Instytut Chemii Bioorganicznej
Polskiej Akademii Nauk

Udział polimorfizmu liczby kopii w kształtowaniu wewnątrzgatunkowej zmienności strukturalnej metabolicznych klastrów genów u *Arabidopsis thaliana*

mgr inż. Małgorzata Marszałek-Zeńczak

Praca doktorska została wykonana w Zakładzie Genomiki Roślin
Promotor: dr hab. Agnieszka Żmieńko, prof. ICHB PAN

Poznań 2023

Składam serdeczne podziękowania:

Promotor dr hab. Agnieszce Żmieńko, prof. ICHB PAN
za opiekę merytoryczną, przekazaną wiedzę, cenne wskazówki i dyskusje
oraz za wsparcie i wyrozumiałość.

Koleżankom i Kolegom z Europejskiego Centurm Bioinformatyki
i Genomiki, Zakładu Genomiki Roślin oraz Zakładu Biologii Molekularnej
i Systemowej za ciekawe rozmowy i dyskusje, życzliwość, chęć pomocy
i miłą atmosferę.

Rodzinie i przyjaciołom za wiarę i wsparcie.

W szczególności dziękuję mojemu mężowi, Michałowi,
za nieocenioną pomoc i wsparcie w każdej sytuacji, wytrwałość, dzielenie się
wiedzą, motywację i za nieustanną wiarę we mnie.

Dziękuję!

Dla moich wspaniałych dzieci,
Zosi i Piotrusia

Przedłożona rozprawa doktorska została sfinansowana
z projektu badawczego Narodowego Centrum Nauki OPUS 7
"Poszukiwanie mechanizmów indukujących polimorfizm liczby kopii
dużych fragmentów genomu oraz związków tego zjawiska
z tworzeniem i ewolucją metabolicznych klastrów genów"
2014/13/B/NZ2/03837

Lista publikacji

wchodzących w skład rozprawy doktorskiej:

1. Samelak-Czajka A, **Marszalek-Zenczak** M, Marcinkowska-Swojak M, Kozlowski P, Figlerowicz M and Zmienko A (2017) MLPA-Based Analysis of Copy Number Variation in Plant Populations. Front. Plant Sci. 8: 222. doi:10.3389/fpls.2017.00222. **5-letni IF** = 4,353

2. Zmienko A, **Marszalek-Zenczak M**, Wojciechowski P, Samelak-Czajka A, Luczak M, Kozlowski P, Karlowski WM, Figlerowicz M (2020) AthCNV: A Map of DNA Copy Number Variations in the Arabidopsis Genome. Plant Cell. 32(6): 1797-1819. doi:10.1105/tpc.19.00640. **5-letni IF** = 12,061

3. **Marszalek-Zenczak M**, Satyr A, Wojciechowski P, Zenczak M, Sobieszczanska P, Brzezinski K, Iefimenko T, Figlerowicz M, Zmienko A (2023) Analysis of Arabidopsis non-reference accessions reveals high diversity of metabolic gene clusters and discovers new candidate cluster members. Front Plant Sci. 14: 1104303. doi:10.3389/fpls.2023.1104303. **5-letni IF** = 7,255

niewchodzacych w skład rozprawy doktorskiej:

1. Butkiewicz D, Krześniak M, Gdowicz-Kłosok A, Giglok M, **Marszałek-Zeńczak M**, Suwiński R (2021) Polymorphisms in EGFR Gene Predict Clinical Outcome in Unresectable Non-Small Cell Lung Cancer Treated with Radiotherapy and Platinum-Based Chemoradiotherapy. Int J Mol Sci. 22(11): 5605. doi:10.3390/ijms22115605. **5-letni IF** = 6,628

2. Samelak-Czajka A, Wojciechowski P, **Marszalek-Zenczak M**, Figlerowicz M, Zmienko A (2023) Differences in the intraspecies copy number variation of Arabidopsis thaliana conserved and nonconserved miRNA genes. Funct Integr Genomics. 23(2): 120. doi:10.1007/s10142-023-01043-x. **5-letni IF** = 3,711

Spis treści

Lista publikacji **4**

Streszczenie **8**

Abstract **10**

1 Wprowadzenie **12**

1.1 Metody detekcji wariantów strukturalnych 14

1.2 Identyfikacja wariantów strukturalnych w genomie człowieka i ich związek z fenotypem 17

1.3 Rola CNV w kształtowaniu zmienności genetycznej i fenotypowej roślin . . 20

1.4 Badania asocjacyjne całego genomu 21

1.5 Arabidopsis jako roślina modelowa w badaniu zmienności wewnątrzgatunkowej 23

1.6 Metabolity wtórne 25

1.7 Roślinne metaboliczne klastry genów 26

1.8 Regulacja metabolicznych klastrów genów 29

2 Cel Pracy **32**

3 Omówienie wyników badań **33**

3.1 Identyfikacja regionów CNV w genomie Arabidopsis 33

3.2 Charakterystyka i analiza porównawcza poziomu zmienności MGC w kontekście ich funkcji 37

3.3 CNV jako użyteczne markery w badaniu struktury populacji i historii demograficznej Arabidopsis 41

3.4 Rola CNV w kształtowaniu zmienności fenotypowej i adaptacji do warunków środowiska 42

4 Podsumowanie i perspektywy **44**

5 Bibliografia **47**

6 Publikacje wchodzące w skład rozprawy doktorskiej 60

Publikacja 1 61

Publikacja 2 79

Publikacja 3 140

7 Oświadczenia współautorów 196

Publikacja 1 197

Publikacja 2 204

Publikacja 3 214

**„Natura jest naszym największym sprzymierzeńcem
i największą inspiracją.”**

David Attenborough

Streszczenie

Sekwencja genomowego DNA jest charakterystyczną cechą każdego żywego organizmu. Nawet genomy indywidualnych osobników w obrębie gatunku różnią się między sobą. Zmienność genetyczna obejmuje różnorodne warianty, od zmian pojedynczych nukleotydów (SNP) aż po duże zmiany strukturalne (SV). Niezbalansowane SV, w których fragment sekwencji DNA jest tracony lub powielony, nazywane są zmiennością liczby kopii (CNV). Większość CNV nie ma znaczącego wpływu na fenotyp osobnika. Jednak niektóre z nich mogą mieć wyraźny - szkodliwy efekt, np. związany z rozwinięciem się choroby, podczas gdy inne ujawniają się lepszą adaptacją osobnika do warunków środowiska. U eukariontów geny zaangażowane we wspólny szlak metaboliczny są zazwyczaj rozproszone w genomie. Jednak w ostatnich latach zidentyfikowano metaboliczne klastry genów (MGC) czyli skupiska niehomologicznych genów zaangażowanych we wspólny szlak metaboliczny. Zarówno CNV jak i MGC często występują w dynamicznych regionach genomu takich jak okolice centromerów czy regiony bogate w transpozony. Oba te zjawiska, choć słabo poznane, wydają się mieć istotny wpływ na kształtowanie genomów roślinnych. W regionach podatnych na rearanżacje strukturalne możliwość połączenia korzystnych zestawów genów jest większa niż w pozostałej części genomu, promując tym samym tworzenie MGC. CNV i MGC mają wiele elementów wspólnych i wydaje się, że udział CNV w kształtowaniu i ewolucji MGC jest duży i znaczący.

Celem prowadzonych badań w ramach mojej rozprawy doktorskiej było określenie wewnątrzgatunkowej zmienności liczby kopii dla rośliny modelowej *Arabidopsis thaliana*, a także analiza, w jakim stopniu CNV wpływają na strukturę i stabilność metabolicznych klastrów genów. Wykorzystując dane z wysokoprzepustowego sekwencjonowania nowej generacji dla ponad 1000 naturalnych linii *A. thaliana*, opracowałam autorskie podejście do integracji wyników z siedmiu programów opartych o trzy główne metody detekcji CNV: analizę głębokości pokrycia, analizę mapowania odczytów sparowanych i analizę mapowania odczytów rozłącznych. Stworzyłam katalog dużych indeli (50-499 pz) oraz wariantów CNV ($\geq$500 pz). Wykazałam, że CNV są ważnymi markerami, które mogą być wykorzystywane w analizach populacyjnych i asocjacji z fenotypem. Następnie skupiłam się na zbadaniu zmienności strukturalnej (głównie CNV) w czterech MGC, w populacji *A. thaliana*. Zaobserwowałam ogromną różnorodność w obrębie badanych klastrów. Klaster genów marneralu wydaje się ustalony na poziomie gatunku. Klaster biosyntezy thalianolu

istnieje w dwóch wersjach. W tym MGC zidentyfikowałam inwersję, obecną aż w 65% badanej populacji, która skutkowała bardziej zwartym klastrem centralnym. Kompaktowa wersja klastra biosyntezy thalianolu była dominująca i bardziej konserwatywna niż wersja nieciągła. Największy, wysoce zmienny i zróżnicowany w populacji jest klaster biosyntezy arabidiolu/baruolu. W tym klastrze zidentyfikowałam dużą (21-27 kpz) insercję genomową obecną w około jednej trzeciej analizowanej populacji. Wewnątrz insercji występuje nowa para genów: *CYP705A2a-BARS2*, przy czym *BARS2* stanowi niereferencyjny gen kodujący nową, dotychczas niescharakteryzowaną syntazę oksydoskwalenu (OSC) w genomie *A. thaliana*. Analiza asocjacji z fenotypem (GWAS) wskazała, że linie z insercją miały wolniejszą dynamikę wzrostu korzeni i były związane z cieplejszym klimatem w porównaniu do linii z referencyjnym układem genów. W korzeniach i liściach, profil ekspresji genów w klastrze arabidiolu/baruolu był różny dla linii z insercją i bez. Dodatkowo, analiza par genów: OSC-oksydazy cytochromu P450 wykazała, że pary genów są bardziej zmienne niż ich niesparowane odpowiedniki.

Uzyskane wyniki badań podkreślają istotny wpływ zmienności genetycznej w formowaniu i kształtowaniu MGC. Duża różnorodność klastrów u *A. thaliana* wskazuje na ich dynamiczną ewolucję zaś wyniki GWAS potwierdzają ich możliwą rolę w różnorodności fenotypowej i adaptacji roślin. Zrozumienie zmienności genetycznej i jej wpływu na organizację genomu jest kluczowe w lepszym zrozumieniu ich funkcji.

Abstract

The genomic DNA sequence is a characteristic feature of every living organism. Even the genomes of individuals within a species vary. Genetic variation ranges from single nucleotide polymorphisms (SNPs) to large structural variations (SVs). Unbalanced SVs, in which a fragment of a DNA sequence is lost or gained, are called copy number variations (CNV). Most CNVs do not have a significant impact on an individual's phenotype. However, some of them may have a deleterious effect e.g. associated with the development of disease, while others contribute to improved adaptation of the individual to environmental conditions. In eukaryotes, genes involved in a common metabolic pathway are usually dispersed throughout the genome. Nevertheless, recent investigations have identified metabolic gene clusters (MGCs) comprising non-homologous genes involved in a shared metabolic pathway. Both CNVs and MGCs are often found in dynamic regions of the genome such as centromere proximity or transposon-rich regions. Both these phenomena, although poorly understood, appear to have important implications for shaping plant genomes. In regions prone to structural rearrangements, the possibility of combining favorable sets of genes is greater than in the rest of the genome, thus promoting the formation of MGCs. CNVs and MGCs have many elements in common, and the contribution of CNVs to the formation and evolution of MGCs seems to be large and significant.

The primary objective of my research was to investigate intraspecific copy number variation in the model plant *Arabidopsis thaliana*, and to analyze to what extent CNVs affect the structure and stability of metabolic gene clusters. Using high-throughput next-generation sequencing data for more than 1,000 natural *A. thaliana* accessions, I developed a pipeline to integrate the results from seven different tools based on three main CNV detection methods: read depth, paired-end mapping and split read. I created a catalog of large indels (50-499 bp) and CNVs ($\geq$500 bp). I demonstrated that CNVs are important markers that can be used in population analyses and in genome-wide association study (GWAS). I then focused on the analysis of structural variation (mainly CNVs) in four MGCs, in the *A. thaliana* population. I observed significant diversity within these studied clusters. The marneral gene cluster appears to be fixed at the species level. The thalianol biosynthesis cluster exists in two versions. In this MGC, I identified an inversion, present in as many as 65% of the studied

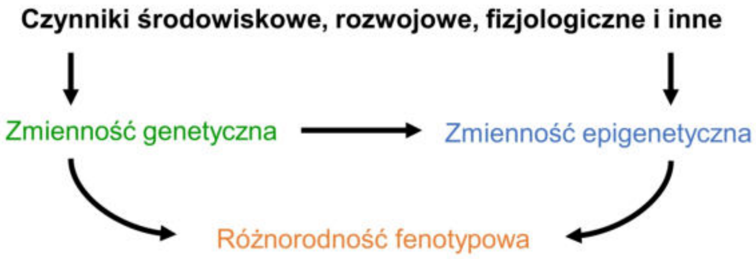
population, which resulted in a more compact central cluster. The compact version of the thalianol biosynthesis cluster was dominant and more conservative than the discontiguous version. The largest, highly variable and diverse in the population is the arabidiol/baruol biosynthesis cluster. In this cluster, I identified a large (21-27 kbp) genomic insertion present in about one-third of the analyzed population. This insertion introduced a new gene pair, *CYP705A2a-BARS2*, where *BARS2* was a non-reference gene encoding a previously uncharacterized oxidosqualene synthase (OSC) in the *A. thaliana* genome. GWAS indicated that accessions with this insertion displayed slower root growth dynamics and were associated with a warmer climate as opposed to accessions with the reference gene arrangement. In roots and leaves, the gene expression profile in the arabidiol/baruol cluster was different for accessions with and without the insertion. In addition, analysis of gene pairs: OSC- cytochrome P450 oxidase showed that the gene pairs were more variable than their unpaired counterparts.

The findings of this research underscore the significant influence of genetic variability on the formation and shaping of MGCs. The high diversity of clusters in *A. thaliana* indicates their dynamic evolution while GWAS results confirm their possible role in phenotypic diversity and plant adaptation to environmental conditions. Understanding genetic variation and its impact on genome organization is crucial to gaining insights into their biological functions.

Rozdział 1

Wprowadzenie

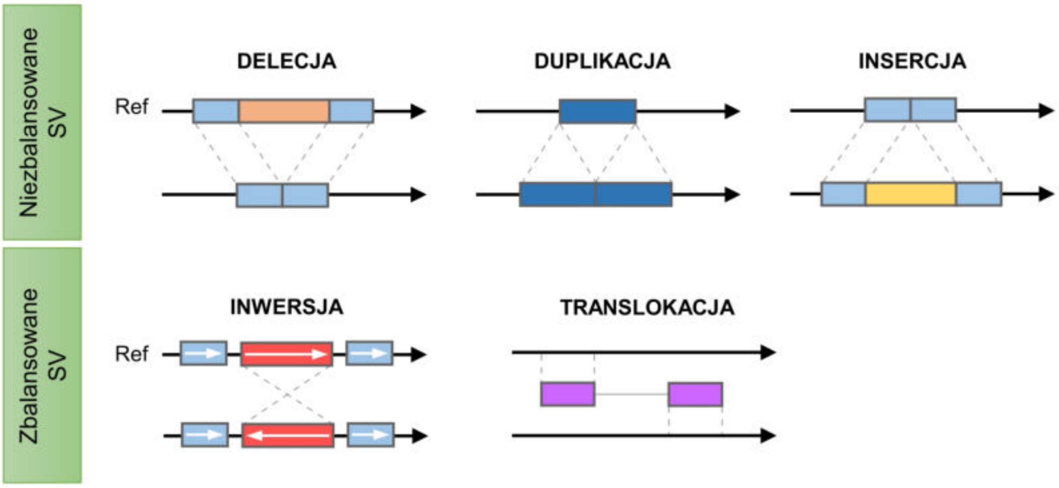
Zrozumienie mechanizmów leżących u podstaw różnorodności fenotypowej, prowadzących do jej powstawania i utrzymywania się w środowisku naturalnym, jest jednym z głównych celów współczesnej biologii. Czynniki wewnątrzkomórkowe oraz rozmaite bodźce środowiskowe mogą prowadzić nie tylko do zmian w sekwencji genomu, ale również powodować zmiany epigenetyczne (odzwierciedlone na poziomie komórkowym m.in. poprzez modyfikacje histonów czy stopień metylacji DNA), prowadząc tym samym do zmian fenotypowych (Lang i wsp. 2016) (Rysunek 1.1).



Rysunek 1.1: Wpływ różnorodnych czynników na zmienność genetyczną i epigenetyczą, które przyczyniają się do różnorodności fenotypowej

Zmienność genetyczna, będąca jednocześnie podstawą doboru naturalnego i procesu ewolucji, wynika z występowania różnic w sekwencjach DNA między osobnikami. Źródłem tej zmienności są mutacje czyli trwałe zmiany w sekwencji genomu. Większość pojawiających się zmian nie wywiera znaczącego wpływu na fenotyp osobnika. Zdarza się jednak, że niosą one ze sobą bardzo wyraźny efekt, który może być szkodliwy, na przykład związany z rozwinięciem się choroby, bądź korzystny, gdy pojawienie się alternatywnych wariantów danej sekwencji genomowej umożliwia adaptację osobników do zmieniających się warunków środowiska, ich przetrwanie, a z czasem pojawienie się nowych cech fenotypowych w populacji. Mutacje występujące w komórkach rozrodczych (germinalnych) mogą zostać przekazane kolejnym pokoleniom, stanowiąc tym samym prawdziwy motor napędowy ewolucji. Gdy dany wariant pojawia się w co najmniej 1% populacji mówimy o zjawisku zwanym „polimorfizmem” (Hollox i wsp. 2022). Dotychczasowe badania nad zmiennością genetyczną ujawniły złożone relacje między genotypem, a fenotypem, przyczyniając się do poszerzenia naszej wiedzy na temat funkcjonowania i różnorodności organizmów.

Zmienność genetyczna obejmuje szeroki wachlarz wielkości wariantów, począwszy od zmian pojedynczych nukleotydów (ang. *Single Nucleotide Polymorphism*, SNP), poprzez małe insercje lub delecje, zwane indelami, aż po duże zmiany strukturalne (ang. *Structural Variations*, SV), które mogą obejmować rozległe segmenty sekwencji DNA, a nawet część lub cały chromosom. Warianty strukturalne stanowią najbardziej zróżnicowaną grupę pod względem typów i wielkości zmian genetycznych. Początkowo tym terminem określano zmiany >1 kpz. Wraz z dynamicznym rozwojem nowych technologii, a zarazem większą czułością i precyzją metod badawczych, granica ta zmniejszyła się i obecnie w literaturze definiuje się je jako zmiany >100 pz, a czasem nawet ≥50 pz (Ho i wsp. 2020, Yuan i wsp. 2021, Hollox i wsp. 2022). SV obejmują regiony DNA, w których obserwuje się różnice w liczbie kopii danego fragmentu sekwencji (delecje, duplikacje, insercje), jego orientacji (inwersje) bądź w położeniu na chromosomie (translokacje) między osobnikami (Rysunek 1.2). Dodatkowo, takie warianty mogą się nakładać lub łączyć tworząc duże i złożone kompleksy, zwane rearanżacjami genomowymi. Występowanie niezbalansowanych wariantów strukturalnych, w których fragment sekwencji DNA jest tracony bądź powielony nazywane jest zmiennością liczby kopii (ang. *Copy Number Variation*, CNV).



Rysunek 1.2: Typy wariantów strukturalnych

Takie regiony w sekwencji genomowej mogą występować zarówno w obszarze niekodującym, jak i kodującym, a także dotyczyć sekwencji regulatorowych genu niosąc ze sobą potencjalne skutki funkcjonalne, takie jak zmiana struktury genu, jego dawki czy też mechanizmów regulacji. Przykładowo, duplikacja całego genu może (choć nie musi) skutkować podwyższeniem poziomu jego ekspresji i produkcji odpowiadających mu białek. Delecja genu pociąga za sobą utratę zarówno całego transkryptu, jak i kodowanego przez niego białka. Z kolei złożone zmiany strukturalne mogą mieć bardziej zróżnicowane skutki.

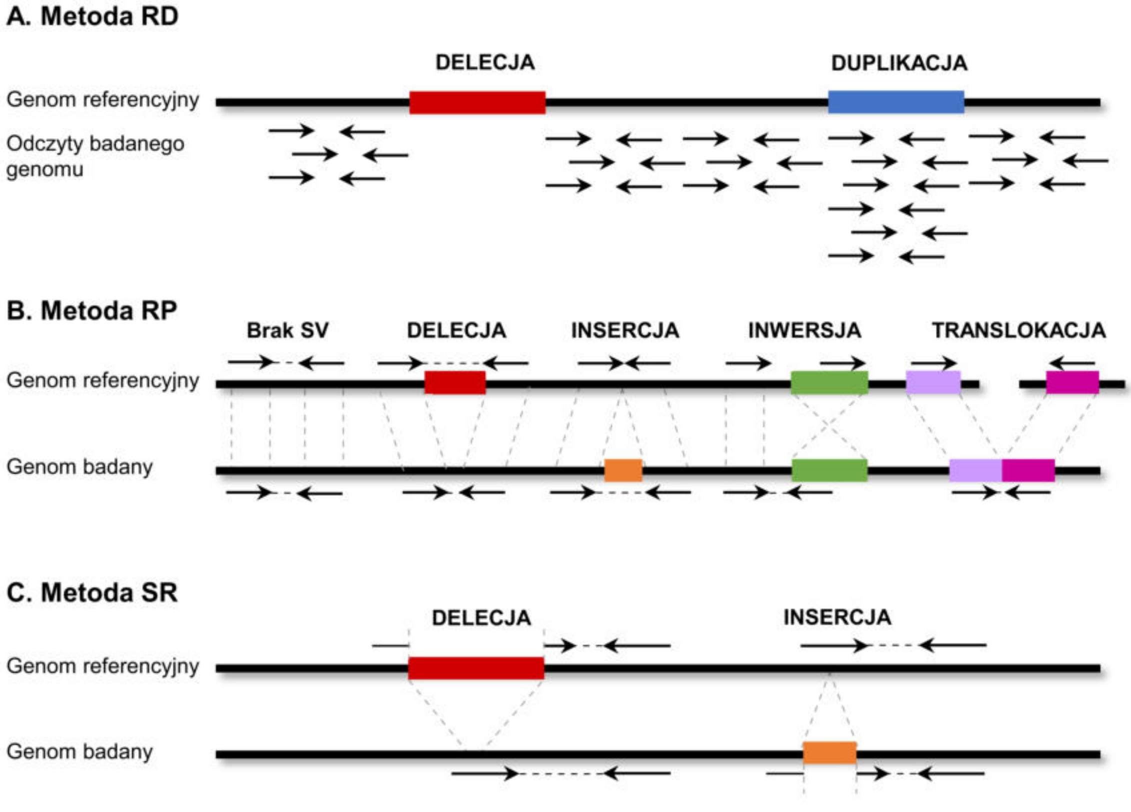
1.1 Metody detekcji wariantów strukturalnych

Badania nad zmiennością genetyczną oraz jej związkami z fenotypem często skupiają się na analizach SNP lub krótkich indeli, zaniedbując przy tym warianty strukturalne. Przyczyną jest duża złożoność problemu badawczego oraz istniejące jeszcze do niedawna istotne ograniczenia technologiczne. Wczesne badania strukturalne genomów prowadzono przy pomocy mikroskopów na poziomie kariotypu, z rozdzielczością >3 Mpz (Feuk i wsp. 2006). Badania wariantów strukturalnych w wyższej rozdzielczości i na szerszą skalę stały się możliwe dzięki opracowaniu metody porównawczej hybrydyzacji genomowej do mikromacierzy (ang. *Array Comparative Genomic Hybridization*, aCGH). Metoda ta pozwala na wykrycie niezbalansowanych różnic (a więc CNV) między dwoma porównywanymi genomami, w wielu loci jednocześnie (Escaramís i wsp. 2015). Niemniej jednak, to podejście nie wykrywa zbalansowanych SV (czyli transolokacji i inwersji), ani też bezwzględnej liczby kopii danego segmentu DNA. Ma również pewne ograniczenia pod względem czułości i swoistości wykrywanych wariantów. Eksperymenty z wykorzystaniem tej metody dostarczyły pierwszych istotnych informacji na temat CNV. Wyniki badań opublikowane w dwóch ważnych pracach przez Sebat i wsp. (2004) i Iafrate i wsp. (2004) wskazywały, że SV niewidoczne na poziomie kariotypu stanowią liczną grupę niezidentyfikowanych jak dotąd zmian w genomie człowieka. Był to bez wątpienia punkt zwrotny w badaniach nad zmiennością genetyczną, który jednocześnie wzbudził szerokie zainteresowanie naukowców analizami CNV. Badania prowadzone w kolejnych latach dowiodły, że SV przyczyniają się co najmniej w takim samym stopniu jak SNP i małe indele do międzyosobniczej zmienności na poziomie genomu (Feuk i wsp. 2006, Escaramís i wsp. 2015).

Prawdziwym przełomem w genetyce był rozwój technik sekwencjonowania DNA. Niespełna 50 lat temu, w 1977 roku, Fred Sanger opracował metodę sekwencjonowania przez syntezę (Sanger i wsp. 1977), otwierając tym samym nową epokę w badaniach nad genomem. W efekcie, na przestrzeni lat poznano sekwencje genomowe licznych organizmów. Z kolei rozwój technik sekwencjonowania nowej generacji (ang. *Next-Generation Sequencing*, NGS) otworzył drogę do masowego, równoległego sekwencjonowania krótkich fragmentów DNA, dając początek analizom opartym na porównaniach międzygatunkowych i wewnątrzgatunkowych. Dynamiczny wzrost ilości generowanych sekwencji pociągnął za sobą problem na etapie przechowywania, obróbki i analizy coraz większych ilości danych, w efekcie czego rozwinęła się bioinformatyka – interdyscyplinarna dziedzina łącząca nauki biologiczne i informatyczne.

Detekcja wariantów strukturalnych przy użyciu krótkich odczytów to złożony problem badawczy. Choć identyfikacja delecji jest stosunkowo łatwa, zarówno duplikacje jak i bardziej złożone rearanżacje stanowią duże wyzwanie. W odpowiedzi na zaistniałą potrzebę nastąpił prężny rozwój algorytmów do detekcji różnych typów SV, w tym również do identyfikacji inwersji i translokacji. W wyniku sekwencjonowania NGS uzyskuje się krótkie sekwencje z jednego końca fragmentu DNA (tzw. pojedyncze odczyty; ang. *single-end reads*), bądź z obu jego końców (tzw. odczyty sparowane; ang. *pair-end reads*). W zależności od zastosowanej technologii odczyty mogą mieć długość od 30 do 600 pz (Giza i wsp. 2021). Do identyfikacji SV można wykorzystać oba typy odczytów, przy czym zastosowanie odczytów sparowanych znacznie rozszerza zarówno możliwości samej detekcji jak i zwiększa dokładność w określaniu końców regionów zmiennych (ang. *breakpoints*). Większość algorytmów bazuje na mapowaniu (przyrównaniu) odczytów badanej próbki do sekwencji genomu referencyjnego. Możliwe jest również wcześniejsze składanie genomów *de novo* (ang. *de novo assembly*, AS), czyli łączenie odczytów w dłuższe fragmenty (tzw. kontigi), a następnie detekcja SV poprzez mapowanie kontigów do genomu odniesienia. Niemniej jednak, w przypadku krótkich odczytów podejście to jest bardzo żmudne i wymagające pod względem jakości danych, a w regionach bogatych w powtórzenia takich jak centromery, telomery czy powtórzenia mikrosatelitarne czasem wręcz niemożliwe do zastosowania.

W przypadku CNV, główne podejścia stosowane do ich detekcji przy użyciu krótkich odczytów to: analiza głębokości pokrycia (ang. *Read Depth*, RD) - identyfikuje CNV zliczając odczyty zmapowane do każdego regionu w genomie; analiza mapowania odczytów sparowanych (ang. *Read Pair*, RP) - wykrywa CNV w oparciu o „niezgodnie zmapowane" odczyty, czyli takie, dla których długość wstawki (odległość między parą odczytów, tzw. *insert size*) znacznie odbiega od wartości średniej z próbki; analiza mapowania rozdzielonych odczytów (ang. *Split Read*, SR) - do identyfikacji CNV wykorzystuje niekompletnie zmapowany odczyt z każdej pary (Rysunek 1.3). Czynniki takie jak głębokość pokrycia, długość odczytu czy średnia odległość między parami odczytu wpływają na dokładność i jakość mapowania, co z kolei bezpośrednio przekłada się na wydajność każdej metody (Yoon i wsp. 2009, Li i Olivier 2013, Singh i wsp. 2021). Metoda RD pozwala na identyfikację duplikacji i delecji o szerokim spektrum wielkości wariantów wykorzystując zarówno pojedyncze jak i sparowane odczyty sekwencjonowania. Dobrze sprawdza się w detekcji dużych zmian, większych niż 1 kpz. Jej słabym punktem jest często niedokładne określanie końców regionu zmiennego, chociaż w zależności od jakości danych oraz przyjętych parametrów podejście to może wykrywać granice z dużą dokładnością. Bez wątpienia dużą zaletą metody RD jest możliwość precyzyjnego określenia liczby kopii. Z kolei podejścia oparte o RP i SR można wykorzystać do identyfikacji wszystkich typów SV, w tym również do inwersji i translokacji. Niemniej jednak wydajność tych metod w regionach zduplikowanych jest bardzo niska, gdyż opierają się one na niezależnym mapowaniu każdego końca. Podejścia te wymagają wysokiej jakości i spójności danych w różnych regionach. Dobrze sprawdzają się w detekcji mniejszych SV (<1 kpz). Oczywistą zaletą tych metod jest duża dokładność i precyzja określenia końców poszczególnych wariantów, a w przypadku SR z rozdzielczością nawet do 1 nukleotydu. Niemniej jednak, wysoka wrażliwość tych podejść skutkuje detekcją bardzo dużej liczby, często wzajemnie nakładających się wariantów, przez co kluczowe jest zastosowanie restrykcyjnych kryteriów filtracji wyników w celu usunięcia regionów fałszywie pozytywnych. Każde z wyżej wymienionych podejść ma pewne ograniczenia pod względem czułości, specyficzności, precyzji w określaniu granic, typów czy wielkości wykrywanych wariantów, stąd często wykorzystuje się podejście integrujące różne metody, które tak naprawdę wzajemnie się uzupełniają (ang. *Combinatorial Approach*, CA) (Zhao i wsp. 2013, Ho i wsp. 2020, Hollox i wsp. 2022) (Tabela 1.1).



Rysunek 1.3: Charakterystyczne wzorce mapowania odczytów dla wybranych SV w trzech metodach detekcji wariantów przy użyciu krótkich odczytów

Tabela 1.1: Porównanie metod do identyfikacji SV w oparciu o analizę krótkich odczytów

Metoda	**Typ SV**	**Wielkość SV**	**Udział wyników fałszywie pozytywnych**	**Rozdzielczość**	**Określenie liczby kopii**
RD	Delecje, duplikacje	Zdolność do identyfikacji dużych zmian (>1 kpz)	Niski	Średnia	Tak
RP	Wszystkie typy SV	Maksymalna wielkość wariantów częściowo uzależniona od długości wstawki, najwyższa czułość przy SV <3000 pz	Średni	Wysoka	Nie
SR	Wszystkie typy SV	Głównie warianty <500 pz	Bardzo wysoki	Bardzo wysoka, nawet do 1 nt	Nie

Technologia sekwencjonowania III generacji zwana również jako sekwencjonowanie długich odczytów (ang. *Long-Read Sequencing*) została uznana przez *Nature Methods* Metodą Roku 2022 (Marx 2023). Polega ona na sekwencjonowaniu pojedynczych molekuł (cząsteczek kwasów nukleinowych) w czasie rzeczywistym, w wyniku czego uzyskuje się tzw. długie odczyty. Podejście to pozwala na uzyskanie odczytów o długości znacznie większej niż poprzednie technologie, przy średnim współczynniku błędu wynoszącym (w przypadku niektórych technologii) zaledwie 1%. Średnia długość odczytów wynosi 10–15 kpz, jednak ultra długie odczyty mogą sięgać kilkuset tysięcy pz, a czasem osiągać długość ponad miliona nukleotydów (Athanasopoulou i wsp. 2021, Marx 2023). Długie odczyty sekwencjonowania umożliwiają analizę i składanie problematycznych dla NGS regionów w genomie m.in. centromerów czy telomerów, zawierających wysoce powtarzalne sekwencje czy bogatych w różne rearanżacje genomowe. Pozwalają one na bardzo skuteczną asemblację prawie kompletnych genomów eukariotycznych, co umożliwia nowe podejście do identyfikacji SV poprzez bezpośrednie porównanie właściwej sekwencji badanych regionów np. między osobnikami, z pominięciem mapowania odczytów do genomu referencyjnego. Na chwilę obecną czynnikiem limitującym jest kosztowność przeprowadzenia takich badań. Ponadto, trudności na etapie analizy, zarówno podczas składania *de novo* samych genomów jak i podczas identyfikacji wariantów wymagają obecnie eksperckiej wiedzy z zakresu bioinformatyki.

1.2 Identyfikacja wariantów strukturalnych w genomie człowieka i ich związek z fenotypem

Z oczywistych względów, najbardziej prężne działania prowadzone są w badaniach nad genomem człowieka. Począwszy od lat 90-tych, najlepsze instytucje badawcze z całego świata łączą swoje siły w ramach konsorcjów m.in. *The 1000 Genomes Project Consortium*, *Telomere-to-Telomere* (T2T), *Human Pangenome Reference Consortium*

(HPRC) czy *Human Genome Structural Variation Consortium* (HGSVC), generując sekwencje genomowe możliwie największej liczby osób, aby jak najlepiej poznać sekwencję ludzkiego genomu oraz scharakteryzować i zrozumieć pełne spektrum zmienności genetycznej człowieka. Ostatni rok okazał się prawdziwym przełomem w tej dziedzinie. W marcu 2022 r., w *Science*, opublikowano kompletną, ciągłą sekwencję ludzkiego genomu haploidalnego (łącznie z centromerami) dla wszystkich chromosomów z wyjątkiem Y, jednocześnie wypełniając luki obecne w poprzedniej wersji genomu referencyjnego (Nurk i wsp. 2022). Miesiąc później, w tym samym czasopiśmie, ukazała się praca charakteryzująca segmentalne duplikacje (ang. *Segmental Duplication*, SD) w nowo złożonym genomie (Vollger i wsp. 2022). SD to bardzo dynamiczne regiony i uważa się, że odgrywają znaczącą rolę w ewolucji. Wyniki badań ujawniły, że większość luk w poprzedniej wersji genomu pokrywała się właśnie z regionami SD, zaś w odniesieniu do nowej, kompletnej sekwencji genomu, regiony te stanowią ok. 7% całości. Najdłuższe powtórzenia zlokalizowano na krótkich ramionach chromosomów akrocentrycznych. Ponadto, globalna analiza ekspresji i metylacji między genami zduplikowanymi i unikalnymi sugeruje, że aż dwie trzecie zduplikowanych genów jest wyciszanych epigenetycznie. Najnowsza publikacja prezentująca wstępną wersję pierwszego ludzkiego pangenomu, która ukazała się w *Nature*, stanowi kolejny krok milowy w badaniach nad zmiennością (Liao i wsp. 2023). Termin „pangenom" został zaproponowany przez Tettelin i współbadaczy w 2005 r. podczas analizy ośmiu szczepów chorobotwórczego paciorkowca *Streptococcus agalactiae* (Tettelin i wsp. 2005). Celem badania pangenomu jest poznanie całości zmienności genetycznej u badanej grupy blisko spokrewnionych osobników np. należących do tego samego gatunku, a także identyfikacja tzw. genów rdzenia (ang. *core genes*), czyli zestawu genów obecnych u wszystkich bądź większości badanych osobników. Analiza polega na porównywaniu sekwencji genomowych wielu osobników, identyfikacji podobieństw i różnic między nimi oraz stworzeniu swoistego katalogu zmienności badanej grupy, który będzie punktem odniesienia w badaniach porównawczych. W ramach pilotażowej fazy projektu pangenomu człowieka, pangenom stworzono z sekwencji 47 osób wyselekcjonowanych z różnych regionów świata. Docelowo projekt zakłada analizę sekwencji DNA pochodzących od 350 osób. Już wstępne wyniki badań pokazują ogromny potencjał pangenomiki. Zastosowanie pangenomu umożliwiło identyfikację dwukrotnie większej liczby SV na osobę, niż było to możliwe przy użyciu liniowego genomu referencyjnego GRCh38.

Spośród genów zlokalizowanych w regionach polimorficznych początkowo niewiele wiadomo było o ich znaczeniu funkcjonalnym. Znaczna część badań skupiała się na poszukiwaniu związków CNV z ryzykiem rozwoju różnych chorób u ludzi (Shaikh 2017). Wykazano, że liczne CNV są zaangażowane w etiologię m.in. zespołu Pradera-Williego i Angelmana (Vogels i Fryns, 2002), złożonych chorób neuropsychiatrycznych, w tym zaburzeń ze spektrum autyzmu (Sebat i wsp. 2007) i schizofrenii (Walsh i wsp.

2008) oraz chorób neurodegeneracyjnych, w tym choroby Parkinsona (Singleton i wsp. 2003) i choroby Alzheimera (Sekine i Makino 2017). Ponadto, w przypadku wieloallelicznych CNV, czyli takich które występują w populacji w wielu wariantach liczby kopii, powiązano obecność pewnych częstych wariantów (>1%) z kilkoma złożonymi chorobami lub cechami np. podatnością na zakażenie wirusem HIV, chorobami autoimmunologicznymi, takimi jak reumatoidalne zapalenie stawów i choroba Leśniowskiego-Crohna, czy cukrzycą typu 2 (Girirajan i wsp. 2011, Shaikh 2017). Wykazano również, że CNV są najczęstszymi mutacjami somatycznymi obserwowanymi w genomach nowotworów, głównie wpływającymi na liczbę kopii genów supresorowych nowotworów i protoonkogenów (Lee i wsp. 2007, Hu i wsp. 2018).

Obecne możliwości technologiczne otwierają zupełnie nowe perspektywy do tego typu badań. Coraz większym zainteresowaniem cieszą się badania dotyczące roli zmienności strukturalnej, w tym polimorfizmu liczby kopii w adaptacji i ewolucji organizmów (Hollox i wsp. 2022). U człowieka, ciekawym przykładem pokazującym elastyczność genomu i jego możliwości adaptacji jest zależność między dietą, a liczbą kopii genów kodujących amylazę - enzym ważny dla trawienia pokarmów bogatych w skrobię, takich jak zboża czy ziemniaki. Amylaza kodowana jest przez dwa geny: *AMY1*, kodujący tzw. amylazę ślinową i *AMY2*, kodujący tzw. amylazę trzustkową. Szacuje się, że w światowej populacji liczba kopii genu *AMY1* wynosi 2–18 kopii, natomiast *AMY2* 0–4 kopii. Badania pokazują, że istnieje pozytywna korelacja zarówno między populacjami z dietą bogatą w skrobię, a wyższą liczbą kopii *AMY1* jak i liczbą kopii obu genów *AMY1* i *AMY2* (Perry i wsp. 2007, Carpenter i wsp. 2015). Biorąc pod uwagę, że archaiczne genomy homininów i genomy szympansów mają tylko jedną kopię *AMY1*, można podejrzewać, że wysoka liczba kopii jest niedawną adaptacją do diety bogatej w skrobię związanej z silnym rozwojem rolnictwa. Coraz więcej badań pokazuje również, że choroby zakaźne m.in. malaria, przyczyniły się do ukształtowania specyficznej dla populacji adaptacji (Williams i wsp. 2005, Louzada i wsp. 2020). Malaria, choroba przenoszona przez komary w tropikalnych i subtropikalnych strefach klimatycznych, jest wywoływana przez pierwotniaki m.in. *Plasmodium falciparum* i *Plasmodium vivax.* Zakażenie malarią stanowi główną, infekcyjną przyczynę śmierci wśród dzieci w Afryce. Składnik hemoglobiny, α-globina, kodowana jest przez dwa geny: *HBA1* i *HBA2.* Zazwyczaj u ludzi występują dwie kopie każdego genu na genom diploidalny, chociaż istnieją zarówno warianty delecyjne jak i duplikacje tych genów. Wykazano, że allele delecyjne są utrzymywane w wyższych częstościach w Afryce Subsaharyjskiej (do 20%) i stanowią selektywną ochronę przed ciężką malarią (Williams i wsp. 2005). Co ciekawe, u człowieka prawie wszystkie SV, dla których zaproponowano adaptacje specyficzne dla populacji, obejmowały adaptacje do diety lub miały związek z podatnością na choroby (Hollox i wsp. 2022).

1.3 Rola CNV w kształtowaniu zmienności genetycznej i fenotypowej roślin

Rośliny, jako organizmy osiadłe, doskonale nadają się do śledzenia i badania zmienności wewnątrzgatunkowej. Poszczególne osobniki pochodzące z określonej lokalizacji reprezentują specyficzne, stosunkowo jednorodne podgrupy (zwane liniami, ekotypami czy też odmianami) zarówno pod względem genetycznym, fizjologicznym jak i morfologicznym. Reprezentują tym samym swoisty zestaw cech wykształconych w wyniku ewolucji i adaptacji do swojego siedliska. Dodatkowo, wiele gatunków roślin, w tym także gatunki modelowe, wykazuje samopylność, a ich genomy są zazwyczaj wysoce homozygotyczne. Dlatego analiza SV u roślin zwykle polega na identyfikacji wariantów np. między liniami czy ekotypami, a następnie określeniu wpływu różnic genetycznych na ich fenotyp i przystosowanie do danych warunków otoczenia. Z kolei całogenomowe porównania są świetną okazją do prowadzenia badań pangenomicznych.

Obecnie wiadomo, że podobnie jak u ludzi i zwierząt, polimorfizm liczby kopii stanowi istotną część zmienności genetycznej również u roślin (Żmieńko i wsp. 2014, Ho i wsp. 2020, Yuan i wsp. 2021, Hollox i wsp. 2022). Początkowo większość badań dotyczyła identyfikacji sekwencji obecnych w jednym genomie, ale całkowicie nieobecnych w innych (ang. *Presence-Absence Variations*; PAV) (Żmieńko i wsp. 2014). Pierwsze analizy CNV na szeroką skalę prowadzone były na kukurydzy (Springer i wsp. 2009, Beló i wsp. 2010, Swanson-Wagner i wsp. 2010), ryżu (Yu i wsp. 2011) oraz soi (Haun i wsp. 2011, McHale i wsp. 2012). Rozwój technik sekwencjonowania znacznie rozszerzył zakres eksploracji CNV u roślin. Liczne prace wykorzystujące dane z sekwencjonowania krótkich lub/i długich odczytów czy też integracji różnych metod przez lata dostarczyły informacji o wariantach strukturalnych m.in. dla kukurydzy (Chia i wsp. 2012, Yang i wsp. 2019, Lin i wsp. 2021, Hufford i wsp. 2021*), ryżu (Xu i wsp. 2011, Wang i wsp. 2018*, Kou i wsp. 2020, Ma i wsp. 2020), soi (Lam i wsp. 2010, Liu i wsp. 2020*, Valliyodan i wsp. 2021*), pszenicy (Saintenac i wsp. 2011, Montenegro i wsp. 2017*, De Oliveira i wsp. 2020, Walkowiak i wsp. 2020*), melona (Zhao i wsp. 2019), brzoskwini (Guo i wsp. 2020), rzepaku (Chawla i wsp. 2021), jęczmienia (Jayakodi i wsp. 2020*), *Medicago truncatula* (Zhou i wsp. 2017*), słonecznika (Hübner i wsp. 2019*), pomidora (Alonge i wsp. 2020*, Gao i wsp. 2019*, Li i wsp. 2023*), jabłka (Sun i wsp. 2020*) oraz wielu innych (Żmieńko i wsp. 2014, Dolatabadian i wsp. 2017, Yuan i wsp. 2021). Warto podkreślić, że prace te mają różnorodny stopień złożoności – od detekcji kilkudziesięciu regionów w genomie aż po analizy pangenomiczne (oznaczone *). Dodatkowo, znaczna większość badań nadal skupia się na identyfikacji PAV, co mimo tak zaawansowanej technologii i możliwości integracji różnych metod badawczych pokazuje jak dużym wyzwaniem wciąż są regiony zduplikowane i rearanżacje strukturalne.

Równolegle do badań nad poznaniem zmienności strukturalnej prowadzono działania mające na celu znalezienie związków pomiędzy CNV, a fenotypem. Zwłaszcza warianty liczby kopii, które pokrywają się albo sąsiadują z genami kodującymi białka (ang. *Protein-Coding Genes*, PCG) lub ich promotorami, mogą być przyczyną zmienności fenotypowej. Już pierwsze analizy u różnych organizmów wskazywały na nierównomierne rozmieszczenie CNV w genomie. Zaobserwowano dużo wyższe zagęszczenie CNV w regionach powtórzeń, a niższe w regionach PCG (Emerson i wsp. 2008, Conrad i wsp. 2010, Chia i wsp. 2012), wskazując na potencjalnie silne oddziaływanie tych ostatnich na organizm. Stąd, liczne prace koncentrują się na poszukiwaniu zależności między konkretnymi genami bądź regionami w genomie, a występowaniem określonego fenotypu. Odkryte jak dotąd związki między CNV, a fenotypem u roślin obejmują zarówno cechy morfologiczne, jak i fizjologiczne oraz odporność na choroby (Żmieńko i wsp. 2014, Dolatabadian i wsp. 2017, Yuan i wsp. 2021). Przykładowo, Díaz i wsp. (2012) wykazali korelację między liczbą kopii genu *Vrn-A1*, kodującego czynnik transkrypcyjny *MADS-box*, a czasem kwitnienia pszenicy. Rośliny z dodatkową kopią tego genu kwitły później. W pszenicy, CNV powiązano również z karłowatym fenotypem (Li i wsp. 2012) czy większą tolerancją na mróz (Sieber i wsp. 2016). W różnych roślinach wykazano, że zmiany w liczbie kopii DNA wiążą się z tolerancją na różne związki chemiczne obecne w glebie. I tak u *Arabidopsis halleri* zwiększenie liczby kopii genu *HMA4*, kodującego białko odpowiedzialne za translokację jonów metali cynku i kadmu z korzenia do pędu, przyczynia się do hiperakumulacji i hipertolerancji tych jonów (Hanikenne i wsp. 2008, 2013). W jęczmieniu, odmiany wykazujące wyższą tolerancję na bor zawierały czterokrotnie więcej kopii genu *Bot1* (gen transportera boru) (Sutton i wsp. 2007), zaś u kukurydzy, tolerancję na glin powiązano z wyższą liczbą kopii genu *MATE1* (Maron et al. 2013). Wskazano również, że zwielokrotnienie kopii genu *MATE1* nastąpiło stosunkowo niedawno i prawdopodobnie było związane z udomowieniem rośliny. Podobną tendencję zaobserwowano u innych roślin uprawnych pokazując, że CNV przyczyniają się do szybkiej adaptacji związanej z ich udomowieniem, co może być kluczowym aspektem w kwestii ulepszania upraw (Lye i wsp. 2019). Liczne przykłady u roślin wskazują również na powiązanie CNV z odpornością na choroby i patogeny. Dowiedziono, że CNV związane są często z genami kodującymi białka klasy NB-LRR (tj. zawierającymi domenę wiążącą nukleotydy oraz powtórzenia bogate w leucynę), o których wiadomo, że biorą udział w mechanizmach związanych z obronnością roślin (Saxena i wsp. 2014, Dolatabadian i wsp. 2017).

1.4 Badania asocjacyjne całego genomu

Poznanie i zrozumienie w jaki sposób SV kształtuje fenotyp wydaje się być kluczowym wyzwaniem współczesnej genetyki. Wzrastająca kompleksowość danych pociąga za sobą rozwój nowych, globalnych metod analizy np. w oparciu o badania asocjacyjne całego genomu (ang. *Genome-Wide Association Study*, GWAS). Analiza GWAS ma charakter przesiewowy. Jej celem jest identyfikacja wariantów genomowych, które są statystycznie powiązane z wystąpieniem określonej cechy. Metoda ta, choć ma ogromny potencjał, stawia przed badaczami sporo wyzwań (Pearson i Manolio 2008). Na pomyślną detekcję asocjacji pomiędzy wariantami genetycznymi, a badaną cechą (z metodologicznego punktu widzenia) ma wpływ kilka czynników m.in. liczba próbek, liczba wariantów używanych w analizie, złożoność analizowanego fenotypu (ilość markerów wpływających na pojawienie się danej cechy), a także zróżnicowanie badanej cechy w populacji. W analizie GWAS uwzględnia się wyłącznie częste polimorfizmy (występujące np. w co najmniej 5% populacji), przez co część informacji o rzadkich wariantach jest tracona. Analiza GWAS obejmuje setki tysięcy, a czasem więcej wykonanych testów statystycznych w jednym eksperymencie, co wymusza stosowanie niskich progów istotności statystycznej (uwzględnionych np. poprzez zastosowanie korekty na wielokrotne testowanie). Dlatego, aby uzyskać wystarczającą moc analizy statystycznej, badana populacja powinna składać się z co najmniej kilku tysięcy próbek. Im więcej, tym lepiej. Przykładowo, *International Consortium for Blood Pressure Genome-Wide Association Studies* poszukiwało markerów związanych z ryzykiem chorób układu krążenia na grupie aż 200 tysięcy osób (International Consortium for Blood Pressure Genome-Wide Association Studies 2011). Kolejnym aspektem jest złożoność badanej cechy. Liczne fenotypy warunkowane są przez kilka, wspólnie występujących wariantów, które – analizowane pojedynczo - często mają stosunkowo mały efekt i przez co są bardzo trudne do identyfikacji. Im bardziej złożona cecha, bądź dyskretny efekt, tym trudniej uzyskać odpowiednią moc statystyczną i tym większa powinna być badana próba. Z kolei im silniejszy efekt, tym większa szansa na identyfikację wariantu z nim związanego. W badaniach GWAS tzw. „ukrytą zmienną", silnie wpływającą na wynik, jest struktura populacji (Barton i wsp. 2019). Każda naturalna populacja jest w jakiś sposób ustrukturyzowana i można w jej obrębie wydzielić podobne genetycznie podgrupy. Przykładowo, jeśli w jednej podgrupie współistnieją: badany marker i niezwiązana z nim cecha fenotypowa, to w efekcie analizy GWAS może dojść do identyfikacji nieprawdziwej, pozornej zależności. Dlatego uwzględnienie struktury populacji w celu zminimalizowania jej wpływu na wyniki GWAS jest niezbędne. Należy pamiętać, że wyniki GWAS reprezentują wyłącznie matematyczne i statystyczne zależności występujące w analizowanym przez nas zbiorze danych. To narzędzie, które skanuje genom w poszukiwaniu potencjalnych „kandydatów" związanych z określonym stanem lub cechą. Jednak uzyskanie istotnej statystycznie asocjacji nie oznacza, że dany wariant wywołuje badany fenotyp. Konieczne są dalsze badania funkcjonalne wyselekcjonowanych markerów, które ujawnią ich prawdziwy związek przyczynowo-skutkowy. W większości przypadków analizy GWAS opierają się na SNP. Możliwość wykorzystania CNV samodzielnie bądź wspólnie z informacją o SNP do tego typu analiz wydaje się niezwykle cenna i może rzucić nowe spojrzenie na związek CNV z fenotypem. W badaniach asocjacyjnych na kukurydzy Chia i wsp. (2012) wykazali, że CNV w dużym stopniu przyczynia się do zmienności analizowanych 5 cech związanych z rozwojem liści i odpornością na choroby oraz dostarcza nowych informacji stanowiąc uzupełnienie wyników uzyskanych ze SNP.

1.5 Arabidopsis jako roślina modelowa w badaniu zmienności wewnątrzgatunkowej

Obiektem moich badań jest roślina modelowa Arabidopsis (*Arabidopsis thaliana*) znana również jako rzodkiewnik pospolity. To pierwsza roślina, której genom zsekwencjonowano. Sekwencja DNA pochodząca z ekotypu Columbia-0 (Col-0) została opublikowana przez konsorcjum *Arabidopsis Genome Initiative* w 2000 r. (Arabidopsis Genome Initiative 2000) i - zaktualizowana - do teraz uznawana jest za sekwencję referencyjną dla tego gatunku. Arabidopsis ma stosunkowo mały, diploidalny genom (5 chromosomów, ∼135 Mpz), a jej łatwość w hodowli i krótki cykl życiowy (około 6 tygodni) sprawiają, że jest popularnym obiektem badań. Jest rośliną samopylną i szeroko rozpowszechnioną na świecie, szczególnie na półkuli północnej. Wśród naturalnych ekotypów Arabidopsis obserwuje się duże zróżnicowanie cech fizjologicznych i morfologicznych. Linie (ang. *accessions*) reprezentujące osobniki pochodzące z określonej lokalizacji geograficznej, często były utrzymywane w warunkach laboratoryjnych przez długi czas. Obecnie, oznaczone unikalnym identyfikatorem wchodzą w skład bogatego banku nasion dla tej rośliny (*The Nottingham Arabidopsis Stock Centre*, NASC, https://arabidopsis.info/). Te lokalnie przystosowane linie są wysoce wsobne i jednorodne genetycznie. Umożliwia to wielokrotne fenotypowanie tego samego genotypu w różnych, kontrolowanych warunkach. Dlatego, Arabidopsis wydaje się być idealnym obiektem do badania interakcji między genotypem, a środowiskiem oraz do prowadzenia badań wewnątrzgatunkowych.

Podobnie jak u innych roślin, początkowo u Arabidopsis większość analiz była ukierunkowana na scharakteryzowanie zjawiska PAV, podczas gdy wykrywanie duplikacji ograniczało się do pojedynczych linii lub niewielkich populacji. Jedną z pierwszych prac obejmujących detekcję CNV na szeroką skalę z danych aCGH oraz NGS była praca Santuari i wsp. (2010). Porównanie przez nich czterech linii w odniesieniu do referencji (linii Col-0) pozwoliło wykryć liczne delecje, które pokrywały się z transpozonami (ang. *Transposable Elements*, TE) i PCG. Co ciekawe, większość zidentyfikowanych wariantów występowała w dwóch lub więcej liniach, wskazując istotny stopień zmienności wewnątrzgatunkowej genomu Arabidopsis.

W 2008 r., powstało konsorcjum w ramach projektu 1001 Genomów (*The 1001 Genomes Project*, https://1001genomes.org/), którego celem było scharakteryzowanie całogenomowej zmienności dla co najmniej 1001 linii Arabidopsis. Trzy lata później, w 2011 r., w ramach pilotażowej fazy projektu zsekwencjonowano i opublikowano dane NGS dla 80 linii, łącznie z analizą SNP i krótkich indeli (Cao i wsp. 2011). Zidentyfikowano również 1059 dłuższych CNV (w odniesieniu do Col-0) o wielkości od 1 do 13 kpz. Ponad 85% wykrytych wariantów występowała w więcej niż jednej linii, a sumarycznie wszystkie warianty obejmowały ∼2% genomu. W efekcie zakończenia drugiej fazy tego projektu, w czerwcu 2016 r. udostępniono całogenomowe dane dla 1135 linii Arabidopsis (1001 Genomes Consortium 2016). Stworzono katalog SNP i krótkich indeli (<50 pz), jednak pominięto identyfikację dłuższych wariantów strukturalnych, w tym również CNV. W oparciu o uzyskane dane m.in. wydzielono grupy genetyczne, scharakteryzowano strukturę populacji i zaproponowano wzorce migracji, co stanowi doskonałe podwaliny do badań nad fenotypem, ewolucją i adaptacją. Równolegle do danych genomowych przedstawiono wyniki analiz metylomów i transkryptomów również dla ponad 1000 linii Arabidopsis (Kawakatsu i wsp. 2016). Publiczny dostęp do trzech tak dużych i wzajemnie uzupełniających się zestawów danych NGS otworzył zupełnie nowe możliwości badawcze. Dzięki temu, przez kilka ostatnich lat powstało wiele zaawansowanych baz danych dla Arabidopsis integrujących różnorodne dane biologiczne, geograficzne, klimatyczne, czy fenotypowe, w tym m.in. niezwykle cenna baza danych fenotypów AraPheno (https://arapheno.1001genomes.org/). Połączenie różnorodnych wyników daje możliwość wzbogacania własnych badań o dodatkowe aspekty bez konieczności przeprowadzania analiz eksperymentalnych od podstaw, co przekłada się na większą kompleksowość analiz, jednakże stawia duże wymagania dla bioinformatyków, gdyż tylko poprzez zrozumienie danych i prawidłowo dobrane, i opracowane ścieżki analizy z surowych danych można uzyskać prawidłowy wynik.

Pomimo zsekwencjonowania setek genomów Arabidopsis wciąż istniała luka dotycząca zmienności wewnątrzgatunkowej. Analizy pangenomiczne wymagają kompletnej sekwencji genomowej dla poszczególnych linii, co z krótkich odczytów pochodzących z sekwencjonowania było bardzo trudne do osiągnięcia. Jiao i Schneeberger (2020) złożyli sekwencje genomowe 7 linii Arabidopsis, z danych uzyskanych przy pomocy długich odczytów. Wzajemna analiza porównawcza ujawniła w każdym z genomów rearanżacje o łącznej długości około 13–17 Mpz i pozwoliła na zidentyfikowanie 5–6 Mpz sekwencji nieobecnej w genomie referencyjnym. Przełożyło się to na różnice liczby kopii w około 5000 genów. Dodatkowo, zidentyfikowano tzw. gorące miejsca rearanżacji czyli regiony bardzo zmienne w populacji Arabidopsis, co przekładało się obserwowany zwiększony odsetek CNV w tych regionach genomu. Obecnie trwają prace nad projektem 1001G+, którego celem jest złożenie jak największej liczby kompletnych sekwencji genomowych naturalnych linii Arabidopsis w oparciu o długie odczyty sekwencjonowania. Takie dane pokażą co kryje się w części genomu, której obecnie nie możemy zobaczyć oraz pozwolą na przeprowadzenie kompleksowej analizy pangenomicznej ujawniając rzeczywisty zakres zmienności i rożnorodności dla Arabidopsis (Wlodzimierz i wsp. 2023, Lian i wsp. 2023).

1.6 Metabolity wtórne

Rośliny mają niesamowitą zdolność do syntezy zróżnicowanych niskocząsteczkowych związków organicznych. W efekcie mogą łatwo dostosowywać się do życia w zmiennych i wymagających warunkach środowiskowych. Ze względu na pełnione funkcje, związki te dzieli się na trzy grupy: metabolity pierwotne – związki chemiczne niezbędne do prawidłowego przebiegu procesów życiowych, które spełniają w organizmie żywym najbardziej podstawowe funkcje (budulcowe, energetyczne, zapasowe); metabolity wtórne - związki istotne, choć nie niezbędne dla przeżycia komórki, które powstają w przebiegu wyspecjalizowanej przemiany materii; oraz hormony, które regulują procesy organizmu i metabolizm (Jasiński i wsp. 2009, Erb i Kliebenstein 2020). Przez dziesięciolecia ta funkcjonalna trychotomia metabolizmu roślin kształtowała teorię i eksperymenty w biologii roślin, mimo że granice między tymi klasami metabolitów były w zasadzie umowne.

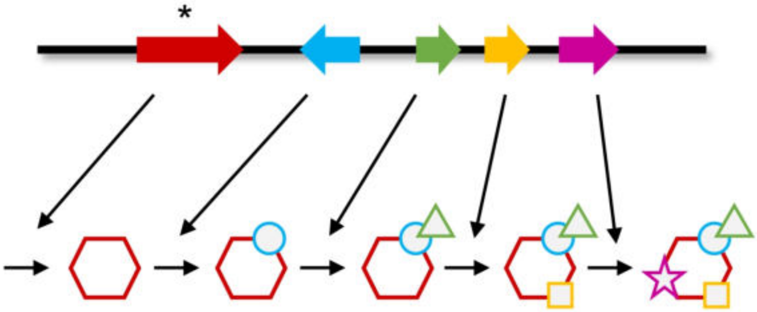
Chociaż metabolity wtórne nie są niezbędne do wzrostu i rozwoju samej rośliny odgrywają dużą rolę w komunikacji między rośliną, a środowiskiem. Do najważniejszych funkcji tych związków należy ochrona rośliny przed roślinożercami czy pasożytami, ochrona przed mikroorganizmami patogennymi, wabienie owadów, regulacja wzrostu i rozwoju rośliny, a także oddziaływań symbiotycznych roślin z innymi organizmami (Jasiński i wsp. 2009, Elshafie i wsp. 2023). Dodatkowo, najnowsze badania podkreślają znacznie obszerniejszą niż wcześniej zakładano rolę tych związków. Wskazują, że metabolity wtórne mogą być wielofunkcyjne i działać zarówno jako regulatory wzrostu roślin, jak i mogą być zaangażowane w metabolizm pierwotny czy ochronę roślin (Isah 2019, Erb i Kliebenstein 2020). Metabolity wtórne stanowią niezwykle zróżnicowaną grupę związków o szerokim zakresie aktywności biologicznej. Poszczególne związki są zwykle swoiste dla danej rośliny i powstają tylko w niektórych komórkach, tkankach lub organach. Związki te magazynowane są głównie w wakuolach, w przestrzeniach międzykomórkowych oraz na powierzchni tkanek (Jasiński i wsp. 2009). Warto podkreślić, że stężenie danego metabolitu w roślinie może być zmienne w zależności od warunków środowiska.

Występujące w świecie roślinnym metabolity wtórne dzielimy na trzy podstawowe grupy: terpenoidy, niebiałkowe związki azotowe (głównie alkaloidy, jak również betalainy, glikozydy cyjanogenne oraz glukozynolany) oraz związki fenolowe (Jasiński i wsp. 2009, Twaij i Hasan 2022, Elshafie i wsp. 2023). Poza oczywistymi korzyściami dla samej rośliny metabolity wtórne są wykorzystywane przez człowieka jako dodatki spożywcze, w medycynie naturalnej, ziołolecznictwie, do produkcji leków i farmaceutyków, kosmetyków oraz w innych gałęziach przemysłu. Spośród powszechnie stosowanych związków można wymienić m.in.: limonen, mircen, linalol, α/β-pinen, kamfen, bajkalinę, apigeninę i wiele innych (Twaij i Hasan 2022, Elshafie i wsp. 2023). Szerokie spektrum wykorzystania tych związków oraz praktycznie nieograniczona ich liczba (biorąc pod uwagę ogromne zróżnicowanie świata roślinnego) pokazuje jak duże znaczenie ma identyfikacja nowych związków i zrozumienie szlaków ich powstawania.

1.7 Roślinne metaboliczne klastry genów

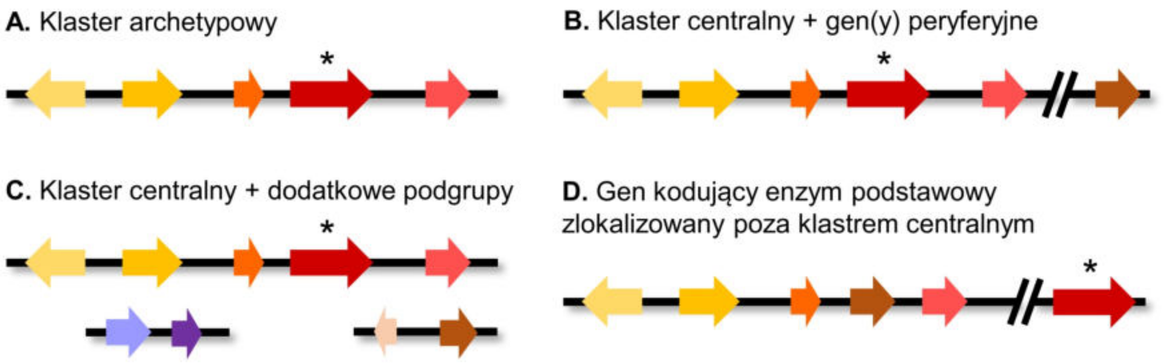
W genomach bakterii powszechne są tzw. operony czyli jednostki zgrupowanych genów (tworzących swoisty klaster), transkrybowanych wspólnie - jako pojedyncza cząsteczka mRNA i kodujących enzymy zaangażowane we wspólny szlak metaboliczny. U eukariontów z kolei geny zaangażowane we wspólny szlak metaboliczny są zazwyczaj rozproszone w genomie. Jednakże, badania prowadzone w ostatnich latach dowiodły istnienia podobnych do operonów jednostek organizacyjnych - funkcjonalnych klastrów złożonych z niehomologicznych genów - w genomach u roślin i grzybów. Te stosunkowo niedawno odkryte jednostki organizacyjne u eukariontów są nazywane metabolicznymi klastrami genów lub klastrami genów biosyntezy (ang. *Metabolic Gene Clusters*, MGC/*Biosynthetic Gene Clusters*) (Nützmann i wsp. 2018). Pomimo częściowych podobieństw do operonów bakteryjnych (fizyczne grupowanie, współregulacja), geny tworzące klaster są transkrybowane oddzielnie (jako odrębne cząsteczki mRNA) (Boycheva i wsp. 2014, Nützmann i wsp. 2018). W miarę nowych, znaczących odkryć definicja MGC jest na bieżąco doprecyzowana. Obecnie, MGC są określane jako grupa trzech lub więcej genów, które kodują co najmniej trzy różne typy enzymów zaangażowanych we wspólny szlak metaboliczny (zwykle szlak biosyntezy). Budowa MGC jest więc zróżnicowana i zależy od wielu czynników. W obrębie MGC wyróżnia się dwa typy enzymów (Nützmann i wsp. 2018). Enzym podstawowy (ang. *signature enzyme*) jest odpowiedzialny za tworzenie chemicznego szkieletu metabolitu. Definiuje on finalną klasę produktów szlaku (np. terpenoidy, alkaloidy) i jest zaangażowany w główny (zazwyczaj pierwszy) etap szlaku metabolicznego. Drugim typem enzymów są różnorodne enzymy kodowane przez pozostałe geny klastra, modyfikujące szkielet podstawowy metabolitu (ang. *tailoring enzymes*) (Rysunek 1.4). Najczęściej są to oksydazy cytochromu P450 (CYP450), transferazy (m.in. acylo-, metylo-, glukozylotransferazy), dehydrogenazy, transaminazy i wiele innych. Warto zaznaczyć, że w przypadku jednego z najczęściej występujących enzymów w obrębie znanych MGC, CYP450, każda rodzina cytochromów jest traktowana jako odrębny typ enzymu (Nützmann i wsp. 2016).

Pod względem architektury wyróżnia się kilka typów MGC (Nützmann i wsp. 2016). U roślin, klaster centralny zawiera większość genów szlaku (zwykle łącznie z genem kodującym enzym podstawowy). Geny te są zlokalizowane w sąsiednich pozycjach genomu tworząc zwartą grupę (Rysunek 1.5 A). Zdarza się jednak, że mogą być one przeplatane ograniczoną liczbą niepowiązanych funkcjonalnie genów. Zazwyczaj, poza klastrem centralnym istnieje co najmniej jeden gen peryferyjny kodujący inne etapy szlaku metabolicznego. Taki gen może być zlokalizowany w pobliżu klastra centralnego



Rysunek 1.4: Dwa typy enzymów w obrębie MGC: podstawowy (*) oraz inne enzymy modyfikujące szkielet podstawowy metabolitu (na podstawie: Nützmann i Osbourn 2014)

bądź w odległym miejscu w genomie, a nawet na innym chromosomie (Nützmann i wsp. 2016, Kautsar i wsp. 2017) (Rysunek 1.5 B). Klaster DIMBOA u kukurydzy związany jest z syntezą benzoksazynoidów (Bx), czyli metabolitów wtórnych, które działają jako pestycydy, insektycydy i wykazują działanie allelopatyczne (Sicker i wsp. 2000). Związki te są produkowane przez wiele gatunków roślin, w tym głównie przez kukurydzę, pszenicę i żyto (Sue i wsp. 2011). U kukurydzy, synteza cyklicznego kwasu hydroksamowego DIBOA wymaga współudziału pięciu genów, które znajdują się w klastrze centralnym (Frey i wsp. 1997, 2003, von Rad i wsp. 2001, Jonczyk i wsp. 2008). Peryferyjny gen *Bx7* kodujący enzym O-metylotransferazę, niezbędną do biosyntezy DIMBOA (czyli 2,4-dihydroksy-7-metoksy-1,4-benzoksazyn-3-onu), znajduje się w odległości 15 Mpz od klastra centralnego (Jonczyk i wsp. 2008). Z kolei gen *Bx9*, który koduje transferazę cukrową aktywną wobec DIBOA/DIMBOA, znajduje się na innym chromosomie (von Rad i wsp. 2001). Badania prowadzone dla pszenicy i żyta ujawniły, że dla tych roślin geny biosyntezy DIBOA znajdują się w dwóch podgrupach zlokalizowanych na różnych chromosomach (Sue i wsp. 2011). Niektóre szlaki metaboliczne są bardziej pofragmentowane. Poza klastrem centralnym mogą istnieć dodatkowe skupiska złożone z dwóch lub trzech genów (ang. *satellite subgroups*) w innym miejscu w genomie (Rysunek 1.5 C). W ostatnim typie organizacji MGC gen kodujący enzym podstawowy jest peryferyjny. Wówczas klaster centralny składa się z co najmniej trzech genów kodujących enzymy zaangażowane w dodatkowe etapy szlaku, które wykazują silną koekspresję z genem kodującym enzym podstawowy (Rysunek 1.5 D). Te dwa ostatnie typy organizacji MGC można zaobserwować w klastrze α-tomatyny u pomidora. Klaster centralny biosyntezy α-tomatyny zlokalizowany jest na chromosomie 7 i składa się z 6 genów. Gen katalizujący pierwszy etap tego szlaku (konwersję cholesterolu do 22-hydroksycholesterolu) (Itkin i wsp. 2013), również znajduje się na chromosomie 7, aczkolwiek jest oddalony o 7,9 Mpz od najbliższego genu klastra centralnego (Tomato Genome Consortium 2012). Dwa dodatkowe, sąsiadujące ze sobą geny kodujące enzymy tego szlaku znajdują się na chromosomie 12. Udział enzymów kodowanych przez geny peryferyjne oraz sieć połączeń między różnymi szlakami biosyntezy metabolitów może skutkować dodatkowym zróżnicowaniem produktów biosyntezy (Huang i wsp. 2019).


A. Klaster archetypowy
*
B. Klaster centralny + gen(y) peryferyjne
*
C. Klaster centralny + dodatkowe podgrupy
*
D. Gen kodujący enzym podstawowy zlokalizowany poza klastrem centralnym
*

Rysunek 1.5: Typy organizacji MGC (na podstawie: Nützmann i wsp. 2016)

MGC u grzybów występują dość powszechnie i u tych organizmów zidentyfikowano klastry zaangażowane zarówno w metabolizm pierwotny (podstawowy) jak i wtórny. Z kolei u roślin obecnie znanych jest ponad 30 MGC i wszystkie związane są z wytwarzaniem metabolitów wtórnych. Ich rozmiary wahają się od 35 kpz do kilkuset kpz i obejmują od 3 do 10 genów (Nützmann i wsp. 2018). Wśród stosunkowo niewielu zidentyfikowanych MGC u roślin zadziwiająca jest tak duża różnorodność i złożoność klastrów, zarówno pod względem budowy jak i produkowanych związków. Badania nad MGC u roślin rozwijają się bardzo dynamicznie, zaś nowe odkrycia na bieżąco poszerzają naszą wiedzę w tej złożonej tematyce. Warto przywołać tu najlepiej jak dotąd poznany klaster biosyntezy thalianolu u Arabidopsis. Przez wiele lat uznawano, że ten MGC zbudowany jest z czterech sąsiadujących ze sobą genów (Field i Osbourn 2008). Wyniki badań w ostatnich latach wskazują nie tylko na istnienie genów peryferyjnych w tym klastrze (Huang i wsp. 2019, Liu i wsp. 2020), ale również dowodzą istnienia inwersji w obrębie klastra centralnego w niektórych liniach Arabidopsis, w efekcie tworząc rozszerzony, ciągły klaster centralny (Liu i wsp. 2020). Z kolei Mugford i wsp. (2009) zidentyfikowali w klastrze awenacyny u owsa dodatkowe zgrupowanie genów wewnątrz klastra centralnego. Ten swoisty moduł złożony z trzech sąsiadujących ze sobą genów tworzył tzw. „moduł acylacyjny” niezbędny do acylowania triterpenów (Mugford i wsp. 2013). Innym, ciekawym zjawiskiem jest liniowe ułożenie genów w obrębie klastra noskapiny w maku (Winzer i wsp. 2012). Zaobserwowano, że geny te zorganizowane są w podgrupy odpowiadające wczesnym, środkowym i późnym etapom szlaku. Chociaż obserwacje te sugerują, że kolejność genów może być istotna dla funkcji szlaku, to współliniowość wydaje się być raczej wyjątkiem niż regułą u roślin (Nützmann i wsp. 2016).

1.8 Regulacja metabolicznych klastrów genów

Większość związków metabolicznych powstaje w ścieżkach, w których geny nie są sklastrowane tylko rozproszone w genomie. Dlatego odkrycie stosunkowo rzadkiego zjawiska grupowania genów wśród eukariontów nasuwa wiele intrygujących pytań m.in.: dlaczego niektóre szlaki metaboliczne są związane z organizacją w klastry, podczas gdy inne nie? Jakie korzyści przynosi takie grupowanie? Jakie są mechanizmy odpowiedzialne za powstawanie tych jednostek organizacyjnych? Na obecnym etapie wiedzy naukowej brakuje odpowiedzi na te kluczowe pytania.

Geny w obrębie MGC często ulegają ekspresji w tym samym czasie i w podobnych warunkach. Wzajemna bliskość genów z pewnością sprzyja koordynacji ekspresji i współregulacji, co przekłada się na efektywną produkcję enzymów lub innych białek niezbędnych do danego procesu metabolicznego. Geny w klastrze mogą dzielić wspólne elementy regulatorowe, które wpływają na ich ekspresję. Wcześniejsze badania prowadzone u grzybów wykazały, że MGC podlegają epigenetycznym mechanizmom regulacji ekspresji genów (Bok i wsp. 2009, Brakhage 2013). Również u roślin coraz więcej prac opisuje znaczenie modyfikacji epigenetycznych w regulacji MGC. Wegel i wsp. (2009) pokazali, że ekspresja klastra awenacyny u owsa związana jest z komórkowo specyficzną dekondensacją chromatyny. Zgodnie z aktualną wiedzą, wszystkie MGC u Arabidopsis powiązano z aktywnością w korzeniach (Huang i wsp. 2019). Yu i wsp. (2016) analizując regiony MGC u tej modelowej rośliny zaobserwowali istotnie wyższy poziom trimetylacji reszty lizyny 27 histonu H3 (H3K27me3), związanej z wyciszeniem genów, w siewkach niż w korzeniach. Wykazano również, że poziom tej modyfikacji był odwrotnie skorelowany z ekspresją genów klastra. Analiza MGC u owsa i kukurydzy wykazała podobne zależności. Dodatkowo, w korzeniach Arabidopsis wykazano wzbogacenie w regionach klastrów biosyntezy marneralu i thalianolu w wariant histonu H2A.Z, biorący udział w aktywacji ekspresji genów. Na podstawie uzyskanych wyników autorzy zaproponowali, że H3K27me3 jest zaangażowany w wyciszanie MGC, a H2A.Z w ich aktywację (Nützmann i Osbourn, 2015, Yu i wsp. 2016).

Fizyczna bliskość genów ułatwia zarówno ich wspólne dziedziczenie, ewolucję jak i utrzymanie określonych szlaków metabolicznych w populacji. Genetyczne sprzężenie genów kodujących złożone cechy, które zapewniają selektywną przewagę, zmniejsza ryzyko zakłócenia tych korzystnych zestawów genów przez rekombinację. Warto podkreślić, że liczne badania zwracają uwagę na istotność MGC jako całości, zaś wszelkie możliwe zaburzenia w kodowanych szlakach metabolicznych mogą skutkować akumulacją toksycznych związków pośrednich i negatywnie wpływać na wzrost, rozwój, a nawet przeżycie rośliny (von Rad i wsp. 2001, Field i Osbourn 2008, Field i wsp. 2011, Itkin i wsp. 2011). Stąd grupowanie może również chronić przed produkcją i kumulacją toksycznych/bioaktywnych produktów pośrednich.

Wyniki badań wskazują, że wiele z dotychczas opisanych MGC powstało stosunkowo niedawno i jest ograniczonych do wąskich podgrup taksonomicznych (Qi i wsp. 2004, Field i wsp. 2011, Matsuba i wsp. 2013). Tak duże zróżnicowanie metaboliczne wśród roślin może odzwierciedlać adaptację do warunków w określonych niszach ekologicznych. Zdolność roślin do tworzenia nowych szlaków metabolicznych jest niezwykła i fascynująca, i pokazuje jak wysoce plastyczne i wszechstronne są genomy tych organizmów. W świetle obecnych badań MGC jawią się jako niezwykle dynamiczne obszary genomów. Można zatem uznać, że ich aktualny „kształt" reprezentuje tylko moment w czasie ich ewolucji. Daje to możliwość wychwycenia takich MGC, które dopiero się formują, takich, które ewoluują i tych, które już się wykształciły, czyli pozwala rozpatrywać klastry w kontekście ich „narodzin, życia i śmierci" (Lind i wsp. 2017). Po znalezieniu korzystnej kombinacji alleli różnych genów szlaku prawdopodobnie korzystne będzie wspólne dziedziczenie tych genów. W rezultacie grupowanie genów może być cechą skrajnej selekcji, dążącej do współdziedziczenia i optymalizacji najlepszych kombinacji genów dla nowych szlaków metabolicznych, które zapewniają selektywne korzyści (Yeaman i Whitlock 2011, Takos i Rook 2012).

Badania locus *Rhg1* w genomie soi, związanego z rozwojem oporności rośliny na nicień węgorka cysty sojowej (*Heterodera glycines*, SCN) dostarczyły fascynujący przykład takiej selekcji pozytywnej powiązanej funkcjonalnie grupy genów (Lee i wsp. 2015, Cook i wsp. 2012, Cook i wsp. 2014). Wewnątrz regionu ∼31 kpz zidentyfikowano trzy niehomologiczne geny, które, jak wykazano, wspólnie przyczyniają się do pełnej odporności soi na patogen (Cook i wsp. 2012). Poszczególne haplotypy locus *Rhg1* związane były z różną liczbą kopii całego segmentu (a nie pojedynczych genów). Odmiany podatne na SCN zawierały tylko jedną kopię 31 kpz segmentu na genom haploidalny, podczas gdy w odmianach odpornych obserwowano nawet do 10 tandemowych kopii (rozszerzając tym samym locus odporności do ∼300 kpz). Chociaż locus *Rhg1* nie jest zaliczane do klasycznych MGC jest to pierwsze odkrycie, w którym pokazano, że za cechę odporności na patogen u roślin może odpowiadać zwielokrotnienie liczby kopii wielogenowego klastra. Wydaje się wysoce prawdopodobne, że wiele innych, złożonych cech w genomach roślinnych jest kontrolowanych właśnie przez tego typu zmienność strukturalną.

Według jednej z hipotez roślinne MGC powstają w wyniku duplikacji i późniejszej neo- lub subfunkcjonalizacji genów zaangażowanych w metabolizm pierwotny, po czym może nastąpić rekrutacja dodatkowych genów do nowo tworzącego się szlaku biosyntezy (Nützmann i Osbourn 2014). Ponadto, MGC są często zlokalizowane w dynamicznych regionach genomu np. w okolicach centromerów czy regionach bogatych w TE, gdzie możliwość połączenia korzystnych zestawów genów poprzez rearanżacje strukturalne jest większa niż w pozostałej części genomu, promując w ten sposób tworzenie MGC (Field i wsp. 2011). Te same czynniki mogą również przyczynić się do dalszych modyfikacji genetycznych i zmiany profilu metabolicznego roślin, czyniąc z takich MGC „gorące miejsca ewolucji”. Wszystkie te cechy, łącznie z liniowym ułożeniem genów klastra w genomie powodują, że udział CNV w kształtowaniu MGC wydaje się być duży i znaczący. Dlatego zbadanie powiązań pomiędzy występowaniem duplikacji/delecji, a strukturą i funkcją MGC pomoże lepiej zrozumieć genetyczne podstawy różnorodności fenotypowej roślin, a w przyszłości może znaleźć praktyczne zastosowanie w hodowli roślin uprawnych i biotechnologii.

Rozdział 2

Cel Pracy

Celem mojej pracy doktorskiej jest ocena wewnątrzgatunkowej zmienności strukturalnej metabolicznych klastrów genów u Arabidopsis oraz zweryfikowanie hipotezy, że polimorfizm liczby kopii ma istotny wpływ na integralność tych MGC, a tym samym może odgrywać rolę w adaptacji roślin do warunków środowiskowych.

Osiągnięcie powyższego celu wymagało realizacji następujących zadań:

1. Opracowanie podejścia bioinformatycznego do kompleksowej identyfikacji regionów wykazujących polimorfizm liczby kopii DNA na podstawie danych z wysokoprzepustowego sekwencjonowania oraz zastosowanie go do detekcji wariantów CNV w genomie Arabidopsis.

2. Szczegółowa charakterystyka zmienności strukturalnej regionów obejmujących znane metaboliczne klastry genów.

3. Poszukiwanie powiązań pomiędzy występowaniem wariantów liczby kopii genów, a zmiennością fenotypową i strukturą populacji, ze szczególnym uwzględnieniem wysoce polimorficznych metabolicznych klastrów genów.

Rozdział 3

Omówienie wyników badań

3.1 Identyfikacja regionów CNV w genomie Arabidopsis

Rozwój sekwencjonowania NGS zapoczątkował nowy etap w badaniach nad zmiennością genetyczną. Jednak w ciągu pierwszej dekady rozwoju tej technologii zdecydowanie większą popularnością cieszyły się analizy SNP i indeli w porównaniu z detekcją SV. Dlatego jeszcze kilka lat temu istniała spora luka w charakterystyce dużych zmian strukturalnych, nawet w przypadku tak dobrze zbadanej modelowej rośliny jak Arabidopsis. Wykorzystanie bogatego zestawu danych z projektu 1001 Genomów obejmujących ponad tysiąc naturalnych linii Arabidopsis, wydawało się zatem doskonałym podejściem do stworzenia katalogu regionów wykazujących zmienność liczby kopii, a tym samym określenia zakresu wewnątrzgatunkowej zmienności dla tej rośliny, otwierając drogę do lepszego zrozumienia tego zjawiska. Badania poświęcone temu zagadnieniu zostały opisane w publikacji pt. „AthCNV: A Map of DNA Copy Number Variations in the Arabidopsis Genome" (Zmienko i wsp. 2020), wchodzącej w skład mojej rozprawy doktorskiej.

Jak wspomniałam we wstępie, detekcja CNV przy pomocy danych NGS jest problemem złożonym i nie ma ustalonej, ani też powszechnie stosowanej ścieżki analizy. W związku z tym, prace rozpoczęliśmy od szerokiego przeglądu literatury dotyczącej metod używanych do detekcji CNV oraz testowania wybranych programów. Warto podkreślić, że znaczna większość programów do detekcji CNV powstała dla genomu człowieka, zaś ich zastosowanie do innych genomów wymagało wcześniejszej optymalizacji i odpowiedniego doboru parametrów, a czasem okazywało się wręcz niemożliwe. Spośród siedmiu ostatecznie wyselekcjonowanych narzędzi trzy oparte były o metodę RD (CNVnator, Control-FREEC, Genome STRiP-CNV), dwa o metodę RP (VariationHunter, BreakDancer), jeden o metodę SR (Pindel), a jeden wykorzystywał podejście hybrydowe, łączące metody RP i RD (Genome STRiP-SV). We wszystkich wybranych programach warianty identyfikowaliśmy indywidualnie dla każdej próbki, w odniesieniu do genomu referencyjnego (linii Col-0) lub całej badanej populacji (w przypadku Genome STRiP-SV oraz Genome STRiP-CNV). Wstępny wybór parametrów dla każdego programu prowadziliśmy na zestawie danych dla 80 genomów Arabidopsis (Cao i wsp. 2011). Po udostępnieniu danych dla 1 135 linii (1001 Genomes Consortium 2016) zarówno końcowa optymalizacja parametrów jak i dalsze analizy prowadziliśmy już na tym nowym, rozszerzonym zestawie danych. W zależności od próbki, dane NGS różniły się między sobą nie tylko średnim pokryciem genomu (5× – 118×), ale również typem odczytów (pojedyncze lub sparowane) i jakością. Zestaw dobrej jakości próbek (1 064 linie) wykorzystaliśmy w dalszych etapach analizy. Warto wspomnieć, że większość usuniętych na tym etapie próbek pochodziła właśnie z wcześniejszego zestawu 80 genomów, co pokazuje jak ulepszenia technologiczne i metodologiczne w przeciągu zaledwie kilku lat przełożyły się na poprawę jakości danych.

Opracowanie ścieżki identyfikacji CNV było zadaniem wieloetapowym. Pierwszy etap obejmował identyfikację wariantów przy użyciu poszczególnych programów. Dla czterech z siedmiu użytych przez nas programów (CNVnator, Control-FREEC, BreakDancer i Pindel), reprezentujących trzy różne metody identyfikacji CNV (RD, RP i SR), przeprowadziłam optymalizację parametrów i wykonałam identyfikację wariantów w genomie Arabidopsis. W następnym kroku, dla wyników uzyskanych z wszystkich siedmiu programów, opracowałam indywidualne kryteria filtracji w celu usunięcia wyników fałszywie pozytywnych (Zmienko i wsp. 2020: sekcja METHODS). Obserwowałam znaczne różnice w wielkości, typie i liczbie wykrytych wariantów nie tylko pomiędzy metodami, ale także między programami opartymi o tę samą metodę. Często dla pojedynczej próbki program identyfikował kilka - kilkanaście wzajemnie nakładających się SV o różnej długości w tym samym regionie. Podejścia RD oraz hybrydowe znajdowały najdłuższe warianty, z najmniejszą frakcją wzajemnie nakładających się regionów. W programach opartych o RP obserwowałam znaczny wzrost udziału dużych i małych indeli, zaś Pindel, oparty o metodę SR, znajdował głównie małe indele oraz, co zaskakujące, większe duplikacje. W tym ostatnim odsetek wzajemnie nakładających się wariantów był największy. Podsumowując, RD wykryło najmniej, ale najdłuższych wariantów, RP zidentyfikowało więcej, aczkolwiek krótszych zmian niż RD, zaś SR bardzo dużo krótkich wariantów (Zmienko i wsp. 2020: Figure 1B-C, Supplemental Figure 1). Delecje stanowiły przeważającą większość zmian we wszystkich programach. Z uwagi na różnice w czułości i specyficzności detekcji poszczególnych podejść, podzieliłam zestaw danych na dwie kategorie: duże indele, czyli warianty o długości od 50 do 499 pz, oraz CNV obejmujące warianty ≥500 pz.

Aby stworzyć katalog zmienności dla Arabidopsis opracowałam autorskie podejście do integracji wyników uzyskanych z różnych programów (Zmienko i wsp. 2020: sekcja METHODS, Figure 1A). Etap integracji danych w obu przypadkach (dla dużych indeli i CNV) obejmował łączenie wariantów w obrębie wyników z tego samego programu, jak i pomiędzy metodami. Początkowo, największe nadzieje pokładaliśmy w podejściu hybrydowym. Jednak już pierwsze wyniki uzyskane z poszczególnych programów ujawniły ogromne różnice w liczbie zidentyfikowanych wariantów. Podczas gdy programy oparte na pojedynczej metodzie znajdowały setki lub tysiące wariantów dla pojedynczej linii, Genome STRiP-SV znajdował ich zaledwie kilkadziesiąt. Naszym najważniejszym celem była identyfikacja duplikacji, których detekcja możliwa jest prawie wyłącznie przy użyciu metody RD. Dodatkowo, podejście to generowało najdłuższe warianty z najniższą frakcją wzajemnie nakładających się zmian, stąd postanowiłam, że właśnie ta metoda oraz metoda hybrydowa stworzą bazę do zbioru CNV. W celu wyselekcjonowania regionów zmiennych i jednocześnie usunięcia wariantów nakładających się, wykonałam dwustopniowy etap łączenia wariantów uzyskanych przy pomocy programów CNVnator, Control FREEC, Genome STRiP-CNV oraz Genome STRiP-SV: najpierw w obrębie programu, zakładając minimalne 50% wzajemne pokrycie (ang. *Reciprocal Overlap*, RO), a następnie pomiędzy programami z RO wynoszącym 80%. Dodatkowo, uwzględniałam wyłącznie regiony obecne w co najmniej 2 liniach. Tak stworzona baza wskazywała regiony CNV, jednak ze względu na niską rozdzielczość metod RD oraz łączenie poszczególnych wariantów, ich granice obarczone były dużym błędem. Z tego względu, kolejny etap obejmował udokładnienie granic regionów. W tym celu, z pozostałych programów wyselekcjonowałam wszystkie warianty, które pokrywały się z uzyskanymi regionami z RO $\geq$80%. Następnie, ustaliłam hierarchię przypisania granic wariantu na podstawie tych nałożeń, biorąc pod uwagę precyzję i rozdzielczość każdego podejścia, tj. największy priorytet przypisałam granicom wyznaczonym przez metody hybrydową i SR, a najmniejszy – granicom określonym przez metody RD. Aby dodatkowo zwiększyć wiarygodność wyników, z uzyskanych 34 368 CNV wybrałam tylko te warianty, które były wykryte przez co najmniej 2 programy. W ten sposób otrzymałam końcowy zestaw 19 003 CNV, które stworzyły katalog nazwany przez nas AthCNV. W przypadku ścieżki integracji dużych indeli zastosowałam analogiczne podejście oparte o łączenie zarówno w obrębie programu jak i między programami. Dla tego zestawu wariantów dysponowałam głównie regionami z metod RP i SR, przy czym RP wybrałam jako bazę podstawową. Ostatecznie otrzymałam 70 137 dużych indeli.

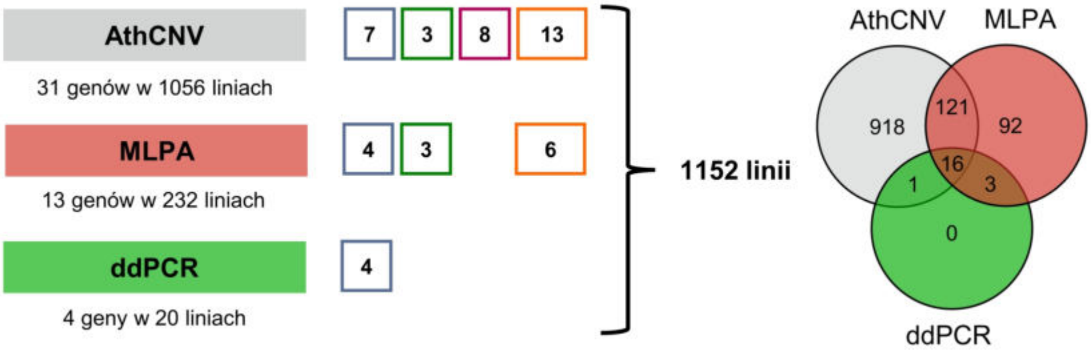
Stworzenie katalogu regionów zmiennych pozwoliło na dalszą analizę uzyskanych danych w kontekście ich rozmieszczenia w genomie, pokrycia z SNP, związku z PCG czy TE. W kolejnych etapach pracy skupiliśmy się wyłącznie na analizie regionów AthCNV. Ponad 90% zidentyfikowanych wariantów było krótszych niż 20 kpz. Pod względem rozmieszczenia w genomie obserwowaliśmy duże zagęszczenie CNV w regionach centromerów i telomerów, jak również zaobserwowaliśmy pozytywną korelację między CNV, a TE oraz negatywną korelację między CNV, a PCG (Zmienko i wsp. 2020: Table 1, Figure 2). Pary PCG-TE, zlokalizowane względem siebie nie dalej niż 2kpz, zazwyczaj miały ten sam status (CNV lub niezmienny). Szczegółowa analiza wzajemnego rozmieszczenia zmiennych i niezmiennych par ujawniła, że PCG-CNV leżą bliżej TE niż PCG niezmienne, podczas gdy TE-CNV znajdowały się dalej od PCG niż TE niezmienne. Może to wskazywać na przeciwnie działający czynnik selekcyjny, a jednocześnie skutkować nierównomiernym rozkładem CNV w genomie (Zmienko i wsp. 2020: Figure 5). W obrębie regionów CNV zidentyfikowałam 7 712 PCG, z czego 5 032 całkowicie zawierała się w regionie zmiennym. Analiza ontologii tych genów wskazała wzbogacenie w terminy związane z interakcją roślin z innymi organizmami oraz obroną i odpowiedzią na stres (Zmienko i wsp. 2020: Figure 4). Identyfikacja tak dużej liczby PCG w regionach CNV skłoniła nas do oszacowania liczby kopii dla wszystkich analizowanych linii. Uzyskane wyniki genotypowania były zbieżne z określonymi przeze mnie regionami AthCNV i zostały dodatkowo przedstawione w formie przeglądarki internetowej (http://athcnv.ibch.poznan.pl/).

