Uniwersytet Rolniczy im. H. Kołłątaja w Krakowie
Wydział Biotechnologii i Ogrodnictwa

Katarzyna Daniela Stelmach

Analiza dystrybucji transpozonów *DcSto* w genomie marchwi oraz wykorzystanie polimorfizmów insercji do analizy struktury genetycznej populacji marchwi uprawnej (*Daucus carota* subsp. *sativus*)

Praca doktorska

Praca wykonana pod kierunkiem
prof. dra hab. inż. Dariusza Grzebelusa

Kraków, maj 2021

Karta dyplomowa

Katarzyna Stelmach
/ Imię i nazwisko autora pracy /

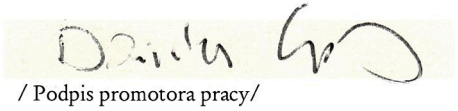
Dariusz Grzebelus
/ Imię i nazwisko promotora pracy /

Wydział Biotechnologii i Ogrodnictwa
/ Wydział - kierunek studiów /

Katedra Biologii Roślin i Biotechnologii
/ Katedra / Instytut /

Doktor
/ Nadawany stopień /

Tytuł pracy w języku polskim	Analiza dystrybucji transpozonów DcSto w genomie marchwi oraz wykorzystanie polimorfizmów insercji do analizy struktury genetycznej populacji marchwi uprawnej (*Daucus carota* subsp. *sativus*)
Słowa kluczowe /maksymalnie 5 słów /	marchew, markery molekularne, transpozony, SNP, zmienność genetyczna
Streszczenie pracy / maksymalnie 1200 znaków /	Celem badań prowadzonych w ramach rozprawy doktorskiej była analiza dystrybucji ruchomych elementów genetycznych *Stowaway*-like w genomie marchwi. Badania koncentrowały się na możliwości wykorzystania zidentyfikowanych insercji transpozonów typu *DcSto* zlokalizowanych w obrębie intronów do opracowania panelu markerów molekularnych typu ILP. Opracowany zestaw markerów posłużył za narzędzie do analizy struktury zmienności genetycznej marchwi uprawnej typu zachodniego.
Tytuł pracy w języku angielskim	The analysis of the distribution of *DcSto* transposable elements in the carrot genome and the use of insertional polymorphism for the assessment of the genetic structure of carrot (*Daucus carota* subsp. *sativus*)
Słowa kluczowe /maksymalnie 5 słów /	carrot, molecular markers, transposons, SNP, genetic diversity
Streszczenie pracy / maksymalnie 1200 znaków /	The aim of the conducted research was to understand and analyse the distribution of *DcSto* transposable elements in the carrot genome. The research focused on the possibility of developing a panel of ILP molecular markers based on the polymorphism of DcSto insertions localised within introns of the carrot genes. The developed panel of markers was then used to evaluate the structure of genetic variability present in the collection of open-pollinated cultivars representing the western carrot gene pool.



/ Podpis promotora pracy/

Oświadczenie autora pracy doktorskiej *Załącznik nr 2a*

Ja, niżej podpisany/-a:
Katarzyna Stelmach
/ Imię i nazwisko /

-
/ Numer albumu /

autor pracy doktorskiej pt.:
Analiza dystrybucji transpozonów *DcSto* w genomie marchwi oraz wykorzystanie polimorfizmów insercji do analizy struktury genetycznej populacji marchwi uprawnej (*Daucus carota* ssp. *sativus*)
/ Tytuł pracy /

Prowadzonego przewodu doktorskiego w Uniwersytecie Rolniczym im. Hugona Kołłątaja w Krakowie
Wydział Biotechnologii i Ogrodnictwa
/ Wydział /

Oświadczam, że ww. praca dyplomowa:

- została przygotowana przeze mnie samodzielnie[1],
- nie narusza praw autorskich w rozumieniu ustawy z dnia 4 lutego 1994 r. o prawie autorskim i prawach pokrewnych (Dz.U.2018. poz. 1191 t.j. z dnia 21.06.2018) oraz dóbr osobistych chronionych prawem cywilnym,
- nie zawiera danych i informacji, które uzyskałem/-am w sposób niedozwolony,

1. Oświadczam również, że treść pracy dyplomowej zamieszczonej przeze mnie w Archiwum Prac Dyplomowych jest identyczna z treścią zawartą w wydrukowanej wersji pracy.
2. W związku z realizowaniem przez Uniwersytet Rolniczy im. Hugona Kołłątaja w Krakowie zadań ustawowych i statutowych, szczególnie w zakresie prowadzenia działalności dydaktycznej i naukowo-badawczej upoważniam Uniwersytet Rolniczy im. Hugona Kołłątaja do archiwizowania i przechowywania w/w pracy utrwalonej w postaci tradycyjnej (papierowej) i elektronicznej - zgodnie z ustawą - prawo o szkolnictwie wyższym i przepisami wykonawczymi do tej ustawy, ustawą o narodowym zasobie archiwalnym i archiwach oraz ustawą o prawie autorskim i prawach pokrewnych.

Jestem świadomy/-a odpowiedzialności karnej za złożenie fałszywego oświadczenia.

Kraków, dn. 26.05.2021 r.
/ Miejsce i data/

Katarzyna Stelmach
/ Podpis autora pracy/

[1] uwzględniając merytoryczny wkład opiekuna/promotora

Umowa licencyjna niewyłączna *Załącznik nr 3*

Zawarta w Krakowie w dniu 26.05.2021 r. między Uniwersytetem Rolniczym im. Hugona Kołłątaja w Krakowie, reprezentowanym przez Prodziekana ds. Studenckich i Dydaktycznych

a Katarzyną Stelmach

autorem pracy doktorskiej pt. Analiza dystrybucji transpozonów *DcSto* w genomie marchwi oraz wykorzystanie polimorfizmów insercji do analizy struktury genetycznej populacji marchwi uprawnej (*Daucus carota* ssp. *sativus*)

realizowanej w Katedrze Biologii Roślin i Biotechnologii

pod kierunkiem prof. dr hab. Dariusza Grzebelusa

/imię i nazwisko promotora/

1. **Niniejszym udzielam Uniwersytetowi Rolniczemu im. Hugona Kołłątaja w Krakowie nieodpłatnej, bezterminowej licencji niewyłącznej do korzystania z w/w pracy na następujących polach eksploatacji:**
   a. **w zakresie obrotu oryginałem pracy lub egzemplarzami, na których pracę utrwalono w postaci tradycyjnej (papierowej) – poprzez wprowadzenie ich do obrotu, użyczanie lub najem egzemplarzy pracy;**
   b. **w zakresie zwielokrotnienia i rozpowszechniania – w ramach wewnętrznej elektronicznej bazy danych prac dyplomowych – w taki sposób, aby każdy korzystający z wewnętrznej sieci Uniwersytetu mógł mieć do pracy dostęp w miejscu i czasie przez siebie wybranym – od dnia, gdy taka baza danych zostanie w Uniwersytecie uruchomiona.**

2. **udzielenie licencji do korzystania przez Uniwersytet Rolniczy z w/w pracy na polach eksploatacji wymienionym w pkt. 1 ograniczam w następujący sposób:**

—

.....
w imieniu UR Prodziekan /podpis/

Katarzyna Stelmach
.....
/czytelny podpis doktoranta/

Umowa licencyjna wyłączna *Załącznik nr 4*

Zawarta w Krakowie w dniu 26.05.2021 r. między Uniwersytetem Rolniczym im. Hugona Kołłątaja w Krakowie, reprezentowanym przez Prodziekana ds. Studenckich i Dydaktycznych

a Katarzyną Stelmach

autorem pracy doktorskiej pt. Analiza dystrybucji transpozonów *DcSto* w genomie marchwi oraz wykorzystanie polimorfizmów insercji do analizy struktury genetycznej populacji marchwi uprawnej (*Daucus carota* ssp. *sativus*)

realizowanej w Katedrze Biologii Roślin i Biotechnologii

pod kierunkiem prof. dr hab. Dariusza Grzebelusa

/imię i nazwisko promotora/

Świadomy, iż wykonana przeze mnie praca dyplomowa jest częścią tematu badawczego realizowanego przez pracowników Uniwersytetu Rolniczego im. Hugona Kołłątaja w Krakowie, niniejszym udzielam Uniwersytetowi nieodpłatnej, bezterminowej licencji wyłącznej na korzystanie z w/w pracy w zakresie rozpowszechniania pracy lub jej fragmentów, a szczególnie wykorzystywania wyników badawczych zamieszczonych w pracy w sposób inny, niż określono w umowie o udzieleniu licencji niewyłącznej, którą zawarłem w dniu 26.05.2021 r. z Uniwersytetem Rolniczym – z zastrzeżeniem nienaruszalności moich autorskich praw osobistych.

.....
w imieniu UR Prodziekan /podpis/

Katarzyna Stelmach
.....
/czytelny podpis doktoranta/

Promotor:
prof. dr hab. inż. Dariusz Grzebelus — Katedra Biologii Roślin i Biotechnologii, Wydział Biotechnologii i Ogrodnictwa, Uniwersytet Rolniczy im. Hugona Kołłątaja w Krakowie

Promotor pomocniczy:
dr inż. Alicja Macko-Podgórni, prof. UR — Katedra Biologii Roślin i Biotechnologii, Wydział Biotechnologii i Ogrodnictwa, Uniwersytet Rolniczy im. Hugona Kołłątaja w Krakowie

Pracę doktorską realizowałam w Katedrze Biologii Roślin i Biotechnologii w latach 2014-2020 w ramach studiów doktoranckich prowadzonych w Studium Doktoranckim Uniwersytetu Rolniczego im. Hugona Kołłątaja w Krakowie. Niniejsza praca została wykonana w ramach programu badań podstawowych na rzecz postępu biologicznego w produkcji roślinnej finansowanego przez Ministerstwo Rolnictwa i Rozwoju Wsi, zadanie badawcze nr 69 pt. „Opracowanie i wykorzystanie wysokowydajnych technik selekcji genomowej w doskonaleniu warzyw".

W trakcie odbywania studiów doktoranckich uzyskałam środki finansowe z Narodowego Centrum Nauki (projekt ETIUDA, nr 2019/32/T/NZ9/00198), które pozwolą na odbycie 6 miesięcznego stażu w University of Wisconsin-Madison w Stanach Zjednoczonych Ameryki Północnej.

Składam serdeczne podziękowania

prof. dr hab. inż. Dariuszowi Grzebelusowi

za opiekę naukową i pomoc w przygotowaniu rozprawy doktorskiej

pracownikom naukowym i technicznym Katedry Biologii Roślin i Biotechnologii

za współpracę i życzliwą pomoc na wielu etapach pracy

Spis treści

Wykaz publikacji wchodzących w skład rozprawy doktorskiej

Wyniki badań prowadzonych w ramach realizacji rozprawy doktorskiej zostały przedstawione w trzech publikacjach naukowych:

1. **Stelmach K.**, Macko-Podgórni A., Machaj G., Grzebelus D. 2017. Miniature inverted repeat transposable element insertions provide a source of intron length polymorphism markers in the carrot (*Daucus carota* L.). *Frontiers in Plant Science*, 8, 725, doi: 10.3389/FPLS.2017.00725 (IF_{2017} = 4.117)

2. Macko-Podgórni A., **Stelmach K.**, Kwolek K., Grzebelus D. 2019. *Stowaway* miniature inverted repeat transposable elements are important agents driving recent genomic diversity in wild and cultivated carrot. *Mobile DNA*, 10, 47, doi: 10.1186/s13100-019-0190-3 (IF_{2020} = 3.161)

3. **Stelmach K.**, Macko-Podgórni A., Allender C., Grzebelus D. 2021. Genetic diversity of western-type carrots. *BMC Plant Biology*, 21(1), 200, doi:10.1186/s12870-021-02980-0 (IF_{2020} = 3.497)

Każda z wymienionych publikacji zawiera odrębny wstęp, opis materiałów i metod oraz wyniki i dyskusję. Z tego względu w niniejszym opracowaniu ograniczyłam się do podania najważniejszych wyników wynikających z każdej publikacji. Podsumowanie uzyskanych rezultatów w przeprowadzonych badań przedstawiłam łącznie dla wszystkich trzech publikacji.

Wykaz używanych skrótów i terminów

DcSto – ruchome element genetyczne typu *Stowaway* zidentyfikowane w genomie marchwi (ang. *Daucus carota Stowaway*-like MITEs)

F_{IS} – współczynnik wsobności

F_{ST} – współczynnik utrwalenia

H_E – oczekiwana heterozygotyczność

H_O – obserwowana heterozygotyczność

IACV – zmienność wewnątrzodmianowa (ang. *Intra-cultivar Variability*)

IECV – zmienność międzyodmianowa (ang. *Inter-cultivar Variability*)

ILP – polimorfizm długości intronów (ang. *Intron Length Polymorphism*)

IS – miejsce insercji TE (ang. *Insertion Site*)

MITE – krótkie ruchome element genetyczne zawierające powtórzone sekwencje odwrócone (ang. *Miniature Inverted-repeat Transposable Elements*)

PIS – miejsce równoległej insercji TE (ang. *Parallel Insertion Site*)

SIR – krótkie odwrócone subterminalne powtórzenia (ang. *Subterminal Inverted Repeat*)

SNP – polimorfizm pojedynczego nukleotydu (ang. *Single Nucleotide Polymorphism*)

SSR – markery mikrosatelitarne (ang. *Simple Sequence Repeats*)

TEASV – zmienność strukturalna związana z aktywnością ruchomych elementów genetycznych (ang. *TE-Associated Structural Variation*)

TE – ruchome elementy genetyczne (ang. *Transposable Elements*)

TIR – krótkie odwrócone terminalne powtórzenia będące integralną częścią transpozonów DNA (ang. *Terminal Inverted Repeats*)

1. Streszczenie

Celem badań prowadzonych w ramach rozprawy doktorskiej była analiza dystrybucji ruchomych elementów genetycznych *Stowaway*-like w genomie marchwi. Badania koncentrowały się na możliwości wykorzystania zidentyfikowanych insercji transpozonów typu *DcSto* zlokalizowanych w obrębie intronów do opracowania panelu markerów molekularnych typu ILP. Opracowany zestaw markerów posłużył za narzędzie do analizy struktury zmienności genetycznej marchwi uprawnej typu zachodniego.

Materiałem roślinnym wykorzystanym w badaniach były rośliny reprezentujące odmiany populacyjne marchwi uprawnej (*Daucus carota* subsp. *sativus*) typu wschodniego i zachodniego oraz rośliny reprezentujące populacje dzikie (*Daucus carota* subsp. *carota*) typu wschodniego i zachodniego.

Do detekcji oraz analizy dystrybucji *DcSto* wykorzystano sekwencje konsensusowe elementów reprezentujących 14 rodzin transpozonów. Insercje transpozonów *DcSto* zidentyfikowane w obrębie intronów genomu referencyjnego marchwi spełniające przyjęte kryteria zwalidowano pod kątem możliwości opracowania markerów molekularnych typu ILP. Spośród 209 insercji *DcSto* wybrano 90 (średnio 10 na chromosom) identyfikujących polimorfizmy długości intronu w sposób powtarzalny i opracowano jednolity panel stanowiący nowy rodzaj markerów molekularnych opartych na aktywności transpozycyjnej elementów należących do nadrodziny MITE.

Opracowany panel *DcS*-ILP posłużył do analizy struktury zmienności genetycznej kolekcji 390 roślin reprezentujących 78 odmian populacyjnych marchwi typu zachodniego. Równocześnie przeprowadzono genotypowanie przez sekwencjonowanie na tej samej kolekcji roślin, uzyskując zestaw 2354 markerów identyfikujących polimorfizmy pojedynczego nukleotydu – SNP. Poprzez porównanie wyników genotypowania zweryfikowano hipotezę mówiącą o tym, że ze względu na krótszy okres ewolucyjny aktywności transpozycyjnej struktura genetyczna identyfikowana w oparciu o polimorficzne insercje ruchomych elementów genetycznych nie jest tożsama ze strukturą wynikającą z analizy polimorfizmów pojedynczego nukleotydu.

2. Summary

The aim of the conducted research was to understand and analyse the distribution of *DcSto* transposable elements in the carrot genome. The research focused on the possibility of developing a panel of ILP molecular markers based on the polymorphism of *DcSto* insertions localised within introns of the carrot genes. The developed panel of markers was then used to evaluate the structure of genetic variability present in the collection of open-pollinated cultivars representing the western carrot gene pool.

Consensus sequences of transposable elements belonging to 14 *DcSto* families were used to identify *DcSto* insertion sites and to analyse their distribution within the 31 resequenced genomes of various origin and status. In total, more than 18,000 *DcSto* insertion sites were detected in the resequenced genomes, 292 of them were recognised as parallel insertion sites able to harbour transposons belonging to at least two different families. Insertions identified within the introns of the carrot reference genome were manually inspected and validated to meet the criteria for the development of intron length polymorphism molecular markers. 90 of 209 candidate *DcSto* insertions (average of 10 per chromosome) yielded expected PCR products. The *DcS*-ILP markers developed in the course of this study are a novel set of publicly available transposon-based markers in the carrot.

The panel of *DcS*-ILP markers was then exploited as a tool for the analysis of the structure of genetic diversity in the collection of 78 open-pollinated cultivars representing the most popular market types of the western gene pool. Subsequently, genotyping-by-sequencing was carried out on the same collection of cultivars resulting in a novel set of SNP markers used for the analysis of genetic diversity. The results obtained for both molecular marker systems were compared, as we assumed that *DcS*-ILPs might be capable of revealing variability which arose more recently as a consequence of the transpositional activity of *DcSto* MITEs.

3. Przegląd literatury

3.1. Ruchome elementy genetyczne i ich rola w ewolucji genomu

Obserwowany na początku XXI wieku postęp technologii sekwencjonowania DNA umożliwił poznanie sekwencji genomowej wielu gatunków organizmów prokariotycznych i eukariotycznych. Coraz większa liczba danych dotyczących zsekwencjonowanych genomów potwierdziła istnienie znacznego zróżnicowania wielkości i złożoności genomów eukariotycznych. Przykładowo, w królestwie protista obserwowane są ponad 31-krotne różnice w wielkości genomu pomiędzy *Tetrahymena thermophila* (wielkość genomu: 199,7 Mpz; Eisen i in., 2006) a *Babesia microti* (6,43 Mpz; Cornillot i in., 2013). Różnice w wielkości genomu są jeszcze bardziej widoczne w gromadzie roślin okrytonasiennych. Genom szachownicy lisiej (*Fritillaria uva-lupis*) złożony jest z 87 400 Mpz (Leitch i in., 2007), podczas gdy genom rzodkiewnika pospolitego (*Arabidopsis thaliana*) to zaledwie 119 Mpz (Chin i in., 2016). Pomimo obserwowanych znacznych różnic w wielkości genomów liczba sekwencji kodujących pozostaje stosunkowo stała i zawarta jest w przedziale 5 tys. do 50 tys. genów w genomach roślin i kręgowców (Wicker, 2012). Czynnikiem, który w głównej mierze przyczynia się do tak istotnych różnic w wielkości genomów nawet blisko spokrewnionych gatunków, jest liczba sekwencji powtórzeniowych. Wśród nich przeważają ruchome elementy genetyczne, inaczej transpozony. Są to fragmenty DNA zdolne do przemieszczania się w obrębie genomu. Występują powszechnie w świecie organizmów prokariotycznych i eukariotycznych znacząco wpływając na zmienność genetyczną, zwłaszcza organizmów roślinnych i zwierzęcych. Genomy niektórych roślin w ponad 2/3 składają się z sekwencji TE, przykładowo genom kukurydzy jest w 77% złożony z transpozonów (Meyers i in., 2001).

Zaproponowany przez Wickera i in. (2007) podział systematyczny TE zakłada istnienie sześciu poziomów klasyfikacji: klasy, podklasy, rzędu, nadrodziny, rodziny oraz podrodziny (tab. 1).

Tabela 1. System klasyfikacji ruchomych elementów genetycznych

takson	kryterium podziału
klasa	obecność intermediatu RNA podczas transpozycji
podklasa	obecność mechanizmu „kopiuj-wklej" lub „wytnij-wklej"
rząd	różnice w: - mechanizmie insercji - ogólnej organizacji struktury

nadrodzina	różnice w: - strukturze domen kodujących i niekodujących - obecności i długości sekwencji TSD
rodzina	podobieństwo sekwencji rejonów kodujących
podrodzina	dane filogenetyczne

Podstawowy podział TE opiera się na rodzaju powstającego w czasie transpozycji kwasu nukleinowego i obejmuje dwie klasy. W obrębie klasy I TE (retrotranspozonów) wszystkie elementy charakteryzują się transpozycją typu „kopiuj-wklej", jednak poszczególne grupy elementów wykorzystują różne mechanizmy mobilizacji. Wicker i in. (2007) wyróżnili pięć rzędów w obrębie klasy I: retrotranspozony LTR, LINEs, SINEs, elemeny podobne do DIRS i elementy podobne do *Penelope*. W większości genomów roślinnych dominującą klasą są właśnie retrotranspozony, a zwłaszcza rząd retrotranspozonów LTR, choć nie jest to regułą. W tabeli 2 przedstawiono porównanie udziału poszczególnych klas i rzędów TE w genomach pięciu gatunków roślin uprawnych, tj. kukurydzy (*Zea mays*), sorgo (*Sorghum bicolor*), soi (*Glycine max*), marchwi (*Daucus carota*) i ryżu (*Oryza sativa*).

Tabela 2. Procentowy udział poszczególnych typów TE w genomie marchwi, soi, ryżu, sorgo i kukurydzy (Iorizzo i in., 2016; Paterson i in., 2009; Schmutz i in., 2010; Schnable i in., 2009; Yu i in., 2002)

gatunek:	***Daucus carota***	***Glycine max***	***Oryza sativa***	***Sorghum bicolor***	***Zea mays***
typ TE:	**% udział w sekwencji genomu**				
klasa I: retroelementy	**34,6**	**42,2**	**25,8**	**54,5**	**79,4**
retrotranspozony LTR	27,4	41,9	23,5	54,4	75,1
retrotranspozony non-LTR	3,6	0,2	1,2	0,04	0,4
niesklasyfikowane	3,6	0,1	1,1	0,06	3,9
klasa II: transpozony DNA	**13,6**	**16,5**	**13,7**	**7,5**	**2,7**

TE klasy II (transpozony DNA) w większości przypadków charakteryzuje mechanizm transpozycji typu „wytnij-wklej" determinowany przez ich budowę. Zbudowane są z dwóch podstawowych elementów, tj. krótkich odwróconych powtórzeń o długości od 10 do 200 pz znajdujących się na końcach 3' oraz 5' oraz genu kodującego transpozazę czyli białko odpowiedzialne za rozpoznanie fragmentów końcowych fragmentów transpozonu, wycięcie

elementu oraz znalezienie sekwencji DNA odpowiedniej do wklejenia wyciętego elementu. W obu klasach TE występują elementy autonomiczne i nieautonomiczne. Elementy autonomiczne posiadają geny kodujące białka niezbędne do ich mobilizacji, natomiast elementy nieautonomiczne nie posiadają tych genów. Elementy nieautonomiczne mogą być mobilizowane przez ich autonomiczne odpowiedniki (Wicker i in, 2007).

3.1.1. Transpozony typu MITE

Transpozony typu MITE (ang. *Miniature Inverted-repeat Transposable Elements*) są liczną grupą krótkich (do 800 pz) nieautonomicznych ruchomych elementów genetycznych posiadających na swoich końcach krótkie odwrócone terminalne powtórzenia (TIRs, ang. *Terminal Inverted Repeats*). Dwie pierwsze rodziny tych elementów, *Tourist* i *Stowaway*, zostały odkryte na początku lat 90. XX w. w genomie kukurydzy (Bureau i Wessler, 1992, 1994). Sekwencje elementów obydwóch rodzin wykazują dużą homologię do autonomicznych transpozonów z nadrodzin *Tc1*/*Mariner* i *PIF/Harbinger*, co może sugerować ich wspólne pochodzenie. Prawdopodobnie na skutek rearanżacji wewnątrzgenomowych elementy typu MITE utraciły fragmenty sekwencji kodujących białka niezbędne do prawidłowej transpozycji (Feschotte i Mouchès, 2000). MITEs mogą być również mobilizowane przez elementy nie wykazujące homologii sekwencji. Przykładowo, zidentyfikowano kilka odmian ryżu, w których nieautonomiczny transpozon mPing wykazuje aktywność transpozycyjną pomimo braku obecności autonomicznego elementu z którego pochodzi – mobilizacja elementu jest katalizowana przez inne transpozazy obecne w genomie (Bureau i Wessler, 1992).

Elementy typu MITE wykazują tendencję do insercji w pobliżu rejonów kodujących. Insercje powyżej, poniżej lub w obrębie genu mogą mieć bezpośredni wpływ na jego ekspresję (Oki i in., 2008; Zerjal i in., 2009). MITE ulegają również transkrypcji tworząc drugorzędowe struktury typu spinki do włosów (ang. *hairpin structures*) rozpoznawane przez kompleksy białkowe DCL (ang. *dicer-like*), tworząc małe cząsteczki RNA (sRNA) posiadające właściwości regulatorowe. Istnienie tego zjawiska potwierdzono u kilku gatunków roślin i zwierząt, m. in. u *A. thaliana, M. musculus* i *H. sapiens* (Piriyapongsa i in., 2007; Piriyapongsa i Jordan, 2008). Te właściwości MITE sprawiają, że gatunkowo-specyficzna analiza dystrybucji zmienności strukturalnej związanej z aktywnością TE (ang. *TE-Associated Structural Variation*; TEASV), w szczególności elementów typu MITE może pomóc w lepszym zrozumieniu mechanizmów ewolucji genomu.

3.1.2. Identyfikacja elementów typu MITE

Metody identyfikacji i klasyfikacji TE wykorzystują trzy podejścia: identyfikację *de novo* (1), metody opierające się na homologii sekwencji (2) oraz na podobieństwie struktury (3) TE. Pierwszy element typu MITE – *Tourist-Zm1* – został zidentyfikowany *de novo* w genomie kukurydzy w obrębie 11 egzonu genu *waxy* (Bureau i Wessler, 1992). Posiadał on sekwencje TIR o długości 14 pz otoczone przez trójnukleotydowe duplikacje 5'-GCA-3' (ang. *Target Site Duplication*; TSD). W obrębie sekwencji *Zm1* były również obecne subterminalne odwrócone powtórzenia (ang. SIR) o długości 5 pz. Obecność TSD, TIR i SIR – charakterystycznych składowych elementów MITE – pozwoliła na identyfikację pierwszych elementów w genomach kolejnych organizmów. Od czasu odkrycia *Tourist-Zm1* znaleziono i opisano tysiące rodzin transpozonów MITE, większość korzystając z homologii sekwencji elementów już opisanych. Wśród roślin użytkowych liczba zidentyfikowanych rodzin MITE jest bardzo zróżnicowana i waha się od 25 w genomie *Brassica oleracea* (Nouroz i in., 2015; Sampath i in., 2014) poprzez 110 w grupie sześciu gatunków z rodzaju *Citrus* L. (Liu i in., 2019) do ponad 6 tysięcy w genomie pszenicy (Crescente i in., 2018).

Do tej pory opracowano co najmniej kilkanaście narzędzi bioinformatycznych stworzonych na potrzeby identyfikacji MITE, zarówno na bazie homologii sekwencji, jak i podobieństwa struktury TE. Narzędzia BLAST (Altschul i in., 1990), HMMER3 (Eddy, 2011) czy RepeatMasker (Smit i in., 1996) wykorzystują podobieństwo pomiędzy sekwencją znanego elementu a przypuszczalnym elementem obecnym w sekwencji genomowej. Wykorzystanie tego podejścia pozwala na wykrycie nawet jednej kopii elementu w przeszukiwanej sekwencji, nie umożliwia natomiast wykrycia nowych, wcześniej nieopisanych elementów. Dodatkowo, często wynikiem wyszukiwania są sekwencje nie w pełni zachowanych elementów, np. pozbawionych fragmentów lub całych TIR. Narzędzia bazujące na wykryciu podobieństwa struktury transpozonów m.in. detectMITE (Ye i in., 2016), MITE-Hunter (Han i Wessler, 2010) pozwalają na identyfikację charakterystycznych elementów MITE, np. TSD i TIR z uwzględnieniem długości elementu, odległości pomiędzy TIR czy długości samej sekwencji TIR. Często wymienianą wadą tego podejścia jest zbyt wysoka specyficzność przeszukiwania sekwencji skutkująca identyfikacją rodzin MITE składających się z jednej kopii elementu – w rzeczywistości stanowiących fałszywy negatywny wynik analizy (Ye i in., 2016).

Istnieje kilka baz zawierających dane sekwencyjne elementów typu MITE zidentyfikowanych w genomach różnych gatunków. Do najbardziej popularnych należy baza Repbase (Jurka i in., 2005) zawierająca sekwencje repetytywne, w tym transpozony pochodzące z genomów

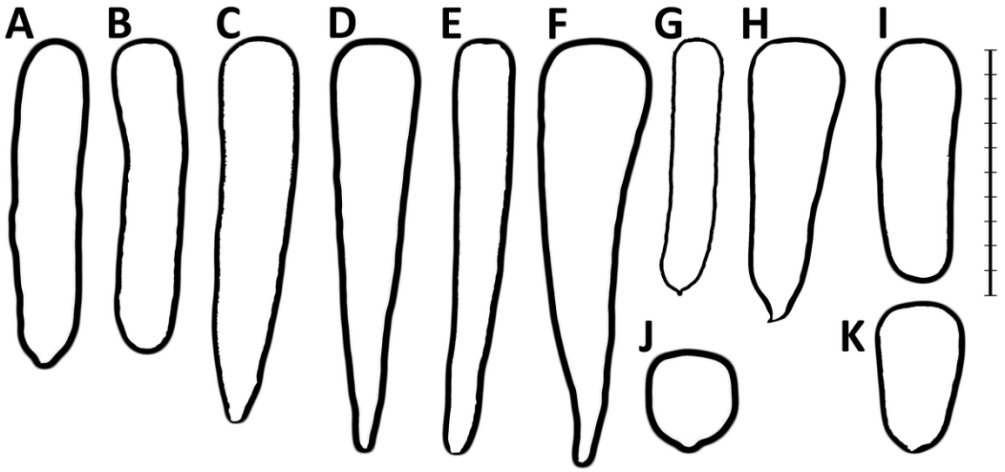
eukariotycznych oraz P-MITE (Chen i in., 2014) dedykowana elementom MITE zidentyfikowanym w ponad 40 genomach roślinnych.

3.1.3. Ruchome elementy genetyczne typu MITE jako markery molekularne

Powszechność występowania w królestwie zwierząt i roślin oraz obecność w genomie w dużej liczbie kopii spowodowała, że TE znajdują potencjalne zastosowanie jako markery molekularne (Kumar i Hirochika, 2001). Mnogość elementów MITE sprawia, że stanowią one cenne źródło zmienności genetycznej. Dotychczas opracowano dwie techniki wykorzystujące sekwencje TIR elementów MITE – MITE Transposon Display oraz Inter-MITE Polymorphism. Znalazły one zastosowanie w analizach zmienności genetycznej licznych gatunków roślin, takich jak: ryż, kukurydza, sorgo czy jęczmień (Casa i in., 2004; Chang i in., 2001; Lee i in., 2005; Park i in., 2003). Insercje elementów MITE, w tym zidentyfikowanych w genomie marchwi elementów *DcSto* (ang. *Daucus carota Stowaway*-like elements), w obrębie intronów generują znaczny stopień zmienności na poziomie genetycznym. Polimorfizm długości intronów (ang. *Intron Length Polymorphism*, ILP), będący również wynikiem insercji TE, może być wykorzystany poprzez opracowanie markerów molekularnych służących m.in. do mapowania genetycznego (Sharma i in., 2020; Wydner i in., 1994) bądź analiz z zakresu genetyki populacji (Lessa, 1992). Podstawą wykorzystania markerów ILP jest znaczna różnica dynamiki zmian zachodzących w obrębie kodujących sekwencji egzonów i niekodujących intronów. Markery molekularne oparte na polimorfizmie insercji ruchomych elementów genetycznych mogą stanowić alternatywę dla innych typów kodominujących markerów molekularnych, w tym zwłaszcza markerów SSR (ang. *Simple Sequence Repeats*) i SNP (ang. *Single Nucleotide Polymorphisms*). W odróżnieniu od wymienionych technik, identyfikacja polimorfizmów insercji wymaga wyłącznie zastosowania standardowych procedur laboratoryjnych - powielania PCR fragmentów genomowego DNA oraz elektroforezy produktów powielania w żelu agarozowym. Umożliwia to szybkie i wiarygodne genotypowanie wielu loci w układzie kodominującym, bez utraty informacji występującej w przypadku technik generujących markery dominujące (np. RAPD, AFLP).

3.2. Marchew uprawna (*Daucus carota* subsp. *sativus*)

Marchew uprawna (*Daucus carota* subsp. *sativus*) należy do najważniejszych gatunków warzyw korzeniowych uprawianych na świecie. W Polsce zajmuje trzecie pod względem powierzchni uprawy miejsce wśród warzyw, po kapuście i cebuli. Powierzchnia uprawy w 2018 roku wynosiła 22,4 tys. ha, stanowiąc 12,7% ogólnej uprawy warzyw gruntowych (Rocznik Statystyczny Rolnictwa, 2019). Marchew należy do rodzaju *Daucus* skupiającego, w zależności od przyjętej klasyfikacji, 21-40 gatunków występujących przede wszystkim na półkuli północnej ale również pojedynczo w Południowej Ameryce oraz Australii. Jest rośliną obcopylną, dwuletnią, miododajną o jadalnym korzeniu spichrzowym. Możliwość wykorzystania korzenia jako potencjalnego surowca spożywczego wpłynęła na zainteresowanie marchwią jako gatunkiem hodowlanym. Szacuje się, że proces udomowienia marchwi rozpoczął się ponad 2 tysiące lat temu (Stolarczyk i Janick, 2011), a centrum udomowienia stanowi Azja Centralna (Iorizzo i in., 2013). W wyniku tego procesu powstały dwie główne grupy: marchew typu wschodniego/azjatyckiego charakteryzująca się fioletowym lub żółtym korzeniem, oraz marchew typu zachodniego obejmująca odmiany wytwarzające korzeń koloru pomarańczowego, ale również czerwonego lub białego (Banga, 1963). Dotychczas prowadzone badania dotyczące analizy zróżnicowania genetycznego marchwi z zastosowaniem markerów molekularnych wskazują na istnienie trzech odrębnych puli genowych odpowiadających marchwi dzikiej (1), oraz uprawnej typu wschodniego (2) i zachodniego (3) (Baranski i in., 2012; Cavagnaro i in., 2011; Grzebelus i in., 2014) potwierdzając tym samym powszechnie akceptowaną teorię dotyczącą udomowienia marchwi. Cechy ważne z punktu widzenia udomowienia marchwi obejmowały między innymi zdolność tworzenia mięsistego korzenia spichrzowego, brak tendencji do jego rozwidlania oraz dwuletni okres wegetacyjny. W dalszej kolejności prowadzono selekcję pod kątem cech związanych z jakością korzenia, między innymi jego kolorem i kształtem. Kształt korzenia spichrzowego jest istotną cechą agronomiczną, u marchwi typu zachodniego bardzo zróżnicowaną (ryc. 1). Wyróżnia się kilkanaście typów kształtu korzenia, począwszy od bardzo krótkich i owalnych (odmiany typu Paris Market), po bardzo długie i cienkie (odmiany typu Imperator), z wieloma typami pośrednimi. Zróżnicowanie kształtu korzenia, choć zależne w pewnym stopniu od warunków wzrostu, determinowane jest również w znacznym stopniu poprzez czynniki genetyczne. Nadal nieznane są mechanizmy warunkujące tak dużą zmienność fenotypową obserwowaną wśród marchwi uprawnej typu zachodniego.



Rycina 1. Typowe kształty korzenia spichrzowego marchwi reprezentujące najpopularniejsze typy hodowlane marchwi zachodniej. A – Nantes, B – Berlikum, C – Autumn King, D – Long Orange, E – Imperator, F – St. Valery, G – Amsterdam, H – Danvers, I – Chantenay, J – Paris Market, K – Guerande (=Oxheart). Długość całkowita podziałki odpowiada 20 cm

3.3. Ruchome elementy genetyczne w genomie marchwi

Opublikowanie sekwencji genomu marchwi uprawnej dostarczyło ważnych informacji dotyczących budowy i organizacji genomu tego gatunku (Iorizzo i in., 2016). Wielkość genomu marchwi jest szacowana na 473 Mpz, z czego ponad 197 Mpz (46% genomu) stanowią sekwencje repetytywne. 98% spośród sekwencji powtarzalnych stanowią ruchome elementy genetyczne klasy I i II. Wśród transpozonów klasy I najliczniej reprezentowane są retrotranspozony LTR (ponad 134 000 elementów), natomiast wśród elementów klasy II najliczniejszą grupę stanowią transpozony *hAT* (ponad 65 000 elementów). Udział ruchomych elementów genetycznych, jest znacznie wyższy niż u innych gatunków charakteryzujących się zbliżoną wielkością genomu (*Vitis vinifera* : 41,4%, *Cucumis melo*: 20%, *Amaranthus hypochondriacus*: 23,1%;Clouse i in, 2016; Cuevas i in., 2008; Jaillon i in., 2007), a analiza dystrybucji współczynników dywergencji TE w genomie marchwi wskazuje na ich stosunkowo niedawne pochodzenie (Iorizzo i in., 2016).

