bioRxiv preprint doi: https://doi.org/10.1101/564278; this version posted March 1, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

The alternative reality of plant mitochondrial DNA

Alexander Kozik[1*], Beth A. Rowan[1*], Dean Lavelle[1], Lidija Berke[2,3], M. Eric Schranz[2], Richard W. Michelmore[1], and Alan C. Christensen[4].

ABSTRACT

Plant mitochondrial genomes are usually assembled and displayed as circular maps based on the widely-held assumption that circular genome molecules are the primary form of mitochondrial DNA, despite evidence to the contrary. Many plant mitochondrial genomes have one or more pairs of large repeats that can act as sites for inter- or intramolecular recombination, leading to multiple alternative genomic arrangements (isoforms). Most mitochondrial genomes have been assembled using methods that were unable to capture the complete spectrum of isoforms within a species, leading to an incomplete inference of their structure and recombinational activity. To document and investigate underlying reasons for structural diversity in plant mitochondrial DNA, we used long-read (PacBio) and short-read (Illumina) sequencing data to assemble and compare mitochondrial genomes of domesticated (*Lactuca sativa*) and wild (*L. saligna* and *L. serriola*) lettuce species. This allowed us to characterize a comprehensive, complex set of isoforms within each species and to compare genome structures between species. Physical analysis of *L. sativa* mtDNA molecules by fluorescence microscopy revealed a variety of linear, branched linear, and circular structures. The mitochondrial genomes for *L. sativa* and *L. serriola* were identical in sequence and arrangement, and differed substantially from *L. saligna,* indicating that the mitochondrial genome structure did not change during domestication. From the isoforms evident in our data, we inferred that recombination occurs at repeats of all sizes at variable frequencies. The differences in genome structure between *L. saligna* and the two other lettuce species can be largely explained by rare recombination events that rearrange the structure. Our data demonstrate that representations of plant mitochondrial DNA as simple, genome-sized circular molecules are not accurate descriptions of their true nature and that in reality plant mitochondrial DNA is a complex, dynamic mixture of forms.

bioRxiv preprint doi: https://doi.org/10.1101/564278; this version posted March 1, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[1] Genome Center and Department of Plant Sciences, University of California, Davis, CA 95616, USA
[2] Wageningen University & Research, Droevendaalsesteeg 1, 6708 PB Wageningen, Gelderland, The Netherlands
[3] Current address: Genetwister Technologies B.V., Nieuwe Kanaal 7b, 6709PA Wageningen, The Netherlands
[4] School of Biological Sciences, University of Nebraska - Lincoln, Lincoln, NE 68588-0666, USA
* Joint first authors.

Author Contributions:
A.K.: Conceptualization, Investigation, Data curation, Formal Analysis, Methodology, Resources, Software, Validation, Visualization, Writing -Original Draft Preparation and Review & Editing
B.A.R.: Conceptualization, Investigation, Resources, Visualization, Writing -Original Draft Preparation and Review & Editing
D.L.: Resources, Data curation, Software
L.B.: Resources
M.E.S.: Resources
R.W.M.: Conceptualization, Funding Acquisition, Project Administration, Supervision, Writing - Review and Editing
A.C.C.: Methodology, Writing -Original Draft Preparation and Review & Editing, Supervision, Conceptualization, Visualization

Short title: The alternative reality of plant mitochondrial DNA

Data Availability:
BioProject: Organellar genomes of cultivated and wild lettuce (*Lactuca*) varieties PRJNA508811 https://www.ncbi.nlm.nih.gov/bioproject/508811 and other accessions as indicated through the text and supplemental data.

Funding
NSF grant MCB-1413152 to ACC and support from UC Davis to RWM.

bioRxiv preprint doi: https://doi.org/10.1101/564278; this version posted March 1, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

INTRODUCTION

Unlike the relatively simple mitochondrial genomes of animals, the genomes of non-parasitic flowering plant mitochondria are large and complex (1–8). They exhibit extensive variation in size (191 kb – 11,319 kb), sequence arrangement, and repeat content, yet coding sequences are highly conserved (typically 24 core genes with 17 variable genes (9–13)). The importance of mitochondria in plants is not only their role in respiration, metabolism, and programmed cell death (similar to animal mitochondria), but also their role in conferring male sterility (14–17).

Because they can often be assembled and mapped as circles, there is the common misconception that they exist *in vivo* as circular molecules (the master circle model) (18–20); however, there is a lack of strong evidence for genome-sized circular molecules in flowering plants and evidence is accumulating for non-circular forms (5,6,21–25). Their replication and recombination mechanisms are still not well understood, nor are the adaptive reasons for the striking differences from animal mitochondrial genomes. Further characterizations of plant mitochondrial genome structures and DNA maintenance pathways are required for understanding their functions, replication, inheritance, and their peculiar evolutionary trajectories.

DNA sequencing and comparisons between sister taxa revealed that plant mitochondrial synonymous mutation rates are substantially lower than in animal mitochondria, or the nucleus (26–29). In contrast, once entire genomes were sequenced, it became clear that there was very little conservation of gene order even between close relatives (30,31), likely due to frequent DNA repair that occurs via homologous recombination and nonhomologous end-joining (32–34). While most plant mitochondrial genomes include a substantial fraction of poorly conserved DNA of unknown function, some also have dramatically increased their genome sizes to millions of nucleotides, while still encoding only dozens of genes (11,35,36). In contrast to their usual representations as master circles, mtDNA exhibits complex and dynamic structures, including linear and branched molecules (which could be intermediates in replication or recombination) and these may represent multiple isoforms of the genome (37).

Plant mitochondrial genomes generally have a small number of non-tandem direct or inverted repeat sequences of several kb in length. These may recombine frequently and symmetrically, isomerizing the genome (38,39). Some genomes assemble into more than one independent molecule (40), although these may also be the consequences of assembly methods and parameters used during the assembly process (11). In many cases, particularly those sequences assembled solely from short-read sequencing,

bioRxiv preprint doi: https://doi.org/10.1101/564278; this version posted March 1, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

recombination and isomerization have simply been assumed or ignored. There are also dispersed repeats of up to a few hundred base pairs that recombine at relatively low frequencies in wild type plants but often recombine more frequently (and asymmetrically) in DNA maintenance and repair mutants (41–45). Such repeats are not always annotated, but their important contribution to the rearrangement and evolution of mitochondrial genomes is starting to emerge (33).

We have employed new technologies to address the challenge of assembling genomes with branches and rearrangements by determining the complete mitochondrial genome sequences of three closely related species in the genus *Lactuca*. In terms of gene content, overall size, and repeat content, these genomes are typical of many flowering plant mitochondrial genomes (34,46). We combined long-read and high coverage, short-read data to determine the sequences, junctions, rearrangements, and stoichiometry with great precision to produce the first high-quality mitochondrial assemblies with detailed information about structures and isoforms for species in the Compositae. This allowed us to evaluate the repeat structure and frequent isomerization by recombination at the large repeats. We propose a model for rare recombination events that rearranged the mitochondrial genome during the divergence of two *Lactuca* lineages. The physical structure of the genomes visualized by fluorescence microscopy showed that the mtDNA of *L. sativa* exists primarily in branched, linear forms and subgenome-sized circles. Our data allowed us to document the diversity of isoforms in great detail, clarify misconceptions about mtDNA structures, explore the best methods for assembly of these dynamic, complex genomes, and examine the evolution of mitochondrial genomes in both a wild and a domesticated descendent from a common *Lactuca* ancestor.

bioRxiv preprint doi: https://doi.org/10.1101/564278; this version posted March 1, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

RESULTS

Overview of assembly of the mitochondrial genomes of three lettuce species

We sequenced and assembled the mitochondrial genomes of three *Lactuca* species: *L. sativa*, *L. serriola*, and *L. saligna.* Because plant mitochondria are known to have dynamic and rearranging genome structures, we knew that there was a possibility of the assembly producing large contigs with multiple connections to other contigs. Therefore, our assembly approach did not make assumptions about linearity or circularity. We only relied on the sequence reads to identify the contiguous segments and how they are connected. After producing and polishing contigs to form the primary structural units of the mitochondrial genome, we used junction information to join primary structural units together to form secondary building blocks. Finally we used the stoichiometry of these blocks to find repeated elements that recombine to generate different isoforms of the genome (Figure S1).

For *L. sativa* and *L. saligna*, we verified our high quality assemblies using a multistep approach involving processing and analysis of data from several sources (Table 1, Figure S1). For *L. serriola,* we compared Illumina paired-end and mate-pair reads to the *L. sativa* assembly to determine its mitochondrial genome structure. Because the mitochondrial genome of *L. serriola* was essentially identical to that of *L. sativa* (Genbank accession pending), we focused the majority of our subsequent analyses on comparing *L. saligna* and *L. sativa.*

Table 1: Types of read data used for mitochondrial genome assembly and analysis

	Illumina paired-end	Illumina mate-pair	Pac Bio	Hi-C
Lactuca sativa	✓	✓	✓	✓
Lactuca serriola	✓	✓		
Lactuca saligna	✓		✓	

Initial assembly of PacBio reads for *L. sativa* and *L. saligna*

For *L. sativa* and *L. saligna*, we assembled each mitochondrial genome using the CLC assembler with PacBio mitochondrial reads (selected from whole genome read data) and polished with Illumina paired-end reads. This process resulted in 11 contigs for *L. sativa* and 10 for *L. saligna* (Figure 1A). Distinct contigs are designated by letters from K to Z and will be referenced in the text accordingly (see contig naming convention in Materials and Methods). The total assembly length of the polished non-redundant

bioRxiv preprint doi: https://doi.org/10.1101/564278; this version posted March 1, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

contigs (primary structural units; see below) was 314,659 bp for *L. sativa* and 323,254 bp for *L. saligna* (Table 2). Eight contigs (K, L, M, T, R, P, Q, and Z) were very similar between the two species (ranging from 99.8% to nearly 100% similarity at the nucleotide level), although there were minor variations around the contig termini. In *L. saligna*, the contig designated UV was a single contig that contained sequences found in U and W of *L. sativa*. We therefore split this contig into two parts (U and V) to simplify and clarify the visualization and interpretation of contig relationships. Upon refining contig sequences we defined them as the primary structural units of the mitochondrial genome, and determined their junctions, relative orientations and copy number.

Table 2: Lengths of primary structural units for *L. sativa* and *L. saligna*

L. sativa		***L. saligna***	
Unit name	Length (bp)	Unit name	Length (bp)
K	78,285	K	79,109
L	44,054	L	43,583
W	53,983	UV	90,023
U	38,613		
M	30,567	M	30,534
P	20,468	P	20,420
Z	19,295	Z	19,278
Q	11,296	Q	11,083
R	10,430	R	10,433
		S	14,743
N	4,116		
T	3,552	T	4,048
TOTAL	314,659	**TOTAL**	323,254

bioRxiv preprint doi: https://doi.org/10.1101/564278; this version posted March 1, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Classification of repeated sequences

Because segments of the genome that are several kb in length are commonly present at two locations in many mitochondrial genomes (7,29,38–41), we expected that any contigs that represent a segment of the genome present in more than one location would be detected as variation in the copy number of that contig. Indeed, our coverage analysis (Figure S2) revealed that three contigs (M, R, and T) had twice the coverage relative to the others; thus, we designated these as large repeats. Repeat pairs of M, R, and T would be expected to recombine continuously and serve essentially as hinges between different isoforms of the genome. We also found intermediate-length (< 1 kb) repeats within contigs (Table S1). For example, a 576 bp repeat that we termed X-01b (yellow triangles on Figure 1) was found in contig N in *L. sativa* and in W/V in both species. In *L. saligna*, this repeat was also present in an inverted orientation at the termini of contig S.

Identification and characterization of primary structural units and secondary building blocks

To infer the structural topology of the mitochondrial genomes of both species, we took advantage of the long PacBio reads to identify the junctions between contigs. We first polished the contigs by correcting minor discrepancies at their termini and precisely defined their boundaries at junction points. After this refinement, we considered the polished contigs as “primary structural units.” Using the set of maximally-informative reads that we selected during the assembly process (see Methods), we employed a “reverse read mapping” approach for determining which primary units were joined to one another. In this approach, the primary structural units were fragmented into 2 kb sliding windows with a 1 kb step size and mapped against the maximally-informative read set for each species (see Materials and Methods). This enabled the identification of secondary building blocks composed of several primary structural units in a specific order (Figure 1 B-F). In many instances, the PacBio reads were long enough to span junctions on both sides of the T and R repeat units (but not the much longer M repeat unit). These basic building blocks revealed the existence of several different arrangements of the primary structural units. For example, the 3’ end of unit R was joined to unit U in one building block and unit P in another in both species. Ultra-long reads that spanned junctions between four units (Figure 1D) confirmed these alternative arrangements for *L. sativa*. After re-assembling the junctions for greater accuracy (see Methods), we mapped 2 kb segments spanning the junctions to PacBio reads to quantify how often they occurred (Figure S3).

bioRxiv preprint doi: https://doi.org/10.1101/564278; this version posted March 1, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Identification of major isoforms

Given the set of all detected secondary building blocks and their stoichiometry, we were able to order the primary structural units into sequences that were the total length of the mitochondrial genome. In both species, there were several different arrangements (isoforms) of primary structural units that were well supported by the data. The two major mitochondrial isoforms for *L. sativa* can be represented as in Figure 2A. The difference between the two major isoforms (α and β) of *L. sativa* can be described as resulting from an exchange of two primary structural units, P and U, between long repeats R and T. These major isoforms have roughly equivalent stoichiometry. With two copies of each of the three large repeats included, these isoforms represent genome lengths of 363,324 bp for *L. sativa* and 368,269 bp for *L. saligna.*

The assembly of *L. saligna* was more complex. We identified an interesting 12 kb primary unit in *L. saligna*. This sequence, unit S, encodes a Type B2 superfamily DNA polymerase, a T3/T7 phage type RNA polymerase, and a 1,218 bp inverted repeat at each end, resembling linear plasmids described in other plants (47). Some plant mitochondria include integrated fragments of these plasmids or autonomous plasmid molecules (48–53); phylogenetic analysis suggests occasional horizontal transfer of these as well (54). Unit S appears to be an integrated copy of a molecule of this type in *L. saligna*. A portion of the sequence in unit S is present in *L. sativa*, but it lacks one inverted repeat and the DNA and RNA polymerase genes, and is presumably degraded and nonfunctional. Other species that carry such plasmids include relatives of *Lactuca*, *Daucus carota* (52) and *Diplostephium hartwegii* (55). We found no evidence for free linear plasmids in either *Lactuca* species. Unit S is equally likely in either orientation in *L. saligna*, indicating that the inverted repeats at the end recombine with each other at a high frequency.

Figure 2B displays a comprehensive set of *L. saligna* mitochondrial isoforms that considers all of the detected secondary building blocks (Figure 1 E,F). The presence of the complete S unit in *L. saligna* prevented us from unambiguously constructing a simple and stoichiometrically symmetrical genome model as in *L. sativa*. Primary unit T, which is a repeat, is present in four arrangements: P-T-Q, P-T-L, K-T-L, and K-T-Q. Likewise, the repeat unit R was found in four distinct blocks: Z-R-P, Z-R-U, L-R-U, and L-R-P. Analysis of ultra-long PacBio reads (Figure 1 F) suggested that all configurations with the R and T repeats are equally represented (as was also the case for *L. sativa*). Primary unit M is flanked by unit S on both sides and can potentially form an inverted repeat configuration of ~32 kb. Recombination via the X-01a repeat, which is part of the inverted S repeat structure, and its counterpart in the V unit can lead to the formation of an M-V junction (Figure 2B). PacBio read-through analysis showed the M-V and V-Z

bioRxiv preprint doi: https://doi.org/10.1101/564278; this version posted March 1, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

junctions in equilibrium (equally represented). There was a noticeable deficiency of S-Z junctions in comparison with M-S junctions, which may reflect that a linear genome structure with unit S at one terminus is more prevalent. These *L. saligna* isoforms are not necessarily in stoichiometric equilibrium (like the major *L. sativa* isoforms) due to anomalies caused by the integrated linear S plasmid.

Annotation of the sequences

Because the representation of the annotations on circular maps has led to the misperception of the existence of a specific circular molecule, we present annotations of each primary unit separately. Importantly, we do not suggest that these maps represent specific linear chromosomes either. Combining the junction data and the physical data (below) suggests a fluid and dynamic genome with multiple isoforms and topologies. The annotation maps shown in Figure 3 are the best way to present a static figure representing the dynamic reality of mitochondrial genomes. Arbitrarily choosing to present one possible isoform among many (which is unfortunately required for the GenBank submission) leads to an overly simplistic picture of the genome.

Annotation of the genomes identified a typical angiosperm set of rRNA, tRNA, and protein-coding genes. Both *L. sativa* and *L. saligna* have all of the core genes (as defined in reference (12): *atp1*, *atp4*, *atp6*, *atp9*, *ccmB*, *ccmC*, *ccmFc*, *ccmFn*, *cob*, *cox1*, *cox2*, *cox3*, *matR*, *mttB*, *nad1*, *nad2*, *nad3*, *nad4*, *nad4L*, *nad5*, *nad6*, *nad7*, and *nad9*. Of the variable genes defined previously (12), both *Lactuca* genomes have the same gene content as *H. annuus*, namely, *rpl5*, *rpl10*, *rpl16*, *rps1*, *rps3*, *rps4*, *rps12*, and *rps13*. *L. sativa*, *L. saligna*, and *H. annus* also have a pseudogene of *sdh4*, identified by a stop codon in the middle. This gene content is unremarkable for a member of the Compositae, except for four duplicated genes described below.

The repeat units in *Lactuca* mitochondrial genomes contain protein-coding genes, as is often the case in plant mitochondrial genomes, although the particular genes duplicated varies. The *ccmB* and *rpl10* genes are both located in repeat unit R, resulting in two identical copies in the full-length genomes of both species. Another gene present in two locations is *atp1*, encoding the alpha subunit of the F_oF_1 ATP synthase (atpα). Nearly the entire gene is in repeat unit T, including the 5' flanking region, the start of translation, and 1,319 nucleotides of coding sequence. In flowering plants, the *atp1* gene encodes a highly conserved protein of between 505 and 512 amino acids in length. One copy in both *Lactuca* species, annotated as *atp1-1*, extends from repeat unit T into unit P and encodes a 511 amino acid protein that is 99.8% identical to the atpα proteins from *Helianthus annuus* and *D. hartwegii* (55,56) and is 97.3% identical to the atpα protein of the distantly related dicot *Vitis vinifera* (57). The second copy, *atp1-2*,

bioRxiv preprint doi: https://doi.org/10.1101/564278; this version posted March 1, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

extends from repeat unit T into unit U in *L. sativa* and into unit K in *L. saligna*. These copies are identical to *atp1-1* through 444 amino acids of coding sequence and then diverge: *atp1-2* of *L. sativa* adds an additional 205 amino acids and *atp1-2* of *L. saligna* adds 216. A similar situation exists in *D. carota* (58) in which the first 1,452 nucleotides of the gene are present in a repeat. One copy (annotated in 58 as a pseudogene) is a 514 amino acid protein that resembles other plant atpα proteins, and the other, annotated as *atp1*, includes the first 485 amino acids of atpα fused to an additional 271 amino acids, suggesting that the *D. carota* annotations are reversed. Similarly, the 5' end of *rps4* is in repeat unit R, joined to units Z and L (in both *Lactuca* species). The gene that extends from R into Z is annotated as *rps4-1* and resembles other plant mitochondrial *rps4* genes. The other copy, *rps4-2*, consists of 189 nucleotides encoding the first 63 amino acids of the rps4 protein, fused to a 268 amino acid open reading frame in unit L, which does not resemble other plant rps4 proteins. Further work will be needed to understand if these novel chimeras are expressed, whether they are selectively neutral or have a novel function, and how multi-subunit protein complex assembly is regulated.

Long distance sequence technologies support mitochondrial genome isoform models

Illumina mate-pair and Hi-C libraries validated the models of the mitochondrial genome generated using PacBio reads. Reads from two different insert size (2.5 and 10 kb) libraries for *L. sativa* and *L. serriola* mapped to the mitochondrial genome isoform alpha confirmed the sequential order of the primary structural units. The distribution of long repeats (R, T, and MN) on two-dimensional distance plots were consistent with the structure of isoforms inferred from the reverse mapping approach using PacBio reads (Figure 4A,B). The distribution of long-distance interactions revealed by mapping reads from the Hi-C genomic library of *L. sativa* to the mitochondrial reference assembly was congruent with the plot with mate-pair libraries (Figure 4C). Hi-C has the potential to detect contacts between regions of the genome that are brought into proximity *in vivo* by proteins or other components of the mitochondrial nucleoid, but we did not find strong evidence for long-distance interactions arising from the organization of mtDNA within the nucleoid (see also Figure S4).

Detailed inspection of distance plots from Illumina mate-pair reads for junction regions revealed the existence of minor isoforms that resulted from recombination between short repeats of less than 600 nt (Figure 5). Presence of these minor isoforms did not affect the equilibrium between the major isoforms of *L. sativa* (Figure 2A). Plots shown in Figures 4 and 5 were generated by mapping 320,000 reads each for 2.5 and 10 kb libraries. To detect recombination between short repeats of less than 300 nt, all

bioRxiv preprint doi: https://doi.org/10.1101/564278; this version posted March 1, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

available (~1 million) reads from the 5 and 10 kb mate-pair libraries for *L. sativa* and the 10 kb library for *L. serriola* were used. To simplify the analysis of recombination at short repeats, which can be ambiguous and complicated due to reads aligning at multiple locations, only unique mappings were selected (Figure S5 A-C). All mappings of reads from *L. sativa* and *L. serriola* generated similar patterns with detectable signals of recombination events between short repeats. The fraction of recombinants detected between short repeats was very low and estimated to be within (1% – 10% for repeats 100 – 500 bp and less than 1% for repeats shorter than 100 bp) of the values on the main diagonal (see example data in Figure S6A). This rarity explained why only a few corresponding recombinants were detected within the PacBio reads.

Evolution of the mitochondrial genomes of *L. sativa* and *L. saligna*

While *L. sativa* and *L. serriola* are identical in their sequence arrangement and repeat content, *L. saligna* has notable differences from *L. sativa*. A major difference in the *L. saligna* lineage is that there has been an inversion around the intermediate-sized repeat X-07. Conceptual reversal of that inversion leads to an intermediate form that can be converted to *L. sativa* by additional rearrangements (Figure 6). Without an outgroup or intermediates in the evolutionary process, it is not currently possible to unambiguously specify the molecular events in each lineage that led to these differences. An additional important difference is the 12-kb sequence unit S integrated in the genome of *L. saligna* and partially deleted in *L. sativa*.

