UNIVERSITY OF TRENTO - Italy

International PhD Program in Biomolecular Sciences
Centre for Integrative Biology
XXVIII Cycle

Phylogeny and chloroplast evolution in Brassicaceae

Tutor
Dr. Claudio Varotto
FEM- Agricultural Institute of San Michele all’Adige
Ecogenomics research group

Ph.D. Thesis of
Hu Shiliang
FEM- IASMA
Academic Year 2014-2015

Declaration

I (Hu Shiliang) confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged.

Student's signature: Hu Shiliang

Date of submission: 14/03/2016

Abstract

Brassicaceae is a large family of flowering plants, characterized by cruciform corolla, tetradynamous stamen and capsular fruit. In light of the important economic and scientific values of Brassicaceae, many phylogenetic and systematic studies were carried out. One recent and important phylogenetic analysis revealed three major lineages (I, II and III), however, classification at different taxonomic levels (tribe, genus, and species) remained problematic and evolutionary relationships among and within these lineages were still largely unclear. This is partly due to the fact that the past studies lacked information, as they mainly utilized the morphological data, nuclear DNA, partial chloroplast (cp) genes and so on. Nowadays, next generation sequencing (NGS) technology provides the possibility to make use of big data in phylogeny and evolutionary studies. Thus, we sequenced the chloroplast genomes of 80 representative species, using additional 15 reference chloroplast genomes from the NCBI database, and carried out both the phylogenetic reconstruction and the study of protein coding genes evolution in this novel dataset with different methods. Several novel results were obtained.

1 Successful application of NGS technology in chloroplast genome sequencing. During the final assembly, I could reconstruct full chloroplast genomes and the structure maps for 14 out of 80 sampled species, while the remaining were assembled nearly completely with only few gaps remaining.

2 Characterization of chloroplast genome structure. Gene number and order, single sequence repeat (SSR) as well as variety and distribution of large repeat sequence were characterized.

3 The difference of codon usage frequency was calculated between *Cardamine resedifolia* and *Cardamine impatiens*. Twelve genes with signatures of positive selection were identified at a family-wide level.

4 Three major lineages (I – III) were confirmed with high support values. Besides, the positions of various tribes were reclassified. Relationships among and within these lineages were highly resolved and supported in the final tree. Most of the tribes in the analyses were inferred to be monophyletic, only Thlaspideae was paraphyletic. Anastaticeae was for the first time classified into position of expanded lineage II, and position of tribe Lepidieae was delimited with relatively low support values in the final phylogenetic tree.

This study was a new and successful application of NGS in large-scale Brassicaceae

phylogeny and evolution, which offered the chance to look in details of the structural and functional features of the chloroplast genome. These results provided a paradigm on how to proceed towards the full elucidation of the evolutionary relationships among various biological species in the tree of life.

Index

Chapter 1: Introduction

1.1 Brief introduction of Brassicaceae: Species, Distribution, and Characteristics

Brassicaceae, also known as the mustards, the crucifers or the cabbage family, is included in the order of Brassicales according to the Angiosperm Phylogeny Group system (APG system). It contains over 372 genera and about 4,060 species [1], constituting a large and economically important family in flowering plants, showing a wide diversity of phenotypes. The most common and large genera that identified include *Draba* with 440 species, *Erysimum* with 261 species, *Lepidium* with 234 species, *Cardamine* with 233 species and *Alyssum* with 207 species. However, the taxonomic circumscriptions of many taxa are still provisional, as many of the genera having fewer number of species sampled to a relatively low depth [2].

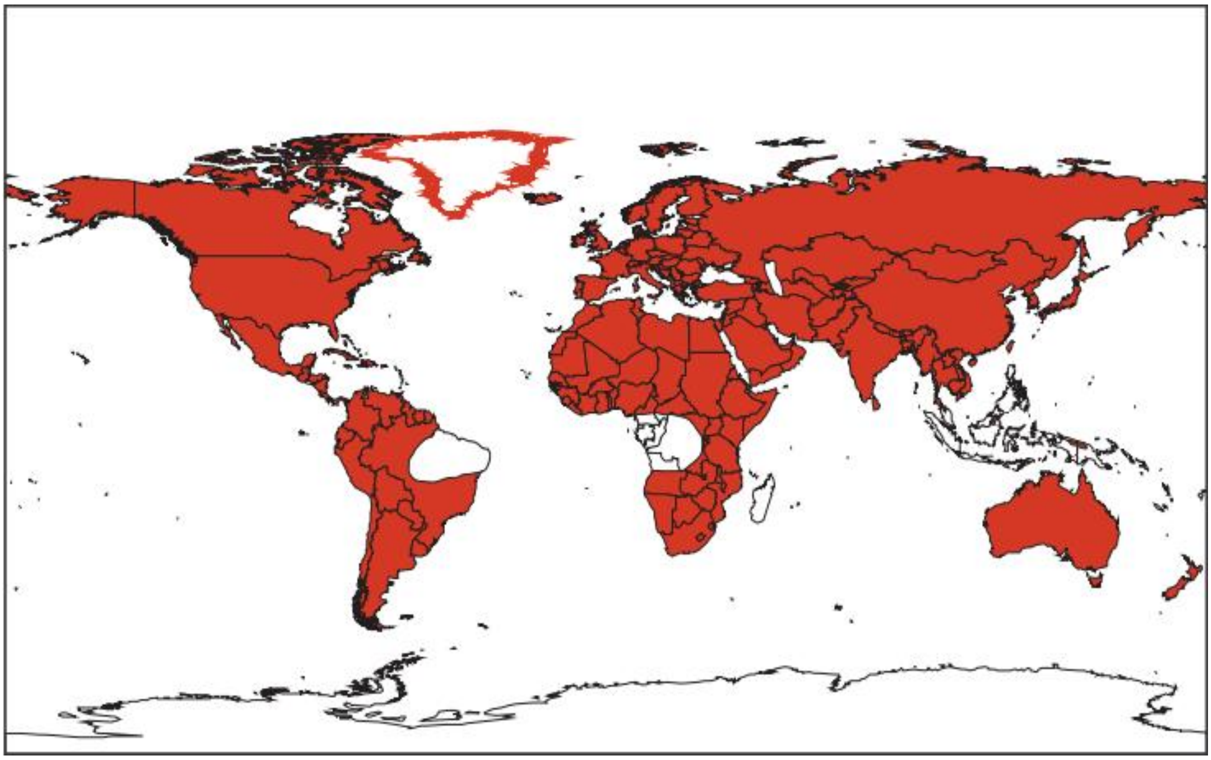


Fig. 1.1 Distribution of the Brassicaceae in the world [3]

Previous research revealed a worldwide distribution of species in this family, as all continents except Antarctica (Fig. 1.1) [2], [3] are potential habitats. Most of the species are found concentrated in the temperate regions of the Northern Hemisphere. However, many genera are more common in the southern hemisphere, such as *Draba*, *Lepidium*, and *Cardamine*. Some species, which were subsumed under a genus *Heliophila* roughly defined by Al-Shehbaz and Mummenhoff, are widely distributed in the southern (especially South African) regions, such as *Brachycarpea*, *Chamira*, *Schlechteria*, and *Silicularia* [4]. Tropics and subtropical regions, mountainous, and alpine regions are also habitats where Brassicaceae could often be found. The species *Arabis alpina*, for instance, is a representative that is widespread worldwide in the northern hemisphere, with a marked preference for mountainous, alpine and arctic habitats, including some high mountain chains in Kenya, Tanzania, Ethiopia and East Africa [5]. This worldwide distribution of Brassicaceae provides an excellent chance for various evolutionary, biogeographic or phylogeographic studies at different taxonomic levels [6].

However, a worldwide distribution inevitably leads to an unequal distribution in different regions. For instance, the Irano-Turanian region and Mediterranean region hold around 1530 highly diversified species of 263 genera. A downtrend of species diversity was found from Asia to America and Africa, from North America to South America, from the northern hemisphere to southern hemisphere [7]. This distribution revealed a potential Irano-Turanian origin of Brassicaceae [8], a place where the family possibly originated and then spread to the other parts of the globe.

Brassicaceae species can possess an annual, biennial or perennial lifespan, and consist for the large part of herbaceous plants. In the Mediterranean region, some wooden shrubs in this family have a height of 1-3 meters, such as *Zilla spinosa* and *Ptilotrichum spinosum* in northern Africa, *Dendralyssum* and *Cramboxylon* in the Dalmatian islands. *Dendrosinapis*, *Descurainia*, *Parolinia*, and *Stanleya* are the representatives of the wooden cruciferous genera in Canarias.

As an important family of the plant kingdom, the most famous and unique morphological feature of Brassicaceae is the structure of the flower, which is rather uniform throughout the family and can easily be used to distinguish it from any other family of vascular plants. Typically, the flower has four free saccate sepals and four clawed free petals, staggered and bilaterally symmetrical distributed (seldom partly zygomorphic). They are entirely disposed in cross-like arrangements (the name Cruciferae originates from this feature). The stamens are also four, with the outer two shorter than inner four (some *Lepidium* species could be different from this general rule) Brassicaceae also possess a bicarpellate and superior ovary. The flowers form ebracteate racemose inflorescences, often apically corymb-like (https://en.wikipedia.org/wiki/Brassicaceae). Only a few species and genera like *Iberis* and *Teesdalia* have an asymmetrical perianth, or *Berteroa* with divided petals [2].

In addition to the flower, the fruit is another important characteristic for species in this family, which was considered to be an important diagnostic character for delimiting and identifying taxa at different levels. In fact, the sizes, shapes, and structures of fruits all showed enormous diversity within the family. A peculiar shape like a capsule named siliqua is what the fruit usually looks like. It has two valves, and the tissue between them with the placenta form the framework, which holds the seeds. The siliqua has a length less than three times as long as its width. When a constriction happens to the segments of the fruit, it then ejects the seeds in an explosive way to increase the dissemination distance from the mother plant [2].

Other important taxonomic characters include the alternate leaves (rarely opposite), embryo characteristics (location of cotyledons and radicle), nectary glands, trichomes, chromosome numbers, growth form, and anatomy and surface of seed coat [2].

1.2 History of phylogeny and evolutionary study in Brassicaceae

1.2.1 Brief history of systematics, phylogenetics, and evolutionary research in Brassicaceae

The whole history of systematics, phylogenetics, and evolutionary research in Brassicaceae family can be divided into three main periods. The first period started from the early nineteenth to the mid-twentieth century. It provided us with comprehensive and artificial taxon descriptions. It also proposed several classification systems based on morphological data, such as replum, flower nectar, the relative position of cotyledon and radicle, pod length and trichomes. In these systems, Brassicaceae included 4–19 tribes and 20–30 subtribes [2]. The next period started from more than 30 years ago, when more species had been described, and various tribes and subtribes had been re-defined [9]. The recent period started from the early 1990s, when isozymes and increasing amount of DNA data were applied to promote significant taxonomic changes [10]. Meanwhile, the position of *Arabidopsis thaliana* as the most prominent model plant got established, which significantly promoted the intense study of the entire Brassicaceae family. In this phase, both molecular biology and DNA sequencing techniques development witnessed a revolutionary process, which had a deep impact on the fields of molecuar systematics and phylogenetics. The following paragraphs will explain in detail about the most recent achievements which have been made in Brassicaceae phylogenetics.

1.2.2 Recent and current phylogenetics in Brassicaceae

New results came with the advent of the new century. In 2001, Stevens revealed that the order Brassicales (extended order Capparales) comprises 17 families, 398 genera, roughly 4,450 species [11]. Overall Brassicales constitutes nearly 2.2% of the eudicot diversity [12] with its earliest fossil record from the Turonian [89.5 million years ago (mya)]. A more recent comprehensive angiosperm phylogeny[1] shown the family Brassicaceae comprises 372 genera, around 4,060 species. According to the strictly morphological studies in 1994, Judd and others pointed out that Brassicaceae is included within the paraphyletic Capparaceae [13]. However, molecular studies supported that Brassicaceae is sister to Cleomaceae and both are sister to Capparaceae [14]–[16]. Therefore, three families were currently recognized in Brassicales.

Inside Brassicaceae, the history of tribal classification systems has been long and well summarized in various reviews. As concluded in 2006 by Koch and Mummenhoff [17], most of the tribes in the Brassicaceae had been artificially delimited and in fact it did not reflect the phylogenetic relationships among investigated genera. In another overview, 25 tribes were newly defined by Al-Shehbaz and others in 2006 [18].

Based on the new classification, Beilstein and others for the first time combined the chloroplast gene *ndhF* and trichomes to infer the phylogeny of 113 species from 17 tribes [19]. The genus *Aethionema* was inferred to be the basal lineage. Besides, three different major, significantly supported lineages had been defined (I–III), these results had been further confirmed by Nuclear phytochrome A sequence data [20]. Later, Bailey and others used 746 nrDNA internal transcribed spacer (ITS) sequences to infer the phylogenetic relationships of Brassicaceae, representing 24 of the 25 previous recognized tribes; 13 tribes and several broadly defined genera were proved to be monophyletic while the others were clearly polyphyletic [21]. A subsequent phylogeny, based on *trnL-F*, dehydrogenase *(ADH*), chalcone synthase (CHS), internal transcribed spacer of nuclear ribosomal DNA (ITS) and plastidic maturase (*matK*), provided a supernetwork for the Brassicaceae. However, conflicting "phylogenetic signal" existed at the deeper nodes of the family tree. For instance, the contradictory placement of Cochlearieae compared to the former analysis based on *ndhF* or ITS data makes the tree not well resolved [22]. A more recent multi-gene method based on mitochondrial *nad4* intron sequence provides more insight [8]; the result was strikingly congruent with the ITS and *ndhF* based studies. Most of the tribes recognized by Al-Shehbaz and others [9] are clearly delimited, however, the support for relationships of different tribes was not high.

Although a complete tribal classification system of Brassicaceae is not yet available, we are gradually approaching this goal. Tribal adjustments based on most comprehensive modern tribal classification were done by Al-Shehbaz and the rest [18]. At the same time, several extra studies provided supplementary information.

Al-Shehbaz and Warwick and the rest [23] [24] showed the Anchonieae and the Euclidieae each separate into two distinct and distant clades (appointed here as Anchonieae I and II and Euclidieae I and II) and were newly defined as the Malcolmieae, Dontostemoneae and the reestablished Buniadeae (Fig. 1.2).

German and Al-Shehbaz [25] proposed the new tribes Aphragmeae and Conringieae, and reestablished Biscutelleae, Calepineae and Erysimeae (tribes 29–33, Fig. 1.2). ITS studies of Bailey [26] and Warwick [27] confirmed the recognition of the last tribe Erysimeae as monophyletic, their findings showed that the tribe Camelineae was weakly supported and paraphyletic, because tribes Boechereae and Halimolobeae were nested inside. These results were inconsistent with the *ndhF* phylogeny of Beilstein [20], but were in full agreement with results from Bailey [26] and the phyA phylogeny [20]. The Camelineae was not supported as monophyletic in the phyA phylogeny and needed to be divided into a few smaller ones, herein recognized as 2 tribes, 34 (2A) and 35 (2B) here (Table 1.1,Fig. 1.2) [20].

Furthermore, the results from Warwick and others [27] fully supported the recent finding that tribes Schizopetaleae and Thelypodieae were two distinct tribes instead of a single tribe, also provided some support for the reestablishment of the tribe Cremolobeae, raising the total number of tribes up to 44 in the family. The supermatrix approach adopted by Couvreur [28] suggested that an early rapid radiation within Brassicaceae led to the unresolved backbone of the phylogenetic tree. These two recent study supported these monophyletic, well-supported lineages, lineage I include 13 tribes, lineage II four tribes and lineage III seven tribes [29].

While in 2012, a further phylogenetic research was carried out by using four plastidic regions (*rpl32-trnL*, *atpI-atpH*, *psbD-trnT*, and *ycf6-psbM*) in the tribe Brassiceae for 89 species. Eight well-supported clades were recognized. Meanwhile, relationships within and between the eight major clades were strongly supported for the first time [30].

Table 1.1 Number of genera and species described within the Brassicaceae family

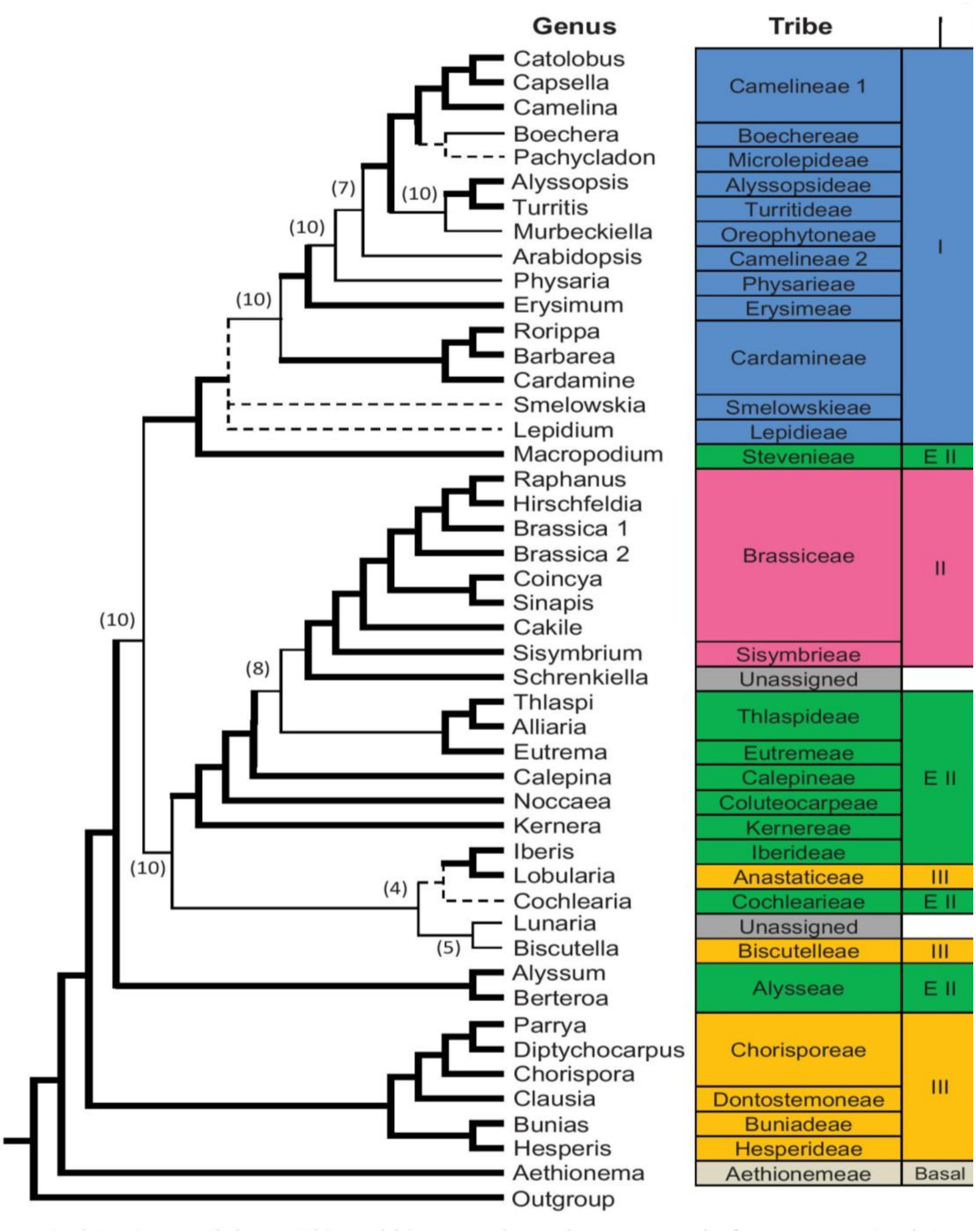
	No.of Genera	No. of Species	References
1Aethionemeae	1	45	Koch and Al-Shehbaz (2009)
2Camelineae	7	35	Koch and Al-Shehbaz (2009)
3Boechereae	7	118	Al-Shehbaz et al. (2006)
4Halimolobeae	5	39	Bailey et al. (2007)
5Physarieae	7	133	Koch and Al-Shehbaz (2009)
6Cardamineae	9	333	Koch and Al-Shehbaz (2009)
7Lepidieae	4	235	Koch and Al-Shehbaz (2009),Warwick et al. (2008)
8Alysseae	15	283	Koch and Al-Shehbaz (2009)
9Desurainieae	6	57	Al-Shehbaz et al. (2006)
10Smelowskieae	1	25	Al-Shehbaz et al. (2006)
11Arabideae	8	470	Koch and Al-Shehbaz (2009)
12Brassiceae	46	230	Al-Shehbaz et al. (2006)
13Schizopetaleae	28	230	Al-Shehbaz et al. (2006)
14Sisymbrieae	1	40	Al-Shehbaz et al. (2006)
15Isatideae	2	65	Koch and Al-Shehbaz (2009)
16Eutremeae	1	26	Warwick and Al-Shehbaz (2006)
17Thlaspideae	7	27	Al-Shehbaz et al. (2006)
18Noccaeeae	3	90	Koch and Al-Shehbaz (2009)
19Hesperideae	1	45	Al-Shehbaz et al. (2006)
20Anchonieae	8	68	Al-Shehbaz and Warwick (2007)
21Euclidieae	13	115	Al-Shehbaz and Warwick (2007)
22Chorisporeae	3	47	Al-Shehbaz and Warwick (2007)
23Heliophileae	1	80	Al-Shehbaz et al. (2006)
24Cochlearieae	1	21	Al-Shehbaz et al. (2006)
25Iberideae	1	27	Al-Shehbaz et al. (2006)
26Malcolmieae	8	37	Al-Shehbaz and Warwick (2007)
27Buniadeae	1	3	Al-Shehbaz and Warwick (2007)
28Dontostemoneae	3	28	Al-Shehbaz and Warwick (2007)
29Biscutelleae	1	53	German and Al-Shehbaz (2008)
30Calepineae	3	8	German and Al-Shehbaz (2008)
31Conringieae	2	9	German and Al-Shehbaz (2008)
32Erysimeae	1	180	German and Al-Shehbaz (2008)
33Aphragmeae	1	11	German and Al-Shehbaz (2008)
34Unnamed(Camelineae2A)	2	5	Koch and Al-Shehbaz (2009)
35Unnamed(Camelineae2B	3	20	Koch and Al-Shehbaz (2009)
Total	212	3,249	

Compiled by Warwick and others and represented nearly two-thirds (62.7%) of the 338 genera and 87.6% of the 3,709 species [31]

Based on complete chloroplast sequences of 29 Brassicaceae species, a comprehensive time-calibrated framework was obtained with important divergence time estimation, which shown the diversification of the Brassicaceae crown group started at the

Eocene-to-Oligocene transition, also, the age of the Arabidopsis thaliana crown group was 6 million years ago. The species richness of the family was well explained by high levels of neopolyploidy and species radiation, paralleled by high levels of neopolyploidization, following genome size decrease, stabilization and genetic diploidization [32]. In 2015, Huang and colleagues used nuclear markers of 55 species spanning 29 out of 51 tribes in Brassicaceae, proposed a highly supported phylogeny with six major clades, from A to F [33].

Fig. 1.2 A summarized Brassicaceae phylogeny in 2015 [33]



An outline of these various tribes and a synopsis of their relationships of Brassicaceae family were presented by Fig. 1.2. However, a high and deep resolution outcome that shows clearly relationship among species in a comprehensive family-wide level is still lacking.

1.3 Brief introduction of data collection and approaches for phylogenetic analysis

1.3.1 Traditional data resource and chloroplast genetics

1.3.1.1 Comparision of traditional types of data used for phylogeny study

A correct understanding of the evolutionary relationship among different forms of organisms is not only the premise of evolutionary biology research, but also the foundation of taxonomy and basis of study in others branches of biology [34].

Early phylogenetic scientists through the study of the fossil record, comparative morphology and physiology, constructed a primary evolution framework for various kinds of species [35]. For instance, length and width ratio of fruit as a unique character had been applied to distinguish the species, however, that had been approved as arbitrary and had no phylogenetic implications [9]. After the 1980s, with the rapid development of molecular biology, phylogeny relied more and more on the molecular biological data, namely the use of biological macromolecules information (e.g DNA Sequence, Amino acid sequence, etc.), to infer the evolutionary history of organisms. Compared to the former comparative morphological study, this method was easier to operate.

Because of the complexity of the nuclear genome, the screening of single copy (or low copy) genes was rather difficult [36]. Sequencing of the nuclear genome of the plant is currently only limited to rather small genomes of model species or species with important economic or ecological values, such as Arabidopsis, Tobacco, Rice, Corn, Snapdragon, etc., So the development of large-scale phylogenomics frameworks is still lagging behind. As for the mitochondrial genome, its application in plant phylogenetic studies has been limited by several inner unique features; the large size variation of the mitochondrial genome in different plant taxa (range of mitochondrial genome size: 300-600 kb), plus the insertion of foreign genes caused by horizontal transfer among genomes are possibly the two main reasons. Also, the mitochondrial genome evolved at a pace which is generally too slow to provide enough informative characters at the taxonomic levels usually investigated (family, genus) in plants [37]. In fact, the nucleotide substitution rate is on average four times slower than that of the chloroplast genome. An additional problem is that the intramolecular recombination phenomenon has also been widely documented in plant mitochondrial genomes [42] [43], which can seriously complicate its reliability for phylogenetic reconstruction. Therefore, the mitochondrial genomes are not the ideal candidate for phylogenomic study in plants. In contrast, the plant chloroplast genome sequence is widely used in molecular evolution and phylogenetic studies, as it has many advantages. First, the chloroplast genome is large enough to contain a significant amount of genetic information. Second, although the nucleotide substitution rate of chloroplast DNA is moderate, the molecular evolution rate between the coding regions and non-coding regions of the chloroplast genome show significant differences, which could then be applied to taxonomic studies at different levels [40]. Also, the chloroplast genome size is moderate (or small) as compared to the mitochondrial and nuclear genomes, respectively. The high number of copies per cell also makes it easy for sequencing. What's more, a good co-linearity exists among different plant groups, which greatly simplify the alignment and comparative analyses. For all these reasons, there is a fast development in the phylogenetic studies based on plant chloroplast genome in recent years [41].

1.3.1.2 Genetics and evolution of chloroplast genome and its application in phylogeny

Chloroplasts are one of many types of organelles in a plant cell. It is the place where both the light and dark phases of photosynthesis take place. The photosynthetic process mediates the fixation of atmospheric carbon dioxide (CO_2) into organic compounds and the production of metabolic energy, which is the basis of green plants' life [42]. Besides, many other fundamental biosynthetic processes also take place here, e.g. the production of lipids, isoprenoids, hormones, cofactors, etc., making them one of the most important cellular compartments [43]. Chloroplasts are considered to be originated through endosymbiosis happened to cyanobacteria [44], It is currently accepted that the origin of chloroplasts happened multiple times during the radiation of plants. According to this view, free-living cyanobacteria entered and were permanently engulfed into an early eukaryotic cell in at least three independent occasions, thus giving rise to a total of three independent chloroplast lineages—the green algae, red algae and glaucophyte algae [45]. During the process of different endosymbiotic events, chloroplasts had transferred varying amounts of their DNA to the nucleus of their host. The *Arabidopsis thaliana* genome sequence revealed that up to 18% of the nuclear genes are of endosymbiotic origin, as they were progressively transferred from the cyanobacterial ancestors of chloroplasts during the progressive sub-functionalization of the engulfed bacterial cells [46]. During this process, however, many genes were lost, resulting in the marked metabolic differences currently observed between angiosperms and cyanobacterial cells [47].

As a separate organelle of bacterial origin, the chloroplast has its own set of genes organized in a genome of its own, namely the plastome. The plastome is paternally inherited in gymnosperms while is mostly maternally inherited in angiosperms. Biparental inheritance is rare in angiosperms, accounting for about 14% of all species [48]. The chloroplast genome is generally a covalently closed circular DNA, existing in the form of multiple copies per organelle. Most chloroplast genomes of the terrestrial plants are highly conserved and organized in a tetrad structure [41], which consists of two inverted repeats (inverted repeat sequence, IR), a large single copy region (large single copy, LSC), and a small single copy region (small single copy, SSC). Only in a few Fabaceae species such as ground clover (*Trifolium subterraneum* L.), terrestris clover Alfalfa (*Medicago truncatula* Gaertn.), chickpea (*Cicer arietinum* L.), the chloroplast genome has a special structure because of a complete loss of the inverted repeat region [49]. The general size of chloroplast genome is around 120-170 kb, encoding about 110-140 genes, mainly involved in gene expression and photosynthesis and some open reading frames with yet unidentified function. The chloroplast genome of land plants is relatively conserved in gene number and order, but some groups have their own unique features. For example, the *ndh* gene family has been completely lost in the *Pinus* chloroplast genome [50]. On the other hand, the chloroplast genome of the parasitic plant species *Epifagus virginiana* (L.) Bart in the Orobanchaceae family contains only 42 genes, as photosynthesis and respiration-related genes and RNA polymerase genes encoded by the chloroplast were lost [51].

The past phylogenetic reconstructions with chloroplast genome were applied mostly to higher taxonomic levels (above order, family or subfamily), and had contributed to solving many problematic evolutionary relationships in molecular systematics. Such chloroplast phylogenomics had, for instance, addressed the relationships of the main branches of the core eudicots in angiosperms [52]. Due to the rather limited availability of chloroplast genomics data, so far, the application of chloroplast genomes to the phylogenetic reconstruction of low taxonomic levels phylogenetic relationships was rarely used. However, with the development of next-generation sequencing technology, the advantage of using cp genome in phylogeny for closely related taxa began to show up. Parks and others sequenced 37 *Pinus* chloroplast genomes and closely related species by next-generation sequencing technology [53]. The new phylogenomic approach resulted in a great improvement of internal branch resolution in the pine genus when compared to the earlier phylogenetic results, which were based on several molecular fragments. Other examples of the successful application of complete chloroplast phylogenomics at low taxonomic levels were those in *Acacia*, *Tanaecium*, and many others [54]–[56]. Thus, phylogenomics based on whole chloroplast sequences have great potential in phylogenetic studies at lower taxa levels. As the number of chloroplast genomes records grows, phylogenetic reconstructions are expected to reflect more and more closely the real evolutionary history of plants, thus providing more evidence for the study of biological evolution and the elucidation of the plant portion of the tree of life [57].

1.3.2 Principle of phylogenetic inference and common phylogeny programs

Phylogenetics is the study of the evolutionary history and relationships among individuals and groups of organisms. The principle underlying phylogenetic inference is quite simple: Analysis of the similarities and differences among biological entities can be used to infer the evolutionary history of those entities. Nowadays, these kinds of analyses are mainly carried out on the sequences with genetic information, which can represent these biological entities. Usually, the first step is to obtain sequences of interested, and these sequences should be homologous. Next critical step is the multiple sequence alignment, which has to reflect the homology (i.e. genetic variants inherited by speciation from a common ancestor) of the aligned characters. Once the multiple sequence alignment has been validated and, in case, manually edited, a suitable phylogenetic reconstruction method should be chosen for analysis. The many and remarkably diverse methods for molecular phylogeny can be classified into the following main categories, according to their specific features: distance, likelihood, parsimony, and bayesian methods. The following paragraphs provide a general outline of these approaches and of the corresponding programs used for analysis.

PAUP (Phylogenetic Analysis Using Parsimony), as a computational phylogenetics program for inferring evolutionary trees (phylogenies), is one of the most widely used package written by David L. Swofford [58]. Originally, as the name implies, PAUP only implemented parsimony, but from version 4.0 it also supports distance matrix and likelihood methods. PAUP was the preferred choice of many phylogenetists [59], but the development of more recent methods and program packages eliminated the nearly exclusive monopoly existing in the past.

More recently, a software, PhyML, estimates maximum likelihood phylogenies from alignments of nucleotide or amino acid sequences has become popular. Its main strength lies in the large number of substitution models, which combined with various options, allowing a thorough search of phylogenetic trees with multiple choices, from very fast and efficient approaches to slower but generally more accurate methods. The capability of PhyML was designed to vary from moderate to large datasets. In theory, datasets with less than 4,000 sequences X 2,000,000 characters can be processed [60].

Another popular choice for phylogenetic reconstruction is MrBayes, a program for Bayesian inference and model selection in a wide range of phylogenetic and evolutionary models. MrBayes uses Markov chain Monte Carlo (MCMC) methods to estimate the posterior distribution of model parameters [61].

CodonPhyML, a fast maximum likelihood package, provides hundreds of different codon models with the largest variety so far for phylogeny inference by maximum likelihood. CodonPhyML was tested with simulated and real data, which convincingly showed the excellent speed and convergence properties offered by the program. In addition, CodonPhyML includes the most recent methods for estimating phylogenetic branch support, which provides an integral framework of models selection, including amino acid and DNA models [62].

In addition to the very popular software packages described above, there are several others. For sake of completeness only the most common among them will be here briefly cited. IQ-TREE provides a fast and effective evolutionary algorithm for inferring maximum likelihood phylogeny. The PAML package applies the maximum likelihood method for inferring phylogeny and identifying signatures of positive selection. QuickTree allows fast phylogenetic reconstructions using the Neighbor-joining method. MEGA (Molecular Evolutionary Genetics Analysis), is an integrated software released by Kumar and collaborator for conducting automatic and manual sequence alignment, inferring phylogenetic trees, mining web-based databases, estimating rates of molecular evolution, inferring ancestral sequences and testing evolutionary hypotheses by distance, parsimony and maximum composite likelihood methods [63].

1.4 New opportunities for in-depth phylogeny and evolution studies

1.4.1 Development of high-throughput sequencing technology and its application in phylogeny and Molecular evolution study

DNA as the main genetic material is an informative macromolecule responsible for passing the information from one generation to the next. Deciphering DNA is of paramount importance for many branches of biological research [64]. Capillary electrophoresis (CE)-based Sanger sequencing brought a revolution in the amount of DNA data which could be produced from virtually any organism of interest. As the most effective tool, this technology is widely established in laboratories around the world. However, nuclear genetic information usually largely exceeds the limited throughput offered by Sanger sequencers, which often hinders the obtainment of sufficiently large datsets. With the progress made in DNA sequencing technology, Next Generation Sequencing (NGS) became the most efficient tool to overcome these shortcomings.

In theory, the concept of NGS technology and CE are similar: Each base of short fragments of DNA is identified by means of a photochemical signal, which is then integrated into a DNA sequence following the order of bases present in the template. The same process will happen with millions of reactions in a massively parallel way, which is the most critical step in NGS and enables rapid sequencing of large genomes. With the latest instruments, it is feasible to produce hundreds of billions of base pairs of data in a single sequencing run.

In light of this obvious potential, NGS should be promoted to take root in phylogenetics as it already did in other fields like metagenomics and disease genetics [65].

So far, the complete genomes of thousands of species have been fully sequenced. According to a recent update in Wikipedia, there are fully sequenced genomes for, eukaryotic organisms, like fungi, plants and animals, as well as for protists, including archaea and eubacteria. For example, the link: https://en.wikipedia.org/wiki/List_of_sequenced_plant_genomes provides a summary can be found by checking “list of sequenced plant genome”.

As for record of the chloroplast genome in the Brassicaceae family, the number of sequenced chloroplast genome has been recorded over a long time. From 2010, the number of plastid genome record in NCBI database has increased even faster than the Moore's law. At present, the number of available chloroplast genomes in the Brassicaceae family is 40

(Fig.1.3). With the obtainment of complete and accurate sequence, these reference genomes can be reliably used as references in the process of new chloroplast genome assembly and phylogenetic tree reconstruction. Just five years after the introduction of NGS technology, there has been a revolutionary change in the protocol for scientists to extract genetic information from different organisms, which significantly accelerated the research and revealed stintless insight about the genome, transcriptome and epigenome of species. This capability has stimulated quantity of key breakthroughs, promoting a wide range of scientific studies from human health to agriculture and more. NGS is currently recognized as an indispensable and universal tool for biological research. With the capability to extract genetic information from any biological entity, the currently available NGS sequencing platforms are already making possible to unlock information never previously imaginable.

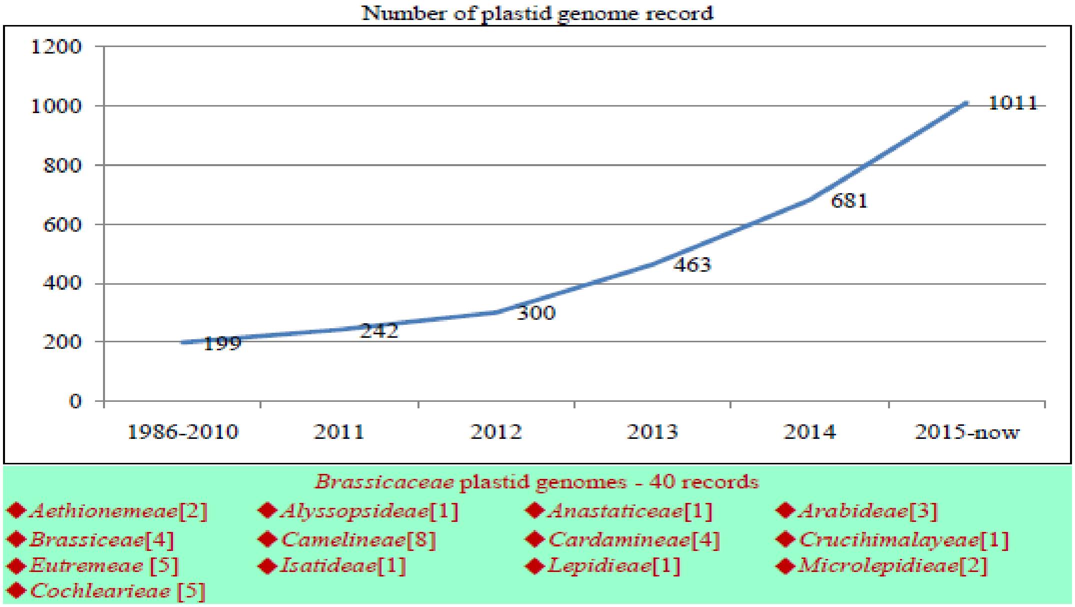


Fig.1.3 Available plastid genome records in NCBI database and Brassicaceae family by 02/2016

1.4.2 Established bioinformatic techniques promoted the phylogeny and molecular evolution studies

When dealing with the sequencing of whole chloroplast genomes, the amount of data produced requires the use of advanced bioinformatic approaches, which will automate at least part of the computational process of the analyses. The processing of the raw reads obtained from the Illumina sequencers is usually the first step in the analysis, as it is necessary to remove the fraction of incomplete and noisy reads that present in any NGS dataset. The random shearing of template DNA prior to library preparation ensures a sufficient coverage of the template DNA. Multiple reads will distribute randomly along the complete length of the plastome. Once the assembly of the plastome sequence is completed, it is normally necessary to check for its reliability using the plastome sequences of closely related species as references.

Genome annotation is another important aspect of genome characters mining. For the chloroplast, several tools have been developed which are dedicated for this purpose, as the dual organellar genome annotator (DOGMA)[66], and CpGAVAS [67], but sometimes the annotation can be carried out also by means of custom and stand alone programming scripts. Among them, DOGMA, being the first tool which automates the annotation process with reference databases for genes from 16 complete genomes of green plants, is the most popular web-based annotation tool for chloroplast genomes.

Common methods for phylogenetic inference involve distance-matrix approaches such as neighbor-joining or UPGMA, which collect the genetic distance from multiple sequence alignments, are easy to carry out, but do not adopt an evolutionary model. The other sequence alignment methods like ClustalW infer trees by adopting the easy algorithms (i.e. those based on distance). Another simple method for estimating phylogenetic trees is maximum parsimony, which implies an implicit model of evolution. Advanced methods apply the optimality criterion of maximum likelihood, usually within a Bayesian framework, which applies an explicit model of evolution to phylogenetic tree estimation [68]. Identifying the optimal tree using many of these techniques is NP (Non-deterministic Polynomial)-hard, so the combination of a heuristic search and optimization methods with tree-scoring functions is usually regarded as a better way to find a reasonably good tree which fits the data [68].

1.4.3 New strategy for the phylogeny and molecular evolution studies

New technologies and methods always promote the study of phylogeny and molecular evolution. In the past, the access to data sources and computational phylogenetics methods were the two main constraints to phylogenetic reconstruction. But now, with the continuous improvement in sequencing technology and the flourishing of bioinformatics, theses limitations are no longer so obvious.

The new high-throughput sequencing technology, which is able to detect the exact nucleotides and bases component of both DNA and RNA sequence, offers us the opportunity to get huge amounts of data. These data include nuclear DNA, chloroplast DNA, mitochondrial DNA, and other transcriptome sequences. Besides, after so many years of development, the common phylogenetic methods have been widely applied to various types of phylogenetic analyses. Especially during the past decade, the development of new statistical methods and advances in computational technology has promoted a remarkable progress in the study of molecular evolution [69].

Despite these, on how to choose the phylogenetic methods and data, different strategies will produce different results.

First, the selection of data set can significantly affects the results of a phylogenetic and evolutionary analysis. Before the invention of high-throughput sequencing technology, simple of nucleic acid sequences or morphological data was the widely choice for phylogenetic analysis. However, phylogenetic hypotheses based on single markers (e.g., plastidial, mitochondrial or nuclear) possessed a limited value [70], [71]. Besides, with the increasing availability of molecular data, the molecular systematic results for the same organisms based on different molecular fragments often turned out to be different. This brought about the realization that single gene trees, although they can reflect the evolutionary history of the organisms to some limited extent, very seldom provide a reliable approximation to species trees [69], [72].

Secondly, software based on different algorithms for phylogenetic and molecular evolution analysis will not always produce consistent results, and sometimes it can lead to very different ones. The problem most often encountered is that it is not clear which method is more suitable for each data set and which is more precise in inferring the evolutionary relationships among taxa. Given the above-mentioned problems, several criteria have been used to assess the tree-building methods [73]. The first is the computational efficiency. In other words, this criterion takes into account the memory and time required by the algorithm for the reconstruction. Of course, this criterion provides only a technical estimation of efficiency, which can be useful in cases where computational resources are limiting. Possibly more importantly, one should assess whether the software used made an efficient use of the data. Besides, the consistency among several simulated repeats of the same method, e.g. through bootstrapping or jackknifing approaches [74], is also a very important criterion on to assess the reliability of the reconstructed phylogenies. Last but not least, how the method deals with the violations of the assumptions and how the violation affects the result of the phylogenetic reconstruction should also be taken into account [73].

For these above-mentioned reasons, what is currently the most up to date strategy for phylogenetic and molecular evolution analysis?

Currently, the technology for the high-throughput sequencing to obtainment of the whole genome sequence for an organism of choice is no longer a problem, but sequencing of an entire genome of a plant is still an expensive practice. Until now it has been achieved only for a limited number of plant species. However, it is quite feasible to determine the sequence of chloroplast or mitochondrial genomes, whose sizes vary from 100 kbp to 1000 kbp. In the past few years, chloroplast sequence has frequently been used in molecular evolution studies in Brassicaceae, as summarized in the synopsis of these works done in 2012 by Renate Schmidt and Ian Bancroft[2]. Besides, the use of multiple loci to infer population and species histories has also been increasingly adopted, especially with the combination of sequences from different genomes of plant cells (nuclear, mitochondrial and chloroplast genomes), the selection of molecular fragments having different functions, or the use of morphological data in the analyses. Multilocus studies in phylogenetics benefited firstly from the decreasing costs of DNA sequencing in the last three decades. More recently, the theoretical justification for incorporating information from multiple loci into estimates of population and species history did provide convincing results [75], and this phenomenon attracted the interest from many practitioners of phylogenetics to spend a significant portion of their time on developing and screening molecular markers suitable for their study system. With the increasing availability of molecular markers for non-model organisms [76], [77], the process of data generation for a multilocus study became less laborious.

Many phylogenetic researchers are already using NGS technologies [78], [79]. With sequencing technologies becoming increasingly sophisticated, more and more genomic sequences have been obtained, thus progressively promoting the transition from a traditional phylogenetic study into a new era - the era of phylogenomics. Phylogenomics is the new discipline resulting from the combination of phylogenetics and genomics. The main tasks include studying the phylogenetic relationships with large-scale molecular biological data on a genomic scale and conversely using the deduced evolutionary relationships to study genome evolutionary mechanisms (such as DNA repair process, annotation of unknown gene functions, expansion and contraction of non-coding sequences, genome rearrangements, etc.). In more detail, phylogenomic methods mainly include: (1) sequence analysis, consistently with the traditional method of phylogenetic analysis, which is still widely used. (2) Non-sequential analysis, which includes the analysis of genome-wide characteristics (whole-genome features, WGFs), namely the analysis of composition of genes, gene order and the oligonucleotide sequence distribution in the genome; rare genomic changes (Rare genomic changes, RGCs), such as indels, presence or absence of introns, transposon insertion, gene fusion and fracture events. These rare genomic variations can be used to construct phylogenetic trees and can also be used to offer special support for certain nodes [80].

1.5 Aim of the study

The phylogenetic reconstruction of the Brassicaceae family and the elucidation of the evolutionary patterns affecting its plastome is important because of the large number of species and the insight they can provide in the dissection of the evolution of this model family. Given the fact that Brassicaceae is extremely important from an economical point of view, these studies also have the potential to have practically important applications, especially when considering the close uses of many of these species in our daily lives.

However, until now, the evolutionary relationships of the species in Brassicaceae are still not fully resolved (as described above). The coming of age of high-throughput sequencing technology brings us quantities of high-quality data. At the same time, the development of phylogeny theory and the availability of more tools also provide new methods to address the unanswered questions about the evolution of the family.

In my PhD project, I took three steps with the final aim to contribute to the elucidation of the phylogenetic relationships. The first step was to focus on two species *Cardamine resedifolia* and *Cardamine impatiens,* which show distinct life history traits and habitats. According to the previous study in our lab [81], there were different selective pressures acting on different functional gene, reflecting a faster evolution in cold-related genes exclusively in the high altitude species *Cardamine resedifolia* . To extend to organelles the knowledge of the positive selection signatures detected in the previous transcriptome-wide analysis, we sequenced the full chloroplast genome of *Cardamine resedifolia* and *Cardamine impatiens.* The structure of the whole chloroplast genome and their gene space and repeat patterns were analyzed, and the patterns of natural (positive) selection and phylogenetic position of the species were determined. These results are reported in chapter two.

The second step takes the lead from the previous one, as both *Cardamine resedifolia* and *Cardamine impatiens* belong to the tribe Cardamineae, which is an important and large tribe in the Brassicaceae family. As a prelude to a larger scale sampling and plastome sequencing in Brassicaceae, 14 species (including *Cardamine resedifolia* and *Cardamine impatiens*) from the Cardamineae were selected here to test the best methods for high-throughput chloroplast genome assembly, to analyze in detail the genome structure and to infer the preliminary phylogeny of this tribe. These will provide a reference for the data mining in chloroplast genome and for selection of the strategy for final phylogenetic tree reconstruction. These resulst will be described in chapter three.

The third step encompasses the phylogenetic reconstruction and the elucidation of chloroplast evolution patterns with samples of 80 new species in Brassicaceae (including the Cardamineae from step two), plus 15 available reference chloroplast genome sequences from the NCBI plastid database. Given the tested analytical methods and procedures provided by the previous two steps, the final large-scale study is believed to bring solid and informative results. This conjecture will be verified and discussed in chapter four.

Chapter 2: Plastome organization and evolution of chloroplast genes in Cardamine species adapted to contrasting habitats

2.1 Introduction

Cardamine resedifolia, found at high altitudes (1500-3500 meters above sea level.; Mixed mating system; Full sunlight), and *Cardamine impatiens*, found at low altitudes (0-1500 meters above sea level; Selfing; Shadow) (http://es.wikipedia.org/), are two species in the Brassicaceae family which show divergent habitat preference [82]. To accommodate to these different living environments, they both evolved with a suit of adaptive responses, which reflects the evolutionary pressure imposed upon them by natural selection. The genetic basis of this adaptation lies in the selective constraints acting on genes related to the traits under selection, but little is known at present about what these genes and traits are. Molecular evolution analyses can be applied to the elucidation of the mechanisms underlying adaptation.

In 2012, a transcriptome-wide molecular evolution analysis was carried out, which focused on genes that are involved in stress responses to two factors differentiating the high- and low-altitude habitats. It revealed important lineage-specific patterns that may be associated with the distinct life history traits and habitats of *Cardamine resedifolia* and *Cardamine impatiens*, and also revealed the difference of selection pressure on the analyzed genes both on transcription and protein expression levels [81]. It further explicitly demonstrated a faster evolution of the cold-related genes (indicating either positive or relaxed selection) and a slower evolution of the photosynthetic genes (indicating purifying selection) exclusively in the high altitude species *Cardamine resedifolia*.

