De Novo Transcriptome Assembly and Development of Novel Microsatellite Markers for the Traditional Chinese Medicinal Herb, *Veratrilla baillonii* Franch (Gentianaceae)

Lei Wang[1], Zhengkun Wang[1], Jianbing Chen[2], Chunyan Liu[1], Wanlong Zhu[1], Liuyang Wang[3] and Lihua Meng[1]

[1]School of Life Sciences, Engineering Research Center of Sustainable Development and Utilization of Biomass Energy Ministry of Education, Key Laboratory of Ecological Adaptive Evolution and Conservation on Animals-Plants in Southwest Mountain Ecosystem of University in Yunnan Province, Yunnan Normal University, Kunming, Yunnan, People's Republic of China. [2]Department of Information Management, Yunnan Normal University, Kunming, Yunnan, People's Republic of China. [3]Department of Molecular Genetics and Microbiology, Duke University Medical Center, Durham, NC, USA.

Supplementary Issue: RNA: An Expanding View of Function and Evolution

ABSTRACT: *Veratrilla baillonii* Franch is an important Chinese medicinal herb for treating liver-related diseases, which has been over-collected in the recent decades. However, the effective conservation and related population genetic study has been hindered because of the lack of genome sequences and genetic markers in the natural population. We have conducted RNA-seq on *V. baillonii*. We performed de novo assembly of these data to characterize the *V. baillonii* transcriptome, resulting in 133,019 contigs with size >200 bp. These contigs were annotated using the NCBI nonredundant database and Gene Ontology (GO) terms. From these contigs, we developed novel microsatellite simple sequence repeat (SSR) markers, identifying a total of 40,885 SSRs. SSRs with repeat motifs of 1–4 bp (mono-, di-, tri-, and tetranucleotides) accounted for 99.8% of all SSRs, with mononucleotide repeats most common, followed by dinucleotide (16.2%) and trinucleotide repeats (14.7%). We selected 151 SSRs for experimental validation, of which 74 were confirmed by polymerase chain reaction. Fourteen SSRs were determined to be polymorphic by screening 40 individuals from six distant populations. The number of alleles per locus ranged from two to four, and the expected heterozygosity varied from 0.2637 to 0.8571, suggesting that these SSR markers are highly polymorphic and effective for further genetic analysis in the nature population. In addition, we explored the genetic structure of *V. baillonii* using five SSRs in four geographic populations and found that the identified genotypes were clustered into two phylogenetic clades: the Mekong River clade and Jinsha River clade. This result indicates that these two regions may harbor highly divergent genetic lineages and enriched genetic diversity. The de novo transcriptome sequences and new SSR markers discovered by this study provide an initial step for understanding the population genetics of *V. baillonii*, and a valuable resource for effective conservation management.

KEYWORDS: transcriptome, Illumina RNA-Seq, microsatellite (SSR) markers, *Veratrilla baillonii*

SUPPLEMENT: RNA: An Expanding View of Function and Evolution

CITATION: Wang et al. De Novo Transcriptome Assembly and Development of Novel Microsatellite Markers for the Traditional Chinese Medicinal Herb, *Veratrilla baillonii* Franch (Gentianaceae). *Evolutionary Bioinformatics* 2015:11(S1) 39–45 doi: 10.4137/EBO.S20942.

RECEIVED: April 15, 2015. **RESUBMITTED:** June 02, 2015. **ACCEPTED FOR PUBLICATION:** June 08, 2015.

ACADEMIC EDITOR: Jike Cui, Associate Editor

TYPE: Original Research

FUNDING: This study was funded by the Natural Science Foundation of China (31160084, 31460096). The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: wallacewly@gmail.com, menglihua@mails.gucas.ac.cn

Introduction

Microsatellites, or simple sequence repeats (SSRs), are short tandem repeats of 1–6 bp nucleotides, which have increasingly been used in population genetics analysis, crop breeding, and conservation genetics. SSRs are widespread in both protein-coding and noncoding regions in plants.[1] Because of their abundance, high levels of allelic variation, and codominant inheritance characteristics, SSRs are considered to be an effective genetic marker in genetic diversity analysis, gene tagging, and conservation biology in plants.[2,3] SSRs are generally categorized into two common groups based on their origins, genomic SSRs from genomic sequences, and expressed sequence tag (EST)-SSRs from transcribed RNA sequences.[4]

The de novo development of genomic SSRs is time consuming and labor intensive, involving the screening of a genomic DNA library screening with specific SSR probes to isolate the microsatellite sequences.[4] In contrast, development of EST-SSR markers has many advantages, including low cost, technical simplicity, and the ability to capture the functional diversity in natural populations or germplasm collections, as well as high transferability to related species.[4] The NCBI dbEST database collects cDNA sequences from a number of organisms. However, dbEST does not include the vast majority of known species, having particular poor coverage of nonmodel organisms, especially those with low economic value. Lack of EST data makes it difficult to develop effective EST-SSR markers.

