ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
"НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
"ВЫСШАЯ ШКОЛА ЭКОНОМИКИ"
ФАКУЛЬТЕТ КОМПЬЮТЕРНЫХ НАУК

Шнайдер Элина Дмитриевна

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

АНАЛИЗ ЭВОЛЮЦИИ СЕМЕЙСТВ ГЕНОВ У АЛЛОТЕТРАПЛОИДА *CAPSELLA BURSA-PASTORIS*

Analysis of the Allotetraploid *Capsella Bursa-pastoris* Gene Families Evolution

по направлению подготовки 01.04.02 Прикладная математика и информатика

образовательная программа «Анализ данных в биологии и медицине»

Студент

_____

Э.Д. Шнайдер

Научный руководитель:

к.б.н., доцент

Пятницкий М.А.

Соруководитель:

М.н.с. лаборатории геномики растений ИППИ РАН

Клепикова А.В.

Москва 2020

Аннотация

Полиплоидизация является важнейшим событием в эволюции высших растений, во многом определяющее их разнообразие и доминирование среди других растительных групп. Однако события, происходящие после полногеномной дупликации в различных популяциях полиплоидных видов, по-прежнему мало изучены. Работа посвящена изучению популяций недавнего аллотетраплоида Пастушья сумка обыкновенная (*Capsella bursa-pastoris L. Medik.*), который входит в число самых распространенных растений на Земле и обладает колоссальной экологической пластичностью. Для этого были проанализированы данные полногеномного секвенирования 68 растений, принадлежащих к трем популяциям. Согласно полученным результатам, все растения на уровне популяций имеют как структурные отличия генома, так и отличия в генах, ответственных за развитие и ответы на биотические и абиотические стрессы. Отдельно был проведен анализ на наличие положительного отбора в генах, входящих в сеть ответа на холодовой стресс, который показал, что все три популяции идут разными эволюционными путями. Полученные результаты могут пролить свет на причины эволюционного успеха и необычайной способности к адаптации *C. bursa-pastoris*, а также на ранние события, имеющие место после полиплоидизации.

Abstract

Polyploidization was a crucial event in the evolution of Embryophyta which in many ways defined their diversity and their dominance among other plant taxa. However, the changes happening after the whole-genome duplication in different populations of polyploid species has not been studied well enough. This study examines the populations of a recent allotetraploid *Capsella bursa-pastoris* L. Medik., which is considered one of the most wide-spread plants on Earth and has extreme ecological plasticity. The study analyses whole-genome sequencing data for 68 plants from three populations. The results demonstrate, that all the populations are characterized by structural genome differences as well as differences between genes involved in development and response to biotic and abiotic stresses. An analysis of the positive selection in genes controlling the cold stress response has shown, that the evolution of the three populations has taken different directions. The results shed light on the reasons *of C. bursa-pastoris* evolutionary success and unique adaptivity as well as on the early stages of its evolution following polyploidization.

Оглавление

Введение

Полиплоидизация является ключевым событием в эволюции высших растений. Тем не менее, вопрос, что происходит с геномом после полногеномной дупликации у представителей различных популяций полиплодиного вида, по-прежнему остается открытым.

Однолетнее травянистое растение Пастушья сумка обыкновенная (*Capsella bursa-pastoris*) – недавний аллотетраплоид, произошедший 100.000–300.000 лет назад в результате скрещивания видов *C. rubella/grandiflora* и *C. orientalis.* На сегодняшний день *C. bursa-pastoris* является наиболее успешным представителем небольшого рода *Capsella* и одним из самых часто встречающихся видов Покрытосеменных на Земле. В отличие от своих диплоидных родительских видов, имеющих небольшой ареал, *C. bursa-pastoris* расселилась по всей планете, во всех климатических поясах, кроме арктического и антарктического. Такое широкое распространение свидетельствует о колоссальной приспособленности и экологической пластичности представителей этого вида.

В современных исследованиях выделяется три большие популяции *C. bursa-pastoris,* которые различаются по морфологическим и генетическим признакам: Дальневосточную, представители которой встречаются на территории Китая и Юго-Восточной Азии; Ближневосточную, обитающие, в основном, по берегам Средиземного моря; Европейскую, распространенную по всей Евразии. Также, данные хлоропластных и митохондриальных геномов данные свидетельствуют об отличном от других популяций происхождении Дальневосточной популяции.

Цель исследования заключалась в поиске свойств геномов и генов растений, происходящих из разных популяций, которые могут помочь в изучении того, каким образом вид *C. bursa-pastoris* смог распространиться по всей планете и приспособиться ко всем климатическим поясам.

1. Обзор литературы

1.1. Полиплоидизация и ее роль в эволюции Покрытосеменных

Полиплоидизация, или кратное увеличение набора хромосом, является частым событием в эволюции растений, влияющим как на размер генома, так и на количество генов. Несмотря на то, что полиплоидия встречается у всех основных групп наземных растений (Jiao et al., 2011), наиболее характерна она для представителей отдела Покрытосеменных (Magnoliophyta), или Цветковых – одной из самых успешных групп наземных растений, в настоящий момент включающей в себя более 295 тысяч видов (Christenhusz et al., 2016; Bell et al., 2010)

Наиболее важными причинами эволюционного успеха Покрытосеменных можно назвать явления полиплоидии и межвидовой гибридизации. Согласно литературе, ~30–35% от всех описанных видов Цветковых являются полиплоидными, 70-80% видов однодольных и двудольных имеют полиплоидное происхождение, 70% предков которых проходили один и более раундов полногеномной дупликации (Lewis., 1980; Stebbins, 1971; Masterson, 1994).

События древней полиплоидизации происходили у предка всех представителей семенных растений (событие ξ) и у раннего предка Цветковых (событие ε) (Jiao et al., 2011). Согласно реконструкции генома самого раннего предка Magnoliophyta показывает, что анцестральный геном был представлен 1175 протогенами и 15 протохромосомами (Murat et al., 2017). В дальнейшем у предка Цветковых произошла редукция и стабилизация генома, сопровождающаяся потерей генов, делециями и транслокациями, неофункциализацией и субфункциализацией, а также слиянием хромосом. Все эти процессы привели к тому, что ранние Цветковые разделились на: отдел Двудольные с семью протохромосомами и отдел Однодольные с пятью протохромосомами.

Значительная часть существующих ныне представителей Magnoliophyta подвергалась полиплоидизации в прошлом и их геномы свидетельствуют о неоднократном прохождении их предками полногеномных дупликаций и вторичной диплоидизации. Так, даже виды с небольшими геномами являются палеополиплоидами. Например, растение с пятью хромосомами *Arabidopsis thaliana*, содержит в своем геноме следы двух недавних полногеномных дупликаций, произошедших внутри семейства Brassicaceae (события α и β) и одного события трипликации (событие γ), произошедшего у предка двудольных (Simillion et al., 2002, De Bodt et al., 2005; Bowers et al., 2003).

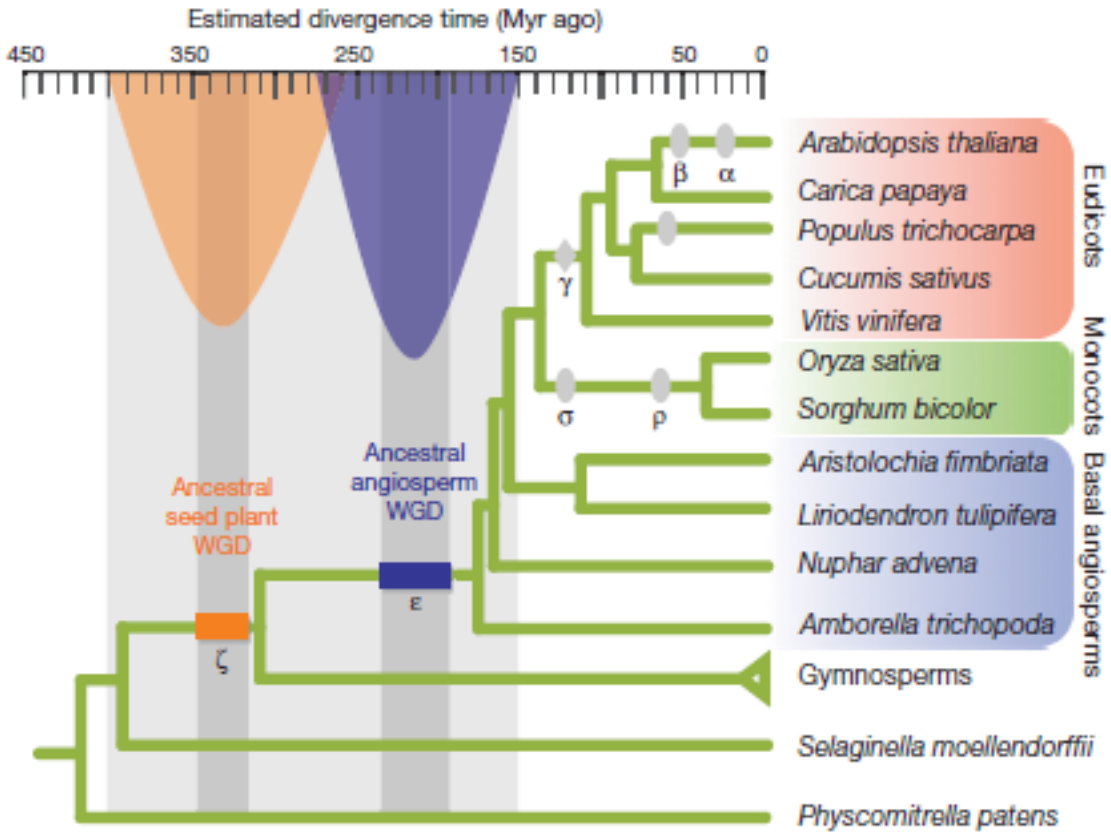


Рисунок 1. Древние события полиплоидизации в истории Покрытосеменных растений (по Jiao et al., 2011).

Полиплоиды могут возникать внутри одного вида вследствие спонтанного события полногеномной дупликации (автополиплоидия), так и в результате межвидовой гибридизации с последующими полногеномными дупликациями (аллополиплоидия). За событиями полиплоидизации следует стабилизация генома, сопровождающаяся потерей функций дублированных генов, делециями, транслокациями, субфункциализацией и неофункциализацией.

Наиболее значительный вклад в видообразование вносит аллополиплоидизация. Как показывают опыты по ресинтезу видов, в первых же поколениях наблюдаются значительные геномные перестройки, а в последующих поколениях могут возникать новые морфологические признаки. Так, анализ генома ресинтезированного аллотетраплоидного арахиса *Arachis hypogaea* выявил высокий уровень гомологичной рекомбинации между двумя геномами, что привело к спонтанному изменению цвета венчика у некоторых линий после нескольких поколений. (Tate et al., 2006; Bertioli et al., 2019; Gaeta et al., 2007) Изучение геномов ранних аллотетраплоидов так же показало, что, помимо геномных перестроек и потерей и появлением новых функций генами, изменения могут быть связаны с приобретением новых и потерей

существующих регуляторных элементов, перемещениями и изменениями в количестве мобильных элементов (Kasianov et al., 2017; Renny-Byfield et al., 2014; Ozkan et al., 2001; Ungerer et al., 2006; Douglas et al., 2015;)

Еще одним следствием полиплоидизации является возрастающая экологическая пластичность, позволяющая заселять нехарактерные для диплоидных видов территории. Одним из примеров подобного географического распределения, является вид *A. arenosa*, имеющий диплоидную и тетраплоидную расы. Так, диплоид имеет ареал обитания, ограничивающийся Карпатами, Балканами, Паннонской равниной и побережьем Балтийского моря, тогда как тетраплоид населяет всю Западную, Центральную и Восточную Европу (по Novikova et al., 2018).

Согласно (Rice et al., 2019), частота встречи полиплоидных растений увеличивается от экватора. Наиболее значительным предиктором их распространения являются климатические условия, особенно температура, и жизненная форма, но при этом зависимость от таксономического состава и видового разнообразия экологического региона отсутствует (рис. 2)

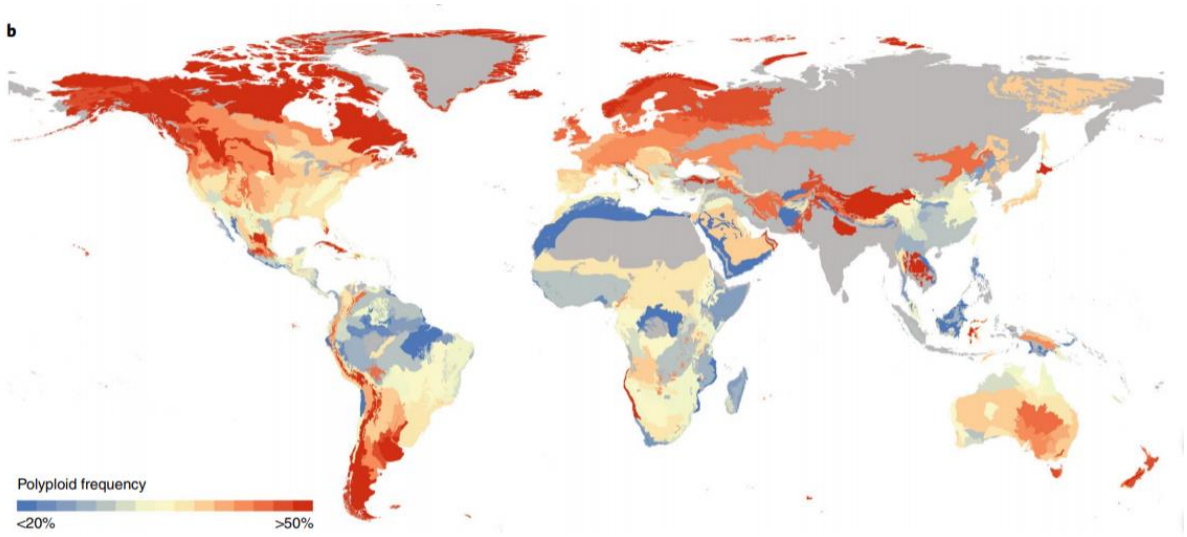


Рисунок 2. Процент полиплоидных видов представлен для наземных экологических регионов. Экологические регионы с недостаточными данными окрашены в серый цвет, в результате чего. Синий цвет обозначает <20% полиплоидных видов, в темно-красный с > 50% (из Rice et al., 2019)

Таким образом, полиплоидизация является основным фактором в эволюции цветковых растений, а также обеспечивает внутривидовое генетическое, морфологическое и экологическое разнообразие.

1.2. Род *Capsella* и популяционная структура *Capsella bursa-pastoris.*

Род *Capsella*, принадлежащий семейству Крестоцветные (Brassicaceae), небольшой и в настоящее время включает в себя четыре признанных вида однолетних травянистых растений, три из которых являются диплоидными (*C. grandiflora, C. rubella* и *C. orientalis*) и один тетраплоидным (*C. bursa-pastoris*) (Sicard & Lenhard., 2018). До недавнего времени так же выделяли пятый вид – *C. thracica*, который считали аллотетраплоидом, произошедшим в результате скрещивания между ранее считавшимся автополиплоидным видом *C. bursa-pastoris* в качестве материнского вида и *C. grandiflora* в качестве отцовского (Hurka et al., 2012), затем *C. thracica* стала считаться синонимом *C. bursa-pastoris*. Так же род *Capsella* является филогенетически наиболее близким к роду *Arabidopsis* – модельному объекту генетики и геномики растений.

