LBNL-59386

A Comparison of the First Two Sequenced Chloroplast Genomes in Asteraceae: Lettuce and Sunflower

Ruth E. Timme[1§], Jennifer V. Kuehl[2], Jeffrey L. Boore[2,3,4], Robert K. Jansen[1]

[1]Section of Integrative Biology and Institute of Cellular and Molecular Biology, University of Texas at Austin, Austin, TX.

[2]DOE Joint Genome Institute and Lawrence Berkeley National Laboratory, Walnut Creek, CA

[3]Department of Integrative Biology, University of California, Berkeley, CA

[4]Genome Project Solutions, Hercules, CA

[§]Corresponding author

Email addresses:

RET: mailto:retimme@mail.utexas.edu

JVK: jvkuehl@lbl.gov

JLB: jlboore@calmail.berkeley.edu

RKJ: jansen@mail.utexas.edu

Background

Asteraceae is the second largest family of plants, with over 20,000 species. For the past few decades, numerous phylogenetic studies have contributed to our understanding of the evolutionary relationships within this family, including comparisons of the fast evolving chloroplast gene, *ndhF* [2], *rbcL*, as well as non-coding DNA from the *trnL* intron plus the *trnL-trnF* intergenic spacer [3, 4], *matK* [5], and, with lesser resolution, *psbA-trnH* [6]. This culminated in a study by Panero and Funk in 2002 [1] that used over 13,000 bp per taxon for the largest taxonomic revision of Asteraceae in over a hundred years. Still, some uncertainties remain, and it would be very useful to have more information on the relative rates of sequence evolution among various genes and on genome structure as a potential set of phylogenetic characters to help guide future phylogenetic structures.

By way of contributing to this, we report the first two complete chloroplast genome sequences from members of the Asteraceae, those of *Helianthus annuus* and *Lactuca sativa*. These plants belong to two distantly related subfamilies, Asteroideae and Cichorioideae, respectively [1]. In addition to these, there is only one other published chloroplast genome sequence for any plant within the larger group called Eusterids II, that of *Panax ginseng* (Araliaceae, 156,318 bps, AY582139).

Early chloroplast genome mapping studies demonstrated that *H. annuus* and *L. sativa* share a 22 kb inversion relative to members of the subfamily Barnadesioideae [7-9]. By comparison to outgroups, this inversion was shown to be derived, indicating that the Asteroideae and

Cichorioideae are more closely related than either is to the Barnadesioideae. Later sequencing study found that taxa that share this 22 kb inversion also contain within this region a second, smaller, 3.3 kb inversion [10].

These sequences also enable an analysis of patterns of shared repeats in the genomes at fine level and of RNA editing by comparison to available EST sequences. In addition, since both of these genomes are crop plants, their complete genome sequence will facilitate development of chloroplast genetic engineering technology, as in recent studies from Daniell's lab [11-15]. Knowing the exact sequence from spacer regions is crucial for introducing transgenes into the chloroplast genome [14].

Results

Size, gene content, order and organization: lettuce and sunflower cp genomes

The lettuce chloroplast genome is 152,772 bp in length (Fig 1) and contains a pair of inverted repeats (IRs) of 25,034 bp each, separated by a large and small single copy region (LSC and SSC) of 84,105 and 18,599 bp, respectively. The sunflower chloroplast genome is 151,104 bp in length, with IRs of 24,633 each, separated by an LSC of 83,530 bp and a SSC of 18,308. The G+C content of both sunflower and lettuce is 38% across the whole cpDNA. Gene content and arrangement are identical in both cpDNAs, but the order is different from tobacco, which has one large and one small inversion relative to these. There are 81 unique protein-coding genes in both genomes, seven of which are duplicated in the IR. The four rRNA genes are contained completely within the IR, so they are doubled in the genome. There are 29 unique tRNA genes, of which seven are in the IR, bringing the total number to 35 in the genome. There are 17 unique intron-containing genes; 15 genes have a single intron and two genes have two introns.

Sequence divergence

The p-distance for the 25 most divergent non-coding regions of cpDNA is listed in Table 1, with values ranging from 0.084 to 0.226. Figure 2 shows the average p-distance for four classes of genomic regions: protein coding genes, introns, intergenic spacers, and RNAs (both rRNA and tRNAs). The intergenic spacer divergence is almost double the next highest class (introns). RNAs hold the lowest sequence divergence, at an average of only 0.8 percent. Table 2 shows the 10 most divergent protein coding sequences, ranging from 0.102 to 0.036. Sequence divergence across the whole genome of *Helianthus* and *Lactuca* is graphically summarized in Fig 3 by a percent identity plot. Tobacco is included for comparison and the annotation from *Helianthus* was used for gene locations.

Repeat Analysis

Because the raw REPuter [16] output contains many redundant repeats, we used the filtering program Comparative Repeat Analysis (CRA), which identifies and excludes repeats that are contained entirely within other repeats. CRA also identifies shared repeats by similarity searching using BLAST to the repeats of other input genomes. The output of the CRA analysis is found in Fig 4a. Most of the repeats are less then 40 bp, with only two larger than 90 bp. Only repeats that are 23 bp or larger were examined by eye for both *Helianthus* and *Lactuca*. Since we are interested in the role of repeats in genome organization, we attempted to categorize these repeats, and arrived at seven classes: (1) Tandem repeats > 2x are identified by CRA as one large repeat, but upon closer examination are actually 3+ smaller repeats back-to-back; (2) Direct repeats dispersed in the genome; (3) Repeats found in reverse complement orientation dispersed in the genome; (4) Hairpin loops with a predicted 2º structure based on mfold [17]; (5) Tandem repeat 2x are true tandem repeats, or any twice repeated sequence back-to-back; (6) Repeats of runs of A's or T's, usually in excess of 12 bp; and (7) Repeats of portions of RNA or protein encoding genes.

Figure 4c shows the histogram of frequency of repeats according to these categories. Figs 4c and 4d describe the "true" repeats, i.e., the collection of repeats excluding polyA, polyT, and gene similarity matches. In Fig 4d, the graph shows the number of repeats shared by tobacco and both Asteraceae genomes, shared only among Asteraceae genomes, and repeats that are unique to *Helianthus* or *Lactuca*. Most of these repeats are located in intergenic spacers, as shown in Fig 4e. A table with specific repeat information is located in the Supplemental Materials.

Variation between coding sequences and cDNAs

ESTs are available in the databases for only a few of the relevant chloroplast genes. Only one lettuce chloroplast gene was present in the EST database and it is a perfect match to the genomic sequence. There were 10 ESTs of *Helianthus* chloroplast genes available and the differences are summarized in Table 3. There are three C-to-U changes, which are thought to be conventional angiosperm RNA editing changes [18]. These changes occur in *psbC*, *psbZ*, and *ndhA*, none of which are homologous to known RNA editing sites in tobacco cpDNA [18]. Of note is a C-to-U change that causes a stop codon in *psbC*. Two insertions of a single nucleotide also are noted. In both cases they cause frame shifts that result in numerous stop codons.