Elementy należące do dwóch nadrodzin transpozonów typu MITE– *Tourist*-like *Krak* i *Stowaway*-like *DcSto* są jedynymi dotychczas scharakteryzowanymi elementami nieautonomicznymi. Transpozony *DcSto* są reprezentowane w genomie marchwi liczniej niż elementy *Krak*. Do tej pory w genomie referencyjnym zidentyfikowano powyżej 4 000 kopii *DcSto* należących do 14 rodzin. Liczba zidentyfikowanych insercji *DcSto* w genomie znacznie przewyższa liczbę elementów typu *Stowaway* w poznanych dotąd genomach roślin

reprezentujących klad astrowych (Iorizzo i in., 2016). Pierwszy element *DcSto* został zlokalizowany w sekwencji pierwszego intronu genu *rs* kodującego izoenzym II inwertazy (Macko-Podgorni i in., 2013). Iorizzo i in. (2016) wykazali, że obecność licznych insercji elementów *DcSto* w obrębie lub bliskim sąsiedztwie genów nie jest jednak wynikiem preferencyjnej insercji lecz wynika raczej ze znacznej liczebności tej nadrodziny transpozonów w genomie. Analiza podobieństwa sekwencji konsensusowych elementów *DcSto* wykazała istnienie znacznych różnic strukturalnych pomiędzy poszczególnymi rodzinami, wskazując na ich równoległą ekspansję w genomie. Dzięki cechom charakterystycznym transpozonów *DcSto*, takim jak: (1) stosunkowo niedawny charakter insercji, (2) duża liczebność w genomie oraz (3) częsta obecność w pobliżu rejonów kodujących, elementy te mogą dostarczyć cennych informacji dotyczących zróżnicowania genetycznego marchwi uprawnej. Polimorfizm insercji *DcSto* może stanowić źródło markerów molekularnych mających zastosowanie w analizie struktury populacji marchwi uprawnej.

4. Hipotezy badawcze i cele prowadzonych badań

W pracy doktorskiej postawiono następujące hipotezy badawcze:

1. Duży polimorfizm insercji transpozonów *DcSto* wynika z ich stosunkowo niedawnej aktywności transpozycyjnej i prowadzi do znacznych różnic strukturalnych w obrębie sekwencji kodujących genomu marchwi.

2. Polimorfizm insercji transpozonów *DcSto* jest istotnym źródłem zróżnicowania genetycznego marchwi uprawnej typu zachodniego i może stanowić narzędzie do charakterystyki zmienności genetycznej roślin reprezentujących zachodnią pulę genetyczną.

3. Struktura zmienności genetycznej marchwi uprawnej identyfikowana w oparciu o polimorficzne insercje ruchomych elementów genetycznych, m. in. *DcSto*, nie jest tożsama ze strukturą wynikającą z analizy polimorfizmów pojedynczego nukleotydu (SNP), ze względu na krótszy okres ewolucyjny aktywności transpozycyjnej tych elementów w porównaniu do akumulacji SNP.

Aby zweryfikować powyższe hipotezy wyznaczono następujące cele naukowe:

1. Analiza dystrybucji insercji elementów *DcSto* w obrębie sekwencji kodujących genomu referencyjnego oraz 31 resekwencjonowanych genomów marchwi.

2. Opracowanie markerów molekularnych opartych na polimorfizmie insercji *DcSto* zidentyfikowanych w genomie referencyjnym oraz w resekwencjonowanych genomach marchwi uprawnej typu zachodniego.

3. Analiza struktury zmienności genetycznej kolekcji odmian populacyjnych marchwi uprawnej typu zachodniego przeprowadzona w oparciu o opracowany panel markerów ILP i SNP.

5. Materiały i metody

Do badań prowadzonych w ramach pracy doktorskiej wykorzystano następujące materiały roślinne:

1. Do analizy dystrybucji transpozonów *DcSto* (Macko-Podgórni i in., 2019): DNA pochodzące z linii DH1 (podwojony haploid), czterech linii wsobnych, czternastu odmian uprawnych typu wschodniego i zachodniego, ośmiu roślin dzikich reprezentujących pulę genową zachodnią i wschodnią oraz pięciu roślin dzikich reprezentujących dwa podgatunki marchwi – tab. 3. Materiał został udostępniony przez Departament Rolnictwa Stanów Zjednoczonych Ameryki Północnej (USDA), dzięki uprzejmości prof. Philippa Simona.
2. Do opracowania markerów *DcS*-ILP (Stelmach i in., 2017): DNA pochodzące z dwudziestu trzech odmian uprawnych typu zachodniego, linii DH oraz czterech roślin reprezentujących dwa podgatunki marchwi – tab. 4. Materiał roślinny został udostępniony przez Warwick Genetic Resources Unit w Wielkiej Brytanii.
3. Do analizy struktury zmienności genetycznej marchwi uprawnej (Stelmach i in., 2021): DNA pochodzące z siedemdziesięciu ośmiu odmian marchwi uprawnej typu zachodniego – tab. 5. Materiał roślinny został udostępniony przez Warwick Genetic Resources Unit w Wielkiej Brytanii.

Tabela 3. Zestawienie materiału roślinnego wykorzystanego w pracy Macko-Podgórni i in. (2019). Stosowane skróty: DH – podwojony haploid, ODW – odmiana typu wschodniego, OTZ – odmiana typu zachodniego, DTW – dzika typu wschodniego, DTZ – dzika typu zachodniego

l.p.	symbol odmiany/linii	podgatunek	status	pochodzenie	numer BioSample NCBI
1	DH1	*D. carota ssp. sativus*	linia DH	NLD	SAMN0321663
2	I1	*D. carota ssp. sativus*	linia wsobna	USA	SAMN03766317
3	I2	*D. carota ssp. sativus*	linia wsobna	USA	SAMN03766318
4	I3	*D. carota ssp. sativus*	linia wsobna	USA	SAMN03766319
5	I4	*D. carota ssp. sativus*	linia wsobna	USA	SAMN03766320
6	C1	*D. carota ssp. sativus*	OTW	AFG	SAMN03766321
7	C2	*D. carota ssp. sativus*	OTW	CHN	SAMN03766322
8	C3	*D. carota ssp. sativus*	OTW	UZB	SAMN03766323
9	C4	*D. carota ssp. sativus*	OTW	AFG	SAMN03766324
10	C5	*D. carota ssp. sativus*	OTW	SYR	SAMN03766325
11	C6	*D. carota ssp. sativus*	OTW	TUR	SAMN03766326
12	C7	*D. carota ssp. sativus*	OTZ	JPN	SAMN03766327
13	C8	*D. carota ssp. sativus*	OTZ	BRA	SAMN03766328
14	C9	*D. carota ssp. sativus*	OTZ	NLD	SAMN03766329

15	C10	*D. carota ssp. sativus*	OTZ	USA	SAMN03766330
16	C11	*D. carota ssp. sativus*	OTZ	NLD	SAMN03766331
17	C12	*D. carota ssp. sativus*	OTZ	FRA	SAMN03766332
18	C13	*D. carota ssp. sativus*	OTZ	USA	SAMN03766333
19	C14	*D. carota ssp. sativus*	OTZ	BEL	SAMN03766334
20	W1	*D. carota ssp. carota*	DTZ	PRT	SAMN03766342
21	W2	*D. carota ssp. carota*	DTZ	PRT	SAMN03766350
22	W3	*D. carota ssp. carota*	DTZ	FRA	SAMN03766336
23	W4	*D. carota ssp. carota*	DTW	CHN	SAMN03766338
24	W5	*D. carota ssp. carota*	DTW	UZB	SAMN03766335
25	W6	*D. carota ssp. carota*	DTW	TUR	SAMN03766343
26	W7	*D. carota ssp. carota*	DTW	TUR	SAMN03766337
27	W8	*D. carota ssp. carota*	DTW	PAK	SAMN03766339
28	Ssp1	*D. carota ssp. gummifer*	dzika	PRT	SAMN03766344
29	Ssp2	*D. carota ssp. gummifer*	dzika	PRT	SAMN03766345
30	Ssp3	*D. carota ssp. gummifer*	dzika	FRA	SAMN03766351
31	Ssp4	*D. carota ssp. gummifer*	dzika	FRA	SAMN03766341
32	Ssp5	*D. carota ssp. capilifolius*	dzika	LBY	SAMN03766340

Tabela 4. Zestawienie materiału roślinnego wykorzystanego w pracy Stelmach i in. (2017). Stosowane skróty: OP – odmiana populacyjna, DH – podwojony haploid, WGRU – Warwick Genetic Resources Unit

l.p.	numer odmiany	podgatunek	status	nazwa odmiany	typ korzenia	pochodzenie	nr ident. WGRU
1	RS33	*D. carota* ssp. *sativus*	OP	Chantenay Royal	Chantenay	FRA	8860
2	RS34	*D. carota* ssp. *sativus*	OP	Chantenay Red Cored	Chantenay	GBR	8847
3	RS35	*D. carota* ssp. *sativus*	OP	Royal Chantenay	Chantenay	USA	3882
4	RS37	*D. carota* ssp. *sativus*	OP	Gold King	Chantenay	USA	5127
5	RS39	*D. carota* ssp. *sativus*	OP	Chantenay Long Type	Chantenay	USA	5090
6	RS41	*D. carota* ssp. *sativus*	OP	Chantenay Rex RS	Chantenay	NLD	5589
7	RS43	*D. carota* ssp. *sativus*	OP	Danvers 126	Danvers	GBR	6487
8	RS44	*D. carota* ssp. *sativus*	OP	Danvers Danro RS	Danvers	NLD	5595
9	RS45	*D. carota* ssp. *sativus*	OP	Danvers Red Cored	Danvers	USA	5128
10	RS49	*D. carota* ssp. *sativus*	OP	Danvers	Danvers	NLD	11144
11	RS50	*D. carota* ssp. *sativus*	OP	Danvers Pride	Danvers	USA	8098
12	RS51	*D. carota* ssp. *sativus*	OP	Danvers Half Long	Danvers	USA	8109
13	RS56	*D. carota* ssp. *sativus*	OP	Paris Market	Early Short Horn	NLD	5596
14	RS57	*D. carota* ssp. *sativus*	OP	Paris Forcing	Early Short Horn	GBR	3966
15	RS59	*D. carota* ssp. *sativus*	OP	French Forcing Horn	Early Short Horn	GBR	6489
16	RS60	*D. carota* ssp. *sativus*	OP	Parijse Markt	Early Short Horn	—	9294

17	RS62	*D. carota* ssp. *sativus*	OP	Parijse Markt (Rubin)	Early Short Horn	—	9296
18	RS71	*D. carota* ssp. *sativus*	OP	Gold Pak	Imperator	USA	3885
19	RS72	*D. carota* ssp. *sativus*	OP	Imperator 408	Imperator	USA	3907
20	RS73	*D. carota* ssp. *sativus*	OP	Imperator	Imperator	NLD	11145
21	RS74	*D. carota* ssp. *sativus*	OP	Imperator 407	Imperator	USA	3891
22	RS75	*D. carota* ssp. *sativus*	OP	Long Imperator 58	Imperator	USA	3917
23	RS76	*D. carota* ssp. *sativus*	OP	Imperator 58	Imperator	USA	3892
24	DH1	*D. carota* ssp. *sativus*	linia DH	—	—	NLD	—
25	CDS15	*D. carota* ssp. *azoricus*	dzika	—	—	ESP	6667
26	CDS39	*D. carota* ssp. *carota*	dzika	—	—	CHE	9226
27	CDS93	*D. carota* ssp. *carota*	dzika	—	—	USA	—
28	CDS40	*D. carota* ssp. *carota*	dzika	—	—	POL	9270

Tabela 5. Zestawienie odmian uprawnych marchwi (*D. carota* ssp. *sativus*) wykorzystanych w pracy Stelmach i in. (2021). Stosowane skróty: WGRU – Warwick Genetic Resources Unit

l.p.	**numer odmiany**	**symbol odmiany**	**nazwa odmiany**	**typ korzenia**	**pochodzenie**	**nr ident. WGRU**
1	RS01	AM1	Foram	Amsterdam	NLD	7123
2	RS03	AM2	Amsterdam Grace	Amsterdam	DNK	3942
3	RS04	AM3	Amsterdam 5564	Amsterdam	GBR	3981
4	RS06	AM4	Amsterdam Forcing	Amsterdam	GBR	5477
5	RS07	AM5	Amsterdamska	Amsterdam	POL	10371
6	RS08	AM6	Pickmo	Amsterdam	SWE	6030
7	RS09	AM7	Amstel	Amsterdam	FRA	3970
8	RS10	AM8	Amsterdammer Bak Normaal	Amsterdam	NLD	11153
9	RS11	AU1	Karotan	Autumn King	GBR	6511
10	RS12	AU2	Flakkee	Autumn King	FRA	10327
11	RS13	AU3	Flakko	Autumn King	NLD	3961
12	RS14	AU4	Autumn King Vita Longa	Autumn King	GBR	6513
13	RS15	AU5	Regol	Autumn King	DNK	6483
14	RS16	AU6	Century	Autumn King	GBR	3938
15	RS17	AU7	Rothild	Autumn King	GER	6021
16	RS19	AU8	Beacon	Autumn King	GBR	6026
17	RS20	AU9	Regulus Imperial	Autumn King	SWE	6031
18	RS21	BN1	Suko	Baby Nantes	GER	8118
19	RS22	BE1	Feonia	Berlicum/Imperator	DNK	6484
20	RS24	BE2	Banta	Berlicum	DNK	3949
21	RS26	BE3	Lange Rote St O Herz 2/Zino	Berlicum	GER	12403
22	RS27	BE4	Berlikum Perfecta	Berlicum	ITA	6085
23	RS29	BE5	Camberley	Berlicum	GBR	6045
24	RS30	BE6	Berlicum Normaal	Berlicum	NLD	11157
25	RS32	BE7	Feonia Nobo	Berlicum	NLD	7256

26	RS33	CH1	Chantenay Royal	Chantenay	FRA	8860
27	RS34	CH2	Chantenay Red Cored	Chantenay	GBR	8847
28	RS35	CH3	Royal Chantenay	Chantenay	USA	3882
29	RS37	CH4	Gold King	Chantenay	USA	5127
30	RS38	CH5	Comet	Chantenay	GBR	11257
31	RS39	CH6	Chantenay Long Type	Chantenay	USA	5090
32	RS40	CH7	Macbeth	Chantenay	GBR	4670
33	RS41	CH8	Chantenay Rex Rs	Chantenay	NLD	5589
34	RS42	CH9	Criolla	Chantenay	ARG	7255
35	RS43	DA1	Danvers 126	Danvers	GBR	6487
36	RS44	DA2	Danvers Danro Rs	Danvers	NLD	5595
37	RS45	DA3	Danvers Red Cored	Danvers	USA	5128
38	RS48	DA4	Danvers Half Long Scarlet Intermediate	Danvers	GBR	8097
39	RS49	DA5	Danvers	Danvers	NLD	11144
40	RS50	DA6	Danvers Pride	Danvers	USA	8098
41	RS51	DA7	Danvers Half Long	Danvers	USA	8109
42	RS52	DA8	Danvers Half Long Danvers 126	Danvers	GBR	10110
43	RS53	PA1	Davanture	Early Short Horn	FRA	7131
44	RS55	PA2	Duwicker	Early Short Horn	-	9308
45	RS56	PA3	Paris Market	Early Short Horn	NLD	5596
46	RS57	PA4	Paris Forcing	Early Short Horn	GBR	3966
47	RS58	PA5	Early Scarlet Horn	Early Short Horn	GBR	6490
48	RS60	PA6	Parijse Markt	Early Short Horn	-	9294
49	RS61	PA7	Early Scarlet Horn	Early Short Horn	-	9311
50	RS64	FL1	Rood Hild	Flakkee	NLD	7258
51	RS65	GU1	Guerande	Guerande	GBR	6491
52	RS66	GU2	Oxheart	Guerande	GBR	13891
53	RS67	GU3	Flattie (Oxheart)	Guerande	-	3983
54	RS68	GU4	Guerande Oxheart	Guerande	-	9312
55	RS69	GU5	Oxheart	Guerande	USA	3906
56	RS70	IM1	Gold Pak 28	Imperator	USA	5814
57	RS71	IM2	Gold Pak	Imperator	USA	3885
58	RS72	IM3	Imperator 408	Imperator	USA	3907
59	RS75	IM4	Long Imperator 58	Imperator	USA	3917
60	RS76	IM5	Imperator 58	Imperator	USA	3892
61	RS78	LC1	Newmarket	Long Chantenay	GBR	3843
62	RS79	LO1	Long Red Surrey	Long Orange	GBR	6102
63	RS80	PA8	Kardinal Marche De Paris	Market De Paris	GER	7121
64	RS81	NA1	Early Nantes	Nantes	GBR	8874
65	RS82	NA2	Delta A Cuore Rosso	Nantes	ITA	7125
66	RS83	NA3	Tip Top	Nantes	NLD	8895
67	RS84	NA4	Tantal	Nantes	FRA	3956
68	RS85	NA5	Nantes Duke	Nantes	DNK	3945
69	RS86	NA6	Juwarot	Nantes	GER	5994
70	RS87	NA7	Touchon	Nantes	FRA	6090

71	RS89	NA8	Sytan	Nantes	-	9328
72	RS90	NA9	Nantskaja	Nantes	RUS	6761
73	RS92	SV1	St Valery	St Valery	GBR	6514
74	RS96	SV2	New Red Intermediate	St Valery	GBR	6163
75	RS97	FO1	Koewortelen	Cow Carrot	SWE	13161
76	RS98	FO2	Egmont Gold	Fodder Carrot	NZL	6330
77	RS99	FO3	Flakee Samo	Fodder Carrot	NLD	10532
78	RS100	FO4	Flavius	Fodder Carrot	-	11290

Szczegółowy opis metod badawczych wykorzystywanych w poszczególnych częściach prowadzonych badań opinano w rozdziale '*Materials and Methods'* w każdej z publikacji wchodzących w skład pracy doktorskiej.

6. Najważniejsze wyniki przeprowadzonych badań

6.1. Opracowanie markerów molekularnych *DcS*-ILP

Stelmach K., Macko-Podgórni A., Machaj G., Grzebelus D. 2017. Miniature inverted repeat transposable element insertions provide a source of intron length polymorphism markers in the carrot (*Daucus carota* L.). *Frontiers in Plant Science*, 8, 725

IF_{2017} = 4.117

4-letni IF = 5.244

Głównym celem badań było opracowanie i walidacja panelu markerów molekularnych opartych na aktywności transpozycyjnej elementów *DcSto* w genomie marchwi. Do opracowania markerów identyfikujących polimorfizm długości intronów (ILP) wykorzystano 209 insercji transpozonów *DcSto* spełniających wymagane kryteria, tj. obecność *DcSto* w obrębie intronu jako jedynego TE (1), długość intronu nie przekraczająca 3700 pz (2) oraz równa dystrybucja w genomie (3). Liczba insercji *DcSto* spełniająca wymienione kryteria wahała się od 18 (chromosom 9) do 32 (chromosom 2). W wyniku amplifikacji 209 wytypowanych intronów z zastosowaniem techniki łańcuchowej reakcji polimerazy (ang. *Polymerase Chain Reaction*, PCR) dla 100 loci otrzymano oczekiwany profil prążkowy odpowiadający polimorfizmowi insercji elementu *DcSto*. W obrębie 10 loci zidentyfikowano dodatkowy amplikon u przynajmniej jednej z genotypowanych roślin. Sekwencjonowanie dodatkowych amplikonów wykazało, że żaden z nich nie był związany z aktywnością transpozycyjną elementów *DcSto* obecnych w genomie referencyjnym marchwi. Potencjalne markery *DcS*-ILP podzielono na klasy uwzględniając długość amplifikowanego intronu, w przedziałach co 600 pz, w zakresie od 400 do 3400 pz. Stosunek loci kandydujących do zwalidowanych był najwyższy dla klasy III (introny o długości 1601-2200 pz) i stanowił 55,6%. Markery klasy V i VI, tj. o długości intronu przekraczającej 2800 pz, charakteryzowały się niejednoznacznym wynikiem amplifikacji i nie zostały wykorzystane w dalszych analizach. Docelowo do opracowania panelu do genotypowania marchwi uprawnej wybrano 90 biallelicznych markerów opartych na polimorfizmie insercji *DcSto*.

Użyteczność opracowanych markerów w analizie struktury zmienności genetycznej marchwi zweryfikowano poprzez genotypowanie kolekcji dwudziestu trzech roślin reprezentujących odmiany uprawne przynależące do czterech typów o zróżnicowanym pochodzeniu i kształcie korzenia spichrzowego (tj. Chantenay, Danvers, Imperator oraz Paris Market) oraz czterech roślin reprezentujących marchew dziką. Łącznie zidentyfikowano 180 alleli, średnio 2,0 na

locus. 2,78% alleli było rzadkich (frekwencja występowania < 0,05). Obserwowana heterozygotyczność (H_O) dla pojedynczych loci wahała się pomiędzy 0,04 a 0,56 (średnia 0,24), natomiast oczekiwana heterozygotyczność (H_E) zawierała się w zakresie 0,04 do 0,51 (średnia 0,34). Współczynnik informacji o polimorfizmie (ang. *Polymorphism Information Content*, PIC) obliczony dla pojedynczych loci wahał się pomiędzy 0,04 a 0,37 (średnia 0,27). Analizę struktury zmienności genetycznej przeprowadzono metodą klastrowania Bayesowskiego z wykorzystaniem oprogramowania STRUCTURE. Wartości ΔK, współczynnika stosowanego do wskazania najbardziej prawdopodobnej liczby klastrów w obrębie analizowanej grupy osobników, wskazały na obecność dwóch klastrów (K=2) skupiających 23 odmiany marchwi uprawnej w grupie 1 (G1) oraz 4 rośliny dzikie w grupie 2 (G2). Zaobserwowane wartości współczynnika przynależności do grupy (ang. *membership coefficient*, Q) były bardzo wysokie. W G1 wahały się między 0,831 a 0,997, natomiast w G2 wynosiły pomiędzy 0,965 a 0,998. Obserwowane wartości dystansu genetycznego pomiędzy osobnikami, mierzone jako H_E, były bardzo zbliżone w obu grupach, odpowiednio 0,31 i 0,29 w G1 i G2. W celu dokładniejszej analizy struktury zmienności genetycznej marchwi uprawnej wykonano ponowną analizę STRUCTURE wyłącznie dla roślin zgrupowanych w G1. Najwyższe wartości ΔK uzyskano dla K=21, K=2 oraz K=4. Przy założeniu K=4 trzy rośliny zostały zgrupowane z podgrupie 1 (PG1), sześć w podgrupie 2 (PG2), pięć w podgrupie 3 (PG3) oraz pięć w podgrupie 4 (PG4). Wartości Q wahały się pomiędzy 0,782 a 0,961. Najwyższe wartości uzyskano w obrębie PG1 skupiającej trzy rośliny reprezentujące typ Chantenay (Q 0,928-0,962). Cztery rośliny wykazywały wysoki poziom admiksji (Q < 0,6) uniemożliwiający przypisanie do żadnej z podgrup. Każda z wyodrębnionych podgrup charakteryzowała się dominującym udziałem odmian danego typu. W PG1 dominował typ Chantenay (średnia wartość Q = 0,605), w PG2 typ Danvers (Q = 0,626), w PG3 typ Imperator (Q = 0,786) oraz w PG4 typ Paris Market (Q = 0,884). Powyższe wyniki grupowania w dużym stopniu odpowiadają proponowanej przez Banga (1963) historii hodowli marchwi typu zachodniego i wskazują na możliwy związek pomiędzy strukturą zmienności genetycznej a obserwowanym kształtem korzenia spichrzowego.

6.2. Analiza dystrybucji transpozonów *DcSto* w genomie marchwi

Macko-Podgórni A., Stelmach K., Kwolek K., Grzebelus D. 2019. *Stowaway* miniature inverted repeat transposable elements are important agents driving recent genomic diversity in wild and cultivated carrot. *Mobile DNA*, 10(1), 47

IF_{2019} = 3.161

5-letni IF = 3.371

Głównym celem badań było wykorzystanie 14 rodzin transpozonów *DcSto* do analizy TEASV w 31 resekwencjonowanych genomach reprezentujących cztery pule genetyczne marchwi, tj. marchew uprawną typu zachodniego (1) i wschodniego (2) oraz marchew dziką o pochodzeniu europejskim (3) i azjatyckim (4). Przeprowadzono szczegółową annotację insercji elementów *DcSto* i przeanalizowano ich dystrybucję w genomie marchwi.

Przy zastosowaniu narzędzia RelocaTE zidentyfikowano 18 518 miejsc insercji (ang. *Insertion Site*, IS) *DcSto* w 31 genomach marchwi. Zróżnicowanie pokrycia genomu (od 10x do 40x) nie wpłynęło na dokładność identyfikacji IS – nie zaobserwowano korelacji pomiędzy liczbą odczytów a liczbą zidentyfikowanych insercji *DcSto*. Zaobserwowano znaczne różnice w liczebności elementów reprezentujących poszczególne rodziny *DcSto*. Najmniej liczną rodziną, z zaledwie 155 zidentyfikowanymi IS była *DcSto11,* podczas gdy insercje elementów *DcSto6* zidentyfikowano 3633 razy. W celu zweryfikowania dokładności detekcji IS *in silico* wybrano losowo 39 loci, dla których opracowano markery *DcS*-ILP. Dla 16 loci wyniki amplifikacji w pełni potwierdziły wyniki identyfikacji *in silico*. W 12 loci sporadycznie zaobserwowano obecność dodatkowych wariantów o innej długości od oczekiwanej lub nie zidentyfikowano obecności żadnego elementu *DcSto*. Wyniki amplifikacji pozostałych 11 IS były niejednoznaczne – uzyskano niespecyficzne amplikony. Analiza amplikonów uzyskanych dla 28 IS dających jednoznaczne produkty PCR wykazała ponad 96% zgodność z predykcją RelocaTE uzyskaną dla konkretnej kombinacji roślina/IS. 292 spośród wszystkich zidentyfikowanych IS było tzw. równoległymi miejscami insercji (ang. *Parallel Insertion Site*, PIS) – w tej samej pozycji w genomie zidentyfikowano insercje co najmniej dwóch różnych elementów *DcSto*. 95% spośród PIS zawierało kopie elementów pochodzących z dwóch różnych rodzin. W pozostałych 5% zidentyfikowano elementy należące do co najmniej trzech rodzin. W celu weryfikacji predykcji *in silico* wytypowano 11 regionów obejmujących PIS, które poddano amplifikacji w 30 resekwencjonowanych roślinach oraz linii DH1 służącej jako genom referencyjny. Otrzymane amplikony sekwencjonowano metodą Sangera. Obecność PIS potwierdzono we wszystkich amplifikowanych regionach. Zaobserwowano rozbieżność

wyników predykcji *in silico* i weryfikacji PCR dla pojedynczych roślin w analizowanych PIS. Przykładowo, obecność amplikonów dłuższych niż oczekiwane „puste" IS w analizie RelocaTE wynikała z obecności dodatkowych rearanżacji, np. insercji niezidentyfikowanego elementu MITE, insercji elementu *DcSto* należącego do rodziny innej niż analizowane w doświadczeniu czy insercji fragmentów retrotranspozonów LTR. Obecność u pojedynczych roślin produktów PCR krótszych niż oczekiwane „puste" IS była prawdopodobnie wynikiem utworzenia krótkich delecji w trakcie mobilizacji elementu *DcSto* obecnego uprzednio w analizowanym locus. Wyniki predykcji *in silico* PIS oraz weryfikacji metodą PCR wskazują na dość powszechne występowanie insercji różnych elementów *DcSto* oraz innych elementów typu MITE w tej samej pozycji w genomu marchwi.

Spośród 18 518 zidentyfikowanych IS, tylko dwa były obecne we wszystkich resekwencjonowanych genomach. Zaledwie 22 IS były współdzielone przez rośliny reprezentujące marchew uprawną. Liczba insercji przypadająca na genom marchwi wahała się od 468 do 1978. Średnia liczba zidentyfikowanych kopii *DcSto* była nieznacznie wyższa dla genomów reprezentujących marchew uprawną (1655 IS – uprawna, 1226 IS – dzika). W obrębie genomów odmian uprawnych (C1-C13) oraz linii wsobnych (I1-I3) zidentyfikowano porównywalną liczbę kopii *DcSto*. Większe zróżnicowanie liczby *DcSto* zaobserwowano w grupie genomów reprezentujących marchew dziką – liczba kopii *DcSto* wśród roślin pochodzących z terenów basenu Morza Śródziemnego, uważanych za centrum bioróżnorodności marchwi, była poniżej średniej, podczas gdy w puli genomów reprezentujących marchew dziką azjatycką (W4-W7) liczba kopii *DcSto* była porównywalna do liczby obserwowanej w genomach marchwi uprawnej. Wyniki analizy PCoA (ang. *Principal Coordinate Analysis*, PCoA) opartej na polimorfizmie insercji *DcSto* w resekwencjonowanych genomach odzwierciedlały podział badanych roślin na cztery wyraźnie odrębne pule genetyczne: uprawną zachodnią (C7-C13, I1-I4), uprawną wschodnią (C1-C6), dziką zachodnią (W1-W3, Ssp1-Ssp5) oraz dziką wschodnią (W4-W8). Dystrybucja poszczególnych rodzin *DcSto* była zróżnicowana w obrębie poszczególnych puli genetycznych. Przykładowo, elementy z rodzin *DcSto1* oraz *DcSto8* występowały rzadziej w genomach marchwi uprawnej i dzikiej typu wschodniego w porównaniu do typu zachodniego. Zaobserwowano istotne różnice w dystrybucji elementów *DcSto7b* w genomach roślin typu wschodniego – genomy roślin dzikich zawierały znacznie mniej elementów z tej rodziny. Największą różnorodność dystrybucji elementów *DcSto* zaobserwowano w puli genetycznej marchwi dzikiej typu zachodniego.

6.3. Analiza struktury zmienności genetycznej marchwi typu zachodniego

Stelmach K., Macko-Podgórni A., Allender C., Grzebelus D. 2021. Genetic diversity of western-type carrots. *BMC Plant Biology*, 21(1), 200

IF_{2020} = 3.497

5-letni IF = 4.494

Głównym celem pracy była detekcja struktury zmienności genetycznej leżącej u podstaw różnic fenotypowych obserwowanych u odmian populacyjnych marchwi typu zachodniego reprezentujących różne typy korzenia spichrzowego. W doświadczeniu wykorzystaliśmy dwa typy kodominujących markerów molekularnych, tj. SNP i *DcS*-ILP, zakładając, że markery *DcS*-ILP umożliwią detekcję zmienności wynikającej ze stosunkowo niedawnej aktywności transpozycyjnej elementów *DcSto* w genomie marchwi. Do genotypowania kolekcji 78 odmian populacyjnych marchwi typu zachodniego użyto 93 markery *DcS*-ILP opracowane przez Stelmach i in. (2017) oraz 2354 markery SNP równo rozdystrybuowane na dziewięciu chromosomach marchwi. Średnia H_O oraz średnia H_E była znacząco wyższa dla markerów SNP i wynosiła odpowiednio 0,323 oraz 0,295. Średnie wartości H_O i H_E uzyskane dla markerów *DcS*-ILP wynosiły odpowiednio 0,253 i 0,239. Na poziomie odmian wartości H_E charakteryzowały się dużą rozbieżnością: od 0,115 (odmiana LC1) do 0,323 (BE7) dla *DcS*-ILP oraz 0,174 (LO1) do 0,350 (GU3) dla SNP. Procent loci polimorficznych dla poszczególnych odmian wahał się od 31,18% (odmiana LC1) do 89,25% (BE7) dla markerów *DcS*-ILP oraz od 45,11% (LC1) do 91,29% (SV1) dla markerów SNP. Duży udział loci polimorficznych wskazuje na wysoki poziom wewnątrzodmianowej zmienności genetycznej (ang. *intra-cultivar variability*, IACV) kontrastującej z obserwowaną stabilnością odmian na poziomie fenotypowym. Średnie wartości współczynnika wsobności (F_{IS}) dla wszystkich analizowanych odmian były ujemne, co również wskazuje na wysoki poziom ICV. Analiza współczynnika utrwalenia (F_{ST}) wskazuje na umiarkowany poziom zróżnicowania genetycznego pomiędzy klasami reprezentującymi poszczególne kształty korzenia spichrzowego (dalej: klasy). Oba systemy markerowe wykazały największe zróżnicowanie pomiędzy odmianami (ang. *inter-cultivar variability,* IECV) reprezentującymi klasę Amsterdam i St. Valery (odpowiednio F_{ST} = 0,260 i F_{ST} = 0,214 dla *DcS*-ILP i SNP). Wartości H_O otrzymane dla poszczególnych klas były wyższe od F_{ST}. Na tej podstawie można wnioskować, że na zmienność całkowitą obserwowaną w obrębie danej klasy większy wpływ ma IACV niż IECV. Wyniki analizy AMOVA również wskazują na dominujący udział IACV

w całkowitej obserwowanej zmienności genetycznej (wartości IACV odpowiednio 71% i 68% dla *DcS*-ILP i SNP).

Analizy struktury zmienności genetycznej przeprowadzono metodą klastrowania Bayesowskiego z wykorzystaniem oprogramowania STRUCTURE, dla całej kolekcji 390 roślin reprezentujących 11 predefiniowanych klas. Wartości ΔK wskazały na obecność trzech, czterech lub siedmiu klastrów w przypadku genotypowania markerami *DcS*-ILP oraz trzech, czterech lub pięciu klastrów przy genotypowaniu markerami SNP. Wartości dla ΔK dla liczby klastrów zbliżonej do liczby analizowanych klas (np. K=10) były bardzo niskie, co wskazywało na grupowanie odmian przynależących do co najmniej kilku klas w obrębie jednego klastra. Przy genotypowaniu markerami SNP obserwowano niższy poziom admiksji. Przykładowo, przy założeniu istnienia trzech klastrów (K=3) 98,7% roślin zostało jednoznacznie przypisanych do jednego z klastrów na podstawie danych otrzymanych dla SNP, podczas gdy wartość ta była niższa dla *DcS*-ILP i wynosiła 78,2%. Odmiany reprezentujące klasę Amsterdam i Chantenay charakteryzowały się wysokimi wartościami Q i tworzyły odrębne klastry niezależnie od wykorzystanego systemu markerowego oraz zakładanej najbardziej prawdopodobnej liczby klastrów. Obserwowane wartości dystansu genetycznego pomiędzy osobnikami, mierzone jako H_E, były znacząco niższe dla klastrów grupujących odmiany klasy Amsterdam i Chantenay w porównaniu do pozostałych klastrów, wskazując na odrębność odmian reprezentujących te klasy.

Wyniki grupowania *de novo* metodą analizy dyskryminacyjnej głównych składowych (ang. *Discriminant Analysis of Principal Components*, DAPC) wskazały na istnienie ośmiu grup genetycznych (K=8) w obrębie analizowanej kolekcji odmian, niezależnie od zastosowanego systemu markerowego. Wartość K=8 wskazuje, podobnie jak w przypadku analizy STRUCTURE, na istnienie grup skupiających odmiany przynależące do co najmniej dwóch predefiniowanych klas. 51 spośród 78 odmian zostało przyporządkowanych do tych samych grup 1-8, dając 65,39% zgodności na poziomie odmianowym. 277 spośród 390 roślin zostało przyporządkowanych do tych samych grup, dając 71,03% zgodności na poziomie osobniczym. Trzy grupy odmian, tj. Amsterdam (grupa 1), Chantenay (grupa 3) i Imperator (grupa 6) charakteryzowały się niskim poziomem heterogeniczności. Zgodność przyporządkowania odmian w grupach 1 i 3 wynosiła 100%, natomiast w grupie 6 71%. Obie metody klastrowania wykazały istnienie stosunkowo jednorodnych grup odmian populacyjnych reprezentujących typ Amsterdam i Chantenay.

7. Podsumowanie

Wyniki uzyskane w ramach przedstawionego cyklu publikacji poszerzyły aktualny stan wiedzy dotyczący dystrybucji ruchomych elementów typu MITE w genomie marchwi oraz wskazały na potencjalną możliwość wykorzystania polimorfizmu insercji elementów *DcSto* w ocenie struktury zmienności genetycznej populacji marchwi.

Analizy *in silico* z wykorzystaniem narzędzia RelocaTE okazały się stosunkowo skuteczną i wydajną metodą identyfikacji miejsc insercji elementów *DcSto* w resekwencjonowanych genomach marchwi charakteryzujących się zmiennym pokryciem genomu. Ponad 70% skuteczność predykcji została potwierdzona poprzez walidację metodą PCR dla 39 losowo wybranych przewidywanych miejsc insercji. Zaobserwowano znaczne różnice w liczebności elementów przynależących do poszczególnych rodzin *DcSto* oraz zróżnicowaną dystrybucję w obrębie czterech wyodrębnionych puli genetycznych marchwi.

