Université du Québec en Abitibi-Témiscamingue

ANALYSE COMPARATIVE DES GÉNOMES CHLOROPLASTIQUES DES ESPÈCES DE MÉLÈZES DU NORD DE L'ASIE ET DU NORD DE L'AMÉRIQUE

Mémoire
Présenté
Comme exigence partielle
de la maîtrise en écologie et aménagement des écosystèmes forestiers

Par
Amal Saidani

Octobre 2024

REMERCIEMENTS

Avec une profonde gratitude que je souhaite, adresser mes remerciements à toutes les personnes qui ont contribué à l'aboutissement de ce travail. J'exprime ma sincère gratitude à mon encadrant, le professeur Mebarek Lamara pour son encadrement avisé, sa disponibilité et ses conseils précieux. Je tiens à exprimer ma reconnaissance envers le professeur Yves Bergeron d'avoir accepté d'être mon encadrant et pour le temps qu'il a consacré à évaluer et orienter mes recherches. Mes vifs remerciements vont également à Raju Soolanayakanahally et Juan Carlos Villarreal en acceptant de prendre part à mon jury de maîtrise, et d'avoir eu l'amabilité d'évaluer ce document. Je tiens à adresser des mots de gratitude profonde à l'âme de mon père, celui qui a été mon pilier, mon mentor, et ma source inestimable d'amour et de soutien tout au long de ma vie. Je remercie aussi ma famille pour son soutien précieux tout au long de mon parcours de maîtrise.

AVANT-PROPOS

Le présent mémoire s'inscrit dans le cadre de mes études de deuxième cycle en écologie forestière à l'Université du Québec en Abitibi-Témiscamingue. Elle se structure en trois chapitres principaux. Le premier chapitre, introduction générale, est consacré à une revue bibliographique en plus du contexte et des objectifs de l'étude. Le deuxième chapitre est présenté sous forme d'un article scientifique dont lequel nous avons exposé notre méthodologie, et nos résultats de recherche. L'article sera soumis à la revue Molecular ecology. Mes directeurs de recherche ont contribué à la conception de cette étude, en guidant le processus de recherche et en fournissant une assistance lors de l'interprétation des résultats obtenus. Le troisième chapitre de ce mémoire représente une conclusion générale qui synthétise les découvertes et les implications de cette recherche.

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

LISTE DES ABRÉVIATIONS

ADEGENET	Analysis of Genetic Data
ADN	Acide Désoxyribonucléique
AFLP	Amplified Fragment Length Polymorphism
AIC	Akaike Information Criterion
APE	Analysis of Phylogenetics and Evolution
ANOVA	Analysis of Variance
ARN	Acide Ribonucléique
BIC	Bayesian Information Criterion
DNA	Deoxyribonucleic Acid
GCPGVIEW	Genome Comparison and Plotting Tools
GTR	General Time Reversible
ITS	Internal Transcribed Spacer
LSC	Large Single Copy
MAFFT	Multiple Alignment using Fast Fourier Transform
MISA	Microsatellite Identification Tool
ML	Maximum Likelihood
NCBI	National Center for Biotechnology Information
NGS	Next-Generation Sequencing
NJ	Neighbor Joining
OGDRAW	Organellar Genome DRAW
PHANGORN	Phylogenetic Analysis with Graphs and Orthogonal polynomials in R.
PCR	Polymerase Chain Reaction
RFLP	Restriction Fragment Length Polymorphism
RNA	Ribonucleic Acid
SAM	Sequence Alignment/Map
SNP	Single Nucleotide Polymorphism
SSC	Small Single Copy
SSR	Simple Sequence Repeat
UPGMA	Unweighted Pair Group Method with Arithmetic Mean

RÉSUMÉ

Le genre *Larix* compte une douzaine d'espèces appartenant à la famille des Pinacées qui sont largement répandues en Amérique du Nord, en Asie et en Europe. Chez le mélèze, le génome chloroplastique (ADN_{cp}) est utilisé comme un code-barre dans les études de diversité génétique ainsi que les études phylogénétiques. Dans cette étude, nous avons assemblé et annoté le génome complet de l'ADN_{cp} de sept mélèzes.

Les résultats obtenus démontrent que tous les génomes chloroplastiques sont relativement conservateurs en termes de contenu génique, d'arrangement et de taille qui varie entre 122,048 à 123,46 pb. Chaque génome présente une paire de régions inversées, avec une variabilité dans la taille de la région de copie unique large allant de 59,760 pb à 76,699 pb, ainsi que dans la taille de la région de copie unique petite qui s'étend de 43,183 pb à 59,760 pb. Le nombre total de gènes chez toutes les espèces se situe entre 95 et 111, codant pour des protéines, 4 gènes d'ARN ribosomique et de 30 à 34 gènes d'ARN de transfert. Le contenu global en GC (Guanine, Cytosine) varie entre 35,55 % et 40,79%. Des variations ont été observées dans le nombre de séquences simples répétées, ainsi que dans la présence des gènes hypothétiques parmi les différentes espèces. L'analyse phylogénétique a mis en évidence des liens étroits entre les espèces, mais également une proximité génétique chez celles géographiquement proches. Cette étude fournit pour la première fois une caractérisation de l'ADN_{cp} de quatre espèces des mélèzes offrant ainsi des données cruciales pour orienter les initiatives de conservation et éclairer les mécanismes évolutifs du genre *Larix* et de ses espèces apparentées.

Mots clés : Mélèze, génome chloroplastique, code-barre, diversité, SSR, phylogénie, Illumina NovaSeq 6000

INTRODUCTION GÉNÉRALE

Présentation du mélèze. Les mélèzes sont des conifères du genre *Larix*, de la famille des Pinaceae et occupent environ 30 % des terres forestières du monde (Volney & Fleming, 2000). Sur le plan taxonomique, ce genre est composé d'une dizaine d'espèces qui sont officiellement reconnues par Ostenfeld et Syrach-Larsen (1930) (Figure 1). En Amérique du Nord, trois espèces sont recensées ; *L. laricina* (Du Roi) K. Koch, *L. lyallii* Parl. et *L. occidentalis* Nutt. En Europe, on observe principalement l'espèce *L. decidua* (Mill.). Quant à l'Asie, elle abrite six espèces distinctes : *L. gmelinii*, *L. sibirica*, *L. griffithiana* (Lindt Gord.) Carr., *L. Kampferi* (Lamb), L. mastersiana (Rehd. Et wils.) et *L. potaninii* (Batalin).

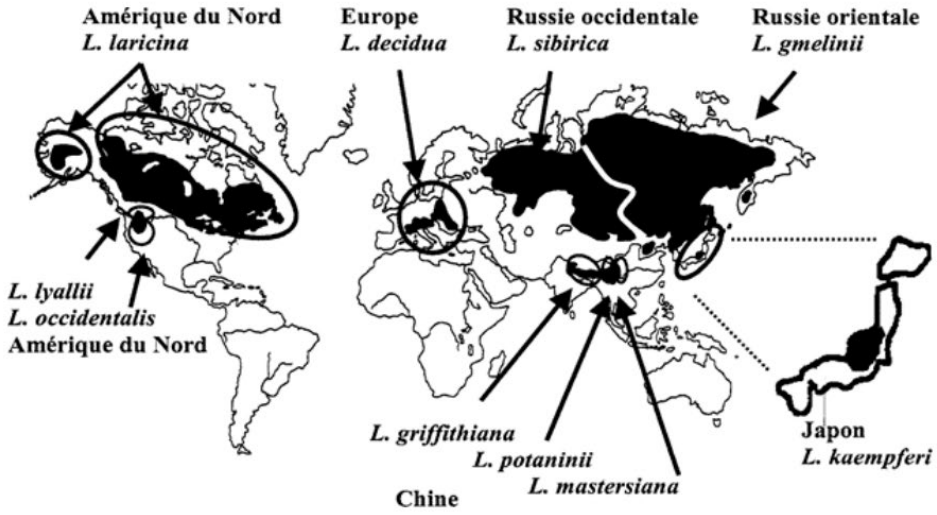


Figure 1
Répartition géographique des mélèzes à l'échelle mondiale
Source : krüssman, 1985

Distribution du mélèze. Originaire de l'Europe centrale, *Larix decidua* se trouve entre 1 400 et 2 400 mètres d'altitude et s'étend aux montagnes des Alpes du sud de la France jusqu'à la frontière de l'Autriche. Il est également présent dans les montagnes de Sudètes, les montagnes Tartras et le centre de la Pologne (McComb, 1955).

Dans les régions subalpines de l'Asie, on trouve le mélèze d'Asie (*Larix gmelinii*), qui s'étend jusqu'aux montagnes Xingan en Chine. Le mélèze japonais (*Larix kaempferi*) pousse à des altitudes comprises entre 1 300 à 2 900 mètres et se trouve principalement dans les régions montagneuses du centre de l'île de Honshu. Il a été introduit en Europe en 1861, en particulier dans les régions côtières au climat océanique, et il se caractérise par une croissance plus rapide que le *Larix decidua* (Isoda *et al*., 2006).

Le mélèze de Sibérie (*Larix sibirica*) couvre la région boréale de l'ouest de la Russie, principalement à des altitudes supérieures à 1 000 mètres. Le mélèze de Sukachev (*Larix sukaczewii*) se trouve au sud-ouest de la Sibérie occidentale et dans le nord-est de l'Europe, coexistant avec d'autres espèces comme le Pin sylvestre (*Pinus sylvestris* L.). Il est séparé de *L. sibirica* sur la base de distinctions morphologiques (Shugart *et al*., 1992).

Dans les hautes régions subalpines de la Chine, *L. mastersiana* couvre la zone Est tandis que *L. potaninii* couvre la région centrale. *L. griffithiana* occupe le sud-ouest de la Chine et s'étend aux montagnes de L'Himalaya (Szmidt, 1987). *L. kaempferi* pousse à 2800 mètres d'altitude dans les régions montagneuses du Japon, y compris les zones volcaniques.

En Amérique du Nord, le mélèze de l'Ouest (*Larix occidentalis*) pousse à des altitudes comprises entre 500 et 2400 mètres. On le trouve au Canada ; au sud-est de la Colombie-Britannique et au sud-ouest de l'Alberta, aussi aux États-Unis ; au nord et au centre ouest de l'Idaho, au nord-est de Washington et dans le centre nord de l'Oregon. Il supporte très bien le froid et préfère les terrains bien drainés non gorgés d'eau (Graham *et al*., 1979). En Amérique du Nord, c'est le mélèze laricin (*Larix laricina*) qui occupe la plus vaste aire de répartition naturelle de tous les conifères. Sa distribution est continue de l'Alaska au nord-est des États-Unis (Klimaszewska *et al*., 1997). L'aire de répartition transcontinentale du mélèze laricin chevauche celle des autres espèces de conifères boréaux largement

distribuées (*Picea mariana (Mill.*)*, Picea glauca* (Moench) *Voss*, *Pinus banksiana Lamb.*, *Abies balsamea* (*L.*) *Mill*) (Warren *et al*., 2016).

Études phylogénétiques du mélèze. Selon une phylogénie controversée, le genre *Larix* contient 10 à 15 espèces (Wei & Wang, 2003 ; Khatab *et al*., 2008 ; Abaimov, 2010). Il a été initialement divisé en deux groupes : un groupe comprenait des espèces caractérisées par de longues bractées (Paucisériales) et l'autre comprenait les espèces à courtes bractées (Multisériales) (Figure 2). Cette classification a été validée par des études morphologiques ultérieures (Bobrov, 1972 ; Farjon, 1990).

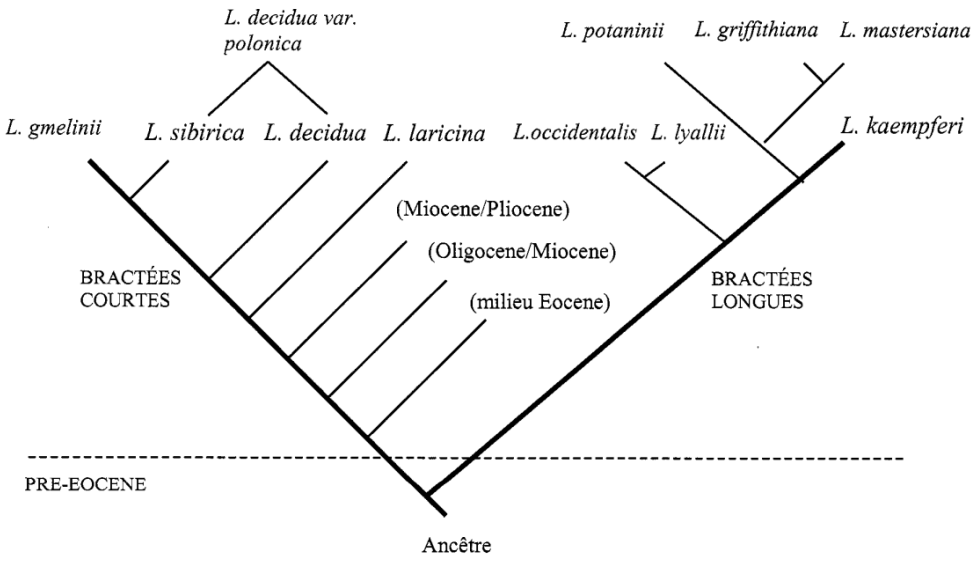


Figure 1
Classification morphologique des espèces du genre *Larix* basée sur la longueur des bractées
Source : LePage & Basinger, 1995

Schorn (1994) n'était pas d'accord avec la division de *Larix* en Paucisériles et Multisériales arguant que la longueur des bractées est un caractère continu d'un point de vue ontogénétiques qu'historique. En effet, il a proposé une division du genre en deux groupes : un groupe I (Aristatus) a une bractée relativement longue qui s'étend bien au-delà des écailles des graines et un groupe II (Laminatus) caractérisé par des bractées plus courtes que les écailles des graines. Toutefois, la

plupart des hypothèses évolutives basées sur le critère des cônes femelles comme caractère fondamental sont en contradiction avec les résultats des analyses moléculaires. Dans ce contexte, des études se sont appuyées sur les marqueurs d'ADN nucléaire (Kisanuki *et al*., 1995 ; Gernandt & Liston 1999 ; Arcade *et al*., 2000 ; Khasa *et al*., 2000 ; Semérikov *et al*., 2003 ; Wei & Wang 2003). D'autres études phylogéniques ont porté sur des marqueurs d'ADN cytoplasmique, principalement d'origine chloroplastique (Qian *et al*., 1995 ; Semérikov et Lascoux 2003 ; Semérikov *et al*., 2003 ; Wei & Wang, 2003 ; Achéré *et al*., 2004). Qian *et al*., (1995) ont apporté la première preuve moléculaire à partir de l'analyse RFLP de l'ADN_{cp} contredisant la classification basée sur des écailles de bractées. Ils ont suggéré une relation étroite entre *L. sibirica* et les mélèzes d'Amérique du Nord, tout en distinguant *L. griffithii* de toutes les autres espèces de mélèzes. Cette classification est sujette à une certaine incertitude en raison de la méthode utilisée pour calculer les fragments d'hybridation plutôt que de considérer la présence ou l'absence de sites de restriction (Wei & Wang, 2003). Kisanuki *et al*., (1995) ont utilisé onze marqueurs RFLP chloroplastiques pour construire la phylogénie de huit espèces et quatre sous-espèces de *Larix*.

Les relations génétiques obtenues sont presque en cohérence avec le patron morphologique obtenu par Le Page & Basinger (1995) à l'exception de la position des espèces de l'Amérique du Nord. Gernandt & Liston (1999) ont construit la phylogénie de *Larix* en analysant la séquence de la région ITS de l'ADN nucléaire et ont constaté que le genre était divisé en un clade nord-américain et eurasien. Ce découpage phylogénétique selon les continents correspondait au résultat de l'analyse des allozymes (Semerikov & Lascoux, 1999). Cependant, certaines espèces chinoises n'ont pas été incluses dans ces études, de plus l'hybridation naturelle se produit fréquemment entre les espèces de *Larix*, telles que *L. sibirica* et *L. gmelinii*, *L. occidentalis* et *L. lyallii*, *L. potaninii* et *L. mastersiana* (Szmidt *et al*., 1987 ; Mogensen, 1996). Un point commun émerge de ces études, à savoir que le genre *Larix* peut être divisé en trois groupes principaux ; Nord-Américain, Nord-Eurasiens et Sud-Asiatiques (Figure 3). Néanmoins, des désaccords persistent en

fonction du type de marqueur utilisé et du génome analysé (Gros-Louis *et al*., 2005). Par exemple, une analyse de la région chloroplastique trnT-trnF a regroupé *L. sibirica* avec des taxons sud-asiatiques (Wei & Wang, 2003), tandis qu'une analyse des allozymes l'a associé à d'autres taxons Nord Eurasiatiques (Semerikov & Lascoux, 1999). Selon l'étude de Gros-Louis *et al*., (2005), la position de *L. sibirica* par rapport aux autres espèces nécessiterait une confirmation par le séquençage de régions supplémentaires du génome chloroplastique en raison de leurs résultats contradictoires. Ils ont trouvé deux hypothèses différentes ; la première suggère que *L. sibirica* est regroupée avec les espèces asiatiques en se basant sur l'ADN_{cp} tandis que la seconde le place avec des espèces nord-eurasiennes en se basant sur l'ADNmt. Il est difficile de reconstruire la phylogénie des espèces du genre en se basant uniquement sur l'analyse des gènes nucléaires en raison de l'évolution réticulée. Le gène du chloroplaste de mélèze est transmis paternellement et par conséquent, une phylogénie résolue de l'ADN_{cp} sera utile pour révéler son histoire évolutive (McGrath *et al*., 2001).

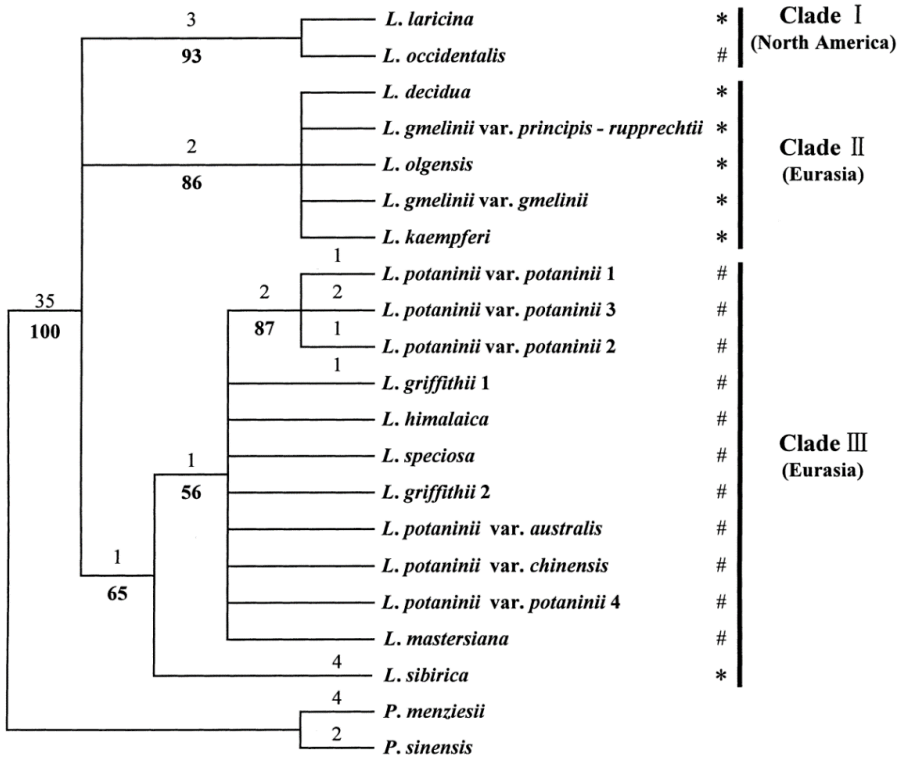


Figure 2
Arbre phylogénétique des espèces du genre *Larix* basé sur l'analyse de la région trnT-trnF du génome chloroplastique
Source : Wei & Wang, 2003

Le génome chloroplastique. Les chloroplastes sont des organites spécifiques des cellules végétales, se distinguent par la présence de pigments photosynthétiques qui captent la lumière. Ils sont plus volumineux et complexes que les mitochondries et remplissent divers rôles dans les cellules végétales au-delà de la simple production d'ATP (Geoffrey, 2017).

En particulier, les chloroplastes jouent un rôle prédominant dans le processus de la photosynthèse, ils sont également impliqués dans la synthèse d'acides gras, de lipides, d'acides aminés, de vitamines et d'autres métabolites. Ils sont essentiels pour la fixation du carbone chez les plantes (Zybailov *et al*., 2008). Le génome des chloroplastes diffère de ceux des mitochondries et du noyau cellulaire, présentant une structure hautement conservée. En conséquence, les taux de mutation dans le génome des chloroplastes sont généralement plus bas que dans le génome nucléaire,

limitant les changements génétiques au fil du temps (Leister, 2003). De plus, le génome chloroplastique est souvent circulaire et compact, avec un nombre restreint de gènes principalement impliqués dans la photosynthèse et d'autres fonctions spécifiques aux chloroplastes (Dobrogojski *et al*., 2020). Il se présente sous forme d'ADN circulaire double brin dont la taille varie, selon l'espèce, de 120 000 à 180 000 paires de bases. Il contient une paire de répétitions inversées (IR), une grande région à copie unique (LSC) et une petite région à copie unique (SSC) (Figure 4) (Chen *et al*., 2020). Il code pour 4 ARN ribosomiques, 30 RNA de transfert et une quarantaine de protéines identifiées (Ohyama *et al*., 1986 ; Ruhlman & Jansen, 2021). Depuis 1909, le génome chloroplastique est reconnu en tant qu'une source d'information génétique autonome avec des traits héréditaires non mendéliens, reflétant ses origines évolutives de bactéries photosynthétiques (Szmidt *et al*., 1987). Ces caractéristiques ont stimulé les études écologiques et évolutives, tout en simplifiant la conception, le séquençage et l'utilisation d'amorces en tant que codes-barres pour l'identification des espèces.