Discussion

Genome organization

Although the sunflower and lettuce chloroplast genomes are identical in gene content and arrangement, they differ in their length and in the extent of their IR regions. The lettuce IR is 401 bp longer than the sunflower IR, which adds twice that length to the whole genome. Although the lettuce genome IR is longer, the sunflower IR actually has a greater expansion along both edges by a total of 140 bp. The sunflower IR extends into *ycf1* with 576 bp (only 471 in sunflower) and into *rps19* with 101 bp (only 60 bp in lettuce). This expansion of boundaries in the sunflower IR is balanced by a deletion of 456 bp in *ycf2* in the sunflower. Other smaller indels across the IR add to this length difference. Since this 401 bp IR length difference is doubled in the genome, it contributes half of the total genome difference, which is 1,668 bp between the two genomes. The extent of the IR in both genomes is similar, although the exact extent into the single-copy genes varies among other published genomes, like *Glycine*, *Nicotiana*, *Atropa*, *Eucalyptus*, and *Panax* [11, 12, 19-21].

There is a 152 amino-acid (aa) deletion in the *ycf2* gene in sunflower. *ycf2* is one of five genes absent in some species' chloroplast genomes; the other genes are *accD*, *ycf1*, *rpl23, infA* [22]. Both *ycf1* and *ycf2* are absent in monocot grains, namely maize, rice, and sugarcane [23-25]. However, knockout studies of *ycf2* have confirmed it as an essential chloroplast gene for survival in tobacco [26]. From this study, we can only hypothesize that the *ycf2* gene in *Helianthus* is functional because the rest of the gene is highly conserved compared to the lettuce copy, with only 1.31% sequence divergence. If the large deletion in the *Helianthus* copy rendered it a pseudogene, we would expect there to be higher sequence divergence with internal stop codons.

The start codon in *accD* gene occurs 15 aa further into the gene than it does in *Lactuca*, a position that matches the annotation in *Lotus* and *Arabidopsis*. *Lactuca* also has a 25 aa insertion in the middle of the *accD* gene. Like with *ycf2*, we assume the gene is still functional since the sequence divergence is otherwise low across the rest of the gene. There are a few other instances where the lengths of genes differ (*matK, rbcL, rpl22, rpl33, rpoC2, ycf1, ycf15*), but the majority of genes between *Helianthus* and *Lactuca* have no indels. The tRNAs are even lower in indel events: one involves a five bp deletion in *trnS*-UGA that is shared between *Helianthus* and *Lactuca*, and two others occur between the two genomes (a one bp indel in both *trnV*-UAC and *trnI*-GAU). We assume that these events do not affect the tRNA function for the same reasons as above.

The differences between DNA sequence and the mRNA sequence (from the EST database) result in many polymorphisms that cause amino-acid changes. Since the EST library only overlaps in a small subset of genes, we cannot draw conclusions about RNA editing across the whole chloroplast genome. Three of these are C-to-U edits, which are possible RNA edited changes [27]. The C-to-U changes in *Helianthus* occur in *ndhA, psbC,* and *psbZ.* Two of these cause an amino-acid change and the third cause change to a stop codon (Table 3). None of these edits correspond to other published, conserved, angiosperm editing sites, but although edited sites can be shared among distantly related taxa [18] others have been acquired later in angiosperm evolution [28]. The other 17 polymorphisms are interesting because the ESTs were made from the exact same strain of plant as was used in the chloroplast genome sequencing. These differences could be due either to polymorphisms or low quality sequence in the ESTs (our stringent phred-phrap requirement across the genomic sequence makes it very unlikely that low quality sequence could be present in the genomic sequence). Lee et al. showed a similar pattern of intra-species polymorphism between DNA and EST sequences in cotton [12], and in their case, none of the polymorphisms were C-to-U edits.

Evolutionary Implications

Past analyses of repeated sequence in chloroplast genomes have focused primarily on Simple Sequence Repeats or SSRs [29-31], which are useful for population level studies. Larger and more complex repeats have been associated with rearranged genomes [10, 11, 32], but tools for identifying and summarizing these more complex repeats have been incomplete in the past. There are two issues that skew repeat results when using the program REPuter [16]. One is the use of hamming-distance (HD) as a measure of determining similarity of repeating sequence. This is a fixed parameter that only allows one user-defined number of differences per repeat, which is the same regardless of length. In effect, this skews the number of smaller repeats found in the genome since a greater percentage of differences for smaller repeats is allowed. Our solution to this problem was to use a sliding window method of increasing the HD with the repeat size, i.e. 2 HD for repeats 21-30 bp, 3 HD for 31-40 bp, 4 HD for 41-50 bp, etc. The second issue with REPuter is that it doesn’t recognize repeats contained within other repeats, which drastically overestimates the number of repeats found in the genome. Stacia Wyman’s Comparative Repeat Analysis (CRA) addresses this second issue by sifting through the REPuter output and excluding repeats contained within others. This more accurate collection of repeats is summarized in Fig 4a.

In order to examine the evolutionary or functional significance of repeats they must be categorized by form. For instance, we assume that a hairpin loop has different evolutionary implications than does tRNA similarity. After manually examining each repeat above 22 bp, we established seven categories (Figure 4b). The four largest repeats from Fig 4a were actually composed of smaller tandem repeats, so were re-categorized as such and summarized in Fig 4c. Two of our categories are not considered 'real' repeats for our purposes: gene similarity and tRNA repeats provide evidence of gene duplication, which is shared among most land plants and poly-A and poly-T runs are actually SSRs. Figure 4d omits these former categories of repeats and identifies which of these remaining ones are shared and unique among the genomes. The four repeats that are shared among *Helianthus*, *Lactuca*, and *Nicotiana* are as follows: a 32 bp tandem repeat in the *rrn4.5-rrn5* spacer; a 42 bp sequence that is dispersed in the 2nd intron of *ycf3*, *ndhA* intron, and the *rps12-ycf15* intergenic spacer; a palindrome in the *accD-psbI* spacer, and another palindrome in the *trnT-trnL* spacer. Finally, most of the repeats are found in non-coding DNA. The greater number of repeats present in spacers vs. introns is likely a function of fewer intron sequences in the genome and not an actual bias for spacers over introns. Since repeats have been implicated in the rearrangement of chloroplast genomes, we looked for them at the our three rearrangement endpoints (Table 4). The 31 bp repeat at positions 12,333 and 31,010 in the *Helianthus* genome is close to two of the 2nd and 3rd rearrangement endpoints. The copy at coordinate 31,010 is in the *trnG* intron, which is only173 bp from the 3rd endpoint. But none of the other repeats stand out as being correlated with the rearrangement. Our analysis only looked at repeats of 21 bp and larger, so a further examination of smaller repeats might reveal a higher density of repeats in this area. All repeats examined in this study are listed in the supplementary data (Sup. ##).

