Central Infrastructure Group – Bioinformatics
Apl. Prof. Dr. Dirk Walther

Investigating the impact of genomic compartments contributing to non-Mendelian inheritance based on high throughput sequencing data

Kumulative Dissertation

zur Erlangung des akademischen Grades

"doctor rerum naturalium" (Dr. rer. nat.)

in der Wissenschaftsdisziplin "Bioinformatik"

eingereicht an der

Mathematisch-Naturwissenschaftlichen Fakultät

Institut für Biologie und Biochemie

der Universität Potsdam

von

Axel Fischer

Potsdam, 23. September 2021

Reviewers:
Apl. Prof. Dr. Dirk Walther
Universitat Potsdam
Potsdam, Germany

Prof. Dr. Volker Knoop
Universitat Bonn
Bonn, Germany

Prof. Dr. Heiko Schoof
Universitat Bonn
Bonn, Germany

Published online on the
Publication Server of the University of Potsdam:
https://doi.org/10.25932/publishup-54900
https://nbn-resolving.org/urn:nbn:de:kobv:517-opus4-549001

Zusammenfassung

Vor mehr als einem Jahrhundert wurde über das Phänomen der Nicht-Mendelschen Vererbung (NMI), welche als jede Art von Vererbungsmuster definiert wird, in denen Eigenschaften nicht nach dem Mendelschen Gesetzen segregieren, zum ersten Mal berichtet. Im Pflanzenkönigreich können drei Zellkompartimente mit ihrem eigenen Erbgut, Nukleus, Chloroplast, und Mitochondrium, an einem solchen Phänomen beteiligt sein. Hochdurchsatz-Sequenzierungstechnologien (HTS) stellen nachweislich eine Schlüsseltechnologie dar, um NMI Phänomene durch die Assemblierung und/oder Resequenzierung von ganzen Genomen zu untersuchen. Jedoch bleibt die Generierung, Analyse sowie die Interpretation solcher Datensätze durch die vielschichtige biologische Komplexität weiterhin eine Herausforderung. Um unser Wissen über NMI zu erweitern habe ich drei Studien durchgeführt in denen unterschiedliche HTS Technologien involviert waren und zwei neue Algorithmen implementiert um sie zu analysieren.

In der ersten Studie habe ich eine neue Postassemblierungspipeline mit dem Namen Semi-Automated Graph-Based Assembly Curator (SAGBAC) entwickelt, welche nicht Graphen-basierte Genomassemblierungen als Graphen visualisiert, rekombinierende Repeatpaare (RRP) identifiziert, und pflanzliche mitochondrielle Genome (PMG) in einem halb-automatisierten Prozess rekonstruiert. Wir haben diese Pipeline auf Assemblierungen von drei *Oenothera* Spezies angewandt, was in gefalteten zirkularisierten Modellen resultierte. Dieses Modell wurde durch PCR und Southern Blot Analysen bestätigt, sowie verwendet, um einen definierten Satz von 70 PMG Isoformen vorherzusagen. Mit Illumina Mate Pair und PacBio RSII Daten wurde die Stöchiometrie der RRPs quantitativ bestimmt, die sich bis zu dreifach unterscheiden.

In der zweiten Studie habe ich einen post-Multiples Sequenzalignment Algorithmus mit dem Namen Correlation Mapping (CM) entwickelt, der eine segmentweise Anzahl von Nukleotidaustauschen mit numerisch erfassbaren Phänotypen korreliert. Wir haben diesen Algorithmus auf 14 Wildtypen und 18 mutagenisierte Plastomassemblierungen aus der Gattung *Oenothera* angewandt und konnten zwei Gene, *accD* und *ycf2* identifizieren, welche für das kompetitive Verhalten von Plastid Genotypen verantwortlich ist. Weiterhin wird die Lipidkomposition der Plastid-Hüllmembranen durch Polymorphismen in den beiden Genen beeinflusst.

Für die dritte Studie habe ich eine Pipeline programmiert, um ein NMI Phänomen, bekannt als Paramutation, in der Tomate mit Hilfe von DNA- und Bisulfitsequenzierungsdaten sowie Microarray Daten zu analysieren. Wir haben das verantwortliche Gen (Solyc02g005200) identifiziert und waren in der Lage den verursachenden Phänotyp durch eine heterologe Komplementation eines paramutationsinsensitiven Transgens des Orthologs aus *Arabidopsis thaliana* vollständig zu unterdrücken. Weiterhin konnten wir zeigen, dass eine Suppressormutante ein global verändertes

DNA-Methylierungsmuster aufweist und es durch eine große Deletion zu einer Genfusion gekommen ist, an der eine Histon-Deacetylase beteiligt ist.

Zusammenfassend sind meine entwickelten und implementierten Algorithmen und Datenanalysen geeignet, um NMI zu untersuchen und haben wie folgt zu neuen Einsichten über solche Phänomene verholfen: (a) durch die Rekonstruktion von PMGs (SAGBAC) als Voraussetzung um Mitochondrien-assoziierte Phänotypen zu studieren, (b) durch die Identifizierung von Genen (CM), welche interplastidäre Kompetition auslösen sowie (c) durch Anwendung einer DNA-/Bisulfit-seq Analysepipeline zur Beantwortung der Ursache eines transgenerational epigenetischen Vererbungsphänomens.

Abstract

More than a century ago the phenomenon of non-Mendelian inheritance (NMI), defined as any type of inheritance pattern in which traits do not segregate in accordance with Mendel’s laws, was first reported. In the plant kingdom three genomic compartments, the nucleus, chloroplast, and mitochondrion, can participate in such a phenomenon. High-throughput sequencing (HTS) proved to be a key technology to investigate NMI phenomena by assembling and/or resequencing entire genomes. However, generation, analysis and interpretation of such datasets remain challenging by the multi-layered biological complexity. To advance our knowledge in the field of NMI, I conducted three studies involving different HTS technologies and implemented two new algorithms to analyze them.

In the first study I implemented a novel post-assembly pipeline, called Semi-Automated Graph-Based Assembly Curator (SAGBAC), which visualizes non-graph-based assemblies as graphs, identifies recombinogenic repeat pairs (RRPs), and reconstructs plant mitochondrial genomes (PMG) in a semi-automated workflow. We applied this pipeline to assemblies of three *Oenothera* species resulting in a spatially folded and circularized model. This model was confirmed by PCR and Southern blot analyses and was used to predict a defined set of 70 PMG isoforms. With Illumina Mate Pair and PacBio RSII data, the stoichiometry of the RRPs was determined quantitatively differing up to three-fold.

In the second study I developed a post-multiple sequence alignment algorithm, called correlation mapping (CM), which correlates segment-wise numbers of nucleotide changes to a numeric ascertainable phenotype. We applied this algorithm to 14 wild type and 18 mutagenized plastome assemblies within the *Oenothera* genus and identified two genes, *accD* and *ycf2* that may cause the competitive behavior of plastid genotypes as plastids can be biparental inherited in *Oenothera*. Moreover, lipid composition of the plastid envelope membrane is affected by polymorphisms within these two genes.

For the third study, I programmed a pipeline to investigate a NMI phenomenon, known as paramutation, in tomato by analyzing DNA and bisulfite sequencing data as well as microarray data. We identified the responsible gene (Solyc02g0005200) and were able to fully repress its caused phenotype by heterologous complementation with a paramutation insensitive transgene of the *Arabidopsis thaliana* orthologue. Additionally, a suppressor mutant shows a globally altered DNA methylation pattern and carries a large deletion leading to a gene fusion involving a histone deacetylase.

In conclusion, my developed and implemented algorithms and data analysis pipelines are suitable to investigate NMI and led to novel insights about such phenomena by reconstructing PMGs (SAGBAC) as a requirement to study mitochondria-associated phenotypes, by identifying genes (CM) causing interplastidial competition as well by applying a DNA/Bisulfite-seq analysis pipeline to shed light in a transgenerational epigenetic inheritance phenomenon.

Table of Contents

List of figures

Abbreviations

3GS	Third generation sequencing
bp	base pair
BSMAP	Bisulfite Sequence MAPping
CM	Correlation mapping
CMS	Cytoplasmic male sterility
CNV	Copy number variant
CRC	Contig-repeat-contig
DBG	De Bruijn graph
DNA	Desoxyribonucleic acid
F1	Filial 1 generation
GC	Guanine cytosine
HTS	High throughput sequencing
Indel	Insertion/deletion
IR	Inverted repeat
IDBA	Iterative De Bruijn Graph Assembler
ISEIS	Iterative Sequence Ends Identity Search
LSC	Large single copy
MIRA	Mimicking Intelligent Read Assembly
MSA	Multiple sequence alignment
mt	mitochondrion
mtDNA	mitochondrial DNA
NGS	Next generation sequencing
NMI	Non-Mendelian inheritance
OLC	Overlay layout consensus
ORF	Open reading frame
PacBio	Pacific Biosciences
PCR	Polymerase chain reaction
PMG	Plant mitochondrial genome
pt	plastid

ROSU	Revertant of *SULFUREA*
RRP	Recombinogenic repeat pair
SAGBAC	Semi-Automated Graph-Based Assembly Curator
sRNA	small RNA
SNP	Single nucleotide polymorphism
SOSU	Suppressor of *SULFUREA*
SSC	Short single copy
Sulf	sulfurea
SV	Structural variant

1 Introduction

1.1 Non-Mendelian inheritance

In 1909 two scientists simultaneously but independently reported for the very first time the phenomenon of non-Mendelian inheritance (NMI) in plant species (Baur, 1909; Correns, 1909). It is defined as any type of inheritance pattern in which traits do not segregate in accordance with Mendel's laws. In the plant kingdom three genomic compartments, the nucleus, chloroplast, and mitochondrion, can participate in such a phenomenon. Common NMI phenomena are e.g. codominance, incomplete dominance and epistasis. A more uncommon NMI phenomenon is transgenerational epigenetic inheritance also known as paramutation in which DNA methylation of the nuclear genome is mechanistically involved. Whereas so called extranuclear (or extrakaryotic) inheritance involves genomic compartments other than the nucleus which are chloroplasts and mitochondria in plants.

1.1.1 Organelle inheritance in *Oenothera*

Since the beginning of the last century, the evening primrose (*Oenothera*), with its special genetic features, provides a unique toolbox for scientists to investigate incompatibilities between the three different genomic sources present in plant cells. In most common plant model species organelles are co-inherited maternally and therefore nearly impossible to genetically separate their cytoplasmic effects from each other (Greiner et al., 2015). In *Oenothera*, however, biparental inheritance of plastids (Cleland, 1972) but uniparental inheritance of mitochondria is observed (Brennicke and Schwemmle, 1984; Greiner et al., 2015; Dotzek, 2016; Ulbricht-Jones et al., 2021). Cytoplasmic effects, and with this also observed NMI phenomena, in reciprocal crosses can therefore unequivocally attributed to one of the two organelle genomes.

Oenothera is historically one of the first plant models in which its mitochondrial DNA was investigated (Brennicke, 1980; Wissinger et al., 1991). Evening primrose geneticists have shown in a classical cross between *Oenothera berteriana* and *O. odorata* that a genetic determinate in the cytoplasm influences floral traits which was later called mitochondrion (Schwemmle, 1938; Ulbricht-Jones et al., 2021). But few additional putative extrakaryotic inheritance patterns of unknown origin are described from the evening primrose (Barthelmess, 1965; Stubbe, 1989a, b; Harte, 1994). A transition from morphological markers to molecular markers is obligatory to clearly identify the causative genomic compartment as well as underlying mechanisms for those extrakaryotic inheritance patterns. Therefore, high-quality plant mitochondrial genome (PMG) sequences – which also include structural information - are highly desirable.

In contrast to most plant species, plastids can be biparentally inherited in *Oenothera* and do neither fuse among each other nor undergo sexual recombination during meiosis (Cleland, 1972). Therefore, chloroplasts compete for cellular resources, which led to the theory of “selfish” cytoplasmic elements, originally developed in evening primrose genetics as naturally occurring aggressive chloroplasts exist in *Oenothera* (Grun, 1976; Greiner et al., 2015). In *Oenothera* five genetically distinguishable chloroplast genome types, designated by Roman numbers (I-V), were identified by extensive crossing studies (Greiner et al., 2008) and were groupable into three classes according to their inheritance strength: strong plastomes (I and III), intermediate (II) and weak plastomes (IV and V). This kind of assertiveness rate reflects their ability to outcompete a second plastid genome in the F1 generation upon biparental transmission. The plastome types were initially identified based on their (in)compatibility with certain nuclear genomes (Stubbe, 1964; Greiner et al., 2011). The determinants which are responsible for the variation in competitive ability of the different chloroplast types can be clearly traced back to the chloroplast genome (Cleland, 1972; Gillham, 1978; Kirk and Tilney-Basset, 1978). However, the genes involved in the control of chloroplast competition are not known yet.

1.1.2 Paramutation in *Solanum lycopersicum*

Paramutation is an epigenetic NMI phenomenon in which heritable changes in gene expression is involved. It is defined by a silencing conferring interaction between a pair of homologous alleles: a paramutable allele and paramutagenic allele. Typically, the paramutagenic allele is silent but has the capability to impose its silence state onto the paramutable allele present in trans. Interestingly the newly silenced (paramutated) allele persists in progeny and can act there as paramutagenic allele itself. Paramutation was initially investigated in plant species (garden peas (Bateson and Pellew, 1920) and evening primroses (Renner, 1938)) and could later on be extended also to animals (Brink, 1973; Rassoulzadegan et al., 2006; Chandler, 2007). The paramutation phenomenon is most intensely studied in maize (Brink 1958) as the molecular identity of genes affected by paramutation are known (Hollick et al., 1997; Dorweiler et al., 2000; Sidorenko and Peterson, 2001). From the timepoint of initial observation until today a complex working model was developed for this organism.

Chromatin structure and DNA methylation analyses of paramutated loci in maize (Lisch et al., 2002; Stam et al., 2002a; Stam et al., 2002b; Chandler and Stam, 2004; Belele et al., 2013) as well as the characterization of several maize mutants revealed a number of candidate genes which are part of an unconventional transcriptional silencing pathway (Dorweiler et al., 2000; Hollick et al., 2005; Alleman et al., 2006; Woodhouse et al., 2006; Hale et al., 2007; Erhard et al., 2009; Barbour et al., 2012). This pathway, which is quite complex, involves transcription, small RNA biogenesis and DNA methylation machineries and is here only briefly summarized (Giacopelli and Hollick, 2015): (i) Transcription of repetitive enhancer elements (b1 locus) by DNA-dependent RNA-polymerases (Pol IV), (ii) generation and processing of dsRNA producing 24 bp small RNAs (sRNA), (iii) forming sRNA-Argonaute complexes, and (iv) recruitment of DNA methyltransferases by formed sRNA-Argonaute complexes to homologous sequences triggering *de novo* cytosine methylation in sequence elements controlling transcription of the paramutable allele.

Nevertheless, paramutation was also observed in tomato generated by X-ray mutagenesis experiments leading to the *sulfurea* mutant (Hagemann, 1958). It is only one of the few paramutagenic loci found in a dicotyledonous model species which is routinely amenable to genetic manipulation by stable transformation. Paramutated sulfurea tissue shows an astonishing yellow, chlorophyll-deficient phenotype (Hagemann, 1958; Ehlert et al., 2008). The *sulfurea* (*sulf*) allele is recessive, but the pigment deficiency appears spontaneously in somatic tissues of heterozygous plants at high frequency leading to variegated leaves with green (*sulf (het-g)*) and yellow portions (*sulf het-y*) and thus fulfills the criteria of paramutation.

In general, high throughput sequencing (HTS) proved to be a key technology to study NMI mechanisms by assembling and/or resequencing entire genomes as first step.

1.2 High Throughput Sequencing

The costs of HTS have dropped tremendously (www.genome.com/sequencingcosts) and there are almost no limits to sequence any genome in a short period of time. In the last 15 years different companies emerged with various library preparation types and sequencing technologies/devices leading to the more specific terms Next Generation Sequencing (NGS) and Third Generation Sequencing (3GS). The latter term is defined by its unique feature to produce ultra-long reads from dozen up to several hundreds of kilobases. While NGS is nowadays predominantly covered by Illumina, 3GS is the campus of Pacific Biosciences (PacBio) (Rhoads and Au, 2015) and Oxford Nanopore (Deamer et al., 2016). Nevertheless, the analysis and interpretation of huge amounts of generated HTS data remains challenging and has become the bottleneck in many project workflows, in resequencing and assembly projects alike.

Resequencing projects in which sequences can be mapped against well-known and annotated reference sequences have different requirements on library construction and sequencing types than *de novo* assembly projects with the goal to generate a high-quality genome from scratch. In both types of projects GC content, genome size and its repetitiveness are the main properties of a genome with the highest impact on decisions made within the workflows. In Figure 1 individual workflows are displayed for the three presented studies which differ in tissue type (seedlings or mature leaves) and compartment/DNA fraction (total DNA or mitochondria-enriched DNA), chosen HTS library types and used HTS technology as well as generated read depth to fulfill needed requirements for sustainable bioinformatic analyses.

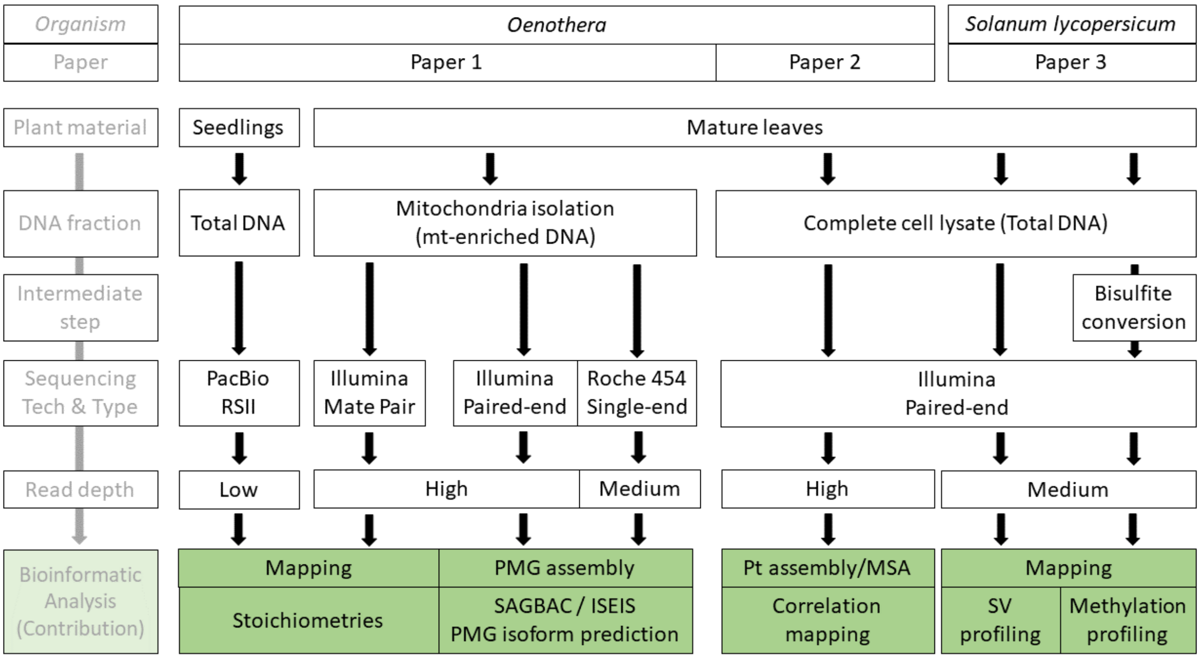


Figure 1. Overview of complete wet lab/dry lab workflows for all HTS datasets generated in all three studies.

On the left the generalized schemata for the workflows is displayed, starting with the plant material type of a specific organism, and ending with the bioinformatic analysis which displays the contributions highlighted within each of the presented studies. ISEIS = Iterative Sequence Ends Identity Search; MSA = Multiple Sequence Alignment; mt = Mitochondria; PMG = Plant mitochondrial genome; Pt = Plastome; SAGBAC = Semi-automated graph-based assembly curator; SV = Structural variants.

Additionally, an overview of all four used HTS libraries preparation types is separately visualized in Figure 2 as they take an important role in *de novo* assembly projects which will discussed later also in regards of spanning repetitive elements.

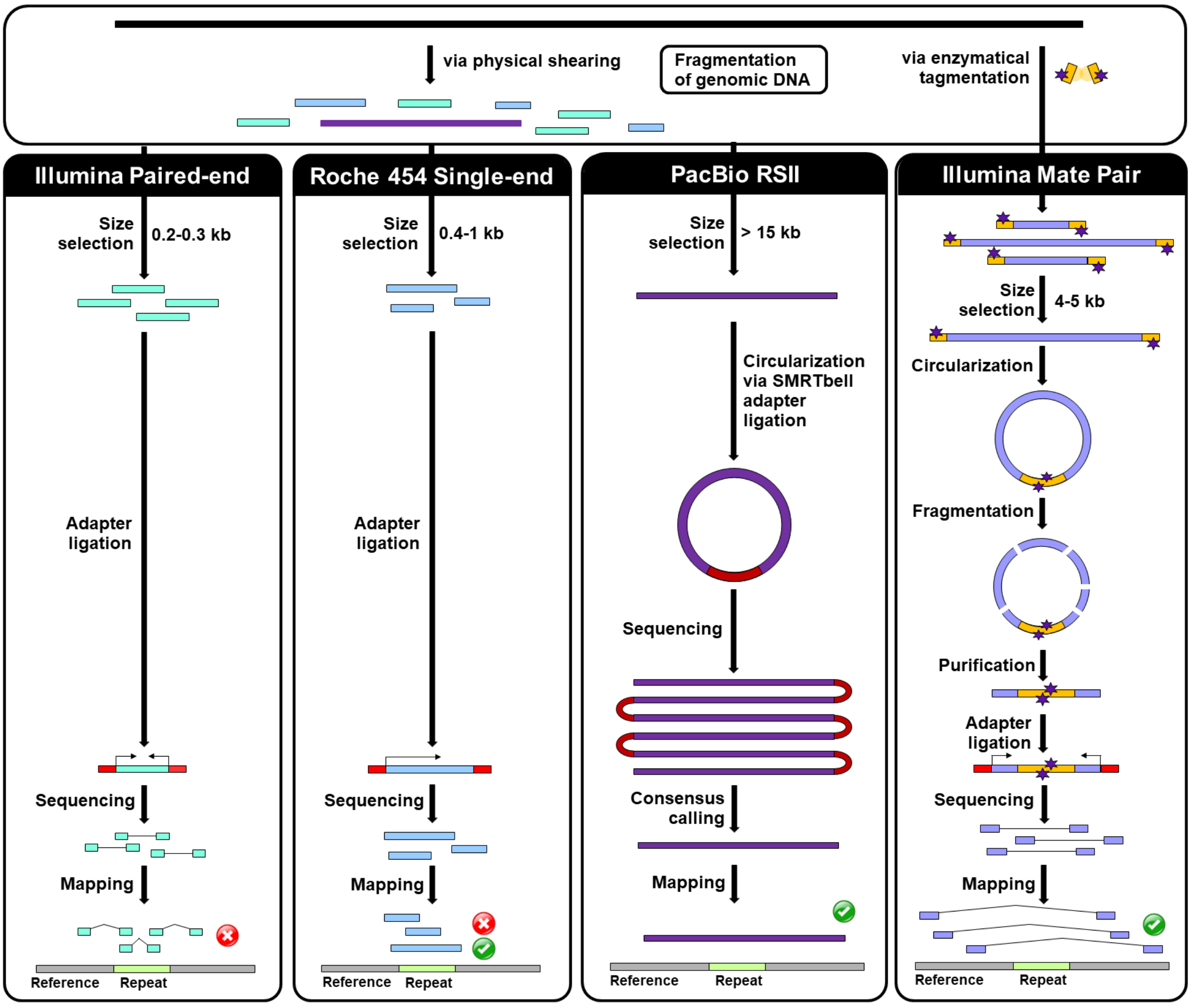


Figure 2. Overview of the wet lab workflows for all four used high throughput sequencing types. Shown are the library construction procedures of all four high throughput sequencing library types used within all three presented studies. At the bottom line is highlighted whether the sequence library type is capable to span repetitive elements (green check mark) or not (red X mark). Both mates of a pair or both ends of a single read need to map on unique genomic sequences in order to span repetitive sequences correctly.

1.3 *De novo* assembler programs

Many *de novo* assemblers were programmed within the NGS community for different purposes: from small and simple to large and complex genomes as well as from single-cell transcriptomes to complex metagenomes. Metagenomics literally means “beyond the genome” and is usually defined as the study of a mixture of genomes from different taxonomies or organisms (Handelsman et al., 1998). But it can also be defined as a mixture of genomes within one organism on the level of genomic compartments: nucleus, chloroplast, and mitochondrion. All three compartments harbor their own genome with their own levels of complexity which have the highest impact on the choice of HTS technology and *de novo* assembly algorithm that should be used. In general, *de novo* assembly algorithms can be divided into two classes: Overlap Layout Consensus (OLC) and De Bruijn Graph (DBG). Both *de novo* assembler algorithm types have in common that they handle unresolvable repeats by essentially leaving them out or report them separately which breaks the assembly into fragments. And as mitochondrial and plastidial genomes have different biological characteristics they also have different requirements in applicable assemblers as follows.

To determinate applicable assemblers for the genome of interest two general aspects should be considered: employability of HTS data and knowledge about biological characteristics of the genome of interest. Figure 3 illustrates how different combinations of wet lab steps and biological factors can influence the ability to assemble a certain organelle genome. Total DNA extracts from seedlings are unsuitable for this task as in this developmental stage chloroplasts are outnumbered (Figure 3A). Whereas total DNA extracts from mature leaves (Figure 3B) have a high abundance of plastidial DNA (Golczyk et al., 2014; Greiner et al., 2020) and delivers a high read coverage. This approach is also referred to the term “genome skimming” as even whole genome sequencing at low coverage (equal to low costs) harbors enough reads to assemble plastidial genomes. Many popular reference-based or *de novo* assem-

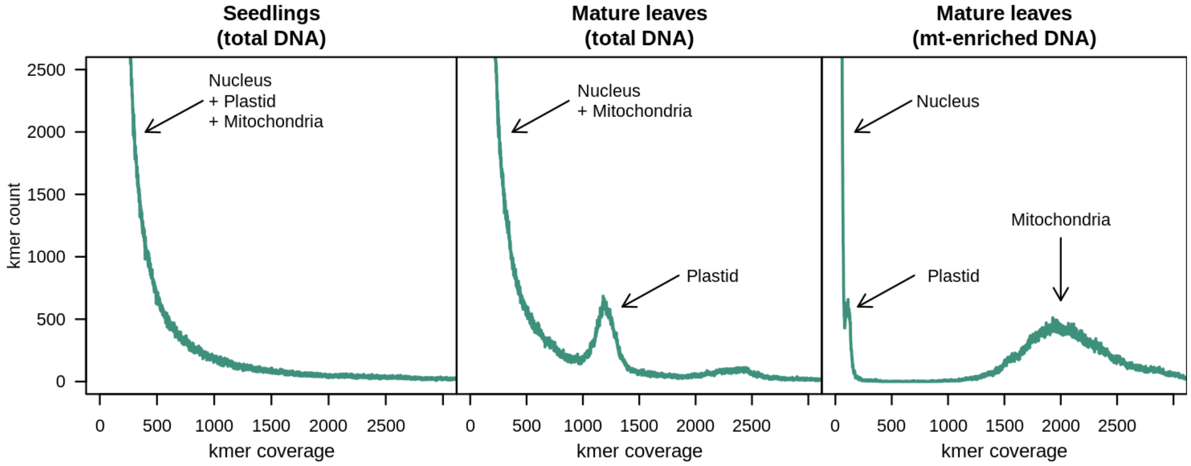


Figure 3. Kmer plots from various tissues, developmental stages and DNA compositions. Displayed are 27-kmer plots generated from three different Illumina Paired-end datasets of equal size (15 mio. reads). They illustrate the changing content of DNA from all three genomic compartments depending on developmental stages of used plant material and DNA composition. mt = mitochondria.

blers can be used afterwards to generate the desired chloroplast genome in a single contig, as they are in almost all cases structurally conserved and repeat-poor (Straub et al., 2011; Zhang et al., 2011; Sloan et al., 2014). As mitochondrial DNA is underrepresented in total DNA extracts from seedlings as well as from mature leaves, it is necessary to enrich for mitochondria before the DNA extraction procedure (Figure 3C).

But besides the importance of mitochondrial enrichment to obtain pure mitochondrial DNA, biological characteristics of plant mitochondrial genomes needed to be considered as well. For *Oenothera* many studies (Brennicke et al., 1985; Hiesel and Brennicke, 1985; Schuster and Brennicke, 1988; Binder et al., 1990) have shown that the mitochondrial genome contains recombinogenic repeat pairs (RRP) which are able to influence the structure by two different recombination mechanisms as shown in Figure 4.

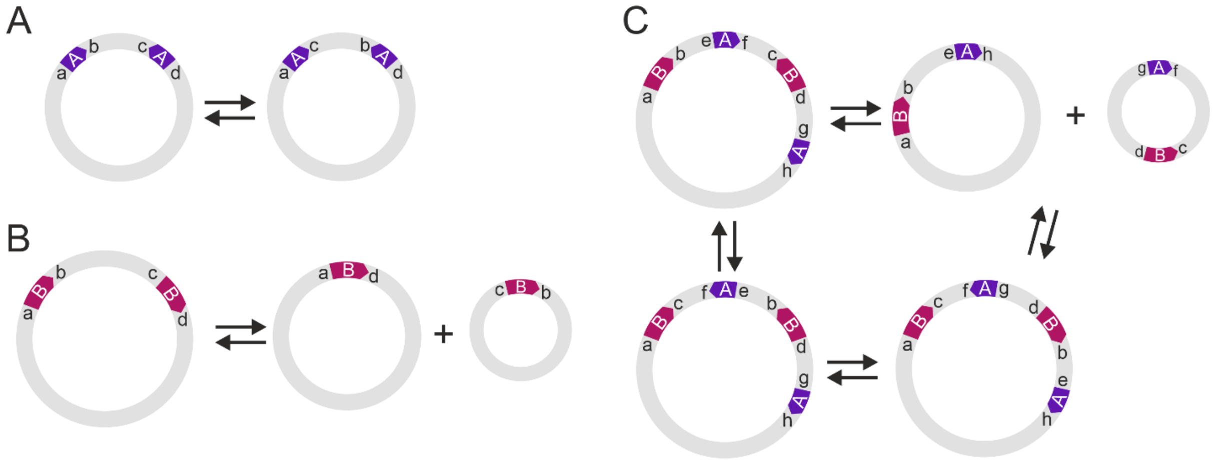


Figure 4. Possible recombination events in plant mitochondrial genomes.
(A) Pair of inverted RRPs, recombination leads to a sequence inversion between the mates of the inverted pair. (B) Pair of directed RRPs, recombination leads to the generation of two circular molecules. (C) A pair of inverted and direct RRPs which are nested. This gives four possible genome configurations and for each RRP four possible contig-repeat-contig (CRC) configurations. The larger the number of direct and inverted repeats the more complex the genome organization becomes (Redrawn image and adapted figure legend is taken from (Lonsdale, 1984).

If an inverted RRP is present in the genome the recombination event leads to an inverted sequence between both mates of the RRP. If a direct RRP exists in the genome the recombination event creates two circular products. Both recombination event types are reversible (Lonsdale, 1984). By the combination of a direct and an inverted RRP four different genome configurations, and with it four different contig-repeat-contig (CRC) combinations for each RRP, are possible (e-A-f; e-A-h; f-A-g; g-A-h for RRP A in Figure 4C). To obtain stoichiometries for RRPs, length of sequenced DNA templates (insert size) as well as suitable HTS library types are important (Figure 2). To span RRPs the insert size of the library needs to be longer than the length of the RRP itself. If the library does not fulfill this criterion stoichiometric proportions cannot be calculated. This can only be achieved by Illumina Mate pair as well as long Roche and PacBio libraries whereas Illumina paired-end libraries cannot span repetitive sequences by its limited insert size to get usable reads (Figure 2). With an increasing number of RRPs genome organization also increases. Such a structure complexity would lead inevitably into a set of discontiguous contigs which necessitates a specialist post assembly program including the possibility to calculate stoichiometries for RRPs.

1.4 DNA-/Bisulfite sequencing

Besides standard DNA-sequencing (e.g. Illumina Paired-end, Figure 2) which is now standard for several years to resequence entire nuclear genomes, it is also possible to directly detect methylated cytosines on single nucleotide level on a genome-wide scale (Lister and Ecker, 2009). This is accomplished by introducing an intermediate step in the library preparation called bisulfite conversion as illustrated in Figure 5 (Frommer et al., 1992).

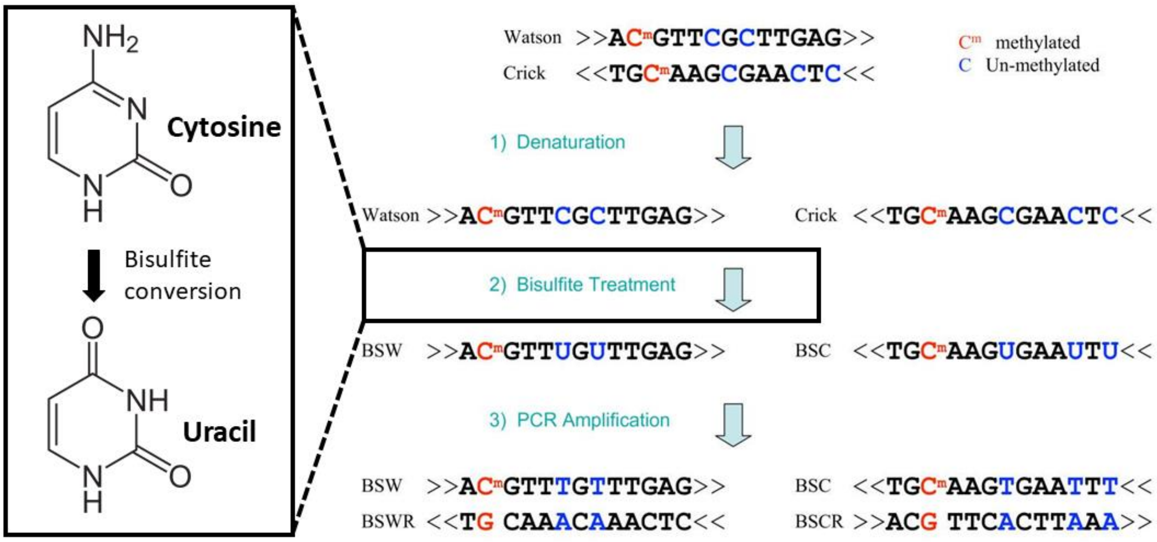


Figure 5. Main wet lab steps of the bisulfite sequencing pipeline.

Displayed is the bisulfite sequencing pipeline in which DNA is denaturated, bisulfite treated and amplified by PCR ending up in four different strands, as different bases of both original DNA strands become converted. Only those cytosines are converted to uracil (intermediate reactions are sulfonation, deamination and desulfonation) which are unmethylated. Methylated cytosines are not susceptible to bisulfite conversion. Original image is taken from (Xi and Li, 2009) and has been edited.

By a treatment with sodium bisulfite unmethylated cytosines are converted to uracil which pairs with adenosine and are replaced by thymidine in the PCR step afterwards. Methylated cytosines cannot be converted and become preserved. As different positions of the original DNA strands are converted four independent strands of DNA are produced after the PCR step. This is also the reason why specialist mapping tools were put in place such as Bismark (Krueger and Andrews, 2011) and BSMAP (Xi and Li, 2009). Methylation status of all cytosines within the genome of interest is then called where cytosines can appear in different base context: CpG, CHG and CHH where H stands for every base excluding cytosine (A=adenosine, G=guanosine and T=thymidine), and determines the regulation of its methylation (Zhang et al., 2018). Differential methylation can afterwards be calculated between two samples e.g. mutant versus wild type.

1.5 Structural variants

Structural variants are changes in the structure of genomic sequences which ranges from deletions and insertion, inversions, translocations over tandem and interspersed duplications to copy number variants. Copy number variation (CNV) is defined as a structural variation which results in a loss or gain of genomic sequence ranging from kilobases to several megabases. CNVs are an important source of genetic variation but was long time neglected especially within the research field of plant domestication (Lye and Purugganan, 2019). To detect larger deletions and amplifications in genomes most often segmentation algorithms are applied to NGS data. All segmentation algorithms have in common that the genome sequence of interest is split into segments of defined size and mapped NGS reads are counted per segment (Pirooznia et al., 2015). Afterwards proportions between two samples (e.g. mutant versus wild type) are calculated, normalized, logarithmized and are finally plotted chromosome-wise. Within a diploid organism, two alleles are present. A deletion event in one copy would cause the loss of one allele. If both boundaries of such a deletion event map to sequences of two distinct genes, a fusion of two genes is formed. From such a gene fusion, if completely in-frame, a gene with properties of both participating genes could be created and the second gene's content comes under the control of the promotor of the upstream gene.

1.6 Aim of the thesis

The task of this thesis was to develop bioinformatic algorithms and pipelines to aid the investigation of NMI mechanisms in plants based on different HTS technologies – a wonderful challenge given the complex nature of plant biology underneath.

First, I aimed to develop a post-assembly pipeline to transfer to and to visualize non-graph-based *de novo* assembler outputs in a graph-based model and to identify repetitive elements which are responsible for recombinatoric events in PMGs. The developed pipeline was needed to be capable to predict PMG isoforms from such a model, to calculate stoichiometries of the determined repetitive elements based on HTS data and to reconstruct PMGs. All these properties of the pipeline to be developed are prerequisites to study mitochondria-associated phenotypes in *Oenothera*. Second, I intended to implement a post-multiple sequence alignment algorithm to correlate nucleotide exchange rates with numeric ascertainable phenotypes to identify genes which are responsible for the competition of different plastid types in *Oenothera*. Third, I planned to design a pipeline to analyze DNA- and Bisulfite sequencing data as well as microarray data to find the causing gene for a paramutation in tomato.

In summary, my thesis centered on the development of algorithms and data analysis pipelines intended to expand the knowledge of NMI mechanisms covering all three genomic compartments present in plant cells based on HTS data.

2 Contributions

Paper 1: Graph-based models of the *Oenothera* mitochondrial genome capture the enormous complexity of higher plant mitochondrial DNA organization

Axel Fischer devised and implemented all bioinformatic pipelines and algorithms and analyzed all Next Generation Sequencing data. Jana Dotzek established the mitochondria enrichment protocol and performed the PCR and Southern blot validation experiments. Axel Fischer wrote the manuscript with guidance and editing by Dirk Walther and Stephan Greiner.

Paper 2: Chloroplast competition is controlled by lipid biosynthesis in evening primroses

Barbara B. Sears and Stephan Greiner designed the study. Johanna Sobanski performed the main experimental work. Patrick Giavalisco, Axel Fischer, Julia M. Kreiner, Mark Aurel Schöttler, Tommaso Pellizzer, Hieronim Golczyk, Toshihiro Obata, Barbara B. Sears, and Stephan Greiner provided supportive data. Axel Fischer and Stephan Greiner developed the correlation mapping approach. Axel Fischer assembled the plastomes and performed the MSA. Julia M. Kreiner implemented PGLS. Johanna Sobanski, Patrick Giavalisco, Axel Fischer, Julia M. Kreiner, Dirk Walther, Mark Aurel Schöttler, Tommaso Pellizzer, Hieronim Golczyk, Toshihiro Obata, Ralph Bock, Barbara B. Sears, and Stephan Greiner analyzed and discussed the data. Johanna Sobanski and Stephan Greiner wrote the manuscript. Patrick Giavalisco, Axel Fischer, Dirk Walther, Mark Aurel Schöttler, Hieronim Golczyk, Ralph Bock, and Barbara B. Sears participated in the writing.

Paper 3: Identification of the paramutated *SULFUREA* locus of tomato and release from epigenetic silencing by spontaneous reversion or genetic suppression

Ralph Bock devised the study and designed research approaches. Britta Ehlert and Axel Fischer designed and performed the experimental and computational work and analyzed the data. Axel Fischer conducted all next generation sequencing data analyses, comprising the DNA-/Bisulfite-sequencing, structural variant and gene fusion analysis as well as the microarray analysis. Ralph Bock analyzed data and wrote the paper with participation of all co-authors.

Apl. Prof. Dr. Dirk Walther

Potsdam, 16.09.2021

2.1 Paper 1: Graph-based models of the *Oenothera* mitochondrial genome capture the enormous complexity of higher plant mitochondrial DNA organization

Axel Fischer †*, Jana Dotzek †, Dirk Walther, Stephan Greiner*

Max Planck Institute of Molecular Plant Physiology, Am Mühlenberg 1, D-14476 Potsdam-Golm, Germany

† These Authors contributed equally

Corresponding authors: afischer@mpimp-golm.mpg.de; greiner@mpimp-golm.mpg.de

Submitted to Journal: NAR Genomics and Bioinformatics, on 8 September 2021

ABSTRACT

Plant mitochondrial genomes display an enormous structural complexity, as recombining repeat pairs lead to the generation of various sub-genomic molecules, rendering these genomes extremely challenging to assemble. Here, we present a novel bioinformatic data-processing pipeline called SAGBAC (Semi-Automated Graph-Based Assembly Curator) that identifies recombinogenic repeat pairs and reconstructs plant mitochondrial genomes. SAGBAC processes non-graph-based assembly outputs using our novel ISEIS (Iterative Sequence Ends Identity Search) algorithm to obtain a graph-based visualization. We applied this approach to three mitochondrial genomes of the evening primrose (*Oenothera*), a model organism for cytoplasmic genetics. We could show that all identified repeat pairs are flanked by two alternative unique sequence-contigs defining so-called "double forks", leading to four contig-repeat-contig combinations for each repeat pair. Based on the structural model, the stoichiometry of the different contig-repeat-contig combinations was analyzed using Illumina mate pair and PacBio RSII data. This uncovered a remarkable structural diversity of the closely related mitochondrial genomes, as well as substantial phylogenetic variation of the underlying repeat units. Our model allows predicting all recombination events and, thus, all possible sub-genomes. In future work, the proposed methodology can contribute to the investigation of the sub-genome organization and dynamics in different tissues and at various developmental stages.

INTRODUCTION

Plant mitochondrial genomes (PMGs) vary enormously in complexity, size, and structure (1,2). In addition to the circular mitochondrial genome - once considered a dogma - new structural types such as linear chromosomes or multiple circles (3-6) have been identified in plants. At the genome molecular sequence level, creation of alternative master- and sub-circles is realized by pairs of repetitive elements (known as recombinogenic repeat pairs or RRPs), which can lead to two different recombination events - depending on the relative orientation of both mates of an RRP to each other (7). Such events result in different genome configurations and can generate a population of different master- and sub-circles within the mitochondrion.

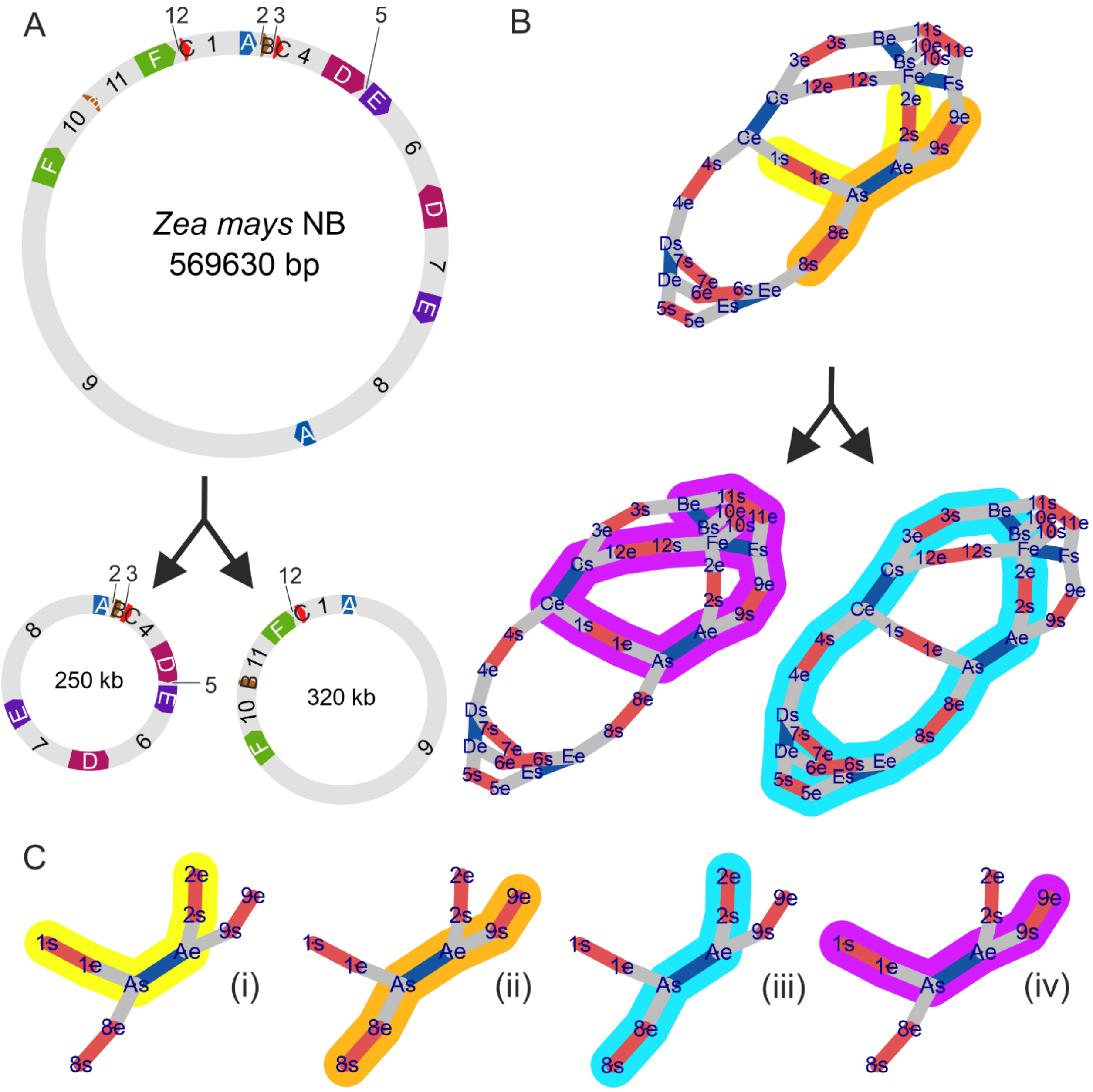


Figure 1. Graph-based visualization of circular genomes. (A) Circular visualization of *Zea mays* NB, redrawn from (122) using AngularPlasmid. Contigs between repeats are indexed numerically starting at 12 o'clock going clockwise; same holds for the RRPs for their first occurrence but using letters. The same letter was used for the second mate of an RRP when going through the master circle sequence. Both mates of RRP_A are flanked by unique sequences as follows: 1-a-2 and 8-a'-9. A recombination event at RRP_A within the master circle leads to the formation of two sub-circles in which RRP_A is now flanked by 1-a-9 and 8-a'-2. (B) Same information content as (A), but visualized as graphs with the following translations: Contig ends are respectively represented by two vertices linked by a red edge. Repeat ends are also represented by two vertices linked by a blue edge. Each transition between two different contigs (independent of its type, contig or repeat) is represented by a gray edge between their respective vertices. Vertices of repeats must be connected to more than one other non-repeat vertex. Color-marked paths highlight used contigs and RRPs to their corresponding circular equivalents in (A). (C) Shown is only RRP_A with its unique flanking sequences on both sides (= "double fork"). i-iv: all four possible contig-repeat-contig (CRC) combinations derived from the graphs in (B) for RRP_A. Again, colors highlight their origin from each of their corresponding graphs in (B).

Nevertheless, PMGs are typically still represented as a single circular genome, thereby not reflecting the complexity of a population of different master- and sub-circles within the mitochondria of a plant cell.

Sequence-structure conversion – from sequence to graph

Reconstructing the complexity of a variable sub-circularizing genome using current sequencing technologies proves challenging. Piecing short sequence reads together and considering alternative topologies requires specialized approaches, as typically, assembly programs tend towards assembling reads into the largest possible contig. Furthermore, the combinatorial complexity of possible assembly paths poses computational challenges. These problems can be addressed by graph-based approaches (8-10), illustrated in Figure 1. In Figure 1a, a sub-circularization event is illustrated, forming two sub-circles derived from the master circle of the *Zea mays* mitochondrial DNA (NC_007982.1) using circular representations. These circular representations can be converted into a graph-based representation as shown in Figure 1b. Comparing the circular and graph-based representation, the interpretability of the graph is much easier than that of a circular representation, as all sub-circles can be readily generated by their corresponding subpath (Figure 1b). Also, the graph illustrates that a repeat is flanked on both sides by a set of two unique sequences, henceforth named "double fork", resulting in four different contig-repeat-contig (CRC) combinations (Figure 1c).

Study system

Historically, the evening primrose (*Oenothera*) represents one of first plant models analyzed for its mitochondrial DNA (11,12). The reason for choosing this model species already in the early days of molecular biology lies in its major importance for the study of extranuclear inheritance (13,14). In a now classical cross between *Oenothera berteriana* and *O. odorata*, evening primrose geneticists could show as early as the 1930 that a genetic determinant in the cytoplasm influences floral traits (15,16). The determinant was later called the mitochondrion, and in fact, the *Oenothera* system is famous for the possibility to separate the genetic effects of chloroplast and mitochondria from each other (13,16,17). This is in contrast to common models or crop species used in plant biology. Those display maternal co-inheritance of their cell organelles (14) and in those systems it is difficult, if not impossible, to genetically separate cytoplasmic effects of the chloroplast from that of the mitochondria (18). In *Oenothera*, however, biparental inheritance of chloroplasts, but uniparental inheritance of mitochondria is observed (14,16,19,20). Cytoplasmic effects in reciprocal crosses can therefore unequivocally be attributed to one of the two organelle genomes. This is one of the reasons, why *Oenothera* has developed into a model system for organelle genetics and population biology, in which, for example, aspects or hybrid incompatibly, organelle mediated adaptation, speciation, or organelle inheritance are studied (e.g. 21,22,23). *Oenothera* is one of the few examples for which plastid-borne cytoplasmic male sterility (CMS) could be demonstrated (24) and is currently developing as a model to study organelle signalling involved in plant development (16). For these reasons, and also because putative extrakaryotic inheritance patterns of unknown origin have been described in *Oenothera* species (25,26,27,28), a high-quality mitochondrial genome sequence - that also includes structural information - is highly desirable.

