Developing SSR Individual Identification System and Tracking the Geographical Origin of Timber for the Endangered Species *Chamaecyparis Formosensis*

Chiun-Jr Huang (d04625001@ntu.edu.tw)
National Taiwan University
Fang-Hua Chu
National Taiwan University
Yi-Shiang Huang
Institute of Biological Chemistry, Academia Sinica
Yu-Ching Tu
Investigation Bureau, Ministry of Justice
Yu-Mei Hung
Investigation Bureau, Ministry of Justice
Yu-Hsin Tseng
Biodiversity Research Center, Academia Sinica
Chang-En Pu
Investigation Bureau, Ministry of Justice
Cheng Te Hsu
Hualien Forest District Office, Forestry Bureau, Council of Agriculture
Chi-Hsiang Chao
Investigation Bureau, Ministry of Justice
Yu-Shyang Chou
Investigation Bureau, Ministry of Justice
Shau-Chian Liu
National Taitung University
Ya Ting You
Biodiversity Research Center, Academia Sinica
Shuo-Yu Hsu
National Taiwan University
Hsiang-Chih Hsieh
National Taiwan University
Chieh-Ting Wang
National Taiwan University
Chi-Tsong Chen
Investigation Bureau, Ministry of Justice

Research Article

Keywords: species, Taiwan, court, evidence, Mahogany tree, European ash, Merbau timber, Brazilian rosewood

Posted Date: May 13th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-493686/v1

Abstract

Chamaecyparis formosensis is an endemic species of Taiwan, threatened from intensive use and illegal felling. An individual identification system for *C. formosensis* is required to provide scientific evidence for court use and deter illegal felling. In this study, 36 polymorphic simple sequence repeat (SSR) markers were developed. By applying up to 28 non-linkage of the developed markers, it is calculated that the cumulative random probability of identity (CP_I) is as low as 1.652×10^{-12}, the combined power of discrimination (*CPD*) is as high as 0.999999999998348, and the identifiable population size is up to 60 million, which is greater than the known *C. formosensis* population size in Taiwan. Biogeographical analysis data show that *C. formosensis* from four geographic areas belong to the same genetic population which can be further divided into three clusters: SY (Eastern Taiwan), HV and GW (Northwestern Taiwan), and MM (Southwestern Taiwan). The developed system was applied to assess the provenance of samples with 88.44% of accuracy rate, and therefore can serve as a pre-screening tool to reduce the range required for comparison. The system developed in this study is a potential crime-fighting tool against illegal felling.

Introduction

Chamaecyparis formosensis Masam., also known as False Cypress, is endemic tree species in Taiwan distributed majorly in the cloud forest, a zonal forest type in the mid elevation (1,700-2,600 m) with extremely high biodiversity[1]. Gigantic *C. formosensis* is known for its superb timber quality. Due to its high quality and market value (4,050 USD/m^3,woodprice.forest.gov.tw), *C. formosensis* is critically threatened by illegal felling. Unplanned felling and poor management had compromised the ecosystem and led these precious species endangered. Similar scenarios also happened to Brazilian rosewood[2], Dalbergia Timber[3], European ash[4], Mahogany tree[5], and Merbau timber[6]. Moreover, timber production countries suffer from illegal felling, particularly in South-east-Asian, African and South American countries[7]. Although suspects were arrested in some occasions, lack of direct scientific evidence to link seized timber and stump had led fail of conviction in majority[8].

In recent years, countries suffering from serious illegal logging have successively begun to develop DNA-based timber individual identification systems that can provide court evidence for conviction. The technologies can be further categorized into SSR (Simple sequence repeat)-based systems including *Chamaecyparis taiwanensis*[8], *Fraxinus excelsior*[4], *Cinnamomum kanehirae*[9], *Entandrophragma cylindricum*[10] and *Intsia palembanica*[6] and SNP (Single nucleotide polymorphism) and INDEL (Insertion/ Deletion)-based system including *Acer macrophyllum*[11].

The SSR individual identification technique was developed more than 30 years and have been widely used in DNA paternity testing, forensic examination, victim identification, and animal individual identification[12,13]. SSR marker is a co-dominant and highly reproducible DNA marker with the following characteristics: it has a high degree of polymorphism; it is abundant and evenly distributed in eukaryotic genome; most of them are not functional, and can be easily and economically tested by PCR (polymerase chain reaction); last, the length is generally short which provides a higher opportunity to be amplified when applied to the lysed sample[13,14]. Therefore, SSR is the most commonly used method for individual identification systems[15,16]. In order to protect our precious cypress resources in Taiwan, we have developed and adapted several polymorphic SSR markers[17] for individual identification.

In the illegal felling crime case reports *C. taiwanensis*[8] and *F. excelsior*[4], it is demonstrated that SSR individual identification system can provide scientific evidences which are considered acceptable by court. In these cases, the individual identification system developed with genetic markers for those species were used to link seized timbers and victim trees, while considering the random probability of the same genotype appearing in the population. Therefore, convincible scientific proof with confidence level close to 100% were accepted as court evidences for crime conviction.

The legality of wood products usually depends on their source[18]. SSR is also often used in genetic diversity and population structure analysis of species[14,19], and can further predict the geographic provenance and distribution of species. Genetic methods have been applied to confirm the source and trade routes of protected species[20,21]. However, filing the DNA of every individual is not feasible even if *C. formosensis* has been listed as an endangered species. Therefore, it is important to analyze the genetic variation and population structure of *C. formosensis*. Genotyping can reveal the provenance of the timber and greatly reduces the range of possible plant sources.

The main purpose of this study is to develop *C. formosensis* SSR individual identification system, which provides high discrimination power against genetic variation, in order to prevent the occurrence of illegal felling. In addition, the biogeographical analysis of *C. formosensis* also facilitate to provide provenance information of the seized timber. In addition to being a long-developed individual identification tool[15,16], SSR is also often used in plant conservation and breeding[22,23]. The SSR markers developed in this study can also support the future *C. formosensis* afforestation by providing the selection basis for mother trees with higher diversity.

Materials And Methods

Development of new SSR markers for *C. formosensis*

In order to develop SSR markers for individual identification, we constructed three DNA libraries (Fig. 1 and Supplementary 1). Three *C. formosensis* individuals from QL (Voucher no. *Chung 4450*) and SY (Voucher no. *Chung 4905, 4906*) were used for DNA library preparation. To build three DNA libraries, genomic DNA was extracted from fresh leaves using the VIOGENE plant DNA extraction kit (VIOGENE, New Taipei City, Taiwan). The DNA libraries were sequenced using the Illumin MiSeq System (2 × 301 bp paired-end; Illumin, San Diego, California, USA) at Tri-I Biotech (New Taipei City, Taiwan).