To extend the knowledge of positive selection signatures observed in the above-mentioned transcriptome-wide analysis and verify whether analogous signatures can be identified also in chloroplast-encoded genes, the sequencing of the whole plastome of both species was carried out. The results on elucidation of plastome organization and chloroplast genes evolution in these *Cardamine* species adapted to contrasting habitats are reported in chapter two and the appended paper. This will provide the technical basis for data analysis and a training set for the subsequent phylogeny and molecular evolution study in the tribe Cardamineae and at the level of the whole family.

2.2 Authors’ specific contributions for the appended paper

BW, DQ and EB helped to carry out lab work and draft the manuscript. ML contributed to conceive and design of the study, carried out all the phases of lab work, helped to draft the manuscript. GS carried out data analyses, drafted the manuscript. HS carried out data analyses, such as chloroplast genome assembly, chloropalst genome annotation and map production, codon usage estimation, molecular evolution and phylogeny reconstruction, and helped to draft the manuscript. RV helped to draft the manuscript. CV conceived, designed and coordinated the study, finalized the manuscript. All authors read and approved the final manuscript.

Hu *et al. BMC Genomics* (2015) 16:306
DOI 10.1186/s12864-015-1498-0

BMC Genomics

RESEARCH ARTICLE Open Access

Plastome organization and evolution of chloroplast genes in *Cardamine* species adapted to contrasting habitats

Shiliang Hu[1†], Gaurav Sablok[1†], Bo Wang[1], Dong Qu[1,2], Enrico Barbaro[1], Roberto Viola[1], Mingai Li[1] and Claudio Varotto[1*]

Abstract

Background: Plastid genomes, also known as plastomes, are shaped by the selective forces acting on the fundamental cellular functions they code for and thus they are expected to preserve signatures of the adaptive path undertaken by different plant species during evolution. To identify molecular signatures of positive selection associated to adaptation to contrasting ecological niches, we sequenced with Solexa technology the plastomes of two congeneric Brassicaceae species with different habitat preference, *Cardamine resedifolia* and *Cardamine impatiens*.

Results: Following in-depth characterization of plastome organization, repeat patterns and gene space, the comparison of the newly sequenced plastomes between each other and with 15 fully sequenced Brassicaceae plastomes publically available in GenBank uncovered dynamic variation of the IR boundaries in the *Cardamine* lineage. We further detected signatures of positive selection in ten of the 75 protein-coding genes of the examined plastomes, identifying a range of chloroplast functions putatively involved in adaptive processes within the family. For instance, the three residues found to be under positive selection in RUBISCO could possibly be involved in the modulation of RUBISCO aggregation/ activation and enzymatic specificty in Brassicaceae. In addition, our results points to differential evolutionary rates in *Cardamine* plastomes.

Conclusions: Overall our results support the existence of wider signatures of positive selection in the plastome of *C. resedifolia*, possibly as a consequence of adaptation to high altitude environments. We further provide a first characterization of the selective patterns shaping the Brassicaceae plastomes, which could help elucidate the driving forces underlying adaptation and evolution in this important plant family.

Keywords: *Cardamine*, Molecular adaptation, Large single copy region (LSC), Small single copy region (SSC), Plastomes, Positive selection, Repeats, Codon usage

Background

Chloroplast genomes, hereafter referred to as plastomes, have been widely used as models for elucidating the patterns of genetic variation in space and time, ranging from colonization to speciation and phylogeny, encompassing both micro- and macro-evolutionary events across all lineages of plants [1]. Understanding the phyletic patterns of chloroplast evolution can also potentially layout the basis of species discrimination [2], as indicated by the fact that the core DNA barcode chosen for plants is composed by the two plastomic regions *rbc*L and *mat*K [3]. In fact, the presence of a high number of plastomes per cell, ease of amplification across the angiosperm phylogeny, and good content in terms of phylogenetic information explain the popularity of these and other plastidial markers for both species identification and phylogenetic reconstruction. The organization of the plastome is remarkably conserved in higher plants, and it is characterized by two usually large inverted repeat regions (IR_A and IR_B) separated by

* Correspondence: claudio.varotto@fmach.it
†Equal contributors
[1]Ecogenomics Laboratory, Department of Biodiversity and Molecular Ecology, Research and Innovation Centre, Fondazione Edmund Mach, Via E. Mach 1, 38010 S Michele all'Adige (TN), Italy
Full list of author information is available at the end of the article

single copy regions of different lengths, called large single copy region (LSC) and small single copy region (SSC; [4]). Both traditional Sanger sequencing and next generation sequencing approaches have been widely employed to elucidate the dynamic changes of these four plastome regions, revealing patterns of evolutionary expansion and contraction in different plant lineages [5,6]. The genes present in plastomes play fundamental functions for the organisms bearing them: they encode the core proteins of photosynthetic complexes, including Photosystem I, Photosystem II, Cytochrome b_6f, NADH dehydrogenase, ATP synthase and the large subunit of *RUBISCO*, tRNAs and ribosomal RNAs and proteins necessary for chloroplast ribosomal assembly and translation, and sigma factors necessary for transcription of chloroplast genes [7]. Plastomes of seed plants typically encode four rRNAs, around 30 tRNAs and up to 80 unique protein-coding genes [6-8]. With the notable exception of extensive photosynthetic gene loss in parasitic plants [9], genic regions are generally conserved across the plastomes of higher plants reported so far; inversions and other rearrangements, however, are frequently reported [5]. In line with the higher conservation of genic versus inter-genic regions, a recent report of plastome from basal asterids indicates the conservation of the repeat patterns in the coding regions, whereas the evolution of the repeats in the non-coding regions is lineage-specific [10]. Due to the endosymbiotic origin of plastomes, several of the genes are coordinately transcribed in operons (e.g. the *psb*B operon) [11,12]. Additionally, chloroplast transcripts undergo RNA editing, especially in ancient plant lineages like ferns and hornworts [13,14].

The *Cardamine* genus represents one of the largest and most polyploid-rich genera of the Brassicaceae, and underwent several recent and rapid speciation events contributing to the divergent evolution of its species [15]. The diversification of *Cardamine* has been driven by multiple events of polyploidization and hybridization, which, together with the high number of species, has till now hindered the obtainment of a comprehensive phylogeny of the genus [16]. Using cpDNA regions, patterns of extensive genetic variation have been previously reported in *Cardamine flexuosa* and related species [17]. The high seed production characterizing several *Cardamine* taxa makes them highly invasive species, which can become noxious in both wild habitats and cultivation. *C. flexuosa* and *C. hirsuta*, for instance, are among the most common weeds in cultivation [17]. *C. impatiens* is rapidly colonizing North America, where it is considered as one of the most aggressive invaders of the understory given its high adaptability to low light conditions [18]. Several *Cardamine* species have been object of growing interest as models for evolutionary adaptive traits and morphological development. *C. hirsuta*, a cosmopolitan weed with fast life cycle, is now a well established model for development of leaf dissection in plants [19]. *C. flexuosa* has been recently used to elucidate the interplay between age and vernalization in regulating flowering [20]. Earlier, in a pioneering study with cross-species microarray hybridization, the whole transcriptome of *C. kokaiensis* provided insights on the molecular bases of cleistogamy and its relationship with environmental conditions, especially chilling temperatures [21].

More recently, using the *Cardamine* genus as a model we demonstrated transcriptome-wide patterns of molecular evolution in genes pertaining to different environmental habitat adaptation by comparative analysis of low altitude, short lived, nemoral species *C. impatiens* to high altitude, perennial, open-habitat dweller *C. resedifolia*, suggesting contrasting patterns of molecular evolution in photosynthetic and cold-tolerance genes [22]. The results explicitly demonstrated faster evolution of the cold-related genes exclusively in the high altitude species *C. resedifolia* [22]. To extend the understanding of positive selection signatures observed in the aforementioned transcriptome-wide analysis to organelles, in this study we carried out the complete sequencing with Solexa technology of the plastome of both species and characterized their gene space and repeat patterns. The comparison of the newly sequenced plastomes between each other and with 15 fully sequenced Brassicaceae plastomes publically available in GenBank uncovered dynamic variation of the IR boundaries in the *Cardamine* lineage associated to generation of lineage-specific pseudogenic fragments in this region. In addition, we could detect signatures of positive selection in ten of the 75 protein-coding genes of the plastomes examined as well as specific *rbcL* residues undergoing intra-peptide co-evolution. Overall our results support the existence of wider signatures of positive selection in the plastome of *C. resedifolia*, possibly as a consequence of adaptation to high altitude environments.

Results and discussion

Genome assembly and validation

In order to further our understanding of selective patterns associated to contrasting environmental adaptation in plants, we obtained and annotated the complete plastome sequence of two congeneric species, high altitude *Cardamine resedifolia* (GenBank accession number KJ136821) and low altitude *C. impatiens* (accession number KJ136822). The primers used amplified an average of 6,2 Kbp, with a minimum and maximum amplicon length of 3,5 and 9,0 Kbp, respectively (Additional file 1: Table S1). In this way, a total of 650335 x100 bp paired-end (PE) reads with a Q30 quality value and mean insert size of 315 bp were obtained for *C. resedifolia*, while 847076 x100 bp PE reads with 325 bp insert size were obtained for *C. impatiens.* Velvet *de-novo* assembly resulted in 36 and 48 scaffolds in *C. resedifolia* and *C. impatiens*, respectively (Table 1). To validate the accuracy of the assembled plastome we carried out Sanger sequencing of PCR amplicons spanning the junction regions (LSC/IR$_A$, LSC/IR$_B$, SSC/ IR$_A$, SSC/IR$_B$). The perfect identity of the sequences to those resulting from assembly confirmed the reliability of assembled plastomes (data not shown). Additionally, we Sanger-sequenced selected regions of the plastome genic space to verify the correct translational frame of the coding regions and to eliminate any Ns still present in the assembly. The finished, high quality organelle genome sequences thus obtained were used for downstream analyses.

Plastome structural features and gene content

The finished plastomes of *C. resedifolia* and *C. impatiens* have a total length of 155036 bp and 155611 bp and a GC content of 36.30% and 36.33%, respectively. These values of GC content suggest an AT-rich plastome organization, which is similar to the other Brassicaceae plastomes sequenced so far (Figures 1 and 2). Quadripartite organization of plastomes, characterized by two large inverted repeats, plays a major role in the recombination and the structural diversity by gene expansion and gene loss in chloroplast genomes [8]. Each plastome assembly displayed a pair of inverted repeats (IR$_A$ and IR$_B$) of 26502 bp and 26476 bp respectively in *C. resedifolia* and *C. impatiens*, demarking large single copy (LSC) regions of 84165 bp and 84711 bp and small single copy (SSC) regions of 17867 bp and 17948 bp in *C. resedifolia* and *C. impatiens* respectively (Table 1, Additional file 2: Table S2). The assembled plastomes contained a total of 85 protein-coding genes, 37 t-RNAs, and 8 r-RNAs in both *C. resedifolia* and *C. impatiens.* We observed a total of 12 protein-coding regions and 6 t-RNAs containing one or more introns (Table 2), which is similar to *Nicotiana tabacum*, *Panax ginseng* and *Salvia miltiorrhiza* [23] but higher than the basal plastomes of the Asterid lineage, where only *ycf*3 and *clpP* have been reported to be protein-coding genes with introns [10]. Of the observed gene space in *C. resedifolia* and *C. impatiens*, 79 protein-coding genes, 30 t-RNA and 4 r-RNAs were found to be unique while 6 protein-coding (*ndhB*, *rpl23*, *rps7*, *rps12*, *ycf2*, *rpl2*), 7 t-RNAs (*trnA-UGC, trnI-CAU, trnI-GAU, trnL-CAA, trnN-GUU, trnR-ACG and trnV-GAC*) and 4 r-RNA genes (*rrn4.5*, *rrn5*, *rrn16*, *rrn23*) were found be duplicated in IR$_A$ and IR$_B$ (Table 2). GC content analysis of the IR, SSC and LSC showed no major fluctuations, with SSC regions accounting for 29.26%/29.16% GC, LSC 34.06%/34.00%, IR$_A$ and IR$_B$ each accounting for 42.36%/42.36% GC in *C. impatiens* and *C. resedifolia*, respectively. Of the observed intron-containing genes, *clpP* and *ycf3* contained two introns. In *rps12* a trans-splicing event was observed with the 5′ end located in the LSC region and the duplicated 3′ end in the IR region as previously reported in *Nicotiana* [24]. In the *trnK-UUU* gene was located the largest intron, harboring the *mat*K gene and accounting for 2552 bp in *C. resedifolia* and 2561 bp in *C. impatiens* (Additional file 3: Table S3).

Table 1 Sequencing statistics and general characteristics of *C. resedifolia* and *C. impatiens* plastome assembly

	C. resedifolia	*C. impatiens*
PE reads with a Q > 30	650335 (315 bp*)	847076 (325 bp*)
Type of Assembler	de-bruijn Graph	de-bruijn Graph
K-mer used	63	63
Number of scaffolds	36	48
Reference species	*Nasturtium officinale*	*Nasturtium officinale*
Assembled plastome size	155036 bp	155611 bp
Number of genes	85(79unique)	85(79unique)
Number of t-RNA	37(30unique)	37(30unique)
Number of r-RNA	8(4unique)	8(4unique)
Length of IRa and IRb	26502 bp	26476 bp
Length of SSC	17867 bp	17948 bp
Length of LSC	84165 bp	84711 bp
Annotation	cpGAVAS, DOGMA	CpGAVAS, DOGMA

*Number in parenthesis indicate the insert size of the PE library.

Pseudogenization events (gene duplication followed by loss of function) have been reported in several plant lineages, e.g., in the plastomes of Anthemideae tribe within the Asteraceae family and *Cocus nucifera*, which belongs to the Arecaceae family [8,25]. Among the genes that underwent pseudogenization there are *ycf68*, *ycf1* and *rps19*, which showed incomplete duplication in the IR$_A$/IR$_B$ and LSC junction regions with loss of function due to accumulation of premature stop codons or truncations. In both *Cardamine* species a partial duplication (106 bp) of the full-length copy of the *rps19* gene (279 bp) located at the IR$_A$/LSC boundary is found in the IR$_B$/LSC region. The fact that only one gene copy is present in the outgroup *N. officinale* indicates that the duplication event leading to *rps19* pseudogenization occurred after the split between *Nasturtium* and *Cardamine*. Sequencing of IR$_B$/LSC regions from additional *Cardamine* species and closely related outgroups will be required to ascertain whether the psedogenization event is genus-specific or not. The conservation of pseudogene length and the close phylogenetic proximity of *Nasturtium* to *Cardamine* [26], however, point to a relatively recent origin of the causal duplication. The basal position of the clade comprising *C. resedifolia* further corroborates the view that the duplication possibly happened early during the radiation of the *Cardamine* genus [15].

Among the coding regions of the sequenced plastomes, the majority of genes have canonical ATG as *bona-fide* start codons. Only 3 genes (*ndhD*, *psbC*, *rps19*) had

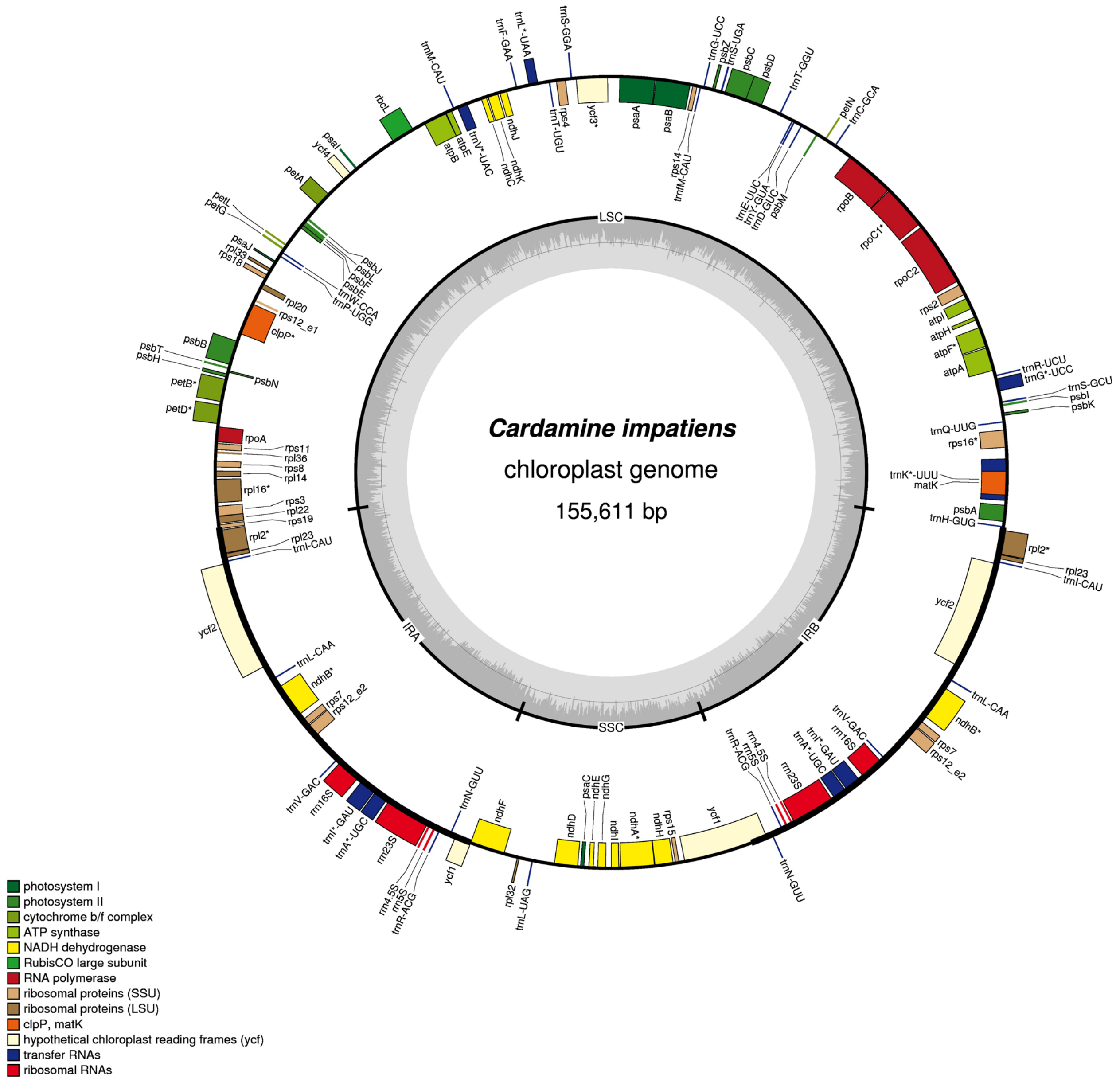


Figure 1 Plastome map of *C. resedifolia*. Genes shown outside of the larger circle are transcribed clockwise, while genes shown inside are transcribed counterclockwise. Thick lines of the smaller circle indicate IRs and the inner circle represents the GC variation across the genic regions.

non-canonical or conflicting starting codon annotations compared to those in the reference plastomes deposited in GenBank, thus requiring manual curation. Previously, RNA editing events of the AUG initiation site to GUG have been reported for *psbC* [27] and *rps19* [8,25]. Analogously (but not observed in our study), RNA editing events contributing to the change of the translational initiation codon to GUG have been reported also in *cemA* [28]. Previous studies on non-canonical translational mechanisms suggest that translational efficiency of GUG codons is relatively high as compared to canonical AUG as initiation codon [29]. It is, therefore, possible that the GTG start codons observed in Brassicaceae *psbC* and *rps19* are required to ensure enhanced translational efficiency for these genes. Also in the case of *ndhD* we identified a *bona fide* non-canonical start codon (ACG), analogously to what observed in other dicotyledonous and monocotyledonous species [8,30,31]. The reported lack of conservation among congeneric *Nicotiana* species [32] and the ability of unedited *ndhD* mRNA to associate to polysomes [33], however, renders the adaptive relevance of this non-canonical start codon in Brassicaceae elusive.

We further analyzed the codon usage frequency and the relative synonymous codon usage frequency (RSCU)

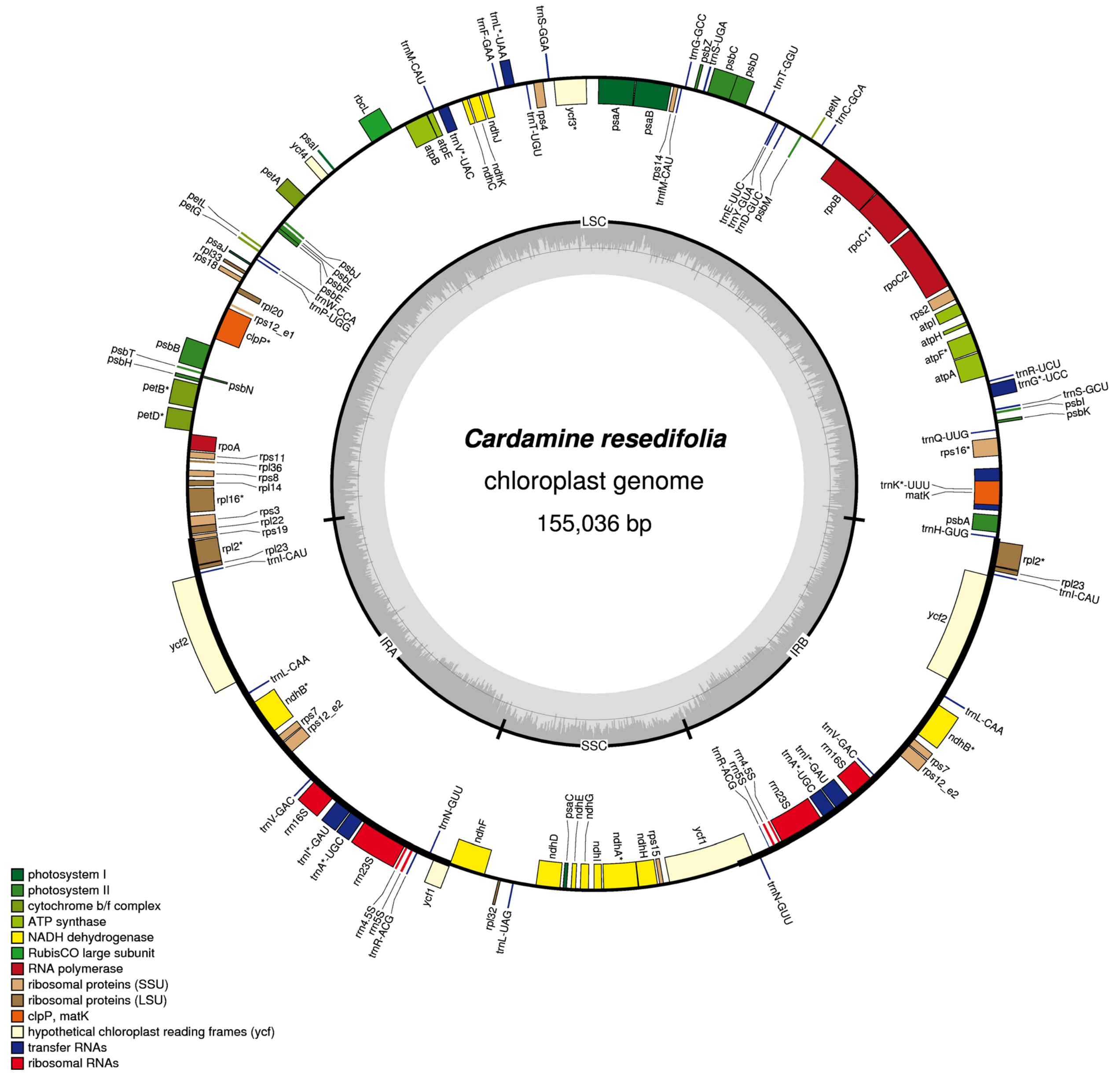


Figure 2 Plastomic map of *C. impatiens*. Genes shown outside of the larger circle are transcribed clockwise, while genes shown inside are transcribed counterclockwise. Thick lines of the smaller circle indicate IRs and the inner circle represents the GC variation across the genic regions.

in the two *Cardamine* plastomes. Mutational bias has been reported as an important force shaping codon usage in both animal and plant nuclear genomes [34,35]. Only few studies addressed the role of mutational bias in plant organelles, and earlier evidence pointed to a comparativley larger effect of natural selection in organellar biased usage of codons [36-38]. More recent studies, however, challenge this view and convincingly show that mutational bias can also be a dominant force in shaping the coding capacity of plant organelles and especially of Poaceace plastomes [39,40]. We, therefore, evaluated Nc plots to estimate the role of mutational bias in shaping the codon usage frequency in *C. resedifolia* and *C. impatiens* and found that most of the genes falls below the expected line of Nc, suggesting a relevant role of mutational bias in *C. resedifolia* and *C. impatiens* (Additional file 4: Figure S1). To provide support for the observed mutational bias, statistical analysis invoking Spearman-rank correlations (ρ) were further implemented between Nc and GC_{3s} and were found to be significant in case of *C. resedifolia* ($\rho = 0.557$, $p < 0.01$) and *C. impatiens* ($\rho = 0.595$, $p < 0.01$). We also evaluated (ρ) between Nc and G_{3s} and positive correlations ($\rho = 0.620$; *C. impatiens*, $\rho = 0.597$, *C. resedifolia*) were observed, which demonstrates

Table 2 List of genes encoded in *C. impatiens* and *C. resedifolia* plastomes

Gene Category	Genes
ribosomal RNAS	§rrn4.5, §rrn5, §rrn16, §rrn23
transfer RNAs	§*trnA-UGC, trnC-GCA, trnD-GUC, trnE-UUC, trnF-GAA, trnfM-CAU, *trnG-UCC, trnG-UCC, trnH-GUG, §trnI-CAU, §*trnI-GAU, *trnK-UUU, §trnL-CAA, *trnL-UAA, trnL-UAG, trnM-CAU, §trnN-GUU, trnP-UGG, trnQ-UUG, §trnR-ACG, trnR-UCU, trnS-GCU, trnS-UGA, trnS-GGA, trnT-UGU, trnT-GGU, *trnV-UAC, §trnV-GAC, trnW-CCA, trnY-GUA
Photosystem I	psaA, psaB, psaC, psaI, psaJ
Photosystem II	psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbI, psbJ, psbK, psbL, psbM, psbN, psbT, psbZ
Cytochrome	petA, *petB, *petD, petG, petL, petN
ATP synthase	atpA, atpB, atpE, *atpF, atpH, atpI
Rubisco	rbcL
NADH dehydrogenase	*ndhA, §*ndhB, ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI, ndhJ, ndhK
Ribosomal protein (large subunit)	§*rpl2, rpl14, *rpl16, rpl20, rpl22, §rpl23, rpl32, rpl33, rpl36
Ribosomal protein (small subunit)	rps2, rps3, rps4, §rps7, rps8, rps11, §*rps12, rps14, rps15, *rps16, rps18, rps19
RNA polymerase	rpoA, rpoB, *rpoC1, rpoC2
ATP-dependent protease	*clpP
Cytochrome c biogenesis	ccsA
Membrane protein	cemA
Maturase	matK
Conserved reading frames	ycf1_short, ycf1_long, §ycf2, *ycf3, ycf4

§Gene completely duplicated in the inverted repeat. *Gene with intron(s).

the role of mutational bias in the biased codon usage frequency in *C. resedifolia* and *C. impatiens*. Taken together, these results indicate that in the two *Cardamine* plastomes sequenced in this study a major role is played by mutational bias, analogously to what suggested in the case of the *Coffea arabica* plastome [41]. Currently we do not have any data on translational efficiency in *Cardamine*, but we cannot exclude it as a possible factor contributing to codon bias in their plastomes as previously suggested in the case of *O. sativa* [42]. Our data, on the other hand, indicate a small fraction of positively selected amino acids (see below), suggesting only marginal contributions of natural selection to codon usage bias in *Cardamine*.

Distribution of repeat content and SSRs analysis

In addition to the larger repeats constituted by IR_A and IR_B, plastid genomes encompass a number of other repeated sequences. We employed REPUTER for the identification of the repeats, which are > 30 bp using a Hamming distance of 90. A total of 49 and 43 repeats were classified in the *C. impatiens* and the *C. resedifolia* plastome (Additional file 5: Table S4), values which are intermediate between those in Poaceae and Arecaceae and the one in Orchidaceae [8]. Among the perfect repeats, we detected four forward repeats, which are located in the LSC (spacer between *trnL* and *trnF*), and two palindromic repeats also localized in the LSC (spacer between *psbT* and *psbN*; Additional file 5: Table S4). Among the imperfect repeats, we annotated a total of 29 forward tandem repeats with a prevalence of them in the spacer between *trnL* and *trnF* and additional 14 palindromic repeats distributed throughout the plastome of *C. impatiens.* In *C. resedifolia*, we observed only two perfect repeats, both palindromic, located in the LSC (spacer between *petN* and *psbM* and spacer between *psbE* and *petL*; Additional file 5: Table S4). All others were imperfect repeats: 15 forward, two reverse and one compound tandem repeats. Interestingly, in *C. resedifolia* we did not observe the large number of repeats found in the *trnL*/*trnF* spacer of *C. impatiens*. As repeat organization and expansion in plastomes may induce recombination and rearrangements (e.g. in Poaceae and Geraniaceae) [8], the *trnL*/*trnF* spacer appears to be a particularly interesting region to reconstruct micro- and macro-evolutionary patterns in *C. impatiens* and closely related species like *C. pectinata* [43].

We further analyzed the distribution of the simple sequence repeats (SSRs), repetitive stretches of 1-6 bp distributed across nuclear and cytoplasmatic genomes, which are prone to mutational errors in replication. Previously, SSRs have been described as a major tool to unravel genome polymorphism across species and for the identification of new species on the basis of the repeat length polymorphism [44]. Since SSRs are prone to slip-strand mispairing, which is demonstrated as a primary source of microsatellite mutational expansion [45], we applied a length threshold greater than 10 bp for mono-, 4 bp for di- and tri- and 3 minimum repetitive units for tetra-, penta- and hexa-nucleotide repeats patterns. We observed a total of 169 SSRs in *C. resedifolia* and 145 SSRs stretches in *C. impatiens* (Additional file 6: Table S5). The observed number of repetitive stretches is in line with the previous results obtained in Brassicaceae [44,46] and other plastomes [23]. Among the observed repeats, the most abundant pattern was found to be stretches of mononucleotides (A/T) accounting for a total of 81 and 61 stretches of polyadenine (polyA) or polythymine (polyT) (A/T) followed by di-nucleotide patterns accounting for a total of 77 and 71 repetitive units in *C. resedifolia* and *C. impatiens*. Interestingly, we observed a higher tendency of longer repeats to occur species-specifically (see e.g. motifs such as AATAG/ATTCT in *C. resedifolia* and AACTAT/AGTTAT in *C. impatiens*; Additional file 6: Table S5), a possible consequence of their rarity [44,46]. Based on the identified SSR stretches, we provide a total of 127 and 114 SSR primer pairs in *C. resedifolia* and in *C. impatiens*, respectively (Additional file 6: Table S5), which can be used for future in-depth studies of phylogeography and population structure in these species.

Synteny conservation and phylogeny of sequenced Brassicaceae plastomes

Among the Brassicaceae species whose plastomes have been fully sequenced so far (a total of 15 at the time of the analyses), only *Nasturtium officinale* and *Barbarea verna* belong to the Cardamineae tribe like *C. impatiens* and *C. resedifolia*. As *Nasturtium* has been indicated as putative sister genus to *Cardamine* [26], the plastome of *N. officinale* was used as reference to calculate average nucleotide identity (ANI) plots using a window size of 1000 bp, step size of 200 bp and a alignment length of 700 bp, 70% identity. As expected by their close relatedness, a high degree of synteny conservation with the reference plastome was observed (Additional file 7: Figure S2). Average nucleotide identity value based on 748 and 568 fragments using one-way and two-way ANI indicated a similarity of 97.76% (SD 2.25%) and 97.55% (SD 2.17%) between *C. resedifolia* and *N. officinale*. Similarly, one-way and two-way ANI values of 98.19% (SD 1.88%) and 98.03% (SD 1.78%) based on 759 fragments and 603 fragments were observed in case of *C. impatiens* and *N. officinale*. Syntenic analysis of the coding regions across Brassicaceae and one outgroup belonging to the Caricaceae family (*Carica papaya*) revealed perfect conservation of gene order along the plastome of the analyzed species (Figure 3). Similarity among plastomes was a function of plastome organization and gene content, with IR and coding regions of fundamental genes being the most highly conserved, as indicated by analysis of pairwise mVISTA plots using *C. impatiens* as reference (Additional file 8: Figure S3).

To precisely determine the phylogenetic position and distance of *C. resedifolia* and *C. impatiens* with respect to the other Brassicaceae with fully sequenced plastome, we performed a concatenated codon-based sequence alignment of the 75 protein coding genes, representing a total of 67698 nucleotide positions. The GTR + I + G model resulted the best fitting model for the matrix according to the JModelTest program using the Akakie information criterion (AIC) and Bayesian information criterion (BIC). Phylogenetic reconstruction was carried out using maximum parsimony (MP), Maximum likelihood (ML) and Bayesian inference (BI). MP analysis resulted in a tree length of 15739, a consistency index of 0.819 and retention index of 0.646. ML analysis revealed a phylogenetic tree with the -lnL of 186099.2 using the GTR + I + G model as estimated using JModelTest. For MP and ML analysis, 1000 bootstrap replicates were evaluated and all the trees obtained were rooted using *Carica papaya* as an outgroup (Figure 4). All phylogenetic methods provided consistent topologies, indicating good reproducibility of the recovered phylogeny. The tree positioning of *Lepidium virginicum*, which lacked resolution in the MP tree, constituted the only exception. As expected, the four taxa from the Cardamineae tribe (genera *Cardamine*, *Nasturtium* and *Barbarea*) formed a well-supported, monophyletic clade with *B. verna* as most basal species. Our phylogenetic reconstruction is in agreement with previous reports on the relationships among Brassicaeacea tribes [47,48], thus indicating that it can be used as a reliable framework for assessment of protein coding gene evolution in the Brassicaeae family in general and *Cardamine* species in particular.

Molecular evolution of Brassicaceae plastomes

Understanding the patterns of divergence and adaptation among the members of specific phylogenetic clades can offer important clues about the forces driving its evolution [49,50]. To pinpoint whether any genes underwent adaptive evolution in Brassicaceae plastomes in general and in the *Cardamine* genus in particular, we carried out the identification of genes putatively under positive selection using Selecton. At the family level, we observed signatures of positive selection in 10 genes (*ycf1, rbcL, rpoC2, rpl14, matK, petD, ndhF, ccsA*, *accD*, and *rpl20*) at a significance level of 0.01 (Table 3). Two of these genes, namely *ycf1* and *accD,* have been reported to undergo fast evolution in other plant lineages as well. *ycf1* is one of the largest plastid genes and it has been classified as the most divergent one in plastomes of tracheophytes [5]. Despite it has been reported to be essential in tobacco [51], it has been lost from various angiosperm groups [52]. Recently, ycf1 was identified as one of the core proteins of the chloroplast inner envelope membrane protein translocon forming a complex (called TIC) with Tic100, Tic56, and Tic20-I [53]. None of the 24 amino acids putatively under positive selection in Brassicaceae are located in predicted transmembrane

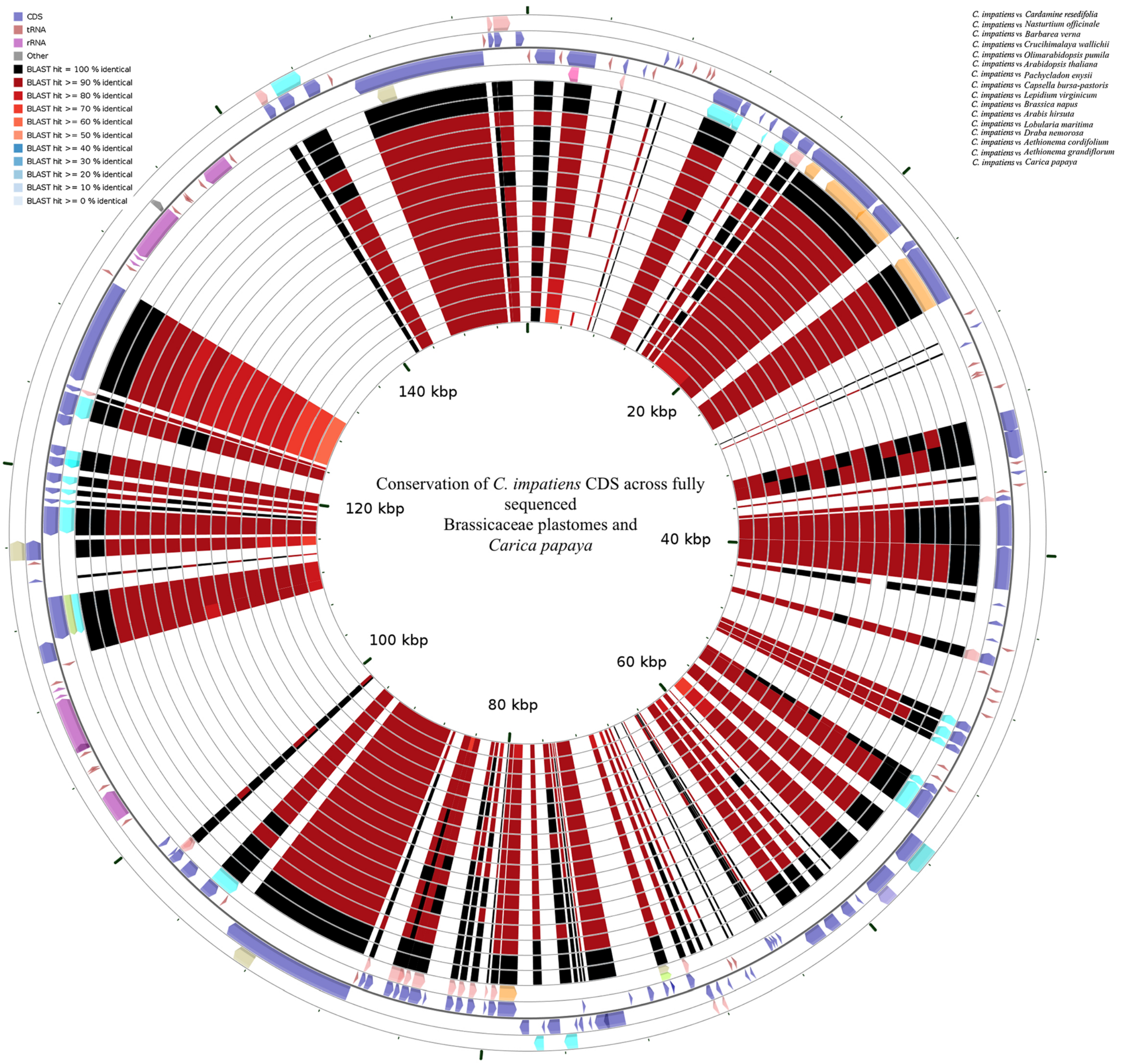


Figure 3 Circular map displaying the conservation of the coding regions across the Brassicacae, the *Cardamine* plastomes sequenced in this study and the outgroup *Carica papaya*.

domains [53], indicating that in Brassicaceae evolution of predicted channel-forming residues is functionally constrained. Analogously to what found for Brassicaceae in our study, in the asterid lineage recent studies also show accelerated rates of evolution in *accD*, a plastid-encoded beta-carboxyl transferase subunit of acetyl-CoA carboxylase (ACCase) [54], which has been functionally re-located to nucleus in the Campanulaceae [55]. As in none of the fully sequenced Brassicaceae re-location of plastidial *accD* to the nuclear genome has been observed, it is likely that the fast evolution of this gene is independent from the genome from which it is expressed. On the other hand, *accD* has been demonstrated to be essential for proper chloroplast and leaf development [54]. Plastidial *accD* together with three nucleus-encoded subunits form the ACCase complex, which been reported to produce the large majority of malonyl CoA required for *de novo* synthesis of fatty acids [56,57] under the regulatory control of the PII protein [58]. Most importantly, there are direct evidences that accD can affect plant fitness and leaf longevity [59]. The signatures of positive selection observed in both Brassicaceae (our study) and asterids [55], therefore, indicate that this gene may have been repeatedly involved in the adaptation to specific ecological niches during the radiation of dicotyledonous plants.

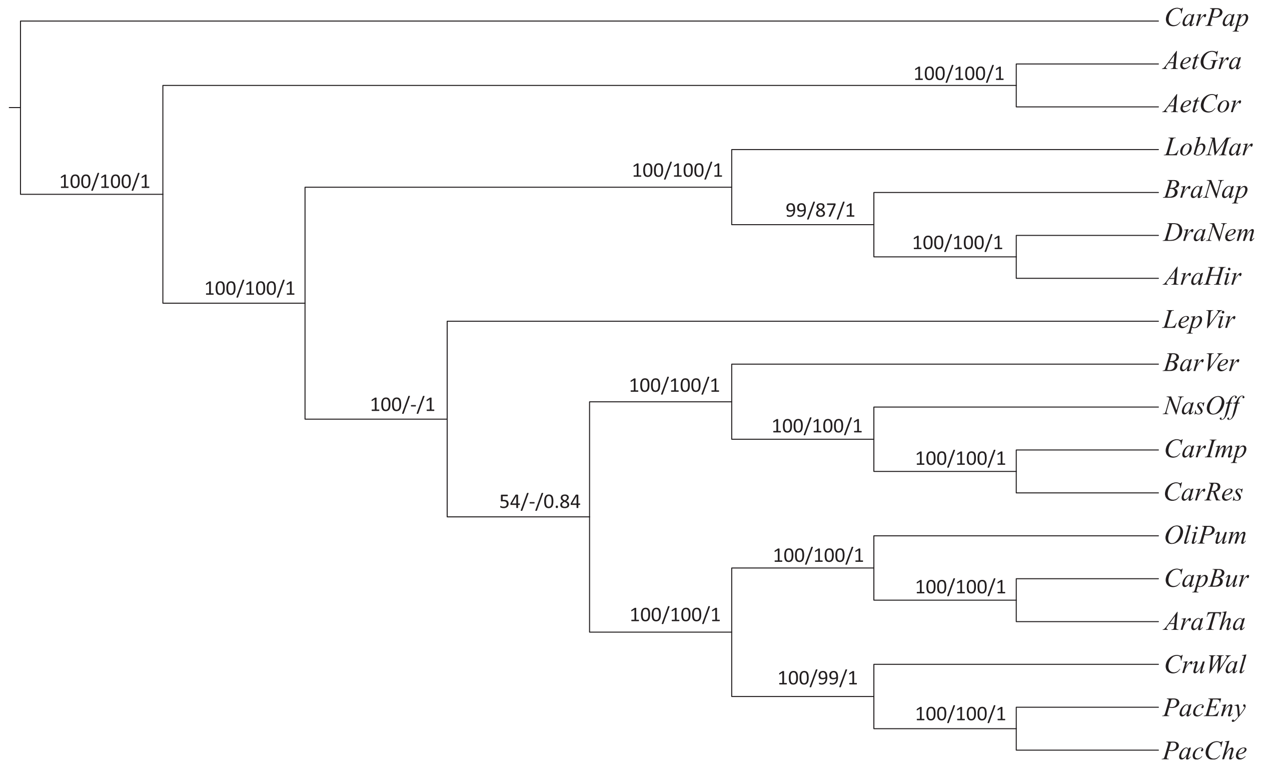


Figure 4 Cladogram of the phylogenetic relationships among Brassicaceae species with fully sequenced plastome used in this study. The cladogram represents the consensus topology of the maximum likelihood (ML), maximum parsimony (MP) and bayesian inference (BI) phylogenetic reconstructions using the concatenated alignment of 75 protein coding genes. Numbers on branches indicate ML/MP/BI support values (bootstrap proportion > 50%). Dashes indicate lack of statistical support. Abbreviation of species names can be found in Additional file 10: Table S7. Phylogenetic tree visualization was done using FigTree.

Given the prominent role that plastid proteins play in the constitution of cores of photosynthetic complexes [60], one could expect that some photosynthetic genes would also be targeted by positive selection. Previous analyses in leptosporangiates, for instance, uncovered a burst of putatively adaptive changes in the *psb*A gene, which is coding for a core subunit of Photosystem II (PSII). Extensive residue co-evolution along with positive Darwinian selection was also detected [61]. However, we did not observe such burst of high rate of evolution in Brassicaceae *psbA*. We instead observed co-evolving residues along with positive signatures of Darwinian selection in *rbcL* (*ribulose-1, 5-bisphosphate carboxylase/ oxygenase*), which codes for RUBISCO, the enzyme

Table 3 Positive selection sites identified with selecton with d.f. =1

Gene	Null	Positive	Putative sites under positive selection *
ycf1	-21668,5	-21647,6	24(343 P, 424 A, 533 D, 565 H, 970 L, 1293 L, 1313 N, 1399 R, 1400 N, 1414 R, 459 W, 564 I, 738 K, 922 F, 928 L, 1081 F, 1113 T, 1235 K, 1259 P, 1343 R, 1428 F, 1475 S, 1477 R, 1533 Y)
rbcL	-3000,07	-2984,64	3(326 V, 472 V, 477 A)
rpoC2	-11431,8	-11423,5	7(490 F, 527 L, 540 P, 541 H, 981 A, 998 L, 1375 Y)
rpl14	-631,147	-623,836	2(18 K, 33 K)
matK	-5014,38	-5007,21	1(51 V)
petD	-1052,21	-1045,47	2(138 V, 139 V)
ndhF	-6497,59	-6491,61	4(65 I, 509 F, 594 Q, 734 M)
ccsA	-3031,79	-3026,12	5(97 H, 100 H, 176 L, 182 E, 184 F)
accD	-4142,84	-4137,43	3(112 F, 167 H, 485 E)
rpl20	-834,791	-831,556	2(80 R, 117 E)

*lower bound > 1.
"Null" and "Positive" columns list likelihood values obtained under the models M8a (null model) and M8 (positive selection), respectively.

catalyzing photosynthetic assimilation of CO_2 and one of the major rate-limiting steps in this process. Positive rates of selection were observed at three sites across Brassicaceae. The observed rates of positive selection on neutral hydrophobic residues such A (alanine) and V (valine) are consistent with previous estimates of selection sites across land plants [62]. As compared to RUBISCO adaptive selection in gymnosperms, where previous reports suggest 7 sites under positive selection (A11V, Q14K, K30Q, S95N, V99A, I133L, and L225I) [63], the low frequency of the sites under positive selection observed in Brassicaceae, which belongs to Angiosperms, could be a consequence of the more recent origin of the latter group. The fact that the long series of geological variations of atmospheric CO_2 concentrations experienced by gymnosperms seem to parallel adaptive bursts of co-evolution between RUBISCO and RUBISCO activase lend support to this view [63]. Recent studies across Amaranthaceae *sensu lato* identified multiple parallel replacements in both monocotyledonous and dicotyledonous C_4 species at two residues (281 and 309), suggesting their association with selective advantages in terms of faster and less specific enzymatic activity (e.g. in C_4 taxa or C_3 species from cold habitats) [64]. We found no evidence of selection in these or other residues in their proximity in the crystal structure of RUBISCO, indicating that in the Brassicaceae species analyzed (including high altitude *C. resedifolia*) this kind of adaptation possibly did not occur. The three residues under positive selection in our study belong to RUBISCO loop 6 (amino acid 326 V) and C-terminus (amino acids 472 V and 477 A). None of these aminoacids belong to the set of highly conserved residues identified among RUBISCO and RUBISCO-like proteins, which are likely under strong purifying selection [65,66]. This result is in agreement with the observation that in monocotyledons adaptive mutations preferentially affect residues not directly involved in catalysis, but either aminoacids in proximity of the active site or at the interface between RUBISCO subunits [67]. The C-terminus of RUBISCO is involved in interactions between large subunits (intra-dimer) and with RUBISCO activase, and amino acid 472 was previously identified among rbcL residues evolving under positive selection [64]. It is, therefore, possible that the mutation in residues 472 and 477 could contribute to modulate the aggregation and/or activation state of the enzyme in Brassicaceae. Also amino acid 326 has consistently been identified as positively selected in different studies, although in relatively few plant groups [64]. This residue is in close proximity to the fourth among the most often positively selected RUBISCO residues in plants (amino acid 328), which has been associated to adaptive variation of RUBISCO active site possibly by modifying the position of H327, the residue coordinating the P5 phosphate of ribulose-1,5-bisphosphate [64,67]. Such "second shell mutations" in algae and cyanobacteria are known to be able to modulate RUBISCO catalytic parameters [68], and were recently shown to be implicated in the transition from C_3 to C_4 photosynthesis in monocotyledons by enhancing conformational flexibility of the open-closed transition [67]. Taken together, these data indicate that in Brassicaceae residue 326 could affect RUBISCO discrimination between CO_2 and O_2 fixation, analogously to what suggested for residue 328 in several other plant groups.

