OPEN ACCESS International Journal of Molecular Sciences ISSN 1422-0067 www.mdpi.com/journal/ijms

Article

First Insights into the Large Genome of *Epimedium sagittatum* (Sieb. et Zucc) Maxim, a Chinese Traditional Medicinal Plant

Di Liu ^{1,2}, Shao-Hua Zeng ³, Jian-Jun Chen ¹, Yan-Jun Zhang ¹, Gong Xiao ^{1,2}, Lin-Yao Zhu ⁴ and Ying Wang ^{1,*}

- ¹ Key Laboratory of Plant Germplasm Enhancement and Specialty Agriculture, Wuhan Botanical Garden, Chinese Academy of Sciences, Wuhan 430074, China; E-Mails: liudi206@mails.ucas.ac.cn (D.L.); jianjunchen@wbgcas.cn (J.-J.C.); yanjunzhang@wbgcas.cn (Y.-J.Z.); gongxiaobio@gmail.com (G.X.)
- ² University of Chinese Academy of Sciences, Beijing 100039, China
- ³ Key Laboratory of Plant Resources Conservation and Sustainable Utilization, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou 510650, China; E-Mail: shhzeng@scib.ac.cn
- ⁴ Wuhan Vegetable Research Station, Wuhan 430065, China; E-Mail: zly01062535@163.com
- * Author to whom correspondence should be addressed; E-Mail: yingwang@wbgcas.cn; Tel.: +86-27-8751-0675; Fax: +86-27-8751-0251.

Received: 7 January 2013; in revised form: 16 May 2013 / Accepted: 6 June 2013 / Published: 27 June 2013

Abstract: *Epimedium sagittatum* (Sieb. et Zucc) Maxim is a member of the Berberidaceae family of basal eudicot plants, widely distributed and used as a traditional medicinal plant in China for therapeutic effects on many diseases with a long history. Recent data shows that *E. sagittatum* has a relatively large genome, with a haploid genome size of ~4496 Mbp, divided into a small number of only 12 diploid chromosomes (2n = 2x = 12). However, little is known about *Epimedium* genome structure and composition. Here we present the analysis of 691 kb of high-quality genomic sequence derived from 672 randomly selected plasmid clones of *E. sagittatum* genomic DNA, representing ~0.0154% of the genome. The sampled sequences comprised at least 78.41% repetitive DNA elements and 2.51% confirmed annotated gene sequences, with a total GC% content of 39%. Retrotransposons represented the major class of transposable element (TE) repeats identified (65.37% of all TE repeats), particularly LTR (Long Terminal Repeat) retrotransposons (52.27% of all TE repeats). Chromosome analysis and Fluorescence *in situ* Hybridization of *Gypsy-Ty3* retrotransposons were performed to survey the

E. sagittatum genome at the cytological level. Our data provide the first insights into the composition and structure of the *E. sagittatum* genome, and will facilitate the functional genomic analysis of this valuable medicinal plant.

Keywords: *Epimedium sagittatum*; Chinese medicinal plant; structural genomics; genome composition; *Gypsy-Ty3* retrotransposon; repetitive elements; FISH; chromosome

Abbreviations: ENS, *Epimedium sagittatum* Nuclear Sequences; TE, Transposable Element; LTR, Long Terminal Repeat; LINE, Long Interspersed Nuclear Element; SINE, Short Interspersed Nuclear Element; MITE, Miniature Inverted-repeat Transposable Element.

1. Introduction

Epimedium L. ("*Yin Yang Huo*" in Chinese) is a genus of the Berberidaceae family, endemic to and widely distributed in China [1]. Species of *Epimedium* have been used as traditional medicinal plants in China and East Asia for more than 2000 years. To date, five species have been recorded in the Chinese Pharmacopoeia as medicinal plants: *E. brevicornu* Maxim, *E. sagittatum* (Sieb. et Zucc) Maxim, *E. pubescens* Maxim, *E. wushanense* T. S. Ying and *E. koreanum* Nakai (Chinese Pharmacopoeia Commission, 2005) [2]. The significant worth of these *Epimedium* species to traditional medicine is attributed largely to their high levels of bioactive chemicals, especially flavonoids [3], which play an important role in cell senescence delay [4] and retarding aging [5]. In addition, *Epimedium* has attracted increasing commercial attention for its use in the treatment of impotence, spermatorrhea, infertility, amenorrhea and menopause symptoms [6].

To date, research into *Epimedium* species has concentrated on taxonomy and phytogeography [7,8], phylogenetic analysis [9–11], and chemical and pharmacological investigations [3–6,8]. More recently, the development of an EST dataset and EST-SSRs in *E. sagittatum* [12] combined with characterization of the genes involved in flavonoid biosynthesis [13,14] and the evolution of carpel and nectary development in this basal eudicot [15], have shed new light on *Epimedium* functional genomics. Fundamentally, however, our understanding of the basic genomic characteristics of the genus *Epimedium*, such as genome size and genome structure, is still lacking

Given that *Epimedium* is a genus of Berberidaceae, a member of Ranunculales lying in a basal clade of eudicots, research into *Epimedium* genomics has the potential to enhance our understanding of the phylogeny and evolution of angiosperms, from which almost all modern cultivated species are exclusively derived. Therefore, given not only the great medicinal value, but also the important placement in evolutionary history, of *Epimedium*, investigation of the *Epimedium* genome is both warranted and timely.

