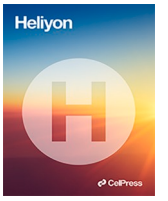
Heliyon 10 (2024) e29104

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

Next-generation DNA sequencing of Panax samples revealed new genotypes: Burrows-Wheeler Aligner, Python-based abundance and clustering analysis

Christopher Oberc, Paul C.H. Li *

Department of Chemistry, Simon Fraser University, 8888 University Drive, Burnaby, BC, V5A 1S6, Canada

A B S T R A C T

Background: There are two major species of the *Panax* genus, namely *Panax ginseng* and *Panax quinquefolius*. Other than the nucleic acid test and nucleic acid amplification test, DNA sequencing can be used to authenticate the species of ginseng samples, especially when their physical forms cannot be used for differentiation.

Method: In this work, next generation sequencing was used to obtain millions of reads from fourteen ginseng samples (root, powder, and granule). Then Gaussian Mixture clustering analysis was applied to analyze the reads from each sample.

Results and Discussion: A new genotype has been revealed in this study. Two samples have been authenticated with certainty, while the others may be hybrid in nature as revealed by the clustering results.

1. Introduction

There are various species of ginseng, for example *Panax ginseng* (Chinese/Korean ginseng) and *Panax quinquefolius* (American/Canadian ginseng). These species have different medicinal properties [1,2] as well as market values [3]. However, although these two species are still possible to differentiate from one another based on their morphological features, the species become very difficult to differentiate when made in commercial products in slices or powder forms. Genetic authentication may overcome these limitations.

There are many genetic methods available that use single nucleotide polymorphisms (SNPs), e.g., nucleic acid test (NAT) using probe-PCR product strand hybridization [4–8], qPCR [9–11], and lesion induced DNA amplification (LIDA) [12–14]. These techniques, which are effective when the DNA samples have known single nucleotide polymorphisms (SNPs), require considerable optimization. For instance, work by Zhou, H et al. on the dammarenediol-II synthase (DS) gene [3,15] has shown that the gene contains several SNP sites which vary between *P. ginseng* and *P. quinquefolius*. Inspired by this work, three known SNP sites have been confirmed for use in ginseng authentication using hybridization [5,6] and LIDA [14].

On the other hand, DNA sequencing does not require any prior knowledge of SNPs that exist within a sequence, but Sanger sequencing demands an additional step of cloning. Therefore, next generation sequencing (NGS) using the Illumina method has been employed to sequence the locus on the section of the DS gene which contains multiple SNP sites. In addition, NGS allows for the discovery of genetic variations (i.e. insertions and deletions) other than point mutations [16–20] and we confirm the findings of new genetic variations through 28 million reads in each ginseng sample. Recent advances in the elucidation of the *Panax* genome are also available [21–26].

* Corresponding author.
E-mail address: paulli@sfu.ca (P.C.H. Li).

https://doi.org/10.1016/j.heliyon.2024.e29104
Received 17 January 2024; Received in revised form 29 March 2024; Accepted 1 April 2024
Available online 7 April 2024

Table 1
The reference sequences of *P. ginseng* and *P. quinquefolius* [3,15]. The three SNP sites, which are coloured in red and assigned the names N1, N2 and N3, correspond to T, C and G respectively for *P. ginseng* or C, T and A respectively for *P. quinquefolius*.

P. ginseng Reference Sequence	TACAGTGAT AATTAAATAT TGTAACTATC TAAAAAAAAA GTATTTTTCA TCTAAATTTT GAATTTGAAA GTGTCTTAAA TTGATTTTCA AAAGTCATAT AATTTGGAAC GGAGGGAGTA ACAAACAT	N1 N2 N3
P. quinquefolius Reference Sequence	TACAGTGAT AATTAAATAT TGTAACTATC TAAAAAAAAA GTATTTCTCA TCTAAATTTT GAATTTGAAA GTGTTTTAAA TTGATTTTCA AAAGTCATAT AATTTAGAAC GGAGGGAGTA ACAAACAT	N1 N2 N3

Table 2
List of all PCR primers used for the Illumina sequencing. To create the 127-nt section of the DS sequence from genomic samples, P8′ and P7 primers are used. To prepare the sequencing library, PCR1 and PCR2 primers are used. For the PCR1 primers, the bold sections, which are locus-specific, are the same as the P8′ and P7 primers, respectively; the italic sections form toeholds on the PCR1 product during the first-round DNA amplification. The complementary sections of these parts of the toeholds (red or blue) anneal to the 3' end of the PCR2 primers (red or blue) in the second-round amplification process which forms the adapters (up to $4 \times 4 = 16$ distinct combinations). These adapters contain the universal primers (italic) and indices (underlined) in the PCR2 primers (FP2 and RP2) that are needed for Illumina sequencing.

