UNIVERSITA’ DEGLI STUDI DI NAPOLI FEDERICO II

DOTTORATO DI RICERCA IN BIOLOGIA XXXI CICLO

APPROCCI BIOINFORMATICI PER L’ANALISI DI DATI DI NEXT GENERATION SEQUENCING IN *OPHRYS* (ORCHIDACEAE)

BIOINFORMATIC TOOLS FOR THE ANALYSIS OF NEXT GENERATION SEQUENCING DATA IN *OPHRYS* (ORCHIDACEAE)

Tutor e Coordinatore del corso **Candidato**

Ch.^{mo} Prof. Salvatore Cozzolino **Dott. Luca Roma**

ANNO ACCADEMICO 2017 – 2018

Sommario

Introduzione

L'ultimo decennio è stato caratterizzato da uno straordinario sviluppo delle tecnologie per il sequenziamento del DNA, che ha consentito la produzione di una mole di informazioni senza precedenti, rivoluzionando così molti campi della ricerca scientifica. Dal 2005 anno in cui l'azienda "Roche" ha messo in commercio il primo sequenziatore di seconda generazione (454 - Roche), sono state sviluppate tecnologie basate su principi molto diversi tra loro, che hanno dato la possibilità di sequenziare interi genomi con tempi e costi ridotti rispetto al sequenziamento tradizionale di prima generazione. Lo sviluppo di queste nuove tecnologie ha reso possibile la realizzazione di imponenti progetti scientifici come "1000 Genomes project", la cui realizzazione sarebbe stata impossibile con i sequenziatori di prima generazione.

Questo recente progresso delle tecniche di sequenziamento, ha richiesto un parallelo sviluppo di metodiche computazionali e bioinformatiche capaci di analizzare la grossa mole di dati prodotta, favorendo così la nascita delle cosiddette scienze "*omiche*" come la *genomica*, la *trascrittomica*, la *proteomica*, la *metagenomica*, la *metabolomica* e la *filogenomica*.

Il ruolo della bioinformatica è stato fondamentale in questo sviluppo. Infatti, grazie ad essa, sono stati sviluppati molteplici algoritmi basati su diversi linguaggi di "programmazione orientata ad oggetti" come AWK, BASH, JAVA, PERL, PYTHON che hanno consentito la manipolazione e l'ideazione di costrutti come tabelle, liste, indici ed espressioni regolari.

Grazie allo sviluppo della *genomica* e della *filogenomica* è oggi possibile utilizzare da centinaia a migliaia di geni nucleari e interi genomi organellari al fine di ricostruire relazioni evolutive e di distinguere specie per le quali problematiche evolutive o ecologiche sono ancora aperte. Avere il vantaggio di sequenziare milioni di sequenze di regioni sconosciute del genoma in tempi molto brevi, sta rivoluzionando il modo di fare ricerca e l'imminente riduzione dei costi renderà accessibile questo tipo di esperimenti anche ai piccoli laboratori, consentendo così un incremento esponenziale delle informazioni nelle banche dati pubbliche. Infatti, uno dei più grandi problemi, soprattutto per le piante, è la mancanza di informazioni fruibili dalle banche dati, che rende questo lavoro arduo e lungo. Lo studio delle specie non modello, come ad esempio quelle del genere *Ophrys*, sta richiedendo un parallelo arricchimento di queste banche dati mediante assemblaggio e annotazione dei genomi nucleari e organellari (mitocondriale e plastidiale). In futuro è ipotizzabile che ciascuna specie vegetale avrà il proprio genoma assemblato e annotato e questo consentirà non solo di ricostruire relazioni evolutive irrisolte, ma anche di monitorare e individuare facilmente specie ad alto rischio di estinzione. Inoltre, il sequenziamento del trascrittoma e dell'esoma potranno consentire la scoperta di geni importanti che regolano la resa e la tolleranza agli stress biotici e abiotici. Pertanto, l'obiettivo di questo dottorato è stato quello di ottimizzare tecniche bioinformatiche in un contesto evolutivo e di assemblare dei genomi per contribuire all'arricchimento delle banche dati.

Capitolo I

Sequenziamenti di seconda e terza generazione e bioinformatica

1.1 Sequenziamento di seconda generazione

Il *pirosequenziamento* è una delle nuove tecniche di sequenziamento ad elevato parallelismo, basata sul principio del *Sequencing By Synthesis* (Fuller et al., 2009). Questa tecnica richiede diverse fasi e si basa sull'utilizzo di una serie di enzimi che, in presenza di ATP, producono luce quando un nucleotide viene incorporato nel nuovo filamento prodotto. Questa tecnologia è alla base dei sequenziatori ILLUMINA-SOLEXA di cui quelli attualmente in produzione sono: MISEQ, NEXTSEQ, HISEQ, HISEQ X e il recentissimo NOVASEQ.

Un sequenziamento ILLUMINA consta di tre fasi principali: la costruzione della libreria rappresentativa (*Reduced Representation Library*), l'amplificazione mediante *bridge amplification* ed il sequenziamento (*Sequencing by Synthesis*).

Nella prima fase viene costruita una libreria rappresentativa ridotta (RRL) in cui il DNA viene prima frammentato attraverso sonicazione e poi a ciascuno dei frammenti ottenuti viene legato un adattatore alle estremità 5' e 3'. Questi frammenti si legheranno alla piastra (*flow cell*) mediante ibridazione tra adattatori e inneschi presenti sulla *flow cell*. L'adattatore in 3' presenta una sequenza identificativa del campione sequenziato chiamata *index*, che consente al bioinformatico di separare dal file di output i diversi campioni mediante il processo del *demultiplexing*.

La seconda fase, chiamata *bridge amplification*, consiste nell'amplificazione dei frammenti attraverso diverse fasi di PCR in cui le reads formano una tipica “struttura a ponte”. Questa fase è indispensabile, perché consente la formazione di clusters della stessa sequenza, consentendo durante la fase del sequenziamento l’aumento dell’intensità della luce emessa da ogni nucleotide, al fine di rendere il segnale abbastanza forte da poter essere catturato dalla camera CCD (charge - coupled device).

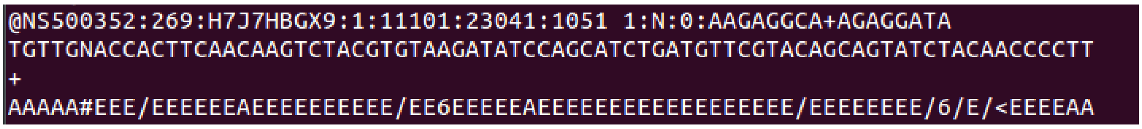
La terza fase è il *Sequencing By Synthesis*, grazie alla quale è possibile determinare la sequenza nucleotidica man mano che vengono aggiunti nuovi desossiribonucleotidi. Questi ultimi sono nucleotidi modificati che presentano dei terminatori di catena reversibili legati all’estremità 3’ OH. I terminatori di catena presentano dei gruppi fluorescenti in grado di emettere un fascio di luce diverso a seconda del tipo di nucleotide che viene rilevato. Questi terminatori di catena sono reversibili perché si staccano in seguito alla scansione con il laser che rileva la frequenza emessa e, allo stesso tempo, libera l’estremità 3’ OH rendendola disponibile per l’aggiunta di un altro nucleotide. Il file generato dal sequenziatore è in formato FASTQ. Un file FASTQ riporta informazioni sul campione, la sequenza nucleotidica e dei simboli ASCII che riportano per ogni nucleotide la qualità del sequenziamento.

Un file FASTQ differisce dal classico file in formato FASTA, grazie alla presenza di una linea supplementare, che riporta la qualità per ciascun nucleotide.

Un file FASTQ è composto da 4 linee per sequenza (Figura 1.1).

- Linea 1 o “Header” inizia sempre per il simbolo'@' ed è seguita da una sequenza identificatrice che riporta le coordinate della sequenza nucleotidica sulla *flowcell*.
- Linea 2 è la sequenza nucleotidica.

- Linea 3 inizia con il simbolo '+'.
- Linea 4 riporta i valori di qualità in codice ASCII (PHRED QUALITY SCORE) di ciascun corrispondente nucleotide in linea 2. PHRED QUALITY SCORE Q sono definiti come una proprietà logaritmicamente correlata alle probabilità di errore di chiamata base P. $Q = -10\log_{10}P$.



```
@NS500352:269:H7J7HBGX9:1:11101:23041:1051 1:N:0:AAGAGGCA+AGAGGATA
TGTTGNACCACTTCAACAAGTCTACGTGTAAGATATCCAGCATCTGATGTTCGTACAGCAGTATCTACAACCCCTT
+
AAAAA#EEE/EEEEEEAEEEEEEEEEE/EE6EEEEEAEEEEEEEEEEEEEEEEEE/EEEEEEEE/6/E/<EEEEAA
```

Figura 1.1. Schema di un file FASTQ. La prima linea è l'header o intestazione, la seconda la sequenza nucleotidica e la quarta riporta un valore di qualità corrispondente a ciascun nucleotide in codice ASCII

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

Tabella 1.1. PHRED QUALITY SCORE e BASE CALLING ACCURACY. Un valore di PHRED uguale a 30 spesso viene usato come cut-off per la qualità di ciascun nucleotide

La linea 4 è molto importante perché ci dà la possibilità di filtrare o tagliare le sequenze in base alla loro qualità attraverso l'utilizzo di programmi che usano approcci diversi, come ad esempio TRIMMOMATIC (Bolger et al., 2014) o BBDUK (https://jgi.doe.gov/data-and-tools/bbtools/). TRIMMOMATIC filtra le reads analizzandole per “finestra”, cioè tagliando quelle con una qualità media all'interno della finestra al di sotto di una soglia che viene specificata dall’utente. Il software BBDUK filtra le reads selezionandole per *k-mers*. Il valore soglia del PHRED QUALITY SCORE che spesso si usa è 30. Questo perché si ha un’accuratezza della reads del 99,9% cioè la probabilità di errore di 1 su 1000 (Tabella 1.1).

Per visualizzare la qualità delle reads è possibile usare il programma FASTQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc), il quale ci dà diverse informazioni statistiche sulla qualità del campione (Figura 1.2). Queste informazioni sono: il numero di reads, percentuale di GC, qualità media delle reads, presenza di *k-mers* e presenza di adattatori interni.

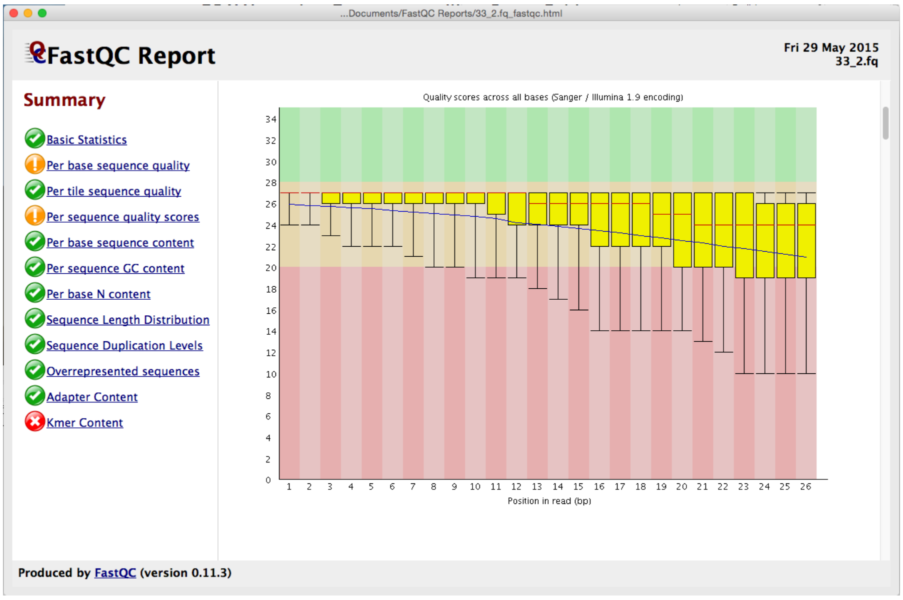

...Documents/FastQC Reports/33_2.fq_fastqc.html
FastQC Report
Fri 29 May 2015
33_2.fq
Summary
Basic Statistics
Per base sequence quality
Per tile sequence quality
Per sequence quality scores
Per base sequence content
Per sequence GC content
Per base N content
Sequence Length Distribution
Sequence Duplication Levels
Overrepresented sequences
Adapter Content
Kmer Content
Quality scores across all bases (Sanger / Illumina 1.9 encoding)
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
Position in read (bp)
Produced by FastQC (version 0.11.3)

Figura 1.2. Report del software FASTQC. Sulle ascisse è riportata la lunghezza delle reads e sulle ordinate il valore della qualità in PHRED QUALITY SCORE. I boxplot in giallo riportano la distribuzione della qualità delle reads per finestra.

I sequenzatori ILLUMINA hanno numerosi vantaggi, tra cui il principale è quello di poter sequenziare un intero genoma in tempi relativamente brevi, producendo miliardi di reads. Un altro vantaggio è quello di poter riprodurre le parti sconosciute del genoma, perché i primers presenti sulla *flow cell* sono complementari agli adattatori inseriti a monte e a valle dei frammenti di DNA generati. Nonostante ciò, i sequenziatori ILLUMINA hanno delle caratteristiche che possono essere ulteriormente migliorate. Infatti, le reads troppo corte spesso rendono difficile l'assemblaggio di regioni ripetute del genoma. Inoltre, una delle caratteristiche che spesso rende difficile l'analisi, è la difficoltà nella standardizzazione del numero di reads per campione.

1.2 Sequenziatori di terza generazione

Le tecniche di sequenziamento di terza generazione, definite anche come *next next generation sequencing* (NNGS), si basano su una tecnologia molto sensibile chiamata *Single Molecule Real-Time* (SMRT), che non prevede la fase di amplificazione dei frammenti di DNA (McCarthy et al., 2010). Uno dei sequenziatori più conosciuti è il PACBIO, prodotto dall'azienda *Pacific Biosciences* Menlo Park, California, USA. Il sequenziatore PACBIO, oltre ad avere la tecnologia SMRT, ha l'enorme vantaggio di produrre reads più lunghe di un normale sequenziamento ILLUMINA, consentendo l'assemblaggio *de novo* anche di genomi con sequenze molto ripetute.

Altri sequenziatori sono le due piattaforme NANOPORE (GridION e MinION) della Oxford Nanopore Technologies. Come suggerisce il nome, questa tecnologia è basata sulla lettura della differenza di potenziale a seguito del passaggio della sequenza nucleotidica, attraverso una membrana costituita da polimeri sintetici e dotata di nanopori proteici. Il DNA da sequenziare è a filamento singolo e la parte carica negativamente si sposta attraversando il nanoporo verso la carica positiva, bloccando così il canale e generando una differenza di potenziale ai lati della membrana.

Un altro sequenziatore è lo ION TORRENT, sviluppato dalla Ion Torrent System (Life Technologies). Il principio su cui si basa è molto innovativo e consiste nel misurare la variazione di pH determinata dallo ione idrogeno liberato durante la sintesi del DNA ad opera della DNA polimerasi. Ad ogni ciclo viene introdotto un diverso nucleotide che, se complementare a quello dello stampo, determina la liberazione di pirofosfato e ione idrogeno provocando una conseguente variazione di *p*H che viene registrata dall’apparecchio.

1.3 Analisi bioinformatica dei dati NGS

1.3.1 Assemblaggio di genomi mediante approccio *de novo*

Uno dei problemi principali delle piante è la mancanza di informazioni fruibili dalle banche dati genomiche, che rende spesso questo lavoro molto difficile. Infatti, quando si lavora con organismi non modello, è necessario prima assemblare e annotare la sequenza. Un assemblaggio *de novo* richiede una maggiore profondità nei dati di sequenziamento, rispetto a quando si ha a disposizione un genoma di riferimento. Bisogna tenere in considerazione, inoltre, che le reads molto corte possono far nascere dei problemi nell’assemblaggio di regioni altamente ripetute o a bassa complessità.

Esistono diversi programmi che consentono di assemblare *de novo* un genoma e sono stati scritti per rilevare sovrapposizioni tra le reads e assemblarle in *contigs*, e sucessivamente combinare questi ultimi in *scaffold* per ottenere una bozza del genoma. Tra i programmi per l'assemblaggio *de novo* si possono citare VELVET (Zerbino et al., 2008), SOAP2 (Li et al., 2009), SPADES (Bankevich et al., 2012) o ABYSS (Simpson et al., 2009). Questi programmi possono assemblare anche genomi molto grandi, ma in questo caso non possono essere usati su computer tradizionali perché richiedono molta memoria RAM e molti processori che lavorano in parallelo. Si possono effettuare diversi controlli al fine di comprendere la qualità del dato prodotto. I parametri da considerare sono N50, allineamento delle reads contro i contigs assemblati e identificazione dei geni altamente conservati mediante confronto con quelli già annotati in banca dati. N50 è un'unità di misura che descrive la qualità dei contigs assemblati e corrisponde alla somma delle lunghezze di tutti i contigs che formano almeno il 50% della sequenza del genoma totale.

1.3.2 Allineamento delle reads con un genoma di riferimento

Quando si ha già un genoma di riferimento, è possibile confrontare le reads dei campioni sequenziati allineandole contro di esso. Anche in questo caso si tratta di un processo molto dispendioso in termini di memoria RAM, in quanto il software deve confrontare ogni reads con il DNA di riferimento. I file SAM (SEQUENCE ALIGNMENT MAP) e BAM (BINARY ALIGNMENT MAP) sono gli standard di riferimento per il salvataggio dei dati ottenuti dall'allineamento per le tecnologie di nuova generazione. I file SAM vengono sempre convertiti nella versione binaria

(BAM) soprattutto per ragioni di spazio e velocità di esecuzione. SAM è un file di testo delimitato da tabulazioni e costituito da un'intestazione e una sezione di allineamento.

Se presente, l'intestazione deve essere precedente agli allineamenti. Le intestazioni si distinguono dagli allineamenti perchè iniziano con il simbolo '@'. Ogni linea di allineamento ha 11 campi obbligatori per informazioni essenziali, tra cui la posizione o della reads sul riferimento e la 'CIGAR STRING' (COMPACT IDIOSYNCRATIC GAPPED ALIGNMENT REPORT), che è una stringa che riassume brevemente come le reads si allineano con il riferimento. I file BAM sono la versione binaria dei SAM e quindi non hanno una un'interfaccia visibile dall'utente.

La maggior parte dei programmi che allineano le reads con una sequenza, usa un metodo basato sull'indicizzazione del riferimento, che rende più veloce la ricerca di posizioni di allineamento contro di esso. I programmi più utilizzati sono BWA (Li et al., 2009) e BOWTIE (Langmead et al., 2012) e sono basati sulla trasformata di *Burrows-Wheeler*, cioè un algoritmo di compressione reversibile che permuta l'ordine dei caratteri, senza cambiarne il valore. È da sottolineare che spesso è possibile riconvertire i files in formato SAM e BAM in FASTQ.

1.3.3 Variant calling

Dopo l'allineamento delle reads, il DNA in analisi può essere confrontato col genoma di riferimento al fine di individuare polimorfismi (SNPs, indels, SSRs). In particolare, le tecnologie NGS permettono la scoperta di diverse migliaia di polimorfismi per singoli nucleotidi (*Single Nucleotide Polymorphisms*, SNPs), cioè differenze tra due individui di un solo nucleotide che hanno una frequenza maggiore o uguale all'1%. La difficoltà in questo caso è quella di saper distinguere polimorfismi da errori di sequenziamento.

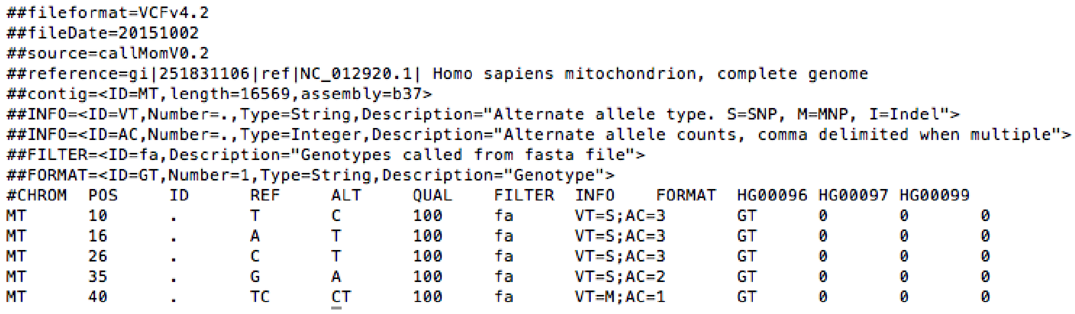
Spesso, i programmi bioinformatici interpretano male inserzioni e delezioni di DNA a causa di cattivi allineamenti. Proprio per risolvere questo problema, anche l'azienda GOOGLE ha sviluppato un programma che si chiama DEEPVARIANT (Poplin et al., 2017) che, oltre ad individuare inserzioni e delezioni, ha un sistema di correzione mediato da un'intelligenza artificiale che riduce così la presenza di falsi positivi. Altra causa del fenomeno sono gli errori dovuti ad una bassa copertura di dati che rendono indistinguibili SNPs da eventuali errori della DNA-polimerasi. Inoltre, è da sottolineare che, una bassa copertura dei dati rende difficilmente distinguibili i polimorfismi eterozigoti.

I programmi capaci di richiamare varianti genetiche, richiedono l'input file in formato BAM e i più comuni sono: GENOME ANALYSIS TOOLKIT (McKenna et al., 2010), SAMTOOLS (Li et al., 2009), FREEBAYES (Garrison et al., 2012), DEEPVARIANT (Poplin et al., 2017) e PLATYPUS (Rimmer et al., 2014).
Il file di output generato da questi programmi è in formato *Variant Calling Format* (VCF) (Danecek et al., 2011).

Il *Variant Calling Format* (VCF) è un formato di file testuale delimitato da tabulazioni che permette di descrivere le varianti di un genoma, insieme alla possibilità di inserire annotazioni (Figura 1.3). Per ciascun locus vengono riportate le seguenti informazioni:

1. la variante genetica per ciascun campione;

2. la qualità della variante genetica riportato in scala PHRED SCORE;
3. l’eterozigosità;
4. il numero di reads che supportano quella variante genetica.



```
##fileformat=VCFv4.2
##fileDate=20151002
##source=callMomV0.2
##reference=gi|251831106|ref|NC_012920.1| Homo sapiens mitochondrion, complete genome
##contig=<ID=MT,length=16569,assembly=b37>
##INFO=<ID=VT,Number=.,Type=String,Description="Alternate allele type. S=SNP, M=MNP, I=Indel">
##INFO=<ID=AC,Number=.,Type=Integer,Description="Alternate allele counts, comma delimited when multiple">
##FILTER=<ID=fa,Description="Genotypes called from fasta file">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  HG00096 HG00097 HG00099
MT      10      .       T       C       100     fa      VT=S;AC=3       GT      0       0       0
MT      16      .       A       T       100     fa      VT=S;AC=3       GT      0       0       0
MT      26      .       C       T       100     fa      VT=S;AC=3       GT      0       0       0
MT      35      .       G       A       100     fa      VT=S;AC=2       GT      0       0       0
MT      40      .       TC      CT      100     fa      VT=M;AC=1       GT      0       0       0
```

Figura 1.3. Prime righe di un file in formato VCF. Le linee che iniziano col simbolo #, sono l’intestazione (header). Le colonne riportano informazioni sul genoma di riferimento, posizione sul genoma di riferimento, allele del riferimento e allele alternativo, qualità dello SNPs, genotipo.

1.4 Il genere *Ophrys*

Il genere *Ophrys* appartiene alla famiglia delle *Orchidaceae* e comprende più di un centinaio di specie che risiedono soprattutto nell’area euro-mediterranea. Le *Ophrys* sono piante geofite bulbose, con apparato radicale costituito da due tuberi tondeggianti, peduncolati. Sono in genere piante esili, con fusto eretto, foglie basali riunite in rosetta, in genere di colore verde glauco, e foglie cauline bratteiformi. I taxa del genere *Ophrys* solitamente sono diploidi, con un numero cromosomico 2n=36. L’impollinazione è basata su un meccanismo di inganno sessuale. Per indurre gli insetti pronubi a visitare i loro fiori, le *Ophrys* adottano diverse strategie visive, tattili e olfattive. Il segnale più importante è sicuramente l’emissione di un “odore” simile ai ferormoni emessi dalla femmina dell’insetto quando giunge il momento dell’accoppiamento. La femmina dell’insetto raggiunge la maturità sessuale in un periodo successivo rispetto all’antesi dell’orchidea. Il maschio, così richiamato, si posa sul fiore e cerca di accoppiarsi (pseudocopula) caricandosi involontariamente delle masse polliniche. In seguito visiterà e quindi impollinerà un nuovo fiore, lasciandosi nuovamente ingannare. Le Ofridi mostrano una particolare propensione a formare ibridi, che spesso sono fertili e, a loro volta, si possono reincrociare. Le motivazioni di questa elevata tendenza a ibridarsi sono da ricercarsi non tanto nelle barriere genetiche, ma nella condivisione di insetti impollinatori che rendono più probabile un’impollinazione interspecifica.

Capitolo II

Assemblaggio e annotazione del genoma plastidiale

2.1 I plastidi

I plastidi sono organelli citoplasmatici semi-autonomi della cellula vegetale responsabili dello svolgimento di molte delle attività metaboliche, come la fotosintesi, la biosintesi degli acidi grassi, degli amminoacidi e dell'amido in tutti gli organismi eucarioti autotrofi. Essi derivano dalla stessa forma embrionale, chiamata *proplastidio*, e si suddividono in *cloroplasti*, *cromoplasti* e *leucoplasti*. I cloroplasti sono di colore verde, grazie all'elevato contenuto di clorofilla. Nelle piante sono presenti diverse copie di cloroplasti per cellula, con dimensioni che variano da 4 a 10 μm ed hanno forma ellissoidale. I cloroplasti internamente sono differenziati in un sistema di membrane, dette *tilacoidi,* immerse in una sostanza amorfa, detta *stroma*. I tilacoidi formano dei sacchi appiattiti disposti uno sull'altro a formare delle pile dette *grana*. Tutti i tilacoidi sono in continuità tra loro costituendo un sistema chiuso di membrane che racchiude una singola camera interna definita *lume*. Nello spessore della membrana dei tilacoidi è localizzato l'apparato fotochimico della fotosintesi. Ciascun cloroplasto nella cellula vegetale presenta una propria copia del genoma di eredità materna. Questo rende i cloroplasti indipendenti dal nucleo nel produrre le proteine che servono per la fotosintesi clorofiliana.

Il genoma plastidiale o *plastoma* è una molecola di DNA circolare aploide a doppio filamento. Esso ha una struttura simile al genoma batterico e per questo si pensa che derivi da fenomeni di endosimbiosi tra batteri e la cellula vegetale avvenuti miliardi di anni fa (McFadden et al., 2001). In seguito al loro ingresso nella cellula ospite, questi organelli hanno perso o trasferito al genoma nucleare gran parte dei loro geni. L'attivazione dei geni citoplasmatici integrati nel genoma nucleare può richiedere milioni di anni ed è influenzata da diversi fattori come la lunghezza del gene, la natura della sua sequenza codificante, la delezione del genoma plastidiale e la localizzazione nel genoma nucleare.

La grandezza del genoma plastidiale è variabile a seconda della specie ed è compresa tra 107 kb (*Cathaya argyrophylla*) e 218 kb (*Pelargonium*) costituendo una piccola parte del DNA totale, con percentuali di DNA plastidiale compresi tra lo 0,3% in *Picea abies* e il 37% in *Asclepias syriaca* (Twyford e Ness, 2016). Anche se la dimensione del genoma nelle piante a fiore varia da 63,6 Mbp in *Genlisea aurea* (Leushkin et al., 2013) a quasi 152,23 Gbp a *Paris japonica* (Pellicer et al., 2010), non sembra esserci una correlazione tra la dimensione del genoma e la percentuale di DNA organellare totale.

Il genoma plastidiale presenta una struttura quadripartita che include due sequenze ripetute di 10-25 kb chiamate *inverted repeat* (IR_A e IR_B) orientate in senso inverso. Queste due sequenze ripetute dividono il genoma in una regione piccola (*Small Single Copy*) e una grande a singola copia (*Large Single Copy)* (Figura 2.1). In molte piante, una delle due IR è stata completamente persa durante l'evoluzione (Wu et al., 2011).

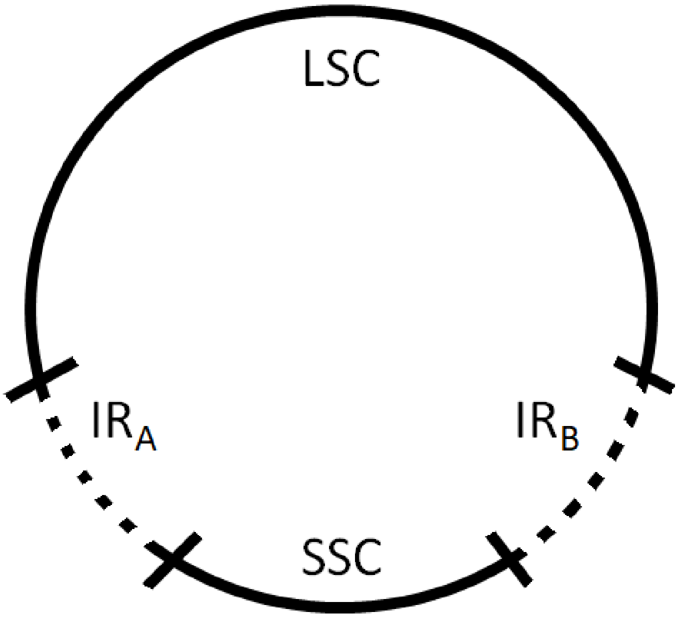


Figura 2.1. Schema riassuntivo della struttura quadripartita del genoma plastidiale.

Il genoma plastidiale include 120-130 geni che principalmente sono coinvolti nella trascrizione e nella traduzione delle proteine necessarie per lo svolgimento della fotosintesi (Jensen et al., 2014). Non sono rari casi di *eteroplasmia,* cioè la presenza di più di un tipo di genoma organellare (DNA mitocondriale o plastidiale) all'interno di una cellula o di un individuo (Scarcelli et al., 2015).

La struttura del genoma plastidiale è generalmente molto conservata nelle piante terrestri, ma durante il processo di evoluzione dall'endosimbiosi al cloroplasto, alcuni geni sono stati persi o trasferiti al nucleo (Jensen et al., 2014). Questi geni sono *inf*A, *rpl*22, *ndh* ed il loro trasferimento intracellulare dal cloroplasto al nucleo o al mitocondrio forniscono preziose informazioni per analisi filogenetiche e studi evolutivi.

Il gene *ndh* è coinvolto nel flusso ciclico degli elettroni fotosintetici al termine della fotosintesi clorofiliana e facilita la clororespirazione. Esso è composto da 11 subunità (A, B, C, D, E, F, G, H, I, J e K). Il gene *ndh*B è generalmente presente nell' *inverted repeat* e per questo è duplicato.

Il numero e le funzioni di questi geni plastidiali sono altamente conservati tra le piante superiori. Questi geni sono omologhi a quelli che codificano per le subunità mitocondriali della NADH deidrogenasi. Nei cloroplasti delle angiosperme, queste proteine *ndh* si associano alle subunità codificate dal DNA nucleare per formare un complesso simile alla deidrogenasi NADH. Questo complesso proteico si associa al fotosistema I per diventare un super-complesso che media il trasporto ciclico di elettroni (Munekage et al., 2004) e produce ATP per bilanciare il rapporto ATP / NADPH facilitando la clororespirazione quando il trasporto ciclico degli elettroni si ferma durante la notte (Peltier et al., 2002). La delezione delle subunità del gene *ndh* è molto frequente in alcune famiglie vegetali ed in particolare nelle Orchidaceae. In alcuni casi è stato visto che le delezioni non sono correlate a relazioni tassonomiche e evolutive (Lin et al., 2015).

Il gene plastidiale NADH deidrogenasi F (ndhF) si trova in tutte le divisioni delle piante vascolari ed è altamente conservato. Il suo frammento di DNA risiede nella piccola regione a copia singola del genoma plastidiale e in *O. iricolor* codifica per una proteina idrofoba contenente 597 amminoacidi. Il gene *ndh*F è stato spesso usato per la ricostruzione filogenetica a diversi livelli tassonomici. Il gene *ndh*F è spesso troncato o deleto nelle specie appartenenti alla famiglia delle orchidaceae e questo riarrangiamento molecolare è spesso correlato con lo spostamento della giunzione tra la *Small Single Copy* e la *Inverted Repeat*.