All *Epimedium* species analysed to date have only 12 chromosomes (2n = 2x = 12) [16], but genome sizes range from 4115.35 Mbp/1C to 4876.4 Mbp/1C, giving a median haploid genome size of ~4496 Mbp [17]. This genome size is relatively large compared to model plant species such as *Arabidopsis thaliana* (156 Mbp/1C) [18], *Oryza sativa* (466 Mbp/1C) [19], *Carica papaya* (372 Mbp/1C) [20], *Vitis vinifera* (475Mbp/1C) [21], *Medicago truncatula* (465 Mbp/1C) [20],

Populus trichocarpa (484 Mbp/1C) [22], *Gossypium raimondii* (880 Mbp/1C) [23], *Solanum lycopersicon* (900 Mbp/1C) [24] and even *Zea mays* (2,665 Mbp/1C) [19]. Therefore, despite advances in genome sequencing technologies and analysis techniques, it remains comparatively expensive to sequence the whole genome of *Epimedium*. Furthermore, genome size in plants is usually positively correlated with an increasing content of repetitive DNA, mainly transposable elements (TEs) [25], making *de novo* assembly of next generation sequence reads particularly difficult for large genomes. Instead, for investigating large genomes, the GeneTrek approach, first used in analyzing the pufferfish (*Fugu*) genome, is considered feasible and cost-effective [26]. This approach works by first randomly sequencing and annotating a small subsection of the genome and then extrapolating this information to estimate the characteristics of the genome in general [27]. This has been widely adopted to provide the first insights into genomes such as maize [27–29], rye [30], cotton [31], papaya [32], bread wheat [33], chickpea [34], woodland strawberry [35], and carrot [36], as well as the medicinal gymnosperm *Taxus mairei* [37].

Here we report the first structural analysis of the *E. sagittatum* genome using the GeneTrek approach combined with chromosome analysis and FISH of *Gypsy-Ty3* retrotransposable elements. The information gained in this study lays the foundations for future research into *Epimedium* genome organization, species evolution and also the functional analysis of genes involved in the complex and valuable metabolism of *Epimedium* species.

2. Results

2.1. Sequence Assembly and Composition

Sequencing of the 672 randomly selected clones generated 1075 assemblies (126 contigs, 949 singlets) with an average length of 674.43 bp, a total length of 725,008 bp and a GC% content of up to 39.00% (Table 1, Figure 1). All sequences were deposited into GenBank (Accession numbers: JY266095-JY267380). Out of the 1075 assemblies, 1023 contained nuclear DNA sequences, while 50 assemblies comprised organellar (chloroplast and mitochondria) DNA (Table 1). After excluding organellar and bacterial DNA sequences, the dataset of 1023 assemblies comprising 690,804 bp of nuclear DNA sequence was obtained and designated as the ENS (*E. sagittatum* nuclear sequence) dataset. This represents 0.0154% of the entire nuclear genome based on the approximated 1C-value of 4496 Mbp. Within the ENS dataset, at least 795 assemblies contained repetitive DNA elements with a calculated total length of 541,625 bp, representing 78.41% of the ENS dataset. The total length of the "confirmed annotated" gene sequences was 17,360 bp (2.51% of the ENS dataset) while the total length of "putative" gene sequences was 50,466 bp (7.3%). No sequence homology was detected for 130 assemblies using the analyses herein, so these were referred to as "unknown" (Table 2, Figure 2A).

	554111011451
Total number of assemblies	1075
Contigs	126
Singlets	949
Total sequence length	725,008 bp

Table	1.	Summary	of the	assemblies
-------	----	---------	--------	------------

Total number of assemblies	1075
Average insert length	674.43 bp
GC content	39.00%
Nuclear DNA insert	1023
Organellic DNA insert (cp, mt) ^a	50
Bacterial DNA insert	2

Table 1. Cont.

^a cp, chloroplast; mt, mitochondrion.

Figure 1. Length distribution of insert sequences of plasmid clones. Sequencing of the 672 randomly selected clones generated 1075 assemblies (126 contigs, 949 singlets). The x-axis indicates the length interval of insert sequences. The y-axis indicates the number of assemblies at each length interval.



Length intervals (bp)

Table 2. Details of the sequence composition of nuclear DNA sequences.

Nuclear DNA insert	Number of assemblies	Calculated sequence length (bp)
Confirmed annotated gene sequences	23	17,360
Putative gene sequences	75	50,466
Repetitive DNA elements	795	541,625
Unknown elements	130	81,353
Total ENS dataset	1023	690,804

^a ENS, *E. sagittatum* Nuclear Sequences.





2.2. Repetitive DNA Elements of E. sagittatum Genome

In total 451,355 bp (65.34% of the ENS dataset) of sequence showed significant homology to the repeats in Repbase or the TE-related proteins in either the NR protein database or the Plant Repeat

Database. These were categorized as "total TE repeats". In addition to the total TE repeats, ribosomal RNA genes (31,002 bp, 4.49%), microsatellite repeats (21,065 bp, 3.05%), telomeric sequences (10,249 bp, 1.48%), and centromeric sequences (11,105 bp, 1.61%), were identified, giving a total of 541,625 bp from 795 assemblies identified as repetitive DNA elements, occupying 78.41% of the ENS dataset (Table 3, Figure 2A). By extrapolation, this suggests that the total repetitive DNA content in the *E. sagittatum* genome may be as high as 78.41%.

	Number of assemblies	Calculated length of repetitive DNA (bp)	Percentage of total repetitive DNA (%)
Total TE ^a repeats	656	451,355	83.33
ClassI retrotransposons	431	295,036	54.47
Class II DNA transposons	171	119,536	22.07
MITE ^b	54	36,783	6.79
Ribosomal RNA genes	47	31,002	5.72
Microsatellite repeats	34	21,065	3.89
Telomeric sequences	16	10,249	1.89
Centromeric sequences	17	11,105	2.05
Other repeat	25	16,849	3.11
Total repetitive DNA elements	795	541,625	100

Table 3. Summary of repetitive DNA elements.

^a TE, Transposable Element; ^b MITE, Miniature Inverted-repeat Transposable Element.

	Number of	Calculated length of	Percentage of TE
	assemblies	TE repeats (bp)	repeats (%)
ClassIretrotransposons	431	295,036	65.37
LTR ^a / <i>Gypsy-Ty3</i>	211	142,593	31.59
Copia-Ty1	136	95,698	21.20
Other LTR	3	2,176	0.48
Non-LTR LINE ^b / <i>RTE</i>	26	17,700	3.92
LINE/L1	16	11,985	2.66
Other LINE	13	8261	1.83
SINE ^c	6	3840	0.85
Other retrotransposon	20	12,783	2.83
ClassII DNA transposons	171	119,536	26.48
CACTA	110	78,494	17.39
hAT	21	15,805	3.50
Mutator	6	3760	0.83
Other DNA transposons	34	21,477	4.76
MITE	54	36,783	8.15
Total TE repeats	656	451,355	100

Table 4. Detailed classification and coomposition of TE repeats.

