

LD-CNV: rapid and simple discovery of chromosomal translocations using linkage disequilibrium between copy number variable loci

Luca Comai ^{1,*}, Kirk R. Amundson,¹ Benny Ordoñez ¹, Xin Zhao,^{1,†} Guilherme Tomaz Braz ², Jiming Jiang ^{2,3} and Isabelle M. Henry ¹

¹Department of Plant Biology and Genome Center, University of California, Davis, Davis, CA 95616, USA,

²Department of Plant Biology, Michigan State University, East Lansing, MI 48824, USA and

³Department of Horticulture, Michigan State University, East Lansing, MI 48824, USA

[†]Present address: Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, Beijing, China.

*Corresponding author: Email: lcomai@ucdavis.edu

Abstract

Large-scale structural variations, such as chromosomal translocations, can have profound effects on fitness and phenotype, but are difficult to identify and characterize. Here, we describe a simple and effective method aimed at identifying translocations using only the dosage of sequence reads mapped on the reference genome. We binned reads on genomic segments sized according to sequencing coverage and identified instances when copy number segregated in populations. For each dosage-polymorphic 1 Mb bin, we tested independence, effectively an apparent linkage disequilibrium (LD), with other variable bins. In nine potato (*Solanum tuberosum*) dihaploid families translocations affecting pericentromeric regions were common and in two cases were due to genomic misassembly. In two populations, we found evidence for translocation affecting euchromatic arms. In cv. PI 310467, a nonreciprocal translocation between chromosomes (chr.) 7 and 8 resulted in a 5–3 copy number change affecting several Mb at the respective chromosome tips. In cv. “Alca Tarma,” the terminal arm of chr. 4 translocated to the tip of chr. 1. Using oligonucleotide-based fluorescent in situ hybridization painting probes (oligo-FISH), we tested and confirmed the predicted arrangement in PI 310467. In 192 natural accessions of *Arabidopsis thaliana*, dosage haplotypes tended to vary continuously and resulted in higher noise, while apparent LD between pericentromeric regions suggested the effect of repeats. This method, LD-CNV, should be useful in species where translocations are suspected because it tests linkage without the need for genotyping.

Keywords: structural variation; chromosome; translocation; DNA sequence; copy number variation; linkage disequilibrium

Introduction

Genomic structural variation (SV) is common within plant populations (Swanson-Wagner *et al.* 2010) and has a profound effect on the phenotype of individuals (Díaz *et al.* 2012; Maron *et al.* 2013; Bastiaanse *et al.* 2019). SV is manifested at multiple scales, from single genes to large chromosomal regions, such as variable heterochromatic blocks in the pericentromeres of potato (*Solanum tuberosum*) (Gong *et al.* 2012; Zhang *et al.* 2014; de Boer *et al.* 2015; Hardigan *et al.* 2016). Small scale SV can alter gene structure and copy number. Large-scale translocations and inversions can alter copy number, recombination, meiotic anaphase patterns, viability of gametes, gene structure, and evolutionary potential (Khush 1973; Rieseberg 2001). Translocations, which tend to be underdominant, i.e., to confer a heterozygous disadvantage (Rieseberg 2001), can persist in nature when they involve reciprocal exchanges of chromosome arms, resulting in copy-neutral, monocentric recombinant

chromosomes. Nonreciprocal translocations are often deleterious because they can involve duplication and deletion of segments of the recombined chromosomes. However, in polyploids their effect is lessened by genomic redundancy.

Detection of translocation is not simple and often requires multiple approaches. SV loci display linkage disequilibrium (LD) with linked regions (Hinds *et al.* 2006; Locke *et al.* 2006). Translocations were originally detected by observing unusual lethality and unexpected linkage in *Drosophila* (Bridges 1923) and in plants (Belling and Blakeslee 1924). Later, they were confirmed cytologically (Belling and Blakeslee 1924; Muller 1929; McClintock 1930; Burnham 1956). The development of fluorescent in situ hybridization (FISH) facilitated their identification because each chromosome in a spread could be identified by cytological markers or by chromosome painting (He *et al.* 2018). Detection by FISH analysis requires chromosome-specific probes (Cremer *et al.* 1988), which can be readily developed using an oligonucleotide-based methodology (Zhang *et al.* 2021).

Genome sequencing can be informative of copy number changes and novel junctions (Alkan et al. 2011). Dosage analysis of sequence reads is informative when unbalanced translocations result in copy number variation (CNV), and when balanced translocations generate copy number variable progeny. CNV can be inferred by count of genomic sequencing reads. Regions with high or low read counts have likely undergone duplication or deletion. Read count, however, does not provide information on the context and position of copy number variable regions. Specifically, a duplication could involve transposition or be in tandem. The software CNVmap exploits heterozygous states arising from duplications to detect and map CNV (Falque et al. 2020). However, SNP between duplicated loci is a function of duplication age and may not be present in newly arisen SV. Evidence of a junction between two different chromosomes can be obtained by detecting sequencing reads that directly or indirectly span the junction. In practical terms, however, identification and characterization of a translocation is not simple, particularly in the absence of prior evidence pointing to its location. DNA analysis is complicated by the presence of many spurious signals in mapped sequenced reads, especially when using short sequencing reads, and by the frequent presence of repetitive DNA at translocation junctions (Chen et al. 2010). Long read technology can be effective at addressing these problems (Sedlazeck et al. 2018), although it is typically more expensive than short-read sequencing.

Polyploid species are more likely to display large SV because their genome buffers the deleterious effects of copy number changes (Comai 2005). In polyploids, gametes carrying a large chromosomal deletion might still be viable because they carry a second normal copy of the affected chromosome. Cultivated potato is an autotetraploid, highly heterozygous organism that is clonally propagated, but capable of elevated outcrossing rates (Brown 1993). The potato genome displays a high density of nucleotide and structural polymorphism, suggesting high plasticity (Hardigan et al. 2016, 2017). From a cultivated tetraploid potato clone, sexual haploid progeny can be extracted by pollination with a haploid inducer (Figure 1A). These haploids inherit a diploid genome consisting of the maternal gametic contribution and are therefore called dihaploids. Their diploid genome simplifies genetic analysis. Further, they are fertile when crossed to many wild diploid accessions, enabling the bridging of germplasm and introduction of valuable traits into cultivated potato (Peloquin et al. 1989; Rokka 2009).

To investigate structural genomic variation, we developed a novel translocation detection approach and demonstrated its application in potato. Our method does not require genotyping or prior identification of polymorphism, such as SNPs or other common genetic markers. Instead, it leverages the reference genome sequence and dosage states inferred from sequencing the genome of related individuals, such as dihaploids of the same parent, in order to identify apparent LD (correlated dosage states) between structurally variant loci. The resolution of the identified SV scales with sequencing coverage, but even low coverage data can lead to interesting findings. Using several potato dihaploid populations, we demonstrate discovery of translocations in two families, as well as regions that could either be translocated and polymorphic, or assembly errors.

