


RESEARCH

Open Access



Extensive sequence duplication in *Arabidopsis* revealed by pseudo-heterozygosity

Benjamin Jaegle¹, Rahul Pisupati¹, Luz Mayela Soto-Jiménez¹, Robin Burns^{1,2}, Fernando A. Rabanal³ and Magnus Nordborg^{1*} 

*Correspondence:
magnus.nordborg@gmi.oeaw.ac.at

¹ Gregor Mendel Institute, Austrian Academy of Sciences, Vienna Biocenter, Vienna, Austria

² Department of Plant Sciences, University of Cambridge, Cambridge, UK

³ Max Planck Institute for Developmental Biology, Tübingen, Germany

Abstract

Background: It is apparent that genomes harbor much structural variation that is largely undetected for technical reasons. Such variation can cause artifacts when short-read sequencing data are mapped to a reference genome. Spurious SNPs may result from mapping of reads to unrecognized duplicated regions. Calling SNP using the raw reads of the 1001 *Arabidopsis* Genomes Project we identified 3.3 million (44%) heterozygous SNPs. Given that *Arabidopsis thaliana* (*A. thaliana*) is highly selfing, and that extensively heterozygous individuals have been removed, we hypothesize that these SNPs reflected cryptic copy number variation.

Results: The heterozygosity we observe consists of particular SNPs being heterozygous across individuals in a manner that strongly suggests it reflects shared segregating duplications rather than random tracts of residual heterozygosity due to occasional outcrossing. Focusing on such pseudo-heterozygosity in annotated genes, we use genome-wide association to map the position of the duplicates. We identify 2500 putatively duplicated genes and validate them using de novo genome assemblies from six lines. Specific examples included an annotated gene and nearby transposon that transpose together. We also demonstrate that cryptic structural variation produces highly inaccurate estimates of DNA methylation polymorphism.

Conclusions: Our study confirms that most heterozygous SNP calls in *A. thaliana* are artifacts and suggest that great caution is needed when analyzing SNP data from short-read sequencing. The finding that 10% of annotated genes exhibit copy-number variation, and the realization that neither gene- nor transposon-annotation necessarily tells us what is actually mobile in the genome suggests that future analyses based on independently assembled genomes will be very informative.

Keywords: Structural variation, Gene duplication, GWAS, SNP calling, Methylation



© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Background

With the sequencing of genomes becoming routine, it is evident that structural variants (SVs) play a major role in genome variation [1]. There are many kinds of SVs, e.g., indels, inversions, and translocations. Of particular interest from a functional point of view is gene duplication, leading to copy number variation (CNV).

Before Next-Generation Sequencing (NGS) was available, genome-wide detection of CNVs was achieved using DNA microarrays. These methods had severe weaknesses, leading to low resolution and problems detecting novel and rare mutations [2, 3]. With the development of NGS, our ability to detect CNVs increased dramatically, using tools based on split reads, paired-end mapping, sequencing depth, or even de novo assembly [4, 5]. In mammals, many examples of CNVs with a major phenotypic effect have been found [6–8]. One example is the duplication of MWS/MLS, associated with better trichromatic color vision [9].

While early investigation of CNV focused on mammals, several subsequent studies have looked at plant genomes. In *Brassica rapa*, gene CNV has been shown to be involved in morphological variation [10] and an analysis of the poplar “pan-genome” revealed at least 3000 genes affected by CNV [11]. It has also been shown that variable regions in the rice genome are enriched in genes related to defense to biotic stress [12]. More recently, the first chromosome-level assemblies of seven accessions of *A. thaliana* based on long-read sequencing were released [13], demonstrating that a large proportion of the genome is structurally variable. Similar studies have also been carried out in maize [14, 15], tomato [16], rice [17], and soybean [18]. These approaches are likely to provide a more comprehensive picture than short-read sequencing, but are also far more expensive.

In 2016, the 1001 Genomes Consortium released short-read sequencing data and SNP calls for 1135 *A. thaliana* accessions [19]. Several groups have used these data to identify large numbers of structural variants using split reads [20–22]. Here we approach this from a different angle. Our starting point is the startling observation that, when calling SNPs in the 1001 Genomes data set, we identified 3.3 million (44% of total) putatively heterozygous SNPs. In a highly selfing organism, this is obviously highly implausible, and these SNPs were flagged as spurious: presumably, products of cryptic CNV, which can generate “pseudo-SNPs” [23, 24] when sequencing reads from non-identical duplicates are (mis-)mapped to a reference genome that does not contain the duplication. Note that allelic SNP differences are expected to exist *ab initio* in the population, leading to instant pseudo-heterozygosity as soon as the duplicated copy recombines away from its template (as a consequence of outcrossing). In this paper, we return to these putative pseudo-SNPs and show that they are indeed largely due to duplications, the position of which can be precisely mapped using GWAS. Our approach is broadly applicable, and we demonstrate that it can reveal interesting biology.

Results

Massive pseudo-heterozygosity in the 1001 Genomes data

Given that *A. thaliana* is highly selfing, a large fraction (44%) of heterozygous SNPs is inherently implausible. Two other lines of evidence support the conclusion that they are

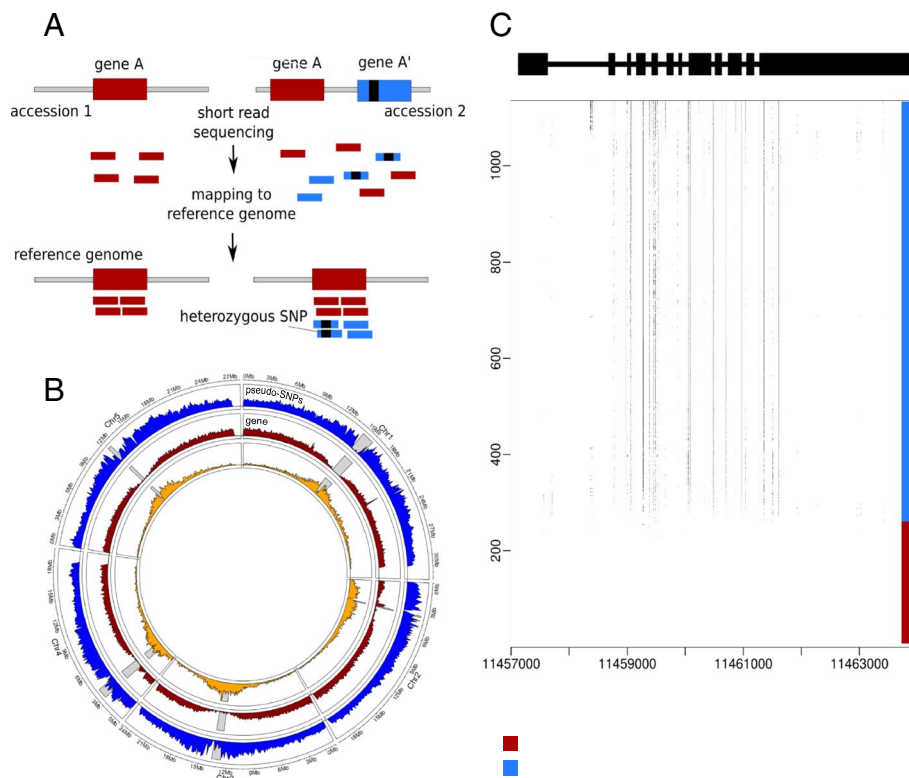


Fig. 1 Pseudo-heterozygosity in the 1001 Genomes dataset. **A** Cartoon illustrating how a duplication can generate pseudo-SNPs when mapping to a reference genome that does not contain the duplication. **B** Genomic density of transposons, genes, and shared heterozygous SNPs. Gray bars represent the position of the centromere for each chromosome. **C** The pattern of putative heterozygosity around AT1G31910 for the 1057 accessions. Dots in the plot represent putative heterozygosity

spurious. First, genuine residual heterozygosity would appear as large genomic tracts of heterozygosity in individuals with recent outcrossing in their ancestry. Being simply a random product of recombination and Mendelian segregation, there is no reason two individuals would share tracts unless they are very closely related. The observed pattern is completely the opposite. While a small number of individuals do show signs of recent outcrossing, this is quite rare (as expected given the low rate of outcrossing in this species, and the fact that the sequenced individuals were selected to be completely inbred). Instead, we find that the same SNP is often heterozygous in multiple individuals. Although the population level of heterozygosity at a given SNP is typically low (Additional file 1: Fig. S1), over a million heterozygous SNPs are shared by at least 5 accessions, and a closer look at the pattern of putative heterozygosity usually reveals short tracts of shared heterozygosity that would be vanishingly unlikely under residual heterozygosity, but would be expected if tracts represent shared duplications, and heterozygosity is, in fact, pseudo-heterozygosity due to mis-mapped reads (Fig. 1). Further supporting the notion that pseudo-heterozygous SNPs involve SNPs that already existed in the population before the duplication occurred, both alleles are also present as homozygotes for 97% of the putatively heterozygous SNPs. Analysis of the distribution of the lengths and the number of putatively heterozygous tracts across accessions shows

that the vast majority of accessions have a large number of very short tracts (roughly 1 kb) of heterozygosity (Additional file 1: Fig. S2). Longer tracts are rare and not shared between accessions.

