



ПЛОВДИВСКИ УНИВЕРСИТЕТ „Паисий Хилендарски“



БИОЛОГИЧЕСКИ ФАКУЛТЕТ

КАТЕДРА ”Физиология на растенията и молекулярна биология“

Здравка Петрова Иванова

А В Т О Р Е Ф Е Р А Т

на дисертационен труд за придобиване на образователна и научна степен “доктор”

ТЕМА:

„Секвениране и анотиране на хлоропластния геном на *Haberlea rhodopensis* Friv.“

Област на висше образование: 4. Природни науки, математика
и информатика

Професионално направление: 4.3. Биологически науки

Докторска програма: Молекулярна биология

Научен ръководител:

Доц.д-р Веселин Баев

Гл.ас.д-р Евелина Даскалова

Пловдив, 2017

Дисертантът е редовен докторант към катедра „Физиология на растенията и молекулярна биология“, ПУ“Паисий Хилендарски“

Дисертационният труд е обсъден и предложен за защита на заседание на катедра „Физиология на растенията и молекулярна биология“ при Биологическия факултет на Пловдивския университет „П.Хилендарски“ проведен на 02.05.2017 г.

Защитата на дисертационния труд ще се състои на 18.09.2017г. от 10.00 часа в гр. Пловдив на открито заседание на научното жури.

УВОД

Родопският силивряк (*Haberlea rhodopensis*) е рядък вид, терциерен реликт и балкански ендемит със забележителна устойчивост към засушаване. *Haberlea rhodopensis* е хомеохлорофилно възкръсващо растение – има способността да оцелява при пълно засушаване, изпадайки в анабиоза, съхранявайки структурната цялост на фотосинтетичния си апарат, който бързо се активира при последваща рехидратация.

В настоящата дисертационна работа са представени резултатите от секвенирането, асемблирането и анотацията на хлоропластния геном на *Haberlea rhodopensis*, както и последващи сравнително-геномни анализи с хлоропластни геноми на еволюционно близки видове.

Резултатите от изследването разкриха, че хлоропластния геном на *Haberlea rhodopensis* има големина от 153,099 bp, състои се от двойка обърнати повтори (IRs) с големина 25,415 bp, и от малък и голям единични региони (SSC и LSC), с размер съответно от 17,826 bp и 84,443 bp. Геномната структура, генният порядък, GC съдържанието и използваемостта на кодоните са подобни на типичните хлоропластни геноми на покритосеменните растения. Анотирани са 137 гена, които заемат 70.66% от пластома и включват: 86 протеин кодиращи гени, 36 гена за транспортни РНК и 8 – за рибозомни РНК. Хлоропластният геном съдържа и 13 интрон-съдържащи гена, от които 8 са протеин кодиращи и 5 тРНК. Два от интрон-съдържащите гени имат два интрона, а останалите 15 гена имат само един интрон.

Сравнителният анализ на пластома с други близкородствени членове на разред *Lamiales*, разкри консервативен генен порядък в IRs и LSC/SSC регионите.

Еволюционните анализи показват, че повечето пластомни протеини на *Haberlea rhodopensis* са подложени на строги функционални ограничения (пречистващ отбор). Бяха открити два гена, кодиращи рибозомни протеини с вероятен позитивен отбор на цялата протеинова секвенция. Това може да се интерпретира като следствие от молекулярна адаптация към стреса от засушаване, осигуряваща еволюционно преимущество на *Haberlea rhodopensis*.

ЦЕЛ И ЗАДАЧИ

Целта на настоящия дисертационен труд е да се секвенира, асемблира и аотира хлоропластния геном на *Haberlea rhodopensis* и да се направят сравнителни геномни анализи с близкородствени видове.

1. Да се проведе качествен контрол и предварителна обработка на данните от масово паралелно секвениране;
2. Избор на алгоритъм и оптимизация за асемблиране на хлоропластни геноми чрез провеждане предварително асемблиране и статистическа оценка на резултатите;
3. Асемблиране и получаване на първоначална версия на хлоропластния геном на *Haberlea rhodopensis*;
4. Аноторане и установяване на геномната структура и генното съдържание на пластома;
5. Провеждане на сравнително геномни анализи:
 - Сравнителен анализ на геномната структура
 - Анализ на IR/SSC и IR/LSC регионите
 - Селекционен анализ на синонимни и несинонимни замени
 - Филогенетичен анализ

МАТЕРИАЛИ И МЕТОДИ

I. Растителен материал и данни от секвениране

1. Материал от *Haberlea rhodopensis*

Растителният материал от *Haberlea rhodopensis*, необходим за изолирането на хлоропластна ДНК, беше събран от планината Родопи (местоположение 42°1'N 24°52'E), където е естествено находище на изследваното растение. Хлоропластната ДНК беше изолирана от листата на 16 индивидуални растения. За оптималния добив на интактни хлоропласти, беше използван специален за целта кит (Chloroplast Izolation Kit – Sigma-Aldrich) с употребата на 40/80% Percoll градиент. Хлоропластната ДНК на *Haberlea rhodopensis* беше извлечена с помощта на DNeasy Plant Mini Kit (QIAGEN). Чрез микроскопски анализ (Epoch Microplate Spectrophotometer) и гел електрофореза беше извършен оглед на качеството и чистотата

2. Данни от секвениране

Изготвянето на библиотеката и секвенирането на двете проби хлоропластна ДНК от *Haberlea Rhodopensis* (HR1 и HR2) бяха извършени от BGI-Shenzhen, Китай. За всяка проба, изолираната ДНК беше използвана за генерирането на данни *paired-end*, с дължина на рийда 100bp и *library insert size* в размер на 170 bp, което е в съответствие със стандартните протоколи за секвениране, прилагани от платформата Illumina Hiseq2000.

II. Биоинформатичен софтуер

1. Софтуер за качествен анализ на данните от секвениране

1.1.FastQC

2. Софтуер за сравняване на секвенции

2.1.BWA

2.2.SAM Tools

3. Софтуер за удължаване на данните от секвениране

3.1.Flash

4. Софтуер за асемблиране

4.1.ABySS

4.2. Edena

- 4.3. SPAdes
- 5. Софтуер за генериране на скафолди**
 - 5.1.SSPACE
- 6. Софтуер за оценка на асемблирането**
 - 6.1.QUAST
- 7. Софтуер за генериране на първичния геном**
 - 7.1.ABACAS
 - 7.2.MUMmer
- 8. Софтуер за запълване на фрагментациите на първичния геном**
 - 8.1.GapFiller
- 9. Софтуер за анотация**
 - 9.1.DOGMA
 - 9.2.Geneious
- 10. Софтуер за генериране на изображения и графики**
 - 10.1.OGDRAW
 - 10.2.BRIG
 - 10.3.mVISTA
- 11. Софтуер за анализи**
 - 11.1.MISA
 - 11.2.REPuter
 - 11.3.Ka/Ks calculator

РЕЗУЛТАТИ И ДИСКУСИЯ

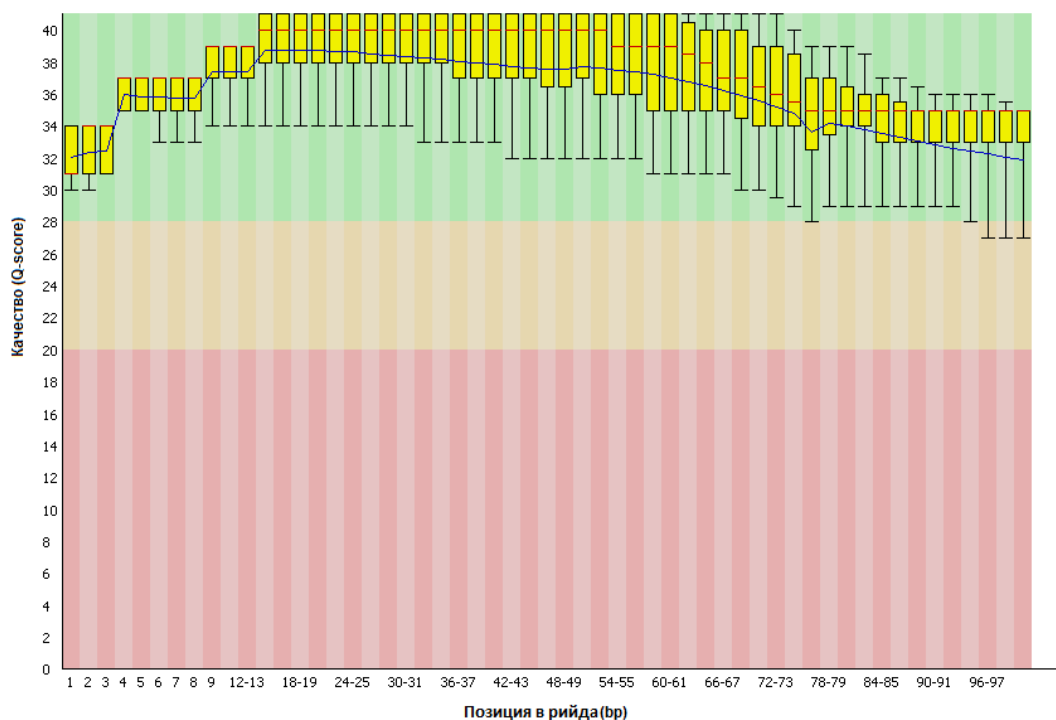
I. Секвениране на хлоропластна ДНК и предварителна обработка на NGS данните