Methods

Potato populations

The potato families tested here were produced by crossing autotetraploid cultivated varieties of *S. tuberosum* to diploid *S. t. Group*

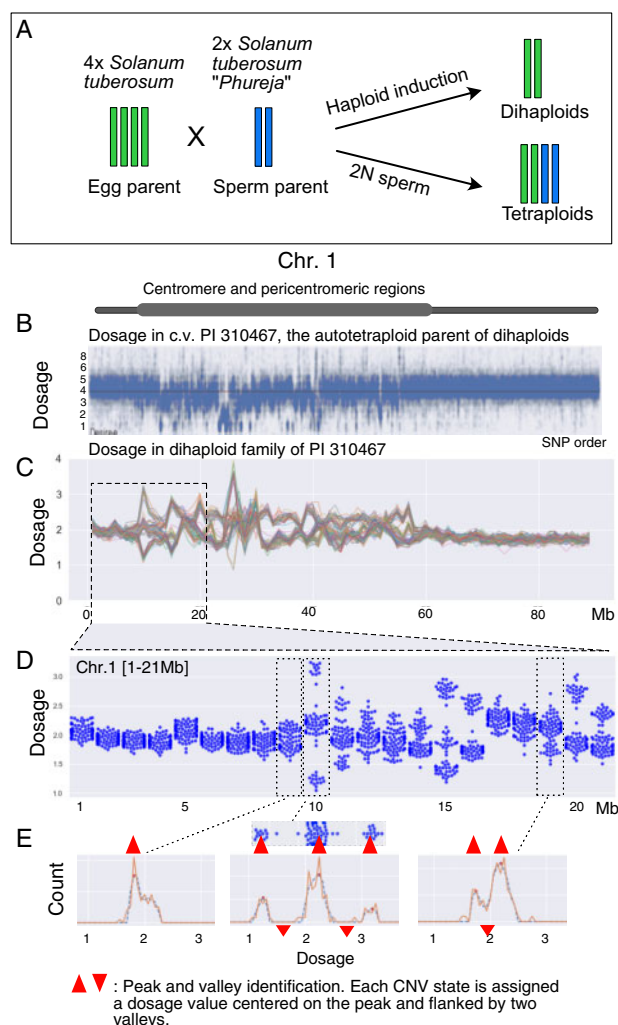


Figure 1 Dosage states of potato chr. 1 in 84 dihaploid siblings. (A) Haploid induction crosses generate progeny of different ploidy. Dihaploids inherit the genomic content of the egg gamete exclusively. Tetraploids are hybrids and receive contribution from both parental genomes. Triploid hybrids are also possible, but they are rare because of the unbalanced endosperm (genome ratio: 4 maternal: 1 paternal instead of 2 m: 1p). (B) Standardized coverage per SNP of chr. 1 for the tetraploid parent PI 310467. (C) Relative dosage states along chr. 1 were derived by sequence coverage binned on 1Mb intervals and standardized on the maternal dosage. 84 maternal dihaploid individuals produced from the haploid induction cross *S. t.* PI 310467 × *S. t. group phureja* IvP48 are overlapped revealing recurring dosage trends and polymorphism, mainly associated with heterochromatic regions along the pericentromere. (D) Swarm plots of dosage states in the first 21 Mb of chr. 1. (E) Cluster derivation by Peakutils. Peaks (upward red arrowheads) are identified by the algorithm, while valleys (downward red arrowheads) are defined as the mid-distance between peaks. Dosage values flanked by two valleys were assigned to the corresponding clusters.

phureja haploid inducers. All, with the exception of the BB dihaploids have been described (Amundson et al. 2020, 2021).

The BB population was derived by crossing tetraploid potato variety PI 310467 to haploid inducer IvP48. Similar to other potato haploid induction crosses, this cross results in high seed lethality with 5–15 live seeds per berry and ~0.5 dihaploid per berry. The rest of the seed are predominantly tetraploids, i.e., the result of 2N pollen fertilizing a 4X central cell and 2X egg. The dihaploids were identified by the lack of the dominant paternal seed spot

marker. Ploidy was confirmed by flow cytometry. The triploid progeny are rare because of lethality caused by unbalanced 4m:1p endosperm and were not used for this analysis (B.O., unpublished results).

DNA sequencing

Information on DNA sequencing for all samples except for the BB samples has been published (Amundson et al. 2020, 2021). Similarly, genomic DNA of the BB cohort was extracted from young leaves (Ghislain et al. 1999) and 750 ng of each genomic DNA sample was sheared to a mean size of 300 bp. Sequencing libraries were constructed using KAPA HyperPrep kit (KAPA Biosystems KK8504) using half-scale reactions, custom 8 bp index adapters, and four amplification cycles. Libraries were sequenced as SR100 on the Illumina HiSeq 4000 or NovaSeq 6000 platform at the University of California, Davis DNA Technologies Core, Vincent Coates Genome Sequencing Laboratory, or University of California San Francisco Center for Advanced Technologies. Libraries were demultiplexed using custom Python scripts available at (allprep-12.py; <https://github.com/Comai-Lab/allprep>, last accessed: 28 August 2021).

Dosage analysis

Single end reads were aligned to the DM1-3 4.04 reference *S. tuberosum* genome with BWA-mem (Li 2013) and only reads with mapping quality ≥ 10 were retained. Standardized coverage values were derived by taking the fraction of mapped reads that aligned to a bin of interest for that sample, normalizing it to the corresponding fraction from the same interval in the parent of the population, if available, or the mean of the population, and multiplying the resulting value by 2 to indicate the expected diploid state, as previously described (Henry et al. 2015). To mitigate mapping bias due to read type and length, paired-end reads from LOP-868 were hard trimmed to 50 nt and only forward mates were used for analysis. Read depth is calculated for fixed-size nonoverlapping windows, set to 250 kb or 1 Mb in this study. Whole chromosome aneuploidies are identified as previously described (Amundson et al. 2020) and withheld from analysis prior to FET probability matrix construction.

Apparent linkage disequilibrium

Linkage disequilibrium is defined as the nonrandom association of alleles at different loci in a given population. Our analysis correlates dosage state, i.e., the cumulative dosage value of a given sequence in the genome and is related to correlation between quantitative traits. The outcome of the analysis resembles that of LD analysis and for this reason we define it as apparent LD. The analysis and figure generating software is available at https://github.com/lcomai/cnv_mapping, last accessed: 28 August 2021. Fisher's exact test (FET) was carried out between pairs of dosage states derived from 1 Mb bins to assess apparent LD between bins (Figure 2). For example, assume that both Bin1 and Bin100 have three dosage states: copy number 1, 2, and 3. To test if Bin1 correlated to Bin100, the following four dihaploid sets were compared in a 2x2 contingency table: (observed in Bin1-CN1)/(observed not in Bin1-CN1), (expected in Bin1-CN1)/(expected not in Bin1-CN1), where the expectation was derived from the assumption of complete independence. Self-comparison and reciprocal comparisons were removed, and the remaining comparisons were controlled at FDR = 0.05. Chromosomal bins that were engaged in at least one statistically significant correlation after these corrections were deemed in LD. The matrix analysis to compare different genome

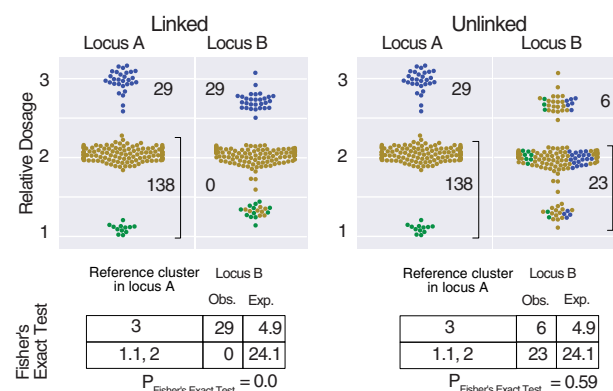


Figure 2 Testing random association between dosage states at different loci. Each dot represents the dosage state at the specified locus for one individual as shown in Figure 1D, and it is colored to mark its cluster occupancy in Locus A. Hypothetical loci A and B are illustrated for linked (left) and unlinked (right) states. The independence test compares dosage states at the two loci. Here, the top cluster in both loci is contrasted to the aggregate of the other two, producing a 2 × 2 contingency table that is subjected to Fisher exact test (FET). The test is repeated for each dosage cluster of both loci. Obs., observed; Exp., expected according to random association.

assemblies (Supplemental Fig. 4) was carried out using the method of Cabanette and Klopp (2018).

