Research article

Copy number variations shape the structural diversity of Arabidopsis metabolic gene clusters and are associated with the climatic gradient

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], Agnieszka Zmienko[1*]

[1]Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poland

[2]Institute of Computing Science, Faculty of Computing and Telecommunications, Poznan University of Technology, Poznan, Poland

[3]National University of Kyiv-Mohyla Academy, Kyiv, Ukraine

[*] Corresponding author:

Agnieszka Zmienko, e-mail akisiel@ibch.poznan.pl

Abstract

Background: Metabolic gene clusters (MGCs) encode at least three different enzymes for a common biosynthetic pathway. Comparative genome analyses highlighted the role of duplications, deletions and rearrangements in MGC formation. We hypothesized that these mechanisms also contribute to MGC intraspecies diversity and play a role in adaptation.

Results: We assessed copy number variations (CNVs) of four *Arabidopsis thaliana* MGCs in a population of 1,152 accessions, with experimental and bioinformatic approaches. The MGC diversity was lowest in marneral gene cluster (one private deletion CNV) and highest in the arabidiol/baruol gene cluster where 811 accessions had gene gains or losses, however, there were no

presence/absence variations of the entire clusters. We found that the compact version of thalianol gene cluster was predominant in Arabidopsis and more conserved than the noncontiguogus version. In arabidiol/baruol cluster we found a large insertion in 35% of analyzed accessions, that contained duplications of the reference genes *CYP705A2* and *BARS1*. The *BARS1* paralog, which we named *BARS*2, encoded a novel oxidosqualene synthase. Unexpectedly, in accessions with the insertion, the arabidiol/baruol gene cluster was expressed not only in roots but also in leaves. Additionally, they presented different root growth dynamics and were associated with warmer climates compared to the reference-like accessions. We also found that paired genes encoding terpene synthases and cytochrome P450 oxidases had higher copy number variability compared to non-paired ones.

Conclusions: Our study highlights the importance of intraspecies variation and nonreference genomes for dissecting secondary metabolite biosynthesis pathways and understanding their role in adaptation and evolution.

Keywords copy number variation, biosynthetic gene cluster, secondary metabolism, coexpression, oxidosqualene cyclase, GWAS

BACKGROUND

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 growth and development. However, they are often multifunctional and may act both as plant growth regulators and be engaged in primary metabolism [1]. These chemicals also provide benefits by protecting the plant against pathogenic bacteria, insects, fungi or herbivores. The ability to produce particular types of compounds is usually restricted to individual species or genera. Therefore, they are enormously diverse and have a wide range of biological activity [2]. Secondary metabolites and the herbs producing them have been widely used in natural medicine and the production of pharmaceuticals, cosmetics or nutrients. Accordingly, there is a continuous interest in identifying new plant-derived bioactive compounds in elucidating their biosynthesis pathways and determining their biological function *in vitro* and *in vivo* [3–7].

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 the so-called operons [2]. Similar gene organization units named biosynthetic gene clusters or metabolic gene clusters (MGCs) have been recently found in numerous plant species. MGCs are classically defined as a group of three or more genes, which 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 [8–10]. 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 cluster genes. The contribution of other enzymes encoded by peripheral genes (located outside the MGC) and

connections between different MGC pathways result in some diversification of the biosynthetic pathway products [9,11]. Currently, there are over 30 known MGCs in plants from various phylogenetic clades and new ones 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.

Plant MGCs are thought to have arisen by duplication and subsequent neo- or sub-functionalization of genes involved in primary metabolism which might have been followed by the recruitment of additional genes to the newly forming biosynthetic pathway [12]. MGCs are frequently located within dynamic chromosomal regions, e.g., sub-telomeric, centromeric or rich in transposable elements (TEs). The possibility of bringing together the beneficial sets of genes by recombination, translocation, copy number changes and rearrangement may be higher in these regions than in the rest of the genome, thus promoting MGC formation. However, the same factors may also contribute to the accumulation of genetic changes in already preformed MGCs, that might be further shaped by selection and evolution. Therefore we hypothesized that plant MGCs display high intraspecific diversity. Exploring this diversity may help us to better understand the biology of MGC-encoded biosynthesis pathways and identify new MGCs or their unknown genetic components. To verify our hypothesis, we evaluated the structural variation of MGCs in *Arabidopsis thaliana* (hereafter Arabidopsis) with a combination of bioinformatic and experimental approaches. All known Arabidopsis MGCs are related to triterpene metabolism. Triterpenes constitute a large and diverse group of natural compounds derived from 2,3-oxidosqualene cyclization in a reaction catalyzed by oxidosqualene cyclases (OSCs) [13]. Five out of 13 OSC genes found so far in the Arabidopsis genome encode the MGC “signature” enzymes – synthases, which produce the triterpene scaffolds: thalianol (*THAS1*), marneral (*MRN1*), tirucalladienol (*PEN3*), arabidiol (*PEN1*) and baruol (*BARS1*). These genes are clustered along with the genes encoding enzymes for the downstream steps of the biosynthesis pathways, mostly cytochrome P450 oxidases (CYPs) and acyltransferases.

They altogether form four MGCs: three on chromosome 5 and one on chromosome 4. Here we present the extent of intraspecies diversity of these MGCs in a population of over 1,000 natural accessions. We focused on gene duplications and deletions since they may directly alter the genetic composition of MGCs and thus have a strong possible impact on the MGC-encoded metabolic pathways.

RESULTS

Arabidopsis MGCs are located in highly variable regions, but only two of them display substantial gene copy number polymorphism

We started our analysis from searching both the literature and published gene coexpression datasets to include confirmed and putative gene members in each MGCs' boundaries, as briefly summarized below and in **Additional file 1: Table S1**. Next, we aligned each MGC with the common copy number variants (CNVs), previously identified by the analysis of short read whole genome sequencing (WGS) data, to assess their intraspecies diversity **(Fig. 1A, Additional file 1: Table S2**) [14].

Thalianol gene cluster was initially identified as a set of four "core" genes: thalianol synthase gene *THAS1*, two CYPs, *CYP708A2* and CYP705A5, as well as BAHD acyltransferase gene family member *AT5G47980,* the latter two being separated from each other by a noncoding transcribed locus *AT5G07035* [15,16]. Recently, additional genes required for thalianol conversion to thalianin were identified [11]. One of them, *AT5G47950,* encodes an acyltransferase and is located close to *AT5G47980* (separated by two intervening genes, *RABA4C* and *AT5G47970*) and the other two (oxidoreductases *SDR4* and *AT1G66800*) are located on different chromosomes and were not included in our analysis. Accordingly, the estimated size of the entire thalianol gene cluster in the reference (Col-0) genome is ~45 kb (**Additional file 2: Fig. S1A)**. We found that this MGC was overlapped by six CNVs of various sizes (0.6 kb to 240.7 kb). The smallest one was located in the

intergenic region. The largest CNV encompassed the whole cluster and additionally extended upstream by 195 kb (regarding plus genome strand). The remaining variants were from 10.1 to 26.6 kb in size and each overlapped with some but not all thalianol cluster genes. CNV_18605 entirely overlapped with *THAS1* and *CYP708A2* and partially with *CYP705A5*. CNV_18604 completely covered *AT5G47980*, *CYP705A5* and the noncoding locus *AT5G07035* located between them. CNV_18606 and CNV_18608 covered single genes, *CYP708A2* and *THAS1*, respectively.

Marneral gene cluster is ~35 kb in size and is the most compact plant MGCs described to date (**Additional file 2: Fig. S1B)**. It is made up of three genes: *MRN1* (encoding marneral synthase), *CYP71A16* (encoding marneral oxidase) and *CYP705A12* (function unknown) [17–19]. *MRN1* is separated from the rest of the cluster by intervening noncoding transcribed loci *AT5G00580*, *AT5G06325* and *AT5G06335*. We found 15 CNVs, each smaller than 10 kb, scattered across this MGC region and together covering nearly 52% of its length. However, the CNV arrangement was strikingly different from any other MGC, since all variants were intergenic and did not overlap with the marneral cluster genes.

The tirucalladienol gene cluster includes five members. Of them, *PEN3* encodes triterpene synthase involved in the production of tirucalla-7,24-dien-3β-ol [20]. Three other genes in this cluster belong to the *CYP716* family. Two of them, *AT5G36130* and *CYP716A2*, have been recently postulated to jointly form one locus and encode a single cytochrome protein with unknown function [21]. The third one, *CYP716A1*, encodes an enzyme that is involved in the hydroxylation of tirucalla-7,24-dien-3β-ol when coexpressed with *PEN3* in *Nicotiana benthamiana* [22]. *SCPL1* - the last putative gene in the tirucalladienol cluster, was identified based on its co-expression with *PEN3* [22,23]. It encodes an uncharacterized acyltransferase. The contiguity of this MGC is interrupted by the presence of intervening genes *CCB3*, *AT5G36125*, *HCF109* and *AT5G36160* and the noncoding locus *AT5G05325*. The estimated size of the tirucalladienol cluster region is ~47 kb (**Additional**

file 2: Fig. S1C). This MGC overlapped with 17 CNVs, which together covered nearly 80% of its length. The longest variant, CNV_17742 (69.7 kb), overlapped with only the 5' end of the cluster, covering *CYP716A1*, *AT5G36130* and the intervening genes between them. CNV_17765 was 19.1 kb in size and overlapped with *CYP716A2*, *PEN3* and a nearby intervening gene. Six CNVs partially overlapped with one cluster gene only (*CYP716A1*, *AT5G36130*, *CYP716A2* or *PEN3*) and sometimes extended to adjacent intervening genes. The remaining variants were small (0.6 kb to 3.4 kb in length) and entirely located in the intergenic regions. It is worth to notice that upstream of the tirucalladienol gene cluster, a region genetically divergent from the surrounding genomic segments, called a hotspot of rearrangements, was previously discovered [24]. Another, smaller hotspot of rearrangements was found between *CYP716A1* and *AT5G36130*. The presence of these two divergent regions stayed in agreement with the higher frequency of CNVs in this location. On the contrary, the 3' region of the tirucalladienol gene cluster, including *SCPL1*, was not overlapped by any CNV.

The arabidiol/baruol gene cluster is the only known MGC located on Arabidopsis chromosome 4. It is a complex region, which encompasses two closely located OSCs (*PEN1* and *BARS1*), sharing 91% similarity at the amino acid level. *PEN1* encodes a synthase involved in the production of arabidiol [25]. Adjacent to *PEN1* is *CYP705A1*, involved in arabidiol degradation upon jasmonic acid treatment, which leads to releasing a volatile homoterpene, (E)-4,8-dimethyl-1,3,7-nonatriene (DMNT) [26]. When *PEN1* and *CYP705A1* were coexpressed in *N. benthamiana* along with distantly located genes *AT1G66800*, *SDR4* and *AT5G47950* (which are also involved in the thalianol biosynthetic pathway), they acted together to convert arabidiol to arabidin [11]. *BARS1* encodes a multifunctional cyclase, which produces baruol as its main product [27]. The genomic region surrounding PEN1 and BARS1 contains seven additional CYPs (*CYP702A2*, *CYP702A3*, *CYP705A2*, *CYP705A3*, *CYP705A4*, *CYP702A5*, *CYP702A6*) and two acyltransferase genes (*AT4G15390*, *BIA1*). Their role in the arabidiol/baruol metabolism has not been determined.

However, they displayed coexpression with either *PEN1* or *BARS1* and are included in the MGC [16,23]. Accordingly, the estimated size of the arabidiol/baruol gene cluster region is 83 kb (**Additional file 2: Fig. S1D)**. There are few intervening loci there, including a protein-coding gene *CSLB06*, two pseudogenes *CYP702A4P* and *CYP702A7P* and one novel transcribed region *AT4G06325*. Arabidiol/baruol gene cluster overlapped with 24 CNVs (0.6 kb to 21 kb), which covered about 53% of its length. CNVs were grouped in three distinct regions, separated by invariable segments. In the first variable region, which also co-localized with a hotspot of rearrangement [24] there were eight CNVs and three partially overlapped with *CYP702A2* or *CYP702A3*. In the second variable region there were four CNVs, three of which had some overlap with cluster genes (*CYP705A2, CYP705A3, BARS1*). Notably, only CNV_14660 fully overlapped all three of them. The third variable region was covered by 12 CNVs, but they were mostly intergenic and only three overlapped with genes (*CYP702A5* or *CYP702A5* and *CYP702A6*). *CYP705A1, PEN1, CYP705A4, AT4G15390* and *BIA1* were not covered by any common CNV.