Insercje elementów *DcSto* obecne w obrębie intronów genomu referencyjnego marchwi posłużyły do opracowania markerów molekularnych opartych na polimorfizmie długości intronów (ILP). Użyteczność opracowanego panelu 90 markerów *DcS*-ILP w ocenie zmienności genetycznej zweryfikowano poprzez genotypowanie kolekcji marchwi dzikiej oraz uprawnej typu zachodniego. Wyniki klastrowania bayesowskiego oraz skalowania wielowymiarowego jednoznacznie wykazały istnienie dwóch odrębnych puli genetycznych obejmujących marchew dziką oraz uprawną. Odmiany reprezentujące marchew uprawną były zgrupowane w czterech klastrach. Wyniki grupowania w dużym stopniu odpowiadały historii hodowli marchwi typu zachodniego i wskazywały na możliwy związek pomiędzy strukturą zmienności genetycznej a obserwowanym kształtem korzenia spichrzowego. W kolejnym etapie genotypowaniu markerami *DcS*-ILP oraz SNP poddano znacznie liczniejszą kolekcję 390 roślin reprezentujących 78 odmian populacyjnych marchwi typu zachodniego. Wysoki udział loci polimorficznych, ujemne wartości współczynnika wsobności w połączeniu z wynikami analizy zmienności molekularnej wskazały na bardzo wysoki poziom zmienności wewnątrzodmianowej kontrastujący z obserwowaną stabilnością fenotypową odmian. Wyniki klastrowania metodą STRUCTURE oraz DAPC wykazały znaczną odrębność genetyczną odmian reprezentujących dwa typy korzenia, Amsterdam oraz Chantenay.

W wyniku przeprowadzonych badań opracowano następujące wnioski:

1. Detekcja *in silico* insercji ruchomych elementów genetycznych oparta na poszukiwaniu homologii pomiędzy sekwencją konsensusową transpozonu a surowymi odczytami stanowi wydajną i precyzyjną metodę identyfikacji transpozonów *DcSto* w genomie marchwi.

2. Obserwowany wysoki poziom polimorfizmu insercji *DcSto* w genomie marchwi wskazuje na stosunkowo niedawną aktywność transpozycyjną niektórych rodzin tych elementów. Analiza zmienności genetycznej wynikającej z obserwowanego polimorfizmu insercji 14 rodzin *DcSto* pozwala na wyodrębnienie czterech puli genetycznych marchwi odzwierciedlających zasięg występowania (marchew typu zachodniego i wschodniego) oraz status udomowienia (marchew dzika i uprawna).

3. Wyniki genotypowania markerami *DcS*-ILP oraz SNP stanowią według naszej wiedzy najpełniejszy obraz rzeczywistej struktury zmienności genetycznej marchwi uprawnej typu zachodniego i wskazują na wysoką zmienność wewnątrzodmianową oraz znaczący poziom admiksji obserwowanej w kolekcji odmian populacyjnych.

4. Struktura zmienności genetycznej odmian populacyjnych marchwi typu zachodniego nie stanowi odzwierciedlenia w klasyfikacji opartej na podobieństwie cech fenotypowych korzenia. Jedynie odmiany typu Amsterdam i Chantenay tworzą stosunkowo jednorodne grupy będące prawdopodobnie wynikiem stosunkowo restrykcyjnego doboru materiału rodzicielskiego w procesie hodowlanym.

5. Markery *DcS*-ILP posiadają wiele zalet typowych dla markerów SSR, m. in. kodominacyjny typ dziedziczenia, wysoką specyficzność oraz wysoki poziom powtarzalności równocześnie charakteryzując się szybszą, wygodniejszą i mniej kosztochłonną detekcją. Z tego względu mogą stanowić dobrą alternatywę dla markerów SSR. Opracowane panele markerów *DcS*-ILP oraz SNP stanowią wydajne i oszczędne narzędzia do analizy zmienności genetycznej marchwi.

8. Spis literatury

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipman, D. J. 1990. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3), 403–410

Banga, O. 1963. Origin and distribution of the western cultivated carrot. *Genetica Agraria*, 17, 357–370

Baranski, R., Maksylewicz-Kaul, A., Nothnagel, T., Cavagnaro, P. F., Simon, P. W., Grzebelus, D. 2012. Genetic diversity of carrot (*Daucus carota* L.) cultivars revealed by analysis of SSR loci. In *Genetic Resources and Crop Evolution*, 59, 163–170

Bureau, T. E., Wessler, S. R. 1992. *Tourist*: a large family of small inverted repeat elements frequently associated with maize genes. *The Plant Cell*, 4(10), 1283–1294

Bureau, T. E., Wessler, S. R. 1994. *Stowaway*: a new family of inverted repeat elements associated with the genes of both monocotyledonous and dicotyledonous plants. *The Plant Cell*, 6(6), 907–916

Casa, A. M., Nagel, A., Wessler, S. R. 2004. MITE display. *Methods in Molecular Biology*, 260, 175–188

Cavagnaro, P. F., Chung, S. M., Manin, S., Yildiz, M., Ali, A., Alessandro, M. S., Iorizzo, M., Senalik, D. A., Simon, P. W. 2011. Microsatellite isolation and marker development in carrot-genomic distribution, linkage mapping, genetic diversity analysis and marker transferability across *Apiaceae*. *BMC Genomics*, 12(1), 386

Chang, R.-Y., O'Donoughue, L. S., Bureau, T. E. 2001. Inter-MITE polymorphisms (IMP): a high throughput transposon-based genome mapping and fingerprinting approach. *Theoretical and Applied Genetics*, 102(5), 773–781

Chen, J., Hu, Q., Zhang, Y., Lu, C., Kuang, H. 2014. P-MITE: a database for plant miniature inverted-repeat transposable elements. *Nucleic Acids Research*, 42, 1176-81

Chin, C. S., Peluso, P., Sedlazeck, F. J., Nattestad, M., Concepcion, G. T., Clum, A., Dunn, C., O'Malley, R., Figueroa-Balderas, R., Morales-Cruz, A., Cramer, G. R., Delledonne, M., Luo, C., Ecker, J. R., Cantu, D., Rank, D. R., Schatz, M. C. 2016. Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods*, 13(12), 1050–1054

Clouse, J. W., Adhikary, D., Page, J. T., Ramaraj, T., Deyholos, M. K., Udall, J. A., Fairbanks, D. J., Jellen, E. N., Maughan, P. J. 2016. The amaranth genome: genome, transcriptome, and physical map assembly. *The Plant Genome*, 9(3)

Cornillot, E., Dassouli, A., Garg, A., Pachikara, N., Randazzo, S., Depoix, D., Carcy, B., Delbecq, S., Frutos, R., Silva, J. C., Sutton, R., Krause, P. J., Mamoun, C. 2013. Whole genome mapping and re-organization of the nuclear and mitochondrial genomes of *Babesia microti* isolates. *PLoS ONE*, 8(9)

Crescente, J. M., Zavallo, D., Helguera, M., Vanzetti, L. S. 2018. MITE Tracker: An accurate approach to identify miniature inverted-repeat transposable elements in large genomes. *BMC Bioinformatics*, 19(1), 348

Cuevas, H. E., Staub, J. E., Simon, P. W., Zalapa, J. E., McCreight, J. D. 2008. Mapping of genetic loci that regulate quantity of beta-carotene in fruit of US western shipping melon (*Cucumis melo* L.). *Theoretical and Applied Genetics*, 117, 1345–1359

Eddy, S. R. 2011. Accelerated profile HMM searches. *PLoS Computational Biology*, 7(10)

Eisen, J. A., Coyne, R. S., Wu, M., Wu, D., Thiagarajan, M., Wortman, J. R., Badger, J. H., Ren, Q., Amedeo, P., Jones, K. M., Tallon, L. J., Delcher, A. L., Salzberg, S. L., Silva, J. C., Haas, B. J., Majoros, W. H., Farzad, M., Carlton, J. M., Smith, R. K., … Orias, E. 2006. Macronuclear genome sequence of the ciliate *Tetrahymena thermophila*, a model eukaryote. *PLoS Biology*, 4(9), 1620–1642

Feschotte, C., Mouchès, C. 2000. Evidence that a family of miniature inverted-repeat transposable elements (MITEs) from the A*rabidopsis thaliana* genome has arisen from a pogo-like DNA transposon. *Molecular Biology and Evolution*, 17(5), 730–737

Grzebelus, D., Iorizzo, M., Senalik, D., Ellison, S., Cavagnaro, P., Macko-Podgorni, A., Heller-Uszynska, K., Kilian, A., Nothnagel, T., Allender, C., Simon, P. W., Baranski, R. 2014. Diversity, genetic mapping, and signatures of domestication in the carrot (*Daucus carota* L.) genome, as revealed by Diversity Arrays Technology (DArT) markers. *Molecular Breeding*, 33, 625–637

Han, Y., Wessler, S. R. 2010. MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Research*, 38(22), e199

Iorizzo, M., Ellison, S., Senalik, D., Zeng, P., Satapoomin, P., Huang, J., Bowman, M., Iovene, M., Sanseverino, W., Cavagnaro, P., Yildiz, M., Macko-Podgórni, A., Moranska, E., Grzebelus, E., Grzebelus, D., Ashrafi, H., Zheng, Z., Cheng, S., Spooner, D., … Simon, P. 2016. A high-quality carrot genome assembly provides new insights into carotenoid accumulation and asterid genome evolution. *Nature Genetics*, 48, 657–666

Iorizzo, M., Senalik, D. A., Ellison, S. L., Grzebelus, D., Cavagnaro, P. F., Allender, C., Brunet, J., Spooner, D. M., van Deynze, A., Simon, P. W. 2013. Genetic structure and domestication of carrot (*Daucus carota* subsp. *sativus*) (*Apiaceae*). *American Journal of Botany*, 100(5), 930–938

Jaillon, O., Aury, J. M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., Choisne, N., Aubourg, S., Vitulo, N., Jubin, C., Vezzi, A., Legeai, F., Hugueney, P., Dasilva, C., Horner, D., Mica, E., Jublot, D., Poulain, J., Bruyère, C., … Wincker, P. 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, 449(7161), 463–467

Jurka, J., Kapitonov, V. v., Pavlicek, A., Klonowski, P., Kohany, O., Walichiewicz, J. 2005. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research*, 110(1–4), 462–467

Kumar, A., Hirochika, H. 2001. Applications of retrotransposons as genetic tools in plant biology. *Trends in Plant Science*, 6, 127–134

Lee, J. K., Kwon, S. J., Park, K. C., Kim, N. S. 2005. Isaac-CACTA transposons: new genetic markers in maize and sorghum. *Genome*, 48(3), 455–460

Leitch, I. J., Beaulieu, J. M., Cheung, K., Hanson, L., Lysak, M. A., Fay, M. F. 2007. Punctuated genome size evolution in *Liliaceae*. *Journal of Evolutionary Biology*, 20(6), 2296–2308

Lessa, E. 1992. Rapid surveying of DNA sequence variation in natural populations. *Molecular Biology and Evolution,* 9(2), 323–330

Liu, Y., Tahir, M., Feng, J. W., Ding, Y., Wang, S., Wu, G., Ke, L., Xu, Q., Chen, L. L. 2019. Comparative analysis of miniature inverted-repeat transposable elements (MITEs) and long terminal repeat (LTR) retrotransposons in six *Citrus* species. *BMC Plant Biology*, 19(1), 140

Macko-Podgórni, A., Nowicka, A., Grzebelus, E., Simon, P. W., Grzebelus, D. 2013. *DcSto*: carrot Stowaway-like elements are abundant, diverse, and polymorphic. *Genetica,* 41, 255–267

Macko-Podgórni, A., Stelmach, K., Kwolek, K., Grzebelus, D. 2019. *Stowaway* miniature inverted repeat transposable elements are important agents driving recent genomic diversity in wild and cultivated carrot. *Mobile DNA*, 10(1), 47

Meyers, B. C., Tingey, S., Morgante, M. 2001. Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. *Genome Research*, 11(10), 1660–1676

Nouroz, F., Noreen, S., Heslop-Harrison, J. S. 2015. Evolutionary genomics of miniature inverted-repeat transposable elements (MITEs) in *Brassica*. *Molecular Genetics and Genomics*, 290(6), 2297–2312

Oki, N., Yano, K., Okumoto, Y., Tsukiyama, T., Teraishi, M., Tanisaka, T. 2008. A genome-wide view of miniature inverted-repeat transposable elements (MITEs) in rice, *Oryza sativa* ssp. *japonica*. *Genes & Genetic Systems*, 83(4), 321–329

Park, K. C., Lee, J. K., Kim, N. H., Shin, Y. B., Lee, J. H., Kim, N. S. 2003. Genetic variation in *Oryza* species detected by MITE-AFLP. *Genes & Genetic Systems*, 78(3), 235–243

Paterson, A. H., Bowers, J. E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., Haberer, G., Hellsten, U., Mitros, T., Poliakov, A., Schmutz, J., Spannagl, M., Tang, H., Wang, X., Wicker, T., Bharti, A. K., Chapman, J., Feltus, F. A., Gowik, U., … Rokhsar, D. S. 2009. The *Sorghum bicolor* genome and the diversification of grasses. *Nature*, 457, 551–556

Piriyapongsa, J., Jordan, I. K. 2008. Dual coding of siRNAs and miRNAs by plant transposable elements. *RNA*, 14(5), 814–821

Piriyapongsa, J., Mariño-Ramírez, L., Jordan, I. K. 2007. Origin and evolution of human microRNAs from transposable elements. *Genetics*, 176(2), 1323–1337

Sampath, P., Murukarthick, J., Izzah, N. K., Lee, J., Choi, H. I., Shirasawa, K., Choi, B. S., Liu, S., Nou, I. S., Yang, T. J. 2014. Genome-wide comparative analysis of 20 miniature inverted-repeat transposable element families in *Brassica rapa* and *B. oleracea*. *PLoS ONE*, 9(4), e94499

Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D. L., Song, Q., Thelen, J. J., Cheng, J., Xu, D., Hellsten, U., May, G. D., Yu, Y., Sakurai, T., Umezawa, T., Bhattacharyya, M. K., Sandhu, D., Valliyodan, B., … Jackson, S. A. 2010. Genome sequence of the palaeopolyploid soybean. *Nature*, 463, 178–183

Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T. A., Minx, P., Reily, A. D., Courtney, L., Kruchowski, S. S., Tomlinson, C., Strong, C., Delehaunty, K., Fronick, C., Courtney, B., … Wilson, R. K. 2009. The B73 maize genome: complexity, diversity, and dynamics. *Science*, 326(5956), 1112–1115

Sharma, H., Bhandawat, A., Rahim, M. S., Kumar, P., Choudhoury, M. P., Roy, J. 2020. Novel intron length polymorphic (ILP) markers from starch biosynthesis genes reveal genetic relationships in Indian wheat varieties and related species. *Molecular Biology Reports*, 47(5), 3485–3500.

Smit, A., Hubley, R., Green, P. 1996. RepeatMasker Open-3.0.

Stelmach, K., Macko-Podgórni, A., Allender, C., Grzebelus, D. 2021. Genetic diversity structure of western-type carrots. *BMC Plant Biology*, 21(1), 200

Stelmach, K., Macko-Podgórni, A., Machaj, G., Grzebelus, D. 2017. Miniature inverted repeat transposable element insertions provide a source of intron length polymorphism markers in the carrot (*Daucus carota* L.). *Frontiers in Plant Science*, 8, 725

Stolarczyk, J., Janick, J. 2011. Carrot: history and iconography. *Chronica Horticulturae*, 51(2), 13–18

Wicker, T. 2012. So many repeats and so little time: how to classify transposable elements. W: Grandbastien, M. A., Casacuberta, J. M. (eds.), *Plant Transposable Elements*. Springer-Verlag, New York

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P., Schulman, A. H. 2007. A unified classification system for eukaryotic transposable elements. *Nature reviews*. *Genetics*, 8, 973–982

Wydner, K. S., Sechler, J. L., Boyd, C. D., Passmore, H. C. 1994. Use of an intron length polymorphism to localize the tropoelastin gene to mouse chromosome 5 in a region of linkage conservation with human chromosome 7. *Genomics*, 23(1), 125–131

Ye, C., Ji, G., Liang, C. 2016. DetectMITE: A novel approach to detect miniature inverted repeat transposable elements in genomes. *Scientific Reports*, 6(1), 1–11

Yu, J., Hu, S., Wang, J., Wong, G. K. S., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y., Zhang, X., Cao, M., Liu, J., Sun, J., Tang, J., Chen, Y., Huang, X., Lin, W., Ye, C., Tong, W., … Yang, H. 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science*, 296(5565), 79–92

Zerjal, T., Joets, J., Alix, K., Grandbastien, M. A., Tenaillon, M. I. 2009. Contrasting evolutionary patterns and target specificities among three tourist-like MITE families in the maize genome. *Plant Molecular Biology*, 71(1–2), 99–114

9. Wydruki publikacji wchodzących w skład rozprawy doktorskiej

frontiers in Plant Science

METHODS
published: 09 May 2017
doi: 10.3389/fpls.2017.00725

Miniature Inverted Repeat Transposable Element Insertions Provide a Source of Intron Length Polymorphism Markers in the Carrot (*Daucus carota* L.)

*Katarzyna Stelmach, Alicja Macko-Podgórni, Gabriela Machaj and Dariusz Grzebelus **

Faculty of Biotechnology and Horticulture, Institute of Plant Biology and Biotechnology, University of Agriculture in Krakow, Krakow, Poland

The prevalence of non-autonomous class II transposable elements (TEs) in plant genomes may serve as a tool for relatively rapid and low-cost development of gene-associated molecular markers. Miniature inverted-repeat transposable element (MITE) copies inserted within introns can be exploited as potential intron length polymorphism (ILP) markers. ILPs can be detected by PCR with primers anchored in exon sequences flanking the target introns. Here, we designed primers for 209 *DcSto* (*Daucus carota Stowaway*-like) MITE insertion sites within introns along the carrot genome and validated them as candidate ILP markers in order to develop a set of markers for genotyping the carrot. As a proof of concept, 90 biallelic *DcS*-ILP markers were selected and used to assess genetic diversity of 27 accessions comprising wild *Daucus carota* and cultivated carrot of different root shape. The number of effective alleles was 1.56, mean polymorphism informative content was 0.27, while the average observed and expected heterozygosity was 0.24 and 0.34, respectively. Sixty-seven loci showed positive values of Wright's fixation index. Using Bayesian approach, two clusters comprising four wild and 23 cultivated accessions, respectively, were distinguished. Within the cultivated carrot gene pool, four subclusters representing accessions from Chantenay, Danvers, Imperator, and Paris Market types were revealed. It is the first molecular evidence for root-type associated diversity structure in western cultivated carrot. *DcS*-ILPs detected substantial genetic diversity among the studied accessions and, showing considerable discrimination power, may be exploited as a tool for germplasm characterization and analysis of genome relationships. The developed set of *DcS*-ILP markers is an easily accessible molecular marker genotyping system based on TE insertion polymorphism.

Keywords: *DcSto*, genetic diversity structure, ILP, *Stowaway*-like MITEs, TEs

OPEN ACCESS

Edited by:
Fulvio Cruciani, Sapienza University of Rome, Italy

Reviewed by:
Christian Parisod, University of Neuchâtel, Switzerland
Tina T. Hu, Princeton University, USA

****Correspondence:***
Dariusz Grzebelus
d.grzebelus@ogr.ur.krakow.pl

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Plant Science

Received: *20 February 2017*
Accepted: *19 April 2017*
Published: *09 May 2017*

Citation:
Stelmach K, Macko-Podgórni A, Machaj G and Grzebelus D (2017) Miniature Inverted Repeat Transposable Element Insertions Provide a Source of Intron Length Polymorphism Markers in the Carrot (Daucus carota L.). Front. Plant Sci. 8:725. doi: 10.3389/fpls.2017.00725

INTRODUCTION

Transposable elements (TEs) are segments of DNA that can move themselves to new chromosomal location. They are prevalent in the genomes of both prokaryotes and eukaryotes, and account for a great subsection of the genetic variation in plants and animals. Some plant genomes are composed of transposable elements in more than two thirds, as the 77% of the maize genome

(Meyers et al., 2001). Miniature inverted-repeat transposable elements (MITEs) are a special type of class II non-autonomous elements with a maximum of a few hundred base pairs in size (Hua-Van et al., 2005). Although they were first discovered in plant genomes (Bureau and Wessler, 1992, 1994), they have been also identified in a wide range of animal, eubacteria and archea genomes (Brügger et al., 2002; Feschotte et al., 2002). The two largest MITE families, *Stowaway* and *Tourist*, were identified as members of the *Tc1/Mariner* and the *PIF/Harbinger* superfamilies, respectively (Jiang et al., 2004). *Stowaway* MITEs were first described in the maize genome (Bureau and Wessler, 1994) as less than 500 bp long, forming a 2 bp TA TSD upon insertion. MITEs are usually present in many thousand copies per genome. 22,000 identified *Stowaway* MITEs were classified into 34 families in the *Oryza sativa* genome (Feschotte et al., 2003), whereas 18,000 MITE insertions were classified into 18 families in the *Triticum* spp. genome (Yaakov et al., 2013).

The ubiquity, genome-wide distribution and high copy numbers have provided genetic markers from both class I and class II TEs (Kumar and Hirochika, 2001). The abundance of MITE copies makes them highly useful source of polymorphism. To date, MITE Transposon Display (MITE-TD) and Inter-MITE Polymorphism (IMP) techniques exploiting the TIR sequences in *Oryza sativa*, *Zea mays*, *Sorghum bicolor, Hordeum vulgare*, and *Daucus carota* MITEs, have been developed (Chang et al., 2001; Park et al., 2003; Casa et al., 2004; Lee et al., 2005; Grzebelus et al., 2007). Some *Stowaway* MITEs identified to date were described as being preferentially inserted or retained in genic regions (Casa et al., 2000; Jiang et al., 2003). However, even though 54% of *DcSto* insertion sites in the carrot genome were located less than 2 kb away from or inside the coding sequences, random distribution of *DcSto* rather than preferential insertions around genes was proposed (Iorizzo et al., 2016).

Insertions within introns may provide a significant polymorphism. Intron polymorphisms, particularly intron length polymorphisms (ILPs), can be exploited as genetic markers used for gene mapping (Wydner et al., 1994) and population genetic surveys (Lessa, 1992). ILP takes advantage of the different rate of evolution of exons and introns that can result in conserved exon nucleotide sequences adjoined to more variable intron sequences. ILP can be detected by the polymerase chain reaction with a pair of primers anchored in the exons flanking the intron of interest (Wang et al., 2005). ILP markers are unique due to their gene-specifity, codominancy, conveniency, reliability and cost-efficiency. Furthermore, ILPs are characterized by high transferability among related plant species (Yang et al., 2007; Gupta et al., 2011). To date, studies on the development of ILP markers in plants have been restricted

TABLE 1 | Description of plant material used in the present study.

Number	Accession	Species	Cultivar name	Root type	Origin	Source	
1	RS33	*Daucus carota* subsp. *sativus*	Chantenay Royal	Chantenay	FRA	HRIGRU	8860
2	RS34	*Daucus carota* subsp. *sativus*	Chantenay Red Cored	Chantenay	GBR	HRIGRU	8847
3	RS35	*Daucus carota* subsp. *sativus*	Royal Chantenay	Chantenay	USA	HRIGRU	3882
4	RS37	*Daucus carota* subsp. *sativus*	Gold King	Chantenay	USA	HRIGRU	5127
5	RS39	*Daucus carota* subsp. *sativus*	Chantenay Long Type	Chantenay	USA	HRIGRU	5090
6	RS41	*Daucus carota* subsp. *sativus*	Chantenay Rex RS	Chantenay	NLD	HRIGRU	5589
7	RS43	*Daucus carota* subsp. *sativus*	Danvers 126	Danvers	GBR	HRIGRU	6487
8	RS44	*Daucus carota* subsp. *sativus*	Danvers Danro RS	Danvers	NLD	HRIGRU	5595
9	RS45	*Daucus carota* subsp. *sativus*	Danvers Red Cored	Danvers	USA	HRIGRU	5128
10	RS49	*Daucus carota* subsp. *sativus*	Danvers	Danvers	NLD	HRIGRU	11144
11	RS50	*Daucus carota* subsp. *sativus*	Danvers Pride	Danvers	USA	HRIGRU	8098
12	RS51	*Daucus carota* subsp. *sativus*	Danvers Half Long	Danvers	USA	HRIGRU	8109
13	RS56	*Daucus carota* subsp. *sativus*	Paris Market	Paris Market	NLD	HRIGRU	5596
14	RS57	*Daucus carota* subsp. *sativus*	Paris Forcing	Paris Market	GBR	HRIGRU	3966
15	RS59	*Daucus carota* subsp. *sativus*	French Forcing Horn	Paris Market	GBR	HRIGRU	6489
16	RS60	*Daucus carota* subsp. *sativus*	Parijse Market	Paris Market	—	HRIGRU	9294
17	RS62	*Daucus carota* subsp. *sativus*	Parijse Market (Rubin)	Paris Market	—	HRIGRU	9296
18	RS71	*Daucus carota* subsp. *sativus*	Gold Pak	Imperator	USA	HRIGRU	3885
19	RS72	*Daucus carota* subsp. *sativus*	Imperator 408	Imperator	USA	HRIGRU	3907
20	RS73	*Daucus carota* subsp. *sativus*	Imperator	Imperator	NLD	HRIGRU	11145
21	RS74	*Daucus carota* subsp. *sativus*	Imperator 407	Imperator	USA	HRIGRU	3891
22	RS75	*Daucus carota* subsp. *sativus*	Long Imperator 58	Imperator	USA	HRIGRU	3917
23	RS76	*Daucus carota* subsp. *sativus*	Imperator 58	Imperator	USA	HRIGRU	3892
24	CDS15	*Daucus carota* subsp. *azoricus*	–	–	ESP	HRIGRU	6667
25	CDS39	*Daucus carota* subsp. *carota*	–	–	CHE	HRIGRU	9226
26	CDS93	*Daucus carota* subsp. *carota*	–	–	USA	USDA	–
27	CDS40	*Daucus carota* subsp. *carota*	–	–	POL	HRIGRU	9270

to few species (Wang et al., 2005; Huang et al., 2008; Chen et al., 2010; Gupta et al., 2011, 2012; Li et al., 2013; Muthamilarasan et al., 2014).

Carrot is the most widely grown member of Apiaceae family. Its progenitor, wild *Daucus carota* L., is a plant commonly occurring in the temperate climatic zones. To date, a range molecular tools facilitating genome analysis in context of evolutionary history of wild and cultivated carrot have been developed, i.e., DArT, SSR, and SNP markers (Cavagnaro et al., 2011; Iorizzo et al., 2013; Grzebelus et al., 2014) and a set of ca. 30 resequenced genomes (Iorizzo et al., 2016). The analyses showed clear evidence for the carrot germplasm separation into three distinct groups of wild, western cultivated (European and American germplasm) and eastern cultivated (Asian germplasm) carrot. The majority of modern cultivars belong to the western group. Several varietal types were distinguished within western carrots, based primarily on the storage root shape and size (Prohens and Nuez, 2008). Despite apparent phenotypic differences, previous studies have indicated absence of any apparent population structure in western carrots, suggesting no significant genetic separation among these varietal types (Bradeen et al., 2002; Iorizzo et al., 2013).

In this study, we performed (1) a genome-wide search for *DcSto* (*Daucus carota Stowaway*-like) MITE insertion-based intron length polymorphism markers, and (2) validation of candidate ILP markers in order to develop a panel for genotyping the carrot by means of applying a simple, cost- and time-efficient polymerase chain reaction.

MATERIALS AND METHODS

Plant Materials

Twenty eight carrot accessions comprising four wild carrots of different origin, 23 western type carrot cultivars representing four types of root shape and a DH1 plant (Iorizzo et al., 2016) as the reference, were used for ILP validation (**Table 1**). Total genomic DNA was isolated from fresh young leaves using commercial DNeasy Plant Mini Kit (Qiagen) and used as the template for PCR amplification.

Development of ILP Markers

Coordinates of 4028 *DcSto* insertions belonging to 14 families were compared to coordinates of ca. 32 thousand genes annotated in the carrot reference DH1 genome assembly (Iorizzo et al., 2016; NCBI accession LNRQ01000000). 609 gene-associated *DcSto* insertion sites localized in introns were identified, of which 209 were manually selected for development of ILP markers. The criteria for initial selection were as followed: insertion sites were (1) free from any other annotated repetitive sequences, (2) present in introns not longer than 3.7 Kb, and (3) evenly distributed over each chromosome. Primer3 (Untergasser et al., 2012) and Primer-BLAST (Ye et al., 2012) were used to design PCR primer pairs anchored in exons flanking introns harboring the selected *DcSto* insertions. Primer pairs were designed to amplify fragments in a 400–3,700-bp range. The optimal annealing temperature was set to 58°C; and the size and GC content ranged from 18 to 23 bases and 40 to 60%, respectively.

Validation and Evaluation of *DcS*-ILP Markers

Candidate ILP markers were selected for experimental evaluation. Amplification was carried out in a 10 μL total

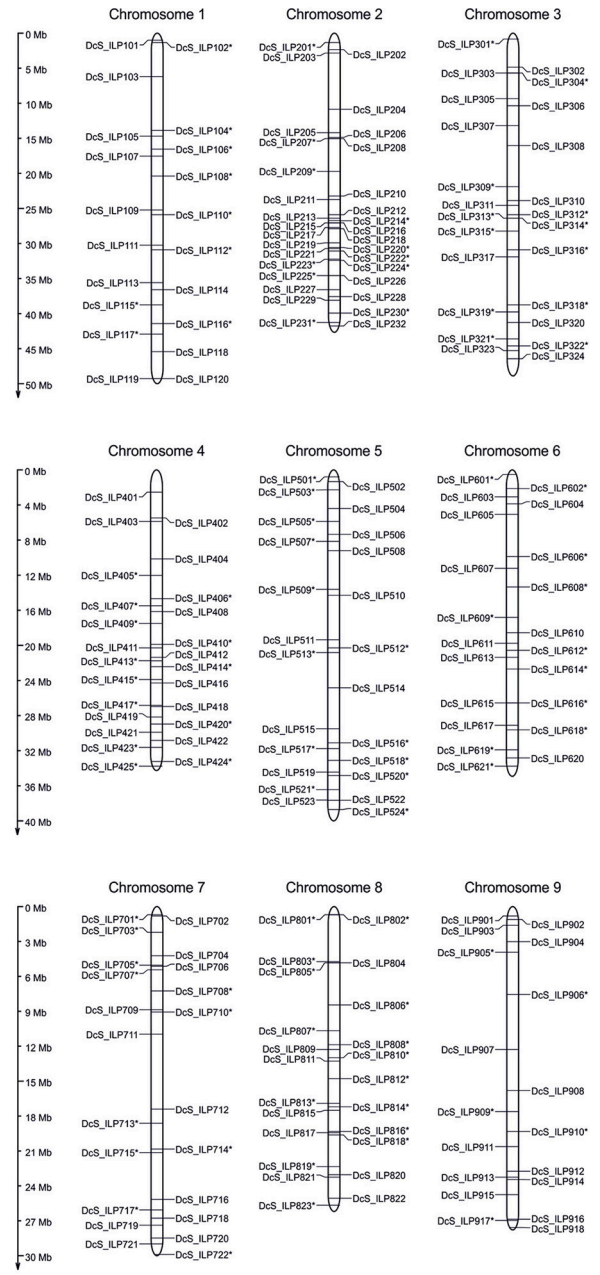


FIGURE 1 | Physical genomic distribution of the 209 developed *DcS*-ILP markers on nine chromosomes of the carrot genome. The vertical bars correspond to the position of introns harboring *DcSto* insertions, selected for a development of ILP markers. Positively validated markers are marked by asterisk.

volume containing 20 ng of genomic DNA, 0.5 μM each of forward and reverse primer, 0.25 mM of each dNTP (Thermo Fisher Scientific), 0.5 U Taq DNA polymerase (Thermo Fisher Scientific) and 1x Taq buffer. The PCR amplifications were performed in an Eppendorf MasterCycler Gradient using the following thermal profile: 94°C (120 s), 30 cycles of 94°C (30 s), 56°C (30 s), 68°C (120 s) and final step of 68°C (600 s). For primers generating ambiguous profiles, the annealing temperature was adjusted to 58, 59, or 60°C. PCR products were separated in 1% agarose gels run in 1x Tris-borate-EDTA buffer (pH 8.0) at a constant current of 5V/cm for about 2 h, stained with Midori Green (Nippon Genetics) and analyzed using GelDoc-It imaging system (UVP). GeneRuler 1 kb and 100 bp$^+$ DNA Ladders (Thermo Fisher Scientific) were used to determine product sizes for each locus. The amplicons representing additional local rearrangements within introns were excised, purified using GenJET™ Gel Extraction Kit (Thermo Fisher Scientific), cloned into T/A cloning vector (Promega Corporation) and transformed into *Escherichia coli,* strain DH10B. Up to five recombinant colonies were selected and cultured overnight at 37°C in culture tubes containing 5 mL of Luria–Bertani medium and ampicillin (100 mg/L). Plasmids were purified using Wizard SV Minipreps KIT (Promega Corporation). Sequencing reactions were set up with universal primers sp6 and T7 using Big Dye terminator chemistry (Applied Biosystems), as recommended by manufacturer. Sequencing was carried out on ABI 3700 capillary sequencer (Applied Biosystems).The sequences were manually edited using BioEdit (Hall, 1999) and aligned to the sequences of predicted genes for which ILP primers were designed.

Recording of Electrophoretic Bands and Statistical Data Analysis

The ILP marker profiles were scored manually. Each allele was scored as: 1 (empty insertion site), 2 (occupied insertion site) or 0 (lack of amplification).The codominant marker matrix with diploid individuals was created (Supplementary Table 1) and used in GenAlEx 6.5 (Peakall and Smouse, 2006) for creating genetic distance matrix and analysis of molecular variance (AMOVA). Expected and observed heterozygosity (H_e and H_o), and fixation index (F_{IS}) were computed using POPGENE 1.32 (Yeh et al., 2000). Polymorphism informative content (PIC) of *n*-allele locus, an indicator of a genetic marker's usefulness introduced by Botstein et al. (1980), was calculated as: $PIC = 1 - \sum_{i=1}^{n} p_i^2 - \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} 2p_i^2 p_j^2$, where p_i and p_j are the population frequency of the *i*th and *j*th allele. Genetic structure was inferred using Bayesian model-based software STRUCTURE 2.2.3 (Pritchard et al., 2008) without information on the accession origin. Ten independent iterations with an admixture and correlated allele frequencies model were performed. The length of the burn-in period and the number of Markov Chain Monte Carlo (MCMC) replications after the burn-in were assigned at 10^5 for each number of clusters (K) set from 1 to 27 and 1 to 23 for further subclustering. The estimation of K was provided by joining the log probability of data [LnP(D)] from STRUCTURE output and an *ad hoc* statistics ΔK (Evanno et al., 2005) based on the second rate of change of the log probability of data with respect to the number of clusters. In addition, CLUMPAK software (Kopelman et al., 2015) was used to confirm the selection of the best K. Based on the chosen K, each carrot accession was assigned to a subpopulation for which its membership value (Q) was higher than 0.6. AMOVA was performed using GenAlEx 6.5 to evaluate differentiation among the subpopulations. Principal coordinate analysis (PCoA) was conducted to visualize genetic diversity of the studied accessions.

RESULTS

Development and Validation of the Candidate ILP Markers

Insertion sites of 209 *DcSto* MITEs within introns of annotated genes were chosen to develop *Daucus carota Stowaway*-like Intron Length Polymorphism (*DcS*-ILP) markers evenly distributed throughout the genome (**Figure 1**). The number of *DcSto* insertion sites evaluated per chromosome varied from 18 (chromosome 9) to 32 (chromosome 2), with an average of 23.22. Their density ranged from 1.37 (chromosome 2) to 2.57 per Mb (chromosome 1), with an average of 1.76.

Upon PCR amplification, 100 of the 209 sites showed the expected *DcSto* insertion-based polymorphism, however, in case of 10 sites at least one additional amplicon was present in at least one accession (**Figure 2**). Sequencing of those amplicons revealed that none of the additional variants was related to the activity of the *DcSto* copy present in the reference genome (data not shown). Of the remaining 109 sites, six did not amplify efficiently; 32 were monomorphic for all tested plants; 13 showed a complex pattern resulting from nonspecific amplification, whereas 58 yielded polymorphic products not associated with *DcSto* insertions (i.e.,

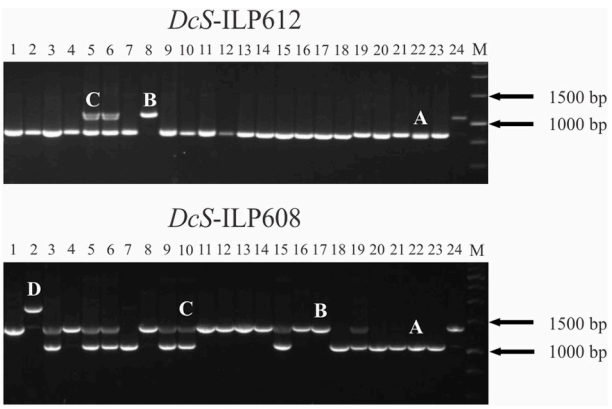


FIGURE 2 | Amplification of *DcS*-ILP612 and *DcS*-ILP608 markers in 23 carrot cultivars and the reference genome. Carrot accessions from 1 to 24 are listed in **Table 1**. *DcS*-ILP612—amplification of two alleles corresponding to empty **(A)** and occupied **(B)** *DcSto* insertion site and heterozygote **(C)**; *DcS*-ILP608—amplification of an additional allele **(D)** resulting from an unclassified rearrangement within the intron. M, 1 kb DNA Ladder.

TABLE 2 | Results of the experimental validation of developed candidate *DcS*-ILP markers.

