Efficient development of highly polymorphic microsatellite markers based on polymorphic repeats in transcriptome sequences of multiple individuals

Molecular Ecology Resources

Vukosavljev, M.; Esselink, G.; Westende, W.P.C.; Cox, P.; Visser, R.G.F. et al

https://doi.org/10.1111/1755-0998.12289

MOLECULAR ECOLOGY
RESOURCES

Molecular Ecology Resources (2015) **15**, 17–27 doi: 10.1111/1755-0998.12289

Efficient development of highly polymorphic microsatellite markers based on polymorphic repeats in transcriptome sequences of multiple individuals

M. VUKOSAVLJEV,*† G. D. ESSELINK,* W. P. C. VAN 'T WESTENDE,* P. COX,‡ R. G. F. VISSER,* P. ARENS* and M. J. M. SMULDERS*

**Wageningen UR Plant Breeding, Wageningen University & Research Centre, P.O. Box 386, NL-6700AJ, Wageningen, the Netherlands, †C.T. de Wit Graduate School for Production Ecology and Resource Conservation (PE&RC), Wageningen, the Netherlands, ‡Roath BV, Eindhoven, the Netherlands*

Abstract

The first hurdle in developing microsatellite markers, cloning, has been overcome by next-generation sequencing. The second hurdle is testing to differentiate polymorphic from nonpolymorphic loci. The third hurdle, somewhat hidden, is that only polymorphic markers with a large effective number of alleles are sufficiently informative to be deployed in multiple studies. Both steps are laborious and still performed manually. We have developed a strategy in which we first screen reads from multiple genotypes for repeats that show the most length variants, and only these are subsequently developed into markers. We validated our strategy in tetraploid garden rose using Illumina paired-end transcriptome sequences of 11 roses. Of 48 tested two markers failed to amplify, but all others were polymorphic. Ten loci amplified more than one locus, indicating duplicated genes or gene families. Completely avoiding duplicated loci will be difficult because the range of numbers of predicted alleles of highly polymorphic single- and multilocus markers largely overlapped. Of the remainder, half were replicate markers (i.e. multiple primer pairs for one locus), indicating the difficulty of correctly filtering short reads containing repeat sequences. We subsequently refined the approach to eliminate multiple primer sets to the same loci. The remaining 18 markers were all highly polymorphic, amplifying on average 11.7 alleles per marker (range = 6–20) in 11 tetraploid roses, exceeding the 8.2 alleles per marker of the 24 most polymorphic markers genotyped previously. This strategy therefore represents a major step forward in the development of highly polymorphic microsatellite markers.

Keywords: microsatellite marker, next-generation sequencing, RNA-seq, simple sequence repeat

Received 18 February 2014; revision received 29 May 2014; accepted 30 May 2014

Introduction

Thanks to their reproducibility, codominant inheritance and abundance, microsatellite (also known as simple sequence repeat – SSR) markers are suitable molecular tools for many applications in genetic analysis and breeding. Additionally, being multi-allelic, they are powerful for parentage analysis and haplotyping, particularly for mapping in polyploids as they allow detecting multiple alleles at the same locus on all homologous chromosomes (Vukosavljev *et al.* 2012). Despite the advent of SNP markers, recent studies in various plant and animal genera, for instance, Cucurbita (Barzegar *et al.* 2013), Euphydryas (Smee *et al.* 2013), Lilium (Yuan *et al.* 2013), Medicago (Zitouna *et al.* 2013), Pinus (Iwaizumi *et al.* 2013), Portunus (Guo *et al.* 2013), Scatophagus (Liu *et al.* 2013), Triticum (Ansari *et al.* 2013) and Vitis (Doulati-Baneh *et al.* 2013) indicate that microsatellite markers are still extensively being developed as a molecular tool for various purposes.

Conventional microsatellite development is a long and costly process. First, many microsatellite repeats need to be sequenced. Second, often as many as 50–100 primer pairs have to be tested to develop 10 polymorphic markers. Third, for many of these polymorphic markers, only few alleles with length differences in the repeat exist in the germplasm. The flanking regions of microsatellite repeats may contain additional SNPs (Xing *et al.* 2005; Zhang *et al.* 2013), but to this day, these cannot be detected routinely with sufficient precision. Practical usage shows that the best microsatellite markers are multi-allelic and have a high effective number of alleles (Ne) in the germplasm. However, only a small portion of

Correspondence: Marinus Smulders, Fax: +31 317 418094; E-mail: rene.smulders@wur.nl

all polymorphic markers published have many alleles and will be widely used.

The development of highly polymorphic microsatellite markers using transcriptomic sequences is an interesting alternative that requires less effort, as sequences are already available or can be generated easily using next-generation sequencing (Jennings *et al.* 2011; Triwitayakorn *et al.* 2011; Wang *et al.* 2012; Wang *et al.* 2013; Nybom *et al.* 2014), and microsatellite repeats can be identified by custom or freely available bioinformatics pipelines, such as PolySSR (Tang *et al.* 2008) and Pal_Finder (Castoe *et al.* 2012). Indeed, recently, several studies reported microsatellite marker development based on expressed sequences from sources such as GenBank or Genome database for Rosaceae (e.g. Durand *et al.* 2010; Park *et al.* 2010; Duran *et al.* 2013) or from custom-made transcriptome sequence libraries (e.g. Blair & Hurtado 2013). However, from the identification step onwards, the process is still slow, as most researchers select random subsets of repeats as a start for marker development (e.g. Liu *et al.* 2013). Legendre *et al.* (2007) developed a model, 'SERV', to predict the potential variability of repeats based on number of repeated units, unit length and purity, which would allow to preselect more promising repeat loci. Tang *et al.* (2008) developed a pipeline to preselect repeat loci for which sequence reads show polymorphism in repeat length between a few genotypes, to exclude monomorphic repeat loci from the marker-testing step.

Although finding many microsatellite repeats makes it possible to test more markers until a set of high-quality markers has been established, it does not speed up the testing process for multi-allelic markers. As one of few new developments for the latter problem, Eschbach and Schöning (2013) screened existing microsatellite markers for within-population polymorphism by scoring differences in sequence reads from a pooled sample of genotypes of the population they studied. Duran *et al.* (2013) developed a pipeline to extract putatively polymorphic microsatellites from EST data generated by Sanger sequencing and present in GenBank. They saw a relationship between the number of different repeats found in the ESTs and the number of different alleles amplified.

To improve the efficiency of developing multi-allelic microsatellites, we have developed a new strategy for these three steps. We first generate transcriptome sequences from multiple genotypes, then screen sequence reads from these genotypes for those repeats that show the most variation in length and move only these to the testing step. This strategy leads to highly polymorphic markers only. We demonstrate the suitability of this approach by developing highly polymorphic markers for garden roses. Garden roses are tetraploids, and for such a situation, microsatellite markers are very appropriate molecular markers. To ensure that the selected markers will have a large effective number of alleles across the garden rose germplasm, we based our marker development on transcriptome sequences from a set of 11 garden roses representing different garden rose cultivar groups (Vukosavljev *et al.* 2013).