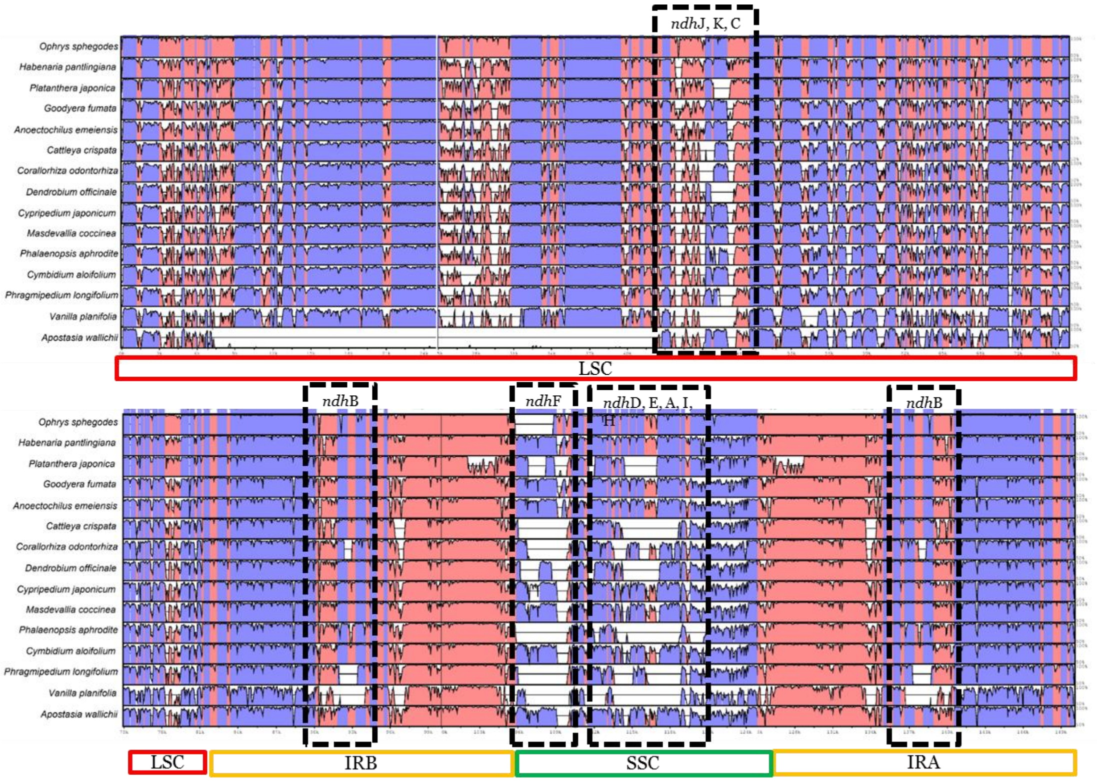


Figura 2.2. Grafico raffigurante la percentuale di omologia tra *Ophrys iricolor* e quindici specie appartenenti alla famiglia delle Orchidaceae. In blu sono riportate le regioni codificanti e in rosa quelle non codificanti. LSC = Large Single Copy, IRB = Inverted Repeat B, IRA = Inverted Repeat A, SSC = Small Single Copy.

2.2 Metodi computazionali per analisi di genomi plastidiali

L'avvento delle tecnologie di sequenziamento di seconda generazione ha accelerato e facilitato il progresso nel campo della genomica del plastidio. Da quando nel 1986 è stato sequenziato il primo genoma plastidiale (nel tabacco), oltre 1000 genomi sono stati sequenziati e assemblati. I genomi plastidiali annotati sono pubblicamente e gratuitamente disponibili nel NATIONAL CENTRE FOR BIOTECHNOLOGY INFORMATION (NCBI) ORGANELLE GENOME DATABASE (https://www.ncbi.nlm.nih.gov/genome/organelle/).

La disponibilità di questa grande mole di informazioni ha dato un contributo significativo agli studi filogenetici di diverse famiglie di piante e alla risoluzione delle relazioni evolutive all'interno di cladi filogenetici. I genomi plastidiali, inoltre, hanno rivelato notevoli variazioni all'interno e tra le specie vegetali anche dal punto di vista strutturale. Questa informazione è stata particolarmente preziosa per la comprensione dell'adattamento climatico di colture economicamente importanti, facilitando l'allevamento di specie strettamente correlate e l'identificazione e conservazione di tratti importanti (Wambugu et al., 2015) (Brozynska et al., 2016).

La comprensione delle variazioni tra i genomi plastidiali, ha anche permesso l'identificazione di trasferimento di geni plastidiali al genoma nucleare o a quello mitocondriale, migliorando la conoscenza sulla relazione tra questi tre genomi nelle piante.

Nonostante ci siano diverse tecniche di arricchimento di DNA plastidiale (Shi et al, 2012), al momento non è possibile separare completamente il DNA nucleare da quello plastidiale con tecniche di laboratorio. Questo spesso è un problema per l'assemblaggio di genomi di organismi non modello.

Esistono diverse pipeline per l'assemblaggio di genomi plastidiali, come ACRE (Wysocki et al., 2014), IOGA (Bakker et al., 2016), NOVOPlasty (Dierckxsens et al., 2017), Fast-Plast [https://github.com/mrmckain/Fast-Plast], un approccio basato su *k-mer* (Izan et al., 2017), ciascuna delle quali è basata su un principio diverso.

I genomi plastidiali possono potenzialmente fornire più segnale filogenetico di regioni intergeniche per la ricostruzione delle relazioni tra specie strettamente correlate (Carbonell-Caballero et al., 2015). Essendo di eredità materna ed avendo i geni delle funzioni molto conservate, ci possono fornire importanti informazioni filogenetiche che consentono di ricostruire relazioni evolutive e di distinguere specie per le quali problematiche ecologiche sono ancora aperte. Nell'attuazione del progetto di un assemblaggio di un genoma plastidiale, ci sono più aspetti da prendere in considerazione.

Un assemblaggio *de novo* completo di genomi plastidiali, spesso richiede una profondità del sequenziamento maggiore rispetto a quello basato sul riferimento (circa 50 – 100 x). Il secondo fattore da considerare è la percentuale relativa di DNA plastidiale rispetto a quello genomico totale.

Le principali difficoltà nell'assemblaggio di un genoma plastidiale sono la presenza di DNA nucleare e mitocondriale e delle *Inverted Repeat*. Ci sono diversi approcci per l'assemblaggio di un genoma plastidiale:

- Utilizzo di un genoma plastidiale di riferimento altamente omologo per separare le reads (FAST-PLAST)
- Se si dispone di un sequenziamento WGS è possibile separare le reads che hanno un più alto coverage e usare un approccio di *seed-extend* ossia quello di ricercare una sola reads plastidiale dal file FASTQ da cui assemblare l'intero genoma plastidiale (NOVOPLASTY).
- Approccio basato sui *k-mer*.

Capitolo III

Genotyping by Sequencing

3.1 Il protocollo “Genotyping By Sequencing”

L'innovativo approccio del G*enotyping By Sequencing* (GBS) è stato ideato come uno strumento a costi ridotti per studi di genetica di popolazione, caratterizzazione del germoplasma, miglioramento genetico e mappatura in organismi ad elevata diversità genetica (Elshire et al., 2011). Il principio su cui si basa questo protocollo è la riduzione dell’elevata complessità genomica dei campioni tramite digestione con endonucleasi di restrizione e successivo sequenziamento con le già citate tecniche NGS. Questo approccio ha il vantaggio di essere altamente specifico, riproducibile e in grado di raggiungere regioni del genoma che, con altre tecnologie, risulterebbero inaccessibili.

Le tecniche GBS e RAD-SEQ in particolare sono l’ideale per lo studio delle radiazioni recenti (Nadeau et al., 2013; Wagner et al., 2013), delimitazione delle specie (Leaché et al., 2014), e introgressione (Dasmahapatra et al., 2012; Eaton e Ree, 2013), dove un ampio campionamento di siti di migliaia di regioni attraverso il genoma può essere usato per caratterizzare l'eterogeneità nella distribuzione di modelli di alberi genetici e fornire potenza statistica per test dell'evoluzione reticolare (Durand et al., 2011; Reddy et al., 2017).

La tecnologia GBS ha permesso di ridurre notevolmente i costi per l’analisi del DNA, permettendo il sequenziamento di un grande pool di individui in un unico campione mediante l’utilizzo di adattatori ILLUMINA modificati, che presentano delle estremità appiccicose complementari al sito di taglio dell’enzima e a valle dell’estremità 5’ delle sequenze *barcodes*, che sono identificative dell’individuo e consentono al bioinformatico di risalire al campione di partenza mediante la fase del *demultiplexing*.

L’elevatissima variabilità delle endonucleasi rende la tecnica GBS estremamente versatile, grazie alla scelta di un’appropriata endonucleasi in base alla frequenza di taglio del DNA. Infatti, se si desidera individuare più marcatori molecolari si sceglie un enzima che taglia più frequentemente, altrimenti se ne sceglie uno con minore frequenza di taglio se si è interessati ad avere una maggiore profondità nell’analisi. Nel sequenziamento di genomi vegetali, la scelta opportuna di endonucleasi sensibili alle basi metilate consente di escludere le regioni ripetute del genoma (Davey et al., 2011).

Esistono diverse varianti del protocollo GBS che differiscono nel numero e nel tipo di enzimi che vengono utilizzati, così come nel tipo di attrezzatura richiesta per la preparazione dei campioni. Il metodo è stato descritto per la prima volta nel 2011 sulla rivista *PlosOne* (Elshire et al., 2011).

In sintesi sono 4 le fasi del sequenziamento dei campioni col protocollo GBS (figura 3.1):

1. I DNA ad alto peso molecolare vengono estratti e digeriti utilizzando un’endonucleasi di restrizione specifica precedentemente definita tagliando frequentemente nella frazione ripetitiva principale del genoma. ApeKI è l’endonucleasi più usata.
2. Gli adattatori, contenenti un diverso *barcodes* per campione ma stesso *index*, vengono quindi ligati alle estremità appiccicose e viene eseguita l'amplificazione mediante PCR.

3. Tutti i campioni preparati vengono uniti e si procede al sequenziamento.
4. Si procede alla fase del *demultiplexing*, cioè la separazione del DNA di ciascun individuo mediante i *barcodes* presenti all' estremità in 5' di ciascuna reads in forward.

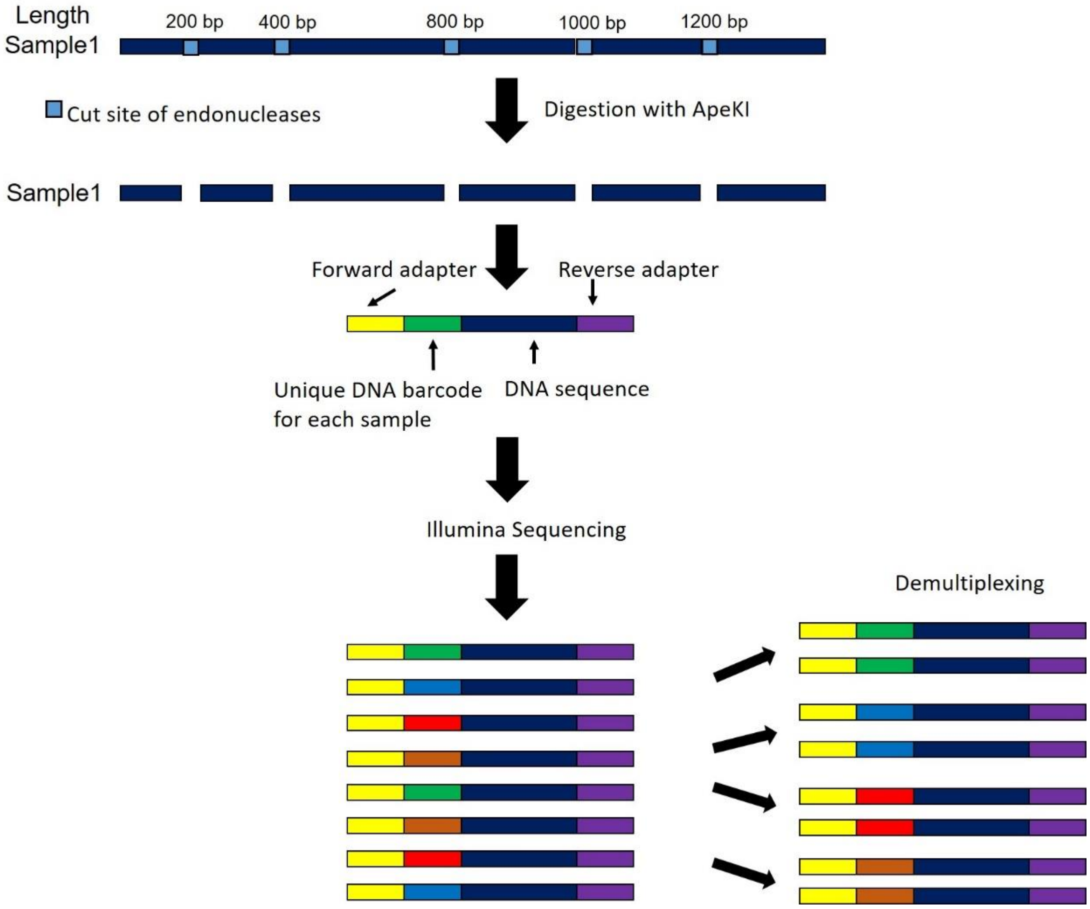


Figura 3.1. Protocollo GBS: Varie fasi di preparazione dei campioni con il protocollo GBS.

3.2 RAD-SEQ

Uno dei problemi principali del protocollo GBS, è la produzione di frammenti troppo lunghi causata spesso dalla perdita del sito di taglio dell'enzima in alcuni campioni. Per ovviare a questi problemi, recentemente sono nate nuove tecniche come il RAD-SEQ. Questa tecnica, a differenza del protocollo GBS, prevede una fase della preparazione dei campioni più lunga al fine di recuperare frammenti di DNA troppo lunghi che altrimenti verrebbero perse. Il protocollo RAD-SEQ, rispetto al GBS, richiede un secondo taglio, selezione e ligazione di adattatori dei frammenti più lunghi.

Per risolvere ulteriormente questo problema, è stata ideata un'altra variante del protocollo RAD-SEQ: il metodo dual digest RAD (DDRAD), che utilizza due enzimi di restrizione al fine di produrre frammenti più piccoli.

3.3 Analisi bioinformatica dei dati GBS

3.3.1 Software per l'analisi di dati GBS

Quando si ha a disposizione un buon genoma di riferimento, i dati GBS possono essere analizzati utilizzando un normale approccio di *variant calling,* come discusso precedentemente. Quando invece non si ha a disposizione un buon genoma di riferimento, l'analisi *de novo* dei dati GBS o RAD-SEQ richiede spesso l'uso di pipeline specifiche, che hanno l'obiettivo di selezionare e assemblare clusters di sequenze mediante una percentuale minima di omologia. Di seguito vengono citati alcuni software progettati per l'analisi dei dati GBS che sono stati largamente usati in progetti di filogenomica e di genetica di popolazione.

PYRAD è una pipeline progettata per l’analisi di dati GBS, RAD-SEQ e DDRAD (Eaton et al., 2014). Essa ha diverse dipendenze tra cui, il software VSEARCH che è un algoritmo di clustering di allineamento, che consente la variazione delle inserzioni e delezioni all'interno e tra i campioni, il software MUSCLE che consente l’allineamento delle sequenze e i pacchetti scritti in linguaggio PYTHON SCIPY e NUMPY. Lo scopo di PYRAD è quello di selezionare i loci che hanno una percentuale minima di omologia e infine di assemblare *de novo* un concatenamero in formato PHYLIP o NEXUS. Questo concatenamero viene usato per analisi filogenetiche ed è l’input dei seguenti programmi: RAXML, MRBAYES, PAUP o BEAST.

PYRAD consente di analizzare il dataset dando la possibilità di includere dati mancanti per ogni locus. Questo ha lo scopo di aumentare il dataset da analizzare al fine di avere un’analisi filogenetica più robusta. Gli output di PYRAD sono in formato VCF, PHYLIP e NEXUS.

STACKS è un’altra pipeline progettata per l’analisi dei dati GBS e RAD-SEQ. Essa è in grado di assemblare e analizzare sia organismi non modello (*de_novo_map.pl*), che modello (*ref_map.pl*). STACKS seleziona loci che hanno una minima percentuale di omologia e li inserisce tutti in una banca dati MYSQL. Infine STACKS genera in output un file in formato VCF e un file contenente gli aplotipi utili per l’analisi filogenetica. Infatti, STACKS a differenza di PYRAD non produce loci concatenati in output ma solo gli aplotipi contenenti i polimorfismi utili per l’analisi filogenetica.

3.4 Problemi principali nell’analisi dei dati GBS

3.4.1 L’importanza dei dati mancanti

Uno dei principali problemi che si ha durante l’analisi bioinformatica dei GBS e RAD-SEQ, è la capacità di gestire la grossa mole di dati mancanti generati dal confronto dei diversi campioni. Spesso può capitare di non avere sequenze omologhe tra i diversi campioni analizzati a causa delle possibili differenze nei siti di taglio dell’enzima o di una non corretta standardizzazione del numero delle sequenze nella fase della preparazione della libreria che precede il sequenziamento.

Selezionando esclusivamente i loci che sono in comune tra tutti i campioni, oltre a perdere una grandissima mole di informazione si rischia di avere un approccio più conservativo che non sempre consente di effettuare una corretta analisi filogenetica. Infatti, selezionando esclusivamente i loci in comune tra i diversi campioni c’è il rischio di filtrare solo quelli più conservati che potrebbero non fornirci nessuna informazione utile dal punto di vista filogenetico. Questo problema ha messo in discussione le tecniche GBS e RAD-SEQ sulla loro applicazione ad analisi filogenetiche più profonde (Rubin et al., 2012; Cariou et al., 2013).

In generale, il miglior approccio è quello di recuperare la maggior parte dei loci in quasi tutti i campioni oggetto di studio. Metodi bioinformatici vengono utilizzati per selezionare loci con una percentuale minima di dati mancanti (Eaton, 2014). Tuttavia, poiché i loci con informazioni mancanti per alcuni taxa possono ancora fornire informazioni filogenetiche per molti altri taxa, la maggior parte dei set di dati consente di combinare il 30-90% di dati mancanti multi-locus (Eaton et al., 2017). In generale, i dati mancanti tendono ad avere un impatto minimo sulla topologia dell’albero filogenetico (Rubin et al., 2012; Mastretta-Yaneset al., 2015) ma possono influire sulla lunghezza dei rami (Ogilvie et al., 2016).

3.4.2 Problema dei geni paraloghi

Come già accennato prima, quando si ha a disposizione un buon genoma di riferimento, i dati GBS possono essere analizzati utilizzando un normale approccio di *variant calling* in cui i geni paraloghi vengono facilmente identificati. Quando invece è richiesto un approccio *de novo* per mancanza di un genoma di riferimento, i clusters GBS devono essere selezionati e assemblati usando una percentuale minima di omologia (default 88%). In questo caso, la paralogia può essere valutata considerando la frequenza e gli eccessi di polimorfismi eterozigoti (Eaton, 2014). È quindi corretto valutare che non ci siano troppi SNPs eterozigoti per locus al fine di scartare eventuali geni paraloghi.

3.4.3 Programmi per l’analisi filogenetica dei dati

Ricostruire la storia evolutiva delle specie è stato da sempre uno dei problemi più importanti in biologia. La filogenesi è la scienza che studia l’evoluzione della specie a partire da allineamenti multipli di sequenze. Il principio su cui si basa, è quello di considerare l’accumulo di mutazioni: più tempo è passato dal momento in cui due geni si sono originati da un comune antenato per speciazione o duplicazione, maggiori sono le differenze che ci aspettiamo di osservare dal confronto dei due prodotti genici odierni. Esistono diversi metodi per ricostruire la filogenesi e i loro punti deboli sono già stati evidenziati e analizzati (Felsestein, 2004). Alcuni di essi sono metodi “algoritmici”, che permettono di avere in output un albero filogenetico. Il più usato è il metodo *neighbor joining* e produce un albero che risponde al principio della *minima evoluzione,* o *massima parsimonia*, cioè l'albero che ipotizza il percorso evolutivo più breve. L'albero viene costruito a partire da una matrice delle distanze che riporta per ogni coppia di sequenze un punteggio che misura il grado di diversità. Altri metodi per la ricerca degli alberi, sono i metodi della *massima verosimiglianza* (*maximum likelihood*) e i metodi *bayesiani*. Entrambi si basano su un particolare modello di evoluzione dei caratteri rappresentati nei dati (solitamente sequenze di DNA). Essi si basano su diversi modelli evolutivi che possono essere scelti in base al grado di divergenza delle specie analizzate e al presunto tasso di sostituzione nucleotidica per quel tratto genico; esistono quelli che assumono che tutte le sostituzioni sono ugualmente probabili e che un tasso di sostituzione costante possa essere stimato sulla base dei dati. Diversamente, esistono altri modelli che potrebbero assumere che diversi tipi di sostituzione avvengano con velocità diverse (modello Kimura a due parametri). Scelto il modello, per ciascuno degli alberi possibili il metodo *maximum likelihood* calcola la probabilità di osservare i dati disponibili secondo quel modello e quella particolare filogenesi. Il metodo statistico più usato, per validare la topologia di un albero filogenetico ottenuto con il metodo della *massima verosimiglianza,* è il *bootstrap*. Il metodo *bootstrap,* è una tecnica statistica che consiste nel ricampionare casualmente l'allineamento multiplo iniziale e generare un numero *N*, generalmente elevato di nuovi allineamenti multipli. Infine ad ogni nodo verrà assegnato un valore da 1 a 100 che riporta in percentuale la probabilità di essere corretto. Il programma più usato per il metodo *maximum likelihood*, è RAXML (Stamatakis, 2014). RAXML richiede in input il file in formato PHYLIP. Il metodo *bayesiano*, sviluppato più recentemente e sempre più diffuso, differisce dal *maximum likelihood* poiché massimizza la probabilità di ottenere un determinato albero, sulla base del modello scelto e dei dati. Il metodo *bayesiano* calcola le probabilità di diversi alberi, cosicchè essi possano essere confrontati. Il programma più usato per il metodo *bayesiano*, è MRBAYES (Huelsenbeck et al., 2001). MRBAYES richiede in input il file in formato NEXUS.

Obiettivo del Dottorato

L'obiettivo di questo dottorato è stato quello di ottimizzare tecniche bioinformatiche per risolvere problematiche in un contesto evolutivo. Questo obiettivo è stato perseguito attraverso l'approfondimento di due casi di studio utlizzando organismi non modello appartenenti al genere *Ophrys* (Orchidaceae). Questo genere è particolarmente interessante e rappresenta una sfida aperta per i ricercatori, in quanto è caratterizzato da una rapida radiazione evolutiva che ha impedito una chiara identificazione delle specie utilizzando le tradizionali tecniche genetiche. Nel capitolo 4 ho utilizzato tecniche bioinformatiche innovative per l'assemblaggio (*seed-extend*) e l'annotazione del genoma plastidiale di due specie (*O. sphegodes* ed *O. iricolor*).

Nel capitolo 5 ho utilizzato tecniche bioinformatiche per analizzare dati GBS. Le metodiche di analisi utilizzate hanno consentito di ottenere una chiara identificazione delle specie ed anche di delineare la storia biogeografica delle specie studiate.

References

Bakker FT, Lei D, Yu J,Mohammadin S,Wei Z, Kerke S,Gravendeel B, Nieuwenhuis M, Staats M, Alquezar DE. 2016. Herbarium genomics: plastome sequence assembly from a range of herbarium specimens using an Iterative Organelle Genome Assembly pipeline. *Biological Journal of the Linnean Society* 117: 33-43.

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

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30: 2114-2120.

Brozynska M, Furtado A, Henry RJ. 2016. Genomics of crop wild relatives: expanding the gene pool for crop improvement. *Plant Biotechnology Journal* 14: 1070-1085.

Carbonell-Caballero J, Alonso R, Ibañez V, Terol J, Talon M, Dopazo J. 2015. A phylogenetic analysis of 34 chloroplast genomes elucidates the relationships between wild and domestic species within the genus Citrus. *Molecular Biology and Evolution* 32: 2015-2035.

Cariou M, Duret L, Charlat S. 2013. Is RAD-seq suitable for phylogenetic inference? An *in silico* assessment and optimization. *Ecology and Evolution* 3: 846-852.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, McVean G. 2011. The variant call format and VCFtools. *Bioinformatics* 27: 2156-2158.

Dasmahapatra KK, Walters JR, Briscoe AD, Davey JW, Whibley A, Nadeau NJ, Zimin AV. 2012. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature* 487: 94.

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics* 12: 499.

Dierckxsens N, Mardulyn P, Smits G. 2017. NOVOPlasty: *de novo* assembly of organelle genomes from whole genome data. *Nucleic Acids Research* 45: 18

Durand EY, Patterson N, Reich D, Slatkin M. 2011. Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution* 28: 2239-2252.

Eaton DAR, Ree RH. 2013. Inferring phylogeny and introgression using RADseq data: An example from flowering plants (*Pedicularis*: Orobanchaceae). *Systematic Biology* 62: 689-706.

Eaton DAR. 2014. PyRAD: Assembly of de novo RADseq loci for phylogenetic analyses. *Bioinformatics* 30: 1844-1849.

Eaton DAR, Spriggs EL, Park B, Donoghue MJ. 2017. Misconceptions on missing data in RAD-seq phylogenetics with a deep-scale example from flowering plants. *Systematic Biology* 66: 399-412.

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PloS one* 6: e19379.

Felsenstein J. 2003. Inferring Phylogeny. *Sinauer Associates*, Sunderland, MA.

Fuller CW, Middendorf LR, Benner SA, Church GM, Harris T, Huang X, Vezenov DV. 2009. The challenges of sequencing by synthesis. *Nature biotechnology* 27: 1013.

Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. 2012. *arXiv preprint arXiv:1207.3907*.

Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogeny. *Bioinformatics* 17:754-755.

Izan S, Esselink D, Visser RG, Smulders MJ, Borm T. 2017. De novo assembly of complete chloroplast genomes from non-model species based on a K-mer frequency-based selection of chloroplast reads from total DNA sequences. *Frontiers in plant science* 8: 1271.

Jensen PE, Leister D. Chloroplast evolution, structure and functions. 2014. *F1000Prime Reports*. 6: 40.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature methods* 9: 357.

Leache AD, Oaks JR. 2017. The utility of single nucleotide polymorphism (SNP) data in phylogenetics. *Annual Review of Ecology, Evolution, and Systematics* 48: 69-84.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics* 25: 2078-2079.

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

Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, Wang J. 2009. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* 25: 1966-1967.

Lin CS, Chen JJ, Huang YT, Chan MT, Daniell H, Chang WJ, Liao CF. 2015. The location and translocation of ndh genes of chloroplast origin in the Orchidaceae family. *Scientific reports* 5: 9040.

Mastretta-Yanes A, Arrigo N, Alvarez N, Jorgensen TH, Pinero D, Emerson BC. 2015. Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. *Molecular Ecology Resources* 15: 28-41.

McCarthy A. 2010. Third generation DNA sequencing: pacific biosciences single molecule real time technology. *Chemistry & biology* 17: 675-676.

McFadden GI. 2001. Chloroplast origin and integration. *Plant Physiology* 125: 50-53.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, DePristo MA. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* 20: 1297-303.

Hashimoto M, Miyake C, Tomizawa K, Endo T, Tasaka M, Shikanai T . 2004. Cyclic electron flow around photosystem I is essential for photosynthesis. *Nature* 429:579-582.Peltier G Cournac L. 2002. Chlororespiration. *Annu. Rev. Plant Biol*. 53:523-550.

Nadeau NJ, Martin SH, Kozak KM, Salazar C, Dasmahapatra KK, Davey JW, Baxter SW, Blaxter ML, Mallet J, Jiggins CD. 2013. Genome-wide patterns of divergence and gene flow across a butterfly radiation. *Molecular Ecology* 22: 814-826.

Ogilvie HA, Heled J, Xie D, Drummond AJ. 2016. Computational performance and statistical accuracy of BEAST and comparisons with other methods. *Systematic Biology* 65: 381-396.

Poplin R, Newburger D, Dijamco J, Nguyen N, Loy D, Gross SS, DePristo MA. 2017. Creating a universal SNP and small indel variant caller with deep neural networks. *BioRxiv* 092890.

Reddy CB, Hickerson MJ, Frantz LAF, Lohse K. 2017. Blockwise site frequency spectra for inferring complex population histories and recombination. *BioRxiv* 77958.

Rimmer A, Phan H, Mathieson I, Iqbal Z, Twigg SRF, WGS500 Consortium, Wilkie AOM, McVean G, Lunter G. 2014. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nature Genetics* 46: 912-918.

Rubin BER, Ree RH, Moreau CS. 2012. Inferring phylogenies from RAD sequence data. *PLoS One* 7: e33394.

Scarcelli N, Mariac C, Couvreur TL, Faye A, Richard D, Sabot F, Berthouly-Salazar C, Vigouroux Y. 2016. Intra-individual polymorphism in chloroplasts from NGS data: where does it come from and how to handle it? *Molecular Ecology Resources* 16: 434-445.

Shi C, Hu N, Huang H, Gao J, Zhao YJ, Gao LZ. 2012. An Improved Chloroplast DNA Extraction Procedure for Whole Plastid Genome Sequencing. *PLoS One* 7: e31468.

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I. 2009. ABySS: a parallel assembler for short read sequence data. *Genome research* gr-089532.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30: 1312–1313.

Twyford AD, Ness RW. 2016. Strategies for complete plastid genome sequencing. *Molecular Ecology Resources* 17: 858-868.

Wagner CE, Keller I, Wittwer S, Selz OM, Mwaiko S, Greuter L, Sivasundar A, Seehausen O. 2013. Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. *Molecular Ecology* 22: 787-798.

Wambugu P, Brozynska M, Furtado A, Waters D, Henry R. 2015. Relationships of wild and domesticated rices (Oryza AA genome species) based upon whole chloroplast genome sequences. *Scientific Reports* 5: 13957.

Wysocki WP, Clark LG, Kelchner SA, Burke SV, Pires JC, Edger PP, Duvall MR. 2014. A multi-step comparison of short-read full plastome sequence assembly methods in grasses. *Taxon* 63: 899-910.

Wu CS, Wang YN, Hsu CY, Lin CP, Chaw SM. 2011. Loss of Different Inverted Repeat Copies from the Chloroplast Genomes of Pinaceae and Cupressophytes and Influence of Heterotachy on the Evaluation of Gymnosperm Phylogeny. *Genome Biology and Evolution* 3: 1284-1295.

Zerbino D, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research* gr-074492.

Capitolo IV

The complete plastid genomes of *Ophrys iricolor* and *O. sphegodes* (Orchidaceae) and comparative analyses with other orchids

4.1 Abstract

Sexually deceptive orchids of the genus *Ophrys* may rapidly evolve by adaptation to pollinators. However, understanding of the genetic basis of potential changes and patterns of relationships is hampered by a lack of genomic information. We report the complete plastid genome sequences of *Ophrys iricolor* and *O. sphegodes,* representing the two most species-rich lineages of the genus *Ophrys*. Both plastomes are circular DNA molecules (146754 bp for *O. sphegodes* and 150177 bp for *O. iricolor*) with the typical quadripartite structure of plastid genomes and within the average size of photosynthetic orchids. 213 Simple Sequence Repeats (SSRs) (31.5% polymorphic between *O. iricolor* and *O. sphegodes*) were identified, with homopolymers and dipolymers as the most common repeat types. SSRs were mainly located in intergenic regions but SSRs located in coding regions were also found, mainly in *ycf1* and *rpoC2* genes. The *Ophrys* plastome is predicted to encode 107 distinct genes, 17 of which are completely duplicated in the Inverted Repeat regions. 83 and 87 putative RNA editing sites were detected in 25 plastid genes of the two *Ophrys* species, all occurring in the first or second codon position. Comparing the rate of nonsynonymous (dN) and synonymous (dS) substitutions, 24 genes (including *rbc*L and *ycf*1) display signature consistent with positive selection. When compared with other members of the orchid family, the *Ophrys* plastome has a complete set of 11 functional *ndh* plastid genes, with the exception of *O. sphegodes* that has a truncated *ndh*F gene. Comparative analysis showed a large co-linearity with other related Orchidinae. However, in contrast to *O. iricolor* and other Orchidinae, *O. sphegodes* has a shift of the junction between the Inverted Repeat and Small Single Copy regions associated with the loss of the partial duplicated gene *ycf*1 and the truncation of the *ndh*F gene. Data on relative genomic coverage and validation by PCR indicate the presence, with a different ratio, of the two plastome types (i.e. with and without *ndh*F deletion) in both *Ophrys* species, with a predominance of the deleted type in *O. sphegodes*. A search for this deleted plastid region in *O. sphegodes* nuclear genome shows that the deleted region is inserted in a retrotransposon nuclear sequence. The present study provides useful genomic tools for studying conservation and patterns of relationships of this rapidly radiating orchid genus.