P8′ Forward Primer	TACAGTGATAATTAAATATTGTAACTATCTAA
P7 Reverse Primer	ATGTTTGTTACTCCCTCCGTT
PCR1 Primers	
Forward PCR1 Primer	*TCGTCGGCAGCGTC**AGATGTGTATAAGAGACAG*****TACAGTGATAATTAA ATATTGTAACTATCTAA**
Reverse PCR1 Primer	*GTCTCGTGGGCTCGG**AGATGTGTATAAGAGACAG*****ATGTTTGTTACTCCCTCCGTT**
PCR2 Forward Primers (FP2)	
FP2-1	*AATGATACGGCGACCACCGAGATCTACAC*AGCGCTTCGTCGGCAGCGTC
FP2-2	*AATGATACGGCGACCACCGAGATCTACAC*GATATCTCGTCGGCAGCGTC
FP2-3	*AATGATACGGCGACCACCGAGATCTACAC*CGCAGATCGTCGGCAGCGTC
FP2-4	*AATGATACGGCGACCACCGAGATCTACAC*TATGAGTCGTCGGCAGCGTC
PCR2 Reverse Primers (RP2)	
RP2- 1	*CAAGCAGAAGACGGCATACGAGAT*GTGAATGTCTCGTGGGCTCGG
RP2- 2	*CAAGCAGAAGACGGCATACGAGAT*ACAGGCGTCTCGTGGGCTCGG
RP2- 3	*CAAGCAGAAGACGGCATACGAGAT*CATAGAGTCTCGTGGGCTCGG
RP2- 4	*CAAGCAGAAGACGGCATACGAGAT*TGCGAGGTCTCGTGGGCTCGG

2. Materials and methods

2.1. Materials

All single-stranded (ss) oligonucleotides were purchased from IDT. A Qiagen plant DNA extraction kit was used to prepare the genomic DNA from fourteen ginseng samples. Six of the samples were (X1-6) powder samples provided by Macan Biotechnologies Ltd. of Macau and have no claimed species identities. The remaining eight ginseng samples were root and granular samples purchased from local herbal stores: three of them were claimed to be *P. quinquefolius* (AmG, AmG2, AmG3) and five were claimed to be *P. ginseng* (ChG, ChG2, ChG3, KorG, KorG2). Therefore, there are six powdery samples, one granular sample and seven root samples, i.e. a total of 14 samples.

The locus used for the NGS is a 127 nucleotide (nt) section within the DS gene which contains three SNP site (shown in Table 1). These reference sequences were used to design the sequencing primers (shown in Table 2).

3. Methods: PCR

To conduct the polymerase chain reaction (PCR), the PCR kit (ABM) was used. In the PCR mixture (50 μL) there are the following

components: 5 μL of 10 × PCR buffer (with 15 mM $MgCl_2$), 1.5 μL each of P8' forward primer and P7 reverse primer (both at 10 μM), template [i.e. 10 μL of genomic DNA (100 ng) or 0.5 μL purified PCR product (30 ng)], 1 μL of $MgSO_4$ (25 mM), 1 μL of dNTPs (10 mM), 1 μL of *Taq* polymerase (5 U/μL), and deionized (DI) water (in the volume of 29 μL for genomic templates or 38.5 μL for PCR product templates). Primer pairs: P8′ and P7 were previously designed and used [5].

Thermocycling for the PCR mix was conducted in a thermocycler (Techne ^{3}Prime). The PCR program consisted of (1) 94 °C initial denaturation for 3 min, (2) 30 cycles of 95 °C denaturation (30 s), 50 °C primer annealing (30 s) and 72 °C elongation (30 s), and (3) final elongation at 72 °C for 3 min.

A PCR purification kit (QIAquick® PCR Purification Kit) was then used to purify the PCR mixtures. Briefly, Buffer PB (250 μL) was mixed with the PCR mix (50 μL) by vortexing. The mixture was subsequently transferred to a QIAquick spin column. The column was centrifuged at 13,000 rpm for 1 min, with the flow-through liquid discarded. Then, Buffer PE (750 μL) was added to wash the spin column, centrifuged for 1 min; after the flow-through was discarded, the spin column was centrifuged again for 2 min. The spin column was then transferred to a clean centrifugal tube (1.5 mL) and the PCR product was eluded using DI water (30 μL) by centrifuging for 1 min. All centrifugations were conducted at 13,000 rpm.

To quantify the PCR product, a UV spectrometer (Thermo Scientific Nanodrop 1000) was used; the quantity was determined based on the absorbance at 260 nm, with the purity determined from the ratio of the 260 nm/280 nm absorbances.

The fourteen PCR products (127 nt) derived from the different ginseng samples were used to prepare the sequencing library. Two rounds of PCR (i.e. PCR1 and PCR2) were used to amplify each sample, based on the same procedure as described above. The PCR1 primer pair (Table 2) was used for the first round of PCR. After the products were purified, the PCR2 forward and reverse primers were used to amplify these products for a second round of PCR. Here, the adapter sequence and a different combination of indices (in the PCR2 forward and reverse primers) were inserted into each ginseng sample; there were 14 different combinations of indices used. The adapter sequence was used in order to allow for bridge amplification to occur in the Illumina sequencing protocol [27].

The red region in the 5′ end of the PCR1 forward primer (the left half of the toehold) is complementary to the PCR2 forward primer (FP2). The blue region in the 5′ end of the PCR1 reverse primer is complementary to the PCR2 reverse primer (RP2). The use of the PCR1 primers is for locus-specificity amplification and the PCR2 primers are generic for any species/locus in order to insert adapters and indices. The use of such a method as opposed to a single set of primers is two-fold: 1) because the strand length in oligonucleotide synthesis is limited, it is hard to synthesize the full length of the adapter sequence with the index. 2) Only the PCR1 primers ought to be custom made to recognize a specific locus of interest. The PCR2 primers are universal, and they can be used for any species/locus (i.e. the primers are generic), see Fig. S1.

After purification of the fourteen PCR2 products, these products were first quantified and then mixed in equimolar amounts to have a cumulative DNA concentration of 4 ng/μL in DI water. The sequencing mixture (10 μL) was sent to Genewiz for Illumina sequencing (HiSeq 2 × 150 bp).