^a LTR, Long Terminal Repeat; ^b LINE, Long Interspersed Nuclear Element; ^c SINE, Short Interspersed Nuclear Element.

TE repeats (Class I retrotransposons, ClassII DNA transposons and MITEs) accounted for 83.33% of the entire repetitive DNA content of *E. sagittatum*, (Table 3, Figure 2B). Of these, retrotransposons (Class I) were the dominant class, accounting for 65.37% of the total TE repeats, which could be further subdivided into *Gypsy-Ty3* LTR retrotransposons (31.59%), *Copia-Ty1* (21.20%), other LTRs (0.48%), Non-LTR/LINE 8.41%, Non-LTR/SINE 0.85% and other retrotransposons 2.83%. DNA transposons (Class II) totaled 119,536 bp, accounting for 26.48% of the total TE repeats, with DNA/*CACTA*, DNA/*hAT*, and DNA/*Mutator* the most abundant (Table 4, Figure 2C).

2.3. Gene Content Analysis and Gene Number Estimation

Gene prediction using BlastX, FGENESH, and EST database support identified 23 "confirmed annotated" gene sequences (2.51% of the ENS dataset) and 75 "putative" gene sequences (7.31% of the ENS dataset) (Table 2). Most of the captured gene sequences were related to plant primary metabolism including an amino acid carrier, a secondary cell wall-related glycosyltransferase, beta-fructofuranosidase and 3-isopropylmalate dehydratase. The combined coding length of the 23 "confirmed annotated" gene sequences was 9275 bp, and was used to extrapolate the predicted protein-coding portion of the *E. sagittatum* nuclear genome. Based on the assumption that the average gene length is 2 kb, similar to Arabidopsis [38], the 9275 bp sequences implies a total of 60.23 Mbp protein-coding sequence within the entire nuclear genome, and a total gene number of 30,114 was predicted.

2.4. Chromosomal Number Analysis of E. sagittatum

Chromosome counts of metaphase *E. sagittatum* cells revealed a chromosome number of 2n = 2x = 12 (Figure 3). This result is consistent with the observations of somatic chromosome number in additional *Epimedium* taxa by Zhang *et al.* (2008) [16].

Figure 3. Root-tip mitotic metaphase chromosomes of *E. sagittatum* 2n = 2x = 12 are shown. Scale bar = 5 µm.



2.5. Chromosomal Distribution of Gypsy-Ty3 Retrotransposons in E. sagittatum

Given the prevalence of Gypsy-Ty3-type sequences in the genome sequence obtained for *E. sagittatum*, FISH to pachytene chromosomes of *E. sagittatum* was conducted using probes derived from Gypsy-Ty3 reverse transcriptase (*rt*), to examine the distribution of major retrotransposable elements in the *E. sagittatum* nuclear genome. The FISH analysis revealed that Gypsy-Ty3 retrotransposons are distributed unevenly on all chromosomes, and some strong signals were detected in the telomeric domains of several chromosomes (Figure 4). The extent of hybridisation suggested that Gypsy-Ty3 retrotransposons are prolific in the *E. sagittatum* genome.

Figure 4. FISH analysis of *Gypsy-Ty3* retrotransposons distributed on pachytene chromosomes of *E. sagittatum*; (a) DAPI stained chromosomes; (b) Dispersed signal; (c) Merged images. Scale bar = $10 \mu m$.



3. Discussion

3.1. Characteristics of E. sagittatum Genome

E. sagittatum is a member of the Berberidaceae family of basal eudicot plant species recorded in the Chinese Pharmacopoeia as one of five *Epimedium* species with valuable medicinal properties. This study used the GenTrek approach, combined with chromosome and FISH analyses, to extrapolate and reveal novel insights into the genome sequence, structure and composition of *E. sagittatum*. This provides the first detailed genomic analysis of this evolutionarily and agriculturally important medicinal plant species.

In total, over 69,0804 bp of nuclear DNA, designated the ENS dataset, was sequenced from 1075 sequence contigs, representing 0.0154% of the *E. sagittatum* nuclear genome. A total of 23 sequences, representing 2.51% of the ENS dataset, were annotated as confirmed protein-coding sequences, while a significantly larger portion of the genome, comprising 78.41% of ENS dataset, was identified as repetitive DNA. Of this, the majority of repetitive DNA comprised transposable elements, and specifically Class I retrotransposons. This is consistent with previous investigations showing that retrotransposons, whose amplification has been the major cause of genome expansion [39,40]. By randomly sequencing and annotating a small fraction of the whole genome, the GeneTrek approach has

provided the first insight into the detailed genomic makeup of *Epimedium* spp, laying the foundations for the further elucidation of genome organization.

3.2. Genome Size Variation of Epimedium Species

The chromosomal number of all *Epimedium* species examined to date is 2n = 2x = 12 [16]. Consistent with this, our chromosome analysis revealed 12 chromosomes (2n = 2x = 12) for *E. sagittatum*. Despite this relatively small chromosome number, *Epimedium* species have intermediately-sized genomes compared to other angiosperms [41], with haploid genome sizes ranging from 4115.35 Mbp to 4876.4 Mbp [17]. According to the plant DNA C-values database [42], angiosperm genomes vary over 2,400-fold in size, from 63 Mbp/1C in *Genlisea aurea* [43] to 148,852 Mbp/1C in *Paris japonica* [44]. Within the Berberidaceae, the largest genome currently recorded is 14,351 Mbp, for *Podophyllum emodi* [45], while other Berberidaceae species vary from 489 Mbp in *Berberis koreana* [46] to 4005 Mbp in *Epimedium alpinum* [47]. Intriguingly, significant genome size variation in *E. sagittatum* was observed between different geographical populations [17]. This is consistent with the belief that intraspecific genome size variation is associated with adaption to different growing conditions or habitat [48].