Materials and methods

Plant material and RNA extraction

For this study, we used a set of 11 tetraploid garden rose cultivars (Table 1), which were bred by different breeders, and belong to different types (Vukosavljev *et al.* 2013) with a large amount of phenotypic variation (e.g. difference in flower colour, fragrance, number of petals, winter hardiness, growth type, presence/absence of recurrent blooming). From each cultivar flowers in three stages (closed buds, half-way open and fully open flowers), young leaves were collected for RNA isolation. Tissues were frozen using liquid nitrogen. Frozen flower material was ground with an IKA mill. Leaf tissue was ground in a mortar. After grinding, powder of leaf and flowers was pooled in equal amounts. RNA was extracted according to the protocol of Cheng *et al.* (1993). Briefly, 1–1.5 g of frozen material was added to a preheated (65 °C) CTAB extraction buffer and mixed thoroughly. After two extractions with chloroform, the RNA was precipitated overnight using LiCl. Next, the pellet was dissolved and the RNA purified further by chloroform extraction and EtOH precipitation. RNA integrity, yield and quality were measured on agarose gel and with NanoDrop (Thermo Scientific).

Microsatellite marker prediction

After RNA extraction, cDNA library preparation and Illumina HiSeq sequencing were performed according to manufacturer specifications (Illumina, San Diego, CA, USA) at GATC Biotech (Konstanz, Germany). For each cultivar, around 40 million 100-bp paired-end (PE) reads were obtained (trimmed read lengths 88.9 ± 7.1 (SD) bp to 89.9 ± 4.5 bp, average 89.3 bp), of which after quality checking between 12.1 million and 16.5 million were analysed for marker selection and development (Table S1, Supporting information).

Microsatellite repeats were detected by Pal_Finder v0.02.04 (http://sourceforge.net/projects/palfinder) in the raw reads, using a minimum repeat number of 4 for tri- and tetranucleotide repeats and 3 for penta- and hexanucleotide repeats. Merging of the reads was not necessary, but quality trimming did improve the speed of the process. Detected repeats were mostly located in one of the read pairs, but as they run until the end of the read,

Table 1 Garden rose varieties used

Cultivar	Type*	Breeder	Ploidy	Flower colour	Winter hardiness zone†	Growth type	Fragrance	Number of petals	Blooming
Morden Centennial	CP	Marshall	4n	Pink	3b	Shrubby	Mild	40–45	Recurrent
Red New Dawn	Cl	Robichon	4n	Pink	6b	Rambling climber	Strong	17–25	Prolific, occasionally repeat blooming
Nipper	MIN	Harkness	4n	Red	6b	Ground cover	Strong		Occasionally repeat blooming
Diamond Border	S	Olesen	4n	White	4b	Shrubby	Mild to none	17–25	Recurrent
Princess of Wales	F	Austin	4n	White	6b		Mild to strong	17–25	Recurrent
Graham Thomas	MOE	Austin	4n	Yellow	5b	Shrub	Strong	35	Recurrent
J.P. Connell	CE	Svejda	4n	White	2b	Shrub	Strong	50	Occasionally repeat blooming
City of London	F	Harkness	4n	Light Pink	6b	Shrub	Strong	15–25	Recurrent
Henry Kelsey	CE	Svejda	4n	Pink	2b	Climber	Spicy scent	5–30	Occasionally repeat blooming
Heritage	MOE	Austin	4n	Light pink	5b	Shrub	Strong	40	Recurrent
Adelaide Hoodless	CP	Marshall	3n‡	Pink	2b	Shrub	Mild	5–30	Recurrent

*CP, Canadian Parkland series; CE, Canadian Explorer series; Cl, Climber rose; MIN, Miniature rose; S, Shrub; F, Floribunda; MOE, Modern English rose.

†Winter hardiness zone; http://planthardiness.ars.usda.gov (accessed 18 July 2013).

‡According to literature, Adelaide Hoodless is a triploid rose, but our flow cytometer result indicates tetraploidy (aneuploidy is still possible).

the exact length is not known. Primers were designed for tri-, tetra-, penta- and hexanucleotide repeats by Primer3 (Rozen & Skaletsky 2000). Dinucleotide repeats were not taken into consideration.

Potential microsatellite markers ('Potentially Amplifiable Loci' or PAL) were thus developed for each cultivar separately, and the results were ordered (in Excel) by number of different alleles across genotypes, in decreasing order. For the top 100, those markers were excluded that had more than four different length variants per individual tetraploid cultivar. A set of 48 potential markers with ten or more predicted alleles were picked from the top of the list (predicted number of alleles among the 11 cultivars: 24 to 16).

For transcriptome assembly, high-quality reads were filtered using PRINSEQ (Schmieder & Edwards 2011). The paired-end reads were merged using FLASH (Fast Length Adjustment of Short Reads to Improve Genome Assemblies; http://www.cbcb.umd.edu/software/flash), producing a read span of 144.6 ± 37.6 bp to 162.2 ± 53.4 bp, average 152.4 bp. Assembly was performed using Trinity (Grabherr *et al.* 2011). The potential markers were screened for duplicates by BLASTn of the primers against the transcriptome of one of the genotypes, cultivar Red New Dawn, as well as against the genome sequence of *Fragaria vesca*. The screening against Red New Dawn identified both duplicate markers that shared forward or reverse primers as well as duplicate markers for which the primer sequences did not overlap.

Validation

Forty-eight potentially highly polymorphic microsatellite markers were tested by genotyping the 11 cultivars (Table 2). Amplification reactions were performed in 10 μL containing 8 ng DNA, 5 μL multiplex kit (QIAGEN, Germany) and 4 pmol of each forward (labelled) and reverse primer. Amplification was under the following condition: an initial denaturation at 95 °C for 15 min following with 30 cycles of 94 °C for 30 s, ramp 1 °C/s to 50 °C, 50 °C for 30 s, ramp 1 °C/s to 72 °C, 72 °C for 120 s and a final extension at 72 °C for 10 min. One μl of 100x diluted PCR product was mixed with Hi-Di formamide (Applied Biosystems) containing GeneScan-500 LIZ size standard (Applied Biosystems) and run on an ABI 3730 DNA analyser. Output from the ABI platform was analysed with GENEMAPPER 4.0