Представители рода *Capsella* различаются не только плоидностью, но и местами обитания и типами опыления. Так, диплоидные виды имеют ограниченные ареалы обитания: самоопылитель *C. rubella* распространен на территории Средиземноморья и точечно встречается в Северной Америке и Австралии. Второй диплоидный самоопыляемый вид – *C. orientalis*, населяющий часть Восточной Европы и Азии. Самое узкое распространение имеет перекрестно опыляемый вид *C. grandiflora* произрастает в западной Греции, на некоторых греческих островах, в Албании и редко в северной Италии. Напротив, самоопыляемый тетраплоид *C. bursa-pastoris* распространен по всему миру во всех климатических зонах, за исключением антарктического и субантарктического. (https://www.gbif.org/) и является одним из самых распространенных растений на планете, что свидетельствуют о большой экологической пластичности (Guo et al., 2009; Hurka et al., 2012; Coquillat, 1951) (рис. 3)

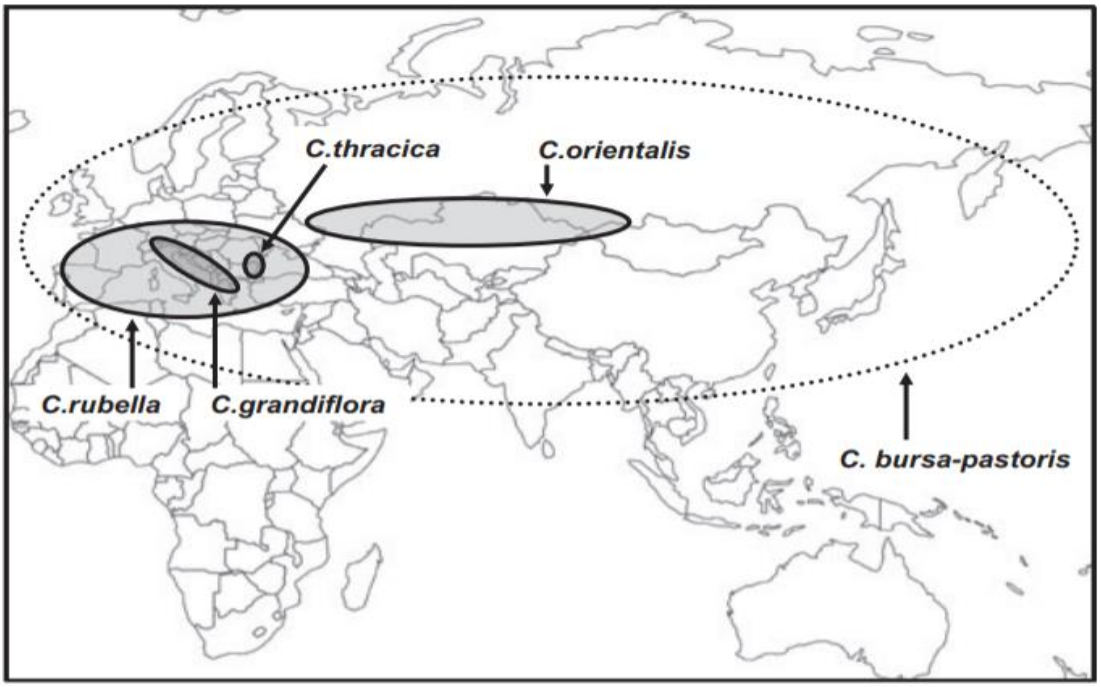


Рисунок 3. Распространение видов рода *Capsella* (Hurka et al., 2012)

Уровень полиморфизма, как по фенотипическим признаком, так и по геномам, внутри популяций *C. grandiflora* и *C. rubella* невысок. (Slotte et al., 2013) Так же в (Hurka et al., 2012; (Hurka et al., 1997; Slotte et al., 2006) показано, что эти два диплоида имеют одного и того же предка и общее происхождение, а их расхождение связано с различиями систем опыления (Brandvain et al., 2013). Для третьего диплоидного вида *C. orientalis* тоже отмечается отсутствие внутривидового полиморфизма, что указывает на единое происхождение этого вида (Hurka et al., 2012)

Однако, для полиплоида *C. bursa-pastoris* характерен высокий уровень полиморфизма между популяциями по морфологии и физиологии, имеющего наследственную природу. (Neuffer et al., 1989; Hurka et al., 1991). При этом растения внутри вида *C. bursa-pastoris* часто более далеки, чем два диплоидных вида *C. grandiflora* и *C. rubella* друг от друга (Hurka et al., 1997; Wu et al., 2015).

О большом внутривидовом полиморфизме у *C. bursa-pastoris* свидетельствуют данные пластидных микросаттелитов (Ceplitis et al., 2005), генотипирования с помощью секвенирования (GBS) и полногеномного секвенирования (WGS). Согласно результатам, полученным с помощью последних двух методов, *C. bursa-pastoris* подразделяется в Евразии на три генетические группы с низким потоком генов между ними: одна из них объединяет

растения преимущественно из Западной Европы и Юго-Восточной Сибири («европейская» популяция), вторая сосредоточена на Ближнем Востоке («ближневосточная» популяция) и третья – на Дальнем Востоке («дальневосточная» популяция) (Cornille et al., 2016; Kryvokhyzha et al., 2019).

1.3.Происхождение аллотетраплоида *C. bursa-pastoris*

Пастушья сумка обыкновенная является недавним аллотетраплоидам, появившимся приблизительно 100.000–300.000 лет назад в результате гибридизации двух диплоидных представителей рода *Capsella* (Douglas et al., 2015). Геном *C. bursa-pastoris,* также, как и геномы большинства известных полиплоидов, претерпевает перестройки, связанные с диплоидизацией, ранние стадии которой сопровождались неофункционализацией, ослаблением отбора и псевдогенизацией, приобретением новых регуляторных элементов и потерей старых. Согласно существующим публикациям (Douglas et al., 2015; Kasianov et al., 2017) прародителями *C. bursa-pastoris* являются передок современных *C. rubella* и C. *grandiflora*, а также *C. orientalis.*

Однако результаты филогенетического анализа (рисунок 4), проведенного мной ранее на основе пластидного и митохондриального геномов и изложенного в курсовой работе «Анализ происхождения аллотетраплоида *Capsella bursa-pastoris* с использованием данных полногеномного секвенирования», свидетельствуют о множественном происхождении этого полиплоида. Так, материнским видом для европейских и ближневосточной групп *C. bursa-pastoris* является предок *C. orientalis*, в то время как для дальневосточных ни этот вид, ни другие ныне известные виды *C. rubella* и C. *grandiflora* материнскими не являются.

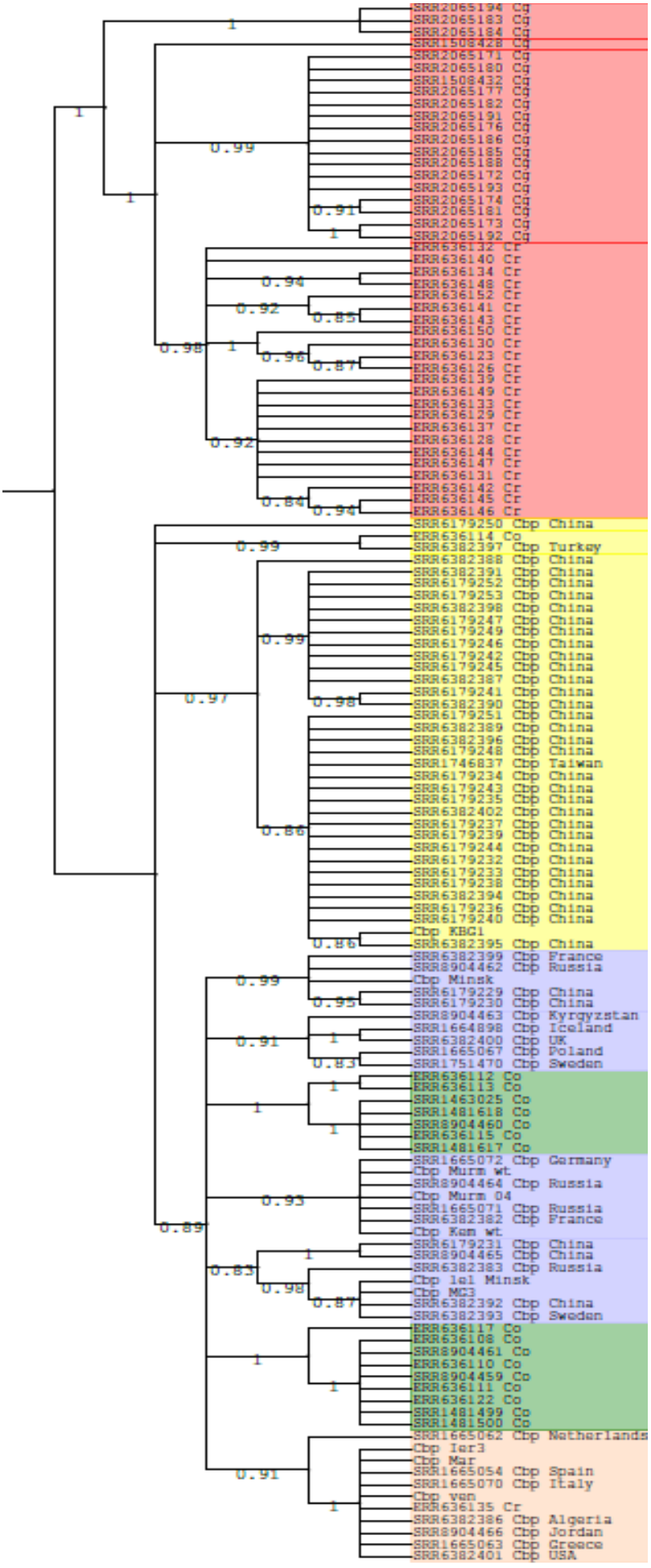


Рисунок 4. Филогенетическое дерево, построенное методом максимального правдоподобия по пластидным данным, с указанием вида и места произрастания в случае *C. bursa-pastoris*

Cg – *C. grandiflora* (помечена красным цветом)

Cr – *C. rubella* (помечена красным цветом)

Cbp – *C. bursa-pastoris*

Co – *C. orientalis* (помечена зеленым цветом)

Желтым цветом обозначены растения, образующие дальневосточную группу, голубым отмечены растения из европейской группы и коричневым – растения, образующие ближневосточную группу.

2. Материалы и методы

2.1.Используемые данные

Для работы были использованы данные полногеномного секвенирования 70 образцов растений *C. bursa-pastoris*, из которых 55 линий были взяты из базы данных SRA (https://www.ncbi.nlm.nih.gov/sra) и 13 были получены из Лаборатории эволюционной геномики Факультета Биоинженерии и Биоинформатики МГУ им М.В. Ломоносова и Лаборатории геномики растений ИППИ РАН. Согласно проведенному мной ранее филогенетическому анализу на основе пластидного генома, каждое растение отнесено к одной из трех популяций: европейской (EUR), среднеазиатской (ME) и дальневосточной (ASI). Таким образом в анализе участвует 25 растений из европейской, 10 из среднеазиатской и 33 растения из дальневосточной популяции. Список линий представлен в таблице № 1 приложения.

Использованные данные полногеномного секвенирования содержали чтения длиной от 100 до 150 нуклеотидов, полученные на платформе Illumina. Инструменты, на которых получены публичные данные из базы SRA: HiSeq 2000, HiSeq 2500, и MiSeq. Инструменты, использованные в Лаборатория эволюционной геномики: HiSeq 2000 и MiSeq. Чтения были триммированы с помощью программы Trimmomatic-0.38 (Bolger et al., 2014) по наличию адаптеров и по качеству (параметры: -phred33, :2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36).

2.2. Картирование на референсный геном и расчет базовых статистик

Для проведения дальнейшей работы все данные для 68 растений были картированы на ядерный референсный геном *C. bursa-pastoris*, взятый из базы NCBI Genome (код сборки GCA_001974645.1), представленный 8186 скаффолдами и 32319 контигами. Файл с аннотация к геному был взят с сайта http://capsella.org/.

Картирование проводилось с помощью программы bwa mem (версия 0.7.12, Li, 2013) со следующими параметрами:

1) -t 12, распределяющим расчеты на 12 параллельных потоков

2) -R "@RG\tID:<имя образца>\tLIB:<имя образца>\tPL:ILLUMINA\tSM:<имя образца>", присваивающий группу каждому чтению

3) -M, для совместимости с используемой далее программой GATK4

Далее результаты картирования перенаправлялись на вход программе samtools view -bS (версия 1.10, https://www.htslib.org/), которая переводит файл с выровненными на геном ридами в бинарный. После всего, с помощью программы samtools flagstat были получены описательные статистики картирования, в том числе процент картированих на геном чтений.

Код для выполнения описанной части анализа представлен ниже:

```
bwa mem -t 12 -M -R \
"@RG\tID:<префикс образца>\tLIB:$F\tPL:ILLUMINA\tSM:<префикс образца>" \
<референсный геном> <файл с чтениями в одну сторону> \
<файл с чтениями в другую сторону> | samtools view -bS - > bwa/<файл с картированием>
```

```
samtools flagstat bwa/<файл с картированием> -O tsv > bwa/stat/<файл со статистикой картирования>
```

2.3.Анализ покрытия

Важной характеристикой выравнивания данных полногеномного секвенирования на референсный геном является покрытие. Расчет среднего покрытия проводился для каждого скаффолда и контига с помощью программы samtools depth с параметром -a для учета всех позиций в геноме. Среднее покрытие для каждого гена было получено при использовании программы bedtools multicov.

```
samtools sort <файл с картированием>  <сортированный файл с картированием>
samtools index <сортированный файл с картированием>
samtools depth -a <сортированный файл с картированием> -o <таблица с покрытием>
bedtools multicov -bams <сортированный файл с картированием> \
-bed <файл с координатами генов> > <файл со средним покрытием для каждого гена>
```

Поскольку каждый из образцов имеет разную глубину секвенирования и размер библиотек сильно варьирует от одного анализируемого растения к другому, необходимо привести нормировку, после которой становится возможным проводить сравнения. Среднее покрытие генов было нормировано на длину (нормировка RPKM). Для полного генома сначала

была посчитана скользящая медиана покрытия внутри каждого скаффолда с длиной более 1000 нуклеотидов, а затем полученные значения каждой позиции в скаффолде были нормированы на размер библиотеки и умножены на 10^6. Код для выполнения описанной части анализа представлен в разделе 2.6

2.4. Функциональная аннотация

Предсказание семейства для каждого гена *C. bursa-pastoris* было выполнено программой pfamScan из пакета HMMER3 (Mistry et al., 2013), которая относит каждый ген к семейству на основании профилей hmm из базы PFAM. Далее была произведена фильтрация по значению E-value < 10^{-5}. В результате 17870 генов были отнесены к одному из 1870 семейств.

Поскольку не для всех генов было найдено семейство напрямую, был сделан поиск ортологов для каждого гена *C. bursa-pastoris* из ближайшего родственника с хорошо собранным и аннотированным геномом *A. thaliana*. Файл со всеми белковыми последовательностями *A. thaliana* был взят из базы Araport, версии 11 (https://www.araport.org/). Соотнесение ортологов *A. thaliana* к генам *C. bursa-pastoris* выполнялось с помощью программы blastp-2.2.31 (Camacho et al., 2008) по белковым последовательностям с фильтрацией по E-value < 10^{-5}.

Далее по ортологичным генам *A. thaliana* были найдены семейства для тех генов *C. bursa-pastoris*, к которым не удалось найти с помощью pfamScan.

```
pfam_scan.pl -fasta <fasta файл с белковыми последовательностями> -dir ./ -o <файл с результатами>

makeblastdb -in < fasta файл с белковыми последовательностями A. thaliana> -dbtype prot

makeblastdb -in < fasta файл с белковыми последовательностями C. bursa-pastoris> -dbtype prot

blastp -query <fasta файл с белками C. bursa-pastoris> faa -db <белки A. thaliana> -out <файл с результатами> -evalue 1E-5 -outfmt 6
```

2.5.Генотипирование и построение консенсусов

Генотипирование каждого из 68 растений проводилось с помощью программы HaplotypeCaller из пакета GATK4 (Poplin et al. 2017; McKenna et al. 2010). Затем программой SelectVariants были отобраны варианты с покрытием менее 6 для того, чтобы в дальнейшем замаскировать такие варианты как недостаточно покрытые. Далее, программой vcftools-0.1.14 (Danecek et al. 2011) однонуклеотидные полимофизмы (SNP) и индели были отфильтрованы в отдельные файлы и из них удалены те, которые имели более двух аллелей, а затем из каждого

файла были удалены варианты с покрытием менее шести ридов с помощью bcftools-1.11 (https://www.htslib.org/).

По файлам с отфильтрованными и маскированными однонуклеотидными полиморфизмами, а также файлу с координатами генов в программе FastaAlternateReferenceMaker из пакета GATK4 были построены консенсусные последовательности генов для каждого из анализируемых растений.

Код для выполнения описанной части анализа представлен ниже:

```
gatk --java-options "-Xmx4g" HaplotypeCaller \
  -R <референсный геном> \
  -I <сортированный файл с картированием> \
  -O <файл с результатами генотипирования> \
  -L <файл с геномными интервалами>

gatk --java-options "-Xmx4g" SelectVariants \
  -R <референсный геном> \
  -V <файл с результатами генотипирования> \
  -O <файл с маскированными вариантами> \
  --select "DP < 6"
vcftools --vcf <файл с результатами генотипирования> --remove-indels --max-alleles 2 --recode --
recode-INFO-all --out <файл с SNP>
vcftools --vcf <файл с результатами генотипирования> --keep-only-indels --max-alleles 2 --recode --
recode-INFO-all --out <файл с инделями>
bcftools view -i 'DP>=6' <файл с SNP> > <файл с SNP, отфильтрованными по покрытию>
bcftools view -i 'DP>=6' <файл с инделями> > <файл с инделями, отфильтрованными по покрытию>


gatk --java-options "-Xmx4g" FastaAlternateReferenceMaker \

  -R <референсный геном> \
  -V <файл с SNP, отфильтрованными по покрытию> \
  -O <консенсус в формате fasta>\
  --snp-mask <файл с маскированными вариантами> \
  -L <файл с геномными интервалами>
```

2.6.Расчет генетического смешения и субпопуляционного индекса фиксации

После этапа генотипирования и фильтрации все vcf-файлы, содержащие только однонуклеотидные полиморфизмы, были объединены с помощью программы bcftools merge. Для расчета субпопуляционного индекса фиксации (F_{st}) полученный файл был подан на вход программе Populations из пакета Stacks-2.54 (Catchen et. al, 2013) с параметрами:

- -M <файл>, указывающий на файл, содержащий информацию о принадлежности каждого образца к одной из трех популяций
- -V <.vcf>, указывающий на объединенный файл с генотипированием
- --fstats для расчета по данным генотипирования F-статистик, в частности F_{st}
- --smooth-fstats, позволяющий применить ядерное сглаживание для уменьшения шума в результатах
- --bootstrap-fst 100, выполняющая 100 бутстреп-реплик для сглаженных значений Fst
- --plink перезаписывающий входной vcf в формат PLINK

В результате работы программы Populations были получены сглаженные значения F_{st} для трех парных сравнений (ASI-ME, ASI-EUR, ME-EUR) в каждом из анализируемых локусов, а так же файл в формате PLINK, который был переформатирован с помощью программы plink-2.0 (Purcell et al., 2007) и отправлен на вход программе Admixture (Alexander et al., 2011) для расчета генетического смешения с параметром -K 3, обозначающим три предковые популяции. Результаты Admixture были визуализированы с помощью языка R.

