Influence of neighboring small sequence variants on functional impact prediction

Jan-Simon Baasner[1], Dakota Howard[1,2] and Boas Pucker[1,3]

1 Genetics and Genomics of Plants, Center for Biotechnology, Bielefeld University, Bielefeld, Germany

2 Biology and Computer Science Department, Furman University, Greenville, South Carolina, USA

3 Evolution and Diversity, Plant Sciences, University of Cambridge, Cambridge, United Kingdom

JSB: janbaas@cebitec.uni-bielefeld.de

DH: dhoward@cebitec.uni-bielefeld.de

BP: bpucker@cebitec.uni-bielefeld.de

Corresponding author: Boas Pucker, bpucker@cebitec.uni-bielefeld.de

ORCIDs:

JSB: 0000-0003-3996-6817

DH: 0000-0002-7674-0385

BP: 0000-0002-3321-7471

Abstract

Once a suitable reference sequence is generated, genomic differences within a species are often assessed by re-sequencing. Variant calling processes can reveal all differences between two strains, accessions, genotypes, or individuals. These variants can be enriched with predictions about their functional implications based on available structural annotations i.e. gene models. Although these functional impact predictions on a per variant basis are often accurate, some challenging cases require the simultaneous incorporation of multiple adjacent variants into this prediction process. Examples are neighboring variants which modify each others' functional impact. Neighborhood-Aware Variant Impact Predictor (NAVIP) considers all variants within a given protein coding sequence when predicting the functional consequences. As a proof of concept, variants between the *Arabidopsis thaliana* accessions Columbia-0 and Niederzenz-1 were annotated. NAVIP is freely available on github: https://github.com/bpucker/NAVIP.

Introduction

Re-sequencing projects e.g. investigating many individuals or accessions of one species [1–3] are gaining relevance in plant research. Approaches similar to genome-wide association studies which are based on mapping-by-sequencing (MBS) were frequently applied [4–6]. They are boosted by an increasing availability of high quality reference genome sequences [7–12] and dropping sequencing costs [13, 14]. *De novo* assemblies are still beneficial for the detection of large structural variants [8, 11, 12, 15–17] and especially to reveal novel sequences [8, 11, 12, 18], but the reliable detection of modifying single nucleotide variants (SNVs) can be achieved based on (short) read mappings.

Once identified, the annotation of sequence variants in most species is performed by predicting their functional implications based on the available annotation of genes. Leading tools like ANNOVAR [19] and SnpEff [20] are currently performing this prediction by focusing on a single variant at a time. An impact prediction facilitates the identification of targets for post-GWAS analyses [21, 22]. Although the effect prediction for single variants is very efficient and usually correct, there is a minority of challenging cases in which predictions cannot be accurate based on a single variant alone. Multiple InDels could either lead to frameshifts or they compensate for each others' effect leaving the sequence with minimal modifications [23–25]. Two SNVs

occurring in the same codon could lead to a different amino acid substitution compared to the apparent effect resulting from an isolated analysis of each of these SNVs.

Here we present a new tool to accurately predict the combined effect of phased variants on annotated coding sequences. Neighborhood-Aware Variant Impact Predictor (NAVIP) was developed to investigate large variant data sets of plant re-sequencing projects, but is not limited to the annotation of variants in plants. As a proof of concept, NAVIP was deployed to identify cases between the *A. thaliana* accessions Columbia-0 (Col-0) and Niederzenz-1 (Nd-1) where an accurate impact prediction needs to consider multiple variants at a time [15].

Materials & Methods

Variant detection

Sequencing reads of Nd-1 [15] were mapped to the Col-0 reference genome sequence (TAIR9) [26] via BWA MEM v.0.7.13 [27] using the –m option to avoid spurious hits. Variant calling was performed via GATK v3.8 [28] based on the developers' recommendation. All processes were wrapped into custom Python scripts (https://github.com/bpucker/variant_calling) to facilitate automatic execution on a high performance compute cluster. An initial variant set was generated based on hard filtering criteria recommended by the GATK developers. The two following variant calling runs considered the set of surviving variants of the previous round as gold standard to avoid the need for hard filtering.

Variant validation

Since a high quality genome sequence assembly of Nd-1 was recently generated [12], we harnessed this sequence to validate all variants identified by short read mapping. Starting at the north end of each chromosome sequence, sorted variants were tested one after the other by taking the upstream sequence from Col-0, modifying it according to all upstream *bona fide* variants, and searching for it in the Nd-1 assembly (AdditionalFile1). Variants were admitted to the following analysis if the assembly supports them. This consecutive inspection of all variants enabled a reliable removal of false positives.

