecol. Evol.*, **25**, 705–12.
2. Pool, J. E., Hellmann, I., Jensen, J. D. and Nielsen, R. 2010, Population genetic inference from genomic sequence variation, *Genome Res.*, **20**, 291–300.
3. Morin, P. A., Luikart, G., Wayne, R. K. and Grp, S. W. 2004, SNPs in ecology, evolution and conservation, *Trends Ecol. Evol.*, **19**, 208–16.
4. Tiffin, P. and Ross-Ibarra, J. 2014, Advances and limits of using population genetics to understand local adaptation, *Trends Ecol. Evol.*, **29**, 673–80.
5. Sachidanandam, R., Weissman, D., Schmidt, S. C., et al. 2001, A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms, *Nature*, **409**, 928–33.
6. Nordborg, M., Hu, T. T., Ishino, Y., et al. 2005, The pattern of polymorphism in Arabidopsis thaliana, *PLoS Biol.*, **3**, e196–1299.
7. Shen, Y. J., Jiang, H., Jin, J. P., et al. 2004, Development of genome-wide DNA polymorphism database for map-based cloning of rice genes, *Plant Physiol.*, **135**, 1198–205.
8. Morrell, P. L., Buckler, E. S. and Ross-Ibarra, J. 2011, Crop genomics: advances and applications, *Nat. Rev. Genet.*, **13**, 85–96.
9. Varshney, R. K., Glaszmann, J. C., Leung, H. and Ribaut, J. M. 2010, More genomic resources for less-studied crops, *Trends Biotechnol.*, **28**, 452–60.
10. Neale, D. B. and Kremer, A. 2011, Forest tree genomics: growing resources and applications, *Nat. Rev. Genet.*, **12**, 111–22.
11. Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M. and Blaxter, M. L. 2011, Genome-wide genetic marker discovery and genotyping using next-generation sequencing, *Nat. Rev. Genet.*, **12**, 499–510.
12. Mamanova, L., Coffey, A. J., Scott, C. E., et al. 2010, Target-enrichment strategies for next-generation sequencing, *Nat. Methods.*, **7**, 111–8.
13. Gasc, C., Peyretaillade, E. and Peyret, P. 2016, Sequence capture by hybridization to explore modern and ancient genomic diversity in model and nonmodel organisms, *Nucleic Acids Res.*, **44**, 4504–18.
14. Gnirke, A., Melnikov, A., Maguire, J., et al. 2009, Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing, *Nat. Biotechnol.*, **27**, 182–9.
15. Walsh, T., Lee, M. K., Casadei, S., et al. 2010, Detection of inherited mutations for breast and ovarian cancer using genomic capture and massively parallel sequencing, *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 12629–33.
16. Fu, Y., Springer, N. M., Gerhardt, D. J., et al. 2010, Repeat subtraction-mediated sequence capture from a complex genome, *Plant J.*, **62**, 898–909.
17. Dasgupta, M. G., Dharanishanthi, V., Agarwal, I. and Krutovsky, K. V. 2015, Development of genetic markers in eucalyptus species by target enrichment and exome sequencing, *PLoS One*, **10**, e0116528.
18. Zhou, L. and Holliday, J. A. 2012, Targeted enrichment of the black cottonwood (*Populus trichocarpa*) gene space using sequence capture, *BMC Genomics*, **13**, 703.
19. Neves, L. G., Davis, J. M., Barbazuk, W. B. and Kirst, M. 2013, Whole-exome targeted sequencing of the uncharacterized pine genome, *Plant J.*, **75**, 146–56.
20. Jones, M. R. and Good, J. M. 2016, Targeted capture in evolutionary and ecological genomics, *Mol. Ecol.*, **25**, 185–202.
21. Nicholls, J., Pennington, R., Koenen, E., et al. 2015, Using targeted enrichment of nuclear genes to increase phylogenetic resolution in the Neotropical rain forest genus Inga (Leguminosae: Mimosoideae), *Front. Plant Sci.*, **6**, 710.