Next-generation sequencing (NGS) is an important tool for generating large quantities of genomic data from nonmodel organisms in a cost-effective manner. In particular, genome-scale transcriptome analysis by RNA-seq enables identification of genes that have high differential expression in response to environment changes, determination of the genetic basis of critical phenotypes, and mapping genomic diversity in nonmodel organisms. RNA-seq can accelerate the development of SSR markers, which is particularly useful for genetic analysis in the nonmodel species. Nonmodel plant species that have benefited from RNA-seq based SSR marker development include *Lilium*,[5] buckwheat (*Fagopyrum esculentum* and *F. tataricum*),[6] and big sagebrush (*Artemisia tridentata*).[7] EST-SSRs have provided a direct and reliable approach for exploring the genetic diversity and evolutionary history in these species.

Veratrilla baillonii Franch (Gentianaceae) is an important Chinese medicinal herb that is distributed only in the eastern Himalayas and Hengduan Mountains of southwest China. *V. baillonii* has been widely used to cure liver-related diseases with a common name of "Jin Bu Huan".[8] Because of the overexploitation, the distribution and natural populations for *V. baillonii* have shrunk substantially in recent years. However, the effective conservation management has been hampered by the lack of genome information and effective genetic markers. In this study, we aimed to obtain a de novo transcriptome for *V. baillonii* via RNA-seq and develop efficient SSR markers. We assembled the resulting genome-scale expression data, generating what is, to the best of our knowledge, the first reported transcriptome sequence for *V. baillonii*. We screened this transcriptome for identifying 40,885 putative EST-SSR markers. We selected 151 of these SSRs for experimental testing, which validated 74 markers. Further examination demonstrated 14 SSR markers to be highly polymorphic across six distant populations. This de novo transcriptome assembly and collection of SSR markers represent a first step toward the large-scale genomic analysis of *V. baillonii*. These resources can be used for further studies on genetic diversity and population demographic history, and as a tool in developing conservation strategies.

Methods

Plant material. *V. baillonii* is an important Chinese medicinal herb and grows only in the high altitudes (3700–4600 m) of the Himalaya–Hengduan Mountains in southwest China. During July 2013, we sampled the fresh leaves of *V. baillonii* from Shangri-la in northwest Yunnan (28°31'0"N, 99°57'0"E, alt. 4514 m) and quickly stored in liquid nitrogen. In addition, 40 individuals from six populations were collected and kept *in silica* gel for DNA extraction, polymerase chain reaction (PCR) amplification, SSR marker validation, and analysis of genetic diversity. Detailed information for the plant materials was listed in Supplementary Table 1.

RNA extraction and sequencing. Total RNA was extracted from the samples, using a CTAB procedure.[9] A260/A280 ratios of the RNA samples dissolved in 10 mm Tris (pH 7.6) ranged from 1.9 to 2.1. The integrity of the RNA samples was examined with an Agilent 2100 Bioanalyzer and their RIN (RNA integrity number) values ranged from 8.6 to 10.0, with no sign of degradation. RNA from each replicate was pooled with equal volumes to obtain enough RNA for RNA-Seq.

A total of 20 μg RNA was used for RNA sequencing, and the mRNA was fragmented into small pieces using divalent cations at an elevated temperature. The cDNA library was constructed via poly-A enriched RNA method, and 200–300 bp fragments were chosen for paired-end sequencing based on Illumina protocols (San Diego, CA, USA). The double strands were synthesized with random hexamer primers. The short fragments were purified with the QIAquick PCR Purification kit (Qiagen Inc.). The purified DNA libraries were first amplified via PCR and then sequenced on Illumina HiSeq™ 2000 platform.

De novo assembly. Raw reads were first filtered by removing the adaptors and reads with >eight ambiguous bases and >50% of the bases with a quality score ≤5 using in-house *Perl* scripts (available upon request). The transcriptome sequence was assembled into distinct contigs by Trinity tools[10] with default parameters. For convenience, we used those unigenes to represent the nonredundant (NR) contigs in the further analysis.