The aim of this study was the assembly and annotation of the mitochondrial genomes of three major experimental strains of the genus *Oenothera*, representing the species *O. villaricae* (referred to as *O. berteriana* in the genetic literature, see above), *O. biennis*, and *O. elata*. The latter two are closely related and belong to the North American subsection *(Eu)Oenothera*, whereas *O. villaricae* is a member of South American subsection *Munzia*, the sister subsection of subsection *(Eu)Oenothera* (Wagner et al. 2007). Assembling plant mitochondrial genomes can lead to a set of discontinuous and unconnected contigs, especially when recombinogenic repeat pairs, RRPs, are present. Typically, insert sizes of paired-end Illumina short NGS reads are shorter than the repeat size and therefore cannot span the repeats entirely. Since, usually, it is desired to generate, “the one and only” mitochondrial genome (configuration), this is considered a disadvantage. However, this perceived “disadvantage” of discontinuous contigs as the outcome of a *de novo* assembler can, in fact, be turned into an advantage. As we will demonstrate here, it allows to highlight and investigate the true complexity of a plant mitochondrial genome. Instead of trying to deduce a circular configuration from a single contig, we are performing a graph-to-sequence conversion (i.e. deducing a much more complex sequence organization from a graph, Figure 1). For this, we have developed and are employing our newly developed Semi-Automated Graph-Based Assembly Curator (SAGBAC) bioinformatics data-processing pipeline. At the core of this pipeline, the novel Iterative Sequence Ends Identity Search (ISEIS) algorithm identifies contigs with identical sequences at their ends from a short-read *de novo* assembly. We assign them by blasting all contigs from the *de novo* assembly against each other. Then, an adjacency list is created. Such a list holds information on which contig ends overlap. This adjacency list is then used to construct an undirected graph, which can be visualized. From this, circular genome variants can be inferred (Figure 1). The obtained genome model is then employed to predict all possible genome configurations produced by the recombinogenic repeat pairs. This new assembly and visualization approach offers a technical solution to the assembly of the highly complex higher plant mitochondrial DNA. Its graph-based visualization provides a higher information content than the classical mini and master circle model.

MATERIAL AND METHODS

Plant material

Plant material used here was derived from the *Oenothera* germplasm collection harbored at the Max Planck Institute of Molecular Plant Physiology, Potsdam-Golm, Germany (29). *Oenothera biennis* strain suaveolens Grado (named hereafter *O. biennis*) (30) and *O. elata* ssp. *hookeri* strain johansen Standard (named hereafter *O. elata*) (31) belong to subsection *Oenothera*. *Oenothera villaricae* strain berteriana Schwemmle (*syn*: *O. berteriana* Erlangen, named hereafter *O. villaricae*) (15) is part of subsection *Munzia*. As abbreviations for the strains/species, the following code was used: berS = *O. villaricae*, suavG = *O. biennis*, johSt = *O. elata*). The line reassembles the original material used by Julius Schwemmle and Axel Brennicke. For details on the taxonomy of them, see (32-34).

Plant cultivation

Seeds were germinated in Petri dishes on wet filter paper supplemented with 0.05% (v/v) of Plant Preservative Mixture (Plant Cell Technology Store, Washington, DC, USA) at 27°C and

100-150 µE m^{-2} s^{-1} To obtain etiolated seedlings, Petri dishes were wrapped with aluminum foil immediately after germination when root tips became visible. After three days, material was harvested and frozen in liquid nitrogen. If older material was needed, plants were grown to the appropriate developmental stage in a glasshouse at 22°C and 300-400 µE m^{-2} s^{-1} in a 16 h photoperiod.

Isolation of mitochondria

Mitochondria were isolated from mature rosette leaves following a modified protocol from (35,36): First, our homogenization buffer was supplemented with 25 mM boric acid and 10 mM EGTA. Both compounds effectively liquefy viscous homogenates from *Oenothera* leaf tissue (Peter Westhoff, personal communication). While boric acid reacts with 1,2-dihydroxy groups of polysaccharides (37), EGTA specifically chelates Ca^{2+} ions. Those are often associated with gelling properties of mucilage (38). In addition, in an essential mitochondria purification step, a triple Percoll density gradient (18%, 23%, 50%) was employed.

During the isolation procedure all steps were performed at 4°C. 100 g of leaves tissue were incubated for approximately 30 min in ice water and dried using a salad spinner. Afterwards, 1 l of BoutHomX homogenization buffer (0.4 M sucrose, 50 mM Tris, 25 mM boric acid, 10 mM EGTA, 10 mM KH_2PO_4, 1% [w/v] fat free BSA, 0.1% [w/v] PVP-40, pH 7.6 with KOH, and 5 mM freshly supplemented β-mercaptoethanol) was added and leaves ground 5 x 5 sec in a razor blade grinder (Waring® Blender 8010E, Waring Commercial, New Hartford, NY, USA). The homogenate was filtered in 100 ml aliquots through two layers of mull (Verbandmull ZZ, Hartmann, Heidenheim, Germany) and one layer of Miracloth (Merck, Darmstadt, Germany), respectively. Then it was centrifuged in three 250 ml aliquots for 15 min at 5,000 x g. Chloroplast containing pellets were discarded and the supernatants centrifuged again for 20 min at 22,000 x g. Mitochondria pellets were then resuspended in 20 ml BoutWashY (0.4 M mannitol, 10 mM KH_2PO_4, 0.1% [w/v] fat free BSA, pH 7.6 with KOH) each, using a 30-cm^2 Potter homogenizer (0.1 to 0.15 mm mill chamber tolerance; Wheaton, Millville, USA). Afterwards, solutions of re-suspended mitochondria were combined, dispensed on four 50 ml centrifugation tubes and volumes adjusted to 50 ml with BoutWashY. Following a centrifugation at 3,000 x g for 5 min the supernatant was used for further purification and centrifuged at 18,000 x g for 15 min. The obtained pellets were re-suspended with a brush in all together 8 ml of 0% Gradient Medium (0.3 M sucrose, 5 mM KH_2PO_4, 0.1% fat free BSA), tubes rinsed with 2 ml 0% Gradient Medium and the two solutions combined. Still unresolved mitochondria were homogenized in a 15-cm^2 Potter homogenizer (0.1 to 0.15

mm mill chamber tolerance; Wheaton, Millville, USA). After this procedure, an additional centrifugation step at 3,000 x g for 5 min was performed. Then, the supernatant of the sample was split into halves and carefully loaded on two three-step density gradients (5 ml of 50% Percoll, 10 ml of 23% Percoll, and 5 ml of 18% Percoll in 0.3 M sucrose, 10 mM KH_2PO_4, and 0.1% fat free BSA, pH 7.6 with KOH; freshly prepared in a 30 ml Corex tube). Gradients were centrifuged with decreased acceleration at 10,000 x g for 40 min and decelerated without using of the centrifuge brake. Intact mitochondria were extracted with a pipette from the bottom of the 23%-50% interphase. For washing, the mitochondria fraction was dissolved in 50 ml BoutWashY and centrifuged four times while reducing the volume in each centrifugation step. The pellet was finally diluted in an appropriated puffer for further analyses in an Eppendorf cap. Purity of the isolated mitochondria fraction were directly assessed by confocal microscopy (employing MitoTracker and DAPI staining to visualize mitochondria and broken nuclei, respectively; chloroplasts were detected based on their autofluorescence) and western blot analyses of marker proteins for the individual genetic compartments (COXII, CF1α/β, and H3ab). Real-time PCR on isolated mtDNA (see below) with appropriated markers probes showed an enrichment of mtDNA from ~1.5% in total DNA isolations to ~95% by our protocol. For details see (20).

Mitochondrial DNA extraction

Mitochondria pellets from above were re-suspended in TENTS buffer (100 mM Tris/HCl at pH 8.0, 50 mM EDTA, 0.5 M NaCl, 0.2% [v/v] Triton X-100; 1% [w/v] SDS) and incubated for 15 min at 60°C while shaking at 400 rpm. After adding 100 µl of a 10 mg/ml RNase A solution (50 U/mg; Roche Diagnostics GmbH, Manheim, Germany) samples were incubated for 1.5 h at 37°C. Subsequently 100 µl of Proteinase K solution (10 mg/mL; Sigma-Aldrich, Steinheim, Germany) were added and samples placed over night at room temperature. Then, 630 µl phenol/chloroform/isoamyl alcohol (25:24:1) were added, probes incubated for 10 min at room temperature and then centrifuged at 18,000 x g for 10 min. After this, the supernatant was removed, 630 µl chloroform added and samples were centrifuged again at 18,000 x g for 10 min. Precipitation of mitochondrial DNA (mtDNA) was performed with 1/10 vol of 5 M NH_4-acetate and 1 vol isopropanol over night at -20°C. After centrifugation for 45 min at 20,000 x g, the pellet was washed two times with 70% v/v and 100% v/v ethanol and re-suspended in 10 µl of 5 mM Tris/HCl, pH 7.6.

Extraction of total DNA

Total high-molecular weight DNA for PacBio sequencing was obtained from etiolated seedlings with a CTAB/phenol-based method. For Southern blotting, an IGEPAL/phenol-based DNA isolation protocol was applied to plants at the early rosette stage (29). Both procedures, as well as subsequent purification of the DNA with anion-exchange columns, were published previously in (39). For PCR reactions, we used total DNA obtained with the DNeasy Plant Mini Kit (Qiagen, Hilden, Germany) applying minor modifications to the manufacture's protocol as reported in (40)**.**

Extraction of total RNA

Total RNA from the emerging fourth leaf of *O. elata* was isolated using TRIzol (Invitrogen, Thermo Fisher Scientific, Waltham, MA, USA) in a protocol adjusted to the specific needs of *Oenothera* tissue that is rich in mucilage and phenolic compounds. In contrast to the previously published RNA isolation protocols for evening primrose (41,42), the procedure described here omits silica membrane columns and allows direct precipitation of RNA from aqueous solutions. Following this protocol, depending on tissue age, 25 - 75 mg leaf material are frozen in liquid nitrogen and ground using a mixer mill. Then 800 µl of IDS buffer (120 mM Tris/HCl at pH 8.0, 120 mM EDTA at pH 8.0, 2.4% IGEPAL [v/v], 1.2% SDS [w/v], 1.2% PVP [w/v]) and 200 µl of β-mercaptoethanol are added and the sample vortexed until the powder has completely dissolved. Subsequently, the homogenizate is incubated for 10 min at 60°C under medium shaking and cell debris removed by centrifugation at room temperature. Subsequently, the supernatant is mixed with 1.0 ml of TRIzol (Invitrogen, Thermo Fisher Scientific, Waltham, MA, USA) and incubated for 10 min at 60°C under medium shaking. Than the sample is incubated on ice for 5 min and centrifuged at 12,000 g for 5 min at 4°C. The upper phase is collected, treated with chloroform:isoamyl alcohol (24:1) once, than repeatedly with acidic phenol:chlorophorm (5:1) at a pH of 4.5 until the interphase was clean, and then again with chloroform:isoamyl alcohol (24:1) twice. RNA is precipitated with 1 vol of isopropanol and washed in 75% of ethanol. To resolve the pellet in ddH_2O, RNA is incubated for 10 min at 60°C under medium shaking.

Standard polymerase chain reaction

PCR reactions were performed from total DNA using standard methods employing DreamTaq polymerase (Thermo Fisher Scientific, Waltham, MA, USA). All primers used in this work are listed in Supplementary Table 2 and were obtained from Eurofins MWG Operon (Ebersberg, Germany).

Detection of radiolabeled DNA via Southern blot

3 µg of total DNA per sample was digested over night with appropriate restriction enzymes and subsequently separated on a 1% agarose gel. DNA was then transferred by a capillary transfer to a nylon membrane (Amersham Hybond-XL, GE Healthcare, UK) using 10x SSC buffer (1.5 M NaCl, 0.15 M sodium citrate, pH 7.0). After crosslinking, the membrane was prehybridized with Church buffer (1% BSA, 1 mM EDTA, 7% SDS, 0.5 N $NaHPO_4$, pH 7.2) for 1 h at 65° C. Radiolabeled DNA probes derived from PCR products were used for detection of the corresponding DNA sequences. Labelling with ^{32}P dCTPs was performed using the Maxiscript Kit (Ambion, Darmstadt, Germany) according to the manufacturer's protocol. Radioactive probes were transferred into hybridization tubes containing the nylon membrane and Church buffer and incubated over night at 65°C. After three washing steps, once for 20 min in Wash Solution I (2x SSC, 0.1% SDS) and twice for 20 min in Wash Solution II (0.5x SSC, 0.1% SDS), the radioactive signal was detected with the Radioisotope Image Analyser Typhoon Trio (GE Healthcare, UK) after 1 day incubation. For detecting very low signals, the membranes were incubated for 24-168 h at -80°C on Amersham HyperfilmTM-ECL (GE Healthcare, UK).

Sanger sequencing

Sanger sequencing of PCR products was done at Eurofins MWG Operon (Ebersberg, Germany).

Next Generation Sequencing

Next Generation Sequencing technologies, libraries, and DNA origin used in this work are summarized in Supplementary Table 1 and detailed in the following paragraphs.

Roche 454 sequencing

454 sequencing was performed at Eurofins MWG Operon (Ebersberg, Germany). 100 ng of isolated mtDNA were pre-amplified using the GenomiPhi HY DNA Amplification Kit (GE Healthcare, Chalfont St Giles, UK). Then, samples were nebulized, emulsionPCR performed and single-end read libraries sequenced on a Roche/GS FLX Titanium platform (Roche Diagnostics GmbH, Manheim, Germany).

Illumina paired-end sequencing of isolated mtDNA from *O. elata*

Library preparation and sequencing was performed at the Max Planck Genome Centre Cologne, Germany. In brief, 100 ng mtDNA were initially fragmented by sonication (Covaris S2; Covaris, Woburn, MA, USA), followed by library preparation with NEBNext Ultra Directional DNA Library Prep Kit for Illumina (New England Biolabs, Ipswich, MA, USA). The latter included 9 cycles of PCR amplification. At all steps, quality and quantity were assessed via capillary electrophoresis (TapeStation; Agilent Technologies, Santa Clara, CA, USA) and fluorometry (Qubit; Thermo Fisher Scientific, Waltham, MA, USA). Libraries were immobilized and processed onto a flow cell with cBot (Illumina, San Diego, CA, USA) and subsequently sequenced on a HiSeq 3000 system (Illumina, San Diego, CA, USA) with 2x 150 bp paired-end reads.

Illumina paired-end sequencing of isolated mtDNA from *O. biennis* and *O. villaricae*

Creation of shotgun libraries was done by using a commercially available kit (NEBNext DNA Sample Prep Master Mix Set 1; New England Biolabs, Ipswich, MA, USA). In brief, genomic DNA was fragmented using a Covaris E210 Instrument (Covaris, Woburn, MA, USA). Than end-repair, A-tailing and ligation of indexed Illumina Adapter, agarose gel size selection and amplification was performed. The resulting fragments were cleaned up, pooled and sequenced on a HiSeq 2000 at Eurofins MWG Operon (Ebersberg, Germany) with 2x 101 bp paired-end reads.

Illumina mate pair sequencing of isolated mtDNA from *O. elata*

A mate pair library was generated from mtDNA for paired end sequencing according to the protocol of the Nextera Mate Pair Library Prep Kit (Illumina, San Diego, CA, USA). Due to the limited input DNA amount of 1 µg, the library was not additionally size-selected by e.g. Blue Pippin or SAGE Science. Sequencing-by-synthesis was performed on a HiSeq 3000 with 2x 150 bp paired-end reads at the Max Planck Genome Centre Cologne, Germany.

Illumina paired-end sequencing of ribosomal-depleted cDNA from *O. elata*

Around 1 µg DNase treated total RNA of *O. elata* was sent for sequencing to the Max Planck Institute for Molecular Genetics (Berlin, Germany). The library preparation was done using Roche KAPA RNA HyperPrep with RiboErase (Roche Diagnostics GmbH, Manheim, Germany). Sequencing was performed on an Illumina HiSeq 4000 system (Illumina, San Diego, CA, USA) with 2x 75 bp paired-end reads.

PacBio sequencing of total DNA from *O. elata*

PacBio sequencing of total DNA of etiolated seedlings of *O. elata* was performed on a PacBio RS II sequencer (Pacific Biosciences, Menlo Park, CA, USA). For this, 5 µg of high molecular weight DNA (between 20 kb and 200 kb in size; see above) were used without further fragmentation to prepare five SMRTbell libraries with PacBio SMRTbell Template Prep Kit 1 (Pacific Biosciences, Menlo Park, CA, USA) according to the manufacturer's recommendations. The libraries were additionally size-selected with BluePippin (Sage Science, Beverly, MA, USA) to enrich for molecules larger than 10, 11 or 15 kb. Recovered libraries were again damage repaired and then sequenced on a total of 138 SMRT cells with P4-C2 or P6-C4v2 chemistry and by MagBead loading on the PacBio RSII system (Pacific Biosciences, Menlo Park, CA, USA) with 360 min movie length.

De novo assembly of isolated mtDNA

Illumina paired-end reads were trimmed with SeqtrimNext v2.0.62 using the plugins “PluginIndeterminants”, “PluginLowQuality”, and “PluginSizeTrim” (https://rubygems.org/gems/seqtrimnext). Before and after trimming, read quality was evaluated with FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc). Uncalled and low-quality bases were removed. The sff_extract software v0.3.0 (https://bioinf.comav.upv.es/sff_extract) was utilized for Roche 454 data to trim 454-specific sequencing adapters, remove low quality bases and to convert sequence reads from SFF to FASTA format. Then, both pre-processed data sets (Illumina and 454) were used as input for four different *de novo* assemblers namely CLC, IDBA, MIRA, and Newbler, with Newbler operating on 454 data as input only. CLC v6.00 (part of the CLC Genomics Workbench, https://digitalinsights.qiagen.com/) and Newbler version 2.9 were executed using their graphical user interfaces with default parameters. IDBA_UD (from now on named “IDBA”) v1.1.1 was performed with Illumina Paired-end reads (-r option), 454 Roche single-end reads (-l option) and a k-mer range between 30 (--mink) and 90 (--maxk) incremented by 10 (--step) (43). MIRA 4.0.2 was run with a reduced Illumina data set (4 mio. pairs), using job mode “genome, denovo, accurate”, setting the parameters option to “-GE:not=20 -DI:trt=/scratch_local -NW:cac=warn -OUT:rtd=yes”, adjusting the template_size option to “200-450” and incorporating the ancillary xml file generated by sff_extract (44).

For the *de novo* assembler evaluation, the definition of high-confidence contigs (HCC) differs to the one for the final pipeline: Sequences of a specific *de novo* assembler were blasted against the sequences of all other three *de novo* assemblers and needed to be found by all assemblers. In detail, a sequence was called found if 90% of the contig bases were covered by any number of (sub)-sequences of another assembler contigs having an e-value lower than 1e-40. For the final pipeline, high-confidence contigs were identified based on read-mapping statistic and length criteria as follows, Illumina data were mapped against the assembled contigs using BWA v0.7.15 (45). SAMtools v1.4 was used to create, sort and index the alignment data in BAM format. Contig-wise coverage was estimated with coverageBed (46,47). HCCs were then defined as those with contig size > 1 kb and coverage > 3000; for details see Results and Figure 5). This change of the HCC definition is necessary to be independent of an inter-assembler comparison and is possible by the upstream mitochondrial isolation protocol.

Afterwards, to build the mtDNA assembly graph, the Iterative Sequence Ends Identity Search (ISEIS) pipeline, developed here, was applied to the assembled contigs. The ISEIS core algorithm takes an all against all BLASTN search result (48) of the assembled contigs (all contigs including high and low quality contigs) and filters for significant end-to-end hits (hitting at the very end of contigs; for the *de novo* assembly evaluation a range of 300 bp at the ends for MIRA, IDBA, and CLC, 600 bp for Newbler were tolerated) with at least 49 bp with the respective termini and orientation/strand combinations within the sequence alignments as mentioned in Supplementary Figure 4). By this, an adjacency list of linked contig ends is created. As entry points for an iterative breadth-first search for connected components, HCCs from above were allowed only. By starting the search on HCCs only, low-confidence contigs will be integrated into the connected component only if connected to HCCs. Otherwise, they are discarded. The obtained graph consists of contig termini as vertices connected via edges, i) when on the same contig, and ii) when connected other contigs via overlapping contig ends. As detailed in the text, from this undirected graph (from now on named “IDBA graph”) consecutive recombinogenic repeat pairs can be identified and the mitochondrial genome sequence reconstructed semi-automatically in the FASTA format. The R-package igraph was used for graph visualization (49).

Naming conventions

De novo assembly contigs generated by IDBA were renamed using the abbreviations berS (*O. villaricae*), suavG (*O. biennis*) and johSt (*O. elata*) for the different strains as introduced at the beginning of the Methods section. Contigs of IDBA were sorted by length in descending order and indexed in ascending order. PCR primers and southern blot probes were named using the indices of the corresponding contigs for which they were designed for.

Validation of mtDNA enrichment

To assess the level of nuclear and chloroplast DNA contamination in the isolated mtDNA samples, paired-end data were mapped with BWA v0.7.12 (45) against the complete set of contigs generated by the IDBA *de novo* assembly of all three species. For coverage analysis, SAMtools v1.4 was used to create, sort and index the alignment data in BAM format as well as to generate read counts per contig statistics using idxstats (46). Contigs that were part of the final IDBA graphs were defined as mitochondrial. Plastidial contigs were identified via BLASTN (48) by comparing them to the available chloroplast sequences (*O. elata*, *O. biennis* and *O. villaricae*; GenBank accessions: AJ271079.4, EU262889.2 and KX118606.1, respectively). The remaining contigs were determined as derived from the nucleus.

Pairwise BLASTN alignments to identify sequence homologies

Pairwise BLASTN alignments were performed using Circoletto v15.10.12 (50), a wrapper program executing legacy NCBI blastall v2.2.26 with default parameters (https://www.ncbi.nlm.nih.gov/books/NBK279671) and visualizing the BLASTN outcome with Circos v0.62.1 (51).

Stoichiometric analysis of recombinatoric repeats at the individual double forks

Nextera tagmentation adapters were removed from the mate pair data of the isolated *O. elata* mtDNA using Nextclip v1.3 (52). Remaining clipped reads were aligned afterwards with BWA v0.7.15 (45) against the contigs of the *O. elata* IDBA graph. SAMtools v1.4 (46) was then used to create, sort and index the alignment data in BAM format. Only pairs where both mates map to different contigs were counted and kept for further analysis. As the contigs within the IDBA graph can vary in size between hundreds of bases and many dozens of kilobases, the mate pair fragments can span more than one contig by their large insert size. To overcome this issue, so-called contig chains were defined by extending the contig-repeat-contig (CRC) at both ends till the next occurring recombinogenic repeat pair (RRP) within the graph (Supplementary Figure 5). With this approach it is possible to count the number of reads spanning one of the four CRC combinations for each identified “double fork” within the IDBA graph.

In addition to the stoichiometric analysis of the mate pair data set, PacBio long-reads generated from a total DNA library from *O. elata* (see above) were taken to calculate the stoichiometric distribution among the different CRC combinations. PacBio-specific bax files containing the information of the polymerase reads were converted with bax2bam to generate a SMRT Link pipeline v5 (Pacific Biosciences, Menlo Park, CA, USA) or higher compatible input. PacBio circular consensus sequences (CCS) were called with pbcss from the polymerase reads to correct for PacBio-specific errors (mostly 1 bp indels) with the following relaxed parameters: --minPredictedAccuracy 0.75, --maxDropFraction 0.5, --minPasses 0. CCS bam files were then converted to FASTA format using bam2fasta to obtain a BLAST compatible sequence format. The programs bax2bam, pbccs and bam2fasta are part of SMRT Link v5.0.1 program suite (Pacific Biosciences, Menlo Park, CA, USA) used in this approach. Afterwards, all CCS reads were blasted against all sequences included in the IDBA graph of *O. elata* employing NCBI blastall v2.2.26 (https://www.ncbi.nlm.nih.gov/books/NBK279671) with default parameters. Three different data sets were filtered from the overall BLASTN outcome remaining only hits longer than 100, 170 and 180 bp, respectively, for further analysis. These different subsets of BLASTN hits is necessary to reduce short hit contamination (100 bp) and to get an estimator of cross-mappings to contig ends of the other CRC combinations (170 bp and 180 bp) in the tab-delimited output table.

To identify and count CCS reads, which are consistent with our model of the IDBA graph and, in particular, fit our predicted CRCs, three steps were essential: (1) As a CCS read can be hit by more than one IDBA-graph contig, resulting in an unsorted multi-row entry, the BLASTN outcome was sorted by the CCS read identifier and the start position of the query sequence where the hit was positioned; bash command: sort -k1,1 -k7,7n blastfile >sorted.blastfile. (2) To facilitate a straightforward search for each CRC string (e.g. “contig_1,repeat_4,contig_2”) BEDTools groupBy v2.20.0 (47) was applied to group the BLASTN outcome by the identifier of each CCS read. By this, the multi-row entry can be condensed into one row by summarizing the following columns with a specific operation in brackets: hit id (collapse), hit id (count_distinct), alignment length (sum) and length of IDBA graph contig (distinct); bash command: groupBy -i sorted.blastfile -g 1 -c 2,2,4,13 -o collapse, count_distinct,sum,distinct >collapsed.blastfile. (3) Finally, a custom Perl script was implemented to collapse identical neighboured IDBA-graph contig identifiers, as both, the “collapse” and the “distinct” operation of BedTools cannot conduct this task. “collapse” on the one hand just concatenates the identifiers (even if neighboring identifiers are the identical) whereas “distinct” does not preserve the order of contigs, which is essential for CRC identification in our approach. Lastly, the percentage of covered CCS sequence by the IDBA graph contigs was calculated. Only those CCSs were kept whose sequences fully align to the

IDBA graph contigs, and, to deal with PacBio sequencing errors, have at least 95% identity and are 5 kb long. CCS reads considered to be of nuclear origin we excluded from further analysis (Supplementary Figure 6). The final, condensed output was then screened by Linux command “grep” for the comma-separated CRC-identifier string (example see above) and counted.

Mitochondrial genome annotation, visualization

Mitochondrial genome sequences were annotated with a complex annotation scheme, illustrated in Supplementary Figure 3 and organized in three stages: (a) Initial data generation, (b) Filtering and cross validation between the generated datasets and (c) Merging filtered and validated data. Tools, which are available at chlorobox.mpimp-golm.mpg.de, assisted in the process of annotating organellar genomes (GeSeq), converting between different file formats (GBSON, a GenBank JSON converter), drawing organelle genome maps (OGDraw) as well as preparing GenBank files for NCBI submission (GB2sequin).

The first stage of the annotation scheme corresponded to the creation of the different datasets which focus on different genome feature types (protein coding genes, pseudogenes, tRNAs, rRNAs as well as open reading frames) from distinct organelle origin (plastid and mitochondrion). First, GeSeq v1.82 was applied on the mitochondrial input sequences in two different ways (53). For the annotation of the mitochondrial genes (GeSeq Mt run), six land plant species were selected covering a variety of angiosperms from within the rosids clade (*Arabidopsis thaliana* NC_037304.1, *Geranium maderense* NC_027000.1 and *Vitis vinifera* NC_012119.1), including the only two species of the myrtales order from which the mitochondrial genomes are known (*Eucalyptus grandis* NC_040010.1, *Lagerstroemia indica* NC_035616.1) as well as a gymnosperm (*Cycas taitungensis* NC_010303.1) as an outgroup. In the same GeSeq run, sequences of the recombinogenic repeats were uploaded as FASTA Nucleotide and tRNA *de novo* prediction with tRNAscan-SE v2.0.5 (Lowe and Chan, 2016) was activated. Plastidial pseudogenes were identified in a second run of GeSeq (GeSeq Pt run) using the respective plastidial genome of the three investigated *Oenothera* species (*O. villaricae* KX118606.1, *O. biennis* EU262889.2 and *O. elata* AJ271079.4) as database. rRNAs were identified by a simple BLASTN search with default parameters but allowing only the best hit using the NCBI entries X61277.1 (rrn5 and rrn18) as well as X02559.1 (rrn26) as queries. An rRNA-depleted Illumina paired-end RNA-seq dataset of *O. elata* was used to construct an RNA editome, to evaluate all exon-exon boundaries of the GeSeq-based gene predictions as well as to generate an expression profile. For that the RNA-seq data were mapped against the final mitochondrial sequence of this species using STAR v2.7.0a (54). To generate an unbiased mapping result, the previously GeSeq-generated annotation was not included during the genome indexing step of STAR (--genomeSAindexNbases 8 --genomeChrBinNbits 18) nor used in the mapping step itself. Nevertheless the coverage of each GeSeq-determined exonic position was computed applying samtools mpileup with the -l option on the exon entries within the GeSeq annotation file only. In parallel, instead of counts per gene, the mitochondrial genome was segmented into 250bp pieces using windowMaker from the bedtools suite (47). Afterwards, coverageBed was applied on the generated 250bp segments bed file and the alignment bam file to count reads per segment. Single nucleotide polymorphisms (SNPs) were called using freebayes v1.0.2 with default parameters, annotated with snpEff v4.3k (55) and filtered exclusively for C>T and G>A SNPs. Open reading frames were predicted using ORFfinder v0.4.3 with default parameters but allowing only ATG as start codon (-s 0).

In the second stage of the annotation scheme, the focus lay on filtering of and on cross-validation between the generated datasets. For single-exonic genes predicted open-reading frames were intersected with the CDS entries of the GeSeq Mt run using intersectBed allowing only those intersections which have the same start and end positions (-f 1 -F 1). Genes that fulfill these criteria were instantly tagged as verified protein-coding genes. For all other genes, where the ORFs are longer or shorter than the BLAT hits, these ORFs were fed into a BLASTP search on NCBI and in parallel examined within the IGV viewer v2.5.3, for example, if stop codons were present within the BLAT hit regions (56). This visualization approach was also used for the evaluation of the exon-exon boundaries of the multi-exonic genes taking the mapped RNA-seq data into account. For the trans-spliced genes, additionally available data on NCBI as well as the information within the corresponding papers were considered as follows: *nad1* (AH003143.2, (12)), *nad2* (AH003694.2, (57)) and *nad5* (exon a & b: X07566, exon c: X60046.1, exon d and e: X6004691, (58)). Because of the shortness of exon c of nad5 this sequence was separately searched on the mitochondrial genomes using BLASTN v0.2.26 with an e-value cutoff of 1e-03. Within the tRNAscanSE-predicted tRNAs, only those were considered to be true which have a score equal or greater than 30, have no undetermined anti-codon (trnNull-NNN) and have not a fragmented BLAT hit by GeSeq. To distinguish between plastidial and mitochondrial originated tRNAs, tRNA entries of both GeSeq runs were intersected via intersectBed. Those intersections, where the plastidial locus is covered 100% by the BLAT search itself, were defined to be derived from plastidial origin and all other tRNAs are set to be of mitochondrial origin. A similar approach was used to discriminate between mitochondrial and plastidial pseudogenes. Only those BLAT hits from the GeSeq Pt runs that do not overlap with genes from the GeSeq Mt run or are shorter than there corresponding BLAT hits from the GeSeq Mt run were designated to be plastidial pseudogenes. Intersected BLAT hits, where the BLAT hits coverage is low in both GeSeq runs, were tagged as mitochondrial pseudogenes.

In the third stage of the annotation pipeline, filtered and cross-validated datasets were collected and re-processed in different ways: All decisions made by all the different investigations within both GeSeq runs were collected and used as input in a custom-developed Perl script in order to manipulate the initial json files that were generated during the GeSeq runs. Also for the RNA editome, a custom perl script was written to extract all relevant information from the VCF file created by snpEff to generate the Supplementary table XY as well as a json file. All three json files (GeSeq Mt and Pt run as well as RNA editome) were combined to serve as input for the GBSON to GenBank converting tool.

The finalized annotation was used to create a mitochondrial genome map using OGDRAW v1.3 (59,60) with a user-defined configuration XML file to include identified plastidial pseudogenes and important repetitive elements. In parallel, the final GenBank annotation file was again pre-processed using GB2sequin (61) for the NCBI submission.

RESULTS

The metagenome assembler IDBA works best for mtDNA assemblies

To reduce the complexity of the input DNA, we generated our sequencing data from highly pure mitochondrial DNA (mtDNA) enriched by cell fractionation (Material and Methods, and below). The intent was to reduce possible contaminations by nuclear mitochondrial DNA (NUMTs;(62)), i.e. segments of mitochondrial genomes, translocated in the nucleus that are commonly present throughout the plant kingdom. Two Next Generation Sequencing (NGS) datasets were generated, Illumina paired-end and Roche 454 single-end. The sequencing data of *O. villaricae* were used to perform a *de novo* assembly evaluation of four different assemblers: CLC (Qiagen), IDBA (43), MIRA (44), and Newbler (Roche). Those were chosen to cover a wide range of implemented algorithms (OLC (Newbler) and De Bruijn graph (IDBA)), data input options (stand-alone (Newbler) or hybrid assemblers (CLC, IDBA and MIRA)), availability (open source (MIRA and IDBA) and commercial software (CLC and Newbler)), and area of application (i.e. meta-genomics, IDBA). The raw assembly outcome ranged from 14 contigs (Newbler; cumulative genome size of 424,082 bp) to 847 contigs (CLC; 969,165 bp), 883 contigs (IDBA, 1,057,763 bp), and up to 1540 contigs (MIRA; 1,243,023 bp).

To assess, which assembler creates valid contigs, i.e. contigs whose sequences overlap at their ends, an all against all sequence alignment was generated with BLASTN. The BLAST output then served as input for our ISEIS algorithm that represents the first step of the SAGBAC pipeline: the starting point for the first iteration were so-called high confidence contigs (HCC), which (for the *O. villaricae* data set) we defined as contigs produced by all assemblers. We identified these contigs in the BLAST-result table and compared their ends to the ends of the remaining ones to find overlaps. In the second iteration, we then searched for overlaps of these new contigs with the contigs left in the BLAST table. The iterative process proceeded until no overlaps to any new contigs were found. As a result, an adjacency list of linked contig ends was created, which can be visualized as a graph with appropriate R packages such as igraph (49). It should be emphasized that the only exit criterion for the data-driven ISEIS algorithm is the absence of any new overlaps between contig ends. Thus, the outcome can be a linear/chromosome-like *or* a circular graph. The resulting graphs coming from the four *de novo* assembly raw outputs differ drastically (Supplementary Figure 1). CLC yielded almost exclusively isolated contigs (HCCs without any sequence overlap with other contigs). Using Newbler, all contigs were connected at least once, although some contigs' ends remained unconnected. MIRA harbors most contigs within its graph. However, it yielded a large number of unconnected contig ends, leading to a complexity that can no longer be inspected by eye. Only IDBA, conceptually designed to assemble meta-genomes, was able to generate a set of contigs that was fully connected to each other. Strikingly, it produced a circular graph. We, therefore, continued to work with IDBA and further generated the IDBA assemblies for the other two *Oenothera* species *O. biennis* and *O. elata*.

Oenothera mtDNA sequence reads assemble into circular graphs

Comparing the raw assemblies of the three *Oenothera* species, it is conspicuous that the number of contigs ranges from 883 for *O. villaricae,* to 4,381 for *O. biennis,* to 20,317 contigs for *O. elata*. The sum of all contig sizes ranged from 1,057,763 bp, to 2,346,337, and 18,986,762 bp, respectively (Table 1). In both assembly metrics, the numbers differed by one order of magnitude for *O. elata*, when compared to *O. villaricae* or *O. biennis*. As discussed below, this is a result of a much higher contamination of the initial DNA, containing segments of the nuclear and chloroplast genomes. Unlike the case for *O. villaricae*, the initial IDBA assemblies of *O. elata* and *O. biennis*, (this time HCCs were defined by read coverage (3000x) and contig length (1 kb); see Material and Methods for details), no circular graph structure was obtained. We therefore compared the IDBA graphs to their respective plastid genomes available from

Table 1. Descriptive statistics for the assembly, annotation, and RNA editing outcomes. [1] RNA editome available for *O. elata* only, as only for that species, rRNA-depleted Illumina data were available. [a] All three mitochondrial rRNAs are localized on the large repeat, which was identified in *O. elata*

	Metric	***O. villaricae***	***O. biennis***	***O. elata***
Assembly	raw assembly [# contigs]	883	4,381	20,317
	raw assembly [bp]	1,057,763	2,346,377	18,986,762
	IDBA graph [# contigs]	21	38	45
	IDBA graph [size]	408,260	419,446	418,451
	Reconstructed Master circle [bp]	408,744	424,132	449,216
	GC content	46.70%	46.70%	46.70%
Annotation	Protein coding genes	35		
	tRNA	14 (Mt origin) 7 (Pt origin)		
	rRNA	3 (3 unique)	6[a] (3 unique)	
	Coding genes with introns	8		
	of which are trans-spliced	3		
RNA editing[1]	# RNA editing sites	NA	681	
	on genes	NA	511	
	Non-synonymous	NA	472	
	of which have gained a premature stop codon	NA	3	
	on double edited codons	NA	22(11)	
	synonymous	NA	39	

NCBI (*O. elata*, *O. biennis* and *O. villaricae*; GenBank accessions: AJ271079.4, EU262889.2 and KX118606.1, respectively), as well as to the contigs that are part of the IDBA graph of *O. villaricae*. From this it was obvious, that plastidial contigs were integrated/connected within/to the two graphs and that some misassemblies have occurred. Curation of the IDBA graphs included (1) removal of plastidial subgraphs, (2) removal of small and low coverage contigs, and (3) correcting obvious misassemblies and falsely connected contigs by removing affected edges from the graph. The filtering steps are illustrated in Figure 2. After curation, the number of contigs were 21 for *O. villaricae*, 38 for *O. biennis*, and 45 for *O. elata*, respectively, with respective cumulative length of 408,260 bp, 424,132 bp, and 418,451 bp, respectively, in line with published size estimations of the *Oenothera* mitochondrial genome of approximately 400 kb (63).

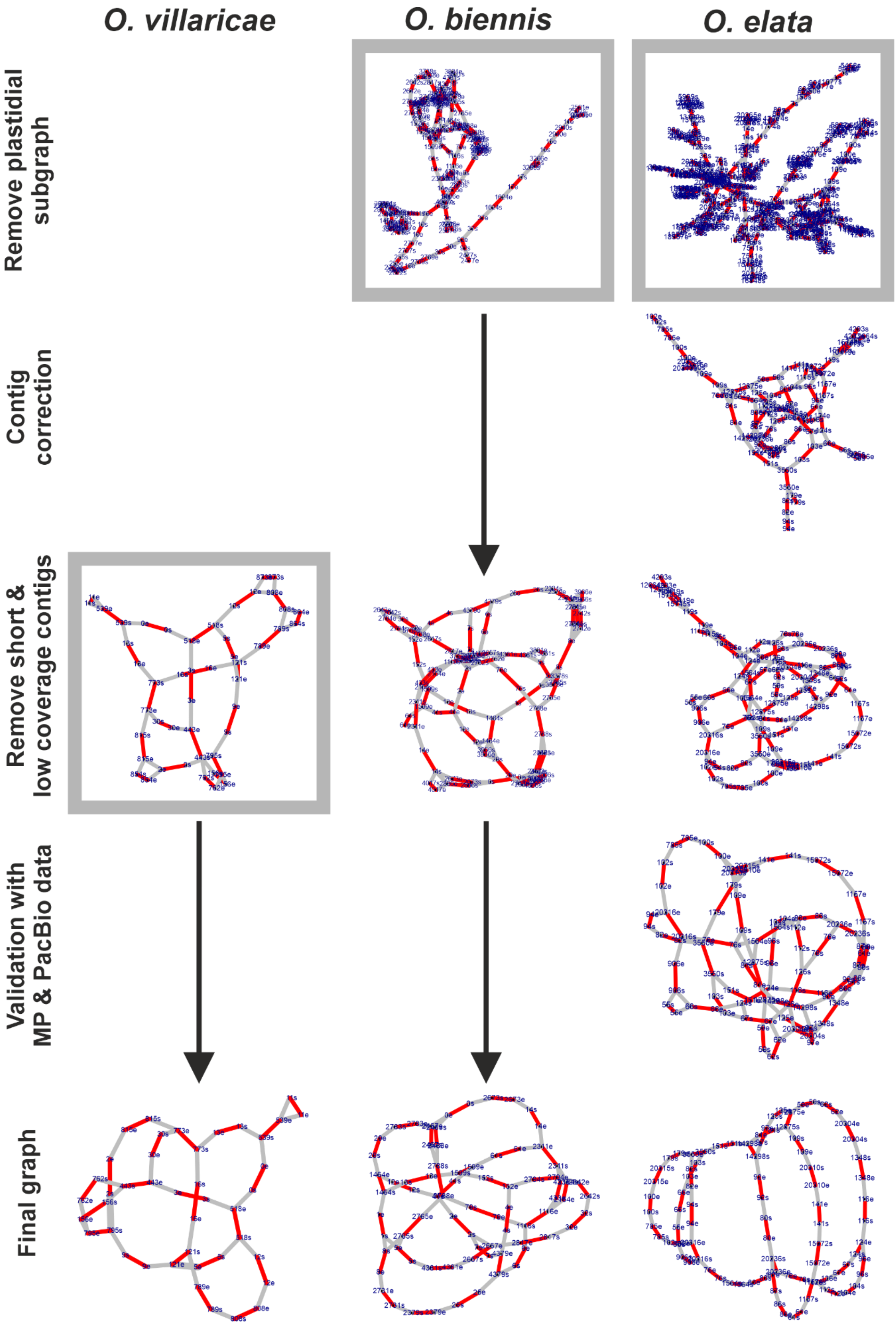

O. villaricae
O. biennis
O. elata
Remove plastidial subgraph
Contig correction
Remove short & low coverage contigs
Validation with MP & PacBio data
Final graph

Figure 2. Graph curation pipeline from raw to final graphs for the IDBA assemblies of three *Oenothera* species. For each IDBA assembly, an adjacency list is generated by the ISEIS algorithm that can be used for the construction of an undirected graph. Each contig is defined by two vertices (s=start, e=end) linked by a red edge. An overlapping event between two different contig ends is represented by a gray edge. Curation steps conducted with ISEIS are as follows: (i) Remove plastidial subgraph: Contigs, identified by a BLASTN search using available *Oenothera* plastomes. (ii) Contig correction: Check for misassemblies, check unlinked contigs for any plausible reasons, concatenate contigs for comparison of RRPs among species. (iii) Remove small and low coverage contigs: Contigs, which are so small that they are fully-covered by their neighboring contigs or have a low coverage and therefore likely originated from the nucleus. (iv) Correction of falsely connected contigs with Illumina Mate-Pair and PacBio data: Grey edges in *O. elata* were removed, where the Illumina Mate Pair and PacBio data do not corroborate the connectivity of the participating contigs. Gray boxes show the start point at which each raw IDBA graph is starting at within the curation pipeline. Black arrows illustrate the skipped steps, which are not necessary for that species.

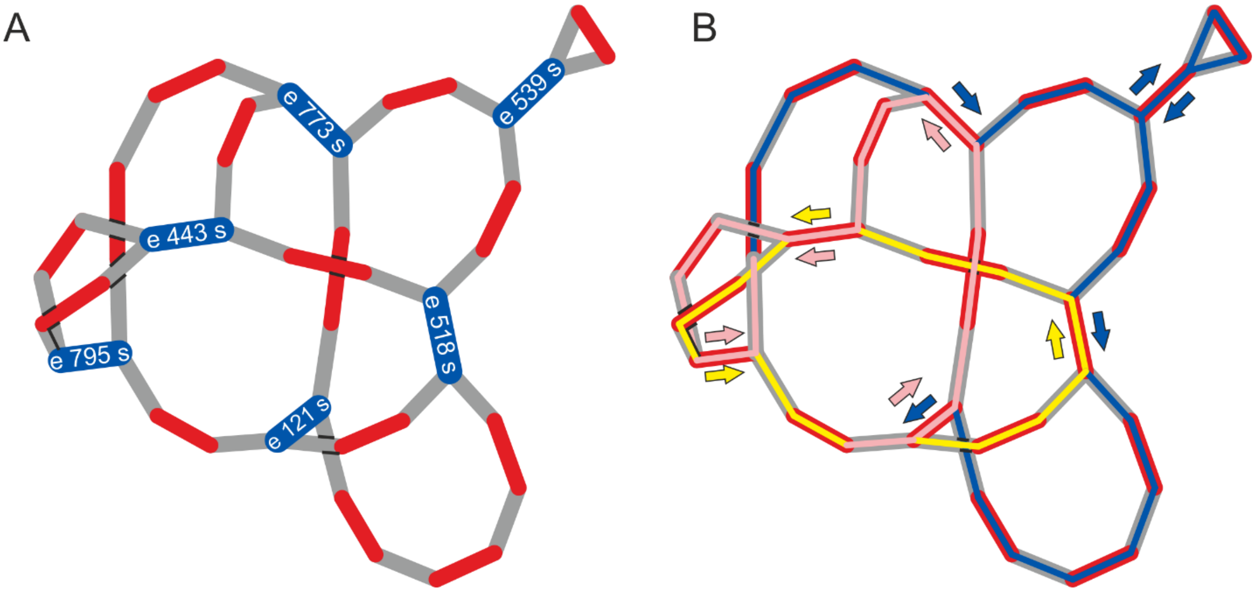


Figure 3. Identification of recombinogenic repeat pairs and master circle reconstruction in *O. villaricae*. (A) Identification of recombinogenic repeat pairs: Each contig is defined by two vertices (s=start, e=end) that are linked by a red edge within the graph. An overlapping event between two different contig ends is represented by a grey edge. Highlighted blue are those contigs that are embedded in a "double fork"-like structure (Figure 1c) and thus represent the recombinogenic repeat pairs. (B) Manual detection of paths through the graph in three steps following the arrows and using two rules: First, end the path at the same contig as you started and, second, each repeat should be traversed only twice. Path: Starting at the star-tagged contig following the blue path, then the yellow and finishing with the pink path ending at the second vertex of the star-tagged contig.

Identification and comparison of repetitive elements in the mitochondrial genome of *Oenothera* species

Using the obtained graphs as starting point, the recombinogenic repeat pairs (RRPs) could now easily be identified from the graph by eye, as illustrated exemplarily in Figure 3a for *O. villaricae*. Their number varies between the three species. *Oenothera villaricae* and *O. biennis* harbor six, while *O. elata* contains five repeat pairs (Table 2). The RRPs can be divided by their length distribution into two groups: intermediate-size repeats ranging from 171 bp to 479 bp and long-size repeats ranging between 825 bp to 1625 bp. In the three species, the contigs

Table 2. Recombinogenic repeat pair sets of three *Oenothera* species. Listed are all recombinogenic repeat pairs (RRP) identified in three different *Oenothera* species and their occurrence within the three species. Black: sequence exists as recombinogenic repeat pair within in a species included in the IDBA graph; blue: sequence remained only once in the IDBA-graph as a singleton contig; Bold rows: shared repeats among all three *Oenothera* species. johSt_3550 represents the *nad6* gene locus (green), which is split into three contigs in *O. villaricae* and *O. biennis* as an alternative contig is present in these two species (Figure 4). For comparison reasons, the contig between the two RRPs is added (purple) including the overall length of the *nad6* locus; LSR=Long-size repeat; ISR=Intermediate-size repeat. p.o.c.= sequence is present only once, which is part of another contig. [a,b] Note, contig suavG_4381 and johSt_1348 originally consist of four and two contigs, respectively, that were concatenated to be comparable to the other two respective species in this table. [c], listed are genes, which are extending into the RRP, or the RRP is, as a whole, part of the gene.

Repeat type	*O. villaricae*	length	*O. biennis*	length	*O. elata*	length	gene[c]
LSR	berS_121	1337	suavG_4381[a]	1316	johSt_1348[b]	1352	*atp9*
	p.o.c.		p.o.c.		johSt_1564	1151	*atp6*
LSR	p.o.c.		suavG_152	1625	p.o.c.		
ISR[1]/	**berS_443[1]**	**475**	**suavG_1116[1]**	**479**			
LSR[2]	**+berS_762 (305)**	**821**	**+suavG_2154 (285)**	**825**	**johSt_3550[2]**	**825**	*nad6*
	berS_795[1]	**239**	**suavG_2311[1]**	**239**			
ISR	berS_518	432	p.o.c.		p.o.c.		
ISR	**berS_539**	**421**	**suavG_1464**	**397**	**johSt_12875**	**397**	
ISR	berS_773	300	suavG_2483 suavG_2457	201 206	johSt_20315 johSt_20310	201 206	*rpl2*
ISR	p.o.c.		suavG_1599	370	johSt_14298	370	
ISR	p.o.c.		suavG_4379	260	johSt_20236	261	*atp1*
ISR	p.o.c.		p.o.c.		johSt_20316	171	*nad5*
sum	*1 LSR* *+ 5 ISR*		*1 LSR* *+ 5 ISR*		*1 LSR* *+ 4 ISR*		

berS_121, suavG_152, and johSt_3550 represent the long-size repeats. Those are considered to confer frequent and reversible recombination events that lead to a simultaneous presence of master- and smaller sub-circles (64). Each of the three *Oenothera* species has its own unique long-size repeat. Furthermore, we found eight intermediate-size repeats of which three are shared among all species (Table 2, marked bold). Mitochondrial intermediate-size repeats have been found to recombine infrequently but are believed to be part of the break-induced replication pathway (BIR) and can lead to further complexity of mtDNA (65). The long-size repeat johSt_3550 of *O. elata* harbors the *nad6* (NADH dehydrogenase subunit 6) gene. Interestingly, in *O. villaricae* and *O. biennis* gene and long-size repeat are split into three contigs (berS_443, berS_762, berS_795 and suavG_1116, suavG_2154, suavG_2311, respectively). Here, genome configurations are possible that incorporate an alternative sequence (berS_156 or suavG_433) into the *nad6* gene by replacing the contigs berS_795 or suavG_2154, respectively (see Figure 4 for details). Interestingly, the loss of the alternative sequence creates the long-size repeat of *O. elata*. The next step after RRP identification was the reconstruction of the master circle.

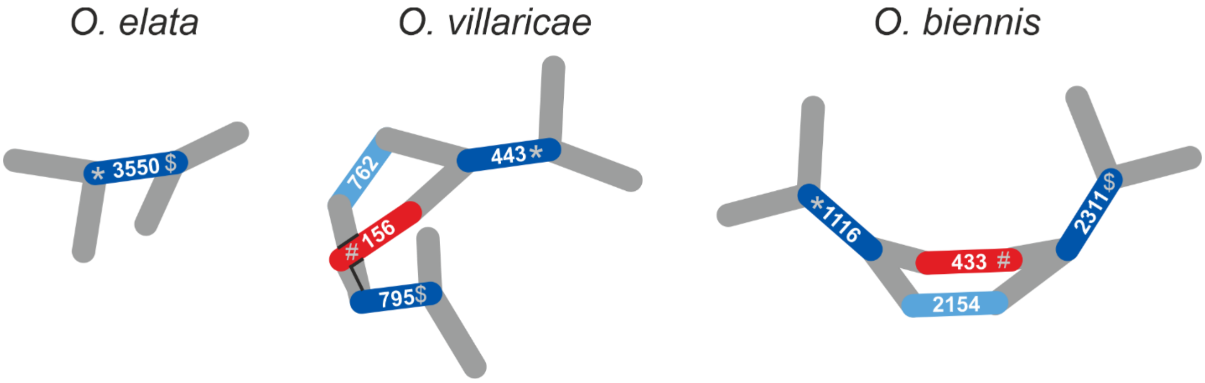


Figure 4. Identification of an alternative *nad6* locus. Occurrence of an alternative sequence (contig) within the *nad6* gene locus, which is directly apparent from the IDBA graph itself. Visualized are the contigs making up the *nad6* gene locus within the final, curated IDBA graphs of the investigated *Oenothera* species. The fully functional *nad6* gene consists of the contigs in blue (representing the RRPs) and light blue, whereas the contig in red represents an alternative sequence which is present in *O. villaricae* and *O. biennis*, but is absent in *O. elata*, where only the native *nad6* gene as RRP exists. (*) indicates shared start codon of both *nad6* isoforms, whereas ($) and (#) mark different stop codons of the fully functional and alternative *nad6* isoform, respectively.