Variant impact prediction

Our Neighborhood-Aware Impact Predictor (NAVIP, https://github.com/bpucker/NAVIP) takes a VCF file containing sequence variants, a FASTA file containing the reference sequence, and a GFF3 file containing the annotation as input. Provided variants need to be homozygous or in a phased state to allow an accurate impact prediction per allele. Effects on all annotated transcripts are assessed per gene by taking the presence of all given variants into account. NAVIP generates a new VCF file with an additional annotation field and additional report files including FASTA files with the resulting sequences (see manual for details: https://github.com/bpucker/NAVIP/wiki).

Assessing predicted premature stop codons and frameshifts

SnpEff [20] was applied to the validated variant data set to predict the effects of single variants. To assess the influence of the underlying annotation, this prediction was performed based on TAIR10 [26] and Araport11 [29]. Predicted premature stop codons with two variants within the same codon were selected for comparison to the NAVIP prediction, because these cases have the potential to show different results.

Transcripts with predicted frameshifts were analyzed to identify downstream insertions/deletions which are compensating each others' effect i.e. the second frameshift is reverting an upstream frameshift. The distance between these events was analyzed by the third module of NAVIP.

Experimental validation of variants

A. thaliana Nd-1 plants were grown as previously described [15]. DNA for PCR experiments was extracted from leaf tissue using a cetyltrimethylammonium bromide (CTAB)-based method as previously described [30]. Oligonucleotides flanking regions with variants of interest were designed manually (AdditionalFile2) and purchased from Metabion (http://www.metabion.com/). Amplification via PCR, analysis of PCR products, purification of PCR products, Sanger sequencing, and evaluation of results was performed as previously described [31].

Results

Variant detection and validation

Nd-1 reads were mapped against the Col-0 reference genome sequence (TAIR9). Based on 124,662,140 mapped paired-end reads, 384,622 variants were detected in the first variant calling round of this study. This initial set was extended over three additional rounds of variant calling leading to over one million of variants. The variant calling was stopped, because no substantial increase in the number of novel variants was observed during the last rounds. An assembly based on independent Single Molecule Real Time (SMRT) sequencing reads supported 772,644 (76.6%) of all variants detected during the last iteration (AdditionalFile3, Fig. 1). On average, one variant was observed every 154 bp between Col-0 and Nd-1. SNV frequencies ranged from one event in 225 bp on Chr5 to one event in 158 bp on Chr4. InDel frequencies ranged from one event in 1,051 bp on Chr5 to one event in 809 bp on Chr4.