SSR locus search, primer acquisition, and validation. The unigenes were used for detecting SSR loci by MicroSAtellite (MISA, http://pgrc.ipk-gatersleben.de/misa).[11] Criteria includes a minimum of five repeats for simple motifs, and three repeats for complex or imperfect repeats, a motif length of 2–10 bp, and for compound SSRs, a maximum interruption distance of 100 bp between different SSRs. To facilitate SSR detection, only 1–6-nucleotide motifs were considered, and the minimum repeat unit was defined as 10 for mono-, 6 for di-, and 5 for tri-, tetra-, penta-, and hexanucleotides. Primer pairs of each unique SSR were designed using Primer 3.0,[12] with target microsatellites containing at least five repeats and yielding PCR products of 80–500 bp. One hundred fifty-one primer pairs were synthesized and used for validation (Supplementary Table 2). Screened primer pairs giving good amplification were subsequently used to characterize polymorphism among 40 individuals from six populations (Supplementary Table 1). PCR was performed in a 25-μL volume, containing 10–40 ng plant DNA. The PCR reactions were carried out under the following conditions: DNA initial denaturation at 94 °C for 4 minutes, 35 cycles of 94 °C for 1 minute 30 seconds, annealing temperature ranging from 45 °C to 60 °C for 50 seconds, 72 °C for 50 seconds, and a final extension at 72 °C for 7 minutes. The PCR products were purified before sequencing to remove excess primers and deoxynucleotide triphosphates using a TIAN quick Midi Purification Kit (Tiangen Biotechnology Co. Ltd.), and then, sequencing reactions were performed using ABI Prism Sequencing Ready

Reaction Kit with the same primers as PCRs and analyzed on the ABI 3730 genetic analyzer (Applied Biosystems).

Functional annotation for unigenes containing SSRs. Functional annotation of SSR-containing coding sequences was conducted using the program Blast2GO.[13] All the SSR-containing unigenes were blasted against the NCBI's NR protein database using BLASTx. The E-value threshold was set as 1e-6. The contig was assigned with gene names according to best BLASTx hits. The distributions of functional categories were plotted with the program WEGO.[14]

Population genetic analyses. We used POPGEN v1.32[15] to calculate the number of alleles and expected and observed heterozygosity. Five primer pairs were selected for phylogenetic analysis using 23 individuals from four populations. We used MEGA6[16] to construct the dendrogram tree using the unweighted pair-group as implemented in the UPGMA method.

Results and Discussion

De novo assembly. A total of 28,483,317 high-quality RNA-Seq reads passed our stringent quality assessment and filtering. We used the Trinity assembler,[10] to assemble these reads, resulting in 133,019 contigs. The length of the contigs in our assembly ranged from 201 to 13830 bp, with an average 1263 bp and median of 850 bp. The N50 value of our assembly is 2104 bp. N50 is an important measurement for quantifying assembly quality and is measured by the length of the contigs for which all contigs of that length or longer contain 50% of the bases in the assembly. The N50 for *V. baillonii* transcriptome assembly is much higher than those from lily[17] and *Gossypium arboreum*.[18] The GC-content of our assembly is 40.6% (Table 1). The distribution of contig sizes follows a power-law-like distribution, with the number of contigs decreasing with increasing contig length (Fig. 1), as observed in *Primula poissonii* and *P. wilsonii*.[19]

Frequency and distribution of different types of SSR markers. MISA was used to analyze the 133,019 contigs, identifying 40,885 putative SSRs (Table 1). The number of SSRs obtained in this study was lower than known model plants, including *Arabidopsis thaliana* (50,092), *Medicago truncatula* (152,461), *Oryza sativa* (135,265), and *Sorghum bicolor* (129,564).[20] The density of SSRs was 243.3 per Mb for *V. baillonii*, which was higher than *Brachypodium distachyon* (191.3 per Mb) and *S. bicolor* (175.4 per Mb).[20] The reduced number and high density of SSRs may indicate unique evolutionary history for *V. baillonii*.

A detailed summary of SSRs, including repeat motif and total number of different repeat motifs, is shown in Table 2. SSRs with repeat motifs of 1–4 bp (mono-, di-, tri-, and tetranucleotides) accounted for 99.8% of the total (Table 2). The frequency of SSRs decreased with the increase of motif length (mono- to hexanucleotide repeats), with mononucleotide being the dominant repeat unit. Proportions of dinucleotide and trinucleotide repeats were 16.2% and 14.7%, respectively. The combined number of tetra-, penta-, and hexanucleotide repeats accounted for 2.7% of SSRs, which was higher than in *Populus trichocarpa* (1.66%), *M. truncatula* (0.94%), *O. sativa* (2.54%), *B. distachyon* (2.45%), and *A. thaliana* (0.53%).[20]