Master circle reconstruction and genome comparison

To obtain a master circle for the mitochondrial genome from the curated IDBA graph, two rules for traversing the graph were set initially. (i) While imposing circularity, start and termination vertex (a vertex is a node within a graph that connects objects) need to be identical, but should be traversed from different edges. (ii) Each repeat should be traversed at least twice. This can be easily realized for the graph of *O. villaricae* (Figure 3b). However, the graphs of *O. biennis* and *O. elata* appeared to be more complex. To close the path for the *O. biennis* graph, a large contig needs to be traversed twice (suavG_41; 4,703 bp). It was not identified as a repeat by the ISEIS algorithm, because it is not embedded within a double, but within a single fork. In addition, one repeat (suavG_1599) was traversed three times. For the graph of *O. elata*, it was necessary to pass a stretch of the three contigs twice (johSt_124, johSt_67, johSt_126). In all cases, involved contigs have the highest read coverage when mapping the initial Illumina paired-end reads to them that were used for the *de novo* assembly. We are therefore confident that the proposed genome models are valid (Supplementary Table 3).

The sizes of the three reconstructed mitochondrial genomes range from 409 to 449 kb with an average GC content of 46.7% (Table 1) for all three. This places them among the top-20 GC-rich species of the 284 mitochondrial land plant genomes published thus far (NCBI as of September 2021). A sequence alignment of all three genomes using AliTV (66) shows nearly 100% sequence identity between *O. biennis* and *O. elata*, not taking structural variation into account, however. Comparison of *O. villaricae* with *O. elata* reveals that 12 kb (3.0%) of *O. elata* are unique to it, whereas 27 kb (6.0%) of *O. villaricae* are not present in *O. elata* (Supplementary Figure 2). Next, we validated the proposed structure by independent wet lab and bioinformatics analyses.

Read coverage of the IDBA graph supports mitochondrial origin

As previously mentioned, the NGS libraries used for the assemblies were derived from highly pure mtDNA obtained by cell fractionation. Due to an improved triple Percoll sucrose gradient the employed mitochondrial fractions were largely depleted from chloroplast and nuclear DNA contaminations (Dotzek, 2016; and Material and Methods for details). Consequently, the contigs integrated into the final IDBA graphs have one or even two orders of magnitude higher coverage than most of the remaining contigs of the assemblies, which, therefore, likely represent the still present residual contamination that was not integrated into the graph. For the chloroplast genomes, this could be verified by BLAST analyses against the corresponding plastome sequence present in NCBI (AJ271079.4, EU262889.2, and KX118606.1; also see above). All other contigs, neither found in the IDBA graph nor mapped against the plastid genome, were considered nuclear DNA contamination (Figure 5).

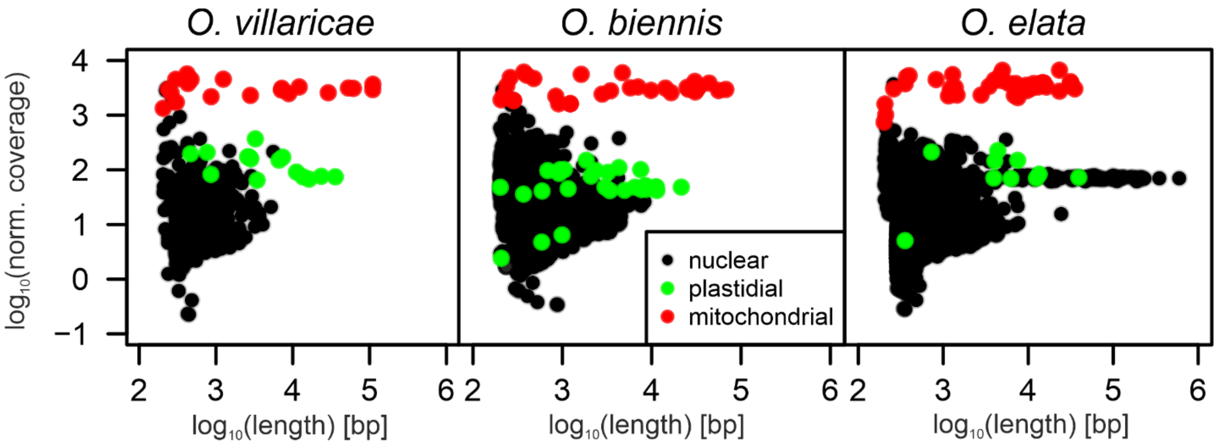


Figure 5. Length-normalized coverage distribution among IDBA contigs for three *Oenothera* species. Shown are the contig-length-normalized coverage distributions for the IDBA contigs after mapping the Illumina paired-end reads to them and as a function of contig length, Black=nuclear; green=plastidial; red=mitochondrial.

The IDBA graph contains the canonical set of mtDNA genes – annotation homology search

Following mitochondrial genome reconstruction we employed GeSeq (53) for sequence annotation of all genomes. Besides determination of coding regions, this proved an important step to validate genome completeness. For this, a complex multi-staged annotation scheme was applied (Supplementary Figure 3) and annotations were visualized with OGDRAW (60) (Figure 6). First, we could show an even distribution of genes over the whole sequence of all three genomes and on both strands. A set of conserved genes, considered essential for the mitochondria genomes of most land plants, was found in all three investigated *Oenothera* species. Specifically, genes encoding the oxidative phosphorylation complexes (*nad1-6, 4L, 7, 9*; *sdh4*; *cob*; *cox1-3*; *atp1, 4, 6, 8, 9*), all three rRNA genes (*rrn5, rrn18,* and *rrn26*), and several ribosomal proteins (*rps1, 3, 4, 13, 14, 19* and *rpl2, 5, 10, 16*) were identified. Also, *mttB* (a membrane transport protein), *matR* (maturase), *ccm* (cytochrome c biogenesis), and additionally, 16 plastidial pseudogenes, psaA, psbB ,psbC, psbD, psbE, psbF, rbcL, rps4, rps11, rps12, rps14, ycf2, ycf3 and all three plastidial rRNAs, were detected. Besides *nad1* and *nad2*, also *nad5* is trans-spliced (12,57,58). Twenty-four different tRNA genes were initially identified with tRNAscan-SE v2.0.5 (67). Four of them were identified as false positives reflected by their corresponding fragmented-only BLAT hits (see below); seven were of plastidic origin and translocated from this genome. The remaining 13 are of true mitochondrial

Table 3. Gene content of all three investigated Oenothera mitochondrial genomes. Genes with multiple exons are denoted with the number of exons shown in parenthesis, trans-spliced genes are indicated with an *. [r] These genes are present in two copies in *O. elata* as they are located on the large repeat identified. [a] *rpl6* carries an internal stop codon but can be transcribed using an alternative start codon GUG downstream from the ATG start codon. [b] *mttB* lacks its normal start codon but an alternative GUG start codon can be created via RNA editing.

Gene set	**Members of the gene set**				
Complex I	*nad1* (5*)	*nad2* (5*)	*nad3*	*nad4* (4)	*nad4L*
	nad5 (5*)	*nad6*	*nad7* (5)	*nad9*	
Complex II	*sdh4*				
Complex III	*cob*				
Complex IV	*cox1*	*cox2*	*cox3*		
Complex V	*atp1*	*atp4*	*atp6*	*atp8*	*atp9*
Cytochrome C biogenesis	*ccmB*	*ccmC*	*ccmFC* (2)	*ccmFN*	
Ribosomal large subunits	*rpl2* (2)	*rpl5*	*rpl10*	*rpl16*[a]	
Ribosomal small subunits	*rps1*	*rps3*	*rps4*	*rps13*	*rps14*
	rps19				
Intron maturase	*matR*				
Protein translocase	*mttB*[b]				
Plastidial pseudogenes	*4.5S rRNA*	*16S rRNA*	*23S rRNA*	*rbcL*	*psaA*
	psbB	*psbC*	*psbD*	*psbE*	*psbF*
	rps4	*rps11*	*rps12*	*rps14*	*ycf2*
	ycf4				
rRNA	*5S rRNA (x3)*[r]	*18S rRNA (x2)*[r]	*26 rRNA (x2)*[r]		
Mt-originated tRNAs	*tRNA-Cys* (GCA)	*tRNA-Gln* (TTG)	*tRNA-Glu* (TTC)	*tRNA-Gly* (GCC)	*tRNA-Ile* (CAT)
	tRNA-Lys (TTT)	*tRNA-fMet* (CAT) (x2)[r]	*tRNA-Phe* (GAA)	*tRNA-Pro* (TGG)	*tRNA-Ser* (GCT)
	tRNA-Ser (TGA)	*tRNA-Tyr* (GTA)			
Pt-originated tRNAs	*tRNA-Asn* (GTT)	*tRNA-Asp* (GTC)	*tRNA-His* (GTG)	*tRNA-Met* (CAU)	*tRNA-Ser* (GGA)
	tRNA-Trp (CCA)	*tRNA-Val* (GAC)			

origin, which could be confirmed by full-length BLAT hits, when blasted against the six chosen plant mitochondrial genomes (for details see material and methods). A list of all genes present on the *Oenothera* mitochondrial genomes is provided in Table 3.

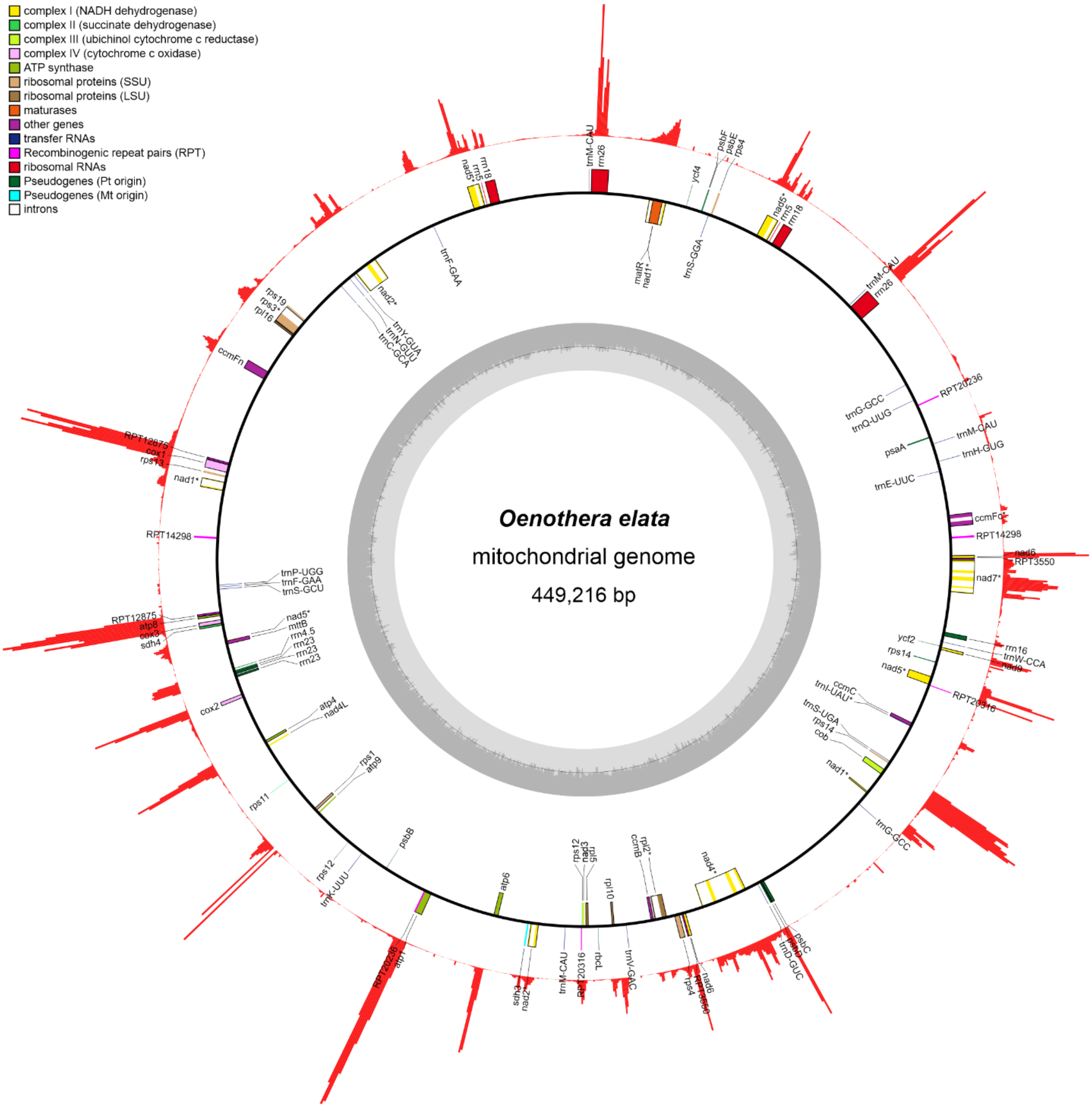


Figure 6. Gene map and gene expression atlas of the *O. elata* mitochondrion. A circular representation of the mitochondrial genome of *O. elata* is shown, generated by OGDRAW. The input sequence starts at "3 o'clock" moving counter-clockwise. All types of essential mitochondrial complexes, the identified recombinogenic repeat pairs as well as mitochondrial and plastidial pseudogenes are described in the legend along with their color code. Outer circle histogram: Gene expression along the mitochondrion (log2 coverage for each 250bp genome segment).

The IDBA graph contains the canonical set of mtDNA genes – expressed genes and RNA editome

In parallel to the annotation based on homology searches or tRNA prediction, an rRNA-depleted Illumina paired-end RNA-seq dataset was used to create an expression atlas (read coverage profile) along the mitochondrial genome of *O. elata* (Figure 6). When compared to the GeSeq annotation, all 33,779 unique exonic positions of protein coding genes were covered by the RNA-seq data. This supports the high accuracy of the homology-search based annotation, as well as that of the assembly.

Besides the expression profile, additionally, an RNA editome was generated (Supplementary Table 4). In total, 681 RNA editing sites were identified in *O. elata*, from which 511 are located in protein-coding genes. 472 non-synonymous editing sites led to non-synonymous changes of amino acids, including three non-sense mutations (gain of stop codons) (*ccmF*CeU1315R*, *atp9*eU223R* and *atp6*eU844Q*). Thirty-nine synonymous RNA editing sites and 11 double-edited codons were observed.

IDBA graph contains the earlier published *Oenothera* mtDNA sequences

As another checkpoint of our analyses we checked to what extent already published sequences of the *Oenothera* mtDNA were found within the IDBA graph. Already very early, particular efforts were invested into the mitochondrial genome of *O. villaricae* (68) and first mitochondrial genes were identified therein (69,70). Therefore, a comparison between these sequences and our IDBA contigs of *O. villaricae* was performed. After downloading available mitochondrial sequences for *O. villaricae* from NCBI (nuccore database entries for taxonomic ids 3941 and 3950), a BLASTN search was performed to compare the IDBA contigs of *O. villaricae* to the 44 retrieved sequences. Strikingly, all of them could be mapped partly or completely to the contig sequences, which are included in the IDBA graph (Figure 7). Contigs berS_2 (110,274 bp), berS_3 (109,856 bp), and berS_5 (61,574 bp) were covered most, which are the largest contigs within the IDBA *de novo* assembly. The sequences AH003143.2 and AH003694.2 map to more than one locus and on different contigs, which is not surprising as they both represent the trans-spliced *nad1* and *nad2* genes, for which the exonic sequences were concatenated by stretches of 100-bp-long Ns (12,57). Additionally, four of the six repeat pairs, identified in *O. villaricae,* could be found among the NCBI sequences.

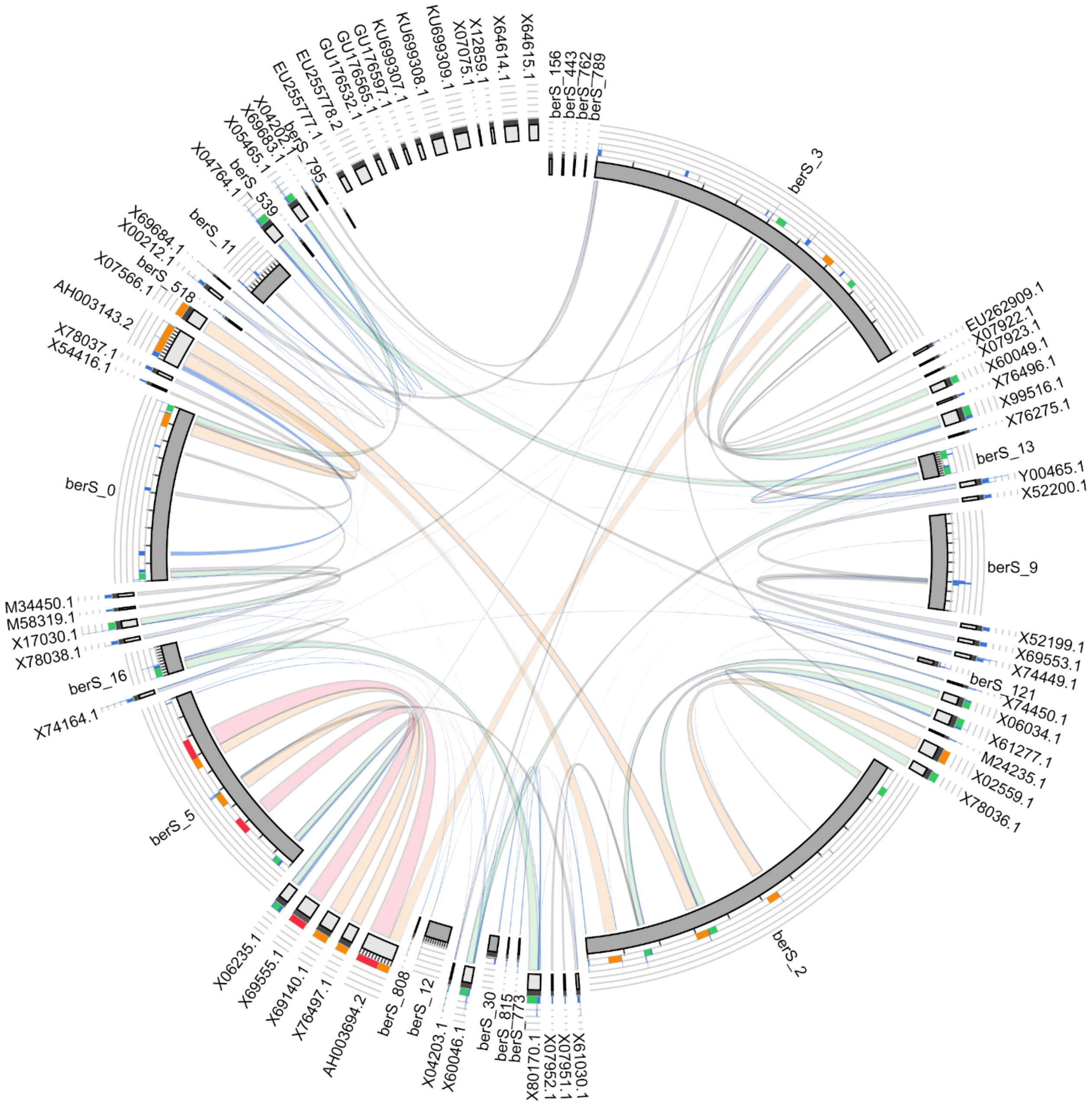


Figure 7. Circos-based visualization of the BLAST result between the NCBI sequences and the contigs that are part of the final mitochondrial sequence of *O. villaricae*. Shown is an untangled Circos plot (with the untangled option) generated by Circoletto for the BLAST search between the IDBA graph contig set of *O. villaricae* and sequences from NCBI harboring the taxonomic ids of *O. villaricae* (3941 and 3950). The quality of the BLAST hits is represented by the link color from red (best bit scores), over orange and green to blue (worst bit score).

PCR and Southern blots validate *in silico* model

As introduced in Figure 1, each RRP in the IDBA-graph model is connected to four different contigs, together building a “double fork”, as mentioned in the introduction. Rearrangement events can potentially lead to four different variants of combined contigs, so-called contig-repeat-contig (CRC) combinations for every RRP (Figure 1c). To test the existence of all four possible CRC variants, we performed PCR experiments targeting all six identified RRPs and, in addition, Southern blot experiments analyzing one of the RRPs. Exemplary results of both, PCR and Southern analysis, can be found in Figure 8 for the “double fork” berS_518.

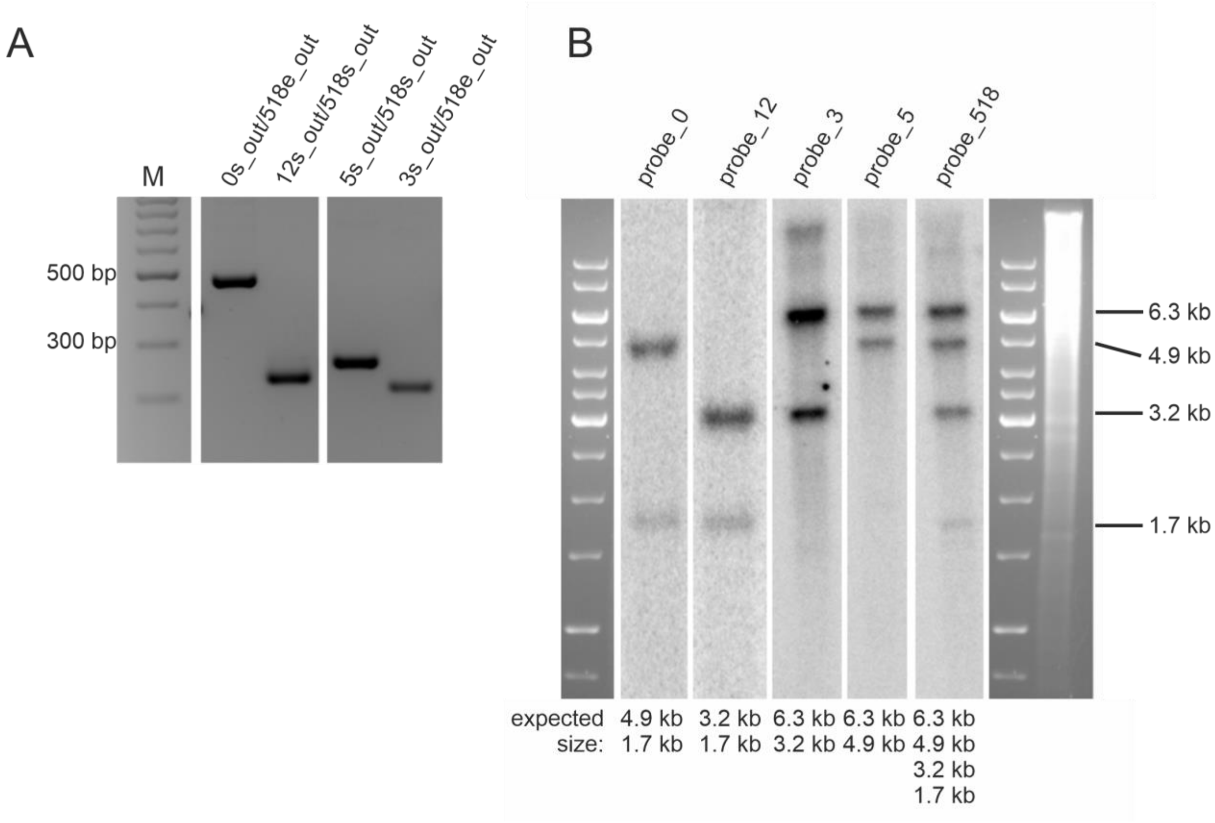


Figure 8. PCR and Southern blot analysis for verification of the predicted mitochondrial genome structure. (A) PCR experiment; *in vitro* fragments for the repeat berS_518 and its adjacent contigs are matching the expected sizes (above each band). This example is representative for all tested contig overlaps. (B) Southern Blot experiment; Genomic DNA was digested with HindIII for berS_518. Probes used for hybridization are indicated above the blot, the expected fragments below the blot. Probe_0, _3, _12, and _5 result in two expected fragments, whereas probe_518 appears in all four possible variants as expected.

In more detail, for the PCR experiments, primer pairs were designed that they either span one of the four contig-repeat boundaries (CRC), or both primer mates target the unique sequences of a CRC, fully spanning the repeat (CRC) (Supplementary Table 2). In all cases, the PCR amplification was tested via gel electrophoresis and resulted in distinct DNA bands that correspond well with the expected fragment length. These results show that the overlaps between contigs and repeats (CRC) as well as all adjacent contig combinations (CRC) exist *in vitro*.

For the Southern blots, as calculated from the *in silico* model, we expected fragments of 1.7 kb and 4.9 kb when berS_518 was analyzed in a *Hind*III digest. The two fragments were detected by hybridization to the two CRCs to probe 0: *Hind*III - berS_0 – berS_518 – berS_12 - *Hind*III (1.7 kb) and *Hind*III - berS_0 – berS_518 – berS_5 - *Hind*III (4.9 kb; Figure 8b). In the first lane of the blot, two bands of the expected size were obtained (Figure 8b). For probes 3, 12 and 5, sizes of expected and verified sequence variants aligned as well. Lastly, probe 518 was designed directly on the repeat and should detect all four possible sequence combinations (1.7 kb, 3.2 kb, 4.9 kb, and 6.3 kb), which is indeed the case. In summary, both wet lab techniques confirmed the graph-derived *in silico* model of the *O. villaricae* mtDNA.

Recombinogenic repeat pairs exist in different stoichiometries

To further substantiate the experimental validations of the predicted CRC configurations of all RRPs, and to bring them into a stoichiometric context, we performed two additional analyses. For this, we used mtDNA-enriched Illumina mate pair (5 kb insert size) and total DNA PacBio RSII (size selection > 5 kb) data that were available for *O. elata*. To be able to calculate read counts for the four CRCs of each RRP, a developed custom data processing pipeline was used to overcome the peculiarities of both datasets after mapping them to the contigs in the final IDBA graph of *O. elata* (for details see Material and Methods). Illumina mate pair: Briefly, with increased insert size of the Illumina mate pair data, it was now possible to span the occurring RRPs. However, mates of read pairs can map on contigs that are not directly neighboring the RRPs, as contig length can vary between hundreds of bases to dozens of kilobases. For this, so called contig chains were defined (Supplemental Figure 5). Contig chains are a subsequent stretch of contigs between two RRPs. With this extension, reads were counted that map on the contig chains, instead of mapping on the contigs flanking the RRP, only. PacBio: For the PacBio dataset, a special filtering strategy of read alignments (Supplemental Figure 6) was necessary to identify and separate nuclear from mitochondrial reads. In brief, to calculate stoichiometry, only PacBio reads that map completely on a subsequent stretch of neighboring contigs without gaps were kept for read counting. PacBio reads that mapped only partially and/or on isolated contigs and/or with gaps were discarded.

Table 4 summarizes the stoichiometric analysis and its statistics, specifically assessing relative abundance of all four alternative CRCs for every given RRP, from which two observations can be deduced. First, the four different CRC configurations that are possible for every RRP, occur in different proportions, by which the RRPs can be divided into two distinct groups. In group 1, one CRC configuration dominates (largest percentage >95%) over all other three configurations, thereby possibly identifying those minor RRPs as false positives (Table 4), namely johSt_1348 and johSt_1564 which were possible long-size repeats in *O. elata*, and also johSt_20304. In group 2, two out of the four CRC configurations are equally abundant, with the other two CRC configurations being underrepresented, but still existing to a significant degree (Table 4).

Table 4. Stoichiometric contig-repeat-contig configuration statistics. Total read statistics as well as the read frequency distributions for all contig-repeat-contig (CRC) configurations (conf) of each recombinogenic repeat pair (RRP) found in *O. elata* spanned by Illumina Mate-Pair and PacBio long reads. Green indicates an agreement between Illumina and PacBio read frequencies; orange not. Usage factor is defined by dividing each sum of total reads by the lowest number occurring in each respect column (Illumina=21,865; PacBio=665) and rounded to the nearest integer. [a] Occurrence by very long insert sizes; PacBio data reveal reads where the outer contigs are the CRC contigs; [b] Nuclear contamination; [c] Can be corrected to near 100% after removing [a] and [b]. [d] Illumina usage factor is 1 instead of 2 for johSt_3550 as two contig chains consist only one contig as after this contig directly a RRP comes next (which is, by definition, the end of a contig chain).

	Contig ID	**1348**	**1564**	**3550**	**12875**	**14298**	**20236**	**20304**	**20310/15**	**20316**
Conf 1	Illumina	1,057	20,863	346	3,293	569	28,545	224	1,035	731
	[%]	5[a]	95	1	5	1	47	1	2	2
	PacBio	0	697	60	12	5	943	1	1	0
	[%]	0	100	4	1	0	52	0	0	0
Conf 2	Illumina	1,000	478	15,538	27,852	28,789	829	26,215	18,131	24,142
	[%]	5[b]	2	49	45	50	1	97	40	50
	PacBio	0	3	579	882	958	8	742	336	585
	[%]	0	0	39	47	52	0	100	30	52
Conf 3	Illumina	78	155	14,865	29,759	26,064	717	62	25,412	23,248
	[%]	0	1	47	48	45	1	0	56	48
	PacBio	1	0	736	994	890	5	0	768	545
	[%]	0	0	50	52	48	0	0	69	48
Conf 4	Illumina	20,173	369	966	556	2,606	30,799	512	660	255
	[%]	90[c]	2	3	1	4	51	2	1	1
	PacBio	664	0	93	6	4	872	0	3	0
	[%]	100	0	6	0	0	48	0	0	0
Read Sums	Illumina	22,308	21,865	31,715	61,460	58,028	60,890	27,013	45,238	48,376
	PacBio	665	700	1,468	1,894	1,857	1,828	743	1,108	1,130
Usage Factor	Illumina	1.0	1.0	1.5[d]	2.8	2.7	2.8	1.2	2.1	2.2
	PacBio	1.0	1.1	2.2	2.8	2.8	2.7	1.1	1.7	1.7

Second, by summing up the reads of all four CRC configurations of each RRP we could calculate the overall usage of each RRP. Depending on the RRP, it ranges between 21,865 to 61,460 reads for the mate pair dataset and between 665 to 1,894 reads for the PacBio dataset, respectively (Table 4). To determine whether recombination at some RRPs was more frequent than at others, read counts were normalized by dividing the total read number by the lowest read count that occurred in each dataset. This yielded the so-called “usage factor”. Overall usage showed an up to threefold difference between the RRPs, which allowed grouping into three fractions (factor 1, 2, and 3; an usage factor of 2 or 3 means that this particular RRP is used twice or three times as often than a RRP with usage factor of 1). The usage factors are presented in Table 4 and are consistent for both data sets (Illumina and PacBio) for all RRPs.

Prediction-based analysis reveals variability of mitochondrial genome isoforms

After the identification of all RRPs, the reconstruction of the master circle, as well as resolving stoichiometry of recombinatoric events, our ultimate goal was to predict all possible rearrangements and all possible (sub)graphs. This was accomplished with a new algorithm to estimate the structural diversity. In brief, the new algorithm is embedded into the SAGBAC pipeline and generates all paths through a graph considering the two rules which were set previously, but with a slight difference. (1) Start and end the path at the same vertex but use a different edge, and (2), traverse each repeat no more than twice. Hence, the difference between this and the original algorithm to reconstruct the master cycle from the graph is that it is not obligatory to traverse all repeats twice. Instead, sub-circles that lack some repeats and/or harbor only one repeat just as a singleton can be generated. Applying these rules, as many as 70 different graphs were predicted for *O. villaricae* (Supplementary Data 1, Supplementary Data 2). Of those 44 graphs represent the master circle in different configurations (all repeats twice, each of all other contig once), six small graphs/sub-circles (no repeat twice, some but not all of the other contigs once) and 20 intermediate-sized graphs/sub-circles (all repeats at least once, some but not all of the other contigs once).

How master- and sub-circles can emerge from each other is illustrated exemplarily in Figure 9. Sub-circles that derive from the four CRC configurations of berS_518 are displayed in Figure 9b and which recombination events are necessary to obtain them (Figure 9a): Sub-circularization of original master circle at berS_518 (event I), inversion of sequence between the berS_795 repeat pair (event II) of original master and down from that a sub-circularization at berS_773 (event III) leads to two other sub-circles.

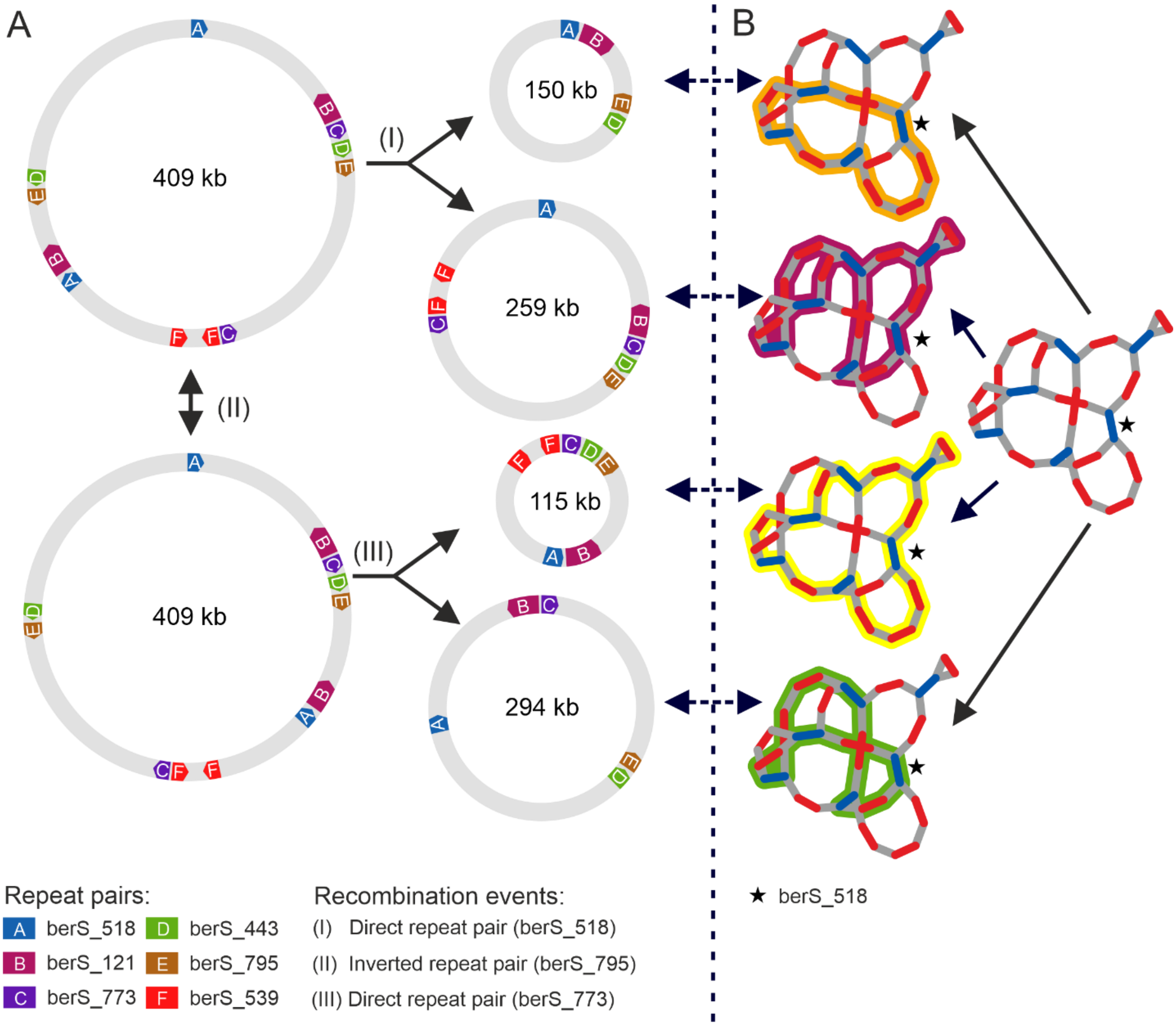
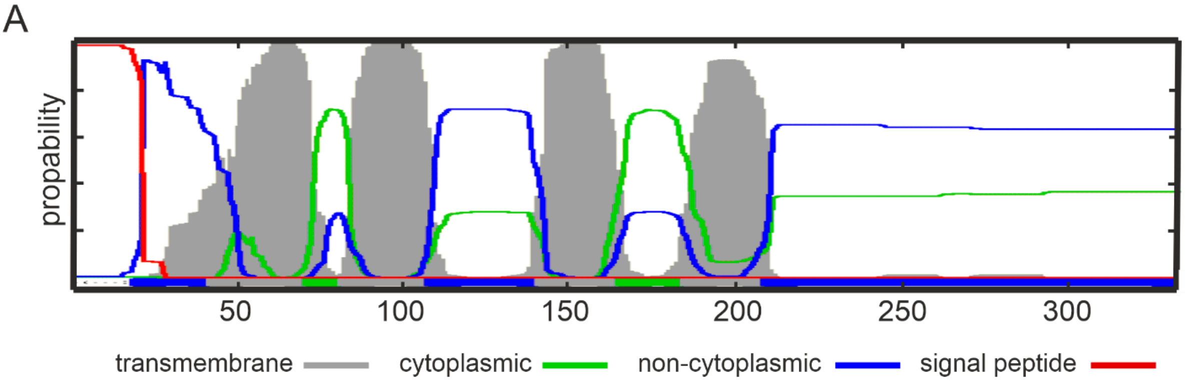


Figure 9. Examples for rearrangement events and their corresponding sub-paths within the IDBA graph of *O. villaricae*. Recombination at inverted repeats, like berS_795 (RRP_E), can lead to an inversion of the sequence between the two mates of the repeat (event I) and therefore to a new orientation of all mates of repeat pairs, which lie on that inverted sequence, which is true for berS_518 (RRP_A), berS_121 (RRP_B) and berS_773 (RRP_C) in that particular case. Direct repeats (here exemplary for berS_518 and berS_773) lead to the formation of two smaller sub-circles in event II (259 kb and 150 kb) and event III (115 kb and 294 kb). Circular genome views were generated using AngularPlasmid. Proportions of repeat and contig sizes was altered to give a proper view of participating RRPs. (B) Illustration of the usage of the four different contig-repeat-contig (CRC) combinations for the identified repeat berS_518 (RRP_A) within the IDBA graph of *O. villaricae* leading to four different sub paths which correspond to one of the sub-circles (dashed arrows) generated in (A). Within the graphs, a pair of two vertices defines a contig (red edge) or a repeat (green edge) and an overlapping event between two different contig ends is represented by a grey edge. Black star: contig representing the identified repeat berS_518 (RRP_A).

The *Oenothera* mitochondrial genome might contain loci for cryptic cytoplasmic male sterility

Beyond the power of the mitochondrial genome reconstruction, our ISEIS algorithm unveiled a potential locus for cryptic cytoplasmic male sterility (CMS). An alternative *nad6* gene is generated that is present in the IDBA graphs of two of the three *Oenothera* species (Figure 4). The alternative sequence is present in *O. villaricae* (berS_156) and *O. biennis* (suavG_433) but is lost in *O. elata*. In comparison to the native Nad6 protein (Figure 10a), the recombinant versions share their N-terminus (80 amino acid residues), whereas further downstream, the protein sequences differ. Here, the stop codon lies in the alternative contig sequence instead of the next contig with identical sequence. Compared to the non-recombined protein, the recombined versions differ in their number of trans-membrane domains (three vs. four) that



```
1     MILSVLSSPA LVSGLMVARA KNPVHSVLFP IPVFRNTSGL LLLLGLDFSA MIFPVVHIGA 60
61    IAVLFLFVVM MFHIQIAEIH ESAMIFPVVH IGAIAVLFLF VVMMFHIQIA EIHESLLRMV 120
121   RMGRGKGNPT GWIARQSTIL LRLWIPCLLL SSTFLFFYLI SFHFGLNFQP LFEKVACSLG 180
181   GRALSFALSK GLGCSGGVAF LISVTLRLVL NSEQDLNLFM DPQPPALREW EQALELPDTQ 240
241   SAPEALRAHV KGELFVLFNI GRRRTLSEKH FYNYIDPLAL DNSSVGFCKA LLAKVDSLQR 300
301   EHWNGGNHRR FVFGTPIERA RLYSIMWDYE SNP
```

B

```
atp9_orig  1   MLEGAKSMGAGAATIASAGAAIGIGNVFSSLIHSVARNPSLAKQLFGYAILGFALTEAIA 60
atp9_alt   1   MLEGAKSMGA--------GAAIGIGNVFSSLIHSVARNPSLAKQLFGYAILGFALTEAIA 52
               **********        ********************************************

atp9_orig  61  LFAPMMAFLILFVFRSVK------------------------   78
atp9_alt   53  LFAPMMAFLILFVFRSVKFRGGCSLFLTKKMRPARKKLKRAG   94
               ******************
```

Figure 10. Putative cryptic CMS loci. Illustrated are two mitochondrial genes, for which, besides their fully functional version, another alternative version was observed. (A) Protein sequence comparison of the alternative *nad6* gene with its native fully functional version in *O. villaricae* and prediction of transmembrane helices. Identical amino acids are in bold font. Transmembrane domains as predicted by Phobius are underlined. (B) Protein sequence alignment for Atp9 for which, besides the original version (atp9_orig), a putative second version (atp9_alt) was identified in *O. elata*. Bold-face amino acids highlight the RNA editing site leading to a truncation of both versions.

leads to a change in topology of the C-terminus. The latter is now found in the intermembrane space, rather than in the matrix (Figure 10a).

Besides this locus, which is directly discernible from the IDBA graph, another locus was identified as a duplication of *atp9* gene in all three investigated *Oenothera* species. With its original version, the alternative one shares the same start codon. The translated alternative protein sequence, in turn, displays a deletion of eight amino acid within the N-terminal sequence and should be 16 amino acid residues longer than the original *atp9* gene. However, the 16-amino acid extension is seemingly not translated, because, as is the original gene, also the alternative *atp9* gene becomes edited at *atp9*eU199R*. This leads to a premature stop codon after amino acid 66 (Figure 10b). If the editing factor of the premature stop codon acts as fertility restorer locus is an interesting hypothesis, which deserves future investigation.

Lastly, 40 bp upstream of the *atp9* variant, a 789 bp long ORF is present. The amino acid sequence of this ORF has a high similarity to ORF873 in *Helianthus annuus*, where it is associated to the MAX1 type of cytoplasmic male-sterility in sunflower.

DISCUSSION

Starting with the publication of the Arabidopsis thaliana mitochondrial genome in 1994 (71,72), and after more than a quarter century of plant mitochondrial genome (PMG) research, there is still an obvious imbalance of published land plant organelle genomes. As of September 2021, the NCBI holds 23 times more chloroplast than mitochondrial genomes of land plants (6530 chloroplast vs. 284 mitochondrial ones). This can be explained by a research focus on chloroplast genomes, especially for phylogenetic studies, but also by technical obstacles associated with sequencing mitochondrial genomes. Genome Skimming is the most popular approach to sequence plastid genomes (73-75) as even whole genome sequencing at low coverage contains enough reads to assemble them. It is most efficient when total DNA is extracted from mature leaves for a standard Illumina paired-end library preparation, as this developmental stage has the highest abundance of plastid DNA, and hence, then delivers the highest read coverage (76,77). Many popular reference-based or *de novo* assemblers can be used afterwards to generate the desired chloroplast genome as a single contig, as they are in almost all cases structurally conserved and repeat-poor (78-80). However, such a straight-forward workflow is not applicable for PMGs as they exhibit a high degree of complexity. Our success, reported in this study, relied on an adapted and optimized workflow and the developed methodology. By approaching the post-assembly task from a graph-based perspective, we were not only successful in constructing a structural model that captures the complexity of PMGs, but also in predicting a defined spectrum of alternative PMG isoforms for the first time.

Our initial test of four *de novo* assemblers revealed that only one (IDBA) was capable to generate a closed circular graph, despite using NGS libraries of mainly mitochondrial origin and a purity level of mtDNA of above 95%. The IDBA assembler has been designed for sequencing data with highly uneven sequencing depth and is therefore capable of distinguishing between nuclear, plastidial, and mitochondrial sequences as these are very unevenly distributed in mitochondria-enriched samples. This is supported by the observation that almost the complete plastid genome exists within the IDBA assembly. Newbler assemblies most likely failed to construct a circular graph, because the coverage of the Roche 454 data with approx. 35x and 74x was possibly too low, i.e. causing the Newbler algorithm to misassemble and/or introduce scaffolding errors. MIRA assemblies have failed to create circular graphs as well, which is surprising, as it is explicitly mentioned in the manual that it can handle repetitive sequences. Given its exorbitant runtime for ultra-deep coverage data sets, MIRA should rather be avoided for such data. It also turned out that only IDBA was able to split assemblies into structural units of repetitive and non-repetitive sequences, which was essential for our post-assembly methodology and is explainable by the implemented De Bruijn graph algorithm (43).

But there are two limitations of IDBA that were uncovered in our graph-based visualization and necessitated manual curation of assemblies and/or graphs. One is related to assembler accuracy, reflected by the occurrence of unconnected vertices. This can be biologically relevant as, on the one hand, they can represent linear isoforms of PMGs, but on the other hand can also be explainable by false-positive contig breaks/ends. The latter seems to have been the case for *O. biennis* and *O. elata*, as in all instances a high number of read pairs were identified within the sequencing data used for the assembly, which went beyond the contig end boundary. The second limitation can be traced back to technical and biological noise within the sequencing data itself. The technical noise is explainable by the impurity of the fraction taken from the sucrose gradient, which was especially observed in the *O. elata* mtDNA dataset. Biological noise can be attributed to so called promiscuous DNA resulting from inter-compartimental DNA exchange. DNA transfer from the plastid into the nucleus (NUPT) and mitochondria (MIPT) occurred often during plant evolution (81-85), and in some cases even from the nucleus to the mitochondrion, as also described very early within the evening primrose (86,87). This contamination resulted in the interconnection of plastidial and mitochondrial subgraphs, introgression of single contigs within the mitochondrial graph, and/or false-positive sequence overlaps in regards of true mitochondrial origin. Hence, it made it necessary to remove sets of vertices and edges (subgraphs), pairs of vertices (contigs) or single edges (sequence overlaps) respectively. As the underlying reasons from which these problems arise are manifold, their corrections can only be done manually or in some cases semi-automatically by the user, as it is challenging to translate human experience and intuition into computational algorithms.

Obviously, promiscuous DNA does interfere with assembly strategies that start from whole genome sequencing data generated from total DNA. An obvious solution, to identify mtDNA reads by unmapped reads, i.e. mapping total sequencing data first against available plastid and nuclear genomes of the species of interest, should be avoided. This would inevitably result in unconnected contigs, as sequences, which are present in both organelle genomes, would be removed beforehand and could be falsely interpreted as biologically relevant. In this context, it also should be mentioned that an available wrapper program called mitoBIM uses MIRA to reconstruct mitochondrial genomes (88). It employs an iterative two-tier approach, by first baiting perceived mitochondrial reads from total DNA sequence with known mitochondrial genes and/or genomic sequences using MIRA in mapper mode. In a second step, it assembles those reads using the MIRA in assembler mode. As, in our hands, MIRA cannot generate a circular graph for mtDNA-enriched DNA (see above), our assumption is that it will most likely run into problems for a total DNA data set as well. However, and probably more importantly, all reference-based assemblers catch nuclear-mitochondrial DNA (NUMT) sequences. By definition, they do not belong to the mitochondrial genome, underlie higher mutation rates and most important different recombination events (89-91). This can lead to uncorrectable/wrong graphs and introduces SNPs/indels, which do not actually belong to the PMG itself.

The need for sequence assembly visualization is as old as sequence assembly itself. For example, Bandage, a tool that has been established as a standard program to investigate graph-based assembler outputs, is frequently used even for large nuclear genomes (9). In most cases it is employed to investigate and solve problematic issues such as loops or bubbles. The Contiguity software can help in visualization of non-graph based assembler outputs, as a reverse mapping step is implemented to find links between different contigs (10). However, this is very time-consuming as one needs to arrange the resulting linked contigs manually, which might take hours. Also, Contiguity is not orientation/strandedness-aware, which, in the context of sequence overlaps (Supplementary Figure 4) is important to be considered. In our study, accounting for orientation did lead to the identification of real, contiguous stretches of contigs. IDBA, our assembler of choice, is a true De Bruijn graph-based assembler, but lacks any kind of supported graph-based output format. Consequently, with our SAGBAC pipeline we implemented those missing features including semi-automatically graph output curation that reduces necessary hands-on time.

The final graphs, generated here, can be best described as spatially folded and circularized models of PMGs, in which the repetitive contigs are easily identifiable by the occurrence of "double forks" (Figure 1c). It is well known that PMG rearrangements appear at RRPs depending on their orientation and size. Repeats larger than 1 kb recombine very frequently, symmetrically, and are reversible, which isomerize the genome (71,72,92,93). By contrast, repeats between 200 and 1000 bp recombine infrequently, asymmetrically, and are part of the break-induced repair (BIR) pathway increasing mtDNA complexity (65). Both groups are represented in the three investigated *Oenothera* species. While sharing four RRPs, each of the species has its own RRPs (especially a unique long-size repeat). This is in line with previous findings in that repetitive sequences can differ between closely related species (7,94).

As introduced in Figure 1, with the concept of RRPs embedded in "double forks" and CRC combinations, all master- and sub-circles can be inferred from our graph-based model. The urgent need for such a model was adamantly recommended by Kozik et al, as more than 50% of the PMGs between the years 2016 and 2019 were presented and published just as circles without even mentioning and considering that other models or variants are possible (5). This resistance to surrendering the one-master-molecule view is surprising, as experiments often failed to recover large circular molecules and studies postulated and provided evidence for the existence of alternative models including linear configurations, various independent mtDNA molecules (95-97), branched structures, circularly permuted linear fragments or even more complex structures (6,7). Nonetheless, most submitted PMGs are still represented by just a master circle (e.g. (4,98), more rarely sub-circles (99) or two or more autonomous circles (3,100,101). Yet, despite the inadequate reflection in databases, it is now commonly accepted that alternative isoforms of PMGs most likely co-exist within each cell. With this study we proposed a coherent, graph-based framework that offers an unified view of the complexity of PMGs.