Bioinformatics analysis was conducted with CLC Genomics Workbench version 10 (QIAGENE, Aarhus, Denmark). The raw reads were prescreened to remove adapter sequences and reads with greater than 0.01 error or with an average quality less than QV20. The trimmed sequences were further subjected to de novo assembly (Fig. 1).

MISA (*MI*cro*SA*tellite v 1.0)[27] was applied to screen the SSR containing sequences from contigs. To design SSR primers, sequences with at least five di-, tri-, tetra-, penta-, and hexa-nucleotide repeats were selected using BatchPrimer3[28], with optimized conditions set length at 18-23 bp, melting temperature 45-62°C, and a product size of 80-300 bp (Fig. 1).

A total of 100 candidate SSR primer pairs were newly designed in this study. These markers were subjected to validation test on 92 samples from four *C. formosenses* geographic areas (MM, HV, GW, SY see Fig.2 and Supplementary 1). The samples DNA used in marker validation were extracted using the VIOGENE plant DNA extraction kit (VIOGENE, New Taipei City, Taiwan). The PCR reaction was conducted with a final volume 20 µL containing 2 ng of genomic DNA, 0.25 µL of 10 µM each primer and 10 µL of Q-Amp 2x Screening Fire Taq Master Mix (Bio-Genesis Technologies, Taipei, Taiwan). The following PCR process was conducted: an initial denaturation of 95°C for 2 min; 30 cycles of 95°C for 45 sec, a primer-specific annealing temperature for 45 sec, and 72°C for 45 sec; followed by a 15-min extension at 72°C (Table 1). The amplified products were evaluated on the ABI 3500 (Applied Biosystems, Waltham, Massachusetts, USA) with GeneScan 600 LIZ Size Standard (Applied Biosystems). Fragment size was determined by using GeneMapper ID-X v1.6 (Applied Biosystems).

The the cross-species transferability of the designed markers were tested in *Chamaecyparis taiwanensis*, for details see Supplementary 2 and 3.

Developing *C. formosensis* individual identification system

Marker analysis was conducted by combining 27 pairs published SSR markers[17] with 9 validated SSR markers abovementioned. GenAlex 6.51b2[44] was used to calculate number of alleles (A), observed heterozygosity (*Ho*), expected heterozygosity (*He*), Hardy-Weinberg equilibrium (HWE). PowerMarker V3.25[51] was used to calculate polymorphism information content or power of information content (*PIC*)[52]. Power of discrimination (*PD*)[35], $PD = 1 - \Sigma P_i^2$, where P_i is the frequency of genotype i. Probability of identity (P_I)[36], $P_I = 1 - PD$. Combined power of discrimination (*CPD*)[35], here we calculated *CPD* of 28 markers. $CPD = 1 - [(1 - PD_1)(1 - PD_2)...(1 - PD_{28})]$.Combined probability of identity (CP_I)[36]. Microsoft Excel (Microsoft Office 2016) was used to calculate *PD*, P_I, *CPD*, CP_I. GENEPOP 4.2[37] was used to test for linkage disequilibrium.

Population genetics analysis

Genetic differentiation and gene flow among 4 geographic areas (MM, HV, GW, SY) were analyzed using F_{st} and Nm by Arlequin 3.5.2.2[53].

The population genetic structure was analyzed using STRUCTURE 2.3.4[46]. The program was run for K = 1 to 5 clusters with 20 independent runs to assess simulation stability. Each simulation was run for an initial 1,000,000 burn-in period followed by 100,000 replications based on the Markov chain Monte Carlo (MCMC)[54]. The best grouping was evaluated by Delta K[54] in Structure Harvester Web v0.6.94[55]. Bar graphs were generated by CLUMPP 1.1.2[56] for K ideal.

Individual provenance simulation was conducted with GENECLASS v. 2.0[50] on each 92 individuals independently. Pairwise simulation was also conducted on the pooled database deducted the sample itself.

Result And Discussion

Development of new SSR markers for *C. formosensis*

For a higher accuracy of court's judgment on illegal felling, it is necessary to establish a complete forensic system. Although some SSR markers of cypress have been published[17,24-26], the reported detection rates for dried timber were only 20-40%. Therefore, it is necessary to develop more SSR markers as a contingency plan. When the sample is in a poor condition, more markers can be applied in order to achieve the threshold of combined power of discrimination (*CPD*) required for successful comparison between seized timbers and victim trees. In order to maximize potential loci, Next generation sequence (NGS) methods were used. In this study 3 DNA library were constructed. We used the Illumina MiSeq platform (2 × 301 bp; Illumina, San Diego, California, USA) to sequence the DNA libraries (Fig. 1 and Supplementary 1).

A total of 70,325,072 raw reads were produced. The raw reads were deposited in the NCBI BioProject (PRJNA454510). After quality-trimming to the raw reads with CLC Genomics Workbench version 10 (QIAGENE, Aarhus, Denmark), 70,319,509 contigs were generated with length between 133 and 146 bp in average. De novo assembly was then conducted with the following parameters: contig number 208,467, minimum length of contigs 18bp, maximum length contigs 108,928bp, and average length contigs 491bp. The sequence was assembled with software CLC Genomics Workbench version 10 and the length of assembly sequence was 102,281,642 bp.

A sum of 78,250 SSR containing sequences were screened by MISA (v 1.0, *Mi*cro*SA*tellite)[27]. We newly designed 100 candidate SSR primer pairs for testing in *C. formosensis* by BatchPrimer3[28].

There are 9 validated SSR markers are polymorphic (success rate 9.00%), which were registered in GenBank in NCBI (Table 1) and passed for cross-species tests (supplementary 2 and 3).

Different from the traditional SSR cloning method, with next generation sequencing technology, it is easy to obtain colossal amount of SSR containing sequences from sequenced genomes[29]. However, transforming candidate SSR primer pairs into validated SSR markers is still a time-consuming and expensive step. Qualified SSR markers need to succeed in PCR amplification, have good peak pattern quality with little stuttering, and be free of non-amplifying (invalid) alleles. The turnover rate from candidate SSR primer pairs to validated SSR markers varies from species to species[29,30]. The success rate in *Chamaecyparis* plants is between 5.24% and 9.27%[8,17,24,25,31].