In addition to the inversion and the integrated linear plasmid, there is also a complex rearrangement and a number of minor differences distinguishing the two *Lactuca* lineages. The block referred to as V in *L. saligna* is instead in an inverted orientation between blocks M and Z in *L. sativa*. A 50 bp sequence at the end of *L. sativa* contig L was found at the end of *L. saligna* contig T. Another short segment of ~200 nt at one end of *L. sativa* contig Q was missing in *L. saligna*, and a ~600 nt segment at the end of *L. saligna* contig K was missing in *L. sativa*. *L. sativa* had a unique contig N, but a small portion of it is present in *L. saligna* contig U. Fragments of the chloroplast genome are integrated into the same location in contig U in both species; remarkably, the integrated sequences are not the same. In *L. sativa,* the integrated sequence is a fragment of a DNA-directed RNA polymerase gene, and in *L. saligna* it is a fragment of a P700 chlorophyll a apoprotein gene. These differences may be due to events that began with a homologous strand invasion, but included additional breakage events in the unique flanking regions that were subsequently repaired by non-homology based mechanisms.

bioRxiv preprint doi: https://doi.org/10.1101/564278; this version posted March 1, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Physical structure of mitochondrial DNA molecules in *L. sativa* confirms multiple forms of the genome

We observed a variety of physical structures of mtDNA molecules using in-gel fluorescence microscopy after staining agarose-embedded mtDNA from *L. sativa* with DNA-binding fluorophores (Figure 7). We quantified the structural forms in multiple microscopic fields (Figure 7A-F). Branched linear forms were most frequently observed (40 of 98 fields); circular molecules were observed in only 22 fields (Figure 7G). Additional examples of branched forms are shown in Figure S7. Furthermore, no supercoiled circular or genome-sized linear molecules were observed by pulsed-field gel electrophoresis (Figure 7H) in five different mitochondrial samples from *L. sativa*. The majority of DNA in the gel was confined to the well, consistent with the immobilities of branched linear and relaxed circular forms. The remainder ran as a small smear of linear fragments ranging in size from ~50 to 100 kb. No band representing the linear DNA form of the mitochondrial genome (isoforms α and β in Figure 2A; ~363 kb) was observed. These results suggest that simple and branched linear forms are the predominant form of mtDNA molecules in *L. sativa*.

DISCUSSION

The complexity of plant mitochondrial genomes has confounded efforts to characterize their sequence, structure, and dynamic evolution. Most plant mitochondrial genomes can be assembled into circular maps; however, there is no evidence that full genome-length circular molecules exist in substantial quantities (4,59) and there is increasing evidence for other configurations (5,6,21–25). Importantly, assumptions have to be made in any genome assembly process about the type of molecule being assembled, and published sequences reflect choices made during the assembly process. By default, the usual goal of assembly for plant mitochondria has been to produce a “master circle,” and successful construction of circular maps has reinforced the near-universal belief in the existence of master circle molecules. Because plant mitochondrial genomes rearrange so frequently and are present in so many complex and non-circular states, assumptions about the genome before assembly have inevitably led to incomplete and/or incorrect structures. In addition to assuming a master circle, sometimes the goal has been to assemble the sequence into unique circular molecules of minimum size, in spite of probable recombination at very large repeats in different molecules (60). Incomplete assemblies can also lack repeats that could combine smaller molecules into larger ones. Genome rearrangements can occur both within species or individuals in real time and between species during their evolution and

bioRxiv preprint doi: https://doi.org/10.1101/564278; this version posted March 1, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

divergence (33); therefore, accurate assembly and identification of multiple isoforms is critical to understanding the evolution of mitochondrial genomes.

Our success was dependent on the exploitation of multiple long read technologies and atypical assembly approaches. Our use of Illumina mate-pair, Hi-C, and long PacBio reads enabled the high resolution analysis of the mitochondrial genomes of three species of *Lactuca*. Long reads at high coverage depth were essential for accurately determining the *in vivo* sequences and arrangements because short paired-end reads would not allow the identification of recombination events that involve repeats longer than the paired-end library fragments. We assembled these genomes with CLC, reverse read mapping, and long distance analysis with mate-pair reads that made no assumptions as to the component structures and redundancies. Conventional assembly software programs, such as Falcon or Canu (61,62), assume a single chromosomal sequence and consequently are limited in their ability to identify all isoforms that may exist in plant mitochondrial genomes and resolve complex nested repeats. Assemblers that can split putative chimeras (isoforms) into distinct structural units in the initial steps of assembly have advantages over those that try to assemble all reads into a single linear or circular genome. Bottom-up construction of contigs and alignment to raw PacBio reads allowed a sensitive and accurate assembly of dynamic mitochondrial genomes and identification of multiple isoforms, resulting in precise representation of the subtle complexities of plant mitochondrial genomes and enabling evolutionary studies. The availability of long read sequences for an increasing number of plant species provides the opportunity for the reassessment of mitochondrial structures and diversity across the plant kingdom using our approach. When we applied our workflow to the mitochondrial genome data of *Leucaena trichandra* (63), we were able to generate a cyclic graph for this genome and connect all segments into contiguous isoforms (Figure S8).

Our structural studies using fluorescence microscopy of mtDNA molecules and Hi-C provided a detailed description of the complexities of plant mitochondrial genome architecture and dynamics that was consistent with our genome assemblies. This provided further evidence for the existence of multiple major and minor isoforms produced by homologous recombination at very large repeats. The prevalence of branched, linear forms of mtDNA likely represents ongoing recombination, perhaps due to recombination-dependent DNA replication (5). *L. sativa* major isoforms had a genome size of ~363 kb, but no linear band of this size was observed by pulsed-field gel electrophoresis (PFGE). Circles only accounted for a small percentage of the DNA forms, yet the assemblies had a circular topology. The interpretation that the branched linear structures contain multi-genomic concatemers of the genome is the only one that

bioRxiv preprint doi: https://doi.org/10.1101/564278; this version posted March 1, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

is consistent with all of these findings. The Hi-C data also suggested no higher-order structure of mtDNA within the nucleoid.

Repeated sequences in plant mitochondria provide the substrates for contemporaneous and long-term structural variation. Large repeats, usually several kb or more in length, recombine frequently and isomerize the genome continuously. Whether the abundance of isoforms varies among cell types or tissues remains to be investigated; however, the tools are now available for the detailed analysis of specificity. There are also infrequently recombining non-tandem repeats, usually less than several hundred base pairs in length (34). Although it is unclear what controls the interconversion of isoforms and their relative stoichiometries, repeat length may be an important factor. Homology-based ectopic recombination for the intermediate-length repeats has been shown to occur following double-strand breaks (44,64–66), when DNA repair functions are impaired by mutations (41,44,67), and occasionally during the evolution and divergence of species (33). We were able to compare our detailed structures of the different species and infer the most parsimonious path between possible ancestral forms and the genomes of the extant species. Ectopic recombination between dispersed repeats could explain some of the structural changes that have become fixed between the two species. We also identified small linear plasmid-like molecules that have been shown to exist autonomously in plant mitochondria in a variety of species and are sometimes integrated into the mitochondrial genome (reviewed in 68). Both *Z. mays* and *D. carota* have integrated plasmids in their mitochondrial genomes which prevented the assembly of the genomes into circular maps (58,69,70). One such linear plasmid is integrated into the *L. saligna* genome but is incomplete in *L. sativa*. We found no evidence for its integration at any other location, nor for its existence as an autonomous molecule. Its presence in the *L. saligna* genome and its frequent inversions complicated the analysis of junctions in *L. saligna* and prevented a simple interpretation of the stoichiometry of the junctions.

This is the most detailed description of the sequence, structure, and dynamics of plant mitochondrial genomes to date. Our comprehensive approach for investigating mitochondrial genomes can be applied to existing and future datasets generated for other plant species. This approach should avoid incomplete assemblies and reveal the complexity of multiple isoforms, non-circular molecules within circularly permuted maps, and recombination events that occur within and between species. This will be facilitated by the advent of even longer reads provided by rapidly advancing single molecule (PacBio) and nanopore sequencing technologies. Understanding the evolutionary changes that can occur with mitochondrial plant genomes, including integration or loss of linear plasmid-like molecules, provides the foundation for future work on plant evolution and taxonomy, as well as mitochondrial genome structure and function.

bioRxiv preprint doi: https://doi.org/10.1101/564278; this version posted March 1, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

MATERIALS AND METHODS

PacBio reads

DNA was extracted from seven-day-old dark-grown seedlings of *L. sativa* cv. Salinas grown in sterile conditions at 15°C using a modified CTAB protocol (71) with two chloroform extractions. DNA was removed after EtOH precipitation using a glass hook and then washed two times in 70% EtOH. The DNA was further processed with a high-salt, phenol-chloroform extraction and precipitation and removal of polysaccharides (https://www.pacb.com/wp-content/uploads/2015/09/Experimental-Protocol-Guidelines-for-Using-a-Salt-Chloroform-Wash-to-Clean-Up-gDNA.pdf). Finally, the DNA was precipitated using EtOH and washed two times with 70% EtOH. The DNA was divided into two samples for library preparation. For the first library, the DNA was sheared using a Megaruptor instrument to 20 kb before PacBio library construction, while the second PacBio library was prepared directly from the non-sheared DNA. SMRTbell libraries were made according to manufacturer's standard protocol (Pacific Biosciences). Libraries were size selected >20 kb using BluePippin (Sage Science). Sequencing on a Pacific Biosciences (PacBio) RSII Instrument generated 39.6 Gb of 2.7 million single pass reads from 28 SMRT cells**.** The raw reads were deposited in GenBank under accession numbers SRX3557844–SRX3557871.

For the *L. saligna* germplasm accession CGN5271, DNA was extracted from leaves of young seedlings grown on soil in a greenhouse under long-day (16 h light) conditions with a temperature of 21°C during the day and 19°C at night using a modified version of the protocol described previously (72). PacBio long read data was obtained from 86 SMRT cells. The selected set of mitochondrial reads (1 Gb, 101,753 reads) is available at SRA under accession number SRX5104332.

Illumina genomic reads

For *L. sativa* cv. Salinas, we used previously published Illumina whole genome libraries (73). Paired-end reads included insert sizes of 175 bp (SRR577192), 475 bp (SRR577183), and 750 bp (SRR577184), and mate pair reads included insert sizes of 2.5 kb (SRR577197), 5 kb (SRR577193), and 10 kb (SRR577207). Paired-end libraries were filtered for high quality reads of uniform length (100 nt), yielding 25 million read pairs from the 175 and 475 bp libraries, and 15 million pairs from the 750 bp insert library. Paired-end genomic reads were aligned (mapped) to the lettuce chloroplast reference sequence (http://www.ncbi.nlm.nih.gov/nuccore/DQ383816) using the CLC Genomics Workbench with stringent parameters (Mismatch cost = 2; Insertion cost = 3; Deletion cost = 3; Length fraction = 0.9; Similarity fraction = 0.9) to find sequences that

bioRxiv preprint doi: https://doi.org/10.1101/564278; this version posted March 1, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

mapped to the chloroplast genome. Unmapped read pairs were collected and compiled into separate files for downstream mitochondrial assembly and analysis. Reads from mate-pair libraries were used without filtering against the lettuce chloroplast genome.

L. serriola acc. US96UC23, whole genome sequence (WGS) paired-end and mate pair genomic libraries were prepared using the DNeasy Plant Mini Kit (Qiagen) from DNA isolated from dark grown seedlings. A paired-end 170 bp insert library was constructed utilizing the NEXTflex PCR-Free approach (BIOO Scientific). The 2.5 kb library was constructed using the Mate Pair Library v2 Kit (Illumina). The 10 kb library was constructed using the Nextera Mate Pair Sample Preparation Kit (Illumina). The raw reads were submitted to the SRA database under accession numbers SRX5097892 (170 bp), SRX5097891 (2.5 kb), and SRX5097890 (10 kb).

Illumina paired-end read data for *L. saligna* (insert size 500 bp) were generated as a part of the International Lettuce Genomics Consortium (ILGC) project (http://lgr.genomecenter.ucdavis.edu/). Selected mitochondrial reads are available at NCBI Sequence Read Archive under accession number SRX5131542.

Dovetail Hi-C libraries

Hi-C libraries from leaves were generated by Dovetail™ using their standard proprietary protocol. Libraries were sequenced using an Illumina HiSeq 4000 at the UC Davis Genome Center. Paired reads were deposited in NCBI GenBank SRA under accession number SRX3973834.

Preliminary Illumina assembly of the mitochondrial genome for *L. sativa*

The plant mitochondrial RefSeq sequences from GenBank were used as a reference to mine mitochondrial contigs from a preliminary assembly of *L. sativa* Illumina reads that was generated using Velvet (74). This preliminary assembly involved six rounds of independent assemblies using Velvet with k-mer values 51, 57, 63, 67, 71, and 75. The Velvet contigs were then joined using CAP3 (75) specifying -o 200 -p 95. A self-BLAST was performed on the CAP3 contigs and a non-redundant set of sequences was selected for downstream analysis.

Optimization of CLC assembly parameters with *L. sativa* chloroplast PacBio genomic reads

The reference chloroplast genome (accession number DQ383816.1) was used to recover all PacBio reads that contained chloroplast fragments for input into the

bioRxiv preprint doi: https://doi.org/10.1101/564278; this version posted March 1, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

assembly with the CLC Genomics Workbench using the default approach of the Genome Finishing Module. Assembly parameters were adjusted until we achieved the expected three contigs for this genome, corresponding to the long and short unique segments and the inverted repeat.

Assembly of the *L. sativa* mitochondrial genome with PacBio reads

The preliminary Illumina-based mitochondrial genome assembly for *L. sativa* was used to query and select PacBio reads containing mitochondrial sequences from the raw PacBio genomic read set. The Illumina reads had already been filtered to remove any plastid DNA reads. The total number of PacBio reads selected for the assembly was adjusted to achieve 250x coverage (~4,000 reads) before input into the first round of assembly with the CLC Genomics Workbench using the approach and parameters identified from the chloroplast genome assembly. After the first round of assembly, chloroplast segments contained in the mitochondrial assembly were masked and the masked contigs were used to mine additional PacBio reads for mitochondrial sequences for a second round of assembly. This process was repeated until all reads containing mitochondrial segments had been recovered.

PacBio assembly of mitochondrial genome of *L. saligna*

The *L. sativa* PacBio mitochondrial genome assembly was used to select PacBio reads from the *L. saligna* data that contained mitochondrial fragments. As before, the amount of reads was adjusted to ~250x coverage (~6,000 reads) for input into the the first round of assembly with the CLC Genomics Workbench, following the same approach as used for *L. sativa*. The PacBio reads were slightly shorter for *L. saligna*, requiring more total reads to achieve ~250X coverage.

Contig naming convention

Upon assembly and preliminary comparison of contigs between species, the following convention was chosen to assign IDs to simplify downstream data analysis and interpretation. Contig naming started with K to avoid first letters of alphabet that could be used to designate figure panels on data presentation. Letters X and Y were skipped because of confusion with human chromosomes, letter O because of visual similarity with 0 (zero). *L. sativa* contigs M and N are frequently adjacent on genome isoforms, so alphabetical ordering keeps them together in data tables. Two letters were assigned to the longest *L. saligna* contig UV because it can be decomposed into two distinct sequences that are similar to *L. sativa* U and W. At the same time, *L. sativa* W is different from *L. saligna* V by having extra segments and thus was “wider.” Alphabetical

bioRxiv preprint doi: https://doi.org/10.1101/564278; this version posted March 1, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

sorting keeps them (U + V) in the right order. For one repeat, the letter R was assigned, and for the shortest one T (tiny). *L. saligna* S contig stands for "special" (integrated linear plasmid). These simple mnemonic rules were helpful for navigating the diverse data in the process of genome assembly because so many steps had to be visualized and resolved manually. This letter coding style was particularly useful in analysis of data using MS Excel spreadsheets and conditional formatting that allowed distinct coloring for different contigs and their segments.

Inference of mitochondrial secondary building blocks using raw PacBio reads

Contiguity of the assemblies (sequential order of contigs or primary structural units) was determined based on the analysis of full length raw PacBio reads. This led to the set of secondary building blocks that were used to infer the comprehensive set of mitochondrial genome isoforms.

For the reverse PacBio read mapping approach, a sliding window of 2 kb with a step size of 1 kb was applied to each contig of the assembly (Figure S9A) to generate a library of overlapping fragments. Overlapping tiling fragments were mapped/aligned back to raw mitochondrial PacBio reads. A BLAST-N search was used to map tiling fragments to PacBio reads using the relaxed parameters to allow longer gaps (compared to default parameters) over contiguous alignment (-V T -F F -e 1e-60 -y 50 -X 75 -Z 500). A custom BLAST-N parser (https://github.com/alex-kozik/atgc-tools/blob/wiki/tcl_blast_parser.md) was used to generate a tab-delimited table that was used for sequential order analysis of primary structural units in PacBio reads. PacBio reads containing at least 10 distinct tiling fragments of the mitochondrial non-redundant contiguous genome sequence for *L. sativa* and 12 distinct tiling fragments of the mitochondrial non-redundant contiguous genome sequence for *L. saligna* were selected and compiled into maximally informative sets. Maximally informative sets resulted in 4,046 and 11,009 PacBio reads for *L. sativa* and *L. saligna*, respectively. These sets were used for the analysis of junctions between different mitochondrial genome primary structural units. The distributions of junctions between different units were first analyzed visually with MS Excel (Figure S9B). Regular expressions with egrep were utilized to quantify the junctions from tables in tab-delimited format.

Re-assembly and validation of junctions

Regions of PacBio reads specific to individual junctions (as identified in each set of secondary building blocks, see above) were compiled as separate, small subsets and reassembled to recover the precise sequence at each junction. For each junction, PacBio reads containing termini of two contigs were trimmed to 4–6 kb so the junction

bioRxiv preprint doi: https://doi.org/10.1101/564278; this version posted March 1, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

region was located in the middle of the trimmed reads. Subsets of the trimmed PacBio reads were assembled with CLC for each junction individually. The alignment of mitochondrial contigs to reassembled junction regions allowed for the determination of precise sequences between mitochondrial contigs.

Error correction using Illumina reads

Paired-end Illumina reads for *L. sativa* and *L. saligna* were used for error correction of the assembled PacBio contigs. Illumina reads from genomic libraries were mapped/aligned to the PacBio contigs to correct homopolymer errors and consensus sequences were selected based on what was supported by the majority of the Illumina reads.

Mitochondrial contig stoichiometry

The stoichiometry of mitochondrial contigs was determined by PacBio read coverage. A sliding window of 2 kb with a step size of 0.5 kb was applied to each contig of the assembly to generate a library of overlapping fragments. BLASTN was performed on each 2 kb tiling fragment versus the set of the most informative PacBio reads. To calculate the coverage, the number of PacBio reads for each tiling fragment was counted if the alignments were longer than 1.8 kb with an identity of 80% or better. Since all non-tandem repeats of medium size are shorter than 1.2 kb, alignments due to repeated sequences were not included. The total number of alignments per tiling fragment were plotted (this reflects coverage per segment across contig) and compared to each other.

The coverage for each junction was calculated for the *L. sativa* and *L. saligna* mitochondrial genomes following methods similar to those above. The junction-spanning assemblies, consisting of 2 kb fragments with 1 kb of sequence on either side of the junction, were used as BLAST queries versus the PacBio reads to verify the assembly and provide numerical values for the fractionation of each junction in the pool of PacBio reads. Each segment that corresponded to a particular junction generated one of two types of alignment. The first type was a completely uninterrupted alignment that was specific for a particular junction. The second type was a fragmented alignment that corresponded to junction components in different locations of the mitochondrial genome. The alignments of different lengths were plotted for each BLAST hit (sorted by alignment length). The transition from complete (1.8 kb) to shorter segmented alignments indicated the fraction of each junction type in the pool of PacBio reads.

bioRxiv preprint doi: https://doi.org/10.1101/564278; this version posted March 1, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Isoform inference and quantification

Mitochondrial contigs generated with the CLC assembler were considered as primary units after polishing. Junctions between primary structural units inferred by reverse PacBio read mapping and analysis (see above) determined their sequential order. Visual inspection of secondary building blocks and analysis of their possible inter-connections taking into account their stoichiometries resulted in the set of mitochondrial genome isoforms.

In this paper, we define contigs as primary structural units; secondary building blocks are combinations of several primary structural units (as detected by reverse read mapping); isoforms are combinations of several secondary building blocks with stoichiometry taken into account. Stoichiometry data were derived from contig coverage with PacBio and Illumina reads.

Validation of the assemblies with Illumina mate-pair libraries

For *L. sativa* and *L. serriola*, we used Illumina mate-pair libraries to validate structural accuracy of the PacBio assembly and inferred mitochondrial isoforms. Illumina mate-pair reads from genomic libraries of *L. sativa* with insert sizes of 2.5 and 10 kb were aligned to the assembled isoforms using BWA [doi: 10.1093/bioinformatics/btp324]. The distances between aligned mate-pair reads were calculated based on their positions/coordinates on the mitochondrial reference sequence. Pairs with distances greater than 1 kb were selected and compiled into subsets of equal size (320,000 pairs) for each library. Two dimensional distance heat plots for 1 kb bins were generated using modified Python scripts (https://github.com/alex-kozik/atgc-uni-cluster) and the Pixelirator visualization program (http://cgpdb.ucdavis.edu/data_pixelirator/).

Higher order arrangement analysis with Hi-C libraries

Read pairs from Hi-C libraries were filtered to eliminate ligation artifacts that do not reflect true long-distance interactions. The use of the Mbo-I restriction enzyme for fragmentation prior to ligation creates GATC-GATC dimer sites upon ligation of the Mbo-I digested termini, which do not exist in the original mitochondrial genome. Reads with GATC-GATC sites were selected and used for the long distance interaction analysis. Each read with a GATC-GATC site was split into pairs of sub-reads and each read of a split pair was mapped separately to the mitochondrial reference. The distances were calculated as in the mate read analysis. A subset of 640,000 pairs with distances greater than 1 kb was randomly selected for the distance analysis and visualization.

bioRxiv preprint doi: https://doi.org/10.1101/564278; this version posted March 1, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Repeat analysis

Repeated sequences were found as described previously (34) using BLAST-N with a word size of 50, ungapped, no masking, reward +1, penalty -20, and e-value 1,000. These parameters identified nearly identical repeats, indicating either a recent duplication, or recent homologous recombination and gene conversion of the two copies. Less similar repeats were presumed to have mutated and drifted without recent gene conversion, indicating that they are no longer engaging in productive homology-based events.

Detection of rare recombination events between repeats less than 1 kb

For the detection of rare recombination events between short repeats, all available mate pair reads of insert size 5 and 10 kb for *L. sativa* and 10 kb for *L. serriola* were used. The mitochondrial fraction of these libraries had ~1 million pairs per library for *L. sativa* and ~0.7 million for *L. serriola*. Mate-pair reads were aligned to the genomic sequences using BWA (76), the same parameters as above, and masking the second copy of the large repeats in the reference genome. Two-dimensional distance heat plots for 1 kb bins were generated as described above. Recombination events between short repeats resulted in distinct patterns of double short diagonals that are separated from the main diagonal. Variable intensity of the short diagonals was an indication of recombination frequency for particular repeats. Numerical values were analyzed in an MS Excel table as shown in Supplemental Figure S6.

Gene content, analysis, and mitochondrial genome annotation

The mitochondrial genome was annotated for expected mitochondrial features using the Mitofy web server (http://dogma.ccbb.utexas.edu/mitofy/). In order to find coding sequences missed by Mitofy, all open reading frames were manually examined. The two different junction open reading frames for *atp1* and *rps4* were translated and compared to other plant mitochondrial atp1 and rps4 protein sequences using BLAST-P (https://blast.ncbi.nlm.nih.gov/Blast.cgi).

Leucaena trichandra dataset

The *L. trichandra* dataset (63) was used to validate our approach to mitochondrial genome assembly with a different species. *L. trichandra* mitochondrial PacBio reads (SRX2719625), Illumina 4 kb mate-pair libraries (SRX2719623 and SRX2719624), and the putative autonomous mitochondrial DNA element (MH717174.1) were reanalyzed

bioRxiv preprint doi: https://doi.org/10.1101/564278; this version posted March 1, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

as described above and compared to the published *L. trichandra* mitochondrial genome assembly (MH717173.1).