Uzyskane wyniki poddaliśmy obszernej walidacji. Lokalizację regionów AthCNV porównaliśmy z wynikami z licznych publikacji (Zmienko i wsp. 2020: Supplemental Data Set 4, Figure 3, Supplemental Figure 2) uzyskując bardzo dużą zgodność wyników. Niezwykle istotnym etapem było również eksperymentalne potwierdzenie opracowanego podejścia. W tym celu równolegle zajmowaliśmy się optymalizacją metody multipleksowej zależnej od ligacji amplifikacji (ang. *Multiplex Ligation-dependent Probe Amplification*, MLPA). W efekcie przedstawiliśmy protokół, który ma charakter uniwersalny i można go zastosować dla różnych roślin, opisany w pracy „MLPA-based Analysis of Copy Number Variation in Plant Populations" (Samelak-Czajka i wsp. 2017), która wchodzi w skład mojej rozprawy doktorskiej. Moje zadanie w tej pracy polegało na określeniu optymalnych zakresów stężeń matrycy DNA używanych do analiz (Samelak-Czajka i wsp. 2017: Figure 2). Metodę MLPA wykorzystaliśmy następnie w naszych badaniach (Zmienko i wsp. 2020, Marszalek-Zeńczak i wsp. 2023, Samelak-Czajka i wsp. 2023). W przypadku walidacji wyników genotypowania wyselekcjonowaliśmy 45 PCG-CNV i 4 geny niezmienne, których liczbę kopii weryfikowaliśmy metodą MLPA aż w 30% wszystkich analizowanych linii (314 próbek) (Zmienko i wsp. 2020: Figure 7, Supplemental Figure 13-15). Wszystkie etapy weryfikacji wyników pokazują dużą wiarygodność stworzonego przez nas katalogu regionów zmiennych dla Arabidopsis.

3.2 Charakterystyka i analiza porównawcza poziomu zmienności MGC w kontekście ich funkcji

Dysponując mapą regionów zmiennych AthCNV i wynikami genotypowania, postanowiliśmy zbadać zmienność strukturalną metabolicznych klastrów genów u Arabidopsis, co opisaliśmy w pracy pt. „Analysis of Arabidopsis non-reference accessions reveals high diversity of metabolic gene clusters and discovers new candidate cluster members" (Marszalek-Zenczak i wsp. 2023), również wchodzącej w skład mojej rozprawy doktorskiej. Jak dotąd, u tej rośliny modelowej znane są cztery MGC: klaster biosyntezy thalianolu, marneralu, tirucalladienolu i arabidiolu/baruolu, związane odpowiednio z syntezą ww. wyspecjalizowanych triterpenów. Triterpeny to związki naturalne powstające w reakcji cyklizacji 2,3-oksydoskwalenu, katalizowanej przez cyklazy oksydoskwalenu (ang. *Oxidosqualene Cyclases*, OSC). W genomie Arabidopsis zidentyfikowano dotychczas 13 genów OSC, z czego pięć (*THAS1*, *MRN1*, *PEN3*, *PEN1*, *BARS1*) znajduje się w obrębie MGC i koduje enzymy podstawowe.

Prace rozpoczęłam od obszernego przeglądu literatury w celu stworzenia szczegółowej charakterystyki tych czterech klastrów, ustalenia zakresu ich wielkości oraz określenia genów je tworzących (Marszalek-Zenczak i wsp. 2023: INTRODUCTION, Table S1). W wyniku analizy wariantów AthCNV w obrębie poszczególnych MGC zaobserwowałam duże zróżnicowanie pokrycia, od ∼50% obszaru obejmującego wyłącznie regiony międzygenowe w klastrze marneralu, aż po całkowite pokrycie klastra thalianolu (Marszalek-Zenczak i wsp. 2023: Figure 1A, Table S2). Następnie, aby przyjrzeć się zmianom dla wszystkich genów zlokalizowanych w badanych MGC na poziomie indywidualnych linii, określiłam liczbę ich kopii poprzez zestawienie danych z genotypowania oraz wyników eksperymentalnych. Te ostatnie obejmowały zarówno analizy MLPA, przeprowadzone dla 13 genów w 232 liniach i bazujące na opracowanym wcześniej protokole, jak i analizy PCR techniką emulsyjną (ang. *Droplet Digital PCR*, ddPCR), które wykonałam dla 4 genów klastra thalianolu w 20 wybranych liniach. Ustaliłam, że 98.8% wyników eksperymentalnych było zgodnych z przewidywaniami bioinformatycznymi, zaś kilka wątpliwych przypadków szczegółowo przeanalizowałam i rozwiązałam indywidualnie (Marszalek-Zenczak i wsp. 2023: Table S7, Figure S4). Ogółem, uzyskałam obszerny i wiarygodny zestaw danych dla 1 152 próbek (Rysunek 3.1). Dla 30% analizowanych linii nie zaobserwowałam zmian w liczbie kopii w żadnym z genów w obrębie MGC. CNV zidentyfikowałam łącznie w 19 genach (w klastrze biosyntezy thalianolu - 4, marneralu - 1, tirucalladienolu – 3, arabidiolu/baruolu – 11) (Marszalek-Zenczak i wsp. 2023: Figure 1C, Table S6).



Rysunek 3.1: Integracja danych bioinformatycznych (AthCNV) i eksperymentalnych (MLPA i ddPCR). Kwadraty z kolorową ramką wskazują liczbę genów danego klastra, dla których posiadałam wyniki przy użyciu poszczególnego podejścia. Niebieski – klaster biosyntezy thalianolu, Zielony – klaster biosyntezy marneralu, Różowy – klaster biosyntezy tirucalladienolu, Pomarańczowy – klaster biosyntezy arabidiolu/baruolu. Diagram Venna obrazuje część wspólną linii w różnych metodach

Klaster thalianolu jest najlepiej opisanym i poznanym MGC u Arabidopsis. Początkowo uważano, że składa się on z czterech sąsiadujących ze sobą genów zlokalizowanych na chromosomie 5. Jednak badania prowadzone w ostatnich latach wykazały nie tylko istnienie genów peryferyjnych, ale również doprowadziły do identyfikacji piątego genu w klastrze centralnym (*AT5G47950*), zaangażowanego w metabolizm thalianolu, oddzielonego dwoma niepowiązanymi z tym szlakiem genami (*RABA4C*, *AT5G47970*) (Huang i wsp. 2019). Kolejne badania ujawniły obecność inwersji obejmującej *AT5G47950* i wspomniane dwa geny w 17 z 22 analizowanych linii (77%) (Liu i wsp. 2020). Efektem takiej inwersji (nieobecnej w genomie referencyjnym) jest uzyskanie ciągłości całego klastra biosyntezy thalianolu. W moich badaniach zaobserwowałam pięć typów wariantów CNV w obrębie tego MGC, występujących łącznie w 54 liniach (Marszalek-Zenczak i wsp. 2023: Figure 1B, Figure 2A). Cztery typy obejmowały delecje różnych genów klastra, a jeden duplikację genu *AT5G47950.* We wszystkich wzorcach zmian zaobserwowałam związek z pochodzeniem geograficznym danej linii. Obecność duplikacji potwierdziliśmy w złożonej de novo sekwencji genomowej linii Mitterberg 2-185, jednocześnie wykrywając w niej wspomnianą inwersję. W związku z tym postanowiłam użyć program BreakDancer do detekcji inwersji i przeanalizować wszystkie linie Arabidopsis, dla których dysponowaliśmy sparowanymi odczytami sekwencjonowania (997 próbek). Zidentyfikowałam inwersję o długości 12.8-15.4 kpz, zgodnie z oczekiwaniami obejmującą wszystkie trzy geny (*AT5G47950*, *RABA4C*, *AT5G47970*) aż w 649 liniach, czyli w ∼65% wszystkich analizowanych próbek (Marszalek-Zenczak i wsp. 2023: Table S8, Figure 2B-E). Ta ciągła, i jak się okazało, dominująca wersja klastra była silnie nadreprezentowana w grupach genetycznych Szwecji oraz Azji, podczas gdy referencyjna wersja klastra stanowiła większość linii reprezentujących USA i Hiszpanię. Porównując grupy o dwóch wersjach organizacji klastra zaobserwowałam, że w wersji referencyjnej (nieciągłej) zmiany liczby kopii genów występowały częściej niż w wersji ciągłej (odpowiednio 12.7% i 1.1%), zaś obecność duplikacji korelowała z obecnością inwersji. Różnice te mogą być skutkiem pozytywnej selekcji w kierunku utrzymania ciągłego klastra genów.

Kolejne dwa MGC: klaster biosyntezy marneralu i tirucalladienolu wykazywały niską zmienność. W tym pierwszym zidentyfikowałam częściową delecję *CYP705A12* w jednej linii (Mir-0), co dodatkowo potwierdziliśmy sekwencjonowaniem Sangera. Klaster genów biosyntezy marneralu wydaje się być ustalony na poziomie gatunku, zgodnie z potwierdzoną, krytyczną rolą enzymu syntazy marneralu w rozwoju roślin (Go i wsp. 2012). Z kolei dla drugiego klastra wykryłam obecność CNV w 15 liniach, pochodzących z kilku konkretnych lokalizacji geograficznych, a zatem miały charakter incydentalny.

Klaster biosyntezy arabidiolu/baruolu jest największym i najbardziej zróżnicowanym MGC pod względem wykrytych CNV. Zmianę liczby kopii w co najmniej jednym genie tego klastra zidentyfikowałam w około 75% linii. W obrębie klastra znajdują się dwa geny OSC kodujące enzymy podstawowe: *PEN1* i *BARS1* oraz liczne geny CYP450. Pierwsza para genów, *CYP705A1* i *PEN1*, była niezmienna pod względem liczby kopii. Wcześniejsze badania wykazały, że geny *PEN1* i *CYP705A1* są zaangażowane w biosyntezę i modyfikację arabidiolu, zaś ich produkty biorą udział w odpowiedzi na traktowanie kwasem jasmonowym i infekcję patogenem *Pythium irregulare* powodującym gnicie korzeni (Sohrabi i wsp. 2015). Z kolei gen *BARS1*, biorący udział w biosyntezie baruolu, jak i otaczające go CYP450 wykazywały bardzo wysoki poziom zmienności. Od samego początku analiza locus *BARS1* była dla mnie problematyczna. Zaobserwowałam liczne rozbieżności między wynikami eksperymentalnymi, a przewidywaniami bioinformatycznymi. Analizując dogłębnie odczyty zmapowane w tym regionie zauważyłam, że w większości linii brakuje odczytów mapujących się do największego intronu, gdzie dodatkowo adnotowany jest TE. Ponadto, w danych NGS zaobserwowaliśmy mieszankę homo- i heterozygotycznych SNP (zwanych dalej heteroSNP). Ponieważ genom Arabidopsis jest wysoce homozygotyczny, tak duża kumulacja heteroSNP wydawała się mało prawdopodobna. Zauważyliśmy również zależność między liczbą heteroSNP w locus *BARS1* i duplikacją *CYP705A2*, którą wykryłam w ponad 37% próbek. Wysnuliśmy hipotezę, że odczyty te pochodzą z innego, homologicznego do *BARS1* locus, którego nie było w genomie referencyjnym, natomiast było obecne w wielu innych liniach. Aby zweryfikować tę hipotezę przeanalizowaliśmy i porównaliśmy cały region klastra w złożonych de novo sekwencjach genomowych siedmiu linii Arabidopsis (Jiao i Schneeberger, 2020). W czterech liniach, poza *BARS1* zidentyfikowaliśmy dodatkowy gen, wykazujący duże podobieństwo do syntazy baruolu 1, który nazwaliśmy *BARS2*. Znaleźliśmy również gen o wysokiej homologii do *CYP705A2* i nazwaliśmy go *CYP705A2a*. Razem, te dwa nowo zidentyfikowane geny wchodziły w skład dużej insercji w sekwencji genomowej, rozszerzającej cały klaster o 21-27 kpz. Porównanie sekwencji genów *BARS1* i *BARS2* na poziomie nukleotydów i białka, identyfikacja domen, analiza filogenetyczna oraz modelowanie 3D potwierdziły, że nowo odkryty gen jest duplikatem *BARS1*. Tym samym *BARS2* stanowi niereferencyjny gen kodujący nową, dotychczas niescharakteryzowaną OSC w genomie Arabidopsis (Marszalek-Zenczak i wsp. 2023: Figure 3).

W kolejnym kroku sprawdziłam w jakiej części próbek występuje nowo wykryta insercja. Łącząc dwie informacje: status duplikacji w genie *CYP705A2* oraz liczbę heteroSNP w locus *BARS1*, dla każdej linii określiłam najbardziej prawdopodobny genotyp (Marszalek-Zenczak i wsp 2023: Table S11, Figure S10-S11). Wyróżniłam cztery grupy genotypów: PP-AA (obecna para referencyjna, brak insercji); PP-PP (obecna para referencyjna i insercja); AA-PP (obecna wyłącznie insercja); AA-AA (brak obu par genów). Referencyjna grupa PP-AA była najliczniejsza i obejmowała łącznie 628 linie. Jednak w prawie jednej trzeciej badanej populacji (326 linii) zidentyfikowałam obie pary genów. Dwie pozostałe grupy pośrednie składały się z linii lokalnego pochodzenia, i tak grupa AA-PP reprezentowana była głównie przez próbki z Azerbejdżanu, a grupa AA-AA przez próbki z północnej Szwecji. Linie posiadające dodatkową parę genów (AA-PP i PP-PP) obecne były we wszystkich grupach genetycznych, przy czym obie pary genów (PP-PP) były nadreprezentowane w grupie Reliktów, Hiszpanii i Włoch, jednak stanowiły mniejszość wśród grupy ze Szwecji i Azji (Marszalek-Zenczak i wsp. 2023: Figure 4A-B).

Badania prowadzone przez Boutanaev i wsp. (2015) wykazały, że geny kodujące syntazy terpenów (ang. *Terpenoid Synthases*, TS; obejmujące także analizowane przez nas geny OSC) często występują w pobliżu CYP450, jako pary TS-CYP450 i stanowią najczęściej pojawiające się pary w regionach MGC. Analizując klastry u Arabidopsis zaobserwowałam, że występujące w obrębie klastrów pary TS-CYP450 miały często podobny status (często były razem duplikowane lub usuwane). Zainspirowana tą obserwacją postanowiłam sprawdzić czy istnieje podobna zależność między innymi parami TS-CYP450 w genomie Arabidopsis. W tym celu, w oparciu o różne źródła literaturowe oraz liczne bazy danych stworzyłam obszerną listę wszystkich TS (48) i CYP450 (242) w genomie Arabidopsis (Marszalek-Zenczak i wsp. 2023: Table S14,S15). Zmienność w >1% linii zaobserwowałam dla 13 TS oraz 33 CYP450 reprezentujących tylko trzy klany: CYP71, CYP85 i CYP72. Dodatkowo, analiza par TS-CYP, położonych we wzajemnej odległości do 30 kpz wykazała, że pary mają większą zmienność niż ich niesparowane odpowiedniki, potwierdzając moje przypuszczenia.

3.3 CNV jako użyteczne markery w badaniu struktury populacji i historii demograficznej Arabidopsis

Wcześniejsze analizy genetyczne wykonane na danych SNP wskazywały silne ustrukturyzowanie populacji Arabidopsis, zaś rozkład wydzielonych podgrup korelował z rozkładem geograficznym próbek (Cao i wsp. 2011, 1001 Genomes Consortium 2016). Korzystając z danych z genotypowania postanowiliśmy sprawdzić użyteczność CNV jako markerów w analizach populacyjnych. W tym celu opracowałam i wykonałam analizę głównych składowych (ang. *Principal Component Analysis*, PCA) na dwóch zestawach danych: 1 050 wybranych genach wykazujących wyraźny polimorfizm liczby kopii oraz na danych SNP (pobranych z Projektu 1001 Genomów). Rozkład grup genetycznych uzyskany dla obu typu markerów nie tylko pokazał, że dane CNV dobrze odzwierciedlają rozkład geograficzny próbek, ale również ujawnił nowe zależności między grupami, m.in. bliskość między grupą reliktów i północną Szwecją, niewidoczne wcześniej na danych SNP (Zmienko i wsp. 2020: Figure 8).

W badaniach nad zmiennością klastra biosyntezy arabidiolu/baruolu postanowiłam wykorzystać wykonane wcześniej analizy PCA na danych SNP, poszerzając je o zaktualizowane informacje na temat statusu par genów *CYP705A2-BARS1* i *CYP705A2a-BARS2*, w celu analizy głównych podgrup (PP-AA i PP-PP) wydzielonych w tym klastrze. Dodatkowo, aby uzyskać bardziej szczegółowy wgląd w historię tych loci na tle pozostałej zmienności genetycznej populacji Arabidopsis, wykonałam analizę dla szerokiego zakresu wartości parametru nierównowagi sprzężeń (ang. *Linkage Disequilibrium*, LD) (Marszalek-Zenczak i wsp. 2023: Figure 4C, S17). Najwyraźniejszy rozdział podgrup dotyczący obecności/braku duplikacji obserwowałam dla niskich wartości LD, gdzie udział alleli ancestralnych w PCA był najwyższy. Sugeruje to związek pomiędzy istnieniem duplikacji, a kształtowaniem się rozmieszczenia geograficznego populacji Arabidopsis w przeszłości. Szczegółowa analiza naturalnego występowania poszczególnych linii wykazała, że linie z insercją *CYP705A2a-BARS2* pochodzą z regionów o niższej szerokości geograficznej niż linie należące do grupy referencyjnej. Co więcej, różnice zaobserwowaliśmy także w poszczególnych państwach, dla próbek z Niemiec, Hiszpanii i Włoch (Marszalek-Zenczak i wsp. 2023: Figure 4D, S18).

3.4 Rola CNV w kształtowaniu zmienności fenotypowej i adaptacji do warunków środowiska

Z uwagi na dużą użyteczność markerów CNV w analizach genetycznych postanowiliśmy zastosować je także w badaniach GWAS. W tym celu opracowałam i zaadaptowałam metodę analizy asocjacji CNV z wybranymi fenotypami. Jak wspomniałam we wstępie, GWAS to potężne narzędzie, jednak stawia przez badaczami sporo wyzwań. Wzorując się na analizach SNP zastosowałam podejście wykorzystujące zakodowanie zmian w sposób dyskretny, przypisując status: delecja/duplikacja/brak zmiany. Należy podkreślić, że podejście to nie pozwala na analizę wieloallelicznych CNV z dokładnym uwzględnieniem ich wartości liczby kopii. Analizie poddałam 2519 PCG-CNV, wykazujących zmienność w co najmniej 1% badanej populacji. Do wykonania analizy GWAS zastosowałam model EMMAX (ang. *Efficient Mixed-Model Association eXpedited*, EMMAX), który uwzględnienia korektę na strukturę populacji. Linie Arabidopsis są wysoce wsobne i jednorodne genetycznie, co w badaniach GWAS ma ogromne, korzystne znaczenie. Takie dane mają bowiem znacznie niższy szum niż heterozygotyczne próbki, w związku z czym możliwa jest skuteczna identyfikacja wariantów dysponując niższą liczbą próbek.

Aby przetestować skuteczność analiz, pobrałam z bazy AraPheno dane dla 23 cech fenotypowych (Atwell i wsp. 2010) związanych ze stresem biotycznym. W 8 przypadkach uzyskałam statystycznie istotną asocjację, spośród których szczególnie interesujące jest powiązanie delecji genów *RPS5* oraz *RPM1* z obniżeniem odporności roślin w odpowiedzi na infekcję szczepami *Pseudomonas* (Zmienko i wsp. 2020: Figure 11). Potwierdzało się to z wcześniejszymi wynikami badań pokazującymi, że *RPS5* i *RPM1* to tak zwane geny odporności R, kodujące receptory zaangażowane w rozpoznawanie bakteryjnych białek wirulencji (odpowiednio avrPhB i avrB), produkowanych przez *Pseudomonas*. Zatem, pomimo niskiej liczebności badanych grup (średnio kilkadziesiąt linii), udało nam się potwierdzić duże możliwości wykorzystania danych CNV w analizach GWAS. Wydaje się, że połączenie wzajemnie uzupełniających się informacji o wariantach SNP i CNV pomoże w identyfikacji nowych związków między genotypem, a fenotypem.

W związku z tym postanowiłam poszukać związku odkrytej zmienności w obrębie klastra biosyntezy arabidiolu/baruolu z cechami fenotypowymi. W tym celu wykonałam analizę GWAS na danych SNP zintegrowanych z analizowanym CNV, na wszystkich fenotypach dostępnych wówczas w bazie AraPheno (516 cech). Ten szeroki screening wyróżnił grupę danych klimatycznych, jak również danych związanych z rozwojem korzeni (Marszalek-Zenczak i wsp. 2023: Figure 5A-D, Table S12). Szczegółowa analiza porównawcza par PP-AA i PP-PP wykazała, że linie z obiema parami genów pochodziły z lokalizacji geograficznej o cieplejszym i bardziej wilgotnym klimacie. Zaobserwowałam również wolniejszą dynamikę wzrostu korzeni u linii z insercją *CYP705A2a-BARS2*.

Wcześniejsze badania wskazywały, że geny klastra biosyntezy arabidiolu/baruolu ulegają najwyższej ekspresji w korzeniach. Potwierdziłam to, analizując ich ekspresję w różnych tkankach linii Col-0. Następnie, korzystając z bogatego zbioru danych RNAseq dla liści, stworzonego przez Kawakatsu i wsp. (2016) wykazałam, że linie z dodatkową parą genów mają istotną statystycznie wyższą ekspresję. W związku z tym, korzystając z dostępnych sekwencji genomowych dla kilku linii, postanowiłam na nowo przeanalizować surowe dane RNAseq ze wspomnianej pracy, tym razem mapując je bezpośrednio do odpowiednich transkryptomów. W dwóch liniach z insercją potwierdziłam wyższą ekspresję genu *BARS2* niż *BARS1* w liściach. Dodatkowo porównałam poziom ekspresji intersujących mnie genów w korzeniach, liściach i pędach linii Col-0 (PP-AA) i Cvi-0 (PP-PP). Zaobserwowałam, że w korzeniach występowała ekspresja obu par genów, przy czym ekspresja referencyjnej pary w Col-0 była wyższa niż w linii Cvi-0. Z kolei w liściach nie wykryłam ekspresji dla *CYP705A2-BARS1* w Col-0, aczkolwiek zaobserwowałam ekspresję obu par genów w Cvi-0 (Marszalek-Zenczak i wsp. 2023: Figure 5E-F oraz S21). Może to wskazywać, że zduplikowane pary genów uległy subfunkcjonalizacji.

Jak wykazałam, klaster biosyntezy thalianolu istnieje w dwóch podstawowych wersjach – nieciągłej (referencyjnej) i ciągłej, w których również zaobserwowałam związek między typem klastra, a szerokością geograficzną. Z tego względu postanowiłam sprawdzić potencjalne interakcje między wariantami w obu zmiennych MGC. Okazało się, że zmienność strukturalna klastra biosyntezy arabidiolu/baruolu lepiej wyjaśniała geograficzne rozmieszczenie próbek na projekcji PCA (Marszalek-Zenczak i wsp. 2023: Figure S19). Co więcej, obie wersje klastra biosynezy thalianolu nie różniły się poziomem ekspresji jego genów w liściach ani nie wykazywały różnic w dynamice wzrostu korzeni (Marszalek-Zenczak i wsp. 2023: Figure S20). Wskazuje to, że wykryty związek insercji *CYP705A2a-BARS2* z tą cechą może mieć charakter adaptacyjny, a nie jest wyłącznie wynikiem np. dryftu genetycznego. Podsumowując, wyniki GWAS pozwoliły na selekcję cech być może związanych z insercją w klastrze arabidiolu/baruolu, co w przyszłości powinno zostać szczegółowo zbadane i potwierdzone eksperymentalnie.

Rozdział 4

Podsumowanie i perspektywy

Badania prowadzone w ramach mojej rozprawy doktorskiej wchodzą w skład trzech omówionych wcześniej publikacji: Samelak-Czajka i wsp. (2017), Zmienko i wsp. (2020) oraz Marszalek-Zenczak i wsp. (2023). Efektem badań prowadzonych w ramach Samelak-Czajka i wsp. (2017) było stworzenie uniwersalnego protokołu MLPA, który może być wykorzystany dla różnych roślin. W Zmienko i wsp. (2020) stworzyliśmy katalog dużych indeli (o długości 50-499 pz) oraz regionów $\geq$500 pz wykazujących zmianę liczby kopii, nazwany AthCNV, w ponad 1000 naturalnych liniach Arabidopsis. Ważnym osiągnieciem tej pracy było opracowanie autorskiego podejścia do integracji danych uzyskanych z różnych metod detekcji CNV na podstawie krótkich odczytów sekwencjonowania. Ustalenie ścieżki postępowania z podkreśleniem kluczowych aspektów jak również staranny dobór kryteriów filtracji wyników z poszczególnych programów pozwala na zaadaptowanie i modyfikację podejścia w zależności od problemu badawczego. Wykorzystanie podejścia przez inne grupy badawcze widać w np. publikacji Xu i wsp. (2023). Określenie liczby kopii genów dla analizowanych linii dało lepszy wgląd w poziom zmienności dla tej rośliny modelowej oraz umożliwiło wykorzystanie danych jako markerów do analiz populacyjnych, asocjacji z fenotypem czy do analizy ekspresji. Stworzony katalog zmienności AthCNV wraz z danymi z genotypowania uzupełniają istniejącą lukę w charakterystyce zmienności wewnątrzgatunkowej dla Arabidopsis i są dobrym punktem wyjścia do dalszych badań nad związkiem między zmiennością genetyczną, a fenotypem i adaptacją do zmiennych warunków środowiskowych. Z tego względu, w badaniach opisanych w publikacji Marszalek-Zenczak i wsp. (2023) wykorzystałam uzyskane wcześniej dane do sprawdzenia zmienności strukturalnej metabolicznych klastrów genów u Arabidopsis. Szczegółowa charakterystyka czterech poznanych jak dotąd MGC u Arabidopsis ujawniła, że wszystkie zlokalizowane są w regionach wysoce zmiennych. Co ciekawe, zaobserwowaliśmy całe spektrum zmienności, począwszy od prawie całkowicie niezmiennego klastra marneralu, aż po liczne i duże zmiany w klastrze arabidiolu/baruolu. Klaster genów marneralu wydaje się być ustalony na poziomie gatunku, zgodnie z krytyczną rolą enzymu syntazy marneralu w rozwoju roślin. Z kolei klaster genów thalianolu istnieje w dwóch głównych wersjach: nieciągłej (referencyjnej, występującej w linii Col-0) i ciągłej. Obie wersje różnią się zarówno częstością jak i poziomem zmienności liczby kopii poszczególnych genów klastra. Uzyskane przez nas wyniki wskazują, że ciągła wersja klastra może być w trakcie utrwalania, w przeciwieństwie do wersji nieciągłej. Największy i najbardziej złożony klaster arabidiolu/baruolu istnieje również w dwóch głównych wersjach, różniących się obecnością/nieobecnością nowo odkrytej pary genów *CYP705A2a-BARS2*. *BARS2* to nowy, niescharakteryzowany jak dotąd OSC, obecny w ponad jednej trzeciej analizowanych linii Arabidopsis. Należy podkreślić, że pomimo to, że detekcję CNV wykonaliśmy w oparciu o genom referencyjny, udało nam się zidentyfikować nową syntazę, nieobecną w genomie referencyjnym. Kluczowym aspektem było wykorzystanie informacji o heterozygotycznych SNP, które dla organizmów homozygotycznych mogą służyć jako czynniki wskazujące obecność duplikacji w sekwencji. Warianty strukturalne mogą przyczyniać się do różnic zarówno w poziomie ekspresji genów *CYP705A2-BARS1* jak i specyficzności tkankowej. Obserwowane przez nas różnice w dynamice wzrostu korzeni i odmiennych preferencjach klimatycznych, dla linii z insercją i bez, wymagają weryfikacji eksperymentalnej. Szczegółowy wgląd w zmienność metabolicznych klastrów genów przyczyni się do lepszego zrozumienia tego zjawiska, mechanizmów jego powstawania i regulacji w genomach roślinnych, co umożliwi identyfikację zarówno nowych ścieżek jak i produkowanych w nich związków.

Ulepszenia technologiczne prowadzą do uzyskania coraz dokładniejszych sekwencji genomowych poszczególnych osobników. Poznanie kompletnej sekwencji DNA ujawnia nieznane wcześniej powiązania, szczególnie w najbardziej dynamicznych regionach sekwencji, a przez to zapewnia pełniejsze zrozumienie organizacji i regulacji genomu. Chociaż sekwencjonowanie krótkich odczytów jest nadal wykorzystywane do wykrywania SV, przyszłość stanowią długie odczyty. Klasyczny genom referencyjny zwykle reprezentuje sekwencję DNA pojedynczego osobnika. Dlatego identyfikacja SV w odniesieniu do sekwencji referencyjnej nie jest wystarczająca do pełnego przedstawienia całej różnorodności genetycznej danego gatunku. Aby uzyskać pełną informację konieczne jest skonstruowanie jego pangenomu. Po opublikowaniu wstępnej wersji pangenomu człowieka, Aimé Lumaka, genetyk na Uniwersytecie w Liège w Belgii i Uniwersytecie w Kinszasie w Demokratycznej Republice Konga, napisał: „To coś, na co wszyscy czekaliśmy. Obecnemu genomowi referencyjnemu brakuje nie tylko części informacji genomowych, ale, co najważniejsze, brakuje mu różnorodności" (Liverpool 2023). Głębsze zrozumienie interakcji między genotypem, a fenotypem daje lepszy wgląd w to, jak działa świat natury, a tym samym otwiera nowe perspektywy do lepszego zrozumienia wpływu czynników środowiskowych na różnorodność organizmów, zwiększając możliwości ochrony różnorodności biologicznej jak i naszej planety. Pociąga to za sobą również korzystny i praktyczny dla ludzkości aspekt w postaci rozwoju wielu dziedzin nauki, w tym genetyki, biologii ewolucyjnej i medycyny, a także szukania nowych rozwiązań w m.in. w hodowli roślin i zwierząt, gdzie poznanie zmienności genetycznej pozwala na uzyskiwanie odmian o lepszych właściwościach, takich jak wyższa wydajność czy większa odporność na choroby.

Rozdział 5

Bibliografia

1001 Genomes Consortium. (2016) 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. Cell. 166(2): 481-491. doi:10.1016/j.cell.2016.05.063.

Alonge M, Wang X, Benoit M, Soyk S, Pereira L et al. (2020) Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato. Cell. 182(1): 145-161.e23. doi:10.1016/j.cell.2020.05.021.

Arabidopsis Genome Initiative. (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature. 408(6814): 796-815. doi:10.1038/35048692.

Athanasopoulou K, Boti MA, Adamopoulos PG, Skourou PC, Scorilas A. (2021) Third-Generation Sequencing: The Spearhead towards the Radical Transformation of Modern Genomics. Life (Basel). 12(1): 30. doi:10.3390/life12010030.

Atwell S, Huang YS, Vilhjálmsson BJ, Willems G, Horton M et al. (2010) Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature. 465(7298): 627-31. doi:10.1038/nature08800.

Barton N, Hermisson J, Nordborg M. (2019) Why structure matters. Elife. 8:e45380. doi:10.7554/eLife.45380.

Beló A, Beatty MK, Hondred D, Fengler KA, Li B et al. (2010) Allelic genome structural variations in maize detected by array comparative genome hybridization. Theor Appl Genet. 120(2): 355-67. doi:10.1007/s00122-009-1128-9.

Bok JW, Chiang YM, Szewczyk E, Reyes-Dominguez Y, Davidson AD et al. (2009) Chromatin-level regulation of biosynthetic gene clusters. Nat Chem Biol. 5(7): 462-4. doi:10.1038/nchembio.177.

Boutanaev AM, Moses T, Zi J, Nelson DR, Mugford ST et al. (2015) Investigation of terpene diversification across multiple sequenced plant genomes. Proc Natl Acad Sci U S A. 112(1): E81-8. doi:10.1073/pnas.1419547112.

Boycheva S, Daviet L, Wolfender JL, Fitzpatrick TB. (2014) The rise of operon-like gene clusters in plants. Trends Plant Sci. 19(7): 447-59. doi:10.1016/j.tplants.2014.01.013.

Brakhage AA. (2013) Regulation of fungal secondary metabolism. Nat Rev Microbiol. 11(1): 21-32. doi:10.1038/nrmicro2916.

Cao J, Schneeberger K, Ossowski S, Günther T, Bender S et al. (2011) Whole-genome sequencing of multiple Arabidopsis thaliana populations. Nat Genet. 43(10): 956-63. doi:10.1038/ng.911.

Carpenter D, Dhar S, Mitchell LM, Fu B, Tyson J et al. (2015) Obesity, starch digestion and amylase: association between copy number variants at human salivary (AMY1) and pancreatic (AMY2) amylase genes. Hum Mol Genet. 24(12): 3472-80. doi:10.1093/hmg/ddv098.

Chawla HS, Lee H, Gabur I, Vollrath P, Tamilselvan-Nattar-Amutha S et al. (2021) Long-read sequencing reveals widespread intragenic structural variants in a recent allopolyploid crop plant. Plant Biotechnol J. 19(2): 240-250. doi:10.1111/pbi.13456.

Chia JM, Song C, Bradbury PJ, Costich D, de Leon N et al. (2012) Maize HapMap2 identifies extant variation from a genome in flux. Nat Genet. 44(7): 803-7. doi:10.1038/ng.2313.

Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O et al. (2010) Origins and functional impact of copy number variation in the human genome. Nature. 464(7289): 704-12. doi:10.1038/nature08516.

Cook DE, Bayless AM, Wang K, Guo X, Song Q et al. (2014) Distinct Copy Number, Coding Sequence, and Locus Methylation Patterns Underlie Rhg1-Mediated Soybean Resistance to Soybean Cyst Nematode. Plant Physiol. 165(2): 630-647. doi:10.1104/pp.114.235952.

Cook DE, Lee TG, Guo X, Melito S, Wang K et al. (2012) Copy number variation of multiple genes at Rhg1 mediates nematode resistance in soybean. Science. 338(6111): 1206-9. doi:10.1126/science.1228746.

De Oliveira R, Rimbert H, Balfourier F, Kitt J, Dynomant E et al. (2020) Structural Variations Affecting Genes and Transposable Elements of Chromosome 3B in Wheats. Front Genet. 11:891. doi:10.3389/fgene.2020.00891.

Díaz A, Zikhali M, Turner AS, Isaac P, Laurie DA. (2012) Copy number variation affecting the Photoperiod-B1 and Vernalization-A1 genes is associated with altered flowering time in wheat (Triticum aestivum). PLoS One. 7(3): e33234. doi:10.1371/journal.pone.0033234.

Dolatabadian A, Patel DA, Edwards D, Batley J. (2017) Copy number variation and disease resistance in plants. Theor Appl Genet. 130(12): 2479-2490. doi:10.1007/s00122-017-2993-2.

Elshafie HS, Camele I, Mohamed AA. (2023) A Comprehensive Review on the Biological, Agricultural and Pharmaceutical Properties of Secondary Metabolites Based-Plant Origin. Int J Mol Sci. 24(4): 3266. doi:10.3390/ijms24043266.

Emerson JJ, Cardoso-Moreira M, Borevitz JO, Long M. (2008) Natural selection shapes genome-wide patterns of copy-number polymorphism in Drosophila melanogaster. Science. 320(5883): 1629-31. doi:10.1126/science.1158078.

Erb M, Kliebenstein DJ. (2020) Plant Secondary Metabolites as Defenses, Regulators, and Primary Metabolites: The Blurred Functional Trichotomy. Plant Physiol. 184(1): 39-52. doi:10.1104/pp.20.00433.

Escaramís G, Docampo E, Rabionet R. (2015) A decade of structural variants: description, history and methods to detect structural variation. Brief Funct Genomics. 14(5): 305-14. doi:10.1093/bfgp/elv014.

Feuk L, Carson AR, Scherer SW. (2006) Structural variation in the human genome. Nat Rev Genet. 7(2): 85-97. doi:10.1038/nrg1767.

Field B, Fiston-Lavier AS, Kemen A, Geisler K, Quesneville H et al. (2011) Formation of plant metabolic gene clusters within dynamic chromosomal regions. Proc Natl Acad Sci U S A. 108(38): 16116-21. doi:10.1073/pnas.1109273108.

Field B, Osbourn AE. (2008) Metabolic diversification-independent assembly of operon-like gene clusters in different plants. Science. 320(5875): 543-7. doi:10.1126/science.1154990.

Frey M, Chomet P, Glawischnig E, Stettner C, Grün S et al. (1997) Analysis of a chemical plant defense mechanism in grasses. Science. 277(5326): 696-9. doi:10.1126/science.277.5326.696.

Frey M, Huber K, Park WJ, Sicker D, Lindberg P et al. (2003) A 2-oxoglutarate-dependent dioxygenase is integrated in DIMBOA-biosynthesis. Phytochemistry. 62(3): 371-6. doi:10.1016/s0031-9422(02)00556-3.

Gao L, Gonda I, Sun H, Ma Q, Bao K et al. (2019) The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. Nat Genet. 51(6): 1044-1051. doi:10.1038/s41588-019-0410-2.

Girirajan S, Campbell CD, Eichler EE. (2011) Human copy number variation and complex genetic disease. Annu Rev Genet. 45: 203-26. doi:10.1146/annurev-genet-102209-163544.

Giza A, Iwan E, Bomba A, Wasyl D. (2021) Podstawy sekwencjonowania wysokoprzepustowego. Med. Weter. 77(11): 530-534. doi:dx.doi.org/10.21521/mw.6594.

Go YS, Lee SB, Kim HJ, Kim J, Park HY et al. (2012) Identification of marneral synthase, which is critical for growth and development in Arabidopsis. Plant J. 72(5): 791-804. doi:10.1111/j.1365-313X.2012.05120.x.

Guo J, Cao K, Deng C, Li Y, Zhu G et al. (2020) An integrated peach genome structural variation map uncovers genes associated with fruit traits. Genome Biol. 21(1): 258. doi:10.1186/s13059-020-02169-y.

Hanikenne M, Kroymann J, Trampczynska A, Bernal M, Motte P et al. (2013) Hard selective sweep and ectopic gene conversion in a gene cluster affording environmental adaptation. PLoS Genet. 9(8): e1003707. doi:10.1371/journal.pgen.1003707.

Hanikenne M, Talke IN, Haydon MJ, Lanz C, Nolte A et al. (2008) Evolution of metal hyperaccumulation required cis-regulatory changes and triplication of HMA4. Nature. 453(7193): 391-5. doi:10.1038/nature06877.

Haun WJ, Hyten DL, Xu WW, Gerhardt DJ, Albert TJ et al. (2011) The composition and origins of genomic variation among individuals of the soybean reference cultivar Williams 82. Plant Physiol. 155(2): 645-55. doi:10.1104/pp.110.166736.

Ho SS, Urban AE, Mills RE. (2020) Structural variation in the sequencing era. Nat Rev Genet. 21(3): 171-189. doi:10.1038/s41576-019-0180-9.

Hollox EJ, Zuccherato LW, Tucci S. (2022) Genome structural variation in human evolution. Trends Genet. 38(1): 45-58. doi:10.1016/j.tig.2021.06.015.

Hu L, Yao X, Huang H, Guo Z, Cheng X et al. (2018) Clinical significance of germline copy number variation in susceptibility of human diseases. J Genet Genomics. 45(1): 3-12. doi:10.1016/j.jgg.2018.01.001.

Huang AC, Jiang T, Liu YX, Bai YC, Reed J et al. (2019) A specialized metabolic network selectively modulates Arabidopsis root microbiota. Science. 364(6440): eaau6389. doi:10.1126/science.aau6389.

Hufford MB, Seetharam AS, Woodhouse MR, Chougule KM, Ou S et al. (2021) De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. Science. 373(6555): 655-662. doi:10.1126/science.abg5289.

Hübner S, Bercovich N, Todesco M, Mandel JR, Odenheimer J et al. (2019) Sunflower pan-genome analysis shows that hybridization altered gene content and disease resistance. Nat Plants. 5(1): 54-62. doi:10.1038/s41477-018-0329-0.

Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK et al. (2004) Detection of large-scale variation in the human genome. Nat Genet. 36(9): 949-51. doi:10.1038/ng1416.

International Consortium for Blood Pressure Genome-Wide Association Studies (2011). Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. Nature. 478(7367): 103-9. doi:10.1038/nature10405.

Isah T. (2019) Stress and defense responses in plant secondary metabolites production. Biol Res. 52(1): 39. doi:10.1186/s40659-019-0246-3.

Itkin M, Heinig U, Tzfadia O, Bhide AJ, Shinde B et al. (2013) Biosynthesis of antinutritional alkaloids in solanaceous crops is mediated by clustered genes. Science. 341(6142): 175-9. doi:10.1126/science.1240230.

Itkin M, Rogachev I, Alkan N, Rosenberg T, Malitsky S et al. (2011) GLYCOALKALOID METABOLISM1 is required for steroidal alkaloid glycosylation and prevention of phytotoxicity in tomato. Plant Cell. 23(12): 4507-25. doi:10.1105/tpc.111.088732.

Jasiński M, Mazurkiewicz E, Rodziewicz P, Figlerowicz M. (2009) Flawonoidy–budowa, właściwości i funkcja ze szczególnym uwzględnieniem roślin motylkowatych. biotechnologia. 2(85): 81-94.

Jayakodi M, Padmarasu S, Haberer G, Bonthala VS, Gundlach H et al. (2020) The barley pan-genome reveals the hidden legacy of mutation breeding. Nature. 588(7837): 284-289. doi:10.1038/s41586-020-2947-8.

Jiao WB, Schneeberger K. (2020) Chromosome-level assemblies of multiple Arabidopsis genomes reveal hotspots of rearrangements with altered evolutionary dynamics. Nat Commun. 11(1): 989. doi:10.1038/s41467-020-14779-y.

Jonczyk R, Schmidt H, Osterrieder A, Fiesselmann A, Schullehner K et al. (2008) Elucidation of the final reactions of DIMBOA-glucoside biosynthesis in maize: characterization of Bx6 and Bx7. Plant Physiol. 146(3): 1053-63. doi:10.1104/pp.107.111237.

Kautsar SA, Suarez Duran HG, Blin K, Osbourn A, Medema MH. (2017) plantiSMASH: automated identification, annotation and expression analysis of plant biosynthetic gene clusters. Nucleic Acids Res. 45(W1): W55-W63. doi:10.1093/nar/gkx305.

Kawakatsu T, Huang SC, Jupe F, Sasaki E, Schmitz RJ et al. (2016) Epigenomic Diversity in a Global Collection of Arabidopsis thaliana Accessions. Cell. 166(2): 492-505. doi:10.1016/j.cell.2016.06.044.

Kou Y, Liao Y, Toivainen T, Lv Y, Tian X et al. (2020) Evolutionary Genomics of Structural Variation in Asian Rice (Oryza sativa) Domestication. Mol Biol Evol. 37(12): 3507-3524. doi:10.1093/molbev/msaa185.

Lam HM, Xu X, Liu X, Chen W, Yang G et al. (2010) Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. Nat Genet. 42(12): 1053-9. doi:10.1038/ng.715.

Lang Z, Xie S, Zhu JK. (2016) The 1001 Arabidopsis DNA Methylomes: An Important Resource for Studying Natural Genetic, Epigenetic, and Phenotypic Variation. Trends Plant Sci. 21(11): 906-908. doi:10.1016/j.tplants.2016.09.001.

Lee C, Iafrate AJ, Brothman AR. (2007) Copy number variations and clinical cytogenetic diagnosis of constitutional disorders. Nat Genet. 39(7 Suppl): S48-54. doi:10.1038/ng2092.

Lee TG, Kumar I, Diers BW, Hudson ME. (2015) Evolution and selection of Rhg1, a copy-number variant nematode-resistance locus. Mol Ecol. 24(8): 1774-91. doi:10.1111/mec.13138.

Li N, He Q, Wang J, Wang B, Zhao J et al. (2023) Super-pangenome analyses highlight genomic diversity and structural variation across wild and cultivated tomato species. Nat Genet. 55(5): 852-860. doi:10.1038/s41588-023-01340-y.

Li W, Olivier M. (2013) Current analysis platforms and methods for detecting copy number variation. Physiol Genomics. 45(1): 1-16. doi:10.1152/physiolgenomics.00082.2012.

Li Y, Xiao J, Wu J, Duan J, Liu Y et al. (2012) A tandem segmental duplication (TSD) in green revolution gene Rht-D1b region underlies plant height variation. New Phytol. 196(1): 282-291. doi:10.1111/j.1469-8137.2012.04243.x.

Lian Q, Hüttel B, Walkemeier B, Mayjonade B, Lopez-Roques C et al. (2023) A pan-genome of 72 Arabidopsis thaliana accessions reveals a conserved genome structure throughout the global species range. PREPRINT (Version 1) available at Research Square. DOI: 10.21203/rs.3.rs-2976609/v1.

Liao WW, Asri M, Ebler J, Doerr D, Haukness M et al. (2023) A draft human pangenome reference. Nature. 617(7960): 312-324. doi:10.1038/s41586-023-05896-x.

Lin G, He C, Zheng J, Koo DH, Le H et al. (2021) Chromosome-level genome assembly of a regenerable maize inbred line A188. Genome Biol. 22(1): 175. doi:10.1186/s13059-021-02396-x.

Lind AL, Wisecaver JH, Lameiras C, Wiemann P, Palmer JM et al. (2017) Drivers of genetic diversity in secondary metabolic gene clusters within a fungal species. PLoS Biol. 15(11): e2003583. doi:10.1371/journal.pbio.2003583.

Liu Y, Du H, Li P, Shen Y, Peng H et al. (2020) Pan-Genome of Wild and Cultivated Soybeans. Cell. 182(1): 162-176.e13. doi:10.1016/j.cell.2020.05.023.

Liu Z, Cheema J, Vigouroux M, Hill L, Reed J et al. (2020) Formation and diversification of a paradigm biosynthetic gene cluster in plants. Nat Commun. 11(1): 5354. doi:10.1038/s41467-020-19153-6.

Liverpool L. (2023) First human 'pangenome' aims to catalogue genetic diversity. Nature. 617(7961): 444–445. doi:10.1038/d41586-023-01576-y.

Louzada S, Algady W, Weyell E, Zuccherato LW, Brajer P et al. (2020) Structural variation of the malaria-associated human glycophorin A-B-E region. BMC Genomics. 21(1): 446. doi:10.1186/s12864-020-06849-8.

Lye ZN, Purugganan MD. (2019) Copy Number Variation in Domestication. Trends Plant Sci. 24(4): 352-365. doi:10.1016/j.tplants.2019.01.003.

Ma X, Fan J, Wu Y, Zhao S, Zheng X et al. (2020) Whole-genome de novo assemblies reveal extensive structural variations and dynamic organelle-to-nucleus DNA transfers in African and Asian rice. Plant J. 104(3): 596-612. doi:10.1111/tpj.14946.

Maron LG, Guimarães CT, Kirst M, Albert PS, Birchler JA et al. (2013) Aluminum tolerance in maize is associated with higher MATE1 gene copy number. Proc Natl Acad Sci U S A. 110(13): 5241-6. doi:10.1073/pnas.1220766110.

Marszalek-Zenczak M, Satyr A, Wojciechowski P, Zenczak M, Sobieszczanska P et al. (2023) Analysis of Arabidopsis non-reference accessions reveals high diversity of metabolic gene clusters and discovers new candidate cluster members. Front Plant Sci. 14: 1104303. doi:10.3389/fpls.2023.1104303.

Marx V. (2023) Method of the year: long-read sequencing. Nat Methods. 20(1): 6-11. doi:10.1038/s41592-022-01730-w.

Matsuba Y, Nguyen TT, Wiegert K, Falara V, Gonzales-Vigil E et al. (2013) Evolution of a complex locus for terpene biosynthesis in solanum. Plant Cell. 25(6): 2022-36. doi:10.1105/tpc.113.111013.

McHale LK, Haun WJ, Xu WW, Bhaskar PB, Anderson JE et al. (2012) Structural variants in the soybean genome localize to clusters of biotic stress-response genes. Plant Physiol. 159(4): 1295-308. doi:10.1104/pp.112.194605.

Montenegro JD, Golicz AA, Bayer PE, Hurgobin B, Lee H et al. (2017) The pangenome of hexaploid bread wheat. Plant J. 90(5): 1007-1013. doi:10.1111/tpj.13515.

Mugford ST, Louveau T, Melton R, Qi X, Bakht S et al. (2013) Modularity of plant metabolic gene clusters: a trio of linked genes that are collectively required for acylation of triterpenes in oat. Plant Cell. 25(3):1078-92. doi:10.1105/tpc.113.110551.

Mugford ST, Qi X, Bakht S, Hill L, Wegel E et al. (2009) serine carboxypeptidase-like acyltransferase is required for synthesis of antimicrobial compounds and disease resistance in oats. Plant Cell. 21(8): 2473-84. doi:10.1105/tpc.109.065870.

Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV et al. (2022) The complete sequence of a human genome. Science. 376(6588): 44-53. doi:10.1126/science.abj6987.

Nützmann HW, Huang A, Osbourn A. (2016) Plant metabolic clusters - from genetics to genomics. New Phytol. 211(3): 771-89. doi:10.1111/nph.13981.

Nützmann HW, Osbourn A. (2015) Gene clustering in plant specialized metabolism. Curr Opin Biotechnol. 26: 91-9. doi:10.1016/j.copbio.2013.10.009.

Nützmann HW, Osbourn A. (2015) Regulation of metabolic gene clusters in Arabidopsis thaliana. New Phytol. 205(2): 503-10. doi:10.1111/nph.13189.

Nützmann HW, Scazzocchio C, Osbourn A. (2018) Metabolic Gene Clusters in Eukaryotes. Annu Rev Genet. 52: 159-183. doi:10.1146/annurev-genet-120417-031237.

Pearson TA, Manolio TA. (2008) How to interpret a genome-wide association study. JAMA. 299(11): 1335-44. doi:10.1001/jama.299.11.1335.

Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H et al. (2007) Diet and the evolution of human amylase gene copy number variation. Nat Genet. 39(10): 1256-60. doi:10.1038/ng2123.

Qi X, Bakht S, Leggett M, Maxwell C, Melton R et al. (2004) A gene cluster for secondary metabolism in oat: implications for the evolution of metabolic diversity in plants. Proc Natl Acad Sci U S A. 101(21): 8233-8. doi:10.1073/pnas.0401301101.

Saintenac C, Jiang D, Akhunov ED. (2011) Targeted analysis of nucleotide and copy number variation by exon capture in allotetraploid wheat genome. Genome Biol. 12(9): R88. doi:10.1186/gb-2011-12-9-r88.

Samelak-Czajka A, Marszalek-Zenczak M, Marcinkowska-Swojak M, Kozlowski P, Figlerowicz M et al. (2017) MLPA-Based Analysis of Copy Number Variation in Plant Populations. Front Plant Sci. 8: 222. doi:10.3389/fpls.2017.00222.

Samelak-Czajka A, Wojciechowski P, Marszalek-Zenczak M, Figlerowicz M, Zmienko A. (2023) Differences in the intraspecies copy number variation of Arabidopsis thaliana conserved and nonconserved miRNA genes. Funct Integr Genomics. 23(2): 120. doi:10.1007/s10142-023-01043-x.

Sanger F, Nicklen S, Coulson AR. (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 74(12): 5463-5467. doi:10.1073/pnas.74.12.5463

Santuari L, Pradervand S, Amiguet-Vercher AM, Thomas J, Dorcey E et al. (2010) Substantial deletion overlap among divergent Arabidopsis genomes revealed by intersection of short reads and tiling arrays. Genome Biol. 11(1): R4. doi:10.1186/gb-2010-11-1-r4.

Saxena RK, Edwards D, Varshney RK. (2014) Structural variations in plant genomes. Brief Funct Genomics. 13(4): 296-307. doi:10.1093/bfgp/elu016.

Sebat J, Lakshmi B, Troge J, Alexander J, Young J et al. (2004) Large-scale copy number polymorphism in the human genome. Science. 305(5683): 525-8. doi:10.1126/science.1098918.

Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C et al. (2007) Strong association of de novo copy number mutations with autism. Science. 316(5823): 445-9. doi:10.1126/science.1138659.

Sekine M, Makino T. (2017) Inference of Causative Genes for Alzheimer's Disease Due to Dosage Imbalance. Mol Biol Evol. 34(9): 2396-2407. doi:10.1093/molbev/msx183.

Shaikh TH. (2017) Copy Number Variation Disorders. Curr Genet Med Rep. 5(4): 183-190. doi:10.1007/s40142-017-0129-2.

Sicker D, Frey M, Schulz M, Gierl A. (2000) Role of natural benzoxazinones in the survival strategy of plants. Int Rev Cytol. 198: 319-46. doi:10.1016/s0074-7696(00)98008-2.

Sieber AN, Longin CF, Leiser WL, Würschum T. (2016) Copy number variation of CBF-A14 at the Fr-A2 locus determines frost tolerance in winter durum wheat. Theor Appl Genet. 129(6): 1087-97. doi:10.1007/s00122-016-2685-3.

Singh AK, Olsen MF, Lavik LAS, Vold T, Drabløs F et al. (2021) Detecting copy number variation in next generation sequencing data from diagnostic gene panels. BMC Med Genomics. 14(1): 214. doi:10.1186/s12920-021-01059-x.

Singleton AB, Farrer M, Johnson J, Singleton A, Hague S et al. (2003) alpha-Synuclein locus triplication causes Parkinson's disease. Science. 302(5646): 841. doi:10.1126/science.1090278.

Sohrabi R, Huh JH, Badieyan S, Rakotondraibe LH, Kliebenstein DJ et al. (2015) In planta variation of volatile biosynthesis: an alternative biosynthetic route to the formation of the pathogen-induced volatile homoterpene DMNT via triterpene degradation in Arabidopsis roots. Plant Cell. 27(3): 874-90. doi:10.1105/tpc.114.132209.

Springer NM, Ying K, Fu Y, Ji T, Yeh CT et al. (2009) Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. PLoS Genet. 5(11): e1000734. doi:10.1371/journal.pgen.1000734.

Sue M, Nakamura C, Nomura T. (2011) Dispersed benzoxazinone gene cluster: molecular characterization and chromosomal localization of glucosyltransferase and glucosidase genes in wheat and rye. Plant Physiol. 157(3): 985-97. doi:10.1104/pp.111.182378.

Sun X, Jiao C, Schwaninger H, Chao CT, Ma Y et al. (2020) Phased diploid genome assemblies and pan-genomes provide insights into the genetic history of apple domestication. Nat Genet. 52(12): 1423-1432. doi:10.1038/s41588-020-00723-9.

Sutton T, Baumann U, Hayes J, Collins NC, Shi BJ et al. (2007) Boron-toxicity tolerance in barley arising from efflux transporter amplification. Science. 318(5855): 1446-9. doi:10.1126/science.1146853.

Swanson-Wagner RA, Eichten SR, Kumari S, Tiffin P, Stein JC et al. (2010) Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. Genome Res. 20(12): 1689-99. doi:10.1101/gr.109165.110.

Takos AM, Rook F. (2012) Why biosynthetic genes for chemical defense compounds cluster. Trends Plant Sci. 17(7): 383-8. doi:10.1016/j.tplants.2012.04.004.

Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D et al. (2005) Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". Proc Natl Acad Sci U S A. 102(39): 13950-5. doi:10.1073/pnas.0506758102.

Tomato Genome Consortium. (2012) The tomato genome sequence provides insights into fleshy fruit evolution. Nature. 485(7400): 635-41. doi:10.1038/nature11119.

Twaij BM, Hasan MN. (2022) Bioactive Secondary Metabolites from Plant Sources: Types, Synthesis, and Their Therapeutic Uses. Int. J. Plant Biol. 13(1): 4-14. doi:10.3390/ijpb13010003.

Valliyodan B, Brown AV, Wang J, Patil G, Liu Y et al. (2021) Genetic variation among 481 diverse soybean accessions, inferred from genomic re-sequencing. Sci Data. 8(1): 50. doi:10.1038/s41597-021-00834-w.

Vogels A, Fryns JP. (2002) The Prader-Willi syndrome and the Angelman syndrome. Genet Couns. 13(4): 385-96. PMID: 12558108.

Vollger MR, Guitart X, Dishuck PC, Mercuri L, Harvey WT et al. (2022) Segmental duplications and their variation in a complete human genome. Science. 376(6588): eabj6965. doi:10.1126/science.abj6965.

von Rad U, Hüttl R, Lottspeich F, Gierl A, Frey M. (2001) Two glucosyltransferases are involved in detoxification of benzoxazinoids in maize. Plant J. 28(6): 633-42. doi:10.1046/j.1365-313x.2001.01161.x.

Walkowiak S, Gao L, Monat C, Haberer G, Kassa MT et al. (2020) Multiple wheat genomes reveal global variation in modern breeding. Nature. 588(7837): 277-283. doi:10.1038/s41586-020-2961-x.

Walsh T, McClellan JM, McCarthy SE, Addington AM, Pierce SB et al. (2008) Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. Science. 320(5875): 539-43. doi:10.1126/science.1155174.

Wang W, Mauleon R, Hu Z, Chebotarov D, Tai S et al. (2018) Genomic variation in 3,010 diverse accessions of Asian cultivated rice. Nature. 557(7703): 43-49. doi:10.1038/s41586-018-0063-9.

Wegel E, Koumproglou R, Shaw P, Osbourn A. (2009) Cell type-specific chromatin decondensation of a metabolic gene cluster in oats. Plant Cell. 21(12): 3926-36. doi:10.1105/tpc.109.072124.

Williams TN, Wambua S, Uyoga S, Macharia A, Mwacharo JK et al. (2005) Both heterozygous and homozygous alpha+ thalassemias protect against severe and fatal Plasmodium falciparum malaria on the coast of Kenya. Blood. 106(1): 368-71. doi:10.1182/blood-2005-01-0313.

Winzer T, Gazda V, He Z, Kaminski F, Kern M et al. (2012) A Papaver somniferum 10-gene cluster for synthesis of the anticancer alkaloid noscapine. Science. 336(6089): 1704-8. doi:10.1126/science.1220757.

Wlodzimierz P, Rabanal FA, Burns R, Naish M, Primetis E et al. (2023) Cycles of satellite and transposon evolution in Arabidopsis centromeres. Nature. 618(7965): 557-565. doi:10.1038/s41586-023-06062-z.

Xu X, Liu X, Ge S, Jensen JD, Hu F et al. (2011) Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. Nat Biotechnol. 30(1): 105-11. doi:10.1038/nbt.2050.

Xu J, Zhang W, Zhang P, Sun W, Han Y et al. (2023) A comprehensive analysis of copy number variations in diverse apple populations. BMC genomics. 24(1): 256. doi:10.1186/s12864-023-09347-9.

Yang N, Liu J, Gao Q, Gui S, Chen L et al. (2019) Genome assembly of a tropical maize inbred line provides insights into structural variation and crop improvement. Nat Genet. 51(6): 1052-1059. doi:10.1038/s41588-019-0427-6.

Yeaman S, Whitlock MC. (2011) The genetic architecture of adaptation under migration-selection balance. Evolution. 65(7): 1897-911. doi:10.1111/j.1558-5646.2011.01269.x.

Yoon S, Xuan Z, Makarov V, Ye K, Sebat J. (2009) Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Res 19(9): 1586–1592. doi:10.1101/gr.092981.109.

Yu N, Nützmann HW, MacDonald JT, Moore B, Field B et al. (2016) Delineation of metabolic gene clusters in plant genomes by chromatin signatures. Nucleic Acids Res. 44(5): 2255-65. doi:10.1093/nar/gkw100.

Yu P, Wang C, Xu Q, Feng Y, Yuan X et al. (2011) Detection of copy number variations in rice using array-based comparative genomic hybridization. BMC Genomics. 12:372. doi:10.1186/1471-2164-12-372.

Yuan Y, Bayer PE, Batley J, Edwards D. (2021) Current status of structural variation studies in plants. Plant Biotechnol J. 19(11): 2153-2163. doi:10.1111/pbi.13646.

Zhao G, Lian Q, Zhang Z, Fu Q, He Y et al. (2019) A comprehensive genome variation map of melon identifies multiple domestication events and loci influencing agronomic traits. Nat Genet. 51(11): 1607-1615. doi:10.1038/s41588-019-0522-8.

Zhao M, Wang Q, Wang Q, Jia P, Zhao Z. (2013) Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. BMC Bioinformatics. 14 (Suppl 11): S1. doi:10.1186/1471-2105-14-S11-S1.

Zhou P, Silverstein KA, Ramaraj T, Guhlin J, Denny R et al. (2017) Exploring structural variation and gene family architecture with De Novo assemblies of 15 Medicago genomes. BMC Genomics. 18(1): 261. doi:10.1186/s12864-017-3654-1.

Zmienko A, Marszalek-Zenczak M, Wojciechowski P, Samelak-Czajka A, Luczak M et al. (2020) AthCNV: A Map of DNA Copy Number Variations in the Arabidopsis Genome. Plant Cell. 32(6): 1797-1819. doi:10.1105/tpc.19.00640.

Żmieńko A, Samelak A, Kozłowski P, Figlerowicz M. (2014) Copy number polymorphism in plant genomes. Theor Appl Genet. 127(1): 1-18. doi:10.1007/s00122-013-2177-7.

Rozdział 6

Publikacje wchodzące w skład rozprawy doktorskiej

Publikacja 1

Samelak-Czajka A, **Marszalek-Zenczak M**, Marcinkowska-Swojak M, Kozlowski P, Figlerowicz M and Zmienko A (2017)

MLPA-Based Analysis of Copy Number Variation in Plant Populations

Front. Plant Sci.8:222. doi: 10.3389/fpls.2017.00222

5-letni IF = 4,353

frontiers in Plant Science

PROTOCOLS
published: 21 February 2017
doi: 10.3389/fpls.2017.00222

MLPA-Based Analysis of Copy Number Variation in Plant Populations

*Anna Samelak-Czajka[1], Malgorzata Marszalek-Zenczak[2], Malgorzata Marcinkowska-Swojak[3], Piotr Kozlowski[3], Marek Figlerowicz[1,2] and Agnieszka Zmienko[1,2]**

[1] Institute of Computing Science, Faculty of Computing, Poznan University of Technology, Poznan, Poland, [2] Department of Molecular and Systems Biology, Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan, Poland, [3] Department of Molecular Genetics, Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan, Poland

Copy number variants (CNVs) are intraspecies duplications/deletions of large DNA segments (>1 kb). A growing number of reports highlight the functional and evolutionary impact of CNV in plants, increasing the need for appropriate tools that enable locus-specific CNV genotyping on a population scale. Multiplex ligation-dependent probe amplification (MLPA) is considered a gold standard in genotyping CNV in humans. Consequently, numerous commercial MLPA assays for CNV-related human diseases have been created. We routinely genotype complex multiallelic CNVs in human and plant genomes using the modified MLPA procedure based on fully synthesized oligonucleotide probes (90–200 nt), which greatly simplifies the design process and allows for the development of custom assays. Here, we present a step-by-step protocol for gene-specific MLPA probe design, multiplexed assay setup and data analysis in a copy number genotyping experiment in plants. As a case study, we present the results of a custom assay designed to genotype the copy number status of 12 protein coding genes in a population of 80 *Arabidopsis* accessions. The genes were pre-selected based on whole genome sequencing data and are localized in the genomic regions that display different levels of population-scale variation (non-variable, biallelic, or multiallelic, as well as CNVs overlapping whole genes or their fragments). The presented approach is suitable for population-scale validation of the CNV regions inferred from whole genome sequencing data analysis and for focused analysis of selected genes of interest. It can also be very easily adopted for any plant species, following optimization of the template amount and design of the appropriate control probes, according to the general guidelines presented in this paper.

Keywords: structural variation, MLPA, 1001 Arabidopsis Genomes project, CNV genotyping, multiplexing

OPEN ACCESS

Edited by:
Scott V. Edwards,
Harvard University, USA

Reviewed by:
Nathan Lewis Clark,
University of Pittsburgh, USA
John Malone,
University of Connecticut, USA
Pierre Baduel,
John Innes Centre (BBSRC), USA

****Correspondence:***
Agnieszka Zmienko
akisiel@ibch.poznan.pl

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

Received: *05 October 2016*
Accepted: *06 February 2017*
Published: *21 February 2017*

Citation:
Samelak-Czajka A, Marszalek-Zenczak M, Marcinkowska-Swojak M, Kozlowski P, Figlerowicz M and Zmienko A (2017) MLPA-Based Analysis of Copy Number Variation in Plant Populations. Front. Plant Sci. 8:222. doi: 10.3389/fpls.2017.00222

INTRODUCTION

The rise of high-throughput genomics techniques – DNA arrays and, more recently, whole-genome sequencing (WGS) – has revealed the structural complexity and dynamics of eukaryotic genomes. In particular, the ability to re-sequence and compare hundreds or even thousands of genomes of individuals within one species has paved the way for the investigation of the extent to which individual genomes differ from each other. One type of structural variation that is ubiquitous in the genomes of humans, animals and plants is copy number variation (CNV). This term refers to intraspecies duplications and deletions of large DNA segments, usually >1 kb [although variants >50 bp have been recently included in this spectrum (Alkan et al., 2011)]. The human genome is the most intensively studied eukaryotic genome in terms of the distribution and functional significance of CNVs and the mechanisms leading to the formation of copy number rearrangements (Zarrei et al., 2015). However, the number of species for which CNV regions have been inferred on the genome-wide scale is growing rapidly. For plants, this list includes maize, rice, sorghum, *Arabidopsis* (*Arabidopsis thaliana)*, soybean, wheat, and barley (Springer et al., 2009; Beló et al., 2010; Swanson-Wagner et al., 2010; Cao et al., 2011; Saintenac et al., 2011; Zheng et al., 2011; McHale et al., 2012; Muñoz-Amatriaín et al., 2013; Duitama et al., 2015; Bai et al., 2016). As in humans, CNV regions in plants are not uniformly distributed across the chromosomes. Although they are more common in the intergenic regions, they also co-localize with hundreds of protein-coding genes (Swanson-Wagner et al., 2010; Beló et al., 2010; McHale et al., 2012; Muñoz-Amatriaín et al., 2013). The ability to alter the gene structure and copy number makes CNV an important factor that influences gene expression (Żmieńko et al., 2014). By the gene dosage effect, CNVs can also affect the interaction of the genes' products within protein and metabolic networks (Hanada et al., 2011; Conant et al., 2014). Quite often, such variation accounts for adaptive traits or - as shown for humans - can underlie disease (Stankiewicz and Lupski, 2010; Zarrei et al., 2015). In plants, a growing number of studies highlight the shaping role of CNVs in genome evolution, phenotypic variation and – sometimes rapid - adaptation to environmental challenges (Gaines et al., 2010; Cook et al., 2012; Maron et al., 2013; Chang et al., 2015; Wang et al., 2015). Therefore, it is anticipated that the number of genetic studies focused on individual CNVs of interest will grow and that new CNV-associated traits will be revealed.

In-depth analysis of individual CNVs in plants has rarely been conducted (Gaines et al., 2010; Cook et al., 2012; Maron et al., 2013). Likewise, in plants for which the CNV regions were inferred from WGS data, the subsequent validation was not conducted or was limited to the PCR-based detection of CNV deletions (Swanson-Wagner et al., 2010; Cao et al., 2011; Tan et al., 2012; Bai et al., 2016). Therefore, there is an urgent need to widen the range of experimental studies of CNV in plants to contribute to the creation of high-confidence CNV maps and enhance association studies linking CNVs with phenotypic traits in plant species. In this context, the lack of validated experimental approaches for the analysis of individual CNVs in plants is apparent, as opposed to the well-established methods and standardized protocols available for the human genome.

The range of popular molecular methods used for DNA copy number genotyping in humans is wide (Ceulemans et al., 2012; Cantsilieris et al., 2013; Bharuthram et al., 2014). Among them, multiplex ligation-dependent probe amplification (MLPA), first introduced in 2002 (Schouten et al., 2002) and later developed by the MRC Holland company, is considered a gold standard in the diagnosis of numerous DNA copy number-related human diseases (Hömig-Hölzel and Savola, 2012). MLPA is a simple and robust method of relative quantification of DNA sequences on a population scale. The standard multiplex assay utilizes up to 50 probes targeting specific DNA regions (e.g., exons in a gene of interest). Each probe is composed of two half-probes (physically separate DNA fragments, one fully synthetic and one clone-derived) that match the target sequence in directly adjacent positions with their target-specific sequences (TSSs). Successful hybridization of both half-probes to the genomic DNA enables their ligation and linear amplification. The amplification products are then analyzed by capillary electrophoresis. Relative quantification of the signal peaks from fragments of unique size, generated by individual probes in the assay, provides information about the template DNA copy number. MLPA requires little genomic DNA input (Schouten et al., 2002). Additionally, the genomic sequence targeted by the probes is quite short (50–70 nt), which enables use of MLPA for the analysis of regions too small to be detected by the FISH method. MLPA has been shown to be superior to qPCR for gene copy number quantification (Perne et al., 2009; Cantsilieris et al., 2014). Additionally, it presents similar performance to droplet digital PCR in accurate quantification of up to eight gene copies, making it suitable for the analysis of multiallelic CNVs, i.e., those that exist in more than two genotypes in a population (Zmienko et al., 2016).

According to PubMed, the seminal MLPA work (Schouten et al., 2002) has been cited almost 450 times (∼220 times within 5 last years). Additionally, ∼2,000 articles in PubMed matched the search keyword "Multiplex Ligation-Dependent Probe Amplification". Among these papers, only 16 also matched the search keyword "plant". Those that actually described plant applications of MLPA involved alternative applications of this method: the detection of genetically modified organisms (GMO-MLPA) (Rudi et al., 2003), single nucleotide polymorphism (SNP) genotyping (Thumma et al., 2009), or gene expression analysis (RT-MLPA) (Li et al., 2009, 2011, 2013). However, none of these papers presented a primary MLPA application of copy number analysis. Several reasons might account for the fact that the MLPA approach has not been adopted by the plant community. One is much later recognition of the intraspecies variation and CNV prevalence in the plant genomes than in humans. Additionally, the commercial MLPA assays are focused on biomedical studies and cover only humans. Therefore, to assess plant genome variation with MLPA, it is necessary to self-design synthetic probes. It should be noted that, over the years, numerous modifications of the MLPA strategy have been introduced that simplify the probe design procedure (Marcinkowska et al., 2010; Ling et al., 2015, and references therein). In the current work, we present the optimized protocol for MLPA-based CNV analysis and provide guidelines for designing and performing MLPA assays in plants. The protocol is based on the MLPA adaptation developed previously by one of us (PK) that involves fully synthetic oligonucleotide probes, 90 to 200 nt in length, and allows for simultaneous genotyping of >30 different positions in the genomic DNA (Kozlowski et al., 2007). The protocol combines MLPA probe design, synthesis, experimental procedures, data preprocessing and analysis stages into one comprehensive procedure. The lack of MLPA-based genotyping studies in plants highlights the need for such an integrated resource. We also provided the probe design template, developed specifically for the presented MLPA variant. It allows for semi-automatic probe sequence setup, clarifies the idea of probe set composition and shortens the design process by days.

High and low copy level duplications may have different effects on the gene dosage and the phenotype, e.g., by triggering differences in gene expression level or inducing the silencing mechanisms in plants. Therefore, an important aspect of plant CNV genotyping studies is to estimate the actual gene copy numbers in the analyzed lines in order to analyze their influence on the trait of interest (Cook et al., 2014). To illustrate the performance of the MLPA method for precise DNA copy number genotyping in plant populations, we present exemplar assays for 12 genes with different levels of copy number diversity in a population of 80 *Arabidopsis* ecotypes, including multiallelic CNVs. We also describe the set of experimentally verified normalization control probes and the results of genomic DNA template amount optimization performed for this model species.

An advantage of the presented approach is that the assay - after it has been standardized for the particular organism – is always performed in the same conditions, regardless of the probe set composition. It may be utilized for the detailed analysis of a genomic region of interest using a set of MLPA probes scattered along this region or for large-scale validation/genotyping studies of WGS-based predicted CNVs, with 1-2 MLPA probes per inferred CNV.

MATERIALS AND EQUIPMENT

Materials

(1) High-quality genomic DNA for each analyzed sample, evaluated using a NanoDrop 2000 spectrophotometer (Thermo Scientific) and with standard gel electrophoresis; the working concentration is typically 0.4 to 50 ng/μl, depending on the species (see the following sections).
 For *Arabidopsis*: We successfully genotyped CNVs using genomic DNA from 3-week-old rosette leaves extracted with a DNeasy Plant Mini Kit (Qiagen).
(2) Self-designed synthetic oligonucleotides (MLPA half-probes; see the following section for the probe design instructions) purchased from Integrated DNA Technologies (or similar provider) as 100 nmol oligo, purified by HPLC (for oligonucleotides up to 100 nt in length) or PAGE (for oligonucleotides over 100 nt in length); the right half-probes should be additionally modified by 5′ phosphorylation.
(3) Nuclease-free water (not DEPC-treated) (Ambion, cat. no. AM9938)
(4) SALSA MLPA EK-1 reagent kit (MRC-Holland, cat. no. EK1-FAM), which includes the following components:
 SALSA MLPA Buffer
 SALSA Ligase-65
 Ligase Buffer A
 Ligase Buffer B
 SALSA PCR Primer MIX
 SALSA Polymerase
(5) Consumables for capillary electrophoresis, depending on the instrument type; here, for the ABI Prism 3130XL Genetic Analyzer:
 HiDi formamide (Thermo Fisher Scientific, cat. no. 4440753)
 GeneScan 600 LIZ Size Standard (Thermo Fisher Scientific, cat. no 4366589)
 POP7 Polymer (Thermo Fisher Scientific, cat. no 4352759).

Equipment

(1) 0.2 ml PCR strips and suitable caps, e.g., 8-Strip PCR tubes (Starlab, cat. no. I1402-3500) and 8-Strip caps (Starlab, cat. no. I1400-0800).
(2) Standard and multichannel pipettes.
(3) Thermocycler with heated lid (e.g., Bio-Rad T100 Thermal Cycler or equivalent).
(4) Vortex mixer (e.g., ELMI V-3 Sky Line or equivalent).
(5) Mini laboratory centrifuge with Eppendorf tube adapter and PCR strip adapter (e.g., Labnet Spectrafuge or equivalent).
(6) Capillary electrophoresis instrument (AppliedBiosystems ABI Prism 3130XL Genetic Analyzer or equivalent) or access to a capillary electrophoresis service provider.
(7) Software tool for the extraction of the intensity data after size-separation of MLPA reaction products (e.g., GeneMarker by SoftGenetics).

STEPWISE PROCEDURES

The general concept of the MLPA strategy is presented in **Figure 1**. The entire procedure involves three main stages: (A) designing the MLPA probes; (B) performing MLPA assay, which involves half-probes hybridization to DNA template, subsequent ligation and amplification; and (C) data collection and analysis, including the estimation of the copy number genotypes.

Stage A: Design the MLPA Probes (Time: Approximately 1 Week + Oligonucleotide Synthesis and Transportation by an External Provider)

The presented MLPA procedure based on fully synthetic oligonucleotide probes allows for simultaneous copy number analysis of ∼30 individual regions in the genomic DNA. Of these, at least 3 to 5 MLPA probes should target the confirmed non-variable control regions, distant from the studied genomic positions. These probes serve as normalization controls in the subsequent analysis of the MLPA data to account for the possible variation of the input DNA template amount and technical issues. The typical targets of the MLPA assays are protein-coding genes, as the changes in their copy number potentially affect the protein level and may contribute to the phenotype. The number of probes designed for each gene and their density in the covered genomic region depend on the user’s requirements.

The procedure for individual MLPA probe design has been graphically presented in Supplementary Figure S1 and is

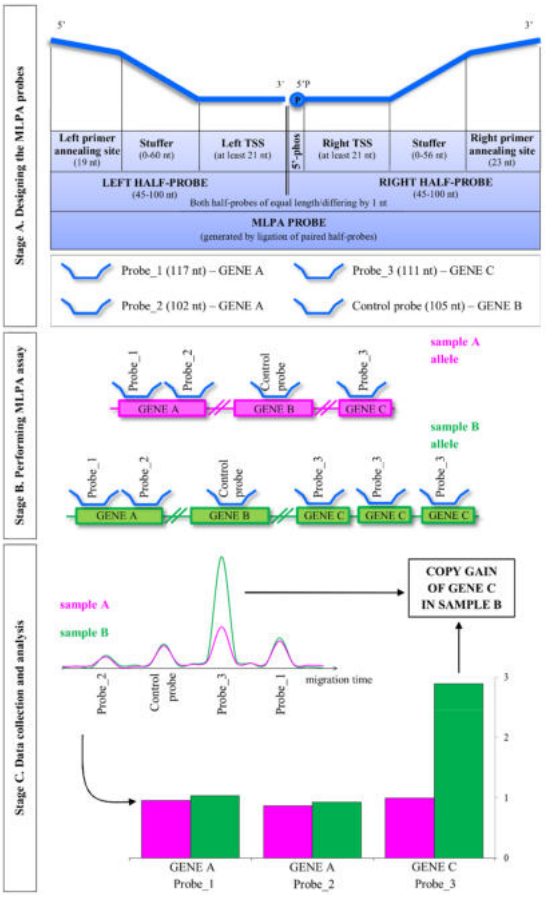

Stage A. Designing the MLPA probes
5'
3'
3'
5'P
P
Left primer annealing site (19 nt)
Stuffer (0-60 nt)
Left TSS (at least 21 nt)
5'-phos
Right TSS (at least 21 nt)
Stuffer (0-56 nt)
Right primer annealing site (23 nt)
LEFT HALF-PROBE (45-100 nt)
RIGHT HALF-PROBE (45-100 nt)
Both half-probes of equal length/differing by 1 nt
MLPA PROBE
(generated by ligation of paired half-probes)
Probe_1 (117 nt) – GENE A
Probe_3 (111 nt) – GENE C
Probe_2 (102 nt) – GENE A
Control probe (105 nt) – GENE B
Stage B. Performing MLPA assay
sample A allele
Probe_1
Probe_2
Control probe
Probe_3
GENE A
GENE B
GENE C
sample B allele
Probe_1
Probe_2
Control probe
Probe_3
Probe_3
Probe_3
GENE A
GENE B
GENE C
GENE C
GENE C
Stage C. Data collection and analysis
sample A
sample B
COPY GAIN OF GENE C IN SAMPLE B
Probe_2
Control probe
Probe_3
Probe_1
migration time
3
2
1
0
GENE A Probe_1
GENE A Probe_2
GENE C Probe_3

FIGURE 1 | Overview of the multiplex ligation-dependent probe amplification (MLPA) method. MLPA is comprised of three main stages: designing the probes, performing the multiplex MLPA assay and data collection and analysis. The three stages are described in the detail in the main text. TSS, target-specific sequence; 5′-phos: phosphorylation at the 5′ end of the oligonucleotide. Note that *Arabidopsis*, as a self-pollinating plant, typically carries pairs of identical alleles. For simplicity, single alleles are depicted.

described in detail in the following sections. We used *Arabidopsis* gene *AT1G01040* encoding Dicer-like 1 protein as an example.

Select TSSs for the MLPA Probes

Step 1. Retrieve the genomic sequence of the gene of interest from the appropriate database, including the exon-intron positions. We recommend localizing the MLPA probes within the exon sequences because they display lower variation than the non-coding regions of genes.

For *Arabidopsis:* Use the gene locus identifier (e.g., *AT1G01040*) to localize that gene in the TAIR10 genomic sequence, available through the *Arabidopsis* genome browser[1], and display its splice variants, when applicable (Protein Coding Gene Models track). In *Arabidopsis*, protein coding genes have five exons on average, each with mean length of ∼240 bp (Koralewski and Krutovsky, 2011). This length is sufficient for selecting two adjacent TSSs (one for each half-probe). Use the GBrowse navigation tools to zoom in to the selected exon and export its DNA sequence as a FASTA file.

Step 2. Ensure your sequence does not include any repetitive elements.

For *Arabidopsis*, rice, maize, wheat, and some other crops: Submit the extracted sequence to the CENSOR software tool (Kohany et al., 2006) that masks the repetitive elements in the query sequence using the collection of repeats for selected animal and plant species. Select a fragment of at least 100 nt that is not interrupted by any masked regions.

Step 3. If possible, check the selected sequence for the presence of SNPs and small indels.

For *Arabidopsis*: Use the 1001 Genomes Project VCF Subset tool[2] to download the subset of VCF files that contain full-genome VCF data for 1135 accessions (as of September 2016) (1001 Genomes Consortium, 2016). Download SNP information for the region and accessions of interest. Evaluate whether the selected sequence is free of common polymorphisms.

Step 4. From the selected region, choose two directly adjacent fragments of at least 21 nt (left and right TSS) and adjust their length and position so that the melting temperature (Tm) of each fragment will be as close as possible to 71°C (calculated with the free RaW program available from MRC Holland[3] with the following settings: method Go-Oli-Go, salt concentration 0.1 M, oligo concentration 1 µm). Avoid long homopolymer tracts and GC tracts of ≥4 bases.

Step 5. Join the adjacent left and right TSSs and use the resulting sequence in a homology search against the genomic sequence of the analyzed species to check for its specificity.

For *Arabidopsis*: Perform a BLAST search against *A. thaliana* NCBI reference genome with the following parameters: blastn algorithm, word size 7, match/mismatch scores 2;-3, gap costs 5;2, no sequence masking and filtering, *E*-value threshold 0.001.

Step 6. Repeat steps 3 to 5 until the pair of adjacent TSSs that satisfies all design criteria is found for a given gene.

Design the Half-Probes

Step 7. Add the respective PCR primer annealing sequence to each TSS and – optionally – the stuffer sequence, in the following order (see **Figure 1**):

for the left half-probe:

5′-left primer annealing sequence – stuffer – left TSS -3′,

where the left primer annealing sequence is GGGTTCCCTAAGGGTTGGA;

for the right half-probe:

5′-right TSS – stuffer – right primer annealing sequence – 3′,

where the right primer annealing sequence is TCTAGATTGGATCTTGCTGGCGC.

For the stuffer, use the fragment of enterobacteria phage M13 sequence (NCBI/GenBank ID V00604, range: 3-119). This

[1] https://gbrowse.arabidopsis.org/cgi-bin/gb2/gbrowse/arabidopsis/

[2] http://tools.1001genomes.org

[3] http://www.mrc-holland.com/

fragment has no significant blastn matches to any eukaryotic genomic sequence deposited in the NCBI/RefSeq Representative Genome Database (accessed July 4th, 2016). It has been successfully applied as a stuffer in our previous MLPA assays performed for *Arabidopsis* and human DNA (Marcinkowska-Swojak et al., 2014; Klonowska et al., 2015; Zmienko et al., 2016).

Note: The addition of the optional stuffer sequence allows the user to adjust the length of the half-probes so that the resulting PCR amplification fragments would be of unique size and differ by 3 nt for probes in the 90-120 nt range and by 4 nt for probes >120 nt long. The length of the two half-probes in the pair should be the same or differ by 1 nt. For example, to obtain the MLPA probe of length 120, the left and right half-probe sequences should each be 60 nt long (and at least 21 nt of each half-probe should constitute TSS).

To facilitate the process of MLPA probe design and combining multiple MLPA probes in one experimental assay, we provided a Microsoft Excel template (Supplementary Table S1). This template includes the formulas that automatically adjust the length of the stuffer sequence and add the required adapter sequences to both the left and right half-probes. As a result, the final sequence of the MLPA probe of the desired length is returned. The user can choose the MLPA probe length. Typically, when fewer than the maximal number of MLPA probes are included in the assay, we recommend designing shorter probes to minimize the oligonucleotide synthesis costs. Often, the MLPA assays contain two or more probes targeting adjacent genomic regions. We recommend randomization of these probe MLPA lengths to minimize the influence of the possible biases or artifacts. Likewise, we recommend distributing the control probe lengths to cover the entire range of the MLPA probes in the assay.

For *Arabidopsis*: We provide pre-designed sequences for five control MLPA probes (ctrl1–ctrl5) that target genes located on chromosomes 1, 2, 4, and 5. The first gene is *DCL1*, coding for a RNA helicase involved in microRNA processing. The second gene encodes an oxidoreductase belonging to a zinc-binding dehydrogenase family protein. The third non-variable gene is *APG10*, coding for a BBMII isomerase involved in histidine biosynthesis. The fourth gene is *PDF5*, coding for a prefoldin, involved in unfolded protein binding. The fifth gene is *PS2*, coding for a pyrophosphate-specific phosphatase. The lengths of the probes cover the entire range of the MLPA assay (Supplementary Table S1). The regions were selected as not copy-number variable in *Arabidopsis* based on WGS data and were experimentally validated in 189 natural accessions (Zmienko et al., 2016).

Order the Oligonucleotide Synthesis

The synthesis of the designed MLPA probes is typically performed by an external service provider, such as Integrated DNA Technologies (IDT).

Step 8. Order the synthesis of left and right half-probes, each as separate oligonucleotides, at a 100-nmol scale. All right half-probes must be additionally modified at their 5′ ends (5′ phosphorylation).

Caution: 5′ phosphorylation of the right half-probes is essential for a successful ligation step (described below). The oligonucleotides designed for MLPA assays should be of high purity; therefore, we recommend selecting a PAGE or HPLC purification option, depending on the oligonucleotide length and according to the oligonucleotide manufacturer's recommendations.

Step 9. Re-dissolve the lyophilized oligonucleotides upon arrival in deionized water to a concentration of 20 μM. Alternatively, the oligonucleotides can be re-dissolved in 10 mM Tris-HCl, pH 8.2.

Step 10. Store the half-probe stocks at –20°C.

Stage B. Perform MLPA Assay (Time: 2 Days)

Note: When performing the MLPA assay, keep all reagents, stock solutions and working solutions on ice. Set up the reactions in PCR tubes or strips (recommended) at room temperature, unless indicated otherwise. Depending on the user's experience, we recommend running assays for 8–32 samples at once in 1–4 PCR strips.

Note: Whenever applicable, prepare the reagent master mixes for all assayed samples with 10% volume surplus to minimize sample-to-sample variation and save pipetting time. Distribute the master mix to eight tubes of a new PCR strip and then transfer the required amount to all PCR strips containing your samples with a multichannel pipette.

Note: Perform all incubation steps in a thermocycler, programmed as specified in **Table 1**.

Caution: Do not vortex the tubes containing Ligase-65 or Salsa Polymerase enzymes. Likewise, do not vortex the master mixes after adding any of these enzymes.

Prepare the MLPA Probe Set Mix

The correctly composed assay should include both half-probes (left and right) for each region of interest. Each pair of half-probes should generate a ligation product of unique length in the assay. The concentration of the MLPA probes in the final reaction mixture is very low (see below); therefore, it is convenient to perform a two-step oligonucleotide dilution during the probe set mix preparation as follows.

Step 1. Melt all half-probe stocks constituting one assay.

Step 2. Dilute each 20 μM stock with water to a 0.2 μM working solution (200 μl).

Step 3. Mix 2 μl of each half-probe working solution and fill to 400 μl with water.

The resulting 1 nM MLPA Probe Set Mix will contain all the desired pairs of half-probes in equal concentrations and is directly applicable in the reaction setup.

Note: MLPA Probe Set Mix can be stored at –20°C until later use.

Hybridize Half-Probes

For each genomic DNA sample, perform the MLPA assay in a separate tube. We recommend running MLPA assays in multiples of 8 in PCR strips with caps.

TABLE 1 | Programmed thermocycler conditions for multiplex ligation-dependent probe amplification (MLPA) assay.

Program		Action
Denaturation (Step 5)		
98°C, 5 min;		Denature samples.
25°C, ∞;		Cool down samples before removing.
Pause		Proceed to Step 6.
Hybridization (Steps 9-10)		
95°C, 1 min;		Hybridize half-probes to their genomic targets.
60°C, 16–20 h;		
54°C, ∞;		Adjust the temperature for the next step.
Pause		Proceed to Step 11.
Ligation (Step 14)		
54°C, 15 min;		Ligate adjacently hybridized half-probes.
98°C, 5 min;		Inactivate the enzyme.
20°C, ∞;		Cool down samples before removing.
Pause		Proceed to Step 15.
Amplification (Step 18)		

35 cycles of:	95°C, 30 s;
	60°C, 30 s;
	72°C, 1 min;
		Amplify the correctly ligated MLPA probes.

72°C, 20 min;		Perform final extension of PCR products.
4°C, ∞;		Cool down samples before removing.
End		Proceed to Step 19.

Caution: Replace the strip caps with new ones at each opening during the entire procedure to prevent cross-contamination.

Step 4. Aliquot 5 μl of genomic DNA (0.4 to 50 ng/μl) to individual strip tubes to obtain a final template amount of 2–250 ng per assay, depending on the species.

Note: We recommend performing template optimization assays for each species.

For *Arabidopsis:* We successfully performed MLPA assays using 2, 5, 10, 15, 30, 60, and 100 ng genomic DNA per assay (see the next section).

Step 5. Insert the samples into the thermocycler. Heat for 5 mins at 98°C then let the samples cool to 25°C.

Step 6. Remove the samples from the thermocycler and centrifuge.

Step 7. Prepare master mix I. Briefly vortex and centrifuge the SALSA MLPA buffer and MLPA Probe Set Mix. Prepare the adequate amount of the master mix I by mixing 1.5 μl of SALSA MLPA buffer and 1.5 μl of 1 nM MLPA Probe Set Mix per sample, with 10% volume surplus. Vortex and centrifuge the tube.

Step 8. Add 3 μl of the master mix I to each denatured DNA sample and mix briefly by pipetting. Close the strips with the new caps and centrifuge. The reaction volume in each tube should be 8 μl.

Step 9. Put the samples back into the thermocycler and incubate for 1 min at 95°C, then for 16 to 18 h at 60°C.

Step 10. Adjust the thermoblock temperature to 54°C before proceeding to the next step.

Caution: Do NOT remove the samples from the thermocycler!

Ligate the Hybridized Half-Probes

Step 11. Prepare master mix II without enzyme. Briefly vortex and centrifuge Ligase Buffer A and Ligase Buffer B. Mix 3 μl of Ligase Buffer A, 3 μl of Ligase Buffer B, and 25 μl of nuclease-free water per sample, with 10% volume surplus. Vortex and centrifuge the tube.

Step 12. Centrifuge the tube containing SALSA Ligase-65 enzyme. Add 1 μl of the enzyme per sample with 10% volume surplus to the master mix II. Mix briefly by pipetting. Centrifuge the tube and store on ice until use. Proceed to the next step without delay.

Step 13. Without removing the strips from the thermocycler, add 32 μl of master mix II to each sample. Mix by pipetting and close the strips with new caps. The reaction volume in each tube should be 40 μl.

Step 14. Incubate the samples for 15 min at 54°C, followed by heat inactivation of the ligase enzyme (5 min at 98°C). Cool the thermoblock to 20°C and remove the samples.

Amplify the Ligated MLPA Probes

Step 15. Prepare master mix III. Briefly vortex and centrifuge the SALSA PCR primer mix. Mix 2 μl of SALSA PCR primer mix and 7.5 μl of nuclease-free water per sample, with 10% volume surplus. Vortex and centrifuge the tube.

Step 16. Centrifuge the tube containing SALSA Polymerase enzyme. Heat the tube in hands for approximately 10 s, then add 0.5 μl of the enzyme per sample with 10% volume surplus to master mix III. Mix briefly by pipetting. Centrifuge the tube and store on ice until use.

Step 17. Add 10 μl of master mix III to each sample and mix by pipetting. Close the strips with new caps and replace in the thermocycler. The final reaction volume in each tube should be 50 μl.

Step 18. Perform the PCR comprising 35 cycles of: 95°C for 30 s; 60°C for 30 s and 72°C for 1 min, followed by a 20 min final elongation at 72°C. Cool the thermoblock to 4°C.

Step 19. Store the samples at 4°C, protected from light, until the product size-separation (1–3 days).

Stage C. Collect and Analyze the Data (Time: 1 Day for the Data Collection, Variable for the Analysis)

Size-Separate the PCR Products by Capillary Electrophoresis

The product separation should be performed under denaturing conditions on any standard capillary DNA analyzer. The specific run parameters must be adjusted according to the recommendations of the instrument manufacturer.

We typically use the services of the local Molecular Biology Techniques facility (at the Department of Biology of Adam Mickiewicz University, Poznan, Poland) and separate the samples in ABI Prism 3130XL Genetic Analyzer (Applied Biosystems), using the following procedure.

Step 1. Each MLPA reaction sample is diluted 20× with nuclease-free water, mixed with 9 μl of HiDi formamide (Thermo

Fisher Scientific) containing GeneScan 600 LIZ Size Standard (Thermo Fisher Scientific) and denatured.

Step 2. Samples are injected at 1.2 kV voltage and separated on ABI Prism 3130XL Genetic Analyzer (Applied Biosystems) at 15 kV, in POP7 separation matrix (Thermo Fisher Scientific).

Analyze the Electropherograms

Evaluate the data quality and extract the signal intensity from the electropherograms. Numerous software tools are appropriate for this purpose. Below, we describe the step-by-step analysis performed with GeneMarker (SoftGenetics) (Supplementary Figure S2).

Note: The GeneMarker functions used here are accessible in the limited demo version of the software, freely downloadable from the manufacturer's web site. The details regarding use of these functions are described in the software manual, also available for download.

Step 3. Load the electropherogram data to GeneMarker.

Step 4. Analyze the raw data files with the MLPA analysis type option and appropriate DNA standard selected (depending on the capillary electrophoresis conditions). Select the size call method and data normalization approach (Supplementary Figure S2A).

Note: GeneMarker software provides two normalization options (intra-sample "Internal Control Probe Normalization" and inter-sample "Population Normalization") that aim to correct for the variation in signal intensity caused by the differences in the lengths of the probes in the multiplex assay. We typically use the intra-sample normalization against our control probes, although at this step it is not critical, because the range of the probe lengths in our assay (96–200 nt) is much smaller than in the case of commercial MLPA assays (130–490 nt).

Caution: Use the same parameter settings for all samples. When applying internal control probe normalization, use the same set of control probes for analysis of all samples in the MLPA assay.