Chromosome	Number of insertion sites	Validated insertion sites					
		Polymorphic with two allelic variants resulting from *DcSto* insertion	Polymorphic with two allelic variants resulting from *DcSto* insertion and an additional variant	Polymorphic with many allelic variants not associated with *DcSto* insertion	Complex amplification pattern	Monomorphic	No amplification
1	20	9	–	7	1	2	1
2	32	11	–	12	2	6	1
3	24	8	4	4	3	4	1
4	25	11	2	6	1	4	1
5	24	10	3	6	1	3	1
6	21	10	1	6	1	3	–
7	22	11	–	5	1	5	–
8	23	15	–	7	–	1	–
9	18	5	–	5	3	4	1
Total	209	90	10	58	13	32	6

TABLE 3 | The intron length-based classification of candidate *DcS*-ILP markers.

Marker class	The range of intron lengths [bp]	Number of candidate *DcS*-ILP markers	Number of positively validated *DcS*-ILP markers
I	400–1,000	75	34
II	1,001–1,600	80	34
III	1,601–2,200	27	15
IV	2,201–2,800	22	7
V	2,801–3,400	4	0
VI	>3,401	1	0

sizes of PCR products did not correspond to the expected sizes of empty or occupied variants) (**Table 2**).

The length of introns harboring the selected *DcSto* insertions varied from 449 to 3,637 bp. Based on the length of amplified introns, the developed markers were divided into six classes; I to V with intron size ranging from 400 to 3,400 bp, each at 600-bp interval, and class VI comprising introns longer than 3,400 bp (**Table 3**). Introns belonging to classes I to IV comprised 97.6% of all the developed markers. Class I and II markers were the most numerous, whereas class III markers showed the highest (55.6%) successful amplification rate indicating the most suitable length of introns considered for ILP markers. DcS-ILP markers of class V and VI were characterized by ambiguous amplification patterns, therefore not considered for further analyses.

Finally, 90 *DcS*-ILP (Supplementary Table 2) markers showing biallelic *DcSto* insertion polymorphism (**Figure 2**) were chosen for development of a panel for genotyping the carrot.

Assessment of Genetic Diversity

The utility of 90 biallelic *DcS*-ILP markers was verified by estimating the genetic diversity of the collection of 27 *D. carota* accessions comprising 23 cultivated and 4 wild populations. In total, 180 alleles were identified with an average of 2.0 per locus. 2.78% of the alleles were rare (frequency <0.05) and the mean effective number of alleles was 1.56. The observed heterozygosity for individual loci ranged from 0.04 to 0.56, with an average of 0.24, whereas the expected heterozygosity ranged from 0.04 to 0.51, with an average of 0.34. Shannon's index was from 0.09 to 0.69, with an average of 0.50. Among all the loci analyzed with the Wright's fixation index, 67 were positive. The PIC values ranged from 0.04 to 0.37, with an average of 0.27 (Supplementary Table 1).

STRUCTURE analysis based on 90 loci representing *DcSto* insertion-derived polymorphisms was performed to evaluate genetic structure of the 27 accessions. The value of ΔK statistics was the highest when two clusters were assumed [$\Delta K_{(2)} = 297.64$]. The increase in the number of assumed clusters resulted in low ΔK value [$\Delta K_{(>2)} = 0.01–52.35$]. Twenty three cultivated accessions were assigned to cluster 1 (C1) with membership coefficients (Q) ranging between 0.831 and 0.997, whereas cluster 2 (C2) comprised exclusively wild accessions with the *Q* value of 0.965–0.998 (**Figure 3A**). The level of genetic diversity within C1 (0.31) was slightly higher than within C2 (0.29).

To evaluate the genetic structure of the 23 cultivated accessions further subclustering was performed on the accessions assigned to C1. The highest ΔK was observed for $K = 21$ [$\Delta K_{(21)} = 22.77$], $K = 2$ [$\Delta K_{(2)} = 17.33$] and $K = 4$ [$\Delta K_{(4)} = 14.55$]. ΔK values for $K = 3$, $K = 5–20$ and $K = 22–23$ were not significant ($\Delta K = 0.164–4.16$). The mean value of log probability of the data was higher for $K = 4$ than for $K = 21$, and $K = 2$ [$\mathrm{LnP(D)}_{K=4} = -1891.7$, $\mathrm{LnP(D)}_{K=2} = -1922.5$, $\mathrm{LnP(D)}_{K=21} = -2703.2$], therefore four subclusters were chosen as the most probable genetic structure of the studied cultivated accessions. With $K = 4$, three accessions were assigned to subcluster SC1 with Q ranging between 0.928 and 0.962, six to subcluster SC2 with Q between 0.746 and 0.908, five to subcluster SC3 with Q between 0.825 and 0.954 and five to subcluster SC4 with Q between 0.782 and 0.922 (**Figure 3B**). Four accessions, namely Chantenay Red Cored, Chentenay Rex RS, Danvers 126, and Danvers could not be assigned to any of the subclusters due

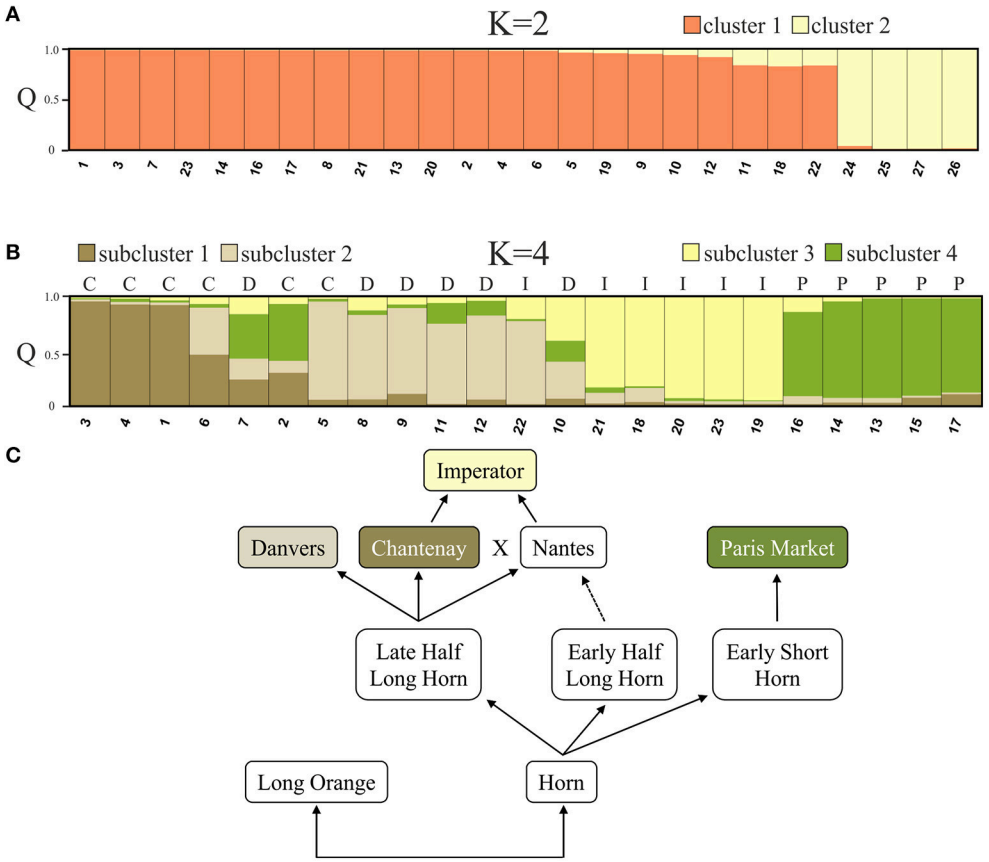


FIGURE 3 | The genetic structure of the studied 27 accessions based on a Bayesian approach assuming two clusters comprising cultivated (cluster 1) and wild (cluster 2) accessions, exclusively (A). The analysis of the genetic structure within first cluster resulted in forming four subclusters, generally comprising accessions representing each of described storage root shapes: C, Chantenay; D, Danvers; I, Imperator; P, Paris Market **(B)**. Assumed four gene pools reflect their breeding history as proposed by Banga (1963) **(C)**. The numbers of accessions correspond to those listed in **Table 1**.

TABLE 4 | The proportion of membership coefficients (Q) of each population defined by the type of root in each of the four subclusters.

Population name	Q proportion for four assumed subclusters				Number of accessions assigned to defined population
	SC1	SC2	SC3	SC4	
Chantenay	0.605	0.253	0.031	0.111	6
Danvers	0.082	0.626	0.136	0.155	6
Imperator	0.014	0.175	0.786	0.024	6
Paris market	0.043	0.034	0.039	0.884	5

to high level of admixture ($Q < 0.6$). The overall Q proportion of each of the four types clearly distinguished ($Q > 0.6$) the membership of Chantenay root type in SC1 ($Q = 0.605$), Danvers root type in SC2 ($Q = 0.626$), Imperator root type in SC3 ($Q = 0.785$), and Paris Market root type in SC4 ($Q = 0.884$) (**Table 4**).

AMOVA attributed 19% ($P = 0.001$) of the total genetic diversity to variation among the root types. The diversity of the 23 cultivated accessions was revealed by PCoA (**Figure 4**). Using the first three axes 31.7% of the total variation could be explained, with the 1st, 2nd, and 3rd axes explaining 12.1, 10.4, and 9.2%, respectively.

The above results suggested four separate groups in the collection of 23 cultivated carrots and the grouping generally corresponded with a postulated breeding history of western carrot types presented by Banga (1963), indicating that Chantenay and Danvers types originated from the Late Half Long Horn group, while Paris Market type descended from the Early Short Horn group. Both historical groups differ in terms of their storage root shape and earliness. In turn, the origin Imperator type was traced back to a cross between Chantenay and Nantes (**Figure 3C**).

DISCUSSION

In the present study, we took advantage of intron length polymorphisms resulting from retained *DcSto* insertions in order to develop a set of ILP markers in the carrot. The *DcSto* elements used in the study comprised mostly two families, *DcSto6* and *DcSto1*, the most numerous in the carrot genome and showing

FIGURE 4 | PCoA of 23 carrot accessions based on 90 *DcS*-ILP markers.

high percentage of insertions within coding regions (20 and 12%, respectively) (Iorizzo et al., 2016). The ubiquity of *DcSto* elements facilitated the selection of evenly distributed insertion sites for analysis, as well as equal coverage of the genome with the developed markers. 62.7% of the candidate markers were successfully amplified and 47.8% of them identified *DcSto* insertion polymorphisms. The success of amplification rate was lower in comparison with ILP markers in other plants, such as *Vigna unguiculata* (89%; Gupta et al., 2012), *Glycine max* (88.2%; Shu et al., 2010), *Solanum lycopersicum* (71%; Wang et al., 2010), probably as a result of high percentage of ambiguous amplification of introns longer than 2,200 bp. The length of intron is considered the main cause of PCR failure and generally, the successful amplification rate decreases with greater length of intron (Wang et al., 2010; Gupta et al., 2012). Polymorphism information content (PIC) has become the most widely used formula to measure the information content of molecular markers (Nagy et al., 2012). The mean PIC value of *DcS*-ILPs obtained for the studied *Daucus carota* accessions was higher compared to many of the developed ILP markers, e.g., *Setaria italica* (Gupta et al., 2011) and *Hevea brasiliensis* (Li et al., 2013), and comparable to study of Gupta et al. (2012) where 16 CILP loci were analyzed in 10 *Vigna unguiculata* accessions, with an average of 2.0 alleles per locus, and PIC value of 0.34. Differences in PIC values might be attributed to the various numbers of markers and accessions exploited in these studies. The average PIC value obtained in study of Huang et al. (2010), where 103 ILP loci were analyzed in 36 *Oryza sativa* accessions, was considerably higher (0.44) due to the higher number of alleles identified by rice ILPs (2.29 alleles per locus). As expected, the mean PIC value of the codominant *DcS*-ILPs was lower than the one obtained for the genomic SSR markers developed for the carrot (Rong et al., 2010; Cavagnaro et al., 2011). Similar results were reported for the comparative analysis of genetic diversity in *Oryza sativa* using ILP and genomic SSR markers (Huang et al., 2010). The developed *DcS*-ILPs showed discriminatory power comparable to that of dominant markers, e.g., DArT (Grzebelus et al., 2014). The values of Wright's fixation index which were significantly higher than zero, as well as the lower mean value of observed heterozygosity, indicated an excess of homozygous allelic states expected in advanced cultivars. *DcS*-ILP-based analysis of genetic structure of the studied accessions showed clear differentiation of wild and cultivated carrot, supporting earlier observations based on DArT, SSR and SNP genotyping (Cavagnaro et al., 2011; Iorizzo et al., 2013; Grzebelus et al., 2014). Bayesian clustering, on both accession and pre-defined population levels, revealed the presence of four gene pools that generally could be attributed to the shape of the storage root, namely: (1) Chantenay, (2) Danvers, (3) Imperator, and (4) Paris Market, and corresponding to their breeding history, as proposed by Banga (1963) (**Figures 3B,C**). Having said that, a substantial level of admixture was apparent for few investigated cultivars, possibly resulting from inter-type crosses aiming to derive an intermediate root morphology, e.g., longer or shorter roots. On the other hand, clear separation between the Paris Market type cultivars and the remaining three types confirms the postulated origin of the former from the Early Short Horn gene pool, opposed to Danvers and Chantenay types originating from the Late Half Long Horn gene pool. It is the first molecular evidence for a possible root-type associated structure of genetic diversity in western cultivated carrot. Nonetheless, a more extensive study ought to be conducted in order to substantiate this hypothesis. The results of PCoA were mostly

consistent with Bayesian clustering indicating the presence of the above-mentioned genetic structure.

CONCLUSION

In this study, we showed that the abundance of class II transposable elements may serve as a tool for relatively rapid and low-cost development of gene-derived molecular markers for effective use in carrot genotyping studies. *DcSto* insertion-derived ILP markers detect substantial variation among carrot plants of different origin and can be exploited in germplasm characterization and analysis of genome relationships. In addition, *DcS*-ILP markers directly reflect the variation within the genes and could be potentially useful in gene tagging and genetic map construction. ILP markers share many advantages of SSR markers, i.e., codominant nature, locus specificity and high reproducibility, but provide more convenient and rapid detection. To our knowledge, the *DcS*-ILP markers developed in this study are a novel set of publicly available transposon-based markers in the carrot.

AUTHOR CONTRIBUTIONS

AM, DG, and KS designed the study; KS, AM, and GM developed *DcS*-ILP markers; KS performed the validation of candidate *DcS*-ILP markers and the assessment of genetic diversity; KS, DG, AM, and GM drafted sections of the manuscript; KS and DG prepared the final version of the paper. All authors read, reviewed and approved the manuscript.

FUNDING

The research was financed from funds for basic research on crop improvement granted by the Polish Ministry of Agriculture and Rural Development in the years 2014–2016.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpls.2017.00725/full#supplementary-material

REFERENCES

Banga, O. (1963). *Main Types of the Western Carotene Carrot and Their Origin*. Zwolle: N.V. Uitgevers-Maatschappij W.E.J. Tjeenk Willink.

Botstein, D., White, R. L., Skolnick, M., and Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.* 32, 314–331.

Bradeen, J. M., Bach, I. C., Briard, M., le Clerc, V., Grzebelus, D., Senalik, D. A., et al. (2002). Molecular diversity analysis of cultivated carrot (*Daucus carota* L.) and wild *Daucus* populations reveals a genetically nonstructured composition. *J. Amer. Soc. Hort. Sci.* 127, 383–391.

Brügger, K., Redder, P., She, Q., Confalonieri, F., Zivanovic, Y., and Garrett, R. A. (2002). Mobile elements in archaeal genomes. *FEMS Microbiol. Lett.* 206, 131–141. doi: 10.1016/S0378-1097(01)00504-3

Bureau, T. E., and Wessler, S. R. (1992). *Tourist*: a large family of small inverted repeat elements frequently associated with maize genes. *Plant Cell* 4, 1283–1294. doi: 10.1105/tpc.4.10.1283

Bureau, T. E., and Wessler, S. R. (1994). *Stowaway*: a new family of inverted repeat elements associated with the genes of both monocotyledonous and dicotyledonous plants. *Plant Cell Online* 6, 907–916. doi: 10.1105/tpc.6.6.907

Casa, A. M., Brouwer, C., Nagel, A., Wang, L., Zhang, Q., Kresovich, S., et al. (2000). The MITE family *Heartbreaker* (*Hbr*): molecular markers in maize. *Proc. Natl. Acad. Sci. U.S.A.* 97, 10083–10089. doi: 10.1073/pnas.97.18.10083

Casa, A. M., Nagel, A., and Wessler, S. R. (2004). MITE display. *Methods Mol. Biol.* 260, 175–188. doi: 10.1385/1-59259-755-6:175

Cavagnaro, P. F., Chung, S.-M., Manin, S., Yildiz, M., Ali, A., Alessandro, M. S., et al. (2011). Microsatellite isolation and marker development in carrot - genomic distribution, linkage mapping, genetic diversity analysis and marker transferability across Apiaceae. *BMC Genomics* 12:386. doi: 10.1186/1471-2164-12-386

Chang, R.-Y., O'Donoughue, L. S., and Bureau, T. E. (2001). Inter-MITE polymorphisms (IMP): a high throughput transposon-based genome mapping and fingerprinting approach. *TAG Theor. Appl. Genet.* 102, 773–781. doi: 10.1007/s001220051709

Chen, X., Zhang, G., and Wu, W. (2010). Investigation and utilization of intron length polymorphisms in conifers. *New For.* 41, 379–388. doi: 10.1007/s11056-010-9229-5

Evanno, G., Regnaut, S., and Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol. Ecol.* 14, 2611–2620. doi: 10.1111/j.1365-294X.2005.02553.x

Feschotte, C., Swamy, L., and Wessler, S. R. (2003). Genome-wide analysis of mariner-like transposable elements in rice reveals complex relationships with *Stowaway* miniature inverted repeat transposable elements (MITEs). *Genetics* 163, 747–758.

Feschotte, C., Zhang, X., and Wessler, S. R. (2002). "Minature inverted-repeat transposable elements (MITEs) and their relationship with established DNA transposons," in *Mobile DNA II*, eds L. Craig, R. Craigie, M. Gellert, and A. Lambowitz (Washington, DC: American Society for Microbiology Press), 1147–1158.

Grzebelus, D., Iorizzo, M., Senalik, D., Ellison, S., Cavagnaro, P., Macko-Podgorni, A., et al. (2014). Diversity, genetic mapping, and signatures of domestication in the carrot (*Daucus carota* L.) genome, as revealed by Diversity Arrays Technology (DArT) markers. *Mol. Breed.* 33, 625–637. doi: 10.1007/s11032-013-9979-9

Grzebelus, D., Jagosz, B., and Simon, P. W. (2007). The *DcMaster* transposon display maps polymorphic insertion sites in the carrot (*Daucus carota* L.) genome. *Gene* 390, 67–74. doi: 10.1016/j.gene.2006.07.041

Gupta, S., Bansal, R., and Gopalakrishna, T. (2012). Development of intron length polymorphism markers in cowpea [*Vigna unguiculata* (L.) Walp.] and their transferability to other *Vigna* species. *Mol. Breed.* 30, 1363–1370. doi: 10.1007/s11032-012-9722-y

Gupta, S., Kumari, K., Das, J., Lata, C., Puranik, S., and Prasad, M. (2011). Development and utilization of novel intron length polymorphic markers in foxtail millet (*Setaria italica* (L.) P. Beauv.). *Genome* 54, 586–602. doi: 10.1139/g11-020

Hall, T. (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleid Acids Symp. Ser.* 41, 95–98.

Huang, M., Xie, F., Chen, L., Zhao, X., Jojee, L., and Madonna, D. (2010). Comparative analysis of genetic diversity and structure in rice using ILP and SSR markers. *Rice Sci.* 17, 257–268. doi: 10.1016/S1672-6308(09)60025-1

Huang, X., Lu, G., Zhao, Q., Liu, X., and Han, B. (2008). Genome-wide analysis of transposon insertion polymorphisms reveals intraspecific variation in cultivated rice. *Plant Physiol.* 148, 25–40. doi: 10.1104/pp.108.121491

Iorizzo, M., Ellison, S., Senalik, D., Zeng, P., Satapoomin, P., Huang, J., et al. (2016). A high-quality carrot genome assembly provides new insights into carotenoid accumulation and asterid genome evolution. *Nat. Genet.* 48, 657–666. doi: 10.1038/ng.3565

Iorizzo, M., Senalik, D. A., Ellison, S. L., Grzebelus, D., Cavagnaro, P. F., Allender, C., et al. (2013). Genetic structure and domestication of carrot (*Daucus carota* subsp. *sativus*) (Apiaceae). *Am. J. Bot.* 100, 930–938. doi: 10.3732/ajb.1300055

Jiang, N., Bao, Z., Zhang, X., Hirochika, H., Eddy, S. R., McCouch, S. R., et al. (2003). An active DNA transposon family in rice. *Nature* 421, 163–167. doi: 10.1038/nature01214

Jiang, N., Feschotte, C., Zhang, X., and Wessler, S. R. (2004). Using rice to understand the origin and amplification of miniature inverted repeat transposable elements (MITEs). *Curr. Opin. Plant Biol.* 7, 115–119. doi: 10.1016/j.pbi.2004.01.004

Kopelman, N. M., Mayzel, J., Jakobsson, M., Rosenberg, N. A., and Mayrose, I. (2015). Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. *Mol. Ecol. Resour.* 15, 1179–1191. doi: 10.1111/1755-0998.12387

Kumar, A., and Hirochika, H. (2001). Applications of retrotransposons as genetic tools in plant biology. *Trends Plant Sci.* 6, 127–134. doi: 10.1016/S1360-1385(00)01860-4

Hua-Van, A., Le Rouzic, A., Maisonhaute, C., and Capy, P. (2005). Abundance, distribution and dynamics of retrotransposable elements and transposons: similarities and differences. *Cytogenet. Genome Res.* 110, 426–440. doi: 10.1159/000084975

Lee, J. K., Kwon, S.-J., Park, K.-C., and Kim, N.-S. (2005). Isaac-CACTA transposons: new genetic markers in maize and sorghum. *Genome* 48, 455–460. doi: 10.1139/g05-013

Lessa, E. (1992). Rapid surveying of DNA sequence variation in natural populations. *Mol. Biol. Evol.* 9, 323–330.

Li, D., Xia, Z., Deng, Z., Liu, X., and Feng, F. (2013). Development, characterization, genetic diversity and cross-species/genera transferability of ILP markers in rubber tree (*Hevea brasiliensis*). *Genes Genomics* 35, 719–731. doi: 10.1007/s13258-013-0122-4

Meyers, B. C., Tingey, S. V., and Morgante, M. (2001). Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. *Genome Res.* 11, 1660–1676. doi: 10.1101/gr.188201

Muthamilarasan, M., Venkata Suresh, B., Pandey, G., Kumari, K., Parida, S. K., and Prasad, M. (2014). Development of 5123 intron-length polymorphic markers for large-scale genotyping applications in foxtail millet. *DNA Res.* 21, 41–52. doi: 10.1093/dnares/dst039

Nagy, S., Poczai, P., Cernák, I., Gorji, A. M., Hegedűs, G., and Taller, J. (2012). PICcalc: an online program to calculate polymorphic information content for molecular genetic studies. *Biochem. Genet.* 50, 670–672. doi: 10.1007/s10528-012-9509-1

Park, K.-C., Lee, J. K., Kim, N.-H., Shin, Y.-B., Lee, J.-H., and Kim, N.-S. (2003). Genetic variation in *Oryza* species detected by MITE-AFLP. *Genes Genet. Syst.* 78, 235–243. doi: 10.1266/ggs.78.235

Peakall, R., and Smouse, P. E. (2006). GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. *Mol. Ecol. Notes* 6, 288–295. doi: 10.1111/j.1471-8286.2005.01155.x

Pritchard, J. K., Wen, X., and Falush, D. (2008). *Structure software: version 2.2.3*. Chicago, IL: Univ. Chicago.

Prohens, J., and Nuez, F. (2008). *Vegetables: Fabaceae, Liliaceae, Solanaceae, and Umbelliferae*. New York, NY: Springer New York. doi: 10.1007/978-0-387-30443-4

Rong, J., Janson, S., Umehara, M., Ono, M., and Vrieling, K. (2010). Historical and contemporary gene dispersal in wild carrot (*Daucus carota* ssp. *carota*) populations. *Ann. Bot.* 106, 285–296. doi: 10.1093/aob/mcq108

Shu, Y., Li, Y., Zhu, Y., Zhu, Z., Lv, D., Bai, X., et al. (2010). Genome-wide identification of intron fragment insertion mutations and their potential use as SCAR molecular markers in the soybean. *Theor. Appl. Genet.* 121, 1–8. doi: 10.1007/s00122-010-1285-x

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., et al. (2012). Primer3–new capabilities and interfaces. *Nucleic Acids Res.* 40:e115. doi: 10.1093/nar/gks596

Wang, X., Zhao, X., Zhu, J., and Wu, W. (2005). Genome-wide investigation of intron length polymorphisms and their potential as molecular markers in rice (*Oryza sativa* L.). *DNA Res.* 12, 417–427. doi: 10.1093/dnares/dsi019

Wang, Y., Chen, J., Francis, D. M., Shen, H., Wu, T., and Yang, W. (2010). Discovery of intron polymorphisms in cultivated tomato using both tomato and *Arabidopsis* genomic information. *Theor. Appl. Genet.* 121, 1199–1207. doi: 10.1007/s00122-010-1381-y

Wydner, K. S., Sechler, J. L., Boyd, C. D., and Passmore, H. C. (1994). Use of an intron length polymorphism to localize the tropoelastin gene to mouse Chromosome 5 in a region of linkage conservation with human chromosome 7. *Genomics* 23, 125–131. doi: 10.1006/geno.1994.1467

Yaakov, B., Ben-David, S., and Kashkush, K. (2013). Genome-wide analysis of *Stowaway*-like MITEs in wheat reveals high sequence conservation, gene association, and genomic diversification. *Plant Physiol.* 161, 486–496. doi: 10.1104/pp.112.204404

Yang, L., Jin, G., Zhao, X., Zheng, Y., Xu, Z., and Wu, W. (2007). PIP: a database of potential intron polymorphism markers. *Bioinformatics* 23, 2174–2177. doi: 10.1093/bioinformatics/btm296

Ye, J., Coulouris, G., Zaretskaya, I., Cutcutache, I., Rozen, S., Madden, T. L., et al. (2012). Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics* 13:134. doi: 10.1186/1471-2105-13-134

Yeh, F., Rongcal, Y., and Boyle, T. (2000). *POPGENE-1.32: A Free Program for the Analysis of Genetic Variation Among and Within Populations Using Co-dominant and Dominant Markers*. Edmonton: Department of Renewable Resources at the University of Alberta.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Macko-Podgórni et al. Mobile DNA (2019) 10:47
https://doi.org/10.1186/s13100-019-0190-3

Mobile DNA

RESEARCH Open Access

Stowaway miniature inverted repeat transposable elements are important agents driving recent genomic diversity in wild and cultivated carrot

Alicja Macko-Podgórni*, Katarzyna Stelmach, Kornelia Kwolek and Dariusz Grzebelus*

Abstract

Background: Miniature inverted repeat transposable elements (MITEs) are small non-autonomous DNA transposons that are ubiquitous in plant genomes, and are mobilised by their autonomous relatives. *Stowaway* MITEs are derived from and mobilised by elements from the *mariner* superfamily. Those elements constitute a significant portion of the carrot genome; however the variation caused by *Daucus carota Stowaway* MITEs (*DcStos*), their association with genes and their putative impact on genome evolution has not been comprehensively analysed.

Results: Fourteen families of *Stowaway* elements *DcStos* occupy about 0.5% of the carrot genome. We systematically analysed 31 genomes of wild and cultivated *Daucus carota*, yielding 18.5 thousand copies of these elements, showing remarkable insertion site polymorphism. *DcSto* element demography differed based on the origin of the host populations, and corresponded with the four major groups of *D. carota,* wild European, wild Asian, eastern cultivated and western cultivated. The *DcStos* elements were associated with genes, and most frequently occurred in 5′ and 3′ untranslated regions (UTRs). Individual families differed in their propensity to reside in particular segments of genes. Most importantly, *DcSto* copies in the 2 kb regions up- and downstream of genes were more frequently associated with open reading frames encoding transcription factors, suggesting their possible functional impact. More than 1.5% of all *DcSto* insertion sites in different host genomes contained different copies in exactly the same position, indicating the existence of insertional hotspots. The *DcSto*7b family was much more polymorphic than the other families in cultivated carrot. A line of evidence pointed at its activity in the course of carrot domestication, and identified *Dcmar*1 as an active carrot *mariner* element and a possible source of the transposition machinery for *DcSto*7b.

Conclusion: *Stowaway* MITEs have made a substantial contribution to the structural and functional variability of the carrot genome.

Keywords: Transposable elements, Insertional polymorphism, TE-gene association, *Mariner*, *DcSto*, *Daucus carota*

Background

Transposable elements (TEs) are discrete segments of DNA capable of changing their genomic location in a process called transposition [1]. Based on the mechanism of transposition, TEs are divided into two classes, class I (retrotransposons), mobilised via an RNA intermediate, use a 'copy and paste' mechanism, while class II (DNA transposons) are mobilised by 'cut and paste' or 'copy and paste' mechanisms of DNA that do not require a reverse transcription step. In both classes, there are autonomous elements that possess enzyme-encoding genes required for mobilisation, non-autonomous elements which can still be mobilised by their autonomous counterparts, and inactive defective copies [2].

Miniature inverted-repeat transposable elements (MITEs) are small in size (< 800 base pairs, bp), usually AT-rich sequences with no coding capacity. They are

* Correspondence: a.macko@urk.edu.pl; d.grzebelus@urk.edu.pl
Institute of Plant Biology and Biotechnology, Faculty of Biotechnology and Horticulture, University of Agriculture in Krakow, 31425 Krakow, Poland

mobilised by related autonomous class II trans-acting elements. Despite their small size, they may account for a significant portion of plant genomes, representing up to 10 and 13.8% for rice and mulberry, respectively [3, 4]. This extremely efficient proliferation of MITEs, as compared to their ancestral autonomous elements, might be caused by their higher affinity for transposase, resulting from lower *cis*-requirements for enzyme recognition and by the presence of subterminal and/or internal enhancers of nucleoprotein complex formation [5, 6]. Some MITEs families, such as *mPing* elements in rice, may preferentially target single-copy gene-rich regions [7], and escape the epigenetic control system because MITE-derived trans-acting siRNAs do not share sequence similarity with the coding region of the source of the transposase [8]. All these features, coupled with their small size, make them abundant in plant genomes, and frequently present in the vicinity of genes.

Currently, the pivotal role of TEs in the evolution of plant genomes is becoming more widely recognised. Mobilisation of TEs leads to structural variations that contribute to the genomic diversity of their host, some of which can be adaptive. Among other effects, the role of TE insertional hotspots in the formation of biosynthetic gene clusters was proposed, based on the analysis of genes of the terpene biosynthesis pathway in eudicots [9]. TE insertions can impact gene expression in many ways. By insertion upstream, within, or downstream of coding regions, TEs may provide new regulatory features that can affect gene expression [10]. In addition, RNA-directed methylation (RdDM), which has a role in repetitive DNA control and defense against viruses, can lead to epigenetic changes upon TE insertion that may produce epialleles for adjacent genes [11]. In crops, such modifications can affect agronomically important traits, such as observed with flowering time variation due to MITE insertions into the quantitative trait locus *Vegetative to generative transition 1 (Vgt1)* [12].

Therefore, one of the main challenges of crop genomics is to critically evaluate the extent of species-wide TE-associated structural variation (TEASV), in order to better understand the dynamics of genome evolution. Several bioinformatics tools have been developed that allow for the identification of TEASV from resequencing data generated by next-generation sequencing (NGS) (reviewed in [13]). However, only a few genome-wide comparative analyses of TEASV have been published, almost exclusively for autogamous species. Moreover, most of these were focused on the global TE landscape, and thus they were biased towards the most numerous TE families.

TE-derived variants are also sources of functional variation, as shown by TE variants linked to changes in DNA methylation in *Arabidopsis thaliana* [14] and flowering traits in maize [15, 16]. A study associating TE-derived variants with phenotypic traits related to maize adaptation to a temperate climate revealed more candidate genes than analysis using single nucleotide polymorphisms [16]. This suggests that TEs may be able to quickly enhance host adaptability under adverse environmental conditions. An adaptive role for TEs has also been suggested for *A. thaliana* [17] and *Capsella rubella* [18].

To date, only a few reports have focused on the global analysis of MITEs at the species level. Mining of MITEs in 19 *A. thaliana* ecotypes yielded a total of 2406 copies grouped into 212 families [19]. In another report, 20 MITE families were annotated in *B. rapa, B. oleracea*, and *A. thaliana*. Of these, only four were present in *A. thaliana*, indicating that amplification and diversification of most *Brassica* MITE families took place after divergence of the *Arabidopsis* and *Brassica* lineages. Moreover, some MITE families were significantly enriched in *B. rapa* and *B. oleracea*; therefore, they were likely activated after divergence of those species [20]. Rice *mJing* elements were frequently inserted into introns (16.67%) and into the 2 kb regions flanking genes (45.83%) [21]. Rice accessions differ dramatically in terms of *mJing* copy number, ranging from 18 to 150 in japonica and African cultivated rice, respectively. This suggested multiple amplification bursts, which most likely occurred before the amplification burst of *mPing*, another well-characterised active MITE in rice [21]. Some rice lines showed a sharp increase of *mPing* copies, from less than 10 copies in indica to 1000 in a temperate japonica cultivar Gimbozu EG4 [22]. Association of MITE insertions with coding regions was also shown for 18 wheat *Stowaway* element families, with 5.1% of more than 19,000 MITEs being transcribed, and 52–63% insertion sites being located within 100 bp of genes [23, 24]. A comparative analysis revealed specific proliferation of two MITE families in the A genome and one in the B genome, suggesting their possible impact on genome diversification during speciation [23].

Carrot (*Daucus carota*) is a diploid species with 2n = 2x = 18, and a relatively small genome of 473 Mb [25]. It is an allogamous species, suffering from inbreeding depression. Cultivated carrot is a biennial root vegetable and the most economically important species of the *Apiaceae* family, and is grown around the world in temperate and subtropical regions [26]. *D. carota* has been domesticated relatively recently, about 1100 years ago. Wild carrot is widespread in temperate regions of the world. While the Mediterranean basin is considered the centre of biodiversity for *Daucus* spp. [27], Central Asia has been identified as the place of origin of domesticated carrots [25, 28]. The species has four major structural groups: European wild *D. carota*, which show

remarkable morphological diversity and are grouped into several subspecies, referred to as *D. carota* complex; Asian wild *D. carota* subsp. *carota*; eastern cultivated carrots, which are mostly primitive landraces, often producing yellow or purple storage roots; and western cultivated carrots, which include advanced orange cultivars. Cultivated and wild carrots can easily hybridise, and a considerable amount of genetic variation is exhibited both between and within the groups, with no apparent signature of a domestication bottleneck [28].

The carrot reference genome assembly of a double haploid plant (DH1) has been published recently [25]. The repetitive fraction constituted 46% of this carrot genome. DNA transposons comprised 13.6% of the genome and 30% of the total repetitive DNA. Approximately 2.3% of the assembled portion of the genome was attributed to MITEs, of which *Stowaway*-like elements constituted around 0.5% [25]. Carrot *Stowaway*-like MITEs (*DcSto*s) had previously been reported to be abundant and highly polymorphic [25, 29]. In this current study, we used 14 *DcSto* families for a systematic genome-wide analysis of TEASVs in 31 resequenced genomes from cultivated and wild carrot accessions. The accessions were representative of the four structural groups of *D. carota*, as described above. *DcSto* insertions were comprehensively annotated and their chromosomal distribution was analysed. In addition, we identified a *DcSto* family likely active in cultivated carrot.

Results

Distribution of *DcSto* elements in *D. carota*

In total, 18,518 *DcSto* insertion sites were identified across 31 genomes of *D. carota* (Table 1 and Additional file 1). Although the coverage of the resequenced genomes ranged from approximately 10× to 40× (50.8–225.3 million reads), this did not affect the sensitivity of insertion detection, as no correlation between the number of reads and the number of identified insertion sites was observed (Spearman rank correlation rho = – 0.12, $p = 0.52$; Additional file 2: Figure S1). In addition, the reference genome was similarly covered by reads from the resequenced accessions, spanning from 93.8 to 96.8% of the assembly [25], with 89.88 to 97.52% of total reads mapped (Additional file 2: Table S1). This indicated that the resequencing data did not show any significant bias, and that they were robust enough to be used for comparative analysis.