Isolation of mitochondria and preparation for in-gel fluorescence microscopy and pulsed-field gel electrophoresis

Whole seven-day-old seedlings (roots and shoots) of *L. sativa* cv. Salinas grown in the dark under sterile conditions at 15°C were harvested and 5–10 g of plant material was flash frozen in liquid nitrogen. The frozen tissue was ground to a fine powder using a mortar and pestle, then ground again after adding 5 mL of High Salt Buffer (HSB: 1.25 M NaCl, 40 mM HEPES pH 7.6, 2 mM EDTA pH 8, 0.1% Bovine Serum Albumin (BSA), 0.1% 2-mercaptoethanol) until the frozen slurry melted. The liquid homogenate was filtered through 1 layer of Miracloth, then centrifuged at 4°C at 3,000 x g to pellet chloroplasts and nuclei. The supernatant was transferred to a fresh set of tubes and centrifuged at 4°C at 20,000 x g to pellet mitochondria. The mitochondrial pellet was resuspended in Liverwort Dilution Buffer (LDB; 0.4 M sorbitol, 1 mM EDTA, 0.1% BSA, 20 mM HEPES-KOH, pH 7.5) at half the volume of HSB used for the initial grinding and centrifuged at 4°C at 3,000 x g to again remove chloroplasts and nuclei by pelleting. The supernatant was transferred to a fresh set of tubes and centrifuged at 4°C at 20,000 x g to pellet mitochondria. The mitochondrial pellet was resuspended in half the original volume of LDB and the 3,000 x g centrifugation, transfer of supernatant and 20,000 x g centrifugation were repeated. The mitochondrial pellet was resuspended using 2 µL of LDB per µL of pelleted mitochondria. For PFGE, low melt point agarose (LMPA) and sorbitol were added to the suspension to achieve final concentrations of 0.7% LMPA and 0.41 M sorbitol before pipetting into an agarose plug mold, which was allowed to cool for 20 min at 4°C. For in-gel fluorescence microscopy, agarose plugs of the mitochondria were prepared as for PFGE, but the mitochondria were first diluted 1:20, 1:100, or 1:200 in LDB before embedding. Samples for PFGE were run on a 1.5% agarose gel in 0.5X Tris-Borate-EDTA (TBE) buffer at 5 V/cm with a pulse time of 30 s for 24 hours and stained in 3X Gel Red (Biotium) before imaging. For in-gel fluorescence microscopy, ¼ of an agarose plug was soaked in BET solution (3% 2-mercaptoethanol, 0.1 ug/mL EtBr, 1X TBE) for 30 mins. The BET solution was removed and the sample was soaked in fresh BET before slide preparation. Alternatively, ¼ of an agarose plug was soaked in GET solution (3% 2-mercaptoethanol, 0.5x QuantiFluor® dye (Promega), 1X TBE) for 30 mins. To prepare slides, ½ of the stained agarose plug was placed on top of a glass slide on a heat block at 60–65°C and mixed with 15 µL ABET (1% LMPA in BET) and a coverslip was placed on top. When the agarose was completely melted underneath the coverslip, the slides were removed from the heat block and sealed with nail polish before microscopy.

bioRxiv preprint doi: https://doi.org/10.1101/564278; this version posted March 1, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Acknowledgments

We thank Luca Comai and Arnold Bendich for helpful comments on the manuscript and Elizabeth Georgian for editorial assistance. We are also grateful to Pauline Sanders for supplying and maintaining seed stocks, and Huaqin Xu for assistance with data submission to NCBI GenBank.

FIGURE LEGENDS

Figure 1. Primary structural units of the *Lactuca* mitochondrial genome and secondary building blocks.

A) Comparison of primary structural units between the *L. sativa* and *L. saligna* mitochondrial genomes. Polished contigs generated from CLC assembly form the primary structural units of the mitochondrial genome and are shown in pairs with *L sativa* on top and *L. saligna* on the bottom. The stoichiometry (1x or 2x) and whether the units were common or unique between *L. sativa* and *L. saligna* are indicated. Different termini of a pair of contigs are labeled with an asterisk. The differences between termini can be up to several hundred nucleotides. A dark green bar indicates a large insertion from the chloroplast genome. The contig from *L.saligna*, UV, was a single contig that correspond to major parts of contigs U and W of *L.sativa* and was conceptually split into two segments (color coding) to simplify and clarify the visualization and interpretation of contig relationships. A segment of the S inverted repeat sequence is duplicated in both species on the termini of units W and V, and in unit N of *L. sativa* (yellow arrows). B-F) Secondary building blocks of mitochondrial genomes with a defined sequential order of primary units determined by reverse read mapping (see Materials and Methods) for *L. sativa* (B-D) and *L. saligna* (E,F). Junctions within boxes indicate the regions that were used to count PacBio reads containing particular junctions.

Figure 2. Isoforms of *Lactuca* mitochondrial genomes.

Putative major isoforms of *L. sativa* (A) and *L. saligna* (B) were derived from analysis of secondary building blocks (Figure 1 B-F) and the primary structural unit stoichiometry (Figure 1A, Figure S2). For *L. sativa*, the dashed green arrow in A shows rearrangement of unit U and the dashed yellow arrow shows rearrangement of unit P between the two major isoforms. For *L. saligna*, putative recombination events that result in a transition from one isoform to another are indicated by solid blue arrows. Isoforms are labeled by the configuration of the units M, S, and Z, with > and < symbols designating where recombination has taken place. Yellow triangles indicate repeat X-

bioRxiv preprint doi: https://doi.org/10.1101/564278; this version posted March 1, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

01. Displayed isoforms are simplified models that fit the sequential order of primary structural units and stoichiometry data and make no assumptions about the underlying form of mtDNA molecules.

Figure 3. *Lactuca* mitochondrial genome annotations

The annotations for genes and other sequence features for *L. sativa* and *L. saligna* are displayed along the primary structural units (see Figure 1A) for each genome, which are indicated by thick gray arrows. Intronless genes are indicated by red arrows, exons of spliced genes are indicated by blue arrows, and plastid insertions are indicated by green arrows. Genes that span junctions between primary units are indicated by a jagged line at the division point. Thin gray arrows show alternative junctions between primary structural units that result in different models for the genes that are split over a junction.

Figure 4. Long distance analysis of mitochondrial genomes using Illumina mate pair and Hi-C libraries.

Plots of distances between read-pairs Illumina mate pair libraries with 2.5 and 10 kb insert sizes in 1 kb bins for *L. sativa* (A) and *L. serriola* (B) exhibit essentially identical long-distance sequence connections. C) Plot of Hi-C contact frequencies for *L. sativa*. Major isoform α was arbitrarily chosen as a reference for visualization of mate-pair distances and Hi-C contact frequency. The color gradient displayed below panels reflects the number of read-pairs (out of a total of 320,000) in each 2-dimensional-1-kb bin and applies to all panels.

Figure 5. Detection of minor isoforms using Illumina mate pair data.

Above, schematic illustrations of the transition between major isoform α and two minor isoforms through recombination at repeat X-01b (yellow triangles and blue arrows). Below, high-resolution-Illumina-mate-pair plot showing sequence connections (indicated by gray ovals) that are expected for the minor isoforms shown above. Mate-pair library sizes were 2.5 and 10 kb. The major mitochondrial isoform of *L. sativa* includes the following order of basic units: M-N-K-M-N-W-Z. There is a medium size repeat X-01b of length ~500 bp (yellow triangle) at the N and W termini that can cause two distinct recombination events as shown on the right side of the figure. Upon X-01b recombination, two new isoforms have three distinct junctions M-W (common for both), W-N-Z, and K-N-Z (specific for each isoform). A long distance plot of mate-pair libraries clearly demonstrated the existence of both isoforms and all three junctions (left part of the figure). Ovals highlight the diagonals of long distance interactions that were evidence of the existence of minor isoforms for the *L. sativa* mitochondrial genome.

bioRxiv preprint doi: https://doi.org/10.1101/564278; this version posted March 1, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Similar patterns were detected using *L. serriola* mate pair libraries (Figure S5A). The color gradient displayed below panels reflects the total number of read-pairs (out of a total of 320,000) in each 2-dimensional-1-kb bin and applies to all panels.

Figure 6. Interconversion between *L. sativa* and *L. saligna* genomes. Linear representation of one of the *L. saligna* major mitochondrial genome isoforms on the top. The middle structure is a putative minor isoform of *L. saligna* derived from recombination at repeat X-07 (black triangles), which leads to a similar arrangement of primary structural units as that of *L. sativa* major isoform α (bottom). Collinear segments (longer than 500 bp) are indicated by gray shading.

Figure 7. Structural analysis of *L. sativa* mitochondrial DNA by in-gel fluorescence microscopy and pulsed-field gel electrophoresis. (A-F) DNA obtained from mitochondria isolated from seven-day-old dark-grown seedlings (roots and shoots) was stained with either ethidium bromide or QuantiFluor®□ dye. Images are representative of branched linear (A), circular (B), linear (C), degraded (D), comet (E), and branched circular (F) structures. The scale bar in panel F applies to panels A-F and is 10 μm, corresponding to the length of approximately 30 kb of DNA. Branched linear: interconnected linear forms with or without a densely staining central core. Circular: closed loop without any additional branches. Linear: linear fiber with no branches. Degraded: many small molecules of undetermined structure. Comet: bright core with short connected fibers and no other visible branch points. Branched Circular: Closed loop structure with linear branches. G) Quantification of the primary structures observed among 98 total microscopic fields. H) Pulsed-field gel electrophoresis. Each lane represents an independent gel run. Sizes (in kb) are determined from the migration of lambda concatemers run in each gel experiment. In some cases, the DNA migrated at a slight angle to the left.

bioRxiv preprint doi: https://doi.org/10.1101/564278; this version posted March 1, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure S1. Mitochondrial genome assembly and analysis strategy. The flowchart summarizes all of the key steps of the *Lactuca* mitochondrial genome project along with the source and type of raw data used in each iteration of the assembly and analysis.

Figure S2. Results of the stoichiometry analysis for primary structural units. A tiling library of overlapping fragments (similar to that used for reverse read mapping) was used to estimate coverage for each fragment (see Contig Stoichiometry in Materials and Methods). The Y axis shows coverage; the X axis displays the coordinates of the primary structural units. Note the elevated coverage in the middle portion in the majority of basic units. Stoichiometry values for each primary structural unit were used for the mitochondrial genome isoform modeling.

Figures S3. Stoichiometry analysis for each detected junction between different primary structural units. The set (library) of all identified junctions (2 kb each, see Materials and Methods, Mitochondrial Contig Stoichiometry) was analyzed for the fraction of uninterrupted alignments within a pool of most informative PacBio reads for *L. sativa* (A) and *L. saligna* **(**B). All BLAST-N alignments are plotted on the X axis and sorted according to alignment lengths. The length of each alignment is shown on the Y axis. Breakpoints (transitions between uninterrupted alignments longer than 1.8 kb) are an indication of the abundance of each particular junction in a pool of PacBio reads. This value (number of uninterrupted alignments) reflects the proportion of any particular junction within isoforms of a mitochondrial genome.

Figure S4. High resolution images shown in Figure 4.

Figure S5. Detection of recombination events at repeats of up to several hundred bp using Illumina mate pair libraries. Plots of distances between read-pairs in 1 kb bins for *L. sativa* mate pair libraries of 5 kb (A) and 10 kb insert sizes (B), and *L. serriola* 10 kb (C). Only unique read mappings were selected from all available reads for the analysis and data interpretation.

Figure S6. Example of numerical values for detection of recombination between short repeats X-01b and X-03 using the 5 kb mate-pair *L. sativa* library (see Figure S5A for the complete 2D plot). Each cell is a 1 kb bin across the mitochondrial genome isoform α. Values within each cell give the number of times mate-pair reads mapped within the same bin. The majority of mate-pair reads are mapped and equally distributed over the main diagonal and within large repeats (R or T). Rare recombination events between short repeats generate distinct shapes (shown as X-01b and X-03) that are located away from the main diagonal. The values reflect the frequency of

bioRxiv preprint doi: https://doi.org/10.1101/564278; this version posted March 1, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

recombination. Thus, for repeat X-01b, it could be estimated that corresponding recombination frequency is ~10% and ~1% for repeat X-03.

Figure S7. Mitochondrial DNA branched linear structures. Additional examples of molecules scored in Figure 7G.

Figure S8. Reanalysis of the mitochondrial genome of *Leucaena trichandra* with PacBio and Illumina mate-pair reads. Fourteen contigs derived after CLC assembly of the selected PacBio reads (SRX2719625) were ordered using the published linear assembly accession MH717173 along with autonomous element X (accession MH717174, contig 10) using additional information about the sequential order of contigs based on read-through data (reverse read mapping) and long distance analysis with mate-pair reads. Long distance analysis with mate-pair reads clearly demonstrated that the organization of *Leucaena trichandra* mitochondrial genome can by represented in the form of a cyclic graph. All genome segments are linked to each other through several repeats with potential complex rearrangements. Former autonomous element X was placed between a set of repeats that have different segmentation in other parts of the genome.

Figure S9. Summary of Reverse Read Mapping approach. Explanation and data interpretation of the reverse read mapping approach. Panel A: Scheme of overlapping tiling fragments for primary structural units and reverse mapping protocol outline. Tiling fragments were used as queries in BLAST-N searches versus a database of mitochondrial PacBio reads. Results of BLAST-N were parsed and exported into an MS Excel table as shown in (B). Visual inspection of the distribution of primary structural units over long PacBio reads in an MS Excel table with subsequent search queries on text files provided information about the sequential order of primary structural units within the PacBio reads. This ultimately led to the identification of secondary building blocks (see example of distinct L-T-P and L-T-U blocks detected on a set of PacBio reads).

bioRxiv preprint doi: https://doi.org/10.1101/564278; this version posted March 1, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

REFERENCES

1. Quetier F, Vedel F. Heterogeneous population of mitochondrial DNA molecules in higher plants. Nature. 1977 Jul 28;268(5618):365–8.

2. Belliard G, Vedel F, Pelletier G. Mitochondrial recombination in cytoplasmic hybrids of Nicotiana tabacum by protoplast fusion. Nature. 1979 Oct 4;281(5730):401–3.

3. Ward BL, Anderson RS, Bendich AJ. The mitochondrial genome is large and variable in a family of plants (cucurbitaceae). Cell. 1981 Sep;25(3):793–803.

4. Bendich AJ. Reaching for the ring: the study of mitochondrial genome structure. Curr Genet. 1993 Oct;24(4):279–90.

5. Oldenburg DJ, Bendich AJ. Size and Structure of Replicating Mitochondrial DNA in Cultured Tobacco Cells. Plant Cell. 1996 Mar;8(3):447–61.

6. Backert S, Börner T. Phage T4-like intermediates of DNA replication and recombination in the mitochondria of the higher plant Chenopodium album (L.). Curr Genet. 2000 May;37(5):304–14.

7. Sloan DB. One ring to rule them all? Genome sequencing provides new insights into the “master circle” model of plant mitochondrial DNA structure. New Phytol. 2013 Dec;200(4):978–85.

8. Oldenburg DJ, Bendich AJ. DNA maintenance in plastids and mitochondria of plants. Front Plant Sci. 2015 Oct 29;6:883.

9. Adams KL, Qiu Y-L, Stoutemyer M, Palmer JD. Punctuated evolution of mitochondrial gene content: high and variable rates of mitochondrial gene loss and transfer to the nucleus during angiosperm evolution. Proc Natl Acad Sci U S A. 2002 Jul 23;99(15):9905–12.

10. Adams KL, Palmer JD. Evolution of mitochondrial gene content: gene loss and transfer to the nucleus. Mol Phylogenet Evol. 2003 Dec;29(3):380–95.

11. Sloan DB, Alverson AJ, Chuckalovcak JP, Wu M, McCauley DE, Palmer JD, et al. Rapid evolution of enormous, multichromosomal genomes in flowering plant mitochondria with exceptionally high mutation rates. PLoS Biol. 2012 Jan;10(1):e1001241.

12. Skippington E, Barkman TJ, Rice DW, Palmer JD. Miniaturized mitogenome of the parasitic plant Viscum scurruloideum is extremely divergent and dynamic and has lost all nad genes. Proc Natl Acad Sci U S A. 2015 Jul 7;112(27):E3515–24.

13. Petersen G, Cuenca A, Zervas A, Ross GT, Graham SW, Barrett CF, et al. Mitochondrial genome evolution in Alismatales: Size reduction and extensive loss of ribosomal protein genes. PLoS One. 2017 May 17;12(5):e0177606.

bioRxiv preprint doi: https://doi.org/10.1101/564278; this version posted March 1, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

14. Van Aken O, Van Breusegem F. Licensed to Kill: Mitochondria, Chloroplasts, and Cell Death. Trends Plant Sci. 2015 Nov;20(11):754–66.

15. Liberatore KL, Dukowic-Schulze S, Miller ME, Chen C, Kianian SF. The role of mitochondria in plant development and stress tolerance. Free Radic Biol Med. 2016 Nov;100:238–56.

16. Kim Y-J, Zhang D. Molecular Control of Male Fertility for Crop Hybrid Breeding. Trends Plant Sci. 2018 Jan;23(1):53–65.

17. Siqueira JA, Hardoim P, Ferreira PCG, Nunes-Nesi A, Hemerly AS. Unraveling Interfaces between Energy Metabolism and Cell Cycle in Plants. Trends Plant Sci. 2018 Aug;23(8):731–47.

18. Palmer JD, Shields CR. Tripartite structure of the Brassica campestris mitochondrial genome. Nature. 1984 Feb 2;307(5950):437–40.

19. Palmer JD, Herbon LA. Tricircular mitochondrial genomes of Brassica and Raphanus: reversal of repeat configurations by inversiton. Nucleic Acids Res. 1986 Dec 9;14(24):9755–64.

20. Lonsdale DM, Hodge TP, Fauron CM-R. The physical map and organisatlon of the mitochondrial genome from the fertile cytoplasm of maize. Nucleic Acids Res. 1984 Dec 21;12(24):9249–61.

21. Bendich AJ. Structural analysis of mitochondrial DNA molecules from fungi and plants using moving pictures and pulsed-field gel electrophoresis. J Mol Biol. 1996 Feb 2;255(4):564–88.

22. Oldenburg DJ, Bendich AJ. The structure of mitochondrial DNA from the liverwort, Marchantia polymorpha. J Mol Biol. 1998 Mar 6;276(4):745–58.

23. Manchekar M, Scissum-Gunn K, Song D, Khazi F, McLean SL, Nielsen BL. DNA recombination activity in soybean mitochondria. J Mol Biol. 2006 Feb 17;356(2):288–99.

24. Mower JP, Case AL, Floro ER, Willis JH. Evidence against equimolarity of large repeat arrangements and a predominant master circle structure of the mitochondrial genome from a monkeyflower (Mimulus guttatus) lineage with cryptic CMS. Genome Biol Evol. 2012;4(5):670–86.

25. Cheng N, Lo Y-S, Ansari MI, Ho K-C, Jeng S-T, Lin N-S, et al. Correlation between mtDNA complexity and mtDNA replication mode in developing cotyledon mitochondria during mung bean seed germination. New Phytol. 2017 Jan;213(2):751–63.

26. Wolfe KH, Li WH, Sharp PM. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc Natl Acad Sci U S A.

bioRxiv preprint doi: https://doi.org/10.1101/564278; this version posted March 1, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

1987 Dec;84(24):9054–8.

27. Palmer JD, Herbon LA. Plant mitochondrial DNA evolved rapidly in structure, but slowly in sequence. 1988 Dec [cited 2017 Jun 13]; Available from: http://hdl.handle.net/2027.42/48042

28. Drouin G, Daoud H, Xia J. Relative rates of synonymous substitutions in the mitochondrial, chloroplast and nuclear genomes of seed plants. Mol Phylogenet Evol. 2008 Dec;49(3):827–31.

29. Richardson AO, Rice DW, Young GJ, Alverson AJ, Palmer JD. The “fossilized” mitochondrial genome of Liriodendron tulipifera: ancestral gene content and order, ancestral editing sites, and extraordinarily low mutation rate. BMC Biol. 2013;11(1):29.

30. Kubo T, Nishizawa S, Sugawara A, Itchoda N, Estiati A, Mikami T. The complete nucleotide sequence of the mitochondrial genome of sugar beet (Beta vulgaris L.) reveals a novel gene for tRNACys(GCA). Nucleic Acids Res. 2000 Jul 1;28(13):2571–6.

31. Satoh M, Kubo T, Nishizawa S, Estiati A, Itchoda N, Mikami T. The cytoplasmic male-sterile type and normal type mitochondrial genomes of sugar beet share the same complement of genes of known function but differ in the content of expressed ORFs. Mol Genet Genomics. 2004 Oct;272(3):247–56.

32. Maréchal A, Brisson N. Recombination and the maintenance of plant organelle genome stability. New Phytol. 2010 Apr;186(2):299–317.

33. Cole LW, Guo W, Mower JP, Palmer JD. High and variable rates of repeat-mediated mitochondrial genome rearrangement in a genus of plants. Mol Biol Evol [Internet]. 2018 Sep 7; Available from: http://dx.doi.org/10.1093/molbev/msy176

34. Wynn EL, Christensen AC. Repeats of Unusual Size in Plant Mitochondrial Genomes: Identification, Incidence and Evolution. G3 . 2019 Feb 7;9(2):549–59.

35. Rice DW, Alverson AJ, Richardson AO, Young GJ, Sanchez-Puerta MV, Munzinger J, et al. Horizontal transfer of entire genomes via mitochondrial fusion in the angiosperm Amborella. Science. 2013 Dec 20;342(6165):1468–73.

36. Shearman JR, Sangsrakru D, Ruang-Areerate P, Sonthirod C, Uthaipaisanwong P, Yoocha T, et al. Assembly and analysis of a male sterile rubber tree mitochondrial genome reveals DNA rearrangement events and a novel transcript. BMC Plant Biol. 2014 Feb 10;14:45.

37. Backert S, Lynn Nielsen B, Börner T. The mystery of the rings: structure and replication of mitochondrial genomes from higher plants. Trends Plant Sci. 1997 Dec 1;2(12):477–83.

bioRxiv preprint doi: https://doi.org/10.1101/564278; this version posted March 1, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

38. Klein M, Eckert-Ossenkopp U, Schmiedeberg I, Brandt P, Unseld M, Brennicke A, et al. Physical mapping of the mitochondrial genome of Arabidopsis thaliana by cosmid and YAC clones. Plant J. 1994 Sep;6(3):447–55.

39. Unseld M, Marienfeld JR, Brandt P, Brennicke A. The mitochondrial genome of Arabidopsis thaliana contains 57 genes in 366,924 nucleotides. Nat Genet. 1997 Jan;15(1):57–61.

40. Shearman JR, Sonthirod C, Naktang C, Pootakham W, Yoocha T, Sangsrakru D, et al. The two chromosomes of the mitochondrial genome of a sugarcane cultivar: assembly and recombination analysis using long PacBio reads. Sci Rep. 2016 Aug 17;6:31533.

41. Shedge V, Arrieta-Montiel M, Christensen AC, Mackenzie SA. Plant mitochondrial recombination surveillance requires unusual RecA and MutS homologs. Plant Cell. 2007 Apr;19(4):1251–64.

42. Arrieta-Montiel MP, Shedge V, Davila J, Christensen AC, Mackenzie SA. Diversity of the Arabidopsis mitochondrial genome occurs via nuclear-controlled recombination activity. Genetics. 2009 Dec;183(4):1261–8.

43. Kühn K, Gualberto JM. Recombination in the Stability, Repair and Evolution of the Mitochondrial Genome. Adv Bot Res. 2012 Jan 1;63:215–52.

44. Janicka S, Kühn K, Le Ret M, Bonnard G, Imbault P, Augustyniak H, et al. A RAD52-like single-stranded DNA binding protein affects mitochondrial DNA repair by recombination. Plant J. 2012 Nov;72(3):423–35.