4.2 Introduction

Plastids such as chloroplasts are important plant organelles involved in the photosynthetic process thus providing essential energy to plants [1]. Plastids have small circular genomes, ranging from 135 to 160 kb [2–4]. Most angiosperm plastid genomes so far annotated have a quadripartite structure containing two copies of Inverted Repeat (IR) regions, separating a Large Single Copy (LSC) and Small Single Copy (SSC) regions [5–7]. Recently, with the extraordinary advances in sequencing platforms, many plastid genomes have been annotated and have provided valuable tools for the understanding of plant phylogenies and genome evolution e.g. [8]. Plastid structure and gene order are generally stable, and the rate of nucleotide substitution is slow [9] so that plastid genomes were traditionally considered to have experienced rearrangements rarely enough to be suitable to demarcate major plant groups [10]. Nonetheless, several angiosperm lineages show extensive gene order changes in plastid genomes that are often correlated with increased rates of nucleotide substitutions and gene and/or multiple intron losses [11, 12]. These rearrangements in the plastid genome have been found to be often associated with repeated sequences [2].

The family Orchidaceae consists of more than 700 genera and approximately 28,000 species [13], which are distributed in a wide variety of habitats. So far, several complete plastid genomes have been annotated in different orchid lineages. These studies revealed that Orchidaceae often underwent accelerated plastome evolution including large inversions, shifts in boundaries between IRs and the two single copies, indels, intron losses, and pseudogene formation by stop codons often associated with shifts from heterotrophy to parasitism/heterotrophism [14,15]. Compared to other angiosperms, photosynthetic orchids were also found particularly variable in the conservation of *NADH* dehydrogenase *(ndh)* genes [16], that encode components of the thylakoid complex involved in the redox level of the cyclic photosynthetic electron transporters.

The number of intact and degraded *ndh* genes present in the orchids plastomes varies even among closely related species suggesting that this specific gene class may be actively degraded in Orchidaceae [17]. This is not surprising as gene transfer from plastid to nucleus is known to occur frequently during evolutionary processes as even the complete loss of some plastid-encoded *ndh* genes seems to not affect the plant life [15]. Indeed, there is no clear-cut evidence of phylogenetic signal in the pseudogenization or loss of the *ndh* genes. For instance, no correlation with phylogeny was found for *ndh* genes loss in the Epidendroideae lineages while related species of Oncidiinae show a consistent loss of two *ndh* genes (*ndh*F and *ndh*K) and pseudogenization by gene truncation of other five genes (*ndh*A, D, H, I and J) [18].

The IR/SC junctions represent another hotspot of orchid plastome evolution, with the rearrangement of flanking regions leading to expansion or contraction of the inverted repeat regions. Different types of junctions have been reported in orchids, with considerable variation particularly in the *ycf*1 gene [19]. It has been hypothesized that the exhibited usage bias of A/T base pairs typical of all known orchid *ycf*1 genes would render less stable the DNA in the *ycf*1 gene thus leading to the higher recombination of IR/SSC junction [20]. This often leads to a consequent partial or complete degradation of the *ndh*F gene, or even, in some case, to its transfer to mitochondrial DNA by intraorganellar recombination [17].

Despite Orchidaceae represents approximately 1/8 of all flowering plants [13], most published plastid sequences belong to tropical orchid lineages, while there is a remarkable dearth of information for the important temperate terrestrial subtribe Orchidinae with only two *Habenaria* and one *Platanthera* species plastomes having been annotated so far [17, 21]. With the aim to fill this gap, we sequenced the complete plastid genomes of *Ophrys iricolor* and *Ophrys sphegodes*. These species are representative of the two main diverging lineages of the Mediterranean *Ophrys,* a sexually deceptive genus belonging to the subtribe Orchidinae characterized by an elevated taxonomic complexity due to a very fast radiation by pollinator shifts [22, 23]. The specific aims of the present study were to (i) annotate the complete plastid genome sequences of two *Ophrys* species, (ii) evaluate the homology between these two plastomes, (iii) investigate any significant characteristics suggesting plastome rearrangement in *Ophrys* and their phylogenetic signal, and (iv) explore significant changes in gene content and gene order in the subtribe Orchidinae compared to other orchid subtribes.

4.3 Materials and methods

4.3.1 Genome sequencing, assembling and annotation

DNA was extracted from a specimen of *Ophrys iricolor* (collected between Miamou and Agios Kyrillos, Crete, Greece; N34.9693, E24.9154; under permit number 118565/3022 issued by the Ministry of Environment and Energy in Athens on 13.02.2015) and from a specimen of *Ophrys sphegodes* (collected between Cagnano Varano and San Nicandro Garganico, Apulia, Italy; N41.9133, E15.6784 under permit number 173 issued by the National Park of Gargano in Monte Sant’Angelo (FG) on 12.01.2016). Whole genomic libraries were sequenced in paired-end mode, 2 x 150 bp, using the Illumina HiSeq 4000 platform (Illumina Inc., San Diego, CA, USA) at the Functional Genomics Centre Zurich (Switzerland). The obtained reads were trimmed using the software TRIMMOMATIC v. 0.36 [24] and the resulting trimmed reads (309,012,252 reads for *O. sphegodes* and 251,959,572 reads for *O. iricolor*) were *de novo* assembled using NOVOPLASTY v. 2.5.2 [25]. The gene annotation of the *Ophrys* plastid genomes was carried out using the software GESEQ v. 1.42 [26] and BLAST v. 2.6.0 [27] searches. From this initial annotation analysis, putative starts and stops of the gene exons, along with the positions of the related introns, were determined based on comparisons to homologous genes in other plastid genomes [28]. All tRNA genes were verified by using tRNAscan-SE server v. 1.3.1 [29]. The physical maps of the plastid circular genomes were drawn using Organellar Genome DRAW (OGDRAW) v. 1.2.1 [30]. The complete plastome sequences of *Ophrys sphegodes* and *O. iricolor* were deposited in the Sequence Reads Archive (NCBI-SRA) database under the accession number SRP148126. BLAST v. 2.6.0 [27, 31] was used to check whether deleted part of the *ndh*F gene in the *O. sphegodes* plastid genome was translocated into the nuclear genome. Reads were realigned against the assembled scaffolds of *O. sphegodes* nuclear genome (unpublished) using BWA v. 0.7.16 and converted in BAM [32] format using SAMtools v. 1.5 [33]. Finally, a BLASTX search was performed to annotate the nuclear *O. sphegodes* scaffold1075174.

4.3.2 Genome structure, deletions validation, and repeat sequences

The software MAFFT v. 7.205 [34] and the Perl script Nucleotide MUMmer (NUCmer) available in MUMmer 3.0 [35] were employed to compare the plastome structures between *O. sphegodes* and *O. iricolor*. To detect putative errors in the *de novo* assemblies, the trimmed reads were mapped to the assembled genomes using the aligner BWA [32], converted to BAM format using SAMtools [33] and finally visualized using the IGV genome browser v. 2.4 [36]. To validate the deletion *in silico*, BAM files were further analysed using the software BEDtools coverage v. 2.21.0 [37] which generated a table in BED format containing an interval “windows” with coverage information across the two *Ophrys* plastomes. The BED file format was in turn used to visualize the sequencing coverage in regions of interest using the software CNView v. 1.0 [38]. To experimentally validate the *ndh*F deletion in *O. sphegodes*/*O. iricolor*, we designed primers for both the flanking and internal regions of *ndh*F from the assembled plastomes (S1 Fig a). With these primers, we PCR amplified DNAs of *O. sphegodes* and *O. iricolor* from different localities and of *O. incubacea* and *O. fusca*, as close relatives to *O. sphegodes* and *O. iricolor*, respectively and *O. insectifera* as distant related. PCR reaction conditions were as described in [39], with 5 ng of total DNA as template. Amplification products were visualized on 2% agarose gel using a 100 bp ladder as standard. PCR products and ladder were stained with ethidium bromide and photographed using a digital camera. Confirmatory sequences of the PCR products were done with ABI3130 automatic sequencer following manufacture instructions. Simple sequence repeats (SSRs) or microsatellites were detected using the MIcroSAtellite (MISA) Perl script v. 1.0 [40]. Thresholds were set at eight repeat units for mononucleotide SSRs, four repeat units for di- and trinucleotide SSRs, and three repeat units for tetra-, penta- and hexanucleotide SSRs as done in [41]. We also analysed tandem repeat sequences from the plastid genomes of *O. sphegodes* and *O. iricolor* and searched for forward, reverse and palindromic repeats by using REPuter [42]. We limited the maximum computed repeats and the minimal repeat size to 50 and 8, respectively and with a Hamming distance equal to 1.

4.3.3 Prediction of RNA editing sites and identification of positive signatures in plastid proteincoding genes

Potential RNA editing sites in protein-coding genes of *Ophrys* plastome were predicted by the program PREPACT v. 2.0 [43] using the following 30 highly homologous reference genes from *Phalaenopsis aphrodite*: *acc*D, *atp*A, *atp*B, *atp*F, *atp*I, *ccs*A, *clp*P, *mat*K, *pet*B, *pet*D, *pet*G, *pet*L, *psa*B, *psa*I, *psb*B, *psb*E, *psb*F, *psb*L, *rpl*2, *rpl*20, *rpl*23, *rpo*A, *rpo*B, *rpo*C1, *rpo*C2, *rps*2, *rps*8, *rps*14, *rps*16, and *ycf*3.

In order to identify putative genes under positive selection, the 67 protein-coding genes present in sixteen Orchidaceae plastomes (*Ophrys iricolor*, AP018716 *O. sphegodes* AP018717, *Cattleya crispata* NC_026568.1, *Corallorhiza odontorhiza* KM390021.1, *Cymbidium aloifolium* NC_021429.1, *Cypripedium japonicum* KJ625630.1, *Goodyera procera* NC_029363.1, *Habenaria pantlingiana* NC_026775.1, *Masdevallia coccinea* NC_026541.1, *Phalaenopsis aphrodite* NC_017609.1, *Anoectochilus emeiensis* NC_033895.1, *Apostasia wallichii* NC_030722.1, *Dendrobium officinale* KX377961.1, *Phragmipedium longifolium* KM032625.1, *Platanthera japonica* MG925368.1, *Vanilla planifolia* KJ566306.1) were downloaded from Genbank. We analysed all coding gene regions, except *ndh* genes, due to their frequent loss across the entire set of orchids listed here.

In order to build a reference phylogenetic tree, all genes were aligned using MAFFT software v. 7.205 [44] and were concatenated using MESQUITE software v. 3.5 [45]. PARTITION FINDER software v. 2.1.0 [46] was used in order to search the best evolution model for each gene and a reference phylogenetic tree was built using RAxML software v. 8.2.10 using 1000 bootstrap replicates [47]. The positive signatures were analysed using SELECTON server v. 2.4 (http://selecton.tau.ac.il/index.html; [48], *Ophrys iricolor* was used as query sequence (i.e. the plastome type without *ndh*F deletion) and codon alignment was done using the software MAFFT v. 7.205 [44] implemented in SELECTON software. The phylogenetic tree was set as input in SELECTON analyses and branch lengths were automatically optimized from the software. The gene divergence was estimated by the sum of total branch lengths that link the operational taxonomical units to the common ancestor of Orchidaceae species sampled here as done in [28]. SELECTON software generated for each gene as output the number of putative sites under positive selection. In order to test whether positive selection is operating on a protein, a Likelihood Ratio Test for positive selection was performed with the comparison of M8 (allows positive selection) against M8a (null model). We consider in our analysis only sites where possible positive selection was inferred (lower bound > 1 and test with probability < 0.01). P-values were adjusted for multiple testing in R (R Core Team) using FDR method in the *p.adjust* function.

4.4 Results and Discussion

4.4.1 Genome organization and features

The plastomes of the two *Ophrys* species are circular DNA molecules of 146,754 bp for *O. sphegodes* and 150,177 bp for *O. iricolor* with the typical quadripartite structure of plastid genomes of flowering plants (Fig 1): a pair of inverted repeats of 25,052 bp and 26,348 bp, respectively, separated by a large single copy (LSC) region (80,471 bp and 80,541, respectively) and a small single copy (SSC) region of 16,179 bp and 16,940 bp, respectively for *O. sphegodes* (DDBJ accession number AP018717) and *O. iricolor* (DDBJ accession number AP018716). The size of the *Ophrys* plastid genome was comparable to other published plastomes of photosynthetic orchids. The plastomes of the two *Ophrys* species are largely collinear with the exception of a large deletion in the *ndh*F gene in *O. sphegodes*.

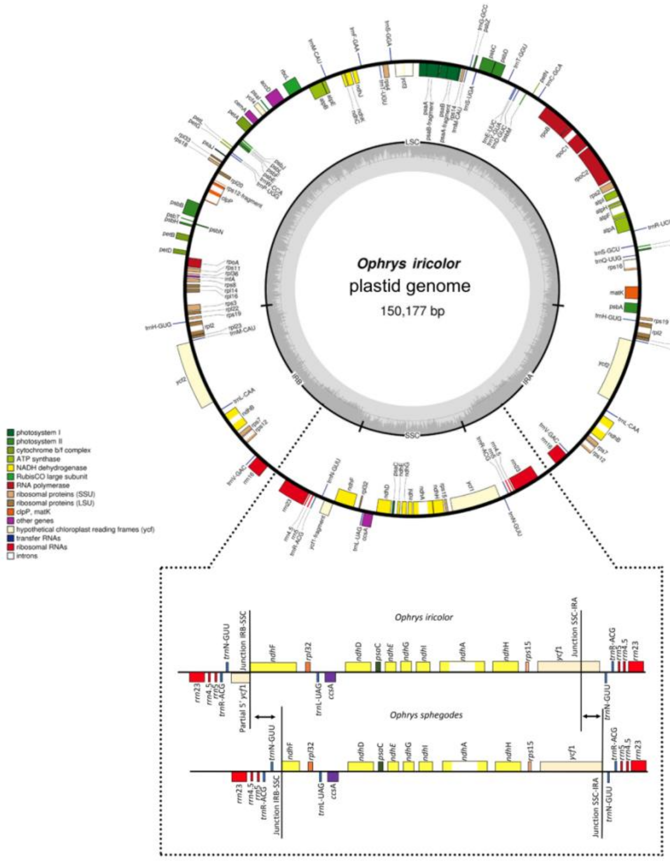


Figure 1. Gene map of *Ophrys sphegodes* and *Ophrys iricolor* plastid genomes. Genes drawn inside the circle are transcribed in the clockwise direction, and genes drawn outside are transcribed in the counter-clockwise direction. Different functional groups of genes are colour-coded. The darker grey in the inner circle corresponds to G/C content, and the lighter grey corresponds to A/T content. LSC, Large Single Copy; SSC, Small Single Copy; IRA/B, Inverted Repeat A/B. The enlargement shows that the loss of the partial duplicated gene of *ycf*1 and the truncation of *ndh*F gene in *O. sphegodes* are correlated with the shift of the junction between the IR and SSC.

The percentage of plastid reads in total WGS data was 5.43 % for *O. sphegodes* and 1.96 % for *O. iricolor*. The lowest average coverage of the assembled plastid genomes used was 13,673x for *O. sphegodes* and 3,816x for *O. iricolor*. The

G/C contents were 37.14 % and 36.4% respectively for *O. sphegodes* and *O. iricolor*, similar to other angiosperms (Table 1). The *Ophrys* plastome is predicted to encode 107 distinct genes, 17 of which are completely duplicated in the IR regions resulting in a total of 124 genes (Table 2). The annotation revealed distinct protein-coding genes (seven of them completely duplicated, namely *ndh*B, *rpl*2, *rpl*23, *rps*7, *rps*12, *rps*19 and *ycf*2), 30 distinct tRNAs genes (five of them duplicated, *trn*H-GUG, *trn*L-CAA, *trn*N-GUU, *trn*R-ACG, *trn*V-GAC and one triplicated *trn*M-CAU), and four distinct rRNA genes (all of them completely duplicated: *rrn*4.5, *rrn*5, *rrn*16 and *rrn*23). A truncated gene *ndh*F, was identified in *O. sphegodes* but not in *O. iricolor*. Ten genes contain one intron (*atp*F, *ndh*A, *ndh*B, *pet*B, *pet*D, *rpl*2, *rpl*16, *rps*12, *rps*16 and *rpo*C1) and two genes (*clp*P and *ycf*3) contain two introns.

	Ophrys sphegodes	*Ophrys iricolor*
Plastid reads	16,782,955 bp	4,952,605 bp
Average plastid coverage	13,673 x	3,816 x
G/C percentage	37.14 %	36.4 %
Large Single Copy Region	80,471 bp	80,541 bp
Small Single Copy Region	16,179 bp	16,940 bp
Inverted Repeats	25,052 bp	26,348 bp

Table 1. Comparison of two *Ophrys* plastid genomes

Group of gene	Name of gene
Ribosomal RNA genes	*rrn*16[a]; *rrn*23[a]; *rrn*4.5[a]; *rrn*5[a]
Transfer RNA Genes	*trn*C-GCA; *trn*D-GUC; *trn*E-UUC; *trn*F-GAA; *trn*G-CCC; *trn*G-GCC; *trn*H-GUG[a]; *trn*L-CAA[a]; *trn*L-UAG; *trn*M-CAU[c]; *trn*N-GUU[a]; *trn*P-UGG; *trn*Q-UUG; *trn*R-ACG[a]; *trn*R-UCU; *trn*S-GCU; *trn*S-UGA; *trn*S-GGA; *trn*T-GGU; *trn*T-UGU; *trn*V-GAC[a]; *trn*W-CCA; *trn*Y-GUA
Small subunit of ribosome	*rps*2; *rps*3; *rps*4; *rps*7[a]; *rps*8; *rps*11; *rps*12[a]; *rps*14; *rps*15; *rps*16; *rps*18; *rps*19[a]
Large subunit of ribosome	*rpl*2[a]; *rpl*14; *rpl*16; *rpl*20; *rpl*22; *rpl*23[a]; *rpl*32; *rpl*33; *rpl*36
DNA-dependent RNA polymerase	*rpo*A; *rpo*B; *rpo*C1; *rpo*C2
Genes for photosynthesis:	
Subunits of photosystem I (PSI)	*psa*A; *psa*B; *psa*C; *psa*I; *psaJ*; *ycf*3; *ycf*4
Subunits of photosystem II (PSII)	*psb*A; *psb*B; *psb*C; *psb*D; *psb*E; *psb*F; *psb*H; *psb*I; *psb*J; *psb*K; *psb*L; *psb*M; psbN, *psb*T; *psb*Z
Subunits of cytochrome b_6f	*pet*A; *pet*B; *pet*D; *pet*G; *pet*L; *pet*N
Subunits of ATP synthase	*atp*A; *atp*B; *atp*E; *atp*F; *atp*H; *atp*I
Subunits of NADH dehydrogenase	*ndh*A; *ndh*B[a]; *ndh*C; *ndh*D; *ndh*E; *ndh*F[b]; *ndh*G; *ndh*H; *ndh*I; *ndh*K; *ndh*J
Large subunits of Rubisco	*rbc*L
Other genes:	
Maturase	*mat*K
Envelope membrane protein	*cem*A
Subunit of acetyl-CoA carboxylase	*acc*D
C-type cytochrome synthesis	*ccs*A

gene	
Protease	*clp*P
Component of TIC complex	*ycf*1[d]
Translation initiation factor IF-1	*inf*A
Genes of unknown function	*ycf*2[a]

[a] Duplicated gene; [b] Truncated in *O. sphegodes*; [c] triplicated gene; [d] partially duplicated in *O. iricolor*

Table 2. List of genes identified in the plastomes of *Ophrys iricolor* and *Ophrys sphegodes*

4.4.2 Repeat sequence detection

The occurrence, type, and distribution of SSRs in *Ophrys* plastomes were analysed. In total, 213 SSRs were identified in *O. sphegodes* and *O. iricolor.* Three of these microsatellites occurred in the sequence portion that is deleted in *O. sphegodes* plastome (S1 Table). Homopolymers and dipolymers were the most common SSRs with, respectively, 71% and 24% occurrence. Seven and nine SSRs were present in compound formation in *O. iricolor* and *O. sphegodes,* respectively. Furthermore, the majority of *O. sphegodes* and *O. iricolor* SSRs are located in IGS regions (56.2% and 55%), followed by coding sequences (38.2% and 38%) and introns (5.6% and 7%), respectively. SSRs located in coding regions were found mainly in *ycf1* and *rpoC2* genes. A comparison of SSRs found in the two *Ophrys* species showed that 67 SSRs (31.5% of the total) were polymorphic between the two species. Among these polymorphic SSRs, 46 were located in the IGS regions, 5 in introns and 16 in genes (S1 Table).

Ophrys sphegodes contains 15 directed repeats, 9 inverted repeats, 3 complementary repeats and 21 palindromic repeats, whose lengths range from 18 to 60 bp. *Ophrys iricolor* contains 15 directed repeats, 27 palindromic repeats, 2 complementary repeats and 5 inverted repeats, whose lengths range from 20 to 60 bp. Most of the *O. iricolor* and *O. sphegodes* repeats were located in IGS regions (65.3% and 66.7 % respectively), others were located in genes (22.4%, in *ycf*2, *pet*G, *ndh*C, *psa*A and 22.9%; in *psb*I, *ndh*C, *ycf*2, *ndh*A respectively) and introns (12.3% and 10.4% in *clp*P and *rps*16 intron respectively).

4.4.3 RNA editing sites prediction and positive signatures of adaptive evolution

The RNA editing is a post-transcriptional modification typical of plastid and mitochondrial DNA. The process originated early during the evolution of land plants and several RNA editing sites have been maintained or lost during angiosperms evolution [49, 50]. In our analysis, PREPACT found a total of 83 and 87 putative RNA editing sites in 25 genes in *O. sphegodes* and *O. iricolor* respectively (S2 Table), in line with previous report for other orchids [51]. The RNA editing sites predicted for plastid genes of *Ophrys sphegodes* and *Ophrys iricolor* occur in the first or second codon position with all nucleotide changes being from cytidine (C) to uridine (U), as very often reported in other angiosperms. In *O. sphegodes* the genes predicted to have RNA editing sites are *mat*K (12 sites), *rpo*C1 (9 sites), *rpo*C2 (8 sites), *rpo*B (8 sites), *acc*D (6 sites), *rpo*A (4 sites), *atp*A (4 sites), *rpl*2 (3 sites) *rpl*20 (3 sites), *atpI* (3 sites), *ccs*A (3 sites), *ycf*3 (3 sites), *clp*P (3 sites), *pet*B (2 sites), *rps*16 (2 sites) and the *atp*F, *pet*D, *pet*L, *psa*B, *psa*I, *psb*F, *rpl*23, *rps*2, *rps*8 and *rps*14 genes with only one site. In *O. iricolor* the genes predicted were the same as *O. sphegodes* with few differences: *ccs*A (5 sites), *atp*A (3 sites), *psa*B (2 sites), *rps*14 (2 sites), *atp*F (2 sites) (S2 Table) which suggest a general conservation of the RNA editing mechanism within *Ophrys* but also that RNA editing evolution accumulated enough differences to differentiate two *Ophrys* species. A previous study has also found that the number of RNA editing sites predicted for protein-coding genes in orchids species is high in comparison with other monocots [51]. Likelihood ratio test between a null model and an alternative model carried out following [52] shows that 24 genes are under positive selection (S3 Table); overall, the most divergent genes have the stronger signatures of positive selection (S2 Fig). In details, the positively selected genes were involved in different essential functions such as photosynthesis, PSII (*psb*A, *psb*B, *psb*E, *psb*H, *psb*M, *psbN* genes), large subunits of rubisco (*rbc*L), ATP synthase (*atp*I gene), cytochrome b6f (*pet*B gene), subunits of RNA polymerase (*rpo*A, *rpo*B, *rpo*C1, *rpo*C2 genes), RNA maturation (*mat*K gene), ribosomal proteins (*rpl*20, *rpl*22, *rpl*32, *rpl*33, *rps*12, *rps*19 genes), fatty acid biosynthesis (*acc*D gene), cytochrome biosynthesis (*ccs*A gene), import of protein in the plastid (*ycf*1 gene), and unknown function (*ycf*2 gene). The high number of genes containing positive signatures (including the *rbc*L gene) among photosynthesis-related genes are consistent with previous observation on other monocots and may be related to the recent increase of diversification rate following adaptation to different ecological conditions. [53]. In particular**,** and as already suggested for other monocots as Arecaceae, many tropical orchid species grow as epiphytes in tropical forests and are shade adapted. The transition to the terrestrial habitus of all temperate orchid lineages (as *Ophrys*) may have promoted a new selective pressure for improving the photosynthesis efficiency under the new terrestrial ecological conditions [52].

Interestingly, some positively selected sites that were identified in our study (e.g., the *acc*D and *ycf*1genes) have been found very variable also in other orchids and flowering plants [54]. In particular *acc*D gene is a conserved plastid gene involved in de novo synthesis of fatty acids [55] and is essential for chloroplast functionality, leaf development and longevity [56]. Therefore *acc*D has been associated in a significant manner with adaptation to the environment, including factors such as temperature, light, humidity, and atmosphere [57].

On the other hand, *ycf*1 is one of the largest plastid genes and it has been found extremely divergent in orchids plastomes [19], likely because of its position at IR/SC junction that generates large variation in sequence length and pseudogenes [58] as also found in our study.

4.4.4 Genomic comparison of *Ophrys* with other orchid plastomes

The *Ophrys* plastid genome is fully collinear both in gene order an gene orientation with the other available Orchidoideae. When compared with representative species belonging to the different subfamilies of the Orchidaceae (i.e., Epidendroideae, Cypripedioideae, Vanilloideae and Apostasioideae), we found that Cypripedioideae, Epidendroideae, Vanilloideae and Orchidoideae are largely collinear in plastid sequence with a few small exceptions: an inversion of the *psb*M - *pet*N gene order in Epidendroideae and a gene inversion in the SSC of *Vanilla* (S3 Fig).

In contrast to these four tribes, large rearrangements in gene order have been found in the supposed basal smaller tribe of Apostasioideae. However, under the assumption that the common plastid types observed in most orchids represent the primitive state, it is likely that the rearrangements found in *Apostasia wallichii* and *Apostasia odorata* (but not in the related *Neuwiedia* [59]) may be due to recent, terminal autoapomorphic changes rather than being representative of the ancestral gene order of the orchid family.

As many *ndh* genes had either truncations or indels, resulting in frameshifts or pseudogenes in several orchid plastomes, we also compared *ndh* genes in the different tribes. *Ophrys iricolor*, like other Orchidoideae, has the complete set of *ndh* plastid genes, i.e. 11 functional genes, which is different from *Apostasia wallichii* and *Vanilla planifolia* in which the *ndh*B gene is truncated and from *Vanilla planifolia* where all other 10 *ndh* subunits are deleted. The presence of *ndh* genes within terrestrial Orchidoideae is ubiquitous, which contrasts with the extensive variation in presence/absence of *ndh* genes found within tropical orchid genera (see *Cymbidium* [17]). The functional role of the *ndh* genes seems closely related to the land adaptation of photosynthesis so they have been conserved in terrestrial, temperate orchid plastomes whereas they are partially lost in epiphytic, tropical orchid plastomes [60].

4.4.5 Boundaries between single copy and inverted repeat regions

Expansion or contraction of the IR region is one of the main causes of size variation among angiosperm plastid genomes [61] and it has found to be variable even among related orchid species as, for instance, within the *Cymbidium* genus [17].

The multiple genome alignment analysis using plastome sequences of *O. sphegodes* and *O. iricolor* revealed the loss of a *ycf*1 fragment in the IR and partial deletion of the *ndh*F gene in *O. sphegodes* (S4 Fig). *In silico* validation confirmed the partial *ndh*F gene loss in *O. sphegodes* and demonstrated that part of the *ycf*1 gene is duplicated in *O. iricolor*, which does not occur in *O. sphegodes* (Fig 2).

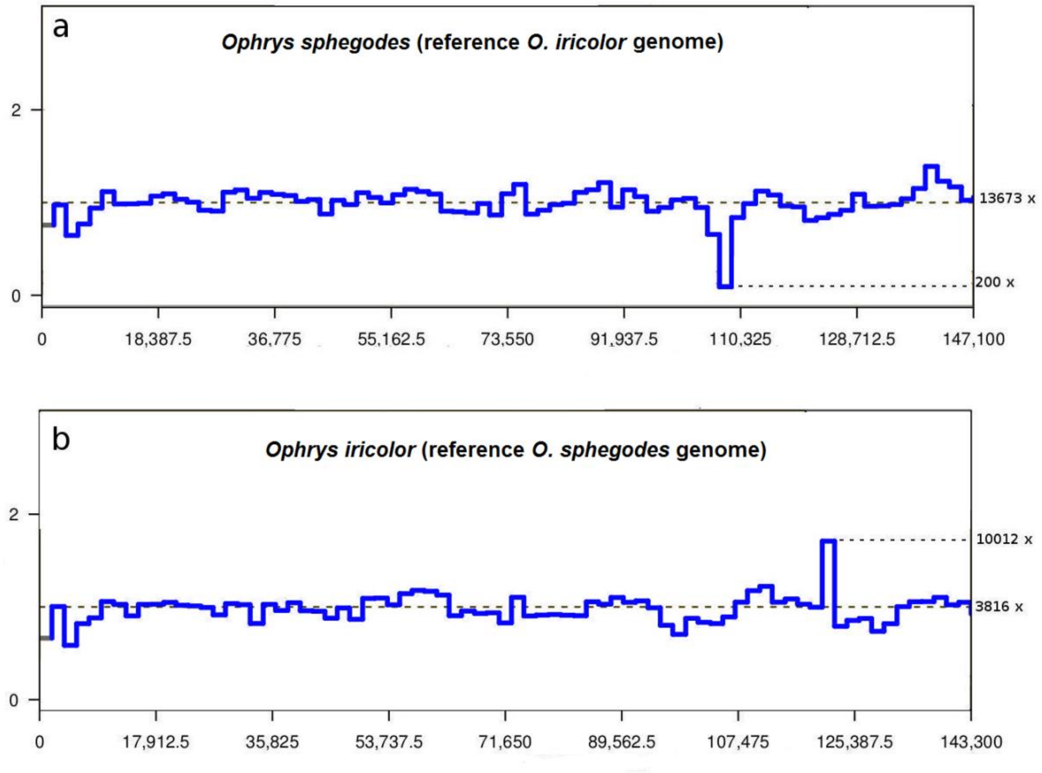


Figure 2. *In silico* validation of *ndh*F deletion (using software CNView) comparing *O. sphegodes* plastid reads against reference genome of *O. iricolor* (a) and *O. iricolor* plastid reads against reference genome of *O. sphegodes* (b). Y-axis represents normalized coverage values.