Of the two possible types of mononucleotide repeats, the most abundant was (A/T)n (97.1%), as in most plants,[20,21] and the (G/C)n contributed 1.90% to total SSRs, which was higher than 0.05% in tree peony.[21] For the dinucleotide repeat category, different species have unique motif frequency distributions. For example, AG/CT repeats were more frequent in *B. distachyon* and *O. sativa* with 50.7% and 41.9% frequencies, respectively, whereas AT/AT repeats were more frequent in *P. trichocarpa* (60.5%) and *M. truncatula* (59.9%). Similar to the latter two species, AT/AT repeats in *V. baillonii* were most abundant, accounting for 49.0% of dinucleotide repeats. The AC/GT and AG/CT repeats were abundant in *V. baillonii*, accounting for 24.6% and 26.3% of dinucleotide repeats, respectively. The CG/CG repeats contributed <0.08%, similar to *P. trichocarpa*, *M. truncatula*, and *A. thaliana*, suggesting that CG-rich motifs were the least preferred in dicot genomes.

Trinucleotide repeats AGC/CGT, AGG/CCT, and CCG/CGG were observed more frequently in all the monocot species, whereas A/T-rich repeats, such as AAC/GTT, AAG/CTT, and AAT/ATT, were preferred in dicots. Similar to the results of Sonah et al.[20], A/T-rich repeats were the dominant trinucleotide SSRs in *V. baillonii*. In addition, AGC/CTG repeats were abundant, accounting for 15.6%. The dominant occurrence of repeat motifs from particular sequences with certain size in plant genomes is generally caused by natural selection. However, like in tree peony, ACG/CGT and CCG/CGG repeats were scarce, which may be because of highly mutable CpG dinucleotide repeats. The CCG repeats may also be under selection, which could be because of specific splicing pattern for the maintenance of other related tandem repeat forms.[22] The overall absence of a particular repeat motif may also indicate the presence of strong selective pressure.[22] Replication slippage is the most common mechanism for creating

Table 1. Summary of assembly and annotation results for *V. baillonii*.

	V. BAILLONII
Total number of high quality reads	28483317
Total number of contigs	133019
Total size of contigs (bp)	168009542
Mean length of contigs	1263
N50 value of contigs	2104
Length range of contigs	201–13830
GC content	40.6%
Total number of identified SSRs	40885
SSRs containing sequences with BLASTx hit	28912 (70.7%)
SSRs containing sequences with annotation	11148 (27.3%)

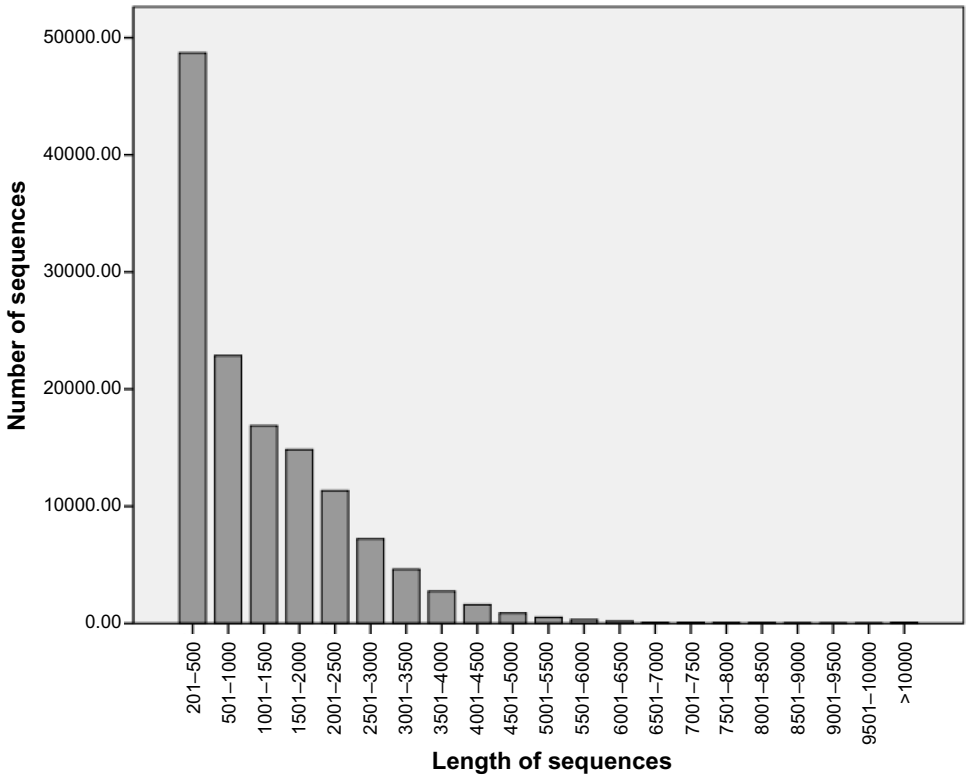

Figure 1. Length distributions for all 133,019 transcriptome contigs for *V. baillonii*, of size >200 bp.