Анализируемые SNP в каждом из трех сравнений были отфильтрованы по значению сглаженного $F_{st} > 0.25$ и сопоставлены с генами. Затем для каждого отобранного гена из каждого сравнения было проверено медианное покрытие, расчеты которого описаны в пункте 2.3. Гены, где часть последовательности была непокрыта совсем в одной из линий или же покрытие было менее 6, из анализа удалялись.

```
bgzip <индивидуальные vcf>
tabix <индивидуальные vcf.gz>
bcftools merge <файлы в формате .vcf.gz > -O vcf -o <объединенный vcf.gz >
populations -V <объединенный vcf.gz > \
        -M popmap \
        --fstats \
        --smooth-fstats \
        --bootstrap-fst 100 \
        --plink \
        -O ./
plink --file <.ped>--recode --out <переформатированный ped>
```

```
admixture -K3 < переформатированный ped>
```

2.7. Отбор генов с высоким F_{st}

После фильтрации генов по покрытию, они были разбиты на три группы: гены из субгенома A, предполагаемый предок которого *C. orientalis*, гены из субгенома B, родственного *C. rubella*, и гены, для которых не известно, к какому субгеному они относятся.

Каждому гену с известной принадлежностью к одному из двух субгеномов был сопоставлен гомеологичный ген из другого генома, если такой был. Затем последовательности каждого отобранного гена и его гомеолога были извлечены из консенсусных последовательностей и записаны как выравнивание в формате fasta. Далее, из каждого гена были удалены интроны. Последовательности были проверены на наличие стартового кодона ATG и при необходимости переведены в комплементарную цепь и записаны в направлении 5'-3'. Код на языке Python3 для выполнения описанных выше действий:

```
import glob
from Bio import AlignIO
import pandas as pd
import os

fasta = glob.glob(path+ directory+"*.fasta")
scaff_names=pd.read_csv(path+"scaf_names_gene_names.txt", sep="\t", names = ['coord', 'gene'])
exons=pd.read_csv(path+"Capsella_bursa_pastoris.gff", sep="\t", names = ['contig','version', 'cds', 'begin',
'end', 'V5', "V6", 'V7', 'gene'])
exons = exons[exons.cds != 'gene']

for i in range(0, len(fasta)):
   try:
      alignment = AlignIO.read(open(fasta[i]), "fasta")
      gene = "Gene="+fasta[i].split("/")[-1][:-6]+';'
      gene_mark = exons.query("gene == @gene")
      gene_mark = gene_mark.reset_index(drop = True)
      f = open(path+ directory+'New/'+fasta[i].split("/")[-1][:-6]+"_new.fna", "w")
      for j in range(0,(len(alignment))):
         f = open(path+ directory+'New/'+fasta[i].split("/")[-1][:-6]+"_new.fna", "a")
         f.write("\n"+">"+alignment[j].id+"\n")
         if gene_mark.shape[0] == 1:
            end = int((gene_mark.end[0]-gene_mark.begin[0]))
            f.write(str(alignment[j].seq[0:end+1]))
         else:
            end = int((gene_mark.end[0]-gene_mark.begin[0]))
            f.write(str(alignment[j].seq[0:end+1]))
            for g in range(1,gene_mark.shape[0]):
               intron_length = (gene_mark.begin[g]-gene_mark.end[g-1])
```

```
                length_last = int((gene_mark.end[g-1]-gene_mark.begin[g-1]))
                end_last = int((gene_mark.end[g-1]-gene_mark.begin[0]))
                begin_next = intron_length+end_last
                length = (gene_mark.end[g]-gene_mark.begin[g])
                end_next = begin_next+length+1
                f.write(str(alignment[j].seq[begin_next:end_next]))
    except ValueError:
        print("file " + fasta[i] + " ok")

arr = os.listdir(path + directory+'New/')
for j in range(0, len(arr)):
    try:
        alignment = AlignIO.read(open(path + directory+'New/' + arr[j]), "fasta")
        record = alignment[0]
        if record.seq[0:3] == "ATG":
            print("good file " + arr[j])
        else:
            f = open(path + directory+'New/' + arr[j], "w")
            for i in range(0,(len(alignment))):
                f = open(path + directory+'New/' + arr[j], "a")
                f.write(">"+alignment[i].id+"\n")
                f.write(str(alignment[i].seq.reverse_complement()+'\n'))
                f.close()
    except ValueError:
        print("file " + arr[j] + " doesn't exist")
```

Выравнивание каждого гена было транслировано в белковые последовательности с помощью пакета EMBOSS (Rice et al., 2000) и просмотрены на наличие несинонимичных замен, отличающихся между популяциями.

2.8. Поиск генов с признаками положительного отбора в сети генов ответа на холодовой стресс

Гены *A. thaliana*, участвующим в ответе на холодовой стресс (Park et al., 2015), были сопоставлены с ортологами в *C. bursa-pastoris*. Последовательности каждого отобранного гена и его гомеолога были извлечены из консенсусных последовательностей и записаны как выравнивание в формате fasta для каждой популяции отдельно. Далее, из каждого гена были удалены интроны, после чего гены были проверены на наличие стартового кодона ATG и при необходимости переведены в комплементарную цепь и записаны в направлении 5’-3’, как описано в предыдущем пункте.

Полученные нуклеотидные выравнивания были поданы на вход программе, предоставленной Надеждой Потаповой, для расчета внутривидовых частот синонимичных (pS)

и несинонимичных (pN) замен. В результате были получены таблицы значений pN, pS и их соотношение для каждого гена во всех трех популяциях по отдельности.

Аннотация генома и набор референсных белок-кодирующих последовательностей *C. bursa-pastoris* несовершенны, поэтому были найдены и удалены из итоговых таблиц гены, содержащие стоп-кодоны в референсе, а также были удалены гены, имевшие pS = 0, как потенциально содержащие нераспознанные интроны или не являющиеся на самом деле кодирующими последовательностями.

3. Результаты и обсуждение

3.1. Описательные статистики для каждой популяции

Для анализа использовались результаты полногеномного секвенирования линий *C. bursa-pastoris*, культивируемых в лаборатории геномики растений ИППИ РАН (линии Ier3, Kem-wt, lel-Mk, Mar, Mich, Minsk, Murm-04, Murm-wt, ven и eng002), а также публично доступные данные из базы SRA (https://www.ncbi.nlm.nih.gov/sra). Полный список использованных в работе линий представлен в Таблице №1 Приложения.

Согласно ранее полученным результатам филогенетического анализа по данным генотипирования и пластидных геномов, каждая линия принадлежит к одной из трех популяций: европейской (EUR), среднеазиатской (ME) и дальневосточной (ASI), одна из которых (дальневосточная) имеет отличное от остальных популяций происхождение (см. Обзор литературы, раздел 1.3 «Происхождение аллотетраплоида *C. bursa-pastoris*»). Распространение анализируемых линий, а также их принадлежность к популяциям приведены на рисунке 5.

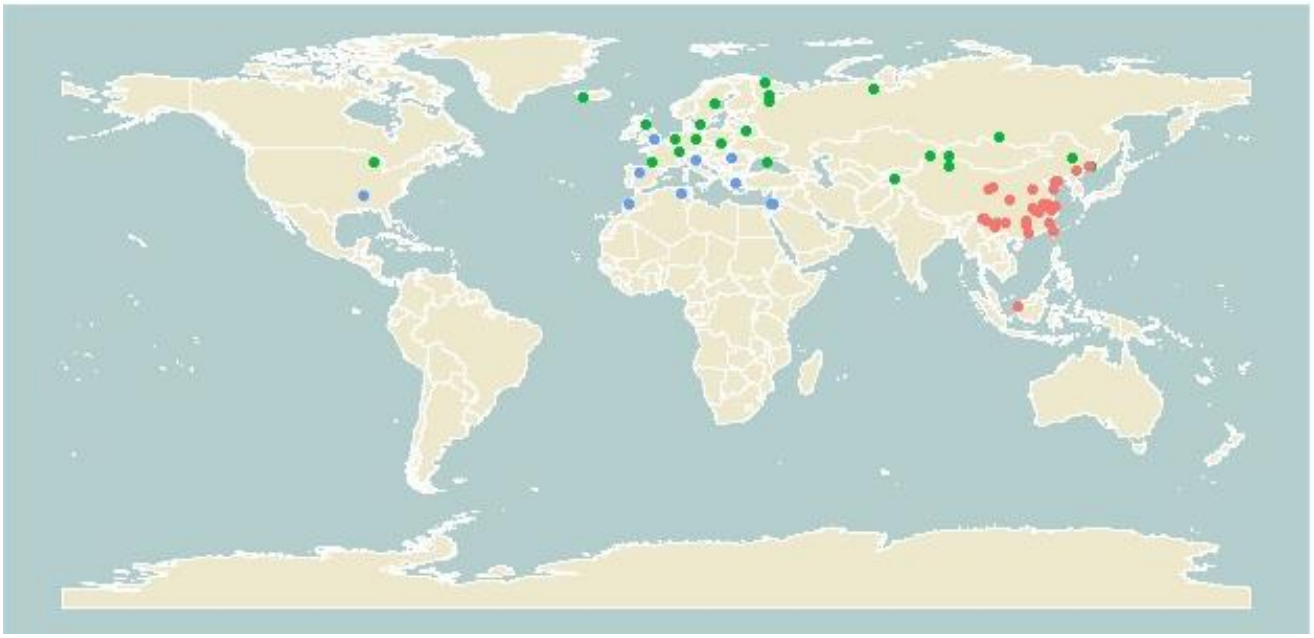


Рисунок 5. Места сбора линий *C. bursa-pastoris*, используемых в анализе. Зеленым обозначены растения, принадлежащие к европейской популяции, синим – к среднеазиатской и красным – к дальневосточной.

Вследствие различий в происхождении и недавнего полиплоидного происхождения *C. bursa-pastoris* можно ожидать различия в структуре генома, такие как: транслокации, крупные делеции и инверсии. Все описанные геномные перестройки, сопровождающие

процесс диплоидизации любого неополиплоида (Gaeta et al., 2007; Douglas et al., 2015; Renny-Byfield et al., 2014), отражаются на том, как данные полногеномного секвенирования картируются на референсный геном. Для прямого анализа таких событий необходимы независимые полногеномные сборки для каждой популяции, однако судить о них косвенно возможно по картированию: по доле выровненных на геном чтений и покрытию.

Для всех популяций наблюдается высокий уровень картирования на референсный геном. Так, медианный процент картированных на геном чтений составил 87.9% для дальневосточной, 91.6% для европейской и 90.7% ближневосточной (рисунок 6а). Однако уровень картирования значимо отличается между растениями из дальневосточной и европейской (U-тест: FDR = 0.0009) и дальневосточной и ближневосточной (U-тест: FDR = 0.0009).

Среднее покрытие на ген не отличается между популяциями, но внутри дальневосточной и европейской группы растений наблюдались выбросы с высоким покрытием, что связано с большей глубиной секвенирования этих образцов (рисунок 6б).

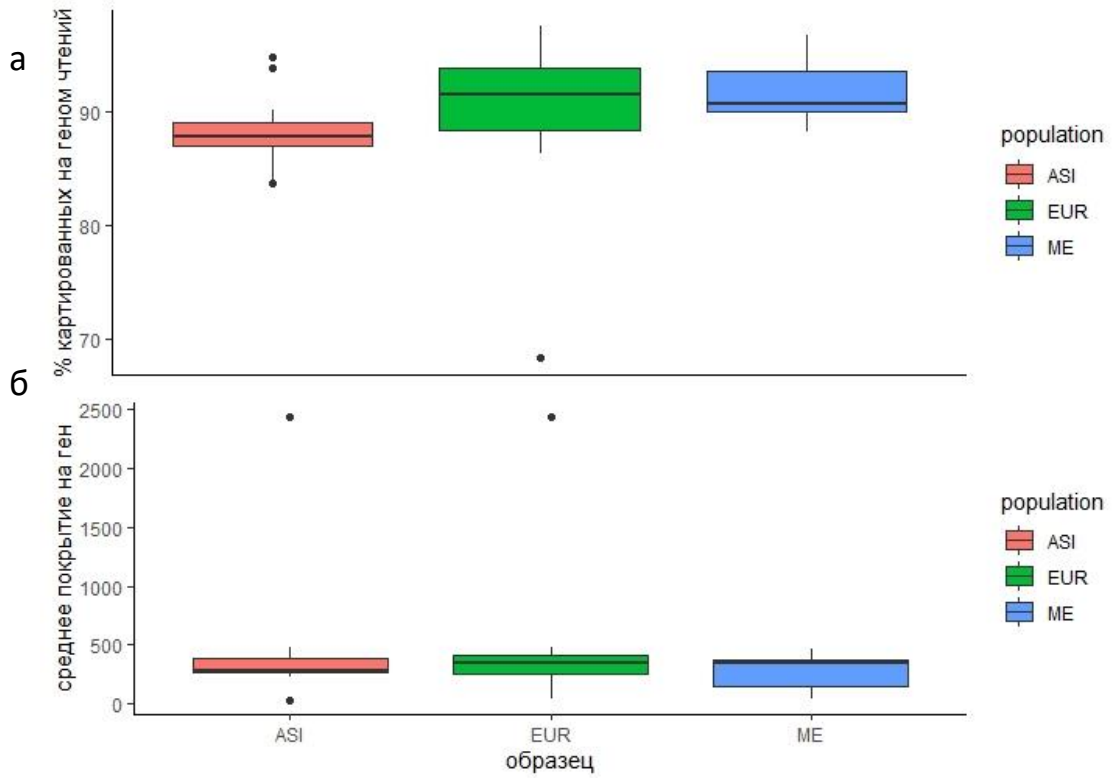


Рисунок 6. Диаграммы размаха для каждой популяции: а) процент картированных чтений на референсный геном, б) среднее покрытие на ген. Цветами обозначена принадлежность к популяции.

Изменения в структуре генома, связанные с процессом диплоидизации, отражаются на покрытии отдельных генов и целых участков хромосом. Для того, чтобы посмотреть, есть ли межпопуляционные различия в покрытии генов для каждой линии, используемой в работе, были взяты данные о покрытии каждого гена во всех, нормированные на длину гена и размер библиотеки (RPKM), затем были посчитаны корреляции Спирмена между всеми возможными парами растений, а так же было проведена иерархическая кластеризация методом WPGMA. Растения из одной популяции хорошо коррелируют между собой и кластеризуются внутри нее. Наибольшая корреляция и наименьшее расстояние наблюдается между дальневосточными растениями, тогда как европейские растения разнообразнее и демонстрируют большие расстояния внутри своей группы (рисунок 7).