Furthermore, the density of shared heterozygous SNPs is considerably higher around the centromeres (Fig. 1), which is again not expected under random residual heterozygosity, but is rather reminiscent of the pattern observed for transposons, where a similar pattern is interpreted as the result of selection removing insertions from euchromatic regions, leading to a build-up of common (shared) transposon insertions near centromere [25]. As we shall see below, it is likely that transposons play an important role in generating cryptic duplications leading to pseudo-heterozygosity (although we emphasize again that the heterozygous SNPs were called taking known repetitive sequences into account).

Despite the evidence for selection against these putative duplications, we found 2570 genes containing 26,647 common pseudo-SNPs (more than 5% [i.e., [26]] pseudo-heterozygous accessions; see (Additional file 1: Fig. S3). Gene-ontology analysis of these genes reveals an enrichment for biological processes involved in response to UV-B, bacteria or fungi (Additional file 1: Fig. S5). In the following sections, we investigate these putatively duplicated genes further.

Mapping common duplications using genome-wide association

If heterozygosity is caused by the presence of cryptic duplications in non-reference genomes, it should be possible to map the latter using GWAS with heterozygosity as a “phenotype” (Imprialou et al 2017). The principle here is that the extra copy that gives rise to pseudo-heterozygosity will be “tagged” by SNPs like any other causal allele (Fig. 2A). We did this for each of the 26,647 SNPs exhibiting shared heterozygosity within genes (Additional file 1: Fig. S3).

Of the 2570 genes that showed evidence of duplication, 2511 contained at least one major association (using a significance threshold of $p < 10^{-20}$; see the “Methods” section). For 708 genes, the association was more than 50 kb away from the pseudo-SNP used to define the phenotype, and for 175 it was within 50 kb (the usual extent of GWAS peaks in this species, see also (Additional file 1: Fig. S4). We will refer to these as *trans*- and *cis*-associations, respectively. The majority of genes, 1628, had both *cis*- and *trans*-associations (Fig. 2), suggesting that both the original and the duplicated were tagged by SNPs.

To validate these results, we assembled 6 non-reference genomes de novo using long-read PacBio sequencing. The GWAS results provide predicted locations of the duplications (the putative causes of pseudo-heterozygosity). We identified the homologous region of each non-reference genome, then used BLAST to search for evidence of duplication. For 84% of the 403 genes predicted to have a duplication present in at least one of the six non-reference genomes, evidence of a duplication was found; for 60%, the occurrence perfectly matched the pattern of heterozygosity across the six genomes. For the remaining 16%, no evidence of a duplication was found, which could be due to the stringent criteria we used to search for evidence of duplication (see [Methods](#)). The

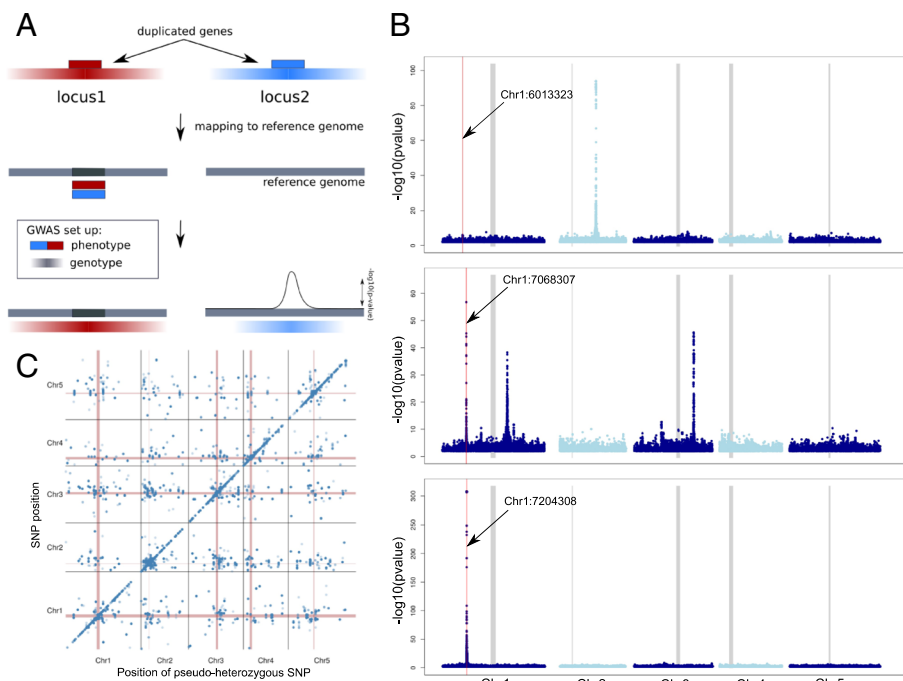


Fig. 2 GWAS of putative duplications. **A** Schematic representation of the principle of how GWAS can be used to detect the position of the duplicated genes based on linkage disequilibrium (LD). As phenotype, heterozygosity at the position of interest is coded as 1 (present) or 0 (absent). As a genotype, the SNPs matrix of the 1001 genome dataset was used (with heterozygous SNPs filtered out). Color gradients represent the strength of LD around the two loci. In this example, the reference genome does not contain locus2. **B** GWAS results for three different genes with evidence of duplication, for illustration. The red lines indicate the position of the pseudo-SNP used for each gene/GWAS and the thick gray lines indicate the centromeres. The top plot shows a *trans*-association, the bottom a *cis*-association, and the middle shows a case with both (*cis* plus two *trans*). **C** Summary of all 26,647 GWAS results

distribution of fragment sizes detected suggests that we capture a mixture of duplicated gene fragments and full genes (Additional file 1: Fig. S6).

Rare duplications

The GWAS approach has no power to detect rare duplications, which is why we restricted the analysis above to pseudo-heterozygous SNPs seen in five or more individuals. Yet most are rarer: 40% are seen only in a single individual, and 16% are seen in two. As it turns out, many of these appear to be associated with more common duplications. Restricting ourselves to genes only, 11.4% of the singleton pseudo-heterozygous SNPs are found in the 2570 genes already identified using common duplications, a significant excess ($p = 2.5e-109$). For doubletons, the percentage is 11.1% ($p = 1.9e-139$). Whether they are caused by the same duplications or reflect additional ones present at lower frequencies is difficult to say. To confirm duplications, we took the reads generating the singleton and doubleton pseudo-heterozygotes and compared the result of mapping them to the reference genome, and to the appropriate genome (derived from the same inbred line). One predicted consequence of the reads mapping at different locations is that mapping coverage around the pseudo-SNPs will be decreased when mapping to the newly assembled PacBio genomes rather than the reference genome. As