Note: At the first analysis of a new MLPA assay, run the analysis for a selection of samples using the "NONE" panel selection. This will allow you to manually create the custom MLPA panel later by indicating the peak positions in your pre-processed samples (see Step 5). If the MLPA panel has already been created, select that panel for the final analysis of all your samples.

Step 5. Perform this step for the new MLPA assay only. Manually create the probe panel with the Panel Editor (Supplementary Figure S2B). Use the pre-processed set of representative MLPA electropherograms (see Step 4) to locate and insert the alleles at the expected positions. Label the alleles with the MLPA probe names. If you want to use the "Internal Control Probe Normalization" option during the analysis, mark the control probes as 1. Repeat Step 4 to re-run all samples using the newly created panel.

Note: In our assays, all peak sizes consistently appeared ∼3 bp shorter than the theoretical length of their attributed MLPA probes. This is not an unexpected result because the migration times of the peak maxima depend on many factors, including the amount of the sample injected, the temperature and the dye used. The capillary electrophoresis systems estimate the relative allele size (using internal standard) and do not necessarily report the true fragment size (McCord, 2003). Therefore, the observed shift is specific to the system and MLPA assay conditions. As long as the peaks are consistently observed at the same positions in all samples under comparison, it does not influence the peak discrimination and subsequent analysis of the MLPA data.

Step 6. Evaluate the quality of individual electropherograms in accordance with the peak pattern of the size standard, the electrophoresis baseline, signal sloping and overall signal intensity. Samples that show abnormalities should be excluded from the analysis.

Step 7. Configure the report layout and copy the results to MS Excel or similar program for further analysis (Supplementary Figure S2C).

Note: The processed data can be reported as the fluorescence intensity (peak height) or the peak area values for each allele. The choice of the output typically does not affect the downstream data analysis and we obtained comparable results with both options. We preferably use the fluorescence intensity data.

Estimate the DNA Copy Number

Step 8. Use the normalization controls to perform within-sample normalization of all your sample data before comparison.

For Arabidopsis: Use at least 3 of the provided control probes (ctrl1–ctrl5) for normalization. Divide each intensity value by the average intensity of the control probes, separately for each sample.

Step 9. For each region analyzed, compare the normalized intensity between the samples. Cluster the samples with the similar intensities and infer the copy numbers from analysis of histograms or two-dimensional plots (see next section). Whenever possible, use the (set of) positive and negative control samples with known copy number status to determine the duplication/deletion intensity thresholds (see the next section for exemplar results).

ANTICIPATED RESULTS

Exemplar MLPA Assay

Based on the available WGS data from 1001 Arabidopsis Genomes Project (1001 Genomes Consortium, 2016) and our own analysis of a subset of this data including 80 accessions, originally described in (Cao et al., 2011), we selected 12 genes that overlapped CNVs with various levels of structural complexity. Genes *AT1G47670* and *AT1G80830* do not present copy number changes. Genes *AT1G32300* and *AT4G19520* are biallelic; more specifically, they display presence-absence variation. The remaining eight genes are multiallelic and present duplications (*AT4G27080*, *AT5G09590*, and *AT5G61700*) or duplications and deletions (*AT1G27570*, *AT1G52950*, *AT3G21960*, *AT4G27080*, and *AT5G54710*). Additionally, gene *AT5G09590* overlaps CNV only partially, whereas *AT1G52950*, *AT5G54710*, and *AT1G27570* are members of multigene families and are localized in the regions of high structural diversity (manifested e.g., by the presence of adjacent or overlapping CNVs, presence of nearby transposable element genes or the presence of clusters of highly similar paralogs). To present the performance of the MLPA approach we set up a multiplex assay Ath.test for these genes (**Table 2**). We evaluated the genes' copy number status in 80 *Arabidopsis* accessions, characterized in the first stage of 1001 *Arabidopsis* Genomes Project (Cao et al., 2011). All seeds were obtained from The European *Arabidopsis* Stock Centre[4] and grown as described previously (Zmienko et al., 2016).

Optimization of the Template Amount

The multiplex MLPA-based strategy presented in this paper was originally developed for CNV genotyping of human DNA (Kozlowski et al., 2007; Marcinkowska et al., 2010). To adjust it for use with the *Arabidopsis* genome, we aimed to optimize the amount of DNA template. For humans, the typical MLPA assays include 50-250 ng genomic DNA per reaction. In our previous study, we successfully performed MLPA-based copy number analysis using 100 ng *Arabidopsis* genomic DNA (Zmienko et al., 2016). However, because the *Arabidopsis* genome is ∼20 times smaller than the human genome, we expected that the template amount could be substantially reduced without affecting the reaction performance. To evaluate the acceptable range of DNA amount for this species, we used the Col-0 accession, performed serial dilutions of the DNA template and performed MLPA assays for each of the following DNA amounts: 100, 60, 30, 15, 10, 5, and 2 ng, in three replicates. We observed that the intensity data showed little variance across all DNA concentrations tested and the peaks showed very good resolution and similar distribution, regardless of the template amount (**Figures 2A–C**; Supplementary Data Sheet S1). The normalized signal intensity data for various template amounts were highly correlated, with the results calculated for 2 ng DNA input showing only slightly lowered correlation than the other amounts (**Figure 2D**). From this comparison, we concluded that the whole range of tested DNA amounts generates valid data. Below, we used the smallest tested amount of DNA (2 ng) to perform the exemplar Ath.test MLPA assay.

Gene Copy Number Analysis

We generated MLPA data, processed it in GeneMarker and exported it to a Microsoft Excel worksheet (Supplementary Data Sheet S1). Three samples were excluded at this stage due to poor data quality. To enable sample-to-sample comparison, we normalized the data within each sample using the mean signal intensity of the control probes ctrl1–ctrl5. The data were then compared and the copy numbers were estimated relative to the Col-0 accession that has the basic copy number of each gene analyzed in this assay ($2n = 2$) and therefore served as the reference sample. To reveal groups of accessions with distinct gene copy numbers, the population data were displayed as dot plots, histograms of the signal intensities or (for genes targeted by two MLPA probes) as 2D plots. We set the duplication/deletion thresholds at <0.7 and >1.3 of the relative intensity, respectively, for all genes in the assay. Subsequently, for each gene, the samples passing the threshold values were clustered and the clusters were manually assigned the copy numbers, as demonstrated previously (Marcinkowska-Swojak et al., 2014; Zmienko et al., 2016).

Non-variable Regions

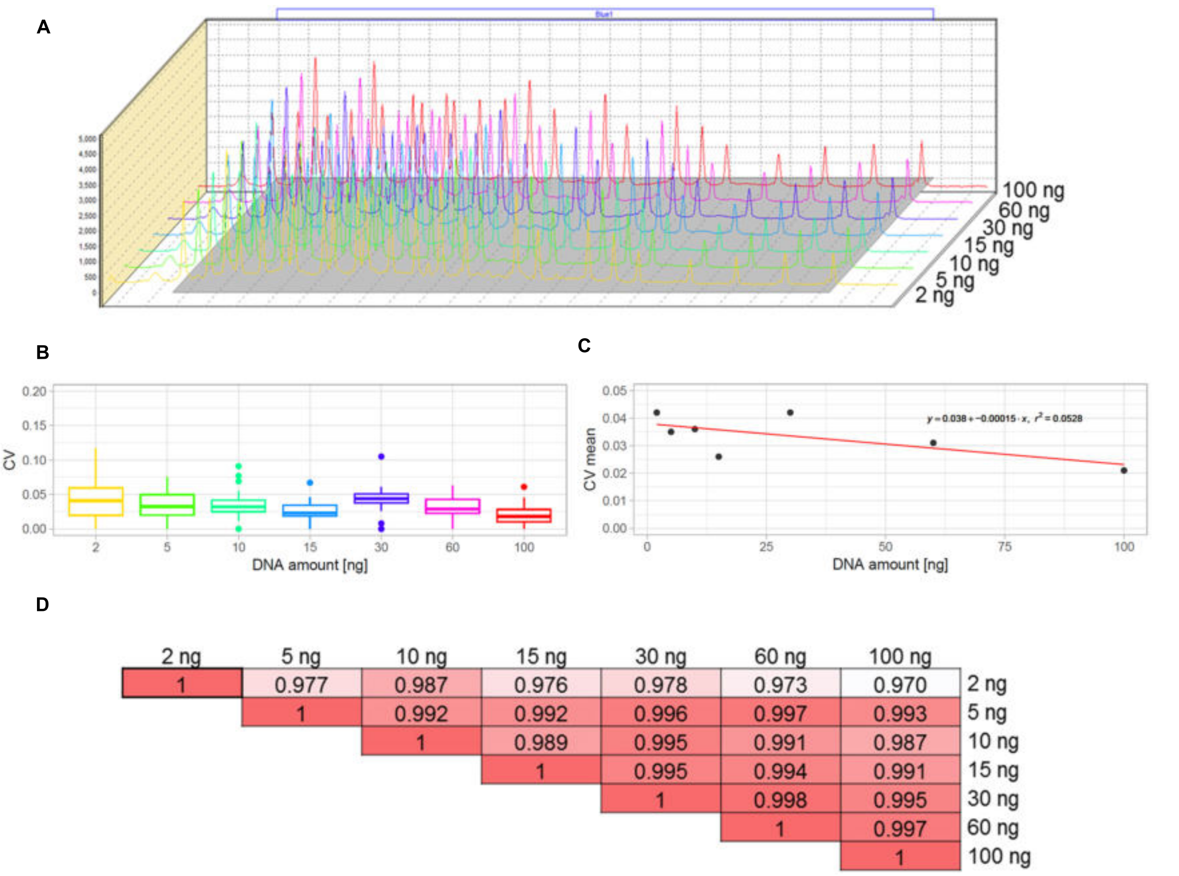
The probes mlpaA, mlpaB1, and mlpaB2 targeted two genes predicted to have the same copy number in all accessions: *AT1G47670*, coding for lysine histidine transporter-like 8 (mlpaA), and *AT1G80830*, coding for NRAMP1 transporter (mlpaB1 and mlpaB2). For all accessions, the relative signals

TABLE 2 | The probe composition and gene targets of Ath.test assay.

Probe name	Probe length	Target genomic site	Locus ID	Predicted CNV status	Source*
ctrl1	96 nt	Chr1:25593..25645	AT1G01040	Non-variable; normalization control	a
ctrl2	111 nt	Chr4:11476533..11476582	AT4G21580	Non-variable; normalization control	a
ctrl3	124 nt	Chr2:15194440..15194490	AT2G36230	Non-variable; normalization control	a
ctrl4	144 nt	Chr5:7847361..7847414	AT5G23290	Non-variable; normalization control	a
ctrl5	172 nt	Chr1:27465468..27465522	AT1G73010	Non-variable; normalization control	a
mlpaA	160 nt	Chr1:17539289..17539343	AT1G47670	Non-variable	b; c
mlpaB1; mlpaB2	90 nt 148 nt	Chr1:30374276..30374321 Chr1:30373647..30373699	AT1G80830	Non-variable	b; c
mlpaC	93 nt	Chr1:11651708..11651754	AT1G32300	Biallelic	b
mlpaD1; mlpaD2	105 nt 114 nt	Chr1:9575624..9575678 Chr1:9577003..9577055	AT1G27570	Multiallelic	b; c
mlpaE1; mlpaE2	136 nt 196 nt	Chr1:19726669..19726721 Chr1:19727385..19727439	AT1G52950	Multiallelic	b; c
mlpaF1; mlpaF2	99 nt 120 nt	Chr3:7737420..7737467 Chr3:7737872..7737929	AT3G21960	Multiallelic	b; c
mlpaG1; mlpaG2	128 nt 164 nt	Chr4:10641616..10641668 Chr4:10644628..10644679	AT4G19520	Biallelic	c
mlpaH	180 nt	Chr4:13592606..13592658	AT4G27080	Multiallelic	b; c
mlpaI	117 nt	Chr4:17705274..17705327	AT4G37685	Multiallelic	b
mlpaJ1; mlpaJ2	108 nt 156 nt	Chr5:2976409..2976464 Chr5:2978013..2978065	AT5G09590	Multiallelic; part of the gene	c
mlpaK1; mlpaK2	188 nt 102 nt	Chr5:22228424..22228479 Chr5:22229438..22229488	AT5G54710	Multiallelic	b; c
mlpaL	132 nt	Chr5:24796111..24796161	AT5G61700	Multiallelic	c

**The initial information about the gene CNV status comes from the following resources: a, Zmienko et al. (2016); b, Arabidopsis 1001 Genomes Project; c, our unpublished analysis of the WGS data originally presented in Cao et al. (2011).*

[4] http://arabidopsis.info/



	2 ng	5 ng	10 ng	15 ng	30 ng	60 ng	100 ng
2 ng	1	0.977	0.987	0.976	0.978	0.973	0.970
5 ng		1	0.992	0.992	0.996	0.997	0.993
10 ng			1	0.989	0.995	0.991	0.987
15 ng				1	0.995	0.994	0.991
30 ng					1	0.998	0.995
60 ng						1	0.997
100 ng							1

FIGURE 2 | Optimization of the DNA template amount for copy number variant (CNV) genotyping in *Arabidopsis*. The data were obtained from three replicates of MLPA assay per each DNA amount tested: 2, 5, 10, 15, 30, 60, and 100 ng. **(A)** comparison of the peak heights and localization on the electropherograms. **(B,C)** Variance of probe intensity measurements across tested DNA concentrations. Boxplots **(B)** present the distribution of coefficients of variation (CV) calculated separately for each probe and each DNA amount; The linear plot **(C)** presents regression analysis (CV vs input amount). For visibility, mean CV values (all probes, each in three replicates) per input amount are displayed on the plot; **(D)** pairwise correlation of results obtained for all template amounts tested, presented as coefficients of determination (R^2) of the MLPA probe intensity data.

from these three probes were at the same level as those in Col-0 (mean intensity 1.01, 1.03, and 0.93, respectively, see **Figure 3A**) and showed very little variance (CV 0.060, 0.089, and 0.064, respectively). Additional evaluation of the mlpaB1 and mlpaB2 probes on a 2D plot revealed that all samples were grouped in one cluster (**Figure 3B**).

Biallelic CNVs

We analyzed two genes with presence-absence variation revealed by the WGS data analysis: *AT1G32300* (coding for D-arabinono-1,4-lactone oxidase family protein) and *AT4G19520* (coding for TIR-NBS-LRR class disease resistance protein). We designed one probe (mlpaC) for *AT1G32300* exon 1 and two probes, mlpaG1 and mlpaG2, for *AT4G19520* exons 3 and 5, respectively. For *AT1G32300*, we observed a dominant population of samples with mean signal intensity 1.08, indicative of two gene copies per diploid genome. The remaining samples formed a distinct group with mean signal intensity 0.09, indicative of the absence of the analyzed gene in the respective accessions (**Figure 4A**). In the case of *AT4G19520*, the combined data for the mlpaG1 and mlpaG2 probes revealed the presence of two compact clusters (**Figure 4B**). One cluster included 29 accessions with no difference in copy number relative to Col-0 (mlpaG1 mean intensity 1.03; mlpaG2 mean intensity 1.01). The other cluster included 47 accessions with substantially reduced intensity (mlpaG1 mean intensity 0.14; mlpaG2 mean intensity 0.12), indicative of the deletion.

Multiallelic CNVs: One MLPA Probe Per Gene

For three genes that overlap multiallelic CNVs we designed 1 MLPA probe per gene in Ath.test assay (**Figure 5A**). Gene *AT4G37685* codes for a hypothetical protein and is targeted by the mlpaI probe. Majority of accessions (39) harbor two copies of this gene. Gene deletion was detected in eight accessions and

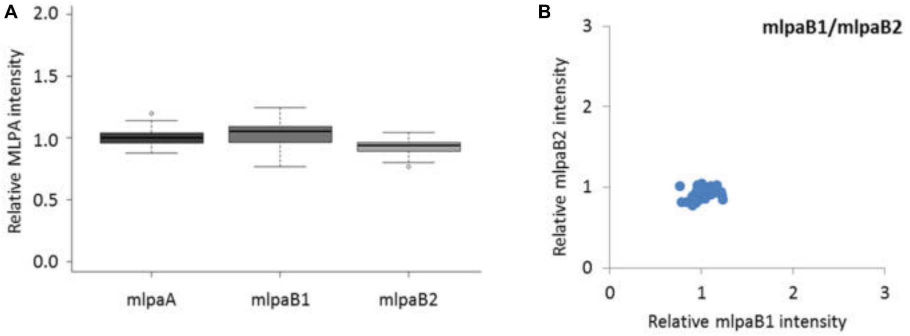


FIGURE 3 | Multiplex ligation-dependent probe amplification assay results for non-variable regions. (A) relative intensity distribution of the mlpaA probe, targeting the *AT1G47670* gene as well as mlpaB1 and mlpaB2 probes, both targeting the *AT1G80830* gene, among the accessions; **(B)** 2D plot of intensity from probes mlpaB1 and mlpaB2.

duplication in 30 accessions. Of the latter, 22 accessions had four copies, seven accessions had six copies, and one harbored a very high-level duplication, most likely ≥12 copies.

Gene *AT5G61700* codes for ATH16, a member of ABC transporter subfamily A and is targeted by probe mlpaL. In most analyzed accessions, the gene exists in two copies per diploid genome. In eight accessions, however, duplications were detected: four copies in three accessions, six copies in two accessions, and ≥10 copies in three accessions. It is worth noting that, in MLPA assays, the signal intensity is non-linearly related to the DNA copy number (Zmienko et al., 2016). This is manifested by reducing the distance between the clusters with different duplication levels for high copy numbers. Consequently, a large number of samples harboring high-level duplications is needed to precisely distinguish the clusters of 8 and more copies from each other.

Gene *AT4G27080* codes for a protein disulfide isomerase that is involved in cell redox homeostasis and is targeted by the mlpaH probe. From the WGS data, we predicted that majority of accessions harbor partial or full duplications of this gene. Likewise, MLPA analysis revealed that only nine accessions harbor two copies of *AT4G27080* gene, while duplications were detected in 68 accessions. Among them, we clearly identified a group of 44 accessions with four copies, but the remaining accessions were less distinctive and formed two heterogeneous groups which we named "medium-level duplications" (10 accessions) and "high-level duplications" (14 accessions). For 12 of these "high-level duplication" accessions, the mlpaH peak intensity counts reached the upper detection limits (see **Notes** section below for additional comments). We concluded that designing two or more MLPA probes targeting this genomic region and repeating the assay with adjusted capillary electrophoresis parameters would be helpful in more accurate distinction of the CNV genotypes or resolution of the structural complexity of the investigated gene.

Multiallelic CNVs: Two MLPA Probes Per Gene

For 2 other genes that overlap multiallelic CNVs we designed two MLPA probes per gene (**Figure 5B**). The *AT5G54710* gene codes for an ankyrin repeat family protein and is positioned between two other ankyrin repeat family protein coding genes, in the region that is highly copy number variable. We used two specific probes (mlpaK1 and mlpaK2), located in the fourth and third exons of *AT5G54710*, respectively, and confirmed that this gene is multiallelic. The high linear correlation of the mlpaK1 and mlpaK2 probe intensities allowed us distinguish several clusters of accessions with distinct copy numbers: 0 copies (2 accessions), 2 copies (54 accessions), 4 copies (8 accessions), 6 copies (6 accessions), and 8 copies (1 accession). We did not assign the integer copy numbers for 6 accessions which displayed uneven duplication level based on the mlpaK1 and mlpaK2 probe signal.

The *AT5G09590* gene, encoding mitochondrial heat shock protein MTHSC70-2, is localized in the breakpoint of a large CNV that encompasses loci *AT5G09590 – AT5G09630*. Consequently, *AT5G09590* is only partially duplicated in several accessions. We designed two probes, localized outside of and within the CNV region (mlpaJ1, targeting fourth exon and mlpaJ2, targeting sixth exon, respectively). The results of the MLPA assay clearly revealed that only the 3′ part of *AT5G09590* (targeted by probe mlpaJ2) is duplicated: 43 accessions harbored four copies, two accessions harbored six copies, and one accession harbored at least 10 copies. The region targeted by probe mlpaJ1 invariantly had two copies in all accessions.

Complex Multiallelic CNVs

Some genomic regions, e.g., these that harbor clustered multigene families, may display high structural diversity in the populations. A gene may be fully duplicated/deleted in some accessions while in the other ones only part of this gene may display copy number alteration. Additionally, the duplicated DNA copies within one sample may differ from each other in length and sequence, which may affect the affinity of the MLPA probe to some (but not all) copies. Consequently, the copy number pattern revealed by the MLPA analysis may be complex. Below we present some examples of MLPA analysis in multiallelic CNVs with a complex structure (**Figure 5C**).

The *AT3G21960* gene is localized in the central part of a ~50 kb CNV, that encompasses 21 genes, mainly members of

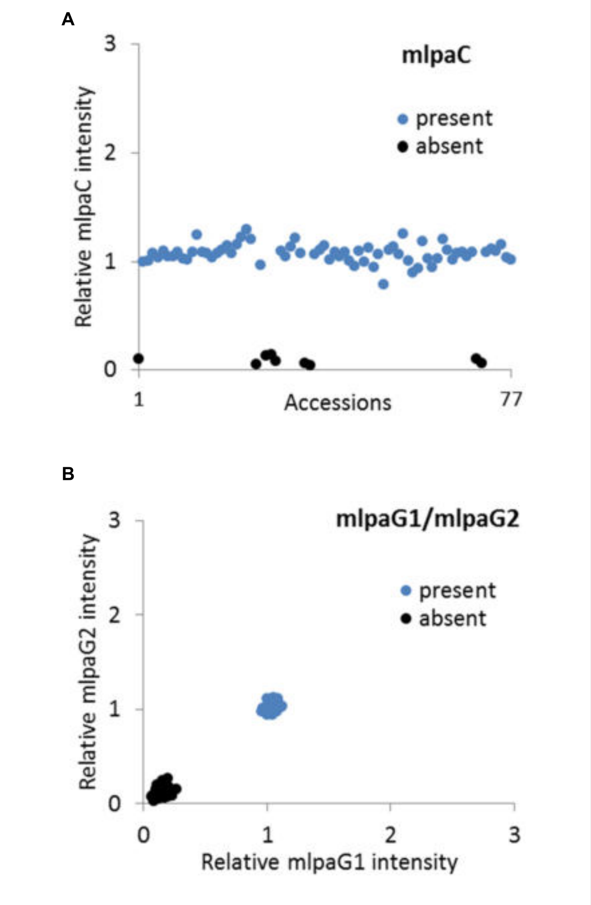


FIGURE 4 | Multiplex ligation-dependent probe amplification assay results for presence-absence CNVs. (A) relative intensity from probe mlpaC targeting the *AT1G32300* gene, in individual accessions; **(B)** 2D plot of relative signal from mlpaG1 and mlpaG2, both targeting the *AT4G19520* gene. Clusters are colored according to the deduced CNV status.

the receptor-like protein kinase-related family and genes coding for proteins with unknown domain DUF26. We assayed the *AT3G21960* gene with specific probes targeting exons 1 and 2 (probes mlpaF1 and mlpaF2, respectively). In 30 samples the signals from these probes were highly correlated and formed 4 distinct groups of: 0 copies (1 accession), 2 copies (26 accessions), 4 copies (1 accession) and 6 copies (2 accessions). In 6 accessions, however, only the mlpaF2 probe intensity was elevated (1.83–6.54), while mlpaF1 intensity was about 1. On the contrary, the remaining 41 accessions formed a compact cluster, with the mlpaF1 intensity below 0.7 (the value that has been set as the deletion threshold), and the mlpaF2 intensity about 1. A brief evaluation of the *AT3G21960* genomic sequence inferred from WGS data[5] (obtained with Pseudogenomes Download Tool) provided evidence that this complex pattern is true, as 519 out of 1135 accessions with available genomic data had 80–100% uncalled sites (Ns) in the exon 1 sequence, while only 3 accessions had 80–100% uncalled sites in exon 2 sequence.

Complex multiallelic CNVs are often related to the activity of mobile genetic elements, which may trigger partial or full deletion/duplication of the nearby genes. Gene *AT1G52950* codes for a nucleic acid-binding OB fold-like protein and is localized within one CNV region with a nearby transposable element gene *AT1G52960* (the two loci are separated by only 3.6 kb distance). We assayed the copy number status of *AT1G52950* using two probes, mlpaE1 to target exon 6 and mlpaE2 target exon 9. For 69 accessions, we detected compact clusters with distinct copy numbers (0 to 6 copies) and a high correlation between the two measurements ($R^2 = 0.9881$). Interestingly, in two cases, the intensity data suggested the existence of one copy and three copies of the *AT1G52950* gene per diploid genome in the surveyed individuals. *Arabidopsis* is a highly self-pollinating species for which most genomic loci are expected to exist in a homozygous state, therefore assaying additional individuals would be necessary to establish the representative gene copy number for these two accessions in a population study. For seven accessions, of which six originated from Southern Tyrol region and 1 was a Spanish relict accession (1001 Genomes Consortium, 2016), the copy number status indicated by probe mlpaE1 was always higher than the copy number status indicated by probe mlpaE2. This effect may have many reasons, e.g., partial duplication or deletion of a gene of interest, sequence divergence in some duplicated copies that affect the hybridization of one MLPA probe, etc. Unambiguous interpretation of these data would require additional region characterization by sequencing. Nevertheless, the signals from both probes were also well correlated ($R^2 = 0.9856$). Finally, one accession displayed an extremely high level of duplications at the mlpaE1 target site while no copy number changes were observed at the mlpaE2 site.

Effect of Non-specific Hybridization on MLPA Signal

To present the effect of compromised probe specificity on the MLPA results, we assayed a gene *AT1G27570*, which encodes the phosphatidylinositol 3- and 4-kinase family protein and is localized within the large multiallelic CNV (over 20 kb). We designed two probes, mlpaD1 and mlpaD2, targeting this gene, of which only mlpaD2 was specific to *AT1G27570*. Probe mlpaD1 had an alternative target site (with only two mismatches in the left TSS and one mismatch in the right TSS, distant from the ligation site) in the nearby gene *AT1G27590*, not copy number variable. As a result, the signal from the mlpaD1 probe was elevated by the background signal from the alternative target site. This background signal was stable (due to unchanged copy number of *AT1G27590* gene in all accessions) therefore the high correlation between the data for mlpaD1 and mlpaD2 probes was preserved (**Figure 6A**). As a rule, we suggest re-designing of the MLPA probes that produce non-specific signal. However, if a set of the control samples that carry confirmed deletion of the gene of interest can be defined, these samples may be used for the data correction. In the present example, we calculated the mean non-specific signal of probe mlpaD1 in the cluster of 15 samples with gene deletions (marked in black color in

[5] http://1001genomes.org

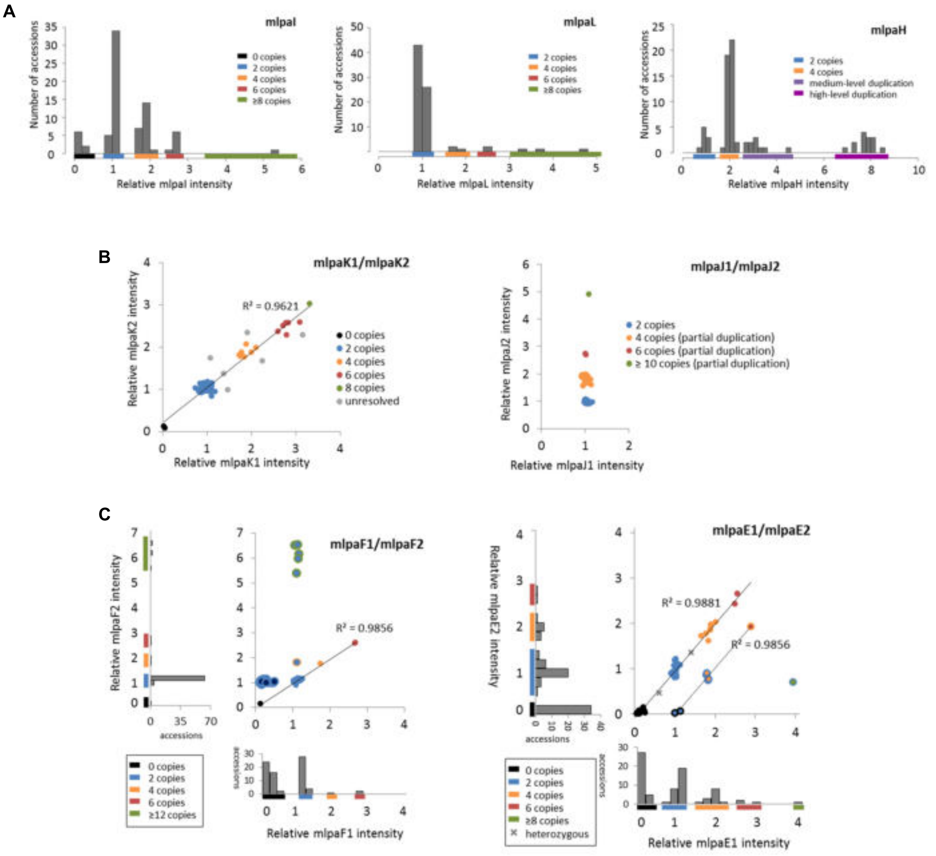


FIGURE 5 | Multiplex ligation-dependent probe amplification results for multiallelic CNVs. (A) CNV genotyping with one MLPA probe per gene. Histograms present the relative signal distribution from probe mlpaI (targeting the *AT4G37685* gene), probe mlpaL (targeting the *AT5G61700* gene), and probe mlpaH (targeting the *AT4G27080* gene). The histogram bin size is 0.2 in all plots; **(B)** CNV genotyping with two MLPA probes per gene. 2D plots present the relative signal from probes mlpaK1and mlpaK2 (both targeting the *AT5G54710* gene) and from probes mlpaJ1 and mlpaJ2 (both targeting the *AT5G09590* gene). Clusters are colored according to deduced CNV status. The coefficient of determination (R^2) is calculated for accessions with assigned copy numbers. **(C)** Genotyping complex multiallelic CNVs. 2D intensity plots present relative signal from probes mlpaF1and mlpaF2 (targeting exon 1 and exon 2 of the *AT3G21960* gene, respectively) and from probes mlpaE1and mlpaE2 (targeting exon 6 and exon 9 of the *AT1G52950* gene, respectively). Clusters are colored according to deduced CNV status. The coefficient of determination (R^2) is calculated for subsets of accessions, as detailed in the main text.

Figure 6A). This value was then subtracted from the probe mlpaD1 signal in each sample, before estimating the intensity ratio relative to Col-0 accession. The correction improved the relative intensity ratio observed for probe mlpaD1 (**Figure 6B**). We note here, that the process of data correction had no effect on the overall correlation between the signals from probes mlpaD1 and mlpaD2. This correlation was high ($R^2 = 0.9386$), therefore allowing to distinguish the copy number clusters on 2D plots pretty easily both before and after data correction.

NOTES

Below we included some notes on the limitations of the procedure, common mistakes and possible artifacts related to the presented application.

Probe Design

Oligonucleotide MLPA probes described in this procedure target specific sequences in the genome, typically 45–75 bp. Regions located outside of the probe's recognition sequence may have different copy number status. If partial gene duplication/deletion or insertion of duplicated sequence is suspected, additional probes, e.g., covering different exons of the gene should be included in the assay.

Compromised ability of MLPA probe to recognize the target sequence may be the source of false positive results. Sequence changes (SNPs, indels, point mutations) in the target sequence detected by a probe can negatively affect or completely prevent probe binding. The critical positions in the TSS sequence are these constituting the ligation site; the presence of a SNP at or near the ligation site will disrupt the ligation step and result in

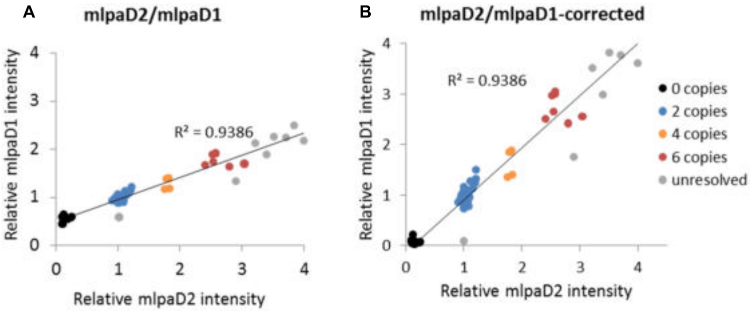


FIGURE 6 | Effect of non-specific probe hybridization on the MLPA results. Probes mlpaD1 and mlpaD2 target the *AT1G27570* gene. Additionally, probe mlpaD1 targets also the *AT1G27590* gene. **(A)** 2D intensity plot of relative signal from probes mlpaD1 and mlpaD2; **(B)** 2D intensity plot of relative signal from probe mlpaD1 corrected for the presence of non-specific hybridization signal (see main text for details) and probe mlpaD2. Clusters are colored according to the deduced CNV status and are identical for each sample on both plots. The coefficient of determination (R^2) is calculated for all accessions. The deletion of *AT1G27570* gene has been confirmed by PCR in accessions from the "0 copies" cluster (not shown).

no signal from the MLPA probe, falsely indicative of deletion of the region in the affected sample (Kim et al., 2016). Note that the MLPA technique can be also used for detecting small mutations (Marcinkowska-Swojak et al., 2016), but these applications are not covered in the present protocol.

The accuracy of the results is also strictly dependent on the MLPA probe specificity. If alternative target site exists in the genome (e.g., in a paralogue or a pseudogene), it will generate non-specific signal (see Effect of Non-specific Probe Hybridization on the MLPA Results Section). To this end, for plants with incomplete genome information we strongly advise designing ≥2 MLPA probes per gene, to minimize this risk.

In the case of newly designed MLPA probes we recommend verifying their performance on a (set of) well characterized reference samples. If no product is observed, make sure that the common mistakes interfering with the experimental steps are avoided (see below). If needed, re-design the MLPA probe.

Assay Design and Performing

Multiplex ligation-dependent probe amplification results may be compromised by multiple factors that will affect the enzymatic reactions and result in reduced peak signals. These factors include but are not limited to: DNA integrity and contamination, presence of PCR inhibitors in the samples, incomplete DNA denaturation, sample evaporation, suboptimal amount of the sample DNA used. In the Section "Stepwise Procedures" we included useful tips regarding the sample preparation and assay setup. Additional comments are given below.

If the DNA sample contamination is a suspected problem, perform new DNA extraction. From our experience, we advise using column-based methods, e.g., DNeasy Plant Mini Kit (Qiagen) for DNA extraction (or purification of DNA extracted with other methods) because they produce samples of high purity and comparable amounts.

Use multichannel pipettes to reduce the pipetting time and avoid sample evaporation.

Reduce sample-to-sample variability by simultaneous performing multiple assays, using strips (preferable) or multiwell PCR plates. Use the same MLPA Probe Set Mix preparation for all samples under comparison.

Replacing the strip caps on each opening minimizes the risk of sample cross-contamination.

Follow the capillary electrophoresis protocols (size standard, sample preparation, injection time and voltage) suitable for the instrument used. Decrease injection time if the peaks are out of range. We recommend prior optimization of the DNA template amount in the assay and capillary electrophoresis conditions on a validated reference sample.

Abnormal pictures after capillary electrophoresis may indicate capillary electrophoresis problems but they also may result from the PCR step troubles. See the MLPA troubleshooting wizard by MRC Holland[6] for common peak pattern problems and possible solutions.

Data Analysis and Copy Number Estimation

It is advisable to manually check the peaks identified by GeneMarker before further data processing. In our assay, we repeatedly observed that the software did not detect the peaks for probe mlpaH in 12 samples and reported "0" intensity for this probe (Supplementary Figure S3). In fact, high intensity peaks from probe mlpaH with their tops flattened (cut) were present in these samples, which indicated that the signal exceeded the capillary electrophoresis system detection limits. We manually corrected the peak localization and used the maximum reported values for copy number calculation, but this likely resulted in underestimation of the gene copy number in these samples in our study (see Section "Multiallelic CNVs: One MLPA Probe Per Gene"). To accurately quantify the probe signal, repeating the electrophoresis with lower injection time would be necessary. The results from high and low injection time electropherograms may be then merged after internal control probe normalization step, to preserve good resolution of the low intensity peaks.

[6] http://www.mlpa.com/elearning/tswizard/

Multiplex ligation-dependent probe amplification is a relative technique, therefore selecting well validated reference samples with basic copy number of the region of interest (usually two copies) is essential for accurate quantification. However, in case of population scale CNV genotyping of numerous independent genomic regions in a multiplex assay (similar to example provided in this paper) such a reference sample may not exist or remains unknown. Providing that sufficiently large number of samples in the population are genotyped, the presented protocol still allows for inferring the cluster copy numbers without a reference sample, under the assumption that the neighboring clusters of accessions/lines differ by two copies and that the distances between these clusters are ∼equal in the range of 0–4 copies (see Zmienko et al., 2016 for further discussion on the distances between the clusters in MLPA assays).

Validation of the Results

Regardless of the number of probes and samples used, we recommend to verify the positive MLPA results with an independent technique. We advise performing droplet digital PCR (ddPCR) on selected samples, as this approach allows for estimating gene copy numbers at the same or even higher range, as the MLPA procedure described in this protocol (Zmienko et al., 2016). Additionally, ddPCR generates amplicons of ∼60–200 bp, therefore allows for genome assaying at similar resolution as MLPA.

CONCLUSION

In this work, we described the protocol for the simple MLPA-based CNV genotyping in plants, with particular emphasis on the model plant *Arabidopsis*. We provided a description of the probe design process, experimental setup, and data analysis. We also discussed the results of the exemplar multiplex assay and showed that the MLPA method is very robust and is a rich source of information regarding the CNV in the analyzed samples. The abundant genomic data obtained for a growing number of species as a part of large-scale sequencing projects, highlight CNV as the major contributor to natural diversity at a genotype level (Zarrei et al., 2015; 1001 Genomes Consortium, 2016; Bai et al., 2016). Gene duplication has been considered the major factor driving long-term evolution and gene birth by sub- and neofunctionalization of the duplicated copies (Conant et al., 2014). Some regions in the genome may be more prone to CNV than the others, due to their specific structural features, that will locally induce the mechanisms leading to CNV formation, e.g., non-allelic recombination (Zmienko et al., 2016). The duplication / deletion events may have also consequences on organism's fitness and contribute to the adaptation to environmental challenges, as well as to coevolutionary interactions between host and pathogen or a symbiont (reviewed in: Kondrashov, 2012, Żmieńko et al., 2014). Remarkably, the protein coding genes displaying CNVs are often related to environmental stress response and pathogen resistance (Cook et al., 2012; Maron et al., 2013). The creation of high-confidence CNV maps and assessing the gene copy number in large populations will enhance the studies on the evolution of genomes in the context of CNV origin, fixation and the impact on the phenotype. These data can be later combined with the results of the transcriptomic, proteomic, metabolomics, protein interaction, phenotyping, and other studies). We recently used the MLPA method to genotype *MSH2*, *AT3G18530*, and *AT3G18535* copy number in a set of 189 natural accessions. Based on these results, we were subsequently able to reveal the recurrent nature of *AT3G18530* and *AT3G18535* duplications/deletions and to dissect the structural features that promoted non-allelic homologous recombination, leading to a widespread occurrence of the *AT3G18530* and *AT3G18535* genes deletion in nature (Zmienko et al., 2016).

This protocol will enable potential users to introduce the MLPA technique in plant genetic and population biology studies. The technique is multiplexable and very well suited for verification of WGS-based analyses or for rapid characterization of copy number status across a region of interest in large populations. Notably, once designed, the individual MLPA probes may be used in various combinations according to one's needs, providing that the lengths of the probes in one assay are unique. We believe that the MLPA protocol presented in the current work will contribute to accelerating the discovery of new associations between CNV and important traits in plants.

AUTHOR CONTRIBUTIONS

AS-C prepared DNA samples, performed MLPA assays, analyzed data, helped prepare figures, and draft the manuscript. MM-Z performed template optimization experiments. MM-S helped design the MLPA probes and set up the assay. PK analyzed the data and helped draft the manuscript. MF contributed to the conception of the work and revised the manuscript. AZ conceived of and designed the study, analyzed data, oversaw the research, prepared figures, and wrote the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was supported by National Centre of Science grants (2011/01/B/NZ2/04816 and 2014/13/B/NZ2/03837). Publication of the results was supported by Polish Ministry of Science and Higher Education, under the KNOW program.

ACKNOWLEDGMENT

We thank Pawel Wojciechowski and Kaja Milanowska for help with WGS data analysis.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpls.2017.00222/full#supplementary-material

REFERENCES

1001 Genomes Consortium (2016). 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell* 166, 1–11. doi: 10.1016/j.cell.2016.05.063

Alkan, C., Coe, B. P., and Eichler, E. E. (2011). Genome structural variation discovery and genotyping. *Nat. Rev. Genet.* 12, 363–376. doi: 10.1038/nrg2958

Bai, Z., Chen, J., Liao, Y., Wang, M., Liu, R., Ge, S., et al. (2016). The impact and origin of copy number variations in the *Oryza* species. *BMC Genomics* 17:261. doi: 10.1186/s12864-016-2589-2

Beló, A., Beatty, M. K., Hondred, D., Fengler, K. A., Li, B., and Rafalski, A. (2010). Allelic genome structural variations in maize detected by array comparative genome hybridization. *Theor. Appl. Genet.* 120, 355–367. doi: 10.1007/s00122-009-1128-9

Bharuthram, A., Paximadis, M., Picton, A. C. P., and Tiemessen, C. T. (2014). Comparison of a quantitative Real-Time PCR assay and droplet digital PCR for copy number analysis of the CCL4L genes. *Infect. Genet. Evol.* 25, 28–35. doi: 10.1016/j.meegid.2014.03.028

Cantsilieris, S., Baird, P. N., and White, S. J. (2013). Molecular methods for genotyping complex copy number polymorphisms. *Genomics* 101, 86–93. doi: 10.1016/j.ygeno.2012.10.004

Cantsilieris, S., Western, P. S., Baird, P. N., and White, S. J. (2014). Technical considerations for genotyping multi-allelic copy number variation (CNV), in regions of segmental duplication. *BMC Genomics* 15:329. doi: 10.1186/1471-2164-15-329

Cao, J., Schneeberger, K., Ossowski, S., Günther, T., Bender, S., Fitz, J., et al. (2011). Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nat. Genet.* 43, 956–963. doi: 10.1038/ng.911

Ceulemans, S., van der Ven, K., and Del-Favero, J. (2012). "Targeted screening and validation of copy number variations," in *Genomic Structural Variants: Methods and Protocols, Methods in Molecular Biology*, ed. L. Feuk (Berlin: Springer Science+Business Media), 369–384. doi: 10.1007/978-1-61779-507-7_18

Chang, C., Lu, J., Zhang, H.-P., Ma, C.-X., and Sun, G. (2015). Copy number variation of cytokinin oxidase gene Tackx4 associated with grain weight and chlorophyll content of flag leaf in common wheat. *PLoS ONE* 10:e0145970. doi: 10.1371/journal.pone.0145970

Conant, G. C., Birchler, J. A., and Pires, J. C. (2014). Dosage, duplication, and diploidization: clarifying the interplay of multiple models for duplicate gene evolution over time. *Curr. Opin. Plant Biol.* 19, 91–98. doi: 10.1016/j.pbi.2014.05.008

Cook, D. E., Bayless, A. M., Wang, K., Guo, X., Song, Q., Jiang, J., et al. (2014). Distinct copy number, coding sequence, and locus methylation patterns underlie Rhg1-mediated soybean resistance to soybean cyst nematode. *Plant Physiol.* 165, 630–647. doi: 10.1104/pp.114.235952

Cook, D. E., Lee, T. G., Guo, X., Melito, S., Wang, K., Bayless, A. M., et al. (2012). Copy number variation of multiple genes at Rhg1 mediates nematode resistance in soybean. *Science* 338, 1206–1209. doi: 10.1126/science.1228746

Duitama, J., Silva, A., Sanabria, Y., Cruz, D. F., Quintero, C., Ballen, C., et al. (2015). Whole genome sequencing of elite rice cultivars as a comprehensive information resource for marker assisted selection. *PLoS ONE* 10:e0124617. doi: 10.1371/journal.pone.0124617

Gaines, T. A., Zhang, W., Wang, D., Bukun, B., Chisholm, S. T., Shaner, D. L., et al. (2010). Gene amplification confers glyphosate resistance in *Amaranthus palmeri*. *Proc. Natl. Acad. Sci. U.S.A.* 107, 1029–1034. doi: 10.1073/pnas.0906649107

Hanada, K., Sawada, Y., Kuromori, T., Klausnitzer, R., Saito, K., Toyoda, T., et al. (2011). Functional compensation of primary and secondary metabolites by duplicate genes in *Arabidopsis thaliana*. *Mol. Biol. Evol.* 28, 377–382. doi: 10.1093/molbev/msq204

Hömig-Hölzel, C., and Savola, S. (2012). Multiplex ligation-dependent probe amplification (MLPA) in tumor diagnostics and prognostics. *Diagn. Mol. Pathol.* 21, 189–206. doi: 10.1097/PDM.0b013e3182595516

Kim, M. J., Cho, S. I., Chae, J. H., Lim, B. C., Lee, J. S., Lee, S. J., et al. (2016). Pitfalls of multiple ligation-dependent probe amplifications in detecting DMD exon deletions or duplications. *J. Mol. Diagn.* 18, 253–259. doi: 10.1016/j.jmoldx.2015.11.002

Klonowska, K., Ratajska, M., Czubak, K., Kuzniacka, A., Brozek, I., Koczkowska, M., et al. (2015). Analysis of large mutations in BARD1 in patients with breast and/or ovarian cancer: the Polish population as an example. *Sci. Rep.* 5:10424. doi: 10.1038/srep10424

Kohany, O., Gentles, A. J., Hankus, L., and Jurka, J. (2006). Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and censor. *BMC Bioinformatics* 7:474. doi: 10.1186/1471-2105-7-474

Kondrashov, F. A. (2012). Gene duplication as a mechanism of genomic adaptation to a changing environment. *Proc. Biol. Sci.* 279, 5048–5057. doi: 10.1098/rspb.2012.1108

Koralewski, T. E., and Krutovsky, K. V. (2011). Evolution of exon-intron structure and alternative splicing. *PLoS ONE* 6:e18055. doi: 10.1371/journal.pone.0018055

Kozlowski, P., Roberts, P., Dabora, S., Franz, D., Bissler, J., Northrup, H., et al. (2007). Identification of 54 large deletions/duplications in TSC1 and TSC2 using MLPA, and genotype-phenotype correlations. *Hum. Genet.* 121, 389–400. doi: 10.1007/s00439-006-0308-9

Li, X., Wu, H. X., Dillon, S. K., and Southerton, S. G. (2009). Generation and analysis of expressed sequence tags from six developing xylem libraries in *Pinus radiata* D. Don. *BMC Genomics* 10:41. doi: 10.1186/1471-2164-10-41

Li, X., Wu, H. X., and Southerton, S. G. (2011). Transcriptome profiling of *Pinus radiata* juvenile wood with contrasting stiffness identifies putative candidate genes involved in microfibril orientation and cell wall mechanics. *BMC Genomics* 12:480. doi: 10.1186/1471-2164-12-480

Li, X., Yang, X., and Wu, H. X. (2013). Transcriptome profiling of radiata pine branches reveals new insights into reaction wood formation with implications in plant gravitropism. *BMC Genomics* 14:768. doi: 10.1186/1471-2164-14-768

Ling, X.-Y., Zhang, G., Pan, G., Long, H., Cheng, Y., Xiang, C., et al. (2015). Preparing long probes by an asymmetric polymerase chain reaction-based approach for multiplex ligation-dependent probe amplification. *Anal. Biochem.* 487, 8–16. doi: 10.1016/j.ab.2015.03.031

Marcinkowska, M., Wong, K.-K., Kwiatkowski, D. J., and Kozlowski, P. (2010). Design and generation of MLPA probe sets for combined copy number and small-mutation analysis of human genes: EGFR as an example. *ScientificWorldJournal.* 10, 2003–2018. doi: 10.1100/tsw.2010.195

Marcinkowska-Swojak, M., Handschuh, L., Wojciechowski, P., Goralski, M., Tomaszewski, K., Kazmierczak, M., et al. (2016). Simultaneous detection of mutations and copy number variation of NPM1 in the acute myeloid leukemia using multiplex ligation-dependent probe amplification. *Mutat. Res.* 786, 14–26. doi: 10.1016/j.mrfmmm.2016.02.001

Marcinkowska-Swojak, M., Klonowska, K., Figlerowicz, M., and Kozlowski, P. (2014). An MLPA-based approach for high-resolution genotyping of disease-related multi-allelic CNVs. *Gene* 546, 257–262. doi: 10.1016/j.gene.2014.05.072

Maron, L. G., Guimarães, C. T., Kirst, M., Albert, P. S., Birchler, J. A., Bradbury, P. J., et al. (2013). Aluminum tolerance in maize is associated with higher MATE1 gene copy number. *Proc. Natl. Acad. Sci. U.S.A.* 110, 5241–5246. doi: 10.1073/pnas.1220766110

McCord, B. (2003). *Troubleshooting Capillary Electrophoresis Systems*. Available at: https://pl.promega.com/resources/profiles-in-dna/2003/troubleshooting-capillary-electrophoresis-systems/

McHale, L. K., Haun, W. J., Xu, W. W., Bhaskar, P. B., Anderson, J. E., Hyten, D. L., et al. (2012). Structural variants in the soybean genome localize to clusters of biotic stress-response genes. *Plant Physiol.* 159, 1295–1308. doi: 10.1104/pp.112.194605

Muñoz-Amatriaín, M., Eichten, S. R., Wicker, T., Richmond, T. A., Mascher, M., Steuernagel, B., et al. (2013). Distribution, functional impact, and origin mechanisms of copy number variation in the barley genome. *Genome Biol.* 14:R58. doi: 10.1186/gb-2013-14-6-r58

Perne, A., Zhang, X., Lehmann, L., Groth, M., Stuber, F., and Book, M. (2009). Comparison of multiplex ligation-dependent probe amplification and real-time PCR accuracy for gene copy number quantification using the beta-defensin locus. *Biotechniques* 47, 1023–1028. doi: 10.2144/000113300

Rudi, K., Rud, I., and Holck, A. (2003). A novel multiplex quantitative DNA array based PCR (MQDA-PCR) for quantification of transgenic maize in food and feed. *Nucleic Acids Res.* 31:e62. doi: 10.1007/s00217-009-1155-4

Saintenac, C., Jiang, D., and Akhunov, E. D. (2011). Targeted analysis of nucleotide and copy number variation by exon capture in allotetraploid wheat genome. *Genome Biol.* 12:R88. doi: 10.1186/gb-2011-12-9-r88

Schouten, J. P., McElgunn, C. J., Waaijer, R., Zwijnenburg, D., Diepvens, F., and Pals, G. (2002). Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. *Nucleic Acids Res.* 30:e57. doi: 10.1093/nar/gnf056

Springer, N. M., Ying, K., Fu, Y., Ji, T., Yeh, C.-T., Jia, Y., et al. (2009). Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. *PLoS Genet.* 5:e1000734. doi: 10.1371/journal.pgen.1000734

Stankiewicz, P., and Lupski, J. R. (2010). Structural variation in the human genome and its role in disease. *Annu. Rev. Med.* 61, 437–455. doi: 10.1146/annurev-med-100708-204735

Swanson-Wagner, R. A., Eichten, S. R., Kumari, S., Tiffin, P., Stein, J. C., Ware, D., et al. (2010). Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. *Genome Res.* 20, 1689–1699. doi: 10.1101/gr.109165.110

Tan, S., Zhong, Y., Hou, H., Yang, S., and Tian, D. (2012). Variation of presence/absence genes among *Arabidopsis* populations. *BMC Evol. Biol.* 12:86. doi: 10.1186/1471-2148-12-86

Thumma, B. R., Matheson, B. A., Zhang, D., Meeske, C., Meder, R., Downes, G. M., et al. (2009). Identification of a cis-acting regulatory polymorphism in a eucalypt COBRA-like gene affecting cellulose content. *Genetics* 183, 1153–1164. doi: 10.1534/genetics.109.106591

Wang, Y., Xiong, G., Hu, J., Jiang, L., Yu, H., Xu, J., et al. (2015). Copy number variation at the GL7 locus contributes to grain size diversity in rice. *Nat. Genet.* 47, 944–948. doi: 10.1038/ng.3346

Zarrei, M., MacDonald, J. R., Merico, D., and Scherer, S. W. (2015). A copy number variation map of the human genome. *Nat. Rev. Genet.* 16, 172–183. doi: 10.1038/nrg3871

Zheng, L.-Y., Guo, X.-S., He, B., Sun, L.-J., Peng, Y., Dong, S.-S., et al. (2011). Genome-wide patterns of genetic variation in sweet and grain sorghum (*Sorghum bicolor*). *Genome Biol.* 12:R114. doi: 10.1186/gb-2011-12-11-r114

Żmieńko, A., Samelak, A., Kozłowski, P., and Figlerowicz, M. (2014). Copy number polymorphism in plant genomes. *Theor. Appl. Genet.* 127, 1–18. doi: 10.1007/s00122-013-2177-7

Zmienko, A., Samelak-Czajka, A., Kozlowski, P., Szymanska, M., and Figlerowicz, M. (2016). *Arabidopsis thaliana* population analysis reveals high plasticity of the genomic region spanning MSH2, AT3G18530 and AT3G18535 genes and provides evidence for NAHR-driven recurrent CNV events occurring in this location. *BMC Genomics* 17:893. doi: 10.1186/s12864-016-3221-1

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.

MATERIAŁY SUPLEMENTARNE

Samelak-Czajka A, **Marszalek-Zenczak M**, Marcinkowska-Swojak M, Kozlowski P, Figlerowicz M and Zmienko A (2017)

MLPA-Based Analysis of Copy Number Variation in Plant Populations

Front. Plant Sci.8:222. doi: 10.3389/fpls.2017.00222

5-letni IF = 4,353

Materiały suplementarne dostępne online:
https://www.frontiersin.org/article/10.3389/fpls.2017.00222/full#supplementary-material

Publikacja 2

Zmienko A, **Marszalek-Zenczak M**, Wojciechowski P, Samelak-Czajka A, Luczak M, Kozlowski P, Karlowski WM, Figlerowicz M.

AthCNV: A Map of DNA Copy Number Variations in the Arabidopsis Genome

Plant Cell. 2020 Jun;32(6):1797-1819. doi: 10.1105/tpc.19.00640.

5-letni IF = 12,061

The Plant Cell, Vol. 32: 1797–1819, June 2020, www.plantcell.org © 2020 ASPB.

LARGE-SCALE BIOLOGY ARTICLE

AthCNV: A Map of DNA Copy Number Variations in the Arabidopsis Genome[OPEN]

Agnieszka Zmienko,[a,b,1] Malgorzata Marszalek-Zenczak,[a] Pawel Wojciechowski,[a,b] Anna Samelak-Czajka,[a] Magdalena Luczak,[a] Piotr Kozlowski,[a] Wojciech M. Karlowski,[c] and Marek Figlerowicz[a,b,1]

[a] Institute of Bioorganic Chemistry, Polish Academy of Sciences, 61-704 Poznan, Poland
[b] Institute of Computing Science, Faculty of Computing Science, Poznan University of Technology, Poznan, Poland
[c] Department of Computational Biology, Institute of Molecular Biology and Biotechnology, Faculty of Biology, Adam Mickiewicz University, 61-614 Poznan, Poland

ORCID IDs: 0000-0002-9128-7996 (A.Z.); 0000-0003-3498-3159 (M.M.-Z.); 0000-0001-8020-9493 (P.W.); 0000-0002-0167-8265 (A.S.-C.); 0000-0002-2182-5699 (M.L.); 0000-0003-3770-7715 (P.K.); 0000-0002-8086-5404 (W.M.K.); 0000-0002-6392-0192 (M.F.)

Copy number variations (CNVs) greatly contribute to intraspecies genetic polymorphism and phenotypic diversity. Recent analyses of sequencing data for >1000 Arabidopsis (*Arabidopsis thaliana*) accessions focused on small variations and did not include CNVs. Here, we performed genome-wide analysis and identified large indels (50 to 499 bp) and CNVs (500 bp and larger) in these accessions. The CNVs fully overlap with 18.3% of protein-coding genes, with enrichment for evolutionarily young genes and genes involved in stress and defense. By combining analysis of both genes and transposable elements (TEs) affected by CNVs, we revealed that the variation statuses of genes and TEs are tightly linked and jointly contribute to the unequal distribution of these elements in the genome. We also determined the gene copy numbers in a set of 1060 accessions and experimentally validated the accuracy of our predictions by multiplex ligation-dependent probe amplification assays. We then successfully used the CNVs as markers to analyze population structure and migration patterns. Finally, we examined the impact of gene dosage variation triggered by a CNV spanning the *SEC10* gene on *SEC10* expression at both the transcript and protein levels. The catalog of CNVs, CNV-overlapping genes, and their genotypes in a top model dicot will stimulate the exploration of the genetic basis of phenotypic variation.

INTRODUCTION

The frequent occurrence of duplications and deletions in eukaryotic genomes is among the most crucial factors that affect adaptation, evolution, and speciation (Kondrashov, 2012; Panchy et al., 2016). There are numerous lines of evidence that at an intraspecies level, these DNA copy number changes contribute to the phenotypic variation of humans, animals, and plants (McHale et al., 2012; Handsaker et al., 2015; Xu et al., 2016). Accordingly, efforts toward developing tools to detect copy number variations (CNVs) and map polymorphic regions have recently intensified. A good example of this trend is the latest advance in CNV discovery in the human genome, which has been empowered by the consecutive release of data from three phases of the 1000 Genomes Project. Remarkably, 60% of CNVs identified in phase 3 of this project (Sudmant et al., 2015) were novel compared to those identified in previous reports by Mills et al. (2011) and the 1000 Genomes Project Consortium et al. (2012), reflecting the methodological improvements and the importance of using large, diversified data sets.

The number of plant species for which CNV regions have been identified at the genome-wide scale has grown rapidly within the last decade (Swanson-Wagner et al., 2010; Chia et al., 2012; Muñoz-Amatriaín et al., 2013; Duitama et al., 2015; Hardigan et al., 2016; Fuentes et al., 2019). However, for Arabidopsis (*Arabidopsis thaliana*), an important model plant (Alonso-Blanco and Koornneef, 2000) with more than 1000 accessions whose genomes have been sequenced with coverage between 5× and 118× (1001 Genomes Consortium et al., 2016), comprehensive genome-wide CNV analysis is still required. Previous CNV analyses in Arabidopsis have been limited to individual lines or small populations and most often focused on characterizing presence-absence variation only. One of the earliest studies of this type combined the results of array-based hybridization and short read–based whole-genome sequencing (WGS) to identify ≥100-bp deletions in the genomes of four Arabidopsis accessions: Eil-0, Lc-0, Sav-0, and Tsu-1 (Santuari et al., 2010). These deletions overlapped with 987 to 1344 protein-coding genes (for simplicity, we refer to them as genes hereafter), and many of them were shared by at least two accessions. A larger study that focused on comparing the genomes of 17 accessions that were sequenced and assembled de

[1] Address correspondence to akisiel@ibch.poznan.pl or marekf@ibch.poznan.pl.

The authors responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantcell.org) are: Agnieszka Zmienko (akisiel@ibch.poznan.pl) and Marek Figlerowicz (marekf@ibch.poznan.pl).

[OPEN]Articles can be viewed without a subscription.

www.plantcell.org/cgi/doi/10.1105/tpc.19.00640

IN A NUTSHELL

Background: The genomes of individuals of a single species are not identical. There are genomic differences of various types (e.g., presence, absence, duplication, sequence alteration, or change in the location of a DNA fragment in one genome compared to another) and sizes (they may involve any DNA fragment from 1 bp to the entire chromosome). Variations in the number of copies of large DNA fragments (typically 500 bp or longer), named CNVs, may directly affect the structure and number of the genes they overlap. This in turn may cause phenotypic variation ranging from disease to increased adaptation of an individual with a specific CNV genotype.

Question: We wanted to identify CNVs in the Arabidopsis genome and evaluate how they affect the structure and genomic distribution of genes and transposable elements. We also wanted to test whether CNVs may be useful for genetic and functional studies.

Findings: We compared the genome sequencing data collected from 1,064 Arabidopsis accessions. We identified numerous CNVs, which are typically shorter than 20 kbp but together cover over one-third of the Arabidopsis genome. CNVs are concentrated in regions that are abundant in transposable elements and poor in protein-coding genes. Nevertheless, over 18% of genes overlap with CNVs. These genes are enriched for functions related to biotic stress responses. We determined the gene copy numbers in each accession and showed that these data are useful for population analysis in Arabidopsis. We used the CNVs to analyze population structure and reveal the genetic similarity of geographically distant accessions. We also demonstrated how variation in the number of specific genes might lead to variation at the gene transcriptional level, protein level, or phenotypic level. Additionally, our observations indicate that selective forces have opposite effects on shaping variation and the relative distribution patterns of genes and transposable elements.

Next steps: The map of CNVs in the Arabidopsis genome will help researchers explore the impact of this type of genetic polymorphism on various phenotypic traits. New gene variants that are not present in the reference genome can now be identified and studied.

novo from WGS data revealed multiple polymorphic regions that could not be mapped to the reference genome (Gan et al., 2011). Based on the same WGS data, Bush et al. (2014) identified numerous exon-overlapping regions in the Arabidopsis genome that were absent from at least one accession. These regions overlapped with 411 genes. A wider study that, in addition to detecting large deletions, also identified duplications and multiallelic CNVs included WGS data from 80 accessions from Europe, Asia, and North Africa (Cao et al., 2011). The identified CNVs covered 1.8% of the reference genome and overlapped with nearly 500 genes. Subsequent copy number genotyping of several genes performed by our group using these 80 accessions indicated, however, that the number of genes affected by CNVs may in fact be much higher (Samelak-Czajka et al., 2017). Another study involved the detection of regions of deletions and duplications among 180 accessions, but these accessions represented a narrow local population from Sweden (Long et al., 2013). In these accessions, more than 7700 regions with duplications of a fixed size (3 kb) were identified. A read depth–based approach for CNV detection was used by both Cao et al. (2011) and Long et al. (2013), without further refinement of the CNV breakpoints.

Recently, WGS data from a global collection of 1135 Arabidopsis accessions were released by the 1001 Genomes Consortium et al. (2016), and a catalog of single-nucleotide polymorphisms (SNPs) as well as insertions and deletions shorter than 50 bp (short indels) was created based on these data. Here, we extended the spectrum of characterized genetic variations in these accessions by calling and analyzing large indels and CNVs. We determined the distribution and genomic content of CNV regions and performed population-scale copy number analysis of genes overlapping with CNVs. We investigated the variation in and relative distributions of genes and transposable elements (TEs). We then successfully used gene copy number estimates as markers to reconstruct the genetic structure of the Arabidopsis population. We also demonstrated that natural changes in gene dosage may lead to variations in transcript and protein levels. The CNV map and copy number genotyping data generated in this study provide a background for further studies on the genetic bases of phenotypic variation in Arabidopsis.

RESULTS

Identification of CNVs and Large Indels

We selected 1064 high-quality WGS data sets from the 1135 data sets available in the 1001 Genomes Project collection and performed an integrated CNV analysis (Figure 1A). To this end, we set up a pipeline that combined the three main types of read signatures that can be used for CNV identification (Alkan et al., 2011). We used three read depth–based tools, namely, CNVnator (Abyzov et al., 2011), Control-FREEC (Boeva et al., 2011), and the Genome STRiP-CNV module (Handsaker et al., 2015); two discordant read pair–based tools, namely, BreakDancer (Chen et al., 2009) and VariationHunter (Hormozdiari et al., 2009); the split read–based tool Pindel (Ye et al., 2009); and a hybrid approach implemented in the Genome STRiP-SV module (Handsaker et al., 2015). Methods relying on read depth signatures are the most sensitive in detecting large size variations (Figure 1B) and are more successful when analyzing regions with segmental duplications (Yoon et al., 2009; Teo et al., 2012). However, their accuracy in estimating CNV breakpoints is low (Figure 1C) and depends on the window size used during the calling step. Tools based on discordant read pair mappings are more precise in setting CNV breakpoints but are unable to detect large variants (Supplemental

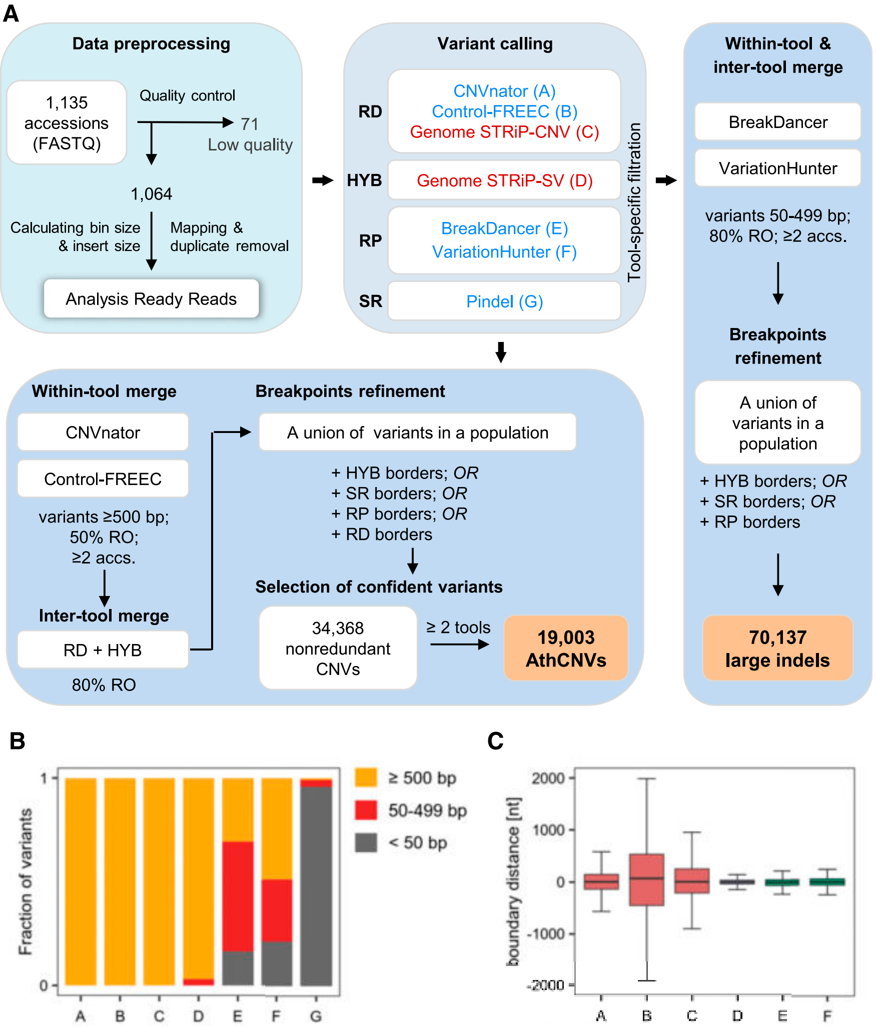


Figure 1. Genome-Wide Structural Variant Discovery in an Arabidopsis Population.

(A) Variant identification pipeline. The analysis involved three main stages: data preprocessing, variant calling, and merging and filtering. Variants were called with seven different tools, based on read depth (RD), read pair (RP), split read (SR), or hybrid (HYB) approach, in individual samples (blue labels) or in the entire population (red labels). The last stage depended on variant length. RO, reciprocally overlapping each other.
(B) Fraction of variants of different size ranges identified by individual callers.
(C) Comparison of the boundaries set by the callers for variants ≥500 bp reciprocally overlapping each other by 80%. Pindel-derived coordinates served as a reference since this tool reports variants at single-nucleotide resolution. Boxplots show median (inner line) and inner quartiles (box). Whiskers extend to the highest and lowest values no greater than 1.5 times the inner quartile range. nt, nucleotides.

Figure 1) or to identify highly duplicated regions. Pindel, which is based on split reads, reports variants at a single-nucleotide resolution but is more sensitive to short indels than to CNVs; additionally, it generates a very large number of predictions with a high false-positive rate (Li et al., 2013). To handle these constraints, we separately processed CNVs (defined here as unbalanced variations at least 0.5 kb in length) and large indels (variants 50 to 499 bp in length).

For CNVs, we selected variants that were detected by at least one read depth–based or hybrid approach (Supplemental Tables 1 and 2). In the next step, whenever possible, we further refined the CNV borders with the additional support of the remaining callers to improve the accuracy of CNV breakpoint predictions. Finally, we included only variants supported by at least two of the seven callers that were used in the list of high-confidence regions that are copy number variable in the Arabidopsis genome, hereafter referred to as the AthCNV data set. This data set consists of 19,003 CNVs that vary in length from 500 to 984,676 bp, 92.1% of which are shorter than 20 kb. These variants are listed in Supplemental Data Set 1, along with 15,365 low-confidence CNVs, which were supported by only one caller and were not further investigated.

We identified large indels by combining 50- to 499-bp-long variants from the read pair–based callers only, followed by redundancy removal, and set boundaries with the support from hybrid- and split read–based callers. As a result, we obtained 70,137 variants (Supplemental Data Set 2). Of these, 4149 exceeded the upper size limit defined in our pipeline as a result of merging and breakpoint refinement. We did not remove them from the final large indel data set since they were identified using a different approach from AthCNVs. Overall, large indels had 56% overlap with AthCNVs.

We then compared the genomic distribution of the newly identified variants with that of the previously identified short indels (1001 Genomes Consortium et al., 2016). All types of variants (short indels, large indels, and AthCNVs) were most abundant in the pericentromeric regions and less abundant in the chromosome arms (Figure 2). However, short indels had moderate overlap with AthCNVs (46%) and very little overlap with large indels (8%). Thus, our results substantially complement the existing catalog of known structural variations present in the Arabidopsis genome.

In the subsequent analysis, we focused on CNVs since this class of variants—due to their size—may directly influence the copy number and dosages of entire functional loci, including genes.

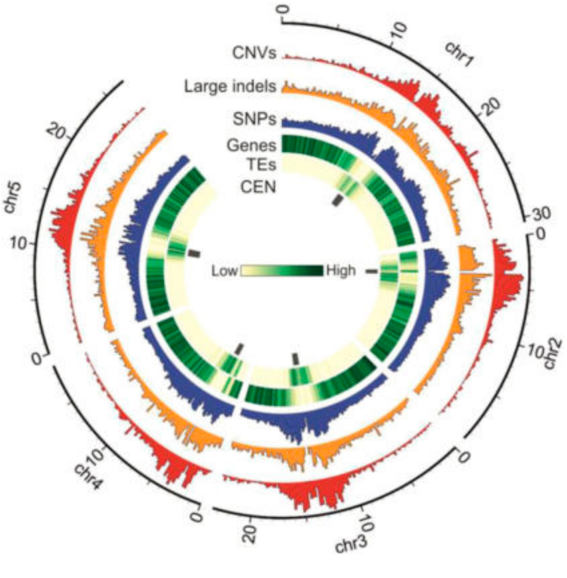


Figure 2. Genomic Distribution of CNVs, Large Indels, and Short Variants in the Arabidopsis Genome.

Histograms are scaled for equal height. Tracks present: CEN, pericentromeric regions; CNVs, confident CNVs discovered in this study; Genes, protein coding genes; Large indels, variants 50 to 499 bp discovered in this study; SNPs, SNPs and short indels from 1001 Genomes Project; TEs, annotated TEs.

Since our data analysis pipeline involved two CNV merging steps (between samples and between tools) that preceded the breakpoint refinement step, we attempted to verify the sensitivity and accuracy of our approach at three levels: species, geographically related accessions, and individual genomes (Figure 3A and 3B). For species-level verification, we used CNVs previously identified in a population of 80 accessions that represented a similar geographic range and were not included in our data set (Cao et al., 2011). Of the 1059 CNVs identified in that study, 87% overlapped with AthCNV regions and 81% were positioned entirely within them. This result was in line with our expectations, since the previously identified CNVs were much shorter.

For verification at the level of geographically related accessions, we evaluated the overlap of the AthCNV data set with the duplications and deletions previously detected in 180 Swedish accessions (Long et al., 2013), 174 of which were also included in our analysis. After merging directly adjacent regions with duplications and removing private variants (since they were also filtered out by our CNV discovery pipeline; see Methods), we obtained 235 deletions and 1487 duplications ≥0.5 kb in length in the Swedish samples. We observed that 76% of deletion regions overlapped with the AthCNVs, and 51% were positioned entirely within them. Likewise, 68% of duplication regions overlapped with AthCNVs, and 50% were located entirely within them.

Finally, we investigated how well the AthCNV data set fit the variants identified in eight genomes representing individual accessions. One genome (KBS-Mac-74 accession) has been assembled to the contig level from Nanopore ultralong reads (Michael et al., 2018). We used the Assemblytics tool (Nattestad and Schatz, 2016) to identify CNVs in this genome (Supplemental Data Set 3). The seven remaining genomes (An-1, C24, Cvi-0, Eri-1, Kyoto, Ler, and Sha accessions) were assembled into five chromosome-level scaffolds from PacBio ultralong reads, and structural variants were identified with the SyRI tool (Jiao and Schneeberger, 2020). Note that both SyRI and Assemblytics rely on the same genome aligner, MUMmer. We selected CNVs ≥0.5 kb in length (the reference genome coordinates were considered in the size evaluation) and compared them with our data set.

In each accession, the majority of CNVs (91 to 99%) were shorter than 20 kb, similar to the AthCNVs. From 88 to 94% of the CNVs in each accession overlapped with the AthCNVs by at least 1 bp. As many as 63 to 77% deletions, but only 22 to 25% duplications overlapped with individual AthCNVs by at least 70% and therefore had similar lengths and breakpoint locations (Supplemental Figure 2). We also observed that the AthCNVs for which the breakpoints best fit the breakpoints of variants found in individual genomes, that is, the localization of one of their borders (left or right) differed by no more than ±10 bp (Figure 3C), were mostly refined using the split reads and hybrid approach (91 to 94%) or the discordant read pair approach (7 to 9%). These observations validate the approach we used to assess CNV borders (the highest priority was given to the information provided by the callers based on discordant read pairs and split reads) and explained the lower accuracy of assessing duplication breakpoints. Taking the above-mentioned information into account, the AthCNV map reliably represents variants present in individual accessions.

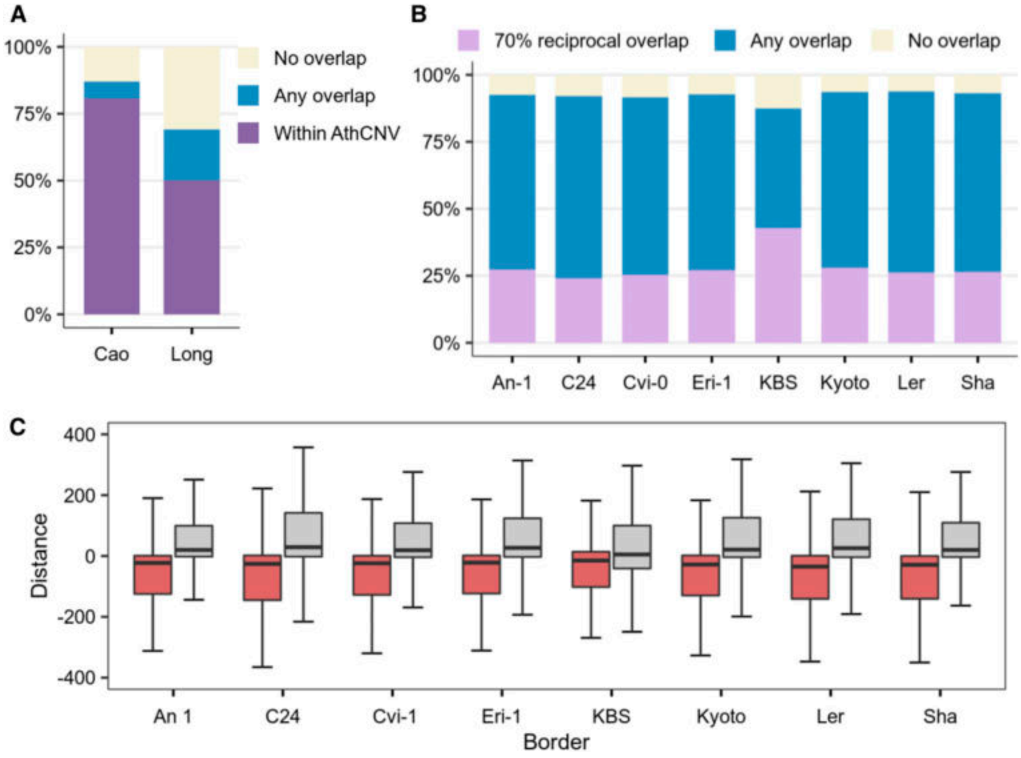


Figure 3. Overlap of the AthCNV Data Set with Variants Identified in Small Populations and Individual Genomes.

(A) Fractions of CNVs identified previously in a small, worldwide population of 80 accessions (Cao data set) and a narrow population of Swedish accessions (Long data set) that overlap with AthCNVs.

(B) Fractions of CNVs detected in the genomes of individual accessions assembled de novo from long reads that overlap with AthCNVs.

(C) Relative distances between the breakpoints in the AthCNVs and the breakpoints in CNVs in eight accessions (each used as a reference for AthCNV distance calculation). Boxplots depict data for pairs of variants with ≥70% reciprocal overlap. Boxplots show median (inner line) and inner quartiles (box). Whiskers extend to the highest and lowest values no greater than 1.5 times the inner quartile range.

We also performed literature mining and found 106 genes for which complete or partial duplications/deletions have been reported and—as an obligatory criterion—experimentally confirmed in Arabidopsis (Supplemental Data Set 4; Grant et al., 1995; Stahl et al., 1999; Xiao et al., 2001; Kroymann et al., 2003; Werner et al., 2005; Balasubramanian et al., 2006; Clark et al., 2007; Staal et al., 2008; Vlad et al., 2010; Smith et al., 2011; Bloomer et al., 2012; Cole and Diener, 2013; Karasov et al., 2014; Vukašinović et al., 2014; Pucker et al., 2016; Zmienko et al., 2016; Samelak-Czajka et al., 2017; Michael et al., 2018). We found that 100 genes overlapped with AthCNVs (Supplemental Figure 3). Four additional genes overlapped with low-confidence variants, which were also detected by our CNV discovery pipeline. Thus, our data are highly consistent with the existing experimental evidence on the distribution of CNVs in the Arabidopsis genome.

Genomic Content in CNV Regions

We observed uneven genome coverage by CNVs (Table 1). From 84 to 99% of the centromeric regions were covered by CNVs, with multiple CNVs of various lengths overlapping with each other (Supplemental Figure 4). In Arabidopsis, the centromeres are rich in 178- to 180-bp repeats and TEs (Minoru, 2013). Additionally, in the noncentromeric parts of the genome, the distribution of CNVs was positively correlated with the distribution of TEs and negatively correlated with the distribution of the genes. Nevertheless, a very large number of genes (7712) overlapped with CNV regions. We hereafter refer to genes and TEs covered by AthCNVs by at least 1 bp as CNV-genes and CNV-TEs, respectively, to distinguish them from NONVAR-genes and NONVAR-TEs, which did not overlap with any CNVs. We then investigated more deeply the fraction of CNV-genes that were covered by CNVs for ≥90% of their length (Figure 4A). These genes were highly represented by orphan genes, that is, genes with no detectable homologues in any other species (497 of the 1170 orphan genes present in the Arabidopsis genome) and species-specific gene families (49 of the 55 families found only in this species; Figure 4B; Supplemental Table 3). They were also significantly overrepresented in genes encoding proteins of an unclassified type (binomial test with Bonferroni-corrected P-value < 0.01; Figure 4C). Similarly, we

Table 1. Arabidopsis Genome Coverage by the Identified CNVs

Region Type	No. of Variants	Mean Coverage (%) of the Given Region Type[a]	Average No. of Variants in Overlapping Segments[b]
Genome	19,003	35.7	3.8
Centromeres	6,584	93.5	7.2
Outside centromeres	12,419	28.0	2.4
Overlapping protein-coding genes	6,326	18.5	1.7
Overlapping pseudogenes	943	59.6	2.6
Overlapping TEs	8,548	94.0	3.7

[a] Calculated from the following formula: coverage in individual region of a given type = number of bases overlapped by any CNV/number of all bases in this region × 100%; average value is reported in the table.

[b] Calculated as the number of CNVs overlapping each region in 1-bp windows. Average number is reported in the table. To remove the bias resulting from different overall coverage of various region types, only the positions with nonzero overlap were counted, for example, for a 1000-bp pseudogene overlapped by several CNVs in a total of 46% of its length; the number of overlapping variants was counted for 460 1-bp intervals covered by any CNV and averaged.

observed significant overrepresentation in CNV-genes that are unclassified based on the Molecular Function, Biological Process, and Cellular Component Gene Ontology (GO) terms. In addition, terms related to plant interactions with other organisms, defense, and stress responses were overrepresented in each category. There were no significantly depleted GO terms, but genes encoding nucleic acid binding proteins, transporters, transferases, and protein kinases were significantly underrepresented in the CNV-genes data set.

A recent comparative study of seven Arabidopsis genomes assembled de novo from long reads revealed multiple regions with strongly decreased collinearity and multiple haplotypes (Jiao and Schneeberger, 2020). These regions were referred to as hotspots of rearrangements and were enriched in TEs and depleted in genes, similar to the CNVs identified in our study. Additionally, similar to our CNV-genes, the genes within the hotspots of rearrangements were enriched for functions related to biotic stress response. In addition, they displayed high CNV and high mutation frequency among the seven accessions. We therefore expected them to be identified as population-level CNVs in our study. Indeed, we found that 98.6% of rearrangement hotspots overlapped with AthCNVs (73.6% were entirely within CNV regions). Of the eight regions without overlap, two were near AthCNVs (less than 250 bp), and four formed a large cluster with numerous adjacent hotspots of rearrangements, which extended for over 212 kb and was flanked by multiple CNVs on both sides. Many hotspots of rearrangements shared a common pattern of almost exclusively forward tandem gene duplications and large indels (Jiao and Schneeberger, 2020), which prompted us to investigate whether AthCNVs were also enriched in tandem duplications. According to the Plaza 4.0 database (Van Bel et al., 2018), 25.3% of genes in the Arabidopsis genome are located in regions of segmental duplications, while 12.8% arose through tandem duplication events (additionally, 8.3% are located in regions with both segmental and tandem duplications). These proportions were reversed among CNV-genes, with 12.4% of these genes localized in regions of segmental duplications and 24.1% in regions of tandem duplications (additionally, 10.7% underwent both segmental and tandem duplications; Figure 4D). Altogether, these observations indicate that the regions of tandem duplications are sites that accumulate rearrangements and, consequently, show high structural diversity.

In the next step, we analyzed CNV-TEs, which constituted 67.5% of all TEs. These TEs were slightly depleted in RC/Helitron TEs and enriched in long terminal repeat/Gypsy TEs (Figure 4E); however, the composition of CNV-TE superfamilies did not change much compared to all TEs (Supplemental Table 4). We also investigated how many CNV-TEs were proximal to genes, that is, overlapped with genes or were located within 2-kb regions flanking the genes. Only 36.2% of CNV-TEs were proximal to genes, and they were slightly enriched in RC/Helitron TEs but severely depleted in long terminal repeat/Gypsy TEs compared to both all CNV-TEs and the entire genome. They were also moderately enriched in DNA/MuDR elements. In contrast to CNV-TEs, genes with TEs in their proximity constituted the majority (64.4%) of the CNV-gene data set.

Interplay between the Copy Number Polymorphism of Genes and TEs

To investigate the relationship between the copy number polymorphism of genes and TEs, we compared the genomic distributions of CNV-genes and CNV-TEs. Both CNV-genes and CNV-TEs were, on average, located closer to the chromosome centromeres than were their NONVAR counterparts, and this tendency was much stronger for TEs than for genes (Figure 5A). However, the average distance between CNV-genes and the nearest TEs was smaller than the average distance between NONVAR-genes and the nearest TEs. The reverse was observed for CNV-TEs, which were, on average, farther from the nearest gene than were NONVAR-TEs (Supplemental Figure 5). Our observations indicated that some selective forces have opposite effects on shaping the relative distribution patterns of CNV-genes and CNV-TEs. The cut-insert and copy-insert mechanisms underlying TE mobility may affect adjacent genes, usually in a negative manner, for example, by interrupting gene coding or regulatory sequences, by gene rearrangement and duplication, or by altering their DNA methylation status (Quadrana et al., 2016; Bourque et al., 2018). Gene proximity may therefore be considered a negative force acting against nearby TE transposition, especially

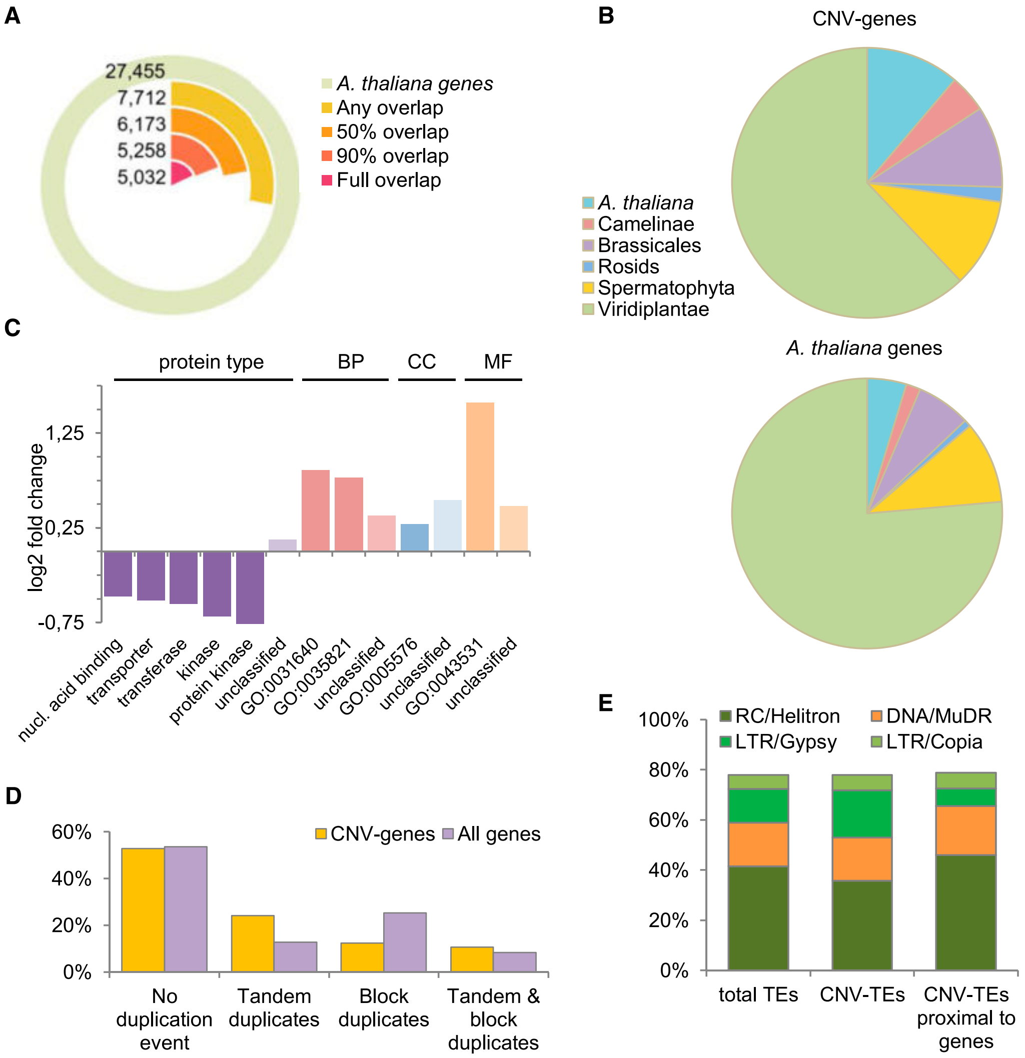


Figure 4. Genomic Content in Regions Overlapped by AthCNVs.

(A) Fractions of annotated Arabidopsis genes with various degrees of overlap with AthCNV variants.
(B) Enrichment of CNV-genes that are overlapped by AthCNVs by at least 90% in the fractions of species-specific and clade-specific genes compared to that of all annotated Arabidopsis genes.
(C) Over- and underrepresented protein types and GO terms among the CNV-genes, in the Biological Process (BP), Cellular Component (CC), and Molecular Function (MF) categories. All terms are either significantly enriched or depleted (binomial test with Bonferroni-corrected P-value < 0.01). The GO terms shown in the chart are killing of cells of other organism (GO:0031640), modification of morphology or physiology of other organism (GO:0035821), extracellular region (GO:0005576), and ADP binding (GO:0043531). nucl., nucleic.
(D) Locations of CNV-genes in regions of tandem and block duplications in the genome compared to those of all genes.
(E) Superfamily composition of Arabidopsis TEs and its comparison with all CNV-TEs and gene-proximal CNV-TEs (located within ±2-kb distance). Top-four most abundant superfamilies are presented. Class I TEs are depicted in orange; class II TEs are in different shades of green. All families are listed in Supplemental Table 4. LTR, long terminal repeat. RC, rolling cycle.

in the case of genes involved in crucial metabolic processes. On the other hand, TE proximity may contribute to increased copy number polymorphism of nearby genes by inducing DNA breaks and genomic instability.

To extend our observations to all genes, we analyzed the distances and compared the CNV statuses of genes and their proximal TEs. We found strong enrichment in pairs where proximal TEs and genes had the same variation statuses (Figure 5B),

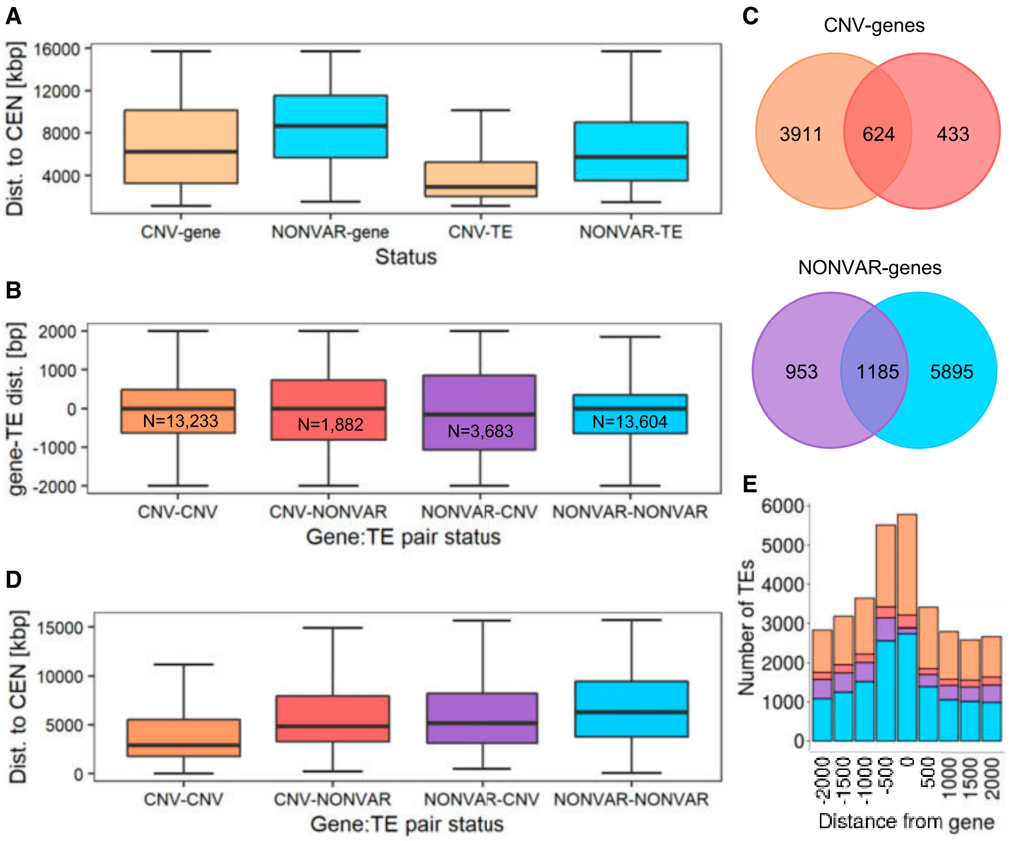


Figure 5. Links between Genes and TE Variation and Localization.

(A) Distance to centromeres of genes and TEs grouped by variation status (determined based on their overlap with AthCNVs). The groups were significantly different (Wilcoxon rank sum test with continuity correction, $P < 0.0001$). Genetic elements localized in the pericentromeric regions were not included. dist., distance.

(B) Relative distances between genes and their proximal TEs, grouped by variation status. For each gene, a proximal TE was defined as each TE overlapping with this gene (distance = 0) or overlapping region located within 2 kb upstream from the gene's 5′ untranslated region (distance < 0) or overlapping region located within 2 kb downstream from 3′ untranslated region (distance > 0). N, number of pairs with a given variation status. dist., distance.

(C) Number of unique CNV-genes and NONVAR-genes with proximal CNV-TEs and NONVAR-TEs and their overlap.

(D) Gene distances to centromeres presented for gene-TE pairs differing by variation status. dist., distance.