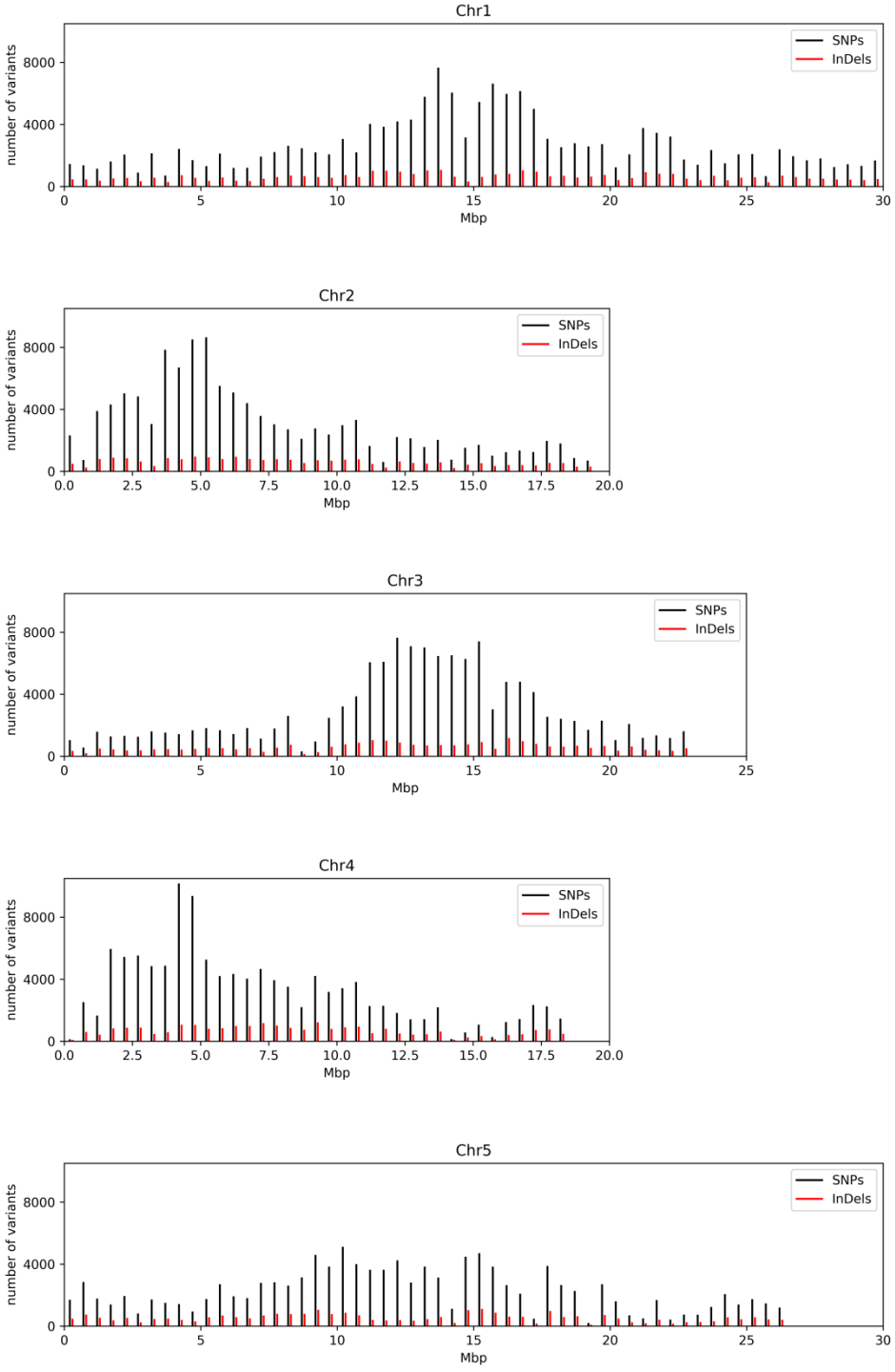


Fig. 1: Genome-wide distribution of sequence variants between Col-0 and Nd-1.

Distributions of SNVs and InDels over the chromosome sequences of Col-0 were visualized as previously described [15].

Although the repeated variant calling processes were intended to increase the sensitivity, we did not observe a substantial improvement between the second and third round. This saturation indicates that no additional variants would be detected in further variant calling rounds. The number of detected variants as well as the validation rate was almost constant (Table 1).

Table 1: Total and validated number of variants.

Variant data set	Total variants	Validated variants
Initial set based on hard filtering	384,617	350,005 (90.1%)
Soft filtering round 1	1,006,920	771,449 (76.6%)
Soft filtering round 2	1,008,610	772,612 (76.6%)
Soft filtering round 3	1,008,629	772,643 (76.6%)

Experimental validation

Randomly selected loci with two SNVs within one codon were experimentally validation via PCR and amplicon sequencing (Table 2). Successful sequencing reactions show a validation rate of >95%.

Table 2: Neighboring SNVs validated in Nd-1 via PCR and amplicon sequencing. Sequences of oligonucleotides used for the amplicon generation are listed in AdditionalFile2.

AGI	Fw primer	Rv primer	Status
At1g30545	N400	N401	Validated
At3g55500	N402	N403	Validated
At3g26770	N406	N407	Validated
At4g30570	N408	N409	Validated
At1g28150	N410	N411	Validated
At1g35430	N412	N413	Validated
At4g27230	N414	N415	One SNV failed
At5g60230	N424	N425	2 validated
At1g31820	N426	N427	4 validated / 1 failed

Relevance of NAVIP

Running NAVIP on this *A. thaliana* data set (AdditionalFile4) took about 5 minutes with a single core and a peak memory usage of about 3 GB RAM. Since SnpEff is one of the most frequently applied tools for the annotation of variants, the NAVIP output was compared with SnpEff predictions. SnpEff was applied to the same data set based on the Araport11 annotation. Interesting cases for comparison are codons containing at least two SNVs. Of 75 premature stop codons predicted in such codons by SnpEff, 73 were predicted as amino acid substitutions by NAVIP (Fig. 2). While a single SNV would cause a premature stop codon, the simultaneous presence of two SNVs results in an amino acid encoding codon. In total, 702 premature stop codons were predicted by SnpEff thus 9.6 % of them were false positives. NAVIP revealed that tyrosine occurs frequently instead of a premature stop codon, because the tyrosine codons are very similar to two of the three stop codons.

