Treball de Fi de Grau

Identificació i estudi filogenètic d'espècies properes mitjançant la tècnica de DNA Barcoding

Marc Solé Estragués

Grau en Biotecnología

Tutor: Josep Bau Macià

Vic, Juny de 2019

FACULTAT
DE CIÈNCIES
I TECNOLOGIA
UVIC | UVIC·UCC

RESUM

Títol: *Identificació i estudi filogenètic d'espècies properes mitjançant la tècnica de DNA Barcoding*
Paraules Clau: *Eilema*, Barcode of Life, PCR, seqüenciació, Cytochrome c oxidase subunit I (COI), Alignment, R, Markdown, filogènia
Autor: Marc Solé Estragués
Tutors: Josep Bau Macià
Data: Juny de 2019

La identificació taxonòmica d'espècies properes requereix en ocasions la intervenció d'un taxònom expert en el grup zoològic en qüestió, i sovint hi ha de dedicar un esforç important.
Actualment, s'han establert diferents tècniques d'anàlisi de seqüència genètica que permeten, no únicament, identificar un individu amb fiabilitat sinó també realitzar estudis taxonòmics i poblacionals amb un esforç relativament petit.

La tècnica anomenada *DNA barcoding*, o *Barcode of Life*, es basa en l'amplificació i seqüenciació d'un o pocs gens que s' utilitzen com a referència (ex. el gen mitocondrial COI) per identificar els individus d'una espècie determinada a partir de la informació que ens aporten. D'aquesta manera podem dir que s'obté un codi de barres amb el que podem obtenir una representació gràfica de la seqüència.

Així doncs, aquest treball ha consistit en utilitzar aquests conceptes com a base per aplicar la tècnica del *Barcoding* per donar solidesa i contribuir a l'estudi taxonòmic morfològic que s'està duent a terme en insectes lepidòpters del gènere *Eilema*. Això implica que s'ha hagut d'optimitzar aquesta tècnica per obtenir bons resultats a nivell de seqüenciació d'individus d'aquest gènere i s'ha analitzat les dificultats d'obtenir resultats en individus que es troben en diferents estats de conservació. També s'ha estudiat les condicions adequades per una correcta amplificació i seqüenciació del gen COI, i també les condicions de diferents gens somàtics que ens aportaran informació complementaria per l'estudi taxonòmic.

S'han utilitzat tant gens mitocondrials com somàtics per obtenir la informació necessària per realitzar un estudi taxonòmic de les mostres. La informació obtinguda a partir de la seqüenciació s'ha analitzat utilitzant diversos programes informàtics que han permès la construcció dels *barcodes* en forma de seqüències consens, i la construcció d'arbres filogenètics que ens permeten estudiar d'una forma visual la proximitat genètica entre les diferents espècies estudiades.

ABSTRACT

Title: *Identification and phylogenetic study of near species by the DNA Barcoding technique*
Keywords: *Eilema*, Barcode of Life, PCR, sequencing, Cytochrome c oxidase subunit I (COI), Alignment, R, Markdown, phylogeny
Author: Marc Solé Estragués
Tutors: Josep Bau Macià
Date: June de 2019

The taxonomic identification of nearby species sometimes requires the intervention of an expert taxon in the zoological group in question, and an important effort must often be devoted to identify them correctly. At the moment different genetic sequence analysis techniques have been established that allow, not only, to identify an individual with reliability but also to carry out taxonomic and population studies with a relatively small effort.

The technique called DNA *Barcoding*, or Barcode of Life, is based on the amplification and sequencing of one or a few genes that are used as a reference (eg the mitochondrial gene COI) to identify individuals of a particular species based on the information they provide. In this way we can say that you get a barcode with which we can obtain a graphic representation of the sequence.

Therefore, this work consisted in using these concepts as the basis for applying the technique of *Barcoding* to give solidity and contribute to the morphological taxonomic study that is being carried out in lepidopteran insects of the genus *Eilema*. This implies that this technique has had to be optimized to obtain good results at the level of sequencing of individuals of this genus and the difficulty of obtaining results in individuals that are in different conservation states has also been studied. We also studied the appropriate conditions for the correct amplification and sequencing of the COI gene, as well as the conditions of different somatic genes that will provide complementary information for the taxonomic study.

Both mitochondrial and somatic genes have been used to obtain the information necessary to perform a taxonomic study of the samples. The information obtained from the sequencing was analyzed using various computer programs that allowed the construction of barcodes in the form of consensus sequences, and the construction of phylogenetic trees that allow us to study genetic proximity between the different species studied.

FACULTAT
DE CIÈNCIES
I TECNOLOGIA
UVIC | UVIC·UCC

Índex

1. Introducció

1.1 El gènere *Eilema* de la família de lepidòpters Erebidae

El gènere *Eilema* és un gènere de la família de lepidòpters Erebidae, anteriorment englobat en la família Arctiidae. Aquesta és una gran família de papallones que es poden localitzar en totes les regions geogràfiques del món: des de l'equador fins a la zona boreal i amb la màxima diversitat en la Amèrica meridional i Àfrica. Consta d'un gran nombre d'espècies que es creu que està entre les 10.000 i les 12.000. D'aquest gran nombre només 65 es troben a la Península Ibèrica i illes Balears. És un grup molt monofilètic, i està integrat per papallones de totes mides i de colors molt variats. Tant poden tenir activitat diürna com crepuscular, però la majoria són nocturnes[1].

La família Erebidae té les següents subfamílies[2] :

1. Lithosiinae: Agrupa les espècies caracteritzades en què les larves posseeixen unes mandíbules amb una base molar molt ampla, que la utilitzen per macerar líquens, fongs, i algues que constitueixen la seva dieta. Les ales manquen ocels (taques arrodonides multicolors que aparenten ser ulls), les ales anteriors són molt estretes i allargades amb un marge exterior molt curt, mentre que les posteriors són curtes però proporcionalment molt grans, fent el doble d'amplada que les anteriors. En repòs les ales anteriors queden adherides longitudinalment al cos, amb les posteriors plegades sota elles.
   Tenen un cap ample, antenes filiformes, espiritrompa ben desenvolupada i palps prims. El seu cos sol ser allargat i esvelt i en les femelles més curt.
   Aquesta subfamília es localitza principalment (un miler d'espècies) als Tròpics, l'Amèrica Central, Sud-Amèrica i el sud-est asiàtic. D'aquestes només 31 agrupades en 11 gèneres es troben a la Península Ibèrica (11 a les Illes Balears).

2. Syntominae: Formada per papallones de talla mitjana o petita i amb el cos força gran, ales anteriors allargades i triangulars amb àpex arrodonit i el marge extern obliquo. Les ales posteriors són petites amb angle anal arrodonit i marge anal molt curt, no té la vena Sc (sinus coronari) i tenen una única vena anal. En repòs les ales queden adherides al cos en forma de teulada.
   Els imagos tenen la espiritrompa ben desenvolupada, ulls grans, palps curts i antenes filiformes, ciliades o doblement pectinades. Tenen potes curtes amb les tíbies posteriors amb quatre o tres esperons.
   Són d'activitat diürna, amb tendència a aproximar-se a flors i a la llum artificial. Consta de 1.200 espècies repartides principalment en regions tropicals. 6 es troben a Europa, tres de les quals a la Península Ibèrica, no se'n troba cap a les balears.

3. Arctiinae: Papallones de talla mitjana amb el cos robust i cobert de pels. Els mascles tenen antenes doblement pectinades i les de les femelles són filiformes: Tenen palps curts,

espiritrompa ben desenvolupada, ulls glabres i ocels presents. Presenten ales anteriors més o menys allargades, àpex arrodonit amb el marge extern sempre més curt que el marge dorsal; ales posteriors relativament amples i amb un angle anal molt arrodonit. Coloració bigarrada de colors vius, sobretot vermell, groc i taronja. En repòs les ales posteriors s'amaguen sotes les anteriors i queden les dues en forma de teules.
Els imagos són d'activitat principalment nocturna. Les erugues estan recobertes de pel i són molt ràpides de moviment. En tot el món existeixen unes 4.400 espècies pertanyents a aquesta subfamília, la meitat es troba a l'Amèrica Central i del Sud. 31 d'aquestes, pertanyents a 23 gèneres, es troben a la Península Ibèrica, 7 es troben a les Illes Balears.

El gènere *Eilema* [1819] ha anat patint diverses modificacions en els últims anys, en un principi estaven incloses en la subfamília Lithosiinae, però el 2011 es va proposar moure'l a la subfamília Arctiinae, i totes les espècies a altres gèneres com ara el *Katha* [3]. En aquest treball s'ha decidit utilitzar la classificació antiga ja que encara no s'ha adoptat formalment aquesta nova classificació i molts estudis continuen utilitzant la existent.
Aquest gènere està format per les següents espècies i subespècies [4,5]:

E. depressa (Esper, 1787)
E. griseola (Hübner, 1803)
E. lurideola (Zincken, 1817)
E. complana (Linnaeus, 1758)
    E. complana iberica (Mentzer 1980)
E. pseudocomplana (Daniel, 1939)
E. caniola (Hübner, 1808)
    E. caniola torstenii (Mentzer 1980)
    E. caniola gibrati (Oberthür, 1922)
E. palliatella (Scopoli, 1763)
E. pygmaeola (Doubleday, 1847)
    E. pygmaeola pallifrons (Zeller, 1847)

E. marcida (Mann, 1859)
E. predotae (Schawerda, 1927)
E. lutarella (Linnaeus, 1758)
    E. lutarella luqueti (Leraut 2006)
E. sororcula (Hufnagel, 1766)
E. uniola (Rambur, 1866)
E. interpositella (Strand, 1920)
E. albicosta (Rogenhofer, 1894)
    E. albicosta witti (Kobes, 1993)
E. rungsi (Toulgoët, 1960)
E. bipuncta (Hübner, 1824)

1.2 DNA barcoding: The Barcode of Life

La gran diversitat de formes de vida que tenim al nostre planeta fa que la seva classificació sigui una tasca molt complexa. Des dels temps immemorials l'ésser humà s'ha dedicat a trobar una lògica en les formes de vida que ens envolten. No va ser fins l'aparició de Carl Von Linné (Systema Naturae, 1735) que no es va classificar de forma clara i científica cada ésser viu. El concepte d'espècie va néixer doncs com una forma senzilla de classificar aquests organismes principalment per la seva aparença fenotípica. A mesura que avançaven els descobriments científics aquest concepte va esdevenir més complex. No ens podem basar en només la definició del que és una espècie: "Conjunt d'individus que tenen característiques semblants i que són capaces de reproduir-se entre elles i tenir descendència fèrtil". Ja que el problema amb aquesta definició és que no és fàcil de comprovar de manera clara i inequívoca, a més és una definició molt simple per a un context tant complex.
La plasticitat fenotípica i la variabilitat genètica del gènere *Eilema* implica la possibilitat d'una incorrecte identificació si s'utilitzen només els caràcters morfològics de les mostres. A més com que molts dels detalls necessaris per identificar l'espècie depenen de l'edat i sexe fa que molts individus no puguin ser identificats. [6]. Això fa que molts taxònoms experts tinguin dificultats a l'hora d'identificar correctament aquestes espècies. Per aquest motiu s'ha desenvolupat el mètode d'identificació del *Barcoding*, un mètode que facilita la identificació d'espècies, ja que ens permet distingir espècies no per la seva aparença fenotípica, sinó per la seva similitud genètica.

El terme "DNA *barcoding*" no és un concepte nou, ja es va utilitzar en un estudi del 1993[7] que no va rebre gaire atenció per part de la comunitat científica. Aquest mètode va néixer a partir de la idea de seqüenciar i comparar alguns gens específics que tenen tots els éssers vius, però que degut al temps d'evolució han patit més o menys mutacions (rati de mutació). A partir d'aquest concepte es poden mesurar doncs les distàncies entre espècies utilitzant diversos algoritmes que calculen les similituds genètiques entre les seqüències i ens donen uns valors matemàtics de la proximitat entre espècies. Podem utilitzar després aquest valors per construir un arbre filogenètic on es veu de forma clara la igualtat entre les diverses espècies estudiades.
Per tant aquests fragments genètics es poden considerar figurativament com a codis de barres útils per identificar cada espècie.

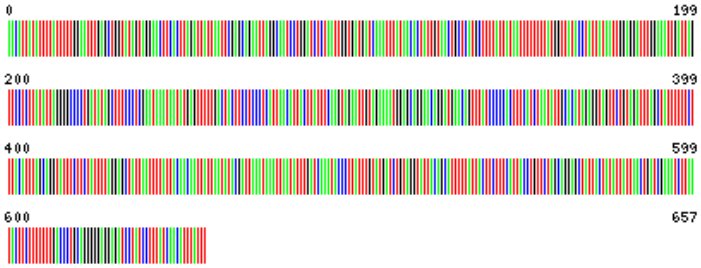


Figura 1 *Barcode il·lustratiu obtingut de BOLD d'una Eilema pygmaeola pallifrons (mostra MSE_16)*[8]

A partir d'aquesta concepte es va crear el Consortium for Barcode of Life (CBOL) que va acordar que els *barcodes* haurien de ser domini públic i s'haurien d'arxivar en una base de dades global i organitzar-los per espècies [9].

També es van acordar una sèrie de requeriments que han d'acompanyar cada *Barcode*:

1- Cada mostra ha de tenir un identificador específic que ens permeti col·locar-lo en un catàleg online i públic juntament amb les seves metadades.
2- Incloure el nom de l'espècie o un nom provisional per espècies no publicades. Això permetrà relacionar la mostra a altres registres d'altres bases de dades.
3- Incloure el codi de país utilitzant el format que fa servir GenBank.
4- El *barcode* ha de provenir d'una regió gènica acceptada pel CBOL.
5- La qualitat ha de ser alta i ha de comprendre com a mínim un 75% de la regió del gen.
6- Incloure el nom de la regió utilitzada.
7- Associar *traces* de la seqüència *forward* i *reverse* penjats a *NCBI Trace Archive* o a l'*Ensembl Trace Server*.
8- Incloure les seqüències i els noms dels *primers* utilitzats, tant *forward* com *reverse*.[9]

Un dels repositoris online mencionat en el punt 1 és el *BOLD: The Barcode of Life Data System* (www.barcodinglife.org).

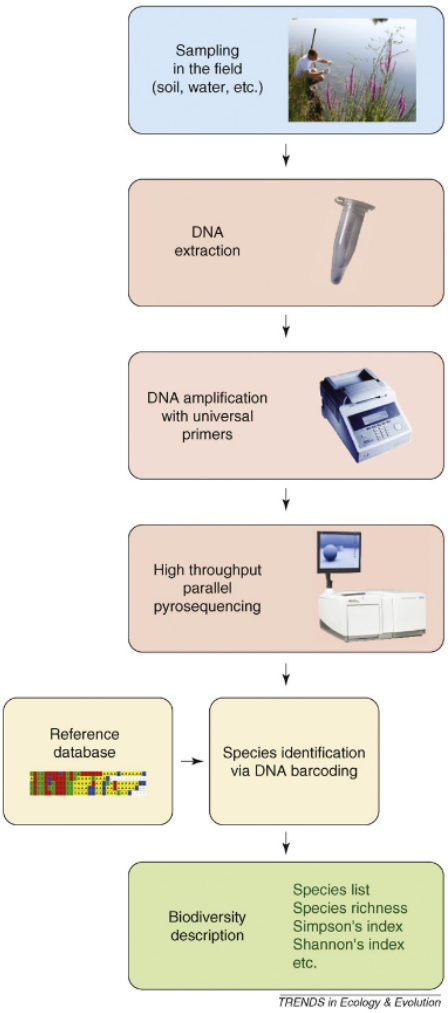


Figura 2 *Metodologia per analitzar biodiversitat de mostres del medi mitjançant barcoding i Next Generation Sequencing. Utilitzant una base de dades de DNA com a referència s'identifica l'organisme d'on provenen.*

Aquesta web és un tauler de treball que vol ajudar en l'adquisició, emmagatzematge, anàlisi i publicació dels registres de *barcodes*. Mitjançant l'assemblatge de dades moleculars, morfològiques i de distribució, cosa que facilita l'anàlisi bioinformàtic i la construcció de *barcodes* de qualitat que compleixin els estàndards establerts [10].
La informació que aporten els *barcodes* es pot utilitzar per millorar les investigacions sobre patrons de la biodiversitat i realitzar nous estudis. Per exemple a partir d'aquestes dades es poden agrupar individus a presumptes espècies anomenades "*operational taxonòmic units*" (OTUs). Aquests OTUs consisteixen en clústers d'individus de la mateixa espècie que

presenten una major similitud genètica entre ells que entre els altres individus de l'espècie. Això permet desenvolupar una sèrie d'algoritmes (ABGD, CROP, GMYC, jMOTU) que generen un registre taxonòmic persistent a nivell d'espècies basat en l'anàlisi de patrons de variació de nucleòtids d'un *barcode*. Aquest *framework* és el que utilitza BOLD per classificar individus semblant d'espècies en BINs (*Barcode Index Number*) [11].

Els principals avantatges del *Barcoding* respecte a altres mètodes són la seva rapidesa i el fet que és un mètode bastant barat. Degut al desenvolupament de noves tecnologies de seqüenciació que redueixen dràsticament el temps necessari per obtenir una seqüència comparat amb el mètode clàssic de seqüenciació mitjançant electroforesi capil·lar s'ha obert pas a la utilització del *DNA-barcoding* per estudis a gran escala. A més cada vegada més bases de dades accepten *barcodes* alguns exemples són (GenBank, EMBL, DDBJ). Aquest augment en l'interès per aquest mètode també està fent que es dissenyin millors primers "universals" [12].

Algunes de les limitacions d'aquest mètode és que sempre hi pot haver algun error, i per tant no totes les seqüencies que es troben a les bases de dades son de bona qualitat. Aquests errors es poden trobar principalment com a errors de seqüenciació, contaminacions, mala identificació de la mostra o problemes taxonòmics. BOLD vol servir com a filtre per evitar aquests errors i així poder tenir una base de dades amb només *barcodes* de qualitat [12].

1.3 Cytochrome C oxidase subunit 1 (*COI*) com a *barcode*

El fet d'utilitzar un únic gen codificant de proteïna ens permet estudiar espècies properes ja que es mantenen constants durant més temps (2% de canvis cada milió d'anys implicaria 12 bases diferents en una seqüencia de 600pb). Per tant un aspecte important a tenir en compte és la longitud del gen que volem amplificar, un gen massa curt podria no ser suficient per trobar diferències entre dues espècies molt properes, i la amplificació d'un gen molt llarg pot ser complicada. Malgrat això no es pot saber en exactitud la longitud necessària per obtenir informació del gen ja que el rati d'evolució varia entre espècies i regions del genoma.

Un altre aspecte a tenir en compte a l'hora de seleccionar el gen a seqüenciar és la presència d'introns. Ja que un gen amb introns pot contenir més variacions en aquesta regió, i sense fer un "*splicing* virtual" podríem obtenir resultats erronis[6].

Per aquest motiu els gens mitocondrials són els més utilitzats en *Barcoding*. Ja que no tenen introns, fan poca recombinació i s'hereten de forma haploide[13].

Un dels gens mitocondrials més utilitzats és el gen del *cytochrome c oxidase I* ja que té diverses avantatges respecte altres gens. Per començar els seus *primers* són molt robusts (funcionen amb un ventall d'espècies ampli), per tant podem amplificar de forma repetida aquest fragment gènic en cada individu. En segon lloc el fragment obtingut és curt (600pb), i força variable. Això vol dir que ens permet, mitjançant algoritmes, identificar espècies molt properes, o fins i tot es poden trobar diferències entre grups d'individus de la mateixa espècie i agrupar-los entre ells , el que es coneix com a un "BIN" (*Barcode Index Number*) [14].

A més aquest és un gen essencial per la vida de molts organismes sinó tots, ja que és una subunitat de la proteïna *cytochrome c oxidase*, altrament coneguda com el complex IV de la cadena respiratòria dels mitocondris (*Figura 3*), que catalitza la reducció d'oxigen a aigua en l'últim pas de la cadena respiratòria mitjançant un complex enzimàtic que forma una bomba de protons que redueix O_2 a H_2O utilitzant el seu centre catalític format per ions de ferro i coure.

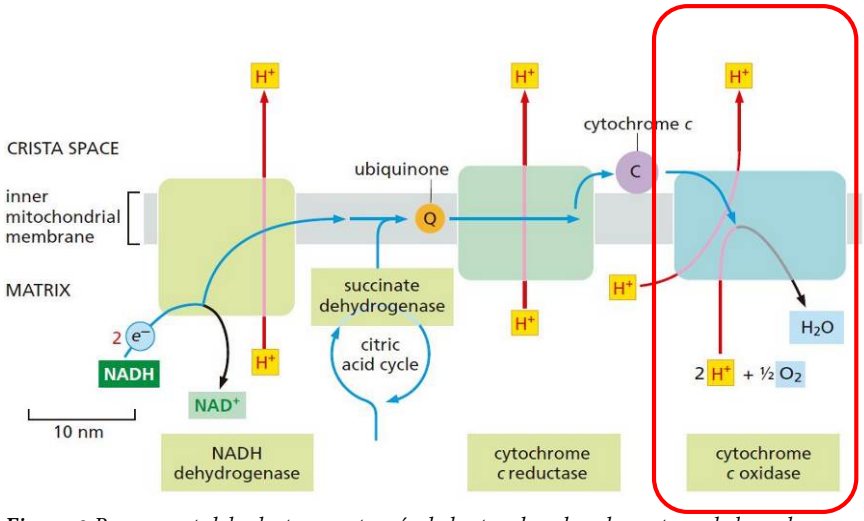


Figura 3 *Recorregut dels electrons a través de les tres bombes de protons de la cadena respiratòria* [15].

La molècula de *cytochrome c oxidase* està formada per diferents subunitats. Entre elles la subunitat 1 (), codificada pel gen del mateix nom (*Figura 5*). Una regió d'aquest gen és la que hem utilitzat com a *barcode*. Aquesta regió de 658pb ens és suficient per poder diferenciar entre **espècies degut a la seva gran variabilitat genètica.**

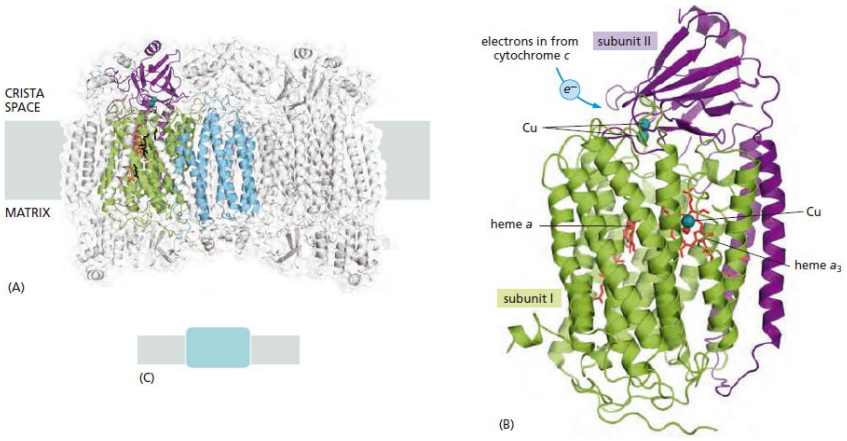


Figura 4 *Estructura del cytochrom c oxidase. En verd podem veure la subunitat 1* [15].