45. Wallet C, Le Ret M, Bergdoll M, Bichara M, Dietrich A, Gualberto JM. The RECG1 DNA Translocase Is a Key Factor in Recombination Surveillance, Repair, and Segregation of the Mitochondrial DNA in Arabidopsis. Plant Cell. 2015 Oct;27(10):2907–25.

46. Skippington E, Barkman TJ, Rice DW, Palmer JD. Comparative mitogenomics indicates respiratory competence in parasitic Viscum despite loss of complex I and extreme sequence divergence, and reveals horizontal gene transfer and remarkable variation in genome size. BMC Plant Biol. 2017 Feb 21;17(1):49.

47. Handa H. The complete nucleotide sequence and RNA editing content of the mitochondrial genome of rapeseed (Brassica napus L.): comparative analysis of the mitochondrial genomes of rapeseed and Arabidopsis thaliana. Nucleic Acids Res. 2003 Oct 15;31(20):5907–16.

48. Paillard M, Sederoff RR, Levings CS III. Nucleotide sequence of the S-1 mitochondrial DNA from the S cytoplasm of maize. EMBO J. 1985 May 1;4(5):1125–8.

49. Levings CS, Sederoff RR. Nucleotide sequence of the S-2 mitochondrial DNA from

bioRxiv preprint doi: https://doi.org/10.1101/564278; this version posted March 1, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

the S cytoplasm of maize. Proc Natl Acad Sci U S A. 1983 Jul 1;80(13):4055–9.

50. Palmer JD, Shields CR, Cohen DB, Orton TJ. An unusual mitochondrial DNA plasmid in the genus Brassica. Nature. 1983 Feb 24;301(5902):301725a0.

51. Pring DR, Conde MF, Schertz KF, Levings CS. Plasmid-like DNAs associated with mitochondria of cytoplasmic male-sterile Sorghum. Mol Gen Genet. 1982 Jul 1;186(2):180–4.

52. Robison MM, Wolyn DJ. A mitochondrial plasmid and plasmid-like RNA and DNA polymerases encoded within the mitochondrial genome of carrot (Daucus carota L.). Curr Genet. 2005 Jan;47(1):57–66.

53. McDermott P, Connolly V, Kavanagh TA. The mitochondrial genome of a cytoplasmic male sterile line of perennial ryegrass (Lolium perenne L.) contains an integrated linear plasmid-like element. Theor Appl Genet. 2008 Aug;117(3):459–70.

54. Warren JM, Simmons MP, Wu Z, Sloan DB. Linear Plasmids and the Rate of Sequence Evolution in Plant Mitochondrial Genomes. Genome Biol Evol. 2016 Jan 11;8(2):364–74.

55. Vargas OM, Ortiz EM, Simpson BB. Conflicting phylogenomic signals reveal a pattern of reticulate evolution in a recent high-Andean diversification (Asteraceae: Astereae: Diplostephium). New Phytol. 2017 Jun;214(4):1736–50.

56. Grassa CJ, Ebert DP, Kane NC, Rieseberg LH. Complete Mitochondrial Genome Sequence of Sunflower (Helianthus annuus L.). Genome Announc [Internet]. 2016 Sep 15;4(5). Available from: http://dx.doi.org/10.1128/genomeA.00981-16

57. Goremykin VV, Salamini F, Velasco R, Viola R. Mitochondrial DNA of Vitis vinifera and the issue of rampant horizontal gene transfer. Mol Biol Evol. 2009 Jan;26(1):99–110.

58. Iorizzo M, Senalik D, Szklarczyk M, Grzebelus D, Spooner D, Simon P. De novo assembly of the carrot mitochondrial genome using next generation sequencing of whole genomic DNA provides first evidence of DNA transfer into an angiosperm plastid genome. BMC Plant Biol. 2012 May 1;12:61.

59. Bendich AJ. The size and form of chromosomes are constant in the nucleus, but highly variable in bacteria, mitochondria and chloroplasts. Bioessays. 2007 May;29(5):474–83.

60. Wu Z, Cuthbert JM, Taylor DR, Sloan DB. The massive mitochondrial genome of the angiosperm Silene noctiflora is evolving by gain or loss of entire chromosomes. Proc Natl Acad Sci U S A. 2015 Aug 18;112(33):10185–91.

61. Chin C-S, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, et al. Phased diploid genome assembly with single-molecule real-time sequencing. Nat

bioRxiv preprint doi: https://doi.org/10.1101/564278; this version posted March 1, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Methods. 2016 Dec;13(12):1050–4.

62. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 2017 May;27(5):722–36.

63. Kovar L, Nageswara-Rao M, Ortega-Rodriguez S, Dugas DV, Straub S, Cronn R, et al. PacBio-based mitochondrial genome assembly of Leucaena trichandra (Leguminosae) and an intrageneric assessment of mitochondrial RNA editing. Genome Biol Evol [Internet]. 2018 Aug 20 [cited 2018 Aug 23]; Available from: https://academic.oup.com/gbe/advance-article/doi/10.1093/gbe/evy179/5076815

64. Cappadocia L, Maréchal A, Parent J-S, Lepage E, Sygusch J, Brisson N. Crystal structures of DNA-Whirly complexes and their role in Arabidopsis organelle genome repair. Plant Cell. 2010 Jun;22(6):1849–67.

65. Parent J-S, Lepage E, Brisson N. Divergent roles for the two PolI-like organelle DNA polymerases of Arabidopsis. Plant Physiol. 2011 May;156(1):254–62.

66. Miller-Messmer M, Kühn K, Bichara M, Le Ret M, Imbault P, Gualberto JM. RecA-dependent DNA repair results in increased heteroplasmy of the Arabidopsis mitochondrial genome. Plant Physiol. 2012 May;159(1):211–26.

67. Davila JI, Arrieta-Montiel MP, Wamboldt Y, Cao J, Hagmann J, Shedge V, et al. Double-strand break repair processes drive evolution of the mitochondrial genome in Arabidopsis. BMC Biol. 2011 Sep 27;9:64.

68. Handa H. Linear plasmids in plant mitochondria: peaceful coexistences or malicious invasions? Mitochondrion. 2008 Jan;8(1):15–25.

69. Iorizzo M, Ellison S, Senalik D, Zeng P, Satapoomin P, Huang J, et al. A high-quality carrot genome assembly provides new insights into carotenoid accumulation and asterid genome evolution. Nat Genet. 2016 Jun;48(6):657–66.

70. Allen JO, Fauron CM, Minx P, Roark L, Oddiraju S, Lin GN, et al. Comparisons among two fertile and three male-sterile mitochondrial genomes of maize. Genetics. 2007 Oct;177(2):1173–92.

71. Rogers SO, Bendich AJ. Extraction of DNA from milligram amounts of fresh, herbarium and mummified plant tissues. Plant Mol Biol. 1985 Mar;5(2):69–76.

72. Chang S, Puryear J, and Cairney J. A Simple and Efficient Method for Isolating RNA from Pine Trees. Plant Molecular Biology Reporter. 1993;11(2):113–6.

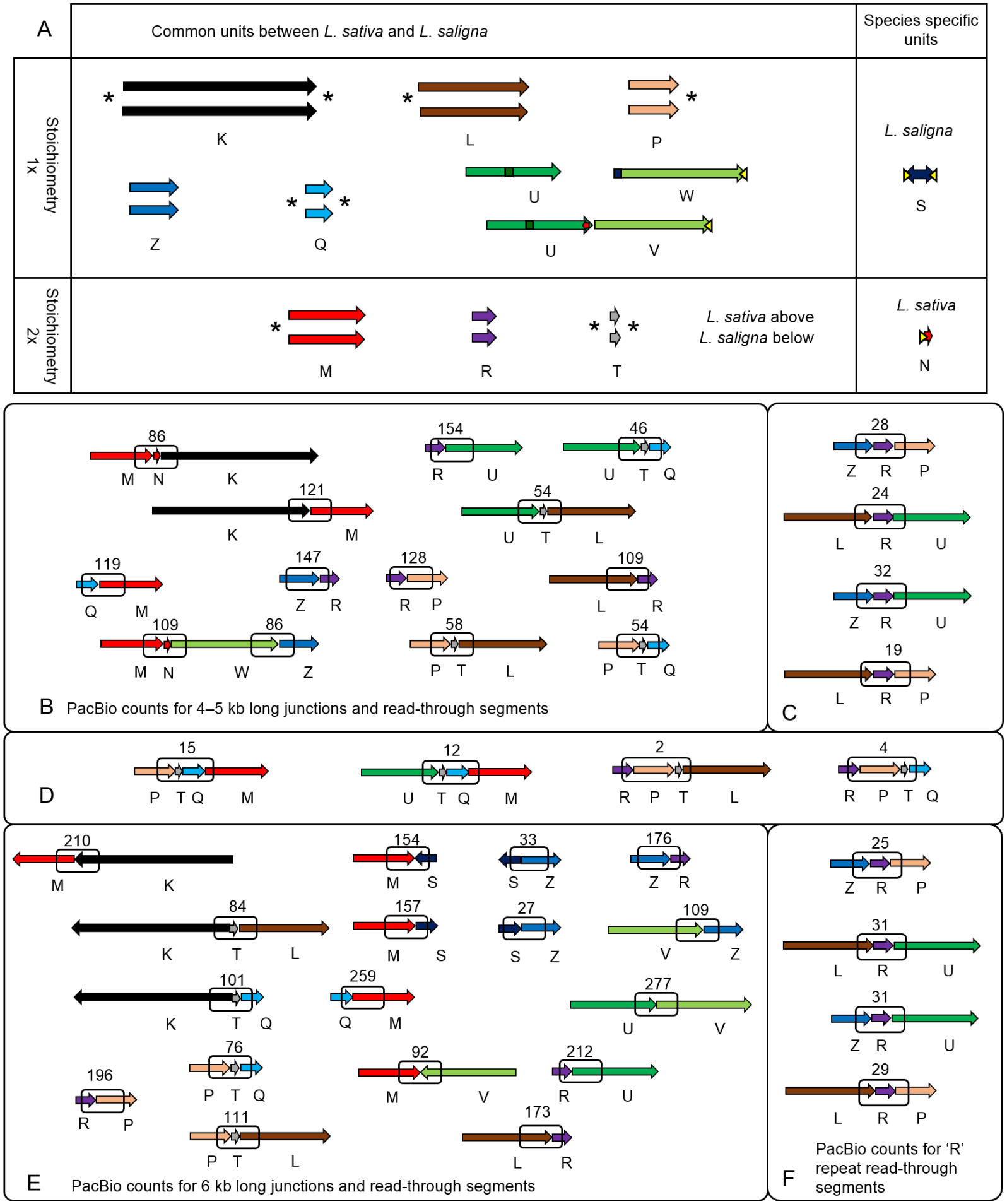
73. Reyes-Chin-Wo S, Wang Z, Yang X, Kozik A, Arikit S, Song C, et al. Genome assembly with in vitro proximity ligation data and whole-genome triplication in lettuce. Nat Commun. 2017 Apr 12;8:14953.

bioRxiv preprint doi: https://doi.org/10.1101/564278; this version posted March 1, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

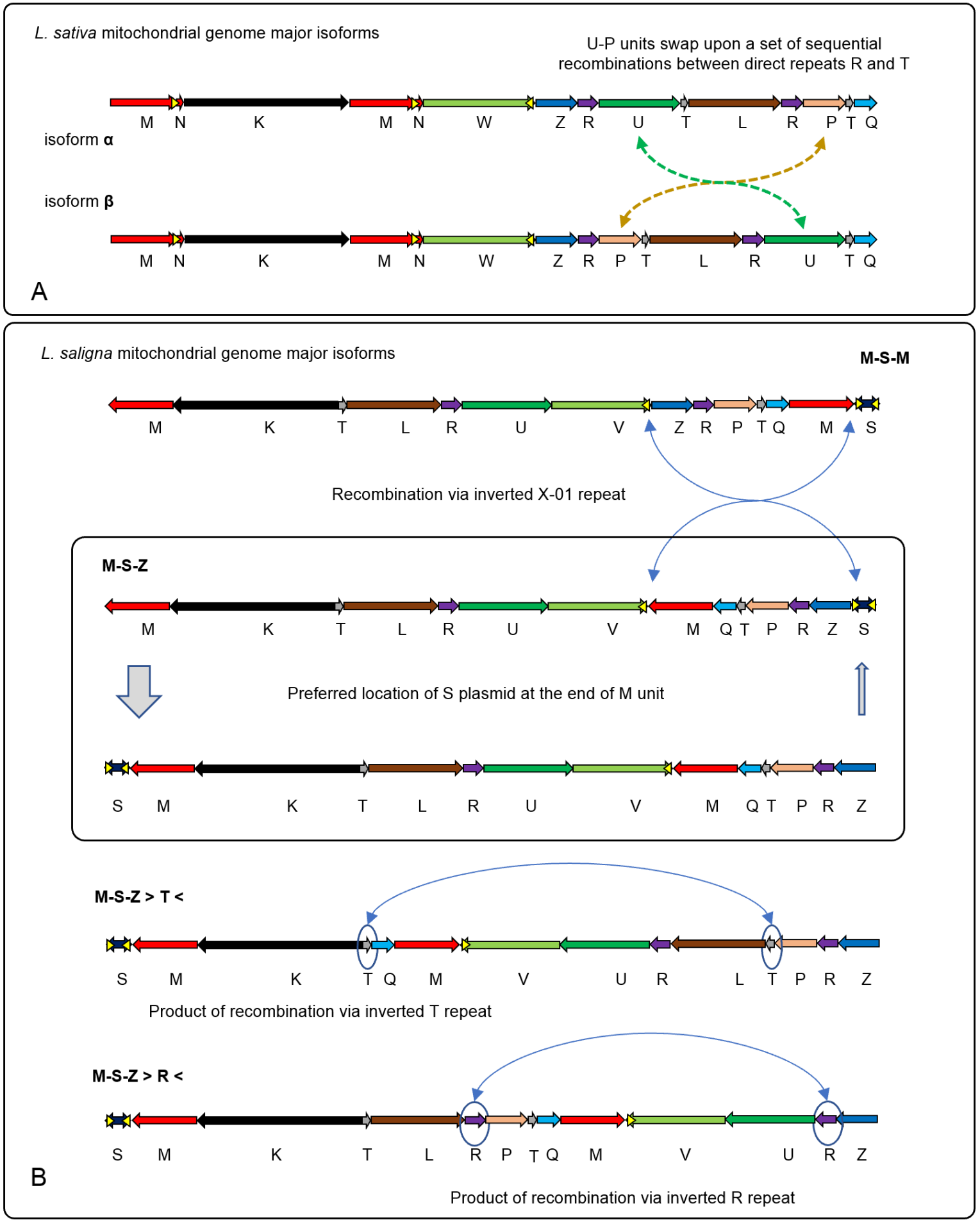
74. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008 May;18(5):821–9.

75. Huang X, Madan A. CAP3: A DNA sequence assembly program. Genome Res. 1999 Sep;9(9):868–77.

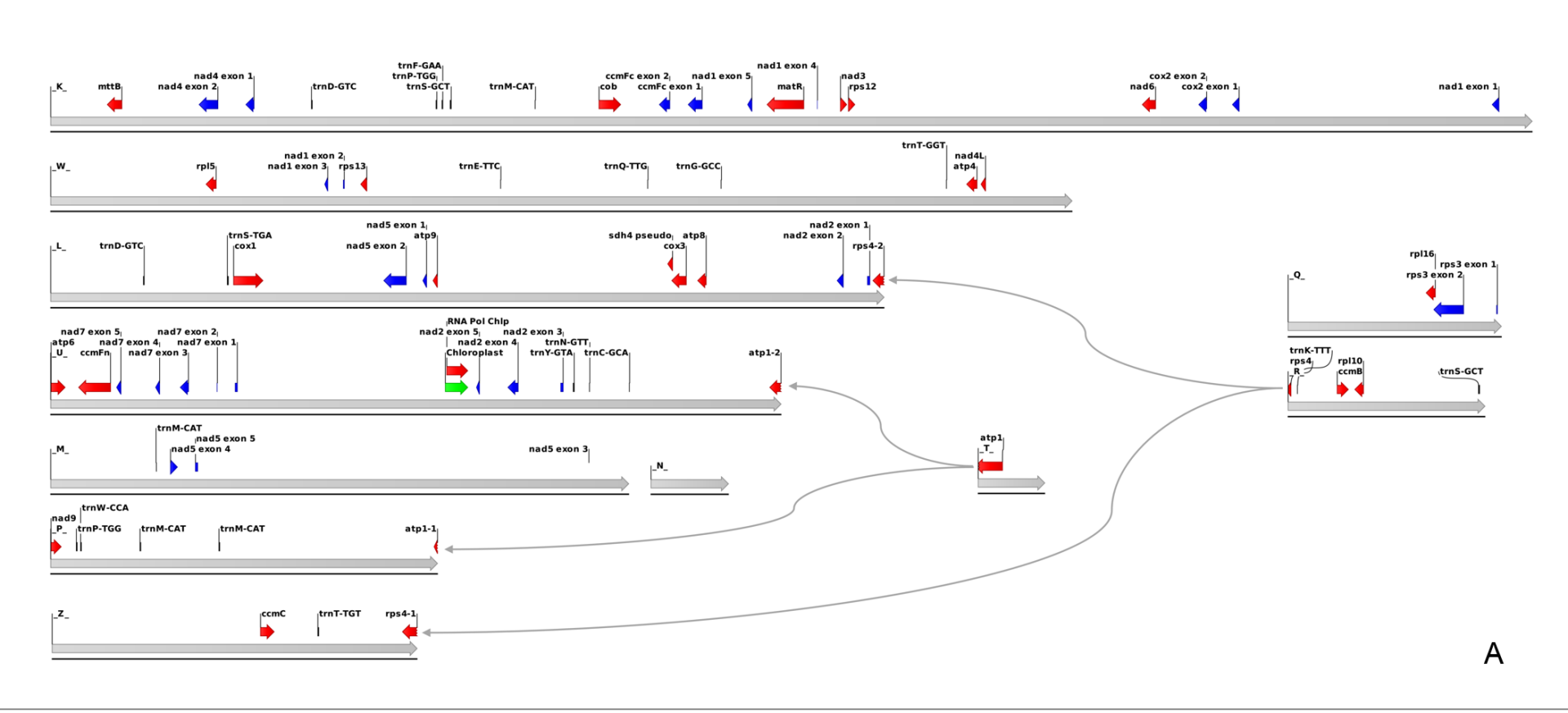
76. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010 Mar 1;26(5):589–95.

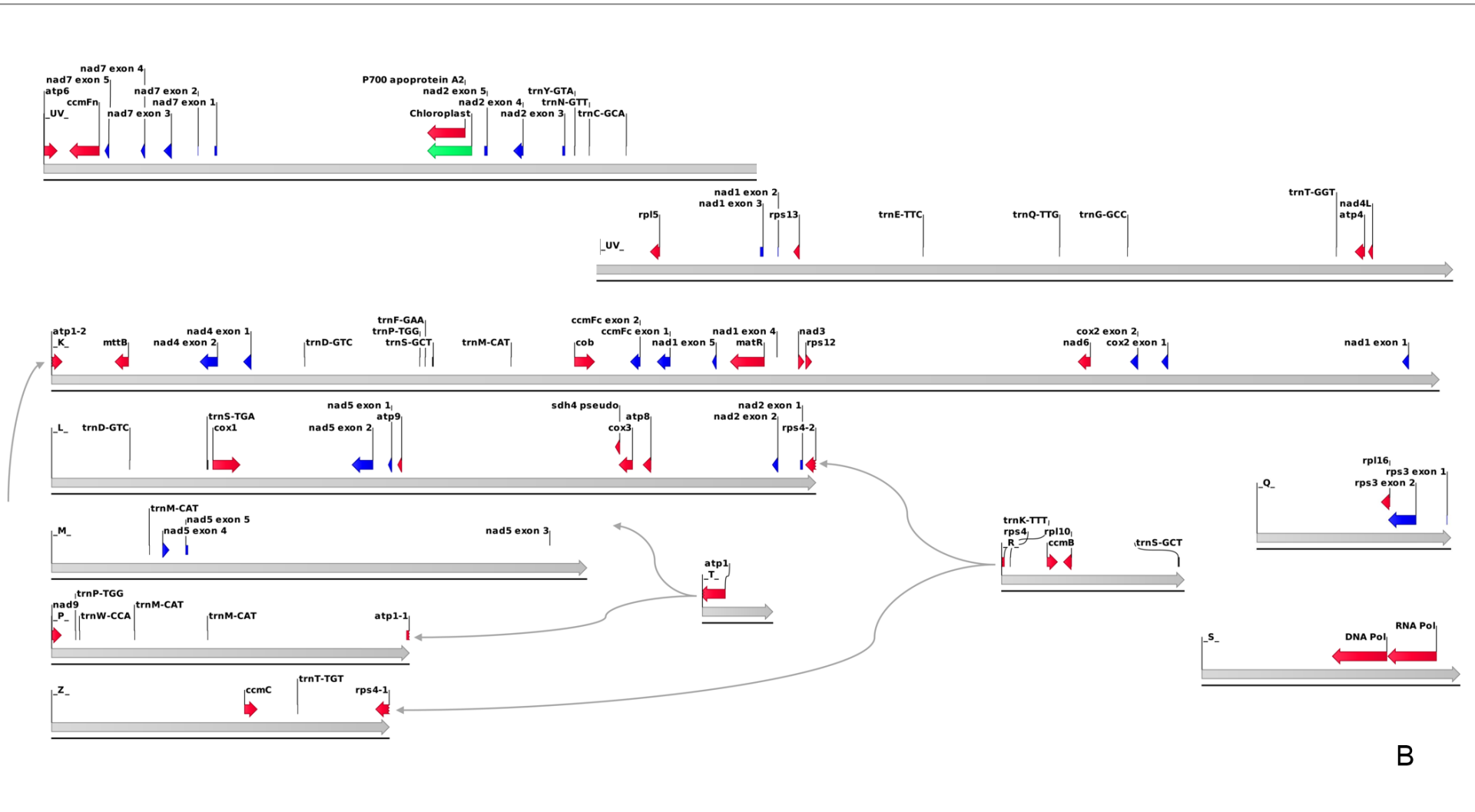

A
Common units between L. sativa and L. saligna
Species specific units
Stoichiometry 1x
K
L
P
L. saligna
S
U
W
Z
Q
U
V
Stoichiometry 2x
M
R
T
L. sativa above
L. saligna below
L. sativa
N
B PacBio counts for 4–5 kb long junctions and read-through segments
C
D
E PacBio counts for 6 kb long junctions and read-through segments
F PacBio counts for 'R' repeat read-through segments

F1


L. sativa mitochondrial genome major isoforms
U-P units swap upon a set of sequential recombinations between direct repeats R and T
isoform α
M N K M N W Z R U T L R P T Q
isoform β
M N K M N W Z R P T L R U T Q
A
L. saligna mitochondrial genome major isoforms
M-S-M
M K T L R U V Z R P T Q M S
Recombination via inverted X-01 repeat
M-S-Z
M K T L R U V M Q T P R Z S
Preferred location of S plasmid at the end of M unit
S M K T L R U V M Q T P R Z
M-S-Z > T <
S M K T Q M V U R L T P R Z
Product of recombination via inverted T repeat
M-S-Z > R <
S M K T L R P T Q M V U R Z
B
Product of recombination via inverted R repeat