(E) Number of proximal TEs within and around genes. Colors in **(B)** to **(E)** are identical for the same groups. Boxplots in **(A)**, **(B)**, and **(D)** show median (inner line) and inner quartiles (box). Whiskers extend to the highest and lowest values no greater than 1.5 times the inner quartile range.

regardless of whether they were both polymorphic (40% pairs) or invariable (42% pairs). Furthermore, 3911 of 4968 unique CNV-genes (79%) had only CNV-TEs in their proximity and 5895 of 8033 unique NONVAR-genes (73%) had only proximal NONVAR-TEs (Figure 5C). Additionally, the gene-TE pairs with the same variation statuses were located closer to each other than pairs with the opposite statuses. Combining the information about the genomic distribution and relative distances of genes and TEs clearly revealed that the localization of polymorphic gene–TE pairs was biased toward centromeres, while the localization of invariable gene–TE pairs was biased toward chromosome ends (Wilcoxon rank sum test with continuity correction for the difference between CNV-CNV and NONVAR-NONVAR groups, P-value < 0.0001; Figure 5D). Moreover, CNV-genes with proximal CNV-TEs were enriched in extracellular proteins and proteins involved in cell disruption, defense responses, and nucleic acid catabolism (Supplemental Data Set 5). At the same time, NONVAR-genes with proximal NONVAR-TEs were enriched in nuclear proteins and proteins involved in nucleic acid metabolism, regulation of fertilization, and transcription factor activity. There was no difference in the chromosomal distribution of pairs displaying opposite variation statuses, and no or few GO terms were enriched in these two groups.

Interestingly, the combined variation status of gene–TE pairs was also apparently related to the position of TEs relative to nearby genes (Figure 5E). All TEs localized in proximity to genes were 1.2

to 1.4 times more often inserted in their upstream flanking regions compared to downstream flanking regions. CNV-TEs very rarely overlapped with NONVAR-genes (3.8% cases) compared to CNV-genes (19.4%) or NONVAR-TEs, which overlapped with both NONVAR-genes and CNV-genes at similar frequencies (20.2 and 17.4%, respectively). The four groups had similar TE family compositions, which indicated that these differences were not caused by insertion bias of any specific TEs. Altogether, our observations confirmed the presence of selective constraints reciprocally imposed on genes and TEs, which is an important factor contributing to their present variation and genomic distribution patterns.

Copy Number Genotyping and Experimental Evaluation of CNV-Genes

After we identified the genomic regions showing copy number polymorphism in Arabidopsis, we used the Genome STRiP SVGenotyper module (Handsaker et al., 2015) to evaluate the copy number statuses of CNV-genes in individual accessions based on read depth estimates. Based on our earlier observations, we decided to directly evaluate the copy numbers of the genes covered by AthCNVs (using the gene coordinates as the input) instead of the AthCNVs themselves. Our motivation was to simplify the subsequent application of the copy number genotyping data in functional analyses. AthCNVs overlapping with each other may have been formed by different molecular mechanisms and may be present in different accessions (Zmienko et al., 2016); however, at the population scale, they collectively contributed to the copy number diversity of the CNV-genes that they covered (Supplemental Figure 6). Accordingly, we observed that the direct genotyping of CNV-genes provided the most accurate information about their copy number statuses in individual accessions. We ultimately genotyped 7324 CNV-genes as well as—for comparison purposes—5060 genes overlapped by low-confidence CNVs and 14,661 NONVAR-genes in 1060 accessions. These data can be accessed through the web interface at http://athcnv.ibch.poznan.pl in the form of user-generated plots.

Genome STRiP SVGenotyper is capable of assigning integer copy numbers to genotyped regions. We found, however, that it frequently assigned the copy number classes to intervals of only one copy; because Arabidopsis is a predominately selfing species, the expected differences between copy number alleles were multiples of two (Supplemental Figure 7). The integer copy number assignment by Genome STRiP SVGenotyper was also disturbed by the presence of CNV-genes that did not form clear, discrete copy number classes or for which the reported copy number was very high (up to many thousands of copies) in most accessions, including Arabidopsis ecotype Columbia (Col-0), which was expected to have the reference diploid copy number (two copies) for each gene. Such problems were commonly encountered when genotyping complex CNVs and CNVs that were mapped to segmental duplications (Conrad et al., 2010; Campbell et al., 2011; Handsaker et al., 2015). For these reasons, we reported unrounded rather than integer copy number outputs. Additionally, we filtered the genotyping data by excluding genes with extreme copy numbers in Col-0 separately for each of the three data sets. In this step, we removed 451 genes from the analysis.

The global distributions of the copy number estimates obtained for CNV-genes significantly differed from those obtained for NONVAR-genes, which were more uniform (interquartile range for NONVAR-genes was 0.23 versus 0.30 for CNV-genes) and much more concentrated around the reference diploid copy number value (kurtosis = 13 for NONVAR-genes versus 120 for CNV-genes). Moreover, CNV-genes had significantly higher copy number variance, larger copy number ranges, and more extreme maximum and minimum copy number values than did NONVAR-genes (Figure 6). Genes covered by low-confidence CNVs had intermediate values, but overall, they were more similar to NONVAR-genes than to CNV-genes.

For 1777 (25.3%) CNV-genes, we observed an unexpectedly small level of variation: for these genes, the copy number difference between any two accessions in the population was <2. One reason for the low level of variation in these CNV-genes was their partial overlap with AthCNVs. In these cases, the reads that mapped to the invariable gene segments contributed to read depth estimates, reducing the observed differences between the accessions with distinct copy number statuses (Supplemental Figure 8). Therefore, for all subsequent analyses, we selected only the 5517 CNV-genes that had ≥50% overlap with AthCNVs. This

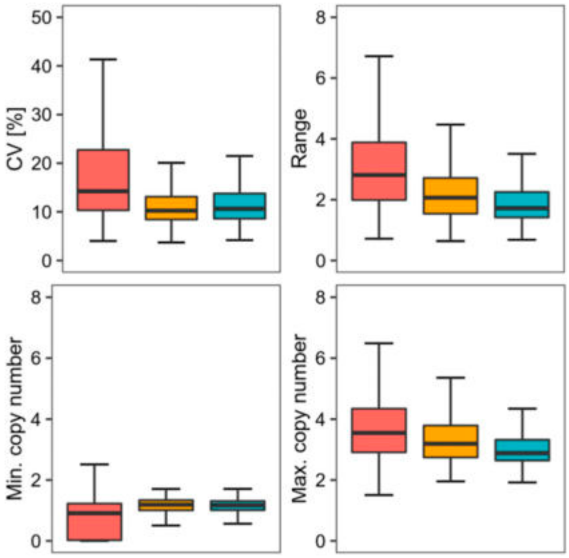


Figure 6. Differences between CNV-Genes, NONVAR-Genes, and Genes Covered by Low-Confidence CNVs in Terms of the Read Depth–Based Copy Number Genotypes.

The genotyping data for 7031 CNV-genes (red), 4482 low-confidence CNV-genes (orange), and 14,877 NONVAR-genes (blue) were compared for four attributes: the coefficient of the CNV (CV; top left), the copy number range in a population represented by 1060 accessions (top right), and the minimum (min.) and maximum (max.) copy number values (bottom left and bottom right, respectively). For each attribute tested, CNV-genes significantly differed from the other groups (Kruskal–Wallis test, P < 0.0001, Dunn–Bonferroni post hoc method P-value < 0.0001). Boxplots show median (inner line) and inner quartiles (box). Whiskers extend to the highest and lowest values no greater than 1.5 times the inner quartile range.

reduced the percentage of CNV-genes with low variation to 17.7%. To further investigate the possible reasons for their low variation, we assigned each CNV-gene to its longest overlapping AthCNV and found that all CNV-genes with little variation in copy number were contained in only 332 AthCNVs. Moreover, 228 of these AthCNVs also encompassed CNV-genes with high CNV (Supplemental Figure 9). This result suggested that some AthCNVs included small nonvariable subregions, presumably not identified during the segmentation step. We further observed that the presence of this mosaicism was related to AthCNV size—CNV-genes with little variation in copy number were covered by very long AthCNVs, with a median size of 183.4 kb. For comparison, the median size of AthCNVs covering CNV-genes with high CNV was 19.9 kb.

We further verified the accuracy of our read depth–based copy number estimates by performing multiplex ligation-dependent probe amplification (MLPA) assays using 314 accessions (i.e., 30% of the genomes genotyped with Genome STRiP). The experiment involved CNV-genes located in 45 nonoverlapping AthCNVs (Supplemental Figure 10) and four NONVAR-genes. While read depth–based genotyping provided copy number estimates for entire CNV-genes, by disregarding factors such as incomplete overlap with AthCNVs, the fine-scale MLPA approach focused on small (<75-nucleotide) target regions within the assayed genes, which made it more precise but also more sensitive to the presence of local sequence variations such as SNPs and indels. After taking these factors into account, we were able to explain most of the discordant results observed in our experiment by the presence of sequence variation in MLPA probe binding sites in the assayed accessions (Supplemental Figures 11 and 12). Overall, the MLPA-based genotyping results were in agreement with the read depth–based estimates for all assayed genes (Supplemental Figures 13 to 15). For numerous multiallelic CNV-genes, the clusters of samples with the same copy number could be clearly distinguished by plotting the read depth–based data against the MLPA data (Figure 7).

Interestingly, the MLPA analysis provided another, although unexpected, piece of evidence supporting the accuracy of our read depth–based genotyping results. Initially, we included 346 accessions in the MLPA assays. However, 32 of them were recently reported as potentially mislabeled in public seed repositories (from which we acquired our seed collection) based on resequencing and SNP analyses, which failed to assign these stocks to the expected strains (Pisupati et al., 2017). In agreement with these findings, we observed a very strong negative effect of these 32 samples on the correlation between the read depth–based and MLPA results (Supplemental Figure 16; Supplemental Table 5). Consequently, we removed them from the MLPA analysis.

Arabidopsis Population Structure Revealed by CNV Markers

The analysis of SNP markers in the 1001 Genomes Project accessions revealed that 95% of Arabidopsis accessions belong to one genetic group composed of several subgroups of accessions sharing a similar geographic origin (Platt et al., 2010; 1001 Genomes Consortium et al., 2016). The remaining 5% of accessions (referred to as relicts) form a few groups that are genetically distant from each other and from the nonrelicts (Lee et al., 2017). We aimed to infer Arabidopsis population structure from CNV markers and verify its consistency with the structure derived from SNP markers. We selected 1050 AthCNVs of various types (deletions, duplications, and multiallelic CNVs) distributed across the genome and used the copy numbers of the representative CNV-genes (one gene per AthCNV) as input for principal component analysis (PCA). We then compared our results to population structure derived from 1001 Genomes Project SNP markers. The first two principal components (PCs) revealed that the population is highly structured and that the accession groupings reflect their geographical distribution (Figure 8A), which is consistent with the SNP-based groupings (Cao et al., 2011; Horton et al., 2012). SNPs better distinguished the genetic subgroups than did the CNVs, which was an expected result, as the subgroups were defined based on SNP variation, and SNPs substantially outnumbered CNV markers (1001 Genomes Consortium et al., 2016). However, the CNV-based analysis better reflected the global distribution of the accessions (the directions of the accessions' separation were consistent with geographical directions, north to south for PC1 and east to west for PC2, after removing clearly unique U.S. accessions; Figure 8B).

Interestingly, CNV-based PCA revealed some similarities between the accessions that were not captured by SNP-based grouping. The third and fourth PCs distinguished the groups from the edges of the natural species range and highlighted the genetic similarity of the northern Sweden accessions to the relict genomes from southern Europe (Figure 8C). Remarkably, this observation is in agreement with the recently proposed two-wave expansion model of Arabidopsis across Eurasia, derived from the analysis of the extent of relict introgression in the nonrelict genomes (Lee et al., 2017). According to this model, the populations from different glacial refugia (relicts) expanded from the south of Europe northward at the end of the last ice age. Subsequently, the ancestors of today's nonrelicts expanded along the east–west axis, probably from the Balkans or the Black Sea area, and replaced the local accessions, except in the north and south of the species range, where large introgressions from the relict genomes (locally adapted) might have helped the nonrelicts colonize the habitats with more severe climatic conditions.

We then compared the extent of CNV-gene copy number changes between 1059 accessions (Col-0 was excluded from this analysis). To this end, we treated all copy number genotypes ≤1 as losses, all copy number genotypes >3 as gains, and all the remaining genotypes as unchanged. These thresholds were justified because the median copy number value for all accessions and all CNV-genes analyzed was 1.98. On average, copy number losses were more frequent in all subgroups (the mean gain-to-loss ratio was 0.5), and their amount differed among the subgroups to a greater extent than did that of copy number gains (Figure 9A). The subgroups least affected by CNVs were Germany (8.2%) and Central Europe (8.6%), while the relicts (11.2%) and northern Sweden (10.0%) subgroups were most affected. This order was in good agreement with the general similarity of the subgroups to the reference genome (the Col-0 accession was assigned to the Germany group) but also confirmed the general rule that the choice of a reference genome is a crucial step that determines the range of variation that may be identified by a mapping-based approach.

In individual accessions, 3.9 to 26.9% of CNV-genes were affected by copy number changes (Figure 9B), and this broad range was mostly caused by the differences in the number of gains (ranging from 88 to 1068) and, to a lesser extent, by the losses (ranging from 114 to 660). The top five accessions in terms of total copy number changes were also the top five in terms of the number of gains and had a gain-to-loss ratio ranging from 0.93 to 2.77. Two of the accessions were from Sweden (Ull2-5 and Sanna-2), while the remaining accessions were U.S. accessions (KBS-Mac-74, KBS-Mac-68, and BRR57).

Gene Dosage, Gene Expression, and Missing Duplications in the Reference Genome: *SEC10* Example

Duplication of *AT5G12370*, encoding the SEC10 protein involved in exocytotic vesicle fusion, was recently discovered in the Col-0 accession (Vukašinović et al., 2014). The *SEC10* duplication is absent from TAIR10 version of the Arabidopsis reference genome (the reference sequence is a chimera of both copies). To determine whether other gene duplications occur in Col-0, we manually searched the genotyping results for the CNV-genes excluded by

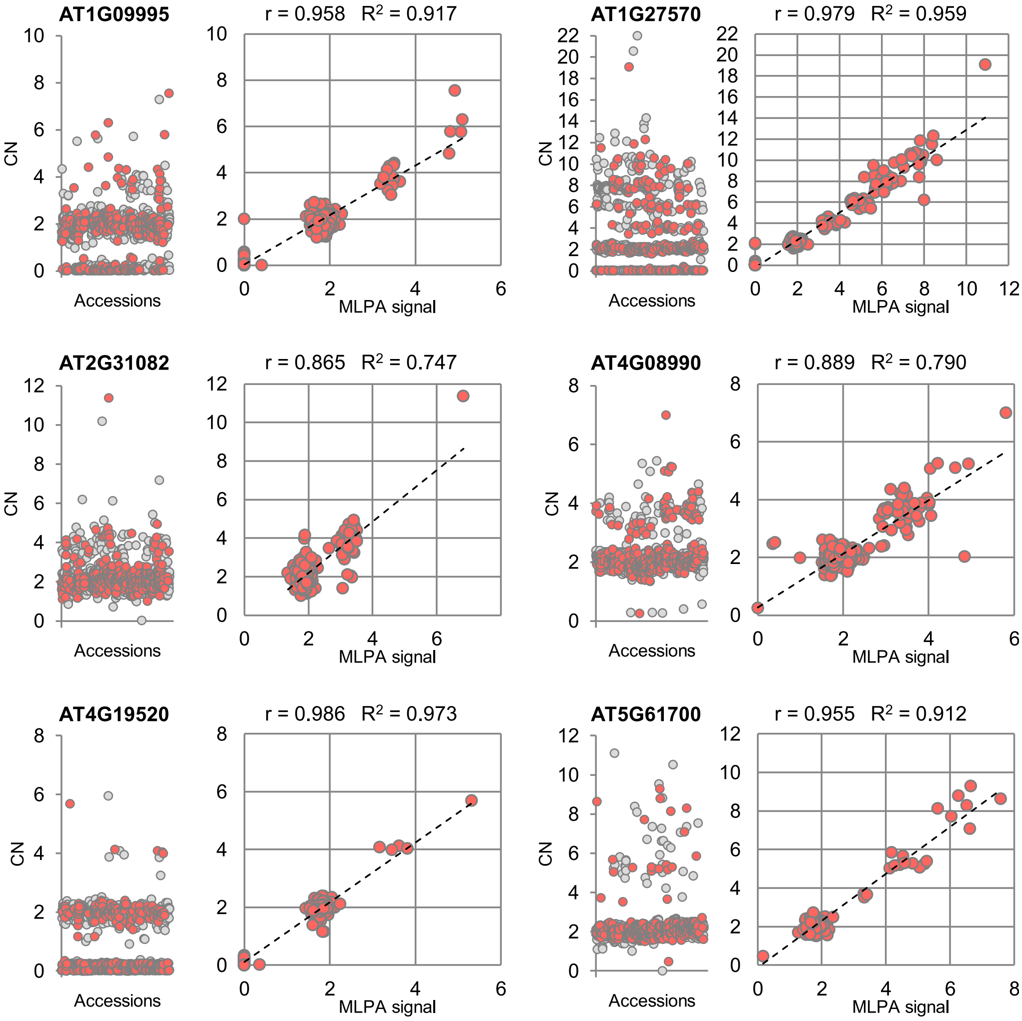


Figure 7. Experimental Validation of Read Depth–Based Copy Number Genotyping Results.

For each CNV-gene, two scatterplots are presented: read depth–based copy numbers (CN) for 1060 accessions (left) and the correlation of the genotype data with the MLPA results for 314 accessions (right). The same set of accessions was used in all MLPA experiments, which are labeled in red in the plots on the left. The MLPA results were scaled for each CNV-gene using Col-0 signal as a reference value (CN = 2). R, Pearson correlation coefficient; R^2, coefficient of determination of linear regression.

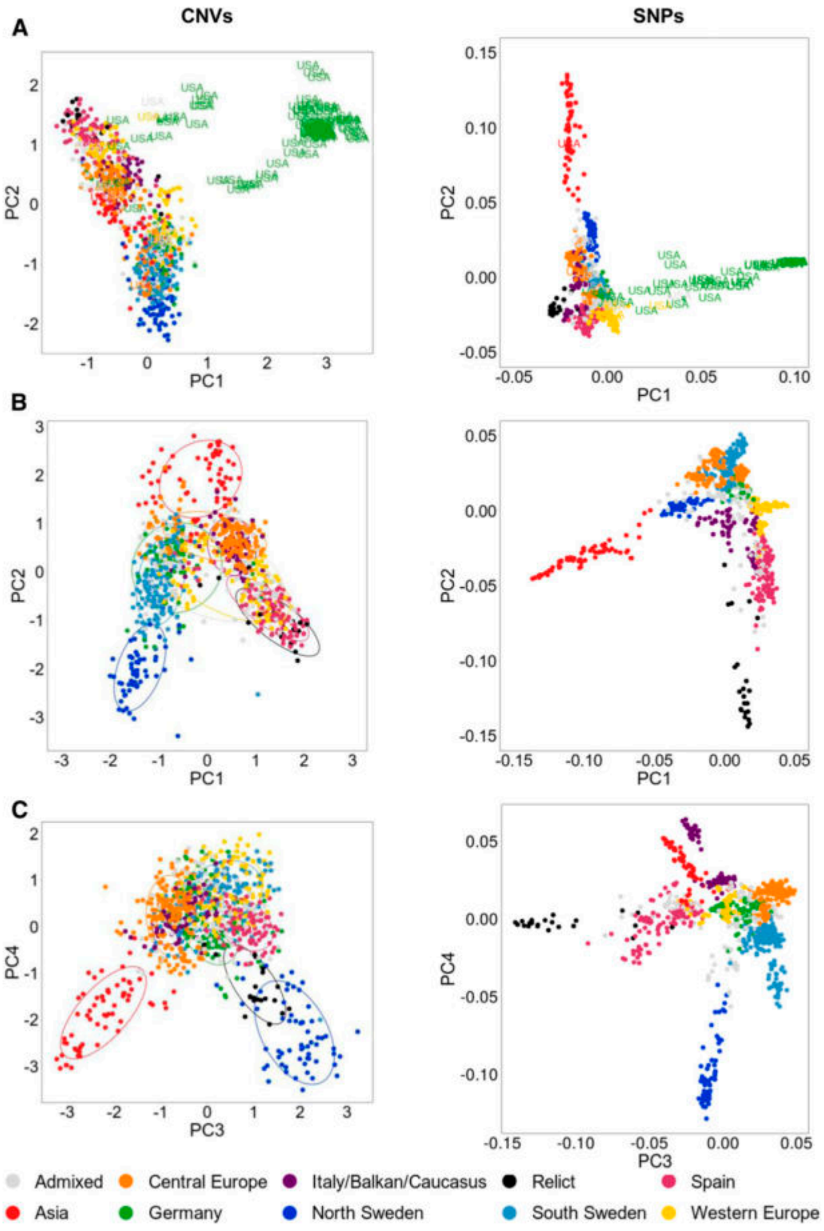


Figure 8. Arabidopsis Population Structure Based on the Analysis of CNV Genotypes.

PCA was performed on 1060 accessions and on genotyping data from 1050 CNV-PCGs (left). For comparison, another PCA was performed on the same set of accessions and 117,232 SNPs from the 1001 Genomes Project (right).

(A) PC1 and PC2 components; all accessions were included. U.S. accessions assigned to the Germany subgroup were distinguished from the other samples.

(B) PC1 and PC2 components; U.S. accessions from the Germany subgroup were excluded from the analysis.

(C) PC3 and PC4 components; all accessions were included. The accessions in PCA plots are colored based on their 1001 Genomes Project grouping.

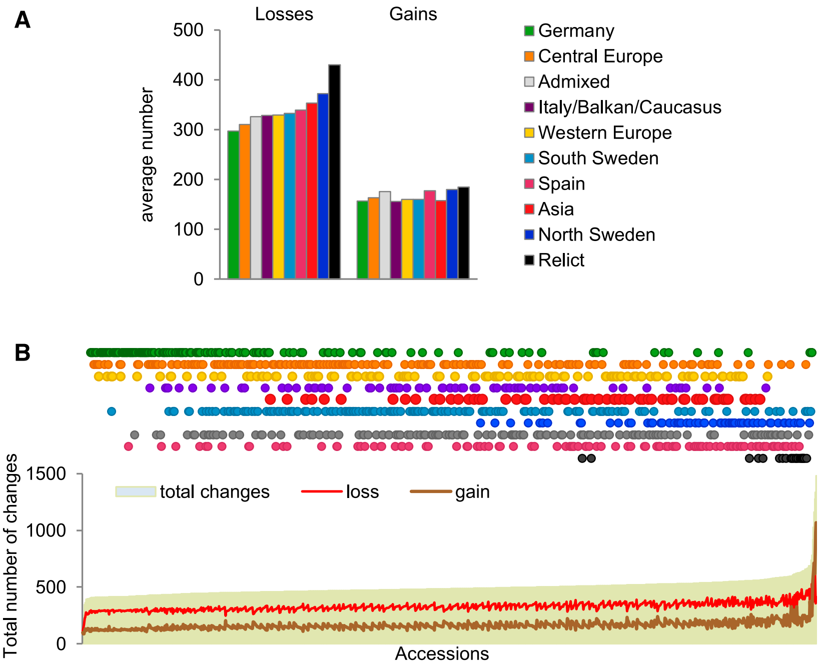


Figure 9. Losses and Gains in Gene Copy Number in Arabidopsis Subgroups.

(A) Average number of gene copy number gains and losses in the subgroups.
(B) Total number of gene copy number changes in individual accessions.

our interquartile range–based filter. As a result, we identified eight candidates that were possibly duplicated in Col-0, including *SEC10* (Supplemental Figure 17). Our genotyping results indicated that the *SEC10* duplication was prevalent in the Arabidopsis population, as four, six, and eight copies were detected in the diploid genomes of 1039 accessions, 14 accessions, and 1 accession, respectively, while two copies were detected in only 6 accessions (0.56%; Figure 10A). We also evaluated *SEC10* expression in 601 accessions using available RNA sequencing (RNA-seq) data (Kawakatsu et al., 2016) and observed that the transcript levels increased in samples with elevated *SEC10* copy numbers (Figure 10B). To determine whether these differences were also reflected at the protein level, we analyzed the SEC10 protein content in 12 accessions representing genotypes with two, four, or six copies of *SEC10*. Indeed, the mean protein level was significantly higher in accessions with four *SEC10* gene copies than in those with two copies (Figure 10C; Supplemental Figure 18). It was also elevated in two of three accessions with six copies compared to samples with no *SEC10* duplication.

Genome-Wide Association Study of CNVs

Several studies have provided evidence that CNVs account for a substantial amount of phenotypic variation. In particular, presence-absence polymorphism of resistance genes that are involved in race-specific recognition of pathogen avirulence determinants (McHale et al., 2006) contributes to plant resistance phenotypes. In Arabidopsis, CNVs affect numerous loci related to biotic responses, including *RPM1*, *RPS5*, *RLM1*, *RLM3*, *RPP1*, *RPP5*, and *RPP7* (Grant et al., 1998; Henk et al., 1999; Yi and Richards, 2009; Roux and Bergelson, 2016). A previous genome-wide association study revealed strong SNP associations for four hypersensitive response phenotypes to *Pseudomonas* elicitor proteins: *AvrPphB*, *AvrB*, *AvrRpm1*, and *AvrRpt2* (Atwell et al., 2010). Single candidate loci encoding known resistance genes could be associated with these SNPs: *RPS5* for *AvrPphB*, *RPM1* for *AvrB* and *AvrRpm1*, and *RPS2* for *AvrRpt2*. According to our results, *RPS2* is not a CNV-gene; therefore, the association for this gene likely resulted from small-scale variation. We wanted to find out, however, whether the remaining two genes, for which the impact of gene deletion on pathogen resistance has been confirmed previously (Grant et al., 1998; Stahl et al., 1999, Karasov et al., 2014), could be directly distinguished in an association analysis using our genotyping data. To test this possibility, we selected 23 defense-related phenotypes from the Atwell et al. (2010) study, including the four hypersensitive response phenotypes mentioned above (Supplemental Data Set 6). This medium-sized data set consisted of 76 to 175 accessions per phenotype, 51 to 117 of which were shared with our study. Using CNV-gene statuses (gain, loss, or no change) as genetic markers, we filtered the CNV-genes using a 1% minor allele frequency threshold, which left only 2519 CNV-genes. We then evaluated their association with each phenotype using a linear mixed model correcting for population structure (efficient mixed-model association

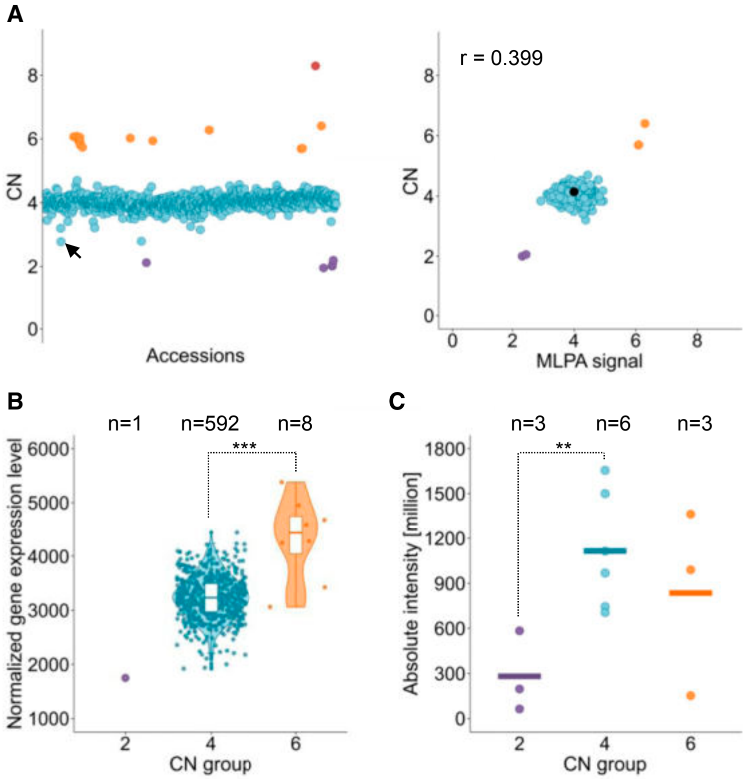


Figure 10. Prevalence of the Duplication of the *SEC10* Gene and Its Effects on Transcript and Protein Levels.

(A) *SEC10* gene copy number in the Arabidopsis population. (Left) Read depth–based copy number (CN) genotypes plotted for 1060 accessions. (Right) Verification of the genotyping data with MLPA assays for 314 accessions. The MLPA signal was scaled to that of the Col-0 accession (marked in black, CN = 4). R, Pearson correlation coefficient.

(B) Distribution of RNA-seq normalized transcript levels among accessions grouped by the copy number class. White boxplots show median (inner line) and inner quartiles (box). Whiskers extend to the highest and lowest values no greater than 1.5 times the inner quartile range, and dots represent the measurements in individual accessions. Asterisks indicate significant differences based on Welch's *t* test (***, $P < 0.01$). Significance was not calculated for the copy number (CN) = 2 group, which included only one sample.

(C) SEC10 protein levels in 3-week-old plants grouped by copy number class. Horizontal lines represent the mean protein level in each group, and the dots represent the measurements in individual accessions. Asterisks indicate significant differences based on Student's *t* test (**, $P < 0.05$). The data were averaged from the measurements of four SEC10 peptide fragments identified by mass spectrometry. The quantification results for individual peptides are presented in Supplemental Figure 18. In each plot, the accessions are colored according to the copy number (CN) classes manually assigned based on the genotyping data: CN = 2 (purple), CN = 4 (blue), CN = 6 (orange), and CN = 8 (red). The accession with the lowest unrounded copy number assigned to the CN = 4 group is KBS-Mac-74 (marked by a black arrow in the left plot); for this accession, the presence of a tandem duplication was confirmed by a BLAST search of the *SEC10* nucleotide sequence against a nanopore-based genomic assembly, confirming the correct group assignment.

expedited). For eight phenotypes, we obtained significant associations with one to eight CNV-genes (Supplemental Figure 19). Among these, the strongest were single-gene associations with three phenotypes of interest: *avrPphB* (*RPS5* gene, $-\log_{10}$ P-value = 16.27), *avrB* (*Rpm1* gene, $-\log_{10}$ P-value = 6.81), and *avrRpm1* (*Rpm1* gene, $-\log_{10}$ P-value = 6.07). These results are in perfect agreement with previous results (Figure 11). This serves as a proof of concept that CNVs can serve as powerful and informative markers for traits where copy number polymorphism is a causative agent of the observed phenotypic variation.

DISCUSSION

Analysis using SNP patterns combined with transcriptomic, proteomic, and phenotypic data has led to the efficient discovery of gene function. However, within the last decade, it has become increasingly clear that variation in gene dosage may also lead to phenotypic diversity within a species. Therefore, copy number genotypes must also be considered when attempting to uncover the genetic basis of many traits (Stankiewicz and Lupski, 2010; Żmieńko et al., 2014). To date, unlike our knowledge about SNPs, our inadequate understanding of CNV locations and frequencies

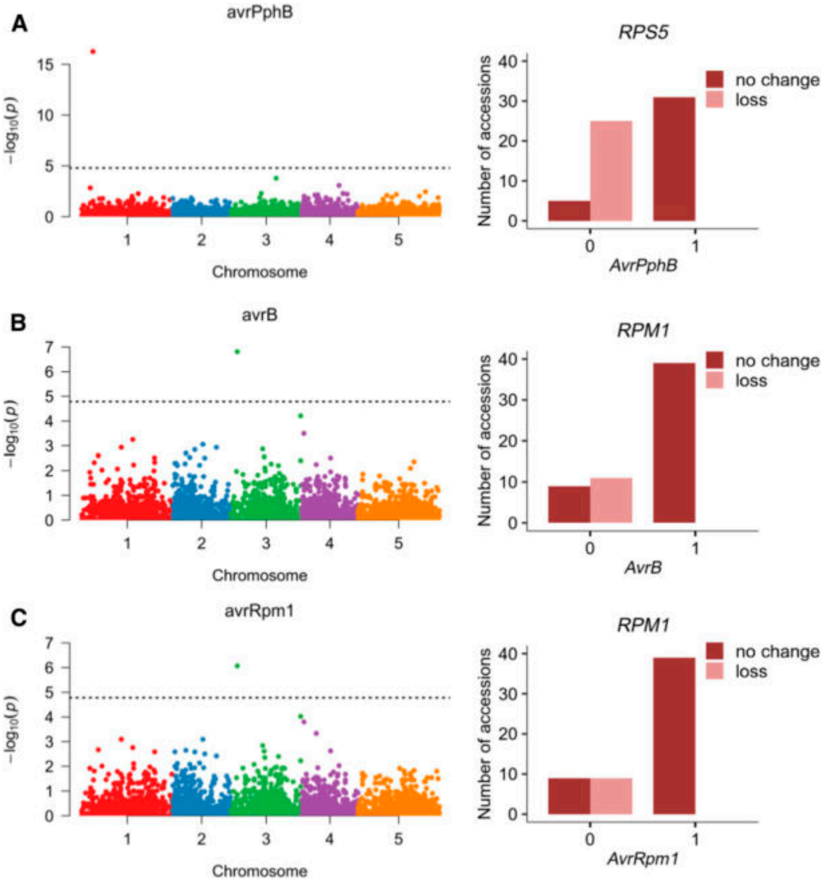


Figure 11. Association of Gene Copy Number Losses in Arabidopsis with Defense Phenotypes.

(A) AvrPphB phenotype.
(B) AvrB phenotype.
(C) AvrRpm1 phenotype. Left panels show Bonferroni-corrected P-values from association analysis; right panels show copy number allele distribution for significantly associated CNV-genes.

in the Arabidopsis 1001 Genomes collection has limited our ability to identify links between genotype and phenotype in this model dicot. Here, we performed an integrative study involving detailed characterization of CNVs in the Arabidopsis genome and their impact on gene dosages. Our map, based on the WGS data for 1064 accessions, substantially extends the list of identified regions with structural variation in this plant obtained from previous studies (Cao et al., 2011; Long et al., 2013). We also performed extensive experimental verification of the genotyping results: we assayed 45 CNV-genes, all in the same set of 314 randomly selected accessions, which guaranteed that the results were not biased toward presenting only a subset of data with the strongest correlations for each CNV. We obtained high concordance between the read depth–based copy numbers and the MLPA signals not only for deletions but also for rare duplications and multiallelic CNV-genes, which is worth noting since experimental verification of duplications has been performed occasionally in large-scale CNV discovery studies in plants (Springer et al., 2009; Swanson-Wagner et al., 2010; Saintenac et al., 2011; Zheng et al., 2011; McHale et al., 2012; Muñoz-Amatriaín et al., 2013; Yu et al., 2013).

Similar to studies involving other plant species (Chia et al., 2012; Muñoz-Amatriaín et al., 2013; Hardigan et al., 2016), we reported high but uneven genome coverage by CNVs in Arabidopsis. We hypothesize that the distribution of CNVs in the genome results from structural and functional constraints on their formation and preservation. The structural constraints may be reflected by the increased representation of tandem duplicates among the CNV-genes identified in our study, which is consistent with the previous finding that CNV regions are hotspots of both past and present large-scale variations (Schuster-Böckler et al., 2010; Jiao and Schneeberger, 2020). The functional constraints might cause highly conserved genes and genes encoding proteins involved in numerous interactions within the cell to be underrepresented in CNV regions due to the usually negative effect of changes in their dosages (Krylov et al., 2003; Platt et al., 2010). In line with this observation, the CNV-genes detected in our study were enriched for less conserved genes, that is, Arabidopsis-specific genes and genes of unknown function. The changes in gene dosage may also provide immediate benefits, for example, a rapid increase in the amount of the enzyme providing drug or herbicide resistance. Indeed, there are several examples highlighting the dynamics of CNV-based adaptation (Harms et al., 1992; Jones et al., 1994; Caretto et al., 1995; Gaines et al., 2010; Kondrashov, 2012). Drawing from nature, processes that induce local changes in DNA copy might therefore be adopted to breed plants with desired traits. However, deeper knowledge about the mechanisms of CNV formation as well as the function of yet-uncharacterized genes is needed to achieve this goal.

AthCNV regions were highly enriched in class I and class II TEs and, similar to the TEs, were unequally distributed across the genome. Indeed, TEs are overrepresented in regions with structural variation (Huang et al., 2008; Cao et al., 2011; Gan et al., 2011; Niu et al., 2019). There is no bias in the localization of newly inserted TEs; however, the deletion of TEs is an ongoing, active, selective process that is largely responsible for the TE distribution pattern in the Arabidopsis genome (Quadrana et al., 2016). A comparison of the genomes of three Arabidopsis accessions, Col-0, Bur-0, and C24, revealed multiple polymorphic TEs for which large deletions were the most common type of variation (93%; Wang et al., 2013). TEs proximal to genes were less variable than distal TEs, suggesting that nearby genes have a negative effect on TE divergence, probably due to stronger selective constraints in these regions. By contrast, TE proximity was positively correlated with the level of small-scale mutations (SNPs and 1- to 3-bp indels) in the genes, pointing to a link between TEs and gene sequence variation. Our observations are in agreement with previous results, and they demonstrate that the variation statuses of genes and TEs are tightly linked and jointly contribute to the unequal distribution of these elements in the genome.

Early studies indicated that the genomes of individual Arabidopsis accessions contain segments not present in the reference genome. The total length of the new sequences in these genomes ranges from 1.3 to 3.3 Mbp (Ossowski et al., 2008; Gan et al., 2011). A recent analysis of the de novo assemblies of seven accessions showed that duplications are the most prevalent type of large CNV (Jiao and Schneeberger, 2020). Because of the limitations of short read–based sequencing (Alkan et al., 2011), we did not use de novo assembly-based approaches for CNV discovery; therefore, our study focused exclusively on regions that were present in the reference genome. Consequently, we detected copy number losses more frequently than copy number gains in most accessions. Nevertheless, by applying population-scale genotyping, we were also able to identify regions missing from the reference genome in our analysis represented by the Col-0 accession, including the recently described duplication of the *SEC10* gene (Vukašinović et al., 2014). Homozygous mutant lines with T-DNA insertions in only one *SEC10* gene had no obvious mutant phenotype; by contrast, introducing mutations in *SEC6* or *SEC8*, which also encode components of the multiprotein exocyst complex, led to defects in pollen-specific transmission. *SEC10* and its duplicate, which share 99% sequence identity, are thought to be functional and complementary (Vukašinović et al., 2014). Here, we showed that the natural duplication of the *SEC10* gene is correlated with the increased transcription and production of *SEC10* protein. Thus, our results strongly support the opinion of Vukašinović et al. (2014) on the role of *SEC10* duplication in the Arabidopsis Col-0 accession. This example also highlights the importance of carefully considering the genetic background in functional and comparative studies. Therefore, we believe that the AthCNV map and the patterns of gene CNV resulting from our study will provide a valuable resource to the Arabidopsis community. They may, for example, guide the selection of the most appropriate sets of accessions for downstream analyses when investigating individual regions in the genome, regardless of whether the presence or lack of variation between these accessions is the main point of interest. As we demonstrated for hypersensitive response phenotypes in Arabidopsis, the copy number data may also complement SNP markers in genome-wide association studies (Fuentes et al., 2019), or to some extent supplement the small number of appropriate plant mutants in comparative functional analyses.

Because of their repetitive nature and the abundance of TE elements, CNV hotspots may accumulate duplications, deletions, and other rearrangements. These rearrangements may be triggered by various mechanisms (Gu et al., 2008; Gabur et al., 2019; Krasileva, 2019). Except for nonallelic homologous recombination events, which lead to recurrent copy number changes with nearly identical breakpoints, the CNV breakpoints in a given region may vary among individuals/accessions. The increasing availability and improvement of the accuracy of long-read DNA sequencing may facilitate more detailed characterizations of such complex CNVs (Michael et al., 2018; Jiao and Schneeberger, 2020). However, the use of population genetics based on chromosome-level sequence assemblies for large numbers of individuals is still a future goal. We observed high consistency between AthCNVs placed at our map, which is a map of merged CNVs and is therefore representative of the entire population rather than individuals, and the variants detected in individual accessions. Thus, we believe that the AthCNV map showing common CNVs in the Arabidopsis genome, combined with the CNV-gene genotyping data, will serve as a useful reference for future studies on variation in Arabidopsis at multiple levels.

METHODS

Data Preprocessing for CNV Discovery and Analysis

The raw reads for 1001 Genomes Project whole-genome shotgun sequence data were downloaded from the National Center for Biotechnology Information Sequence Read Archive repository (PRJNA273563; https://www.ncbi.nlm.nih.gov/bioproject/PRJNA273563). Processed RNA-seq data (normalized counts) for 728 accessions were downloaded from the Gene Expression Omnibus repository (PRJNA319904; https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA319904). The CNV and large indels discovery pipeline was set up based on freely available published tools.

Data Filtering and Quality Analysis

FastQC v.0.11.5 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc) and Trimmomatic v.0.36 (Bolger et al., 2014) were used for read quality analysis and preprocessing. Briefly, the Illumina/Nextera adapter sequences were removed, and the leading and trailing sequences with low base quality (<15) were trimmed. Reads with <30 bases and an average quality score <20 were removed. Finally, reads with a local drop in base quality (average quality <15 measured with a four-base sliding window) were removed. For 45 accessions, fewer than 50% reads or 5,000,000 reads remained following the quality-based filtering, and these accessions were excluded from further analysis (Supplemental Data Set 7). The sequencing data for most rejected accessions were generated during the early stage of the 1001 Genomes Project (Cao et al., 2011), and we decided to remove all data generated at that stage (26 additional accessions) due to their overall lower quality and variable read lengths. The final data set for 1064 accessions was further processed with mapping and CNV detection tools following program-specific parameter optimization, as described below. For 23 accessions, we were unable to extract information about read pairs from the downloaded files; therefore, they were analyzed with read depth–based methods only.

Read Mapping and Marking Duplicates

The genomic reads were mapped to the TAIR10 reference genome assembly using BWA-MEM v.07.15 (Li and Durbin, 2010) and mrsFAST v.3.3.0 (Hach et al., 2014) with default parameters. For mrsFAST mapping, all reads within one sample were first trimmed to obtain a uniform length, and the final read length was calculated separately for each sample based on the largest value that allowed at least 80% of the reads to be kept after trimming. Picard Tools v.2.7.1 (http://broadinstitute.github.io/picard/) and SAMTools v.1.3.1 (Li et al., 2009) were used for data sorting and duplicate removal, respectively. For Genome STRiP analysis, the duplicates were marked, but not removed, to ensure that no unpaired reads remained after the duplicate removal step, since Genome STRiP requires the availability of only paired reads in the input data.

Calculating the Window Size for Read Depth–Based Methods

The number and lengths of the CNV calls when read depth–based methods are used depend on the window/bin size selected for the data-partitioning step. The bin size is a function of coverage, read length, and data quality. To account for all these variables, a bin size evaluation step was performed before the CNV calling step. For CNVnator, the suggested optimal bin size was that for which the ratio of the average read depth signal to its SD was ~4 to 5. We calculated statistics for a wide range of bin sizes (100 to 1500 bases, with 100-base increments) for all samples (Supplemental Data Set 8). The selection of a very small bin size (100 bases) to ensure the highest sensitivity and resolution was justified for multiple samples, but not for all. Because large discrepancies in the CNV lengths and number between the samples might interfere with the subsequent merging process, we narrowed the acceptable bin size range to 400 to 800 bases. The final bin size was then selected for each sample within this range by determining the smallest value for which the ratio of the average read depth to its SD would be at least 4. For 174 samples, the ratio did not reach the threshold, and they were analyzed with a maximal bin size (800 bases). For Control-FREEC, to evaluate the optimal window size, the coefficient of variation for the read depth data was calculated for a wide range of window sizes, as suggested in a previous report (Boeva et al., 2011). For the final analysis, an overlapping sliding window of 800 bases with a step size of 400 bases was chosen. When this window size was used, the coefficient of variation was below 0.1 for 1025 of 1064 samples (the suggested threshold was 0.05 to 0.1; Supplemental Table 6). We noticed that the optimal window size was similar to the CNVnator bin size parameter, therefore enabling the subsequent comparison and merging of the outputs of the two programs.

Calculating the Insert Size Distributions for the Methods Relying on Paired-End Reads

BreakDancer, VariationHunter, and Pindel require insert size range thresholds as input parameters. The insert size distribution in each sequencing library was therefore evaluated with Picard Tools. At this step, 44 accessions were removed from analyses with these callers due to the bimodal distribution of the insert sizes (Supplemental Figure 20; Chen et al., 2009). The upper and lower threshold cutoffs were then calculated for the remaining libraries using two alternative approaches based on either the mean insert size ± 4 SD or the median insert size ± 5 median absolute deviation, and the maximum result of the two approaches was chosen.

CNV and Large Indels Discovery Pipeline

Variants were called by three read depth–based callers (CNVnator, Control-FREEC, and Genome STRiP-CNV pipeline), two discordant read pair–based tools (BreakDancer and VariationHunter), a split read–based tool (Pindel), and a combination of the above-mentioned approaches (the Genome STRiP-SV pipeline). CNV calling was performed with each tool as specified below. Subsequently, a common filter based on size (50 to 499 bp for large indels and at least 0.5 kb for CNVs) and genomic location was applied to the outputs of each caller. Specifically, variants overlapping with assembly gaps larger than 50 bp (with 50-bp borders) or regions close to the chromosome ends (<1 kb) were discarded. Additional filters specific for each CNV calling algorithm are described below.

CNVnator

BWA-MEM alignments were used to call duplications and deletions with CNVnator v.0.33 (Abyzov et al., 2011) based on read mapping density, separately for each accession, with nonoverlapping windows. The read depth signals were corrected for GC bias with a script implemented in the tool. The raw duplication and deletion calls were filtered based on variant size and genomic location. Additionally, to select the calls with the highest confidence, we applied a q0 filter (q0 describes the fraction of reads with a mapping quality of 0 in the called CNV; a high q0 indicates mapping uncertainty due to a lack of uniqueness in the region). Calls with a q0 ≥ 0.5 were removed. Finally, the read depth threshold was applied to remove uncertain calls (i.e., deletion calls with a normalized read depth >0.5 and duplication calls with a normalized read depth <1.5).

Control-FREEC

Aligned BWA-MEM BAM files for each sample were used to detect regions with gains and losses with Control-FREEC v.9.3 (Boeva et al., 2011) using sliding windows. The average GC content of the Arabidopsis genome varies from 32% in the noncoding regions to 44% in the coding regions; therefore, we set the parameters for GC normalization as follows: minExpectedGC = 0.3 and maxExpectedGC = 0.45. The telocentromeric parameter was set to 0 because it was included in our common filter. The breakPointThreshold value for the segmentation of normalized profiles was set to 0.6 (default is 0.8) to increase sensitivity and obtain more segments (and thus more predicted CNVs). The normalized read depth thresholds for CNV detection were ≤0.25 for loss and ≥1.75 for gain.

BreakDancer

The BreakDancerMax program from the BreakDancer package v.1.3.6 (Chen et al., 2009) was used to detect CNVs in each of 997 samples with paired-end data. Calls were made separately for each sample and each chromosome. The raw results that were indicative of CNVs (deletions or insertions) were filtered by a method-specific filter based on the number of supporting read pairs and the confidence score value. Calls with five or more supporting read pairs and confidence scores >30 were retained. For calls supported by less than five read pairs, the confidence score threshold was raised to 90.

VariationHunter

DIVET files with mrsFAST read alignments were used as the input data (for each sample separately) for VariationHunter v.0.04 (Hormozdiari et al., 2009). The analysis consisted of two main steps: the first step involved the clustering of discordant paired-end read mappings. This was performed with the default parameter values, which resulted in read pairs with more than 500 alternative mapping positions being discarded (-x 500) and low-quality ambiguous mapping alternatives being removed with a pruning parameter (-p 0.001). The required genome.satellite.bed and genome.gap.bed files were prepared with in-house scripts from the RepeatMasker v.4.0.7 output. The second step of VariationHunter analysis was the selection of variants from the created clusters. This was performed with a mismatch score (-ms 0.1) to increase the penalty for reads that were not mapping perfectly; additionally, a heuristic algorithm (-wh) was used with the conflict resolution version (-cr) instead of the greedy algorithm, since this algorithm preferred calls that had reads with decreased multiple mapping, and for reads that had multiple mapping, the mapping with a lower edit distance was preferred. A high number of calls were produced as an initial output (-t 10,000) that were subsequently pruned based on the supporting reads information. Additionally, only regions with an average edit distance (AvgEditDits) ≤ 3 were retained. Eventually, all insertion calls were removed after applying the common filter (Supplemental Table 1) because they were shorter than the lower size threshold.

Pindel

Pindel v.0.2.5b8 (Ye et al., 2009) was used for CNV detection (deletions, insertions, and tandem duplications) in individual samples from BWA-MEM alignments of paired-end reads with the following parameters. The maximum size of the structural variations and the window size were set to the default values (-x 5 -w 10), the balance cutoff was set to 0 (-B 0), and the median of the insert size was calculated for each sample (see above). All insertion calls were shorter than 500 bp, and they were eventually removed with the common filter (Supplemental Table 1).

Genome STRiP

BWA-MEM alignments of all 1064 samples were used as input for Genome STRiP v.2.00.1774 (Handsaker et al., 2015). The software required the precomputing of reference metadata based on the ArabidopsisTAIR10 genome sequence, as described in the software documentation (http://software.broadinstitute.org/software/genomestrip/node_ReferenceMetadata.html). All required information was generated according to this documentation except for the lcmask.fasta file (low-complexity mask), where the regions marked as Low complexity, Satellites, and Simple repeat were obtained from RepeatMasker results. Additionally, the TAIR10 reference sequence contained ambiguous nucleotides, which were not permitted by the CNVDiscoveryPipeline script. Therefore, the positions with nucleotides other than A, C, G, T, or N were changed to N and masked in the genome alignability mask (svmask file) by our own scripts. CNV discovery in Genome STRiP was performed with two separate modes, both of which were preceded by summary metadata computations (SVPreprocess script). This step was run with the default values. Large deletions were then identified in the entire population using the SVDiscovery script with the minimum (-minimumSize) and maximum (-maximumSize) event sizes set to 500 and 1,000,000, respectively. The SVDiscovery pipeline scanned the genome for polymorphic sites with large deletions only. The method was initially seeded with aberrantly spaced read pairs and used the read depth as secondary support for the variant sites. All types of CNVs (biallelic duplications, biallelic deletions, and multiallelic variants) were detected separately with the CNVDiscoveryPipeline script in the entire population with the following parameters: -tilingWindowSize 1000, -tilingWindowOverlap 500, -maximumReferenceGapLength 1000, -boundaryPrecision 100, and -minimumRefinedLength 500. The CNVDiscovery Pipeline script implemented a pipeline for discovering CNVs by seeding based on the read depth of the coverage. CNVs that passed through all read signature filters were retained. The outputs of both pipelines were treated as separate data sets.

Variant Merging and Breakpoint Refinement for CNV Discovery

The CNVs were merged, and the breakpoints were refined as follows. (1) Within-tool merge. Variants ≥0.5 kb detected in individual samples by CNVnator and Control-FREEC were merged separately for each caller and for each CNV type (gains and losses) with 50% reciprocal overlap as a criterion. CNVs detected in fewer than two accessions were subsequently discarded. This step eliminated the initial data redundancy and enabled the subsequent comparison of population-based and sample-based CNV calls. (2) Inter-tool merge. A union of all CNVs detected with read depth and hybrid approaches was created by combining the merged-CNVnator, merged-Control-FREEC, Genome STRiP-CNV pipeline, and Genome STRiP-SV pipeline outputs. To remove redundancy, the variants were merged using reciprocal overlap ≥80% as a criterion, which resulted in 34,366 CNVs. (3) CNV breakpoint refinement. The breakpoints of the merged variants were refined by prioritizing the information obtained from the most accurate methods. Individual variants from BreakDancer, VariationHunter, and Pindel that reciprocally overlapped the merged CNVs by at least 80% were used in this step (Supplemental Table 2). If any variants called by the hybrid method (which combines information from the split reads and discordant read pairs at the population level) supported the merge, the maximal coordinates of these variants were used. For the remaining CNVs, if the split read–based variants supported the merge, the maximal coordinates of these variants were used. For any CNVs remaining after this step, if any discordant read pair–based variants supported the merge, the maximal coordinates of these variants were used. Finally, for the CNVs that still remained, the averaged boundaries of the variants predicted by read depth–based methods were set. (4) CNV selection. We selected 19,003 high-confidence CNVs (supported by two or more different callers) for the final AthCNV data set (Supplemental Table 1). Unless otherwise indicated, these CNVs were analyzed further.

Variant Merging and Breakpoint Refinement for Large Indel Discovery

Large indels were merged, and the breakpoints were refined as follows. (1) Within-tool merge. Variants 50 bp to 499 bp detected in individual samples by BreakDancer and VariationHunter were merged separately for each caller with 80% reciprocal overlap as a criterion. Variants detected in fewer than two accessions were subsequently discarded. This step eliminated the initial data redundancy. (2) Inter-tool merge and breakpoints refinement. Variants overlapping each other by at least 80% were merged and their breakpoints were set by prioritizing the information obtained from the most accurate methods, in the same manner as for CNVs. As a result, we obtained 70,137 variants.

Detection of CNVs in the KBS-Mac-74 Genome Assembly

The KBS-Mac-74 genomic assembly based on Oxford Nanopore long reads was downloaded from the European Nucleotide Archive Genome Assembly Database (PRJEB21270; https://www.ebi.ac.uk/ena/data/view/PRJEB21270). We aligned this assembly to the reference genome (TAIR10) with the nucmer aligner in the MUMmer package (Marçais et al., 2018), followed by variant detection with Assemblytics (Nattestad and Schatz, 2016). For comparison with the AthCNV data set, 1551 KBS-Mac-74 variants that were at least 500 bp long were selected and paired with the best matching AthCNVs.

CNV Genotyping with Genome STRiP SVGenotyper

The genome STRiP SVGenotyper module was used to genotype genes in each accession. Prior to genotyping, the nonunique segments in the reference genome were identified by creating subsequence strings with 40-bp sliding windows and a 1-bp step and aligning them with the reference genome; the nonunique segments were masked. This approach was shown to be successful for distinguishing between highly similar paralogs and resulted in more accurate genotyping (Handsaker et al., 2015). All variants in the input vcf files were marked with a SVTYPE tag specifying a general copy number variant ("CNV"). The genotyping failed for 4 of 1064 accessions, and these data were removed. We ultimately obtained the genotyping data for 26,845 genes. A comparison of the unrounded copy numbers and integer copy number genotypes with the results of the MLPA assays for a subset of CNV-genes indicated that the copy number genotypes were frequently not correctly assigned by the SVGenotyper. Therefore, we did not use the genotype confidence filter integrated into the software. Instead, a custom filter based on the unrounded copy number distribution in the Col-0 accession was used to mark and remove outliers, defined as genes falling below (lower quartile minus 3* SD) value or above (upper quartile plus 3* SD) the value of the copy number range distribution in this accession. The threshold values were calculated separately for CNV-genes, genes overlapped by low-confidence CNVs, and NONVAR-genes. This step resulted in 7031 CNV-genes (5517 of them had at least 50% overlap with the CNVs), 4482 genes overlapped by low-confidence CNVs (2874 overlapped by at least 50%), and 14,877 genes not overlapped by any CNVs in the genotyping data.

Annotation and Analysis of CNV-Genes

The centromere positions were defined as described previously (Clark et al., 2007). The genes and noncoding elements in the CNV regions were located using Araport 11 annotations (Cheng et al., 2017). GO analysis was performed with Panther Tools (Panther database v.13.1; Mi et al., 2013). The classification of the gene duplication types (tandem versus block) and gene family specificity analysis were conducted based on information retrieved from the Plaza v.4.0 database (Van Bel et al., 2018). For PCA, 1050 CNV-genes were manually selected based on the distribution of the copy number genotypes (at least two visibly distinguishable copy number classes) and the genomic location (one CNV-gene represented one AthCNV variant; selected AthCNVs were located throughout the entire genome: 390 in chromosome 1 [Chr1], 153 in Chr2, 203 in Chr3, 129 in Chr4, and 175 in Chr5). The analyses were performed with the R-3.5.0 package prcomp(). Graphical representations of CNVs and genes in the genome were prepared with IGV v.2.3.90 (Robinson et al., 2011), circos-0.69.6 (Krzywinski et al., 2009), and TAIR Chromosome Map Tool (https://www.arabidopsis.org/jsp/ChromosomeMap/tool.jsp).

SNP Analysis

SNP data (1001genomes_snp-short-indel_only_ACGTN_v3.1.vcompared withsnpeff file) were downloaded from the 1001 Genomes Project server. PLINK v.1.90b3w program (https://www.cog-genomics.org/plink2) was used for data preprocessing. Only SNP data for 1060 accessions for which we also had CNV genotyping data were used. Variants with missing call rates exceeding value 0.5 as well as variants with minor allele frequency below 3% were filtered out. The LD parameter for linkage disequilibrium-based filtration was set as follows: –indep-pairwise 200'kb' 25 0.3. The resulting 117,232 SNPs were used for PCA analysis with EIGENSOFT v.7.2.1 (Price et al., 2006). The ggbiplot and ggplot2 packages were used for data visualization in the R version 3.6.1 environment.

Genome-Wide Association Study of CNV Data

Defense-related phenotypes (Atwell et al., 2010) were downloaded from the Arapheno database (Togninalli et al., 2019). For the genome-wide association study, we treated all copy number genotypes $\leq$1 as losses, all copy number genotypes >3 as gains, and all the remaining genotypes as unchanged. After filtering the CNV-gene data set with a 1% minor allele frequency threshold, 2519 CNV-genes remained in the analysis. Input files were preprocessed with PLINK v.1.90b3w. The IBS kinship matrix was calculated using SNPs for 1060 accessions. Association analysis was performed for each phenotype using a mixed model correcting for population structure using Efficient Mixed-Model Association eXpedited, version emmax-beta-07Mar2010 (Kang et al., 2010). To declare the threshold for significant association, we used Bonferroni correction. Results were further processed using the qqman package in R.

Experimental Procedures

Plant Materials and Growth Conditions

Arabidopsis seeds were obtained from The Nottingham Arabidopsis Stock Centre. The seeds were surface-sterilized, vernalized for 3 d, and grown on Jiffy pellets in ARASYSTEM containers (BETATECH) in a growth chamber (Percival Scientific). A light intensity of 175 μmol m^{-2} s^{-1} with proportional blue, red, and far red light was provided by a combination of fluorescent lamps (Philips) and GroLEDs red/far red LED Strips (CLF PlantClimatics). Plants were grown for 3 weeks under a 16-h light (22°C)/8-h dark (18°C) cycle, at 70% RH, with nourishment from Murashige and Skoog medium, 0.5× (Serva). A list of accessions used in the experiments is available in Supplemental Data Set 7.

DNA Extraction and MLPA Assays

DNA was extracted from leaves with a DNeasy Plant Mini Kit (Qiagen). The MLPA assays were performed as described previously (Samelak-Czajka et al., 2017) using 5 ng of DNA template with the SALSA MLPA reagent kit FAM (MRC-Holland). The MLPA products were separated by capillary electrophoresis in an ABI Prism 3130XL analyzer at the Molecular Biology Techniques Facility in the Department of Biology at Adam Mickiewicz University, Poznan, Poland. The results were analyzed with GeneMarker v.2.4.2 (SoftGenetics). Whenever possible, to minimize the risk of incorporating SNPs and indels that might affect the probe hybridization step for some accessions, the MLPA probes were designed within regions of minimal sequence variation, as verified by examining vcf files for 1135 accessions obtained from the 1001 Genomes Project website (1001 Genomes Consortium et al., 2016). The genomic target sequence coordinates for the MLPA probes are provided in Supplemental Table 7.

Protein Extraction and Quantification

Proteins were extracted using the phenol method (Hurkman and Tanaka, 1986). The protein pellet was solubilized in 100 mM ammonium bicarbonate for 2 h with three cycles of sonication using a sonic bath every 0.5 h. The protein concentration was determined using a bicinchoninic acid assay (Pierce). For quantification, 10 μg of total protein was reduced, alkylated, and digested with trypsin (Luczak et al., 2016). Each sample was prepared for digestion in duplicate. For each run, 1.5 μg of protein digest was subjected to nano-liquid chromatography–tandem mass spectrometry analysis using a Dionex UltiMate 3000 chromatograph and a Q-Exactive Orbitrap mass spectrometer (Thermo Fisher Scientific) as described previously (Luczak et al., 2016). After each liquid chromatography–tandem mass spectrometry run, the raw files were analyzed by MaxQuant (Cox and Mann, 2008). Quantitative analysis of the experimental groups was based on the label-free quantification intensities. The statistical analyses were performed using Perseus v.1.6.1.3.

Accession Numbers

A detailed list of the accessions and individual data sets used for CNV discovery is provided in Supplemental Data Set 7. The genomic coordinates of CNVs identified in the current study are listed in Supplemental Table 2. The genotyping results for the genes can be accessed through the web interface at http://athcnv.ibch.poznan.pl as user-generated scatterplots that present the copy number values and their distribution across the different genetic groups.

Supplemental Data

Supplemental Figure 1. Comparison of the variants generated by the callers prior to data merging.

Supplemental Figure 2. Fractions of large duplications and deletions detected in the genomes of individual accessions assembled *de novo* from long reads that overlap with AthCNVs.

Supplemental Figure 3. Chromosome map of 100 genes with evidence for duplication/deletion in *A. thaliana* that overlap with AthCNVs.

Supplemental Figure 4. Differences in the number of CNVs overlapping with various genetic elements in the *A. thaliana* genome.

Supplemental Figure 5. Relative distances between genes and TEs and their relationship with CNV status.

Supplemental Figure 6. The accuracy of gene copy number estimates in a complex CNV region calculated for CNV-gene intervals versus AthCNV intervals.

Supplemental Figure 7. Differences between automatic and manual assignment of CNV-gene integer copy numbers from sequencing data.

Supplemental Figure 8. Read depth-based copy number estimates for CNV-genes partially overlapping with CNV regions.

Supplemental Figure 9. Example of a long CNV with a non-uniform pattern of variation of CNV-genes overlapped by this variant.

Supplemental Figure 10. Chromosome map of CNV-genes subjected to experimental verification with MLPA.

Supplemental Figure 11. Intermediate copy number values reported by Genome STRiP for a gene partially covered by CNV.

Supplemental Figure 12. The influence of small-scale sequence variations on oligonucleotide MLPA probe signal and concordance with read depth-based data.

Supplemental Figure 13. Experimental validation of copy number genotypes for NONVAR-genes.

Supplemental Figure 14. Experimental validation of copy number genotypes for CNV-genes with rare duplications (<1%).

Supplemental Figure 15. Experimental validation of copy number genotypes for CNV-genes with common (≥1%) copy number polymorphism.

Supplemental Figure 16. The effect of stock misidentification on the correlation of sequencing-based (source data from the 1001 Genomes Project) and in-house experimental genotyping results.

Supplemental Figure 17. Histograms of gene copy number distribution for CNV-genes that are likely duplicated in the Col-0 accession.

Supplemental Figure 18. Results of mass spectrometry-based identification of SEC10 peptides.

Supplemental Figure 19. Results from GWAS of defense-related phenotypes and CNV-gene data.

Supplemental Figure 20. Insert size distributions in paired-end libraries.

Supplemental Table 1. Variants >0.5 kb in size considered to be copy number changes discovered by each caller in the *A. thaliana* population.

Supplemental Table 2. CNVs resulting from the inter-tool merging of variants (80% RO) and their support by individual callers.

Supplemental Table 3. Gene family specificity of CNV-genes.

Supplemental Table 4. Superfamily composition of *A. thaliana* TEs and its comparison with CNV-TEs and CNV-TEs located within +/− 2 kb distance from the genes.

Supplemental Table 5. Effect of excluding suspicious stocks on the correlation of read depth-based and MLPA-based genotyping results.

Supplemental Table 6. Coefficients of variation (CVs) of read depth values in Control-FREEC analysis.

Supplemental Table 7. List of genomic regions targeted by MLPA probes.

Supplemental Data Set 1. CNVs detected in the *A. thaliana* genome.

Supplemental Data Set 2. Large indels detected in the *A. thaliana* genome.

Supplemental Data Set 3. CNVs at least 0.5 kb long identified in the KBS-Mac-74 genome assembly.

Supplemental Data Set 4. Genes with previous experimental evidence of CNV among *A. thaliana* ecotypes and their overlap with AthCNV variants.

Supplemental Data Set 5. Gene Ontology terms enrichment and protein domain enrichment among groups of genes with proximal TEs depending on their variation status.

Supplemental Data Set 6. List of defense-related phenotypes and identified associations with CNV-genes from GWAS.

Supplemental Data Set 7. List of samples and sequencing data used in this study.

Supplemental Data Set 8. Read depth statistics and bin size selection for CNVnator.

ACKNOWLEDGMENTS

We thank Michal Zenczak for help with drawing Circos plots and Ireneusz Stolarek for valuable comments and discussions. The computations were performed using PL Grid infrastructure resources. This work was supported by the Polish National Centre of Science (grants 2014/13/B/NZ2/03837 to M.F., 2017/01/X/NZ2/00144 to A.Z., and 2017/26/D/NZ2/01079 to A.Z.).

AUTHOR CONTRIBUTIONS

A.Z. and M.F. conceived the study. A.Z., P.W., M.M.-Z., and W.M.K. performed methods optimization tests. A.Z., M.M.-Z., and P.W. performed bioinformatics analyses and analyzed the data. A.Z., A.S.-C., and M.L.

performed experiments and analyzed the data. P.K. and M.F. contributed to the critical interpretation of the results. P.W. prepared the web interface for data visualization. A.Z. wrote the article. M.M.-Z., P.K., and M.F. revised the article. A.Z. and M.M.-Z. prepared the figures. M.F. supervised the study.

Received August 19, 2019; revised March 9, 2020; accepted March 30, 2020; published April 7, 2020.

REFERENCES

1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. Nature **491:** 56–65.

1001 Genomes Consortium (2016). 1,135 Genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. Cell **166:** 481–491.

Abyzov, A., Urban, A.E., Snyder, M., and Gerstein, M. (2011). CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. **21:** 974–984.

Alkan, C., Coe, B.P., and Eichler, E.E. (2011). Genome structural variation discovery and genotyping. Nat. Rev. Genet. **12:** 363–376.

Alonso-Blanco, C., and Koornneef, M. (2000). Naturally occurring variation in Arabidopsis: An underexploited resource for plant genetics. Trends Plant Sci. **5:** 22–29.

Atwell, S., et al. (2010). Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. Nature **465:** 627–631.

Balasubramanian, S., Sureshkumar, S., Lempe, J., and Weigel, D. (2006). Potent induction of Arabidopsis thaliana flowering by elevated growth temperature. PLoS Genet. **2:** e106.

Bloomer, R.H., Juenger, T.E., and Symonds, V.V. (2012). Natural variation in GL1 and its effects on trichome density in *Arabidopsis thaliana*. Mol. Ecol. **21:** 3501–3515.

Boeva, V., Zinovyev, A., Bleakley, K., Vert, J.-P., Janoueix-Lerosey, I., Delattre, O., and Barillot, E. (2011). Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. Bioinformatics **27:** 268–269.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics **30:** 2114–2120.

Bourque, G., Burns, K.H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., Imbeault, M., Izsvák, Z., Levin, H.L., Macfarlan, T.S., Mager, D.L., and Feschotte, C. (2018). Ten things you should know about transposable elements. Genome Biol. **19:** 199.

Bush, S.J., Castillo-Morales, A., Tovar-Corona, J.M., Chen, L., Kover, P.X., and Urrutia, A.O. (2014). Presence-absence variation in *A. thaliana* is primarily associated with genomic signatures consistent with relaxed selective constraints. Mol. Biol. Evol. **31:** 59–69.

Campbell, C.D., Sampas, N., Tsalenko, A., Sudmant, P.H., Kidd, J.M., Malig, M., Vu, T.H., Vives, L., Tsang, P., Bruhn, L., and Eichler, E.E. (2011). Population-genetic properties of differentiated human copy-number polymorphisms. Am. J. Hum. Genet. **88:** 317–332.

Cao, J., et al. (2011). Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. Nat. Genet. **43:** 956–963.

Caretto, S., Giardina, M.C., Nicolodi, C., and Mariotti, D. (1995). Acetohydroxyacid synthase GENE amplification induces clorsulfuron resistance in *Daucus carota* L. In Current Issues in Plant Molecular and Cellular Biology. Current Plant Science and Biotechnology in Agriculture, M. Terzi, R. Cella, and A. Falavigna, eds (Dordrecht: Springer), **Vol. 22: pp.** 235–240.

Chen, K., et al. (2009). BreakDancer: An algorithm for high-resolution mapping of genomic structural variation. Nat. Methods **6:** 677–681.

Cheng, C.-Y., Krishnakumar, V., Chan, A.P., Thibaud-Nissen, F., Schobel, S., and Town, C.D. (2017). Araport11: A complete reannotation of the *Arabidopsis thaliana* reference genome. Plant J. **89:** 789–804.

Chia, J.M., et al. (2012). Maize HapMap2 identifies extant variation from a genome in flux. Nat. Genet. **44:** 803–807.

Clark, R.M., et al. (2007). Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*. Science **317:** 338–342.

Cole, S.J., and Diener, A.C. (2013). Diversity in receptor-like kinase genes is a major determinant of quantitative resistance to *Fusarium oxysporum* f.sp. *matthioli*. New Phytol. **200:** 172–184.

Conrad, D.F., et al.; Wellcome Trust Case Control Consortium. (2010). Origins and functional impact of copy number variation in the human genome. Nature **464:** 704–712.

Cox, J., and Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat. Biotechnol. **26:** 1367–1372.

Duitama, J., Silva, A., Sanabria, Y., Cruz, D.F., Quintero, C., Ballen, C., Lorieux, M., Scheffler, B., Farmer, A., Torres, E., Oard, J., and Tohme, J. (2015). Whole genome sequencing of elite rice cultivars as a comprehensive information resource for marker assisted selection. PLoS One **10:** e0124617.

Fuentes, R.R., Chebotarov, D., Duitama, J., Smith, S., De la Hoz, J.F., Mohiyuddin, M., Wing, R.A., McNally, K.L., Tatarinova, T., Grigoriev, A., Mauleon, R., and Alexandrov, N. (2019). Structural variants in 3000 rice genomes. Genome Res. **29:** 870–880.

Gabur, I., Chawla, H.S., Snowdon, R.J., and Parkin, I.A.P. (2019). Connecting genome structural variation with complex traits in crop plants. Theor. Appl. Genet. **132:** 733–750.

Gaines, T.A., et al. (2010). Gene amplification confers glyphosate resistance in *Amaranthus palmeri*. Proc. Natl. Acad. Sci. USA **107:** 1029–1034.

Gan, X., et al. (2011). Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. Nature **477:** 419–423.

Grant, M.R., Godiard, L., Straube, E., Ashfield, T., Lewald, J., Sattler, A., Innes, R.W., and Dangl, J.L. (1995). Structure of the Arabidopsis RPM1 gene enabling dual specificity disease resistance. Science **269:** 843–846.

Grant, M.R., McDowell, J.M., Sharpe, A.G., de Torres Zabala, M., Lydiate, D.J., and Dangl, J.L. (1998). Independent deletions of a pathogen-resistance gene in *Brassica* and *Arabidopsis*. Proc. Natl. Acad. Sci. USA **95:** 15843–15848.

Gu, W., Zhang, F., and Lupski, J.R. (2008). Mechanisms for human genomic rearrangements. PathoGenetics **1:** 4.

Hach, F., Sarrafi, I., Hormozdiari, F., Alkan, C., Eichler, E.E., and Sahinalp, S.C. (2014). mrsFAST-Ultra: A compact, SNP-aware mapper for high performance sequencing applications. Nucleic Acids Res. **42:** W494–W500.

Handsaker, R.E., Van Doren, V., Berman, J.R., Genovese, G., Kashin, S., Boettger, L.M., and McCarroll, S.A. (2015). Large multiallelic copy number variations in humans. Nat. Genet. **47:** 296–303.

Hardigan, M.A., et al. (2016). Genome reduction uncovers a large dispensable genome and adaptive role for copy number variation in asexually propagated *Solanum tuberosum*. Plant Cell **28:** 388–405.

Harms, C.T., et al. (1992). Herbicide resistance due to amplification of a mutant acetohydroxyacid synthase gene. Mol. Gen. Genet. **233:** 427–435.

Henk, A.D., Warren, R.F., and Innes, R.W. (1999). A new *Ac*-like transposon of Arabidopsis is associated with a deletion of the *RPS5* disease resistance gene. Genetics **151:** 1581–1589.

Hormozdiari, F., Alkan, C., Eichler, E.E., and Sahinalp, S.C. (2009). Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. Genome Res. **19:** 1270–1278.

Horton, M.W., et al. (2012). Genome-wide patterns of genetic variation in worldwide *Arabidopsis thaliana* accessions from the RegMap panel. Nat. Genet. **44:** 212–216.

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.

Hurkman, W.J., and Tanaka, C.K. (1986). Solubilization of plant membrane proteins for analysis by two-dimensional gel electrophoresis. Plant Physiol. **81:** 802–806.

Jiao, W.-B., and Schneeberger, K. (2020). Chromosome-level assemblies of multiple *Arabidopsis* genomes reveal hotspots of rearrangements with altered evolutionary dynamics. Nat. Commun. **11:** 989.

Jones, J.D., Weller, S.C., and Goldsbrough, P.B. (1994). Selection for kanamycin resistance in transformed petunia cells leads to the co-amplification of a linked gene. Plant Mol. Biol. **24:** 505–514.

Kang, H.M., Sul, J.H., Service, S.K., Zaitlen, N.A., Kong, S.Y., Freimer, N.B., Sabatti, C., and Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. Nat. Genet. **42:** 348–354.

Karasov, T.L., et al. (2014). The long-term maintenance of a resistance polymorphism through diffuse interactions. Nature **512:** 436–440.

Kawakatsu, T., et al.; 1001 Genomes Consortium. (2016). Epigenomic diversity in a global collection of *Arabidopsis thaliana* accessions. Cell **166:** 492–505.

Kondrashov, F.A. (2012). Gene duplication as a mechanism of genomic adaptation to a changing environment. Proc. Biol. Sci. **279:** 5048–5057.

Krasileva, K.V. (2019). The role of transposable elements and DNA damage repair mechanisms in gene duplications and gene fusions in plant genomes. Curr. Opin. Plant Biol. **48:** 18–25.

Kroymann, J., Donnerhacke, S., Schnabelrauch, D., and Mitchell-Olds, T. (2003). Evolutionary dynamics of an Arabidopsis insect resistance quantitative trait locus. Proc. Natl. Acad. Sci. USA **100** (Suppl 2): 14587–14592.

Krylov, D.M., Wolf, Y.I., Rogozin, I.B., and Koonin, E.V. (2003). Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. Genome Res. **13:** 2229–2235.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. (2009). Circos: An information aesthetic for comparative genomics. Genome Res. **19:** 1639–1645.

Lee, C.-R., Svardal, H., Farlow, A., Exposito-Alonso, M., Ding, W., Novikova, P., Alonso-Blanco, C., Weigel, D., and Nordborg, M. (2017). On the post-glacial spread of human commensal *Arabidopsis thaliana*. Nat. Commun. **8:** 14458.

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics **26:** 589–595.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics **25:** 2078–2079.

Li, S., Li, R., Li, H., Lu, J., Li, Y., Bolund, L., Schierup, M.H., and Wang, J. (2013). SOAPindel: Efficient identification of indels from short paired reads. Genome Res. **23:** 195–200.

Long, Q., et al. (2013). Massive genomic variation and strong selection in *Arabidopsis thaliana* lines from Sweden. Nat. Genet. **45:** 884–890.

Luczak, M., Suszynska-Zajczyk, J., Marczak, L., Formanowicz, D., Pawliczak, E., Wanic-Kossowska, M., and Stobiecki, M. (2016). Label-free quantitative proteomics reveals differences in molecular mechanism of atherosclerosis related and non-related to chronic kidney disease. Int. J. Mol. Sci. **17:** 1–18.

Marçais, G., Delcher, A.L., Phillippy, A.M., Coston, R., Salzberg, S.L., and Zimin, A. (2018). MUMmer4: A fast and versatile genome alignment system. PLOS Comput. Biol. **14:** e1005944.

McHale, L.K., Haun, W.J., Xu, W.W., Bhaskar, P.B., Anderson, J.E., Hyten, D.L., Gerhardt, D.J., Jeddeloh, J.A., and Stupar, R.M. (2012). Structural variants in the soybean genome localize to clusters of biotic stress-response genes. Plant Physiol. **159:** 1295–1308.

McHale, L., Tan, X., Koehl, P., and Michelmore, R.W. (2006). Plant NBS-LRR proteins: Adaptable guards. Genome Biol. **7:** 212.

Mi, H., Muruganujan, A., Casagrande, J.T., and Thomas, P.D. (2013). Large-scale gene function analysis with the PANTHER classification system. Nat. Protoc. **8:** 1551–1566.

Michael, T.P., Jupe, F., Bemm, F., Motley, S.T., Sandoval, J.P., Lanz, C., Loudet, O., Weigel, D., and Ecker, J.R. (2018). High contiguity *Arabidopsis thaliana* genome assembly with a single nanopore flow cell. Nat. Commun. **9:** 541.

Mills, R.E., et al.; 1000 Genomes Project. (2011). Mapping copy number variation by population-scale genome sequencing. Nature **470:** 59–65.

Minoru, M. (2013). Arabidopsis centromeres. In Plant Centromere Biology, J. Jiang, and and J.A. Birchler, eds (New York: Wiley), pp. 1–14.

Muñoz-Amatriaín, M., et al. (2013). Distribution, functional impact, and origin mechanisms of copy number variation in the barley genome. Genome Biol. **14:** R58.

Nattestad, M., and Schatz, M.C. (2016). Assemblytics: A web analytics tool for the detection of variants from an assembly. Bioinformatics **32:** 3021–3023.

Niu, X.M., Xu, Y.C., Li, Z.W., Bian, Y.T., Hou, X.H., Chen, J.F., Zou, Y.P., Jiang, J., Wu, Q., Ge, S., Balasubramanian, S., and Guo, Y.L. (2019). Transposable elements drive rapid phenotypic variation in *Capsella rubella*. Proc. Natl. Acad. Sci. USA **116: 6908–6913**.

Ossowski, S., Schneeberger, K., Clark, R.M., Lanz, C., Warthmann, N., and Weigel, D. (2008). Sequencing of natural strains of *Arabidopsis thaliana* with short reads. Genome Res. **18:** 2024–2033.

Panchy, N., Lehti-Shiu, M., and Shiu, S.-H. (2016). Evolution of gene duplication in plants. Plant Physiol. **171:** 2294–2316.

Pisupati, R., Reichardt, I., Seren, Ü., Korte, P., Nizhynska, V., Kerdaffrec, E., Uzunova, K., Rabanal, F.A., Filiault, D.L., and Nordborg, M. (2017). Verification of Arabidopsis stock collections using SNPmatch, a tool for genotyping high-plexed samples. Sci. Data **4:** 170184.

Platt, A., et al. (2010). The scale of population structure in *Arabidopsis thaliana*. PLoS Genet. **6:** e1000843.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. **38:** 904–909.

Pucker, B., Holtgräwe, D., Rosleff Sörensen, T., Stracke, R., Viehöver, P., and Weisshaar, B. (2016). A de novo genome sequence assembly of the *Arabidopsis thaliana* accession Niederzenz-1 displays presence/absence variation and strong synteny. PLoS One **11:** e0164321.

Quadrana, L., Bortolini Silveira, A., Mayhew, G.F., LeBlanc, C., Martienssen, R.A., Jeddeloh, J.A., and Colot, V. (2016). The *Arabidopsis thaliana* mobilome and its impact at the species level. eLife **5:** e15716.

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nat. Biotechnol. **29:** 24–26.

Roux, F., and Bergelson, J. (2016). The genetics underlying natural variation in the biotic interactions of *Arabidopsis thaliana*: The challenges of linking evolutionary genetics and community ecology. Curr. Top. Dev. Biol. **119:** 111–156.

Saintenac, C., Jiang, D., and Akhunov, E.D. (2011). Targeted analysis of nucleotide and copy number variation by exon capture in allotetraploid wheat genome. Genome Biol. **12:** R88.

Samelak-Czajka, A., Marszalek-Zenczak, M., Marcinkowska-Swojak, M., Kozlowski, P., Figlerowicz, M., and Zmienko, A. (2017). MLPA-based analysis of copy number variation in plant populations. Front Plant Sci **8:** 222.

Santuari, L., Pradervand, S., Amiguet-Vercher, A.-M., Thomas, J., Dorcey, E., Harshman, K., Xenarios, I., Juenger, T.E., and Hardtke, C.S. (2010). Substantial deletion overlap among divergent Arabidopsis genomes revealed by intersection of short reads and tiling arrays. Genome Biol. **11:** R4.

Schuster-Böckler, B., Conrad, D., and Bateman, A. (2010). Dosage sensitivity shapes the evolution of copy-number varied regions. PLoS One **5:** e9474.

Smith, L.M., Bomblies, K., and Weigel, D. (2011). Complex evolutionary events at a tandem cluster of *Arabidopsis thaliana* genes resulting in a single-locus genetic incompatibility. PLoS Genet. **7:** e1002164.

Springer, N.M., et al. (2009). Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. PLoS Genet. **5:** e1000734.

Staal, J., Kaliff, M., Dewaele, E., Persson, M., and Dixelius, C. (2008). RLM3, a TIR domain encoding gene involved in broad-range immunity of Arabidopsis to necrotrophic fungal pathogens. Plant J. **55:** 188–200.

Stahl, E.A., Dwyer, G., Mauricio, R., Kreitman, M., and Bergelson, J. (1999). Dynamics of disease resistance polymorphism at the Rpm1 locus of Arabidopsis. Nature **400:** 667–671.

Stankiewicz, P., and Lupski, J.R. (2010). Structural variation in the human genome and its role in disease. Annu. Rev. Med. **61:** 437–455.

Sudmant, P.H., et al. (2015). An integrated map of structural variation in 2,504 human genomes. Nature **526:** 75–81.

Swanson-Wagner, R.A., Eichten, S.R., Kumari, S., Tiffin, P., Stein, J.C., Ware, D., and Springer, N.M. (2010). Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. Genome Res. **20:** 1689–1699.

Teo, S.M., Pawitan, Y., Ku, C.S., Chia, K.S., and Salim, A. (2012). Statistical challenges associated with detecting copy number variations with next-generation sequencing. Bioinformatics **28:** 2711–2718.

Togninalli, M., Seren, Ü., Freudenthal, J.A., Monroe, J.G., Meng, D., Nordborg, M., Weigel, D., Borgwardt, K., Korte, A., and Grimm, D.G. (2019). AraPheno and the AraGWAS catalog 2020: A major database update including RNA-seq and knockout mutation data for *Arabidopsis thaliana*. Nucleic Acids Res. **23:** gkz925.

Van Bel, M., Diels, T., Vancaester, E., Kreft, L., Botzki, A., Van de Peer, Y., Coppens, F., and Vandepoele, K. (2018). PLAZA 4.0: An integrative resource for functional, evolutionary and comparative plant genomics. Nucleic Acids Res. **46** (D1): D1190–D1196.

Vlad, D., Rappaport, F., Simon, M., and Loudet, O. (2010). Gene transposition causing natural variation for growth in *Arabidopsis thaliana*. PLoS Genet. **6:** e1000945.

Vukašinović, N., Cvrčková, F., Eliáš, M., Cole, R., Fowler, J.E., Žárský, V., and Synek, L. (2014). Dissecting a hidden gene duplication: The *Arabidopsis thaliana* SEC10 locus. PLoS One **9:** e94077.

Wang, X., Weigel, D., and Smith, L.M. (2013). Transposon variants and their effects on gene expression in *Arabidopsis*. PLoS Genet. **9:** e1003255.

Werner, J.D., Borevitz, J.O., Warthmann, N., Trainer, G.T., Ecker, J.R., Chory, J., and Weigel, D. (2005). Quantitative trait locus mapping and DNA array hybridization identify an FLM deletion as a cause for natural flowering-time variation. Proc. Natl. Acad. Sci. USA **102:** 2460–2465.

Xiao, S., Ellwood, S., Calis, O., Patrick, E., Li, T., Coleman, M., and Turner, J.G. (2001). Broad-spectrum mildew resistance in *Arabidopsis thaliana* mediated by RPW8. Science **291:** 118–120.

Xu, L., Hou, Y., Bickhart, D.M., Zhou, Y., Hay, H.A., Song, J., Sonstegard, T.S., Van Tassell, C.P., and Liu, G.E. (2016). Population-genetic properties of differentiated copy number variations in cattle. Sci. Rep. **6:** 23161.

Ye, K., Schulz, M.H., Long, Q., Apweiler, R., and Ning, Z. (2009). Pindel: A pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics **25:** 2865–2871.

Yi, H., and Richards, E.J. (2009). Gene duplication and hypermutation of the pathogen resistance gene *SNC1* in the Arabidopsis *bal* variant. Genetics **183:** 1227–1234.

Yoon, S., Xuan, Z., Makarov, V., Ye, K., and Sebat, J. (2009). Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Res. **19:** 1586–1592.

Yu, P., Wang, C.H., Xu, Q., Feng, Y., Yuan, X.P., Yu, H.Y., Wang, Y.P., Tang, S.X., and Wei, X.H. (2013). Genome-wide copy number variations in *Oryza sativa* L. BMC Genomics **14:** 649.

Zheng, L.-Y., Guo, X.-S., He, B., Sun, L.-J., Peng, Y., Dong, S.-S., Liu, T.-F., Jiang, S., Ramachandran, S., Liu, C.-M., and Jing, H.-C. (2011). Genome-wide patterns of genetic variation in sweet and grain sorghum (*Sorghum bicolor*). Genome Biol. **12:** R114.

Żmieńko, A., Samelak, A., Kozłowski, P., and Figlerowicz, M. (2014). Copy number polymorphism in plant genomes. Theor. Appl. Genet. **127:** 1–18.

Zmienko, A., Samelak-Czajka, A., Kozlowski, P., Szymanska, M., and Figlerowicz, M. (2016). *Arabidopsis thaliana* population analysis reveals high plasticity of the genomic region spanning MSH2, AT3G18530 and AT3G18535 genes and provides evidence for NAHR-driven recurrent CNV events occurring in this location. BMC Genomics **17:** 893.

MATERIAŁY SUPLEMENTARNE

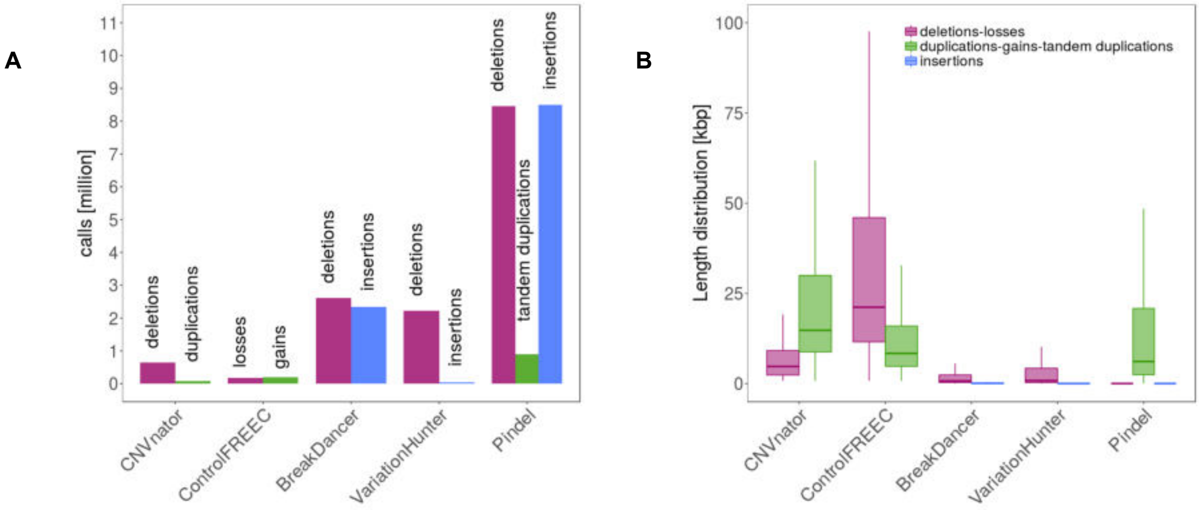
Zmienko A, **Marszalek-Zenczak M**, Wojciechowski P, Samelak-Czajka A, Luczak M, Kozlowski P, Karlowski WM, Figlerowicz M.

AthCNV: A Map of DNA Copy Number Variations in the Arabidopsis Genome

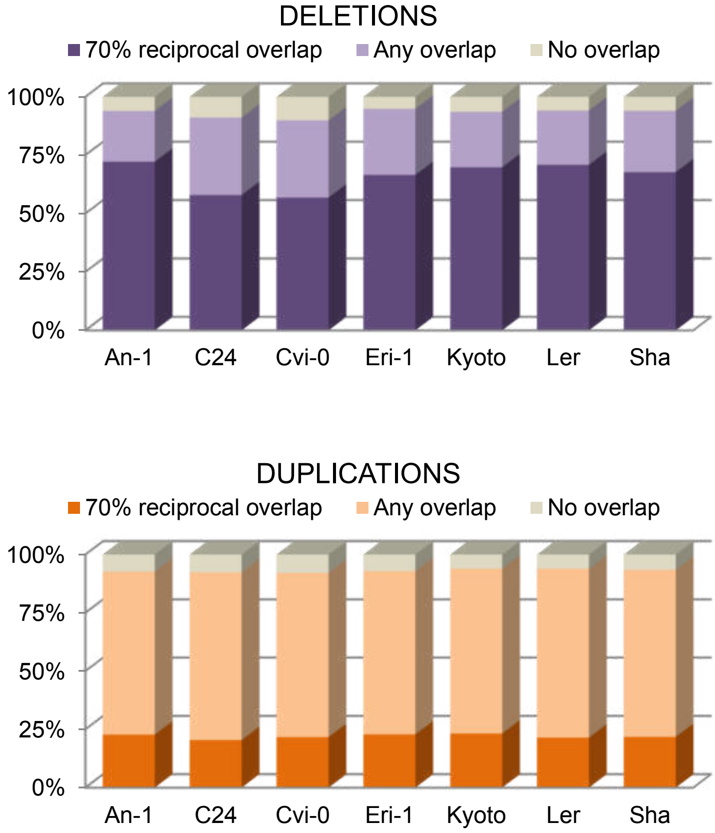
Plant Cell. 2020 Jun;32(6):1797-1819. doi: 10.1105/tpc.19.00640.

5-letni IF = 12,061

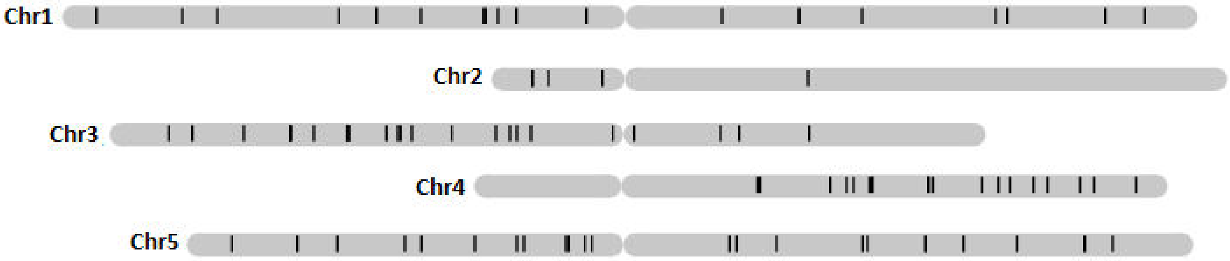
Tabele suplementarne dostępne online:
https://academic.oup.com/plcell/article/32/6/1797/6115649#supplementary-data



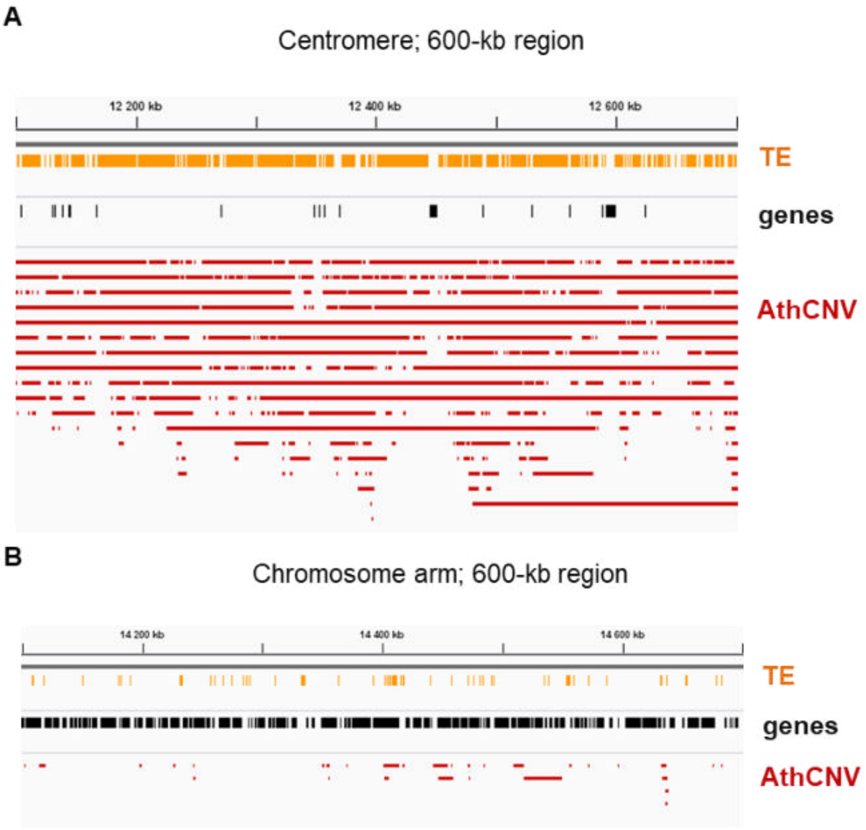
Supplemental Figure 1. Comparison of the variants generated by the callers prior to data merging (Supports Figure 1). A, The total number of raw calls identified in all samples per caller; B, Variant type and length distribution of the raw calls. Boxplots show median (inner line) and inner quartiles (box). Whiskers extend to the highest and lowest values no greater than 1.5 times the inner quartile range.