A large diversity of CNVs and their overlap with MGC genes indicated there might be considerable differences in triterpene biosynthesis pathways encoded by various Arabidopsis accessions. In order to assess this putative diversity, we analyzed the deletions and duplications of individual genes. We retrieved copy number data for 31 genes (clustered and intervening genes), each from 1,056 accessions (listed as RD dataset in **Additional file 1: Table S3**) [14]. We validated and supplemented these data with multiplex ligation-dependent amplification (MLPA) and droplet digital PCR-based genotyping (ddPCR) assays. MLPA covered 13 genes from thalianol, marneral and arabidiol/baruol clusters and were performed for 232 accessions, of which 136 were in common with RD analysis (MLPA dataset in **Additional file 1: Table S4; Additional file 2: Fig. S2-S3**) while ddPCR covered four genes from thalianol cluster and were performed for 20 accessions, all in common with RD and/or MLPA analyses (ddPCR dataset in **Additional file 1: 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 (**Additional file 1: Table S6**). Out of 1,784 genotypes assigned with two or three approaches, 1,763 were fully concordant (98.8% congruency) and 19 out of 21 discrepancies were resolved manually (see **Additional file 1: Table S7**, **Additional file 2: Fig. S4** and **Methods**). The combined genotyping data were further used to asses and compare the levels of MGC variation.

Among the 1,152 accessions, only 28.6% had no gene gains or losses in any MGC (**Fig. 1B**). This included 65% of assayed accessions from the German genetic group and 39% of accessions from the Central Europe group. In turn, 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 of them, all genes in the tirucalladienol cluster and some in the arabidiol/baruol cluster were not genotyped. Altogether, 19 MGC 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 (**Fig. 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, six genes were multiallelic (with both gains and losses) and for 11 genes we observed only copy losses. As expected, genes overlapping the previously defined CNV regions showed the highest variability in our assays. Remarkably, we did not observe complete loss or gain of 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 thalianol gene cluster is more common and has lower variation than the reference-like noncontiguous version

Survey in 1,152 accessions with a combination of RD, MLPA and ddPCR approaches revealed 49 accessions with gene deletions and 5 accessions with gene duplications in the thalianol gene cluster.

Except for *AT5G47950*, which was invariable, all genes in this cluster were affected by CNVs, which followed five distinct patterns (**Fig. 2A**). Three of these patterns (deletions only) were previously identified by browsing the polymorphisms listed in 1001 Genomes Project database [28,29]. These deletions were found in 21 accessions (~2%) and 19 of them were also detected in our study (we did not analyze the other two accessions). The most common variant (variant A) was the deletion of a region encompassing *AT5G47980* and *CYP705A5*, combined with the deletion of *THAS1*. We detected it 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. The first one (variant B) was a large deletion spanning *AT5G47980*, *CYP705A5* and *CYP708A2*. We found it in two accessions from Germany (Bch-1, Sp-0), one from Italy (Etna-2) and one from Spain (IP-Mon-5). The second rare variant (variant C) was a deletion of a single gene, *CYP708A2*, which we found in four accessions from Spain (Can-0, Ped-0, IP-Her-12 and Nac-0) and one from Portugal (IP-Mos-1). Of note, four accessions in this group represented the genetically differentiated lineage of Relicts. We also found a new type of deletion in two Relicts (IP-Rel-0 and Con-0) and one non-Relict (IP-All-0), all originating from Spain. This deletion (variant D) spanned *CYP705A5*, *CYP708A2* and *THAS1* and was detected by RD and MLPA assays as well as confirmed by read pair analysis of WGS data **(Additional file 2: Fig. S5**). The last variant we identified (variant E) was a duplication of acyltransferase gene *AT5G47980* in four accessions from Italy (Mitterberg-1-179, Mitterberg-1-180, Mitterberg-1-183, Mitterberg-2-185) and one from Greece (Olympia-2). The duplication was detected with both RD and MLPA assays and we confirmed it by sequence analysis of the *de novo* genomic assembly for Mitterberg-2-185, which we generated with Nanopore sequencing data **(Additional file 2: Fig. S6**). The duplicated region was ~3 kb in size. It spanned *AT5G47980* and its flanks (0.5 kb upstream and 0.7 kb downstream). In Mitterberg-2-185, the sequence copies were positioned in tandem and differed from each other by only two mismatches and a 1-bp gap. Accordingly, duplicated genes encoded identical proteins with 99% identity to the reference *AT5G47980* acyltransferase. Although they

were shorter than the reference protein (404 aa versus 443 aa), they possessed a complete transferase domain (pfam02458).

AT5G47950 - the gene encoding another acyltransferase in the thalianol gene cluster was stable in copy number. However, we detected a chromosomal inversion (with respect to the reference genome orientation), spanning this gene and two intervening genes, *RABA4C* and *AT5G47970*, in Mitterberg-2-185 *de novo* assembly. This inversion resulted in a more contiguous and compact organization of the entire cluster (**Fig. 2B**). Similar inversions were previously detected in PacBio-based assemblies of 17 other accessions (out of 22 analyzed) [29], which indicated that the compact version of thalianol gene cluster might be predominant in Arabidopsis. To verify this possibility, we set up a bioinformatic pipeline for detecting genomic inversions based on paired-end WGS data analysis of 997 accessions. In 649 of them (65%), including Mitterberg-2-185, we found inversions 12.8 kb to 15.4 kb in size, spanning *AT5G47950*, *RBAA4C* and A*T5G47970* genes (**Additional file 1: Table S8**). Out of 15 accessions in which inversions were previously confirmed by PacBio sequencing [29], we correctly identified 12, which indicated good sensitivity of our method. Based on this result, we concluded that the compact thalianol gene cluster was present in most of the Arabidopsis population. It was dominant among the accessions from South and North Sweden genetic groups as well as from Asia group (83.6% to 88.9%) (**Fig. 2C**). It was also more frequent in Central and Western Europe groups as well as in Germany group, except for U.S.A accessions, in which we mainly observed the discontiguous version. Apart from USA, the reference-like cluster was slightly more abundant also in Spain genetic group. In Relicts – the group of individuals considered to represent different glacial refugia, genetically diverted from the majority of worldwide accessions [28], there was similar frequency of discontiguous and compact versions (12 and 10 accessions, respectively).

Interestingly, the CNV frequency substantially differed between the accessions with and without inversion (**Fig. 2D, E**). Copy number changes affected only 1.1% of accessions with the compact cluster version, including four out of five cases of variant E (*AT5G47980* duplication) and three out of four cases of variant B (*AT5G47980*, *CYP705A5* and *CYP708A2* deletion). The remaining CNVs, including all deletions spanning *THAS1*, were found exclusively among the accessions with the reference-like cluster type. Altogether, 12.7% of accessions with the discontiguous cluster type were affected by CNVs.

Marneral and tirucalladienol gene clusters display little structural variation

Copy number analysis of 1,152 accessions combined with RD and MLPA approaches confirmed exceptionally low variability of marneral cluster genes. The sole variant we detected was 1.2 kb in size and spanned the first exon of *CYP705A12* gene with an unknown function in the marneral biosynthetic pathway [19]. We found this deletion in one accession, Mir-0, and we confirmed it by Sanger sequencing (**Additional file 2: Fig. S7**). The predicted truncated protein product was only 96 amino acids long and lacked the signal peptide and the N-terminal part of the cytochrome P450 conserved domain pfam0006. Apart from that, we did not detect any common gene duplications or deletions within this MGC.

Likewise, we observed low copy number polymorphism for tirucalladienol cluster genes, which were analyzed in 1,056 accessions with RD approach. There were no copy number changes of *CYP716A1* and *SCPL1*. Low variation (1.4%) was detected only in the region spanning adjacent genes *AT5G36130*, *CYP716A2* and *PEN3*, where deletions or duplications occurred in a small number of accessions from Czech (2), Spain (5), Georgia (3), Bulgaria (1), Germany (1), Netherlands (1), United Kingdom (1) and Sweden (1) (**Additional file 2: Fig. S8**). In seven accessions, all three genes were absent. This included Ty-1, for which we analyzed de novo genomic assembly and confirmed the gene loss, as well as Draha2 and Jl-3, for which we observed

numerous read pairs spanning the deletion site. In seven other accession, *AT5G36130* or *CYP716A2* or both were missing, including Dolna-1-40, for which we confirmed *AT5G36130* loss with the Nanopore sequencing and analysis of *de novo* genomic assembly (**Additional file 2: Fig. S9**). According to a recent study, *AT5G36130* and *CYP716A2* gene models are misannotated and they jointly encode a single full-length protein of the CYP716A subfamily with cytochrome oxidase activity [21]. Therefore, it should be noted that all 15 accessions with CNVs in the tirucalladienol gene cluster, lacked a functional *CYP716A2* gene.

Structural variation analysis of arabidiol/baruol gene cluster reveals a novel OSC gene in the Arabidopsis genome

Arabidiol/baruol gene cluster, which we analyzed with a combination of RD and MLPA approaches, was the most heterogeneous of all MGCs. Consistent with the segmental CNV coverage, there were apparent differences in the variation frequency between the clustered genes. Two *CYP702A* family members located at the cluster's 5'end were highly variable. *CYP702A2* was duplicated in 50 accessions and *CYP702A3* was deleted in 564 accessions, including about 70% of all analyzed accessions from Sweden and Spain. On the contrary, CYPs located at the 3' end of the cluster showed little variation. *CYP702A5* was deleted in 35 accessions, *CYP705A4* was deleted in 2 accessions and no changes in *CYP702A6* copy number were observed. The acyltransferase gene *BIA1* was missing in only one accession and another acyltransferase, *AT4G15390*, was invariable in copy number.

The two OSCs, *PEN1* and *BARS1*, differed in variation between each other but had similar variation frequencies as their neighboring CYPs. *CYP705A1* and *PEN1*, both implicated in the arabidiol biosynthesis pathway, were stable in copy number, in agreement with their localization outside any common CNVs. Both genes were deleted only in IP-Deh-1 accession (Spain), while in Qui-0 (also Spain) *PEN1* was deleted and *CYP705A1* status was not evaluated. Additionally, in Kyoto both

genes were partially deleted, which we confirmed by analyzing available PacBio-based *de novo* genomic assembly of this accession [24].

BARS1, *CYP705A2* and *CYP705A3* were all located within the central CNV segment. All three genes were deleted in some accessions originating from Sweden. We also observed smaller deletions or duplications in this segment, of which the most remarkable was the duplication of *CYP705A2*, detected in 433 (37.6%) accessions. However, genotypic data for *CYP705A2* and *BARS1* were noisy and we reasoned there was more variation in this region than could be revealed by our standard genotyping. Therefore, we manually inspected WGS reads which mapped in this region (examples are presented in **Additional file 2: Fig. S10**). We noticed that in most accessions *BARS1* lacked the largest intron, where *ATREP11* TE (RC/Helitron super family) is annotated. This explained the decrease in the level of RD values for *BARS1* compared to other genes (see **Additional file 2: Fig. S3**). Surprisingly, we also observed a mix of reads which mapped 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 get more information on their diversity. As expected, 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 the mismatches originated from different loci (gene duplicates), with similarity to *CYP705A2* and *BARS1*, respectively. WGS reads from the accessions with the duplications would then generate pseudo-heterozygous SNP calls when mapped to the reference gene models. Additionally, since we were able to detect copy gain of *CYP705A2* in hundreds of accessions but found only one duplication of *BARS1*, we concluded that the sequence differences between *BARS1* and its duplicate prevented its detection by RD or MLPA assays. In support of our hypothesis, in accessions with *CYP705A2* duplication, we found some heterozygous SNPs both at *CYP705A2* and *BARS1* loci. Out of 394 accessions with *CYP705A2* duplication, 357 (90.6%) had at least one heterozygous SNP called in this locus, compared to only 69 out of 646 (10.7%) accessions without

changes in *CYP705A2* copy number (Wilcoxon rank sum test with continuity correction, p.value $<2.2\times10^{-16}$) (**Additional file 2: Fig. S11A**). Moreover, the numbers of heterozygous SNPs at *CYP705A2* and *BARS1* were strongly correlated (Pearson's r = 0.86) (**Additional file 2: Fig. S11B**). We also reasoned that the absence of reference *CYP705A2* and *BARS1* genes and the presence of their nonreference homologs in some accessions, would result in nonzero read coverage and homozygous instead of heterozygous SNP calls. Indeed, we observed such polymorphisms in 14 accessions with intermediate RD values for *CYP705A2* (RD_{mean} = 1.5) and *BARS1* (RD_{mean} = 0.6) and with loss of *CYP705A3* (RD_{mean} = 0), which stayed in agreement with our hypothesis.