1. Секвениране на хлоропластната ДНК

Технологията с прилагането на метода за секвениране на данните тип *paired-end* (секвениране от двата края на фрагмента ДНК) е нова и предполага повишаване на точността на асемблиране. Двете проби хлоропластна ДНК бяха секвенирани на платформа Illumina. Генерирани бяха общо 8 365 536 и 7 129 508 секвенции с 40% GC съдържание, съответно за първа и втора проба. При дължина на рийда 100bp и *library insert size* в размер на 170 bp, се получи припокриване на рийдовете от секвенирането с приблизително 30 bp. Данните бяха записани в FastQ файлов формат.

2. Анализ на качеството на суровите данни от секвениране (FastQC pre-processing)

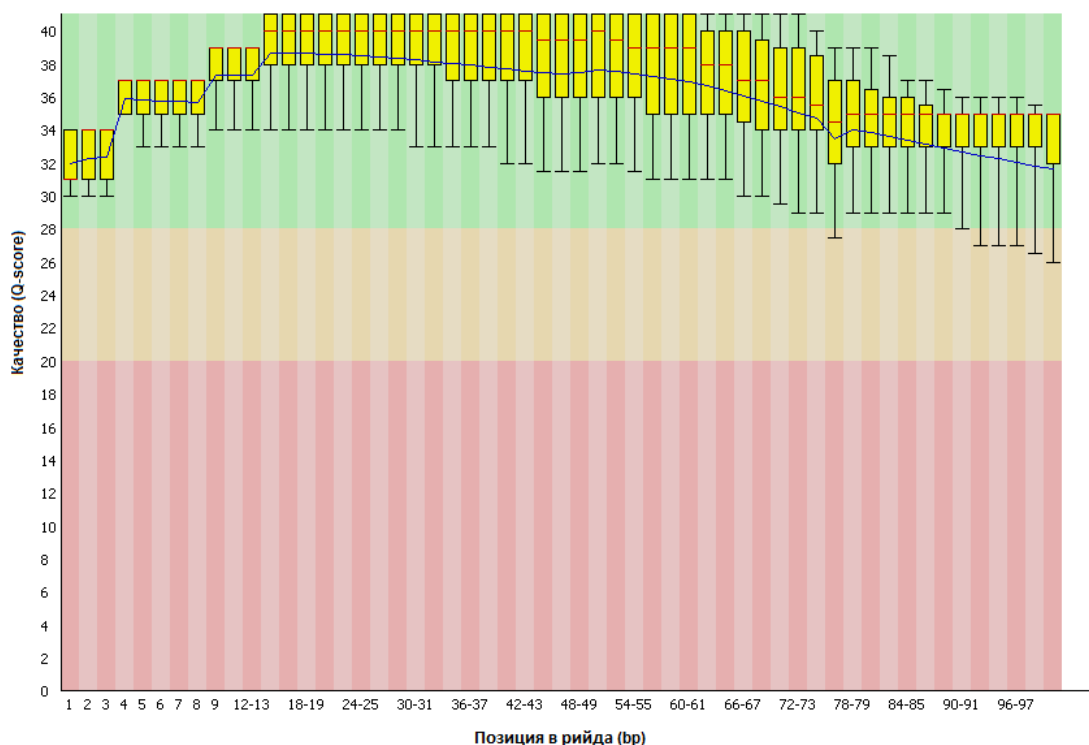
Анализът на данните от секвенирането е задължителна стъпка след етапа секвениране и необходимо и необходимо условие за провеждането на етапа асемблиране. Оценката на качеството (*Q-score* или *Phred score*) е изчисляване на вероятността за грешка в *base calling* (определяне на секвенцията от нуклеотиди в ДНК молекулата). Високата стойност на тази оценка означава, че *base call* е надежден и има малка вероятност за грешка. Скалата започва от Q10 (вероятност за грешка 0.1). Данните от двете проби бяха анализирани посредством софтуерната програма FastQC.



Фигура 1. Анализ на качеството на данните от секвениране за проба HR1 по бази чрез FastQC

Най-важната графика от анализа на качеството на данните е представена за първа и втора проба, съответно на Фигура 1 и Фигура 2. На графиките е показан интервала в който попадат стойностите на качеството на данните за всички бази на всяка позиция от рийда. На диаграмата има ясно очертани три полета, които отговарят на качеството, оцветени в зелено (много добро качество с *Q-score* над 28), оранжево (средно качество с *Q-score* между 20 и 28) и червено (много лошо качество с *Q-score* под 20).

Жълтите правоъгълници отразяват интервала на 25-75% от данните. Червената линия показва медианата на стойността на качеството, която попада в зелената област и е индикатор за много добро качество в интервала Q31-Q40 и за двете проби. Синята линия показва средно аритметичната стойност на качеството за всички бази и също попада в зелената област. Това потвърждава високото качество на данните и за двете проби.



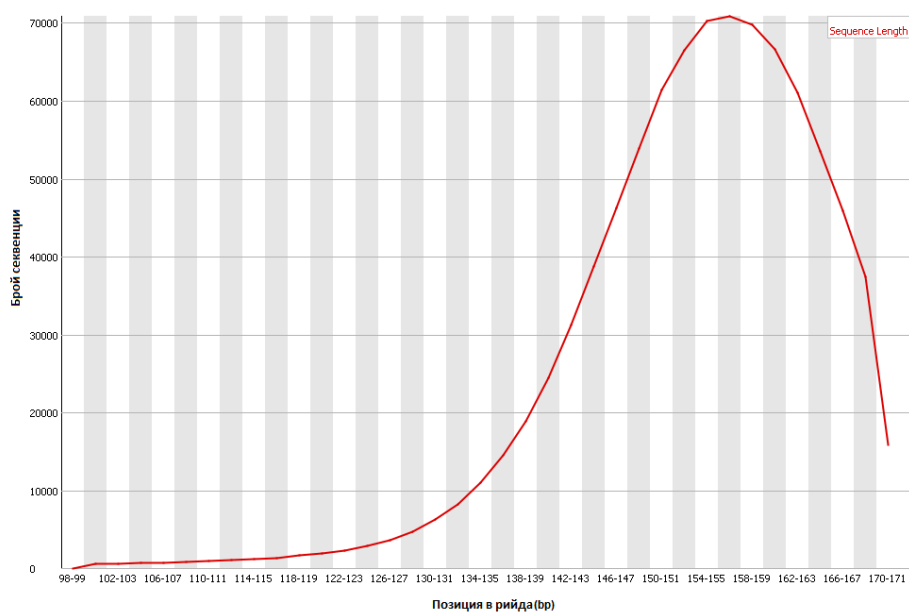
Фигура 2. Анализ на качеството на данните от секвениране за проба HR2 по бази чрез FastQC

II. Асемблиране и анотиране на хлоропластния геном на *Haberlea rhodopensis*

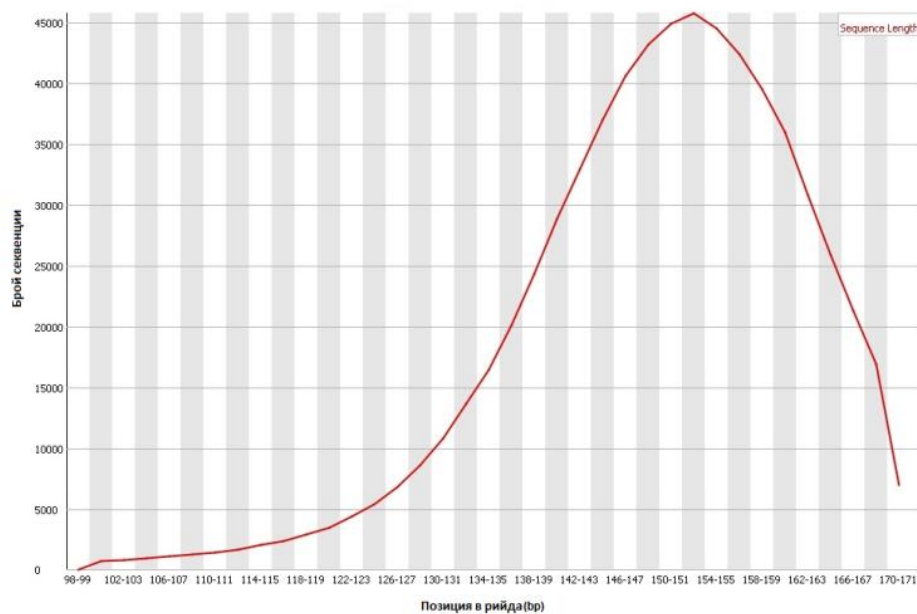
1. Асемблиране на хлоропластния геном на *Haberlea rhodopensis*

Непосредствено преди самото *de novo* асемблиране бе извършено картиране на рийдовете от секвенирането посредством BWA към всички хлоропластни геноми на Viridiplantae, налични в биоинформатичната базата от данни NCBI. Целта бе отфилтриране на нехлоропластните рийдове (Li & Durbin, 2010) и избягване на замърсяването от митохондриална и ядрена ДНК.