The other genes displaying signature of positive selection in our study belong to 4 main functional classes: transcription and transcript processing (*rpoC2*, *matK*), translation (*rps14* and *rpl20*), photosynthetic electron transport and oxidoreduction (*pet*D, *ndh*F), cytochrome biosynthesis (*ccsA*). The broad spectrum of candidate gene functional classes affected indicate that natural selection target different chloroplast functions, supporting the possible involvement of plastid genes in adaptation and speciation processes in the Brassicaceae family [69].

To obtain a more precise picture of the phylogenetic branch(es), where the putatively adaptive changes took place, the rate of substitution mapping on each individual branch was estimated by the MapNH algorithm [70]. Focusing on the Cardamineae tribe and using a branch length threshold to avoid bias towards shorther branches, we found that genes under positive selection in the *Cardamine* lineage (*accD*, *ccsA*, *matK*, *ndhF*, *rpoC2*) evolved faster in *C. resedifolia* as compared to *C. impatiens*, suggesting that adaptive changes may have occurred more frequently in response to the highly selective conditions of high altitude habitats (Additional file 9: Table S6). These results are in line with the accelerated evolutionary rates of cold-related genes observed for *C. resedifolia* in the transcriptome-wide comparison of its transcriptome to that of *C. impatiens* [22]. Given the different genomic inheritance and low number of genes encoded in the chloroplast, it is unfortunately difficult to directly compare the evolutionary patterns observed for photosynthetic plastid genes in this study with the strong purifying selection identified for nuclear-encoded photosynthetic genes of *C. resedifolia* [22]. It is, however, worth of note that the genes with larger differences in evolutionary rates between *C. resedifolia* and *C. impatiens* are not related to photosynthetic light reactions, suggesting that this function is likely under intense purifying selection also for plastidial subunits in *Cardamine* species (Additional file 9: Table S6). Given the relatively few studies available and the complex interplay among the many factors potentially affecting elevational adaptation in plants [71,72], however, additional studies will be needed to specifically address this point.

Conclusion

In conclusion, the comparative analysis of the *de-novo* sequences of *Cardamine* plastomes obtained in our study identified family-wide molecular signatures of positive selection along with mutationally biased codon usage frequency in Brassicaceae chloroplast genomes. We additionally found evidence that the plastid genes of *C. resedifolia* experienced more intense positive selection than those of the low altitude *C. impatiens*, possibly as a consequence of adaptation to high altitude environments. Taken together, these results provide a series of candidate plastid genes to be functionally tested for elucidating the driving forces underlying adaptation and evolution in this important plant family.

Methods

Illumina sequencing, plastome assembly, comparative plastomics and plastome repeats

Genomic DNA was extracted from young leaves of *Cardamine impatiens* and *C. resedifolia* using the DNeasy Plant Mini kit (Qiagen GmbH, Hilden, Germany) and Long PCR amplification with a set of 22 primer pairs was carried out using Advantage 2 polymerase mix (Clontech Laboratories Inc., Mountain View, CA, USA) according to manufacturer's instructions. We chose to use a long-PCR whole plastome amplification approach to maximize the number of reads to be used for assembly. The primer pairs used are listed in Additional file 1: Table S1. Amplicons from each species were pooled in equimolar ratio, sheared with Covaris S220 (Covaris Inc., Woburn, MA, USA) to the average size of 400 bp and used for illumina sequencing library preparation. Each library was constructed with TruSeq DNA sample preparation kits V2 for paired-end sequencing (Illumina Inc., San Diego, CA) and sequenced on a HiSeq 2000 at The Genome Analysis Centre (Norwich, UK). Subsequently, the reads were quality filtered using a Q30 quality value cutoff using FASTX_Toolkit available from http://hannonlab.cshl.edu/fastx_toolkit/. After subsequent quality mapping on the Brassicaceae plastomes, contaminating reads were filtered off. Specifically, raw reads were mapped on the publicly available Brassicaceae plastomes (Additional file 10: Table S7) using the Burrows-Wheeler Aligner (BWA) programusing -n 2, -k 5 and -t 10. SAM and BAM files obtained as a result were consecutively filtered for the properly paired end (PE) reads using SAMtools [73].

To obtain the *de novo* plastome assembly, properly PE reads were assembled using Velvet assembler [74]. In Velvet, N50 and coverage were evaluated for all *K-mers* ranging from 37 to 73 in increments of 4. Finally, the plastome assembly with *K-mer* = 65 was used for all subsequent analyses in both species. The selected Velvet assembly was further scaffolded using optical read mapping as implemented in Opera [75]. Assembled scaffolds were further error corrected using the SEQUEL software by re-mapping the reads and extending/correcting the ends of the scaffolded regions [76]. Gap filling was performed using the GapFiller program with parameters –m 80 and 10 rounds of iterative gap filling [77]. All the given computational analysis was performed on a server equipped with 128 cores and a total of 512 GB.

Following scaffolding and gap filling, *C. resedifolia* and *C. impatiens* scaffolds were systematically contiguated based on the *Nasturtium officinale* plastome (AP009376.1, 155,105 bp) using the nucmer and show-tiling programs of the MUMmer package [78]. Finally, mummer plot from the same package was used to evaluate the syntenic plots and the organization of the inverted repeats by pairwise comparison between the *N. officinale* and *C. resedifolia* and *C. impatiens* plastomes. Due to assembler's insufficient accuracy in assembly of repeat regions, manual curations of the IRs were carried out using the BLAST2Seq program by comparison of the scaffolded regions with the *N. officinale* plastome. To test assembly quality and coverage, average nucleotide identity plots were calculated. Additionally, the junctions of the IRs and all remaining regions containing Ns were amplified by PCR using the primers listed in Additional file 1: Table S1 and Sanger sequenced. The finished *C. resedifolia* and *C. impatiens* chloroplast sequences have been deposited to GenBank with accession numbers KJ136822 and KJ136821, respectively.

To assess the levels of plastid syntenic conservation, the assembled plastomes of *C. resedifolia* and *C. impatiens* were compared to all publicly available plastomes of Brassicaceae using CGview by computing pairwise similarity [79]. Additionally, mVISTA plots were constructed using the annotated features of *C. resedifolia* and *C. impatiens* plastomes with a rank probability of 0.7 (70% alignment conservation) to estimate genome-wide conservation profiles [80]. To identify the stretches of the repetitive units, the REPUTER program was used with parameters -f –p –r –c –l 30 –h 3 –s and the repeat patterns along with the corresponding genomic co-ordinates were tabulated [81]. Additionally, we mined the distribution of perfect and compound simple sequence repeats using MISA (http://pgrc.ipk-gatersleben.de/misa/). In our analysis, we defined a minimum repetitive stretch of 10 nucleotides as mono-nucleotide, a consecutive stretch of 4 repeats units to be classified as di- and tri-nucleotide, and a stretch of 3 repeat units for each tetra-, penta- and hexanucleotide stretches as simple sequence repeats (SSRs).

Chloroplast genome annotation and codon usage estimation

The assembled plastome of *C. resedifolia* and *C. impatiens* was annotated using cpGAVAS [82] and DOGMA (Dual Organellar GenoMe Annotator) [83]. Manual curation of start and stop codons was carried out using the 20 available reference Brassicaceae plastomes. The predicted coding regions were manually inspected and were re-sequenced with

Sanger chemistry whenever large differences in conceptually translated protein sequences were detected compared with the reference plastome of *N. officiale* (Additional file 10: Table S7). GenomeVx [84] was used for visualization of plastome maps. Transfer-RNAs (t-RNAs) were identified using the t-RNAscan-SE software using the plastid genetic code and the covariance models of RNA secondary structure as implemented in cove algorithm [85]. Only coding regions longer than 300 bp from *Cardamine* and the other Brassicaceae plastomes were used for estimation of codon usage in CodonW with translational table = 11 (available from codonw.sourceforge.net). We further tabulated additional codon usage measures such as Nc (effective number of codons), GC_{3s} (frequency of the GC at third synonymous position). GC, GC_1, GC_2 and GC_3 were calculated with in-house Perl scripts. Estimation of the standard effective number of codon (Nc) was tabulated using the equation N(c) = 2 + s + 29/(s(2) + (1-s)(2)), where s denotes GC_{3s} [86].

Molecular evolution in *Cardamine* plastomes

For evaluating the patterns of molecular evolution, codon alignment of the coding regions was created using MACSE, which allows the identification of frameshift events [87]. Model selection was performed using the JmodelTest 2 [88]. Phylogenetic reconstruction was performed using PhyML with 1000 bootstrap replicates [89]. To identify the role of selection on the evolution of plastid genes, MACSE codon alignments were analysed using Selecton [90] allowing for two models: M8 (model of positive selection) and M8a (null model) and likelihood scores were compared for each gene set followed by a chi-square test with 1 degree of freedom. Only tests with probability lower than 0.01 were considered significant and were classified as genes under positive selection. We further mapped the substitution rate on the phylogeny of the Brassicaceae species using MapNH [70] with a threshold of 10 to provide a reliable estimation of the braches under selection.

Availability of supporting data

The data set supporting the results of this article are available in the GenBank repository, *Cardamine resedifolia* plastome (GenBank accession number KJ136821) and *C. impatiens* (accession number KJ136822). The phylogenetic matrix and trees are available from Treebase (http://purl.org/phylo/treebase/phylows/study/TB2:S17255).

Additional files

Additional file 1: Table S1. Long-range PCR primers used for tiled whole-plastome amplification.

Additional file 2: Table S2. Summary of distribution and localization of genes in the *C. resedifolia* and *C. impatiens* plastomes.

Additional file 3: Table S3. Genes with introns in *C. resedifolia* (^{a}) and *C. impatiens* (^{b}) plastome and length of exons and introns.

Additional file 4: Figure S1. Nc plot showing the distribution of the genes >300 bp in *C. resedifolia* and *C. impatiens*. The black line in the curve represents the standard effective number of codons (Nc) calculated using the equation N(c) = 2 + s + 29/(s(2) + (1-s)(2)), where s denotes GC_{3s} (Wright [86]).

Additional file 5: Table S4. Distribution and localization of repeat sequences in cpDNA of *C. impatiens and C. resedifolia.*

Additional file 6: Table S5. Cumulative SSR frequency and corresponding primer pairs in *C. resedifolia* and *C. impatiens*.

Additional file 7: Figure S2. Average nucleotide identity plots of the *C. resedifolia* and *C. impatiens* against *Nasturtium officinale*.

Additional file 8: Figure S3. mVISTA plots showing genome-wise similarity between *C. resedifolia*, *C. impatiens* and *N. officinale* with rank probability of 70% and window size of 100 bp. The annotations displayed are derived from the *C. impatiens* plastome.

Additional file 9: Table S6. Phylogenetic distribution map of substitution rates using probabilistic substitution mapping under the homogenous model of sequence evolution.

Additional file 10: Table S7. Accessions and references for fully sequenced plastomes used in phylogenetic reconstruction and genome comparison in this study.

Abbreviations

RUBISCO: Ribulose-1, 5-bisphosphate carboxylase/oxygenase; IR: Inverted repeat region; LSC: Large single copy region; SSC: Small single copy region; Bp: Base pair; Nc: Effective number of codons used in a gene; GC: Guanine-cytosine; SSR: Simple sequence repeat; ANI: Average nucleotide identity.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

BW, DQ and EB helped to carry out lab work and draft the manuscript. ML contributed to conceive and design of the study, carried out all the phases of lab work, helped to draft the manuscript. GS carried out data analyses, drafted the manuscript. HS carried out data analyses and helped to draft the manuscript. RV helped to draft the manuscript. CV conceived, designed and coordinated the study, finalized the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by: the Autonomous Province of Trento (Italy) through core funding of the Ecogenomics group (EB, GS, ML, RV and CV) and the ACE-SAP project (regulation number 23, 12 June 2008, of the Servizio Universita' e Ricerca Scientifica); the China Scholarship Council (BW, DQ, HS).

Author details

[1]Ecogenomics Laboratory, Department of Biodiversity and Molecular Ecology, Research and Innovation Centre, Fondazione Edmund Mach, Via E. Mach 1, 38010 S Michele all'Adige (TN), Italy. [2]College of Horticulture, Northwest Agricultural and Forest University, 712100 Yangling, Shaanxi, PR China.

Received: 22 November 2014 Accepted: 27 March 2015
Published online: 17 April 2015

References

1. Wu J, Liu B, Cheng F, Ramchiary N, Choi SR, Lim YP, et al. Sequencing of chloroplast genome using whole cellular DNA and Solexa sequencing technology. Front Plant Sci. 2012;3:243.
2. Waters DLE, Nock CJ, Ishikawa R, Rice N, Henry RJ. Chloroplast genome sequence confirms distinctness of Australian and Asian wild rice. Ecol Evol. 2012;2:211–7.
3. Plant C, Group W. A DNA barcode for land plants. Proc Natl Acad Sci U S A. 2009;106:12794–7.
4. Sugiura M. The chloroplast genome. Plant Mol Biol. 1992;19:149–68.

5. Kim K-J, Lee H-L. Complete chloroplast genome sequences from Korean ginseng (*Panax schinseng* Nees) and comparative analysis of sequence evolution among 17 vascular plants. DNA Res. 2004;11:247–61.
6. Yang M, Zhang X, Liu G, Yin Y, Chen K, Yun Q, et al. The complete chloroplast genome sequence of date palm (*Phoenix dactylifera* L.). PLoS One. 2010;5:e12762.
7. Green BR. Chloroplast genomes of photosynthetic eukaryotes. Plant J. 2011;66:34–44.
8. Huang Y-Y, Matzke AJM, Matzke M. Complete sequence and comparative analysis of the chloroplast genome of coconut palm (*Cocos nucifera*). PLoS One. 2013;8:e74736.
9. Braukmann T, Kuzmina M, Stefanović S. Plastid genome evolution across the genus *Cuscuta* (Convolvulaceae): two clades within subgenus *Grammica* exhibit extensive gene loss. J Exp Bot. 2013;64:977–89.
10. Ku C, Hu JM, Kuo CH. Complete plastid genome sequence of the basal asterid *Ardisia polysticta* Miq. and comparative analyses of asterid plastid genomes. PLoS One. 2013;8:e62548.
11. Westhoff P, Herrmann RG. Complex RNA maturation in chloroplasts: the *psbB* operon from spinach. Eur J Biochem. 1988;171:551–64.
12. Barkan A. Expression of plastid genes: organelle-specific elaborations on a prokaryotic scaffold. Plant Physiol. 2011;155:1520–32.
13. Kugita M. RNA editing in hornwort chloroplasts makes more than half the genes functional. Nucleic Acids Res. 2003;31:2417–23.
14. Wolf PG, Hasebe M, Rowe CA. High levels of RNA editing in a vascular plant chloroplast genome: analysis of transcripts from the fern *Adiantum capillus-veneris*. Gene. 2004;339:89–97.
15. Carlsen T, Bleeker W, Hurka H, Elven R, Brochmann C. Biogeography and phylogeny of *Cardamine* (Brassicaceae). Ann Missouri Bot Gard. 2009;96:215–36.
16. Marhold K, Lihová J. Polyploidy, hybridization and reticulate evolution: lessons from the Brassicaceae. Plant Syst Evol. 2006;259:143–74.
17. Lihová J, Marhold K. Worldwide phylogeny and biogeography of *Cardamine flexuosa* (Brassicaceae) and its relatives. Am J Bot. 2006;93:1206–21.
18. Huffman KM. Investigation into the potential invasiveness of the exotic narrow-leaved bittercress, (*Cardamine impatiens* L.), Brassicaceae. Master's Thesis. Virginia Polytechnic Institute and State University, Biological Sciences Department. 2008.
19. Canales C, Barkoulas M, Galinha C, Tsiantis M. Weeds of change: *Cardamine hirsuta* as a new model system for studying dissected leaf development. J Plant Res. 2010;123:25–33.
20. Zhou C-M, Zhang T-Q, Wang X, Yu S, Lian H, Tang H, et al. Molecular basis of age-dependent vernalization in *Cardamine flexuosa*. Science. 2013;340:1097–100.
21. Morinaga SI, Nagano AJ, Miyazaki S, Kubo M, Demura T, Fukuda H, et al. Ecogenomics of cleistogamous and chasmogamous flowering: genome-wide gene expression patterns from cross-species microarray analysis in *Cardamine kokaiensis* (Brassicaceae). J Ecol. 2008;96:1086–97.
22. Ometto L, Li M, Bresadola L, Varotto C. Rates of evolution in stress-related genes are associated with habitat preference in two *Cardamine* lineages. BMC Evol Biol. 2012;12:7.
23. Qian J, Song J, Gao H, Zhu Y, Xu J, Pang X, et al. The complete chloroplast genome sequence of the medicinal plant Salvia miltiorrhiza. PLoS One. 2013;8:e57607.
24. Hildebrand M, Hallick RB, Passavant CW, Bourque DP. Trans-splicing in chloroplasts: the rps 12 loci of *Nicotiana tabacum*. Proc Natl Acad Sci U S A. 1988;85:372–6.
25. Liu Y, Huo N, Dong L, Wang Y, Zhang S, Young HA, et al. Complete Chloroplast genome sequences of Mongolia medicine *Artemisia frigida* and phylogenetic relationships with other plants. PLoS One. 2013;8:e57533.
26. Sweeney PW, Price RA. Polyphyly of the genus *Dentaria* (Brassicaceae): evidence from *trnL* intron and *ndhF* sequence data. Syst Bot. 2000;25:468–78.
27. Kuroda H, Suzuki H, Kusumegi T, Hirose T, Yukawa Y, Sugiura M. Translation of *psbC* mRNAs starts from the downstream GUG, not the upstream AUG, and requires the extended Shine-Dalgarno sequence in tobacco chloroplasts. Plant Cell Physiol. 2007;48:1374–8.
28. Moore MJ, Bell CD, Soltis PS, Soltis DE. Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. Proc Natl Acad Sci U S A. 2007;104:19363–8.
29. Rohde W, Gramstat A, Schmitz J, Tacke E, Prüfer D. Plant viruses as model systems for the study of non-canonical translation mechanisms in higher plants. J Gen Virol. 1994;75:2141–9.
30. Neckermann K, Zeltz P, Igloi GL, Kössel H, Maier RM. The role of RNA editing in conservation of start codons in chloroplast genomes. Gene. 1994;146:177–82.
31. Hirose T, Sugiura M. Both RNA editing and RNA cleavage are required for translation of tobacco chloroplast *ndhD* mRNA: a possible regulatory mechanism for the expression of a chloroplast operon consisting of functionally unrelated genes. EMBO J. 1997;16:6804–11.
32. Sasaki T, Yukawa Y, Miyamoto T, Obokata J, Sugiura M. Identification of RNA editing sites in chloroplast transcripts from the maternal and paternal progenitors of tobacco (*Nicotiana tabacum*): comparative analysis shows the involvement of distinct trans-factors for *ndhB* editing. Mol Biol Evol. 2003;20:1028–35.
33. Zandueta-Criado A, Bock R. Surprising features of plastid *ndhD* transcripts: addition of non-encoded nucleotides and polysome association of mRNAs with an unedited start codon. Nucleic Acids Res. 2004;32:542–50.
34. Kawabe A, Miyashita NT. Patterns of codon usage bias in three dicot and four monocot plant species. Genes Genet Syst. 2003;78:343–52.
35. Plotkin JB, Kudla G. Synonymous but not the same: the causes and consequences of codon bias. Nat Rev Genet. 2011;12:32–42.
36. Liu Q, Feng Y, Xue Q. Analysis of factors shaping codon usage in the mitochondrion genome of *Oryza sativa*. Mitochondrion. 2004;4:313–20.
37. Liu Q, Xue Q. Comparative studies on codon usage pattern of chloroplasts and their host nuclear genes in four plant species. J Genet. 2005;84:55–62.
38. Zhang W, Zhou J, Li Z, Wang L, Gu X, Zhong Y. Comparative analysis of codon usage patterns among mitochondrion, chloroplast and nuclear genes in *Triticum aestivum* L. Acta Botanica Sinica. 2007;49:246–54.
39. Sablok G, Nayak KC, Vazquez F, Tatarinova TV. Synonymous codon usage, GC_3, and evolutionary patterns across plastomes of three pooid model species: emerging grass genome models for monocots. Mol Biotechnol. 2011;49:116–28.
40. Zhou M, Li X. Analysis of synonymous codon usage patterns in different plant mitochondrial genomes. Mol Biol Rep. 2009;36:2039–46.
41. Nair RR, Nandhini MB, Monalisha E, Murugan K, Nagarajan S, Surya N, et al. Synonymous codon usage in chloroplast genome of *Coffea arabica*. Bioinformation. 2012;8:1096–104.
42. Morton BR, So BG. Codon usage in plastid genes is correlated with context, position within the gene, and amino acid content. J Mol Evol. 2000;50:184–93.
43. Kučera J, Lihová J, Marhold K. Taxonomy and phylogeography of *Cardamine impatiens* and *C. pectinata* (Brassicaceae). Bot J Linn Soc. 2006;152:169–95.
44. Sablok G, Mudunuri SB, Patnana S, Popova M, Fares MA, La Porta N. Chloromitossrdb: open source repository of perfect and imperfect repeats in organelle genomes for evolutionary genomics. DNA Res. 2013;20:127–33.
45. Schlötterer C, Harr B. Microsatellite instability. Encycl life Sci. 2001:1–4.
46. Gandhi SG, Awasthi P, Bedi YS. Analysis of SSR dynamics in chloroplast genomes of Brassicaceae family. Bioinformation. 2010;5:1–5.
47. Couvreur TLP, Franzke A, Al-shehbaz IA, Bakker FT, Koch A, Mummenhoff K. Molecular phylogenetics, temporal diversification, and principles of evolution in the mustard family (Brassicaceae). Mol Biol Evol. 2010;27:55–71.
48. Franzke A, Lysak MA, Al-Shehbaz IA, Koch MA, Mummenhoff K. Cabbage family affairs: the evolutionary history of Brassicaceae. Trends Plant Sci. 2011;16:108–16.
49. Duchene D, Bromham L. Rates of molecular evolution and diversification in plants: chloroplast substitution rates correlated with species-richness in the Proteaceae. BMC Evol Biol. 2013;13:65.
50. Wicke S, Schäferhoff B, Depamphilis CW, Müller KF. Disproportional plastome-wide increase of substitution rates and relaxed purifying selection in genes of carnivorous Lentibulariaceae. Mol Biol Evol. 2014;31:529–45.
51. Drescher A, Stephanie R, Calsa T, Carrer H, Bock R. The two largest chloroplast genome-encoded open reading frames of higher plants are essential genes. Plant J. 2000;22:97–104.
52. Huang JL, Sun GL, Zhang DM. Molecular evolution and phylogeny of the angiosperm *ycf2* gene. J Syst Evol. 2010;48:240–8.
53. Kikuchi S, Bédard J, Hirano M, Hirabayashi Y, Oishi M, Imai M, et al. Uncovering the protein translocon at the chloroplast inner envelope membrane. Science. 2013;339:571–4.
54. Kode V, Mudd EA, Iamtham S, Day A. The tobacco plastid *accD* gene is essential and is required for leaf development. Plant J. 2005;44:237–44.
55. Rousseau-Gueutin M, Huang X, Higginson E, Ayliffe M, Day A, Timmis JN. Potential functional replacement of the plastidic acetyl-CoA carboxylase subunit (*accD*) gene by recent transfers to the nucleus in some angiosperm lineages. Plant Physiol. 2013;161:1918–29.
56. Ohlrogge J, Browse J. Lipid biosynthesis. Plant Cell. 1995;7:957–70.
57. Sasaki Y, Nagano Y. Plant acetyl-CoA carboxylase: structure, biosynthesis, regulation, and gene manipulation for plant breeding. Biosci Biotechnol Biochem. 2004;68:1175–84.

58. Feria Bourrellier AB, Valot B, Guillot A, Ambard-Bretteville F, Vidal J, Hodges M. Chloroplast acetyl-CoA carboxylase activity is 2-oxoglutarate-regulated by interaction of PII with the biotin carboxyl carrier subunit. Proc Natl Acad Sci U S A. 2010;107:502–7.
59. Madoka Y, Tomizawa K, Mizoi J, Nishida I, Nagano Y, Sasaki Y. Chloroplast transformation with modified *accD* operon increases acetyl- CoA carboxylase and causes extension of leaf longevity and increase in seed yield in tobacco. Plant Cell Physiol. 2002;43:1518–25.
60. Allen JF, de Paula WBM, Puthiyaveetil S, Nield J. A structural phylogenetic map for chloroplast photosynthesis. Trends Plant Sci. 2011;16:645–55.
61. Sen L, Fares M, Su Y-J, Wang T. Molecular evolution of *psbA* gene in ferns: unraveling selective pressure and co-evolutionary pattern. BMC Evol Biol. 2012;12:145.
62. Wang M, Kapralov MV, Anisimova M. Coevolution of amino acid residues in the key photosynthetic enzyme Rubisco. BMC Evol Biol. 2011;11:266.
63. Sen L, Fares MA, Liang B, Gao L, Wang B, Wang T, et al. Molecular evolution of *rbcL* in three gymnosperm families: identifying adaptive and coevolutionary patterns. Biol Direct. 2011;6:29.
64. Kapralov MV, Smith JAC, Filatov DA. Rubisco evolution in C4 eudicots: an analysis of Amaranthaceae *sensu lato*. PLoS One. 2012;7:e52974.
65. Tabita FR, Hanson TE, Satagopan S, Witte BH, Kreel NE. Phylogenetic and evolutionary relationships of RubisCO and the RubisCO-like proteins and the functional lessons provided by diverse molecular forms. Philos Trans R Soc Lond B Biol Sci. 2008;363:2629–40.
66. Tabita FR, Hanson TE, Li H, Satagopan S, Singh J, Chan S. Function, structure, and evolution of the RubisCO-like proteins and their RubisCO homologs. Microbiol Mol Biol Rev. 2007;71:576–99.
67. Studer RA, Christin P-A, Williams MA, Orengo CA. Stability-activity tradeoffs constrain the adaptive evolution of RubisCO. Proc Natl Acad Sci U S A. 2014;111:2223–8.
68. Parry MAJ. Manipulation of Rubisco: the amount, activity, function and regulation. J Exp Bot. 2003;54:1321–33.
69. Greiner S, Bock R. Tuning a ménage à trois: co-evolution and co-adaptation of nuclear and organellar genomes in plants. Bioessays. 2013;35:354–65.
70. Romiguier J, Figuet E, Galtier N, Douzery EJP, Boussau B, Dutheil JY, et al. Fast and robust characterization of time-heterogeneous sequence evolutionary processes using substitution mapping. PLoS One. 2012;7:e33852.
71. Gale J. Plants and altitude–revisited. Ann Bot. 2004;94:199.
72. Shi Z, Liu S, Liu X, Centritto M. Altitudinal variation in photosynthetic capacity, diffusional conductance and $\delta^{13}C$ of butterfly bush (*Buddleja davidii*) plants growing at high elevations. Physiol Plant. 2006;128:722–31.
73. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.
74. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008;18:821–9.
75. Gao S, Sung WK, Nagarajan N. Opera: reconstructing optimal genomic scaffolds with high-throughput paired-end sequences. J Comput Biol. 2011;18:1681–91.
76. Ronen R, Boucher C, Chitsaz H, Pevzner P. sEQuel: improving the accuracy of genome assemblies. Bioinformatics. 2012;28:i188–96.
77. Boetzer M, Pirovano W. Toward almost closed genomes with GapFiller. Genome Biol. 2012;13:R56.
78. Delcher AL, Phillippy A, Carlton J, Salzberg SL. Fast algorithms for large-scale genome alignment and comparison. Nucleic Acids Res. 2002;30:2478–83.
79. Grant JR, Arantes AS, Stothard P. Comparing thousands of circular genomes using the CGView Comparison Tool. BMC Genomics. 2012;13:202.
80. Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I. VISTA: computational tools for comparative genomics. Nucleic Acids Res. 2004;32:W273–9.
81. Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R. REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res. 2001;29:4633–42.
82. Liu C, Shi L, Zhu Y, Chen H, Zhang J, Lin X, et al. CpGAVAS, an integrated web server for the annotation, visualization, analysis, and GenBank submission of completely sequenced chloroplast genome sequences. BMC Genomics. 2012;13:715.
83. Wyman SK, Jansen RK, Boore JL. Automatic annotation of organellar genomes with DOGMA. Bioinformatics. 2004;20:3252–5.
84. Conant GC, Wolfe KH. GenomeVx: simple web-based creation of editable circular chromosome maps. Bioinformatics. 2008;24:861–2.
85. Schattner P, Brooks AN, Lowe TM. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. Nucleic Acids Res. 2005;33:W686–9.
86. Wright F. The “effective number of codons” used in a gene. Gene. 1990;87:23–9.
87. Ranwez V, Harispe S, Delsuc F, Douzery EJP. MACSE: Multiple Alignment of Coding SEquences accounting for frameshifts and stop codons. PLoS One. 2011;6:e22594.
88. Darriba D, Taboada GL, Doallo R, Posada D. jModelTest 2: more models, new heuristics and parallel computing. Nat Methods. 2012;9:772.
89. Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 2006;22:2688–90.
90. Stern A, Doron-Faigenboim A, Erez E, Martz E, Bacharach E, Pupko T. Selecton 2007: advanced models for detecting positive and purifying selection using a Bayesian inference approach. Nucleic Acids Res. 2007;35:W506-11.

Chapter 3: Comparative analysis and structural feature exploration with 12 newly assembled chloroplast genomes in tribe Cardamineae

3.1 A brief introduction of tribe Cardamineae

Cardamineae, comprising more than 340 annual or perennial species from ten genera, is an economically and phylogenetically important tribe in the Brassicaceae family [9]. The main genus is *Cardamine,* which includes about 200 spcies including those previously incuded in *Dentaria*. There are other two main genera in Cardamineae with more than 20 recognized species each, namely *Rorippa* (86 species) and *Barbarea* (25 species). Species in thise genera are distributed on all continents except for *Barbarea*, which does not occur in South America. Minor genera in Brassicaceae are *Nasturtium*, which includes five species, with two native to Mexico and the USA; the North American *Iodanthus* with 1, *Leavenworthia* with 8, *Ornithocarpa* with 2, *Planodes* with 1, and *Selenia* with 5 species. Two species of *Subularia* are currently recognized, one in Africa and the other in North America, but as of today they have not been validated by molecular proofs. Most of the Cardamineae species often prefer mesic or aquatic habitats. Most of the species in this tribe are featured with simple trichomes or are glabrous, with alternated leaves, accumbent cotyledons and a haploid chromosome number of x = 8 [83]. Besides the abundance of phenotypes and the great species diversity, the tribe has been attracting increasing attention, as it includes several economically important members with high commercial and ornamental value. One of the most economic values of Cardamineae is the production of natural medicine, like the one made from the whole plant of *Cardamine impatiens* and *Cardamine trifoliolata*. The other most economically important aspect is food, which has the longest history of cultivation and utilization in China, not only for ordinary dishes but also for edible oil used primarily in cooking (e.g. species *Cardamine pratensis* and *Cardamine limprichtiana*). Moreover, the Cardamineae species are of great ornamental value, particularly represented by *Cardamine circaeoides*, *Cardamine angustata*, and *Cardamine bulbosa* [84].

Due to the frequent hybridization and polyploidization of its species, genera of the Cardamineae tribes are taxonomically and phylogenetically regarded as some of the most challenging taxa in plants. Traditional classification of species by adopting a morphology-based system, usually affected by environmental factors, is frequently dynamic and unreliable. The shortage of suitable DNA fragments or polymorphic genetic markers for phylogenetic analysis has long hindered the obtainment of a reliable phylogeny. Moreover, the controversies about taxonomic classification have constituted an obstacle towards a clear understanding the diversification and evolution of the tribe Cardamineae. For instance, some previous studies by utilizing the simple sequence repeats (SSRs) [86], amplified fragment length polymorphisms (AFLPs) [87], random amplified polymorphic DNA (RAPD) [87], internal transcribed spacer (ITS) [86], [88] and several DNA loci [89], [90], provided some insights into the taxonomy and phylogeny of the *Cardamine* species, but still a satisfactory resolution is lacking.

Also a recent effort using ITS, *trnL* and *trnL-F* sequences of 38 *Cardamine* species has generated useful information but still failed to determine their phylogenetic relationships with a higher resolution [91].

In light of the moderate sequence divergence between plant species and individuals, the chloroplast genomes could provide valuable information for taxonomic classification and the reconstruction of a reliable phylogeny. Owing to the maternal transmission and absence of recombination, the chloroplast genomes are helpful for tracing source populations [92], [93] and for resolving complex evolutionary relationships [52], [53], [94]. This is particularly true in the case of Cardamineae, as the nuclear genomes are large, and thus nuclear data are not easily applicable to infer phylogenetic relationships [88]. In alternative, cp-derived markers, e.g. *rpl32-trnL*, *atpI-atpH*, *psbD-trnT*, *ycf6-psbM*, *ndhF* were previously successfully employed to study evolutionary relationships between plants [95], [96]. Repetitive sequences within the chloroplast genomes are also potentially useful for ecological and evolutionary studies of plants [97]. The advent of next-generation sequencing techniques makes now more convenient to obtain cp genome sequences than nuclear genes, thus allowing the transition from gene-based phylogenetics to phylogenomics.

In this study, we sequenced 12 Cardamineae chloroplast genomes using next-generation Illumina genome analyzer platform. Chloroplast genome sequence of other two species, namely Cardamine resedifolia and Cardamine impatiens, already obtained in a previous study in our laboratory and 20 other species in Brassicaceae available in NCBI plastid database were jointly used for analysis. This study aims to further examine patterns of structural variation in the Cardamineae cp genomes and to reconstruct phylogenetic relationships among the representative species. The complete cp genome sequences of Cardamineae reported here are an important prerequisite for classifying the “difficult taxa” and could also potentially be applied to modifying these economic important plants by chloroplast genetic engineering techniques.

3.2 Sampling and bioinformatic pipeline

A total of 12 species in the tribe Cardamineae, namely *Cardamine alpina*, *Cardamine asarifolia*, *Cardamine enneaphyllos*, *Cardamine flexuosa*, *Cardamine hirsuta*, *Cardamine pentaphyllos*, *Cardamine pratensis*, *Cardamine trifolia*, *Leavenworthia exigua*, *Leavenworthia uniflora*, *Rorippa austriaca*, *Rorippa sylvestris* and one close species *Descurainia bourgaeana*, were chosen for sequencing mainly because of their economic significance and phylogenetic placement in a recent study [98].

Young leaves were collected from plants grown in the greenhouse at the Ecogenomics laboratory (Research and Innovation Center, Fondazione Edmund Mach, Italy). Genomic DNA was extracted with the DNeasy Plant Mini kit (Qiagen GmbH, Hilden, Germany) according to manufacturer's instructions. Purified DNA was fragmented by nebulization with compressed nitrogen gas, and then short-insert (300 bp) libraries were constructed according to the manufacturer's protocol (Inc., San Diego, CA). Tags and adapters were attached to the small DNA fragments, then, sent to the Illumina's Genome Analyzer for sequencing. Above works were carried out by laboratory staff.

The first step when the reads were ready was to remove the sequencing primer by FASTX-Toolkit version 0.7 [99] available from http://hannonlab.cshl.edu/fastx_toolkit/. Then the sequence reads mixed with DNA from the nucleus and mitochondria were filtered by mapping to chloroplast reference genome. The quality assessment of rest reads of the chloroplast genome from Illumina sequencing platform was carried out by FastaQC Version 0.11.2 [100]; Raw reads with Q-value ≤ 30, namely poor quality reads were removed by FASTX-Toolkit as above.

Filtered PE reads were assembled using NGS assemblers Velvet 1.2.10 [101] for denovo assembly and contigs were produced; The longest contig was blasted to the NCBI plastid database for identification of the best reference genome for the next step. The assembled contigs were reordered and concatenated by MUMmer3.23 [102] according to the chosen reference genome from NCBI plastid database; Sequenced reads were mapped to the newly assembled plastome to fill the gaps by using GapCloser from SOAPdenovo [103]; The newly assembled chloroplast genome was separated according to its four regions and aligned with the reference sequences in the Brassicaceae family for validation.

Genome annotation, alignment, and visualization

The assembled chloroplast genome was submitted to the online plant chloroplast annotation software DOGMA (http://dogma.ccbb.utexas.edu/) for annotation [104], using default parameters. Protein coding sequences were extracted by the application on the website. Transfer RNA (tRNA) genes were detected by DOGMA and tRNAscan-SE [105]. Start and stop codons of protein-coding genes were checked one by one manually with blast hits against 15 cp reference genomes in DOGMA. OGDraw (version 1.2) was used for visualization of the plastome maps [106]. Global alignment of the 12 newly sequenced genomes with 22 reference genomes was carried out by MAFFT (version 7) [107] and adjusted manually when necessary. Full alignments with annotations were visualized by mVISTA [108]. Detection of various types of repeats is provided by REPuter [109], an evaluation of the significance and interactive visualization was calculated, default parameters were chosen for the settings while the minimal size for a repeat was limited to 30 bp. SSRs (simple sequence repeat, microsatellite) were detected by MISA (Microsatellite identification tool) available from http://pgrc.ipk-gatersleben.de/misa/.

Nucleotide substitution model calculation and phylogenetic analysis

Global alignment of the 12 newly sequenced genomes with 22 reference genomes (Additional file 16 Table C3-5) were done by MAFFT (version 7) [107]. The optimal nucleotide substitution model for the dataset was assessed by jModelTest 2.1.5 [110], and the best model selected was then used both in the maximum likelihood (ML) and Bayesian inference (BI) phylogenetic analyses.

For parsimony analysis, individual bases were considered multistate, unordered characters of equal weight; MP analyses were implemented in PAUP 4.0b10 [111]. Tree-bisection-reconnection (TBR) branch swapping was used. Consistency indices (CI) and retention indices (RI) were calculated to evaluate the amount of homoplasy.

Maximum likelihood analysis and ML bootstrapping (MLB) was performed by using the program PhyML version 3 [60]. The total number of bootstrap replicates was set to 100. For Bayesian inference (BI), the optimal model of sequence evolution was calculated by jModeltest and chosen according to the Akaike information criterion (AIC).

Bayesian analyses were conducted using MrBayes v.3.2.2 [112] allowing setting nucleotide substitution model for the dataset. Two independent runs of 20,000,000 generations were completed with four chains each (three heated, one cold), using a chain temperature of 0.2 and uniform priors. Trees were sampled every 1000 generations, and the first 25% of runs were discarded as burn-in. Likelihood-by-generation plots were created. A majority-rule consensus tree was produced from the remaining trees from the two runs, and posterior probabilities (PP) were collected.

All the phylogenetic trees were rooted with *Aethionema cordifolium* and *Aethionema grandiflorum* as the outgroup (Franzke, Lysak,Al-shehbaz, Koch, & Mummenhoff, 2011).

3.3 Feature exploration and a further phylogenetic analysis within tribe Cardamineae

Chloroplast genome sequencing and assembly

Using the Illumina genome analyzer platform HiSeq2000, we sequenced cp genomes of 12 species of Cardamineae (Table 3-1). These cp genomes used in our study were assembled and checked by two following steps: 1) Reference-guided assembly and gap filling; 2) Global alignment for all the 12 species with another 22 reference genomes were aligned, boundaries of large single copy (LSC), small single copy (SSC) and two inverted repeats (IRa and IRb) were checked by visual inspection.

Table 3-1 Summary of assembly and features for the 12 chloroplast genomes

Species number	N50	Contig Mean length	Contig Max length	IR (bp)	SSC (bp)	LSC (bp)	Final cp genome length
Rorippa.sylvestris	835 bp	501 bp	9.06 Kbp	26184	17869	83662	153.899Kbp
Cardamine.hirsuta	2.76 Kbp	1.01 Kbp	8.27 Kbp	26451	17792	83219	153.913Kbp
Cardamine.alpina	6.34 Kbp	735 bp	36.17 Kbp	26485	17872	84146	154.988Kbp
Cardamine.flexuosa	2.84 Kbp	1.07 Kbp	11.49 Kbp	26289	17673	83947	154.198Kbp
Rorippa.austriaca	19.18 Kbp	775 bp	157.28 Kbp	26460	18023	83315	154.258Kbp
Cardamine.enneaphyllos	2.92 Kbp	1.24 Kbp	11.41 Kbp	26465	17920	83905	154.755Kbp
Cardamine.pentaphyllos	326 bp	327 bp	25.96 Kbp	26477	17909	84373	155.236Kbp
Leavenworthia.uniflora	2.94 Kbp	1.17 Kbp	12.78 Kbp	26472	17854	83110	153.908Kbp
Leavenworthia.exigua	7.36 Kbp	758 bp	25.89 Kbp	26452	17864	83367	154.135Kbp
Cardamine.asarifolia	305 bp	315 bp	71.43 Kbp	26290	17694	83983	154.257Kbp
Cardamine.trifolia	1.84 Kbp	489 bp	37.00 Kbp	26136	17583	83216	153.071Kbp
Cardamine.pratensis	6.99 Kbp	558 bp	34.51 Kbp	26102	17629	83873	153.706Kbp

Illumina paired-end (read length 100 bp) sequencing produced big data sets for individual species. Paired-end reads were mapped to the reference cp genome, reaching a 200X coverage on average across these cp genomes.

After de novo and reference-guided assembly as described in chapter 2, we obtained 12 complete cp genomes with minor corrections. The four junction regions of each cp genome were validated by global alignment with corresponding area from 22 reference cp genomes (Additional file 11: Figure C3-1).

Conservation of Cardamineae chloroplast genomes

All twelve completely assembled Cardamineae cp genomes were revealed to have some identical sequences as the reference genome both at the start and end. They possessed the typical quadripartite structure of most angiosperms, including the large single copy (LSC), the small single copy (SSC) and a pair of inverted repeats (IRa and IRb). There were no obvious sequence inversions or genomic rearrangements (Figure 3-1).

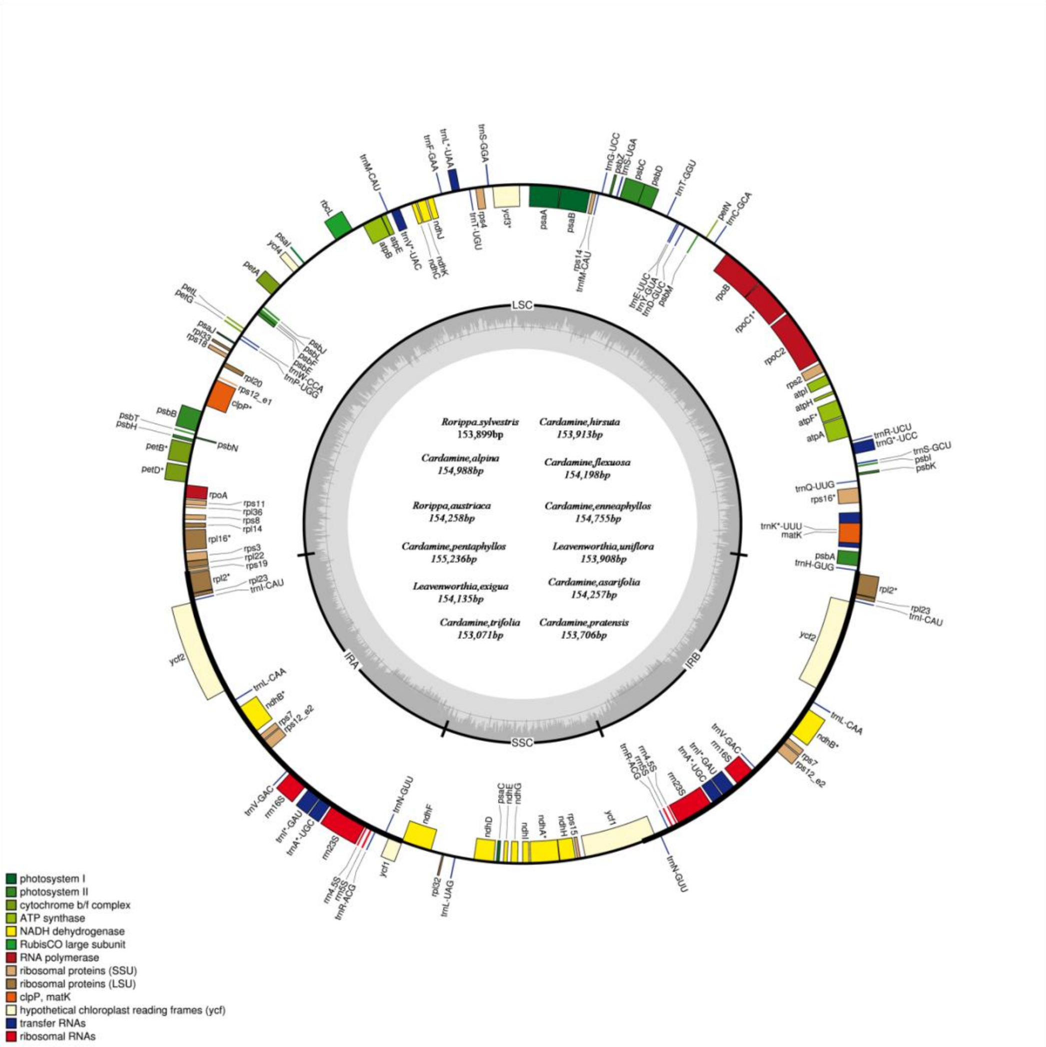


Figure 3-1 Chloroplast map of the twelve Cardamineae species cp genomes. Genes shown outside the outer circle are transcribed clockwise and those inside are transcribed counterclockwise. Genes belonging to different functional groups are color coded. Dashed area in the inner circle indicates the GC content of the chloroplast genome

Among these cp genomes, the complete size ranged from 153,071 bp (*C. trifolia*) to 155,236 bp (*C. pentaphyllos*). As for the quadripartite structure, the length varied from

83,216bp (*C.trifolia*) to 84,373 bp (*C.pentaphyllos*) in the LSC region, from 17,583 bp (*C.trifolia*) to 18,023 bp (*Rorippa austriaca*) in the SSC region, from 26,102 bp (*C.pratensis*) to 26,485 bp (*C.alpina*) in IR region. Each cp genome was found to contain a total of 130 genes, including 85 protein-coding genes, 37 transfer RNA (tRNA) genes and 8 ribosomal RNA (rRNA) genes (Table 3-2). Of them, we identified 12 protein-coding genes, 14 tRNA-coding genes, and 8 rRNA coding genes were located within two IRs. The LSC region contained 60 protein-coding and 22 tRNA genes while the SSC region contained 10 protein-coding and one tRNA gene. The *rps12* gene was a unique gene divided with the 5′end exon located in the LSC region while two copies of the 3′ end exon and intron were located in the IRs. The *ycf1* was located at the boundary regions between IRa/SSC/IRb, leading to the incomplete duplication of the gene within IRs. There were 18 intron-containing genes, including 6 tRNA genes and 12 protein-coding genes, almost all of which were single-intron genes except for *ycf3* and *clpP*, which independently had two introns. *matK* was located within the intron of *trnK*-UUU, the largest intron.

Table 3-2 List of genes encoded in 12 newly sequenced *Cardamineae* chloroplast genome

Gene Category	**Genes**
ribosomal RNAS	§_rrn4.5_, § _rrn5_, §_rrn16_, §_rrn23_
transfer RNAs	§*_trnA-UGC_, _trnC-GCA_, _trnD-GUC_, _trnE-UUC_, _trnF-GAA_, _trnfM-CAU_, *_trnG-UCC_, _trnG-UCC_, _trnH-GUG_, §_trnI-CAU_, §*_trnI-GAU_, *_trnK-UUU_, §_trnL-CAA_, *_trnL-UAA_, _trnL-UAG_, _trnM-CAU_, §_trnN-GUU_, _trnP-UGG_, _trnQ-UUG_, §_trnR-ACG_,_trnR-UCU_, _trnS-GCU_, _trnS-UGA_, _trnS-GGA_, _trnT-UGU_, _trnT-GGU_, *_trnV-UAC_, §_trnV-GAC_, _trnW-CCA_, _trnY-GUA_
Photosystem I	_psaA_, _psaB_, _psaC_, _psaI_, _psaJ_
Photosystem II	_psbA_, _psbB_, _psbC_, _psbD_, _psbE_, _psbF_, _psbH_, _psbI_, _psbJ_, _psbK_, _psbL_, _psbM_, _psbN_, _psbT_, _psbZ_
Cytochrome	_petA_, *_petB_, *_petD_, _petG_, _petL_, _petN_
ATP synthase	_atpA_, _atpB_, _atpE_, *_atpF_, _atpH_, _atpI_
Rubisco	_rbcL_
NADH dehydrogenase	*_ndhA_,§*_ndhB_, _ndhC_, _ndhD_, _ndhE_, _ndhF_, _ndhG_, _ndhH_, _ndhI_, _ndhJ_, _ndhK_
Ribosomal protein (large subunit)	§*_rpl2_, _rpl14_, *_rpl16_, _rpl20_, _rpl22_,§_rpl23_, _rpl32_, _rpl33_, _rpl36_
Ribosomal protein (small subunit)	_rps2_, _rps3_, _rps4_, §_rps7_, _rps8_, _rps11_, §*_rps12_, _rps14_, _rps15_, *_rps16_, _rps18_, _rps19_
RNA polymerase	_rpoA_, _rpoB_, *_rpoC1_, _rpoC2_
ATP-dependent protease	*_clpP_
Cytochrome c biogenesis	_ccsA_
Membrane protein	_cemA_
Maturase	_matK_
Conserved reading frames	_ycf1_short_, _ycf1_long_, §_ycf2_, *_ycf3_, _ycf4_
Pseudogenes	_accD_

§Gene completely duplicated in the inverted repeat. *Gene with intron(s)

It was found that *ycf1*, *accD*, *rpl23* and *ycf2* were often absent in plants [113], but they were detected in the reported Cardamineae cp genomes in this study. As in other higher plants, one pair of genes, *atpB-atpE*, were observed to overlap with each other by 4 bp. However, *psbC-psbD* had a 52 bp overlapped region in the Cardamineae cp genomes, the same as the case in *Camellia* [114], but different from the situation in *Gossypium*, where there was a 53 bp overlapped area [115]. The overall GC content was approximately 36.36%, which is almost identical to that present in the twelve complete Cardamineae cp genomes (Additional file: 12 Table C3-1).