The validation of our model by molecular biology techniques using PCR and Southern blots confirmed all contig-repeat-contig combinations for all RRPs in *O. villaricae*. Additionally, we employed two more advanced technologies, Illumina Mate Pair and PacBio RSII data, not only to verify the qualitative, but also quantitative status of all RRPs of *O. elata*. This goes beyond earlier approaches, in which researchers only used them (i) to scaffold contigs further (Mate-pair; e.g. (100)) or (ii) to assemble the PMGs with additional sequencing sequencing data (PacBio; (3)). In both, exemplarily selected publications, they were unsuccessful in creating one master circle and predicted instead two or more autonomous circles. Interestingly, in both cases some of the autonomous circles carry a subset of exons of a trans-spliced gene each.

With our integrated investigation of structure and stoichiometry of RRPs as well as PMG isoform prediction, we believe to contribute to the advancement of PMG research by further providing evidence for the existence of different PMG isoforms that are simultaneously present at various amounts. Only few publications so far also reported on PMG isoform stoichiometries, which were reported to be similar (99) or different (5). In 2020 it was reported that knock-out mutants for *msh1, recA3* and *polIb* in *Arabidopsis thaliana* showed distinct patterns of large and repeatable shifts in abundance of the mitochondrial genome (2). All three genes are involved in cytoplasmic DNA replication, recombination, and repair. As msh1 is also involved in controlling recombination activity in mitochondria (102), it appears plausible that the stoichiometry is directly or indirectly influenced by this gene, too. This hypothesis is further corroborated by several other studies, which reported changing amounts of mitochondrial genome variants (94) and tight time control of stoichiometry (103) during plant development as well as changing copy number of mitochondrial genes that differ between various genes and tissues (104).

To generate PMG sequences for database submission, rules for traversing the obtained graphs were needed, as only individual linear or circular individual molecules are accepted, but not graphs with different traversal paths. To accomplish this, it was necessary to break our traversal rules (see above) as repeats can occur more than twice, as also shown before (102). Also large repeats, as identified in *O. elata* harboring *rRNAs* and *nad5* exons, had to be tolerated, which, however, is supported by their existence in other plant species within the asterid clade: *Rhazya stricta* (105) (repeat size 36.3 kb, NC_024293.1) and Daucus carota (98) (repeat size 14.2 kb, NC_017855.1). But so far, *Oenothera* is the only genus from the rosid clade, in which this kind of repeat is present or absent within a single genus.

The performed homology-based gene search revealed the presence of all essential gene families (6) and tRNA sets of mitochondrial and plastidial origin (106) found in a wide range of PMGs. But it should be emphasized that our study is one of the few (107) in which homology-based annotation was completely supported by a set of previously published genes for *O. villaricae*, but, more importantly, by a RNA-seq based gene expression atlas for *O. elata*. By a subsequent SNP analysis, we were, moreover, able to lift the concept of RNA editing investigated in *Oenothera* (108) to a mitochondrial-wide RNA editome level, which, with 681 RNA editing sites, is as large as expected for a genus from the angiosperms clade (109).

Three unexpected findings are particularly noteworthy. Two pertain to potential loci for cryptic CMS (110-112) as they encode for alternative variants of Nad6 and Atp9 proteins. And, surprisingly, in both cases they do not originate from a fusion event between the original gene and a downstream ORF, which is the common definition of a cryptic CMS. Only the mid-part of the nad6 gene is exchanged, but leads to a change in topology of its encoded protein, which may hinder the electron transport itself. So far CMS candidates, in which *nad6* participate, were only identified in *Mimulis guttatus* (113,114). CMS loci involving the *atp9* gene were reported many times in the literature, first in Petunia (115,116). But also here our reported CMS locus deviates from the standard CMS definition by involving an RNA editing event to neutralize it, which was also reported previously to occur (117). Regarding the third finding we wish to highlight, some genes are extending into some of the RRPs or the RRPs are, as a whole, part of them (Table 2). As three of those six genes represent the top-3 in a ranked-by-frequency list of CMS loci in major crops (112), these RRPs are perhaps involved in the creation of cryptic CMS loci. Some mitochondria-associated phenotypes (118-120) were observed in crosses between phylogenetically distant *Oenothera* species. which might lead to an incompatibility of RRP sets and nuclear regulators. Such incompatibilities potentially may have played an important role in the evolution of *Oenothera* species and may act as speciation barriers.

CONCLUSION

With the newly developed SAGBAC pipeline and its ISEIS core algorithm it is now possible to systematically investigate the overall mitochondrial status in different *Oenothera* species as well as in various organs and developmental stages and to understand mechanistically the influence of RRPs on the creation of novel, biologically relevant/active open reading frames (cryptic CMS loci), which might explain some mitochondria-associated phenotypes observed within the *Oenothera* genus. In addition, our methodology will possibly allow to identify different mitochondrion types and integrate them as a third dimension into the nuclear-plastome-compatibility chart (121), which was created in the past, linking them to the evolutionary context of speciation.

DATA AVAILABILITY

All NGS-related data are submitted to the Short Read Archive under BioProject PRJNA757974. Final mitochondrial genomes are available on the NCBI Nucleotide database under the following accessions: *O. villaricae* (MZ934755), *O. biennis* (MZ934756), *O. elata* ssp. *hookeri* (MZ934757). Raw assembly outputs are downloadable on figshare ***(****doi:10.6084/m9.figshare.16496814)* and SAGBAC scripts are accessible through github ***(****https://www.github.com/AxelMacFoly/SAGBAC****).***

ACKNOWLEDGEMENT

We thank Werner Dietrich and Wilfried Stubbe, who compiled the comprehensive *Oenothera* collection that allowed us to use the original material of the Renner school; the Max Planck Institute of Molecular Plant Physiology Green Team for their support; the Max-Planck Genome Center (MPGC) as well as the Sequencing Core Facility at the Max Planck Institute for Molecular Genetics (MPIMG) for their efforts in providing the required sequencing data; Liliya Yaneva-Roder for technical assistance and RNA isolation protocol establishment; and Christine Schütt for her preliminary assembly work.

FUNDING

None.

CONFLICTS OF INTEREST

None.

REFERENCES

1. Smith, D.R. and Keeling, P.J. (2015) Mitochondrial and plastid genome architecture: Reoccurring themes, but significant differences at the extremes. *Proc Natl Acad Sci U S A*, **112**, 10177-10184.
2. Wu, Z.Q., Liao, X.Z., Zhang, X.N., Tembrock, L.R. and Broz, A. (2020) Genomic architectural variation of plant mitochondria—A review of multichromosomal structuring. *J Syst Evol*.
3. Varré, J.S., D'Agostino, N., Touzet, P., Gallina, S., Tamburino, R., Cantarella, C., Ubrig, E., Cardi, T., Drouard, L., Gualberto, J.M. *et al.* (2019) Complete Sequence, Multichromosomal Architecture and Transcriptome Analysis of the. *Int J Mol Sci*, **20**.
4. Alverson, A.J., Zhuo, S., Rice, D.W., Sloan, D.B. and Palmer, J.D. (2011) The mitochondrial genome of the legume Vigna radiata and the analysis of recombination across short mitochondrial repeats. *PLoS One*, **6**, e16404.
5. Kozik, A., Rowan, B.A., Lavelle, D., Berke, L., Schranz, M.E., Michelmore, R.W. and Christensen, A.C. (2019) The alternative reality of plant mitochondrial DNA: One ring does not rule them all. *PLoS Genet*, **15**, e1008373.
6. Sloan, D.B., Alverson, A.J., Chuckalovcak, J.P., Wu, M., McCauley, D.E., Palmer, J.D. and Taylor, D.R. (2012) Rapid evolution of enormous, multichromosomal genomes in flowering plant mitochondria with exceptionally high mutation rates. *PLoS Biol*, **10**, e1001241.

7. Gualberto, J.M., Mileshina, D., Wallet, C., Niazi, A.K., Weber-Lotfi, F. and Dietrich, A. (2014) The plant mitochondrial genome: dynamics and maintenance. *Biochimie*, **100**, 107-120.
8. West, D. (1996) Introduction to graph theory Prentice Hall Upper Saddle River NJ USA.
9. Wick, R.R., Schultz, M.B., Zobel, J. and Holt, K.E. (2015) Bandage: interactive visualization of *de novo* genome assemblies. *Bioinformatics*, **31**, 3350-3352.
10. Sullivan, M.J., Zakour, N.L.B., Forde, B.M., Stanton-Cook, M. and Beatson, S.A. (2015). PeerJ PrePrints.
11. Brennicke, A. (1980) Mitochondrial DNA from *Oenothera* berteriana: Purification and properties. *Plant Physiol*, **65**, 1207-1210.
12. Wissinger, B., Schuster, W. and Brennicke, A. (1991) Trans splicing in *Oenothera* mitochondria: *nad1* mRNAs are edited in exon and trans-splicing group II intron sequences. *Cell*, **65**, 473-482.
13. Hagemann, R. (2010) The foundation of extranuclear inheritance: plastid and mitochondrial genetics. *Mol. Genet. Genom.*, **283**, 199-209.
14. Greiner, S., Sobanski, J. and Bock, R. (2015) Why are most organelle genomes transmitted maternally? *Bioessays*, **37**, 80-94.
15. Schwemmle, J., Haustein, E., Sturm, A. and Binder, M. (1938) Genetische und zytologische Untersuchungen an *Eu-Oenotheren*: Teil I bis VI. *Mol. Gen. Genet.*, **75**, 358-800.
16. Ulbricht-Jones, E.S., Dotzek, J. and Greiner, S. (2021) Maternal inheritance of mitochondria and biparental inheritance of chloroplasts allows separation of cytoplasmic effects in the evening primrose (*Oenothera*). *In Preparation*.
17. Kirk, J. and Tilney-Basset, R. (1978) The Plastids: Their Chemistry, Structure, Growth and inheritance, *Amsterdam-New York-Oxford: Elsevier/North Holland Biomedical Press.*
18. Greiner, S. and Bock, R. (2013) Tuning a ménage à trois: co-evolution and co-adaptation of nuclear and organellar genomes in plants. *BioEssays*, **35**, 354-365.
19. Brennicke, A. and Schwemmle, B. (1984) Inheritance of mitochondrial DNA in *Oenothera berteriana* and *Oenothera odorata* Hybrids. *Zeitschrift für Naturforschung*, **39c**, 191-192.
20. Dotzek, J. (2016) Mitochondria in the genus *Oenothera* – Non-Mendelian inheritance patterns, in vitro structure and evolutionary dynamics. *PhD Thesis*, University of Potsdam, Potsdam.
21. Greiner, S., Rauwolf, U., Meurer, J. and Herrmann, R.G. (2011) The role of plastids in plant speciation. *Mol. Ecol.*, **20**, 671-691.
22. Zupok, A., Kozul, D., Schöttler, M.A., Julia, N., Garbsch, F., Liere, K., Malinova, I., Bock, R. and Greiner, S. (2021) A photosynthesis operon in the chloroplast genome drives speciation in evening primroses. *bioRxiv*, 2020.2007.2003.186627 (The Plant Cell, in revision).
23. Sobanski, J., Giavalisco, P., Fischer, A., Kreiner, J.M., Walther, D., Schöttler, M.A., Pellizzer, T., Golczyk, H., Obata, T., Bock, R. *et al.* (2019) Chloroplast competition is controlled by lipid biosynthesis in evening primroses. *Proceedings of the National Academy of Sciences*, **116**, 5665-5674.
24. Stubbe, W. and Steiner, E. (1999) Inactivation of pollen and other effects of genome-plastome incompatibility in *Oenothera*. *Plant Systematics and Evolution*, **217**, 259-277.
25. Stubbe, W. (1989) The falcifolia syndrome of *Oenothera*: III. The general pattern of its non-Mendelian inheritance. *Mol. Gen. Genet.*, **218**, 499-510.
26. Stubbe, W. (1989) The falcifolia syndrome of *Oenothera*: IV. Loss of falcifolia-determining factors. *Mol. Gen. Genet.*, **218**, 511-515.
27. Bartehlmess, A. (1965) Grundlagen der Vererbung. Akademische Verlagsgesellschaft Athenaion, Frankfurt am Main.
28. Harte, C. (1994) *Oenothera* - Contributions of a Plant to Biology. 1st ed. Springer, Berlin, Heidelberg, New York.
29. Greiner, S. and Köhl, K. (2014) Growing evening primroses (*Oenothera*). *Front. Plant Sci.*, **5**, 38.
30. Stubbe, W. (1953) Genetische und zytologische Untersuchungen an verschiedenen Sippen von *Oenothera suaveolens*. *Z. indukt. Abstamm. Vererbungsl.*, **85**, 180-209.
31. Cleland, R.E. (1935) Cyto-taxonomic studies on certain *Oenothera*s from California. *Proc. Am. Philos. Soc.*, **75**, 339-429.

32. Dietrich, W. (1977) The South American species of *Oenothera* sect. *Oenothera* (Raimannia, Renneria; Onagraceae). *Annals of the Missouri Botanical Garden*, 425-626.

33. Dietrich, W., Wagner, W.L. and Raven, P.H. (1997) Systematics of *Oenothera* section *Oenothera* subsection *Oenothera* (Onagraceae). Sistemática de *Oenothera* sección *Oenothera* subsección *Oenothera* (Onagraceae). *Systematic Botany*, **50**, 12-34.

34. Wagner, W.L., Hoch, P.C. and Raven, P.H. (2007) Revised classification of the Onagraceae. *Systematic Botany Monographs*.

35. Michalecka, A.M., Agius, S.C., Moller, I.M. and Rasmusson, A.G. (2004) Identification of a mitochondrial external NADPH dehydrogenase by overexpression in transgenic Nicotiana sylvestris. *Plant J*, **37**, 415-425.

36. Rodiger, A., Baudisch, B. and Klosgen, R.B. (2010) Simultaneous isolation of intact mitochondria and chloroplasts from a single pulping of plant tissue. *J Plant Physiol*, **167**, 620-624.

37. Deuel, H., Neukom, H. and Weber, F. (1948) Reaction of boric acid with polysaccharides. *Nature*, **161**, 96-97

38. Sepúlveda, E., Sáenz, C., Aliaga, E. and Aceituno, C. (2007) Extraction and characterization of mucilage in *Opuntia* spp. *Journal of Arid Environments*, **68**, 534-545.

39. Massouh, A., Schubert, J., Yaneva-Roder, L., Ulbricht-Jones, E.S., Zupok, A., Johnson, M.T.J., Wright, S.I., Pellizzer, T., Sobanski, J., Bock, R. *et al.* (2016) Spontaneous chloroplast mutants mostly occur by replication slippage and show a biased pattern in the plastome of *Oenothera*. *Plant Cell*, **28**, 911-929.

40. Rauwolf, U. (2008) PhD thesis, Ludwig-Maximilians-University, Munich.

41. Hollister, J.D., Greiner, S., Wang, W., Wang, J., Zhang, Y., Wong, G.K., Wright, S.I. and Johnson, M.T. (2015) Recurrent loss of sex is associated with accumulation of deleterious mutations in *Oenothera*. *Mol Biol Evol*, **32**, 896-905.

42. Massouh, A., Schubert, J., Yaneva-Roder, L., Ulbricht-Jones, E.S., Zupok, A., Johnson, M.T., Wright, S.I., Pellizzer, T., Sobanski, J., Bock, R. *et al.* (2016) Spontaneous Chloroplast Mutants Mostly Occur by Replication Slippage and Show a Biased Pattern in the Plastome of *Oenothera*. *Plant Cell*, **28**, 911-929.

43. Peng, Y., Leung, H.C., Yiu, S.M. and Chin, F.Y. (2012) IDBA-UD: a *de novo* assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, **28**, 1420-1428.

44. Chevreux, B., Wetter, T. and Suhai, S. (1999), *German conference on bioinformatics*. Citeseer, Vol. 99, pp. 45-56.

45. Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, **25**, 1754-1760.

46. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. and Genome Project Data Processing, S. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078-2079.

47. Quinlan, A.R. and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841-842.

48. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403-410.

49. Csárdi, G. and Nepusz, T. (2006) The igraph software package for complex network research. *InterJournal, Complex Systems*, **1695**, 1-9.

50. Darzentas, N. (2010) Circoletto: visualizing sequence similarity with Circos. *Bioinformatics*, **26**, 2620-2621.

51. Krzywinski, M., Schein, J., Birol, İ., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J. and Marra, M.A. (2009) Circos: An information aesthetic for comparative genomics. *Genome Research*, **19**, 1639-1645.

52. Leggett, R.M., Clavijo, B.J., Clissold, L., Clark, M.D. and Caccamo, M. (2014) NextClip: an analysis and read preparation tool for Nextera Long Mate Pair libraries. *Bioinformatics*, **30**, 566-568.

53. Tillich, M., Lehwark, P., Pellizzer, T., Ulbricht-Jones, E.S., Fischer, A., Bock, R. and Greiner, S. (2017) GeSeq - versatile and accurate annotation of organelle genomes. *Nucleic Acids Res*.

54. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. and Gingeras, T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15-21.

55. Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X. and Ruden, D.M. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)*, **6**, 80-92.

56. Thorvaldsdottir, H., Robinson, J.T. and Mesirov, J.P. (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform*, **14**, 178-192.

57. Binder, S., Marchfelder, A., Brennicke, A. and Wissinger, B. (1992) RNA editing in trans-splicing intron sequences of nad2 mRNAs in *Oenothera* mitochondria. *J Biol Chem*, **267**, 7615-7623.

58. Knoop, V., Schuster, W., Wissinger, B. and Brennicke, A. (1991) Trans splicing integrates an exon of 22 nucleotides into the nad5 mRNA in higher plant mitochondria. *EMBO J*, **10**, 3483-3493.

59. Lohse, M., Drechsel, O., Kahlau, S. and Bock, R. (2013) OrganellarGenomeDRAW--a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. *Nucleic Acids Res*, **41**, W575-581.

60. Greiner, S., Lehwark, P. and Bock, R. (2019) OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes. *Nucleic Acids Res*, **47**, W59-W64.

61. Lehwark, P. and Greiner, S. (2019) GB2sequin - A file converter preparing custom GenBank files for database submission. *Genomics*, **111**, 759-761.

62. Richly, E. and Leister, D. (2004) NUPTs in sequenced eukaryotes and their genomic organization in relation to NUMTs. *Mol Biol Evol*, **21**, 1972-1980.

63. Binder, S. and Brennicke, A. (1993) Transcription initiation sites in mitochondria of *Oenothera* berteriana. *J Biol Chem*, **268**, 7849-7855.

64. Logan, D.C. (2010) Mitochondrial fusion, division and positioning in plants. *Biochem Soc Trans*, **38**, 789-795.

65. Davila, J.I., Arrieta-Montiel, M.P., Wamboldt, Y., Cao, J., Hagmann, J., Shedge, V., Xu, Y.Z., Weigel, D. and Mackenzie, S.A. (2011) Double-strand break repair processes drive evolution of the mitochondrial genome in Arabidopsis. *BMC Biol*, **9**, 64.

66. Ankenbrand, M.J., Hohlfeld, S., Hackl, T. and Förster, F. (2017) AliTV—interactive visualization of whole genome comparisons. *PeerJ Computer Science*, **3**, e116.

67. Chan, P.P., Lin, B.Y., Mak, A.J. and Lowe, T.M. (2019) tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *BioRxiv*, 614032.

68. Brennicke, A. and Blanz, P. (1982) Circular mitochondrial DNA species from *Oenothera* with unique sequences. *Molecular and General Genetics MGG*, **187**, 461-466.

69. Schuster, W., Combettes, B., Flieger, K. and Brennicke, A. (1993) A plant mitochondrial gene encodes a protein involved in cytochrome c biogenesis. *Molecular and General Genetics MGG*, **239**, 49-57.

70. Schuster, W., Hiesel, R., Issac, P.G., Leaver, C.J. and Brennicke, A. (1986) Transcript termini of messenger RNAs in higher plant mitochondria. *Nucleic acids research*, **14**, 5943-5954.

71. Klein, M., Eckert-Ossenkopp, U., Schmiedeberg, I., Brandt, P., Unseld, M., Brennicke, A. and Schuster, W. (1994) Physical mapping of the mitochondrial genome of Arabidopsis thaliana by cosmid and YAC clones. *Plant J*, **6**, 447-455.

72. Unseld, M., Marienfeld, J.R., Brandt, P. and Brennicke, A. (1997) The mitochondrial genome of Arabidopsis thaliana contains 57 genes in 366,924 nucleotides. *Nat Genet*, **15**, 57-61.

73. Straub, S.C., Parks, M., Weitemier, K., Fishbein, M., Cronn, R.C. and Liston, A. (2012) Navigating the tip of the genomic iceberg: Next-generation sequencing for plant systematics. *Am J Bot*, **99**, 349-364.

74. Hao, W., Fan, S., Hua, W. and Wang, H. (2014) Effective extraction and assembly methods for simultaneously obtaining plastid and mitochondrial genomes. *PLoS One*, **9**, e108291.

75. Weitemier, K., Straub, S.C., Cronn, R.C., Fishbein, M., Schmickl, R., McDonnell, A. and Liston, A. (2014) Hyb-Seq: Combining target enrichment and genome skimming for plant phylogenomics. *Appl Plant Sci*, **2**.
76. Golczyk, H., Greiner, S., Wanner, G., Weihe, A., Bock, R., Börner, T. and Herrmann, R.G. (2014) Chloroplast DNA in mature and senescing leaves: a reappraisal. *Plant Cell*, **26**, 847-854.
77. Greiner, S., Golczyk, H., Malinova, I., Pellizzer, T., Bock, R., Börner, T. and Herrmann, R.G. (2020) Chloroplast nucleoids are highly dynamic in ploidy, number, and structure during angiosperm leaf development. *Plant J*, **102**, 730-746.
78. Straub, S.C., Fishbein, M., Livshultz, T., Foster, Z., Parks, M., Weitemier, K., Cronn, R.C. and Liston, A. (2011) Building a model: developing genomic resources for common milkweed (Asclepias syriaca) with low coverage genome sequencing. *BMC Genomics*, **12**, 211.
79. Zhang, T., Zhang, X., Hu, S. and Yu, J. (2011) An efficient procedure for plant organellar genome assembly, based on whole genome data from the 454 GS FLX sequencing platform. *Plant Methods*, **7**, 38.
80. Sloan, D.B., Triant, D.A., Forrester, N.J., Bergner, L.M., Wu, M. and Taylor, D.R. (2014) A recurring syndrome of accelerated plastid genome evolution in the angiosperm tribe Sileneae (Caryophyllaceae). *Mol Phylogenet Evol*, **72**, 82-89.
81. Ellis, J. (1982) Promiscuous DNA--chloroplast genes inside plant mitochondria. *Nature*, **299**, 678-679.
82. Lin, X., Kaul, S., Rounsley, S., Shea, T.P., Benito, M.I., Town, C.D., Fujii, C.Y., Mason, T., Bowman, C.L., Barnstead, M. *et al.* (1999) Sequence and analysis of chromosome 2 of the plant Arabidopsis thaliana. *Nature*, **402**, 761-768.
83. Knoop, V. (2004) The mitochondrial DNA of land plants: peculiarities in phylogenetic perspective. *Curr Genet*, **46**, 123-139.
84. Leister, D. (2005) Origin, evolution and genetic effects of nuclear insertions of organelle DNA. *Trends Genet*, **21**, 655-663.
85. Wang, D., Wu, Y.W., Shih, A.C., Wu, C.S., Wang, Y.N. and Chaw, S.M. (2007) Transfer of chloroplast genomic DNA to mitochondrial genome occurred at least 300 MYA. *Mol Biol Evol*, **24**, 2040-2048.
86. Schuster, W. and Brennicke, A. (1987) Plastid, nuclear and reverse transcriptase sequences in the mitochondrial genome of *Oenothera*: is genetic information transferred between organelles via RNA? *EMBO J*, **6**, 2857-2863.
87. Kleine, T., Maier, U.G. and Leister, D. (2009) DNA transfer from organelles to the nucleus: the idiosyncratic genetics of endosymbiosis. *Annu Rev Plant Biol*, **60**, 115-138.
88. Hahn, C., Bachmann, L. and Chevreux, B. (2013) Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads--a baiting and iterative mapping approach. *Nucleic Acids Res*, **41**, e129.
89. Wolfe, K.H., Li, W.H. and Sharp, P.M. (1987) Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proc Natl Acad Sci U S A*, **84**, 9054-9058.
90. Drouin, G., Daoud, H. and Xia, J. (2008) Relative rates of synonymous substitutions in the mitochondrial, chloroplast and nuclear genomes of seed plants. *Mol Phylogenet Evol*, **49**, 827-831.
91. Michalovova, M., Vyskot, B. and Kejnovsky, E. (2013) Analysis of plastid and mitochondrial DNA insertions in the nucleus (NUPTs and NUMTs) of six plant species: size, relative age and chromosomal localization. *Heredity (Edinb)*, **111**, 314-320.
92. Abdelnoor, R.V., Yule, R., Elo, A., Christensen, A.C., Meyer-Gauen, G. and Mackenzie, S.A. (2003) Substoichiometric shifting in the plant mitochondrial genome is influenced by a gene homologous to MutS. *Proc Natl Acad Sci U S A*, **100**, 5968-5973.
93. Logan, D.C. (2006) The mitochondrial compartment. *J Exp Bot*, **57**, 1225-1243.
94. Sugiyama, Y., Watase, Y., Nagase, M., Makita, N., Yagura, S., Hirai, A. and Sugiura, M. (2005) The complete nucleotide sequence and multipartite organization of the tobacco mitochondrial genome: comparative analysis of mitochondrial genomes in higher plants. *Mol Genet Genomics*, **272**, 603-615.

95. Oldenburg, D.J. and Bendich, A.J. (1996) Size and Structure of Replicating Mitochondrial DNA in Cultured Tobacco Cells. *Plant Cell*, **8**, 447-461.

96. Oldenburg, D.J. and Bendich, A.J. (2001) Mitochondrial DNA from the liverwort Marchantia polymorpha: circularly permuted linear molecules, head-to-tail concatemers, and a 5' protein. *J Mol Biol*, **310**, 549-562.

97. Backert, S. and Börner, T. (2000) Phage T4-like intermediates of DNA replication and recombination in the mitochondria of the higher plant Chenopodium album (L.). *Curr Genet*, **37**, 304-314.

98. Iorizzo, M., Senalik, D., Szklarczyk, M., Grzebelus, D., Spooner, D. and Simon, P. (2012) *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*, **12**, 61.

99. Guo, W., Grewe, F., Fan, W., Young, G.J., Knoop, V., Palmer, J.D. and Mower, J.P. (2016) Ginkgo and Welwitschia Mitogenomes Reveal Extreme Contrasts in Gymnosperm Mitochondrial Evolution. *Mol Biol Evol*, **33**, 1448-1460.

100. Guo, W., Zhu, A., Fan, W. and Mower, J.P. (2017) Complete mitochondrial genomes from the ferns Ophioglossum californicum and Psilotum nudum are highly repetitive with the largest organellar introns. *New Phytol*, **213**, 391-403.

101. Shearman, J.R., Sonthirod, C., Naktang, C., Pootakham, W., Yoocha, T., Sangsrakru, D., Jomchai, N., Tragoonrung, S. and Tangphatsornruang, S. (2016) The two chromosomes of the mitochondrial genome of a sugarcane cultivar: assembly and recombination analysis using long PacBio reads. *Sci Rep*, **6**, 31533.

102. Arrieta-Montiel, M.P., Shedge, V., Davila, J., Christensen, A.C. and Mackenzie, S.A. (2009) Diversity of the Arabidopsis mitochondrial genome occurs via nuclear-controlled recombination activity. *Genetics*, **183**, 1261-1268.

103. Paszkiewicz, G., Gualberto, J.M., Benamar, A., Macherel, D. and Logan, D.C. (2017) Arabidopsis Seed Mitochondria Are Bioenergetically Active Immediately upon Imbibition and Specialize via Biogenesis in Preparation for Autotrophic Growth. *Plant Cell*, **29**, 109-128.

104. Preuten, T., Cincu, E., Fuchs, J., Zoschke, R., Liere, K. and Börner, T. (2010) Fewer genes than organelles: extremely low and variable gene copy numbers in mitochondria of somatic plant cells. *Plant J*, **64**, 948-959.

105. Park, S., Ruhlman, T.A., Sabir, J.S., Mutwakil, M.H., Baeshen, M.N., Sabir, M.J., Baeshen, N.A. and Jansen, R.K. (2014) Complete sequences of organelle genomes from the medicinal plant Rhazya stricta (Apocynaceae) and contrasting patterns of mitochondrial genome evolution across asterids. *BMC Genomics*, **15**, 405.

106. Warren, J.M. and Sloan, D.B. (2020) Interchangeable parts: The evolutionarily dynamic tRNA population in plant mitochondria. *Mitochondrion*, **52**, 144-156.

107. Grimes, B.T., Sisay, A.K., Carroll, H.D. and Cahoon, A.B. (2014) Deep sequencing of the tobacco mitochondrial transcriptome reveals expressed ORFs and numerous editing sites outside coding regions. *BMC Genomics*, **15**, 31.

108. Hiesel, R., Wissinger, B., Schuster, W. and Brennicke, A. (1989) RNA editing in plant mitochondria. *Science*, **246**, 1632-1634.

109. Takenaka, M., Zehrmann, A., Verbitskiy, D., Härtel, B. and Brennicke, A. (2013) RNA editing in plants and its evolution. *Annu Rev Genet*, **47**, 335-352.

110. Chase, C.D. (2007) Cytoplasmic male sterility: a window to the world of plant mitochondrial-nuclear interactions. *Trends Genet*, **23**, 81-90.

111. Rieseberg, L.H. and Blackman, B.K. (2010) Speciation genes in plants. *Ann Bot*, **106**, 439-455.

112. Chen, L. and Liu, Y.G. (2014) Male sterility and fertility restoration in crops. *Annu Rev Plant Biol*, **65**, 579-606.

113. Case, A.L. and Willis, J.H. (2008) Hybrid male sterility in Mimulus (Phrymaceae) is associated with a geographically restricted mitochondrial rearrangement. *Evolution*, **62**, 1026-1039.

114. Mower, J.P., Case, A.L., Floro, E.R. and Willis, J.H. (2012) 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*, **4**, 670-686.

115. Folkerts, O. and Hanson, M.R. (1991) The male sterility-associated pcf gene and the normal atp9-1 gene in Petunia are located on different mitochondrial DNA molecules. *Genetics*, **129**, 885-895.

116. Boeshore, M.L., Lifshitz, I., Hanson, M.R. and Izhar, S. (1983) Novel composition of mitochondrial genomes in Petunia somatic hybrids derived from cytoplasmic male sterile and fertile plants. *Molecular and General Genetics MGG*, **190**, 459-467.

117. Gallagher, L.J., Betz, S.K. and Chase, C.D. (2002) Mitochondrial RNA editing truncates a chimeric open reading frame associated with S male-sterility in maize. *Curr Genet*, **42**, 179-184.

118. Arnold, C. (1967) Berichte der deutschen Botanischen Gesellschaft. *Gustav Fischer Verlag* Villengang 2, D-07745 Jena, Germany, Vol. 80, pp. 124-&.

119. Arnold, C.G. (1970) Extrakaryotic inheritance of pollen sterility in *Oenothera*. *Theor Appl Genet*, **40**, 241-244.

120. Schwemmle, J. (1938) Genetische und zytologische Untersuchungen an Eu-Oenotheren. *Zeitschrift für induktive Abstammungs-und Vererbungslehre*, **75**, 486-660.

121. Greiner, S., Wang, X., Herrmann, R.G., Rauwolf, U., Mayer, K., Haberer, G. and Meurer, J. (2008) The complete nucleotide sequences of the five genetically distinct plastid genomes of *Oenothera*, subsection *Oenothera*: I. Sequence evaluation and plastome evolution. Nucleic Acids Res 36, 2366-2378.

122. Clifton, S.W., Minx, P., Fauron, C.M., Gibson, M., Allen, J.O., Sun, H., Thompson, M., Barbazuk, W.B., Kanuganti, S., Tayloe, C. *et al.* (2004) Sequence and comparative analysis of the maize NB mitochondrial genome. *Plant Physiol*, **136**, 3486-3503.

2.2 Paper 2: Chloroplast competition is controlled by lipid biosynthesis in evening primroses

Johanna Sobanski[a], Patrick Giavalisco[b], Axel Fischer[c], Julia M. Kreiner[d], Dirk Walthe [c], Mark Aurel Schöttler[a], Tommaso Pellizzer[a], Hieronim Golczyk[e], Toshihiro Obata[f], Ralph Bock[a], Barbara B. Sears[g], and Stephan Greiner[a]

[a] Department Organelle Biology, Biotechnology and Molecular Ecophysiology, Max Planck Institute of Molecular Plant Physiology, 14476 Potsdam-Golm, Germany

[b] Department Molecular Physiology, Max Planck Institute of Molecular Plant Physiology, 14476 Potsdam-Golm, Germany

[c] Department Metabolic Networks, Max Planck Institute of Molecular Plant Physiology, 14476 Potsdam-Golm, Germany

[d] Department of Ecology & Evolutionary Biology, University of Toronto, ON M5S 3B2, Canada

[e] Department of Molecular Biology, Institute of Biotechnology, John Paul II Catholic University of Lublin, Konstantynów 1I, 20-708, Poland

[f] Center for Plant Science Innovation and Department of Biochemistry, University of Nebraska-Lincoln, Lincoln, NE 68588

[g] Department of Plant Biology, Michigan State University, East Lansing, MI 48824-1312

Journal: Proceedings of the National Academy of Sciences of the United States of America

Received: 6 July 2018

Accepted: 4 February 2019

Published: 4 March 2019

doi: 10.1073/pnas.1811661116

Chloroplast competition is controlled by lipid biosynthesis in evening primroses

Johanna Sobanski[a], Patrick Giavalisco[b], Axel Fischer[c], Julia M. Kreiner[d], Dirk Walther[c], Mark Aurel Schöttler[a], Tommaso Pellizzer[a], Hieronim Golczyk[e], Toshihiro Obata[f], Ralph Bock[a], Barbara B. Sears[g], and Stephan Greiner[a,1]

[a]Department Organelle Biology, Biotechnology and Molecular Ecophysiology, Max Planck Institute of Molecular Plant Physiology, 14476 Potsdam-Golm, Germany; [b]Department Molecular Physiology, Max Planck Institute of Molecular Plant Physiology, 14476 Potsdam-Golm, Germany; [c]Department Metabolic Networks, Max Planck Institute of Molecular Plant Physiology, 14476 Potsdam-Golm, Germany; [d]Department of Ecology & Evolutionary Biology, University of Toronto, ON M5S 3B2, Canada; [e]Department of Molecular Biology, Institute of Biotechnology, John Paul II Catholic University of Lublin, Konstantynów 1I, 20-708, Poland; [f]Center for Plant Science Innovation and Department of Biochemistry, University of Nebraska-Lincoln, Lincoln, NE 68588; and [g]Department of Plant Biology, Michigan State University, East Lansing, MI 48824-1312

Edited by Daniel B. Sloan, Colorado State University, Fort Collins, CO, and accepted by Editorial Board Member Joseph R. Ecker February 4, 2019 (received for review July 6, 2018)

In most eukaryotes, organellar genomes are transmitted preferentially by the mother, but molecular mechanisms and evolutionary forces underlying this fundamental biological principle are far from understood. It is believed that biparental inheritance promotes competition between the cytoplasmic organelles and allows the spread of so-called selfish cytoplasmic elements. Those can be, for example, fast-replicating or aggressive chloroplasts (plastids) that are incompatible with the hybrid nuclear genome and therefore maladaptive. Here we show that the ability of plastids to compete against each other is a metabolic phenotype determined by extremely rapidly evolving genes in the plastid genome of the evening primrose *Oenothera*. Repeats in the regulatory region of *accD* (the plastid-encoded subunit of the acetyl-CoA carboxylase, which catalyzes the first and rate-limiting step of lipid biosynthesis), as well as in *ycf2* (a giant reading frame of still unknown function), are responsible for the differences in competitive behavior of plastid genotypes. Polymorphisms in these genes influence lipid synthesis and most likely profiles of the plastid envelope membrane. These in turn determine plastid division and/or turnover rates and hence competitiveness. This work uncovers cytoplasmic drive loci controlling the outcome of biparental chloroplast transmission. Here, they define the mode of chloroplast inheritance, as plastid competitiveness can result in uniparental inheritance (through elimination of the "weak" plastid) or biparental inheritance (when two similarly "strong" plastids are transmitted).

biparental inheritance | selfish cytoplasmic elements | correlation mapping | acetyl-CoA carboxylase | *ycf2*

Most organelle genomes are inherited from the mother (1, 2), but biparental transmission of plastids has evolved independently multiple times. Approximately 20% of all angiosperms contain chloroplasts in the pollen generative cell, indicating at least a potential for biparental transmission (2–4). Although reasons for this are controversial (2, 5–9), a genetic consequence of the biparental inheritance patterns is a genomic conflict between the two organelles. When organelles are transmitted to the progeny by both parents, they compete for cellular resources. As the plastids do not fuse and hence their genomes do not undergo sexual recombination, selection will favor the organelle genome of the competitively superior plastid (2, 10–13). In a population, the ensuing "arms race" can lead to evolution and spread of selfish or aggressive cytoplasmic elements that potentially could harm the host cell. The mechanisms and molecular factors underlying this phenomenon are elusive, yet there is solid evidence for a widespread presence of competing organelles (2). Heteroplasmic cells can be created from cell fusion events, mutation of organelle DNA, or sexual crosses (14–17), but only very few cases have been studied in some detail in model organisms. One them is *Drosophila*, in which mitochondrial competition experiments can be set up via cytoplasmic microinjections (18, 19). Another is the evening primrose (genus *Oenothera*) (20, 21). This plant genus is certainly the model organism of choice to study chloroplast competition; the original theory of "selfish" cytoplasmic elements is based on evening primrose genetics (10): in the *Oenothera*, biparental plastid inheritance is the rule (22), and the system is a prime example of naturally occurring aggressive chloroplasts (2, 10). Based on extensive crossing studies, five genetically distinguishable chloroplast genome (plastome) types were shown to exist. Those were designated by Roman numerals (I–V) and grouped into three classes according to their inheritance strength or assertiveness rates in crosses (strong, plastomes I and III; intermediate, plastome II; and weak, plastomes IV and V), reflecting their ability to outcompete a second chloroplast genome in the F1 generation upon biparental transmission (20, 22, 23). The plastome types were initially identified based on their (in)compatibility with certain nuclear genomes (24, 25). Strong plastomes provide "hitchhiking" opportunities for

Significance

Plastids and mitochondria are usually uniparentally inherited, typically maternally. When the DNA-containing organelles are transmitted to the progeny by both parents, evolutionary theory predicts that the maternal and paternal organelles will compete in the hybrid. As their genomes do not undergo sexual recombination, one organelle will "try" to outcompete the other, thus favoring the evolution and spread of aggressive cytoplasms. The investigations described here in the evening primrose, a model species for biparental plastid transmission, have discovered that chloroplast competition is a metabolic phenotype. It is conferred by rapidly evolving genes that are encoded on the chloroplast genome and control lipid biosynthesis. Because of their high mutation rate, these loci can evolve and become fixed in a population very quickly.

Author contributions: J.S. performed the main experimental work; P.G., A.F., J.M.K., D.W., M.A.S., T.P., H.G., T.O., B.B.S., and S.G. provided supportive data; A.F. and S.G. developed the correlation mapping approach; J.M.K. implemented PGLS; J.S., P.G., A.F., J.M.K., D.W., M.A.S., T.P., H.G., T.O., R.B., B.B.S., and S.G. analyzed and discussed the data; B.B.S. and S.G. designed the study; J.S. and S.G. wrote the manuscript; and P.G., A.F., D.W., M.A.S., H.G., R.B., and B.B.S. participated in writing.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. D.B.S. is a guest editor invited by the Editorial Board.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. EU262890.2, EU262891.2, KT881170.1, KT881171.1, KT881172.1, KT881176.1, KU521375.1, KX687910.1, KX687913.1, KX687914.1, KX687915.1, KX687916.1, KX687917.1, and KX687918.1).

[1]To whom correspondence should be addressed. Email: greiner@mpimp-golm.mpg.de.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1811661116/-/DCSupplemental.

Published online March 4, 2019.

loci that result in incompatible or maladaptive chloroplasts, and these would be viewed as selfish cytoplasmic elements (2, 10). It was further shown that the loci, which determine the differences in competitive ability of the chloroplast, are encoded by the chloroplast genome itself (20, 22, 23).

Results

To pinpoint the underlying genetic determinants, we developed an association mapping approach that correlates local sequence divergence to a phenotype. We analyzed 14 complete chloroplast genomes from *Oenothera* WT lines, whose inheritance strength had been previously classified in exhaustive crossing analyses (20, 22, 26). This enabled us to correlate the experimentally determined inheritance strengths to sequence divergence in a given alignment window (*Materials and Methods* and *SI Appendix*, *SI Text*). The Pearson's and Spearman's applied correlation metrics associated the genetic determinants of inheritance strength with four major sites: the regulatory region of the fatty acid biosynthesis gene *accD* (promoter, 5′-UTR, and protein N terminus), the *origin of replication B* (*oriB*), *ycf1*, and *ycf2* (including its

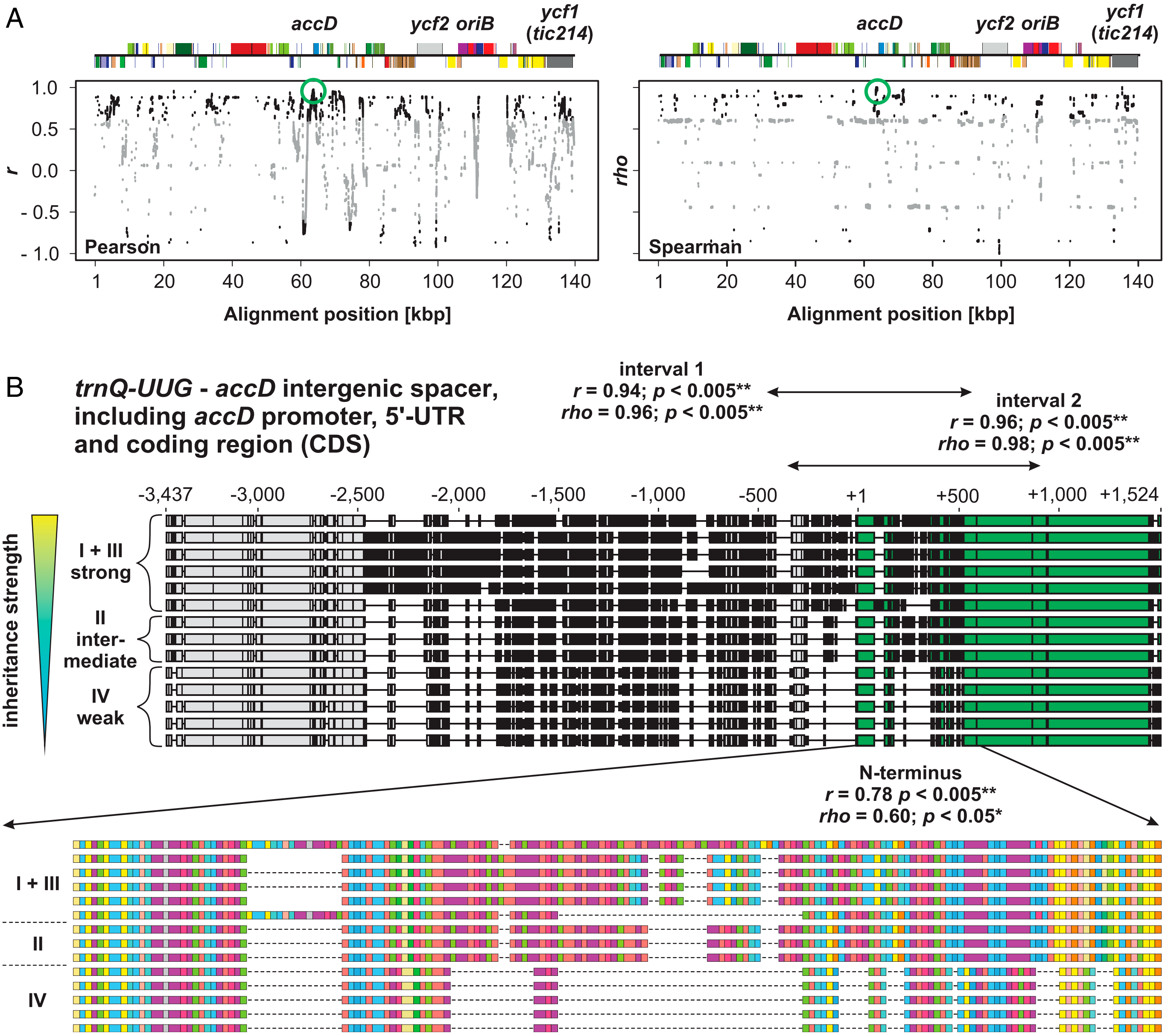


Fig. 1. Correlation mapping to identify chloroplast loci for inheritance strength in the WT chloroplast genomes. (*A*) Spearman and Pearson correlation to inheritance strength plotted against alignment windows of the WT plastomes. Relevant genes or loci with significant correlation are indicated in the linear plastome maps above. The region displayed in *B* is highlighted by green circles. Significant correlations ($P < 0.05$) are shown in black. Correlations to *k*-means classes are shown. Further details are provided in *Materials and Methods*, *SI Appendix*, *SI Text*, and main text. (*B*) Correlation to inheritance strength at the *accD* region in the WT plastomes. Individual sequences are sorted according to their competitive ability. Polymorphic regions are indicated in black, and thin lines represent gaps mostly resulting from deletions. (*Upper*) Alignment of the *trnQ–UUG–accD* intergenic spacer (−3,437 to −1) and the *accD* gene, including promoter, 5′-UTR, and CDS. The *accD* CDS, starting from +1, is highlighted in green. Regions marked by "interval 1" and "interval 2" display nearly absolute correlation to inheritance strength (*SI Appendix*, *SI Text* and Dataset S1). Note that these sequence intervals span the promotor region, the 5′-UTR, and the 5′-end of *accD*. All three are considered to play a regulatory role (29, 75–78). (*Lower*) Amino acid sequence of the AccD N terminus and correlation to inheritance strength. Colors indicated different amino acids. Most variation in the sequence is conferred by repeats encoding glutamic acid-rich domains marked in purple (*SI Appendix*, *SI Text*).

promoter/5′-UTR; Fig. 1 and *SI Appendix*, *SI Text*, Fig. S1, and Dataset S1). The *ycf1* and *ycf2* genes are two ORFs of unknown function; *ycf1* (or *tic214*) has tentatively been identified as an essential part of the chloroplast protein import machinery (27), but this function has been questioned (28). The *accD* gene encodes the β-carboxyltransferase subunit of the plastid-localized plant heteromeric acetyl-CoA carboxylase (ACCase). The enzyme is responsible for catalyzing the initial tightly regulated and rate-limiting step in fatty acid biosynthesis. The other three required ACCase subunits, α-carboxyltransferase, biotin-carboxyl carrier protein, and biotin carboxylase, are encoded by nuclear genes (29). The polymorphisms detected through our correlation mapping represent large insertions/deletions (indels), which are in frame in all coding sequences (CDSs; Fig. 1*B* and *SI Appendix*, *SI Text*, Figs. S2–S6, and Dataset S2). Several other polymorphisms in intergenic spacers of photosynthesis and/or chloroplast translation genes were correlated with plastome assertiveness rates (Fig. 1*A* and *SI Appendix*, *SI Text* and Dataset S1). However, both of those gene classes are unlikely to affect chloroplast inheritance, as other studies have shown that mutations in genes that result in chlorophyll-deficient chloroplasts do not alter their inheritance strengths in the evening primrose (20, 30, 31) (*SI Appendix*, *SI Text*). By contrast, the large ORFs of unknown function, the origins of replication, and a central gene in lipid metabolism such as *accD* (29) are serious candidates to encode factors involved in chloroplast competition.