Table 2 Characteristics of the microsatellite markers

Single-locus micro-satellite marker	Forward primer (5′–3′)	Reverse primer (5′–3′)	Expected number of alleles†	Detected number of alleles	Effective number of alleles‡	Allele length range (bp)	Median allele length (bp)	Quality§	Blast match to the strawberry genome¶		Repeat motif(s)	European Nucleotide Archive accession nos
									Protein	e-value		
WGR04	ATTCTTAAAATGCATCCC	AGTCGCAAGTACTACCATCC	24	12	7.3	150–199	169	1	—		TGT	HG934830
WGR05	CAAACATGCACCTACAAGG	CAAGAAGAAGGAGGACGG	24	10	4.6	152–197	183	2	Uncharacterized LOC101309484 (79%)	0.002	GCT	HG934831
WGR07	TTTTAGAGAAGAAGCAGAGAGC	CTCTCGGTTTCAGAGAAGC	24	6	3.0	173–278	187.5	1	Auxin-induced protein AUX28-like, AUX_IAA superfamily (96%)	1e-119	GGA	HG934832
WGR11	GAGAAAGTTCAAACTCGCC	TCTTGGTGATGTTCTTCTCG	18	11	8.5	260–284	274	2	RAMP4 superfamily	Isoform 1 2e-36 Isoform 2 3e-36	CAA	HG934834
WGR12	AAAACCCAGAAGACGTTAGC	AGTTTAAATGCTTGCTTTCG	17	11	8.0	161–206	188	1	Cytochrome c oxidase subunit 5b-2,mitochondrial-like isoform, Cyt_c_Oxidase_V5 superfamily (95%)	Isoform 1 2e-87 Isoform 2 4e-75	TAA, GAA	HG934835
WGR17	TCATCATCAGCAACAGCC	TTGTTTGAGCTACGTTTGG	16	13	7.9	130–215	186	1	Transcription factor TCP20-like (90%)	7e-131	AAC, AAC	HG934836
WGR18	TTCTTCATCCTCTGCATCC	GGAGATGAGGACTTCTCAGG	16	9	4.4	166–220	194	1	Uncharacterized LOC101301324 (86%)	5e-27	CAT	HG934837
WGR20	ATTGGTTTAGCTGAGAGACG	TTAATTAATCCAAACATGGC	15	18	12	188–263	213	1	Plastid-lipid-associated protein, chloroplast-like (89%)	7e-173	GAA	HG934838
WGR22	CCTTACAAGCCTCCTACTCC	AACCCCTTTCTTTATTTTGG	14	7	3.5	225–277	273	1			GAA	HG934840
WGR28	TCCTCAAAGTGAGAAGAAGG	ATATTTGTTTAGCTCGCACG	13	12	5.8	146–180	162	1	Uncharacterized LOC101299879 (92%)	0	CAA	HG934843
WGR31	CACTCTTTCTCCTCAACCG	TAATCCTTGCCCTTCTTGG	12	20	17.3	168–260	239	1			CAA	HG934844
WGR32	GTAAACCATTCCGTGTTCC	CCTCATTACCTTCTTCATCG	12	19	13.7	170–222	204	2	Uncharacterized C119.09c-like, ORMDL superfamily (99%)	6e-109	CCT	HG934845
WGR34	TCGAACTTCTCCTTATCCG	CTTCCGGTACTCTCGTCG	12	11	8.3	265–305	279	1	—		GGA	HG934846
WGR37	AGCTGAGGATGATGGTGG	CAACAGCAACAATTTATACACC	11	14	8.6	186–243	210	1	G-type lectin S-receptor-like serine/threonine-protein kinase CES 101-like (38%)		TGC	HG934848
WGR39	CACCATCACGTACCAACC	CTACCTAGCTCATCCTGTGC	11	11	5.7	154–196	180	1	Uncharacterized LOC101298395 (91%)	3e-28	ACA	HG934849

Table 2 (Continued)

Single-locus microsatellite marker	Forward primer (5′–3′)	Reverse primer (5′–3′)	Expected number of alleles†	Detected number of alleles	Effective number of alleles‡	Allele length range (bp)	Median allele length (bp)	Quality§	Blast match to the strawberry genome¶		Repeat motif(s)	European Nucleotide Archive accession nos
									Protein	e-value		
WGR42	CTCAAACATTCTTGAATTTCC	CTGGTTAATAGTCGCATTCC	11	9	5.1	173–200	194	1	Uncharacterized LOC101311909 (81%)	1e-168	ACT	HG934850
WGR44	GAAAACCAGAGAGAGAGAAGC	AAGTGGAGACTTCAACAACC	11	8	5.4	117–295	244	1	PLN03107		GGA	LK392375
WGR47	AAAAGAACCAGAGACACACC	TTATGGATCATGTCTAGGGG	10	10	6.0	243–261	252.5	1	—		AAG	HG934851

†Estimated based on the sequence reads, using a minimum of 3 identical reads per putative allele.

‡Ne (effective number of alleles) was estimated as the reciprocal of Σpi2, where pi is the frequency of the allele in the 11 varieties examined. In tetraploids (with dominant scoring), this is an approximation.

§Quality 1, no or weak stutter peaks, well scorable; 2, stutter peaks present, but product still scorable (Smulders *et al.* 1997).

¶The predicted protein sequence derived from garden rose cDNA was used for a BLASTx search in the Fragaria vesca (diploid strawberry) strawberry.

software (Applied Biosystems). For each microsatellite marker, presence or absence of individual alleles was scored (dominant scoring).

Multigene markers

A high level of polymorphism may also be associated with multilocus microsatellites, and thus, we tested whether an additional step of checking could be implemented. For this, we used the predicted protein sequence derived from the cDNA sequence to search protein databases for the likelihood of dealing with a member of a multigene protein family by BLASTx (http://blast.ncbi.nlm.nih.gov/Blast.cgi) against the closely related strawberry genome.

Results

Microsatellite repeat and motif overview

Microsatellites with tri-, tetra-, penta- and hexanucleotide repeats were identified among the sequences for each cultivar separately. Dinucleotide repeats were not analysed. The total number of reads with microsatellite repeats per cultivar varied from 259 749 in ‘Adelaide Hoodless’ to 341 719 in ‘Princess of Wales’ (Table S1). All cultivars showed the same trend in motif frequency distributions; trinucleotide repeats were most abundant (65.1–69.3%), followed by tetranucleotides (16.3–20.5%) and hexanucleotides (9.3–11.6%). Pentanucleotide repeats were the least frequent motif type (4.8–5.4%).

Among the trinucleotide repeats (Fig. 1), TTC/GAA was the most abundant motif (30.9%), followed by TCC/GGA (14.1%) and ATC/GAT (13.3%). Among tetranucleotide repeats, CTTT/AAAG was the most common motif (21.6%); among the pentanucleotides, it was CTTTT/AAAAG (13.3%); and among hexanucleotides, it was TTCCTC/GAGGAA (7.2%).