Alternative to FET-based analysis of copy number dosage states, one can derive and plot a heatmap of correlation coefficients such as Pearson's R as demonstrated in the Github notebooks (see above) and in Supplementary Figures S3 and S4. Gating the resulting matrix to display only the strongly correlated values helps in the analysis (Supplementary Figures S3 and S4). Measuring correlation with a utility such as `DataFrame.corr()` (Pandas) or `scipy.stats.pearsonr()` avoids the bin-by-bin FET, which becomes computationally intensive with increasing numbers of CNV bins. Nonetheless, we found the approach based on FET manageable and clearer in its output. Categorization in peaks (dosage clusters) helped the subsequent analysis of candidate regions.

Chromosome spread preparation

Potatoes were planted from culture tubes into a greenhouse planting mix and allowed to grow for 3–5 days. The root tips were collected and treated with iced water for 24 h. Then, the root tips were fixed directly into Carnoy's solution (3 ethanol: 1 acetic acid) and stored at −20°C until use. The root tips were squashed on the microscope slide with the same fixative solution after digestion with 2% cellulose (Sigma, USA) and 1% pectolyase (Sigma, USA) at 37°C for 2 h.

Oligo-FISH painting

We used oligonucleotide-based FISH probes (Oligo-FISH) (Han et al. 2015) to specifically paint the chr. 7 and chr. 8 of potato. Probe labeling and FISH were performed following published protocols (Braz et al. 2018, 2020). Biotin and digoxigenin-labeled probes were detected by anti-biotin fluorescein (Vector Laboratories, Burlingame, CA, USA) and anti-digoxigenin rhodamine (Roche Diagnostics, Indianapolis, IN, USA), respectively. Chromosomes were counterstained with 4,6-diamidino-2-phenylindole in VectaShield antifade solution (Vector Laboratories). The chromosome spreads were imaged using a QImaging Retiga EXi Fast 1394 CCD camera (Teledyne Photometrics, Tucson, AZ, USA) attached to an Olympus BX51 epifluorescence microscope.

Images were processed with META IMAGING SERIES 7.5 software and their final contrast was processed using Adobe PHOTOSHOP software (Adobe, San Jose, CA, USA).

Results

Identification and apparent LD analysis of CNV loci in a dihaploid population of potato

We generated a family of diploids and tetraploids by crossing tetraploid potato variety PI 310467 to male haploid inducer IvP48, producing a population called BB (Figure 1A; Supplementary Table S1; see Methods). PI 310467 is listed as Desiree in the GRIN germplasm collection but, although related, it differs from Desiree in its SNP profile (K.A., unpublished results). In the BB progeny, we selected and analyzed 84 2X haploids (dihaploids) by flow cytometry and low-pass genome sequencing. For comparison, we also analyzed 78 4X hybrid siblings. Mapped reads were counted in nonoverlapping, 1 Mb genomic intervals along the reference genome, and these counts were standardized to a mean of 2 using the corresponding maternal variety counts. When these standardized dosage values from all individuals were overplotted, CNV polymorphisms were readily visible, as displayed for chr.1 (Figure 1, Supplementary Figure S1). Invariant regions form unimodal distributions, while polymorphic regions form bimodal or multimodal distributions (Figure 1, B and C). For each of these regions, individuals could be clustered based on their corresponding read counts (Figure 1D, Supplementary Figures S1 and S2). We used the Python utility Peakutils (<https://peakutils.readthedocs.io/>, last accessed: 28 August 2021) to identify these clusters and assign each individual to a genomic dosage state, which can be viewed as a stepwise, quantitative phenotype resulting from additive alleles.

Consistent with previous reports (de Boer et al. 2015; Hardigan et al. 2016; Amundson et al. 2020), dosage variation was common in pericentromeric regions (Figure 1B, Supplementary Figure S1). To further characterize the associated structural changes, we asked whether dosage variation at different loci was independent. Linked loci, such as those in the same pericentromeric regions, would be expected to co-vary, i.e., to exhibit an apparent LD (henceforth LD for brevity). Dosage states of unlinked loci

should be independent. Unexpected LD suggests either epistasis or novel linkage, such as resulting from a translocation or genome assembly error. Therefore, for each SV locus, we tested the null hypothesis that its dosage states were independent of dosage states at another SV locus (Figure 2).

After correction for multiple testing (Benjamini and Hochberg 1995), associations between bins were plotted on a genomic matrix that displays significant probabilities of FET, thus called FET probability matrix (Figure 3). The potato genome was partitioned in 731 1 Mb bins, of which 246 (33% of genome) displayed distinct dosage states. Most, 236, were in LD with at least another bin (corrected $P < 0.05$). This is not surprising, since LD is expected for physically linked loci. However, 65 bins were in LD with bins on other chromosomes, indicating that 26.7% of copy variable sequences display unexpected linkage. Loci in LD detected with $FDR = 0.05$ are marked in blue and black in the matrix (Figure 3).

The FET probability matrix displays binary comparisons for which the probability of FET is significant, thus highlighting bins that display correlated dosage states. Loci displaying linkage, such as adjacent ones, form a diagonal line. Blocks that correspond to low recombination intervals are positioned around the centromeres (Figure 3). In addition to intrachromosomal interactions, cases of strong interchromosomal ones are evident, suggesting physical linkage. For example, a region of 1 Mb in chr. 1 displays an association to the entire pericentromeric region of chr. 8. Similar interactions are visible between chr. 4 and chr. 7 and between chr. 11 and chr. 12. These could be translocated heterochromatic blocks, which have been demonstrated in potato (Zhang et al. 2014; de Boer et al. 2015), or genomic assembly errors (see below).

An alternative approach to determining Fisher's exact probabilities, is correlation based on Pearson's R, which while less sensitive and less interpretable could often provide equivalent information because translocation signals are strong (Methods; Supplementary Figures S3 and S5).