Because of the lack of measurable sexual recombination in seed plant chloroplast genomes (2), our correlation mapping method first established associations of polymorphic loci that are fixed in the slow plastome type IV with weak inheritance strength, regardless of their functional relevance (*SI Appendix*, *SI Text*). Ideally, this problem can be partially circumvented if phylogenetic independence of the correlation between inheritance strength and a sequence's window comprising a candidate locus can be shown. Therefore, to correct for phylogeny in our correlation mapping analysis, we implemented phylogenetic generalized least squares (PGLS). This, however, yielded insignificant results after *P* value adjustment when applied to all sequence windows of the WT plastomes (*SI Appendix*, *SI Text*, Fig. S1, and Dataset S1). Because the correlation mapping results did not withstand controlling for phylogeny and multiple testing, we aimed to generated mutants from the candidate loci to test whether those loci affect inheritance strength. For this, we conducted a genetic screen for weak chloroplast mutants derived from the strong chloroplast genome I by employing a *plastome*

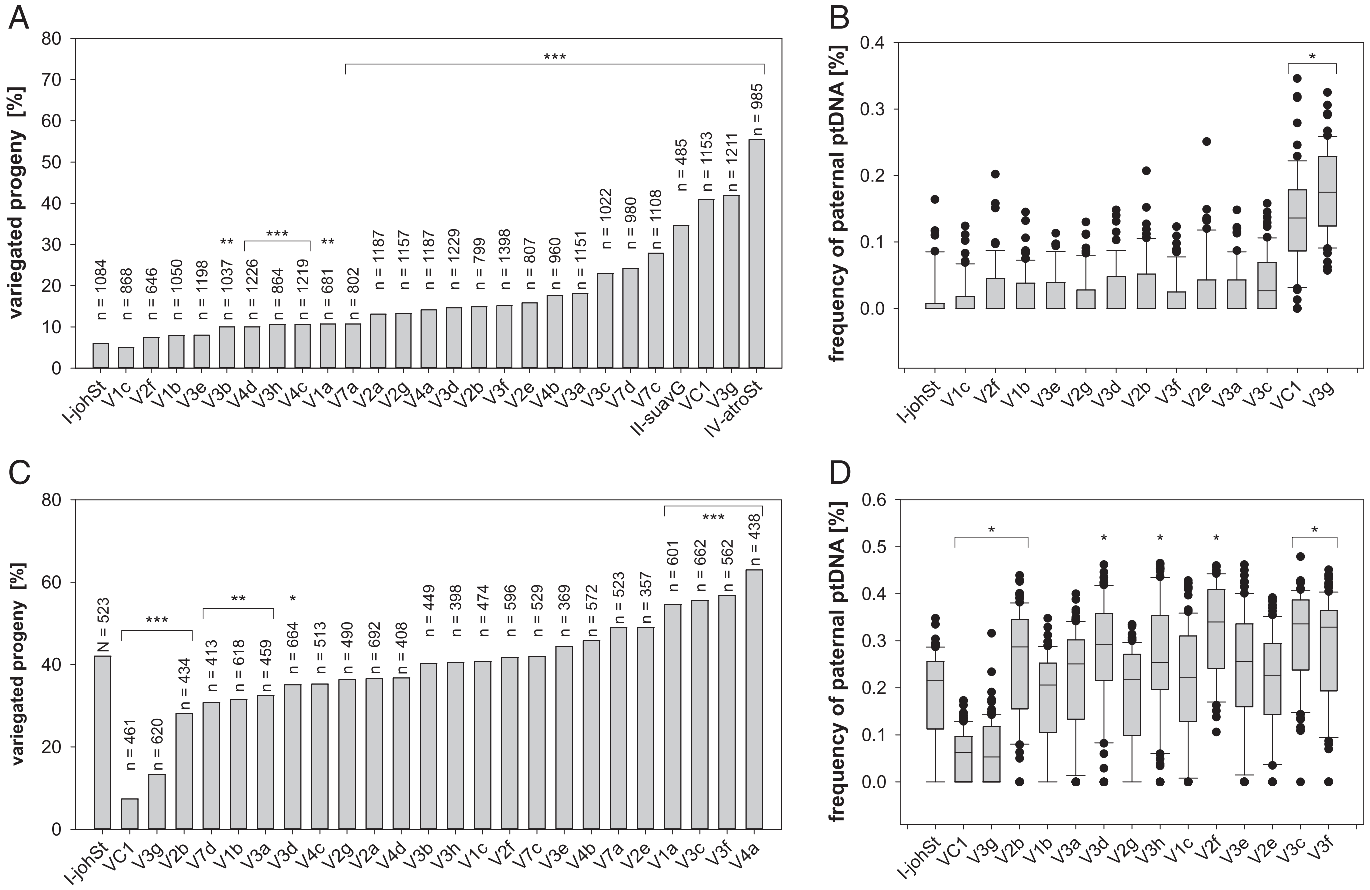


Fig. 2. Transmission efficiencies of plastome I variants and the WT plastomes I-johSt (strong), II-suavG (intermediate), and IV-atroSt (weak) as determined by crosses with I-chi/I-hookdV (strong) and IV-delta/IV-atroSt (weak). (*A* and *C*) Classical approach based on counting of variegated seedlings. (*B* and *D*) MassARRAY assay quantifying maternal and paternal ptDNA. WTs and green variants were crossed as the seed parent to I-chi as male parent (*A*) and as the pollen parent to IV-delta as female parent (*C*). The *X*-axis depicts the average percentage of variegated progeny (percentage biparental inheritance) obtained from three seasons (2013, 2014, and 2015). Fisher's exact test determined significance of differences between variants and their progenitor I-johSt (***$P < 0.0001$, **$P < 0.001$, *$P < 0.01$). Crosses of I-johSt and variants with I-hookdV as male (*B*) and with IV-atroSt as female parent (*D*). Box plots represent the transmission frequencies of the paternal plastomes measured by MassARRAY. To account for significant differences vs. I-johSt, Kruskal–Wallis one-way ANOVA on ranks was performed (*$P < 0.05$). Because of the detection threshold of the MassARRAY (5–10%), most variants show the same or slightly decreased transmission efficiency as their WT I-johSt (*B* and *D*). Only for the weak variants VC1 and V3g is the difference of the ratio of paternal and maternal ptDNA in the pool large enough to result in the detection of a significantly lower assertiveness rate in both crossing directions. Altogether, the classical approach using bleached chloroplast mutants gives more reliable results and allows a much finer discrimination of transmission efficiencies (cf. *A* vs. *B* and *C* vs. *D*). Further details are provided in *SI Appendix*, *SI Text*.

mutator (*pm*) allele (*Materials and Methods*). The *pm*-based mutagenesis approach yields indels in repetitive regions (32), similar to those identified by our association mapping (*SI Appendix, SI Text*). This led to the isolation of 24 plastome I variants with altered inheritance strength (Fig. 2 and *SI Appendix, SI Text*). As we selected for green and photosynthetically competent chloroplasts in the mutagenesis (*Materials and Methods*), none of those variants differed from the WT in their photosynthetic parameters, chloroplast size, or chloroplast volume per cell. The plants with the variant plastomes did not display any growth phenotype (*SI Appendix, SI Text* and Figs. S7–S9). Sequence analysis of 18 variants, spanning the range of observed variation in inheritance strength, revealed an average of seven mutation events per variant. The analysis included one additional variant (VC1) with more background mutations, which resulted from its isolation after being under the mutagenic action of the *pm* for several generations (*Materials and Methods* and *SI Appendix, SI Text*). Most of the *pm*-induced mutations are composed of single base pair indels at oligo(N) stretches in intergenic regions, i.e., not of functional relevance, or larger in-frame indels at the highly repetitive sites of *accD*, *ycf1*, *ycf2*, or *oriB* (*SI Appendix,* Figs. S2–S6, Table S1, and Dataset S2). Correlation analysis to inheritance strengths at these sites confirmed the relevance of *accD* and *ycf2* in chloroplast inheritance (r = 0.72, P = 0.05 for the 5′ end of AccD; r = 0.91, P < 0.0005 for the promotor/5′-UTR of *ycf2*; and r = 0.70, P < 0.005 for a mutated site in Ycf2; Fig. 3 and *SI Appendix, SI Text* and Figs. S2–S6 and S10). These findings were confirmed by two additional very weak mutants derived from the strong plastome III (*Materials and Methods* and *SI Appendix, SI Text*). Based on the full chloroplast genome sequences of these lines (*SI Appendix,* Dataset S2), it appeared that the promotor/5′-UTR of *accD* is affected in one of the lines. In addition, in both lines, the *ycf2* gene is most heavily mutated compared with all other (weak) materials sequenced so far. This strongly advocates for *ycf2* being involved in chloroplast competition (*SI Appendix,* Table S2). The very weak variants of plastome III, as well as correlation analysis of *oriB* and *ycf1* in the plastome I variants, did not support an involvement of these regions in the inheritance phenotype (Fig. 3 and *SI Appendix, SI Text*, Figs. S6 and S10, and Tables S1 and S2). Furthermore, the second replication origin (*oriA*) was found to be nearly identical within all sequenced WT or mutant plastomes.

These data argue against plastid DNA (ptDNA) replication, per se, being responsible for differences in chloroplast competitiveness. This conclusion is in line with previous analyses of the *Oenothera* replication origins, which had suggested that their variability does not correlate with the competitive strength of the plastids (33, 34) (*SI Appendix, SI Text*). We further confirmed this by determining the relative ptDNA amounts of chloroplasts with different inheritance strengths in a constant nuclear background. No significant variation of ptDNA amounts was observed over a developmental time course in these lines, thus excluding ptDNA stability and/or turnover as a potential mechanism (*SI Appendix, SI Text* and Fig. S11). Moreover, no significant differences in nucleoid numbers per chloroplast or nucleoid morphology was observed, as judged by DAPI staining (*SI Appendix, SI Text* and Figs. S12 and S13).

Next, we conducted a more detailed analysis of *accD* and *ycf2*. In a constant nuclear background, the weak WT plastome IV appeared to be an *accD* overexpressor compared with the strong WT plastome I, as judged from Northern blot analyses. However, this overexpression could not be detected in the plastome I variants that have a weakened competitive ability. Similar results were obtained for *ycf2*, which has an RNA of approximately 7 kb, reflecting the predicted size of the full-length transcript (*SI Appendix, SI Text* and Fig. S14). Interestingly, lower bands, probably reflecting transcript processing and/or degradation intermediates, differ between the strong WT plastome I and the weak WT plastome IV, with the latter being similar to the weak plastome I variants.

As these analyses did not allow conclusions about the functionality of AccD or Ycf2 in our lines, we decided to determine the ACCase activity in chloroplasts isolated from a constant nuclear background. As shown in Fig. 4*A*, the presence of large mutations/polymorphisms in the N terminus of the *accD* reading frame co-occurs with higher levels of ACCase enzymatic activity. Surprisingly, mutations/polymorphisms in *ycf2* also have an influence on ACCase activity, as revealed by lines that are not affected by mutations in *accD*. The molecular nature of this functional connection between Ycf2 and ACCase activity is currently unclear, although Ycf2 shares weak homologies to the FtsH protease (35, 36), which has a regulatory role in lipopolysaccharide synthesis in *Escherichia coli* (37). In any case, a simple relationship of ACCase activity and competitive ability of plastids is not present, but alterations in the earliest step of fatty acid biosynthesis can conceivably result in various changes in lipid metabolism (*SI Appendix, SI Text*).

To examine the alterations in lipid biosynthesis, we determined the lipid composition of seedlings harboring strong and weak WT chloroplasts as well as the variants that differ in chloroplast inheritance strength (*SI Appendix, SI Text* and Table S3). Then, we employed a least absolute shrinkage and selection operator (LASSO) regression model to predict competitive ability of a given chloroplast (Fig. 4*B*). As chloroplast inheritance strength is independent of photosynthetic competence (*SI Appendix, SI Text*), we included pale lines. The aim was to enrich the lipid signal responsible for inheritance strength, i.e., to deplete for structural lipids of the photosynthetic thylakoid membrane, which is a major source of chloroplast lipids (38). Indeed, of 102 lipids analyzed, 20 predictive ones for inheritance strengths were identified (Fig. 4*C* and *SI Appendix, SI Text* and Table S4). Strikingly, the signal is independent of greening (i.e., an intact thylakoid membrane system and the photosynthetic capacity), which is in line with the genetic data (*SI Appendix, SI Text*). This result hints at the chloroplast envelope determining assertiveness rates, a view that is supported by the fact that half of the predictive lipids come from lipid classes present in plastidial membranes and abundant in the chloroplast envelope, such as MGDG, DGDG, PG, and PC (39). The remaining predictive lipids mostly represent storage lipids (TAG). This might be a result of an altered fatty acid pool (*SI Appendix, SI Text*). Statistical significance of enrichment of a given class could not be established as a result of low numbers (*SI Appendix, SI Text* and Tables S4 and S5), although, especially in the lipid class PC, which is the dominant phospholipid class in the chloroplast outer envelope (40), 4 of 13 detected lipids were found to be predictive. In the chloroplast, PCs are specific to the envelope membrane and essentially absent from thylakoids. This makes it very likely that the lipid composition of the envelope membrane affects chloroplast competition. A possible explanation could be that strong and weak chloroplasts differ in division rates, for example, as a result of differential effectiveness in recruiting chloroplast division rings, which are anchored by membrane proteins (41). Alternatively, chloroplast stability might depend on envelope membrane composition. Strictly speaking, we cannot exclude a (further) involvement of extraplastidial membranes to the inheritance phenotype (most of the total cellular PCs are found in the ER and the plasma membrane) (38), but this would require a much more complicated mechanistic model.

Discussion

The present work explains plastid competition following biparental inheritance from a mechanistic perspective and points to genetic loci that appear to be responsible for these differences. Moreover, as chloroplast competition can result in uniparental inheritance through the elimination of weak

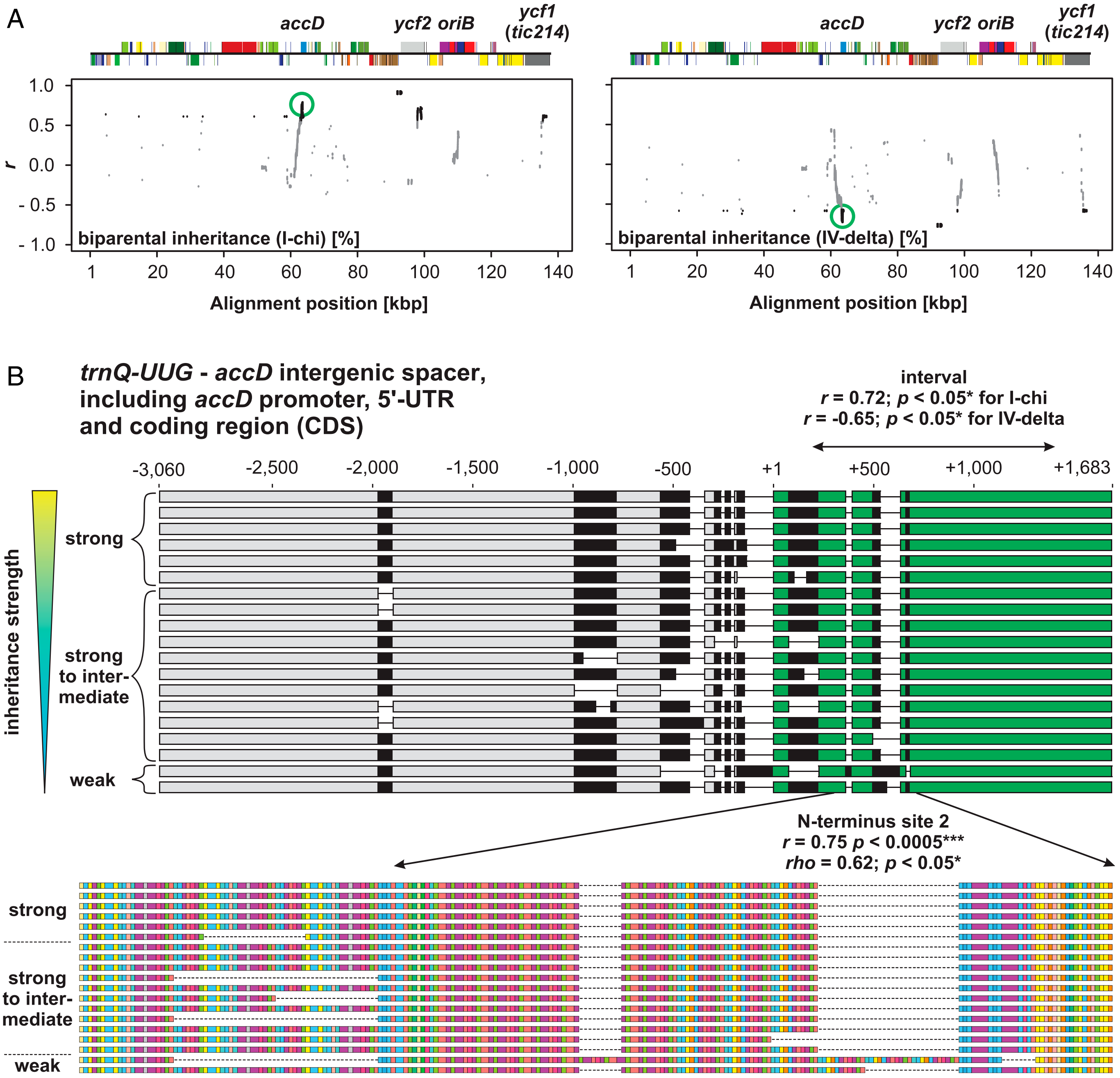


Fig. 3. Correlation mapping to identify chloroplast loci for inheritance strength in the plastome I variants. (*A*) Pearson correlation to inheritance strength plotted against alignment windows of the variants' plastomes. Relevant genes or loci with significant correlation are indicated in the linear plastome maps above. The region displayed in *B* is highlighted by green circles. Significant correlations ($P < 0.05$) are shown in black. Correlations to I-chi and IV-delta crosses (Fig. 2 *A* and *C*) are shown. Further details are provided in *Materials and Methods*, *SI Appendix*, *SI Text*, and main text. (*B*) Correlation to inheritance strength at the *accD* region in the variants' plastomes. Individual sequences are sorted according to their competitive ability. Polymorphic regions are indicated in black, and thin lines represent gaps mostly resulting from deletions. (*Upper*) Alignment of the *trnQ–UUG–accD* intergenic spacer (–3,060 to –1) and the *accD* gene, including promoter, 5′-UTR, and CDS. The *accD* CDS, starting from +1, is highlighted in green. The region marked by "interval" within the *accD* CDS displays high correlation to inheritance strength (*SI Appendix*, *SI Text* and Dataset S1). (*Lower*) Amino acid sequence of the AccD N terminus and correlation to inheritance strength. Colors indicate different amino acids. Most variation in the sequence is conferred by repeats encoding glutamic acid-rich domains marked in purple (*SI Appendix*, *SI Text*).

chloroplasts (9, 20, 30), at least for the evening primrose, the mechanistic explanation can be extended to uniparental transmission. Because approximately 20% of all angiosperms contain ptDNA in the sperm cell, it is likely that this mechanism is present in other systems (3, 5, 42). However, it should be emphasized that uniparental inheritance can be achieved by multiple mechanisms (2), and nuclear loci controlling the mode of organelle inheritance still need to be identified.

Arguably, the most surprising finding from our work is the discovery that chloroplast competition in evening primroses is essentially a metabolic phenotype and not directly connected to ptDNA replication or copy number (43). The underlying molecular loci are rapidly evolving genes throughout the plant kingdom. In general, the Ycf1 and Ycf2 proteins as well as the N terminus of AccD are highly variable in angiosperms (44–46). Interactions between them were repeatedly suggested (46, 47);

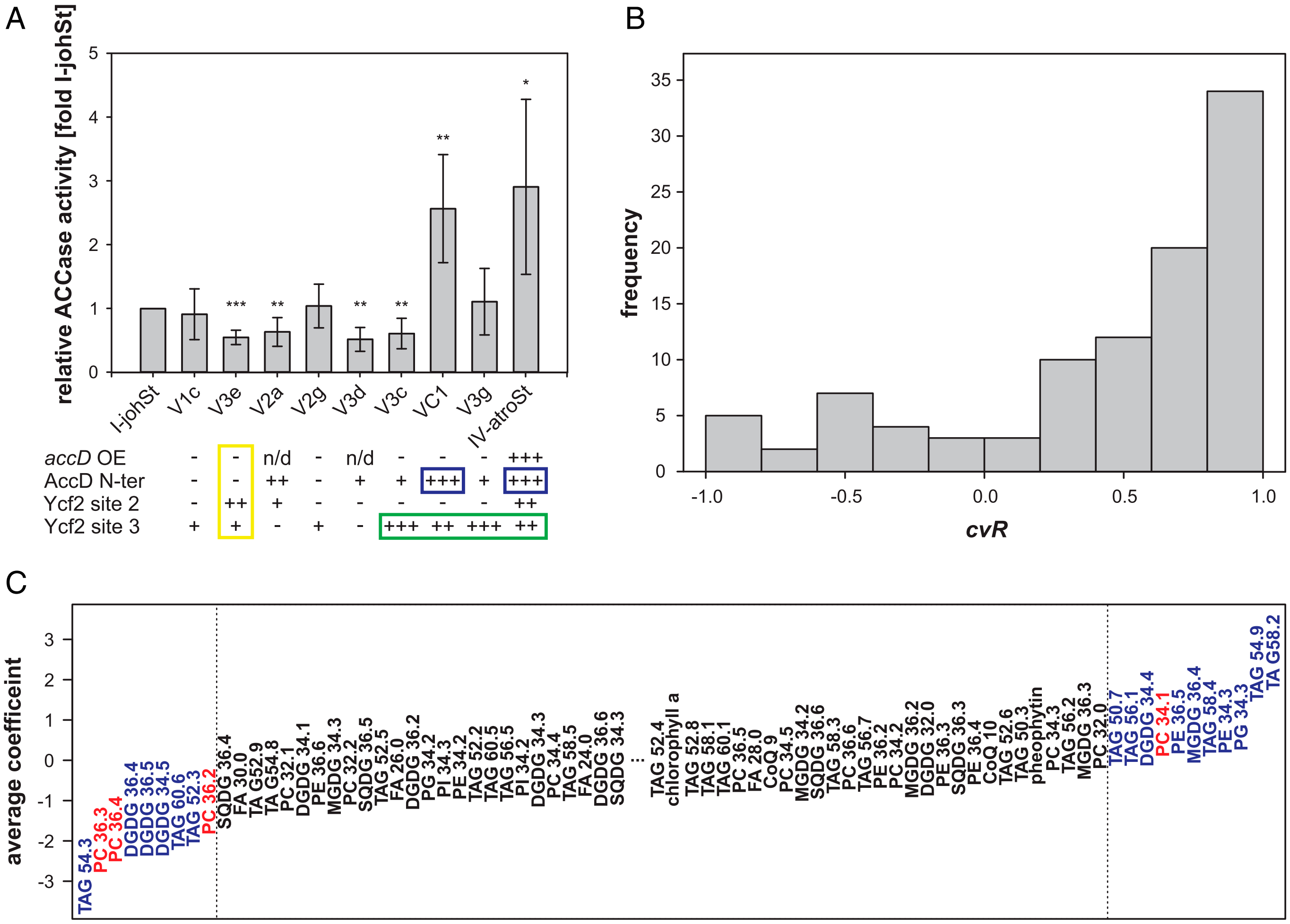


Fig. 4. ACCase activity and prediction of inheritance strengths based on lipid level. (*A*) ACCase activity of chloroplasts with difference inheritance strengths sorted according to competitive abilities (percentage biparental inheritance I-chi; cf. Fig. 2); *accD* OE, *accD* overexpressor; AccD N-ter, AccD N terminus (*SI Appendix, SI Text*). Compared with I-johSt, −, not affected; +, mildly affected; ++, intermediately affected; +++, strongly affected; n/d, not determined (cf. *SI Appendix*, Figs. S1–S5 and S14). Note the influence of mutations in *ycf2* on AccD activity in a nonmutated *accD* background (yellow box), the striking co-occurrence of mutations in the AccD N terminus with ACCase activity (blue boxes), and co-occurrence of large mutations/polymorphisms in site 3 of Ycf2 with inheritance strengths (green box). Significance of difference compared with I-johSt was calculated by using a paired two-tailed *t test* (***$P < 0.0005$, **$P < 0.005$, and *$P < 0.05$). (*B*) Histogram of Pearson correlation coefficient *cvR* between actual and predicted inheritance strengths obtained from 100 cross-validation runs. Note the shift toward 1.0, pointing to the predictive power of lipid data in the LASSO regression model. (*C*) Average linear LASSO model coefficients of the 102 lipids/molecules available for analysis (cf. *SI Appendix*, Table S4). The 20 identified predictive lipids are marked in color. PCs, the dominant phospholipids of the chloroplast outer envelope, are indicated in red. In general, large absolute values show predictive power negatively or positively correlating with inheritance strength. Predictive lipids were designated when their absolute average weight was greater than 1 SD of all weight values obtained for the 102 lipids. For better presentability, lipids/molecules with an absolute average weight ~0.00 were removed from the figure.

e.g., the loss of *ycf1*, *ycf2*, and *accD* genes from the plastome of grasses (48), as well as their common retention in many plastomes of nonphotosynthetic parasites (49), provides room for speculation about functional interaction (46). In addition, *Silene*, a genus in which biparental chloroplast transmission is described (50), exhibits accelerated evolution of *accD* (51). In contrast, *Campanulastrum*, again displaying biparental chloroplast inheritance (9), has lost *accD* from the chloroplast genome, and its *ycf2* is under accelerated evolution (52). In pea, which usually shows uniparental chloroplast inheritance, one strain with biparental chloroplast inheritance exists, and biparental transmission is accompanied by chloroplast/nuclear genome incompatibility in the resulting hybrid (53), with repeats in *accD* implicated as being responsible (54). Analysis of a very similar chloroplast/nuclear genome incompatibility in *Oenothera* also identifies repeats in *accD* as causative (55). Hence, correlation of the highly divergent *accD* gene (and/or *ycf2*) with the presence or absence of biparental inheritance and/or chloroplast incompatibility is certainly worth investigating on a broader phylogenetic scale. Connected with that, the *accD* plastome variants represent a superb material to study ACCase regulation and those of its nuclear counterparts. Interaction effects in different nuclear genomic background should be observable.

In the plastome of the evening primrose, the loci involved in chloroplast competition are very sensitive to replication slippage as a result of the presence of highly repetitive sequences, and that process appears to be the major mechanism of spontaneous chloroplast mutation (31, 44) (*SI Appendix, SI Text*). This result is somewhat reminiscent of recent findings in *Drosophila* in which sequence variation in the noncoding regulatory region of the mitochondrial genome, containing the origins of replication, was associated with different competitive abilities. Similar to *Oenothera*, these sequences are highly repetitive and hypervariable and contribute to cytoplasmic drive. In *Drosophila*, they are among the most divergent ones in Metazoa, pointing to their positive selection (19) (*SI Appendix, SI Text*).

Our analyses show that, as a result of their high mutation rates, cytoplasmic drive loci can evolve and become fixed in a population very quickly: in *Oenothera*, the current view on the evolutionary history of the plastome is IV → II → I/III (56), with an estimated divergence time of ~830,000 y (44). This is largely based on chloroplast/nuclear incompatibility and can explain why plastomes IV and II (although compatible with the nuclear genomes of many *Oenothera* species) were outcompeted by the newly evolved aggressive plastomes I and III in many species. Interestingly, the extant plastomes I and III do not form a phylogenetic clade. The recently evolved strong plastome I clusters with the intermediately strong plastome II, with recently evolved plastome III as outgroup (*SI Appendix*, Fig. S1*B*). The divergence time of II and III, however, is only approximately 420,000 y (44). Hence, evolution and fixation of an aggressive cytoplasm happened twice independently within a very short time frame.

Materials and Methods

Plant Material. Throughout this work, the terms "*Oenothera*" or "evening primrose" refer to subsection *Oenothera* (genus *Oenothera* section *Oenothera*) (56). Plant material used here is derived from the *Oenothera* germplasm collection harbored at the Max Planck Institute of Molecular Plant Physiology (Potsdam-Golm, Germany) (57). Part of this collection is the so-called Renner Assortment, a collection of lines thoroughly characterized by the genetic school of Otto Renner (22, 58). Therefore, the original material of Schötz (21, 26), which determined the classes of chloroplast replication speeds, was available for correlation mapping (as detailed later). For all other genetic or physiological work presented here, the nuclear genetic background of *Oenothera elata* subsp. *hookeri* strain johansen Standard (59) was used. The employed chloroplast (genomes) are native or were introgressed into that race by author S.G. or Wilfried Stubbe. The WT chloroplast genomes (I-johSt, I-hookdV, II-suavG, and IV-atroSt) are compatible with, and hence green when combined with, the johansen Standard nucleus. The chloroplast genome III-lamS confers a reversible bleaching, so-called virescent, phenotype in this genetic background (56, 60) (*SI Appendix*, Fig. S15). The white chloroplast mutants I-chi and IV-delta (*SI Appendix*, Fig. S15) are part of the huge collection of spontaneous plastome mutants compiled by Stubbe and coworkers (31, 61, 62). Both mutants harbor a similar single locus mutation in the *psaA* gene (encoding a core subunit of photosystem I) and derive from the strong and weak WT plastomes I-hookdV and IV-atroSt, respectively (31). *SI Appendix*, Tables S6–S11 provide a summary of all strains and origins of the chloroplast genomes, including the *pm*-induced variants that are subsequently described.

Plastome Mutator Mutagenesis. The *pm* line is a descendant of the original isolate E-15–7 of Melvin D. Epp. The nuclear *pm* allele was identified after an ethyl methanesulfonate mutagenesis (63) in johansen Standard. When homozygous, the *pm* causes a 200–1,000× higher rate of chloroplast mutants compared with the spontaneous frequency. The underlying mutations mostly represent indels resulting from replication slippage events (32, 63–66).

Johansen Standard plants newly restored to homozygosity for the nuclear *pm* allele (*pm*/*pm*) were employed to mutagenize the chloroplast genome (I-johSt) as described previously (35). Homozygous *pm* plants were identified when new mutant chlorotic sectors were observed on them. On those plants, flowers on green shoots were backcrossed to the WT *PM* allele as pollen donor. In the resulting *pm*/*PM* populations, the chloroplast mutations were stabilized. This led (after repeated backcrosses with the *PM* allele and selection with appropriate markers against the paternal chloroplast) to homoplasmic green variants derived from the strong plastome I-johSt. The variants differ by certain indels or combination of indels, and the material was designated V1a, V1b, V2a, etc., where "V" stands for variant, the Arabic number for the number of the backcrossed plant in the experiment, and the small Latin letter for the shoot of a given plant. An additional line, named VC1, was derived from a similar *pm* mutagenesis of I-johSt, but the mutagenesis was conducted over several generations. Therefore, VC1, which is also a green variant, carries a much larger number of background mutations than do variants V1a, V1b, V2a, etc. (*SI Appendix*, Table S10). The two variant chloroplast genomes III-V1 and III-V2 (*SI Appendix*, Table S11) have a derivation similar to VC1. They are derived from the strong WT chloroplast genome III-lamS, which displays a reversible bleaching (virescent phenotype) in the johansen Standard nuclear genetic background. To mutagenize this chloroplast genome, it was introgressed into the *pm*/*pm* background of johansen Standard by Wilfried Stubbe and self-pollinated for a number of generations. When stabilized with the *PM* allele, it still displayed a virescent phenotype that is comparable to the original WT plastome III-lamS (*SI Appendix*, Fig. S15). Both lines quite likely go back to the same ancestral *pm*/*pm*-johansen Standard III-lamS plant, i.e., experienced a common mutagenesis before separating, although the number of (independent) mutagenizing generations is unclear.

Determination of Plastid Inheritance Strength. In the evening primroses, biparental transmission of plastids shows maternal dominance, i.e., F1 plants are homoplasmic for the maternal chloroplast or heteroplasmic for the paternal and maternal chloroplasts. If, in such crosses, one of the chloroplasts is marked by a mutation, resulting in a white phenotype, the proportion of variegated (green/white; i.e., heteroplasmic) seedlings can be used to determine chloroplast inheritance strength (as percentage of biparental inheritance). Moreover, if, in such crosses, one of the crossing partners is kept constant, the inheritance strength of all tested chloroplasts with respect to the constant one can be determined (20, 21, 26, 30). For example, in the I-chi crosses (in which the strong white plastid is donated by the father, as detailed later), more variegated seedlings are found in the F1, indicating that more paternal (white) chloroplasts were able to outcompete the dominating maternal green chloroplasts. Hence, in this crossing direction, small biparental percentage values indicate strong (i.e., assertive) plastomes from the maternal parent and high biparental values indicate weak variants contributed by the maternal parent. The situation is reversed in the reciprocal cross in which the white chloroplast is donated by the mother, as is the case in the IV-delta crosses. Here, the weak white chloroplast is maternal, and strong green variants contributed by the pollen give high fractions of variegated plants in the F1, whereas low percentages of biparental progeny result when weak green variants are carried by the pollen donor.

Crossing Studies. All crossing studies between chloroplast genomes were performed in the constant nuclear background of the highly homozygous johansen Standard strain (as described earlier). Germination efficiency in all populations was 100% (*SI Appendix*, *Supplementary Materials and Methods*). Transmission efficiencies of the green plastome I variants (V1a, V1b, V2a, etc.) were determined by using the white chloroplast I-chi (strong inheritance strength) and IV-delta (weak inheritance strength) as crossing partners, respectively. This allows the determination of the inheritance strength of a given green chloroplast relative to a white one based on quantification of the biparental (variegated) progeny among the F1, as progeny that inherit chloroplasts from only the paternal parent are virtually nonexistent (*SI Appendix*, *SI Text*). The fraction of variegated (green/white) seedlings was assessed in the I-chi crosses in which the white mutant was contributed by the paternal parent. Similarly, variegated (white/green) seedlings were quantified in the IV-delta crosses in which the white mutant was donated by the maternal parent (21, 26, 30) (as described earlier). In the I-chi crosses, the green plastome I variants, as well as the WT chloroplast genomes I-johSt (strong inheritance strength; native in the genetic background of johansen Standard and the original WT chloroplast genome used for mutagenesis), II-suavG (intermediate inheritance strength), and IV-atroSt (weak inheritance strength) were crossed as female parent to I-chi in three following seasons (2013, 2014, and 2015). In the IV-delta crosses, green variants and I-johSt were crossed as male parent to IV-delta, again in three independent seasons, 2013, 2014, and 2015. From each cross of each season, randomized populations of 100–300 plants were grown twice independently, followed by visual assessment of the number of variegated seedlings/plantlets 14–21 d after sowing (DAS; I-chi crosses) or 7–14 DAS (IV-delta crosses). Based on these counts, the percentage of variegated progeny was calculated for each individual cross. To determine statistically significant differences between the transmission efficiencies of the plastome I variants and I-johSt, the numbers from all three seasons were summed for a particular cross and a Fisher's exact test was employed.

A very similar experiment was performed to determine the inheritance strength of the two variants III-V1 and III-V2 (*SI Appendix*, *SI Text*), which derive from the strong chloroplast genomes III-lamS. Here, in two independent seasons (2015 and 2016), the WT I-johSt was used as pollen donor to induce variegation between the maternal plastome III (giving rise to a virescent phenotype) and the green plastome I (native in the background of johansen Standard as described earlier).

To determine transmission efficiencies independent of white chloroplast mutants or other bleached material, the plastome I variants (including their WT I-johSt) were crossed to the green WT plastomes IV-atroSt (weak inheritance strength) as female parent and to I-hookdV (strong inheritance

GENETICS

strength) as male parent in two independent seasons (2013 and 2014). F1 progeny was harvested at 6 DAS by pooling 60–80 randomized seedlings, and the ratios of the plastome types in the pool were analyzed via MassARRAY (Agena Bioscience) as described in the following section.

MassARRAY: Multiplexed Genotyping Analysis Using iPlex Gold. SNP genotyping to distinguish plastome I-johSt and I-hookdV/I-chi or I-johSt and IV-atroSt/IV-delta and subsequent quantification of their plastome ratios in appropriate F1s was carried out with the MassARRAY system (Agena Bioscience). The system was used to analyze chloroplast transmission efficiencies in different crosses. For this, total DNA was prepared from 60–80 randomized pooled plantlets at 6 DAS. Then, 10 SNPs distinguishing the plastomes I-johSt and I-hookdV/I-chi (I/I assay) and 15 SNPs between I-johSt and IV-atroSt/IV-delta (I/IV assay) were selected. Two appropriate primers flanking the SNP and one unextended primer (binding an adjacent sequence to the SNP) were designed by using MassARRAY Assay Design v4.0 (Agena Bioscience). Primer sequences and SNPs and their positions in I-johSt are listed in *SI Appendix*, Table S12. Plastome regions were amplified in a 5-μL PCR containing PCR buffer (2 mM $MgCl_2$, 500 μM dNTP mix, 1 U HotStartTaq; Agena Bioscience), 10 ng DNA, and 10 (I/I assay) or 15 (I/IV assay) PCR primer pairs, respectively, at concentrations ranging from 0.5 to 2.0 μM. The reaction mix was incubated for 2 min at 95 °C in 96-well plates, followed by 45 cycles of 30 s at 95 °C, 30 s at 56 °C, and 60 s at 72 °C, and a final elongation for 5 min at 72 °C. Excess nucleotides were removed by adding 0.5 U shrimp alkaline phosphatase (SAP) enzyme and SAP buffer (Agena Bioscience), followed by an incubation for 40 min at 37 °C and 5 min at 85 °C. For the primer extension reaction, the iPLEX reaction mixture (containing Buffer Plus, Thermo Sequenase, and termination mix 96; Agena Bioscience) and, depending on the primer, the extension primers at a concentration of 7–28 μM were added. Sequence-specific hybridization and sequence-dependent termination were carried out for 30 s at 94 °C, followed by 40 cycles of 5 s at 94 °C plus five internal cycles of 5 s at 52 °C and 5 s at 80 °C, and finally 3 min at 72 °C. After desalting with CLEAN resin (Agena Bioscience), the samples were spotted on 96-pad silicon chips preloaded with proprietary matrix (SpectroCHIP; Agena Bioscience) by using the Nanodispenser RS1000 (Agena Bioscience). Subsequently, data were acquired with a MALDI-TOF mass spectrometer MassARRAY Analyzer 4 (Agena Bioscience) and analyzed with the supplied software. To identify significant differences in the frequencies of paternal ptDNA, Kruskal–Wallis one-way ANOVA on ranks was performed.

***k*-Means Clustering to Classify Inheritance Strength.** For the WT chloroplasts, inheritance strength was classified by using the biparental transmission frequencies (percentage of variegated plants in F1, as detailed earlier) of the chloroplasts "biennis white" and "blandina white" according to Schötz (26) (*SI Appendix*, *SI Text*). Both crossing series included the same 25 WT chloroplasts, 14 of which had fully sequenced genomes and were employed for correlation mapping (*SI Appendix*, Table S7 and as detailed later). The original data are provided in the study of Schötz (26) and summarized by Cleland (ref. 22, p. 180) and in *SI Appendix*, Table S13. Based on the two transmission frequencies, the WT plastomes were clustered by using the *k*-means algorithm with Euclidean distance as distance dimension. The optimal number of centers was calculated with the pamk function of the fpc package, as implemented in R v.3.2.1 (67). Strikingly, essentially the same three classes (strong, intermediate, and weak) were obtained that had been previously determined by Schötz (20, 22) (*SI Appendix*, *SI Text* and Fig. S16).

For the variants, we used the transmission frequencies from I-chi and IV-delta crosses obtained from this work (Fig. 2 and *SI Appendix*, *SI Text* and Table S10). As the data-driven determination of the optimal number of clusters (*k* = 2, as detailed earlier) does not reflect the biological situation, upon repeated *k*-means runs, we chose the number of centers with the best trade-off between lowest swapping rate of the samples between the clusters and the biological interpretability. This approach resulted in four classes (*SI Appendix*, *SI Text* and Fig. S16).

Correlation Mapping. For correlation mapping in WTs, 14 completely sequenced plastomes (GenBank accession nos. EU262890.2, EU262891.2, KT881170.1, KT881171.1, KT881172.1, KT881176.1, KU521375.1, KX687910.1, KX687913.1, KX687914.1, KX687915.1, KX687916.1, KX687917.1, and KX687918.1) with known inheritance strength (20, 22, 26) (*SI Appendix*, Table S7) were employed. The eight chloroplast genomes assigned to accession numbers starting with KU or KX were newly determined in the course of this work. Mapping of genetic determinants in the green variants was done in 18 fully sequenced mutagenized plastomes (V1a, V1b, V1c, V2a, V2b, V2g, V3a, V3b, V3c, V3d, V3e, V3f, V3g, V3h, V4b, V4c, V7a, and VC1) as well as their WT reference (I-johSt; GenBank accession no. AJ271079.4).

In both sequence sets, divergence at a given alignment window was correlated to the experimentally determined inheritance strengths of a chloroplast genome. For the WTs, inheritance strength was measured by using the paternal transmission frequencies (i.e., percentage of variegated plants in the F1) of the chloroplasts biennis white or blandina white according to Schötz (26) (*SI Appendix*, *SI Text*) or *k*-means classes combining the two datasets by clustering (*SI Appendix*, *SI Text* and Table S13 and as described earlier). For the variants, we used the transmission frequencies from the I-chi and IV-delta crosses determined in this work, as well as the *k*-means classes that were obtained from them (Fig. 2, *SI Appendix*, *SI Text* and Table S10, and as described earlier).

For correlation of these transmission frequencies to loci on the chloroplast genome, the redundant inverted repeat A (IR_A) was removed from all sequences. Then, plastomes were aligned with ClustalW (68) and the alignments were curated manually (*SI Appendix*, Dataset S2). Subsequently, by using a script in R v3.2.1 (67) (*SI Appendix*, Dataset S3), nucleotide changes (SNPs, insertions, and deletions) relative to a chosen reference sequence plastome [I-hookdV (KT881170.1) for WT set and I-johSt (AJ271079.4) for variant set; *SI Appendix*, *SI Text*] were counted window-wise by two approaches: (*i*) segmenting the reference sequence in overlapping windows by using a sliding-window approach with a window size of 1 kb and a step size of 10 bp, yielding a matrix of 13,912 × 13 (WT set) and 13,668 × 18 (variants), respectively; or (*ii*) defining regions of interest with correspondingly chosen window sizes. Then, Pearson's and Spearman's correlation coefficients were calculated between (*i*) the total count of nucleotide changes for every plastome in the aligned sequence window compared with the reference (i.e., total sequence divergence) and (*ii*) the determined inheritance strength of the plastomes (*SI Appendix*, Fig. S17 and Dataset S1). For the sliding-window approach, *P* values were adjusted for multiple testing by using Benjamini–Hochberg correction. To reduce the number of *P* value adjustments, adjacent alignment windows with identical count vectors were collapsed into one. To annotate the correlation mapping script output file, gene annotation of the consensus sequence (*SI Appendix*, Dataset S2) was converted into bed format (*SI Appendix*, Dataset S3) and combined with the correlation bins by using intersectBed and groupBy of the bedtools package (69). For visualization (Figs. 1*A* and 3*A* and *SI Appendix*, Figs. S1 and S10), correlation coefficients obtained for every alignment window were plotted as a function of the alignment position. Correlation values greater than the *P* value threshold (>0.05) are grayed out. Linear chloroplast genome maps were derived from the annotation of the consensus of both sequence sets (*SI Appendix*, Dataset S2) and drawn by OrganellarGenomeDRAW v1.2 (70) in a linear mode by using a user-defined configuration XML file. Alignments of selected plastome regions were visualized in Geneious v10.2.3 (71) and subsequently, as the output of OrganellarGenomeDRAW, edited in CorelDraw X8 (Corel).

ACCase Activity Assay. ACCase activity was measured in isolated chloroplast suspensions (72, 73) diluted to 400 μg chlorophyll per milliliter (*SI Appendix*, *Supplementary Materials and Methods*). To validate equilibration to chlorophyll, protein concentration using a Bradford assay (Quick Start Bradford 1× Dye Reagent; Bio-Rad; with BSA solutions of known concentrations as standards) and chloroplast counts per milliliter of suspension were determined for the same samples. For chloroplast counting, the suspension was further diluted 1:10, with 15 μL subsequently loaded on a Cellometer Disposable Cell Counting Chamber (Electron Microscopy Sciences) and analyzed under a Zeiss Axioskop 2 microscope (Zeiss). For each sample, six "B squares" were counted, and chloroplast concentration was calculated as chloroplasts per milliliter = 10 × average count per B square/4 × 10^{-6}. All three equilibration methods gave comparable results.

ACCase activity was measured as the acetyl-CoA–dependent fixation of $H^{14}CO_3^-$ into acid-stable products. For each plant line (I-johSt, V1c, V3e, V2a, V2g, V3d, V3c, VC1, V3g, and IV-atroSt), three independent chloroplast isolations (i.e., three biological replicates) were analyzed in triplicate, including individual negative controls (minus acetyl-CoA) for each measurement. A total of 10 μL of chloroplast suspensions were incubated with 40 μL of reagent solution, with a final concentration of 100 mM Tricine KOH, pH 8.2, 100 mM potassium chloride, 2 mM magnesium chloride, 1 mM ATP, 0.1 mM Triton X-100, 10 mM sodium bicarbonate, 0.5 mM acetyl-CoA, and 40 mM radioactively labeled sodium bicarbonate ($NaH^{14}CO_3$, ca. 4,000 dpm/nmol; Amersham) at room temperature for 20 min. For the negative control, acetyl-CoA in the reaction mixture was replaced by water. Reactions were stopped by adding 50 μL 2 M hydrochloric acid. The sample was transferred to a scintillation vial, and acid labile radioactivity (i.e., remaining $H^{14}CO_3^-$) was evaporated by heating for 20 min at 85 °C. After addition of 3 mL scintillation mixture (Rotiszint eco plus; Carl Roth), the acid-stable radioactivity from incorporation of $H^{14}CO_3^-$ (^{14}C dpm) was detected by liquid scintillation counter (LS6500; Beckman Coulter). ACCase activity is represented as the ^{14}C incorporation rate into acid-stable fraction (dpm per minute) calculated by dividing the total fixed radioactivity by 20 min. The rates in three replicated reactions were averaged, and corresponding values from negative control samples were subtracted and normalized by the number of chloroplasts to gain ACCase activity in individual samples. The average rates were calculated for each line. To combine all measurements, relative ACCase activities were calculated for each experiment as relative to the I-johSt line, and significant differences between each line and the WT were identified by using a two-tailed paired *t* test, followed by *P* value adjustment by using the Benjamini–Hochberg procedure.

Predictability of Inheritance Strength Based on Lipid-Level Data as Explanatory Variables. Lipidomics data from *Oenothera* seedlings of the strain johansen Standard, harboring chloroplast genomes with different assertiveness rates (*SI Appendix, Supplementary Materials and Methods*), were analyzed jointly to test for predictability of inheritance strength based on lipid levels. For this, 33 probes representing 16 genotypes whose chloroplast genomes ranged from inheritance strength class 1 to 5 (*SI Appendix, SI Text*) were measured in five replicates in three independent experimental series (*SI Appendix, SI Text* and Table S3). In this dataset, a total of 184 different lipids/molecules could be annotated (*SI Appendix*, Dataset S4 and as described earlier). To normalize across experiments, the data from each series were log-transformed and median-centered based on genotypes with inheritance strengths of 1, i.e., for every lipid/molecule, its median level across all "inheritance strengths = 1 genotypes" was determined and subtracted from all genotypes tested in the respective experimental series. Inheritance strength 1 was then selected to serve as a common reference across all three experimental series. Subsequently, the three experimental series were combined into a single set. Only those lipids/molecules for which level data were available across all three datasets were considered further, leaving 102 lipids/molecules for analysis (*SI Appendix*, Dataset S4)

LASSO regression model. Inheritance strength was predicted based on the median-centered lipid-level data by using LASSO, a regularized linear regression approach (74), as implemented in the "glmnet" R software package (R v3.2.1) (67). glmnet was invoked with parameter α set to 1 to perform LASSO regression (*SI Appendix*, Dataset S4). The penalty parameter λ was determined from the built-in cross-validation applied to training set data (i.e., all but two randomly selected genotypes) and set to the 1-SE estimate deviation from the optimal (minimal error) value and assuming Gaussian response type. All other parameters were taken as their default values.

Predictive lipids. As a regularized regression method, LASSO aims to use few predictor variables, which allows better identification of truly predictive lipids. Summarized from all 100 cross-validation runs performed, lipids/molecules were ranked by their mean linear model coefficients assigned to them in the LASSO regression, with their absolute value indicating influence strength and their sign indicating positive or negative correlation of their abundance to inheritance strength.

Test for enrichment of predictive lipids/molecules in lipid classes. Across all 100 cross-validation runs, the importance of each of the 102 molecules was assessed based on their average absolute weight factor (avgW) by which they entered the 100 LASSO models. Molecules with avgW of greater than 1 SD obtained across all 102 molecules were considered important. Then, all lipids/molecules were assigned to their respective class (MGDG, DGDG, SQDG, PG, PC, PI, PE, FA, PE, TAG, CoQ, chlorophyll, and pheophytin), and every class was tested for enrichment in the lipid/molecule set considered to be important. This was done by employing a Fisher's exact test, yielding *P* values and odds ratios. The *P* values express enrichment, and the odds ratio express the relative enrichment or depletion of a particular class among the set of important lipids.

ACKNOWLEDGMENTS. We thank Werner Dietrich and Wilfried Stubbe, who compiled the comprehensive *Oenothera* collection that allowed us to use the original material of the Renner school; John B. Ohlrogge for advice regarding the ACCase activity assay; Michael Tillich for help with the MassArray system; Stephen I. Wright for fruitful discussion on PGLS; the Max Planck Institute of Molecular Plant Physiology GreenTeam for their support; the Ornamental Plant Germplasm Centre for providing the line chicaginensis de Vries; and Liliya Yaneva-Roder for technical assistance. This research was supported by the Max Planck Society (S.G., P.G., D.W., M.A.S., and R.B.). DAPI fluorescent patterns were analyzed with microscopy equipment funded by Polish National Science Centre Grant 2015/19/B/N22/01692 (to H.G.).

1. Birky CW, Jr (2001) The inheritance of genes in mitochondria and chloroplasts: Laws, mechanisms, and models. *Annu Rev Genet* 35:125–148.
2. Greiner S, Sobanski J, Bock R (2015) Why are most organelle genomes transmitted maternally? *BioEssays* 37:80–94.
3. Hu Y, Zhang Q, Rao G, Sodmergen (2008) Occurrence of plastids in the sperm cells of *Caprifoliaceae*: Biparental plastid inheritance in angiosperms is unilaterally derived from maternal inheritance. *Plant Cell Physiol* 49:958–968.
4. Birky CW, Jr (1995) Uniparental inheritance of mitochondrial and chloroplast genes: Mechanisms and evolution. *Proc Natl Acad Sci USA* 92:11331–11338.
5. Zhang Q, Sodmergen (2010) Why does biparental plastid inheritance revive in angiosperms? *J Plant Res* 123:201–206.
6. Bendich AJ (2013) DNA abandonment and the mechanisms of uniparental inheritance of mitochondria and chloroplasts. *Chromosome Res* 21:287–296.
7. Reboud X, Zeyl C (1994) Organelle inheritance in plants. *Heredity* 72:132–140.
8. Jansen RK, Ruhlman TA (2012) Plastid genomes of seed plants. *Genomics of Chloroplasts and Mitochondria*, Advances in Photosynthesis and Respiration, eds Bock R, Knoop V (Springer, Dordrecht, The Netherlands), Vol 35, pp 103–126.
9. Barnard-Kubow KB, McCoy MA, Galloway LF (2017) Biparental chloroplast inheritance leads to rescue from cytonuclear incompatibility. *New Phytol* 213:1466–1476.
10. Grun P (1976) *Cytoplasmic Genetics and Evolution* (Columbia Univ Press, New York), 1st Ed, p 446.
11. Hoekstra RF (1990) Evolution of uniparental inheritance of cytoplasmic DNA. *Organizational Constrains of the Dynamics of Evolution*, eds Smith MJ, Vida J (Manchester Univ Press, Manchester, UK), pp 269–278.
12. Cosmides LM, Tooby J (1981) Cytoplasmic inheritance and intragenomic conflict. *J Theor Biol* 89:83–129.
13. Eberhard WG (1980) Evolutionary consequences of intracellular organelle competition. *Q Rev Biol* 55:231–249.
14. Hoekstra RF (2011) Nucleo-cytoplasmic conflict and the evolution of gamete dimorphism. *The Evolution of Anisogamy*, eds Togashi T, Cox PA (Cambridge Univ Press, Cambridge, UK), pp 111–130.
15. Rand DM, Haney RA, Fry AJ (2004) Cytonuclear coevolution: The genomics of cooperation. *Trends Ecol Evol* 19:645–653.
16. Barr CM, Neiman M, Taylor DR (2005) Inheritance and recombination of mitochondrial genomes in plants, fungi and animals. *New Phytol* 168:39–50.
17. Greiner S (2012) Plastome mutants of higher plants. *Genomics of Chloroplasts and Mitochondria*, Advances in Photosynthesis and Respiration, eds Bock R, Knoop V (Springer, Dordrecht, The Netherlands), Vol 35, pp 237–266.
18. De Stordeur E (1997) Nonrandom partition of mitochondria in heteroplasmic *Drosophila*. *Heredity (Edinb)* 79:615–623.
19. Ma H, O'Farrell PH (2016) Selfish drive can trump function when animal mitochondrial genomes compete. *Nat Genet* 48:798–802.
20. Kirk JTO, Tilney-Bassett RAE (1978) *The plastids. Their Chemistry, Structure, Growth and Inheritance* (Elsevier, Amsterdam), 2nd Ed.
21. Schötz F (1954) Über Plastidenkonkurrenz bei *Oenothera*. *Planta* 43:182–240.
22. Cleland RE (1972) *Oenothera–Cytogenetics and Evolution* (Academic, London), 1st Ed, p 370.
23. Gillham NW (1978) *Organelle Heredity* (Raven, New York), 1st Ed, p 602.
24. Stubbe W (1964) The role of the plastome in evolution of the genus *Oenothera*. *Genetica* 35:28–33.
25. Greiner S, Rauwolf U, Meurer J, Herrmann RG (2011) The role of plastids in plant speciation. *Mol Ecol* 20:671–691.
26. Schötz F (1968) Über die Plastidenkonkurrenz bei *Oenothera* II. *Biol Zentralbl* 87: 33–61.
27. Kikuchi S, et al. (2013) Uncovering the protein translocon at the chloroplast inner envelope membrane. *Science* 339:571–574.
28. Bölter B, Soll J (2017) Ycf1/Tic214 is not essential for the accumulation of plastid proteins. *Mol Plant* 10:219–221.
29. Salie MJ, Thelen JJ (2016) Regulation and structure of the heteromeric acetyl-CoA carboxylase. *Biochim Biophys Acta* 1861:1207–1213.
30. Chiu W-L, Stubbe W, Sears BB (1988) Plastid inheritance in *Oenothera*: Organelle genome modifies the extent of biparental plastid transmission. *Curr Genet* 13: 181–189.
31. Massouh A, et al. (2016) Spontaneous chloroplast mutants mostly occur by replication slippage and show a biased pattern in the plastome of *Oenothera*. *Plant Cell* 28: 911–929.
32. Stoike LL, Sears BB (1998) *Plastome mutator*-induced alterations arise in *Oenothera* chloroplast DNA through template slippage. *Genetics* 149:347–353.
33. Sears BB, Stoike LL, Chiu WL (1996) Proliferation of direct repeats near the *Oenothera* chloroplast DNA origin of replication. *Mol Biol Evol* 13:850–863.
34. Chiu W-L, Sears BB (1992) Electron microscopic localization of replication origins in *Oenothera* chloroplast DNA. *Mol Gen Genet* 232:33–39.
35. Wolfe KH (1994) Similarity between putative ATP-binding sites in land plant plastid ORF2280 proteins and the FtsH/CDC48 family of ATPases. *Curr Genet* 25:379–383.
36. De Las Rivas J, Lozano JJ, Ortiz AR (2002) Comparative analysis of chloroplast genomes: Functional annotation, genome-based phylogeny, and deduced evolutionary patterns. *Genome Res* 12:567–583.
37. Langklotz S, Baumann U, Narberhaus F (2012) Structure and function of the bacterial AAA protease FtsH. *Biochim Biophys Acta* 1823:40–48.
38. Kobayashi K, Wada H (2016) Role of lipids in chloroplast biogenesis. *Lipids in Plant and Algae Development*, eds Nakamura Y, Li-Beisson Y (Springer, Cham, Switzerland), pp 103–125.