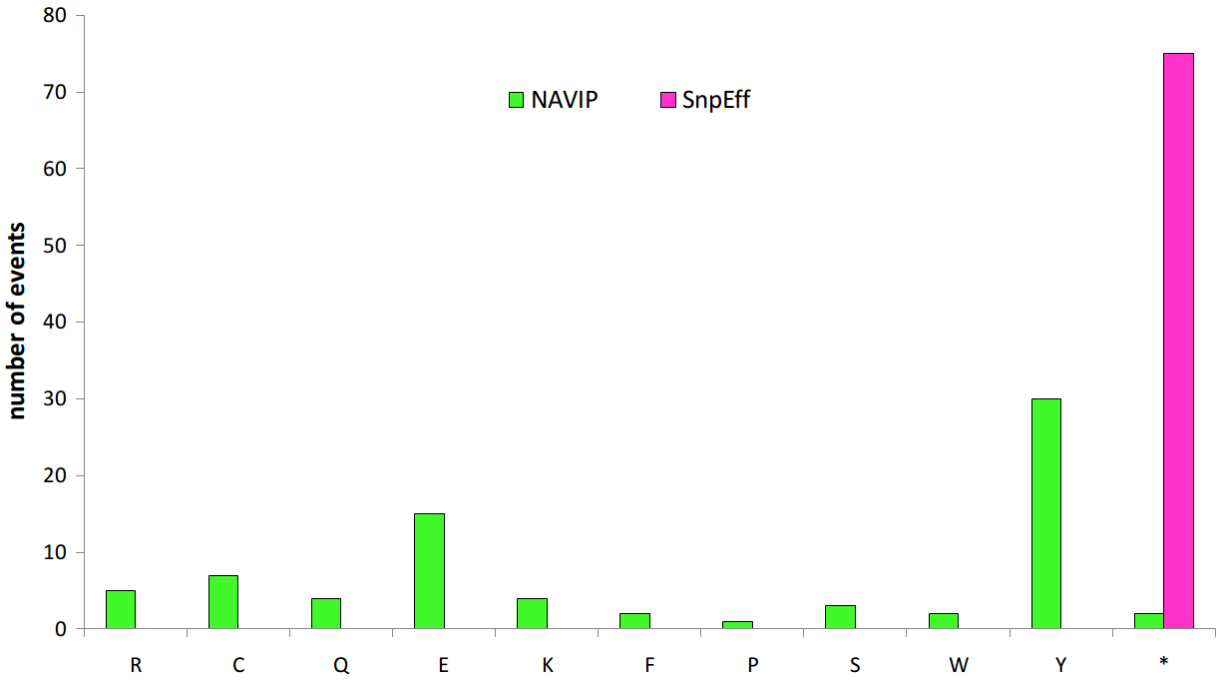


Fig. 2: Second site variants turn predicted premature stop codons into amino acid substitutions.

Premature stop codons predicted by SnpEff (pink) are frequently amino acid substitutions if a second variant is located within the same codon. NAVIP revealed 73 false positive predictions of premature stop codons by SnpEff which are in fact amino acid substitutions (green).

InDels can compensate each others' frameshift when occurring together. Since premature stop codons can emerge by chance following a frameshift, the distance between such InDels was analyzed. This length distribution revealed that most compensating InDels (cInDels) occur within a short distance of 2-8 bp (Fig. 3). Multiples of three are more frequent than other distances of a similar size.