La diversitat en la seqüència aminoacídica codificada per aquest gen mitocondrial ens dona suficient informació per col·locar espècies en les seves categories taxonòmiques corresponents. A més aquesta divergència genètica pel que fa a la seqüència del COI és suficient per discriminar espècies properes de lepidòpteres, tot i ser demostrat que la variabilitat en lepidòpters és menor que en altres espècies estant un 60.4% de les seqüencies al 4-8% de divergència, mentre que hi ha altres fílums com els anèl·lids que el 70.3% de la seqüència té entre un 16 i un 32% de divergència (*Taula 1*¡**Error! No se encuentra el origen de la referencia.**), això vol dir que aquest percentatge del DNA ha acumulat més mutacions durant el temps [16].

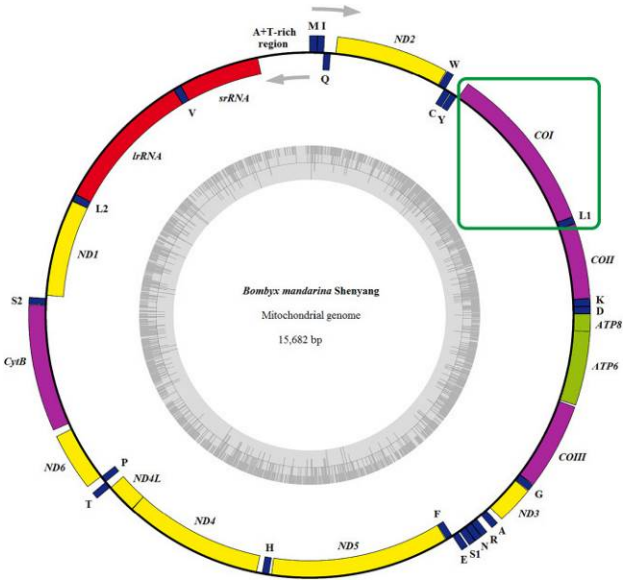


Figura 5 *Genoma mitocondrial de Bombyx mandarin podem observar emmarcat en verd la posició del cytochrome c oxydase subunit 1* [17].

Taula 1 *Mitjana (mean) i desviació estàndard (s.d) del percentatge de divergència de la seqüència de COI per 13.320 parelles d'espècies congènere en 11 fílums. (També es mostra en percentatge l'estimació de divergència d'un rang en particular. n indica el nombre de parelles congènere examinades en cada grup)* [16]

				COI sequence divergence (%)						
phylum	*n*	mean	s.d.	0–1	1–2	2–4	4–8	8–16	16–32	32+
Annelida	128	15.7	6.2	6.3	1.6	—	3.9	18.0	70.3	—
Arthropoda										
Chelicerata	1249	14.4	3.6	—	0.2	0.2	2.0	50.8	46.8	—
Crustacea	1781	15.4	6.6	0.1	0.3	4.3	13.4	18.0	63.8	0.1
Coleoptera	891	11.2	3.8	2.2	1.6	3.0	8.0	74.2	11.0	—
Diptera	1429	9.3	3.5	0.9	2.1	4.1	14.0	76.2	2.7	—
Hymenoptera	2993	11.5	3.8	0.2	—	0.3	3.3	93.0	3.2	—
Lepidoptera	882	6.6	2.2	1.0	2.8	7.3	60.4	28.5	—	—
other orders	1458	10.1	4.9	0.5	1.6	8.4	35.5	41.8	12.1	—
Chordata	964	9.6	3.8	1.2	0.7	4.3	19.2	61.7	12.9	—
Cnidaria	17	1.0	1.2	88.2	5.9	5.9	—	—	—	—
Echinodermata	86	10.9	4.9	1.2	1.2	5.8	39.5	44.2	8.1	—
Mollusca	1155	11.1	5.1	1.2	1.9	4.0	15.0	67.5	10.0	0.4
Nematoda	49	11.0	2.9	—	2.0	—	22.4	73.5	2.0	—
Platyhelminthes	84	14.4	5.4	8.3	—	—	4.8	44.0	42.9	—
minor phyla	154	13.3	9.7	0.6	1.3	2.6	39.6	38.3	16.9	0.7
overall	13 320	11.3	5.3	0.9	1.0	3.4	16.2	59.4	19.0	0.1

Així doncs aquest gen pot servir com a *barcode* per defecte per obtenir un sistema taxonòmic final no només en *Eilema*, sinó també en altres espècies.

1.4 Us de gens nuclears com a complement

Apart de la seqüenciació del COI es va decidir utilitzar altres gens per complementar l'estudi i donar més solidesa als resultats obtinguts, ja que el fet d'estudiar un únic gen ens pot limitar la informació obtinguda, i per tant podem quedar limitats pel que fa a les dades necessàries per estudiar la filogènia d'*Eilema*. Els gens que es van decidir utilitzar es van treure de l'estudi de Wahlberg et al [18]. ja que va estudiar diferents primers per trenta gens diferents i les seves possibilitats per ser utilitzats en *Lepidoptera (Taula 2)*.

Taula 2 *Informació de diversos gens recomanats per Whalberg et al. que es van utilitzar en aquest treball per la seva amplificació en Lepidoptera. Inclou la longitud del fragment, el percentatge de mostres que van sortir correctes en el seu l'estudi, i el GeneID de cada gen per Bombyx* [18].

Gen	Llargada del fragment (pb)	Percentatge mostres correctes (Whalberg et al.)	GeneID de *Bombyx*
ArgK	388	80	BGIBMGA005812
Ca-ATPase	444	77	BGIBMGA000408
DDX23	303	80	BGIBMGA003429
ProSup	432	73	BGIBMGA004645
PSb	366	77	BGIBMGA000201
SSU72	249	77	BGIBMGA000925
CAD	826	80	
IDH	722	77	
MDH	407	77	
Wingless	400	67	

De la llista anterior de gens ens vam centrar principalment en els següents:

SSU72:
Es va decidir utilitzar aquest gen ja que s'ha utilitzat en altres estudis i s'ha demostrat que es poden obtenir bons resultats i que és un candidat viable per *barcoding* [18].
Aquest gen nuclear un cop amplificat ens dona un fragment de 249pb, que tot i ser menor que el que obtenim pel gen COI, ens pot donar una informació molt valuosa si complementem les informacions dels dos.

Aquest gen codifica per la proteïna *RNA polymerase II subinit A C-terminal domain phosphatase SSU72*. Aquesta fosfatasa catalitza la desfosforilació del domini C-terminal de la RNA polymerase II, i per tant juga un paper important en el processament i la terminació 3' de l'RNA[19].

Arginine Kinase:
Un altre gen nuclear que ens pot ser servir com a complement és el gen que codifica per l'enzim Arginina Kinase (ArgK), ja que ha set recomanat [18] per varis estudis degut a la seva elevada taxa d'èxit en *Lepidoptera*. Aquest gen nuclear codifica per una fosfoquinasa que emmagatzema energia en forma d'arginina fosfatasa. Si amplifiquem aquest gen obtindrem un fragment de 388pb que ens pot ser útil per l'anàlisi per *barcoding* [20].

Wingless:
El gen Wingless (wg) ha estat molt estudiat en *Drosophila sp*. S'uneix a un lligand dels receptors de la família *frizzle-7-transmembrane*. Es creu que pot tenir funció de factor de creixement ja que pot regular gens de les cèl·lules properes en que s'expressa, i també s'ha detectat la seva expressió durant el desenvolupament de l'epidermis. Però la seva principal funció és la del desenvolupament i formació de la tràquea i el tronc dorsal d'aquests insectes[21]. Aquest gen és considerat com a un bon marcador per *barcoding* en altres estudis relacionats amb nimfàlids. En aquests estudis es va observar que aquest gen nuclear pateix menys saturació genètica que el DNA mitocondrial, però tenen un rati de mutació semblant[22].

1.5 Anàlisi filogenètic

Els anàlisis filogenètics de seqüències de DNA o de proteïnes s'han convertit en una eina important per l'estudi de l'historia evolutiva de tots els organismes. Com que el rati d'evolució de les seqüències varia extensivament entre gens i altres fragments de DNA, es pot estudiar la relació de virtualment tots els nivells de classificació d'organismes (regnes, fílums, gèneres, etc.) mitjançant aquests fragments genètics. Aquests estudis també es poden utilitzar per analitzar el patró d'evolució de famílies de gens [23].

Utilitzant diversos algoritmes informàtics com MUSCLE, ClustalW, ClustalOmega, podem obtenir un alineament entre els *barcodes* de totes les mostres que ens interessa estudiar. És a partir d'aquest alineament que es pot construir un arbre filogenètic que ens indicarà d'una forma visual les distàncies entre les diferents espècies.

S'ha de tenir molt clar quina regió del DNA es vol amplificar, ja que els canvis evolutius varien molt de regió en regió, no tenen el mateix rati evolutiu regions codificants de proteïnes, introns, exons, regions flanquejants i regions no codificants. Les seqüencies descendents d'una seqüència ancestral van divergent gradualment ja que es van produint substitucions de nucleòtids. Es pot mesurar d'una forma simple la divergència de la seqüència utilitzant la següent formula:

$$\hat{p} = n_d/n$$

P mesura la proporció de nucleòtids en que les dues seqüencies divergeixen. $\mathbf{n_d}$ representa el nombre de nucleòtids diferents entre les dues seqüències i **n** representa el nombre total de nucleòtids examinats. Aquesta proporció és el que es coneix com la "***p distance***" de les seqüències de nucleòtids [24].

Com que les seqüències de DNA estan formades per 4 nucleòtids això vol dir que hi pot haver 16 combinacions diferents. A cada combinació se li assigna un símbol matemàtic i una classe: *identical* (O), *transition* (P), *transversion* (Q). (*Taula 3*)

Taula 3. *Setze tipus diferents de parelles de nucleòtids entre les dues seqüències* [24].

Class	Nucleotide Pair				
Identical nucleotides	AA	TT	CC	GG	Total
Frequency	O_1	O_2	O_3	O_4	O
Transition-type pair	AG	GA	TC	CT	Total
Frequency	P_{11}	P_{12}	P_{21}	P_{22}	P
Transversion-type pair	AT	TA	AC	CA	
Frequency	Q_{11}	Q_{12}	Q_{21}	Q_{22}	
	TG	GT	CG	GC	Total
Frequency	Q_{31}	Q_{32}	Q_{41}	Q_{42}	Q

Si la substitució és aleatòria entre els quatre nucleòtids tenint en compte que p = P + Q, s'espera que Q sigui el doble que P quan p es baixa. Però a la pràctica això no passa, ja que P passa més sovint que Q. Igual que amb els aminoàcids la *p distance* ens dona una estimació del nombre de substitucions que pateixen els nucleòtids quan les seqüències estan estretament relacionades. Però quan la p és gran ens dona una estimació incorrecta ja que no té en compte substitucions inverses i paral·leles. Per aquest motiu és necessari un model matemàtic que estimi el nombre de substitucions.

Molts autors diferents han presentat models representats aquí com a matrius de substitució (*Taula 4*) [24]:

Taula 4 *Models de substitució de nucleòtids* [24].

	A	T	C	G	A	T	C	G
	(A) Jukes-Cantor model				(E) HKY model			
A	-	α	α	α	-	βg_T	βg_C	αg_G
T	α	-	α	α	βg_A	-	αg_C	βg_G
C	α	α	-	α	βg_A	αg_T	-	βg_G
G	α	α	α	-	αg_A	βg_T	βg_C	-
	(B) Kimura model				(F) Tamura-Nei model			
A	-	β	β	α	-	βg_T	βg_C	$\alpha_1 g_G$
T	β	-	α	β	βg_A	-	$\alpha_2 g_C$	βg_G
C	β	α	-	β	βg_A	$\alpha_2 g_T$	-	βg_G
G	α	β	β	-	$\alpha_1 g_A$	βg_T	βg_C	-
	(C) Equal-input model				(G) General reversible model			
A	-	αg_T	αg_C	αg_G	-	$a g_T$	$b g_C$	$c g_G$
T	αg_A	-	αg_C	αg_G	$a g_A$	-	$d g_C$	$e g_G$
C	αg_A	αg_T	-	αg_G	$b g_A$	$d g_T$	-	$f g_G$
G	αg_A	αg_T	αg_C	-	$c g_A$	$e g_T$	$f g_C$	-
	(D) Tamura model				(H) Unrestricted model			
A	-	$\beta\theta_2$	$\beta\theta_1$	$\alpha\theta_1$	-	a_{12}	a_{13}	a_{14}
T	$\beta\theta_2$	-	$\alpha\theta_1$	$\beta\theta_1$	a_{21}	-	a_{23}	a_{24}
C	$\beta\theta_2$	$\alpha\theta_2$	-	$\beta\theta_1$	a_{31}	a_{32}	-	a_{34}
G	$\alpha\theta_2$	$\beta\theta_2$	$\beta\theta_1$	-	a_{41}	a_{42}	a_{43}	-

Mitjançant aquests models podem calcular les distancies entre seqüències i a partir d'aquí construir un arbre filogenètic.

<u>Existeixen diversos tipus d'arbres filogenètics:</u>

Amb arrel o sense arrel: El patró d'on surten les branques d'un arbre es coneix com la topologia.

Si el nombre de unitats taxonòmiques de l'arbre (**taxa**) és quatre, un arbre amb arrel té 15 possibles topologies diferents, mentre que un sense arrel en té 4.

En teoria, una seqüència de DNA es parteix en dues seqüències durant la duplicació. Per aquest motius els arbres es bifurquen [23].

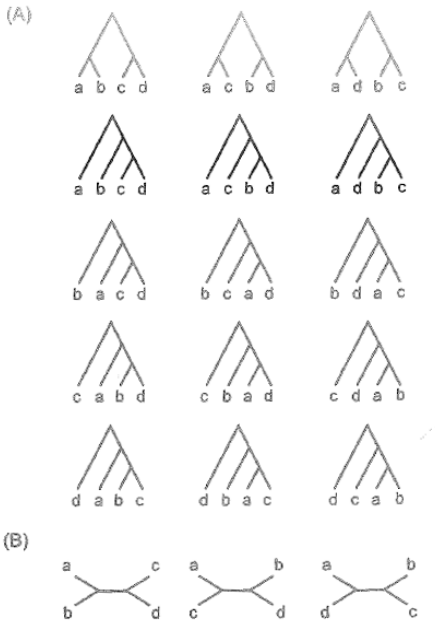


Figura 6 *(A) Quinze arbres amb arrel possibles i (B) tres possibles arbres sense arrel* [23]

Arbres de gens i arbres d'espècies: Un arbre d'espècies o poblacional és un tipus d'arbre que representa l'historia evolutiva d'un grup d'espècies o poblacions. En aquest tipus d'arbre el temps de divergència entre dues espècies consisteix en el temps que van tardar les espècies a ser aïllades pel que respecta a la reproducció. Un arbre de gens en canvi pot no assemblar-se a un arbre d'espècies. Per exemple en el cas d'al·lels polimòrfics per un locus els temps de divergència dels gens pot ser més llarg que el que respecta a l'espècie, i per tant les branques es formen abans [23].

***Expected* i *Realized Trees*:** Un arbre que pot ser construït fent servir seqüències de llargada infinita, o el nombre esperat de substitucions per cada branca es coneix coma un "*expected tree*" (arbre esperat). Mentre que un arbre basat en el nombre real de substitucions es coneix com a un "*realized tree*" (arbre real) [23].

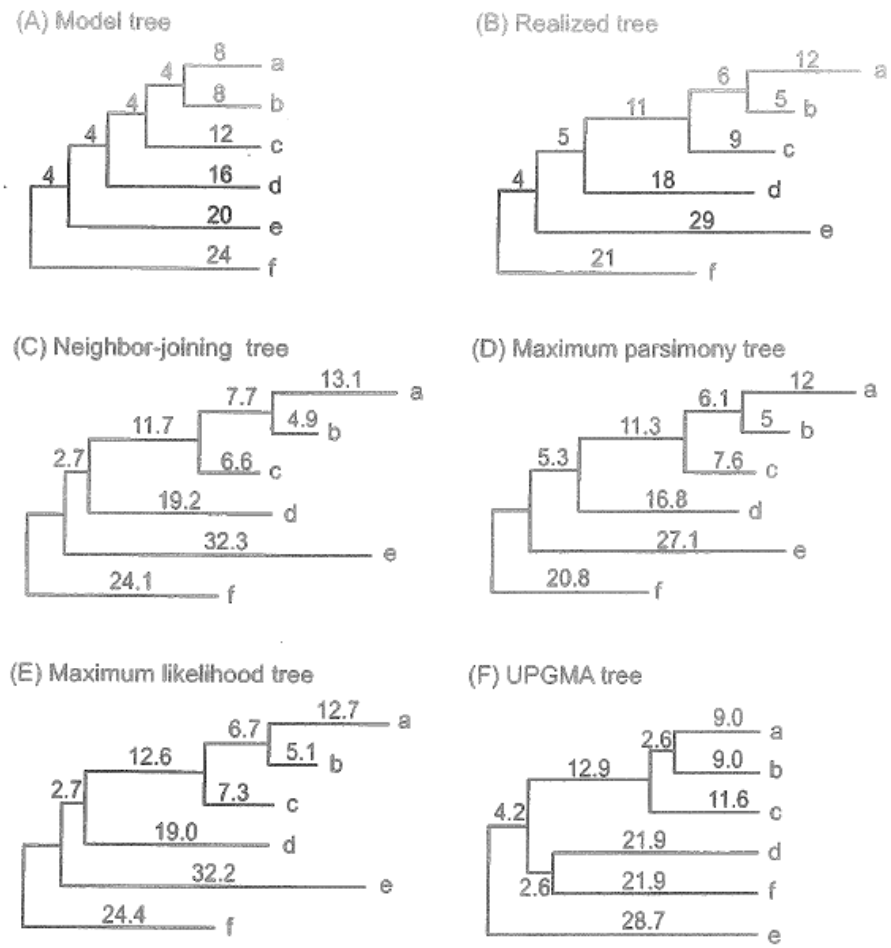


Figura 7 *Diferents models d'arbres* [23]

Definició de "distància topològica" i com es calcula:

La topologia és l'estructura que formen les branques d'un arbre. És d'una gran importància biològica ja que indica patrons en la relació entre taxes. Per tant arbres amb la mateixa topologia i arrel tenen la mateixa interpretació biològica.

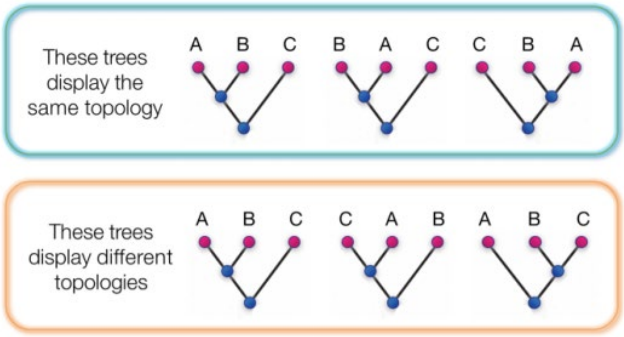


Figura 8 *Exemple d'arbres amb la mateixa (quadre de dalt) i diferent (quadre de baix) topologies* [25]

Per exemple els tres arbre de la capsa de dalt de la (*Figura 8*) tenen la mateixa topologia. En tots tres es compleix que A i B estan més estretament relacionats que C. En la capsa de sota tots tenen topologies diferents i per tant les relacions que s'observen són diferents [25].
No hi ha cap mètode estadístic general per escollir a l'hora de calcular les distancies entre taxes i construir la topologia, però a partir de simulacions informàtiques i estudis empírics s'han establert un seguit de punts per obtenir aquesta topologia. En un principi es recomana utilitzar el mètode *Jukes-Cantor*, i si aquest mètode no dona els resultats desitjats després se'n poden utilitzar d'altres [26].

Un dels mètodes més utilitzats en la generació d'arbres filogenètics és el ***"neighbor-joining method"***. Aquet procediment ens permet reconstruir arbres a partir de les dades de distància evolutiva obtinguda a partir d'un alineament (com ara *MUSCLE*, *ClustalOmega*, *ClustalW*). El *neighbor-joining method* consisteix en trobar parelles d'OTUs que minimitzin la llargada de la branca en cada pas de *clustering* d'aquests OTUs, començant per un arbre en forma d'estrella. Aquest mètode és molt ràpid per obtenir la distància topològica i s'ha comprovat que funciona millor que altres mètodes com els: *Farris*,, i *Li*. I té una eficàcia igual de bona que el mètode *Sattath Tversky's* del 1977, però el *neighbor-joining* és molt més ràpid [27].

2. Objectius

La tècnica de *barcoding* ha estat utilitzada per identificar moltes espècies diferents, però per cada espècie s'han hagut de buscar les millors condicions per obtenir els millors resultats. És a dir, la finalitat és trobar un mètode que ens doni resultats consistents. Això vol dir optimitzar les condicions d'extracció, amplificació i anàlisi de tal manera que s'obtinguin dades òptimes.
A més la majoria d'estudis de *barcoding* utilitzen un únic gen, el COI. En el nostre cas volem aconseguir no només informació pel que respecta a aquest gen, sinó que també volem obtenir informació de la seqüencia de gens nuclears que ens serviran per complementar la informació obtinguda a partir de les seqüències del *cytochrom c oxidase subunit* 1.
L'objectiu final del treball és construir arbres filogenètics a partir de les seqüències de COI i de gens nuclears. Amb aquests arbres podrem comprovar visualment la proximitat entre aquestes espècies.

Per tant podem resumir els objectius del treball en tres punts:

- Optimitzar la tècnica de DNA *Barcoding* pel gènere de Lepidòpters *Eilema* sp. (Fam. Erebidae).

- Posar a punt la seqüenciació d'un gen mitocondrial, i estudiar la possibilitat d'utilitzar alguns gens nuclears.

- Realitzar un estudi taxonòmic sobre mostres de diverses espècies properes del gènere *Eilema* sp.

3. Material i Mètodes

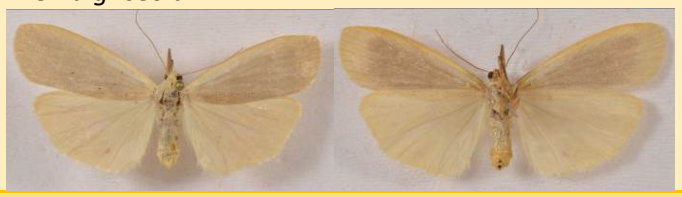
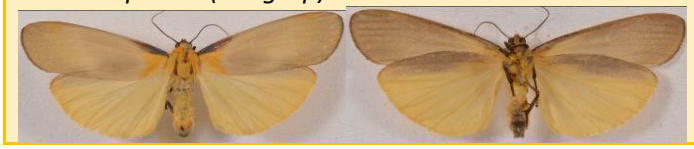
3.1 Material Biològic

Per la realització de l'estudi s'ha treballat amb diferents espècies del gènere *Eilema*. En total s'han estudiat 26 espècies diferents d'aquest gènere i un *outgrup* format per *Lithosia quadra*. Les diferents espècies estudiades es mostren a la *Taula* 5.