Table 2. Counts of various SSR types with different repeat motifs in *V. baillonii*.

REPEATS	COUNTS
Mono-nucleotide	**27136**
A/T	26360
C/G	776
Bi-nucleotide	**6620**
AC/GT	1631
AG/CT	1739
AT/AT	3245
CG/CG	5
Tri-nucleotide	**6010**
AAC/GTT	215
AAG/CTT	1828
AAT/ATT	572
ACC/GGT	612
ACG/CGT	63
ACT/AGT	197
AGC/CTG	1036
AGG/CCT	768
ATC/ATG	594
CCG/CGG	125
Tetra-nucleotide	**1014**
Penta-nucleotide	**105**
Hexa-nucleotide	**0**

and/or mutating microsatellites, involving changes of motif repeats. However, the molecular mechanisms for the origin of microsatellites are still not clear.

SSR loci were divided into two groups based on the size of SSR tracts and possible informative genetic markers: Class I, or hypervariable markers, consisted of SSRs ≥20 bp, and Class II, or potentially variable markers, consisted of SSRs ≥12 bp and <20 bp. Class I microsatellites are generally highly polymorphic and more informative because of the large size and long repeats, which was first observed in human beings,[23] and subsequently confirmed by studies in a number of other organisms, including rice.[22,24] Class II microsatellites are less variable, representing mutations that have accumulated recently because of sporadic SSR expansion. In *V. baillonii*, 6.8% of SSRs were categorized as Class I microsatellites and 93.2% as Class II microsatellites. The small fraction of Class I microsatellites may make it difficult to further develop efficient and polymorphic microsatellite markers in *V. baillonii*. The preponderance of short SSRs could reflect from the characteristics of physiology or development, or may simply be a result of genetic drift.

SSR-containing coding sequences annotation. A large number of the contigs in our assembly (28,912, 70.7%) had more than one hit in the NCBI NR database with an E-value of 1e-6. This percentage was higher than *P. poissonii* (65.6%) or *P. wilsonii* (65.1%).[19] We performed Gene Ontology (GO) annotation on unigenes containing the SSRs. The number of SSR-containing genes in each GO classification is shown in Figure 2. A total of

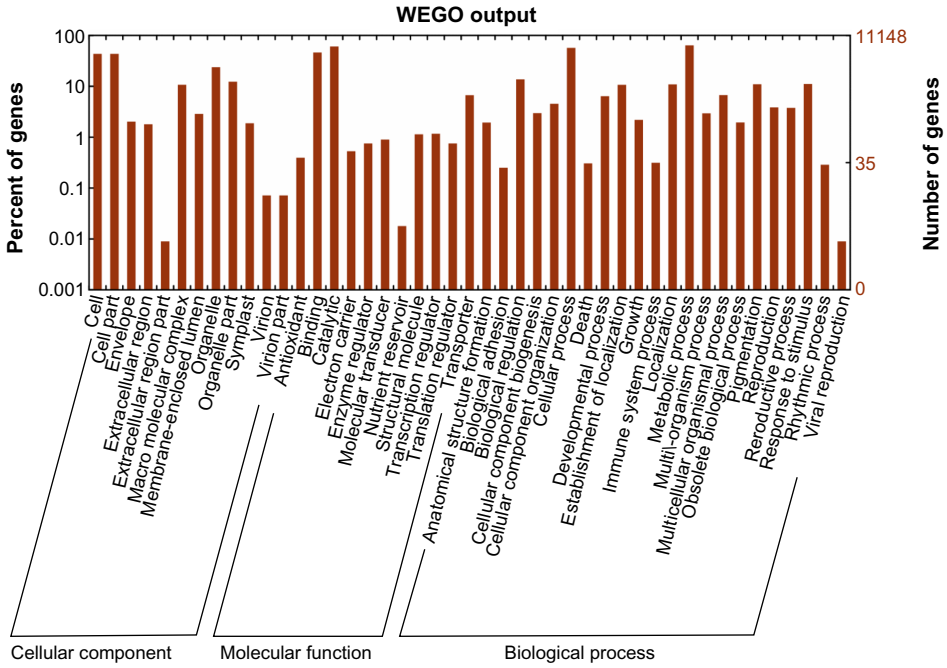

Figure 2. GO classification of SSRs in coding regions. The *x*-axis refers to the three functional classes. The *y*-axis indicates the percentage (left) and number (right) of genes that contain an SSR, which belong to each functional class.