To identify the hidden *BARS1* duplicate, we analyzed PacBio-based genomic assemblies of seven accessions: An-1, Cvi-0, Kyoto, Ler-0, C24, Eri-1 and Sha [24], four of which have been also genotyped in our study (**Fig. 3A**). In our assays, An-1 and Kyoto had no changes in *CYP705A2* and *BARS1* copy number and no heterozygous SNPs were found, while in Cvi-0 and Ler-0 *CYP705A2* was duplicated and heterozygous SNPs in both genes were present. We re-annotated the entire cluster region in each accession and compared it with the reference (**Additional file 1: Table S9**). In six accessions, *BARS1* lacked the largest intron present in the reference gene model, in agreement with our observations (**Additional file 2: Fig. S12**). In Cvi-0, Eri-1 and Ler-0 we identified a nonreference gene, encoding a protein with ~91% identity to baruol synthase 1 (**Additional file 2: Fig. S13**). In C24, this gene was present but interrupted by ATCOPIA52 retrotransposon insertion, resulting in two shorter ORFs, of which only one has been annotated previously. Based on phylogenetic analysis and sequence comparison with *BARS1* ortholog from a related species *Arabidopsis lyrata*, we concluded that the identified gene was a *BARS1* duplicate and we named it *BARS2* (**Fig. 3B**). At the nucleotide level, the differences between *BARS1* and *BARS2* accumulated in the introns, which likely prevented mapping of many *BARS2*-derived reads to *BARS1* locus. The differences in the exons very well matched the heterozygous SNP positions, which we observed (**Additional file 2: Fig. S14**). MLPA probe targeting *BARS1* was located in a highly divergent

region which prevented it from binding to the *BARS2* genomic sequence. These observations explained why we did not detect *BARS1* duplication in our assays.

Full-length 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 (**Fig. 3C**). We performed three-dimensional (3D) modelling of baruol synthase 1 from Col-0 (reference) and Cvi-0 (which is a Relict) as well as putative baruol synthase 2 from Cvi-0, using a machine learning approach, implemented in ColabFol software (**Additional file 2: Supplementary Methods**) and found that all structures were highly similar (**Additional file 1: Table S10).** Additionally, we superposed our 3D models with the experimental,crystal structure of human OSC, available in a complex with its reaction product lanosterol [30,31]. This approach allowed us to identify potential substrate-binding cavities (**Fig. 3D; Additional file 3**). Notably, the catalytic aspartate residue D455 present in human cyclase had its counterparts in plant OSCs: D493 in Col-0 baruol synthase 1 (isoform NP_193272.1) and D490 in Col-0 baruol synthase 1 (isoform NP_001329547.1), Cvi-0 BARS1 (ATCVI-4G38020) and Cvi-0 BARS2 (ATCVI-4G38110). This indicated that *BARS2* encoded a novel, so far uncharacterized OSC in the Arabidopsis genome. As expected, we also found *CYP705A2* duplicates in C24, Cvi-0, Eri-1 and Ler-0 assemblies (84% identity at the nucleotide level, 79% identity and 88% similarity at the protein level) (**Additional file 2: Fig. S15**). This gene, which we named *CYP705A2a*, was adjacent to *BARS2* and located on the minus strand of the large genomic sequence insertion between *CYP702A6* and *BIA* genes. The inserted sequence also contained a ~5 kb long interspersed nuclear element 1 (LINE-1) retrotransposon and some shorter, undefined ORFs. The insertion increased the size of the entire arabidiol/baruol gene cluster MGC by 21-27 kb in each accession.

Copy number changes of *CYP705A2* and *BARS1* are associated with climatic gradient and root growth variation

In the next step, we combined the genotyping results with the SNP analysis for all accessions, in order to determine the copy number status of *CYP705A2* and *BARS1* and their duplicates (see **Additional file 1: Table S11** and **Methods** for details). We distinguished four groups of accessions (**Fig. 4A, Additional file 2: Fig. S10**). The reference-like group, with *CYP705A2* + *BARS1* gene pair presence and *CYP705A2a* + *BARS2* absence (PP-AA), was the largest one, with 628 accessions. Nearly one-third of the population (326 accessions) had both *CYP705A2* + *BARS1* and *CYP705A2a* + *BARS2* pairs (PP-PP). Apart from that, we separated two intermediate groups with the local range of occurrence. The first one included 14 accessions with only the nonreference *CYP705A2a* + *BARS2* gene pair (AA-PP). Six of these accessions originated from Azerbaijan; the remaining ones were from Spain, Bulgaria, Russia, Serbia and U.S.A. The last group, where all four genes were absent (AA-AA), comprised 15 accessions. The majority of them (13) were North Swedish accessions from the Bothnian Bay coast collection site. In the remaining 73 accessions, the gene pair status was uncertain and left unassigned. We confirmed the accuracy of group assignments for accessions with available *de novo* genomic assemblies: An-1, Kyoto, Mitterberg-2-185 and Kn-0 (representing PP-AA group) as well as Cvi-0, Ler-0, Dolna-1-40 and Ty-1 (representing PP-PP group). Additionally, we performed PCR amplification with gene-specific primers and genomic DNA template for a subset of 36 accessions from all four groups. The PCR results confirmed the differences between these groups (**Additional file 2: Fig. S16**). We noted that we did not detect *BARS2*-specific products in many samples from the AA-PP group. However, the band for *CYP705A2a* was clearly present in these accessions. We suppose that the *BARS2* sequence might further diverge in this group. Genome sequencing and assembling for representative AA-AA and AA-PP accessions will be needed to further evaluate the structure of arabidiol/baruol cluster region in these two minor groups.

The accessions with the nonreference gene pair (AA-PP and/or PP-PP) were present among all genetic groups but with a different frequency (**Fig. 4B**). They dominated among the Relicts (81%),

Spain (60%) and Italy/Balkan/Caucasus (89.6%) groups but constituted the minority of accessions in the groups from north and east margins of the species range (North Sweden 18.6%, South Sweden 16%, Asia 9.4%). They were also mostly absent among U.S.A accessions. The presence of *CYP705A2a* and *BARS2* genes in the majority of Relicts suggested that the gene duplication event preceded the recent massive species migration, which took place in the post-glacial period and shaped the current Arabidopsis population structure [32]. We visualized the four groups in the principal component analysis (PCA) plots, generated with genome-wide bi-allelic SNPs, by applying varying linkage disequilibrium (LD) parameter (**Additional file 2: Fig. S17**). At low LD, when the contribution of the ancestral alleles to PCA was highest, we observed a clear convergence of the PC1 and PC2 components with the presence/absence of duplicate gene pair (**Fig. 4C)**. We also observed that the accessions possessing *CYP705A2a* and *BARS2* genes were associated with 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)) (**Fig. 4D**). In case of two most abundant groups, namely PP-PP and PP-AA, this difference was noticeable even within individual countries and was significant for Germany, Spain and Italy (**Additional file 2: Fig. S18**). The reverse trend was observed only in Russia, where PP-AA accessions were in great excess (88%) and France. However, it is worth to notice that PP-AA accessions outnumbered PP-PP accessions in the Pyrenees, Alps and Tian Shan mountain ranges, as visualized in the accession distribution map [33].

Motivated by the above observations, we decided to evaluate the possible association between arabidiol/baruol gene cluster variation and environmental conditions. Since the climate is a substantial selection factor, we also checked for the phenotypic variability between PP-AA and PP-PP accessions. To this end, we performed two-group comparisons of 516 phenotypic and climatic variables retrieved from the Arapheno database [34] and focused on these which significantly differed between both groups (Wilcoxon rank sum test with continuity correction, p.value <0.05)

(**Additional file 1: Table S12**). The most prominent were 88 variables describing the climatic conditions at the accessions' sites of origin (**Fig. 5A**), especially maximal and minimal temperature conditions, precipitation and evapotranspiration [35]. These results confirmed that the differential geographical distribution of accessions between PP-AA and PP-PP groups was strongly associated with the climatic gradient.

Apart from the climate data, 40 diverse phenotypes varied significantly between both groups. At this point, we cannot exclude that these differences might be influenced by another genetic factor, independent from the arabidiol/baruol cluster structure. For example, the different flowering times between both groups could be related to the known latitudinal allelic variation of flowering time-associated genes [36]. However, we paid special attention to root growth-related phenotypes since all Arabidopsis MGCs are considered to have root-specific expression. Moreover, the mutation of *AT5G47950* acyltransferase gene (involved in the thalianin and arabidin biosynthesis) was previously shown to negatively impact the root growth [37]. We observed significant differences in root growth dynamics between PP-AA and PP-PP groups during the first week after germination [38]. More specifically, roots of PP-PP accessions elongated slower compared to PP-AA accessions (**Fig. 5B**). Additionally, PP-PP accessions showed a significantly lower rate of root organogenesis from explants, compared to PP-AA accessions, under one of three growth conditions tested in another study (**Fig. 5C**) [39]. We also checked for the genetic interactions between thalianol and arabidiol/baruol clusters for these phenotypes since the distribution of discontiguous and compact versions of this cluster was also related to geography. However, despite the strong association between the thalianol gene cluster contiguity and the latitude, the variation in thalianol gene structure did not explain the variation in root growth phenotypes, contrary to the presence vs. absence variation of genomic insertion with *CYP705A2a* and *BARS2* genes in the arabidiol/baruol gene cluster (**Additional file 2: Fig. S19**).

In a highly structured population, such as that of Arabidopsis, the correlation of allele frequency with geography and phenotypic traits may not only result from the selection but may be also influenced by the genetic similarities of neighboring populations, resulting from the gene flow and genetic drift [40]. Therefore, we next applied a linear mixed model in a Genome-Wide Association Study (GWAS) on the same 516 phenotypes, in order 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, which we supplemented with a SNP-like encoded information about *CYP705A2* and *BARS1* duplication status (only PP-AA and PP-PP accessions were taken into consideration). Although this comprehensive screening showed no statistical significance for any variable we tested, still the lowest p-values were obtained for PP-AA and PP-PP groups' association with the climatic data and root organogenesis phenotypes (**Fig. 5D, Additional file 1: Table S12**).

In the reference accession Col-0, all genes in the arabidiol/baruol cluster are low-expressed and active almost exclusively in roots (**Additional file 2: Fig. S20A**). In search of the additional links between arabidiol/baruol gene cluster structure and phenotypic variation, we investigated whether there were any differences in *CYP705A2* and *BARS1* expression profiles between PP-AA and PP-PP accessions or whether the expression of *CYP705A2a* and *BARS2* differed from their paralogues in PP-PP accessions. We downloaded raw RNA-Seq data from NCBI/SRA database for studies in which Col-0 and Cvi-0 accessions were grown in parallel under the same (standard) conditions and used for gene expression profiling in roots, shoots and leaves [41,42]. For each accessions we mapped the data to the respective (Col-0 or Cvi-0) annotated genome and compared the expression profiles of the genes in the arabidiol/baruol cluster (**Fig. 5E; Additional File 1: Table S13**). In both accessions, the clusters were silenced in shoots, except for acyltransferase gene *AT4G15390*, which displayed low activity in Cvi-0. Also, in both accessions, the clusters were active in roots, where the expression of *AT4G15390* was much stronger than the remaining genes. In Cvi-0, genes located in

the genomic insertion (*CYP705A2a*, *BARS2* and *ATCVI-4G38100* which encodes a protein with partial similarity to acyltransferase) were also expressed, although at lower level compared to the rest of the cluster. Surprisingly, we detected expression of arabidiol/baruol clusters genes in leaves of Cvi-0, which indicated different tissue specificity of its arabidiol/baruol gene cluster compared to Col-0. Outside of the genomic insertion, the clustered gene expression level was much lower compared to the roots, and the transcripts of *CYP705A2*, *CYP705A3* and *BARS1* were barely detectable. However, in Cvi-0, the transcription levels of *ATCVI-4G38100*, *CYP705A2a* and *BARS2* genes localized in the insertion were similar in the leaves and roots. The expression of *CYP705A2a* and *BARS2* in leaves of this accession was higher than the expression of *CYP705A2* and *BARS1*. Accordingly, it can not be excluded that the metabolic products of arabidiol/baruol gene cluster activity in the roots and leaves of Cvi-0 accession are not identical.