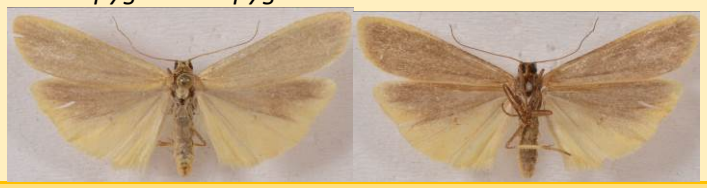
Taula 5 *Informació visual de cada espècie estudiada, inclou la localització de cada espècie, i l'ID de les mostres estudiades de cada espècie.*

ID Mostres	Fotografia (vista dorsal/vista ventral)	Distribució[4]
MSE_44 **MSE_48** **MSE_66**	*Eilema albicosta albicosta* 	
MSE_22 **MSE_58**	*Eilema albicosta witti* 	
MSE_02 **MSE_19** **MSE_46** **MSE_52**	*Eilema caniola* 	
MSE_30 **MSE_51** **MSE_67**	*Eilema caniola gibrati* 	
MSE_24 **MSE_53** **MSE_54**	*Eilema caniola torstenii* 	
MSE_32 **MSE_68**	*Eilema cereola* 	Ailefroide, França 

28

MSE_12 **MSE_18**	*Eilema complana* 	
MSE_55 **MSE_49** **MSE_69** **MSE_70**	*Eilema complana iberica* 	En clar *complana iberica*, en fosc *complana complana*
MSE_35 **MSE_37** **MSE_47**	*Eilema costalis* 	*Xipre i Balcans*  29
MSE_01 **MSE_09**	*Eilema depressa* 	
MSE_29 **MSE_50**	*Eilema bipuncta* 	
MSE_20 **MSE_59**	*Eilema griseola* 	
MSE_03 **MSE_04** **MSE_06** **MSE_10**	*Eilema interpositella* 	
MSE_36 **MSE_60** **MSE_72** **MSE_73**	*Lithosia quadra (Outgrup)* 	

MSE_21 **MSE_61**	*Eilema lurideola* 	
MSE_31 **MSE_71**	*Eilema lutarella lutarella* 	
MSE_25 **MSE_39** **MSE_45**	*Eilema lutarellla luqueti* 	
MSE_23 **MSE_43**	*Eilema marcida* 	
MSE_27 **MSE_34** **MSE_41** **MSE_56**	*Eilema muscula* 	Ailefroide, França  29
MSE_13 **MSE_62**	*Eilema palliatella* 	
MSE_15 **MSE_17**	*Eilema predotae* 	
MSE_26 **MSE_33** **MSE_40** **MSE_57**	*Eilema pseudocomplana* 	

MSE_16 **MSE_63**	*Eilema pygmaeola pallifrons* 	
MSE_28 **MSE_42**	*Eilema pygmaeola pygmaeola* 	
MSE_05 **MSE_11** **MSE_64**	*Eilema rungsi* 	
MSE_08 **MSE_38**	*Eilema sororcula* 	
MSE_14 **MSE_65**	*Eilema uniola* 	

Les mostres van ser capturades per en Ramon Macià Vilà. Cada insecte consta d'un seguit de metadades que ens donen una sèrie d'informacions: Data de captura, Lloc de captura, Coordenades GPS, Sexe.

Les mostres es troben en dos estats de conservació diferents. 15 mostres provenen de les pràctiques de l'assignatura de *Tècniques de Biologia Molecular*. Aquestes mostres es troben en millor estat que les altres ja que un cop capturades l'any passat 2018 es van posar al congelador, per tant el seu DNA està poc degradat i ha set més fàcil obtenir resultats.

La resta de mostres que va portar en Ramon Macià es troben seques i preparades per exposició (*Figura 9*). Aquestes mostres es van capturar en dates i zones molt diferents (del

Figura 9 *Mostres seques en exposició analitzades*

1994 al 2018), per tant el seu estat de conservació varia de mostra a mostra. A més no totes les espècies de la col·lecció tenen el mateix nombre d'individus, per exemple *E. interpositella* en tenim sis, mentre que de *E. bipuncta* només hi ha un exemplar. Per aquest motiu es va començar estudiant les que n'hi havia més i després es va passar a les més rares.

Cada insecte que s'ha analitzat se li ha donat un ID format per les inicials de qui ha fet l'extracció i un numero (MSE_XX) acompanyat de una C o K si s'ha realitzat una extracció amb chelex o amb un kit comercial. A més s'ha posat la data de l'extracció en el tub que conté l'extracte de DNA.
S'ha portat en tot moment un recompte de totes les mostres per saber quantes s'han analitzat i quantes queden per analitzar.

3.2 Procediment

El procediment va consistir principalment en realitzar extraccions de DNA de les diferents mostres, i amplificar mitjançant PCR regions de COI. Es va començar per les mostres de *Molecular Biology* Techniques, ja que van ser capturades recentment i es troben en molt bon estat de conservació. Un cop es va determinar quines eren les millors condicions per obtenir els millors resultats, i quin era el procediment necessari per obtenir les millors possibilitats d'èxit (*Figura 10*), es va decidir realitzar el mateix procediment aquesta vegada utilitzant mostres seques en exposició. Apart de COI es va realitzar el mateix procediment utilitzant gens nuclears.

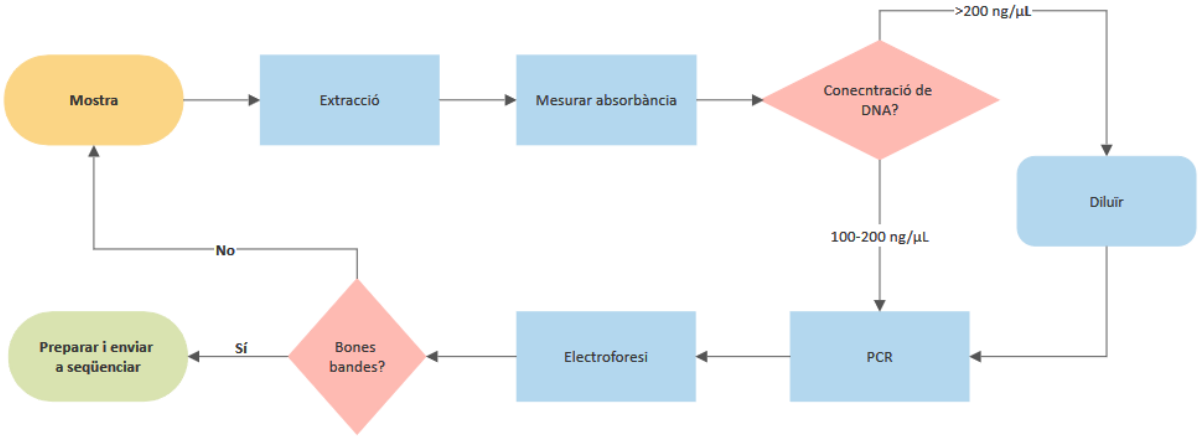


Figura 10 *Diagrama de flux del procediment realitzat al laboratori* [30].
També es va programar un script per obtenir la seqüència consens i construir un arbre filogenètic.

3.3 Extracció de DNA

S'han utilitzat dos mètodes diferents per obtenir l'extracte de DNA de les mostres. La utilització d'un o altre mètode es va decidir conforme la dificultat d'obtenir resultats en la mostra. Per defecte es va realitzar l'extracció amb Chelex 100, si utilitzant aquest mètode no s'obtenien resultats es passava a una extracció amb un kit comercial. Es va utilitzar el kit c, ja que vam comprovar que obteníem els mateixos resultats a millor preu.

Chelex 100:

El chelex 100 és un copolímer que conté ions iminodiacetat aparellats que actuen com a agents quelats. Aquests agents s'uneixen a cations Mg^{2+} que son elements cofactors per DNases. Això fa que protegeixi les mostres de DNases.

El chelex conté unes boletes (*beads*) polars (*Figura 11*) que s'uneixen a components cel·lulars polars després de la lisi amb 5µL de proteïnasa K i força mecànica, mentre que el DNA queda en solució a l'aigua. [31]

Figura 11 *Chelex amb els beads al fons de l'eppendorf.*

El protocol que s'ha utilitzat per l'extracció amb chelex és el de l'assignatura de *Tècniques de Biologia Molecular* i que es troba a l'annex, que recomana utilitzar una pota per cada insecte.

Kits comercials:

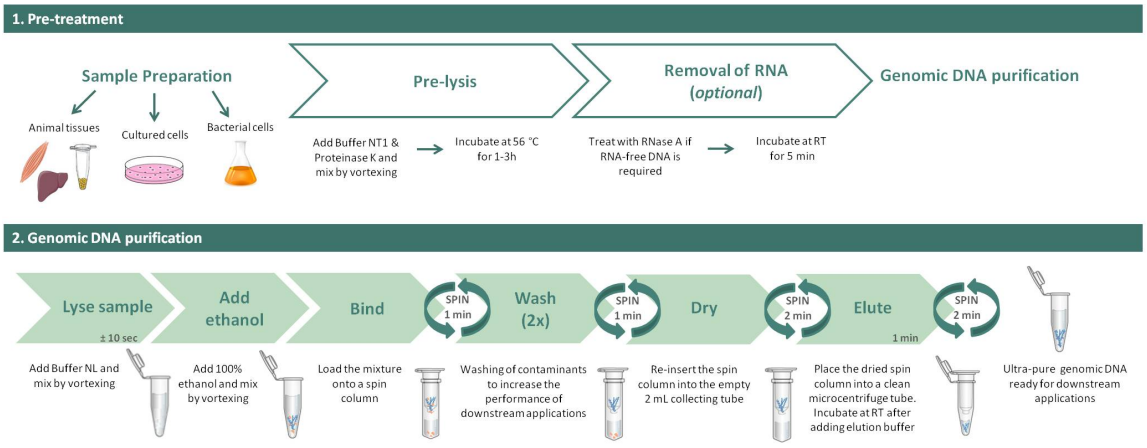


Figura 12 *Resum visual del protocol d'extracció del kit NZY® Tissue gDNA Isolation Kit, es pot observar que hi ha dos passos dividits en altres passos, el primer és un pre-tractament de la mostra que consisteix principalment en la lisi de les cèl·lules. El pas 2 consisteix en la purificació del DNA del lisat mitjançant diversos reactius i l'us de columnes amb membrana de sílice i varies centrifugacions. Al final s'obté DNA purificat lliure de contaminant i altres restes cel·lulars.* [33]

Per dur a terme l'extracció de DNA en mostres més complicades hem utilitzat el *DNeasy® Blood & Tissue kit de Qiagen* [32] i el *NZY® Tissue gDNA Isolation Kit de NZYtech* [33], seguint el protocol en que venen. Aquests kits es basen en la lisi del teixit i posterior purificació d'aquests lisat utilitzant columnes amb membrana de sílice i varis reactius que netegen la mostra per poder obtenir DNA purificat d'alta qualitat (*Figura 12*). Per mostres seques en què amb una pota no s'obtenien resultats per aquest mètode es va decidir realitzar una extracció de DNA a partir de l'abdomen.

Malgrat que l'extracció amb chelex surt més econòmica i és més ràpida que el kit les mostres extretes amb chelex es degraden més ràpidament, mentre que l'DNA obtingut a partir de l'extracció amb un kit comercial aguanta molt més temps. Per aquest motiu les mostres que no sortien amb chelex es tornava a fer una extracció utilitzant un kit comercial per així poder realitzar més proves. Un cop realitzada l'extracció es comprova la qualitat i quantitat de DNA mesurant l'absorbància utilitzant un nanofotòmetre. Per millors resultats es busca una concentració de 100 a 200 ng/µL. Ja que s'utilitza massa DNA en la PCR pot no produir-se amplificació (*Taula 6*).

Taula 6 *Resultats obtinguts del nanofotòmetre. Com es pot observar la qualitat és força bona, però la concentració és massa elevada. Aquestes mostres no van donar resultat en la PCR, després de diluïres es va obtenir banda.*

Mostra	A260/A280 (Puresa)	A260/230 (Contaminants)	Concentració (ng/µL)
MSE_08K	1,461	1,453	1328
MSE_09K	1,844	1,934	578

3.4 *Primers*

Per l'amplificació s'han utilitzat diferents *primers* depenent de les regions que es volien ampliar. En el cas del gen COI es va començar per utilitzar la parella de *primers* LepF1/R1 ja que són un del parell de *primers* més utilitzats en estudis de *barcoding* en *Lepidoptera*. Després de realitzar varies proves es va determinar que la parella de *primers* per COI L*COI*490/HCO2198 donava millor resultat que LepF1/R1, per tant es va decidir utilitzar aquests darrers per defecte a l'hora d'amplificar per COI. A continuació es pot veure la comparació de les dues seqüencies d'aquests dos parells de *primers* per COI (*Taula 7*).

***Taula 7** Primers utilitzats per l'amplificació de COI*

Primer	Seqüència	Fwd-Rev	Descripció
LCO1490	GGTCAACAAATCATAAAGATATTGG	FWD	CO1 - Cytochrome Oxydase 1
HCO2198	TAAACTTCAGGGTGACCAAAAAATCA	REV	CO1 - Cytochrome Oxydase 1
LepF1	ATTCAACCAATCATAAAGATATTGG	FWD	CO1 - Cytochrome Oxydase 1
LepR1	TAAACTTCTGGATGTCCAAAAAATCA	REV	CO1 - Cytochrome Oxydase 1

Aquests primers ens donen uns fragments d'uns 658 pb del gen COI. La temperatura recomanada per amplificar aquests primers és de 50ºC i es va aplicar dos protocols per la PCR basant-nos en els resultats obtinguts de Herbert et al. (2002) [6]. Un d'aquests protocols utilitza pre-amplificació per intentar augmentar la quantitat de DNA obtingut amb el handicap d'obtenir més soroll.

Pel que fa als gens nuclears es va utilitzar *primers* específics per cada un dels següents gens nuclears: DDX23, ArgK, Wingless, Psb, SSU72, Prosup, CAD, IDH, MDH, Calcium ATPase. Aquest *primers* incorporen una cua T7/T3 ja que agilitza el procés de preparar les mostres per enviar a seqüenciar degut a que tots fan servir el mateix primer *forward* i el mateix primer *reverse*.

A continuació es poden veure les seqüències dels *primers* (sense les cues universals) juntament amb el gen que codifiquen (*Taula 8*).

Taula 8 *Primers nuclears utilitzats*

Primer	Seqüència	Fwd-Rev	Descripció
T7Fwd	TAATACGACTCACTATAGGG	FWD	T7 Promoter - Reverse Sequencing *primer*
T3Rev	ATTAACCCTCACTAAAGGG	REV	T3 Promoter - Reverse Sequencing *primer*
T7-ArgK_F	yGAyCCsATCATyGAGGACTACC	FWD	ArgK - Arginine Kinase
T3-ArgK_R	AGrTGGTCCTCCTCrTTGCACCAvA	REV	ArgK - Arginine Kinase
T7-DDX23_F	ACAAAAGATAAAGAACGTgargargargcha	FWD	ATP-dependent RNA helicase DDX23
T3-DDX23_R	TGATCTTTTTCAgaccartghckrtcatccc	REV	ATP-dependent RNA helicase DDX23
LepWG1_F	GARTGYAARTGYCAYGGYATGTCTGG	FWD	Wingless Nuclear Gene
LepWG2_R	ACTICGCRCACCARTGGAATGTRCA	REV	Wingless Nuclear Gene
LepWG2a_R	ACTICGCARCACCARTGGAATGTRCA	REV	Wingless Nuclear Gene
Ca-ATPase_F	GAAtacgarccbgaaatgggwaarg	FWD	Calcium ATPase
Ca-ATPase_R	cdccrtgrgcggggtcgttraagt	REV	Calcium ATPase
ProSup_F	GACAACAATCGACtggcayccnaaya	FWD	ProSup gene
ProSup_R	CTGTCCAGTgactggaayttyttcatdg	REV	ProSup gene
PSb_F	GCTGGGAGCTACTggvtgytggtgygay	FWD	Proteasome Subunit beta type-1
PSb_R	AGATGCAGTCTCCAGTGTAGatrtcdckyt	REV	Proteasome Subunit beta type-1
Ssu72_F	CAGCTGACAGACCTaaytgttaygarttyg	FWD	RNA polymerase II subunit A
Ssu72_R	CCGATTGTAGCTTCTtcrtgrttrtcyt	REV	RNA polymerase II subunit A
MDHf	GGAYATNGCNCCNATGATGGGNG	FWD	Malate Dehydrogenase 1
MDHr	AGNCCYTCNACDATYTTCCAYT	REV	Malate Dehydrogenase 1
CAD743f	GGGNGTNACNACNGCNTGYTTYGARC	FWD	Carbamoyl-Phosphate Synthetase 2, Aspartate
CAD1028r	TTRTTNGGNARYTGNCCNCCCA	REV	Carbamoyl-Phosphate Synthetase 2, Aspartate
IDHdeg27F	GGGWGAYGARATGACNAGRATHATHTG	FWD	Isocitrate Dehydrogenase
IDHdegR	TTYTTRCAIGCCCANACRAANCCNC	REV	Isocitrate Dehydrogenase

En el cas del gen Wingless vam utilitzar dos *primers reverse* diferents, el LepWG2_R i el LepWG2a_R, ja que hi ha estudis que diuen que el LepWG2_R funciona millor. Per aquest motiu vàrem utilitzar els dos per veure quin ens donava millors resultats.
Es va decidir utilitzar com a referència les temperatures i cicles de PCR utilitzats pel grup de recerca de nimfàlids [34]. Això vol dir una temperatura d'anellament de 55ºC per aquests gens nuclears.

3.5 Amplificació per PCR

Per realitzar l'amplificació per PCR es van utilitzar dos volums diferents (*Taula 9*)

Taula 9 *Diferents volums utilitzats per la PCR*

Reactiu	Volum final de 25µL	Volum final de 50µL
Taq-Ready Mix	12,5 µL	25 µL
FWD Primer	1 µL	1 µL
REV Primer	1 µL	1 µL
H_2O	8,5 µL	21 µL
Mostra	2 µL	2 µL

A mesura que es feien les PCRs s'anava optimitzant la quantitat de mostra que s'afegia, ja que a vegades amb 2µL no era suficient per obtenir una quantitat d'DNA per enviar a seqüenciar. Per aquest motiu en diverses PCRs es va arribar a posar fins a 5µL de mostra (reduint respectivament el volum d'H_20).
Es van utilitzar dos Taq polimerasa diferents, la primera polimerasa (*Taq-ReadyMix*) provinent de *Sigma*[35] ens va donar mals resultats, i per això després vam utilitzar la *NZYTaq II DNA polymerase* provinent d'*NZTtech*[36].

Es van utilitzar diferents programes de PCR (*Taula 10*) depenent del gen que volíem ampliar. Es van anar optimitzant les temperatures i temps per cada gen fins a trobar les que donaven millors resultats. Apart d'aquests programes també es van realitzar altres estudis utilitzant gradients per intentar determinar, sobretot en el cas de gens nuclears, quina temperatura era la ideal per l'amplificació. Els gradients més utilitzats eren de 50 a 57ºC o de 50 a 55ºC.

Taula 10 *Programes del termocycler utilitzats per la PCR*

Nom del programa	Paràmetres	Descripció
LEPCOI2 Pream + 50ºC	[1 min at 94ºC, 1,5 min at 45ºC, 1,5 min at 72ºC]x5 cycles [1 min at 94ºC, 1,5 min at 50ºC, 1 min at 72ºC]x35 cycles (Herbert et al., 2002)	Utilitzat per amplificar COI. Fa u na primera preamplificació per obtenir més quantitat de DNA al cost de una pitjor qualitat.
LEPCOI2 50ºC	[1 min at 94ºC, 1,5 min at 50ºC, 1 min at 72ºC]x35 cycles	Utilitzat per amplificar COI. Sense preamplificació.
Touchdown LEPWGLS	3 min at 94ºC [0,5 min at 94ºC, 1 min at 60ºC - 0,4ºC/cycle, 1,5 min at 72ºC]x24 cycles [0,5 min at 94ºC, 1 min at 50ºC, 1,5 min at 72ºC]x16 cycles, 10min 72ºC	Utilitzat per amplificar el gen Wingless. Es va decidir utilitzar una Touchdown PCR per intentar millorar l'especificitat de l'anellament de primers per Wingless.
LepCOI04	1 min 94ºC [0,5 min at 94ºC, 1,5 min at 45ºC, 1,5 min at 72ºC]x5 cycles [1,5 min at 94ºC, 1,5 min at 50ºC, 1 min at 72ºC]x35 cycles, 5 min 72ºC	Utilitzat per amplificar COI. Fa una preamplificació, i millora una mica els temps de LEPCOI2 Preamp. Es va utilitzar aquesta en mostres que la anterior no funcionava.
LEPNUC 53ºC	[1 min at 94ºC, 1,5 min at 53ºC, 1 min at 72ºC]x35 cycles	Utilitzat per amplificar gens nuclears.
LEPNUC 55ºC	[1 min at 94ºC, 1,5 min at 55ºC, 1 min at 72ºC]x35 cycles	Utilitzat per amplificar gens nuclears.

Un cop quantificat i valorat es fa una PCR utilitzant el primer específic pel nostre gen per amplificar la regió del gen en específic. S'utilitza un primer cicle d'amplificació a baixa temperatura (45ºC) que permet una primera amplificació imprecisa que millora el resultat final quan s'amplifica a 50ºC.

Per comprovar els resultats de la PCR i veure la longitud dels fragments del DNA es fa una electroforesis amb un gel al 1% d'agarosa. S'utilitza 2,5µL de colorant SYBR™ per tenyir el gel, i cada mostra es tenyeix amb 2,5µL de buffer 5x, es posen 10µL de mostra a cada pou.

3.6 Purificació i Seqüenciació

A partir dels resultats obtinguts de l'electroforesi després es decideix si es pot enviar el producte de PCR a seqüenciar o no. Si per exemple es veu que l'electroforesi conté bandes no desitjades es pot purificar la única banda que ens interessa o es pot decidir tornar a repetir la PCR canviant els paràmetres del *termocycler*. A més es pot comprovar la qualitat de la banda utilitzant l'espectrofotòmetre. Aquest aparell mesura la densitat òptica de la mostra i obtenim la concentració de DNA en ng/µL. A més ens dona una sèrie de ratis de l'absorbància a diferents longituds d'ona que ens indiquen diversos aspectes sobre la qualitat:

A260/280: Aquest rati ens diu la puresa del nostre DNA, amb un rati del voltant de 1.8 es considera que tenim una mostra pura [37], per tant si el nostre producte amplificat o ja purificat per enviar a seqüenciar es troba al voltant d'aquest valor, obtindrem segurament un bon resultat i és recomanable enviar-lo a seqüenciar.
A260/230: Aquest rati ens indica la quantitat de contaminants presents a la mostra, es considera lliure de contaminants si està al voltant de 2.0 a 2.2. Si aquest rati dona un valor baix pot indicar que tenim contaminants que absorbeixen a 230nm [37]. Aquest valor més baix pot ser degut a que els kits comercials d'extracció utilitzen guanidina en un dels tampons, si es fa una bona centrifugació final no hauríem de tenir-ne.

Aquests ratis i la quantitat de DNA s'utilitza també per mesurar la quantitat i la qualitat del DNA ja purificat que s'envia a seqüenciar, ja que es recomana enviar entre 80 i 100 ng/µL.

El producte provinent de la PCR ha de ser netejat abans d'enviar-lo a seqüenciar, ja que conté primers no anellats, nucleòtids, i restes de polimerasa.
S'han utilitzat dos mètodes diferents per purificar el producte de la PCR: utilitzant un kit exprés per la neteja i purificació del DNA amplificat i utilitzant uns enzims que netegen l'DNA.

Kits comercials:
S'han utilitzat dos kits comercials de dos marques diferents per realitzar la purificació: el kit de *Qiagen QIAquick® Gel Extraction kit* [38], i més endavant el kit *d'NZYtech NZYGelpure®* [39], ja que es va comprovar que donava els mateixos resultats que el de Qiagen a millor preu.
Aquests kits són molt semblants als kits d'extracció i purificació de DNA de les respectives cases comercials, però sense el pas d'extracció. Estan dissenyats principalment per purificar gels d'agarosa que contenen DNA, però també es poden utilitzar per purificar productes de PCR. Utilitzen columnes amb membranes de sílice i varis buffers (*binding buffer, wash*

buffer, elution buffer) per netejar tot aquest producte de tots els possibles contaminats i deixar només el DNA de doble cadena utilitzant varies centrifugacions. S'ha fet servir el protocol de cada un d'aquests kits comercials que es troben a l'annex.