Perhaps the most directly practical data to emerge from this analysis is the identification of new genomic regions for use in phylogenetic studies. The study of evolutionary relationships in Asteraceae has ballooned over the past 15 years, with studies focused both within and between the genera. Most studies use a combination of several chloroplast regions and one or two nuclear genes. Usually the chloroplast regions used have a lower rate of evolution than the nuclear DNA, so more sequencing is needed to achieve equivalent resolution. Panero and Crozier [33] reviewed the phylogenetic utility of different chloroplast regions specifically for Asteraceae, while a more recent Shaw et al. [34] review covered phylogenetic utility of cpDNA across flowering plants. Interestingly, these only partially overlap with our results. We listed the 25 most divergent regions between *Helianthus* and *Lactuca* in Table 1 along with their length and number of indels. Our p-distance measure excludes any position containing a gap, so indels are not included in the divergence calculation. Of these most divergent regions longer than 300 bp, 11 have not been widely utilized, if ever, for phylogenetic inference. This is a promising finding, since plant systematists are constantly searching for more variable chloroplast sequences for resolving species level relationships. Three of these regions are currently being utilized for phylogenetic analyses in *Helianthus* (RET unpublished) and the primer sequences used to amplify these regions are listed in Table 6.

Methods

Chloroplast isolation, amplification, and sequencing

Fresh leaf material from lettuce (*Lactuca sativa* strain Salinas) and sunflower (*Helianthus annuus* line HA383) was used for the chloroplast isolation. These strains are the same ones used in the EST and nuclear genome sequencing efforts of the Compositae Genome Project [35]. Chloroplasts were isolated from the fresh leaves by the sucrose-gradient method [36]. They were then lysed and amplified using the REPLI-g™ whole genome amplification kit (Molecular Staging). The product was then digested with *Eco*RI and *Bst*BI and the clear banding pattern ensured that amplification product was indeed chloroplast and not nuclear DNA. A detailed description of these steps is outlined in Jansen et al. [37]. Purified cpDNA was sheared by serial passage through a narrow aperture using a Hydroshear device (Gene Machines), then these fragments were enzymatically repaired to blunt ends and gel purified, then ligated into pUC18 plasmids. These clones were introduced into *E. coli* by electroporation, plated onto nutrient agar with antibiotic selection, and grown overnight. Colonies were randomly selected and robotically processed through rolling circle amplification of plasmid clones, sequencing reactions using BigDye chemistry (Applied Biosystems), reaction cleanup using solid-phase reversible immobilization, and sequencing determination using an ABI 3730 XL automated DNA sequencer. Detailed protocols are available at http://www.jgi.doe.gov/sequencing/protocols/protsproduction.html.

Genome Assembly and Annotation

Sequences from randomly chosen clones were processed using PHRED and assembled based on overlapping sequence into a draft genome sequence using PHRAP [38]. Quality of sequence determination and assembly were verified by eye using the program Consed [39]. PCR and sequencing at the University of Texas at Austin were used to bridge gaps and mend low quality areas of the genome. Additional sequences were added until a completely contiguous consensus was created representing the entire cpDNA. Throughout the entire consensus, we verified that all regions had a quality of Q40 or greater and including at least two sequencing reads. For both lettuce and sunflower, most of the genome far exceeds these minimum requirements.

The beginning of each genome was standardized for gene annotation to be the first bp after the IRa. (In this case both started right before *trnH.*) The program DOGMA (Dual Organellar GenoMe Annotator, [40]) was used to assist in fully annotating all genes, identifying coding sequence, rRNAs, and tRNAs using the plastid/bacterial genetic code.

Calculating sequence divergence

The whole genome sequence and annotation of lettuce and sunflower were compared to the reference genome, tobacco, by a Percent Identity Plot (PIP) produced by the program MultiPipMaker [41]. The individual genes, rRNAs, tRNAs, introns, and intergenic spacers were also exported from both genomes in DOGMA and aligned by hand in MacClade [42] for a more detailed quantification of sequence divergence. Since we are only comparing two genomes, we quantified sequence divergence as the proportion (*p*) of aligned nucleotide sites within a specified region that are different (p-distance). A perl script was written to call PAUP [43] on each nexus file, calculate the p-distance between each region, and write out to a tab-delimited file.

Examination of repeat structure

Shared and unique repeats were characterized for both lettuce and sunflower genomes and compared to the reference genome of tobacco (*Nicotiana tabacum*) using Comparative Repeat Analysis (CPA) [44]. This program filters the redundant output of REPuter [16] and identifies shared repeats among the input genomes. For repeat identification, the following constraints were set in CPA: (i) minimum repeat size of 21 bp, and (ii) 90% or greater sequence identity for each 10 bp bin (i.e. hamming distance (HD) was set to 2 for 21-30 bp, HD = 3 for 31-40, HD = 4 for 41-50 etc., until no further repeats were found). All repeats above 22 bp were examined by eye and placed into author-defined repeat categories.

Variation between coding sequences and cDNAs

Expressed sequence tag (EST) databases for both lettuce and sunflower were downloaded from the Compositae Genome Project Database (CGPDB). The complete set of coding sequences from our direct sequencing of lettuce and sunflower were searched for similarity by BLAST against their respective EST database. Significant hits were examined by eye for base-pair differences and summarized in a table as possible RNA edited sites.

Authors' contributions

RET extracted and isolated chloroplasts from *Helianthus* and *Lactuca*, performed amplification of cpDNA, made the genomic library at the DOE Joint Genome Institute (JGI), performed PCR for closing gaps and improving low quality regions across the genome, annotated both genomes, performed all analyses and drafted this manuscript; JVK and JLB performed the genome sequencing and draft assembly. JLB also assisted with writing of the manuscript; RKJ assisted with wet chloroplast isolations, annotation, analyses, and revised several versions of this manuscript.

Acknowledgements

Leaf material was provided by Steve Knapp (*Helianthus annuus*) and Richard Michelmore (*Lactuca sativa*). Zhengqiu Cai assisted with the perl scripts. This research was supported in part by a grant from the National Science Foundation (DEB 0120709) and the Sidney F. and Doris Blake Centennial Professorship in Systematic Botany to RKJ. Part of this work was performed under the auspices of the U.S. Department of Energy, Office of Biological and Environmental Research, by the University of California, Lawrence Berkeley National Laboratory, under contract No. DE-AC02-05CH11231.