11,148 contigs containing SSRs were annotated, representing 27.3% of total identified SSRs. The unigenes were annotated with the three categories: cellular component, molecular function, and biological processes. For the cellular component category, the two most common were cell and cell part. For the molecular function category, catalytic activity was abundant, followed by binding. The biological processes category, cellular process, and metabolic process were the most abundant terms (Fig. 2). These results may give hints of function constraints in *V. baillonii*, and similar results were reported in other plants, such as tree peony and *P. poissonii and P. wilsonii*.[19]

Validation of SSR assays and UPGMA analysis. RNA-seq is a labor- and cost-effective method to identify a large number of candidate SSRs for organisms such as *V. bailloni*, which are not represented in dbEST. RNA-seq for SSRs development is particularly attractive in comparison to the cost and labor-intensive process of identifying genomic SSRs. In order to validate the identified SSRs, we designed 151 pairs of SSRs primers for PCR screening in six natural populations. PCR amplification was successful for 74 of these SSRs (Supplementary Table 2). Microsatellites containing the (AT)n repeats represented the most abundant and polymorphic type among all markers, but we only obtained PCR products from a few of these markers, similar to rice.[25] The validated primers were subsequently used to characterize genetic polymorphism among 40 individuals from six populations (Supplementary Table 1). Of these, 14 were found to be polymorphic. The number of alleles per locus ranged from two to four, and the expected heterozygosity varied between 0.2637 and 0.8571, which was higher than tree peony of which observed heterozygosity ranged from 0.0000 to 0.8751 (Table 3). Then, we identified eight distinct genotypes among 23 individuals from four populations using five primer pairs (Supplementary Table 3). The eight genotypes were grouped into two major clusters in the UPGMA dendrogram (Fig. 3). Cluster I was only composed of the YN Deqin population from Mekong River. Cluster II contained five genotypes collected from YN Dali, Lijiang, and SC Jiulong populations, which generally fall into the Jinsha River region. The phylogenetic relationship is generally in agreement with the geographic locations, indicating that the river formation might play a critical role in shaping present genetic structure in *V. baillonii*. However, we cannot rule out the effect of past climate changes, for example, the Pleistocene glacial–interglacial climate cycles. Elucidating this will require further assessment with additional population.

Plant genomes are very complex and contain large amounts of repetitive DNA, including microsatellites, which has immediate practical implications for the success of SSR marker development. The 14 SSRs identified in this study represent high-quality and polymorphic genomic loci, which will allow us to further explore the genetic diversity and genetic structure in wild populations of *V. baillonii*. These markers will be an important tool in characterizing the phylogeographic patterns and evolutionary history of *V. baillonii* and related species of *Veratrilla* and will help guide development of a conservation strategy for this traditional medicinal plant.

Conclusions

In this study, we used RNA-seq to determine a de novo transcriptome for *V. baillonii*. We identified a number of

Table 3. Fourteen SSR primers, size, and summary statistic across four populations in *V. baillonii*.

LOCUS	REPEAT	FORWARD PRIMER (5′–3′)	REVERSE PRIMER (5′–3′)	Ta (°C)	SIZE(BP)	Na	He	Ho
HQJ19	(TA)6	TTTGCTTACCGTTTGTCC	AATGCTTCCAGCCTATCC	52	187–192	2	0.3310	0.0000
HQJ34	(TGG)6	CGTTACGGTCTTTCCTTG	AATACCTCACTCCTCCACAT	58	191–197	4	0.6095	0.1111
HQJ35	(TAAAA)4	CCGAACAAACAACTCATT	TCCTGTATTCACCCTCCT	54	163–178	2	0.3692	0.0000
HQJ37	(AAGA)5	GCTCGTTTCGTTTGTTTC	GTCGGTTATGAGATTCCATC	58	105–129	4	0.6841	0.1667
HQJ40	(GAT)7	AGCGTCTATTGGGCAGTG	AAAAGCAGAGTGAAGAAACATC	58	52–149	4	0.5270	0.3333
HQJ45	(AT)6	CAGCCTCACGCTCAACAA	CGACGGCCTACCATCTTT	55	195–199	3	0.7750	0.3750
HQJ56	(TA)9	CTAAAAATGATGAACTCCCGAAAAA	ACTGAGCAGCACAGCACAAC	58	98–106	4	0.7033	0.8751
HQJ63	(TA)5	ACGGAGGACATCACGAGC	TGGCAGGGCAAACCATAT	52	116–118	2	0.2637	0.0000
HQJ79	(AGA)6	CAGCTTGCGAGGATACGG	CTTCCCAAACTGCGAGGC	58	157–162	2	0.6667	0.0000
HQJ99	(GCC)6	GAGCAATCAGGAGGAGGG	GGGAAATGAACAGCGACTT	60	156–168	2	0.3556	0.0000
HQJ103	(ACTC)5	TGACTCCTTGACTGACCCTC	TGCAGCAGCTTGCTTTAT	58	122–133	3	0.5333	0.0000
HQJ115	(AT)10	GTTCTGTTGCTACCTGTG	TTGTCTCATTTTGCTTTC	56	200–209	2	0.8571	1.0000
HQJ134	(TA)9	TCCTCCTCCTTTATCACA	GTGCAGTATTAAGCGTTG	57	365–377	3	0.5455	1.0000
HQJ137	(AC)9	TTTCACGCTCATCTTTTA	CCTTTTGGCAGTCATTAT	52	231–233	2	0.5455	1.0000