Once the sequencing task was completed, twenty eight data files (each of 2 GB in FASTQ format) were received, corresponding to the forward read (R1) and reverse read (R2) of the 14 ginseng samples. The data files, each corresponding to *ca.* 28 million reads, were first analyzed using the Burrows-Wheeler Aligner (BWA) [28], which aligned the R1 (sense) sequences according to the sequence standard: *P. ginseng* sense strand (see Table 1); this alignment generated the sequence alignment map (SAM) file. To visualize the newly formed SAM files, the Integrated Genomic Viewer (IGV) was used [29]. Thereafter, the *P. ginseng/P. quinquefolius* ratios were determined (see Table S1) to identify a good pair of samples to serve as training samples for the clustering algorithm.

3.1. Abundance analysis

To determine the abundance of each genotype, a custom-written Python program was written and used. This script searched for each SNP site by finding the conserved five to seven nt sequences directly in front of each of the SNP sites to locate the SNP site. This was followed by recording the SNP sites.

This forms the basis for performing the clustering analysis.

3.2. Clustering model

A Gaussian Mixture clustering model written in Python [30] was built to identify ginseng genotypes. This is achieved by first selecting a pair of sequencing samples to serve as *P. ginseng* and *P. quinquefolius* references and then using them to train the clustering model. Then this model was used to characterize the remaining 12 ginseng samples.

To prepare the DNA sequences for clustering, three modifications were first made to the sequences:

1) The extra nt from the sequencing primers were removed
2) The "N"s were replaced by either "*A*", "*T*", "*G*" or "*C*", based on whichever one is the most common at that position in similar sequences
3) The DNA sequence reads were converted into k-mers, which are sub-sequences of k characters in strings (or k nucleotides in DNA sequences) (length k = 6)

In a preliminary analysis, 10,000 reads were first used to determine the k-mers.

For building the clustering model, 200,000 reads were taken from each of the two reference samples and converted into k-mer vectors using countVectorizer, which is a method to convert nucleotide texts into numeral data. The clustering was performed using