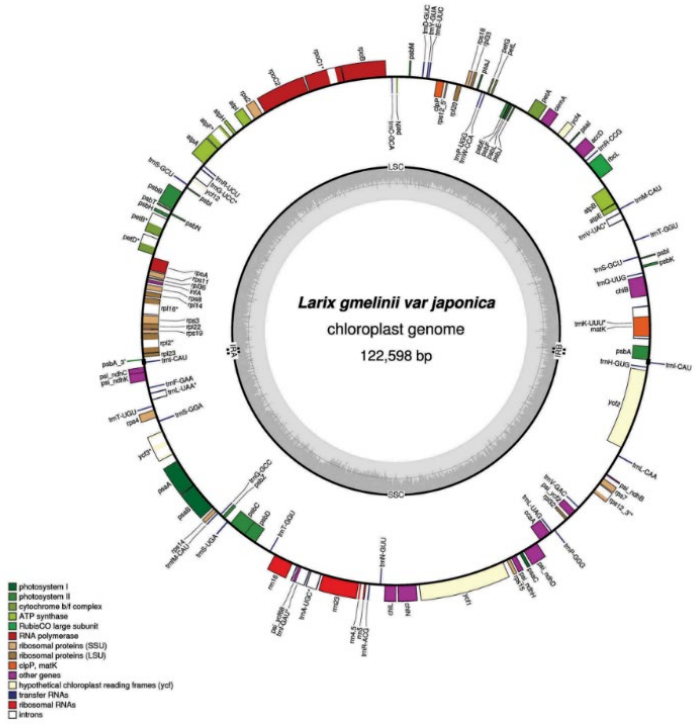


Figure 3
Carte génétique du génome chloroplastique de *Larix gmelinii* var. *japonica*
Source : Ishizuka *et al*., 2017

Les séquences inversées. Une répétition inversée est une séquence d'ADN ou ARN suivie, en aval, de sa séquence complémentaire inversée. Les deux parties de la séquence répétée inversée peuvent être séparées par un nombre variable de nucléotides (Gros-Louis *et al*., 2005). Ces répétitions inversées (IR) sont des éléments cruciaux dans le génome du chloroplaste pour la plupart des végétaux, contribuant au maintien de la structure conservée des séquences d'ADN chloroplastique (ADN_{cp}). Des études antérieures ont rapporté que la longueur des répétitions inversées reste généralement constante parmi les espèces de gymnospermes. Contrairement aux angiospermes, l'ADN du chloroplaste des conifères ne comporte pas de grandes répétitions inversées, mais plutôt de l'ADN répétitif dispersé qui est associé à des réarrangements structurels. En plus des grandes séquences répétées dispersées, l'ADN_{cp} de conifère possède également un certain nombre de petites répétitions. Il englobe des répétitions en tandem de taille variant de 124 à 150 pb, qui sont associées à la région polymorphe réarrangée près de trnK-psbA. L'ensemble du génome chloroplastique est d'une utilité significative pour les études phylogénétiques (Nock *et al*., 2011).

Les introns. Les introns se trouvent dans les gènes de la plupart des organismes ainsi que des organelles telles que le chloroplaste. Ils peuvent être présents dans de gènes, qui génèrent des protéines, de l'ARN ribosomique (ARNr) et de l'ARN de transfert (ARNt) (Qian *et al*., 1995). Ils sont des régions non codantes d'une transcription d'ARN, ou d'ADN qui sont éliminées lors du processus d'épissage avant la traduction. Les séquences qui restent dans l'ARN mature final après l'épissage sont des exons (Ali *et al*., 2019). Afin de mieux comprendre les relations systématiques au sein du genre *Larix*, diverses études se sont concentrées sur l'analyse des introns. L'étude de Gros-Louis *et al*., (2005) a examiné diverses espèces de mélèzes et a révélé une divergence significative dans les séquences de l'intron trnK entre *Larix sukaczewii* et *Larix sibirica.*

Utilité du séquençage de génomes chloroplastiques dans les études évolutives et phylogénétiques des mélèzes. Le séquençage du génome chloroplastique est essentiel pour l'étude de l'évolution, de la diversité génétique et des relations phylogénétiques de mélèzes. De nombreuses études scientifiques se sont penchées sur le génome chloroplastique des différentes espèces de mélèzes. Les deux espaceurs intergéniques trnT (UGU) -trnL (UAA) et trnL (UAA) -trnF (GAA), ainsi que l'intron trnL (UAA) dans la région trnT-F de L'ADN_{cp} ont été largement utilisés dans les études portant sur les relations phylogénétiques tant au niveau interspécifique qu'intraspécifique (Fujii *et al*., 1995, 1997 ; Böhle e*t al.*, 1996 ; Gielly *et al*., 1996 ; Bakker *et al*., 2000 ; Fukuda *et al*., 2001).

Les arbres phylogénétiques dérivés de l'ADN_{cp} ont repéré trois grands groupes : 1) taxons du sud de l'Asie (*L. griffithiana*, *L. mastersiana* et *L. potaninii*), 2) taxons Nord eurasiens (*L. decidua*, *L. gmelinii* et *L. kaempferi*), comprenant *L. sibirica* dans quelques études (Wei & Wang 2003) et 3) taxons nord-américains (*L. laricina*, *L. lyalii* et *L. occidentalis*) (Gros-Louis *et al*., 2005). Ces trois groupes correspondent à la dispersion géographique décrite par LePage & Basinger (1995). Qian *et al*., (1995) ont étudié les relations phylogénétiques entre huit espèces et trois variétés du genre *Larix* en analysant le polymorphisme de longueur des fragments de restriction dans l'ADN chloroplastique. Ils n'ont pas détecté des variations de l'ADN_{cp} au sein des taxons et les niveaux moyens de divergence des nucléotides entre les espèces étaient faibles. Ils ont classé *L. griffithiana* dans un groupe distinct, car elle présentait des différences génétiques marquées par rapport à tous les autres taxons. Un deuxième groupe était composé de *Larix sibirica* et deux espèces nord-américaines, *Larix laricina* et *Larix occidentalis*. Ils ont classé les taxons Chinois et japonais (*Larix gmelinii*, *Larix potaninii, Larix kaempferi*) avec le mélèze d'Europe (*Larix decidua*). Kim *et al*., (2018) ont étudié *L. kaempferi* et *L. olgensis*. Ils ont séquencé les génomes chloroplastiques et ont constaté que le génome de *L. kaempferi* mesurait 122,158 pb et avait deux régions répétées inversées de 436 pb chacune. Le génome de *L. olgensis* était légèrement plus long, mesurant 122,573 pb, et avait également deux régions répétées inversées de 436 pb. L'analyse phylogénétique a révélé que toutes les espèces échantillonnées de

Pinaceae formaient un clade monophylétique avec une valeur de bootstrap élevée. Le genre *Larix* est étroitement lié au genre Pseudotsuga. Dans le même cadre, la séquence complète du génome chloroplastique de *Larix potaninii* a été déterminée par Han *et al*., (2017). L'ADN_{cp} mesurait 122,492 pb et contenait une paire de régions répétées inversées (IR) de 435 pb. L'analyse phylogénétique basée sur 36 génomes des chloroplastes indique que *L. potaninii* var. *chinensis* est étroitement apparenté *à L. decidua*. Bondar *et al*., (2019) ont réalisé le séquençage, l'assemblage et l'annotation du génome chloroplastique de l'une des principales espèces de conifères de la forêt boréale de Sibérie, le mélèze de Sibérie. Les résultats ont révélé que la longueur du génome assemblé du chloroplaste était de 122,561 pb, ce qui est proche à celle du mélèze d'Europe (*Larix decidua* Mill.) dont le génome chloroplastique mesure 122,474 pb.

Également les recherches sur les génomes chloroplastiques permettent le développement des marqueurs moléculaires diagnostiques pour l'identification des hybrides, quel que soit leur stade de développement. À cet égard, on peut citer l'exemple de l'étude menée par Acheré *et al*., (2004) sur les mélèzes d'Europe et du Japon. Ils ont testé la combinaison de marqueurs hérités de la mère du génome mitochondrial (ADN_{mt}) et de marqueurs hérités du père du génome du chloroplaste. Les hybrides ont été identifiés ultérieurement par la présence d'une séquence mitochondriale héritée d'une espèce parentale et d'une séquence chloroplastique héritée de l'autre espèce parentale. Ces marqueurs ont été utilisés pour l'évaluation de la proportion d'hybrides dans un lot de semences issues de vergers à graines ; celle-ci a été évaluée entre 43% et 53% selon l'espèce parentale. Les espèces parentales mâles et femelles ont pu être déterminées pour chaque descendance. Guo *et al*., (2021) ont utilisé la technique de séquençage nouvelle génération pour séquencer les génomes chloroplastiques de cinq espèces de mélèzes (*L. grifthii*, *L. speciose*, *L. himalaica*, *L. kongboensis* et *L. potaninii* var. *australis*). Les résultats ont montré que les génomes sont conservateurs en ce qui concerne leur contenu génétique, leur taille et leur arrangement. Ils contenaient peu de séquences répétées simples avec une faible variabilité de nucléotides. Ainsi ces résultats seront utiles pour d'autres recherches sur la génétique de population.

Objectif général. L'objectif général de cette étude est d'assembler, et d'annoter les génomes chloroplastiques de sept mélèzes, ainsi que de clarifier leurs relations phylogénétiques à l'aide du séquençage de nouvelle génération.

Objectifs spécifiques

a. Séquencer et assembler des génomes chloroplastiques de mélèze.
b. Effectuer des annotations génomiques pour identifier les gènes, les régions codantes et non codantes, les séquences répétées et d'autres éléments fonctionnels dans les génomes chloroplastiques assemblés.
c. Effectuer des analyses statistiques et structurales comparatives pour évaluer le l'organisation des génomes chloroplastiques et le contenu en gènes.
d. Utiliser les variations dans les génomes chloroplastiques dans des analyses phylogénétiques pour différencier les espèces de mélèzes.

Hypothèses

- Le séquençage de l'ADN chloroplastique permettra de définir la trajectoire évolutive des génomes chloroplastiques, clarifiant ainsi les relations phylogénétiques entre les différentes espèces de mélèzes.
- Le séquençage de l'ADN chloroplastique permettra de révéler l'effet de la distribution géographique sur la structure phylogénétique et que les espèces les plus proches sont génétiquement plus similaires, indiquant une divergence évolutive influencée par la répartition géographique.

1. MANUSCRIPT

1.1 Abstract

The chloroplast genome (cpDNA) in conifers, including larch species, retains a paternal transmission via pollen and often used for evolutionary studies. In this study, the whole chloroplast genomes of seven *Larix* from North America (*L. laricina*, *L. occidentalis*) and North Asia (*L. sibirica*, *L. sukaczewii, L. gmelinii* var. *japonica*, *L. gmelinii* var. *olgensis*, *L. gmelinii* var. *kamchatica*) were analyzed by high-throughput sequencing techniques. The circular complete chloroplast genomes were 122,048 to 123,460 bp in length. They exhibit a typical quadripartite structure and conserved arrangement. We showed that each genome features a pair of inverted regions, with variability in the size of the large single copy region ranging from 59,760 bp to 76,699 bp, as well as in the size of the small single copy region which ranges from 43,183 bp to 59,760 bp. The total number of genes across species ranges from 95 to 111, encoding proteins, 4 ribosomal RNA genes and 30 to 34 transfer RNA genes. Overall GC (Guanine, Cytosine) content varies between 35.55% and 40.79%. While maintaining these similarities, variations emerged notably in simple sequence repeats, gene introns, and the presence of hypothetical genes. Phylogenetic analysis revealed close relationships between species, as well as genetic proximity among those geographically close. These findings provide crucial genetic insights vital for comprehending the evolution and conservation of the genus *Larix* and its related species.

Keywords : cpDNA, conifers, paternal transmission, larch, phylogeny, SSR, Illumina NovaSeq 6000.

1.2 Introduction

Larch (*Larix spp*.) are renowned for their adaptability to diverse ecological sites and northern climates, characterized by rapid juvenile growth (Isebrands & Hunt, 1975), they are prominent fast growers within forest ecosystems. Despite their ecological significance, our understanding of the genetic characteristics and phylogenetic relationships within larch forests remains largely unexplored.

The genus *Larix* exhibits a wide distribution across North America, Asia, and Europe, representing one of the primary boreal tree species that covers approximately 30% of all global forested lands (Volney & Fleming, 2000). Previous studies have been conducted to assess its genetic diversity and population structure using various markers and genomes. These methods included techniques such as PCR–restriction fragment length polymorphisms (RFLPs) of chloroplast DNA (Semerikov *et al*., 2003) and the analysis of the chloroplast trnT-trnF region (Wei & Wang, 2003). However, the emergence of next-generation sequencing (NGS) technology has significantly advanced research in this field, providing a robust method for comparative genomics studies.

NGS enables the rapid generation of highly accurate data, facilitating the identification of conserved regions, the dissection of gene content and the decryption of structural variations within the chloroplast DNA. Chloroplasts contain a genome that is distinct from the nuclear genome, which encodes for specific proteins (Olmstead and Palmer, 1994) and have a highly conserved genomic structure (Ohyama *et al*., 1986). Generally, ranging in size from 120,000 to 180,000 base pairs (bp), depending on the species. They encode for 4 ribosomal RNAs, 30 transfer RNAs and around 40 identified proteins (Ruhlman and Jansen, 2021). The chloroplast genome in conifers, including larch species, retains a paternal transmission via pollen (Hipkins *et al*., 1994). This characteristic allows the study of population differentiation and gene flow. Thus far, the chloroplast genome of numerous species of *Larix* has been sequenced; *L. decidua* (Wu *et al*., 2011), *L. potaninii var. chinensis* (Han *et al*., 2017), *L. gmelinii var. japonica* (Ishizuka *et al*., 2017), *L. potaninii var. macrocarpa* (Qiu *et al*., 2017), *L. sibirica*

(Bondar *et al*., 2019), *L. gmelinii* (Zimmermann *et al*., 2019), *L. Kaempferi* (Chen *et al*., 2020), *L. griffithii* (Guo *et al*., 2021), and *L. himalaica* (Guo *et al*., 2021). While the number of larch species with fully sequenced chloroplast DNA using NGS is increasing, there are still other species that have not been sequenced. *L. laricina* holds immense ecological and biological importance, yet it remains noteworthy that there is a scarcity of research on its chloroplast DNA.

To date, only a few studies have delved into this North American taxon based on chloroplast DNA polymorphisms. Notably, Warren *et al*., (2016) conducted a comprehensive analysis, utilizing molecular and fossil data, to explore the genetic relationships and distribution patterns of *Larix laricina*. Their findings unveiled genetically distinct lineages that emerged during the glacial period. Additional studies have been conducted on various Eurasian larch species such as *Larix decidua* (Wu *et al*., 2011), *L. Kaempferi* (Chen *et al*., 2020), *L. gmelinii* (Zimmermann *et al*., 2019). These studies revealed the total chloroplast DNA sequences, which provide valuable insights into the evolutionary relationships among these species. However, as of now, some species and varieties remain unsequenced. Araki *et al*., (2008) conducted a study to assess the genetic divergence between populations of *Larix sibirica* in western Russia and *Larix sukaczewii* populations situated west of the Irtysh and Ob rivers. Their results demonstrated a significant level of divergence (Fixation index (F_{ST}) = 0.531) between the two species based on analyzing nucleotide variations at three nuclear genes. This contrasts with other studies considering these populations as a single species (Kullman 1998; Semerikov *et al*., 1999, Wei & Wang, 2003). Their findings provided partial support for the hypothesis of taxonomic distinction between these two taxa. However, sequencing the entire chloroplast genome of *Larix sukaczewii* is essential to reinforce this classification. Currently, our comprehension of the intraspecific divergence of *Larix gmelinii* is limited, as there have been relatively few studies conducted on this topic. Chen *et al*., (2020) demonstrated that the *L. gmelinii* var. *japonica* and *L. gmelinii* var. *olgensis* form a monophyletic group. Thus, sequencing additional *Larix gmelini* subspecies, including the Northeast Russian; *Larix gmelinii* var. *kamchatica* would provide a broader representation of

its genetic diversity. In the current study, we describe for the first time the whole chloroplast genome sequence of *Larix sukaczewii, Larix laricina, Larix gmelinii. var. kamchatica and Larix gmelinii. var. olgensis.* Furthermore, we analyze their phylogenetic status.

1. 3 Material and Methods

1.3.1 Plant materials

In October 2019, fresh needle leaves were collected from a total of seven larch trees aged 2 years and cultivated in the greenhouse. These selected trees were chosen to represent distinct geographical regions (North Asia and North America) as illustrated in (Figure 5). Detailed accession numbers for each tree are provided in (Table 1). To preserve the integrity of the collected samples, they were immediately stored at a temperature of -80 °C until DNA extraction was performed.

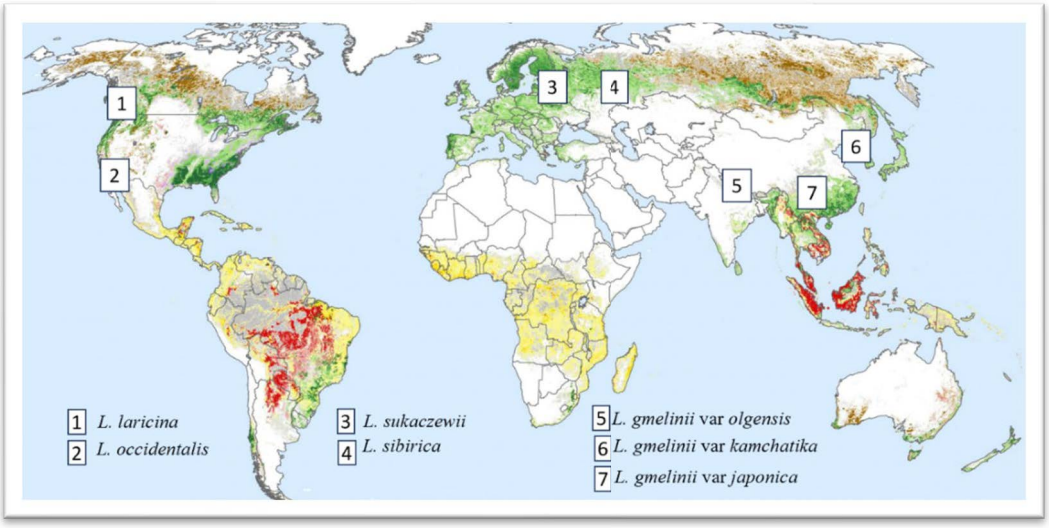


Figure 5
Study site locations of larch species from North America and Asia

Table 1
Species' names and their accession number

Species name	Accession number
L. gmelini japonica	7415
L. gmelini kamchatica	7477
L. gmelini olgensis	4671
L. laricina	8220092.0
L. occidentalis	20007432.0
L. sibirica	4431
L. sukaczewi	7072

1.3.2 DNA extraction and genome sequencing

The samples were shipped to Bio S&T *Inc*. (Montreal, QC, Canada) for chloroplast DNA extraction. Briefly, to maintain sample integrity and prevent contamination with nuclear and mitochondrial DNA, chloroplasts were isolated from 60 g of tissue to obtain 10 µg of high-quality purified chloroplast DNA per sample. Bio S&T employed an improved method involving liquid nitrogen sucrose (LN) density gradient centrifugation. All procedures were performed at 4 °C, and all centrifugation was performed in a CP80NX ultracentrifuge (Hitachi). Extracted chloroplast DNA was sent to Genome Quebec (Montreal, Canada) for sequencing on an Illumina NovaSeq 6000 platform.

1.3.3 DNA quality assessment

1.3.3.1 DNA samples quality assessment

A quantity and quality assessment of the extracted DNA was carried out for each sample. This assessment involved two main steps; first, the concentration of chloroplast DNA in each sample was determined using a ND-2000 spectrometer from Nanodrop Technologies in Wilmington, DE, USA. The second visual examination was performed using gel electrophoresis.

1.3.3.2 DNA sequences quality assessment

Before proceeding with chloroplast genome assembly, the data quality was verified using the FastQ software (Andrew, 2010). The raw sequencing reads were trimmed with an average quality Q5 to remove all the containing ambiguous bases and adapters using Trimmomatic v.0.3.2 (Bolger *et al*., 2014). This refined quality control process ensured that only high-quality data were used for subsequent chloroplast genome assembly steps.

1.3.4 Genome annotation and sequence alignment

To assemble the genome sequences, the Bowtie2 software (Langmead and Salzberg, 2012) was used to map the short reads to the available complete chloroplast genome sequences of *Larix occidentalis* Nutt. (NCBI Genbank accession numbers A FJ899578.1). Second, the aligned reads were assembled by SPAdes software (Bankevich et *al*., 2012). The annotation of the obtained chloroplast genome was performed using GeSeq programs (Tillich *et al*., 2017). The structural features of chloroplast genome maps were drawn using Organellar Genome DRAW (Lohse *et al*., 2013), CPGAVAS2 (Shi *et al*., 2019) and GCPGVIEW (Liu *et al*., 2023).

1.3.5 Analysis of simple sequence repeats

To identify simple sequence repeats (SSRs) in the chloroplast genome, we used the MIcroSAtellite (MISA-web) identification tool developed by Beier *et al*., (2017). This web-based tool is specifically designed for the detection and characterization of microsatellites, including SSRs. The search parameters used in MISA-web have been defined as follows: “1–10 2–6 3–5 4–5 –65”.

These parameters define the minimum and maximum lengths of the SSR patterns that were considered during the analysis. The numbers before the dashes represent the minimum repeat unit length, while the numbers after the dashes represent the maximum repeat unit length.