F2

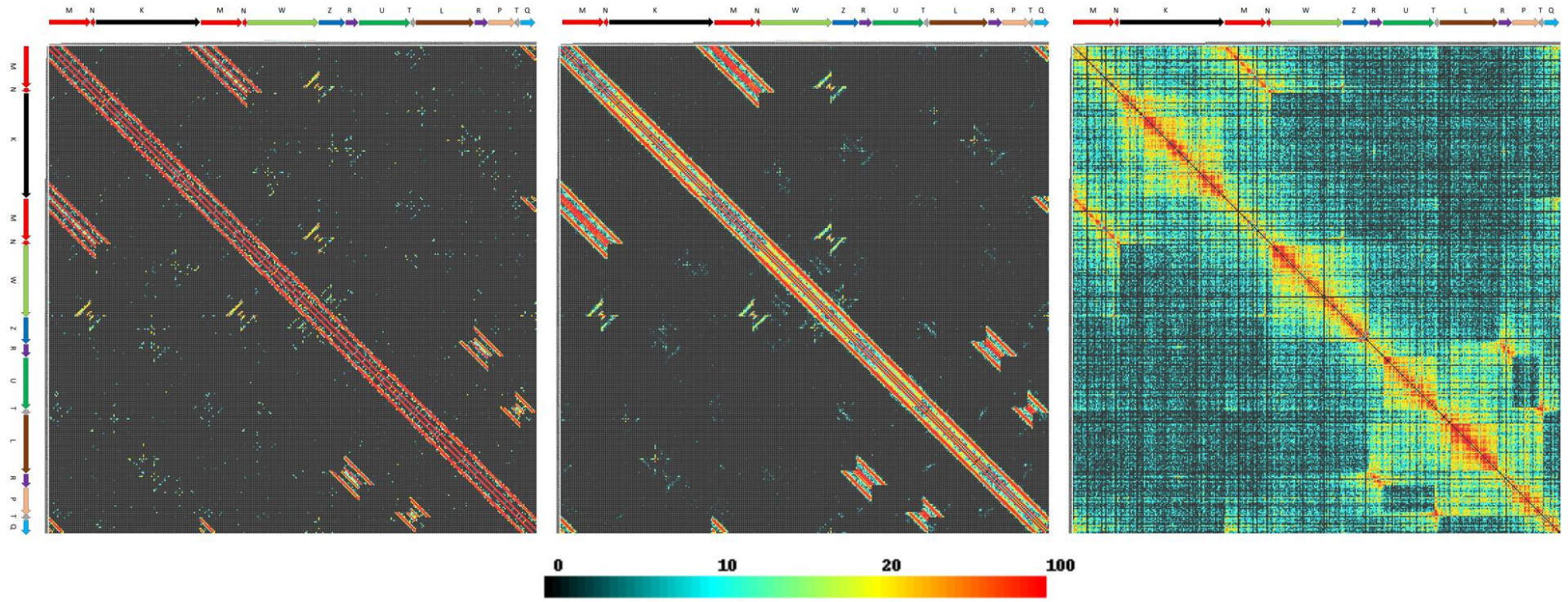

A


B

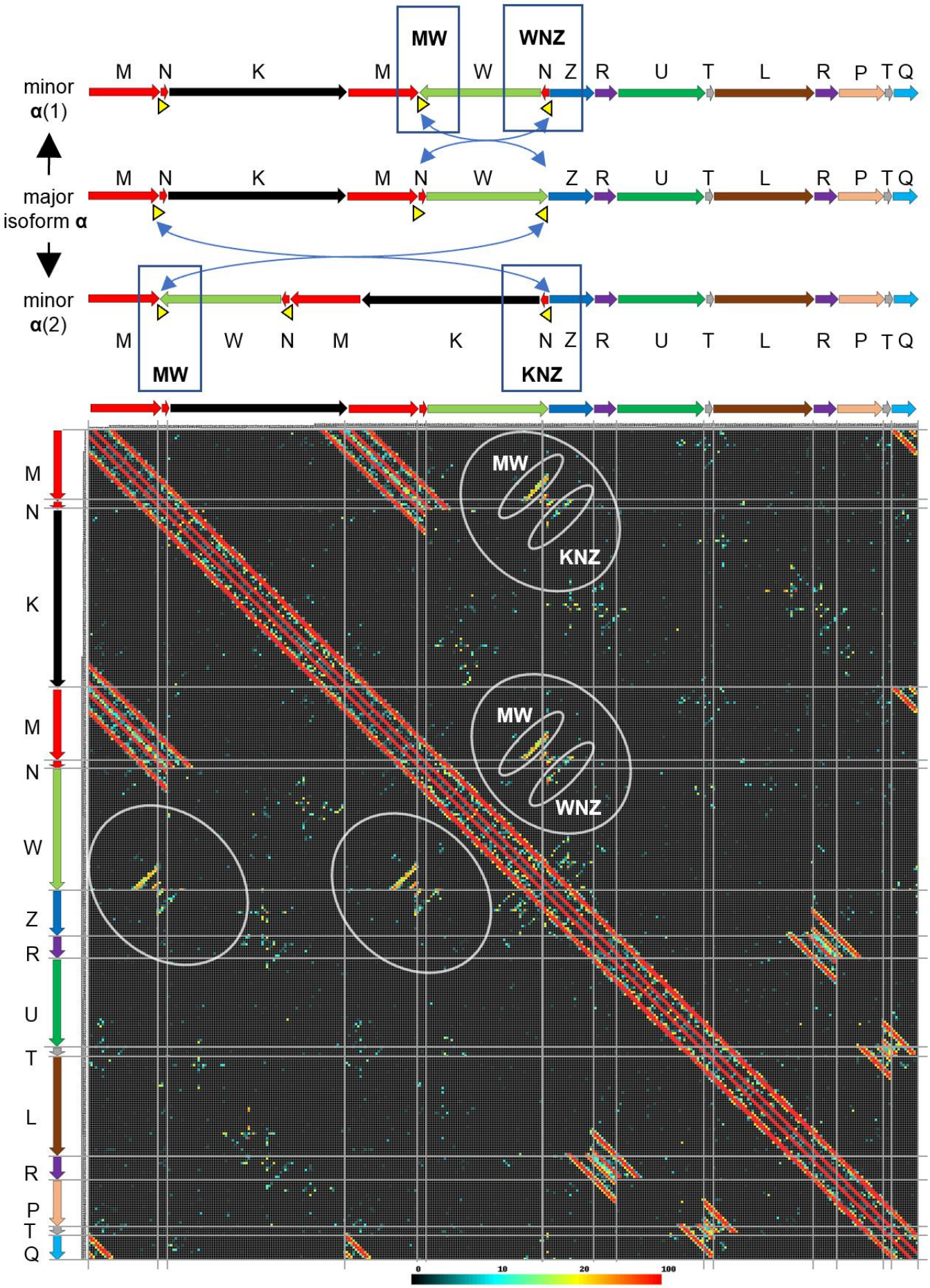
L. sativa Illumina mate-pair libraries
(2.5 + 10 kb)

L. serriola Illumina mate-pair libraries
(2.5 + 10 kb)

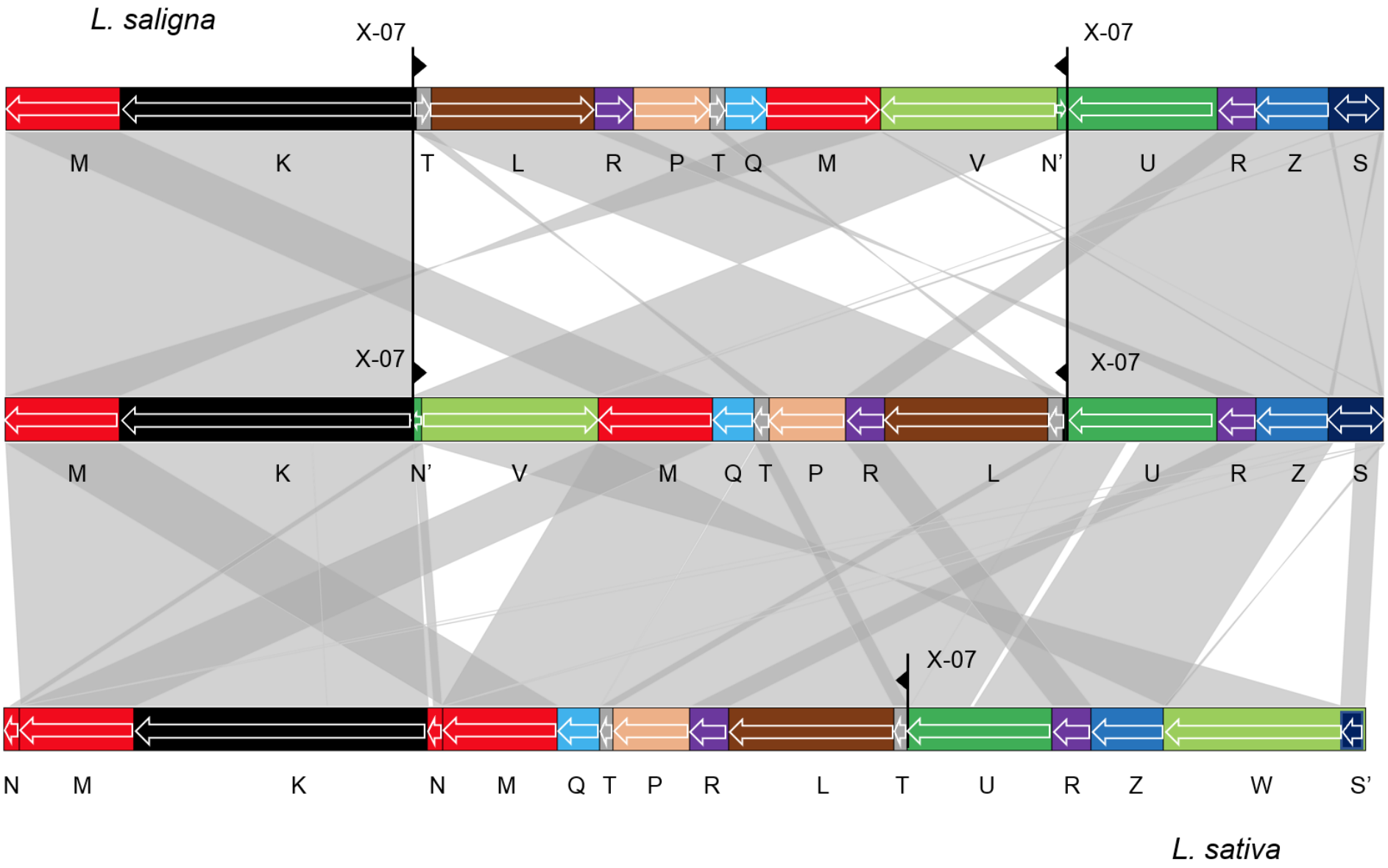
L. sativa Hi-C library



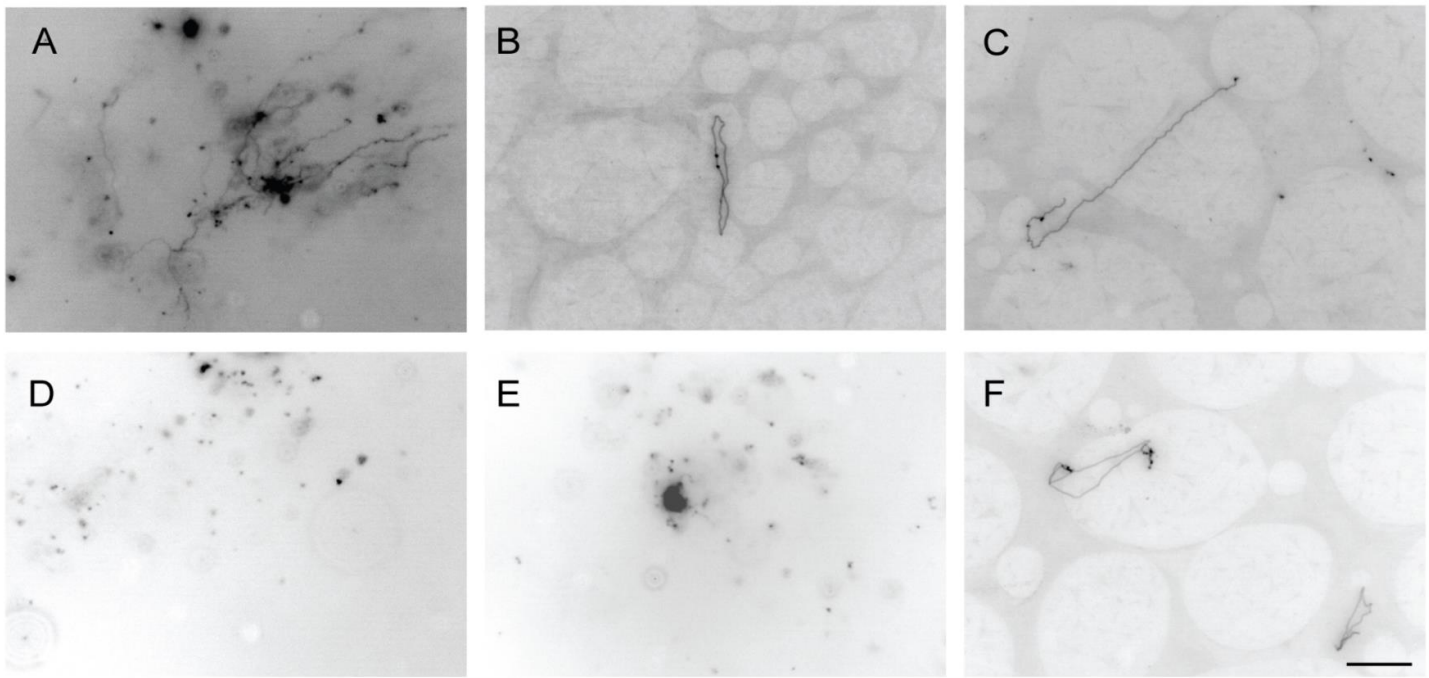
F4

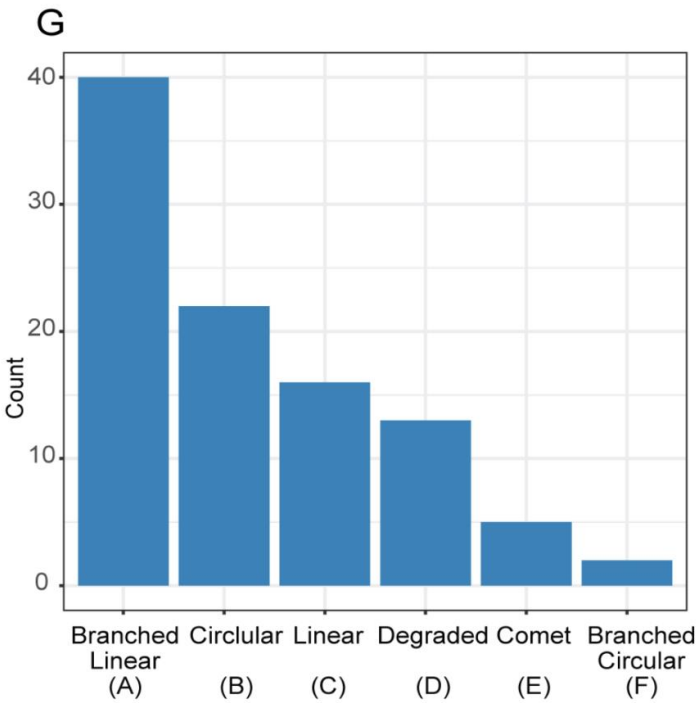

MW
WNZ
minor α(1)
major isoform α
minor α(2)
MW
KNZ
MW
KNZ
MW
WNZ
M N K M W N Z R U T L R P T Q
M N K M N W Z R U T L R P T Q
M W N M K N Z R U T L R P T Q
0 10 20 100

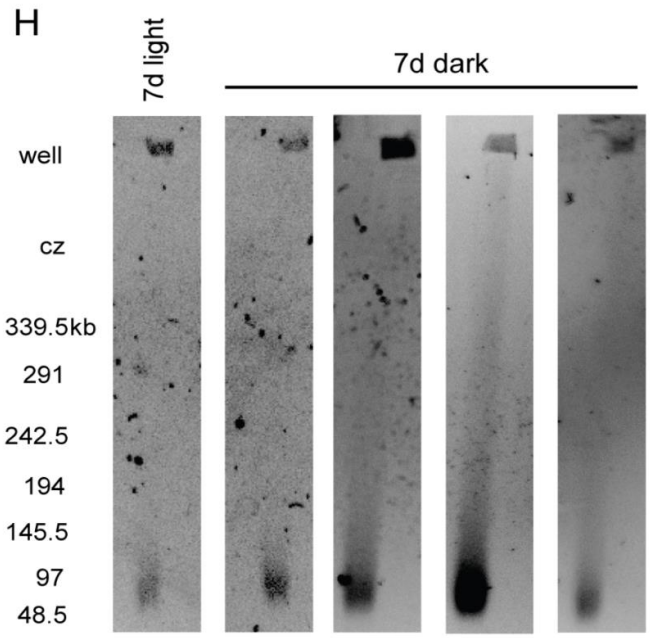
F5


L. saligna
X-07
X-07
M K T L R P T Q M V N' U R Z S
X-07
X-07
M K N' V M Q T P R L U R Z S
X-07
N M K N M Q T P R L T U R Z W S'
L. sativa

F6


A
B
C
D
E
F


G
40
30
20
10
0
Count
Branched Linear (A)
Circlular (B)
Linear (C)
Degraded (D)
Comet (E)
Branched Circular (F)


H
7d light
7d dark
well
cz
339.5kb
291
242.5
194
145.5
97
48.5

F7

bioRxiv preprint doi: https://doi.org/10.1101/564278; this version posted March 1, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table S1

***L. sativa* mitochondrial repeats less than 3kb**

Repeat_ID	Length	Sequence
X-01a	877	CGAAAAGACCAGCTAAGCATCATTGGCTTGGTCAGCCTTTACCTGACCAACTACCTAATACTACGCAG GCTCATCAAACAGCGCTTTTTAGCTTTCTTCAGGATTTGGCCCGAACTGTTCGGCAGATTCCCACGCG TTACGCACCCGTTCGCCACTTTGTTCTCAACTCTTCTCACCTCCTGGGCGAGACAAGCTACCTTTAGC TAGGAGCCTCTTTTCCTTCTGCCCAGCTCCCCGAAAACAACGTTCGACTTGCATGTGTTAAGCATATA GCTAGCGTTCCTTCTGAGCCAGGATCAAACTCTTCTTTTGACTATGATTTGGCCCTACAGTGGTAGAA CCTGGTGAACCGGGCGTACTACTTCTCAACCTTCTGTGAACTTTTCTTCTCTTATGATTTTTGCTACT TTGTTTAGTTTAGTGATTGATATCTTAGAGAAGCTGGTAAAGTAAAGACAGACTCTTTCGAAAGTAAT GGTGGCTAGGAATGGCTTCAGTCAGGTCATCGATTCACCTTATGGTCGACTTGCTTGAATCGAAGAAG AAGAAACCCCCGGACTAGAGCCTTCAGTGCTTCTCCCCCATGAATGTGATCAACAAGAATTGATGTAC CCACCATTCTAGATAGGAAAGAGATTGATGGTAAGAACCAGTACTAATCTTATAGCAAGCCAAAATAG AGTTTAGATAGGTAAGAACAAGTGCCTATTTCTTGCTAGCCAAAAAACAAGTTTGAGTTGTCGGGGTT GGATCTGAAGCATCAAGCTTCTCCTATACTTTCTCAAAGAAATGCCCTTATCTCGATCTTTATTAGAG GCTAGAGTCTAACCATTAAAGAAGGCTTGAACAGGCTTTTGAGTAAAGAAAAAAGTCACCC
X-01b	576	CGAAAAGACCAGCTAAGCATCATTGGCTTGGTCAGCCTTTACCTGACCAACTACCTAATACTACGCAG GCTCATCAAACAGCGCTTTTTAGCTTTCTTCAGGATTTGGCCCGAACTGTTCGGCAGATTCCCACGCG TTACGCACCCGTTCGCCACTTTGTTCTCAACTCTTCTCACCTCCTGGGCGAGACAAGCTACCTTTAGC TAGGAGCCTCTTTTCCTTCTGCCCAGCTCCCCGAAAACAACGTTCGACTTGCATGTGTTAAGCATATA GCTAGCGTTCCTTCTGAGCCAGGATCAAACTCTTCTTTTGACTATGATTTGGCCCTACAGTGGTAGAA CCTGGTGAACCGGGCGTACTACTTCTCAACCTTCTGTGAACTTTTCTTCTCTTATGATTTTTGCTACT TTGTTTAGTTTAGTGATTGATATCTTAGAGAAGCTGGTAAAGTAAAGACAGACTCTTTCGAAAGTAAT
X-02	215	TTGATTCAGTGATGTGACTTCTCCTGCTCCAGTAAATCAGTCTGGAGACTTGCTAATTCAATTCAATT GCCTTGGTTTTTCCAAGAAATACATGGGAAAAATAGGGAGAAAGAAATGGGATTGTGTCAACAGGCCC CTATTCCGACGAAAAACAGCCCCTTTTCTTTTTGTTTAGCCTGTTCCTATCAGATAAATTCTTAAGAG AAGAGGAAAAA
X-03	186	CCTCACTCATCCCATCTCCCCTTCCTCCAACTGCGCTTAGACCGACAGTGGCAAGAGCTTCTTCTTTG TTCACAGCTTGAGCGGATTCGATTCCGAACCTTTTTTTTGAGTAACTATAGTGATTCCTTCCCTTTCT CTTCCTTCTTTCCCAGAGCTACTGGCTCATTTCACTCTCTTTAAATAAGA
X-04	137	AACGAAGATCTTCGAGAACTCATATTGGACCGGGAATCAAAAAGGGACAAAAGTAAAGTATAAGCGCA TATGGCACGAAAAGGAAATCCAATTTCGGTAAGACTTGGTCTGAATCGTAGTTCAGATTCAAGTCGGT T
X-05	131	AAAGAGAATAGAAAGCGTACTGACTCTGACTGCTATCTACGATGCGATCAGGGGGGAATAGCGCAAGG GCTTAAGTCATCGATTCAAATCCTATCTTTTTTTCGGTATGCCGCTCCGCGAGCAAGGAGCGA
X-06	123	CCTGGGATTGTAGTTCAATCGGTCAGAGCACCGCCCTGTCAAGGCGGAAGTTGCGGGTTCGAGCCCCG TCAGTCCCGACGCCGGATTCAAAAAAGACATCAATTCCTCTCTCCATTTTCTGTG
X-07u	122	CCTTCCATCCTCTCGCAAAGCTCGAGAACTACGTACCTCGATCTTCTGATACAGCAAGGGTCGCCTCC CTCCCCCTTACCACGCTTTTCCACCCTCTGGCCATTCAACCCCCTCTCCCCGGG
X-08	119	GCGAGAACTACTAGAAAATGGTGAGATCATTAGATCATTAGTGAAAGAAGGACCAACGACGATCCTAT CCTAAAGGAGAAGAAGGAGTAGGAGGAGTCAGTTTTCTTTGAAGATCGAGG
X-09	117	CGCCTCTCTAGGCTTCGCTATCGCTCATGACTGGTATATGGATCTCTCTATGTAGTGGTCGGCCTTCT AGAAGCTTCGCCAGAAGCGACTAGTCGCTTCCCGAATGCCCCTTTCCTT
X-10	97	ATTAACGTCGGCTTATCTGTCAGTCGTGTTGGGTCTGCCGCTCAGTTGAAAACTATGAAACAAGTCTG CGGTAGTTCAAAACTGGAATTGGCACAAT
X-11	88	TGGAAATGCCATAAAGCGCGAACCAAGATCCGTGAGACGAAAACCAAAATAAGAATGAGGAAGAAAAA GATATCGTGATGTAAGTCTA
X-12	68	AAAAAAAAAAAAAAGAATCTAACGCTCTAGTCTAATTGAAGACTTTCAACGCTCATGCTATTTGAAAG
X-13	65	TTTTGGTTTTCGTCTCACGGATCTTGGTTCGCGCTTTATGGCATTTCCACTATAATAAGTGTAGG
X-14	64	GAATGAATGGTAGTTCGCTGGGCCTGTCTTTTGCTCCCAGAGGTGCTGGTTCGAATCCAGTTCG
X-15	62	ATACGTAGATGACTTTTGGCTGGTACAACCACTCGATTGTCCACTTCTAATAAACGTAATTG