Оценка на качеството на получените след филтрирането хлоропластни рийдове бе извършено чрез FastQC и отново показва много добро качество. *Paired-end* рийдовете с дължина 100 bp бяха обработени със софтуерния инструмент FLASH (Magoc & Salzberg, 2011). Резултатът от тази процедура беше свързване на рийдовете, чрез използването на информацията за *library insert size*, и получаване на сравнително по-голяма дължина на рийдовете. Целта бе да се увеличи ефективността и качеството при асемблиране. Разпределението на дължините на новосинтезираните рийдове е представено чрез Фигура 3 и Фигура 4. От фигурите може да се види, че болшинството от рийдовете обработени с FLASH, имат дължина в интервала 140-160 bp.



Фигура 3. Разпределение на дължините на секвенциите за проба HR1 след обработката с FLASH



Фигура 4. Разпределение на дължините на секвенциите за проба HR2 след обработката с FLASH

За асемблирането на хлоропластния геном на *Haberlea rhodopensis* бяха избрани три от най-използваните асемблатори, работещи на различен принцип – Edena (на принципа на припокриването), ABySS и SPAdes (на принципа на графите на Брюйн).

Асемблаторите, базирани на графите на Брюйн бяха подложени на тестово асемблиране, в широк спектър от k-мер-и. След анализиране на резултатите, установихме, че с най-добри показатели бяха асемблиранията с k-mer=75 при SPAdes и k-mer =55 при ABySS. За асемблатор Edena бяха проведени тестове с различна стойност на припокриването, като най-добър резултат беше установен при m=60.

Като едни от основните критерии за анализ на качеството на асемблиранията е високо N50, голяма дължина на контига или скафолда, и по-малък общ брой генерирани контиги или скафолди. Анализът на резултатите от QUAST са представени в Таблица 1 и показват, че с най-добри показатели по отношение на дължина на скафолдите и N50 е асемблатор ABySS.

	Edena (m=60)	SPAdes (k-mer=75)	AbySS (k-mer=55)
Брой скафолди	56	129	71
Най-голям скафолд	11555	12657	27644
GC%	38.13	37.67	37.0
N50	2656	3922	20267
N75	1449	1645	16015

Таблица 1. Сравнение на резултатите от QUAST за трите асемблатора

Допълнителен анализ на фрагментирането на асемблиранията беше проведен, а резултатите от него обобщени в Таблица 2. С най-добри показатели отново е асемблатор ABySS, което означава, че асемблиранията с този софтуер са по-малко фрагментирани.

	Edena	SPAdes	ABySS
Брой празнини	55	70	19
Най-голяма празнина (bp)	15018	12279	7982
Обща дължина на празнините (bp)	48401	35739	25821

Таблица 2. Анализ на фрагментирането на асемблирането за трите асемблатора

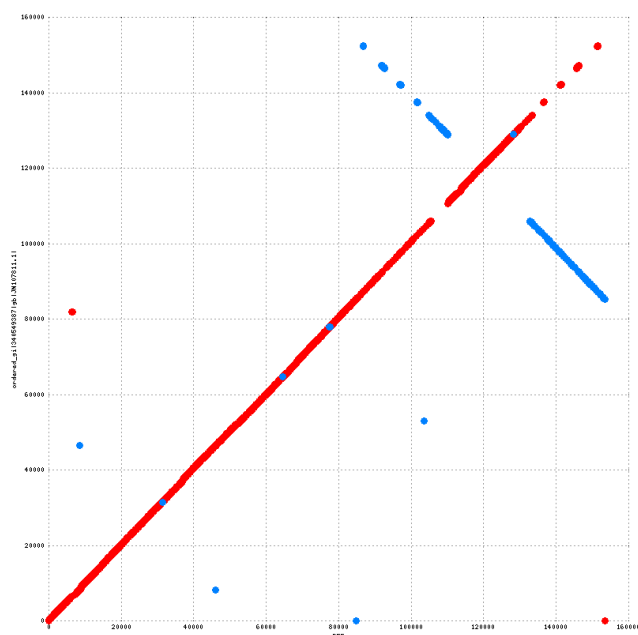
Въз основа на данните от тестовите и сравненията между параметрите на различните асемблатори, като основен асемблатор беше приет ABySS, версия 1.9.0, със стойност на параметъра на основното асемблиране k-mer =55.

Асемблирането беше извършено на два етапа – получаване на контиги и впоследствие получените контиги след асемблирането бяха пренесени към SSPACE 3.0 (Boetzer, Henkel, Jansen, Butler, & Pirovano, 2011) и обработени така, че да се получат скафолди. Етапите бяха анализирани чрез софтуерния инструмент за статистически анализ на асемблирания QCAST (Gurevich, Saveliev, Vyahhi, & Tesler, 2013). Параметрите N50, L50 , както и всички свързани с асемблирането статистики, са показани в Таблица 3.

	След ABYSS, контиги	След SSPACE , scaffolds
Контиги, брой	123	71
Най-голям контиг	14595	27644
N50	8354	20267
N75	3339	16015
L50	6	3
L75	13	5

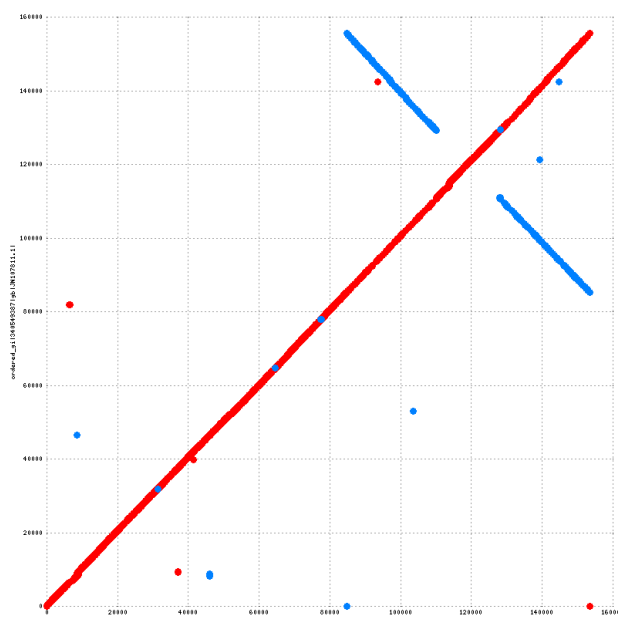
Таблица 3. Резултати от QCAST след асемблиране и след генериране на скафолди

Подреждането на скафолдите и формирането на първичния геном (Фигура 5) беше извършено чрез използването на референтния геном на *Boea hygrometrica* и с помощта на софтуерните програми Abacas 1.3.1 (Assefa, Keane, Otto, Newbold, & Berriman, 2009) и MUMmer (Kurtz et al., 2004).



Фигура 5. MUMmer плот на първичния геном на *Haberlea rhodopensis* към референтния геном на *Boea hygrometrica*

Получената първична версия на генома беше с лека фрагментация, затова се наложи запълване на празнините, образувани по време на подреждането. За целта бе използван софтуерния инструмент GapFiller (Nadalin, Vezzi, & Policriti, 2012) и чрез няколко цикъла празнините бяха запълнени. Крайният, финален геном (Фигура 6) отново бе сравнен чрез алайнмънт към референтния геном на *Boea hygrometrica* и потвърди наличието на четирите основни участъка в генома.