References

1. Panero JL, Funk VA: **Toward a phylogenetic subfamilial classification for the Compositae (Asteraceae)**. *Proceedings Of The Biological Society Of Washington* 2002, **115**(4):909-922.
2. Kim K-J, Jansen RK: **ndhF sequence evolution and the major clades in the sunflower family**. *Proceedings of the National Academy of Sciences of the United States of America* 1995, **92**(22):10379-10383.
3. Bayer RJ, Starr JR: **Tribal phylogeny of the Asteraceae based on two non-coding chloroplast sequences, the trnL intron and trnL/trnF intergenic spacer**. *Annals Of The Missouri Botanical Garden* 1998, **85**(2):242-256.
4. Jansen RK, Kim KJ: **Implications of chloroplast DNA data for the classifications and phylogeny of the Asteraceae**. In: *Compositae: Systematics, Proceedings of the International Compositae Conference: 1994; Kew*: Royal Botanic Gardens, Kew; 1996: 317-339.
5. Denda T, Watanabe K, Kosuge K, Yahara T, Ito M: **Molecular phylogeny of Brachycome (Asteraceae)**. *Plant Systematics and Evolution* 1999, **217**(3-4):299-311.
6. Kim SC, Crawford DJ, Jansen RK, Santos-Guerra A: **The use of a non-coding region of chloroplast DNA in phylogenetic studies of the subtribe Sonchinae (Asteraceae: Lactuceae)**. *Plant Systematics and Evolution* 1999, **215**(1-4):85-99.
7. Jansen RK, Palmer JD: **Chloroplast DNA from lettuce and *Barnadesia* (Asteraceae): structure, gene localization, and characterization of a large inversion.** *Current Genetics* 1987, **11**:553-564.

8. Heyraud F, Serror P, Kuntz M, Steinmetz A, Hieizmann P: **Physical map and gene localization on sunflower (*Helianthus annuus*) chloroplast DNA: evidence for an inversion of a 32.5-kbp segment in the large single copy region.** *Plant Molecular Biology* 1987, **9**:485-496.

9. Jansen RK, Palmer JD: **A chloroplast DNA inversion marks an ancient evolutionary split in the sunflower family (Asteraceae)**. *Proceedings of the National Academy of Sciences* 1987, **84**:5818-5822.

10. Kim H-G, Choi K-S, Jansen RK: **Two chloroplast DNA inversions originated simultaneously during the early evolution of the sunflower family (Asteraceae)**. *Molecular Biology and Evolution* 2005, **22**:1-10.

11. Saski C, Lee S, Daniell H, Wood T, Tomkins J, Kim H-G, Jansen RK: **Complete chloroplast genome sequence of *Glycine max* and comparative analyses with other legume genomes.** *Plant Molecular Biology* 2005, **59**:309-322.

12. Lee S-B, Jansen RK, Hostetler JB, Tallon LJ, Town CD, Daniell H: **The complete chloroplast genome sequence of *Gossypium hirstum*: organization and phylogenetic relationships to other angiosperms**. *BMC Genomics* in press.

13. Ruiz O, Daniell H: **Engineering cytoplasmic male sterility via the chloroplast genome**. *Plant Physiology* 2005, **138**:1232-1246.

14. Daniell H, Kumar S, Ruiz O: **Breakthrough in chloroplast genetic engineering of agronomically important crops**. *Trends in Biotechnology* 2005, **23**:238-245.

15. Daniell H, Dhimgra A, Ruiz O: **Chloroplast genetic engineering to confer desired plant traits.** *Methods in Molecular Biology* 2004, **286**(111-137).

16. 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 Research* 2001, **29**:4633-4642.

17. Zuker M: **Mfold web server for nucleic acid folding and hybridization prediction**. *Nucleic Acids Research* 2003, **31**(13):3406-3415.

18. Hirose T, Kusumegi T, Tsudzuki T, Sugiura M: **RNA editing sites in tobacco chloroplast transcripts: editing as a possible regulator of chloroplast RNA polymerase activity**. *Molecular Biology and Evolution* 1999, **262**:462-467.

19. Kim KJ, Lee HL: **Complete chloroplast genome sequences from Korean ginseng (*Panax schinseng* Nees) and comparative analysis of sequence evolution among 17 vascular plants**. *DNA Research* 2004, **11**(4):247-261.

20. Wakasugi T, Sugita M, Tsudzuki T, Sugiura M: **Updated gene map of tobacco chloroplast DNA**. *Plant Molecular Biology Reporter* 1998, **16**:231-241.

21. Steane DA: **Complete nucleotide sequence of the chloroplast genome from Tasmanian Blue Gum, *Eucalyptus globulus* (Myrtaceae)**. *DNA Research* 2005, **12**(3):215-220.

22. Millen RS, Olmstead RG, Adams KL, Palmer JD, Lao NT, Heggie L, Kavanagh TA, Hibberd JM, Gray JC, Morden CW *et al*: **Many parallel losses of *infA* from chloroplast DNA during angiosperm evolution with multiple independent transfers to the nucleus**. *Plant Cell* 2001, **13**(3):645-658.

23. Asano T, Tsudzuki T, Takahashi S, Shimada H, Kadowaki K: **Complete nucleotide sequence of the sugarcane (*Saccharum officinarum*) chloroplast genome: A**

comparative analysis of four monocot chloroplast genomes. *DNA Research* 2004, **11**(2):93-99.

24. Matsuoka Y, Yamazaki Y, Ogihara Y, Tsunewaki K: **Whole chloroplast genome comparison of rice, maize, and wheat: Implications for chloroplast gene diversification and phylogeny of cereals**. *Molecular Biology and Evolution* 2002, **19**(12):2084-2091.

25. Maier RM, Neckermann K, Igloi GL, Kossel H: **Complete sequence of the maize chloroplast genome: gene content, hotspots of divergence and fine tuning of genetic information by transcript editing.** *Journal of Molecular Biology* 1995, **251**:614-628.

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

27. Gott JM, Emeson RB: **Functions and mechanisms of RNA editing**. *Annual Review Of Genetics* 2000, **34**:499-531.

28. Tsudzuki T, Wakasugi T, Sugiura M: **Comparative analysis of RNA editing sites in higher plant chloroplasts**. *Journal Of Molecular Evolution* 2001, **53**(4-5):327-332.

29. Provan J, Powell W, Hollingsworth PM: **Chloroplast microsatellites: new tools for studies in plant ecology and evolution**. *TRENDS in Ecology and Evolution* 2001, **16**(3):142-148.

30. Powell W, Morgante M, McDevitt R, Vendramin GG, Rafalski JA: **Polymorphic simple sequence repeat regions in chloroplast genomes: applications to the population genetics of pines**. *Proceedings of the National Academy of Sciences* 1995, **92**(17):7759-7763.

31. Marshall HD, Newton C, Ritland K: **Sequence-repeat polymorphisms exhibit the signature of recombination in Lodgepole Pine chloroplast DNA**. *Molecular Biology and Evolution* 2001, **18**(11):2136-2138.