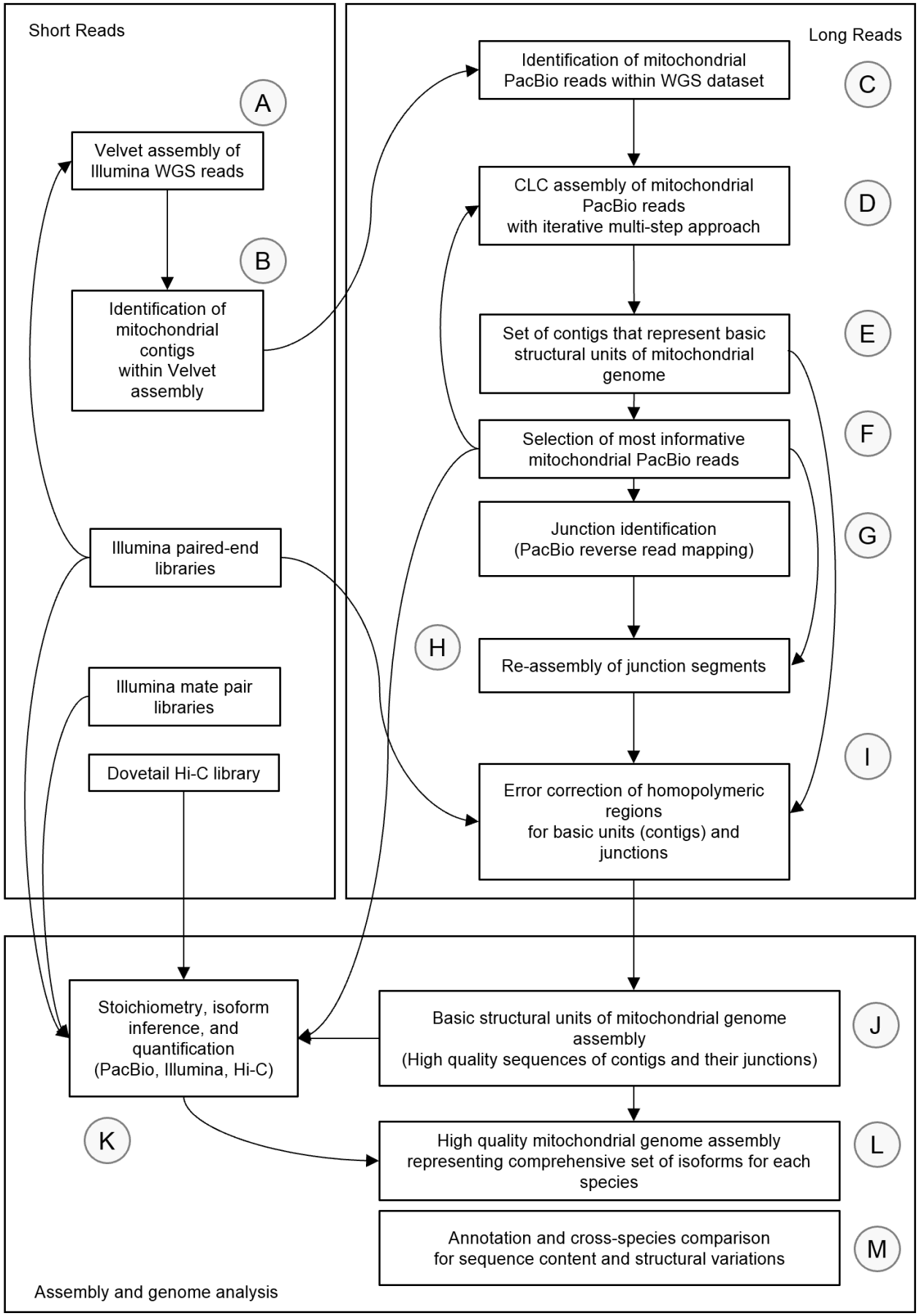
bioRxiv preprint doi: https://doi.org/10.1101/564278; this version posted March 1, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

***L. saligna* mitochondrial repeats less than 3kb**

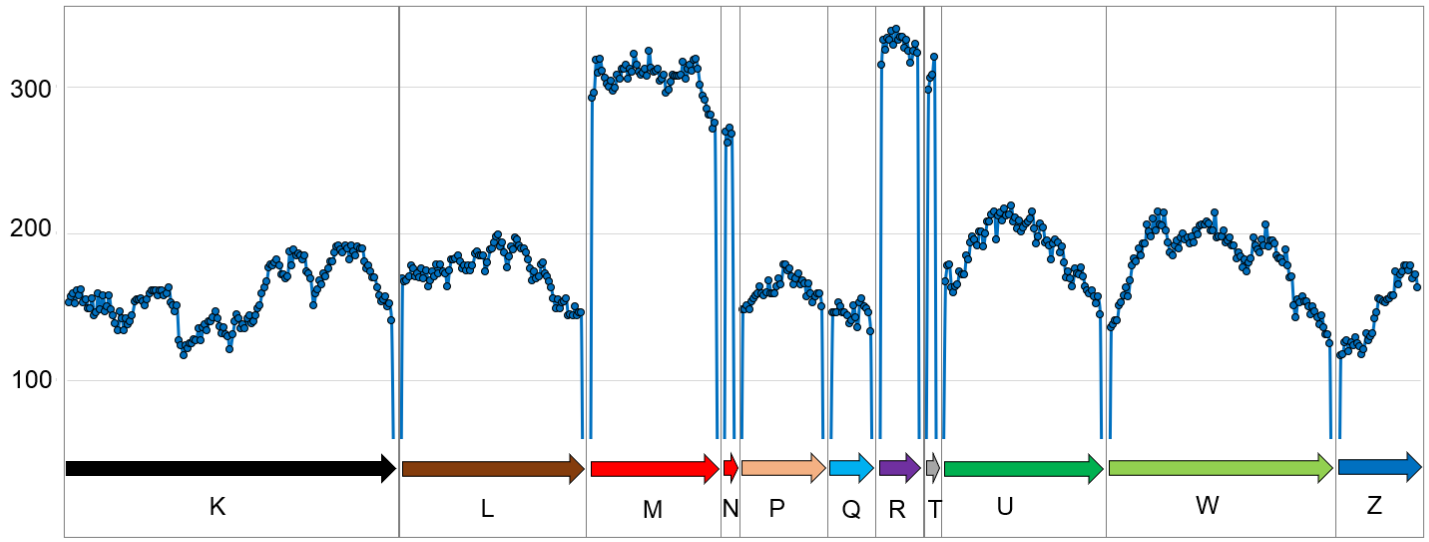
Repeat_ID	Length	Sequence
S-01-IR	1218	CGAAAAGACCAGCTAAGCATCATTGGCTTGGTCAGCCTTTACCTGACCAACTACCTAATACTACGCAG GCTCATCAAACAGCGCTTTTTAGCTTTCTTCAGGATTTGGCCCGAACTGTTCGGCAGATTCCCACGCG TTACGCACCCGTTCGCCACTTTGTTCTCAACTCTTCTCACCTCCTGGGCGAGACAAGCTACCTTTAGC TAGGAGCCTCTTTTCCTTCTGCCCAGCTCCCCGAAAACAACGTTCGACTTGCATGTGTTAAGCATATA GCTAGCGTTCCTTCTGAGCCAGGATCAAACTCTTCTTTTGACTATGATTTGGCCCTACAGTGGTAGAA CCTGGTGAACCGGGCGTACTACTTCTCAACCTTCTGTGAACTTTTCTTCTCTTATGATTTTTGCTACT TTGTTTAGTTTAGTGATTGATATCTTAGAGAAGCTGGTAAAGTAAAGACAGACTCTTTCGAAAGTAAT GGTGGCTAGGAATGGCTTCAGTCAGGTCATCGATTCACCTTATGGTCGACTTGCTTGAATCGAAGAAG AAGAAACCCCCGGACTAGAGCCTTCAGTGCTTCTCCCCCATGAATGTGATCAACAAGAATTGATGTAC CCACCATTCTAGATAGGAAAGAGATTGATGGTAAGAACCAGTACTAATCTTATAGCAAGCCAAAATAG AGTTTAGATAGGTAAGAACAAGTGCCTATTTCTTGCTAGCCAAAAAACAAGTTTGAGTTGTCGGGGTT GGATCTGAAGCATCAAGCTTCTCCTATACTTTCTCAAAGAAATGCCCTTATCTCGATCTTTATTAGAG GCTAGAGTCTAACCATTAAAGAAGGCTTGAACAGGCTTTTGAGTAAAGAAAAAAGTCACCCCCCCTCC CCCAGGGTCTTTAGAGTAATACCGCTGTCTGTCCCCTGGAGGTTATCCTAAACCATTTTCTAGGGGAG AAATACTGGCTCAGGGTAAAGGCAGCCCCCCGTCGGGTAAAATCCCATAATGCCTGTCCCAAGGCCGA ATAAACATGGCTCTGTGCCTTGTTCCTCTCATGGTCTGGGTTGAGTCTAGTATTTCTGTGGGTAAAGG CACCTCATATACTCTTATGTCCTACGGAATCGAAAATCTATCCTACGGAATCGAGATTTTCCCTTAAG
X-01a	877	CGAAAAGACCAGCTAAGCATCATTGGCTTGGTCAGCCTTTACCTGACCAACTACCTAATACTACGCAG GCTCATCAAACAGCGCTTTTTAGCTTTCTTCAGGATTTGGCCCGAACTGTTCGGCAGATTCCCACGCG TTACGCACCCGTTCGCCACTTTGTTCTCAACTCTTCTCACCTCCTGGGCGAGACAAGCTACCTTTAGC TAGGAGCCTCTTTTCCTTCTGCCCAGCTCCCCGAAAACAACGTTCGACTTGCATGTGTTAAGCATATA GCTAGCGTTCCTTCTGAGCCAGGATCAAACTCTTCTTTTGACTATGATTTGGCCCTACAGTGGTAGAA CCTGGTGAACCGGGCGTACTACTTCTCAACCTTCTGTGAACTTTTCTTCTCTTATGATTTTTGCTACT TTGTTTAGTTTAGTGATTGATATCTTAGAGAAGCTGGTAAAGTAAAGACAGACTCTTTCGAAAGTAAT GGTGGCTAGGAATGGCTTCAGTCAGGTCATCGATTCACCTTATGGTCGACTTGCTTGAATCGAAGAAG AAGAAACCCCCGGACTAGAGCCTTCAGTGCTTCTCCCCCATGAATGTGATCAACAAGAATTGATGTAC CCACCATTCTAGATAGGAAAGAGATTGATGGTAAGAACCAGTACTAATCTTATAGCAAGCCAAAATAG AGTTTAGATAGGTAAGAACAAGTGCCTATTTCTTGCTAGCCAAAAAACAAGTTTGAGTTGTCGGGGTT GGATCTGAAGCATCAAGCTTCTCCTATACTTTCTCAAAGAAATGCCCTTATCTCGATCTTTATTAGAG GCTAGAGTCTAACCATTAAAGAAGGCTTGAACAGGCTTTTGAGTAAAGAAAAAAGTCACCC
X-01b	576	CGAAAAGACCAGCTAAGCATCATTGGCTTGGTCAGCCTTTACCTGACCAACTACCTAATACTACGCAG GCTCATCAAACAGCGCTTTTTAGCTTTCTTCAGGATTTGGCCCGAACTGTTCGGCAGATTCCCACGCG TTACGCACCCGTTCGCCACTTTGTTCTCAACTCTTCTCACCTCCTGGGCGAGACAAGCTACCTTTAGC TAGGAGCCTCTTTTCCTTCTGCCCAGCTCCCCGAAAACAACGTTCGACTTGCATGTGTTAAGCATATA GCTAGCGTTCCTTCTGAGCCAGGATCAAACTCTTCTTTTGACTATGATTTGGCCCTACAGTGGTAGAA CCTGGTGAACCGGGCGTACTACTTCTCAACCTTCTGTGAACTTTTCTTCTCTTATGATTTTTGCTACT TTGTTTAGTTTAGTGATTGATATCTTAGAGAAGCTGGTAAAGTAAAGACAGACTCTTTCGAAAGTAAT GGTGGCTAGGAATGGCTTCAGTCAGGTCATCGATTCACCTTATGGTCGACTTGCTTGAATCGAAGAAG AAGAAACCCCCGGACTAGAGCCTTCAGTGCTT
X-02	215	TTGATTCAGTGATGTGACTTCTCCTGCTCCAGTAAATCAGTCTGGAGACTTGCTAATTCAATTCAATT GCCTTGGTTTTTCCAAGAAATACATGGGAAAAATAGGGAGAAAGAAATGGGATTGTGTCAACAGGCCC CTATTCCGACGAAAAACAGCCCCTTTTCTTTTTGTTTAGCCTGTTCCTATCAGATAAATTCTTAAGAG AAGAGGAAAAA
X-03	186	CCTCACTCATCCCATCTCCCCTTCCTCCAACTGCGCTTAGACCGACAGTGGCAAGAGCTTCTTCTTTG TTCACAGCTTGAGCGGATTCGATTCCGAACCTTTTTTTTGAGTAACTATAGTGATTCCTTCCCTTTCT CTTCCTTCTTTCCCAGAGCTACTGGCTCATTTCACTCTCTTTAAATAAGA
X-04q	137	AACGAAGATCTTCGAGAACTCATATTGGACCGGGGATCAAAAAGGGACAAAAGTAAAGTATAAGCGCA TATGGCACGAAAAGGAAATCCAATTTCGGTAAGACTTGGTCTGAATCGTAGTTCAGATTCAAGTCGGT T
X-04z	137	AACGAAGATCTTCGAGAACTCATATTGGACCGGGAATAAAAAAGGGACAAAAGTAAAGTATAAGCGCA TATGGCACGAAAAGGAAATCCAATTTCGGTAAGACTTGGTCTGAATCGTAGTTCAGATTCAAGTCGGT T

bioRxiv preprint doi: https://doi.org/10.1101/564278; this version posted March 1, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

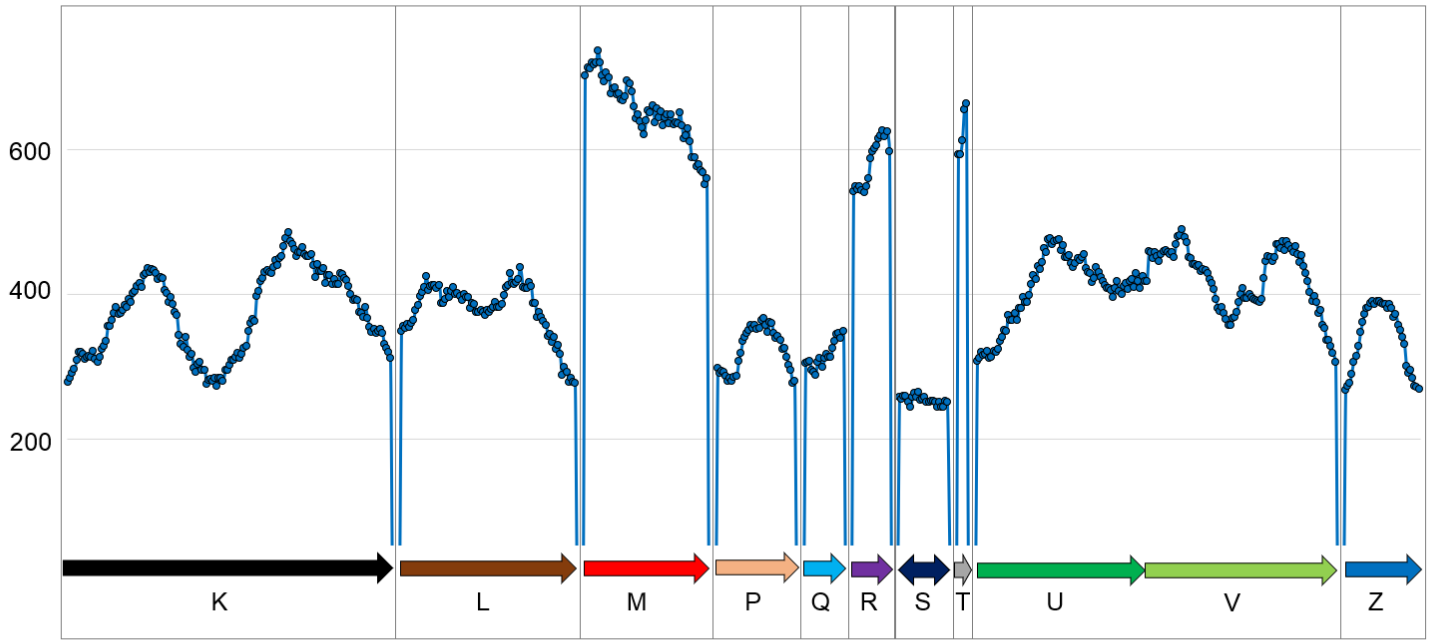
X-05	131	AAAGAGAATAGAAAGCGTACTGACTCTGACTGCTATCTACGATGCGATCAGGGGGGAATAGCGCAAGG GCTTAAGTCATCGATTCAAATCCTATCTTTTTTTCGGTATGCCGCTCCGCGAGCAAGGAGCGA
X-06	123	CCTGGGATTGTAGTTCAATCGGTCAGAGCACCGCCCTGTCAAGGCGGAAGTTGCGGGTTCGAGCCCCG TCAGTCCCGACGCCGGATTCAAAAAAGACATCAATTCCTCTCTCCATTTTCTGTG
X-07	122	CCTTCCATCCTCTCGCAAAGCTCGAGAACTACGTACCTCGATCTTCTGATACAGCAAGGGTCGCCTCC CTCCCCCTTACCACACTTTTCCACCCTCTGGCCATTCAACCCCCTCTCCTCGGG
X-08	119	GCGAGAACTACTAGAAAATGGTGAGATCATTAGATCATTAGTGAAAGAAGGACCAACGACGATCCTAT CCTAAAGGAGAAGAAGGAGTAGGAGGAGTCAGTTTTCTTTGAAGATCGAGG
X-09	117	CGCCTCTCTAGGCTTCGCTATCGCTCATGACTGGTATATGGATCTCTCTATGTAGTGGTCGGCCTTCT AGAAGCTTCGCCAGAAGCGACTAGTCGCTTCCCGAATGCCCCTTTCCTT
X-10	97	ATTAACGTCGGCTTATCTGTCAGTCGTGTTGGGTCTGCCGCTCAGTTGAAAACTATGAAACAAGTCTG CGGTAGTTCAAAACTGGAATTGGCACAAT
X-11	88	TGGAAATGCCATAAAGCGCGAACCAAGATCCGTGAGACGAAAACCAAAATAAGAATGAGGAAGAAAAA GATATCGTGATGTAAGTCTA
X-12	68	AAAAAAAAAAAAAAGAATCTAACGCTCTAGTCTAATTGAAGACTTTCAACGCTCATGCTATTTGAAAG
X-13	65	TTTTGGTTTTCGTCTCACGGATCTTGGTTCGCGCTTTATGGCATTTCCACTATAATAAGTGTAGG
X-14	64	GAATGAATGGTAGTTCGCTGGGCCTGTCTTTTGCTCCCAGAGGTGCTGGTTCGAATCCAGTTCG
X-15	62	ATACGTAGATGACTTTTGGCTGGTACAACCACTCGATTGTCCACTTCTAATAAACGTAATTG


Short Reads
Long Reads
A
Velvet assembly of Illumina WGS reads
B
Identification of mitochondrial contigs within Velvet assembly
Illumina paired-end libraries
Illumina mate pair libraries
Dovetail Hi-C library
C
Identification of mitochondrial PacBio reads within WGS dataset
D
CLC assembly of mitochondrial PacBio reads with iterative multi-step approach
E
Set of contigs that represent basic structural units of mitochondrial genome
F
Selection of most informative mitochondrial PacBio reads
G
Junction identification (PacBio reverse read mapping)
H
Re-assembly of junction segments
I
Error correction of homopolymeric regions for basic units (contigs) and junctions
Stoichiometry, isoform inference, and quantification (PacBio, Illumina, Hi-C)
J
Basic structural units of mitochondrial genome assembly (High quality sequences of contigs and their junctions)
K
L
High quality mitochondrial genome assembly representing comprehensive set of isoforms for each species
M
Annotation and cross-species comparison for sequence content and structural variations
Assembly and genome analysis

L. sativa mitochondrial genome units coverage by PacBio reads

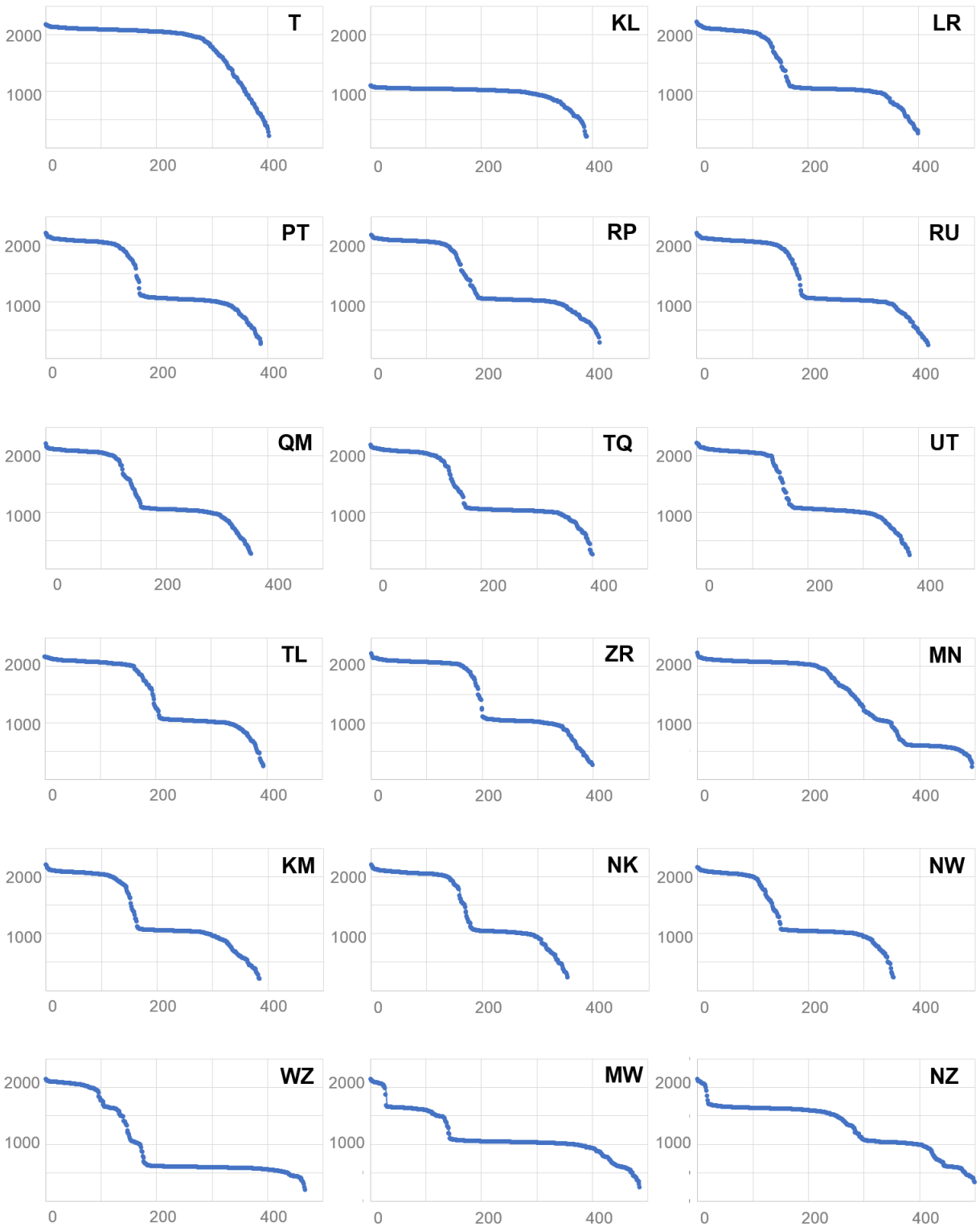


L. saligna mitochondrial genome units coverage by PacBio reads



X axis - 2000 bp long tiling overlapping segments across mitochondrial genome units
Y axis - number of PacBio read alignments to the tiling queries