3.3. Repetitive DNA and Retrotransposable Elements

As mentioned above, 78.41% of the randomly sampled nuclear genomic sequences for E. sagittatum were identified as repetitive DNA. Therefore, by extrapolation it is suggested that at least 78.41% of the whole nuclear genome of E. sagittatum is repetitive DNA. Given that E. sagittatum is a relatively under-studied plant species, it is likely that this is an underestimation due to insufficient annotations of potentially novel species-specific repeats. Previous research demonstrated that repetitive DNA elements are important components of eukaryotic genomes and play a significant role in genome size variation and genome evolution. For example, it is well established that the differences in genome size observed in the plant kingdom are accompanied by variations in the amount of repetitive DNA. In maize for example, TE families accounted for 70% of the genome size variation between the cultivated B73 inbred line and its related species Zea luxurians [49]. More specifically, this has been attributed to the amplification and deletion of LTR retrotransposons, suggesting that these elements are important players in the evolution of plant genome size and polyploidy [40,50–52]. In B73 maize, TEs constitute over 85% of the genome, with 75% of this attributed to LTR retrotransposons [53]. LTR retrotransposons also account for the majority of the nuclear genome in many other plant species [30,31,33], including species with relatively small genomes and proportionally less repetitive DNA [32,35,54]. Consistent with these findings, LTR retrotransposons comprised the majority of repetitive DNA in the sampled E. sagittatum genome. Of these, Gypsy-Ty3 retrotransposons were the dominant type, being almost two times more abundant than Copia-Tv1 type sequences (Table 4, Figure 2C). In cotton and rice, the ratio of Gypsy-Ty3 to Copia-Ty1 retrotransposons was reported to be around 2:1 [31,55], while it was found to be around 1:1 in Arabidopsis [38] and maize [28]. Therefore, it appears that the ratio of Gypsy-Ty3 to Copia-Ty1 retrotransposons is similar in E. sagittatum to cotton and rice.

In tomato BACs [56] and the genus *Helianthus* and other Asteraceae [57], *Gypsy-Ty3* has been shown to preferentially localize to centromeric or pericentromeric chromosome regions. However, *Copia-Ty1* retrotransposons were distributed unevenly and mostly at the precentromeric and terminal heterochromatin regions [17]. The different content and localization of *Gypsy-Ty3* and *Copia-Ty1* retrotransposons in plant genomes suggests that these two retrotransposons effect genome size variation and genome evolution independently and may have distinct roles in the evolutionary history of genome expansion. Analysis of the diversity of *copia*-RT fragments in *Epimedium* revealed relatively low *copia*-RT sequence heterogeneity, suggesting that *Copia-Ty1* retrotransposons experienced bursts of activation followed by deactivation during *Epimedium* genome size [17]. This study has contributed our understanding of the TE composition of the *Epimedium* genome, and the potential roles of various TEs in genome size evolution, however more in-depth efforts are needed to explore retrotransposon behavior in detail and the impact of these TEs on plant genome size, composition and evolution.

3.4. Gene Content Analysis

Extrapolation of the number and cumulative length of "confirmed annotated" genic sequences from the sampled genomic data predicts the total gene number of *E. sagittatum* to be 30,114. Comparisons to species with known genome sizes and gene numbers suggest this is a reasonable estimate for *Epimedium* spp. For example, 27,411 genes are estimated in Arabidopsis (TAIR, version 10, Department of Plant Biology, Carnegie Institution: Stanford, CA, USA, available on http://www.arabidopsis.org/index.jsp), 39,045 genes in rice [58], 24,746 genes in papaya [59], 45,555 genes in black cottonwood [22], 30,434 genes in grapevine [60], 32,000 genes in maize[53], 34,809 genes in woodland strawberry[61], 40,976 genes in diploid cotton [62] and 34,727 genes in tomato [24]. Nonetheless, gene prediction is challenging in plants and other complex eukaryotic genomes and many transposable elements in plant BACs and genomes have been annotated as hypothetical genes, leading to a consistent over-prediction of gene number [63]. As such relatively stringent criteria were used to predict the total gene number in *E. sagittatum*.

The consistency in repetitive element and gene content estimates between *Epimedium* herein and published plant genomes supports suggestions that the GeneTrek approach is a cost-effective and efficient way to gain a global insight into relatively large plant genomes. Previous GeneTrek analysis of the maize genome predicted 37,000 genes [27] and at least 66% repetitive DNA [29]. Accordingly, the draft sequence of the whole maize genome predicted over 32,000 genes, and revealed that nearly 85% of the genome is composed of transposable elements [53]. In future, the sampling of more *E. sagittatum* genomic sequences may further narrow-down the range of estimated gene number for *E. sagittatum* [35]. Furthermore, based on the sequence data for the *E. sagittatum* genome provided herein, future work could focus on the mining of genes involved in secondary metabolite biosynthesis and examining their regulatory factors.

4. Experimental Section

4.1. Plant Material

E. sagittatum was grown in Spring (March–June) in the field at the Wuhan Botanical Garden, CAS, China. Several fully-expanded leaves were harvested for DNA preparation from 25 days old plants. Root tips of the same plants were collected for chromosome analysis and young panicles were harvested for FISH analysis.

4.2. DNA Preparation and Construction of the Shotgun Insertion Clones

DNA was isolated following Rabinowicz's protocol [64] and sonicated at 20% energy for 6 s and at 100% energy for 60 s in succession (Sonopuls GM 200, Bandelin, Berlin, Germany). Sonicated DNA was incubated with T4 DNA polymerase (Roche, Basel, Switzerland) and dNTPs (Takara, Dalian, China) at 16 °C for 18 h to create blunt-ended fragments. DNA fragments were size separated on a 1% TAE agarose gel and fragments ranging from 0.3 kb to 2 kb were purified (Qiagen) and ligated at 16 °C for 18 h into dephosphorylated, *Sma*I-digested pBluescriptKS (+) vector (National Center for Gene Research, CAS, Shanghai, China) using the T4 DNA ligation kit (Roche, Basel, Switzerland). Recombinant plasmids were electroporated into *Escherichia coli* DH10B competent cells (National Center for Gene Research, CAS, Shanghai, China). Positive transformants were selected on LB-agar plates containing ampicillin (100 μ g mL⁻¹), IPTG (60 μ g mL⁻¹) and X-gal (24 μ g mL⁻¹), then incubated in glycerol storage medium in 384-well plates for storage at -80 °C. In total, 1536 plasmid clones were constructed.