We further validated the results of in silico predictions for 39 randomly chosen *DcSto* insertion sites, using intron length polymorphism (DcS-ILP) genotyping, as described by Stelmach et al. [30]. For 16 sites, the results of DcS-ILP genotyping fully supported RelocaTE predictions. At 12 sites, other allelic variants were occasionally present, differing in size from the predicted *DcSto* insertion or the empty site, while for the remaining 11 sites no scorable polymerase chain reaction (PCR) products were produced. For the 28 sites yielding unambiguous PCR products, more than 96% RelocaTE predictions for

Table 1 Abundance and distribution of the 14 *DcSto* families in *D. carota*

DcSto		Number of *DcSto* insertion sites							
	Total	UIS[a]	PrC[b]	2 kb upstream	5′UTR[c]	cds[d]	intron	3′UTR	2 kb downstream
DcSto1	1685	1145	12.70	468	40	4	388	32	265
DcSto2	2594	1739	12.20	763	92	10	450	45	441
DcSto3	821	489	10.39	238	28	4	166	24	145
DcSto4	315	153	6.88	70	9	1	87	4	62
DcSto5	1385	916	11.16	419	53	0	204	27	260
DcSto6	3633	2512	13.23	932	87	8	903	86	702
DcSto7a	1484	983	12.10	456	38	9	296	37	266
DcSto7b	2887	2284	17.50	972	168	17	428	91	527
DcSto7c	256	143	9.85	70	9	0	48	7	55
DcSto8	857	637	12.16	233	35	6	210	13	145
DcSto9	266	140	9.82	73	11	1	57	7	39
DcSto10	301	184	9.87	85	11	0	34	11	49
DcSto11	155	72	6.33	41	4	1	23	3	35
DcSto12	1587	1068	11.18	393	60	6	323	52	336
PIS[e]	292		–	83	10	0	30	3	58
Total/average	18,518	12,464	11.10	5296	655	67	3647	442	3385

[a]*UIS* Unique insertion sites, [b]*PrC* Proliferation coefficient (total number of insertions/average number of insertions per plant), [c]*UTR* Untranslated region, [d]*cds* coding sequence, [e]*PIS* Parallel insertion sites

the accession and site combinations were confirmed by the DcS-ILP assay (Table 2). This demonstrated that the applied in silico strategy reliably identified *DcSto* insertions.

All *DcSto* families had similar densities in all the nine carrot chromosomes (Additional file 2: Figure S2); however, they differed in terms of their copy number, from 155 copies for *DcSto*11 to 3633 copies for *DcSto*6 (Table 1). The differences likely reflected their ability to proliferate once integrated in the genomes of *D. carota*. The proliferation coefficient (PrC), i.e., the proportion of the total number of insertions per family divided by the average number of insertions per genome, ranged from 6.33 for *DcSto*11 to 17.50 for *DcSto*7b (Table 1). PrC values correlated with the intra-family similarity (Additional file 2: Figure S3), indicating that PrC was a good measure of the recent expansion of particular *DcSto* families.

Insertional polymorphism of *DcStos*

Among the 18,518 insertion sites, only two of them harboured the same element insertion in all 31 genomes of *D. carota* (Additional file 2: Table S2), while 22 insertions (0.12%) were present in all genomes of cultivated carrots (Additional file 2: Table S3). We observed a high proportion of insertions in only one of the 31 *D. carota* genomes (66.2%; Additional file 2: Table S4), which we subsequently referred to as unique insertion sites (UIS). It was important to note that most UIS were likely not 'unique' in absolute terms, but only in relation to the collection of 31 plants of different origin investigated in this current study. Thus, the majority of them represented *DcSto* insertions occurring less frequently, but likely still shared among populations of *D. carota*. In general, the number of UIS in the cultivated carrot accessions was relatively more uniform than in those representing the wild carrots. In addition, the wild carrots, especially those of European origin, had a higher proportion of UIS per genome, as compared to the cultivated carrots.

DcStos and the structure of genetic diversity for *D. carota*

The average number of *DcSto* insertion sites per *D. carota* genome was greater than 1400, ranging from 468 to 1978 (Fig. 1a). In the cultivated carrots, the number of *DcSto* copies was slightly higher (1655; ranging from 876 to 1978) than in the wild carrots (1226; ranging from 468 to 1920). As there was no apparent bias resulting from different coverage of the resequenced genomes, the four-fold difference was likely a biological phenomenon. Accessions of cultivated carrots (C1–C13) and inbreds (I1–I3) shared roughly similar numbers of *DcSto* copies. A purple carrot inbred line, B7262 (I4), was the only exception, as it carried far fewer *DcStos* copies. More pronounced differences were observed in the wild *D. carota* gene pool. A group of accessions mostly originating from the West Mediterranean (the centre of biodiversity for *D. carota*) carried less than the average number of *DcSto* copies, while *DcSto* abundance in most wild Asian accessions (W4-W7) was similar to that of the cultivated accessions. This suggested variable dynamics of *DcStos* elements in geographically separated wild populations.

DcSto families differed in terms of their contribution to the total copy number in individual genomes (p-value = 2.2e-16). The differences in *DcSto* distribution reflected the classification of the investigated accessions into four major groups, European wild (Ssp1-Ssp5 and W1-W3), Asian wild (W4-W8), eastern cultivated (C1-

Table 2 Verification of in silico results for 28 *DcSto* insertion sites in the carrot genome by DcS-ILP

Comparison of in silico / DcS-ILP results	Accession / insertion site combinations	
	number	%
empty / homozygous empty	541	64.4%
occupied / homozygous occupied	128	15.2%
occupied / heterozygous (empty + occupied)	94	11.2%
empty / homozygous variant of different size ('empty' for *DcSto*)	30	3.6%
empty / heterozygous (empty + variant of different size)	13	1.6%
occupied / heterozygous (occupied + variant of different size)	1	0.1%
Total correct calls	807	96.1%
empty / heterozygous (empty + occupied)	11	1.3%
empty / homozygous occupied	8	1.0%
occupied / homozygous empty	1	0.1%
Total incorrect calls	20	2.4%
No amplification	13	1.5%
Total	840	100%

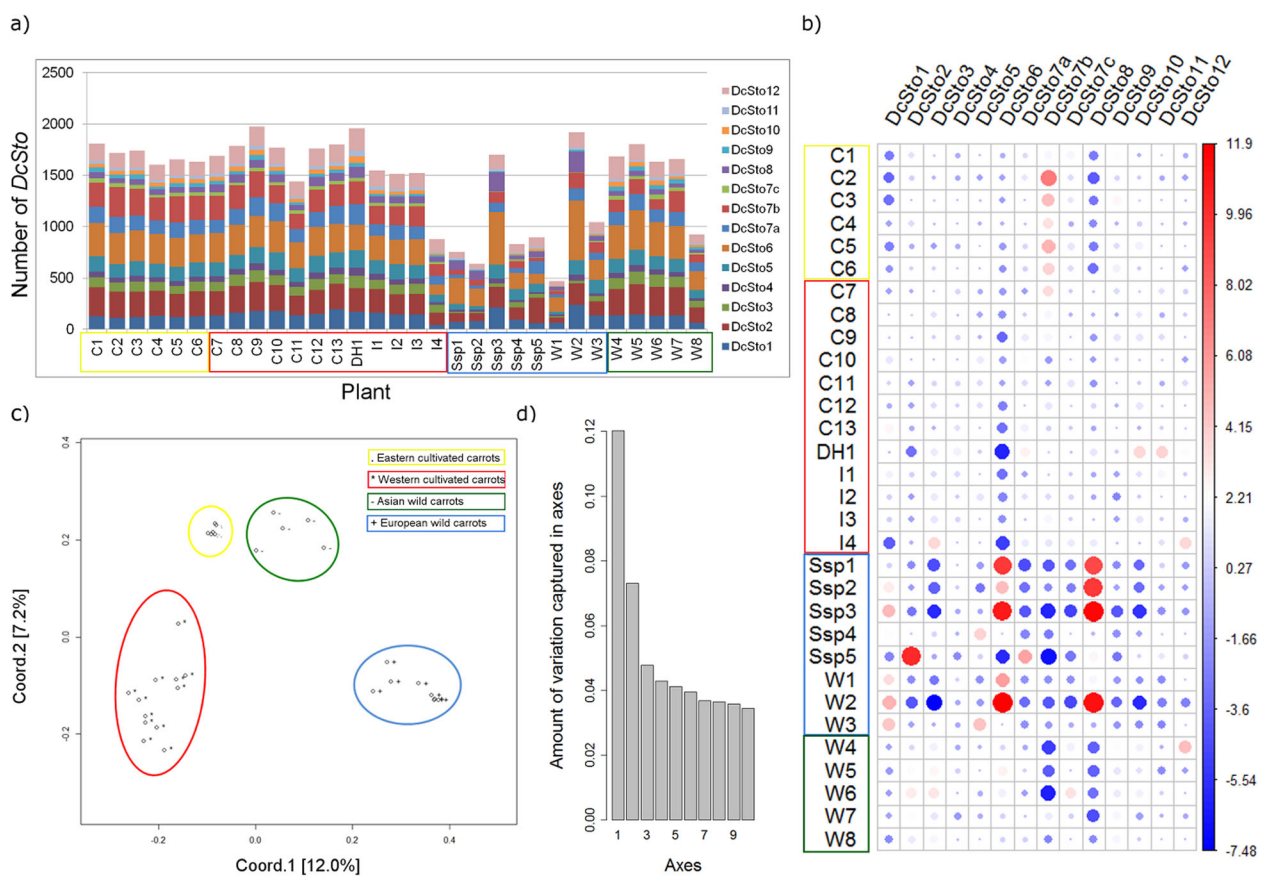


Fig. 1 Distribution of *DcSto* copies among the 31 genomes of *D. carota*. **a** Contribution of families to the total number of *DcSto* insertion sites. **b** Relative abundance of *DcSto* families in 31 *D. carota* genomes ($p = 2.2e\text{-}16$). Colour scale reflects deviations from the average value. Circle size is proportional to the contribution of each test to the total Pearson chi-squared score. **c** Principal coordinate analysis (PCoA) plot based on insertion polymorphisms of the 14 *DcSto* families among the 31 *D. carota* accessions. **d** The variation explained by the first 10 axes of the PCoA. Western and eastern cultivated carrots, and European and Asian wild carrots are marked with red, yellow, blue, and green circles, respectively. Genetic distances were calculated using the Jaccard coefficient

C6), and western cultivated (C7-C13 and I1-I4; Fig. 1b), in agreement with the previously reported population structure of *D. carota* [25, 28]. This was further confirmed by the genetic diversity structure inferred from global *DcSto* insertion polymorphisms. The four major groups were clearly distinguishable as non-overlapping clusters (Fig. 1c and d).

The eastern cultivated carrots were characterised by fewer *DcSto*1 and *DcSto*8 copies, while they carried more *DcSto*7b copies. Different proportions were observed in western cultivated carrots, which had slightly less *DcSto*6 copies. Within the Asian wild carrots, as in the case of the eastern cultivated carrots, *DcSto*1 and *DcSto*8 families were less numerous. By contrast, the eastern cultivated and the Asian wild accessions largely differed in the number of *Dcsto*7b copies, which were overrepresented in the former and underrepresented in the latter. Generally, European wild carrots were the most diverse in terms of *DcSto* distribution. Within this group, accessions Ssp1, Ssp2, Ssp3, and W2 had more *DcSto*6, *DcSto*8, and *DcSto*1 copies, while Ssp5 (*D. carota* subsp. *capillifolius*) was characterised by more *DcSto*2 and *DcSto*7a copies (Fig. 1b).

Localisation of *DcSto* copies in relation to genes was non-random and family-specific

More than 73% of *DcSto* insertions were localised in genic regions, defined as insertions in genes and sequences 2 kb up- or downstream (Table 1). In absolute numbers, *DcSto*s elements were most frequently present in 2 kb upstream regions (28.4%), introns (21.7%), and 2 kb downstream regions (18.3%), while they were virtually absent in exons (Table 1, Fig. 2a and c). The number of insertion sites adjusted for the cumulative length of each defined genic region segment indicated enrichment of *DcSto* insertions in 5′ and 3′ untranslated regions (UTRs), with about 11 insertions per 100 kb of UTR, as compared to 4.5 insertions per 100 kb of introns (Fig. 2d).

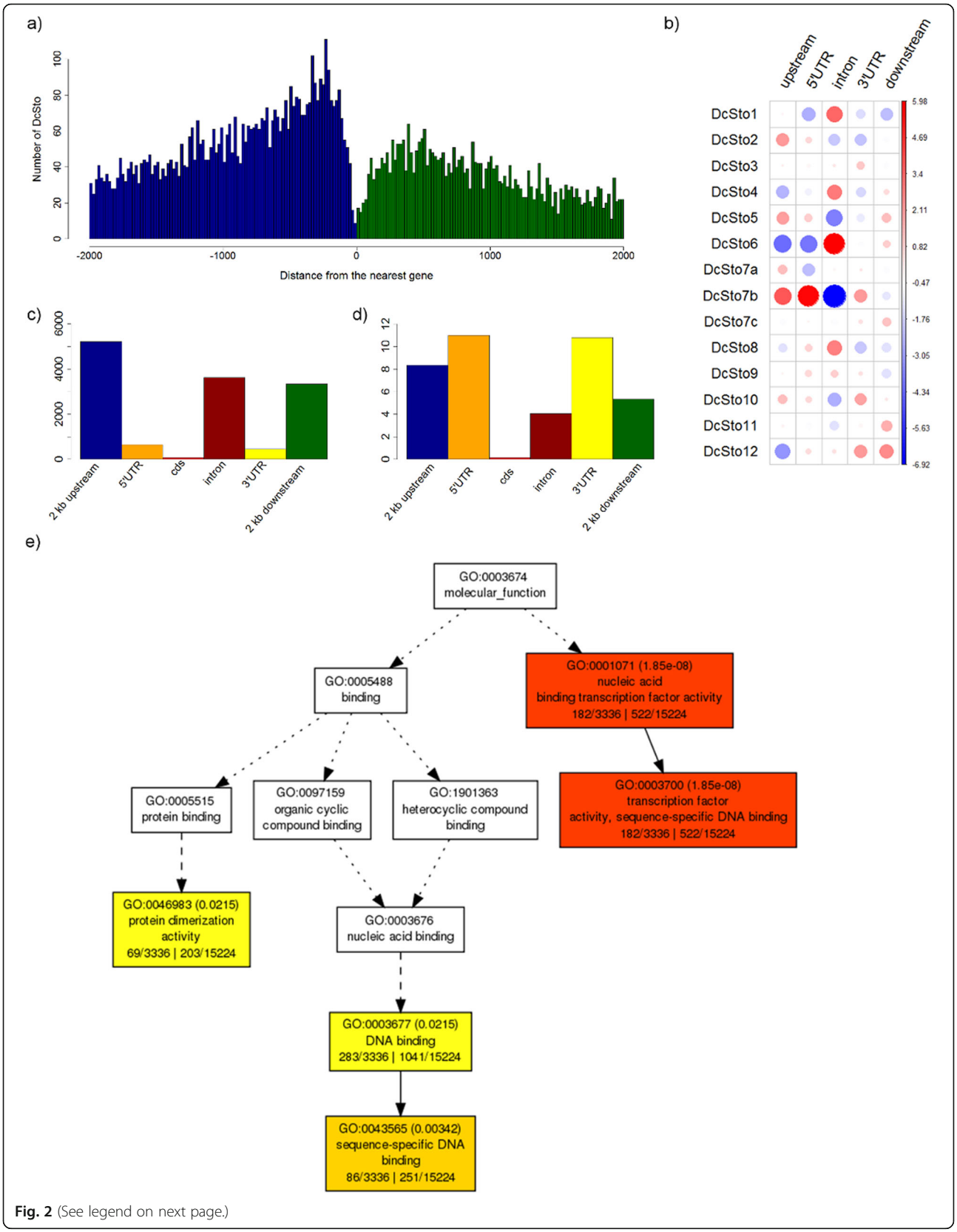


Fig. 2 (See legend on next page.)

(See figure on previous page.)
Fig. 2 The number, distribution, and functional annotation of *DcSto* insertion sites within genic regions. **a** The number of *DcSto* insertions within 2 kb of the nearest gene in windows of 20 bp, with the up- and downstream regions being coloured blue and green, respectively. **b** The number of *DcSto* insertions in up- and downstream regions, exons (cds), introns and UTRs. **c** The number of *DcSto* insertions per 100 kb (standardised to the cumulative length of each region). **d** Differences in the distribution of *DcSto* families within genic regions ($p = 5e\text{-}4$), with cds regions not being included in the analysis. Colour scale reflects deviation from the average value. The size of circles is proportional to the contribution of each test to the total Pearson chi-squared score, and the number inside each cell is the Pearson's residual. **e** Singular enrichment analysis (SEA) of all *DcSto*-associated genes, using AgriGO to define molecular functions

DcSto families differed with respect to their distribution within the genic region (Pearson's chi-squared test, $p = 5e\text{-}04$). *DcSto*7b showed a higher than average proportion of insertions upstream of genes and within 5'UTRs, and a lower than average proportion of insertions within introns. By contrast, the most numerous family, *DcSto6*, showed the opposite pattern, being overrepresented within introns and underrepresented upstream of genes and within 5'UTRs (Fig. 2b).

Gene ontology (GO) enrichment analysis revealed that *DcSto* copies inserted in upstream or downstream regions of genes were significantly associated with those involved in the regulation of transcription (biological process, p-value = 4.06e-13; Additional file 2: Figure S4) and transcription factor activity (molecular function, p-value = 1.85e-08; Fig. 2e). By contrast, *DcSto*s elements inserted in introns did not show an association with any particular GO term, except for marginally significant family-specific signals not related to transcription regulation. For UTRs, the number of *DcSto* insertions were too low to find reliable associations, except for *DcSto*7b insertions in 5'UTR regions, which were significantly associated with genes encoding transcription factors (Additional file 2: Table S5).

DcSto insertion hotspots

Within all identified insertion sites, 292 (1.6%) were parallel insertion sites (PIS), i.e., insertion sites of different *DcSto*s into precisely the same genomic position. Within PIS, 95% harboured insertions of *DcSto* copies from two different families, while the remaining 5% carried alternative insertions of three or more different copies. More

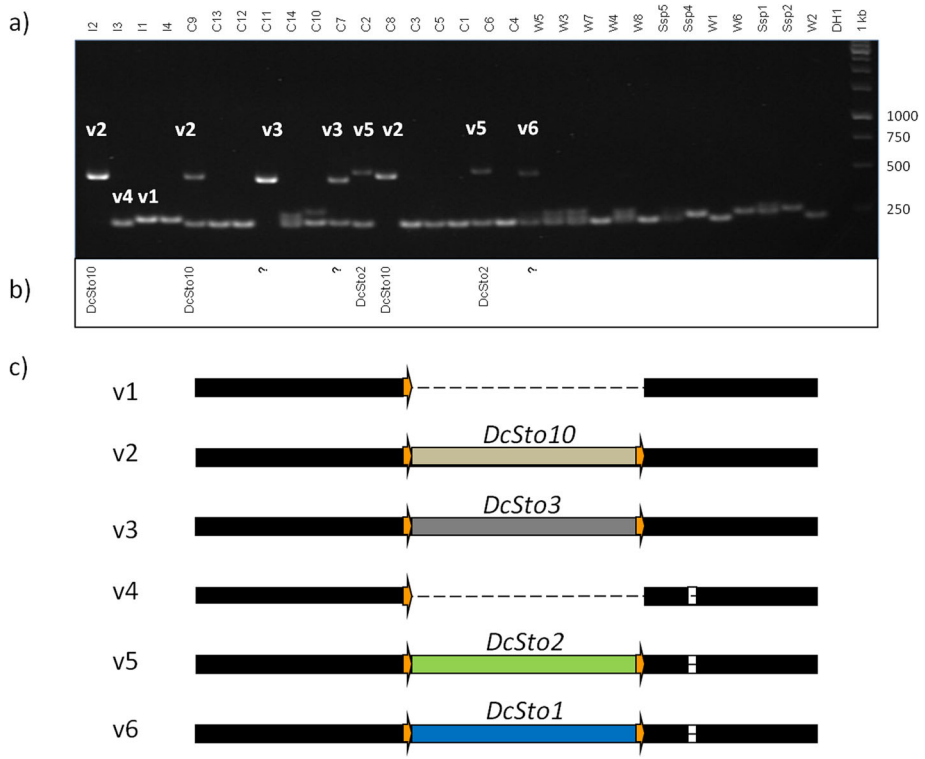


Fig. 3 Verification of parallel insertions (PIS) at the DcS-MIS309 site. **a** PCR amplification profiles for variants (v1 to v6) labeled according to the schematic representation in (**c**). **b** Insertions identified by RelocaTE analysis. **c** Schematic representation of all identified insertion variants. White boxes show insertions and deletions (InDels) in the flanking region. The target site (TS) is represented by an orange arrow

than 63% of PIS were localised in the vicinity of genes or within the body of genes (Additional file 3).

To validate these in silico results, 11 PIS regions were PCR-amplified in the 30 resequenced accessions, and in the DH1 line as a reference, and the resulting amplicons were sequenced by the Sanger method. Presence of the in silico predicted PIS was confirmed in all instances (Fig. 3). Amplicons longer than the expected 'empty' variant were present in some accessions that were qualified as 'empty' by RelocaTE analysis. They carried additional rearrangements, e.g., an insertion of an unrecognised *Stowaway*-like MITE with terminal inverted repeat (TIR) sequences differing from the *DcSto* consensus or another unidentified insertion (Additional file 2: Figure S5), other *DcSto*s, or solo long terminal repeats (LTRs) in nearby positions within the amplicon (Additional file 2: Figure S6). PCR fragments shorter than the expected 'empty' fragment might represent deletion footprints created upon excision of a *DcSto* copy (Additional file 2: Figure S7). In addition, almost all PIS identified in silico as heterozygotes for particular individuals (each variant carrying a different *DcSto* copy) were positively verified by nucleotide sequencing (Additional file 2: Figure S8–S12). In silico identification of MITE insertion sites might be expected to be less reliable for PIS, as observed for the DcS-MIS309 site, where RelocaTE analysis failed to identify insertions in three plants, as revealed by the PCR screen (Fig. 3). Nevertheless, the combined results of in silico prediction and PCR verification suggested that insertions of different *DcSto*s elements and other MITEs into exactly the same genomic sites were quite common.

The co-occurrence of copies from two different *DcSto* families in PIS was positively correlated ($p = 7.03e\text{-}12$) with the cumulative number of all insertion sites of those families, and was negatively correlated ($p = 8.80e\text{-}03$) with the genetic distance between terminal inverted repeats (TIRs) of those families (Additional file 2: Figure S13, Additional file 2: Table S6). Thus, families with more copies were more frequent in PIS; however, *DcSto*s elements carrying more similar TIRs were also relatively more frequently inserted into the same site.

*DcSto7*b elements have been active in the course of carrot domestication

In cultivated carrots (both eastern and western), UIS of elements belonging to the *DcSto7*b family were exceptionally frequent (Fig. 4), accounting for an average of 38% (range 9–59%) of all insertions produced by the family. By contrast, UIS attributed to other *DcSto* families in the cultivated carrots ranged from 0 to 23%, with the average of 8% (Fig. 4b and c). In the reference genome (DH1), the *DcSto7*b family was characterised by the highest within-family similarity (96%) and a unimodal

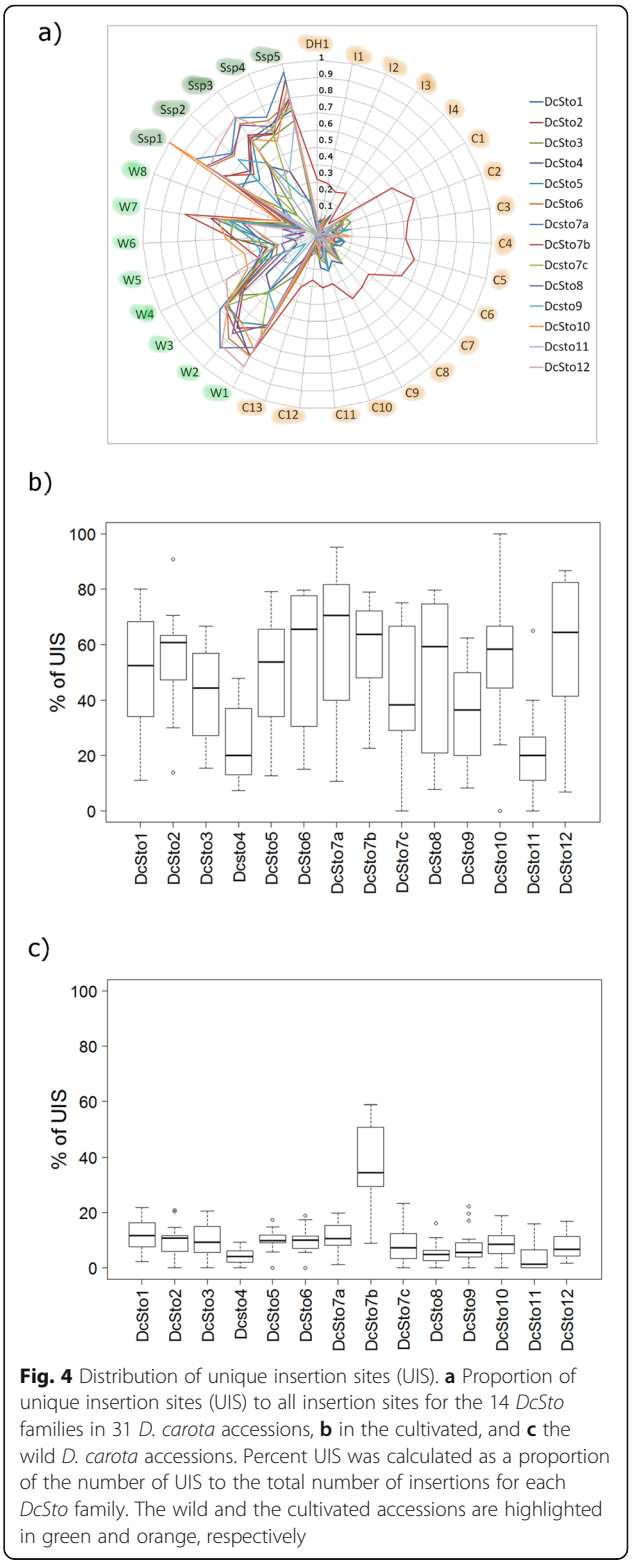


Fig. 4 Distribution of unique insertion sites (UIS). **a** Proportion of unique insertion sites (UIS) to all insertion sites for the 14 *DcSto* families in 31 *D. carota* accessions, **b** in the cultivated, and **c** the wild *D. carota* accessions. Percent UIS was calculated as a proportion of the number of UIS to the total number of insertions for each *DcSto* family. The wild and the cultivated accessions are highlighted in green and orange, respectively

distribution of pairwise distances [25], suggesting a very recent burst in its activity. Combined with the present evidence, including (1) the high proportion of UIS in the genomes of cultivated accessions, (2) the highest PrC

value, and (3) the unique pattern of insertion in relation to genes, as described above, it was likely that *DcSto7*b elements had been mobile in the cultivated carrot gene pool in the course of domestication.

*Dcmar*1 might provide the transposition machinery for *DcSto7*b

The evidence for recent mobilisation of *DcSto*s elements described above prompted us to search for autonomous elements that could have been involved in the process. Eleven copies of *mariner*-like elements were found in the carrot DH1 genome (Additional file 2: Table S7), ranging from 1922 bp (*Dcmar*9) to 4940 bp (*Dcmar*6) and carrying 24- to 32-nt-long TIRs. One mismatch between the 5′ and 3′ TIRs was present in *Dcmar*1, *Dcmar*5, and *Dcmar*10, while TIRs of the remaining elements carried more mismatches (Additional file 2: Table S7). The C-terminal part of the predicted transposases of eight *Dcmar*s had a complete DD39D motif, characteristic of *mariner* elements. Three elements lacking the conserved region of the MLE domain (*Dcmar*9, *Dcmar*10 and *Dcmar*11) were classified as internally truncated and were not further considered.

The first two aspartic acids of the DD39D motif were predicted to be Mg^{2+} binding sites for all eight *Dcmar*s elements, while the helix-turn-helix (HTH) DNA binding motif was predicted for six of them, with at least a 90% probability (Additional file 2: Table S7). However, all features required for *mariner* transposition, as defined by Claeys Bouuaert and Chalmers [31], were only found with *Dcmar*1, a 4353 bp-long element inserted in chromosome 8 (position 25,189,375–25,193,731 in the reference genome DH1).

We investigated the transcriptional status of *Dcmar*s elements, using RNAseq reads of DH1 [25]. The *Dcmar*1 transposase was expressed in four of 20 tissues, callus, whole opened flowers (2 cm umbels at anthesis), bracts (2 cm umbels), and flower buds, while no transcripts attributed to other *Dcmar*s elements were found.

*Dcmar*1 and *DcSto7*b elements were the most similar with respect to their 100 nucleotide (nt) terminal sequences (Fig. 5). Both families shared 31 nt-long TIRs (5′ CTC CCT CCG TCC CTW TTT ATC TGT CCA HTT T 3′). Interestingly, most accessions harbouring a copy of *Dcmar*1 carried more *DcSto7*b copies, as compared to those lacking the autonomous element (Additional file 2: Figure S14a). However, not all accessions carrying *Dcmar*1 elements showed a *DcSto7*b copy number increase, indicating that the presence of *Dcmar*1 elements were essential; however, their activity likely depended on other factors, e.g., chromosomal position of the autonomous element. Therefore, the combined structural, transcriptomic and phylogenetic evidence suggested that *Dcmar*1 elements might have provided

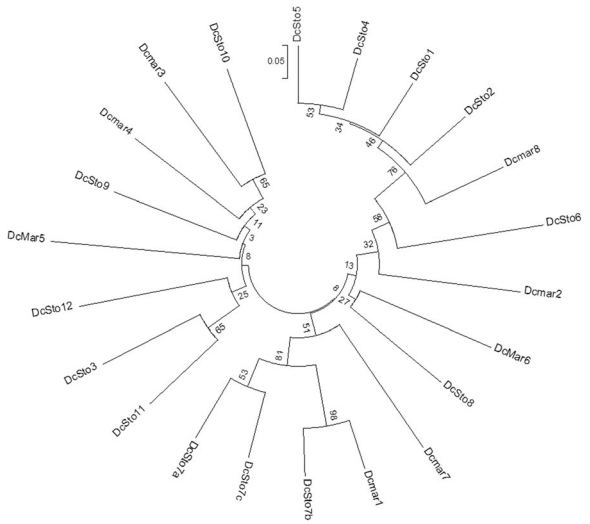


Fig. 5 Neighbour joining tree showing evolutionary relationships of carrot *mariner*-like elements (*Dcmar*s) and *Stowaway* MITEs (*DcSto*s). The tree was generated based on the alignment of 100 nt-long 5′ and 3′ terminal sequences of elements. Numbers show bootstrap values

the transposition machinery for *DcSto7*b elements, driving their recent mobilisation in the gene pool of cultivated carrot.

At least one copy of *Dcmar*1 was present in 14 of 19 genomes of cultivated carrots (74%), but only 4 of 12 wild carrot genomes (33%; Additional file 2: Figure S14b). However, PCR amplification of the region spanning the *Dcmar*1 insertion site in the reference genome DH1 revealed an absence of the element at that genomic location in all other *D. carota* genomes investigated in this study (Additional file 2: Figure S14c). In silico identification of the position of *Dcmar1* elements was largely consistent with results from PCR assays. For all plants but one, we identified at least one putative insertion site, at 17 different genomic locations. Of these, eight *Dcmar*1 insertion sites were associated with genes. In the case of the eastern cultivated accession C4 from Afghanistan, probably only one insertion site was present in the genome; however, due to the insertion of *Dcmar*1 into an intron of one version of paralogues with high similar sequences, we were not able to determine its precise position (Additional file 2: Figure S14d). Only in case of wild carrot accession W2 from Portugal was the PCR assay positive; however, the presence of a *Dcmar*1 copy was not confirmed by in silico analysis. The data suggested that *Dcmar*1 elements had been actively transposing.

Discussion

Transposable elements have been recognised as major drivers of the evolution of eukaryotic genomes. They have been involved in the creation of structural and

functional novelty, making them an essential and long-standing part of genomes [32]. In particular, MITEs and LTR retrotransposons, being ubiquitous in plants, have dynamically shaped genome structure and altered gene function in a variety of ways [33]. A number of recent reports have provided clues of functional interactions of MITEs with host genes [4, 34–41]. The rice genome has been used as a model for investigating MITE-host interactions, as despite its moderate size it harbours an exceptionally rich and diverse collection of almost 180,000 MITE copies divided into 338 families [42], including *mPing*, the most thoroughly studied currently active MITE family [7]. However, less than 15% of all full-length MITE insertions were polymorphic between subspp. *Indica* and *japonica* [42]. Being a dicot species, carrot provides an alternative plant host model to rice. Our results point at a much higher level of insertional polymorphism of carrot *Stowaway* MITEs, perhaps more similar to that of active *mPing* elements in rice [22]. This might be attributed to the recent or ongoing activity of some *DcSto* families, but also to the fact that contrary to rice, carrot is an allogamous species. De la Chaux et al. [43] reported a major reduction of TE abundance in autogamous *A. thaliana*, as compared to its close allogamous relative *A. lyrata*, consistent with such theoretical expectations [44].

DcSto mining strategy

Fourteen *DcSto* families were selected as they were the most abundant in the carrot reference genome [25]. A global annotation of MITEs in carrot also revealed the presence of several other *Stowaway*-like MITE families that were usually less numerous (data not shown). It was possible that some of them acquired much higher copy numbers in the resequenced genomes, as suggested by up to four-fold differences in the global number of analysed *DcSto*s elements, as well as within-family differences.

The genome-wide comparative analysis yielded a catalogue of 18,518 structural variants caused by MITE copies belonging to 14 *DcSto* families across 31 *D. carota* accessions. Of these, less than 2000 copies were attributed to the reference genome DH1. Previously, Iorizzo et al. [25] identified around 4000 *DcSto* copies in the carrot genome. However, in this current study a more stringent approach was used for *DcSto* mining, which resulted in a generally lower number of catalogued insertion sites. Iorizzo et al. [25] used the web tool TIRfinder [45], which identifies all sequences that share common structural features, i.e., specified target site duplications (TSD) and TIRs in assembled sequences. All elements meeting these criteria, regardless of their genomic location, were reported. By contrast, RelocaTE analysis [46] was used in the current study. It retrieved insertion sites from raw sequencing reads based on a similarity search using stringent cut-off parameters, filtering out reads mapping to genomic regions comprising repetitive sequences. Therefore, elements too divergent from the TE family consensus and those inserted into repetitive regions or in the vicinity of structural variants were not reported. This was why in this current study only complete MITEs residing in unique genomic regions were mined, and consequently, the total number of *DcSto* elements obtained for the reference genome was lower than previously reported. Nevertheless, the same approach was systematically applied to all carrot accessions, and the results were comparable. In addition, the in silico mining results were extensively validated by PCR, with the reliability of the mapping tool RelocaTE being largely confirmed.

DcSto distribution and the genetic diversity of *D. carota*

DcSto insertions were extremely polymorphic between cultivated and wild carrots, and within both groups. Global *DcSto* insertion polymorphism revealed a genetic diversity that mirrored previous reports using single nucleotide polymorphisms [25, 47]. It showed that informative *DcSto* insertional polymorphisms, i.e., those present in at least two genomes, allowed grouping of the accessions. The high rate of unique insertion sites observed might have resulted from insufficient sampling, especially for the wild *D. carota* group; however, it might also indicate current transpositional activity. This latter option was supported by the observation that copy numbers of particular *DcSto* families differed among accessions belonging to the same group. For example, the Portuguese accessions (members of the wild European group) were enriched in *DcSto*6 and *Dcsto*8 elements, while subsp. *capillifolius* carried more copies of the *DcSto*2 family. This might suggest amplification bursts of different MITE families in geographically separated populations of wild carrots. In addition, while both eastern cultivated carrots and eastern wild carrots had similarly low frequencies of *DcSto*1 and *DcSto*8, they differed with respect to the numbers of *DcSto*7b elements. The sharp increase in the number of *DcSto*7b copies in eastern cultivated carrots, as compared to the sister clade of Asian wild carrots, suggested the activity of these elements might have been significant during the early stages of domestication.

DcSto insertions in the context of the carrot genome

DcSto copies showed similar distribution patterns across all carrot chromosomes typical for MITEs, i.e. depletion around centromeres and enrichment in genic regions. Previously, fluorescence in situ hybridisation had revealed *DcSto* signals along chromosome arms, and their absence at centromeres, telomeres, and nucleolar

organiser regions [48]. Iorizzo et al. reported that *DcSto*s elements did not show any deviation from a random distribution across the carrot reference genome [25]. Even so, the current study indicated that individual *DcSto* families were characterised by contrasting distribution patterns in terms of their association with genes. Enrichment of *DcSto*7b copies was observed within the 2 kb region upstream of transcription start sites (TSSs) and in 5′ UTRs, and depletion in introns, while *DcSto*6 showed the opposite tendency. This might reflect a genuine preference for insertion of different families into specific sections of genic regions, or random insertions followed by selection acting on non-neutral insertions. The latter scenario would change frequencies of insertions observed in different genic segments, depending on the age of the insertions. *DcSto*7b copies are very similar and the family has been shown to have had a single very recent peak of activity [25]. By contrast, *DcSto*6 elements are an older family, which has likely experienced several peaks of activity [25]. If the selection hypothesis is true, it would imply selection has acted against *DcSto*6 insertions in sequences upstream of genes and in 5′ UTRs and/or retention of insertions in introns.

To date, comparative analyses of TE distribution has usually been generalised for larger groups. A global distribution of MITEs in *Citrus* resembled that reported in this current study for *DcSto*s elements in carrot [49]. However, analyses focusing on particular TE families have revealed specific patterns [4, 50]. Notably, in contrast to carrot *Stowaways*, mulberry *Tc1*/*mariner* superfamily MITEs were the only group that was not preferentially inserted near genes. Nevertheless, they had the highest ratio of the total number of transcribed MITEs to the total number of genes [4]. This supported our hypothesis that detailed analysis of individual families was essential for better understanding of the impact of TEs on the host genome.