As we have shown above, a substantial fraction of the Arabidopsis population represents the PP-PP group. We wanted to check whether the gene expression profile we observed in leaves of Cvi-0 was ubiquitous among the accessions from this group. Therefore, we evaluated available gene expression data from the leaves of 552 accessions with known *CYP705A -BARS* duplication status, using normalized counts that were previously generated by mapping the reads from each accession to the reference genome (Araport 11 annotation) [41] and we compared the BARS1 expression level between AA-PP, PP-PP and PP-AA groups. Strikingly, it was significantly higher in accessions from the AA-PP and PP-PP groups, in contrast to 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) (**Fig. 5F**). This observation stayed in agreement with our predictions that some reads from *BARS2* transcripts mapped to *BARS1* locus, elevating the measured expression level in leaves of accessions from these two groups. To confirm that, we additionally re-mapped the raw RNA-Seq reads from four accessions: Ty-1 and Cdm-1 (both representing the PP-PP group) as well as Kn-0 and Sha (the PP-AA group) to their respective genomic assemblies, in which we annotated the genes of interest

(**Additional file 2: Figure S20B**). We analyzed the transcription levels of *CYP705A2*, *BARS1*, *CYP705A2a* and *BARS2* and confirmed that in the leaves of PP-PP accessions *BARS2* was expressed, while *BARS1* was not.

The proximity of terpenoid synthase and cytochrome P450 genes in the Arabidopsis genome is associated with increased copy number diversity

In many plant genomes, genes encoding terpenoid synthases (TSs, this includes 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 [22]. According to our results, TS-CYP pairs located in MGCs had similar (either high or low) diversity in copy number and the paired genes were frequently duplicated or deleted together. We wanted to check whether copy number polymorphism is a common feature of TS-CYP pairs in the Arabidopsis genome. We created a comprehensive list of 48 TSs and 242 CYPs in this plant, relying on the information from previous studies and dedicated databases [22,43–45]. We obtained the genomic localization for each gene and retrieved information about its copy number diversity among 1,056 accessions (**Additional file1: Tables S14-S15**). For 13 TSs, including *THAS1* and *BARS1*, we observed gains or losses in at least 1% of accessions. Among CYPs, only 33 CYPs showed variability in at least 1% of accessions and these CYPs represented three clans: CYP71 (the most abundant clan in Arabidopsis; 26 variable genes out of 151), CYP85 (6 variable genes out of 29) and CYP72 (1 variable gene out of 19). The remaining clans, including the second most abundant clan CYP86, showed very low variability. Next, for each TS, we selected all CYPs within +/- 30-kb distance, resulting in 38 pairs between 18 TSs and 27 CYPs. These pairs included genes from thalianol, marneral, tirucalladienol and arabidiol/baruol gene clusters and also TS-CYP pairs contained within other putative secondary metabolism clusters, as listed in plantiSMASH resource [10]. Subsequent group comparison revealed that TSs and CYPs occurring in pairs were more variable than their non-paired counterparts (Wilcoxon rank sum test with continuity correction p.value<0.01 for TSs,

p.value<0.001 for CYPs). Moreover, the CNV rates and duplication/deletion patterns of paired genes were similar for numerous TS-CYP gene pairs, not only the ones located in the thalianol, marneral, tirucalladienol and arabidiol/baruol gene clusters (**Additional file 2: Figure S21**).

DISCUSSION

According to our current understanding of the MGC formation phenomenon, nonrandom gene clustering in eukaryotes is linked with the highly dynamic chromosomal regions. Presence/absence variations, tandem and segmental gene duplications, recombination, gene shuffling and TE-mediated variations are considered as the main genetic drivers of metabolic profiles diversity and MGC evolution in plants [29,46-48]. Recently, population-scale studies highlighted the importance of intraspecies genetic variation in elucidating the role of these factors in the formation of gene clusters. Resequencing the genomes of nine opium poppy accessions with a range of alkaloid production profiles revealed that CNVs affected the majority of known benzylisoquinoline alkaloid pathway-related genes and their paralogs [49]. Pathway genes located within clusters were about three times more likely to exhibit CNVs than those outside of clusters. Morevover, the cultivars with similar CNV profiles in their benzylisoquinoline alkaloid pathway genes also tended to have similar profiles in the production of alkaloids, e.g. noscapine, morphine and codeine. In Arabidopsis, in the roots of accessions which lacked one member of the thalianol gene cluster (*CYP708A2*), the end product of thalianol biosynthesis pathway, thalianin, could not be detected [46]. In rice, genomic sequence analysis and metabolite-based GWAS performed in 424 accessions representing two subpopulations, *indica* and *japonica*, enabled identification of a new gene cluster *DGC7*, involved in the production of casbene-type diterpenoids, that are almost exclusively restricted to *japonica* varieties [50]. Phylogenetic analysis revealed that all three gene members of this cluster origined by duplications in *Oryza* genus, which was followed by their organizing into a cluster. Moreover, intact *DGC7* was was located in a selective sweep region of *japonica* subspecies and was highly enriched in this subpopulation, while in many *indica* accessions it was absent or

incomplete, suggesting positive selection of *DGC7* in *japonica* but not *indica* during domestication. These studies suggested that plant MGCs are dynamically evolving and the genetic mechanisms that originally lead to their formation may be captured at the intraspecies genetic variation level. Similar conclusions were drawn from the previous study of a filamentous fungus *Aspergillus fumigatus*, in which the secondary metabolic pathway genes are commonly organied into clusters. In that species, substantial MGC genetic diversity was observed among 66 strains. The main variants included, apart from point mutations and horizontal gene transfer, also the gain and loss polymorhpisms that affected both individual genes and the entire gene clusters [51]. The same mechanisms were previously hypothesized to contribute to the secondary metabolite gene cluster diversity across entire fungal classes and phyla, suggesting that the intraspecies analysis enabled the authors to "catch "the genetic factors involved in the MGCs macroevolution, "in the act". During evolution, new biochemical pathways are tuned and tested by many rounds of natural selection. Thus, the analysis of the intraspecies MGC variants, which are more recent than variants found in interspecies comparisons, may provide an important insight into the formation of the clustered gene architectures and the plant metabolic diversity, in a small evolutionary time frame.

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 Brassicaceae lineage ~23-43 Mya [19]. 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 the genes from CYP705, CYP708 and CYP702 clans, and with genes from the acyltransferase IIIa subfamily [46]. Bioinformatic studies also revealed that TSs and CYPs are paired in plant genomes more frequently than expected [22]. By CNV analysis in ~1,000 accessions we found that in Arabidopsis, the physical proximity of CYPs and TSs was associated with increased rates of these genes' duplications and deletions compared to non-paired ones. It 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 [10]. We also revealed that Arabidopsis MGCs had different levels of variation and these differences reflected the phylogeny of clade II OSCs that we analyzed. Of them, MRN1 is most divergent in amino acid sequence. It is also mono-functional, i.e., catalyzes the cyclization of 2,3-oxidosqualene into one specific product - marneral [17]. Functional studies indicated an important role of marneral synthase in Arabidopsis development. Knock-out *mrn1* mutant displayed round-shaped leaves, late flowering and delayed embryogenesis [18]. Consistent with these findings, *MRN1* was the only clustered OSC gene, which was not affected by deletions or duplications, in any accession. Also, the neighboring CYPs were stable in copy number, apart from a private partial deletion of *CYP705A12*. Our results indicate that marneral gene cluster is fixed in Arabidopsis genome.

Thalianol gene cluster, which has been most intensively studied so far, was the second most variable MGC in our analysis [11,16,37]. The first evidence for the copy number and structural diversity of this cluster comes from the study of Liu et al. [29], who analyzed the frequency of SNPs, small indels and gene deletions within the thalianol cluster genes. The authors showed that SNPs and indels were the most common polymorphisms, however the changes were mostly benign. They also identified large gene deletions in ~2% of the studied accessions. Since our approach was specifically focused on CNV analysis and was duplication-aware, it was more sensitive to large structural variations. As a result, we observed over two times higher rate of copy number alterations within the thalianol gene cluster (4.7%), with 49 accessions carrying gene deletions and 5 accessions with gene duplications. We identified two new CNV patterns: one of them was a large deletion of *CYP705A5*, *CYP708A2* and *THAS1* and the other one was *AT5G47980* duplication. Apart from that, by the large-scale analysis of chromosomal inversions, performed for nearly 1000 accessions, we validated earlier assumptions, based on only 22 accessions, that the compact (contiguous) cluster version is predominant in Arabidopsis. It is also better conserved compared to the discontiguous version. However, it remains to be investigated whether the tighter clustering of thalianol gene cluster leads

to altered gene expression or the diversification of the metabolite profiles and whether it may be advantageous in certain environmental conditions.

The arabidiol/baruol was the largest and the most variable MGC in Arabidopsis. This region comprises only few gene subfamilies, but it is significantly expanded compared to the sister species *A. lyrata*, including the presence of two highly similar OSCs, *PEN1* and *BARS1*, orthologous to *A. lyrata* gene LOC9306317. It is suggestive of recent duplications and stays in agreement with our observations of an exceptionally high rate of intraspecific gene gains and losses within this MGC. Of note, within this cluster there was a clear segmentation into variable and invariable gene blocks. *PEN1* and *CYP705A1* genes constituted the almost invariable segment, while the adjacent group of C*YP705A2*, *CYP705A3* and *BARS1* genes displayed high rate of copy number changes. Segmentation may result from the ongoing process of selection-driven fixation of the arabidiol sub-cluster. The products of *PEN1* and *CYP705A1* are involved in response to jasmonic acid treatment and infection with the root-rot pathogen *Pythium irregulare* [26]. Moreover, arabidiol may be further converted to arabidin in the pathway involving acyltransferase encoded by *AT5G47950*. Crispr mutants with disrupted *AT5G47950* gene had significantly shorter roots than the wild-type plants and arabidin did not accumulate in these roots [37]. *AT5G47950* is the only member of the thalianol gene cluster which was invariable in copy number, which may indicate the biological significance of the arabidin biosynthesis pathway. Interestingly, *A. lyrata* is able to convert apo-arabidiol (the product of arabidiol degradation) into downstream compounds, despite the lack of arabidiol synthase [54]. This indicates modularity of the triterpene (and possibly other metabolite) pathways, which might enable the plants to easily assembly a biosynthesis network and increase the repertoire of produced secondary metabolites. Understanding the complexity of this network will require further in-depth investigation and may be supported by the intraspecies analysis of MGC variants.

The diversity of OSCs determines the initial diversity of 2,3-oxidosqualene cyclization products generated by the plant [55]. So far, the Arabidopsis genome was considered to include 13 OSCs and it has been over a decade since the biochemical profile of *PEN3* - which was thought to be the last unexamined one - was described [20]. Here we report the discovery of *BARS2* gene, which was found in numerous accessions but was absent from Col-0, hence it is also missing from the current reference genome. We performed sequence and expression analyses and applied structural modelling, which indicated that *BARS2* encodes a functional clade II OSC. It may be involved in the biosynthesis of one product or a range of them, presumably different from those generated by its closest paralog, BARS1. It is worth to notice that BARS1 has the lowest product specificity among plant OSCs [55]. It catalyzes the formation of at least 23 triterpene scaffolds: mono-, tri-, tetra- and pentacyclic ones, among which baruol is the major product [27]. Why some OSCs are highly multi-functional is not well understood. It has been suggested that low product specificity might reflect ongoing evolution towards mono-functionality. It was also demonstrated that as few as two amino acid changes in the OSC sequence may lead to the complete transformation of its product specificity [56]. Future biochemical characterization of BARS2 protein may thus contribute to increasing the knowledge of natural triterpene biosynthetic pathways in Arabidopsis and revealing the role of particular amino acids in acquiring specificity for given products.