Фигура 6. MUMmer плот на финалния, първичен геном на *Haberlea rhodopensis*

2. Аотиране на хлоропластния геном на *Haberlea rhodopensis*

Асемблираните хлоропластни секвенции бяха аотирани, използвайки веб-базираната платформа за автоматично аотиране DOGMA (Wyman, Jansen, & Boore, 2004), чрез ръчна корекция на старт и стоп кодоните. Програмата аотира по голяма част от гена, но не успява да идентифицира с точност старт и стоп кодоните. Вместо това генерира алайнмънти с други видове за конкретния ген и дава възможност за ръчно определяне (Фигура 7). Има случаи, в които DOGMA дава изместена рамка на четене и невъзможност да аотира с точност даден ген, което е едно от несъвършенствата на тази програма.

psbI

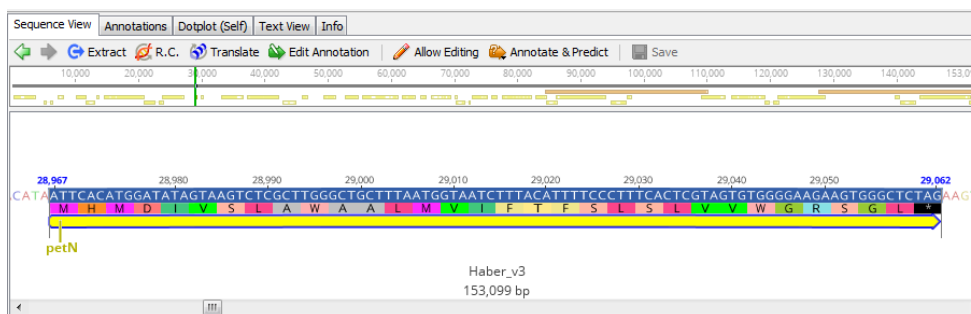
		Erodium texanum	M	L	T	L	K	L	F	V	Y	T	V	V	I	F
		Pelargonium x hortorum	M	L	T	L	K	L	F	V	Y	T	V	V	I	F
		Phoenix x thurifera	M	L	T	L	K	L	F	V	Y	T	V	V	I	F
Pinus thunbergii	M	I	Y	S	C	C	T	Y	N	Q	D	S	G	D	Y	V
Phoenix dactylifera	M	Y	S	L	F	Q	K	K	K	D	L	G	D	C	V	M
Lilian_final	I	Y	S	L	F	F	Q	K	N	H	L	G	D	C	V	M

ATAGGATATGTGGTAGAAAAATGGAAGATCTATTCTCTTTTTCCAAAAAATCATCTTGAGATTTGTAAATGCTTACTCTGAAACCTCTTCGTTTATACCGTAGTGATAATTTT
TATCCTATACACCATCTTTTACCTCTCAGATAAGAGAAAAAAGGTTTTTTATGAGAACCCTTAACACATTACGAATGAGACTTTGAGAAGCAAAATATGGCATCACTATAAA

Фигура 7. Анотиране с DOGMA

Поради тези факти, в процеса на анотация и за създаването на файла GFF (general feature format, файлов формат, който се използва за описанието на гени или други характеристики на ДНК, РНК и протеинови секвенции) използвахме паралелно програмата Geneious (Фигура 8).

За анотацията на протеин кодиращите гени използвахме кодировката translational table=11, която се използва основно за бактерии и пластиди. При тази кодировка, стартовия кодон на част от гените е различен от стандартния AUG (Фигура 8).



Фигура 8. Анотиране с Geneious

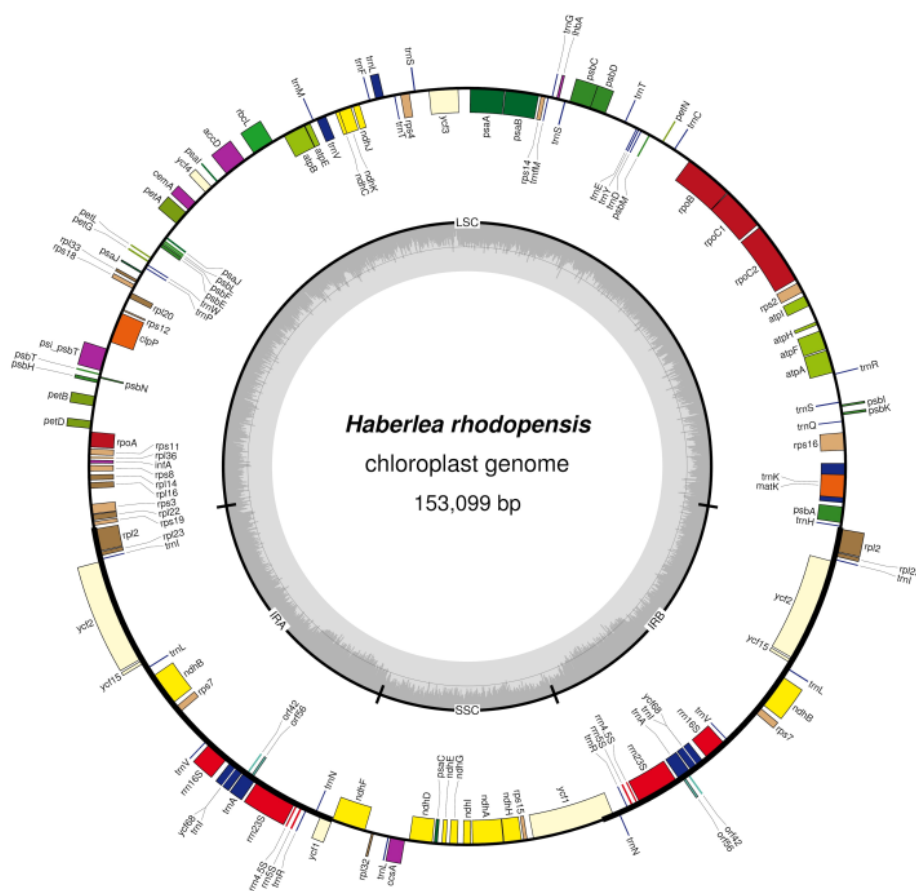
Валидация на анотираните гени беше извършена чрез Sequin (<http://www.ncbi.nlm.nih.gov/Sequin>) инструмента, наличен в биоинформатичната база от данни NCBI.

Номенклатурата на хлоропластните гени беше установена съгласно хлоропластната геномна база от данни (<http://chloroplast.cbio.psu.edu>), както и информацията от вече анотирани и публикувани хлоропластни геноми.

Гените за транспортни РНК бяха идентифицирани чрез DOGMA и впоследствие валидирани чрез tRNAscan-SE програмата, версия 1.21 (Schattner, Brooks, & Lowe, 2005). Кръговото представяне на хлоропластния геном на *Haberlea rhodopensis* беше генерирано чрез веб-базирания софтуер OGDRAW.

3. Организация на хлоропластия геном на *Haberlea rhodopensis* и генно съдържание

Хлоропластният геном на *Haberlea rhodopensis* (качен в GenBank под номер за достъп KX657870) е с дължина 153,099bp и притежава типичната за висшите растения структура, състояща се от четири части. Геномът включва региона LSC, с дължина 84,443bp (заемайки 55.2% от генома) и региона SSC (заемайки 11.6% от генома, които са разделени един от друг от регионите на два обърнати повтора IRs, с дължина 25,415bp (заемайки 33.20% от генома) (Фигура 9).



Фигура 9. Хлоропластния геном на *Haberlea rhodopensis*

Хлоропластния геном на *Haberlea rhodopensis* има GC% съдържание от 37.8%, като във регионите IRs то е значително по-високо (43.3%), в сравнение с останалите два региона – LSC (35,7%) и SSC(31,7%) . Повечето от секвенциите на хлоропластния геном кодират протеини, тРНК, рРНК (съответно 51.56%, 1.78%, 5.90%). Останалите региони, които не са протеин кодиращи, включват интрони, интергенни региони и псевдогени. Отрити са

137 гени в генома, като 86 от тях са протеин кодиращи, 4 тРНК, 4 псевдогена и 21 дублициращи гена (Таблица 4).