32. Hupfer H, Swaitek M, Hornung S, Herrmann RG, Maier RM, Chiu WL, Sears B: **Complete nucleotide sequence of the *Oenothera elata* plastid chromosome, representing plastome1 of the five distinguishable *Euoenthera* plastomes.** *Molecular Genomics and Genetics* 2000, **263**:581-585.

33. Panero JL, Crozier BS: **Primers for PCR amplification of Asteraceae chloroplast DNA**. *Lundellia* 2003, **6**:1-9.

34. Shaw J, Lickey EB, Beck JT, Farmer SB, Liu W, Miller J, Siripun KC, Winder CT, Schilling EE, Small RL: **The tortoise and the hare II: relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis**. *American Journal of Botany* 2005, **92**(1):142-166.

35. Michelmore R, Knapp SJ, Bradford KJ, Rieseberg LH, Jackson LE, Kesseli RV: **Compositae Genome Project Database**. In.: http://cgpdb.ucdavis.edu/sitemap.html; 2006.

36. Palmer JD: **Isolation and structural analysis of chloroplast DNA**. In: *Methods in Enzymology*. Edited by Weissbach A, Weissbach H, vol. 118. New York: Academic Press; 1986: 167-186.

37. Jansen RK, Raubeson LA, Boore JL, dePamphilis CW, Chumley TW, Haberle RC, Wyman SK, Alverson AJ, Peery R, Herman SJ *et al*: **Methods for obtaining and analyzing whole chloroplast genome sequences**. In: *Methods in Enzymology*. vol. 395; 2005: 348-384.

38. Ewing B, Green P: **Base-calling of automated sequencer traces using phred. II Error probabilities**. *Genome Research* 1998, **8**:186-194.

39. Gordon D, Abajian C, Green P: **Consed: a graphical tool for sequence finishing**. *Genome Research* 1998, **8**:195-202.

40. Wyman SK, Boore JL, Jansen RK: **Automatic annotation of organellar genomes with DOGMA**. *Bioinformatics* 2004, **20**(3252-3255).

41. Schwartz S, Z. Z, Frazer KA, Smit A, Riemer C, Bouck J, Gibbs R, Hardison R, Miller W: **PipMaker -- A web server for aligning two genomic DNA sequences**. *Genome Research* 2000, **10**(4):577-586.

42. Maddison DR, Maddison WP: **MacClade: analysis of phylogeny and character evolution**. In., 4.05 edn: Sunderland: Sinauer Asociates; 2002.

43. Swofford DL: **PAUP*: phylogenetic analysis using parsimony (* and other methods)**. In., 4.0b 10 edn. Sunderland, Massachusetts.: Sinauer Associates, Inc.; 2003.

44. Wyman SK, Jansen RK: **Comparative Repeat Analysis**. *http://bugmasterjgi-psforg/repeats/* in prep.

Table 1. The 25 genomic regions with the largest p-distance between *Lactuca* and *Helianthus* genomes. Checked are the regions that have been highlighted in other review papers: Shaw at al. [34] (see their Fig. 5) and Panero and Crozier [33]. Note: Phylogenetic utility is not directly correlated with p-distance.

non-coding	**length**	**p-distance**	**# indels**	**Shaw et al. 2005**	**Panero and Crozier 2003**
trnE-rpoB	1003	0.226	184		
trnY-trnE	169	0.223	66	✓partial	
trnL-rpl32	925	0.214	238		
5'rps12-clpP	172	0.166	17		
ndhC-trnV	1174	0.130	395		
trnH-psbA	418	0.117	58	✓	
trnL-trnF	361	0.109	29	✓	✓
trnR-trnG	203	0.107	16	✓	
trnD-trnY	113	0.106	0	✓partial	
trnC-petN	870	0.106	109		
trnT-trnL	571	0.100	45	✓	
ycf1-rps15	599	0.100	234		
ndhD-ccsA	303	0.098	41		
rpl32-ndhF	1083	0.096	258		
psbI-trnS	156	0.096	16		
ycf3-trnS	948	0.095	181		
3'trnK-matK	285	0.094	20	✓	✓
ndhI-ndhG	395	0.093	86		✓

trnG-trnT	171	0.091	31		
petN-psbM	549	0.088	88		
rps16-trnQ	1021	0.088	49		
trnG-trnfM	213	0.087	30	✓partial	
rpl36-infA	122	0.087	8		
psbZ-trnG	333	0.086	35	✓partial	
trnM-atpE	213	0.084	13		

Table 2. The 10 most divergent coding regions between *Lactuca* and *Helianthus* genomes.

Genes	**Length**	**p-distance**
ycf1	5343	0.102
psbT	102	0.059
petL	96	0.052
ndhF	2235	0.049
ccsA	969	0.047
psbH	222	0.045
matK	1521	0.043
accD	1593	0.042
rps15	279	0.039
rpl32	165	0.036

Table 3. Base pair differences between genomic sequences and processed mRNA in the form of expressed sequence tags (ESTs) for *Helianthus* and *Lactuca*.

Gene name	**EST contig Name**	**gene length**	**EST overlap**	**position**	**bp change**	**AA genome**	**AA cDNA**
Helianthus							
atpA	QHL20C13.yg.ab1	1527	573	648	G-->A	K(aaG)	K(aaA)
atpB	QHL3D12.yg.ab1	1497	394	44 bp diffs - only 87% similar			
ndhH	QH_CA_Contig4939	1182	421	no bp changes			
ndhA	QH_CA_Contig4939	1092	553	107	**C-->U**	P(cCt)	L(cTt)
psbC	QH_CA_Contig1086	1422	513	786	U-->C	R(cgT)	R(cgC)
				896	G-->C	S(aGt)	T(aCt)
				901	c insert	causes frameshift	
				956	U-->C	V(gTt)	A(gCt)
				959	G-->C	R(aGa)	T(aCa)
				1096, 1097	U-->C	L(TTa)	P(CCa)
				1131	c insert	causes frameshift	
				1162	**C-->U**	Q(Caa)	STOP(Taa)
psbI	QHM17F18.yg.ab1	111	111	44	U-->C	L(tTt)	S(tCt)
				67	U-->C	F(Ttc)	L(Ctc)
psbZ	QH_CA_Contig5109	189	189	50	**C-->U**	S(tCa)	L(tTa)
				74	U-->C	V(gTt)	A(gCt)
				152	U-->G	V(gTc)	G(gGc)
				165	G-->A	G(gGt)	D(gAt)
rpl2	QHK5F11.yg.ab1	825	394	816	U-->G	R(cgT)	R(cgG)
				819	U-->G	S(agT)	R(agG)
rpl14	QH_CA_Contig6111	369	175	no bp changes			

rpl16	QH_CA_Contig5085	411	411	44	G-->C	R(aGa)	T(aCa)
				186	U-->C	G(ggT)	G(ggC)
				247	c insert	causes frameshift	
				256	U-->C	G(ggT)	G(ggC)
				261	U-->C	G(ggT)	G(ggC)
Lactuca							
ycf4	QGH3h12.yg.ab1	555	193	no bp changes			

Table 4. Repeats that are in close approximation to genome rearrangement – locations stated for *Helianthus* genome based on estimated rearrangement endpoints, from Lee et al. [10].