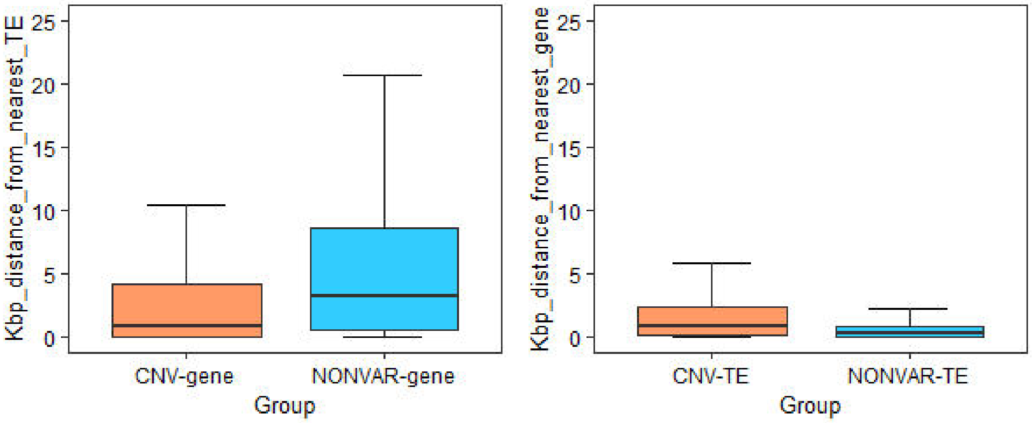
Supplemental Figure 2. Fractions of large duplications and deletions detected in the genomes of individual accessions assembled *de novo* from long reads that overlap with AthCNVs (Supports Figure 3).



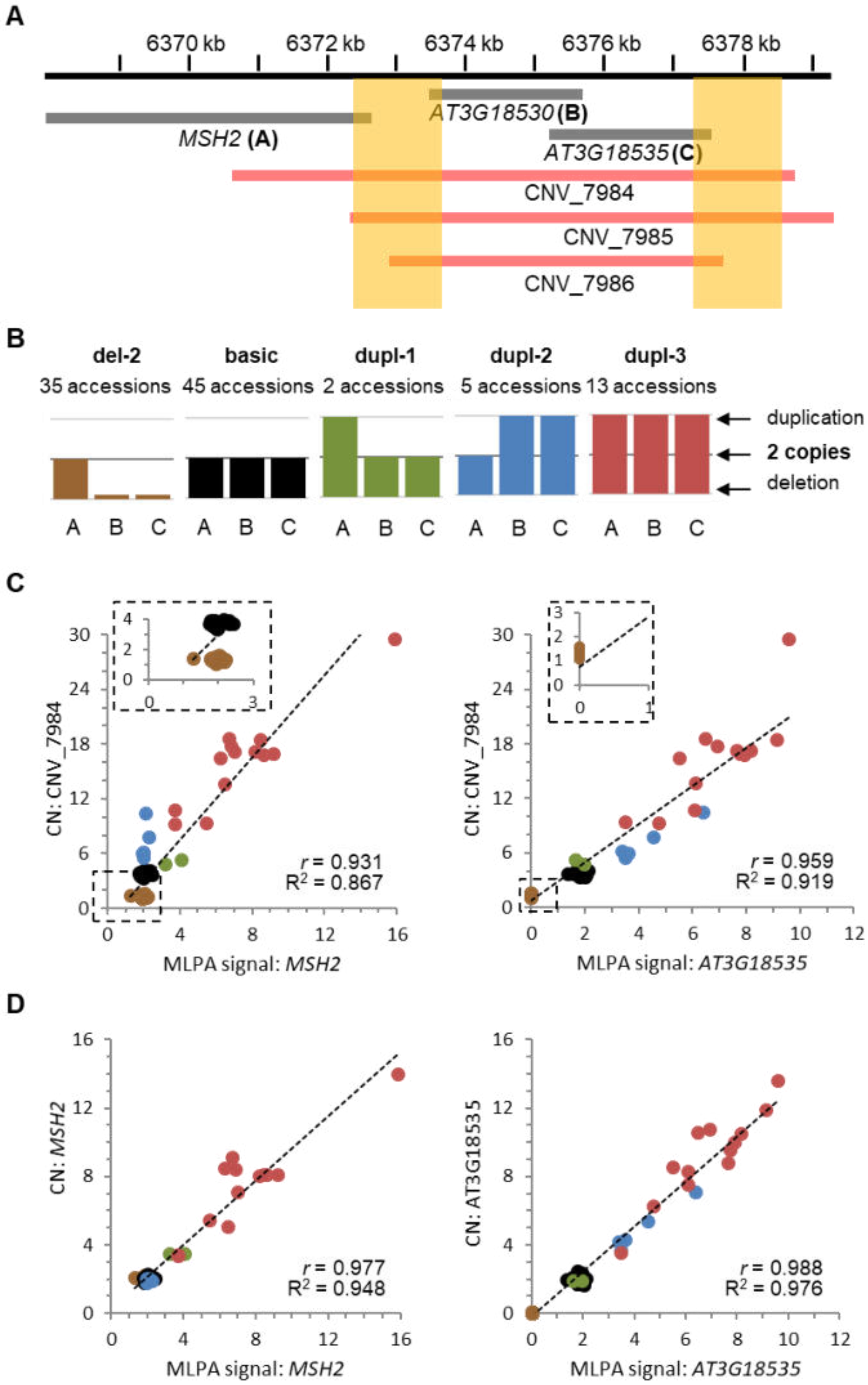
Supplemental Figure 3. Chromosome map of 100 genes with the evidence for duplication/deletion in *A. thaliana* that overlap with AthCNVs (Supports Figure 3).



Supplemental Figure 4. Differences in the number of CNVs overlapping with various genetic elements in the *A. thaliana* genome (Supports Figure 2). IGV screenshots of 600-kbp long regions of (A) centromere (Chr3:12,100,000-12,700,000) and (B) chromosome arm (Chr2:14,100,000-14,70,0000) are presented. AthCNV – CNVs identified in this study; TE – transposable elements; genes – Araport11-annotated protein coding genes.



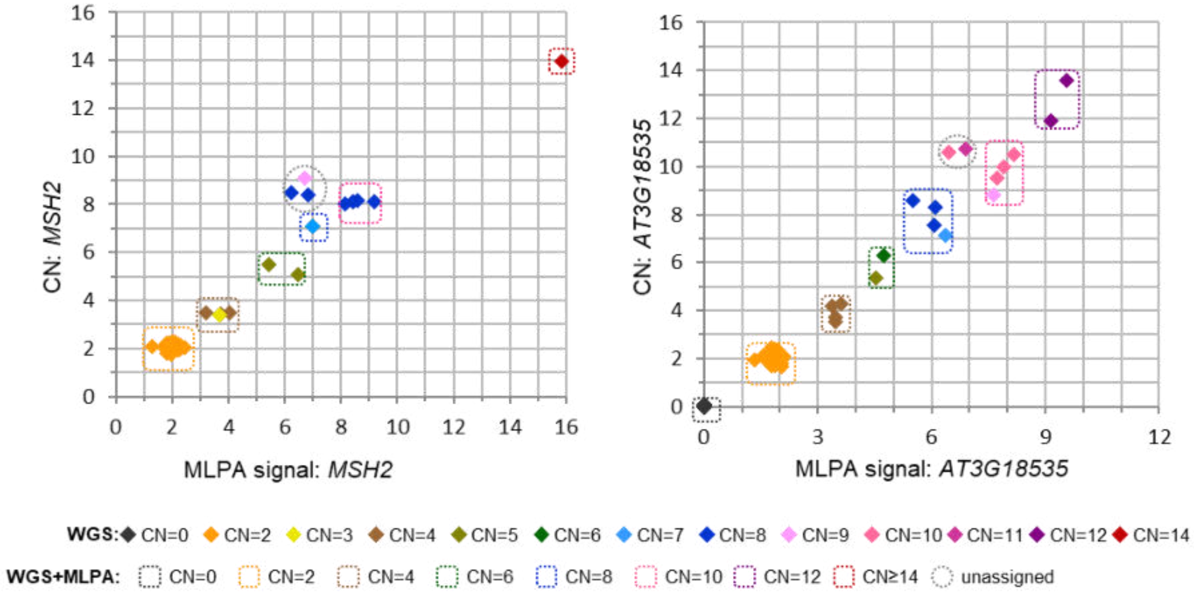
Supplemental Figure 5. Relative distances between genes and TEs and their relationship with CNV status (Supports Figure 5). Only genes and TEs located outside the centromeres were counted. Boxplots show median (inner line) and inner quartiles (box). Whiskers extend to the highest and lowest values no greater than 1.5 times the inner quartile range. For both comparisons, groups were significantly different - Wilcoxon rank sum test with continuity correction, p<0.0001.



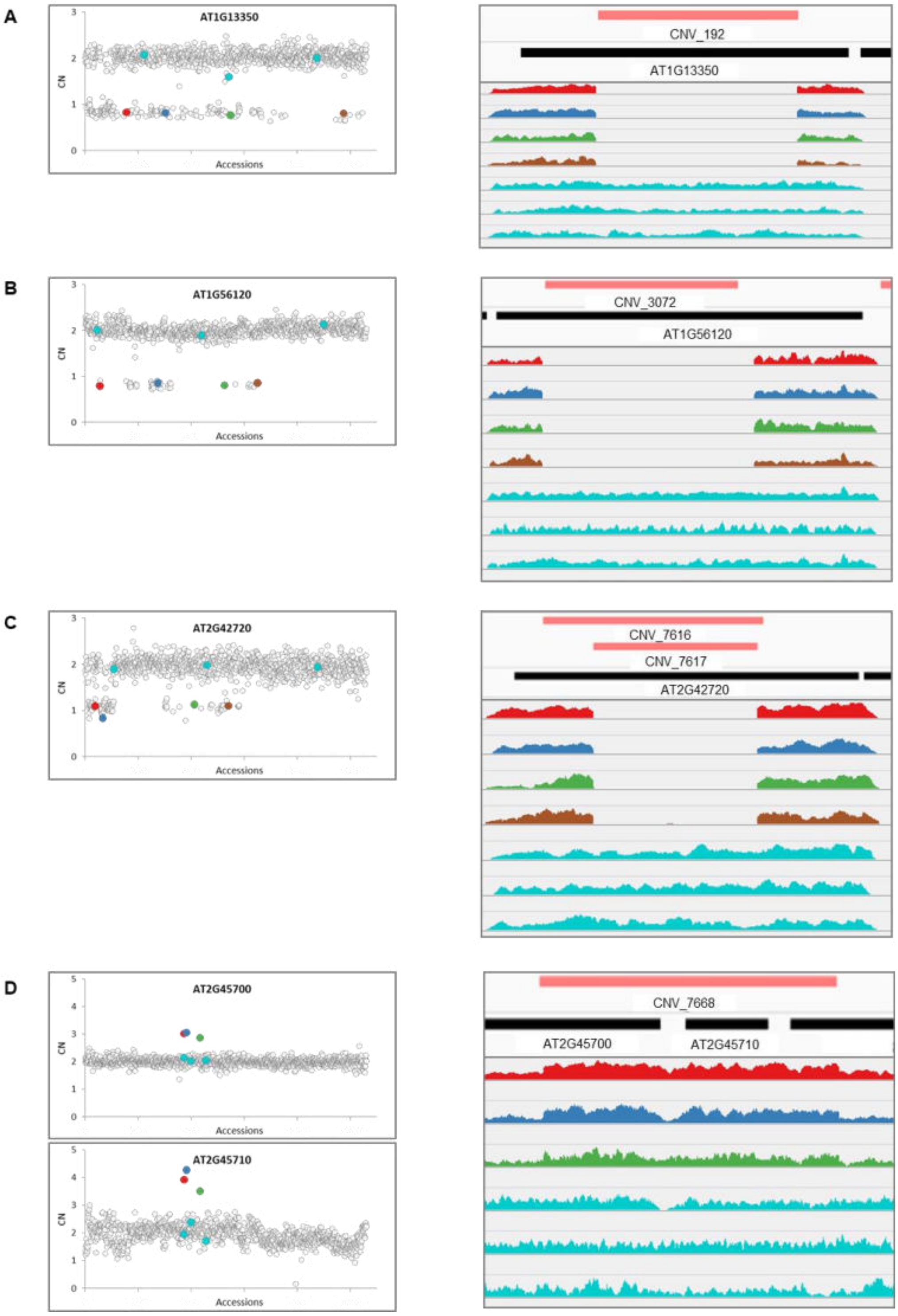
Supplemental Figure 6.

On previous page:

Supplemental Figure 6. The accuracy of gene copy number estimates in a complex CNV region calculated for CNV-gene intervals versus AthCNV intervals (Supports Figure 6). A, The genomic region that encompasses 3 CNV-genes: *MSH2* (*AT3G18524*), *AT3G18530* and *AT3G18535*, is covered by overlapping CNVs: CNV_7984, CNV_7985 and CNV_7986. B, Copy number genotypes of *MSH2* (marked A), *AT3G18530* (marked B) and *AT3G18535* (marked C) in 100 accessions, assigned based on MLPA analysis. C, Correlation of read depth-based copy number estimates in CNV_7984 interval with MLPA-based data for *MSH2* and *AT3G18535*. Variants CNV_7985 and CNV_7986 are embedded within CNV_7984 and influence the copy number estimates for *AT3G18530* and *AT3G18535*, therefore the genotyping results based on CNV_7984 are skewed in accessions with del-2 and dupl-2 genotypes. D, Correlation of read depth-based copy number estimates calculated directly for *MSH2* and *AT3G18535* genomic intervals, with MLPA-based data for these CNV-genes. Orange blocks in (A) mark the regions of 99% similarity, involved in non-allelic homologous recombination. The colors on plots (C-D) correspond to the genotypes depicted in (B). For each gene, the MLPA signals from all accessions were scaled to get the value of 2 in Col-0 accession, that is representative of the reference genotype. CN- copy number. *r* – Pearson correlation coefficient; R^2 – coefficient of determination of linear regression line.



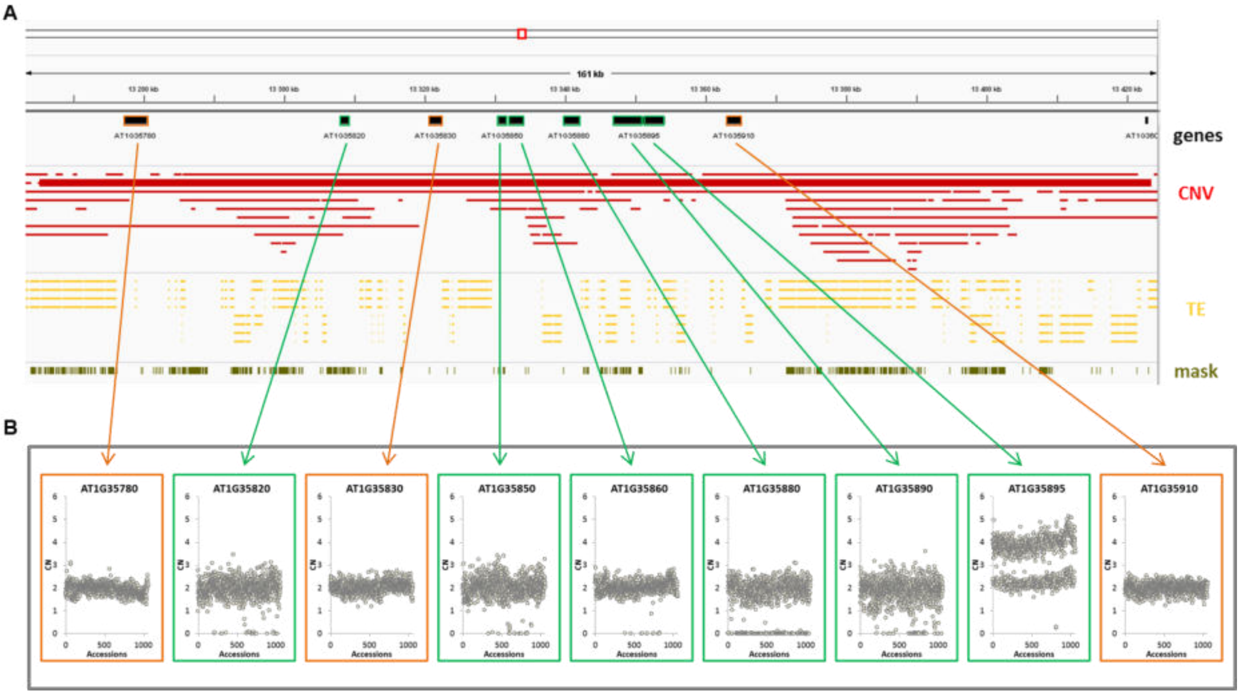
Supplemental Figure 7. Differences between automatic and manual assignment of CNV-gene integer copy numbers from sequencing data (Supports Figure 6)**.** Unrounded CN estimates for *MSH2* and *AT3G18535* in 100 accessions obtained with Genome STRiP SVGenotyper (y axis) are plotted against MLPA signal (scaled to the value of 2 in Col-0 accession). Diamonds are colored according to integer copy numbers assigned by Genome STRiP's clustering script. Dotted loops are colored according to integer copy numbers manually assigned from the intersection of read depth and MLPA data, with the assumption of allele homozygosity, since *A. thaliana* is mostly self-pollinating.



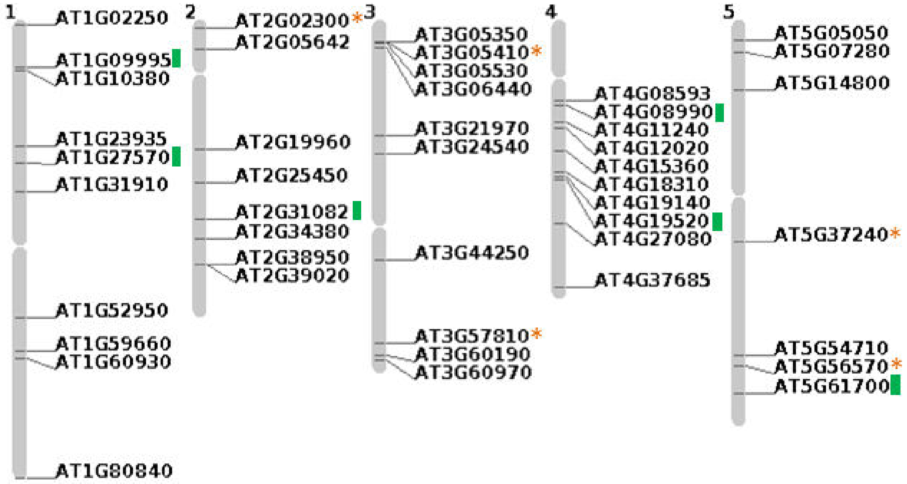
Supplemental Figure 8.

On previous page:

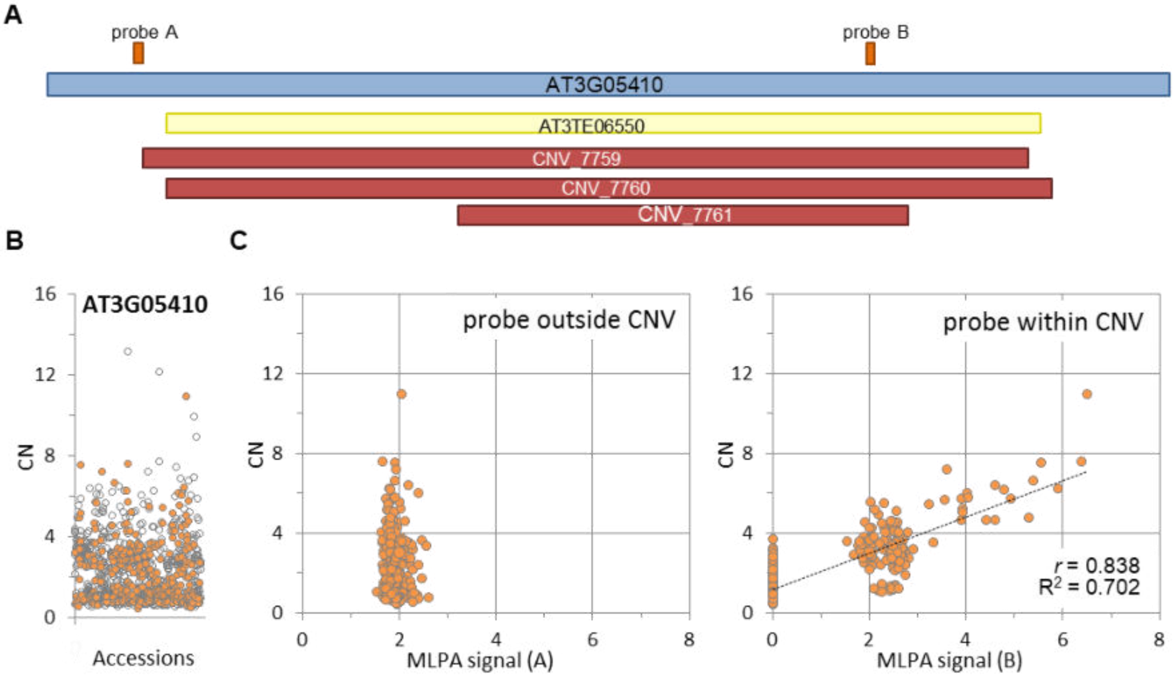
Supplemental Figure 8. Read depth-based copy number estimates for CNV-genes partially overlapping with CNV regions (Supports Figure 6). A, Locus *AT1G13350*, partially overlapping with CNV_192 (61%). B, Locus *AT1G56120*, partially overlapping with CNV_3072 (53%). C, Locus *AT2G42720*, partially overlapping with CNV_7616 and CNV_7617 (64% total). D, Loci *AT2G45700* and *AT2G45710*, both overlapped by CNV_7668 (60% and 100% overlap, respectively). For each CNV-gene two panels are presented: left – read depth-based copy number estimates in 1,060 accessions, right - the IGV screenshots indicating the positions of CNV-genes and AthCNVs and the short read coverage tracks for selected accessions. The colors for these accessions are identical with the colors in the left panels.



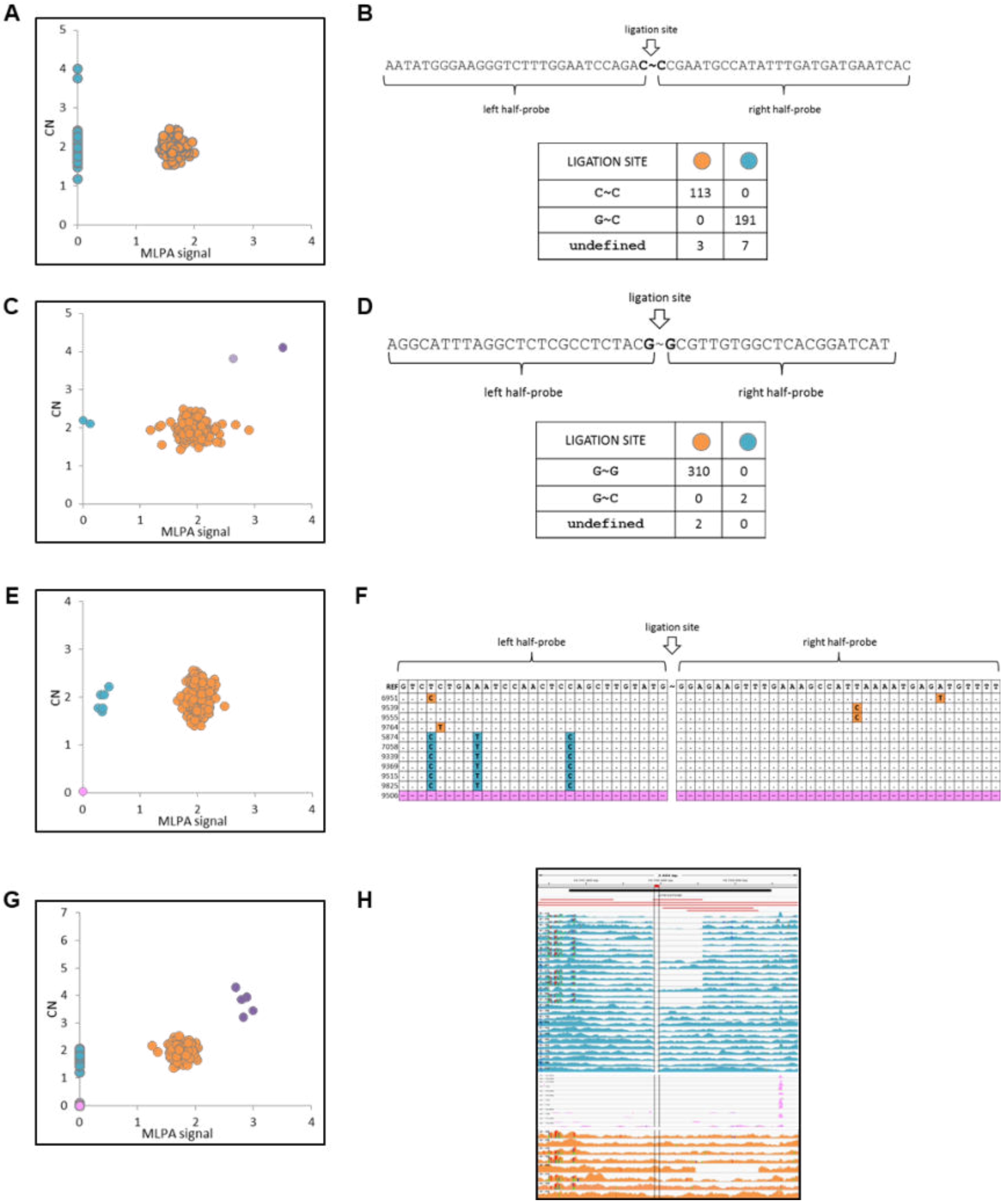
Supplemental Figure 9. Example of a long CNV with a non-uniform pattern of variation of CNV-genes overlapped by this variant (Supports Figure 6). A, An IGV screenshot of a 161-kbp subgenomic region on chromosome 1 covered by CNV_1210 (wide red bar) and numerous shorter CNVs (narrow red bars). This CNV overlaps 9 CNV-genes (black) and multiple TEs (orange). The non-unique segments in the reference genome are marked in olive (mask); these regions were not used for copy number genotyping. B, The CNV-gene genotypes in 1,060 accessions inferred from WGS data. CNV-genes with the copy number difference value between any two accessions < 2 are marked by orange frames, the remaining CNV-genes are marked by green frames.



Supplemental Figure 10. Chromosome map of CNV-genes subjected to experimental verification with MLPA (Supports Figure 7). The results of validation are presented in Supplemental Figures 13-15, except for results for CNV-genes marked with orange asterisks, which are shown separately, in Supplemental Figures 11-12. Genes marked with green rectangles are also displayed in Figure 7 in the Main Text.



Supplemental Figure 11. Intermediate copy number values reported by Genome STRiP for a gene partially covered by CNV (Supports Figure 7). A, *AT3G05410* locus is partially (80%) covered by 3 CNVs that also overlap with the transposable element *AT3TE06550.* Two probes were used for MLPA-based genotyping of this CNV-gene, localized outside (probe A) and within CNV region (probe B). B, Non-discrete distribution of copy numbers (CN) among 1,060 accessions is observed in read depth-based genotyping experiment; C. MLPA results stay in agreement with the genomic localization of CNV variants. The results for probe B are positively correlated with read depth-based measurements. Increased CN values in read depth data can be attributed to incomplete overlap by AthCNVs. The results for 314 accessions used for MLPA analysis are colored orange on each plot. The MLPA signals were scaled to the value of 2 in Col-0 accession, that is representative of the reference genotype. *r* – Pearson correlation coefficient; R^2 – coefficient of determination of linear regression line.

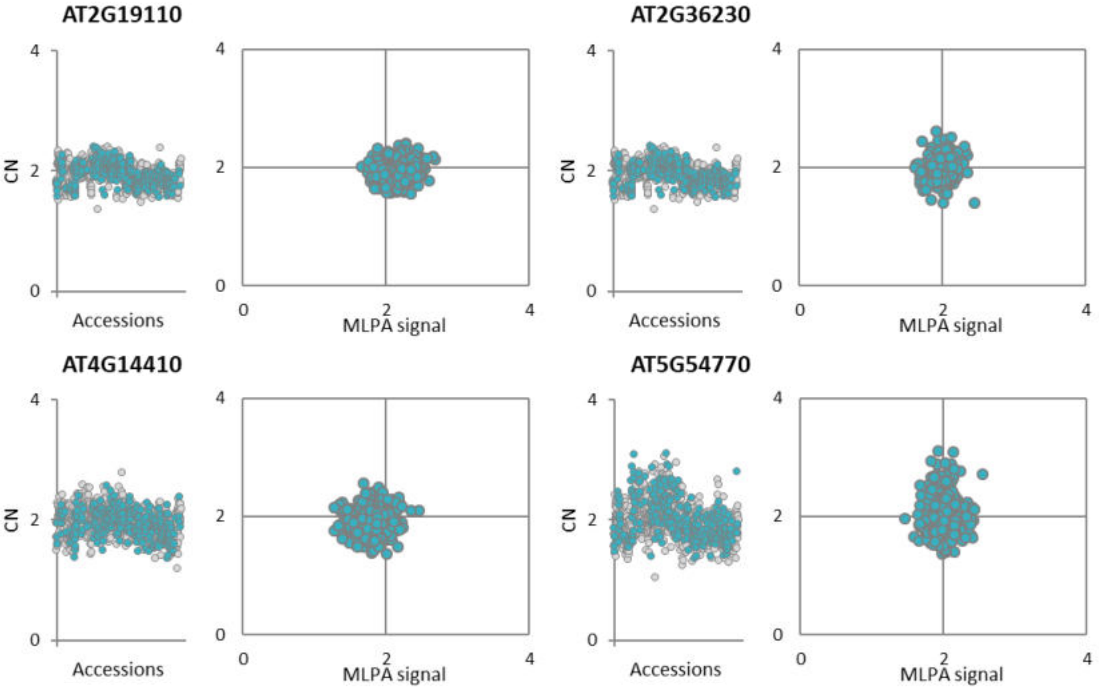


LIGATION SITE	●	●
C~C	113	0
G~C	0	191
undefined	3	7

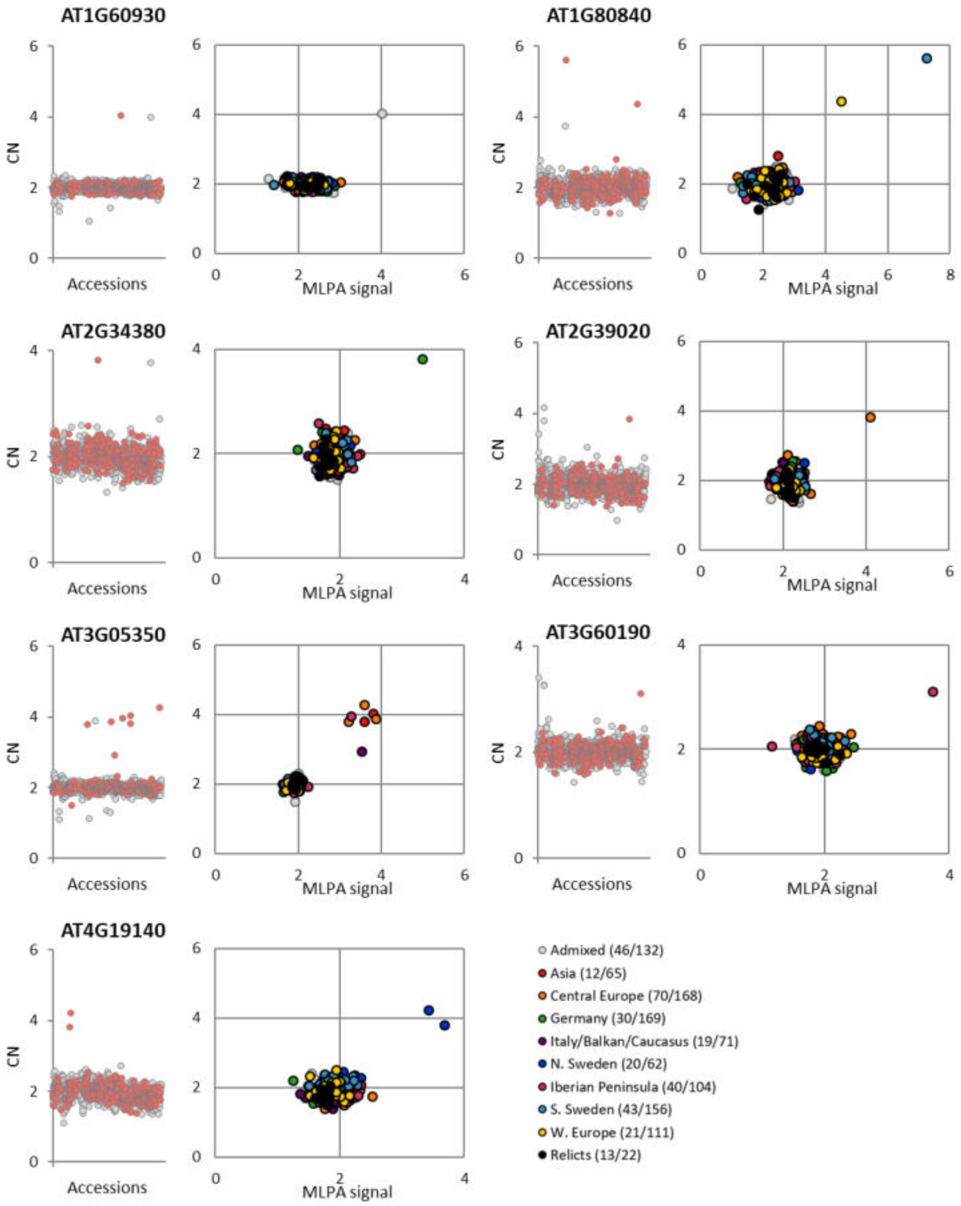
LIGATION SITE	●	●
G~G	310	0
G~C	0	2
undefined	2	0

Supplemental Figure 12.

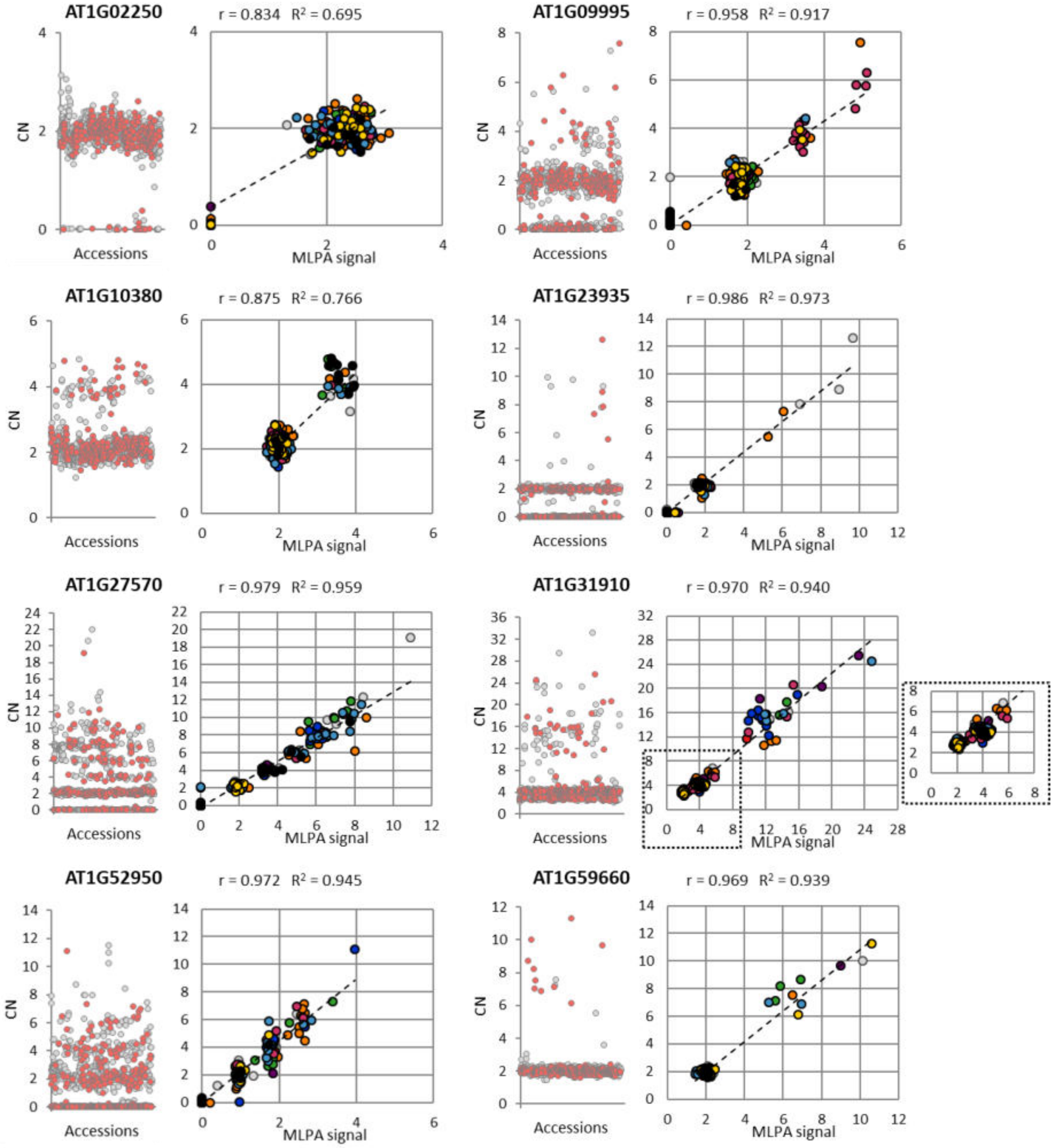
On previous page:
Supplemental Figure 12. The influence of small-scale sequence variations on oligonucleotide MLPA probe signal and concordance with read depth-based data (Supports Figure 7). A,C,E,G, The two-dimensional plots of read depth-based copy numbers (CN) versus MLPA assay-derived signals for *AT5G56570, AT3G57810, AT2G02300* and *AT5G37240* loci, respectively, in 314 accessions. The MLPA signal was scaled to the value of 2 in Col-0 accession, that is representative of the reference genotype. The accessions are marked with colors indicating their copy number group and concordance of read depth-based and MLPA results: orange - unchanged copy number relatively to Col-0, pink – concordant deletion, purple – concordant duplication, blue - discordant results. B,D,F,H, The sequence variations in MLPA probes affecting their hybridization explain the presence of discordant data. SNP information was obtained from .vcf files downloaded from the 1001 Genomes Project server.
For *AT5G56570* (A,B) no signal from MLPA assay was detected in 198 samples, while the gene was indicated present in unchanged or increased copy number state by read depth analysis. The discrepancy between read depth and MLPA analysis was correlated with the presence of a SNP (C→G) in the sequence targeted by MLPA probe. The SNP compromised the left half-probe affinity at the ligation site, critical for MLPA performance. Colors in the table in (B) match the colors in the scatterplot in (A).
For *AT3G57810* (C,D) no signal from MLPA assay was detected in 2 samples, while the gene was indicated present in unchanged copy number state by read depth analysis. The discrepancy between read depth and MLPA analysis was correlated with the presence of a SNP (G→C) in the sequence targeted by MLPA probe. The SNP compromised the right half-probe affinity at the ligation site, critical for MLPA performance. Colors in the table in (D) match the colors in the scatterplot in (C).
For *AT2G02300* (E,F) decreased signal from MLPA assay was detected in 6 samples, while the gene was indicated present in unchanged copy number state by read depth analysis. The discrepancy between read depth and MLPA analysis was fully correlated with the presence of 3 SNPs in these accessions, that collectively affected the affinity of left half-probe to the target sequence. Colors in the table in (F) match the colors in the scatterplot in (E).
For *AT5G37240* (G,H) no signal from MLPA assay was detected in 28 samples, while the gene was indicated present by read depth analysis. Close inspection of mapping reads revealed the presence of a gap overlapping the sequence targeted by MLPA probe in all accessions with discordant results. On IGV screenshot: black – *AT5G37240* locus, green – the region targeted by the MLPA probe, red – CNVs overlapping the *AT5G37240* gene; the colors of reads of individual accessions in (H) match the colors in scatterplot in (G).



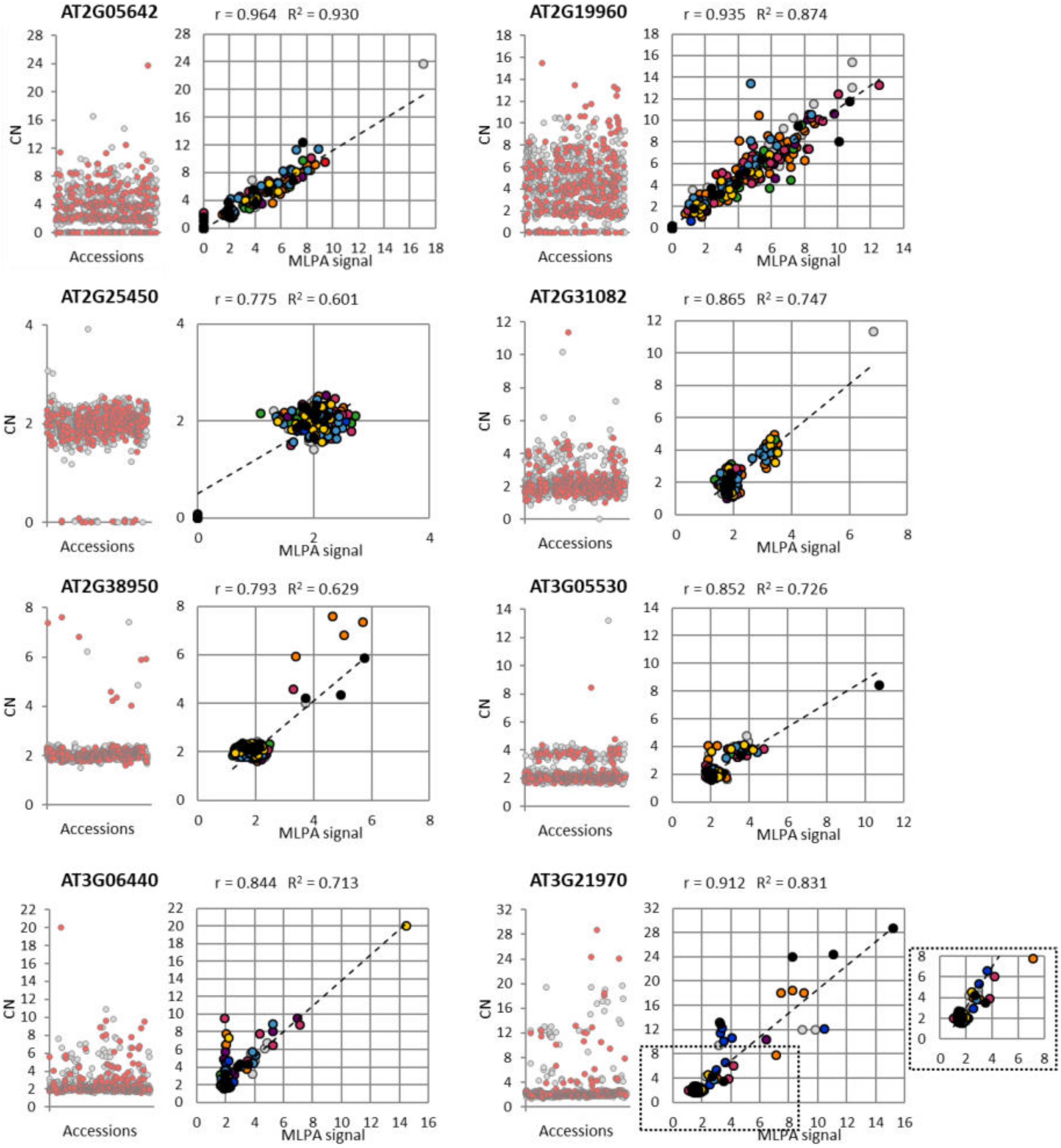
Supplemental Figure 13. Experimental validation of copy number genotypes for NONVAR-genes (Supports Figure 7). For each gene, two plots are presented. The left scatterplot presents the read depth-based copy number (CN) data for 1,060 accessions. The right scatterplot compares read depth data with the results of MLPA assays. For each gene, the MLPA signals from all accessions were scaled to the value of 2 in Col-0 accession, that is representative of the reference genotype. The same set of 314 accessions was used in all validation experiments (blue).



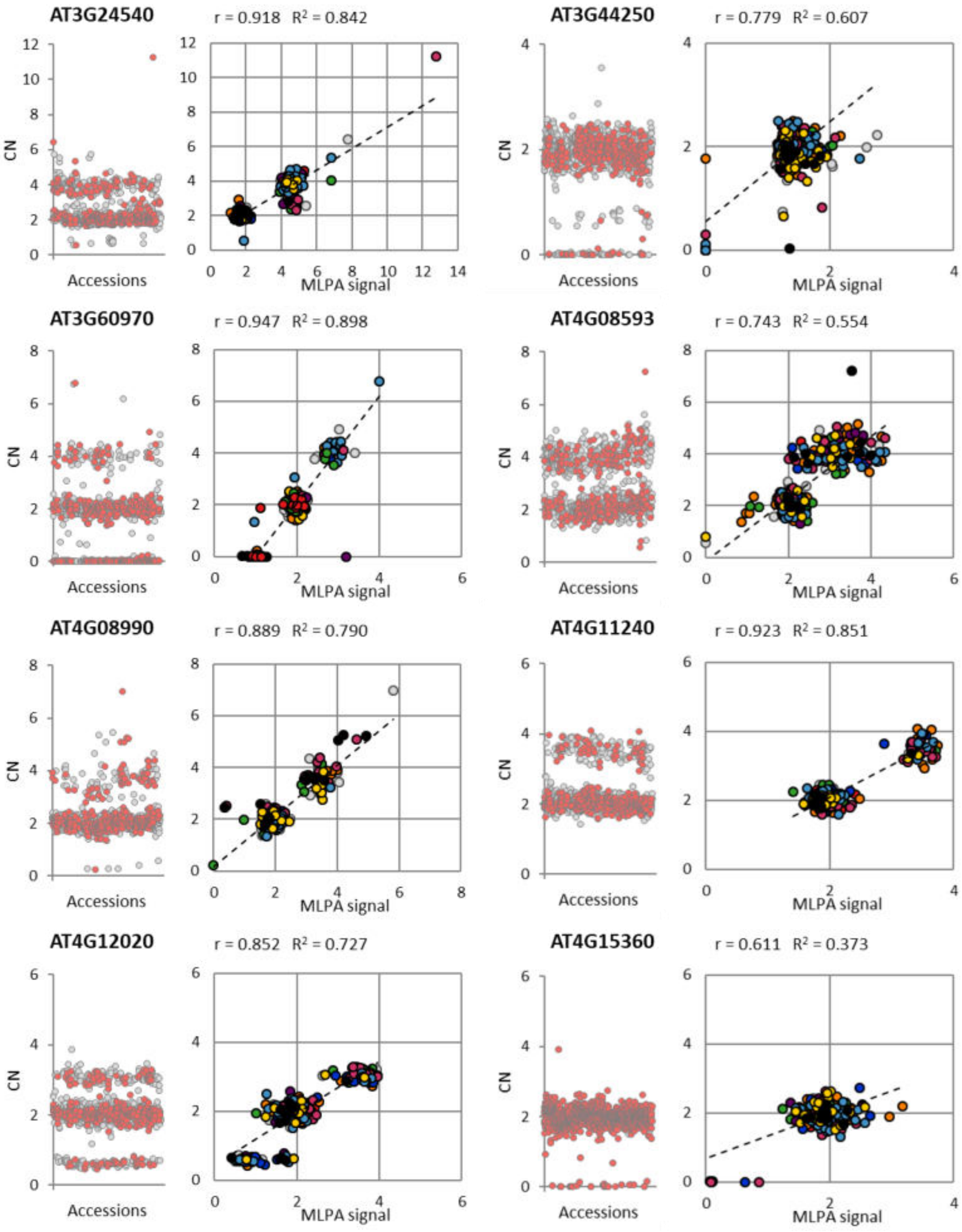
Supplemental Figure 14. Experimental validation of copy number genotypes for CNV-genes with rare duplications (<1%) (Supports Figure 7). For each gene, two plots are presented. The left scatterplot presents read depth-based copy number (CN) data for 1,060 accessions. The right scatterplot compares read depth data with the results of MLPA assays. The same set of 314 accessions was used in all validation experiments; these accessions are colored red in the left plots, while in the right plots they are colored according to the genetic group. MLPA signals were scaled to the value of 2 in Col-0 accession. The legend shows the number of accessions with read depth-based/MLPA-based data, respectively.



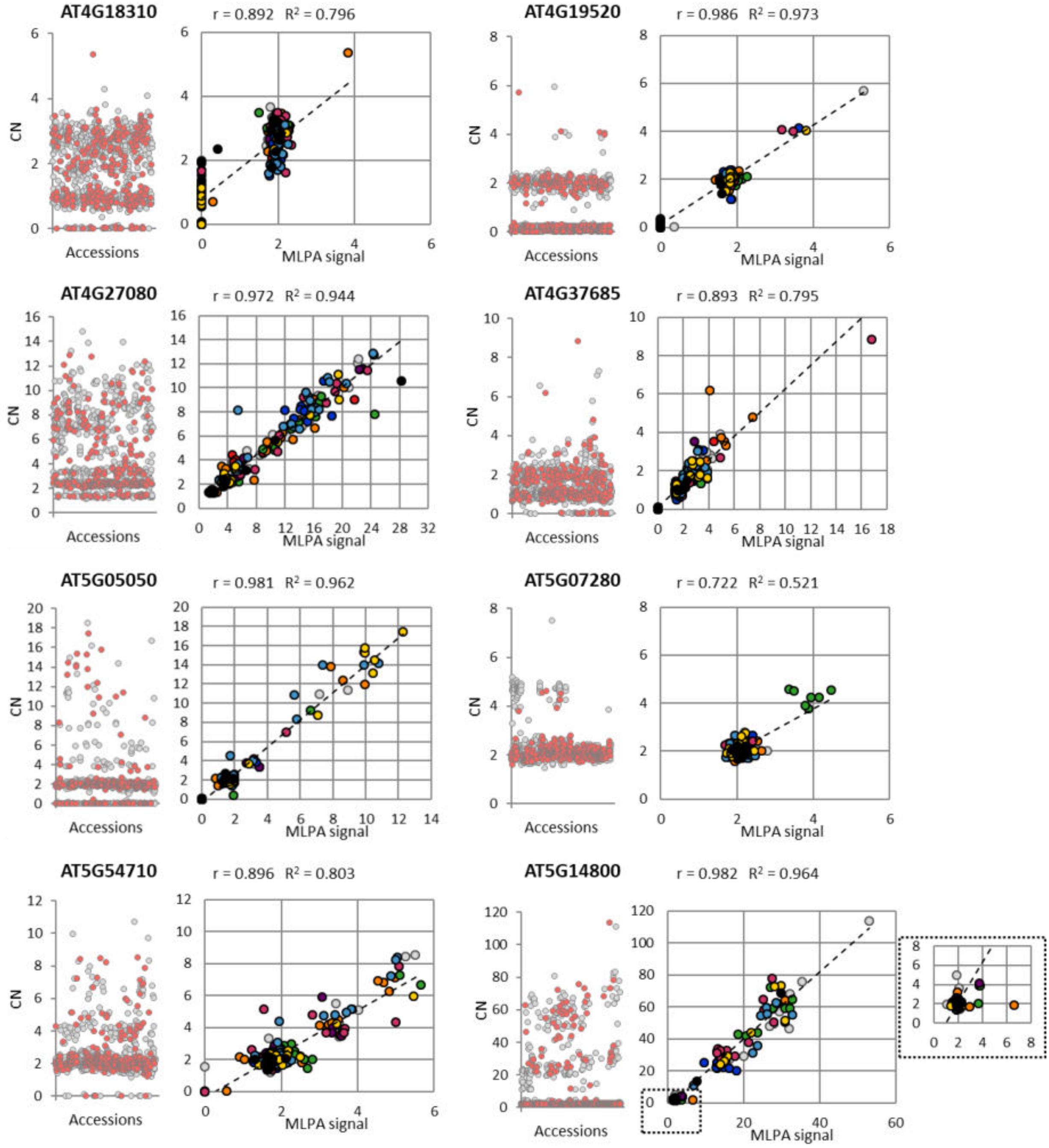
Supplemental Figure 15. (Continued on next page)



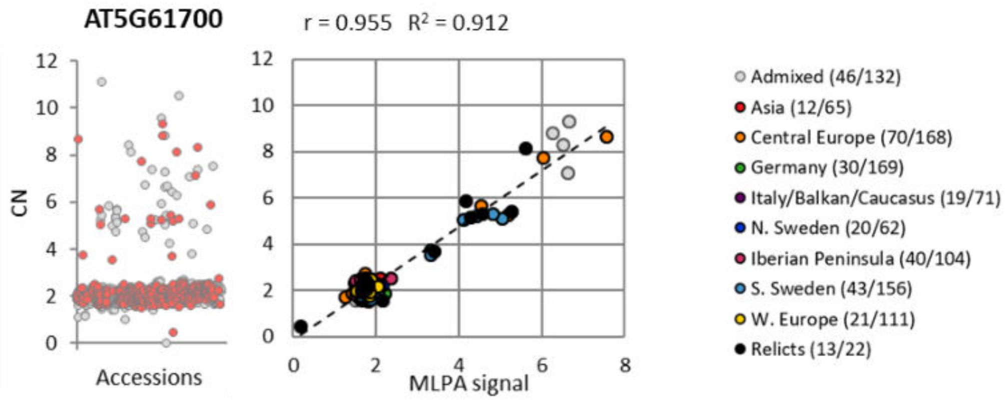
Supplemental Figure 15. (Continued on next page)



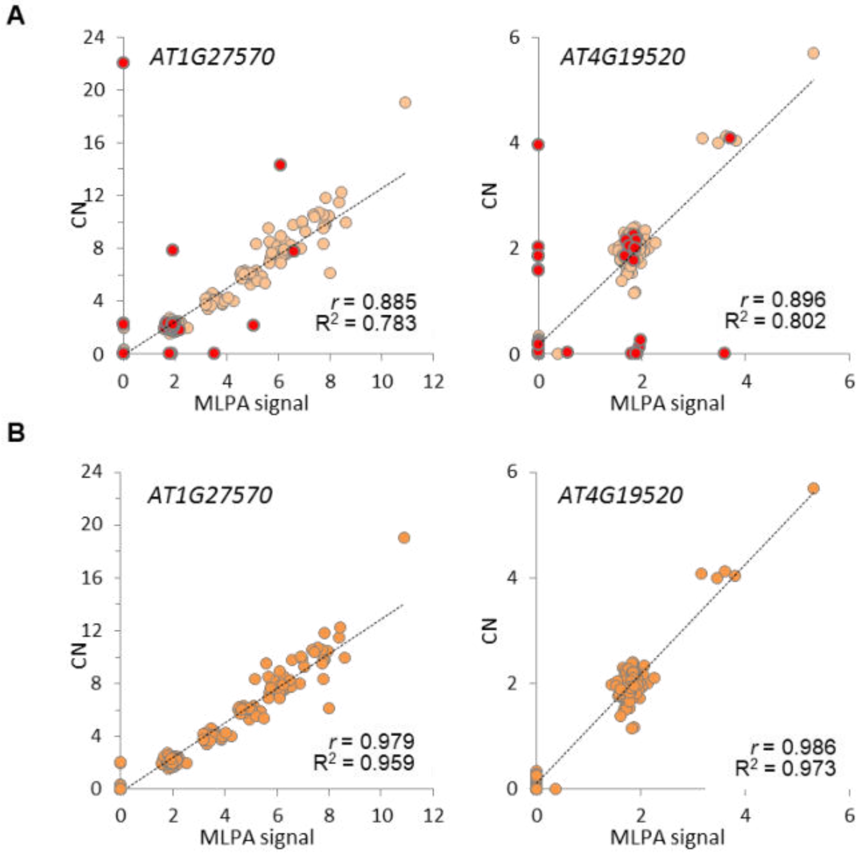
Supplemental Figure 15. (Continued on next page)



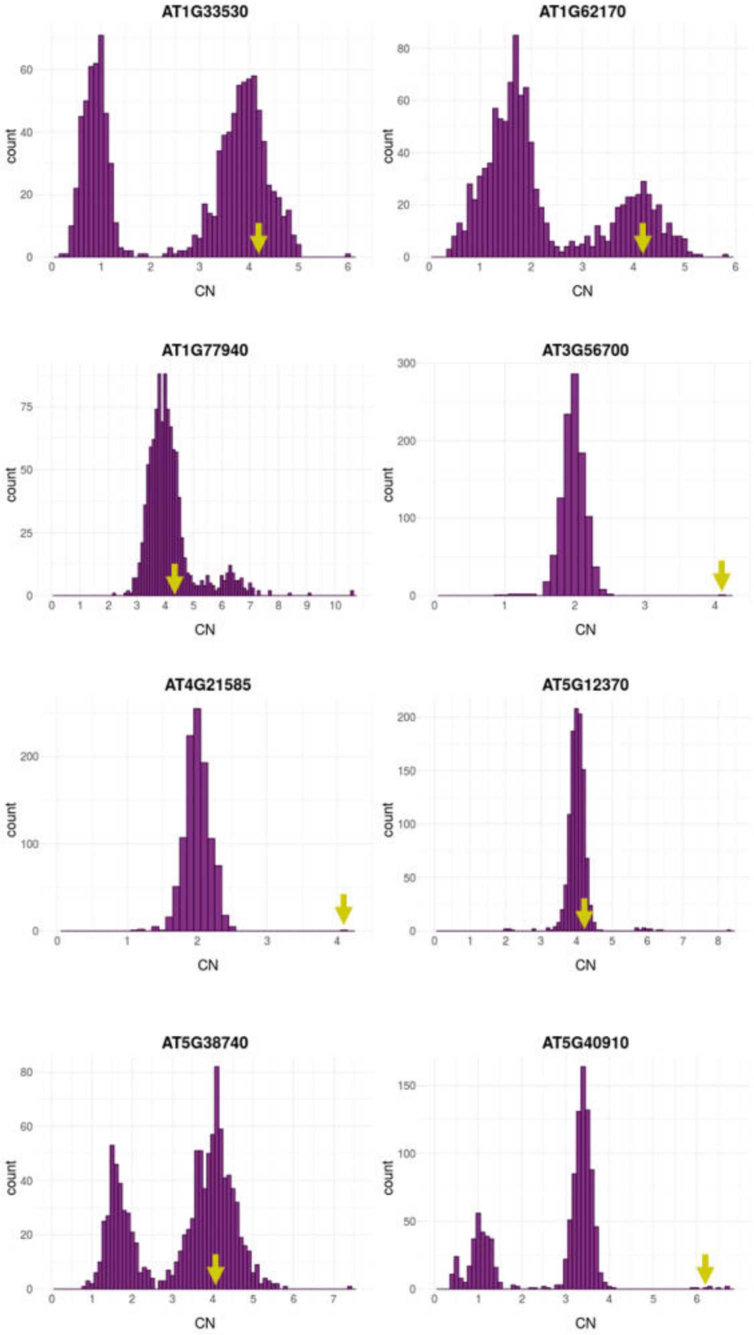
Supplemental Figure 15. (Continued on next page)



On this and previous pages:
Supplemental Figure 15. Experimental validation of copy number genotypes for CNV-genes with common (≥1%) copy number polymorphism (Supports Figure 7). For each gene, two plots are presented. The left scatterplot presents read depth-based copy number (CN) data for 1,060 accessions. The right scatterplot presents the correlation of read depth data with MLPA assays. The same set of 314 accessions was used in all validation experiments; these accessions are colored red in the left plots, while in the right plots they are colored according to the genetic group. MLPA signals were scaled to the value of 2 in Col-0 accession. The legend shows the number of accessions with read depth-based/MLPA-based data, respectively, in each genetic group. r- Pearson correlation coefficient. R^2 - coefficient of determination of linear regression line.



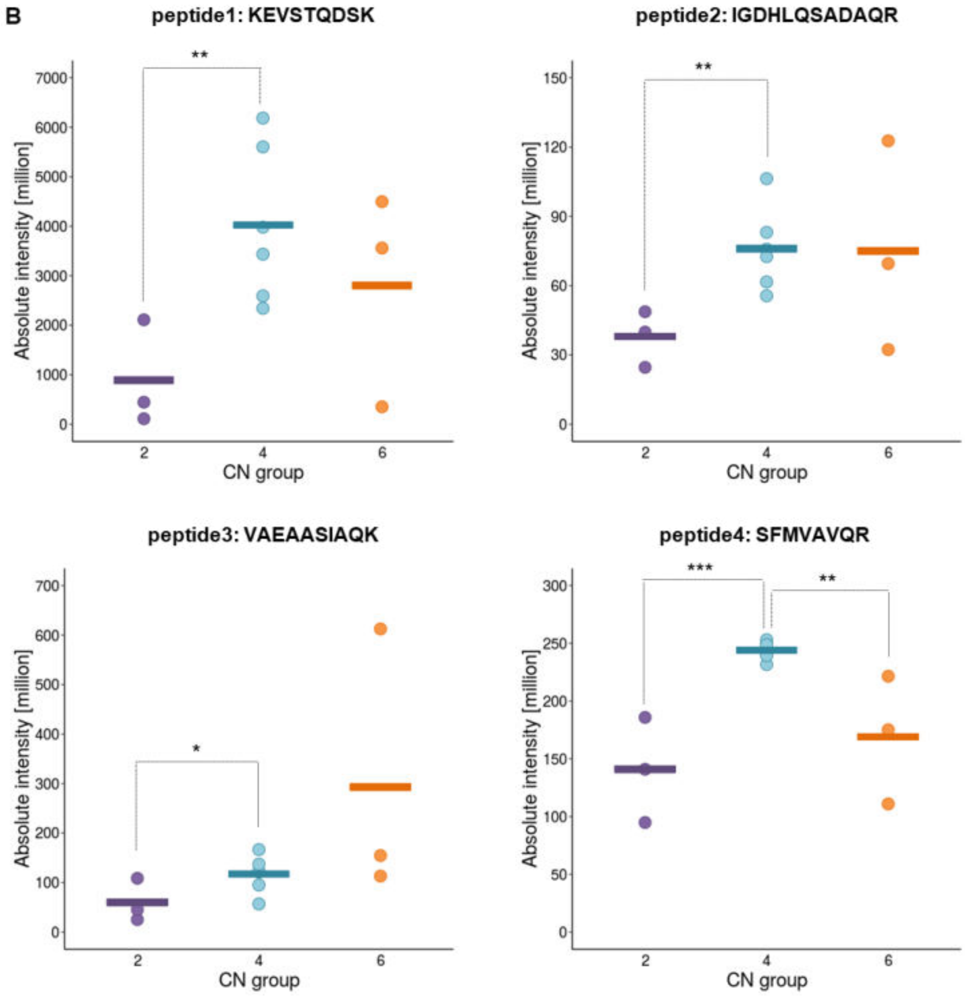
Supplemental Figure 16. The effect of stock misidentification on the correlation of sequencing-based (source data from the 1001 Genomes Project) and in-house experimental genotyping results (Supports Figure 7). Out of 1,135 *A. thaliana* stocks distributed by ABRC/NASC as 1001 Genomes Collection, 74 have been reported as „suspicious", following their resequencing and analysis with SNPmatch tool. Among 346 samples initially used for MLPA experiment, 32 belonged to the „suspicious" list. A, The correlation of MLPA data and read depth-based copy number estimates for 346 accessions. Accessions marked in red were previously reported as „suspicious". B, Removal of all „suspicious" samples improved the correlation between both genotyping methods. CN – copy number estimate, derived from sequencing data. MLPA signal was scaled to the value of 2 in Col-0 accession, that is representative of the reference genotype**.** *r* – Pearson correlation coefficient; R^2 – coefficient of determination of linear regression line.



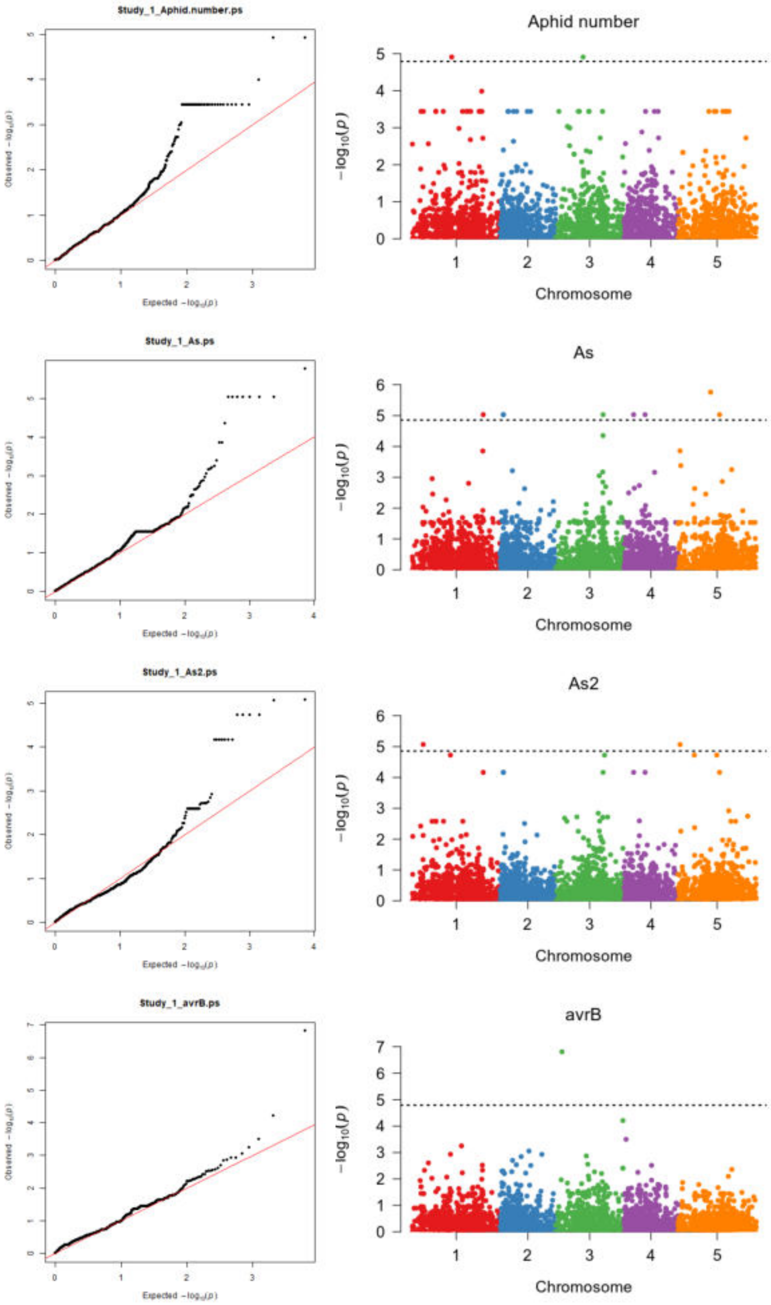
Supplemental Figure 17. Histograms of gene copy number distribution for CNV-genes that are likely duplicated in the Col-0 accession (Supports Figure 10)**.** Arrows point to bins including Col-0 accession.

A

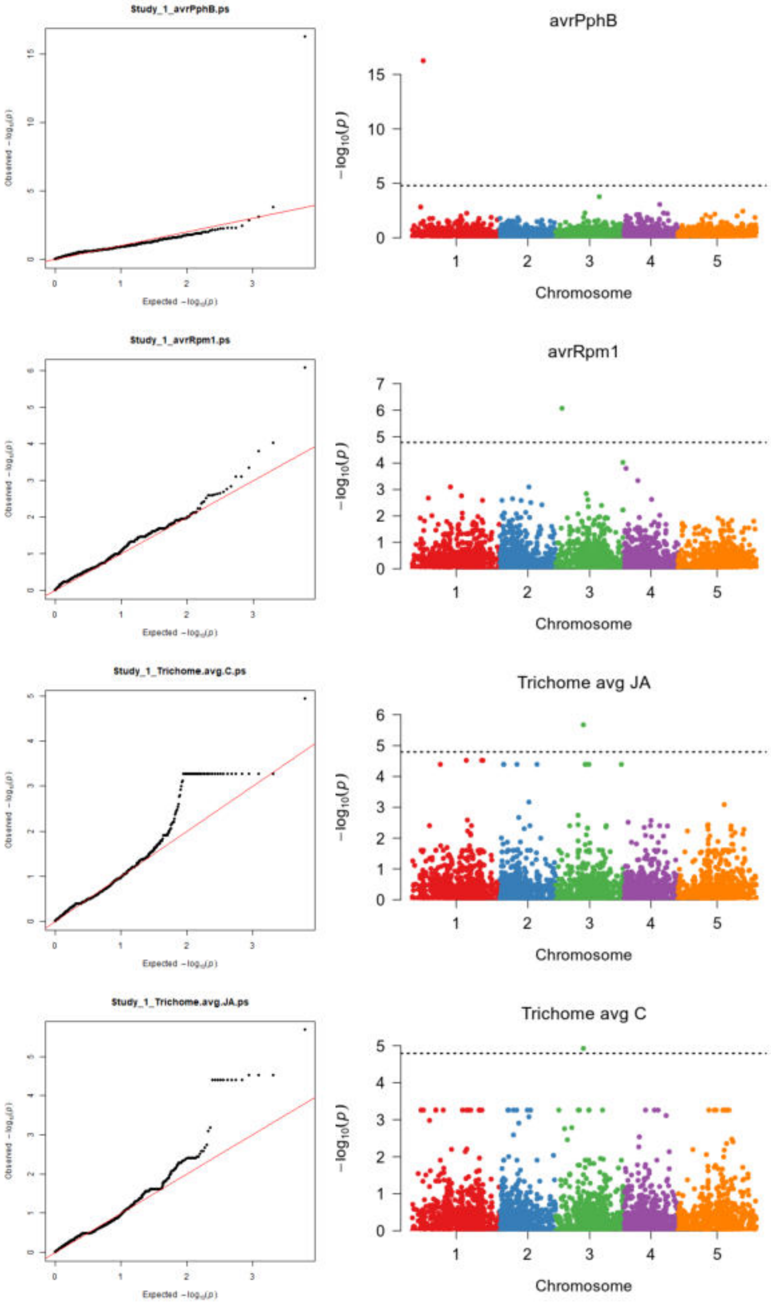
>SEC10_ARATH
MTEGIRARGPRSSSVNSVPLILDIEDFKGDFSFDALFGNLVNDLLPSFLDEEADSGDGHGNIAGVDGLTNGHLRGQSAPLSSAPFFPEVDGLLSLFK
DACKELVDLRKQVDGRLNTLK**KEVSTQDSK**HRKTLTEIEKGVDGLFESFARLDGRISSVGQTAAK**IGDHLQSADAQR**ETASQTIDLIKYLMEFNGSP
GDLMELSALFSDDSR**VAEAASIAQK**LRSFAEEDIGRQGASAAAGNATPGRGLEVAVANLQDYCNELENRLLSRFDAASQRRDLSTMSECAKILSQFN
RGTSAMQHYVATRPMFIDVEVMNSDIRLVLGDHGSQPSPSNVARGLSALFKEITDTVRKEAATITAVFPTPNEVMAILVQRVLEQRVTGILDKILAK
PSLMSPPPVQEGGLLLYLRMLAVAYERTQELAKDLRAVGCGDLDVEDLTESLFSSHKDEYPEHERASLKQLYQAKMEELRAESQQVSESSGTIGRSK
GASISSSLQQISVTVVTDFVRWNEEAITRCTLFSSQPATLAANVKAIFTCLLDQVSVYITEGLERARDSLSEAAALRERFVLGRRVAAAAASAAEAA
AAAGESSFK**SFMVAVQR**CGSSVAIVQQYFANSISRLLLPVDGAHAASCEEMSTALSKAEAAAYKGLQQCIETVMAEVDRLLSSEQKSTDYRSTDDGI
ASDHRPTNACIRVVAYLSRVLESAFTALEGLNKQAFLTELGNRLEKLLLTHWQKFTFNPSGGLRLKRDLNEYVGFVKSFGAPSVDEKFELLGIIANV
FIVAPDSLPTLFEGSPSIRKDAQRFIQLREDYKSAKLATKLSSLWPSLS



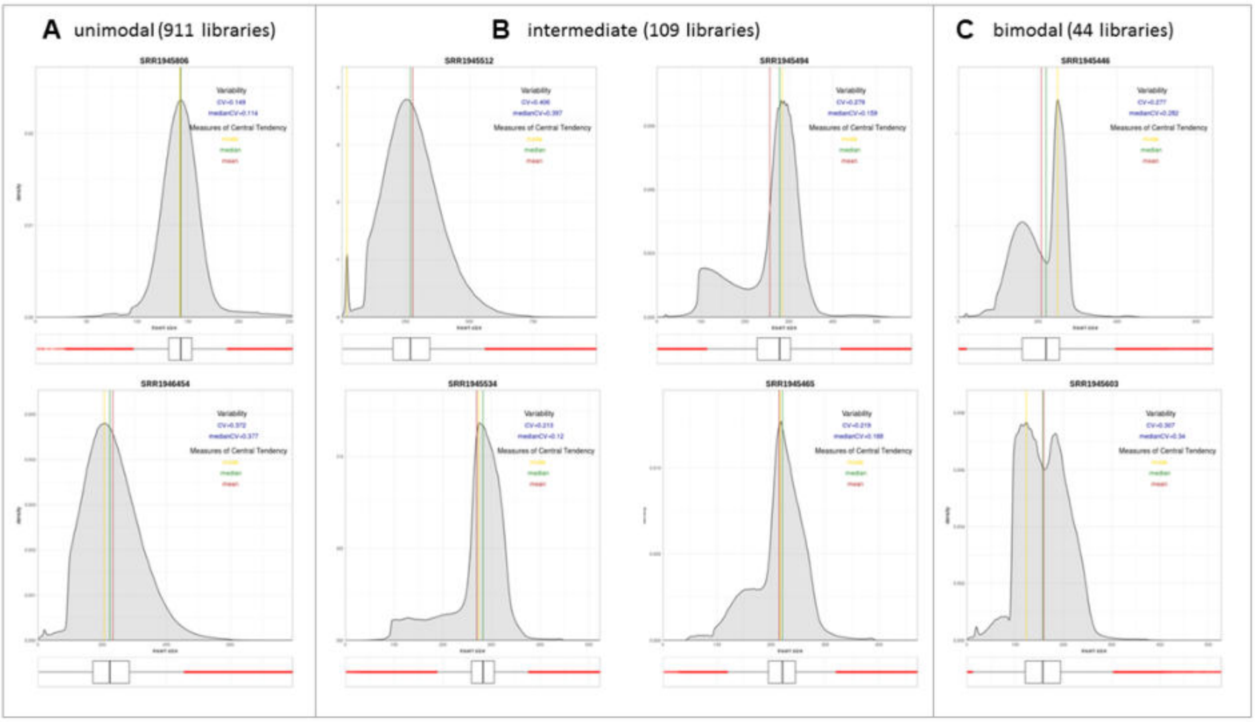
Supplemental Figure 18. Results of mass spectrometry-based identification of SEC10 peptides (Supports Figure 10). A, SEC10 peptides identified by mass spectrometry analysis. B, quantification results of each peptide in individual accessions (dots) and averaged in groups with distinct copy numbers (horizontal lines). Groups are colored according to copy number classes manually assigned from read depth-based genotyping data: CN=2 (purple), CN=4 (blue) and CN=6 (orange). Asterisks indicate significant Student's t-test results for the indicated groups (*p<0.1, **p<0.05, ***p<0.01).



Supplemental Figure 19. (Continued on the next page)



Supplemental Figure 19. Results from GWAS of defense-related phenotypes and CNV-gene data. (Supports Figure 11). Phenotypes for which significant associations were found are presented. Left – QQ=plots of P-value distributions. Right – Manhattan plots of Bonferroni-corrected P-values from EMMAX. Manhattan plots for AvrPphB, avrB, avrRpm1 phenotypes are also presented in Figure 11 in the main text.



Supplemental Figure 20. Insert size distributions in paired-end libraries (Supports Figure 1). Exemplar plots for libraries with A, unimodal; B, intermediate and C, bimodal distribution. Libraries with clear bimodal distribution (C) were excluded from the analysis with BreakDancer, VariationHunter and Pindel tools, which required insert size ranges as an input. Boxplots show median (inner line) and inner quartiles (box).

Supplemental Table 1. Variants >0.5 kb in size considered to be copy number changes discovered by each caller in the *A. thaliana* population.

	CNVnator	**Control-FREEC**	**BreakDancer**	**Variation Hunter**	**Pindel**	**Genome STRiP- CNV**[#]	**Genome-STRiP-SV**
Discovery mode	read depth	read depth	read pair	read pair	split read	read depth	hybrid
No. of accessions	1,064	1,064	997	997	997	1,064	1,064
Callset generation	per sample	per sample	per sample	per sample	per sample	population	population
Detected variants:	total/merged	total/merged	total/merged[$]	total/merged[$]	total/merged[$]		
duplications-gains-mCNVs	63,868/1,992	179,080/3,564	-	-	-	8,671	-
deletions-losses	532,384/12,089	164,397/2,554	911,205/14,766	927,173/21,378	1,302,725/14,237	1,674	14,985
insertions	-	-	7,312/1,235	0/0	0/0	-	-
tandem duplications	-	-	-	-	705,691/8,579	-	-

[#] GenomeStrip's CNV pipeline classifies the identified variants into three types, of which bi-allelic duplications and mCNVs (regions with either multiple copy-number gain alleles or regions with both gains and losses detected among samples) are reported as duplications-gains-mCNVs in this comparison table while bi-allelic deletions only are reported as deletions-losses

[$] For BreakDancer, VariationHunter and Pindel the between-samples merging of total raw calls detected by individual tool was performed only for this comparison purposes, while individual variants were used in the final breakpoint refinement.

Supplemental Table 2. CNVs resulting from the inter-tool merging of variants (80% RO) and their support by individual callers

SOURCE	No. of variants	≥2 callers	CNVnator	Control-FREEC	Genome STRiP-CNV	Genome-STRiP-SV	Break-Dancer	Variation-Hunter	Pindel
CNVnator	13,751	63.6%	100.0%	16.4%	5.1%	23.0%	41.0%	45.1%	28.1%
ControlFreec	5,728	67.7%	39.4%	100.0%	2.0%	15.0%	28.5%	39.2%	29.7%
GenomeStrip-CNV	9,664	19.1%	7.3%	1.2%	100.0%	5.3%	9.2%	8.9%	11.7%
GenomeStrip-SV	12,005	94.3%	26.4%	7.1%	4.3%	100.0%	77.5%	78.5%	40.2%
MERGED	34,368	19,003							

Supplemental Table 3. Gene family specificity of CNV-genes

	Gene Families		**Genes**	
	All$	CNV-genes	All$	CNV-genes
orphan genes:			**1170**	**497**
gene families:#				
A. thaliana	55	49 (89.1%)	123	93
Camelinae (3 species)	378	185 (48.9%)	+473	+238
Brassicales (8 species)	1347	448 (33.3%)	+1784	+513
Rosids (33 species)	1438	479 (33.3%)	+230	+93
Spermatophyta (50 species)	2488	733 (29.5%)	+2701	+559
Viridiplantae (55 species)	8662	2313 (26.7%)	+20964	+3265

#For each clade indicated in the Table, gene families present in at least one species within this clad, but absent from outer species, were counted. Counts were made based on Plaza v.4.0 database

$All protein coding genes encoded in the nuclear genome, based on the most recent (Araport 11) annotation of *A. thaliana* genome

Supplemental Table 4 . Superfamily composition of *A. thaliana* TEs and its comparison with CNV-TEs and CNV-TEs located within +/- 2kb distance from the genes

Superfamily	total TEs	CNV-TEs	CNV-TEs proximal to genes
RC/Helitron	41.51%	35.72%	45.88%
DNA/MuDR	17.35%	17.33%	19.68%
LTR/Gypsy	13.41%	18.69%	7.00%
DNA	5.86%	5.56%	3.40%
LTR/Copia	5.71%	6.26%	6.33%
LINE/L1	4.38%	4.62%	4.93%
DNA/HAT	3.32%	3.31%	4.33%
DNA/En-Spm	3.02%	3.71%	2.19%
DNA/Harbinger	1.22%	1.40%	1.76%
DNA/Pogo	1.10%	0.92%	1.28%
RathE1_cons	0.68%	0.49%	0.67%
DNA/Mariner	0.48%	0.35%	0.48%
SINE	0.42%	0.38%	0.47%
Unassigned	0.41%	0.51%	0.52%
RathE3_cons	0.33%	0.24%	0.33%
DNA/Tc1	0.30%	0.20%	0.27%
LINE?	0.26%	0.17%	0.25%
RathE2_cons	0.24%	0.15%	0.23%

Supplemental Table 5. Effect of excluding suspicious stocks on the correlation of read depth-based and MLPA-based genotyping results.

Gene	Coefficient of determination of linear regression (R^2)		
	346 accessions	314 accessions (suspicious stocks excluded)	change
AT5G05050	0.746	0.962	+0.216
AT5G61700	0.708	0.912	+0.204
AT1G31910	0.739	0.94	+0.201
AT1G27570	0.783	0.959	+0.176
AT4G19520	0.802	0.973	+0.171
AT5G14800	0.801	0.964	+0.163
AT1G23935	0.848	0.973	+0.125
AT3G24540	0.717	0.842	+0.125
AT1G59660	0.816	0.939	+0.123
AT4G18310	0.678	0.796	+0.118
AT3G60970	0.781	0.898	+0.117
AT4G11240	0.735	0.851	+0.116
AT2G31082	0.632	0.748	+0.115
AT1G02250	0.591	0.695	+0.105
AT4G15360	0.275	0.373	+0.098
AT2G19960	0.784	0.874	+0.090
AT4G27080	0.855	0.944	+0.089
AT1G52950	0.858	0.945	+0.087
AT3G05530	0.646	0.726	+0.080
AT5G54710	0.724	0.804	+0.080
AT2G25450	0.529	0.601	+0.072
AT2G05642	0.86	0.93	+0.069
AT4G08593	0.537	0.604	+0.067
AT3G44250	0.546	0.607	+0.061
AT4G12020	0.674	0.727	+0.053
AT1G09995	0.868	0.917	+0.049
AT4G08990	0.741	0.79	+0.049
AT1G10380	0.723	0.766	+0.044
AT4G37685	0.756	0.798	+0.042
AT3G06440	0.681	0.713	+0.032
AT2G38950	0.605	0.629	+0.024
AT3G21970	0.825	0.831	+0.006
AT5G07280	0.566	0.521	-0.045

Supplemental Table 6. Coefficients of variation (CVs) of read depth values in Control-FREEC analysis.

CV	No. of samples
<0.05	106
0.05-0.1	919
>0.1	39

CVs were calculated for a sliding window of 800 nt with 400 nt step.

Supplemental Table 7. List of genomic regions targeted by MLPA probes

Locus	Overlapping variants from AthCNV datset	Target region (Araport 11 coordinates)	Length
AT1G02250	CNV_16	Chr1:439093-439144	52
AT1G09995	CNV_117; CNV_118	Chr1:3262555-3262623	69
AT1G10380	CNV_130	Chr1:3401635-3401692	58
AT1G23935	CNV_447; CNV_448; CNV_450; CNV_451; CNV_452; CNV_453	Chr1:8463614-8463665	52
AT1G27570	CNV_557; CNV_558; CNV_559; CNv_560	Chr1:9577003-9577055	53
AT1G31910	CNV_742; CNV_746; CNV_747; CNV_748; CNV_749; CNV_750; CNV_751	Chr1:11460205-11460259	55
AT1G52950	CNV_2975; CNV_2976	Chr1:19726669-19726721	53
AT1G59660	CNV_3235; CNV_3236; CNV_3237	Chr1:21926568-21926627	60
AT1G60930	CNV_3322	Chr1:22435705-22435763	59
AT1G80840	CNV_3994	Chr1:30385066-30385114	49
AT2G02300	CNV_4066; CNV_4067	Chr2:606776-606839	64
AT2G05642	CNV_4473; CNV_4476; CNV_4486; CNV_4488; CNV_4489	Chr2:2099659-2099717	59
AT2G19110	---	Chr2:8281024-8281080	57
AT2G19960	CNV_7024; CNV_7025	Chr2:8622418-8622487	70
AT2G25450	CNV_7275; CNV_7276	Chr2:10830834-10830899	66
AT2G31082	CNV_7421	Chr2:13241594-13241648	55
AT2G34380	CNV_7484	Chr2:14512989-14513046	58
AT2G36230	---	Chr2:15194440-15194490	51
AT2G38950	CNV_7546; CNV_7548	Chr2:16262634-16262693	60
AT2G39020	CNV_7551	Chr2:16295593-16295646	54
AT3G05350	CNV_7758	Chr3:1529078-1529139	62
AT3G05410	CNV_7759; CNV_7760; CNV_7761	Chr3:1555222-1555272	51
AT3G05410	---	Chr3:1559701-1559756	56
AT3G05530	CNV_7764; CNV_7766	Chr3:1604428-1604483	56
AT3G06440	CNV_7775; CNV_7778; CNV_7781; CNV_7782;	Chr3:1975155-1975212	58
AT3G18524	CNV_7984; CNV_7985	Chr3:6371322-6371371	50
AT3G18535	CNV_7984; CNV_7985; CNV_7986	Chr3:6375946-6376000	55
AT3G21970	CNV_8073; CNV_8080; CNV_8083; CNV_8084; CNV_8085; CNV_8086	Chr3:7743031-7743081	51
AT3G24540	CNV_8192; CNV_8193	Chr3:8953616-8953677	62
AT3G44250	CNV_11089; CNV_11164; CNV_11171; CNV_11178; CNV_11180; CNV_11181	Chr3:15948801-15948854	54
AT3G57810	CNV_11902	Chr3:21416393-21416438	46
AT3G60190	CNV_11970	Chr3:22246645-22246697	53
AT3G60970	CNV_11979; CNV_11986; CNV_11987; CNV_11992; CNV_11993; CNV_11994	Chr3:22558207-22558263	57
AT4G08593	CNV_13992	Chr4:5470316-5470386	71
AT4G08990	CNV_14088; CNV_14098; CNV_14090; CNV_14091	Chr4:5766958-5767009	52
AT4G11240	CNV_14348; CNV_14349	Chr4:6848068-6848130	63
AT4G12020	CNV_14403; CNV_14404; CNV_14405; CNV_14406; CNV_14407; CNV_14408	Chr4:7205074-7205137	64
AT4G14410	---	Chr4:8300694-8300752	59
AT4G15360	CNV_14660; CNV_14661	Chr4:8771306-8771369	64
AT4G18310	CNV_14836	Chr4:10122445-10122498	54
AT4G19140	CNV_14861	Chr4:10470991-10471053	63
AT4G19520	CNV_14878; CNV_14883; CNV_14884; CNV_14885	Chr4:10641616-10641668	53
AT4G27080	CNV_15127; CNV_15128; CNV_15133; CNV_15134	Chr4:13592606-13592658	53
AT4G37685	CNV_15294	Chr4:17705274-17705327	54
AT5G05050	CNV_15397; CNV_15405	Chr5:1491828-1491900	73

AT5G07280	CNV_15435	Chr5:2285229-2285273	45
AT5G14800	CNV_15511	Chr5:4786508-4786556	49
AT5G37240	CNV_17860; CNV_17861; CNV_17862; CNV_17863; CNV_17864; CNV_17865	Chr5:14737918-14737982	65
AT5G54710	CNV_18769; CNV_18772; CNV_18773; CNV_18774; CNV_18775	Chr5:22228424-22228479	56
AT5G54770	---	Chr5:22246711-22246760	50
AT5G56570	CNV_18833	Chr5:22902935-22902991	57
AT5G61700	CNV_18925; CNV_18927	Chr5:24796111-24796161	51

Publikacja 3

Marszalek-Zenczak M, Satyr A, Wojciechowski P, Zenczak M, Sobieszczanska P, Brzezinski K, Iefimenko T, Figlerowicz M, Zmienko A.
Analysis of Arabidopsis non-reference accessions reveals high diversity of metabolic gene clusters and discovers new candidate cluster members
Front Plant Sci. 2023 Jan 26;14:1104303. doi: 10.3389/fpls.2023.1104303.

5-letni IF = 7,255

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 26 January 2023
DOI 10.3389/fpls.2023.1104303

Analysis of Arabidopsis non-reference accessions reveals high diversity of metabolic gene clusters and discovers new candidate cluster members

Malgorzata Marszalek-Zenczak[1], Anastasiia Satyr[1], Pawel Wojciechowski[1,2], Michal Zenczak[1], Paula Sobieszczanska[1], Krzysztof Brzezinski[1], Tetiana Iefimenko[3], Marek Figlerowicz[1] and Agnieszka Zmienko[1]*

[1]Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan, Poland, [2]Institute of Computing Science, Faculty of Computing and Telecommunications, Poznan University of Technology, Poznan, Poland, [3]Department of Biology, National University of Kyiv-Mohyla Academy, Kyiv, Ukraine

OPEN ACCESS

EDITED BY
Li Wang,
Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, China

REVIEWED BY
Mei Yang,
Wuhan Botanical Garden, Chinese Academy of Sciences (CAS), China
Aalt-Jan Van Dijk,
Wageningen University and Research, Netherlands
Ancheng Huang,
Southern University of Science and Technology, China

*CORRESPONDENCE
Agnieszka Zmienko
akisiel@ibch.poznan.pl

SPECIALTY SECTION
This article was submitted to Plant Metabolism and Chemodiversity, a section of the journal Frontiers in Plant Science

RECEIVED 21 November 2022
ACCEPTED 11 January 2023
PUBLISHED 26 January 2023

CITATION
Marszalek-Zenczak M, Satyr A, Wojciechowski P, Zenczak M, Sobieszczanska P, Brzezinski K, Iefimenko T, Figlerowicz M and Zmienko A (2023) Analysis of Arabidopsis non-reference accessions reveals high diversity of metabolic gene clusters and discovers new candidate cluster members. *Front. Plant Sci.* 14:1104303. doi: 10.3389/fpls.2023.1104303

Metabolic gene clusters (MGCs) are groups of genes involved in a common biosynthetic pathway. They are frequently formed in dynamic chromosomal regions, which may lead to intraspecies variation and cause phenotypic diversity. We examined copy number variations (CNVs) in four *Arabidopsis thaliana* MGCs in over one thousand accessions with experimental and bioinformatic approaches. Tirucalladienol and marneral gene clusters showed little variation, and the latter was fixed in the population. Thalianol and especially arabidiol/baruol gene clusters displayed substantial diversity. The compact version of the thalianol gene cluster was predominant and more conserved than the noncontiguous version. In the arabidiol/baruol cluster, we found a large genomic insertion containing divergent duplicates of the *CYP705A2* and *BARS1* genes. The *BARS1* paralog, which we named *BARS2*, encoded a novel oxidosqualene synthase. The expression of the entire arabidiol/baruol gene cluster was altered in the accessions with the duplication. Moreover, they presented different root growth dynamics and were associated with warmer climates compared to the reference-like accessions. In the entire genome, paired genes encoding terpene synthases and cytochrome P450 oxidases were more variable than their nonpaired counterparts. Our study highlights the role of dynamically evolving MGCs in plant adaptation and phenotypic diversity.

KEYWORDS

copy number variation, biosynthetic gene cluster, secondary metabolism, oxidosqualene cyclase, triterpenes, cytochrome P450

Introduction

Plants are able to produce a variety of low molecular weight organic compounds, which enhance their ability to compete and survive in nature. Secondary metabolites are not essential for plant growth and development. However, they are often multifunctional and may act both as plant growth regulators and be engaged in primary metabolism or plant protection (Isah, 2019; Erb and Kliebenstein, 2020). The ability to produce particular types of compounds is usually restricted to individual species or genera. Therefore, these compounds are enormously diverse and have a wide range of biological activities. In plants, genes involved in a common metabolic pathway are typically dispersed across the genome. In contrast, functionally related genes that encode the enzymes involved in specialized metabolite biosynthesis in bacteria and fungi are frequently coexpressed and organized in so-called operons (Boycheva et al., 2014; Nützmann et al., 2018). Similar gene organization units called biosynthetic gene clusters or metabolic gene clusters (MGCs) have recently been found in numerous plant species. MGCs have typically been defined as a group of three or more genes that i) encode a minimum of three different types of biosynthetic enzymes, ii) are involved in the consecutive steps of a specific metabolic pathway and iii) are localized in adjacent positions in the genome or are interspersed by a limited number of intervening (i.e., not functionally related) genes (Nützmann and Osbourn, 2014; Kautsar et al., 2017). A typical MGC contains a "signature" enzyme gene involved in the major (usually first) step of a biosynthetic pathway. In this step, the metabolite scaffold is generated that determines the class of the pathway products (e.g., terpenes or alkaloids). This scaffold is further modified by "tailoring" enzymes encoded by other clustered genes, e.g., cytochrome P450 oxidases (CYPs), acyltransferases or alcohol dehydrogenases. The contribution of other enzymes encoded by peripheral genes (i.e., located outside the MGC), and the connection network between different metabolite biosynthesis pathways may result in additional diversification of the biosynthetic products (Huang et al., 2019). Currently, there are over 30 known MGCs in plants from various phylogenetic clades, and new MGCs are being discovered. Their sizes range from 35 kb to several hundred kb. However, clusters of functionally related nonhomologous genes are still considered unusual in plant genomes.