Microsatellite marker prediction and primer development

With Primer3 we designed primers around each potentially amplifiable microsatellite repeat in each of the sequence reads. As our aim was to develop polymorphic markers, we sorted the read data based on the forward primer of the potential microsatellite marker and selected primer pairs that corresponded to reads with multiple repeat length variants in each of the eleven cultivars, but not more than four different alleles per tetraploid cultivar. This ordering was a technically simple solution for the problem of identifying multiple alleles of the same locus among paired-end reads in which a relatively large proportion of the sequence information is taken up

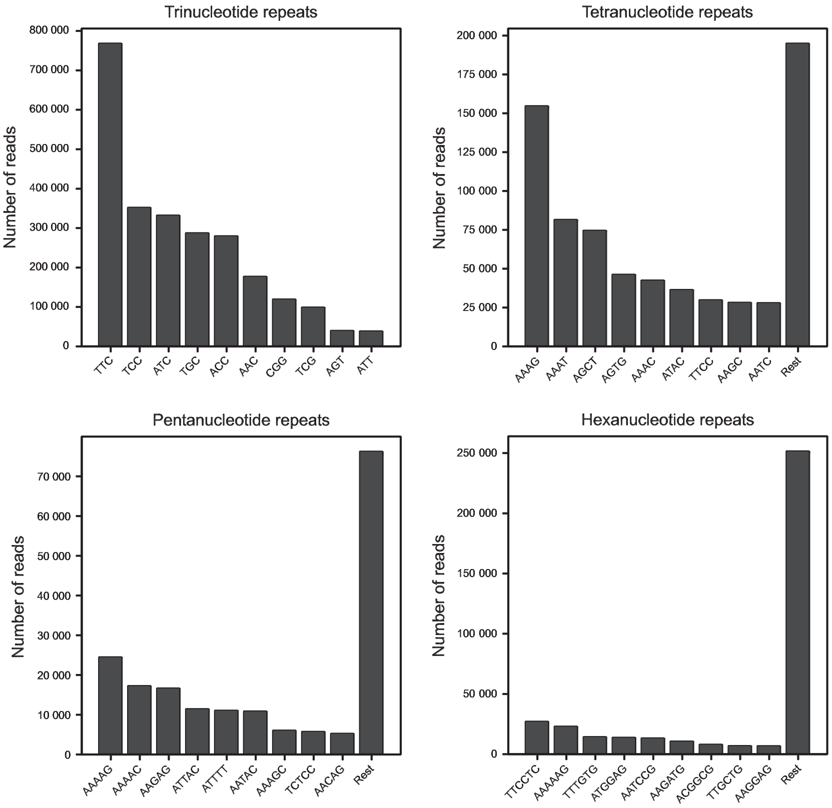

Fig. 1 Number of reads for different tri-, tetra-, penta- and hexanucleotide repeats motifs found in 11 garden rose cultivars.

by simple sequence repeats (but with the risk of not combining all reads of one locus together, see below). Of a total of 1797 developed markers, 48 trinucleotide repeat microsatellite markers were taken from the top of the list.

Polymorphism testing for validation

The selected microsatellite markers were amplified in the 11 cultivars. Two did not give amplification. All other markers were polymorphic, and allele presence/absence was scored and compared with the predicted number of alleles. In 10 markers more than four alleles per cultivar were amplified. A careful analysis of the electropherograms of these multilocus microsatellites showed the occurrence of multiple allele patterns (with and without stutter bands), amplification success (strong and weak amplification) and/or differences in allele length [two groups of alleles that differed one or two repeat units within the group but 20–40 bp between groups, which in theory could be used as a tool for assigning alleles to different loci (not shown)].

Thirty-six markers were putative single-locus markers, showing four or less clearly distinguished alleles per genotype. Analysis of their electropherograms did not detect any difference in amplification rate, stutter band pattern, nor shifts in allele lengths, which is consistent with a single-locus marker. They were all polymorphic, but upon close scrutiny, ten of them were replicate markers that shared some of the primer sequences, and an additional eight were from replicate loci but did not share any primer sequence (see below). Hence, the net result was a set of 18 unique microsatellite markers, all highly polymorphic as they amplified between 6 and 20 different alleles each in the 11 tetraploid varieties (on average 11.7 different alleles per marker; Table 2). WGR44 has a large allele size range (between 117 and 295 bp). The effective number of alleles in a large set of varieties is unknown, but an approximation, by calculating it for these 11 varieties, gives values from 2 to 17.3.

For evaluation we compared the level of polymorphism with a set of 143 microsatellites developed from genomic and EST sequences in rose (Rajapakse *et al.* 2001; Esselink *et al.* 2003; Yan *et al.* 2005; Kimura *et al.* 2006; Zhang *et al.* 2006; Hibrand Saint Oyant *et al.* 2008; Meng *et al.* 2009; Spiller *et al.* 2010) that were tested on the same set of 11 cultivars. All 143 markers have previously been successfully tested in various rose species and cultivars. After removing microsatellites that did not amplify in our set (10), had low amplification (1), showed no polymorphism (2) and multilocus ones (23), the 107 polymorphic markers amplified on average 5.1 alleles per marker. The 24 most polymorphic markers of this set of 107 markers (16.8%) were used in the diversity study of Vukosavljev *et al.* (2013). These amplified on average 8.2 alleles/marker in the 11 cultivars. This comparison shows that our new set of highly polymorphic microsatellites have more alleles per marker.

Possible improvements to the strategy

We initially screened for duplicate markers by comparing the primer sequences of the selected markers in the list. This procedure, which should take into account reverse complement and slightly shifted primer sequences, can be performed in Excel, but it is not fully conclusive as duplicate markers may have completely different sets of primer sequences. We found that the most straightforward and conclusive screening for replicate markers was to BLASTx the primer sequences against an assembly of the transcriptome of one of the genotypes. Replicate markers were identified by a hit to the same contig. In our test set of PALs with many alleles, 25 of 48 markers were replicates, of which 8 replicate loci that had no primer sequence in common. In comparison, a BLASTn search against the related genome sequence of *Fragaria vesca* was much less effective. It only discovered eight of the 25 replicates, and the others did not have primer sequence matches.

We tested whether we could have predicted which marker is multilocus based on the number of sequence length variants observed. The prediction of the number of alleles per marker based on observed sequenced length variants was imprecise (Table 2). At a cut-off of three or more reads per length variant to predict an allele, the single-locus markers had 10–24 predicted alleles in the 11 cultivars, while 6–20 were amplified. The multilocus markers were predicted to have 11–25 alleles, while 11–27 were amplified. Although the average number of amplified alleles of the single-locus markers (11.6) was much lower than the average of the multilocus markers across these cultivars (19), the overlap in the range was so large that a prediction of multilocus markers based on overall number of length variants did not work. The same was the case when we used the number of length variants per cultivar. Of the eight markers with four or fewer length variants in every cultivar, five were multilocus and only three were single-locus markers. Only one marker (WGR28) passed the more stringent threshold for a single-locus marker of maximally three predicted alleles in every cultivar. Thus, on basis of the predictive number of alleles, no effective distinction can be made between single and multilocus markers.