Developing *C. formosensis* individual identification system

In this study, newly developed 9 validated SSR markers and other 27 validated SSR markers[17] polymorphic SSR markers were analyzed against 92 individuals from 4 geographic areas (MM, HV, GW, SY, Fig. 2 and Supplementary 1). The results of developed 36 SSR markers are summarized in Table 2. Among the 92 individuals in this study, each number of alleles of SSR is between 2 and 27, with average of 7.916. The levels of observed heterozygosity (*Ho*)[32] are from 0.000 to 0.891, with average of 0.414. The levels of expected heterozygosity (*He*)[32] are ranged from 0.103 to 0.906, with average of 0.565. Significant ($P < 0.001$) deviations of Hardy-Weinberg equilibrium (HWE)[33] were detected in 23 SSR loci: Cred47, 225, 231, 236, 242, 248, 249, 250, 253, 260, 262, 276, 277, 280, 603, 610, 628, 640, 641, 674, 678, 682, 683. *Ho* is the actual proportion of heterozygous individuals contained in each locus within the population, whereas the *He* is the expected value estimated per HWE. *Ho* and *He* are the most popular parameters used in estimating genetic diversity in population. The population structural and even historical information can be obtained from *Ho* and *He*. Most marker within the 36 tested ones have higher *He* than *Ho* (all except Cred211, Cred220, Cred225, Cred248, Cred276, Cred281, Cred297,

Cred298), suggesting that different loci have lower genetic diversity in *C. formosensis* and the population of *C. formosensis* is an inbred. HWE describes that under ideal conditions, gene frequency does not change over time or generation. However, there will always be one or more interfering factors affecting gene frequency in nature. Therefore, HWE is difficult to achieve in nature. In this study, 23 loci out of 36 markers deviated from HWE (63.89% deviation rate). The reason for this deviation could be artificial selection, non-panmixia or genetic drift.

Polymorphism information content or power of information content (*PIC*). *PIC* value is an index of relative ability of the SSR marker's genetic variability. The higher the polymorphism of marker's genotype, the higher the *PIC* value[34]. Polymorphic markers were highly informative (*PIC* > 0.50), reasonably informative (0.50 > *PIC* > 0.25), and slightly informative (*PIC* < 0.25). Power of discrimination (*PD*)[35] refers to the ability of genetic markers to distinguish individuals within a population. Obviously, in a population with more allele types and evenly distributed genotypes, the low probability of two random individuals have the same genotype, and the greater probability of two random individuals can be identified by the system. Probability of identity (P_I)[36] is the probability of two individuals with the same genotype. $PD = 1\text{-}P_I$. The value of *PIC*, *PD* and P_I of individual markers reflects its identification ability in the individual identification system. The greater *PIC* and *PD*, the lower P_I in value, suggesting the higher identification ability of the marker, and vice versa. The levels of *PIC* range from 0.097 to 0.876, with average 0.528. The levels of *PD* range from 0.102 to 0.885, with average 0.567. The levels of P_I range from 0.114 to 0.897, with average 0.431. There were 19 out of 36 markers with *PIC* greater than 0.5, and the mean of these 36 markers *PIC* values was greater than 0.5, suggesting the markers have a high identification ability. The results of *PD* and P_I correspond to those of *PIC*. The highly informative markers presented in *PIC* also show higher identification ability in *PD* and P_I.

Significant linkages ($P < 0.001$) were detected among Cred35/229/277, Cred47/298, Cred231/249/253/262, Cred281/297, Cred603/683 and Cred640/678/682 with GENEPOP 4.2[37], suggesting the abovementioned group located in the same linkage group (Table 2). When identifying several independent polymorphic genetic markers at the same time (polymorphic markers located in different linkage groups), the combined probability of identity (CP_I) is the product of the P_I of each genetic marker. At this time, CP_I will be greatly reduced, and combined power of discrimination (*CPD*) will become very high. As defined above, $CP_I + CPD = 1$. The credibility of the individual identification system is calculated based on "Random match probability in population size and confidence levels"' published by Budowle et al (2000)[38]. Confidence levels (CL) = $(1 - CP_I)^N$ where N is number of individuals.

The individual identification system was applied to illegal felling cases[8]. When the seized timber and the victim tree are identified to be the same individual plant, under the considerations of fairness and objective, the court usually adopts 99.99%, 99% or 95% confidence level as the credibility standard[39] (Wall 2002, ISO ISO/IEC 17025). In this study, the locus with the lowest P_I within a linkage group was used to calculate the CP_I (Table 3). The CP_I decreased along with the accumulation of loci, and with the P_I of each locus sorted in ascending order. The system can accumulate up to 28 loci without linkage. When reaching its maximum capability, even under most strict standard (confidence level 99.99%) by dictated by court, this system can be used to identify 60 million *C. formosensis*, with CP_I as low as 1.652×10^{-12}, *CPD* as high 0.999999999998348 (almost equal to 1), beyond the known population size 32.06 ± 3.20 million of *C. formosensis*[40]. Under ideal condition, a minimal of 6 loci can be applied to the system, with an identifiable *C. formosensis* population of 2,900 under 95% confidence level. The CP_I is as low as 1.728×10^{-5}, and *CPD* is as high as 0.999982712603209 (Table 3).

Population genetics analysis

The fixation index (Fst) is a measure of population differentiation due to genetic structure[41]. A higher Fst value means a higher degree of difference between populations. When F_{st} is less than 0.05, there is no differentiation among populations. When F_{st} is between 0.05 and 0.15, there is low differentiation among populations. On the other hand, estimation of number of migrants (Nm) is a measure of gene flow value[41]. When Nm is greater than one, it represents genes frequently exchange which counteracts the genetic drift and prevents the population differentiation[42]. When Nm is greater than 4, it would be a random mating population[43]. Fst and Nm analysis on the 4 geographic areas were conducted with GeneAlex 6.503[44] (Table 4). The Fst value of HV and GW is 0.035, suggesting no inter-population differentiation. The highest Fst value 0.074 was found between HV and MM. The Fst values range from 0.056 to 0.065 were found between the rest geographic areas, indicating a low differentiation among geographic areas. The highest Nm value 6.832 was found between HV and GW, whereas the lowest value 3.141 between HV and MM. The Nm values of 4

geographic areas were all greater than 1 (between 3.141 and 6.832), suggesting a frequent gene exchange between the 4 geographic areas, which offsets genetic drift and prevents population differentiation. For GW/MM (Nm = 4.022), GW/HV (Nm = 6.382), the Nm values of the population are greater than 4, suggesting that they are random mating populations.

STRUCTURE analysis[45,46] was used to analyze the population genetic structure of *C. formosensis* (Fig. 3), and the Delta K value was calculated to obtain the optimal number of clusters. K and Delta K were show in Fig. 3a and K=2, 3 and 4 are drawn respectively in Fig. 3b. These individuals of *C. formosensis* present the most likely three clusters: The SY located in Eastern Taiwan is an independent cluster, the MM located in Southwestern Taiwan is another standalone cluster, whereas the two HV and GW geographic areas are in the same genetic cluster. Combing the results of Fst (Table 4), Nm (Table 4) and STRUCTURE analysis (Fig. 3), the data show that the *C. formosensis* of the four geographical areas belongs to the same genetic population. The Fst and STRUCTURE analysis data suggests that the samples fall into three clusters. The hypothesis that Taiwan Island is one of the plant refuges during the Quaternary glaciation[47,48] may help to explain the results. In the research on the historical biogeography and phylogeny of cypress[48], *C. formosensis* in Taiwan Island was separated from *Chamaecyparis* in Japan at 2.9 million years ago. The arrival of the Quaternary glaciation caused the extinction of many species and led a continued retreat of species to lower latitudes, thus Taiwan Island became a refuge for many ancient species. After the glaciation, species spread from the refuge to the surrounding area and formed species diversity on a latitude gradient[49]. In our results, the low polymorphism of *C. formosensis* in the four geographic areas' individuals may be from the same big population of *C. formosensis* during the glaciation. After the glacial retreat, *C. formosensis* spread out from the refuge. Due to the influence of terrain and distance, *C. formosensis* in different geographical areas has a tendency to form different clusters.