Rearrangement location in *Helianthus*	**repeat location**	**repeat type**	**length of repeat**
8897-8901, *trnS-trnC*	8850 8891	palindrome polyA run	30 (13bp stem) 21
	9048, 66394	dispersed repeat	21
12183-12693, *trnE-rpoB*	12333, 31010	dispersed repeat	31
	12094, 47516	dispersed palindrome	21 (5bp stem)
	12431, 47612	dispersed repeat	21
31183, *trnG-trnT*	none		

Table 5. Primer sequences for three chloroplast spacer regions – Primers amplify across *Helianthus* and *Lactuca*. *internal sequencing primer.

cp DNA region	length	Primer name	5'- 3' primer sequence
ndhC-trnV	1248	ndhCretF	AAGTTTCTCCGGTCCTTTGC
		trnVretR	TCTACGGTTCGAGTCCGTATAG
trnL-rpl32	998	trnLretF	TACCGATTTCACCATAGCGG
		rpl32retR	AGGAAAGGATATTGGGCGG
trnY-trnE-rpoB	1185	trnYretF	CGAATTTACAGTCCGTCCCC
		trnYintF*	TAGATTAGGTATATCCGCG
		rpoBretR	GGACATTGCGTCTATCCC

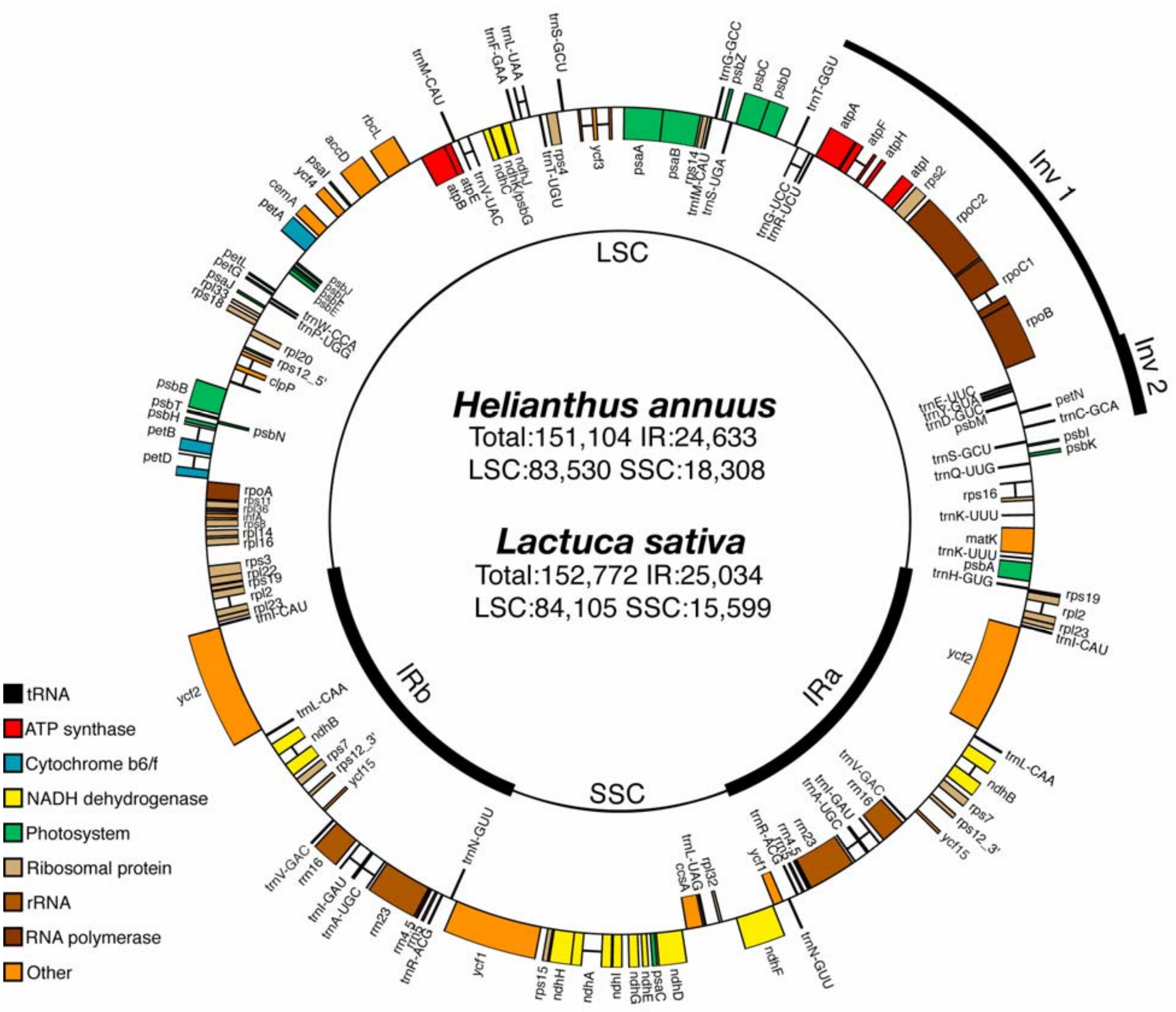
Figure Legends

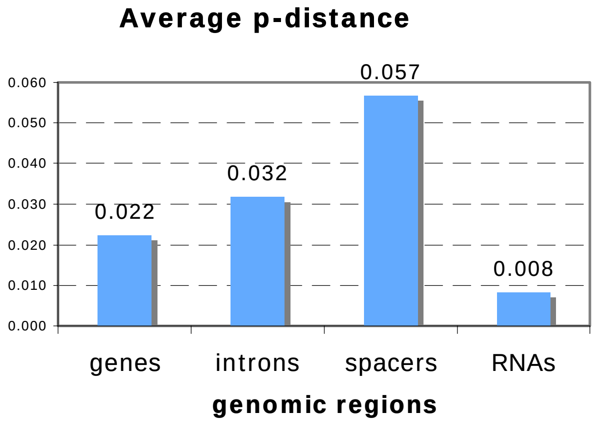
Figure 1. Chloroplast genome map for *Helianthus* and *Lactuca*. Gene order and content is the same in both genomes – they differ slightly in their extent of the IR. Thick lines in inner circle indicate extent of inverted repeats (IRa and IRb). Genes on outside of the map or transcribed in the clockwise direction and genes on the inside are transcribed in the counterclockwise direction.