Although genome size and overall genomic structure, including gene number and gene order, are highly conserved, the IR expansion/contraction is common among plant cp genomes. For example, the end of two genes, *ndhH* and *ndhF*, were reported to have repeatedly migrated into and outside of the adjacent IRs in grasses [116]. The whole *rps19* was located within the LSC region in the majority of *Gossypium* cp genomes, but was not found in the cp genome of *G. raimondii* D5 [115]. Kim and colleagues considered that the length of angiosperm cp genomes was variable primarily due to the expansion and contraction happened between the inverted repeat IR region and the single-copy boundary regions [117]. The IR/SC boundary regions of our 12 newly sequenced Cardamineae cp genomes and another 22 reference genomes were compared, showing slight differences at junction positions (Additional file 11 Figure C3-1).

The junction positions were generally conserved across all the Cardamineae cp genomes, however, small differences also exist. At the border between IRa and SSC, the incomplete *ycf1 5′* stretches crossed the border and had an average 30 bp extension, and overlapped with the end of the *ndhF* gene. For the *ndhF*, it extended into the IRa area in most of the genomes, except for *Cardamine enneaphyllos*, where the termini of *ndhF* were still limited in SSC area. At the junction between IRa and SSC, the sequence is also highly conserved and only some base pairs shift happened in several genomes.

At the border between IRb and LSC, the incomplete copy of the *rps9* gene extended from IRb to LSC but just stopped at the border. There is no obvious contraction or expansion, except for a 9 bp gap in species *Cardamine trifolia* when aligned with other cp genomes, but this situation was the same as cp genomes of species *Arabis hirsuta* and *Lobularia maritima*.

The junction between LSC and Ira was included in the *rps19* gene stretches. This area of the 34 aligned sequences was highly conserved and there were no more than 3 bp expansion in the LSC area to IRa in the species *Cardamine trifolia*. The same was observed when compared to the reference species *Arabis hirsuta* and *Lobularia maritima*.

In the last junction between SSC and IRb, the intergenic area had more variations among plastomes. There are around 20 bp and 50 bp gaps in *Cardamine hirsuta* and *Cardamine enneaphyllos* when aligned to other plastomes. Except the two, there were only several base pairs different at the termini of the SSC area for the rest of the Cardamineae plastomes, and the same was observed for the termini of IRb near the border (Additional file 11 Figure C3-1).

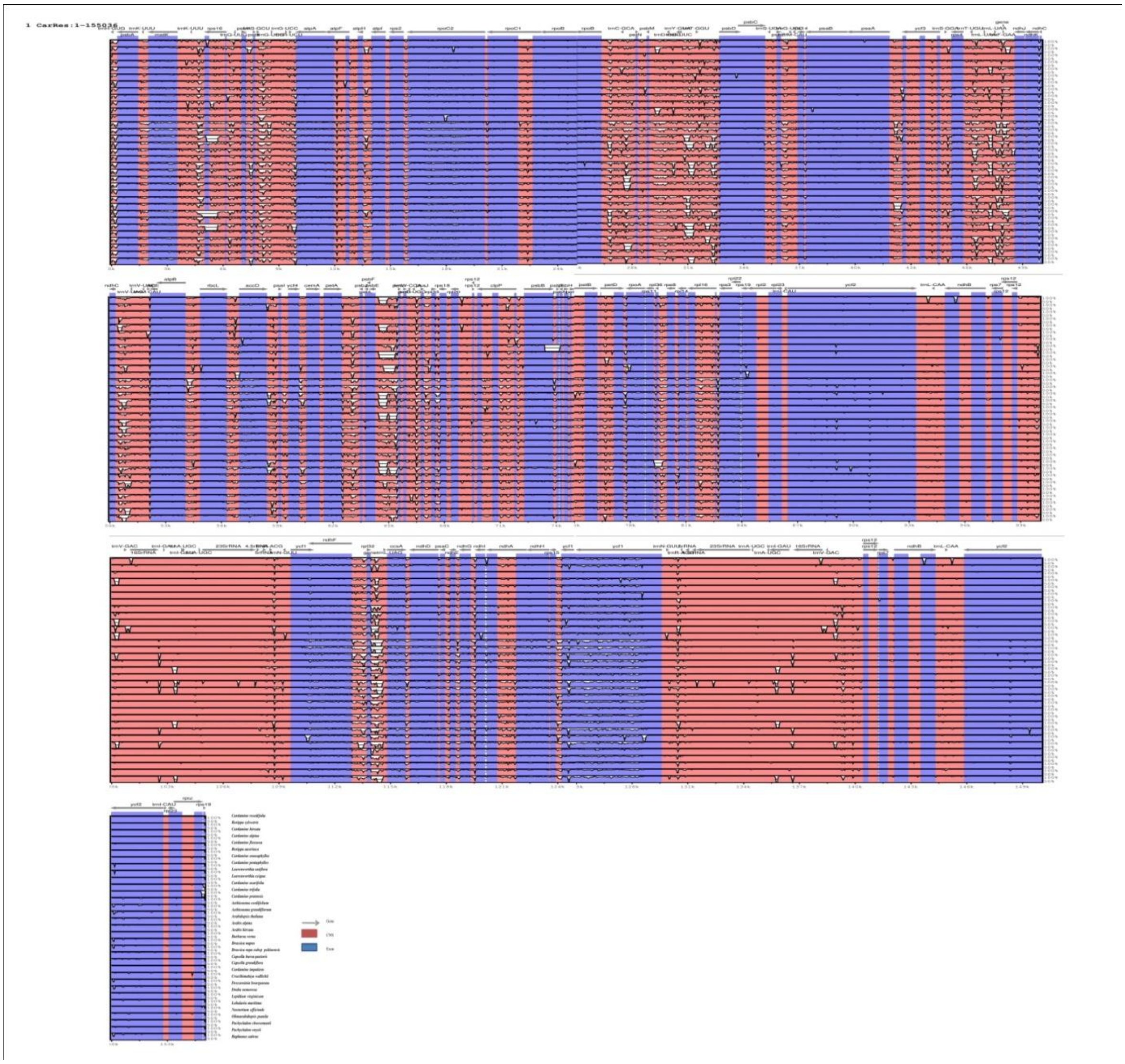


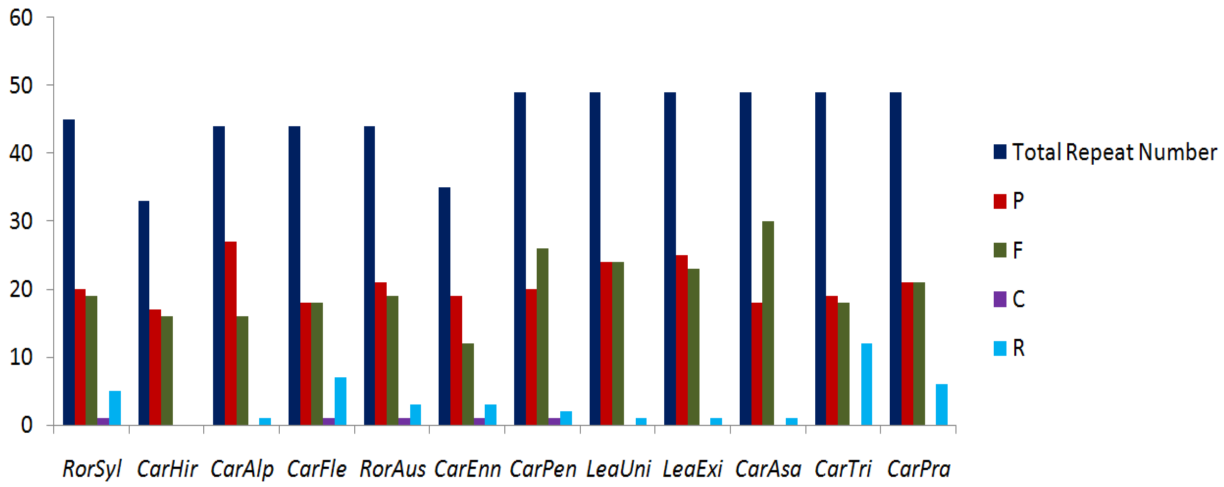
Figure 3-2 Visualization alignment of 34 chloroplast genome. VISTA-based identity plots shown sequence identity between the 34 chloroplast genomes. Genome regions were color-coded as protein coding, rRNA coding, tRNA coding or conserved noncoding sequences (CNS).

To investigate the levels of genome divergence, alignments of 12 newly sequenced Cardamineae cp genome sequences and 22 reference cp genomes in the Brassicaceae family were performed with *Cardamine resedifolia* as a reference. We plotted sequence identity using VISTA [118]. The results revealed high sequence similarity across the 34 chloroplast genomes, especially in the Cardamineae tribe, suggesting that Cardamineae cp genomes were rather conserved. Differences between the Cardamineae cp genomes and other plants did exist but they were minor and not worth a detailed description. As expected, the IRs was more conserved than single-copy regions, and coding regions were more conserved than noncoding regions. The most divergent coding regions were *matK*, *rpoC2*, *accD*, *rps19*, and *ycf1* (Figure 3-2).

Repetitive sequences

We divided repeats into four categories: forward, reverse, complement and palindromic repeats. For all repeat types, the minimal cut-off for identifying two copies was set to 90%. The minimal copy size screened was 30 bp for all. In total, around 50 repeats were detected in the Cardamineae cp genomes by REPuter [119] (Table 3-3).

Table 3-3 Analysis of repeated sequences in the 12 Cardamineae chloroplast genomes



Among the four types identified, palindromic repeats and tandem repeats were the most common ones, accounting for 91% of total repeats on average. Besides, except the large IR repeat region, the longest repeat in most of the cp genomes was no more than 70 bp, indicating that the presence of large repeats was under relatively strong negative selection. Especially a 67 bp repeat was shared in most of the Cardamineae cp genomes analyzed (Additional file 14 Table C3-3). Numbers for the four types of repeat detected in these 12 cp genomes were very close. In light of the high similarity of the cp sequences, one can expect that their overall repeat distribution in the cp genome should be highly conserved. However, in *Cardamine hirsuta*, no complement or reverse repeat longer than 30 bp was detected. Besides, the total repeat number turned out to be less than in other cp genomes.

Even though the analyzed cp genomes contained a similar pattern of repeats, the number for each category of repeat in each species was unique, pointing out the most common sources of sequence variation in cp genomes of Cardamineae tribe. The pattern of repeats was so variable that it could be used as a reference for species identification, or also serve as an important source of species-specific genetic marker for phylogenetic and population genetic studies.

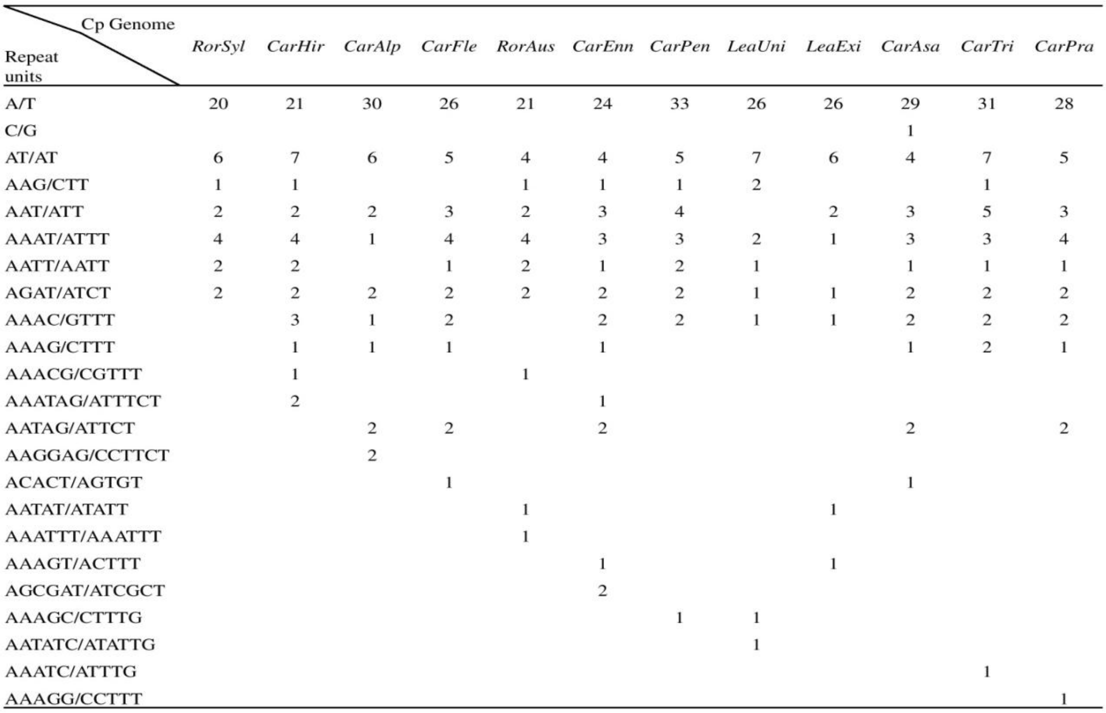
SSR polymorphisms

Compared to other neutral DNA regions, SSRs usually have a higher mutation rate due to slipped-DNA strands. They thus were often treated as genetic markers, possessing useful information concerning plant population genetic structuring in ecological and evolutionary studies due to their non-recombinant, haploid and uniparentally inherited nature[120], [121].

Using a threshold of 12 base pairs for mononucleotide, 6 for dinucleotide, 4 for trinucleotide and 3 for tetranucleotide, pentanucleotide, and hexanucleotide in MISA (MicroSAtellite Identification Tool), mononucleotide to hexanucleotide repeats were detected in the 12 Cardamineae cp genomes. On average, more than 37 SSRs were found for each cp genome (Additional file 15 Table C3-4). The repeat unit A/T was found to be the most abundant even with a threshold over 12 (Table 3-4), this finding was consistent with the previous discovery that cp SSRs were dominated by A or T mononucleotide repeats [122].

Some of the SSRs identified were extremely rare while others were very common. For instance, only one mononucleotide (C/G) repeat was found in *Cardamine asarifolia* cp genome; one pentanucleotide (AAATC/ATTTG) was found in *Cardamine trifolia*, another pentanucleotide (AAAGG/CCTTT) was only found in *Cardamine pratensis*. In the case of hexanucleotide repeat, one (AATATC/ATATTG) was only found in *Leavenworthia uniflora*. On the other hand, four kinds of SSR, one mononucleotide (A/T), one dinucleotide (AT/AT) and two tetranucleotides (AAAT/ATTT and AGAT/ATCT) were commonly shared among all the cp genomes. Mononucleotide to hexanucleotide repeats were mainly composed of A or T base pairs, which contributed to the overall A-T richness in the cp genomes [123], [124]. The variations of number and length in SSRs had been reported to be useful in species identification and studies of varieties and population genetics [125], [126]. As in the case of longer repeats mentioned in the former paragraph, the SSRs characterized in our cp genome showed a unique distribution with different number and length in each species, which should be useful in their application to polymorphisms study on population-level and phylogenetic relationships comparison among close organism at the species level or below [114] (Table 3-4).

Table 3-4 Simple sequence repeats (SSRs) in the12 Cardamineae chloroplast genomes



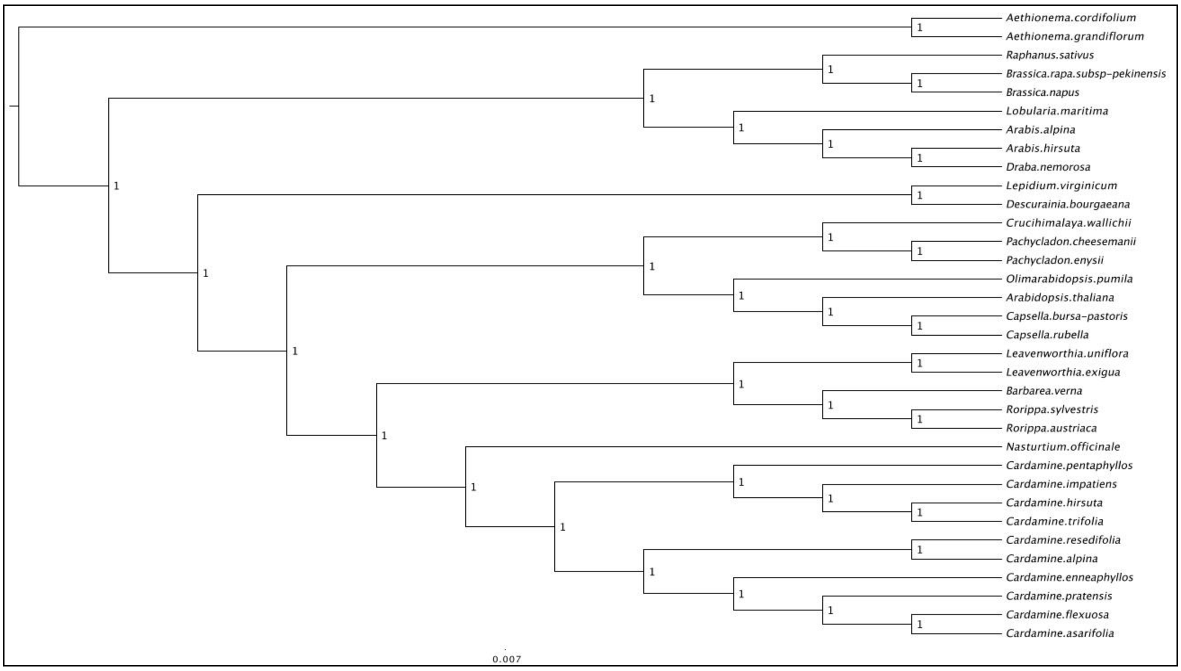
Cp Genome Repeat units	*RorSyl*	*CarHir*	*CarAlp*	*CarFle*	*RorAus*	*CarEnn*	*CarPen*	*LeaUni*	*LeaExi*	*CarAsa*	*CarTri*	*CarPra*
A/T	20	21	30	26	21	24	33	26	26	29	31	28
C/G										1		
AT/AT	6	7	6	5	4	4	5	7	6	4	7	5
AAG/CTT	1	1			1	1	1	2			1	
AAT/ATT	2	2	2	3	2	3	4		2	3	5	3
AAAT/ATTT	4	4	1	4	4	3	3	2	1	3	3	4
AATT/AATT	2	2		1	2	1	2	1		1	1	1
AGAT/ATCT	2	2	2	2	2	2	2	1	1	2	2	2
AAAC/GTTT		3	1	2		2	2	1	1	2	2	2
AAAG/CTTT		1	1	1		1				1	2	1
AAACG/CGTTT		1			1							
AAATAG/ATTTCT		2				1						
AATAG/ATTCT			2	2		2				2		2
AAGGAG/CCTTCT			2									
ACACT/AGTGT				1						1		
AATAT/ATATT					1				1			
AAATTT/AAATTT					1							
AAAGT/ACTTT						1			1			
AGCGAT/ATCGCT						2						
AAAGC/CTTTG							1	1				
AATATC/ATATTG								1				
AAATC/ATTTG											1	
AAAGG/CCTTT												1

Phylogenetic analyses

The application of cp genome sequence in phylogenetic studies had successfully addressed several phylogenetic issues in angiosperm[94], [123], [127], but the situation in tribe Cardamineae was not good, even the number of genera was not fixed [2], [9], therefore, our new phylogeny was a useful attempt.

In the reconstructed tree, the 12 newly sequenced species with *Barbarea verna, Nasturtium officinale, Cardamine resedifoia and Cardamine impatiens* were successfully clustered into one clade, which was tribe *Cardamineae.* Within it, it was notable that *Leavenworthia uniflora* and *Leavenworthia exigua* together from the same genera *Leavenworthia* were the sister group to the branch consisting of *Barbarea verna* and two species from genera *Rorippa,* namely *Rorippa sylvestris and Rorippa austriaca. Nasturtium officinale* was the sister group to the genus *Cardamine,* which included all other ten species and were well supported as monophyletic. The relationships among all the species got full pp (Posterior probability = 1) support inside the tribe Cardamineae. Besides, the relationship between Cardamineae and other species in Brassicaceae were also solved and fully supported.

Figure 3-3 Phylogenetic relationships of the twelve newly sequenced species and twenty-two reference cp genome in Brassicaceae constructed by Mrbayes method, with all branches fully supported.



3.4 Conclusions

We sequenced 12 complete cp genomes in the tribe Cardamineae via a combination of de novo and reference-guided assembly based on Illumina sequencing technology. These cp genomes were found highly conserved. We investigated the variation of repeat sequences, SSRs among the 12 complete Cardamineae cp genomes, which presented a wide diversity in the tribe Cardamineae. The unique long repeat and SSR could serve as potential molecular markers for further species identification.

As the first well-supported phylogenomic analyses of Cardamineae, our results indicate that the use of cp genome in the phylogenetic study can classify well the Cardamineae species, and provided well-supported evolutionary relationships among speices. The obtained cp genomes may facilitate the development of biotechnological applications for these economically important plants, also offer useful genetic information for purposes related to phylogenetics, taxonomy and species identification in the Brassicaceae family. Thus a further taxon sampling and more complete cp genomes in Brassicaceae are necessary for in-depth analyses of trait evolution and adaptation in this family.

Chapter 4: Molecular Phylogeny in Brassicaceae based on 71 chloroplast protein-coding genes: Species relocation and taxonomic implications

4.1 Short introduction of phylogeny in Brassicaceae

Brassicaceae (Cruciferae or mustard family) is a worldwide distributed family with approximately 338 genera and 3,709 species [128], and is of special interest as it includes many economically important crop plants, ornamentals as well as model organisms (Such as *Arabidopsis thaliana*, *Brassica napus*, *Arabis alpina* and few others) in the plant sciences [2]. The broad distribution (except Antarctica) and high species diversity made the evolutionary study of Brassicaceae perhaps one of the most enigmatic, problematic and fascinating issues in recent plant evolutionary biology.

The Mediterranean region, as the most important center of species diversification, provides an excellent basis to perform various evolutionary, biogeographic or phylogeographic studies at different taxonomic levels [3]. Reconstructing the phylogenetic relationships among members of the family is essential for understanding the taxonomy of this ecologically and economically important group of angiosperms. During the past few years, the most dynamic period of significant taxonomic changes started with isozymes and continued with increasing DNA data [129].

Meanwhile, the number of tribes recognized in the mustard family has explosively increased from 25 [130] to 49 [131]. Three major lineages (I, II, III) have been recognized in the Brassicaceae phylogeny with chloroplast and nuclear markers [130], [132]. Subsequent studies, ITS-based phylogeny [21], supernetwork (ADH, CHS, ITS, *matK*) phylogeny [133] and mitochondrial *nad4* intron phylogeny [134] provided substantial support to the new tribal system, also were mostly in congruence with each other.

However, at the deeper nodes of the family tree, some results were contradictory. Such as the ancestral position of the Cochlearieae [133], which was not supported by the *ndhF* [130], [132] or ITS data [21]. Besides, most of the tribes recognized by Al-Shehbaz and others [131] were clearly delimited, however, some tribes were still roughly delimited or were paraphyletic and needed further splitting [135]. Moreover, much less significant support was available for the relationships between the various tribes, and several genera within the Brassicaceae were also poorly circumscribed. A reliable phylogenetic framework is required to restructure the classification of members in the family, in particular, the most species-rich and polyphyletic genera.

Progress toward resolving the family phylogeny and establishing the monophyly of its genera has been slow down due to firstly the limited selection of informative molecular markers, as phylogenetic hypotheses based on single markers (e.g., plastidic, mitochondrial or nuclear) possess a limited value [133], [136]. Secondly, classification schemes proposed solely on morphological characters were not fully supported by modern molecular systematic data.

In this study, the main aims were to (1) resolve the relationships within and among the main clades comprising the family Brassicaceae, and (2) place previously unsampled genera and species. To achieve these goals, the taxonomic sampling within Brassicaceae will increase to a number, which could be representatives of the current species at generic and subgeneric levels. In addition, sequence data from all commonly shared protein-coding sequences was used. Parsimony, likelihood, and bayesian analyses were performed to infer the phylogeny. The taxonomic implications for the discovery of novel clades were discussed.

4.2 Sampling and bioinformatic pipeline

Taxon sampling, DNA extraction, amplification, and sequencing

The taxon sampling included 80 representative species of the family, including around 36 tribes in Brassicaceae. Sampling was carried out across Trentino (Italy) by laboratory staff. 15 species from NCBI database belonging to this family served as cp reference genome were combined in this study (Additional file 17 Table C4-1).

Plants were grown in the greenhouses at the Ecogenomics laboratory (Research and Innovation Center, Fondazione, Italy). Herbarium vouchers were collected when plants flowered. Germination time and flowering time were recorded for all specimens.

Leaf materials for DNA extraction were collected from all the 80 species, immediately dried with silica gel and preserved at a low temperature until next step or directly frozen at -80 ℃ in the freezer for a backup. The dried and frozen tissue was finely ground in liquid nitrogen with a ceramic pestle. Genomic DNA was extracted with the DNeasy Plant Mini kit (Qiagen GmbH, Hilden, Germany) according to manufacturer’s instructions. Quality and quantification inspection of DNA was assessed with 1% agarose gels. Purified DNA was fragmented by nebulization with compressed nitrogen gas, and then constructed into short-insert (300 bp) libraries according to the manufacturer’s protocol (Inc., San Diego, CA). Tags for index and adapters were attached to the small DNA fragments, and then sequenced on the Illumina’s Genome Analyzer HiSeq2000 of the Next Generation Sequencing facility at CIBIO (Povo, Trento, Italy). This part of the work was done by laboratory staff.

Data processing, cp genome assembly, and annotation

Raw reads were checked for quality control before assembly; the first step was to sort the reads according to the tags and remove the adapter by FASTX-Toolkit (version 0.7)[99]. The DNA from the nucleus and mitochondria were filtered by mapping all the reads to the cp reference genome to exclude contaminants. The selected reads were checked by FastQC (Version 0.11.2)[100], and raw reads with Q-value ≤ 30, were removed by FASTX-Toolkit. Filtered PE reads were processed by the de novo assembler software Edena (V3)[137] for de novo assembly and produced the preliminary contigs. The longest contig was chosen to blast against the NCBI plastid database to find the best reference genome for next step. With the reference cp genome, the MUMmer (3.23) [102] package could reorder and rearrange all the contigs and produce a pseudomolecule, with discontinuous areas (gaps) being represented by N. These gaps were filled by GapCloser from the SOAPde novo package by mapping formerly cleaned reads to the pseudomolecule [103].

Assembled chloroplast genomes were submitted to online plant chloroplast annotation software DOGMA (http://dogma.ccbb.utexas.edu/) [104], using default parameters. Protein coding sequences were extracted by the application on the website. Transfer RNA (tRNA) genes were detected by DOGMA and tRNAscan-SE [105]. Stop and start codons of protein-coding genes were confirmed one by one manually with blast hits against 15 cp reference genomes in DOGMA.

Protein coding sequence extraction, arrangement, and alignment

The protein-coding sequences were extracted from the newly assembled chloroplast genomes, also from another 15 cp reference genomes. The sequences of commonly shared protein-coding genes by the 95 species were extracted and sorted by the gene names. Each gene was aligned by codon within MACSE [138]. Given the uniparental inheritance and lack of recombination in the chloroplast genome, aligned gene sequences were concatenated by a Perl script. The concatenated sequences formed the final supermatrix dataset, encompassing a total of 71 genes out of 95 species.

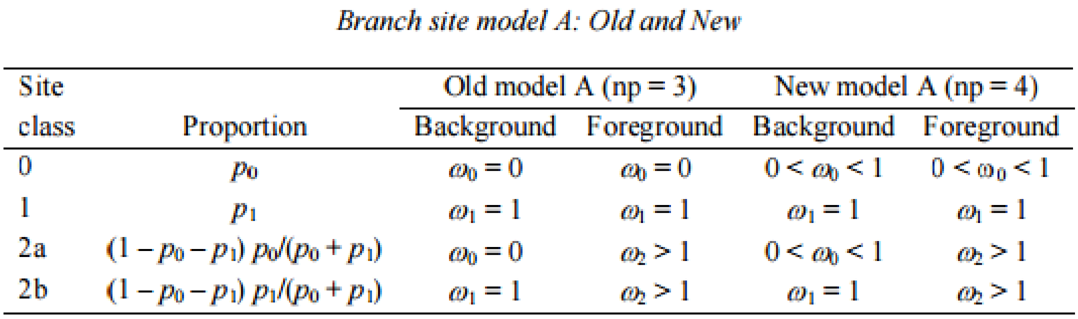
Sequence variation and selective pressure analysis of 71 chloroplast protein-coding genes

Sequence variation analysis was carried out for each of the 71 chloroplast fragments, the number of parsimony informative sites was calculated, as well as transitional and transversional pairs.

During the evolution of each taxon, certain traits or alleles of genes segregating within a population may be subjected to selection. When these traits were associated with a genetic basis, selection can increase the prevalence of those traits in the population. If the selection is persistent and intense, adaptive traits will become universal to the population or species and fixed. A typical feature of positive selection is a high non-synonymous substitution rate, dN (leading to amino acids change), compared to synonymous substitution, dS. The ratio between dN and dS, also called ω (omega), is therefore used to detect signatures of selection acting on specific coding sequences. When ω is larger than 1, the majority of the mutations affecting the corresponding codons are non-conservative, indicating that positive or relaxed selection took place. In our case, the software Selecton was chosen for the detection of the site-specific positive selection of genes.

Based on the result of Selecton, genes with positive selection sites were further analyzed under the Branch–site model with codeml within PAML package. Branch-site model [139] implemented two models, namely A and B, which allowed the ω ratio to vary among different sites and different lineages. The models attempt to detect positive selection, which affects only a few sites along a few lineages. The parameter settings for model A are model = 2 NSsites = 2, for model B are model = 2 NSsites = 3. A change was made for model A compared to before, as detailed below [140]. The new ω_0 is estimated from the data, which varied from 0 to 1, not equal to 0 as before. Within this new branch-site model A, the comparison between M1 and new M1a (NearlyNeutral) model will form a likelihood ratio test, with d.f. ≈ 2. This is called test 1. This test can mistake relaxed selective constraint on the foreground branches as positive selection. Hence, a significant result does not necessarily mean positive selection. Another test, called test 2 or branch-site test of positive selection, uses the same alternative model A, but the null model is model A with ω_2 = 1 fixed (fix_omega = 1 and omega = 1 in codeml.ctl). Test 2 appears to be a robust test of positive selection on the foreground branches and is called the branch-site test of positive selection.

Similarly, both the NEB and BEB methods for calculating posterior probabilities for site classes were implemented for the modified branch-site model A. Model A in combination with the BEB procedure were chosen for analysis.



Branch site model A: Old and New

Site class	Proportion	Old model A (np = 3)		New model A (np = 4)	
		Background	Foreground	Background	Foreground
0	p_0	$\omega_0 = 0$	$\omega_0 = 0$	$0 < \omega_0 < 1$	$0 < \omega_0 < 1$
1	p_1	$\omega_1 = 1$	$\omega_1 = 1$	$\omega_1 = 1$	$\omega_1 = 1$
2a	$(1 - p_0 - p_1)\ p_0/(p_0 + p_1)$	$\omega_0 = 0$	$\omega_2 > 1$	$0 < \omega_0 < 1$	$\omega_2 > 1$
2b	$(1 - p_0 - p_1)\ p_1/(p_0 + p_1)$	$\omega_1 = 1$	$\omega_2 > 1$	$\omega_1 = 1$	$\omega_2 > 1$

(quoted from the user guide of PAML package (version 4))

Nucleotide substitution, model calculation, and phylogeny inference

As for the concatenated sequence, it was submitted to jModelTest 2.1.5 for calculating the best nucleotide substitution model [110], this model will be used in the maximum likelihood (ML) and Bayesian inference (BI) phylogenetic analysis.

Phylogenetic relationship was inferred from the nucleotide data using maximum parsimony (MP). Phylogenetic trees were rooted using *Chleome hirta* and *Chleome spynosa* as out-group [29]. For parsimony analysis, individual bases were considered multistate, unordered characters of equal weight; MP analyses were implemented in PAUP 4.0b10 [111]. Tree-bisection-reconnection (TBR) branch swapping was adopted. Retention indices (RI) and consistency indices (CI) were calculated to evaluate the amount of homoplasy.

Maximum likelihood analysis and ML bootstrapping (MLB) were performed using the program RAxML version 8 [141]. The number of bootstraps replicates was set to 100. For Bayesian methods (BI), the optimal model of sequence evolution was calculated by jModeltest, chosen according to the Akaike information criterion (AIC).

MrBayes v.3.2.2 was adopted for the bayesian analyses [112], which allows setting nucleotide substitution model for the dataset. Two independent runs of 20,000,000 generations were completed with four chains each (three heated, one cold), by using a chain temperature of 0.2 and uniform priors. Trees were sampled every 1000 generations, and the first 25% of runs were discarded as burn-in. Likelihood-by-generation plots were created. A majority-rule consensus tree from the remaining trees produced by the two runs was inferred, posterior probabilities values (PP) were collected.

4.3 Result of cp genome assembly and new phylogenetic reconstruction in Brassicaceae family

Chloroplast genome sequencing, assembly, and annotation

In total, including *Cardamine resedifolia* and *Cardamine impatiens*, 80 species were newly sampled and sequenced. The length of the paired-end reads was 100 bp. The average number of reads for each plastome is 2,507,996. The average size of chloroplast genome is around 155,000 (154,694) bp. The average depth of sequencing is thus about 800 (817) (Table. 4-1).

Table 4-1 Summary of assembly of chloroplast genome for 80 species in Brassicaceae

Species Name	Reads Number	ContigsNumber	N50	Estimated Depth	Pseudomolecular
Berteroa incana	2999844	1942	266	977.368529	128457
Alyssum alissoides	3062700	5153	135	997.8474194	144812
Fibigia clypeata	2728212	3863	201	888.869071	156087
Matthiola fruticulosa	2594036	3028	213	845.1536645	128868
Bunias orientalis	2629820	4544	196	856.8123226	128464
Draba verna	1941988	1857	452	632.7122194	153914
Draba dubia	1887084	6190	290	614.8241419	137886
Arabis alpina	2438728	3149	296	794.5533161	153890
Arabis hirsuta Aggreg	1569304	5327	393	511.2893677	156459
Arabis nova	3042836	1238	592	991.3756	153529
Arabis soyeri subsp subcoriacea	1418444	4523	309	462.1382065	128243
Arabis turrita	3087084	2434	529	1005.791884	131848
Boechera gracilipes	3349244	2466	403	1091.205303	155429
Phoenicaulis cheiranthoides	2593912	3663	392	845.1132645	129552
Polyctenium fremontii	2548096	3531	540	830.1861161	154981
Diplotaxis tenuifolia	1001488	5119	264	326.2912516	133661
Brassica repanda susp baldensis	2404168	2947	255	783.2934452	133577
Hirschfeldia incana	1145660	3014	308	373.2634194	146737
Camelina microcarpa	2394012	1974	306	779.9845548	155554
Capsella grandiflora	1811176	865	1511	590.0928258	155306
Erysimum aurantiacum	1480972	4124	335	482.5102323	149916
Erysimum rhaeticum	2518844	20107	193	820.6556258	154424
Erysimum sylvestre	2817132	4374	346	917.8397806	154459
Erysimum virgatum	2102120	3718	190	684.8842581	138547
Neslia paniculata	2340736	1729	384	762.6268903	129431
Rorippa sylvestris	2497488	3362	379	813.6977032	154437
Cardamine hirsuta	3099920	2877	283	1009.973935	159840
Cardamine alpina	2500352	2265	323	814.6308129	130179
Cardamine flexuosa	2914608	2751	335	949.5980903	133018
Rorippa austriaca	1754756	2213	390	571.7108258	129067
Dentaria enneaphyllos	3047592	5505	265	992.9251355	130418
Dentaria pentaphyllos	2074116	8077	265	675.7603742	150642

Leavenworthia uniflora	2763736	1069	820	900.4430194	155675
Leavenworthia exigua	3344420	3242	322	1089.633613	129360
Cochlearia officinalis	3915604	5005	288	1275.729045	137123
Descurainia bourgaeana	1922576	1455	366	626.3876645	147834
Descurainia sofia	2790740	327	1286	909.2410968	128917
Hornungia petraea	2150672	1151	912	700.7028129	139713
Hutchinsia alpina	2896852	20649	143	943.813071	161800
Hutchinsia brevicaulis	2633200	2262	350	857.9135484	155251
Hymenolobus pauciflorus	2103652	2427	445	685.3833935	154327
Malcolmia littorea	3207388	4098	273	1044.987703	131467
Morettia philaeana	1695472	507	308	552.3957161	128582
Thellungiella halophila	2828928	1662	457	921.6829935	154310
Halimolobos pubens	3475880	3959	485	1132.464129	130113
Heliophila coronopifolia	3117228	1098	645	1015.612994	154245
Hesperis matronalis	2980812	4698	177	971.1677806	154767
Iberis amara	1297980	3558	248	422.8902581	147147
Isatis tinctoria	2878588	2185	340	937.8625419	128148
Lepidium campestris	2935484	2705	238	956.3996258	129598
Cardaria draba	1062324	809	1883	346.1120129	140835
Noccaea precox	3387912	3723	405	1103.803587	128861
Noccaea rotundifolium	3114292	6053	335	1014.656426	127776
Lesquerella montana	2056408	4967	260	669.9909935	154409
Nerisyrenia camporum	1962748	1487	466	639.4759613	154838
Stanleya pinnata	2861628	1659	264	932.3368645	129562
Thelypodium laciniatum	2720716	3819	279	886.4268258	153465
Ochthodium aegyptiacum	1695540	5429	278	552.417871	128625
Sisymbrium officinale	2414628	4781	253	786.7013806	128705
Smelowskia calycina	3414796	2689	319	1112.562568	111816
Thlaspi perfoliatum	2874836	356	11804	936.6401161	127979
Peltaria angustifolia	1610204	3516	323	524.6148516	129661
Biscutella laevigata	3122740	4543	210	1017.408839	155726
Biscutella prealpina	1625844	4863	212	529.7104645	136671
Calepina irregularis	1658292	1372	601	540.2822323	155009
Kernera saxatilis	1506884	10903	171	490.952529	128574
Lunaria annua	1173784	3222	291	382.4264	160455
Cleome spynosa	3045992	4758	529	992.4038452	158130
Cleome hirta	1116868	2360	1223	363.8828	162131
Alyssum dasycarpum	3088728	4210	226	1006.32751	127468
Draba aizoides	3284752	2786	338	1070.193394	127707
Turritis glabra	2953208	4352	367	962.1742194	154562
Cardamine pentaphyllos	3289460	4913	270	1071.72729	133255
Cardamine asarifolia	2551672	6145	244	831.3512	154399
Cardamine Trifolia	3127504	3429	303	1018.960981	150884
Cardamine pratensis	3416204	6146	221	1113.021303	126670
Aethionema saxatile	2977624	2595	519	970.1291097	157400
Arabidopsis halleri	3495144	3240	446	1138.740465	155607
Cardamine impatiens	As in Chapter 2 table 2-1				155611
Cardamine resedifolia					155036

Newly 80 assembled plastomes were submitted to DOGMA for annotation. Annotation information for 15 cp reference genomes were downloaded from NCBI. According to the annotation, 78 coding genes were found commonly shared in all the plastomes, 71 well-assembled genes were managed to be extracted from all the sequences, then, classified into 15 categories according to their functions (Table.4-2).

Table.4-2 List of protein coding genes extracted from 95 cp genomes by DOGMA

Gene Category	Genes
Photosystem I	*psaA, psaB, psaC, psaI, psaJ*
Photosystem II	*psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbJ, psbK, psbL, psbM, psbN, psbT, psbZ*
Cytochrome	*petA, *petB, *petD, petG, petL, petN*
ATP synthase	*atpA, atpB, atpE, *atpF, atpH, atpI*
Rubisco	*rbcL*
NADH dehydrogenase	**ndhA,* §**ndhB, ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI, ndhJ, ndhK*
Ribosomal protein (large subunit)	§**rpl2, rpl14, *rpl16, rpl20, rpl22,*§*rpl23, rpl33, rpl36*
Ribosomal protein (small subunit)	*rps2, rps3, rps4,* §*rps7, rps8, rps11,* §**rps12, rps14, rps15, rps18, rps19*
RNA polymerase	*rpoA, rpoB, *rpoC1, rpoC2*
ATP-dependent protease	**clpP*
Cytochrome c biogenesis	*ccsA*
Membrane protein	*cemA*
Maturase	*matK*
Conserved reading frames	*ycf4*
Pseudogenes	*accD*

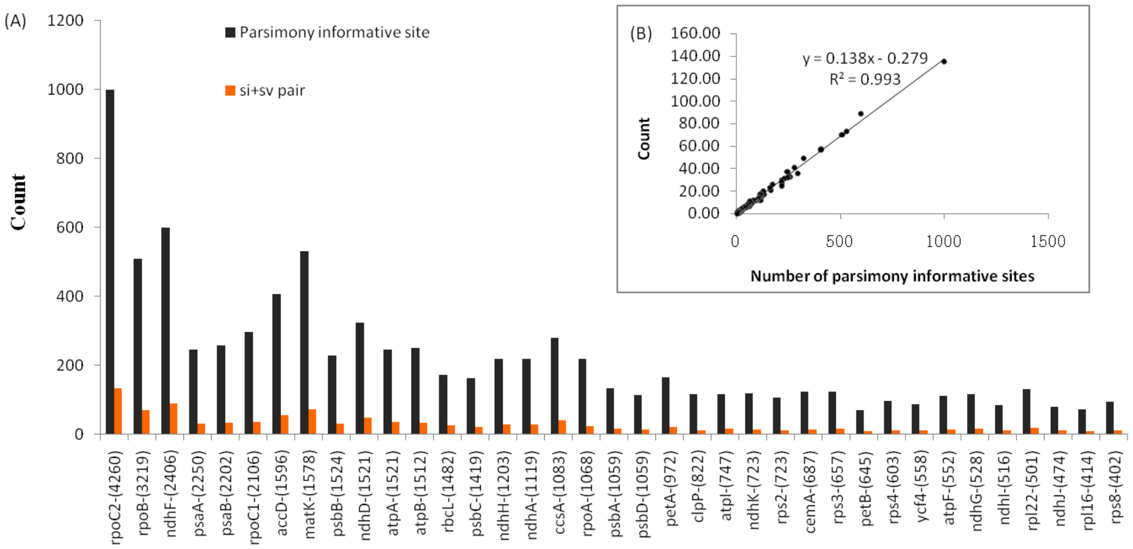
§Gene completely duplicated in the inverted repeat. *Gene with intron(s)

Phylogenetic analysis of 95 species in Brassicaceae

In addition to the 80 sampled species, the plastomes of other 15 species from the NCBI plastid database were collected, representing around 36 tribes from Brassicaceae. In total, 71 plastidic protein-coding genes out of the 95 species were chosen for the Phylogenetic analysis.

The 71 plastidic protein-coding genes varied from 87 (*petN*) to 4260 (*rpoC2*) aligned nucleotides in length and varied in the numbers of useful parsimony informative site. The correlation coefficient between the numbers of parsimony informative sites and the number of transitional and transversional pairs indicates a high correlation (R^2 = 0.99) (Table .4-3), thus suggesting that point mutations were the main source of information collected for phylogenetic inference.

Table .4-3 Degrees of variation of the phylogenetic utility of the plastidic protein-coding gene used in this study (38/71genes). Data are based on Parsimony informative site calculated among Brassicaceae taxa. (A) The number of calculated parsimony informative site, the transitional and transversional pair for different plastid protein-coding gene. (B) Scatterplot and regression line showing the relationship between parsimony informative site and point mutation.



The final aligned data matrix was 55710 bp long, of which 9348 bp were parsimony informative, 5621 bp were variable but parsimony-uninformative. The nucleotide evolution models generated by jModeltest for this concatenated sequence was : GTR + I + G, −lnL = 321102.5742. This substitution model was chosen for maximum likelihood and Bayesian inference.

Four phylogenetic trees were obtained individually by different methods. MP analysis of the cp-DNA phylogeny recovered a tree of length 35620, with a CI of 0.546 and RI of 0.678. ML analysis by PhyML produced a ML tree with -lnL = 309891.88787. Bootstrap support values from ML (MLB) were generally lower than BI posterior probability values (Fig .4-1).

Based on the same dataset of 71 concatenated protein-coding genes, topologies of the phylogenetic trees inferred by maximum likelihood (PhyML and CodonPhyML), maximum parsimony (PAUP) and Bayesian (MrBayes) analysis were highly consistent with each other (Fig .4-1). Bootstrap values and posterior probabilities both supported the three main lineages division in the new phylogeny, only a few clades were weakly supported, like between tribes Hesperideae and Buniadeae, Kernereae and Heliophileae.

As the out-group, the Cleomaceae family is sister to the Brassicaceae family, which was consistent with previous studies. Besides, three major lineages (I – III) were also delimited and defined. In our phylogeny, the lineage I consists of the tribes Smelowskieae, Descurainieae, Lepidieae, Cardamineae, Physarieae, Halimoloeae, Boechereae, Crucihmalayeae, Microlepidieae, Alyssopsideae, Erysimum, Turritideae, Camelineae. It included species from *Smelowskiacalycina* to *Capsella bursa-pastoris* in the tree (Fig . 4-1); Lineage II consists of the tribes Calepineae, Eutremeae, Arabideae, Alysseae, Coluteocarpeae, Thlaspideae, Iberideae, Cochlearieae, Heliophileae, Kernereae,Isatideae, Sisymbrieae, Thelyodieae, Brassiceae, which includes species from *Kernera saxatilis* to *Brassicanapus*; Lineage III consists of the tribes Anchonieae, Hesperideae, Buniadeae, that is from species *Matthiola fruticulosa to Bunias orientalis.*

According to the phylogenetic tree, the lineage I was monophyletic and was highly supported in the final tree. The tribes Aethionemeae, Biscutelleae, Anastaticeae, Coluteocarpeae, Alysseae, Arabideae, Thelypodieae, Brassiceae, Descurainieae, Lepidieae, Cardamineae, Physarieae, Boechereae, Microlepidieae, Erysimum and Camelineae were monophyletic in topologies generated from all methods while Thlaspideae were not monophyletic in the tree (Fig. 4-1).

The monophyly of other 16 tribes, like Anchonieae, Hesperideae, Buniadeae, Kernereae, Heliophileae, Cochlearieae, Iberideae, Eutremeae, Calepineae, Isatideae, Sisymbrieae, Smelowskieae, Halimolobeae, Crucihimalayeae, Turritideae and Alyssopsideae cannot be assessed due to insufficient sampling. Two other species, *Lunaria annua* and *Ochthodum aegyptiacum* were not assigned to any tribe yet.