In *Arabidopsis thaliana* (hereafter Arabidopsis), four MGCs have been discovered thus far (Supplemental Table S1). They are involved in the metabolism of specialized triterpenes: thalianol, marneral, tirucalladienol, arabidiol and baruol. Triterpenes constitute a large and diverse group of natural compounds derived from 2,3-oxidosqualene cyclization in a reaction catalyzed by oxidosqualene cyclases (OSCs) (Thimmappa et al., 2014). Out of 13 OSC genes known in the Arabidopsis genome, five (*THAS1*, *MRN1*, *PEN3*, *PEN1*, *BARS1*) are located within MGCs and encode the "signature" enzymes of the MGCs (Field and Osbourn, 2008; Field et al., 2011; Boutanaev et al., 2015). The thalianol gene cluster contains five members involved in thalianol production and in its conversion to another triterpene, thalianin (Fazio et al., 2004; Field and Osbourn, 2008; Huang et al., 2019). In the reference genome, this MGC is ~45 kb in size. The thalianol synthase gene *THAS1* as well as *CYP708A2*, *CYP705A5* and *AT5G47980* (BAHD acyltransferase) genes are tightly clustered together, with only one noncoding transcribed locus (*AT5G07035*) between them. The fifth member, acyltransferase *AT5G47950*, is separated from the rest of the cluster by *RABA4C* and *AT5G47970* intervening genes. The marneral gene cluster is ~35 kb in size and is the most compact plant MGC described to date. It is made up of three members: the marneral synthase gene *MRN1*, the marneral oxidase gene *CYP71A16* and the gene *CYP705A12*, whose function is unknown (Xiong et al., 2006; Field et al., 2011; Go et al., 2012). Additionally, there are three noncoding transcribed loci (*AT5G00580*, *AT5G06325* and *AT5G06335*) located between *CYP701A16* and *MRN1*. The tirucalladienol gene cluster is ~47 kb in size and includes five members: tirucalla-7,24-dien-3β-ol synthase gene *PEN3*, an uncharacterized acyltransferase gene *SCPL1*, which was identified based on its coexpression with *PEN3*, *CYP716A1*, which is involved in the hydroxylation of tirucalla-7,24-dien-3β-ol, as well as *AT5G36130* and *CYP716A2* (Morlacchi et al., 2009; Boutanaev et al., 2015; Wisecaver et al., 2017). The contiguity of this MGC is interrupted by four intervening genes (*CCB3*, *AT5G36125*, *HCF109* and *AT5G36160*) and the noncoding locus *AT5G05325*. The arabidiol/baruol gene cluster is most complex and has an estimated size of 83 kb. It encompasses two closely located OSCs, *PEN1* and *BARS1*, sharing 91% similarity at the amino acid level. *BARS1* encodes a multifunctional cyclase that produces baruol as its main product (Lodeiro et al., 2007). *PEN1* encodes arabidiol synthase and is adjacent to *CYP705A1*, which is involved in arabidiol degradation upon jasmonic acid treatment (Xiang et al., 2006; Castillo et al., 2013; Sohrabi et al., 2015). The role of the remaining genes in the arabidiol/baruol gene cluster (*CYP702A2*, *CYP702A3*, *CYP705A2*, *CYP705A3*, *CYP705A4*, *CYP702A5*, *CYP702A6* as well as acyltransferases *AT4G15390* and *BIA1*) has not been determined; however, they displayed coexpression with either *PEN1* or *BARS1* (Wada et al., 2012; Wisecaver et al., 2017). There are few intervening loci in the arabidiol/baruol gene cluster, including a protein-coding gene *CSLB06*, two pseudogenes *CYP702A4P* and *CYP702A7P* and one novel transcribed region *AT4G06325*.

Plant MGCs are thought to have arisen by duplication and subsequent neo- or subfunctionalization of genes involved in primary metabolism, which might have been followed by the recruitment of additional genes to the newly forming biosynthetic pathway (Nützmann and Osbourn, 2014). MGCs are frequently located within dynamic chromosomal regions, e.g., subtelomeric regions, centromeric regions or regions rich in transposable elements (TEs), where the possibility of bringing together the beneficial sets of genes by structural rearrangements may be higher than in the rest of the genome, thus promoting MGC formation (Field et al., 2011). However, the same factors may also contribute to further genetic modifications and alteration of the plant metabolic profile, thus making such MGCs "evolutionary hotspots". To verify this scenario, we evaluated the intraspecific diversity of Arabidopsis MGCs and examined whether this diversity is associated with trait variation. Here, we present a detailed picture of MGC copy number variations (CNVs), describe the discovery of novel, nonreference genes in the arabidiol/baruol gene cluster and reveal the links between the variation in MGC structure and plant adaptation to different natural environments.

Results

MGCs differ in levels of copy number polymorphism

We started our analysis by aligning each MGC with the common CNVs in the Arabidopsis genome, which were identified previously (Zmienko et al., 2020). As expected, each MGC had a substantial overlap with the variable regions: 100% for the thalianol gene cluster, 79.6% for the tirucalladienol gene cluster, 53.1% for the arabidiol/baruol gene cluster, and 52.8% for the marneral gene cluster (Figure 1A). However, the potential impact of CNVs on the clustered genes differed among the MGCs (Supplemental Figure S1; Supplemental Table S2). In the thalianol gene cluster, most CNVs were grouped in the region spanning *AT5G47980*, *CYP705A5*, *CYP708A2* and *THAS1*, while *AT5G47950* was covered only by the largest variant CNV_18592 (241 kb in size), which encompassed the entire cluster. In the arabidiol/baruol gene cluster, the CNVs (0.6 kb to 21 kb in size) were grouped into three distinct regions separated by invariable segments. The first variable region overlapped with *CYP702A2* and *CYP702A3*. The second variable region overlapped with *CYP705A2*, *CYP705A3* and *BARS1*. The CNVs in the third variable region were mostly intergenic and overlapped with only two genes, *CYP702A5* and *CYP702A6*. *CYP705A1, PEN1*, *CYP705A4*, *AT4G15390* and *BIA1* were not covered by any common CNV. In the tirucalladienol gene cluster, the CNVs accumulated in the 5' part of the cluster, and none of them overlapped with *SCPL1*. Notably, upstream of the tirucalladienol gene cluster, a region genetically divergent from the surrounding genomic segments, called a hotspot of rearrangements, was previously described (Jiao and Schneeberger, 2020). Smaller hotspots of rearrangements were also found between *CYP716A1* and *AT5G36130* in the same MGC as well as in one variable segment of the arabidiol/baruol gene cluster. It was demonstrated that the hotspots of rearrangements are highly variable in the Arabidopsis population, which was in agreement with the observed increased CNV rate in these genomic regions. The CNV arrangement in the marneral gene cluster was strikingly different from that in any other MGC in that all variants were intergenic and did not overlap with the marneral cluster genes.

For each MGC, there were CNVs that overlapped only part of the cluster. This indicated that in some accessions, gene deletions/duplications might have altered MGC composition and consequently affected the entire biosynthetic pathway. To evaluate this possibility, we retrieved copy number data for 31 genes (clustered

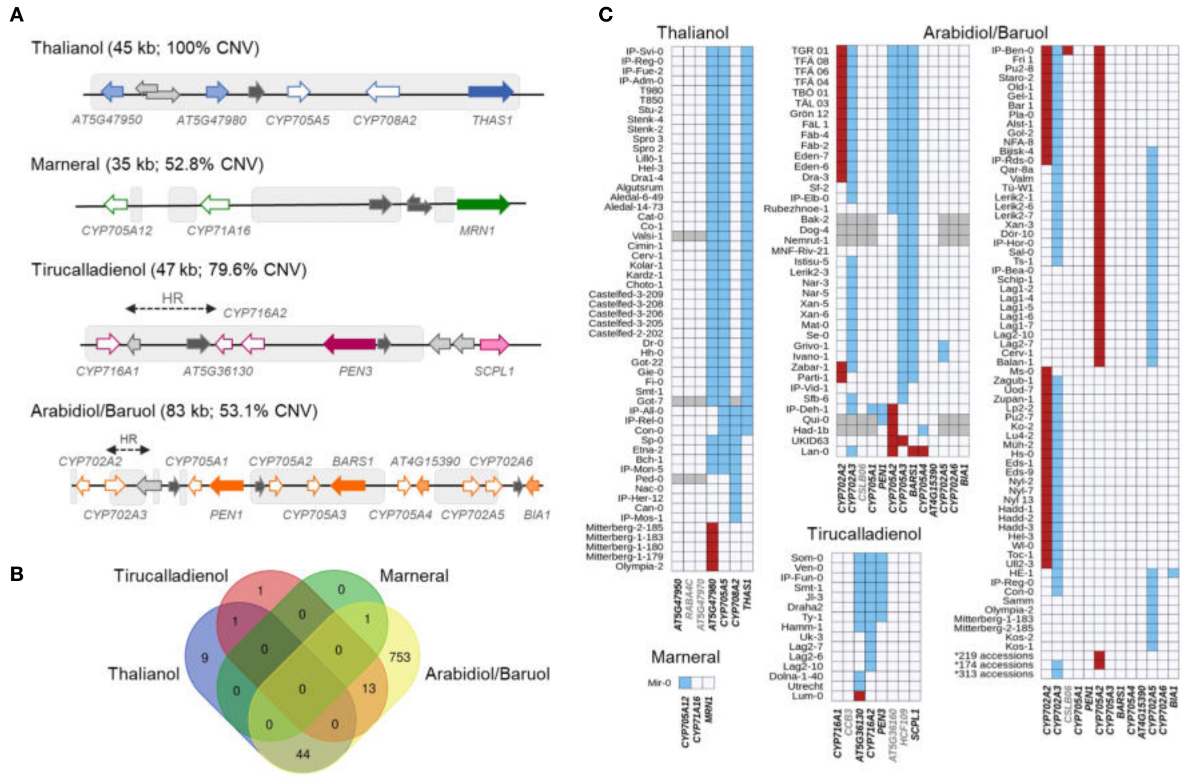


FIGURE 1
Copy number variation of Arabidopsis metabolic gene clusters. **(A)** MGC overlap with CNV regions. Colored arrows with white filling denote CYPs. Arrows with dark color filling denote OSCs. Arrows with light color filling denote other types of MGC genes. Intervening genes are in light grey. Noncoding genes are in dark grey. Grey boxes indicate overlap with CNV regions. HR – hotspot of rearrangements; **(B)** Number and overlap among the accessions with detected gene copy numbers in each of four MGCs; **(C)** Patterns of gene copy number variation in each MGC. Red – gain; blue – loss, grey – no assignment. Names of the genes considered as MGC members are in black; names of the intervening genes are in grey. Source data for histograms are in Supplemental Table S6.

and intervening genes in all MGCs), each from 1,056 accessions (RD dataset; Supplemental Table S3), and supplemented them with multiplex ligation-dependent amplification assays for 232 accessions (MLPA dataset; Supplemental Table S4) and droplet digital PCR-based genotyping assays for 20 accessions (ddPCR dataset; Supplemental Table S5). We defined the thresholds for detecting duplications and deletions for each data type. Next, we assigned the copy number status of each gene in each accession ("REF", "LOSS" or "GAIN") by combining all three datasets (Supplemental Table S6). Out of the genotypes assigned with two or three approaches, 98.8% were fully concordant, and most of the remaining discrepancies could be resolved manually (Supplemental Figures S2-S4; Supplemental Table S7). The combined genotyping data for 1,152 accessions were further used to assess and compare MGC variation at the gene level.

Only 28.6% of the assayed accessions had no gene gains or losses in any MGC (Figure 1B). This included 65% of accessions from the German genetic group and 39% of accessions from the Central Europe group. In contrast, the vast majority (at least 90%) of accessions from groups known to be genetically distant from the reference genome (North Sweden, Spain, Italy-Balkan-Caucasus, and Relict groups) displayed gene CNV in at least one MGC. We note that the real number of invariable accessions could be even lower since for 96 accessions, some MGC genes were not genotyped. Altogether, 19 genes were affected: four in the thalianol cluster, one in the marneral cluster, three in the tirucalladienol cluster and 11 in the arabidiol/baruol cluster (Figure 1C). The latter was also most variable in terms of the number of accessions carrying CNVs and the diversity of CNV patterns. For two genes, we detected only copy gains, and for 11, we detected only losses, while six genes were multiallelic (with both gains and losses). As expected, these genes resided in the previously defined variable regions. Remarkably, we did not observe complete loss or gain of the entire MGC in any accession. In the next step, we inspected in more detail the level of diversity of each MGC.

The compact version of the thalianol gene cluster is predominant and more conserved than the reference-like noncontiguous version

A survey with a combination of RD, MLPA and ddPCR approaches revealed 54 accessions with copy number changes in the thalianol gene cluster, which followed five distinct patterns, and *AT5G47950* was the only invariant gene in all accessions (Figure 2A). The most common (variant A) was the deletion of a region encompassing *AT5G47980* and *CYP705A5*, combined with the deletion of *THAS1*. We detected this variant in 37 accessions from six countries: Sweden (13), Italy (8), Germany (6), Spain (5), Bulgaria (3) and Portugal (2). We also confirmed the existence of two previously reported rare variants (Liu et al., 2020a). One of them (variant B) was a large deletion spanning *AT5G47980, CYP705A5* and *CYP708A2*. We found this variant in two accessions from Germany (Bch-1, Sp-0), in one from Italy (Etna-2) and in one from Spain (IP-Mon-5). The other one (variant C) was a deletion of a single gene, *CYP708A2*, which we found in five accessions, mainly Relicts, originating from Spain (Can-0, Ped-0, IP-Her-12 and Nac-0) and Portugal (IP-Mos-1). We also found a new type of deletion (variant D) in two Spanish Relicts (IP-Rel-0 and Con-0) and one non-Relict (IP-All-0). The deletion spanned *CYP705A5*, *CYP708A2* and *THAS1* (Supplemental Figure S5). The last variant (variant E) was a duplication of the acyltransferase gene *AT5G47980*, which was found in four accessions from Italy (Mitterberg-1-179, Mitterberg-

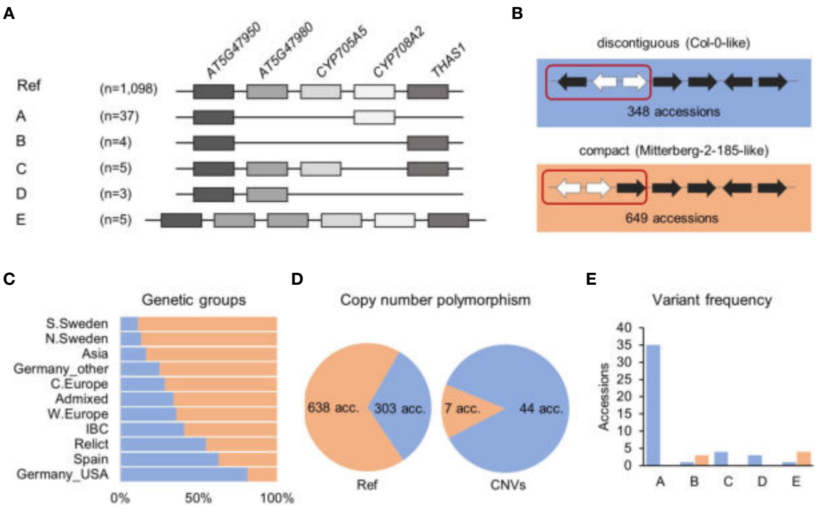


FIGURE 2
Structural variation of thalianol gene cluster. **(A)** Five types of CNVs that change the number of thalianol cluster genes. The position of intervening genes is ignored and they are not shown. Gene orientation is disregarded. **(B)** Two versions of thalianol gene cluster organization. Clustered genes are in black; interfering genes are in white. **(C)** The frequency of the two thalianol gene cluster versions (discontiguous and compact) among the genetic groups. **(D)** Rate of copy number polymorphism within discontiguous and compact clusters. **(E)** Frequency of variants presented in **(A)** among the accessions with different cluster organizations. The number of presented accessions in panels is 1,152 for **(A)** – genotyping, 997 for **(B, C)** – inversion detection and 992 for **(D, E)** – the intersection of the above.

1-180, Mitterberg-1-183, Mitterberg-2-185) and one from Greece (Olympia-2). The presence of a tandem duplication ~3 kb in size in Mitterberg-2-185 was confirmed by sequence analysis of its *de novo* genomic assembly (Supplemental Figure S6). The duplication spanned the entire *AT5G47980* and its flanks (0.5 kb upstream and 0.7 kb downstream) and differed from its copy only by two mismatches and a 1-bp gap. The predicted protein products of both gene copies were identical and shorter than the reference acyltransferase (404 aa versus 443 aa), but they possessed complete transferase domains (pfam02458).

In the Mitterberg-2-185 assembly, we also detected a chromosomal inversion (with respect to the reference genome orientation) spanning *AT5G47950* and two intervening genes, *RABA4C* and *AT5G47970*. This resulted in a more compact cluster organization compared to the reference (Figure 2B). Similar inversions were previously detected in 17 other accessions (out of 22 analyzed), which indicated that the compact version of the thalianol gene cluster might be predominant in Arabidopsis (Liu et al., 2020a). To verify this possibility, we set up a bioinformatic pipeline for detecting genomic inversions based on paired-end genomic read analysis in 997 accessions. We correctly detected inversions in 12 out of 15 previously analyzed accessions, which indicated the good sensitivity of our method. Altogether, we found inversions, 12.8 kb to 15.4 kb in size, spanning the *AT5G47950*, *RBAA4C* and A*T5G47970* genes in 649 accessions (65%), which fully confirmed our predictions (Supplemental Table S8). The compact version of the thalianol gene cluster was dominant in the South and North Sweden genetic groups as well as in the Asia group (83.6% to 88.9%), while the discontiguous version was mainly observed among the U.S.A. accessions and was also slightly more abundant in the Spain genetic group (Figure 2C). There was a similar frequency of discontiguous and compact versions among the Relicts (12 and 10 accessions, respectively). Interestingly, the CNV frequency substantially differed between the accessions with different cluster organization (Figures 2D, E). The compact cluster was more conserved; copy number changes (variants B and E) affected only 1.1% of the accessions in this group. The remaining variants, including deletions spanning the *THAS1* signature gene, were found exclusively among the accessions with the reference-like cluster type. Altogether, 12.7% of accessions with discontiguous clusters were affected by CNVs.

Marneral and tirucalladienol gene clusters display little structural variation

Analysis of RD and MLPA data confirmed exceptionally low variability of marneral cluster genes. One private variant, which we detected in Mir-0 and confirmed by Sanger sequencing, was 1.2 kb in size and spanned the first exon of the *CYP705A12* gene, which resulted in the truncation of its predicted protein product (Supplemental Figure S7). Apart from that, we did not detect any common gene duplications or deletions within this MGC. Likewise, we observed low variation in the tirucalladienol gene cluster. In 15 accessions (1.4%), deletions or duplications occurred in the region spanning the *AT5G36130*, *CYP716A2* and *PEN3* genes and affected one, two or all of them. Differences between the countries indicated that these structural variants were of local origin (Supplemental Figure 8). Sequence analysis of *de novo* genomic assemblies for Ty-1 and Dolna-1-40 confirmed the predicted deletion patterns in these accessions. It should be noted that, according to a recent study, *AT5G36130* and *CYP716A2* gene models are misannotated, and they jointly encode a single protein of the CYP716A subfamily with cytochrome oxidase activity (Yasumoto et al., 2016) (Supplemental Figure S9). Therefore, a full-length gene was absent from all 15 accessions with CNVs in the tirucalladienol gene cluster (Figure 1C).

Intraspecies variation in the arabidiol/baruol gene cluster reveals a novel OSC gene

The arabidiol/baruol gene cluster was the most heterogeneous of all the MGCs. Consistent with the segmental CNV coverage, there were apparent differences in the variation frequency between the genes. At the cluster's 5' end, *CYP702A2* was duplicated in 50 accessions, and *CYP702A3* was deleted in 564 accessions, including approximately 70% of all analyzed accessions from Sweden and Spain. In contrast, genes located at the 3' end of the cluster showed little variation. There were *CYP702A5* deletions in 35 accessions, *CYP705A4* deletions in two accessions, and *BIA1* deletion in one accession, while *CYP702A6* and *AT4G15390* were invariable in copy number.

The two OSCs, *PEN1* and *BARS1*, were located in segments with opposite variation levels. *PEN1* and the neighboring gene *CYP705A1*, both implicated in the arabidiol biosynthesis pathway, were stable in copy number, except for three accessions with full or partial gene deletions: the Qui-0 and IP-Deh-1 accessions from Spain and the Kyoto accession from Japan. In the latter, we confirmed partial deletion of both genes by analysis of its *de novo* genomic assembly (Jiao and Schneeberger, 2020). In contrast, *BARS1*, *CYP705A2* and *CYP705A3* were all deleted in several accessions originating from Sweden. We also observed smaller deletions or duplications in this genomic segment, of which the most remarkable was the duplication of *CYP705A2*, detected in 433 (37.6%) accessions. Since the genotypic data for *CYP705A2* and *BARS1* were noisy and indicated more variation than could be revealed by our standard genotyping, we manually inspected short read genomic data that mapped in this region (examples are presented in Supplemental Figure S10). In most accessions, *BARS1* lacked the largest intron, where the *ATREP11* TE (RC/Helitron superfamily) is annotated, which might explain the lower RD values for *BARS1* compared to other genes (see Supplemental Figure S3). Surprisingly, we also observed a mix of reads mapping to *CYP705A2* and *BARS1* loci with and without mismatches in a large number of accessions. Thus, we called SNPs in the coding sequences of both genes to obtain more information on their diversity. Numerous heterozygous SNPs were called in both genes in the above accessions. Because Arabidopsis is a self-pollinating species and therefore highly homozygous, we hypothesized that the reads with mismatches originated from duplicated loci, which showed similarity to *CYP705A2* and *BARS1* and mapped to the reference gene models, resulting in heterozygous SNP calls. In support of this hypothesis, we detected heterozygous SNPs at the *CYP705A2* locus in 90.6% of accessions with this gene's duplication but only in 10.7% of accessions without changes in its copy number (Wilcoxon rank sum test with continuity correction, *p* value $<2.2\times10^{-16}$; Supplemental Figure S11A). Additionally, heterozygous SNPs at the *BARS1* locus were present in the same accessions (Pearson's correlation coefficient r = 0.86; Supplemental Figure S11B), although we found only one duplication of *BARS1* with our genotyping methods. We concluded that the sequence differences between *BARS1* and its duplicate prevented its detection by RD or MLPA assays. We also observed low but nonzero read coverage and homozygous SNPs at both loci in some accessions with intermediate RD values for *CYP705A2* (RD_{mean} = 1.5) and *BARS1* (RD_{mean} = 0.6) and with the clear loss of *CYP705A3* (RD_{mean} = 0). In agreement with the gene duplication scenario, this could be explained by the presence of *CYP705A2* and *BARS1* duplicates but absence of the entire region spanning the reference genes *CYP705A2*, *CYP705A3* and *BARS1*.

To identify the cryptic *BARS1* duplication, we analyzed genomic assemblies of seven accessions: An-1, Cvi-0, Kyoto, Ler-0, C24, Eri-1 and Sha (Jiao and Schneeberger, 2020), four of which were also genotyped in our study (Figure 3A). We reannotated the entire arabidiol/baruol cluster region in each accession and compared it with the reference (Supplemental Table S9). In six accessions, *BARS1* lacked the largest intron, as indicated earlier by short read data (Supplemental Figure S12). In the Cvi-0, Eri-1 and Ler-0 accessions,

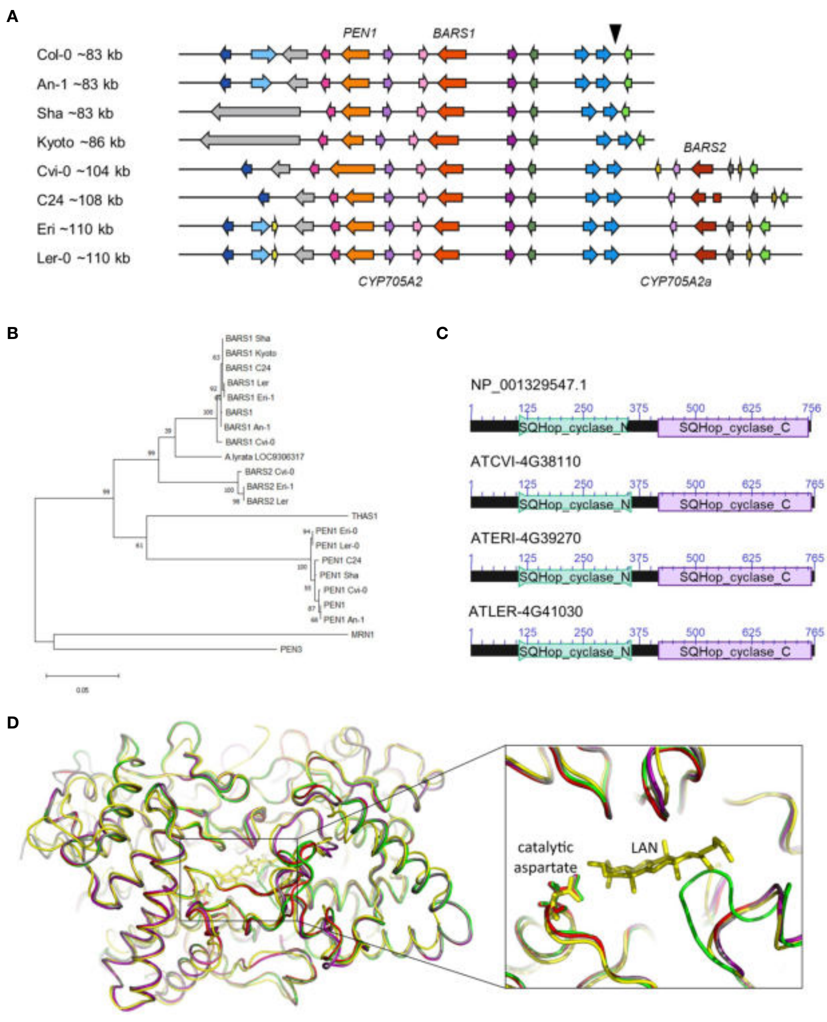


FIGURE 3
BARS2 is a *BARS1* duplicate absent from the reference genome and encodes oxidosqualene synthase. **(A)** Organization of arabidiol/baruol gene cluster in Col-0 and seven nonreference accessions. The genomic insertion including *CYP705A2a* and *BARS2* genes is marked with a triangle above the reference cluster. **(B)** Phylogeny of amino acid sequences of clade II OSCs residing in clusters. BARS1 ortholog from *A.lyrata* (LOC9306317) is included. The maximum likelihood tree was generated using the MEGA11 package with Jones-Taylor (JTT) substitution matrix and uniform rates among sites. Values along branches are frequencies obtained from 1000 bootstrap replications. **(C)** Conserved protein domains encoded in *BARS1* (Col-0) and *BARS2* (Cvi-0, Eri-1, Ler-0) genes. SQHop_cyclase_N - squalene-hopene cyclase N-terminal domain (Pfam 13249). SQHop_cyclase_C - squalene-hopene cyclase C-terminal domain (pfam13243) **(D)** 3D models of baruol synthase proteins encoded by *BARS1* and *BARS2*, predicted by ColabFold software, superposed with the crystal structure of human oxidosqualene cyclase in a complex with lanosterol (LAN). The enlargement box highlights the positions of the catalytic aspartate residue in the predicted models. Colors mark superposed models: green (Col-0 BARS1 isoform NP_193272.1), red (Col-0 BARS1 isoform NP_001329547.1), purple (Cvi-0 BARS1 ATCVI-4G38020), grey (Cvi-0 BARS2 ATCVI-4G38110) and yellow (human OSC PDB ID: 1W6K).

we identified a nonreference gene encoding a protein with ~91% identity to baruol synthase 1 (Supplemental Figure S13). In C24, it was also present but interrupted by ATCOPIA52 retrotransposon insertion, resulting in two shorter ORFs. Based on phylogenetic analysis, we concluded that the identified gene was indeed a *BARS1* duplicate, and we named it *BARS2* (Figure 3B). The differences in the exons of the *BARS1* and *BARS2* sequences matched the heterozygous SNP positions very well (Supplemental Figure S14). Their introns were much more divergent, which likely affected RD genotyping. Likewise, the probe targeting the *BARS1* locus was located in a highly divergent region, which prevented us from detecting this duplication with MLPA.

The proteins encoded by *BARS2* in Cvi-0, Eri-1 and Ler-0 possessed both N-terminal and C-terminal squalene-hopene cyclase domains, typical for OSCs (Figure 3C). We performed three-dimensional (3D) modeling of two reference (Col-0) isoforms of baruol synthase 1 (the product of *BARS1*) and its counterpart from Relict Cvi-0 as well as putative baruol synthase 2 (the product of *BARS2*) from Cvi-0 using ColabFold software. Next, we superposed these models with the experimental crystal structure of human OSC, available in a complex with its reaction product lanosterol (Thoma et al., 2004; Jumper et al., 2021) (Supplemental Information). All structures were highly similar, and we were able to identify potential substrate-binding cavities in the plant enzymes (Figure 3D; Supplemental Table S10). Notably, the catalytic aspartate residue D455 present in the human cyclase had its counterparts in the plant OSCs: D493 in the reference isoform NP_193272.1 and D490 in the remaining proteins (Supplemental Data 1-5). Together, our data indicated that *BARS2* encoded a novel, thus far uncharacterized OSC. As expected, we also found *CYP705A2* duplication in the C24, Cvi-0, Eri-1 and Ler-0 assemblies, and we named it *CYP705A2a*. It had 84% identity with *CYP705A2* at the nucleotide level and 88% similarity at the protein level (Supplemental Figure S15). *CYP705A2a* and *BARS2* were adjacent to each other and located on the minus strand of the large genomic sequence insertion between *CYP702A6* and *BIA* genes (Figure 3A), next to an ~5 kb long interspersed nuclear element 1 (LINE-1) retrotransposon and some shorter, undefined ORFs. The presence of the insertion increased the size of the entire arabidiol/baruol gene cluster by 21-27 kb.

Structural diversity of the arabidiol/baruol gene cluster is associated with the climatic gradient and root growth variation

In the next step, we used the results from the SNP analysis to evaluate the presence/absence variation of both reference (*CYP705A2* + *BARS1*) and nonreference (*CYP705A2a* + *BARS2*) gene pairs in the Arabidopsis population (Supplemental Table S11). The group with only the reference gene pair present was the largest (PP-AA; 628 accessions). Nearly one-third of the population had both gene pairs (PP-PP; 326 accessions). We also separated two smaller groups with the local range of occurrence. The first one, with only the nonreference gene pair, was found in Azerbaijan, Spain, Bulgaria, Russia, Serbia and the U.S.A. (AA-PP; 14 accessions). The last group, where we did not detect any of these genes, was mostly observed at the Bothnian Bay coast collection site in North Sweden (AA-AA; 15 accessions). For 73 accessions, the data were inconclusive. The accuracy of group assignments was validated by sequence analysis of *de novo* genomic assemblies for An-1, Kyoto, Mitterberg-2-185 and Kn-0 (PP-AA group) as well as Cvi-0, Ler-0, Dolna-1-40 and Ty-1 (PP-PP group). Additionally, the results of PCR amplification with gene-specific primers and genomic DNA template for a subset of 36 accessions from all four groups confirmed the differences between them (Supplemental Figure S16). We could not detect *BARS2*-specific products in many samples from the AA-PP group; however, we did detect the band for *CYP705A2a*. We suppose that the *BARS2* sequence might further diverge in this minor group.

The accessions with the nonreference gene pair (AA-PP; PP-PP) dominated among Relicts (81%) and among the Spain (60%) and Italy/Balkan/Caucasus (89.6%) genetic groups but constituted the minority at the northern and eastern margins of the species range (North Sweden 18.6%, South Sweden 16%, Asia 9.4%; Figure 4B). They were also mostly absent among U.S.A. accessions. The widespread presence of *CYP705A2a* and *BARS2* genes in Relicts suggested that the duplication event preceded the recent massive species migration, which took place in the postglacial period and shaped the current Arabidopsis population structure (Lee et al., 2017). We next visualized the four groups in principal component analysis (PCA) plots generated with genome-wide biallelic SNPs (1001 Genomes Consortium, 2016; Zmienko et al., 2020). At a low linkage disequilibrium parameter, where the contribution of the ancestral alleles to PCA was highest, there was a clear convergence of the PC1 and PC2 components with the presence/absence of gene duplication (Figure 4C; Supplemental Figure S17). This suggested that the presence/absence of the genomic insertion containing *CYP705A2a* and *BARS2* genes had some impact on the current geographic distribution of the Arabidopsis accessions. We then evaluated the accessions' latitudes of origin and found that accessions with the nonreference gene pair originated from significantly lower latitudes compared to the remaining accessions (one-way rank-based analysis of variance, ANOVA, p value<0.001, followed by Dunn's test with BH correction, p value<0.001) (Figure 4D). This difference was noticeable even within individual countries and was significant for Germany, Spain and Italy (Supplemental Figure S18). We observed the reverse trend in Russia, where PP-AA accessions were in great excess (88%), and in France; however, we also noticed that PP-AA accessions outnumbered PP-PP accessions in the Pyrenees, Alps and Tian Shan mountain ranges (Supplemental Information). This result suggested that there was an association between arabidiol/baruol gene cluster variation and environmental conditions; therefore, we decided to investigate this in the next step. Since climate is a substantial selection factor, we also checked for phenotypic variability between the most abundant PP-AA and PP-PP groups. To this end, we performed two-group comparisons of 516 phenotypic and climatic variables retrieved from the Arapheno database (Seren et al., 2017; Togninalli et al., 2020) and focused on those that significantly differed between both groups (Wilcoxon rank sum test with continuity correction, p value <0.05) (Supplemental Table S12). Notably, we observed differences in 88 climatic variables (Exposito-Alonso et al., 2019), especially maximal and minimal temperature conditions, precipitation and evapotranspiration (Figure 5A). Apart from the climate data, 40 diverse phenotypes varied significantly

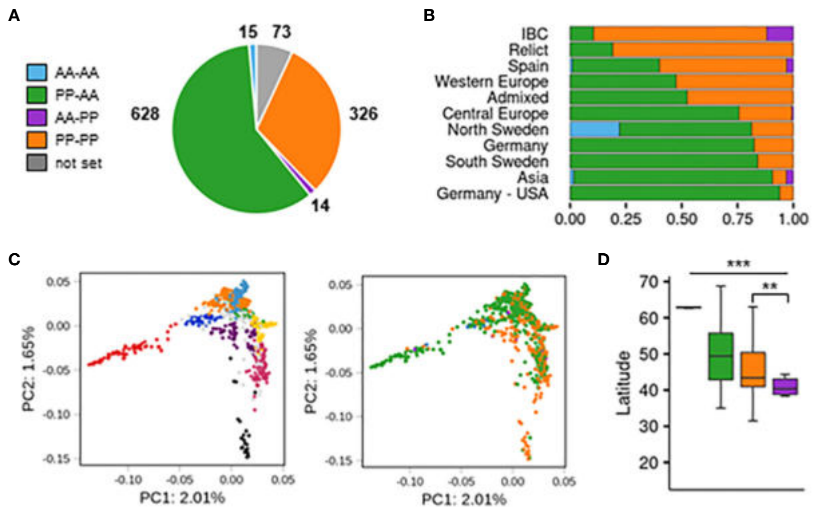


FIGURE 4
Population-scale diversity of *CYP705A2* and *BARS1* duplication status. **(A)** The sizes of four groups differing by the presence (PP)/absence (AA) of *CYP705A2-BARS1* and *CYP705A2-BARS2* gene pairs. **(B)** Group distribution among the genetic groups. U.S. accessions from the German group were separated from the remaining accessions. **(C)** Principal component analysis (PCA) plots, generated at linkage disequilibrium LD = 0.3. The first two components are presented. Accessions are colored according to their genetic group (left) or CYP-BARS status (right). U.S. accessions were not included in the analysis, in order to better visualize the remaining groups. PCA plots with other LD parameters are in Supplemental Figure S17. **(D)** Latitudes of accessions' sites of origin, grouped by CYP-BARS status. One-way rank-based analysis of variance ANOVA, *p* value<0.001, followed by Dunn's test with BH correction, ***p* value<0.05 (PP-PP vs AA-PP); ****p* value < 0.001 (all the other pairwise comparisons).

between both groups. Although some of these differences, e.g., flowering-related phenotypes, might be influenced by another genetic factor, independent from the arabidiol/baruol gene cluster structure (Li et al., 2010), we paid special attention to root growth-related phenotypes, since all Arabidopsis MGCs are considered to have root-specific expression (Huang et al., 2019). We observed significant differences between the PP-AA and PP-PP groups in root growth dynamics, which was analyzed during the first week after germination by Bouain et al. (2018). More specifically, the roots of PP-PP accessions elongated slower than those of PP-AA accessions (Figure 5B). Additionally, PP-PP accessions showed a significantly lower rate of root organogenesis from explants under one of three growth conditions tested in another study (Lardon et al., 2020) (Figure 5C). We next applied a linear mixed model in a genome-wide association study on the same 516 phenotypes to independently evaluate the significance of our observations after correction for the population structure and multiple testing. We used a genome-wide matrix of over 250 thousand biallelic SNPs supplemented with SNP-like encoded information about the gene duplication status (only PP-AA and PP-PP groups were analyzed). Although the association of *CYP705A2a* and *BARS2* presence/absence variation was not statistically significant for any variable we tested, we again obtained the lowest *p* values for the climatic data and root organogenesis phenotypes (Figure 5D, Supplemental Table S12). We then checked for the genetic interactions between the thalianol and arabidiol/baruol clusters to exclude the possibility that they affected our results, since the distribution of discontiguous and compact versions of the thalianol gene cluster was also strongly associated with latitude (Supplemental Figure S19). However, structural variation of the arabidiol/baruol gene cluster better explained the geographical distribution of the accessions. Moreover, variation in thalianol gene cluster organization did not affect the expression of the thalianol biosynthesis genes and had little impact on root growth phenotypic variation (Supplemental Figure S20).

In the reference accession Col-0, all genes in the arabidiol/baruol cluster were expressed at low levels and were active almost exclusively in roots (Supplemental Figure S21). In search of the possible links between arabidiol/baruol gene cluster structure and phenotypic variation, we investigated *CYP705A2*, *BARS1*, *CYP705A2a* and *BARS2* expression profiles in Col-0 and Cvi-0. We used RNA-Seq data from roots, shoots and leaves, which we retrieved from the studies where these accessions were grown in parallel under standard conditions (Kawakatsu et al., 2016; van Veen et al., 2016). We mapped the data to the respective (Col-0 or Cvi-0) annotated genome and compared the gene expression profiles (Figure 5E; Supplemental Table S13). In both accessions, the arabidiol/baruol gene cluster was silenced in shoots, except for the low activity of acyltransferase gene *AT4G15390*, detected in Cvi-0. Additionally, in both accessions, the clusters were active in roots, and the expression of *AT4G15390* was much stronger than that of the remaining genes. In Cvi-0, genes located in the genomic insertion (*CYP705A2a*, *BARS2* and *ATCVI-4G38100*, the latter encoding a protein with partial similarity to acyltransferase) were also expressed, although at a lower level, compared to the rest of the cluster. Surprisingly, in leaves of Cvi-0, but not Col-0, we also detected transcriptional activity within the arabidiol/baruol gene cluster. Most clustered genes were expressed at lower levels than in Cvi-0 roots, and the transcripts of *CYP705A2*, *CYP705A3* and *BARS1* were barely detectable. However, *ATCVI-4G38100*, *CYP705A2a* and *BARS2* had similar expression in leaves and roots. Taking these observations into account, it should not be

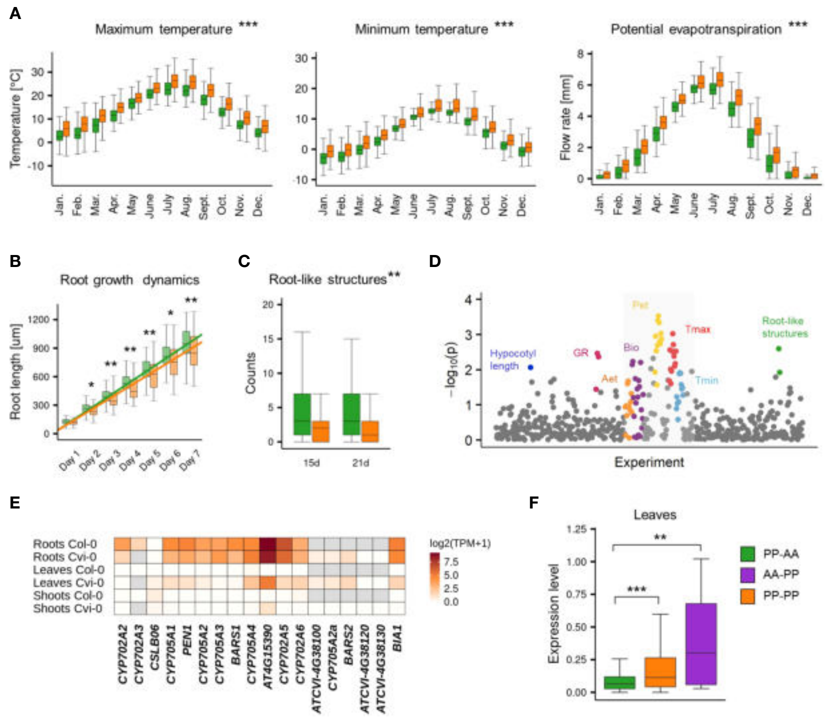


FIGURE 5
Phenotypic variation of PP-AA and PP-PP groups. **(A-C)** Two-group comparisons of climatic **(A)**, root growth dynamics **(B)** and root organogenesis **(C)** data between PP-AA (green) and PP-PP (orange) accessions. Stars denote the significance of Wilcoxon rank sum test with continuity correction, *p.value<0.1, **p.value<0.05, ***p.value<0.001. **(D)** Results of a genome-wide association study for PP-AA/PP-PP allelic variation. Study with climatic data is in the grey box **(E)** Tissue specificity of arabidiol/baruol gene cluster expression in Col-0 and Cvi-0. **(F)** Population-level differences in gene expression in leaves among the PP-AA, PP-PP and AA-PP groups. Expression levels are shown as $\log_2$(TPM+1). Stars denote the significance of one-way rank-based analysis of variance ANOVA, p.value<0.001, followed by Dunn's test with BH correction, **p.value<0.05, ***p.value<0.001. Source data are available in the Arapheno database (plots **A-C**), Supplemental Table S12 (plot **D**) and Supplemental Table S13 (plots **E-F**).

excluded that the metabolic products of arabidiol/baruol gene cluster activity in the roots and leaves of the Cvi-0 accession are not identical.

Since the PP-PP group represented a substantial fraction of the Arabidopsis population, we wanted to check whether the gene expression profile, which we observed in leaves of Cvi-0, was ubiquitous among the accessions from this group. To this end, we analyzed RNA-Seq data for 552 accessions mapped against the reference genome (Kawakatsu et al., 2016), and we compared the *BARS1* expression level between the AA-PP, PP-PP and PP-AA groups. It was significantly higher in accessions with the *CYP705A2a* + *BARS2* gene pair than in the PP-AA group (one-way rank-based analysis of variance, ANOVA, *p* value<0.001, followed by Dunn's test with BH correction, *p* value <0.05) (Figure 5F), in agreement with our predictions that *BARS2* was expressed in the leaves of these accessions and that reads from *BARS2* transcripts mapped to the *BARS1* locus, elevating its measured expression level. We also remapped the raw RNA-Seq reads from the Ty-1 and Cdm-1 accessions (PP-PP group), as well as from the Kn-0 and Sha (PP-AA group) accessions to their respective genomic assemblies and separately measured the expression levels of *BARS1* and *BARS2.* As expected, *BARS2* was expressed in the leaves of PP-PP accessions, while *BARS1* was not (Supplemental Figure S21).

Paired terpenoid synthase and cytochrome P450 genes are more variable than nonpaired genes

In many plant genomes, genes encoding terpenoid synthases (TSs, including the OSCs analyzed in our study) are positioned in the vicinity of CYPs more often than expected by chance. Therefore, they frequently exist as TS-CYP pairs (Boutanaev et al., 2015). TS-CYP pairs located in MGCs had similar (either high or low) copy number diversity and were frequently duplicated or deleted together. We wanted to check whether this observation could be extended to other TS-CYP pairs in the Arabidopsis genome. Therefore, we created a comprehensive list of 48 TSs and 242 CYPs based on trusted sources (Paquette et al., 2000; Bak et al., 2011; Nelson and Werck-Reichhart, 2011; Boutanaev et al., 2015). We then retrieved information about each gene's copy number diversity among 1,056 accessions (Supplemental Tables S14-S15). For 13 TSs, including *THAS1* and *BARS1*, we observed gains or losses in at least 1% of accessions. Only 33 CYPs showed such variability, and they represented three clans: CYP71 (26 variable genes out of 151), CYP85 (6 variable genes out of 29) and CYP72 (1 variable gene out of 19). The remaining clans showed very low variability. Next, for each TS, we selected all CYPs within +/- 30-kb distance, which produced 38 pairs between 18 TSs and 27 CYPs, including pairs in thalianol, marneral, tirucalladienol and arabidiol/baruol gene clusters, as well as other putative secondary metabolism clusters, listed in the plantiSMASH resource (Kautsar et al., 2017). Subsequent group comparisons revealed that TSs and CYPs occurring in pairs were more variable than their nonpaired counterparts (Wilcoxon rank sum test with continuity correction, *p* value<0.01 for TSs, *p* value<0.001 for CYPs).

Discussion

According to our current understanding of the MGC formation phenomenon, nonrandom gene clustering in eukaryotes is linked with highly dynamic chromosomal regions. Numerous studies have highlighted that structural variations are the main genetic drivers of metabolic profile diversity and MGC evolution in plants (Fan et al., 2020; Li et al., 2020; Liu et al., 2020a; Liu et al., 2020b; Zhan et al., 2020; Katz et al., 2021). These studies suggested that plant MGCs are dynamically evolving and that the genetic mechanisms that originally led to their formation may be captured at the intraspecies genetic variation level. Similar conclusions were drawn from a previous study of the filamentous fungus *Aspergillus fumigatus*, in which secondary metabolic pathway genes were commonly organized into clusters (Lind et al., 2017). During evolution, new biochemical pathways are tuned and tested by many rounds of natural selection. The analysis of intraspecies MGC variants, which are more recent than the variants found in interspecies comparisons, may provide important insight into the formation of clustered gene architectures and plant metabolic diversity in a small evolutionary time frame. Accordingly, in our study we established that the mechanisms driving gene duplications and deletions contributed to the formation of Arabidopsis MGC in their present form and that they are still involved in shaping their structures. The dynamics of these mechanisms is e.g. marked by the observed extensive variation of the thalianol gene cluster and the arabidiol/baruol gene cluster.

The four MGCs in Arabidopsis are implicated in the biosynthesis of structurally diverse triterpenes and are dated after the α whole-genome duplication event, which occurred in the Brassicaceae lineage ~23-43 Mya (Field et al., 2011). These MGCs are assembled around the gene(s) encoding clade II OSCs. It has been shown that in various Brassicaceae genomes, clade II OSCs are often colocalized with genes from the CYP705, CYP708 and CYP702 clans and with genes from the acyltransferase IIIa subfamily (Liu et al., 2020b). Bioinformatic studies have also revealed that TSs and CYPs are paired in plant genomes more frequently than expected (Boutanaev et al, 2015). We found that in Arabidopsis, the physical proximity of CYPs and TSs was associated with increased CNV rates for these genes compared to the nonpaired ones. This might suggest that the occurrence of such a specific gene mix, combined with the structural instability of its genomic neighborhood, boosted the potential to produce novel metabolic pathways. The four Arabidopsis MGCs had different levels of variation (Figure 1), which generally reflected the phylogeny of clade II OSCs contained in these clusters (Figure 3C). Of them, MRN1 is most divergent in amino acid sequence. It is also mono-functional, i.e., catalyzes the formation of one specific product – marneral (Xiong et al., 2006). Functional studies have indicated a critical role of marneral synthase in Arabidopsis development (Go et al., 2012). Consistent with these findings, *MRN1* was the only clustered OSC gene, which was not affected by deletions or duplications, in any accession. Additionally, the neighboring CYPs were stable in copy number. Our results indicate that the marneral gene cluster is fixed in the Arabidopsis genome.

The arabidiol/baruol gene cluster was the most variable MGC. It comprises few gene subfamilies but is significantly expanded compared to the sister species *A. lyrata*, which is suggestive of recent duplications. For example, *PEN1* and *BARS1* have only one ortholog in *A. lyrata*, *LOC9306317*. Accordingly, we observed an exceptionally high rate of intraspecific gene gains and losses within this MGC. The segmentation of the arabidiol/baruol gene cluster into variable and invariable gene blocks may result from the ongoing process of selection-driven fixation of the arabidiol subcluster. The products of *PEN1* and *CYP705A1* are involved in the response to jasmonic acid treatment and infection with the root-rot pathogen *Pythium irregulare* (Sohrabi et al., 2015). Moreover, arabidiol may be further converted to arabidin in the pathway involving acyltransferase encoded by *AT5G47950*, which is located in the thalianol gene cluster (Huang et al., 2019) and which was also invariable in copy number in the present study. The fixation of genes involved in arabidin biosynthesis may indicate the biological significance of this pathway. CRISPR mutants with a disrupted *AT5G47950* gene has been shown to have significantly shorter roots than wild-type plants, and arabidin did not accumulate in these roots (Bai et al., 2021). Interestingly, *A. lyrata* is able to convert apo-arabidiol (the product of arabidiol degradation) into downstream compounds, despite the lack of arabidiol synthase (Sohrabi et al., 2017). This indicates that there may be modularity of the biosynthetic pathways in plants. This modularity might facilitate the assembly of a biosynthesis network and lead to an increase in the repertoire of secondary metabolites produced by the plant. Understanding the complexity of this network may be supported by in-depth analysis of MGC intraspecies variation.

The initial diversity of 2,3-oxidosqualene cyclization products generated by the plant is determined by OSC diversity. Here, we report the discovery of the *BARS2* gene, which was found in numerous accessions but was absent from Col-0; hence, it was absent from the reference genome (Figure 3A). Our data indicated that *BARS2* encodes a functional clade II OSC (Figures 3C, D). Notably, baruol synthase 1 encoded by its closest paralog, *BARS1*, has the lowest product specificity among plant OSCs (Lodeiro et al., 2007; Ghosh, 2016). Why some OSCs are highly multifunctional is not well understood. It has been suggested that they are undergoing evolution toward increased product specificity. It has been demonstrated that only two amino acid changes in cycloartenol synthase lead to its conversion into an accurate lanosterol synthase (Lodeiro et al., 2005). Biochemical characterization of baruol synthase 2 and its comparison with baruol synthase 1 may help reveal the role of particular amino acids in acquiring specificity for given products.

According to our data, the *BARS2* and *CYP705A2a* gene pair may be present in nearly one-third of the Arabidopsis population (Figure 4A), and their presence/absence variation is associated with the climatic gradient and root growth dynamics (Figures 5A-D). In Col-0, MGCs are embedded in local hotspots of three-dimensional chromatin interactions. Their activation in roots and repression in leaves is combined with the distinct chromatin condensation states and nuclear repositioning of MGC regions between these tissues (Nützmann et al., 2020). Loss of the histone mark H3K27me3 in the *clf/swn* mutant resulted in the loss of interactive domains associated with the thalianol, marneral and arabidiol/baruol cluster regions, indicating that different transcriptional states of these MGCs are strictly regulated by the switches in their conformation. Curiously, in accessions with *CYP705A2a* and *BARS2*, we observed some transcriptional activity of arabidiol/baruol cluster genes in leaves (Figures 5E, F). The presence of an ~25-kb insertion in the arabidiol/baruol gene cluster may alter its structure and affect the epigenetic regulation of its activity. Thus, variation at the epigenetic and transcriptional level might lead to phenotypic differences, which could in turn contribute to local adaptation and eventually affect the global distribution of Arabidopsis accessions. However, additional studies are needed to assess whether the association between *BARS2* and *CYP705A2a* presence/absence variation and the global distribution of Arabidopsis accessions may be linked to the expression of these two genes or to the differences in transcriptional activity of the entire cluster (Wegel et al., 2009; Yu et al., 2016; Roulé et al., 2022).

The thalianol gene cluster was the second most variable MGC in our analysis. The first evidence for its structural diversity comes from the study of Liu et al. (2020a), who found large deletions affecting thalianol biosynthesis genes in ~2% of the studied accessions. Since our approach was specifically focused on CNV analysis and was duplication-aware, we were able to detect over two times more CNVs in a similar population (4.7%), with 49 accessions carrying gene deletions and five accessions with gene duplications (Figure 2A). Apart from the identification of two new variants – one large deletion and a duplication – we validated earlier assumptions that the nonreference compact version of the thalianol gene cluster is predominant in Arabidopsis (Figure 2B). Moreover, it is also better conserved than the discontiguous version (Figures 2D, E). It remains to be investigated whether tighter clustering of the thalianol gene cluster may be advantageous in certain environmental conditions or whether it is just less prone to structural variation due to physical constraints.

Triterpenes are high-molecular-weight nonvolatile compounds that are likely to act locally. However, they may be further processed and generate various breakdown products, both volatile and nonvolatile, which may be biologically active (Sohrabi et al., 2015; Sohrabi et al., 2017). Compounds of plant origin may also be metabolized by plant-associated microbiota. A recent study demonstrated that various combinations of thalianin, thalianyl fatty acid esters and arabidin attracted or repelled various microbial communities present in the soil and participated in the plant's active selection of root microbiota (Huang et al., 2019). In fact, a small but significant effect of Arabidopsis genotype on the root microbiome has been demonstrated previously (Bulgarelli et al., 2012; Lundberg et al., 2012). In a recent study by Karasov et al. (2022), bacterial communities that colonized the leaves of 267 local Arabidopsis populations, assessed at various localizations in Europe, formed two distinct groups strongly associated with the latititude. Specifically, a significant latitudinal cline was observed for the strains of the *Sphingomonas* genus, which is commonly associated with Arabidopsis (Bodenhausen et al., 2013). Various *Sphingomonas* species possess a range of biodegradative and biosynthetic capabilities (Mohn et al., 1999; Asaf et al., 2020). *Sphingomonas* is implicated in promoting Arabidopsis growth, increasing drought resistance and protecting plants against the leaf-pathogenic *Pseudomonas syringae* (Innerebner et al., 2011; Luo et al., 2019). Notably, in the study by Karasov et al. (2022), the host plant genotype alone could explain 52% to 68% of the observed variance in the phyllosphere microbiota. Moreover, the microbiome type was strongly associated with the dryness index of the local environment based on recent precipitation and temperature data. We propose that the genetic diversity of terpenoid metabolism pathways in Arabidopsis may be interdependent on the diversity of soil bacterial communities present in various environments, and this relationship might play a role in Arabidopsis adaptation to climate-driven selective pressures. Further exploration of MGC diversity may help us understand these biotic interactions.

Currently, the bioinformatic identification of new MGC candidates is mainly based on the combination of physical gene grouping and coexpression analyses. The accuracy and sensitivity of such approaches strongly depend on the abundance of data from various tissues, time points, and environmental conditions (Wisecaver et al., 2017). We suggest that the analysis of intraspecies genetic and transcriptomic variation may provide a valuable addition to MGC studies. The genome of one individual may not be representative enough to reveal the entire complexity of a given pathway, not to mention the metabolic diversity of the entire species (Kawakatsu et al., 2016; Shirai et al., 2017; Zmienko et al., 2020; Katz et al., 2021). With the rapid increase in the number of near-to-complete assemblies of individuals' genomes facilitated by the development of third-generation sequencing technologies, we are now entering the era of intense exploration of the impressive plasticity of plant metabolic pathways.

Materials and methods

Plant material and DNA samples

Arabidopsis seeds were obtained from The Nottingham Arabidopsis Stock Centre. The seeds were surface-sterilized, vernalized for 3 days, and grown on Jiffy pellets in ARASYSTEM containers (BETATECH) in a growth chamber (Percival Scientific). A light intensity of 175 mmol m^{-2} s^{-1} with proportional blue, red, and the far red light was provided by a combination of fluorescent lamps (Philips) and GroLEDs red/far red LED Strips (CLF PlantClimatics). Plants were grown for 3 weeks under a 16-h light (22°C)/8-h dark (18°C) cycle, at 70% RH, nourished with half-strength Murashige & Skoog medium (Serva). Genomic DNA for MLPA and ddPCR assays was extracted from 100 mg leaves with a DNeasy Plant Mini Kit (Qiagen), according to manufacturer's protocol, which included RNase A treatment step.

RD assays

To determine the boundaries of each MGC, the relevant literature and gene coexpression datasets were surveyed (Fazio et al., 2004; Xiong et al., 2006; Xiang et al., 2006; Lodeiro et al., 2007; Field and Osbourn, 2008; Morlacchi et al., 2009; Field et al., 2011; Go et al., 2012; Thimmappa et al., 2014; Sohrabi et al., 2015; Yasumoto et al., 2016; Wisecaver et al., 2017). TAIR10 genome version and Araport 11 annotations (Cheng et al., 2017) were used as a reference in all analyses. Short read sequencing data from Arabidopsis 1001 Genomes Project (1001 Genomes Consortium, 2016) were downloaded from National Center for Biotechnology Information Sequence Read Archive repository (PRJNA273563), processed and mapped to the reference genome as described in (Zmienko et al., 2020). The gene copy number estimates based on read-depth analysis of short reads (RD dataset) were generated previously and are available at http://athcnv.ibch.poznan.pl. Accessions BRR57 (ID 504), KBS-Mac-68 (ID 1739), KBS-Mac-74 (ID 1741) and Ull2-5 (ID 6974), which we previously identified as harboring unusually high level of duplications, were removed from the analysis.

MLPA assays

MLPA probes were designed according to a procedure designed previously and presented in detail in (Samelak-Czajka et al., 2017). Probe genomic target coordinates are listed in Supplemental Table S16. The MLPA assays were performed using 5 ng of DNA template with the SALSA MLPA reagent kit FAM (MRC-Holland). The MLPA products were separated by capillary electrophoresis in an ABI Prism 3130XL analyzer at the Molecular Biology Techniques Facility in the Department of Biology at Adam Mickiewicz University, Poznan, Poland. Raw electropherograms were quality-checked and quantified with GeneMarker v.2.4.2 (SoftGenetics), with peak intensity and internal control probe normalization options enabled. Data were further processed in Excel (Microsoft). To allow easy comparison of RD and MLPA values, the MLPA results were normalized to a median of all samples' intensities and then multiplied by 2, separately for each gene/MLPA probe.

ddPCR assays

Genomic DNA samples were digested with XbaI (Promega). DNA template (2.5 ng) was mixed with 1× EvaGreen ddPCR Supermix (Bio-Rad), 200 nM gene-specific primers (Supplemental Table S17) and 70 µl of Droplet Generation Oil (Bio-Rad), then partitioned into approximately 18,000 droplets in a QX200 Droplet Generator (Bio-Rad), and amplified in a C1000 Touch Thermal Cycler (Bio-Rad), with the following cycling conditions: 1× (95°C for 5 min), 40× (95°C for 30 s, 57°C for 30 s, 72°C for 45 s), 1× (4°C for 5 min, 90°C for 5 min), with 2°C/s ramp rate. Immediately following end-point amplification, the fluorescence intensity of the individual droplets was measured using the QX200 Droplet Reader (Bio-Rad). Positive and negative droplet populations were automatically detected by QuantaSoft droplet reader software (Bio-Rad). For each accession and each gene, the template CNs [copies/µl PCR] were calculated using Poisson statistics, background-corrected based on the no-template control sample and normalized against the data for previously verified non-variable control gene *DCL1*.

PCR assays

Genomic DNA samples (5 ng) were used as templates in 20 µl reactions performed with PrimeSTAR GXL DNA Polymerase (TaKaRa), according to the manufacturer's instructions, in a three-step PCR. Amplicons (10 ul) were analyzed on 1% agarose with 1kb Gene Ruler DNA ladder (Fermentas). Primer sequences are listed in Supplemental Table S17. Primer pairs for *BARS1-BARS2* and *CYP705A2-CYP705A2a* were designed in corresponding genomic regions, that assured primer divergence between the paralogs. However, primers designed for *CYP705A2* produced unspecific bands of ~5kb in many samples. Therefore, this gene was excluded from the analysis.

Genotype assignments

For MLPA dataset, genotypes were assigned to each gene and each accession based on normalized MLPA values of ≤1 for LOSS genotype and >3 for GAIN genotype. The remaining cases were assigned REF genotype. For RD dataset, the respective RD thresholds were ≤1 for LOSS genotype and >3.4 for GAIN genotype, except for *BARS1*, for which both thresholds were lowered by 0.2. The remaining cases were assigned REF genotype. For ddPCR, genes with normalized CN=0 were assigned LOSS genotype and genes with normalized CN=2 were assigned REF genotype. The RD, MLPA and ddPCR datasets were then combined using the following procedure. For genes and accessions covered by multiple datasets, the final genotype was assigned based on all data. Discordant genotype assignments (21 out of 1,784 covered by multiple datasets) were manually investigated and 19 of them were resolved (Supplemental Figure S4; Supplemental Table S7). Out of the remaining 32,000, which were assayed with one method only, the genotype was manually corrected in 13 cases with values very close to the arbitrary threshold, based on population data distribution. Final genotype assignments for each gene and each accession are listed in Supplemental Table S6.

Sanger sequencing

The genomic DNA of Mir-0 accession (ID 8337) was used as a template (2 ng) for amplification using PrimeSTAR® GXL DNA Polymerase (TaKaRa), in a 40-µl PCR reaction with 0.3 µM primers OP009 and OP010, according to general manufacturer instructions. The amplified product, of ~8 kb in length, was purified with DNA Clean & Concentrator (ZYMO Research) and checked by gel electrophoresis and analysis on NanoDrop™ 2000 Spectrophotometer. The purified product (110 ng) was mixed with 1 ul of sequencing primer Mar02_R and sequenced on ABI Prism 3130XL analyzer at the Molecular Biology Techniques Facility in the Department of Biology at Adam Mickiewicz University, Poznan, Poland. Sequencing files were analyzed with Chromas Lite v. 2.6.6. (Technelysium) software.

De novo genomic assemblies generation, annotation and analysis

Mitterberg-2-185 and Dolna-1-40 genomic sequences were extracted, sequenced on 1 MinION flowcell (*Oxford Nanopore Technologies*) each and assembled *de novo* with Canu. Genomic sequences of interest (corresponding to thalianol gene cluster for Mitterberg-2-185 and tirucalladienol gene cluster for Dolna-1-40) were then retrieved with megablast (blast-2.10.0+ package) using TAIR10 reference genomic sequence as a query. The remaining *de novo* assemblies were retrieved from the following public databases. The PacBio-based genomic assemblies, gene annotations and orthogroups for An-1, C24, Cvi-0, Eri-1, Kyoto, Ler-0 and Sha accessions, as well as the reference genome coordinates of the hotspots of rearrangements, were downloaded from Arabidopsis 1001 Genomes Project Data Center (MPIPZJiao2020) or retrieved from the corresponding paper (Jiao and Schneeberger, 2020). Assembled genomic sequences of Ty-1 (PRJEB37258), Cdm-0 (PRJEB40125) and Kn-0 (PRJEB37260) accessions were retrieved from NCBI/Assembly database (Sayers et al., 2022). Gene prediction was performed with Augustus v.3.3.3 (Stanke and Morgenstern, 2005) with the following settings: "Species *Arabidopsis thaliana*", "both strands", "few alternative transcripts" or "none alternative transcripts", "predict only complete genes". These parameters were first optimized by gene prediction in the corresponding TAIR 10 genomic sequence and comparison with Araport 11 annotation. For previously annotated assemblies, we added information about the newly predicted genes to existing annotations. The protein sequences of *de novo* predicted genes and the information about their best blast hit in the reference genome are available in Supplemental Information. The search for conserved domain organization was performed with the online NCBI search tool against Pfam v.33.1 databases. Protein sequence alignment was done with Multalin or EMBL online tools (Corpet, 1988; Madeira et al., 2019). TEs were annotated with RepeatMasker software version 4.1.2 (http://www.repeatmasker.org), using homology-based method with TAIR10-transposable-elements reference library.

Identification of chromosomal inversions

The BreakDancerMax program from the BreakDancer package v.1.3.6 (Chen et al., 2009) was used to detect inversions in each of 997 samples with paired-end data and unimodal insert size distribution. Variants were called separately for each accession and each chromosome. Only calls with lengths within the range 0.5 kbp – 50 kbp and with the Confidence Score >35 were retained. Since BreakDancerMax output included numerous overlapping calls for individual accessions, we first minimized its redundancy. From the overlapping regions, we kept one variant with i) the highest Confidence Score, and ii) the highest number of supporting reads. If two or more overlapping variants had the same score and the number of supporting reads number, maximized coordinates of these variants were used. This step was carried out in two iterations, considering the 50% reciprocal overlap of the variants. Then, the inversions that overlapped with the thalianol gene cluster were selected from each genome-wide dataset.

SNP calling at *CYP705A2* and *BARS1* genes

Variants (SNPs and short indels) were called with DeepVariant v.1.3.0 in WGS mode and merged with GLnexus (Yun et al., 2021). Analysis was performed for *CYP705A2* and *BARS1* genomic loci. The results were further filtered to include only biallelic variants, that were located in the exons of each gene (for *BARS1*, exon intersections from two transcript models were used). The number of heterozygous positions was then calculated for each accession and each gene. The same procedure was repeated by taking into account only biallelic variants with at least 1% frequency, which resulted in nearly identical results. Both types of analysis led to the selection of the same set of accessions with duplication at both loci.

Genome-wide SNP analysis

Variants for 983 accessions with known *CYP705A2* + *BARS1* and *CYP705A2a* + *BARS2* pair status were downloaded from the 1001 Genomes Project Data Center (1001genomes_snp-short-indel_only_ACGTN_v3.1.vcf.snpeff file) (1001 Genomes Consortium, 2016). Data preprocessing was performed using PLINK v.1.90b3w (https://www.cog-genomics.org/plink/1.9/; Chang et al., 2015). Variants with missing call rates exceeding value 0.5 and variants with minor allele frequency below 3% were filtered out. The LD parameter for linkage disequilibrium-based filtration was set as follows: indep-pairwise 200'kb' 25 0.3. For PCA analysis with EIGENSOFT v.7.2.1 (Price et al., 2006; Patterson et al., 2006) at least 130,000 SNPs were used. PCA for a wide LD range between 0.3 - 0.9 was then calculated in a similar manner. U.S.A accessions which only recently separated geographically from the rest of the population (Lee et al., 2017) were excluded, to ensure better visibility of the remaining accessions. The ggplot2 package was used for data visualization in R v4.0.4 (https://www.r-project.org; Wickham, 2016).

Genome-wide association study and phenotype analysis

The entire set of 516 phenotypes from 26 studies was downloaded from the Arapheno database on 26 April 2022 (Seren et al., 2017; Togninalli et al., 2020). The above genome-wide SNP dataset, to which we added a biallelic variant representing PP-AA or PP-PP group assignment, was used. The IBS kinship matrix was calculated on 954 accessions. Association analysis was performed for each phenotype using a mixed model correcting for population structure using Efficient Mixed-Model Association eXpedited, version emmax-beta-07Mar2010 (Kang et al., 2008). Input file generation and analysis of the results were performed with PLINK v.1.90b3w and R v4.0.4.

Analysis of RNA-Seq data

Processed RNA-seq data from leaves for 728 accessions (552 in common with our study) mapped to the reference transcriptome (Kawakatsu et al., 2016) were downloaded from NCBI/SRA (PRJNA319904), normalized and used to compare *BARS1* expression levels between PP-AA, PP-PP and AA-PP groups. Additionally, raw RNA-Seq reads from leaves were downloaded from the same source for accessions-specific mapping and analysis of Cdm-0, Col-0, Cvi-0, Kn-0, Ty-1 and Sha accessions. Raw RNA-Seq reads from roots and shoots of Col-0 and Cvi-0 accessions were retrieved from BioProject PRJEB14092 (van Veen et al., 2016). SRA Toolkit v2.8.2. (https://github.com/ncbi/sra-tools) and FastQC v0.11.4 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) were used for downloading the raw reads and for the quality analysis. For Cdm-0, Kn-0 and Ty-1 genomes.gtf files were generated based on Augustus results, that included the annotations for the genes of interest (provided as Supplemental Information). Raw reads were mapped to the respective genomes using the STAR aligner version 2.7.8a (Dobin et al., 2013). STAR indices were generated with parameters: "–runThreadN 24 –sjdbOverhang 99 –genomeSAindexNbases 12". The following parameters were used for the mapping step: "–runThreadN 24 –quantMode GeneCounts –outFilterMultimapNmax 1 –outSAMtype BAM SortedByCoordinate –outSAMunmapped Within". Bioinfokit v1.0.8 https://zenodo.org/record/3964972#.Yyw6oRzP1hE) was used to convert.gff3 to.gtf files. Transcripts per million (TPM) values and fragments per kilobase exon per million reads (FPKM) with total exon length for each gene were computed in R v4.0.4.

Analysis of TS-CYP pairs

A list of Arabidopsis CYP genes was created by collecting information from previous studies and acknowledged website resources (Arabidopsis Cytochromes P450; Paquette et al., 2000; Ehlting et al., 2008; Nelson, 2009; Bak et al., 2011; Nelson and Werck-Reichhart, 2011; Boutanaev et al., 2015) (http://www.p450.kvl.dk/p450.shtml). Genes marked in Araport 11 as pseudogenes were excluded from the further analysis. Genes were assigned to clans and families according to the information from the above resources. A list of TS genes was created based on a previous study (Boutanaev et al., 2015) and restricted to genes with valid Araport 11 locus. Genotypes were assigned based on criteria defined for RD dataset: (CN =< 1 as losses, CN >=3.4 as gains, the remaining genotypes were classified as unchanged). Genes from thalianol, tirucalladienol, arabidiol/baruol and marneral gene clusters were already genotyped. Gene coordinates were downloaded from Araport 11. All CYP genes positioned at a distance +/- 30 kb from TS gene borders were classified as paired with a given TS gene. Information about predicted secondary metabolism clusters was retrieved from plantiSMASH resource (Kautsar et al., 2017).

Prediction and analysis of BARS1 and BARS2 3D protein structures

The three-dimensional structures of the reference baruol synthase 1 proteins NP_193272.1, NP_001329547.1, as well as Cvi-0 proteins encoded by *ATCVI-4G38020* (*BARS1*) and *ATCVI-4G38110* (*BARS2*), were predicted from their amino acid sequences using the AlphaFold2 code through the ColabFold software (Jumper et al., 2021; Mirdita et al., 2022). The modeling studies were performed for a single amino acid chain. A crystal structure of human OSC in a complex with lanosterol (ID 1W6K) was retrieved from the Protein Data Bank (Thoma et al., 2004; Berman et al., 2007). The SSM algorithm implemented in COOT was used for superpositions of protein models (Krissinel and Henrick, 2004; Emsley et al., 2010) (Supplemental Information).

Data availability statement

Publicly available datasets were analyzed in this study. Sequence data can be found at the National Center for Biotechnology Information (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA273563/, https://www.ncbi.nlm.nih.gov/bioproject/PRJEB31147/; https://www.ncbi.nlm.nih.gov/bioproject/PRJEB37258/; https://www.ncbi.nlm.nih.gov/bioproject/PRJEB40125/; https://www.ncbi.nlm.nih.gov/bioproject/PRJEB37260/; https://www.ncbi.nlm.nih.gov/bioproject/PRJNA319904/; and https://www.ncbi.nlm.nih.gov/bioproject/PRJEB14092/). Genomic variants can be found in the 1,001 Genomes Project resources (https://1001genomes.org/data/GMI-MPI/releases/v3.1/1001genomes_snpeff_v3.1/). All phenotyping data and the associated metadata can be found in the AraPheno database (https://arapheno.1001genomes.org/static/database.zip). Individual phenotypes with their DOI identifiers can be additionally accessed and downloaded from https://arapheno.1001genomes.org/phenotypes/. The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Author contributions

Conceptualization: AZ. Methodology: MM-Z, PW, and AZ. Investigation: MM-Z, AS, PW, PS, KB, and TI. Software: MM-Z, AS, PW, and MZ. Visualization: MM-Z, KB, and AZ. Formal analysis: MM-Z. Writing – original draft: MM-Z, and AZ. Writing – review and editing: MM-Z, KB, MF, MZ, and AZ. Supervision: MF, and AZ. Project administration: AZ. Funding acquisition: MF, and AZ. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the National Science Centre (Poland) grants 2014/13/B/NZ2/03837 to MF and 2017/26/D/NZ2/01079 to AZ. TI obtained funding from the support program for Ukrainian researchers under the Agreement between the Polish Academy of Sciences and the U.S. National Academy of Sciences. The funding agencies had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Acknowledgments

We thank Piotr Kozłowski for fruitful discussions and comments on the manuscript. Computations were supported in part by PLGrid Infrastructure.

Conflict of interest

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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1104303/full#supplementary-material

SUPPLEMENTAL FILE 1
Supplemental Tables S1-S17.

SUPPLEMENTAL FILE 2
Supplemental information and Supplemental Figures S1-S21.

SUPPLEMENTARY DATA SHEET 15
Superposed 3D models of BARS1, BARS2 and human oxidosqualene cyclase proteins.

References

1001 Genomes Consortium (2016). 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell* 166, 1–11. doi: 10.1016/j.cell.2016.05.063

Asaf, S., Numan, M., Khan, A. L., and Al-Harrasi, A. (2020). *Sphingomonas*: from diversity and genomics to functional role in environmental remediation and plant growth. *Crit. Rev. Biotechnol.* 40, 138–152. doi: 10.1080/07388551.2019.1709793

Bai, Y., Fernández-Calvo, P., Ritter, A., Huang, A. C., Morales-Herrera, S., Bicalho, K. U., et al. (2021). Modulation of arabidopsis root growth by specialized triterpenes. *New Phytol.* 230, 228–243. doi: 10.1111/nph.17144

Bak, S., Beisson, F., Bishop, G., Hamberger, B., Höfer, R., Paquette, S., et al. (2011). Cytochromes p450. *Arabiopsis Book* 9, e0144. doi: 10.1199/tab.0144

Berman, H., Henrick, K., Nakamura, H., and Markley, J. L. (2007). The worldwide protein data bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res.* 35, D301–D303. doi: 10.1093/nar/gkl971

Bodenhausen, N., Horton, M. W., and Bergelson, J. (2013). Bacterial communities associated with the leaves and the roots of *Arabidopsis thaliana*. *PloS One* 8, e56329. doi: 10.1371/journal.pone.0056329

Bouain, N., Satbhai, S. B., Korte, A., Saenchai, C., Desbrosses, G., Berthomieu, P., et al. (2018). Natural allelic variation of the AZI1 gene controls root growth under zinc-limiting condition. *PloS Genet.* 14, e1007304. doi: 10.1371/journal.pgen.1007304

Boutanaev, A. M., Moses, T., Zi, J., Nelson, D. R., Mugford, S. T., Peters, R. J., et al. (2015). Investigation of terpene diversification across multiple sequenced plant genomes. *Proc. Natl. Acad. Sci.* 112, E81–E88. doi: 10.1073/pnas.1419547112

Boycheva, S., Daviet, L., Wolfender, J. L., and Fitzpatrick, T. B. (2014). The rise of operon-like gene clusters in plants. *Trends Plant Sci.* 19, 447–459. doi: 10.1016/j.tplants.2014.01.013

Bulgarelli, D., Rott, M., Schlaeppi, K., Ver Loren van Themaat, E., Ahmadinejad, N., Assenza, F., et al. (2012). Revealing structure and assembly cues for arabidopsis root-inhabiting bacterial microbiota. *Nature* 488, 91–95. doi: 10.1038/nature11336

Castillo, D. A., Kolesnikova, M. D., and Matsuda, S. P. (2013). An effective strategy for exploring unknown metabolic pathways by genome mining. *J. Am. Chem. Soc* 135, 5885–5894. doi: 10.1021/ja401535g

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., and Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 4, 7. doi: 10.1186/s13742-015-0047-8

Cheng, C. Y., Krishnakumar, V., Chan, A. P., Thibaud-Nissen, F., Schobel, S., and Town, C. D. (2017). Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome. *Plant J.* 89, 789–804. doi: 10.1111/tpj.13415

Chen, K., Wallis, J. W., McLellan, M. D., Larson, D. E., Kalicki, J. M., Pohl, C. S., et al. (2009). BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods* 6, 677–681. doi: 10.1038/nmeth.1363

Corpet, F. (1988). Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res.* 16, 10881–10890. doi: 10.1093/nar/16.22.10881

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21. doi: 10.1093/bioinformatics/bts635

Ehlting, J., Sauveplane, V., Olry, A., Ginglinger, J. F., Provart, N. J., and Werck-Reichhart, D. (2008). An extensive (co-)expression analysis tool for the cytochrome P450 superfamily in *Arabidopsis thaliana*. *BMC Plant Biol.* 8, 47. doi: 10.1186/1471-2229-8-47

Emsley, P., Lohkamp, B., Scott, W. G., and Cowtan, K. (2010). Features and development of coot. *Acta Crystallogr. D. Biol. Crystallogr.* 66, 486–501. doi: 10.1107/S0907444910007493

Erb, M., and Kliebenstein, D. J. (2020). Plant secondary metabolites as defenses, regulators, and primary metabolites: The blurred functional trichotomy. *Plant Physiol.* 184, 39–52. doi: 10.1104/pp.20.00433

Exposito-Alonso, M.500 Genomes Field Experiment Team, , Burbano, H. A., Bossdorf, O., Nielsen, R., Weigel, D. (2019). Natural selection on the *Arabidopsis thaliana* genome in present and future climates. *Nature* 573, 126–129. doi: 10.1038/s41586-019-1520-9

Fan, P., Wang, P., Lou, Y. R., Leong, B. J., Moore, B. M., Schenck, C. A., et al. (2020). Evolution of a plant gene cluster in *Solanaceae* and emergence of metabolic diversity. *Elife* 9, e56717. doi: 10.7554/eLife.56717.sa2

Fazio, G. C., Xu, R., and Matsuda, S. P. T. (2004). Genome mining to identify new plant triterpenoids. *J. Am. Chem. Soc* 126, 5678–5679. doi: 10.1021/ja0318784

Field, B., Fiston-Lavier, A. S., Kemen, A., Geisler, K., Quesneville, H., and Osbourn, A. E. (2011). Formation of plant metabolic gene clusters within dynamic chromosomal regions. *Proc. Natl. Acad. Sci.* 108, 16116–16121. doi: 10.1073/pnas.1109273108

Field, B., and Osbourn, A. E. (2008). Metabolic diversification–independent assembly of operon-like gene clusters in different plants. *Science* 320, 543–547. doi: 10.1126/science.1154990

Ghosh, S. (2016). Biosynthesis of structurally diverse triterpenes in plants: the role of oxidosqualene cyclase. *Proc. Indian Natl. Sci. Acad.* 82, 1189–1210. doi: 10.16943/ptinsa/2016/48578

Go, Y. S., Lee, S. B., Kim, H. J., Kim, J., Park, H. Y., Kim, J. K., et al. (2012). Identification of marneral synthase, which is critical for growth and development in arabidopsis. *Plant J.* 72, 791–804. doi: 10.1111/j.1365-313X.2012.05120.x

Huang, A. C., Jiang, T., Liu, Y. X., Bai, Y. C. Y., Reed, J., Qu, B., et al. (2019). A specialized metabolic network selectively modulates arabidopsis root microbiota. *Science* 364, eaau6389. doi: 10.1126/science.aau6389

Innerebner, G., Knief, C., and Vorholt, J. A. (2011). Protection of *Arabidopsis thaliana* against leaf-pathogenic *Pseudomonas syringae* by *Sphingomonas* strains in a controlled model system. *Appl. Environ. Microbiol.* 77, 3202–3210. doi: 10.1128/AEM.00133-11

Isah, T. (2019). Stress and defense responses in plant secondary metabolites production. *Biol. Res.* 52, 39. doi: 10.1186/s40659-019-0246-3

Jiao, W. B., and Schneeberger, K. (2020). Chromosome-level assemblies of multiple arabidopsis genomes reveal hotspots of rearrangements with altered evolutionary dynamics. *Nat. Commun.* 11, 989. doi: 10.1038/s41467-020-14779-y

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589. doi: 10.1038/s41586-021-03819-2

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., et al. (2008). Efficient control of population structure in model organism association mapping. *Genetics* 178, 1709–1723. doi: 10.1534/genetics.107.080101

Karasov, T. L., Neumann, M., Shirsekar, G., Monroe, G.PATHODOPSIS Team, , Weigel, D., et al. (2022) (Accessed November 2, 2022).

Katz, E., Li, J. J., Jaegle, B., Ashkenazy, H., Abrahams, S. R., Bagaza, C., et al. (2021). Genetic variation, environment and demography intersect to shape arabidopsis defense metabolite variation across Europe. *Elife* 10, e67784. doi: 10.7554/eLife.67784.sa2

Kautsar, S. A., Suarez Duran, H. G., Blin, K., Osbourn, A., and Medema, M. H. (2017). plantiSMASH: automated identification, annotation and expression analysis of plant biosynthetic gene clusters. *Nucleic Acids Res.* 45, W55–W63. doi: 10.1093/nar/gkx305

Kawakatsu, T., Huang, S. S. C., Jupe, F., Sasaki, E., Schmitz, R. J., Urich, M. A., et al. (2016). Epigenomic diversity in a global collection of *Arabidopsis thaliana* accessions. *Cell* 166, 492–505. doi: 10.1016/j.cell.2016.06.044

Krissinel, E., and Henrick, K. (2004). Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr. D. Biol. Crystallogr.* 60, 2256–2268. doi: 10.1107/S0907444904026460

Lardon, R., Wijnker, E., Keurentjes, J., and Geelen, D. (2020). The genetic framework of shoot regeneration in arabidopsis comprises master regulators and conditional fine-tuning factors. *Commun. Biol.* 3, 549. doi: 10.1038/s42003-020-01274-9

Lee, C. R., Svardal, H., Farlow, A., Exposito-Alonso, M., Ding, W., Novikova, P., et al. (2017). On the post-glacial spread of human commensal *Arabidopsis thaliana*. *Nat. Commun.* 8, 14458. doi: 10.1038/ncomms14458

Li, Y., Huang, Y., Bergelson, J., Nordborg, M., and Borevitz, J. O. (2010). Association mapping of local climate-sensitive quantitative trait loci in *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. U. S. A.* 107, 21199–21204. doi: 10.1073/pnas.1007431107

Lind, A. L., Wisecaver, J. H., Lameirasm, C., Wiemann, P., Palmer, J. M., Keller, N. P., et al. (2017). Drivers of genetic diversity in secondary metabolic gene clusters within a fungal species. *PloS Biol.* 15, e2003583. doi: 10.1371/journal.pbio.2003583

Li, Q., Ramasamy, S., Singh, P., Hagel, J. M., Dunemann, S. M., Chen, X., et al. (2020). Gene clustering and copy number variation in alkaloid metabolic pathways of opium poppy. *Nat. Commun.* 11, 1190. doi: 10.1038/s41467-020-15040-2

Liu, Z., Cheema, J., Vigouroux, M., Hill, L., Reed, J., Paajanen, P., et al. (2020a). Formation and diversification of a paradigm biosynthetic gene cluster in plants. *Nat. Commun.* 11, 5354. doi: 10.1038/s41467-020-19153-6

Liu, Z., Suarez Duran, H. G., Harnvanichvech, Y., Stephenson, M. J., Schranz, M. E., Nelson, D., et al. (2020b). Drivers of metabolic diversification: how dynamic genomic neighbourhoods generate new biosynthetic pathways in the brassicaceae. *New Phytol.* 227, 1109–1123. doi: 10.1111/nph.16338

Lodeiro, S., Schulz-Gasch, T., and Matsuda, S. P. T. (2005). Enzyme redesign: two mutations cooperate to convert cycloartenol synthase into an accurate lanosterol synthase. *J. Am. Chem. Soc* 127, 14132–14133. doi: 10.1021/ja053791j

Lodeiro, S., Xiong, Q., Wilson, W. K., Kolesnikova, M. D., Onak, C. S., and Matsuda, S. P. T. (2007). An oxidosqualene cyclase makes numerous products by diverse mechanisms: a challenge to prevailing concepts of triterpene biosynthesis. *J. Am. Chem. Soc* 129, 11213–11222. doi: 10.1021/ja073133u

Lundberg, D. S., Lebeis, S. L., Paredes, S. H., Yourstone, S., Gehring, J., Malfatti, S., et al. (2012). Defining the core *Arabidopsis thaliana* root microbiome. *Nature* 488, 86–90. doi: 10.1038/nature11237

Luo, Y., Wang, F., Huang, Y., Zhou, M., Gao, J., Yan, T., et al. (2019). *Sphingomonas* sp. Cra20 increases plant growth rate and alters rhizosphere microbial community structure of *Arabidopsis thaliana* under drought stress. *Front. Microbiol.* 10, 1221. doi: 10.3389/fmicb.2019.01221

Madeira, F., Park, Y. M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., et al. (2019). The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res.* 47, W636–W641. doi: 10.1093/nar/gkz268

Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., and Steinegger, M. (2022). ColabFold: making protein folding accessible to all. *Nat. Methods* 19, 679–682. doi: 10.1038/s41592-022-01488-1

Mohn, W. W., Yu, Z., Moore, E. R., and Muttray, A. F. (1999). Lessons learned from *Sphingomonas* species that degrade abietane triterpenoids. *J. Ind. Microbiol. Biotechnol.* 23, 374–379. doi: 10.1038/sj.jim.2900731

Morlacchi, P., Wilson, W. K., Xiong, Q., Bhaduri, A., Sttivend, D., Kolesnikova, M. D., et al. (2009). Product profile of PEN3: The last unexamined oxidosqualene cyclase in *Arabidopsis thaliana*. *Org. Lett.* 11, 2627–2630. doi: 10.1021/ol9005745

Nelson, D. R. (2009). The cytochrome p450 homepage. *Hum. Genomics* 4, 59–65. doi: 10.1186/1479-7364-4-1-59

Nelson, D., and Werck-Reichhart, D. (2011). A P450-centric view of plant evolution. *Plant J.* 66, 194–211. doi: 10.1111/j.1365-313X.2011.04529.x

Nützmann, H. W., Doerr, D., Ramírez-Colmenero, A., Sotelo-Fonseca, J. E., Wegel, E., Di Stefano, M., et al. (2020). Active and repressed biosynthetic gene clusters have spatially distinct chromosome states. *Proc. Natl. Acad. Sci. U. S. A.* 117, 13800–13809. doi: 10.1073/pnas.1920474117

Nützmann, H. W., and Osbourn, A. (2014). Gene clustering in plant specialized metabolism. *Curr. Opin. Biotechnol.* 26, 91–99. doi: 10.1016/j.copbio.2013.10.009

Nützmann, H. W., Scazzocchio, C., and Osbourn, A. (2018). Metabolic gene clusters in eukaryotes. *Annu. Rev. Genet.* 52, 159–183. doi: 10.1146/annurev-genet-120417-031237

Paquette, S. M., Bak, S., and Feyereisen, R. (2000). Intron-exon organization and phylogeny in a large superfamily, the paralogous cytochrome P450 genes of *Arabidopsis thaliana*. *DNA Cell Biol.* 19, 307–317. doi: 10.1089/10445490050021221

Patterson, N., Price, A. L., and Reich, D. (2006). Population structure and eigenanalysis. *PloS Genet.* 2, e190. doi: 10.1371/journal.pgen.0020190

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 38, 904–909. doi: 10.1038/ng1847

Roulé, T., Christ, A., Hussain, N., Huang, Y., Hartmann, C., Benhamed, M., et al. (2022). The lncRNA MARS modulates the epigenetic reprogramming of the marneral cluster in response. *Mol. Plant* 15, 840–856. doi: 10.1016/j.molp.2022.02.007

Samelak-Czajka, A., Marszalek-Zenczak, M., Marcinkowska-Swojak, M., Kozlowski, P., Figlerowicz, M., and Zmienko, A. (2017). MLPA-based analysis of copy number variation in plant populations. *Front. Plant Sci.* 8, 222. doi: 10.3389/fpls.2017.00222

Sayers, E. W., Bolton, E. E., Brister, J. R., Canese, K., Chan, J., Comeau, D. C., et al. (2022). Database resources of the national center for biotechnology information. *Nucleic Acids Res.* 50, D20–D26. doi: 10.1093/nar/gkab1112

Seren, Ü., Grimm, D., Fitz, J., Weigel, D., Nordborg, M., Borgwardt, K., et al. (2017). AraPheno: a public database for *Arabidopsis thaliana* phenotypes. *Nucleic Acids Res.* 45, D1054–D1059. doi: 10.1093/nar/gkw986

Shirai, K., Matsuda, F., Nakabayashi, R., Okamoto, M., Tanaka, M., Fujimoto, A., et al. (2017). A highly specific genome-wide association study integrated with transcriptome data reveals the contribution of copy number variations to specialized metabolites in *Arabidopsis thaliana* accessions. *Mol. Biol. Evol.* 34, 3111–3122. doi: 10.1093/molbev/msx234

Sohrabi, R., Ali, T., Harinantenaina Rakotondraibe, L., and Tholl, D. (2017). Formation and exudation of non-volatile products of the arabidiol triterpenoid degradation pathway in arabidopsis roots. *Plant Signal. Behav.* 12, e1265722. doi: 10.1080/15592324.2016.1265722

Sohrabi, R., Huh, J. H., Badieyan, S., Rakotondraibe, L. H., Kliebenstein, D. J., Sobrado, P., et al. (2015). In planta variation of volatile biosynthesis: an alternative biosynthetic route to the formation of the pathogen-induced volatile homoterpene DMNT *via* triterpene degradation in arabidopsis roots. *Plant Cell* 27, 874–890. doi: 10.1105/tpc.114.132209

Stanke, M., and Morgenstern, B. (2005). AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* 33, W465–W467. doi: 10.1093/nar/gki458

Thimmappa, R., Geisler, K., Louveau, T., O'Maille, P., and Osbourn, A. (2014). Triterpene biosynthesis in plants. *Annu. Rev. Plant Biol.* 65, 225–257. doi: 10.1146/annurev-arplant-050312-120229

Thoma, R., Schulz-Gasch, T., D'Arcy, B., Benz, J., Aebi, J., Dehmlow, H., et al. (2004). Insight into steroid scaffold formation from the structure of human oxidosqualene cyclase. *Nature* 432, 118–122. doi: 10.1038/nature02993

Togninalli, M., Seren, Ü., Freudenthal, J. A., Monroe, J. G., Meng, D., Nordborg, M., et al. (2020). AraPheno and the AraGWAS catalog 2020: a major database update including RNA-seq and knockout mutation data for *Arabidopsis thaliana*. *Nucleic Acids Res.* 48, D1063–D1068. doi: 10.1093/nar/gkz925

van Veen, H., Vashisht, D., Akman, M., Girke, T., Mustroph, A., Reinen, E., et al. (2016). Transcriptomes of eight *Arabidopsis thaliana* accessions reveal core conserved, genotype- and organ-specific responses to flooding stress. *Plant Physiol.* 172, 668–689. doi: 10.1104/pp.16.00472

Wada, M., Takahashi, H., Altaf-Ul-Amin, M., Nakamura, K., Hirai, M. Y., Ohta, D., et al. (2012). Prediction of operon-like gene clusters in the *Arabidopsis thaliana* genome based on co-expression analysis of neighboring genes. *Gene* 503, 56–64. doi: 10.1016/j.gene.2012.04.043

Wegel, E., Koumproglou, R., Shaw, P., and Osbourn, A. (2009). Cell type-specific chromatin decondensation of a metabolic gene cluster in oats. *Plant Cell* 21, 3926–3936. doi: 10.1105/tpc.109.072124

Wickham, H. (2016). *ggplot2* (Springer Cham). 2nd ed. doi: 10.1007/978-3-319-24277-4

Wisecaver, J. H., Borowsky, A. T., Tzin, V., Jander, G., Kliebenstein, D. J., and Rokas, A. (2017). A global coexpression network approach for connecting genes to specialized metabolic pathways in plants. *Plant Cell* 29, 944–959. doi: 10.1105/tpc.17.00009

Xiang, T., Shibuya, M., Katsube, Y., Tsutsumi, T., Otsuka, M., Zhang, H., et al. (2006). A new triterpene synthase from *Arabidopsis thaliana* produces a tricyclic triterpene with two hydroxyl groups. *Org. Lett.* 8, 2835–2838. doi: 10.1021/ol060973p

Xiong, Q., Wilson, W. K., and Matsuda, S. P. T. (2006). An arabidopsis oxidosqualene cyclase catalyzes iridal skeleton formation by grob fragmentation. *Angew. Chem. Int. Ed. Engl.* 45, 1285–1288. doi: 10.1002/anie.200503420

Yasumoto, S., Fukushima, E. O., Seki, H., and Muranaka, T. (2016). Novel triterpene oxidizing activity of *Arabidopsis thaliana* CYP716A subfamily enzymes. *FEBS Lett.* 590, 533–540. doi: 10.1002/1873-3468.12074

Yun, T., Li, H., Chang, P. C., Lin, M. F., Carroll, A., and McLean, C. Y. (2021). Accurate, scalable cohort variant calls using DeepVariant and GLnexus. *Bioinformatics* 36, 5582–5589. doi: 10.1093/bioinformatics/btaa1081

Yu, N., Nützmann, H. W., MacDonald, J. T., Moore, B., Field, B., Berriri, S., et al. (2016). Delineation of metabolic gene clusters in plant genomes by chromatin signatures. *Nucleic Acids Res.* 44, 2255–2265. doi: 10.1093/nar/gkw100

Zhan, C., Lei, L., Liu, Z., Zhou, S., Yang, C., Zhu, X., et al. (2020). Selection of a subspecies-specific diterpene gene cluster implicated in rice disease resistance. *Nat. Plants* 6, 1447–1454. doi: 10.1038/s41477-020-00816-7

Zmienko, A., Marszalek-Zenczak, M., Wojciechowski, P., Samelak-Czajka, A., Luczak, M., Kozlowski, P., et al. (2020). AthCNV: A map of DNA copy number variations in the arabidopsis genome. *Plant Cell* 32, 1797–1819. doi: 10.1105/tpc.19.00640

MATERIAŁY SUPLEMENTARNE

Marszalek-Zenczak M, Satyr A, Wojciechowski P, Zenczak M, Sobieszczanska P, Brzezinski K, Iefimenko T, Figlerowicz M, Zmienko A.