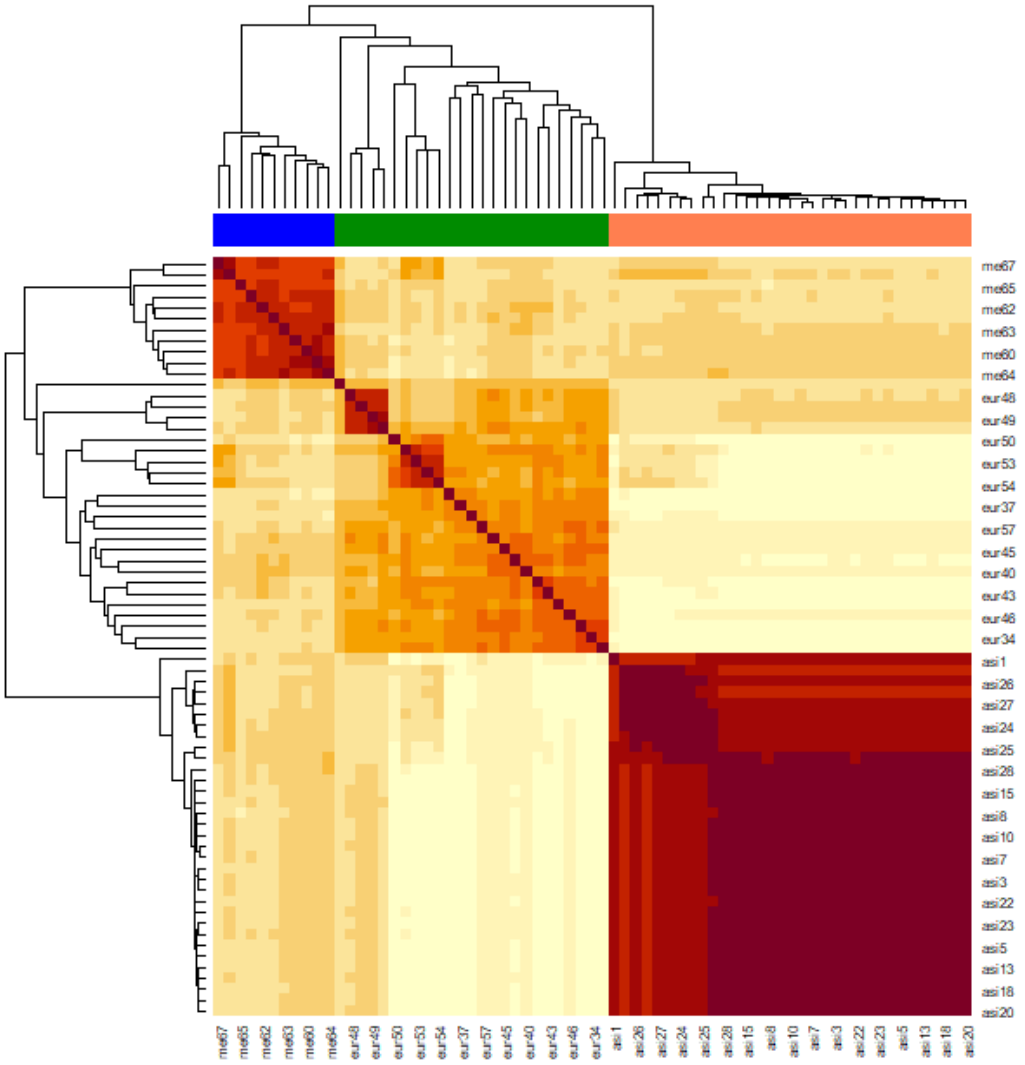


Рисунок 7. Корреляционная матрица и иерархическая кластеризация (алгоритм WPGMA) по нормированным на длину гена покрытиям. Оттенками красного цвета обозначены значения корреляции Спирмена, чем выше значение, тем более темный цвет используется.

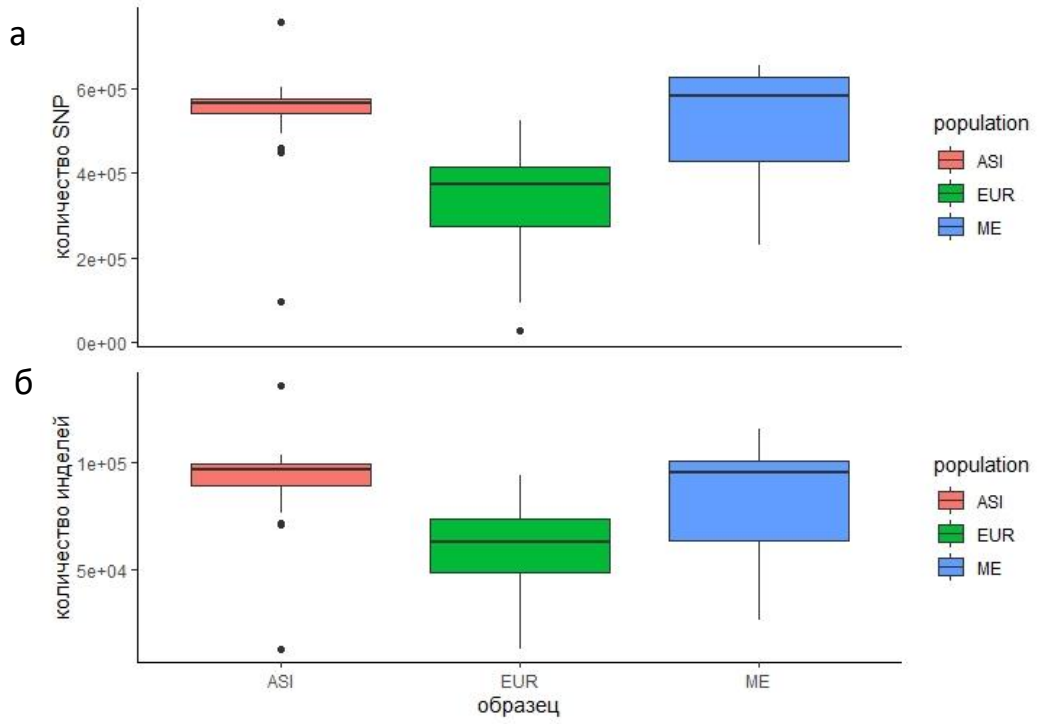


Рисунок 7. Диаграммы размаха для каждой популяции: а) количество однонуклеотидных полиморфизмов, б) количество инделей. Цветами обозначена принадлежность к популяции.

3.2.Анализ генетического смешения

Поскольку *C. bursa-pastoris* является синантропным видом, расселяющимся с помощью человека (Neuffer et al., 2010), нельзя считать ее популяции полностью географически и репродуктивно изолированными. Данные генотипирования позволяют выявить популяционную структуру и отследить генетическое смешение, произошедшее в результате скрещивания между представителями разных популяций.

Анализ генетического смешения с помощью программы admixture основывался на предположении о наличии трех передковых популяций, которые дали начало современным и вносят или не вносят вклад в каждое из анализируемых растений. Результаты анализа показали, что дальневосточная популяция является наиболее генетически обособленной и однородной, тогда как европейские растения, наоборот, демонстрируют большое разнообразие внутри популяции. Так же результаты указывают на возможность скрещиваний между растениями из разных популяций. На рисунках 8 и 9 представлен вклад предполагаемых предковых популяций в каждое из участвующих в анализе растение, а также их географическое положение.

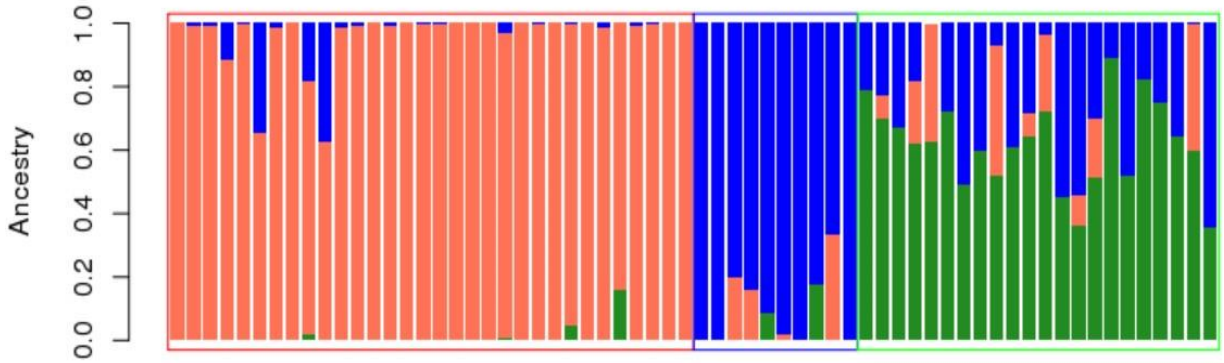


Рисунок 8. Вклад каждой из предполагаемых предковых популяций в растения, участвующие в анализе. Зеленым обозначен вклад европейской предковой популяции, синим – ближневосточной и красным – к дальневосточной.

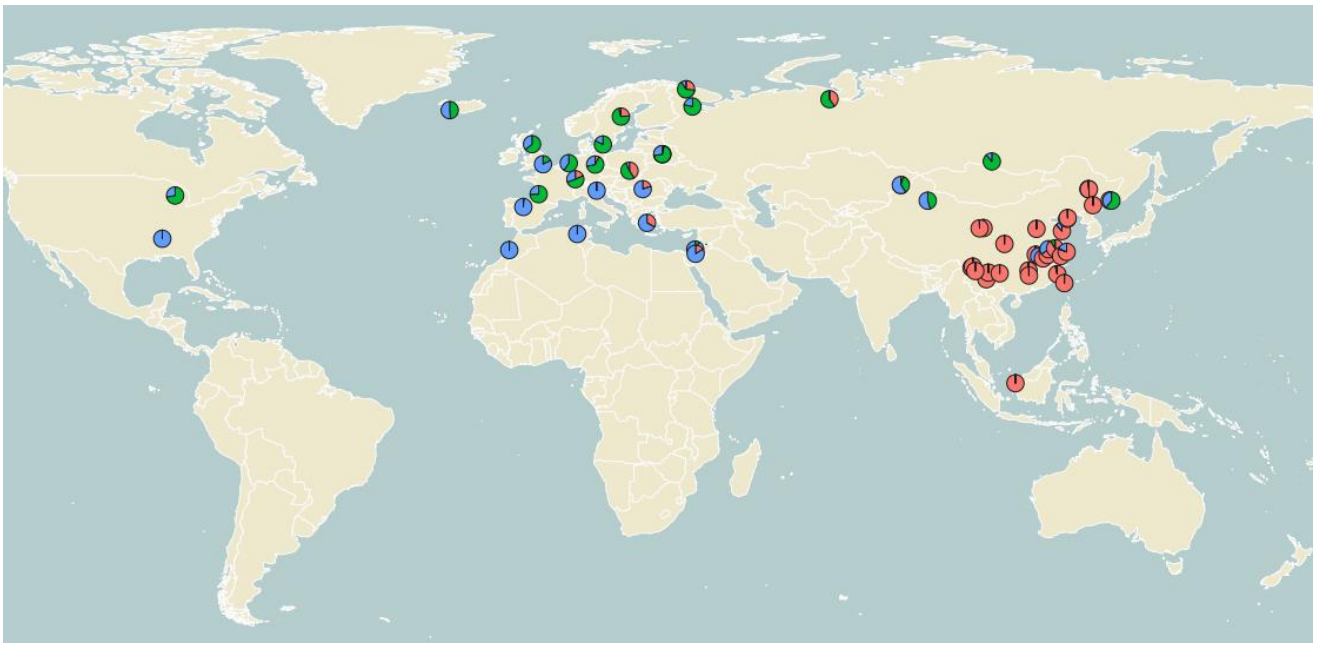


Рисунок 9. Места происхождения каждого образца. Круговыми диаграммами обозначен вклад каждой из предполагаемых предковых популяций. Зеленым обозначен вклад европейской предковой популяции, синим – ближневосточной и красным – к дальневосточной.

Таким образом, географическое положение действительно влияет на структуру популяций *C. bursa-pastoris*, однако не вызывает полную изоляцию. Наиболее интенсивный обмен аллелями наблюдается между европейской и ближневосточной популяциями, что связано как с географической близостью, историей расселения (Neuffer et al., 2010) и общим

происхождением. Схожие результаты наблюдаются и в (Cornille et al., 2016), где использовался другой тип данных (Genotyping-By-Sequencing) и *C. orientalis* в качестве референса.

3.3.Анализ генов с высоким F_{st}

Одной из часто используемых мер для оценки генетического расхождения популяций организмов одного вида является субпопуляционный индекс фиксации (F_{st}), который принимает значения от 0, означающего отсутствие дивергенции, одинаковые частоты аллелей и панмиксию, до 1, указывающей на генетическую изоляцию исследуемых популяций. Чем выше Fst, тем более изолированными друг от друга можно считать популяции: значения до 0.05 показывают небольшую дифференциацию, начиная от значения 0.25 популяции считаются значительно изолированными (Hartl and Clark, 1997).

Средние и максимальные значения F_{st}для каждой пары популяций *C. bursa-pastoris*, а также количество генов с $F_{st} > 0.25$ представлены в таблице 1. Полученные результаты указывают на слабую дифференциацию популяций, однако в каждом из сравнений присутствуют гены, показывающие высокие значения F_{st}.

Таблица 1. Средние и максимальные значение F_{st} для каждой из трех пар популяций *C.bursa-pastoris* и количество генов с $F_{st} > 0.25$.

	Среднее значение Fst	Максимальное значение Fst	Количество генов с Fst > 0.25
ASI-ME	0.049	0.49	108
ASI-EUR	0.056	0.48	112
ME-EUR	0.047	0.47	27

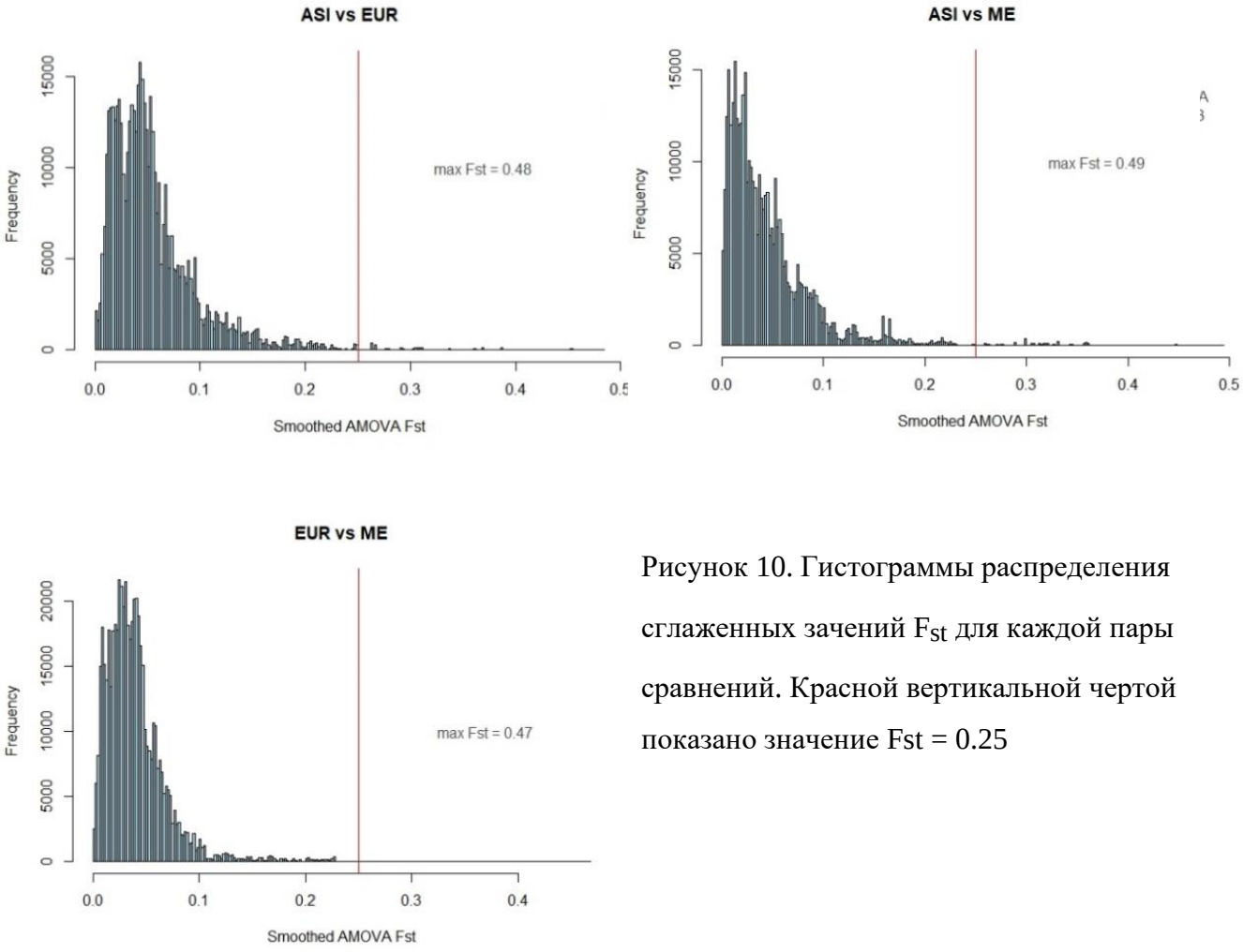


Рисунок 10. Гистограммы распределения сглаженных зачений F_{st} для каждой пары сравнений. Красной вертикальной чертой показано значение Fst = 0.25

Небольшое количество генов, имеющих F_{st} большего порогового значения 0.25, препятствует проведению анализа обогащения категориями Gene Ontology, однако в каждом сравнении были обнаружены гены, обладающие сходными функциями. В сравнении растений из европейской и дальневосточной популяции 20 генов связаны с ответом на биотические и абиотические стрессы, 18 с различными метаболическими процессами, три с организацией клеточной стенки, два гена, связанные с регуляцией фитохромов и ответом на изменение спектра видимого света, и два, участвующие в развитии органов и клеточном делении. Среди них следует отдельно отметить ортолог гена *AT5G48540.1 A. thaliana*, участвующего в ответе на каррикины – класс регуляторных веществ, выделяемых во время сгорания растения. Каррикины способствуют прорастанию семян многих видов растений, в особенности тех, которые обитают в местах частых природных пожаров (Chiwocha et al., 2009). У всех представителей дальневосточной популяции наблюдается четыре несинонимичные замены,

три из которых являются консервативными (с лейцина на валин, с аспарагина на серин и с валина на аланин) и одна радикальной (с серина на фенилаланин).