GENECLASS v. 2.0[50] was applied to analyse the provenance of 92 individuals independently and the probability of samples returning to the correct provenance is 95.00% (MM), 88.00% (HV), 69.57% (GW), and 100.00% (SY), with an overall mean correct rate of 88.04% (Table 5). Three HV individuals were mis-assigned to GW and four GW individuals were mis-assigned to HV, corresponding to the observation that HV and GW are the same cluster. However, 3 GW individuals were mis-assigned to MM, possibly because the geographic location of GW is between HV and MM, and therefore GW has characteristics of north and south at the same time. Likelihood, one MM was mis-assigned to GW, further supporting the inference that there is partial gene exchange between MM and GW. Overall, our data shows that the populations in eastern (SY) and western Taiwan (the rest populations) have distinct differences in genotype. Within the western populations, the northern (HV) and the southern one (MM) have obvious differences. Therefore, when seizing timbers in the futures, the genotype can be served as a prefilter to infer the geographic area of the victim tree if the provenance is found to be MM, HV or SY. A further inspection is required if the provenance is found to be GW because of the existence of gene exchange between nearby geographic area.

Conclusions

In this study, a *C. formosensis* individual identification system was built with 36 polymorphic SSR markers. When 28 non-linkage SSR markers are applied, the system is capable to identify 60 million *C. formosensis* individuals with confidence level of 99.99%. The lowest CP_I is 1.652×10^{-12} and the highest *CPD* is 0.999999999998348. This system can provide the scientific evidence to link seized timbers and victim trees required the illegal felling court cases, and facilitate the future legal sales by profiling timbers. Through population genetics analysis, the system is capable to provide provenance information, which would significantly enhance the efficiency by reducing the range required for investigation. The polymorphic markers developed in this study can be further applied to the conservation and breeding of the precious species *C. formosensis*.

Declarations

Acknowledgements

The authors would like to thank Dr. Kuo-Fang Chung (Biodiversity Research Center, Academia Sinica, Taiwan) for his support in research funding and facilities. The authors would like to thank Dr. Chaolun Allen Chen and Dr. Shu-Miaw Chaw (Biodiversity Research Center, Academia Sinica, Taiwan) for their support in research facilities. This work was financially supported by Ministry of Science and Technology, Taiwan (grant no. MOST 104-2321-B-002-056) and Ministry of Justice, Taiwan (grant no. 109-1301-05-17-02).

Author contributions

C.J.H. conceived, designed and conducted the experiments, and wrote the main manuscript text, drew the figures and tables, sample collection, funding application and manuscript submitted. F.H.C. edited the manuscript. Y.S.H. edited the manuscript and assisted to draw figures and tables. Y.C.T. and Y.M.H. performed the experiments. Y.H.T. edited the manuscript. C.E.P. data analysis. C.T.H. drew the figures and sample collection. C.H.C. edited the manuscript. Y.S.C. data analysis. S.C.L., Y.T.Y., S.Y.H. and H.C.H. sample collection and performed the experiments. C.T.W. sample collection. C.T.C. funding application. All authors reviewed the manuscript.

Author statement

In regard to legislation compliance of experimental materials, we hereby declare that all of our experimental research and field studies on plants, either cultivated or wild, including the collection of plant material, comply with relevant institutional, national, and international guidelines and legislation.

References

1. Hwang, S. Y., Lin, H. W., Kuo, Y. S. & Lin, T. P. RAPD variation in relation to population differentiation of Chamaecyparis formosensis and Chamaecyparis taiwanensis. *Botanical Bulletin of Academia Sinica.* **42**, 173–179 (2001).
2. Kite, G. C. *et al.* Dalnigrin, a neoflavonoid marker for the identification of Brazilian rosewood (Dalbergia nigra) in CITES enforcement. *Phytochemistry.* **71**, 1122–1131 https://doi.org/10.1016/j.phytochem.2010.04.011 (2010).
3. Espinoza, E. O., Wiemann, M. C., Barajas-Morales, J., Chavarria, G. D. & McClure, P. J. Forensic analysis of CITES-protected Dalbergia timber from the Americas. *IAWA journal.* **36**, 311–325 (2015).
4. Tereba, A., Woodward, S., Konecka, A., Borys, M. & Nowakowska, J. A. Analysis of DNA profiles of ash (Fraxinus excelsior L.) to provide evidence of illegal logging. *Wood Science and Technology.* **51**, 1377–1387 https://doi.org/10.1007/s00226-017-0942-5 (2017).
5. Cabral, E. C. *et al.* Wood typification by Venturi easy ambient sonic spray ionization mass spectrometry: the case of the endangered Mahogany tree. *J Mass Spectrom.* **47**, 1–6 https://doi.org/10.1002/jms.2016 (2012).
6. Lowe, A. J., Wong, K. N., Tiong, Y. S., Iyerh, S. & Chew, F. T. A DNA Method to verify the integrity of timber supply chains; confirming the legal sourcing of merbau timber from logging concession to sawmill. *Silvae Genetica.* **59**, 263–268 https://doi.org/10.1515/sg-2010-0037 (2010).
7. Dormontt, E. E. *et al.* Forensic timber identification: It's time to integrate disciplines to combat illegal logging. *Biol. Conserv.* **191**, 790–798 https://doi.org/10.1016/j.biocon.2015.06.038 (2015).
8. Huang, C. J. *et al.* Development and technical application of SSR-based individual identification system for Chamaecyparis taiwanensis against illegal logging convictions. *Scientific reports.* **10**, 1–14 (2020).
9. Hung, K. H., Lin, C. H. & Ju, L. P. Tracking the geographical origin of timber by DNA fingerprinting: a study of the endangered species Cinnamomum kanehirae in Taiwan. *Holzforschung.* **71**, 853–862 (2017).
10. Jolivet, C. & Degen, B. Use of DNA fingerprints to control the origin of sapelli timber (Entandrophragma cylindricum) at the forest concession level in Cameroon. *Forensic Sci Int Genet.* **6**, 487–493 (2012).
11. Dormontt, E. *et al.* Forensic validation of a SNP and INDEL panel for individualisation of timber from bigleaf maple (Acer macrophyllum Pursch). *Forensic Science International: Genetics.* **46**, 102252 (2020).
12. Jeffreys, A. J., Wilson, V. & Thein, S. L. Individual-specific 'fingerprints' of human DNA. *Nature.* **316**, 76–79 (1985).
13. Robinson, A. J., Love, C. G., Batley, J., Barker, G. & Edwards, D. Simple sequence repeat marker loci discovery using SSR primer. *Bioinformatics.* **20**, 1475–1476 (2004).
14. Ali, A. *et al.* Genetic diversity and population structure analysis of Saccharum and Erianthus genera using microsatellite (SSR) markers. *Sci Rep.* **9**, 395 https://doi.org/10.1038/s41598-018-36630-7 (2019).
15. Jobling, M. A. & Gill, P. Encoded evidence: DNA in forensic analysis. *Nat Rev Genet.* **5**, 739–751 https://doi.org/10.1038/nrg1455 (2004).
16. Butler, J. M. Genetics and genomics of core short tandem repeat loci used in human identity testing. *Journal of forensic sciences.* **51**, 253–265 (2006).