Analysis of Arabidopsis non-reference accessions reveals high diversity of metabolic gene clusters and discovers new candidate cluster members

Front Plant Sci. 2023 Jan 26;14:1104303. doi: 10.3389/fpls.2023.1104303.

5-letni IF = 12,061

Tabele suplementarne oraz Supplementary Data Sheet 15 dostępne online:
https://www.frontiersin.org/articles/10.3389/fpls.2023.1104303/full#supplementary-material

Supplemental Figures

Figure S1. IGV screens of genomic regions covering Arabidopsis MGCs

Figure S2. Copy number analysis of genes in thalianol (A), marneral (B) and tirucalladienol (C) gene clusters

Figure S3. Copy number analysis of genes in arabidiol/baruol gene cluster

Figure S4. Evidence supporting manual correction of genotype assignments in individual genes and accessions

Figure S5. WGS data-based evidence for a new type of deletion in the thalianol gene cluster spanning *CYP705A5*, *CYP708A2* and *THAS1*

Figure S6. Duplication of acyltransferase gene in Mitterberg-2-185

Figure S7. Partial deletion of *CYP705A12* in Mir-0

Figure S8. Differences between the countries in read coverage and mapping indicate that structural variants in tirucalladienol cluster genes are of local origin

Figure S9. Alternative *CYP716A2* gene models

Figure S10. Variation in WGS data coverage and mapping in the region spanning *CYP705A2*, *CYP705A3* and *BARS1* genes

Figure S11. *CYP705A2* duplication detected by RD assay correlates with the occurrence of pseudo-heterozygous SNPs in *CYP705A2* and *BARS1* loci

Figure S12. Multiple sequence alignment of *BARS1* genomic sequences reveals a common lack of the largest intron

Figure S13. Comparison of baruol synthase 1 protein NP_001329547.1 with proteins encoded by *BARS2* genes in Cvi-0, Eri-1 and Ler-0

Figure S14. Heterozygous SNPs in Cvi-0 co-localize with sequence differences between BARS1and its duplicate

Figure S15. Sequence comparison of *CYP705A2* and its duplicate *CYP705A2a*

Figure S16. PCR verification of group assignments based on the presence/absence of *CYP705A2*, *BARS1*, *CYP705A2a* and *BARS2* genes

Figure S17. Spread of PP-AA and PP-PP variants of arabidiol/baruol gene cluster in Arabidopsis population

Figure S18. Latitudes of origin among accessions with and without *CYP705A2a-BARS2* genes divided by country

Figure S19. Variability of the arabidiol/baruol gene cluster organization better explains latitudinal distribution of Arabidopsis accessions compared to variability of the thalianol gene cluster

Figure S20. Structural variation of the thalianol gene cluster has little impact on gene expression and root growth phenotypic variation

Figure S21. Differences in expression of arabidiol/baruol gene cluster among the accessions

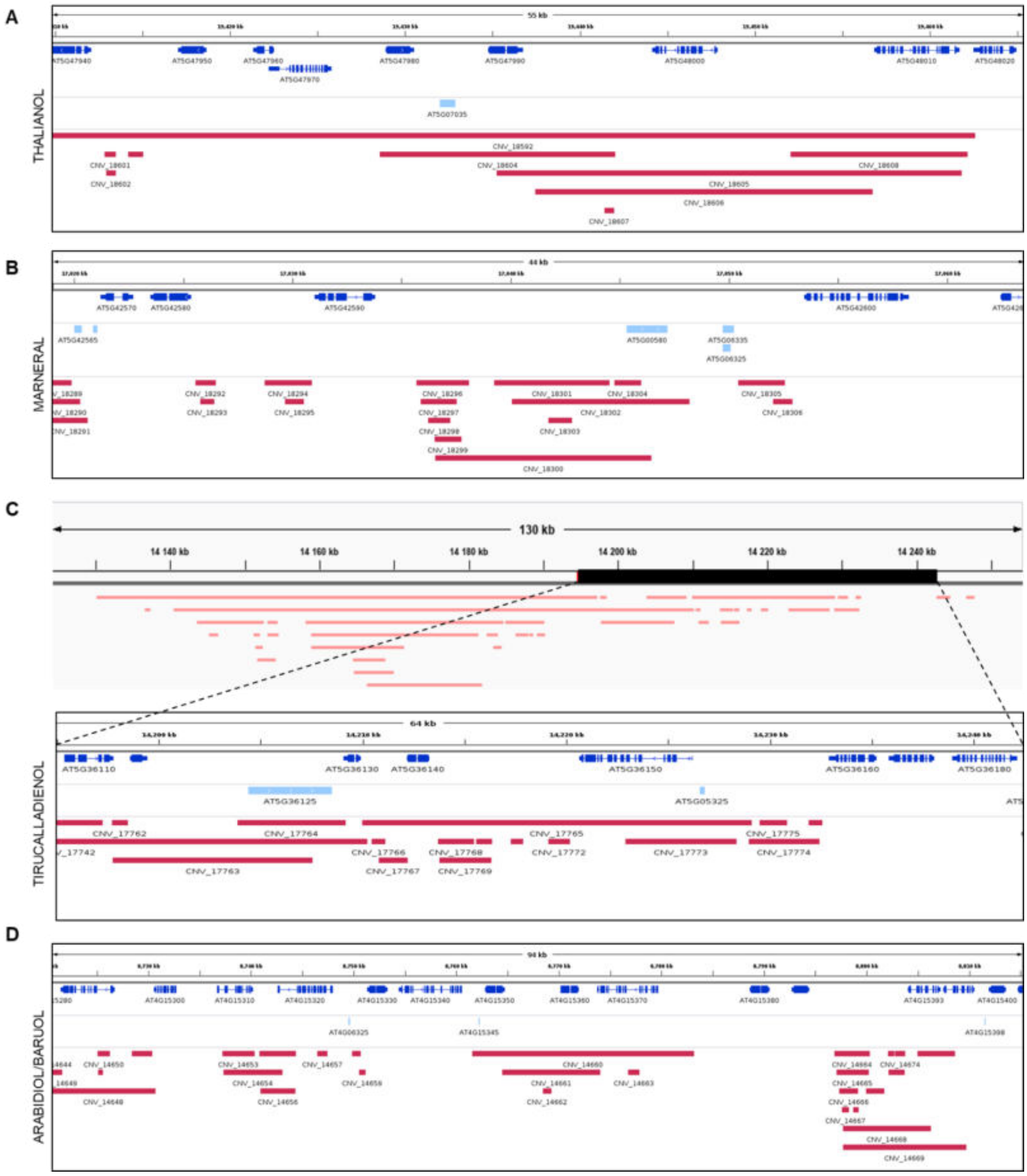


Figure S1. IGV screens of genomic regions covering Arabidopsis MGCs. A) Thalianol gene cluster; B) Marneral gene cluster; C) Tirucalladienol gene cluster; D) Arabidiol/baruol gene cluster. Dark blue – protein-coding genes. Araport 11 genomic loci are denoted. Light blue – noncoding genes. Red – CNVs extracted from AthCNV atlas (http://athcnv.ibch.poznan.pl/).

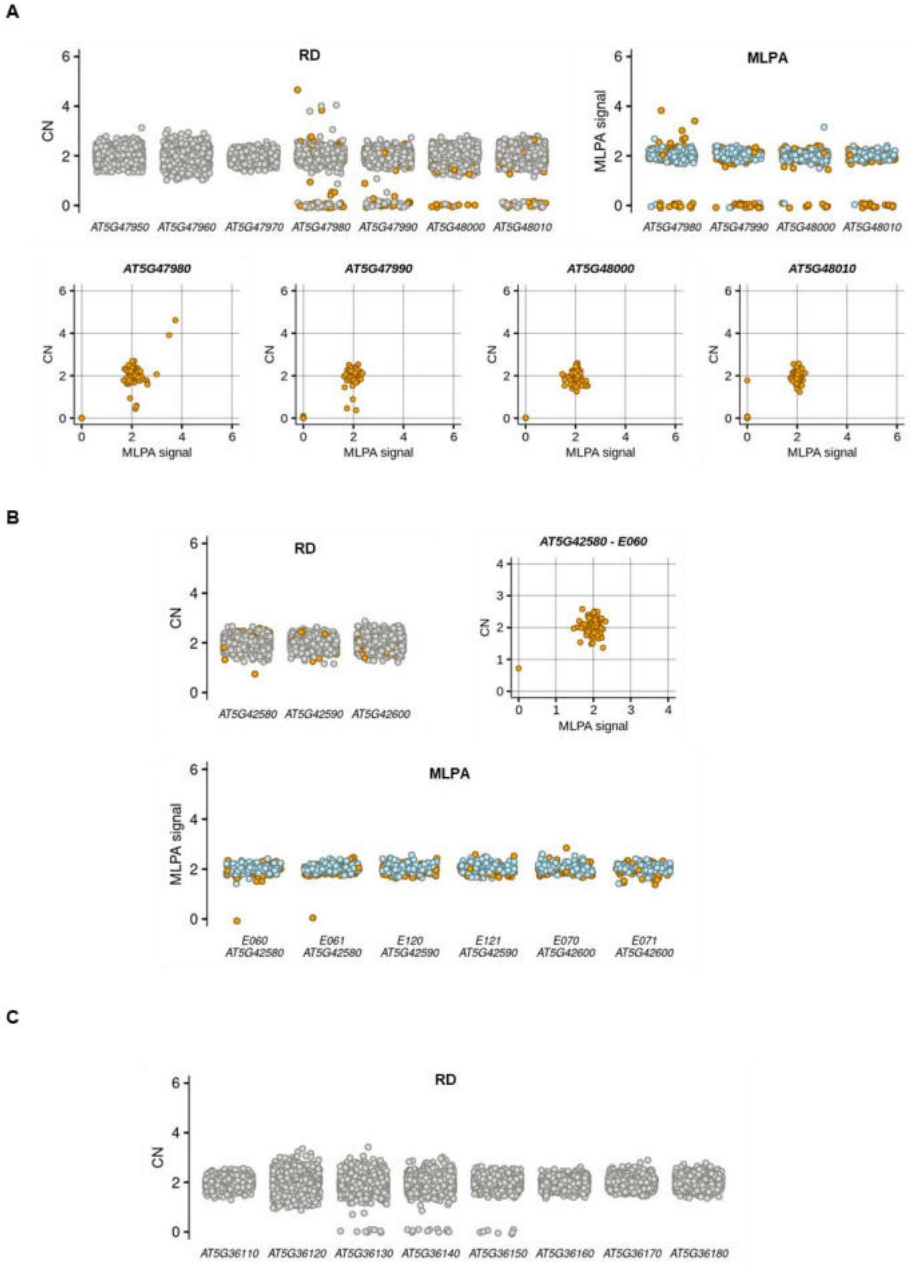


Figure S2. Copy number analysis of genes in thalianol (A), marneral (B) and tirucalladienol (C) gene clusters. Dots indicate individual accessions. Colours indicate a method of analysis: grey – RD only, blue – MLPA only, orange – both. The sets of analysed accessions differ between the methods (1,056 for RD; 232 for MLPA) but are identical for each gene.

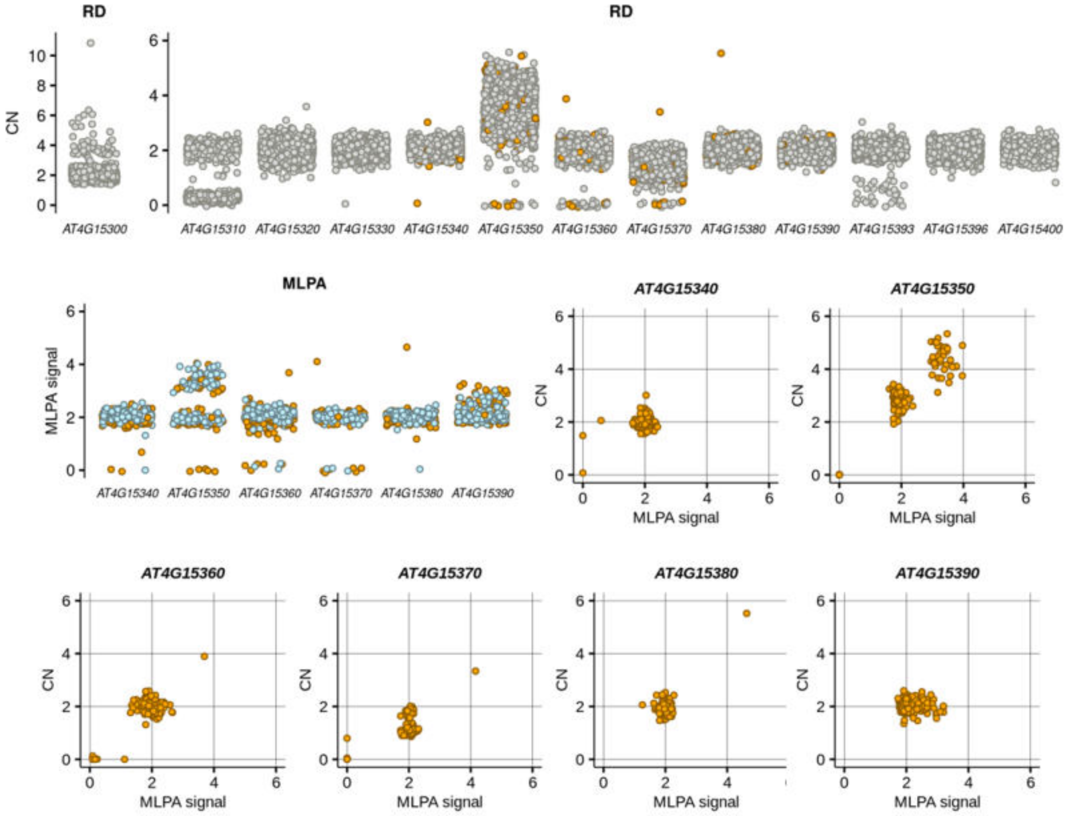


Figure S3. Copy number analysis of genes in arabidiol/baruol gene cluster. Dots indicate individual accessions. Colours indicate a method of analysis: grey – RD only, blue – MLPA only, orange – both. The sets of analysed accessions differ between the methods (1,056 for RD; 232 for MLPA) but are identical for each gene.

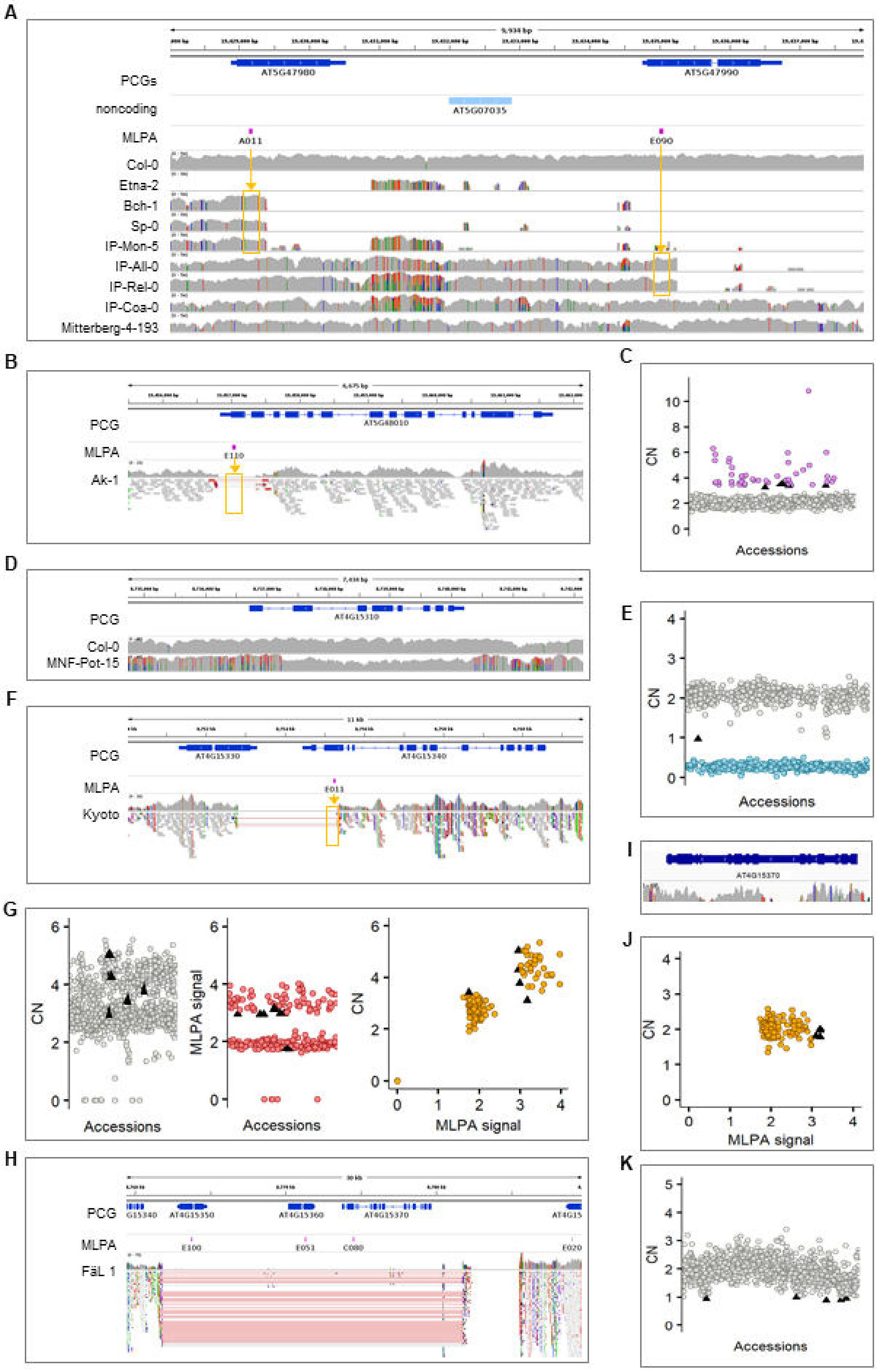


Figure S4. Evidence supporting manual correction of genotype assignments in individual genes and accessions. Details are presented in Supplemental Table S7.

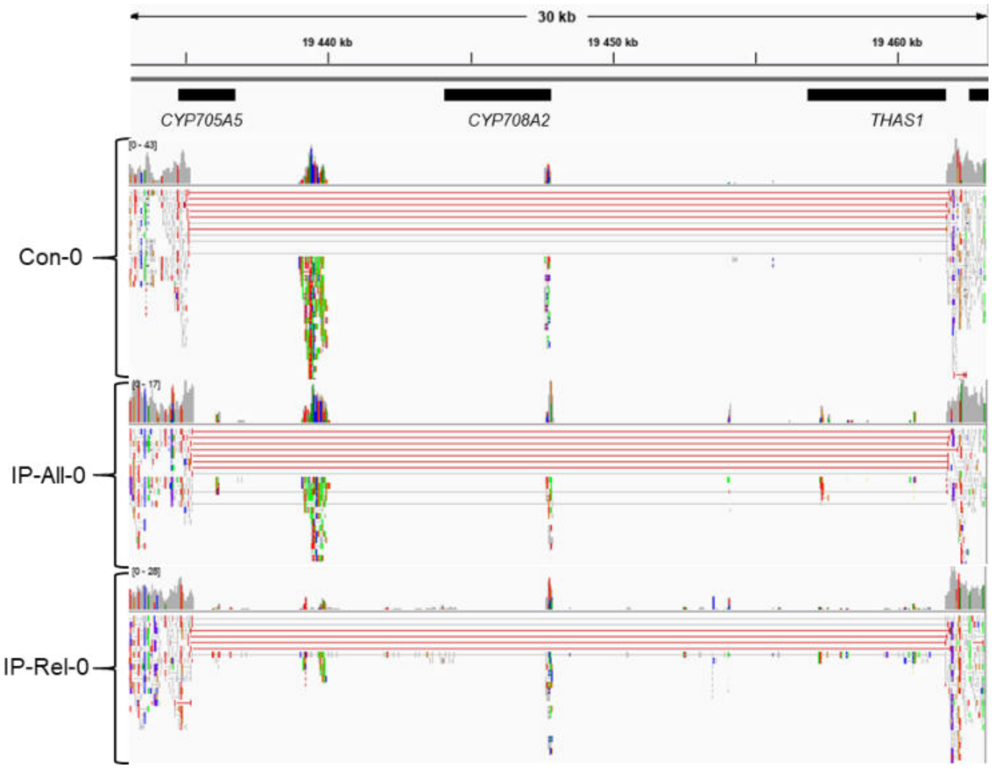
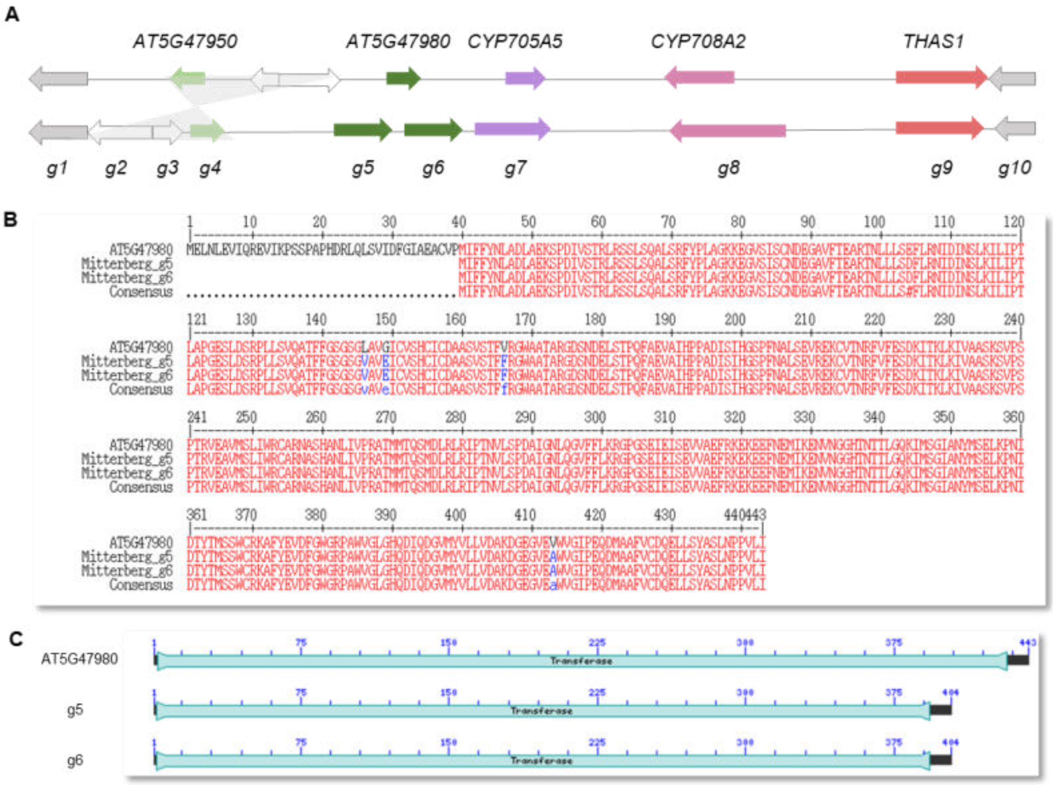


Figure S5. WGS data-based evidence for a new type of deletion in the thalianol gene cluster spanning *CYP705A5*, *CYP708A2* and *THAS1*.



```
            1        10        20        30        40        50        60        70        80        90       100       110       120
AT5G47980   MELNLEVIQREVIKPSSPAPHDRLQLSVIDFGIAEACVPMIFFYNLADLAEKSPDIVSTRLRSSLSQALSRFYPLAGKKEGVSISCNDEGAVFTEARTNLLLSEFLRNIDINSLKILIPT
Mitterberg_g5                                          MIFFYNLADLAEKSPDIVSTRLRSSLSQALSRFYPLAGKKEGVSISCNDEGAVFTEARTNLLLSDFLRNIDINSLKILIPT
Mitterberg_g6                                          MIFFYNLADLAEKSPDIVSTRLRSSLSQALSRFYPLAGKKEGVSISCNDEGAVFTEARTNLLLSDFLRNIDINSLKILIPT
Consensus   .......................................MIFFYNLADLAEKSPDIVSTRLRSSLSQALSRFYPLAGKKEGVSISCNDEGAVFTEARTNLLLS#FLRNIDINSLKILIPT

            121      130       140       150       160       170       180       190       200       210       220       230       240
AT5G47980   LAPGESLDSRPLLSVQATFFGSGSGLAVGICVSHCICDAASVSTFVRGWAATARGDSNDELSTPQFAEVAIHPPADISIHGSPFNALSEVREKCVTNRFVFESDKITKLKIVAASKSVPS
Mitterberg_g5   LAPGESLDSRPLLSVQATFFGSGSGVAVEICVSHCICDAASVSTFFRGWAATARGDSNDELSTPQFAEVAIHPPADISIHGSPFNALSEVREKCVTNRFVFESDKITKLKIVAASKSVPS
Mitterberg_g6   LAPGESLDSRPLLSVQATFFGSGSGVAVEICVSHCICDAASVSTFFRGWAATARGDSNDELSTPQFAEVAIHPPADISIHGSPFNALSEVREKCVTNRFVFESDKITKLKIVAASKSVPS
Consensus   LAPGESLDSRPLLSVQATFFGSGSGvAVeICVSHCICDAASVSTFfRGWAATARGDSNDELSTPQFAEVAIHPPADISIHGSPFNALSEVREKCVTNRFVFESDKITKLKIVAASKSVPS

            241      250       260       270       280       290       300       310       320       330       340       350       360
AT5G47980   PTRVEAVMSLIWRCARNASHANLIVPRATMMTQSMDLRLRIPTNVLSPDAIGNLQGVFFLKRGPGSEIEISEVVAEFRKEKEEFNEMIKENVNGGHTNTTLGQKIMSGIANYMSELKPNI
Mitterberg_g5   PTRVEAVMSLIWRCARNASHANLIVPRATMMTQSMDLRLRIPTNVLSPDAIGNLQGVFFLKRGPGSEIEISEVVAEFRKEKEEFNEMIKENVNGGHTNTTLGQKIMSGIANYMSELKPNI
Mitterberg_g6   PTRVEAVMSLIWRCARNASHANLIVPRATMMTQSMDLRLRIPTNVLSPDAIGNLQGVFFLKRGPGSEIEISEVVAEFRKEKEEFNEMIKENVNGGHTNTTLGQKIMSGIANYMSELKPNI
Consensus   PTRVEAVMSLIWRCARNASHANLIVPRATMMTQSMDLRLRIPTNVLSPDAIGNLQGVFFLKRGPGSEIEISEVVAEFRKEKEEFNEMIKENVNGGHTNTTLGQKIMSGIANYMSELKPNI

            361      370       380       390       400       410       420       430      440443
AT5G47980   DTYTMSSWCRKAFYEVDFGWGRPAWVGLGHQDIQDGVMYVLLVDAKDGEGVEVWVGIPEQDMAAFVCDQELLSYASLNPPVLI
Mitterberg_g5   DTYTMSSWCRKAFYEVDFGWGRPAWVGLGHQDIQDGVMYVLLVDAKDGEGVEAWVGIPEQDMAAFVCDQELLSYASLNPPVLI
Mitterberg_g6   DTYTMSSWCRKAFYEVDFGWGRPAWVGLGHQDIQDGVMYVLLVDAKDGEGVEAWVGIPEQDMAAFVCDQELLSYASLNPPVLI
Consensus   DTYTMSSWCRKAFYEVDFGWGRPAWVGLGHQDIQDGVMYVLLVDAKDGEGVEaWVGIPEQDMAAFVCDQELLSYASLNPPVLI
```

Figure S6. Duplication of acyltransferase gene in Mitterberg-2-185. A) Thalianol gene cluster organization in Col-0 (upper) and Mitterberg-2-185 *de novo* assembly (lower). Corresponding genes have the same colors. The region of inversion is marked in grey. B) Sequence alignment of reference AT5G47980 protein with predicted proteins from Mitterberg-2-185. C) Conserved domain prediction. Transferase – transferase domain (pfam02458).

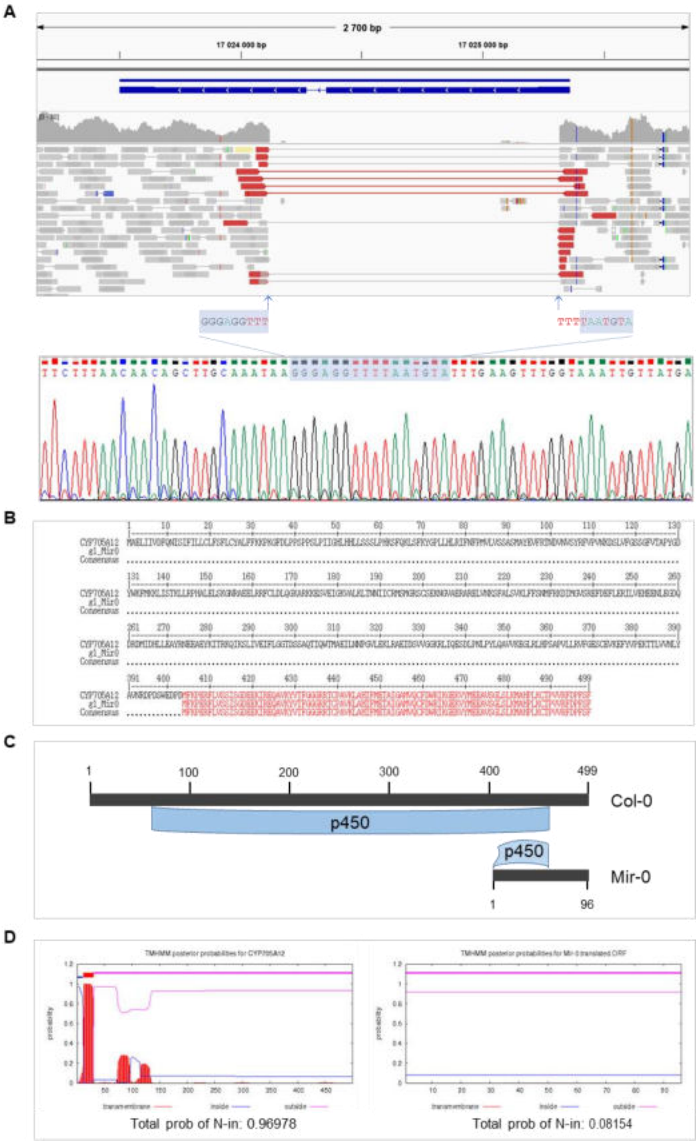


Figure S7. Partial deletion of *CYP705A12* in Mir-0. A) WGS-based and Sanger sequencing-based data concordantly indicate there is a 1,202-bp deletion in *CYP705A12* gene sequence in Mir-0. The reference *CYP705A12* model is presented in blue. B) Predicted truncated protein in Mir-0. Multiple protein alignment made with Multalin. C) Full (reference) and partial (Mir-0) cytochrome P450 superfamily domain present in the respective proteins. D) TMHMM server prediction probabilities for CYP705A12 (left) and its truncated homolog in Mir-0 (right).

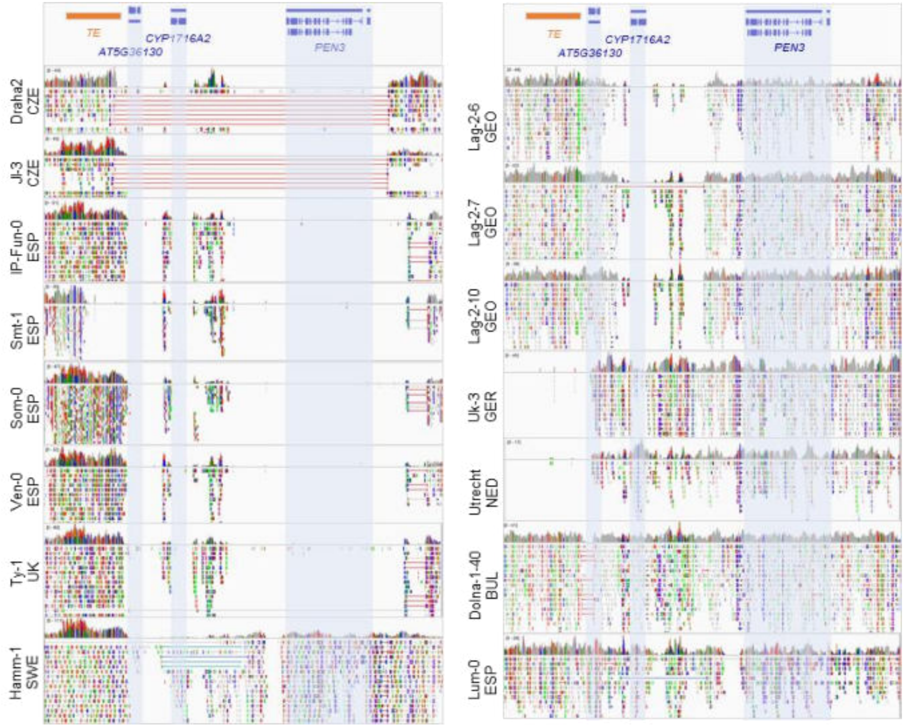


Figure S8. Differences between the countries in read coverage and mapping indicate that structural variants in tirucalladienol cluster genes are of local origin. The picture presents data for all 15 accessions with detected copy number changes in tirucalladienol cluster genes.

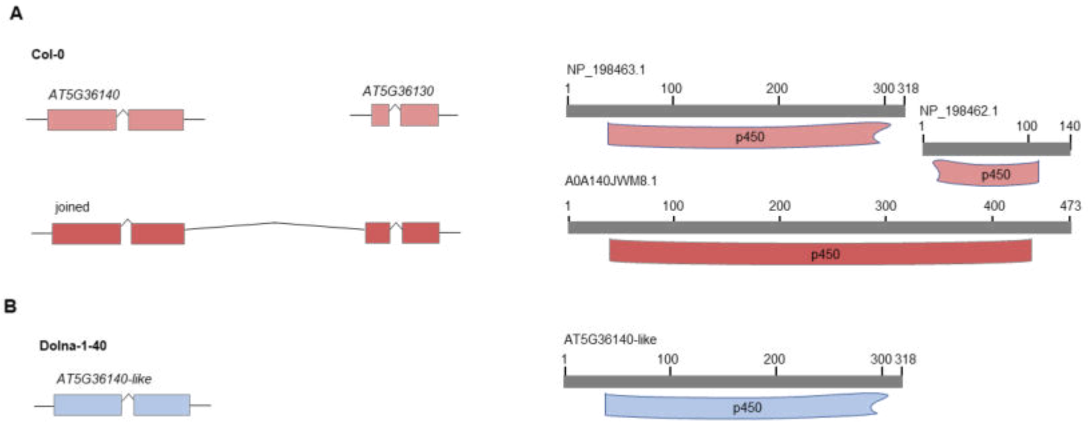


Figure S9. Alternative *CYP716A2* gene models. A) According to Araport 11 annotation, there are two separate genes, *AT5G36130* and *AT5G36140* (*CYP716A2*) in the genome. The Augustus tool predicted the same gene models and additionally an alternative joint model. The joint gene encodes a protein identical to the one predicted by Yatsumoto et al. (2016), for which full-length cDNA had been isolated. B) In Dolna-1-40 *AT5G36130* is absent, as indicated by the analysis of its *de novo* genomic assembly. The predicted ORF encodes a protein identical to *AT5G36140*, which lacks C-part of p450 superfamily domain (cl12078).

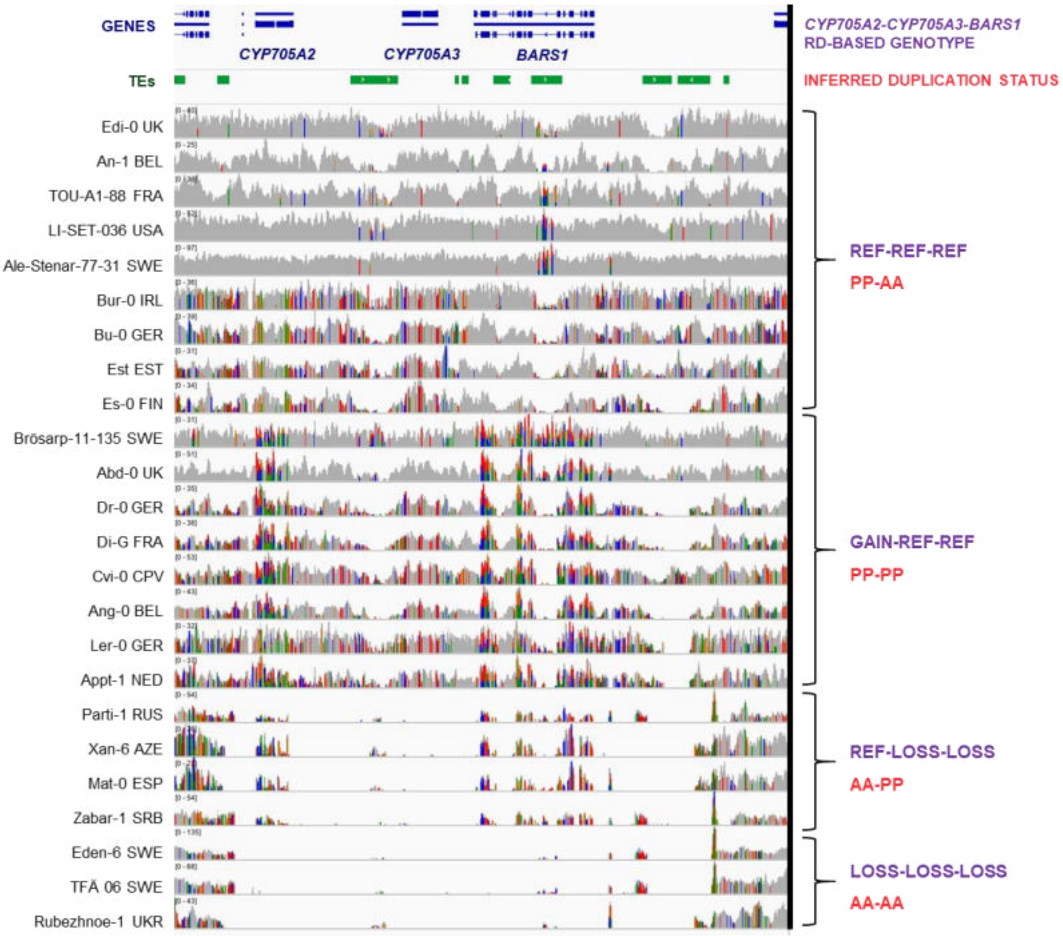


Figure S10. Variation in WGS data coverage and mapping in the region spanning *CYP705A2*, *CYP705A3* and *BARS1* genes. The presence (P) / absence (A) of *CYP705A2, BARS1* and their duplicates was inferred based on a combination of RD genotyping results and the SNP analysis at *CYP705A2* and *BARS1* loci (see Supplemental Table S11 and Methods for details).

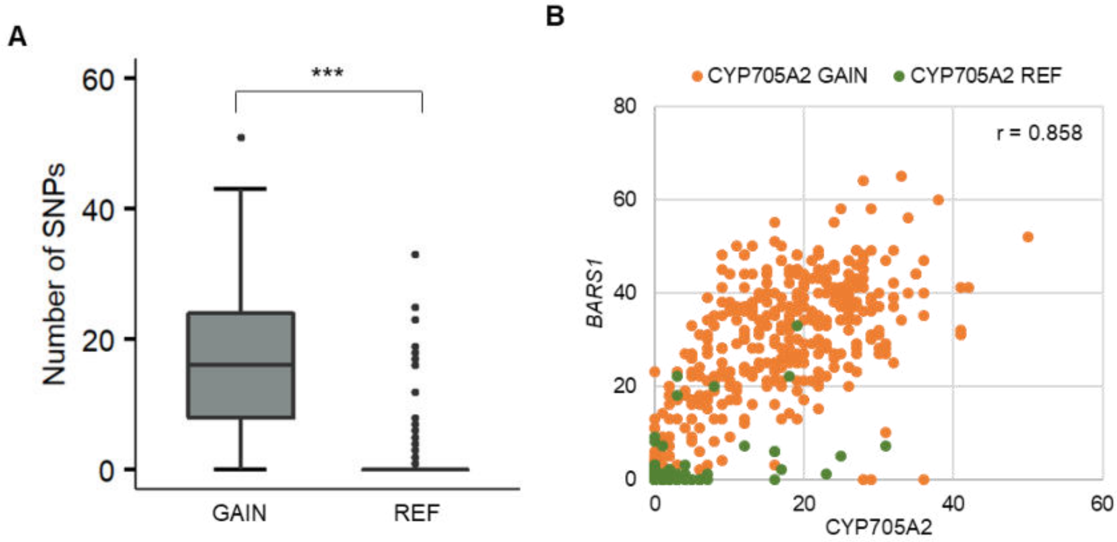


Figure S11. *CYP705A2* duplication detected by RD assay correlates with the occurrence of heterozygous SNPs at *CYP705A2* and *BARS1* loci. A) Number of heterozygous SNPs found in *CYP705A2* coding sequence in accessions with GAIN and REF genotypes. Boxplots show median (inner line) and inner quartiles (box). Whiskers extend to the highest and lowest values no greater than 1.5 times the inner quartile range. Asterisks indicate statistical significance (Wilcoxon rank sum test with continuity correction, ***p.value <0.001). B) Correlation of heterozygous SNP frequency between *CYP705A2* and *BARS1*. Colors indicate accessions with varying *CYP705A2* copy numbers. r – Pearson's correlation coefficient.

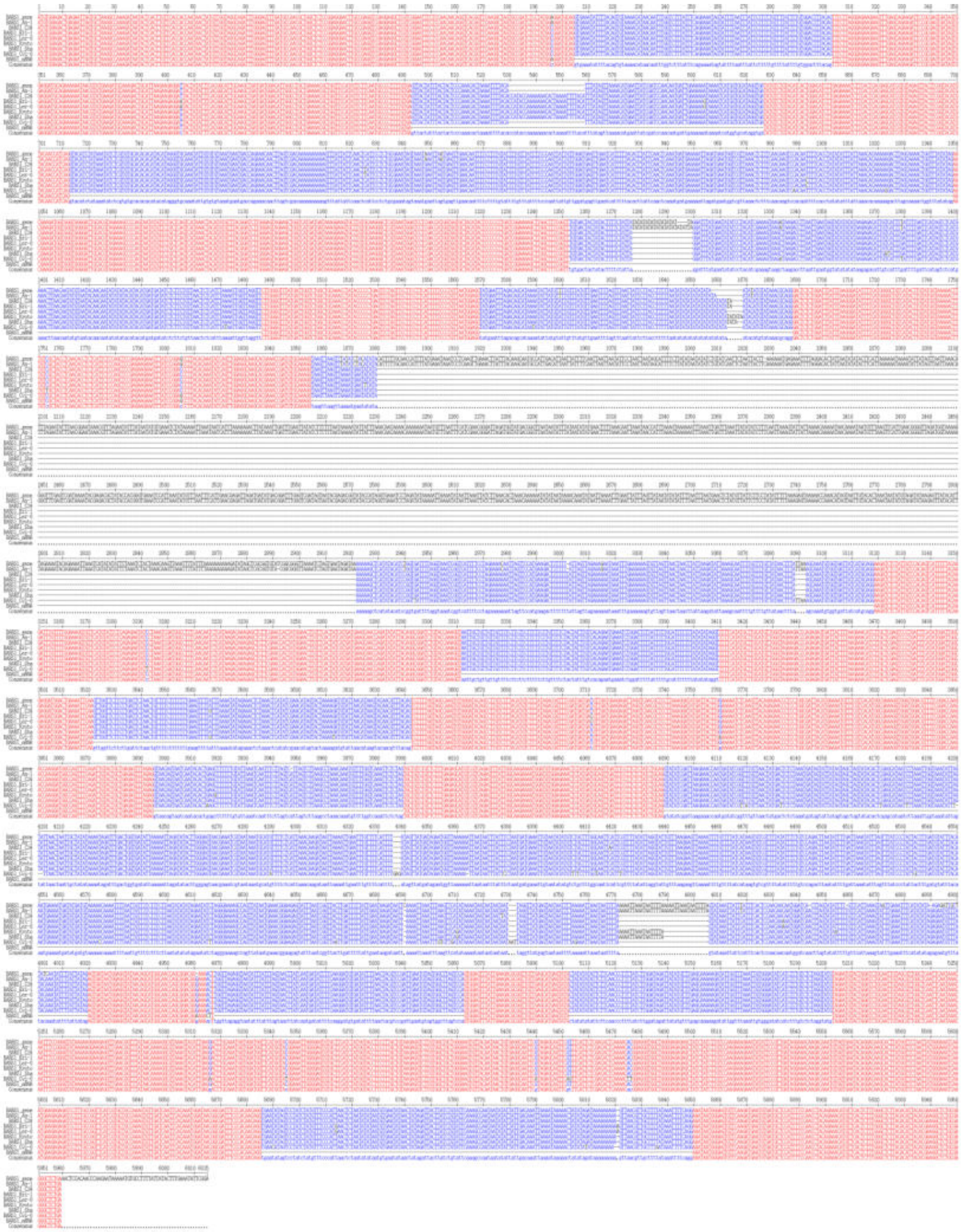


Figure S12. Multiple sequence alignment of *BARS1* genomic sequences reveals a common lack of the largest intron. Sequence order from the top: Col-0; An-1; C24; Eri-1; Ler-0; Kyoto; Sha; Cvi-0; BARS1_mRNA (to indicate exon-intron organization) and consensus.

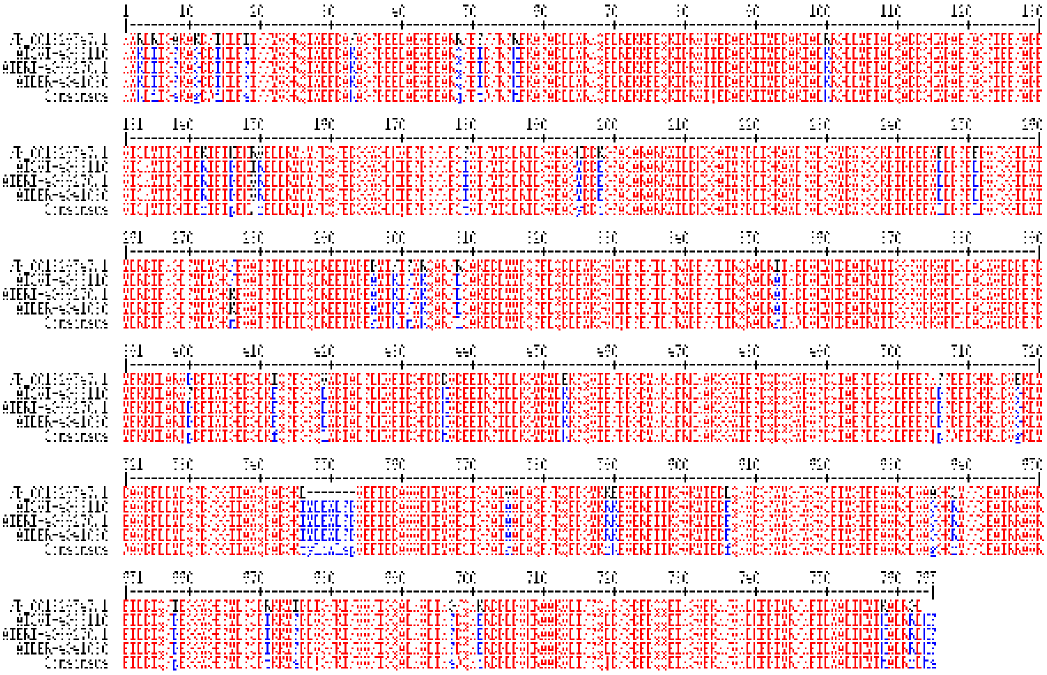
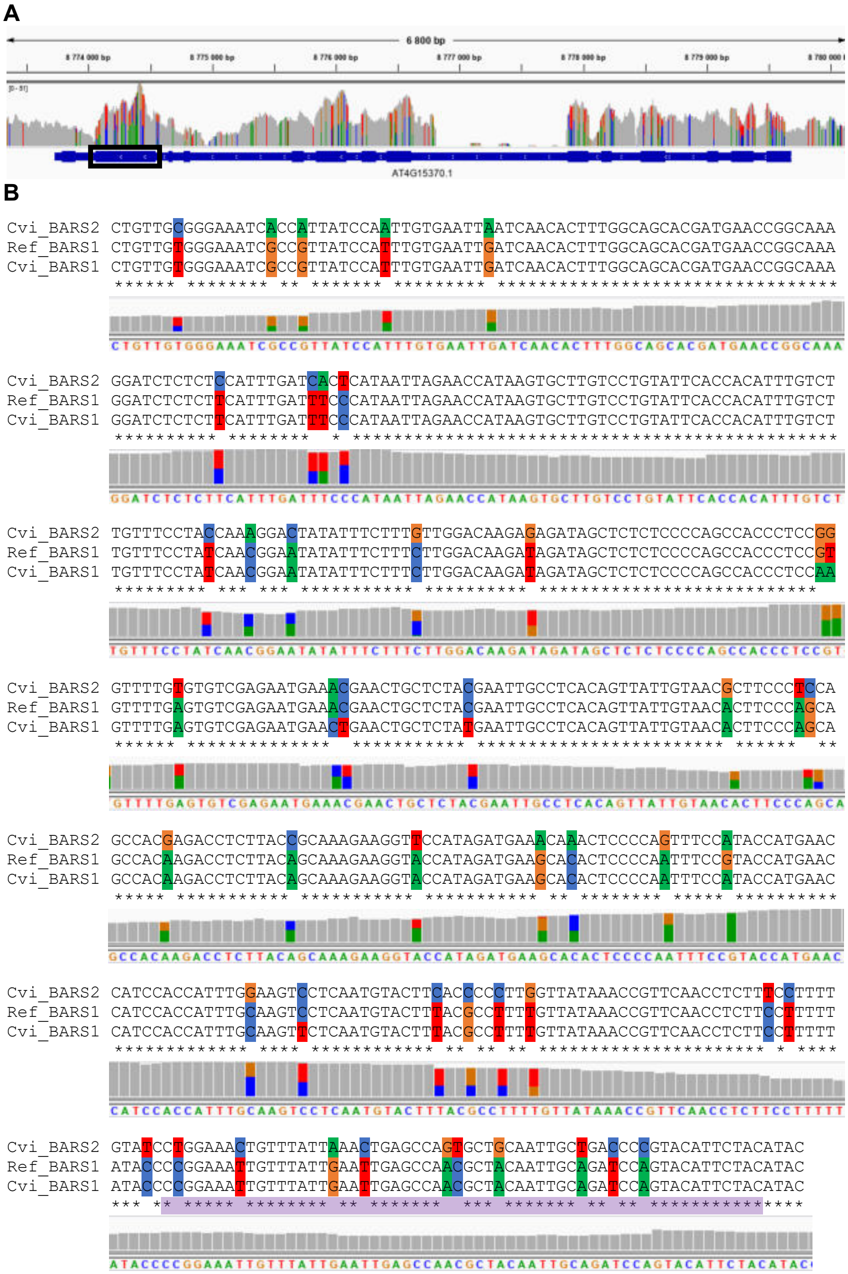


Figure S13. Comparison of baruol synthase 1 protein NP_001329547.1 with proteins encoded by *BARS2* genes in Cvi-0, Eri-1 and Ler-0. Multiple sequence alignment generated with Multalin with default parameters.


A
6 800 bp
8 774 000 bp
8 775 000 bp
8 776 000 bp
8 777 000 bp
8 778 000 bp
8 779 000 bp
8 780 000 b
[0 - 51]
AT4G15370.1
B
CTGTTGTGGGAAATCGCCGTTATCCATTTGTGAATTGATCAACACTTTGGCAGCACGATGAACCGGCAAA
GGATCTCTCTTCATTTGATTTCCCATAATTAGAACCATAAGTGCTTGTCCTGTATTCACCACATTTGTCT
TGTTTCCTATCAACGGAATATATTTCTTTCTTGGACAAGATAGATAGCTCTCTCCCCAGCCACCCTCCGT
GTTTTGAGTGTCGAGAATGAAACGAACTGCTCTACGAATTGCCTCACAGTTATTGTAACACTTCCCAGCA
GCCACAAGACCTCTTACAGCAAAGAAGGTACCATAGATGAAGCACACTCCCCAATTTCCGTACCATGAAC
CATCCACCATTTGCAAGTCCTCAATGTACTTTACGCCTTTTGTTATAAACCGTTCAACCTCTTCCTTTTT
ATACCCCGGAAATTGTTTATTGAATTGAGCCAACGCTACAATTGCAGATCCAGTACATTCTACATAC

```
Cvi_BARS2 CTGTTGCGGGAAATCACCATTATCCAATTGTGAATTAATCAACACTTTGGCAGCACGATGAACCGGCAAA
Ref_BARS1 CTGTTGTGGGAAATCGCCGTTATCCATTTGTGAATTGATCAACACTTTGGCAGCACGATGAACCGGCAAA
Cvi_BARS1 CTGTTGTGGGAAATCGCCGTTATCCATTTGTGAATTGATCAACACTTTGGCAGCACGATGAACCGGCAAA
          ****** ******** ** ******* ********* ********************************

Cvi_BARS2 GGATCTCTCTCCATTTGATCACTCATAATTAGAACCATAAGTGCTTGTCCTGTATTCACCACATTTGTCT
Ref_BARS1 GGATCTCTCTTCATTTGATTTCCCATAATTAGAACCATAAGTGCTTGTCCTGTATTCACCACATTTGTCT
Cvi_BARS1 GGATCTCTCTTCATTTGATTTCCCATAATTAGAACCATAAGTGCTTGTCCTGTATTCACCACATTTGTCT
          ********** ********  * ***********************************************

Cvi_BARS2 TGTTTCCTACCAAAGGACTATATTTCTTTGTTGGACAAGAGAGATAGCTCTCTCCCCAGCCACCCTCCGG
Ref_BARS1 TGTTTCCTATCAACGGAATATATTTCTTTCTTGGACAAGATAGATAGCTCTCTCCCCAGCCACCCTCCGT
Cvi_BARS1 TGTTTCCTATCAACGGAATATATTTCTTTCTTGGACAAGATAGATAGCTCTCTCCCCAGCCACCCTCCAA
          ********* *** *** *********** ********** ***************************  

Cvi_BARS2 GTTTTGTGTGTCGAGAATGAAACGAACTGCTCTACGAATTGCCTCACAGTTATTGTAACGCTTCCCTCCA
Ref_BARS1 GTTTTGAGTGTCGAGAATGAAACGAACTGCTCTACGAATTGCCTCACAGTTATTGTAACACTTCCCAGCA
Cvi_BARS1 GTTTTGAGTGTCGAGAATGAACTGAACTGCTCTATGAATTGCCTCACAGTTATTGTAACACTTCCCAGCA
          ****** **************  *********** ************************ ******  **

Cvi_BARS2 GCCACGAGACCTCTTACCGCAAAGAAGGTTCCATAGATGAAACAAACTCCCCAGTTTCCATACCATGAAC
Ref_BARS1 GCCACAAGACCTCTTACAGCAAAGAAGGTACCATAGATGAAGCACACTCCCCAATTTCCGTACCATGAAC
Cvi_BARS1 GCCACAAGACCTCTTACAGCAAAGAAGGTACCATAGATGAAGCACACTCCCCAATTTCCATACCATGAAC
          ***** *********** *********** *********** ** ******** ***** **********

Cvi_BARS2 CATCCACCATTTGGAAGTCCTCAATGTACTTCACCCCCTTGGTTATAAACCGTTCAACCTCTTTCCTTTT
Ref_BARS1 CATCCACCATTTGCAAGTCCTCAATGTACTTTACGCCTTTTGTTATAAACCGTTCAACCTCTTCCTTTTT
Cvi_BARS1 CATCCACCATTTGCAAGTTCTCAATGTACTTTACGCCTTTTGTTATAAACCGTTCAACCTCTTCCTTTTT
          ************* **** ************ ** ** ** ********************** * ****

Cvi_BARS2 GTATCCTGGAAACTGTTTATTAAACTGAGCCAGTGCTGCAATTGCTGACCCCGTACATTCTACATAC
Ref_BARS1 ATACCCCGGAAATTGTTTATTGAATTGAGCCAACGCTACAATTGCAGATCCAGTACATTCTACATAC
Cvi_BARS1 ATACCCCGGAAATTGTTTATTGAATTGAGCCAACGCTACAATTGCAGATCCAGTACATTCTACATAC
          *** ** ***** ******** ** *******  *** ******* ** ** ***************
```

Figure S14. Heterozygous SNPs in Cvi-0 co-localize with sequence differences between BARS1and its duplicate. A) Read coverage at BARS1 locus for Cvi-0 (reads mapped to the reference genome). B) Multiple sequence alignment of the *BARS1* fragment (region chr4:8774060-8774546, marked by the black box in A) with *BARS1* and its duplicate (denoted as *BARS2*) from Cvi-0. IGV screenshots presenting Cvi-0 read coverage and the SNP positions are overlaid. The purple block indicates the position of the C080 MLPA probe.

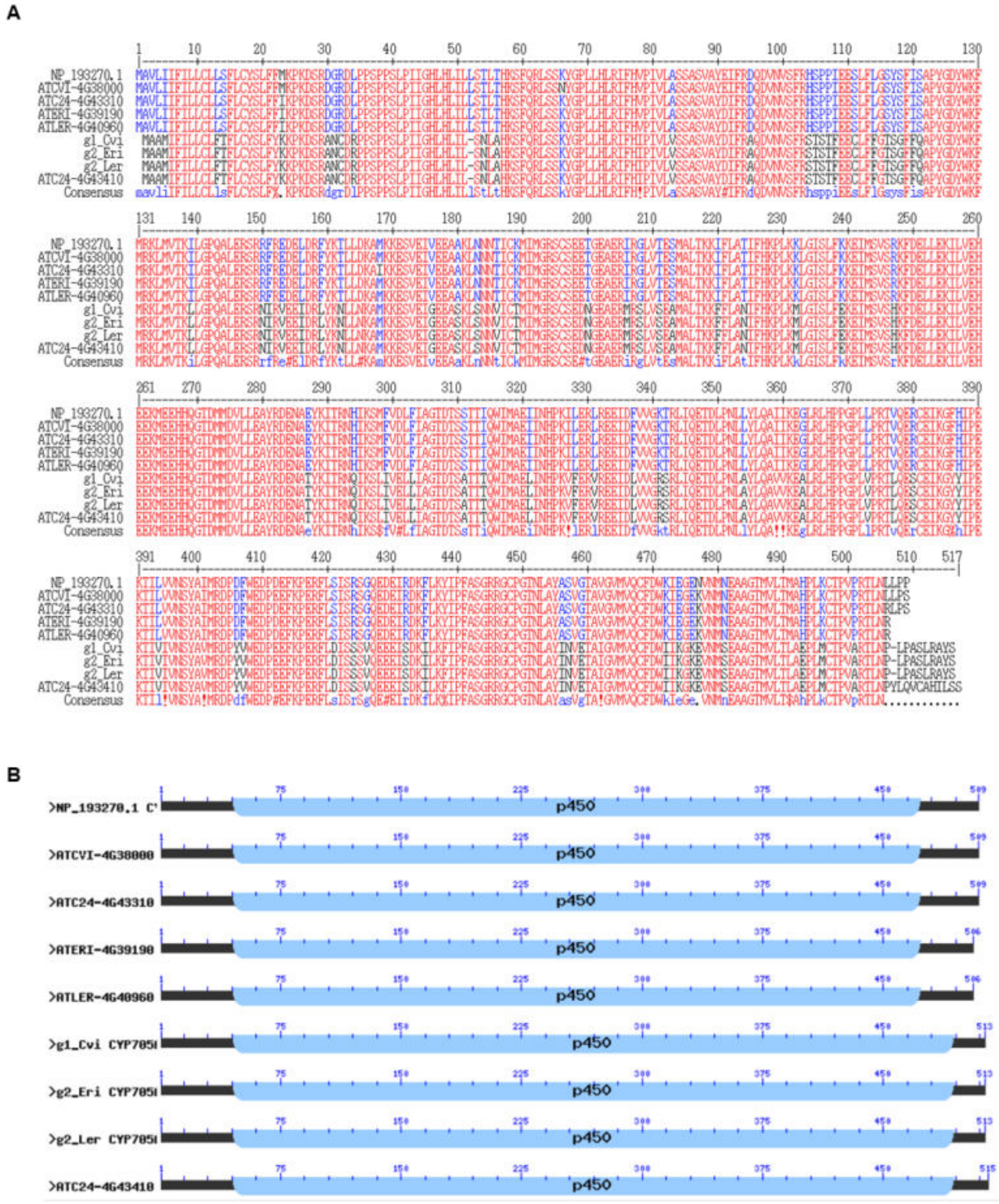


Figure S15. Sequence comparison of *CYP705A2* and its duplicate *CYP705A2a*. A) Multiple protein alignment of the proteins encoded by *CYP705A2* in Col-0, Cvi-0, C24, Eri-1 and Ler-0 (NP_193270.1, ATCVI-4G38000, ATC24-4G43310, ATERI-4G39190, ATLER-4G40960, respectively) and proteins encoded by CYP705A2 in Cvi-0, Eri-1, Ler-0 and C24 (g1_Cvi, g2_Eri, g2_Ler, ATC24-4G43410, respectively). B) Conserved protein domains in CYP705A2 and CYP705A2a sequences found by searching the Pfam database.

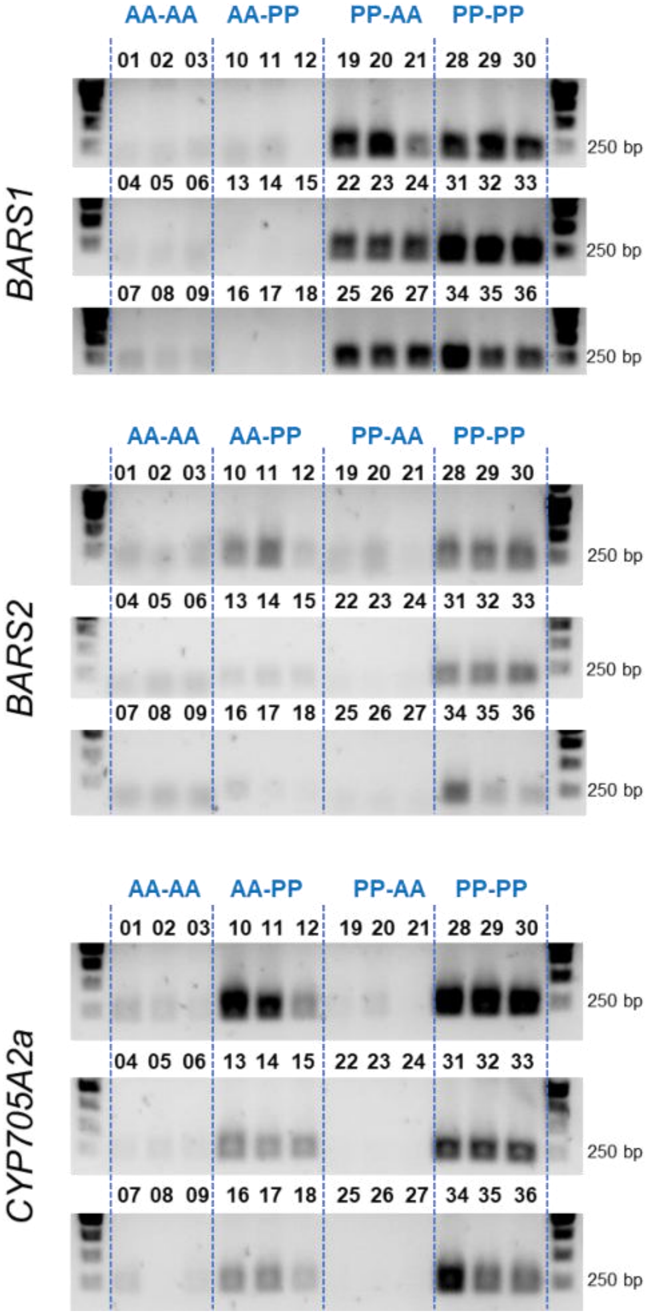


Figure S16. PCR verification of group assignments based on the presence/absence of *BARS1*, *CYP705A2a* and *BARS2* genes. Sample identities are provided in Supplemental Table S11.

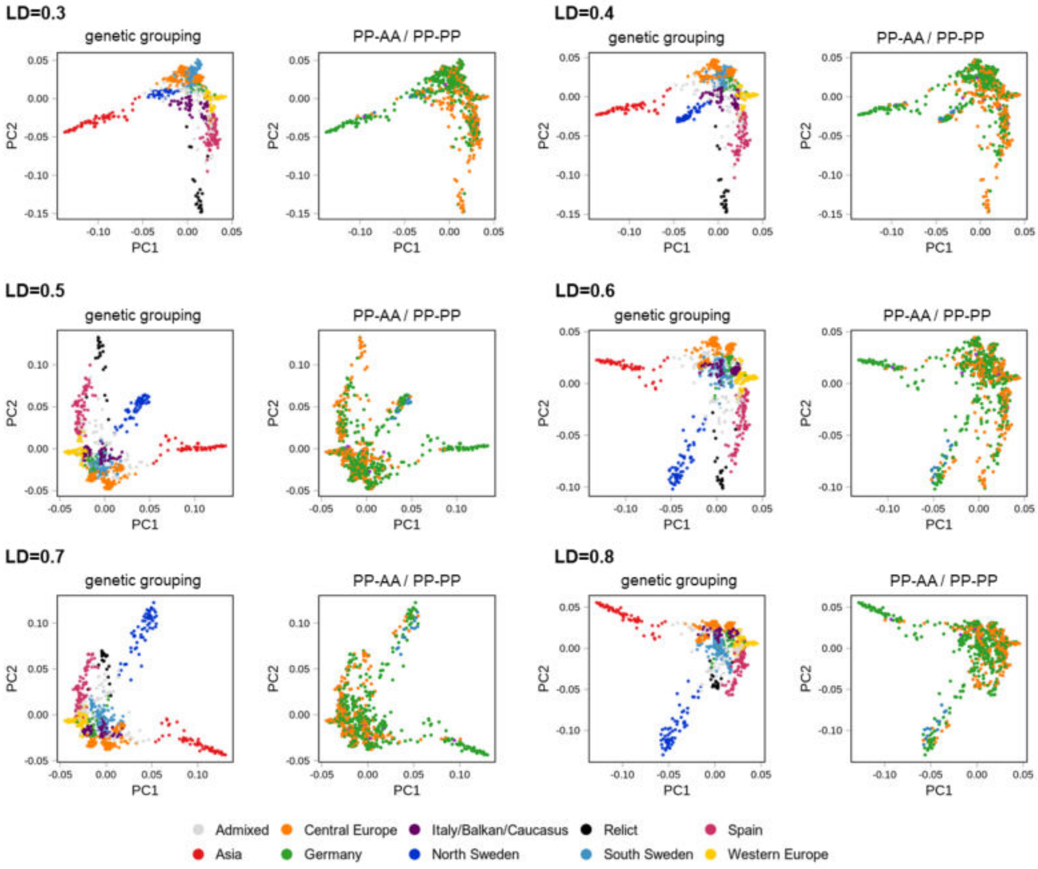


Figure S17. Spread of PP-AA and PP-PP variants of arabidiol/baruol gene cluster in Arabidopsis population. Principal component analysis (PCA) plots were generated with varying LD parameter. U.S.A. accessions were excluded from the analysis to better visualize other groups. Plots are colored according to main genetic groups (left) or *CYP705A2-BARS1* duplication status (right). Plots generated at LD=0.3 are also presented in Fig. 4 in the main text.

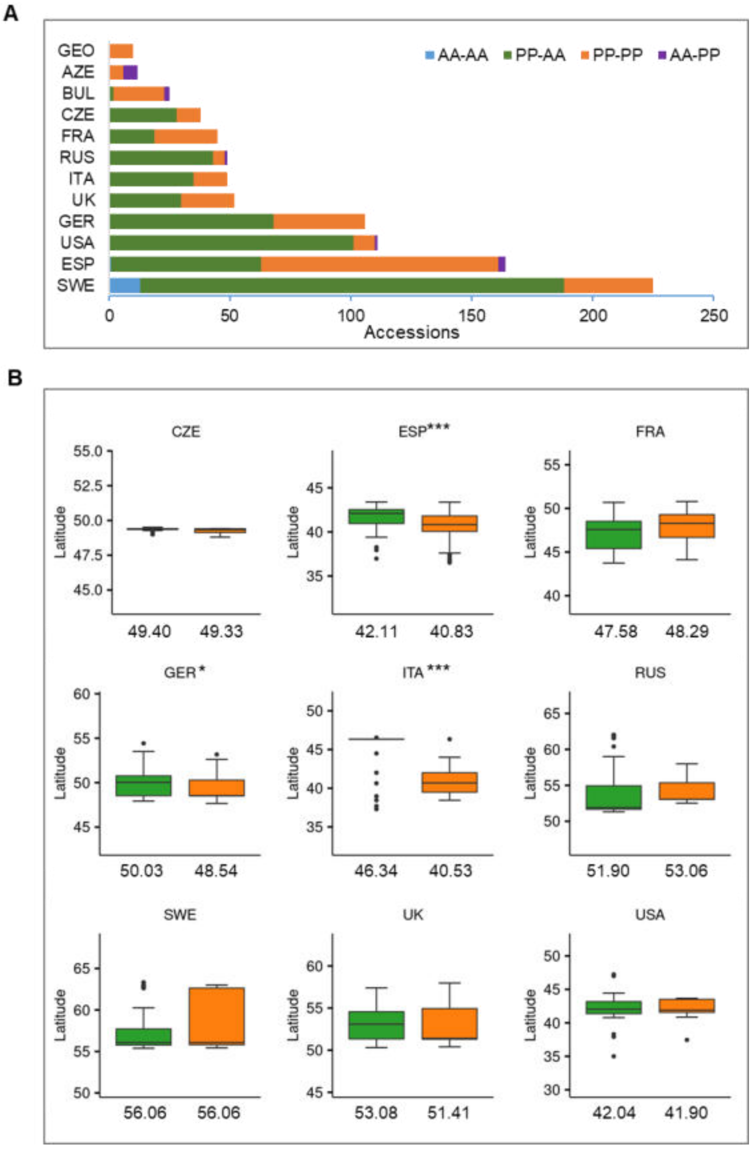


Figure S18. Latitudes of origin among accessions with and without *CYP705A2a-BARS2* genes divided by country. A) Frequency of four groups in individual countries. B) Boxplots presenting collection site latitudes of accessions from PP-AA and PP-PP groups. Only countries with ≥5 accessions within each group are presented. Median values are presented below the boxplots. Boxplots show median and inner quartiles. Whiskers extend to the highest and lowest values no greater than 1.5 times the inner quartile range. Asterisks indicate statistical significance (Wilcoxon rank sum test with continuity correction, *p.value<0.05, ***p.value <0.001).

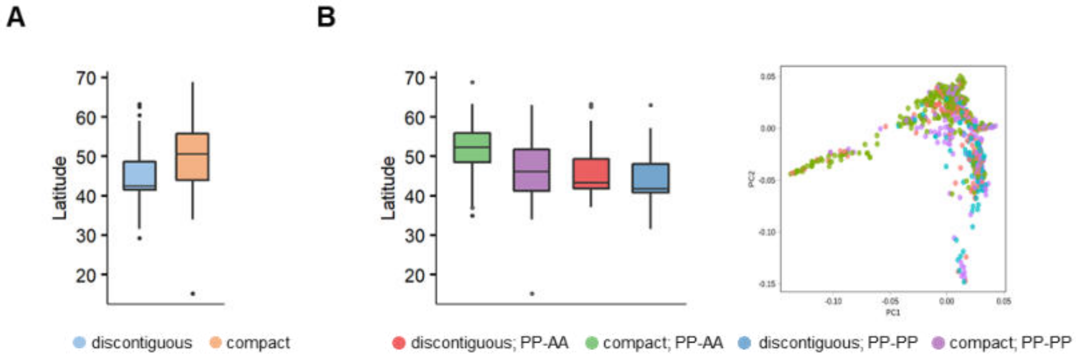


Figure S19. Variability of the arabidiol/baruol gene cluster organization better explains latitudinal distribution of Arabidopsis accessions compared to variability of the thalianol gene cluster. A) Differences in latitudinal distribution of accessions with the compact and discontiguous versions of the thalianol gene cluster. B) PCA on >200k SNPs and LD = 0.3 (left) and latitudes of origin (right) of accessions divided by both thalianol and arabidiol/baruol cluster type.

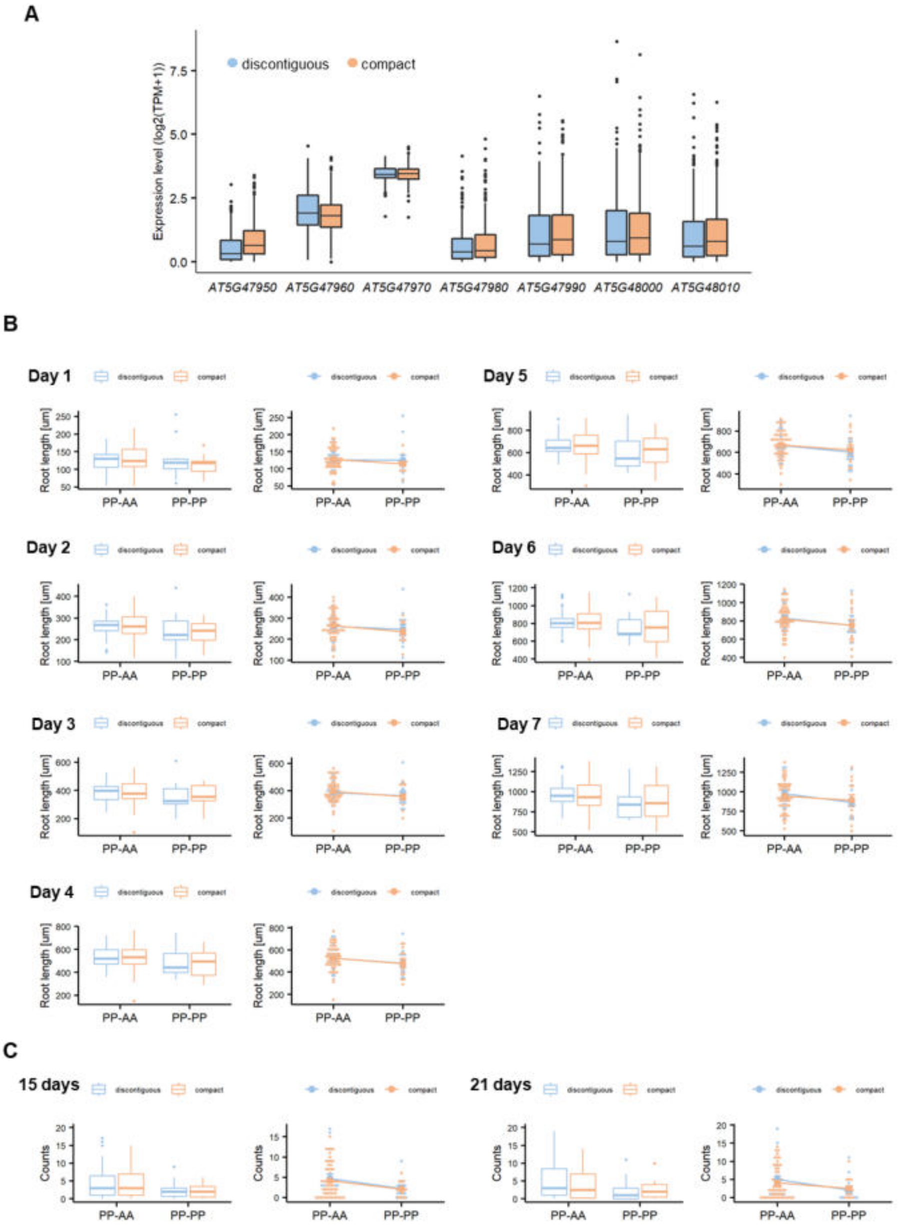


Figure S20. Structural variation of the thalianol gene cluster has little impact on gene expression and root growth phenotypic variation. A) Expression of the thalianol cluster genes in leaves in accessions with discontiguous and compact versions. B, C) Root growth phenotypes presented in the main text in Fig. 5B and 5C, respectively, divided by both thalianol and arabidiol/baruol cluster type (left), along with two-way ANOVA plot (right).

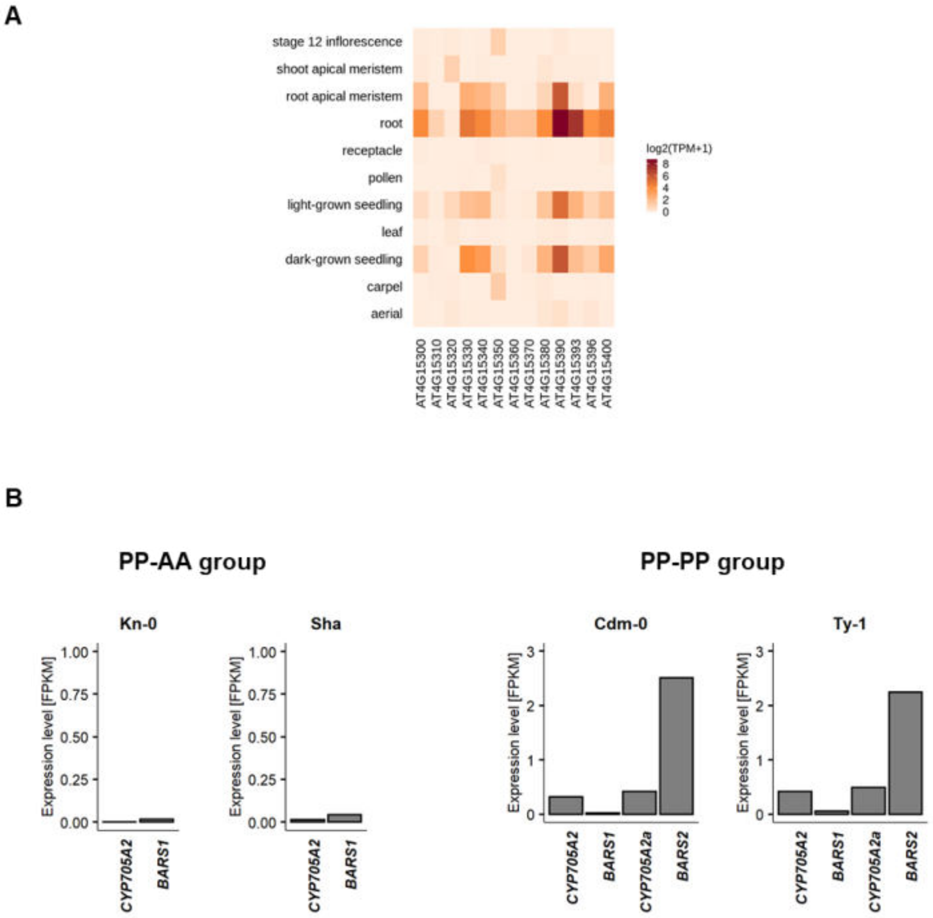


Figure S21. Differences in expression of arabidiol/baruol gene cluster among the accessions. A) Tissue-specific expression in Col-0 accession (PP-AA group). B) Expression of *CYP705A2*, *BARS1*, *CYP705A2a* and *BARS2* genes in accessions from PP-AA and PP-PP groups. For each accession, RNA-Seq data were mapped to the respective genomic assemblies. Where necessary, the region of interest was annotated with Augustus and the gene models were used for FPKM calculations.

Supplemental information

Prediction and analysis of BARS1 and BARS2 3D protein structures

A theoretical 3D model of a plant baruol synthase 1 isoform (NP_193272.1) obtained by the AlphaFold2 algorithm is available (Uniprot ID: O23390). However, an experimental 3D model of any plant OSC is elusive. Thus, we attempted to obtain 3D models of the reference (Col-0) baruol synthase 1 protein (isoforms NP_193272.1, NP_001329547.1) as well as BARS proteins from Cvi-0 encoded by gene duplicates *ATCVI-4G38020* (BARS1) and *ATCVI-4G38110* (BARS2). A comparison of NP_193272.1 models generated with ColabFold software (this study) with that from the UniProt database indicated a very high agreement of their geometrical parameters. Both models superposed with rmsd of ~0.42 Å for 754 Cα atoms, out of overall 759 Cα atoms. Also, the prediction quality for all models we generated with ColabFold was characterized by high pLDDT measures. For the first best predictions, these measures were equal to 91.8 (NP_193272.1), 94.3 (ATCVI-4G38020) and 94.9 (NP_001329547.1 and ATCVI-4G38110). These results indicated that our predictions could be treated with confidence. Therefore, we used 3D protein models obtained with the ColabFold software for further comparative analyses.

Amino acid sequences of Col-0 and Cvi-0 homologs were very similar. Not surprisingly, their structural models superimposed each other very well, with rmsd values from 0.26 Å to 0.42 Å for the Cα atoms (**Supplemental Table S10**). On the other hand, a simple structural comparison of these plant enzymes did not provide any information about their active site. Therefore, we surveyed the Protein Data Bank in search of any OSC experimental structure determined in a complex with its ligand. Using an amino acid sequence of the Col-0 baruol synthase 1 enzyme NP_193272.1, we identified a crystal structure of human OSC (~35% sequence identity) in a complex with lanosterol (ID 1W6K). Superposition of our 3D models of BARS homologs with human OSC revealed their structural similarity, with rmsd values around 1.2 Å for Cα atoms. Moreover, the presence of lanosterol molecule in the active site of the human enzyme allowed us to identify potential substrate-binding cavities in plant homologs (**Supplemental data 1-5**). It is of note that the catalytic aspartate residue D455 present in human cyclase had its counterparts in plant BARS homologs (D493 in Col_BARS1 NP_193272.1, D490 in NP_001329547.1 as well as in ATCVI-4G38020 and ATCVI-4G38110).