In *O. sphegodes*, the loss of the partial duplicated gene of *ycf*1 and the partial deletion of *ndh*F gene are correlated with the shift of the junction between the IR and SSC (Fig 1) with a pattern very similar to some *Cymbidium* species [17]. High sequence variability, especially in the *ycf*1 gene at IR-SSC junction, have been frequently observed as a result of expansion and contractions events by gene conversion [62, 63]. While *in silico* validation by CNVIEW largely confirms the occurrence of the *ndh*F deletion in *O. sphegodes*, however, approximately 2% of *O. sphegodes* reads map on the plastid region corresponding to *O. iricolor* plastome type (i.e. where complete *ndh*F occurs). At the same time, IGV also reveals that 888 of *O. iricolor* reads map on the junction with *ndh*F deletion (i.e. corresponding to *O. sphegodes* plastome type). Thus, to confirm the occurrence of *ndh*F deletion in *O. sphegodes*/*O. iricolor,* we amplified DNA with primers for both the flanking and internal regions of *ndh*F. Further, to rule out any possible cross contamination (during the NGS steps) as cause of presence of both plastome types in both *Ophrys* species, different accessions were used in PCR validation. PCR amplifications with primers flanking *ndh*F yielded two amplicons in *O. iricolor*: a small one (0.25 Kb), corresponding to the plastid fragment with the *ndh*F deletion, and a larger amplicon (3.25 Kb) containing the undeleted *ndh*F gene. Only the small plastid fragment with the *ndh*F deletion (primers F1/R1) was detected in *O. sphegodes.* To exclude, in *O. sphegodes*, that the small fragment was selectively amplified due to its shorter size and higher copy number, we also amplified *O. sphegodes* and *O. iricolor* (as control) with primers located within the *ndh*F deletion (primers F2/R2). Contrary to expectation (i.e. no amplification in *O. sphegodes*) both species successfully amplified a 1.2 Kb fragment. However, the two species differed in their amplicon yield, i.e. we obtained a stronger amplification band in *O. iricolor* compared to *O. sphegodes* (S1 Fig b). Taken together, this suggest that both species contained copies with and without the *ndh*F deletion but with a different relative representation (high proportion of deletions in *O. sphegodes* and low in *O. iricolor*). The fact that all examined members of *O. sphegodes* and *O. iricolor* lineages (including the basal *O. insectifera*) share a similar PCR amplification pattern suggests that the deletion of *ndh*F has likely occurred only once during the early evolution of the genus *Ophrys*, i.e. immediately before the separation of the two main lineages. The presence of two plastome types (with a different relative representation) across the two lineages represents an unusual case of maintenance of plastid heteroplasmy likely established as consequence of retention of ancestral polymorphism or of plastid capture by hybridization. Both processes have been commonly suggested to explain the unusual genomic admixture detected among *Ophrys* species as they are characterized by very rapid radiation and recurrent hybridization [64, 65].

4.4.6 Genomic localization of deleted *ndh*F gene in *O. sphegodes* nuclear genome

BLAST search of the assembly for the deleted *ndh*F region from the plastid genome of *O. sphegodes* found the nuclear scaffold1075174 (length 5,436 bp) with a score of 924 and e-value of 0.0. Reads of whole genome sequencing were mapped against scaffold1075174 to check whether some reads overlap with the junction between plastid deleted region and the remaining part of this scaffold. A total of 124,961 reads mapped on the scaffold. BLASTX search for the scaffold1075174 (after excluding the deleted plastid region) revealed the presence of a reverse transcriptase, a GAG pre-integrase domain, and the gag-polypeptide of LTR copia-type. Twelve reads map on the junction between *ndh*F and the reverse transcriptase so confirming the connection between the two parts. This result represents a clear indication that the deleted plastid region has been inserted in a retrotransposon nuclear sequence of *O. sphegodes* (Fig 3). Most of the repetitive DNA in available orchid genomes are gypsy- and copia-like retrotransposons [66] and their activity is likely to significantly contributed to the orchid large genome size [67].

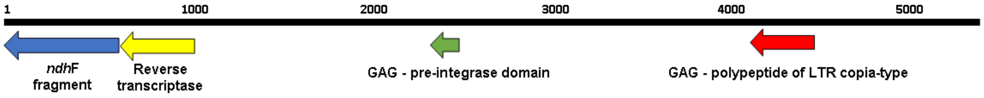

1
1000
2000
3000
4000
5000
*ndh*F fragment
Reverse transcriptase
GAG - pre-integrase domain
GAG - polypeptide of LTR copia-type

Figure 3. Results of BLASTX search of scaffold1075174 (length 5,436 bp): putative domain hits are indicated by the colored arrows

4.5 Conclusions

The complete plastid genomes provided here for two taxa from the rapidly evolving orchid genus *Ophrys* represents a source of novel information that can help resolve evolutionary questions. While the plastid gene order and organization reveal the signal of phylogenetic relationships among main species groups in this genus, the highly variable SSRs and tandem repeats with suitable level of intraspecific variation can be used as markers in phylogeographic and speciation studies among those closely related species. These relationships can now be explored with the novel genomic resources available today.

References

1. Raven JA, Allen JF (2003) Genomics and chloroplast evolution: what did cyanobacteria do for plants? Genome Biology 4: 209.

2. Palmer JD (1991) Plastid chromosomes: structure and evolution. In: Vasil LK, Bogorad L, editors. Cell Culture and Somatic Cell Genetics in Plants, the Molecular Biology of Plastid 7A. San Diego: Academic Press; pp. 5–53.

3. Downie SR, Palmer JD (1992) Use of chloroplast DNA rearrangements in reconstructing plant phylogeny. In: Soltis PS, Soltis DE, Doyle JJ, editors. Molecular Systematics of Plants. Springer US; pp. 14–35.

4. Judd WS, Campbell CS, Kellogg EA, Stevens PF, Donoghue MJ (2002) Plant systematics: a phylogenetic approach. 2nd ed. Sunderland, Massachusetts: Sinauer Associates.

5. Chaney L, Mangelson R, Ramaraj T, Jellen EN, Maughan PJ (2016) The complete chloroplast genome sequences for four *Amaranthus* species (Amaranthaceae). Applications in Plant Sciences 4.

6. Cho KS, Cheon KS, Hong SY, Cho JH, Im JS, Mekapogu M, et al. (2016) Complete chloroplast genome sequences of *Solanum commersonii* and its application to chloroplast genotype in somatic hybrids with *Solanum tuberosum*. Plant Cell Reports 35: 2113–2123.

7. Fu J, Liu H, Hu J, Liang Y, Liang J, Wuyun T, Tan X (2016) Five complete chloroplast genome sequences from diospyros: genome organization and comparative analysis. PloS One 11: e0159566.

8. Xu J-H, Liu Q, Hu W, Wang T, Xue Q, Messing J (2015) Dynamics of chloroplast genomes in green plants. Genomics 106: 221–231.

9. Wolfe K, Li W, Sharp P (1987) Rates of Nucleotide Substitution Vary Greatly among Plant Mitochondrial, Chloroplast, and Nuclear DNAs. Proceedings of the National Academy of Sciences USA 84: 9054–9058.

10. Shaw J, Lickey EB, Schilling EE, Small RL (2007) Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: the tortoise and the hare III. American Journal of Botany 94: 275–288.

11. Kang JS, Lee BY, Kwak M (2017) The complete chloroplast genome sequences of *Lychnis wilfordii* and *Silene capitata* and comparative analyses with other Caryophyllaceae genomes. PloS one 12: e0172924.

12. Jansen RK, Cai Z, Raubeson LA, Daniell H, Leebens-Mack J, Müller KF et al. (2007) Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proceedings of the National Academy of Sciences USA 104: 19369–19374.

13. Christenhusz, MJM, Byng JW (2016) The number of known plants species in the world and its annual increase. Phytotaxa 261: 201–217.

14. Barrett CF, Freudenstein JV, Li J, Mayfield-Jones DR, Perez L, Pires JC, Santos C (2014) Investigating the path of plastid genome degradation in an early-transitional clade of heterotrophic orchids, and implications for heterotrophic angiosperms. Molecular Biology and Evolution 31: 3095–3112.

15. Lin CS, Chen JJ, Huang YT, Chan MT, Daniell H, Chang WJ et al. (2015) The location and translocation of *ndh* genes of chloroplast origin in the Orchidaceae family. Scientific Reports 5: 9040.

16. Luo J, Hou BW, Niu ZT, Liu W, Xue QY, Ding XY (2014) Comparative chloroplast genomes of photosynthetic orchids: insights into evolution of the Orchidaceae and development of molecular markers for phylogenetic applications. PLoS One 9: e99016.

17. Kim HT, Chase MW (2017) Independent degradation in genes of the plastid ndh gene family in species of the orchid genus *Cymbidium* (Orchidaceae; Epidendroideae). PLoS One 12: e0187318.

18. Wu FH, Chan MT, Liao DC, Hsu CT, Lee YW, Daniell H et al. (2010) Complete chloroplast genome of *Oncidium* Gower Ramsey and evaluation of molecular markers for identification and breeding in Oncidiinae. BMC Plant Biology 10: 68.

19. Neubig KM, Whitten WM, Carlsward BS, Blanco MA, Endara L, Williams NH, Moore M (2009) Phylogenetic utility of ycf1 in orchids: a plastid gene more variable than matK. Plant Systematics and Evolution 277: 75–84.

20. Wang RJ, Cheng CL, Chang CC, Wu CL, Su TM, Chaw SM (2008) Dynamics and evolution of the inverted repeat-large single copy junctions in the chloroplast genomes of monocots. BMC Evolutionary Biology 8: 36.

21. Dong WL, Wang RN, Zhang NY, Fan WB, Fang MF, Li ZH (2018) Molecular Evolution of Chloroplast Genomes of Orchid Species: Insights into Phylogenetic Relationship and Adaptive Evolution. International Journal of Molecular Sciences 19: 716.

22. Devey DS, Bateman RM, Fay MF, Hawkins JA (2008) Friends or relatives? Phylogenetics and species delimitation in the controversial European orchid genus *Ophrys*. Annals of Botany 101: 385–402.

23. Breitkopf H, Onstein RE, Cafasso D, Schlüter PM, Cozzolino S (2015) Multiple shifts to different pollinators fuelled rapid diversification in sexually deceptive *Ophrys* orchids. New Phytologist 207: 377–389.

24. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina Sequence Data. Bioinformatics 30: 2114–2120.

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

26. Tillich M, Lehwark P, Pellizzer T, Ulbricht-Jones ES, Fischer A, Bock R, Greiner S (2017) GeSeq – versatile and accurate annotation of organelle genomes. *Nucleic Acids Research* 45: W6–W11.

27. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. Journal of Molecular Biology 215: 403–410.

28. de Santana Lopes A, Pacheco TG, do Nascimento Vieira L, Guerra MP, Nodari RO, de Souza EM et al. (2018) The *Crambe abyssinica* plastome: Brassicaceae phylogenomic analysis, evolution of RNA editing sites, hotspot and microsatellite characterization of the tribe Brassiceae. Gene 671: 36–49.

29. Lowe TM, Eddy SR (1997) tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Research 25: 955–964.

30. 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 Research 41: W575–W581.

31. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J et al. (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience 1: 18.

32. Li H, Durbin R (2009) The short read alignment component (bwa-short) has been published: Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics 25: 1754–1760.

33. 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 (SAM) format and SAMtools. Bioinformatics 25: 2078–2079.

34. Katoh K, Rozewicki J, Yamada KD (2017) MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Briefings in Bioinformatics bbx108.

35. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL (2004) Versatile and open software for comparing large genomes. *Genome Biology* 5: R12.

36. Thorvaldsdóttir H, Robinson JT, Mesirov JP (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration**.** Briefings in Bioinformatics 14: 178–192.

37. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26: 841–842.

38. Collins RL, Stone MR, Brand H, Glessner JT, Talkowski ME (2016) CNView: a visualization and annotation tool for copy number variation from whole-genome sequencing. bioRxiv 049536.

39. Cozzolino S, Cafasso D, Pellegrino G, Musacchio A, Widmer A (2003) Molecular evolution of a plastid tandem repeat locus in an orchid lineage. Journal of Molecular Evolution 57: S41–S49.

40. Thiel T, Michalek W, Varshney R, Graner A (2003) Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). Theoretical and Applied Genetics 106: 411–422.

41. de Santana Lopes A, Pacheco TG, dos Santos KG, do Nascimento Vieira L, Guerra MP, Nodari RO et al. (2018). The *Linum usitatissimum* L. plastome reveals atypical structural evolution, new editing sites, and the phylogenetic position of Linaceae within Malpighiales. Plant cell reports, 37: 307–328.

42. Kurtz S, Schleiermacher C (1999) REPuter: fast computation of maximal repeats in complete genomes. Bioinformatics 15: 426–427.

43. Lenz H, Rüdinger M, Volkmar U, Fischer S, Herres S, Grewe F, Knoop V (2010) Introducing the plant RNA editing prediction and analysis computer tool PREPACT and an update on RNA editing site nomenclature. Current Genetics 56: 189–201.

44. Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Research 30: 3059–3066.

45. Maddison WP, Maddison DR (2018) Mesquite: a modular system for evolutionary analysis. Version 3.40 http://mesquiteproject.org.

46. Lanfear R, Frandsen PB, Wright AM, Senfeld T, Calcott B (2016) PartitionFinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. Molecular Biology and Evolution 34: 772–773.

47. Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30: 1312–1313.

48. Stern A, Doron-Faigenboim A, Erez E, Martz E, Bacharach E, Pupko T (2007) Selecton 2007: advanced models for detecting positive and purifying selection using a Bayesian inference approach. Nucleic Acids Research 35: W506–W511.

49. Tillich M, Lehwark P, Morton BR, Maier UG (2006) The evolution of chloroplast RNA editing. Molecular Biology and Evolution 23: 1912–1921.

50. Takenaka M, Zehrmann A, Verbitskiy D, Härtel B, Brennicke A (2013) RNA editing in plants and its evolution. Annual Review of Genetics 47: 335–352.

51. Chen TC, Liu YC, Wang X, Wu CH, Huang CH, Chang CC (2017) Whole plastid transcriptomes reveal abundant RNA editing sites and differential editing status in *Phalaenopsis aphrodite* subsp. *formosana*. Botanical Studies 58: 38.

52. de Santana Lopes A, Gomes Pacheco G, Nimz T, do Nascimento Vieira L, Guerra MP, Nodari RO et al. (2018). The complete plastome of macaw palm [*Acrocomia aculeata* (Jacq.) Lodd. ex Mart.] and extensive molecular analyses of the evolution of plastid genes in Arecaceae. Planta 247: 1011–1030.

53. Piot A, Hackel J, Christin PA, Besnard G (2017) One-third of the plastid genes evolved under positive selection in PACMAD grasses. Planta 247: 255–266.

54. Givnish TJ, Spalink D, Ames M, Lyon SP, Hunter SJ, Zuluaga A et al. (2015) Orchid phylogenomics and multiple drivers of their extraordinary diversification. Proceedings of the Royal Society of London B 282: 2108–2111.

55. Feria Bourrellier AB, Valot B, Guillot A, Ambard-Bretteville F, Vidal J, Hodges M (2010) Chloroplast acetyl-CoA carboxylase activity is 2-oxoglutarateregulated by interaction of PII with the biotin carboxyl carrier subunit. Proceedings of the National Academy of Science U S A 107: 502–507.

56. Mizoi J, Nishida I, Nagano Y, Sasaki Y (2002) Chloroplast transformation with modified accD operon increases acetyl- CoA carboxylase and causes extension of leaf longevity and increase in seed yield in tobacco. Plant Cell Physiology 43: 1518–1525.

57. Hu S, Sablok G, Wang B, Qu D, Barbaro E, Viola R et al. (2015) Plastome organization and evolution of chloroplast genes in *Cardamine* species adapted to contrasting habitats. BMC Genomics 16: 306.

58. Jheng CF, Chen TC, Lin JY, Chen TC, Wu WL, Chang CC (2012) The comparative chloroplast genomic analysis of photosynthetic orchids and developing DNA markers to distinguish *Phalaenopsis* orchids. Plant Science 190: 62–73.

59. Niu Z, Pan J, Zhu S, Li L, Xue Q, Liu W, Ding X (2017) Comparative analysis of the complete plastomes of *Apostasia wallichii* and *Neuwiedia singapureana* (Apostasioideae) reveals different evolutionary dynamics of IR/SSC boundary among photosynthetic orchids. Frontiers in Plant Science 8: 1713.

60. Martín M, Sabater B (2010) Plastid ndh genes in plant evolution. Plant Physiology and Biochemistry 48: 636–645.

61. Ravi V, Khurana JP, Tyagi AK, Khurana P (2008) An update on chloroplast genomes. Plant Systematics and Evolution 271: 101–122.

62. Goulding SE, Olmstead RG, Morden CW, Wolfe KH (1996) Ebb and flow of the chloroplast inverted repeat. Molecular and General Genetics 252: 195–206.

63. Zhu A, Guo W, Gupta S, Fan W, Mower JP (2016) Evolutionary dynamics of the plastid inverted repeat: the effects of expansion, contraction, and loss on substitution rates. New Phytologist 209: 1747–1756.

64. Vereecken NJ, Streinzer M, Ayasse M, Spaethe J, Paulus HF, Stoekl J, et al. (2011) Integrating past and present studies on *Ophrys* pollination–a comment on Bradshaw et al. Botanical Journal of the Linnean Society 165: 329–335.

65. Sedeek KEM, Scopece G, Staedler YM, Schönenberger J, Cozzolino S, Schiestl FP, Schlüter PM (2014) Genic rather than genome-wide differences between sexually deceptive *Ophrys* orchids with different pollinators. Molecular Ecology 23: 6192–6205.

66. Hsu CC, Chung YL, Chen TC, Lee YL, Kuo YT, Tsai WC et al. (2011). An overview of the *Phalaenopsis* orchid genome through BAC end sequence analysis. BMC Plant Biology 11: 3.

67. Leitch IJ, Kahandawala I, Suda J, Hanson L, Ingrouille MJ, Chase MW, Fay MF (2009) Genome size diversity in orchids: consequences and evolution. Annals of Botany, 104: 469–481.

Supporting information

SSR nr.	**Gene**	**SSR *O. sphegodes***	**start-end *O. sphegodes***	**SSR *O. iricolor***	**start-end *O. iricolor***
1	IGS *psb*A - *mat*K	(A)11	1294-1304	(A)10	1294-1303
2	IGS *psb*A - *mat*K	(T)11A(T)10	1317-1338		
3	IGS *psb*A - *mat*K			(T)14	1317-1330
4	IGS *psb*A - *mat*K	(T)10	1600-1609	(T)17	1582-1598
5	*mat*K	(AT)4	2473-2480	(AT)4	2462-2469
6	*mat*K	(T)9	2782-2790	(T)9	2771-2779
7	*mat*K	(A)9	2935-2943	(A)9	2924-2932
8	IGS *mat*K - *rps*16	(T)10	3352-3361	(T)11	3349-3359
9	IGS *mat*K - *rps*16	(A)10	3676-3685	(A)11	3683-3693
10	IGS *mat*K - *rps*16	(A)13	3890-3902		
11	IGS *mat*K - *rps*16	(T)11	3967-3977	(T)14	3975-3988
12	IGS *mat*K - *rps*16	(T)8	3987-3994	(T)8	3998-4005
13	IGS *mat*K - *rps*16	(A)10(AAT)4	4163-4182	(AAT)6	4179-4196
14	IGS *mat*K - *rps*16	(A)12	4246-4257	(A)9	4260-4268
15	IGS *mat*K - *rps*16	(T)12	4287-4298	(T)10	4295-4304
16	IGS *mat*K - *rps*16	(G)12	4583-4594		
17	IGS *mat*K - *rps*16			(C)10(T)8	4574-4591
18	IGS *mat*K - *rps*16	(A)9	4614-4622	(A)9	4623-4631
19	IGS *mat*K - *rps*16	(T)10	4715-4724		
20	*rps*16			(A)11	5187-5197
21	*rps*16	(T)8	5245-5252	(T)10	5254-5263
22	*rps*16	(TA)6	5546-5557	(TA)5	5558-5567
23	IGS *rps*16 - *trn*Q-UUG	(T)11	6052-6062	(T)10	6062-6071

24	IGS *rps*16 - *trn*Q-UUG	(A)9	6072-6080	(A)9	6081-6089
25	IGS *trn*Q-UUG - *psb*K	(A)10	6497-6506	(A)11	6506-6516
26	IGS *trn*Q-UUG - *psb*K	(A)18	6595-6612	(A)13	6605-6617
27	*psb*K	(T)9	6756-6764	(T)9	6761-6769
28	IGS *psb*K - *psb*I	(AT)4	6814-6821	(AT)4	6819-6826
29	IGS *psb*K - *psb*I	(GT)4	6933-6940	(GT)4	6940-6947
30	IGS *psb*K - *psb*I	(A)13	6965-6977	(A)13	6972-6984
31	*trn*S-GCU	(GA)4	7243-7250	(GA)4	7250-7257
32	IGS *trn*S-GCU - *trn*R-UCU	(TA)6	7596-7607	(TA)6	7596-7607
33	IGS *trn*S-GCU - *trn*R-UCU	(T)12	7738-7749	(T)11	7742-7752
34	IGS *trn*S-GCU - *trn*R-UCU	(T)9	7948-7956	(T)11	7951-7961
35	IGS *trn*S-GCU - *trn*R-UCU	(A)8	8074-8081	(A)10	8079-8088
36	IGS *trn*S-GCU - *trn*R-UCU	(T)8	8219-8226	(T)8	8226-8233
37	*atp*A	(GTCT)3	9969-9980	(GTCT)3	9969-9980
38	*atp*F	(T)8	11330-11337	(T)10	11330-11339
39	*atp*F	(A)9	11395-11403	(A)8	11396-11403
40	*atp*F	(A)8	11832-11839	(A)8	11833-11840
41	IGS *atp*H - *atp*I	(A)8	12331-12338	(A)8	12332-12339
42	IGS *atp*H - *atp*I	(A)10	12453-12462	(A)9	12454-12462
43	IGS *atp*H - *atp*I			(AT)4	12683-12690
44	IGS *rps*2 - *rpo*C2			(T)11	14517-14527
45	IGS *rps*2 - *rpo*C2	(A)10	14662-14671	(A)10	14686-14695
46	*rpo*C2	(T)10	16545-16554	(T)10	16559-16568
47	*rpo*C2	(T)10	16612-16621	(T)10	16626-16635
48	*rpo*C2	(T)11	16718-16728	(T)11	16732-16742
49	*rpo*C2	(A)9	16866-16874	(A)9	16880-16888
50	*rpo*C2	(AT)4	18016-18023	(AT)4	18030-18037
51	*rpo*C2	(AT)5	18106-18115	(AT)5	18120-18129
52	*rpo*C1	(A)8	20568-20575	(A)8	20582-20589

53	*rpo*C1	(T)8	21761-21768	(T)8	21775-21782
54	*rpo*B	(T)8	24435-24442	(T)8	24449-24456
55	*rpo*B	(T)9	25495-25503	(T)10	25501-25510
56	IGS *rpo*B - *trn*C-GCA	(A)13	26090-26102	(A)10	26146-26155
57	IGS *pet*N - *psb*M	(A)14	27078-27091	(A)11	27131-27141
58	IGS *pet*N - *psb*M	(TA)4	27130-27137	(TA)4	27180-27187
59	IGS *pet*N - *psb*M	(AT)4	27239-27246	(AT)4	27289-27296
60	IGS *pet*N - *psb*M	(TA)5	27403-27412	(AT)6	27399-27410
61	IGS *pet*N - *psb*M			(TA)5	27461-27470
62	IGS *psb*M - *trn*D-GUC			(T)11	28192-28202
63	IGS *psb*M - *trn*D-GUC	(T)9(TTTA)3*	28126-28143		
64	IGS *trn*E-UUC – *trn*T-GGU	(TA)4	29157-29164	(TA)4	29215-29222
65	IGS *trn*E-UUC – *trn*T-GGU	(A)10	29477-29486	(A)12	29531-29542
66	IGS *trn*E-UUC – *trn*T-GGU	(AT)5	29736-29745	(AT)5	29778-29787
67	IGS *trn*T-GGU - *psb*D	(A)8	29939-29946	(A)8	29981-29988
68	IGS *trn*T-GGU – *psb*D	(TAAA)3	30195-30206	(TAAA)3	30237-30248
69	IGS *psb*B - *trn*S-UGA	(TA)4(T)9	33202-33218	(TA)10(T)10	33211-33240
70	IGS *psb*Z - *trn*G-GCC	(AT)4	33944-33951	(AT)4	33966-33973
71	IGS *trn*G-GCC - *trn*M-CAU	(AT)4	34234-34241	(AT)4	34256-34263
72	IGS *trn*G-GCC - *trn*M-CAU	(A)9	34270-34278	(A)10	34292-34301
73	IGS trnG-GCC - *trn*M-CAU			(A)9	34322-34330
74	IGS *trn*M-CAU - *rps1*4	(AT)4	34449-34456	(AT)4	34474-34481
75	IGS *rps*14 - *psa*B	(A)12	34976-34987	(A)10	35001-35010
76	IGS *psb*30 - *ycf*3	(A)8	39724-39731	(A)8	39747-39754
77	IGS *psb*30 - *ycf*3	(T)11	39869-39879	(T)10	40596-40605
78	*ycf*3	(T)9	40571-40579	(T)12	41894-41905
79	*ycf*3	(T)11	41868-41878		
80	*trn*S-GGA	(CT)4	42468-42475	(CT)4	42495-42502
81	IGS *rps*4 - *trn*T-UGU	(TA)4	43494-43501	(TA)4	43521-43528

82	IGS *trn*T-UGU - *trn*F-GAA	(AT)4G(A)9	43912-43929	(AT)4G(A)10	43912-43930
83	IGS *trn*T-UGU - *trn*F-GAA	(T)13	44169-44181	(T)10	44060-44069
84	IGS *trn*T-UGU - *trn*F-GAA	(A)12	44235-44246	(A)16	44307-44322
85	IGS *trn*T-UGU - *trn*F-GAA	(T)13AT(A)13	44679-44706	(T)10AT(A)13	44748-44772
86	IGS *trn*T-UGU - *trn*F-GAA	(A)10	44802-44811	(A)9	44866-44874
87	IGS trnT-UGU - *trn*F-GAA			(TTATA)3	45073-45087
88	IGS *trn*T-UGU - *trn*F- GAA	(AG)4	45173-45180	(AG)4	45234-45241
89	IGS *trn*F-GAA - *ndh*J	(AT)4	45638-45645	(AT)4	45692-45699
90	IGS *trn*F-GAA - *ndh*J	(T)10	45873-45882	(T)12	45923-45934
91	IGS *ndh*C - *trn*M-CAU	(A)11	48357-48367	(A)11	48416-48426
92	IGS *ndh*C - *trn*M-CAU	(T)10	48510-48519	(T)13	48569-48581
93	IGS *ndh*C - *trn*M-CAU	(TA)4	48549-48556	(TA)4	48606-48613
94	IGS *atp*B - *rbc*L	(T)9	51739-51747	(T)9	51787-51795
95	IGS *atp*B - *rbc*L			(T)8	51934-51941
96	IGS *atp*B - *rbc*L	(AT)4	51983-51990	(AT)4	52096-52103
97	IGS *rbc*L - *acc*D	(T)10	53711-53720	(T)9	53816-53824
98	IGS *rbc*L - *acc*D	(T)8	54150-54157	(T)10	54254-54263
99	IGS *rbc*L - *acc*D	(A)11	54265-54275	(A)10	54371-54380
100	*acc*D	(T)9(TAA)4*	54708-54727	(T)9(TAA)4*	54813-54832
101	*acc*D	(TG)4	55397-55404	(TG)4	55502-55509
102	IGS *acc*D - *psa*I	(CT)4	56003-56010	(CT)4	56109-56116
103	IGS *acc*D - *psa*I	(A)11	56020-56030		
104	IGS *acc*D - *psa*I	(TA)4	56207-56214	(TA)4	56310-56317
105	IGS *psa*I - *ycf*4	(AT)4	56508-56515		
106	*ycf*4	(T)8	56906-56913	(T)8	56926-56933
107	*cem*A	(A)8	57866-57873	(A)8	57879-57886
108	*cem*A	(AATG)3	58531-58542	(AATG)3	58544-58555
109	*pet*A	(C)8	59122-59129	(C)8	59135-59142
110	IGS *psb*E - *pet*L	(T)8	61622-61629	(T)8	61631-61638

111	IGS *psb*E - *pet*L	(AT)4	61732-61739	(AT)4	61741-61748
112	IGS *pet*G - *trn*W-CCA	(T)8	62184-62191	(T)8	62193-62200
113	IGS *pet*G - *trn*W-CCA			(A)8	62229-62236
114	IGS *trn*P - UGG - *psa*J	(TA)5	62662-62671	(TA)5	62671-62680
115	IGS *psa*J - *rpl*33	(T)10	63160-63169	(T)9	63169-63177
116	IGS psaJ - *rpl*33			(T)11	63544-63554
117	IGS *psa*J - *rpl*33	(A)8	63330-63337	(A)12	63593-63604
118	IGS *psa*J - *rpl*33	(T)12	63520-63531		
119	IGS *psa*J - *rpl*33	(A)10	63547-63556		
120	IGS *rpl*33 - *rps*18	(AT)4	63856-63863	(AT)4	63904-63911
121	IGS *rpl*33 - *rps*18	(A)8	63905-63912	(A)10	63953-63962
122	IGS *rpl*33 - *rps*18	(TA)4	63926-63933		
123	IGS *rpl*33 - *rps*18			(TA)4TTT(TA)4	63976-63994
124	*rps*18	(AAAT)3	64245-64256	(AAAT)3	64306-64317
125	*rpl*20	(T)8	64836-64843	(T)8	64897-64904
126	IGS *rpl*20 - *clp*P	(T)14	65588-65601	(T)12	65649-65660
127	(intron) *clp*P	(A)12	66224-66235	(A)10	66283-66292
128	(intron) *clp*P	(A)9	66394-66402	(A)10	66451-66460
129	(intron) *clp*P	(A)10	67175-67184	(A)9	67233-67241
130	(intron) *clp*P	(T)15	67212-67226	(T)11	67269-67279
131	(intron) *clp*P	(A)9	67266-67274	(A)9	67322-67330
132	(intron) *clp*P	(T)10	67461-67470	(T)10	67517-67526
133	(intron) *clp*P	(T)8	67719-67726	(T)10	67781-67790
134	IGS *clp*P - *psb*B	(AT)4	68210-68217	(AT)4	68274-68281
135	IGS *clp*P - *psb*B	(TA)4	68293-68300	(TA)4	68365-68372
136	*psb*B	(T)8	69336-69343	(T)8	69408-69415
137	IGS *psb*C - *psb*T	(TA)4	70405-70412	(TA)4	70477-70484
138	IGS *psb*C - *psb*T	(T)8	70589-70596	(T)8	70661-70668

139	*psb*N	(T)10	70832-70841	(T)9	70903-70911
140	IGS *psb*H - *pet*B	(A)8	71527-71534	(A)8	71265-71272
141	IGS *psb*H - *pet*B	(A)10	71759-71768	(A)8	71597-71604
142	(Intron) *pet*B			(A)11	71828-71838
143	IGS *pet*B - *pet*D	(AT)4	72716-72723	(AT)4	72786-72793
144	(intron) *pet*D	(T)9	73144-73152	(T)8	73214-73221
145	IGS *pet*D - *rpo*A	(A)12	74338-74349		
146	*rps*11	(T)8	75468-75475	(T)9	76059-76067
147	*rps*8	(T)8	76766-76773	(T)11	76247-76257
148	IGS *rps*8 - *rpl*14	(T)11	77038-77048	(T)8	76828-76835
149	IGS *rps*8 - *rpl*14	(T)11	77055-77065	(T)8	76878-76885
150	IGS *rps*8 - *rpl*14	(T)10	77651-77660	(T)10	77117-77126
151	*rpl*14	(T)9	78241-78249	(T)10	77722-77731
152	(intron) rpl16			(T)8	78312-78319
153	(Intron) *rpl*16	(TA)4	78279-78286	(TA)4	78349-78356
154	(Intron) *rpl*16	(T)9	78807-78815	(T)10	78877-78886
155	(intron) *rpl*16	(T)9	79016-79024	(T)10	79087-79096
156	(intron) *rpl*16	(T)10	79179-79188	(T)9	79258-79266
157	IGS *rpl*16 - *rps*3	(T)9AT(A)12	79317-79339	(T)10AT(A)10	79395-79416
158	*rps*3	(TA)4	79829-79836	(TA)4	79899-79906
159	*rps*3	(T)10	79911-79920	(T)10	79981-79990
160	IGS *rps*3 - *rpl*22	(T)9	80072-80080	(T)9	80142-80150
161	*rpl*22	(T)8	80186-80193	(T)8	80256-80263
162	IGS *rpl*22 - *rps*19	(AAAAT)3	80607-80621		
163	*rps*19	(T)9	80961-80969	(T)9	81043-81051
164	IGS *rps*19 - *rpl*2	(T)11	80997-81007	(T)11	81079-81089
165	*ycf*2	(GA)4	84329-84336	(GA)4	84402-84409
166	*ycf*2	(A)8	85210-85217	(A)8	85283-85290
167	*ycf*2	(A)9	86451-86459	(A)9	86524-86532