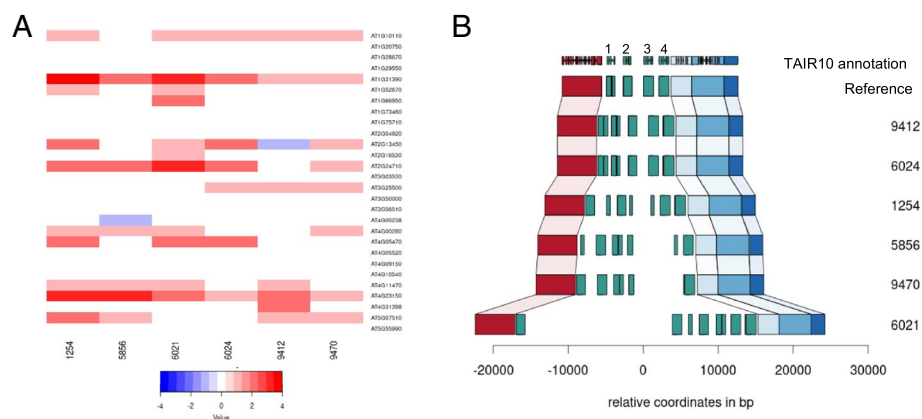


Fig. 3 Confirmation of tandem duplications. **A** The distribution of estimated copy number (based on sequencing coverage) across 6 PacBio genomes for 28 genes predicted to be involved in tandem duplications based on the analyses of this paper. **B** The duplication pattern observed in these genomes for the gene AT1G31390, as an example the reference genome contains four copies, shown as numbered green boxes. Other colored boxes denote other genes

expected, a high proportion of the SNPs tested have lower coverage when mapping to the PacBio genomes (Additional file 1: Fig. S7-8). In addition to a decrease in coverage, we were also able to confirm many duplications directly by demonstrating that reads map to multiple locations in their own genome, and that the putative heterozygosity disappeared when we did so. Of doubleton pseudo-heterozygotes, 41.5% are found in regions that are revealed as duplicated simply by mapping to the right genome instead of the reference genome (Additional file 1: Fig. S7-9).

Local duplications

If duplications arise via tandem duplications, they will not give rise to pseudo-SNPs until the copies have diverged via mutations. This is in contrast to unlinked copies, which will lead to pseudo-SNPs due to existing allelic variation as soon as recombination has separated copy from the original. We should thus expect the approach taken here to be biased against detecting local duplications. Nonetheless, GWAS revealed 175 genes with evidence only for a *cis* duplication. 28 of these were predicted to be present in at least one of the six new genomes, and 14 could be confirmed to have a local variation of copy number relative to the reference (Fig. 3A).

The local structure of the duplications can be complex. An example is provided by the gene AT1G31390, annotated as a member of MATH/TRAF-domain genes, and which appears to be present in 4 tandem copies in the reference genome, but which is highly variable between accessions, with one of our accessions carrying at least 6 copies (Fig. 3B). However, there are no copies elsewhere in any of the new genomes for this gene (Additional file 1: Fig. S10).

A transposon-driven duplication

Transposons are thought to play a major role in gene duplications, capturing and moving genes or gene fragments around the genome [27, 28]. While confirming the *trans* duplications in the PacBio genomes, we found a beautiful example of this process. The

gene AT1G20400 (annotated, based on sequence similarity, to encode a myosin-heavy chain-like protein) was predicted to have multiple *trans*-duplications. The 944 bp coding region contains 125 putatively heterozygous SNPs with striking haplotype structure characteristic of structural variation (Fig. 4C). We were able to identify the duplication predicted by GWAS in the six new genomes (Fig. 4). Four of the newly assembled genomes have only one copy of the gene, just like the reference genome, but one has 3 copies and one has 4 copies. However, none of the 6 new genomes has a copy in the same place as in the reference genome (Additional file 1: Fig. S11).

In the reference genome, AT1G20400 is closely linked to AT1G20390, which is annotated as a Gypsy element. This element also contains many pseudo-SNPs, and GWAS revealed duplication sites overlapping those for AT1G20400 (Fig. 4B). This suggested that the putative gene and putative Gypsy element transpose together, i.e., that both are misannotated, and that the whole construct is effectively a large transposable element. Further analysis of the PacBio genomes confirmed that AT1G20400 and AT1G20390 were always found together, and we were also able to find conserved Long Terminal Repeat sequences flanking the whole construct, as would be expected for a retrotransposon (Additional file 1: Fig. S12-13). We did not find any evidence for expression of AT1G20400 in RNAseq from seedlings in any of the accessions. Available bisulfite

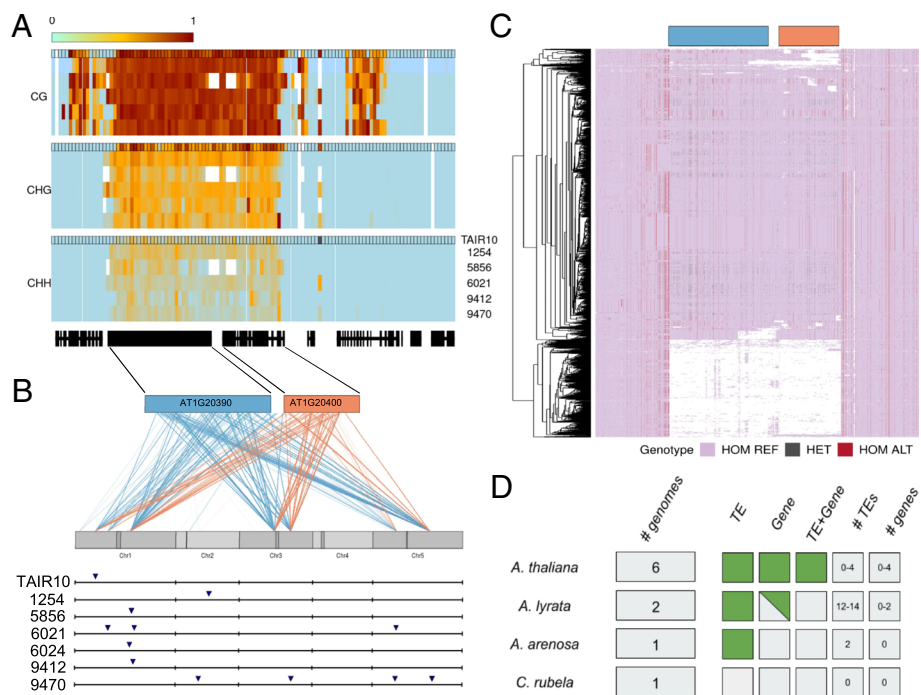


Fig. 4 A Gypsy element (AT1G20390) and a gene transpose (AT1G20400) together. **A** Methylation levels on regions containing AT1G20390 and AT1G20400 for 6 accessions, calculated in 200 bp windows after mapping reads to the TAIR10 reference genome (annotation outline in black). **B** GWAS results for the putatively heterozygous SNPs in AT1G20390 and AT1G20400. Each line represents the link between the position of the pseudo-SNP and a GWAS hit position in the genome. The lower part shows the presence of the new transposable element in the 6 PacBio genomes as well as in the reference genome. **C** SNP haplotypes around the AT1G20400 region in the 1001 genomes data. White represents a lack of coverage. **D** Presence of the gene and the transposon in related species

sequencing data [29] showed that the whole region is heavily methylated, as expected for a transposon (Fig. 4). In an attempt to look at the methylation pattern for each insertion separately, we tried mapping the bisulfite reads to the corresponding genome, but the coverage was too low and noisy to observe a difference in methylation between the multiple insertions (Additional file 1: Fig. S14).

Having located precise insertions in the six new genomes, we attempted to find them using short-read data in the 1001 Genomes dataset. Except for one insertion that was shared by 60% of accessions, the rest were found in less than 20%, suggesting that this new element has no fixed insertions in the genome — including the insertion found in the TAIR10 reference genome, which was only found in 17.4% of the accessions (Additional file 1: Fig. S15). We also looked for the element in the genomes of *A. lyrata* (two different genomes), *A. suecica* [a tetraploid containing an *A. thaliana* and an *A. arenosa* subgenome; see 29], and *Capsella rubella* [30]. The gene and the Gypsy element were only found together in *A. thaliana* (including the *A. thaliana* sub-genome of the allopolyploid *A. suecica*). The Gypsy element alone is present in the other *Arabidopsis* species, and the gene alone is present in *A. lyrata*, but only in one of two genomes. In *Capsella rubella*, neither the transposon nor the gene could be detected (Additional file 1: Fig. S16). Thus, the transposon and gene appear to be specific to the genus *Arabidopsis*, while their co-transposition is specific to *A. thaliana*, suggesting that the new transposable element evolved since the divergence of *A. thaliana* from the other member of the genus.