We also tested whether we could have distinguished single- from multilocus microsatellites based on the type of genes in which they resided, using BLASTx against the related *Fragaria vesca* genome sequence. Some of the multilocus markers indeed had hits with members of a superfamily or stress-associated proteins. For example, one of the markers that turned out to be multilocus based on the banding patters had hits with the R3H-associated superfamily. Additionally, another marker had highly significant hits with two different isoforms of the same protein (stress-associated endoplasmic reticulum protein 2-like isoform-1 and -2). However, as only 14 (30%) of the repeat-containing contigs we tested had a hit with known genes, this selection criterion may not be very effective.

Discussion

An efficient strategy for polymorphic marker development

The main problem for developing microsatellite markers nowadays is not generating repeat-containing sequences, as next-generation sequencing generates more repeat-containing sequences than needed, but it is the testing and selecting of those that are highly polymorphic as a marker, as this is still performed manually. We have developed an efficient strategy in which we deploy next-generation sequencing of multiple genotypes and select only those repeat loci for marker development that already show a range of different repeat lengths within the set of sequence reads. This selection does not predict the actual number of alleles precisely, but it proved to be very efficient for preselecting highly polymorphic markers (at least six and up to 24 alleles in 11 tetraploid garden rose cultivars).

The strategy makes efficient use of the strength of next-generation sequencing, namely that sequencing is cheap and that sequencing multiple genotypes does not require a lot more manual activities. Thus, we save on labour-intensive screening activities by generating sequences from multiple genotypes. For marker development, many studies use next-generation sequencing of multiple genotypes for SNP retrieval. Although many recent studies have been published on microsatellite marker development in which such sequences are mined (e.g. Cardoso *et al.* 2013; Lance *et al.* 2013), most studies do not make use of the full potential of the sequencing data in combination with multiple genotypes to predict the most polymorphic microsatellite markers. To our knowledge, only the recent study by Hoffman and Nichols (2011) utilized a similar approach to our study to identify polymorphic microsatellite markers from 454 sequences of the Antarctic fur seal (*Arctocephalus gazella*). Their approach rendered promising results (21 polymorphic markers from 50 tested) and had some success in predicting the number of alleles amplified from those found in the reads.

Prediction of allele number and comparison with SNP discovery

The prediction of the number of alleles based on variations in repeat length among our Illumina sequence paired-end reads was very imprecise, as both too many (e.g. WGR04, WGR05 and WGR11) and too few alleles (e.g. WGR31, WGR32) were predicted for some markers. Too many apparent alleles can be the result of mistakes made by the DNA polymerase during PCR amplification prior to next-generation sequencing. The frequency depends partly on the repeat type, length and whether the repeat is perfect or imperfect. This type of mistake is also visible as the relative number and height of stutter bands during detection on an acrylamide gel. One stutter band was present for WGR04 and WGR11, but not for other markers for which too many alleles were predicted (e.g. WGR11). With regard to predicting too few alleles, two possible reasons can be envisaged. First, only the minimum length of the repeats was known, as the repeats extended up to the end of one of the reads obtained in paired-end sequencing. Only sequencing technologies that produce longer reads can solve this problem. Second, our bioinformatics approach was simple and straightforward, but often did not collect all reads of one locus into one contig, as exemplified by the number of replicate markers. Here, again longer reads would make it easier to optimize this step. Prediction of the number of alleles based on paired-end short reads is not an easy task. Cao *et al.* (2014) developed a Bayesian method, STRViper, to predict repeat length variation. Using data from Arabidopsis strains, it outperformed all other methods.

Our results indicate that, even though the prediction of exact allele number was imprecise, the strategy for finding a set of polymorphic markers was very efficient, as all unique markers produced here are highly polymorphic (six alleles or more). A random subset of studies using traditional microsatellite marker development in polyploid species produced between 0% and 34% highly polymorphic markers (Table S2, Supporting information) irrespective of the use of NGS. This indicates that it is efficient to sequence more genotypes at lower depth and select those repeats with a large number of predicted alleles for further marker development.

It is interesting to note that the imprecision in allele calling based on Illumina reads appears to be a smaller problem for selecting microsatellite repeats than it is for calling SNPs, where wrong calling usually means that it is a false SNP, and great care has to be taken to avoid them, for example by focussing on identifying reliable haplotypes (Tang *et al.* 2006; Shahin *et al.* 2012; Nijveen *et al.* 2013). Nevertheless, some mistakes are better avoided for both types of markers: polymorphisms between paralogs in gene families, and (in polyploids) polymorphisms between subgenomes. Taking all this into account is possible, as, for example, implemented in the IStraw90 90k Axiom array for strawberry, which excludes all SNPs between the four subgenomes of octaploid strawberry (Bassil *et al.* 2014), but this is time-consuming.

Replicated markers

The single most important screening step in our strategy is identifying replicate markers. More than half of our potential markers with many alleles were replicates. Apparently the sequence information in the short paired-end reads was insufficient to always link the markers of the same locus. Identifying the replicates worked best by BLASTx to a custom-assembled transcriptome. It even enabled identifying 8 replicate markers (32% of the duplicates) that shared no primer information. It was about three times as efficient as a BLASTx to the genome sequence of the related species *Fragaria vesca*, which did not even identify all replicates with overlapping primers, that is, it was not better than careful manual screening of primers and reads that have the same repeat (provided one screens all variants in forward and in reverse complement directions). In our strategy (Fig. 2), we have included the transcriptome assembly therefore as an option to improve replicate detection. If laboratories have no possibility to do it, manual screening of replicates will do, as long it is accepted that some replicate markers will end up being tested before being identified from similar genotype patterns.

Degree of polymorphism for repeats in coding regions

The rate of successful microsatellite amplification (46 of 48; 96%) in our study is higher compared with studies in tetraploid rose that were based on genomic DNA repeats, that is, mostly located in noncoding DNA [Esselink *et al.* 2003 (89%); Kimura *et al.* 2006 (85%); Park *et al.* 2010 (92%)] or in other tetraploid species, such as cotton (86%; Han *et al.* 2004) and peanut (87%; Liang *et al.* 2009). The high level of successful PCR amplification of microsatellites from transcriptome sequences is attributed to their nature: their primers are developed from gene sequences (Saha *et al.* 2006).