17. Huang, C. J. *et al.* Isolation and characterization of SSR and EST-SSR loci in Chamaecyparis formosensis (Cupressaceae). *Applications in plant sciences.* **6**, e01175 https://doi.org/10.1002/aps3.1175 (2018).
18. Finch, K. N. *et al.* Predicting the geographic origin of Spanish Cedar (Cedrela odorata L.) based on DNA variation. *Conserv. Genet.* **21**, 625–639 (2020).
19. Dorji, J., Tamang, S., Tshewang, T., Dorji, T. & Dorji, T. Y. Genetic diversity and population structure of three traditional horse breeds of Bhutan based on 29 DNA microsatellite markers. *PloS one.* **13**, e0199376 (2018).
20. Paredes-Villanueva, K. *et al.* Nuclear and plastid SNP markers for tracing Cedrela timber in the tropics.*Conservation Genetics Resources*,1–6(2019).
21. Blanc-Jolivet, C., Yanbaev, Y., Kersten, B. & Degen, B. A set of SNP markers for timber tracking of Larix spp. in Europe and Russia. *Forestry: An International Journal of Forest Research.* **91**, 614–628 (2018).
22. Penha, J. *et al.* Estimation of natural outcrossing rate and genetic diversity in Lima bean (Phaseolus lunatus L. var. lunatus) from Brazil using SSR markers: implications for conservation and breeding. *Genetic Resources and Crop Evolution.* **64**, 1355–1364 (2017).
23. Yang, H., Zhang, R., Jin, G., Feng, Z. & Zhou, Z. Assessing the genetic diversity and genealogical reconstruction of cypress (Cupressus funebris Endl.) breeding parents using SSR markers. *Forests.* **7**, 160 (2016).
24. Matsumoto, A. *et al.* Development and polymorphisms of microsatellite markers for hinoki (Chamaecyparis obtusa). *Mol. Ecol. Notes.* **6**, 310–312 https://doi.org/10.1111/j.1471-8286.2006.01212.x (2006).
25. Nakao, Y., Iwata, H., Matsumoto, A., Tsumura, Y. & Tomaru, N. Highly polymorphic microsatellite markers in Chamaecyparis obtusa. *Canadian journal of forest research.* **31**, 2248–2251 https://doi.org/10.1139/cjfr-31-12-2248 (2001).
26. Kim, Y. M., Shin, Y. S. & Jeong, J. H. Development and characterization of microsatellite primers for Chamaecyparis obtusa (Cupressaceae). *Applications in plant sciences.* **4**, 1500136 (2016).
27. Thiel, T., Michalek, W., Varshney, R. & Graner, A. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare L.). *Theoretical and applied genetics.* **106**, 411–422 (2003).
28. You, F. M. *et al.* BatchPrimer3: a high throughput web application for PCR and sequencing primer design. *BMC Bioinformatics.* **9**, 253 https://doi.org/10.1186/1471-2105-9-253 (2008).
29. Zalapa, J. E. *et al.* Using next-generation sequencing approaches to isolate simple sequence repeat (SSR) loci in the plant sciences. *American journal of botany.* **99**, 193–208 (2012).
30. Gardner, M. G., Fitch, A. J., Bertozzi, T. & Lowe, A. J. Rise of the machines–recommendations for ecologists when using next generation sequencing for microsatellite development. *Molecular Ecology Resources.* **11**, 1093–1101 (2011).
31. Iwaizumi, M., Watanabe, A. & Isoda, K. Primer Note: Development of Highly Polymorphic Nuclear Microsatellite Markers for Hinoki (Chamaecyparis obtusa). *Silvae Genetica.* **60**, 62–65 (2011).
32. Nei, M. Analysis of gene diversity in subdivided populations. Proceedings of the National Academy of Sciences 70, 3321–3323(1973).
33. Rodriguez, S., Gaunt, T. R. & Day, I. N. Hardy-Weinberg equilibrium testing of biological ascertainment for Mendelian randomization studies. *American journal of epidemiology.* **169**, 505–514 (2009).
34. Pan, Y. B. Highly polymorphic microsatellite DNA markers for sugarcane germplasm evaluation and variety identity testing. *Sugar Tech.* **8**, 246–256 (2006).
35. Fisher, R. Standard calculations for evaluating a blood-group system. *Heredity.* **5**, 95 (1951).
36. Jones, D. A. & Blood Samples Probability of Discrimination. *Journal of the Forensic Science Society.* **12**, 355–359 https://doi.org/10.1016/s0015-7368(72)70695-7 (1972).
37. Raymond, M. & Rousset, F. GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. *Journal of Heredity.* **86**, 248–249 (1995).
38. Budowle, B., Chakraborty, R., Carmody, G. & Monson, K. L. Source attribution of a forensic DNA profile. *Forensic Science Communications.* **2**, 6 (2000).
39. Wall, W. *Genetics & DNA technology: legal aspects* (Routledge-Cavendish, 2002).
40. Qiu, L. W., Huang, Q. X., Wu, C. C. & Hsieh, H. T. (Taipei, 2015).
41. Wright, S. Isolation by distance. *Genetics.* **28**, 114 (1943).