Категория	Група гени	Име на гена				
Гени за фотосинтеза	Субединици на фотосистема I	<i>psaA</i>	<i>psaB</i>	<i>psaC</i>	<i>psaI</i>	<i>psaJ</i>
	Субединици на фотосистема II	<i>psbA</i>	<i>psbC</i>	<i>psbD</i>	<i>psbE</i>	<i>psbF</i>
		<i>psbH</i>	<i>psbI</i>	<i>psbJ</i>	<i>psbK</i>	<i>psbL</i>
		<i>psbM</i>	<i>psbN</i>	<i>psbT</i>	<i>psi_psbT</i>	
	Субединици на цитохром b/f комплекс	<i>petA</i>	<i>petB</i>	<i>petD</i>	<i>petG</i>	<i>petL</i>
	Субединици на АТФ синтаза	<i>atpA</i>	<i>atpB</i>	<i>atpE</i>	<i>atpF^b</i>	<i>atpH</i>
		<i>atpI</i>				
	Голяма субединица на Rubisco	<i>rbcL</i>				
	Субединици за NADH дехидрогеназа	<i>ndhA^b</i>	<i>ndhB^{a,b}</i>	<i>ndhC</i>	<i>ndhD</i>	<i>ndhE</i>
		<i>ndhF</i>	<i>ndhG</i>	<i>ndhH</i>	<i>ndhI</i>	<i>ndhJ</i>
Гени за репликация	Рибозомни РНК гени	<i>rrn4.5^a</i>	<i>rrn5^a</i>	<i>rrn16^a</i>	<i>rrn23^a</i>	
	Транспортни РНК гени	<i>trnA-UGC_{a,b}</i>	<i>trnC-GCA</i>	<i>trnD-GUC</i>	<i>trnE-UUC</i>	<i>trnF-GAA</i>
		<i>trnG-UCC</i>	<i>trnH-GUG</i>	<i>trnI-CAU_a</i>	<i>trnI-GAU_{a,b}</i>	<i>trnK-UUU_b</i>
		<i>trnL-CAA^a</i>	<i>trnL-UAA_b</i>	<i>trnL-UAG</i>	<i>trnM-CAU</i>	<i>trnM-CAU</i>
		<i>trnN-GUU^a</i>	<i>trnP-UGG</i>	<i>trnQ-UUG</i>	<i>trnR-ACG^a</i>	<i>trnR-UCU</i>
		<i>trnS-GCU</i>	<i>trnS-GGA</i>	<i>trnS-UGA</i>	<i>trnT-GGU</i>	<i>trnT-UGU</i>
		<i>trnV-GAC^a</i>	<i>trnV-UAC_b</i>	<i>trnW-CCA</i>	<i>trnY-GUA</i>	
	Рибозомни протеини (SSU)	<i>rps2</i>	<i>rps3</i>	<i>rps4</i>	<i>rps7^a</i>	<i>rps8</i>
		<i>rps11</i>	<i>rps12</i>	<i>rps14</i>	<i>rps15</i>	<i>rps16^b</i>
		<i>rps18</i>	<i>rps19</i>			
	Рибозомни протеини (LSU)	<i>rpl2^{a,b}</i>	<i>rpl14</i>	<i>rpl16</i>	<i>rpl20</i>	<i>rpl22</i>
		<i>rpl23^a</i>	<i>rpl32</i>	<i>rpl33</i>	<i>rpl36</i>	
Други гени	РНК полимераза	<i>rpoA</i>	<i>rpoB</i>	<i>rpoC1^b</i>	<i>rpoC2</i>	
	Инициращ фактор за транслация	<i>infA</i>				
	Матураза	<i>matK</i>				
	Мембранен протеин	<i>cemA</i>				
	LHC PSII фактор I	<i>lhbA</i>				
	Субединица на ацетил-CoA	<i>accD</i>				
	Ц-тип цитохром синтазен ген	<i>ccsA</i>				
	Протеаза	<i>clpP^c</i>				
	Хипотетични ОРЧ	<i>ycf1^{a,d}</i>	<i>ycf2^a</i>	<i>ycf3^c</i>	<i>ycf4</i>	<i>ycf15^a</i>
		<i>ycf68^{a,d}</i>				
		<i>orf56^{a,d}</i>	<i>orf42^{a,d}</i>			

а) Две копия на гена в IRs; b) ген, съдържащ единичен интрон; c) ген, съдържащ два интрона d) псевдоген

Таблица 4. Генно съдържание на хлоропластния геном на *Haberlea rhodopensis*

От всички открити гени, 80 протеин кодиращи гена, 29 тРНК и 4 рРНК са установени, че са неповтарящи се в генома, докато 6 протеин кодиращи гена (*rpl2, rpl23, ycf2, ycf15, ndhB, rps7*), 7 тРНК (*trnI-CAU, trnL-CAA, trnV-GAC, trnI-GAU, trnA-*

UGC, *trnR-ACG*, *trnN-GUU*), и 4 рРНК гена (*rrn4.5*, *rrn5*, *rrn16*, *rrn23*) са открити, че се дуплицират в IRa и IRb регионите (Таблица 4).

Също така, идентифицирани са 4 псевдогена (*ycf1*, *ycf68*, *orf42*, *orf56*). Сред гените, които са претърпели псевдогенизация, само *ycf1* показва непълна дупликация (в IRa - SSC региона на свързка), докато останалите – *ycf68*, *orf42* и *orf56* показват напълно дупликация в IRa/IRb , предполагайки загуба на функция, дължаща се на натрупването на стоп кодони и изрязвания на гена.

Хлоропластният геном съдържа и 13 интрон-съдържащи гена (Таблица 5), от които 8 са протеин кодиращи (*rps16*, *atpF*, *rpoC1*, *ycf3*, *clpP*, *rpl2*, *ndhB*, *ndhA*) и 5 тРНК (*trnK-UUU*, *trnL-UAA*, *trnV-UAC*, *trnI-GAU*, *trnA-UGC*). Два от интрон-съдържащите гени имат два интрона (*clpP*, *ycf3*), а останалите 15 имат само 1 интрон.

Ген	Местопол.	Екзон I (bp)	Интрон I (bp)	Екзон II (bp)	Интрон II (bp)	Екзон II (bp)
<i>rps16</i>	LSC	48	914	213		
<i>atpF</i>	LSC	144	656	471		
<i>rpoC1</i>	LSC	456	781	1620		
<i>ycf3</i>	LSC	129	703	228	714	153
<i>clpP</i>	LSC	69	808	291	615	228
<i>rpl2</i> *	IR	393	669	435		
<i>ndhB</i> *	IR	777	680	756		
<i>ndhA</i>	SSC	558	941	540		
<i>trnK-UUU</i>	LSC	37	2527	26		
<i>trnL-UAA</i>	LSC	37	475	50		
<i>trnV-UAC</i>	LSC	38	582	37		
<i>trnI-GAU</i> *	IR	42	936	35		
<i>trnA-UGC</i> *	IR	38	823	35		

* Идентични дуплициращи се гени, съдържащи интрони в IR региона не са включени

Таблица 5. Интрон-съдържащи гени в хлоропластния геном на *Haberlea rhodopensis*

В допълнение, открити са 4 случая на припокриващи се гени (*psbD-psbC*, *ndhK-ndhC*, *atpE-atpB*, *ycf1-ndhF*).

III. Сравнителна геномика на пластома на *Haberlea rhodopensis*

1. Съпоставка на геномна структура и генен порядък на пластома на *Haberlea rhodopensis* с близкородствени видове

За по-задълбочено проучване, проведохме множествени секвенционни анализи между шест близкородствени хлоропластни генома (*Boea hygrometrica*, *Salvia miltiorrhiza*, *Olea europea*, *Sesamum indicum*, *Rosmarinus officinalis*, *Lavandula angustifolia*), с помощта на софтуера mVISTA (Poliakov, Foong, Brudno, & Dubchak, 2014) (Фигура 10). Сравняването на секвенции по двойки между хлоропластния геном на *Haberlea rhodopensis* и останалите геноми разкрива висока степен на консервативност на генния порядък, което предполага еволюционна консервативност на ниво геном, въпреки че съществува дивергенция в някои региони. Както при много други цъфтящи растения (Nie et al., 2012), кодиращите региони са много по-консервативни, в сравнение с не кодиращите региони. Най-голяма разлика в кодиращите участъци е установена при следните гени - *ycf1*, *ndhF*, *accD*, *ccsA*, *rbcL*, *ycf2* и *rps19* (Фигура 10). Установено е, че кодиращите участъци на гените *ycf1*, *rbcL* и *accD* показват значителна дивергенция в пластомите на други покритосеменни (Nie et al., 2012), което прави тези гени надеждни маркери за филогенетични анализи (Nazareno, Carlsen, & Lohmann, 2015). В геномите на *S.indicum* и *S.miltiorrhiza* генът *rps19* е дублициран, при *Haberlea rhodopensis* е анотиран като протеин-кодиращ ген, докато единствено при *Boea hygrometrica* този ген е открит като псевдоген (T. Zhang et al., 2012).