В сравнении растений из ближневосточной и дальневосточной популяций 10 генов связаны с ответом на биотические и абиотические стрессы, 24 с различными метаболическими процессами, четыре с организацией клеточной стенки и шесть генов, участвующие в развитии органов и клеточном делении. Кроме того, высоким F_{St} в сравнении ближне- и дальневосточных популяций характеризуется ген *Cbp40396* из субгенома B, имеющий 3 замены. Используя экспрессионные данные, полученные для *C. bursa-pastoris* и *A. thaliana* (Kasianov et al., 2017; Klepikova et al., 2016), было показано, что паттерн экспрессии ортолога этого гена *AT2G15020* у *A. thaliana* покрывает тычинки цветка и розеточные листья при холодовом стрессе, в то время как у *C. bursa-pastoris* гомеолог из субгенома B экспрессируется в тычинках, а гомеолог субгеноме A (*Cbp19252*) отвечает на стресс в семядолях (рисунок 11), что указывает на возможную субфункциализацию этих двух гомеологов.

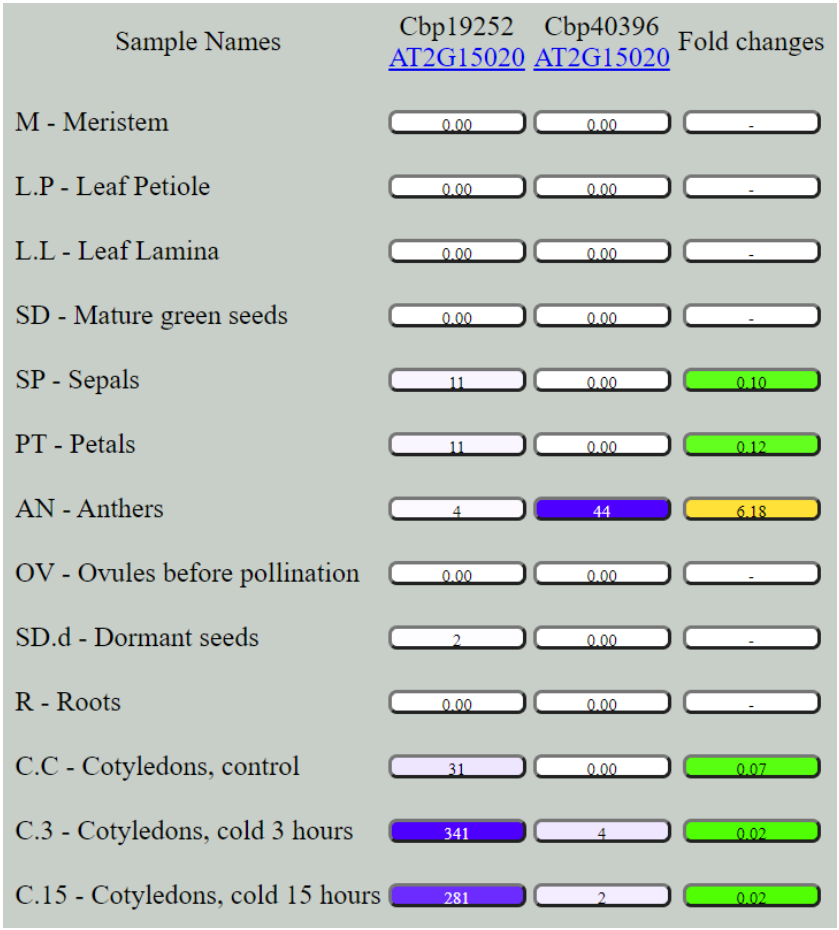


Рисунок 11. Отличия в паттерне экспрессии двух гомеологичных генов. Информация из базы TraVA (http://travadb.org/)

В сравнении растений из ближневосточной и европейской популяции 12 генов связаны с различными метаболическими процессами, один с организацией клеточной стенки и один с метилированием. Среди них следует отметить ортолог гена *AT5G42400.8 A. thaliana,* продукт которого участвует в активации гена *FLC,* регулирующего переход к цветению (Berr et al., 2009; Strange et al., 2011). Наблюдается четыре несинонимичные замены у большинства представителей европейской популяции, одна из которых консервативна (аспарагиновая кислота в глутаминовую) и три радикальные (глутаминовая кислота на лизин, глутамин в лейцин, цистеин в фенилаланин). У представителей ближневосточной популяции две несинонимичные, характерные только для них: замена тирозина в фенилаланин и неполярного изолейцина на полярный треонин.

Известны фенотипы растений *A. thaliana* с мутациями в этом гене: в одних случаях наблюдаются удлиненные листья с рассеченной листовой пластинкой, в других – позднее зацветание. Между растениями *C. bursa-pastoris* из европейской и ближневосточной популяций есть схожие фенотипические различия: у ближневосточных растений очень длинные листья с рассеченным краем, тогда как у европейский, несмотря на разнообразие в морфологии и размерах листьев, подобной длины и формы не наблюдается. Также, различается и время зацветания: согласно данным, полученным в лаборатории геномики растений ИППИ РАН, переход к цветению у европейских линий *C. bursa-pastoris* происходит на третью-четвертую неделю после прорастания, тогда как ближневосточные растения зацветают раньше, на второй неделе.

Для более подробного изучения полиморфных генов была проведена функциональная аннотация генов *C. bursa-pastoris* с помощью программы pfamScan из пакета HMMER3, относящей каждый ген к семейству на основании профилей hmm из базы PFAM. Однако не для всех генов было достоверно найдено семейство, поэтому для таких генов с использованием лучшего хита программы BLAST были найдены ортологи хорошо изученного и аннотированного растения *A. thaliana*, а затем по ним найдены семейства.

В сравнении дальневосточной популяции с европейской 18 генов не принадлежат к какому-либо из известных семейств, для одного из них известна локализация в субгеноме B (рисунок 12 а). В сравнении дальневосточной популяции с ближневосточной 26 генов не принадлежат к какому-либо из известных семейств, пять из них расположены в субгеноме B, тогда как локализация остальных 21 не известна (рисунок 12 б). В сравнении ближневосточной

и европейской популяций для восьми генов не известна принадлежность к семейству, их принадлежность к субгеному неизвестна (рисунок 12 в).

По всему геному в целом доминируют по количеству семейство F-box (Приложение № 5), которое так же является самым крупным семейством у *A. thaliana* (Gagne et al., 2002). Среди генов с высоким Fst в сравнениях растений из дальневосточной и ближневосточной, дальневосточной и европейской популяций, так же часто встречаются гены с F-box доменом, функции которых не известны. Так же в каждом из сравнений есть гены, образующие семейство мобильных ретроэлементов (Retrotransposon gag), вызывающих хромосомные перестройки (сайт EMBL-EBI по Pyret et al., 2001). Однако явной перепредставленности в семействах генов не наблюдается.

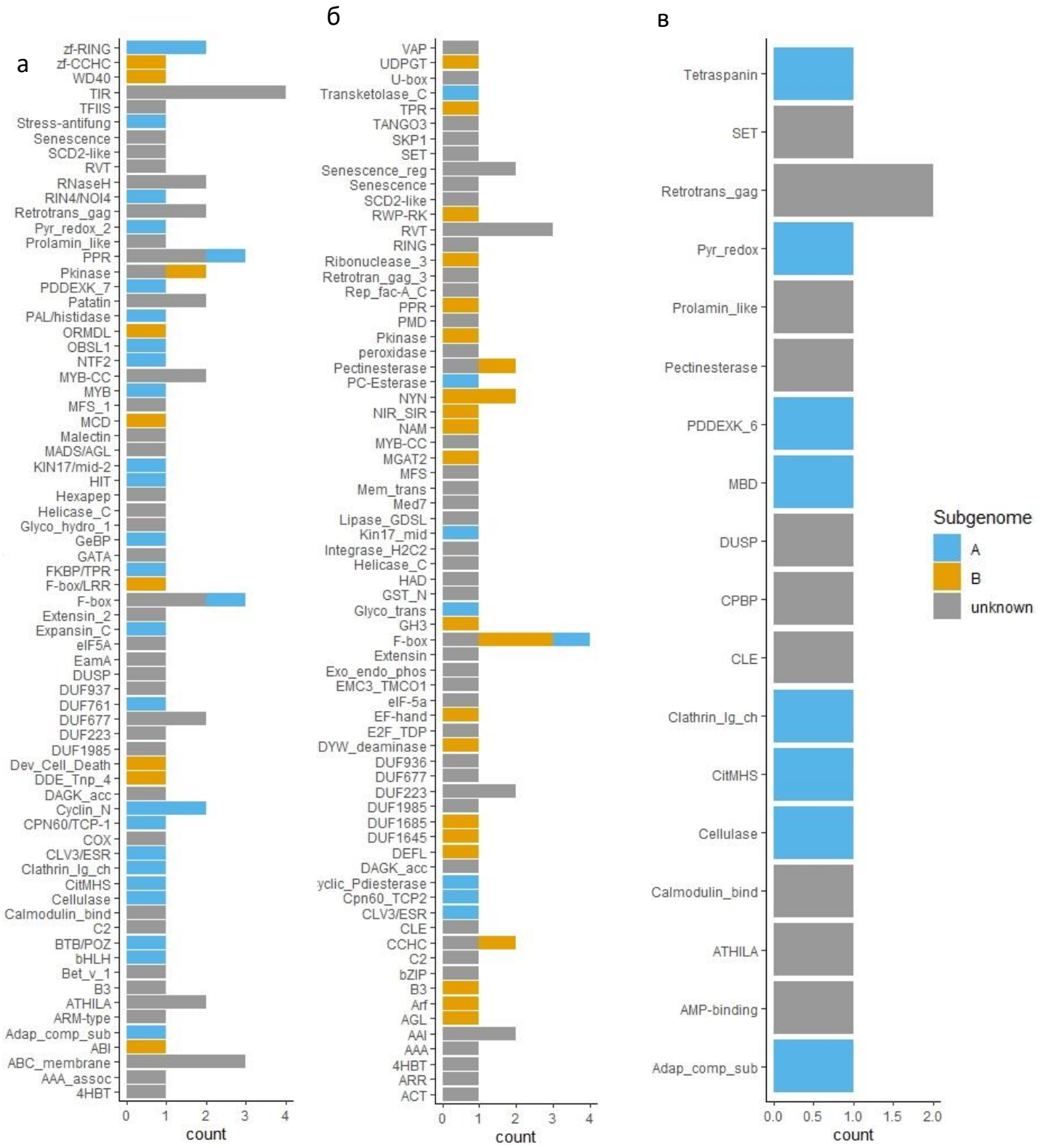


Рисунок 12. Семейства генов с высоким Fst в трех парных сравнениях популяций: а) дальневосточной с европейской, б) дальневосточной с ближневосточной, в) ближневосточной с европейской.

3.4. Положительный отбор в генах, участвующих в ответе на холодовой стресс

Среди генов, полиморфных между разными популяциями *C. bursa-pastoris*, был обнаружен ряд генов, участвующих в ответе на биотические и абиотические стрессы, что может отражать адаптацию различных линий *C. bursa-pastoris* к разнообразным условиям жизни в самых различных биомах. На протяжении своего жизненного цикла растения постоянно сталкиваются с различными биотическими и абиотическими стрессами, демонстрируя невероятные способности к адаптации (Takahashi et al., 2020). Молекулярные основы ответа, на некоторые типы стресса, а также гены, участвующие в нем и их сети взаимодействия, довольно хорошо изучены (Gong ei al., 2020). Одним из наиболее изученных на уровне генов и их взаимодействий является ответ на холодовой стресс у ближайшего родственника Пастушьей сумки *A. thaliana* (Park et al., 2015).

Существуют различные подходы для определения типа отбора по белок-кодирующим последовательностям. Одним из наиболее популярных является расчет соотношения частот несинонимичных замен (dN) к частотам синонимичных замен (dS) между изучаемыми популяциями и видами. Если соотношение dN/dS менее 1, то отбор можно считать отрицательным, если более 1 – положительным. При dN/dS = 1 считается, что отбора нет (Li et al., 1985). В настоящей работе был использован внутрипопуляционный аналог соотношения dN/dS – соотношение частот несинонимичных полиморфизмов (pN) к частотам синонимичных полиморфизмов (pS). Значения pN/pS интерпретируются так же, как и dN/dS.

Для каждой популяции *C. bursa-pastoris* были найдены гены, ортологичные 1352 генам *A. thaliana*, участвующим в ответе на холодовой стресс (Park et al., 2015). Затем, для таких генов было рассчитано соотношение частот несинонимичных полиморфизмов к частотам синонимичных полиморфизмов pN/pS. Количество генов варьировало между популяциями из-за фильтрации по стоп-кодонам и значению pS (см. пункт 2.8 в «Материалы и методы»). Всего было проанализировано 1417 генов в ближневосточной, 1600 в европейской и 1525 в дальневосточной популяциях. На рисунке 13 изображены распределения pN/pS в дальневосточной (а), европейской (в) и ближневосточной (д) популяциях. В основном, значения pN/pS невысокие, среднее значение в дальневосточной популяции составляет 0.44, в европейской – 0.33 и в ближневосточной – 0.27. Однако есть гены, в которых соотношение выше 1, то есть, наблюдается положительный отбор: 167 генов найдено для дальневосточной популяции, 99 для европейской и 83 для ближневосточной. Распределение генов по субгеномам представлено в таблице 2.

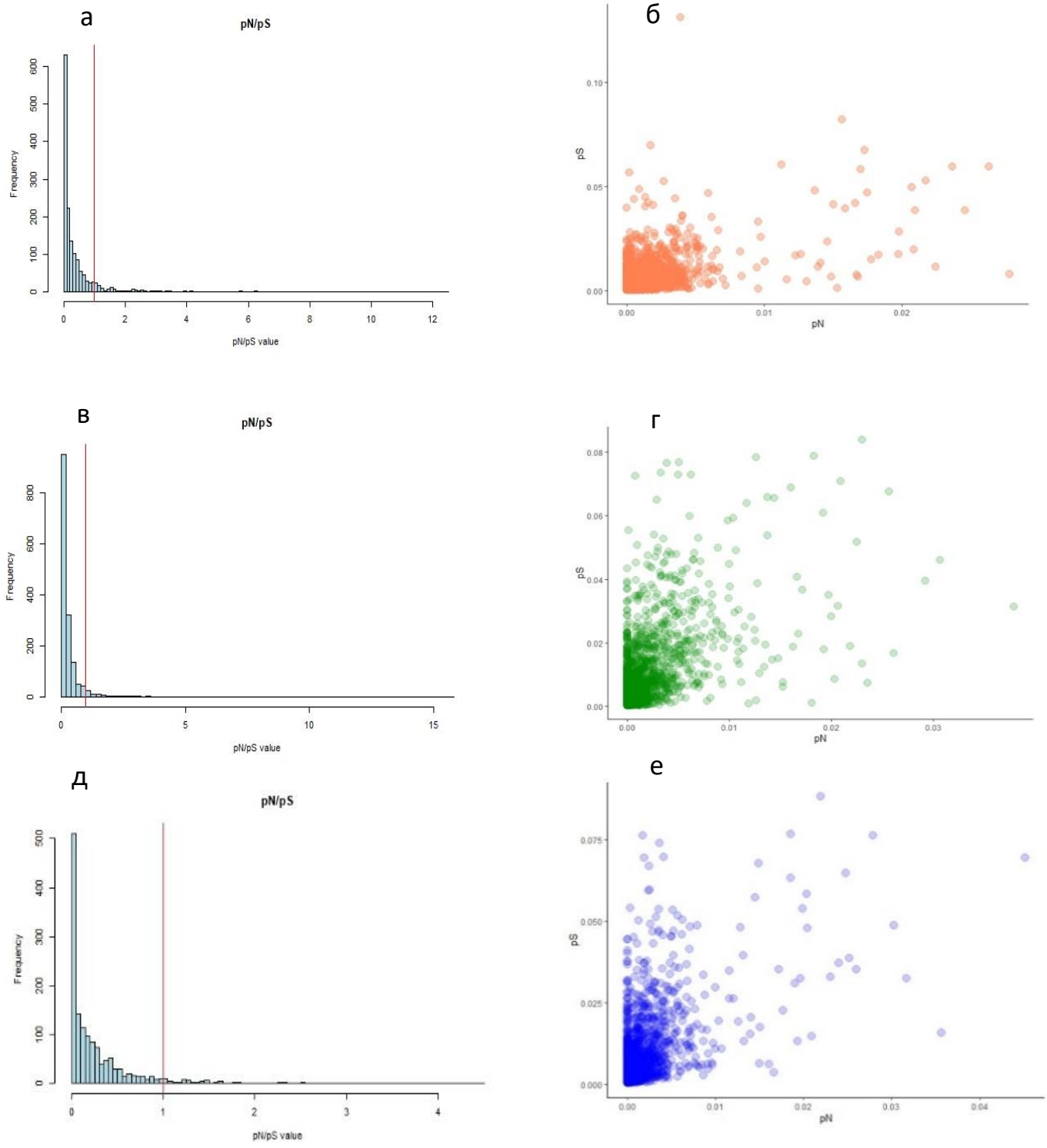


Рисунок 13. Распределения значений pN/pS и диаграммы рассеяния для дальневосточной (а, б), европейской (в, г) и ближневосточной (д, е) популяций. Также, семь генов среди найденных общие для всех популяций, но набор генов с положительным отбором по большей части являются уникальным для каждой популяции (Рисунок 14)

Таблица 2. Распределение генов по субгеномам в каждой популяции.