BARS2 is a part of a large ~25kb insertion located at the end of the arabidiol/baruol gene cluster, extending its region significantly. Within this insert, we also found a duplication of *CYP705A2*, namely *CYP705A2a*. According to our estimations, *BARS2* and *CYP705A2a* gene pair may be present in nearly one-third of the Arabidopsis population. Interestingly, we observed a significant association between the presence/absence of this segment and the climates of accessions' origin as well as the root growth dynamics. Unfortunately, little is known about the role of baruol and its modification products in Arabidopsis, which currently limits our ability to explain this association. Additional studies will be also required to assess whether it is related to the expression of individual

genes or to the differences in transcriptional activity of the entire cluster, e.g. by the epigenetic mechanisms [57–59]. It has been shown that Arabidopsis MGCs are embedded in local hotspots of three-dimensional chromatin interactions. Loss of the histone mark H3K27me3 in *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 MGCs are strictly regulated by the switches in their conformation [60]. The presence of ~25-kb insertion in the arabidiol/baruol gene cluster may alter its structure and lead to changes in epigenetic regulation of its activity. Curiously, in accessions with *CYP705A2a* and *BARS2*, we observed some transcriptional activity of arabidiol/baruol cluster genes in leaves, while in Col-0 all MGCs studied so far are active in roots but repressed in leaves, which is combined with distinct chromatin condensation state and nuclear repositioning of MGC regions between these tissues. The putative biological role of tissue-specific gene expression differences between the accessions with and without the insertion in the baruol gene cluster needs to be determined.

A recent study demonstrated that thalianin, thalianyl fatty acid esters and arabidin modulate Arabidopsis root microbiota [11]. Various triterpene cocktails attracted or repelled various microbial communities present in the soil. Moreover, some of the applied triterpenes were metabolized by specific bacterial strains, indicating their role in the plant's active selection of root microbiota. In fact, a small but significant effect of Arabidopsis genotype on the root microbiome has been demonstrated previously [52,53]. Additionally, a strong association was observed between Arabidopsis phyllosphere microbiota and the host genotype, in a recent study by Karasov et al. [61]. In that study, bacterial communities that colonized the leaves of 267 Arabidopsis populations, assessed at various localizations in Europe, formed two distinct groups. These groups were strongly associated with the latititude, in a phylotype-specific manner, namely a significant latitudinal cline was observed for the most abundant strains of *Sphingomonas* genus. *Sphingomonas* are commonly associated with both roots and leaves across Arabidopsis populations [62]. The species belonging to

this genus possess a range of biodegradative and biosynthetic capabilities [63,64]. Plant-colonizing *Sphingomonas* strains were also implicated in promoting Arabidopsis growth and increasing drought resistance [65] or protecting the plants against the leaf-pathogenic *Pseudomonas syringae* [66]. Prominently, in the study by Karasov et al., the host plant genotype alone could explain 52% to 68% of the observed variance in the phyllosphere microbial composition [61]. Moreover, the microbiome type was strongly associated with the index of the dryness of the local environment, based on recent precipitation and temperature data. Taking into account that the presence/absence variation of genomic insertion containing *CYP705A2a* and *BARS2* was also associated with these climatic variables and that the geographic distribution of the compact version of thalianol gene cluster was skewed towards higher latitudes, we hypothesize that the genetic diversity of triterpene biosynthesis pathways in Arabidopsis may be interdependent on the diversity of soil bacterial communities present in various environments and this relationship might play some role in Arabidopsis adaptation to climate-driven selective pressures. It is worth to note that triterpenes are high-molecular-weight non-volatile compounds, which are likely to act locally. However, they may be further processed and generate various breakdown products, both volatile and non-volatile, which may be biologically active [26,54]. There is also a growing evidence that intraspecies variation in secondary metabolites accumulating in plants is shaped by local adaptation [48,67,68]. Exploring the diversity of Arabidopsis terpenoid metabolism pathways may help understand how local adaptation shapes intraspecies variation in secondary metabolites in plants and contribute to identification of new useful compounds of plant origin [69].

The bioinformatic identification of new MGC candidates is mainly based on the combination of physical gene grouping and their coexpression, which are the prominent features of functional gene clusters. MGC prediction is, however, often complicated by the low expression of genes for specialized metabolite biosynthesis, therefore its accuracy and sensitivity strongly depend on the abundance of data from various tissues, time points, and environmental conditions [23]. Based on the

presented results, we suggest that the analysis of intraspecies genetic and transcriptomic variation will provide additional contributions 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 [14,41,48,70]. 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.

CONCLUSIONS

In the current study, we created a detailed picture of CNVs and structural rearrangements, driving the diversity of triterpene biosynthesis-related genes in Arabidopsis. We revealed new associations between MGC structure, root growth phenotypes and the geographic distribution of Arabidopsis accessions. Our results provided insight into the recent evolution of the OSCs gene family and demonstrated the differences in the dynamics of MGC regions. This study sets the grounds for future investigation of the apparently complex genetic network underlying triterpene metabolism in Arabidopsis and the mechanisms of its regulation.

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 [13,15–21,23,25–27]. TAIR10 genome version and Araport 11 annotations [71] were used as a reference in all analyses. Short read sequencing data from Arabidopsis 1001 Genomes Project [28] were downloaded from National Center for Biotechnology Information Sequence Read Archive repository (PRJNA273563), processed and mapped to the reference genome as described in [14]. The gene copy number estimates based on read-depth analysis of short reads (RD dataset) were generated previously and are available at [72]. 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 [73]. Probe genomic target coordinates are listed in **Additional file1: 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 (**Additional file1:**

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 a template in 20 µl reactions performed with PrimeSTAR GXL DNA Polymerase (TaKaRa), according to the manufacturer's instructions, in a three-step PCR. 10 ul Amplicons were analyzed on 1% agarose with 1kb Gene Ruler DNA ladder (Fermentas). Primer sequences are listed in **Additional file 1: 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 (**Additional file 1: Figure S4; Additional file 2: 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 **Additional file 2: 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. The assembly details will be described in a separate paper. 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 [24] Assembled genomic sequences of Ty-1 (PRJEB37258), Cdm-0 (PRJEB40125) and Kn-0 (PRJEB37260) accessions were retrieved from NCBI/Assembly database [74]. Gene prediction was performed with Augustus v.3.3.3 [75] 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 **Additional file 1: Supplementary Methods**. 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 [76,77]. TEs were annotated with RepeatMasker software (version 4.1.2) using homology-based method with TAIR10-transposable-elements reference library [78].

Identification of chromosomal inversions

The BreakDancerMax program from the BreakDancer package v.1.3.6 [79] 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 attempted to minimize this 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 [80]. 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* & *CYP705A2a-BARS2* pair status were downloaded from the 1001 Genomes Project Data Center (1001genomes_snp-short-indel_only_ACGTN_v3.1.vcomparedwithsnpeff file) [28]. Data preprocessing was performed using PLINK v.1.90b3w [81,82]. 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 0.5. Over 250,000 SNPs were used for PCA analysis with EIGENSOFT v.7.2.1 [83,84]. PCA for a wide LD range between 0.3 - 0.9 was calculated. U.S.A accessions which only recently separated geographically from the rest of the population [32] were excluded to ensure better visibility of the remaining accessions. The ggplot2 package was used for data visualization in R v4.0.4 [85,86].

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 [87]. 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 [88]. Input file generation and analysis of the results were performed with PLINK v.1.90b3w and R v4.0.4 [81,82,86].

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 [41] 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 [42]. SRA Toolkit v2.8.2 [89] and FastQC v0.11.4 [90] 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. **(Additional file 4**). Raw reads were mapped to the respective genomes using the STAR aligner version 2.7.8a [91]. 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 was used to convert .gff3 to .gtf files [92]). 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 [86].

Analysis of TS-CYP pairs

A list of Arabidopsis CYP genes was created by collecting information from previous studies and acknowledged website resources [22,43–45,93–95]. Genes marked in Araport 11 as pseudogenes were excluded from the further analysis. Genes were assigned to clans and families according to the

above resources. A list of TS genes was created based on a previous study [22] 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 [10].

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 [30,96]. 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 [31,97]. The SSM algorithm implemented in COOT was used for superpositions of protein models [98,99] (**Additional file 1: Supplementary Methods**).

ADDITIONAL FILES

Additional file 1. Supplementary Tables S1-S17. Format .xlsx

Additional file 2. Supplementary Methods and Supplementary Figs. S1-S21. Format .pdf

Additional file 3. Superposed 3D models of BARS1, BARS2 and human oxidosqualene cyclase proteins. Format .pdb, zipped.

Additional file 4. Gene models for selected genes for Cdm-0, Kn-0 and Ty-1 genomes. Format .gtf, zipped.

DECLARATIONS

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

Data generated or analyzed during this study are included in this published article and its supplementary information files. Public sources of previously published WGS, RNA-Seq datasets and catalogs of genetic variants (SNPs and CNVs) are detailed in the Methods section. The geographic distribution map of accessions with and without *CYP705A2-BARS1* duplication [33] was generated based on accession information from the 1001 Genomes Project resource page.

Competing interests

The authors declare that they have no competing interests.

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.

Authors' contributions

AZ conceived the project. MM-Z, AS, PS, KB and TI generated the data. MM-Z, PW, MZ and AZ analyzed the data. MF contributed to the discussion and interpretation of the results. MM-Z and AZ wrote the paper and designed the figures. All authors contributed to the final version of the manuscript.

Acknowledgments

The authors thank Piotr Kozłowski for fruitful discussions and comments on the manuscript. This research was supported in part by PLGrid Infrastructure.

REFERENCES

1. Erb M, Kliebenstein DJ. Plant Secondary Metabolites as Defenses, Regulators, and Primary Metabolites: The Blurred Functional Trichotomy. Plant Physiol. 2020;184:39-52.

2. Nützmann H-W, Scazzocchio C, Osbourn A. Metabolic Gene Clusters in Eukaryotes. Annu Rev Genet. 2018;52:159-83.

3. Pinto DCGA, Simões MAM, Silva AMS. *Genista tridentata* L.: A Rich Source of Flavonoids with Anti-inflammatory Activity. Medicines (Basel). 2020;7:31.

4. Rathaur P, Johar KSR. Metabolism and Pharmacokinetics of Phytochemicals in the Human Body. Curr Drug Metab. 2019;20:1085-102.

5. Veeresham C. Natural products derived from plants as a source of drugs. J Adv Pharm Technol Res. 2012;3:200-1.

6. He J-H, Chen L-X, Li H. Progress in the discovery of naturally occurring anti-diabetic drugs and in the identification of their molecular targets. Fitoterapia. 2019;134:270-89.

7. Isah T. Stress and defense responses in plant secondary metabolites production. Biol Res. 2019;52:39.

8. Nützmann H-W, Osbourn A. Gene clustering in plant specialized metabolism. Curr Opin Biotechnol. 2014;26:91-9.

9. Nützmann H-W, Huang A, Osbourn A. Plant metabolic gene clusters - from genetics to genomics. New Phytol. 2016;211:771-89.

10. Kautsar SA, Hernando GSD, Blin K, Osbourn A, Medema MH. plantiSMASH: automated identification, annotation and expression analysis of plant biosynthetic gene clusters. Nucleic Acids Res. 2017;45:W55-W63.

11. Huang AC, Jiang T, Liu Y-X, Bai Y-C, Reed J, Qu B, et al. A specialized metabolic network selectively modulates *Arabidopsis* root microbiota. Science. 2019;364:eaau6389.

12. Boycheva S, Daviet L, Wolfender J-L, Fitzpatrick TB. The rise of operon-like gene clusters in plants. Trends Plant Sci. 2014;19:447-59.

13. Thimmappa R, Geisler K, Louveau T, O'Maille P, Osbourn A. Triterpene Biosynthesis in Plants. Annu Rev Plant Biol. 2014;65:225-57.

14. Zmienko A, Marszalek-Zenczak M, Wojciechowski P, Samelak-Czajka A, Luczak M, Kozlowski P, et al. AthCNV: A Map of DNA Copy Number Variations in the Arabidopsis Genome. Plant Cell. 2020;32:1797-819.

15. Fazio GC, Xu R, Matsuda SPT. Genome mining to identify new plant triterpenoids. J Am Chem Soc. 2004;126:5678-9.

16. Field B, Osbourn AE. Metabolic Diversification—Independent Assembly of Operon-Like Gene Clusters in Different Plants. Science. 2008;320:543-7.

17. Xiong Q, Wilson WK, Matsuda SPT. An Arabidopsis oxidosqualene cyclase catalyzes iridal skeleton formation by Grob fragmentation. Angew Chem Int Ed Engl. 2006;45:1285-8.