1.3.6 Construction of the phylogenetic tree

To infer the phylogenetic relationships among the sequenced genomes, we evaluated the accuracy of different methods. Initially, we compared the neighbor-

joining method (NJ) (Saitou and Nei, 1987) and the unweighted pair group method with arithmetic mean (UPGMA) (Sokal and Michener, 1958) in RStudio with the ape (Paradis *et al.*, 2004), phangorn (Schliep, 2011), and adegenet (Jombart, 2008) packages. Subsequently, the selected method was compared to the maximum-likelihood method (ML) (Felsenstein, 1981) using an ANOVA test. Additionally, we calculated the Akaike Information Criterion (AIC) for each method. Multiple alignments were conducted using MAFFT version 7.222 (Katoh and Standley, 2011) with default parameters and the model was chosen based on MEGA 11 test (Tamura *et al*., 2021). Models with the lowest BIC scores (Bayesian Information Criterion) are considered to describe the substitution pattern the best.

1.4 Results

1.4.1 DNA sequence quality assessment

DNA samples exhibited significant levels of concentration, ranging from 1.53 ng/ul to 2.80 ng/µl. These results imply that the DNA sequences obtained from these samples are of high quality and they can be used for further analysis.

For the quality check, FastQ analysis revealed that the obtained sequences were primarily situated within the green range, with a slight extension into the orange region. This indicates the absence of systematic errors or biases in the sequencing data, as supported by a substantial average score of 35 (Annexe A).

In general, all the modules in FastQC indicate entirely normal results, suggesting that the sequencing data meets the expected quality standards. However, to further enhance the data quality, we employed Trimmomatic for raw reads trimming and adapters elimination (Annnexe B).

1.4.2 Genome annotation and sequence alignment

1.4.2.1 Characteristics of *Larix sibirica* chloroplast genome

The NovaSeq platform generated a total of 71,455,531 reads. Subsequent assembly of the Siberian larch chloroplast genome revealed a final length of 122,595 bp with a GC content of 38.74%.

For the chloroplast genome annotation, we conducted a comparative analysis with available data for *L. occidentalis*, identifying a total of 111 genes, comprising 34 tRNA genes, 4 rRNA genes, and 73 protein-coding genes. Three hypothetical chloroplast reading frames were identified (*ycf1*, *ycf2*, *ycf4*) (Table 2). The small single-copy (SSC) region measures 53,935 bp, while the long single-copy region (LSC) covers 64,010 bp. A gene map depicting the arrangement of these genes within the chloroplast genome was generated using CPGAVAS2 (Figure 6).

Figure 6
Visualization of features identified in the *L. sibirica* chloroplast genome. The map has four rings, from the center going outward. The first circle shows the forward and reverse repeats connected with red and green arcs, respectively. The second shows the tandem repeats marked with dashes. The third circle shows the microsatellite sequences identified using MISA. The fourth circle shows the gene structure of the chloroplast genome. The colors of these genes are classified according to their function, as shown in the lower left

Among the protein-coding genes, six (*rps2*, *rps4*, *rps7*, *rps14*, *rps15*, *rps18*) encode small ribosomal subunit proteins, and seven (*rpl2*, *rpl14*, *rpl22*, *rpl23*, *rpl32*, *rpl33*, *rpl36*) encode large ribosomal subunit proteins. Additionally, 25 genes are related to photosynthesis proteins, three (*rpoA*, *rpoB*, *rpoC*) encode DNA-dependent RNA polymerase, and four (*accD*, *ccsA*, *cemA*, *matK*) encode other proteins.

Furthermore, thirteen genes containing one intron, and two exons were identified. The *trnK* gene was found to have the largest intron (70,394 bp), while the *trnV* gene had the smallest intron (543 bp) (Table 3). *rpl2*, *PetB*, *PetD*, *ycf3*, *psb* are recognized as cis-splicing genes (Figure 7).

Table 2
Gene composition in *L. sibirica* chloroplast genome

Category	Group of genes	Name of genes
Self-replication	Large subunit of ribosomal proteins	*rpl2, rpl14, rpl22, rpl23, rpl32, rpl33, rpl36*
	Small subunit of ribosomal proteins	*rps14, rps15, rps18, rps2, rps4, rps7*
	DNA-dependent RNA polymerase	*rpoA, rpoB, rpoC1*
	rRNA genes	*rrn16S, rrn23S, rrn4S, rrn5S*
	tRNA genes	*trnL-CAA, trnL-UAG, trnS-GGA, trnD-GUC, trnF-GAA, trnV-GAC, trnG-UCC, trnRCCG, trnM-CAU, trnY-GUA, trnE-UUC, trnN-GUU, trnR-UCU, trnL-UAA, trnT-UGU, trnT-GGU, trnI-GAU, trnA-UGC, trnG-GCC, trnfM-CAU, trnR-ACG, trnS-UGA, trnP-UGG, trnW-CCA, trnA-UGC, trnQ-UUG,*
Photosynthesis	Photosystem I	*psaA, psaC*
	Photosystem II	*psbA, psbB, psbC, psbD, psbE, psbF, psbI, psbK, psbL, psbM, psbN, psbT psbZ, ycf3*
	Cytochrome b6/ f complex	*petA, petB, petD, petN*
	ATP synthase	*atpA, atpB, atpE, atpF, atpH, atpI*
	Rubisco	*rbcl*
Other genes	Maturase	*matK*
	Protease	*clpP*
	Envelope membrane protein	*cemA*

	Subunit acetyl-CoA-carboxylase	*accD*
Unkown conserved open reading frames		*ycf1, ycf2, ycf4*

Table 3
Lengths of introns and exons in split genes of Siberian larch

Gene	Strand	Start	End	ExonI	IntronI	ExonII
rpl2	-	26061	27369	198	682	429
trnK	-	33596	104061	37	70394	35
trnL	-	37988	86686	50	48612	37
trnV	+	46759	47375	39	543	35
PetB	+	56118	57636	5	938	576
PetD	+	57851	59099	7	712	530
atpF	+	63009	64334	145	771	410
ycf3	+	81504	84055	228	2168	156
trnI	+	92842	93907	37	994	35
psbD	+	91808	97506	815	4637	247
psbC	+	97454	107567	139	8716	1259
trnA	+	93987	109504	38	15442	38
rpl2	-	120739	122076	402	672	264

Figure 7
Schematic map of the cis-splicing genes in the Siberian chloroplast genome.

The genes are arranged from top to bottom based on their order on the chloroplast genome. The gene names are shown on the left, and the gene structures are on the right. The exons are shown in black; the introns are shown in white. The arrow indicates the direction

Two distinct categories of SSR loci were detected within the chloroplast genome. Specifically, a total of 17 loci were identified to possess mononucleotide repeat motifs, while 2 loci were found with dinucleotide repeat motifs. The search parameters employed did not yield any SSR loci featuring trinucleotide, tetranucleotide, pentanucleotide, or hexanucleotide repeat motifs (Table 4).

Table 4
Simple sequence repeats in *Larix sibirica*

Start nucleotide position	Motif and number of repeats	Location
5514	(A)10	Non-coding region between *rbcl* and *atpH* genes
7422	(T)15	Non-coding region between *trnC-GCA* and trnT-GGU
18229	(T)10	Non-coding region between *ccsA* and *rps4* genes
20414	(T)10	Non-coding region between *trnS-GGA* and *rps2* genes
38019	(A)12	Non-coding region between *trnF-GAA* and *ycf12* genes
39107	(T)12	Non-coding region between *trnV-GAC* and *accD* genes
45271	(T)12	Non-coding region between *clpP* and *trnR-CCG* genes
69816	(T)13	Non-coding region between *clpP* and *trnY-GUA* genes
72638	(A)11	Non-coding region between *trnE-UUC* and *rpoB* genes
72777	(T)16	Non-coding region between *trnE-UUC* and *rpoB* genes
78114	(A)11	trnS-CGA gene
78637	(A)11	Non-coding region between trnS-CGA and trnR-UCU genes
80313	(A)11	Non-coding region between *atpA* and *ycf3* genes
84347	(AT)7	Non-coding region between *rpl36* and *psaA* genes

94738	(AT)7	Non-coding region between *trnE-UUC* and *tRNA-Glu* genes
98959	(T)14	Non-coding region between rps19 and *trnR-ACG* genes
102941	(C)10	*trnV* gene
111637	(C)11	*ycf4* gene
114296	(T)15	Non-coding region between *petA* and *psB* genes

1.4.2.2 Characteristics of *Larix sukaczewii* DyL. chloroplast genome

The NovaSeq platform generated a total of 56,716,387 reads. The total length of *Larix sukaczewii* chloroplast genome assembly was 122,048 bp with 40.54% GC content. The annotation through comparison with the available data for *L. sibirica* identified 108 genes, from which 32 represented *tRNA* genes, 4 *rRNA*, and 72 protein-coding genes. A gene map of the genome was generated using CPG and is presented in (Figure 8). The small single-copy region measures 58,102 bp, while the long single-copy region covers 59,760 bp. Seven genes (*rps15, rps18, rps19, rps2, rps3, rps4, rps7*) are responsible for coding small ribosomal subunits, while six genes (*rpl14, rpl2, rpl23, rpl32, rpl33, rpl36*) encode large ribosomal subunits. Additionally, a group of 24 genes are associated with photosynthesis-related proteins, while three genes (*rpoA, rpoB, rpoC1*) code for DNA-dependent RNA polymerase. Another set of four genes (*accD, ccsA, cemA, matK*) contributes to the synthesis of various other proteins. Two hypothetical reading frames were identified (*ycf3*, *ycf4*) (Table 5).

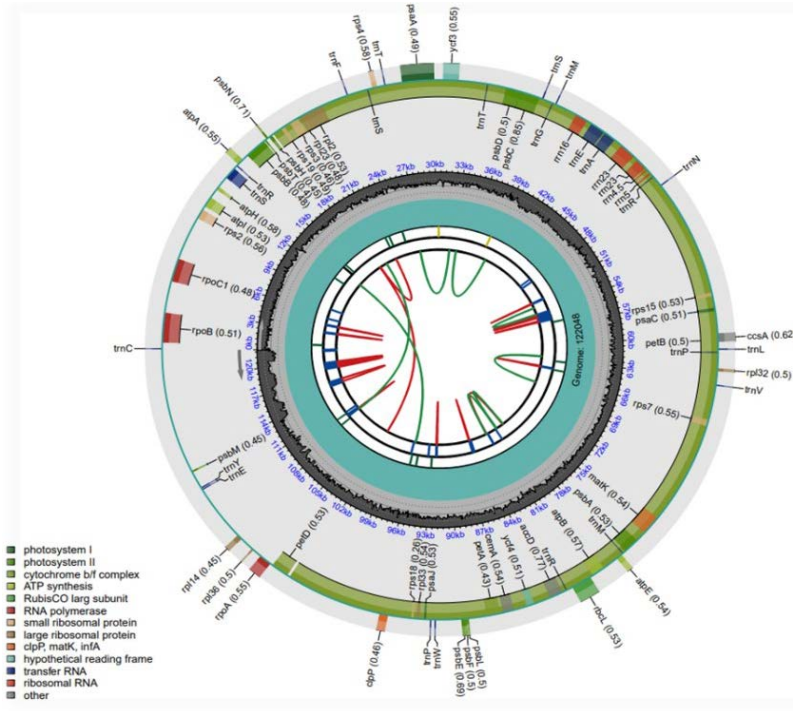


Figure 8

Visualization of features identified in *Larix sukazewii* chloroplast genome using CPGVIEW. The map contains six tracks in default. From the center outward, the first track shows the dispersed repeats. The dispersed repeats consist of direct (D) and Palindromic (P) repeats, connected with red and green arcs. The second track shows the long tandem repeats as short blue bars. The third track shows the short tandem repeats or microsatellite sequences as short bars with different colors. The colors, the type of repeat they represent, and the description of the repeat types are as follows. Black: c (complex repeat); Green: p1 (repeat unit size = 1); Yellow: p2 (repeat unit size = 2); Purple: p3 (repeat unit size = 3); Blue: p4 (repeat unit size = 4); Orange: p5 (repeat unit size = 5); Red: p6 (repeat unit size = 6). The small single-copy (SSC), inverted repeat (IRa and IRb), and large single-copy (LSC) regions are shown on the fourth track. The base frequency at each site along the genome will be shown between the fourth and fifth tracks. The genes are shown on the sixth track. The optional codon usage bias is displayed in the parenthesis after the gene name. Genes are color-coded by their functional classification. The transcription directions for the inner and outer genes are clockwise and anticlockwise, respectively. The functional classification of the genes is shown in the bottom left corner

Table 5
Gene composition in *Larix sukazewii* chloroplast genome

Category	Group of genes	Name of genes
Self-replication	Large subunit of ribosomal proteins	*rpl14, rpl2, rpl23, rpl32, rpl33, rpl36,*
	Small subunit of ribosomal proteins	*rps15, rps18, rps19, rps2, rps3, rps4, rps7*
	DNA-dependent RNA polymerase	*rpoA, rpoB, rpoC1*
	rRNA genes	*rrn16S, rrn23S, rrn4S, rrn5S*
	tRNA genes	*trnC-GCA, trnR-UCU, trnF-GAA, trnS-GGA, trnT-UGU, trnT-GGU, trnS-UGA, trnG-GCC, trnA-UGC, trnR-ACG, trnN-GUU, trnL-UAG, trnP-GGG, trnV-GAC, trnM-CAU, trnR-CCG, trnW-CCA, trnP-UGG, trnE-UUC, trnY-GUA*
Photosynthesis	Photosystem I	*psbA, psbC, psbJ*
	Photosystem II	*psbA, B, C, D, E, F, L, M, N, T, ycf3*
	Cytochrome b6/f complex	*petA, petB, petD,*
	ATP synthase	*atpA, atpB, atpE, atpF, atpH, atpI*
	Rubisco	*Rbcl*
Other genes	Maturase	*matK*
	Protease	*clpP*
	Envelope membrane protein	*cemA*
	Subunit acetyl-CoA-carboxylase	*accD*
Unkown conserved open reading frames		*ycf4*

Furthermore, the chloroplast genome contains 7 genes that exhibit one intron and two exons. Notably, the *petB* gene consists of the largest intron, spanning 84,034 base pairs, while the rpl2 gene possesses the smallest intron, measuring 685 base

pairs (Table 6). Specific genes such as *rpl2*, *PetD*, *ycf3*, are recognized as cis-splicing genes, as showing in (Figure 9).

Table 6
Lengths of introns and exons in split genes of *Larix sukazewii*

Gene	Strand	Start	End	ExonI	IntronI	ExonII
trnS-CGA	+	13455	14321	32	775	60
rpl2	+	20487	22002	400	685	431
ycf3	-	30719	31817	228	715	156
trnE-UUC	+	41708	42776	33	994	42
trnA-UGC	+	42854	43701	37	775	36
petB	+	18022	102637	6	84034	576
petD	+	102852	104100	8	712	529

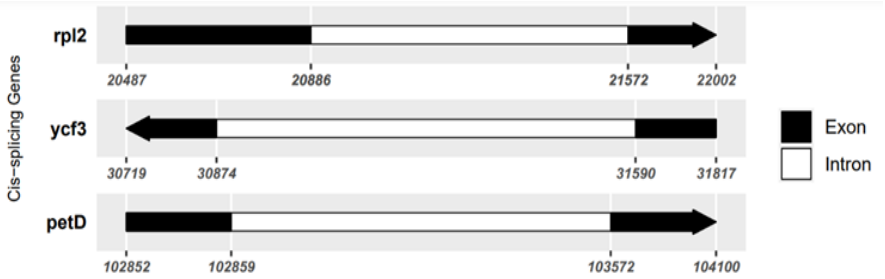


Figure 9
Schematic map of the cis-splicing genes in the chloroplast genome of *Larix sukazewii*

In addition to the other findings, our research identified a total of 17 SSR loci. Out of these loci, 15 were characterized by mononucleotide repeat motifs, while the remaining 2 exhibited dinucleotide repeat motifs (Table 7). No SSR loci with trinucleotide, tetranucleotide, pentanucleotide, or hexanucleotide repeat motifs were detected in the genome.

Table 7
Simple sequence repeats in *Larix sukazewii*

Start nucleotide position	Motif and number of repeats	Location
362	(T)16	Non-coding region between *rbcl* and *atpH* genes
11856	(A)10	Non-coding region between *trnC-GCA* and *trnT-GGU*
14013	(A)11	Non-coding region between *ccsA* and *rps4* genes
14536	(A)10	Non-coding region between trnS-*GGA* and *rps2* genes
22021	(A)14	Non-coding region between trnF-*GAA* and *ycf12* genes
22649	(G)11	Non-coding region between *trnV-GAC* and *accD* genes
24800	(A)10	Non-coding region between *clpP* and *trnR-CCG* genes
30414	(AT)7	Non-coding region between clpP and *trnY-GUA* genes
38549	(AT)7	Non-coding region between *trnE-UUC* and *rpoB* genes
56558	(A)13	Non-coding region between *trnE-UUC* and *rpoB* genes
63945	(A)13	*trnS-CGA* gene
65025	(A)11	Non-coding region *between trnS-CGA* and *trnR-UCU* genes
92776	(G)10	Non-coding region between atpA and *ycf3* genes
95101	(T)17	Non-coding region between *rpl36* and *psaA* genes
96067	(T)10	Non-coding region between *trnE-UUC* and *tRNA-Glu* genes
108474	(T)11	Non-coding region between *rps19* and *trnR-ACG* genes
109517	(T)11	*atpI* gene

1.4.2.3 Characteristics of *Larix gmelinii var. japonica* chloroplast genome

The NovaSeq platform generated a total of 62,308,733 reads. The total length of *Larix japonica* chloroplast genome was 122,339 bp.

The GC content was evaluated by 35.55%. The annotation through comparison with the available data for *L. sibirica* identified 108 genes, from which 34 represented *tRNA* genes, 4 *rRNA*, and 70 protein-coding genes. A gene map was generated using CPGVIEW and is presented in (Figure 10).

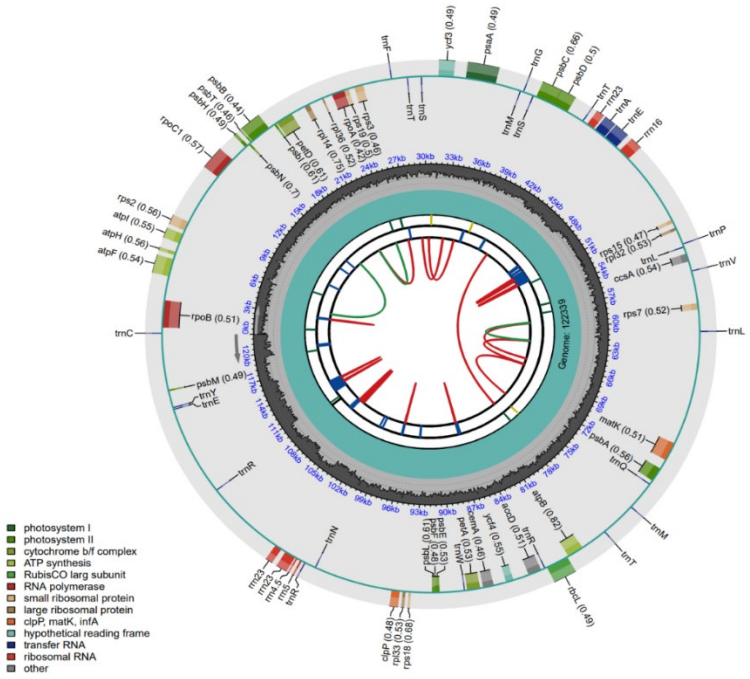


Figure 10
Schematic map of overall features of *Larix gmelinii* var. *japonica* chloroplast genome

The long single-copy region measures 62,102 bp, while the small single-copy region covers 55,760 bp. Among the protein-coding genes, eleven genes (*rps2, rps3, rps4, rps7, rps8, rps11, rps14, rps15, rps16, rps18, rps19*) code for small ribosomal subunits, while eight genes (*rpl2, rpl14, rpl16, rpl20, rpl22, rpl32, rpl33, rpl36*) code for for large ribosomal subunit proteins. Additionally, thirty genes related to photosynthesis proteins were identified, four genes (*rpoA, rpoB, rpoC1, rpoC2*) encoding DNA-dependent RNA polymerase, and four genes *(accD, ccsA,*

cemA, matK) responsible for other proteins. Two hypothetical reading frames were found (*ycf2*, *ycf4*) (Table 8).

Table 8
Gene composition in *L. gmelinii* var. *japonica* chloroplast genome

Category	Group of genes	Name of genes
Self-replication	Large subunit of ribosomal proteins	*rpl14, rpl16, rpl2, rpl20, rpl22, rpl32, rpl33, rpl33, rpl36*
	Small subunit of ribosomal proteins	*rps11, rps14, rps15, rps15, rps16, rps18, rps19, rps19, rps2, rps3, rps3, rps4, rps7, rps7, rps8*
	DNA-dependent RNA polymerase	*rpoA, rpoB, rpoC1, rpoC2*
	rRNA genes	*rrn16S, rrn23S, rrn4.5S, rrn5S*
	tRNA genes	*trnC, trnT, trnL, trnV, trnR, trnM trnY, trnE, trnN, trnT, trnI, trnA, trnG trnfM, trnS, trnP, trnW, trnQ*
Photosynthesis	Photosystem I	*PsaA, psaB, psaC, psaI, psaJ*
	Photosystem II	*psbA, psbB, psbC, psbD, psbE, psbF, psbI, psbJ, psbK, psbL, psbM, psbN, psbT, psbZ, ycf3*
	Cytochrome b6/f complex	*petA, petB, petD, petG, petL, petN*
	ATP synthase	*atpA, atpB, atpE, atpF, atpH, atpI*
	Rubisco	*rbcl*
Other genes	Maturase	*matK*
	Protease	*clpP*
	Envelope membrane protein	*cemA*
	Subunit acetyl-CoA-carboxylase	*accD*
Unkown conserved open reading frames		*ycf2*, *ycf4*

Furthermore, twelve genes are containing one intron and only one gene (*ycf3)* is containing two introns. The *trnK* gene was found to have the largest intron (2702 bp), while the *trnL* gene had the smallest intron (582 bp) (Table 9). *ycf3, atpf, rpoC1, rpl33, PetB, PetD*, *psb* are recognized as cis-splicing genes (Figure 11).