Purificació enzimàtica:

Un altre mètode de purificació que es va aplicar més endavant i amb el que obtenim resultats tan bons com els dels kits comercials és utilitzar un parell d'enzims que "netegin" el nostre producte de PCR.

Aquest mètode de purificació desenvolupat per *New England BioLabs Inc.* Consisteix en utilitzar Exonucleasa I i *Shrimp Alkaline Phosphatase* (fosfatasa alcalina de gamba), altrament coneguda com rSAP, per eliminar els fragments de primers i dNTPs que es troben al producte de PCR. [40]

L'exonucleasa I degrada els primers residuals que no s'han anellat ja que catalitza l'eliminació de fragments de DNA d'una sola cadena en direcció 3' a 5' . Això vol dir que no afecta el nostre DNA ja amplificat que és de doble cadena. [41]

La rSAP elimina els fragments dNTP del producte ja que els hidrolitza i els desfosforila.

Apart de sortir més barat que utilitzar un kit, també és més ràpid ja que simplement s'ha de posar les quantitats indicades (1µL de rSAP i 0,5µL de Exo I per cada 5µL de mostra) i s'ha de posar a incubar uns 15 minuts a 37ºC i 15 minuts més a 80ºC per així inactivar els enzims. [42]

Un cop realitzada la purificació del producte i comprovat que la quantitat de DNA provinent de la PCR és la recomanada, mitjançant la lectura de l'absorbància, s'envia a seqüenciar a *Eurofins Genomics*. El producte que enviem conté entre 5µL i 7,5µL de mostra, i 2,5µL de primer *Forward* o *Reverse*. El volum de mostra el determinem a partir de la mesura del nanofotòmetre.

3.7 Obtenció de la seqüència a partir del cromatograma Sanger

El resultat de la seqüenciació arriba per correu en forma de diferents formats d'arxius entre ells dos FASTA, un amb el tros de la seqüència amb més qualitat, i un amb tota la seqüència (*unclipped*). Però el format que ens interessa és el .ab1, aquest tipus d'arxiu conté molta més informació que un FASTA. Aquest arxiu conté la informació necessària per construir un cromatograma de la nostra seqüència. El podem obrir utilitzant el programa *Chromas* [43] i observem el següent:

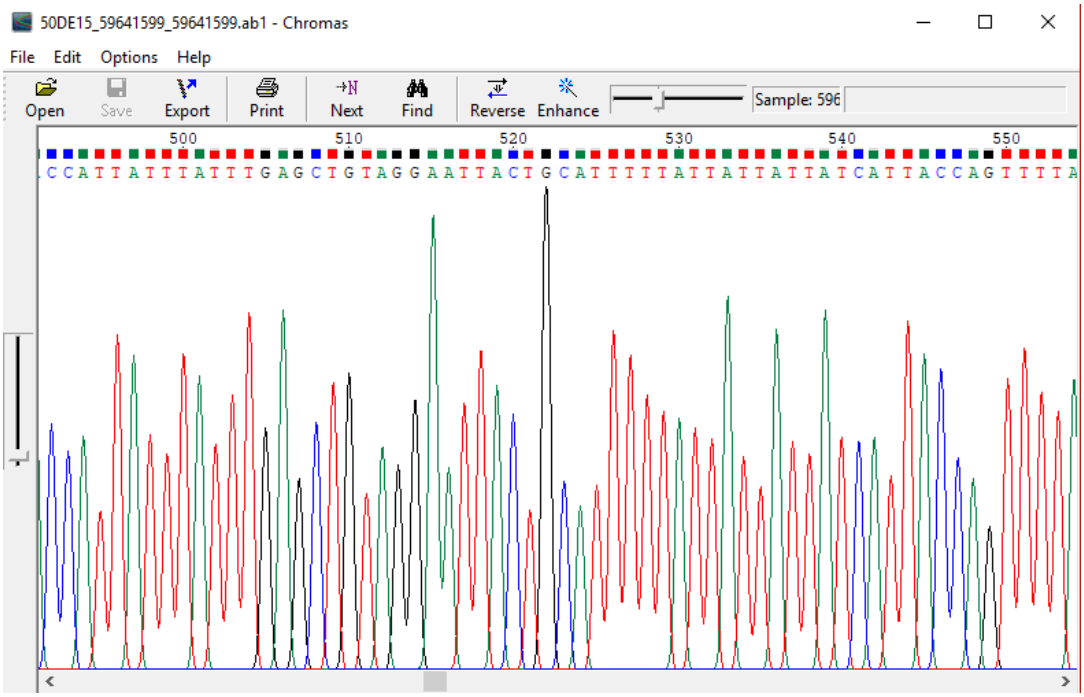


Figura 12 *Arxiu .ab1 obert amb el programa "Chromas".*

Cada base està representada per un color diferent, i la seva qualitat està representada per la forma dels pics. Com més alts són els pics més "intensitat" té aquella banda, és a dir, més probabilitats hi ha que aquella banda sigui la base que ens indica. Els quadres de dalt indiquen la qualitat, com més plens del color de la base més qualitat.

Aquest arxiu conté a més la informació de possibles al·lels, ja que registra la intensitat de totes les quatre bases en cada posició de la seqüència.

És a partir d'aquestes dades que podem construir doncs la seqüència consens a partir d'un arxiu .ab1 que contingui la informació del gen en direcció "*Forward*" i un altre arxiu .ab1 que contingui la informació del gen en direcció "*Reverse*".

Us d' un Script amb el llenguatge R per poder construir la seqüència:
Utilitzant un script d'R[44] amb l'ajuda del programa informàtic R Studio[45] podem construir amb facilitat una seqüència consens entre els dos arxius *forward* i *reverse*, a més podem construir un cromatograma que ens mostri les *basecalls* primàries i secundaries, és a dir les primeres i segones bases amb més intensitat de la seqüenciació.
L'script utilitzat ha set dissenyat per poder obtenir seqüències consens de qualitat de tots els arxius .ab1 que tinguem, és a dir funciona a mode de *batch*.

L'script utilitza els paquets i funcions següents:

R Markdown [46,47]**:** És un marc de creació personalitzat per l'estudi de dades. Aquest tipus de fitxer .Rmd permet construir codi d'una manera molt entenedora, ja que podem separar seccions de l'arxiu en: Metadades, Text i "*Code Chunks*", o sigui blocs de codi. A més té l'avantatge que pots exportar automàticament el teu script a HTML al clic d'una tecla. A més ens ha permès mantenir l'script actualitzat constantment entre ordinadors mitjançant Git i vinculant el teu script a GitHub.

SangerseqR [48]**:** És el paquet amb més pes de tot l'script, ja que és el que s'utilitza per llegir la informació que contenen els nostres *traces* .ab1 mitjançant la funció *read.abif()*. Mitjançant la funció *sangerseq()* del mateix paquet, convertim l'*abif* a un objecte *sangerseq*. Aquest tipus d'objecte és necessari per poder realitzar funcions que ens donin els cromatogrames, i també per trobar les *basecalls* primàries i/o secundaries.
Aquest paquet també conté la funció *chromatogram()* que ens serveix per imprimir un cromatograma d'un dels nostres *traces* per així observar on es troben les bases amb millor qualitat. A més aquesta funció et permet crear un cromatograma en què es poden veure les *basecalls* primàries i secundaries. *(Figura 13)*

Trim.mott() [49] **:** Aquesta funció ens dona les posicions per on s'ha de retallar els extrems de les seqüències que tenim en forma d'objecte *sangerseq* per així quedar-nos només amb el fragment de més qualitat, ja que el principi i el final de un *trace* sol tenir pitjor qualitat que la part central. Aquesta funció es basa d'un rati (per defecte 0.33) que li serveix per detectar aquelles bases amb menys qualitat. Com més baix és el rati menys estricte és. Aquesta funció necessita a més els paquets següents: *Seqinr* [50], *ape* [51], *phangorn* [52,53], *stringi* [54], *stringr* [55], *DECIPHER* [56].

Figura 13 *Cromatograma de la regió de COI amplificada amb primers LCO/HCO d'una E. predotae (MSE_17) que mostra les "primary basecalls" i les "seconadry basecalls", així com la zona de baixa qualitat que ha eliminat la funció trim.mott en línies obliqües vermelles, i en ombrejat blau els possibles polimorfismes.*

Biostrings [57]**:** Aquest paquet conté la funció *pairwiseAlignment()* que ens permet fer un alineament entre les seqüències *Forward* i *Reverse* un cop hem eliminat els extrems que tenen pitjor qualitat utilitzant la funció *trim.mott()*. També conté la funció *consensusString()* que la utilitzem per obtenir la seqüència consens *(Figura 14)* a partir d'un alineament, que ens servirà per construir l'arbre filogenètic juntament amb la resta de seqüències consens de les altres mostres.

```
>MSE_36-Lithosia quadra-COI-5P
----------------------------------------------------------------------
TCATTAAGATTATTAATTCGAGCAGAATTAGGAAATCCAGGATCCTTAATTGGAGATGATCAAATTTATAATACTATTGTAACTGCTCATGCTTTTATTATAATTTTTTTTATAG
TTATACCTATTATAATTGGAGGATTTGGAAATTGATTAGTTCCTTTAATATTAGGAGCTCCTGATATAGCATTCCCTCGAATAAATAATATAAGTTTTTGATTATTACCCCCCTC
TTTAACCCTTTTAATTTCAAGAAGAATTGTAGAAAATGGAGCAGGAACAGGATGAACAGTTTATCCCCCACTCTCATCAAATATTGCCCATAGAGGTAGTTCCGTAGATTTAGCC
ATTTTTTCTTTACATTTAGCAGGTATTTCTTCTATTTTAGGAGCTATTAATTTTATTACCACAATCATTAATATACGATTAAATAAATTAATATTTGATCAAATACCTCTATTTG
TATGAGCTGTAGGAATTACAGCATTTTTATTACTTTTATCATTACCTGTATTAGCTGGAGCTATTACTATACTTCTAACAGATCGAAACCTCAAT
>MSE_02-Eilema caniola-COI-5P
----------------------------------------------------------------------
TTAGATTATTAATTCGAGCAGAATTAGGAAATCCTGGATCATTAATTGGAGATGATCAAATCTATAATACTATTGTAACTGCTCATGCTTTTATTATAATTTTTTTTATAGTTAT
ACCTATTATAATCGGAGGATTTGGAAATTGACTAGTTCCTCTTATATTAGGGGCCCCTGACATAGCATTCCCTCGAATAAATAACATAAGTTTTTGACTACTTCCCCCCTCTTTA
ACATTACTTATCTCAAGTAGAATTGTAGAAAATGGGGCAGGAACTGGATGAACAGTTTATCCCCCACTTTCATCTAATATTGCCCATAGAGGTAGTTCTGTAGACTTAGCTATTT
TTTCTTTACATTTAGCAGGTATTTCTTCTATTTTAGGAGCTATTAATTTTATTACAACAATTATTAATATACGATTAAATAAATTAATATTTGATCAAATACCTTTATTTGTATG
AGCCGTAGGGATTACAGCATTTTTACTACTTTTATCTTTACCAGTTTTAGCAGGAGCTATTACAATACTATTAACTGACCGAAATCTTAAT
```

Figura 14 Seqüències consens obtingudes a partir de l'script

3.8 Generació d'arbres filogenètics

Per poder generar els arbres filogenètics s'utilitza un altre script d'R Markdown que ens fa un alineament entre totes les seqüències consens i ens calcula doncs les distàncies. És a partir d'aquestes distàncies que després podem construir l'arbre filogenètic.

L'script utilitza els paquets i funcions següents:

Biostrings [57]**:** S'utilitza la funció *readDNAStringSet()* per llegir l'arxiu .fasta que conté les nostres seqüències consens que hem generat amb l'altre script juntament amb altres seqüències verificades (de cada una de les nostres espècies) descarregades de BOLD.

MSA [58]**:** Aquest paquet conté diversos algoritmes necessaris per poder realitzar l'alineament múltiple de seqüències. Conté ClustalW, ClustalOmega, i MUSCLE. Utilitzant la funció *msa()* podem alinear les nostres seqüències consens utilitzant l'algoritme que desitgem. Per defecte hem utilitzat ClustalOmega ja és el que ens ha donat millors resultats.

Seqinr [50]**:** Utilitzem aquest paquet utilitza *pairwise* per crear una matriu de distancies de les seqüències ja alineades. Ho fem amb la funció *dist.alignment()*.

Ape [51]**:** Si utilitzem la funció *njs()* d'aquest paquet amb el resultat de la matriu obtinguda a partir de *seqinr* podem generar un arbre filogenètic simple.

4. Resultats i Discussió

4.1 Optimització

Taq Polimerasa:

Es va començar per realitzar diverses amplificacions de COI de mostres molt fresques obtingudes de l'assignatura de *Molecular Biology Technics* utilitzar la *Taq-ReadyMix* provinent de *Sigma*[35] . Al veure que la amplificació no donava resultat incús en les mostres més fresques (*Figura 15*) es va decidir canviar de polimerasa i utilitzar la *NZYTaq II DNA polymerase* provinent d'*NZTtech*[36]. Aquest reactiu va donar millors resultats que la de *Sigma* i per aquest motiu es va decidir utilitzar aquesta polimerasa en la resta de reaccions. Tot i que el control positiu va sortir correcte en la de *Sigma*, no es va observar cap amplificació dels gens, en canvi si ho comparem amb la Taq polimerasa de *NZYtech* es pot observar que hi ha hagut amplificació.

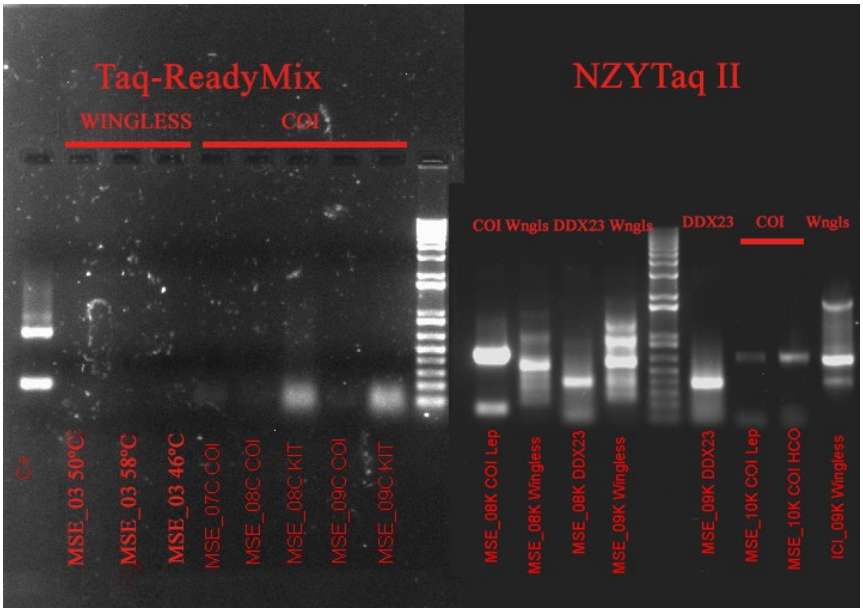


Figura 15 *Comparació entre la polimerasa provinent de Sigma i la de NZYtech.*

L'extracció utilitzant un kit comercial té més possibilitats d'èxit que una extracció utilitzant el mètode del chelex:

Per mostres fresques de *lepidòptera* capturades i de seguida guardades es poden obtenir bons resultats si es realitza una extracció amb chelex d'únicament dues extremitats de l'insecte. Per mostres seques es va comprovar que els kits comercials donaven millors resultats que l'extracció amb chelex.

Després de realitzar un total de 119 PCRs utilitzant la tècnica d'extracció amb chelex i kits comercials podem observar que els kits comercials tenen una lleugera avantatge pel que fa a nombre d'èxits. El percentatge d'èxits amb chelex és del 41,18% mentre que el percentatge amb kit és del 48,53%. Per tant tot i sortir més car hi ha més possibilitats d'èxit, i és recomanable utilitzar el kit si vols analitzar una mostra valuosa o única, ja que utilitzant chelex és més probable que no s' obtinguis resultats i es perdi alguna mostra.

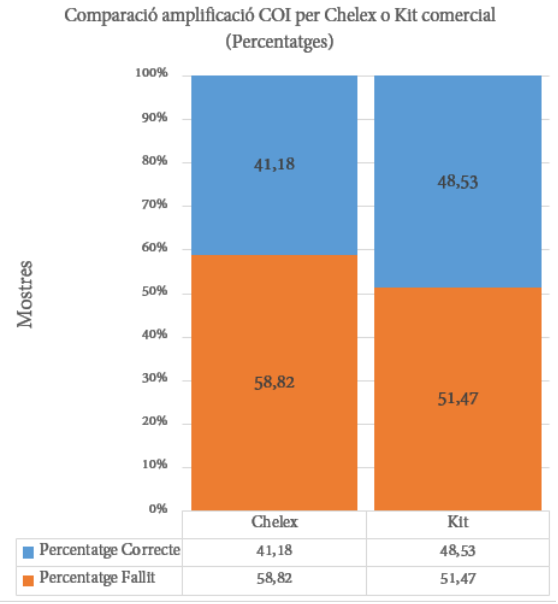


Figura 16 *Comparació entre l'extracció amb Chelex i Kit comercial mitjançant el percentatge d'èxits en l'amplificació de COI*

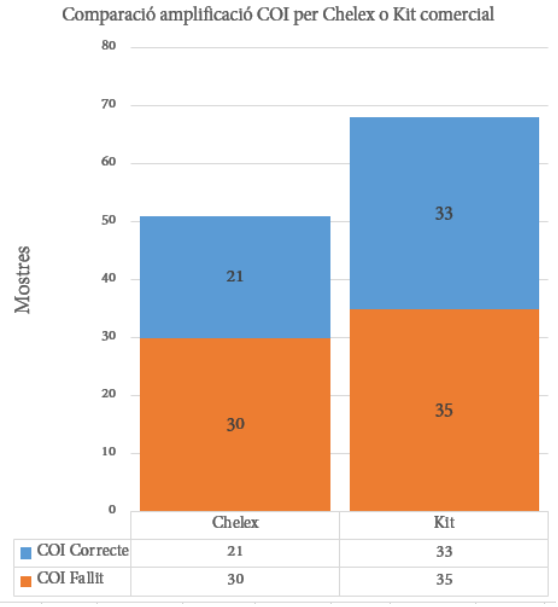


Figura 17 *Gràfica comparativa entre les amplificacions mitjançant PCR utilitzant extracció amb chelex i kit comercial*

L'estat de conservació i l'edat, un factor crucial per obtenir bones amplificacions de COI per PCR:

Les mostres de col·lecció analitzades varien molt tant en edat com en estat de conservació. Les més joves van ser capturades el 2018, mentre que les més velles es van capturar el 1994. Aquesta diferència d'edat fa pensar que les possibilitats d'èxit de les mostres velles són molt més baixes que les mostres noves.

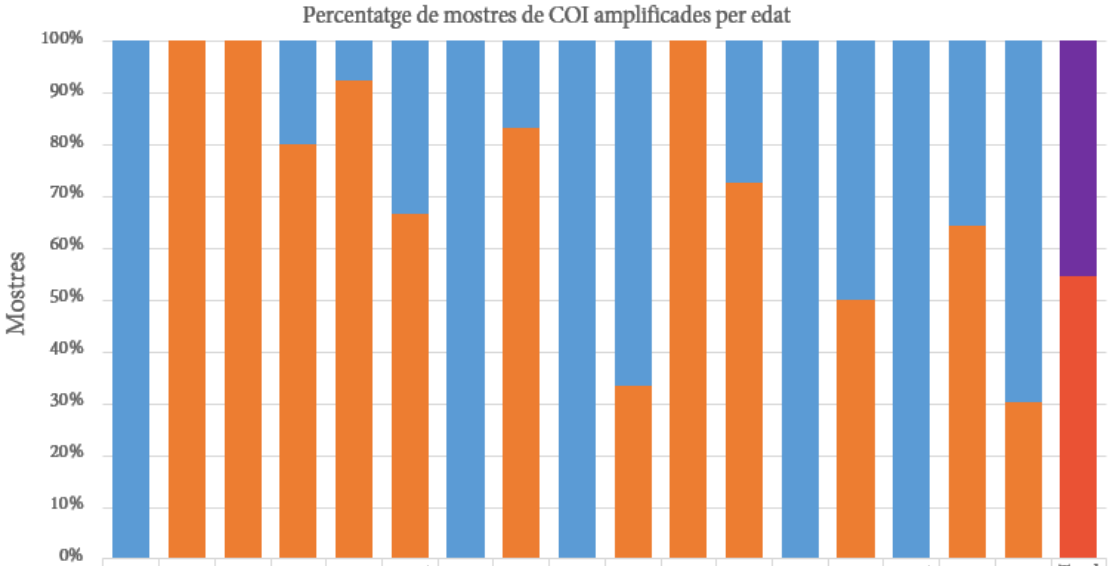
Es realitza una gràfica per comprovar-ho (*Figura 18*). S'ha de tenir en compte el factor nombre de mostres, ja que hi ha més mostres noves que velles en tota la col·lecció.

Així s'observa que mostres de fa més de deu anys, com per exemple les del 2004 tenen un percentatge d'èxit molt baix comparat amb les que van ser capturades el 2018.

Curiosament es va poder amplificar correctament la mostra més antiga, del 1994. Aquest resultat deu ser degut al fet d'utilitzar el mètode més dràstic d'extracció DNA, extraient-lo de l'abdomen amb un kit comercial.

El percentatge d'èxits total és del 45,38% i el de fracàs del 54,62%, per tant podem dir que hi ha un 50% de probabilitats d'aconseguir un resultat sigui de la mostra que sigui.

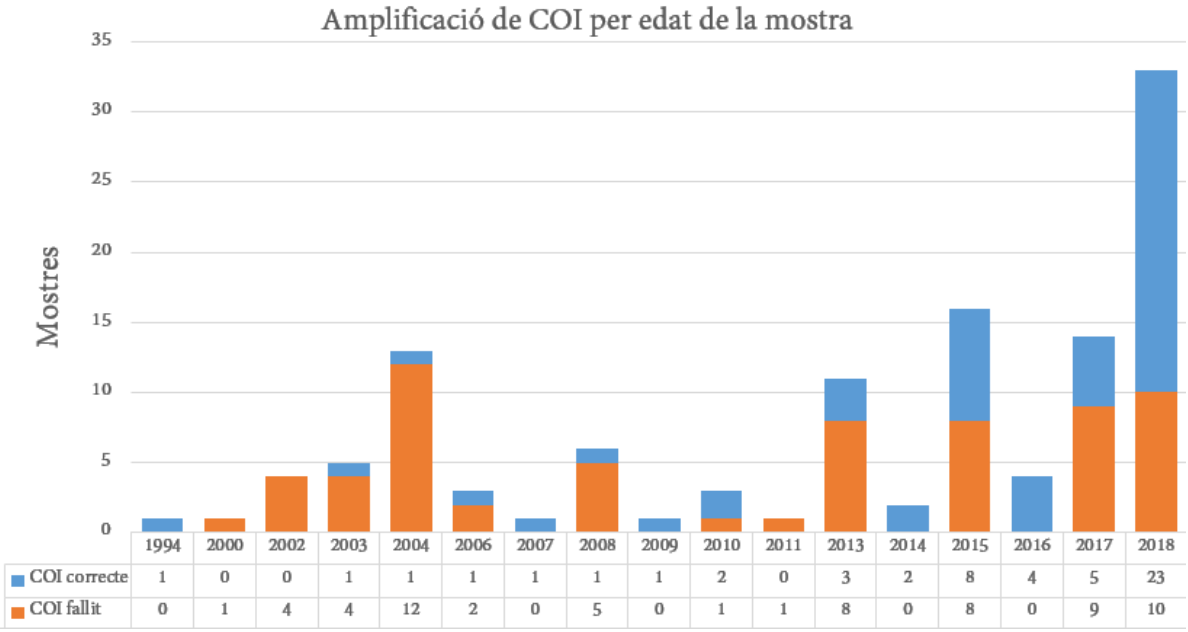
Si mirem les mostres del 2018 podem observar que el percentatge d'èxit es dispara fins al 70%, i es pot veure clarament que a mesura que es va incrementant l'edat de la mostra també comença a augmentar el percentatge de fracassos.