18. Go YS, Lee SB, Kim HJ, Kim J, Park H-Y, Kim J-K, et al. Identification of marneral synthase, which is critical for growth and development in Arabidopsis. Plant J. 2012;72:791-804.

19. Field B, Fiston-Lavier A-S, Kemen A, Geisler K, Quesneville H, Osbourn AE. Formation of plant metabolic gene clusters within dynamic chromosomal regions. Proc Natl Acad Sci U S A. 2011;108:16116-21.

20. Morlacchi P, Wilson WK, Xiong Q, Bhaduri A, Sttivend D, Kolesnikova MD, et al. Product profile of PEN3: The last unexamined oxidosqualene cyclase in Arabidopsis thaliana. Org Lett. 2009;11:2627-30.

21. Yasumoto S, Fukushima EO, Seki H, Muranaka T. Novel triterpene oxidizing activity of Arabidopsis thaliana CYP716A subfamily enzymes. FEBS Lett. 2016;590:533-40.

22. Boutanaev AM, Moses T, Zi J, Nelson DR, Mugford ST, Peters RJ, et al. Investigation of terpene diversification across multiple sequenced plant genomes. Proc Natl Acad Sci U S A. 2014;112:E81-E88.

23. Wisecaver JH, Borowsky AT, Tzin V, Jander G, Kliebenstein DJ, Rokas A. A Global Coexpression Network Approach for Connecting Genes to Specialized Metabolic Pathways in Plants. Plant Cell. 2017;29:944-59.

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

25. Xiang T, Shibuya M, Katsube Y, Tsutsumi T, Otsuka M, Zhang H, et al. A new triterpene synthase from Arabidopsis thaliana produces a tricyclic triterpene with two hydroxyl groups. Org Lett. 2006;8:2835-8.

26. Sohrabi R, Huh J-H, Badieyan S, Rakotondraibe LH, Kliebenstein DJ, Sobrado P, et al. 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. 2015;27:874-90.

27. Lodeiro S, Xiong Q, Wilson WK, Kolesnikova MD, Onak CS, Matsuda SPT. An oxidosqualene cyclase makes numerous products by diverse mechanisms: a challenge to prevailing concepts of triterpene biosynthesis. J Am Chem Soc. 2007;129:11213-22.

28. 1001 Genomes Consortium. 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. Cell. 2016;166:481-91.

29. Liu Z, Cheema J, Vigouroux M, Hill L, Reed J, Paajanen P, et al. Formation and diversification of a paradigm biosynthetic gene cluster in plants. Nat Commun. 2020;11:5354.

30. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596:583-9.

31. Thoma R, Schulz-Gasch T, D'Arcy B, Benz J, Aebi J, Dehmlow H, et al. Insight into steroid scaffold formation from the structure of human oxidosqualene cyclase. Nature. 2004;432:118-22.

32. Lee C-R, Svardal H, Farlow A, Exposito-Alonso M, Ding W, Novikova P, et al. On the post-glacial spread of human commensal Arabidopsis thaliana. Nat Commun. 2017;8:14458.

33. Distribution of accessions with varying CYP705A2-BARS1 copy number status. 2022. https://www.google.com/maps/d/edit?mid=1ZaAMX-EDYlbBtKbKdBsS06HMjvc&usp=sharing Accessed 12 July 2022.

34. Seren Ü, Grimm D, Fitz J, Weigel D, Nordborg M, Borgwardt K, et al. AraPheno: a public database for Arabidopsis thaliana phenotypes. Nucleic Acids Res. 2017;45:D1054-D1059.

35. Exposito-Alonso M, 500 Genomes Field Experiment Team, Burbano HA, Bossdorf O, Nielsen R, Weigel D. Natural selection on the Arabidopsis thaliana genome in present and future climates. Nature. 2019;573:126-9.

36. Li Y, Huang Y, Bergelson J, Nordborg M, Borevitz JO. Association mapping of local climate-sensitive quantitative trait loci in Arabidopsis thaliana. Proc Natl Acad Sci U S A. 2010;107:21199-204.

37. Bai Y, Fernández-Calvo P, Ritter A, Huang AC, Morales-Herrera S, Bicalho KU, et al. Modulation of Arabidopsis root growth by specialized triterpenes. New Phytol. 2021;230:228-43.

38. Bouain N, Satbhai SB, Korte A, Saenchai C, Desbrosses G, Berthomieu P, et al. Natural allelic variation of the AZI1 gene controls root growth under zinc-limiting condition. PLoS Genet. 2018;14:e1007304.

39. Lardon R, Wijnker E, Keurentjes J, Geelen D. The genetic framework of shoot regeneration in Arabidopsis comprises master regulators and conditional fine-tuning factors. Commun Biol. 2020;3:549.

40. Coop G, Witonsky D, Di Rienzo A, Pritchard JK. Using environmental correlations to identify loci underlying local adaptation. Genetics. 2010;185:1411-23.

41. Kawakatsu T, Huang S-SC, Jupe F, Sasaki E, Schmitz RJ, Urich MA, et al. Epigenomic Diversity in a Global Collection of Arabidopsis thaliana Accessions. Cell. 2016;166:492-505.

42. van Veen H, Vashisht D, Akman M, Girke T, Mustroph A, Reinen E, et al. Transcriptomes of Eight Arabidopsis thaliana Accessions Reveal Core Conserved, Genotype- and Organ-Specific Responses to Flooding Stress. Plant Physiol. 2016;172:668-89.

43. Nelson D, Werck-Reichhart D. A P450-centric view of plant evolution. Plant J. 2011;66:194-211.

44. Bak S, Beisson F, Bishop G, Hamberger B, Höfer R, Paquette S, et al. Cytochromes p450. Arabidopsis Book. 2011;9:e0144.

45. Paquette SM, Bak S, Feyereisen R. Intron-exon organization and phylogeny in a large superfamily, the paralogous cytochrome P450 genes of Arabidopsis thaliana. DNA Cell Biol. 2000;19:307-17.

46. Liu Z, Suarez Duran HGS, Harnvanichvech Y, Stephenson MJ, Schranz ME, Nelson D, et al. Drivers of metabolic diversification: how dynamic genomic neighbourhoods generate new biosynthetic pathways in the Brassicaceae. New Phytol. 2020;227:1109-23.

47. Fan P, Wang P, Lou YR, Leong BJ, Moore BM, Schenck CA, et al. . Evolution of a plant gene cluster in Solanaceae and emergence of metabolic diversity. Elife. 2020;9:e56717.

48. Katz E, Li J-J, Jaegle B, Ashkenazy H, Abrahams SR, Bagaza C, et al. Genetic variation, environment and demography intersect to shape Arabidopsis defense metabolite variation across Europe. Elife. 2021;10:e67784.

49. Li Q, Ramasamy S, Singh P, Hagel JM, Dunemann SM, Chen X, et al. Gene clustering and copy number variation in alkaloid metabolic pathways of opium poppy. Nat Commun. 2020;11:1190.

50. Zhan C, Lei L, Liu Z, Zhou S, Yang C, Zhu X, et al. Selection of a subspecies-specific diterpene gene cluster implicated in rice disease resistance. Nat Plants. 2020;6:1447-1454.

51. Lind AL, Wisecaver JH, Lameiras C, Wiemann P, Palmer JM, Keller NP, et al. Drivers of genetic diversity in secondary metabolic gene clusters within a fungal species. PLoS Biol. 2017;15:e2003583.

52. Lundberg DS, Lebeis SL, Paredes SH, Yourstone S, Gehring J, Malfatti S, et al. Defining the core *Arabidopsis thaliana* root microbiome. Nature. 2012;488:86-90.

53. Bulgarelli D, Rott M, Schlaeppi K, Ver Loren van Themaat E, Ahmadinejad N, Assenza F, et al. Revealing structure and assembly cues for *Arabidopsis* root-inhabiting bacterial microbiota. Nature. 2012;488:91-5.

54. Sohrabi R, Ali T, Harinantenaina Rakotondraibe L, Tholl D. Formation and exudation of non-volatile products of the arabidiol triterpenoid degradation pathway in Arabidopsis roots. Plant Signal Behav. 2017;12:e1265722.

55. Ghosh S. Biosynthesis of Structurally Diverse Triterpenes in Plants: the Role of Oxidosqualene Cyclase. Proc Indian Natl Sci Acad. 2016;82:1189-210

56. Lodeiro S, Schulz-Gasch T, Matsuda SPT. Enzyme redesign: two mutations cooperate to convert cycloartenol synthase into an accurate lanosterol synthase. J Am Chem Soc. 2005;127:14132–3.

57. Wegel E, Koumproglou R, Shaw P, Osbourn A. Cell type-specific chromatin decondensation of a metabolic gene cluster in oats. Plant Cell. 2009;21:3926–36.

58. Yu N, Nützmann H-W, MacDonald JT, Moore B, Field B, Berriri S, et al. Delineation of metabolic gene clusters in plant genomes by chromatin signatures. Nucleic Acids Res. 2016;44:2255-65.

59. Roulé T, Christ A, Hussain N, Huang Y, Hartmann C, Benhamed M, et al. The lncRNA MARS modulates the epigenetic reprogramming of the marneral cluster in response to ABA. Mol Plant. 2022;15:840-56.

60. Nützmann H-W, Doerr D, Ramírez-Colmenero A, Sotelo-Fonseca JE, Wegel E, Di Stefano M, et al. Active and repressed biosynthetic gene clusters have spatially distinct chromosome states. Proc Natl Acad Sci U S A. 2020;117:13800-9.

61. Karasov TL, Neumann M, Shirsekar G, Monroe G, PATHODOPSIS Team, Weigel D, et al. Drought selection on Arabidopsis populations and their microbiomes. bioRxiv. 2022; doi: 10.1101/2022.04.08.487684.

62. Bodenhausen N, Horton MW, Bergelson J. Bacterial communities associated with the leaves and the roots of *Arabidopsis thaliana*. PLoS One. 2013;8:e56329.

63. Asaf S, Numan M, Khan AL, Al-Harrasi A. Sphingomonas: from diversity and genomics to functional role in environmental remediation and plant growth. Crit Rev Biotechnol. 2020; 40:138-152.

64. Mohn WW, Yu Z, Moore ER, Muttray AF. Lessons learned from *Sphingomonas* species that degrade abietane triterpenoids. J Ind Microbiol Biotechnol. 1999;23:374-379.

65. Luo Y, Wang F, Huang Y, Zhou M, Gao J, Yan T, et al. *Sphingomonas* sp. Cra20 Increases Plant Growth Rate and Alters Rhizosphere Microbial Community Structure of *Arabidopsis thaliana* Under Drought Stress. Front Microbiol. 2019; 10:1221.

66. Innerebner G, Knief C, Vorholt JA. Protection of *Arabidopsis thaliana* against leaf-pathogenic *Pseudomonas syringae* by *Sphingomona*s strains in a controlled model system. Appl Environ Microbiol. 2011;77:3202-10.

67. Snoeren TAL, Kappers IF, Broekgaarden C, Mumm R, Dicke M, Bouwmeester HJ. Natural variation in herbivore-induced volatiles in Arabidopsis thaliana. J Exp Bot. 2010;61:3041-56.

68. Kooke R, Kruijer W, van Eekelen HDLM, Becker FFM, Wehrens R, Hall RD, et al. Local adaptation shapes metabolic diversity in the global population of Arabidopsis thaliana. bioRxiv. 2021. doi:10.1101/2021.09.13.460026.

69. Zhou X, Liu Z. Unlocking plant metabolic diversity: A (pan)-genomic view. Plant Commun. 2022;3:100300.

70. Shirai K, Matsuda F, Nakabayashi R, Okamoto M, Tanaka M, Fujimoto A, et al. 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. 2017;34:3111-22.

71. Cheng C-Y, Krishnakumar V, Chan AP, Thibaud-Nissen F, Schobel S, Town CD. Araport11: a complete reannotation of the Arabidopsis thaliana reference genome. Plant J. 2017;89:789-804.

72. Zmienko A, Wojciechowski P. Copy number variation of *Arabidopsis thaliana* genes. http://athcnv.ibch.poznan.pl/. Accessed 29 August 2022.