Table 9
Gene composition in *L. gmelinii* var. *japonica* chloroplast genome

Gene	Strand	Start	End	ExonI	IntronI	ExonII	IntronII
trnA-UGC	-	24884	25777	38	821	35	
trnI-GAU	-	25 837	26 829	37	921	35	
rpl2	+	42754	44156	334	641	428	
trnK-UUU	-	50198	52971	37	2702	35	
ycf3	-	55367	57387	124	760	230	748
atpF	-	60743	62136	145	842	407	
rpoC1	-	71609	74351	433	664	1646	
trnL-UAA	+	94571	95244	33	582	59	
trnV-UAC	-	98556	99283	37	643	48	
trnG-GCC	-	101506	102294	24	717	48	
rpl33	-	117960	118191	116	85	31	
petB	+	124835	126281	6	799	642	
petD	+	126478	127770	8	810	475	
rpl16	-	147782	149241	9	1052	399	

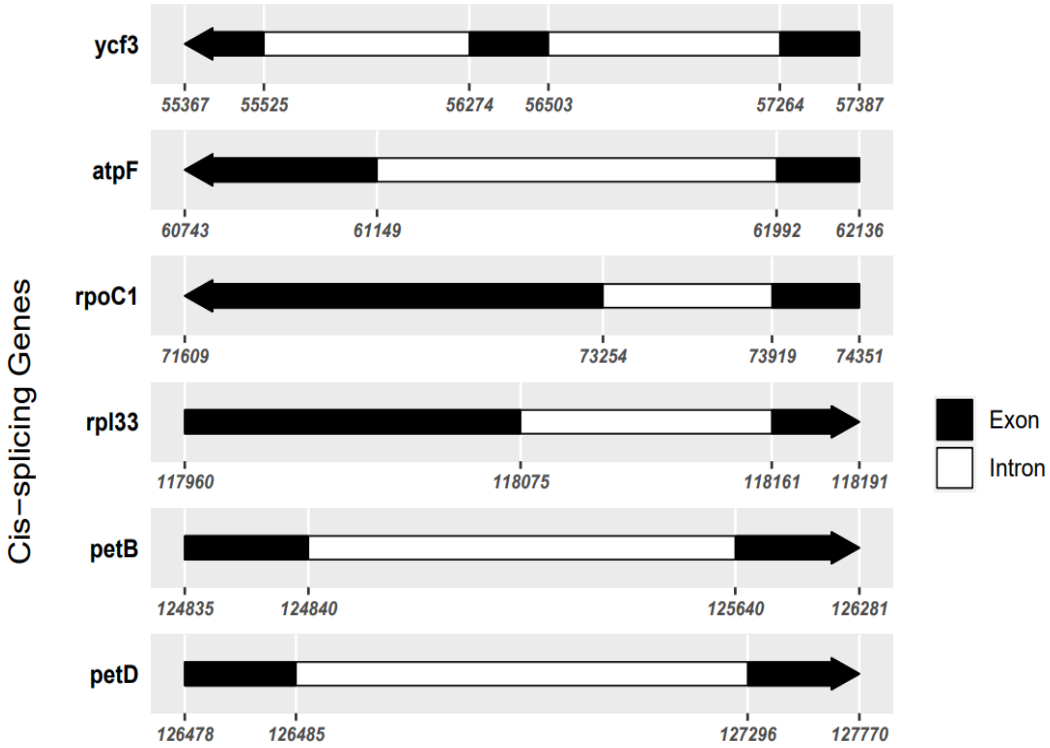


Figure 11
Schematic map of the cis-splicing genes in *L. gmelinii* var. *japonica* chloroplast genome.

We identified a total of 14 SSR loci. Out of these loci, eleven were characterized by mononucleotide repeat motifs, while three exhibited dinucleotide repeat motifs (Table 10). No SSR loci with trinucleotide, tetranucleotide, pentanucleotide, or hexanucleotide repeat motifs were detected in the genome.

Table 10
Schematic map of the cis-splicing genes in *L. gmelinii* var. *japonica* chloroplast genome

Start nucleotide position	Motif and number of repeats	Location
362	(T)14	Non-coding region between *rbcl* and *atpH* genes
6810	(T)11	Non-coding region between *trnC-GCA* and *trn-GGU*
24192	(A)14	Non-coding region between *ccsA* and *rps4* genes
25944	(A)11	Non-coding region between *trnS-GGA* and *rps2* genes
30838	(AT)6	Non-coding region between *trnF-GAA* and *ycf12* genes
37481	(AT)6	Non-coding region between *trnV-GAC* and *accD* genes
51899	(A)15	Non-coding region between *clpP* and *trnR-CCG* genes
56267	(A)12	Non-coding region between clpP and *trnY-GUA* genes
57719	(A)10	Non-coding region between trnE-UUC and rpoB genes
76286	(TA)8	Non-coding region between *trnE-UUC* and *rpoB* genes
78015	(T)10	*trnS-CGA* gene
109143	(A)11	Non-coding region between *trnS-CGA* and *trnR-UCU* genes
109665	(A)10	Non-coding region between atpA and *ycf3* genes
118979	(T)11	Non-coding region between *rpl36* and *psaA* genes

1.4.2.4 Characteristics of *Larix gmelinii var. kamchatica* chloroplast genome

The NovaSeq platform generated a total of 63,146,993 reads. Subsequent assembly of the chloroplast genome revealed a final length of 123,351 bp and GC content of 49.79%. It is composed of two short IR (IRA, IRB) regions, a small copy and a large copy region. The length of SSC and LSC region was 64,175 bp and 58,699 bp respectively. The annotation through comparison with the available data for *L.*

sibirica identified 95 genes, from which 30 represented *tRNA* genes, 4 *rRNA*, and 63 protein-coding genes. A gene map of the genome was generated using OGDRAW and is presented in (Figure 12).

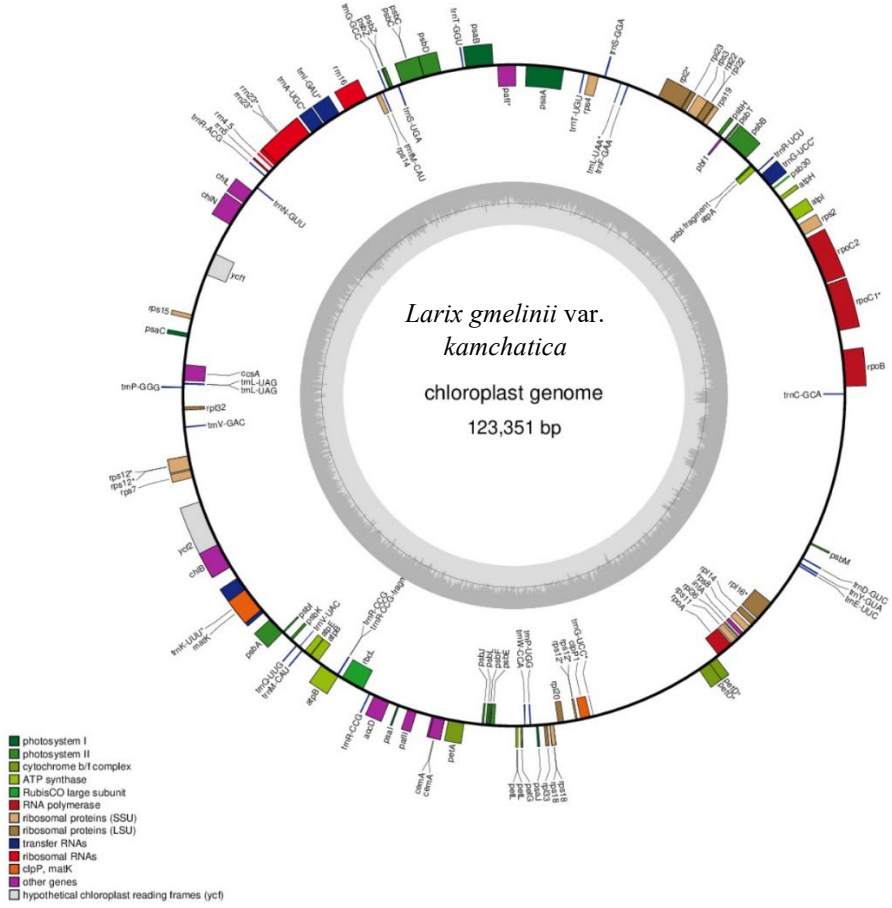


Figure 12
Schematic map of overall features of *Larix gmelinii* var. *kamchatica* chloroplast genome

Among the protein-coding genes, twelve (*rps2, rps3, rps4, rps7, rps8, rps11, rps12, rps14, rps15, rps16, rps18, rps19*) encode small ribosomal subunits, while nine (*rpl2*, *rpl14, rpl16, rpl20, rpl22, rpl23, rpl32, rpl33, rpl36*) encode large ribosomal subunits. Additionally, there are thirty-three genes associated with photosynthesis proteins, four genes (*rpoA, rpoB, rpoC1, rpoC2*) responsible for DNA-dependent RNA polymerase, and four genes (*accD, ccsA, cemA, matK,*) is responsible for other proteins (Table 11). Furthermore, five genes were identified to contain one

intron and two exons. Notably, the trnE-UUC gene possesses the largest intron at 994 bp while the *ycf3* gene has the smallest intron, measuring 715 base pairs (Table 12). Additionally, *atpf, rpl16, pafl* are classified as cis-splicing genes (Figure 13). Two hypothetical chloroplast reading frames were identified (*ycf2*, *ycf4*).

Table 11
Gene composition in *Larix gmelinii* var. *kamchatica* chloroplast genome

Category	Group of genes	Name of genes
Self-replication	Large subunit of ribosomal proteins	*rpl14, rpl16, rpl2, rpl20, rpl22, rpl23, rpl32, rpl33, rpl36*
	Small subunit of ribosomal proteins	*rps11, rps12, rps14, rps15, rps16, rps18, rps19, rps2, rps3, rps4, rps7, rps8*
	DNA-dependent RNA polymerase	*rpoA, rpoB, rpoC1, rpoC2*
	rRNA genes	*rrn16, rrn23, rrn4.5, rrn5*
	tRNA genes	*trnD-GUC, trnR-UCU, trnG-UCC, trnS-GCU, trnI-CAU, trnF-GAA, trnL-UAA, trnT-UGU, trnS-GGA, trnG-GCC, trnS-UGA, trnT-GGU, trnI-GAU, trnA-UGC, trnR-ACG,trnN-GUU, trnL-UAG, trnP-GGG, trnV-GAC, trnL-CAA, trnH-GUG, trnK-UUU, trnQ-UUG, trnS-GCU, trnV-UAC, trnM-CAU, trnR-CCG, trnW-CCA, trnP-UGG, trnE-UUC, trnY-GUA*
Photosynthesis	Photosystem I	*psaA, psaB, psaC, psaJ*
	Photosystem II	*psbA, psbB, psbC, psbD, psbE, psbF, psbI, psbJ, psbK, psbL, psbM, psbN, psbT, psbZ, ycf3*
	Cytochrome b6/f complex	*petA, petB, petD, petG, petL, petN*
	ATP synthase	*atpA, atpB, atpE, atpF, atpH, atpI*
	Rubisco	*rbcl*
	Protease	*clpP*
Other genes	Maturase	*matK*
	Protease	*clpP*
	Envelope membrane protein	*cemA*

	Subunit acetyl-CoA-carboxylase	*accD*
Unkown conserved open reading frames		*ycf2*, *ycf4*

Table 12
Lengths of introns and exons in split genes of *Larix gmelinii* var. *kamchatica*

Gene	Strand	Start	End	ExonI	IntronI	ExonII
atpF	-	5206	6526	145	766	410
trnS-CGA	+	16192	17058	32	775	60
ycf3	-	32167	33265	228	715	156
trnE-UUC	+	41765	42833	33	994	42
trnA-UGC	+	42911	43758	37	775	36

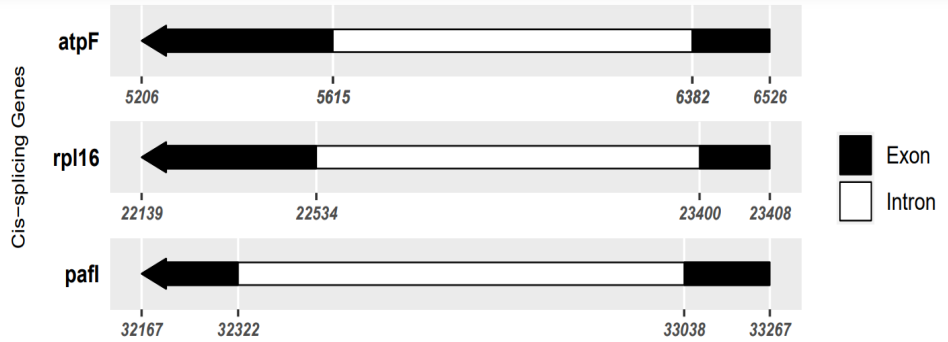


Figure 13
Schematic map of the cis-splicing genes in *Larix gmelinii* var. *kamchatica* chloroplast genome

We identified a total of 14 SSR loci. Out of these loci, eleven were characterized by mononucleotide repeat motifs, while three exhibited dinucleotide repeat motifs (Table 13). No SSR loci with trinucleotide, tetranucleotide, pentanucleotide, or hexanucleotide repeat motifs were detected in the genome.

Table 13
Simple sequence repeats in *Larix gmelinii* var. *kamchatica.*

Start nucleotide position	Motif and number of repeats	Location
2179	(A)14	Non-coding region between *rbcl* and *atpH* genes
7589	(T)10	Non-coding region between *trnC-GCA* and *trnT-GGU*
16749	(A)10	Non-coding region between *ccsA* and *rps4* genes
23153	(T)11	Non-coding region between *trnS-GGA* and *rps2* genes
24829	(G)12	Non-coding region between *trnF-GAA* and *ycf12* genes
31846	(AT)6	Non-coding region between *trnV-GAC* and *accD genes*
34740	(AT)6	Non-coding region between clpP and *trnR-CCG* genes
56684	(T)10	Non-coding region between clpP and *trnY-GUA* genes
68596	(G)10	Non-coding region between *trnE-UUC* and *rpoB* genes
71777	(T)10	Non-coding region between *trnE-UUC* and *rpoB* genes
72670	(A)10	*trnS-CGA* gene
73614	(TA)8	Non-coding region between *trnS-CGA* and *trnR-UCU* genes
79232	(A)13	Non-coding region between *atpA* and *ycf3* genes
81560	(C)10	Non-coding region between *rpl36* and *psaA* genes

1.4.2.5 Characteristics of *Larix gmelinii var. olgensis* chloroplast genome

The draft chloroplast genome of *Larix gmelinii* var. *olgensis* is circular with 123,160 bp length and GC content of 39.50%. It is composed of two short IR (IRA, IRB) regions, a small copy (SSC) and a large copy (SSC) region. The length of SSC and LSC region was 43,183 bp and 76,699 bp respectively (Figure 14). The cp genome contains 108 genes comprise 30 transfer RNA (*tRNA*) genes, 4 ribosomal RNA (*rRNA)* genes, and 74 coding genes. Most genes occur as a single copy, whereas two genes were found to be duplicated: *trnS-GGA* and *trnT-GGU*.

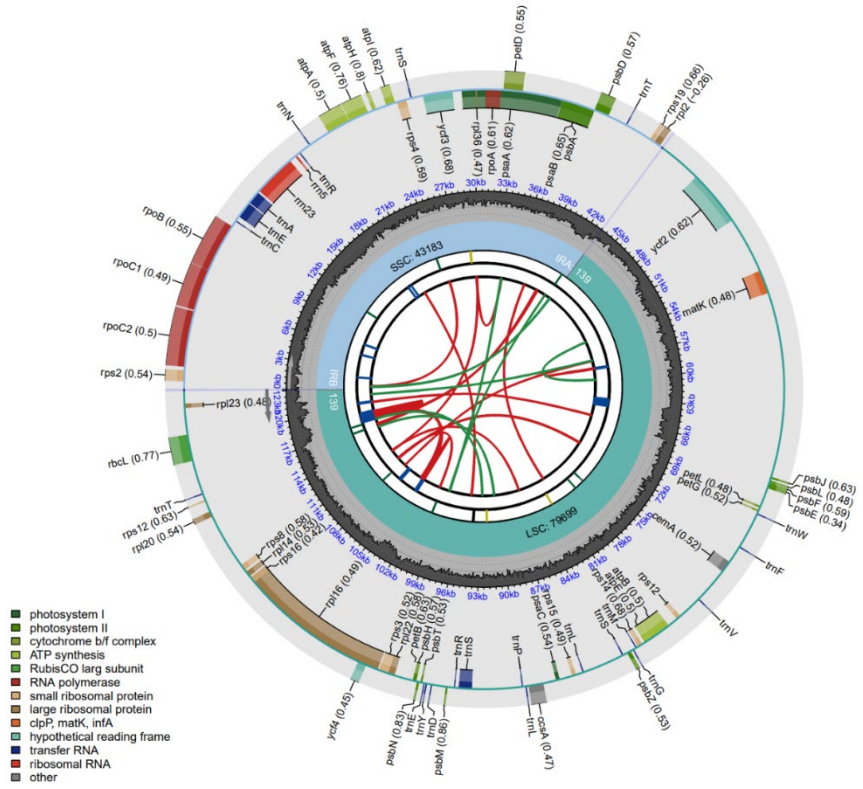


Figure 14
Schematic map of overall features of *Larix gmelinii* var. *olgensis* chloroplast genome

Among the protein-coding genes, nine (*rps2, rps3, rps4, rps8, rps12, rps14, rps15, rps16, rps19,*) encode small ribosomal subunits, while seven (*rpl2, rpl14, rpl16, rpl20, rpl22, rpl23, rpl36*) encode large ribosomal subunits. Additionally, there are twenty-six genes associated with photosynthesis proteins, four genes (*rpoA, rpoB, rpoC1, rpoC2*) responsible for DNA-dependent RNA polymerase, and eight genes (*accD, ccsA, cemA, matK*) responsible for other proteins. Two hypothetical reading frames were found (*ycf2*, *ycf4*) (Table 14). Furthermore, eight genes were identified to contain one intron and two exons. One gene (*ycf3*) contains notably two introns, two exons, and one gene (*psaA*) contains two introns and three exons. The largest intron, measuring 9949 bp, was found in the *rpl16* gene while the *trnY-AUA* gene has the smallest intron, measuring 521 bp (Table 15). Additionally, *rpoC1*, *atpf, ycf3*, *petD, rpl16* are classified as cis-splicing genes, as shown in (Figure 2.15).

Table 14
Gene composition in *Larix gmelinii* var. *olgensis* chloroplast genome

Category	Group of genes	Name of genes
Self-replication	Large subunit of ribosomal proteins	*rpl2, rpl14, rpl16, rpl20, rpl22, rpl23, rpl36*
	Small subunit of ribosomal proteins	*rps12, rps14, rps15, rps16, rps19, rps2, rps3, rps4, rps8*
	DNA-dependent RNA polymerase	*rpoA, rpoB, rpoC1, rpoC2*
	rRNA genes	*rrn16, rrn23, rrn4.5, rrn5*
	tRNA genes	*trnR-ACG, trnN-GUU, trnS-GGA, trnT-GGU, trnK-UUU, trnQ-UUG, trnW-CCA, trnF-GAA, trnL-IS, trnV-GAC, trnfM-CAU, trnG-GCC, trnS-UGA, trnL-CAA, trnP-GGG, trnV-UAC, trnG-UCC, trnR-UCU, trnD-GUC, trnY-GUA, trnE-UUC.*
Photosynthesis	Photosystem I	*psaA, psaB, psaC*
	Photosystem II	*psbA, psbD, psbE, psbF, psbI, psbL, psbM, psbN, psbT, psbZ, ycf3*
	Cytochrome b6/f complex	*petB, petD, petG, petL*
	ATP synthase	*atpA, atpB, atpE, atpF, atpH, atpI*
	Rubisco	*rbcl*
Other genes	Maturase	*matK*
	Protease	*clpP*
	Envelope membrane protein	*cemA*
	Subunit acetyl-CoA-carboxylase	*accD*
Unknown conserved open reading frames		*ycf2*, *ycf4*

Table 15
Lengths of introns and exons in split genes of *L. gmelinii* var. *olgensis*

Gene	Strand	Start	End	ExonI	IntronI	ExonII	IntronII	ExonIII
rpoC1	-	5234	7984	432	660	1659		
trnE-UU	+	12032	13100	33	994	42		
trnA-UGC	+	13178	14025	37	775	36		
atpF	-	21683	23003	145	766	410		
ycf3	+	26768	28710	126	721	225	156	
petD	-	32067	33315	7	712	530		
psaA	+	9316	35940	361	4366	1506	33	359
trnY-AUA	-	91880	92496	35	521	61		
trnS-CGA	+	93173	94038	32	774	60		
rpl16	+	99494	109847	8	9949	397		

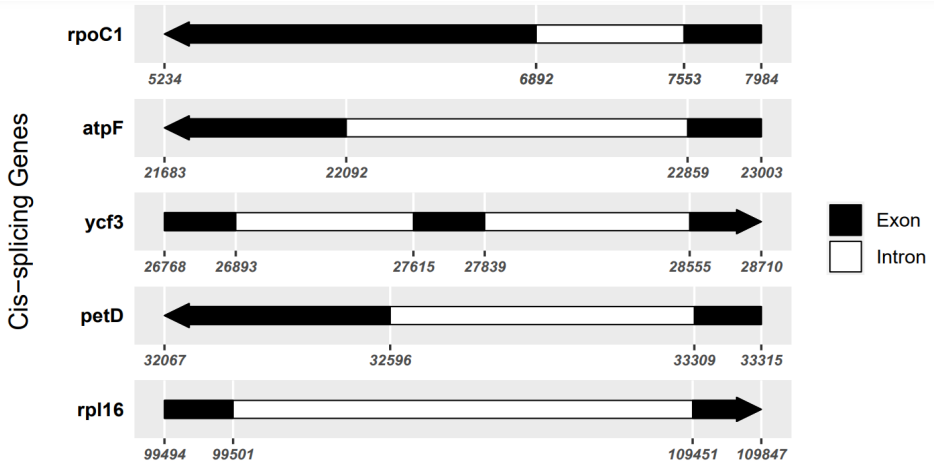


Figure 15
Schematic map of the cis-splicing genes in *L. gmelinii* var. *olgensis* chloroplast genome

We identified a total of 13 SSR loci. Out of these loci, ten were characterized by mononucleotide repeat motifs, while three exhibited dinucleotide repeat motifs (Table 16). No SSR loci with trinucleotide, tetranucleotide, pentanucleotide, or hexanucleotide repeat motifs were detected in the genome.