Translocation of a euchromatic 6 Mb region between chr. 7 and chr. 8

We noted a strong signal between the 6 Mb euchromatic right arm of chr. 7 and the 5 Mb right arm of chr. 8 (Figure 3). In the

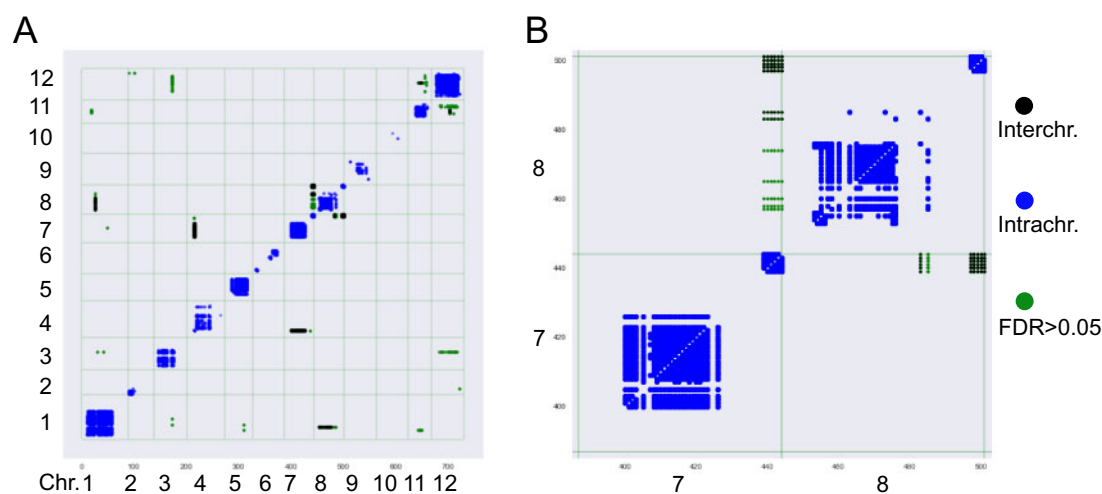


Figure 3 FET probability matrix illustrates correlated genomic dosage states. Each dot in the matrix indicates that in the genomic bin x bin comparison a significant probability value led to rejection of the null hypothesis of independence. Most dots are overplotted because each dosage cluster in each bin is subjected to FET (Figure 2) and, typically, several tests are significant for each compared bin pair. (A) Twelve-chromosome matrix displays significant interactions across the genome. (B) Zoomed view on two chrs.: 7 and 8. Note that only CNV loci can be tested. To illustrate that most interactions are statistically significant, green dots illustrate the relatively few interactions that had $FET P < 0.05$, but did not pass the multiple test correction.

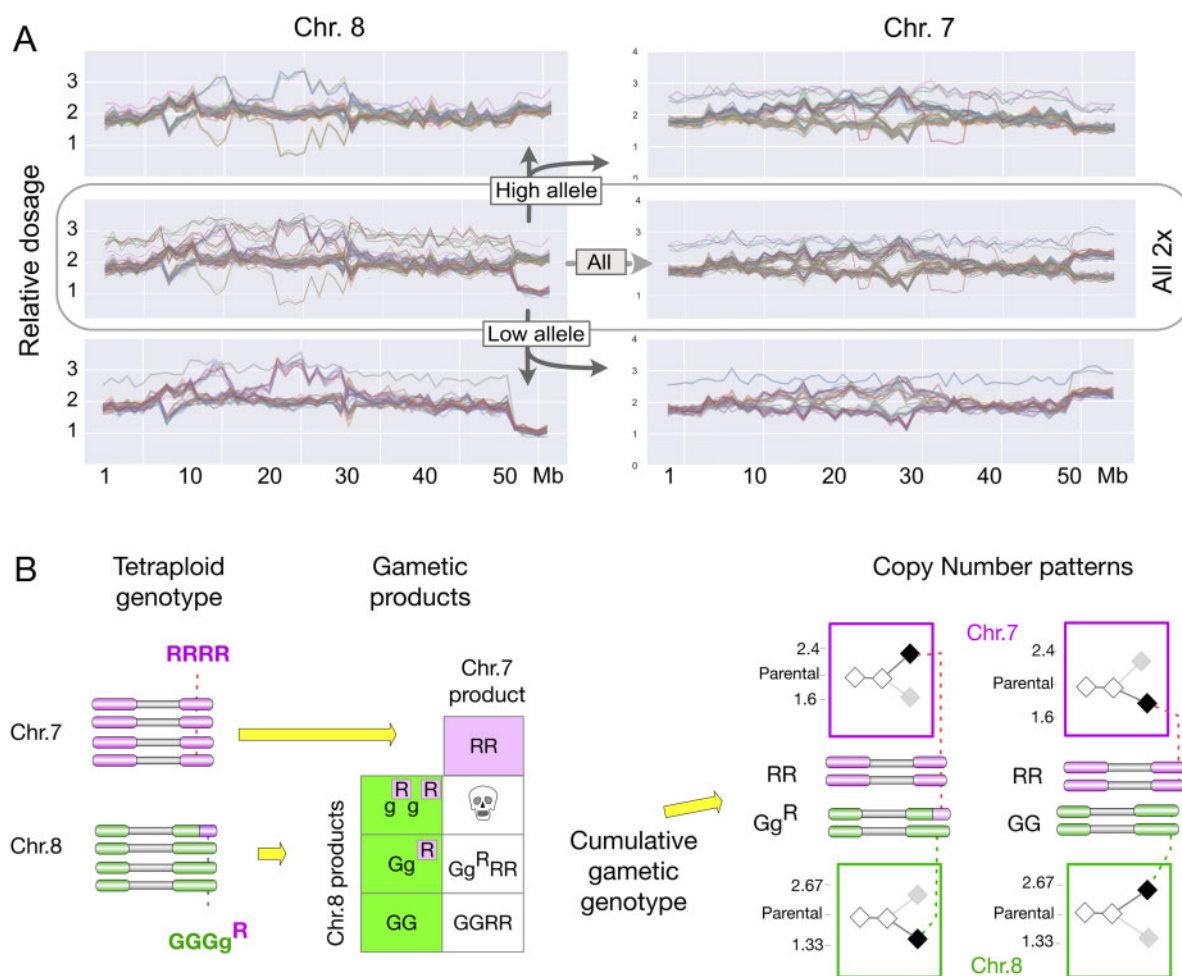


Figure 4 Analysis of CNV states in chrs. 7 and 8 of potato clone PI 310467. (A) Perfect correlation of CNV state between chr. 8 and chr. 7 in the population of dihaploids. Chr. 8 and chr. 7 relative dosages are plotted using all 2x individuals (center row) or filtered sets (top and bottom row), relative to the parental genome of PI 310467. High allele: Selection of individuals where chr. 8 distal right arm copy number > 1.5. Trisomic individuals (high dosage tracks) are removed. Low allele: Selection of individuals where chr. 8 distal right arm copy number < 1.5. The dosage of each individual was calculated by dividing standardized reads per bin by the corresponding count for PI 310467. The dosage shown is therefore not absolute, but relative to that of the maternal dosage. (B) Model of Tr.8-7 in autotetraploid PI 310467, meiotic transmission pattern of chr. 7 and chr. 8 into gametes. On the right, the genotype of dihaploids is displayed together with the resulting copy number pattern for the terminal, right arm region of chr. 7 and chr. 8. The values are relative. For example, the terminal region of 7 is present in 3 copies in the RRGg dihaploid and 5 in the tetraploid parent. The relative dosage = $3/5 \times 4 = 2.4$.

dosage plots (Figure 4), both chromosomes display distinct copy number polymorphism with absolute correlation between dosage states: higher dosage of chr. 7 was always associated with lower dosage of chr. 8. Translocation of the terminal segment of chr. 7 and chr. 8 with concurrent loss of the 5 Mb terminal region of 8 (Figure 4B) could explain the observations, as illustrated in the inheritance and dosage model (Figure 4B, Supplementary Figure S2). The dosage plots in Figure 4 were constructed by standardizing read counts to those of parent variety PI 310467, i.e., they are relative to the control. To test the above hypothesis, we plotted raw genomic dosage of putative single copy regions corresponding to SNP (Figure 5C, blue track) (see Methods). We also used SNP count ratios to provide an independent measure of copy number. Figure 5, B and C demonstrates that there are 5 copies for chr. 7 and three copies of chr. 8 in the involved regions. To provide conclusive validation for this translocation, we conducted FISH using two oligonucleotide-based chromosome painting probes specific for chr. 7 and chr. 8, respectively (Figure 5). We observed four normal copies of chr. 7 and three normal copies of chr. 8. One additional copy of chr. 8 carried a region of chr. 7 (Figure 5D). We

concluded that the PI 310467 clone has an unbalanced translocation between chr. 7 and chr. 8. The presence of this translocation was verified by obtaining and sequencing a second, independent sample of PI 310467 from the USDA Potato Genebank (data not shown).