Protein sequences of genes predicted with Augustus in de novo genomic assemblies

Ecotype Mir-0

Marneral gene cluster

```
>g1_Mir0 96 aa; Best reference protein match NP_199072.1 AT5G42580 499 aa
MFKPERFLVSSISGDEEKIREQAVKYVTFGGGRRTCPAVKLAHIFMETAIGAMVQCFDWR
IKGEKVYMEEAVSGLSLKMAHPLKCTPVVRFDPFSF
```

Ecotype Mitterberg-2-185

Thalianol gene cluster

```
>g1_Mitterberg 749 aa; Best reference protein match NP_199605.1 AT5G47940 749
aa
MGASNHDNDFNSTTNWKLVDGTLIDAISFESSFTANPESDDGIISAAVDHVTKSPLLLLP
PVPNGEPCEITITFAQEHELRQIYIRSSARVYEVYYTKKRRHDKEYLCTVRCGVAIRDEE
VLQIPLTESADSKPVKDLIERKVTDNGNGRTSEDDWVEVKASDDSLLNNEKQDFYEATAE
INDAEPCTSITVRLLSLQDKRCALVDEVYVFADPVDPSESEKEEATGTGNSSSSSLMAMF
MPALLQLSRGKDVRKERDIQVSDKSNSTDPVAIGNTDQIGVSSPVLVDTVAKQVDAATRV
SGEESKPSISCNNVETIMDQLVKKVSMIETILIRFEDQMLKPINSIDARLQLVEKKLEQL
GNKSFESDLGFRKKIPNQDSLRSDTDKTPDTDESDGLTKNTDVVPDSSSIDNSEDCAVVL
PKNRLDNILSKSVELESENSSISGNEMISAEPEISNEEVGHSFEEKPKYSLSINDALASA
LAGLLSSHSITDGKYSQALVITAPEFSSEDDVEIEQKPGTSAHPDDSQVAAEESENRYSS
SLESSTSSQKEPGITPDDSHGTMYGVFKKLDDSFGGDEEAETVVSVSDNALDEEMVTSST
KADCYTERKNLSYKPTEPDSLIHELESSNVTTAKCKGEPSMDDVLKSVLGFQPTTSSVDF
LTPVLDVKFNLENKDSDSKYFFEVLFTGESKTYLDCKNDVFDDNLVSVEDEEELKGPPTD
TLSSVEMNHYATNEMPIHWNGEISEASLI

>g2_Mitterberg 387 aa; Best reference protein match NP_001078731.1 AT5G47970
387 aa
MTVSEAYSPPLFSIAPMMGWTDNHYRTLARLITKHAWLYTEMLAAETIVYQEDNLDSFLA
FSPDQHPIVLQIGGRNLENLAKATRLANAYAYDEINFNCGCPSPKVSGRGCFGALLMLDP
KFVGEAMSVIAANTNAAVTVKCRIGVDDHDSYNELCDFIHIVSSLSPTKHFIIHSRKALL
SGLSPSDNRRIPPLKYEFFFALLRDFPYLKFTINGGINSVVEADAALRSGAHGVMLGRAV
YYNPWHILGHVDTVIYGSPSSGITRRQVLEKYKVYGESVLGKYGKGRPNLRDIVRPLINL
FHSESGNGQWKRRTDAALLHCTTLQSFLDEVLPAIPDYVLDSSAVKEATGREDLFADVQR
LLPPPYEKESLKALERMPTRPVILDEE

>g3_Mitterberg 223 aa; Best reference protein match NP_199607.1 AT5G47960 223
aa
MSKFQSNFNQKIDYVFKVVLIGDSAVGKSQLLARFSRNEFSIESKATIGVEFQTRTLEID
RKTIKAQIWDTAGQERYRAVTSAYYRGAVGAMLVYDITKRQSFDHVARWLEELRGHADKN
IVIMLIGNKTDLGTLRAVPTEDAKEFAQRENLFFMETSALDSNNVEPSFLTVLTEIYRIV
SKKNLVANEEGESGGDSSLLQGTKIVVAGEETESKGKGCCGTS

>g4_Mitterberg 426 aa; Best reference protein match NP_199606.1 AT5G47950 426
aa
MDTMKVETIGKEIIKPSATTPNDLPTLQLSIMDILMPPVYAVAFLFYTKDDLISQEQTSH
TLKTSLSEILTKFHPLAGRVNGVTIKSTDEGAVFVEARVDNCDLSGFLRSPDTESLKQLL
PVDDEPAPTWPLLVVKATYFQCGGMAIGLCISHRLADAASLSIFLQAWAATARGESDSVA
SPEFCSTKLYPAANEAIKIPGEVVKRTSVTKRFVFVASKIEELRKKVASDVVPRPTRVQS
VTSLIWKCAVTASTDKIREKALFQPANLRTKIPSLLSENQIGNFLFNSLTLDGKAGVDIV
ETVKELQKRAEELSGLVQHEEGSSMTIGSRLFGEIINSKFNFELHDMHSVTSWCKIPLYD
ACFGWGSPVWVAGSVSPDLENVTVLIDSKDGQGIEAWVTLHQDNMLLFEQSTELLAFASP
NPSVLI

>g5_Mitterberg 404 aa; Best reference protein match NP_199609.1 AT5G47980 443
aa
MIFFYNLADLAEKSPDIVSTRLRSSLSQALSRFYPLAGKKEGVSISCNDEGAVFTEARTN
LLLSDFLRNIDINSLKILIPTLAPGESLDSRPLLSVQATFFGSGSGVAVEICVSHCICDA
ASVSTFFRGWAATARGDSNDELSTPQFAEVAIHPPADISIHGSPFNALSEVREKCVTNRF
VFESDKITKLKIVAASKSVPSPTRVEAVMSLIWRCARNASHANLIVPRATMMTQSMDLRL
RIPTNVLSPDAIGNLQGVFFLKRGPGSEIEISEVVAEFRKEKEEFNEMIKENVNGGHTNT
TLGQKIMSGIANYMSELKPNIDTYTMSSWCRKAFYEVDFGWGRPAWVGLGHQDIQDGVMY
VLLVDAKDGEGVEAWVGIPEQDMAAFVCDQELLSYASLNPPVLI
```

```
>g6_Mitterberg 404 aa; Best reference protein match NP_199609.1 AT5G47980 443
aa
MIFFYNLADLAEKSPDIVSTRLRSSLSQALSRFYPLAGKKEGVSISCNDEGAVFTEARTN
LLLSDFLRNIDINSLKILIPTLAPGESLDSRPLLSVQATFFGSGSGVAVEICVSHCICDA
ASVSTFFRGWAATARGDSNDELSTPQFAEVAIHPPADISIHGSPFNALSEVREKCVTNRF
VFESDKITKLKIVAASKSVPSPTRVEAVMSLIWRCARNASHANLIVPRATMMTQSMDLRL
RIPTNVLSPDAIGNLQGVFFLKRGPGSEIEISEVVAEFRKEKEEFNEMIKENVNGGHTNT
TLGQKIMSGIANYMSELKPNIDTYTMSSWCRKAFYEVDFGWGRPAWVGLGHQDIQDGVMY
VLLVDAKDGEGVEAWVGIPEQDMAAFVCDQELLSYASLNPPVLI

>g7_Mitterberg 511 aa; Best reference protein match NP_199610.1 AT5G47990 511
aa
MASMITVDFENCFIFLLLCLFSRLSYDLFFRKTKDLRAGCALPPSPPSLPIIGHLHLILF
VPIHQSFKNISSKYGPLLHLRFFNFPIVLVSSASTAYEIFKAQDVNVSSRPPPPIEESLI
LGSSSFINTPYGDYSKFMKKFMVQKLLGPQALQRSRNIRADELERFYKTLLDKAMKKQTV
EIRNEAMKLTNNTICKMIMGRSCSEENGEAETVRGLVTESIFLTKKHFLGAMFHKPLKKL
GISLFAKELMNVSNRFDELLEKILVEHEEKLQEHHQTSDMLDMLLEAYGDENAEYKITRD
QIKSLFVDLFSAGTEASANTIQWTMAEIIKNPKICERLREEIDSVVGKTRLVQETDLPNL
PYLQAIVKEGLRLHPPGPVVRTFKETCEIKGFYIPEKTRLFVNVYAIMRDPDFWEDPEEF
KPERFLASSRLGEEDEKREDMLKYIPFGSGRRACPGSHLAYTVVGSVIGMMVQHFDWIIK
GEKINMKEGGTMTLTMAHPLKCTPVPRNLNT

>g8_Mitterberg 471 aa; Best reference protein match NP_001032030.1 AT5G48000
477 aa
MSFVWSAAVWVIAVAAVVISKWLYRWSNPKCNGKLPPGSMGLPIIGETCDFFEPHGLYEI
SPFVKKRMLKYGPLFRTNIFGSNTVVLTEPDIIFEVFRQENKSFVFSYPEAFVKPFGKEN
VFLKHGNIHKHVKQISLQHLGSEALKKKMIGEIDRVTYEHLRSKANQGSFDAKEAVESVI
MAHLTPKIISNLKPETQATLVDNIMALGSEWFQSPLKLTTLISIYKVFIARRDALQVIKD
VFTRRKASREMCGDFLDTMVEEGEKEDVIFNEESAINLIFAILVVAKESTSSVTSLAIKF
LAENHKALAELKREHAAILQNRNGKGAGVSWEEYRHQMTFTNMKGCERCRNQRQVSLKHG
STRSYTIPAGWIVAVIPPAVHFNDAIYENPLEFNPWRWEGKELRSGSKTFMVFGGGVRQC
VGAEFARLQISIFIHHLVTTYDFSLAQESEFIRAPLPYFPKGLPIKISQSL

>g9_Mitterberg 764 aa; Best reference protein match NP_001078733.1 AT5G48010
766 aa
MWRLRTGPKAGEDTHLFTTNNYAGRQIWEFDANAGSPQEIAEVEDARHKFSDNTSRFKTT
ADLLWRMQFLREKKFEQKIPRVIIEDARKIKYEDAKKALKRGLLYFTALQADDGHWPAEN
SGPNFYTPPFLICLYITGHLEKIFTPEHVKELLRHIYNMQNEDGGWGLHVESHSVMFCTV
INYVCLRIVGEEVGHDDQRNGCAKAHKWIMDHGGATYTPLIGKALLSVLGVYDWSGCNPI
PPEFWLLPSSFPVNGGTLWIYLRDTFMGLSYLYGKKFVATPTPLILQLREELYPEPYAKI
NWTQTRNRCGKEDLYYPRSFLQDLFWKSVHMFSESILDRWPLNKLIRQRALQSTMALIHY
HDESTRYITGGCLPKAFHMLACWIEDPKSDYFKKHLARVREYIWIGEDGLKIQSFGSQLW
DTALSLHALLDGIDDHDVDDEIKTTLVKGYDYLKKSQITENPRGDHFKMFRHKTKGGWTF
SDQDQGWPVSDCTAESLECCLFFESMPSELIEKKMDVEKLYDAVDYLLYLQSDNGGIAAW
QPVEGKAWLEWLSPVEFLEDTIVEYVECTGSAIAALTQFNKQFPGYKNVEVKRFITKAAK
YIEDMQTVDGSWYGNWGVCFIYGTFFAVRGLVAAGKTYSNCEAIRKAVRFLLDTQNTEGG
WGESFLSCPSKKYTPLKGNSTNVVQTAQALMVLIMGDQMERDPLPVHRAAQVLINSQLDN
GDFPQQEIMGTFMRTVMLHFPTYRNTFSLWALTHYTHALRRLLP

>g10_Mitterberg 355 aa; Best reference protein match NP_568689.1; AT5G48020
355 aa
MELPVVDLSRYLDFSGDELGSDLLESCRQVSRILKETGALIVKDPRCCAQDNDRFIDMME
NYFEKPDDFKRLQQRPNLHYQVGATPEGVEVPRSLVDEEMQEKFKTMPNEYKPHIPKGPD
HKWRYMWRVGPRPSNTRFKELNSEPVIPEGFPEWEEVMDSWGFKMISAVEVVAEMAAIGF
GLPKDAFTSLMKQGPHLLAPTGSDLNCYNEEGTIFAGYHYDLNFLTIHGRSRFPGLYIWL
RNGEKVAVKVPVGCLLIQAGKQIEWLTAGECIAGMHEVVVTSKTKDAITLAKEQNRSLWR
VSSTLFAHIASDAELKPLGHFAESSLASKYPAIPAGEYVEQELSVINLKGNKGFS
```

Ecotype Dolna-1-40

Tirucalladienol gene cluster

```
>g1_Dolna 318 aa; Best reference protein match NP_198463.1 AT5G36140 318 aa
MYLTIIFLFISSIIFPLLFFLGKHLSNFRYPNLPPGKIGFPLIGETLSFLSAGRQGHPEK
FVTDRVRHFSSGIFKTHLFGSPFAVVTGASGNKFLFTNENKLVISWWPDSVNKIFPSSTQ
TSSKEEAIKTRMLLMPSMKPEALRRYVGVMDEIAQKHFETEWANQDQLIVFPLTKKFTFS
IACRLFLSMDDLERVRKLEEPFTTVMTGVFSIPIDLPGTRFNRAIKASRLLSKEVSTIIR
QRKEELKAGKVSVEQDILSHMLMNIGETKDEDLADKIIALLIGGHDTTSIVCTFVVNYLA
EFPHIYQRVLEGMQIPLL
```

Ecotype Cvi-0

Arabidiol/baruol gene cluster

```
>g1_Cvi 513 aa; Best reference protein match NP_193270.1 AT4G15350 509 aa;
Augustus predicted gene (chr4:8591612-8593232 complement)
MAAMIFILLCLFTFLCYSLFYKKPKDSRANCDRPPSPPSLPIIGHLHLILSNLAHKSFQR
LSSKYGPLLHLRIFHIPIVLVSSASVAYDIFRAQDVNVSFRSTSTFEECLFFGTSGFFQA
PYGDYWKFMRKLMVTKLLGPQALERSRNIRVEEIDRLYKNLLNKAMKKESVEIGEEASKL
SNNVICTMIMGRSCSEDNGEAERMRSLVSEAMALTKKFFLANIFHKPLKMLGISLFEKEI
MSVSHKFDELLEKILVEHEEKMEEHHQGTDMMDVLLEAYRDENATYKITRNQIKSLIVEL
LIAGTDTSATTTQWIMAELINHPKVFERVREEIDLVVGRSRLIQETDLPNLAYLQAVVKE
ALRLHPPGPLVPRTLQESCEIKGYYIPEKTIVIVNSYAVMRDPYVWEDPEEFKPERFLDI
SSSVQEEEISDKILKFIPFASGRRGCPGTNLAYINVETAIGVMVQCFDWIIKGKEVNMSE
AAGTMVLTLAEPLMCTPVARTLNPLPASLRAYS
```

Ecotype Eri-1

Arabidiol/baruol gene cluster

```
>g1_Eri 763 aa; Best reference protein match NP_001329547.1 AT4G15370 756 aa;
Augustus predicted gene (chr4:8744883-8749727 complement)
MWRLRIGAKAKDNTHLFTTNNYVGRQIWEFDANAGSPEELAEVEEARRNFSNNRSRFKAS
ADLLWRMQFLREKKFEQKIPRVIVEDAEKITYEDAKTALRRGILYFTALQADDGHWPAEN
AGSIFFNAPFVICLYITGHLEKIFTHEHRVELLRYMYNHQNEDGGWGLHVESPSNMFCSV
INYICLRILGVEAGHDDKGSACARARKWILDHGGATYSPLIGKAWLSVLGVYDWSGCKPI
PPEFWFLPSFFPVNGGTLWIYLRDIFMGLSYLYGKNFVATSTPLILQLREEIYPDPYTNI
SWRQARNRCAKEDLYYPQSFLQDLFWKGVHVFSENILNRWPFNNLIRQRALRTTMELVHY
HDEATRYITGGSVPKVFHMLACWVEDPESDYFKKHLARVPDFIWIGEDGLKIQSFGSQVW
DTALSLHVFIDGFDDDVDEEIRSTLLKGYDYLEKSQVTENPPGDYMKMFRHMAKGGWTFS
DQDQGWPVSDCTAESLECCLFFESMSSEFIGKKMDVEKLYDAVDFLLYLQSDNGGITAWQ
PADGKTWLEWLSPVEFIEDAVVEHEYVECTGSAIVALAQFNKQFPGYKKEEVERFITKGV
KYIEDLQMVDGSWYGNWGVCFIYGTFFAVRGLVAAGKCYNNCEAIRRAVRFILDTQNTEG
GWGESYLSCPRKKYIPLIGNKTNVVNTGQALMVLIMGNQMKRDPLPVHRAAKVLINSQMD
NGDFPQQEIMGVFKMNVMLHFPTYRNMFTLWALTHYTKALRGL
```

```
>g2_Eri 513 aa; Best reference protein match NP_193270.1 AT4G15350 509 aa;
Augustus predicted gene (chr4:8792845-8794465 complement)
MAAMIFILLCLFTFLCYSLFYKKPKDSRANCDRPPSPPSLPIIGHLHLILSNLAHKSFQR
LSSKYGPLLHLRIFHIPIVLVSSASVAYDIFRAQDVNVSFRSTSTFEECLFFGTSGFFQA
PYGDYWKFMRKLMVTKLLGPQALERSRNIRVEEIDRLYKNLLNKAMKKESVEIGEEASKL
SNNVICTMIMGRSCSEDNGEAERMRSLVSEAMALTKKFFLANIFHKPLKMLGISLFEKEI
MSVSHKFDELLEKILVEHEEKMEEHHQGTDMMDVLLEAYRDENATYKITRNQIKSLIVEL
```

```
LIAGTDTSATTTQWIMAELINHPKVFERVREEIDLVVGRSRLIQETDLPNLAYLQAVVKE
ALRLHPPGPLVPRTLQESCEIKGYYIPEKTIVIVNSYAVMRDPYVWEDPEEFKPERFLDI
SSSVQEEEISDKILKFIPFASGRRGCPGTNLAYINVETAIGVMVQCFDWIIKGKEVNMSE
AAGTMVLTLAEPLMCTPVARTLNPLPASLRAYS
```

Ecotype Ler-0

Arabidiol/baruol gene cluster

```
>g1_Ler 763 aa; Best reference protein match NP_001329547.1 AT4G15370 756 aa;
Augustus predicted gene (chr4:9215522-9220364 complement)
MWRLRIGAKAKDNTHLFTTNNYVGRQIWEFDANAGSPEELAEVEEARRNFSNNRSRFKAS
ADLLWRMQFLREKKFEQKIPRVIVEDAEKITYEDAKTALRRGILYFTALQADDGHWPAEN
AGSIFFNAPFVICLYITGHLEKIFTHEHRVELLRYMYNHQNEDGGWGLHVESPSNMFCSV
INYICLRILGVEAGHDDKGSACARARKWILDHGGATYSPLIGKAWLSVLGVYDWSGCKPI
PPEFWFLPSFFPVNGGTLWIYLRDIFMGLSYLYGKNFVATSTPLILQLREEIYPDPYTNI
SWRQARNRCAKEDLYYPQSFLQDLFWKGVHVFSENILNRWPFNNLIRQRALRTTMELVHY
HDEATRYITGGSVPKVFHMLACWVEDPESDYFKKHLARVPDFIWIGEDGLKIQSFGSQVW
DTALSLHVFIDGFDDDVDEEIRSTLLKGYDYLEKSQVTENPPGDYMKMFRHMAKGGWTFS
DQDQGWPVSDCTAESLECCLFFESMSSEFIGKKMDVEKLYDAVDFLLYLQSDNGGITAWQ
PADGKTWLEWLSPVEFIEDAVVEHEYVECTGSAIVALAQFNKQFPGYKKEEVERFITKGV
KYIEDLQMVDGSWYGNWGVCFIYGTFFAVRGLVAAGKCYNNCEAIRRAVRFILDTQNTEG
GWGESYLSCPRKKYIPLIGNKTNVVNTGQALMVLIMGNQMKRDPLPVHRAAKVLINSQMD
NGDFPQQEIMGVFKMNVMLHFPTYRNMFTLWALTHYTKALRGL
```

```
>g2_Ler 513 aa; Best reference protein match NP_193270.1 AT4G15350 509 aa;
Augustus predicted gene (chr4:9263484-9265104 complement)
MAAMIFILLCLFTFLCYSLFYKKPKDSRANCDRPPSPPSLPIIGHLHLILSNLAHKSFQR
LSSKYGPLLHLRIFHIPIVLVSSASVAYDIFRAQDVNVSFRSTSTFEECLFFGTSGFFQA
PYGDYWKFMRKLMVTKLLGPQALERSRNIRVEEIDRLYKNLLNKAMKKESVEIGEEASKL
SNNVICTMIMGRSCSEDNGEAERMRSLVSEAMALTKKFFLANIFHKPLKMLGISLFEKEI
MSVSHKFDELLEKILVEHEEKMEEHHQGTDMMDVLLEAYRDENATYKITRNQIKSLIVEL
LIAGTDTSATTTQWIMAELINHPKVFERVREEIDLVVGRSRLIQETDLPNLAYLQAVVKE
ALRLHPPGPLVPRTLQESCEIKGYYIPEKTIVIVNSYAVMRDPYVWEDPEEFKPERFLDI
SSSVQEEEISDKILKFIPFASGRRGCPGTNLAYINVETAIGVMVQCFDWIIKGKEVNMSE
AAGTMVLTLAEPLMCTPVARTLNPLPASLRAYS
```

Ecotype C24

Arabidiol/baruol gene cluster

```
>g1_C24 78 aa; Best reference protein match NP_001329547.1 AT4G15370 756 aa;
Augustus predicted gene (chr4:9580438-9581743 complement)

MWKLIIGSKAGDDIHLFSTNNYVGRQIWEFDAKAGSPEELAEVEEARQNFTDNRSHFKAS
ADLLWRMQFLREKKFEQKIPRVIIEDAEKITYEDAKTALKRGLLYFTALQADDGHWPAEN
AGSIFFNAPFVICMYITGHLERIFTPEHVRELLRYLYNHQNEDGGWGLHIESPSNMFCTV
INYICLRILGVEAGYDDEGSACARARKWILDHGGATYSPLIGKAWLSVLGVYDWSGCKPI
PPEFWLLPSFLPVNGGLKLEGL
```

Distribution of accessions with varying CYP705A2-BARS1 copy number status

Map available at:
https://www.google.com/maps/d/edit?mid=1ZaAMX-EDYlbBtKbKdBsS06HMjvc&usp=sharing
Last accessed: September 21, 2022.

Annotations used for evaluation of gene expression in individual accessions

Cdm-0

```
LR881469.1        MAN      gene        10564284 10571454 .        -        .        gene_id
"g1_Cdm"; gene_name ""; gene_source "MAN";
LR881469.1        MAN      transcript  10564284 10571454 .        -        .        gene_id
"g1_Cdm"; transcript_id "g1_Cdm.1"; gene_source "MAN";
LR881469.1        MAN      exon        10564284 10564400 .        -        .        gene_id
"g1_Cdm"; transcript_id "g1_Cdm.1"; exon_number "1"; gene_name ""; gene_source "MAN";
LR881469.1        MAN      CDS         10564284 10564400 .        -        0        gene_id
"g1_Cdm"; transcript_id "g1_Cdm.1"; cds_number "1"; gene_name ""; gene_source "MAN";
LR881469.1        MAN      start_codon          10571452 10571454 .        -        0
          gene_id "g1_Cdm"; transcript_id "g1_Cdm.1"; gene_name ""; gene_source "MAN";
LR881469.1        MAN      exon        10564603 10565089 .        -        .        gene_id
"g1_Cdm"; transcript_id "g1_Cdm.1"; exon_number "2"; gene_name ""; gene_source "MAN";
LR881469.1        MAN      CDS         10564603 10565089 .        -        1        gene_id
"g1_Cdm"; transcript_id "g1_Cdm.1"; cds_number "2"; gene_name ""; gene_source "MAN";
LR881469.1        MAN      exon        10565181 10565227 .        -        .        gene_id
"g1_Cdm"; transcript_id "g1_Cdm.1"; exon_number "3"; gene_name ""; gene_source "MAN";
LR881469.1        MAN      CDS         10565181 10565227 .        -        0        gene_id
"g1_Cdm"; transcript_id "g1_Cdm.1"; cds_number "3"; gene_name ""; gene_source "MAN";
LR881469.1        MAN      exon        10565313 10565369 .        -        .        gene_id
"g1_Cdm"; transcript_id "g1_Cdm.1"; exon_number "4"; gene_name ""; gene_source "MAN";
LR881469.1        MAN      CDS         10565313 10565369 .        -        0        gene_id
"g1_Cdm"; transcript_id "g1_Cdm.1"; cds_number "4"; gene_name ""; gene_source "MAN";
LR881469.1        MAN      exon        10567717 10567815 .        -        .        gene_id
"g1_Cdm"; transcript_id "g1_Cdm.1"; exon_number "5"; gene_name ""; gene_source "MAN";
LR881469.1        MAN      CDS         10567717 10567815 .        -        0        gene_id
"g1_Cdm"; transcript_id "g1_Cdm.1"; cds_number "5"; gene_name ""; gene_source "MAN";
LR881469.1        MAN      exon        10567896 10568147 .        -        .        gene_id
"g1_Cdm"; transcript_id "g1_Cdm.1"; exon_number "6"; gene_name ""; gene_source "MAN";
LR881469.1        MAN      CDS         10567896 10568147 .        -        0        gene_id
"g1_Cdm"; transcript_id "g1_Cdm.1"; cds_number "6"; gene_name ""; gene_source "MAN";
LR881469.1        MAN      exon        10568297 10568410 .        -        .        gene_id
"g1_Cdm"; transcript_id "g1_Cdm.1"; exon_number "7"; gene_name ""; gene_source "MAN";
LR881469.1        MAN      CDS         10568297 10568410 .        -        0        gene_id
"g1_Cdm"; transcript_id "g1_Cdm.1"; cds_number "7"; gene_name ""; gene_source "MAN";
LR881469.1        MAN      exon        10568488 10568630 .        -        .        gene_id
"g1_Cdm"; transcript_id "g1_Cdm.1"; exon_number "8"; gene_name ""; gene_source "MAN";
LR881469.1        MAN      CDS         10568488 10568630 .        -        2        gene_id
"g1_Cdm"; transcript_id "g1_Cdm.1"; cds_number "8"; gene_name ""; gene_source "MAN";
LR881469.1        MAN      exon        10569276 10569449 .        -        .        gene_id
"g1_Cdm"; transcript_id "g1_Cdm.1"; exon_number "9"; gene_name ""; gene_source "MAN";
LR881469.1        MAN      CDS         10569276 10569449 .        -        2        gene_id
"g1_Cdm"; transcript_id "g1_Cdm.1"; cds_number "9"; gene_name ""; gene_source "MAN";
LR881469.1        MAN      exon        10570301 10570385 .        -        .        gene_id
"g1_Cdm"; transcript_id "g1_Cdm.1"; exon_number "10"; gene_name ""; gene_source "MAN";
LR881469.1        MAN      CDS         10570301 10570385 .        -        0        gene_id
"g1_Cdm"; transcript_id "g1_Cdm.1"; cds_number "10"; gene_name ""; gene_source "MAN";
LR881469.1        MAN      exon        10570464 10570667 .        -        .        gene_id
"g1_Cdm"; transcript_id "g1_Cdm.1"; exon_number "11"; gene_name ""; gene_source "MAN";
LR881469.1        MAN      CDS         10570464 10570667 .        -        0        gene_id
"g1_Cdm"; transcript_id "g1_Cdm.1"; cds_number "11"; gene_name ""; gene_source "MAN";
```

```
LR881469.1      MAN     exon    10570761 10570850 .     -       .       gene_id
"g1_Cdm"; transcript_id "g1_Cdm.1"; exon_number "12"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     CDS     10570761 10570850 .     -       0       gene_id
"g1_Cdm"; transcript_id "g1_Cdm.1"; cds_number "12"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     exon    10570951 10571136 .     -       .       gene_id
"g1_Cdm"; transcript_id "g1_Cdm.1"; exon_number "13"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     CDS     10570951 10571136 .     -       0       gene_id
"g1_Cdm"; transcript_id "g1_Cdm.1"; cds_number "13"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     exon    10571251 10571454 .     -       .       gene_id
"g1_Cdm"; transcript_id "g1_Cdm.1"; exon_number "14"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     CDS     10571251 10571454 .     -       0       gene_id
"g1_Cdm"; transcript_id "g1_Cdm.1"; cds_number "14"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     stop_codon      10564284 10564286 .     -       0
        gene_id "g1_Cdm"; transcript_id "g1_Cdm.1"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     gene    10573610 10575227 .     +       .       gene_id
"g2_Cdm"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     transcript  10573610 10575227 .     +       .       gene_id
"g2_Cdm"; transcript_id "g2_Cdm.1"; gene_source "MAN";
LR881469.1      MAN     exon    10573610 10574509 .     +       .       gene_id
"g2_Cdm"; transcript_id "g2_Cdm.1"; exon_number "1"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     CDS     10573610 10574509 .     +       0       gene_id
"g2_Cdm"; transcript_id "g2_Cdm.1"; cds_number "1"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     start_codon     10573610 10573612 .     +       0
        gene_id "g2_Cdm"; transcript_id "g2_Cdm.1"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     exon    10574598 10575227 .     +       .       gene_id
"g2_Cdm"; transcript_id "g2_Cdm.1"; exon_number "2"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     CDS     10574598 10575227 .     +       0       gene_id
"g2_Cdm"; transcript_id "g2_Cdm.1"; cds_number "2"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     stop_codon      10575225 10575227 .     +       0
        gene_id "g2_Cdm"; transcript_id "g2_Cdm.1"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     gene    10583720 10588582 .     -       .       gene_id
"g3_Cdm"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     transcript  10583720 10588582 .     -       .       gene_id
"g3_Cdm"; transcript_id "g3_Cdm.1"; gene_source "MAN";
LR881469.1      MAN     exon    10583720 10583830 .     -       .       gene_id
"g3_Cdm"; transcript_id "g3_Cdm.1"; exon_number "1"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     CDS     10583720 10583830 .     -       0       gene_id
"g3_Cdm"; transcript_id "g3_Cdm.1"; cds_number "1"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     start_codon     10588580 10588582 .     -       0
        gene_id "g3_Cdm"; transcript_id "g3_Cdm.1"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     exon    10583994 10584480 .     -       .       gene_id
"g3_Cdm"; transcript_id "g3_Cdm.1"; exon_number "2"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     CDS     10583994 10584480 .     -       1       gene_id
"g3_Cdm"; transcript_id "g3_Cdm.1"; cds_number "2"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     exon    10584581 10584627 .     -       .       gene_id
"g3_Cdm"; transcript_id "g3_Cdm.1"; exon_number "3"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     CDS     10584581 10584627 .     -       0       gene_id
"g3_Cdm"; transcript_id "g3_Cdm.1"; cds_number "3"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     exon    10584702 10584758 .     -       .       gene_id
"g3_Cdm"; transcript_id "g3_Cdm.1"; exon_number "4"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     CDS     10584702 10584758 .     -       0       gene_id
"g3_Cdm"; transcript_id "g3_Cdm.1"; cds_number "4"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     exon    10585566 10585664 .     -       .       gene_id
"g3_Cdm"; transcript_id "g3_Cdm.1"; exon_number "5"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     CDS     10585566 10585664 .     -       0       gene_id
"g3_Cdm"; transcript_id "g3_Cdm.1"; cds_number "5"; gene_name ""; gene_source "MAN";
```

```
LR881469.1      MAN     exon    10585760 10586011 .     -       .       gene_id
"g3_Cdm"; transcript_id "g3_Cdm.1"; exon_number "6"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     CDS     10585760 10586011 .     -       0       gene_id
"g3_Cdm"; transcript_id "g3_Cdm.1"; cds_number "6"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     exon    10586133 10586246 .     -       .       gene_id
"g3_Cdm"; transcript_id "g3_Cdm.1"; exon_number "7"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     CDS     10586133 10586246 .     -       0       gene_id
"g3_Cdm"; transcript_id "g3_Cdm.1"; cds_number "7"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     exon    10586345 10586469 .     -       .       gene_id
"g3_Cdm"; transcript_id "g3_Cdm.1"; exon_number "8"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     CDS     10586345 10586469 .     -       2       gene_id
"g3_Cdm"; transcript_id "g3_Cdm.1"; cds_number "8"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     exon    10587041 10587125 .     -       .       gene_id
"g3_Cdm"; transcript_id "g3_Cdm.1"; exon_number "9"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     CDS     10587041 10587125 .     -       0       gene_id
"g3_Cdm"; transcript_id "g3_Cdm.1"; cds_number "9"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     exon    10587336 10587536 .     -       .       gene_id
"g3_Cdm"; transcript_id "g3_Cdm.1"; exon_number "10"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     CDS     10587336 10587536 .     -       0       gene_id
"g3_Cdm"; transcript_id "g3_Cdm.1"; cds_number "10"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     exon    10587871 10587960 .     -       .       gene_id
"g3_Cdm"; transcript_id "g3_Cdm.1"; exon_number "11"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     CDS     10587871 10587960 .     -       0       gene_id
"g3_Cdm"; transcript_id "g3_Cdm.1"; cds_number "11"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     exon    10588095 10588280 .     -       .       gene_id
"g3_Cdm"; transcript_id "g3_Cdm.1"; exon_number "12"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     CDS     10588095 10588280 .     -       0       gene_id
"g3_Cdm"; transcript_id "g3_Cdm.1"; cds_number "12"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     exon    10588379 10588582 .     -       .       gene_id
"g3_Cdm"; transcript_id "g3_Cdm.1"; exon_number "13"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     CDS     10588379 10588582 .     -       0       gene_id
"g3_Cdm"; transcript_id "g3_Cdm.1"; cds_number "13"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     stop_codon      10583720 10583722 .     -       0
        gene_id "g3_Cdm"; transcript_id "g3_Cdm.1"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     gene    10630272 10631892 .     -       .       gene_id
"g4_Cdm"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     transcript  10630272 10631892 .     -       .       gene_id
"g4_Cdm"; transcript_id "g4_Cdm.1"; gene_source "MAN";
LR881469.1      MAN     exon    10630272 10630919 .     -       .       gene_id
"g4_Cdm"; transcript_id "g4_Cdm.1"; exon_number "1"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     CDS     10630272 10630919 .     -       0       gene_id
"g4_Cdm"; transcript_id "g4_Cdm.1"; cds_number "1"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     start_codon     10631890 10631892 .     -       0
        gene_id "g4_Cdm"; transcript_id "g4_Cdm.1"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     exon    10630999 10631892 .     -       .       gene_id
"g4_Cdm"; transcript_id "g4_Cdm.1"; exon_number "2"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     CDS     10630999 10631892 .     -       0       gene_id
"g4_Cdm"; transcript_id "g4_Cdm.1"; cds_number "2"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     stop_codon      10630272 10630274 .     -       0
        gene_id "g4_Cdm"; transcript_id "g4_Cdm.1"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     gene    10634775 10638900 .     -       .       gene_id
"g5_Cdm"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     transcript  10634775 10638900 .     -       .       gene_id
"g5_Cdm"; transcript_id "g5_Cdm.1"; gene_source "MAN";
LR881469.1      MAN     exon    10634775 10634891 .     -       .       gene_id
"g5_Cdm"; transcript_id "g5_Cdm.1"; exon_number "1"; gene_name ""; gene_source "MAN";
```

```
LR881469.1      MAN     CDS     10634775 10634891 .     -       0       gene_id
"g5_Cdm"; transcript_id "g5_Cdm.1"; cds_number "1"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     start_codon     10638898 10638900 .     -       0
        gene_id "g5_Cdm"; transcript_id "g5_Cdm.1"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     exon    10635022 10635508 .     -       .       gene_id
"g5_Cdm"; transcript_id "g5_Cdm.1"; exon_number "2"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     CDS     10635022 10635508 .     -       1       gene_id
"g5_Cdm"; transcript_id "g5_Cdm.1"; cds_number "2"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     exon    10635605 10635651 .     -       .       gene_id
"g5_Cdm"; transcript_id "g5_Cdm.1"; exon_number "3"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     CDS     10635605 10635651 .     -       0       gene_id
"g5_Cdm"; transcript_id "g5_Cdm.1"; cds_number "3"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     exon    10635727 10635783 .     -       .       gene_id
"g5_Cdm"; transcript_id "g5_Cdm.1"; exon_number "4"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     CDS     10635727 10635783 .     -       0       gene_id
"g5_Cdm"; transcript_id "g5_Cdm.1"; cds_number "4"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     exon    10636211 10636309 .     -       .       gene_id
"g5_Cdm"; transcript_id "g5_Cdm.1"; exon_number "5"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     CDS     10636211 10636309 .     -       0       gene_id
"g5_Cdm"; transcript_id "g5_Cdm.1"; cds_number "5"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     exon    10636397 10636648 .     -       .       gene_id
"g5_Cdm"; transcript_id "g5_Cdm.1"; exon_number "6"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     CDS     10636397 10636648 .     -       0       gene_id
"g5_Cdm"; transcript_id "g5_Cdm.1"; cds_number "6"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     exon    10636725 10636838 .     -       .       gene_id
"g5_Cdm"; transcript_id "g5_Cdm.1"; exon_number "7"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     CDS     10636725 10636838 .     -       0       gene_id
"g5_Cdm"; transcript_id "g5_Cdm.1"; cds_number "7"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     exon    10636946 10637137 .     -       .       gene_id
"g5_Cdm"; transcript_id "g5_Cdm.1"; exon_number "8"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     CDS     10636946 10637137 .     -       0       gene_id
"g5_Cdm"; transcript_id "g5_Cdm.1"; cds_number "8"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     exon    10637355 10637521 .     -       .       gene_id
"g5_Cdm"; transcript_id "g5_Cdm.1"; exon_number "9"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     CDS     10637355 10637521 .     -       2       gene_id
"g5_Cdm"; transcript_id "g5_Cdm.1"; cds_number "9"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     exon    10637618 10637702 .     -       .       gene_id
"g5_Cdm"; transcript_id "g5_Cdm.1"; exon_number "10"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     CDS     10637618 10637702 .     -       0       gene_id
"g5_Cdm"; transcript_id "g5_Cdm.1"; cds_number "10"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     exon    10637929 10638129 .     -       .       gene_id
"g5_Cdm"; transcript_id "g5_Cdm.1"; exon_number "11"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     CDS     10637929 10638129 .     -       0       gene_id
"g5_Cdm"; transcript_id "g5_Cdm.1"; cds_number "11"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     exon    10638216 10638305 .     -       .       gene_id
"g5_Cdm"; transcript_id "g5_Cdm.1"; exon_number "12"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     CDS     10638216 10638305 .     -       0       gene_id
"g5_Cdm"; transcript_id "g5_Cdm.1"; cds_number "12"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     exon    10638407 10638592 .     -       .       gene_id
"g5_Cdm"; transcript_id "g5_Cdm.1"; exon_number "13"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     CDS     10638407 10638592 .     -       0       gene_id
"g5_Cdm"; transcript_id "g5_Cdm.1"; cds_number "13"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     exon    10638697 10638900 .     -       .       gene_id
"g5_Cdm"; transcript_id "g5_Cdm.1"; exon_number "14"; gene_name ""; gene_source "MAN";
LR881469.1      MAN     CDS     10638697 10638900 .     -       0       gene_id
"g5_Cdm"; transcript_id "g5_Cdm.1"; cds_number "14"; gene_name ""; gene_source "MAN";
```

```
LR881469.1	MAN	stop_codon	10634775 10634777 .	-	0
	gene_id "g5_Cdm"; transcript_id "g5_Cdm.1"; gene_name ""; gene_source "MAN";
```

Ty-1

```
LR797800.1	MAN	gene	9989227	9995059	.	-	.	gene_id
"g1_Ty1"; gene_name ""; gene_source "MAN";
LR797800.1	MAN	transcript	9989227	9995059	.	-	.	gene_id
"g1_Ty1"; transcript_id "g1_Ty1.1"; gene_source "MAN";
LR797800.1	MAN	exon	9989227	9989343	.	-	.	gene_id
"g1_Ty1"; transcript_id "g1_Ty1.1"; exon_number "1"; gene_name ""; gene_source "MAN";
LR797800.1	MAN	CDS	9989227	9989343	.	-	0	gene_id
"g1_Ty1"; transcript_id "g1_Ty1.1"; cds_number "1"; gene_name ""; gene_source "MAN";
LR797800.1	MAN	start_codon	9995057	9995059	.	-	0
	gene_id "g1_Ty1"; transcript_id "g1_Ty1.1"; gene_name ""; gene_source "MAN";
LR797800.1	MAN	exon	9989546	9990032	.	-	.	gene_id
"g1_Ty1"; transcript_id "g1_Ty1.1"; exon_number "2"; gene_name ""; gene_source "MAN";
LR797800.1	MAN	CDS	9989546	9990032	.	-	1	gene_id
"g1_Ty1"; transcript_id "g1_Ty1.1"; cds_number "2"; gene_name ""; gene_source "MAN";
LR797800.1	MAN	exon	9990131	9990177	.	-	.	gene_id
"g1_Ty1"; transcript_id "g1_Ty1.1"; exon_number "3"; gene_name ""; gene_source "MAN";
LR797800.1	MAN	CDS	9990131	9990177	.	-	0	gene_id
"g1_Ty1"; transcript_id "g1_Ty1.1"; cds_number "3"; gene_name ""; gene_source "MAN";
LR797800.1	MAN	exon	9990263	9990319	.	-	.	gene_id
"g1_Ty1"; transcript_id "g1_Ty1.1"; exon_number "4"; gene_name ""; gene_source "MAN";
LR797800.1	MAN	CDS	9990263	9990319	.	-	0	gene_id
"g1_Ty1"; transcript_id "g1_Ty1.1"; cds_number "4"; gene_name ""; gene_source "MAN";
LR797800.1	MAN	exon	9991370	9991468	.	-	.	gene_id
"g1_Ty1"; transcript_id "g1_Ty1.1"; exon_number "5"; gene_name ""; gene_source "MAN";
LR797800.1	MAN	CDS	9991370	9991468	.	-	0	gene_id
"g1_Ty1"; transcript_id "g1_Ty1.1"; cds_number "5"; gene_name ""; gene_source "MAN";
LR797800.1	MAN	exon	9991550	9991801	.	-	.	gene_id
"g1_Ty1"; transcript_id "g1_Ty1.1"; exon_number "6"; gene_name ""; gene_source "MAN";
LR797800.1	MAN	CDS	9991550	9991801	.	-	0	gene_id
"g1_Ty1"; transcript_id "g1_Ty1.1"; cds_number "6"; gene_name ""; gene_source "MAN";
LR797800.1	MAN	exon	9991951	9992064	.	-	.	gene_id
"g1_Ty1"; transcript_id "g1_Ty1.1"; exon_number "7"; gene_name ""; gene_source "MAN";
LR797800.1	MAN	CDS	9991951	9992064	.	-	0	gene_id
"g1_Ty1"; transcript_id "g1_Ty1.1"; cds_number "7"; gene_name ""; gene_source "MAN";
LR797800.1	MAN	exon	9992133	9992275	.	-	.	gene_id
"g1_Ty1"; transcript_id "g1_Ty1.1"; exon_number "8"; gene_name ""; gene_source "MAN";
LR797800.1	MAN	CDS	9992133	9992275	.	-	2	gene_id
"g1_Ty1"; transcript_id "g1_Ty1.1"; cds_number "8"; gene_name ""; gene_source "MAN";
LR797800.1	MAN	exon	9992933	9993106	.	-	.	gene_id
"g1_Ty1"; transcript_id "g1_Ty1.1"; exon_number "9"; gene_name ""; gene_source "MAN";
LR797800.1	MAN	CDS	9992933	9993106	.	-	2	gene_id
"g1_Ty1"; transcript_id "g1_Ty1.1"; cds_number "9"; gene_name ""; gene_source "MAN";
LR797800.1	MAN	exon	9993909	9993993	.	-	.	gene_id
"g1_Ty1"; transcript_id "g1_Ty1.1"; exon_number "10"; gene_name ""; gene_source "MAN";
LR797800.1	MAN	CDS	9993909	9993993	.	-	0	gene_id
"g1_Ty1"; transcript_id "g1_Ty1.1"; cds_number "10"; gene_name ""; gene_source "MAN";
LR797800.1	MAN	exon	9994072	9994275	.	-	.	gene_id
"g1_Ty1"; transcript_id "g1_Ty1.1"; exon_number "11"; gene_name ""; gene_source "MAN";
LR797800.1	MAN	CDS	9994072	9994275	.	-	0	gene_id
"g1_Ty1"; transcript_id "g1_Ty1.1"; cds_number "11"; gene_name ""; gene_source "MAN";
```

```
LR797800.1      MAN     exon        9994365  9994454  .     -     .     gene_id
"g1_Ty1"; transcript_id "g1_Ty1.1"; exon_number "12"; gene_name ""; gene_source "MAN";
LR797800.1      MAN     CDS         9994365  9994454  .     -     0     gene_id
"g1_Ty1"; transcript_id "g1_Ty1.1"; cds_number "12"; gene_name ""; gene_source "MAN";
LR797800.1      MAN     exon        9994556  9994741  .     -     .     gene_id
"g1_Ty1"; transcript_id "g1_Ty1.1"; exon_number "13"; gene_name ""; gene_source "MAN";
LR797800.1      MAN     CDS         9994556  9994741  .     -     0     gene_id
"g1_Ty1"; transcript_id "g1_Ty1.1"; cds_number "13"; gene_name ""; gene_source "MAN";
LR797800.1      MAN     exon        9994856  9995059  .     -     .     gene_id
"g1_Ty1"; transcript_id "g1_Ty1.1"; exon_number "14"; gene_name ""; gene_source "MAN";
LR797800.1      MAN     CDS         9994856  9995059  .     -     0     gene_id
"g1_Ty1"; transcript_id "g1_Ty1.1"; cds_number "14"; gene_name ""; gene_source "MAN";
LR797800.1      MAN     stop_codon           9989227  9989229  .     -     0
        gene_id "g1_Ty1"; transcript_id "g1_Ty1.1"; gene_name ""; gene_source "MAN";
LR797800.1      MAN     gene        9997423  9999033  .     +     .     gene_id
"g2_Ty1"; gene_name ""; gene_source "MAN";
LR797800.1      MAN     transcript  9997423  9999033  .     +     .     gene_id
"g2_Ty1"; transcript_id "g2_Ty1.1"; gene_source "MAN";
LR797800.1      MAN     exon        9997423  9998322  .     +     .     gene_id
"g2_Ty1"; transcript_id "g2_Ty1.1"; exon_number "1"; gene_name ""; gene_source "MAN";
LR797800.1      MAN     CDS         9997423  9998322  .     +     0     gene_id
"g2_Ty1"; transcript_id "g2_Ty1.1"; cds_number "1"; gene_name ""; gene_source "MAN";
LR797800.1      MAN     start_codon          9997423  9997425  .     +     0
        gene_id "g2_Ty1"; transcript_id "g2_Ty1.1"; gene_name ""; gene_source "MAN";
LR797800.1      MAN     exon        9998404  9999033  .     +     .     gene_id
"g2_Ty1"; transcript_id "g2_Ty1.1"; exon_number "2"; gene_name ""; gene_source "MAN";
LR797800.1      MAN     CDS         9998404  9999033  .     +     0     gene_id
"g2_Ty1"; transcript_id "g2_Ty1.1"; cds_number "2"; gene_name ""; gene_source "MAN";
LR797800.1      MAN     stop_codon           9999031  9999033  .     +     0
        gene_id "g2_Ty1"; transcript_id "g2_Ty1.1"; gene_name ""; gene_source "MAN";
LR797800.1      MAN     gene        10008336 10014219 .     -     .     gene_id
"g3_Ty1"; gene_name ""; gene_source "MAN";
LR797800.1      MAN     transcript  10008336 10014219 .     -     .     gene_id
"g3_Ty1"; transcript_id "g3_Ty1.1"; gene_source "MAN";
LR797800.1      MAN     exon        10008336 10008446 .     -     .     gene_id
"g3_Ty1"; transcript_id "g3_Ty1.1"; exon_number "1"; gene_name ""; gene_source "MAN";
LR797800.1      MAN     CDS         10008336 10008446 .     -     0     gene_id
"g3_Ty1"; transcript_id "g3_Ty1.1"; cds_number "1"; gene_name ""; gene_source "MAN";
LR797800.1      MAN     start_codon          10014217 10014219 .     -     0
        gene_id "g3_Ty1"; transcript_id "g3_Ty1.1"; gene_name ""; gene_source "MAN";
LR797800.1      MAN     exon        10008610 10009096 .     -     .     gene_id
"g3_Ty1"; transcript_id "g3_Ty1.1"; exon_number "2"; gene_name ""; gene_source "MAN";
LR797800.1      MAN     CDS         10008610 10009096 .     -     1     gene_id
"g3_Ty1"; transcript_id "g3_Ty1.1"; cds_number "2"; gene_name ""; gene_source "MAN";
LR797800.1      MAN     exon        10009197 10009243 .     -     .     gene_id
"g3_Ty1"; transcript_id "g3_Ty1.1"; exon_number "3"; gene_name ""; gene_source "MAN";
LR797800.1      MAN     CDS         10009197 10009243 .     -     0     gene_id
"g3_Ty1"; transcript_id "g3_Ty1.1"; cds_number "3"; gene_name ""; gene_source "MAN";
LR797800.1      MAN     exon        10009318 10009374 .     -     .     gene_id
"g3_Ty1"; transcript_id "g3_Ty1.1"; exon_number "4"; gene_name ""; gene_source "MAN";
LR797800.1      MAN     CDS         10009318 10009374 .     -     0     gene_id
"g3_Ty1"; transcript_id "g3_Ty1.1"; cds_number "4"; gene_name ""; gene_source "MAN";
LR797800.1      MAN     exon        10010182 10010280 .     -     .     gene_id
"g3_Ty1"; transcript_id "g3_Ty1.1"; exon_number "5"; gene_name ""; gene_source "MAN";
LR797800.1      MAN     CDS         10010182 10010280 .     -     0     gene_id
"g3_Ty1"; transcript_id "g3_Ty1.1"; cds_number "5"; gene_name ""; gene_source "MAN";
```

```
LR797800.1      MAN     exon        10010376 10010627 .      -      .       gene_id
"g3_Ty1"; transcript_id "g3_Ty1.1"; exon_number "6"; gene_name ""; gene_source "MAN";
LR797800.1      MAN     CDS         10010376 10010627 .      -      0       gene_id
"g3_Ty1"; transcript_id "g3_Ty1.1"; cds_number "6"; gene_name ""; gene_source "MAN";
LR797800.1      MAN     exon        10010749 10010862 .      -      .       gene_id
"g3_Ty1"; transcript_id "g3_Ty1.1"; exon_number "7"; gene_name ""; gene_source "MAN";
LR797800.1      MAN     CDS         10010749 10010862 .      -      0       gene_id
"g3_Ty1"; transcript_id "g3_Ty1.1"; cds_number "7"; gene_name ""; gene_source "MAN";
LR797800.1      MAN     exon        10010961 10011152 .      -      .       gene_id
"g3_Ty1"; transcript_id "g3_Ty1.1"; exon_number "8"; gene_name ""; gene_source "MAN";
LR797800.1      MAN     CDS         10010961 10011152 .      -      0       gene_id
"g3_Ty1"; transcript_id "g3_Ty1.1"; cds_number "8"; gene_name ""; gene_source "MAN";
LR797800.1      MAN     exon        10012414 10012580 .      -      .       gene_id
"g3_Ty1"; transcript_id "g3_Ty1.1"; exon_number "9"; gene_name ""; gene_source "MAN";
LR797800.1      MAN     CDS         10012414 10012580 .      -      2       gene_id
"g3_Ty1"; transcript_id "g3_Ty1.1"; cds_number "9"; gene_name ""; gene_source "MAN";
LR797800.1      MAN     exon        10012690 10012774 .      -      .       gene_id
"g3_Ty1"; transcript_id "g3_Ty1.1"; exon_number "10"; gene_name ""; gene_source "MAN";
LR797800.1      MAN     CDS         10012690 10012774 .      -      0       gene_id
"g3_Ty1"; transcript_id "g3_Ty1.1"; cds_number "10"; gene_name ""; gene_source "MAN";
LR797800.1      MAN     exon        10013000 10013200 .      -      .       gene_id
"g3_Ty1"; transcript_id "g3_Ty1.1"; exon_number "11"; gene_name ""; gene_source "MAN";
LR797800.1      MAN     CDS         10013000 10013200 .      -      0       gene_id
"g3_Ty1"; transcript_id "g3_Ty1.1"; cds_number "11"; gene_name ""; gene_source "MAN";
LR797800.1      MAN     exon        10013537 10013626 .      -      .       gene_id
"g3_Ty1"; transcript_id "g3_Ty1.1"; exon_number "12"; gene_name ""; gene_source "MAN";
LR797800.1      MAN     CDS         10013537 10013626 .      -      0       gene_id
"g3_Ty1"; transcript_id "g3_Ty1.1"; cds_number "12"; gene_name ""; gene_source "MAN";
LR797800.1      MAN     exon        10013732 10013917 .      -      .       gene_id
"g3_Ty1"; transcript_id "g3_Ty1.1"; exon_number "13"; gene_name ""; gene_source "MAN";
LR797800.1      MAN     CDS         10013732 10013917 .      -      0       gene_id
"g3_Ty1"; transcript_id "g3_Ty1.1"; cds_number "13"; gene_name ""; gene_source "MAN";
LR797800.1      MAN     exon        10014016 10014219 .      -      .       gene_id
"g3_Ty1"; transcript_id "g3_Ty1.1"; exon_number "14"; gene_name ""; gene_source "MAN";
LR797800.1      MAN     CDS         10014016 10014219 .      -      0       gene_id
"g3_Ty1"; transcript_id "g3_Ty1.1"; cds_number "14"; gene_name ""; gene_source "MAN";
LR797800.1      MAN     stop_codon           10008336 10008338 .      -      0
        gene_id "g3_Ty1"; transcript_id "g3_Ty1.1"; gene_name ""; gene_source "MAN";
LR797800.1      MAN     gene        10055797 10057417 .      -      .       gene_id
"g4_Ty1"; gene_name ""; gene_source "MAN";
LR797800.1      MAN     transcript  10055797 10057417 .      -      .       gene_id
"g4_Ty1"; transcript_id "g4_Ty1.1"; gene_source "MAN";
LR797800.1      MAN     exon        10055797 10056444 .      -      .       gene_id
"g4_Ty1"; transcript_id "g4_Ty1.1"; exon_number "1"; gene_name ""; gene_source "MAN";
LR797800.1      MAN     CDS         10055797 10056444 .      -      0       gene_id
"g4_Ty1"; transcript_id "g4_Ty1.1"; cds_number "1"; gene_name ""; gene_source "MAN";
LR797800.1      MAN     start_codon          10057415 10057417 .      -      0
        gene_id "g4_Ty1"; transcript_id "g4_Ty1.1"; gene_name ""; gene_source "MAN";
LR797800.1      MAN     exon        10056524 10057417 .      -      .       gene_id
"g4_Ty1"; transcript_id "g4_Ty1.1"; exon_number "2"; gene_name ""; gene_source "MAN";
LR797800.1      MAN     CDS         10056524 10057417 .      -      0       gene_id
"g4_Ty1"; transcript_id "g4_Ty1.1"; cds_number "2"; gene_name ""; gene_source "MAN";
LR797800.1      MAN     stop_codon           10055797 10055799 .      -      0
        gene_id "g4_Ty1"; transcript_id "g4_Ty1.1"; gene_name ""; gene_source "MAN";
LR797800.1      MAN     gene        10060334 10064607 .      -      .       gene_id
"g5_Ty1"; gene_name ""; gene_source "MAN";
```

```
LR797800.1	MAN	transcript	10060334	10064607	.	-	.	gene_id
"g5_Ty1"; transcript_id "g5_Ty1.1"; gene_source "MAN";
LR797800.1	MAN	exon	10060334	10060450	.	-	.	gene_id
"g5_Ty1"; transcript_id "g5_Ty1.1"; exon_number "1"; gene_name ""; gene_source "MAN";
LR797800.1	MAN	CDS	10060334	10060450	.	-	0	gene_id
"g5_Ty1"; transcript_id "g5_Ty1.1"; cds_number "1"; gene_name ""; gene_source "MAN";
LR797800.1	MAN	start_codon	10064605	10064607	.	-	0
	gene_id "g5_Ty1"; transcript_id "g5_Ty1.1"; gene_name ""; gene_source "MAN";
LR797800.1	MAN	exon	10060581	10061067	.	-	.	gene_id
"g5_Ty1"; transcript_id "g5_Ty1.1"; exon_number "2"; gene_name ""; gene_source "MAN";
LR797800.1	MAN	CDS	10060581	10061067	.	-	1	gene_id
"g5_Ty1"; transcript_id "g5_Ty1.1"; cds_number "2"; gene_name ""; gene_source "MAN";
LR797800.1	MAN	exon	10061164	10061210	.	-	.	gene_id
"g5_Ty1"; transcript_id "g5_Ty1.1"; exon_number "3"; gene_name ""; gene_source "MAN";
LR797800.1	MAN	CDS	10061164	10061210	.	-	0	gene_id
"g5_Ty1"; transcript_id "g5_Ty1.1"; cds_number "3"; gene_name ""; gene_source "MAN";
LR797800.1	MAN	exon	10061286	10061342	.	-	.	gene_id
"g5_Ty1"; transcript_id "g5_Ty1.1"; exon_number "4"; gene_name ""; gene_source "MAN";
LR797800.1	MAN	CDS	10061286	10061342	.	-	0	gene_id
"g5_Ty1"; transcript_id "g5_Ty1.1"; cds_number "4"; gene_name ""; gene_source "MAN";
LR797800.1	MAN	exon	10061770	10061868	.	-	.	gene_id
"g5_Ty1"; transcript_id "g5_Ty1.1"; exon_number "5"; gene_name ""; gene_source "MAN";
LR797800.1	MAN	CDS	10061770	10061868	.	-	0	gene_id
"g5_Ty1"; transcript_id "g5_Ty1.1"; cds_number "5"; gene_name ""; gene_source "MAN";
LR797800.1	MAN	exon	10061956	10062207	.	-	.	gene_id
"g5_Ty1"; transcript_id "g5_Ty1.1"; exon_number "6"; gene_name ""; gene_source "MAN";
LR797800.1	MAN	CDS	10061956	10062207	.	-	0	gene_id
"g5_Ty1"; transcript_id "g5_Ty1.1"; cds_number "6"; gene_name ""; gene_source "MAN";
LR797800.1	MAN	exon	10062527	10062718	.	-	.	gene_id
"g5_Ty1"; transcript_id "g5_Ty1.1"; exon_number "7"; gene_name ""; gene_source "MAN";
LR797800.1	MAN	CDS	10062527	10062718	.	-	0	gene_id
"g5_Ty1"; transcript_id "g5_Ty1.1"; cds_number "7"; gene_name ""; gene_source "MAN";
LR797800.1	MAN	exon	10062935	10063101	.	-	.	gene_id
"g5_Ty1"; transcript_id "g5_Ty1.1"; exon_number "8"; gene_name ""; gene_source "MAN";
LR797800.1	MAN	CDS	10062935	10063101	.	-	2	gene_id
"g5_Ty1"; transcript_id "g5_Ty1.1"; cds_number "8"; gene_name ""; gene_source "MAN";
LR797800.1	MAN	exon	10063198	10063282	.	-	.	gene_id
"g5_Ty1"; transcript_id "g5_Ty1.1"; exon_number "9"; gene_name ""; gene_source "MAN";
LR797800.1	MAN	CDS	10063198	10063282	.	-	0	gene_id
"g5_Ty1"; transcript_id "g5_Ty1.1"; cds_number "9"; gene_name ""; gene_source "MAN";
LR797800.1	MAN	exon	10063636	10063836	.	-	.	gene_id
"g5_Ty1"; transcript_id "g5_Ty1.1"; exon_number "10"; gene_name ""; gene_source "MAN";
LR797800.1	MAN	CDS	10063636	10063836	.	-	0	gene_id
"g5_Ty1"; transcript_id "g5_Ty1.1"; cds_number "10"; gene_name ""; gene_source "MAN";
LR797800.1	MAN	exon	10063923	10064012	.	-	.	gene_id
"g5_Ty1"; transcript_id "g5_Ty1.1"; exon_number "11"; gene_name ""; gene_source "MAN";
LR797800.1	MAN	CDS	10063923	10064012	.	-	0	gene_id
"g5_Ty1"; transcript_id "g5_Ty1.1"; cds_number "11"; gene_name ""; gene_source "MAN";
LR797800.1	MAN	exon	10064114	10064299	.	-	.	gene_id
"g5_Ty1"; transcript_id "g5_Ty1.1"; exon_number "12"; gene_name ""; gene_source "MAN";
LR797800.1	MAN	CDS	10064114	10064299	.	-	0	gene_id
"g5_Ty1"; transcript_id "g5_Ty1.1"; cds_number "12"; gene_name ""; gene_source "MAN";
LR797800.1	MAN	exon	10064404	10064607	.	-	.	gene_id
"g5_Ty1"; transcript_id "g5_Ty1.1"; exon_number "13"; gene_name ""; gene_source "MAN";
LR797800.1	MAN	CDS	10064404	10064607	.	-	0	gene_id
"g5_Ty1"; transcript_id "g5_Ty1.1"; cds_number "13"; gene_name ""; gene_source "MAN";
```

```
LR797800.1	MAN	stop_codon	10060334 10060336 .	-	0
	gene_id "g5_Ty1"; transcript_id "g5_Ty1.1"; gene_name ""; gene_source "MAN";
```

Kn-0

```
LR797810.1	MAN	gene	9446623	9452539	.	-	.	gene_id
"g1_Kn0"; gene_name ""; gene_source "MAN";
LR797810.1	MAN	transcript	9446623	9452539	.	-	.	gene_id
"g1_Kn0"; transcript_id "g1_Kn0.1"; gene_source "MAN";
LR797810.1	MAN	exon	9446623	9446739	.	-	.	gene_id
"g1_Kn0"; transcript_id "g1_Kn0.1"; exon_number "1"; gene_name ""; gene_source "MAN";
LR797810.1	MAN	CDS	9446623	9446739	.	-	0	gene_id
"g1_Kn0"; transcript_id "g1_Kn0.1"; cds_number "1"; gene_name ""; gene_source "MAN";
LR797810.1	MAN	start_codon	9452537	9452539	.	-	0
	gene_id "g1_Kn0"; transcript_id "g1_Kn0.1"; gene_name ""; gene_source "MAN";
LR797810.1	MAN	exon	9446942	9447428	.	-	.	gene_id
"g1_Kn0"; transcript_id "g1_Kn0.1"; exon_number "2"; gene_name ""; gene_source "MAN";
LR797810.1	MAN	CDS	9446942	9447428	.	-	1	gene_id
"g1_Kn0"; transcript_id "g1_Kn0.1"; cds_number "2"; gene_name ""; gene_source "MAN";
LR797810.1	MAN	exon	9447525	9447571	.	-	.	gene_id
"g1_Kn0"; transcript_id "g1_Kn0.1"; exon_number "3"; gene_name ""; gene_source "MAN";
LR797810.1	MAN	CDS	9447525	9447571	.	-	0	gene_id
"g1_Kn0"; transcript_id "g1_Kn0.1"; cds_number "3"; gene_name ""; gene_source "MAN";
LR797810.1	MAN	exon	9447657	9447713	.	-	.	gene_id
"g1_Kn0"; transcript_id "g1_Kn0.1"; exon_number "4"; gene_name ""; gene_source "MAN";
LR797810.1	MAN	CDS	9447657	9447713	.	-	0	gene_id
"g1_Kn0"; transcript_id "g1_Kn0.1"; cds_number "4"; gene_name ""; gene_source "MAN";
LR797810.1	MAN	exon	9448851	9448949	.	-	.	gene_id
"g1_Kn0"; transcript_id "g1_Kn0.1"; exon_number "5"; gene_name ""; gene_source "MAN";
LR797810.1	MAN	CDS	9448851	9448949	.	-	0	gene_id
"g1_Kn0"; transcript_id "g1_Kn0.1"; cds_number "5"; gene_name ""; gene_source "MAN";
LR797810.1	MAN	exon	9449030	9449281	.	-	.	gene_id
"g1_Kn0"; transcript_id "g1_Kn0.1"; exon_number "6"; gene_name ""; gene_source "MAN";
LR797810.1	MAN	CDS	9449030	9449281	.	-	0	gene_id
"g1_Kn0"; transcript_id "g1_Kn0.1"; cds_number "6"; gene_name ""; gene_source "MAN";
LR797810.1	MAN	exon	9449431	9449544	.	-	.	gene_id
"g1_Kn0"; transcript_id "g1_Kn0.1"; exon_number "7"; gene_name ""; gene_source "MAN";
LR797810.1	MAN	CDS	9449431	9449544	.	-	0	gene_id
"g1_Kn0"; transcript_id "g1_Kn0.1"; cds_number "7"; gene_name ""; gene_source "MAN";
LR797810.1	MAN	exon	9449613	9449755	.	-	.	gene_id
"g1_Kn0"; transcript_id "g1_Kn0.1"; exon_number "8"; gene_name ""; gene_source "MAN";
LR797810.1	MAN	CDS	9449613	9449755	.	-	2	gene_id
"g1_Kn0"; transcript_id "g1_Kn0.1"; cds_number "8"; gene_name ""; gene_source "MAN";
LR797810.1	MAN	exon	9450413	9450586	.	-	.	gene_id
"g1_Kn0"; transcript_id "g1_Kn0.1"; exon_number "9"; gene_name ""; gene_source "MAN";
LR797810.1	MAN	CDS	9450413	9450586	.	-	2	gene_id
"g1_Kn0"; transcript_id "g1_Kn0.1"; cds_number "9"; gene_name ""; gene_source "MAN";
LR797810.1	MAN	exon	9451389	9451473	.	-	.	gene_id
"g1_Kn0"; transcript_id "g1_Kn0.1"; exon_number "10"; gene_name ""; gene_source "MAN";
LR797810.1	MAN	CDS	9451389	9451473	.	-	0	gene_id
"g1_Kn0"; transcript_id "g1_Kn0.1"; cds_number "10"; gene_name ""; gene_source "MAN";
LR797810.1	MAN	exon	9451552	9451755	.	-	.	gene_id
"g1_Kn0"; transcript_id "g1_Kn0.1"; exon_number "11"; gene_name ""; gene_source "MAN";
LR797810.1	MAN	CDS	9451552	9451755	.	-	0	gene_id
"g1_Kn0"; transcript_id "g1_Kn0.1"; cds_number "11"; gene_name ""; gene_source "MAN";
```

```
LR797810.1      MAN     exon    9451845 9451934 .       -       .       gene_id
"g1_Kn0"; transcript_id "g1_Kn0.1"; exon_number "12"; gene_name ""; gene_source "MAN";
LR797810.1      MAN     CDS     9451845 9451934 .       -       0       gene_id
"g1_Kn0"; transcript_id "g1_Kn0.1"; cds_number "12"; gene_name ""; gene_source "MAN";
LR797810.1      MAN     exon    9452036 9452221 .       -       .       gene_id
"g1_Kn0"; transcript_id "g1_Kn0.1"; exon_number "13"; gene_name ""; gene_source "MAN";
LR797810.1      MAN     CDS     9452036 9452221 .       -       0       gene_id
"g1_Kn0"; transcript_id "g1_Kn0.1"; cds_number "13"; gene_name ""; gene_source "MAN";
LR797810.1      MAN     exon    9452336 9452539 .       -       .       gene_id
"g1_Kn0"; transcript_id "g1_Kn0.1"; exon_number "14"; gene_name ""; gene_source "MAN";
LR797810.1      MAN     CDS     9452336 9452539 .       -       0       gene_id
"g1_Kn0"; transcript_id "g1_Kn0.1"; cds_number "14"; gene_name ""; gene_source "MAN";
LR797810.1      MAN     stop_codon      9446623 9446625 .       -       0
        gene_id "g1_Kn0"; transcript_id "g1_Kn0.1"; gene_name ""; gene_source "MAN";
LR797810.1      MAN     gene    9454903 9456513 .       +       .       gene_id
"g2_Kn0"; gene_name ""; gene_source "MAN";
LR797810.1      MAN     transcript 9454903 9456513 .    +       .       gene_id
"g2_Kn0"; transcript_id "g2_Kn0.1"; gene_source "MAN";
LR797810.1      MAN     exon    9454903 9455802 .       +       .       gene_id
"g2_Kn0"; transcript_id "g2_Kn0.1"; exon_number "1"; gene_name ""; gene_source "MAN";
LR797810.1      MAN     CDS     9454903 9455802 .       +       0       gene_id
"g2_Kn0"; transcript_id "g2_Kn0.1"; cds_number "1"; gene_name ""; gene_source "MAN";
LR797810.1      MAN     start_codon     9454903 9454905 .       +       0
        gene_id "g2_Kn0"; transcript_id "g2_Kn0.1"; gene_name ""; gene_source "MAN";
LR797810.1      MAN     exon    9455884 9456513 .       +       .       gene_id
"g2_Kn0"; transcript_id "g2_Kn0.1"; exon_number "2"; gene_name ""; gene_source "MAN";
LR797810.1      MAN     CDS     9455884 9456513 .       +       0       gene_id
"g2_Kn0"; transcript_id "g2_Kn0.1"; cds_number "2"; gene_name ""; gene_source "MAN";
LR797810.1      MAN     stop_codon      9456511 9456513 .       +       0
        gene_id "g2_Kn0"; transcript_id "g2_Kn0.1"; gene_name ""; gene_source "MAN";
LR797810.1      MAN     gene    9465744 9471645 .       -       .       gene_id
"g3_Kn0"; gene_name ""; gene_source "MAN";
LR797810.1      MAN     transcript 9465744 9471645 .    -       .       gene_id
"g3_Kn0"; transcript_id "g3_Kn0.1"; gene_source "MAN";
LR797810.1      MAN     exon    9465744 9465854 .       -       .       gene_id
"g3_Kn0"; transcript_id "g3_Kn0.1"; exon_number "1"; gene_name ""; gene_source "MAN";
LR797810.1      MAN     CDS     9465744 9465854 .       -       0       gene_id
"g3_Kn0"; transcript_id "g3_Kn0.1"; cds_number "1"; gene_name ""; gene_source "MAN";
LR797810.1      MAN     start_codon     9471643 9471645 .       -       0
        gene_id "g3_Kn0"; transcript_id "g3_Kn0.1"; gene_name ""; gene_source "MAN";
LR797810.1      MAN     exon    9466018 9466504 .       -       .       gene_id
"g3_Kn0"; transcript_id "g3_Kn0.1"; exon_number "2"; gene_name ""; gene_source "MAN";
LR797810.1      MAN     CDS     9466018 9466504 .       -       1       gene_id
"g3_Kn0"; transcript_id "g3_Kn0.1"; cds_number "2"; gene_name ""; gene_source "MAN";
LR797810.1      MAN     exon    9466604 9466650 .       -       .       gene_id
"g3_Kn0"; transcript_id "g3_Kn0.1"; exon_number "3"; gene_name ""; gene_source "MAN";
LR797810.1      MAN     CDS     9466604 9466650 .       -       0       gene_id
"g3_Kn0"; transcript_id "g3_Kn0.1"; cds_number "3"; gene_name ""; gene_source "MAN";
LR797810.1      MAN     exon    9466725 9466781 .       -       .       gene_id
"g3_Kn0"; transcript_id "g3_Kn0.1"; exon_number "4"; gene_name ""; gene_source "MAN";
LR797810.1      MAN     CDS     9466725 9466781 .       -       0       gene_id
"g3_Kn0"; transcript_id "g3_Kn0.1"; cds_number "4"; gene_name ""; gene_source "MAN";
LR797810.1      MAN     exon    9467606 9467704 .       -       .       gene_id
"g3_Kn0"; transcript_id "g3_Kn0.1"; exon_number "5"; gene_name ""; gene_source "MAN";
LR797810.1      MAN     CDS     9467606 9467704 .       -       0       gene_id
"g3_Kn0"; transcript_id "g3_Kn0.1"; cds_number "5"; gene_name ""; gene_source "MAN";
```

```
LR797810.1	MAN	exon	9467800	9468051	.	-	.	gene_id
"g3_Kn0"; transcript_id "g3_Kn0.1"; exon_number "6"; gene_name ""; gene_source "MAN";
LR797810.1	MAN	CDS	9467800	9468051	.	-	0	gene_id
"g3_Kn0"; transcript_id "g3_Kn0.1"; cds_number "6"; gene_name ""; gene_source "MAN";
LR797810.1	MAN	exon	9468173	9468286	.	-	.	gene_id
"g3_Kn0"; transcript_id "g3_Kn0.1"; exon_number "7"; gene_name ""; gene_source "MAN";
LR797810.1	MAN	CDS	9468173	9468286	.	-	0	gene_id
"g3_Kn0"; transcript_id "g3_Kn0.1"; cds_number "7"; gene_name ""; gene_source "MAN";
LR797810.1	MAN	exon	9468385	9468576	.	-	.	gene_id
"g3_Kn0"; transcript_id "g3_Kn0.1"; exon_number "8"; gene_name ""; gene_source "MAN";
LR797810.1	MAN	CDS	9468385	9468576	.	-	0	gene_id
"g3_Kn0"; transcript_id "g3_Kn0.1"; cds_number "8"; gene_name ""; gene_source "MAN";
LR797810.1	MAN	exon	9469838	9470004	.	-	.	gene_id
"g3_Kn0"; transcript_id "g3_Kn0.1"; exon_number "9"; gene_name ""; gene_source "MAN";
LR797810.1	MAN	CDS	9469838	9470004	.	-	2	gene_id
"g3_Kn0"; transcript_id "g3_Kn0.1"; cds_number "9"; gene_name ""; gene_source "MAN";
LR797810.1	MAN	exon	9470112	9470196	.	-	.	gene_id
"g3_Kn0"; transcript_id "g3_Kn0.1"; exon_number "10"; gene_name ""; gene_source "MAN";
LR797810.1	MAN	CDS	9470112	9470196	.	-	0	gene_id
"g3_Kn0"; transcript_id "g3_Kn0.1"; cds_number "10"; gene_name ""; gene_source "MAN";
LR797810.1	MAN	exon	9470426	9470626	.	-	.	gene_id
"g3_Kn0"; transcript_id "g3_Kn0.1"; exon_number "11"; gene_name ""; gene_source "MAN";
LR797810.1	MAN	CDS	9470426	9470626	.	-	0	gene_id
"g3_Kn0"; transcript_id "g3_Kn0.1"; cds_number "11"; gene_name ""; gene_source "MAN";
LR797810.1	MAN	exon	9470963	9471052	.	-	.	gene_id
"g3_Kn0"; transcript_id "g3_Kn0.1"; exon_number "12"; gene_name ""; gene_source "MAN";
LR797810.1	MAN	CDS	9470963	9471052	.	-	0	gene_id
"g3_Kn0"; transcript_id "g3_Kn0.1"; cds_number "12"; gene_name ""; gene_source "MAN";
LR797810.1	MAN	exon	9471158	9471343	.	-	.	gene_id
"g3_Kn0"; transcript_id "g3_Kn0.1"; exon_number "13"; gene_name ""; gene_source "MAN";
LR797810.1	MAN	CDS	9471158	9471343	.	-	0	gene_id
"g3_Kn0"; transcript_id "g3_Kn0.1"; cds_number "13"; gene_name ""; gene_source "MAN";
LR797810.1	MAN	exon	9471442	9471645	.	-	.	gene_id
"g3_Kn0"; transcript_id "g3_Kn0.1"; exon_number "14"; gene_name ""; gene_source "MAN";
LR797810.1	MAN	CDS	9471442	9471645	.	-	0	gene_id
"g3_Kn0"; transcript_id "g3_Kn0.1"; cds_number "14"; gene_name ""; gene_source "MAN";
LR797810.1	MAN	stop_codon	9465744	9465746	.	-	0
	gene_id "g3_Kn0"; transcript_id "g3_Kn0.1"; gene_name ""; gene_source "MAN";
```

Rozdział 7

Oświadczenia współautorów

Publikacja 1

OŚWIADCZENIA WSPÓŁAUTORÓW

Samelak-Czajka A, **Marszalek-Zenczak M**, Marcinkowska-Swojak M, Kozlowski P, Figlerowicz M and Zmienko A (2017)

MLPA-Based Analysis of Copy Number Variation in Plant Populations.

Front. Plant Sci.8:222. doi: 10.3389/fpls.2017.00222

Poznań, 03.02.2018

Anna Samelak-Czajka

Instytut Chemii Bioorganicznej Polskiej Akademii Nauk
w Poznaniu
Instytut Informatyki, Wydział Informatyki
Politechnika Poznańska

Tytuł publikacji: MLPA-based analysis of copy number variations in plant populations

Autorzy: Anna Samelak-Czajka, Małgorzata Marszalek-Zenczak, Małgorzata Marcinkowska-Swojak, Piotr Kozlowski, Marek Figlerowicz, Agnieszka Zmienko

Rok opublikowania: 2017

Czasopismo: Frontiers in Plant Science (Front Plant Sci. 2017; 8: 222. doi: 10.3389/fpls.2017.00222)

OŚWIADCZENIE

Oświadczam, że mój wkład w powstanie powyższej pracy polegał na opracowaniu i przetestowaniu procedury MLPA do badania zmienności liczby kopii u *Arabidopsis thaliana*, zaprojektowaniu specyficznych sond MLPA, przygotowaniu materiału do badań (hodowla roślin, ekstrakcja DNA), przeprowadzeniu eksperymentów MLPA, analizie uzyskanych wyników, udziale w tworzeniu treści protokołu i rysunków do manuskryptu oraz korekcie tekstu.

Anna Samelak-Czajka

Poznań, 03.02.2018

Małgorzata Marszałek-Zeńczak

Instytut Chemii Bioorganicznej Polskiej Akademii Nauk
w Poznaniu

Tytuł publikacji: MLPA-based Analysis of Copy Number Variations in Plant Populations

Autorzy: Anna Samelak-Czajka, Małgorzata Marszałek-Zenczak, Małgorzata Marcinkowska-Swojak, Piotr Kozłowski, Marek Figlerowicz, Agnieszka Zmienko

Rok opublikowania: 2017

Czasopismo: Frontiers in Plant Science

OŚWIADCZENIE

Oświadczam, że mój wkład w powstanie powyższej pracy polegał na wykonaniu i analizie eksperymentu optymalizacji ilości matrycy do metody MLPA oraz na przeczytaniu i zaakceptowaniu treści manuskryptu.

Małgorzata Marszałek - Zeńczak

POLSKA AKADEMIA NAUK

INSTYTUT CHEMII BIOORGANICZNEJ

ul. Noskowskiego 12/ 14, 61-704 Poznań, Polska
tel.: +48-61 852 85 03, sekretariat 852 89 19
fax: +48-61 8520532 e-mail: office@ibch.poznan.pl

Poznań, 03.02.2018

Małgorzata Marcinkowska-Swojak

Instytut Chemii Bioorganicznej Polskiej Akademii Nauk
w Poznaniu

Tytuł publikacji: MLPA-based Analysis of Copy Number Variations in Plant Populations

Autorzy: Anna Samelak-Czajka, Małgorzata Marszałek-Zenczak, Małgorzata Marcinkowska-Swojak, Piotr Kozłowski, Marek Figlerowicz, Agnieszka Zmienko

Rok opublikowania: 2017

Czasopismo: Frontiers in Plant Science

OŚWIADCZENIE

Oświadczam, że mój wkład w powstanie powyższej pracy polegał na konsultacji sposobu przeprowadzenia reakcji MLPA oraz na przeczytaniu i zaakceptowaniu treści manuskryptu.

Marcinkowska-Swojak

POLSKA AKADEMIA NAUK

INSTYTUT CHEMII BIOORGANICZNEJ

ul. Noskowskiego 12/ 14, 61-704 Poznań, Polska
tel.: +48-61 852 85 03, sekretariat 852 89 19
fax: +48-61 8520532 e-mail: office@ibch.poznan.pl

Poznań, 03.02.2018

Piotr Kozłowski

Instytut Chemii Bioorganicznej Polskiej Akademii Nauk
w Poznaniu

Tytuł publikacji: MLPA-based Analysis of Copy Number Variations in Plant Populations

Autorzy: Anna Samelak-Czajka, Małgorzata Marszałek-Zenczak, Małgorzata Marcinkowska-Swojak, Piotr Kozłowski, Marek Figlerowicz, Agnieszka Zmienko

Rok opublikowania: 2017

Czasopismo: Frontiers in Plant Science

OŚWIADCZENIE

Oświadczam, że mój wkład w powstanie powyższej pracy polegał na: konsultacji wyników genotypowania oraz treści manuskryptu.

Piotr Kozłowski

ICHB PAN
POZNAŃ

Instytut Chemii Bioorganicznej
Polskiej Akademii Nauk

Poznań, 8.02.2018

Prof. dr hab. Marek Figlerowicz
Instytut Chemii Bioorganicznej
Polskiej Akademii Nauk
Ul. Noskowskiego 12/14
61-704 Poznań
marekf@ibch.poznan.pl

OŚWIADCZENIE

Oświadczam, że w przypadku wymienionej poniżej publikacji uczestniczyłem w tworzeniu ogólnej koncepcji badań, dyskusji otrzymanych wyników oraz edycji manuskryptu.

Tytuł publikacji: MLPA-based Analysis of Copy Number Variations in Plant Populations

Autorzy: Anna Samelak-Czajka, Małgorzata Marszałek-Zenczak, Małgorzata Marcinkowska-Swojak, Piotr Kozłowski, Marek Figlerowicz, Agnieszka Zmienko

Rok opublikowania: 2017

Czasopismo: Frontiers in Plant Science

POLSKA AKADEMIA NAUK

INSTYTUT CHEMII BIOORGANICZNEJ

ul. Noskowskiego 12/ 14, 61-704 Poznań, Polska
tel.: +48-61 852 85 03, sekretariat 852 89 19
fax: +48-61 8520532 e-mail: office@ibch.poznan.pl

Poznań, 03.02.2018

Agnieszka Żmieńko

Instytut Chemii Bioorganicznej Polskiej Akademii Nauk
w Poznaniu

Tytuł publikacji: MLPA-based Analysis of Copy Number Variations in Plant Populations

Autorzy: Anna Samelak-Czajka, Małgorzata Marszałek-Zenczak, Małgorzata Marcinkowska-Swojak, Piotr Kozłowski, Marek Figlerowicz, Agnieszka Zmienko

Rok opublikowania: 2017

Czasopismo: Frontiers in Plant Science

OŚWIADCZENIE

Oświadczam, że mój wkład w powstanie powyższej pracy polegał na: określeniu koncepcji i kształtu pracy, opracowaniu i interpretacji wyników (wspólnie z ASCz), napisaniu manuskryptu, przygotowaniu rysunków oraz suplementów, a także opiece naukowej nad członkami zespołu badawczego (ASCz i MMZ).

Agnieszka Żmieńko

Publikacja 2

OŚWIADCZENIA WSPÓŁAUTORÓW

Zmienko A, **Marszalek-Zenczak M**, Wojciechowski P, Samelak-Czajka A, Luczak M, Kozlowski P, Karlowski WM, Figlerowicz M.

AthCNV: A Map of DNA Copy Number Variations in the Arabidopsis Genome.

Plant Cell. 2020 Jun;32(6):1797-1819. doi: 10.1105/tpc.19.00640.

Poznań, 05.07.2023 r.

dr hab. Agnieszka Żmieńko, prof. ICHB PAN
Instytut Chemii Bioorganicznej PAN
ul. Noskowskiego 12/14
61-704 Poznań

Politechnika Poznańska
Wydział Informatyki i Telekomunikacji, Instytut Informatyki
ul. Piotrowo 3
60-965 Poznań

Tytuł publikacji: AthCNV: A Map of DNA Copy Number Variations in the Arabidopsis Genome

Autorzy: Agnieszka Zmienko*, Malgorzata Marszalek-Zenczak, Pawel Wojciechowski, Anna Samelak-Czajka, Magdalena Luczak, Piotr Kozlowski, Wojciech M. Karlowski, Marek Figlerowicz*
*współautorzy korespondencyjni

Rok opublikowania: 2020

Czasopismo: The Plant Cell; https://doi.org/10.1105/tpc.19.00640

OŚWIADCZENIE

Oświadczam, że mój wkład w powstanie powyższej pracy polegał na:

- udziale w określeniu koncepcji pracy;
- planowaniu i wykonaniu (wspólnie z innymi współautorami) bioinformatycznej części analiz, koordynowaniu ich i weryfikacji wyników;
- zaprojektowaniu, wykonaniu i analizie wyników eksperymentów MLPA;
- analizie i krytycznej interpretacji pozostałych wyników;

- koncepcji, zaprojektowaniu układu i opracowaniu danych wejściowych dla stworzenia strony do wizualizacji zmienności genów;
- zaprojektowaniu i opracowaniu rysunków, tabel oraz suplementów, przy udziale pozostałych współautorów;
- napisaniu manuskryptu, przy udziale pozostałych współautorów;
- redagowaniu ostatecznej wersji manuskryptu, wspólnie z pozostałymi współautorami.

Agnieszka Żmieńko

dr hab. Agnieszka Żmieńko, prof. ICHB PAN

Poznań, 04.07.2023 r.

mgr inż. Małgorzata Marszałek-Zeńczak
Instytut Chemii Bioorganicznej PAN
ul. Noskowskiego 12/14
61-704 Poznań

Tytuł publikacji: AthCNV: A Map of DNA Copy Number Variations in the Arabidopsis Genome

Autorzy: Agnieszka Zmienko*, Malgorzata Marszalek-Zenczak, Pawel Wojciechowski, Anna Samelak-Czajka, Magdalena Luczak, Piotr Kozlowski, Wojciech M. Karlowski, Marek Figlerowicz*

*współautorzy korespondencyjni

Rok opublikowania: 2020

Czasopismo: The Plant Cell; https://doi.org/10.1105/tpc.19.00640

OŚWIADCZENIE

Oświadczam, że mój wkład w powstanie powyższej pracy polegał na:

1. Udziale w planowaniu analiz bioinformatycznych
2. Wykonaniu większości analiz bioinformatycznych, w tym:
   - optymalizacji parametrów programów oraz identyfikacji i analizie CNV czterema programami: CNVnator, Control-FREEC, BreakDancer, Pindel;
   - opracowaniu indywidualnych kryteriów filtracji wygenerowanych CNV dla wszystkich metod i programów użytych w projekcie;
   - opracowaniu autorskiego podejścia integracji danych z różnych programów w celu stworzenia katalogu AthCNV;
   - analizie statystycznej wszystkich wyników;

- przygotowaniu i wykonaniu analiz PCA w oparciu o dane SNP pobrane z bazy *1001 Genomes Project* oraz dane CNV (dane własne);

- testach i optymalizacji protokołu analizy GWAS, przygotowaniu danych wejściowych, wykonaniu analiz;

3. Udziale w analizie i interpretacji wyników
4. Udziale w pisaniu manuskryptu, w szczególności sekcji *Methods* oraz opisu wyników analiz bioinformatycznych
5. Współtworzeniu rycin i suplementów.

Małgorzata Marszałek- Zeńczak

mgr inż. Małgorzata Marszałek-Zeńczak

Poznań, 10.07.2023 r.

dr inż. Paweł Wojciechowski
Politechnika Poznańska
Wydział Informatyki i Telekomunikacji, Instytut Informatyki
ul. Piotrowo 3
60-965 Poznań

Instytut Chemii Bioorganicznej PAN
ul. Noskowskiego 12/14
61-704 Poznań

Tytuł publikacji: AthCNV: A Map of DNA Copy Number Variations in the Arabidopsis Genome

Autorzy: Agnieszka Zmienko*, Malgorzata Marszalek-Zenczak, Pawel Wojciechowski, Anna Samelak-Czajka, Magdalena Luczak, Piotr Kozlowski, Wojciech M. Karlowski, Marek Figlerowicz*
*współautorzy korespondencyjni

Rok opublikowania: 2020

Czasopismo: The Plant Cell; https://doi.org/10.1105/tpc.19.00640

OŚWIADCZENIE

Oświadczam, że mój wkład w powstanie powyższej pracy polegał na:

- technicznym wsparciu obróbki danych sekwencyjnych;
- wygenerowaniu wstępnej listy zmian strukturalnych i wyników genotypowania z użyciem programów VariationHunter, Genome STRiP-CNV oraz Genome STRiP-SV i Genome STRiP SVGenotyper;
- stworzeniu przeglądarki internetowej do wizualizacji wyników genotypowania;
- przeczytaniu i zaakceptowaniu ostatecznej wersji manuskryptu.

Paweł Wojciechowski
dr inż. Paweł Wojciechowski

Poznań, 04.07.2023 r.

dr Anna Samelak-Czajka
Instytut Chemii Bioorganicznej PAN
ul. Noskowskiego 12/14
61-704 Poznań

Tytuł publikacji: AthCNV: A Map of DNA Copy Number Variations in the Arabidopsis Genome

Autorzy: Agnieszka Zmienko*, Malgorzata Marszalek-Zenczak, Pawel Wojciechowski, Anna Samelak-Czajka, Magdalena Luczak, Piotr Kozlowski, Wojciech M. Karlowski, Marek Figlerowicz*
*współautorzy korespondencyjni

Rok opublikowania: 2020

Czasopismo: The Plant Cell; https://doi.org/10.1105/tpc.19.00640

OŚWIADCZENIE

Oświadczam, że mój wkład w powstanie powyższej pracy polegał na:

- hodowli roślin i izolacji DNA;
- przeczytaniu i zaakceptowaniu ostatecznej wersji manuskryptu.

dr Anna Samelak-Czajka

Anna Samelak-Czajka

Poznań, 04.07.2023 r.

dr hab. Magdalena Łuczak, prof. ICHB PAN
Instytut Chemii Bioorganicznej PAN
ul. Noskowskiego 12/14
61-704 Poznań

Tytuł publikacji: AthCNV: A Map of DNA Copy Number Variations in the Arabidopsis Genome

Autorzy: Agnieszka Zmienko*, Malgorzata Marszalek-Zenczak, Pawel Wojciechowski, Anna Samelak-Czajka, Magdalena Luczak, Piotr Kozlowski, Wojciech M. Karlowski, Marek Figlerowicz*
*współautorzy korespondencyjni

Rok opublikowania: 2020

Czasopismo: The Plant Cell; https://doi.org/10.1105/tpc.19.00640

OŚWIADCZENIE

Oświadczam, że mój wkład w powstanie powyższej pracy polegał na:

- wykonaniu analiz proteomicznych oraz opracowaniu uzyskanych wyników;
- przeczytaniu i zaakceptowaniu ostatecznej wersji manuskryptu.

Magdalena Łuczak

dr hab. Magdalena Łuczak, prof. ICHB PAN

Poznań, 05.07.2023 r.

prof. dr hab. Piotr Kozłowski
Instytut Chemii Bioorganicznej PAN
ul. Noskowskiego 12/14
61-704 Poznań

Tytuł publikacji: AthCNV: A Map of DNA Copy Number Variations in the Arabidopsis Genome

Autorzy: Agnieszka Zmienko*, Malgorzata Marszalek-Zenczak, Pawel Wojciechowski, Anna Samelak-Czajka, Magdalena Luczak, Piotr Kozlowski, Wojciech M. Karlowski, Marek Figlerowicz*

*współautorzy korespondencyjni

Rok opublikowania: 2020

Czasopismo: The Plant Cell; https://doi.org/10.1105/tpc.19.00640

OŚWIADCZENIE

Oświadczam, że mój wkład w powstanie powyższej pracy polegał na:

- analizie i krytycznej interpretacji wyników genotypowania;
- edycji ostatecznej wersji manuskryptu.

Piotr Kozłowski

prof. dr hab. Piotr Kozłowski

Poznań, 05.07.2023 r.

prof. dr hab. Marek Figlerowicz
Instytut Chemii Bioorganicznej PAN
ul. Noskowskiego 12/14
61-704 Poznań

Tytuł publikacji: AthCNV: A Map of DNA Copy Number Variations in the Arabidopsis Genome

Autorzy: Agnieszka Zmienko*, Malgorzata Marszalek-Zenczak, Pawel Wojciechowski, Anna Samelak-Czajka, Magdalena Luczak, Piotr Kozlowski, Wojciech M. Karlowski, Marek Figlerowicz*

*współautorzy korespondencyjni

Rok opublikowania: 2020

Czasopismo: The Plant Cell; https://doi.org/10.1105/tpc.19.00640

OŚWIADCZENIE

Oświadczam, że mój wkład w powstanie powyższej pracy polegał na:

- udziale w określeniu koncepcji pracy;
- pozyskaniu częściowego finansowania badań;
- krytycznej analizie zebranych danych;
- opiece merytorycznej nad projektem;
- współredagowaniu ostatecznej wersji manuskryptu.

prof. dr hab. Marek Figlerowicz

Publikacja 3

OŚWIADCZENIA WSPÓŁAUTORÓW

Marszalek-Zenczak M, Satyr A, Wojciechowski P, Zenczak M, Sobieszczanska P, Brzezinski K, Iefimenko T, Figlerowicz M, Zmienko A. **Analysis of Arabidopsis non-reference accessions reveals high diversity of metabolic gene clusters and discovers new candidate cluster members.** Front Plant Sci. 2023 Jan 26;14:1104303. doi: 10.3389/fpls.2023.1104303.

Poznań, 07.04.2023 r.

mgr inż. Małgorzata Marszałek-Zeńczak
Instytut Chemii Bioorganicznej PAN
ul. Noskowskiego 12/14
61-704 Poznań

Tytuł publikacji: Analysis of Arabidopsis non-reference accessions reveals high diversity of metabolic gene clusters and discovers new candidate cluster members

Autorzy: Malgorzata Marszalek-Zenczak, Anastasiia Satyr, Pawel Wojciechowski, Michal Zenczak, Paula Sobieszczanska, Krzysztof Brzezinski, Tetiana Iefimenko, Marek Figlerowicz, Agnieszka Zmienko*
*autor korespondencyjny

Rok opublikowania: 2023

Czasopismo: Frontiers in Plant Science; https://doi.org/10.3389/fpls.2023.1104303

OŚWIADCZENIE

Oświadczam, że mój wkład w powstanie powyższej pracy polegał na:

1. Zaplanowaniu i wykonaniu analiz, w tym:
   - określeniu liczby kopii genów wchodzących w skład badanych MGC poprzez integrację danych genotypowania z danymi eksperymentalnymi dla >1000 ekotypów;
   - identyfikacji inwersji chromosomowych w regionie klastra thalianolu (wybór metody, przygotowanie danych do analizy oraz jej wykonanie, dobór kryteriów filtracji uzyskanych wyników, opracowanie i analiza końcowych wyników);
   - obróbce danych SNP pobranych z bazy danych projektu *1001 Genomes*, danych fenotypowych pobranych z bazy *Arapheno* oraz danych RNA-Seq pobranych z bazy *NCBI/SRA* i ich przygotowaniu do dalszych analiz;

- wykonaniu analizy PCA oraz interpretacji uzyskanych wyników;
- wykonaniu analizy GWAS oraz interpretacji uzyskanych wyników;
- wykonaniu analizy ekspresji genów oraz interpretacji uzyskanych wyników;
- zebraniu i opracowaniu informacji dotyczących genów oksydaz cytochromowych z dostępnych źródeł oraz analizie zmienności genów par TS-CYP;
- wykonaniu analiz ddPCR;
- wykonaniu wszystkich analiz statystycznych;
- udziału w analizie i krytycznej interpretacji pozostałych wyników;

2. Zaprojektowaniu rycin i suplementów wspólnie z promotor dr hab. Agnieszką Żmieńko oraz wykonaniu większości z nich;
3. Współredagowaniu pierwszej i ostatecznej wersji manuskryptu.

Małgorzata Marszałek-Zeńczak

mgr inż. Małgorzata Marszałek-Zeńczak

Poznań, 04.07.2023 r.

mgr inż. Anastasiia Satyr
Instytut Chemii Bioorganicznej PAN
ul. Noskowskiego 12/14
61-704 Poznań

Tytuł publikacji: Analysis of Arabidopsis non-reference accessions reveals high diversity of metabolic gene clusters and discovers new candidate cluster members

Autorzy: Malgorzata Marszalek-Zenczak, Anastasiia Satyr, Pawel Wojciechowski, Michal Zenczak, Paula Sobieszczanska, Krzysztof Brzezinski, Tetiana Iefimenko, Marek Figlerowicz, Agnieszka Zmienko*
*autor korespondencyjny

Rok opublikowania: 2023

Czasopismo: Frontiers in Plant Science; https://doi.org/10.3389/fpls.2023.1104303

OŚWIADCZENIE

Oświadczam, że mój wkład w powstanie powyższej pracy polegał na:

- wykonaniu nanoporowego sekwencjonowania DNA oraz zasemblowaniu *de novo* sekwencji genomowych dla linii Mitterberg-2-185 oraz Dolna-1-40;
- przeczytaniu i zaakceptowaniu ostatecznej wersji manuskryptu.

mgr inż. Anastasiia Satyr

Anastasiia Satyr

Poznań, 10.07.2023 r.

dr inż. Paweł Wojciechowski
Politechnika Poznańska
Wydział Informatyki i Telekomunikacji, Instytut Informatyki
ul. Piotrowo 3
60-965 Poznań

Instytut Chemii Bioorganicznej PAN
ul. Noskowskiego 12/14
61-704 Poznań

Tytuł publikacji: Analysis of Arabidopsis non-reference accessions reveals high diversity of metabolic gene clusters and discovers new candidate cluster members

Autorzy: Malgorzata Marszalek-Zenczak, Anastasiia Satyr, Pawel Wojciechowski, Michal Zenczak, Paula Sobieszczanska, Krzysztof Brzezinski, Tetiana Iefimenko, Marek Figlerowicz, Agnieszka Zmienko*
*autor korespondencyjny

Rok opublikowania: 2023

Czasopismo: Frontiers in Plant Science; https://doi.org/10.3389/fpls.2023.1104303

OŚWIADCZENIE

Oświadczam, że mój wkład w powstanie powyższej pracy polegał na:

- udziale w asemblowaniu *de novo* sekwencji genomowych dla linii Mitterberg-2-185 oraz Dolna-1-40;
- detekcji SNP i krótkich indeli w obrębie genów *CYP705A2* i *BARS1* w liniach *A. thaliana* z projektu 1001 Genomów Arabidopsis;
- przeczytaniu i zaakceptowaniu ostatecznej wersji manuskryptu.

Paweł Wojciechowski

dr inż. Paweł Wojciechowski

Poznań, 04.07.2023 r.

mgr inż. Michał Zeńczak
Instytut Chemii Bioorganicznej PAN
ul. Noskowskiego 12/14
61-704 Poznań

Tytuł publikacji: Analysis of Arabidopsis non-reference accessions reveals high diversity of metabolic gene clusters and discovers new candidate cluster members

Autorzy: Malgorzata Marszalek-Zenczak, Anastasiia Satyr, Pawel Wojciechowski, Michal Zenczak, Paula Sobieszczanska, Krzysztof Brzezinski, Tetiana Iefimenko, Marek Figlerowicz, Agnieszka Zmienko*
*autor korespondencyjny

Rok opublikowania: 2023

Czasopismo: Frontiers in Plant Science; https://doi.org/10.3389/fpls.2023.1104303

OŚWIADCZENIE

Oświadczam, że mój wkład w powstanie powyższej pracy polegał na:

- wsparciu technicznemu na etapie obróbki danych;
- udziale w interpretacji wyników GWAS;
- przeczytaniu i zaakceptowaniu ostatecznej wersji manuskryptu.

Michał Zeńczak

mgr inż. Michał Zeńczak

Poznań, 04.07.2023 r.

mgr Paula Sobieszczańska
Instytut Chemii Bioorganicznej PAN
ul. Noskowskiego 12/14
61-704 Poznań

Tytuł publikacji: Analysis of Arabidopsis non-reference accessions reveals high diversity of metabolic gene clusters and discovers new candidate cluster members

Autorzy: Malgorzata Marszalek-Zenczak, Anastasiia Satyr, Pawel Wojciechowski, Michal Zenczak, Paula Sobieszczanska, Krzysztof Brzezinski, Tetiana Iefimenko, Marek Figlerowicz, Agnieszka Zmienko*
*autor korespondencyjny

Rok opublikowania: 2023

Czasopismo: Frontiers in Plant Science; https://doi.org/10.3389/fpls.2023.1104303

OŚWIADCZENIE

Oświadczam, że mój wkład w powstanie powyższej pracy polegał na:

- wykonaniu eksperymentów MLPA;
- przeczytaniu i zaakceptowaniu ostatecznej wersji manuskryptu.

Sobieszczańska

mgr Paula Sobieszczańska

Poznań, 04.07.2023 r.

dr hab. Krzysztof Brzeziński, prof. ICHB PAN
Instytut Chemii Bioorganicznej PAN
ul. Noskowskiego 12/14
61-704 Poznań

Tytuł publikacji: Analysis of Arabidopsis non-reference accessions reveals high diversity of metabolic gene clusters and discovers new candidate cluster members

Autorzy: Malgorzata Marszalek-Zenczak, Anastasiia Satyr, Pawel Wojciechowski, Michal Zenczak, Paula Sobieszczanska, Krzysztof Brzezinski, Tetiana Iefimenko, Marek Figlerowicz, Agnieszka Zmienko*
*autor korespondencyjny

Rok opublikowania: 2023

Czasopismo: Frontiers in Plant Science; https://doi.org/10.3389/fpls.2023.1104303

OŚWIADCZENIE

Oświadczam, że mój wkład w powstanie powyższej pracy polegał na:

- wykonaniu analizy filogenetycznej genów OSC;
- wygenerowaniu i analizie modeli 3D przewidywanych białek kodowanych przez geny *BARS1* i *BARS2;*
- współredagowaniu opisu ww. analiz oraz przeczytaniu i zaakceptowaniu ostatecznej wersji manuskryptu.

dr hab. Krzysztof Brzeziński, prof. ICHB PAN

Poznań, 05.07.2023 r.

prof. dr hab. Marek Figlerowicz
Instytut Chemii Bioorganicznej PAN
ul. Noskowskiego 12/14
61-704 Poznań

Tytuł publikacji: Analysis of Arabidopsis non-reference accessions reveals high diversity of metabolic gene clusters and discovers new candidate cluster members

Autorzy: Malgorzata Marszalek-Zenczak, Anastasiia Satyr, Pawel Wojciechowski, Michal Zenczak, Paula Sobieszczanska, Krzysztof Brzezinski, Tetiana Iefimenko, Marek Figlerowicz, Agnieszka Zmienko*
*autor korespondencyjny

Rok opublikowania: 2023

Czasopismo: Frontiers in Plant Science; https://doi.org/10.3389/fpls.2023.1104303.

OŚWIADCZENIE

Oświadczam, że mój wkład w powstanie powyższej pracy polegał na:

- pozyskaniu częściowego finansowania badań;
- krytycznej analizie zebranych danych;
- współredagowaniu ostatecznej wersji manuskryptu.

prof. dr hab. Marek Figlerowicz

Poznań, 04.07.2023 r.

dr hab. Agnieszka Żmieńko, prof. ICHB PAN

Instytut Chemii Bioorganicznej PAN
ul. Noskowskiego 12/14
61-704 Poznań

Tytuł publikacji: Analysis of Arabidopsis non-reference accessions reveals high diversity of metabolic gene clusters and discovers new candidate cluster members

Autorzy: Malgorzata Marszalek-Zenczak, Anastasiia Satyr, Pawel Wojciechowski, Michal Zenczak, Paula Sobieszczanska, Krzysztof Brzezinski, Tetiana Iefimenko, Marek Figlerowicz, Agnieszka Zmienko*
*autor korespondencyjny

Rok opublikowania: 2023

Czasopismo: Frontiers in Plant Science; https://doi.org/10.3389/fpls.2023.1104303

OŚWIADCZENIE

Oświadczam, że mój wkład w powstanie powyższej pracy polegał na:

- określeniu koncepcji pracy;
- pozyskaniu częściowego finansowania badań;
- udziale w planowaniu eksperymentów i analiz;
- krytycznej analizie uzyskanych danych;
- udziale w projektowaniu niektórych rycin i suplementów;
- współredagowaniu pierwszej i ostatecznej wersji manuskryptu;
- koordynowaniu pracy zespołu.

Agnieszka Żmieńko

dr hab. Agnieszka Żmieńko, prof. ICHB PAN