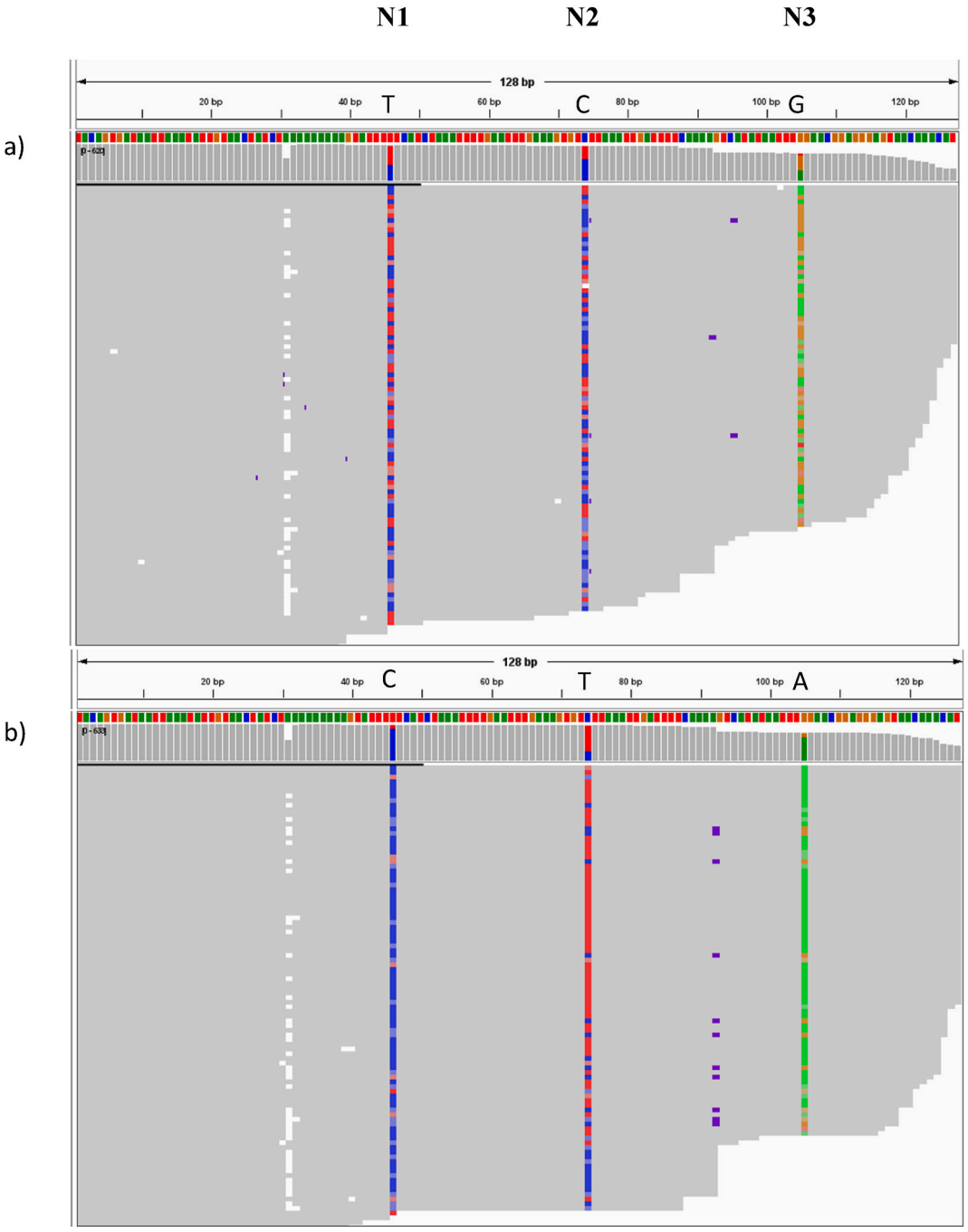


Fig. 1. Genotyping results (IGV images) of the ChG (a) and KorG (b) sequences (127 nt). The genotyping results are listed in the order of decreasing lengths (The ChG image contain 620 reads and the KorG image contains 633 reads.) The *P. ginseng* reference sequence (see Table 1) is represented by the top row of coloured pixels. The genotyping results below, are labeled with the SNP sites: N1, N2 and N3. From the results, additional genetic variations such as deletions and insertions are revealed. The coloured boxes [i.e. C (blue), T (red), G (orange), A (green), deletion (white) and insertion (purple)] represent any position that deviates by $\geq$ 20% relative to the reference. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the Gaussian Mixture model, which accounts for variance in sequences in underlying samples. In order to reduce dimensionality of the sequence vectors, k-mers with the highest variances were selected as clustering features. The model uses default parameters (number of clusters = 3, Gin, Quin, CCG). The model was saved (as a .pkl file) and it was used to perform clustering analysis for the remaining 12 ginseng sequences.

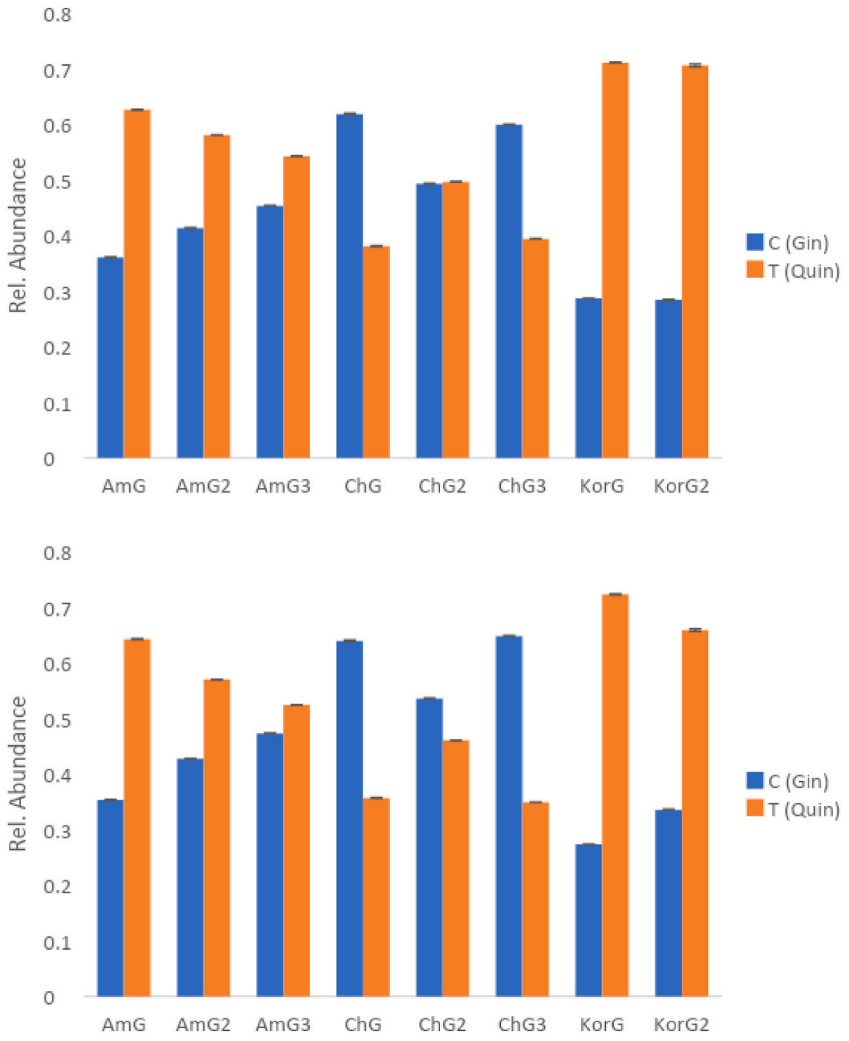


Fig. 2. NGS genotyping results of the N2 site for AmG, AmG2, AmG3, ChG, ChG2, ChG3, KorG, and KorG2 using (a) BWA/IGV and (b) the Python program on the data in the FASTQ files where *P. ginseng* contains a C nucleotide, and *P. quinquefolius* contains a T nucleotide. The abundances have been normalized to the total number of sequences for each sample to generate the relative abundance values. The error bars are calculated as an average of the uncertainties for each read of the N2 nucleotide.

4. Results and discussion

4.1. Burrows-Wheeler Aligner (BWA)

The FASTQ files were first analyzed using the Burrows-Wheeler Aligner (BWA) which generated the sequence alignment map (SAM) files, and they were visualized using the Integrated Genomic Viewer (IGV). The IGV image of the most *P. ginseng*-like sample (ChG) is presented in Fig. 1*a,* and the most *P. quinquefolius*-like sample (KorG) is depicted in Fig. 1*b*. The three coloured columns, which depict nucleotides with high variance among the sequences shown in the invariant grey areas, represent the nucleotides at the three SNP sites of N1, N2 and N3.