	Дальневосточная (ASI)	Европейская (EUR)	Ближневосточная (ME)
Субгеном A	74	49	36
Субгеном B	82	44	42
Всего генов с положительным отбором	167	99	83

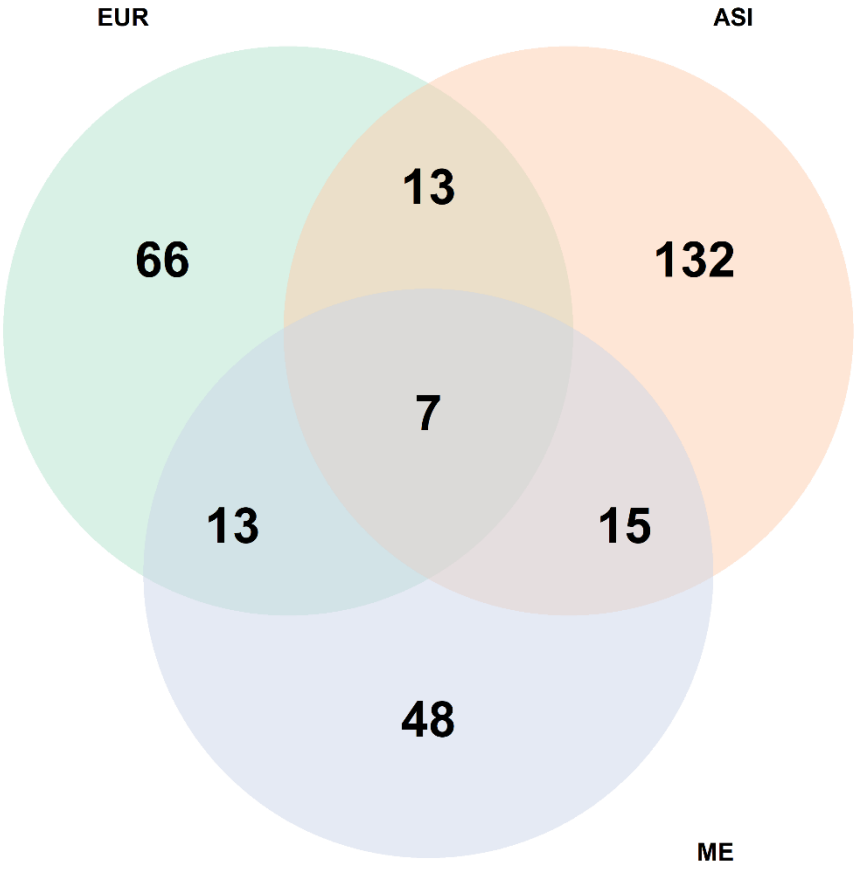


Рисунок 14. Диаграмма Эйлера для генов с положительным отбором

Небольшое количество генов с положительным отбором препятствует проведению анализа обогащения категориями Gene Ontology. Однако среди найденных генов в дальневосточной и европейской популяциях есть те, которые образуют пары гомеологов,

локализующихся в разных субгеномов. Таким образом, можно предположить, что оба гена из таких пар являются важными для растений этой популяции.

Для дальневосточной популяции таких пар найдено пять, среди них есть гены кодирующие фитохромы и участвующие в прорастании семян в ответ на световые стимулы (*Cbp18956, Cbp39111*), участвующие в ответе на действие каррикинов (см. выше) (*Cbp2567, Cbp8493*), фотосинтезе (*Cbp29038, Cbp6898*) и репарации ДНК (*Cbp41981, Cbp53339*).

В европейской популяции найдено три пары, одна из которых (*Cbp30926 и Cbp5353*) является ортологичной гену *LONGIFOLIA2* из *A. thaliana*, участвующему в регуляции развития и роста в длину листовой пластинки (Lee et al., 2006).

Для ближневосточной популяции пар не найдено, однако интересен ортолог гена *CCL* из *A. thaliana*, локализующийся в субгеноме А (*Cbp10376*) и участвующий в регуляции циркадных ритмов (Lidder et al., 2005).

Таким образом, для каждой популяции характерен свой уникальный набор генов с положительным отбором, участвующих в ответе на холодовой стресс. Такие результаты указывают на разные направления в эволюции каждой из популяций *C. bursa-pastoris*, а также могут быть причиной колоссальной экологической пластичности этого вида.

4. Выводы

1) Анализ покрытия линий *Capsella bursa-pastoris* показал, что геномы линий, принадлежащих к дальневосточной (азиатской) популяции, имеют наибольшие отличия от линий остальных популяций, что соответствует ранее полученным данным о независимом происхождении дальневосточной популяции.
2) Популяции *Capsella bursa-pastoris* географически и репродуктивно не изолированы, и между ними происходит поток генов.
3) Сравнение уровня генного полиморфизма в линиях *Capsella bursa-pastoris* позволил обнаружить гены, наиболее различающиеся между популяциями. Среди них наиболее часто встречались гены, участвующие в ответе на биотические и абиотические стрессы, в формировании клеточной стенки и метаболических процессах, что может объяснять морфологические и физиологические различия растений, принадлежащих разным популяциям.
4) Все три популяции *Capsella bursa-pastoris* имеют разные направления эволюции сети ответа на холодовой стресс, что может являться одной из причин широкого ареала вида *Capsella bursa-pastoris* и его экологической пластичности.

Список литературы.

1) Alexander DH, Lange K "Enhancements to the ADMIXTURE algorithm for individual ancestry estimation." BMC Bioinformatics 2011 https://doi.org/10.1186/1471-2105-12-246
2) Bell, C.D., Soltis, D.E. & Soltis, P.S. The age and diversification of the angiosperms-re-revisited. Am. J. Bot. 97, 1296–1303 (2010).
3) Bertioli, D. J., Jenkins, J., Clevenger, J., Dudchenko, O., Gao, D., Seijo, G., … Cameron, C. (2019). The genome sequence of segmental allotetraploid peanut Arachis hypogaea. *Nature Genetics*, *51*(May). https://doi.org/10.1038/s41588-019-0405-z
4) Berr, A., Xu, L., Gao, J., Cognat, V., Steinmetz, A., Dong, A., and Shen, W.H. (2009). SET DOMAIN GROUP 25 encodes a histone methyltransferase and is involved in FLC activation and repression of flowering. Plant Physiol., in press.
5) Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014 Aug 1;30(15):2114-20. doi: 10.1093/bioinformatics/btu170. Epub 2014 Apr 1. PMID: 24695404; PMCID: PMC4103590.
6) Bowers JE, Chapman BA, Rong JK, Paterson AH. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events, Nature , 2003, vol. 422 (pg. 433-438)
7) Brandvain Y, Slotte T, Hazzouri KM, Wright SI, Coop G. (2013) Genomic identification of founding haplotypes reveals the history of the selfing species Capsella rubella. PLoS Genet
8) Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. BMC Bioinformatics. 2009 Dec 15;10:421. doi: 10.1186/1471-2105-10-421. PMID: 20003500; PMCID: PMC2803857.
9) Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA. Stacks: an analysis tool set for population genomics. Mol Ecol. 2013;22(11):3124-3140. doi:10.1111/mec.12354
10) Chiwocha, S., Dixon, K., Flematti, G., Ghisalberti, E., Merritt, D., Nelson, D., Riseborough, J. M., Smith, S., & Stevens, J. (2009). Karrikins: A new family of plant growth regulators in smoke. Plant Science, 177(4), 252-256. https://doi.org/10.1016/j.plantsci.2009.06.007
11) Ceplitis et al. (2005). Bayesian inference of evolutionary history from chloroplast microsatellites in the cosmopolitan weed Capsella bursa - pastoris (Brassicaceae), 4221–4233. https://doi.org/10.1111/j.1365-294X.2005.02743.x
12) Christenhusz, M. J. M., & Byng, J. W. (2016). The number of known plants species in the world and its annual increase. *Phytotaxa*, *261*(3), 201–217. https://doi.org/10.11646/phytotaxa.261.3.1

13) Cornille, A., Salcedo, A., Lascoux, M., Centre, E. B., & Biology, E. (2016). Genomic signature of successful colonization of Eurasia by the allopolyploid shepherd ' s purse (Capsella bursa- pastoris), 616–629. https://doi.org/10.1111/mec.13491

14) Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R; 1000 Genomes Project Analysis Group. The variant call format and VCFtools. Bioinformatics. 2011 Aug 1;27(15):2156-8. doi: 10.1093/bioinformatics/btr330. Epub 2011 Jun 7. PMID: 21653522; PMCID: PMC3137218.

15) De Bodt S, Maere S, Van de Peer Y. Genome duplication and the origin of angiosperms, Trends Ecol Evol , 2005, vol. 20 (pg. 591-597)

16) Douglas, G. M., Gos, G., Steige, K. A., Salcedo, A., Holm, K., Josephs, E. B., … Wright, S. I. (2015). Hybrid origins and the earliest stages of diploidization in the highly successful recent polyploid *Capsella bursa-pastoris*. *Proceedings of the National Academy of Sciences*, *112*(9), 2806–2811. https://doi.org/10.1073/pnas.1412277112

17) Gaeta Robert T., J. Chris Pires, Federico Iniguez-Luy, Enrique Leon, Thomas C. Osborn. Genomic Changes in Resynthesized Brassica napus and Their Effect on Gene Expression and Phenotype. The Plant Cell Nov 2007, 19 (11) 3403-3417; DOI: 10.1105/tpc.107.054346

18) Gagne JM, Downes BP, Shiu SH, Durski AM, Vierstra RD(2002) TheF-box subunit of the SCF E3 complex is encoded by a diverse super-family of genes inArabidopsis. Proc Natl Acad Sci USA99:11519–11524

19) Gong Z, Xiong L, Shi H, Yang S, Herrera-Estrella LR, Xu G, Chao DY, Li J, Wang PY, Qin F, Li J, Ding Y, Shi Y, Wang Y, Yang Y, Guo Y, Zhu JK. Plant abiotic stress response and nutrient use efficiency. Sci China Life Sci. 2020 May;63(5):635-674. doi: 10.1007/s11427-020-1683-x. Epub 2020 Mar 31. PMID: 32246404.

20) Guo Y-L, et al. (2009) Recent speciation of Capsella rubella from Capsella grandiflora, associated with loss of self-incompatibility and an extreme bottleneck. Proc Natl Acad Sci USA106(13):5246–5251

21) Hartl, D, Clark, AG (1997). Principles of Population Genetics 3rd edn. Sinauer Associates: Sunderland.

22) HERBERT HURKA and BARBARA NEUFFER (1997). Evolutionary processes in the genus Capsella (Brassicaceae)

23) Hurka H, Neuffer B. (1991) Colonizing success in plants: genetic variation and phenotypic plasticity in life history traits in Capsella bursa-pastoris. In: Esser G, Overdieck D (eds) Modern ecology: basic and applied aspects. Elsevier, Amsterdam, pp 77–96

24) Hurka, N. Friesen, D. A. German, A. Franzke, and B. Neuffer (2012), "'Missing link' species Capsella orientalis and Capsella thracica elucidate evolution of model plant genus Capsella (Brassicaceae)," Molecular Ecology, vol. 21, no. 5, pp. 1223–1238
25) HURKA, NIKOLAI FRIESEN, DMITRY A. GERMAN, ANDREAS FRANZKE and BARBARA NEUFFER (2012). ' Missing link ' species Capsella orientalis and Capsella thracica elucidate evolution of model plant genus Capsella (Brassicaceae)
26) Jiao, Y., Wickett, N. J., Ayyampalayam, S., Chanderbali, A. S., Landherr, L., Ralph, P. E., … Depamphilis, C. W. (2011). Ancestral polyploidy in seed plants and angiosperms. *Nature*, *473*(7345), 97–100. https://doi.org/10.1038/nature09916
27) Kasianov, A. S., Klepikova, A. V., Kulakovskiy, I. V., Gerasimov, E. S., Fedotova, A. V., Besedina, E. G., … Penin, A. A. (2017). High-quality genome assembly of Capsella bursa-pastoris reveals asymmetry of regulatory elements at early stages of polyploid genome evolution. Plant Journal, 91(2), 278–291.
28) Kryvokhyzha D, Salcedo A, Eriksson MC, Duan T, Tawari N, et al. (2019) Parental legacy, demography, and admixture influenced the evolution of the two subgenomes of the tetraploid Capsella bursa-pastoris (Brassicaceae). PLOS Genetics 15(2): e1007949.
29) Lewis W.H. (1980) Polyploidy in Species Populations. In: Lewis W.H. (eds) Polyploidy. Basic Life Sciences, vol 13. Springer, Boston, MA
30) Lee, Gyung-Tae Kim, In-Jung Kim, Jeongmoo Park, Sang-Soo Kwak, Giltsu Choi, Won-Il Chung LONGIFOLIA1 and LONGIFOLIA2, two homologous genes, regulate longitudinal cell elongation in Arabidopsis Development 2006 133: 4305-4314; doi: 10.1242/dev.02604
31) Li H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v1 [q-bio.GN]
32) Lidder, Rodrigo A. Gutiérrez, Patrice A. Salomé, C. Robertson McClung, Pamela J. Green Circadian Control of Messenger RNA Stability. Association with a Sequence-Specific Messenger RNA Decay Pathway. Plant Physiology Aug 2005, 138 (4) 2374-2385; DOI: 10.1104/pp.105.060368
33) McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA, The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data, 2010 GENOME RESEARCH 20:1297-303
34) Mistry Jaina, Robert D. Finn, Sean R. Eddy, Alex Bateman, Marco Punta, Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions, Nucleic Acids Research, Volume 41, Issue 12, 1 July 2013, Page e121, https://doi.org/10.1093/nar/gkt263

35) Murat, F., Armero, A., Pont, C., Klopp, C., & Salse, J. (2017). Reconstructing the genome of the most recent common ancestor of flowering plants. *Nature Genetics*, *49*(4), 490–496.
36) Neuffer, Barbara et al. Evolutionary History of the Genus Capsella (Brassicaceae) - Capsella orientalis, New for Mongolia. Mongolian Journal of Biological Sciences, [S.l.], v. 12, n. 1-2, p. 3-18, dec. 2014. ISSN 2225-4994.
37) Novikova, P. Y., Hohmann, N., & Van de Peer, Y. (2018). Polyploid Arabidopsis species originated around recent glaciation maxima. Current Opinion in Plant Biology, 42, 8–15.
38) Ozkan H, Levy AA, Feldman M. (2001). Allopolyploidy-induced rapid genome evolution in the wheat (Aegilops–Triticum) group. Plant Cell 13: 1735–1747.
39) Park S, Lee CM, Doherty CJ, Gilmour SJ, Kim Y, Thomashow MF. Regulation of the Arabidopsis CBF regulon by a complex low-temperature regulatory network. Plant J. 2015;82:193–207.
40) Pyret, a Ty3/Gypsy retrotransposon in Magnaporthe grisea contains an extra domain between the nucleocapsid and protease domains. Nakayashiki H, Matsuo H, Chuma I, Ikeda K, Betsuyaku S, Kusaba M, Tosa Y, Mayama S. Nucleic Acids Res. 29, 4106-13, (2001)
41) Poplin Ryan, Valentin Ruano-Rubio, Mark A. DePristo, Tim J. Fennell, Mauricio O. Carneiro, Geraldine A. Van der Auwera, David E. Kling, Laura D. Gauthier, Ami Levy-Moonshine, David Roazen, Khalid Shakir, Joel Thibault, Sheila Chandran, Chris Whelan, Monkol Lek, Stacey Gabriel, Mark J. Daly, Benjamin Neale, Daniel G. MacArthur, Eric Banks. Scaling accurate genetic variant discovery to tens of thousands of samples. 2017 bioRxiv
42) Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ & Sham PC (2007) PLINK: a toolset for whole genome association and population-based linkage analysis. American Journal of Human Genetics, 81.
43) Renny-Byfield S, Gallagher JP, Grover CE, Szadkowski E, Page JT, Udall JA, et al. (2014). Ancient gene duplicates in Gossypium (Cotton) exhibit near-complete expression divergence. Genome Biol Evol. 2014;6:559–71.
44) Rice, A., Šmarda, P., Novosolov, M., Drori, M., Glick, L., Sabath, N., … Mayrose, I. (2019). The global biogeography of polyploid plants. Nature Ecology & Evolution, 3(2), 265–273. doi:10.1038/s41559-018-0787-9
45) Rice P., Longden I. and Bleasby A. EMBOSS: The European Molecular Biology Open Software Suite. Trends in Genetics. 2000 16(6):276-277
46) Schranz ME, TC Osborn, Novel flowering time variation in the resynthesized polyploid *Brassica napus*, *Journal of Heredity*, Volume 91, Issue 3, May 2000, Pages 242–246

47) Sicard Adrien and Lenhard Michael. Capsella. Quick guide. Current Biology 28, R909–R930, September 10, 2018

48) Simillion, Klaas Vandepoele, Marc C. E. Van Montagu, Marc Zabeau, Yves Van de Peer. The hidden duplication past of Arabidopsis thaliana. Proceedings of the National Academy of Sciences Oct 2002, 99 (21) 13627-13632; DOI: 10.1073/pnas.212522399

49) Slotte, T. S., Eplitis, A. C., Euffer, B. N., & Urka, H. H. (2006). THE ORIGIN OF THE TETRAPLOID C . BURSA - PASTORIS BASED ON CHLOROPLAST AND NUCLEAR DNA SEQUENCES 1, *93*(11), 1714–1724. https://doi.org/10.3732/ajb.93.11.1714

50) Slotte, T. S., Eplitis, A. C., Euffer, B. N., & Urka, H. H. (2006). THE ORIGIN OF THE TETRAPLOID C . BURSA - PASTORIS BASED ON CHLOROPLAST AND NUCLEAR DNA SEQUENCES 1, *93*(11), 1714–1724. https://doi.org/10.3732/ajb.93.11.1714

51) Strange A, Li P, Lister C, Anderson J, Warthmann N, Shindo C, Irwin J, Nordborg M, Dean C. Major-effect alleles at relatively few loci underlie distinct vernalization and flowering variation in Arabidopsis accessions. PLoS ONE. 2011;6:e19949.