Spurious methylation polymorphism

Just like cryptic duplications can lead to spurious genetic polymorphisms, they can lead to spurious cytosine methylation polymorphisms. Indeed, given the well-established connection between gene duplication and gene silencing [e.g., [31], they may be more likely to do so. To investigate this, we re-examined the methylation status of genes previously reported by the 1001 Genomes Project [29] as having complex patterns of methylation involving both CG and CHG methylation. In our six sequenced accessions, we found 19530 genes that had been reported as having CG methylation (in at least one accession) and 2556 genes that had been reported as having CHG methylation (in at least one accession). 2473 genes were part of both sets. Out of these, 619, or 24%, had been detected as duplicated in the analyses presented above (a massive enrichment compared to the genome-wide fraction of roughly 10%). To understand these patterns better, we mapped the original bisulfite data to the appropriate genome as well as to the reference genome. In any given accession, roughly 7% of the 2473 genes could not be compared because the homologous copy could not be found (this is presumably mostly because they contain structural variants that prevent them being located by BLAST), and roughly 30% exhibited copy number variation (Table 1). The remaining genes had a single match, almost always in the same location as in the reference genome. These categories are shared across accessions: 1294 of the 2367 genes appeared to be single-copy in all six new genomes, for example (Table 1; Additional files 1-8).

Table 1 Number of copies of the 2367 genes identified in each new genome (and Araport11, as control)

Target	Number of copies identified		
	0	1	>1
1254	138	1563	772
5856	174	1566	733
6021	131	1577	765
6024	152	1554	767
9412	147	1567	759
9470	142	1589	742
<i>Intersection</i>	37	1294	610
Araport11	0	1721	752

Table 2 Fraction of differentially methylated genes when comparing bisulfite reads mapped to reference TAIR genome and to its respective PacBio genome, separated by gene copy number

Target	Number of copies identified			
	1		>1	
	CG (%)	CHG (%)	CG (%)	CHG (%)
1254	3.0	4.4	33.3	21.6
5856	1.2	3.7	27.8	42.9
6021	2.4	3.2	39.3	24.2
6024	3.0	4.2	41.2	29.5
9412	2.0	2.5	37.0	27.1
9470	2.1	4.7	36.0	26.2

Turning to the methylation patterns, the effect of cryptic copy number variation was obvious (Table 2). For the genes with a single match in both the reference and accession genome, methylation status calls based on mapping bisulfite sequencing reads to either genome were largely concordant (roughly 2.5% disagreement), whereas for genes with copy number variation, roughly one third of calls were wrong.

As an illustration for why this occurs, consider the methylation status of AT1G30140 (Fig. 5). When mapped to the reference genome, 5 out of 6 accessions were found to be both CG and CHG methylated, with accession 6021 having no methylation. When mapped to the appropriate genome, we see that this pattern can be quite misleading. In accession 1254, for example, we found three apparent copies of the gene, only two of which are methylated, neither of which is the copy corresponding to the copy present in the reference genome. In accession 5856, the copy corresponding to the reference genome cannot be identified, but a copy on a different chromosome is identified, and it is methylated. In both cases, mapping to the reference genome leads to incorrect methylation status for AT1G30140.

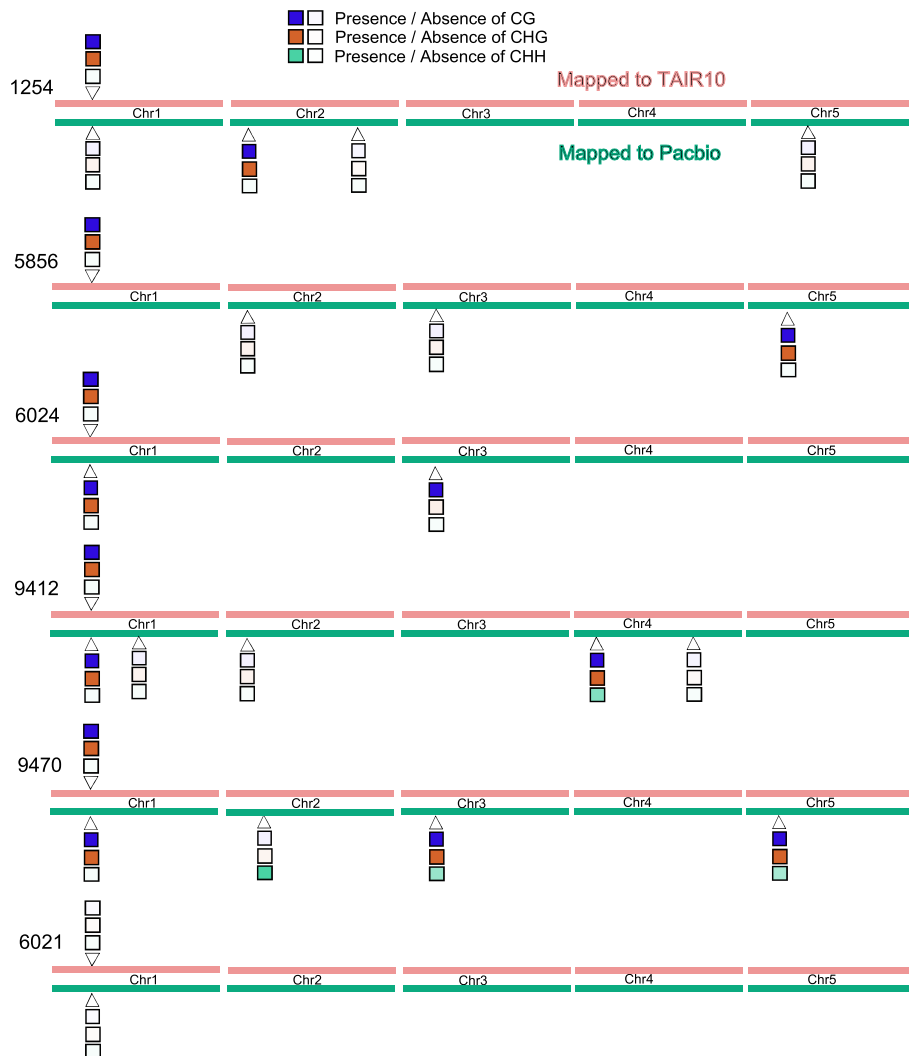


Fig. 5 The effect of calling methylation status for AT1G30140 by mapping to a reference genome vs. the appropriate genome. Locations on the chromosomes are approximate, for illustration only

Discussion

A duplication can lead to pseudo-SNPs when SNPs are identified by mapping short reads to a reference genome that does not contain the duplication. Typically pseudo-SNPs have to be identified using deviations from Hardy-Weinberg proportions, or via non-Mendelian segregation patterns in families or crosses, but in inbred lines, they can be identified solely by their heterozygosity. The overwhelming majority of the 3.3 million heterozygous SNPs identified by our SNP-calling of the 1001 Genomes Project data are likely to be pseudo-SNPs. Assuming this, we used (pseudo-)heterozygosity as a “phenotype,” and tried to map its cause, i.e., the duplication, using a simple but powerful GWAS approach. Focusing on annotated genes, we find that over 2500 (roughly 10% of the total) harbor pseudo-SNPs and show evidence of duplication. Using 6 new long-read assemblies, we were able to confirm 60% of these duplications using conservative criteria (see the “Methods” section). Most of the remaining duplications are located in

pericentromeric regions where SNP-calling has lower quality, and which are difficult to assemble even with long-read (Additional file 1: Fig. S17).