In summary, analysis of covarying dosage states identified regions in LD (i.e., genetically linked) including the unbalanced translocation between chr. 7 and chr. 8.

Analysis of tetraploids in the BB family of potato

To probe the robustness of the method, we analyzed the tetraploid, hybrid BB progeny. These have genomes formed by the 2x egg of PI 310467 and accidental 2N (=2x) sperms from IvP48. Their dosage states derived from the additional action of alleles, two maternal and two paternal, resulting in a more complex outcome. By FET, we detected comparable numbers of intrachromosomal interaction (474 in the 2x group vs 525 in the 4x group), but fewer interchromosomal interactions (764 in the 2x group vs 169 in the 4x group), as indicated by the FET probability matrix. For example, the matrix no longer displayed the chr. 11 and chr. 12

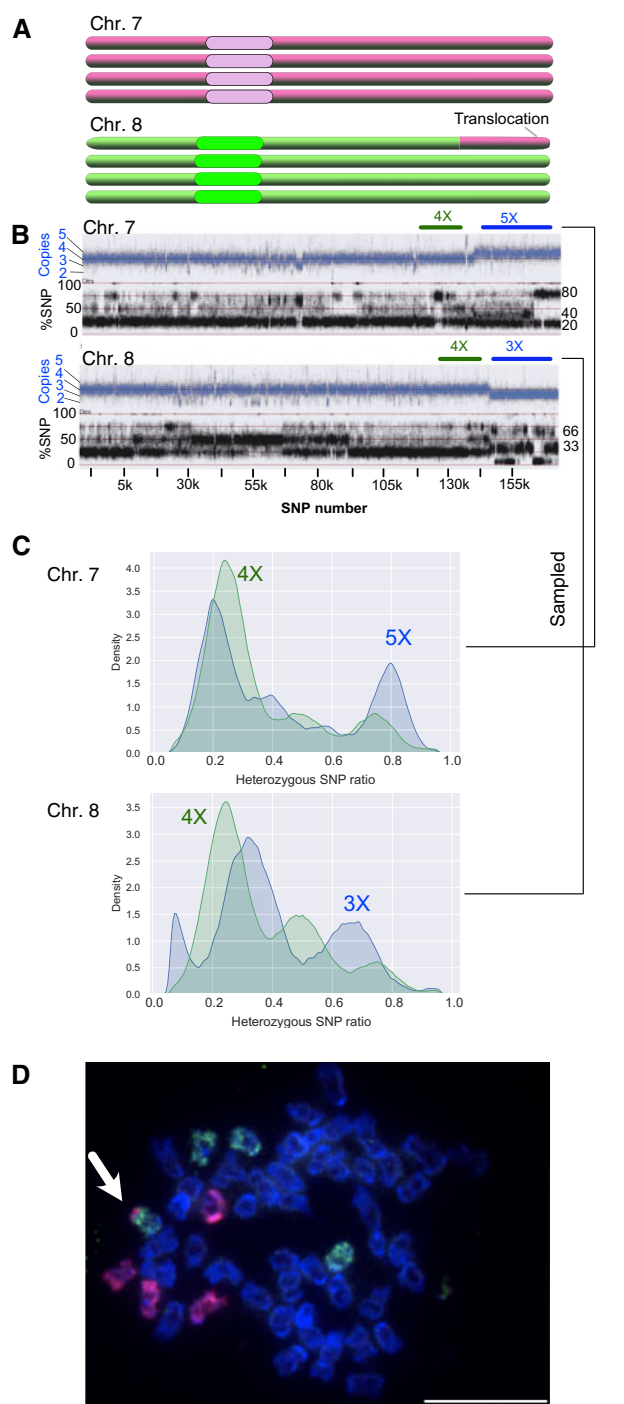


Figure 5 Molecular and cytological evidence for translocation 8-7. The genomic status of chr. 7 and chr. 8 is demonstrated by read coverage (copy number), SNP ratio, and oligonucleotide-based fluorescent in situ hybridization painting probes (oligo-FISH). (A) Karyotype of chr. 7 and chr. 8 in PI 310467 according to the following evidence. (B) The blue tracks display the DNA copies at each SNP locus along the length of the shown chromosome. The multiple black tracks illustrate the allele specific read depth ratio of SNP loci. Four DNA copies yield heterozygous SNP ratios of 25%–50%–75%. For example, the simplex genotype *Aaaa* corresponds to 25%. Three DNA copies yield heterozygous SNP ratios of 33%–66%, five copies of 20%–40%–60%–80%. (C) SNP ratio analysis. For chr. 7 it indicates 4 and 5 copies. For chrs. 8, it indicates 4 and 3 copies. (D) Oligo-FISH painting of a mitotic metaphase cell prepared from PI 310467. The arrow points to the translocation chromosome. Red: chr. 7, Green: chr. 8.

putative translocation visible in the 2x analysis. Tr.8-7, however, was evident (Supplementary Figure S3). We concluded that CNV resulting from duplication or deletion are more difficult to identify in a tetraploid, but events such as Tr8-7 remain distinct.

Analysis of additional dihaploid families of potato

We asked if the method could be useful in eight comparable, but unrelated dihaploid families generated through haploid induction crosses (Supplementary Table S1). Seven of these dihaploid families (Amundson et al. 2021) revealed the presence of translocated or misplaced heterochromatic and pericentromeric blocks, but no arm translocations (Supplementary Figure S4). In the last dihaploid population, called LOP, an interesting arm translocation was evident that affected their tetraploid seed parent, *S. tuberosum andigena* variety Alca Tarma (Velásquez et al. 2007; Amundson et al. 2020). Interestingly, a segment representing several Mb of euchromatin on the short arm of chr. 4 displayed a distinct dosage polymorphism that was in strong LD with the end of chr. 1 (Figure 6, A and B). Similarly, to the T8-7 translocation identified in the BB population, this suggested a translocation of the euchromatic segment to the end of chr. 1, which correspondingly lost a short terminal segment. Dosage analysis in Alca Tarma demonstrated neutrality, i.e., four copies of the short arm of chr. 4. Together with the inheritance pattern observed, this suggested that the translocated chromosome T1-4 and the terminally truncated version of chr. 1 were present as a single copy in Alca Tarma (Figure 6C). By pooling reads from dihaploids sharing chr. 4 dosage states and rerunning the analysis with a smaller bin size, we narrowed the junction of the translocated region to a 10Kb interval between chr04.repeat.3471 and chr04.repeat.3473 (Supplementary Figure S6, A and B). This region, however, consists of N-nucleotides that could not be sequenced during the potato genome project (Supplementary Figure S6C). A model that explains the inheritance pattern and the dosage profiles of the dihaploid progeny is provided in Figure 6C.