39. Block MA, Dorne AJ, Joyard J, Douce R (1983) Preparation and characterization of membrane fractions enriched in outer and inner envelope membranes from spinach chloroplasts. II. Biochemical characterization. *J Biol Chem* 258:13281–13286.
40. Botella C, Jouhet J, Block MA (2017) Importance of phosphatidylcholine on the chloroplast surface. *Prog Lipid Res* 65:12–23.
41. Osteryoung KW, Pyke KA (2014) Division and dynamic morphology of plastids. *Annu Rev Plant Biol* 65:443–472.
42. Corriveau JL, Coleman AW (1988) Rapid screening method to detect potential biparental inheritacne of plastid DNA and results for over 200 angiosperm species. *Am J Bot* 75:1443–1458.
43. Nishimura Y, Stern DB (2010) Differential replication of two chloroplast genome forms in heteroplasmic *Chlamydomonas reinhardtii* gametes contributes to alternative inheritance patterns. *Genetics* 185:1167–1181.
44. Greiner S, et al. (2008) The complete nucleotide sequences of the five genetically distinct plastid genomes of *Oenothera*, subsection *Oenothera*: I. Sequence evaluation and plastome evolution. *Nucleic Acids Res* 36:2366–2378.
45. Wicke S, Schneeweiss GM, dePamphilis CW, Müller KF, Quandt D (2011) The evolution of the plastid chromosome in land plants: Gene content, gene order, gene function. *Plant Mol Biol* 76:273–297.
46. de Vries J, Sousa FL, Bölter B, Soll J, Gould SB (2015) YCF1: A green TIC? *Plant Cell* 27: 1827–1833.
47. Bölter B, Soll J (2016) Once upon a time–Chloroplast protein import research from infancy to future challenges. *Mol Plant* 9:798–812.
48. Jansen RK, et al. (2007) Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. *Proc Natl Acad Sci USA* 104:19369–19374.
49. Wicke S, Naumann J (2018) Molecular evolution of plastid genomes in parasitic flowering plants. *Advances in Botanical Research*, eds Chaw S-M, Jansen RK (Academic, New York), Vol 85, pp 315–347.
50. Newton WCF (1931) Genetical experiments with *Silene otities* and related species. *J Genet* 24:109–120.
51. Rockenbach K, et al. (2016) Positive selection in rapidly evolving plastid-nuclear enzyme complexes. *Genetics* 204:1507–1522.
52. Barnard-Kubow KB, Sloan DB, Galloway LF (2014) Correlation between sequence divergence and polymorphism reveals similar evolutionary mechanisms acting across multiple timescales in a rapidly evolving plastid genome. *BMC Evol Biol* 14:268.
53. Bogdanova VS, Galieva ER, Kosterin OE (2009) Genetic analysis of nuclear-cytoplasmic incompatibility in pea associated with cytoplasm of an accession of wild subspecies *Pisum sativum* subsp. *elatius* (Bieb.) Schmahl. *Theor Appl Genet* 118:801–809.
54. Bogdanova VS, et al. (2015) Nuclear-cytoplasmic conflict in pea (*Pisum sativum* L.) is associated with nuclear and plastidic candidate genes encoding acetyl-CoA carboxylase subunits. *PLoS One* 10:e0119835.
55. Ulbricht-Jones ES (2017) The virescent and narrow leaf phenotype of a plastome-genome-incompatible *Oenothera* hybrid is associated with the plastid gene *accD* and fatty acid synthesis. PhD thesis (Univ Potsdam, Potsdam, Germany).
56. Dietrich W, Wagner WL, Raven PH (1997) *Systematics of Oenothera Section Oenothera Subsection Oenothera (Onagraceae)* (American Society of Plant Taxonomists, Laramie, WY), 1st Ed, p 234.
57. Greiner S, Köhl K (2014) Growing evening primroses (*Oenothera*). *Front Plant Sci* 5:38.
58. Harte C (1994) *Oenothera–Contributions of a Plant to Biology* (Springer, Berlin), 1st Ed.
59. Cleland RE (1935) Cyto-taxonomic studies on certain Oenotheras from California. *Proc Am Philos Soc* 75:339–429.
60. Stubbe W (1989) *Oenothera*–An ideal system for studying the interaction of genome and plastome. *Plant Mol Biol Rep* 7:245–257.
61. Kutzelnigg H, Stubbe W (1974) Investigation on plastome mutants in *Oenothera*: 1. General considerations. *Subcell Biochem* 3:73–89.
62. Stubbe W, Herrmann RG (1982) Selection and maintenance of plastome mutants and interspecific genome/plastome hybrids from Oenothera. *Methods in Chloroplast Molecular Biology*, eds Edelman M, Hallick RB, Chua N-H (Elsevier, Amsterdam), pp 149–165.
63. Epp MD (1973) Nuclear gene-induced plastome mutations in *Oenothera hookeri*: I. Genetic analysis. *Genetics* 75:465–483.
64. Chang T-L, et al. (1996) Characterization of primary lesions caused by the plastome mutator of *Oenothera*. *Curr Genet* 30:522–530.
65. Chiu W-L, et al. (1990) *Oenothera* chloroplast DNA polymorphisms associated with plastome mutator activity. *Mol Gen Genet* 221:59–64.
66. Sears BB, Sokalski MB (1991) The *Oenothera* plastome mutator: Effect of UV irradiation and nitroso-methyl urea on mutation frequencies. *Mol Gen Genet* 229:245–252.
67. R Core Team (2015) R: A Language and Environment for Statistical Computing. Version 3.2.1 (R Foundation for Statistical Computing, Vienna).
68. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22:4673–4680.
69. Quinlan AR, Hall IM (2010) BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* 26:841–842.
70. Lohse M, Drechsel O, Kahlau S, Bock R (2013) OrganellarGenomeDRAW–A suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. *Nucleic Acids Res* 41:W575–W581.
71. Kearse M, et al. (2012) Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28: 1647–1649.
72. Hunter SC, Ohlrogge JB (1998) Regulation of spinach chloroplast acetyl-CoA carboxylase. *Arch Biochem Biophys* 359:170–178.
73. Thelen JJ, Ohlrogge JB (2002) The multisubunit acetyl-CoA carboxylase is strongly associated with the chloroplast envelope through non-ionic interactions to the carboxyltransferase subunits. *Arch Biochem Biophys* 400:245–257.
74. Tibshirani R (1996) Regression shrinkage and selection via the Lasso. *J R Stat Soc B* 58: 267–288.
75. Hirata N, Yonekura D, Yanagisawa S, Iba K (2004) Possible involvement of the 5′-flanking region and the 5′UTR of plastid accD gene in NEP-dependent transcription. *Plant Cell Physiol* 45:176–186.
76. Hajdukiewicz PT, Allison LA, Maliga P (1997) The two RNA polymerases encoded by the nuclear and the plastid compartments transcribe distinct groups of genes in tobacco plastids. *EMBO J* 16:4041–4048.
77. Swiatecka-Hagenbruch M, Liere K, Börner T (2007) High diversity of plastidial promoters in *Arabidopsis thaliana*. *Mol Genet Genomics* 277:725–734.
78. Kozaki A, Mayumi K, Sasaki Y (2001) Thiol-disulfide exchange between nuclear-encoded and chloroplast-encoded subunits of pea acetyl-CoA carboxylase. *J Biol Chem* 276:39919–39925.

2.3 Paper 3: Identification of the paramutated *SULFUREA* locus of tomato and release from epigenetic silencing by spontaneous reversion or genetic suppression

Britta Ehlert *, Axel Fischer *, and Ralph Bock [1]

Max-Planck-Institut für Molekulare Pflanzenphysiologie, Am Mühlenberg 1, D-14476 Potsdam-Golm, Germany

* These two authors contributed equally to this work.

[1] Corresponding author: rbock@mpimp-golm.mpg.de

Submission-ready draft version

Synopsis:

This work reports the identification of *SULFUREA*, a classical paramutated locus in tomato, provides insights into the mechanisms involved in paramutation and its suppression, and makes a dicotyledonous plant accessible as a new model to study the molecular basis of paramutation.

SUMMARY

Paramutation is a heritable epigenetic phenomenon involving inactivation of an expressed (paramutable) allele by a homologous inactive (paramutagenic) allele. One of the classical paramutated loci is the tomato *SULFUREA* locus where paramutation causes loss of photosynthesis. Although intensely studied for over 50 years, the identity of *SULFUREA* and the mechanism of its inactivation by paramutation have remained elusive. Here we identified the *SULFUREA* gene and show that the paramutation phenotype can be fully repressed by heterologous complementation with a paramutation-insensitive transgene derived from the orthologous *Arabidopsis* gene. Genome and methylome sequencing identified a striking difference in promoter methylation between the paramutagenic and the paramutable alleles. We also isolated several spontaneous revertants in which *SULFUREA* regained activity. The loss of silencing was somatically stable, but paramutation reappeared in the next generation, indicating resetting during meiosis. Finally, we characterized a suppressor mutant and found it to exhibit aberrant DNA methylation patterns likely caused by a genomic deletion fusing part of a chaperone gene to a histone deacetylase gene. Our findings shed new light on the role of chromatin structure and DNA methylation in paramutation and make a readily transformable dicotyledonous model amenable to studies into the molecular basis of transgenerational epigenetic inheritance.

Keywords: paramutation; epigenetics; *Solanum lycopersicum*; suppressor mutation; DNA methylation; DNA demethylation; *sulfurea*

INTRODUCTION

Paramutation is an epigenetic phenomenon that involves heritable changes in gene expression. It is characterized by a silencing-conferring interaction between a pair of homologous alleles: a paramutable allele and a paramutagenic allele. The paramutagenic allele is typically silent and can impose its silenced state onto the paramutable (i.e., susceptible) allele. The most intriguing feature of paramutation is that this allelic *trans*-inactivation converts the paramutable allele into a paramutagenic allele which then can silence other paramutable alleles when introduced by appropriate genetic crosses. Thus, paramutation can be rationalized as a "contagious" process in which a paramutagenic allele infects a paramutable allele which then itself becomes infectious (i.e., paramutagenic). For this reason, the phenomenon was initially referred to as "somatic conversion" (Renner, 1938), before the term paramutation became widely accepted (cf. Hagemann, 1958).

Paramutation was first discovered and studied in evening primrose (*Oenothera sp.*) by Otto Renner (Renner, 1938). Subsequently, paramutated loci were found and analyzed genetically in two other plant species: in maize (*Zea mays*) by Royal Alexander Brink and in tomato (*Solanum lycopersicum*, formerly *Lycopersicon esculentum*) by Rudolf Hagemann (for review, see, e.g., Hagemann and Berg, 1977; Chandler and Stam, 2004; Chandler and Alleman, 2008; Stam, 2009). Although for a long time only known in plants, it is now clear that paramutation phenomena also occur in animals and other eukaryotes (Rassoulzadegan et al., 2006; Chandler, 2007). The well-documented cases of paramutation nearly exclusively involve pigment changes that can be easily recognized when occurring somatically. It, therefore, can be assumed that paramutation is likely to be much more widespread and often goes undetected, for example, due to metabolic complementation by adjacent wild-type tissue.

Most studies into the mechanisms underlying paramutation have been conducted in maize. This is because, only in maize, the molecular identity of genes affected by paramutation is currently known (Hollick et al., 1997; Dorweiler et al., 2000; Sidorenko and Peterson, 2001). Structural analyses of paramutated loci in maize as well as analysis of chromatin structure and DNA methylation patterns provided strong support of a transcriptional gene silencing mechanism underlying the paramutation phenomenon (Sidorenko and Peterson, 2001; Lisch et al., 2002; Stam et al., 2002; Stam et al., 2002; Chandler and Stam, 2004; Belele et al., 2013). However, why silencing by paramutation is transgenerationally stable (i.e, heritable) and how it differs mechanistically from conventional types of transcriptional silencing (Amedeo et al., 2000; Probst et al., 2004; Stam and Mittelsten Scheid, 2005) remains to be elucidated. The isolation and characterization of maize mutants that are affected in the maintenance of the paramutated state have uncovered a number of factors relevant to gene inactivation by paramutation (Dorweiler et al., 2000; Hollick et al., 2005; Alleman et al., 2006; Woodhouse et al., 2006; Hale et al., 2007; Erhard Jr. et al., 2009; Barbour et al., 2012). The identification of DNA-dependent RNA polymerase IV (Pol IV) and an RNA-dependent RNA polymerase (RdRP) provided strong evidence for the involvement of small RNAs (sRNAs) in establishing and/or maintaining the paramutated state (reviewed, e.g., in Giacopelli and Hollick, 2015). In particular, an intact pathway of 24 nt sRNA biogenesis seems to be required for paramutation. 24 nt sRNAs are a class of non-coding RNAs that, in association with Argonaute proteins, mediate RNA-directed DNA methylation in *Arabidopsis thaliana*. Thus, the currently available data suggest a working model for paramutation in which sRNA-Argonaute complexes recruit a DNA methyltransferase to homologous sequences in nascent transcripts which in turn triggers *de novo* cytosine methylation in sequence elements controlling transcription of the paramutable allele (Giacopelli and Hollick, 2015). However, the mechanistic details of how gene inactivation by paramutation occurs, how the inactive state is imposed from the paramutagenic allele onto the paramutable allele, and how the inactive state is stably transmitted across generations are currently unknown.

The tomato *sulfurea* mutant represents one of the first discovered cases of paramutation in plants (Hagemann, 1958) and one of the very few paramutagenic loci in a dicotyledonous model species that is routinely amenable to genetic manipulation by stable transformation. The original *sulfurea* mutant occurred in an X-ray mutagenesis experiment (Hagemann, 1958; Hagemann and Berg, 1977; Hagemann, 1993). A second, independent paramutagenic allele was later isolated from a tomato tissue culture and probably originated from somaclonal variation (Wisman et al., 1993). Paramutation at the tomato *SULFUREA* locus leads to two visibly distinguishable phenotypes: yellow leaf sectors or branches (referred to as *sulf^{pura}* allele), or yellow-green variegated or speckled sectors (*sulf^{vag}* allele). Paramutated *sulfurea* tissue displays a striking yellow, chlorophyll-deficient phenotype (Hagemann, 1958; Ehlert et al., 2008; Figure 1). The *sulfurea* allele is recessive, but, fulfilling the hallmark of paramutation, the pigment deficiency appears spontaneously in somatic tissues of heterozygous plants at high frequency.

Extensive genetic mapping work has localized the *SULFUREA (SULF)* gene to the centromeric heterochromatin of tomato chromosome 2 (Hagemann, 1993). This location and the lack of genetic markers in close proximity (Tanksley et al., 1992) have made the isolation of the *SULF* gene extremely challenging. The pigment deficiency could suggest a gene function related to the photosynthetic apparatus (e.g., photosystem biogenesis or chlorophyll synthesis), but chlorotic phenotypes can also occur as secondary consequence of primary defects that are completely unrelated to photosynthesis (e.g., disturbed mineral nutrition).

In our previous work, we suggested a candidate locus that is affected by paramutation in the *sulfurea* mutant (Ehlert et al., 2008). The gene encodes a putative orthologue of the chloroplast protein ATAB2, a nucleus-encoded factor involved in the light-regulated translation of the plastid-encoded photosystem I reaction center protein PsaB in *Arabidopsis* (Barneche et al., 2006). A recent study has provided additional evidence for the tomato ATAB2 orthologue being a candidate gene for the *sulfurea* locus. The authors detected small interfering RNAs (siRNAs) derived from the promoter of the gene in paramutated tissue (Gouil et al., 2016), a finding that would be compatible with current sRNA-based working models for paramutation in maize (Giacopelli and Hollick, 2015). However, as discussed previously (Ehlert et al., 2008), there are several aspects of the *sulfurea* phenotype that are inconsistent with the phenotype reported for the *Arabidopsis atab2* mutant (Barneche et al., 2006). These include (i) the inability of homozygous *sulfurea* plants to grow heterotrophically on sucrose-containing medium, whereas T-DNA knockout mutants of *ATAB2* grow normally under heterotrophic conditions, and (ii) the severe auxin deficiency of homozygous *sulfurea* plants (Ehlert et al., 2008), a hormonal defect that so far has not been associated with primary defects in photosynthesis-related genes (Barneche et al., 2006). Therefore, the identity of the *SULFUREA* gene remains to be ultimately clarified.

Here we present compelling evidence that a tomato *ATAB2* homologue is indeed the *SULF* locus. We demonstrate that the *Arabidopsis ATAB2* gene is not paramutable (i.e., is insusceptible to inactivation by paramutation) and, therefore, can stably complement the paramutation phenotype of the tomato *sulfurea* mutant when introduced transgenically. We have also sequenced the genomes and methylomes of paramutable and paramutagenic *SULF* alleles and characterized a genetic suppressor mutant that suggests a mechanism how silencing by paramutation can be released. Our work paves the way to the exploitation of a readily transformable dicotyledonous plant as a new model to study the molecular mechanisms of paramutation.

RESULTS

Identification of a candidate gene for the tomato *sulfurea* locus

The strong phenotype of the mutant and the complete loss of chlorophyll observed in paramutated tissue suggest that paramutation leads to very efficient silencing of the affected gene. We, therefore, reasoned, that any candidate gene for *SULFUREA* should display strongly reduced gene expression in homozygous *sulfurea (sulf (hom))* seedlings compared to wild-type tissue. However, the severe phenotype of the paramutated tissue also indicates that gene inactivation by paramutation likely causes massive secondary effects, due to the loss of photosynthesis and, presumably, the resulting oxidative damage in the chloroplasts. To minimize these secondary effects, we generated large amounts of seeds from heterozygous *SULF/sulf* plants, germinated them under low-light conditions and phenotyped the seedlings very early after germination, when the first pair of true leaves emerged. At this stage, the homozygous plants were already recognizable by their light-green leaf color. Photosynthetic pigment accumulation in these very young, low-light grown leaflets was much stronger than in mature (yellow) leaves and leaves grown under standard light conditions (Figure 1; Ehlert et al., 2008), indicating that they suffer less from photooxidative damage and hence should show less severe secondary effects on gene expression that are unrelated to the primary, paramutation-caused defect.

To identify candidate genes that are potentially silenced in paramutated tissue, RNA samples extracted from very young leaf material were hybridized to microarrays. Analysis of differentially expressed genes revealed a short list of nine genes that are expressed to lower levels in *sulf* tissue compared to wild-type tissue and a longer list of 36 genes that are upregulated in paramutated tissue (Supplemental Table 1). While the upregulated genes are likely induced in response to the photosynthetic defect and the resulting metabolic deficiencies in the *sulf* mutant, the list of downregulated genes may contain candidate genes for the *SULF* locus. Interestingly, the gene showing the strongest expression difference between *sulf* and wild-type tissue, *Solyc02g005200,* was the only gene in the list of downregulated genes located in proximity to the centromere region of chromosome 2, the region to which *SULF* had been mapped genetically (Hagemann, 1993; Supplemental Table 1; Supplemental Figure 1). In extensive genetic crosses conducted by Rudolf Hagemann, the marker *S* (compound inflorescence; *Solyc02g077390*) was identified as the closest freely recombining marker to the *SULF* locus (Hagemann, 1993; Tanksley et al., 1992). We, therefore, used this marker to delimit the genomic region on chromosome 2 in which the *sulf* locus can potentially reside. The distance between the *S* locus and the end of chromosome 2 is approximately 42.3 Mb (Tanksley et al., 1992; The Tomato Genome Consortium, 2012), thus defining the potential target region where *SULF* could be located.

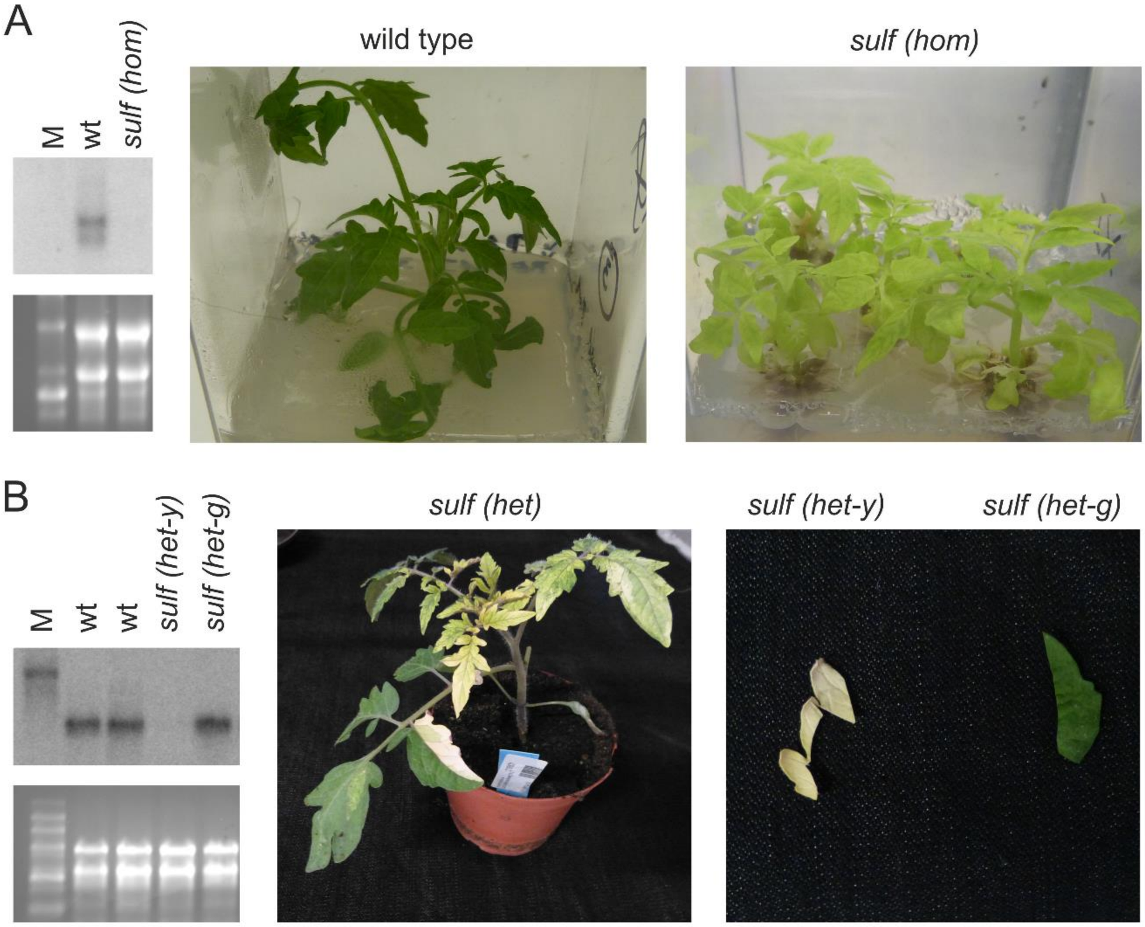


Figure 1. Expression analysis of the candidate gene *Solyc02g005200* in wild type (wt) and sulfurea (*sulf*) tissue. Northern blots were hybridized to a probe specific for *Solyc02g005200*. (**A**) Northern blot analysis of leaf tissue from plantlets grown under aseptic conditions. Wild type and *sulf (hom)* plantlets were grown under low-light condition on auxin-containing growth medium (Ehlert et al., 2008). (**B**) Northern blot analysis of leaf material from greenhouse-grown plants. Two wild-type plants and one heterozygous *sulf* plant (*sulf (het)*) were analyzed. The leaves of the *sulf (het)* plant were separated into green non-paramutated *sulf (het-g)* and yellow paramutated *sulf (het-y)* sectors. M: molecular weight marker.

Given that the microarray analysis had revealed only a single gene within the target region as downregulated in *sulf*, we directly tested this candidate locus, *Solyc02g005200*, for being silenced in paramutated tissue by RNA gel blot analysis. Using a probe against *Solyc02g005200*, northern blot experiments clearly demonstrated the lack of detectable expression of the *Solyc02g005200* gene in *sulf* tissue (Figure 1). The gene appeared to be fully silenced both in homozygous *sulf (hom)* plantlets growing on sucrose-containing and auxin-supplemented medium (Ehlert et al., 2008) under aseptic conditions and in yellow sectors (*sulf* (*het-y)*) excised from variegated heterozygous *sulf (het)* plants growing autotrophically in soil (Figure 1). By contrast, *Solyc02g005200* expression was readily detectable in both wild-type tissue and non-paramutated green sectors from heterozygous plants (*sulf* (*het-g)*; Figure 1). *Solyc02g005200* encodes a tomato homologue of the *Arabidopsis* gene *ATAB2* (*At3g08010*) that encodes a chloroplast-localized translational activator protein controlling expression of reaction center subunits of the photosystems in response to light (Barneche et al., 2006). The gene was originally identified as *TAB2* in in the unicellular green alga *Chlamydomonas reinhardtii* where it translationally regulates the expression of the plastid-encoded photosystem I reaction center protein PsaB (Dauvillée et al., 2003). The TAB2/ATAB2 protein has RNA-binding activity with a preference for A/U-rich motifs. The putative protein encoded by the tomato *Solyc02g005200* gene shares 73 % amino acid sequence identity with the protein encoded by the *Arabidopsis ATAB2* (*At3g08010*) gene. This high degree of conservation and the absence of other putative *ATAB2* homology from the tomato genome suggest that the tomato *Solyc02g005200* gene represents a true orthologue of *ATAB2*.

Transgenic expression of the *Arabidopsis* homologue of *Solyc02g005200* complements the paramutation phenotype of *sulfurea*

To ultimately prove the identity of the *SULFUREA* gene and confirm that it is identical with the tomato gene encoding TAB2, complementation of the mutant phenotype is required. We reasoned that this is unlikely achievable with the tomato gene which is paramutable and hence, may undergo epigenetic silencing when introduced transgenically into the *sulf* mutant. Genetic experiments with aneuploid tomato lines have demonstrated that paramutation is largely independent of the gene dosage in that a single paramutagenic allele can inactivate several paramutable alleles (Hagemann and Berg, 1977; Hagemann, 1993). We further reasoned that the putative *Arabidopsis* orthologue of *Solyc02g005200*, *ATAB2*, has a higher probability of being non-paramutable, especially when used as a cDNA and driven by a heterologous promoter. We, therefore, attempted a genetic complementation experiment with the *Arabidopsis ATAB2* cDNA to examine whether or not the paramutation phenotype of the tomato *sulf* mutant can be rescued. To this end, the cDNA of *At3g08010* was cloned into a binary vector, placed under the control of the CaMV 35S promoter and the *nos* terminator, and transformed into green heterozygous *sulf (het)* cotyledons by *Agrobacterium*-mediated transformation. Five independent transgenic lines were obtained, regenerated into plants and grown to maturity in the greenhouse. Interestingly, none of the lines showed any evidence of paramutation whereas all heterozygous control lines showed the expected variegated paramutation phenotype. To follow the correlation between presence of the transgene and lack of the paramutation phenotype over the next generations, large amounts of seeds were produced from two transgenic T_0 plants (lines 5 and 8) to perform segregation analyses by selfing and back-crosses to the wild type.

Due to the paramutagenicity of the *sulf* allele in the heterozygous *sulf (het)* seedling tissue used for transformation, it is possible that the genetic background of the generated lines changed from *sulf (het)* to *sulf (hom)* during the transformation procedure and/or the subsequent cultivation of the transgenic plants. If our assumptions (that *TAB2* is identical with *SULF* and *ATAB2* is not paramutable) were correct, the transition from *sulf (het)* to *sulf (hom)* would be masked by the expressed transgene and consequently, not become phenotypically visible. Therefore, depending on whether or not paramutation at the endogenous *SULF* locus has occurred in an individual transgenic plant, the next generation can exhibit different segregation ratios (Supplemental Table 2). The genotype of a fruit that produced a given set of seeds can be recognized in the next generation: If the cotyledons of all germinating seedlings that lack the transgene (i.e., 25% of the total progeny) are yellow, the fruit was *sulf (hom)*, if the cotyledons in only one quarter of the seedlings that lack the transgene (i.e., 6.25% of the total progeny) are yellow, the fruit was *sulf (het)*. Similarly, the genotype of the parental plant can be recognized by the different fractions of transgene-free progeny seedlings with true leaves that remain green versus true leaves that develop the variegated phenotype typical of paramutation (Supplemental Table 2). Consistent with these theoretical considerations, the two different frequencies of the occurrence of *sulf (hom)* seedlings with yellow cotyledons were indeed observed, with the values being very close to the theoretically expected values (22% and 5.7% versus the expected 25% and 6.25%; Supplemental Table 3). In 18.7 % of the progeny of the back-cross with the wild type of plants hemizygous for the transgene and heterozygous for *sulfurea* (*sulf (het)*), the paramutation phenotype (variegated leaves) was visible (Supplemental Table 3B). This correlates with the expected absence of the transgene from one quarter of the offspring. The slightly lower number of variegated plants is likely due to some heterozygous *sulf (het-g)* plants not having undergone paramutation yet at the time of phenotypic scoring (which, due to the large number of plants analyzed, was done 6 weeks after germination). Similarly, back-crosses with the wild type of transgenic plants that were homozygous for *sulfurea* (i.e., had *sulf (het-y)* reproductive tissue which is functionally equivalent to *sulf (hom)*) yielded 41.7% variegated plants, again a value that is slightly lower than the theoretically expected value of 50% (Supplemental Tables 2 and 3A), presumably due to the presence of some heterozygous plants that had not undergone paramutation after 6 weeks of growth. Taken together, these genetic data suggest strongly that the *ATAB2* transgene complements the *sulf* mutant and, moreover, indicate that the *ATAB2* transgene is not paramutable.

To provide ultimate proof of the *ATAB2* complementing the paramutation phenotype of the *sulf* mutant, we sought to confirm the strict co-segregation of the *ATAB2* transgene with the absence of the paramutation phenotype. To this end, we genotyped and phenotyped 1086 progeny plants of hemizygous transgenic lines that had been either selfed or back-crossed to the wild type (Figure 2; Table 1). Transgene presence was assayed by PCR and exemplarily confirmed by Southern blotting (see Methods; Table 1), and the three observable phenotypes were correlated to the genetic constitution of the plants concerning the *SULF* locus and the transgenic locus (Figure 2; Table 1). Consistent with the segregation ratios expected if the two assumptions made (*Solyc02g005200* is *SULF* and *ATAB2* is insensitive to paramutation; Supplemental Table 2) were correct, in none of the 169 *sulf (hom)* seedlings and in none of the 218 paramutated heterozygous *sulf (het)* plants (identified by their obvious leaf variegation phenotype; Figure 2), the *ATAB2* transgene could be detected (Table 2; Supplemental Table 2). Moreover, the observed segregation ratios were very similar to the theoretically expected values (cf. Supplemental Tables 2 and 3; Table 1) for both selfed plants and the back-crosses to the wild type.

Table 1. Segregation ratios in the progeny of T1 plants heterozygous for the *ATAB2* transgene *At3g08010* and homozygous for the paramutated *sulfurea* allele (*sulf (hom)* background). Transgene presence was detected by PCR. The number of plants observed for each phenotype and the segregation ratios (in %) are given.

	Yellow cotyledons		**Green cotyledons**			
			Green true leaves		**Variegated true leaves**	
Transgene **Cross**	-	+	-	+	-	+
Selfing	169 (27.4%)	0	0	446 (72.4%)	1 (0.002%)	0
Backcross to wild type	0	0	16 (3.4%)	237 (50.4%)	217 (46.2%)	0

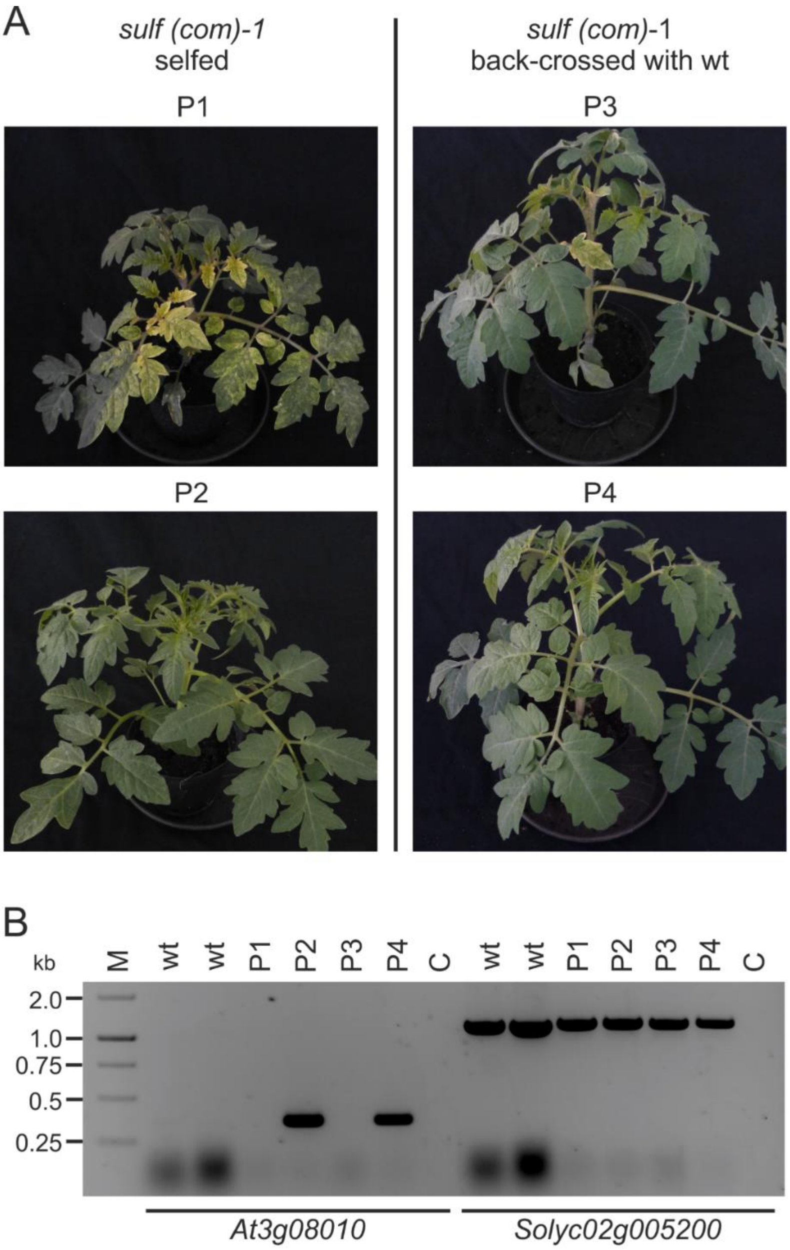


Figure 2. Complementation of the paramutation phenotype of the *sulf* mutant with a transgene derived from the *ATAB2* gene from *Arabidopsis thaliana* (*At03g08010*). (**A**) Reappearance of the variegated *sulfurea* phenotype in the progeny of a hemizygous complemented line (*sulf (com)-1*). The line was either selfed (left panels) or back-crossed to a wild-type plant (right panels). The offspring of both crosses segregates into individuals displaying a clear paramutation phenotype (upper panels; plants P1 and P3) and plants lacking the phenotype and being indistinguishable from the wild type (lower panels; plants P2 and P4). For segregation analyses, see Supplemental Tables 2 and 3. (**B**) Dependence of the reappearance of the paramutation phenotype on outcrossing of the *ATAB2* transgene. Shown here are representative PCR analyses using genomic DNA as template and primer pairs specific for *ATAB2* (*At03g08010*; left part) and the candidate *SULF* gene (*Solyc02g005200*; right part), respectively. Two wild-type plants were used as control. C: negative controls with no template DNA added to the PCR reaction. For segregation analysis, see Table 1.

Finally, we also analyzed expression of the *ATAB2* transgene and the *SULF* candidate locus *Solyc02g005200* in the progeny of complemented *sulf (com)* lines by RNA gel blot analyses (Figure 3). As expected, all plants displaying a wild type-like phenotype and showing no paramutation express the *ATAB2* transgene or the endogenous *Solyc02g005200* gene or both. Importantly, in several plants the candidate *SULF* gene *Solyc02g005200* was fully silenced (*sulf (het-y)*), but yet the plants show no paramutation phenotype. Since all of these plants expressed the *ATAB2* transgene, this finding provides strong additional support for expression of *ATAB2* complementing the paramutation phenotype (Figure 3).

In conclusion, presence and active expression of the *ATAB2 (At3g08010)* transgene fully complements the *sulfurea* phenotype, ultimately confirming that (i) *Solyc02g005200* is the tomato orthologue of *ATAB2*, and (ii) *Solyc02g005200* is identical with the *SULF* gene.

Genome sequencing and methylation sequencing reveal different cytosine methylation patterns at *Solyc02g005200*

Previous work on paramutation in maize has implicated RNA-dependent DNA methylation in the establishment of the paramutated state (e.g., Barbour et al., 2012; Regulski et al., 2013; Bond and Baulcombe, 2014; Giacopelli and Hollick, 2015). We, therefore, wanted to determine whether the wild-type *SULF* allele and the paramutated (epi)allele differ in their DNA methylation patterns. To this end, we sequenced both the genomes and the methylomes of paramutated (*sulf (het-y)*) tissue and non-paramutated green tissue (*sulf (het-g)*) and compared them with the wild-type tomato cultivar from which the *sulf* mutant was originally derived (*S. lycopersicum* cv. Lukullus). The *sulf (het-g)* and *sulf (het-y)* tissues used for sequencing were harvested from the same (phenotypically chimeric) heterozygous *sulf (het)* plants to minimize genetic variation. Genomic DNA was extracted and used for genome resequencing and for bisulfite conversion followed by sequencing (Cokus et al., 2008). In an attempt to obtain further insights into the mechanism of paramutation, we additionally sequenced the genome and the methylome of the spontaneous suppressor mutant *SOSU1* (*SUPPRESSOR OF SULFUREA 1*) which had been isolated as a genetic suppressor of paramutation at the *SULF* locus (Ehlert et al., 2008; Supplemental Tables 4 and 5; see also Materials and Methods).

We first assembled the genomic sequences from the wild type (cultivar Lukullus) and from non-paramutated *sulf (het-g)* and paramutated *sulf (het-y)* leaves and aligned them to the tomato reference genome (version SL2.50; The Tomato Genome Consortium, 2012; Supplemental Table 4). The aligned genome sequences were analyzed for potential structural variations (insertions, deletions and copy number variations; see Materials and Methods). No obvious differences in or close to the centromeric region of chromosome 2 were identified upon comparison of non-paramutated *sulf (het-g)* and paramutated *sulf (het-y)* tissue with the corresponding wild-type cultivar or the sequenced reference cultivar (Supplemental Figure 2). This result is consistent with the epigenetic nature of the paramutation phenomenon.

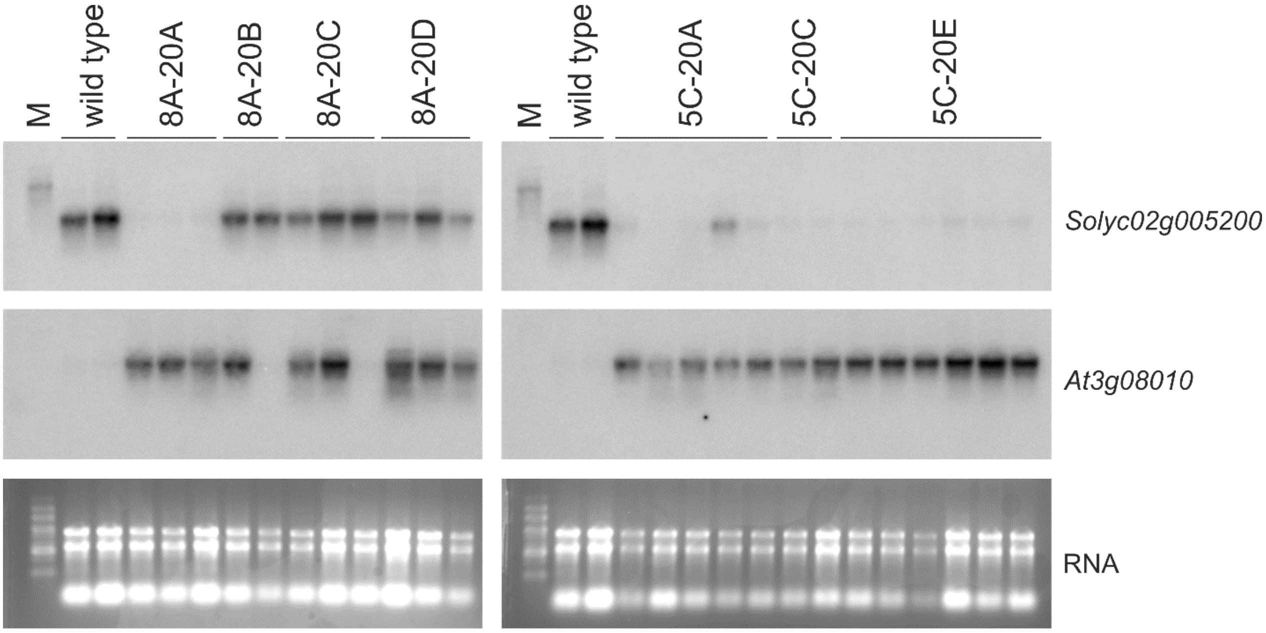


Figure 3. Northern blot analyses of *ATAB2* and *Solyc02g005200* expression in transgenic *sulf (com)* lines. RNA samples extracted from young leaves of greenhouse-grown plants (representing offspring of two independently generated transgenic *sulf (com)* lines: 8A and 5C) were separated by denaturing agarose gel electrophoresis, blotted and hybridized to probes against *Solyc02g005200* (upper panel) and *ATAB2* (*At3g08010*; middle panel). All plants investigated displayed a wild type-like phenotype and showed no evidence of paramutation. The lower panel shows the ethidium bromide-stained agarose gel prior to blotting as a control for RNA integrity and loading. Note that, in several plants (e.g., all three 8A-20A plants, 4 out of 5 5C-20A plants and both 5C-20C plants), the candidate *SULF* gene *Solyc02g005200* is fully silenced, but yet the plants show no paramutation, strongly suggesting phenotypic complementation by *ATAB2*. The two 8A plants lacking *ATAB2* expression represent progeny that lost the transgene due to Mendelian segregation.

Next, we performed bisulfite conversion of genomic DNA followed by next-generation sequencing to determine potential differences in the cytosine methylation patterns in paramutated (*sulf (het-y)*) versus non-paramutated tissue (*sulf (het-g)*) tissue (for quality parameters and mapping statistics of the four sequenced samples, see Supplemental Table 5). The sequences of bisulfite-treated genomic DNA of the four samples were mapped to the reference genome, methylation ratios at single-basepair resolution were calculated and differential cytosine methylation was determined (see Materials and Methods).

Large differences in the cytosine methylation patterns at locus *Solyc02g005200* could be detected upon comparison of bisulfite-converted wild-type DNA with bisulfite-converted *sulf (het)* DNA (Supplemental Datasets 1 and 2). Especially the promotor region and the first two exons differ markedly in their cytosine methylation patterns between the wild type and *sulf (het)* (Figure 4). No other region in chromosome 2 exhibited similarly extensive changes in DNA methylation. While both CpG and CHG methylation levels were very low in the wild type, much higher levels of both methylation types are found in *sulf (het)* (Figure 4A). Interestingly, this was the case for both *sulf (het-g)* and *sulf (het-y)* tissue suggesting that many of the methylated sites are not directly involved in the establishment of paramutation, although we currently cannot exclude the possibility that they contribute to facilitating the transition to the silenced state. Any candidate site that is causally involved in silencing via paramutation should be hypermethyated in *sulf (het-y)* tissue compared to *sulf (het-g)* tissue. Indeed, this was found to be the case for a single methylated position upstream of the *SULF* reading frame (position 7,269,217; Figure 4). This position is 100% methylated in *sulf (het-y)* tissue but only 50% methylated in *sulf (het-g)* tissue (Supplemental Dataset 1; Figure 4B). This difference is strikingly consistent with the expectation that, if methylation triggers paramutation, the presence of one silenced (paramutated) allele and one unsilenced (not yet paramutated) allele in *sulf (het-g)* tissue should result in a methylation level of 50% at any site(s) causally involved in gene inactivation. This finding raises the interesting possibility that methylation at a single cytosine in the putative promoter/enhancer region of the *SULF* gene mediates the switch from the active to the inactive (paramutated) state. In contrast to CpG methylation, the overall CHG methylation levels within the putative promotor region were found to be slightly lower in paramutated *sulf (het-y)* tissue than in *sulf (het-g)* tissue (Supplemental Dataset 2; Figure 4). The possible significance of this difference is currently unclear.

Isolation and phenotypic characterization of revertants and suppressors of paramutation

We have previously described a tissue culture-based screen for suppressors of paramutation at the *sulfurea* locus (Ehlert et al., 2008). As tissue culture conditions are known to induce alterations in DNA methylation patterns (Kaeppler and Phillips, 1993) that can result in stable epigenetic changes (Stroud et al., 2013), we initiated a screen that is based on large scale *in vitro* regeneration from paramutated leaf material followed by visual inspection of regenerated

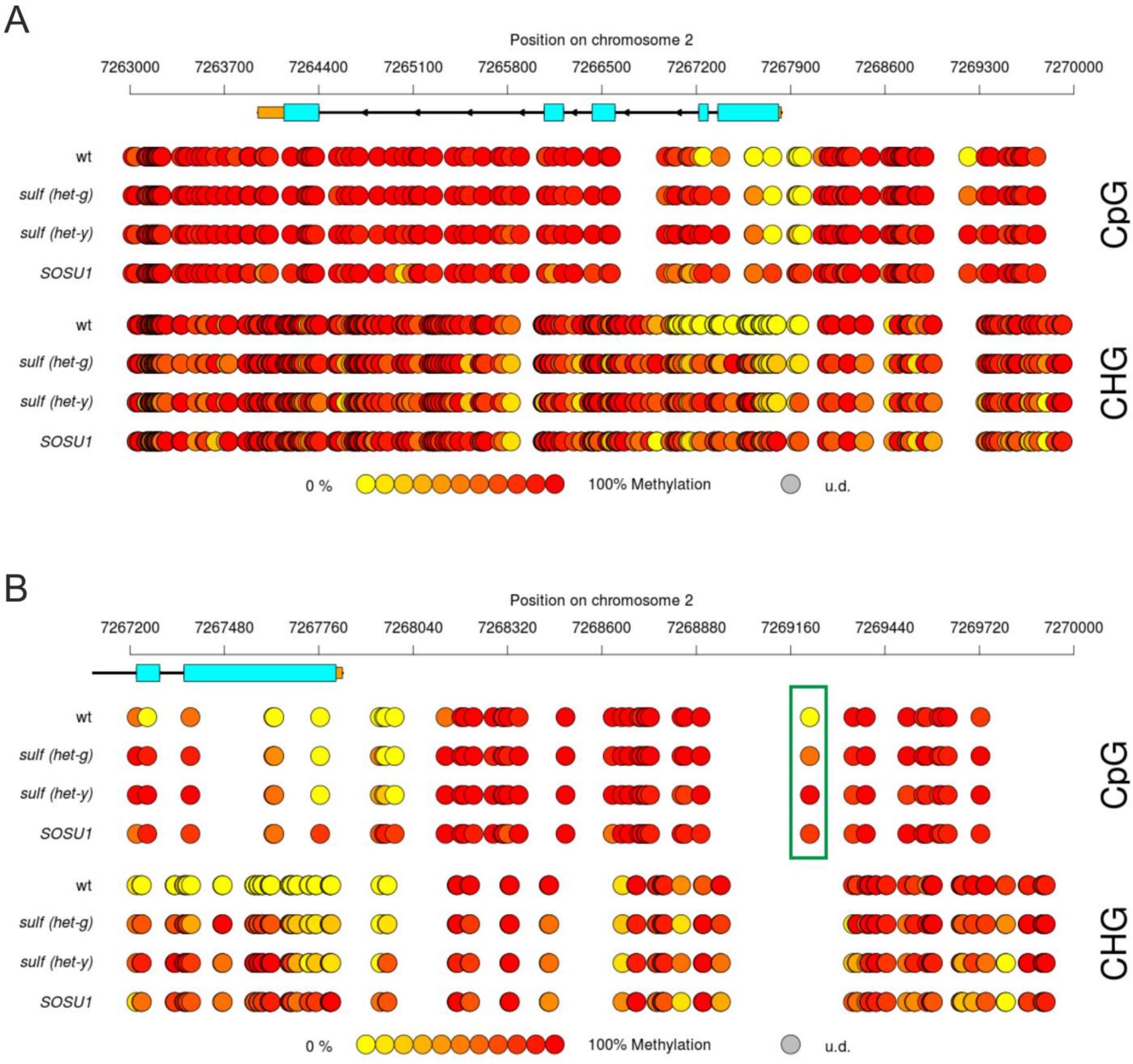


Figure 4. Comparative analysis of DNA methylation at locus *Solyc02g005200*. Data were derived from methylome sequencing of wild-type leaves (wt), yellow paramutated *sulf (het-y)* leaf sectors, green non-paramutated *sulf (het-g)* leaf sectors and leaves of the suppressor mutant *SOSU1* (cf. Supplemental Table 5). The structure of the *SULF* locus (*Solyc02g005200*) is schematically depicted above the methylation charts. Exons are represented as light blue boxes, untranslated regions as orange boxes and introns as black lines. The direction of transcription is from right to left. (**A**) Cytosine methylation levels (CpG and CHG methylation) at locus *Solyc02g005200* and the surrounding genomic sequences upstream and downstream of the coding region. (**B**) Single-nucleotide-level resolution of the cytosine methylation levels (CpG and CHG methylation) in the promotor region and the first two exons of the gene. A candidate site that could mediate the switch from the active to the paramutated state is boxed in green. See text for details. u.d.: site with undetermined methylation status.

plantlets for appearance of dark green spots. Excision of such leaf sectors with wild type-like pigmentation and their regeneration into plants then provides material that can be analyzed and, most importantly, tested for regain of autotrophic growth. This screen led to the isolation of *SOSU1* (*SUPPRESSOR OF SULFUREA 1*), a suppressor mutant that overaccumulates auxin (while paramutated *sulfurea* tissue is auxin deficient; Ehlert et al., 2008). *SOSU1* plants exhibit a striking phenotype many aspects of which (epinastic and strongly retarded growth, aberrant flower morphology with homeotic organ transformations, sterility) are likely explained by the presence of excess amounts of auxin (Kawano et al., 2003; Zhao, 2008).