52) Stebbins, G.L. (1971) Chromosomal Evolution in Higher Plants. Edward Arnold LTD, London, 87-89.

53) Takahashi, F., & Shinozaki, K. (2019). Long-distance signaling in plant stress response. Current Opinion in Plant Biology, 47, 106–111. doi:10.1016/j.pbi.2018.10.006

54) Tate JA, Ni ZF, Scheen AC, Koh J, Gilbert CA, Lefkowitz D, Chen ZJ, Soltis PS, Soltis DE. (2006). Evolution and expression of homeologous loci in Tragopogon miscellus (Asteraceae), a recent and reciprocally formed allopolyploid. Genetics 173: 1599–1611.

55) Ungerer M.C., Strakosh S.C., Zhen Y. Genome expansion in three hybrid sunflower species is associated with retrotransposon proliferation // Curr. Biol. 2006. V. 16. P. R872–R873.

56) Wu, Z., & Ma, Q. (2015). Limited variation across two chloroplast genomes with finishing chloroplast genome of Capsella grandiflora. *Mitochondrial DNA*, *00*(00), 1–2.

57) https://www.gbif.org/

58) https://www.ncbi.nlm.nih.gov/sra

59) https://www.htslib.org/

60) https://www.araport.org

61) https://www.ebi.ac.uk/

62) http://travadb.org/

Приложения

Приложение № 1. Список использованных линий со описательными статистиками картирования

линия	% картированных ридов	Среднее покрытие на ген	Количество SNP с покрытием более 6	Количество инделей с покрытием более 6	Популяция по пластидному геному	Место произрастания
Ier3	96.47	92.22913	410592	60938	ME	Israel
KBG1	94.94	29.99741	96175	13703	ASI	China
Kem_wt	96.47	68.64823	221250	33012	EUR	Russia
lel_Minsk	95.02	374.7164	394971	69383	EUR	Belarus
Mar	96.84	82.13734	403608	59876	ME	Marocco
MG3	97.57	47.76199	99025	13737	EUR	Russia
Minsk	95.68	430.4402	432260	75943	EUR	Belarus
Murm_04	97.50	68.28141	216570	31815	EUR	Russia
Murm_wt	68.33	57.2105	165254	24588	EUR	Russia
ven	94.62	392.1529	652702	115195	ME	Italy
Eng_002	88.21	47.25756	231385	27180	ME	UK
Mich	95.60	401.8151	412869	75078	EUR	USA
SRR1664898	91.50	343.5298	456836	76529	EUR	Iceland
SRR1665054	90.75	344.9171	576776	93279	ME	Spain
SRR1665062	93.30	349.1252	404203	66768	EUR	Netherlands
SRR1665063	90.70	372.6822	626351	99880	ME	Greece
SRR1665067	92.49	318.0836	414149	67715	EUR	Poland
SRR1665070	90.91	334.9788	585870	97371	ME	Italy
SRR1665071	91.62	360.444	379974	60269	EUR	Russia
SRR1665072	93.91	288.5708	369297	62736	EUR	Germany
SRR1746837	93.93	2437.56	754561	135394	ASI	China
SRR1751470	91.53	2435.218	484526	86050	EUR	Sweden
SRR6179229	87.56	255.4316	413049	71737	EUR	China
SRR6179230	86.31	259.9429	28646	71260	EUR	China
SRR6179231	88.41	345.8314	453577	79250	EUR	China
SRR6179232	90.23	256.2792	556941	95966	ASI	China
SRR6179233	87.89	288.973	579159	99923	ASI	China
SRR6179234	85.01	248.2932	540097	92905	ASI	China
SRR6179235	86.51	285.1436	576114	99692	ASI	China
SRR6179236	88.86	291.2448	601552	102965	ASI	China
SRR6179237	88.79	266.0515	566994	97714	ASI	China
SRR6179238	85.79	293.1528	582752	99895	ASI	China
SRR6179239	87.25	264.4382	570642	97419	ASI	China
SRR6179240	87.96	225.5065	539078	91900	ASI	China
SRR6179241	89.59	275.7372	576329	98402	ASI	China
SRR6179242	87.66	270.3718	563565	97105	ASI	China
SRR6179243	89.60	277.1981	572509	98946	ASI	China
SRR6179244	87.03	244.9887	543610	93759	ASI	China
SRR6179245	87.68	278.5926	570884	98847	ASI	China
SRR6179246	86.34	262.5911	553085	95839	ASI	China
SRR6179247	87.15	270.0192	563222	97300	ASI	China
SRR6179248	87.45	283.8113	574738	98733	ASI	China

SRR6179249	90.07	249.0961	548438	94653	ASI	China
SRR6179251	89.31	282.994	582668	100210	ASI	China
SRR6179252	89.35	282.1599	582351	100138	ASI	China
SRR6179253	85.82	270.7371	560537	96496	ASI	China
SRR6382382	86.64	266.6677	274928	45579	EUR	France
SRR6382383	91.00	343.5867	306478	48645	EUR	Russia
SRR6382386	90.14	333.4585	474570	71283	ME	Algeria
SRR6382387	87.36	456.2008	535910	84206	ASI	China
SRR6382388	83.72	424.2031	539894	82223	ASI	China
SRR6382389	87.84	464.4961	554177	88846	ASI	China
SRR6382390	88.86	424.0877	493196	76458	ASI	China
SRR6382391	85.94	346.5925	594185	97889	ASI	China
SRR6382392	86.50	416.2942	439613	73870	EUR	China
SRR6382393	90.98	448.3898	323998	50559	EUR	Sweden
SRR6382394	89.09	370.8854	459482	71171	ASI	China
SRR6382395	88.19	479.0833	522832	80998	ASI	China
SRR6382396	88.08	468.302	561821	89342	ASI	China
SRR6382398	88.27	379.5524	449655	71843	ASI	China
SRR6382399	90.57	407.6333	356354	55953	EUR	France
SRR6382400	87.62	477.9869	373814	58328	EUR	UK
SRR6382401	88.73	465.9124	622757	100358	ME	USA
SRR6382402	84.01	409.6228	570137	87583	ASI	Russia
SRR8904463	92.47	41.43247	93786	13557	EUR	Kyrgyztan
SRR8904464	91.80	178.2152	349895	61605	EUR	Russia
SRR8904465	92.89	441.9238	524557	93929	EUR	China
SRR8904466	90.03	363.2969	643172	112587	ME	Jordan

Приложение № 2А. Сравнения ближневосточных и дальневосточных линий. Список генов из генома А.

Ген с сигналом	Гомеолог из субгенома В	Ортолог из A. thaliana	Семейство	Несинонимичные замены (координата ME>ASI)	GO биологическая функция
Cbp10200	Cbp14569	AT3G28223.1	F-box	2 V>E, 39 V>L, 49 F>L, 55 H>Q, 94 E>Q, 128 F>V, 136 Y>F, 151 L>F, 203 M>V, 206 F>L, 208 S>N	SCF-dependent proteasomal ubiquitin-dependent protein catabolic process
Cbp10205	Cbp14577	AT3G28180.1	Glyco_trans	нет различий	cell wall organization
Cbp10206	Cbp14579	AT3G28150.1	PC-Esterase	35 L>F, 68 P>R, 121 G>S123 F>L, 207 A>E, 274 D>Y, 282 N>D, 392 N>S	xyloglucan metabolic process
Cbp10207	Cbp14580	AT5G40190.1	Cyclic_Pdiesterase	122 L>V, 158 G>A, 178 G>D	Не известна
Cbp12271	Cbp37488	AT1G55510.1	Transketolase_C	82 N>S, 211 E>Q	response to nutrient, branched-chain

Cbp12277	Cbp37486	AT1G55490.5	Cpn60_TCP2	6 T>S, 12 G>S, 48 S>P	amino acid catabolic process
Cbp12279	Cbp37483	AT1G55460.1	Kin17_mid	205 K>N, 207 D>G	response to cold, systemic acquired resistance, cell death
Cbp12655	Cbp32540	AT1G73965.1	CLV3/ESR	15 E>G, 21 F>S, 38 T>A, 64 E>K, 84 I>V	cellular response to DNA damage stimulus cell-cell signaling involved in cell fate commitment

Приложение № 2Б. Сравнения ближневосточных и дальневосточных линий. Список генов из генома B.

Ген с сигналом	Гомеолог из субгенома A	Ортолог из A. thaliana	Семейство	Несинонимичные замены (координата ME>ASI)	GO биологическая функция
Cbp19252	Cbp40396	AT2G15020.1	unknown	75 T>P, 522 F>V, 524 W>C	не известна
Cbp19254	Cbp265	AT4G07666.1	unknown	нет различий	не известна
Cbp19255	Cbp264	AT3G42160.1	Pectinesterase	93 S>F, 233 F>Y, 340 P>S	cell wall modification
Cbp19256	Cbp40399	AT2G14960.1	GH3	446 M>V, 468 A>T	response to auxin
Cbp19257	нет	AT3G21790.1	UDPGT	потерялся старт	не известна
Cbp19258	Cbp40196	AT4G20480.2	NYN	потерялся старт	regulation of gene expression
Cbp19259	нет	AT1G05400.2	NYN	нет	regulation of gene expression
Cbp19270	нет	AT2G15310.1	Arf	нет	intracellular protein transport

Cbp19272	нет	AT3G56530.1	NAM	нет	regulation of transcription, DNA-templated
Cbp19289	Cbp13229	AT2G15590.2	DUF1685	38 K>R, 41 P>A,	не известна
Cbp19291	Cbp13226	AT2G05320.1	MGAT2	176 T>P, 177 N>S, 180 R>K	oligosaccharide biosynthetic process
Cbp19292	Cbp13224	AT2G15620.1	NIR_SIR	79 S>R, 441 A>G, 437 T>S	response to nitrate
Cbp19293	Cbp13223	AT2G15630.1	PPR	2 K>R, 3 S>K, 9 F>L, 27 V>A, 35 D>E, 44 D>N, 47 L>F, 51 N>K, 62 S>Y, 65 L>F, 85 L>F, 86 G>D, 97 D>N, 108 I>V, 140 A>V, 142 A>S, 207 V>I, 309 V>E, 382 R>K, 400 I>V, 416 H>Y, 493 H>N, 590 D>E, 615 S>N, 617 S>A, 622 D>S	не известна
Cbp19294	нет	AT2G16210.1	B3	нет	не известна
Cbp19295	Cbp13221	AT2G15660.1	**AGL**	3 S>L, 86 V>I, 116 L>V, 136 N>K, 137 L>V, 167 P>Q, 170 A>P, 172 L>F, 190 E>D, 226 T>P, 239 I>M, 263 C>Y	не известна
Cbp19303	нет	AT2G16450.1	F-box	нет	не известна
Cbp19304	Cbp13213	AT2G15680.1	EF-hand	122 I>T	не известна

Cbp19305	нет	AT2G15690.1	DYW_deaminase	нет	embryo development ending in seed dormancy
Cbp19307	Cbp13211	AT2G15730.1	unknown	91 G>E	не известна
Cbp19308	Cbp46770	AT2G15760.1	DUF1645	116 P>H, 139 L>I, 184 N>H	не известна
Cbp19309	Cbp13207	AT5G45150.3	Ribonuclease_3	98 F>L, 455 D>N, 533 E>D, 537M>L, 538 C>Y, 540 V>L, 550 K>Q, 763 Q>R, 1160 K>E, 1178 L>I	RNA processing
Cbp19311	Cbp13204	AT3G44060.1	F-box	95 L>I, 113 R>S, 135 M>R	не известна
Cbp19313	нет	AT2G20465.1	DEFL	нет	defense response to fungus
Cbp19314	Cbp13203	AT2G15790.1	TPR	196 V>I,	floral meristem determinacy, vegetative phase change
Cbp29554	Cbp33495	AT2G41590.1	CCHC	31 R>G, 105W>G, 111 D>H	не известна
Cbp30288	Cbp13084	AT2G17150.6	RWP-RK	нет различий	не известна
Cbp35485	Cbp35	AT5G01920.2	Pkinase	нет различий	photosystem II stabilization
Cbp19312	нет	нет	unknown	нет	не известна
Cbp30305	нет	нет	unknown	нет	не известна

Приложение № 2В. Сравнения ближневосточных и дальневосточных линий. Список генов с неизвестной локализацей.

Ген с сигналом	Ортолог из A. thaliana	Семейство	GO биологическая функция
Cbp10193	AT3G28270.2	DUF677	cellular response to water deprivation
Cbp14251	AT5G52547.4	unknown	не известна
Cbp16136	AT2G05642.1	DUF223	не известна
Cbp18698	AT4G25070.1	SCD2-like	cytokinesis by cell plate formation
Cbp19833	AT1G79790.2	HAD	не известна
Cbp20823	AT1G52950.1	Rep_fac-A_C	DNA repair
Cbp20826	AT3G43837.1	RVT	Reverse transcriptase
Cbp20827	AT2G41590.1	E2F_TDP	regulation of cell cycle
Cbp20830	AT3G47290.2	C2	lipid catabolic process
Cbp22859	AT2G31920.1	DUF936	не известна
Cbp24432	AT5G48950.1	4HBT	phylloquinone biosynthetic process
Cbp24464	AT4G35985.2	Senescence	не известна
Cbp24808	AT5G62920.1	ARR	response to cytokinin
Cbp27189	AT1G26630.1	eIF-5a	defense response to bacterium
Cbp28736	AT5G54110.1	VAP	response to osmotic stress

Cbp28737	AT5G54130.2	Exo_endo_phos	не известна
Cbp29911	AT4G07666.1	unknown	не известна
Cbp32241	AT4G12590.1	EMC3_TMCO1	не известна
Cbp32242	AT1G50820.2	PMD	не известна
Cbp33350	AT1G34070.1	Retrotran_gag_3	не известна
Cbp33351	AT1G15680.1	F-box	не известна
Cbp3397	AT5G42150.1	GST_N	cell redox homeostasis
Cbp3398	AT5G42170.1	Lipase_GDSL	lipid catabolic process
Cbp3399	AT5G42180.1	peroxidase	response to oxidative stress
Cbp3400	AT5G42190.1	SKP1	embryo development ending in seed dormancy
Cbp3401	AT3G42160.1	Pectinesterase	cell wall modification
Cbp3402	AT4G23160.3	RVT	не известна
Cbp3413	AT5G34870.1	CCHC	не известна
Cbp3414	AT4G09490.1	RVT	не известна
Cbp3415	ATMG00860.1	Integrase_H2C2	не известна
Cbp3420	AT5G42200.1	RING	cellular response to hypoxia
Cbp3421	AT3G32960.1	DUF1985	не известна
Cbp34649	AT5G03500.6	Med7	regulation of transcription by RNA polymerase II

Cbp34650	AT5G03230.1	Senescence_reg	не известна
Cbp34651	AT3G52590.1	Senescence_reg	не известна
Cbp34888	AT3G19380.1	U-box	response to chitin
Cbp34906	AT4G34200.1	ACT	embryo development ending in seed dormancy, pollen development
Cbp38973	AT4G30340.2	DAGK_acc	defense response
Cbp40368	AT4G22517.1	AAI	не известна
Cbp40370	AT4G22490.1	AAI	не известна
Cbp41578	AT4G10970.7	unknown	не известна
Cbp42017	AT1G20680.2	TANGO3	не известна
Cbp44117	AT5G42400.8	SET	methylation
Cbp44126	AT5G03110.1	unknown	не известна
Cbp4715	AT5G42210.2	MFS	не известна
Cbp4717	AT1G52950.1	DUF223	DNA repair
Cbp47680	AT3G06480.1	Helicase_C	rRNA processing
Cbp52214	AT1G73590.1	Mem_trans	auxin polar transport, flower development
Cbp52352	AT3G54590.2	Extensin	plant-type cell wall organization
Cbp52362	AT5G29000.2	MYB-CC	не известна

Cbp8498	AT5G03340.1	AAA	cell division
Cbp9017	AT4G06598.2	bZIP	не известна
Cbp942	AT5G52547.4	unknown	не известна
Cbp12646	нет	unknown	не известна
Cbp19683	#Н/Д	CLE	Phytohormone action.signalling peptides.NCRP
Cbp33354	нет	unknown	не известна
Cbp3416	нет	unknown	не известна
Cbp3417	нет	unknown	не известна
Cbp3418	нет	unknown	не известна
Cbp3419	нет	unknown	не известна
Cbp34634	нет	unknown	не известна
Cbp34655	нет	unknown	не известна
Cbp45615	нет	unknown	не известна
Cbp4582	нет	unknown	не известна
Cbp4713	нет	unknown	не известна
Cbp47665	нет	unknown	не известна
Cbp48435	нет	unknown	не известна
Cbp48780	нет	unknown	не известна

Cbp8172	нет	unknown	не известна
Cbp8454	нет	unknown	не известна

Приложение № 3А. Сравнения дальневосточных и европейских линий. Список генов из генома A.