Prevalence of low frequency *DcSto* insertions

We observed a high level of low frequency *DcSto* insertions, with most of them referred to as UIS if present in only one of the 31 investigated genomes, a phenomenon also reported for other species [17, 51, 52]. As proposed by Uzunović et al., localisation of TEs in genic region may be limited due to negative selection [53]. On the other hand, the prevalence of UIS may result from an ongoing TE activity. For rice retrotransposon families, Carpentier et al. suggested that the presence of both low- and high-frequency insertion sites indicated continuous transposition, while high numbers of low-frequency insertions indicated their recent mobilisation [52]. In general, most carrot *DcSto* families produced higher proportions of UIS in wild carrots, as compared to the cultivated carrots. With the absence of any significant domestication bottleneck in carrots [25, 28], such a difference was not expected, unless the transpositional activity of *DcSto*s had been elevated in the wild genepool. Alternatively, one might speculate that TE insertional polymorphisms are a more sensitive indicator of a domestication bottleneck than SNPs, due to non-neutrality of some gene-associated insertions.

DcSto insertion hotspots

This current study showed that more than 1.5% of all *DcSto* insertion sites were occupied by more than one *DcSto* element in exactly the same position in different genomes, which we named parallel insertion sites (PIS). The occurrence of different *Stowaway* MITE insertions in orthologous positions has been reported previously, e.g., it was studied for the β-amylase gene in Poaceae [54, 55]. However, it has never been addressed in the context of whole genomes. In this current study, we showed that it was a relatively frequent phenomenon and new insertions appeared fast enough to produce a series of insertion variants within the species. Notably, different copies at the same insertion site usually came from families sharing more similar terminal sequences. This might suggest that they utilised the same source of transposase, which resulted in parallel targeting to the same chromosomal positions. In carrots, 63% of all PIS were located within 2 kb of the nearest gene. As such, it will be important to reveal if these variants show functional variability in these genes. Recently, the importance of variation sources resulting in the occurrence of parallel mutations has been highlighted [56].

*DcSto*s as a source of variation in genic regions

Carrot *DcSto*s elements, like other MITEs, were frequently associated with genes. The current study showed that 73% of all *DcSto* copies were inserted in the vicinity of genes, and particular *DcSto* families differed in their distribution within genic regions. A similar distribution of MITEs, enriched upstream of TSSs and depleted within the body of genes, was observed for *Stowaways* elements in potato [57] and *mPing* elements in rice [58]. Some 9738 carrot genes, including 61 tRNA genes, were associated with at least one *DcSto* element. On average, 3% of all annotated genes were associated with *DcSto*s elements in an individual carrot genome, ranging from 337 genes for accession W1 to 1490 genes for accession C9 (Additional file 2: Table S8). It was likely that these insertions were important for the fine-tuning of the expression of these associated genes [58]. Indeed, the non-random association of *DcSto* insertions with particular groups of genes, most notably transcription factors, indicated functional importance of these associations. However, none of the gene-associated *DcSto* insertions was fixed in *D. carota*. Nevertheless, they might provide a

rich source of variability in the fine-tuning of certain regulatory networks and constitute a basis for selection. A more extensive sampling across carrot germplasm will be required to verify if some of these insertions show signatures of selection during domestication.

A recent genome-wide analysis of TEs showed that they were very important for rapid genome modifications, providing phenotypic variability important for adaptation. In *A. thaliana* ecotypes, genes carrying polymorphic TE insertions were enriched for defense and immune response functions important for adaptation to new ecological niches [17]. At least two genes with TE insertions were likely positively selected, contributing to the adaptation of that species. Similarly, TEs were involved in the rapid adaptation and the "genetic paradox of invasion" of *C. rubella*. In comparison to its outcrossing relative, *C. grandiflora*, *C. rubella* promoter regions were enriched in TE sequences [18]. Variability resulting from polymorphic insertion sites of *Stowaway* MITEs and an altered methylation status of surrounding sequences may impact adaptation to local environmental conditions, as reported for wild emmer wheat [59, 60]. It was likely that *DcSto*s, especially *DcSto*7b, had contributed to the phenotypic variation of cultivated carrots. Carrot has been domesticated relatively recently [61]; however, it shows a remarkable diversity of cultivar types and storage root traits [62]. It was tempting to speculate that at least some of the observed variability among carrot cultivars could have resulted from selection on variants resulting from the insertion of *DcSto* elements.

*DcSto*7b elements were activated upon domestication

To date, only a few active MITEs have been described for which an accompanying autonomous class II element was proposed. These include *Stowaway* family *dTStu1* element in potato [63], and *Tourist* and *hAT*-related MITE families in rice [7, 21, 64–67]. The current study indicated that the *DcSto*7b family had been mobilised in the course of carrot domestication and might still be active in cultivated carrots. The high proportion of UIS of *DcSto*7b elements in cultivated carrots was notable in relation to the opposite trend for the remaining *DcSto* families. Several lines of evidence suggested very recent activity by *DcSto*7b elements, namely the highest PrC value (Table 1), more copies in the cultivated carrot accessions (Fig. 1b), and more UIS as compared to other *DcSto* families (Fig. 4). This was further supported by the highest intra-family similarity of individual copies of *DcSto*7b in the DH1 reference genome, as reported previously [25].

We hypothesised that *Dcmar*1, a related autonomous *Mariner*-like element, provided the transposition machinery for the mobilisation of *DcSto*7b elements. *Dcmar*1 was present only in a subset of the studied carrot accessions, which showed higher *DcSto*7b copy numbers, being more frequently present in genomes of cultivated carrots. The insertion site of *Dcmar*1 in the DH1 reference genome was unique, with the same position being empty in all the remaining 30 plants. Therefore, *Dcmar*1 itself, was likely a currently active element.

Conclusions

This current study described the landscape of carrot *Stowaway* MITEs, providing insight into their importance in shaping the structural and functional variability of the carrot genome. Extreme insertional polymorphism of carrot *Stowaways* was identified, likely resulting from their recent mobilisation, as well as diversification from amplification bursts among carrot accessions. In particular, the *DcSto*7b family had likely been active in the course of domestication. Moreover, *DcSto* insertions were commonly present within genic regions, and were non-randomly associated with specific groups of genes, including those encoding transcription factors, with independent insertions of MITEs in the same genomic positions being relatively common events (comprising 1.6% of all insertion sites). Further analyses of carrot MITEs will be needed to understand the mechanisms responsible for their successful amplification and the extent of their functional impact on genes and on the phenotype of carrots.

Methods

Plant materials

To identify *DcSto* insertions, we used sequencing data from 31 resequenced genomes of *D. carota* (NCBI Sequence Read Archive, accession SRP062070, under umbrella project PRJNA285926; Additional file 2: Table S1), comprising 13 wild and 18 cultivated carrot accessions, along with the assembled carrot reference genome and its raw reads [25]. DNA from the 31 resequenced plants (excluding Ssp3 and including C14) was amplified using a REPLI-g Mini Kit (Qiagen), following the manufacturer's protocol.

In silico mining of *DcSto* insertions

Raw reads were pre-processed by removing low quality reads and trimming adapters using Trimmomatic version 0.35 [68], with parameters minqual = 28, minlen = 50, LEADING:28, TRAILING:28, SLIDINGWINDOW:10:28, and MINLEN:50, and quality was controlled using fastqc [69].

To identify insertion sites of the 14 *DcSto* families we used RelocaTE [46] with consensus sequences representing *DcSto* families [25]. RelocaTE allowed identification of TE insertions from unassembled short reads. In brief, short reads were aligned to a reference/consensus TE sequence, matching reads were trimmed to remove the TE

sequence, and the remaining read fragments were aligned to the reference genome to identify the regions flanking the TE insertions [46]. The following RelocaTE parameters were use: -bm 12, –bt 11, –m 0.2 and -r 1. As the method included a mapping step, we first examined whether there were differences in the percentage of reads aligning to the reference genome. The mapping quality was evaluated with bwa-mem [70], using previously described parameters [71]. Next, files containing information about insertion sites for each *DcSto* family/genome combination were merged and converted into a binary matrix using a custom script, with absence and presence of a TE insertion being scored as 0 and 1, respectively.

Due to differences in genome coverage, we calculated correlation between the depth of coverage and the number of identified insertion sites. The Shapiro-Wilk's normality test was performed, and the non-parametric correlation was tested using Spearman's rank-based correlation, with the results were plotted in R using the 'ggpubr' package v.0.2 [72].

A binary matrix for the 31 accessions was used to calculate the number of *DcSto* insertion sites, UIS, i.e., those present in only a single accession, and the number of PIS, i.e., those with different copies of *DcSto* elements inserted in different genomes at exactly the same position. Genomic distribution of *DcSto* insertion sites and genes was plotted using the 'ggplot2' R package [73].

The presence of *DcSto* insertions in the context of genic regions, divided into five categories of 2 kb upstream sequences, 5'UTRs, coding sequence (cds), introns, 3'UTRs, and 2 kb downstream sequences, were determined based on the *National Center for Biotechnology Information* (NCBI) carrot genome annotation file GCF_001625215.1_ASM162521v1_genomic.gff, using BEDTools v.2.26.0 [74]. The same resource was used to calculate the total length of each of the five genic categories. Singular enrichment analysis (SEA) of the *DcSto*-associated genes was carried out using the Phytozome annotation file (Dcarota_388_v2.0.annotation_info.txt) and AgriGO v.2.0 [75], to define biological processes (BP), cellular components (CC) and molecular functions (MF).

All correlation tests were calculated and plotted using the 'Corrplot' R package [76]. Family distribution of *DcSto* insertion sites within the five genic categories was calculated based on a contingency table of data representing the number of occurrences of each *DcSto* family in defined segments, using the Pearson chi-squared test. Due to a low number of *DcSto*11 insertions, a simulated *p*-value based on 2000 replicates was used. Pearson residuals were calculated using a contingency table containing data representing the total number of insertion sites of each *DcSto* family in individual genomes. The matrix of Pearson's correlation coefficients was calculated to test interconnection between the sum of copy numbers for families that were inserted into the same position (PIS), the number of their common occurrences in PIS, and the genetic distance between each pair of *DcSto* consensus sequences. Intra-family genetic distance was calculated for all copies representing each family identified in the DH1 genome, as reported by Iorrizo et al. [25].

The binary matrix for the 31 genomes was used to calculate the genetic distance based on the Jaccard coefficient, with the 'vegan' R package [77]. This was a conservative approach, where only the presence of a common insertion was considered informative. The values were used for principal coordinate analysis (PCoA) using the 'ape' package in R [78].

Finally, a gff3 file was prepared, where for each insertion the 'start position' referred to the second nucleotide (A) of the target site (TA), while the 'end position' referred to the first nucleotide of the *DcSto* element, in the case of insertions present in the reference genome DH1, or to the first nucleotide following the target site, in the case of insertions not mapped to DH1. For each insertion, a note containing information about its genomic position was given, as well as the LOC number of the adjacent gene, when the *DcSto* copy was inserted in a genic region (less than 2 kb from the gene). The ID field contained information about the *DcSto* family to which the copy was attributed, and comma separated codes of accessions carrying the insertion.

Identification of autonomous elements

Autonomous *mariner*-like elements were mined from the DH1 reference genome assembly using TIRfinder [45], with tirMask: CTCCCTYYSKYMC, tsdMask: TA, tirSeqMismatches: 1, tsdSeqMismatches: 0, tirMaskMismatches: 0 and tsdMaskMismatches: 0. Coordinates and sequences of identified elements were manually inspected to remove redundant sequences. FGENESH [79], GENEID [80] and Augustus [81] gene prediction tools were used to identify coding regions in all mined TE sequences.

Predicted proteins in TEs were aligned with transposase sequences of known plant *mariner*-like elements, from *Ppmar*1 (NCBI accession no. HM581665), *Soymar*1 (NCBI accession no. AF078934) and *OSMAR*1 (Repbase accession no. AC135425), using ClustalW [82]. The presence of a highly conserved fragment of the *mariner* transposase starting from the first two aspartic acids of the DDD motif, previously used for phylogenetic analysis of plant *mariner*-like transposases [83], was manually inspected. Elements lacking the DDD motif were removed from further analysis. For the remaining proteins of putative autonomous elements, HTH motifs [84] and

iron binding sites [85] were identified. The basic local alignment search tool (BLAST) was used to compare the corresponding mRNAs with carrot DH1 RNAseq short reads from 20 tissues (Sequence Read Archive SRP062159) [25].

Phylogenetic analysis of putative autonomous *mariner*-like and *DcSto*s elements was conducted with Mega v.6.06 software [86]. Evolutionary distances were computed based on 50 nt-long sequences of both TIRs using the p-distance method [87], and were used to calculate a neighbour joining tree [88]. Bootstrap values were obtained based on 1000 replicates.

To identify genomic positions of *Dcmar*1 elements in the resequenced genomes, cleaned Illumina reads were analysed by the TRACKPOSON method [52]. One cultivated carrot accession, I4, was not included in the analysis, as only forward reads were available. To avoid false positives from *DcSto* MITEs, TIRs were removed from the *Dcmar*1 query sequence prior to analysis, leaving only the internal portion of the sequence specific to the *Dcmar*1 element. In order to precisely determine genomic positions, sequences flanking *Dcmar*1 elements were reconstructed from unmapped paired reads, manually verified, and aligned with the DH1 carrot reference genome using BLAST analysis. The presence of *Dcmar*1 TIRs in the reconstructed sequences provided a confirmation of the results of in silico mining.

Experimental verification of *DcSto* insertion sites identified by RelocaTE analysis

Thirty-nine *DcSto* insertion sites located in introns, and six sites characterised by parallel insertions, were selected for validation. For PCR, site-specific primers were as described by Stelmach et al. [30], or were designed de novo using Primer3 [89] (Additional file 2: Table S9). Reaction mixes contained about 20 ng REPLI-g-amplified genomic DNA, 1 mM forward and reverse primers, 0.25 mM dNTPs (Thermo Fisher Scientific), 0.5 U Taq DNA polymerase (Thermo Fisher Scientific), and 1x Taq buffer with $MgCl_2$ (Thermo Fisher Scientific). Amplification took place at 94 °C for 1 min, followed by 30 cycles of 94 °C for 30 s, 56 °C/58 °C for 30 s, and 68 °C for 2 min, and finally 68 °C for 6 min. Products were separated by 1% agarose gel electrophoresis, and were purified with a GeneJET Gel extraction kit (Thermo Fisher Scientific), and cloned into pGEM-T (Promega). Cloned DNAs were extracted using the Wizard Plus SV Miniprep DNA Purification System (Promega) and sequenced by the Sanger method (Genomed SA, Poland). Nucleotide sequences were manually aligned using BioEdit [90].

The presence of *Dcmar*1 elements in *D. carota* accessions was verified using a pair of primers, DcMar1_499_F: 5′ GCC GAC ATA CGA ATC CTG TCA 3′ and DcMar1_499_R: 5′ TTG TGG CTT CCT TCT GCT GTA 3′, anchored in the DDD domain of the *Dcmar*1 element. The presence of *Dcmar*1 in the DH1 insertion site was screened across *D. carota* accessions with one of the above DDD-anchored primers in combination with a corresponding forward or reverse primer flanking the insertion (DcMar1_499_flank_F: 5′ TGT TCT TAG CAG CGG TAG CAC and DcMar1_499_flank_R: 5′ GTT GGT GTT TAC ACT GGA GGT TG 3′). As a positive control for the PCRs, a single-copy carrot genomic fragment was amplified with primers CULT-q-orf6-F 5′ CTT CTC GTA CAA CTG AGC C 3′ and CULT-q-orf6-R 5′ GCT TAG CAA GTA CAA GGG AA 3′ [71]. Fragments were amplified in 10 μl reactions containing 20 ng REPLI-g-amplified genomic DNA, 1 mM forward and reverse primer, 1 mM forward and reverse control primer, 0.25 mM dNTPs (Thermo Fisher Scientific), 0.5 U Taq DNA polymerase (Thermo Fisher Scientific), and 1x Taq buffer with $MgCl_2$ (Thermo Fisher Scientific). Amplification took place at 94 °C for 1 min, followed by 30 cycles of 94 °C for 30 s, 56 °C for 30 s, and 68 °C for 2 min for the DDD test and 10 min for the DH1 site, and then 68 °C for the final elongation for 6 min for the DDD test and 20 min for the DH1 site.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s13100-019-0190-3.

Additional file 1: *DcSto* annotation: contains gff3 annotations of 18.5 K *DcSto* insertions identified in 31 carrot genomes.

Additional file 2: Supplementary Figures and Tables: contains supplementary figures and tables referenced in the main manuscripts.

Additional file 3: Supplementary Table: contains a list of gene-associated parallel insertion sites (PIS) and their functional annotations.

Abbreviations

BP: Biological processes; CC: Cellular components; DcS-ILP: *DcSto*-intron length polymorphism; *DcSto*: *Daucus carota Stowaway*; DH1: Carrot double haploid plant used for reference genome assembly; GO: Gene ontology; HTH: Helix-turn-helix; MF: Molecular functions; MITEs: Miniature inverted repeat transposable elements; MLE: *Mariner*-like elements; PIS: Parallel insertion sites; PrC: Proliferation coefficient; RdDM: RNA-directed methylation; SEA: Singular enrichment analysis; TEASV: TE-associated structural variation; TEs: Transposable elements; TIR: Terminal inverted-repeat; TSD: Target site duplication; TSS: Transcription start site; UIS: Unique insertion sites; UTR: Untranslated region

Acknowledgements

Not applicable.

Authors' contributions

AM-P and DG designed the study; AM-P performed in silico analyses; AM-P, KS and KK performed laboratory verification; AM-P and DG edited the manuscript. All authors read, reviewed, and approved the final manuscript.

Funding

The research was financed from (1) funds for basic research on crop improvement granted by the Polish Ministry of Agriculture and Rural Development, (2) MINIATURA1 (grant number 2017/01/X/NZ9/00930) granted by the Polish National Science Center, and (3) funds for the statutory activity of the Faculty of Biotechnology and Horticulture, University

of Agriculture in Krakow, granted by the Polish Ministry of Science and Higher Education.

Availability of data and materials

Reads of 31 resequenced genomes of *D. carota* were downloaded from the NCBI Sequence Read Archive database under the project number: SRP062070. Carrot genome annotation files were downloaded from the NCBI database (GCF_001625215.1_ASM162521v1_genomic.gff) and the Phytozome database (Dcarota_388_v2.0.annotation_info.txt). DH1 transcriptome reads were downloaded from the NCBI Sequence Read Archive database, project number SRP062159. All data generated during this study are included in the published article and its supplementary information files or are available from corresponding authors on reasonable request.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 23 August 2019 Accepted: 21 November 2019
Published online: 27 November 2019

References

1. Piégu B, Bire S, Arensburger P, Bigot Y. A survey of transposable element classification systems – a call for a fundamental update to meet the challenge of their diversity and complexity. Mol Phylogenet Evol. 2015;86:90–109.
2. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 2007;8:973–82.
3. Chen J, Hu Q, Zhang Y, Lu C, Kuang H. P-MITE: a database for plant miniature inverted-repeat transposable elements. Nucleic Acids Res. 2013;42(D1):D1176–81.
4. Xin Y, Ma B, Xiang Z, He N. Amplification of miniature inverted-repeat transposable elements and the associated impact on gene regulation and alternative splicing in mulberry (*Morus notabilis*). Mob DNA. 2019;10(1):27.
5. Loot C, Santiago N, Sanz A, Casacuberta JM. The proteins encoded by the pogo-like *Lemi 1* element bind the TIRs and subterminal repeated motifs of the *Arabidopsis Emigrant* MITE: consequences for the transposition mechanism of MITEs. Nucleic Acids Res. 2006;34(18):5238–46.
6. Yang G, Nagel DH, Feschotte C, Hancock CN, Wessler SR. Tuned for transposition: molecular determinants underlying the hyperactivity of a *Stowaway* MITE. Science. 2009;325:1391–4.
7. Jiang N, Bao Z, Zhang X, Hirochika H, Eddy SR, McCouch SR, et al. An active DNA transposon family in rice. Nature. 2003;421(6919):163.
8. Lu L, Chen J, Robb SM, Okumoto Y, Stajich JE, Wessler SR. Tracking the genome-wide outcomes of a transposable element burst over decades of amplification. Proc Natl Acad Sci U S A. 2017;114(49):E10550–9.
9. Boutanaev AM, Osbourn AE. Multigenome analysis implicates miniature inverted-repeat transposable elements (MITEs) in metabolic diversification in eudicots. Proc Natl Acad Sci U S A. 2018;115(28):E6650–8.
10. Vitte C, Fustier MA, Alix K, Tenaillon MI. The bright side of transposons in crop evolution. Brief Funct Genomics. 2014;13(4):276–95.
11. Martin A, Troadec C, Boualem A, Rajab M, Fernandez R, Morin H, Pitrat M, et al. A transposon-induced epigenetic change leads to sex determination in melon. Nature. 2009;461(7267):1135.
12. Salvi S, Sponza G, Morgante M, Tomes D, Niu X, Fengler KA, et al. Conserved noncoding genomic sequences associated with a flowering-time quantitative trait locus in maize. Proc Natl Acad Sci U S A. 2007;104(27):11376–81.
13. Goerner-Potvin P, Bourque G. Computational tools to unmask transposable elements. Nat Rev Genet. 2018;19:688–704.
14. Stuart T, Eichten SR, Cahn J, Karpievitch YV, Borevitz JO, Lister R. Population scale mapping of transposable element diversity reveals links to gene regulation and epigenomic variation. eLife. 2016;5:e20777.
15. Zerjal T, Rousselet A, Mhiri C, Combes V, Madur D, Grandbastien MA, et al. Maize genetic diversity and association mapping using transposable element insertion polymorphisms. Theor Appl Genet. 2012;124(8):1521–37.
16. Lai X, Schnable JC, Liao Z, Xu J, Zhang G, Li C, et al. Genome-wide characterization of non-reference transposable element insertion polymorphisms reveals genetic diversity in tropical and temperate maize. BMC Genomics. 2017;18(1):702.
17. Li ZW, Hou XH, Chen JF, Xu YC, Wu Q, González J, et al. Transposable elements contribute to the adaptation of *Arabidopsis thaliana*. Genome Biol Evol. 2018;10(8):2140–50.
18. Niu XM, Xu YC, Li ZW, Bian YT, Hou XH, Chen JF, et al. Transposable elements drive rapid phenotypic variation in *Capsella rubella*. Proc Natl Acad Sci U S A. 2019;116(14):6908–13.
19. Guo C, Spinelli M, Ye C, Li QQ, Liang C. Genome-wide comparative analysis of miniature inverted repeat transposable elements in 19 *Arabidopsis thaliana* ecotype accessions. Sci Rep. 2017;7(1):2634.
20. Sampath P, Murukarthick J, Izzah NK, Lee J, Choi HI, Shirasawa K, et al. Genome-wide comparative analysis of 20 miniature inverted-repeat transposable element families in *Brassica rapa* and *B. oleracea*. PLoS One. 2014;9(4):e94499.
21. Tang Y, Ma X, Zhao S, Xue W, Zheng X, Sun H, et al. Identification of an active miniature inverted-repeat transposable element *mJing* in rice. Plant J. 2019;98(4):639–53.
22. Naito K, Cho E, Yang G, Campbell MA, Yano K, Okumoto Y, et al. Dramatic amplification of a rice transposable element during recent domestication. Proc Natl Acad Sci U S A. 2006;103(47):17620–5.
23. Yaakov B, Ben-David S, Kashkush K. Genome-wide analysis of stowaway-like MITEs in wheat reveals high sequence conservation, gene association, and genomic diversification. Plant Physiol. 2013;161(1):486–96.
24. Keidar-Friedman D, Bariah I, Kashkush K. Genome-wide analyses of miniature inverted-repeat transposable elements reveals new insights into the evolution of the *Triticum-Aegilops* group. PLoS One. 2018;13(10):e0204972.
25. Iorizzo M, Ellison S, Senalik D, Zeng P, Satapoomin P, Huang J, Bowman M, et al. A high-quality carrot genome assembly provides new insights into carotenoid accumulation and asterid genome evolution. Nat Genet. 2016;48(6):657.
26. Simon P.W. Economic and Academic Importance. In: Simon P, Iorizzo M, Grzebelus D, Baranski R, editors. The Carrot Genome. C. Cole, series editor. Compendium of Plant Genomes. Springer Nature Switzerland AG; 2019:1–8.
27. Spooner D.M. *Daucus*: Taxonomy, Phylogeny, Distribution. In: Simon P, Iorizzo M, Grzebelus D, Baranski R, editors. The Carrot Genome. C. Cole, series editor. Compendium of Plant Genomes. Springer Nature Switzerland AG; 2019:9–26.
28. Iorizzo M, Senalik DA, Ellison SL, Grzebelus D, Cavagnaro PF, Allender C, Brunet J, et al. Genetic structure and domestication of carrot (*Daucus carota* subsp. *sativus*)(*Apiaceae*). Am J Bot. 2013;100(5):930–8.
29. Macko-Podgórni A, Nowicka A, Grzebelus E, Simon PW, Grzebelus D. *DcSto*: carrot *Stowaway*-like elements are abundant, diverse, and polymorphic. Genetica. 2013;141(4–6):255–67.
30. Stelmach K, Kruk M, Macko-Podgórni A, Grzebelus D. Miniature Inverted Repeat Transposable Element Insertions Provide a Source of Intron Length Polymorphism Markers in the Carrot (*Daucus carota* L.). Front Plant Sci. 2017;8:725.
31. Claeys Bouuaert C, Chalmers R. A single active site in the mariner transposase cleaves DNA strands of opposite polarity. Nucleic Acids Res. 2017;45(20):11467–78.
32. Hua-Van A, Le Rouzic A, Boutin TS, Filée J, Capy P. The struggle for life of the genome's selfish architects. Biol Direct. 2011;6:19.
33. Wessler SR, Bureau TE, White SE. LTR-retrotransposons and MITEs: important players in the evolution of plant genomes. Curr Opin Genet Dev. 1995;5:814–21.
34. Oki N, Yano K, Okumoto Y, Tsukiyama T, Teraishi M, Tanisaka T. A genome-wide view of miniature inverted-repeat transposable elements (MITEs) in rice, *Oryza sativa* ssp. *japonica*. Genes Genet Syst. 2008;83(4):321–9.
35. Kuang H, Padmanabhan C, Li F, Kamei A, Bhaskar PB, Ouyang S, et al. Identification of miniature inverted-repeat transposable elements (MITEs) and biogenesis of their siRNAs in the *Solanaceae*: new functional implications for MITEs. Genome Res. 2009;19(1):42–56.
36. Zhou L, Zhang J, Yan J, Song R. Two transposable element insertions are causative mutations for the major domestication gene teosinte branched 1 in modern maize. Cell Res. 2011;21(8):1267.

37. Yang Q, Li Z, Li W, Ku L, Wang C, Ye J, et al. CACTA-like transposable element in ZmCCT attenuated photoperiod sensitivity and accelerated the postdomestication spread of maize. Proc Natl Acad Sci U S A. 2013;110(42): 16969–74.
38. Wei L, Gu L, Song X, Cui X, Lu Z, Zhou M, et al. Dicer-like 3 produces transposable element-associated 24-nt siRNAs that control agricultural traits in rice. Proc Natl Acad Sci U S A. 2014;111(10):3877–82.
39. Henaff E, Vives C, Desvoyes B, Chaurasia A, Payet J, Gutierrez C, et al. Extensive amplification of the E2F transcription factor binding sites by transposons during evolution of *Brassica* species. Plant J. 2014;77(6):852–62.
40. Mao H, Wang H, Liu S, Li Z, Yang X, Yan J, et al. A transposable element in a NAC gene is associated with drought tolerance in maize seedlings. Nat Commun. 2015;6:8326.
41. Morata J, Marín F, Payet J, Casacuberta JM. Plant lineage-specific amplification of transcription factor binding motifs by miniature inverted-repeat transposable elements (MITEs). Genome Biol Evol. 2018;10(5):1210–20.
42. Chen J, Lu C, Zhang Y, Kuang H. Miniature inverted-repeat transposable elements (MITEs) in rice were originated and amplified predominantly after the divergence of *Oryza* and *Brachypodium* and contributed considerable diversity to the species. Mob Genet Elem. 2012;2(3):127–32.
43. De la Chaux N, Tsuchimatsu T, Shimizu KK, Wagner A. The predominantly selfing plant *Arabidopsis thaliana* experienced a recent reduction in transposable element abundance compared to its outcrossing relative *Arabidopsis lyrata*. Mob DNA. 2012;3:2.
44. Boutin TS, Le Rouzic A, Capy P. How does selfing affect the dynamics of selfish transposable elements? Mob DNA. 2012;3:5.
45. Gambin T, Startek M, Walczak K, Paszek J, Grzebelus D, Gambin A. TIRfinder: a web tool for mining class II transposons carrying terminal inverted repeats. Evol Bioinforma. 2013;9:17.
46. Robb SM, Lu L, Valencia E, Burnette JM, Okumoto Y, Wessler SR, et al. The use of RelocaTE and unassembled short reads to produce high-resolution snapshots of transposable element generated diversity in rice. G3-Genes Genome Genet. 2013;3(6):949–57.
47. Ellison SL, Luby CH, Corak KE, Coe KM, Senalik D, Iorizzo M, et al. Carotenoid presence is associated with the *Or* gene in domesticated carrot. Genetics. 2018;210(4):1497–508.
48. Nowicka A, Grzebelus E, Grzebelus D. Precise karyotyping of carrot mitotic chromosomes using multicolour-FISH with repetitive DNA. Biol Plant. 2016; 60(1):25–36.
49. Liu Y, Tahir ul Qamar M, Feng JW, Ding Y, Wang S, Wu G, et al. Comparative analysis of miniature inverted–repeat transposable elements (MITEs) and long terminal repeat (LTR) retrotransposons in six *Citrus* species. BMC Plant Biol. 2019;19:140.
50. Quadrana L, Silveira AB, Mayhew GF, LeBlanc C, Martienssen RA, Jeddeloh JA, et al. The *Arabidopsis thaliana* mobilome and its impact at the species level. eLife. 2016;5:e15716.
51. Wei B, Liu H, Liu X, Xiao Q, Wang Y, Zhang J, et al. Genome-wide characterization of non-reference transposons in crops suggests non-random insertion. BMC Genomics. 2016;17(1):536.
52. Carpentier MC, Manfroi E, Wei FJ, Wu HP, Lasserre E, Llauro C, et al. Retrotranspositional landscape of Asian rice revealed by 3000 genomes. Nat Commun. 2019;10(1):24.
53. Uzunović J, Josephs EB, Stinchcombe JR, Wright SI. Transposable elements are important contributors to standing variation in gene expression in *Capsella grandiflora*. Mol Biol Evol. 2019;36(8):1734–45.
54. Mason-Gamer RJ. Multiple homoplasious insertions and deletions of a *Triticeae* (*Poaceae*) DNA transposon: a phylogenetic perspective. BMC Evol Biol. 2007;7(1):92.
55. Minaya M, Pimentel M, Mason-Gamer R, Catalan P. Distribution and evolutionary dynamics of *Stowaway* miniature inverted repeat transposable elements (MITEs) in grasses. Mol Phylogenet Evol. 2013;68(1):106–18.
56. Press MO, Hall AN, Morton EA, Queitsch C. Substitutions are boring: some arguments about parallel mutations and high mutation rates. Trends Genet. 2019. https://doi.org/10.1016/j.tig.2019.01.002.
57. Marand AP, Jansky SH, Zhao H, Leisner CP, Zhu X, Zeng Z, et al. Meiotic crossovers are associated with open chromatin and enriched with *Stowaway* transposons in potato. Genome Biol. 2017;18(1):203.
58. Naito K, Zhang F, Tsukiyama T, Saito H, Hancock CN, Richardson AO, et al. Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. Nature. 2009;461(7267):1130.
59. Venetsky A, Levy-Zamir A, Khasdan V, Domb K, Kashkush K. Structure and extent of DNA methylation-based epigenetic variation in wild emmer wheat (*T. turgidum* ssp. *dicoccoides*) populations. BMC Plant Biol. 2015;15(1):200.
60. Domb K, Keidar D, Yaakov B, Khasdan V, Kashkush K. Transposable elements generate population-specific insertional patterns and allelic variation in genes of wild emmer wheat (*Triticum turgidum ssp. dicoccoides*). BMC Plant Biol. 2017;17(1):175.
61. Ellison S. Carrot domestication. In: Simon P, Iorizzo M, Grzebelus D, Baranski R, editors. The Carrot Genome. C. Cole, series editor. Compendium of Plant Genomes. Springer Nature Switzerland AG; 2019:77–92.
62. Simon PW, Freeman RE, Vieira JV, Boiteux LS, Briard M, Nothnagel T, et al. Carrot. In: Prohens J, Carena MJ, Nuez F, editors. Handbook of crop breeding, vol 1. Vegetable breeding. Heidelberg: Springer; 2008. p. 327–57.
63. Momose M, Abe Y, Ozeki Y. Miniature inverted-repeat transposable elements of *Stowaway* are active in potato. Genetics. 2010;186(1):59–66.
64. Fujino K, Sekiguchi H, Kiguchi T. Identification of an active transposon in intact rice plants. Mol Gen Genomics. 2005;273(2):150–7.
65. Moon S, Jung KH, Lee DE, Jiang WZ, Koh HJ, Heu MH, et al. Identification of active transposon *dTok*, a member of the *hAT* family, in rice. Plant Cell Physiol. 2006;47(11):1473–83.
66. Huang J, Zhang K, Shen Y, Huang Z, Li M, Tang D, et al. Identification of a high frequency transposon induced by tissue culture, *nDaiZ*, a member of the *hAT* family in rice. Genomics. 2009;93(3):274–81.
67. Dong HT, Zhang L, Zheng KL, Yao HG, Chen J, Yu FC, et al. A gaijin-like miniature inverted repeat transposable element is mobilized in rice during cell differentiation. BMC Genomics. 2012;13(1):135.
68. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–20.
69. Andrews S. FastQC: a quality control tool for high throughput sequence data. 2010. http://www.bioinformatics.babraham.ac.uk/projects/fastqc of subordinate document. Accessed 20 Aug 2019.
70. Li H, Durbin R. Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics. 2009;25:1754–60.
71. Macko-Podgórni A, Machaj G, Stelmach K, Senalik D, Grzebelus E, Iorizzo M. Characterization of a genomic region under selection in cultivated carrot (*Daucus carota subsp. sativus*) reveals a candidate domestication gene. Front Plant Sci. 2017;8:12.
72. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2014. http://www.R-project.org/. Accessed 20 Aug 2019
73. Wickham H. ggplot2: Elegant Graphics for Data Analysis. New York: Springer-Verlag; 2009.
74. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841–2.
75. Tian T, Liu Y, Yan H, You Q, Yi X, Du Z, et al. agriGO v2. 0: a GO analysis toolkit for the agricultural community, 2017 update. Nucleic Acids Res. 2017; 45(W1):W122–9.
76. Wei T, Simko V. R package "corrplot":Visualization of a Correlation Matrix (Version 0.85). 2018. https://github.com/taiyun/corrplot. Accessed 20 Aug 2019.
77. Oksanen J, Blanchet G, Kindt R, Legendre P, Minchin PR, O'Hara RB, et al. vegan: Community Ecology Package. R package version 2.3–5. 2016. http://CRAN.R-project.org/package=vegan. Accessed 20 Aug 2019.
78. Paradis E, Claude J, Strimmer K. APE: analyses of phylogenetics and evolution in R language. Bioinformatics. 2004;20:289–90.
79. Salamov A, Solovyev V. *Ab initio* gene finding in *Drosophila* genomic DNA. Genome Res. 2000;10:516–22.
80. Blanco E, Parra G, Guigó R. Using geneid to identify genes. Curr Protoc Bioinformatics. 2007;18(1):4–3.
81. Stanke M, Diekhans M, Baertsch R, Haussler D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics. 2008;24(5):637–44.
82. Thompson JD, Higgins DG, Gibson TJ. Improving the sensitivity of progressive multiple sequence management, analysis, and homology determination. Nucleic Acids Res. 1994;22:4673–80.
83. Feschotte C, Wessler SR. Mariner-like transposases are widespread and diverse in flowering plants. Proc Natl Acad Sci U S A. 2002;99(1):280–5.
84. Dodd IB, Egan JB. Improved detection of helix-turn-helix DNA-binding motifs in protein sequences. Nucleic Acids Res. 1990;18(17):5019–26.
85. Hu X, Dong Q, Yang J, Zhang Y. Recognizing metal and acid radical ion-binding sites by integrating ab initio modeling with template-based transferals. Bioinformatics. 2016;32(21):3260–9.

86. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: molecular evolutionary genetics analysis version 6.0. Mol Biol Evol. 2013;30(12):2725–9.
87. Nei M, Kumar S. Molecular evolution and phylogenetics. Oxford: Oxford University Press; 2000.
88. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 1987;4(4):406–25.
89. Rozen S, Skaletsky H. Primer3 on the WWW for general users and for biologist programmers. In: Bioinformatics methods and protocols. Totowa: Humana Press; 2000. p. 365–86.
90. Hall TA. BioEdit: a user-friendly biological sequence alignment editor and analysis program for windows 95/98/NT. Nucleic Acids Symp Ser. 1999; 41(41):95–8.