Table 16
Simple Sequence Repeats of *L. gmelinii* var. *olgensis*

Start nucleotide position	Motif and number of repeats	Location
11384	(A)14	Non-coding region between *trnC-GCA* and *trnT-GGU*
24070	(T)11	Non-coding region between *ccsA* and *rps4* genes
29002	(AT)6	Non-coding region between *trnS-GGA* and *rps2* genes
42569	(T)13	Non-coding region between *trnF-GAA* and *ycf12* genes
76700	(A)11	Non-coding region between *trnV-GAC* and *accD* genes
78151	(A)10	Non-coding region between clpP and *trnR-CCG* genes
82236	(AT)6	Non-coding region between clpP and *trnY-GUA* genes
91834	(AT)8	Non-coding region between *trnE-UUC* and *rpoB* genes
93730	(A)10	Non-coding region between *trnE-UUC* and rpoB genes
99739	(A)10	*trnS-CGA* gene
108131	(G)13	Non-coding region between *trnS-CGA* and *trnR-UCU* genes
116123	(T)10	Non-coding region between *atpA* and *ycf3* genes
117016	(A)10	Non-coding region between *rpl36* and *psaA* genes

1.4.2.6 Characteristics of *Larix laricina* chloroplast genome

The NovaSeq platform generated a total of 64,275,989 reads. The chloroplast genome revealed a final length of 122,699 bp with a GC content of 39.26%. The annotation through comparison with the available data for *L. sibirica* identified 101 genes, from which 30 represented *tRNA* genes, 4 *rRNA*, and 67 protein-coding genes. The small single-copy region measures 53,935 bp, while the long single-copy region covers 64,010 bp. A gene map of the genome was generated using CPGVIEW and is presented in (Figure 16).

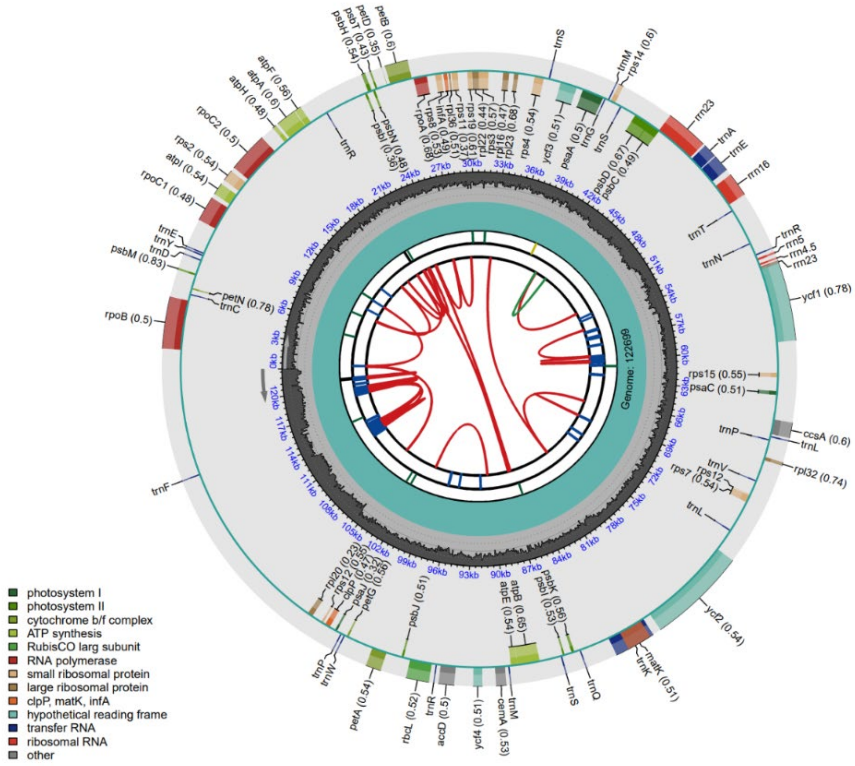


Figure 16
Schematic map of overall features of *Larix laricina* chloroplast genome

Among the protein-coding genes, ten (*rps11, rps12, rps14, rps15*, *rps19, rps2, rps3, rps4, rps7, rps8*) encode small ribosomal subunits, while six (*rpl16, rpl20, rpl22, rpl23, rpl32, rpl36*) encode large ribosomal subunits. Additionally, there are thirty-three genes associated with photosynthesis proteins, four genes (*rpoA, rpoB, rpoC1, rpoC2*) responsible for DNA-dependent RNA polymerase, and four genes (*accD, ccsA, cemA, matK*) responsible for other proteins. Three hypothetical reading frames were identified *(ycf1, ycf2, ycf4*) (Table 17).

Table 17
Gene composition in the chloroplast genome of *Larix laricina*

Category	Group of Genes	Name of Genes
Self-replication	Large subunit of ribosomal proteins	*rpl16, rpl20, rpl22, rpl23, rpl32, rpl36*
	Small subunit of ribosomal proteins	*rps11, rps12, rps14, rps15, rps19, rps2, rps3, rps4, rps7, rps8*
	DNA-dependent RNA polymerase	*rpoA, rpoB, rpoC1, rpoC2*
	rRNA genes	*rrn16, rrn23, rrn4.5, rrn5*
	tRNA genes	*trnR-ACG, trnN-GUU, trnS-GGA, trnT-GGU, trnK-UUU, trnK-UUU, trnQ-UUG trnW-CCA, trnF-GAA, trnL-IS, trnV-GAC, trnfM-CAUtrnG-GCC, trnS-UGA, trnL-CAA, trnP-GGG, trnV-UAC, trnV-UAC, trnG-UCC, trnR-UCU, trnD-GUC, trnY-GUAtrnE-UUC, trnT-GGU*
Photosynthesis	Photosystem I	*psaA, psaC, psaJ*
	Photosystem II	*psbC, psbD, psbI, psbI, psbJ, psbK, psbM, psbN, psbT, ycf3*
	Cytochrome b6/f complex	*petB, petD, petG, petN*
	ATP synthase	*atpA, atpB, atpE, atpF, atpH, atpI*
	Rubisco	*rbcl*
Other genes	Maturase	*matK*
	Protease	*clpP*
	Envelope membrane protein	*cemA*
	Subunit acetyl-CoA-carboxylase	*accD*
unknown conserved open reading frames		*ycf1*, *ycf2*, *ycf4*

Seven genes were identified to contain one intron and two exons. Notably, the *trnK* gene possesses the largest intron at 2513 bp, while the *rpoC2* gene has the smallest intron, measuring 30 bp (Table 18). Additionally, *rpoC1, atpf, ycf3, petD, rpl16*, are classified as cis-splicing genes (Figure 17).

Table 18
Lengths of introns and exons in split genes of *Larix laricina* chloroplast genome

Gene	Strand	Start	End	ExonI	IntronI	ExonII	IntronII
rpoC2	-	13437	15956	1408	30	1082	
petB	-	24775	26216	6	749	642	
Ycf 3	+	36077	37169	126	718	249	
trnA-UGC	-	46103	46950	37	775	36	
trnE-UUC	-	47028	48096	33	994	42	
Ycf1	-	54564	59579	1731	1189	36	2030
trnk-UUU	-	80708	83294	38	2513	36	

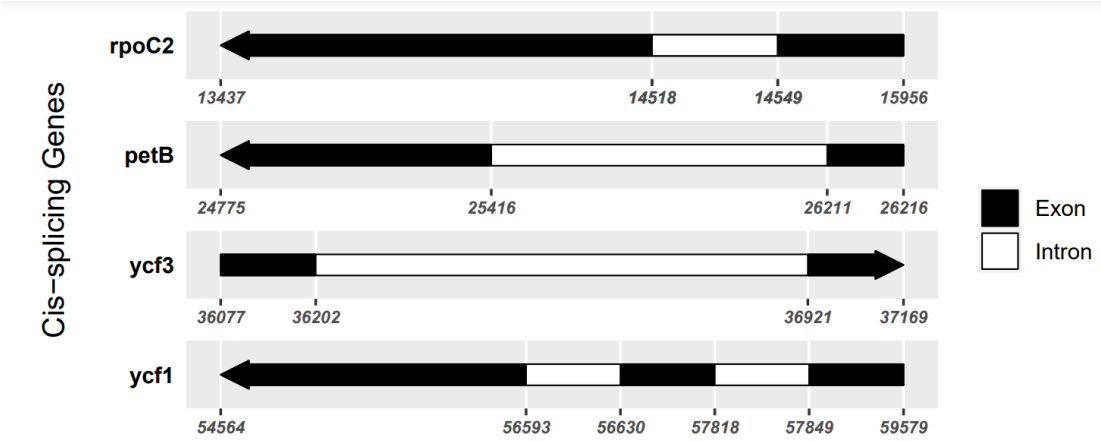


Figure 17
Schematic map of the cis-splicing genes in *Larix laricina* chloroplast genome

Two distinct categories of SSRs loci were detected within the chloroplast genome. All the twelve SSRs were identified to possess mononucleotide repeat motifs, excepted one locus was found with dinucleotide repeat motifs. The search parameters employed did not yield any SSR loci featuring trinucleotide, tetranucleotide, pentanucleotide, or hexanucleotide repeat motifs (Table 19).

Table 19
Simple Sequence Repeats in *Larix laricina* chloroplast genome

Start nucleotide position	Motif and number of repeats	Location
4649	(A)13	Non-coding region between *rbcl* and *atpH* genes
8109	(A)11	Non-coding region between *trnC-GCA* and *trnT-GGU*
11211	(T)10	Non-coding region between *ccsA* and *rps4* genes
19617	(A)10	Non-coding region between *trnS-GGA* and *rps2* genes
20139	(A)10	Non-coding region between *trnF-GAA* and *ycf12* genes
29861	(A)14	Non-coding region between *trnV-GAC* and *accD* genes
31613	(A)10	Non-coding region between *clpP* and *trnR-CCG* genes
39110	(AT)7	Non-coding region between *clpP* and *trnY-GUA* genes
61310	(A)14	Non-coding region between *trnE-UUC* and *rpoB* genes
85670	(C)11	Non-coding region between *trnE-UUC* and *rpoB* genes
101993	(G)10	*trnS-CGA* gene
102703	(A)15	*trnS-CGA* gene

1.4.2.7 Characteristics of *Larix occidentalis* chloroplast genome

A total of 71,455,531 reads were generated. Subsequent assembly of the chloroplast genome revealed a final length of 122,409 bp with a GC content of 39.4%. The annotation through comparison with the available data for *L. sibirica* identified 112 genes, from which 34 represented t*RNA* genes, 4 *rRNA*, and 74 protein-coding genes. The small single-copy region measures 53,935 bp, while the long single-copy region covers 64,010 bp. A gene map of the genome was generated using CPGVIEW and is presented in (Figure 18).

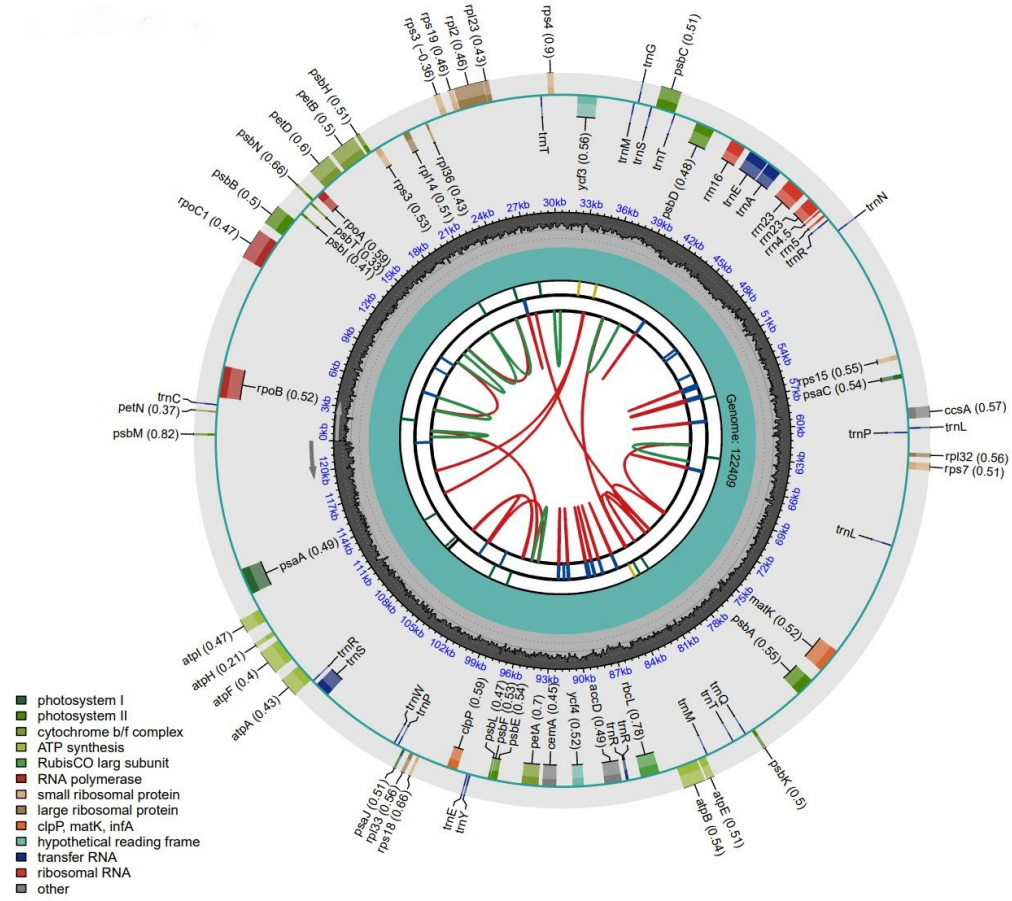


Figure 18
Schematic map of overall features of *Larix occidentalis* chloroplast genome

Among the protein-coding genes, twelve genes (*rps2, rps3, rps4, rps7, rps8, rps11, rps12, rps14, rps15, rps16, rps18, rps19*) code for small ribosomal subunits, while nine genes (*rpl2, rpl14, rpl26, rpl20, rpl22, rpl23, rpl32, rpl33, rpl36*) code for large ribosomal subunits. Additionally, there are 33 genes related to photosynthesis proteins, three genes *(rpoA, rpoB, rpoC1, rpoC2*) encoding DNA-dependent RNA polymerase, and four genes (*accD*, *ccsA*, *cemA*, *matK*) responsible for other proteins. Three hypothetical reading frames were identified (*ycf1*, *ycf2*, *ycf4*) (Table 20). Furthermore, fourteen genes containing one intron, and two exons were identified (Table 2.21). The *trnR-UCU* gene was found to have the largest intron (2036 bp), while the *psaA* gene had the smallest intron (33 bp). *PetD*, *PetB, rpl16, rpl1*, *ycf3*, *ycf1*, *rbcl*, *ycf4*, *atpf, psaA* are recognized as cis-splicing genes (Figure 19).

Table 20
Schematic map of overall features of *Larix occidentalis* chloroplast genome

Category	Group of genes	Name of genes
Self-replication	Large subunit of ribosomal proteins	*rpl14, rpl16, rpl2, rpl20, rpl22, rpl23, rpl32, rpl33, rpl36*
	Small subunit of ribosomal proteins	*rps11, rps12, rps14, rps15, rps16, rps18, rps19, rps2, rps3, rps4, rps7, rps8*
	DNA-dependent RNA polymerase	*rpoA, rpoB, rpoC1, rpoC2*
	rRNA genes	*rrn16, rrn23, rrn4.5, rrn5*
	tRNA genes	*trnD-GUC, trnR-UCU, trnG-UCC, trnS-GCU, trine-CAU, trnF-GAA, trnL-UAA, trnT-UGU, trnS-GGA, trnG-GCC, trnS-UGA, trnT-GGU, trnI-GAU, trnA-UGC, trnR ACG, trnN-GUU, trnL-UAG, trnP-GGG, trnV-GAC, trnL-CAA, trnH-GUG, trnK-UUU, trnQ-UUG, trnS-GCU, trnV-UAC, trnM-CAU, trnR-CCG, trnW-CCA, trnP-UGG, trnE-UUC, trnY-GUA*
Photosynthesis	Photosystem I	*psaA, psaB, psaC, psaJ*
	Photosystem II	*psbA, psbB, psbC, psbD, psbE, psbF, psbI, psbJ, psbK, psbL, psbM, psbN, psbT, psbZ, ycf3*
	Cytochrome b6/f complex	*petA, petB, petD, petG, petL, petN*
	Photosystem I	*psaA, psaB, psaC, psaJ*
	ATP synthase	*atpA, atpB, atpE, atpF, atpH, atpI*
	Rubisco	*Rbcl*
Other genes	Maturase	*matK*
	Protease	*clpP*
	Envelope membrane protein	*cemA*

	Subunit acetyl-CoA-carboxylase	*accD*
Unknown conserved open reading frames	*Ycf1, Ycf2, Ycf4*	

Table 21
Lengths of introns and exons in split genes in *Larix occidentalis* chloroplast genome

Gene	Strand	Start	End	ExonI	IntronI	ExonII
petD	-	15956	17203	7	711	530
petB	-	17418	18859	5	794	643
rpl16	+	19896	21165	8	865	397
ycf3	+	30635	32577	126	721	225
trnE-UUC	+	41833	42978	33	1071	42
trnA-UGC	+	43056	43903	37	775	36
ycf1	+	52346	55779	1474	77	1883
trnR-UCU	+	55462	57590	32	2036	61
trnY-AUA	-	82538	83154	35	521	61
rbcl	+	86258	87762	898	77	530
ycf4	+	90585	91216	133	77	422
trnS-CGA	+	106311	107177	32	775	60
atpF	-	109153	110478	145	771	410
psaA	+	113542	114981	1048	33	359

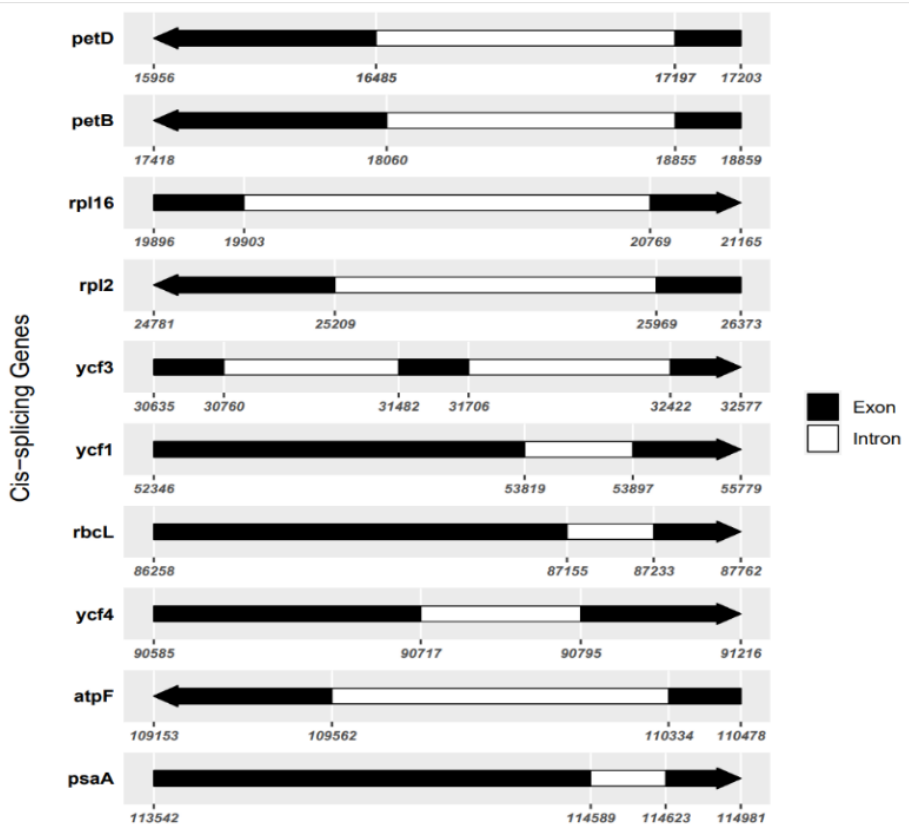


Figure 19
Schematic map of the cis-splicing genes in *Larix occidentalis* chloroplast genome

Two distinct categories of SSRs loci were detected within the chloroplast genome. Specifically, a total of 13 loci were identified to possess mononucleotide repeat motifs, while 3 loci were found with dinucleotide repeat motifs. The search parameters employed did not yield any SSR loci featuring trinucleotide, tetranucleotide, pentanucleotide, or hexanucleotide repeat motifs (Table 22).