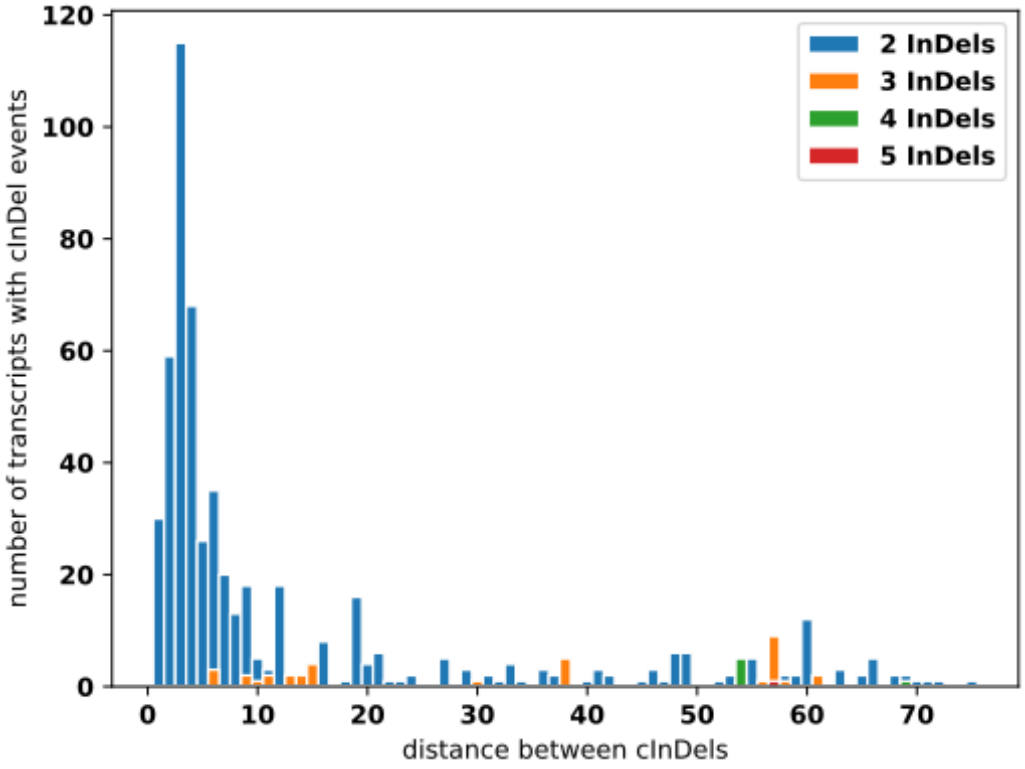


Fig. 3: Distance between compensating InDels (cInDels).

An InDel can compensate the frameshift caused by an upstream InDel. Distances between such cInDels are short and frequently multiples of three. In total, 484 genes contain cInDels.

Discussion

Variant validation, frequency, and distribution

Although differentiation between *bona fide* variants (true positives) and false positives based on a high quality genome sequence assembly worked very well, false negatives were not taken into account and might even bias this classification approach by preventing the validation of neighboring variants (AddtionalFile1). If a variant is missed by the initial variant calling, its presence in the flanking sequence used during the validation process will prevent a proper match. Therefore, the number of variants could be slightly higher than reported here. Nevertheless, this conservative approach was selected to minimize the risk of keeping false positive variants. There is always a trade-off between sensitivity and specificity in the variant calling process [32] and our approach is in strong favor of specificity. However, the number of identified and validated variants exceeds previous reports of 485,887 variants between Col-0 and Nd-1 [15]. Instead the observed variant frequency is closer to the results of a comparison between Bur-0 and Col-0 [33]. Despite the difference in total numbers, the distribution on the chromosome scale is similar to the previous comparison of Col-0 and Nd-1 [15]. It seems that Chr4 is the most variable one, while Chr5 is the least variable one between both compared accessions.

Successful validation via PCR and amplicon sequencing supported the presence of two SNVs within one codon. Although these variants are perceived as two SNVs, the underlying mechanism could be a multiple nucleotide polymorphism (MNP). It would be interesting to see if these SNVs occur independently in other accessions in the *A. thaliana* population.

Functional implications of variants

We developed NAVIP to assess the impact of neighboring variants on protein coding sequences. The presence of the 557 cases described here for the comparison of two *A. thaliana* accessions demonstrates the necessity to have such a tool at hand. NAVIP revealed the

presence of second site mutations that compensate other variants e.g. turning a premature stop codon into an amino acid substitution or compensation of a frameshift. The purpose of NAVIP is not to replace existing tools, but to add novel functionalities to established tools like SnpEff [20]. This could boost the power of re-sequencing studies by opening up the field of compensating or in general mutually influencing variants. Such variants have the potential to reveal new insights into patterns of molecular evolution and especially co-evolution of sites. Although the number of cases is probably small, the consideration of multiple variants during the effect prediction could reveal novel targets in GWAS-like approaches. The remaining challenge is now the reliable detection of sequence variants prior to the application of NAVIP. For heterozygous species phasing of these variants is another task that needs to be addressed. The correct prediction of functional implications relies on the correct assignment of variants to respective haplophases. If provided with accurately phased variants, NAVIP can perform predictions for highly heterozygous and even polyploid species.