As shown in Fig. 1a, the ChG sample appears to have more of the *P. ginseng* genotype (i.e. T (red), C (blue), G (orange)) than the *P. quinquefolius* genotype: C, T, A (blue, red, green), which confirms that ChG is the most *P. ginseng*-like sample. Nonetheless, the sample genotype does appear to be heterozygous. In Fig. 1b, the sample has shown its apparent heterozygous nature, although KorG has shown its more *P. quinquefolius*-like nature with the CTA genotype. In both images, a new CCG genotype is revealed (see Table S2), and this genotype has not been previously reported [3].

What is also seen in Fig. 1a is that there is a 1-nucleotide deletion (shown in white) prior to the N1 SNP site in many of the ChG sequence variants. Furthermore, between N2 and N3, there is a 14-nucleotide insertion (shown in purple), which is AACAAACAATA/GATT/C, where the two underlined sections are either A or G and either T or C, respectively. It is also noteworthy to see that the same insertions (shown in purple) are also found in the KorG image, as were seen in the ChG image (between N2 and N3). Moreover, these

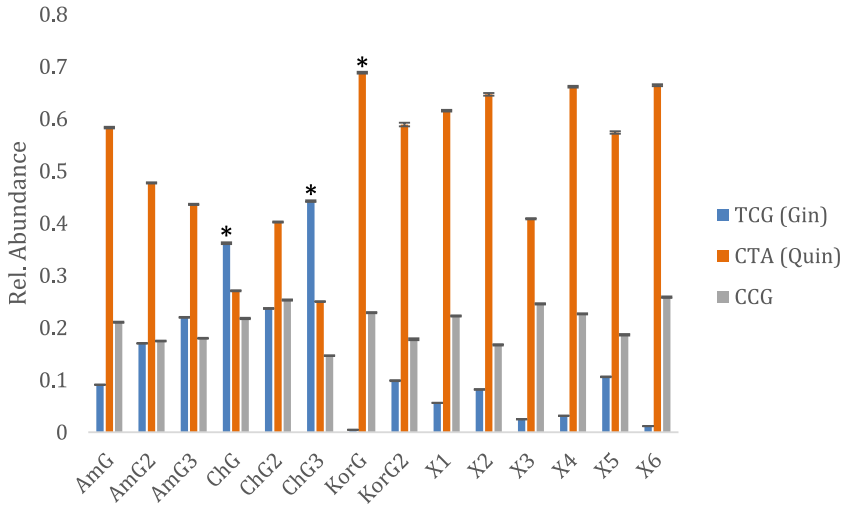


Fig. 3. Abundance-based genotyping results on FASTQ data of the N1, N2 and N3 SNP sites of X1-X6 (powder), and AmG, AmG2, AmG3, ChG, ChG2, ChG3, KorG, KorG2 (root). *P. ginseng* consists of the TCG genotype and *P. quinquefolius* comprises the CTA genotype. The new genotype is labeled as CCG. In order to generate the relative abundance values, the intensities have been normalized to the total number of sequences for each sample. Error bars are calculated as an average of the uncertainties for each read of the N1, 2 and 3 nucleotides.

14-nt insertions are found on sequences that consist of the new CCG genotype, in additional to TCG and CTA. This newly discovered genotype has the 1-nucleotide deletion in front of the N1 SNP site too. (This CCG genotype that contains such an insertion was also observed in the top image.)

The identification of the CCG genotype prompted us to ask how abundant this genotype was in comparison to the TCG and CTA genotypes. To address this question, a Python program was written in order to identify the three SNP sites by searching the conserved 6–8 nt sequence directly in front on the respective SNP site in the FASTQ files.

In order to verify that the Python program is a viable method for analyzing the FASTQ files, the N2 sites was first used. The relative abundances of the two genotypes determined for the same samples were compared using BWA and the Python program (Fig. 2). We are pleased to see that there is a general agreement between the results obtained from the two methods of data analysis, though there are slight differences in some abundances, especially for ChG, ChG2, ChG3, which have higher *P. ginseng* compositions.

4.2. Abundance analysis

It is concluded that the Python program is a viable approach for classifying the different genotypes based on the results presented in Fig. 2. This program was employed to determine the abundance for the new CCG genotype in the 14 samples.

The results are shown in Fig. 3. We found ChG and ChG3 are the most prominent *P. ginseng* with the TCG genotype. Then, we found that the new CCG genotype is actually more abundant than the TCG genotype in X1-6, AmG, KorG, KorG2, see Fig. 3, most of them have concurrently abundant CTA genotypes. The only samples that have a higher relative abundance of the TCG genotype (compared to CTA) are ChG and ChG3, unaffected by the presence of the CCG genotype.

On the other hand, some samples (e.g., X6 and KorG) have very little *P. ginseng* genotype. But the existence of the CCG genotype leads to an apparently higher C (Gin) character; this is reflected in Fig. 2 for KorG when the classification is based on one nucleotide, i.e. C of the N2 site, alone.

The CCG genotype is found in a significant portion in the reads of the fourteen samples. Therefore, the combination of the CCG and TCG genotypes, has made the N2 site sufficient to produce conclusive results for many ginseng samples.

The success was previously shown when the N2 site was used for ginseng species differentiation using NAT based on hybridization [5], and NAAT based on ligase-based amplification [14].

Based on Fig. 3, the sample with the most *P. ginseng* genotype (TCG) is ChG3 and the sample with the most *P. quinquefolius* genotype (CTA) is KorG. This is quantified using the data from Fig. 3, in which ChG3 show the highest Gin/Quin ratio and KorG has the lowest ratio, see Table S1. Therefore, ChG3 and KorG were now selected to be the reference samples for the clustering analysis of *P. ginseng* and *P. quinquefolius,* respectively.