Publisher's Note

Stelmach et al. BMC Plant Biology (2021) 21:200
https://doi.org/10.1186/s12870-021-02980-0

BMC Plant Biology

RESEARCH Open Access

Genetic diversity structure of western-type carrots

Katarzyna Stelmach[1], Alicja Macko-Podgórni[1], Charlotte Allender[2] and Dariusz Grzebelus[1*]

Abstract

Background: Carrot is a crop with a wide range of phenotypic and molecular diversity. Within cultivated carrots, the western gene pool comprises types characterized by different storage root morphology. First western carrot cultivars originated from broad-based populations. It was followed by intercrosses among plants representing early open-pollinated cultivars, combined with mass phenotypic selection for traits of interest. Selective breeding improved root uniformity and led to the development of a range of cultivars differing in root shape and size. Based on the root shape and the market use of cultivars, a dozen of market types have been distinguished. Despite their apparent phenotypic variability, several studies have suggested that western cultivated carrot germplasm was genetically non-structured.

Results: Ninety-three *DcS*-ILP markers and 2354 SNP markers were used to evaluate the structure of genetic diversity in the collection of 78 western type open-pollinated carrot cultivars, each represented by five plants. The mean percentage of polymorphic loci segregating within a cultivar varied from 31.18 to 89.25% for *DcS*-ILP markers and from 45.11 to 91.29% for SNP markers, revealing high levels of intra-cultivar heterogeneity, in contrast to its apparent phenotypic stability. Average inbreeding coefficient for all cultivars was negative for both *DcS*-ILP and SNP, whereas the overall genetic differentiation across all market classes, as measured by F_{ST}, was comparable for both marker systems. For *DcS*-ILPs 90–92% of total genetic variation could be attributed to the differences within the inferred clusters, whereas for SNPs the values ranged between 91 to 93%. Discriminant Analysis of Principal Components enabled the separation of eight groups cultivars depending mostly on their market type affiliation. Three groups of cultivars, i.e. Amsterdam, Chantenay and Imperator, were characterized by high homogeneity regardless of the marker system used for genotyping.

Conclusions: Both marker systems used in the study enabled detection of substantial variation among carrot plants of different market types, therefore can be used in germplasm characterization and analysis of genome relationships. The presented results likely reveal the actual genetic diversity structure within the western carrot gene pool and point at possible discrepancies within the cultivars' passport data.

Keywords: *Daucus carota*, Genetic diversity, Population structure, Market classes, Root shape, SNP, *DcSto*, DAPC

* Correspondence: d.grzebelus@urk.edu.pl
[1]Department of Plant Biology and Biotechnology, University of Agriculture in Krakow, Al. 29 Listopada 54, 31-425 Kraków, Poland
Full list of author information is available at the end of the article

Background

Cultivated carrot is a biennial root vegetable grown around the world in temperate and subtropical regions. It is an outcrossing diploid with a relatively small genome of ca. 473 Mb organized into 2n = 18 chromosomes [1]. Carrot is important nutritionally, placed among the most significant sources of β-carotene in the human diet. It is among the top ten vegetables in terms of global production [2]. Carrot is indigenous to Europe, Asia and North America, with Central Asia identified as the place of origin of cultivated carrots [3]. Carrot was most probably domesticated as a root crop around 1100 years ago in Central Asia. Early domesticated carrots were purple and yellow [4]. The majority of cultivated types that formed the basis of modern commercial cultivars were developed in Asia Minor (Turkey) and temperate regions of Europe. Thus, the above-mentioned regions are considered as the secondary centre of origin for carrot [5]. To date, several molecular approaches, such as isoenzymes, ALFPs, RFLPs and SNPs have been used to examine genetic relationships within *D. carota* [3, 6–8]. Population structure comprising four major groups has been commonly observed within *D. carota* species [3, 9–11]. European wild *D. carota* group is characterized by a high level of diversity and includes several *D. carota* subspecies; Asian wild group is less complex and comprises mostly *D. carota* subsp. *carota*. Western cultivated carrots form a numerous group characterized by high level of diversity in terms of storage root characteristics but are generally orange. Eastern cultivated carrots have more uniform root characteristics but display more variability in terms of colour, as they usually have yellow or purple roots. Further geographic structure was identified by Arbizu et al. [12], leading to the separation of an additional two groups: wild carrots of the Iberian Peninsula and Morocco (1) and landraces of the Balkan Peninsula, Middle East and North Africa (excluding Morocco) (2). Investigation of nearly 120 accessions representing Chinese cultivars and western cultivated carrot carried out by Ma et al. [13] showed clear separation of both gene pools and suggested independent processes of carotenoid-based root pigmentation in the history of eastern cultivars development.

Carrot is a crop with a wide range of phenotypic and genotyping variation that might be of use to breeders. Since the seventeenth century, a lot of breeding efforts have been focused on root traits such as shape, smoothness of the root surface or root integrity [14, 15]. First carrot cultivars originated from broad-based populations. It was followed by intercrosses among plants representing early open-pollinated (OP) varieties, combined with mass phenotypic selection for traits of interest. Discovery of cytoplasmic male sterility in the late 1940s led to a shift from OP to hybrid cultivars characterized by higher level of uniformity. Nonetheless, OP cultivars are still a valuable source of genetic diversity and represent a large portion of plant materials freely available to breeders worldwide through gene banks and public breeding programs [16].

Selective breeding improved root uniformity and led to the development of a range of cultivars differing in root shape and size. Based on the root shape and the market use of carrot cultivars, a dozen of market types (or varietal groups) have been distinguished [17] (Fig. 1). Some older market types were typically bred and developed in Europe (e.g. Long Orange, Amsterdam or Paris Market), while others were characteristic for the U.S. market (e.g. Imperator or Danvers). The work of Banga [14] has been the most comprehensive description of the main western cultivated carrot types available to date. He described key characteristics and use of modern types of western carotene carrot and discussed their connection with well-established original varieties. Despite apparent phenotypic variability observed among market classes, several studies have suggested that western cultivated carrot germplasm was genetically non-structured [3, 9, 10]. Later studies carried out by Stelmach et al. [18] provided the first molecular evidence for a possible root-type associated structure of genetic diversity in western cultivated carrot. They showed that *Daucus carota Stowaway (DcSto)* Miniature Inverted Repeat Transposable Element (MITE) based molecular markers (*DcS*-ILP) detected substantial variation among carrot plants of different origin and could be exploited in germplasm characterization and analysis of genome relationships. MITEs are non-autonomous DNA transposons requiring the presence of a related autonomous element to be a donor of a transposase inducing their transposition *in trans*. Global analysis of *DcSto* MITEs provided evidence for their recent mobility and identified a candidate autonomous element, *DcMar1*, as a possible source of transposase [19].

In the present study, we investigated a collection of plants from a range of OP western-type carrot cultivars producing roots of different shapes and representing several varietal groups. We aimed to detect possible genetic structure underlying apparent phenotypic differences among well-established market types. We exploited and compared two codominant molecular marker systems, *DcS*-ILPs and SNPs, as we assumed that the former system might be capable of revealing variability which arose more recently, resulting from the transpositional activity of *DcSto* MITEs.

Results

Genetic diversity revealed by DcS-ILP and SNP genotyping

A total of 93 *DcS*-ILP markers and 2354 SNP markers, distributed along the nine carrot chromosomes

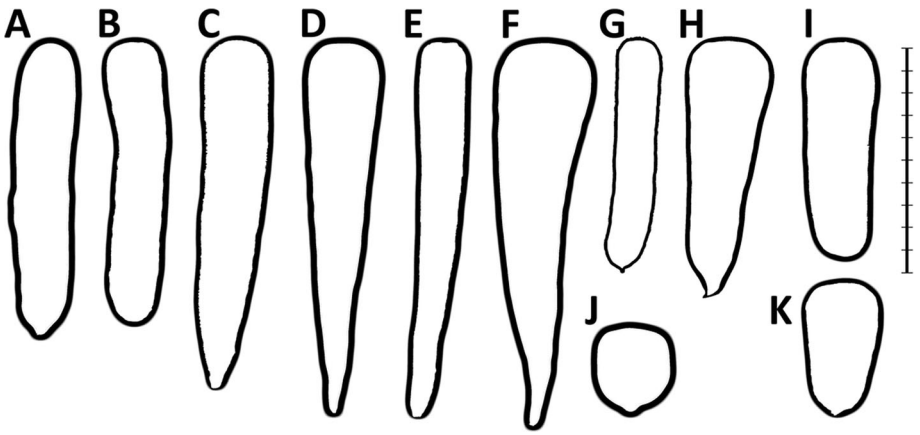


Fig. 1 Typical shapes of carrot storage roots representing some of the most popular market types. A - Nantes, B - Berlikum, C - Autumn King (= Flakkee), D – Long Orange, E - Imperator, F – St. Valery, G - Amsterdam, H - Danvers, I - Chantenay, J – Paris Market, K - Guérande (=Oxheart). Bar length – 20 cm

(Additional file 1: Figure S1), were used to evaluate genetic diversity in the collection of 78 western type carrot cultivars. For *DcS*-ILP genotyping the number of alleles N_a was 1.676 and the number of effective alleles N_e was 1.411, whereas for SNP genotyping the corresponding N_a and N_e values were 1.783 and 1.512, respectively. The observed heterozygosity H_O was higher for SNPs (0.323) than for *DcS*-ILPs (0.253), as well as the expected heterozygosity H_E (0.295 for SNPs and 0.239 for *DcS*-ILPs). H_E estimates of the cultivars ranged from 0.115 (LC1) to 0.323 (BE7) for *DcS*-ILPs and from 0.174 (LO1) to 0.350 (GU3) for SNPs. (Additional file 2: Table S1 and Additional file 3: Table S2). The mean percentage of polymorphic loci segregating within a cultivar varied from 31.18% (LC1) to 89.25% (BE7) for *DcS*-ILP markers and from 45.11% (LC1) to 91.29% (SV1) for SNP markers (Additional file 4: Table S3) revealing high levels of intra-cultivar heterogeneity, in contrast to their apparent phenotypic stability. Cultivars belonging to the Amsterdam market class were characterized by the lowest mean within-group H_O (0.255 and 0.316 for *DcS*-ILP and SNP, respectively), whereas cultivars of the Imperator market class were characterized by the highest mean within-group H_O (0.277 for *DcS*-ILPs and 0.329 for SNPs; Additional file 5: Table S4). The within-group genetic diversity (measured as H_E) was generally lower than H_O and ranged between 0.204 (Chantenay) and 0.267 (Berlikum) for *DcS*-ILP and between 0.269 (Chantenay) and 0.336 (St. Valery) for SNPs. Average inbreeding coefficient F_{IS} for all cultivars was negative for both *DcS*-ILP (– 0.055) and SNP (– 0.097), again indicating high levels of intra-cultivar heterogeneity. The overall genetic differentiation across all market classes, as measured by F_{ST}, was comparable for both marker systems (0.294 for *DcS*-ILPs and 0.279 for SNPs; Table 1).

The F_{ST} analysis suggests a moderate level of differentiation between market classes. The strongest differentiation was observed between the cultivars belonging to the Amsterdam and St. Valery market classes (0.260 for *DcS*-ILP and 0.214 for SNPs; Additional file 5: Table S4). The within-group H_O values were higher than the pairwise F_{ST} values, indicating that within-cultivar genetic variability had greater contribution to overall diversity than between-cultivar variability. Pairwise F_{ST} estimates computed for pairs of cultivars were comparable for both marker systems and ranged from 0.046 (IM3 vs. IM5) to 0.332 (LC1 vs. AM1) for *DcS*-ILP markers, and from 0.052 (NA2 vs. NA3) to 0.323 (LC1 vs. AM1) for SNPs (Additional file 6: Table S5 and Additional file 7: Table S6). The highest percent of pairwise-F_{ST} ranged between 0.1 and 0.15 (47.5% for DcS-ILP and 55.6% for SNPs; Fig. 2). AMOVA of both *DcS*-ILP and SNP genotyping data showed there was more genetic variation observed within the studied cultivars (71% for *DcS*-ILP and 68% for SNPs) than among them (29% for *DcS*-ILP and 32% for SNPs), further underlying the presence of significant amounts of heterogeneity within carrot OP cultivars.

Assessment of genetic structure using a model-based approach

The possible genetic structure within western cultivated carrots was inferred without any prior classification. The

Table 1 F-statistics over all cultivars for all loci resulting from *DcS*-ILP and SNP genotyping

***DcS*-ILP**				**SNP**			
	Fis	**Fit**	**Fst**		**Fis**	**Fit**	**Fst**
Mean	−0.055	0.252	0.294	**Mean**	−0.097	0.209	0.279
SE	0.014	0.013	0.007	**SE**	0.002	0.002	0.001

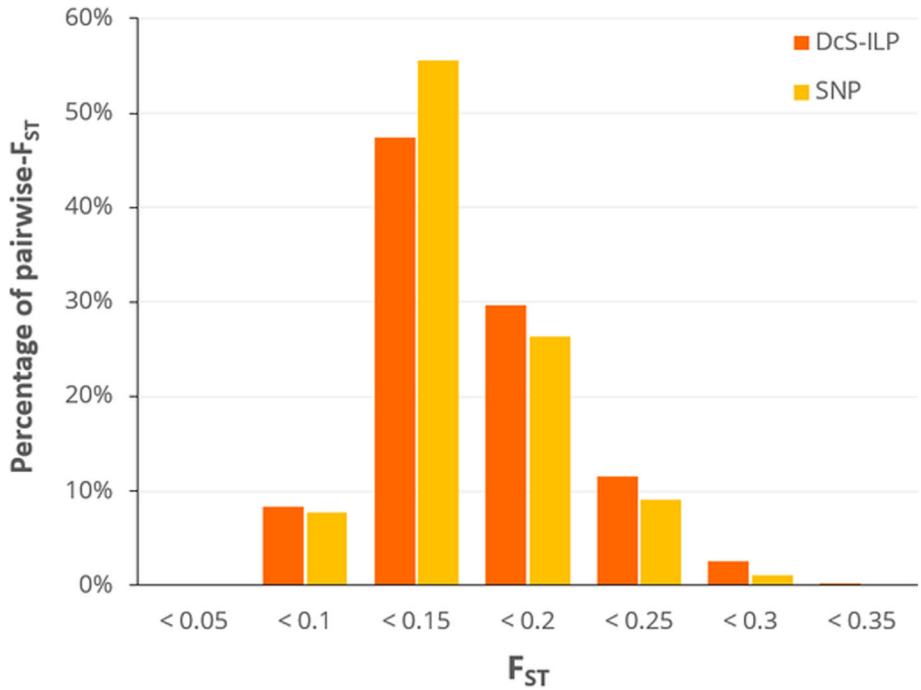


Fig. 2 Comparison of the distribution of estimated pairwise F_{ST} between the 78 studied OP carrot cultivars

entire collection of 390 carrot plants was analyzed using an admixture model-based clustering (Fig. 3). ΔK values (Additional file 8: Table S7) suggested that the most likely number of clusters for *DcS*-ILP genotyping was three, four or seven, with K = 4 being the most probable (ΔK = 61.498), while for SNP genotyping it was three, four and five, with K = 3 being the most probable (ΔK = 133.541). Therefore, a more detailed assessment and comparison of the genetic structure was conducted for K ranging from three to five, and for K = 7.

In general, the genetic structure inferred from the SNP data was characterized by a greater number of cultivars assigned to the assumed clusters (member coefficient values (Q) ≥ 0.5) as compared to *DcS*-ILP data. The percentage of populations assigned was as followed: 98.7% for K = 3, 88.5% for K = 4, 78.2% for K = 5 and 74.4% for K = 7. For *DcS*-ILP data the percentage was generally 10 to 20 points lower and varied from 69.2% for K = 5 to 78.2% for K = 3. However, the number of cultivars attributed to clusters with high confidence (> 0.7) was higher for the *DcS*-ILP data set when K was larger than 3, as compared to the SNP data set (Table 2, Fig. 4).

SNP markers

For K = 3, the most probable number of clusters, clear separation of populations representing the Amsterdam and Baby Nantes market types was observed (group K1, Fig. 5). This pattern was noticeable regardless of the

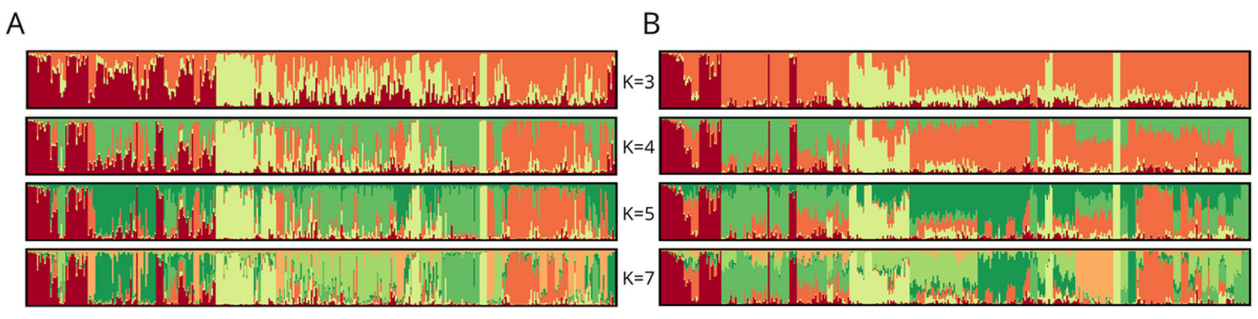


Fig. 3 Estimated genetic structure of the 390 carrot plants representing 78 cultivars. A – genetic structure inferred using 93 *DcS*-ILP markers for K = 3–5 and K = 7; B – genetic structure inferred using 2354 SNP markers for K = 3–5 and K = 7. Each plant is represented by vertical line divided into colored segments representing the membership fractions in the K clusters

Table 2 The number of cultivars assigned to clusters (K) based on the value of membership coefficient (Q)

	SNP			*DcSto*		
	Q ≥ 0.7	0.7 > Q ≥ 0.5	Q < 0.5	Q ≥ 0.7	0.7 > Q ≥ 0.5	Q < 0.5
	number of cultivars					
K = 3	61	16	1	35	26	17
K = 4	34	35	9	44	12	22
K = 5	30	31	17	35	19	24
K = 7	33	25	20	40	20	18

Legend: Q ≥ 0.7 indicates low level of admixture; 0.7 > Q ≥ 0.5 indicates moderate level of admixture, whereas Q < 0.5 indicates high level of admixture. Cultivars with Q < 0.5 were not assigned to any of the inferred clusters

increasing number of the assumed clusters (up to K = 7). The average value of Q within K1 was very high (0.89; Table 3B). Only one cultivar of Amsterdam type (AM4) was characterized by high level of admixture, whereas cultivar AM5 was consequently grouped with cultivars belonging to the Nantes market type. The other clearly separated cluster (K3), that virtually did not change despite of increasing number of K, consisted of eight cultivars belonging to the Chantenay type. The average Q value within this group was high as well (0.78). Group K2 consisted of populations representing diverse market types, e.g. Autumn King (AU), Berlikum (BE) and Imperator (IM) characterized by long, stump roots; or Paris Market (PA), Guerande (GU), Danvers (DA) and Nantes (NA) typically of shorter (short to medium), thicker conical roots. The average distances between individuals within the assumed clusters (measured by H_E) were highest for K2 (0.40), whereas for the clusters more homogeneous with respect to the origin of cultivars - K1 and K3, they were 0.31 and 0.30, respectively (Table 4B).

Increasing the number of clusters to four resulted in the separation of K2 into two separate clusters: the new cluster K2 comprised of AU/BE/FO and cluster K4 comprised of PA/GU/DA/NA. When assuming five clusters, six cultivars belonging to the Nantes market type were separated from K4 creating cluster K5. When the number of clusters was increased to seven, five cultivars belonging to the Imperator type were clearly separated (K6) with Q values above 0.85. The seventh group (K7) consisted mostly of cultivars attributed to the Danvers and St. Valery market types. When K = 7, for four cultivars at least one the plants representing the cultivar was assigned to the different cluster than the majority, thus represented different gene pool. Within another eight cultivars at least one of the plants representing the cultivar was admixed (Q < 0.5) and therefore could not be assigned to any of the defined clusters.

DcS-ILP markers

When the presence of three clusters was assumed, two clusters (K1 and K2) grouped numerous populations of diverse origin such as Amsterdam, Berlikum, Autumn King in group K1 or Nantes, Imperator and St. Valery in group K2. Within K1 Amsterdam cultivars were characterized by high values of Q with mean of 0.85, whereas cultivars of other types were characterized by lower values of Q, not exceeding 0.77. The overall mean Q value within K1 was 0.74 (Table 3A). Eight out of ten analyzed cultivars of Chantenay market type were grouped with two cultivars of Guerande type and two cultivars of Paris Market type in one cluster K3 (Fig. 6). The mean Q for Chantenay cultivars within K3 was 0.89, whereas the overall mean Q for K3 was 0.79. When four clusters were assumed, the previous cluster K1 was reduced mainly to cultivars of the Amsterdam type, thus

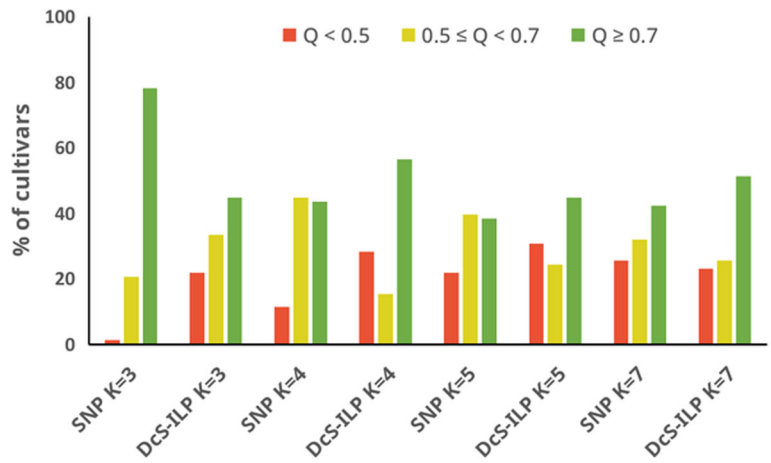


Fig. 4 Percentage of cultivars assigned to the assumed clusters. Assignment is based on different thresholds (Q ≥ 0.5 and Q ≥ 0.7) of the membership coefficients using SNP and *DcS*-ILP markers for the most probable numbers of K: K = 3–5 and K = 7. Q < 0.5 indicates a high level of admixture within a cultivar

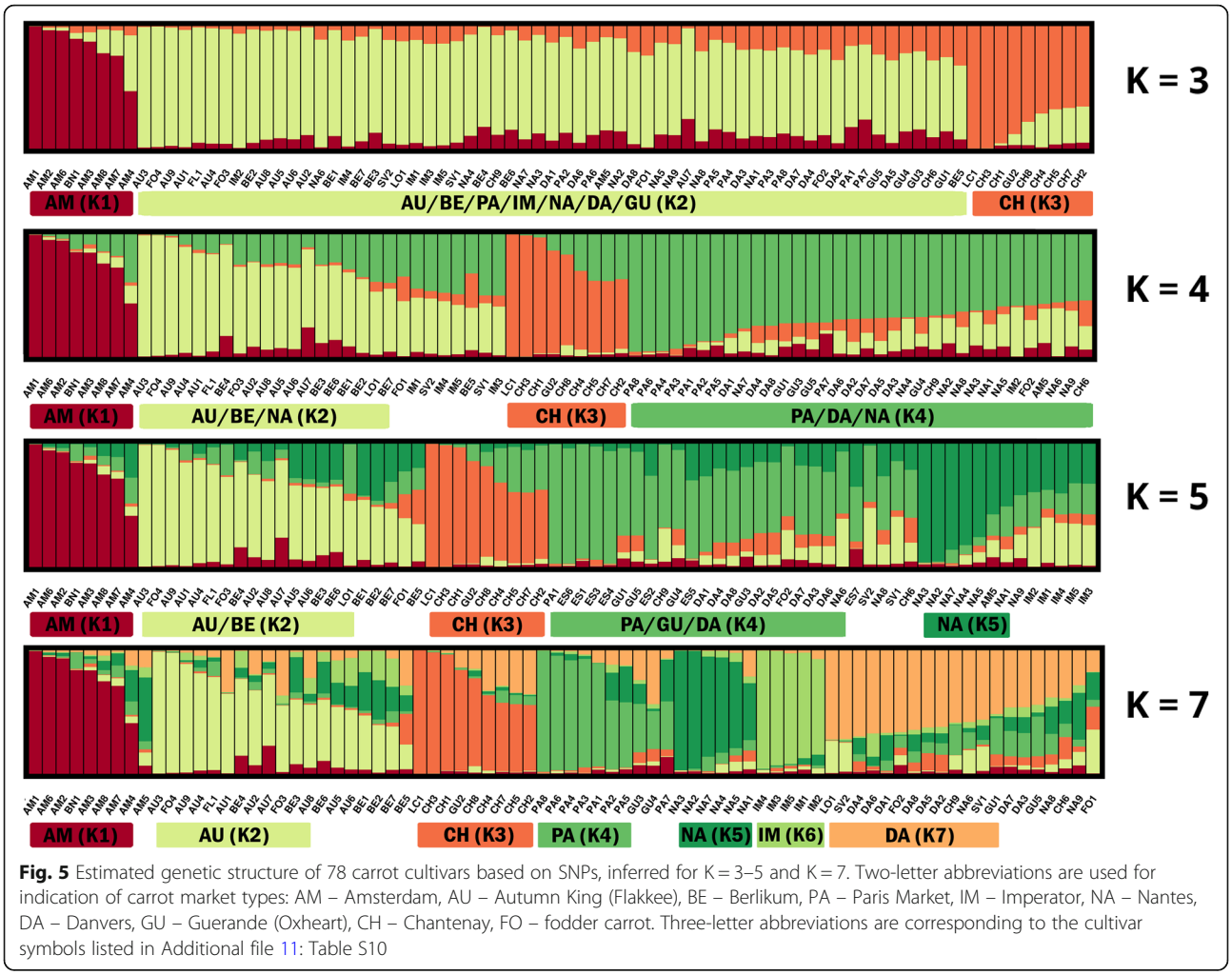


Fig. 5 Estimated genetic structure of 78 carrot cultivars based on SNPs, inferred for K = 3–5 and K = 7. Two-letter abbreviations are used for indication of carrot market types: AM – Amsterdam, AU – Autumn King (Flakkee), BE – Berlikum, PA – Paris Market, IM – Imperator, NA – Nantes, DA – Danvers, GU – Guerande (Oxheart), CH – Chantenay, FO – fodder carrot. Three-letter abbreviations are corresponding to the cultivar symbols listed in Additional file 11: Table S10

Table 3 Mean values of membership coefficient (Q) within inferred groups (*K1-K7*) of carrot cultivars

A	Group						
No. of clusters	***K1***	***K2***	***K3***	***K4***	***K5***	***K6***	***K7***
K = 3	0.74	0.70	0.79	–	–	–	–
K = 4	0.85	0.75	0.87	0.74	–	–	–
K = 5	0.84	0.64	0.87	0.74	0.75	–	–
K = 7	0.83	0.73	0.82	0.76	0.67	0.80	0.77
B	**Group**						
No. of clusters	***K1***	***K2***	***K3***	***K4***	***K5***	***K6***	***K7***
K = 3	0.89	0.79	0.81	–	–	–	–
K = 4	0.87	0.72	0.77	0.72	–	–	–
K = 5	0.86	0.72	0.76	0.68	0.78	–	–
K = 7	0.86	0.70	0.74	0.80	0.84	0.88	0.64

Legend: A - membership coefficient values obtained from *DcS*-ILP genotyping; B - membership coefficient values obtained from SNP genotyping; columns *K1* to *K7* represent groups of carrot cultivars; rows K = 3 to K = 7 represent the most probable number of groups inferred in the course of STRUCTURE analysis

Table 4 Average distances between individuals in the inferred groups (*K1-K7*) of carrot cultivars

A	Group						
No. of clusters	***K1***	***K2***	***K3***	***K4***	***K5***	***K6***	***K7***
K = 3	0.32	0.34	0.25	–	–	–	–
K = 4	0.28	0.33	0.24	0.33	–	–	–
K = 5	0.27	0.34	0.23	0.33	0.31	–	–
K = 7	0.26	0.34	0.22	0.28	0.28	0.31	0.33
B	**Group**						
No. of clusters	***K1***	***K2***	***K3***	***K4***	***K5***	***K6***	***K7***
K = 3	0.31	0.40	0.30	–	–	–	–
K = 4	0.31	0.37	0.28	0.39	–	–	–
K = 5	0.31	0.38	0.28	0.38	0.34	–	–
K = 7	0.31	0.35	0.27	0.34	0.33	0.35	0.34

Legend: A – average genetic distance obtained from *DcS*-ILP genotyping; B - average genetic distance obtained from SNP genotyping; columns *K1* to *K7* represent groups of carrot cultivars; rows K = 3 to K = 7 represent the most probable number of groups inferred in the course of STRUCTURE analysis

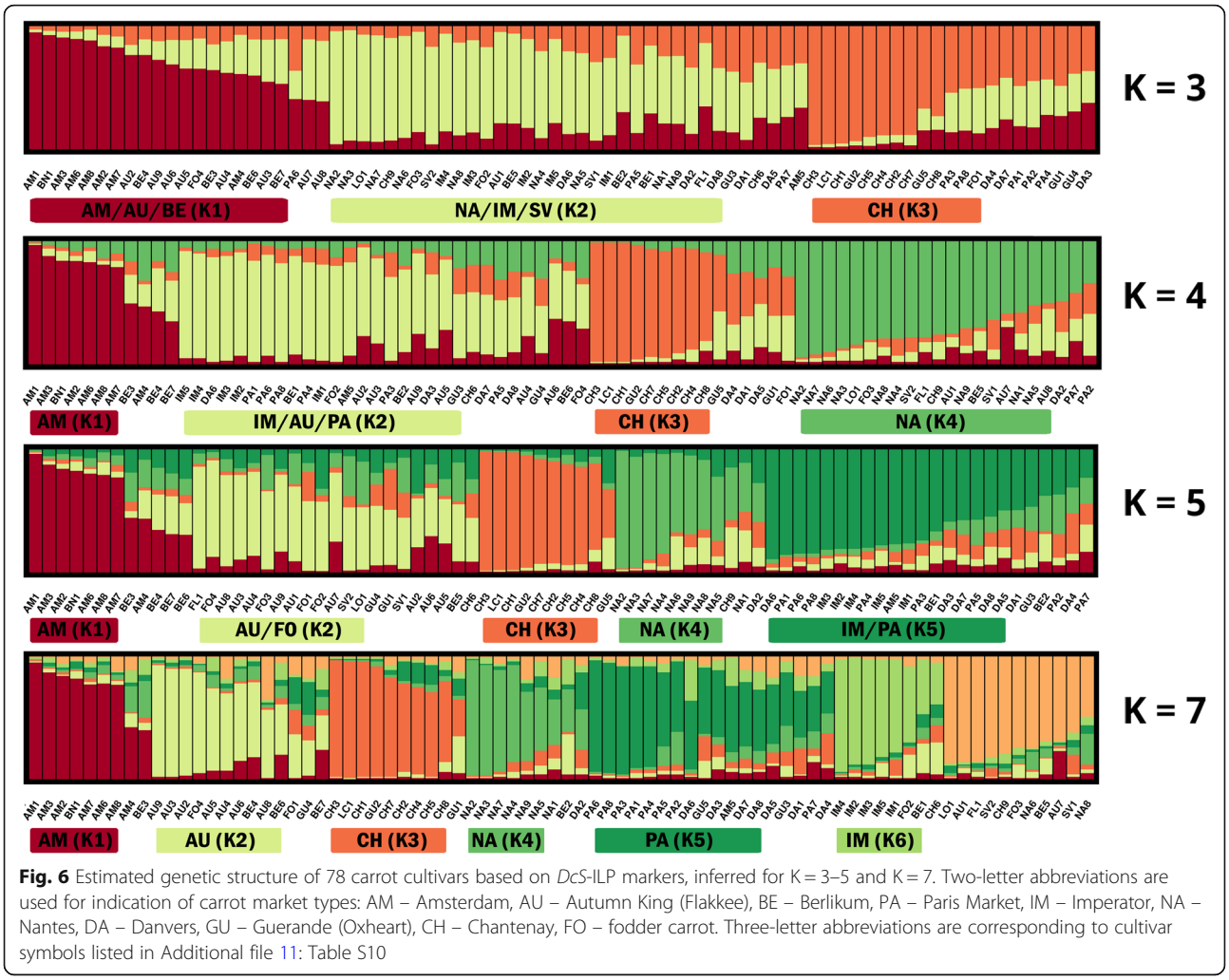


Fig. 6 Estimated genetic structure of 78 carrot cultivars based on *DcS*-ILP markers, inferred for K = 3–5 and K = 7. Two-letter abbreviations are used for indication of carrot market types: AM – Amsterdam, AU – Autumn King (Flakkee), BE – Berlikum, PA – Paris Market, IM – Imperator, NA – Nantes, DA – Danvers, GU – Guerande (Oxheart), CH – Chantenay, FO – fodder carrot. Three-letter abbreviations are corresponding to cultivar symbols listed in Additional file 11: Table S10

the mean Q value has increased to 0.85. Group K2 consistently comprised populations representing various root types such as Imperator, Autumn King and Paris Market with the mean Q value of 0.75. Group K3 was reduced to seven cultivars of Chantenay type and one belonging to Guerande market type and as such remained in spite of the increasing number of clusters (with mean Q = 0.87). For K = 4, newly separated group K4 consisted of 19 cultivars, nine of which belonged to the Nantes type (mean Q = 0.74). Increasing the assumed number of clusters to five resulted in separation of K5 (with mean Q = 0.75) from K2 group. K5 comprised of cultivars belonging to the Imperator and Paris Market types. For K = 5, K2 grouped seven cultivars belonging to the Autumn King market type, four populations of fodder carrot and single population of both Long Orange and St. Valery. The mean Q value for this cluster was 0.64. K4 was restricted exclusively to cultivars representing the Nantes type (mean Q = 0.74). When seven clusters were assumed, populations representing the Imperator type were grouped into K6 along with one population of fodder carrot (Q = 0.80). Group K7 was heterogeneous and comprised eleven cultivars of diverse market types. The mean Q value for that group was 0.77. For K = 7 17 cultivars were of mixed cluster assignment, whereas within another 22 cultivars at least one of the plants representing the cultivar was admixed (Q < 0.5) and therefore could not be assigned to any of the defined clusters. The average H_E between individuals within the assumed clusters were highest for K2 (0.34), whereas for the clusters more homogeneous with respect to the origin of cultivars - K1 and K3, they were of lower values (0.28 and 0.24, respectively; Table 4A).

For the most probable numbers of clusters, K = 3, K = 4 and K = 7 AMOVA analysis was carried out on both *DcS*-ILP and SNP genotyping data. In general, the majority of total genetic variation resulted from differences between cultivars assigned to the predefined clusters.

For *DcS*-ILPs 90–92% of total genetic variation could be attributed to the differences within the clusters, whereas for SNPs the values ranged between 91 to 93%. For K = 3 and K = 4 the percentage of total variation resulting from differences between the assumed clusters was 7% for SNPs and 8% for *DcS*-ILPs. When the number of the predefined clusters was increased to seven, the percentage has risen to 9 and 10% for SNPs and *DcS*-ILPs, respectively.