Фигура 11. Сравнение на регионите IR/SSC и IR/LSC на хлоропластния геном на *Haberlea rhodopensis* с три близкородствени генома

Счита се, че регионите IR са едни от най-консервативните региони във хлоропластните геноми сред всички видове. Разширяването и стесняването на границите на IR регионите се разглежда като еволюционно събитие и е основната причина за вариацията в размера на хлоропластните геноми. Връзките между IR и съответно SSC, LSC са показател на еволюцията на хлоропластните геноми (H. Zhang, Li, Miao, & Xiong, 2013). Във връзка с това твърдение, беше направено сравнение на регионите IR/LSC и IR/SSC в генома на *Haberlea rhodopensis* със съответните региони в геномите на близкородствените *Boea hygrometrica*, *Olea europea* и *Sesamum indicum* (Фигура 11). Както беше установено от предишни проучвания на хлоропластни геноми (Nazareno et al., 2015), разширяването и стесняването на IR в разред Lamiales, е довело до структурни промени в хлоропластните геноми, в резултат на което са се формирали псевдогени. От Фигура 11 ясно се вижда разположението на гена *ycf1*, лежащ точно на границата IRa/SSC, което вероятно е довело до псевдогенизирането му. Твърдението е валидно и за четирите сравнявани геноми, където генът се намира на една и съща позиция и е загубил своята функция като протеин-кодиращ, т.е. аотиран е като псевдоген. Също така, е установено припокриване между псевдогена *ycf1* и *ndhF* гена в геномите на *Haberlea rhodopensis*, *Boea hygrometrica* и *Olea europea*. Втората граница IRb/SSC попада в CDS областта на протеин кодиращия ген *ycf1*, и това е валидно за четирите изследвани генома. Генът *trnH* се намира изцяло в региона LSC и при четирите сравнявани генома. Разликата единствено е в отнемването спрямо

границата IRb/LSC, което варира в интервала 0-44bp. Интересното при гена *rps19* е, има малки вариации в разположението си при сравняваните геноми. При *Boea hygrometrica* и *Sesamum Indicum*, по-голяма си част от него е разположена в областта LSC, а само малка част преминава в областта IRa. При геномите на *Haberlea rhodopensis* и *Olea europaea* се наблюдава разположение на гена изцяло в областта LSC, с малка разлика в отстоянието до границата LCS/IRa.

2. Анализ на прости повтори (SSR) в генома на *Haberlea rhodopensis*

Чрез програмата MISA анализирахме разпределението на SSR в генома на *Haberlea rhodopensis*. Открихме общо 71 броя SSR, показващи общо 10 типа съставни мотиви на повтарящата се единица (Таблица 6).

Размер на изследваната секвенция (bp):	153099
Общ брой идентифицирани SSRs:	71
Брой съставни SSR типа:	10
SSR тип	Честота
A/T	24
C/G	1
AC/GT	4
AG/CT	15
AT/AT	19
AAG/CTT	1
AAT/ATT	1
AAAC/GTTT	1
AAAT/ATTT	2
AACT/AGTT	1
AATC/ATTG	1
AGAT/ATCT	1

Таблица 6. Идентифицирани SSR в хлоропластния геном на *Haberlea rhodopensis*

Сред идентифицираните повтарящи се единици, динуклеотидната единица (AG/CT и AT/AT) е с най-голям брой повторения. Общо 29 SSR са открити в генните региони на *Haberlea rhodopensis*, като седем гена (*atpF*, *rpoC2*, *rpoA*, *rpl2*, *ycf2*, *ndhB*, *rrn23*) от тях съдържат най-малко по два броя SSR. Интересно е да се отбележи, че броят на идентифицираните SSR е по-нисък в сравнение с предишни анализи на SSR в геноми на органели като *Boea hygrometrica*, *O.europeae* и *S.indicum* (Mariotti, Cultrera, Díez, Baldoni, & Rubini, 2010; Yi & Kim, 2012; T. Zhang et al., 2012). В допълнение, трябва да посочим, че не бяха установени голям брой от три- и тетра- нуклеотидни повторения, което

контрастира с предишни изследвания (Qian et al., 2013). Също така, не бяха открити никакви penta- или hexa- нуклеотидни повтарящи се единици в генома на *Haberlea rhodopensis*.

В допълнение към анализа за идентифициране на SSR, ние проведохме изследване за идентифициране на дълги повтори чрез софтуерния продукт REPuter (Kurtz et al., 2001).

Размер (bp)	Позиция 1	Позиция 2	Тип	Местоположение	Регион	Е-стойност
59	109798	109857	F	ycf1	IRA	1.98E-26
59	109857	127696	P	ycf1	SSC	1.98E-26
56	84299	84443	F	rps19, IGS	LSC	1.27E-24
56	84299	153043	P	rps19, IGS(rpl2, trnH-GUG)	IRA, IRB	1.27E-24
55	36105	36159	F	IGS(trnS-UGA, lhbA)	LSC	5.08E-24
41	98642	119664	F	IGS(rps7, trnV-GAC), ndhA	IRA, SSC	1.36E-15
41	119664	138870	P	ndhA, IGS(trnV-GAC, ndhB)	SSC, IRB	1.36E-15
40	59562	59562	P	IGS(accD, psaI)	LSC	5.45E-15
37	57457	57494	F	IGS(rbsL, accD)	LSC	3.49E-13
44	38927	41151	F	psaB, psaA	LSC	7.62E-12
37	6418	80983	F	IGS(rps16, trnQ-UUG), IGS(rps8, rpl14)	LSC	3.87E-11
33	113317	113349	F	IGS(rpl32, trnL-UAG)	SSC	8.93E-11
42	43919	119663	F	ycf3, ndhA	LSC, SSC	1.06E-10
39	43922	138870	P	ycf3, IGS(trnV-GAC, rps7)	LSC, IRB	1.45E-10
39	43922	98644	F	ycf3, IGS(rps7, trnV-GAC)	LSC, IRA	1.45E-10
31	162	199	P	IGS(trnH-GUG, psbA)	LSC	1.43E-09
30	8063	45627	P	IGS(psbI, trnS-GCU)	LSC	5.72E-09

33	63845	63845	P	IGS(petA, psbJ)	LSC	8.84E-09
36	91664	145820	P	ycf2	IRA,IRB	2.69E-07
36	91664	91682	F	ycf2	IRA	2.69E-07
36	91682	145838	P	ycf2	IRA,IRB	2.69E-07
36	145820	145838	F	ycf2	IRB	2.69E-07
35	94093	94093	P	IGS(ycf15, trnL-CAA)	IRA	9.87E-07
35	94093	143408	F	IGS(ycf15,trnL-CAA), CAA,ycf15) IGS(trnL-	IRA,IRB	9.87E-07
35	143408	143408	P	IGS(trnL-CAA, ycf15)	IRB	9.87E-07
32	8501	8501	P	IGS(trnS-GCU,trnR-UCU)	LSC	1.60E-06
34	74589	74599	P	IGS(psbT, psbN)	LSC	3.61E-06
31	20927	20929	P	rpoC1	LSC	5.98E-06
30	13264	13264	P	IGS(atpF, atpH)	LSC	2.24E-05
31	54809	66123	P	IGS(atpB,rbcl), IGS(psbE,petL)	LSC	1.73E-04
30	9660	36718	F	IGS(trnS-GCU,trnR-UCU), trnG-UCC	LSC	6.27E-04
30	43934	138867	P	ycf3, IGS(trnV-GAC, rps7)	LSC, IRB	6.27E-04
30	43934	98656	F	ycf3, IGS(rps7, trnV-GAC)	LSC, IRA	6.27E-04
30	89251	148215	P	ycf2	IRA, IRB	6.27E-04
30	89251	89293	F	ycf2	IRA	6.27E-04
30	89293	148257	P	ycf2	IRA, IRB	6.27E-04
30	91675	145815	P	ycf2	IRA, IRB	6.27E-04
30	91675	91693	F	ycf2	IRA	6.27E-04
30	91693	145833	P	ycf2	IRA, IRB	6.27E-04
30	148215	148257	F	ycf2	IRB	6.27E-04

Таблица 7. Идентифицирани дълги повтори в хлоропластния геном на *Haberlea rhodopensis*

Идентифицирахме общо 40 повтора, от които 18 прави и 22 палиндромни повтора (Таблица 7). Повторите са с дължина, която варира между 30bp и 59 bp за дадена единица. Около 40% от тези повтори попадат в интергенното пространство, докато 60% са разположени в гените или на техните граници. За разлика от някои близкородствени видове като *S.Indicum*, *B.hygrometrica*, *O.europeae*, където са открити съответно 15, 8 и 3 повтора (Nazareno et al., 2015), генома на *Haberlea rhodopensis* съдържа доста голям брой от повтарящи се елементи.