42. Wright, S. Evolution and the genetics of populations: Vol. 2. The theory of gene frequencies(1969).
43. Hartl, D. & Clark, A. Principles of population genetics. Sinauer Assoc. Inc, Sunderland, Massachusetts(1989).
44. Peakall, R. & Smouse, P. E. GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research—an update. *Bioinformatics.* **28**, 2537–2539 (2012).
45. Falush, D., Stephens, M. & Pritchard, J. K. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics.* **164**, 1567–1587 (2003).
46. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics.* **155**, 945–959 (2000).
47. Qian, H. & Ricklefs, R. E. Large-scale processes and the Asian bias in species diversity of temperate plants. *Nature.* **407**, 180–182 (2000).
48. Wang, W. P., Hwang, C. Y., Lin, T. P. & Hwang, S. Y. Historical biogeography and phylogenetic relationships of the genus Chamaecyparis (Cupressaceae) inferred from chloroplast DNA polymorphism. *Plant Systematics and Evolution.* **241**, 13–28 https://doi.org/10.1007/s00606-003-0031-0 (2003).
49. Qian, H. A comparison of the taxonomic richness of temperate plants in East Asia and North America. *American Journal of Botany.* **89**, 1818–1825 (2002).
50. Piry, S. *et al.* GENECLASS2: a software for genetic assignment and first-generation migrant detection. *Journal of heredity.* **95**, 536–539 (2004).
51. Liu, K. & Muse, S. V. PowerMarker: an integrated analysis environment for genetic marker analysis. *Bioinformatics.* **21**, 2128–2129 (2005).
52. Botstein, D., White, R. L., Skolnick, M. & Davis, R. W. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American journal of human genetics.* **32**, 314 (1980).
53. Excoffier, L., Laval, G. & Schneider, S. Arlequin (version 3.0): an integrated software package for population genetics data analysis.*Evolutionary bioinformatics*,47–50(2005).
54. Evanno, G., Regnaut, S. & Goudet, J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol Ecol.* **14**, 2611–2620 https://doi.org/10.1111/j.1365-294X.2005.02553.x (2005).
55. Earl, D. A. & vonHoldt, B. M. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources.* **4**, 359–361 https://doi.org/10.1007/s12686-011-9548-7 (2011).
56. Jakobsson, M. & Rosenberg, N. A. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics.* **23**, 1801–1806 https://doi.org/10.1093/bioinformatics/btm233 (2007).

Tables

Table 1. Characteristics of 9 SSR loci developed in *Chamaecyparis formosensis*.

Locus	Primer sequences (5′ -3′)	Repeat motif	Fluorescent label	Allele size (bp)	*Ta* (°C)	GenBank accession no.	Putative function [organism]
Cred603	TTGCTACATTAGCACTAGATAGCAAAGAAA	(AAG)13	6-FAM	106	60	MW052386	No hit
	ACTGAAGATACTGAGGATATTGAAGAGGAA						
Cred610	TGAGATATACATGTGTGAAAGAGAGTGAAGC	(GTAT)5	PET	166	60	MW052387	No hit
	TGCAATAATTTCTTCAGTGTTACCACTACC						
Cred628	GCTGGAGTCATTATAGTGCCATGTCTTTGT	(GCCC)3	6-FAM	142	60	MW052388	No hit
	TTTTCAAATAGCCGACCGACCTATGTAGAG						
Cred640	ACCCATATCTTCCTTCCCAACCATTAAGAT	(TCTT)5	6-FAM	137	60	MW052389	No hit
	CTTTCAGTGGAATGGAAGAAAGCCCTACTA						
Cred641	ACTTCTAATGAATCCCCATGCCGAATTGTA	(GC)19	VIC	193	60	MW052390	No hit
	CTGTTCGCGATAAGATAATTGGCTAGTGTG						
Cred674	TAAAGAGGCTCTGCTACTGGCTTTTCAACT	(GGGC)4	NED	147	60	MW052391	No hit
	GTGGGTGGCCCTCTATTCTATTGTTGAT						
Cred678	GGTCCATATCCTGGAGTAGAACCTCCCTAC	(GGGC)5	PET	162	60	MW052392	No hit
	GTGTCGCAGGCATAGACTTCTCCCTATATT						
Cred682	CCGCCCTTCTAATAACAGGGAAGATAAGTT	(CCCT)5	NED	147	60	MW052393	No hit
	CCGCCCTTCTAATAACAGGGAAGATAAGTT						
Cred683	GCAGCCTAAATAAACAATAGGGGGATTGAT	(GCCT)4	NED	146	60	MW052394	No hit
	CATGTTACGTATAGAATCGAGTGCAGGTCA						

Table 2. Genetic characterization of 36 polymorphic SSR loci of 92 *Chamaecyparis formosensis* individuals. *A* number of alleles, *Ho* observed heterozygosity, *He* expected heterozygosity, *PIC* polymorphism information content or power of information content, *PD* power of discrimination, P_I probability of identity, *PD* is equal to 1 – P_I. *Highly significant from Hardy-Weinberg equilibrium (P < 0.001). Significant linkage disequilibrium (P < 0.001) was detected in the same colored pairs.

Locus	*A*	*Ho*	*He*	*PIC*	*PD*	P_I
Cred35	6	0.500	0.525	0.481	0.524	0.475
Cred47	15	0.304*	0.653	0.639	0.653	0.346
Cred88	3	0.446	0.478	0.368	0.478	0.521
Cred211	4	0.674	0.598	0.530	0.594	0.405
Cred220	6	0.620	0.588	0.539	0.588	0.411
Cred224	10	0.728	0.811	0.785	0.811	0.188
Cred225	14	0.859*	0.844	0.829	0.844	0.155
Cred226	6	0.609	0.686	0.635	0.685	0.314
Cred229	8	0.511	0.531	0.499	0.531	0.468
Cred231	12	0.576*	0.826	0.807	0.825	0.174
Cred236	27	0.837*	0.906	0.857	0.868	0.131
Cred242	6	0.435*	0.668	0.615	0.668	0.331
Cred248	9	0.598*	0.555	0.533	0.555	0.444
Cred249	12	0.380*	0.771	0.743	0.771	0.228
Cred250	6	0.435*	0.527	0.486	0.527	0.472
Cred253	14	0.717*	0.851	0.836	0.850	0.149
Cred260	10	0.707*	0.711	0.679	0.711	0.288
Cred262	21	0.837*	0.885	0.876	0.885	0.114
Cred264	13	0.620	0.686	0.652	0.686	0.313
Cred276	3	0.891*	0.537	0.439	0.536	0.463
Cred277	10	0.272*	0.668	0.766	0.772	0.227
Cred280	5	0.250*	0.255	0.244	0.255	0.744
Cred281	3	0.326	0.300	0.276	0.300	0.699
Cred295	4	0.489	0.554	0.469	0.553	0.446
Cred297	2	0.109	0.103	0.097	0.102	0.897
Cred298	15	0.859	0.818	0.810	0.824	0.175
Cred299	5	0.304	0.359	0.314	0.359	0.640
Cred603	6	0.000*	0.356	0.338	0.356	0.643
Cred610	2	0.000*	0.488	0.368	0.487	0.512
Cred628	4	0.000*	0.180	0.170	0.180	0.819
Cred640	5	0.011*	0.597	0.520	0.596	0.403
Cred641	2	0.000*	0.141	0.130	0.140	0.859
Cred674	2	0.011*	0.168	0.153	0.167	0.832
Cred678	3	0.000*	0.520	0.460	0.519	0.480
Cred682	4	0.000*	0.494	0.415	0.494	0.505
Cred683	8	0.022*	0.721	0.680	0.720	0.279
average	7.916	0.414	0.565	0.528	0.567	0.431

Table 3. The discrimination power in SSR marker combination. CP_I cumulative random probability of identity, CL = $(1- CP_I)^N$, N number of individuals.