4.3. Sequencing and Sequence Assembly

Plasmid DNA isolation was performed as per Sambrook and Russell (2001) [65]. In total, 672 randomly selected clones were sequenced bidirectionally with SP6 (5'-ATT TAG GTG ACA CTA TAG-3') and T7 (5'-TAA TAC GAC TCA CTA TAG GG-3') primers using the BigDye[®] Terminator v3.1 Cycle Sequencing Kit (Applied BioSystems, Foster City, CA, USA) and an ABI 3730 capillary sequencer (National Center for Gene Research, CAS, Shanghai, China). Basecalling was performed with PHRED [66], sequence trimming and vector sequence removal were performed with TRIMMING [67] and SeqClean (http://compbio.dfci.harvard.edu/tgi/software/) using default settings, respectively. Sequence reads were assembled using CAP3 (Version Date: 12/21/07) [68]. The computational biology and functional genomics laborory

4.4. Sequence Annotation

4.4.1. Identification of Repetitive DNA Elements

Given the lack of a reference repeat database for *Epimedium*, repetitive DNA elements were identified using Protein-Based Repeat-Masking in RepeatMasker [69] and also using CENSOR [70]. In addition, sequences of TE elements or TE related proteins which may not be detected by RepeatMasker [69] and CENSOR [70] were identified by a default Blastn search against the Plant

Repeat Database at Michigan State University (http://plantrepeats.plantbiology.msu.edu/index.html) [71], and default BlastX (Expect value $\leq 10^{-10}$) searches against the National Center for Biotechnology Information non-redundant protein database [72]. The fraction of the genome that each class of repetitive DNA element represented was calculated as the ratio of the total length of repeat sequence to the total length of the *E. sagittatum* nuclear sequence (ENS) assembly, excluding the organellic sequences.

4.4.2. Analysis of Gene Content

Gene content analysis was performed in three stages. First, a default BlastX (Expect value $\leq 10^{-10}$) search was performed against the NCBI (http://www.ncbi.nlm.nih.gov) non-redundant protein database (NR) [72]. Secondly, *ab initio* gene prediction was performed on the ENS dataset using the FGENESH feature (Dicot plants-Arabidopsis) of the MolQuest software package (softberry) [73]. Thirdly, a local BLAST to an in-house EST database of *E. sagittatum* was executed (Expect value $\leq 10^{-10}$). All sequence results were manually examined and evaluated and any showing homology to TE elements, TE related proteins or organelle proteins in the NR protein database were omitted. Among the remaining sequences, those both with significant homology in the NR protein database (Expect value $\leq 10^{-10}$, identity $\geq 60\%$, alignment length of amino acids ≥ 67), and supported by either the in-house EST search or FGENESH prediction, were designated as "confirmed annotated" gene sequences. Those sequences with significant homology in the NR protein database but without the EST or FGENESH support, and *visa versa*, were designated as "putative" gene sequence.

4.5. Chromosomal Analysis

E. sagittatum root tips were pretreated in a saturated solution of paradichlorobenzene (PDB) for 4 h, then fixed in Carnoy's liquid fixative (ethanol:glacial acetic acid = 3:1) for 30 min at 4 °C before being macerated in 1*N*-hydrochloric acid for 6 min at 60 °C, stained with Carbol Fuchsin and then squashed for cytological observation.

4.6. FISH Analysis

Young panicles were fixed in Carnoy's liquid fixative (ethanol:glacial acetic acid = 3:1) for 30 min at 4 °C. Microsporocytes at the pachytene stage were macerated in a 1:1 mixture of pectinase:macerozyme for 30 min and then squashed in 45% glacial acetic acid. The slides were dipped immediately into liquid nitrogen and stored at -70 °C. The probes originated from the reverse transcriptase (*rt*) sequence of *Gypsy-Ty3*, and were amplified from *E. sagittatum* DNA using degenerate primers (forward primer: 5'-TAYCCNHTNCCNCGNATHGA-3'; reverse primer: 5'-ARCATRTCRTCNACRTA-3'). Probes from pooled *E. sagittatum Gypsy-Ty3 rt* PCR fragments were labeled by the DIG-Nick Translation Mix Kit (Roche). FISH was carried out as per Jiang *et al.* (1995) [74]. Sheep Anti-Digoxigenin (Roche) was used to detect the Digoxigenin-labeled probes, and amplified with FITC-conjugated Anti-sheep IgG (Vector Laboratories, Burlingame, CA, USA). Chromosomes were counterstained in 30 μ L 4',6-diamidino-phenylindole (2 μ g mL⁻¹; DAPI, Sigma, St. Louis, MO, USA) in antifade solution (Vector). Images were captured by an Olympus BX61

fluorescence microscope with a Photometrics SenSys CCD (charge coupled device) 1400E. Greyscale images were merged by Metamorph software (Version 7.0, Molecular Devices: Sunnyvale, CA, USA).

5. Conclusions

E. sagittatum is a Chinese traditional medicinal plant with great potential in the development of modern and natural pharmaceuticals. Thus further research into this species is warranted in order to elucidate the genomic and biochemical basis of its medicinal properties and exploit these properly. Here we provided the first insights into the composition, structure and evolution of the *E. sagittatum* genome. Our results indicate that at least 78.41% of the whole genome consists of repetitive DNA elements, with LTR retrotransposons dominating, and thus may play a significant role in *Epimedium* genome evolution. Furthermore, we provide a reasonable gene number estimation of 30,114 genes. This study will pave the way for further functional genomic analysis of this valuable medicinal plant with regards to genome organization, species evolution and the function of genes involved in its complex metabolism.

Acknowledgments

We gratefully acknowledge the National Center for Gene Research, CAS, and would like to thank Changbao Li and Biao Wang for their help in constructing the genomic shotgun clones and sequencing, as well as Alice Hayward and Yaowu Yuan for their helpful revision on this manuscript. This work was supported by the National Natural Science Foundation of China (30800624 and 31270340), CAS/SAFEA International Partnership Program for Creative Research Teams Project, Knowledge Innovation Project of Chinese Academy of Sciences [KSCX2-EW-J-20], and the South China Botanical Garden Startup Fund (201039).