Discriminant analysis of principal components

The analysis was carried out to provide an alternative non-model de novo grouping of the studied cultivars. The optimal number of groups (K) was found to be eight for both data sets resulting in comparable results of the classification of cultivars (Fig. 7). 65.39% assignment accuracy was observed on the cultivar level, i.e. 51 cultivars were attributed to the same groups 1–8. 277 of 390 plants were assigned to the same groups, resulting in 71.03% assignment accuracy at the plant level. Generally, DAPC enabled the separation of cultivars depending on their market type affiliation. Three groups of cultivars were characterized by high homogeneity regardless of the marker system used for genotyping: Amsterdam (group 1), Chantenay (group 3) and Imperator (group 6). Assignment accuracy on the cultivar level was 71% for group 6 and 100% for group 1 and group 3. Remaining groups were more heterogeneous with regard to classification of specific market types. For *DcS*-ILPs group 4 was the most numerous (Additional file 9: Table S8). Within this group Danvers, Paris Market and Guerande types prevailed. It comprised six cultivars of the Danvers type, four of the Paris Market type, four of the Guerande type together with two cultivars of the Nantes type NA1 and NA5), one of the Amsterdam type (AM5) and one cultivar of fodder carrot (FO2). The second most numerous group was comprised of seven cultivars of the Autumn King type, four cultivars of the Berlikum type and one cultivar of the fodder carrot (FO4). Group 7 was comprised of seven cultivars of the Nantes type. Group 8 comprised cultivars of the St. Valery type together with one Long Orange cultivar (LO1), three cultivars of the Autumn King type (AU1, AU7 and FL1), one cultivar of

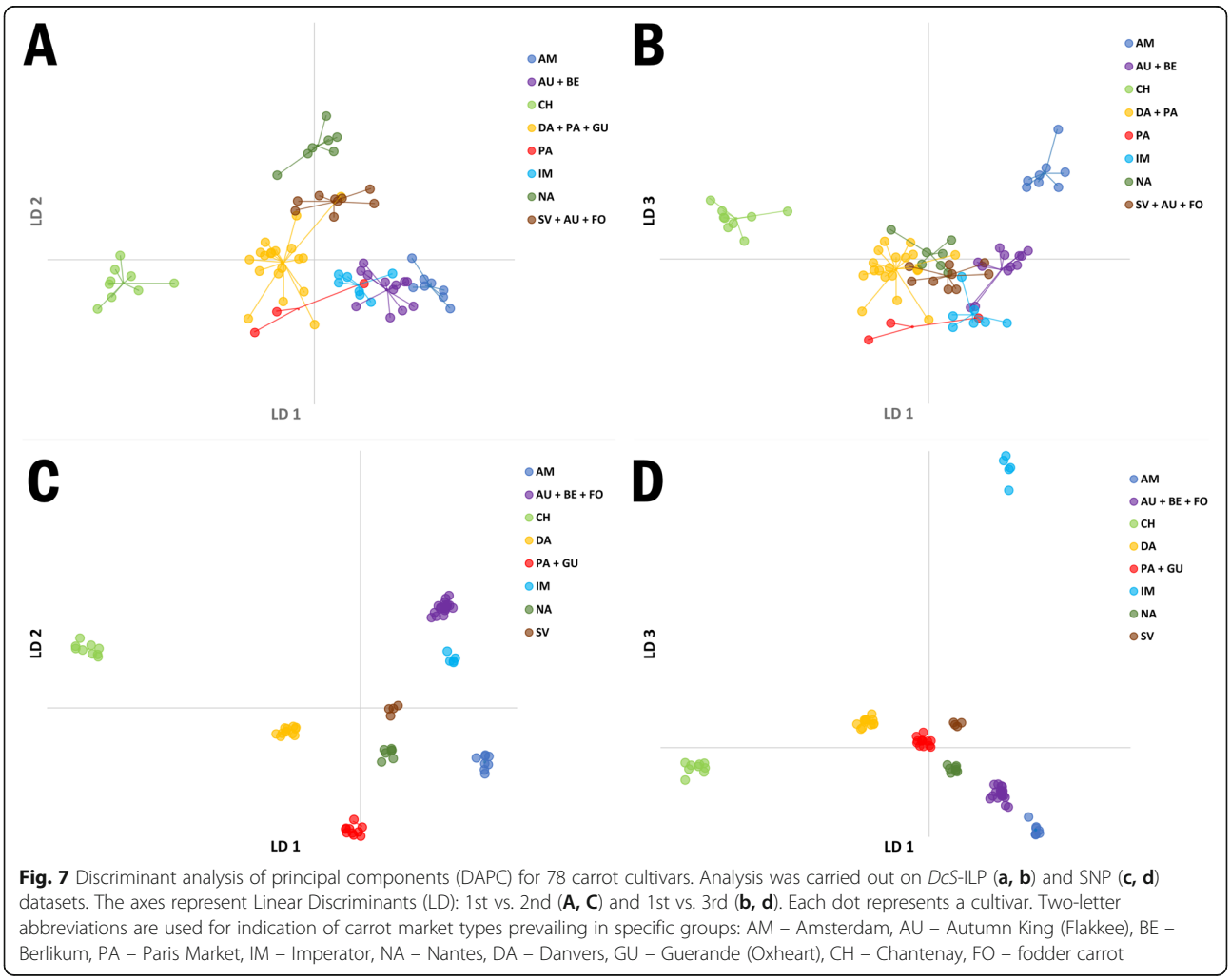


Fig. 7 Discriminant analysis of principal components (DAPC) for 78 carrot cultivars. Analysis was carried out on *DcS*-ILP (**a, b**) and SNP (**c, d**) datasets. The axes represent Linear Discriminants (LD): 1st vs. 2nd (**A, C**) and 1st vs. 3rd (**b, d**). Each dot represents a cultivar. Two-letter abbreviations are used for indication of carrot market types prevailing in specific groups: AM – Amsterdam, AU – Autumn King (Flakkee), BE – Berlikum, PA – Paris Market, IM – Imperator, NA – Nantes, DA – Danvers, GU – Guerande (Oxheart), CH – Chantenay, FO – fodder carrot

Berlikum type (BE5), one cultivar of Chantenay type (CH9) and one cultivar of fodder carrot (FO4). The least numerous group comprised only three cultivars, two of which were of the Paris Market type (PA1, PA8) and one of the Danvers type (DA6).

For SNPs group 2 was the most numerous and comprised 19 cultivars of the Autumn King and Berlikum types together with fodder carrot (Additional file 10: Table S9). Thirteen cultivars of the Paris Market and Guerande types and one cultivar of the Chantenay type (CH9) were placed in group 5, whereas 8 cultivars of the Danvers type together with four cultivars, each of different type (BE5, CH6, NA8, FO2), were placed in group 4. Seven cultivars belonging to the Nantes type were grouped together with one cultivar of the Amsterdam type (AM5). Group 8 comprised only four cultivars: two of them of the St. Valery type, one of the Nantes type (NA6) and one of the Long Orange type (LO1).

Discussion

Previous studies carried out on diverse collections of wild and cultivated carrots suggested there was no distinctive genetic structure within neither wild nor cultivated (western and eastern) carrot [3, 8, 9, 11, 13, 16, 20]. Nonetheless, the above-mentioned studies were carried out on large sets of very diverse carrot germplasm. Because of the apparent and very distinctive genetic differences between wild and cultivated gene pools, they might not have been able to detect the structure of genetic diversity present within the group of western OP cultivars. Historically, the first of the market classes of the western cultivated carrot were selected in the eighteenth century from groups of cultivars showing similar storage root morphologies [21]. As such, the cultivars classified to one market class might have been of relatively close kin. However, as previously reported by Iorizzo et al. [1], no significant bottleneck was observed in the cultivated carrot, pointing at a possible continuous gene flow between the wild and the cultivated pools, and very likely also among different cultivars. Conventional selection based on the plant morphology, leading to the development of the existing types of carrot cultivars showing distinct morphological and agronomic characteristics, apparently allowed retention of significant amounts of genetic heterogeneity within OP cultivars.

The codominant *DcS*-ILP marker system exploited in the present study might reveal genetic variability which arose more recently as *DcSto* MITEs show extreme insertional polymorphism within the carrot genome [19, 22]. Possibly, their recent mobilisation could have led to the genetic diversification within the western carrot gene pool. Despite many advantages, high throughput molecular marker systems, such as SNPs or DArTs are not able to detect transposable element (TE) insertion-derived variability. The resolving power of the *DcS*-ILP panel was previously demonstrated on the collection of 23 OP cultivars of western type carrot [18]. The panel of high quality SNPs bears advantages of cost-efficient throughput sequencing-derived markers but is reduced to ca. 2300 loci evenly-distributed across the genome and referred to the high-quality genome assembly [1], thus providing time- and computing efficiency when exploited for the evaluation of genetic structure within the larger datasets. Both panels of markers can easily be extended by additional loci to gain extra biological information or to possibly modify the resolution of population structure.

The results of the AMOVA, together with high values of H_O, on both cultivar and market class level indicated a significantly higher level of intra-cultivar genetic diversity, mainly contributing to the overall genetic diversity observed in the investigated collection of cultivars. This observation is in accordance with previous studies of Maksylewicz and Baranski [10], as they indicated that almost two third of the of total variation observed in highly diverse collection of carrot cultivars and landraces was attributed to intra-population variation. The values of inbreeding coefficients in the collections of both advanced cultivars and landraces in the studies carried out by Baranski et al. [9] and Maksylewicz and Baranski [10] indicated the excess of homozygous loci that could suggest the repeated selfing during breeding programs aimed at the production of uniform, advanced cultivars. However, using a much more robust set of polymorphisms we did not observe positive inbreeding coefficients in our collection of OP cultivars.

Cultivars classified as Chantenay, Amsterdam and Paris Market types showed lower gene diversity (measured as H_E) among the 11 predefined market classes, whereas cultivars classified to the Berlikum, St. Valery and Imperator types were among the most heterogenous. STRUCTURE clustering led to the decrease in the most probable number of groups from 11 predefined market classes to the most probable three to five or seven clusters. Non-model DAPC grouping also indicated the lesser number of relatively homogenous groups in the examined collection of cultivars. The choice of the most probable number of clusters was more ambiguous for *DcS*-ILP markers as the differences in the ΔK value were relatively small, but generally the increase of the number of defined clusters resulted in lower fraction of the unassigned cultivars together with the increase of the average membership coefficient (Q) within the clusters. This tendency was reversed in the case of SNP genotyping data. The more clusters were extracted, the lower mean values of Q within the groups were observed. Nonetheless, most of the cultivars belonging to the Amsterdam and Chantenay types were always clearly separated from other varieties and characterized by the highest values of within-group Q together

with the lowest within-group H_E. According to the classification of carrot market types proposed by Banga [14] both the Amsterdam and the Chantenay types represent the 'Horn' group that comprises a vast selection of high-quality carrot cultivars subdivided to at least eight market classes (Fig. 8). The Amsterdam market type refers to forcing carrot cultivars grown under covers or for early production in the open. The use of Amsterdam cultivars was originally limited to an early production under covers. Breeders were seeking plant material characterised by a high yield and considerable length, together with the vigour being the key feature in forcing varieties. In the nineteenth century only few varieties could meet the expectations, with *Utrecht Forcing* among them. The modern Amsterdam OP cultivars are believed to be direct descendants of the *Utrecht Forcing* [14].

To date, there is no evidence pointing out that any other breeding material was used to develop Amsterdam cultivars. It is in accordance with our results of STRUCTURE and DAPC clustering, indicating a relatively strong distinctiveness of Amsterdam carrots from other market types, possibly allowing preservation of its specific agricultural characteristics. The admixed nature of *Amsterdamska* cv. possibly reflects the use of Nantes breeding material in the course of the cultivar development. The Chantenay market type comprises cultivars developed for the production between half-summer and late winter carrots. This type is considered as a deviation from main types representing 'Late Half Long Horn' group of market types and was developed as a parallel selection to Guerande type and are believed to originate from *la race de Hollande* cv [14].. However, the majority of Chantenay cultivars was grouped in one clearly distinctive cluster with very high membership values. Regardless of the molecular marker genotyping system, two cultivars, i.e. *Chantenay Long Type* and *Criolla*, were characterized by high levels of admixture. Interestingly, in both cultivars the major genetic components were derived from the market classes with shared ancestor according to the classification proposed by Banga [14]. The two major genetic components of *Chantenay Long Type* cultivar pointed towards the cross between breeding materials belonging to the Danvers and Imperator types. Since *Chantenay Long Type* originated in the U.S., the breeding material of Danvers type could be introduced in the course of the cultivar development, especially because this market type originated in the U.S. in 1870s, and is still used for bunching [14]. The relatively high proportion of genetic components originating from Imperator might be the effect of the crosses between Chantenay and Imperator aimed at obtaining longer storage root. Similarly, the reason for clustering of *Criolla* cultivar (originally classified as Chantenay) with cultivars of the Danvers type could be in the use of the most easily accessible parental breeding material originating from North America. It highlights possible discrepancies between the passport data attributing a cultivar to a particular market class relative to its phenotype and its actual pedigree. Ma et al. [13] reported that Chinese orange carrots, sharing many morphological characteristics with western orange carrots, clustered with Chinese red carrots, suggesting that Chinese orange cultivars and landraces could have emerged from original Asian carrot varieties. It shows that similar phenotypic traits can result from selection from different gene pools. Thus, the presented results possibly reveal the actual genetic diversity within the western carrot gene pool, coupled with remarkable intra-cultivar heterogeneity and significant levels of admixture.

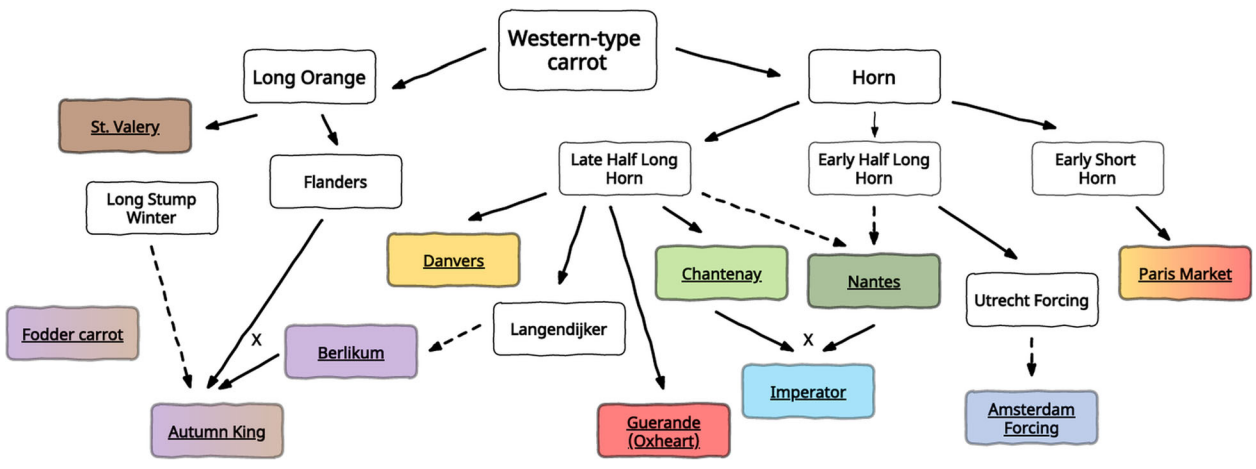


Fig. 8 Schematic representation of the origin of main types of Western carrot as proposed by Banga [14]. Solid arrows show direction of the development of new cultivar types. Punctuated arrows indicate possible origin of the particular type. Underlined names indicate the types of cultivars investigated in the present study. Colours of the boxes represent particular market types or groups of market types clustered using DAPC method for both *DcS*-ILP and SNP data (see Fig. 7)

Conclusions

The aim of the current study was to detect possible genetic structure underlying phenotypic differences among market types of western carrot. We exploited two molecular marker systems, SNPs and TE insertion-derived DcS-ILPs, to provide the tool for time- and cost-efficient evaluation of larger datasets. Both marker systems enabled detection of substantial variation among carrot plants of different market types, therefore can be used in germplasm characterization and analysis of genome relationships. Both model-based STRUCTURE clustering and non-model DAPC grouping indicated the reduction in the number of relatively homogenous groups of OP cultivars in comparison with the classification based primarily on phenotypic traits. The presented results likely reveal the actual genetic diversity within the western carrot gene pool and point at possible discrepancies within the cultivars' passport data.

Methods

Plant material and DNA extraction

Carrot open-pollinated cultivars used in this study were obtained from the Warwick Genetic Resource Unit (WGRU) (Additional file 11**: Table S10**). A total of 390 plants representing 78 OP western-type carrot cultivars (five plants per cultivar) of various tap root shape and market types were grown in the field in Gołębiew (Poland), in 2014 and 2016, under standard agricultural practice, optimally irrigated, fertilized and protected from pathogens. DNA was isolated from fresh young leaves using a modified CTAB protocol (Briard et al., 2000).

DcS-ILP genotyping and SNP discovery by GBS

Genotyping with the use of 93 *DcS*-ILP markers (Additional file 12: Table S11) was performed as described by Stelmach et al. [18]. The *DcS*- ILP marker profiles were scored manually. Each allele was scored as: 1 (empty insertion site), 2 (occupied insertion site) or 0 (lack of amplification). The codominant marker matrix with diploid individuals was created (Additional file 13: Table S12). For SNP discovery plants were genotyped using genotyping-by-sequencing (GBS) at the University of Wisconsin-Madison Biotech Center, carried out as described by Ellison et al. [23]. SNPs were called using Tassel 5.2.31 [24, 25] and the reference carrot genome LNRQ00000000.1 [1]. Polymorphisms were filtered using the VCFtools [26]. Only high quality SNPs (parameters: --max-alleles 2 --max-missing-count 95 --maf 0.1 --minDP 8) were retained and the resulting vcf file was subsequently recoded to the STRUCTURE format matrix using plink 1.9 software [27] (Additional file 14: Table S1).

Data analysis

Genetic diversity indices such as: number of alleles (N_a), number of effective alleles (N_e), observed heterozygosity (H_O) and expected heterozygosity (H_E) were calculated for both *DcS*-ILP and SNP codominant marker systems using GenAlEx 6.51 [28]. Pairwise F_{ST} was estimated using FinePop2 R package [29]. Genetic diversity structure was investigated with STRUCTURE 2.3 [30]. Bayesian clustering was carried out on both *DcS*-ILP and SNP genotyping data matrices. The length of the burn-in period was set to 100,000 and the number of Markov Chain Monte Carlo (MCMC) replications after the burn-in were assigned at 500,000 for each number of clusters (K) set from 2 to 11 (the number of the predefined market classes). Five independent iterations with an admixture and correlated allele frequencies model were performed for each simulated value of *K. no* prior knowledge about the origin of the analyzed populations was used. The most informative K was identified using ΔK value as described by Evanno et al. [31]. To evaluate differentiation among the most probable number of subpopulations, Nei's genetic distances and pairwise F_{ST} estimations were calculated in GenAlEx 6.51. Analysis of molecular variance (AMOVA) was also carried out on codominant genotyping distance matrices in GenAlEx 6.51 with 999 permutations. To provide an assessment of genetic diversity of the studied collection without prior assumptions on the population structure, we conducted Principal Component Analysis (PCA) followed by Discriminant Analysis of Principal Components (DAPC) using *adegenet* 2.1.1 R package [32]. Analyses were carried out on both *DcS*-ILP and SNP datasets. In DAPC, the optimal numbers of retained principal components were determined using a cross-validation (*xval* function implemented in *adegenet*). 150 and 60 principal components explaining 76% (SNP) and 89% (*DcS*-ILP) of the total variance were retained for the DAPC analysis of SNP and *DcS*-ILP datasets, respectively. The number of groups was determined de novo using *find.clusters()* function implemented in *adegenet*. The optimal K was selected based on the decreasing values of Bayesian Information Criterion (BIC). Individuals were then assigned to the clusters.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12870-021-02980-0.

Additional file 1: Figure S1. Genomic distribution of SNP and *DcS*-ILP markers on nine chromosomes of the carrot genome; Legend: The black vertical bars correspond to the position of SNP and DcS-ILP markers; the names of *DcS*-ILPs are bold maroon.

Additional file 2: Table S1. *DcS*-ILP panel summary statistics for the analysed collection of carrot cultivars.

Additional file 3: Table S2. SNP panel summary statistics for the analysed collection of carrot cultivars.

Additional file 4: Table S3. Comparison of the percentage of polymorphic loci observed within analysed cultivars.

Additional file 5: Table S4. Mean H_O and H_E estimates and pairwise F_{ST} values obtained for the predefined market classes.

Additional file 6: Table S5. Pairwise cultivar estimates of F_{ST} based on 93 *DcS*-ILP markers.

Additional file 7: Table S6. Pairwise cultivar estimates of F_{ST} based on 2354 SNP markers.

Additional file 8: Table S7. Estimation of the optimum number of clusters based on Evanno's ΔK method.

Additional file 9: Table S8. Results of DAPC grouping carried out on *DcS*-ILP genotyping data.

Additional file 10: Table S9. Results of DAPC grouping carried out on SNP genotyping data.

Additional file 11: Table S10. Characteristics of the 78 carrot cultivars used in the study.

Additional file 12: Table S11. Description of 93 *DcS*-ILP markers used for genotyping the collection of 78 OP carrot cultivars.

Additional file 13: Table S12. The codominant *DcS*-ILP marker matrix obtained for the collection of 390 carrot plants.

Additional file 14: Table S13. The codominant SNP marker matrix obtained for the collection of 390 carrot plants.

Acknowledgements
Not applicable.

Authors' contributions
KS and DG designed the study; KS prepared plant material and preformed DcS-ILP genotyping; AM-P and KS performed in silico analysis of GBS data; KS performed the assessment of genetic diversity; KS, DG and CA wrote the manuscript. All authors read, reviewed, and approved the final manuscript.

Funding
The research was financed from funds for basic research on biological progress in agriculture for years 2014–2020 granted by the Polish Ministry of Agriculture and Rural Development; KS was supported by the Polish National Science Centre (Narodowe Centrum Nauki), programme ETIUDA 7, project no. 2019/32/T/NZ9/00198.

Availability of data and materials
The datasets supporting the conclusions of this article are included within the article and its additional files or are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Plant Biology and Biotechnology, University of Agriculture in Krakow, Al. 29 Listopada 54, 31-425 Kraków, Poland. [2]School of Life Sciences, University of Warwick, Warwick, UK.

Received: 9 March 2021 Accepted: 15 April 2021
Published online: 26 April 2021

References
1. Iorizzo M, Ellison S, Senalik D, Zeng P, Satapoomin P, Huang J, et al. A high-quality carrot genome assembly provides new insights into carotenoid accumulation and asterid genome evolution. Nat Genet. 2016;48(6):657–66. https://doi.org/10.1038/ng.3565.
2. Simon PW. Economic and academic importance. In: Simon P, Iorizzo M, Grzebelus D, Baranski R, editors. The Carrot Genome. C. Cole, series editor. Compendium of Plant Genomes. Springer Nature Switzerland AG; 2019. p. 1–8.
3. Iorizzo M, Senalik DA, Ellison SL, Grzebelus D, Cavagnaro PF, Allender C, et al. Genetic structure and domestication of carrot (Daucus carota subsp. sativus) (Apiaceae). Am J Bot. 2013;100(5):930–8. https://doi.org/10.3732/ajb.1300055.
4. Stolarczyk J, Janick J. Carrot: history and iconography. Chron Horticult. 2011; 51:13–8.
5. Banga O. Origin and distribution of the Western cultivated carrot. Genet Agrar. 1963;17:357–70.
6. Vivek BS, Simon PW. Linkage relationships among molecular markers and storage root traits of carrot (Daucus carota L. ssp. sativus). Theor Appl Genet. 1999;99(1-2):58–64. https://doi.org/10.1007/s001220051208.
7. Shim SI, Jørgensen RB. Genetic structure in cultivated and wild carrots (*Daucus carota* L.) revealed by AFLP analysis. Theor Appl Genet. 2000;101:227–33.
8. Bradeen JM, Bach IC, Briard M, le Clerc V, Grzebelus D, Senalik DA, et al. Molecular Diversity Analysis of Cultivated Carrot (*Daucus carota* L.) and Wild Daucus Populations Reveals a Genetically Nonstructured Composition. J Amer Soc Hort Sci. 2002;127:383–91.
9. Baranski R, Maksylewicz-Kaul A, Nothnagel T, Cavagnaro PF, Simon PW, Grzebelus D. Genetic diversity of carrot (*Daucus carota* L.) cultivars revealed by analysis of SSR loci. Genetic Resources Crop Evolution. 2012;59:163–70.
10. Maksylewicz A, Baranski R. Intra-population genetic diversity of cultivated carrot (*Daucus carota* L.) assessed by analysis of microsatellite markers. Acta Biochim Pol. 2013;60:753–60.
11. Grzebelus D, Iorizzo M, Senalik D, Ellison S, Cavagnaro P, Macko-Podgorni A, et al. Diversity, genetic mapping, and signatures of domestication in the carrot (Daucus carota L.) genome, as revealed by diversity arrays technology (DArT) markers. Mol Breed. 2014;33(3):625–37. https://doi.org/10.1007/s11032-013-9979-9.
12. Arbizu CI, Ellison SL, Spooner DM, Senalik D, Simon PW. Genotyping-by-sequencing provides the discriminating power to investigate the subspecies of Daucus carota (Apiaceae). BMC Evol Biol. 2016;16(1):234. https://doi.org/10.1186/s12862-016-0806-x.
13. Ma ZG, Kong XP, Liu LJ, Ou CG, Sun TT, Zhao ZW, et al. The unique origin of orange carrot cultivars in China. Euphytica. 2016;212(1):37–49. https://doi.org/10.1007/s10681-016-1753-8.
14. Banga O. Main types of the western carotene carrot and their origin. Zwolle: N.V. Uitgevers-Maatschappij W.E.J. Tjeenk Willink; 1963.
15. Simon PW, Freeman R, Vieira J, Boiteux LS, Nothnagel T, Michalik B, et al. Carrot. In: Prohens J, Nuez F, editors. Vegetables: Fabaceae, Liliaceae, Solanaceae, and Umbelliferae. New York: Springer New York; 2008. p. 327–57. https://doi.org/10.1007/978-0-387-74110-9_8.
16. Luby CH, Goldman IL. Improving freedom to operate in carrot breeding through the development of eight open source composite populations of carrot (*Daucus carota* L. var. sativus). Sustain. 2016;8:479.
17. Rubatzky VE, Quiros CF, Simon PW. Carrots and related vegetable Umbelliferae. New York: CAB International, Wallingford; 1999.
18. Stelmach K, Macko-Podgórni A, Machaj G, Grzebelus D. Miniature Inverted Repeat Transposable Element Insertions Provide a Source of Intron Length Polymorphism Markers in the Carrot (*Daucus carota* L.). Front Plant Sci. 2017;8:725.
19. Macko-Podgórni A, Stelmach K, Kwolek K, Grzebelus D. Stowaway miniature inverted repeat transposable elements are important agents driving recent genomic diversity in wild and cultivated carrot. Mob DNA. 2019;10(1):47. https://doi.org/10.1186/s13100-019-0190-3.
20. Arbizu C, Ruess H, Senalik D, Simon PW, Spooner DM. Phylogenomics of the carrot genus (Daucus, Apiaceae). Am J Bot. 2014;101:1–20.
21. Banga O. The development of the original European carrot material. Euphytica. 1957;6:64–76.
22. Macko-Podgorni A, Nowicka A, Grzebelus E, Simon PW, Grzebelus D. DcSto: carrot stowaway-like elements are abundant, diverse, and polymorphic. Genetica. 2013;141(4-6):255–67. https://doi.org/10.1007/s10709-013-9725-6.
23. Ellison SL, Luby CH, Corak KE, Coe KM, Senalik D, Iorizzo M, et al. Carotenoid presence is associated with the or gene in domesticated carrot. Genetics. 2018;210(4):1497–508. https://doi.org/10.1534/genetics.118.301299.
24. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. TASSEL: software for association mapping of complex traits in diverse

samples. Bioinformatics. 2007;23(19):2633–5. https://doi.org/10.1093/bioinformatics/btm308.
25. Glaubitz JC, Casstevens TM, Lu F, Harriman J, Elshire RJ, Sun Q, et al. TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. PLoS One. 2014;9(2):e90346. https://doi.org/10.1371/journal.pone.0090346.
26. Danecek P, Auton A, Abecasis G, Albers C, Banks E, DePristo M, et al. The variant call format and VCFtools. Bioinformatics. 2011;27(15):2156–8. https://doi.org/10.1093/bioinformatics/btr330.
27. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559–75. https://doi.org/10.1086/519795.
28. Smouse PE, Banks SC, Peakall R. Converting quadratic entropy to diversity: both animals and alleles are diverse, but some are more diverse than others. PLoS One. 2017;12(10):e0185499. https://doi.org/10.1371/journal.pone.0185499.
29. Nakamichi R, Kishino H, Kitada S. FinePop: Fine-Scale Population Analysis. 2020. https://cran.r-project.org/web/packages/FinePop. Accessed 27 Dec 2020.
30. Pritchard JK, Wen X, Falush D. Structure software: version 2.2.3. Univ Chicago, Chicago; 2008.
31. Evanno G, Regnaut S, Goudet J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol Ecol. 2005;14(8):2611–20. https://doi.org/10.1111/j.1365-294X.2005.02553.x.
32. Jombart T. Adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics. 2008;24(11):1403–5. https://doi.org/10.1093/bioinformatics/btn129.

Publisher's Note

10. Oświadczenia dotyczące udziału kandydata i współautorów

UNIWERSYTET ROLNICZY
im. Hugona Kołłątaja w Krakowie

Wydział Biotechnologii i Ogrodnictwa
Katedra Biologii Roślin i Biotechnologii

Oświadczenia o udziale współautorów w publikacji

Oświadczam, że w publikacji Stelmach K., Macko-Podgórni A, Machaj G, Grzebelus D. 2017. Miniature inverted repeat transposable element insertions provide a source of intron length polymorphism markers in the carrot (*Daucus carota* L.). *Frontiers in Plant Science*, 8:725. doi: 10.3389/fpls.2017.00725 mój udział związany był z sformułowaniem problemu badawczego, opracowaniem markerów *DcS*-ILP, wykonaniem analiz laboratoryjnych, interpretacją wyników oraz przygotowaniem wstępnej wersji i opracowaniem ostatecznej wersji manuskryptu.

Katarzyna Stelmach

mgr inż. Katarzyna Stelmach

PL 31-425 Kraków, Al. 29 Listopada 54
tel. +48 12 662 51 87, e-mail: kbrib@urk.edu.pl

UNIWERSYTET ROLNICZY
im. Hugona Kołłątaja w Krakowie

Wydział Biotechnologii i Ogrodnictwa
Katedra Biologii Roślin i Biotechnologii

Oświadczenia o udziale współautorów w publikacji

Oświadczam, że w publikacji Stelmach K., Macko-Podgórni A, Machaj G, Grzebelus D. 2017. Miniature inverted repeat transposable element insertions provide a source of intron length polymorphism markers in the carrot (*Daucus carota* L.). *Frontiers in Plant Science*, 8:725. doi: 10.3389/fpls.2017.00725 mój udział związany był z zaprojektowaniem doświadczenia, opracowaniem markerów *DcS*-ILP oraz przygotowaniem wstępnej wersji manuskryptu.

Alicja Macko-Podgórni

dr inż. Alicja Macko-Podgórni, prof. UR

PL 31-425 Kraków, Al. 29 Listopada 54
tel. +48 12 662 51 87, e-mail: kbrib@urk.edu.pl

UNIWERSYTET ROLNICZY
im. Hugona Kołłątaja w Krakowie

Wydział Biotechnologii i Ogrodnictwa
Katedra Biologii Roślin i Biotechnologii

Oświadczenia o udziale współautorów w publikacji

Oświadczam, że w publikacji Stelmach K., Macko-Podgórni A, Machaj G, Grzebelus D. 2017. Miniature inverted repeat transposable element insertions provide a source of intron length polymorphism markers in the carrot (*Daucus carota* L.). *Frontiers in Plant Science*, 8:725. doi: 10.3389/fpls.2017.00725 mój udział związany był z opracowaniem markerów *DcS*-ILP oraz przygotowaniem wstępnej wersji manuskryptu do publikacji.

mgr inż. Gabriela Machaj

PL 31-425 Kraków, Al. 29 Listopada 54
tel. +48 12 662 51 87, e-mail: kbrib@urk.edu.pl

UNIWERSYTET ROLNICZY
im. Hugona Kołłątaja w Krakowie

Wydział Biotechnologii i Ogrodnictwa
Katedra Biologii Roślin i Biotechnologii

Oświadczenia o udziale współautorów w publikacji

Oświadczam, że w publikacji Stelmach K., Macko-Podgórni A, Machaj G, Grzebelus D. 2017. Miniature inverted repeat transposable element insertions provide a source of intron length polymorphism markers in the carrot (*Daucus carota* L.). *Frontiers in Plant Science*, 8,725; doi: 10.3389/fpls.2017.00725 mój udział związany był z sformułowaniem problemu badawczego oraz przygotowaniem ostatecznej wersji manuskryptu do publikacji.

prof. dr hab. Dariusz Grzebelus

PL 31-425 Kraków, Al. 29 Listopada 54
tel. +48 12 662 51 87, e-mail: kbrib@urk.edu.pl

UNIWERSYTET ROLNICZY
im. Hugona Kołłątaja w Krakowie

Wydział Biotechnologii i Ogrodnictwa
Katedra Biologii Roślin i Biotechnologii

Oświadczenia o udziale współautorów w publikacji

Oświadczam, że w publikacji Macko-Podgórni A, Stelmach K., Kwolek K., Grzebelus D. 2019. *Stowaway* miniature inverted repeat transposable elements are important agents driving recent genomic diversity in wild and cultivated carrot. *Mobile DNA*, 10:47. doi: 10.1186/s13100-019-0190-3 mój udział związany był z zaprojektowaniem doświadczenia, wykonaniem analiz *in silico*, wykonaniem analiz laboratoryjnych oraz przygotowaniem ostatecznej wersji manuskryptu.

Alicja Macko-Podgórni

dr inż. Alicja Macko-Podgórni, prof. UR

PL 31-425 Kraków, Al. 29 Listopada 54
tel. +48 12 662 51 87, e-mail: kbrib@urk.edu.pl

UNIWERSYTET ROLNICZY
im. Hugona Kołłątaja w Krakowie

Wydział Biotechnologii i Ogrodnictwa
Katedra Biologii Roślin i Biotechnologii

Oświadczenia o udziale współautorów w publikacji

Oświadczam, że w publikacji Macko-Podgórni A, Stelmach K., Kwolek K., Grzebelus D. 2019. *Stowaway* miniature inverted repeat transposable elements are important agents driving recent genomic diversity in wild and cultivated carrot. *Mobile DNA*, 10:47. doi: 10.1186/s13100-019-0190-3 mój udział związany był z wykonaniem analiz laboratoryjnych.

Katarzyna Stelmach

mgr inż. Katarzyna Stelmach

PL 31-425 Kraków, Al. 29 Listopada 54
tel. +48 12 662 51 87, e-mail: kbrib@urk.edu.pl

UNIWERSYTET ROLNICZY
im. Hugona Kołłątaja w Krakowie

Wydział Biotechnologii i Ogrodnictwa
Katedra Biologii Roślin i Biotechnologii

Oświadczenia o udziale współautorów w publikacji

Oświadczam, że w publikacji Macko-Podgórni A, Stelmach K., Kwolek K., Grzebelus D. 2019. *Stowaway* miniature inverted repeat transposable elements are important agents driving recent genomic diversity in wild and cultivated carrot. *Mobile DNA*, 10:47. doi: 10.1186/s13100-019-0190-3 mój udział związany był z wykonaniem analiz laboratoryjnych.

Kornelia Kwolek

mgr inż. Kornelia Kwolek

PL 31-425 Kraków, Al. 29 Listopada 54
tel. +48 12 662 51 87, e-mail: kbrib@urk.edu.pl

UNIWERSYTET ROLNICZY
im. Hugona Kołłątaja w Krakowie

Wydział Biotechnologii i Ogrodnictwa
Katedra Biologii Roślin i Biotechnologii

Oświadczenia o udziale współautorów w publikacji

Oświadczam, że w publikacji Macko-Podgórni A, Stelmach K., Kwolek K., Grzebelus D. 2019. *Stowaway* miniature inverted repeat transposable elements are important agents driving recent genomic diversity in wild and cultivated carrot. *Mobile DNA*, 10,47; doi: 10.1186/s13100-019-0190-3 mój udział związany był z zaprojektowaniem doświadczenia oraz przygotowaniem ostatecznej wersji manuskryptu.

prof. dr hab. Dariusz Grzebelus

PL 31-425 Kraków, Al. 29 Listopada 54
tel. +48 12 662 51 87, e-mail: kbrib@urk.edu.pl

UNIWERSYTET ROLNICZY
im. Hugona Kołłątaja w Krakowie

Wydział Biotechnologii i Ogrodnictwa
Katedra Biologii Roślin i Biotechnologii

Oświadczenia o udziale współautorów w publikacji

Oświadczam, że w publikacji Stelmach K., Macko-Podgórni A., Allender C., Grzebelus D. 2021. Genetic diversity structure of western-type carrots. *BMC Plant Biology*, 21:200. doi:10.1186/s12870-021-02980-0 mój udział związany był z zaprojektowaniem doświadczenia, przeprowadzeniem analiz *in silico*, przeprowadzeniem analiz laboratoryjnych oraz przygotowaniem wstępnej wersji i opracowaniem ostatecznej wersji manuskryptu.

Katarzyna Stelmach

mgr inż. Katarzyna Stelmach

PL 31-425 Kraków, Al. 29 Listopada 54
tel. +48 12 662 51 87, e-mail: kbrib@urk.edu.pl

UNIWERSYTET ROLNICZY
im. Hugona Kołłątaja w Krakowie

Wydział Biotechnologii i Ogrodnictwa
Katedra Biologii Roślin i Biotechnologii

Oświadczenia o udziale współautorów w publikacji

Oświadczam, że w publikacji Stelmach K., Macko-Podgórni A., Allender C., Grzebelus D. 2021. Genetic diversity structure of western-type carrots. *BMC Plant Biology*, 21:200. doi:10.1186/s12870-021-02980-0 mój udział związany był z przeprowadzeniem części analiz *in silico*.

Alicja Macko-Podgórni

dr inż. Alicja Macko-Podgórni, prof. UR

PL 31-425 Kraków, Al. 29 Listopada 54
tel. +48 12 662 51 87, e-mail: kbrib@urk.edu.pl

UNIWERSYTET ROLNICZY
im. Hugona Kołłątaja w Krakowie

Wydział Biotechnologii i Ogrodnictwa
Katedra Biologii Roślin i Biotechnologii

Oświadczenia o udziale współautorów w publikacji

Oświadczam, że w publikacji Stelmach K., Macko-Podgórni A., Allender C., Grzebelus D. 2021. Genetic diversity structure of western-type carrots. *BMC Plant Biology*, 21,200; doi:10.1186/s12870-021-02980-0 mój udział związany był zaprojektowaniem doświadczenia oraz opracowaniem ostatecznej wersji manuskryptu.

prof. dr hab. Dariusz Grzebelus

PL 31-425 Kraków, Al. 29 Listopada 54
tel. +48 12 662 51 87, e-mail: kbrib@urk.edu.pl

Declaration of author contribution

Stelmach K., Macko-Podgórni A., Allender C., Grzebelus D. 2021. Genetic diversity structure of western-type carrots. *BMC Plant Biology*, 21:200. doi:10.1186/s12870-021-02980-0

I declare that my contribution was following: critical revision of the manuscript.

dr Charlotte Allender

Ponieważ p. dr Ch. Allender
była niedostępna, potwierdzam
Jej udział w publikacji w zakresie
wskazanym w oświadczeniu

Dariusz Grzebelus