Loci#	CP_I	Confidence levels (CL)			Comment
		99.99%	99%	95%	
1	1.140×10^{-2}				
2	1.493×10^{-2}				
3	2.314×10^{-3}				
4	4.050×10^{-4}				
5	7.615×10^{-5}				Miniature identifiable population size
6	1.728×10^{-5}			2.90×10^{3}	
7	4.823×10^{-6}		2.00×10^{3}	1.00×10^{4}	Small identifiable population size
8	1.389×10^{-6}		7.20×10^{3}	3.60×10^{4}	
9	4.347×10^{-7}		2.30×10^{4}	1.10×10^{5}	Moderate identifiable population size
10	1.365×10^{-7}		7.30×10^{4}	3.70×10^{5}	
11	4.518×10^{-8}		2.20×10^{5}	1.10×10^{6}	Large identifiable population size
12	1.821×10^{-8}	5.40×10^{3}	5.50×10^{5}	2.80×10^{6}	
13	8.741×10^{-9}	1.10×10^{4}	1.10×10^{6}	5.80×10^{6}	
14	4.414×10^{-9}	2.20×10^{4}	2.20×10^{6}	1.10×10^{7}	Gigantic identifiable population size
15	1.787×10^{-9}	5.50×10^{4}	5.60×10^{6}	2.80×10^{7}	
16	7.347×10^{-10}	1.30×10^{5}	1.30×10^{7}	6.90×10^{7}	
17	3.262×10^{-10}	3.00×10^{5}	3.00×10^{7}	1.50×10^{8}	
18	1.455×10^{-10}	6.80×10^{5}	6.90×10^{7}	3.50×10^{8}	
19	6.736×10^{-11}	1.40×10^{6}	1.40×10^{8}	7.60×10^{8}	
20	3.179×10^{-11}	3.10×10^{6}	3.10×10^{8}	1.60×10^{9}	
21	1.628×10^{-11}	6.10×10^{6}	6.10×10^{8}	3.10×10^{9}	
22	8.482×10^{-12}	1.10×10^{7}	1.10×10^{9}	6.00×10^{9}	
23	5.428×10^{-12}	1.80×10^{7}	1.80×10^{9}	9.40×10^{9}	
24	3.794×10^{-12}	2.60×10^{7}	2.60×10^{9}	1.30×10^{10}	
25	2.823×10^{-12}	3.50×10^{7}	3.50×10^{9}	1.80×10^{10}	
26	2.312×10^{-12}	4.30×10^{7}	4.30×10^{9}	2.20×10^{10}	
27	1.923×10^{-12}	5.10×10^{7}	5.20×10^{9}	2.60×10^{10}	
28	1.652×10^{-12}	6.00×10^{7}	6.00×10^{9}	3.10×10^{10}	

Table 4. Pairwise F_{st} and Nm among four populations of *Chamaecyparis formosensis* using 36 simple sequence repeat (SSR) data. $F_{st} < 0.05$, no differentiation among populations. $0.05 < F_{st} < 0.15$, low differentiation among populations. The gene flow value (Nm), Nm > 1 represents genes frequently exchange which counteracts the genetic drift and prevents the population differentiation[42]. When Nm is greater than 4, it would be a random mating population[43].

F_{st} \ Nm	MM (*N*=20)	HV (*N*=25)	GW (*N*=23)	SY (*N*=24)
MM (*N*=20)	-	3.141	4.022	3.798
HV (*N*=25)	0.074	-	6.832	3.603
GW (*N*=23)	0.059	0.035	-	4.199
SY (*N*=24)	0.062	0.065	0.056	-

Table 5. *Chamaecyparis formosensis* individual provenance simulation result. A total of 92 samples composed of 20 MM, 25 HV, 23 GW, and 24 SY individuals were subjected to provenance simulation. Correct provenance is 95.00% (MM), 88.00% (HV), 69.57% (GW), and 100.00% (SY), with an overall mean correct rate of 88.04%.

	MM (*N*=20)	HV (*N*=25)	GW (*N*=23)	SY (*N*=24)	Over all (*N*=92)
Sum of correct samples	19	22	16	24	81
Correct cluster (%)	95.00	88.00	69.57	100.00	88.04

Figures

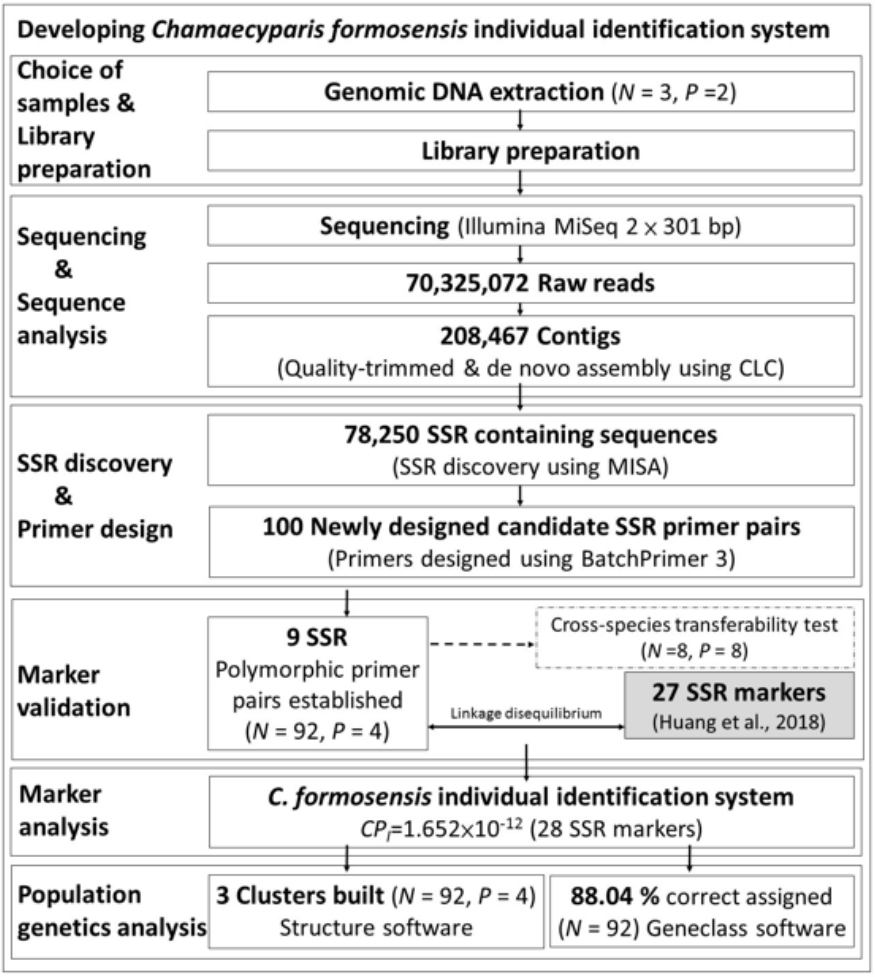

Figure 1

Flowchart of Chamaecyparis formosensis individual identification system development.