Conflict of Interest

The authors declare no conflict of interest.

References

- 1. Ying, T. Petal evolution and distribution patterns of *Epimedium* L. (Berberidaceae). *Acta Phytotaxon. Sin.* **2002**, *40*, 481–489.
- 2. *Pharmacopoeia of the People's Republic of China*; Chinese Pharmacopoeia Commission: Beijing, China 2005.
- Liang, H.; Siren, H.; Jyske, P.; Reikkola, M.; Vuorela, P.; Vuorela, H.; Hiltunen, R. Characterization of flavonoids in extracts from four species of Epimedium by micellar electrokinetic capillary chromatography with diode-array detection. J. Chromatogr. Sci. 1997, 35, 117–125.
- 4. Hu, Z.W.; Shen, Z.Y.; Huang, J.H. Experimental study on effect of epimedium flavonoids in protecting telomere length of senescence cells HU. *Zhongguo Zhong Xi Yi Jie He Za Zhi* **2004**, *24*, 1094–1097.

- 5. Huang, J.H.; Shen, Z.Y.; Wu, B. Effect and mechanism of *Epimedium* flavanoids for aging retardation from viewpoint of transcriptomics and metabonomics. *Zhongguo Zhong Xi Yi Jie He Za Zhi* **2008**, *28*, 47–50.
- 6. Wu, H.; Lien, E.; Lien, L.; Medicines, C. Chemical and pharmacological investigations of *Epimedium* species: A survey. *Prog. Drug Res.* **2003**, *60*, 1–57.
- 7. Kim, Y.D.; Kim, S.H.; Landrum, L.R. Taxonomic and phytogeographic implications from ITS phylogeny in *Berberis* (Berberidaceae). *J. Plant Res.* **2004**, *117*, 175–182.
- 8. Shen, P.; Guo, B.L.; Gong, Y.; Hong, D.Y.; Hong, Y.; Yong, E.L. Taxonomic, genetic, chemical and estrogenic characteristics of *Epimedium* species. *Phytochemistry* **2007**, *68*, 1448–1458.
- Sun, Y.; Fung, K.P.; Leung, P.C.; Shaw, P.C. A phylogenetic analysis of *Epimedium* (Berberidaceae) based on nuclear ribosomal DNA sequences. *Mol. Phylogenet. Evol.* 2005, 35, 287–291.
- Zhang, M.L.; Uhink, C.H.; Kadereit, J.W. Phylogeny and biogeography of *Epimedium/Vancouveria* (Berberidaceae): Western North American-East Asian disjunctions, the origin of European mountain plant taxa, and East Asian species diversity. *Syst. Bot.* 2007, *32*, 81–92.
- 11. Sun, Y.; Fung, K.; Leung, P.; Shi, D.; Shaw, P. Characterization of medicinal *Epimedium* species by 5S rRNA gene spacer sequencing. *Planta Med.* **2004**, *70*, 287–288.
- Zeng, S.; Xiao, G.; Guo, J.; Fei, Z.; Xu, Y.; Roe, B.; Wang, Y. Development of a EST dataset and characterization of EST-SSRs in a traditional Chinese medicinal plant, *Epimedium sagittatum* (Sieb. Et Zucc.) Maxim. *BMC Genomics* 2010, *11*, 94.
- Huang, W.; Sun, W.; Wang, Y. Isolation and molecular characterisation of flavonoid 3'-hydroxylase and flavonoid 3',5'-hydroxylase genes from a traditional Chinese medicinal plant, *Epimedium sagittatum. Gene* 2012, 497, 125–130.
- Huang, W.; Sun, W.; Lv, H.; Xiao, G.; Zeng, S.; Wang, Y. Isolation and molecular characterization of thirteen R2R3-MYB transcription factors from *Epimedium sagittatum*. *Int. J. Mol. Sci.* 2012, *14*, 594–610.
- Sun, W.; Huang, W.; Li, Z.; Lv, H.; Huang, H.; Wang, Y. Characterization of a Crabs Claw gene in basal eudicot species *Epimedium sagittatum* (Berberidaceae). *Int. J. Mol. Sci.* 2013, 14, 1119–1131.
- 16. Zhang, Y.; Dang, H.; Meng, A.; Li, J.; Li, X. Karyomorphology of *Epimedium* (Berberidaceae) and its phylogenetic implications. *Caryologia* **2008**, *61*, 283–293.
- 17. Chen, J.; Li, L.; Wang, Y. Diversity of genome size and Ty1-*copia* in *Epimedium* species used for traditional Chinese medicines. *HortScience* **2012**, *47*, 979–984.
- Bennett, M.; Leitch, I.; Price, H.; Johnston, J. Comparisons with *Caenorhabditis* (~100 Mb) and *Drosophila* (~175 Mb) using flow cytometry show genome size in Arabidopsis to be~157 Mb and thus~25% larger than the Arabidopsis Genome Initiative estimate of ~125 Mb. *Ann. Bot.* 2003, *91*, 547–557.
- 19. Bennett, M.D.; Smith, J.B. Nuclear DNA amounts in Angiosperms. *Phil. Trans. Roy. Soc. (Lond.) B* 1991, *334*, 309–345.
- 20. Arumuganathan, K.; Earle, E.D. Nuclear DNA content of some important plant species. *Plant Mol. Biol. Rep.* **1991**, *9*, 208–218.