	1994	2000	2002	2003	2004	2006	2007	2008	2009	2010	2011	2013	2014	2015	2016	2017	2018	Total
Percentatge correcte	100	0	0	20	8	33	100	17	100	67	0	27	100	50	100	36	70	45,38
Percentatge fallit	0	100	100	80	92	67	0	83	0	33	100	73	0	50	0	64	30	54,62

Figura 18 *Percentatge d'èxit d'amplificació de COI relacionat amb l'edat de la mostra.*

S'ha de tenir en compte, però, que els es van analitzar més mostres del 2017-2018 que mostres més antigues, com les del 1994, 2000, 2006, 2007,2011 i 2014, que només es va analitzar una mostra. A la figura *Figura 19* es pot observar que hi ha moltes més mostres noves que velles, i per tant els percentatges del 100% d'èxit del 1994, el 2007 i el 2009, per exemple, no són fiables.

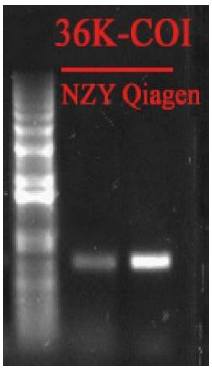


	1994	2000	2002	2003	2004	2006	2007	2008	2009	2010	2011	2013	2014	2015	2016	2017	2018
COI correcte	1	0	0	1	1	1	1	1	1	2	0	3	2	8	4	5	23
COI fallit	0	1	4	4	12	2	0	5	0	1	1	8	0	8	0	9	10

Figura 19 *Nombre de mostres amplificades correctament o malament relacionat amb l'any de captura.*

Diferents kits comercials, mateixos resultats:

Tant per l'extracció com per la purificació es van utilitzar dos kits comercials de cases diferents. Per l'extracció es va utilitzar el *DNeasy® Blood & Tissue kit de Qiagen* [32] *i el NZY® Tissue gDNA Isolation Kit de NZYtech*[33], per la purificació el kit de *Qiagen QIAquick® Gel Extraction kit* [38], i més endavant el kit d'*NZYtech NZYGelpure®* [39].



Pel que respecta a l'extracció es va observar que les mostres extretes utilitzant el kit de Qiagen donaven uns millors resultats en la PCR que les obtingudes a partir del kit comercial de NZYtech (*Figura 20*). Però no es poden treure conclusions definitives ja que aquest resultat poden ser degut a que la quantitat de DNA present en la pota extreta era menor en NZYtech que en Qiagen, per aquest motiu es va utilitzar principalment el kit de Qiagen, però quan aquest es va acabar es va utilitzar el kit d'NZYtech i no va donar cap problema.

Figura 20 *PCR on es compara la qualitat de l'extracte del kit d'extracció d'NZYtech i el kit de Qiagen*

Pel que respecta a la purificació del producte de PCR es va comprovar que tots dos kits comercials donaven els mateixos resultats (*Taula 11*), una quantitat força baixa de DNA en tots dos, però de bona qualitat. La baixa concentració de DNA en aquestes mostres és esperable degut a l'edat i l'estat de conservació.

En el cas de *NZYtech NZYGelpure®* [39], es pot observar que el rati A260/A280 és més elevat, el que indica una concentració elevada de contaminants com ara proteïnes. Aquest valor, però, està influenciat també pel valor del pH, per tant no podem assegurar que tingui contaminants.

Taula 11 *Programes del termocycler utilitzats per la PCR*

Kit de purificació Comercial	A260/A280 (Puresa)	A260/230 (Contaminants)	Concentració (ng/µL)
QIAquick ® Gel Extraction kit [37]	2	0,02	11
NZYtech NZYGelpure® [38]	2,455	0,012	13,5

Purificació utilitzant el mètode enzimàtic:

Pel que respecta a la purificació enzimàtica utilitzant Exonucleasa I i *Shrimp Alkaline Phosphatase*, es va realitzar una purificació utilitzant aquest mètode i es van obtenir els mateixos resultats que utilitzant un kit comercial (*Figura 21*).

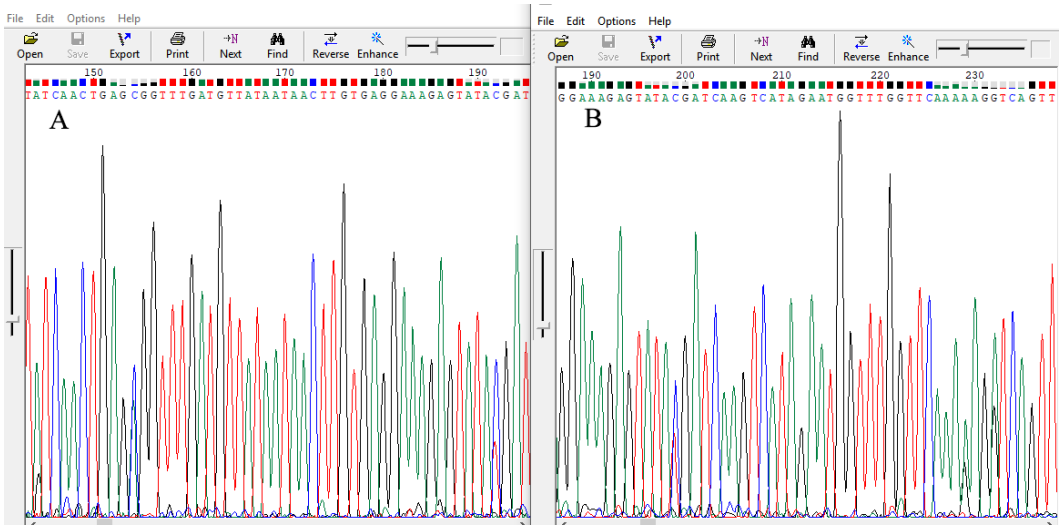


Figura 21 *Comparació entre dues seqüències .ab1. La seqüència A va ser purificada utilitzant un kit comercial, la figura B va ser purificada utilitzant el mètode enzimàtic.*

***COI* pre-amplificació o sense:** Es van utilitzar dos protocols de PCR per l'amplificació del COI, un d'aquests conté una etapa de pre-amplificació a 45ºC durant 5 cicles i 35 cicles a 50ºC. Es va voler comprovar si aquest pas augmentava la quantitat de DNA amplificat, per tant es va comparar amb un altre protocol de PCR que elimina aquesta pre-amplificació i fa només 35 cicles a 50ºC.

Com es pot observar en el gel (***Figura 22***) hi ha poca diferència entre l'un i l'altre, les bandes que surten més fortes sense pre-amplificació tenen una mica més d'intensitat amb pre-amplificació, mentre que passa el contrari amb les bandes més fluixes.
Si per exemple ens mirem la mostra 68K podem observar que no s'observa cap banda amb pre-amplificació, en canvi quan es fa una PCR de la mateixa mostra utilitzant un protocol sense els 5 cicles a 45ºC s'observa que apareix una banda molt fina. Sens dubte aquesta banda no és suficient per enviar a seqüenciar i obtenir bons resultats, però ens indica que el protocol sense pre-amplificació pot tenir una lleugera avantatge sobre l'altre protocol.

Per aquest motiu és recomanable no utilitzar pre-amplificació en mostres difícils, ja que es pot es pot estalviar temps utilitzant un protocol sense pre-amplificació.

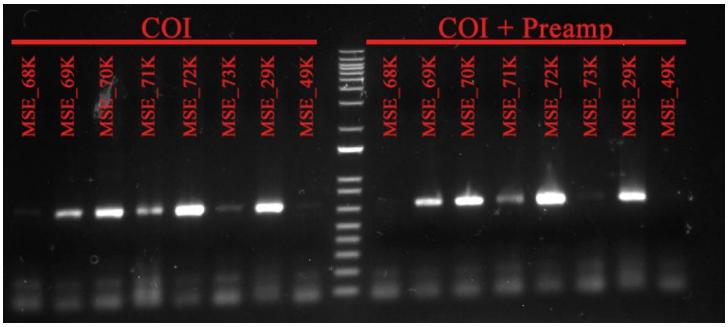


Figura 22 Gel que mostra el resultat de una PCR sense pre-amplificació del DNA, i una utilitzant el protocol que realitza 5 cicles a 45ºC.

També es va provar d'utilitzar un protocol per gens nuclears amb 5 cicles de pre-amplificació a 45ºC i 35 cicles a 55ºC per amplificar el gen nuclear ArgK. Els resultats obtinguts de la PCR observats amb un gel d'agarosa (Figura 23) ens indiquen que no és recomanable utilitzar un pas de pre-amplificació ja que augmenten molt els anellaments no específics i després s'obtenen resultats poc definits.

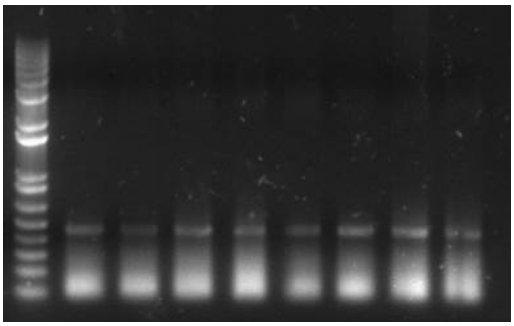


Figura 23 *Gel d'agarosa que mostra les bandes obtingudes després de fer 5 cicles de pre-amplificació a 45ºC i 35 cicles a 55ºC per amplificar el gen nuclear ArgK.*

Comparació entre els primers de COI LepF1/LepR1 i LCO/HCO:

Es van utilitzar dos parells de primers per l'amplificació del COI: LepF1/LepR1 i LCO/HCO. Tots dos amplifiquen regions semblants d'aquest gen, i obtens productes de la mateixa llargada d'unes 600pb.
Es va comprovar que en alguns casos LCO/HCO donava millors resultats que LepF1/LepR1.

En el gel es pot veure com en la mostra MSE_08K s'obtenen millors resultats utilitzant el parell de primers LCO/HCO que utilitzant LepF1/LepR1, en canvi la resta de mostres es pot veure que no hi ha diferencies visibles. Per tant es va decidir utilitzar-los per defecte i fer servir els altres en mostres concretes en que LCO/HCO no donessin resultat.

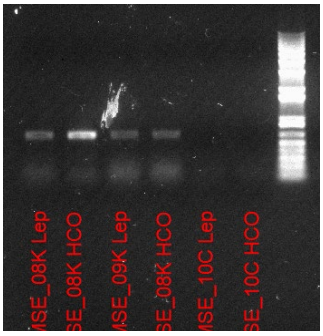


Figura 24 *Comparació entre els dos parells de primers per COI utilitzats.*

Possibilitat d'utilitzar el gen Wingless en *Eilema:*
Es van dur a terme una sèrie d'experiments per determinar si aquest gen podria servir com a candidat per l'estudi. Es van utilitzar tres *primers* diferents, WinglesFWD i Wingless2 i 2a pel REV.
Es van fer diverses PCRs amb diferents paràmetres i es va comprovar que la qualitat de l'amplificació no era prou bona per poder-lo utilitzar com a *barcode*, principalment degut a amplificacions de regions errònies degut a un mal anellament dels *primers*.

Degut als primers resultats que ens indicaven un mal anellament dels *primers* es va decidir fer una Touchdown PCR per intentar augmentar l'especificitat d'aquests. Aquests tipus de PCR consisteixen en fer uns cicles a una temperatura d'anellament elevada i anar-la disminuint fins a la temperatura d'amplificació desitjada. Al augmentar la temperatura l'anellament es fa més específic, fet que redueix el nombre de *primers* mal anellats.

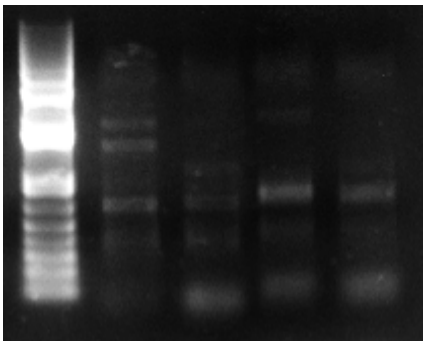


Figura 25 *Touchdown PCR pel gen Wingless*

Malgrat tot encara obteníem anellaments incorrectes (*Figura 25*). Per tant vam decidir abandonar aquest gen i centrar-nos en altres gens nuclears com ara el DDX23, ArgK, PSB, ProSup, CAD, SSU72.

Banc de gens nuclears :

Després de descartar el gen Wingless es va realitzar un seguit de PCRs per determinar possibles gens nuclears que es podrien utilitzar com a complement per l'estudi.

Es va començar per estudiar la possibilitat d'utilitzar els gens nuclears ArgK i DDX23. En algunes mostres es va aconseguir amplificació de DDX23 utilitzant un programa de PCR a 55ºC, però en la majoria de mostres en què es va realitzar una PCR per amplificar DDX23 es va observar que no donava bons resultats.

En canvi ArgK era més fiable que DDX23, ja que com es pot veure en el gel obtenim amplificació d'aquest gen en les mostres que no hem obtingut amplificació de DDX23.

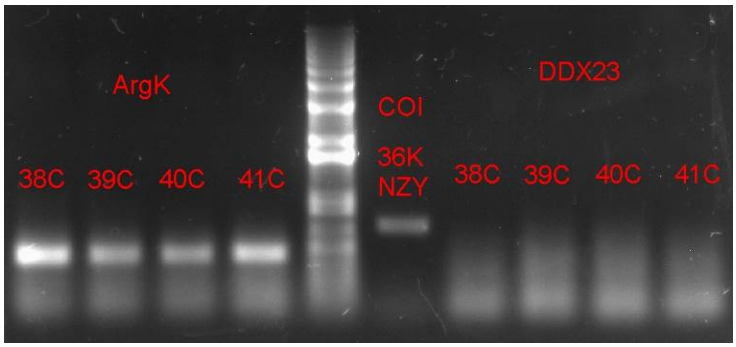


Figura 26 *Gel d'agarosa en que es veu el resultat de l'amplificació de ArgK i DDX23 per les mateixes mostres.*

Més endavant es va observar que ArgK no era prou consistent (*Figura 27*), ja que tot i que no hi ha problema en amplificar aquest gen en mostres fresques, quan s'utilitza en mostres més velles de la col·lecció l'amplificació es redueix molt i pràcticament no s'obté amplificació. Es van provar varies temperatures (52, 53, 54, 55, 56, 57ºC), però no es va obtenir cap resultat que indiqués que la temperatura tenia alguna influència en la falta de resultats.

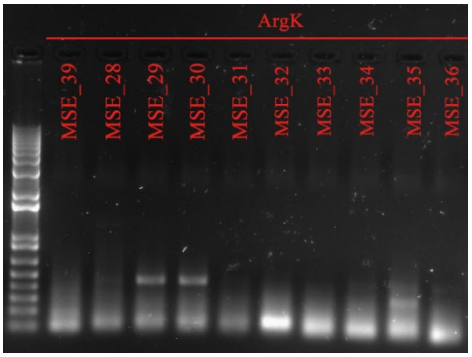


Figura 27 *Gel d'agarosa on es mostra el resultat de l'amplificació d'ArgK de diverses mostres d'Eilema.*

Es va realitzar un banc de tots els gens nuclears disponibles (Psb, SSU72, CAD, IDH, MDH, Ca-ATPasa, ProSup). Es va utilitza la mostra MSE_08K, una *Eilema sororcula* extreta utilitzant kit i que va haver de ser diluïda degut a la seva gran quantitat de DNA. Aquesta mostra de tanta qualitat ens va servir per realitzar una PCR amb tots els primers d'aquests gens. Es va utilitzar el protocol LEPNUC: [1 min a 94ºC, 1,5 min a 55ºC, 1 min a 72ºC]x35 cicles.

El resultat obtingut d'aquesta PCR (Figura 28) ens mostra tres possibles candidats, el gen Psb, el SSU72 i el ProSup.

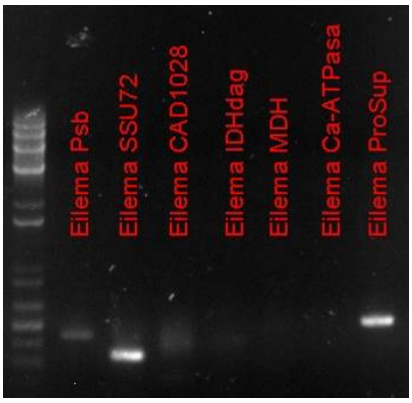


Figura 28 *Resultats obtinguts del banc de nuclears realitzat amb la mostra*

Sabent això es van realitzar varies extraccions de mostres seques per comprovar la seva viabilitat en aquestes mostres més complicades.

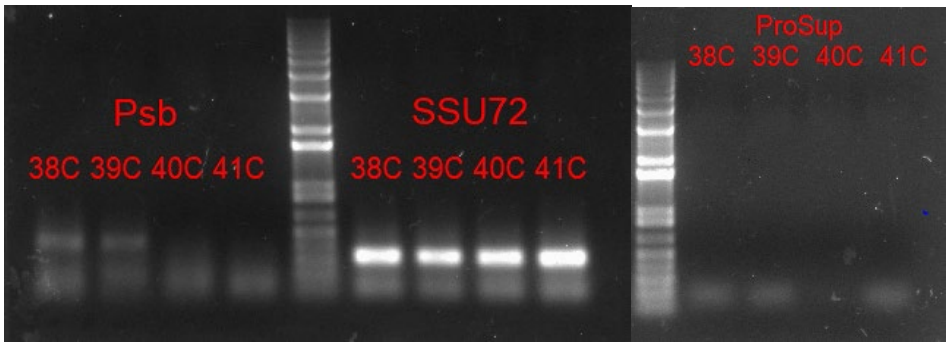


Figura 29 *Resultat de PCR en que es veu l'amplificació de Psb, SSU72 i ProSup per les mateixes mostres.*

Es pot observar (***Figura 29***) que l'únic gen que va donar bona amplificació amb totes és l'SSU72. Per aquest motiu es va decidir utilitzar aquest per complementar l'estudi sobre el COI. Malgrat els seus bons resultats algunes mostres sequenciades ens van arribar en mal estat, molt probablement degut a un error en la purificació. Per aquest motiu moltes mostres a les que es va realitzar l'amplificació d'aquest ls hi falta la seqüència.

4.2 Estudi bioinformàtic de les dades:

Es van obtenir un total de 114 *traces*, formats per seqüències *forward* i *reverse*. Aquestes seqüències es van alinear utilitzant l'script d'R Markdown [59]. A partir d'aquests *traces* es van obtenir un total de 57 seqüències consens, juntament amb 57 cromatogrames (un per cada una). A partir del cromatograma es va poder observar la qualitat de cada una d'elles i descartar les que es tenien una qualitat baixa o es veia alguna anomalia (pics irregulars, molt poca seqüència conservada després de retallar-la per la qualitat).

Com a exemple, en el cromatograma de la mostra MSE_02 podem observar que cada base té un color assignat, i també veiem que s'hi representen les dues basecalls a sobre dels pics, aquestes són les primeres i segones bases més probables de trobar en aquell punt de la seqüència. Allà on hi ha un sombrejat blau és on hi ha una diferencia entre aquests basecalls, i això pot indicar o bé un polimorfisme o un error durant la sequenciació. Als extrems es pot observar unes regions ratllades en vermell, aquestes regions són les que l'script elimina mitjançant la informació de la funció *trim.mott()* , que ens indica les regions a on es troba la nostra seqüència de qualitat, i les regions de pitjor qualitat que poden ser eliminats.

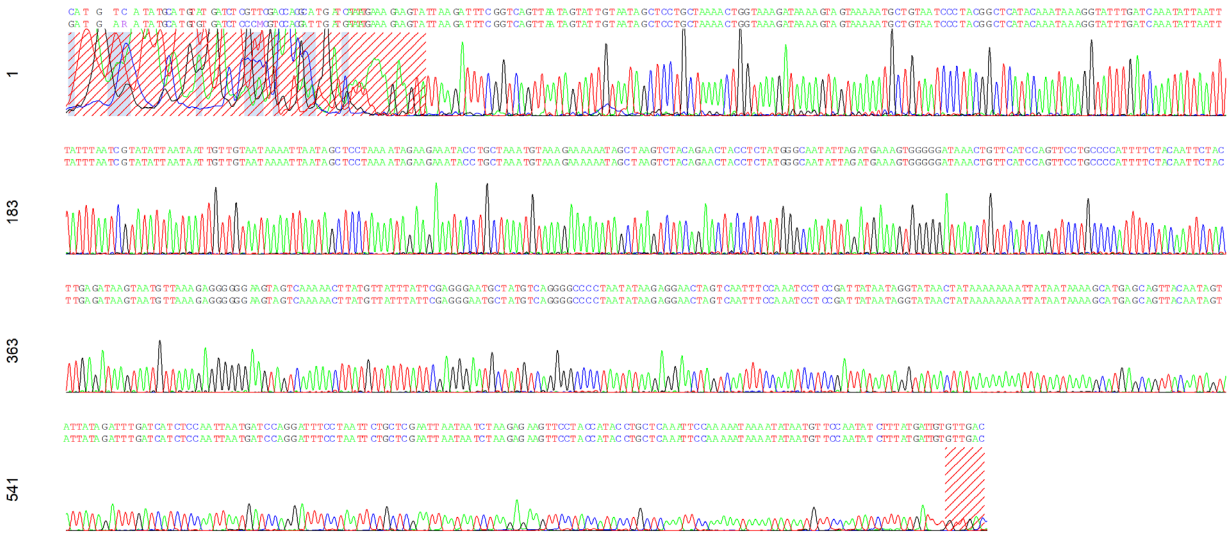


Figura 30 *Cromatograma generat per l'script d'R Markdown per la seqüència de COI de la mostra MSE_02 (Eilema caniola)*

4.3 Arbre filogenètic

A partir d'aquestes seqüències concens es va realitzar un alineament amb ClustalW utilitzant R Markdawn [60], i es va generar un arbre filogenètic per cada un dels gens que teniem: COI, ArgK, SSU72.

L'arbre filogenètic construït a partir dels *barcodes* del COI es va generar a partir de 63 seqüències consens provinents de 19 especies diferents, 37 de les seqüències provenen de BOLD, ja que ens serveixen com a referència. Es van haver d'eliminar algunes seqüències per la seva qualitat per poder obtenir un millor arbre. L'arbre té una línia a escala que mostra un 5% de diferència.