Extensive screens for additional suppressor mutants of paramutation with *sulf (hom)* tissue grown under sterile conditions on auxin-supplemented medium yielded a set of four potential suppressor mutants that were preliminarily named *SOSU2-5* (Figure 5A; Supplemental Figure 3A). Green spots or leaf sectors were excised and regenerated into plantlets. After transfer to soil and growth to maturity under greenhouse conditions, all plants remained homogeneously green and showed no sign of paramutation (Figure 5A; Supplemental Figure 3A).

To confirm that loss of paramutation correlates with regain of *SULF* gene expression, northern blot experiments were conducted (Supplemental Figure 3B). Indeed, all five putative suppressor lines expressed the *SULF* gene to wild type-like levels (Supplemental Figure 3B), indicating that the paramutated locus had become spontaneously reactivated. In contrast to *SOSU1* (see above), *SOSU2-5* plants were indistinguishable from wild-type plants (Figure 5A; Supplemental Figure 3A) and developed normal fruits that harbored fertile seeds. This enabled us to assess the stability of the loss of paramutation in the next generation. To this end, *SOSU* plants were either selfed or crossed to wild-type plants and the progeny was assayed for their paramutation phenotype. Surprisingly, despite the non-paramutated phenotype of the parental *SOSU* plant, the progeny segregated like that of a heterozygous plant with one paramutagenic allele (Supplemental Table 6). Selfed *SOSU* plants produced 25% homozygous *sulf* progeny (as evidenced by their yellow cotyledon phenotype), 50% heterozygous progeny (undergoing somatic paramutation and thus becoming variegated) and 25% wild-type plants (remaining homogeneously green). When crossed to the wild type, half of the progeny was variegated and the other half was wild type, again consistent with a heterozygous status of the *SOSU* parent. These segregation data strongly suggest that the *SOSU2-5* lines are not suppressor mutants, but rather spontaneous somatic revertants of the paramutated state. Remarkably, this reversion is not transmitted through the germline and, instead, one of the two *SULF* alleles remains paramutagenic or regains paramutagenicity in the next generation.

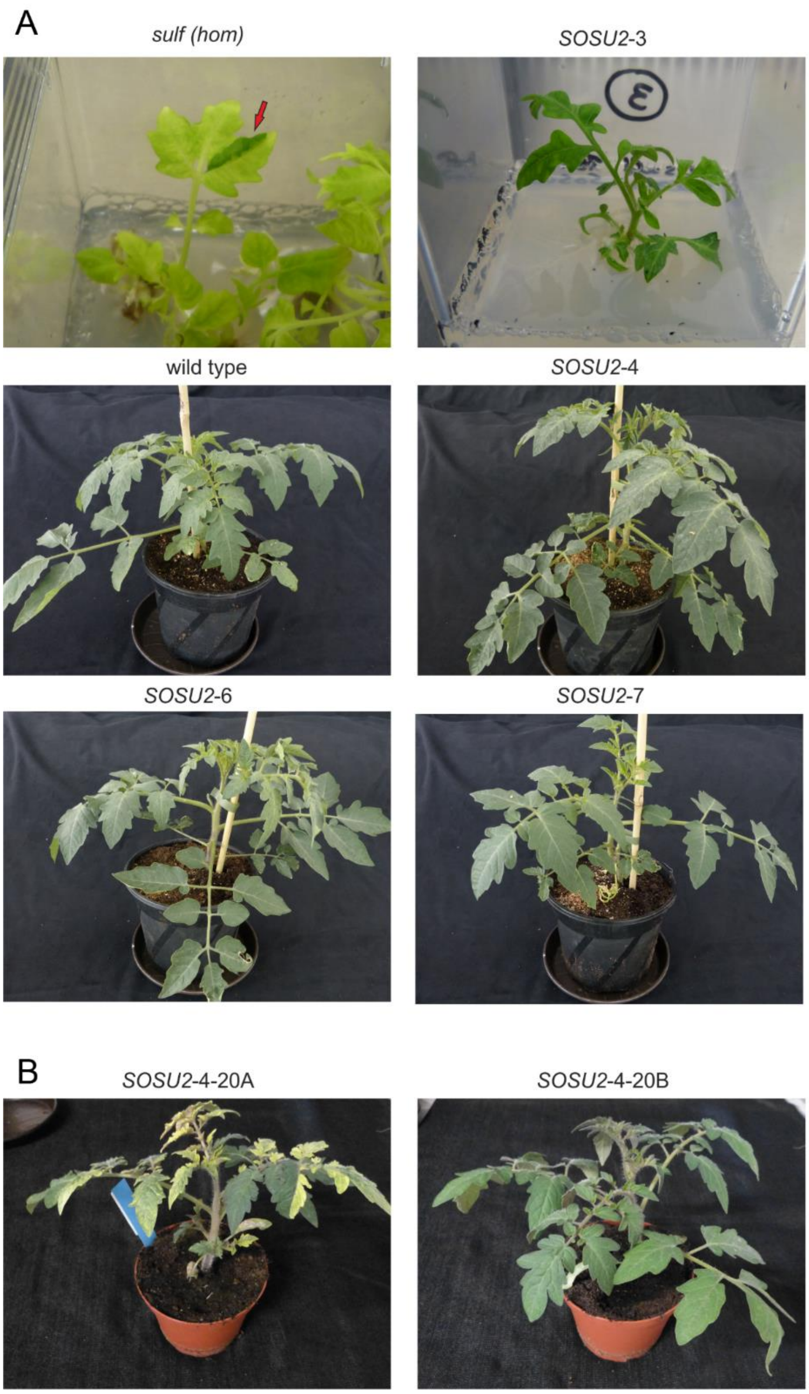


Figure 5. Isolation of *SOSU2*, its phenotype and segregation for paramutation in the next generation. (**A**) Identification of the putative suppressor of paramutation *SOSU2*. A dark green sector appearing spontaneously in a *sulf (hom)* leaf tissue (red arrow in upper left photograph) was excised and regenerated. All regenerated plantlets showed wild type-like dark green pigmentation, both in *in vitro* culture (upper right photograph) and upon growth in soil under greenhouse condition. *SOSU2-3, 4, 6* and *7* represent individual plants regenerated from the same somatic event (red arrow in upper left photograph). (**B**) Reappearance of the paramutation phenotype in the next generation. Two representative progeny plants of *SOSU2-4* are shown (20A and 20B). While plant 20A (left) is an F1 plant exhibiting a clear paramutation phenotype, plant 20B represents a wild-type segregant and, therefore, shows no paramutation.

To ultimately verify that the reappearance of the paramutation phenotype in the next generation was due to loss of *SULF* gene expression, green and variegated progeny was analyzed by northern blot experiments. As expected, green progeny and green sectors from variegated progeny showed *SULF* gene expression, whereas no expression was detectable in yellow leaf sectors (Supplemental Figure 4). This result confirms the presence of a paramutable *SULF* allele in the segregating F1 population.

Taken together, these findings strongly suggest that the *SOSU2-5* lines are revertants rather than suppressors of paramutation. For clarity, we, therefore, renamed these lines *ROSU2-5* (for *<u>R</u>EVERTANT <u>O</u>F <u>SU</u>LFUREA*).

Identification of a candidate suppressor gene of paramutation

In contrast to *SOSU2-5* which turned out to be repressors of paramutation (*ROSU2-5*), several lines of evidence support the conclusion that *SOSU1* represents a real suppressor of paramutation at the *SULFUREA* locus. In particular, the mutant phenotype of *SOSU1* plants (many aspects of which are related to auxin overaccumulation; Ehlert et al., 2008) excludes the possibility that a simple reversion of paramutation has occurred. Based on the origin of *SOSU1* as a spontaneously arisen mutant, we reasoned that the suppressor mutation is likely dominant and, most probably, represents a gain-of-function mutation. However, due to the (male and female) sterility of *SOSU1*, this suspected dominance cannot be established directly by genetic crosses

Suppressors can arise either from a spontaneous point mutation or a chromosomal rearrangement. To identify the potential cause that underlies the suppression of paramutation in *SOSU1*, the genome and the methylome of *SOSU1* were sequenced and compared to the sequences of the wild type, non-paramutated *sulf (het-g)* tissues and paramutated *sulf (het-y)* tissue. Sequencing of bisulfite-treated DNA revealed that *SOSU1* plants exhibit strong alterations in their DNA methylation patterns (Figure 6). While the number of sites that were hypermethylated relative to the wild type was comparably low as in *sulf (het-g)* and paramutated *sulf (het-y)* tissues (Figure 6B), the number of hypomethylated sites was drastically elevated in *SOSU1* plants (Figure 6A). 6094 exons, 5204 introns and 1423 promoter regions displayed hypomethylation at CpG motifs in *SOSU1*, indicating that the suppressor mutant may have a problem with maintaining cytosine methylation patterns (Figure 6).

.

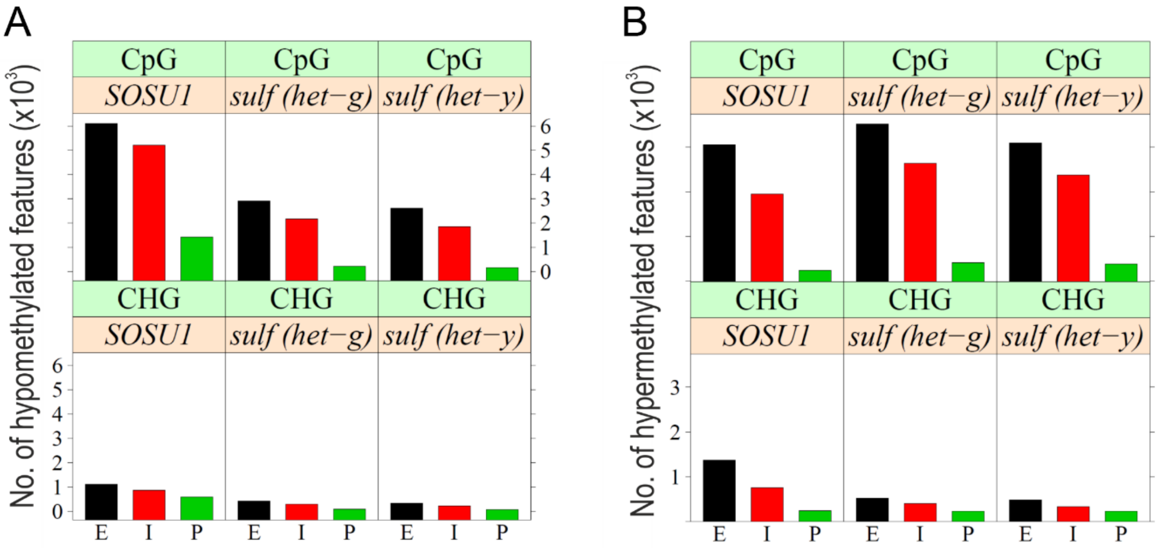


Figure 6. Altered cytosine methylation patterns in the *SUPPRESSOR OF SULFUREA 1* (*SOSU1*) mutant. (**A**) Hypomethylated features. CpG (upper panel) and CHG (lower panel) methylation are shown. The bars represent the number of gene features that are hypomethylated relative to the wild type (with more than 25% difference) in *SOSU1*, non-paramutated *sulf (het-g)* tissue and paramutated *sulf (het-y)* tissue. Exons (E) are represented in black, introns (I) in red and promoter regions (P) in green. (**B**) Hypermethylated features.

Next, we inspected the genomic sequence of *SOSU1* in an attempt to identify a candidate gene encoding the suppressor. Analysis of copy number variation (CNV) in *SOSU1* using two detection tools (cnv-seq and DELLY; see Materials and Methods) identified a single region on chromosome 7 as being hemizygous (i.e., present only in one copy) in *SOSU1* (Supplemental Figure 5). The hemizygous region represents a 375 kb deletion (SL2.50ch07: 66,904,790-67,264,135) that encompasses 61 putative genes (Supplemental Table 7). As most deletions are recessive and it also seemed unlikely that a heterozygous deletion causes the strong mutant phenotype of the *SOSU1* suppressor mutant, we analyzed the breakpoints of the deletion in more detail to test the possibility that an aberrant fusion gene is formed by the deletion which could act as a dominant suppressor. Indeed, the genomic deletion in *SOSU1* generates a potential fusion gene, because the two breakpoints are located within genes that have the same orientation on the chromosome. The first breakpoint is in the fifth intron of gene Solyc07g064950 (encoding a putative trigger factor-like protein) and the second breakpoint is in the first intron of locus Solyc07g065550 (encoding an NAD-dependent histone deacetylase; Supplemental Table 7). The fusion of these two loci results in a potential fusion protein that comprises the first half of the putative trigger factor (a ribosome-associated molecular chaperone with peptidyl-prolyl cis/trans isomerase activity that is involved in co-translational protein folding; Kramer et al., 2009; Pechmann et al., 2013) and the nearly complete NAD-dependent histone deacetylase (Figure 7).

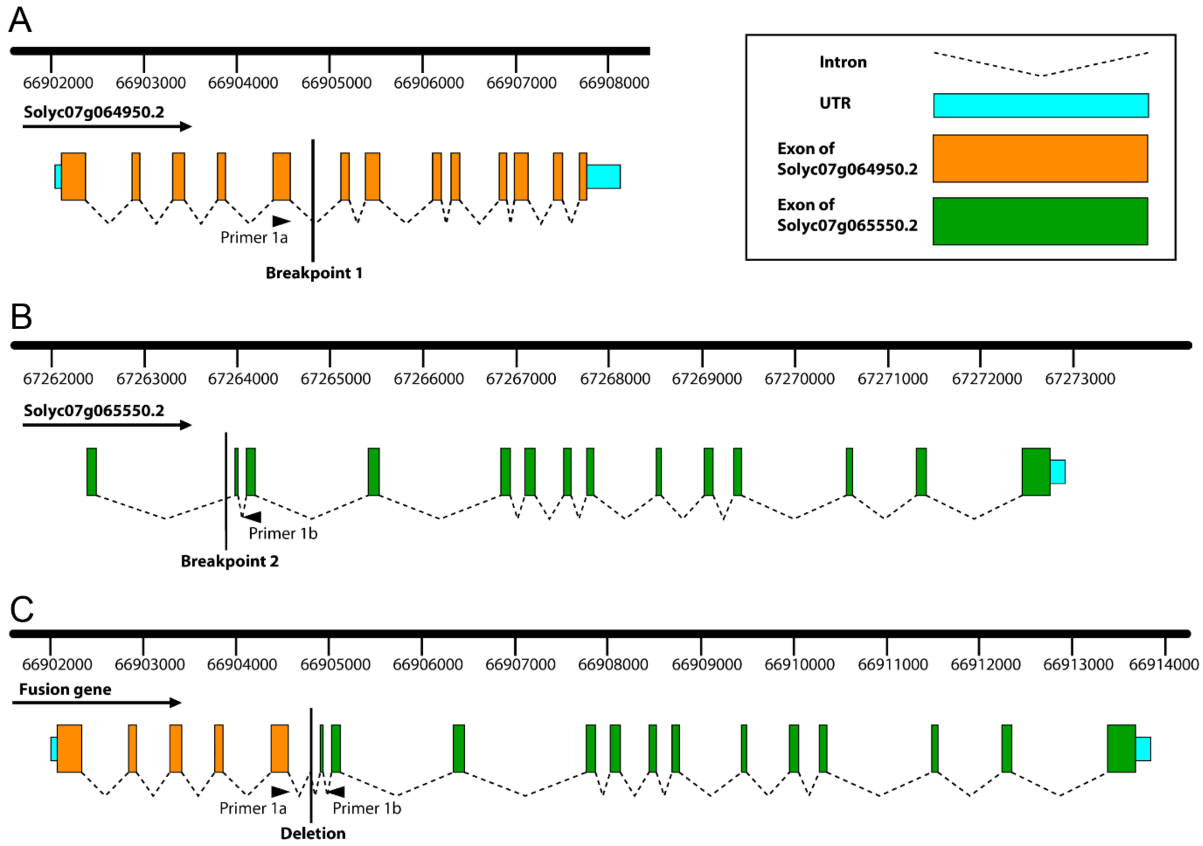


Figure 7. Gene model of the fusion gene identified in *SOSU1*. The breakpoints of the deletion in the two genes are indicated by horizontal lines. The exons of the putative trigger factor-like gene are represented as orange boxes, the exons of the histone deacetylase gene as green boxes. Introns are shown as dotted lines (and not drawn to scale). The binding sites and orientations of primers used for verification of the fusion gene and confirmation of its expression are depicted by arrowheads (cf. Figure 8).

The involvement of a histone deacetylase in the creation of the fusion gene is particularly interesting, because histone deacetylation has been implicated in maintenance of DNA methylation patterns and epigenetic inheritance. For example, in *Arabidopsis thaliana*, histone deacetylation was shown to be involved in the maintenance of CpG methylation by the DNA methyltransferase MET1 (Blevins et al., 2014). Thus, the formation of the aberrant fusion gene in *SOSU1* could conceivably explain the drastically altered methylation patterns (Figures 4 and 6) and, perhaps, also the loss of paramutation. However, the changes in DNA methylation at the *SULF* locus do not include the cytosine that distinguished the paramutated from the non-paramutated allele (position 7,269,217; Figure 4), raising the possibility that the suppression could occur directly through altered histone modification patterns and the concomitant changes in chromatin structure. Whether the altered DNA methylation patterns are solely a secondary consequence of that or, instead, contribute actively to the suppression mechanism, remains to be determined. To verify the existence of the fusion gene, PCR primers flanking the deletion were designed. The primers were derived from exon sequences of both genes involved in the fusion (Figure 7), to be able to also analyze expression of the fusion gene at the mRNA level. Using genomic DNA as template for PCR amplifications, the fusion gene could be readily detected in *SOSU1* plants (Figure 8A). PCR assays with cDNA as template revealed that the fusion gene is expressed. The detection of spliced mRNA (Figure 8B) revealed that the chimeric intron formed at the junction of the two genes (Figure 7) is faithfully spliced out, providing further evidence for the fusion gene giving rise to a functional gene product.

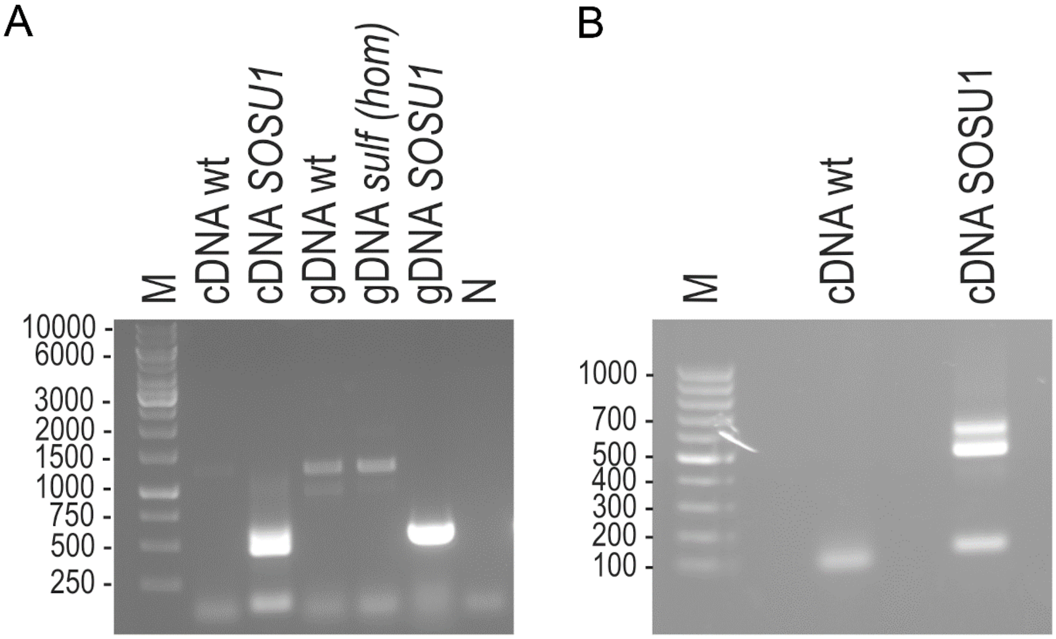


Figure 8. Detection of the fusion gene in the *SOSU1* suppressor mutant and confirmation of its expression. Primers flanking the deletion (primers 1a and 1b; cf. Figure 7) were used for PCR and RT-PCR assays. (**A**) Detection of the fusion gene and its transcript. Templates for PCR reactions were cDNA or gDNA (genomic DNA) of the wild type (wt), paramutated *sulf* tissue and *SOSU1* tissue. Weak larger bands in the assays with wt and *sulf* DNA likely represent non-specific amplification products that were not further characterized. (**B**) Separation of the cDNA amplification products at higher resolution. The same primer combination as in (A) was used for RT-PCR. The upper band represents the unspliced transcript, the lower band (531 bp product) represents the spliced mRNA. Its detection confirms that the chimeric intron (cf. Figure 7) can be faithfully spliced out. M: molecular weight marker (fragment sizes indicated in bp); N: negative control (no template DNA added).

Taken together, these data make the histone deacetylase fusion gene formed in the *SOSU1* mutant a strong candidate locus for encoding the suppressor of paramutation. While several loss-of-function mutants with defects in paramutation have been characterized in maize (e.g., Dorweiler et al., 2000; Alleman et al., 2006), to our knowledge, the *SOSU1* gene represents the first gain-of-function mutation that suppresses paramutation.

DISCUSSION

In this work, we have identified the *SULFUREA* gene of tomato, one of the classical paramutated loci that led to the genetic characterization of paramutation phenomena. The *SULF* locus encodes the tomato homologue of the *ATAB2* gene from *Arabidopsis* (*At3g08010*). The encoded protein is a chloroplast-localized RNA-binding protein that acts as a translational activator by controlling the expression of reaction center subunits of the photosystems in response to light, with the photosystem I reaction center protein PsaB being most strongly affected in *atab2* mutants (Dauvillée et al., 2003; Barneche et al., 2006). The identification of the *SULFUREA* gene (*Solyc02g005200*) as a homologue of *ATAB2* is consistent with our previously reported physiological analysis of paramutated tissue that had revealed strongly reduced photosystem I amounts, while photosystem II was only mildly affected (Ehlert et al., 2008).

Based on physiological analyses and comparative gene expression analyses, the *Solyc02g005200* locus had been suggested as a strong candidate for the *SULF* gene (Ehlert et al., 2008; Gouil et al., 2016). In the course of this work, we unambiguously demonstrated by genetic complementation with the homologous gene from *Arabidopsis* that the tomato homologue of the *ATAB2* gene is identical with the *SULF* locus. Moreover, the successful stable complementation demonstrates that the *Arabidopsis* gene is insusceptible to inactivation by paramutation (i.e., not paramutable). Together with the ease with which tomato can be genetically transformed, this finding provides the exciting opportunity to test both natural alleles and synthetic (mutated or chimeric) alleles for their susceptibility to inactivation by paramutation in straightforward transgenic assays. In the future, this will allow the precise dissection of the *cis*-acting sequence requirements involved in the establishment of the paramutagenic state.

Our methylome analyses revealed a single candidate cytosine whose methylation could be responsible for the inactivation of the *SULF* gene by paramutation (Figure 4; Supplemental Dataset S1). Since the site represents the only cytosine in the *SULF* region that is 100%

methylated in in *sulf (het-y)* tissue, 50% methylated in *sulf (het-g)* tissue and nearly fully unmethylated in wild-type tissue (Supplemental Dataset 1; Figure 4B), it is tempting to speculate that this methylation event is causally involved in gene inactivation.

The original paramutable *sulfurea* epiallele arose through X-ray mutagenesis (Hagemann, 1958). It is conceivable that the irradiation caused DNA damage in this region the repair of which triggered *de novo* cytosine methylation that fortuitously resulted in a paramutation-facilitating methylation pattern. However, how the methylated cytosine of the paramutated epiallele imposes its methylation status onto the wild-type allele remains to be determined. A recent study reported small changes in siRNAs derived from the promoter region of *Solyc02g005200* in paramutated versus non-paramutated tissue (Gouil et al., 2016). Whether or not these changes are causally involved in the establishment of the paramutated state needs to be further investigated.

Unlike most other epigenetic gene silencing phenomena, paramutation is heritable. This meiotic stability is most readily explained by changes in DNA methylation patterns that are stably inherited (Cubas et al., 1999; Mittelsten Scheid et al., 2003). Previous work has also shown that, in autotetraploids of *Arabidopsis thaliana*, epialleles can interact with each other in *trans* thus leading to heritable gene silencing. The inactive status of the gene can persist after genetic segregation from the inactivating epiallele (Mittelsten Scheid et al., 2003), a feature consistent with the hallmark of paramutation. However, whether or not the mechanism underlying this *trans*-inactivation phenomenon in tetraploids is fully equivalent to paramutation, remains to be clarified.

A remaining puzzling aspect of the *sulf* mutant phenotype is the severe auxin deficiency and the partial rescue of the seedling lethality by exogenously applied auxin (Ehlert et al., 2008). The auxin deficiency cannot readily be explained by the known functions of the *ATAB2* gene product. Also, while *sulf* mutant plants cannot grow on sucrose-containing medium, our analysis of *atab2* knock-out mutants in *Arabidopsis* revealed that they grow heterotrophically on sucrose without the requirement for auxin supplementation. Interestingly, our comparative microarray analysis revealed that an auxin efflux carrier (encoded by *Solyc02g082450*) is upregulated in paramutated *sulf* tissue (Supplemental Table 1B), raising the possibility that the observed auxin deficiency represents a secondary effect of deregulated nuclear gene expression in response to defective gene expression in the chloroplast.

Suppressor screens with paramutated *sulf (hom)* tissue in sterile culture with auxin supplementation yielded a series of putative suppressor mutants of *sulfurea* in which wild type-like pigmentation was restored. Our genetic analyses revealed that, with a single exception, these were spontaneous revertants (*ROSU*) rather than second site suppressors (*SOSU*). Interestingly, while the green phenotype of all *ROSU* lines remained somatically stable, paramutation reoccurred in the next generation, suggesting that, while the mechanism of reversion is likely of epigenetic nature, the paramutable state of the *sulf* epiallele is not fully erased in the *ROSU* plants. As an unfortunate consequence of the lack of meiotic stability, *ROSU* plants can only be propagated vegetatively. Why the reversion is somatically stable but lost during meiosis is currently completely unclear and will be interesting to investigate in more detail in future studies.

The only real suppressor of *sulfurea* (*SOSU1*) appears to be a dominant suppressor mutant that, due to a genomic deletion, expresses an aberrant fusion gene involving a gene that encodes an NAD-dependent histone deacetylase (Figures 7 and 8). Consistent with a globally deregulated chromatic structure in the *SOSU1* genome, the mutant has a pleiotropic phenotype and is male and female sterile. The suppression of the paramutation at the *SULF* locus is completely stable and no reoccurrence of leaf variegations was observed within several years of continuous cultivation of (vegetatively propagated) *SOSU1* plants in sterile culture and in the greenhouse. Consistent with the suspected massive changes in chromatin structure, also the DNA methylation patterns in *SOSU1* were dramatically changed (Figure 4; Supplemental Datasets 1 and 2). As methylation at the discriminatory cytosine (Figure 4B) is unaffected in *SOSU1*, it seems possible that the reactivation of the *SULF* gene in the suppressor occurs through changes in histone modification (as mediated by the aberrant NAD-dependent histone deacetylase fusion gene) rather than through removal of the repressive cytosine methylation. It will be interesting to further analyse the mode of action of the gene product of the histone deacetylase gene fusion and determine the mechanism of how it releases the paramutated state.

It is noteworthy that, despite extensive field cultivation of the *sulfurea* mutant for many decades (Hagemann, 1958; Hagemann, 1993), reversion or suppression of the paramutated phenotype was never observed. We attribute this to the fact that our screen for suppressors and revertants was conducted in tissue culture. Tissue culture conditions are known to trigger changes in DNA methylation patterns (Kaeppler and Phillips, 1993; Smulders and de Klerk, 2011; Miguel and Marum, 2011; Stroud et al., 2013) which may be a major cause of somaclonal variation, but also result in elevated mutation rates (Jiang et al., 2011). We, therefore, speculate that screening for suppressors and revertants under tissue culture conditions greatly increases the probability of recovering such events. Due to the unsurmountable difficulties with regenerating maize efficiently from leafy tissues, the tomato system offers the unique attraction of being able to conduct such screens at large scale.

In sum, the identification of the tomato *SULFUREA* gene, the discovery of a cytosine methylation event upstream of the reading frame which distinguishes paramutated from non-paramutated tissue, and the isolation of a dominant suppressor and a series of revertants of paramutation at the *SULF* locus provide a new entry point into the study of paramutation in plants. Moreover, our work reported here makes a tractable and easily transformable dicot model amenable to detailed investigations into the molecular mechanism of paramutation.

MATERIALS AND METHODS

Plant material, growth conditions, and analysis of paramutation phenotypes

Tomato seeds (*Solanum lycopersicum* cv. Lukullus) were germinated in soil. Seedlings were cultivated in a growth chamber at a light intensity of 100 μmol quanta m^{-2} s^{-1} for 3 weeks, then transferred to the greenhouse and grown under standard conditions (average light intensity: 250 μmol quanta m^{-2} s^{-1}). Leaf material was harvested and subsequently frozen in liquid nitrogen. In the case of homozygous *sulf (hom)* seedlings, both cotyledons of one-week old seedlings were harvested (because no further growth occurred). Heterozygous *sulf (het)* and wild-type seedlings developed true leaves. Phenotypes were documented for up to six weeks. Paramutation events in heterozygous plants were easily recognizable as leaf variegations (i.e., the appearance of pale green or yellow sectors).

In vitro cultivation of *sulf (hom)* and isolation of *SOSU* and *ROSU* lines

As autotrophic and heterotrophic growth of *sulf (hom)* seedlings is restricted to the cotyledon stage, surface-sterilized seeds were germinated on MS medium (with 2% sucrose; Murashige and Skoog, 1962). *sulf (hom)* seedlings (identified by their pale green to yellow cotyledons) were continued to be propagated on MS medium with 2% sucrose and 5 µM IAA (Duchefa) under low-light conditions (2.5 μmol quanta m^{-2} s^{-1}) and multiplied via stem cuttings (Ehlert et al., 2008). To isolate revertants (*ROSU*) and suppressors (*SOSU*) of paramutation, large-scale *in vitro* cultures of *sulf (hom)* tissue were set up and visually searched for spontaneously appearing dark green spots. Identified green spots were cut out and regenerated into plantlets on MS medium containing 1 mg/L zeatin (Duchefa) under a light intensity of 100 μmol quanta

m^{-2} s^{-1}. Regenerated plantlets of putative *SOSU* and *ROSU* plants were rooted under sterile conditions on MS medium containing 2% sucrose, transferred to the greenhouse and grown to maturity under standard conditions (250 µmol quanta m^{-2} s^{-1}).

Genetic complementation of *sulfurea*

A cDNA clone of *At3g08010* (pda07273) was obtained from the RIKEN collection of full-length *Arabidopsis* cDNA clones (Seki et al., 2002). The cDNA fragment was excised from the modified pBluescript vector (λZAPIII) by digestion with the restriction enzyme SfiI, ligated into plant transformation vector pBI121 and transformed into *Agrobacterium tumefaciens* strain GV2260. Heterozygous *sulf (het)* cotyledons were used for Agrobacterium-mediated plant transformation using standard transformation protocols. Selection of transformed cells was performed on MS medium containing kanamycin (50 mg/L), zeatin (1 mg/L) and 2% sucrose. Regenerated plantlets were transferred to hormone-free MS medium with 2% sucrose to induce rooting. Confirmed transgenic lines, referred to as *sulf (com)*, were subsequently grown to maturity under standard greenhouse conditions.

RNA extraction and RNA gel blot analysis

Total RNA was isolated from samples of ~100 mg leaf material using the peqGold Trifast reagent (PeqLab) and following the manufacturer’s protocol. Samples of 4 µg RNA were denatured and electrophoretically separated in formaldyhyde-containing 1% (w/v) agarose gels followed by transfer onto Hybond^{TM} XL nylon membranes (GE Healthcare) by capillary blotting. Hybridization was performed at 65°C using [α-^{32}P]dCTP-labeled probes produced by random priming (Megaprime^{TM} DNA labeling system; GE Healthcare). 25 ng of purified PCR products were radiolabeled and used as hybridization probes. For primer sequences, see Supplemental Table 8.

cDNA synthesis

To remove residual contaminating DNA, extracted total RNA was treated with Turbo DNase (Thermo Fisher Scientific) according to the manufacturer’s protocol. Samples of 2 µg RNA were used for first-strand reverse transcription into cDNA with SuperScript III reverse transcriptase (Thermo Fisher Scientific). 1 µl of the cDNA synthesis reaction was used for amplification of target sequences by PCR. Amplification products were separated be electrophoresis in 1% (w/v) agarose gels. DNA bands of interest were cut out of the gel and eluted using the NucleoSpin Gel and PCR Clean-up kit (Macherey-Nagel).

DNA isolation and polymerase chain reactions (PCR)

Total plant DNA was isolated from samples of ~100 mg leaf material using a CTAB-based method (Doyle and Doyle, 1990). Extracted DNAs were treated with RNase A for 30 min at 37°C and subsequently purified by precipitation with isopropanol. PCR reactions were performed with DreamTaq Polymerase (Thermo Fisher Scientific) following the protocol of the manufacturer. Primers used for amplification of specific gene sequences are listed in Supplementary Table 8.

Microarray analysis of gene expression

Microarray analysis was performed using the Potato Gene Expression Microarray (POCI Array; AMADID 015425; https://earray.chem.agilent.com/earray/PublishDesignLogin.do?eArrayAction=showPreviewForLogin&publishdesignid=PD408321788). Briefly, isolated total RNA was used for amplification with Low RNA Input Linear Amp-Kit (Agilent Technologies). 1 µg RNA was reverse transcribed into cDNA, and subsequently converted into cRNA and fluorescently labelled. Samples of 1.63 µg labelled cRNA were hybridized to microarrays following the Agilent 60-mer oligo microarray processing protocol (MACS Molecular). Microarrays were scanned using the Agilent Microarray Scanner system (Agilent Technologies).

Next-generation genomic sequencing and methylation sequencing of bisulfite-treated genomic DNA

5 g of very young frozen leaf material of *sulf (het-g)*, *sulf (het-y)*, *SOSU1* and wild-type seedlings were used for genomic sequencing and methylation sequencing. *sulf (het-g)* and *sulf (het-y)* tissue from two heterozygous *sulf (het)* plants was pooled to obtain sufficient amounts of young leaf tissue for DNA extraction. DNA isolation, bisulfite treatment and next-generation DNA sequencing were performed according to published protocols (Schmitz et al., 2011) using the tomato genome (The Tomato Genome Consortium, 2012) as scaffold. Paired-end Illumina sequencing was performed on a HiSeq instrument (IGA Technology Services, Udine, Italy). Sequencing data were deposited in the European Nucleotide Archive (ENA) under the BioProject PRJEB15487.

Processing of DNA sequencing data

Quality control (QC) of all sequencing data was performed with FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). After QC review, genomic sequencing (DNA-seq) data were trimmed at the 3’ end by five bases and bisulfite sequencing (BS-seq) data at the 3’ end by six bases and at the 5’ end by five bases. Undefined bases at the ends were removed. A minimal read length of 95 and 89 bp was imposed on DNA-seq and BS-seq data, respectively, as different numbers of bases were trimmed.

After trimming, a two-tier mapping approach was performed using BWA version 0.712 (Li and Durbin, 2009) in aln mode and BSMAP version 2.90 (Xi and Li, 2009) for DNA-seq and BS-seq data, respectively. In both cases, data were aligned in the first mapping step to the plastid (NC_007898.3; Kahlau et al., 2006) and mitochondrial (AFYB00000000.1) genomes of *Solanum lycopersicum* and only those pairs were kept, where both mates remained unmapped. In the second mapping step, both datasets were aligned to chromosomes 1 to 12 of the Sol Genomics Network *Solanum lycopersicum* genome version SL2.50 (Fernandez-Pozo et al., 2015). Samtools v1.0 was used to create, sort, index and merge the alignment data sample-wise in BAM format (Li et al., 2009). The Picard (http://broadinstitute.github.io/picard) plugins FixMateInformation, CleanSam and MarkDuplicates were applied to the merged data in order to remove duplicates from the samples.

All remaining DNA-seq reads were then processed by the CNV pipeline cnv-seq version 2014/08/12 (Xie and Tammi, 2009) and the Delly deletion pipeline version 0.7.1 (Rausch et al., 2012) to detect copy number variations and putative deletion events, respectively. Both pipelines were run with default parameters, a window size set to 25 kb and an expected genome size of 802.1 Mb for the cnv-seq Perl script. Genomic segments in the CNV-seq results which had a p-value lower than 0.001 (reflecting the significance cut-off recommended in the cnv-seq manual) and a log2 value above or below the 99.75% quantile were merged with mergeBed of BEDTools (Quinlan and Hall, 2010) and kept only if the merged region was larger than 100 kb. The resulting regions were checked for confirmation by the Delly output and visualized using a custom R script. Genes affected by deletion events were illustrated with FancyGene (Rambaldi and Ciccarelli, 2009).

BS-seq data were processed using the Python script meth_ratio.py included in the BSMAP package to calculate base-wise methylation values reporting additionally zero methylation values (-z option). Afterwards, methylKit was utilized to calculate differential methylation values for the genome features exons, introns and promoters (1 kb upstream of the coding region) of the ITAG2.4 annotation by Sol Genomics Network (Fernandez-Pozo et al.,

2015) and tile-wise for 500 bp windows overlapping by 250 bp over the whole genome according to the recommended workflow described in the methylKit manual (Akalin et al., 2012). Differential methylation statistics for the genomic features were visualized with the R package lattice (barchart). Base-wise methylation patterns of genes of interest were visualized with a custom R script.

Supplemental Data

Supplemental Table 1. Lists of genes deregulated in young *sulf (hom)* tissue.

Supplemental Table 2. Expected occurrence of the paramutation phenotype in the progeny of complemented lines that were hemizygous for the transgene.

Supplemental Table 3. Summary of the occurrence of the different phenotypes in the segregating T_2 generation heterozygous for the transgene *At3g08010*.

Supplemental Table 4. Descriptive filtering and mapping statistics of the four samples subjected to DNA resequencing.

Supplemental Table 5. Descriptive filtering and mapping statistics of the four samples subjected to methylome sequencing.

Supplemental Table 6. Summary of the occurrence of the different phenotypes in the segregating F1 generation of *SOSU* plants that were either selfed or crossed to the wild type.

Supplemental Table 7: List of genes affected by the 375 kb genomic deletion in chromosome 7 that was identified in the suppressor mutant *SOSU1*.

Supplemental Table 8. List of oligonucleotides used in this study.

Supplemental Figure 1. Comparative analysis of gene expression levels of all genes on tomato chromosome 2 represented on the microarray.

Supplemental Figure 2. CNV analyses of chromosome 2 for *sulf (het-g)* and *sulf (het-y)* tissue compared to the wild type.

Supplemental Figure 3. Isolation of the putative suppressor mutants *SOSU3-5* and expression analysis of the *SULF* gene *Solyc02g005200*.

Supplemental Figure 4. Analysis of the progeny of *SOSU3*.

Supplemental Figure 5. Detection of a 375 kb genomic deletion in chromosome 7 of the suppressor mutant *SOSU1* by CNV analysis.

Supplemental Dataset 1. CpG methylation levels at the *SULFUREA* locus and its adjacent regions as measured by methylome sequencing.

Supplemental Dataset 2. CHG methylation levels at the *sulfurea* locus and its adjacent regions as measured by methylome sequencing.

ACKNOWLEDGMENTS

We are grateful to Professor Rudolf Hagemann (Halle/S., Germany) and the IPK Gatersleben (Germany) for providing seeds of the *sulfurea* mutant. We thank Dr. Dirk Walther for advice on bioinformatic analyses, Sandra Stegemann, Steffi Seeger und Helga Kulka for excellent technical assistance, and the MPI-MP Green Team for plant care and cultivation. This paper is dedicated to Professor Rudolf Hagemann, discoverer of paramutation in tomato, on the occasion of his 85th birthday on October 21, 2016.

AUTHOR CONTRIBUTIONS

B.E. and A.F. designed and performed research and analyzed data. R.B. conceived of the study, designed research, analyzed data and wrote the paper with input from all co-authors.

REFERENCES

Akalin, A., Kormaksson, M., Li, S., Garrett-Bakelman, F.E., Figueroa, M.E., Melnick, A. and Mason, C.E. (2012). methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. Genome Biol. **13**, R87.

Alleman, M., Sidorenko, L., McGinnis, K., Seshadri, V., Dorweiler, J.E., White, J., Sikkink, K. and Chandler, V.L. (2006). An RNA-dependent RNA polymerase is required for paramutation in maize. Nature **442**, 295-298.

Amedeo, P., Habu, Y., Afsar, K., Mittelsten Scheid, O. and Paszkowski, J. (2000). Disruption of the plant gene MOM releases transcriptional silencing of methylated genes. Nature **405**, 203-206.

Barbour, J.-E.R., Liao, I.T., Stonaker, J.L., Lim, J.P., Lee, C.C., Parkinson, S.E., Kermicle, J., Simon, S.A., Meyers, B.C., Williams-Carrier, R., Barkan, A. and Hollick, J.B. (2012). required to maintain repression2 is a novel protein that facilitates locus-specific paramutation in maize. Plant Cell **24**, 1761-1775.

Barneche, F., Winter, V., Crèvecoeur, M. and Rochaix, J.-D. (2006). ATAB2 is a novel factor in the signalling pathway of light-controlled synthesis of photosystem proteins. EMBO J. **25**, 5907-5918.

Belele, C.L., Sidorenko, L., Stam, M., Bader, R., Arteaga-Vazquez, M.A. and Chandler, V.L. (2013). Specific tandem repeats are sufficient for paramutation-induced trans-generational silencing. PLoS Genet. **9**, e1003773.

Blevins, T., Pontvianne, F., Cocklin, R., Podicheti, R., Chandrasekhara, C., Yerneni, S., Braun, C., Lee, B., Rusch, D., Mockaitis, K., Tang, H. and Pikaard, C.S. (2014). A two-step process for epigenetic inheritance in *Arabidopsis*. Mol. Cell **54**, 30-42.

Bond, D.M. and Baulcombe, D.C. (2014). Small RNAs and heritable epigenetic variation in plants. Trends Cell Biol. **24**, 100-107.

Chandler, V. and Alleman, M. (2008). Paramutation: Epigenetic instructions passed across generations. Genetics **178**, 1839-1844.

Chandler, V.L. (2007). Paramutation: From maize to mice. Cell **128**, 641-645.

Chandler, V.L. and Stam, M. (2004). Chromatin conversations: mechanisms and implications of paramutation. Nature Rev. **5**, 532-544.

Cokus, S.J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C.D., Pradhan, S., Nelson, S.F., Pellegrini, M. and Jacobsen, S.E. (2008). Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. Nature **452**, 215-219.

Cubas, P., Vincent, C. and Coen, E. (1999). An epigenetic mutation responsible for natural variation in floral symmetry. Nature **401**, 157-161.

Dauvillée, D., Stampacchia, O., Girard-Bascou, J. and Rochaix, J.-D. (2003). Tab2 is a novel conserved RNA binding protein required for translation of the chloroplast psaB mRNA. EMBO J. **22**, 6378-6388.

Dorweiler, J.E., Carey, C.C., Kubo, K.M., Hollick, J.B., Kermicle, J.L. and Chandler, V.L. (2000). mediator of paramutation1 is required for establishment and maintenance of paramutation at multiple maize loci. Plant Cell **12**, 2101-2118.

Doyle, J.J. and Doyle, J.L. (1990). Isolation of plant DNA from fresh tissue. Focus **12**, 13-15.

Ehlert, B., Schöttler, M.A., Tischendorf, G., Ludwig-Müller, J. and Bock, R. (2008). The paramutated SULFUREA locus of tomato is involved in auxin biosynthesis. J. Exp. Bot. **59**, 3635-3647.

Erhard Jr., K.F., Stonaker, J.L., Parkinson, S.E., Lim, J.P., Hale, C.J. and Hollick, J.B. (2009). RNA polymerase IV functions in paramutation in Zea mays. Science **323**, 1201-1205.

Fernandez-Pozo, N., Menda, N., Edwards, J.D., Saha, S., Tecle, I.Y., Strickler, S.R., Bombarely, A., Fisher-York, T., Pujar, A., Foerster, H., Yan, A. and Mueller, L.A. (2015). The Sol Genomics Network (SGN) - from genotype to phenotype to breeding. Nucleic Acids Res. **43**, D1036-1041.

Giacopelli, B.J. and Hollick, J.B. (2015). Trans-homolog interactions facilitating paramutation in maize. Plant Physiol. **168**, 1226-1236.

Gouil, Q., Novák, O. and Baulcombe, D.C. (2016). SLTAB2 is the paramutated SULFUREA locus in tomato. J. Exp. Bot. **67**, 2655-2664.

Hagemann, R. (1958). Somatic conversion in Lycopersicon esculentum Mill.. Z. Vererbungsl. **89**, 587-613.

Hagemann, R. (1993). Studies towards a genetic and molecular analysis of paramutation at the sulfurea locus of Lycopersicon esculentum Mill. Molecular Biology of Tomato. Yoder, J. I. (ed.), Technomic Publ. Co., Davis, Lancaster, Basel, 75-82.

Hagemann, R. and Berg, W. (1977). Vergleichende Analyse der Paramutationssystem bei höheren Pflanzen. Biol. Zbl. **96**, 257-301.

Hale, C.J., Stonaker, J.L., Gross, S.M. and Hollick, J.B. (2007). A novel Snf2 protein maintains trans-generational regulatory states established by paramutation in maize. PLoS Biol. **5**, 2156-2165.

Hollick, J.B., Dorweiler, J.E. and Chandler, V.L. (1997). Paramutation and related allelic interactions. Trends Genet. **13**, 302-308.

Hollick, J.B., Kermicle, J.L. and Parkinson, S.E. (2005). Rmr6 maintains meiotic inheritance of paramutant states in Zea mays. Genetics **171**, 725-740.

Jiang, C., Mithani, A., Gan, X., Belfield, E.J., Klingler, J.P., Zhu, J.-K., Ragoussis, J., Mott, R. and Harberd, N.P. (2011). Regenerant *Arabidopsis* lineages display a distinct genome-wide spectrum of mutations conferring variant phenotypes. Curr. Biol. **21**, 1385-1390.

Kaeppler, S.M. and Phillips, R.L. (1993). Tissue culture-induced DNA methylation variation in maize. Proc. Natl. Acad. Sci. USA **90**, 8773-8776.

Kahlau, S., Aspinall, S., Gray, J.C. and Bock, R. (2006). Sequence of the tomato chloroplast DNA and evolutionary comparison of solanaceous plastid genomes. J. Mol. Evol. **63**, 194-207.

Kawano, N., Kawano, T. and Lapeyrie, F. (2003). Inhibition of the indole-3-acetic acid-induced epinastic curvature in tobacco leaf strips by 2,4-dichlorophenoxyacetic acid. Ann. Bot. **91**, 465-471.

Kramer, G., Boehringer, D., Ban, N. and Bukau, B. (2009). The ribosome as a platform for co-translational processing, folding and targeting of newly synthesized proteins. Nature Struct. Mol. Biol. **16**, 589-597.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics **25**, 1754-1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics **25**, 2078-2079.

Lisch, D., Carey, C.C., Dorweiler, J.E. and Chandler, V.L. (2002). A mutation that prevents paramutation in maize also reverses Mutator transposon methylation and silencing. Prot. Natl. Acad. Sci. USA **99**, 6130-6135.

Miguel, C. and Marum, L. (2011). An epigenetic view of plant cells cultured in vitro: somaclonal variation and beyond. J. Exp. Bot. **62**, 3713-3725.

Mittelsten Scheid, O., Afsar, K. and Paszkowski, J. (2003). Formation of stable epialleles and their paramutation-like interaction in tetraploid *Arabidopsis thaliana*. Nature Genet. **34**, 450-454.

Murashige, T. and Skoog, F. (1962). A revised medium for rapid growth and bio assays with tobacco tissue culture. Physiol. Plant. **15**, 473-497.

Pechmann, S., Willmund, F. and Frydman, J. (2013). The ribosome as a hub for protein quality control. Mol. Cell **49**, 411-421.

Probst, A.V., Fagard, M., Proux, F., Mourrain, P., Boutet, S., Earley, K., Lawrence, R.J., Pikaard, C.S., Murfett, J., Furner, I., Vaucheret, H. and Mittelsten Scheid, O. (2004). *Arabidopsis* histone deacetylase HDA6 is required for maintenance of transcriptional gene silencing and determines nuclear organization of rDNA repeats. Plant Cell **16**, 1021-1034.

Quinlan, A.R. and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics **26**, 841-842.

Rambaldi, D. and Ciccarelli, F.D. (2009). FancyGene: dynamic visualization of gene structures and protein domain architectures on genomic loci. Bioinformatics **25**, 2281-2282.

Rassoulzadegan, M., Grandjean, V., Gounon, P., Vincent, S., Gillot, I. and Cuzin, F. (2006). RNA-mediated non-mendelian inheritance of an epigenetic change in the mouse. Nature **441**, 469-474.

Rausch, T., Zichner, T., Schlattl, A., Stütz, A.M., Benes, V. and Korbel, J.O. (2012). DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics **28**, i333-i339.