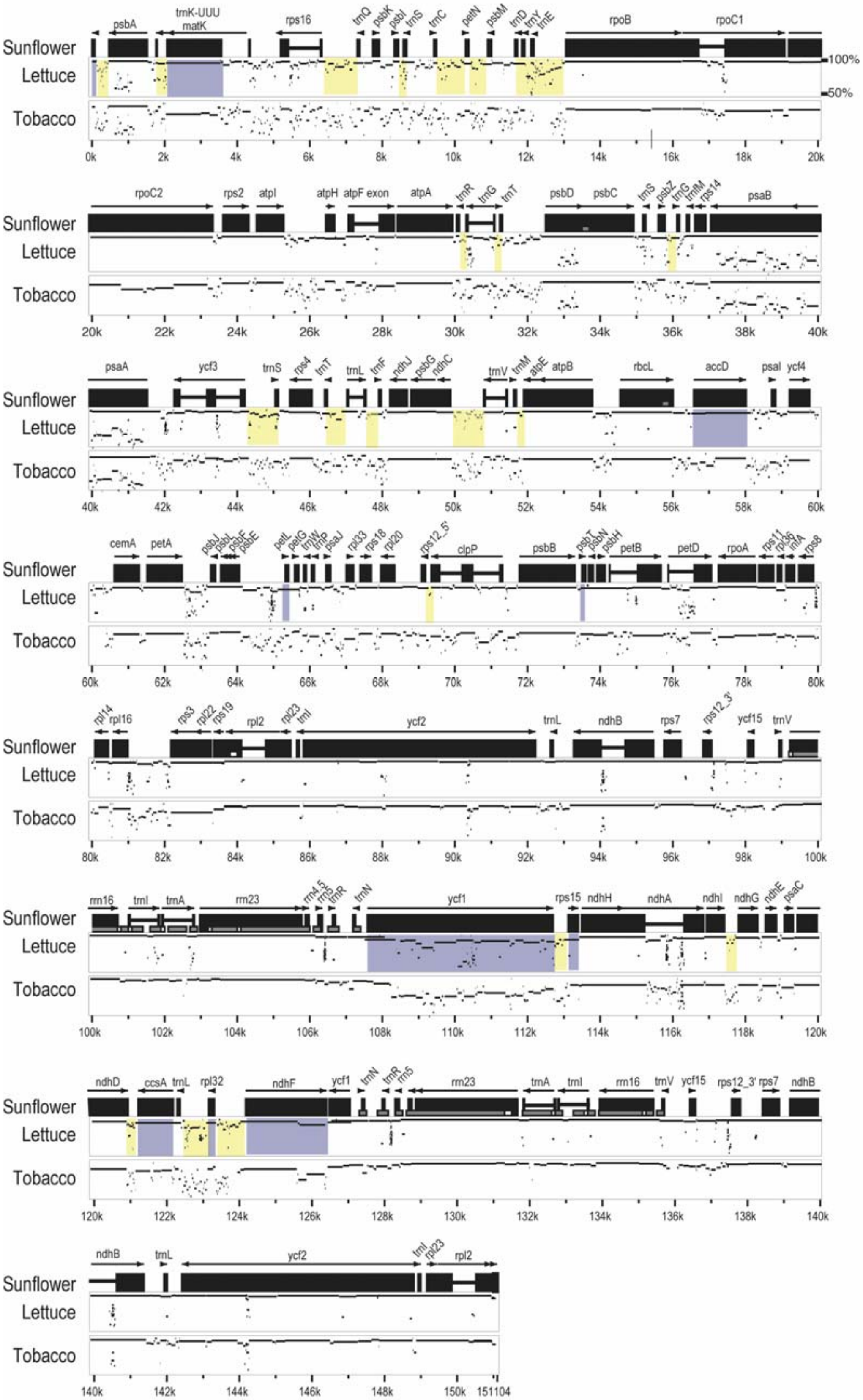
Figure 2. Average p-distance across four classes of genomic regions between *Lactuca* and *Helianthus*.

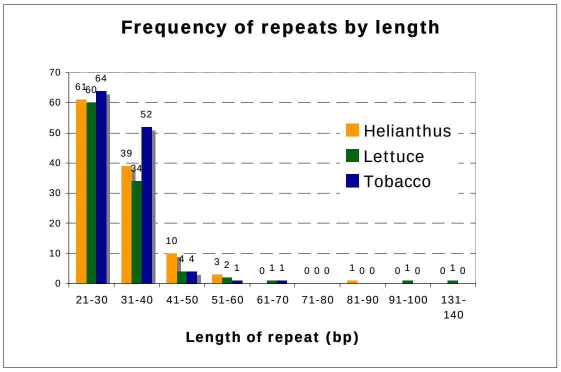
Figure 3. The detailed percent identity plot above shows the aligned regions as horizontal bars indicating average percent identity between 50-100% (shown on right of graph). As expected, the introns and intergenic spacers are most divergent, but the graph also shows variable regions within coding sequences. Highlighted in yellow are the intergenic spacers in Table 1, in blue are the genes in Table 2. Double bars show repeated sequence.

Figure 4. Repeat analyses. a) REPuter output filtered by Stacia Wyman's CPA program for repeats 21 bp or larger given a 90% sequence similarity. b) Manual examination of 23 bp or larger repeats educe seven repeat categories. c) Adjustment of the frequency histogram for *Helianthus* and *Lactuca* after manual examination and reassignment of some repeats (some larger repeats were actually composed of smaller tandem repeats). d) A summary of shared repeats among *Helianthus* (HEL), *Lactuca* (LAC), and *Nicotiana* (NIC). e) Location of repeats from Fig 4.d.