Abbreviations: Size, size of cloned allele; Ta, annealing temperature; Na, number of alleles; He, expected heterozygosity; Ho, observed heterozygosity.

microsatellite markers from the de novo transcriptome sequences and carried out experimental validations for 151 SSRs. Finally, 14 SSR markers were found to be polymorphic, making them useful tools for future studies in this species. The de novo transcriptome and SSR markers identified in this study provide an initial step and valuable resources for understanding the genetic diversity and population history of *V. baillonii* and for the development of a conservation strategy for this species.

Acknowledgments

The authors are grateful to the staff in the Department of Information Management of Yunnan Normal University for providing high-performance computing cluster and Dr. Yuanwen Duan for the field sampling. They thank Joshua Granek for critical reading of this manuscript. This study was funded by the Natural Science Foundation of China (31160084, 31460096).

Author Contributions

Carried out the laboratory experiments and statistical analysis: LW. Participated in the sample collections and the statistical analysis: ZKW. Assisted with bioinformatics tools: JBC. Guided the appropriateness of the tools: CYL, WLZ. Performed bioinformatics analysis and participated

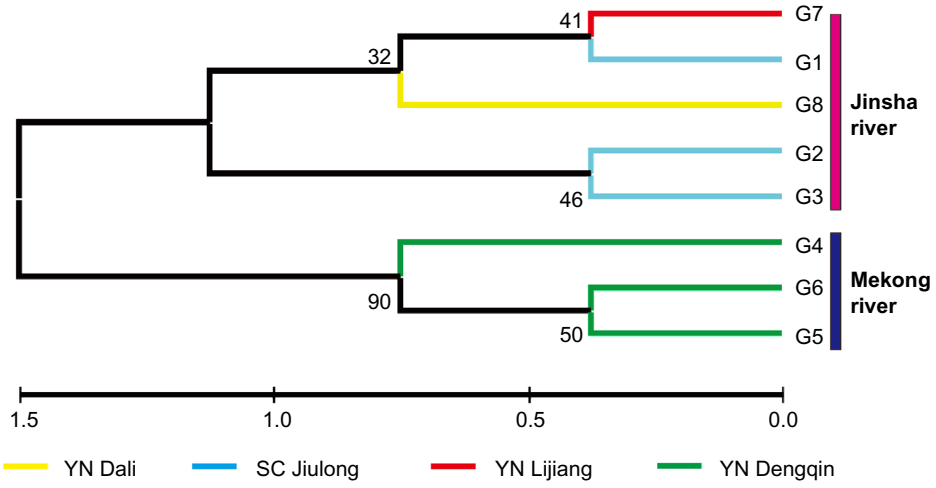

Figure 3. UPGMA dendrogram constructed based on eight genotypes from four representative populations and five SSR markers developed in this study. Two clusters were identified, generally corresponding to two geologic locations: the Mekong River and the Jinsha River (YN, Yun Nan Province; SC, Si Chuan Province).

in writing: LYW. Conceived the idea, guided the study, performed bioinformatics analysis, and participated in writing: LHM. All the authors participated in the editing of the manuscript, reviewed, and approved the final manuscript.

Supplementary Materials

Supplementary Table 1. Locations of populations of sampled *V. baillonii*.

Supplementary Table 2. Primer pairs synthesized and used for validation based on SSR sequences from *V. baillonii*.

Supplementary Table 3. The frequency of repeat motif for eight genotypes observed in *V. baillonii*.