22. Prado, D. E. and Gibbs, P. E. 1993, Patterns of species distributions in the dry seasonal forests of South America, *Ann. Missouri Bot. Gard.*, **80**, 902–27.

23. Collevatti, R. G., Terribile, L. C., Lima-Ribeiro, M. S., et al. 2012, A coupled phylogeographical and species distribution modelling approach recovers the demographical history of a Neotropical seasonally dry forest tree species, *Mol. Ecol.*, **21**, 5845–63.
24. Maddison, W. P. and Knowles, L. L. 2006, Inferring phylogeny despite incomplete lineage sorting, *Syst. Biol.*, **55**, 21–30.
25. Schulze, M., Grogan, J., Uhl, C., Lentini, M. and Vidal, E. 2008, Evaluating ipe (Tabebuia, Bignoniaceae) logging in Amazonia: sustainable management or catalyst for forest degradation? *Biol. Conserv.*, **141**, 2071–85.
26. Shafer, A. B. A., Wolf, J. B. W., Alves, P. C., et al. 2015, Genomics and the challenging translation into conservation practice, *Trends Ecol. Evol.*, **30**, 78–87.
27. Silva-Junior, O. B., Grattapaglia, D., Novaes, E. and Collevatti, R. G. 2018, Genome assembly of the Pink Ipê (*Handroanthus impetiginosus*, Bignoniaceae), a highly-valued ecologically keystone Neotropical timber forest tree and, a natural product producer. *Gigascience*, **7**, 16.
28. DePristo, M. A., Banks, E., Poplin, R., et al. 2011, A framework for variation discovery and genotyping using next-generation DNA sequencing data, *Nat. Genet.*, **43**, 491–8.
29. Iqbal, Z., Caccamo, M., Turner, I., Flicek, P. and McVean, G. 2012, De novo assembly and genotyping of variants using colored de Bruijn graphs, *Nat. Genet.*, **44**, 226–32.
30. Li, H. and Durbin, R. 2009, Fast and accurate short read alignment with Burrows-Wheeler transform, *Bioinformatics*, **25**, 1754–60.
31. ICGMC 2015, High-resolution linkage map and chromosome-scale genome assembly for Cassava (*Manihot esculenta* Crantz) from 10 populations, *G3 (Bethesda)*, **5**, 133.
32. Quinlan, A. R. and Hall, I. M. 2010, BEDTools: a flexible suite of utilities for comparing genomic features, *Bioinformatics*, **26**, 841–2.
33. McCormick, R. F., Truong, S. K. and Mullet, J. E. 2015, RIG: recalibration and Interrelation of Genomic Sequence Data with the GATK, *G3 (Bethesda)*, **5**, 655–65.
34. Cingolani, P., Platts, A., Wang, L. L., et al. 2012, A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff, *Fly*, **6**, 80–92.
35. Chang, C. C., Chow, C. C., Tellier, L. C. A. M., Vattikuti, S., Purcell, S. M. and Lee, J. J. 2015, Second-generation PLINK: rising to the challenge of larger and richer datasets, *Gigascience*, **4**, 16.
36. Watterson, G. A. 1975, On the number of segregating sites in genetical models without recombination, *Theor. Popul. Biol.*, **7**, 256–76.
37. Baird, N. A., Etter, P. D., Atwood, T. S., et al. 2008, Rapid SNP discovery and genetic mapping using sequenced RAD markers, *PLoS One*, **3**, e3376.
38. Elshire, R. J., Glaubitz, J. C., Sun, Q., et al. 2011, A robust, simple genotyping-by-sequencing (GbS) approach for high diversity species, *PLoS One*, **6**, e19379.
39. Myles, S. 2013, Improving fruit and wine: what does genomics have to offer? *Trends Genet.*, **29**, 190–6.
40. Lowry David, B., Hoban, S., Kelley Joanna, L., et al. 2016, Breaking RAD: an evaluation of the utility of restriction site-associated DNA sequencing for genome scans of adaptation, *Mol. Ecol. Resour.*, **17**, 142–52.