168	*ycf*2	(GA)5	86472-86481	(GA)5	86545-86554
169	IGS *trn*L-CAA - *ndh*B	(TA)5	91320-91329	(TA)5	91449-91458
170	*ndh*B	(AG)4	92003-92010	(AG)4	92182-92189
171	IGS *rps*12 - *trn*V-GAC	(T)12	96643-96654	(T)11	96860-96870
172	IGS *rps*12 - *trn*V-GAC	(A)8	96694-96701	(A)8	96910-96917
173	*rrn*23	(CT)4	102777-102784	(CT)4	103001-103008
174	IGS *trn*N - GUU - *ndh*F	(A)12	105847-105858		
175	IGS *trn*N - GUU - *ndh*F	(A)11	105959-105969		
176	*IGS trn*N-GUU - *ndhF*	(T)8	106334-106341		
177	*ndh*F			(A)10	107254-107263
178	*ndh*F			(T)8	107622-107629
179	*ndh*F			(A)9	107843-107851
180	IGS *ndh*F - *rpl*32			(T)12	109056-109067
181	*rpl*32	(T)9	106421-106429		
182	IGS *ndh*F - *rpl*32			(A)9	109108-109116
183	IGS *ndh*F - *rpl*32			(A)9	109218-109226
184	IGS *rpl*32 - *trn*L-UAG	(A)10	106551-106560		
185	IGS *rpl*32 - *trn*L-UAG	(A)11	106574-106584	(A)11	109778-109788
186	*rpl*32			(T)9	109657-109665
187	IGS *rpl*32 - *trn*L-UAG			(A)8	109828-109835
188	IGS *trn*L-UAG - *ccs*A	(AT)4	106944-106951		
189	*ccs*A	(T)9	107530-107538		
190	IGS *ccs*A - *ndh*D	(T)9	108151-108159	(T)9	111364-111372
191	IGS *psa*C - *ndh*E	(TTTA)3	110381-110392		
192	*psa*C	(TTGA)3	110443-110454	(TTGA)3	113678-113689
193	*ndh*G	(A)11	111125-111135	(A)10	114348-114357
194	IGS *ndh*G - *ndh*I	(A)8	111547-111554	(A)8	114774-114781
195	IGS *ndh*I - *ndh*A			(TTA)4	115384-115395
196	*ndh*I	(T)12	111811-111822		

197	(intron) *ndh*A			(T)10	116752-116761
198	*ndh*A	(T)11	113497-113507		
199	*ndh*A	(A)13	113582-113594	(A)17	116833-116849
200	*ndh*A			(ATTT)3	116875-116886
201	IGS *rps*15 - *ycf*1	(A)15	115603-115617	(A)11	118859-118869
202	*ycf*1	(AT)4	116122-116129	(AT)4	119364-119371
203	*ycf*1	(T)12	117092-117103	(T)12	120331-120342
204	*ycf*1	(T)10	117237-117246	(T)11	120476-120486
205	*ycf*1	(T)8	117653-117660	(T)8	120883-120890
206	*ycf*1	(T)9	117879-117887	(T)9	121109-121117
207	*ycf*1	(A)10	118286-118295	(A)10	121528-121537
208	*ycf*1	(T)8	118349-118356	(T)8	121591-121598
209	*ycf*1	(AT)4(T)12*	118775-118793		
210	*ycf*1			(T)10	122022-122031
211	*ycf*1	(T)8	118828-118835	(T)8	122070-122077
212	*ycf*1	(T)12	119079-119090	(T)11	122321-122331
213	*ycf*1	(T)8	119450-119457	(T)8	122692-122699
214	*ycf*1	(A)9	119550-119558	(A)9	122792-122800
215	*ycf*1	(T)15	120185-120199	(T)15	123427-123441
216	*ycf*1	(T)9	120217-120225	(T)10	123458-123467
217	*ycf*1	(T)10	120306-120315	(T)10	123548-123557
218	*ycf*1	(A)12	120461-120472	(A)12	123703-123714
219	*ycf*1	(A)9	120555-120563	(A)9	123797-123805
220	*rrn*23	(AG)4	124442-124449	(AG)4	127711-127718
221	IGS *trn*V-GAC - *rps*12	(T)8	130525-130532	(T)8	133802-133809
222	IGS *trn*V-GAC - *rps*12	(A)12	130572-130583	(A)11	133849-133859
223	*ndh*B	(CT)4	135216-135223	(CT)4	138530-138537
224	IGS *ndh*B - *trn*L-CAA	(TA)5	135897-135906	(TA)5	139261-139270
225	*ycf*2	(TC)5	140745-140754	(TC)5	144165-144174

226	*ycf*2	(T)9	140767-140775	(T)9	144187-144195
227	*ycf*2	(T)8	142009-142016	(T)8	145429-145436
228	*ycf*2	(TC)4	142890-142897	(TC)4	146310-146317
229	IGS *trn*H-GUG – *rps*19	(A)11	146219-146229	(A)11	149630-149640
230	IGS *trn*H-GUG - *rps*19	(A)9	146257-146265	(A)9	149668-149676
231	*rps*19	(TATTT)3	146604-146618		

S1 Table. Distribution of simple sequence repeat (SSR) in *Ophrys sphegodes* and *O. iricolor* plastid genomes. IGS: intergenic spacer.

Gene	Nt pos.	AA pos.	*O. sphegodes*	*O. iricolor*	AA change
*acc*D	748	250	CAC→UAC	CAC→UAC	H→Y
	1184	395	UCA→UUA	UCA→UUA	S→L
	1306	436	CAC→UAC	CAC→UAC	H→Y
	1370	57	UCA→UUA	UCA→UUA	S→L
	1376	459	GCG→GUG	GCG→GUG	A→V
	1412	471	CCA→CUA	CCA→CUA	P→L
*atp*A	773	258	UCA→UUA	UCA→UUA	S→L
	914	305	UCA→UUA	-	S→L
	1148	383	UCA→UUA	UCA→UUA	S→L
	1493	498	ACC→AUC	ACC→AUC	T→I
atpB	-	-	-	-	-
*atp*F	92	31	CCA→CUA	CCA→CUA	P→L
	248	83	-	GCU→GUU	A→V
*atp*I	428	143	CCC→CUC	CCC→CUC	P→L
	437	146	GCG→GUG	GCG→GUG	A→V
	629	210	UCA→UUA	UCA→UUA	S→L
*ccs*A	122	41	UCA→UUA	UCA→UUA	S→L
	266	89	CCG→CUG	CCG→CUG	P→L
	280	94	CAU→UAU	CAU→UAU	H→Y
	511	171	-	CUU→UUU	L→F
	553	185	-	CUU→UUU	L→F
*clp*P	82	28	CAU→UAU	CAU→UAU	H→Y
	263	88	UCA→UUA	UCA→UUA	S→L
	559	187	CAU→UAU	CAU→UAU	H→Y
matK	331	111	CCA→UCA	CCA→UCA	P→S
	472	158	CAU→UAU	CAU→UAU	H→Y
	656	219	UCU→UUU	UCU→UUU	S→F
	722	241	ACA→AUA	ACA→AUA	T→I
	872	291	GCU→GUU	GCU→GUU	A→V
	913	305	CAU→UAU	CAU→UAU	H→Y
	916	306	CUU→UUU	CUU→UUU	L→F
	953	318	UCU→UUU	UCU→UUU	S→F
	1124	375	UCU→UUU	UCU→UUU	S→F

	1186	396	CCA→UCA	CCA→UCA	P→S
	1261	421	CAC→UAC	CAC→UAC	H→Y
	1460	487	CCU→CUU	CCU→CUU	P→L
*pet*B	418	140	CGG→UGG	CGG→UGG	R→W
	611	204	CCA→CUA	CCA→CUA	P→L
*pet*D	416	139	GCA→GUA	GCA→GUA	A→V
*pet*G	-	-	-	-	-
*pet*L	5	2	CCU→CUU	CCU→CUU	P→L
*psa*B	680	227	-	ACG→AUG	T→M
	2132	711	GCC→GUC	GCC→GUC	A→V
*psa*I	80	27	UCU→UUU	UCU→UUU	S→F
*psb*B	-	-	-	-	-
*psb*E	-	-	-	-	-
*psb*F	77	26	UCU→UUU	UCU→UUU	S→F
*psb*L	-	-	-	-	-
*rpl*2	2	1	ACG→AUG	ACG→AUG	T→M
	31	11	CCG→UCG	CCG→UCG	P→S
	217	73	CCU→UCU	CCU→UCU	P→S
*rpl*20	241	81	CUC→UUC	CUC→UUC	L→F
	287	96	ACA→AUA	ACA→AUA	T→I
	352	118	CAA→UAA	CAA→UAA	Q→*
*rpl*23	71	24	UCU→UUU	UCU→UUU	S→F
*rpo*A	200	67	UCU→UUU	UCU→UUU	S→F
	368	123	UCA→UUA	UCA→UUA	S→L
	778	260	CUU→UUU	CUU→UUU	L→F
	830	277	UCA→UUA	UCA→UUA	S→L
*rpo*B	29	10	UCC→UUC	UCC→UUC	S→F
	179	60	GCA→GUA	GCA→GUA	A→V
	338	113	UCU→UUU	UCU→UUU	S→F
	551	184	UCA→UUA	UCA→UUA	S→L
	623	208	CCG→CUG	CCG→CUG	P→L
	1736	579	GCC→GUC	GCC→GUC	A→V
	1747	583	CGC→UGC	CGC→UGC	R→C
	2426	809	UCA→UUA	UCA→UUA	S→L
*rpo*C1	41	14	CCA→CUA	CCA→CUA	P→L
	182	61	UCC→UUC	UCC→UUC	S→F
	257	86	UCU→UUU	UCU→UUU	S→F
	488	163	UCA→UUA	UCA→UUA	S→L
	617	206	UCG→UUG	UCG→UUG	S→L
	787	263	CGG→UGG	CGG→UGG	R→W
	1622	541	GCA→GUA	GCA→GUA	A→V
	1742	581	CCG→CUG	CCG→CUG	P→L
	1948	650	CGU→UGU	CGU→UGU	R→C
*rpo*C2	767	256	CCA→CUA	CCA→CUA	P→L
	1628	543	ACC→AUC	ACC→AUC	T→I
	1970	657	ACG→AUG	ACG→AUG	T→M
	2078	693	GCU→GUU	GCU→GUU	A→V
	2596	866	CGU→UGU	CGU→UGU	R→C
	3011	1004	UCA→UUA	UCA→UUA	S→L
	3725	1242	UCA→UUA	UCA→UUA	S→L
	4016	1339	ACU→AUU	ACU→AUU	T→I

*rps*2	134	45	ACA→AUA	ACA→AUA	T→I
*rps*8	182	61	UCA→UUA	UCA→UUA	S→L
*rps*14	80	27	UCA→UUA	UCA→UUA	S→L
	149	50	-	CCA→CUA	P→L
*rps*16	143	8	UCA→UUA	UCA→UUA	S→L
	202	68	CAU→UAU	CAU→UAU	H→Y
*ycf*3	44	15	UCU→UUU	UCU→UUU	S→F
	185	62	ACG→AUG	ACG→AUG	T→M
	191	64	CCA→CUA	CCA→CUA	P→L

S2 Table. List of RNA editing sites predicted in protein-coding genes of *Ophrys* plastomes using PREPACT program. High dashes indicate absence of RNA editing, * stop codon.

Gene	Null	Positive	Putative sites under positive selection *
***acc*D**	-9228,01	-9205,28	95 (1 M, 2 E, 4 C, 5 W, 8 L, 9 M, 10 L, 11 S, 12 N, 13 K, 18 R, 20 G, 25 K, 30 A, 32 A, 36 T, 44 L, 47 A, 48 E, 50 K, 52 P, 54 W, 55 G, 56 S, 57 Y, 59 L, 63 H, 65 L, 67 S, 68 F, 71 S, 75 W, 86 R, 95 V, 100 E, 102 Q, 113 L, 121 F, 122 N, 124 N, 126 S, 127 G, 129 L, 142 R, 145 P, 149 F, 152 T, 155 R, 159 E, 167 Y, 169 G, 170 I, 171 E, 172 N, 173 Y, 175 T, 180 A, 183 I, 188 D, 189 E, 191 L, 194 S, 196 S, 197 F, 199 R, 200 R, 201 E, 202 I, 206 F, 208 I, 221 E, 222 T, 229 R, 230 S, 237 H, 253 F, 257 G, 269 M, 282 R, 286 Y, 290 I, 302 I, 304 R, 309 P, 318 Q, 374 S, 379 S, 380 N, 386 V, 436 S, 438 A, 442 L, 456 L, 475 Q, 476 G)
***atp*I**	-1868,92	-1864,95	12 (8 I, 9 K, 26 L, 29 Q, 54 V, 63 T, 67 D, 80 R, 143 P, 146 A, 154 S, 162 G)
***ccs*A**	-4187,25	-4177,65	37 (4 V, 5 T, 23 M, 32 Y, 38 R, 42 E, 50 L, 87 R, 89 P, 91 L, 92 G, 94 H, 106 A, 113 A, 125 A, 134 Q, 149 A, 166 R, 168 N, 170 N, 171 L, 173 L, 175 N, 176 K, 179 F, 184 F, 193 F, 197 G, 199 V, 202 K, 206 L, 209 Q, 219 R, 280 S, 282 L, 286 G, 298 I)
***mat*K**	-5982,85	-5970,71	74 (6 L, 11 V, 13 Y, 16 Q, 18 I, 20 W, 21 G, 25 Y, 29 H, 31 Y, 35 L, 49 S, 50 Q, 52 V, 54 S, 57 E, 63 S, 64 Q, 81 S, 97 M, 119 L, 120 I, 127 L, 130 V, 136 D, 139 V, 143 R, 144 N, 146 R, 149 W, 159 C, 162 I, 164 L, 169 Q, 186 H, 189 V, 193 Y, 195 R, 203 Q, 204 R, 209 L, 212 P, 223 A, 226 A, 235 N, 238 K, 241 L, 245 W, 264 P, 271 M, 278 L, 281 T, 298 M, 299 T, 300 K, 306 V, 319 Q, 324 L, 342 L, 344 R, 363 Q, 398 G, 402 E, 405 M, 410 A, 413 L, 419 R, 420 P, 422 P, 425 G, 426 L, 438 R, 439 I, 445 H)
***pet*B**	-1772,77	-1759,86	6 (1 I, 2 N, 123 I, 140 R, 163 S, 204 P)
***psb*A**	-2650,24	-2640,4	1 (346 V)
***psb*B**	-4099,82	-4085,44	5 (296 L, 345 F, 352 R, 494 T, 504 R)
***psb*E**	-570,09	-565,51	2 (59 S, 78 E)
***psb*H**	-660,86	-651,56	5 (5 T, 16 R, 18 G, 45 V, 72 M)
***psb*M**	-246,99	-243,32	9 (6 L, 8 L, 12 A, 13 L, 24 I, 26 Y, 30 A, 33 N, 34 N)
***psb*N**	-275,24	-270,24	1 (32 Q)
***rbc*L**	-3913,24	-3891,32	14 (89 A, 142 P, 219 L, 225 L, 226 Y, 251 M, 375 L, 443 K, 449 S, 461

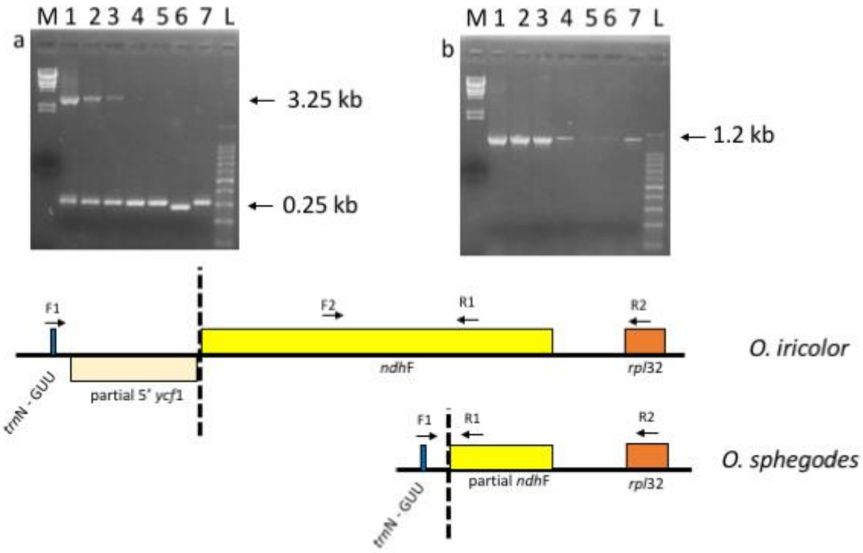
			I, 470 D, 471 P, 475 L, 477 K)
_rpl_20	-1609,01	-1603,49	7 (75 F, 80 K, 81 L, 83 H, 112 F, 116 I, 118 Q)
_rpl_22	-1487,53	-1479,43	16 (7 S, 8 E, 10 S, 22 R, 26 F, 53 R, 71 N, 91 A, 93 M, 96 L, 98 P, 106 M, 110 T, 120 E, 122 S, 124 I)
_rpl_32	-825,55	-819,47	8 (19 L, 40 T, 42 Q, 49 R, 53 V, 54 L, 55 E, 57 S)
_rpl_33	-866,07	-861,89	9 (20 R, 23 V, 26 E, 27 S, 28 T, 45 R, 48 L, 49 K, 53 R)
_rpo_A	-3546,32	-3534,9	42 (9 S, 14 Q, 25 K, 33 V, 34 M, 61 C, 64 C, 66 T, 71 L, 97 R, 105 D, 115 Y, 145 L, 146 C, 152 N, 154 D, 158 R, 164 N, 166 H, 167 D, 182 V, 190 G, 201 L, 237 M, 239 F, 240 E, 246 W, 250 P, 256 R, 260 L, 261 K, 266 G, 282 R, 283 T, 301 Y, 307 M, 309 M, 310 E, 311 Y, 313 C, 319 H, 322 S)
_rpo_B	-8918,47	-8910,72	56 (2 L, 5 V, 24 C, 32 A, 36 Q, 54 V, 58 Q, 63 L, 85 V, 101 L, 158 L, 189 L, 212 E, 250 K, 268 R, 275 S, 325 F, 335 A, 348 L, 380 R, 451 E, 459 E, 463 E, 467 V, 468 F, 489 R, 581 A, 583 R, 589 Y, 596 V, 597 F, 604 L, 613 R, 625 Q, 627 R, 632 I, 638 I, 640 G, 699 S, 708 E, 746 T, 748 N, 753 A, 775 L, 796 G, 799 Y, 801 S, 803 R, 879 N, 910 Q, 935 L, 938 Q, 1020 L, 1023 M, 1027 S, 1065 I)
_rpo_C1	-6089,39	-6079,39	31 (21 R, 61 S, 76 V, 79 T, 83 D, 84 P, 129 L, 139 G, 148 N, 150 S, 154 S, 156 V, 210 S, 231 S, 259 I, 267 R, 423 V, 432 S, 446 Q, 548 M, 564 Y, 569 F, 573 T, 575 D, 585 P, 603 N, 608 L, 636 Y, 642 H, 644 Q, 680 R)
_rpo_C2	-14495,1	-14479,8	153 (8 V, 19 M, 26 L, 42 V, 44 T, 49 R, 223 I, 231 L, 233 G, 238 I, 256 P, 278 R, 379 L, 384 L, 386 I, 394 L, 424 A, 427 S, 434 R, 436 R, 453 G, 457 P, 480 L, 481 C, 486 V, 497 M, 504 F, 512 L, 533 K, 535 I, 536 D, 542 R, 543 T, 545 S, 551 L, 554 P, 557 F, 561 D, 563 Y, 566 S, 568 A, 587 N, 589 L, 592 C, 601 R, 622 S, 632 V, 637 M, 648 G, 649 T, 659 Q, 664 Q, 680 P, 699 L, 712 E, 718 M, 730 P, 732 E, 733 M, 738 R, 747 G, 749 E, 753 S, 759 F, 769 A, 771 T, 775 Y, 783 I, 798 S, 799 Q, 821 G, 848 R, 882 S, 890 T, 891 A, 895 L, 897 S, 899 S, 900 E, 904 I, 905 H, 906 I, 913 V, 916 Q, 917 S, 919 P, 922 R, 924 G, 926 F, 932 R, 936 C, 937 K, 940 I, 948 F, 950 T, 951 G, 952 P, 964 E, 965 A, 968 I, 969 I, 970 S, 974 L, 977 P, 986 V, 988 F, 989 C, 992 Y, 1000 V, 1003 K, 1006 L, 1007 S, 1017 V, 1022 T, 1024 K, 1032 R, 1033 R, 1035 Y, 1040 C, 1042 K, 1045 W, 1047 L, 1049 H, 1058 D, 1059 Y, 1060 Y, 1063 G, 1064 W, 1068 N, 1080 L, 1107 P, 1138 S, 1174 D, 1179 K, 1220 R, 1245 S, 1249 L, 1330 K, 1344 L, 1346 I, 1348 K, 1349 K, 1351 I, 1357 R, 1362 H, 1366 L, 1368 C, 1371 G, 1373 K, 1375 F, 1377 E, 1379 S, 1380 N)
_rps_12	-850,70	-830,32	8 (13 Q, 16 K, 18 I, 57 L, 88 K, 116 S, 117 A, 118 L)
_rps_19	-682,64	-679,31	10 (16 S, 17 E, 26 E, 27 E, 65 R, 78 L, 81 V, 82 R, 84 A, 88 N)
_ycf_1	-29977,7	-29875,2	505 (3 F, 7 L, 27 L, 32 L, 43 F, 48 R, 54 S, 66 A, 102 F, 104 W, 107 H, 117 T, 139 F, 148 T, 180 L, 193 R, 196 H, 200 S, 236 V, 243 T, 253 S, 259 Y, 271 S, 277 S, 283 E, 284 E, 292 H, 294 K, 295 E, 297 R, 307 S, 309 L, 311 T, 314 E, 316 W, 317 K, 318 L, 319 G, 321 P, 326 R, 327 I, 328 N, 329 I, 330 N, 331 K, 332 K, 333 I, 334 D, 335 I, 336 I, 337 Y, 338 L, 339 W, 340 V, 342 K, 345 I, 348 F, 353 R, 364 D, 383 K, 389 K, 402 S, 405 L, 407 R, 408 K, 410 S, 416 K, 418 L, 419 L, 427 T, 433 C, 435 L, 441 S, 444 Q, 446 L, 451 R, 453 P, 455 L, 462 N, 473 C, 474 L, 477 A, 482 L, 485 P, 489 T, 490 I, 493 L, 496 R, 497 T, 501 T, 503 T, 507 D, 508 L, 513 L, 528 L, 529 C, 530 R, 532 S, 534 L, 536 S, 542 S, 544 N, 545 K, 546 E, 548 Y, 549 L, 552 L, 553 F, 558 T, 559 H, 562 D,

			563 Q, 565 I, 566 M, 568 K, 569 K, 570 S, 572 V, 574 K, 575 R, 577 E, 578 V, 584 Q, 591 E, 593 F, 596 E, 600 F, 601 T, 605 S, 606 G, 608 N, 612 A, 614 R, 615 T, 616 I, 619 E, 621 A, 622 N, 623 P, 628 T, 631 I, 632 T, 635 N, 637 S, 640 F, 661 C, 662 N, 663 L, 668 L, 671 P, 680 T, 681 D, 684 L, 685 F, 686 F, 691 K, 694 L, 695 L, 696 F, 699 W, 700 M, 701 G, 702 I, 706 D, 712 K, 714 E, 715 E, 717 K, 718 D, 720 N, 722 E, 724 E, 726 S, 728 I, 731 A, 733 L, 735 T, 737 A, 740 S, 741 F, 742 T, 744 L, 745 I, 749 L, 754 I, 757 L, 761 A, 765 L, 770 L, 772 I, 775 W, 776 H, 779 F, 786 K, 791 T, 800 T, 803 P, 804 Q, 808 T, 809 D, 816 I, 817 H, 827 S, 829 V, 830 R, 832 H, 833 H, 834 I, 836 Q, 837 M, 841 K, 844 Q, 845 N, 854 T, 856 T, 857 K, 858 I, 859 P, 862 S, 863 L, 866 K, 868 L, 871 K, 877 L, 878 K, 880 I, 884 V, 886 N, 887 K, 889 F, 890 Q, 892 I, 894 F, 895 L, 898 K, 899 R, 901 L, 904 K, 908 I, 910 W, 911 V, 912 I, 915 I, 916 R, 922 I, 925 I, 928 V, 930 L, 932 L, 933 F, 936 L, 944 P, 945 N, 948 N, 951 L, 955 N, 962 P, 968 M, 970 W, 971 L, 973 Y, 978 R, 980 I, 988 I, 991 K, 993 Q, 996 Q, 997 T, 998 E, 1000 E, 1007 Y, 1011 I, 1012 L, 1013 K, 1015 Y, 1017 H, 1018 L, 1019 W, 1024 R, 1034 H, 1037 I, 1050 V, 1052 S, 1054 T, 1055 F, 1056 F, 1058 I, 1060 A, 1063 L, 1069 N, 1072 N, 1074 S, 1076 Y, 1079 K, 1080 R, 1082 Q, 1083 K, 1084 K, 1086 G, 1088 N, 1090 I, 1092 Q, 1094 K, 1097 L, 1098 I, 1099 L, 1104 F, 1106 T, 1112 E, 1114 R, 1116 Q, 1118 S, 1120 I, 1121 Y, 1122 W, 1126 S, 1141 L, 1142 F, 1145 Y, 1147 L, 1149 P, 1156 N, 1167 Y, 1168 C, 1172 G, 1176 P, 1178 S, 1181 K, 1184 H, 1197 S, 1203 I, 1205 Q, 1209 R, 1211 I, 1215 W, 1216 R, 1219 R, 1221 K, 1222 L, 1223 R, 1233 L, 1239 F, 1240 N, 1241 S, 1246 D, 1249 A, 1257 D, 1259 C, 1274 N, 1275 P, 1276 P, 1278 S, 1281 S, 1282 E, 1288 K, 1289 E, 1290 A, 1291 K, 1296 H, 1297 F, 1299 T, 1300 S, 1317 L, 1320 K, 1322 I, 1325 S, 1329 L, 1336 C, 1340 S, 1341 I, 1342 C, 1344 R, 1348 E, 1350 W, 1351 T, 1353 A, 1356 R, 1357 R, 1358 N, 1360 Y, 1365 T, 1368 H, 1370 N, 1374 M, 1376 H, 1377 Q, 1378 K, 1381 P, 1382 C, 1386 R, 1387 N, 1396 K, 1398 R, 1403 E, 1406 H, 1407 A, 1411 T, 1418 G, 1421 F, 1423 V, 1425 K, 1427 K, 1430 I, 1438 L, 1439 N, 1441 D, 1442 A, 1443 N, 1444 E, 1451 R, 1457 L, 1459 V, 1460 G, 1462 F, 1463 E, 1467 H, 1468 E, 1470 Q, 1471 N, 1473 G, 1476 V, 1477 L, 1480 L, 1483 Q, 1484 N, 1486 K, 1487 A, 1493 R, 1494 K, 1495 F, 1497 M, 1500 S, 1501 K, 1507 T, 1511 M, 1515 S, 1519 N, 1520 S, 1527 W, 1528 I, 1529 N, 1530 F, 1531 S, 1533 E, 1534 K, 1540 R, 1541 T, 1548 V, 1549 K, 1551 I, 1555 A, 1558 S, 1560 K, 1562 D, 1566 L, 1570 F, 1573 K, 1574 D, 1576 V, 1578 K, 1586 F, 1587 L, 1592 C, 1593 L, 1596 R, 1598 D, 1599 G, 1601 S, 1612 V, 1613 D, 1616 H, 1620 N, 1623 T, 1629 E, 1631 G, 1632 E, 1633 L, 1634 K, 1636 Y, 1638 V, 1639 R, 1640 H, 1642 N, 1645 F, 1647 G, 1651 N, 1654 F, 1656 I, 1672 R, 1675 L, 1677 S, 1678 K, 1680 C, 1683 A, 1687 P, 1690 C, 1702 L, 1704 E, 1705 D, 1712 E, 1714 N, 1715 L, 1716 M, 1718 L, 1744 S)
***ycf*2**	-18981,9	-18817,4	654 (3 R, 6 F, 7 K, 8 S, 11 F, 13 F, 22 L, 28 K, 30 N, 39 F, 46 M, 54 W, 55 S, 63 R, 66 T, 67 S, 72 T, 74 K, 76 V, 77 V, 80 V, 81 V, 82 V, 84 L, 85 I, 86 S, 92 K, 99 L, 103 G, 108 P, 115 I, 124 W, 125 S, 128 R, 139 P, 141 G, 143 K, 144 I, 145 S, 146 D, 148 C, 150 M, 157 W, 158 V, 159 L, 161 I, 163 Q, 165 C, 175 R, 181 N, 182 R, 183 Y, 184 F, 185 G, 186 K, 187 T, 190 Q, 192 L, 196 V, 208 S, 214 L, 217 S, 229 W, 231 F, 243 I, 253 E, 258 D, 259 L, 261 C, 266 A, 271 R, 274 H, 275 F, 276 L, 279 Q,

302 W, 309 C, 310 A, 311 Q, 330 Y, 335 L, 345 W, 362 G, 369 Q, 373 T, 374 R, 379 Q, 387 K, 389 S, 390 Y, 400 S, 402 R, 404 E, 416 E, 418 Q, 420 L, 431 F, 434 T, 438 E, 448 L, 451 S, 458 F, 462 E, 466 N, 475 E, 476 E, 487 Y, 498 L, 502 P, 506 S, 507 T, 508 I, 510 Q, 512 L, 514 K, 515 K, 519 V, 522 V, 523 P, 527 V, 529 N, 530 Q, 536 F, 544 N, 546 V, 554 D, 555 P, 556 G, 557 C, 559 M, 564 E, 571 N, 576 L, 577 N, 580 P, 581 F, 582 F, 583 D, 584 F, 585 F, 586 H, 588 F, 589 H, 590 D, 591 R, 592 N, 593 K, 594 G, 595 G, 596 Y, 597 A, 598 L, 599 R, 600 H, 603 F, 624 Y, 625 H, 628 S, 634 K, 635 K, 636 F, 647 S, 649 N, 652 L, 667 S, 671 I, 673 K, 674 S, 675 V, 690 T, 691 A, 692 V, 702 V, 704 Q, 723 R, 728 R, 731 L, 740 E, 745 R, 758 I, 762 T, 765 R, 767 L, 770 F, 772 N, 773 S, 780 P, 783 S, 784 R, 785 T, 787 R, 791 W, 795 A, 800 W, 803 G, 804 S, 809 E, 818 P, 820 Q, 824 A, 828 R, 830 R, 831 I, 833 Q, 835 S, 841 A, 845 E, 846 D, 847 L, 848 S, 850 S, 854 F, 858 S, 860 P, 864 V, 873 R, 877 H, 878 I, 881 L, 884 P, 889 C, 891 Q, 895 S, 905 K, 909 F, 910 L, 917 S, 922 F, 926 G, 930 L, 933 L, 937 I, 941 M, 943 D, 952 G, 954 S, 958 T, 961 Y, 962 F, 965 I, 972 W, 978 P, 985 I, 987 S, 989 Y, 1002 H, 1010 R, 1014 D, 1021 N, 1044 C, 1045 A, 1049 K, 1050 D, 1051 L, 1057 T, 1061 I, 1072 N, 1074 F, 1081 T, 1084 L, 1088 L, 1091 P, 1093 G, 1096 P, 1100 R, 1106 A, 1110 A, 1111 T, 1114 T, 1118 I, 1119 V, 1123 R, 1125 Y, 1128 P, 1139 R, 1140 N, 1145 Y, 1151 N, 1157 T, 1158 P, 1161 E, 1163 Y, 1165 P, 1166 S, 1175 C, 1177 K, 1184 Q, 1186 Y, 1188 T, 1189 F, 1190 Q, 1198 L, 1206 T, 1211 F, 1227 T, 1230 D, 1231 P, 1234 I, 1237 S, 1240 K, 1254 I, 1255 L, 1256 R, 1257 P, 1259 T, 1261 K, 1264 T, 1266 W, 1267 T, 1268 L, 1271 E, 1276 C, 1277 L, 1278 Q, 1281 L, 1282 L, 1283 S, 1284 E, 1286 M, 1290 K, 1295 I, 1297 L, 1299 W, 1300 A, 1303 R, 1307 A, 1312 Y, 1316 F, 1320 V, 1324 L, 1325 V, 1326 R, 1330 L, 1332 V, 1334 R, 1335 A, 1336 S, 1338 E, 1345 K, 1348 S, 1350 M, 1352 P, 1354 Y, 1356 M, 1358 F, 1359 R, 1360 K, 1361 L, 1362 L, 1370 L, 1372 S, 1375 L, 1381 V, 1382 V, 1383 L, 1384 E, 1385 Q, 1387 G, 1389 S, 1391 E, 1392 E, 1395 G, 1396 S, 1397 A, 1398 S, 1399 G, 1400 G, 1404 W, 1405 G, 1406 G, 1407 A, 1409 G, 1410 V, 1413 I, 1415 S, 1417 K, 1418 K, 1420 W, 1421 K, 1431 I, 1437 R, 1439 I, 1442 R, 1451 S, 1454 I, 1459 R, 1461 R, 1464 V, 1465 N, 1466 G, 1467 D, 1468 W, 1471 E, 1475 F, 1476 W, 1477 V, 1479 N, 1481 D, 1482 S, 1484 D, 1485 D, 1486 E, 1488 R, 1489 E, 1490 F, 1492 V, 1496 T, 1502 R, 1505 K, 1506 I, 1514 D, 1517 S, 1518 K, 1529 P, 1533 S, 1537 L, 1542 K, 1554 C, 1566 Q, 1569 A, 1570 Y, 1574 S, 1575 C, 1576 G, 1577 A, 1578 N, 1585 P, 1590 R, 1592 A, 1593 L, 1594 S, 1595 P, 1607 T, 1615 Y, 1620 S, 1622 V, 1629 P, 1630 N, 1632 F, 1633 L, 1639 G, 1640 Y, 1641 P, 1642 I, 1645 S, 1646 D, 1647 D, 1648 T, 1650 I, 1652 D, 1653 S, 1655 D, 1656 T, 1658 I, 1659 D, 1661 S, 1662 D, 1664 I, 1665 Y, 1668 G, 1669 S, 1670 D, 1671 D, 1672 D, 1673 L, 1677 T, 1678 E, 1679 L, 1680 L, 1681 T, 1684 M, 1685 T, 1686 P, 1687 N, 1688 I, 1689 D, 1690 Q, 1691 F, 1692 D, 1693 I, 1694 T, 1695 L, 1699 L, 1702 A, 1708 I, 1717 H, 1718 V, 1720 E, 1723 Y, 1724 L, 1725 S, 1726 L, 1727 G, 1730 E, 1737 C, 1741 S, 1755 Q, 1756 K, 1760 A, 1765 K, 1767 L, 1773 I, 1775 K, 1777 L, 1778 L, 1782 R, 1787 T, 1799 R, 1803 T, 1807 G, 1808 S, 1809 I, 1815 A, 1816 R, 1819 V, 1820 A, 1821 L, 1824 E, 1825 A, 1831 T, 1833 K, 1835 Y, 1839 T, 1845 A, 1846 L, 1848 R, 1849 K, 1852 D, 1855 S,

	1856 Q, 1860 V, 1862 D, 1872 R, 1874 V, 1875 A, 1879 L, 1883 C, 1893 K, 1894 K, 1895 N, 1897 C, 1898 K, 1899 G, 1902 S, 1903 D, 1912 G, 1913 T, 1914 S, 1917 K, 1918 F, 1942 P, 1946 N, 1947 W, 1949 T, 1954 V, 1958 S, 1967 L, 1970 L, 1972 V, 1974 G, 1976 P, 1978 L, 1979 A, 1980 G, 1982 S, 1987 D, 1990 Q, 2000 L, 2002 S, 2007 Q, 2012 Q, 2015 S, 2018 T, 2019 V, 2021 Q, 2022 R, 2024 L, 2027 K, 2028 Y, 2029 E, 2030 S, 2031 E, 2036 A, 2037 L, 2039 P, 2040 Q, 2041 Q, 2042 I, 2044 E, 2045 D, 2046 L, 2049 H, 2055 R, 2063 E, 2065 P, 2071 P, 2073 W, 2074 I, 2078 R, 2081 R, 2082 I, 2084 S, 2089 E, 2091 Q, 2104 Q, 2106 Q, 2107 T, 2108 R, 2111 S, 2112 S, 2113 K, 2114 E, 2115 Q, 2116 G, 2117 F, 2118 F, 2119 R, 2120 T, 2121 S, 2125 W, 2128 A, 2130 P, 2131 L, 2135 F, 2136 K, 2138 Q, 2139 P, 2140 F, 2141 V, 2143 V, 2156 S, 2158 G, 2160 I, 2163 Q, 2164 T, 2166 P, 2167 P, 2170 M, 2173 R, 2180 Q, 2191 Q, 2192 R, 2194 F, 2203 G, 2209 T, 2211 S, 2220 L, 2225 G, 2234 T, 2239 R, 2241 L, 2243 P, 2254 G, 2257 F)

S3 Table. Positive selection sites identified with selecton with d.f. = 1. “Null” and “positive” columns list likelihood values obtained under the models M8a (null model) and M8 (positive selection), respectively.