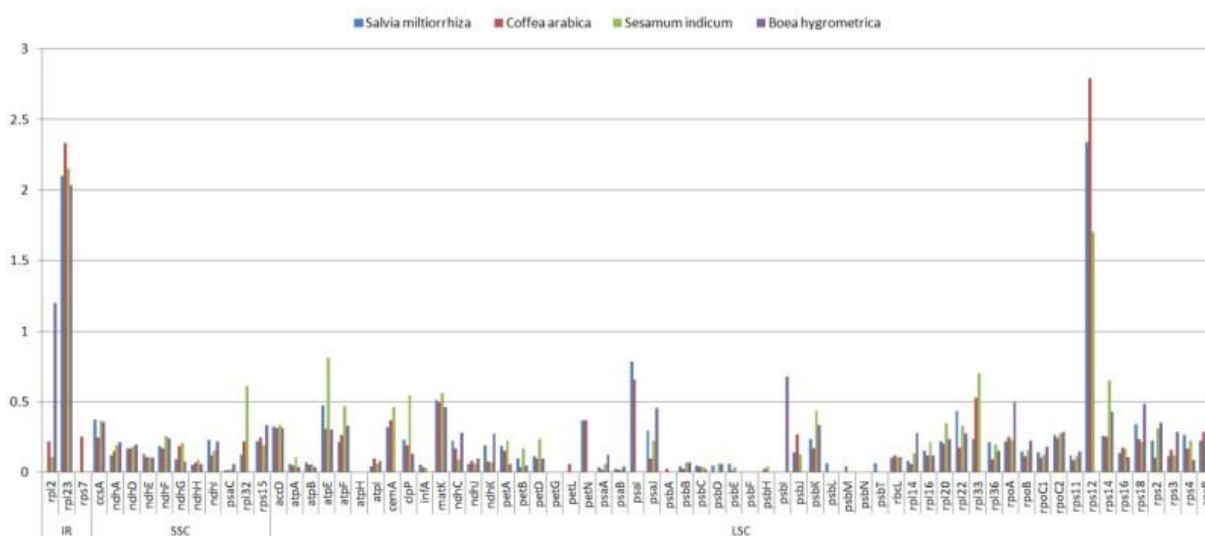
3. Анализ на синонимни и несинонимни замени (Ka/Ks анализ)

Синонимните и несинонимни нуклеотидни замени са важни маркери в еволюционните изследвания. В повечето гени, с изключение на тези, които еволюират много бързо, несинонимните нуклеотидни замени се случват много по-рядко в сравнение с синонимните нуклеотидни замени, което се дължи на пречистваща селекция. Следователно, съотношението $Ka/Ks < 1$ (особено, когато е по-малко от 0.5) означава пречистваща селекция; съотношението $Ka/Ks > 1$ означава позитивен отбор, докато съотношението Ka/Ks близо или равно на 1 означава релаксиран отбор (Kimura, 1983).

Ние изчислихме съотношението Ka/Ks на хлоропластия геном на *Haberlea rhodopensis* към три близкородствени генома от разред Lamiales- *B.hygrometrica*, *S.indicum*, *S.miltiorrhiza*, както и към генома на *Coffea arabica*. Нашият анализ показва, че съотношението Ka/Ks не е регион – специфично (регионите IR, SSC и LSC показват сходни стойности), а по-скоро заключаваме, че е ген-специфично. Средната стойност на Ka/Ks за 72 протеин-кодиращи гена, изчислена в 5 генома е 0.2139. Най-консервативните гени, със стойност на Ka/Ks между 0 и 0.01, с индикация за пречистващ отбор са следните: *atpH*, *psaA*, *psaB*, *psaC*, *psbA*, *psbE*, *psbF*, *spbH*, *spbL*, *psbM*, *psbN*, *petG*, *petL* и *infA*.

Средно аритметичния модел в Ka/Ks калкулатора показва съотношението $Ka/Ks > 1$ за гена *rps12* (кодиращ голямата субединица на рибозомалния протеин 23) и за гена *rpl23* (кодиращ рибозомалния протеин S12). Увеличеното съотношение Ka/Ks е индикатор за позитивен отбор, което би могло да е отражение на специфичния селекционен натиск на *Haberlea rhodopensis*. Другият ген – кандидат за позитивна селекция е генът *rps12*. Той показва висока средна стойност (2.15) на Ka/Ks съотношението в анализа по двойки между *Haberlea rhodopensis* и останалите 4 вида. В сравнение, средната стойност на Ka/Ks по двойки между останалите видове дава стойност 0.633. Стойностите на Ka/Ks в

интервала от 0.5 до 1.0 (указвайки релаксиран отбор) са отчетени за гените *rpl2*, *rpl32*, *atpE*, *psaI*, *psbI*, *matK* и *clpP* (Фигура 12).



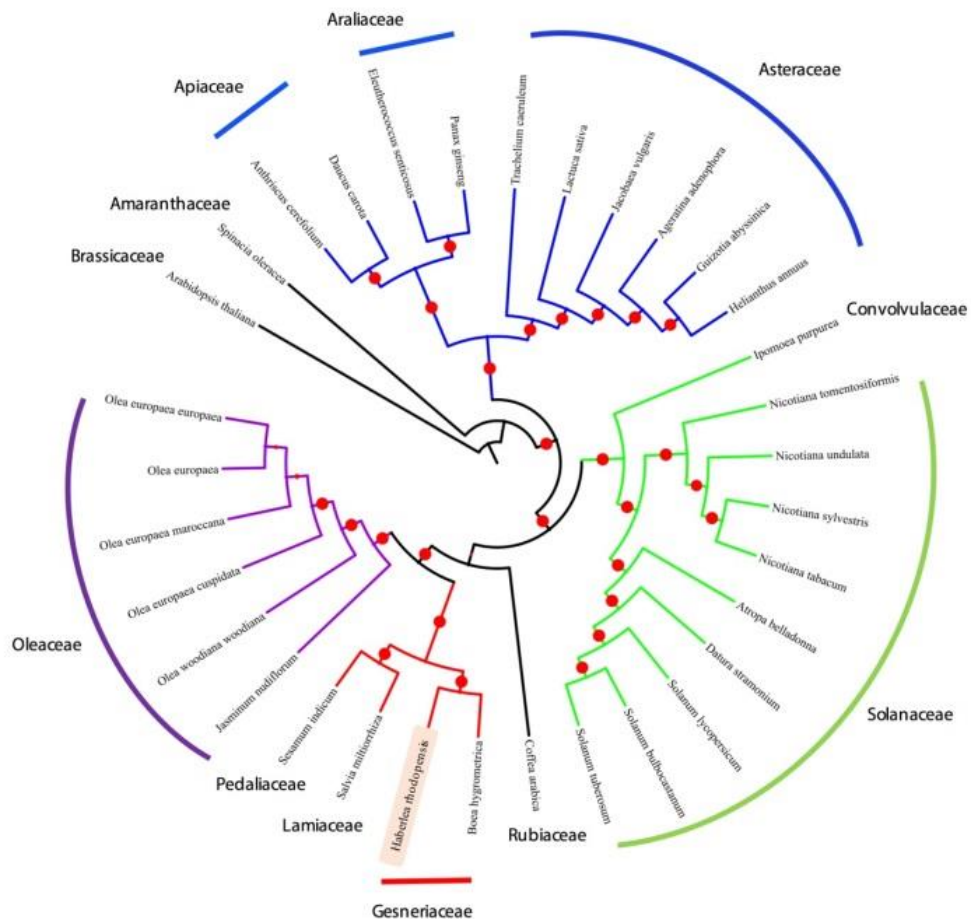
Фигура 12. Ka/Ks анализ

Особен интерес при *Haberlea rhodopensis* представлява нейния възкръсващ фенотип с бързо възстановяване на нивата на фотосинтеза след рехидратация. Възможен е натиск на релаксиран отбор върху гените *psaI* и *psbI*, кодиращи протеини за тилакоидната мембрана, което говори за адаптация на фотосинтетичния апарат на този вид към повтарящо се засушаване и дехидратация. Релаксирания отбор за АТФ синтаза субединица *atpE*, също би могло да бъде свързано с адаптацията към стрес от засушаване.

4. Филогенетичен анализ

За изследването на еволюцията на кодиращите региони е извършен алайнмънт на кодиращи секвенции на *Haberlea rhodopensis* с 32 други хлоропластни генома. Изборът на тези геноми за сравнителния анализ е основан на факта, че всички те споделят обща еволюционна класа, което ги прави близки във филогенетично отношение. В допълнение, избрани са хлоропластни геноми от разред Lamiales като близка група до изследвания геном на *Haberlea rhodopensis*. Добавен е също така и вида *Arabidopsis thaliana*, който се явява външна група за секвенирания геном. За целта са подбрани *gene* *atpA*, *atpB*, *atpE*, *atpF*, *atpH*, *atpI*, *ndhC*, *ndhD*, *ndhE*, *ndhF*, *ndhH*, *ndhI*, *ndhJ*, *petA*, *petB*, *petD*, *petG*, *psaA*, *psaB*, *psaC*, *psaI*, *psbA*, *psbB*, *psbC*, *psbD*, *psbE*, *psbH*, *psbK*, *psbL*, *psbM*, *psbN*, *rbcL*, *rpoA*, *rpoB*, *rpoC2*.