It has been suggested that repeats in coding regions would be less polymorphic than those from random genomic sequences (Dufresnes *et al.* 2014). It should be noted that such a difference in degree of polymorphism only holds for a random set of repeats. As our strategy was aimed at producing a subset of highly polymorphic markers, one would not expect them to be substantially less polymorphic than a set of highly polymorphic nuclear DNA-based microsatellite markers. Indeed, the

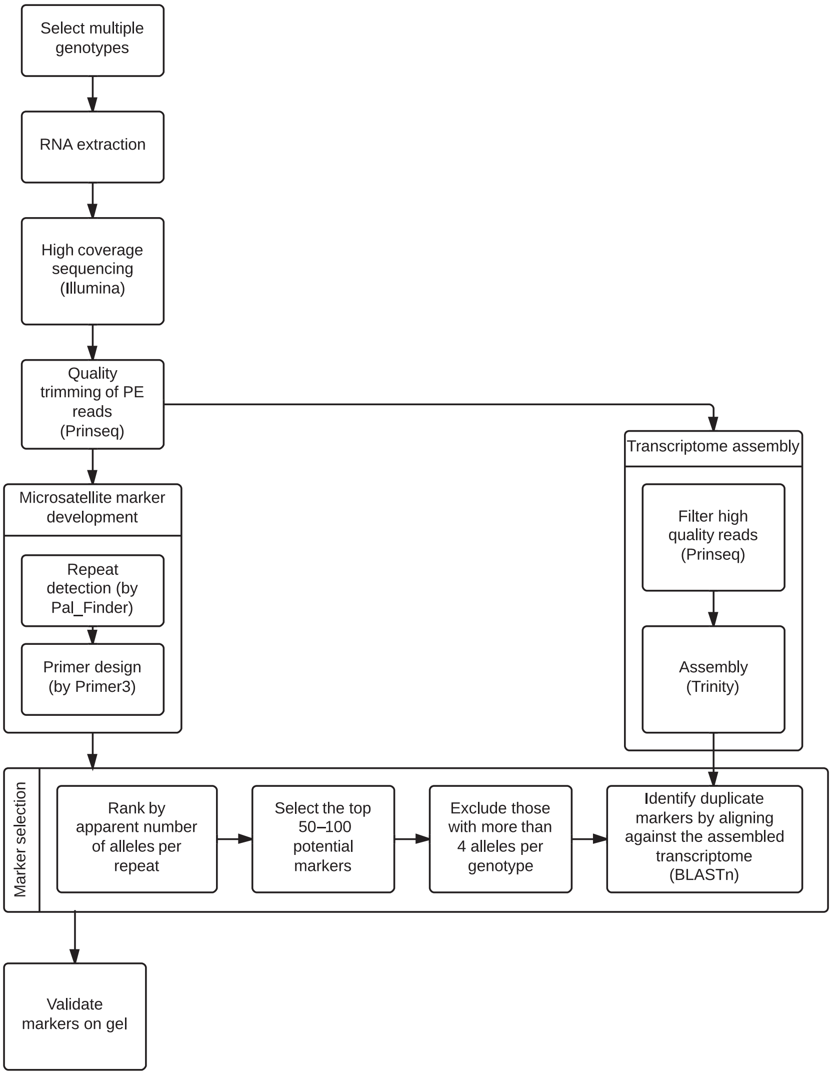
Select multiple genotypes
RNA extraction
High coverage sequencing (Illumina)
Quality trimming of PE reads (Prinseq)
Microsatellite marker development
Repeat detection (by Pal_Finder)
Primer design (by Primer3)
Transcriptome assembly
Filter high quality reads (Prinseq)
Assembly (Trinity)
Marker selection
Rank by apparent number of alleles per repeat
Select the top 50–100 potential markers
Exclude those with more than 4 alleles per genotype
Identify duplicate markers by aligning against the assembled transcriptome (BLASTn)
Validate markers on gel

Fig. 2 The strategy followed to efficiently develop highly polymorphic microsatellite markers.

24 most polymorphic markers selected from the range of publications on genomic DNA microsatellite markers in rose, as used by Vukosavljev *et al*. (2013), amplified on average 8.2 alleles/marker in these 11 cultivars, compared with 11.7 alleles for our set of gene-based markers. As the latter are located in genes and hence their flanking sequences are conserved, such markers are transferrable to related species and therefore form the marker of choice for comparative mapping, and also to tag functional and positional candidate genes to study their colocation with quantitative trait loci (QTL) (Durand *et al.* 2010).

Multilocus markers

In the set of 48 selected microsatellites, 10 amplified more than 1 locus. The presence of multilocus microsatellites in this study may be attributed to the fact that microsatellites have been chosen on the basis of a maximum number of alleles. We have not tested our strategy on genomic DNA sequences. It may be feasible to use our strategy on genomic DNA in species with small genome size, or with the use of appropriate complexity reduction methods, as are also used for SNP development (Smulders *et al.* 2012). Note, however, that the degree of amplification of duplicated repeat loci in noncoding sequences is much higher than that of genes families in our RNA-seq approach, and such highly repetitive loci must be excluded. PAL_Finder, which was designed for identifying microsatellites in genomic DNA, counts the occurrence of primer pairs to be able to select against such repeat families (Castoe *et al.* 2012). We did not employ this counter here, but it may be used in a variant of our strategy.

Conclusion

Highly polymorphic markers can be developed very efficiently by screening transcriptome sequences from multiple genotypes. Such sequence data can be generated on purpose, but often they may be produced for SNP development and highly polymorphic microsatellites can be identified as additional markers. Few studies have used the polymorphism in reads, and we are not aware of any that used RNA-seq reads of multiple genotypes. The microsatellite length data obtained from Illumina paired-end reads are imperfect, but contain sufficient information to make microsatellite development more efficient, notably to develop highly polymorphic microsatellite markers. This strategy can also be used to select markers for specific parental combinations.

Acknowledgements

This research was partly supported by the TTI Green Genetics project 'Hyperrose' and by the TKI Polyploids project (BO-26.03-002-001).

References

Ansari MJ, Al-Ghamdi A, Kumar R *et al.* (2013) Characterization and gene mapping of a chlorophyll-deficient mutant clm1 of *Triticum monococcum* L. *Biologia Plantarum*, **57**, 442–448.

Ashkenazi V, Chani E, Lavi U, Levy D, Hillel J, Veilleux RE (2001) Development of microsatellite markers in potato and their use in phylogenetic and fingerprinting analyses. *Genome*, **44**, 50–62.

Baruah A, Naik V, Hendre S, Rajkumar R, Rajendrakumar P, Aggarwal RK (2003) Isolation and characterization of nine microsatellite markers from *Coffea arabica* L., showing wide cross-species amplifications. *Molecular Ecology Notes*, **3**, 647–650.

Barzegar R, Peyvast G, Ahadi AM, Rabiei B, Ebadi AA, Babagolzadeh A (2013) Biochemical systematic, population structure and genetic variability studies among Iranian Cucurbita (*Cucurbita pepo* L.) accessions, using genomic microsatellites and implications for their breeding potential. *Biochemical Systematics and Ecology*, **50**, 187–198.