Si ens mirem l'arbre (*Figura 31*) des de dalt observarem el següent:
1 - La *Eilema marcida* provinent de BOLD està relacionada amb les dues *Eilema caniola* (MSE_46, MBT_10) extretes al laboratori ja que comparteixen una branca comuna, és possible ja que les dues són molt semblants físicament, però hi pot haver un error d'identificació.
2 - La *Eilema predotae* MSE_17 s'agrupa correctament amb la mostra provinent de BOLD.
3 - MSE_21 s'agrupa correctament amb la *Eilema lurideola* provinent de BOLD.
4 - Totes les *Eilema interpositella* s'agrupen correctament juntament amb la provinent de BOLD.
5 - La *Eilema albicosta* s'agrupa amb la seva parella de BOLD, i la *Eilema albicosta witti*, una subespècie, queda alineada en una branca molt propera, però es pot observar que hi ha una lleugera diferència.
6 - Les *Eilema uniola* s'agrupen entre elles, per tant això és correcte.
7 - Unes altres *Eilema caniola* s'agrupen separades de més amunt, això pot tenir dues explicacions, o bé les dues *Eilema caniola* que s'agrupen prop d'una *eilama marcida* són en realitat *marcida*. O bé formen part d'un altre BIN. Però és més probable la primera opció ja que únicament existeix un bin per *caniola* i físicament les *Eilema marcida* i *Eilema caniola* són molt semblants entre elles i un error d'identificació no seria estrany. A més les *caniola torstenii,* una subespècie, s'agrupen correctament amb el grup de *canioles* que inclou la de BOLD.
8 - La *Eilema palliatella* s'agrupa correctament amb el seu homòleg de BOLD, per tant aquesta també podem dir que és correcta.
9 - Les *Eilema complana* i *pseudocomplana* tenen un nòdul en comú, juntament també amb la subespècie *complana iberica* per tant podem concloure que estan aparellades correctament.

10 - La *Eilema pygmaeola pallifrons*, una subespècie de la *Eilema pygmaeola*, s'aparella amb una *pygmaeola* de BOLD, que pot ser que sigui de la mateixa subespècie, però també pot ser que la nostra subespècie no sigui *pallifrons* i sigui *pygmaeola pygmaeola*.
11 - Les *Eilema lutarella* de BOLD s'agrupen entre elles.
12 - Un grup d'*Eilema interpositella* provinents de les pràctiques de l'assignatura *Molecular Biology Techniques* s'agrupa per separat de les de BOLD i les que han set analitzades en aquest treball. I curiosament comparteixen una semblança bastant important amb *Eilema rungsi* ja que comparteixen node.
13 - Un grup d'*Eilema sororcula* s'agrupa correctament juntament amb la de BOLD, hi ha una separació entre les dos branques, segurament degut a que *sororcula* té diversos BINs, i potser alguna de les nostres mostres correspon a BINs diferents.
14 - La mostra MSE_08 hauria de ser una *Eilema sororcula,* però es classifica juntament amb les *Eilema depressa*, obtenim el mateix resultat si ens mirem l'arbre de l'ArgK (*Figura 32*). Per tant aquesta mostra és possible que sigui una *depressa* tot i que visualment sembla una *sororcula*.
15 - Les mostres MSE_01 i MSE_71 es troben separades de la resta segurament degut a la qualitat de la seqüència, aquestes mostres contenien molt poc DNA amplificat.
16 - Les *Eilema costalis* s'agrupen entre elles a una distància molt gran de la resta. Això s'explica perquè la qualitat de la seqüència era bastant baixa, però tot i això s'agrupen.
17 - La *Eilema griseola* s'agrupa amb la seva homòloga de BOLD, per tant el resultat és l'esperat.
18 - L'*outgrup* format per *Lithosia quadra* conté un *Eilema bipuncta*. Aquest resultat tant inusual fa pensar que potser aquesta *Eilema bipuncta* podria ser una *Lithosia quadra*, ja que les femelles de *Lithosia quadra* presenten dos taques ocelars a les ales igual que la *bipuncta,* si bé que la mida de l'insecte és molt més gran.

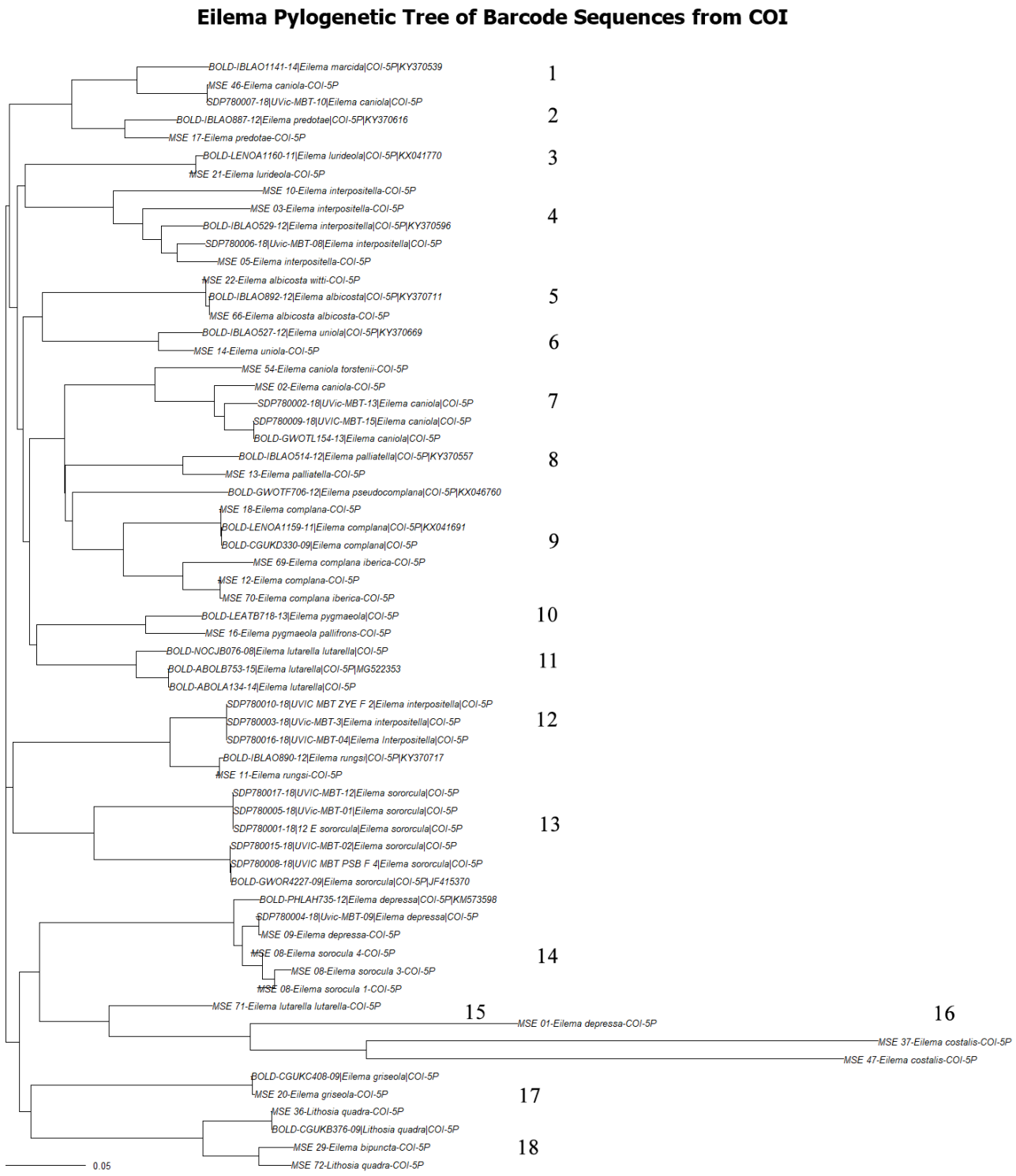


Figura 31 *Arbre filogenètic de les diferents espècies generat per R. Construït a partir de les seqüències consens de COI obtingudes.*

Els arbres dels gens nuclears com SSU72 i ArgK no ens acaben de donar els resultats desitjats (*Figura 32*), ja que ens falten seqüències i les que tenim ens donen uns resultats no gaire bons. En el cas de l'SSU72 a l'arbre s'hi ha incorporat els resultats *forward* de diverses mostres en què no tenim el *reverse*. Podem observar que les dues mostres MSE_50, *bipuncta*, s'alinean correctament com és d'esperar ja que són la mateixa mostra, però diferents PCRs. Les dues *caniola* MSE_52 i MSE_53 s'alineen correctament l'una amb l'altra. La resta de mostres pateixen errors d'alineament. Els mateixos errors es poden observar en l'arbre d'ArgK.

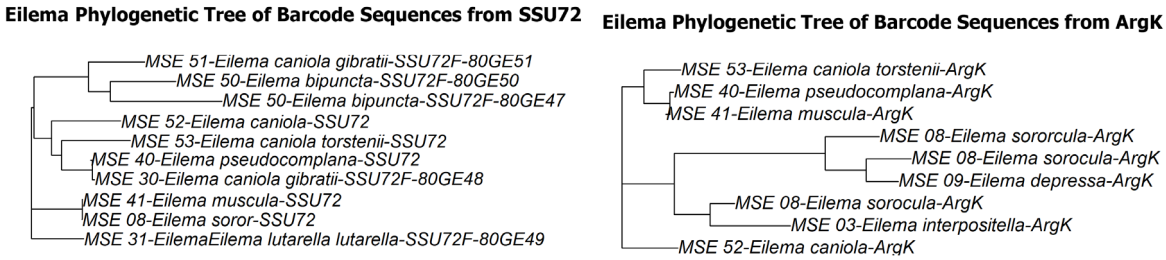


Figura 32 *Arbres filogenètics construïts per les seqüències de SSU72 i ArgK*

5. Conclusions

Les millors condicions per a realitzar una amplificació del gen COI en *Eilema* són les següents:

1- Utilitzar mostres capturades recentment. Utilitzar chelex en aquestes mostres ja que és molt més econòmic que un kit comercial.
2- En el cas que la mostra sigui de col·lecció i/o vella, és recomanable començar per una extracció amb chelex i si no s'obtenen resultats després es pot utilitzar un kit comercial per realitzar l'extracció de DNA de la mostra, ja que pot donar millors resultats en aquests tipus de casos.
3- Utilitzar els primers LCO/HCO ja que s'ha comprovat que donen millors resultats que LepF1/LepR1 aquest gènere.
4- Per la PCR és recomanable utilitzar un protocol sense pre-amplificació com ara el següent: [1 min a 94ºC, 1,5 min a 50ºC, 1 min a 72ºC]x35 cicles. Ja que dona els mateixos resultats o sinó millors que utilitzant pre-amplificació, i a més dura menys.
5- Per la purificació de la mostra es pot utilitzar un kit comercial, però s'obtenen els mateixos resultats utilitzant una purificació enzimàtica, i és més ràpid i econòmic.

Apart de COI es pot utilitzar el gen ArgK com a diana per amplificar mostres en bon estat de conservació. El gen SSU72 té una taxa d'èxit molt elevada en mostres velles comparat amb ArgK que és molt més difícil d'amplificar.

Utilitzant un script de R Markdown es poden generar seqüències consens, cromatogrames, alineaments i arbres amb molta facilitat i en *batch*. Per tant és una solució molt útil si tens moltes seqüències que has d'analitzar.

6. Agraïments

A en Pep Bau, en Ramòn Macià, a la Montse Masoliver, a l'Ignasi Calba, i a tots els tècnics, personal, i investigadors dels laboratoris de la UVic.

7. Bibliografia

1. J, Ylla., Macià, R., Gastón, F. J. La familia Arctiidae Leach [1815]. in *Manual de identificación y guia de campo de los Árctidos de Península Ibérica y Baleares* (ed. Argania editio) 11 (2010).
2. J, Ylla., Macià, R., Gastón, F. J. Subfamilias. in *Manual de Identificación y guía de campo de los árctidos de la Península Ibèrica y Baleares* (ed. Editio, A.) 290 (2010).
3. Dubatolov, V. V. & Zolotuhin, V. V. Does Eilema Hübner, [1819] (Lepidoptera, Arctiidae, Lithosiinae) present one or several genera? *Euroasian Entomol. J.* (2011).
4. J, Ylla., Macià, R., Gastón, F. J. Lithosiinae: Lithosiini. in *Manual de identificación y guia de campo de los Árctidos de Península Ibérica y Baleares* (ed. Editio, A.) 290, 63, 67, 70, 73, 76, 80, 84, 88, 91, 95, 99, 1 (2010).
5. Consulta de especie Eilema albicosta witti Kobes, 1993. *Banco de Datos de Biodiversidad de Canarias* (2019). Available at: http://www.biodiversidadcanarias.es/atlantis/admin/adminEspecieConsulta.jsf?especieCode=A00330. (Accessed: 31st May 2019)
6. Hebert, P. D. N., Cywinska, A., Ball, S. L. & DeWaard, J. R. Biological identifications through DNA barcodes. *Proc. R. Soc. B Biol. Sci.* **270**, 313–321 (2003).
7. Arnot, D. E., Roper, C. & Bayoumi, R. A. L. Digital codes from hypervariable tandemly repeated DNA sequences in the Plasmodium falciparum circumsporozoite gene can genetically barcode isolates. *Mol. Biochem. Parasitol.* (1993). doi:10.1016/0166-6851(93)90154-P
8. V3.boldsystems.org. BOLD Systems: Management & Analysis. (2019). Available at: v3.boldsystems.org/index.php/MAS_DataRetrieval_OpenSequence?selectedrecordid=9961309. (Accessed: 1st June 2019)
9. Hanner, R. *BARCODE, Data Standards for (BRIs), Records in INSDC*. (2009).
10. Ratnasingham, S. & Hebert, P. D. N. BOLD: The Barcode of Life Data System: Barcoding. *Mol. Ecol. Notes* (2007). doi:10.1111/j.1471-8286.2007.01678.x
11. Ratnasingham, S. & Hebert, P. D. N. A DNA-Based Registry for All Animal Species: The Barcode Index Number (BIN) System. *PLoS One* **8**, e66213 (2013).
12. Valentini, A., Pompanon, F. & Taberlet, P. DNA barcoding for ecologists. *Trends in Ecology and Evolution* (2009). doi:10.1016/j.tree.2008.09.011
13. Saccone, C., De Giorgi, C., Gissi, C., Pesole, G. & Reyes, A. Evolutionary genomics in Metazoa: The mitochondrial DNA as a model system. *Gene* (1999). doi:10.1016/S0378-1119(99)00270-X
14. BOLD. BARCODE INDEX NUMBERS. *2014* (2019). Available at: http://www.boldsystems.org/index.php/Public_BarcodeIndexNumber_Home. (Accessed: 29th May 2019)
15. Alberts, B. A. J. *et al. Molecular Biology of the Cell.* (Published by Garland Science, Taylor & Francis Group, LLC, an informa business, 711 Third Avenue, New York, NY 10017, US, 2015).
16. Hebert, P. D. N., Ratnasingham, S. & de Waard, J. R. Barcoding animal life: cytochrome *c*

oxidase subunit 1 divergences among closely related species. *Proc. R. Soc. London. Ser. B Biol. Sci.* **270**, S96-9 (2003).

17. Chen, D. Bin *et al.* Comparative mitochondrial genomes provide new insights into the true wild progenitor and origin of domestic silkworm Bombyx mori. *Int. J. Biol. Macromol.* (2019). doi:10.1016/j.ijbiomac.2019.03.002
18. Wahlberg, N., Peña, C., Ahola, M., Wheat, C. W. & Rota, J. PCR primers for 30 novel gene regions in the nuclear genomes of Lepidoptera. *Zookeys* **596**, 129–141 (2016).
19. Uniprot. RNA polymerase II subunit A C-terminal domain phosphatase SSU72. Available at: https://www.uniprot.org/uniprot/P53538. (Accessed: 29th May 2019)
20. Chen, X. *et al.* Isolation of arginine kinase from Apis cerana cerana and its possible involvement in response to adverse stress. *Cell Stress Chaperones* (2015). doi:10.1007/s12192-014-0535-2
21. Uniprot. P09615 (WNTG_DROME). Available at: https://www.uniprot.org/uniprot/P09615. (Accessed: 29th May 2019)
22. Brower, A. V. Z. & DeSalle, R. Patterns of mitochondrial versus nuclear DNA sequence divergence among nymphalid butterflies: the utility of wingless as a source of characters for phylogenetic inference. *Insect Mol. Biol.* **7**, 73–82 (1998).
23. Nei, M. & Sudhir, K. Phylogenetic Trees. in *Molecular Evolution and Phylogenetics* 348, 73–83 (Oxford University Press, 2000).
24. Nei, M. & Sudhir, K. Evolutionary Change of DNA sequences. in *Molecular Evolution and Phylogenetics* 384, 33–41 (Cambridge University Press).
25. EMBL-EBI. Topology | EMBL-EBI Train online. (2019). Available at: https://www.ebi.ac.uk/training/online/course/introduction-phylogenetics/what-phylogeny/aspects-phylogenies/topology. (Accessed: 2nd June 2019)
26. Nei, M. Distance Measure. in *Molecular Evolution and Phylogenetics* 348, 112–113 (Oxford University Press, 2000).
27. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–25 (1987).
28. Geology.com. france-map.gif. (2019). Available at: https://geology.com/world/france-map.gif. (Accessed: 31st May 2019)
29. Geographicguide.net. cyprus-europe.gif. (2019). Available at: http://www.geographicguide.net/europe/maps-europe/maps/cyprus-europe.gif. (Accessed: 31st May 2019)
30. Smartdraw.com. SmartDraw - Create Flowcharts, Floor Plans, and Other Diagrams on Any Device. (2019). Available at: https://www.smartdraw.com.
31. Bio-rad.com. (2019). Available at: https://www.bio-rad.com/en-us/product/chelex-100-resin?ID=6448ab3e-b96a-4162-9124-7b7d2330288e. (Accessed: 25th May 2019)
32. Qiagen. DNeasy Blood & Tissue Kits. (2019). Available at: https://www.qiagen.com/fi/products/discovery-and-translational-research/dna-rna-purification/dna-purification/genomic-dna/dneasy-blood-and-tissue-kit/#orderinginformation. (Accessed: 3rd June 2019)

33. NZYtech. NZY Tissue gDNA Isolation kit | Genomic DNA Purification | NZYTech. (2019). Available at: https://www.nzytech.com/products-services/molecular-biology/dnarna-purification/genomic-dna-purification/mb135/. (Accessed: 2nd June 2019)
34. Group, T. N. S. Molecular methods. (2019). Available at: http://www.nymphalidae.net/Molecular.htm. (Accessed: 14th February 2019)
35. JumpStart™ Taq ReadyMix™ for High Throughput Quantitative PCR Ready-to-use 2x mix for qPCR with ROX | Sigma-Aldrich. Available at: https://www.sigmaaldrich.com/catalog/product/sigma/d6442?lang=es&region=ES. (Accessed: 3rd June 2019)
36. NZYTaq II DNA polymerase | DNA Polymerases | NZYTech. Available at: https://www.nzytech.com/products-services/molecular-biology/end-point-pcr/dna-polymerases/mb354/. (Accessed: 3rd June 2019)
37. Wilmington, D. U. *T042-TECHNICAL BULLETIN NanoDrop Spectrophotometers. Thermo Fisher Scientific - NanoDrop products* (2007). doi:10.1002/jobm.19770170116
38. Qiagen. QIAquick Gel Extraction Kit. (2019). Available at: https://www.qiagen.com/fi/products/discovery-translational-research/dna-rna-purification/dna-purification/dna-clean-up/qiaquick-gel-extraction-kit/#orderinginformation. (Accessed: 3rd May 2019)
39. NZYGelpure | DNA Clean-up | NZYTech. Available at: https://www.nzytech.com/products-services/molecular-biology/dnarna-purification/nucleic-acids-clean-up/dna-clean-up/mb011/. (Accessed: 3rd June 2019)
40. New England BioLabs Inc. Exonuclease I (E. coli) | NEB. Available at: https://international.neb.com/products/m0293-exonuclease-i-e-coli#Product Information. (Accessed: 2nd June 2019)
41. New England BioLabs Inc. Shrimp Alkaline Phosphatase (rSAP) | NEB. Available at: https://www.neb.com/products/m0371-shrimp-alkaline-phosphatase-rsap#Product Information. (Accessed: 2nd June 2019)
42. New England BioLabs Inc. Enzymatic PCR Cleanup Protocol | NEB. Available at: https://international.neb.com/protocols/2017/07/10/enzymatic-pcr-cleanup-protocol. (Accessed: 2nd June 2019)
43. Chromas | Technelysium Pty Ltd. Available at: https://technelysium.com.au/wp/chromas/. (Accessed: 4th June 2019)
44. Team, R. C. R: A Language and Environment for Statistical Computing. (2019).
45. RStudio Team. RStudio 1.2.1335 Integrated Development Environment for R. (2018).
46. And, J. A. and Y. X. and J. M., Hadley, J. L. and K. U. and A. A. and & Iannone, W. and J. C. and W. C. and R. rmarkdown: Dynamic Documents for R. (2019).
47. Grolemund, Y. X. and J. J. A. and G. *R Markdown: The Definitive Guide*. (Chapman and Hall/CRC, 2018).
48. Hill, J. *et al.* Poly peak parser: Method and software for identification of unknown indels using sanger sequencing of polymerase chain reaction products. (2014).
49. Lanfear, R. sangeranalyseR/trim.mott.R at master · roblanf/sangeranalyseR. *GitHub*

Available at: https://github.com/roblanf/sangeranalyseR/blob/master/R/trim.mott.R. (Accessed: 3rd June 2019)

50. Charif, D. & Lobry, J. R. Seqin{R} 1.0-2: a contributed package to the {R} project for statistical computing devoted to biological sequences retrieval and analysis. (2007).
51. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in {R}. *Bioinformatics* **35**, 526–528 (2018).
52. Schliep, K. P. phangorn: phylogenetic analysis in R. *Bioinformatics* **27**, 592–593 (2011).
53. Schliep *et al.* Intertwining phylogenetic trees and networks. *Methods Ecol. Evol.* **8**, 1212–1220 (2017).
54. Gagolewski, M. R package stringi: Character string processing facilities. (2019).
55. Wickham, H. Stringr: Simple, Consistent wrappers for common String Operations. (2019).
56. Wright, E. A toolset for deciphering and managing biological sequences. *R J.* **8**, 352–359 (2016).
57. Pagès, H., Aboyoun, P., Gentleman, R. & DebRoy, S. Biostrings: Efficient manipulation of biological strings. (2019).
58. Bodenhofer, U., Bonatesta, E., Horejš-Kainrath, C. & Hochreiter, S. msa: an R package for multiple sequence alignment. *Bioinformatics* **31**, btv494 (2015).
59. PepBau/SangerSeq: Sanger Sequencing data analysis for Barcoding. Available at: https://github.com/PepBau/SangerSeq. (Accessed: 4th June 2019)
60. PepBau/Barcode_Align: Barcode DNA sequencies alignment and phylogenetic tree drawing. Available at: https://github.com/PepBau/Barcode_Align. (Accessed: 4th June 2019)

ANNEX

Taula 12 *Taula de les mostres estudiades, inclou un ID únic per cada mostra, l'espècie, i la data de captura.*