These numbers nearly certainly underestimate the true extent of duplication, which has been known to be common in *A. thaliana* for over a decade [32–34]. While unlinked *trans*-duplications are fairly likely to give rise to pseudo-SNPs, local *cis*-duplications will only do so once sufficient time has passed for substantial sequence divergence to occur, or if they arise via non-homologous recombination in a heterozygous individual (which is less likely in *A. thaliana*). As for the GWAS approach, it lacks statistical power to detect rare duplications and can be misled by allelic heterogeneity (due to multiple independent duplications). Finally, duplications are just a subset of structural variants, and it is therefore not surprising that other short-read approaches to detect such variants have identified many more using the 1001 Genomes data [20–22].

Pseudo-SNPs are not the only problem with relying on a reference genome. Our analysis uncovered a striking example of the potential importance of the “mobileome” in shaping genome diversity [35]: we show that an annotated gene and an annotated transposon are both part of a much larger mobile element, and the insertion in the reference genome is missing from most other accessions. When short reads from another accession are mapped to this “gene” using the reference genome, you are neither mapping to a gene, nor to the position you think. One possible consequence of this is incorrect methylation polymorphism calls, as we demonstrate above, but essentially any method that relies on mapping sequencing data to a reference genome could be affected (e.g., RNA-seq). It is important that users of such methods be aware of the biases that may result from mapping reads to the “wrong” genome.

Time (and more independently assembled genomes) will tell how significant this problem is, but the potential for artifactual results is clearly substantial, and likely depends on the amount of recent transposon activity [35]. It is also important to realize that the artifactual nature of the 44% heterozygous SNPs was only apparent because we are working with inbred lines. Other researchers working on inbred lines have reached similar conclusions and used various methods to eliminate them, e.g., *Zea* [36–38] and *Brachypodium* [39]. In human genetics, SNP-calling relies heavily on large sample sizes and Hardy-Weinberg assumptions (or even on family trios), but in outcrossing organisms where this is not possible, less precise approaches must be used (e.g., considering the pattern of linkage disequilibrium between putative closely linked SNPs, or variable sequencing coverage). In species with multiple reference genomes, one may be used for quality control, to give an idea of the magnitude of the problem. Ultimately, however, the only real solution to the problem discussed here is moving away from genotyping using short sequence fragments that are prone to mismapping. Our increasing ability to sequence complex genomes will allow population analyses to avoid using such methods and will reveal new mechanisms of genome evolution in the process.

Conclusions

We have shown that the massive number of heterozygous SNPs identified in the 1001 Arabidopsis Genomes Projects are generally pseudo-heterozygous, mostly due to mis-mapping of reads from segregating duplications, and that many of these duplications can be mapped using GWAS with (pseudo-)heterozygosity as a phenotype. We exemplify the phenomenon by demonstrating that an annotated gene is in fact part of a nearby annotated transposon and that the two generally jumped together, resulting in extensive variation. We note that segregating duplications can bias many analyses and that great caution is needed when mapping short-read sequencing data to the “wrong” genome.

Methods

Long-read sequencing of six *A. thaliana*

We sequenced six Swedish *A. thaliana* lines that are part of the 1001 Genomes collection [19], ecotype ids: 1254, 5856, 6021, 6024, 9412, and 9470. Plants were grown in the growth chamber at 21 C in long-day settings for 3 weeks and dark-treated for 24-48 hours before being collected. DNA was extracted from ~20 g of frozen whole seedling material following a high molecular weight DNA extraction protocol adapted for plant tissue [40]. All six genomes were sequenced with PacBio technology, 6021 with PacBio RSII, and the rest with Sequel. The N50 reads length of the accessions 1254, 5856, 6021, 6024, 9412, and 9470 are respectively 27,875, 14,140, 20,989, 28,613, 25,243, and 30,129 bp. Accession 9412 was sequenced twice and 6024 was additionally sequenced with Nanopore (4.1 Gbp sequenced, 376 K reads. All data were used in the assemblies.

MinION sequencing of two *A. lyrata*

We sequenced two North American *A. lyrata* accessions, 11B02 and 11B21. Both individuals come from the 11B population of *A. lyrata*, which is self-compatible and located in Missouri [41] (GPS coordinates 38° 28' 07.1" N; 90° 42' 34.3" W). Plants were bulked for 1 generation in the lab and DNA was extracted from ~20g of 3-week-old seedlings, grown at 21°C, and dark treated for 3 days prior to tissue collection. DNA was extracted using a modified protocol for high molecular weight DNA extraction from plant tissue. DNA quality was assessed with a Qubit fluorometer and a Nanodrop analysis. We used a Spot-ON Flow Cell FLO-MIN106D R9 Version with a ligation sequencing kit SQK-LSK109. Bases were called using guppy version 3.2.6 (<https://nanoporetech.com/community>). The final output of MinION sequencing for 11B02 was 13.67 Gbp in 763,800 reads and an N50 of 31.15 Kb. The final output of MinION sequencing for 11B21 was 17.55 Gb, 1.11 M reads with an N50 of 33.26 Kb.

Genome assembly, polishing, and scaffolding

The six *A. thaliana* genomes (ecotype ids 1254, 5856, 6021, 6024, 9412, and 9470) were assembled using Canu (v 1.7.1) [42] with default settings, except for genome size. Previous estimates of flow cytometry were used for this parameter [43] when available or 170m was used. The values were 170m, 178m, 135m, 170m, 170m and 170m, respectively. The assemblies were corrected with two rounds of arrow (PacBio's SMRT Link software release 5.0.0.6792) and one of Pilon [44]. For arrow, the respective long reads

were used and for Pilon, the 1001 Genomes DNA sequencing data, plus PCR-free Illumina 150bp data that was generated for accessions 6024 and 9412; lines 5856, 6021, 9470 had available PCR-free data (250bp reads generated by David Jaffe, Broad Institute). This resulted in 125.6Mb, 124.3Mb, 124.5Mb, 124.7Mb, 127.1 Mb, and 128 Mb assembled bases, respectively; contained in 99, 436, 178, 99, 109, and 124 contigs, respectively. For the accessions 1254, 5856, 6024, 9412, and 9470, the average contigs length is 68,880, 25,852, 85,789, 18,7562, and 85,104 bp, respectively. The polished contigs were ordered and scaffolded with respect to the Col-0 reference genome, using RaGOO [45].

We assembled the genome of the two *A. lyrata* accessions 11B02 and 11B21 using Canu [42] (v 1.8) with default settings and a genome size set to 200 Mb. The genomes of 11B02 and 11B21 were contained in 498 and 265 contigs, respectively. The contig assemblies were polished using Racon [46] (v 1.4) and ONT long reads were mapped using nglmr [47] (v 0.2.7). Assemblies were further polished by mapping PCR-free Illumina 150bp short reads (~100X for 11B02 and ~88X for 11B21) to the long-read corrected assemblies. Short-read correction of assembly errors was carried out using Pilon [44] (v1.23). Contigs were scaffolded into pseudo-chromosomes using RaGOO (Alonge et al. 2019) and by using the error-corrected long reads from Canu and the *A. lyrata* reference genome [48] and the *A. arenosa* subgenome of *A. suecica* [49] as a guide followed by manual inspection of regions. The assembly size for 11B02 was 213Mb and 11B21 was 202Mb. Genome size was estimated using findGSE [50] with a resulting estimated genome size of ~256Mb for 11B02 and ~237Mb for 11B21.

SNPs calling / extraction

We downloaded short-read data for 1,057 accessions from the 1001 Genomes Project [19]. Raw paired-end reads were processed with cutadapt (v1.9) [51] to remove 3' adapters, and to trim 5'-ends with quality 15 and 3'-ends with quality 10 or N-endings. All reads were aligned to the *A. thaliana* TAIR10 reference genome [52] with BWA-MEM (v0.7.8) [53], and both Samtools (v0.1.18) and Sambamba (v0.6.3) were used for various file format conversions, sorting and indexing [54, 55], while duplicated reads were by marked by Markduplicates from Picard (v1.101; <http://broadinstitute.github.io/picard/>). Further steps were carried out with GATK (v3.4) functions [26, 56]. Local realignment around indels were done with “RealignerTargetCreator” and “IndelRealigner,” and base recalibration with “BaseRecalibrator” by providing known indels and SNPS from The 1001 Genomes Consortium [19]. Genetic variants were called with “HaplotypeCaller” in individual samples followed by joint genotyping of a single cohort with “GenotypeGVCFs.” An initial SNP filtering was done following the variant quality score recalibration (VQSR) protocol. Briefly, a subset of ~181,000 high-quality SNPs from the RegMap panel [57] was used as the training set for VariantRecalibrator with a priori probability of 15 and four maximum Gaussian distributions. Finally, only bi-allelic SNPs within a sensitivity tranche level of 99.5 were kept, for a total of 7,311,237 SNPs.