Detection of positive selection sites

Selecton was used to detect positively selected sites under codon model, while PAML package was chosen to check whether there were signatures of positive selection at sites on specific branches. Codon model in Selecton allows the selection pressure to vary at different sites along the gene sequence, but not at different branches. In Selecton, M8 and M8a were both models allowing for variation of ω at different sites, but the M8a model is

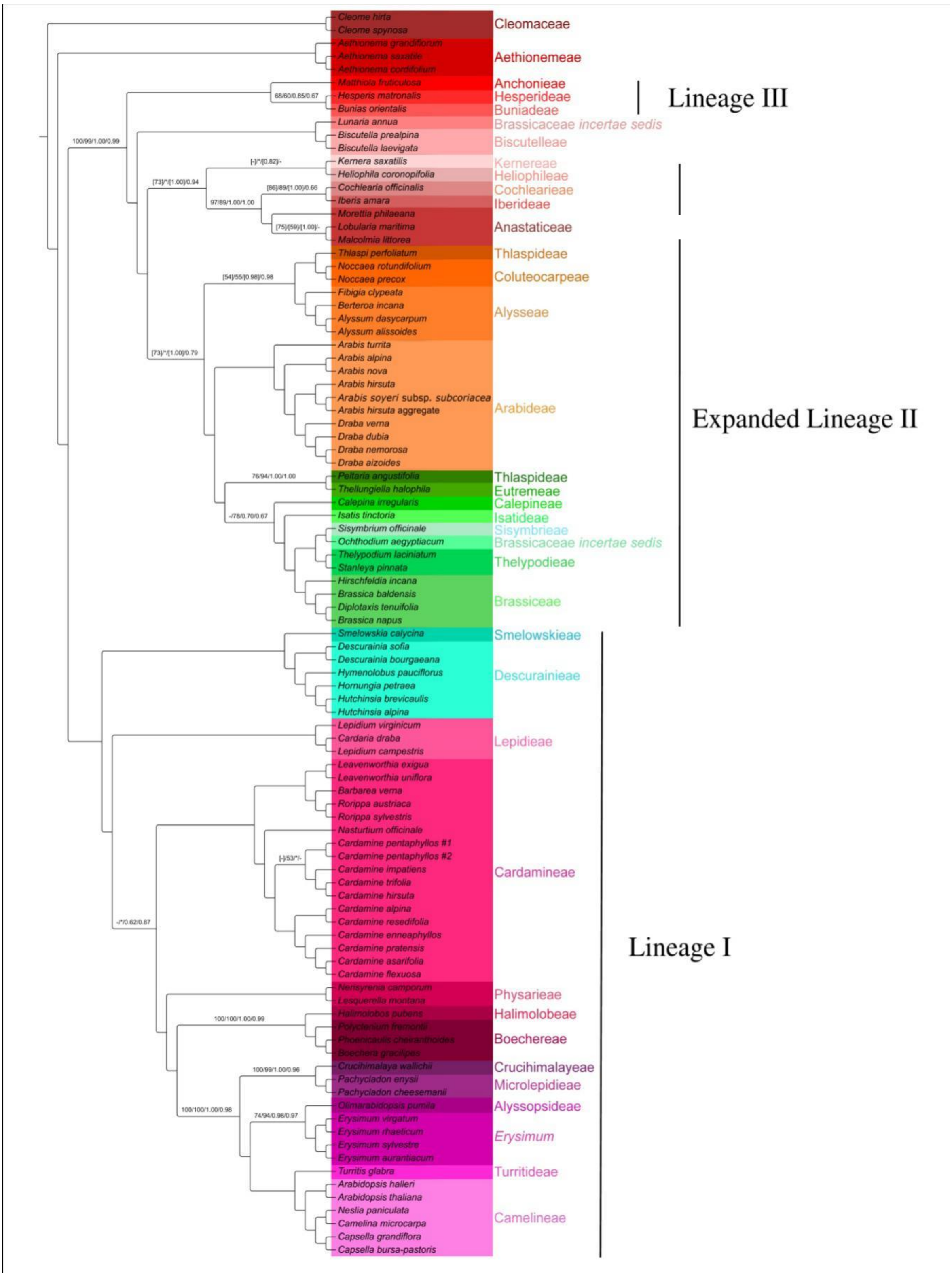


Fig .4-1 Phylogeny of Brassicaceae from the analysis of 71 concatenated plastidic coding genes. Topology based on codonphyml result. Bootstrap support values from the ML analysis, MP analysis, Bayesian posterior probabilities and aLRT(SH-like) were mapped onto branches. Bootstrap value estimates of 100 or PP or P estimates of 1.0 are omitted. Inconsistent indicated with an asterisk (*). No numbers are indicated above branches means "*/*/*/*"

modified on the basis of M8 model. In M8a, the additional category ω is set to 1. Comparison of lnL between M8 and M8a was used to verify the compatibility of the right model. The BEB (Bayes empirical Bayes) approach was used to check the confidence of the result.

Table.4-4 Genes identified with positive selection among 71 genes by Selecton

Gene	Positive selected position and amino acid
rpoC2	430Y 490L 507V 519V 531S 534P 535D 679R 713P 721V 750A 848P 857H 858M 876N 927A 928S 934T 951K 962L 964Q 978G 981C 1027V 1290L 1294T 1297F
accD	9L 75Q 76K 85V 106P 131H 134K 155Y 162I 183A 236R 302F
matK	2E 46A 49D 109L 110L 131L 179D 184S 238V 246S 265C 374S
rbcL	281A 326I 362I 445I 466R 472I 474K
ndhF	138E 151L 324T 401F 419C 429K 436S 449K 490A 507F 519T 562L
rpl20	72M 73E 80R
rpl2	17Y 34A 195C 230A
rps14	33K 55A
petD	135A 136V
rpoC1	84P 518R 522Q 524E 525R 605C
ccsA	26L 28L 47V 58F 119Q 162V 167Y 169K 172F 177V 179Y 183R 260G
rps4	27R 38S

According to the report of Selecton, 12 genes, *rpoC2*, *accD*, *matK*, *rbcL*, *ndhF*, *rpl20*, *rpl2*, *rps14*, *petD*, *rpoC1*, *ccsA*, *rps4* were detected with positive sites inside the gene sequence. Among them, *rpoC1* was the one with the most signatures of positive selection. According to the function and role they played in photosynthesis, these positively selected genes were mainly concentrated in the categories of “Transcription and Translation”, “Carbon assimilation and biosynthesis”, “Electron transport and ATP synthesis”.

Table 4-5 Sequence diversity of 12 positive selected genes along three different lineages

	Carbon assimilation and bosynthesis	Transcription and Translation	Electron transport and ATP synthesis
Lineage I	rbcL (0.013) accD (0.028)	matK (0.033) rpl2 (0.006) rpl20 (0.020) rpoC1 (0.014) rpoC2 (0.023) rps4 (0.013) rps14 (0.012)	petD (0.018) ccsA (0.028) ndhF (0.028)
Expanded Lineage II	rbcL (0.018) accD (0.033)	matK (0.056) rpl2 (0.007) rpl20 (0.030) rpoC1 (0.018) rpoC2 (0.035) rps4 (0.023) rps14 (0.019)	petD (0.022) ccsA (0.036) ndhF (0.045)
Lineage III	rbcL (0.018)accD (0.022)	matK (0.033) rpl2 (0.005) rpl20 (0.019) rpoC1 (0.013) rpoC2 (0.020) rps4 (0.012) rps14 (0.004)	petD (0.019) ccsA (0.026) ndhF (0.023)

Sequence diversity of these 12 genes along the lineages revealed that gene sequence diversity of lineage III was the lowest in most of the cases. The Expanded lineage II showed the highest level of sequence diversity for all the genes. Besides, *matK* showed the highest level of sequence diversity among the 12 genes along all the three lineages.

Based on the result of Selecton, the 12 genes with positive selection sites were further analyzed under Branch–site model with PAML package. According to the setting of PAML, the branch-site model was designed to detect positive selection, which influences only a few sites along a few lineages. Site models allow the ω ratio to vary among sites (among codons or amino acids for protein sequence) [142], [143] for calculating positive selection occurring at particular sites. Branch models allow the ω ratio to vary among branches in the phylogeny, which was designed to detect positive selection acting on particular lineages [143], [144]. Under the Branch-site model, each lineage was chosen as the foreground branch, the rest two as the background branches.

Model A and Model A-null were used independently for the analyses, the lnl for each was used for the chi-squared test to find the most compatible model.

The lnL values were collected as follows:

Table 4-6 lnL values for each lineage with different models

Foreground	Model	Genes chosen for selection analysis with Branch-site model											
		accD	ccsA	matK	ndhF	petD	rbcL	rpl2	rpl20	rpoC1	rpoC2	rps4	rps14
Lineage I	Model A	-9605.99	-7396.73	-15031.85	-12663.34	-2189.53	-5974.71	-2086.61	-2068.52	-9060.75	-27047.07	-2986.75	-1443.24
	Model A null	-9605.99	-7396.73	-15031.85	-12663.34	-2188.13	-5973.79	-2086.61	-2068.52	-9060.75	-27047.07	-2986.75	-1443.47
Lineage II	Model A	-9605.64	-7396.73	-15031.85	-12663.34	-2189.53	-5974.71	-2086.61	-2068.52	-9061.14	-27047.07	-2986.75	-1445.18
	Model A null	-9605.81	-7396.73	-15031.85	-12663.34	-2189.53	-5974.71	-2086.61	-2068.52	-9060.75	-27047.07	-2986.75	-1445.18
Lineage III	Model A	-9605.96	-7395.97	-15028.5	-12663.34	-2187.97	-5973.79	-2086.25	-2068.52	-9060.47	-27046.33	-2986.75	-1443.78
	Model A null	-9605.98	-7396.41	-15029.73	-12663.34	-2188.35	-5974.71	-2086.25	-2068.52	-9060.07	-27046.35	-2986.75	-1443.89

According to the lnL values, Model A did not show more fitness to our data compared to Model A-null. A further validation was carried out with the Chi-Square test. The P –values were more than 5% for all the analyses, indicating the two models have no significant difference. Therefore, the Model A -null should not be rejected.

Table 4-7 p-values for Chi-Squared-test

Foreground	Genes chosen for selection analysis with Branch-site model											
	accD	ccsA	matK	ndhF	petD	rbcL	rpl2	rpl20	rpoC1	rpoC2	rps4	rps14
Lineage I	1	1	1	1	0.09	0.17	1	1	1	1	1	0.501
Lineage II	0.56	1	1	1	1	1	1	1	0.38	1	1	1
Lineage III	0.84	0.35	0.11	1	0.38	0.17	1	1	0.18	0.84	1	0.64

Note: significance level ($p < 5\%$)

For calculating posterior probabilities for site classes, BEB methods were implemented for the modified branch-site model A. The posterior probabilities were collected and shown in table 4-7. However, no gene has a site on foreground branch, except *rps14*, which was detected under positive selection with high pp support. Even for *rps14*, the Model A and Model A Null used for branch-site analysis did not show a significant difference, which means the null hypothesis should not be rejected.

Table 4-8 Positive selected sites approved by Bayes Empirical Bayes (BEB) analysis

Foreground	Model	Genes chosen for selection analysis with Branch-site model											
		accD	ccsA	matK	ndhF	petD	rbcL	rpl2	rpl20	rpoC1	rpoC2	rps4	rps14
Lineage I	Model A	102S(0.700)											18E(0.982)
	Model A null	Not allowed											
Lineage II	Model A												
	Model A null	Not allowed											
Lineage III	Model A		56F(0.684), 257(0.678), 275V(0.69)	385E(0.941), 468N(0.722)	9L(0.501), 22C(0.534), 127V(0.516), 451L(0.501)	44N(0.716)		157I(0.772)		89G(0.615), 391N(0.679), 569S(0.664), 606V(0.674)	302F(0.652), 433R(0.653), 470V(0.650), 570A(0.657), 579K(0.647), 626K(0.605), 660R(0.531), 818A(0.657), 873S(0.6560, 888F(0.645), 952M(0.620)		18E(0.976)
	Model A null	Not allowed											

Amino acids with positively selected sites detected in the BEB analysis with posterior probabilities >95% are colored.

As a summary, in the 12 genes, no site of a specific branch among the analyzed species underwent positive selection.

4.4 Discussion

Sequencing and assembly of chloroplast genomes

The easier access and unique structure of the chloroplast genome made it popular in plant phylogenies at different levels. However, the complete cp genome is not always easy to obtain. In the past, it was mainly limited by the sequencing technology and cost, while nowadays, the Next-Generation Sequencing (NGS) technologies increased the data output and reduced the cost at the same time. This promoted the significant increase of plastid genome records in the NCBI database, especially in recent years. However, the number is still too small when compared to the number of plant species, as large-scale sequencing of chloroplast genome is still not common. The addition of the hundreds of records during the last years, in fact, was from many scattered works. The thirteen *Camellia* chloroplast genomes (eight complete cp genomes and five drafts) obtained by Huang and colleagues was a successful attempt of a systematic approach to plastome sequencing [114]. In 2015, 34 chloroplast genomes（22 completed）were assembled for elucidating the relationships between wild and domestic species within the genus *Citrus* [145], while an assembly of 47 (twelve completed) chloroplast genomes of apple were already done in 2013 [146]. However, even though the NGS technology is under high speed of development, no more than 100 chloroplast genomes of different species in one order were assembled at one time. The quality of assembled chloroplast genome mainly depends on two factors, one is the sequence reads, which should be of good quality and possess enough sequencing depth. Second is the assembly protocol, whose performance can significantly vary among different assemblers. Velvet [147], SPAdes [148], SOAPde novo2 [149], Edena [137] are all suitable assemblers for small genome assembly. In our study, results from four assemblers have been compared in the assembly process; Edena proved to have a better result as it was specially designed to focus on millions of short reads produced by Illumina sequencing platform, which is also supported by the performance comparison among different assemblers [150]. In addition to the above factors, the unique structure of the chloroplast genome could also explain the difficulty of successful assembly. The two inverted repeats always cannot be assembled completely at the same time. As during the mapping process, one IR region will attract most of the reads including the one corresponding to another IR area. In the meantime, the low conservation of intergenic regions makes the assembly producing inconsistent sequences even with a reference genome, which is not easy to validate but mainly depends on the assembly strategy. Our study was the first attempt that tries to assemble the chloroplast genomes for more than 80 species. Due to the imperfect coverage of the reads over the chloroplast genome, 14 out of 80 species were completed assembled; the rest contains gaps of varying sizes. Based on above results, we successfully extracted 71 commonly shared protein-coding genes among all the sampled and reference chloroplast genomes. This was the first time for collecting such a large cp DNA dataset in the Brassicaceae family.

Phylogenetic reconstruction and taxonomic implications

In recent years, phylogenetic research relies more and more on the genetic data, such as nDNA, mtDNA, plus cpDNA for plants. In our study, we make use of cp genome sequence, in light of the unique maternal inheritance and high copy number of the plastomes.

From a theoretical viewpoint, more data is likely to provide a better result when you are solving a complex phylogenetic problem. So making use of 100% of the required data is the ideal approach. However, in actual practice, only a minor part of the sequence information was used to infer the phylogenies of organisms, because in the real world, datasets containing the same sequences regions and taxa are few, and their combinations are difficult, which could be due to a restricted and unequal taxon and gene sampling.

Concatenating multiple alignments to produce a supermatrix is a very popular approach to making use of available data [151], [152]. In our study, not all cp genomes were completely assembled, and filling the gaps for the incomplete cp genome for phylogeny is not easy. In this situation, family wide phylogenomic study with whole chloroplast genome was no longer feasible. A supermatrix including most of the commonly shared protein-coding gene, which contained the representative information from all over the quadripartite organizations, turned out to be a good choice for this kind of studies.

The advantage of this method is that the resulting tree is created directly from the aligned sequences, without the necessity for any intermediate step, such as the combination of multiple single gene trees. The disadvantage of this method includes the inability to handle clearly missing data (gaps) leading to a situation where we do not know the impact of different levels of missing data on the results, though there are some attempts to estimate it [153]. Finally, the super-matrix approach requires much more computational power than normal approaches. Due to this critical limitation, sometimes one cannot obtain the tree from a concatenated long alignment.

Despite this, the supermatrix approach provides the only realistic possibility of using 100% of the information to reconstruct the tree of life, which could join the nDNA, mtDNA information, morphological data and any other forms of data.

Our study applied a supermatrix approach by concatenating 71 protein-coding gene sequences from 95 species. Compared to the carried out on six nuclear markers by Huang in 2015, which phylogeny was reconstructed at the level of major subdivisions of a relatively large family. Their result supported the results obtained by Beilstein in 2006 and 2008, but also defined new relationships [33]. Our results supported the three previously recognized lineages that discovered with chloroplast ndhF sequence [154]. Moreover, the relationships among the three main lineages were fully resolved (Fig .4-1) and large genera were delimited to corresponding lineages. For instance, the genus Draba (440 species) and Alyssum (207 species) belong to the Lineage II, while the Lineage I include the genera Erysmum (261 species), Lepidium (234 species), Cardamine (233 species). In addition to the approval of the main three lineages of the Brassicaceae, the relationships among genera and species were also clearly depicted for the first time.

However, cp genome data alone cannot reflect the real evolutionary history of the plant, as their inheritance is maternal inherited in Brassicaceae. Therefore, it cannot fully record the high frequency of hybridization from ancient to recent [155], [156], and also the polyploidization events can be missed by cp genome data [2]. Then, before considering to make taxonomic changes to the tribal structure of the family, the plastid phylogeny presented here should be validated with nuclear genes [72]. Besides, minor disagreements in topologies (Fig.4-1) among MP, ML and BI analyses also need further validation with data from other sources.

Huang and his colleagues applied the low-copy orthologous nuclear genes and reconstructed the phylogeny in Brassicaceae in 2015, a comprehensive phylogenetic tree was proposed by combining the results of multiple previous studies [33]. Transcriptome data was also applied in the molecular phylogeny of Brassicaceae species by Kagale in 2014 [157]. A comparison between our phylogeny and the summarized Brassicaceae phylogeny by Huang was shown as below, the topologies and evolutionary relationships were mostly consistent with each other.

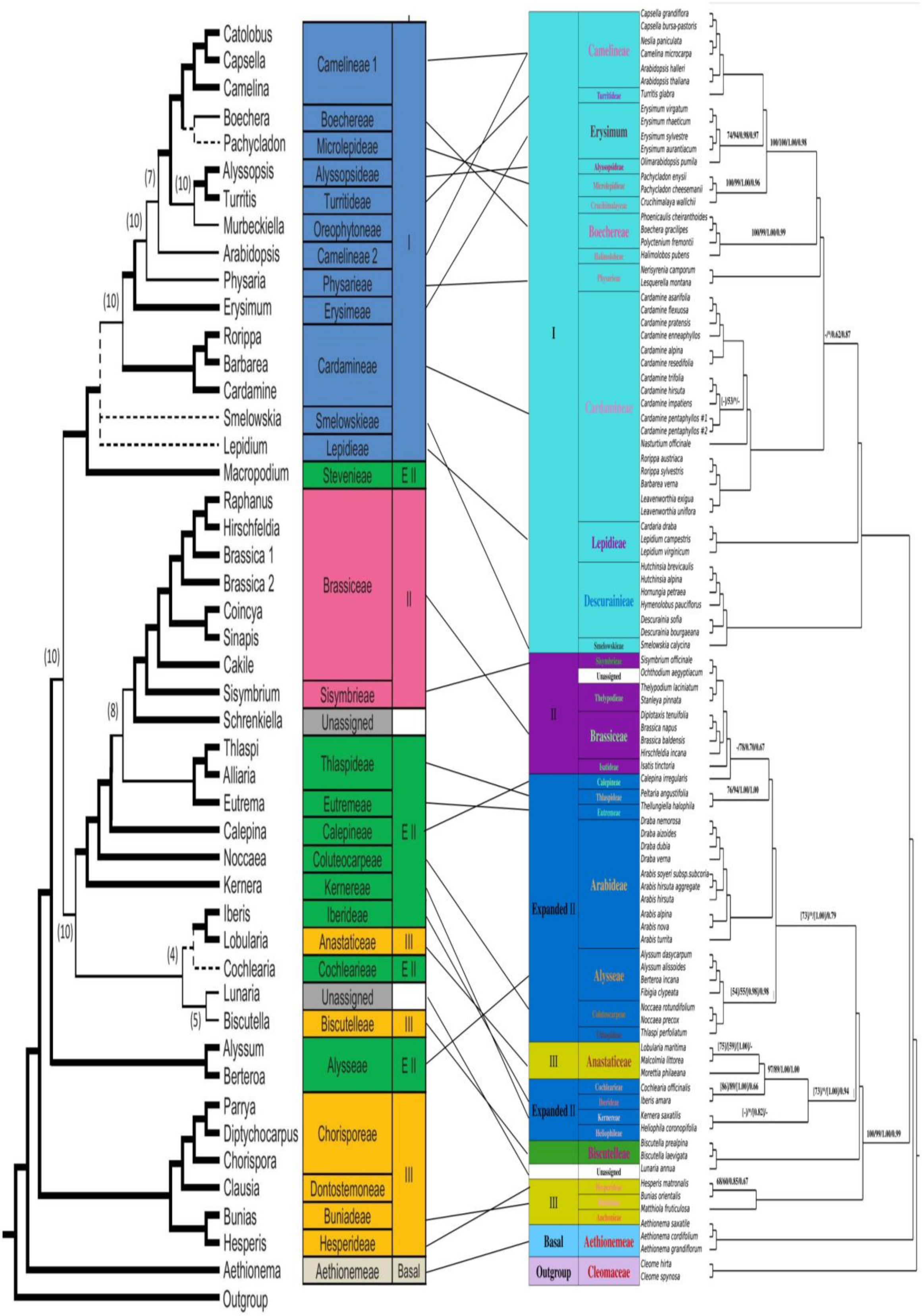


Fig .4-2 A topology comparison between our phylogeny (right) and the summarized Brassicaceae phylogeny (left) by Huang [33], labels on the branch in our phylogeny were the same as indicated in Fig.4-1.

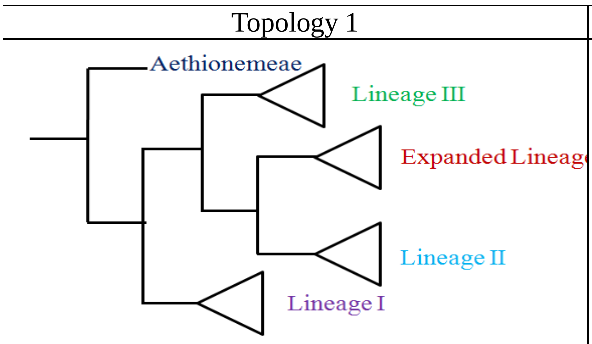
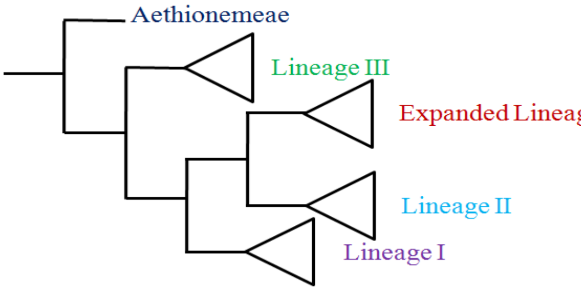
Taxonomy, Systematics and evolutionary history of the Brassicaceae is an ongoing study, has long been controversial because of the complex diversity of this family. In particular, the number of genera and tribes has been changing all the times.

In earlier phylogenetic studies based on non-molecular data, tribal classifications relied on a few morphological characters to delimit tribes, which often neglected the extensive homoplasy presented in data. Later a wealth of molecular phylogenetic studies focused on a few dozen genera to address certain taxonomic problems. Until 2006, Al-shehbaz synthesized the former studies and introduced the first comprehensive phylogenetic tribal classification of the family, 25 recognized monophyletic tribes were delimited. Later the family was revised to contain 49 monophyletic tribes and 20 unassigned genera in 2012 [9], [131]. In 2006, Beilstein proposed a “3 lineages” concept based on cp gene *ndhF*, which has been validated and adjusted many times by later studies [131]. The most recent revision increased the number to 51 tribes [32], with only 22 species and 15 genera were not yet assigned to any tribe.

Our results supported the four major infrafamiliar evolutionary lineages which had repeatedly been described (lineages I to III with sister group Aethionemeae; [8], [19], [22], [26], [28]). However, when compared to the synthesized phylogenetic tree in BrassiBase [158], there are still some differences about the location of some tribes. Two sampled species *Biscutella prealpina* and *Biscutella laevigata* in Biscutelleae, together with *Lunaria annua* became the sister group to Lineage II in the new phylogeny, but not consistent to the position from BrassiBase, which was one of the basal polytomies of Brassicaceae outside these three lineages. In our new phylogeny, *Lunaria annua* as one species in genus *Lunaria* had not been assigned to any tribe yet. We proved that it was closely related to the tribe Biscutelleae in the evolutionary history. Another formerly unassigned species *Ochthodium aegyptiacum* was found to group closely next to the tribe Sisymbrieae.

However, the most important discrepancy among the different trees lies in the evolutionary relationships among the three main lineages. Overviews [9] [159] [160] [2] [158] in the past several years, in fact, did not clearly delimit their relationships. Currently, there are two kinds of relationships available in the following simplified topologies. Studies which supported the topology were listed below (Table 4-9). It is worth noting that data resources used for inferring the phylogeny generally can be divided into two categories, corresponding to the two topologies. One category supported the topology 1, mainly included data from chloroplast sequence, ITS sequence, and morphological data. Except our study, most of them were collected in or before 2010. The second category of data mainly from nuclear genome and transcriptome, corresponding studies were carried out just in the past few years. Besides, our results are also consistent with the results of chloroplast phylogeny. Thus, we could speculate that the different sources of data may be the ultimate cause of the different phylogenetic inference of the three main lineages. During the divergence time of the three lineages, the chloroplast and nuclear genome may record different evolution histories.

Table 4-9 Simplifed topologies of phylogenetic relationships among the three main lineages and corresponding studies

Topology 1	Topology 2
 Aethionemeae Lineage III Expanded Lineag Lineage II Lineage I	 Aethionemeae Lineage III Expanded Lineag Lineage II Lineage I
(Our phylogeny tree) 71 chloroplast coding genes, Using parsimony, likelihood, and Bayesian methods,	[33], 113 Low-Copy Orthologous Nuclear Genes, Using likelihood, and Bayesian methods,
[19], ndhF and Trichomes Using parsimony, likelihood, and Bayesian methods,	[157], 213 orthologous genes in a concatenated alignment of 84,727 bp, Using likelihood methods,
[22], trnL intron and trnLF intergenic spacer sequences, Using parsimony methods,	[161], ITS region, Using likelihood methods,
[130], nuclear phytochrome A (PHYA) gene, Using parsimony, likelihood, and Bayesian methods,	
[21],Supermatrix analysis of data from adh 1, atpB, chalcone synthase, ITS, matK, ndhF, pistillata intron 1, rbcL, leafy, and trnL-F for 65 taxa, Using parsimony methods,	
[162], Nad4 Intron 1 Mitochondrial Marker Data, Using likelihood methods,	
[161], ITS region, Using parsmony methods,	
[163], nuclear ribosomal DNA sequences, Using Bayesian methods,	

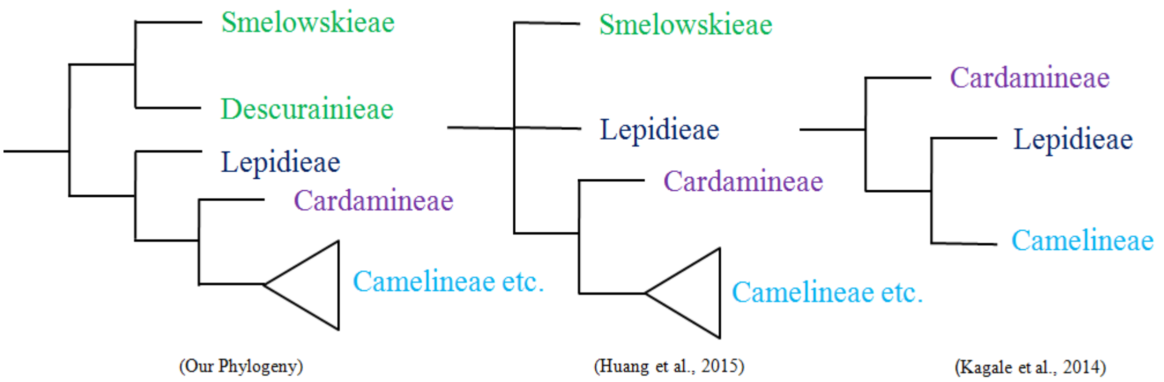
Calepina irregularis in tribe Calepineae was assigned as sister group to the branch consisting of Isatideae, Sisymbrieae, Thelypodieae, Brassiceae and the single unassigned species *Ochthodium aegypthiacum*. However, this tribe was only one branch in the expanded lineage II in the synthesized phylogenetic tree from BrassiBase, and the support of relationship between Calepineae and other tribes was relatively low.

The tribe Thlaspideae was problematic as it was not monophyletic. The two sampled species were clustered into one lineage, but *Thlaspi perfoliatum* was inferred to take a position close to the place that was taken by the tribe nonthalyspideae in the phylogenetic tree from BrassiBase, the same situation was observed in Huang's summarized phylogeny.

Besides, another difference is the tribe Anastaticeae, as it was surrounded by tribes belonging to expanded lineage II in the new phylogeny while the position for this tribe in BrassiBase was in Lineage III. But in Huang's summary, the tribe Anastaticeae was also located inside the range of expanded lineage II. The supported from both cp and nuclear data suggested a reclassification of the taxonomic position of this tribe.

Moreover, the phylogenetic position of Lepidieae was also uncertain with conflicts among all the recent phylogenetic studies in Brassicaceae. However, the separation of Lepidieae before Cardamineae from most other tribes was supported by our study, which was the same in Brassibase and the work of Huang, only different with work of Kagale [33], [157], [158] (Fig 4-3). These conflicts should be validated by further studies (with multiple types of DNA, morphological data, and other data) before a firm taxonomic decision regarding the circumscription.

Fig 4-3 Incongruent phylogenetic positions of tribe Lepidieae in three studies



Molecular evolution of chloroplast protein-coding genes

To determine whether any genes have undergone adaptive evolution in Brassicaceae plastomes in general and in the genus Cardamine in particular, we made the identification of genes putatively under positive selection using Selecton in chapter two and four, which will potentially improve our understanding of driving forces behind patterns of divergence and adaptation among the members of specific phylogenetic clades. The tests were carried out with two different data sets. The first was the one made from 18 species and used for analysis in chapter two. The second was collected from all the 95 species. From the two results, the same genes were detected under positive selection in the function categories of “Carbon assimilation and biosynthesis” and “Electron transport and ATP synthesis”, in another two categories, different genes were detected (Table 4-10).

Table 4-10 Chloroplast genes detected under positive selection in two independent tests [164]

	Carbon assimilation and biosynthesis	Transcription and Translation	Electron transport and ATP synthesis	Hypothetical chloroplast open reading frame
In 95 species	*rbcL, accD*	*matK, rpl2, rpl20, rpoC1, rpoC2, rps4, rps14,*	*petD, ccsA, ndhF,*	
In 18 species	*rbcL, accD*	*matK, rpl20, rpoC2, rpl14,*	*petD, ccsA, ndhF,*	*ycf1*

ycf1 gene was not included in the second data set because of the incomplete assembly among the 95 species, as it was detected as the most variable sequence in chloroplast genome [165]. However, it was approved as an essential gene in higher plants for cell survival[166], moreover, it was detected with positively selected sites in our former test. Thus, the *ycf1* can be considered of great potential and deserves further exploration in the whole family in the next step, on the basis of the complete assembly.

The other difference existed in the category of “Transcription and Translation”, *rpoC1*, *rpl2*, *rps4*, and *rps14* were four new genes detected under positive selection. *rpoC1*, namely DNA-directed RNA polymerase beta' subunit-1, has been reported to catalyze the transcription of DNA into RNA using the four ribonucleoside triphosphates as substrates. *rpl2* (ribosomal protein L2), *rps14* protein (Chloroplast-encoded ribosomal protein S4) and *rps4* protein (Chloroplast-encoded ribosomal protein S4) were the structural constituents of the ribosome. Together with *matK*, *rpl20*, *rpoC2*, and *rpl14,* these genes made the category of “Transcription and Translation” the most prominent in molecular evolution. As several genes involved in coding proteins for structural constituents of the ribosome, therefore, the ribosome could have played an important role in the adaptation to specific ecological habitats in the evolutionary history of the higher plant.

Besides, as described in chapter two, *rbcL*, *accD* were also detected with signatures of positive selection in the wider sampling range. The numbers of detected sites increased, but positions were largely consistent, indicating that some amino acids could play the crucial role in the adaptation process from a nature selection. However, this process is always not easy to be verified, additional effort will be needed to contribute on this issue.

Conclusion

This study assembled the chloroplast genomes of 80 species in Brassicaceae sequenced with NGS technology. Together with 15 reference chloroplast genomes in NCBI database, we carried out the phylogenetic reconstruction and molecular evolution analyses. The main conclusions obtained from above chapters were as follows:

1 Application of NGS technology in large-scale chloroplast genome sequencing in our study was feasible and efficient.

2 The chloroplast genome structures were highly conservative in Brassicaceae. With the global alignment of sequenced genomes, a conservative tetrad structure was found in all the species. Gene orders and numbers shown high co-linearity, but the type and distribution of repeats in each species was a unique feature of the single species.

3 Signatures of positive selection had been identified at different sites of 12 protein-coding genes at a family-wide scale. These positively selected genes were mainly concentrated in the categories of “Transcription and Translation”, “Carbon assimilation and biosynthesis”, “Electron transport and ATP synthesis”. Codon usage frequency in each Brassicaceae species varied slightly. No indication of lineage-specific events of positive selection was obtained.

4 The new phylogeny in Brassicaceae supported the three lineages division in Brassicaceae that was proposed by Beilstein, but the phylogenetic status of some tribes like Anastaticeae in our study still needs further validation, and the *Lepidum* clade was the one less supported as indicated by the low support value on its branch in the final phylogenetic tree.

5 Most of the tribes in the analysis were inferred to be monophyletic, only Thlaspideae was paraphyletic, but needs future validation.

Future prospect:

Chloroplast sequence is now widely applied in phylogenetic analysis. Our new study supported this application, and revealed that the complete chloroplast genome provided more information than all coding gene sequence when compared the tribe Cardamineae in chapter 3 and 4. Therefore, to obtain all the full chloroplast genome could be the next step, especially most of the 80 assembled chloroplast genomes contain few gaps, which should be solved by future lab work.

Besides, our final phylogenetic tree still has some conflicts with previous studies. In light of the late origin of Brassicaceae, also the wide hybridization and polyploidy, this means that an approach with only sequences from the chloroplast genome may not be suffcient to solve the complete phylogeny of the family, especially for most problematic clades. For a complete understanding of the evolutionary history of the Brassicaceae family, the combined analysis with the nuclear genome should be a promising approach.

The types and distribution of repeat sequences are diverse and unique in single species; this information can be applied into future species identification, and can be also integrated into further phylogenetic studies.

The functions of genes encoded by chloroplast genome are highly conserved in Brassicaceae. Positive selection pressure only happens on a minority genes. Whether the patterns of putatively positive selection observed in this study are significantly associated to the differential adaptation to various habits of different lineages or species remains an open question. The results obtained in this thesis lay a strong fundation for the future elucidation of this fundamental question concerning the biology and the evolution of Brassicaceae.

Acknowledgements

I would like to thank Dr. Claudio Varotto for his unwavering support, collegiality, and mentorship over the entire period of this project; Thanks for the support from the Ph.D program in Biomolecular Sciences of Trento University, which provides me great chances to communicate and learn; I wish to express my sincere thanks to “Fondazione Edmund Mach di San Michele all'Adige” for providing me with necessary facilities; Thanks for the financial support from the Chinese scholarship council (CSC), also the kindly help from all our lab members; I would like to extend my thanks to those who paved the path of phylogenetics and evolution study in Brassicaceae, also to experts and referees for their precise and accurate comments for prompting me to prepare a better thesis.

Bibliography:

[1] B. Bremer, K. Bremer, M. W. Chase, M. F. Fay, J. L. Reveal, L. H. Bailey, D. E. Soltis, P. S. Soltis, P. F. Stevens, and T. A. P. Group, “An update of the angiosperm phylogeny group classification for the orders and families of flowering plants: APG III,” *Botanical Journal of the Linnean Society*, vol. 161, no. 2, pp. 105–121, 2009.

[2] S. Renate and I. Bancroft, “Genetics and Genomics of the Brassicaceae,” *Plant Genetics and Genomics:Crops and Models*, vol. 9, 2011.

[3] M. a. Koch and C. Kiefer, “Molecules and migration: biogeographical studies in cruciferous plants,” *Plant Systematics and Evolution*, vol. 259, no. 2–4, pp. 121–142, Jun. 2006.

[4] I. A. Al-Shehbaz and K. Mummenhoff, “Transfer of the South African genera Brachycarpaea, Cycloptychis, Schlechteria, Silicularia, and Thlaspeocarpa to Heliophila (Brassicaceae),” *Novon*, vol. 15, no. 3, pp. 385–389, 2005.

[5] M. a. Koch, C. Kiefer, D. Ehrich, J. Vogel, C. Brochmann, and K. Mummenhoff, “Three times out of Asia Minor: The phylogeography of Arabis alpina L. (Brassicaceae),” *Molecular Ecology*, vol. 15, no. 3, pp. 825–839, 2006.

[6] M. a. Koch and C. Kiefer, “Molecules and migration: Biogeographical studies in cruciferous plants,” *Plant Systematics and Evolution*, vol. 259, no. 2–4, pp. 121–142, 2006.

[7] I. a. Al-Shehbaz and S. L. O’Kane, “Taxonomy and Phylogeny of Arabidopsis (Brassicaceae),” *The Arabidopsis Book*, vol. 6, no. 1, p. 1, 2002.

[8] A. Franzke, D. German, I. A. Al-Shehbaz, and K. Mummenhoff, “Arabidopsis family ties: molecular phylogeny and age estimates in Brassicaceae,” *Taxon*, vol. 58, no. 2, pp. 425–437, 2009.

[9] I. a. Al-Shehbaz, M. a. Beilstein, and E. a. Kellogg, “Systematics and phylogeny of the Brassicaceae (Cruciferae): An overview,” *Plant Systematics and Evolution*, vol. 259, no. 2–4, pp. 89–120, 2006.

[10] M. Koch, I. A. Al-shehbaz, and K. Mummenhoff, “Molecular Systematics, Evolution, and Population Biology in the Mustard Family (Brassicaceae),” *Annals of the Missouri Botanical Garden*, vol. 90, no. 2, pp. 151–171, 2003.

[11] P. F. Stevens, “Angiosperm Phylogeny Website,” *Version 12, July 2012*, 2012.

[Online]. Available: http://www.mobot.org/MOBOT/research/APweb/.

[12] S. Magallon, P. R. Crane, and P. S. Herendeen, “Phylogenetic Pattern, Diversity, and Diversification of Eudicots,” *Annals of the Missouri Botanical Garden*, vol. 86, no. 2, pp. 297–372, 1999.

[13] W. S. Judd, R. W. Sanders, and Mi. J. Donoghue, “Angiosperm family pairs: preliminary phylogenetic analyses,” *Harvard Papers in Botany*, vol. 1, no. 5, pp. 1–51, 1994.

[14] J. C. Hall, H. H. Iltis, and K. J. Sytsma, “Molecular phylogenetics of core brassicales, placement of orphan genera Emblingia, Forchhammeria, Tirania, and character evolution,” *Systematic Botany*, vol. 29, no. 3, pp. 654–669, 2004.

[15] J. C. Hall, K. J. Sytsma, and H. H. Iltis, “Phylogeny of Capparaceae and Brassicaceae based on chloroplast sequence data,” *American Journal of Botany*, vol. 89, no. 11, pp. 1826–1842, 2002.

[16] M. E. Schranz and T. Mitchell-Olds, “Independent ancient polyploidy events in the sister families Brassicaceae and Cleomaceae.,” *The Plant cell*, vol. 18, no. 5, pp. 1152–1165, 2006.

[17] M. A. Koch and K. Mummenhoff, “Editorial: Evolution and phylogeny of the Brassicaceae,” *Plant Systematics & Evolution*, vol. 259, no. 2–4, p. 81, 2006.

[18] I. a. Al-Shehbaz, M. a. Beilstein, and E. a. Kellogg, “Systematics and phylogeny of the Brassicaceae (Cruciferae): an overview,” *Plant Systematics and Evolution*, vol. 259, no. 2–4, pp. 89–120, Jun. 2006.

[19] M. a. Beilstein, I. a. Al-Shehbaz, and E. a. Kellogg, “Brassicaceae phylogeny and trichome evolution,” *American Journal of Botany*, vol. 93, no. 4, pp. 607–619, 2006.

[20] M. a. Beilstein, I. a. Al-Shehbaz, S. Mathews, and E. a. Kellogg, “Brassicaceae phylogeny inferred from phytochrome A and ndhF sequence data: Tribes and trichomes revisited,” *American Journal of Botany*, vol. 95, no. 10, pp. 1307–1327, 2008.

[21] C. D. Bailey, M. a Koch, M. Mayer, K. Mummenhoff, S. L. O’Kane, S. I. Warwick, M. D. Windham, and I. a Al-Shehbaz, “Toward a global phylogeny of the Brassicaceae.,” *Molecular biology and evolution*, vol. 23, no. 11, pp. 2142–60, Nov. 2006.

[22] M. A. Koch, C. Dobes, C. Kiefer, R. Schmickl, L. Klimes, and M. A. Lysak, “Supernetwork Identifies Multiple Events of Plastid trnF(GAA) Pseudogene Evolution in the Brassicaceae,” *Molecular Biology and Evolution*, vol. 24, no. 1, pp.

63–73, 2007.

[23] S. I. Warwick, C. a. Sauder, I. a. Al-Shehbaz, and F. Jacquemoud, “Phylogenetic Relationships in the Tribes Anchonieae, Chorisporeae, Euclidieae, and Hesperideae (Brassicaceae) Based on Nuclear Ribosomal Its Dna Sequences,” *Annals of the Missouri Botanical Garden*, vol. 94, no. 1, pp. 56–78, 2007.

[24] S. I. W. Ihsan A. Al-Shehbaz, “TWO NEW TRIBES (DONTOSTEMONEAE AND MALCOLMIEAE) IN THE BRASSICACEAE (CRUCIFERAE),” *Harvard Papers in Botany*, vol. 12, no. 2, pp. 429–433, 2007.

[25] I. A. A.-S. Dmitry A. German, “FIVE ADDITIONAL TRIBES (APHRAGMEAE, BISCUTELLEAE, CALEPINEAE, CONRINGIEAE, AND ERYSIMEAE) IN THE BRASSICACEAE (CRUCIFERAE),” *Harvard Papers in Botany*, vol. 13, no. 1, pp. 165–170, 2008.

[26] C. D. Bailey, M. a. Koch, M. Mayer, K. Mummenhoff, S. L. O&apos;Kane, S. I. Warwick, M. D. Windham, and I. a. Al-Shehbaz, “Toward a global phylogeny of the Brassicaceae,” *Molecular Biology and Evolution*, vol. 23, no. 11, pp. 2142–2160, 2006.

[27] S. I. Warwick, K. Mummenhoff, C. a. Sauder, M. a. Koch, and I. a. Al-Shehbaz, “Closing the gaps: Phylogenetic relationships in the Brassicaceae based on DNA sequence data of nuclear ribosomal ITS region,” *Plant Systematics and Evolution*, vol. 285, no. 3, pp. 1–24, 2010.

[28] T. L. P. Couvreur, A. Franzke, I. A. Al-shehbaz, F. T. Bakker, A. Koch, and K. Mummenhoff, “Molecular Phylogenetics , Temporal Diversification , and Principles of Evolution in the Mustard Family (Brassicaceae),” vol. 27, no. 1, pp. 55–71, 2010.

[29] A. Franzke, M. A. Lysak, I. A. Al-shehbaz, M. A. Koch, and K. Mummenhoff, “Cabbage family affairs : the evolutionary history of Brassicaceae,” *Trends in Plant Science*, vol. 16, no. 2, pp. 108–116, 2011.

[30] T. Arias and J. Chris Pires, “A fully resolved chloroplast phylogeny of the brassica crops and wild relatives (Brassicaceae: Brassiceae): Novel clades and potential taxonomic implications,” *Taxon*, vol. 61, no. 5, pp. 980–988, 2012.

[31] S. I. Warwick, a. Francis, and I. a. Al-Shehbaz, “Brassicaceae: Species checklist and database on CD-Rom,” *Plant Systematics and Evolution*, vol. 259, pp. 249–258, 2006.

[32] N. Hohmann, E. M. Wolf, M. a. Lysak, and M. a. Koch, “A Time-Calibrated Road Map of Brassicaceae Species Radiation and Evolutionary History,” *The Plant Cell*,

vol. 27, no. October, p. tpc.15.00482, 2015.

[33] C.-H. Huang, R. Sun, Y. Hu, L. Zeng, N. Zhang, L. Cai, Q. Zhang, M. A. Koch, I. Al-Shehbaz, P. P. Edger, J. C. Pires, D.-Y. Tan, Y. Zhong, and H. Ma, "Resolution of Brassicaceae Phylogeny Using Nuclear Genes Uncovers Nested Radiations and Supports Convergent Morphological Evolution.," *Molecular biology and evolution*, vol. 33, no. 2, p. msv226–, 2015.

[34] P. Pontarotti, *Evolutionary biology: Mechanisms and trends*, vol. 9783642304. 2013.

[35] M. Nei and S. Kumar, *Molecular Evolutionand Phylogenetics*, vol. 154. 2000.

[36] R. L. Small, R. C. Cronn, and J. F. Wendel, "Use of nuclear genes for phylogeny reconstruction in plants," *Australian Systematic Botany*, vol. 17, no. 2. pp. 145–170, 2004.

[37] K. H. Wolfe, W. H. Li, and P. M. Sharp, "Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 84, no. 24, pp. 9054–9058, 1987.

[38] J. D. Palmer and L. a Herbon, "Plant mitochondrial DNA evolves rapidly in structure, but slowly in sequence.," *Journal of molecular evolution*, vol. 28, pp. 87–97, 1988.

[39] D. M. Lonsdale, T. Brears, T. P. Hodge, S. E. Melville, and W. H. Rottmann, "The Plant Mitochondrial Genome: Homologous Recombination as a Mechanism for Generating Heterogeneity," *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, vol. 319, pp. 149–163, 1988.

[40] M. T. Clegg, B. S. Gaut, G. H. Learn, and B. R. Morton, "Rates and patterns of chloroplast DNA evolution.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 91, no. 15, pp. 6795–6801, 1994.

[41] R. K. Jansen, L. A. Raubeson, J. L. Boore, C. W. DePamphilis, T. W. Chumley, R. C. Haberle, S. K. Wyman, A. J. Alverson, R. Peery, S. J. Herman, H. M. Fourcade, J. V. Kuehl, J. R. McNeal, J. Leebens-Mack, and L. Cui, "Methods for obtaining and analyzing whole chloroplast genome sequences," *Methods in Enzymology*, vol. 395, pp. 348–384, 2005.

[42] T. Cavalier-Smith, "Chloroplast evolution: Secondary symbiogenesis and multiple losses," *Current Biology*, vol. 12, no. 2. 2002.

[43] B. Liu and J. Harada, *Advances in Plant Biology*. 2011.

[44] J. B. Campbell, N.A., & Reece, *Biology (8th Edition)*. 2008.

[45] P. J. Keeling, “Diversity and evolutionary history of plastids and their hosts,” *American Journal of Botany*, vol. 91, no. 10. pp. 1481–1493, 2004.

[46] W. Martin, T. Rujan, E. Richly, A. Hansen, S. Cornelsen, T. Lins, D. Leister, B. Stoebe, M. Hasegawa, and D. Penny, “Evolutionary analysis of Arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus.,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 19, pp. 12246–12251, 2002.

[47] R. S. Millen, R. G. Olmstead, K. L. Adams, J. D. Palmer, N. T. Lao, L. Heggie, T. A. Kavanagh, J. M. Hibberd, J. C. Gray, C. W. Morden, P. J. Calie, L. S. Jermiin, and K. H. Wolfe, “Many Parallel Losses of infA from Chloroplast DNA during Angiosperm Evolution with Multiple Independent Transfers to the Nucleus,” *The Plant cell*, vol. 13, no. 3, pp. 645–58, 2001.

[48] J. L. Corriveau and A. W. Coleman, “Rapid screening method to detect potential biparental inheritance of plastid DNA and results for over 200 angiosperm species,” *American Journal of Botany*, vol. 75, no. 10, pp. 1443–1458, 1988.

[49] Z. Cai, M. Guisinger, H. G. Kim, E. Ruck, J. C. Blazier, V. McMurtry, J. V. Kuehl, J. Boore, and R. K. Jansen, “Extensive reorganization of the plastid genome of Trifolium subterraneum (Fabaceae) is associated with numerous repeated sequences and novel DNA insertions,” *Journal of Molecular Evolution*, vol. 67, pp. 696–704, 2008.

[50] T. Wakasugi, J. Tsudzuki, S. Ito, K. Nakashima, T. Tsudzuki, and M. Sugiura, “Loss of all ndh genes as determined by sequencing the entire chloroplast genome of the black pine Pinus thunbergii.,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 91, no. 21, pp. 9794–9798, 1994.