S1 Figure. PCR validation of the *ndh*F deletion. PCR amplifications using (a) F1 and R1 primers; (b) F2 and R2 primers. M = marker II (λ DNA / *Hind* III digested); 1 = *O. fusca* Campania, 2 = *O. fusca* Tuscany, 3 = *O. iricolor* Greece; 4 = *O. sphegodes* Campania; 5 = *O. sphegodes* Apulia; 6 = *O. incubacea* Apulia; 7 = *O. insectifera* Spain. A dotted line represents the IRB-SSC junction.

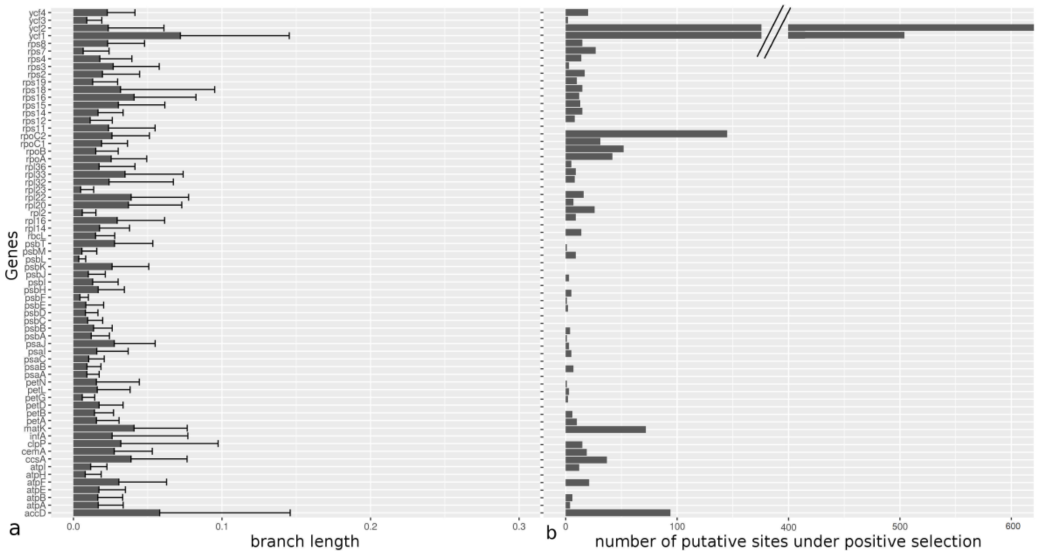
Primer sequences:

F1: 5' - GCTCCGTTCCATGCCTCATT – 3'

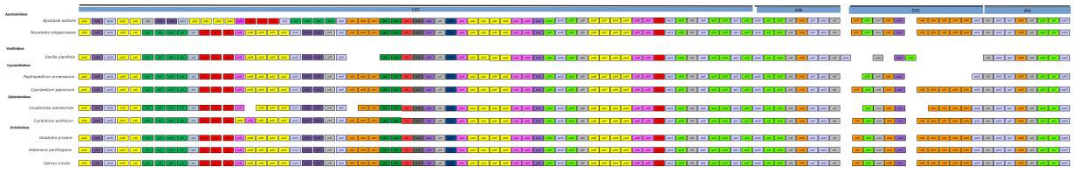
R1: 5' – TCGTCGTATGTGGGCTTTCC – 3'

F2: 5' – TTAGCAATTGCACCGACAAA – 3'

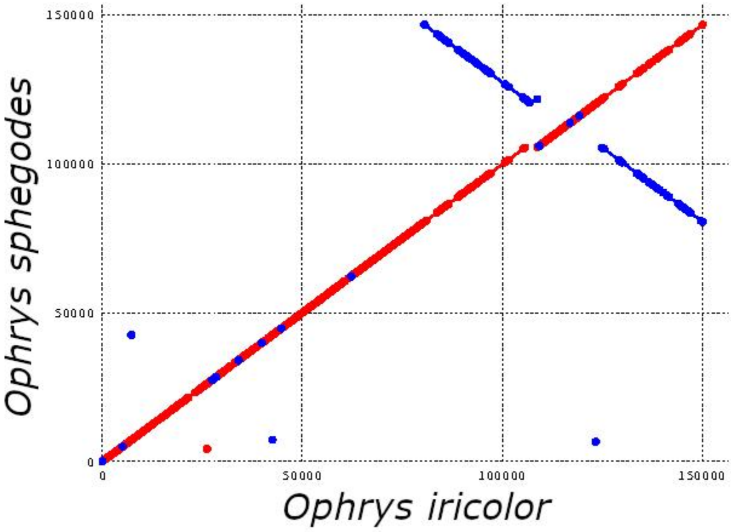
R2: 5' – TCTGTTTCCACCGGACAG – 3'



S2 Figure. Molecular evolution analyses of *Ophrys* plastid genes: a) divergence of protein-coding genes (gene divergence was estimated by the sum of total branch lengths in each gene tree inferred, mean ± SD); b) number of putative sites under positive selection.



S3 Figure. Comparison of gene rearrangements in the plastid genomes among 10 species representative of the five Orchid subfamilies. Genes are indicated in the colored boxes. Boxes colors represent gene families: purple = photosystem I; yellow = photosystem II; orange = NADH-dehydrogenase; light green = ribosome large subunit; light blue = ribosome small subunit; light red = rubisco subunit; red = RNA polymerase; green = ATP synthase; pink = cytochrome b/f complex; dark grey = acetyl-CoA carboxylase; light grey = hypothetical plastid reading frame (ycf series), protease, translation initiation factor IF-1; dark purple = maturase; dark blue = envelope membrane protein.



S4 Figure. Dot-plot analyses of *Ophrys sphegodes* and *O. iricolor* plastid genomes using Mummer software. A positive slope indicates that compared sequences are in the same orientation; a negative slope indicates that compared sequences can be aligned, but their orientation is opposite. Red: Sequences in the same direction; Blue: inversions.

Capitolo V

Different filtering strategies of Genotyping-By-Sequencing data provide complementary resolutions of species boundaries and relationships in a clade of sexually deceptive orchids

5.1 Abstract

Ongoing hybridization and retained ancestral polymorphism in rapidly radiating lineages may mask recent cladogenetic events. This presents a challenge for the application of molecular phylogenetic methods to resolve differences between closely related taxa. We analyzed Genotyping-by-Sequencing (GBS) data to infer the phylogeny of four species within the *Ophrys sphegodes* complex, a recently radiated clade of orchids. We employed different data filtering approaches to detect different types of information contained in the dataset generated by GBS and estimated their effects on Maximum Likelihood (ML) trees, global F_{ST} and bootstrap support values. We obtained a single ML tree with high bootstrap support separating the species by using a large dataset based on loci shared by at least 30% of accessions. Bootstrap and F_{ST} values progressively decreased when filtering for loci shared by a higher number of accessions. However, when filtering more stringently to retain ancestral loci, i.e. homozygous loci shared with at least 70% of accessions, we identified two strongly supported clades. These clades group individuals independently from their *a priori* species assignment but corresponding to two plastid/mitochondrial haplotypes clusters. We infer that a less stringent filtering preferentially retains rapidly evolving

lineage-specific loci, which may better delimit lineages. In contrast, when using highly represented homozygous loci, organellar DNA loci are preferentially selected and the signature of a putative hybridization event in the lineage prevails over the most recent phylogenetic signal. These results show that using differing filtering strategies of GBS data may improve insights into relationships between closely related species.

5.2 Introduction

Understanding the evolutionary relationships in closely related, recently diverged lineages often presents a methodological challenge (Maddison, 1997). Rapidly diverging taxa highlight the limit of the phylogenetic application of molecular markers as these lineages can be at the interface between incipient species and divergent ecotypes (Feder et al., 2012). Plastid DNA (cpDNA) has been widely applied in plant phylogenetic studies and lineage delimitation thanks to the ease of amplification and sequencing that come with its high copy number (Gielly & Taberlet, 1994). Plastid DNA markers are predominantly uniparentally inherited (including in orchids, Cafasso et al., 2004). Effective population size for such organellar markers is smaller than that of nuclear markers, thereby leading to greater genetic drift and resulting in faster coalescence times than diploid nuclear DNA (Petit et al., 2005; Hernández-León et al., 2013). However, the low evolutionary rate and the haploid nature of cpDNA severely limit its application in closely related species particularly when introgression (and consequent plastid capture) and incomplete lineage sorting are suspected (Sang et al., 1997). The use of diploid nuclear gene data is often thought to overcome the shortcomings of organellar gene genealogies, as nuclear genes have been reported to evolve up to five times faster (Wolfe, 1989; Ossowski, 2010; Schlüter et al., 2007). Nevertheless, disadvantages in the use of nuclear genes stem from their frequent occurrence in gene families (paralogy), recombination, and a general lack of available markers for non-model organisms (e.g. Doyle, 1997; Posada & Crandall, 2002) although the latter problem has been alleviated to a certain degree by the arrival of next generation sequencing (NGS) technology. The analysis of recently diverged taxa is further complicated by the frequent existence of retained ancestral polymorphism, when ancestral allelic variants are maintained in both descendant species following neutral expectations. However, coalescent theory predicts that the noise produced by incomplete lineage sorting can be reduced by sampling multiple genes per species (Edwards & Beerli, 2000). But it also results in genealogies that may differ in their topologies, because unlinked nuclear genes are differently affected by introgression and intragenic recombination (Degnan et al., 2009).

The use of large multilocus datasets, such as those consisting of sequence data from multiple, unrelated genomic regions can improve phylogenetic inferences by accounting for the stochasticity in the coalescent process (Knowles & Maddison, 2002; Knowles, 2009; Carstens et al., 2013). Indeed, analyzing multiple genes and alleles per species increases the probability to approximate the underlying species tree supported by the majority of the data (Small et al., 2004). This may help overcome the typical limitations of using single/few genes to assess phylogenetic relationships and demographic history of species (Edwards & Beerli, 2000; Edwards, 2009; Hipp et al., 2014). Recent advances of NGS tools and multilocus analyses have been applied for successful reconstruction of phylogenetic relationships and for delimitation of boundaries between closely related species within species complexes. Amongst the more common genomic methods, reduced-representation methods (reviewed in Davey & Blaxter, 2010), such as restriction-site associated DNA sequencing (RADseq; Miller et al., 2007; Baird et al., 2008; Rowe et al., 2011), or genotyping-by-sequencing (GBS; Elshire et al., 2011) identify sequence fragments of DNA that flank the recognition sites of restriction enzymes in an individual's genome (Baird et al., 2008; Miller et al., 2007) by using high-throughput sequencing technologies. This selection of DNA fragments, scattered throughout the individual genome, allows orthologous sequences to be targeted across multiple samples to identify a large number of genetic markers. These methods provide a useful tool particularly for surveying the genome of non-model organisms (Ellegren, 2014).

Most applications of genomic reduced-representation methods have been within species (e.g. Lewis et al., 2007; Emerson et al., 2010; Hohenlohe et al., 2010; Bruneaux et al., 2013; Wang et al., 2013) or among closely related species (e.g. Stölting et al., 2013; Wagner et al., 2013). This is because the primary challenge in applying these methods to reconstructing interspecific phylogenies lies in confidently identifying and assembling orthologous loci amongst the relatively short (i.e. usually 100 to 200 bp), usually non-coding sequence fragments produced with the NGS technologies (Rubin et al., 2012). This problem stems from the fact that: (1) the number of restriction sites that are conserved among taxa is expected to decrease with increased time since divergence; (2) the ability to compare orthologous loci is expected to decrease with phylogenetic distance due the progressive accumulation of mutations. These caveats indicate that such genotyping data are expected to be particularly valuable for recently diverged and closely related clades (Wagner et al. 2013).

The Mediterranean orchid genus *Ophrys* has not only attracted the interest of taxonomists since Darwin (e.g. Darwin, 1862; Kullenberg, 1961), but it has also become a useful system to study speciation and reproductive isolation (Scopece et al., 2007; Xu et al., 2011). Nevertheless, it also represents evidence of fast evolving clades very recalcitrant to most methods for phylogenetic analyses (Breitkopf et al., 2015). The genus can be merged or split in a large number of lineages that are at least locally and temporally reproductively isolated enough to establish some morphological differences (Bateman et al., 2001; Vereecken et al., 2011). As post-zygotic barriers are effectively absent within closely related groups, reproductive isolation in sympatry is almost exclusively based on floral isolation through specific male pollinators that are lured by the floral scent, a copy of the sexual pheromone of con-specific females, to repeatedly copulate on flowers of only a single *Ophrys* species, leading to cross-pollination (Kullenberg, 1961). An accelerated diversification rate in terminal clades has been explained by the exploitation of a novel, species-rich and diverse groups of pollinators resulting in a recent and rapid radiation that is characterized by dynamic speciation processes due to repeated pollinator shifts (Breitkopf et al., 2015). Previous molecular studies in *Ophrys* (Devey et al., 2008; Breitkopf et al., 2015) support at least 10 main lineages that presumably give rise to species flocks by the adoption of pollinators from large diversified bee genera, such as *Eucera* and *Andrena*. Among these, the *O. sphegodes* complex represents one of the most species-rich groups in *Ophrys*, diversified only in the last 1 million years by exploiting different *Andrena* and, to a lesser degree, *Colletes* bees as pollinators (Breitkopf et al., 2015; Delforge, 2016). Despite intensive past research, phylogenetic patterns and species diversity within this complex remain highly contentious. Both plastid and nuclear phylogenies – including the use of a dataset of multiple nuclear genes – failed to identify species relationships and to delimit species within the *O. sphegodes* complex (Soliva et al., 2001; Bateman et al., 2001; Breitkopf et al., 2015). Thus, this complex represents an ideal group for testing the application of NGS-based multilocus analyses for inference of relationships and species delimitation. The RADseq method has very recently been applied to the phylogeny of *Ophrys* at the level of the ~10 main lineages, confirming the suitability of NGS methods and approaches for phylogenetic purposes in taxonomically complex groups (Bateman et al., 2018). However, only one attempt to employ multilocus NGS approaches has previously been made at the within-species-complex level at the transition zone between species/population levels. Specifically, Sedeek et al. (2014) employed GBS data to present a UPGMA tree based on overall pairwise genotypic distances between individuals of the *O. sphegodes* complex. In this analysis, none of the internal nodes separating the species received any bootstrap support. Similarly, a STRUCTURE analysis, run on the same dataset, indicated a large proportion of shared polymorphism and found K=6 ancestry clusters as the most probable inference, at least in the employed dataset (88 individuals and 1233 loci with 1 SNP analyzed per locus). Here, we re-analyzed genome-wide sequence/SNP data collected by Sedeek et al. (2014) by using different criteria of locus selection in order to: (1) delimit the species boundaries within a group of four sympatric southern Italian species of the O. *sphegodes* complex, (2) infer a well-supported pattern of relationships/descendance for these species, and (3) identify the signature of past events affecting lineage divergence in this group.

5.3 Material and methods

5.3.1 Study system and GBS data source

Here, we investigated all four members of the *O. sphegodes* species complex co-growing in the National Park of Gargano (Apulia, Italy), i.e. *O. exaltata* subsp. *archipelagi* (Gölz & H.R. Reinhard) Del Prete, *O. garganica* Nelson ex O. & E. Danesch, *O. incubacea* Bianca and *O. sphegodes* Miller. These four species are pollinated through sexual deception by three different *Andrena* (*A. pilipes*, *A. morio*, *A. nigroaenea*, for *O. garganica*, *O. incubacea* and *O. sphegodes*, respectively) and a *Colletes* species (*Colletes cunicularis* for *O. exaltata*) (Paulus & Gack, 1990). The four investigated species co-flower in spring (from March to April) and occur in close proximity to each other in the study area (Xu et al., 2011; Sedeek et al., 2014).

We here re-analyze trimmed and demultiplexed GBS Illumina reads generated by Sedeek et al. (2014). From the full dataset of Sedeek et al. (2014), encompassing 127 accessions, we filtered the data according to the number of reads per accession. To maximize the number of reads per accession, we used a more conservative approach than Sedeek et al. (2014) by selecting only samples with at least 800 000 reads per accession (a total of 54 individuals) roughly corresponding to the median value of reads per accession in the original dataset. However, we also compared the results to datasets including accessions with at least 500 000 and 300 000 number of reads per accessions.

5.3.2 Plastid and mitochondrial haplotype network analysis

Plastid reads were identified by mapping GBS reads for each individual against the *O. iricolor* and *O. sphegodes* plastid genomes (Roma et al., 2018) using BWA MEM v. 0.7 software (Li, 2013) with the option -M that marks shorter split hits as secondary (as required by GATK software). Variant calling analysis was performed using Genome Analysis ToolKit v. 3.5 according to the GATK Best Practices workflow (McKenna et al. 2010). After the SNP and indel recalibration, a BAM format file was generated for each sample. Finally, a VCF file was generated with GATK package HaplotypeCaller with the option -ploidy 1. Plastid haplotypes network analysis was performed using POPART v. 1.7 software (Leigh & Bryant, 2015) by only using informative SNPs.

Mitochondrial reads were identified by blasting (BLAST 2.6.0) against the Organelle Genome Resources database (https://www.ncbi.nlm.nih.gov/genome/organelle/). We used informative mitochondrial SNPs shared by all 54 individuals to reconstruct a mitochondrial haplotype network using POPART v. 1.7 software.

5.3.3 GBS data assembly

In contrast to Sedeek et al. (2014), we used the software pipeline PYRAD v.1.2 (Eaton, 2014) to process the GBS reads instead of Stacks (Catchen et al., 2011). We choose this approach because it allows to build supermatrices with different minimum percentages of shared loci. Nucleotide base calls with a quality score below 20 were replaced with N, and sequences having more than five Ns were discarded from edited FASTA files created by PYRAD. Clustering was performed in

VSEARCH v. 1.0.16 (Edgar, 2010), using the forward reads faster version without reverse complement clustering because of the low overlap between forward and reverse reads.

Only Single Nucleotide Polymorphisms (SNPs) were used and their distribution per cluster checked in order to avoid markers with more SNPs potentially biasing the inference when treating each locus as one independent marker. Clusters with coverage less than five reads per locus and more than five heterozygous sites were discarded. Consensus sequences were then clustered across accessions at 88% similarity (the PYRAD default setting) and aligned using MUSCLE v. 3.8 (Edgar, 2004). We then applied a supermatrix approach in which all selected clusters were concatenated into a single alignment using PYRAD v.1.2. Missing data symbols (Ns) were inserted into the data matrix for loci without data for a given individual (Wagner et al., 2013).

5.3.4 Phylogenetic inference

To infer phylogeny from the GBS data, we built different supermatrices by selecting loci shared by at least 10 %, 30%, 50%, 70% and 90% of accessions and reconstructed Maximum Likelihood (ML) trees. ML analyses were conducted in RAXML v. 8.2.10 software using the GTRGAMMA nucleotide substitution model (an inclusive model accounting for a large proportion of missing data; see Roure et al., 2013) and with bootstrap support estimated from 1000 replicate searches. Phylogenetic trees were drawn using FIGTREE v. 1.4.3 software. To test for the effect of heterozygosity, we also reconstructed phylogenetic trees using RRHS software v. 1.0.0.2 (Lischer et al., 2014) on the supermatrix with loci shared by at least 30% of accessions.

For each ML tree, we calculated a mean bootstrap support value by averaging the bootstrap values over all tree nodes. Following Sedeek et al. (2014), for each locus, we calculated ‘global’ F_{ST} among all four species using BayeScan 2.1 (Foll & Gaggiotti, 2008), treating orchid species as four different populations. Then, we plotted average bootstrap support values and F_{ST} values averaged over all loci against the percentage of shared loci among accessions.

We also selected more ancestral and less variable loci (by filtering out the heterozygous loci) from the supermatrices built with loci shared by at least 70% and 90% of accessions. Additionally, we performed the same analysis by discarding from these supermatrices both plastid and mitochondrial reads. Plastid reads were discarded by using the BAM file previously generated for plastid haplotype search. Unmapped reads were retained by using SAMTOOLS v 1.5 (Li et al., 2009) with the parameters view and -f4 and then converted in FastQ format using SAMTOOLS Bam2fq. Mitochondrial reads, identified by blasting, were discarded using a custom Perl script.

On two supermatrices, i.e. (1) the one with 30% of shared loci and (2) the one with 70% of shared loci only including homozygous loci, we also performed analyses of population structure. First, pairwise distances between individuals based on unphased diploid SNP calls were calculated as described in Sedeek et al. (2014) by using a custom Delphi program using the biOP library (https://sourceforge.net/p/biop/). The advantage of this approach is that it avoids global threshold-based exclusion of loci from the dataset and utilizes the maximum number of data points available for any given pairwise comparison. Distance matrices were used for building Neighbor Joining (NJ) trees in FAMD 1.31 (Schlüter & Harris 2006). Second, we used the Bayesian clustering approach as implemented in STRUCTURE v 2.3.4 (Pritchard et al., 2000). Following the method described in Evanno et al. (2005), we tested K from 1 to 7 with a burn-in of 10 000 steps followed by 10 000 Markov chain Monte Carlo iterations with 3 replicates to confirm stabilization of the summary statistics.

5.4 Results

By selecting plastid loci from the GBS data shared by all individuals we identified six distinct haplotypes belonging to two phyletic clusters (A-D and E-F) according to the haplotype network analysis (Fig. 1). The two clusters are separated by two mutational steps. Within each cluster, the haplotypes were separated by single mutation steps (Fig. 1). By selecting four shared mitochondrial SNPs we identified six distinct haplotypes in the network analysis (Fig. S1).

SNP distribution per locus showed that the majority of loci (76%) included a maximum of three SNPs (Fig. S2). The ML phylogenetic analysis with the supermatrices with loci shared at least among 10% and 30% of individuals (Table 1) shows species-specific clades (Fig. S3 and Fig. 2). All four *Ophrys* species are reciprocally monophyletic. However, only in the supermatrix with loci shared at least among 30%, they all have bootstrap support above 70% (Fig. 2). Indeed, the tree built with the supermatrix with loci shared at least among 10% has higher bootstrap support for terminal clades, but the placement and monophyly of *O. garganica* was weakly supported. Analysis with RRHS software, which accounts for heterozygosity, yielded results consistent with these ML results (Fig. S4).

We observed similar phylogenetic relationships but a progressive decay in the bootstrap support when using datasets with fewer reads (at least 500 000 and 300 000 of reads) per samples (76 and 93 individuals, respectively, Fig. S9 and S10). Thus, all following analyses were performed with the dataset including 54 accessions with at least 800 000 reads per accession.

In a reduced supermatrix (with loci shared at least among 50% individuals), resolution of the four species clades slightly decreases as does bootstrap support for the placement of *O. garganica* as sister species of remaining taxa (Table 1; Fig. S5). By progressively reducing the number of loci shared among individuals (loci shared at least among 70% and 90% individuals) we observe a further progressive decay of bootstrap support across clades in the tree (Table 1, Fig. 3A). In these last analyses, individuals of the same species do not form monophyletic clades (Fig. S6, S7). Like bootstrap support, F_{ST} values also decrease progressively as the number of shared loci increases and as the dimensions of the supermatrices are reduced (Table 1; Fig. 3B).

The phylogenetic analysis using the small supermatrix with only homozygous loci (i.e. with homozygous loci shared at least among 70% individuals; 185 019 base pairs in width, 253 informative SNP) again produced a tree topology with high bootstrap support, but only for the main basal nodes (Fig. 4). With this supermatrix, we identified main lineages (bootstrap support above 90%) that group accessions independently from their species assignment but instead according to the plastid and mitochondrial clusters identified in the haplotype network analyses (Fig. 1, Fig. S1). However, after removing both plastid (2.4% of the total) and putative mitochondrial (7.9% of the total) SNPs from the 253 informative SNPs of this supermatrix, the bootstrap support of the main basal nodes decays, though the general grouping is maintained. When using the reduced dataset with 70% of shared homozygotes loci, the NJ trees based upon pairwise SNP distances identified two main lineages corresponding to the two haplotype clusters (cf. Fig. 1) (Fig. S8A). Accordingly, Bayesian analysis on this dataset identified K = 2 (Fig. S8A) as the most probable.

Instead, when using the large dataset with 30% of shared loci, the NJ tree confirmed the ML tree topology: a clear delimitation of the four *Ophrys* species is evident (Fig. S8B). However, Bayesian analysis on this dataset identified K = 5 as the most probable K, mostly corresponding to species assignment, but with *O. incubacea* divided in two groups (Fig. S8B). After we removed both plastid and mitochondrial loci from the dataset with 30% of shared loci, the resulting Bayesian analysis identified K = 2 as the most probable K. However, the plot of delta K also shows a peak at K = 4 fully corresponding to the four species assignment recognized by the corresponding ML and NJ trees (Fig. S8C).

5.5 Discussion

Despite of the great deal of attention the phylogeny of the Mediterranean orchid genus *Ophrys* has received over the past twenty years, relationships among closely related species are still unresolved when using traditional phylogenetic nuclear and plastid markers (Bateman et al., 2001; Soliva et al., 2001; Devey et al., 2008; Breitkopf et al., 2015). Here we show that multilocus GBS data, when properly filtered, can provide a useful tool to assess the degree of genetic separateness/togetherness and pattern of relationship among four species of the *Ophrys sphegodes* complex that are treated as separate species, subspecies, varieties or populations depending on contrasting taxonomic treatments (Bateman et al., 2011; Delforge, 2016; Vereecken et al., 2011). Previous studies employed plastid and/or nuclear genes to infer phylogenetic relationships in *Ophrys* and included in their analysis multiple accessions from the *O. sphegodes* complex (Soliva et al., 2001; Devey et al., 2008; Breitkopf et al., 2015). Results of these studies supported the monophyly of the *O. sphegodes* complex, but patterns of relationships within the species complex were largely unresolved. The application of high-throughput sequencing generating a large multilocus dataset enabled to resolve fine-scale genetic divergence among members of *O. sphegodes* complex. Individuals of the same species (at least based on morphologic traits and scent emission) form well-supported, and reciprocally monophyletic, clades suggesting that insufficient informative characters in previous studies were the major cause for poor resolution and confirm the power and efficiency of multilocus approaches to identify species borders and patterns of relationships among closely related species in *Ophrys*.

Higher resolution of multilocus dataset was already detected between the nuclear single-copy *LFY* gene and AFLP markers for resolving the phylogeny of the *Ophrys fusca* group (Schlüter et al., 2011a). However, compared to the AFLP approach, GBS data overcomes the difficulties associated with AFLP data of assessing fragment homology in the absence of knowledge about the underlying sequence (Althoff et al., 2007). Additionally, GBS data allow for a more robust assessment of relationships because of the larger size of the available input data matrix and the fact that they provide a codominant source of data.