Heterozygous stretches analysis

From the VCF, Plink was used to generate .ped and .map files. (<http://pngu.mgh.harvard.edu/purcell/plink/>) [58]. To detect and characterize the stretches of heterozygosity the package “detectRUNS” in R was then used. (<https://github.com/bioinformatics-ptp/>

`detectRUNS/tree/master/detectRUNS`). We used the function `slidingRuns.run` with the following parameters: `WindowSize=10`, `threshold=0.05`, `RoHet=True`, `minDensity=1/100`, rest as default.

SNP filtering

From the raw VCF files SNP positions containing heterozygous labels were extracted using GATK VariantFiltration. From the 3.3 million of heterozygous SNPs extracted, two filtering steps were then applied. Only SNPs with a frequency of at least 5% of the population and located in TAIR10-annotated coding regions were kept. After those filtering steps a core set of 26,647 SNPs were retained for further analysis (see Additional file 1: Fig. S17). Gene names and features containing those pseudo-SNPs were extracted from the TAIR10 annotation.

GWAS

The presence and absence of pseudo-heterozygosity at a given site (coded as 1 and 0 respectively) was used as a phenotype to run GWAS. As a genotype, the matrix published by the 1001 Genomes Consortium containing 10 million SNPs was used [19]. To run all the GWAS, the `pygwas` package [<https://github.com/timeu/PyGWAS>; see [59]] with the `amm` (accelerated mixed model) option was used. The raw output containing all SNPs was filtered, removing all SNPs with a minor allele frequency below 0.05 and/or a $-\log_{10}(p\text{-value})$ below 4.

For each GWAS performed, the p -value as well as the position was used to call the peaks using the Fourier transform function in R (`filterFFT`), combined with the peak detection function (`peakDetection`), from the package `NucleR` 3.13, to automatically retrieve the position of each peak across the genome. From each peak, the highest SNPs within a region of ± 10 kb around the peak center were used (see the example in Additional file 1: Fig. S18). Using all 26647 SNPs, a summary table was generated with each pseudo-heterozygous SNP and each GWAS peak detected (Additional file 2). This matrix was then used to generate Fig. 2C, applying thresholds of $-\log_{10}(p\text{-value})$ of 20 and a minor allele frequency of 0.1.

Confirmation of GWAS results

To confirm the detected duplications, a combination of BLAST and synteny was used on the denovo-assembled genome. Only the insertions that segregate in the 6 new genomes were used (398). For each gene, the corresponding sequence from the TAIR10 annotation was located in the target genome using BLAST (see Additional file 1: Fig. S6). A threshold of 70% sequence identity as well as 70% of the initial sequence length was used. The presence of a match within 20kb of the predicted peak position was interpreted as confirmation.

Gene ontology

Out of the 2570 genes detected to be duplicated, 2396 have a gene ontology annotation. PLAZA.4 [60] was used to perform a gene enrichment analysis using the full genome as background. Data were then retrieved and plotted using R.

Coverage and methylation analysis

Bisulfite reads for the accessions were taken from 1001 methylomes (Kawakatsu et al. 2016). Reads were mapped to PacBio genomes using an nf-core pipeline (<https://github.com/rbpisupati/methylseq>). We filtered for cytosines with a minimum depth of 3. Their methylation levels were calculated either on the gene-body or on 200bp windows using custom python scripts following guidelines from Schultz et al. [61]. Weighted methylation levels were used, i.e., if there are three cytosines with a depth of t1, t2, and t3 and number of methylated reads are c1, c2, and c3, the methylation level was calculated as $(c1+c2+c3)/(t1+t2+t3)$. We called a gene “differentially methylated” if the difference in weighted methylation level was more than 0.05 for CG and 0.03 for CHG.

The sequencing coverage for each accession was extracted using the function `bamCoverage` (windows size of 50bp) from the program DeepTools [62]. The Bigwig files generated were then processed in R using the package `rtracklayer`. No correlation between the mean sequencing coverage and the number of pseudo-SNPs detected was observed (Additional file 1: Fig. S18).

Multiple sequence alignment

For each insertion of the AT1G20390–AT1G20400 (Transposon+gene) fragment, a fasta file including 2kb on each side of the fragment was extracted from each genome, using the `getfasta` function from `bedtools` [63]. Multiple alignment was performed using KALIGN [64]. Visualization and comparison were done using Jalview 2 [65].

Structural variation analysis

To control the structure of the region around duplicated genes, the sequence from 3 genes upstream and downstream of the gene of interest was extracted. Each sequence was then BLAST to each of the genomes and the position of each BLAST result was retrieved. NCBI BLAST [66] was used with a percentage of identity threshold of 70% and all other parameters as default. From each blast results fragments with at least 50% of the input sequence length have been selected and plotted using R.

Frequency of the insertions in the 1001 Genomes dataset

The same sequences used for the multiple alignment were used to confirm the presence or absence of each insertion in the 1001 Genomes dataset. We used each of those sequences as a reference to map short reads using `minimap 2` [67]. For each insertion, only paired-end reads having both members of the pair mapping to the region were retained. An insertion was considered present in an accession if at least 3 pairs of reads spanned the insertion border (see Additional file 1: Fig. S12).

Multiple species comparison

We used the *Capsella rubella* and *A.arenosa* genomes [30, 49] to search for the new Transposon+gene element, just like in the *A. thaliana* genomes. For *A. arenosa* we used the subgenome of *A. suecica*. We located the transposon+gene fragments, extracted from the TAIR10 annotation, using NCBI BLAST as above. For *A.lyrata* two newly assembled genomes were assembled using MinION sequencing.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13059-023-02875-3>.

Additional file 1: Fig S1. Distribution of heterozygosity across individuals and the genome. **Fig S2.** Tract-length distribution. **Fig S3.** SNP filtering scheme. The SNPs matrix we started with contains 7 million SNPs. Of those, 3.3 million were detected as heterozygous in at least one line. We selected the 48,799 SNPs that we called heterozygous in at least 5 % of the lines, and focused on the 26,647 found in coding regions (according to the TAIR10 annotation). **Fig S4.** The distribution of distances between the “position” of a pseudo-SNPs and the corresponding GWAS peak. Out of the 41 1688 peaks we found that 236494 (57%) are on a different chromosome. The plots show the distribution of distances for the 175194 peaks that were on the same chromosome as the corresponding pseudo-SNPs. Roughly 50% of there are more distance than 50kb. **Fig S5.** Gene ontology analysis of the putatively duplicated genes. **Fig S6.** Pipeline to confirm GWAS peaks. **Fig S7.** Mapping reads tagging singleton pseudo-SNP. First, all reads overlapping the position of a specific pseudo-SNP were extracted based on mapping to the reference genome (TAIR10). This set of reads were then re-mapped to the appropriate Pacbio genome. Reads mapping to multiple regions indicate the presence of a duplicated segment. A decrease in coverage compared to the mapping to the reference genome is also a confirmation that reads map at different positions. An example is presented in **Fig S9**. **Fig S8.** Comparison of mapping coverage between reference genome and PacBio genomes for all regions surrounding pseudo-heterozygous doubleton positions. **Fig S9.** Mapping of reads overlapping singletons (example). **Fig S10.** Position of AT1G31390 in the reference (accession 6909) and each of the six newly assembled genomes. BLAST-thresholds of 70% identity were used, and only fragments of length greater than 50% of the original gene length are shown. **Fig S11.** Position of AT1G20400 in the reference (accession 6909) and each of the six newly assembled genomes. Cf. **Fig S10**. **Fig S12.** Dot-plots of the end of insertion B. LTR repeats can be detected on each side of the insertion. Cf. **Fig S13**. **Fig S13.** Dot plot of the ends of all insertions. Cf. **Fig S12**. **Fig S14.** Methylation profile across all copies of the TE+gene insertions. Bisulfite reads were mapped to the appropriate genomes and profiles extracted based on the inferred locations. Colors distinguish the different insertions found in each genome. **Fig S15.** Frequency of insertions of the new element in the 1001 Genomes. **Fig S16.** Mapping of the AT1G20400 region in multiple species. The rectangle corresponds to annotated genes around AT1G20400 in the *A. thaliana* reference genome. **Fig S17.** Chromosomal position of putatively duplicated genes that could not be confirmed using full genome sequence. The figure shows the distribution of these genes compared to the distribution of all genes, and RepeatMasker annotation. **Fig S18.** Illustration of GWAS peak calling. The p-values from GWAS were used to run the filterFFT function from the NucleR package in R, using 0.05 for pKeepComp option. **Fig S19.** The relationship between the number of pseudo-SNPs and sequencing coverage. Each dot represents an accession.