Regulski, M., Lu, Z., Kendall, J., Donoghue, M.T.A., Reinders, J., Llaca, V., Deschamps, S., Smith, A., Levy, D., McCombie, W.R., Tingey, S., Rafalski, A., Hicks, J., Ware, D. and Martienssen, R.A. (2013). The maize methylome influences mRNA splice sites and reveals widespread paramutation-like switches guided by small RNA. Genome Res. **23**, 1651-1662.

Renner, O. (1938). Über Oenothera atrovirens Sh. et Bartl. und über somatische Konversion im Erbgang des cruciata-Merkmals der Oenotheren. Z. indukt. Abst. u. Vererbungsl. **74**, 91-124.

Schmitz, R.J., Schultz, M.D., Lewsey, M.G., O'Malley, R.C., Urich, M.A., Libiger, O., Schork, N.J. and Ecker, J.R. (2011). Transgenerational epigenetic instability is a source of novel methylation variants. Science **334**, 369-373.

Seki, M., Narusaka, M., Kamiya, A., Ishida, J., Satou, M., Sakurai, T., Nakajima, M., Enju, A., Akiyama, K., Oono, Y., Muramatsu, M., Hayashizaki, Y., Kawai, J., Carninci, P., Itoh, M., Ishii, Y., Arakawa, T., Shibata, K., Shinagawa, A. and Shinozaki, K. (2002). Functional annotation of a full-length *Arabidopsis* cDNA collection. Science **296**, 141-145.

Sidorenko, L. and Peterson, T. (2001). Transgene-induced silencing identifies sequences involved in the establishment of paramutation of the maize p1 gene. Plant Cell 13, 319-335.

Smulders, M.J.M. and de Klerk, G.J. (2011). Epigenetics in plant tissue culture. Plant Growth Regul. **63**, 137-146.

Stam, M. (2009). Paramutation: a heritable change in gene expression by allelic interactions in trans. Mol. Plant **2**, 578-588.

Stam, M. and Mittelsten Scheid, O. (2005). Paramutation: an encounter leaving a lasting impression. Trends Plant Sci. **10**, 283-290.

Stam, M., Belele, C., Dorweiler, J.E. and Chandler, V.L. (2002). Differential chromatin structure within a tandem array 100 kb upstream of the maize b1 locus is associated with paramutation. Genes & Dev. **16**, 1906-1918.

Stam, M., Belele, C., Ramakrishna, W., Dorweiler, J.E., Bennetzen, J.L. and Chandler, V.L. (2002). The regulatory regions required for B´ paramutation and expression are located far upstream of the maize b1 transcribed sequences. Genetics **162**, 917-930.

Stroud, H., Ding, B., Simon, S.A., Feng, S., Bellizzi, M., Pellegrini, M., Wang, G.-L., Meyers, B.C. and Jacobsen, S.E. (2013). Plants regenerated from tissue culture contain stable epigenome changes in rice. eLife **2**, e00354.

Tanksley, S.D., Ganal, M.W., Prince, J.P., de Vicente, M.C., Bonierbale, M.W., Broun, P., Fulton, T.M., Giovannoni, J.J., Grandillo, S., Martin, G.B., Messeguer, R., Miller, J.C., Miller, L., Paterson, A.H., Pineda, O., Röder, M.S., Wing, R.A., Wu, W. and Young, N.D. (1992). High density molecular linkage maps of the tomato and potato genomes. Genetics **132**, 1141-1160.

The Tomato Genome Consortium (2012). The tomato genome sequence provides insights into fleshly fruit evolution. Nature 485, 635-641.

Wisman, E., Ramanna, M.S. and Koorneef, M. (1993). Isolation of a new paramutagenic allele of the sulfurea locus in the tomato cultivar Moneymaker following in vitro culture. Theor. Appl. Genet. **87**, 289-294.

Woodhouse, M.R., Freeling, M. and Lisch, D. (2006). Initiation, establishment, and maintenance of heritable MuDR transposon silencing in maize are mediated by distinct factors. PLoS Biol. **4**, 1678-1688.

Xi, Y. and Li, W. (2009). BSMAP: whole genome bisulfite sequence MAPping program. BMC Bioinform. **10**, 232.

Xie, C. and Tammi, M.T. (2009). CNV-seq, a new method to detect copy number variation using high-throughput sequencing. BMC Bioinform. **10**, 80.

Zhao, Y. (2008). The role of local biosynthesis of auxin and cytokinin in plant development. Curr. Op. Plant Biol. **11**, 16-22.

3 Discussion

With the rise of HTS technologies in the beginning of the 21th century and its vastly dropping costs, the postgenomic era became even more accelerated and is now challenged by analyzing and interpreting the huge amounts of HTS data properly. As in many research fields, NMI research depends nowadays on availability and quality of nuclear and organelle genomes for the organism of interest. Considering the chronological appearance of all three genomes of *Arabidopsis thaliana* (1994 PMG; 1999 Plastome; 2000 Nuclear genome) and that organelle genome types are more or less equal in size (within the same order of magnitude) the number of available land plant genomes categorized by the genomic compartment is quite unbalanced. (NCBI records September 2021: 6,662 chloroplasts versus 1,663 nuclear genomes versus 288 mitochondrial genomes). This could be explained by research interest being focused on chloroplast genomes, especially for phylogenetic studies, but mainly by the minor technical obstacles. While genome skimming and the structurally conserved and repeat-poor nature of chloroplast genomes makes it easy, in normal case, to obtain a plastome sequence, it lacks applicable post-Multiple sequence alignment tools linking polymorphic genes/regions to numeric phenotypic measurements. To address this issue, I developed a post-multiple sequence alignment algorithm, called correlation mapping (CM), which correlates segment-wise numbers of nucleotide changes to a numeric ascertainable phenotype to identify genes that may cause the competitive behavior of plastid genotypes (Paper 2).

For PMGs, it is necessary to even take a step back in the context of standardized workflows. Such a straight-forward workflow as for chloroplast assemblies is not applicable for PMGs as they exhibit a high degree of complexity. Our success in this regard (Paper 1) is based on an adapted and optimized workflow and the developed methodology. By approaching the post-assembly task from a graph-based perspective, we were not only able to construct a spatial model that captures the complexity of PMGs, but additionally to predict a defined spectrum of alternative PMG isoforms for the very first time, to the best of my knowledge.

It becomes even more complex if NMI phenomena like paramutation are to be investigated which acts on multiple biological levels including DNA methylation of the nuclear genome. Here, I implemented a pipeline to analyze DNA and bisulfite sequencing data as well as microarray data to identify the gene which is causing for the phenotype (Paper 3).

3.1 Limitations of assembler programs

By the assemblathon initiative (https://assemblathon.org/), which brought bioinformaticians together from around the world, it quickly became clear that large algorithmic differences of available assembler programs led to various results if applied to the same NGS dataset (Earl et al., 2011; Bradnam et al., 2013). This might be frustrating at first sight but can also be beneficial to find a program within this pool of diverse algorithms/programs which can handle specific biological peculiarities in an organism/genome of interest.

3.1.1 Minimizing biology-derived biases

Besides choosing the right assembler, a minimization of biology-derived biases paramount to wet lab decisions but unfortunately is often ignored or at least neglected. In our study (Paper 1), initial tests revealed that only one out of four *de novo* assemblers, namely IDBA, met our criteria to proceed with. On the wet lab side, using optimized NGS libraries generated with insightful decisions (DNA-extraction from isolated mitochondria of mature leaves) ended up in purity levels of mtDNA above 95%. But the impossibility to reach 100% can be explained by two types of noise, technical and biological. For example, technical noise might arise by difficulties in isolating a pure fraction from the sucrose gradient in the mitochondria isolation process. Biological noise, although its manifestation, impurity, is similar, has a different origin which is promiscuous DNA that resulted from intercompartimental DNA exchange. DNA transfer occurred very often during plant evolution from the plastid into the nucleus and mitochondria (Ellis, 1982; Knoop, 2004; Leister, 2005; Wang et al., 2007) as well as in some cases even from the nucleus to the mitochondrion, as also described very early within the evening primrose (Schuster and Brennicke, 1987; Kleine et al., 2009). Obviously, promiscuous DNA does interfere in assembly strategies when starting with whole genome sequencing data generated from total DNA extract. Available reference-based such as mitoBIM (Hahn et al., 2013) and NOVOPlasty (Dierckxsens et al., 2017) as well as also kmer-based assemblers such as Norgal (Al-Nakeeb et al., 2017) affirm to be able to assemble mitochondrial genomes. For example, mitoBIM is a wrapper program, that executes MIRA in an iterative two-tier approach to reconstruct mitochondrial genomes (Hahn et al., 2013). First, by baiting perceived mitochondrial reads from total DNA sequence with known mitochondrial genes and/or genomic sequences using MIRA in mapper mode. In a second step it assembles those reads using the MIRA in assembler mode. Using MIRA, we were not able to generate a circular graph for mtDNA-enriched DNA, and we assume it will most likely run into problems for a total DNA dataset as well.

However, and probably more important, all reference-based and kmer-based assemblers catch nuclear mitochondrial DNA sequences independently of advantageous kmer distributions in generated HTS data. Per definition, they do not belong to the mitochondrial genome, underlie higher mutation rates and most important different recombination events (Wolfe et al., 1987; Drouin et al., 2008; Michalovova et al., 2013). This can lead to incorrect assemblies and/or introduces SNP/indels which do not belong to the PMG itself. The authors of the Norgal program explicitly mention that assembly outcomes from NGS data in which kmer distributions are inseparable, should be handled with caution (Al-Nakeeb et al., 2017).

3.1.2 Benefits of discontiguous mitochondrial genome assemblies

By explicitly using mainly Illumina paired-end short reads for the *de novo* assembly, we deliberately generated a fragmented and discontiguous assembly output. It turned out, that only IDBA was able to split assemblies into structural units of repetitive and non-repetitive sequences, which was essential for our post-assembly methodology and can be explained by the implemented De Bruijn graph algorithm (Peng et al., 2012). There is a simple explanation not to use PacBio and Illumina Mate pair data for initial assemblies: Because available assembly programs used for PacBio data, such as Falcon or Canu (Chin et al., 2016; Koren et al., 2017), are not designed to work with the smallest structural components of an assembly. Consequently, those assemblers could not resolve complex nested repeats and are incapable to generate all distinct isoforms which may exist in PMGs. The same is true for incorporating mate pair data within *de novo* assembly or scaffold processes itself. To present but two examples, researchers tried to use mate pair data to scaffold contigs further (Guo et al., 2017) or to assemble a PMG with just PacBio data (Varré et al., 2019). Both groups were unsuccessful to reach one master circle and predicted instead two or more autonomous circles. Interestingly in both cases some of the autonomous circles carry a subset of exons of the same trans-spliced gene each.

3.2 Rethinking assembly visualization – supporting more than one purpose

Besides the need of implementing a new assembly visualization tool because of incapabilities of and missing features in available programs (Bandage, Contiguity, and missing output formats in IDBA, for details see discussion paper 1), another demand motivated such an implementation: The capability to predict PMG isoforms with defined parameters. With such a unified toolset, we are at the tip of the spear of PMG research by hardening the existence of different PMG isoforms in various amounts at the same time. Only a few publications have so far reported about PMG isoform stoichiometries which were similar (Guo et al., 2016) or different (Kozik et al., 2019). Additionally, I would like to stress that SAGBAC can be become also relevant for the chloroplast research community in the future as it was previously shown that chloroplast genomes exist which are more complex regarding their repetitiveness and duplicated gene content (Barnard-Kubow et al., 2014).

3.3 Organelle genome reconstruction and harmonization for comparative genomics

Plastidial and mitochondrial genomes have their own biological characteristics which also reflected in different ways to obtain them from assembly outputs. Plastidial genomes are mostly assembled indeed in one contig because of their conserved and repeat-poor nature. But two minor post assembly steps are necessary to obtain a well-annotated plastome genome: (a) restoring the canonical quadripartite structure (Gordon et al., 1981, 1982) by bringing its pair of inverted repeats (IR) which separates the two single copy units, namely large single copy (LSC) and the small single copy (SSC), into the right order and orientation. This is necessary, because of the circularity of and the widely present IR in chloroplast genomes, *de novo* assemblers are order- and orientation-insensitive in regards of the three main genomic regions leading to 8 different combinations. (b) Determining the IR boundaries as they can vary, even between subspecies/cultivars of the same species (Zhu et al., 2016). But more important IR boundaries define the start of the LSC and the end of the SSC and pinpoint the true start and end of the genome sequence itself within the single contig assembly. Both post assembly curation steps are needed to harmonize plastidial genomes for a multiple sequence alignment in order to perform comparative genomics analyses (Paper 2).

Such a harmonization of PMGs is not possible in the classical sense as by the evolutionary occurrence of different sets of RRPs, there are different main genome configurations (Paper 1). Possible solutions would be programs like minigraph or an implementation of so called 2,5D graph-overlay algorithms (Eades and Feng, 1996; Brandes et al., 2004). Minigraph (https://github.com/lh3/minigraph) is a sequence-to-graph mapper, which was built for pan genome investigation as well as core and shell genome identification but is still being tested. An implementation of 2,5D graph overlay algorithms is not trivial. To be able to superimpose a set of graphs generated by SAGBAC, all contigs from all participating graphs needed to be fragmented to their smallest common denominators.

3.4 Linking comparative genomics to phenotypes

To perform a comparative genome analysis a multiple sequence alignment (MSA) is recommended. Classic hierarchical clustering of protein or coding sequences would only be meaningful when all sequences are clustering together each according to its phenotypic classifier (e.g. inheritance strength). A disadvantage would be that promotors and intergenic spacers would be ignored in such an analysis. The small genome size of chloroplast genomes compared to nuclear genomes, the low number of polymorphisms as well as the inadequately low sample size made it inadequate to perform a genome-wide association study (Nishino et al., 2018). Instead, as implemented in the 2nd publication, by applying a segmentation algorithm normally used in CNV analyses, the complete chloroplast sequence can be addressed including promotors and intergenic spacers. Additionally, such an approach is complete polymorphism-type independent, and any combination of single nucleotide polymorphisms, deletions and insertions can be considered in the calculations. By acquiring segment-wise numbers of nucleotide changes, a correlation to a numeric ascertainable phenotype is feasible. As the correlations of every segment (Pearson as well as Spearman) are not independent, an obligatory p value adjustment for multiple testing correction (Benjamini/Hochberg) is applied.

In the 2nd publication, our correlation mapping approach identified genetic loci that appear to be responsible for the difference in chloroplast competition. Outcompeting the weaker chloroplast results in uniparental inheritance at least in evening primrose (Kirk and Tilney-Basset, 1978; Chiu et al., 1988; Barnard-Kubow et al., 2017). Interestingly, instead of a connection to chloroplast DNA replication or copy number (Nishimura and Stern, 2010), chloroplast competition is essentially a metabolic phenotype by an altered lipid composition. The underlying molecular loci are rapidly evolving genes known throughout the plant kingdom. Ycf1, Ycf2 and also the N-terminus of AccD are highly variable in (Greiner et al., 2008; Wicke et al., 2011; de Vries et al., 2015). These loci, which are involved in chloroplast competition, are very sensitive to replication slippage, the main mechanism for the occurrence of spontaneous chloroplast mutations (Greiner et al., 2008; Massouh et al., 2016). Because of their high mutation rates, cytoplasmic drive loci can evolve and become fixed in a population very quickly. As the newly evolved aggressive plastomes I and III in *Oenothera* do not form a phylogenetic clade, evolution and fixation of both genotypes must have happened independently within a very short time (Greiner et al., 2008).

3.5 Bisulfite-sequencing as paramutation loci detector

In the third publication, we identified Solyc02g005200 as the *SULFUREA* gene of tomato that encodes for the *ATAB2* gene in *Arabidopsis thaliana* (At3g08010) which is a chloroplast-localized RNA-binding protein. In *Arabidopsis thaliana* it functions as translational activator by controlling the expression of the reaction center subunits of the photosystems in response to light (Barneche et al., 2006). Previous physiological analysis revealed strongly reduced amounts of photosystem I compared to just a lightly affected photosystem II (Ehlert et al., 2008). By stable genetic complementation with the homologous gene of *Arabidopsis*, we not only determined without doubt that the homologue of the *ATAB2* gene is identical with the *SULF* locus but also demonstrated that the *ATAB2* gene is not paramutable.

Bisulfite-sequencing analysis revealed a candidate cytosine upstream of the *SULF* locus that exactly showed the expected methylation difference between *sulf (het-y)*, *sulf (het-g)* and wild type with 100%, 50 % and 0% methylation respectively. Single CpG regulation was also found in human (Tsuboi et al., 2017). Almost all other methylated cytosines in CpG and CHG context are shared between *sulf (het-g)* and *sulf(het-y)* and therefore, cannot be responsible for the switch between the active and inactive state. In our suppressor mutant SOSU1 (Suppressor of *SULFUREA* 1), this candidate cytosine is unchanged but interestingly it shows a genome-wide changed DNA methylation pattern. As the deletion event on chromosome 7 resulted in a fusion gene involving a histone deacetylase (Solyc07g065550), this globally altered methylome can be explained by a deregulated histone modification machinery.

3.6 Repetitive elements: a blessing curse - biological and bioinformatical

Repetitive elements especially transposons are a blessing curse at the same time in respect of the dual nature of transposable elements which is better reviewed elsewhere (Dubin et al., 2018). Here I just want to point out that in all three presented studies, different types of repetitive elements are present and rather are a curse for a proper analyze of the underlying NGS data.

In publication 1, recombinogenic repeat pairs are not only a source of recombination per se but serving another functionality. Recombination dependent DNA repair maintains genome integrity as within the matrix of mitochondria present oxidative radicals are damaging DNA. Which repeats act as recombiners is controlled by the nucleus (Arrieta-Montiel et al., 2009). To distinguish between repeats and true RRPs, the contig connectivity of RRPs in regards of their CRCs in “double forks” needs be preserved as accomplished by our post assembly pipeline SAGBAC. The result of losing that information can lead to more than one autonomous circle as previously pointed out.

Within the study presented in publication 2, a pair of large inverted repeats are present which are part of the canonical quadripartite structure in most plastomes (Gordon et al., 1981, 1982). But their boundaries and with them the true ends of plastome sequences can only be achieved by analyzing so called split reads. NGS reads are cut in two parts in order to align both parts properly instead of being discarded entirely. Another type of repeats in chloroplast genomes of *Oenothera* is a major source of spontaneous chloroplast mutations which is caused by a mechanism called replication slippage (Massouh et al., 2016). Slippage conferred by misalignments of or formed hairpin structures at tandem/direct repeats or palindrome/inverted repeats are resulting in deletions or duplications when occurring at leading or lagging (nascent) strand respectively. Long Sanger sequencing reads were necessary to span these stretches of repetitive elements to fully resolve them.

In publication 3, repetitive elements upstream of the identified gene (Soly02g005200) are potential enhancer elements but are influencing the mappability of short NGS reads and with that the interpretability of methylation calls. Methylation calling algorithms often recommend to focus on uniquely mapped and properly paired read pairs and do not consider uneven coverage present at repetitive loci. A good example is the b1 locus in maize where seven copies of a repetitive enhancer element are present. With short NGS reads it can neither determined how many copies are existing, nor which copies of the repetitive element are methylated or unmethylated.

3.7 Future directions

3.7.1 mtDNA sequencing - beyond PMG reconstruction

With our proof-of-concept study (paper 1) we were able to show that the newly developed SAGBAC pipeline and its ISEIS core algorithm reconstructs PMGs. Having accomplished this bioinformatic prerequisite, it would be now possible to systematically investigate the overall mitochondrial status in different *Oenothera* species as well as in various organs and developmental stages. Besides performing a PMG screening for several *Oenothera* species, SAGBAC would also be capable to shed light into some putative mitochondria-associated inheritance patterns in crosses between phylogenetically distant *Oenothera* species which were previously described (Schwemmle, 1938; Barthelmess, 1965; Stubbe, 1989b, a). Here, it would be most important to sequence affected and unaffected tissue parts separately to detect any differences in downstream SAGBAC-based analyses. Furthermore, stoichiometric analyses would reveal differences in the abundance of CRC combinations for the RRPs sets of the crossing partners. As our SAGBAC pipeline identified RRPs which are partially or completely part of coding genes, newly created CRC combinations can lead to the creation of novel, biologically relevant/active open reading frames (cryptic CMS loci) resulting in incompatibilities which may potentially play an important role in the evolution of *Oenothera* species and may act as speciation barrier. Additionally, the identification of different mitochondrion genotypes would extend the nuclear-plastome-compatibility chart in *Oenothera* (Stubbe, 1989c; Greiner et al., 2008) by integrating mitochondrion genotypes as a third dimension and transform it into a three-dimensional nuclear-plastome-mitochondrion compatibility “cube”.

Our current recommendation for *de novo* assembly of PMGs is still to perform standard Illumina paired-end libraries but an alternative for the discontinued Roche 454 technology would be needed. Illumina MiSeq devices are able to sequence libraries with 2x 300 bp paired-end (https://www.illumina.com/systems/sequencing-platforms/miseq.html). An insert size distribution between 450-550 bp would be optimal to merge mates of read pairs and to reduce the overlap of mates’ sequences (loss of effective read depth). Illumina mate pair libraries for stoichiometric analyses will be hopefully available in the future as other technologies such as PacBio HiFi reads are capable to replace them as they are even longer than Sanger sequencing reads with the same or even better base accuracy (Hon et al., 2020). Also, nowadays, with the appearance of Oxford Nanopore, one of the key players in 3GS technologies, it would also be possible to sequence entire mtDNA molecules (PMG isoforms) and thus abundance of PMG isoforms would be directly detectable. But to sequence entire PMG isoforms they need to be linearized as PMGs are hypothetically circular. But to retain entire DNA molecules, protocols are necessary which cut circular molecules only once. This would be achievable by restriction enzymes that cut only once preferably on RRP sequences or if they do not exist by designing CRISPR-Cas9 probes targeting only single RRPs.

3.7.2 High throughput plastome assembling and annotation

Taking our investigation in publication 2 to the next level by broadening the phylogenetic scale for multiple sequence alignment would need an increased number of plastomes to be assembled and annotated. An implementation of a (completely) automated harmonization pipeline for plastome assemblies would be an important step as manual curation is very time-consuming. Such a new approach also needs to take care of handling fluent IR boundaries (Zhu et al., 2016) as well as reorganizing assembled plastome contigs regarding order and orientation of genomic regions forming the canonical quadripartite structure (Gordon et al., 1981, 1982). Additionally, an incorporation of sequences generated by Sanger sequencing should also be considered if standard Illumina paired-end libraries generated on HiSeq or NovaSeq devices are still used. By sequencing of Illumina libraries on MiSeq devices, paired-end data can be extended to 2x 300 bp, which can help to overcome the issues of the notorious repetitive regions present in *Oenothera* plastomes.

3.7.3 Expanding the knowledge on the *SULFUREA* paramutation phenomenon

As previously introduced, paramutation phenomena result from a complex interaction of several cellular processes including transcription, 24bp sRNA biogenesis and DNA methylation machineries. Besides the identification of the chromosomal locus, which becomes methylated in trans, little is known about genes from the different interacting processes in tomato. By the isolation of a dominant suppressor (SOSU1) and a series of revertants (ROSU2-6) from tissue culture (see publication 3), detailed investigations into the molecular mechanism of paramutation are now possible.

Genetic interaction in trans depends often on repeat number and therefore it is most important to clearly resolve the copy number of the long terminal repeats of gypsy and copia type which are present at the *SULF* locus (Giacopelli and Hollick, 2015). As *SULFUREA* was generated by X-ray experiments, these repetitive elements become may be activated (Dubin et al., 2018). That such a resolution is important can be easily illustrated at the b1 locus which is paramutated in maize (Stam et al., 2002a; Belele et al., 2013). Seven copies of a repeat are present at this locus but if compared to reference genomes deposited in databases like NCBI (https://www.ncbi.nlm.nih.gov/genome), ensemblPlants (http://plants.ensembl.org) or phytosome (https://phytozome-next.jgi.doe.gov) only one copy is present highlighting the importance of structural variant detection with 3GS technologies.

By performing small RNA sequencing for SOSU1 but also for the several ROSUs that have appeared, we would extend and maybe could fundamentally confirm the results from Gouil et al which investigated 24bp sRNA biogenesis but only in wild type and *sulf(het-y)* plant material (Gouil et al., 2016). Epigenetic mechanisms besides DNA-Methylation, namely chromatin modifications, should also be considered in SOSU1 as the identified fusion gene consists of domains of a histone deacethylase. This can be achieved by Chip-seq experiments.

Furthermore, Hi-C-seq experiments would reveal changes in global chromatin interactions as previously shown in *Arabidopsis* mutants (Feng et al., 2014).

There is no doubt that 3GS technologies will gain more importance in the future to resolve repetitive structures and to investigate them properly in their biological relevance especially in NMI phenomena. Ultimately, another joined effort, named “Telomer2Telomer” aims to assemble entire chromosomes including the centromeric regions which are the most repetitive regions known so far. This kind of effort was now realized for the very first time for the human genome (Nurk et al., 2021) and will surely become standard within the next decade – exiting times ahead.

4 References

Al-Nakeeb, K., Petersen, T.N., and Sicheritz-Pontén, T. (2017). Norgal: extraction and *de novo* assembly of mitochondrial DNA from whole-genome sequencing data. BMC Bioinformatics **18,** 510.

Alleman, M., Sidorenko, L., McGinnis, K., Seshadri, V., Dorweiler, J.E., White, J., Sikkink, K., and Chandler, V.L. (2006). An RNA-dependent RNA polymerase is required for paramutation in maize. Nature **442,** 295-298.

Arrieta-Montiel, M.P., Shedge, V., Davila, J., Christensen, A.C., and Mackenzie, S.A. (2009). Diversity of the Arabidopsis mitochondrial genome occurs via nuclear-controlled recombination activity. Genetics **183,** 1261-1268.

Barbour, J.E., Liao, I.T., Stonaker, J.L., Lim, J.P., Lee, C.C., Parkinson, S.E., Kermicle, J., Simon, S.A., Meyers, B.C., Williams-Carrier, R., Barkan, A., and Hollick, J.B. (2012). Required to maintain repression2 is a novel protein that facilitates locus-specific paramutation in maize. Plant Cell **24,** 1761-1775.

Barnard-Kubow, K.B., Sloan, D.B., and Galloway, L.F. (2014). Correlation between sequence divergence and polymorphism reveals similar evolutionary mechanisms acting across multiple timescales in a rapidly evolving plastid genome. BMC Evol Biol **14,** 268.

Barnard-Kubow, K.B., McCoy, M.A., and Galloway, L.F. (2017). Biparental chloroplast inheritance leads to rescue from cytonuclear incompatibility. New Phytol **213,** 1466-1476.

Barneche, F., Winter, V., Crèvecoeur, M., and Rochaix, J.D. (2006). ATAB2 is a novel factor in the signalling pathway of light-controlled synthesis of photosystem proteins. EMBO J **25,** 5907-5918.

Barthelmess, A. (1965). Grundlagen der Vererbung. (Frankfurt am Main: Akademische Verlagsgesellschaft Athenaion).

Bateson, W., and Pellew, C. (1920). The genetics of “rogues” among culinary peas (*Pisum sativum*). P R Soc Lond B-conta **91,** 186-195.

Baur, E. (1909). Das Wesen und die Erblichkeitsverhältnisse der “Varietates albomarginatae hort." von *Pelargonium zonale*. Mol Gen Genet.

Belele, C.L., Sidorenko, L., Stam, M., Bader, R., Arteaga-Vazquez, M.A., and Chandler, V.L. (2013). Specific tandem repeats are sufficient for paramutation-induced trans-generational silencing. PLoS Genet **9,** e1003773.

Binder, S., Schuster, W., Grienenberger, J.M., Weil, J.H., and Brennicke, A. (1990). Genes for tRNA(Gly), tRNA(His), tRNA(Lys), tRNA(Phe), tRNA(Ser) and tRNA(Tyr) are encoded in *Oenothera* mitochondrial DNA. Curr Genet **17,** 353-358.

Bradnam, K.R., Fass, J.N., Alexandrov, A., Baranay, P., Bechner, M., Birol, I., Boisvert, S., Chapman, J.A., Chapuis, G., Chikhi, R., Chitsaz, H., Chou, W.C., Corbeil, J., Del Fabbro, C., Docking, T.R., Durbin, R., Earl, D., Emrich, S., Fedotov, P., Fonseca, N.A., et al. (2013). Assemblathon 2: evaluating *de novo* methods of genome assembly in three vertebrate species. Gigascience **2,** 10.

Brandes, U., Dwyer, T., and Schreiber, F. (2004). Visual understanding of metabolic pathways across organisms using layout in two and a half dimensions. J Integr Bioinform **1,** 11-26.

Brennicke, A. (1980). Mitochondrial DNA from *Oenothera berteriana*: Purification and properties. Plant Physiol **65,** 1207-1210.

Brennicke, A., and Schwemmle, B. (1984). Inheritance of mitochondrial DNA in *Oenothera berteriana* and *Oenothera odorata* hybrids. Zeitschrift für Naturforschung **39c,** 191-192.

Brennicke, A., Möller, S., and Blanz, P.A. (1985). The 18S and 5S ribosomal RNA genes in *Oenothera* mitochondria: Sequence rearrangments in the 18S and 5S rRNA genes of higher plants. Mol Gen Genet **198,** 404-410.

Brink, R.A. (1973). Paramutation. Annu Rev Genet **7,** 129-152.

Chandler, V.L. (2007). Paramutation: from maize to mice. Cell **128,** 641-645.

Chandler, V.L., and Stam, M. (2004). Chromatin conversations: mechanisms and implications of paramutation. Nat Rev Genet **5,** 532-544.

Chin, C.S., Peluso, P., Sedlazeck, F.J., Nattestad, M., Concepcion, G.T., Clum, A., Dunn, C., O'Malley, R., Figueroa-Balderas, R., Morales-Cruz, A., Cramer, G.R., Delledonne, M.,

Luo, C., Ecker, J.R., Cantu, D., Rank, D.R., and Schatz, M.C. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. Nat Methods **13,** 1050-1054.

Chiu, W.-L., Stubbe, W., and Sears, B.B. (1988). Plastid inheritance in *Oenothera*: organelle genome modifies the extent of biparental plastid transmission. Curr Genet **13,** 181-189.

Cleland, R.E. (1972). *Oenothera*. Cytogenetics and evolution**,** 370.

Correns, C. (1909). Vererbungsversuche mit blass(gelb)grünen und buntblättrigen Sippen bei *Mirabilis jalapa*, *Urtica pilulifera* and *Lunaria annua*. Mol Gen Genet**,** 1,291-329.

de Vries, J., Sousa, F.L., Bölter, B., Soll, J., and Gould, S.B. (2015). YCF1: A Green TIC? Plant Cell **27,** 1827-1833.

Deamer, D., Akeson, M., and Branton, D. (2016). Three decades of nanopore sequencing. Nat Biotechnol **34,** 518-524.

Dierckxsens, N., Mardulyn, P., and Smits, G. (2017). NOVOPlasty: *de novo* assembly of organelle genomes from whole genome data. Nucleic Acids Res **45,** e18.

Dorweiler, J.E., Carey, C.C., Kubo, K.M., Hollick, J.B., Kermicle, J.L., and Chandler, V.L. (2000). Mediator of paramutation1 is required for establishment and maintenance of paramutation at multiple maize loci. Plant Cell **12,** 2101-2118.

Dotzek, J. (2016). Mitochondria in the genus *Oenothera* - Non-Mendelian inheritance patterns, in vitro structure and evolutionary dynamics. In Faculty of Mathematics and Natural Sciences of the University of Potsdam (Potsdam: University of Potsdam), pp. 143.

Drouin, G., Daoud, H., and Xia, J. (2008). Relative rates of synonymous substitutions in the mitochondrial, chloroplast and nuclear genomes of seed plants. Mol Phylogenet Evol **49,** 827-831.

Dubin, M.J., Mittelsten Scheid, O., and Becker, C. (2018). Transposons: a blessing curse. Curr Opin Plant Biol **42,** 23-29.

Eades, P., and Feng, Q.-W. (1996). Multilevel visualization of clustered graphs. In International symposium on graph drawing (Springer), pp. 101-112.

Earl, D., Bradnam, K., St John, J., Darling, A., Lin, D., Fass, J., Yu, H.O., Buffalo, V., Zerbino, D.R., Diekhans, M., Nguyen, N., Ariyaratne, P.N., Sung, W.K., Ning, Z., Haimel, M., Simpson, J.T., Fonseca, N.A., Birol, I., Docking, T.R., Ho, I.Y., et al. (2011). Assemblathon 1: a competitive assessment of *de novo* short read assembly methods. Genome Res **21,** 2224-2241.

Ehlert, B., Schöttler, M.A., Tischendorf, G., Ludwig-Müller, J., and Bock, R. (2008). The paramutated *SULFUREA* locus of tomato is involved in auxin biosynthesis. J Exp Bot **59,** 3635-3647.

Ellis, J. (1982). Promiscuous DNA - chloroplast genes inside plant mitochondria. Nature **299,** 678-679.

Erhard, K.F., Stonaker, J.L., Parkinson, S.E., Lim, J.P., Hale, C.J., and Hollick, J.B. (2009). RNA polymerase IV functions in paramutation in *Zea mays*. Science **323,** 1201-1205.

Feng, S., Cokus, S.J., Schubert, V., Zhai, J., Pellegrini, M., and Jacobsen, S.E. (2014). Genome-wide Hi-C analyses in wild-type and mutants reveal high-resolution chromatin interactions in *Arabidopsis*. Mol Cell **55,** 694-707.

Frommer, M., McDonald, L.E., Millar, D.S., Collis, C.M., Watt, F., Grigg, G.W., Molloy, P.L., and Paul, C.L. (1992). A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. Proc Natl Acad Sci USA **89,** 1827-1831.

Giacopelli, B.J., and Hollick, J.B. (2015). Trans-homolog interactions facilitating paramutation in maize. Plant Physiol **168,** 1226-1236.

Gillham, N.W. (1978). Organelle heredity. (Raven Press).

Golczyk, H., Greiner, S., Wanner, G., Weihe, A., Bock, R., Börner, T., and Herrmann, R.G. (2014). Chloroplast DNA in mature and senescing leaves: a reappraisal. Plant Cell **26,** 847-854.

Gordon, K.H., Crouse, E.J., Bohnert, H.J., and Herrmann, R.G. (1981). Restriction endonuclease cleavage site map of chloroplast DNA from *Oenothera parviflora* (Euoenothera plastome IV). Theor Appl Genet **59,** 281-296.

Gordon, K.H., Crouse, E.J., Bohnert, H.J., and Herrmann, R.G. (1982). Physical mapping of differences in chloroplast DNA of the five wild-type plastomes in *Oenothera* subsection Euoenothera. Theor Appl Genet **61,** 373-384.

Gouil, Q., Novák, O., and Baulcombe, D.C. (2016). SLTAB2 is the paramutated *SULFUREA* locus in tomato. J Exp Bot **67,** 2655-2664.

Greiner, S., Sobanski, J., and Bock, R. (2015). Why are most organelle genomes transmitted maternally? Bioessays **37,** 80-94.

Greiner, S., Rauwolf, U., Meurer, J., and Herrmann, R.G. (2011). The role of plastids in plant speciation. Mol ecol **20,** 671-691.

Greiner, S., Golczyk, H., Malinova, I., Pellizzer, T., Bock, R., Börner, T., and Herrmann, R.G. (2020). Chloroplast nucleoids are highly dynamic in ploidy, number, and structure during angiosperm leaf development. Plant J **102,** 730-746.

Greiner, S., Wang, X., Rauwolf, U., Silber, M.V., Mayer, K., Meurer, J., Haberer, G., and Herrmann, R.G. (2008). The complete nucleotide sequences of the five genetically distinct plastid genomes of *Oenothera*, subsection *Oenothera*: I. Sequence evaluation and plastome evolution. Nucleic acids research **36,** 2366-2378.

Grun, P. (1976). Cytoplasmic genetics and evolution. (Columbia University Press).

Guo, W., Zhu, A., Fan, W., and Mower, J.P. (2017). Complete mitochondrial genomes from the ferns *Ophioglossum californicum* and *Psilotum nudum* are highly repetitive with the largest organellar introns. New Phytol **213,** 391-403.

Guo, W., Grewe, F., Fan, W., Young, G.J., Knoop, V., Palmer, J.D., and Mower, J.P. (2016). Ginkgo and Welwitschia mitogenomes reveal extreme contrasts in gymnosperm mitochondrial evolution. Mol Biol Evol **33,** 1448-1460.

Hagemann, R. (1958). Somatic conversion in *Lycopersicon esculentum* Mill. Mol Gen Genet **89,** 587-613.

Hahn, C., Bachmann, L., and Chevreux, B. (2013). Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads - a baiting and iterative mapping approach. Nucleic Acids Res **41,** e129.

Hale, C.J., Stonaker, J.L., Gross, S.M., and Hollick, J.B. (2007). A novel Snf2 protein maintains trans-generational regulatory states established by paramutation in maize. PLoS Biol **5,** e275.

Handelsman, J., Rondon, M.R., Brady, S.F., Clardy, J., and Goodman, R.M. (1998). Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem Biol **5,** R245-249.

Harte, C. (1994). *Oenothera* - Contributions of a Plant to Biology. (Berlin, Heidelberg, New York: Springer).

Hiesel, R., and Brennicke, A. (1985). Overlapping reading frames in *Oenothera* mitochondria. FEBS letters **193,** 164-168.

Hollick, J.B., Dorweiler, J.E., and Chandler, V.L. (1997). Paramutation and related allelic interactions. Trends Genet **13,** 302-308.

Hollick, J.B., Kermicle, J.L., and Parkinson, S.E. (2005). Rmr6 maintains meiotic inheritance of paramutant states in *Zea mays*. Genetics **171,** 725-740.

Hon, T., Mars, K., Young, G., Tsai, Y.C., Karalius, J.W., Landolin, J.M., Maurer, N., Kudrna, D., Hardigan, M.A., Steiner, C.C., Knapp, S.J., Ware, D., Shapiro, B., Peluso, P., and Rank, D.R. (2020). Highly accurate long-read HiFi sequencing data for five complex genomes. Sci Data **7,** 399.

Kirk, J.T.O., and Tilney-Basset, R.A.E. (1978). The Plastids: Their Chemistry, Structure, Growth, and Inheritance. (Amsterdam-New York-Oxford: Elsevier/North Holland Biomedical Press).

Kleine, T., Maier, U.G., and Leister, D. (2009). DNA transfer from organelles to the nucleus: the idiosyncratic genetics of endosymbiosis. Annu Rev Plant Biol **60,** 115-138.

Knoop, V. (2004). The mitochondrial DNA of land plants: peculiarities in phylogenetic perspective. Curr Genet **46,** 123-139.

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., and Phillippy, A.M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res **27,** 722-736.

Kozik, A., Rowan, B.A., Lavelle, D., Berke, L., Schranz, M.E., Michelmore, R.W., and Christensen, A.C. (2019). The alternative reality of plant mitochondrial DNA: One ring does not rule them all. PLoS Genet **15,** e1008373.

Krueger, F., and Andrews, S.R. (2011). Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics **27,** 1571-1572.

Leister, D. (2005). Origin, evolution and genetic effects of nuclear insertions of organelle DNA. Trends Genet **21,** 655-663.

Lisch, D., Carey, C.C., Dorweiler, J.E., and Chandler, V.L. (2002). A mutation that prevents paramutation in maize also reverses Mutator transposon methylation and silencing. Proc Natl Acad Sci USA **99,** 6130-6135.

Lister, R., and Ecker, J.R. (2009). Finding the fifth base: genome-wide sequencing of cytosine methylation. Genome Res **19,** 959-966.

Lonsdale, D.M. (1984). A review of the structure and organization of the mitochondrial genome of higher plants. Plant Mol Biol **3,** 201-206.

Lye, Z.N., and Purugganan, M.D. (2019). Copy number variation in domestication. Trends Plant Sci **24,** 352-365.

Massouh, A., Schubert, J., Yaneva-Roder, L., Ulbricht-Jones, E.S., Zupok, A., Johnson, M.T.J., Wright, S.I., Pellizzer, T., Sobanski, J., Bock, R., and Greiner, S. (2016). Spontaneous chloroplast mutants mostly occur by replication slippage and show a biased pattern in the plastome of *Oenothera*. Plant Cell **28,** 911-929.

Michalovova, M., Vyskot, B., and Kejnovsky, E. (2013). Analysis of plastid and mitochondrial DNA insertions in the nucleus (NUPTs and NUMTs) of six plant species: size, relative age and chromosomal localization. Heredity (Edinburg) **111,** 314-320.

Nishimura, Y., and Stern, D.B. (2010). Differential replication of two chloroplast genome forms in heteroplasmic *Chlamydomonas reinhardtii* gametes contributes to alternative inheritance patterns. Genetics **185,** 1167-1181.

Nishino, J., Ochi, H., Kochi, Y., Tsunoda, T., and Matsui, S. (2018). Sample size for successful genome-wide association study of major depressive disorder. Front Genet **9,** 227.

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A.V., Mikheenko, A., Vollger, M.R., Altemose, N., Uralsky, L., Gershman, A., Aganezov, S., Hoyt, S.J., Diekhans, M., Logsdon, G.A., Alonge, M., Antonarakis, S.E., Borchers, M., Bouffard, G.G., Brooks, S.Y., Caldas, G.V., et al. (2021). The complete sequence of a human genome. bioRxiv.

Peng, Y., Leung, H.C., Yiu, S.M., and Chin, F.Y. (2012). IDBA-UD: a *de novo* assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics **28,** 1420-1428.

Pirooznia, M., Goes, F.S., and Zandi, P.P. (2015). Whole-genome CNV analysis: advances in computational approaches. Front Genet **6,** 138.

Rassoulzadegan, M., Grandjean, V., Gounon, P., Vincent, S., Gillot, I., and Cuzin, F. (2006). RNA-mediated non-mendelian inheritance of an epigenetic change in the mouse. Nature **441,** 469-474.

Renner, O. (1938). Über *Oenothera atrovirens* Sh. et Bartl. und über somatische Konversion im Erbgang des *cruciata*-Merkmals der *Oenotheren*. Mol Gen Genet **74,** 91-124.

Rhoads, A., and Au, K.F. (2015). PacBio sequencing and its applications. GPB **13,** 278-289.

Schuster, W., and Brennicke, A. (1987). Plastid, nuclear and reverse transcriptase sequences in the mitochondrial genome of *Oenothera*: is genetic information transferred between organelles via RNA? EMBO J **6,** 2857-2863.

Schuster, W., and Brennicke, A. (1988). Interorganellar sequence transfer: plant mitochondrial DNA is nuclear, is plastid, is mitochondrial. Plant Sci **54,** 1-10.

Schwemmle, J. (1938). Genetische und zytologische Untersuchungen an Eu-Oenotheren. Mol Gen Genet **75,** 486-660.

Sidorenko, L.V., and Peterson, T. (2001). Transgene-induced silencing identifies sequences involved in the establishment of paramutation of the maize p1 gene. Plant Cell **13,** 319-335.

Sloan, D.B., Triant, D.A., Forrester, N.J., Bergner, L.M., Wu, M., and Taylor, D.R. (2014). A recurring syndrome of accelerated plastid genome evolution in the angiosperm tribe Sileneae (Caryophyllaceae). Mol Phylogenet Evol **72,** 82-89.

Stam, M., Belele, C., Dorweiler, J.E., and Chandler, V.L. (2002a). Differential chromatin structure within a tandem array 100 kb upstream of the maize b1 locus is associated with paramutation. Genes Dev **16,** 1906-1918.

Stam, M., Belele, C., Ramakrishna, W., Dorweiler, J.E., Bennetzen, J.L., and Chandler, V.L. (2002b). The regulatory regions required for B' paramutation and expression are located far upstream of the maize b1 transcribed sequences. Genetics **162,** 917-930.

Straub, S.C., Fishbein, M., Livshultz, T., Foster, Z., Parks, M., Weitemier, K., Cronn, R.C., and Liston, A. (2011). Building a model: developing genomic resources for common milkweed (*Asclepias syriaca*) with low coverage genome sequencing. BMC Genomics **12,** 211.

Stubbe, W. (1964). The role of the plastome in evolution of the genus *Oenothera*. Genetica **35,** 28-33.

Stubbe, W. (1989a). The falcifolia syndrome of *Oenothera*: IV. Loss of falcifolia-determining factors. Mol Gen Genet **218,** 511-515.

Stubbe, W. (1989b). The falcifolia syndrome of *Oenothera*: III. The general pattern of its non-Mendelian inheritance. Mol Gen Genet **218,** 499-510.

Stubbe, W. (1989c). *Oenothera* - An ideal system for studying the interactions of genome and plastome. Plant Mol Biol Rep **7,** 245-257.

Tsuboi, K., Nagatomo, T., Gohno, T., Higuchi, T., Sasaki, S., Fujiki, N., Kurosumi, M., Takei, H., Yamaguchi, Y., Niwa, T., and Hayashi, S.I. (2017). Single CpG site methylation controls estrogen receptor gene transcription and correlates with hormone therapy resistance. J Steroid Biochem Mol Biol **171,** 209-217.

Ulbricht-Jones, E.S., Dotzek, J., and Greiner, S. (2021). Maternal inheritance of mitochondria and biparental inheritance of chloroplasts allows separation of cytoplasmic effects in the evening primrose (*Oenothera*). In Preparation.

Varré, J.S., D'Agostino, N., Touzet, P., Gallina, S., Tamburino, R., Cantarella, C., Ubrig, E., Cardi, T., Drouard, L., Gualberto, J.M., and Scotti, N. (2019). Complete sequence, multichromosomal architecture and transcriptome analysis of the *Solanum tuberosum* mitochondrial genome. Int J Mol Sci **20**.

Wang, D., Wu, Y.W., Shih, A.C., Wu, C.S., Wang, Y.N., and Chaw, S.M. (2007). Transfer of chloroplast genomic DNA to mitochondrial genome occurred at least 300 MYA. Mol Biol Evol **24,** 2040-2048.

Wicke, S., Schneeweiss, G.M., dePamphilis, C.W., Müller, K.F., and Quandt, D. (2011). The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. Plant Mol Biol **76,** 273-297.

Wissinger, B., Schuster, W., and Brennicke, A. (1991). Trans splicing in *Oenothera* mitochondria: *nad1* mRNAs are edited in exon and trans-splicing group II intron sequences. Cell **65,** 473-482.

Wolfe, K.H., Li, W.H., and Sharp, P.M. (1987). Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc Natl Acad Sci USA **84,** 9054-9058.

Woodhouse, M.R., Freeling, M., and Lisch, D. (2006). Initiation, establishment, and maintenance of heritable MuDR transposon silencing in maize are mediated by distinct factors. PLoS Biol **4,** e339.

Xi, Y., and Li, W. (2009). BSMAP: whole genome bisulfite sequence MAPping program. BMC Bioinformatics **10,** 232.

Zhang, H., Lang, Z., and Zhu, J.K. (2018). Dynamics and function of DNA methylation in plants. Nat Rev Mol Cell Biol **19,** 489-506.

Zhang, T., Zhang, X., Hu, S., and Yu, J. (2011). An efficient procedure for plant organellar genome assembly, based on whole genome data from the 454 GS FLX sequencing platform. Plant Methods **7,** 38.

Zhu, A., Guo, W., Gupta, S., Fan, W., and Mower, J.P. (2016). Evolutionary dynamics of the plastid inverted repeat: the effects of expansion, contraction, and loss on substitution rates. New Phytol **209,** 1747-1756.

5 Eidesstattliche Erklärung

Hiermit versichere ich an Eides statt, dass meine hinsichtlich der früheren Teilnahme an Promotionsverfahren gemachten Angaben richtig sind und, dass die eingereichte Arbeit oder wesentliche Teile derselben in keinem anderen Verfahren zur Erlangung eines akademischen Grades vorgelegt worden sind.

Ich versichere darüber hinaus, dass bei der Anfertigung der Dissertation die Grundsätze zur Sicherung guter wissenschaftlicher Praxis der DFG eingehalten wurden, die Dissertation selbständig und ohne fremde Hilfe, insbesondere für die allgemeine Einleitung und Diskussion, verfasst wurde, andere als die von mir angegebenen Quellen und Hilfsmittel nicht benutzt worden sind und die den benutzten Werken wörtlich oder sinngemäß entnommenen Stellen als solche kenntlich gemacht wurden.

Einer Überprüfung der eingereichten Dissertation bzw. der eingereichten Schriften mittels einer Plagiatsprüfungssoftware stimme ich zu.

Potsdam, den ____________________ Unterschrift ________________________

6 Danksagung

Zuallererst will ich mich bei Herrn apl. Prof. Dr. Dirk Walther bedanken, der es mir ermöglicht hat, während meiner Service Tätigkeit in seiner Arbeitsgruppe unter seiner Betreuung diese Doktorarbeit anzufertigen.

Bei Herrn Prof. Ralph Bock und Herrn Dr. Stephan Greiner will ich mich für die Möglichkeit bedanken, an ihren Projekten mitzuwirken, welche, wie alle anderen Projekte, zwar in meiner Servicetätigkeit begannen, jedoch nach und nach zu etwas Größerem herangewachsen sind, was schlussendlich in diese kumulative Doktorarbeit mündete.

Weiterhin will ich Dr. Stephan Greiner danken, der es mir nicht nur ermöglicht hat, eine weitere Zeit lang am MPI zu arbeiten, um diese Arbeit vollenden zu können, sondern mich auch über die vielen Jahre in die faszinierende Welt der Nachtkerzengewächse eingeführt hat.

Herrn Prof. Dr. Volker Knoop und Herrn Prof. Dr. Schoof will ich danken, dass sie sich bereit erklärt haben, als Gutachter meiner Doktorarbeit zu fungieren.

Jana Dotzek will ich für die etlichen Unterhaltungen und Diskussionen rund um die Nachtkerzen Mitochondrien danken. Unsere Zusammenarbeit war eine echte Symbiose. Gleichermaßen danke ich Britta Ehlert für die mehr als konstruktive Zusammenarbeit im Zuge des Paramutationsprojektes.

Ein herzlicher Dank geht an alle aktuellen und ehemaligen Mitarbeiter der beiden Arbeitsgruppen AG Walther und AG Greiner. Sie haben das Arbeiten in den letzten Jahren lebenswert gemacht und mich durch etliche Diskussionen inspiriert. Im speziellen will ich Stephanie Schaarschmidt nennen, die vor, aber gerade auch während der Corona-Zeit immer ein offenes Ohr für mich hatte.

An dieser Stelle will ich auch Frau Prof. Dr. Dr. Michal-Ruth Schweiger, ihren Mitarbeitern und hier speziell Dr. Martin Kerick danken. Ohne sie hätte ich nie zu der Welt des Next Generation Sequencings gefunden.

Meiner Familie und meinen Freunden bin ich unendlich dankbar, immer für mich da gewesen zu sein.

Meinem Ehemann Marco Fischer-Hoffmann will ich dafür danken, dass er stets an meiner Seite war. Er glaubte stets an mich und ermutigte mich, meine Träume zu leben.