Table 22
Simple sequence repeats in *Larix occidentalis* chloroplast genome

Start nucleotide position	Motif and number of repeats	Location
2305	(T)13	Non-coding region between *clpP* and *trnR-CCG* genes
20141	(A)11	Non-coding region between *ccsA* and *rps4* genes
24749	(T)14	Non-coding region between *trnS-GGA* and *rps2* genes
27934	(C)10	Non-coding region between *trnF-GAA* and *ycf12* genes
32869	(AT)6	Non-coding region between *trnV-GAC* and *accD* genes
34970	(AT)7	Non-coding region between *clpP* and *trnR-CCG* genes
55995	(A)14	Non-coding region between *clpP* and *trnY-GUA* genes
63881	(T)11	Non-coding region between *trnE-UUC* and *rpoB* genes
80568	(G)10	Non-coding region between *trnE-UUC* and *rpoB* genes
81797	(A)10	*trnS-CGA* gene
82492	(AT)8	Non-coding region between *trnS-CGA* and *trnR-UCU* genes
98502	(A)21	Non-coding region between *atpA* and *ycf3* genes
100838	(C)10	Non-coding region between *rpl36* and *psaA* genes
107392	(A)10	*trnS-CGA* gene
111545	(T)10	Non-coding region between *trnE-UUC* and *rpoB* genes

We have compiled key characteristics of plastid genomes from various *Larix* species in the table below. It includes their accession number, the number of Illumina read clusters, plastid genome size in base pairs (bp), overall GC content percentages (%), the total number of genes, and the count of unique *tRNA* genes (Table 23).

Table 23
Summary of chloroplast genome characteristics in *Larix* species

Genome features	Number of Illumina reads	Genome size (bp)	Overall GC content (%)	Total number of genes	Number of unique tRNA genes
Larix gmelinii var. olgensis	55,856,690	122,489	49.79	95	30
Larix gmelinii var. olgensis	55,856,690	123,46	39.50	105	30
Larix gmelinii var. japonica	62,308,733	122,339	35.55	108	34
Larix occidentalis	71,455,531	122,409	39.14	112	34
Larix sukaczewii	56,716,387	122,048	40.54	108	30
Larix sibirica	66,519,011	122,595	39.57	111	32
Larix laricna	64,275,989	122,699	39.26	101	30

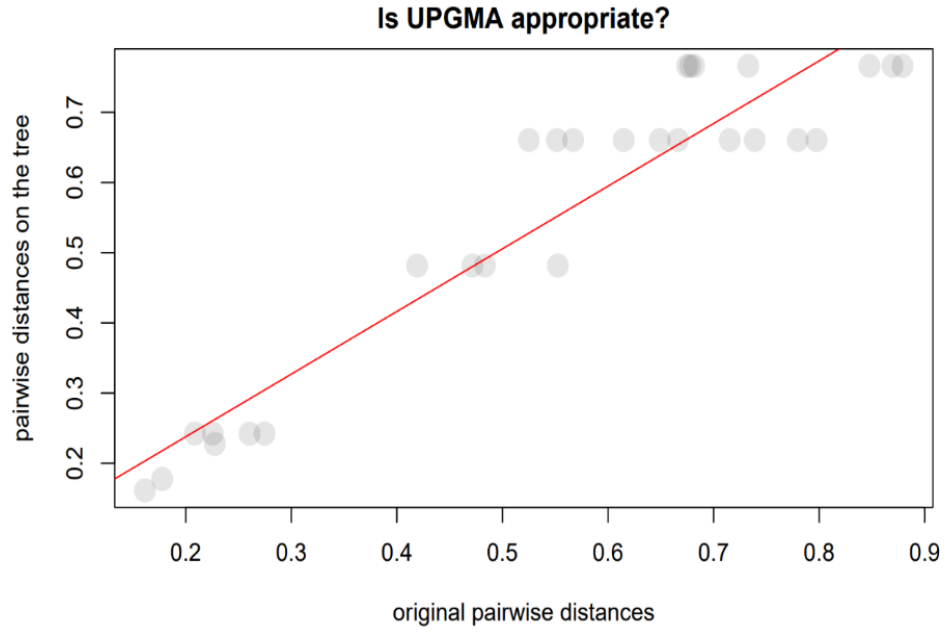
1.4.3 Construction of the phylogenetic tree

1.4.3.1 Neighbor-joining versus unweighted pair group method

The results showed that, in the case of (NJ) method. The data points are closely aligning with the regression line with some deviations. This alignment suggests a good fit with minor discrepancies between the tree distances and the original distances. Consequently, we can infer that the tree generated by the (NJ) method accurately depicts the evolutionary relationships within our dataset. Conversely, the unweighted pair group method with arithmetic mean yielded notably different results. The data points exhibited significant deviations from the regression line when analyzed in the context of this method. These deviations provide compelling

evidence that the (UPGMA) method is not the ideal choice for constructing the phylogenetic tree in this analysis (Figure 20).

A



B

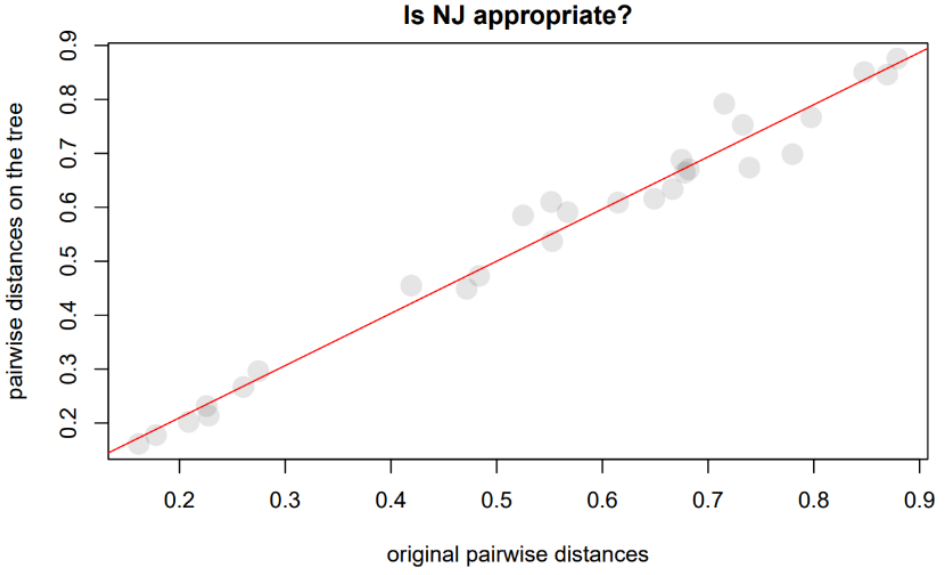


Figure 20
Linear regression analysis of phylogenetic distance preservation in (A) neighbor-joining (NJ) and (B) unweighted pair group method with arithmetic (UPGMA) methods

1.4.3.2 Neighbor-joining versus maximum likelihood method

The ANOVA test shows a significant difference between the two methods. The results indicated that (ML) analysis provide superior accuracy and data representation, supported by the statistical significance of the p-value (2^{e-16}). The results revealed that the (ML) model had a lower AIC (93247.04) compared to the (NJ) method (93247.04). Furthermore, the (ML) model displayed a lower AIC (93247.04) compared to the (NJ) method (93247.04), highlighting its ability to achieve a better balance between accurate data fitting and model parsimony. According to MEGA 11 test, the model GTR+I (General Time Reversible model with a proportion of Invariable sites) had the lowest BIC scores (Annexe C). In fact, Phylogenetic analyses were completed on an alignment of concatenated nucleotide sequences of all chloroplast genomes based on the (ML) method and GTR+I model as shown in (Figure 21).

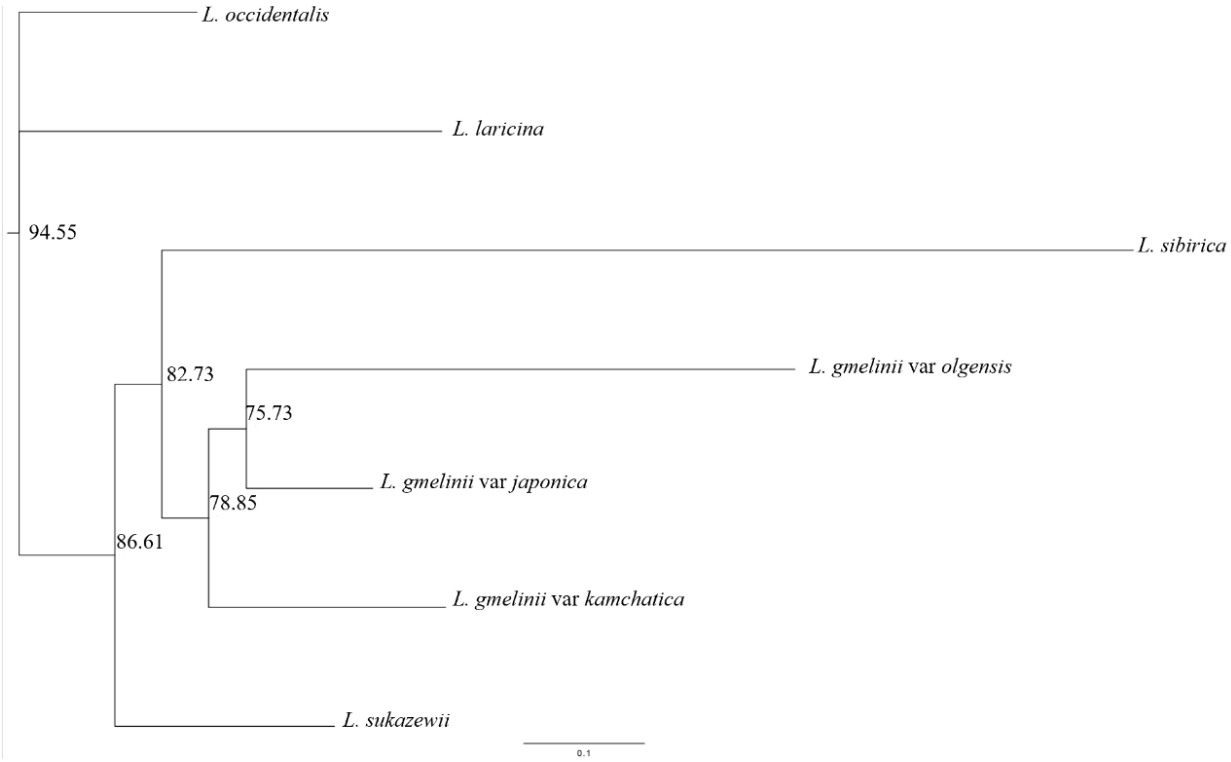


Figure 2.21
The maximum likelihood phylogenetic tree constructed based on the seven chloroplast genomes of *Larix* used in this study

The phylogenetic tree showed that the Asian species form a monophyletic group. *L. gmelinii* var. *japonica*, *L. gmelinii* var. *olgensis*, *L. gmelinii* var. *kamchatica* grouped into one clade. *Larix sukaczewii* is closely related to all the Asian species, but it is genetically distant from *Larix sibirica*. On the other hand, the relationships among the North American species are indeterminate, forming an unresolved polytomy with no synapomorphies shared by them.

1.5 Discussion

1.5.1 Characteristics of the chloroplast genome of *Larix sibirica*

The comparison between the chloroplast genome of *Larix sibirica* in this study and that of Bondar *et al*., (2019) reveals a slight difference of 35 bp in genome length (122,565 bp compared to 122,595 bp in our results). Both studies report an identical number of genes, coding sequences, tRNA genes, and rRNA genes, indicating a high degree of similarity in the overall genomic content and organization. Notably, both identify the same set of unknown conserved open reading frames (*ycf1*, *ycf2*, *ycf3*, *ycf4*). Both studies agree on the location of *ycf3* within the photosystem genes. This consistency highlights the potential significance of these additional open reading frames in the chloroplast genome, justifying further exploration into their functional roles and implications for photosynthesis and chloroplast function. Additionally, Bondar *et al*., (2019) reported 23 chloroplast microsatellites in the Siberian larch chloroplast genome, while our study identified 19. This discrepancy may be attributed to variations in the parameters used to detect these repetitions. Besides, Ebert & Peakall (2009) affirmed that most variations in the chloroplast genome are associated with microsatellite loci. No SSR loci with tri-, tetra-, penta-, and hexanucleotide repeats were found with both search parameters used in both studies. Additionally, our study identified SSRs within the *ycf4* region, in contrast to their studies that found SSRs primarily in coding regions rather than in hypothetical gene frames. Our results align with previous findings which also observed tandem repeats in the *ycf1* region of other conifers, such as *Cryptomeria japonica* (Hirao *et al.*, 2009).

1.5.2 Characteristics of *Larix sukazchewii*

The complete chloroplast genome of *Larix sukaczewii* has been successfully sequenced for the first time, revealing a high degree of similarity in overall genomic content and organization when compared to other larch species. Initially, authors recognized only one species in the western Urals and western Siberia, namely *Larix sibirica*, as reported in different studies by (Bobrov 1972, 1978; Farjon 1990; Milyutin & Vishnevetskaia, 1995). Conversely, other researchers identified two distinct species, *L. sibirica* and *L. sukaczewii*, in the same region, as noted by Abaimov *et al*., (1998, 2002), and Bashalkhanov *et al*., (2003).

In the study of Khatab *et al*., (2008), the levels and patterns of nucleotide variation in two nuclear gene regions: 4-coumarate coenzyme A ligase (4CL) and coumarate 3-hydroxylase (C3H) were investigated and their results indicated that certain haplotypes were unique to either *L. sukaczewii* or *L. sibirica*, thus confirming the genetic differentiation between these two species. Building upon this prior research, our analysis further substantiates the separation between these two species. Notably, the chloroplast genome of Siberian larch exhibited a greater number of genes, particularly those related to self-replication and photosystem functions. Specifically, *Larix sukaczewii* displayed only seven genes containing introns, while Siberian larch contained thirteen. Moreover, our analysis identified one conserved open reading frame of unknown function in Sukachev's larch, whereas three were detected in the Siberian larch. We also identified 17 simple sequence repeats in Sukachev's larch, while 19 SSRs were found in Siberian larch.

1.5.3 Characteristics of the chloroplast genome of *Larix gmelinii var. japonica*

In their analysis, Ishizuka *et al.,* (2017), explored the genome of the *L. gmelinii* var. *japonica*, reporting a length range of 122,553 to 122,598 bp. Our present study is consistent with these previously reported results, as we determined the genome length of this species to be 122,339 bp. Same to their studies, we reported an identical type, number and order of genes, coding sequences, tRNA genes, and rRNA gene. All the SSR repeats were detected only in the non-coding sequences which is different from what was found by Ishizika et *al*., (2017) where SSRs in the

coding sequences (*trnA* region, *ycf1, ycf2* genes) were detected. In addition, the same tRNA genes containing one intron were found except the *trnG-UCC*. In fact, we reported *trnG-GCC* to harbour one intron instead of *trnG-UCC*. This result is identical to what is found by Chen *et al*., (2020). Also, in their study about the *Larix kaempferi* (Japanese larch) chloroplast genome including the isolated population at the northern limit of the range (Manokami larch); initially Chen et *al*., (2020) believed that it is *Larix gmelinii* var. j*aponica*, based on the morphological traits. Our study showed genetic similarities between *Larix kaemferi* or more specifically could be the isolated population (Manokami larch) and *L. gmelinii* var. *japonica*, in fact, *trnG-UCC* could be a characteristic shared between the two species.

This discrepancy highlights that, despite the overall genomic similarity, there are variations in specific genes within the population of *Larix gmelinii var. japonica*. The study of Ishizuka *et al*., (2017) identified two presumed lineages, correspond to the geographically isolated origins: the Chishima lineage from the Kurile Islands and the Karafuto lineage from Sakhalin Island. These variations are believed to be genetically controlled and associated with the differences in the origin of seeds and/or seedlings (Kurinobu, 2005).

1.5.4 Characteristics of the chloroplast genome of *Larix gmelinii* var. *olgensis*

The chloroplast genome length of *Larix gmelinii* var. *olgensis* in our study closely aligns with the findings of Kim *et al*., (2018), although a slightly larger size of 123,160 bp was observed in our investigation. This discrepancy may stem from factors such as sequence methodology used, sample variation, reference genome version, or genomic variability. Despite this, the recorded length falls within the average range, considering that chloroplast DNA typically varies around 120 bp (Turudić *et al*., 2021). Notably, our study identified two additional copies of certain genes (*psbI*), other than *trnI-CAU, trnS-GCU*, and *trnT-GGU*. The presence of this additional *psbI* gene prompts further investigation into the underlying reasons. Explanations may include genomic rearrangements, insertions, or variations in tandem repeats that could lead to the expansion of specific gene regions. Understanding the mechanisms responsible for such variations can provide valuable

insights into the genomic dynamics of *Larix gmelinii* var. *olgensis* and contribute to a comprehensive characterization of its chloroplast genome.

1.5.5 Characteristics of the chloroplast genome of *Larix gmelinii* var. *kamchatica*

The chloroplast DNA sequencing of *Larix gmelinii* var. *kamchatica* reveals distinctive features compared to other *Larix gmelinii* species. Genome annotation identifies the fewest number of genes. We reported the presence of five genes, including *ycf3* containing one intron, a contrast to *Larix gmelinii* var. *japonica* and *Larix gmelinii* var. *olgensis*, which exhibit two introns in this gene. In their study, Khatab *et al.,* (2008) showed a significative differentiation between *Larix gmelinii* population based on patterns of nucleotide variation suggesting that *Larix gmelinii* from central Siberia is very different from the Russian Far East and Kamchatka Peninsula population.

Our result suggests a closer genetic relationship between *L. gmelinii* var. *japonica* and *L. gmelinii* var. *olgensis* compared to their relationship with *L. gmelinii* var. *kamchatica.* We did not detect the *ycf*1 in all *Larix gmelinii* population and previous research has reported insertions or deletions involved in these genes among *Larix gmelinii* (Ishizuka *et al.*, 2017). Although, it was considered a possibility that the *ycf1* might be a nonfunctional pseudogene. Another study indicated that *ycf1* is a functional gene and encodes a product essential for cell survival (Drescher *et al*., 2000) but Dong *et al*., (2015) revealed that the divergence of the *ycf1* was obvious in gymnosperms.

1.5.6 Characteristics of the chloroplast genome of *Larix occidentalis*

Parks *et al.,* (2009) conducted partial chloroplast DNA sequencing of Western larch, revealing 105 genes, including 27 tRNA, 68 protein-coding sequence, and 4 rRNA genes. Gernandt *et al*., (2018) advanced the understanding with whole-genome sequencing, identifying 112 genes with 34 tRNA, 4 rRNA, and 74 protein-coding genes; in concordance with our results which showed a total of 112 genes, with 34 tRNA, 4 rRNA, and 74 protein-coding genes. In our study, the chloroplast genome length of Western larch was 122,409 and is close to all other larch species. Notably, *ycf1*, *ycf2*, *ycf4* identified as conserved open reading frames. The

hypothetical gene *ycf1* was only detected in the American larch species and the Siberian larch cp genomes. In their study, Parks *et al.,* (2009) determined a disproportionate amount of phylogenetic information resides in two loci (*ycf1*, *ycf2*), highlighting their unusual evolutionary properties. In fact, our finding may explain close evolutionary events between *L. sibirica* and American larch species.

1.5.7 Characteristics of the chloroplast genome of *Larix laricina*

For the first time, the complete chloroplast genome of *L. laricina* is sequenced. The results indicated a similarity in size and gene order to those of *L. decidua* (Wu *et al*., 2011), *L. gmelinii* (Ishizuka *et al*., 2017)*, L. potaninii* (Han *et al*., 2017) and *L. sibirica* (Bondar *et al*., 2019). Given the geographical proximity of *L. laricina* to *L. occidentalis*, we expected a close genetic structure. Our results indicated both similarities and differences between them. From the fourteen genes found to harbor introns and exons in *L. occidentalis,* five genes are also present in *L. laricina* which suggests a degree of conservation in gene content and structure between these two species.

This conservation is further supported by the presence of the same conserved unknown frame genes (*ycf1*, *ycf2*, *ycf4*) in both species. These findings prove that certain elements of the chloroplast genome have remained stable and conserved throughout their evolutionary history. Additionally, the similar number of SSR markers in the noncoding sequences of both species, apart from one found in a *tRNA* gene, suggests a degree of similarity in their noncoding regions. These markers are known to be highly variable and can be used for genetic diversity studies.

1.5.8 Species phylogeny based on chloroplast genomes.

The phylogenetic tree derived from cpDNA sequences showed that all the Asian species are genetically close. Genetic studies showed the split of the genus into three main groups: the North American taxa, the South Asian taxa and the North Eurasian taxa. These three groups align with continental geography and reflect the three dispersal patterns outlined by LePage & Basinger (1995). This finding also aligns with earlier studies based on PCR-RFLPs of cpDNA, ITS sequences of

nuclear DNA, and multilocus AFLP data (Semerikov & Lascoux, 2003; Wei & Wang, 2003; Gros-Louis *et al*., 2005).

Among the Asian clade, *Larix gmelinii* var. *japonica* and *Larix gmelinii* var. *olgensis* formed one cluster while *Larix gmelinii* var. *kamchatica* was genetically close to them but formed a distinct cluster. This result is consistent with Chen *et al.*, (2020) finding based on the maximum-likelihood phylogenetic tree of 14 chloroplast genomes. However, their study excluded the newly sequenced species from Kamchatka.

Larix sibirica is close to *Larix gmelinii*. It exhibits the longest branch length among Asian species, indicating a significant genetic divergence within the clade, notably forming a distinct cluster within the phylogenetic tree. In the study of Gros-Louis *et al*., (2005), this divergence is explained by the discovery of one RAPD and two nuclear gene sequence polymorphisms specific to populations of *Larix sibirica* suggest that divergent evolution may be occurring in these populations due to mutation or drift, likely influenced by their remote geographic and high-altitude location in southwestern Russia (Semerikov *et al*., 1999). Despite this divergence, the Siberian larch is genetically close to other Asian species. This suggests that while *L. sibirica* has considerable evolutionary divergence, it shares a common evolutionary history with other North Asian species. This finding aligns with previous studies showing unclear position depending on the chosen genetic marker of *Larix sibirica.* It was grouped with South Asian taxa on a tree derived from an analysis of the chloroplast trnT-trnF region (Wei & Wang 2003). But it was grouped with other North Eurasian taxa on nuclear gene trees, those derived from phylogenetic analysis of nuclear *4CL* genes (Wei & Wang 2003), allozymes (Semerikov & Lascoux 1999), amplified fragment-length polymorphism (AFLP) data (Semerikov *et al*., 2003).