Availability of data

The data sets supporting the results of this article are included within the article and its additional files. Python scripts developed and applied for this study are available on github: https://github.com/bpucker/NAVIP (https://doi.org/10.5281/zenodo.2620396) https://github.com/bpucker/variant_calling (https://doi.org/10.5281/zenodo.2616418).

Authors' contribution

BP designed research. JSB wrote the NAVIP code. JSB, DH, and BP conducted bioinformatic analyses. DH and BP performed experimental validation. BP wrote the manuscript. All authors read and approved the final version.

Acknowledgements

We acknowledge support by members of Genetics and Genomics of Plants, Bioinformatics Resource Facility, and Sequencing Core Facility at the Center of Biotechnology. We thank Hanna Schilbert for critical reading of the manuscript.

References

1. Alonso-Blanco C, Andrade J, Becker C, Bemm F, Bergelson J, Borgwardt KM, et al. 1,135 Genomes Reveal the Global Pattern of Polymorphism in *Arabidopsis thaliana*. Cell. 2016;166:481–91.

2. Duan N, Bai Y, Sun H, Wang N, Ma Y, Li M, et al. Genome re-sequencing reveals the history of apple and supports a two-stage model for fruit enlargement. Nat Commun. 2017;8:249.

3. Lobaton JD, Miller T, Gil J, Ariza D, de la Hoz JF, Soler A, et al. Resequencing of Common Bean Identifies Regions of Inter–Gene Pool Introgression and Provides Comprehensive Resources for Molecular Breeding. Plant Genome. 2018;11. doi:10.3835/plantgenome2017.08.0068.

4. James GV, Patel V, Nordström KJ, Klasen JR, Salomé PA, Weigel D, et al. User guide for mapping-by-sequencing in Arabidopsis. Genome Biol. 2013;14:R61.

5. Mascher M, Jost M, Kuon J-E, Himmelbach A, Aßfalg A, Beier S, et al. Mapping-by-sequencing accelerates forward genetics in barley. Genome Biol. 2014;15:R78.

6. Wu Y, Zheng Z, Visscher PM, Yang J. Quantifying the mapping precision of genome-wide association studies using whole-genome sequencing data. Genome Biol. 2017;18:86.

7. Dohm JC, Minoche AE, Holtgräwe D, Capella-Gutiérrez S, Zakrzewski F, Tafer H, et al. The genome of the recently domesticated crop plant sugar beet (*Beta vulgaris*). Nature. 2014;505:546–9.

8. Zapata L, Ding J, Willing E-M, Hartwig B, Bezdan D, Jiao W-B, et al. Chromosome-level assembly of *Arabidopsis thaliana* Ler reveals the extent of translocation and inversion polymorphisms. Proc Natl Acad Sci. 2016;113:E4052–60.

9. Jarvis DE, Ho YS, Lightfoot DJ, Schmöckel SM, Li B, Borm TJA, et al. The genome of *Chenopodium quinoa*. Nature. 2017;542:307–12.

259 10. Lightfoot DJ, Jarvis DE, Ramaraj T, Lee R, Jellen EN, Maughan PJ. Single-molecule sequencing and Hi-
260 C-based proximity-guided assembly of amaranth (Amaranthus hypochondriacus) chromosomes provide
261 insights into genome evolution. BMC Biol. 2017;15:74.

262 11. Michael TP, Jupe F, Bemm F, Motley ST, Sandoval JP, Lanz C, et al. High contiguity *Arabidopsis*
263 *thaliana* genome assembly with a single nanopore flow cell. Nat Commun. 2018;9:541.

264 12. Pucker B, Holtgraewe D, Stadermann KB, Frey K, Huettel B, Reinhardt R, et al. A Chromosome-level
265 Sequence Assembly Reveals the Structure of the *Arabidopsis thaliana* Nd-1 Genome and its Gene Set.
266 PLOS ONE. 2019: e0216233. doi:10.1371/journal.pone.0216233.