[51] K. H. Wolfe, C. W. Morden, and J. D. Palmer, “Function and evolution of a minimal plastid genome from a nonphotosynthetic parasitic plant.,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 89, no. 22, pp. 10648–10652, 1992.

[52] M. J. Moore, P. S. Soltis, C. D. Bell, J. G. Burleigh, and D. E. Soltis, “Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots.,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 10, pp. 4623–4628, 2010.

[53] M. Parks, R. Cronn, and A. Liston, “Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes.,”

BMC biology, vol. 7, p. 84, 2009.

[54] A. G. Nazareno, M. Carlsen, and L. G. Lohmann, “Complete chloroplast genome of Tanaecium tetragonolobum: The first Bignoniaceae plastome,” *PLoS ONE,* vol. 10, no. 6, 2015.

[55] H.-Y. Kwon, J.-H. Kim, S.-H. Kim, J.-M. Park, and H. Lee, “The complete chloroplast genome sequence of Hibiscus syriacus.,” *Mitochondrial DNA*, pp. 1–2, 2015.

[56] A. V. Williams, J. T. Miller, I. Small, P. G. Nevill, and L. M. Boykin, “Integration of complete chloroplast genome sequences with small amplicon datasets improves phylogenetic resolution in Acacia,” *Molecular Phylogenetics and Evolution*, vol. 96, pp. 1–8, 2016.

[57] C. X. Chan and M. a Ragan, “Next-generation phylogenomics.,” *Biology direct*, vol. 8, p. 3, 2013.

[58] D. L. Swofford, “Phylogenetic Analysis Using Parsimony,” *Options*, vol. 42, pp. 294–307, 2003.

[59] R. Dean, “Phylogenetic Trees Made Easy. A How-to Manual, Second Edition,” *Economic Botany*, vol. 59. pp. 204–204, 2005.

[60] S. Guindon, J. F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, and O. Gascuel, “New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0,” *Systematic Biology*, vol. 59, no. 3, pp. 307–321, 2010.

[61] F. Ronquist, M. Teslenko, P. Van Der Mark, D. L. Ayres, A. Darling, S. Höhna, B. Larget, L. Liu, M. a. Suchard, and J. P. Huelsenbeck, “Mrbayes 3.2: Efficient bayesian phylogenetic inference and model choice across a large model space,” *Systematic Biology*, vol. 61, no. 3, pp. 539–542, 2012.

[62] M. Gil, M. S. Zanetti, S. Zoller, and M. Anisimova, “CodonPhyML: fast maximum likelihood phylogeny estimation under codon substitution models.,” *Molecular biology and evolution*, vol. 30, no. 6, pp. 1270–80, 2013.

[63] S. Kumar, K. Tamura, I. B. Jakobsen, and M. Nei, “MEGA2: molecular evolutionary genetics analysis software,” *Bioinformatics*, vol. 17, no. 12, pp. 1244–1245, 2001.

[64] J. E. McCormack, S. M. Hird, A. J. Zellmer, B. C. Carstens, and R. T. Brumfield, “Applications of next-generation sequencing to phylogeography and phylogenetics,” *Molecular Phylogenetics and Evolution*, vol. 66, no. 2, pp. 526–538, 2013.

[65] E. R. Mardis, “The impact of next-generation sequencing technology on genetics.,” *Trends in genetics : TIG*, vol. 24, no. 3, pp. 133–41, 2008.

[66] S. K. Wyman, R. K. Jansen, and J. L. Boore, “Automatic annotation of organellar genomes with DOGMA,” *Bioinformatics*, vol. 20, no. 17, pp. 3252–3255, 2004.

[67] C. Liu, L. Shi, Y. Zhu, H. Chen, J. Zhang, X. Lin, and X. Guan, “CpGAVAS, an integrated web server for the annotation, visualization, analysis, and GenBank submission of completely sequenced chloroplast genome sequences.,” *BMC genomics*, vol. 13, no. 1, p. 715, 2012.

[68] J. Felsenstein, “Inferring Phylogenies,” *American journal of human genetics*, vol. 74, no. 5, p. 1074, 2004.

[69] R. H. Thomas, *Molecular Evolution and Phylogenetics.*, vol. 86. 2001.

[70] M. Koch, B. Haubold, and T. Mitchell-Olds, “Molecular systematics of the brassicaceae: Evidence from coding plastidic matK and nuclear Chs sequences,” *American Journal of Botany*, vol. 88, no. 2, pp. 534–544, 2001.

[71] M. A. Koch, C. Dobeš, C. Kiefer, R. Schmickl, L. Klimeš, and M. A. Lysak, “Supernetwork identifies multiple events of plastid trnF(GAA) pseudogene evolution in the Brassicaceae,” *Molecular Biology and Evolution*, vol. 24, pp. 63–73, 2007.

[72] J. Wendel and J. Doyle, “Phylogenetic incongruence: window into genome history and molecular evolution,” *Molecular systematics of plants II*, p. 265–296, 1998.

[73] D. Penny, M. D. Hendy, and M. a Steel, “Progress with methods for constructing evolutionary trees.,” *Trends in ecology & evolution*, vol. 7, no. 3, pp. 73–79, 1992.

[74] J. Shi, Y. Zhang, H. Luo, and J. Tang, “Using jackknife to assess the quality of gene order phylogenies.,” *BMC bioinformatics*, vol. 11, p. 168, 2010.

[75] S. V Edwards, “Is a new and general theory of molecular systematics emerging?,” *Evolution; international journal of organic evolution*, vol. 63, no. 1, pp. 1–19, 2009.

[76] R. C. Thomson, I. J. Wang, and J. R. Johnson, “Genome-enabled development of DNA markers for ecology, evolution and conservation.,” *Molecular ecology*, vol. 19, no. 11, pp. 2184–95, 2010.

[77] S. V. Edwards, “A smörgåsbord of markers for avian ecology and evolution,” *Molecular Ecology*, vol. 17, pp. 945–946, 2008.

[78] R. Ekblom and J. Galindo, “Applications of next generation sequencing in molecular ecology of non-model organisms.,” *Heredity*, vol. 107, no. 1, pp. 1–15, 2010.

[79] H. R. L. Lerner, R. C. Fleischer, and S. Url, “Prospects for the Use of Next-Generation Sequencing Methods in Ornithology Special Reviews in Ornithology Prospects for the Use of Next-Generation Sequencing Methods in Ornithology,” vol. 127, no. 1, pp. 4–15, 2012.

[80] F. Delsuc, H. Brinkmann, and H. Philippe, “Phylogenomics and the reconstruction of the tree of life.,” *Nature reviews. Genetics*, vol. 6, pp. 361–375, 2005.

[81] L. Ometto, M. Li, L. Bresadola, and C. Varotto, “Rates of evolution in stress-related genes are associated with habitat preference in two Cardamine lineages.,” *BMC evolutionary biology*, vol. 12, no. 1, p. 7, Jan. 2012.

[82] T. Carlsen, W. Bleeker, H. Hurka, R. Elven, and C. Brochmann, “Biogeography and phylogeny of Cardamine (Brassicaceae),” *Ann. Missouri Bot. Gard*, vol. 96, no. 215–236, pp. 215–236, 2009.

[83] Surinder kumar Gupta, *Biology and Breeding of Crucifers*, vol. 1. 2009.

[84] T.-N. Ho and J. S. Pringle, “Gentianaceae [Flora of China],” in *Flora of China, vol. 16*, vol. 16, 1995, pp. 1–139.

[85] J. L. Bennetzen, “Comparative Sequence Analysis of Plant Nuclear Genomes: Microcolinearity and Its Many Exceptions,” *THE PLANT CELL ONLINE*, vol. 12, pp. 1021–1030, 2000.

[86] K. Marhold, J. Lihová, M. Perný, and W. Bleeker, “Comparative ITS and AFLP analysis of diploid Cardamine (Brassicaceae) taxa from closely related polyploid complexes,” *Annals of Botany*, vol. 93, pp. 507–520, 2004.

[87] B. Neuffer and P. Jahncke, “RAPD analyses of hybridization events inCardamine (Brassicaceae),” *Folia Geobotanica*, vol. 32, no. 1, pp. 57–67, 1997.

[88] A. Franzke and K. Mummenho, “Recent hybrid speciation in Cardamine (Brassicaceae) + conversion of nuclear ribosomal ITS sequences in statu nascendi,” *Theoretical and Applied Genetics (TAG)*, vol. 98, pp. 831–834, 1999.

[89] A. Franzke, K. Pollmann, W. Bleeker, R. Kohrt, and H. Hurka, “Molecular systematics ofCardamine and allied genera (Brassicaceae): Its and non-coding chloroplast DNA,” *Folia Geobotanica*, vol. 33, no. 3, pp. 225–240, 1998.

[90] J. Lihová, J. F. Aguilar, K. Marhold, and G. N. Feliner, “Origin of the disjunct tetraploid Cardamine amporitana (Brassicaceae) assessed with nuclear and chloroplast DNA sequence data,” *American Journal of Botany*, vol. 91, no. 8, pp. 1231–1242, 2004.

[91] M. a Ali, J. Lee, S. Y. Kim, and F. M. a Al-Hemaid, “Molecular phylogenetic study of Cardamine amaraeformis Nakai using nuclear and chloroplast DNA markers.,” *Genetics and molecular research : GMR*, vol. 11, no. 3, pp. 3086–90, 2012.

[92] D. E. McCauley, J. E. Stevens, P. A. Peroni, and J. A. Raveill, “The spatial distribution of chloroplast DNA and allozyme polymorphisms within a population of Silene alba (Caryophyllaceae),” *American Journal of Botany*, vol. 83, no. 6, pp. 727–731, 1996.

[93] R. L. Small, R. C. Cronn, and J. F. Wendel, “Use of nuclear genes for phylogeny reconstruction in plants,” *Australian Systematic Botany*, vol. 17, no. 2, pp. 145–170, 2004.

[94] R. K. Jansen, Z. Cai, L. a Raubeson, H. Daniell, C. W. Depamphilis, J. Leebens-Mack, K. F. Müller, M. Guisinger-Bellian, R. C. Haberle, A. K. Hansen, T. W. Chumley, S.-B. Lee, R. Peery, J. R. McNeal, J. V Kuehl, and J. L. Boore, “Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns.,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 49, pp. 19369–19374, 2007.

[95] T. Arias and J. C. Pires, “A fully resolved chloroplast phylogeny of the brassica crops and wild relatives (Brassicaceae : Brassiceae): Novel clades and potential taxonomic implications,” vol. 61, no. October, pp. 980–988, 2012.

[96] M. a Beilstein, I. a Al-Shehbaz, S. Mathews, and E. a Kellogg, “Brassicaceae phylogeny inferred from phytochrome A and ndhF sequence data: tribes and trichomes revisited.,” *American journal of botany*, vol. 95, no. 10, pp. 1307–27, Oct. 2008.

[97] J. Provan, W. Powell, and P. M. Hollingsworth, “Chloroplast microsatellites: new tools for studies in plant ecology and evolution.,” *Trends in ecology & evolution*, vol. 16, no. 3, pp. 142–147, 2001.

[98] S. Hu, G. Sablok, B. Wang, D. Qu, E. Barbaro, R. Viola, M. Li, and C. Varotto, “Plastome organization and evolution of chloroplast genes in Cardamine species adapted to contrasting habitats,” *BMC Genomics*, vol. 16, pp. 1–14, 2015.

[99] D. Blankenberg, A. Gordon, G. Von Kuster, N. Coraor, J. Taylor, A. Nekrutenko,

and G. Team, “Manipulation of FASTQ data with galaxy,” *Bioinformatics*, vol. 26, pp. 1783–1785, 2010.

[100] S. Andrews, “FastQC: A quality control tool for high throughput sequence data,” *babraham bioinformatics*, p. 1, 2010.

[101] D. R. Zerbino and E. Birney, “Velvet: algorithms for de novo short read assembly using de Bruijn graphs.,” *Genome research*, vol. 18, pp. 821–829, 2008.

[102] S. Kurtz, A. Phillippy, A. L. Delcher, M. Smoot, M. Shumway, C. Antonescu, and S. L. Salzberg, “Versatile and open software for comparing large genomes.,” *Genome biology*, vol. 5, p. R12, 2004.

[103] R. Li, Y. Li, K. Kristiansen, and J. Wang, “SOAP: Short oligonucleotide alignment program,” *Bioinformatics*, vol. 24, pp. 713–714, 2008.

[104] S. K. Wyman, R. K. Jansen, and J. L. Boore, “Automatic annotation of organellar genomes with DOGMA,” *Bioinformatics*, vol. 20, pp. 3252–3255, 2004.

[105] T. M. Lowe and S. R. Eddy, “tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence.,” *Nucleic acids research*, vol. 25, no. 5, pp. 955–64, 1997.

[106] M. Lohse, O. Drechsel, and R. Bock, “OrganellarGenomeDRAW (OGDRAW): A tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes,” *Current Genetics*, vol. 52, pp. 267–274, 2007.

[107] K. Katoh and D. M. Standley, “MAFFT multiple sequence alignment software version 7: improvements in performance and usability.,” *Molecular biology and evolution*, vol. 30, no. 4, pp. 772–80, 2013.

[108] K. A. Frazer, L. Pachter, A. Poliakov, E. M. Rubin, and I. Dubchak, “VISTA: computational tools for comparative genomics.,” *Nucleic Acids Research*, vol. 32, p. W273, 2004.

[109] S. Kurtz and C. Schleiermacher, “REPuter: fast computation of maximal repeats in complete genomes.,” *Bioinformatics (Oxford, England)*, vol. 15, no. 5, pp. 426–7, 1999.

[110] D. Posada, “jModelTest: Phylogenetic model averaging,” *Molecular Biology and Evolution*, vol. 25, pp. 1253–1256, 2008.

[111] D. L. Swofford, “PAUP*: phylogenetic analysis using parsimony, version 4.0b10,” *21 Libro*, p. 11pp, 2003.

[112] J. P. Huelsenbeck and F. Ronquist, “MRBAYES: Bayesian inference of phylogeny,” *Bioinformatics*, vol. 17, pp. 754–755, 2001.

[113] Y. Matsuoka, Y. Yamazaki, Y. Ogihara, and K. Tsunewaki, “Whole Chloroplast Genome Comparison of Rice, Maize, and Wheat: Implications for Chloroplast Gene Diversification and Phylogeny of Cereals,” *Molecular Biology and Evolution*, vol. 19, no. 12, pp. 2084–2091, 2002.

[114] H. Huang, C. Shi, Y. Liu, S.-Y. Mao, and L.-Z. Gao, “Thirteen Camellia chloroplast genome sequences determined by high-throughput sequencing: genome structure and phylogenetic relationships.,” *BMC evolutionary biology*, vol. 14, no. 1, p. 151, 2014.

[115] Q. Xu, G. Xiong, P. Li, F. He, Y. Huang, K. Wang, Z. Li, and J. Hua, “Analysis of complete nucleotide sequences of 12 Gossypium chloroplast genomes: origin and evolution of allotetraploids.,” *PloS one*, vol. 7, no. 8, p. e37128, 2012.

[116] J. I. Davis and R. J. Soreng, “Migration of endpoints of two genes relative to boundaries between regions of the plastid genome in the grass family (Poaceae).,” *American journal of botany*, vol. 97, no. 5, pp. 874–892, 2010.

[117] K.-J. Kim and H.-L. Lee, “Complete chloroplast genome sequences from Korean ginseng (Panax ginseng Nees) and comparative analysis of sequence evolution among 17 vascular plants.,” *DNA Research*, vol. 11, no. 4, pp. 247–61, 2004.

[118] C. Mayor, M. Brudno, J. R. Schwartz, a Poliakov, E. M. Rubin, K. a Frazer, L. S. Pachter, and I. Dubchak, “VISTA : visualizing global DNA sequence alignments of arbitrary length.,” *Bioinformatics (Oxford, England)*, vol. 16, no. 11, pp. 1046–1047, 2000.

[119] S. Kurtz, J. V Choudhuri, E. Ohlebusch, C. Schleiermacher, J. Stoye, and R. Giegerich, “REPuter: the manifold applications of repeat analysis on a genomic scale.,” *Nucleic acids research*, vol. 29, no. 22, pp. 4633–4642, 2001.

[120] C. S. Echt, L. L. DeVerno, M. Anzidei, and G. G. Vendramin, “Chloroplast microsatellites reveal population genetic diversity in red pine, Pinus resinosa Ait.,” *Molecular Ecology*, vol. 7, pp. 307–316, 1998.

[121] W. Powell, M. Morgante, C. Andre, J. W. McNicol, G. C. Machray, J. J. Doyle, S. V Tingey, and J. a Rafalski, “Hypervariable microsatellites provide a general source of polymorphic DNA markers for the chloroplast genome.,” *Current biology : CB*, vol. 5, no. 9, pp. 1023–1029, 1995.

[122] D.-Y. Kuang, H. Wu, Y.-L. Wang, L.-M. Gao, S.-Z. Zhang, and L. Lu, “Complete

chloroplast genome sequence of Magnolia kwangsiensis (Magnoliaceae): implication for DNA barcoding and population genetics.," *Genome / National Research Council Canada = Genome / Conseil national de recherches Canada*, vol. 54, pp. 663–673, 2011.

[123] Y.-J. Zhang, P.-F. Ma, and D.-Z. Li, "High-throughput sequencing of six bamboo chloroplast genomes: phylogenetic implications for temperate woody bamboos (Poaceae: Bambusoideae).," *PloS one*, vol. 6, no. 5, p. e20596, 2011.

[124] D.-K. Yi and K.-J. Kim, "Complete chloroplast genome sequences of important oilseed crop Sesamum indicum L.," *PloS one*, vol. 7, no. 5, p. e35872, 2012.

[125] D. H. Xu, J. Abe, J. Y. Gai, and Y. Shimamoto, "Diversity of chloroplast DNA SSRs in wild and cultivated soybeans: Evidence for multiple origins of cultivated soybean," *Theoretical and Applied Genetics*, vol. 105, pp. 645–653, 2002.

[126] A. J. Garris, T. H. Tai, J. Coburn, S. Kresovich, and S. McCouch, "Genetic structure and diversity in Oryza sativa L.," *Genetics*, vol. 169, no. 3, pp. 1631–8, 2005.

[127] M. J. Moore, C. D. Bell, P. S. Soltis, and D. E. Soltis, "Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 49, pp. 19363–19368, 2007.

[128] S. I. Warwick, a. Francis, and I. a. Al-Shehbaz, "Brassicaceae: Species checklist and database on CD-Rom," *Plant Systematics and Evolution*, vol. 259, no. 2–4, pp. 249–258, Jun. 2006.

[129] M. Koch, I. A. Al-Shehbaz, and K. Mummenhoff, "Molecular Systematics, Evolution, and Population Biology in the Mustard Family (Brassicaceae)," *Annals of the Missouri Botanical Garden*, vol. 90, pp. 151–171, 2003.

[130] M. A. Beilstein, I. A. Al-shehbaz, S. Mathews, and A. Elizabeth, "B RASSICACEAE PHYLOGENY INFERRED FROM PHYTOCHROME A AND NDH F SEQUENCE DATA : TRIBES AND TRICHOMES," vol. 95, no. 10, pp. 1307–1327, 2008.

[131] I. Al-Shehbaz, "A generic and tribal synopsis of the Brassicaceae (Cruciferae)," *Taxon*, vol. 61, pp. 931–954, 2012.

[132] M. A. Beilstein, I. A. Al-Shehbaz, and E. A. Kellogg, "Brassicaceae phylogeny and trichome evolution.," *American journal of botany*, vol. 93, pp. 607–619, 2006.

[133] M. a Koch and M. Matschinger, "Evolution and genetic differentiation among relatives of Arabidopsis thaliana.," *Proceedings of the National Academy of*

Sciences of the United States of America, vol. 104, no. 15, pp. 6272–7, Apr. 2007.

[134] A. Franzke, D. German, I. A. Al-shehbaz, and K. Mummenhoff, “Arabidopsis family ties : molecular phylogeny and age estimates in Brass icaceae,” vol. 58, no. May, pp. 425–437, 2009.

[135] S. I. Warwick, K. Mummenhoff, C. A. Sauder, M. A. Koch, and I. A. Al-Shehbaz, “Closing the gaps: Phylogenetic relationships in the Brassicaceae based on DNA sequence data of nuclear ribosomal ITS region,” *Plant Systematics and Evolution*, vol. 285, pp. 1–24, 2010.

[136] M. Koch, B. Haubold, and T. Mitchell-Olds, “Molecular systematics of the Brassicaceae: evidence from coding plastidic matK and nuclear Chs sequences.,” *American journal of botany*, vol. 88, pp. 534–544, 2001.

[137] D. Hernandez, P. François, L. Farinelli, M. Østerås, and J. Schrenzel, “De novo bacterial genome sequencing: Millions of very short reads assembled on a desktop computer,” *Genome Research*, vol. 18, no. 5, pp. 802–809, 2008.

[138] V. Ranwez, S. Harispe, F. Delsuc, and E. J. P. Douzery, “MACSE: Multiple alignment of coding SEquences accounting for frameshifts and stop codons,” *PLoS ONE*, vol. 6, 2011.

[139] Z. Yang and R. Nielsen, “Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages.,” *Molecular biology and evolution*, vol. 19, no. 6, pp. 908–917, 2002.

[140] J. Zhang, R. Nielsen, and Z. Yang, “Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level,” *Molecular Biology and Evolution*, vol. 22, no. 12, pp. 2472–2479, 2005.

[141] A. Stamatakis, “RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies.,” *Bioinformatics (Oxford, England)*, vol. 30, pp. 1312–3, 2014.

[142] Z. Yang, “Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus A.,” *Journal of molecular evolution*, vol. 51, no. 5, pp. 423–432, 2000.

[143] R. Nielsen and Z. Yang, “Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene,” *Genetics*, vol. 148, no. 3, pp. 929–936, 1998.

[144] Z. Yang, “Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution.,” *Molecular biology and evolution*, vol. 15, pp.

568–573, 1998.

[145] J. Carbonell-caballero, J. Terol, M. Talon, R. Alonso, V. Iba, and J. Dopazo, “A Phylogenetic Analysis of 34 Chloroplast Genomes Elucidates the Relationships between Wild and Domestic Species within the Genus Citrus,” vol. 32, no. 8, pp. 2015–2035, 2015.

[146] S. V. Nikiforova, D. Cavalieri, R. Velasco, and V. Goremykin, “Phylogenetic analysis of 47 chloroplast genomes clarifies the contribution of wild species to the domesticated apple maternal line,” *Molecular Biology and Evolution*, vol. 30, no. 8, pp. 1751–1760, 2013.

[147] D. R. Zerbino and E. Birney, “Velvet: algorithms for de novo short read assembly using de Bruijn graphs.,” *Genome research*, vol. 18, no. 5, pp. 821–9, 2008.

[148] A. Bankevich, S. Nurk, D. Antipov, A. a. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, S. I. Nikolenko, S. Pham, A. D. Prjibelski, A. V. Pyshkin, A. V. Sirotkin, N. Vyahhi, G. Tesler, M. a. Alekseyev, and P. a. Pevzner, “SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing,” *Journal of Computational Biology*, vol. 19, no. 5, pp. 455–477, 2012.

[149] R. Luo, B. Liu, Y. Xie, Z. Li, W. Huang, J. Yuan, G. He, Y. Chen, Q. Pan, Y. Liu, J. Tang, G. Wu, H. Zhang, Y. Shi, Y. Liu, C. Yu, B. Wang, Y. Lu, C. Han, D. W. Cheung, S.-M. Yiu, S. Peng, Z. Xiaoqian, G. Liu, X. Liao, Y. Li, H. Yang, J. Wang, T.-W. Lam, and J. Wang, “SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler.,” *GigaScience*, vol. 1, no. 1, p. 18, 2012.

[150] W. Zhang, J. Chen, Y. Yang, Y. Tang, J. Shang, and B. Shen, “A practical comparison of de novo genome assembly software tools for next-generation sequencing technologies.,” *PloS one*, vol. 6, no. 3, p. e17915, 2011.

[151] A. Rokas, B. L. Williams, N. King, and S. B. Carroll, “Genome-scale approaches to resolving incongruence in molecular phylogenies.,” *Nature*, vol. 425, no. 6960, pp. 798–804, 2003.

[152] F. D. Ciccarelli, T. Doerks, C. von Mering, C. J. Creevey, B. Snel, and P. Bork, “Toward automatic reconstruction of a highly resolved tree of life.,” *Science (New York, N.Y.)*, vol. 311, no. 5765, pp. 1283–1287, 2006.

[153] H. Philippe, E. a Snell, E. Bapteste, P. Lopez, P. W. H. Holland, and D. Casane, “Phylogenomics of eukaryotes: impact of missing data on large alignments.,” *Molecular biology and evolution*, vol. 21, no. 9, pp. 1740–52, 2004.

[154] M. a. Beilstein, I. a. Al-Shehbaz, and E. a. Kellogg, “Brassicaceae phylogeny and

trichome evolution," *American Journal of Botany*, vol. 93, no. 4, pp. 607–619, 2006.

[155] T. P. Hauser, R. B. Jorgensen, and H. Ostergard, "Fitness of backcross and F2 hybrids between weedy Brassica rapa and oilseed rape (B. napus)," *Heredity*, vol. 81, no. May, pp. 436–443, 1998.

[156] T. P. Hauser, R. B. Jorgensen, and H. Ostergard, "Preferential exclusion of hybrids in mixed pollinations between oilseed rape (Brassica napus) and weedy B. campestris (Brassicaceae)," *American Journal of Botany*, vol. 84, no. 6, pp. 756–762, 1997.

[157] S. Kagale, S. J. Robinson, J. Nixon, R. Xiao, T. Huebert, J. Condie, D. Kessler, W. E. Clarke, P. P. Edger, M. G. Links, A. G. Sharpe, and I. A. P. Parkin, "Polyploid evolution of the Brassicaceae during the Cenozoic era.," *The Plant cell*, vol. 26, no. 7, pp. 2777–91, 2014.

[158] M. Kiefer, R. Schmickl, D. a. German, T. Mand ákov á, M. A. Lysak, I. A. Al-Shehbaz, A. Franzke, K. Mummenhoff, A. Stamatakis, and M. A. Koch, "BrassiBase: Introduction to a novel knowledge database on brassicaceae evolution," *Plant and Cell Physiology*, vol. 55, no. 1, pp. 1–9, 2014.

[159] M. A. Beilstein, N. S. Nagalingum, M. D. Clements, S. R. Manchester, and S. Mathews, "Dated molecular phylogenies indicate a Miocene origin for Arabidopsis thaliana," 2010.

[160] I. A. Al-shehbaz, M. B. Garden, and S. Louis, "Brassicaceae (Mustard Family)," 2011.

[161] S. I. Warwick, K. Mummenhoff, C. a. Sauder, M. a. Koch, and I. a. Al-Shehbaz, "Closing the gaps: phylogenetic relationships in the Brassicaceae based on DNA sequence data of nuclear ribosomal ITS region," *Plant Systematics and Evolution*, vol. 285, no. 3–4, pp. 209–232, Apr. 2010.

[162] T. L. P. Couvreur, A. Franzke, I. A. Al-Shehbaz, F. T. Bakker, M. A. Koch, and K. Mummenhoff, "Molecular phylogenetics, temporal diversification, and principles of evolution in the mustard family (Brassicaceae)," *Molecular Biology and Evolution*, vol. 27, pp. 55–71, 2010.

[163] A. R. Khosravi, S. Mohsenzadeh, and K. Mummenhoff, "Phylogenetic relationships of Old World Brassicaceae from Iran based on nuclear ribosomal DNA sequences," *Biochemical Systematics and Ecology*, vol. 37, no. 2, pp. 106–115, 2009.

[164] J. F. Allen, W. B. M. de Paula, S. Puthiyaveetil, and J. Nield, "A structural phylogenetic map for chloroplast photosynthesis.," *Trends in plant science*, vol. 16, no. 12, pp. 645–655, Dec. 2011.

[165] W. Dong, C. Xu, C. Li, J. Sun, Y. Zuo, S. Shi, T. Cheng, J. Guo, and S. Zhou, “ycf1, the most promising plastid DNA barcode of land plants,” *Sci Rep,* vol. 5, p. 8348, 2015.

[166] A. Drescher, R. Stephanie, T. Calsa, H. Carrer, and R. Bock, “The two largest chloroplast genome-encoded open reading frames of higher plants are essential genes,” *Plant Journal*, vol. 22, pp. 97–104, 2000.

Appendix

Additional file 1: Table S1. Long-range PCR primers used for tiled plastome amplification.

Name of primer	Primer sequence: 5’ to 3’
CarPlastome_01F	GAGATCCAGAAACAGGTTCACGA
CarPlastome_01R	GTGTATGGACCAAATATAATTCTCTCA
CarPlastome_02F	CTAGACGCACTTAAAAGCCGAGT
CarPlastome_02R	GTAGCAGGAATCGAACCCGCATC
CarPlastome_03F	CATGTTCGGTTTTGAATTAGAGACG
CarPlastome_03R	CTCGTTTTTTATCAGATGCTTGTG
CarPlastome_04F	CCGTAGTGGACCAATTTGATAACAT
CarPlastome_04R	GATGTGGAGTTGTATTTGTTGATTCT
CarPlastome_05F	GTTTTATGTATCCCATTTGTTATCTTCG
CarPlastome_05R	GGACGATGCCCGAGCGGTTAATG
CarPlastome_06F	CGAGCAGGATTTGAACCAGCGTAG
CarPlastome_06R	GCGTAATAGTCCACCTACACGTCT
CarPlastome_07F	GTCTCGTTAGTTAGCTCTCGGTCT
CarPlastome_07R	CAGTAACGGATGTCGGCTCAATC
CarPlastome_08F	GGTATGTTCCCCATTACTTGTATG
CarPlastome_08R	CTACTATTGGATTTGAACCTATGACTC
CarPlastome_09F	CTCAGTGGTTAGAGTATTGCTTTCAT
CarPlastome_09R	GTTCAATTACTCTTTTACCCGCAA
CarPlastome_10F	GTGGAGTGACAGTTAGTTTTGGTATG
CarPlastome_10R	CGCTCTTAGTTCAGTTCGGTAG
CarPlastome_11F	CACGCTCTGTAGGATTTGAACC
CarPlastome_11R	CCCAATATACCCAATGCCAAATAGC
CarPlastome_12F	CAATGTGGAGACAAGGTATGTTCGT
CarPlastome_12R	GTTCAAGCAAGTTTCAACAATACCAT
CarPlastome_13F	GTCTACAACGATTATGTGGCATAGG
CarPlastome_13R	CGCAATGGAGCCGTAGACAGTCA
CarPlastome_14F	CCTCTGACATTACGACCTTTACCAC
CarPlastome_14R	CACTCGTTCATTATCAAACTGACTGC
CarPlastome_15F	GAGCACTTCTTATGGATTCGTTGAG
CarPlastome_15R	CACTGCTTATAGACCTGGTATTGGC
CarPlastome_16F	CTCCGACAGCATCTAGGGTTCC
CarPlastome_16R	CAACTCCCCGTAGCATTTCGTCG
CarPlastome_17F	GATACCAAGGCACCCAGAGACG
CarPlastome_17R	CGGCTCTTATACATGCTGCTACTA
CarPlastome_18F	CGAATGAATAATGAATCCAGATCCTA
CarPlastome_18R	CATATTTGCTGTGATGTTGATGAATG
CarPlastome_19F	GTTGACTATTACTTATTACATCTTGC
CarPlastome_19R	GAGTCTTACGATGAGTTTGAATGGG
CarPlastome_20F	GGATTCTGTCATTTCGCTAAGTCGT
CarPlastome_20R	GGATGTAAAGGATTGGAAACGTGAA
CarPlastome_21F	GATTCTGTTTCGGATAGTTGAACCC
CarPlastome_21R	GAACAACACCAATCCATCCCGAACTT
CarPlastome_22F	CGCAATGGAGCCGTAGACAGTCA
CarPlastome_22R	GCACTGAAAACCGTCATTACATTGG

Additional file 2: Table S2. Summary of distribution and localization of genes in the *C. resedifolia* and *C. impatiens* plastomes.

Region	**Gene**
LSC	*rps12_e2,trnH-GUG,psbA,trnK*-UUU,matK,rps16*,trnQ-UUG,psbK,psbI,trnS-GCU,trnG*-UCC,trnR-UCU,atpA,atpF*,atpH,atpI,rps2,rpoC2,rpoC1*,rpoB,trnC-GCA,petN,psbM,trnD-GUC,trnY-GUA,trnE-UUC,trnT-GGU,psbD,psbC,trnS-UGA,psbZ,trnG-UCC,trnfM-CAU,rps14,psaB,psaA,ycf3*,trnS-GGA,rps4,trnT-UGU,trnL*-UAA,trnF-GAA,ndhJ,ndhK,ndhC,trnV*-UAC,trnM-CAU,atpE,atpB,rbcL,accD,psaI,ycf4,cemA,petA,psbJ,psbL,psbF,psbE,petL,petG,trnW-CCA,trnP-UGG,psaJ,rpl33,rps18,rpl20,rps12_e1,clpP*,psbB,psbT,psbN,psbH,petB*,petD*,rpoA,rps11,rpl36,rps8,rpl14,rpl16*,rps3,rpl22*
SSC	*ndhF,rpl32,trnL-UAG,ccsA,ndhD,psaC,ndhE,ndhG,ndhI,ndhA*,ndhH,rps15*
IRA	*rpl2*,rpl23,trnI-CAU,ycf2,trnL-CAA,ndhB*,rps7,trnV-GAC,rrn16S,trnI*-GAU,trnA*-UGC,rrn23S,rrn4.5S,rrn5S,trnR-ACG,trnN-GUU*
IRB	*trnN-GUU,trnR-ACG,rrn5S,rrn4.5S,rrn23S,trnA*-UGC,trnI*-GAU,rrn16S,trnV-GAC,rps12_e2,rps7,ndhB*,trnL-CAA,ycf2,trnI-CAU,rpl23,rpl2**
LSC_IRA	*rps19*
IRA_SSC	*ycf1_short*
SSC_IRB	*ycf1_long*

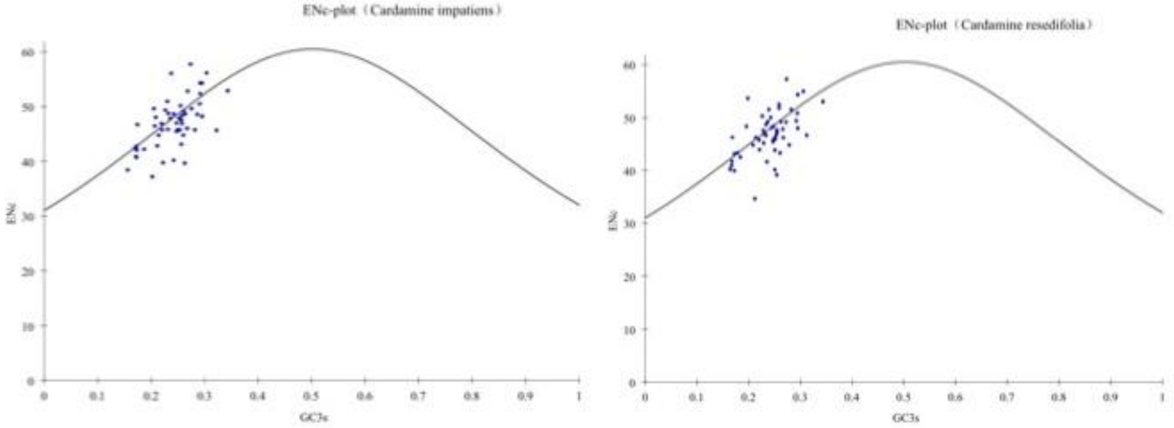
Additional file 3: Table S3.

Genes with introns in *C. resedifolia* ([a]) and *C. impatiens* ([b]) plastome and length of exons and introns.

Gene	**Location**	**Exon I (bp)**	**Intron I (bp)**	**Exon II (bp)**	**Intron II (bp)**	**Exon III (bp)**
atpF	LSC	410[a]/410[b]	679[a]/714[b]	145[a]/145[b]		
clpP	LSC	228[a]/228[b]	576a/573b	292[a]/292[b]	898[a]/897[b]	71[a]/71[b]
ndhA	SSC	530[a]/530[b]	1063a/1072b	553[a]/553[b]		
ndhB	IR	762[a]/762[b]	685a/685b	723[a]/723[b]		
petB	LSC	6[a]/6[b]	794a/794b	642[a]/642[b]		
petD	LSC	8[a]/8[b]	728a/710b	475[a]/475[b]		
rpl16	LSC	399[a]/399[b]	1090a/1110b	9[a]/9[b]		
rpl2	IR	435[a]/435[b]	682a/682b	390[a]/390[b]		
rpoC1	LSC	1611[a]/1611[b]	800a/794b	432[a]/432[b]		
*rps12**	LSC	114[a]/114[b]	-/-	26[a]/26[b]	537[a]/536[b]	232[a]/232[b]
rps16	LSC	227[a]/227[b]	872a/883b	40[a]/40[b]		
trnA-UGC	IR	38[a]/38[b]	800a/800b	35[a]/35[b]		
trnG-UCC	LSC	23[a]/23[b]	716a/716b	49[a]/49[b]		
trnI-GAU	IR	42[a]/42[b]	941a/941b	35[a]/35[b]		
trnK-UUU	LSC	35[a]/35[b]	2552a/2561b	37[a]/37[b]		
trnL-UAA	LSC	35[a]/35[b]	514a/499b	50[a]/50[b]		
trnV-UAC	LSC	35[a]/35[b]	606a/604b	39[a]/39[b]		
ycf3	LSC	153[a]/153[b]	789a/782b	228[a]/228[b]	703[a]/721[b]	126[a]/126[b]

**rps12* is a trans-spliced gene with the 5' end located in the LSC region and the duplicated 3' end in the IR regions

Additional file 4: Figure S1. Nc plot showing the distribution of the genes >300 bp in C. resedifolia and C. impatiens. The black line in the curve represents the standard effective number of codons (Nc) calculated using the equation N(c) = 2 + s + 29/(s(2) + (1-s)(2)), where s denotes GC3s (Wright [86] in Chapter 2).



Additional file 5: Table S4. Distribution and localization of repeat sequences in cpDNA of *C. impatiens* and *C. resedifolia.*

Size(bp)	**Start position**		**Type**	**Repeat sequence**	**Region**
73	38680	40904	F*	ctatacatatgacccgc[at]at[gt]aggaaaagaattgcgatagctaaatgatgatgtgc[ct]atatcggttaaccata	LSC; *psaB* gene, *psaA* gene
65	47939	48048	F*	aaatgatacttc[ga]gtaatggtcgacatagctt[ag][ga]ttgcagaggactgaaaatccttatgtcacca	LSC;spacer between *trn*L and *trnF*
62	47941	48139	F*	atgatacttc[ga]gtaatggt[ct]g[ag]catagcttagttgcagaggactgaaaatccttatgtcacc	LSC;spacer between *trn*L and *trnF*
59	47861	48221	F*	atgatacttc[ga]gtaatggtcggcatagctca[gc]ttggtagagcagaggact[gc]aaaatcct	LSC;spacer between *trn*L and *trnF*
54	47901	48084	F*	gcagaggactgaaaatcctt[ga]tgtcaccac[ac]tttagtaaaatgatacttcggta	LSC;spacer between *trn*L and *trnF*
54	47952	48150	F*	gtaatggt[ct]g[ag]catagcttagttgcagaggactgaaaatccttatgtcacc[at]tt	LSC;spacer between *trn*L and *trnF*
53	47877	48282	F*	gg[tc]cgg[cg]atagctcagttggtagagcagaggactgaaaatcct[tc]gtgtcacca	LSC;spacer between *trn*L and *trnF*
52	38701	40925	F*	aggaaaagaattgcgatagctaaatgatgatgtgc[ct]atatcggttaaccata	LSC; *psaB* gene, *psaA* gene
52	47952	48061	F*	gtaatggtcgacatagctt[ag][ga]ttgcagaggactgaaaatccttatgtcacca	LSC;spacer between *trnL* and *trnF*
48	47872	48232	F*	gtaatggtcggcatagctca[gc]ttggtagagcagaggact[gc]aaaatcct	LSC;spacer between *trnL* and *trnF*
45	89599	89623	F*	tttgtc[tc]aagt[ct]acttcgtttctttttgtccaagttacttc[gt]ttt	IRA; *ycf2*
45	150654	150678	F*	aaa[ac]gaagtaacttggacaaaaagaaacgaagt[ag]actt[ga]gacaaa	IRB; *ycf2*
45	114571	114571	P*	taaagatctttgatttactcat[at]atgagtaaatcaaagatcttta	SSC; spacer between *rpl32* and *trnL*
45	89599	150654	P*	tttgtc[tc]aagt[ct]acttcgtttctttttgtccaagttacttc[gt]ttt	IRA; *ycf2* IRB; *ycf2*
45	89623	150678	P*	tttgtc[ct]aagt[tc]acttcgtttctttttgtccaagttacttc[tg]ttt	IRA; *ycf2* IRB; *ycf3*
44	74746	74746	P	ttgacgtaatcagcctccaaatatttggaggctgattacgtcaa	LSC;spacer between *psbT* and *psbN*
42	47848	48005	F*	tattaaaatgataatgatacttcggtaatggtcg[ga]catagct	LSC;spacer between *trn*L and *trn*F
42	48240	48285	F*	cgg[cg]atagctca[cg]ttggtagagcagaggact[cg]aaaatcctcg	LSC;spacer between *trn*L and *trn*F(include part of trnF)
41	9424	9424	P*	tagcaattgtgtattgaa[tg]t[cg]a[ca]ttcaatacacaattgcta	LSC; spacer between *trn*G and *trn*R
40	47963	48161	F	catagcttagttgcagaggactgaaaatccttatgtcacc	LSC;spacer between *trn*L and *trn*F
40	48072	48161	F*	catagctt[ga][ag]ttgcagaggactgaaaatccttatgtcacc	LSC;spacer between *trn*L and *trn*F
40	28617	28617	P	gctagtatggtagaaagagatctctttctaccatactagc	LSC; spacer between *pet*N and *psb*M
39	43758	99346	F*	cagaaccgta[tc][ga]tgagattttca[tc]ctcatacggctcctc	LSC; ycf3 IRA;spacer between *rps*12 and *trn*V
39	43758	140937	P*	cagaaccgta[tc][ga]tgagattttca[tc]ctcatacggctcctc	LSC; ycf3 IRB; spacer between *trn*V and *rps*7
37	99349	121316	F*	aaccgtacatgag[ag]t[tc]ttc[ag]cctcatacggctcctcg	IRA;spacer between *rps*12 and *trn*V SSC; *ndh*A
37	121316	140936	P*	aaccgtacatgag[ga]t[ct]ttc[ga]cctcatacggctcctcg	SSC; *ndh*A IRB; spacer between *trn*V and *rps*7
36	47860	47940	F*	aatgatacttcggtaatggtcg[ga]catagct[ct]agttg	LSC;spacer between *trn*L and *trn*F
36	48138	48220	F*	gatgatacttcagtaatggt[tc]ggcatagct[tc]a[gc]ttg	LSC;spacer between *trn*L and *trn*F
35	47861	48139	F*	atgatacttc[ga]gtaatggt[ct]ggcatagct[ct]agttg	LSC;spacer between *trn*L and *trn*F

35	60707	60707	P*	aa[ga]aaaaaaagaaagaa[ta]ttctttcttttttt[tc]tt	LSC; spacer between *ycf*4 and *cem*A
33	47922	48105	F*	tgtcaccac[ac]tttagtaaaatgatacttcggta	LSC;spacer between *trn*L and *trn*F
33	108487	108519	F*	cat[at]gttcaactctttgacaaca[ct]gaaaaaacc	IRA;*rrn*5S
33	131770	131802	F*	ggttttttc[ag]tgttgtcaaagagttgaac[at]atg	IRB;*rrn*5S
33	48047	48136	F*	ta[ag]atgatacttcagtaatggt[ct]g[ag]catagctt	LSC;spacer between *trn*L and *trn*F
33	108487	131770	P*	cat[at]gttcaactctttgacaaca[ct]gaaaaaacc	IRA;spacer between *rrn*4.5S and *rrn*5S IRB;spacer between *rrn*4.5S and *rrn*5S IRB
33	108519	131802	P*	cat[ta]gttcaactctttgacaaca[tc]gaaaaaacc	IRA;spacer between *rrn*4.5S and *rrn*5S IRB;spacer between *rrn*4.5S and *rrn*5S IRB
32	47940	48017	F	aatgatacttcggtaatggtcgacatagctta	LSC;spacer between *trn*L and *trn*F
32	89645	89666	F*	tttttgtccaagttacttct[tc]tttttgtc[ct]aa	IRA;*ycf*2
32	150624	150645	F*	tt[ag]gacaaaaa[ga]agaagtaacttggacaaaaa	IRB;*ycf*2
32	89645	150624	P*	tttttgtccaagttacttct[tc]tttttgtc[ct]aa	IRA;*ycf*2 IRB;*ycf*2
32	89666	150645	P*	tttttgtccaagttacttct[ct]tttttgtc[tc]aa	IRA;*ycf*2 IRB;*ycf*2
32	129121	129121	P*	taaaaaaaaaa[ag]aggatcct[ct]tttttttttta	SSC;*ycf*1
31	47973	48082	F	ttgcagaggactgaaaatccttatgtcacca	LSC;spacer between *trn*L and *trn*F
31	48017	48049	F*	aatgatacttc[ga]gtaatggtcgacatagctt	LSC;spacer between *trn*L and *trn*F
31	48253	48298	F*	ttggtagagcagaggact[cg]aaaatcctcg[gt]g	LSC;spacer between *trn*L and *trn*F (include part of *trn*F)
31	118678	118703	F*	ta[ta]tatatatgcaaatttcaatctataat[at]t	SSC; spacer between *psa*C and *ndh*E
30	48082	48171	F	ttgcagaggactgaaaatccttatgtcacc	LSC;spacer between *trn*L and *trn*F
30	7816	44869	P*	a[gc]ggaaagagagggattcgaaccctcggta	LSC; *trn*S
30	35406	44807	P*	gcc[at]tcaaccactcggccatctctccga[ac]a	LSC; spacer between *pet*N and *psb*M

* imperfect repeat

F= Forward repeat

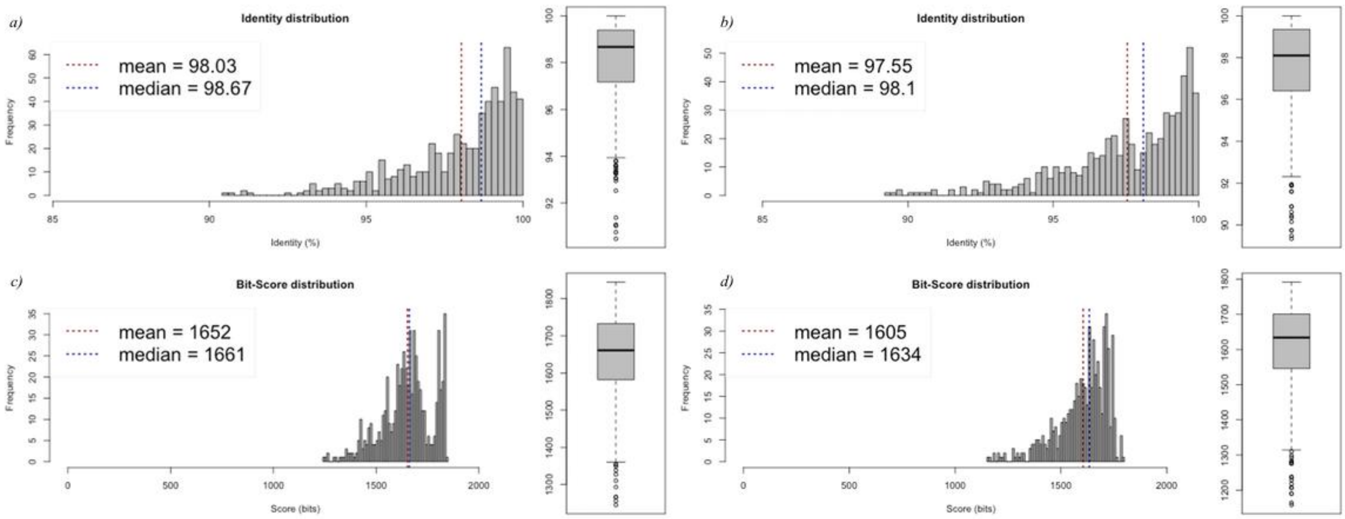
P= Perfect repeat

Additional file 6: Table S5. Cumulative SSR frequency and corresponding primer pairs in *C. resedifolia* and *C. impatiens.*

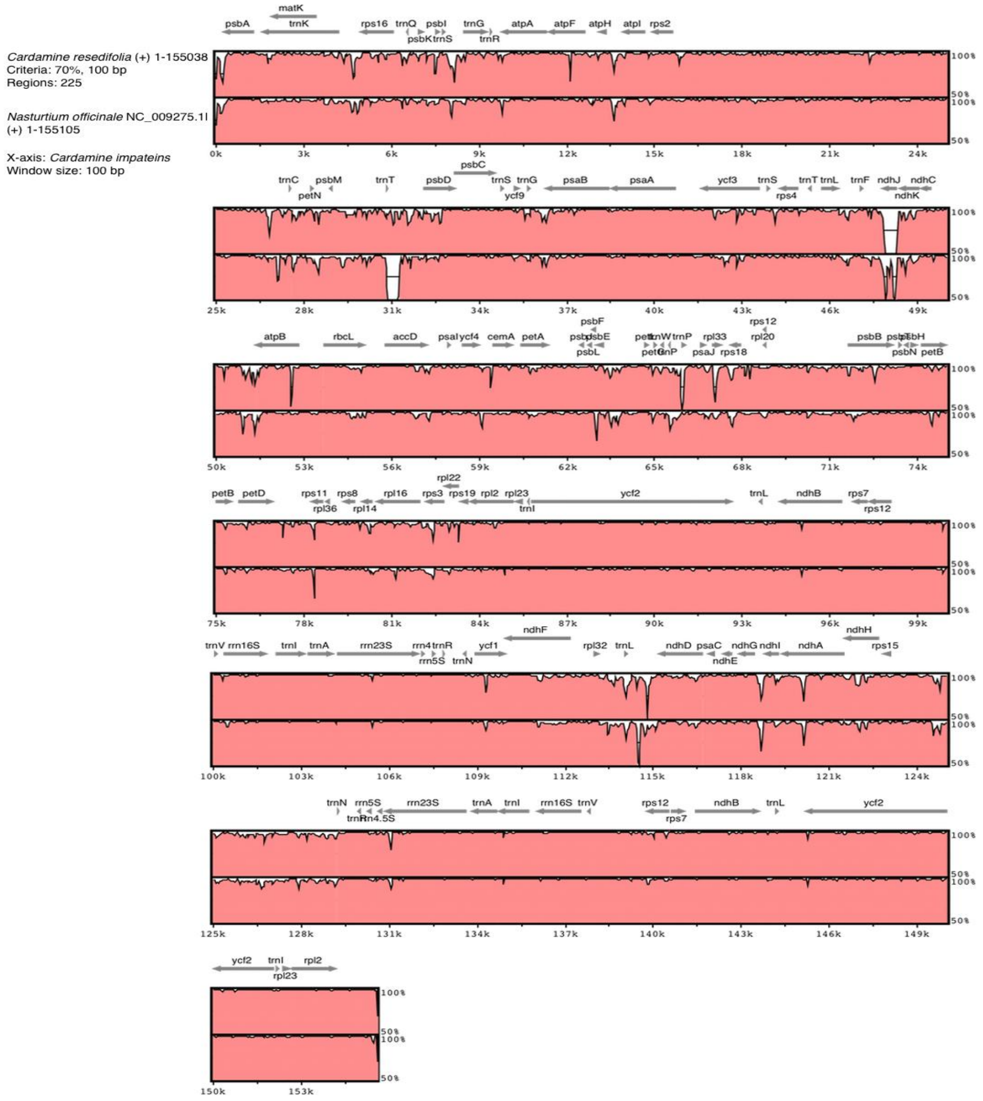
SSR type	*C. resedifolia*	*C. impatiens*
A/T	81	61
C/G	1	2
AC/GT	3	2
AG/CT	17	20
AT/AT	57	49
AAT/ATT	2	2
AAG/CTT	0	1
AAAC/GTTT	1	2
AAAT/ATTT	1	2
AAAG/CTTT	1	1
AGAT/ATCT	2	1
ATCC/ATGG	1	0
AATAG/ATTCT	2	0
AACTAT/AGTTAT	0	2

SSR search parameters: 1-10; 2-4; 3-4; 4-3; 5-3; 6-3 where 1, 2, 3, 4, 5, 6 indicate the mono- di-, tri-, tetra-, penta- and hexa- nucleotide repeats

Additional file 7: Figure S2. Average nucleotide identity plots of the C. resedifolia and C. impatiens against Nasturtium officinale.