4.3. Clustering analysis

A preliminary clustering analysis was first performed by the Gaussian Mixture algorithm using 10,000 reads from both ChG3 and KorG based on two, three or four clusters. When two clusters were used, the *P. ginseng* (TCG) and CCG genotypes were grouped together in one cluster and the *P. quinquefolius* genotype (CTA) was placed in the other cluster. When three clusters were used, each of the three genotypes in Fig. 3 were placed in a separate cluster. When four clusters were used, each of the three genotypes were placed in separate

Table 3
List of ten k-mers (k = 6) selected for the clustering analysis.

I	Gin N1	ATTTTT
II	Quin N1	ATTTCT
III	Gin N2	CTTAAA
IV	Quin N2	TTTAAA
V	Gin N3	AATTTG
VI	Quin N3	AATTTA
VII	N4 with no IX	AAAAGT
VIII	N4 with IX	AAAAAT
IX	Insertion	ACAAAC
X	Deletion	AAAAAA

Table 4
Covariance table of the training samples (ChG3 and KorG) using the ten k-mers (see Table 3 for definitions of Roman numerals) chosen for the analysis where more positive values appear darker and indicate a more positive covariance between the two k-mers, and more negative values appear lighter indicated a more negative covariance between the two k-mers.

	I	II	III	IV	V	VI	VII	VIII	IX	X
I										
II	-0.7676									
III	0.4252	-0.4001								
IV	-0.286	0.3368	-0.7047							
V	0.2704	-0.1943	0.5303	-0.4614						
VI	-0.3177	0.3343	-0.5459	0.5091	-0.559					
VII	0.1224	-0.0409	-0.1392	-0.0185	-0.0185	0.3096				
VIII	-0.1351	0.2131	0.2941	0.2131	0.2131	-0.296	-0.5699			
IX	-0.127	0.1801	0.2529	-0.2117	0.3344	-0.1793	-0.214	0.3926		
X	0.153	-0.0552	-0.0256	0.1159	0.0557	0.0781	0.3469	-0.0915	-0.0917	

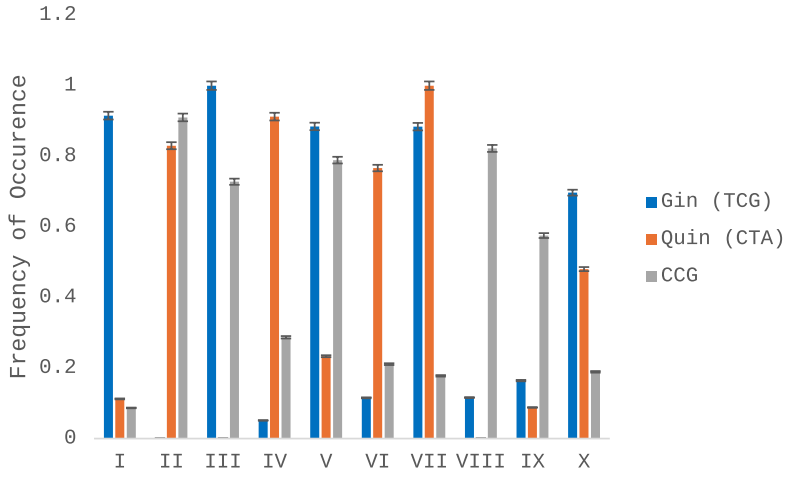


Fig. 4. Frequency of occurrences of the ten k-mer. It should be noted that the k-mers of V and VII already appeared once in a conserved locus in addition to the SNP sites, as a result, these two k-mers have their occurrences inflated by one. Likewise, since the k-mer of X pertains to a repeat of 5–6 adenosines (*As*), there are 3–4 occurrences of the k-mer, which results in its frequency being inflated by 3. Both inflations have been subtracted from the respective k-mers. Error bars are calculated by running a simulation of all 14 samples with the read uncertainties of the kmers and comparing the simulated results to the actual results. The simulated sample was generated used random number generation with probabilities specified in the FASTQ files. Original and simulated sequences were run through clustering and the differences in classification were counted as errors.

clusters and outlier reads were placed in the fourth cluster. Based on these results, the three-cluster system was chosen to be used. First, it had to be determined which k-mers to use. This was a necessary step since analyzing all possible k-mers on all reads would require too much computer time.

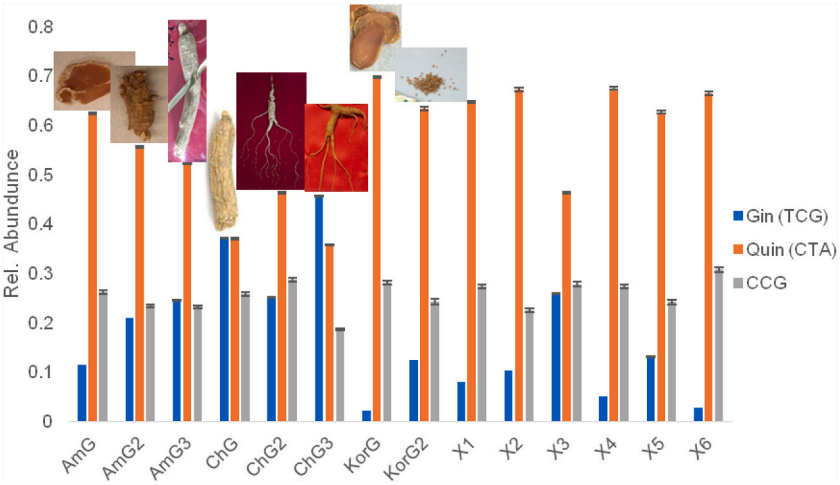


Fig. 5. Gaussian Mixture clustering results (three clusters) to analyze the 14 samples, see Figs. 3 and 4 for notations. All intensities have been normalized to the total number of sequences for each sample to generate the relative abundances. Error bars are calculated by running a simulation of the samples with the read uncertainties of the k-mers and by comparing the simulated and actual results. The simulated sample was generated used random number generation with probabilities specified in the FASTQ files. Original and simulated sequences were run through clustering and the differences in classification were counted as errors.