Our study revealed a clear distinction between Siberian larch (*Larix sibirica*) and Sukachev's larch (*Larix sukaczewii*). Initially, many authors, such as Bobrov (1978) and Milyutin & Vishnevetskaia (1995), rejected the classification of *L. sukaczewii* as a separate species, arguing that it could not be distinguished from *L. sibirica*.

However, subsequent evidence has shown that *L. sukaczewii* differs from *L. sibirica* in several morphological and biochemical traits. A comparative analysis of genetic distances, estimated using the maximum likelihood method for trnK intron sequences, revealed that the genetic distance between *Larix sibirica* and *Larix sukaczewii* was similar to the interspecific distances observed among other larch species pairs (Bashalkhanov *et al*., 2003). Later, the study of Khatab *et al*., (2008) reinforced the separation between the two species based on patterns of nucleotide variation of two nuclear gene regions: the 4-coumarate coenzyme A ligase (4CL) and the coumarate 3-hydroxylase (C3H).

Our cpDNA phylogenetic tree put the two North American species as a sister group to the North Asian group. Our results agree with the phylogenetic analysis of ITS sequences of nuclear ribosomal DNA and that of PCR-RFLPs of cpDNA, which placed the North American taxa in the basal position (Semerikov *et al*., 2003). In examining the North American larch species, our cpDNA tree suggests a close relationship between *L. laricina* and *L. occidentalis*, such finding is supported by Gros-Louis *et al*., (2005) study based on specific sites at positions 224 and 260 of the *Sb51* gene. However, we were unable to determine their precise genetic relationship since we found unresolved polytomy with no synapomorphies shared by them. This unresolved classification was also noticed in Gros-Louis *et al*., (2005) study who combined the mtDNA and cpDNA dataset to the construct a phylogenetic tree representing total cytoplasmic DNA evidence and showing the genus split into American and Asian groups.

The observed incongruence could be due to two main factors. On one hand, it may reflect genuine evolutionary complexity, such as recent divergence or ongoing gene flow between *L. laricina* and *L. occidentalis* (Davies *et al*., 2012). On the other hand, it might be an issue related to the genetic indicators themselves.

The cpDNA may not provide sufficient resolution to clearly distinguish these species or may not capture the true evolutionary signatures due to methodological constraints or the nature of the sequences.

1.6 Conclusion

The next-generation sequencing approach, combined with the assembly and annotation processes, enabled us to analyze the chloroplast genomes of seven *Larix* species derived from North America and North Asia. The chloroplast genome size, GC content, and gene number, and order among seven species are highly similar to each other. The results revealed a genome length extending from 122,048 to 123,460 bp. An average of 105 genes including four ribosomal RNA genes and 30 tRNA genes identified within each genome and GC content ranges between 35 – 40%.

Some punctual changes were observed, such as SSC/IR and LSC/IR boundaries, hypothetical gene frames and SSR number. Repeat sequences may play an important role in chloroplast genome arrangement and sequence divergence. They often contain highly polymorphic variations within a population of conifers. Our result showed that among all the species the SSR motifs number varied between 14 to 19 located all in intergenic space. We identified only one SSR located in the *ycf4* gene in the case of *Larix sibrica.* Additionally, we identified two SSRs in the trnA region in the case of *Larix sukaczewii* and *Larix occidentalis*. The majority of the detected SSR motifs were mononucleotide motifs, of which the SSR motif of mononucleotide T was the most frequent, followed by mononucleotide A and mononucleotide G. In this study, only Siberian larch and the North American species are containing the hypothetical gene *ycf1* and previous research has reported insertions or deletions in this gene.

Phylogenetic analysis based on the chloroplast genomes put the North American species as a sister group to both the North Eurasian groups. Among the Asian clade *Larix gmelinii* populations formed a monophylogenetic group. *Larix gmelinii* var. *japonica* and *Larix gmelinii* var. *olgensis* formed one cluster while *Larix gmelinii* var. *kamchatica* was genetically close to them but formed a distinct cluster.

CONCLUSION GÉNÉRALE

La présente étude, axée sur le genre *Larix*, a permis de séquencer entièrement les génomes chloroplastiques de sept espèces de mélèzes d'Amérique du Nord et d'Asie du Nord. Le génome chloroplastique du mélèze est de forme circulaire, composé de quatre parties : deux copies de larges répétitions inversées (IRs) séparées par une grande copie unique (LSC) et une petite copie unique (SSC) de la région. Le cœur des IRs code pour quatre ARN ribosomiques (16S, 23S, 4,5S et 5S) (Francisconi *et al*., 2023).
La leucine était la plus codée (690,143 (10 %)) alors que la cystéine était la moins codée (131,099 (2 %)) en acides aminés (Annexe E). Des ratios similaires d'acides aminés ont été trouvés dans les génomes chloroplastiques précédemment rapportés (Asaf *et al*., 2018).

Parmi les différentes espèces, la taille du génome, le contenu et l'arrangement génique sont relativement conservateurs. Les génomes chloroplastiques varient de 122 à 123 kb. *Larix occidentalis* avait la taille la plus élevée avec 123,699 pb, tandis que *Larix sukaczewii* avait la taille du génome la plus petite, avec 122,048 pb. Le génome chloroplastique de *L. sukaczewii* a été séquencé pour la première fois et nos résultats ont montré que cette espèce est génétiquement différent de *L. sibirica*. Cette dernière avait une taille de génome plus grande, un contenu en GC plus élevé et plus de répétitions SSR. De plus, *Larix sibrica* présentait 13 introns, dont seulement deux étaient communs à *L. sukaczewii*. En outre, *L. sibirica* contenait plus de gènes hypothétiques (*ycf1*, *ycf2*, *ycf4*) que *L. sukaczewii*, qui n'avait qu'un seul gène hypothétique (*ycf4*). Sur la base de la région ITS, Khatab *et al*., (2007) ont montré la distinction entre les deux espèces, rejetant ainsi l'hypothèse d'autres auteurs selon laquelle les espèces de la Russie centrale et de l'ouest de la Russie étaient regroupées en une seule espèce (Wei & Wang, 2004).

Une variation génétique intraspécifique a été observée chez les variétés de *Larix gmelinii*. En effet, *Larix kamchatica* présentait la plus petite taille de génome et le nombre le plus réduit de gènes, tout en ayant le contenu en GC le plus élevé. Cette variété ne possédait que 5 introns, retrouvés également chez *L. gmelinii* var.

olgensis. Les gènes hypothétiques *ycf2* et *ycf4* étaient présents dans les trois variétés, tandis que *ycf1* était absent chez celles-ci. Chen et *al*., (2020) ont montré que les gènes *ycf1* étaient inclus dans la région hypervariable. En étudiant la diversité génétique au sein des populations de *L. gmelinii*, l'étude de Ishizuka *et al*., (2017) a signalé des insertions ou des suppressions dans le *ycf1* de *L. gmelinii*. Bien qu'il ait été considéré comme possible que le *ycf1* soit un pseudogène non fonctionnel, une autre étude a indiqué que le *ycf1* est un gène fonctionnel et code un produit essentiel à la survie cellulaire. Cependant, étant donné que cela représentait une déviation par rapport à une structure hautement conservée, cette différence peut être expliquée par le fait que les populations de *Larix* en Eurasie auraient divergé génétiquement les unes des autres il y a peu de temps. Cette divergence récente peut être due à des événements tels que des migrations, des changements climatiques ou d'autres facteurs qui ont entraîné une séparation et une évolution génétique distincte des populations de *Larix* (Semerikov & Lascoux, 2003).

Parmi les sept espèces de *Larix* séquencées dans cette étude, un total de 1014 répétitions en séquence simple (SSR) a été identifié (Annexe D. 1). Le plus grand nombre de SSR a été observé chez *L. sibirica*, *L. gmelinii var. japonica* et *L. sukachezwii* (208, 199 et 199, respectivement), suivis des deux espèces américaines (*L. occidentalis et L. laricina*) (143 et 177, respectivement). Les deux autres *Larix gmelinii* présentaient le nombre de SSR le plus bas, avec *L. gmelinii var. olgeniss* à 132 SSR et *L. gmelinii var. kamchatica* à 132 SSR. Toutes les espèces présentaient un plus grand nombre de mononucléotides de type SSR, suivis de dinucléotides (Annexe D. 2).

Aucun tétranucléotide n'a été observé (Chen *et al*., 2020). De plus, toutes les espèces présentaient les motifs A/T et AT. Certains motifs étaient uniques à certaines espèces de mélèzes, tel que le motif C se produisant uniquement chez *L. sibirica*, *L. gmelinii var. kamchatica* et les deux espèces d'Amérique du Nord. En outre, le motif G était présent chez toutes les espèces à l'exception de *L. gmelinii var. japonica*, *L. sibirica* et *L. occidentalis* (Annexe D.1).

L’arbre phylogénétique basé sur les séquences des génomes chloroplastiques a révélé une relation génétique étroite entre les espèces de *Larix* asiatiques, en accord avec des études antérieures indiquant trois groupes principaux : les taxons nord-américains, sud-asiatiques et Nord-eurasiens (Semerikov & Lascoux, 2003 ; Wei & Wang, 2003 ; Gros-Louis *et al*., 2005). Au sein du clade asiatique, les populations de *Larix gmelinii* étaient monophylétiques, avec *L. gmelinii* var. *japonica* et *L. gmelinii* var. *olgensis* plus proches l'une de l'autre, tandis que *L. gmelinii* var. *kamchatica* était plus éloignée, mais restait étroitement liée. *Larix sibirica* a montré une divergence génétique significative au sein du clade asiatique, formant un groupe distinct, mais restant génétiquement proche des autres espèces asiatiques, suggérant une histoire évolutive commune. La séparation entre le mélèze de Sibérie et le mélèze de Sukachev a été confirmée, avec des preuves soutenant leur statut d’espèces distinctes. Les espèces nord-américaines ont montré une distance génétique par rapport aux espèces nord-asiatiques, mais les relations génétiques précises au sein du groupe nord-américain sont restées non résolues.

Généralement, les conifères qui sont géographiquement proches présentent une faible différenciation génétique en raison de leur fécondation croisée et de leur pollinisation par le vent, favorisant ainsi le brassage génétique (Loveless & Hamrick, 1984). Cependant, dans le cas du genre *Larix*, la situation est différente, car son pollen n’est pas équipé de sacs aériens, ce qui limite sa dispersion sur de longues distances. De même, les graines de *Larix* ne se dispersent pas facilement et sont généralement dispersées sur des distances équivalentes à moins de deux hauteurs d’arbre. Ainsi, l’isolement géographique est considéré comme une barrière significative au brassage génétique des populations de *Larix*.

L’assemblage et l’annotation des génomes de conifères représentent une ressource importante pour des études de diversité génétique ultérieures. Ces résultats contribueront à approfondir notre compréhension du genre *Larix*, incluant son évolution, sa conservation et son utilisation durable. Notre approche d’analyse comparative des génomes chloroplastiques ouvre la voie au développement de marqueurs génétiques utiles pour d’autres groupes d’espèces. Il est recommandé

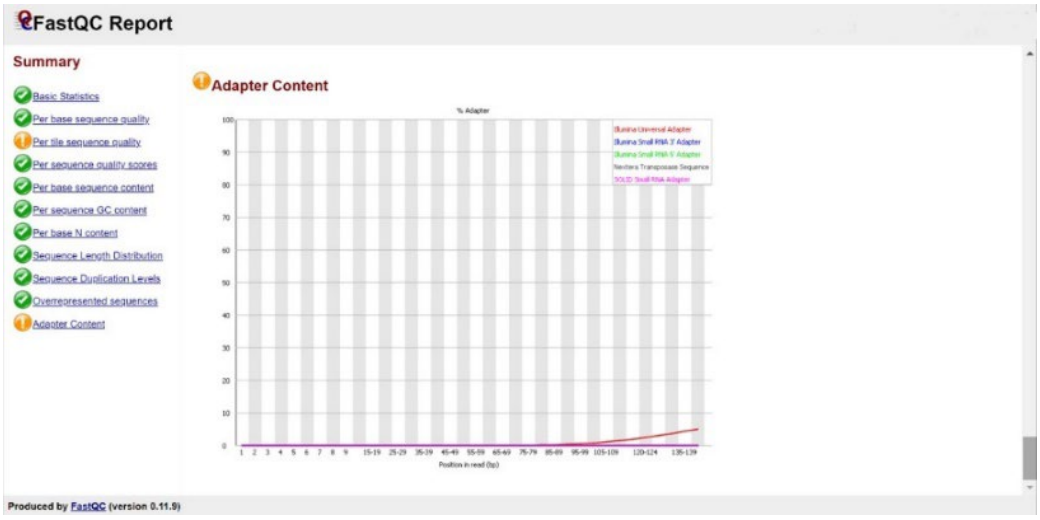
d’approfondir l’étude de la diversité génétique des mélèzes en incluant d’autres espèces, car la diversité génétique au sein de la même espèce est peu étudiée et mérite une attention particulière pour une meilleure compréhension de l’évolution et de l’adaptation de ces arbres. De plus, une phylogénie basée sur l’ADN chloroplastique des autres espèces d’Asie du Sud et d’Europe pourrait aider à clarifier les relations taxonomiques entre les différentes populations et sous-espèces du genre *Larix*.

ANNEXES

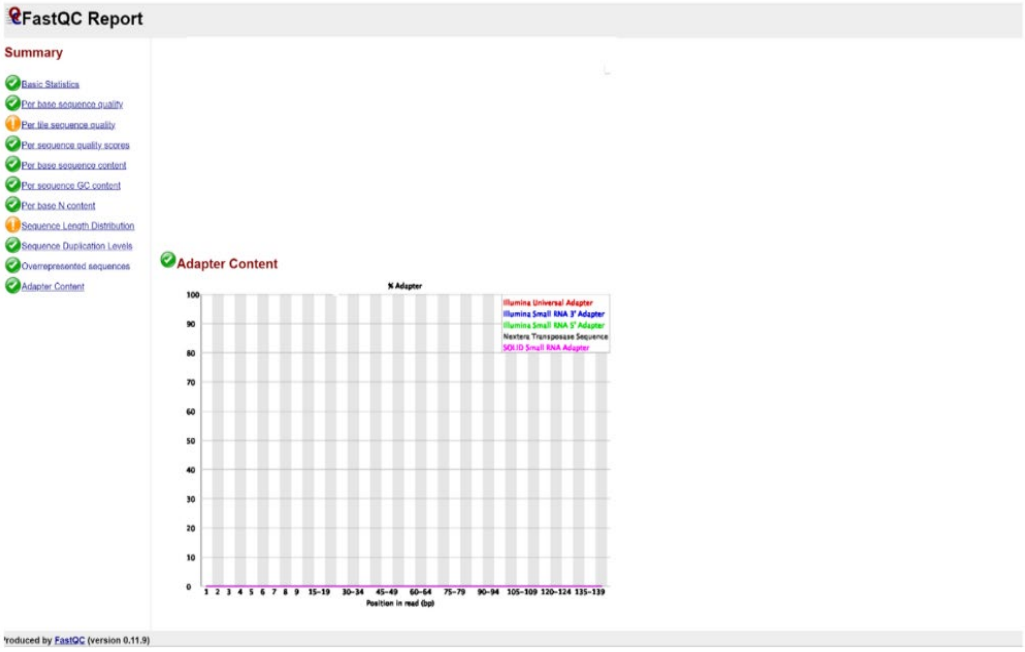
Annexe A - FastQC report depicting module-specific results

Annexe B - Comparison of adapter content before (A) and after (B) Trimmomatic analysis

A



B

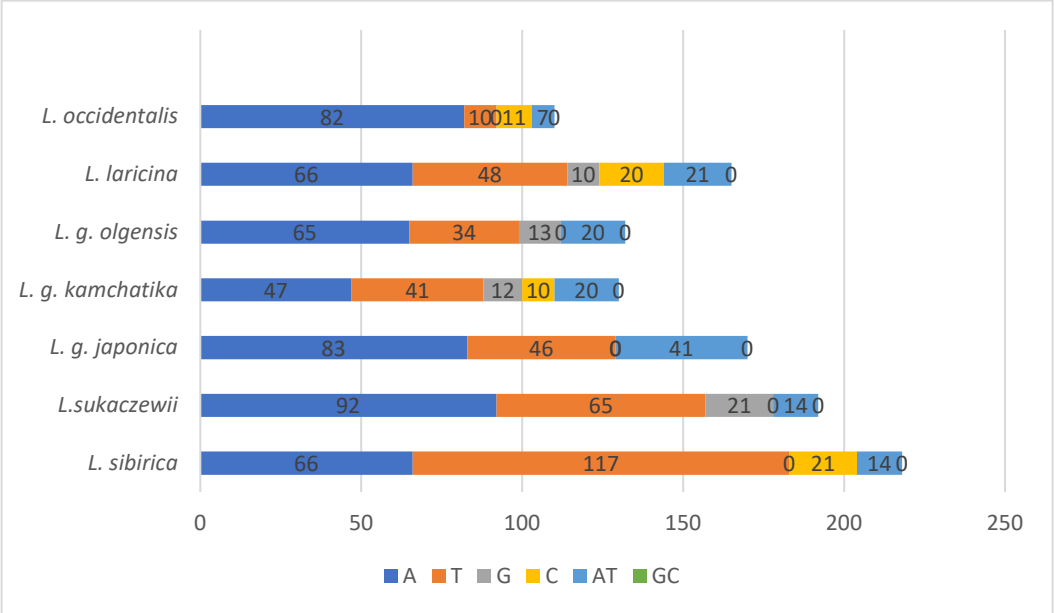


Annexe C - Comparative analysis table that lists 24 different nucleotide substitution models analyzed using maximum likelihood in MEGA 11. It highlights key details such as the model names, BIC scores for model evaluation, frequencies of nucleotides (A, T, C, G), and rates of nucleotide substitution between pairs

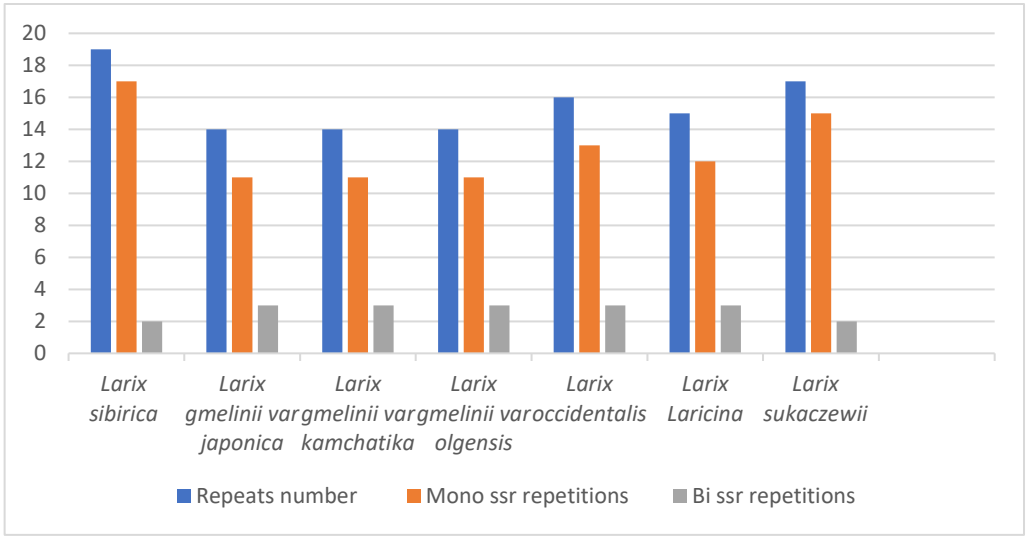
Model	BIC	f(A)	f(T)	f(C)	f(G)	r(AT)	r(AC)	r(AG)
GTR+I	1594154.191	0.301	0.299	0.197	0.203	0.059	0.042	0.114
GTR+G+I	1594170.371	0.301	0.299	0.197	0.203	0.059	0.042	0.114
GTR+G	1594218.667	0.301	0.299	0.197	0.203	0.058	0.042	0.114
GTR	1594278.818	0.301	0.299	0.197	0.203	0.059	0.042	0.114
T92	1594437.915	0.300	0.300	0.200	0.200	0.064	0.043	0.115
TN93	1594449.034	0.301	0.299	0.197	0.203	0.064	0.042	0.114
HKY	1594454.634	0.301	0.299	0.197	0.203	0.064	0.042	0.116
T92+G	1594502.030	0.300	0.300	0.200	0.200	0.060	0.040	0.119
TN93+G	1594513.301	0.301	0.299	0.197	0.203	0.060	0.040	0.118
HKY+G	1594518.724	0.301	0.299	0.197	0.203	0.060	0.040	0.121
T92+I	1595097.561	0.300	0.300	0.200	0.200	0.057	0.038	0.124
TN93+I	1595109.301	0.301	0.299	0.197	0.203	0.057	0.037	0.123
HKY+I	1595114.648	0.301	0.299	0.197	0.203	0.057	0.037	0.126
T92+G+I	1596612.869	0.300	0.300	0.200	0.200	0.052	0.035	0.131
TN93+G+I	1596624.927	0.301	0.299	0.197	0.203	0.052	0.034	0.129
HKY+G+I	1596630.141	0.301	0.299	0.197	0.203	0.052	0.034	0.132
K2	1609338.262	0.250	0.250	0.250	0.250	0.053	0.053	0.144
K2+G	1609547.306	0.250	0.250	0.250	0.250	0.050	0.050	0.150
K2+I	1610353.302	0.250	0.250	0.250	0.250	0.047	0.047	0.155
K2+G+I	1612228.114	0.250	0.250	0.250	0.250	0.043	0.043	0.163
JC+I	1629010.187	0.250	0.250	0.250	0.250	0.083	0.083	0.083
JC+G+I	1629032.479	0.250	0.250	0.250	0.250	0.083	0.083	0.083
JC+G	1629135.549	0.250	0.250	0.250	0.250	0.083	0.083	0.083
JC	1629158.074	0.250	0.250	0.250	0.250	0.083	0.083	0.083

Annexe D - Single sequence repeats (SSRs) among larch species

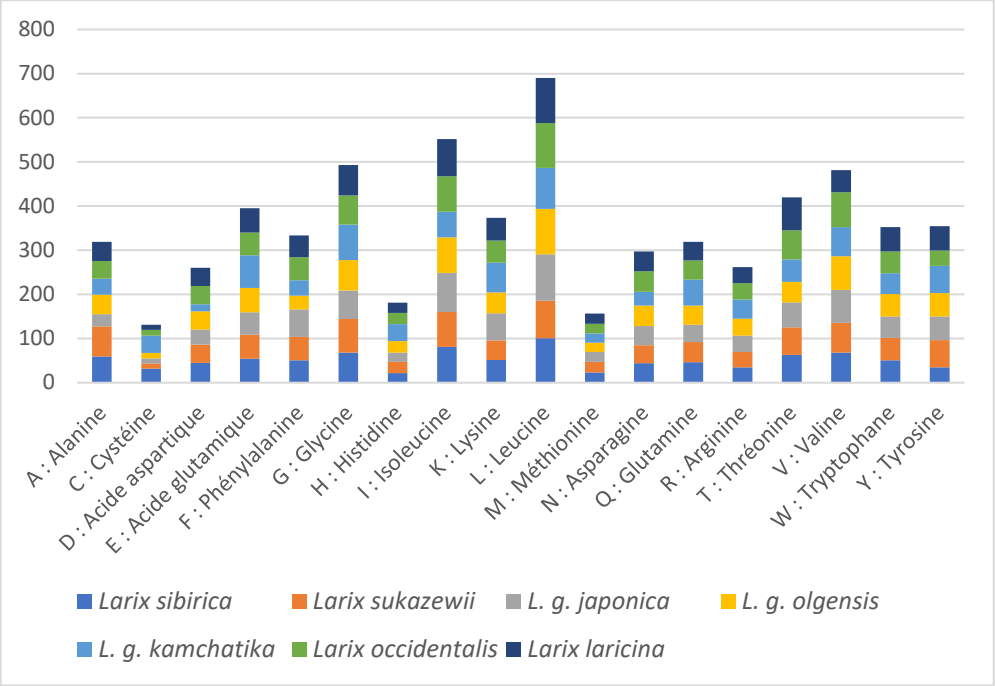
Annexe D.1 - Nucleotide variation frequency in SSRs among larch species



Annexe D.2 - Distribution and classification (mono-, di-) of SSRs among larch species



Annexe E - Comparative analysis of amino acid frequencies in the chloroplast genomes of seven larch species



LISTE DE RÉFÉRENCES

Abaimov, A. P. (2010). Geographical distribution and genetics of Siberian larch species. *Permafrost ecosystems: Siberian larch forests*, 41–58.