ID Mostra	Espècie	Data de Captura	ID Mostra	Espècie	Data de Captura
MSE_01	E. depressa	1/6/2018	MSE_38	E. sororcula	13/6/2015
MSE_02	E. caniola	8/6/2018	MSE_39	E. lutarella luqueti	29/7/2015
MSE_03	E. interpositella	11/6/2018	MSE_40	E. pseudocomplana	2/8/2004
MSE_04	E. interpositella	11/6/2018	MSE_41	E. muscula	12/6/2002
MSE_05	E. rungsi	11/6/2018	MSE_42	E. pygmaeola pygmaeola	juny-04
MSE_06	E. interpositella	11/6/2018	MSE_43	E. marcida	12/8/2007
MSE_08	E. sororcula	26/9/2018	MSE_44	E. albicosta albicosta	21/12/2003
MSE_09	E. depressa	13/9/2015	MSE_45	E. lutarella luqueti	29/7/2009
MSE_10	E. interpositella	16/5/2015	MSE_46	E. caniola	11/6/2018
MSE_11	E. rungsi	22/10/2014	MSE_47	E. costalis	3/9/2003
MSE_12	E. complana	9/8/2016	MSE_48	E. albicosta albicosta	2/2/2004
MSE_13	E. palliatella	11/8/2018	MSE_49	E. complana iberica	27/7/2017
MSE_14	E. uniola	11/8/2018	MSE_50	E. bipuncta	6/2/2018
MSE_15	E. predotae	4/9/2013	MSE_51	E. caniola gibrati	2/6/2000
MSE_16	E. pygmaeola pallifrons	14/7/2018	MSE_52	E. caniola	18/8/2017
MSE_17	E. predotae	4/9/2013	MSE_53	E. caniola torstenii	6/10/2015
MSE_18	E. complana	25/7/2017	MSE_54	E. caniola torstenii	5/9/2018
MSE_19	E. caniola	14/6/2017	MSE_55	E. complana iberica	18/7/2018
MSE_20	E. griseola	24/8/2016	MSE_56	E. muscula	28/8/2003
MSE_21	E. lurideola	6/7/2017	MSE_57	E. pseudocomplana	1/8/1994
MSE_22	E. albicosta witti	1/6/2015	MSE_58	E. albicosta witti	des-18
MSE_23	E. marcida	11/6/2008	MSE_59	E. griseola	24/8/2016
MSE_24	E. caniola torstenii	4/9/2013	MSE_60	Lithosia quadra	17/9/2018
MSE_25	E. lutarella luqueti	25/7/2017	MSE_61	E. lurideola	25/7/2017
MSE_26	E. pseudocomplana	14/6/2017	MSE_62	E. palliatella	11/8/2018
MSE_27	E. muscula	11/6/2008	MSE_63	E. pygmaeola pallifrons	11/8/2018
MSE_28	E. pygmaeola pygmaeola	ag-04	MSE_64	E. rungsi	13/6/2018
MSE_29	E. bipuncta	5/10/2013	MSE_65	E. uniola	11/8/2018
MSE_30	E. caniola gibrati	20/4/2006	MSE_66	E. albicosta albicosta	16/7/2018
MSE_31	E. lutarella lutarella	30/7/2002	MSE_67	E. caniola gibrati	26/5/2017
MSE_32	E. cereola	14/7/2004	MSE_68	E. cereola	11/7/2008
MSE_33	E. pseudocomplana	12/8/2010	MSE_69	E. complana iberica	18/7/2018
MSE_34	E. muscula	21/10/2011	MSE_70	E. complana iberica	11/8/2018
MSE_35	E. costalis	28/8/2003	MSE_71	E. lutarella lutarella	28/7/2011
MSE_36	Lithosia quadra	25/7/2017	MSE_72	Lithosia quadra	17/9/2018
MSE_37	E. costalis	1/8/2004	MSE_73	Lithosia quadra	17/9/2018

Taula 13 *Taula de les extraccions, inclou el codi de la mostra, informació de l'espècie i la data de l'extracció.*

Codi Mostra	Espècie	Data Captura	Lloc Captura	Recolector	Barcode_ID	Data d'extracció
MSE-01	Eilema depressa Mostra nº11	1-juny-18	Els Massets, Pobla de Claramunt, Anoia.	Ramon Macià Vilà	UVic_MSE_01	22/1/2019
MSE-02	Eilema caniola Mostra nº14	8-juny-18	Miracle El Riner, Solsonès	Ramon Macià Vilà	UVic_MSE_02	28/1/2019
MSE-03	Eilema interposita Mostra nº5	11-juny-18	Barranco de Mazarra, Baza (Granada)	Ramon Macià Vilà	UVic_MSE_03	4/2/2019
MSE-04	Eilema interposita Mostra nº6	11-juny-18	Barranco de Mazarra, Baza (Granada)	Ramon Macià Vilà	UVic_MSE_04	4/2/2019
MSE-05	Eilema rungsi Mostra nº7	11-juny-18	Barranco de Mazarra, Baza (Granada)	Ramon Macià Vilà	UVic_MSE_05	4/2/2019
MSE-07	Eilema interposita Mostra nº6	11-juny-18	Barranco de Mazarra, Baza (Granada)	Ramon Macià Vilà	UVic_MSE_07	11/2/2019
MSE-08	Eilema sororcula (seca)	26/9/2018	Santa Perpètua, Gurb.	Ramon Macià Vilà	UVic_MSE_08	11/2/2019
MSE-09	Eilema depressa (seca)	13/9/2015	Plana de Vic	Ramon Macià Vilà	UVic_MSE_09	11/2/2019
MSE-10	Eilema interpositella (seca)	16/5/2015	Tabernas (Almeria)	Ramon Macià Vilà	UVic_MSE_10	18/2/2019
MSE-11	Eilema rungsi (seca)	22/10/2014	Delta del Llobregat	Ramon Macià Vilà	UVic_MSE_11	4/3/2019
MSE-12	Eilema complana (seca)	9/8/2016	Paridera del carmen Albarracín (Teruel)	Ramon Macià Vilà	UVic_MSE_12	4/3/2019
MSE-13	Eilema palliatella (seca)	11/8/2018	Valle de valdevecar, Albarracín (Teruel)	Ramon Macià Vilà	UVic_MSE_13	4/3/2019
MSE-14	Eilema uniola (seca)	11/8/2018	Valle de valdevecar, Albarracín (Teruel)	Ramon Macià Vilà	UVic_MSE_14	4/3/2019
MSE-15	Eilema predotae (seca)	4/9/2013	Puntal del ahorcado camino rural de Moscardón a Royuela	Ramon Macià Vilà	UVic_MSE_15	4/3/2019
MSE-16	Eilema pygmaeola (seca)	14/7/2018	Camí de Moscardón a El Valleoillo, La Peguera El Valleolilo	Ramon Macià Vilà	UVic_MSE_16	4/3/2019
MSE-17	Eilema predotae (seca)	4/9/2013	Puntal del ahorcado camino rural de Moscardón a Royuela	Ramon Macià Vilà	UVic_MSE_17	25/3/2019
MSE-18	Eilema complana (seca)	25/7/2017	Bósc de Tredós Val d'Aran	Ramon Macià Vilà	UVic_MSE_18	11/3/2019
MSE-19	Eilema caniola (seca)	14/6/2017	El Cerro Gredilla la Polera	Ramon Macià Vilà	UVic_MSE_19	11/3/2019
MSE-20	Eilema griseola (seca)	24/8/2016	Camí de Vilavella, Vidrà, Osona.	Ramon Macià Vilà	UVic_MSE_20	11/3/2019
MSE-21	Eilema lurideola (seca)	6/7/2017	Bósc Nere, Banhs de Tredós. Val d'Aran.	Ramon Macià Vilà	UVic_MSE_21	11/3/2019
MSE-22	Eilema Albicota Witti (seca)	1/6/2015	Gran Canaria, Canarias	Ramon Macià Vilà	UVic_MSE_22	11/3/2019
MSE-23	Eilema marcida (seca)	11/6/2008	Porto Cristo, Manacor, (Mallorca)	Ramon Macià Vilà	UVic_MSE_23	11/3/2019
MSE-24C	Eilema caniola torstenii (seca)	4/9/2013	Puntal del ahorcado, Moscardón, Teruel.	Ramon Macià Vilà	UVic_MSE_24	25/3/2019
MSE-25C	Eilema luterella luqueti (seca)	25/7/2017	Bósc de Tredòs, Val d'Aran.	Ramon Macià Vilà	UVic_MSE_25	25/3/2019
MSE-26C	Eilema pseudocomplana (seca)	14/6/2017	El Cerro, Gredilla, La Polera, Burgos.	Ramon Macià Vilà	UVic_MSE_31	25/3/2019
MSE-27C	Eilema muscula (seca)	11/6/2008	Porto Cristo, Manacor.	Ramon Macià Vilà	UVic_MSE_32	25/3/2019
MSE-28	Eilema pygmaeola (seca)	ag-04	Kiskunfelegyháza, Hungary	Ramon Macià Vilà	UVic_MSE_28	18/3/2019
MSE-29	Eilema bipuncta	5/10/2013	Kenitra, Marroc	Ramon Macià Vilà	UVic_MSE_29	18/3/2019
MSE-30	Eilema caniola gibrati	20/4/2006	Rambla Granatilla, Soplamo (Almeria)	Ramon Macià Vilà	UVic_MSE_30	18/3/2019
MSE-31	Eilema luterella	30/7/2002	Nove Zámky, Slovakia	Ramon Macià Vilà	UVic_MSE_31	18/3/2019
MSE-32	Eilema cereola	14/7/2004	Ailefroide, France 05	Ramon Macià Vilà	UVic_MSE_32	18/3/2019
MSE-33	Eilema pseudocomplana	12/8/2010	Remolon, France 05	Ramon Macià Vilà	UVic_MSE_33	18/3/2019
MSE-34	Eilema muscula	19-21/10/2011	Antalya, Kemer, Turkey	Ramon Macià Vilà	UVic_MSE_34	18/3/2019
MSE-35	Eilema costalis	28/8/2003	Agyia, Zypem	Ramon Macià Vilà	UVIC_MSE_35	18/3/2019
MSE-36	Lithosia quadra	25/7/2017	Banhs de Tredós, Val d'Aran	Ramon Macià Vilà	UVIC_MSE_36	1/4/2019
MSE-37	Eilema costalis (seca)	1/8/2004	Kiskunfelegyháza, Hongria.	Ramon Macià Vilà	UVIC_MSE_37	25/3/2019

Taula 14 *Taula de les extraccions, inclou el codi de la mostra, informació de l'espècie i la data de l'extracció.*

Codi Mostra	Espècie	Data Captura	Lloc Captura	Recolector	Barcode_ID	Data d'extracció
MSE-38	Eilema sororcula (seca)	13/6/2015	Ermita de Sta. Margarita de Villatella (Barna).	Ramon Macià Vilà	UVIC_MSE_38	18/3/2019
MSE-39	Eilema luterella luqueti (seca)	29/7/2015	Banhs de Tredòs, Val d'Aran.	Ramon Macià Vilà	UVIC_MSE_39	18/3/2019
MSE-40C	Eilema pseudocomplana (seca) 2 potes	2/8/2004	Ampus, France.	Ramon Macià Vilà	UVic_MSE_40	1/4/2019
MSE-41C	Eilema muscula (seca) 2 potes	12/6/2002	Kidasi, West Zypern	Ramon Macià Vilà	UVic_MSE_41	1/4/2019
MSE-42K	Eilema pygmaeola pygmaeola (seca)	juny-04	Kiskunflekegyháza, Hungria.	Ramon Macià Vilà	UVic_MSE_42	1/4/2019
MSE-43K	Eilema marcida (seca)	12/8/2007	Villalazán, Zamora.	Ramon Macià Vilà	UVic_MSE_43	1/4/2019
MSE-44K	Eilema albicosta albicosta (seca)	21/12/2003	Los Canarios, La Palma.	Ramon Macià Vilà	UVic_MSE_44	1/4/2019
MSE-45K	Luterella luqueti (seca)	29/7/2009	Moscardón, Pino Gordo, Teruel.	Ramon Macià Vilà	UVic_MSE_45	1/4/2019
MSE-46K	Eilema caniola (fresca) 2 potes	11/6/2018	Miracle El Riner, Solsonès	Ramon Macià Vilà	UVic_MSE_46	4/4/2019
MSE-47K	Eilema costalis (seca) abdomen	3/9/2003	Adelfoi Forest, Zypern.	Ramon Macià Vilà	UVic_MSE_47	4/4/2019
MSE-48K	Albicosta albicosta	2/2/2004	Los Canarios, La Palma.	Ramon Macià Vilà	UVic_MSE_48	4/4/2019
MSE-49K	Complana iberica	27/7/2017	Camí de Moscardón a El Vallecillo, La Peguera. El Vallecillo (Teruel).	Ramon Macià Vilà	UVic_MSE_49	4/4/2019
MSE-50K	Eilema bipuncta potes	6/2/2018	Palacio Doñana, Huelva.	Ramon Macià Vilà	UVic_MSE_50	24/4/2019
MSE-51K	Eilema caniola gibrati potes	2/6/2000	Maison Forestier, Marroc.	Ramon Macià Vilà	UVic_MSE_51	24/4/2019
MSE-52C	Eilema caniola (seca) 2 potes	18/8/2017	Villanueva de Sigena, Huesca	Ramon Macià Vilà	UVic_MSE_52	1/4/2019
MSE-53C	Eilema caniola torstenii (seca) 2 potes	6/10/2015	Puig de s'Aritjar. Serra d'Alfàbia Bunyola. (Mallorca)	Ramon Macià Vilà	UVic_MSE_53	1/4/2019
MSE-54K	Eilema caniola torstenii, abdomen	5/9/2018	Es Calò, Formentera.	Ramon Macià Vilà	UVic_MSE_54	29/4/2019
MSE-55K	Eilema complana iberica, abdomen	18/7/2018	Camí de Moscardón. El Vallecillo. La Peguera. El Vallecillo (Teruel).	Ramon Macià Vilà	UVic_MSE_55	29/4/2019
MSE-56K	Eilema muscula, abdomen	28/8/2003	Zypern, nächst Agyia-Picnic Forststr. 400m Pafos-Forest leg. Hentscholek.	Ramon Macià Vilà	UVic_MSE_56	29/4/2019
MSE-57K	Eilema pseudocomplana, abdomen	1/8/1994	Sant Jaume Jou (Lleida).	Ramon Macià Vilà	UVic_MSE_57	29/4/2019
MSE-58K	Eilema albicosta witti, abdomen	des-18	Montaña de la Horca, Las Aguilas, Tenerife.	Ramon Macià Vilà	UVic_MSE_58	29/4/2019
MSE-59K	Eilema griseola, abdomen	24/8/2016	Camí de Vilavella, Vidrà, Osona.	Ramon Macià Vilà	UVic_MSE_59	29/4/2019
MSE-60K	Lithosia quadra (2 potes)	17/9/2018	Vic, centre ciutat. Osona.	Ramon Macià Vilà	UVic_MSE_60	29/4/2019
MSE-61K	Eilema lurideola Abdomen	25/7/2017	Bósc Nere, Banhs de Tredós. Val d'Aran.	Ramon Macià Vilà	UVic_MSE_61	29/4/2019
MSE-62K	Eilema palliatella Abdomen	11/8/2018	Valle de Valdevecar, Albarracín (Teruel).	Ramon Macià Vilà	UVic_MSE_62	29/4/2019
MSE-63K	Eilema pygmaeola pallifrons, abdomen	11/8/2018	Valle de Valdevecar, Albarracín (Teruel).	Ramon Macià Vilà	UVic_MSE_63	29/4/2019
MSE-64K	Eilema rungsi, abdomen	13/6/2018	Punta Entinas, El Sabinar, El Ejido. (Almería)	Ramon Macià Vilà	UVic_MSE_64	29/4/2019
MSE-65K	Eilema Uniola, abdomen	11/8/2018	Valle de Valdevecar, Albarracín (Teruel).	Ramon Macià Vilà	UVic_MSE_65	29/4/2019
MSE-66K	Eilema albicosta albicosta (potes)	16/7/2018	Hermigua, La Gomera (Canarias).	Ramon Macià Vilà	UVic_MSE_66	29/4/2019
MSE-67K	Eilema caniola gibrati ? (potes)	26/5/2017	2km Est de Midar, Driouch, Marroc.	Ramon Macià Vilà	UVic_MSE_67	29/4/2019
MSE-68K	Eilema cereola (abdomen)	11/7/2008	France, H. A. Ailefroide	Ramon Macià Vilà	UVic_MSE_66	6/5/2019
MSE-69K	Eilema complana iberica (abdomen)	18/7/2018	Camí de Moscardón a El Vallecillo, La Peguera. El Vallecillo (Teruel).	Ramon Macià Vilà	UVic_MSE_67	6/5/2019
MSE-70K	Eilema complana iberica (abdomen)	11/8/2018	Valle de Valdevecar, Albarracín (Teruel).	Ramon Macià Vilà	UVic_MSE_54	6/5/2019
MSE-71K	Eilema lutarella lutarella, abdomen	28/7/2011	Janikow, na swiatio	Ramon Macià Vilà	UVic_MSE_55	6/5/2019
MSE-72K	Lithosia quadra (3 potes)	17/9/2018	Vic, Centre Ciutat (Osona)	Ramon Macià Vilà	UVic_MSE_56	6/5/2019
MSE-73K	Lithosia quadra (3 potes)	17/9/2018	Vic, Centre Ciutat (Osona)	Ramon Macià Vilà	UVic_MSE_57	6/5/2019

Figura 33 *Mostres de la col·lecció analitzades (ID en vermell).*

Taula 15 *Resultats de l'amplificació per PCR de cada mostra, cada gen, amb chelex i kit comercial. Correcte/Incorrecte*

	Chelex				Kit Comercial					
ID Mostra	CO1	DDX23	ArgK	Wingless	CO1	DDX23	ArgK	Wingless	SSU72	CAD
MSE_01	1|3									
MSE_02	1|2									
MSE_03	1|1	1|1	1|1	0|3	1|1					
MSE_04	0|2	0|1	0|1							
MSE_05	1|1	1|1	1|1							
MSE_06					1|1	1|1	1|1			
MSE_08	1|2		1|1		2|3	1|1	0|1	0|3	1|1	0|1
MSE_09	1|2				3|4	1|1	0|1	0|10	2|2	
MSE_10	0|1				2|4				1|1	0|1
MSE_11	2|2	0|1								
MSE_12	2|2	0|1								
MSE_13	2|2	0|1								
MSE_14	2|2	0|1								
MSE_15	0|3	0|1			0|1				1|1	
MSE_16	2|2	0|1								
MSE_17	1|1									
MSE_18	0|1				1|1				1|1	
MSE_19	0|1				0|1					
MSE_20	1|1									
MSE_21	1|1									
MSE_22	1|1									
MSE_23	0|1				0|1				1|1	
MSE_24	0|2			0|1						
MSE_25	0|1			0|1						
MSE_26	0|2									
MSE_27	0|2									
MSE_28					0|3	0|1	0|3		2|2	
MSE_29					2|4	0|1	0|2		2|2	0|1
MSE_30					1|3	0|1	0|3		2|2	
MSE_31	0|1	0|1	0|1		0|3		0|2		2|2	
MSE_32	0|1	0|1	0|1		0|2		0|2		2|2	
MSE_33	0|1	0|1	0|1							
MSE_34	0|1	0|1	0|1							
MSE_35	0|1	0|1	0|1							
MSE_36	0|1	0|1	0|1		2|2					
MSE_37	1|2									

Taula 16 *Resultats de l'amplificació per PCR de cada mostra, cada gen, amb chelex i kit comercial. Correcte/Incorrecte*

	Chelex						Kit Comercial		
ID Mostra	CO1	DDX23	ArgK	Psb	SSU72	ProSup	CO1	ArgK	SSU72
MSE_38	0|1		1|1						
MSE_39	0|1						1|2	0|1	1|1
MSE_40		0|1	1|1	0|1	1|1	0|1			
MSE_41		0|1	1|1	0|1	1|1	0|1			
MSE_42							0|3		1|1
MSE_43							1|1		1|1
MSE_44							0|1		
MSE_45							1|1		1|1
MSE_46							1|1		
MSE_47							0|2	0|1	1|1
MSE_48							0|2	0|1	1|1
MSE_49							0|2	0|1	1|1
MSE_50							0|1	0|1	1|1
MSE_51							0|1	0|1	1|1
MSE_52		0|1		1|1	1|1	0|1			
MSE_53		0|1		0|1	1|1	0|1			
MSE_54									1|1
MSE_55									1|1
MSE_56							1|1		1|1
MSE_57							1|1		1|1
MSE_58									1|1
MSE_59							1|1		1|1
MSE_60									1|1
MSE_61									1|1
MSE_62									1|1
MSE_63									1|1
MSE_64									1|1
MSE_65									1|1
MSE_66							1|1		1|1
MSE_67							1|1		1|1
MSE_68							1|2		
MSE_69							2|2		
MSE_70							2|2		
MSE_71							2|2		
MSE_72							1|2		
MSE_73							1|2		

Script d'R-Markdown per obtenir la seqüència consens

```
---
title: "SangerSeq"
Comments: After Walkthrough for using the sangerseqR package
Description: Tools for Sanger Sequencing Data in R.
https://bioconductor.org/packages/release/bioc/html/sangerseqR.html
output:
  word_document: default
  html_notebook: default
  pdf_document: default
  html_document:
    df_print: paged
Author: Pep Bau and Marc Solé
---


###Global Options:
1. Clears Memory
2. Defines File Paths
3. Sets Global Options for all code chunks
4. Defines Global Variables
```{r Set Global Options, warning=FALSE}
#Clear Memory
rm(list=ls(all=TRUE))

#Define Global Variables Here:
primer_file<-"Primers_Tots.fa"

#Sets File Path
datapath<-"./SampleData/"

#Sets Global Options
knitr::opts_chunk$set(root.dir=datapath, echo=FALSE, warning=FALSE,
message=FALSE)

```


###Prepares execution:
1. Installs Packages
```{r Prepare Execution}
#sangerseqR only runs under R version 3.5 or higher
#Installing/loading the latest installr package from GUI (not from
RStudio):
#install.packages("installr")
#library(installr)

#Install Bioconductor
#BiocManager::install("BiocInstaller")

#Install Packages
packages <- c("rmarkdown","seqinr", "Biostrings","sangerseqR",
"ape","phangorn","stringi","stringr","DECIPHER")
if (length(setdiff(packages, rownames(installed.packages()))) > 0) {
  install.packages(setdiff(packages, rownames(installed.packages())))
}

if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
```
```

```
BiocManager::install("DECIPHER")


```

2. Loads Required Libraries
```{r}
#loads libraries
library(Biostrings)
library(seqinr)
library(sangerseqR)

library(ape)
library(phangorn)
library (stringi)
library (stringr)
library(DECIPHER)

#Additional functions source
  source("funcions.r")

#Tutorial:
#browseVignettes("sangerseqR")
```
###File load:
1. Loads files from Data Folder
2. Loads Primer sequences from multi-fasta file
```{r}
#Reads files to proces from file (traces_to_process.csv)
files2pr<-read.csv2(paste0(datapath,"traces_to_process.csv"))

#Loads primer sequences from multi-fasta
primers<-readDNAStringSet(paste0(datapath, primer_file), format="fasta",
nrec=-1L, skip=0L, seek.first.rec=FALSE, use.names=TRUE)

#loads Forward and Reverse Sequences into lists
fwd.abif.ls<-list()
rev.abif.ls<-list()
for (row in nrow(files2pr)) {
  print (paste("Loading Sample",row))
  fwd.abif.ls[[row]]<-sangerseqR::read.abif(paste0(datapath,
files2pr[row,"Fwd"]))
  rev.abif.ls[[row]]<-sangerseqR::read.abif(paste0(datapath,
files2pr[row,"Rev"]))
}