However, species resolution and phylogenetic relationships (with high bootstrap support) with GBS data have been obtained mainly when selecting the larger supermatrices with a higher number of missing data (loci shared at least between 10, 30 or 50% accessions, Fig. 2, Fig. S3, S5). Interestingly, a progressive decay in species resolution occurs when selecting loci with higher representation, i.e. fewer loci but less missing data by increasing the number of individuals sharing the loci from 70% to 90%) (Fig. S6, S7). This is mirrored by a corresponding decrease both of average bootstrap support of resulting trees and of between-lineage differentiation of the employed loci as measured by global F_{ST} (Fig. 3)

There is a debate on how both the size of the matrix and the data matrix properties (i.e. the number of missing loci and whether they are randomly distributed) may contribute to successfully disclose patterns of relationship (Lee et al., 2018). Recent empirical studies have confirmed that larger data matrices, despite their large amount of missing data (SNPs called in a lower number of accessions), result in better resolution in delimiting very closely related species (as in Lake Victoria cichlid fishes, see Wagner et al., 2013) and simulations have shown that a higher proportion of missing data in larger data matrices does not adversely affect phylogenetic accuracy as long as there is no systematic bias (Rubin et al., 2012). The most likely explanation provided was that a less stringent filtering (i.e., inclusion of loci shared by fewer samples) preferentially retains lineage-specific loci, which may allow coalescent methods to better delimit lineages (Huang & Knowles, 2014). Accordingly, Huang and Knowles (2014) by using simulated data showed that low tolerance to missing data leads to a disproportionately high exclusion rate of loci with high mutation rate/substitution rate. These latter loci, with a higher amount of missing data, are therefore those that have differentiated among very recently diverged lineages (increased F_{ST}) and may be especially informative for phylogenetic analyses. Instead, when loci with missing data are excluded in favor of more highly represented (i.e. more ancestral) loci across the dataset, there is a shift in the *spectrum* of mutation rates that negatively affects the power of phylogenetic resolution. Indeed, loci conserved between distant relatives are expected to be slowly evolving. This translates into a disproportionately low number of SNPs and consequently a weak phylogenetic signal (Leaché et al., 2015). Furthermore, those loci that increase in differentiation among species are more likely to be under divergent/positive selection and fast evolving, whereas the slowly evolving loci may be more likely to be neutral and thus particularly prone to be retained as ancestral polymorphisms, a phenomenon particularly relevant in very recent divergent species such as those belonging to the *O. sphegodes* complex (Breitkopf et al., 2015). More ancestral loci (shared among many accessions) are also those with lower F_{ST} values (i.e. the more stringently-filtered data set have lower global F_{ST} value and, correspondingly, less bootstrap support). This is consistent with the idea that pollinator-driven ecological speciation in *Ophrys* may first result in divergent selection and accelerated evolution upon few large-effect genes in the genome that are linked to pollinator-mediated reproductive isolation (Schlüter et al., 2011b; Sedeek et al., 2014).

The supermatrices with a high number of loci (shared by at least 10% and 30% of accessions) allow differentiating the four species with *O. garganica* sister to a clade with the remaining three species. Here *O. incubacea* is sister to the inner lineage of *O. sphegodes* and *O. exaltata*. This pattern of relationships suggests a transition of pollinators from basal *Andrena* species (in *O. garganica, O. incubacea* and *O. sphegodes*) to *Colletes* (in *O. exaltata*), a scenario congruent with a pollinator-mediated progenitor–derivative speciation (Schlüter et al., 2011a) driven by genetic change affecting flower odour emission (as hypothesized by Xu & Schlüter, 2015; see also Sedeek et al., 2016). Basal relationships among species have higher support in the supermatrices with loci shared by at least 30% of accessions than in the larger supermatrix with loci shared by at least 10% of accessions that, in contrast, has higher support in the terminal clades. A potential explanation for this discrepancy is that this latter supermatrix includes a high number of loci that are shared by two individuals only (i.e. roughly 10% of the accessions) so increasing the strength of terminal relationships at the expense of basal relationships (Fig. S3).

By using a more stringent approach, i.e. by selecting homozygotes loci shared by many accessions (at least 70%), the resulting smaller supermatrix generates a phylogenetic tree identifying two main supported clades (bootstrap support ≥ 90%) (Fig. 4). In these clades, individuals cluster independently from their taxonomic attribution. Instead, individuals cluster according to their plastid and mitochondrial haplotypes. For instance, individuals characterized by cpDNA haplotype E belong to all four distinct species. Conversely, five distinct cpDNA haplotypes (A, C, D, E and F) are attributed to *O. garganica* individuals (Fig. 4).

The two most common cpDNA haplotypes (E and D) are five mutations different from each other and would therefore be considered as two independent evolutionary units based on the haplotype network analysis. All four *Ophrys* species contain at least one cpDNA haplotype from each of the haplotype clusters. Network analysis of mtDNA identifies two main haplotype lineages. Almost all accessions carrying cpDNA haplotypes of lineage A-D have the mtDNA haplotype of lineage 1-4. By blasting the smaller supermatrix for plastid and mitochondrial reads, we discovered that these more conserved organellar reads strongly contributed to support the resulting ML tree. In other words, in the small supermatrix, there is a strong contribution of (haploid) organellar DNA to defining overall tree topology. By selecting homozygous loci common to at least 70% of individuals we favor retaining both organellar DNA loci and of more ancestral (fixed) nuclear loci. Retention of organellar (haploid) loci in the original dataset employed by Sedeek et al. (2014), may explain the finding of K=6 in their Bayesian population STRUCTURE analyses. Notably, the analysis performed by Sedeek et al. (2014) is congruent with our Bayesian analysis (K=5) on the larger dataset (loci shared at least between 30% accessions) with *O. incubacea* split in two groups corresponding to the two haplotype clusters. Thus, we argue that their stringent setting (≤55% missing data per individual and ≤10% missing data per locus, 1233 loci with only 1 SNP analyzed per locus) and the retention of haploid organellar loci (coupled with presence of two different haplotypes in *O. incubacea),* explain the different results between the analysis presented by Sedeek et al. (2014) and in the present study.

The different gene genealogies of rapidly evolving nuclear loci compared to the slower organellar and ancestral nuclear loci may explain the incongruence between the tree topologies we observe from our large and stringent supermatrices. In a rapid radiation, there has not been enough time for lineage coalescence of conserved organellar and nuclear loci in each new species (Neigel & Avise, 1986).

Indeed, when filtering out the plastid and mitochondrial loci, the remaining (ancestral) nuclear loci do not support any phylogenetic pattern (data not shown). This clearly indicates that these more ancestral and fixed nuclear loci are those more prone to be retained as ancestral polymorphism among species (and are those displaying the lower F_{ST} values averaged over all loci).

While *Ophrys* is relatively old (7.1–2.9 Ma), some of the species complexes (including the *O. sphegodes* complex) are estimated to be extremely young (Breitkopf et al., 2013, 2015).

This consistent pattern of variation found at organellar DNA loci implies that phylogenetic reconstructions based on fewer conserved loci may be regarded as gene genealogies representing the older evolutionary history of the *Ophrys* lineage rather than a phylogeny that reflects the most recent organismal history (i.e. the *O. spegodes* complex). Even though cpDNA (and mtDNA) phylogenetic distributions can lack concordance with species boundaries when species are very recently separated it still may bear the signature of phylogeographic history of the lineages. The observed haplotype patterns, in particular the fact that the two main cpDNA haplotypes shared among the four species were more strongly divergent from each other than from (derived) haplotypes restricted to a single species, suggests admixture of two *Ophrys* lineages in a common ancestor of the investigated species group (i.e. the retention of haplotype diversity that was present prior to speciation in the descendant species). These two distinct lineages may for instance have segregated (and diverged) in different refugia and later hybridized in secondary contact zones (Widmer & Lexer, 2001) prior to radiation within the *O. sphegodes* complex. This is consistent with the low amounts of differentiation among actual species, only detectable when using the most variable nuclear loci or pollinator-relevant phenotypic traits. Both ancestral polymorphism and signature of old hybridization are more evident in conserved than in fast evolving regions, polymorphism at which likely emerged after the ancestral hybridization event. These rapidly evolving regions (with higher substitution rate) are those preferentially retained in the large supermatrices (as 10 % and 30% of shared loci) and that largely contributed to species resolution in the ML analysis. This highlights that fast-evolving nuclear loci such as those employed here in the large data matrix are likely to be the most important tool for detecting the very recent phylogenetic signal among extremely young species (Wagner et al., 2013).

Past hybridization has been advocated at the bases of recent species radiations as in the Hawaiian silverswords (Barrier et al., 1999) and in African cichlid fishes (Meier et al. 2017). Hybridization occurring when allopatric lineages come into secondary contact may fuel the onset of an adaptive radiation by providing a new genetic background for novel trait combinations or for increasing genotypic diversity (Abbot et al., 2013) that, in *Ophrys*, can allow the exploitation of new available pollinator niches and, consequently, the evolution of premating isolation (Breitkopf et al., 2013, 2015; Vereecken et al., 2010). Although incomplete lineage sorting is difficult to distinguish from reticulation, our results including fast-evolving loci suggest that current hybridization is at least unlikely to occur frequently among the four species in the sympatric study region. This has been further corroborated by local experimental studies confirming premating isolation among the four *Ophrys* species due to pollinator isolation (Xu et al., 2011; Sedeek et al., 2014).

In conclusion, we present a well-resolved phylogenetic tree from a group that has represented a challenge due to its recent origin and weak genomic differentiation (Breitkopf et al., 2013). While the different levels of information contained in GBS loci with different substitution rate and genealogy should be properly accounted for, our finding that these sympatric *Ophrys* species form well-supported lineages highlights the power that NGS-based data holds for resolving species boundaries, particularly in groups with complex evolutionary histories.

References

Abbott R, Albach D, Ansell S, Arntzen JW, Baird SJ, Bierne N, Boughman J, Brelsford A, Buerkle CA, Buggs R, Butlin RK, Dieckmann U, Eroukhmanoff F, Grill A, Cahan SH, Hermansen JS, Hewitt G, Hudson AG, Jiggins C, Jones J, Keller B, Marczewski T, Mallet J, Martinez-Rodriguez P, Möst M, Mullen S, Nichols R, Nolte AW, Parisod C, Pfennig K, Rice AM, Ritchie MG, Seifert B, Smadja CM, Stelkens R, Szymura JM, Väinölä R, Wolf JBW, Zinner D. 2013. Hybridization and speciation. *Journal of Evolutionary Biology* 26: 229–246.

Althoff DM, Gitzendanner MA, Segraves KA. 2007. The Utility of Amplified Fragment Length Polymorphisms in Phylogenetics: A Comparison of Homology within and between Genomes. *Systematic Biology* 56: 477–484.

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA. 2008. Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *PloS One* 3: e3376.

Barrier M, Baldwin BG, Robichaux, RH, Purugganan MD. 1999. Interspecific hybrid ancestry of a plant adaptive radiation: alloploidy of the Hawaian silversword alliance (Asteraceae) inferred from floral homeotic gene duplication. *Molecular Biology and Evolution* 16: 1105–1113.

Bateman RM, Hollingsworth PM, Preston J, Yi-Bo L, Pridgeon AM, Chase MW. 2003. Molecular phylogenetics and evolution of Orchidinae and selected Habenariinae (Orchidaceae). *Botanical Journal of the Linnean Society* 142: 1–40.

Bateman RM, Hollingsworth PM, Preston J, Yi-Bo L, Pridgeon AM, Chase MW. 2003. Molecular phylogenetics and evolution of Orchidinae and selected Habenariinae (Orchidaceae). *Botanical Journal of the Linnean Society* 142: 1–40.

Bateman RM, Bradshaw E, Devey DS, Glover BJ, Malmgren S, Sramko G, Murphy Thomas M, Rudall PJ. 2011. Species arguments: clarifying competing concepts of species delimitation in the pseudo-copulatory orchid genus *Ophrys*. *Botanical Journal of the Linnean Society* 165: 336–347.

Bateman RM, Sramkó G, Paun O. 2018. Integrating restriction site-associated DNA sequencing (RAD-seq) with morphological cladistic analysis clarifies evolutionary relationships among major species groups of bee orchids. *Annals of Botany* 121: 85–105.

Breitkopf H, Schlüter PM, Xu S, Schiestl FP, Cozzolino S, Scopece G. 2013. Pollinator shifts between *Ophrys sphegodes* populations: might adaptation to different pollinators drive population divergence? *Journal of Evolutionary Biology* 26: 2197–2208.

Breitkopf H, Onstein RE, Cafasso D, Schlüter PM, Cozzolino S. 2015. Multiple shifts to different pollinators fuelled rapid diversification in sexually deceptive *Ophrys* orchids. *New Phytologist* 207: 377–389.

Bruneaux M, Johnston SE, Herczeg G, Merilä J, Primmer CR, Vasemägi A. 2013. Molecular evolutionary and population genomic analysis of the nine-spined stickleback using a modified restriction-site-associated DNA tag approach. *Molecular Ecology* 22: 565–582.

Cafasso D, Widmer A, Cozzolino S. 2004. Chloroplast DNA inheritance in the orchid *Anacamptis palustris* using single-seed polymerase chain reaction. *Journal of Heredity* 96: 66–70.

Carstens BC, Pelletier TA, Reid NM, Satler JD. 2013. How to fail at species delimitation. *Molecular Ecology* 22: 4369–4383.

Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA. 2013. Stacks: an analysis tool set for population genomics. *Molecular Ecology* 22: 3124–3140.

Darwin C. 1862. *On the Various Contrivances by Which British and Foreign Orchids are Fertilised by Insects: And on the Good Effect of Intercrossing*. Cambridge (UK): Cambridge Library Collection.

Davey JW, Blaxter ML. 2010. RADSeq: next-generation population genetics. *Briefings in Functional Genomics* 9: 416–423.

Degnan JH, Rosenberg NA. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology & Evolution* 24: 332–340.

Delforge P. 2016. *Guide des orchidées d'Europe, d'Afrique du Nord et du Proche-Orient*. Paris: Delachaux et Niestle.

Devey DS, Bateman RM, Fay MF, Hawkins JA. 2008. Friends or relatives? Phylogenetics and species delimitation in the controversial European orchid genus *Ophrys*. *Annals of Botany* 101: 385–402.

Doyle JJ. 1997. Trees within trees: Genes and species, molecules and morphology. *Systematic Biology* 46: 537–553.

Eaton DAR. 2014. PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. *Bioinformatics* 30: 1844–1849.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* 32: 1792–1797.

Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26: 2460–2461.

Edwards SV. 2009. Is a new and general theory of molecular systematics emerging? *Evolution* 63: 1–19.

Edwards SV, Beerli P. 2000. Perspective: gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies. *Evolution* 54: 1839–1854.

Ellegren H. 2014. Genome sequencing and population genomics in non-model organisms. *Trends in Ecology & Evolution* 29: 51–63.

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PloS One* 6: e19379.

Emerson KJ, Merz CR, Catchen JM, Hohenlohe PA, Cresko WA, Bradshaw WE, Holzapfel CM. 2010. Resolving postglacial phylogeography using high-throughput sequencing. *Proceedings of the National Academy of Sciences USA* 107: 16196–16200.

Evanno G, Regnaut S, Goudet J. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology* 14: 2611–2620.

Feder JL, Egan SP, Nosil P. 2012. The genomics of speciation-with-gene-flow. *Trends in Genetics* 28: 342–350.

Foll M, Gaggiotti OE. 2008. A genome scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics* 180: 977–993.

Gielly L, Taberlet P. 1994. The use of chloroplast DNA to resolve plant phylogenies: noncoding versus rbcL sequences. *Molecular Biology and Evolution* 11: 769–777.

Hernández-León S, Gernandt DS, de la Rosa JAP, Jardón-Barbolla L. 2013. Phylogenetic relationships and species delimitation in *Pinus* section Trifoliae inferrred from plastid DNA. PloS One 8: e70501.

Hipp AL, Eaton DAR, Cavender-Bares J, Fitzek E, Nipper R, Manos PS. 2014. A Framework Phylogeny of the American Oak Clade Based on Sequenced RAD Data. *PloS One* 9: e93975.

Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA. 2010. Population Genomics of Parallel Adaptation in Threespine Stickleback using Sequenced RAD Tags. *PloS Genetics* 6: e1000862.

Huang H, Knowles LL. 2014. Unforeseen Consequences of Excluding Missing Data from Next-Generation Sequences: Simulation Study of RAD Sequences. *Systematic Biology* 65: 357–365.

Knowles LL, Maddison WP. 2002. Statistical phylogeography. *Molecular Ecology* 11: 2623–2635.

Knowles LL. 2009 Statistical phylogeography. *Annual Review of Ecology Evolution and Systematics* 40: 593–612.

Kullenberg B. 1961. Studies in *Ophrys* pollination. *Zoologiska Bidrag Fran Uppsala* 34: 1–340.

Lee KM, Kivelä SM, Ivanov V, Hausmann A, Kaila L, Wahlberg N, Mutanen M. 2018. Information Dropout Patterns in RAD Phylogenomics and a Comparison with Multilocus Sanger Data in a Species-rich Moth Genus. *Systematic Biology* syy029, doi.org/10.1093/sysbio/syy029.

Leaché AD, Banbury BL, Felsenstein J, De Oca, ANM, Stamatakis A. 2015. Short tree, long tree, right tree, wrong tree: new acquisition bias corrections for inferring SNP phylogenies. *Systematic Biology* 64: 1032–1047.

Leigh JW, Bryant D. 2015. PopART: Full-feature software for haplotype network construction. *Methods in Ecology and Evolution* 6: 1110–1116.

Lewis ZA, Shiver AL, Stiffler N, Miller MR, Johnson EA, Selker, EU. 2007. High-Density Detection of Restriction-Site-Associated DNA Markers for Rapid Mapping of Mutated Loci in Neurospora. *Genetics* 177: 1163–1171.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*: 1303.3997.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* 25: 2078–2079.

Lischer HEL, Excoffier L, Heckel G. 2014. Ignoring Heterozygous Sites Biases Phylogenomic Estimates of Divergence Times: Implications for the Evolutionary History of Microtus Voles. *Molecular Biology and Evolution* 31 817–831.

Maddison, Wayne P. 1997. Gene trees in species trees. *Systematic Biology* 46: 523–536.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* 20: 1297–1303.

Meier JI, Marques DA, Mwaiko S, Wagner CE, Excoffier L, Seehausen O. 2017. Ancient hybridization fuels rapid cichlid fish adaptive radiations. *Nature communications* 8: 14363.

Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA. 2007. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research* 17: 240–248.

Neigel JE, Avise JC. 1986. Phylogenetic relationships of mitochondrial DNA under various demographic models of speciation. In: Karlin S, Nevo E. Eds. *Evolutionary Processes and Theory*. New York: Academic Press. 515–534.

Ossowski S, Schneeberger K, Lucas-Lledó JI, Warthmann N, Clark RM, Shaw RG, Weigel D, Lynch M. 2010. The Rate and Molecular Spectrum of Spontaneous Mutations in *Arabidopsis thaliana*. *Science* 327: 92–94.

Paulus HF, Gack C. 1990. Pollinators as prepollinating isolation factors: evolution and speciation in *Ophrys* (Orchidaceae). *Israel Journal of Botany* 39: 43–79.

Petit RJ, Duminil J, Fineschi S, Hampe A, Salvini D, Vendramin GG. 2005. Comparative organization of chloroplast, mitochondrial and nuclear diversity in plant populations. *Molecular Ecology* 14: 689–701.

Posada D, Crandall KA. 2002. The effect of recombination on the accuracy of phylogeny estimation. *Journal of Molecular Evolution* 54: 396–402.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155: 945–959.

Roma L, Cozzolino S, Schlüter PM, Scopece G, Cafasso D. 2018. The complete plastid genomes of *Ophrys iricolor* and *O. sphegodes* (Orchidaceae) and comparative analyses with other orchids. *PloS One* 13: e0204174.

Roure B, Baurain D, Philippe H. 2013. Impact of missing data on phylogenies inferred from empirical phylogenomic data sets. *Molecular Biology and Evolution* 30: 197–214.

Rowe HC, Renaut S, Guggisberg A. 2011. RAD in the realm of next-generation sequencing technologies. *Molecular Ecology* 20: 3499–3502.

Rubin BER, Ree RH, Moreau CS. 2012. Inferring phylogenies from RAD sequence data. *PloS One* 7: e33394.

Sang T, Crawford DJ, Stuessy TF. 1997. Chloroplast DNA phylogeny, reticulate evolution, and biogeography of *Paeonia* (Paeoniaceae). *American Journal of Botany* 84: 1120–1136.

Schlueter PM, Harris SA. 2006. Analysis of multilocus fingerprinting data sets containing missing data. *Molecular Ecology Notes* 6: 569–572.

Schlüter PM, Kohl G, Stuessy TF, Paulus HF. 2007. A screen of low-copy nuclear genes reveals the LFY gene as phylogenetically informative in closely related species of orchids (*Ophrys*). *Taxon* 56: 493–504.

Schlüter PM, Ruas PM, Kohl G, Ruas CF, Stuessy TF, Paulus HF. 2011a. Evidence for progenitor–derivative speciation in sexually deceptive orchids. *Annals of Botany* 108: 895–906.

Schlüter PM, Xu S, Gagliardini V, Whittle E, Shanklin J, Grossniklaus U, Schiestl FP. 2011b. Stearoyl-acyl carrier protein desaturases are associated with floral isolation in sexually deceptive orchids. *Proceedings of the National Academy of Sciences USA* 201013313.

Scopece G, Musacchio A, Widmer A, Cozzolino S. 2007. Patterns of reproductive isolation in Mediterranean deceptive orchids. *Evolution* 61: 2623–2642.

Sedeek KE, Scopece G, Staedler YM, Schönenberger J, Cozzolino S, Schiestl FP, Schlüter PM. 2014. Genic rather than genome-wide differences between sexually deceptive *Ophrys* orchids with different pollinators. *Molecular Ecology* 23: 6192–6205.

Sedeek KE, Whittle E, Guthörl D, Grossniklaus U, Shanklin J, Schlüter PM. 2016. Amino acid change in an orchid desaturase enables mimicry of the pollinator’s sex pheromone. *Current Biology* 26: 1505–1511.

Small RL, Cronn RC, Wendel JF. 2004. Use of nuclear genes for phylogeny reconstruction in plants. *Australian Systematic Botany* 17: 145–170.

Soliva M, Kocyan A, Widmer A. 2001. Molecular phylogenetics of the sexually deceptive orchid genus *Ophrys* (Orchidaceae) based on nuclear and chloroplast DNA sequences. *Molecular Phylogenetics and Evolution* 20: 78–88.

Stölting KN, Nipper R, Lindtke D, Caseys C, Waeber S, Castiglione S, Lexer C. 2013. Genomic scan for single nucleotide polymorphisms reveals patterns of divergence and gene flow between ecologically divergent species. *Molecular Ecology* 22: 842–855.

Vereecken NJ, Cozzolino S, Schiestl FP. 2010. Hybrid floral scent novelty drives pollinator shift in sexually deceptive orchids. *BMC Evolutionary Biology* 10: 103.

Vereecken NJ, Streinzer M, Ayasse M, Spaethe J, Paulus HF, Stoekl J, Cortis P, Schiestl FP. 2011. Integrating past and present studies on *Ophrys* pollination–a comment on Bradshaw et al. *Botanical Journal of the Linnean Society* 165: 329–335.

Wagner CE, Keller I, Wittwer S, Selz OM, Mwaiko S, Greuter L, Sivasundar A, Seehausen O. 2013. Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. *Molecular Ecology* 22: 787–798.

Wang N, Thomson M, Bodles WJA, Crawford RMM, Hunt HV, Featherstone AW, Pellicer J, Buggs RJA. 2013. Genome sequence of dwarf birch (*Betula nana*) and cross-species RAD markers. *Molecular Ecology* 22: 3098–3111.

Widmer A, Lexer C. 2001. Glacial refugia: sanctuaries for allelic richness, but not for gene diversity. *Trends in Ecology & Evolution* 16: 267–269.

Wolfe KH, Sharp PM, Li WH. 1989. Rates of synonymous substitution in plant nuclear genes. *Journal of Molecular Evolution* 29: 208–211.

Xu S, Schlüter PM, Scopece G, Breitkopf H, Gross K, Cozzolino S, Schiestl FP. 2011. Floral isolation is the main reproductive barrier among closely related sexually deceptive orchids. *Evolution* 65: 2606–2620.

Xu S, Schlüter PM. 2015. Modeling the two-locus architecture of divergent pollinator adaptation: how variation in SAD paralogs affects fitness and evolutionary divergence in sexually deceptive orchids. *Ecology and Evolution* 5: 493–502.

Minimum percentage of shared loci	Informative SNPs	Average Bootstrap value	Global F_{ST} value	Plastid SNPs	Mitochondrial SNPs
10%	123080	78.00	0.220	93	132
30%	59435	74.78	0.152	35	53
50%	31272	66.68	0.110	21	35
70%	16710 (253*)	58.76	0.087	6	20
90%	6210	39.01	0.076	3	8

Table 1. Number of informative SNPs, average bootstrap value, global F_{ST} value, number of plastid SNPs and number of mitochondrial SNPs in the different supermatrices built by using different percentage of shared loci. * Number of informative loci after filtering for heterozygous loci.

Figure legends

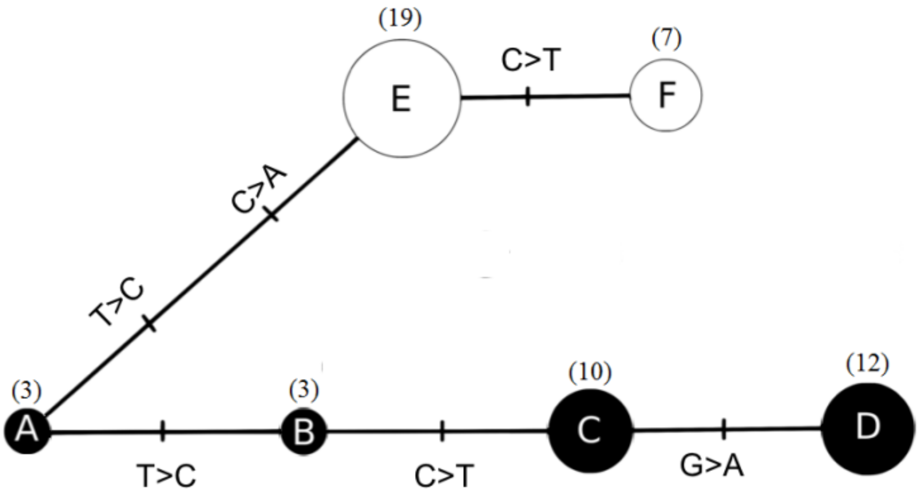


Fig. 1. Statistical parsimony haplotype network based on plastid loci filtered from the GBS data and shared by all individuals. Circle size is proportional to haplotype frequency. Black and with circles indicate the two plastid haplotype lineages identified in the network analysis. In parentheses the number of individuals.

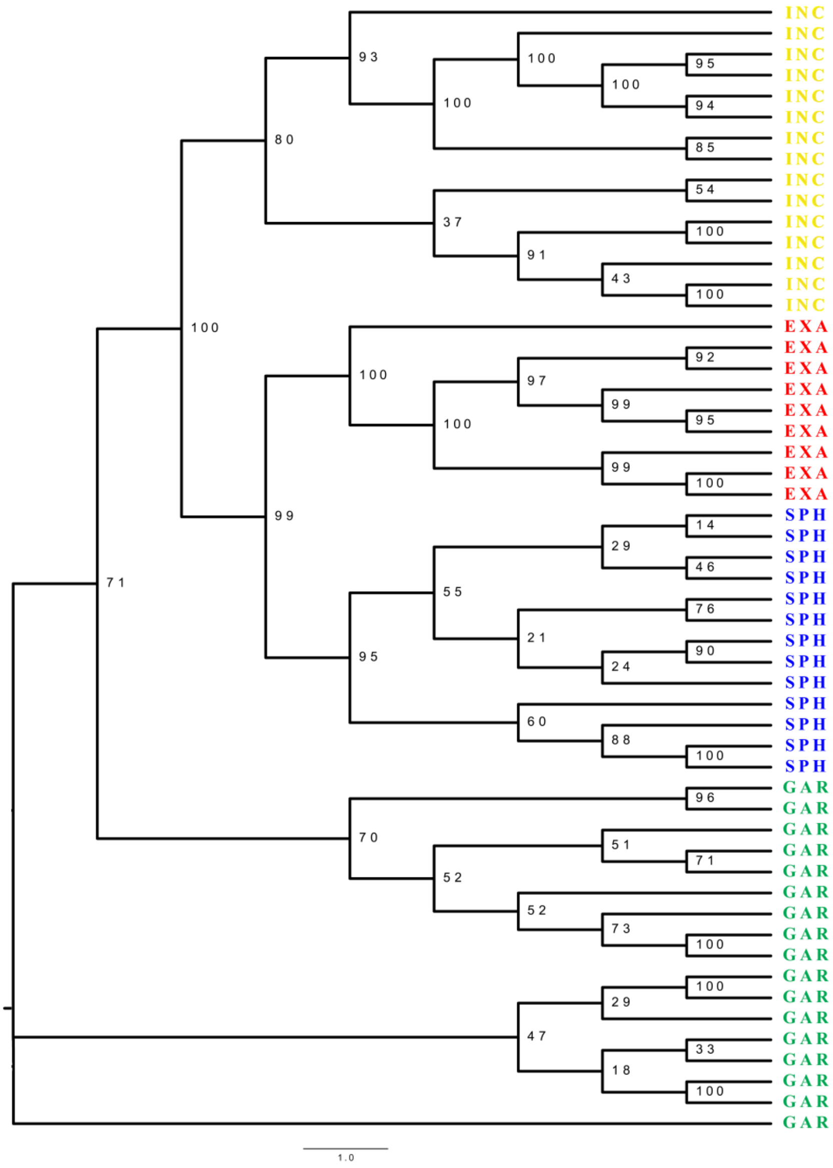


Fig. 2. RAXML tree obtained by using the larger supermatrix (loci shared at least among 30% individuals). EXA = *Ophrys exaltata*, GAR = *O. garganica*, INC = *O. incubacea*, SPH = *O. sphegodes*. Support values are derived from 1 000 bootstrap replicates.

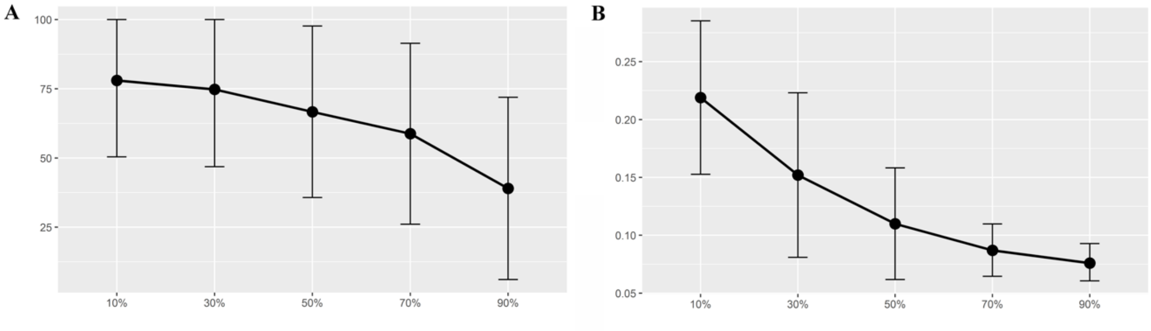


Fig. 3. (A) Average Bootstrap support values in RAXML tree obtained by using the supermatrices with loci shared at least among 10%, 30%, 50%, 70% and 90% of individuals. (B) Global F_{ST} values among the four *Ophrys* species by using the supermatrices with loci shared at least among 10%, 30%, 50%, 70% and 90% of individuals.