From the preliminary results, it was found that six k-mers (Table 3, entry I to VI) associated with the three SNP sites has high covariances or high variability, as expected given what was presented in Fig. 1. In addition, four k-mers from other loci (e.g. deletion, insertion) have notable covariances. One of these loci was the insertion (IX) located between the N2 and N3 sites that was observed in Fig. 1 (in green in Table S2). The second locus was the one-nt deletion (X) in the nine "A" repeat in the 127 nt sequence shown (see Table 1 and Table S2). The third locus was a SNP site that was not previously reported, located at the three nucleotides (in blue, see Table S2) prior to the insertion sequence (VII to VIII).

A covariance table was generated. The positive covariances (when two k-mers are likely to be found on the same read) and negative covariances (when two k-mers are likely to be found on different reads) were listed in Table 4. There are positive covariances between the genotype at the three SNP sites; for instance, the covariance of Gin N2 (IV) and Gin N3 (V) is 0.5303. There is also a positive covariance (0.3926) between the genotype at this new SNP site N4 (VIII) and the insertion (IX).

Next, the ten k-mers were used to train the Gaussian Mixture algorithm using 200,000 reads of ChG3 and KorG. The results are presented in Fig. 4 indicated that the algorithm has found the frequency of occurrences of the three clusters of the sequences based on the three genotypes of TCG (Gin), CTA (Quin) and CCG.

The trained Gaussian Mixture algorithm was next used to genotype the 14 ginseng samples and the resulting three clusters in each sample are shown in Fig. 5. It is observed that the graph in Fig. 5 is similar to Fig. 3; this is anticipated since six of the ten k-mers (I–VI) used in the algorithm pertain to the N1, N2 and N3 SNP sites.

A BLAST search was performed in the National Center for Biotechnology Information (NCBI) website using the three genotypes (shown in Fig. 4 and Table S2). The DS genotypes of *P. ginseng* (TCG) and *P. quinquefolius* (CTA) were found to be an exact match to a locus in chromosome 2 [21] (accession number: JAINUU010000002.1) of their corresponding species. On the other hand, the new CCG genotype was confirmed to be novel, as it had no close match within the *P. quinquefolius* genome, and had a near match to a locus in *P. ginseng,* but on the wrong chromosome (i.e. 17, see Fig. S2).

5. Conclusions

In this work, the CCG genotype is discovered along with other genetic variations such as deletions and insertions, which have not previously been reported [3]. This was achieved by BWA, abundance and clustering on the data obtained by next-generation sequencing (NGS).

It is confirmed that ChG3 is *P. ginseng* and KorG is *P. quinquefolius*. The ambiguities found in other samples may be due to the hybrid nature of the *Panax* species.

While a new genotype of CCG was revealed, N2 remains the robust SNP site for ginseng species authentication, though there are better loci to be discovered.

CRediT authorship contribution statement

Christopher Oberc: Writing – review & editing, Writing – original draft, Methodology, Formal analysis. **Paul C.H. Li:** Writing – review & editing, Supervision, Funding acquisition, Formal analysis.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Paul C.H. Li reports financial support was provided by NSERC of Canada. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We would like to thank Margaret Oberc for her assistance in performing the clustering models and NSERC of Canada for funding of this work.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e29104.