73. Samelak-Czajka A, Marszalek-Zenczak M, Marcinkowska-Swojak M, Kozlowski P, Figlerowicz M, Zmienko A. MLPA-Based Analysis of Copy Number Variation in Plant Populations. Front Plant Sci. 2017;8:222.

74. Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, et al. Database resources of the national center for biotechnology information. Nucleic Acids Res. 2022;50:D20-B26.

75. Stanke M, Morgenstern B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Res. 2005;33:W465-W467.

76. Corpet F. Multiple sequence alignment with hierarchical clustering. Nucleic Acids Res. 1988;16:10881-90.

77. Madeira F, Park YM, Lee J, Buso N, Gur T, Madhusoodanan N, et al. The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res. 2019;47:W636-W641.

78. Smit, AFA, Hubley, R & Green, P. RepeatMasker Open-4.0. 2013-2015. http://www.repeatmasker.org. Accessed 12 June 2022.

79. Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat Methods. 2009;6:677-81.

80. Yun T, Li H, Chang P-C, Lin MF, Carroll A, McLean CY. Accurate, scalable cohort variant calls using DeepVariant and GLnexus. Bioinformatics. 2021;36:5582-9.

81. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4:7.

82. Chang C. PLINK 1.90 beta. 2022. https://www.cog-genomics.org/plink/1.9/. Accessed 12 June 2022.

83. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006;38:904-9.

84. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet. 2006;2:e190.

85. Wickham H. ggplot2. Elegant Graphics for Data Analysis. 2nd ed. Springer Cham. 2016. Available at https://doi.org/10.1007/978-3-319-24277-4. Accessed 29 August 2022.

86. R Foundation. The R Project for Statistical Computing. https://www.r-project.org. Accessed 29 August 2022.

87. Togninalli M, Seren Ü, Freudenthal JA, Monroe JG, Meng D, Nordborg M, et al. AraPheno and the AraGWAS Catalog 2020: a major database update including RNA-Seq and knockout mutation data for Arabidopsis thaliana. Nucleic Acids Res. 2020;48:D1063-D1068.

88. Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, et al. Efficient control of population structure in model organism association mapping. Genetics. 2008;178:1709-23.

89. The SRA Toolkit. https://github.com/ncbi/sra-tools. Accessed 06 March 2022.

90. Babraham Bioinformatics. FASTQC - A quality control tool for high throughput sequence data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Accessed 12 May 2022

91. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15-21.

92. Bedre R. reneshbedre/bioinfokit: Bioinformatics data analysis and visualization toolkit. 2021. https://zenodo.org/record/4422035#.Ys1LIVTP1hE_Accessed 6 Jan 2021.

93. Nelson DR. The Cytochrome P450 Homepage. Hum Genomics. 2009;4:59-65.

94. Ehlting J, Sauveplane V, Olry A, Ginglinger J-F, Provart NJ, Werck-Reichhart D. An extensive (co-)expression analysis tool for the cytochrome P450 superfamily in Arabidopsis thaliana. BMC Plant Biol. 2008;8:47.

95. Arabidopsis Cytochromes P450. 2010. http://www.p450.kvl.dk/p450.shtml. Accessed 29 Nov 2021.

96. Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: making protein folding accessible to all. Nat Methods. 2022;19:679-82.

97. Berman H, Henrick K, Nakamura H, Markley JL. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. Nucleic Acids Res. 2007;35:D301-D303.

98. Krissinel E, Henrick K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Crystallogr D Biol Crystallogr. 2004;60:2256-68.

99. Emsley P, Lohkamp B, Scott WG, Cowtan K. Features and development of Coot. Acta Crystallogr D Biol Crystallogr. 2010;66:486-501.

Figure legends

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 Additional file 1: Table S6.

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.

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).

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 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. The legend of genetic groups' colors is in **Additional file 2: FigS17**. D) Latitudes of accessions' sites of origin, grouped by CYP-BARS status. Wilcoxon rank sum test with continuity correction, **p.value<0.05 (PP-PP vs AA-PP); ***p.value < 0.001 (all the other pairwise comparisons).

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), Additional file 1: Table S12 (plot D) and Additional file 1: Table S13 (plots E-F).

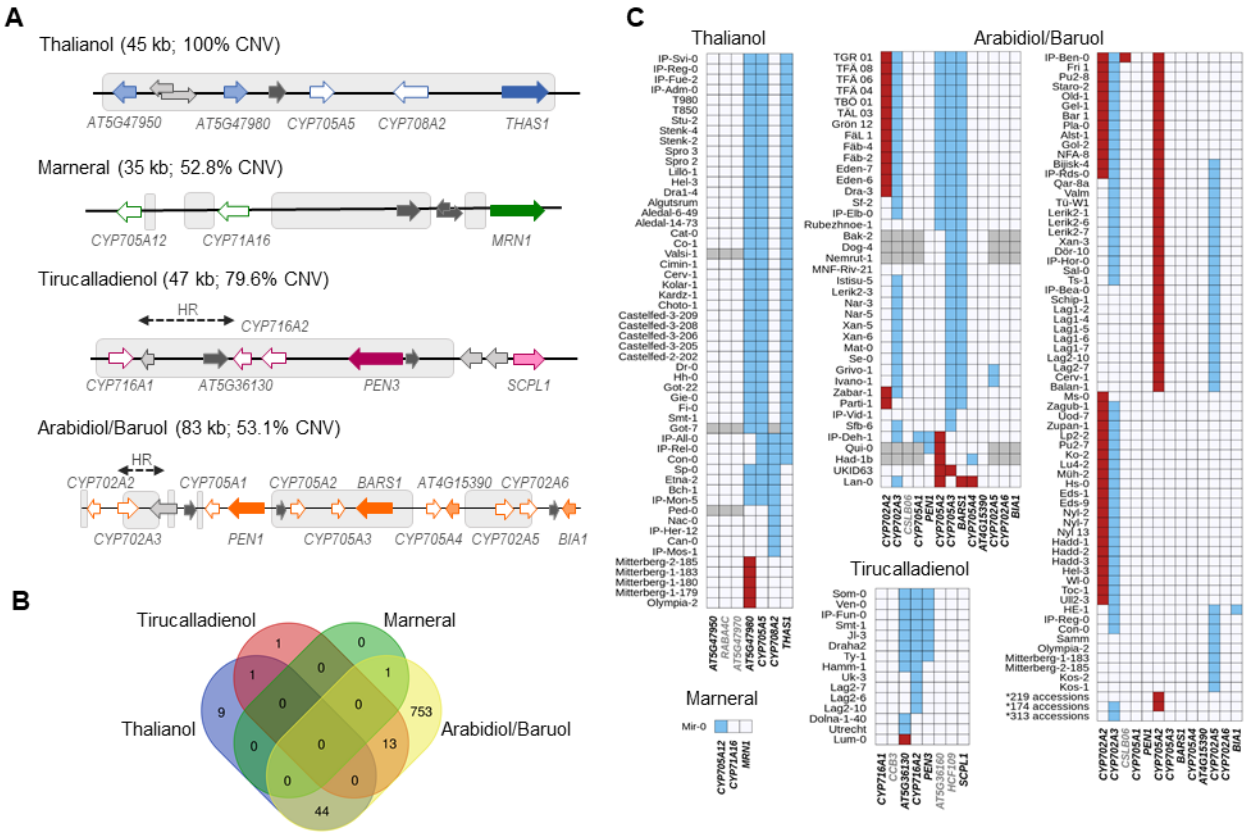
A
Thalianol (45 kb; 100% CNV)
AT5G47950
AT5G47980
CYP705A5
CYP708A2
THAS1
Marneral (35 kb; 52.8% CNV)
CYP705A12
CYP71A16
MRN1
Tirucalladienol (47 kb; 79.6% CNV)
HR
CYP716A2
CYP716A1
AT5G36130
PEN3
SCPL1
Arabidiol/Baruol (83 kb; 53.1% CNV)
HR
CYP702A2
CYP705A1
CYP705A2
BARS1
AT4G15390
CYP702A6
CYP702A3
PEN1
CYP705A3
CYP705A4
CYP702A5
BIA1
B
Tirucalladienol
Marneral
Thalianol
Arabidiol/Baruol
1
1
0
0
1
9
0
0
753
0
0
13
0
0
44
C
Thalianol
Arabidiol/Baruol
Tirucalladienol
Marneral
Mir-0
*219 accessions
*174 accessions
*313 accessions

Figure 1

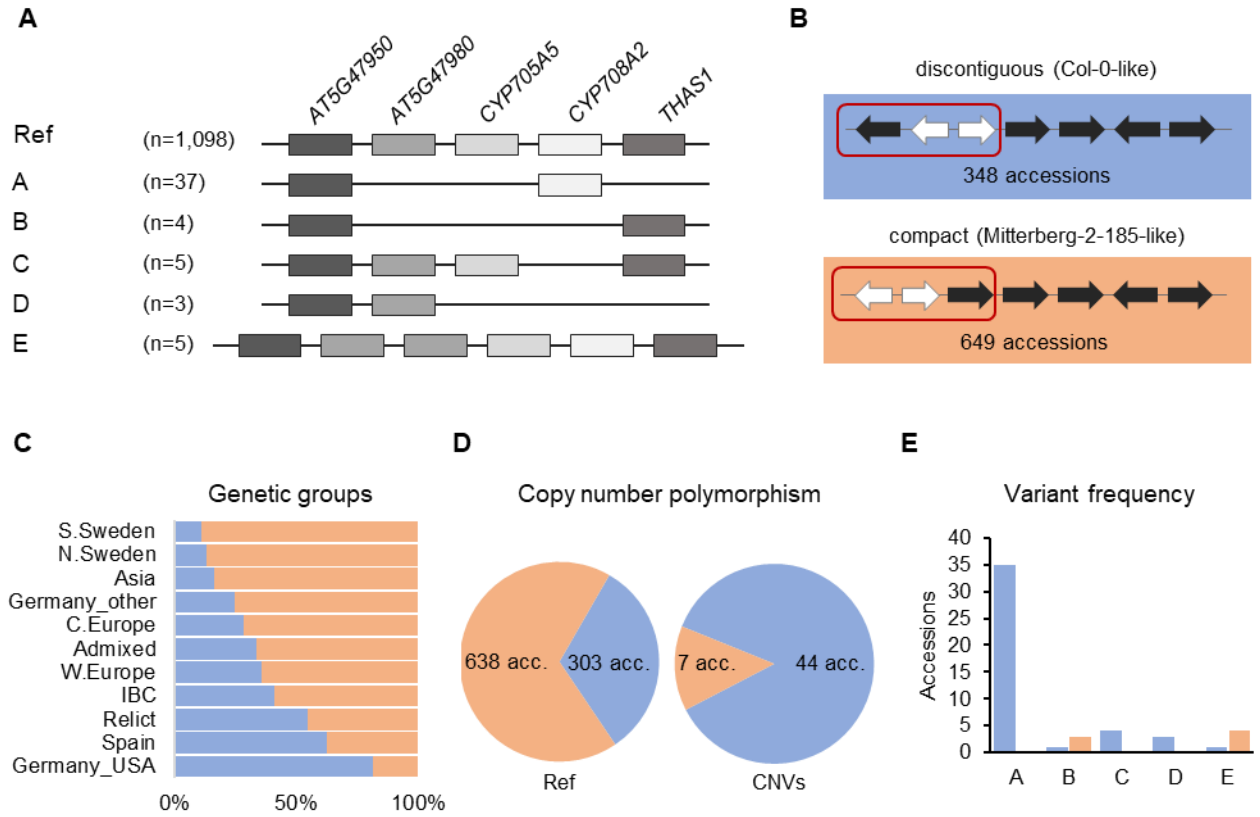
A
AT5G47950
AT5G47980
CYP705A5
CYP708A2
THAS1
Ref (n=1,098)
A (n=37)
B (n=4)
C (n=5)
D (n=3)
E (n=5)
B
discontiguous (Col-0-like)
348 accessions
compact (Mitterberg-2-185-like)
649 accessions
C
Genetic groups
S.Sweden
N.Sweden
Asia
Germany_other
C.Europe
Admixed
W.Europe
IBC
Relict
Spain
Germany_USA
0%
50%
100%
D
Copy number polymorphism
638 acc.
303 acc.
7 acc.
44 acc.
Ref
CNVs
E
Variant frequency
Accessions
40
35
30
25
20
15
10
5
0
A
B
C
D
E