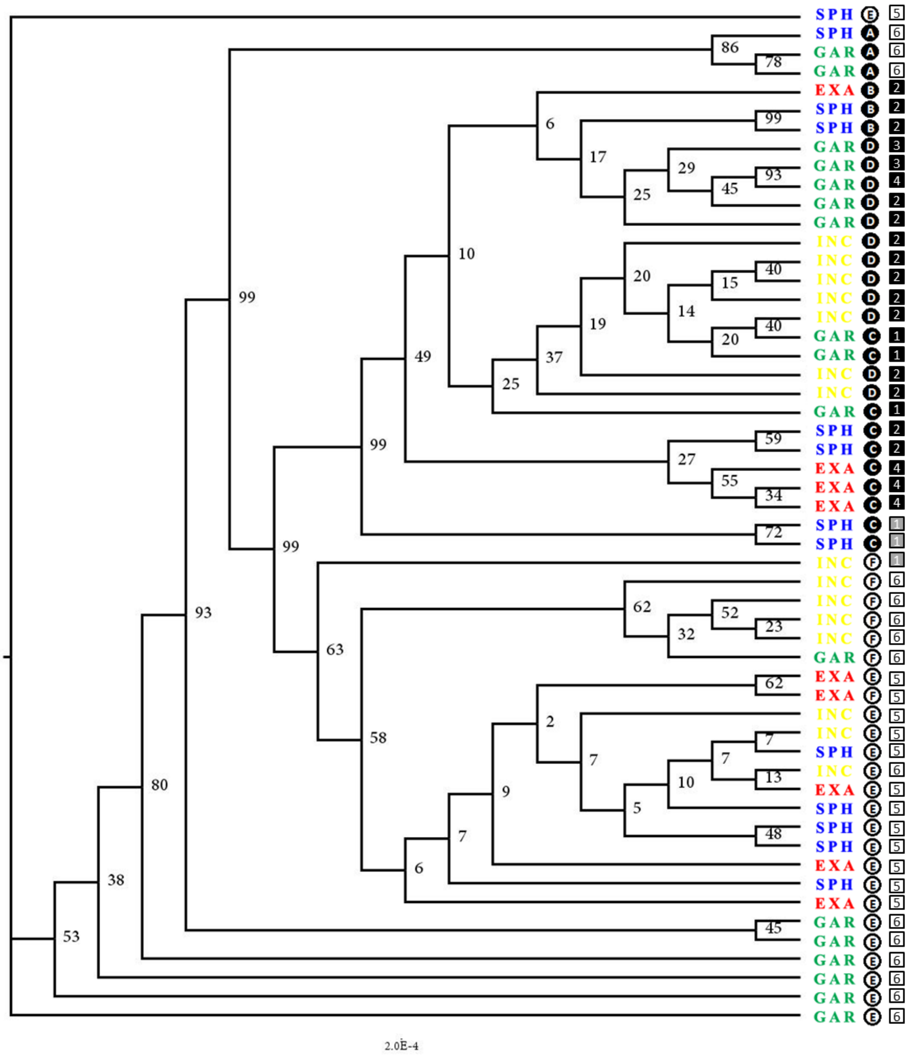


Fig. 4. RAXML tree obtained by using the supermatrix with loci shared at least among 70% individuals and including only homozygous loci. EXA = *Ophrys exaltata*, GAR = *O. garganica*, INC = *O. incubacea*, SPH = *O. sphegodes*. Support values are derived from 1000 bootstrap replicates. Letters in the circles represent plastid haplotypes; numbers in the squares represent mitochondrial haplotypes.

Supplementary material

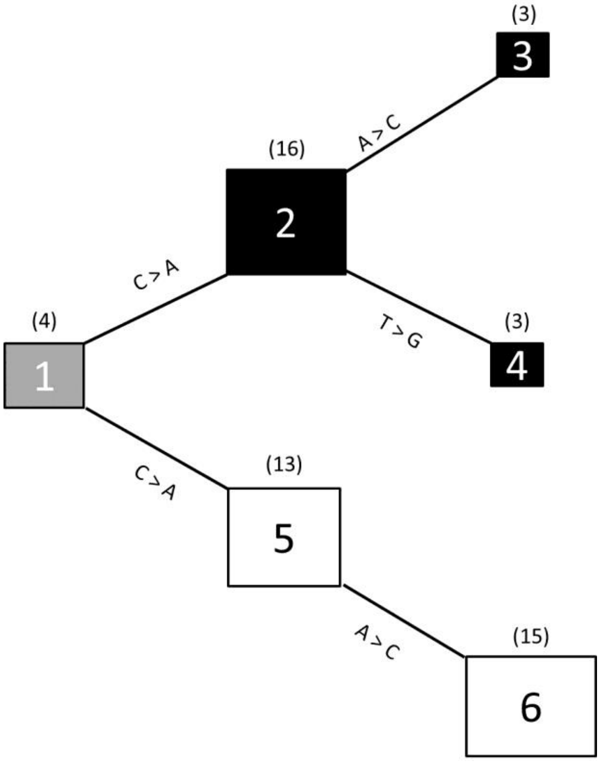


Fig. S1. Statistical parsimony haplotype network based on four mitochondrial loci filtered from the GBS data and shared by all individuals. Square size is proportional to haplotype frequency. In parentheses the number of individuals.

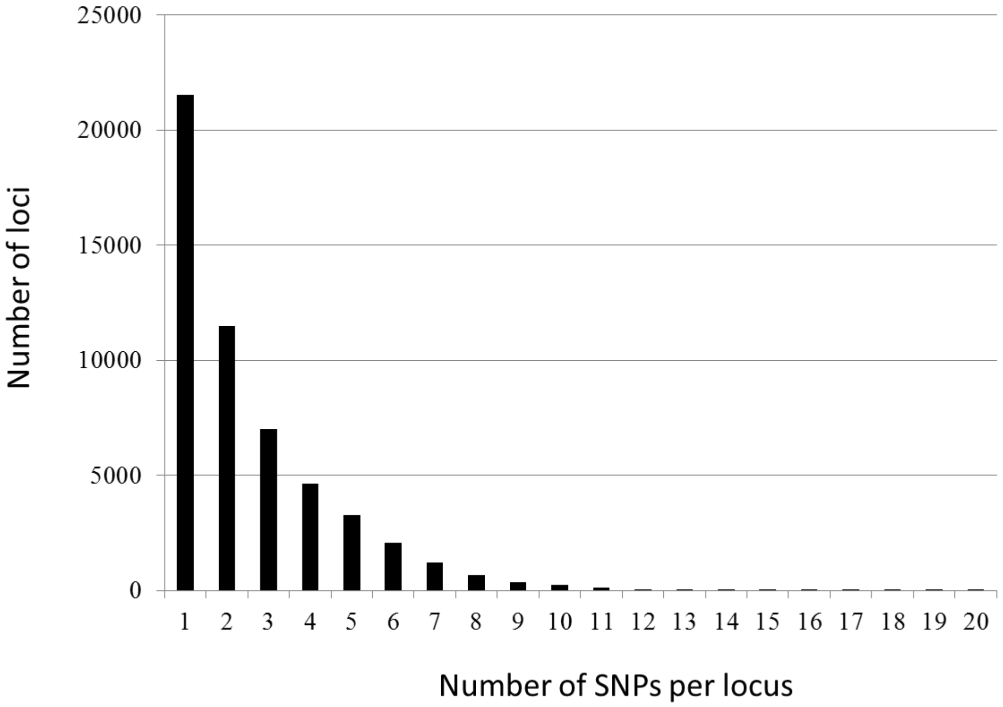


Fig. S2. Distribution of number of SNPs per locus in the 54 individuals from the original dataset from Sedeek et al. (2014).

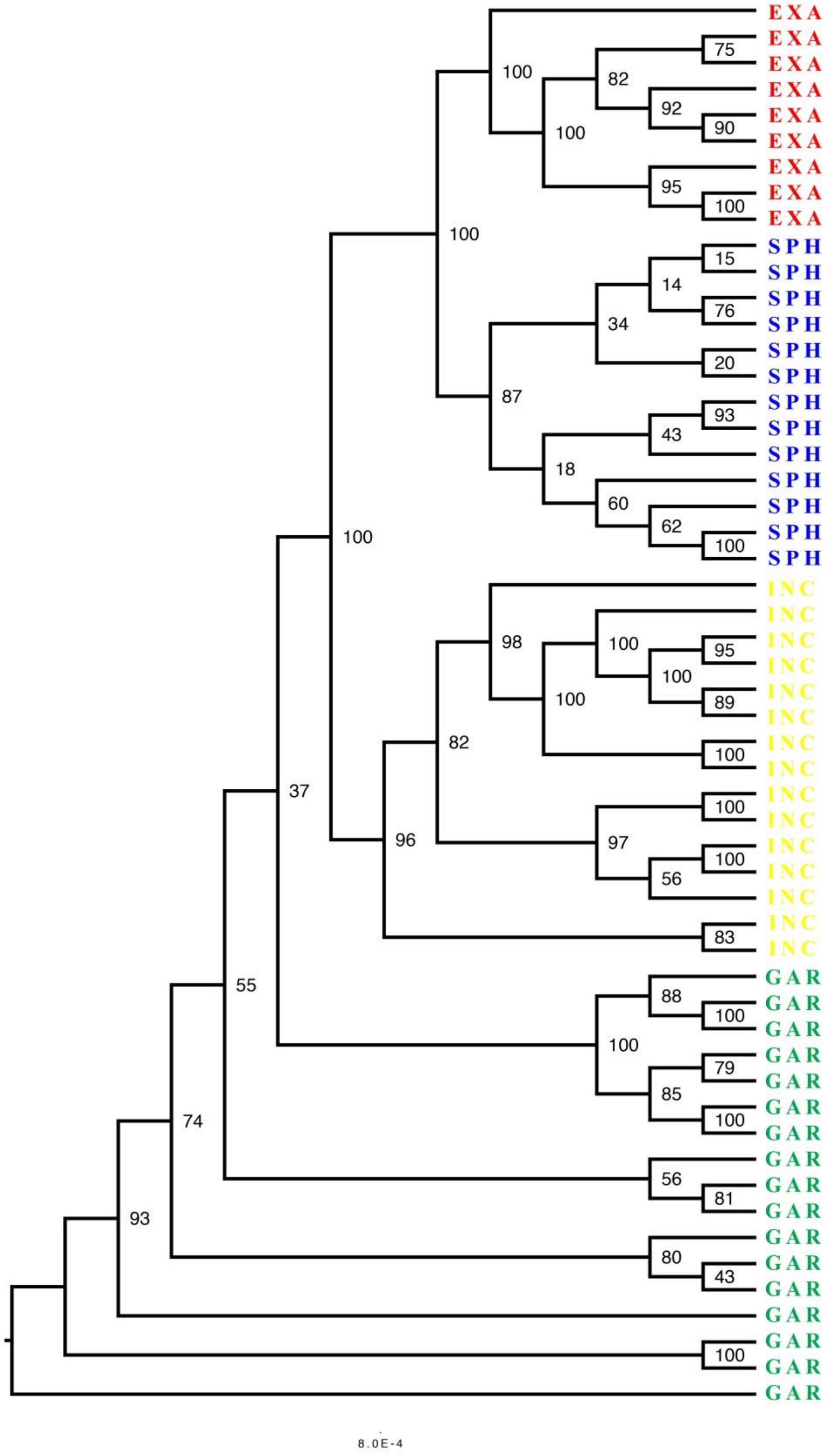


Fig. S3. RAXML tree obtained by using the supermatrix with loci shared at least among 10% individuals). EXA = *Ophrys exaltata*, GAR = *O. garganica*, INC = *O. incubacea*, SPH = *O. sphegodes*. Support values are derived from 1000 bootstrap replicates.

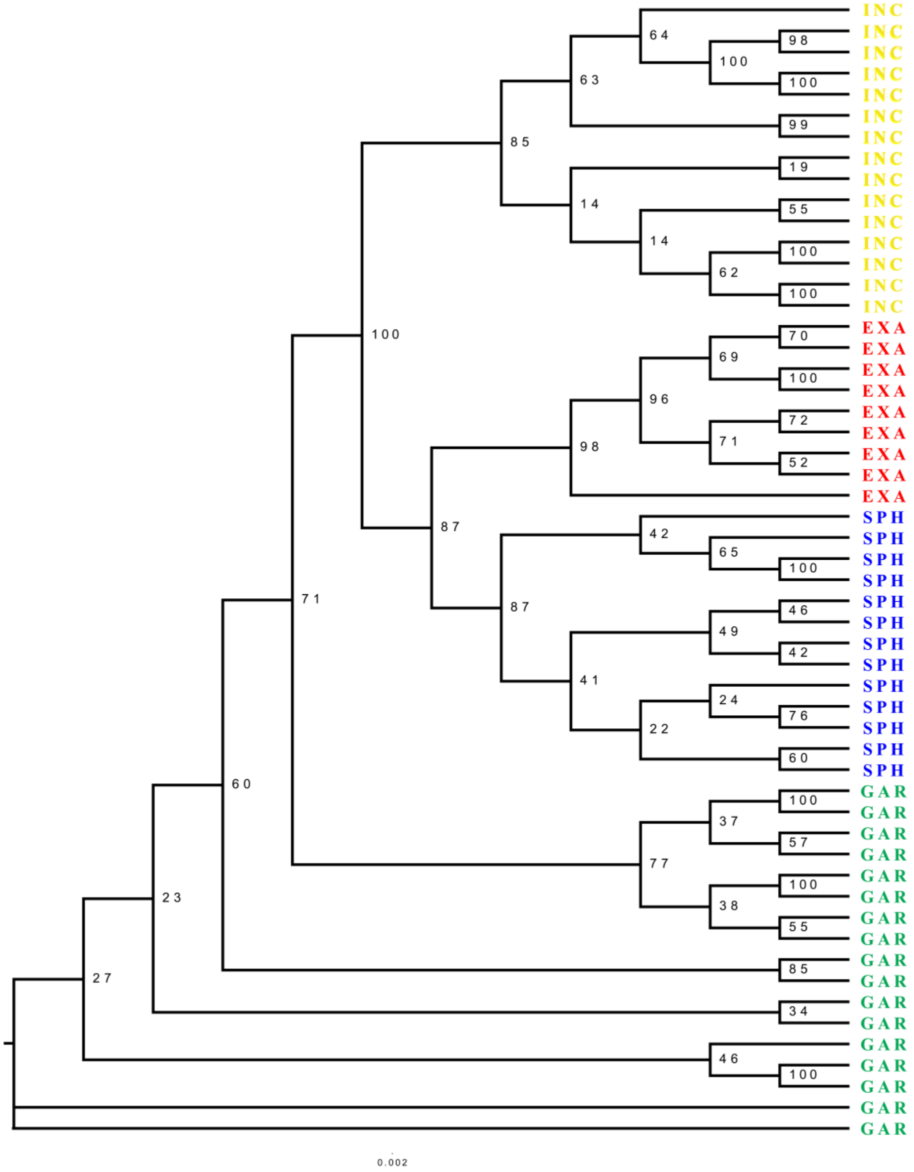


Fig. S4. RAXML tree obtained by using the software RRHS on the larger supermatrix with loci shared at least among 30% individuals. EXA = *Ophrys exaltata*, GAR = *O. garganica*, INC = *O. incubacea*, SPH = *O. sphegodes*. Support values are derived from 1 000 bootstrap replicates.

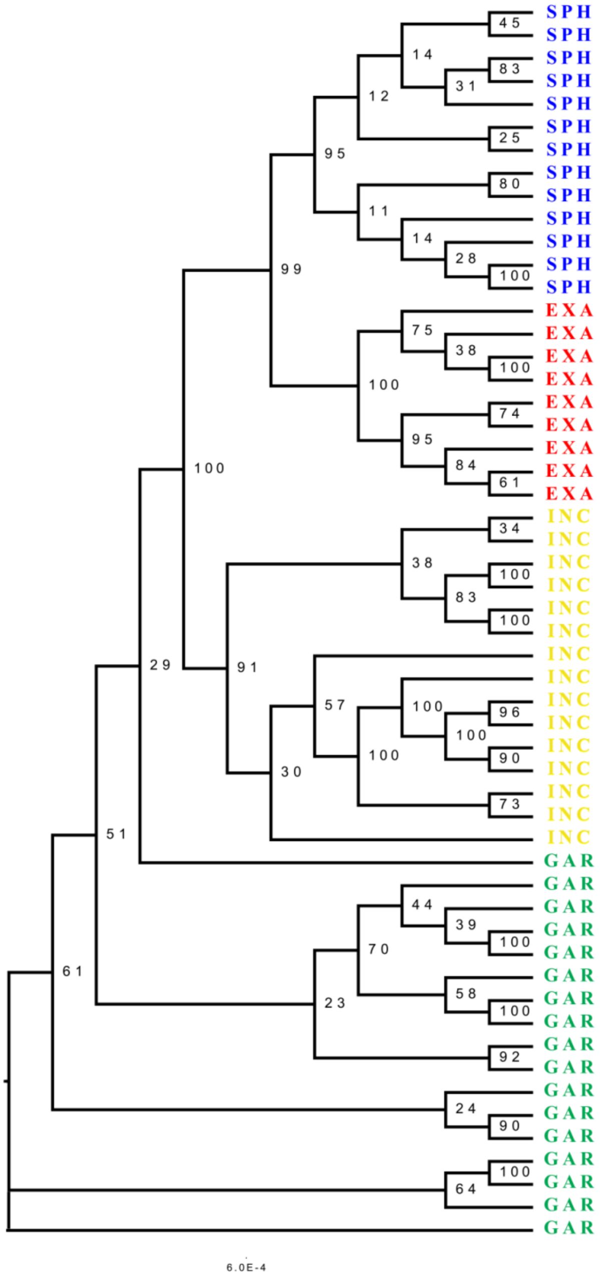


Fig. S5. RAXML tree obtained by using the supermatrix with loci shared at least among 50% individuals). EXA = *Ophrys exaltata*, GAR = *O. garganica*, INC = *O. incubacea*, SPH = *O. sphegodes*. Support values are derived from 1000 bootstrap replicates.

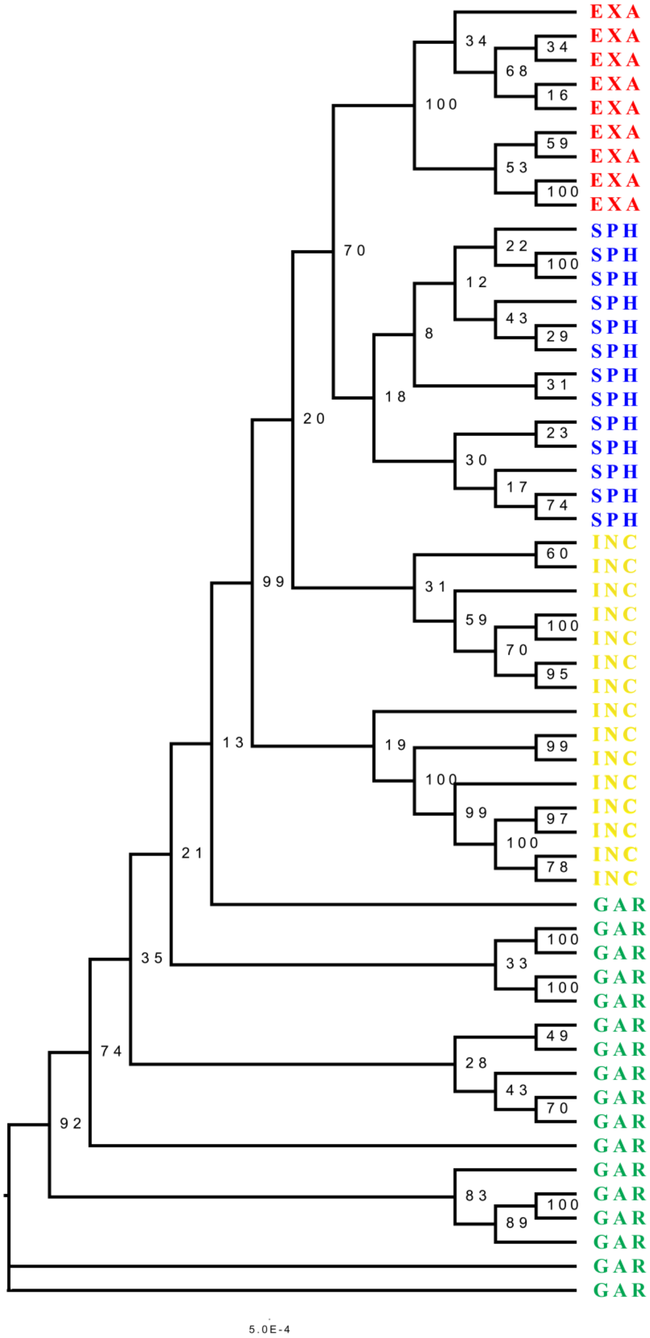


Fig. S6. RAXML tree obtained by using the supermatrix with loci shared at least among 70% individuals. EXA = *Ophrys exaltata*, GAR = *O. garganica*, INC = *O. incubacea*, SPH = *O. sphegodes*. Support values are derived from 1000 bootstrap replicates.

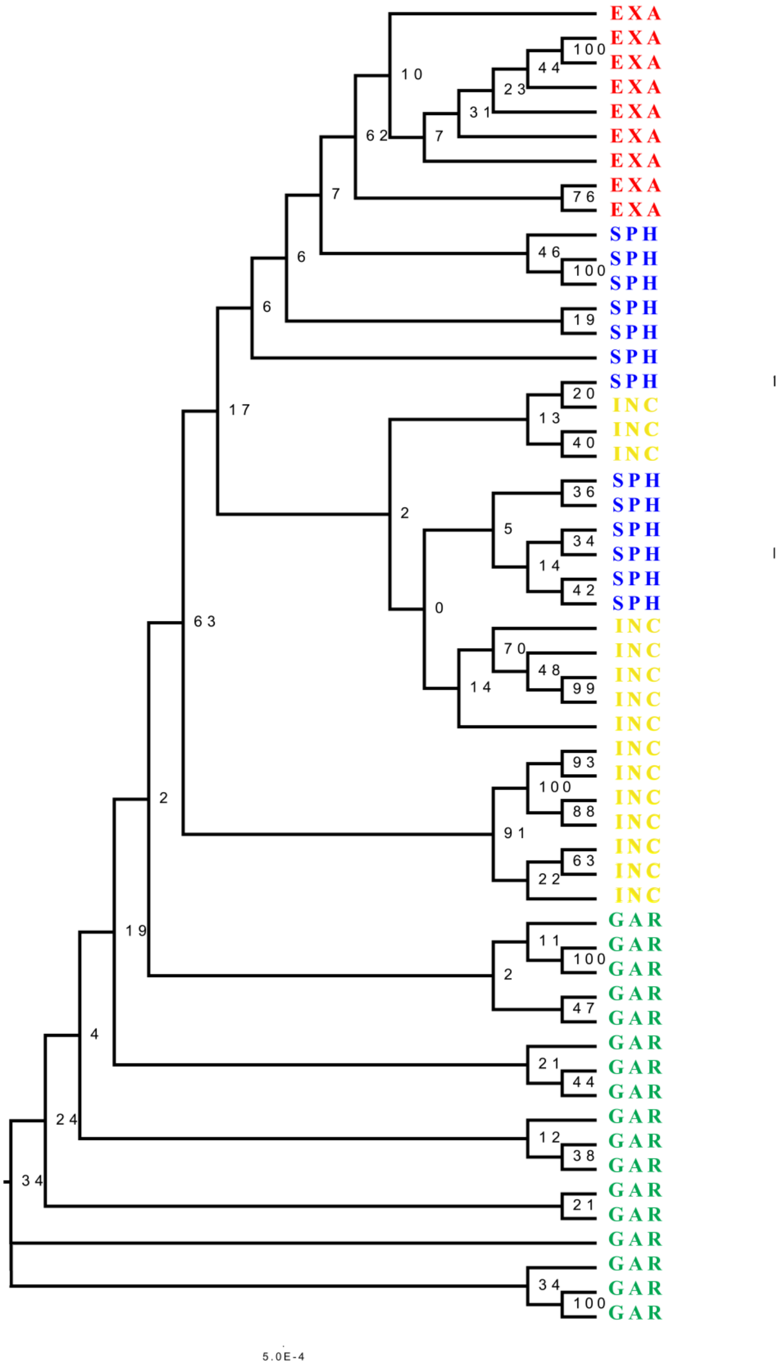


Fig. S7. RAXML tree obtained by using the supermatrix with loci shared at least among 90% individuals. EXA = *Ophrys exaltata*, GAR = *O. garganica*, INC = *O. incubacea*, SPH = *O. sphegodes*. Support values are derived from 1 000 bootstrap replicates.

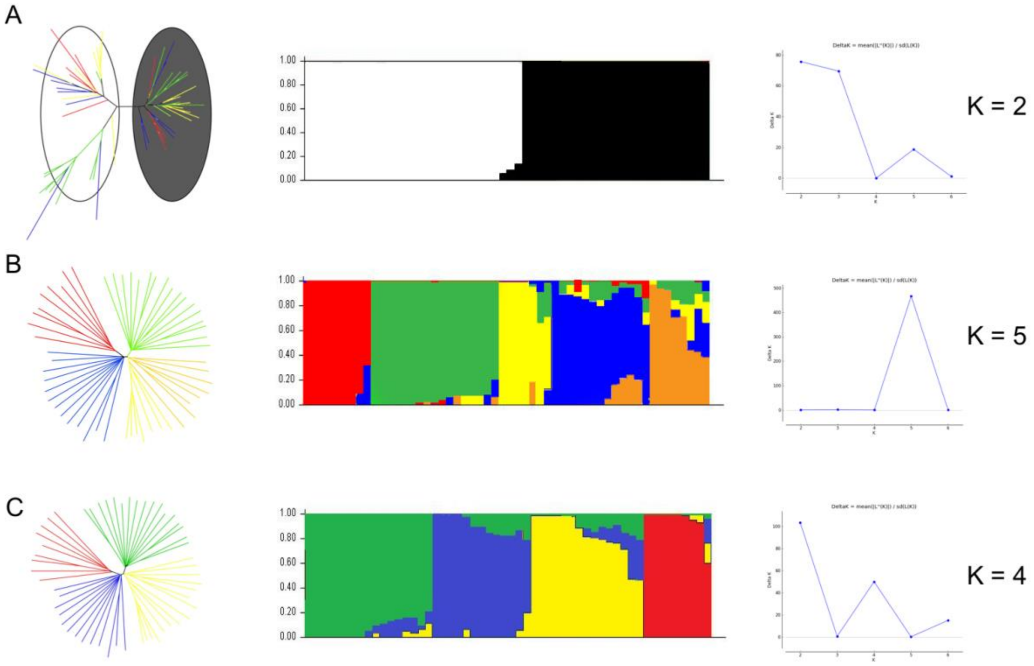


Fig. S8. Neighbor Joining (NJ) tree, Bayesian assignment bar graph and Plot of delta K values from the Structure analyses based on (A) the supermatrix with loci shared at least among 70% individuals; (B) the supermatrix with loci shared at least among 30% individuals; (C) on the supermatrix with loci shared at least among 30% individuals after filtering plastid and mitochondrial reads. Red = *Ophrys exaltata*; Green = *O. garganica*; Yellow/Orange = *O. incubacea*; Blue = *O. sphegodes*. Grey and white circles represent the two plastid haplotype lineages identified in the network analysis.

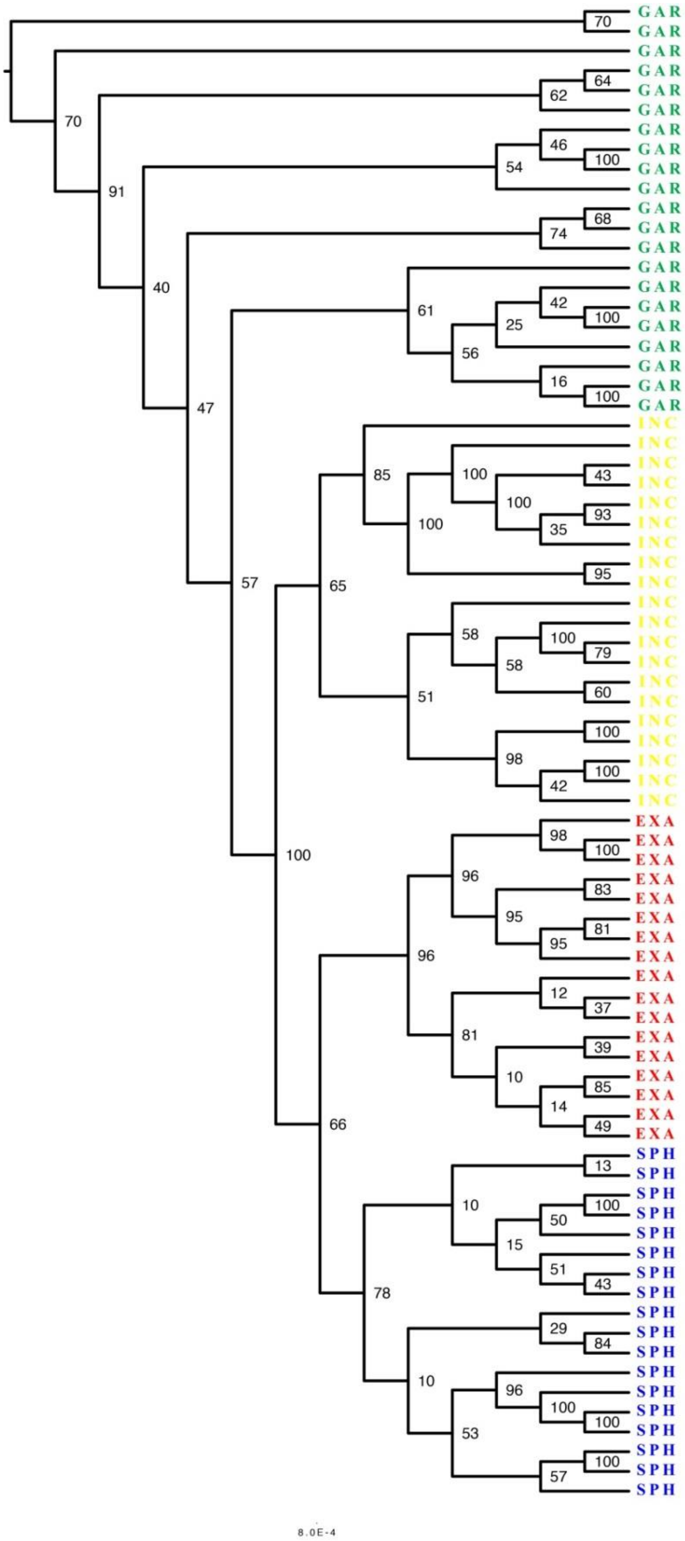


Fig. S9. RAXML tree obtained by using the larger supermatrix composed by accession with at least 500000 number of reads (loci shared at least among 30% individuals). EXA = *Ophrys exaltata*, GAR = *O. garganica*, INC = *O. incubacea*, SPH = *O. sphegodes*. Support values are derived from 1 000 bootstrap replicates.

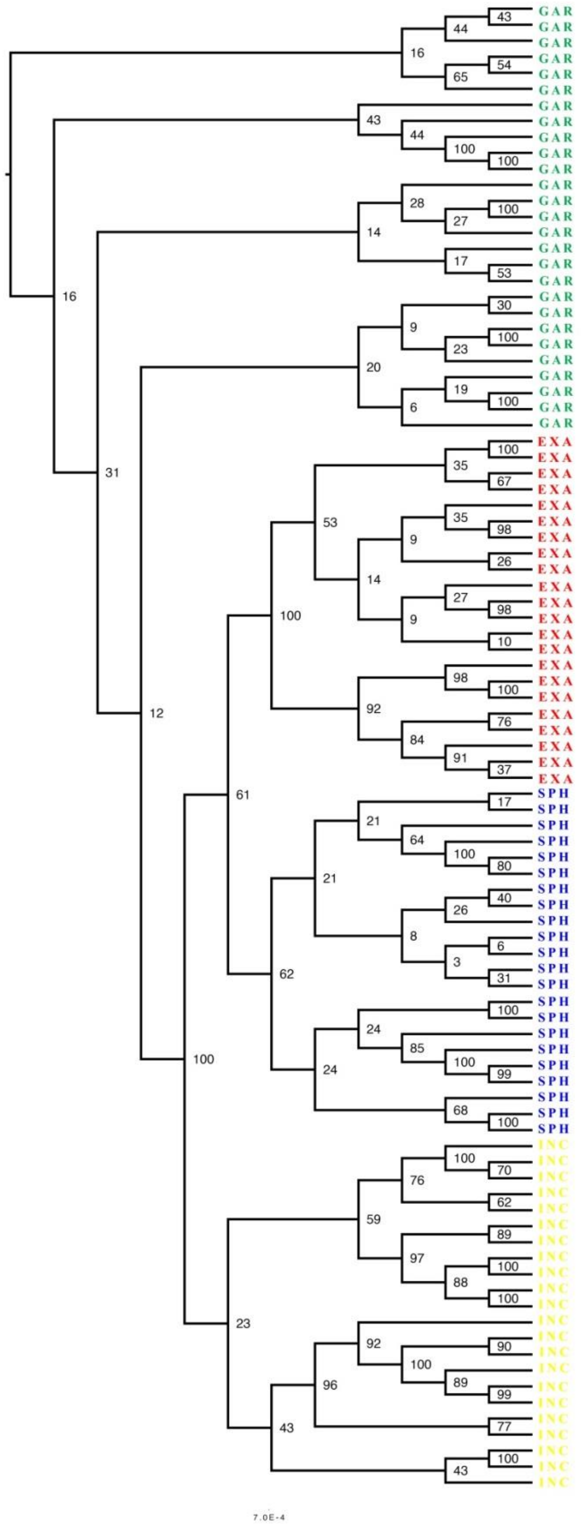


Fig. S10. RAXML tree obtained by using the larger supermatrix composed by accessions with at least 300000 number of reads (loci shared at least among 30% individuals). EXA = *Ophrys exaltata*, GAR = *O. garganica*, INC = *O. incubacea*, SPH = *O. sphegodes*. Support values are derived from 1 000 bootstrap replicates.