Additional file 8: Figure S3. mVISTA plots showing genome-wise similarity between C. resedifolia, C. impatiens and N. officinale with rank probability of 70% and window size of 100 bp. The annotations displayed are derived from the C. impatiens plastome.



Additional file 9: Table S6. Phylogenetic distribution map of substitution rates using probabilistic substitution mapping under the homogenous model of sequence evolution.

Gene	**CarImp**	**CarRes**	**BarVer**	**NasOff**	**Tree**
accD	0	0.139186225625	0.312793762671	0.260268259266	(AetCor:0,AetGra:0.222231246624,((LobMar:0.525004297406,(BraNap:0.219964807142,(AraHir:0.436119963076,DraNem:0.258148369671):0):0):0,(LepVir:0.278151075658,(((OliPum:0.319263292652,(CapBur:0.166486766826,AraTha:0.170754874891):0):0,(CruWal:0,(PacEny:0,PacChe:0):0):0):0.384400208413,(BarVer:0.312793762671,(NasOff:0.260268259266,(CarImp:0,CarRes:0.139186225625):0):0):0):0):0):0.668151928242);
ccsA	0	0.229251725643	0	0	(AetCor:0.296128258205,AetGra:0.37571264642,((LobMar:0.44476366256,(BraNap:0.209095704856,(AraHir:0.543179015978,DraNem:0.545842887058):0.177874662313):0):0,(LepVir:0.196191185602,(((OliPum:0,(CapBur:0,AraTha:0):0):0,(CruWal:0.173741446395,(PacEny:0,PacChe:0):0):0):0,(BarVer:0,(NasOff:0,(CarImp:0,CarRes:0.229251725643):0):0):0):0):0):0.411323952553);
matK	0	0.569008541414	0.352331309841	0.358056113737	(AetCor:0.289302428541,AetGra:0.284115284778,((LobMar:0.156137497574,(BraNap:0.207885738502,(AraHir:0.30926275678,DraNem:0.29996154388):0.200312975129):0):0,(LepVir:0.40998161541,(((OliPum:0.262665790343,(CapBur:0.359563125374,AraTha:0.337151873191):0):0,(CruWal:0.414091208053,(PacEny:0,PacChe:0):0):0):0,(BarVer:0.352331309841,(NasOff:0.358056113737,(CarImp:0,CarRes:0.569008541414):0):0):0):0):0):0.467655516186);
ndhF	0.404600024572	0.601989001837	0.179846871295	0.108474269897	(AetCor:0.140357261759,AetGra:0.270671163546,((LobMar:0.253484542993,(BraNap:0.148485426492,(AraHir:0.322809193618,DraNem:0.178464035593):0.155514171655):0):0,(LepVir:0.0898268448695,(((OliPum:0.106566168676,(CapBur:0.208278689572,AraTha:0.117550045856):0):0,(CruWal:0.208019796345,(PacEny:0,PacChe:0):0):0):0.0702589158037,(BarVer:0.179846871295,(NasOff:0.108474269897,(CarImp:0.404600024572,CarRes:0.601989001837):0):0):0.0234788648028):0):0):0.331425842469);
rbcL	0	0	0	0	(AetCor:0,AetGra:0,((LobMar:0,(BraNap:0.109325439645,(AraHir:0,DraNem:0.0487025970445):0):0):0,(LepVir:0.0391270834061,(((OliPum:0.490826719256,(CapBur:0,AraTha:0):0):0,(CruWal:0,(PacEny:0,PacChe:0):0):0):0,(BarVer:0,(NasOff:0,(CarImp:0,CarRes:0):0):0):0):0):0):0.158657720875);
rpoC2	0.20507754271	0.497943540729	0.345025182223	0.22689666813	(AetCor:0.23215305481,AetGra:0.243588768793,((LobMar:0.149674990391,(BraNap:0.166950633148,(AraHir:0.317768034557,DraNem:0.171236049893):0.222929191281):0):0,(LepVir:0.180277693078,(((OliPum:0.356437807257,(CapBur:0.200350565576,AraTha:0.201212422548):0):0,(CruWal:0.261765236194,(PacEny:0.408130370771,PacChe:0.366150764634):0):0):0.217328386131,(BarVer:0.345025182223,(NasOff:0.22689666813,(CarImp:0.20507754271,CarRes:0.497943540729):0):0.0717810260585):0.143905199841):0):0.188905051823):0.31143993231);

Additional file 10: Table S7. Accessions and references for fully sequenced plastomes used in phylogenetic reconstruction and genome comparison in this study.

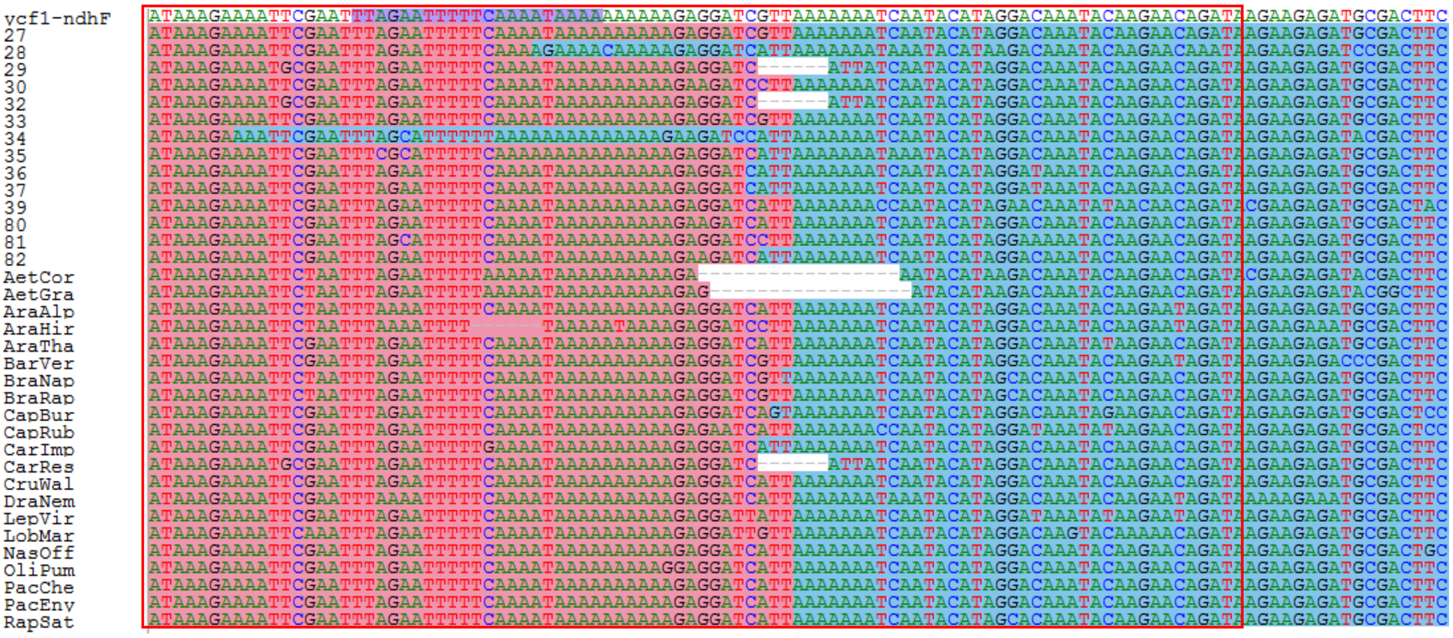
Taxon	Abbreviation	GenBank accession number	Reference
Aethionema cordifolium	AetCor	NC_009265	Hosouchi T., Tsuruoka H., Kotani H (2007) Sequencing analysis of Aethionema coridifolium chloroplast DNA.
Aethionema grandiflorum	AetGra	NC_009266	Hosouchi T., Tsuruoka H., Kotani H (2007) Sequencing analysis of Aethionema coridifolium chloroplast DNA.
Arabidopsis thaliana	AraTha	NC_000932	Sato S, Nakamura Y, Kaneko T, Asamizu E, Tabata S (1999) Complete structure of the chloroplast genome of Arabidopsis thaliana.DNA Res.29:283-90.
Arabis hirsuta	AraHir	NC_009268	Hosouchi T., Tsuruoka H., Kotani H (2007) Sequencing analysis of Aethionema coridifolium chloroplast DNA.
Barbarea verna	BarVer	NC_009269	Hosouchi T., Tsuruoka H., Kotani H (2007) Sequencing analysis of Aethionema coridifolium chloroplast DNA.
Brassica napus	BraNap	NC_016734	Zhi-Yong Hu, Wei Hua, Shun-Mou Huang, Han-Zhong Wang (2011) Complete chloroplast genome sequence of rapeseed (Brassica napus L.) and its evolutionary implications. Genetic Resources and Crop Evolution. 58: 875-887
Capsella bursa-pastoris	CapBur	NC_009270	Hosouchi T., Tsuruoka H., Kotani H (2007) Sequencing analysis of Aethionema coridifolium chloroplast DNA.
Crucihimalaya wallichii	CruWal	NC_009271	Hosouchi T., Tsuruoka H., Kotani H (2007) Sequencing analysis of Aethionema coridifolium chloroplast DNA.
Draba nemorosa	DraNem	NC_009272	Hosouchi T., Tsuruoka H., Kotani H (2007) Sequencing analysis of Aethionema coridifolium chloroplast DNA.
Lepidium virginicum	LepVir	NC_009273	Hosouchi T., Tsuruoka H., Kotani H (2007) Sequencing analysis of Aethionema coridifolium chloroplast DNA.
Lobularia maritima	LobMar	NC_009274	Hosouchi T., Tsuruoka H., Kotani H (2007) Sequencing analysis of Aethionema coridifolium chloroplast DNA.
Nasturtium officinale	NasOff	NC_009275	Hosouchi T., Tsuruoka H., Kotani H (2007) Sequencing analysis of Aethionema coridifolium chloroplast DNA.
Olimarabidopsis pumila	OliPum	NC_009267	Hosouchi T., Tsuruoka H., Kotani H (2007) Sequencing analysis of Aethionema coridifolium chloroplast DNA.
Pachycladon cheesemanii	PacChe	NC_021102	Becker M., Gruenheit N., Deusch O., Voelckel C., Lockhart P.J (2012) "Nunatak survival in the Central Southern Alps of New Zealand."
Pachycladon enysii	PacEny	NC_018565	Becker M., Gruenheit N., Deusch O., Voelckel C., Lockhart P.J (2012) "Nunatak survival in the Central Southern Alps of New Zealand."
Carica papaya	CarPap	NC_010323	Rice D.W., Saw J.J., Yu Q.Q., Feng Y.Y., Wang W.L., Wang L.L., Alam M.M., Palmer J.D (2008) The chloroplast and mitochondrial genomes of papaya. Genome Res. 0:0-0

Additional file 11 Figure C3-1.Boundry checking among LSC,SSC and IRs for 34 cp genome

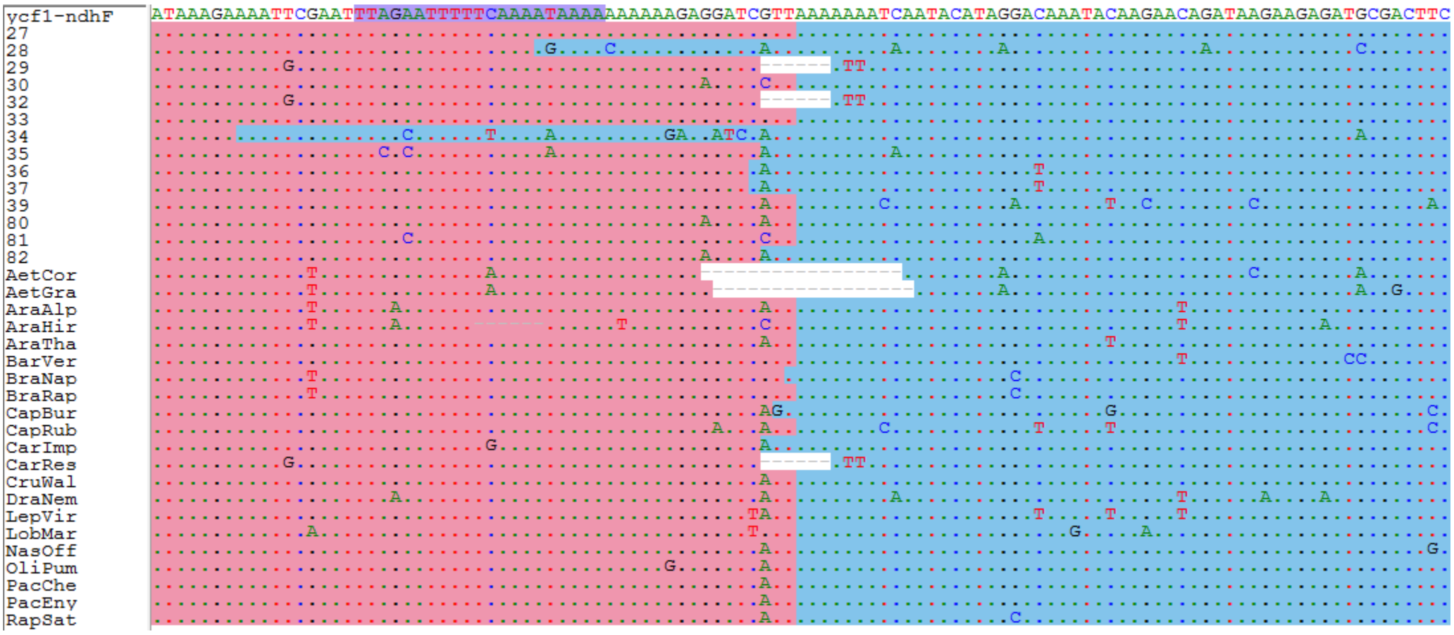
Number to corresponding species:

Species number	Species number
27	*Rorippa.sylvestris*
28	*Cardamine.hirsuta*
29	*Cardamine.alpina*
30	*Cardamine.flexuosa*
32	*Cardanime.resedifolia*
33	*Rorippa.austriaca*
34	*Cardamine.enneaphyllos*
35	*Cardamine.pentaphyllos*
36	*Leavenworthia.uniflora*
37	*Leavenworthia.exigua*
39	*Descurainia.bourgaeana*
80	*Cardamine.asarifolia*
81	*Cardamine.trifolia*
82	*Cardamine.pratensis*

1. IRa-SSC

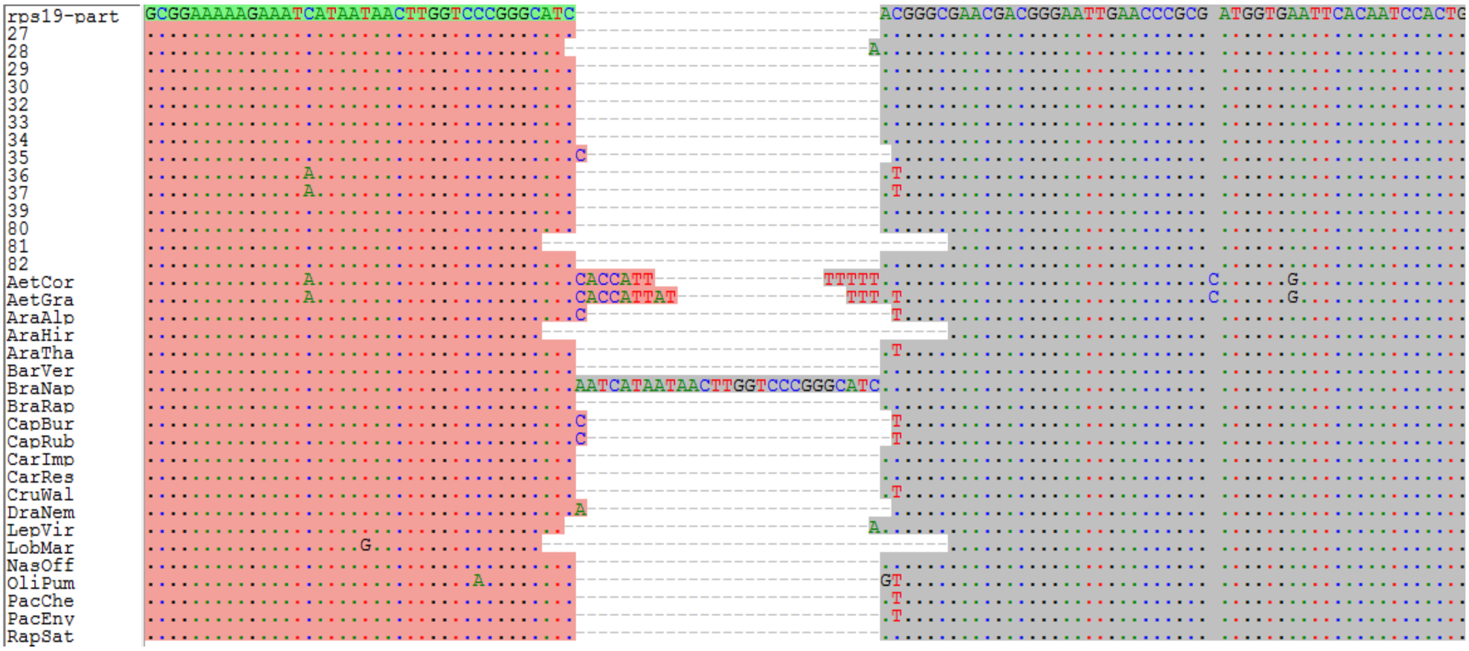


```
ycf1-ndhF ATAAAGAAAATTCGAATTTAGAATTTTTCAAAATAAAAAAAAAAGAGGATCGTTAAAAAAATCAATACATAGGACAAATACAAGAACAGATAAGAAGAGATGCGACTTC
27        ATAAAGAAAATTCGAATTTAGAATTTTTCAAAATAAAAAAAAAAGAGGATCGTTAAAAAAATCAATACATAGGACAAATACAAGAACAGATAAGAAGAGATGCGACTTC
28        ATAAAGAAAATTCGAATTTAGAATTTTTCAAAAGAAAACAAAAAGAGGATCATTAAAAAAATAAATACATAAGACAAATACAAGAACAAATAAGAAGAGATCCGACTTC
29        ATAAAGAAAATGCGAATTTAGAATTTTTCAAAATAAAAAAAAAAGAGGATC------ATTATCAATACATAGGACAAATACAAGAACAGATAAGAAGAGATGCGACTTC
30        ATAAAGAAAATTCGAATTTAGAATTTTTCAAAATAAAAAAAAAAGAAGATCCTTAAAAAAATCAATACATAGGACAAATACAAGAACAGATAAGAAGAGATGCGACTTC
32        ATAAAGAAAATGCGAATTTAGAATTTTTCAAAATAAAAAAAAAAGAGGATC------ATTATCAATACATAGGACAAATACAAGAACAGATAAGAAGAGATGCGACTTC
33        ATAAAGAAAATTCGAATTTAGAATTTTTCAAAATAAAAAAAAAAGAGGATCGTTAAAAAAATCAATACATAGGACAAATACAAGAACAGATAAGAAGAGATGCGACTTC
34        ATAAAGAAAATTCGAATTTAGCATTTTTTAAAAAAAAAAAAAAGAAGATCCATTAAAAAAATCAATACATAGGACAAATACAAGAACAGATAAGAAGAGATACGACTTC
35        ATAAAGAAAATTCGAATTTCGCATTTTTCAAAAAAAAAAAAAAAGAGGATCATTAAAAAAATAAATACATAGGACAAATACAAGAACAGATAAGAAGAGATGCGACTTC
36        ATAAAGAAAATTCGAATTTAGAATTTTTCAAAATAAAAAAAAAAGAGGATCATTAAAAAAATCAATACATAGGATAAATACAAGAACAGATAAGAAGAGATGCGACTTC
37        ATAAAGAAAATTCGAATTTAGAATTTTTCAAAATAAAAAAAAAAGAGGATCATTAAAAAAATCAATACATAGGATAAATACAAGAACAGATAAGAAGAGATGCGACTTC
39        ATAAAGAAAATTCGAATTTAGAATTTTTCAAAATAAAAAAAAAAGAGGATCATTAAAAAAACCAATACATAGAACAAATATAACAACAGATACGAAGAGATGCGACTAC
80        ATAAAGAAAATTCGAATTTAGAATTTTTCAAAATAAAAAAAAAAGAAGATCATTAAAAAAATCAATACATAGGACAAATACAAGAACAGATAAGAAGAGATGCGACTTC
81        ATAAAGAAAATTCGAATTTAGCATTTTTCAAAATAAAAAAAAAAGAGGATCCTTAAAAAAATCAATACATAGGAAAAATACAAGAACAGATAAGAAGAGATGCGACTTC
82        ATAAAGAAAATTCGAATTTAGAATTTTTCAAAATAAAAAAAAAAGAAGATCATTAAAAAAATCAATACATAGGACAAATACAAGAACAGATAAGAAGAGATGCGACTTC
AetCor    ATAAAGAAAATTCTAATTTAGAATTTTTAAAAATAAAAAAAAAAGA-----------------AATACATAAGACAAATACAAGAACAGATACGAAGAGATACGACTTC
AetGra    ATAAAGAAAATTCTAATTTAGAATTTTTAAAAATAAAAAAAAAAGAG-----------------ATACATAAGACAAATACAAGAACAGATAAGAAGAGATACGGCTTC
AraAlp    ATAAAGAAAATTCTAATTTAAAATTTTTCAAAATAAAAAAAAAAGAGGATCATTAAAAAAATCAATACATAGGACAAATACAAGAATAGATAAGAAGAGATGCGACTTC
AraHir    ATAAAGAAAATTCTAATTTAAAATTTT------TAAAAATAAAAGAGGATCCTTAAAAAAATCAATACATAGGACAAATACAAGAATAGATAAGAAGAAATGCGACTTC
AraTha    ATAAAGAAAATTCGAATTTAGAATTTTTCAAAATAAAAAAAAAAGAGGATCATTAAAAAAATCAATACATAGGACAAATATAAGAACAGATAAGAAGAGATGCGACTTC
BarVer    ATAAAGAAAATTCGAATTTAGAATTTTTCAAAATAAAAAAAAAAGAGGATCGTTAAAAAAATCAATACATAGGACAAATACAAGAATAGATAAGAAGAGACCCGACTTC
BraNap    ATAAAGAAAATTCTAATTTAGAATTTTTCAAAATAAAAAAAAAAGAGGATCGTTAAAAAAATCAATACATAGCACAAATACAAGAACAGATAAGAAGAGATGCGACTTC
BraRap    ATAAAGAAAATTCTAATTTAGAATTTTTCAAAATAAAAAAAAAAGAGGATCGTTAAAAAAATCAATACATAGCACAAATACAAGAACAGATAAGAAGAGATGCGACTTC
CapBur    ATAAAGAAAATTCGAATTTAGAATTTTTCAAAATAAAAAAAAAAGAGGATCAGTAAAAAAATCAATACATAGGACAAATAGAAGAACAGATAAGAAGAGATGCGACTCC
CapRub    ATAAAGAAAATTCGAATTTAGAATTTTTCAAAATAAAAAAAAAAGAGAATCATTAAAAAAACCAATACATAGGATAAATATAAGAACAGATAAGAAGAGATGCGACTCC
CarImp    ATAAAGAAAATTCGAATTTAGAATTTTTGAAAATAAAAAAAAAAGAGGATCATTAAAAAAATCAATACATAGGACAAATACAAGAACAGATAAGAAGAGATGCGACTTC
CarRes    ATAAAGAAAATGCGAATTTAGAATTTTTCAAAATAAAAAAAAAAGAGGATC------ATTATCAATACATAGGACAAATACAAGAACAGATAAGAAGAGATGCGACTTC
CruWal    ATAAAGAAAATTCGAATTTAGAATTTTTCAAAATAAAAAAAAAAGAGGATCATTAAAAAAATCAATACATAGGACAAATACAAGAACAGATAAGAAGAGATGCGACTTC
DraNem    ATAAAGAAAATTCGAATTTAAAATTTTTCAAAATAAAAAAAAAAGAGGATCATTAAAAAAATAAATACATAGGACAAATACAAGAATAGATAAAAAGAAATGCGACTTC
LepVir    ATAAAGAAAATTCGAATTTAGAATTTTTCAAAATAAAAAAAAAAGAGGATTATTAAAAAAATCAATACATAGGATAAATATAAGAATAGATAAGAAGAGATGCGACTTC
LobMar    ATAAAGAAAATTCAAATTTAGAATTTTTCAAAATAAAAAAAAAAGAGGATTGTTAAAAAAATCAATACATAGGACAAGTACAAAAACAGATAAGAAGAGATGCGACTTC
NasOff    ATAAAGAAAATTCGAATTTAGAATTTTTCAAAATAAAAAAAAAAGAGGATCATTAAAAAAATCAATACATAGGACAAATACAAGAACAGATAAGAAGAGATGCGACTGC
OliPum    ATAAAGAAAATTCGAATTTAGAATTTTTCAAAATAAAAAAAAAGGAGGATCATTAAAAAAATCAATACATAGGACAAATACAAGAACAGATAAGAAGAGATGCGACTTC
PacChe    ATAAAGAAAATTCGAATTTAGAATTTTTCAAAATAAAAAAAAAAGAGGATCATTAAAAAAATCAATACATAGGACAAATACAAGAACAGATAAGAAGAGATGCGACTTC
PacEny    ATAAAGAAAATTCGAATTTAGAATTTTTCAAAATAAAAAAAAAAGAGGATCATTAAAAAAATCAATACATAGGACAAATACAAGAACAGATAAGAAGAGATGCGACTTC
RapSat    ATAAAGAAAATTCGAATTTAGAATTTTTCAAAATAAAAAAAAAAGAGGATCATTAAAAAAATCAATACATAGCACAAATACAAGAACAGATAAGAAGAGATGCGACTTC
```

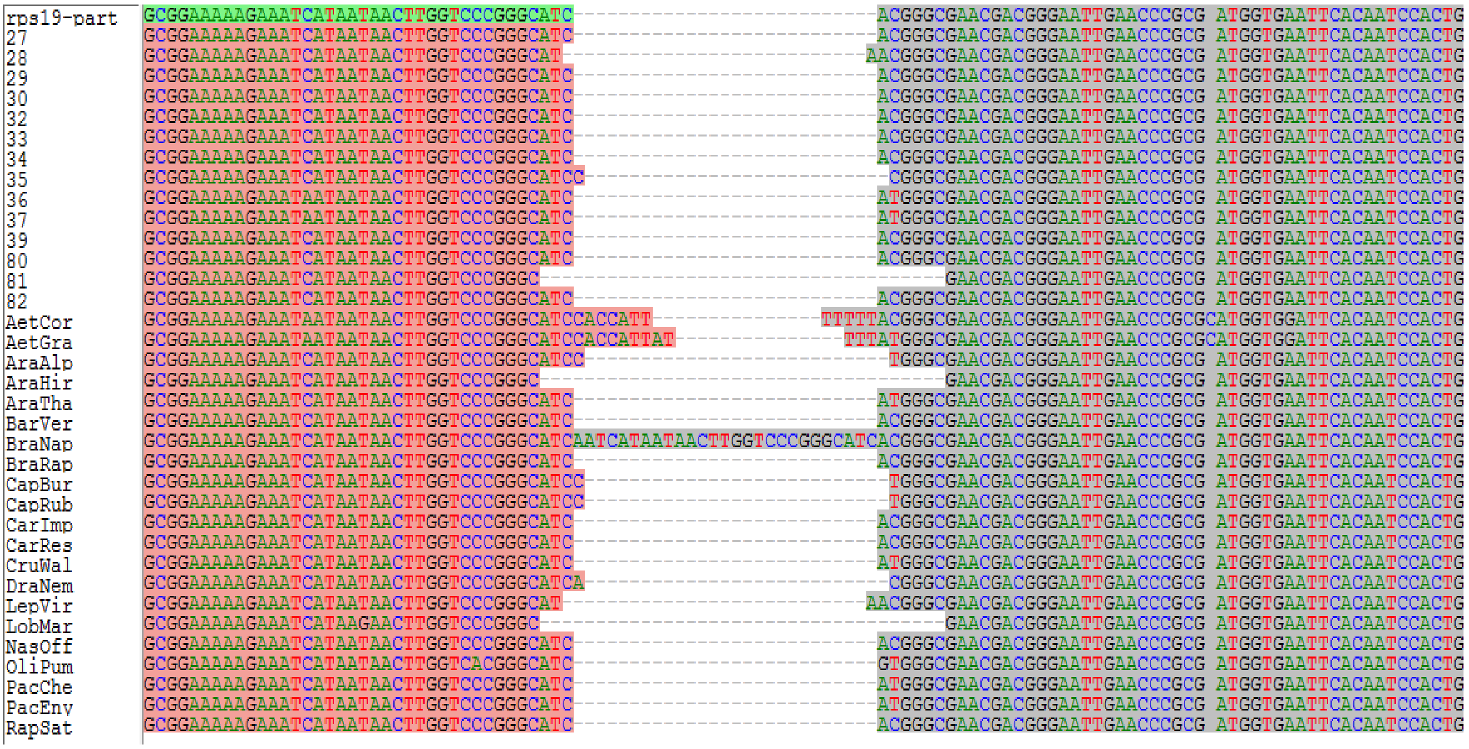


```
ycf1-ndhF ATAAAGAAAATTCGAATTTAGAATTTTTCAAAATAAAAAAAAAAGAGGATCGTTAAAAAAATCAATACATAGGACAAATACAAGAACAGATAAGAAGAGATGCGACTTC
27        .............................................................................................................
28        .................................G....C............A..........A........A................A............C.......
29        ...........G.......................................------.TT.................................................
30        ..............................................A....C.........................................................
32        ...........G.......................................------.TT.................................................
33        .............................................................................................................
34        .....................C......T....A.........GA..ATC.A.................................................A.......
35        ...................C.C...........A.................A..........A..............................................
36        ...................................................A......................T..................................
37        ...................................................A......................T..................................
39        ...................................................A.........C..........A.......T..C........C..............A.
80        ..............................................A....A.........................................................
81        .....................C.............................C......................A..................................
82        ..............................................A....A.........................................................
AetCor    .............T..............A.................-----------------........A....................C........A.......
AetGra    .............T..............A..................-----------------.......A.............................A..G....
AraAlp    .............T......A..............................A..................................T......................
AraHir    .............T......A......------......T...........C..................................T...........A..........
AraTha    ...................................................A............................T............................
BarVer    ......................................................................................T.............CC.......
BraNap    .............T..........................................................C....................................
BraRap    .............T..........................................................C....................................
CapBur    ...................................................AG...........................G..........................C.
CapRub    ...............................................A...A.........C............T.....T..........................C.
CarImp    ............................G......................A.........................................................
CarRes    ...........G.......................................------.TT.................................................
CruWal    ...................................................A.........................................................
DraNem    ....................A..............................A..........A.......................T......A....A..........
LepVir    ..................................................TA......................T.....T.....T......................
LobMar    .............A....................................T..........................G.....A.........................
NasOff    ...................................................A.......................................................G.
OliPum    ...........................................G.......A.........................................................
PacChe    ...................................................A.........................................................
PacEny    ...................................................A.........................................................
RapSat    ...................................................A....................C....................................
```

2. IRb-LSC

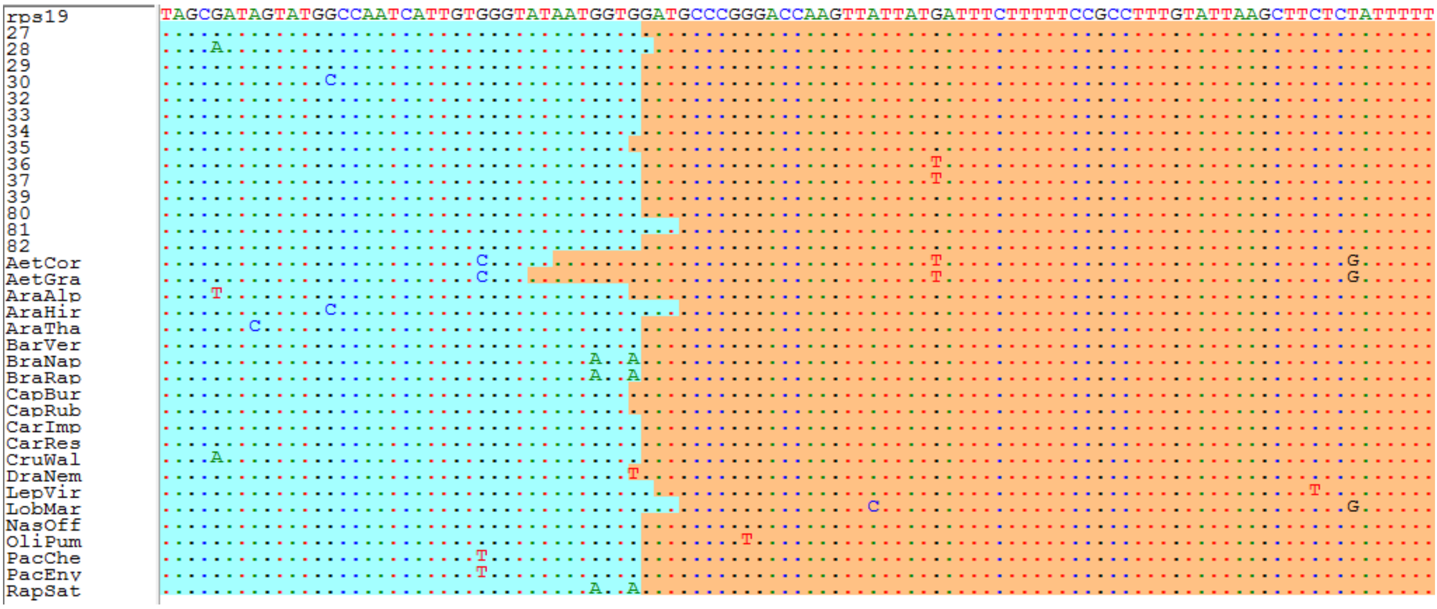


```
rps19-part  GCGGAAAAAGAAATCATAATAACTTGGTCCCGGGCATC---------------------------ACGGGCGAACGACGGGAATTGAACCCGCG ATGGTGAATTCACAATCCACTG
27          ......................................---------------------------............................. ......................
28          .....................................---------------------------A............................. ......................
29          ......................................---------------------------............................. ......................
30          ......................................---------------------------............................. ......................
32          ......................................---------------------------............................. ......................
33          ......................................---------------------------............................. ......................
34          ......................................---------------------------............................. ......................
35          ......................................C---------------------------............................ ......................
36          ..............A.......................---------------------------.T........................... ......................
37          ..............A.......................---------------------------.T........................... ......................
39          ......................................---------------------------............................. ......................
80          ......................................---------------------------............................. ......................
81          ...................................------------------------------------....................... ......................
82          ......................................---------------------------............................. ......................
AetCor      ..............A.......................CACCATT---------------TTTTT.............................C......G...............
AetGra      ..............A.......................CACCATTAT---------------TTT.T...........................C......G...............
AraAlp      ......................................C---------------------------T............................ ......................
AraHir      ...................................------------------------------------....................... ......................
AraTha      ......................................---------------------------.T........................... ......................
BarVer      ......................................---------------------------............................. ......................
BraNap      ......................................AATCATAATAACTTGGTCCCGGGCATC............................. ......................
BraRap      ......................................---------------------------............................. ......................
CapBur      ......................................C---------------------------T............................ ......................
CapRub      ......................................C---------------------------T............................ ......................
CarImp      ......................................---------------------------............................. ......................
CarRes      ......................................---------------------------............................. ......................
CruWal      ......................................---------------------------.T........................... ......................
DraNem      ......................................A---------------------------............................ ......................
LepVir      .....................................---------------------------A............................. ......................
LobMar      ...................G...............------------------------------------....................... ......................
NasOff      ......................................---------------------------............................. ......................
OliPum      .............................A........---------------------------GT........................... ......................
PacChe      ......................................---------------------------.T........................... ......................
PacEnv      ......................................---------------------------.T........................... ......................
RapSat      ......................................---------------------------............................. ......................
```

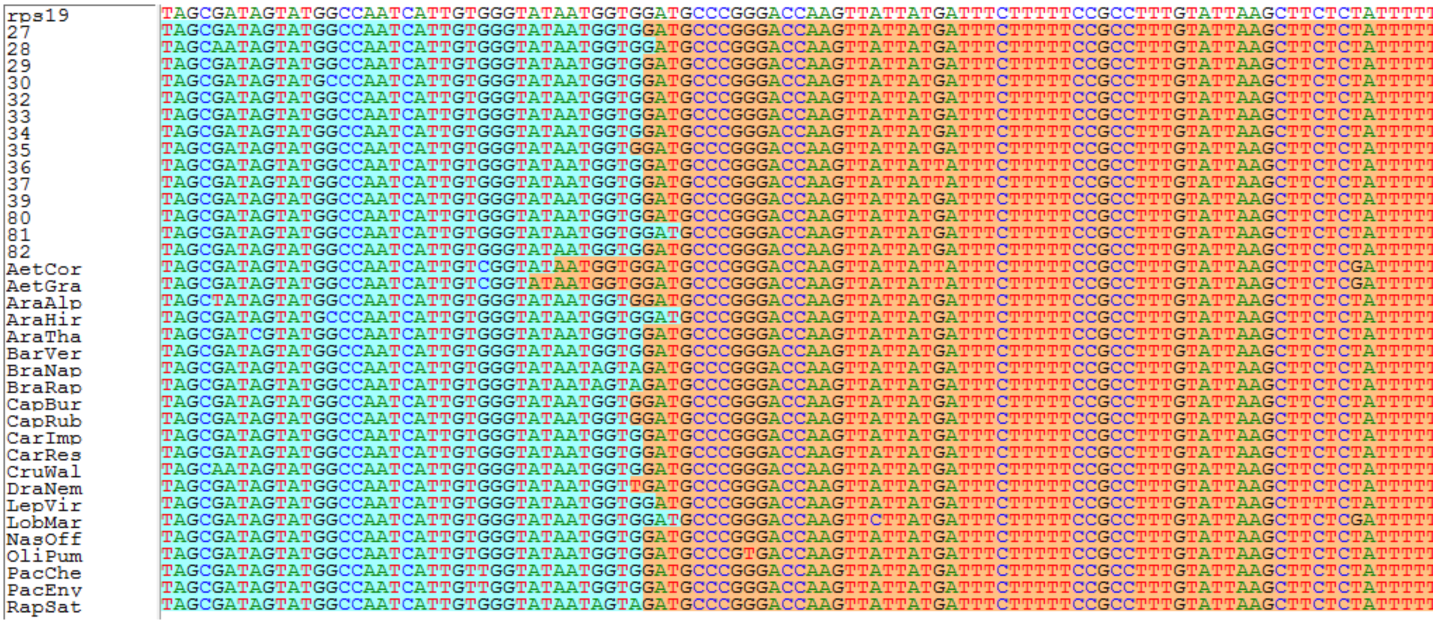


```
rps19-part  GCGGAAAAAGAAATCATAATAACTTGGTCCCGGGCATC---------------------------ACGGGCGAACGACGGGAATTGAACCCGCG ATGGTGAATTCACAATCCACTG
27          GCGGAAAAAGAAATCATAATAACTTGGTCCCGGGCATC---------------------------ACGGGCGAACGACGGGAATTGAACCCGCG ATGGTGAATTCACAATCCACTG
28          GCGGAAAAAGAAATCATAATAACTTGGTCCCGGGCAT---------------------------AACGGGCGAACGACGGGAATTGAACCCGCG ATGGTGAATTCACAATCCACTG
29          GCGGAAAAAGAAATCATAATAACTTGGTCCCGGGCATC---------------------------ACGGGCGAACGACGGGAATTGAACCCGCG ATGGTGAATTCACAATCCACTG
30          GCGGAAAAAGAAATCATAATAACTTGGTCCCGGGCATC---------------------------ACGGGCGAACGACGGGAATTGAACCCGCG ATGGTGAATTCACAATCCACTG
32          GCGGAAAAAGAAATCATAATAACTTGGTCCCGGGCATC---------------------------ACGGGCGAACGACGGGAATTGAACCCGCG ATGGTGAATTCACAATCCACTG
33          GCGGAAAAAGAAATCATAATAACTTGGTCCCGGGCATC---------------------------ACGGGCGAACGACGGGAATTGAACCCGCG ATGGTGAATTCACAATCCACTG
34          GCGGAAAAAGAAATCATAATAACTTGGTCCCGGGCATC---------------------------ACGGGCGAACGACGGGAATTGAACCCGCG ATGGTGAATTCACAATCCACTG
35          GCGGAAAAAGAAATCATAATAACTTGGTCCCGGGCATCC---------------------------CGGGCGAACGACGGGAATTGAACCCGCG ATGGTGAATTCACAATCCACTG
36          GCGGAAAAAGAAATAATAATAACTTGGTCCCGGGCATC---------------------------ATGGGCGAACGACGGGAATTGAACCCGCG ATGGTGAATTCACAATCCACTG
37          GCGGAAAAAGAAATAATAATAACTTGGTCCCGGGCATC---------------------------ATGGGCGAACGACGGGAATTGAACCCGCG ATGGTGAATTCACAATCCACTG
39          GCGGAAAAAGAAATCATAATAACTTGGTCCCGGGCATC---------------------------ACGGGCGAACGACGGGAATTGAACCCGCG ATGGTGAATTCACAATCCACTG
80          GCGGAAAAAGAAATCATAATAACTTGGTCCCGGGCATC---------------------------ACGGGCGAACGACGGGAATTGAACCCGCG ATGGTGAATTCACAATCCACTG
81          GCGGAAAAAGAAATCATAATAACTTGGTCCCGGGC------------------------------------GAACGACGGGAATTGAACCCGCG ATGGTGAATTCACAATCCACTG
82          GCGGAAAAAGAAATCATAATAACTTGGTCCCGGGCATC---------------------------ACGGGCGAACGACGGGAATTGAACCCGCG ATGGTGAATTCACAATCCACTG
AetCor      GCGGAAAAAGAAATAATAATAACTTGGTCCCGGGCATCCACCATT---------------TTTTTACGGGCGAACGACGGGAATTGAACCCGCGCATGGTGGATTCACAATCCACTG
AetGra      GCGGAAAAAGAAATAATAATAACTTGGTCCCGGGCATCCACCATTAT---------------TTTATGGGCGAACGACGGGAATTGAACCCGCGCATGGTGGATTCACAATCCACTG
AraAlp      GCGGAAAAAGAAATCATAATAACTTGGTCCCGGGCATCC---------------------------TGGGCGAACGACGGGAATTGAACCCGCG ATGGTGAATTCACAATCCACTG
AraHir      GCGGAAAAAGAAATCATAATAACTTGGTCCCGGGC------------------------------------GAACGACGGGAATTGAACCCGCG ATGGTGAATTCACAATCCACTG
AraTha      GCGGAAAAAGAAATCATAATAACTTGGTCCCGGGCATC---------------------------ATGGGCGAACGACGGGAATTGAACCCGCG ATGGTGAATTCACAATCCACTG
BarVer      GCGGAAAAAGAAATCATAATAACTTGGTCCCGGGCATC---------------------------ACGGGCGAACGACGGGAATTGAACCCGCG ATGGTGAATTCACAATCCACTG
BraNap      GCGGAAAAAGAAATCATAATAACTTGGTCCCGGGCATCAATCATAATAACTTGGTCCCGGGCATCACGGGCGAACGACGGGAATTGAACCCGCG ATGGTGAATTCACAATCCACTG
BraRap      GCGGAAAAAGAAATCATAATAACTTGGTCCCGGGCATC---------------------------ACGGGCGAACGACGGGAATTGAACCCGCG ATGGTGAATTCACAATCCACTG
CapBur      GCGGAAAAAGAAATCATAATAACTTGGTCCCGGGCATCC---------------------------TGGGCGAACGACGGGAATTGAACCCGCG ATGGTGAATTCACAATCCACTG
CapRub      GCGGAAAAAGAAATCATAATAACTTGGTCCCGGGCATCC---------------------------TGGGCGAACGACGGGAATTGAACCCGCG ATGGTGAATTCACAATCCACTG
CarImp      GCGGAAAAAGAAATCATAATAACTTGGTCCCGGGCATC---------------------------ACGGGCGAACGACGGGAATTGAACCCGCG ATGGTGAATTCACAATCCACTG
CarRes      GCGGAAAAAGAAATCATAATAACTTGGTCCCGGGCATC---------------------------ACGGGCGAACGACGGGAATTGAACCCGCG ATGGTGAATTCACAATCCACTG
CruWal      GCGGAAAAAGAAATCATAATAACTTGGTCCCGGGCATC---------------------------ATGGGCGAACGACGGGAATTGAACCCGCG ATGGTGAATTCACAATCCACTG
DraNem      GCGGAAAAAGAAATCATAATAACTTGGTCCCGGGCATCA---------------------------CGGGCGAACGACGGGAATTGAACCCGCG ATGGTGAATTCACAATCCACTG
LepVir      GCGGAAAAAGAAATCATAATAACTTGGTCCCGGGCAT---------------------------AACGGGCGAACGACGGGAATTGAACCCGCG ATGGTGAATTCACAATCCACTG
LobMar      GCGGAAAAAGAAATCATAAGAACTTGGTCCCGGGC------------------------------------GAACGACGGGAATTGAACCCGCG ATGGTGAATTCACAATCCACTG
NasOff      GCGGAAAAAGAAATCATAATAACTTGGTCCCGGGCATC---------------------------ACGGGCGAACGACGGGAATTGAACCCGCG ATGGTGAATTCACAATCCACTG
OliPum      GCGGAAAAAGAAATCATAATAACTTGGTCACGGGCATC---------------------------GTGGGCGAACGACGGGAATTGAACCCGCG ATGGTGAATTCACAATCCACTG
PacChe      GCGGAAAAAGAAATCATAATAACTTGGTCCCGGGCATC---------------------------ATGGGCGAACGACGGGAATTGAACCCGCG ATGGTGAATTCACAATCCACTG
PacEnv      GCGGAAAAAGAAATCATAATAACTTGGTCCCGGGCATC---------------------------ATGGGCGAACGACGGGAATTGAACCCGCG ATGGTGAATTCACAATCCACTG
RapSat      GCGGAAAAAGAAATCATAATAACTTGGTCCCGGGCATC---------------------------ACGGGCGAACGACGGGAATTGAACCCGCG ATGGTGAATTCACAATCCACTG
```