41. Gautier, M., Gharbi, K., Cezard, T., et al. 2013, The effect of RAD allele dropout on the estimation of genetic variation within and between populations, *Mol. Ecol.*, **22**, 3165–78.
42. Arnold, B., Corbett-Detig, R. B., Hartl, D. and Bomblies, K. 2013, RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling, *Mol. Ecol.*, **22**, 3179–90.
43. Harvey, M. G., Smith, B. T., Glenn, T. C., Faircloth, B. C. and Brumfield, R. T. 2016, Sequence capture versus restriction site associated DNA sequencing for shallow systematics, *Syst. Biol.*, **65**, 910–24.
44. Hoffberg, S. L., Kieran, T. J., Catchen, J. M., et al. 2016, RADcap: sequence capture of dual-digest RADseq libraries with identifiable duplicates and reduced missing data, *Mol. Ecol. Resour.*, **16**, 1264–78.
45. Portik, D., Smith, L. and Bi, K. 2016, An evaluation of transcriptome-based exon capture for frog phylogenomics across multiple scales of divergence (Class: Amphibia, Order: Anura), *Mol. Ecol. Resour.*, **16**, 1069–83.
46. Sims, D., Sudbery, I., Ilott, N. E., Heger, A. and Ponting, C. P. 2014, Sequencing depth and coverage: key considerations in genomic analyses, *Nat. Rev. Genet.*, **15**, 121–32.
47. Lunter, G. and Goodson, M. 2011, Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads, *Genome Res.*, **21**, 936–9.
48. Schneeberger, K., Hagmann, J., Ossowski, S., et al. 2009, Simultaneous alignment of short reads against multiple genomes, *Genome Biol.*, **10**, R98.
49. Marcus, S., Lee, H. and Schatz, M. C. 2014, SplitMEM: a graphical algorithm for pan-genome analysis with suffix skips, *Bioinformatics*, **30**, 3476–83.
50. Huang, L., Popic, V. and Batzoglou, S. 2013, Short read alignment with populations of genomes, *Bioinformatics*, **29**, i361–70.
51. Beller, T. and Ohlebusch, E. 2016, A representation of a compressed de Bruijn graph for pan-genome analysis that enables search, *Algorithms Mol. Biol.*, **11**, 20.
52. Hwang, S., Kim, E., Lee, I. and Marcotte, E. M. 2016, Systematic comparison of variant calling pipelines using gold standard personal exome variants, *Sci. Rep.*, **5**, 17875.
53. Dapprich, J., Ferriola, D., Mackiewicz, K., et al. 2016, The next generation of target capture technologies - large DNA fragment enrichment and sequencing determines regional genomic variation of high complexity, *BMC Genomics*, **17**, 486.
54. van der Werf, I. M., Kooy, R. F. and Vandeweyer, G. 2015, A robust protocol to increase NimbleGen SeqCap EZ multiplexing capacity to 96 samples, *PLoS One*, **10**, e0123872.
55. Syring, J. V., Tennessen, J. A., Jennings, T. N., Wegrzyn, J., Scelfo-Dalbey, C. and Cronn, R. 2016, Targeted capture sequencing in whitebark pine reveals range-wide demographic and adaptive patterns despite challenges of a large, repetitive genome, *Front. Plant Sci.*, **7**, 484.
56. Dutoit, L., Burri, R., Nater, A., Mugal, C. F. and Ellegren, H. 2017, Genomic distribution and estimation of nucleotide diversity in natural populations: perspectives from the collared flycatcher (*Ficedula albicollis*) genome, *Mol. Ecol. Resour.*, **17**, 586–97.
57. Wang, J., Street, N. R., Scofield, D. G. and Ingvarsson, P. K. 2016, Natural selection and recombination rate variation shape nucleotide polymorphism across the genomes of three related populus species, *Genetics*, **202**, 1185–200.
58. de Melo, W. A., Lima-Ribeiro, M. S., Terribile, L. C. and Collevatti, R. G. 2016, Coalescent simulation and paleodistribution modeling for *Tabebuia rosealba* do not support South American dry forest refugia hypothesis, *PLoS One*, **11**, e0159314.