Additional file 2. Methylation value per gene of all accessions mapped to the reference genome. CG and CHG weighted average per genes of the 6 accessions analyzed. Row names correspond to the gene ID and column name to the CG and CHG for each accession.

Additional file 3. Methylation value per gene of all accessions mapped to the corresponding genome. CG and CHG weighted average per genes of the 6 accessions analyzed. Row names correspond to the gene ID. (the “_” corresponds to the multiple copies detected). The column name to the CG and CHG for each accession.

Additional file 4. Methylation value per gene of all accessions mapped to the corresponding genome. CG and CHG weighted average per genes of the 6 accessions analyzed. Row names correspond to the gene ID. (the “_” corresponds to the multiple copies detected). The column name to the CG and CHG for each accession.

Additional file 5. Methylation value per gene of all accessions mapped to the corresponding genome. CG and CHG weighted average per genes of the 6 accessions analyzed. Row names correspond to the gene ID. (the “_” corresponds to the multiple copies detected). The column name to the CG and CHG for each accession.

Additional file 6. Methylation value per gene of all accessions mapped to the corresponding genome. CG and CHG weighted average per genes of the 6 accessions analyzed. Row names correspond to the gene ID. (the “_” corresponds to the multiple copies detected). The column name to the CG and CHG for each accession.

Additional file 7. Methylation value per gene of all accessions mapped to the corresponding genome. CG and CHG weighted average per genes of the 6 accessions analyzed. Row names correspond to the gene ID. (the “_” corresponds to the multiple copies detected). The column name to the CG and CHG for each accession.

Additional file 8. Methylation value per gene of all accessions mapped to the corresponding genome. CG and CHG weighted average per genes of the 6 accessions analyzed. Row names correspond to the gene ID. (the “_” corresponds to the multiple copies detected). The column name to the CG and CHG for each accession.

Additional file 9. Methylation value per gene of all accessions mapped to the corresponding genome. CG and CHG weighted average per genes of the 6 accessions analyzed. Row names correspond to the gene ID. (the “_” corresponds to the multiple copies detected). The column name to the CG and CHG for each accession.

Additional file 10. Review history.

Acknowledgements

We thank numerous people on Twitter for providing feedback on the bioRxiv version.

Review history

The review history is available as Additional file 10.

Peer review information

Wenjing She was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions

BJ and MN developed the project. BJ performed all analyses. LMS and RB assembled the *A.thaliana* and *A.lyrata* genomes, respectively. FR generated the SNP matrix. RP performed the methylation analyses. BJ and MN wrote the manuscript, with input from all authors.

Funding

This project received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 789037), as well as from the ERA-CAPS program (FWF I 3684-B25).

Availability of data and materials

All genome assemblies and raw reads were deposited under the BioProject ID: PRJNA779205 <https://www.ncbi.nlm.nih.gov/bioproject/PRJNA779205> [68].

Scripts used are available under Github with a GPL-3.0 license link: <https://github.com/benjj212/duplication-paper.git> [69] as well deposited <https://doi.org/10.5281/zenodo.7555970> [70]. All scripts are publicly available.

The full GWAS matrix is available at <https://doi.org/10.5281/zenodo.5702395> [71].

Declarations**Ethics approval and consent to participate**

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 23 November 2021 Accepted: 13 February 2023

Published online: 09 March 2023

References

1. Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. *Nat Rev Genet.* 2011;12:363–76.
2. Carter NP. Methods and strategies for analyzing copy number variation using DNA microarrays. *Nat Genet.* 2007;39:S16–21.
3. Snijders AM, Nowak N, Segreaves R, Blackwood S, Brown N, Conroy J, et al. Assembly of microarrays for genome-wide measurement of DNA copy number. *Nat Genet.* 2001;29:263–4.
4. Shendure J, Ji H. Next-generation DNA sequencing. *Nat Biotechnol.* 2008;26:1135–45.
5. Zhao M, Wang Q, Wang Q, Jia P, Zhao Z. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics.* 2013;14(Suppl 11):S1.
6. Gonzalez E, Kulkarni H, Bolivar H, Mangano A, Sanchez R, Catano G, et al. The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science.* 2005;307:1434–40.
7. Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, et al. Diet and the evolution of human amylase gene copy number variation. *Nat Genet.* 2007;39:1256–60.
8. Handsaker RE, Korn JM, Nemesh J, McCarroll SA. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat Genet.* 2011;43:269–76.
9. Miyahara E, Pokorny J, Smith VC, Baron R, Baron E. Color vision in two observers with highly biased LWS/MWS cone ratios. *Vis Res.* 1998;38:601–12.
10. Lin K, Zhang N, Severing EI, Nijveen H, Cheng F, Visser RGF, et al. Beyond genomic variation - comparison and functional annotation of three Brassica rapagenomes: a turnip, a rapid cycling and a Chinese cabbage. *BMC Genomics.* 2014;15:250.
11. Pinosio S, Giacomello S, Fèvre-Rampant P, Taylor G, Jorge V, Le Paslier MC, et al. Characterization of the poplar pan-genome by genome-wide identification of structural variation. *Mol Biol Evol.* 2016;33:2706–19.
12. Yao W, Li G, Zhao H, Wang G, Lian X, Xie W. Exploring the rice dispensable genome using a metagenome-like assembly strategy. *Genome Biol.* 2015;16:187.
13. Jiao W-B, Schneeberger K. Chromosome-level assemblies of multiple Arabidopsis thaliana accessions reveal hot-spots of genomic rearrangements. *bioRxiv.* 2019:738880. Available from: <https://www.biorxiv.org/content/10.1101/738880v1>. Cited 2019 Sep 13.
14. Li C, Xiang X, Huang Y, Zhou Y, An D, Dong J, et al. Long-read sequencing reveals genomic structural variations that underlie creation of quality protein maize. *Nat Commun.* 2020;11:17.
15. Hufford MB, Seetharam AS, Woodhouse MR, Chougule KM, Ou S, Liu J, et al. De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes: Cold Spring Harbor Laboratory; 2021. p. 2021.01.14.426684. Available from: <https://www.biorxiv.org/content/10.1101/2021.01.14.426684v1>. Cited 2021 Jan 22.
16. Alonge M, Wang X, Benoit M, Soyk S, Pereira L, Zhang L, et al. Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell.* 2020. <https://doi.org/10.1016/j.cell.2020.05.021>.
17. Zhou Y, Chebotarov D, Kudrna D, Llaca V, Lee S, Rajasekar S, et al. A platinum standard pan-genome resource that represents the population structure of Asian rice. *Sci Data.* 2020;7:113.