3. LSC-IRa

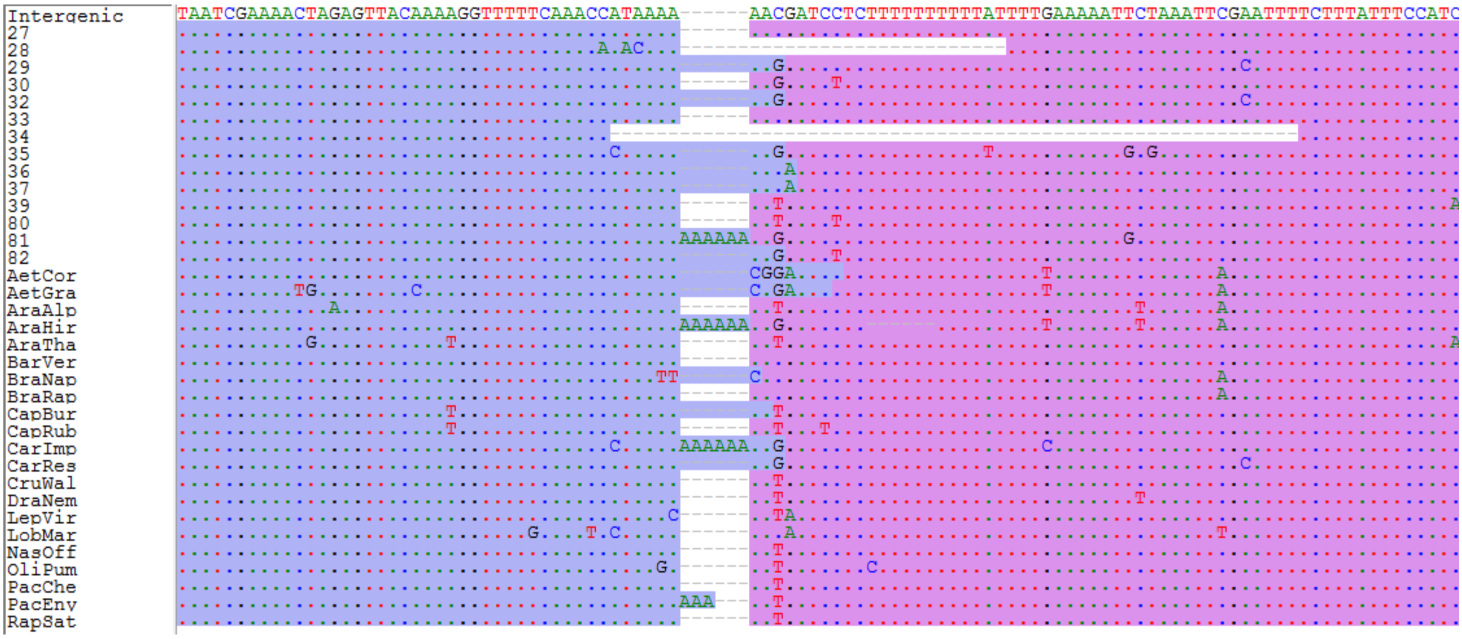


```
rps19   TAGCGATAGTATGGCCAATCATTGTGGGTATAATGGTGGATGCCCGGGACCAAGTTATTATGATTTCTTTTTCCGCCTTTGTATTAAGCTTCTCTATTTTT
27      .....................................................................................................
28      ....A................................................................................................
29      .....................................................................................................
30      .............C.......................................................................................
32      .....................................................................................................
33      .....................................................................................................
34      .....................................................................................................
35      .....................................................................................................
36      .............................................................T.......................................
37      .............................................................T.......................................
39      .....................................................................................................
80      .....................................................................................................
81      .....................................................................................................
82      .....................................................................................................
AetCor  .........................C...................................T................................G......
AetGra  .........................C...................................T................................G......
AraAlp  ....T................................................................................................
AraHir  .............C.......................................................................................
AraTha  .......C.............................................................................................
BarVer  .....................................................................................................
BraNap  ..................................A..A...............................................................
BraRap  ..................................A..A...............................................................
CapBur  .....................................................................................................
CapRub  .....................................................................................................
CarImp  .....................................................................................................
CarRes  .....................................................................................................
CruWal  ....A................................................................................................
DraNem  .....................................T...............................................................
LepVir  ...........................................................................................T.........
LobMar  ........................................................C.....................................G......
NasOff  .....................................................................................................
OliPum  ..............................................T......................................................
PacChe  .........................T...........................................................................
PacEnv  .........................T...........................................................................
RapSat  ..................................A..A...............................................................
```

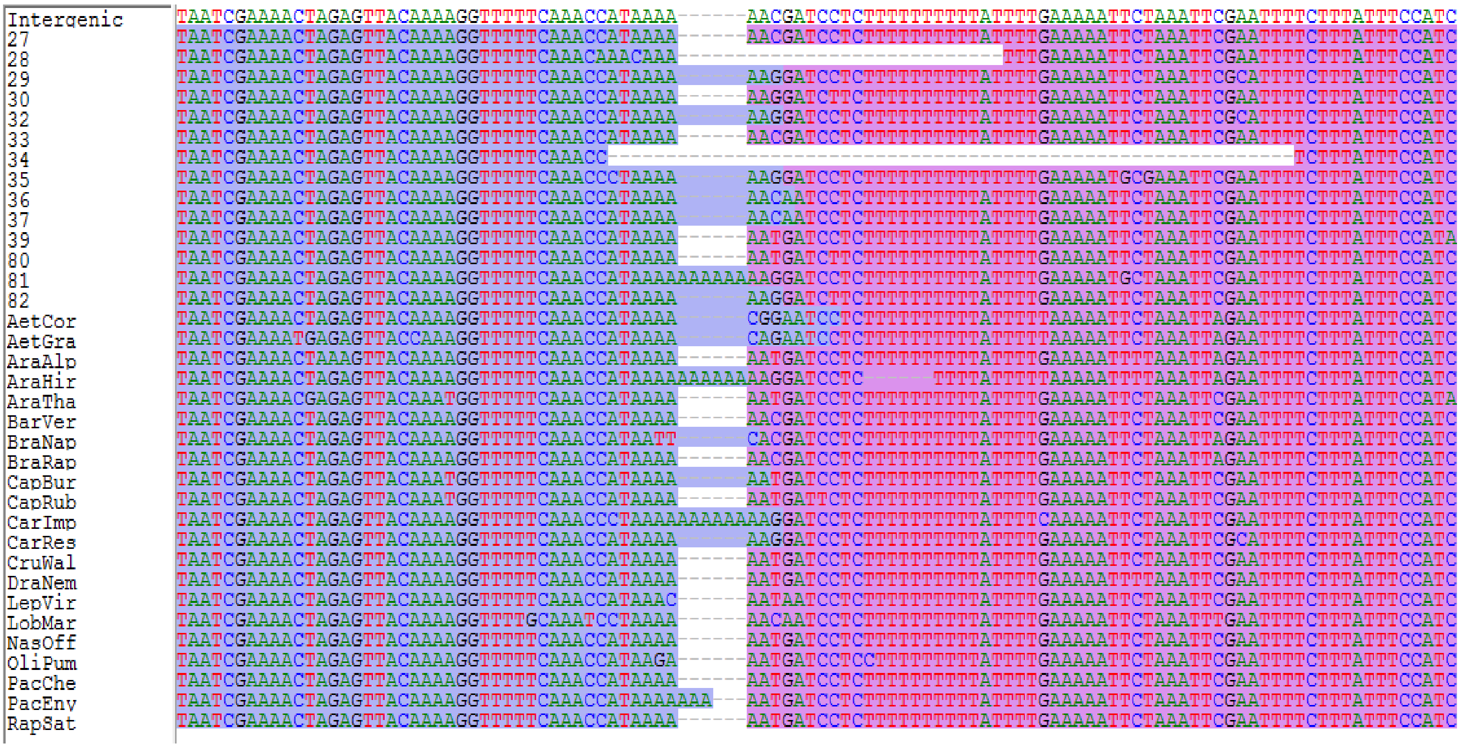


```
rps19   TAGCGATAGTATGGCCAATCATTGTGGGTATAATGGTGGATGCCCGGGACCAAGTTATTATGATTTCTTTTTCCGCCTTTGTATTAAGCTTCTCTATTTTT
27      TAGCGATAGTATGGCCAATCATTGTGGGTATAATGGTGGATGCCCGGGACCAAGTTATTATGATTTCTTTTTCCGCCTTTGTATTAAGCTTCTCTATTTTT
28      TAGCAATAGTATGGCCAATCATTGTGGGTATAATGGTGGATGCCCGGGACCAAGTTATTATGATTTCTTTTTCCGCCTTTGTATTAAGCTTCTCTATTTTT
29      TAGCGATAGTATGGCCAATCATTGTGGGTATAATGGTGGATGCCCGGGACCAAGTTATTATGATTTCTTTTTCCGCCTTTGTATTAAGCTTCTCTATTTTT
30      TAGCGATAGTATGCCCAATCATTGTGGGTATAATGGTGGATGCCCGGGACCAAGTTATTATGATTTCTTTTTCCGCCTTTGTATTAAGCTTCTCTATTTTT
32      TAGCGATAGTATGGCCAATCATTGTGGGTATAATGGTGGATGCCCGGGACCAAGTTATTATGATTTCTTTTTCCGCCTTTGTATTAAGCTTCTCTATTTTT
33      TAGCGATAGTATGGCCAATCATTGTGGGTATAATGGTGGATGCCCGGGACCAAGTTATTATGATTTCTTTTTCCGCCTTTGTATTAAGCTTCTCTATTTTT
34      TAGCGATAGTATGGCCAATCATTGTGGGTATAATGGTGGATGCCCGGGACCAAGTTATTATGATTTCTTTTTCCGCCTTTGTATTAAGCTTCTCTATTTTT
35      TAGCGATAGTATGGCCAATCATTGTGGGTATAATGGTGGATGCCCGGGACCAAGTTATTATGATTTCTTTTTCCGCCTTTGTATTAAGCTTCTCTATTTTT
36      TAGCGATAGTATGGCCAATCATTGTGGGTATAATGGTGGATGCCCGGGACCAAGTTATTATTATTTCTTTTTCCGCCTTTGTATTAAGCTTCTCTATTTTT
37      TAGCGATAGTATGGCCAATCATTGTGGGTATAATGGTGGATGCCCGGGACCAAGTTATTATTATTTCTTTTTCCGCCTTTGTATTAAGCTTCTCTATTTTT
39      TAGCGATAGTATGGCCAATCATTGTGGGTATAATGGTGGATGCCCGGGACCAAGTTATTATGATTTCTTTTTCCGCCTTTGTATTAAGCTTCTCTATTTTT
80      TAGCGATAGTATGGCCAATCATTGTGGGTATAATGGTGGATGCCCGGGACCAAGTTATTATGATTTCTTTTTCCGCCTTTGTATTAAGCTTCTCTATTTTT
81      TAGCGATAGTATGGCCAATCATTGTGGGTATAATGGTGGATGCCCGGGACCAAGTTATTATGATTTCTTTTTCCGCCTTTGTATTAAGCTTCTCTATTTTT
82      TAGCGATAGTATGGCCAATCATTGTGGGTATAATGGTGGATGCCCGGGACCAAGTTATTATGATTTCTTTTTCCGCCTTTGTATTAAGCTTCTCTATTTTT
AetCor  TAGCGATAGTATGGCCAATCATTGTCGGTATAATGGTGGATGCCCGGGACCAAGTTATTATTATTTCTTTTTCCGCCTTTGTATTAAGCTTCTCGATTTTT
AetGra  TAGCGATAGTATGGCCAATCATTGTCGGTATAATGGTGGATGCCCGGGACCAAGTTATTATTATTTCTTTTTCCGCCTTTGTATTAAGCTTCTCGATTTTT
AraAlp  TAGCTATAGTATGGCCAATCATTGTGGGTATAATGGTGGATGCCCGGGACCAAGTTATTATGATTTCTTTTTCCGCCTTTGTATTAAGCTTCTCTATTTTT
AraHir  TAGCGATAGTATGCCCAATCATTGTGGGTATAATGGTGGATGCCCGGGACCAAGTTATTATGATTTCTTTTTCCGCCTTTGTATTAAGCTTCTCTATTTTT
AraTha  TAGCGATCGTATGGCCAATCATTGTGGGTATAATGGTGGATGCCCGGGACCAAGTTATTATGATTTCTTTTTCCGCCTTTGTATTAAGCTTCTCTATTTTT
BarVer  TAGCGATAGTATGGCCAATCATTGTGGGTATAATGGTGGATGCCCGGGACCAAGTTATTATGATTTCTTTTTCCGCCTTTGTATTAAGCTTCTCTATTTTT
BraNap  TAGCGATAGTATGGCCAATCATTGTGGGTATAATAGTAGATGCCCGGGACCAAGTTATTATGATTTCTTTTTCCGCCTTTGTATTAAGCTTCTCTATTTTT
BraRap  TAGCGATAGTATGGCCAATCATTGTGGGTATAATAGTAGATGCCCGGGACCAAGTTATTATGATTTCTTTTTCCGCCTTTGTATTAAGCTTCTCTATTTTT
CapBur  TAGCGATAGTATGGCCAATCATTGTGGGTATAATGGTGGATGCCCGGGACCAAGTTATTATGATTTCTTTTTCCGCCTTTGTATTAAGCTTCTCTATTTTT
CapRub  TAGCGATAGTATGGCCAATCATTGTGGGTATAATGGTGGATGCCCGGGACCAAGTTATTATGATTTCTTTTTCCGCCTTTGTATTAAGCTTCTCTATTTTT
CarImp  TAGCGATAGTATGGCCAATCATTGTGGGTATAATGGTGGATGCCCGGGACCAAGTTATTATGATTTCTTTTTCCGCCTTTGTATTAAGCTTCTCTATTTTT
CarRes  TAGCGATAGTATGGCCAATCATTGTGGGTATAATGGTGGATGCCCGGGACCAAGTTATTATGATTTCTTTTTCCGCCTTTGTATTAAGCTTCTCTATTTTT
CruWal  TAGCAATAGTATGGCCAATCATTGTGGGTATAATGGTGGATGCCCGGGACCAAGTTATTATGATTTCTTTTTCCGCCTTTGTATTAAGCTTCTCTATTTTT
DraNem  TAGCGATAGTATGGCCAATCATTGTGGGTATAATGGTTGATGCCCGGGACCAAGTTATTATGATTTCTTTTTCCGCCTTTGTATTAAGCTTCTCTATTTTT
LepVir  TAGCGATAGTATGGCCAATCATTGTGGGTATAATGGTGGATGCCCGGGACCAAGTTATTATGATTTCTTTTTCCGCCTTTGTATTAAGCTTTTCTATTTTT
LobMar  TAGCGATAGTATGGCCAATCATTGTGGGTATAATGGTGGATGCCCGGGACCAAGTTCTTATGATTTCTTTTTCCGCCTTTGTATTAAGCTTCTCGATTTTT
NasOff  TAGCGATAGTATGGCCAATCATTGTGGGTATAATGGTGGATGCCCGGGACCAAGTTATTATGATTTCTTTTTCCGCCTTTGTATTAAGCTTCTCTATTTTT
OliPum  TAGCGATAGTATGGCCAATCATTGTGGGTATAATGGTGGATGCCCGTGACCAAGTTATTATGATTTCTTTTTCCGCCTTTGTATTAAGCTTCTCTATTTTT
PacChe  TAGCGATAGTATGGCCAATCATTGTTGGTATAATGGTGGATGCCCGGGACCAAGTTATTATGATTTCTTTTTCCGCCTTTGTATTAAGCTTCTCTATTTTT
PacEnv  TAGCGATAGTATGGCCAATCATTGTTGGTATAATGGTGGATGCCCGGGACCAAGTTATTATGATTTCTTTTTCCGCCTTTGTATTAAGCTTCTCTATTTTT
RapSat  TAGCGATAGTATGGCCAATCATTGTGGGTATAATAGTAGATGCCCGGGACCAAGTTATTATGATTTCTTTTTCCGCCTTTGTATTAAGCTTCTCTATTTTT
```

4. SSC-IRb



```
Intergenic  TAATCGAAAACTAGAGTTACAAAAGGTTTTTCAAACCATAAAA------AACGATCCTCTTTTTTTTTTATTTTGAAAAATTCTAAATTCGAATTTTCTTTATTTCCATC
27          ...........................................------.............................................................
28          ....................................A.AC...----------------------------.......................................
29          ...........................................------..G.......................................C..................
30          ...........................................------..G....T.....................................................
32          ...........................................------..G.......................................C..................
33          ...........................................------.............................................................
34          .....................................-----------------------------------------------------------..............
35          .....................................C.....------..G.................T...........G.G..........................
36          ...........................................------...A.........................................................
37          ...........................................------...A.........................................................
39          ...........................................------..T.........................................................A
80          ...........................................------..T....T.....................................................
81          ...........................................AAAAAA..G.............................G............................
82          ...........................................------..G....T.....................................................
AetCor      ...........................................------CGGA.....................T..............A....................
AetGra      ..........TG........C......................------C.GA.....................T..............A....................
AraAlp      .............A.............................------..T..............................T......A....................
AraHir      ...........................................AAAAAA..G.......------.........T.......T......A....................
AraTha      ...........G...........T...................------..T.........................................................A
BarVer      ...........................................------.............................................................
BraNap      .........................................TT------C.......................................A....................
BraRap      ...........................................------........................................A....................
CapBur      .......................T...................------..T..........................................................
CapRub      .......................T...................------..T...T......................................................
CarImp      .....................................C.....AAAAAA..G......................C...................................
CarRes      ...........................................------..G.......................................C..................
CruWal      ...........................................------..T..........................................................
DraNem      ...........................................------..T..............................T...........................
LepVir      ..........................................C------..TA.........................................................
LobMar      ..............................G....T.C.....------...A....................................T....................
NasOff      ...........................................------..T..........................................................
OliPum      .........................................G.------..T.......C..................................................
PacChe      ...........................................------..T..........................................................
PacEnv      ...........................................AAA---..T..........................................................
RapSat      ...........................................------..T..........................................................
```



```
Intergenic  TAATCGAAAACTAGAGTTACAAAAGGTTTTTCAAACCATAAAA------AACGATCCTCTTTTTTTTTTATTTTGAAAAATTCTAAATTCGAATTTTCTTTATTTCCATC
27          TAATCGAAAACTAGAGTTACAAAAGGTTTTTCAAACCATAAAA------AACGATCCTCTTTTTTTTTTATTTTGAAAAATTCTAAATTCGAATTTTCTTTATTTCCATC
28          TAATCGAAAACTAGAGTTACAAAAGGTTTTTCAAACAAACAAA----------------------------TTTGAAAAATTCTAAATTCGAATTTTCTTTATTTCCATC
29          TAATCGAAAACTAGAGTTACAAAAGGTTTTTCAAACCATAAAA------AAGGATCCTCTTTTTTTTTTATTTTGAAAAATTCTAAATTCGCATTTTCTTTATTTCCATC
30          TAATCGAAAACTAGAGTTACAAAAGGTTTTTCAAACCATAAAA------AAGGATCTTCTTTTTTTTTTATTTTGAAAAATTCTAAATTCGAATTTTCTTTATTTCCATC
32          TAATCGAAAACTAGAGTTACAAAAGGTTTTTCAAACCATAAAA------AAGGATCCTCTTTTTTTTTTATTTTGAAAAATTCTAAATTCGCATTTTCTTTATTTCCATC
33          TAATCGAAAACTAGAGTTACAAAAGGTTTTTCAAACCATAAAA------AACGATCCTCTTTTTTTTTTATTTTGAAAAATTCTAAATTCGAATTTTCTTTATTTCCATC
34          TAATCGAAAACTAGAGTTACAAAAGGTTTTTCAAACC-----------------------------------------------------------TCTTTATTTCCATC
35          TAATCGAAAACTAGAGTTACAAAAGGTTTTTCAAACCCTAAAA------AAGGATCCTCTTTTTTTTTTTTTTTGAAAAATGCGAAATTCGAATTTTCTTTATTTCCATC
36          TAATCGAAAACTAGAGTTACAAAAGGTTTTTCAAACCATAAAA------AACAATCCTCTTTTTTTTTTATTTTGAAAAATTCTAAATTCGAATTTTCTTTATTTCCATC
37          TAATCGAAAACTAGAGTTACAAAAGGTTTTTCAAACCATAAAA------AACAATCCTCTTTTTTTTTTATTTTGAAAAATTCTAAATTCGAATTTTCTTTATTTCCATC
39          TAATCGAAAACTAGAGTTACAAAAGGTTTTTCAAACCATAAAA------AATGATCCTCTTTTTTTTTTATTTTGAAAAATTCTAAATTCGAATTTTCTTTATTTCCATA
80          TAATCGAAAACTAGAGTTACAAAAGGTTTTTCAAACCATAAAA------AATGATCTTCTTTTTTTTTTATTTTGAAAAATTCTAAATTCGAATTTTCTTTATTTCCATC
81          TAATCGAAAACTAGAGTTACAAAAGGTTTTTCAAACCATAAAAAAAAAAAGGATCCTCTTTTTTTTTTATTTTGAAAAATGCTAAATTCGAATTTTCTTTATTTCCATC
82          TAATCGAAAACTAGAGTTACAAAAGGTTTTTCAAACCATAAAA------AAGGATCTTCTTTTTTTTTTATTTTGAAAAATTCTAAATTCGAATTTTCTTTATTTCCATC
AetCor      TAATCGAAAACTAGAGTTACAAAAGGTTTTTCAAACCATAAAA------CGGAATCCTCTTTTTTTTTTATTTTTAAAAATTCTAAATTAGAATTTTCTTTATTTCCATC
AetGra      TAATCGAAAATGAGAGTTACCAAAGGTTTTTCAAACCATAAAA------CAGAATCCTCTTTTTTTTTTATTTTTAAAAATTCTAAATTAGAATTTTCTTTATTTCCATC
AraAlp      TAATCGAAAACTAAAGTTACAAAAGGTTTTTCAAACCATAAAA------AATGATCCTCTTTTTTTTTTATTTTGAAAAATTTTAAATTAGAATTTTCTTTATTTCCATC
AraHir      TAATCGAAAACTAGAGTTACAAAAGGTTTTTCAAACCATAAAAAAAAAAAGGATCCTC------TTTTATTTTTAAAAATTTTAAATTAGAATTTTCTTTATTTCCATC
AraTha      TAATCGAAAACGAGAGTTACAAATGGTTTTTCAAACCATAAAA------AATGATCCTCTTTTTTTTTTATTTTGAAAAATTCTAAATTCGAATTTTCTTTATTTCCATA
BarVer      TAATCGAAAACTAGAGTTACAAAAGGTTTTTCAAACCATAAAA------AACGATCCTCTTTTTTTTTTATTTTGAAAAATTCTAAATTCGAATTTTCTTTATTTCCATC
BraNap      TAATCGAAAACTAGAGTTACAAAAGGTTTTTCAAACCATAATT------CACGATCCTCTTTTTTTTTTATTTTGAAAAATTCTAAATTAGAATTTTCTTTATTTCCATC
BraRap      TAATCGAAAACTAGAGTTACAAAAGGTTTTTCAAACCATAAAA------AACGATCCTCTTTTTTTTTTATTTTGAAAAATTCTAAATTAGAATTTTCTTTATTTCCATC
CapBur      TAATCGAAAACTAGAGTTACAAATGGTTTTTCAAACCATAAAA------AATGATCCTCTTTTTTTTTTATTTTGAAAAATTCTAAATTCGAATTTTCTTTATTTCCATC
CapRub      TAATCGAAAACTAGAGTTACAAATGGTTTTTCAAACCATAAAA------AATGATTCTCTTTTTTTTTTATTTTGAAAAATTCTAAATTCGAATTTTCTTTATTTCCATC
CarImp      TAATCGAAAACTAGAGTTACAAAAGGTTTTTCAAACCCTAAAAAAAAAAAGGATCCTCTTTTTTTTTTATTTTCAAAAATTCTAAATTCGAATTTTCTTTATTTCCATC
CarRes      TAATCGAAAACTAGAGTTACAAAAGGTTTTTCAAACCATAAAA------AAGGATCCTCTTTTTTTTTTATTTTGAAAAATTCTAAATTCGCATTTTCTTTATTTCCATC
CruWal      TAATCGAAAACTAGAGTTACAAAAGGTTTTTCAAACCATAAAA------AATGATCCTCTTTTTTTTTTATTTTGAAAAATTCTAAATTCGAATTTTCTTTATTTCCATC
DraNem      TAATCGAAAACTAGAGTTACAAAAGGTTTTTCAAACCATAAAA------AATGATCCTCTTTTTTTTTTATTTTGAAAAATTTTAAATTCGAATTTTCTTTATTTCCATC
LepVir      TAATCGAAAACTAGAGTTACAAAAGGTTTTTCAAACCATAAAC------AATAATCCTCTTTTTTTTTTATTTTGAAAAATTCTAAATTCGAATTTTCTTTATTTCCATC
LobMar      TAATCGAAAACTAGAGTTACAAAAGGTTTTGCAAATCCTAAAA------AACAATCCTCTTTTTTTTTTATTTTGAAAAATTCTAAATTTGAATTTTCTTTATTTCCATC
NasOff      TAATCGAAAACTAGAGTTACAAAAGGTTTTTCAAACCATAAAA------AATGATCCTCTTTTTTTTTTATTTTGAAAAATTCTAAATTCGAATTTTCTTTATTTCCATC
OliPum      TAATCGAAAACTAGAGTTACAAAAGGTTTTTCAAACCATAAGA------AATGATCCTCCTTTTTTTTTATTTTGAAAAATTCTAAATTCGAATTTTCTTTATTTCCATC
PacChe      TAATCGAAAACTAGAGTTACAAAAGGTTTTTCAAACCATAAAA------AATGATCCTCTTTTTTTTTTATTTTGAAAAATTCTAAATTCGAATTTTCTTTATTTCCATC
PacEnv      TAATCGAAAACTAGAGTTACAAAAGGTTTTTCAAACCATAAAAAAA---AATGATCCTCTTTTTTTTTTATTTTGAAAAATTCTAAATTCGAATTTTCTTTATTTCCATC
RapSat      TAATCGAAAACTAGAGTTACAAAAGGTTTTTCAAACCATAAAA------AATGATCCTCTTTTTTTTTTATTTTGAAAAATTCTAAATTCGAATTTTCTTTATTTCCATC
```

Additional file 12 Table C3-1 GC content of 12 newly assembled cp genomes

Species number	Species number	A	C	G	T	GC	AT
27	*Rorippa.sylvestris*	31.36	18.5	17.84	32.3	36.34%	63.66%
28	*Cardamine.hirsuta*	31.37	18.51	17.91	32.21	36.42%	63.58%
29	*Cardamine.alpina*	31.45	18.42	17.86	32.27	36.28%	63.72%
30	*Cardamine.flexuosa*	31.38	18.49	17.9	32.23	36.38%	63.62%
33	*Rorippa.austriaca*	31.33	18.52	17.87	32.28	36.39%	63.61%
34	*Cardamine.enneaphyllos*	31.37	18.46	17.87	32.29	36.34%	63.66%
35	*Cardamine.pentaphyllos*	31.38	18.49	17.86	32.27	36.35%	63.65%
36	*Leavenworthia.uniflora*	31.46	18.45	17.81	32.28	36.26%	63.74%
37	*Leavenworthia.exigua*	31.37	18.49	17.86	32.28	36.35%	63.65%
80	*Cardamine.asarifolia*	31.34	18.48	17.92	32.26	36.40%	63.60%
81	*Cardamine.trifolia*	31.33	18.52	17.89	32.26	36.42%	63.58%
82	*Cardamine.pratensis*	31.36	18.5	17.92	32.22	36.42%	63.58%
	Average					36.36%	63.64%

Additional file 13 Table C3-2 A collection of reference genomes for assembly from NCBI

Number	Abbreviation	Species Name
1	AetCor	*Aethionema cordifolium*
2	AetGra	*Aethionema grandiflorum*
3	AraTha	*Arabidopsis thaliana*
4	AraAlp	*Arabis alpina*
5	AraHir	*Arabis hirsuta*
6	BarVer	*Barbarea verna*
7	BraNap	*Brassica napus*
8	BraRap	*Brassica rapa subsp. pekinensis*
9	CapBur	*Capsella bursa-pastoris*
10	CapGra	*Capsella grandiflora*
11	CarImp	*Cardamine impatiens*
12	CarRes	*Cardamine resedifolia*
13	CruWal	*Crucihimalaya wallichii*
14	DeaBou	*Descurainia bourgaeana*
15	DraNem	*Draba nemorosa*
16	LepVir	*Lepidium virginicum*
17	LobMar	*Lobularia maritima*
18	NasOff	*Nasturtium officinale*
19	OliPum	*Olimarabidopsis pumila*
20	PacChe	*Pachycladon cheesemanii*
21	PacEny	*Pachycladon enysii*
22	RapSat	*Raphanus sativus*

Additional file 14 Table C3-3. Repeat analysis for 12 newly assembled cp genomes (L=length, P=Position, T= Repeat type)

A27			A28			A29			A30			A33			A34		
L	P	T	L	P	T	L	P	T	L	P	T	L	P	T	L	P	T
30	4760	P	32	7566	F	30	244	P	30	1667	P	30	4777	P	30	1633	P
38	6364	P	30	7568	P	30	1636	P	30	4335	F	30	4785	P	30	6350	R
31	6366	C	41	9205	P	34	4642	P	30	4335	R	40	6477	P	30	6350	C
31	6374	R	40	28140	P	31	4646	P	32	7797	F	31	6477	P	32	7837	F
32	7803	F	30	34629	P	31	4646	P	30	7799	P	34	6489	F	30	7839	P
30	7805	P	30	34697	P	30	4646	P	31	7960	R	30	7942	P	53	28626	P
30	8384	P	46	35592	P	34	4649	F	31	7960	C	33	8513	P	31	32000	P
41	9408	P	67	37991	F	32	7829	F	30	8086	R	41	9540	P	30	35231	P
32	22274	R	55	38009	F	30	7831	P	31	8472	P	32	22402	R	30	35299	P
32	28420	P	39	43063	F	30	8155	F	41	9418	P	32	28472	P	30	36204	P
40	28562	P	39	43063	P	41	9496	P	40	28560	P	40	28616	P	67	38570	F
30	35221	P	30	43075	F	31	26926	P	37	31800	F	30	31329	C	55	38591	F
30	35289	P	30	43075	P	30	35344	P	31	31811	F	30	35329	P	39	43649	F
67	38503	F	41	47132	F	30	35412	P	30	35270	P	56	38541	F	39	43649	P
55	38521	F	31	47144	F	30	36334	P	30	35338	P	55	38562	F	30	43661	F
39	43584	F	30	47220	F	67	38694	F	35	36235	P	39	43642	F	30	43661	P
39	43584	P	38	47310	F	55	38715	F	67	38605	F	39	43642	P	33	50360	P
30	43596	F	31	47319	F	39	43777	F	53	38623	F	30	43654	F	40	50368	P
30	43596	P	44	73311	P	39	43777	P	31	42421	R	30	43654	P	30	81601	R
35	47599	F	45	88120	F	30	43789	F	30	42427	R	35	47489	F	30	81603	R
36	47622	F	45	88120	P	30	43789	P	39	43685	F	36	47512	F	45	88807	F
35	47639	F	45	88144	P	45	47889	F	39	43685	P	35	47529	F	45	88807	P
62	47673	F	32	88166	F	31	47905	F	30	43697	F	62	47563	F	45	88831	P
51	47684	F	32	88166	P	40	50532	P	30	43697	P	51	47574	F	32	88853	F
31	47707	F	32	88187	P	30	65051	P	38	47760	F	31	47597	F	32	88853	P
30	47787	F	33	107007	F	44	74138	P	38	47869	F	30	47692	F	32	88874	P
40	50392	P	33	107007	P	32	77011	P	31	47878	F	35	64054	R	37	98565	F
35	64391	R	33	107039	P	45	89048	F	45	88849	F	30	64058	R	33	107715	F
30	64395	R	45	113091	P	45	89048	P	45	88849	P	44	73342	P	33	107715	P
44	73680	P	30	124560	P	45	89072	P	45	88873	P	41	88217	F	33	107747	P
41	88564	F	33	130072	F	32	89094	F	32	88895	F	41	88217	P	45	113824	P
41	88564	P	32	148925	F	32	89094	P	32	88895	P	41	88241	P	37	120478	P
41	88588	P	45	148955	F	32	89115	P	32	88916	P	32	88263	F	33	130894	F
32	88610	F				37	98818	F	37	98617	F	32	88263	P	32	149768	F
32	88610	P				33	107951	F	33	107561	F	32	88284	P	45	149798	F
32	88631	P				33	107951	P	33	107561	P	37	97972	F			
37	98131	F				33	107983	P	33	107593	P	32	107083	F			
32	107145	F				45	114061	P	45	113637	P	32	107083	P			
32	107145	P				37	120695	P	31	115655	R	32	107115	P			
32	107177	P				32	124541	R	30	116961	R	46	116781	P			
32	112289	R				30	125617	P	37	120141	P	37	119963	P			
37	119884	P				33	131132	F	33	130533	F	32	130440	F			
32	130366	F				32	150001	F	32	149211	F	32	149271	F			
32	148912	F				45	150031	F	45	149241	F	41	149305	F			
41	148946	F															

A35			A36			A37			A39			A80			A81			A82		
L	P	T	L	P	T	L	P	T	L	P	T	L	P	T	L	P	T	L	P	T
30	1570	P	30	1647	P	37	1947	P	61	143	P	34	204	F	30	1607	P	30	1632	P
37	6250	P	37	1918	P	30	4710	P	32	6227	P	31	300	P	36	6360	P	32	7789	F
32	7722	F	40	4650	F	32	6335	P	46	9286	P	30	1697	P	30	7857	P	30	7791	P
30	7724	P	32	7818	F	30	7767	P	40	28460	P	32	7665	F	39	8143	R	31	7951	R
31	8000	R	30	7820	P	41	9364	P	67	38364	F	30	7667	P	47	8145	P	31	7951	C
31	8003	C	41	9451	P	32	9581	P	55	38385	F	41	9285	P	32	9854	F	41	9343	P
31	8005	F	32	28413	P	36	22225	R	32	42237	P	32	26640	F	32	26449	P	30	12726	R
30	8017	P	32	31688	P	32	28393	P	39	43079	F	30	27756	F	32	28606	P	30	13496	F
41	9328	P	30	35211	P	40	28531	P	39	43079	P	40	28440	P	36	28746	P	35	26645	F
31	13428	F	31	36093	R	44	31771	P	41	47097	F	37	31684	F	30	35271	P	34	26651	R
40	28424	P	30	36112	P	33	37136	P	41	47097	F	31	31695	F	30	36179	P	38	26659	R
30	35195	P	50	38476	F	50	38518	F	41	47097	F	30	35209	P	67	38506	F	40	28435	P
30	36096	P	52	38497	F	46	38545	F	77	47107	F	67	38475	F	55	38527	F	37	31660	F
67	38462	F	39	43461	F	39	43531	F	43	47128	F	53	38493	F	39	43577	F	30	31858	P
55	38483	F	39	43461	P	39	43531	P	34	47137	F	39	43559	F	39	43577	P	30	35137	P
39	43549	F	30	43473	F	32	47024	P	126	47145	F	39	43559	P	45	47825	F	30	35205	P
39	43549	P	30	43473	P	63	47597	F	50	47145	F	30	43571	F	31	47841	F	35	36093	P
30	43561	F	31	46532	F	52	47608	F	50	47145	F	30	43571	P	44	73321	P	67	38463	F
30	43561	P	32	46912	P	43	47617	F	39	47145	F	44	47643	F	45	88095	F	53	38481	F
35	46094	F	63	47486	F	77	47763	F	105	47174	F	69	47645	F	45	88095	P	39	43549	F
31	47673	F	56	47497	F	66	47763	F	177	47187	F	62	47652	F	45	88119	P	39	43549	P
45	47683	F	47	47506	F	55	47763	F	84	47187	F	54	47663	F	32	88141	F	30	43561	F
34	47694	F	37	47547	F	44	47763	F	83	47196	F	42	47672	F	32	88141	P	30	43561	P
53	47699	F	32	47676	F	33	47763	F	51	47196	F	40	47674	F	32	88162	P	31	47701	F
52	47723	F	31	50323	P	40	50494	P	80	47208	F	39	47736	F	37	97854	F	72	47781	F
31	47723	F	44	76637	P	38	73347	P	52	47208	F	54	47746	F	33	106759	F	38	47781	F
31	47744	F	69	88012	F	69	88270	F	39	47221	F	51	47757	F	33	106759	P	63	47790	F
46	47782	F	69	88012	P	69	88270	P	37	47251	F	55	47768	F	33	106791	P	31	47790	F
31	47800	F	41	88012	F	41	88270	F	127	47272	F	31	47780	F	37	119109	P	63	47801	F
30	47857	F	41	88012	P	41	88270	P	51	47272	F	35	47832	F	31	124553	R	40	50462	P
31	65471	R	69	88036	P	69	88294	P	128	47284	F	38	47932	F	31	124553	F	44	73874	P
34	71721	F	36	88048	F	36	88306	F	52	47284	F	40	50568	P	30	124553	R	30	77387	P
44	74394	P	36	88048	P	36	88306	P	39	47297	F	44	74127	P	30	124553	F	45	88657	F
45	89285	F	41	88060	P	41	88318	P	46	47327	F	45	88885	F	31	124553	R	45	88657	P
45	89285	P	36	88072	P	36	88330	P	47	47365	F	45	88885	P	31	124553	F	45	88681	P
45	89309	P	32	88082	F	32	88340	F	39	47373	F	45	88909	P	32	124555	R	32	88703	F
32	89331	F	32	88082	P	32	88340	P	46	47403	F	32	88931	F	31	124555	R	32	88703	P
32	89331	P	32	88103	P	32	88361	P	57	47449	F	32	88931	P	31	124555	F	32	88724	P
32	89352	P	37	97784	F	37	98048	F	37	47469	F	32	88952	P	30	124555	R	37	98427	F
37	99027	F	32	106888	F	32	107125	F	45	47523	F	37	98653	F	30	124555	F	33	107372	F
33	108174	F	32	106888	P	32	107125	P	36	47532	F	33	107598	F	31	124556	R	33	107372	P
33	108174	P	32	106920	P	32	107157	P	44	73321	P	33	107598	P	31	124556	R	33	107404	P
33	108206	P	37	119583	P	37	119811	P	41	87715	F	33	107630	P	30	124557	R	30	112842	R
45	114254	P	30	123034	F	30	123253	F	41	87715	P	30	113131	R	30	124557	R	30	112909	R
37	120941	P	32	130080	F	32	130327	F	41	87739	P	45	113678	P	31	124558	R	45	113404	P
30	125858	P	32	148897	F	32	149123	F	37	97467	F	37	120183	P	34	126927	P	37	119799	P

33	131386	F	69	148927	F	69	149153	F	45	112442	P	33	130591	F	33	129471	F	33	130156	F
32	150241	F	41	148931	F	41	149157	F	37	119135	P	32	149270	F	32	148101	F	32	148837	F
45	150271	F	35	148961	F	35	149187	F	41	148160	F	45	149300	F	45	148131	F	45	148867	F

Additional file 15 Table C3-4. Small simple repeat analysis for 12 newly assembled cp genomes

Cp Genome Repeat units	*RorSyl*	*CarHir*	*CarAlp*	*CarFle*	*RorAus*	*CarEnn*	*CarPen*	*LeaUni*	*LeaExi*	*CarAsa*	*CarTri*	*CarPra*
A/T	20	21	30	26	21	24	33	26	26	29	31	28
C/G										1		
AT/AT	6	7	6	5	4	4	5	7	6	4	7	5
AAG/CTT	1	1			1	1	1	2			1	
AAT/ATT	2	2	2	3	2	3	4		2	3	5	3
AAAT/ATTT	4	4	1	4	4	3	3	2	1	3	3	4
AATT/AATT	2	2		1	2	1	2	1		1	1	1
AGAT/ATCT	2	2	2	2	2	2	2	1	1	2	2	2
AAAC/GTTT		3	1	2		2	2	1	1	2	2	2
AAAG/CTTT		1	1	1		1				1	2	1
AAACG/CGTTT		1			1							
AAATAG/ATTTCT		2				1						
AATAG/ATTCT			2	2		2				2		2
AAGGAG/CCTTCT			2									
ACACT/AGTGT				1						1		
AATAT/ATATT					1				1			
AAATTT/AAATTT					1							
AAAGT/ACTTT						1			1			
AGCGAT/ATCGCT						2						
AAAGC/CTTTG							1	1				
AATATC/ATATTG								1				
AAATC/ATTTG											1	
AAAGG/CCTTT												1
sun	37	46	47	47	39	47	53	42	39	49	55	49

Additional file 16 Table C3-5.34 species selected for phylogeny analysis in tribe Cardamineae

Number	Abbreviation	Species Name
27	RorSyl	*Rorippa sylvestris*
28	CarHir	*Cardamine hirsuta*
29	CarAlp	*Cardamine alpina*
30	CarFle	*Cardamine flexuosa*
33	RorAus	*Rorippa austriaca*
34	CarEnn	*Cardamine enneaphyllos*
35	CarPen	*Cardamine pentaphyllos*
36	LeaUni	*Leavenworthia uniflora*
37	LeaExi	*Leavenworthia exigua*
80	CarAsa	*Cardamine asarifolia*
81	CarTri	*Cardamine trifolia*
82	CarPra	*Cardamine pratensis*
1	AetCor	*Aethionema cordifolium*
2	AetGra	*Aethionema grandiflorum*
3	AraTha	*Arabidopsis thaliana*
4	AraAlp	*Arabis alpina*
5	AraHir	*Arabis hirsuta*
6	BarVer	*Barbarea verna*
7	BraNap	*Brassica napus*
8	BraRap	*Brassica rapa subsp pekinensis*
9	CapBur	*Capsella bursa-pastoris*
10	CapGra	*Capsella grandiflora*
11	CarImp	*Cardamine impatiens*
12	CarRes	*Cardamine resedifolia*
13	CruWal	*Crucihimalaya wallichii*
14	DeaBou	*Descurainia bourgaeana*
15	DraNem	*Draba nemorosa*
16	LepVir	*Lepidium virginicum*
17	LobMar	*Lobularia maritima*
18	NasOff	*Nasturtium officinale*
19	OliPum	*Olimarabidopsis pumila*
20	PacChe	*Pachycladon cheesemanii*
21	PacEny	*Pachycladon enysii*
22	RapSat	*Raphanus sativus*

Additional file 17 Table C4-1.Summary of sampling for the final phylogeny analysis in Brassicaceae

Serial number		Species
CU.1	1	*Berteroa incana*
CU.2	2	*Alyssum alissoides*
CU.3	3	*Fibigia clypeata*
CU.4	4	*Matthiola fruticulosa*
CU.5	5	*Bunias orientalis*
CU.6	6	*Draba verna*
CU.7	7	*Draba dubia*
CU.8	8	*Arabis alpina*
CU.10	9	*Arabis hirsuta Aggreg*
CU.11	10	*Arabis nova*
CU.12	11	*Arabis soyeri subsp subcoriacea*
CU.13	12	*Arabis turrita*
CU.14	13	*Boechera gracilipes*
CU.15	14	*Phoenicaulis cheiranthoides*
CU.16	15	*Polyctenium fremontii*
CU.17	16	*Diplotaxis tenuifolia*
CU.18	17	*Brassica repanda susp baldensis*
CU.19	18	*Hirschfeldia incana*
CU.20	19	*Camelina microcarpa*
CU.21	20	*Capsella grandiflora*
CU.22	21	*Erysimum aurantiacum*
CU.23	22	*Erysimum rhaeticum*
CU.24	23	*Erysimum sylvestre*
CU.25	24	*Erysimum virgatum*
CU.26	25	*Neslia paniculata*
CU.27	26	*Rorippa sylvestris*
CU.28	27	*Cardamine hirsuta*
CU.29	28	*Cardamine alpina*
CU.30	29	*Cardamine flexuosa*
CU.33	30	*Rorippa austriaca*
CU.34	31	*Dentaria enneaphyllos*
CU.35	32	*Dentaria pentaphyllos*
CU.36	33	*Leavenworthia uniflora*
CU.37	34	*Leavenworthia exigua*
CU.38	35	*Cochlearia officinalis*
CU.39	36	*Descurainia bourgaeana*
CU.40	37	*Descurainia sofia*
CU.41	38	*Hornungia petraea*
CU.42	39	*Hutchinsia alpina*
CU.43	40	*Hutchinsia brevicaulis*
CU.44	41	*Hymenolobus pauciflorus*
CU.45	42	*Malcolmia littorea*
CU.46	43	*Morettia philaeana*
CU.47	44	*Thellungiella halophila*
CU.48	45	*Halimolobos pubens*
CU.49	46	*Heliophila coronopifolia*
CU.50	47	*Hesperis matronalis*
CU.51	48	*Iberis amara*
CU.52	49	*Isatis tinctoria*
CU.53	50	*Lepidium campestris*
CU.54	51	*Cardaria draba*
CU.55	52	*Noccaea precox*
CU.56	53	*Noccaea rotundifolium*
CU.57	54	*Lesquerella montana*
CU.58	55	*Nerisyrenia camporum*
CU.59	56	*Stanleya pinnata*
CU.60	57	*Thelypodium laciniatum*
CU.61	58	*Ochthodium aegyptiacum*
CU.62	59	*Sisymbrium officinale*
CU.63	60	*Smelowskia calycina*
CU.64	61	*Thlaspi perfoliatum*
CU.65	62	*Peltaria angustifolia*
CU.68	63	*Biscutella laevigata*
CU.69	64	*Biscutella prealpina*
CU.70	65	*Calepina irregularis*
CU.71	66	*Kernera saxatilis*
CU.72	67	*Lunaria annua*
CU.74	68	*Cleome spynosa*
CU.75	69	*Cleome hirta*
CU.76	70	*Alyssum dasycarpum*
CU.77	71	*Draba aizoides*
CU.78	72	*Turritis glabra*
CU.79	73	*Cardamine pentaphyllos*
CU.80	74	*Cardamine asarifolia*
CU.81	75	*Cardamine Trifolia*
CU.82	76	*Cardamine pratensis*
CU.84	77	*Aethionema saxatile*
CU.86	78	*Arabidopsis halleri*
CarImp	79	*Cardamine impatiens*
CarRes	80	*Cardamine resedifolia*
AetCor	81	*Aethionema cordifolium*
AetGra	82	*Aethionema grandiflorum*
AraTha	83	*Arabidopsis thaliana*
AraHir	84	*Arabis hirsuta*
BarVer	85	*Barbarea verna*
BraNap	86	*Brassica napus*
CapBur	87	*Capsella bursa-pastoris*
CruWal	88	*Crucihimalaya wallichii*
DraNem	89	*Draba nemorosa*
LepVir	90	*Lepidium virginicum*
LobMar	91	*Lobularia maritima*

NasOff	92	*Nasturtium officinale*
OliPum	93	*Olimarabidopsis pumila*
PacEny	94	*Pachycladon enysii*
PacChe	95	*Pachycladon cheesemanii*