Recurring translocations detected by LD-CNV result from genome misassembly

Comparison of LD in the nine dihaploid families highlighted frequent LD of certain heterochromatic regions across the populations (Supplementary Figures S3–S5). For example, a region in the left arm of chr. 1 was linked to the whole pericentromeric region of chr. 8. Similar translocations were evident for chrs. 4–7 and chrs. 11 and 12. Release of the *S. tuberosum* genome v.6.1 provided a test for the hypothesis of misassembly. A similarity dot matrix between these genome v.6.1 and genome v.4.04 (used here) displayed discordance for the translocated 1–8 and 4–7 regions (Supplementary Figure S4) demonstrating that the detected misassembly was fixed independently in the update genome release.

Detection of related CNV patterns in an *A. thaliana* population

We asked if CNV in LD could be identified in an unstructured natural population of *A. thaliana* (1001 Genomes Consortium 2016). We subjected 192 randomly chosen *A. thaliana* accessions that differ in geographic origin to the CNV-LD analysis (Supplementary Table S2 and Figure S7). Overplotting CNV states displayed mostly continuous, substoichiometric variation without the frequent and distinct dosage states found in potato. We detected significant LD between chromosomes, but these affected predominantly the pericentromeric regions and were likely

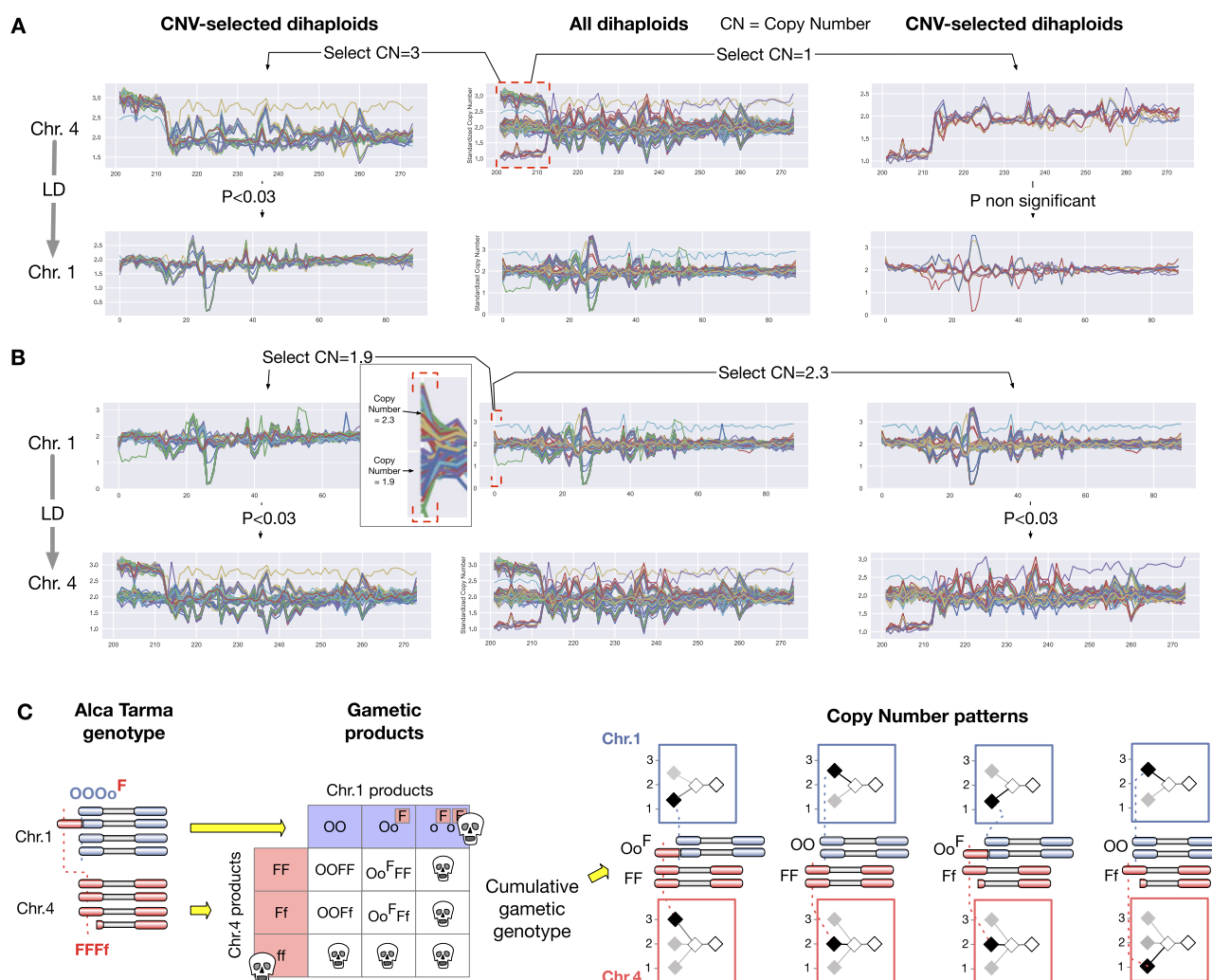


Figure 6 Apparent Linkage Disequilibrium (LD) consistent with translocation of a chr. 4 segment to chr. 1 in the LOP population. (A, B) Each line represents the dosage profile of an individual relative to the parental control. (A). Individuals that share either high or low cluster **Copy Number** dosage state (CN) on the proximal arm of chr. 4 were selected (left, right). The chr. 1 profiles displayed by the selected individuals confirm the LD with CN = 3, the duplicated state. (B). Individuals that share either high or low CN on the left arm of chr. 1 are selected. In this region, a single 1 Mb bin is polymorphic, corresponding to segregation of a terminal deletion. Because the dosage measured through a single bin is likely biased, we report the observed dosage value without calculating its expected standardized value. The corresponding profiles for chr. 4 confirm LD. The left arm tip of chr. 1 is enlarged to display the two dosage states. (C). Translocation model explaining two dosage states in chr.1 and three in chr. 4. A hypothetical genotype of Alca Tarma postulates that the short arm of 4 (F) translocated to the terminal 1 region (O) producing the haplotype of^F. Gametes carrying homozygous deletion or duplication are likely to be dead or impaired. The resulting dihaploids display CN patterns consistent with those observed. See Figure 2 for an explanation of the different CN patterns.

determined by repetitive elements because each affected loci on multiple chromosomes.

Discussion

We developed a method that we named LD-CNV, aimed at identifying correlated dosage states, i.e., apparent LD, between copy number variable loci in the absence of genotyping information. It uses mapped reads from individuals in segregating populations. The reads are binned in genomic intervals, whose size depends on sequence coverage, but typically ranges from 0.1 to 1 Mb. Read counts are normalized to the value of a reference, for example the parent, scaling the mean to overall ploidy. We identify dosage polymorphic bins by detecting variation and clustered dosage values among the tested individuals (Figures 1 and 2). We then test for association between dosage states of different variable

bins using FET (Figures 2 and 3). An alternative approach is to simply measure correlation using the input dosage values or the peak-centered processed values. Regardless of the statistical tool chosen, the method is simple, aiming to identify discrepancies from available genomic models. A related method, CNVmap, differs from ours by leveraging heterozygous SNP as a proxy for duplicated loci and mapping them in segregating populations (Falque et al. 2020). Therefore, it detects small-scale duplications, but only when their sequence is divergent. The independence of LD-CNV from genotyping makes it applicable in a fully homozygous system or in any system lacking genotyping information.