- Adam-Blondon, A.; Bernole, A.; Faes, G.; Lamoureux, D.; Pateyron, S.; Grando, M.; Caboche, M.; Velasco, R.; Chalhoub, B. Construction and characterization of BAC libraries from major grapevine cultivars. *Theor. Appl. Genet.* 2005, *110*, 1363–1371.
- Tuskan, G.; Difazio, S.; Jansson, S.; Bohlmann, J.; Grigoriev, I.; Hellsten, U.; Putnam, N.; Ralph, S.; Rombauts, S.; Salamov, A. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* 2006, *313*, 1596–1604.
- 23. Hendrix, B.; Stewart, J.M. Estimation of the nuclear DNA content of *Gossypium* species. *Ann. Bot.* 2005, *95*, 789–797.
- 24. The Tomato Genome Consortium. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* **2012**, *485*, 635–641.
- 25. Flavell, R.B.; Bennett, M.D.; Smith, J.B.; Smith, D.B. Genome size and the proportion of repeated nucleotide sequence DNA in plants. *Biochem. Genet.* **1974**, *12*, 257–269.
- Brenner, S.; Elgar, G.; Sandford, R.; Macrae, A.; Venkatesh, B.; Aparicio, S. Characterization of the pufferfish (*Fugu*) genome as a compact model vertebrate genome. *Nature* 1993, 366, 265–268.
- 27. Liu, R.; Vitte, C.; Ma, J.; Mahama, A.A.; Dhliwayo, T.; Lee, M.; Bennetzen, J.L. A GeneTrek analysis of the maize genome. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 11844–11849.
- Messing, J.; Bharti, A.K.; Karlowski, W.M.; Gundlach, H.; Kim, H.R.; Yu, Y.; Wei, F.; Fuks, G.; Soderlund, C.A.; Mayer, K.F.; *et al.* Sequence composition and genome organization of maize. *Proc. Natl. Acad. Sci. USA* 2004, *101*, 14349–14354.
- Haberer, G.; Young, S.; Bharti, A.K.; Gundlach, H.; Raymond, C.; Fuks, G.; Butler, E.; Wing, R.A.; Rounsley, S.; Birren, B.; *et al.* Structure and architecture of the maize genome. *Plant Physiol.* 2005, *139*, 1612–1624.
- Bartos, J.; Paux, E.; Kofler, R.; Havrankova, M.; Kopecky, D.; Suchankova, P.; Safar, J.; Simkova, H.; Town, C.D.; Lelley, T.; *et al.* A first survey of the rye (*Secale cereale*) genome composition through BAC end sequencing of the short arm of chromosome 1R. *BMC Plant Biol.* 2008, *8*, 95–106.
- 31. Guo, W.; Cai, C.; Wang, C.; Zhao, L.; Wang, L.; Zhang, T. A preliminary analysis of genome structure and composition in *Gossypium hirsutum*. *BMC Genomics* **2008**, *9*, 314–331.
- Lai, C.W.; Yu, Q.; Hou, S.; Skelton, R.L.; Jones, M.R.; Lewis, K.L.; Murray, J.; Eustice, M.; Guan, P.; Agbayani, R.; *et al.* Analysis of papaya BAC end sequences reveals first insights into the organization of a fruit tree genome. *Mol. Genet. Genomics* 2006, 276, 1–12.
- Devos, K.; Ma, J.; Pontaroli, A.; Pratt, L.; Bennetzen, J. Analysis and mapping of randomly chosen bacterial artificial chromosome clones from hexaploid bread wheat. *Proc. Natl. Acad. Sci.* USA 2005, 102, 19243–19248.
- Rajesh, P.N.; O'Bleness, M.; Roe, B.A.; Muehlbauer, F.J. Analysis of genome organization, composition and microsynteny using 500 kb BAC sequences in chickpea. *Theor. Appl. Genet.* 2008, 117, 449–458.
- Pontaroli, A.C.; Rogers, R.L.; Zhang, Q.; Shields, M.E.; Davis, T.M.; Folta, K.M.; SanMiguel, P.; Bennetzen, J.L. Gene content and distribution in the nuclear genome of *Fragaria vesca*. *Plant Gen.* 2009, *2*, 93–101.

- Cavagnaro, P.; Chung, S.M.; Szklarczyk, M.; Grzebelus, D.; Senalik, D.; Atkins, A.; Simon, P. Characterization of a deep-coverage carrot (*Daucus carota* L.) BAC library and initial analysis of BAC-end sequences. *Mol. Genet. Genomics* 2009, 281, 273–288.
- 37. Hao, D.; Yang, L.; Xiao, P. The first insight into the *Taxus* genome via fosmid library construction and end sequencing. *Mol. Genet. Genomics* **2011**, *285*, 197–205.
- 38. Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **2000**, *408*, 796–815.
- SanMiguel, P.; Tikhonov, A.; Jin, Y.K.; Motchoulskaia, N.; Zakharov, D.; Melake-Berhan, A.; Springer, P.S.; Edwards, K.J.; Lee, M.; Avramova, Z.; *et al.* Nested retrotransposons in the intergenic regions of the maize genome. *Science* 1996, *274*, 765–768.
- Piegu, B.; Guyot, R.; Picault, N.; Roulin, A.; Saniyal, A.; Kim, H.; Collura, K.; Brar, D.S.; Jackson, S.; Wing, R.A.; *et al.* Doubling genome size without polyploidization: Dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. *Genome Res.* 2006, 16, 1262–1269.
- 41. Soltis, D.E.; Soltis, P.S.; Bennett, M.D.; Leitch, I.J. Evolution of genome size in the angiosperms. *Am. J. Bot.* **2003**, *90*, 1596–1603.
- 42. Bennett, M.; Leitch, I. *Angiosperm DNA C-Values Database*, version 8.0. Available online: http://www.kew.org/cvalues/ (accessed on 7 May 2013).
- Greilhuber, J.; Borsch, T.; Müller, K.; Worberg, A.; Porembski, S.; Barthlott, W. Smallest angiosperm genomes found in *Lentibulariaceae*, with chromosomes of bacterial size. *Plant Biol.* 2006, *8*, 770–777.
- Pellicer, J.; Fay, M.F.; Leitch, I.J. The largest eukaryotic genome of them all? *Bot. J. Linnean Soc.* 2010, *164*, 10–15.
- 45. Veselý, P.; Bureš, P.; Šmarda, P.; Pavlíček, T. Genome size and DNA base composition of geophytes: the mirror of phenology and ecology? *Ann. Bot.* **2012**, *109*, 65–75.
- Olszewska, M.J.; Osiecka, R. The relationship between 2 C DNA content, systematic position, and the level of nuclear DNA endoreplication during differentiation of root parenchyma in some dicotyledonous shrubs and trees. comparison with *Herbaceous* species. *Biochem. Physiol. Pflanzen.* 1984, 179, 641–657.
- Siljak-Yakovlev, S.; Pustahija, F.; Oli, M.E.; Boguni, F.; Muratovi, E.; Ba, N.; Catrice, O.; Brown, S.C. Towards a genome size and chromosome number database of balkan flora: C-values in 343 Taxa with novel values for 242. *Adv. Sci. Lett.* 2010, *3*, 190–213.
- 48. Ohri, D. Genome size variation and plant systematics. Ann. Bot. 1998, 82, 75-83.
- 49. Tenaillon, M.I.; Hufford, M.B.; Gaut, B.S.; Ross-Ibarra, J. Genome size and transposable element content as determined by high-throughput sequencing in Maize and *Zea luxurians*. *Genome Biol. Evol.* **2011**, *3*, 219–229.
- 50. Vitte, C.; Panaud, O. LTR retrotransposons and flowering plant genome size: Emergence of the increase/decrease model. *Cytogenet. Genome Res.* **2005**, *110*, 91–107.
- 51. Vitte, C.; Bennetzen, J. Analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 17638–17643.
- 52. Bennetzen, J.; Ma, J.; Devos, K. Mechanisms of recent genome size variation in flowering plants. *Ann. Bot.* **2005**, *95*, 127–132.