Bassil N, Davis T, Amaya I *et al.* (2014) Development and Preliminary Evaluation of the IStraw90 Axiom® Array in the Cultivated Strawberry (*Fragaria ×ananassa*). Abstract W318, Plant and Animal Genome XXII, San Diego, CA, USA. https://pag.confex.com/pag/xxii/webprogram/Paper9546.html

Blair MW, Hurtado N (2013) EST-microsatellite markers from five sequenced cDNA libraries of common bean (*Phaseolos vulgaris* L.) comparing three bioinformatic algorithms. *Molecular Ecology Resources*, **13**, 688–695.

Cao MD, Tasker E, Willadsen K *et al.* (2014) Inferring short tandem repeat variation from paired-end short reads. *Nucleic Acids Research*, **42**, e16. doi:10.1093/nar/gkt1313.

Cardoso SD, Gonçalves D, Robalo JI, Almada VC, Canário AVM, Oliveira RF (2013) Efficient isolation of polymorphic microsatellites from high-throughput sequence data based on number of repeats. *Marine Genomics*, **11**, 11–16.

Castoe TA, Poole AW, de Koning APJ *et al.* (2012) Rapid microsatellite identification from Illumina paired-end genomic sequencing in two birds and a snake. *PLoS ONE*, **7**, e30953. doi:10.1371/journal.pone.0030953.

Cheng S, Puryear J, Cairney J (1993) A simple and efficient method for isolating RNA from pine trees. *Plant Molecular Biology Reporter*, **11**, 113–116.

Doulati-Baneh H, Mohammadi SA, Labra M (2013) Genetic structure and diversity analysis in *Vitis vinifera* L. cultivars from Iran using microsatellite markers. *Scientia Horticulturae*, **160**, 29–36.

Dufresnes C, Brelsford A, Béziers P, Perrin N (2014) Stronger transferability but lower variability in transcriptomic- than in anonymous microsatellites: evidence from Hylid frogs. *Molecular Ecology Resources*, **14**, 716–725.

Duran C, Singhania R, Raman H, Batley J, Edwards D (2013) Predicting polymorphic EST-microsatellites in silico. *Molecular Ecology Resources*, **13**, 538–545. doi:10.1111/1755-0998.12078.

Durand J, Bodénès C, Chancerel E *et al.* (2010) A fast and cost-effective approach to develop and map EST-microsatellite markers: oak as a case study. *BMC Genomics*, **11**, 570.

Eschbach E, Schöning S (2013) Identification of high-resolution microsatellites without a priori knowledge of genotypes using a simple scoring approach. *Methods in Ecology and Evolution*, **4**, 1076–1082.

Esselink GD, Smulders MJM, Vosman B (2003) Identification of cut rose (*Rosa hybrida*) and rootstock varieties using robust sequence tagged microsatellite site markers. *Theoretical Applied Genetics*, **106**, 277–286.

Grabherr MG, Haas BJ, Yassour M *et al.* (2011) Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nature Biotechnology*, **29**, 644–652. doi:10.1038/nbt.1883.

Guo E, Cui Z, Wu D, Hui M, Liu Y, Wang H (2013) Genetic structure and diversity of *Portunus trituberculatus* in Chinese population revealed by microsatellite markers. *Biochemical Systematics and Ecology*, **50**, 313–321.

Han Z-G, Guo W-Z, Song X-L, Zhang T-Z (2004) Genetic mapping of EST-derived microsatellites from the diploid *Gossypium arboreum* in allotetraploid cotton. *Molecular Genetics and Genomics*, **272**, 308–327.

Hibrand Saint Oyant L, Crespel L, Rajapakse S, Zhang L, Foucher F (2008) Genetic linkage maps of rose constructed with new microsatellite markers and locating QTL controlling flowering traits. *Tree Genetics and Genomes*, **4**, 11–23.

Hoffman JI, Nichols HJ (2011) A novel approach for mining polymorphic microsatellite markers In Silico. *PLoS ONE*, **6**, e23283.

Iwaizumi MG, Tsuda Y, Ohtani M, Tsumura Y, Takahashi M (2013) Recent distribution changes affect geographic clines in genetic diversity and structure of Pinus densiflora natural populations in Japan. *Forest Ecology and Management*, **304**, 407–416.

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

Kimura T, Nishitani C, Iketani H, Ban Y, Yamamoto T (2006) Development of microsatellite markers in rose. *Molecular Ecology Notes*, **63**, 810–812.

Lance SL, Love CN, Nunziata SO *et al.* (2013) 32 species validation of a new Illumina paired-end approach for the development of microsatellites. *PLoS ONE*, **8**, e81853.

Legendre M, Pochet N, Pak T, Verstrepen KJ (2007) Sequence-based estimation of minisatellite and microsatellite repeat variability. *Genome Research*, **17**, 1787–1796.

Lian C, Hogetsu T (2002) Development of microsatellite markers in black locust (*Robinia pseudoacacia*) using a dual-supression-PCR technique. *Molecular Ecology Notes*, **2**, 211–213.

Liang X, Chen X, Hong Y *et al.* (2009) Utility of EST-derived microsatellite in cultivated peanut (*Arachis hypogaea* L.) and *Arachis* wild species. *BMC Plant Biology*, **9**, 35.

Liu H, Li S, Hu P *et al.* (2013) Isolation and characterization of EST-based microsatellite markers for *Scatophagus argus* based on transcriptome analyses. *Conservation Genetics Resources*, **5**, 483–485. doi:10.1007/s12686-012-9833-0.

Ma K-H, Jang D-H, Dixit A *et al.* (2007) Characterization of 30 new microsatellite markers, developed from enriched genomic DNA library of zoysiagrass *Zoysia japonica* Steud. *Molecular Ecology Notes*, **7**, 1323–1325.

Meng J, Li D, Yi T, Yang J, Zhao X (2009) Development and characterization of microsatellite loci for *Rosa odorata* var. gigantea Rehder & EH Wilson (Rosaceae). *Conservation Genetics*, **10**, 1973–1976.

Nijveen H, van Kaauwen M, Esselink DG, Hoegen B, Vosman B (2013) QualitySNPng: a user-friendly SNP detection and visualization tool. *Nucleic Acids Research*, **41**, W587–W590. doi:10.1093/nar/gkt333.

Nordström S, Hedrén M (2007) Development of polymorphic nuclear microsatellite markers for polyploid and diploid members of the orchid genus *Dactylorhiza*. *Molecular Ecology Notes*, **7**, 644–647.

Nybom H, Weising K, Rotter B (2014) DNA fingerprinting in botany: past, present, future. *Investigative Genetics*, **5**, 1. doi:10.1186/2041-2223-5-1.

Park YH, Ahn SG, Choi YM *et al*. (2010) Rose (*Rosa hybrida* L.) EST-derived microsatellite markers and their transferability to strawberry (*Fragaria* spp.). *Scientia Horticulturae*, **125**, 733–739.