Figure 2

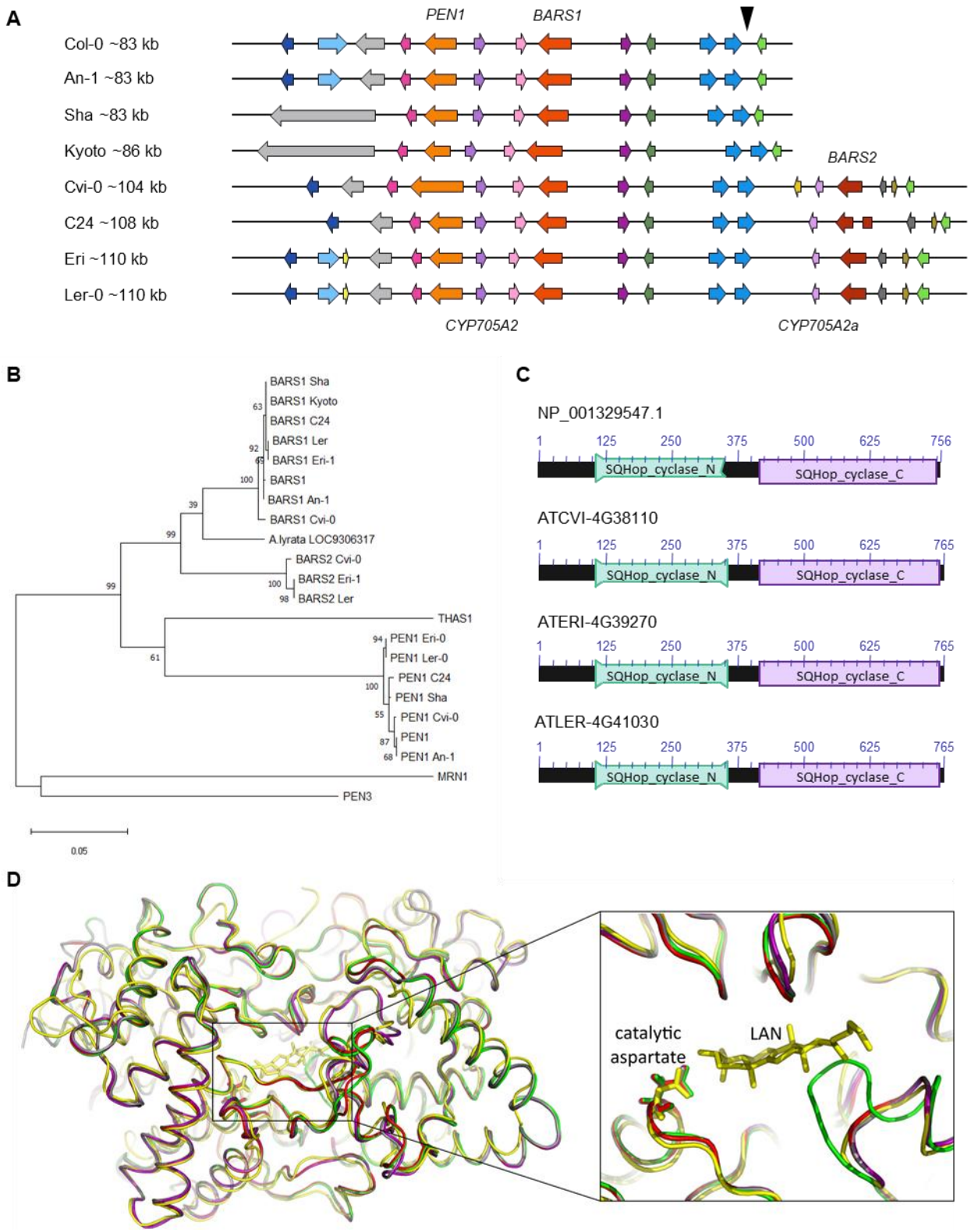
A
PEN1
BARS1
Col-0 ~83 kb
An-1 ~83 kb
Sha ~83 kb
Kyoto ~86 kb
Cvi-0 ~104 kb
C24 ~108 kb
Eri ~110 kb
Ler-0 ~110 kb
BARS2
CYP705A2
CYP705A2a
B
BARS1 Sha
BARS1 Kyoto
BARS1 C24
BARS1 Ler
BARS1 Eri-1
BARS1
BARS1 An-1
BARS1 Cvi-0
A.lyrata LOC9306317
BARS2 Cvi-0
BARS2 Eri-1
BARS2 Ler
THAS1
PEN1 Eri-0
PEN1 Ler-0
PEN1 C24
PEN1 Sha
PEN1 Cvi-0
PEN1
PEN1 An-1
MRN1
PEN3
0.05
C
NP_001329547.1
ATCVI-4G38110
ATERI-4G39270
ATLER-4G41030
SQHop_cyclase_N
SQHop_cyclase_C
D
catalytic aspartate
LAN

Figure 3

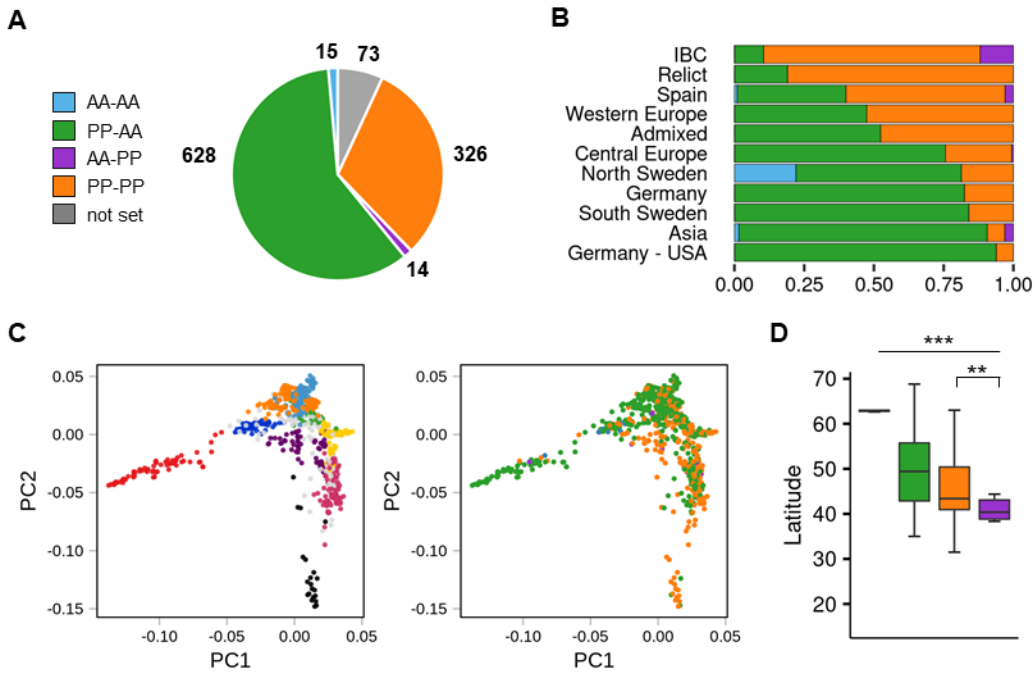
A
AA-AA
PP-AA
AA-PP
PP-PP
not set
15
73
628
326
14
B
IBC
Relict
Spain
Western Europe
Admixed
Central Europe
North Sweden
Germany
South Sweden
Asia
Germany - USA
0.00
0.25
0.50
0.75
1.00
C
PC2
PC1
D
Latitude
***
**

Figure 4

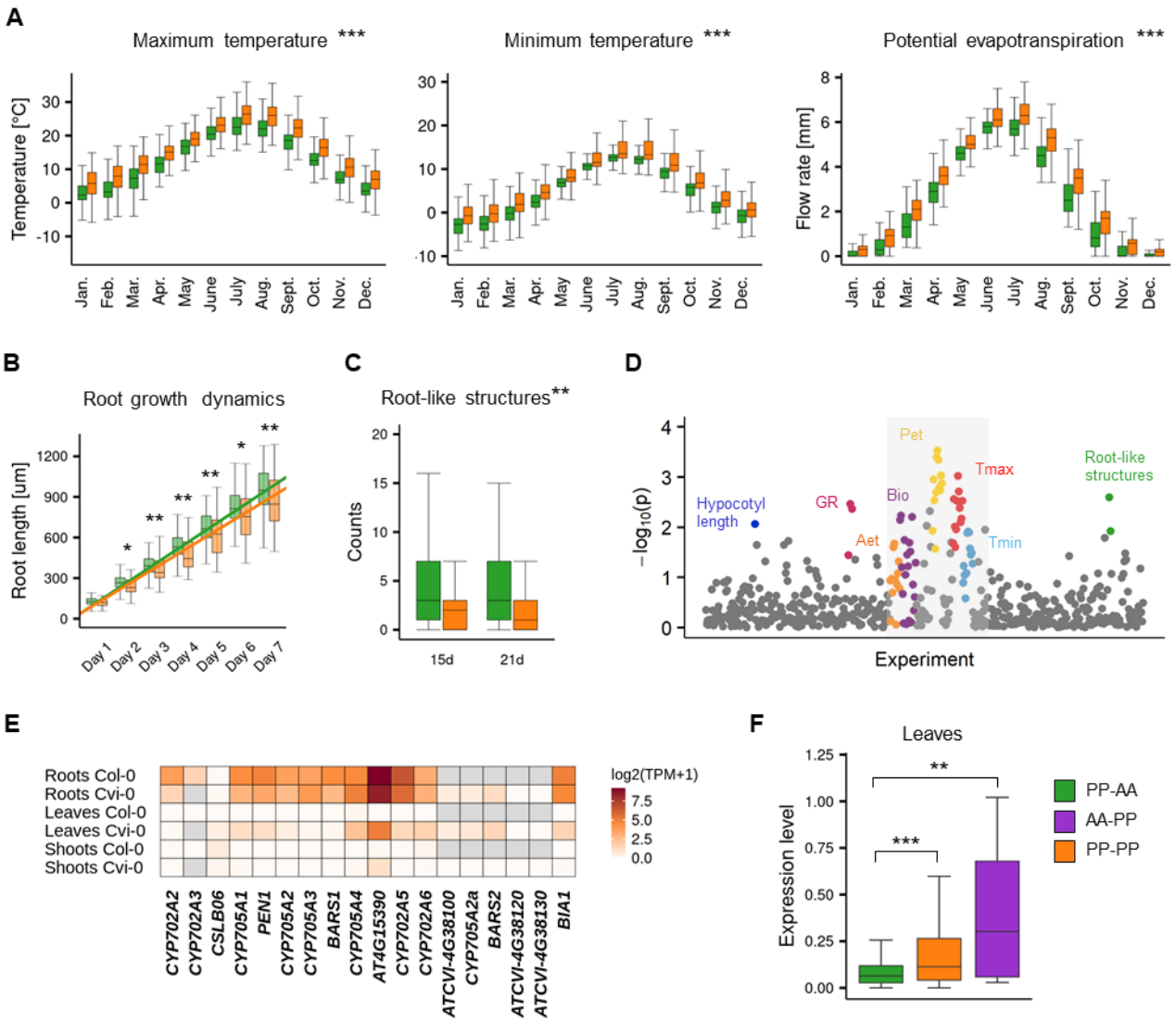
A
Maximum temperature ***
Minimum temperature ***
Potential evapotranspiration ***
Temperature [°C]
Flow rate [mm]
Jan. Feb. Mar. Apr. May June July Aug. Sept. Oct. Nov. Dec.
B
Root growth dynamics
Root length [um]
Day 1 Day 2 Day 3 Day 4 Day 5 Day 6 Day 7
C
Root-like structures**
Counts
15d
21d
D
Pet
Tmax
Root-like structures
Hypocotyl length
GR
Bio
Aet
Tmin
−log10(p)
Experiment
E
Roots Col-0
Roots Cvi-0
Leaves Col-0
Leaves Cvi-0
Shoots Col-0
Shoots Cvi-0
log2(TPM+1)
CYP702A2 CYP702A3 CSLB06 CYP705A1 PEN1 CYP705A2 CYP705A3 BARS1 CYP705A4 AT4G15390 CYP702A5 CYP702A6 ATCVI-4G38100 CYP705A2a BARS2 ATCVI-4G38120 ATCVI-4G38130 BIA1
F
Leaves
Expression level
PP-AA
AA-PP
PP-PP

Figure 5