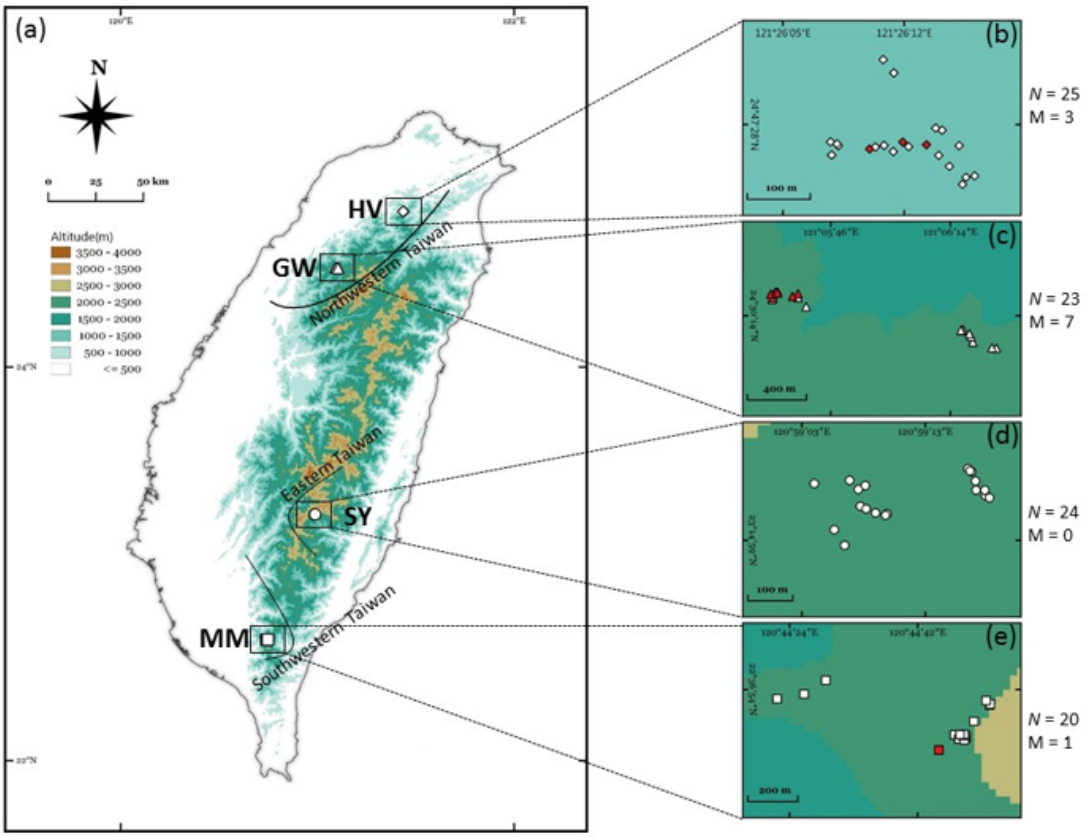

Figure 2

The biogeographic information of Chamacyparis formosensis in this study. A total of 92 samples composed of 20 MM, 25 HV, 23 GW, and 24 SY individuals were analyzed (a)Biogeographic analysis data suggests that the samples fall into three genetical categories: SY (Eastern Taiwan), HV & GW (Northwestern Taiwan), and MM (Southwestern Taiwan) (b)-(e) The red spots represent the individuals that have been mis-assigned (denoted as M in figure legend) from provenance simulation result. Note: The designations employed and the presentation of the material on this map do not imply the expression of any opinion whatsoever on the part of Research Square concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.

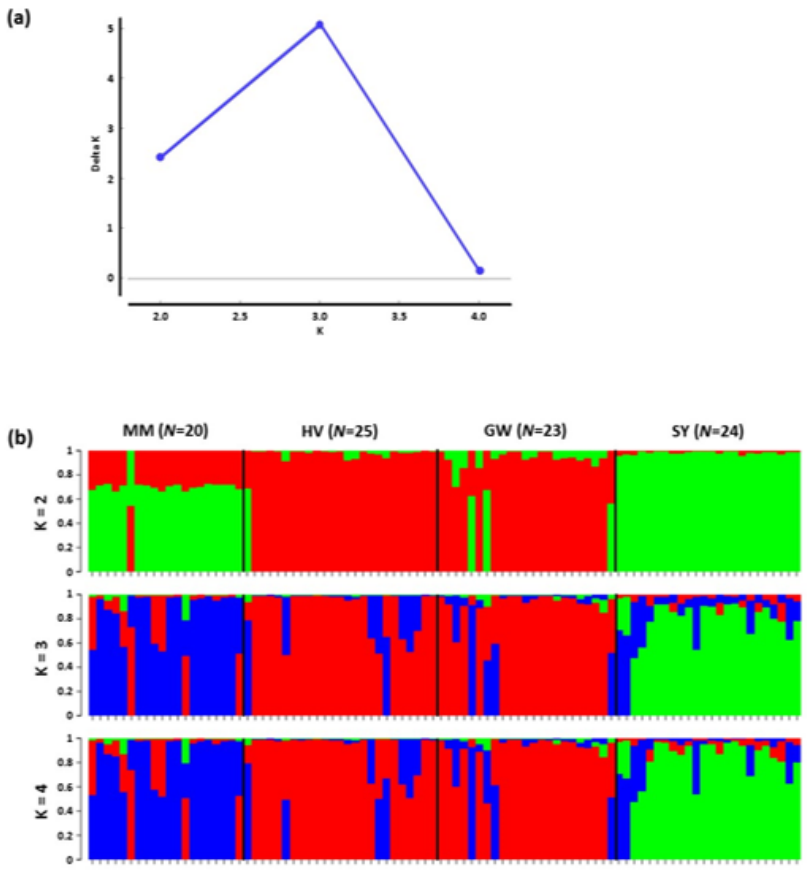

Figure 3

Genetic composition of Chamaecyparis formosensis. (a) The scatter plots of Delta K. (b) The 2, 3 and 4 clusters obtained from STRUCTURE analyses.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Individualidentificationsystemofredcypresssupplementary.docx