References

[1] S. Shibata, O. Tanaka, K. Sôma, Y. Iida, T. Ando, H. Nakamura, Studies on saponins and sapogenins of ginseng the structure of panaxatriol, Tetrahedron Lett. 6 (1965) 207–213.
[2] S.J. Fulder, The growth of cultured human fibroblasts treated with hydrocortisone and extracts of the medicinal plant Panax ginseng, Exp. Gerontol. 12 (1977) 125–131.
[3] W. Hu, N. Liu, Y. Tian, L. Zhang, Molecular cloning, expression, purification, and functional characterization of dammarenediol synthase from Panax ginseng, BioMed Res. Int. 2013 (2013) 1–7.
[4] X. Weng, H. Jiang, D. Li, Microfluidic DNA hybridization assays, Microfluid. Nanofluidics 11 (4) (2011) 367–383.
[5] C. Oberc, A. Sedighi, P.C.H. Li, The genetic authentication of Panax ginseng and Panax quinquefolius based on using single nucleotide polymorphism (SNP) conducted in a nucleic acid test chip, Anal. Bioanal. Chem. 414 (13) (2022) 3987–3998.
[6] C. Oberc, P. Brar, P.C.H. Li, Centrifugal dynamic hybridization conducted in a microfluidic chip for signal enhancement in nucleic acid tests, Anal. Biochem. 658 (2022).
[7] K. Krawczyk, B. Uszczyńska-Ratajczak, A. Majewska, N. Borodynko-Filas, DNA microarray-based detection and identification of bacterial and viral pathogens of maize, J. Plant Dis. Prot. 124 (6) (2017) 577–583.
[8] Y. Liu, X. Wang, L. Wang, X. Chen, X. Pang, J. Han, A nucleotide signature for the identification of American ginseng and its products, Front. Plant Sci. 7 (MAR2016) (2016).
[9] C. Hurth, J. Yang, M. Barrett, C. Brooks, A. Nordquist, S. Smith, et al., A miniature quantitative PCR device for directly monitoring a sample processing on a microfluidic rapid DNA system, Biomed. Microdevices 16 (6) (2014) 905–914.
[10] T.H. Fang, N. Ramalingam, D. Xian-Dui, T.S. Ngin, Z. Xianting, A.T. Lai Kuan, et al., Real-time PCR microfluidic devices with concurrent electrochemical detection, Biosens. Bioelectron. 24 (7) (2009) 2131–2136.
[11] Y. Liu, C. Li, Z. Li, S.D. Chan, D. Eto, W. Wu, et al., On-chip quantitative PCR using integrated real-time detection by capillary electrophoresis, Electrophoresis 37 (3) (2016) 545–552.
[12] A. Kausar, C.J. Mitran, Y. Li, J.M. Gibbs-Davis, Rapid, isothermal DNA self-replication induced by a destabilizing lesion, Angew. Chem. Int. Ed. 52 (40) (2013) 10577–10581.
[13] B. Safeenaz Alladin-Mustan, Y. Liu, Y. Li, D.R.Q. de Almeida, J. Yuzik, C.F. Mendes, et al., Reverse transcription lesion-induced DNA amplification: an instrument-free isothermal method to detect RNA, Chem 1149 (2020).
[14] C. Oberc, P. Sojoudi, P.C.H. Li, Nucleic acid amplification test (NAAT) conducted in a microfluidic chip to differentiate between various ginseng species, Analyst 148 (3) (2022) 525–531.
[15] C. Cheng, W. Wu, B. Huang, L. Liu, P. Luo, H. Zhou, SNPs of dammarenediol synthase gene were associated with the accumulation of ginsenosides in DAMAYA ginseng, a cultivar of Panax ginseng C. A. Mey, Phytochem. Lett. 17 (2016) 194–200.
[16] M. Huang, Y. Bai, S.L. Sjostrom, B.M. Hallström, Z. Liu, D. Petranovic, et al., Microfluidic screening and whole-genome sequencing identifies mutations associated with improved protein secretion by yeast, Proc. Natl. Acad. Sci. U.S.A. 112 (34) (2015) E4689–E4696.
[17] J. Korlach, K.P. Bjornson, B.P. Chaudhuri, R.L. Cicero, B.A. Flusberg, J.J. Gray, et al., Real-time DNA sequencing from single polymerase molecules, Methods Enzymol. 472 (2010) 431–455.
[18] W.S. Pearman, N.E. Freed, O.K. Silander, Testing the advantages and disadvantages of short- and long- read eukaryotic metagenomics using simulated reads, BMC Bioinf. 21 (1) (2020) 220.
[19] A. Bruno, A. Sandionigi, G. Agostinetto, L. Bernabovi, J. Frigerio, M. Casiraghi, et al., Food tracking perspective: DNA metabarcoding to identify plant composition in complex and processed food products, Genes 10 (3) (2019).
[20] Y.T. Lo, P.C. Shaw, Application of next-generation sequencing for the identification of herbal products, Biotechnol. Adv. 37 (8) (2019).
[21] Z.H. Wang, X.F. Wang, T. Lu, M.R. Li, P. Jiang, J. Zhao, et al., Reshuffling of the ancestral core-eudicot genome shaped chromatin topology and epigenetic modification in Panax, Nat. Commun. 13 (1) (2022).
[22] F.X. Shi, M.R. Li, Y.L. Li, P. Jiang, C. Zhang, Y.Z. Pan, et al., The impacts of polyploidy, geographic and ecological isolations on the diversification of Panax (Araliaceae), BMC Plant Biol. 15 (1) (2015).
[23] N.H. Kim, M. Jayakodi, S.C. Lee, B.S. Choi, W. Jang, J. Lee, et al., Genome and evolution of the shade-requiring medicinal herb Panax ginseng, Plant Biotechnol. J. 16 (11) (2018) 1904–1917.
[24] Y.J. Zuo, J. Wen, S.L. Zhou, Intercontinental and intracontinental biogeography of the Eastern Asian – Eastern North American disjunct Panax (the ginseng genus, Araliaceae), emphasizing its diversification processes in eastern Asia, Mol. Phylogenet. Evol. 117 (2017) 60–74.
[25] Y. Wang, Y. Wang, Y. Chen, W. Wang, Z. He, Z. Lin, et al., Analysis of Panax ginseng miRNAs and their target prediction based on high-throughput sequencing, Planta Med. 85 (14–15) (2019) 1168–1176.
[26] V. Manzanilla, A. Kool, L. Nguyen Nhat, H. Nong Van, H. Le Thi Thu, H.J. De Boer, Phylogenomics and barcoding of panax: toward the identification of ginseng species, BMC Evol. Biol. 18 (1) (2018).
[27] Y. Shin, J. Kim, T.Y. Lee, A solid phase-bridge based DNA amplification technique with fluorescence signal enhancement for detection of cancer biomarkers, Sensor. Actuator. B Chem. 199 (2014) 220–225.

[28] L. Heng, D. Richard, Fast and accurate short read alignment with Burrows-Wheeler Transform, Bioinformatics 25 (14) (2009) 1754–1760.
[29] J.T. Robinson, H. Thorvaldsdóttir, A.M. Wenger, A. Zehir, J.P. Mesirov, Variant review with the integrative genomics viewer, Cancer Res. 77 (21) (2017) e31–e34.
[30] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A and Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, F. Pedregosa, G. Varoquaux, A. Gramfort, et al., Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.