Quantification of *L.sativa* mitochondrial genome junctions

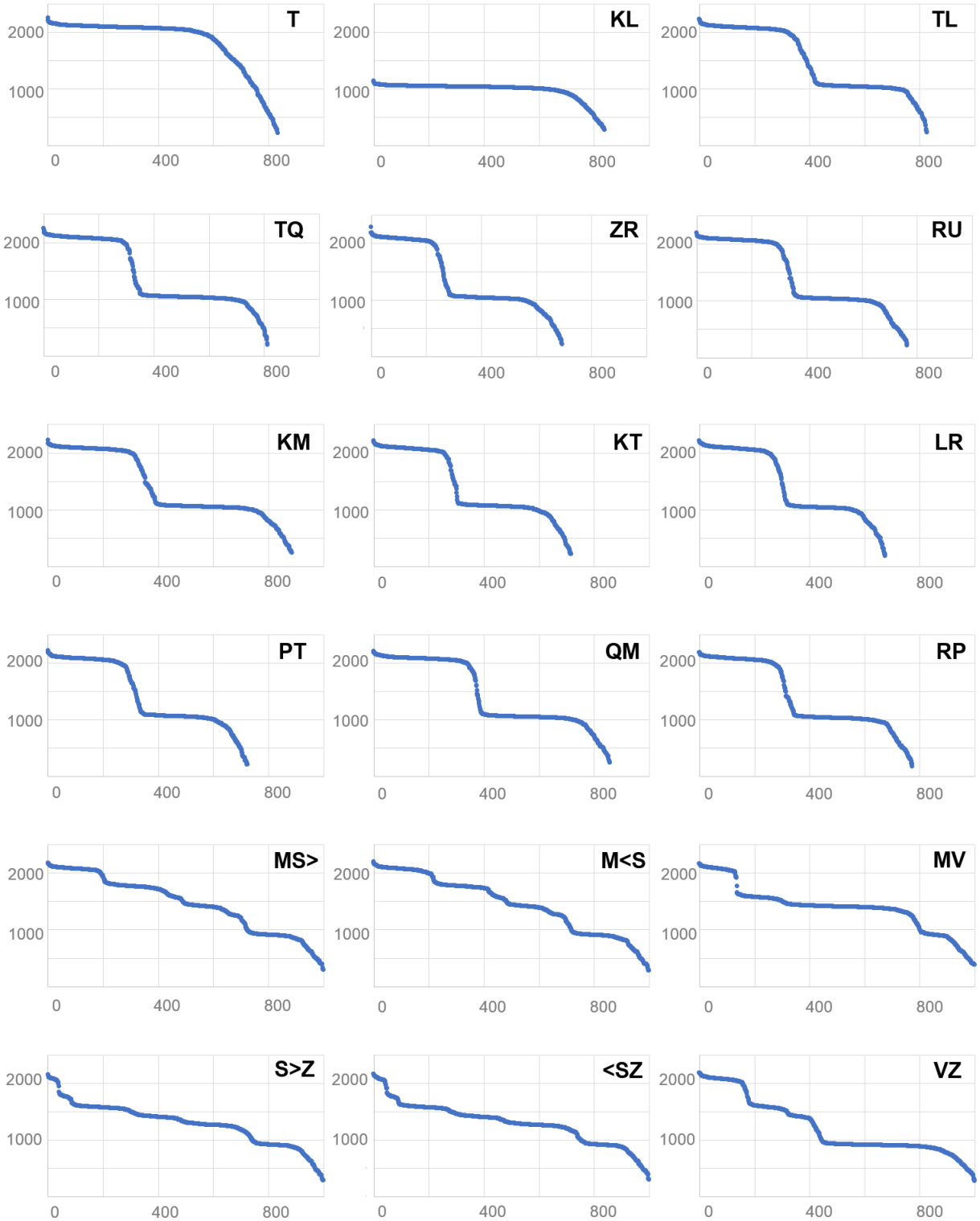


X axis - PacBio reads aligned to 2000 bp junction query and sorted by alignment length
Y axis - alignment length to 2000 bp junction query

Figure S3 A

S3A

Quantification of *L.saligna* mitochondrial genome junctions



X axis - PacBio reads aligned to 2000 bp junction query and sorted by alignment length
Y axis - alignment length to 2000 bp junction query

Figure S3 B

S3B

L.sativa Illumina mate-pair libraries (2.5 + 10 kb)

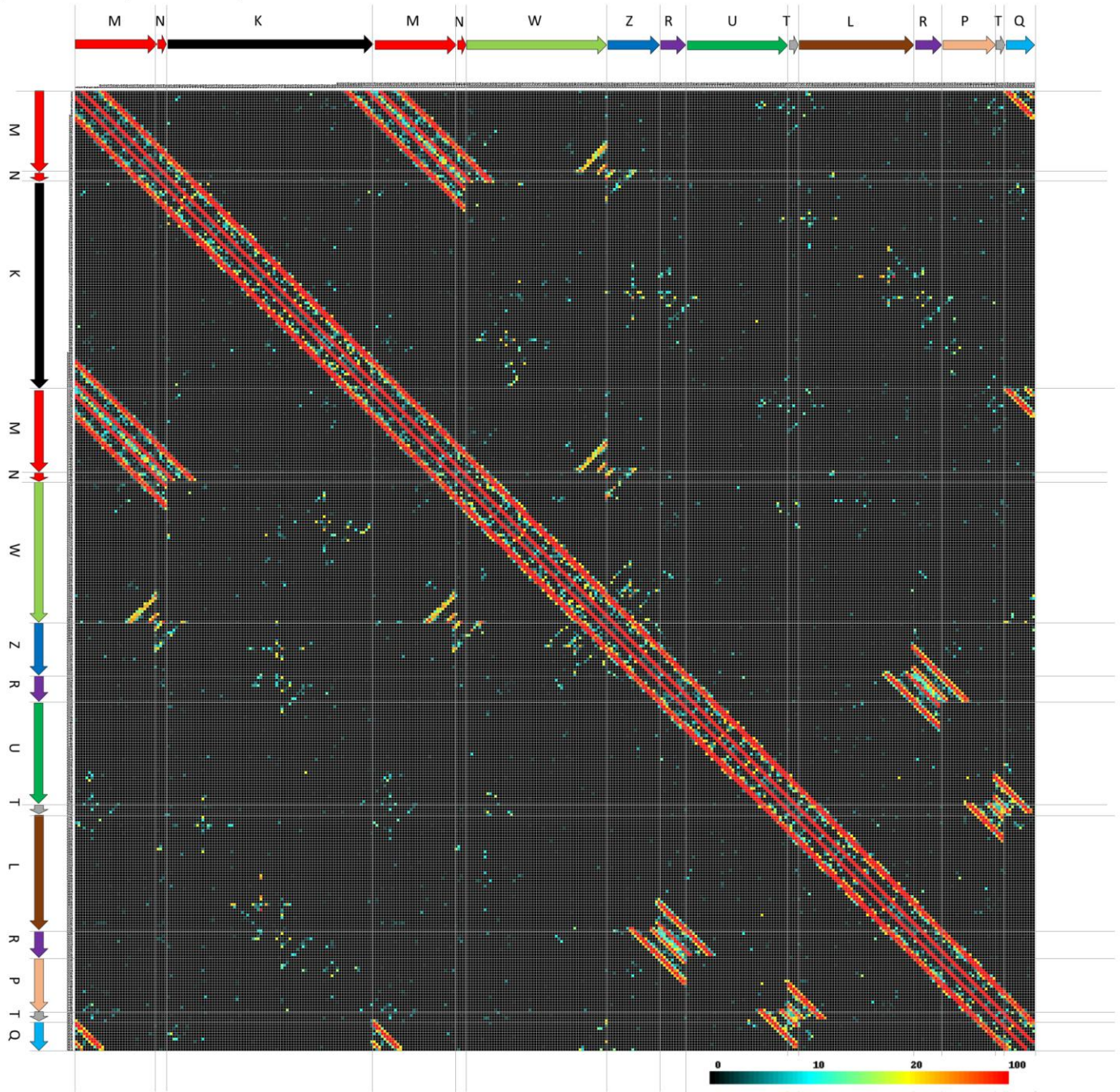


Figure S4 A

S4A

L.serriola Illumina mate-pair libraries (2.5 + 10 kb)