We provide the software as annotated Jupyter Notebooks. These are Python-based data analysis tools that facilitate pipeline journaling, annotation, and modification for flexible analysis (Parker 2018). In addition, plotting and display of data series from the data frames is accessible at any step of the execution

(Waskom 2021; Hunter 2007). Software dependencies can be easily installed and managed via the Anaconda package manager (www.anaconda.com). Complexity and computing power of these tools are scalable, but for most analyses, such as those described here, consumer hardware is sufficient. These features should make the tool both readily usable and modifiable for ad hoc applications.

We originally developed this method to identify epistatic relationships between SV loci in potato, assuming that they would result in LD. We found, however, that the strongest signals were the result of actual, but unexpected linkage. In our potato dihaploid progeny populations, LD-CNV analysis identified two heterochromatic blocks that were misassembled in the original reference genome and corrected in the latest one. Providing further validation, in two tetraploid seed mother accessions out of the nine tested in this analysis, we found convincing evidence of chromosomal translocations involving euchromatic arms. We detected LD between the right arms of chr. 7 and chr. 8 in the BB population (seed parent var. PI 310467), and the left arms of chr. 1 and chr. 4 in the LOP population (seed parent var. Alca Tarma). This information was combined with dosage analysis to generate translocation models. In the BB seed parent PI 310467, 5 and 3 copies, respectively, of the terminal regions of chr. 7 and chr. 8, suggested that the terminal right arm of chr. 7 translocated to chr. 8 with concurrent loss of the terminal region of chr. 8 (Figure 6, A and B). The BB population included a tetraploid set, which enabled testing the method in a genomic scenario more challenging than diploidy. In the tetraploids, haplotypes are contributed by the tetraploid seed mother and the diploid pollen parent. Nonetheless, detection of the 8-7 translocation was still possible. The predicted rearrangement was confirmed by oligo-FISH analysis (Han et al. 2015; Braz et al. 2018) in the BB population parent. A second translocation was detected in Alca Tarma, the seed parent of LOP. In this clone, chr. 1 and chr. 4 appear copy-number neutral. This information, together with the profiles of the dihaploid progeny, supports a model in which the terminal region of chr. 4 translocated near to the end of chr. 1 (Figure 6C). Furthermore, the discovery of 2 translocations among 9 tetraploid varieties suggests that chromosomal rearrangements are not uncommon in cultivated potato germplasm.

We wondered whether this new tool could inform us about SV in natural populations as well, and tested a set of 192 *A. thaliana* accessions. Comparison of the results obtained with pedigree families to this natural population shows the effect of an important feature: sibs in segregating populations form distinct dosage clusters when inheriting CNV haplotypes that are heterozygous in the parent. Large haplotypes are conserved in most individuals because recombination is rare. Large genomic bins in natural populations, however, vary continuously because of the independent behavior of multiple DNA regions and elements, hindering formation of distinct clusters. The CNV observed in the *A. thaliana* population corresponds most likely to repeated regions, while those observed in the potato families includes single-copy sequences. The analysis should work in natural populations if the bin unit examined is very small, such as gene-size. On that scale, presence-absence of a DNA segment should cluster nicely on two dosage states. That level of granularity would require more computing power than provided by consumer hardware. If one is interested in new variation, however, LD would likely be confined to only a few bins and easily analyzed.

In conclusion, the method enables exploration of sequence datasets from segregating families in species with a reference genome to identify relationships among SV loci. Mb scale variation

can be detected with as little as 0.2X coverage since one can expand bin size to increase the number of reads to a threshold of statistical confidence. Low-pass, whole-genome sequencing is a convenient approach toward genotyping that is becoming rapidly more affordable (DePristo et al. 2011). Once sequence reads have been mapped, the method is easily implemented without the need to identify a set of informative SNP. Knowledge of translocations, whether real or caused by genome mis-assembly, is critical to the use of a genetic population for genetic studies and for breeding because of the dramatic effect they can have on outcome and interpretation.

Data availability

Sequence data have been obtained from or deposited in the National Center for Biotechnology Information Sequence Read Archive with the following Bio-Project identifier: for the LOP dihaploids, PRJNA408137; BB, PRJNA750855; sequence reads data for the dihaploids of WA077, LR00014, LR00022, LR00026, 93003, C91640, C93154, are at NCBI Bioproject PRJNA699631. The raw and standardized genomic dosage values for the BB dihaploids are available at DryadData.com <https://doi.org/10.25338/B88D2V>. The Arabidopsis sequence listed in Supplementary Table S2 was obtained from PRJNA273563 “1001 Genomes: A Catalog of *Arabidopsis thaliana* Genetic Variation.” Supplementary material is available at figshare: <https://doi.org/10.25386/genetics.14810961>.

Acknowledgments

The authors thank Meric Lieberman for bioinformatic assistance. L.C. designed experiments with input from K.R.A. and I.M.H. B.O. developed the BB population. L.C., K.R.A., and B.O. performed genomic experiments. X.Z., G.T.B., and J.J. performed cytological experiments. L.C. analyzed data and wrote the manuscript with input from all authors.

Funding

This work was supported by: the National Science Foundation Plant Genome Integrative Organismal Systems (IOS) Grants 1444612 (Rapid and Targeted Introgression of Traits via Genome Elimination) and 1956429 (RESEARCH-PGR: Variants and Recombinants without Meiosis) to L.C. and I.M.H; a grant from the Innovative Genome Institute at UC Berkeley to LC, and the United States-Israel Binational Agricultural Research and Development Funds IS-5038-17C and IS-5317-20C to J.J.

Conflicts of interest

The authors declare that there is no conflict of interest.

Literature cited

- 1001 Genomes Consortium. 2016. 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell*. 166: 481–491. doi:10.1016/j.cell.2016.05.063.
- Alkan C, Coe BP, Eichler EE. 2011. Genome structural variation discovery and genotyping. *Nat Rev Genet*. 12:363–376.
- Amundson KR, Ordoñez B, Santayana M, Nganga ML, Henry IM, et al. 2021. Rare instances of haploid inducer DNA in potato dihaploids and ploidy-dependent genome instability. *Plant Cell*. 33: 2149–2163. doi:10.1093/plcell/koab100.