- Schnable, P.S.; Ware, D.; Fulton, R.S.; Stein, J.C.; Wei, F.; Pasternak, S.; Liang, C.; Zhang, J.; Fulton, L.; Graves, T.A.; *et al.* The B73 maize genome: Complexity, diversity, and dynamics. *Science* 2009, *326*, 1112–1115.
- 54. Zuccolo, A.; Sebastian, A.; Talag, J.; Yu, Y.; Kim, H.; Collura, K.; Kudrna, D.; Wing, R. Transposable element distribution, abundance and role in genome size variation in the genus *Oryza*. *BMC Evol. Biol.* **2007**, *7*, 152–166.
- 55. International Rice Genome Sequencing Project. The map-based sequence of the rice genome. *Nature* **2005**, *436*, 793–800.
- 56. Wang, Y.; Tang, X.; Cheng, Z.; Mueller, L.; Giovannoni, J.; Tanksley, S. Euchromatin and pericentromeric heterochromatin: comparative composition in the tomato genome. *Genetics* **2006**, *172*, 2529–2540.
- Natali, L.; Santini, S.; Giordani, T.; Minelli, S.; Maestrini, P.; Cionini, P.; Cavallini, A. Distribution of Ty3-gypsy-and Ty1-copia-like DNA sequences in the genus Helianthus and other Asteraceae. *Genome* 2006, 49, 64–72.
- Kawahara, Y.; de la Bastide, M.; Hamilton, J.; Kanamori, H.; McCombie, W.; Ouyang, S.; Schwartz, D.; Tanaka, T.; Wu, J.; Zhou, S.; *et al.* Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. *Rice* 2013, *6*, 1–10.
- Ming, R.; Hou, S.; Feng, Y.; Yu, Q.; Dionne-Laporte, A.; Albert, H.; Suzuki, J.; Tripathi, S.; Moore, P.; Gonsalves, D. The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature* 2008, 452, 991–996.
- Jaillon, O.; Aury, J.; Noel, B.; Policriti, A.; Clepet, C.; Casagrande, A.; Choisne, N.; Aubourg, S.; Vitulo, N.; Jubin, C. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 2007, 449, 463–467.
- Shulaev, V.; Sargent, D.J.; Crowhurst, R.N.; Mockler, T.C.; Folkerts, O.; Delcher, A.L.; Jaiswal, P.; Mockaitis, K.; Liston, A.; Mane, S.P.; *et al.* The genome of woodland strawberry (*Fragaria vesca*). *Nat. Genet.*2011, *43*, 109–116.
- 62. Wang, K.; Wang, Z.; Li, F.; Ye, W.; Wang, J.; Song, G.; Yue, Z.; Cong, L.; Shang, H.; Zhu, S.; *et al.* The draft genome of a diploid cotton Gossypium raimondii. *Nat. Genet.* **2012**, *44*, 1098–1103.
- 63. Bennetzen, J.L.; Coleman, C.; Liu, R.; Ma, J.; Ramakrishna, W. Consistent over-estimation of gene number in complex plant genomes. *Curr. Opin. Plant Biol.* **2004**, *7*, 732–736.
- Rabinowicz, P. Constructing gene-enriched plant genomic libraries using methylation filtration technology. In *Methods in Molecular Biology: Plant Functional Genomics*; Grotewold, E., Ed.; Humana Press: Totowa, NJ, USA, 2003; Volume 236, pp. 21–36.
- 65. Sambrook, J.; Russell, D.W. *Molecular Cloning: A Laboratory Manual*; Cold Spring Harbor Laboratory press: New York, NY, USA, 2001.
- 66. Ewing, B.; Hillier, L.; Wendl, M.; Green, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **1998**, 8, 175–185.
- 67. Chou, H.H.; Holmes, M.H. DNA sequence quality trimming and vector removal. *Bioinformatics* **2001**, *17*, 1093.
- 68. Huang, X.; Madan, A. CAP3: A DNA sequence assembly program. Genome Res. 1999, 9, 868-877.
- 69. *RepeatMasker*, Version 4.0.2. Available online: http://www.repeatmasker.org/cgi-bin/ WEBRepeatMasker (accessed on 8 May 2013).

- 70. Kohany, O.; Gentles, A.; Hankus, L.; Jurka, J. Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinforma*. **2006**, *7*, 474–480.
- 71. Ouyang, S.; Buell, C.R. The TIGR Plant Repeat Databases: A collective resource for the identification of repetitive sequences in plants. *Nucleic Acids Res.* **2004**, *32*, D360–D363.
- 72. Altschul, S.; Madden, T.; Schaffer, A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389–3402.
- 73. *MolQuest*, Version 2.1.1. Available online: http://www.molquest.com/molquest.phtml (accessed on 25 March 2010).
- Jiang, J.; Gill, B.; Wang, G.; Ronald, P.; Ward, D. Metaphase and interphase fluorescence *in situ* hybridization mapping of the rice genome with bacterial artificial chromosomes. *Proc. Natl. Acad. Sci. USA* 1995, *92*, 4487–4491.

 \bigcirc 2013 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/3.0/).