```

###Trim Low Quality:
```{r}
trim.cutoff<-0.0001

#Sangerseq Format lists
fwd.sangerseq.ls<-list()
rev.sangerseq.ls<-list()

#BaseCall lists
fwd.calls.ls<-list()
rev.calls.ls<-list()
```

```
#Final trimmed sequences
fwd.final<-list()
rev.final<-list()

for (i in nrow(files2pr)){

  # #Calculates trimming positions
   fwd.tr<-trim.mott(fwd.abif.ls[[i]], cutoff = trim.cutoff)
   rev.tr<-trim.mott(rev.abif.ls[[i]], cutoff = trim.cutoff)

  #converts abif data to sangerseq format
  fwd.sangerseq.ls[[i]]<-sangerseq(fwd.abif.ls[[i]])
  rev.sangerseq.ls[[i]]<-sangerseq(rev.abif.ls[[i]])

  #Base Calls (only different if ratio (secondary/primary) is above 0.33:
  fwd.calls.ls[[i]]<-makeBaseCalls(fwd.sangerseq.ls[[i]],ratio=0.33)
  rev.calls.ls[[i]]<-makeBaseCalls(rev.sangerseq.ls[[i]],ratio=0.33)

   tr5f<- fwd.tr$start
   seqlenf<-length(primarySeq(fwd.calls.ls[[i]])) #non-trimmed seq length
   trlenf<-fwd.tr$finish-fwd.tr$start #trimmed seq length
   tr3f<-tr5f+trlenf
   if ((trlenf+tr5f)>seqlenf) tr3f<-seqlenf
   fwd.final[[i]]<-primarySeq(fwd.calls.ls[[i]])[tr5f:tr3f]

  #Plots Chromatogram showing only primary calls (highest peak)
  chromatogram(fwd.calls.ls[[i]],
showcalls="primary",width=200,height=4,trim5 =tr5f,trim3=seqlenf-
tr3f,showtrim = TRUE)

   tr5r<- rev.tr$start
   seqlenr<-length(primarySeq(rev.calls.ls[[i]])) #non-trimmed seq length
   trlenr<-rev.tr$finish-rev.tr$start #trimmed seq length
   tr3r<-tr5r+trlenr
   if ((trlenr+tr5r)>seqlenr) tr3r<-seqlenr
   rev.final[[i]]<-primarySeq(rev.calls.ls[[i]])[tr5r:tr3r]

  #Plots Chromatogram showing only primary calls (highest peak)

png(file=paste0(datapath,paste(files2pr$ID[i],files2pr$Specie[i],files2pr
$Gene[i],sep="-"," chromatogram.png")),width=3500,height=1500,res=250)
  chromatogram(rev.calls.ls[[i]],
showcalls="both",width=180,height=2,trim5 =tr5r,trim3=seqlenr-
tr3r,showtrim = TRUE)
  dev.off()

}


```

###Forward and Reverse Consensus:
1. pairwise alignemnt of both calls:
```{r}

consens.ls<-list()

for (i in nrow(files2pr)){
```

```
  #Reverse complement of rev sequence
  rev.final.RC<-reverseComplement(DNAString(rev.final[[i]]))

  #Alignment of Forward and Reverse Sequences
  fr.pa<-pairwiseAlignment(fwd.final[[i]],rev.final.RC,type="global-
local")
  summary(fr.pa)

  writePairwiseAlignments(fr.pa)

  consens.ls[[i]]<-consensusString(fr.pa)
  consens.ls[[i]]

}

write.fasta(consens.ls,file.out =
"Consens.fa",names=paste(files2pr$ID,files2pr$Specie,files2pr$Gene,sep="-
"))
```

```
###Remove Primers
```{r}

for (i in 1:nrow(files2pr)){

  #gets primer sequences
  fwd_primer<-primers[files2pr$F_Primer[i]]
  fwd_primer.rc<-reverseComplement(fwd_primer)
  rev_primer<-primers[files2pr$R_Primer[i]]
  rev_primer.rc<-reverseComplement(rev_primer)


  #aligns Forward primer
  fwdprimer.pa<-
pairwiseAlignment(primarySeq(fwd.sangerseq.ls[[i]]),rev_primer.rc,type="g
lobal-local")
  writePairwiseAlignments(fwdprimer.pa)
  fwd.init<-start(pattern(fwdprimer.pa)) #position where rev primer ends
  fwd.init

  #aligns Reverse primer
  revprimer.pa<-
pairwiseAlignment(primarySeq(fwd.sangerseq.ls[[i]]),rev_primer.rc,type="g
lobal-local")

  writePairwiseAlignments(revprimer.pa)
  rev.init<-start(pattern(revprimer.pa)) #position where rev primer ends

}
```
```

Script d'R-Markdown per obtenir un arbre a partir de les seqüències consens

```
---
Author: Josep Bau Macia i Marc Solé
Afiliation: Biosciences Department, University of Vic-Central University
of Catalonia
Comments: null
```

```
Create date: January 21st, 2019
Purpose: Align DNA sequenciation data and draw phylogenetic tree
Revised: null
output:

  pdf_document: default
  html_document: default
---

##BARCODE_ALIGN
###Aligns DNA sequenciation data and draws phylogenetic tree
Josep Bau Macia
Biosciences Department - University of Vic-Central University of
Catalonia

###Global Options:
1. Clears Memory
2. Defines File Paths
3. Sets Global Options for all code chunks
4. Defines Global Variables

```{r Set Global Options, warning=FALSE}
#Clear Memory
rm(list=ls(all=TRUE))

#Define Global Variables Here:
SeqFile<-"Tots COIs meus i mbt per alinear.fasta"  #Name of the File that
contains fasta sequences
TaxID<-"Eilema"                #Taxonomic Identifier for the current
analysis



#Sets File Path
inPath<-paste0("~/Barcode_Align/Data/",TaxID)
setwd("C:\\Users\\Marc\\Documents\\Barcode_Align")

#Sets Global Options
knitr::opts_chunk$set(root.dir=inPath, echo=FALSE, warning=FALSE,
message=FALSE)



```


###Prepares execution:
1. Installs Packages
2. Loads Required Libraries

```{r Prepare Execution}
#Install Packages
packages <- c("rmarkdown","synchrony","ape")
if (length(setdiff(packages, rownames(installed.packages()))) > 0) {
  install.packages(setdiff(packages, rownames(installed.packages())))
}

# If msa is not installed through install.packages, then run the
following two lines:
```
```

```
source("http://www.bioconductor.org/biocLite.R")
# biocLite("msa")
biocLite("seqinr")
biocLite("msa")
#Load Libraries
library(plyr)    #needed for the 'count' function
library(msa)     #Multiple Sequence Alignment package (msa)
library(ape)     #Analyses of Phylogenetics and Evolution (ape)
library(seqinr)
library(phytools)
```

```

###File load:
1. Loads files from Data Folder

```{r}
#retrieves fasta file names (if necessary)
#datFiles <- list.files(path=inPath,pattern="*.fa")

#loads fasta from SeqFile
fa<-readDNAStringSet(paste0(inPath,"/", SeqFile), format="fasta", nrec=-
1L, skip=0L, seek.first.rec=FALSE, use.names=TRUE)

```

###Sequence Alignment
1. Aligns with msa
2. Converts alignment to seqinr format

```{r}
nseq<-length(fa)
myAlignment<-msa(fa)
#print(myAlignment, showNames=FALSE, show="complete")
myConv_Alignment<-msaConvert(myAlignment, type="seqinr::alignment")
str(myConv_Alignment)

```

###Distance Calculation
1. Calculates Genetic Distance Matrix
2. Draws phylogenetic tree
```{r}
Treat<-"quelcom"
d <- dist.alignment(myConv_Alignment, "identity")
save(d,file=paste0(Treat,"Genetic Distances",nseq,"_seqs"))
#load(file=paste0(Treat,"Genetic Distances",nseq,"_seqs"))
#str(d)

myTree <- njs(d)

#install.packages("devtools")
#devtools::install_github("USCBiostats/rphyloxml")
#library(ape)
```
```

```
#library(rphyloxml)

#z <- write_phyloxml(myTree)
#library(XML)
#saveXML(z)


save(myTree,file=paste0(Treat,"Tree",nseq,"_seqs"))
#load(file=paste0(Treat,"Tree",nseq,"_seqs"))
#str(myTree)
png(file=paste0(inPath," ClustalOmega
COI.png"),width=1600,height=1800,res=80)

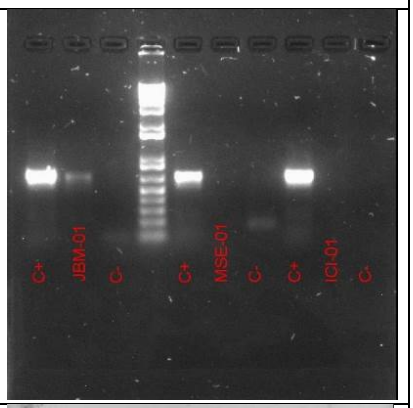
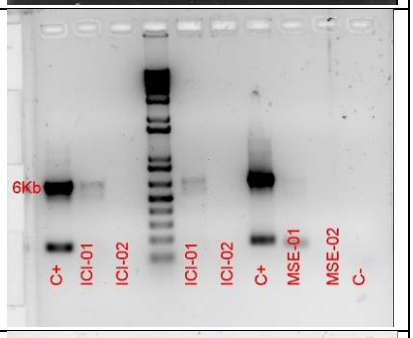
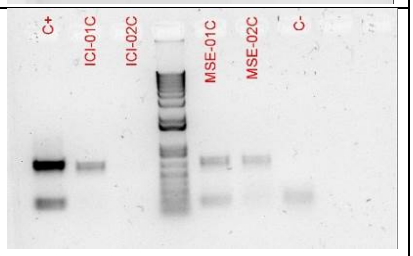
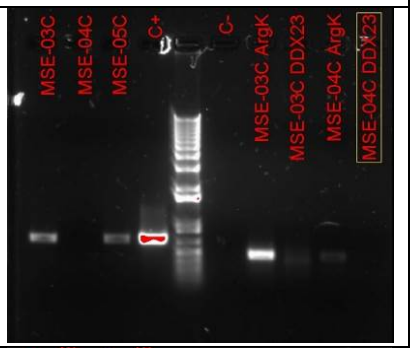
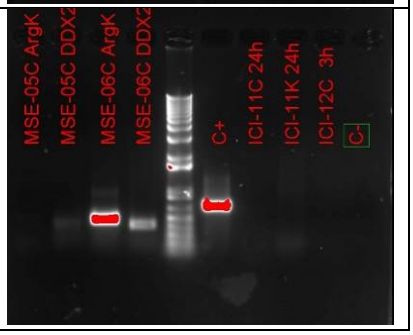


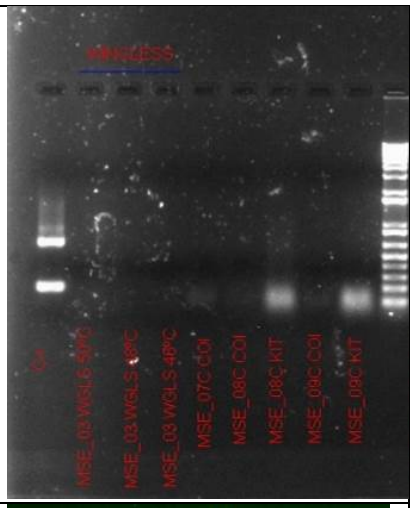
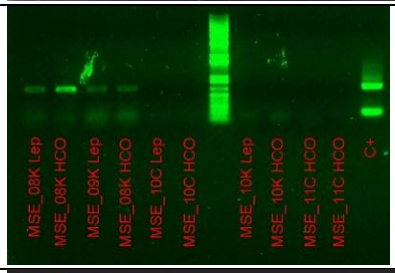
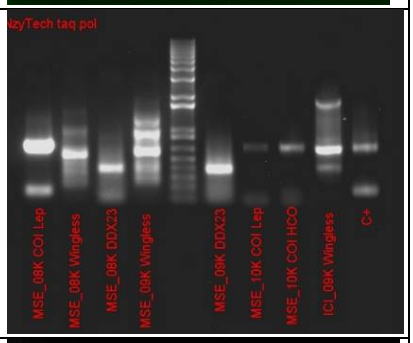
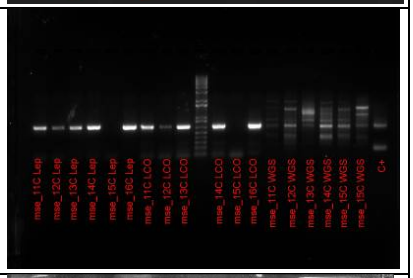
plot(myTree, main=paste0(TaxID,"Phylogenetic Tree of Barcode Sequences"))

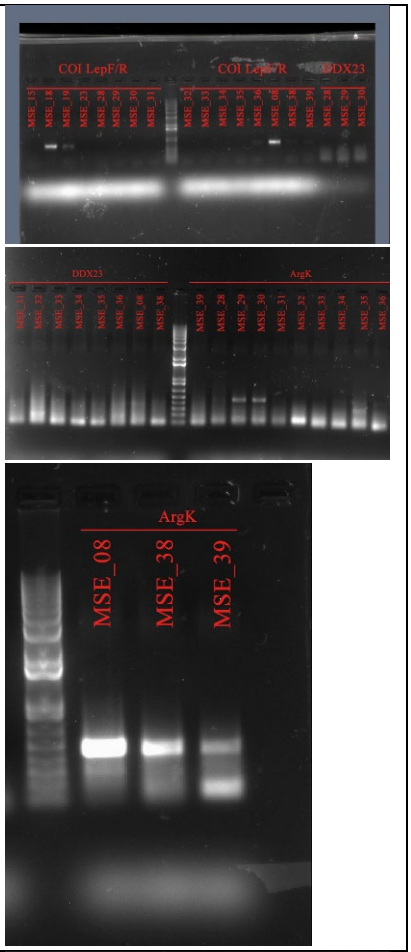
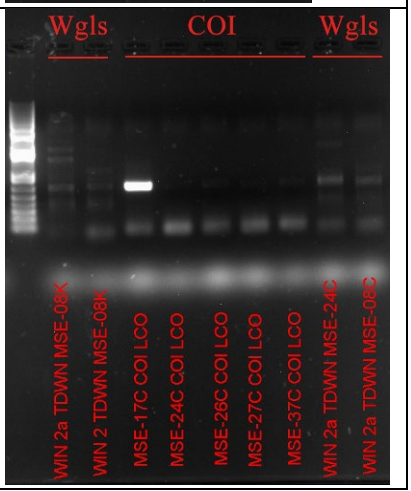
add.scale.bar(length=0.05)
dev.off()
```

```
###Pretty Alignment
1. Creates a .tex file of our alignment which can be viewed using
the software texworks
There is a faster way to do that converting .tex to .pdf with R using
pdflatex,
but it doesn't work.

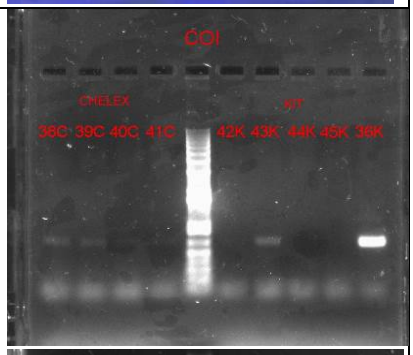
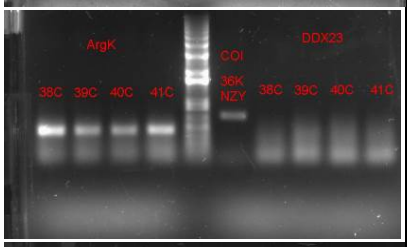
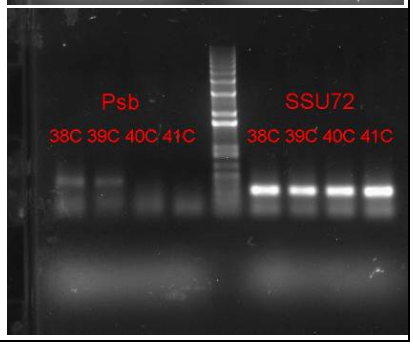
```{r}
msaPrettyPrint(myAlignment, output="tex",
showNames="none",showLogo="none", askForOverwrite=FALSE, verbose=FALSE)
```
```

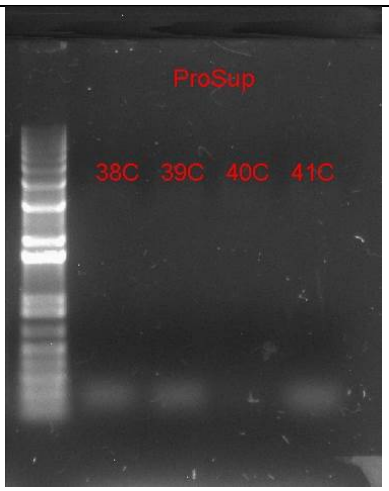
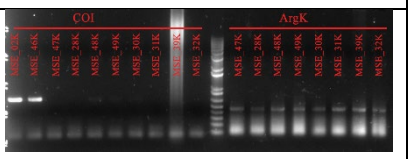
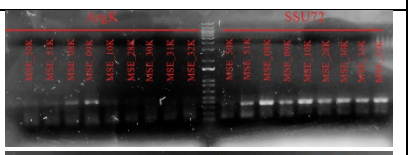
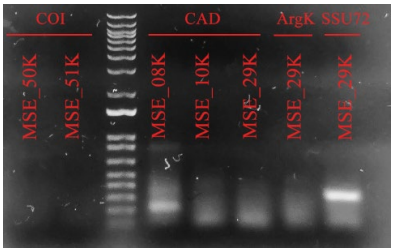
Taula 17 *Imatges dels gels d'agarosa i dels resultats de la PCR, inclou data de l'extracció i informació sobre la PCR.*

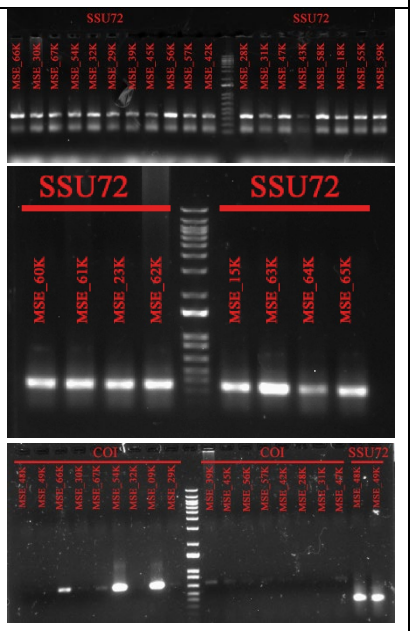
Mostra	Extracció	PCR	Electro	GEL
MSE_01C	22/01/19 1 pota	23/01/19 COI Resultats anòmals degut a PCR incorrecta	23/01/19	
MSE_01C	28/01/19 1 pota	29/01/19 COI Resultats anòmals degut a PCR incorrecta	29/01/19	
MSE_02C	28/01/19 1 pota			
MSE_01C	31/01/19 Mateixes mostre	31/01/19 COI Augmentem a 4µL la quantitat de mostra	31/01/19	
MSE_02C				
MSE_03C MSE_04C MSE_05C	04/02/19 2 potes	07/02/19 COI: LepFR ArgK DDX23	07/02/19	
MSE_05C MSE_06K	04/02/19 2 potes MSE_03=AT_0 Probablement extreta amb kit a MBT	07/02/19 ArgK DDX23		

MSE_03C MSE_07C MSE_08C MSE_08K MSE_09C MSE_09K	11/02/19 Extracció nova d la mse_04 = mse_07 amb C. Mostres seques amb C i K (pota abdomen)	14/02/19 Gradient de T per wingles COI per les seques.	14/02/19	
MSE_08K MSE_09K MSE_10C MSE_10K MSE_11C	18/02/19 MSE_10C MSE_10K MSE_11C	25/02/19 50ºC amb Preamp.	25/02/19	
MSE_08K MSE_09K MSE_10K	11/02/19 18/02/19	28/02/19 50ºC COI 53ºC DDX23 Preamp. Nova taq.	28/02/19	
MSE_11C MSE_12C MSE_13C MSE_14C MSE_15C MSE_16C	04/03/19 2 potes	06/03/19 COI: LepFR i LCO/HCO Wingless	7/03/19	
MSE_15C MSE_18C MSE_19C MSE_20C MSE_21C MSE_22C MSE_23C	11/03/19 2 potes	13/03/19 COI: LCO/HCO Wingless: Gradient 50-54ºC Primers 2a i 2	14/03/19	

MSE_15K MSE_18K MSE_19K MSE_23K MSE_28K MSE_29K MSE_30K MSE_31C MSE_32C MSE_33C MSE_34C MSE_35C MSE_36C MSE_08C MSE_38C MSE_39C	17/03/19 MSE_15K MSE_18K MSE_19K MSE_23K Kit Abdomen MSE_28K MSE_29K MSE_30K Kit 2 potes La resta 2 potes chelex	18/03/19 MSE_15K MSE_18K MSE_19K MSE_23K LepFR La resta LepFR, DDX23 i ArgK	21/03/19	
MSE_08K MSE_15C MSE_24C MSE_25C MSE_26C MSE_27C MSE_37C	25/03/19 Extracció dos potes chelex	26/03/19 Touchdown PCR per Wingless (MSE_08K, MSE_24C, MSE_26C) COI LCO per la resta de chelex	28/03/19	

MSE_08K	Mostra extreta amb kit l'11/02/	30/03/19 Banc nuclear Psb SSU72 CAD1028 IDHdag MDH Ca-ATPasa ProSup COI	1/04/19	
MSE_38C MSE_39C MSE_40C MSE_41C MSE_42K MSE_43K MSE_44K MSE_45K MSE_36K MSE_36K – NZYtech	1/04/19 Chelex i Kit 2 potes	3/04/19 Tots COI. Els chelex també: ArgK, DDX23, Psb, SSU72 i ProSup.	4/04/19	  

				
MSE_02K MSE_46K MSE_47K MSE_28K MSE_48K MSE_49K MSE_30K MSE_31K MSE_39K MSE_32K	08/04/19 Kit 2 Potes: MSE_02K MSE_46K Kit Abdomen: MSE_47K MSE_28K MSE_48K MSE_49K MSE_30K MSE_31K MSE_39K MSE_32K	09/04/19 COI: MSE_02K MSE_46K COI, ArgK: MSE_47K MSE_28K MSE_48K MSE_49K MSE_30K MSE_31K MSE_39K MSE_32K	11/04/19	
MSE_50K MSE_51K MSE_08K MSE_10K MSE_29K MSE_09K MSE_28K MSE_30K MSE_31K MSE_32K	24/04/19; Potes kit: MSE_50 MSE_51 La resta són extraccions anteriors.	25/04/19 COI, ArgK,SSU72: MSE_50K MSE_51K ArgK,SSU72, CAD: MSE_08K MSE_10K MSE_29K ArgK,SSU72 MSE_09K MSE_28K MSE_30K MSE_31K MSE_32K	26/04/19	 

MSE_66K MSE_67K MSE_54K MSE_55K MSE_56K MSE_57K MSE_58K MSE_59K MSE_60K MSE_61K MSE_62K MSE_63K MSE_64K MSE_65K MSE_66K MSE_48K MSE_49K MSE_30K MSE_32K MSE_09K MSE_29K MSE_39K MSE_45K MSE_42K MSE_28K MSE_31K MSE_47K MSE_43K MSE_18K MSE_23K MSE_15K	29/04/19 Potes: MSE_66K MSE_67K Abdomen: MSE_54K MSE_55K MSE_56K MSE_57K MSE_58K MSE_59K MSE_60K MSE_61K MSE_62K MSE_63K MSE_64K MSE_65K MSE_66K La resta són extraccions anteriors.	30/04/19 Totes SSU72, COI només: MSE_66K MSE_67K MSE_54K MSE_56K MSE_57K MSE_48K MSE_49K MSE_30K MSE_32K MSE_09K MSE_29K MSE_39K MSE_45K MSE_42K MSE_28K MSE_31K MSE_47K	02/05/19	
MSE_68K MSE_69K MSE_70K MSE_71K MSE_72K MSE_73K MSE_29K MSE_42K	06/05/19 Kit abdomen: MSE_68K MSE_69K MSE_70K MSE_71K MSE_72K MSE_73K	07/05/19 Tots COI, i COI+preamp.	09/05/19