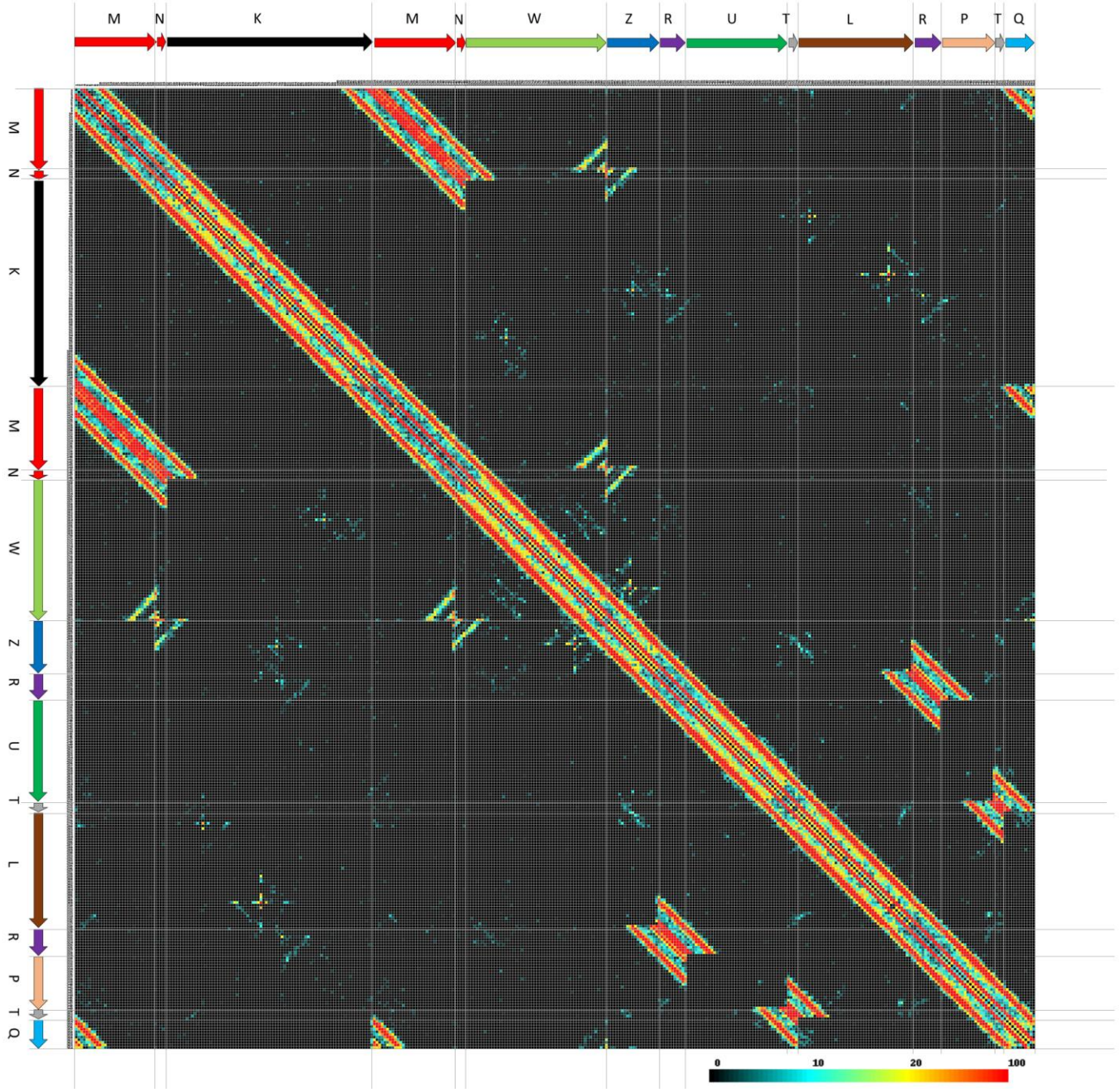


Figure S4 B

S4B

L.sativa Hi-C library

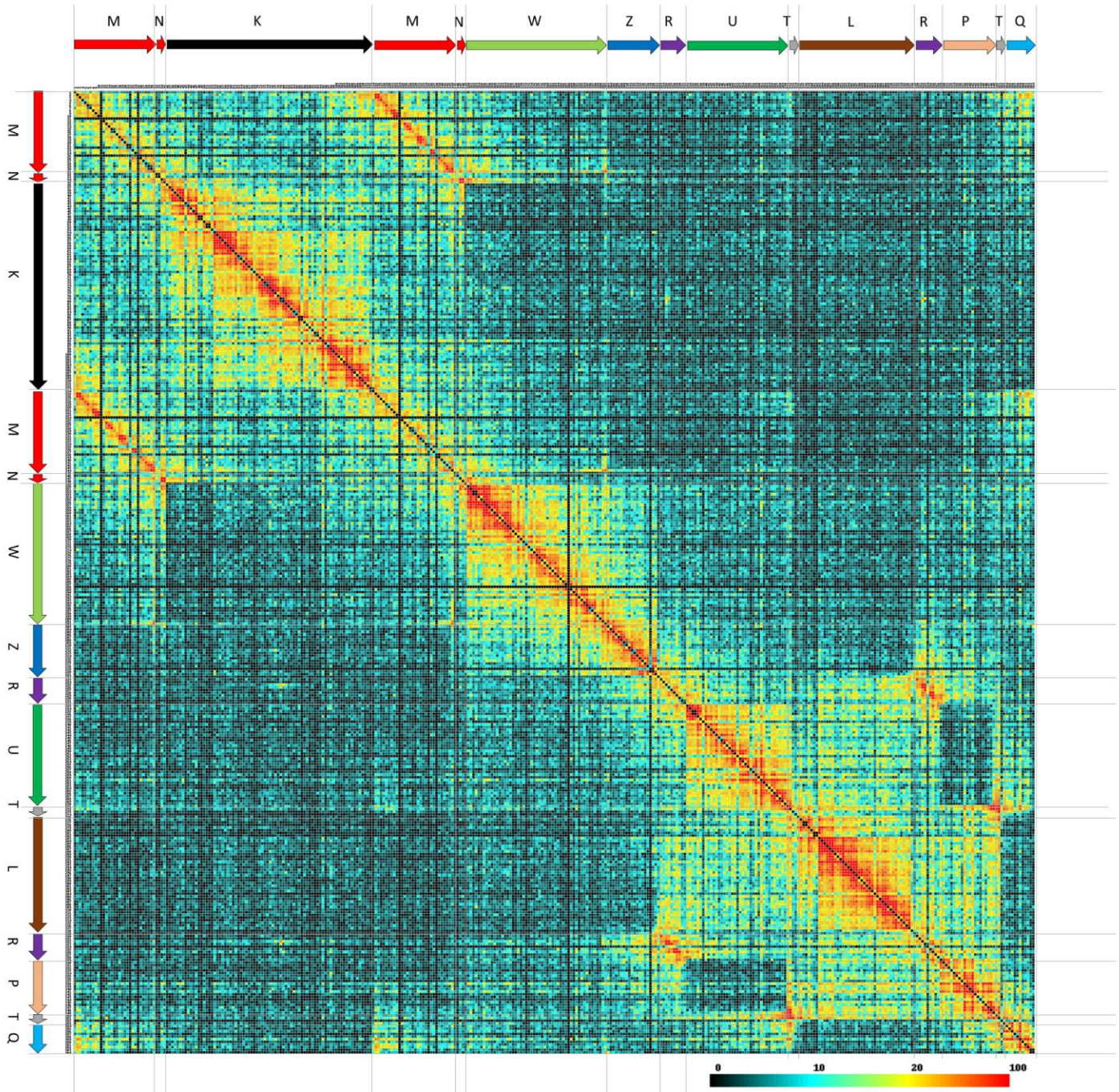


Figure S4 C

S4C

L.sativa 5 kb mate-pair library

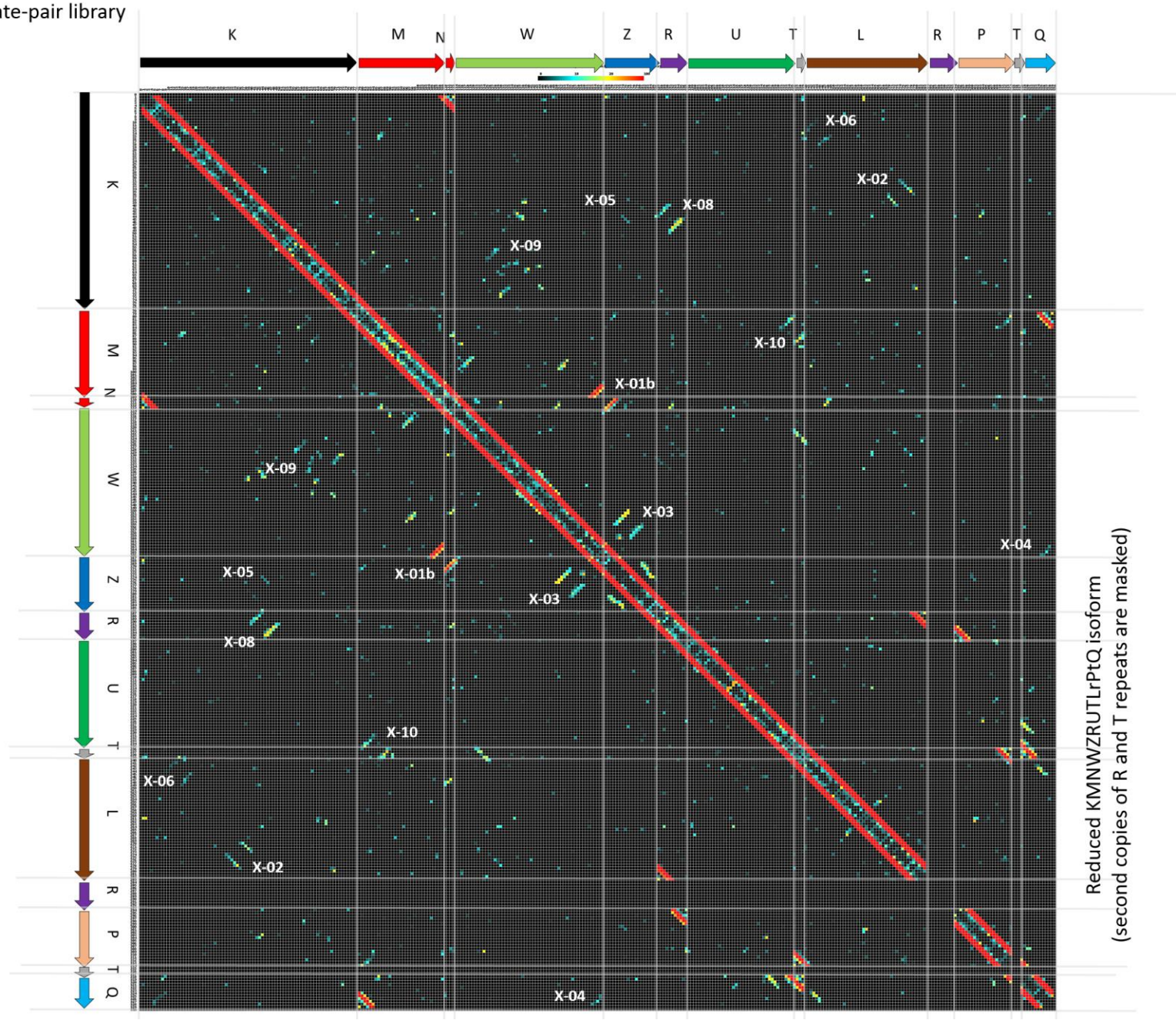


Figure S5 A

S5A

L.sativa 10 kb mate-pair library

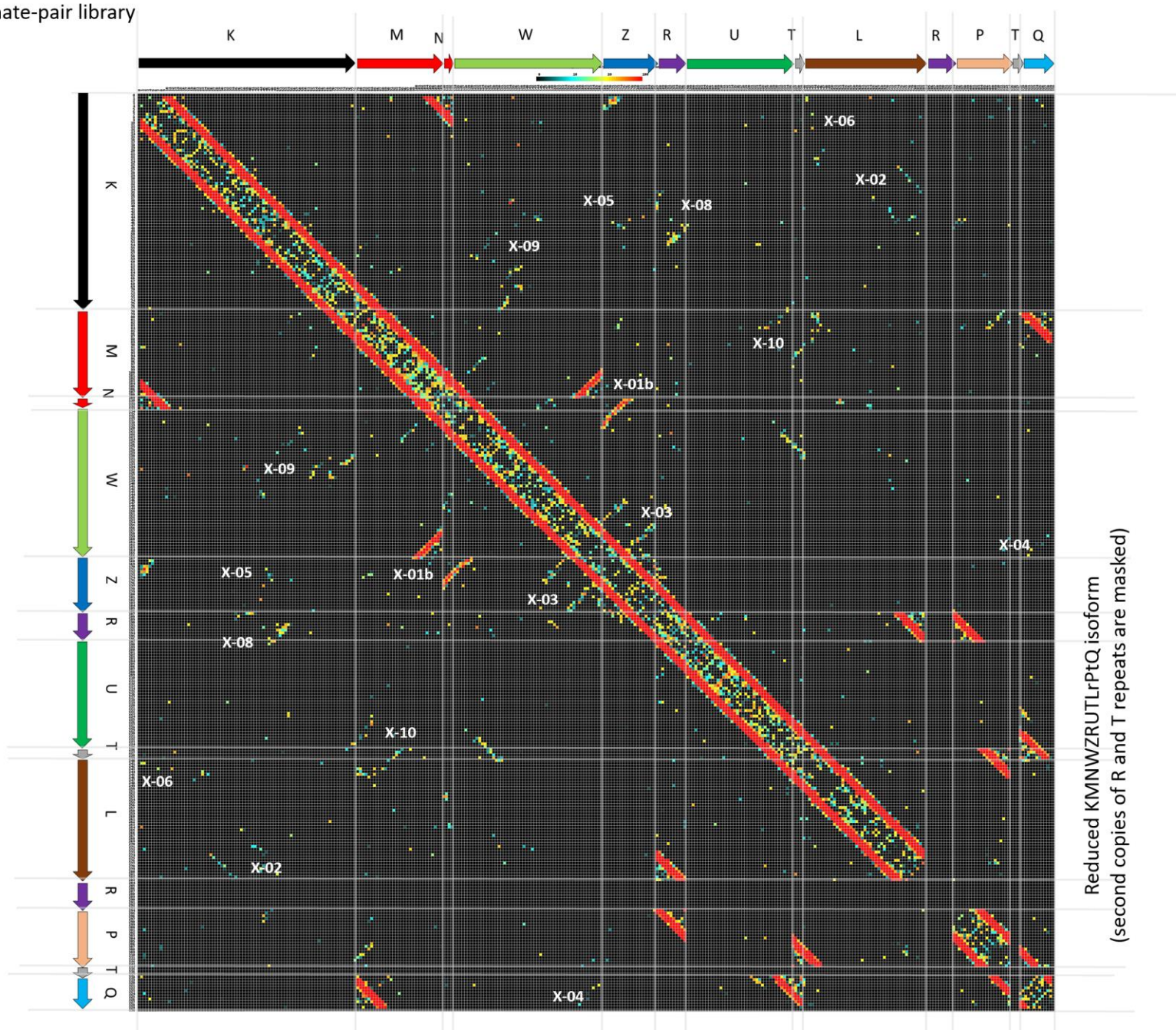


Figure S5 B

S5B

L.serriola 10 kb mate-pair library

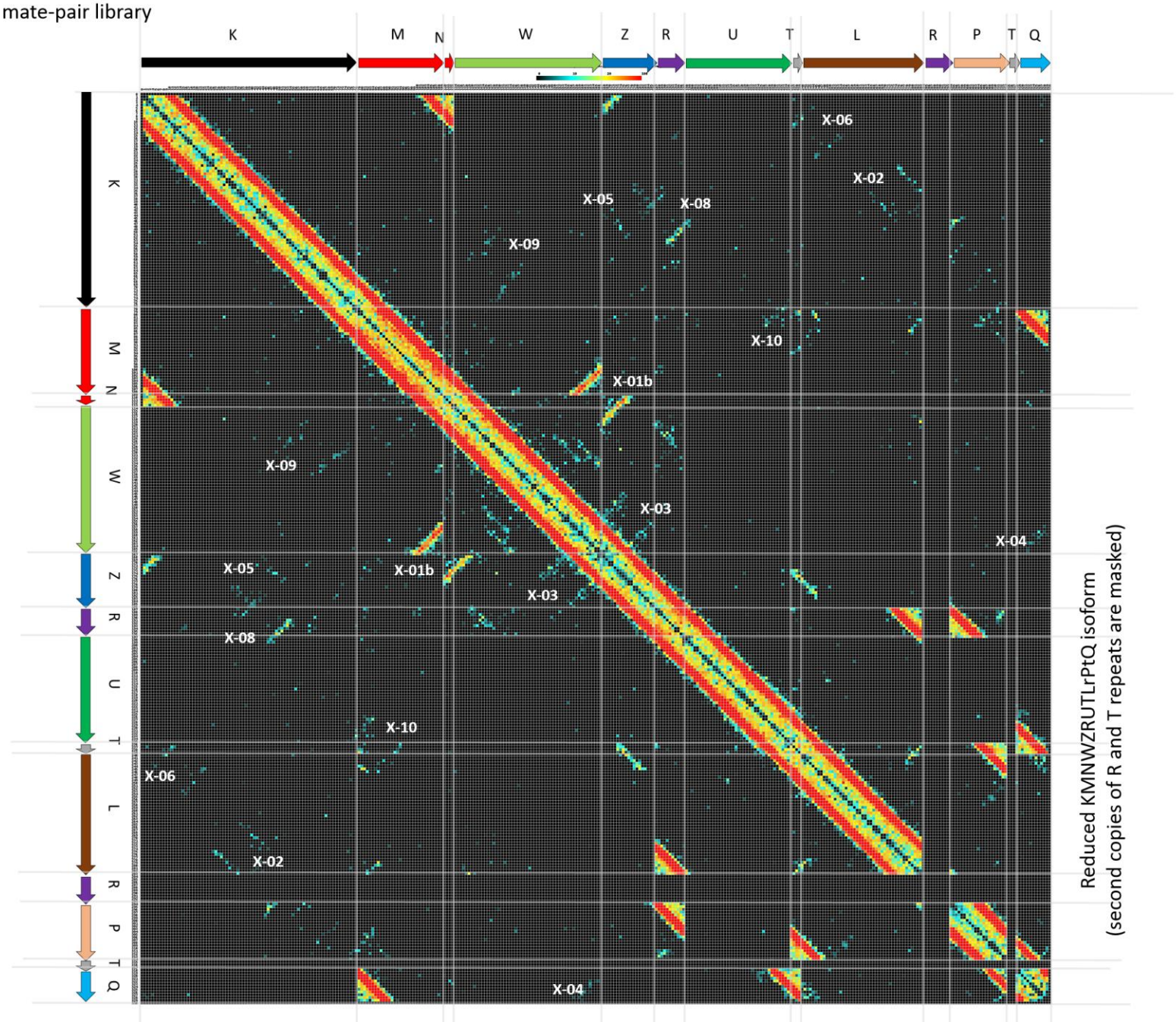
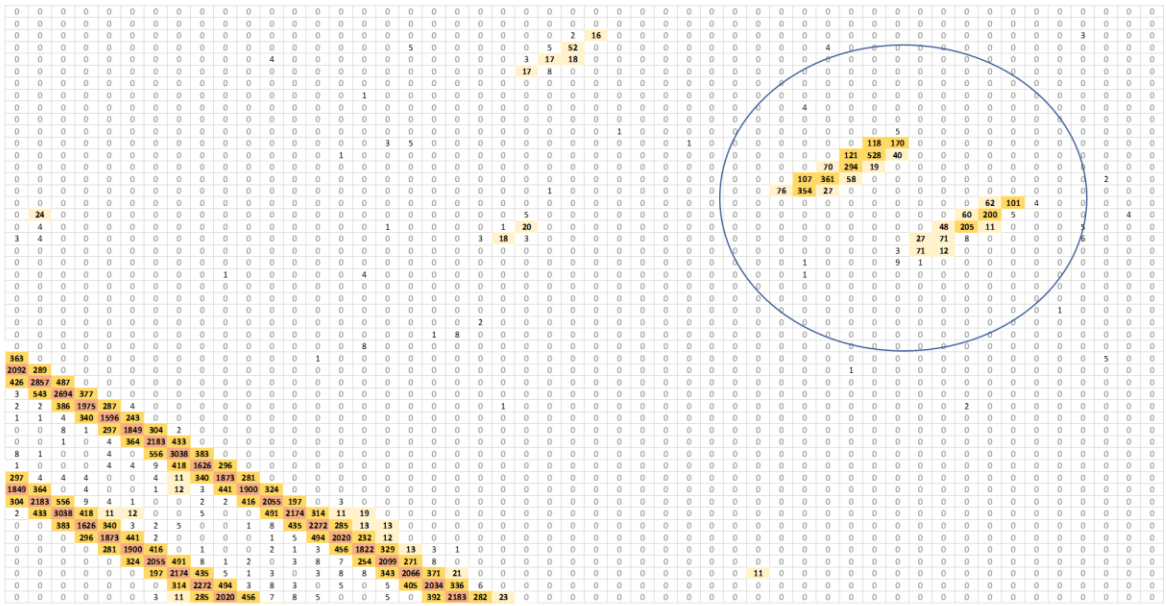


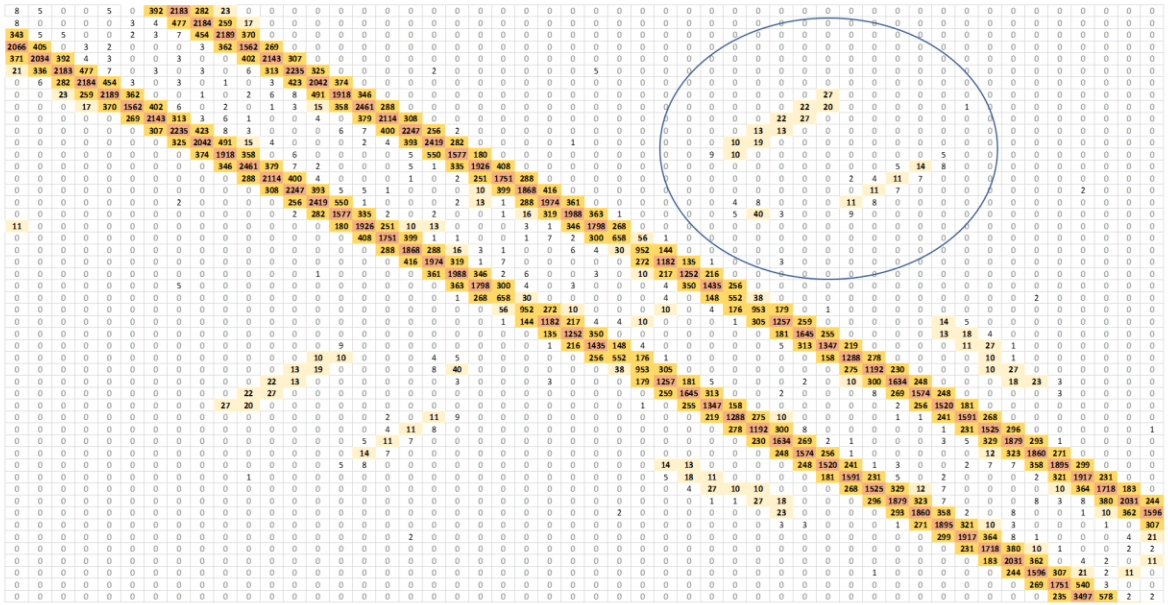
Figure S5 C

S5C

Example of numerical values for detection of recombination between short repeats X-01b and X-03 using the 5 kb mate-pair *L. sativa* library



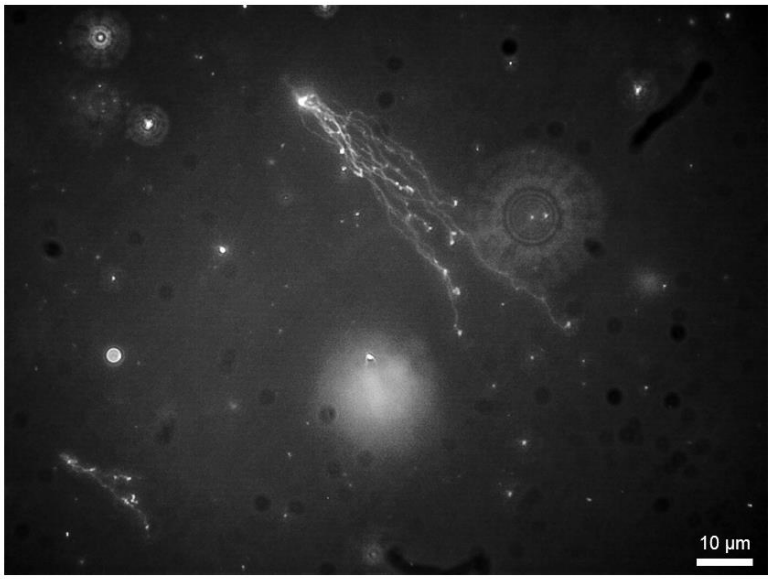
X-01b

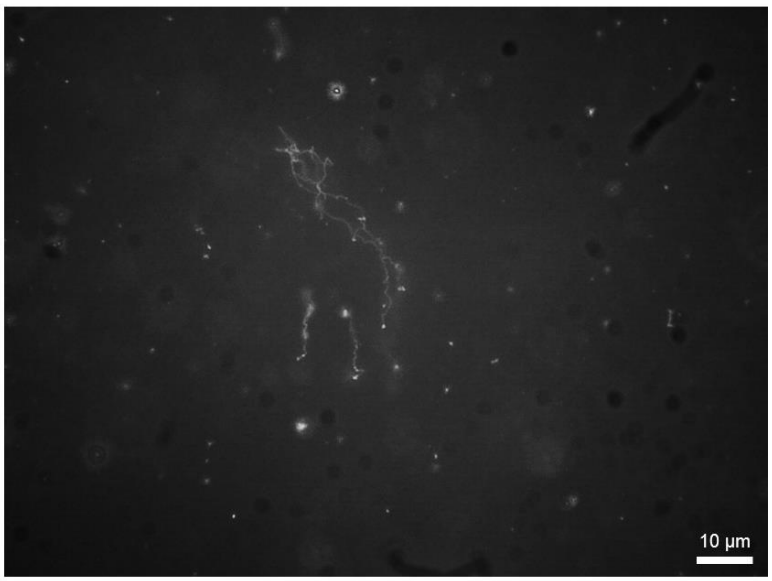


X-03

Figure S6

Mitochondrial DNA branched linear structures





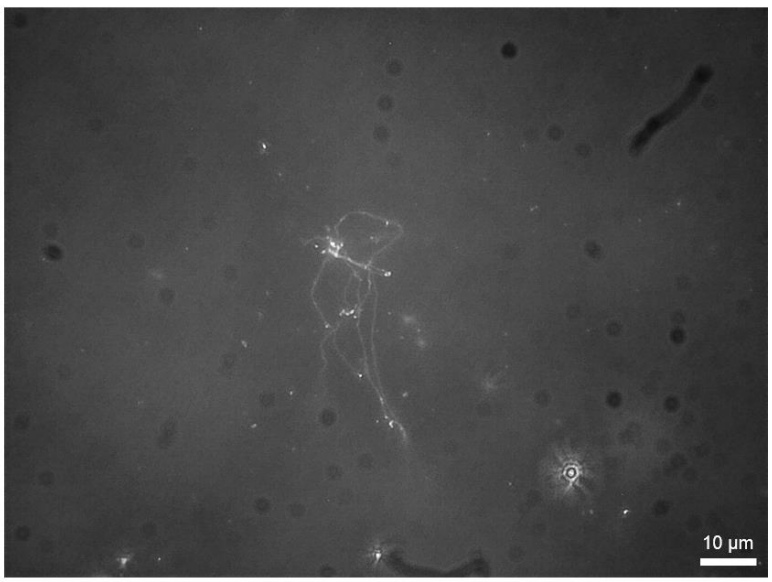


Figure S7

Reanalysis of the mitochondrial genome of *Leucaena trichandra* with PacBio and Illumina mate-pair reads.

Contig ID	Copy#	Length
1	1	186,341
2	1	172,341
3	1	63,894
4	1	56,950
5	1	29,925
6	1	32,568
7	2	43,847
8	2	31,854
9	1	15,237
10	1	7,954
11	3	7,686
12	2	6,421
13	3	1,653
14	2	12,500

Fourteen contigs derived after CLC assembly of the selected PacBio reads (SRX2719625) were ordered using the published linear assembly accession MH717173 along with autonomous element X (accession MH717174, contig 10) using additional information about the sequential order of contigs based on read-through data (reverse read mapping) and long distance analysis with mate-pair reads. Long distance analysis with mate-pair reads clearly demonstrated that the organization of *Leucaena trichandra* mitochondrial genome can by represented in the form of a cyclic graph. All genome segments are linked to each other through several repeats with potential complex rearrangements. Former autonomous element X was placed between a set of repeats that have different segmentation in other parts of the genome.

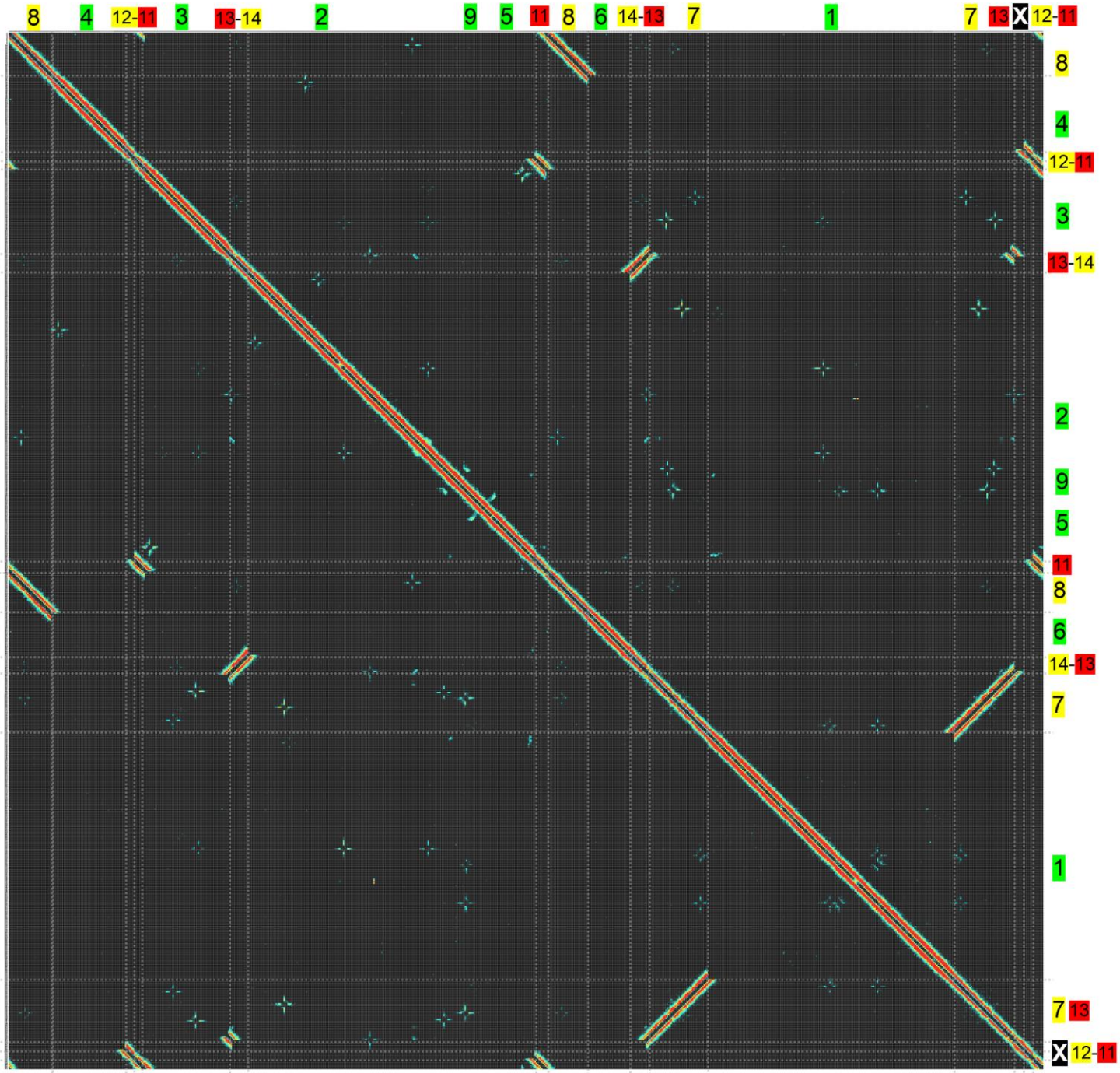
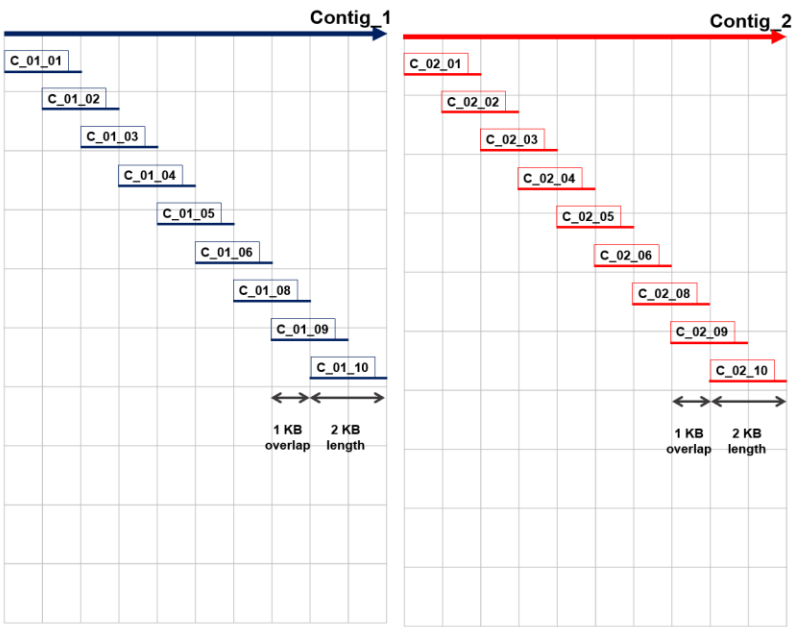
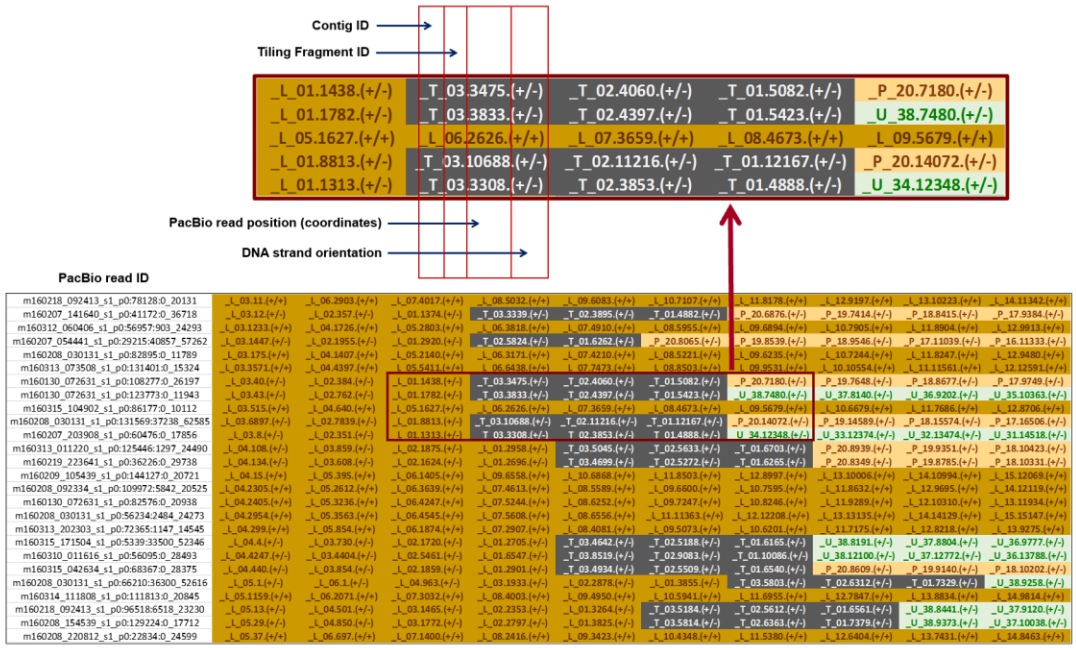


Figure S8

A **Scheme of overlapping tiling fragments for primary structural units**



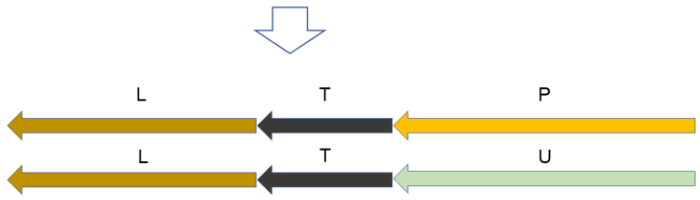
B **Reverse Read Mapping coordinates and inference of secondary building blocks**



Parsed BLAST-N data with detailed info about positioning of tiling fragments over PacBio reads

Reverse Read Mapping protocol outline

1. Construction of primary structural units (contigs) of mitochondrial genome with CLC PacBio assembler
2. Overlapping tiling library of 2 kb long segments for primary structural units (contigs)
3. Alignment/mapping of overlapping tiling library to raw PacBio reads (PacBio reads is a reference)
4. Analysis of sequential order of primary structural units over PacBio reads
5. Compilation of a library of secondary building blocks and inference of isoforms.



Inferred secondary building blocks

Figure S9