Rajapakse S, Byrne DH, Zhang L, Anderson N, Arumuganathan K, Ballard RE (2001) Two genetic linkage maps of tetraploid roses. *Theoretical and Applied Genetics*, **103**, 575–585.

Rozen S, Skaletsky HJ (2000) Primer3 on the WWW for general users and for biologist programmers. In: *Bioinformatics Methods and Protocols: Methods in Molecular Biology* (eds Krawetz S, Misener S), pp. 365–386. Humana Press, Totowa, New Jersey.

Saha MC, Cooper JD, Rouf Mian MA, Chekhovskiy K, May GD (2006) Tall fescue genomic microsatellite markers: development and transferability across multiple grass species. *Theoretical and Applied Genetics*, **113**, 1449–1458. doi:10.1007/s00122-006-0391-2.

Schmieder R, Edwards R (2011) Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, **27**, 863–864.

Shahin A, van Kaauwen M, Esselink D *et al*. (2012) Generation and analysis of expressed sequence tags in the extreme large genomes *Lilium* and *Tulipa*. *BMC Genomics*, **13**, 640.

Smee MR, Pauchet Y, Wilkinson P *et al*. (2013) Microsatellites for the marsh fritillary butterfly: de novo transcriptome sequencing, and a comparison with amplified fragment length polymorphism (AFLP) markers. *PLoS ONE*, **8**, e54721.

Smulders MJM, Bredemeijer G, Rus-Kortekaas W, Arens P, Vosman B (1997) Use of short microsatellites from database sequences to generate polymorphism among *Lycopersicon esculentum* cultivars and accessions of other *Lycopersicon* species. *Theoretical and Applied Genetics*, **94**, 264–272.

Smulders MJM, Vukosavljev M, Shahin A, van de Weg WE, Arens P (2012) High throughput marker development and application in horticultural crops. *Acta Horticulturae (ISHS)*, **961**, 547–551 http://www.actahort.org/books/961/961_72.htm

Spiller M, Linde M, Hibrand-Saint OL *et al*. (2010) Towards a unified genetic map for diploid roses. *Theoretical and Applied Genetics*, **122**, 489–500.

Swarts ND, Sinclair EA, Dixon KW (2007) Characterization of microsatellite loci in the endangered grand spider orchid *Caladenia huegelii* (Orchidaceae). *Molecular Ecology Notes*, **7**, 1141–1143.

Tang J, Vosman B, Voorrips RE, van der Linden GC, Leunissen JAM (2006) QualitySNP: a pipeline for detecting single nucleotide polymorphisms and insertions/deletions in EST data from diploid and polyploid species. *BMC Bioinformatics*, **7**, 438. doi:10.1186/1471-2105-7-438.

Tang J, Baldwin SJ, Jacobs JM *et al*. (2008) Large-scale identification of polymorphic microsatellites using an in silico approach. *BMC Bioinformatics*, **9**, 374. doi:10.1186/1471-2105-9-374.

Tong Z, Yang Z, Chen X *et al*. (2012) Large-scale development of microsatellite markers in *Nicotiana tabacum* and construction of a genetic map of flue-cured tobacco. *Plant Breeding*, **131**, 674–680. doi:10.1111/j.1439-0523.2012.01984.x.

Triwitayakorn K, Chatkulkawin P, Kanjanawattanawong S *et al*. (2011) Transcriptome sequencing of *Hevea brasiliensis* for development of microsatellite markers and construction of a genetic linkage map. *DNA Research*, **18**, 471–482.

Vukosavljev M, Di Guardo M, van de Weg WE, Arens P, Smulders MJM (2012) Quantification of Allele Dosage in tetraploid Roses. *ScienceMED (Bologna)*, **3**, 277–282.

Vukosavljev M, Zhang J, Esselink GD *et al*. (2013) Genetic diversity and differentiation in roses: a garden rose perspective. *Scientia Horticulturae*, **162**, 320–332 http://dx.doi.org/10.1016/j.scienta.2013.08.015

Wang B, Ekblom R, Castoe TA *et al*. (2012) Transcriptome sequencing of black grouse (*Tetrao tetrix*) for immune gene discovery and microsatellite development. *Open Biology*, **2**, 120054.

Wang JY, Song XM, Li Y, Hou XL (2013) In-silico detection of EST-microsatellite markers in three *Brassica* species and transferability in *B. rapa*. *Journal of Horticultural Science & Biotechnology*, **88**, 135–140.

Xiao J, Wu K, Fang DD, Stelly DM, Yu J, Cantrell RG (2009) New microsatellite markers for use in cotton (*Gossypium* spp.) improvement. *Journal of Cotton Science*, **13**, 75–157.

Xing C, Schumacher FR, Xing G, Lu Q, Wang T, Elston RC (2005) Comparison of microsatellites, single-nucleotide polymorphisms (SNPs) and composite markers derived from SNPs in linkage analysis. *BMC Genetics*, **6**(Suppl 1), S29.

Yan Z, Denneboom C, Hattendorf A *et al*. (2005) Construction of an integrated map of rose with AFLP, microsatellite, PK, RGA, RFLP, SCAR and morphological markers. *Theoretical and Applied Genetics*, **110**, 766–777.

Yuan S, Ge L, Liu C *et al*. (2013) The development of EST-microsatellite markers in *Lilium regale* and their cross-amplification in related species. *Euphytica*, **189**, 393–419.

Zhang LH, Byrne DH, Ballard RE, Rajapakse S (2006) Microsatellite development in rose and its application in tetraploid mapping. *Journal of the American Society of Horticultural Science*, **131**, 380–387.

Zhang J, Esselink GD, Che D, Fougère-Danezan M, Arens P, Smulders MJM (2013) The diploid origins of allopolyploid rose species studied using single nucleotide polymorphism haplotypes flanking a microsatellite repeat. *Journal of Horticultural Science and Biotechnology*, **88**, 85–92 http://www.jhortscib.org/Vol88/88_1/11.htm

Zitouna N, Marghali S, Gharbi M, Chennaoui-Kourda H, Haddioui A, Trifi-Farah N (2013) Mediterranean *Hedysarum* phylogeny by transferable microsatellites from *Medicago*. *Biochemical Systematics and Ecology*, **50**, 129–135.

M.J.M.S., M.V., R.G.F.V., P.C. and P.A. conceived the study. P.C. collected the plant material. M.V. and W.P.C.V.T.W. extracted RNA and made cDNA. G.D.E. did the bioinformatics analysis. M.V. and W.P.C.V.T.W. tested the markers. M.V. and G.D.E. analysed the data. M.V., P.A. and M.J.M.S. drafted the manuscript. All authors have read and approved the final version.

Data accessibility

Sequences in ENA (HG934830-HG934851).

Supporting Information

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

Table S1 Reads produced and microsatellite motifs found

Table S2 Overview of studies reporting microsatellite development in polyploids