- Amundson KR, Ordoñez B, Santayana M, Tan EH, Henry IM, et al. 2020. Genomic outcomes of haploid induction crosses in potato (*Solanum tuberosum* L.). *Genetics*. 214:369–380. doi:10.1534/genetics.119.302843.
- Bastiaanse H, Zinkgraf M, Canning C, Tsai H, Lieberman M, et al. 2019. A comprehensive genomic scan reveals gene dosage balance impacts on quantitative traits in *Populus* trees. *Proc Natl Acad Sci USA*. 116:13690–13699. doi:10.1073/pnas.1903229116.
- Belling J, Blakeslee AF. 1924. The configurations and sizes of the chromosomes in the trivalents of 25-chromosome daturas. *Proc Natl Acad Sci USA*. 10:116–120. doi:10.1073/pnas.10.3.116.
- Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B Stat Methodol*. 57:289–300.
- Braz GT, He L, Zhao H, Zhang T, Semrau K, et al. 2018. Comparative Oligo-FISH mapping: an efficient and powerful methodology to reveal karyotypic and chromosomal evolution. *Genetics*. 208:513–523. doi:10.1534/genetics.117.300344.
- Braz GT, Yu F, do Vale Martins L, Jiang J. 2020. Fluorescent in situ hybridization using oligonucleotide-based probes. *Methods Mol Biol*. 2148:71–83. doi:10.1007/978-1-0716-0623-0_4.
- Bridges CB. 1923. The translocation of a section of chromosome II upon chromosome III in *Drosophila*. *Anat Rec*. 24:426–427.
- Brown CR. 1993. Outcrossing rate in cultivated autotetraploid potato. *Am Potato J*. 70:725–734. doi:10.1007/BF02848678.
- Burnham CR. 1956. Chromosomal interchanges in plants. *Bot Rev*. 22:419–552. doi:10.1007/BF02872484.
- Cabanettes F, Klopp C. 2018. D-GENIES: dot plot large genomes in an interactive, efficient and simple way. *PeerJ*. 6:e4958. doi:10.7717/peerj.4958.
- Chen W, Ullmann R, Langnick C, Menzel C, Wotschovsky Z, et al. 2010. Breakpoint analysis of balanced chromosome rearrangements by next-generation paired-end sequencing. *Eur J Hum Genet*. 18:539–543. doi:10.1038/ejhg.2009.211.
- Comai L. 2005. The advantages and disadvantages of being polyploid. *Nat Rev Genet*. 6:836–846. doi:10.1038/nrg1711.
- Cremer T, Lichter P, Borden J, Ward DC, Manuelidis L. 1988. Detection of chromosome aberrations in metaphase and interphase tumor cells by in situ hybridization using chromosome-specific library probes. *Hum Genet*. 80:235–246. doi:10.1007/bf01790091.
- de Boer JM, Datema E, Tang X, Borm TJA, Bakker EH, et al. 2015. Homologues of potato chromosome 5 show variable collinearity in the euchromatin, but dramatic absence of sequence similarity in the pericentromeric heterochromatin. *BMC Genomics*. 16:374. doi:10.1186/s12864-015-1578-1.
- DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*. 43:491–498.
- Díaz A, Zikhali M, Turner AS, Isaac P, Laurie DA. 2012. Copy number variation affecting the Photoperiod-B1 and Vernalization-A1 genes is associated with altered flowering time in wheat (*Triticum aestivum*). *PLoS One*. 7:e33234. doi:10.1371/journal.pone.0033234.
- Falque M, Jebreen K, Paux E, Knaak C, Mezrouk S, et al. 2020. CNVmap: a method and software to detect and map copy number variants from segregation data. *Genetics*. 214:561–576. doi:10.1534/genetics.119.302881.
- Ghislain M, Zhang DP, Herrera MR. 1999. *Molecular Biology Laboratory Protocols Plant Genotyping: Training Manual*. Lima, Peru: International Potato Center.
- Gong Z, Wu Y, Koblízková A, Torres GA, Wang K, et al. 2012. Repeatless and repeat-based centromeres in potato: implications for centromere evolution. *Plant Cell*. 24:3559–3574. doi:10.1105/tpc.112.100511.
- Han Y, Zhang T, Thammapichai P, Weng Y, Jiang J. 2015. Chromosome-specific painting in cucumis species using bulked oligonucleotides. *Genetics*. 200:771–779. doi:10.1534/genetics.115.177642.
- Hardigan MA, Crisovan E, Hamilton JP, Kim J, Laimbeer P, et al. 2016. Genome reduction uncovers a large dispensable genome and adaptive role for copy number variation in asexually propagated *Solanum tuberosum*. *Plant Cell*. 28:388–405.
- Hardigan MA, Laimbeer FPE, Newton L, Crisovan E, Hamilton JP, et al. 2017. Genome diversity of tuber-bearing *Solanum* uncovers complex evolutionary history and targets of domestication in the cultivated potato. *Proc Natl Acad Sci USA*. 114:E9999–E10008. doi:10.1073/pnas.1714380114.
- He L, Braz GT, Torres GA, Jiang J. 2018. Chromosome painting in meiosis reveals pairing of specific chromosomes in polyploid *Solanum* species. *Chromosoma*. 127:505–513. doi:10.1007/s00412-018-0682-9.
- Henry IM, Zinkgraf MS, Groover AT, Comai L. 2015. A system for dosage-based functional genomics in poplar. *Plant Cell*. 27:2370–2383. doi:10.1105/tpc.15.00349.
- Hinds DA, Klok AP, Jen M, Chen X, Frazer KA. 2006. Common deletions and SNPs are in linkage disequilibrium in the human genome. *Nat Genet*. 38:82–85. doi:10.1038/ng1695.
- Hunter JD. 2007. Matplotlib: a 2D graphics environment. *Comput Sci Eng*. 9:90–95. doi:10.1109/MCSE.2007.55.
- Khush GS. 1973. *Cytogenetics of Aneuploids*. New York, NY: Academic Press.
- Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]*.
- Locke DP, Sharp AJ, McCarroll SA, McGrath SD, Newman TL, et al. 2006. Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. *Am J Hum Genet*. 79:275–290. doi:10.1086/505653.
- Maron LG, Guimarães CT, Kirst M, Albert PS, Birchler JA, et al. 2013. Aluminum tolerance in maize is associated with higher MATE1 gene copy number. *Proc Natl Acad Sci USA*. 110:5241–5246. doi:10.1073/pnas.1220766110.
- McClintock B. 1930. A cytological demonstration of the location of an interchange between two non-homologous chromosomes of *Zea mays*. *Proc Natl Acad Sci USA*. 16:791–796.
- Muller HJ. 1929. The first cytological demonstration of a translocation in *Drosophila*. *Am Nat*. 63:481–486. doi:10.1086/280282.
- Peloquin SJ, Jansky SH, Yerk GL. 1989. Potato cytogenetics and germplasm utilization. *Am Potato J*. 66:629–638. doi:10.1007/BF02853983.
- Perkel JM. 2018. Why Jupiter is data scientists' computational notebook of choice. *Nature*. 563:145–146. doi:10.1038/d41586-018-07196-1.
- Rieseberg LH. 2001. Chromosomal rearrangements and speciation. *Trends Ecol Evol*. 16:351–358. doi:10.1016/s0169-5347(01)02187-5.
- Rokka V-M. 2009. Potato Haploids and Breeding. In: A Touraev, BP Forster, SM, editors. *Jain Advances in Haploid Production in Higher Plants*. Dordrecht, Netherlands: Springer. p. 199–208.
- Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, et al. 2018. Accurate detection of complex structural variations using single-molecule sequencing. *Nat Methods*. 15:461–468. doi:10.1038/s41592-018-0001-7.
- Swanson-Wagner RA, Eichten SR, Kumari S, Tiffin P, Stein JC, et al. 2010. Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. *Genome Res*. 20:1689–1699.

- Velásquez AC, Mihovilovich E, Bonierbale M. 2007. Genetic characterization and mapping of major gene resistance to potato leafroll virus in *Solanum tuberosum* ssp. *andigena*. *Theor Appl Genet.* 114:1051–1058. doi:10.1007/s00122-006-0498-5.
- Waskom M. 2021. An Introduction to Seaborn—Seaborn 0.10.0 Documentation. Seaborn.pydata.org.
- Zhang H, Koblížková A, Wang K, Gong Z, Oliveira L, et al. 2014. Boom-bust turnovers of megabase-sized centromeric DNA in *solanum* species: rapid evolution of DNA sequences associated with centromeres. *Plant Cell.* 26:1436–1447. doi:10.1105/tpc.114.123877.
- Zhang T, Liu G, Zhao H, Braz GT, Jiang J. 2021. Chorus2: design of genome-scale oligonucleotide-based probes for fluorescence in situ hybridization. *Plant Biotechnol J.* doi:10.1111/pbi.13610.

Communicating editor: J. B. Endelman