Ген с сигналом	Гомеолог из субгенома B	Ортолог из A. thaliana	Семейство	Несинонимичные замены (координата EUR>ASI)	GO биологическая функция
Cbp12277	Cbp37486	AT1G55490.5	CPN60/TCP-1	48 S>P	response to cold, systemic acquired resistance
Cbp12279	Cbp37483	AT1G55460.1	KIN17/mid-2	208 D>G, 205 K>N	cellular response to copper ion starvation
Cbp12654	Cbp32541	AT1G73970.1	OBSL1	29 K>M, 37 R>K, 97 A>E, 101 L>I, 161 G>R, 229 V>F, 462 T>I, 541 L>I, 608 K>R, 619 V>A, 798 L>R	не известна
Cbp12655	Cbp32540	AT1G73965.1	CLV3/ESR	различий нет	cell-cell signaling involved in cell fate commitment
Cbp24982	Cbp8122	AT5G48510.1	BTB/POZ	171 E>V, T>M	protein ubiquitination
Cbp24987	Cbp8116	AT3G10120.1	PAL/histidase	154 D>S, 168 H>P	L-phenylalanine catabolic process
Cbp24988	Cbp8115	AT5G48540.1	Stress-antifung	91 S>F, 253 N>S, 254 V>A	response to karrikin

Cbp24989	Cbp8114	AT5G48545.1	HIT	20 P>L, 162 E>Q	sulfur compound metabolic process
Cbp24990	Cbp8113	AT5G48550.1	F-box	5 Q>R, 19 F>S, 25 V>I, 29 H>R	не известна
Cbp24991	Cbp8112	AT5G48560.1	bHLH	277 L>S, 422 L>F,	response to blue light
Cbp24992	Cbp8110	AT5G48570.1	FKBP/TPR	10 T>K, 14 T>I, 25 V>A, 65 L>I, 101 E>D, 116 V>E, 117 D>A	cellular heat acclimation
Cbp24993	Cbp8108	AT5G48590.1	DUF761	205 A>T, 291 S>N	не известна
Cbp24994	Cbp8107	AT5G48610.5	MYB	84 E>D, 297 E>A, 362 T>M, 432 S>I	не известна
Cbp24995	Cbp45911	AT5G48630.1	Cyclin_N	11 Y>C, 243 H>N	cell division
Cbp24996	Cbp8103	AT5G48640.1	Cyclin_N	различий нет	cell division
Cbp24997	Cbp8102	AT5G48650.1	NTF2	384 V>L	negative regulation of defense response to bacterium
Cbp24998	Cbp8101	AT5G48655.6	zf-RING	различий нет	response to chitin
Cbp24999	Cbp8100	AT5G48657.3	RIN4/NOI4	5 R>S, 47 N>I, 130 V>I,	innate immune response-activating signal transduction
Cbp47641	Cbp33188	AT1G49810.1	CitMHS	3 F>L, 52 N>S, 271 K>Q	sodium ion transport
Cbp521	Cbp39165	AT2G20650.2	zf-RING	220 R>G	plant-type cell wall modification

Cbp523	Cbp39167	AT2G20670.1	PDDEXK_7	245 G>S	не известна
Cbp524	Cbp39168	AT2G20680.1	Cellulase	175 S>P, 400 S>P	mannan metabolic process
Cbp526	Cbp39171	AT2G20720.1	PPR	различий нет	не известна
Cbp532	Cbp39176	AT1G66420.1	GeBP	7 S>N, 39 W>R, 41 E>K, 43 D>A, 63 E>K, 68 D>N, 80 I>T, 89 V>F, 147 A>V, 159 T>I, 173 R>G, 215 Y>S, 230 F>I	regulation of transcription, DNA-templated
Cbp533	Cbp39178	AT2G20750.1	Expansin_C	9 L>P, 23 C>S, 34 R>H, 51 P>A	cell wall organization
Cbp534	Cbp39179	AT2G20760.1	Clathrin_lg_ch	65 G>A, 266 A>V	clathrin-dependent endocytosis
Cbp537	Cbp39182	AT2G20790.1	Adap_comp_sub	52 I>T, 100 C>F, 128 A>T, 135 Y>F, 316 K>R, 510 I>S	endosomal transport
Cbp538	Cbp39183	AT2G20800.1	Pyr_redox_2	8 Q>K, 439 K>N, 457 L>F	oxidation-reduction process

Приложение № 3Б. Сравнения дальневосточных и европейских линий. Список генов из генома B.

Ген с сигналом	Гомеолог из субгенома A	Ортолог из A. thaliana	Семейство	Несинонимичные замены (координата EUR>ASI)	GO биологическая функция
Cbp26116	Cbp22370	AT5G59960.3	unknown	оба не функциональны	не известна
Cbp29554	Cbp33495	AT2G41590.1	zf-CCHC	нет различий	не известна
Cbp48466	нет	AT5G41840.1	F-box/LRR	нет	не известна
Cbp48467	Cbp33477	AT5G41980.1	DDE_Tnp_4	нет различий	не известна
Cbp48468	Cbp33478	AT5G41990.1	Pkinase	нет различий	protein kinase activity
Cbp48469	Cbp33479	AT5G42000.1	ORMDL	нет различий	defense response to bacterium, leaf senescence, response to oxidative stress

Cbp48470	Cbp33480	AT5G42010.1	WD40	93 Y>N	не известна
Cbp48472	нет	AT4G04320.2	MCD	нет	fatty acid biosynthetic process
Cbp48477	нет	AT5G42030.1	ABI	нет	regulation of actin and microtubule organization
Cbp48478	Cbp33488	AT5G42050.1	Dev_Cell_Death	нет различий	cellular response to hypoxia

Приложение № 3В. Сравнения европейских и дальневосточных линий. Список генов с неизвестной локализацией.

Ген с сигналом	Ортолог из A. thaliana	Семейство	GO биологическая функция
Cbp10179	AT3G28415.2	ABC_membrane	transmembrane transport
Cbp10181	AT3G28390.1	ABC_membrane	transmembrane transport
Cbp10183	AT3G28370.4	unknown	не известна
Cbp10184	AT3G28345.1	ABC_membrane	transmembrane transport
Cbp10189	AT3G28270.2	DUF677	cellular response to water deprivation
Cbp10193	AT3G28270.2	DUF677	cellular response to water deprivation
Cbp10194	AT2G18750.3	Calmodulin_bind	regulation of salicylic acid biosynthetic process
Cbp14251	AT5G52547.4	unknown	не известна
Cbp16136	AT2G05642.1	DUF223	не известна
Cbp16443	AT1G21890.1	EamA	не известна
Cbp17045	AT1G73400.1	PPR	не известна
Cbp18698	AT4G25070.1	SCD2-like	cytokinesis by cell plate formation
Cbp20828	AT2G26560.1	Patatin	cellular response to hypoxia, defense response to virus, plant-type hypersensitive response, response to cadmium ion
Cbp20829	AT2G26560.1	Patatin	cellular response to hypoxia, defense response to virus, plant-type hypersensitive response, response to cadmium ion
Cbp20830	AT3G47290.2	C2	intracellular signal transduction
Cbp22859	AT2G31920.1	DUF937	не известна
Cbp24432	AT5G48950.1	4HBT	phylloquinone biosynthetic process
Cbp24464	AT4G35985.2	Senescence	response to cold, response to oomycetes, response to salt stress
Cbp26884	AT4G36140.2	TIR	defense response
Cbp26886	AT5G45200.1	TIR	signal transduction
Cbp26887	AT5G45210.4	TIR	defense response
Cbp27189	AT1G26630.1	eIF5A	defense response to bacterium, response to cadmium ion, wounding
Cbp31538	AT1G57775.1	Prolamin_like	не известна
Cbp33353	AT2G07240.1	DUF1985	не известна
Cbp34377	AT1G55920.1	Hexapep	cellular response to sulfate starvation, response to cold
Cbp34832	AT1G53440.1	Malectin	protein autophosphorylation
Cbp35427	AT1G47350.1	Retrotrans_gag	не известна
Cbp37775	AT1G24010.1	Bet_v_1	defense response
Cbp38972	AT1G65150.2	F-box	не известна
Cbp38973	AT4G30340.2	DAGK_acc	leaf development, root development
Cbp41940	AT1G49150.1	unknown	не известна

Cbp41957	AT5G07940.5	Pkinase	не известна
Cbp45606	AT2G16210.2	B3	не известна
Cbp46278	AT1G47600.1	Glyco_hydro_1	glucosinolate metabolic process
Cbp46819	AT2G40930.2	DUSP	ubiquitin-dependent protein catabolic process
Cbp4711	AT5G42210.1	MFS_1	не известна
Cbp47631	AT4G14180.1	ARM-type	meiotic DNA double-strand break formation
Cbp47659	AT5G25490.2	TFIIS	transcription, DNA-templated
Cbp47680	AT3G06480.1	Helicase_C	не известна
Cbp52352	AT3G54590.2	Extensin_2	plant-type cell wall organization
Cbp52354	AT3G28510.1	AAA_assoc	не известна
Cbp52362	AT5G29000.2	MYB-CC	не известна
Cbp600	AT4G29090.1	RVT	не известна
Cbp6960	AT3G30520.1	MYB-CC	не известна
Cbp8166	AT1G67020.1	TIR	defense response
Cbp19246	нет	unknown	не известна
Cbp20592	нет	ATHILA	не известна
Cbp20826	нет	RNaseH	не известна
Cbp20827	нет	unknown	не известна
Cbp23583	нет	COX	Cellular respiration
Cbp23585	нет	MADS/AGL	не известна
Cbp25794	нет	RNaseH	не известна
Cbp26885	нет	unknown	не известна
Cbp29603	нет	F-box	не известна
Cbp29617	нет	unknown	не известна
Cbp32478	нет	unknown	не известна
Cbp33354	нет	unknown	не известна
Cbp34371	нет	unknown	не известна
Cbp34381	нет	unknown	не известна
Cbp36424	нет	unknown	не известна
Cbp38971	нет	unknown	не известна
Cbp43421	нет	Retrotrans_gag	не известна
Cbp46096	AT1G10270	PPR	cell division, embryo development ending in seed dormancy
Cbp4713	нет	unknown	не известна
Cbp6949	нет	unknown	не известна
Cbp7	нет	unknown	не известна
Cbp77	нет	GATA	не известна
Cbp8	нет	unknown	не известна
Cbp8174	нет	ATHILA	не известна

Приложение № 4А. Сравнения ближневосточных и европейских линий. Список генов из генома A.

Ген с сигналом	Гомеолог из	Ортолог из A. thaliana	Семейство	Несинонимичные замены (координата ME>EUR)	GO биологическая функция

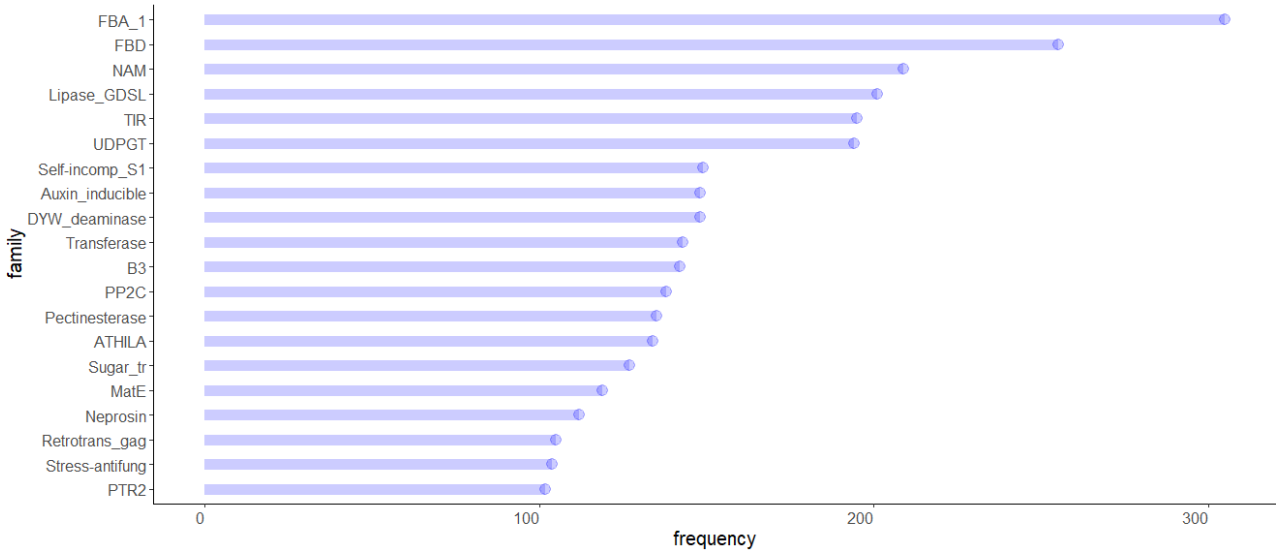
	субгенома B				
Cbp47641	Cbp33188	AT1G49810.1	CitMHS	3 L>F	sodium ion transport
Cbp523	Cbp39167	AT2G20670.1	PDDEXK_6	нонсенс в A у всех	не известна
Cbp524	Cbp39168	AT2G20680.1	Cellulase	91 D>G	mannan metabolic process
Cbp530	Cbp39174	AT2G20740.1	Tetraspanin	нет различий	не известна
Cbp534	Cbp39179	AT2G20760.1	Clathrin_lg_ch	65 A>G	clathrin-dependent endocytosis
Cbp537	Cbp39182	AT2G20790.1	Adap_comp_sub	52 T>I, 128 T>A, 135 F>Y, 316 R>K, 510 S>I	endosomal transport
Cbp538	Cbp39183	AT2G20800.1	Pyr_redox	8 K>Q, 201 S>I, 457 F>L	oxidation-reduction process
Cbp544	Cbp39189	AT4G00416.1	MBD	нет различий	methyl-CpG binding

Приложение № 4Б. Сравнения ближневосточных и европейских линий. Список генов с неизвестной локализацией.

Ген с сигналом	Ортолог из A. thaliana	Семейство	GO биологическая функция
Cbp10194	AT2G18750.3	Calmodulin_bind	regulation of salicylic acid biosynthetic process
Cbp21221	AT1G47350.1	Retrotrans_gag	не известна
Cbp31538	AT1G57775.1	Prolamin_like	не известна
Cbp35427	AT1G47350.1	Retrotrans_gag	

Cbp40298	AT3G43270.1	Pectinesterase	cell wall modification, pectin catabolic process
Cbp41957	AT5G07940.5	unknown	не известна
Cbp44117	AT5G42400.8	SET	methylation
Cbp46819	AT2G40930.2	DUSP	ubiquitin-dependent protein catabolic process
Cbp4883	AT5G16370.1	AMP-binding	fatty acid metabolic process
Cbp53498	AT5G24165.2	unknown	не известна
Cbp109	#Н/Д	unknown	не известна
Cbp13957	#Н/Д	unknown	не известна
Cbp19683	#Н/Д	CLE	Phytohormone action.signalling peptides.NCRP (non-cysteine-rich-peptide) category.CLE-peptide activity.CLE-precursor protein
Cbp20592	#Н/Д	ATHILA	не известна
Cbp20807	#Н/Д	unknown	не известна
Cbp29603	#Н/Д	unknown	не известна
Cbp29617	#Н/Д	unknown	не известна
Cbp35843	#Н/Д	unknown	не известна
Cbp526	AT2G20725.1	CPBP	CAAX-box protein processing

Приложение № 5. Топ-20 наиболее часто встречающихся семейств генов в геноме *C. bursa-pastoris*


FBA_1
FBD
NAM
Lipase_GDSL
TIR
UDPGT
Self-incomp_S1
Auxin_inducible
DYW_deaminase
Transferase
B3
PP2C
Pectinesterase
ATHILA
Sugar_tr
MatE
Neprosin
Retrotrans_gag
Stress-antifung
PTR2
family
0
100
200
300
frequency