Abaimov, A. P., Milyutin, L. I., Lesinski, J. A., & Martinsson, O. (1998). Variability and ecology of Siberian larch species.

Acheré, V., Faivre Rampant, P., Pâques, L. E., & Prat, D. (2004). Chloroplast and mitochondrial molecular tests identify European Japanese larch hybrids. *Theoretical and Applied Genetics*, *108*(8), 1643–1649.

Ali, A., Pan, Y. B., Wang, Q. N., Wang, J. D., Chen, J. L., & Gao, S. J. (2019). Genetic diversity and population structure analysis of *Saccharum* and *Erianthus* genuses using microsatellite (SSR) markers. *Scientific reports*, *9*(1), 1–10.

Andrew, S. (2010). FastQC, a quality control tool for high throughput sequence data.

Arcade, A., Anselin, F., Rampant, P. F., Lesage, M. C., Paques, L. E., & Prat, D. (2000). Application of AFLP, RAPD and ISSR markers to genetic mapping of European and Japanese larch. *Theoretical and Applied Genetics*, *100*(2), 299–307.

Araki, N. H., Khatab, I. A., Hemamali, K. K., Inomata, N., Wang, X. R., & Szmidt, A. E. (2008). Phylogeography of *Larix sukaczewii* Dyl. and *Larix sibirica L.* inferred from nucleotide variation of nuclear genes. *Tree Genetics & Genomes*, *4*, 611–623.

Asaf, S., Khan, A. L., Khan, M. A., Shahzad, R., Lubna, Kang, S. M., Al-Harrasi, A., Al-Rawahi, A., & Lee, I.-J. (2018). Complete chloroplast genome sequence and comparative analysis of loblolly pine (*Pinus taeda* L.) with related species. *PLoS One*, *13*(3).

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., & Pevzner, P. A. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology*, *19*(5), 455–477.

Bashalkhanov, S. I., Konstantinov, Y. M., Verbitskii, D. S., & Kobzev, V. F. (2003). Reconstruction of phylogenetic relationships of larch *Larix* sukaczewii Dyl. based on chloroplast DNA trnK intron sequences. *Russian Journal of Genetics*, *39*, 1116–1120.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120.

Böhle, U. R., Hilger, H. H., & Martin, W. F. (1996). Island colonization and evolution of the insular woody habit in Echium *L*. *Boraginaceae*. *Proceedings of the National Academy of Sciences*, *93*(21), 11740–11745.

Bakker, F. T., Culham, A., Gomez-Martinez, R., Carvalho, J., Compton, J., Dawtrey, R., & Gibby, M. (2000). Patterns of nucleotide substitution in angiosperm cpDNA trnL (UAA)-trnF (GAA) regions. *Molecular Biology and Evolution*, *17*(8), 1146–1155.

Beier, S., Thiel, T., Münch, T., Scholz, U., & Mascher, M. (2017). MISA-web: a web server for microsatellite prediction. *Bioinformatics*, *33*(16), 2583–2585.

Bobrov, E. G. (1972). History and systematics of *Larix*. *Journal of Botany*, *60*(6), 797–805

Bobrov, E. G. (1978). Forest-forming conifers of the USSR. *Nauka Publishing Leningrad*, 189.

Bondar, E. I., Putintseva, Y. A., Oreshkova, N. V., & Krutovsky, K. V. (2019). Siberian larch (*Larix sibirica Ledeb*.) chloroplast genome and development of polymorphic chloroplastmarkers. *BMC Bioinformatics*, *20*(1), 47–52.

Chen, S., Ishizuka, W., Hara, T., & Goto, S. (2020). Complete chloroplast genome of Japanese larch (*Larix kaempferi*): insights into intraspecific variation with an isolated northern limit population. *Forests*, *11*(8), 884.

Davies, T. J., Kraft, N. J., Salamin, N., & Wolkovich, E. M. (2012). Incompletely resolved phylogenetic trees inflate estimates of phylogenetic conservatism. *Ecology*, *93*(2), 242–247.

Dobrogojski, J., Adamiec, M., & Luciński, R. (2020). The chloroplast genome: a review. *Acta Physiologiae Plantarum*, 42(6), 98.

Dong, W., Xu, C., Li, C., Sun, J., Zuo, Y., Shi, S., Cheng, T., Guo, J., & Zhou, S. (2015). *ycf1*, the most promising plastid DNA barcode of land plants. *Scientific reports*, *5*(1), 1–5.

Drescher, A., Ruf, S., Calsa Jr, T., Carrer, H., & Bock, R. (2000). The two largest chloroplast genome-encoded open reading frames of higher plants are essential genes. *The Plant Journal*, *22*(2), 97–104.

Ebert, D., & Peakall, R. O. D. (2009). Chloroplast simple sequence repeats (cpSSRs): technical resources and recommendations for expanding cpSSR discovery and applications to a wide array of plant species. *Molecular Ecology Resources*, *9*(3), 673–690.

Felsenstein, J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of molecular evolution*, *17*, 368–376.

Farjon, A. (1990) Pinaceae. Koeltz Scientific Books, Königstein, Germany.

Fujii, N. (1995). Intraspecific sequence variation in chloroplast DNA of *Primula cuneifolia* Ledeb. (Primulaceae). *Journal of Phytogeography and Taxonomy*, *43*, 15–24.

Fukuda, T., Yokoyama, J., & Ohashi, H. (2001). Phylogeny and biogeography of the genus *Lycium* (Solanaceae): inferences from chloroplast DNA sequences. *Molecular Phylogenetics and Evolution*, *19*(2), 246–258.

Francisconi, A. F., Marroquín, J. A. M., Cauz-Santos, L. A., van den Berg, C., Martins, K. K., Costa, M. F., Picanço-Rodrigues, D., de Alencar, L. D., Zanello, C. A., & Colombo, C. A. (2023). Complete chloroplast genomes of six Neotropical palm tree species: genome structure comparison, identification of repeats sequences and evolutionary dynamic patterns.

Gros-Louis, M. C., Bousquet, J., Pâques, L. E., & Isabel, N. (2005). Species-diagnostic markers in *Larix* spp. based on RAPDs and nuclear, cpDNA, and mtDNA gene sequences, and their phylogenetic implications. *Tree Genetics & Genomes*, *1*, 50–63.

Geoffrey M. Williams, Andrew S. Nelson, & David L.R. Affleck. (2017). Vertical distribution of foliar biomass in western larch (*Larix occidentalis*). *Canadian Journal of Forest Research*, *48*(1), 42–57.

Gernandt, D. S., & Liston, A. (1999). Internal transcribed spacer region evolution in *Larix* and *Pseudotsuga* (Pinaceae). *American Journal of Botany*,*86*(5),711–723.

Gernandt, D. S., Reséndiz Arias, C., Terrazas, T., Aguirre Dugua, X., & Willyard, A. (2018). Incorporating fossils into the Pinaceae tree of life. *American Journal of Botany*, *105*(8), 1329–1344.

Gielly, L., Yuan, Y. M., Küpfer, P., & Taberlet, P. (1996). Phylogenetic use of non coding regions in the genus *Gentiana*: chloroplast trnL (UAA) intron versus nuclear ribosomal internal transcribed spacer sequences. *Molecular Phylogenetics and Evolution*, *5*(3), 460–466.

Graham, R. T., & Tonn, J. R. (1979). Response of grand fir, western hemlock, western white pine, western larch, and Douglas-fir to nitrogen fertilizer in northern Idaho. *USDA Forest Service*, Intermountain Forest and Range Experiment Station.

Gros-Louis, M.C., Bousquet, J., Pâques, L. E., & Isabel, N. (2005). Species-diagnostic markers in *Larix* spp. Based on RAPDs and nuclear, cpDNA,

and mtDNA gene sequences, and their phylogenetic implications. *Tree Genetics & Genomes*, *1*(2), 50–63.

Guo, Q., Li, H., Qian, Z., Lu, J., & Zheng, W. (2021). Comparative study on the chloroplast genomes of five *Larix* species from the Qinghai-Tibet Plateau and the screening of candidate DNA markers. *Journal of Forestry Research*, *32*(5), 2219–2226.

Han, K., Li, J., Zeng, S., & Liu, Z. L. (2017). Complet chloroplast genome sequence of Chinese larch (*Larix* potaninii var. *chinensis*), an endangered conifer endemic to China. *Conservation genetics resources*, *9*(1), 111–113.

Hipkins, V. D., Krutovskii, K. V., & Strauss, S. H. (1994). Organelle genomes in conifers: structure, evolution, and diversity. *Forest Genetics*, *1*(4), 179–189.

Hirao, T., Watanabe, A., Kurita, M., Kondo, T., & Takata, K. (2009). A frameshift mutation of the chloroplast mat K coding region is associated with chlorophyll deficiency in the *Cryptomeria japonica* virescent mutant Wogon-Sugi. Current genetics, *55*, 311–321.

Ishizuka, W., Tabata, A., Ono, K., Fukuda, Y., & Hara, T. (2017). Draft chloroplast genome of *Larix gmelinii* var. *japonica*: insight into intraspecific divergence. *Journal of Forest Research*, *22*(6), 393–398.

Isebrands, J. G., & Hunt, C. M. (1975). Growth and wood properties of rapid-grown Japanese larch. *Wood and Fiber Science*, 119–128.

Isoda, K., & Watanabe, A. (2006). Isolation and characterization of microsatellite loci from *Larix* kaempferi. *Molecular Ecology Notes*, *6*(3), 664–666.

Jombart, T. (2008). adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics*, *24*(11), 1403–1405.

Katoh, K., & Standley, D. M. (2016). A simple method to control over-alignment in the MAFFT multiple sequence alignment program. *Bioinformatics*, *32*(13), 1933–1942.

Khasa, P. D., Newton, C. H., Rahman, M. H., Jaquish, B., & Dancik, B. P. (2000). Isolation, characterization, and inheritance of microsatellite loci in alpine larch and western larch. *Genome*, *43*(3), 439–448.

Khatab, I. A., Ishiyama, H., Inomata, N., Wang, X. R., & Szmidt, A. E. (2008). Phylogeography of Eurasian *Larix* species inferred from nucleotide variation in two nuclear genes. *Genes & genetic systems*, *83*(1), 55–66.

Klimaszewska, K., Devantier, Y., Lachance, D., Lelu, M. A., & Charest, P. J. (1997). *Larix laricina* (tamarack): somatic embryogenesis and genetic transformation. *Canadian Journal of Forest Research*, *27*(4), 538–550.

Kim, S.C., Lee, J.W., Lee, M.W., Baek, S.H., & Hong, K.N. (2018). The complete chloroplast genome sequences of *Larix kaempferi* and *Larix olgensis var. Koreana* (Pinaceae). *Mitochondrial DNA PartB*, *3*(1), 36–37.

Kisanuki H., Kurahashi A., Kato H., Terauchi R., Kawano S., Ide Y., Watanabe S. (1995). Interspecific relationship of the genus *Larix* inferred from RFLPs of Chloroplast *DNA. Journal of the Japanese Forestry Society, 77*, 83–85.

Krüssmann, G., & Warda, H. D. (1985). Manual of cultivated conifers.

Kullman, L. (1998). Palaeoecological, biogeographical and palaeoclimatological implications of early Holocene immigration of *Larix sibirica* Ledeb. into the Scandes Mountains, Sweden. *Global Ecology and Biogeography Letters*, 181–188.

Kurinobu, S. (2005). Forest tree breeding for Japanese larch. *Eurasian Journal of Forest Research*, *8*(2), 127-134.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods*, *9*(4), 357–359.

LePage, B. A., & Basinger, J. F. (1995). The evolutionary history of the genus *Larix* (Pinaceae). Ecology and management of *Larix* forests*:* a look ahead*, 319,*19– 29.

Leister, D. (2023). Enhancing the light reactions of photosynthesis: strategies, controversies, and perspectives. *Molecular Plant*, *16*(1), 4-22.

Liu, S., Ni, Y., Li, J., Zhang, X., Yang, H., Chen, H., & Liu, C. (2023). CPGView: a package for visualizing detailed chloroplast genome structures. *Molecular ecology resources*, *23*(3), 694–704.

Lohse, M., Drechsel, O., Kahlau, S., & Bock, R. (2013). Organellar Genome DRAW a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizingexpression dataset. *Nucleic acids research*, *41*(1), 575–581.

Loveless, M. D., & Hamrick, J. L. (1984). Ecological determinants of genetic structure in plant populations. *Annual review of ecology and systematics*, 65-95.

McComb, A. L. (1955). The European larch: its races, site requirements and characteristics. *Forest Science*, *1(4)*, 298–318.

McGrath, J. A., Duijf, P. H. G., Doetsch, V., Irvine, A. D., Waal, R. de, Vanmolkot, K. R. J., & van Bokhoven, H. (2001). Hay—Wells syndrome is caused by heterozygous missense mutations in the SAM domain of p63. *Human Molecular Genetics*, *10*(3), 221–230.

Milyutin, L. I., and Vishnevetskaia, K. D. (1995) Larch and Larch Forest in Siberia. Ecology and management of *Larix* forests: a look ahead. 19–29.

Mogensen, H. L. (1996). Invited special paper: the hows and whys of cytoplasmic inheritance in seed plants. *American Journal of Botany*, *83*(3), 383–404.

Nock, C. J., Waters, D. L. E., Edwards, M. A., Bowen, S. G., Rice, N., Cordeiro, G. M., & Henry, R. J. (2011). Chloroplast genome sequences from total DNA for plant identification. *PlantBiotechnology Journal*, *9*(3), 328–333.

Ohyama, K., Fukuzawa, H., Kohchi, T., Shirai, H., Sano, T., Sano, S., Ozeki, H. (1986). Chloroplast gene organization deduced from complete sequence of liverwort *Marchantia polymorpha* chloroplast DNA. *Nature*, *322* (6079), 572–574.

Ostenfeld, C. H., & Larsen, C. S. (1930). The species of the genus *Larix* and their geographical distribution. *Kongelige Danske Vid enskabernes Selskab Biologiske Meddelelser, 9*, 1–106.

Paradis, E., Claude, J., & Strimmer, K. (2004). APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics*, *20(2)*, 289–290.

Parks, M., Cronn, R., & Liston, A. (2009). Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. *BMC biology*, *7*, 1–17.

Qian, T., Ennos, R. A., & Helgason, T. (1995). Genetic relationships among larch species basedon analysis of restriction fragment variation for chloroplast DNA. *Canadian Journal of Forest Research*, *25*(7), 1197–1202.

Ruhlman, T. A., & Jansen, R. K. (2021). Plastid genomes of flowering plants: essential principles. *Chloroplast biotechnology: methods and protocols*, 3–47

Saitou, N., & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, *4*(4), 406–425.

Schliep, K. P. (2011). phangorn: phylogenetic analysis in R. *Bioinformatics*, *27*(4), 592–593

Schorn, H. E. (1994). A preliminary discussion of fossil larches (*Larix*, Pinaceae) from the Arctic. *Quaternary International*, *22*, 173–183.

Shi, L., Chen, H., Jiang, M., Wang, L., Wu, X., Huang, L., & Liu, C. (2019). CPGAVAS2, an integrated plastome sequence annotator and analyzer. *Nucleic acids research*, *47*(1), 65–73.

Shugart, H. H., Shugart, H. H., Leemans, R., & Bonan, G. B. (Eds.). (1992*).* A systems analysisof the global boreal forest. *Cambridge University Press,78.*

Semerikov, V. L., & Lascoux, M. (2003). Nuclear and cytoplasmic variation within and between Eurasian *Larix* (Pinaceae) species. *American Journal of Botany*, *90*(8), 1113–1123.

Semerikov, V. L., Semerikov, L. F., & Lascoux, M. (1999). Intra-and interspecific allozyme variability in Eurasian *Larix* Mill. Species. *Heredity*, *82*(2), 193–204.

Semerikov, V. L., Zhang, H., Sun, M., & Lascoux, M. (2003). Conflicting phylogenies of *Larix* based on cytoplasmic and nuclear DNA. *Molecular Phylogenetics and Evolution*, *27*(2), 173–184.

Sokal, Robert R., & Charles D. Michener. A statistical method for evaluating systematic relationships. (1958). *University of Kansas, Science Bulletin*, *38*, 1409–1438.

Szmidt, A. E., Aldén, T., & Hällgren, J. E. (1987). Paternal inheritance of chloroplast DNA in *Larix*. *Plant Molecular Biology*, *9*(1), 59–64.

Tamura, K., Stecher, G., & Kumar, S. (2021). MEGA11: molecular evolutionary genetics analysis version 11. *Molecular biology and evolution*, *38*(7), 3022–3027.

Tillich, M., Lehwark, P., Pellizzer, T., Ulbricht-Jones, E. S., Fischer, A., Bock, R., & Greiner, S. (2017). GeSeq–versatile and accurate annotation of organelle genomes. *Nucleic acids research*, *45*(1), 6–11.

Turudić, A., Liber, Z., Grdiša, M., Jakše, J., Varga, F., & Šatović, Z. (2021). Towards the well-tempered chloroplast DNA sequences. *Plants*, *10*(7), 1360.

Volney, W. J. A., & Fleming, R. A. (2000). Climate change and impacts of boreal forest insects. *Agriculture, ecosystems & environment*, *82*(13), 283–294.

Warren, E., de Lafontaine, G., Gérardi, S., Senneville, S., Beaulieu, J., Perron, M., & Bousquet, J. (2016). Joint inferences from cytoplasmic DNA and fossil data provide evidence for glacial vicariance and contrasted post-glacial dynamics in tamarack, a transcontinental conifer. *Journal of Biogeography*, *43*(6), 1227–1241.

Wei, X-X., & X-Q. Wang. (2003). Phylogenetic split of *Larix*: evidence from paternally inherited cpDNA trnT-trnF region. *Plant Systematics and Evolution 239*, 67–77.

Wei X-X, Wang X-Q (2004) Evolution of 4-coumarate: coenzyme A ligase (4CL) gene and divergence of *Larix* (Pinaceae). *Molecular Phylogenetics and Evolution*, *31*(2), 542–553.

Wu, C. S., Wang, Y. N., Hsu, C. Y., Lin, C. P., & Chaw, S. M. (2011). Loss of different inverted repeat copies from the chloroplast genomes of Pinaceae and cupressophytes and influence of heterotachy on the evaluation of gymnosperm phylogeny. *Genome Biology and Evolution*, *3*, 1284–1295.

Zhang, Q., Wei, X., Liu, W., Liu, N., Zhang, Y., Xu, M., Liu, S., Zhang, Y., Ma, X., & Dong, W. (2018). The genetic relationship and structure of some natural interspecific hybrids in *Prunus* subgenus *Prunophora*, based on nuclear and chloroplast simple sequence repeats. *Genetic resources and crop evolution*, *65*, 625–636.

Zimmermann, H. H., Harms, L., Epp, L. S., Mewes, N., Bernhardt, N., Kruse, S., & Herzschuh, U. (2019). Chloroplast and mitochondrial genetic variation of larches at the Siberian tundra-taiga ecotone revealed by de novo assembly. *PLoS One*, *14*(7), e0216966.

Zybailov, B., Rutschow, H., Friso, G., Rudella, A., Emanuelsson, O., Sun, Q., & Wijk, K. J. van. (2008). Sorting signals, N-terminal modifications and abundance of the chloroplast proteome. *PLoS One*, *3*(4), e19