REFERENCES

1. Lawson MJ, Zhang L. Distinct patterns of SSR distribution in the *Arabidopsis thaliana* and rice genomes. *Genome Biol*. 2006;7(2):R14.
2. Li D, Deng Z, Qin B, Liu X, Men Z. De novo assembly and characterization of bark transcriptome using Illumina sequencing and development of EST-SSR markers in rubber tree (*Hevea brasiliensis* Muell. Arg.). *BMC Genomics*. 2012;13:192.
3. Silva PIT, Martins AM, Gouvea EG, Pessoa-Filho M, Ferreira ME. Development and validation of microsatellite markers for *Brachiaria ruziziensis* obtained by partial genome assembly of Illumina single-end reads. *BMC Genomics*. 2013;14:17.
4. Varshney RK, Graner A, Sorrells ME. Genic microsatellite markers in plants: features and applications. *Trends Biotechnol*. 2005;23(1):48–55.
5. Du F, Wu Y, Zhang L, et al. De novo assembled transcriptome analysis and SSR marker development of a mixture of six tissues from lilium oriental hybrid 'sorbonne'. *Plant Mol Biol Rep*. 2014:1–13.
6. Logacheva MD, Kasianov AS, Vinogradov DV, et al. De novo sequencing and characterization of floral transcriptome in two species of buckwheat (*Fagopyrum*). *BMC Genomics*. 2011;12:30.
7. Bajgain P, Richardson BA, Price JC, Cronn RC, Udall JA. Transcriptome characterization and polymorphism detection between subspecies of big sagebrush (*Artemisia tridentata*). *BMC Genomics*. 2011;12:370.
8. Ho TN, Pringle JS. Gentianaceae. In: Wu ZY, Raven PH, eds. *Floral of China*. Vol. 16. Beijing, St. Louis: Science Press and Missouri Botanical Garden; 1995.
9. Chang S, Puryear J, Cairney J. A simple and efficient method for isolating RNA from pine trees. *Plant Mol Biol Rep*. 1993;11(2):113–6.
10. Grabherr MG, Haas BJ, Yassour M, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol*. 2011;29(7):644-U130.
11. Sharma PC, Grover A, Kahl G. Mining microsatellites in eukaryotic genomes. *Trends Biotechnol*. 2007;25(11):490–8.
12. Rozen S, Skaletsky H. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol*. 2000;132:365–86.
13. Conesa A, Gotz S. Blast2GO: a comprehensive suite for functional analysis in plant genomics. *Int J Plant Genomics*. 2008;2008:619832.
14. Ye J, Fang L, Zheng H, et al. WEGO: a web tool for plotting GO annotations. *Nucleic Acids Res*. 2006;34:W293–97.
15. Nei M. Estimation of average heterozygosity and genetic distance from small number of individuals. *Genetics*. 1978;89(3):583–90.
16. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol Biol Evol*. 2013;30(12):2725–9.
17. Shahin A, van Kaauwen M, Esselink D, et al. Generation and analysis of expressed sequence tags in the extreme large genomes *Lilium* and *Tulipa*. *BMC Genomics*. 2012;13:640.
18. Zhang X, Yao D, Wang Q, et al. mRNA-seq analysis of the *Gossypium arboreum* transcriptome reveals tissue selective signaling in response to water stress during seedling stage. *PLoS One*. 2013;8(1):e54762.
19. Zhang L, Yan HF, Wu W, Yu H, Ge XJ. Comparative transcriptome analysis and marker development of two closely related Primrose species (*Primula poissonii* and *Primula wilsonii*). *BMC Genomics*. 2013;14:329.
20. Sonah H, Deshmukh RK, Sharma A, et al. Genome-wide distribution and organization of microsatellites in plants: an insight into marker development in *Brachypodium*. *PLoS One*. 2011;6(6):e21298.
21. Gao Z, Wu J, Liu Z, Wang L, Ren H, Shu Q. Rapid microsatellite development for tree peony and its implications. *BMC Genomics*. 2013;14:886.
22. Cho YG, Ishii T, Temnykh S, et al. Diversity of microsatellites derived from genomic libraries and GenBank sequences in rice (*Oryza sativa* L.). *Theor Appl Genet*. 2000;100(5):713–22.
23. Weber JL. Informativeness of human (dC-dA)n (dG-dT)n polymorphisms. *Genomics*. 1990;7(4):524–30.
24. Temnykh S, Park WD, Ayres N, et al. Mapping and genome organization of microsatellite sequences in rice (*Oryza sativa* L.). *Theor Appl Genet*. 2000;100(5):697–712.
25. Temnykh S, DeClerck G, Lukashova A, Lipovich L, Cartinhour S, McCouch S. Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length variation, transposon associations, and genetic marker potential. *Genome Res*. 2001;11(8):1441–52.