Построяването на филогенетичното дърво беше извършено чрез програмата IQTree, на което ясно могат да бъдат проследени еволюционните взаимоотношения между изследваните хлоропластни генома, като с червено е отбелязано мястото на *Haberlea rhodopensis* в семейство Gesneriaceae (Фигура 13).



Фигура 13. Филогенетично дърво на базата на алайнмънт на кодиращите секвенции на *Haberlea rhodopensis* с 32 други хлоропластни генома

ИЗВОДИ

1. За повишаване на качеството на суровите данни от масовото паралелно секвениране на хлоропластния геном на *Haberlea rhodopensis*, беше извършено обединяване на paired-end данните като получените секвенции бяха с дължина ~170bp, като това доведе до повишаване на качество им до Q65-Q70.
2. Тестовото асемблиране на хлоропластните секвенции със софтуери Edena, SPAdes и ABySS установи, че последният е с най-добри статистически показатели по отношение на анализирания асемблираня.
3. Асемблирането с ABySS генерира 123 контига с параметри N50 = 8 354 и най-дълъг контиг = 14 595 bp. След премахване на фрагментацията на асемблирането чрез GapFiller беше получена циклична псевдомолекула с дължина 153, 099 bp.
4. Анотацията на хлоропластия геном на *Haberlea rhodopensis* показва наличие на общо 137 гена, от които 86 протеин-кодиращи (80 основни, 6 повтарящи се), 8 рРНК (4 рРНК основни, 4 повтарящи се) , 36 тРНК (29 тРНК основни и 7 повтарящи се) и 4 псевдогена.
5. Най-голяма разлика в кодиращите участъци на хлоропластия геном на *Haberlea rhodopensis* е установена при следните гени - *ycf1*, *ndhF*, *accD*, *ccsA*, *rbcL*, *ycf2* и *rps19*. Гените *psbI* и *rps12* са открити като псевдогени при най-близкородствения вид *Boea hygrometrica* докато при *Haberlea rhodopensis* са функционални гени.
6. Анализа на синонимни и несинонимни кодони показва най-висока консервативност при *atpH*, *psaA*, *psaB*, *psaC*, *psbA*, *psbE*, *psbF*, *spbH*, *spbL*, *psbM*, *psbN*, *petG*, *petL* и *infA*, а най-висока изменчивост при *rps12* и *rpl23*.

ИЗПОЛЗВАНА ЛИТЕРАТУРА В АВТОРЕФЕРАТА

- Assefa, S., Keane, T. M., Otto, T. D., Newbold, C., & Berriman, M. (2009). ABACAS: Algorithm-based automatic contiguation of assembled sequences. *Bioinformatics*, 25(15), 1968–1969. <https://doi.org/10.1093/bioinformatics/btp347>
- Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D., & Pirovano, W. (2011). Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics*, 27(4), 578–579. <https://doi.org/10.1093/bioinformatics/btq683>
- Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: Quality assessment tool for genome assemblies. *Bioinformatics*, 29(8), 1072–1075. <https://doi.org/10.1093/bioinformatics/btt086>
- Kimura, M. (1983). The neutral theory of molecular evolution and the world view of the neutralists.
- Kurtz, S., Choudhuri, J. V., Ohlebusch, E., Schleiermacher, C., Stoye, J., & Giegerich, R. (2001). REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Research*, 29(22), 4633–4642. <https://doi.org/10.1093/nar/29.22.4633>
- Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., & Salzberg, S. L. (2004). Versatile and open software for comparing large genomes. *Genome Biology*, 5(2), R12. <https://doi.org/10.1186/gb-2004-5-2-r12>
- Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, 26(5), 589–595. <https://doi.org/10.1093/bioinformatics/btp698>
- Magoc, T., & Salzberg, S. L. (2011). FLASH: Fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, 27(21), 2957–2963. <https://doi.org/10.1093/bioinformatics/btr507>
- Mariotti, R., Cultrera, N. G. M., Díez, C. M., Baldoni, L., & Rubini, A. (2010). Identification of new polymorphic regions and differentiation of cultivated olives (*Olea europaea* L.) through plastome sequence comparison. *BMC Plant Biology*, 10, 211. <https://doi.org/10.1186/1471-2229-10-211>
- Nadalin, F., Vezzi, F., & Policriti, A. (2012). GapFiller: a de novo assembly approach to fill the gap within paired reads. *BMC Bioinformatics*, 13 Suppl 1(Suppl 14), S8. <https://doi.org/10.1186/1471-2105-13-S14-S8>
- Nazareno, A. G., Carlsen, M., & Lohmann, L. G. (2015). Complete chloroplast genome of *Tanaecium tetragonolobum*: The first Bignoniaceae plastome. *PLoS ONE*, 10(6), 1–18. <https://doi.org/10.1371/journal.pone.0129930>
- Nie, X., Lv, S., Zhang, Y., Du, X., Wang, L., Biradar, S. S., ... Weining, S. (2012). Complete chloroplast genome sequence of a major invasive species, crofton weed (*Ageratina adenophora*). *PLoS ONE*, 7(5). <https://doi.org/10.1371/journal.pone.0036869>
- Poliakov, A., Foong, J., Brudno, M., & Dubchak, I. (2014). GenomeVISTA-an integrated software package for whole-genome alignment and visualization. *Bioinformatics*, 30(18), 2654–2655. <https://doi.org/10.1093/bioinformatics/btu355>
- Qian, J., Song, J., Gao, H., Zhu, Y., Xu, J., Pang, X., ... Chen, S. (2013). The Complete Chloroplast Genome Sequence of the Medicinal Plant *Salvia miltiorrhiza*. *PLoS ONE*, 8(2). <https://doi.org/10.1371/journal.pone.0057607>
- Schattner, P., Brooks, A. N., & Lowe, T. M. (2005). The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Research*, 33(SUPPL. 2), 686–689. <https://doi.org/10.1093/nar/gki366>
- Wyman, S. K., Jansen, R. K., & Boore, J. L. (2004). Automatic annotation of organellar genomes with DOGMA. *Bioinformatics*, 20(17), 3252–3255. <https://doi.org/10.1093/bioinformatics/bth352>

- Yi, D. K., & Kim, K. J. (2012). Complete chloroplast genome sequences of important oilseed crop *Sesamum indicum* L. *PLoS ONE*, 7(5). <https://doi.org/10.1371/journal.pone.0035872>
- Zhang, H., Li, C., Miao, H., & Xiong, S. (2013). Insights from the complete chloroplast genome into the evolution of *Sesamum indicum* L. *PLoS ONE*, 8(11). <https://doi.org/10.1371/journal.pone.0080508>
- Zhang, T., Fang, Y., Wang, X., Deng, X., Zhang, X., Hu, S., & Yu, J. (2012). The complete chloroplast and mitochondrial genome sequences of *Boea hygrometrica*: Insights into the evolution of plant organellar genomes. *PLoS ONE*, 7(1). <https://doi.org/10.1371/journal.pone.0030531>

СЪКРАЩЕНИЯ

IR	Обърнат повтор
LSC	Голямо единично копие
SSC	Малко единично копие
cpDNA	Хлоропластна ДНК
pPHK	Рибозомна РНК
тPHK	Транспортна РНК
NGS	Секвениране от ново поколение
WGS	Секвениране на цели геноми
SSR	Прости повтори, микросателити
BLAST	Локален алгоритъм за сравняване на секвенции
CDS	Кодираща област
Codon	Кодон (триплет)
ORF	Отворена рамка на четене
Bp	Базови двойки
Kbp	Килобази (1000 базови двойки)
Nt	Нуклеотиди
GFF	Основен файлов формат за аотиране

**НАУЧНИ ПРИНОСИ,
свързани с дисертационния труд**

I. Приноси, съдържащи оригинални, нови за науката факти

- За първи път е секвениран и асемблиран хлоропластния геном на *Haberlea rhodopensis*.
- За първи път е описана структурата на хлоропластния геном на *Haberlea rhodopensis*.
- За първи път са анотирани гените, тяхната позиция и структура в хлоропластния геном на *Haberlea rhodopensis*.

II. Приноси с методичен характер

- Софтуерът за асемблиране ABySS е установен, като оптимален за асемблиране на хлоропластния геном на *Haberlea rhodopensis*.

**СПИСЪК НА НАУЧНИТЕ ПУБЛИКАЦИИ,
свързани с дисертационния труд**

1. **Ivanova Z. et al.** (2017) Chloroplast Genome Analysis of Resurrection Tertiary Relict *Haberlea rhodopensis* Highlights Genes Important for Desiccation Stress Response. *Frontiers in Plant Science* 8:204. (IF 4.495)
2. **Ivanova Z. et al.** (2017) Chloroplast genome assembly approaches from NGS data. Journal of BioScience and Biotechnologies. (in press)