18. Liu Y, Du H, Li P, Shen Y, Peng H, Liu S, et al. Pan-genome of wild and cultivated soybeans. *Cell*. 2020. <https://doi.org/10.1016/j.cell.2020.05.023>.
19. 1001 Genomes Consortium. 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell*. 2016;166:481–91.
20. Göktaş M, Fulgione A, Hancock AM. A new catalogue of structural variants in 1301 *A. thaliana* lines from Africa, Eurasia and North America reveals a signature of balancing at defense response genes. *Mol Biol Evol*. 2020. <https://doi.org/10.1093/molbev/msaa309>.
21. Zmienko A, Marszałek-Zenczak M, Wojciechowski P, Samelak-Czajka A, Luczak M, Kozłowski P, et al. AthCNV: a map of DNA copy number variations in the *Arabidopsis* genome. *Plant Cell*. 2020;32:1797–819.
22. Liu D-X, Rajaby R, Wei L-L, Zhang L, Yang Z-Q, Yang Q-Y, et al. Calling large indels in 1047 *Arabidopsis* with Indel-Ensembler. *Nucleic Acids Res*. 2021. <https://doi.org/10.1093/nar/gkab904>.
23. Ranade K, Chang MS, Ting CT, Pei D, Hsiao CF, Olivier M, et al. High-throughput genotyping with single nucleotide polymorphisms. *Genome Res*. 2001;11:1262–8.
24. Hurler M. Are 100,000 “SNPs” useless? *Science*. 2002;298(5598):1509. <https://doi.org/10.1126/science.298.5598.1509a>.
25. Quadrona L, Silveira AB, Mayhew GF, LeBlanc C, Martienssen RA, Jeddloh JA, et al. The *Arabidopsis thaliana* mobilome and its impact at the species level. *ELife Sci*. 2016;5:e15716 eLife Sciences Publications Limited.
26. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*. 2011;43:491–8.
27. Woodhouse MR, Pedersen B, Freeling M. Transposed genes in *Arabidopsis* are often associated with flanking repeats. *PLoS Genet*. Public Library of Science. 2010;6:e1000949.
28. Lisch D. How important are transposons for plant evolution? *Nat Rev Genet*. 2013;14:49–61.
29. Kawakatsu T, Huang S-SC, Jupe F, Sasaki E, Schmitz RJ, Urlich MA, et al. Epigenomic diversity in a global collection of *Arabidopsis thaliana* accessions. *Cell*. 2016;166:492–505 Elsevier.
30. Slotte T, Hazzouri KM, Ågren JA, Koenig D, Maumus F, Guo Y-L, et al. The *Capsella rubella* genome and the genomic consequences of rapid mating system evolution. *Nat Genet*. 2013;45:831–5 Nature Publishing Group.
31. Melquist S, Luff B, Bender J. *Arabidopsis* PAI gene arrangements, cytosine methylation and expression. *Genetics*. 1999;153:401–13.
32. Cao J, Schneeberger K, Ossowski S, Günther T, Bender S, Fitz J, et al. Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nat Genet*. 2011;43:956–63.
33. Gan X, Stegle O, Behr J, Steffen JG, Drewe P, Hildebrand KL, et al. Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. *Nature*. 2011;477:419–23.
34. Schneeberger K, Ossowski S, Ott F, Klein JD, Wang X, Lanz C, et al. Reference-guided assembly of four diverse *Arabidopsis thaliana* genomes. *Proc Natl Acad Sci U S A*. 2011;108:10249–54.
35. Morgante M, Brunner S, Pea G, Fengler K, Zuccolo A, Rafalski A. Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. *Nat Genet*. 2005;37:997–1002.
36. Chia J-M, Song C, Bradbury PJ, Costich D, de Leon N, Doebley J, et al. Maize HapMap2 identifies extant variation from a genome in flux. *Nat Genet*. 2012;44:803–7.
37. Lu F, Romay MC, Glaubitz JC, Bradbury PJ, Elshire RJ, Wang T, et al. High-resolution genetic mapping of maize pan-genome sequence anchors. *Nat Commun*. 2015;6:6914.
38. Bukowski R, Guo X, Lu Y, Zou C, He B, Rong Z, et al. Construction of the third-generation *Zea mays* haplotype map. *Gigascience*. 2018;7:1–12.
39. Stritt C, Gimmi EL, Wyler M, Bakali AH, Skalska A, Hasterok R, et al. Migration without interbreeding: Evolutionary history of a highly selfing Mediterranean grass inferred from whole genomes. *Mol Ecol*. 2021. <https://doi.org/10.1111/mec.16207>.
40. Cristina Barragan A, Collenberg M, Schwab R, Kerstens M, Bezrukov I, Bemm F, et al. Homozygosity at its Limit: Inbreeding Depression in Wild *Arabidopsis arenosa* Populations. *bioRxiv*. 2021:2021.01.24.427284. Available from: <https://www.biorxiv.org/content/10.1101/2021.01.24.427284v1>. Cited 2021 Nov 15.
41. Griffin PC, Willi Y. Evolutionary shifts to self-fertilisation restricted to geographic range margins in North American *Arabidopsis lyrata*. *Ecol Lett*. 2014;17:484–90.
42. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res*. 2017;27:722–36.
43. Long Q, Rabanal FA, Meng D, Huber CD, Farlow A, Platzer A, et al. Massive genomic variation and strong selection in *Arabidopsis thaliana* lines from Sweden. *Nat Genet*. 2013;45:884–90.
44. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*. 2014;9:e112963.
45. Alonge M, Soyk S, Ramakrishnan S, Wang X, Goodwin S, Sedlazeck FJ, et al. RaGOO: fast and accurate reference-guided scaffolding of draft genomes. *Genome Biol*. 2019;20:224.
46. Vaser R, Sović I, Nagarajan N, Šikić M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res*. 2017;27:737–46.
47. Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, et al. Accurate detection of complex structural variations using single-molecule sequencing. *Nat Methods*. 2018;15:461–8.
48. Hu TT, Pattyn P, Bakker EG, Cao J, Cheng J-F, Clark RM, et al. The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nat Genet*. 2011;43:476–81.
49. Burns R, Mandáková T, Gunis J, Soto-Jiménez LM, Liu C, Lysak MA, et al. Gradual evolution of allopolyploidy in *Arabidopsis suecica*. *Nat Ecol Evol*. 2021;5:1367–81.
50. Sun H, Ding J, Piednoël M, Schneeberger K. findGSE: estimating genome size variation within human and *Arabidopsis* using k-mer frequencies. *Bioinformatics*. 2018;34:550–7.
51. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J*. 2011;17:10–2.
52. *Arabidopsis* Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*. 2000;408:796–815.

53. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv. 2013:1303.3997 [q-bio.GN]. Available from: <http://arxiv.org/abs/1303.3997>.
54. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25:2078–9.
55. Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. Sambamba: fast processing of NGS alignment formats. *Bioinformatics*. 2015;31:2032–4.
56. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics*. 2013;11:11.10.1–11.10.33.
57. Horton MW, Hancock AM, Huang YS, Toomajian C, Atwell S, Auton A, et al. Genome-wide patterns of genetic variation in worldwide *Arabidopsis thaliana* accessions from the RegMap panel. *Nat Genet*. 2012;44:212–6 Nature Publishing Group.
58. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007;81:559–75.
59. Seren Ü, Vilhjálmsson BJ, Horton MW, Meng D, Forai P, Huang YS, et al. GWAPP: a web application for genome-wide association mapping in *Arabidopsis*. *Plant Cell*. 2012;24:4793–805.
60. Van Bel M, Diels T, Vancaester E, Kreft L, Botzki A, Van de Peer Y, et al. PLAZA 4.0: an integrative resource for functional, evolutionary and comparative plant genomics. *Nucleic Acids Res*. 2018;46:D1190–6.
61. Schultz MD, Schmitz RJ, Ecker JR. “Leveling” the playing field for analyses of single-base resolution DNA methylomes. *Trends Genet*. 2012;28:583–5.
62. Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, et al. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res*. 2016;44:W160–5.
63. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26:841–2.
64. 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:W636–41.
65. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics*. 2009;25:1189–91.
66. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990;215:403–10.
67. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. 2018;34:3094–100.
68. BioProject. Available from: <https://www.ncbi.nlm.nih.gov/bioproject/PRJNA779205>. Cited 2023 Jan 20.
69. Jaegle B. R-Script for Duplication paper. 2023. Available from: <https://github.com/benjj212/duplication-paper.git>.
70. Jaegle B. R-Script for Duplication paper. 2023. Available from: <https://zenodo.org/record/7555970>.
71. Jaegle B. GWAS Matrix. 2021. Available from: <https://zenodo.org/record/5702395>.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