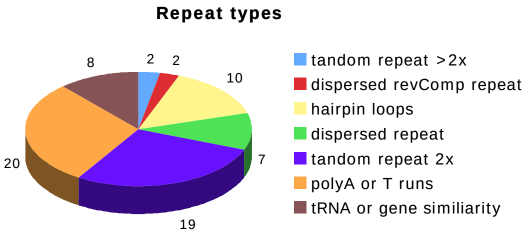

Helianthus annuus
Total:151,104 IR:24,633
LSC:83,530 SSC:18,308
Lactuca sativa
Total:152,772 IR:25,034
LSC:84,105 SSC:15,599
LSC
IRa
SSC
IRb
Inv 1
Inv 2
trnM-CAU
trnF-GAA
trnL-UAA
trnS-GCU
trnG-GCC
psbZ
psbC
psbD
trnT-GGU
atpA
atpF
atpH
atpI
rps2
rpoC2
rpoC1
rpoB
trnE-UUC
trnY-GUA
trnD-GUC
psbM
petN
trnC-GCA
psbI
psbK
trnS-GCU
trnQ-UUG
rps16
trnK-UUU
matK
trnK-UUU
psbA
trnH-GUG
rps19
rpl2
rpl23
trnI-CAU
ycf2
trnL-CAA
ndhB
rps7
rps12_3'
ycf15
trnV-GAC
rrn16
trnI-GAU
trnA-UGC
rrn23
rrn4.5
rrn5
trnR-ACG
ycf1
trnN-GUU
ndhF
ndhD
psaC
ndhE
ndhG
ndhI
ndhA
ndhH
rps15
ycf1
rpl32
trnL-UAG
ccsA
trnN-GUU
trnR-ACG
rrn5
rrn4.5
rrn23
trnA-UGC
trnI-GAU
rrn16
trnV-GAC
ycf15
rps12_3'
rps7
ndhB
trnL-CAA
ycf2
trnI-CAU
rpl23
rpl2
rps19
rpl22
rps3
rpl16
rpl14
rps8
infA
rpl36
rps11
rpoA
petD
petB
psbH
psbT
psbN
psbB
clpP
rps12_5'
rpl20
trnP-UGG
trnW-CCA
rps18
rpl33
psaJ
petG
petL
psbE
psbF
psbL
psbJ
petA
cemA
ycf4
psaI
accD
rbcL
atpB
atpE
trnV-UAC
ndhC
ndhK/psbG
ndhJ
trnT-UGU
rps4
ycf3
psaA
psaB
rps14
trnfM-CAU
trnS-UGA
trnG-UCC
trnR-UCU
⬛ tRNA
🟥 ATP synthase
🟦 Cytochrome b6/f
🟨 NADH dehydrogenase
🟩 Photosystem
🟫 Ribosomal protein
🟫 rRNA
🟫 RNA polymerase
🟧 Other


Average p-distance
0.060
0.050
0.040
0.030
0.020
0.010
0.000
0.022
0.032
0.057
0.008
genes
introns
spacers
RNAs
genomic regions

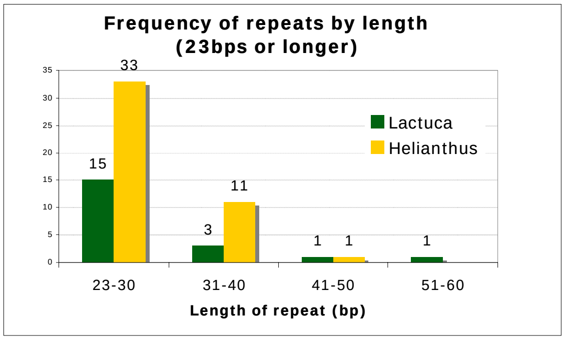

Sunflower
Lettuce
Tobacco
100%
50%
psbA
trnK-UUU
matK
rps16
trnQ
psbK
psbI
trnS
trnC
petN
psbM
trnD
trnY
trnE
rpoB
rpoC1
0k
2k
4k
6k
8k
10k
12k
14k
16k
18k
20k
rpoC2
rps2
atpI
atpH
atpF exon
atpA
trnR
trnG
trnT
psbD
psbC
trnS
psbZ
trnG
trnfM
rps14
psaB
20k
22k
24k
26k
28k
30k
32k
34k
36k
38k
40k
psaA
ycf3
trnS
rps4
trnT
trnL
trnF
ndhJ
ndhK
ndhC
trnV
trnM
atpE
atpB
rbcL
accD
psaI
ycf4
40k
42k
44k
46k
48k
50k
52k
54k
56k
58k
60k
cemA
petA
psbJ
psbL
psbF
psbE
petL
petG
trnW
trnP
psaJ
rpl33
rps18
rpl20
rps12_5'
clpP
psbB
psbT
psbN
psbH
petB
petD
rpoA
rps11
rpl36
infA
rps8
60k
62k
64k
66k
68k
70k
72k
74k
76k
78k
80k
rpl14
rpl16
rps3
rpl22
rps19
rpl2
rpl23
trnI
ycf2
trnL
ndhB
rps7
rps12_3'
ycf15
trnV
80k
82k
84k
86k
88k
90k
92k
94k
96k
98k
100k
rrn16
trnI
trnA
rrn23
rrn4.5
rrn5
trnR
trnN
ycf1
rps15
ndhH
ndhA
ndhI
ndhG
ndhE
psaC
100k
102k
104k
106k
108k
110k
112k
114k
116k
118k
120k
ndhD
ccsA
trnL
rpl32
ndhF
ycf1
trnN
trnR
rrn5
rrn23
trnA
trnI
rrn16
trnV
ycf15
rps12_3'
rps7
ndhB
120k
122k
124k
126k
128k
130k
132k
134k
136k
138k
140k
ndhB
trnL
ycf2
trnI
rpl23
rpl2
140k
142k
144k
146k
148k
150k
151104


Frequency of repeats by length
70
60
50
40
30
20
10
0
61
60
64
39
34
52
10
4
4
3
2
1
0 1 1
0 0 0
1 0 0
0 1 0
0 1 0
Helianthus
Lettuce
Tobacco
21-30
31-40
41-50
51-60
61-70
71-80
81-90
91-100
131-140
Length of repeat (bp)

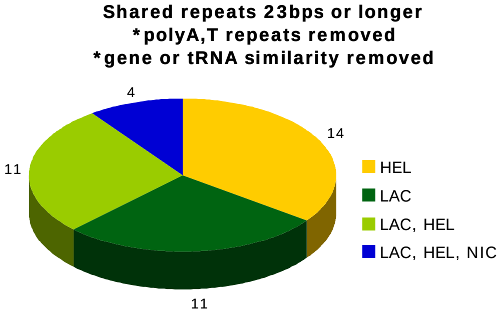
a.


Repeat types
8
2
2
10
7
19
20
🟦 tandom repeat >2x
🟥 dispersed revComp repeat
🟨 hairpin loops
🟩 dispersed repeat
🟪 tandom repeat 2x
🟧 polyA or T runs
🟫 tRNA or gene similiarity

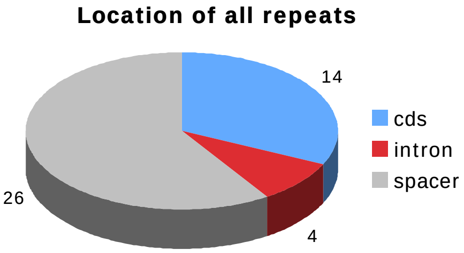
b.


Frequency of repeats by length
(23bps or longer)
35
30
25
20
15
10
5
0
15
33
3
11
1
1
1
Lactuca
Helianthus
23-30
31-40
41-50
51-60
Length of repeat (bp)

c.


Shared repeats 23bps or longer
*polyA,T repeats removed
*gene or tRNA similarity removed
4
14
11
11
🟨 HEL
🟩 LAC
🟩 LAC, HEL
🟦 LAC, HEL, NIC

d.


Location of all repeats
14
26
4
🟦 cds
🟥 intron
⬜ spacer

e.