267 13. Stein LD. The case for cloud computing in genome informatics. Genome Biol. 2010;11:207.

268 14. Christensen KD, Dukhovny D, Siebert U, Green RC. Assessing the Costs and Cost-Effectiveness of
269 Genomic Sequencing. J Pers Med. 2015;5:470–86.

270 15. Pucker B, Holtgräwe D, Sörensen TR, Stracke R, Viehöver P, Weisshaar B. A *De Novo* Genome
271 Sequence Assembly of the *Arabidopsis thaliana* Accession Niederzenz-1 Displays Presence/Absence
272 Variation and Strong Synteny. PLOS ONE. 2016;11:e0164321.

273 16. Fan X, Chaisson M, Nakhleh L, Chen K. HySA: a Hybrid Structural variant Assembly approach using
274 next-generation and single-molecule sequencing technologies. Genome Res. 2017;27:793–800.

275 17. Wala JA, Bandopadhayay P, Greenwald NF, O'Rourke R, Sharpe T, Stewart C, et al. SvABA: genome-
276 wide detection of structural variants and indels by local assembly. Genome Res. 2018;28:581–91.

277 18. Zhou Y, Massonnet M, Sanjak JS, Cantu D, Gaut BS. Evolutionary genomics of grape (Vitis vinifera ssp.
278 vinifera) domestication. Proc Natl Acad Sci. 2017;114:11715–20.

279 19. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-
280 throughput sequencing data. Nucleic Acids Res. 2010;38:e164.

281 20. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and
282 predicting the effects of single nucleotide polymorphisms, SnpEff. Fly (Austin). 2012;6:80–92.

283 21. Hou L, Zhao H. A review of post-GWAS prioritization approaches. Front Genet. 2013;4.
284 doi:10.3389/fgene.2013.00280.

285 22. Ries D, Holtgräwe D, Viehöver P, Weisshaar B. Rapid gene identification in sugar beet using deep
286 sequencing of DNA from phenotypic pools selected from breeding panels. BMC Genomics. 2016;17.
287 doi:10.1186/s12864-016-2566-9.

288 23. Cao J, Schneeberger K, Ossowski S, Günther T, Bender S, Fitz J, et al. Whole-genome sequencing of
289 multiple *Arabidopsis thaliana* populations. Nat Genet. 2011;43:956–63.

290 24. Gan X, Stegle O, Behr J, Steffen JG, Drewe P, Hildebrand KL, et al. Multiple reference genomes and
291 transcriptomes for *Arabidopsis thaliana*. Nature. 2011;477:419–23.

292 25. Schneeberger K, Ossowski S, Ott F, Klein JD, Wang X, Lanz C, et al. Reference-guided assembly of four
293 diverse *Arabidopsis thaliana* genomes. Proc Natl Acad Sci U S A. 2011;108:10249–54.

294 26. Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, et al. The *Arabidopsis* Information
295 Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res. 2012;40 Database
296 issue:D1202–10.

297 27. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.
298 ArXiv13033997 Q-Bio. 2013. http://arxiv.org/abs/1303.3997. Accessed 16 Oct 2018.

299 28. Auwera GAV der, Carneiro MO, Hartl C, Poplin R, Angel G del, Levy-Moonshine A, et al. From FastQ
300 Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. Curr Protoc
301 Bioinforma. 2013;43:11.10.1-11.10.33.

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

304 30. Rosso MG, Li Y, Strizhov N, Reiss B, Dekker K, Weisshaar B. An Arabidopsis thaliana T-DNA
305 mutagenized population (GABI-Kat) for flanking sequence tag-based reverse genetics. Plant Mol Biol.
306 2003;53:247–59.

31. Pucker B, Holtgräwe D, Weisshaar B. Consideration of non-canonical splice sites improves gene prediction on the *Arabidopsis thaliana* Niederzenz-1 genome sequence. BMC Res Notes. 2017;10. doi:https://doi.org/10.1186/s13104-017-2985-y.

32. Olson ND, Lund SP, Colman RE, Foster JT, Sahl JW, Schupp JM, et al. Best practices for evaluating single nucleotide variant calling methods for microbial genomics. Front Genet. 2015;6. doi:10.3389/fgene.2015.00235.

33. Ossowski S, Schneeberger K, Clark RM, Lanz C, Warthmann N, Weigel D. Sequencing of natural strains of *Arabidopsis thaliana* with short reads. Genome Res. 2008;18:2024–33.

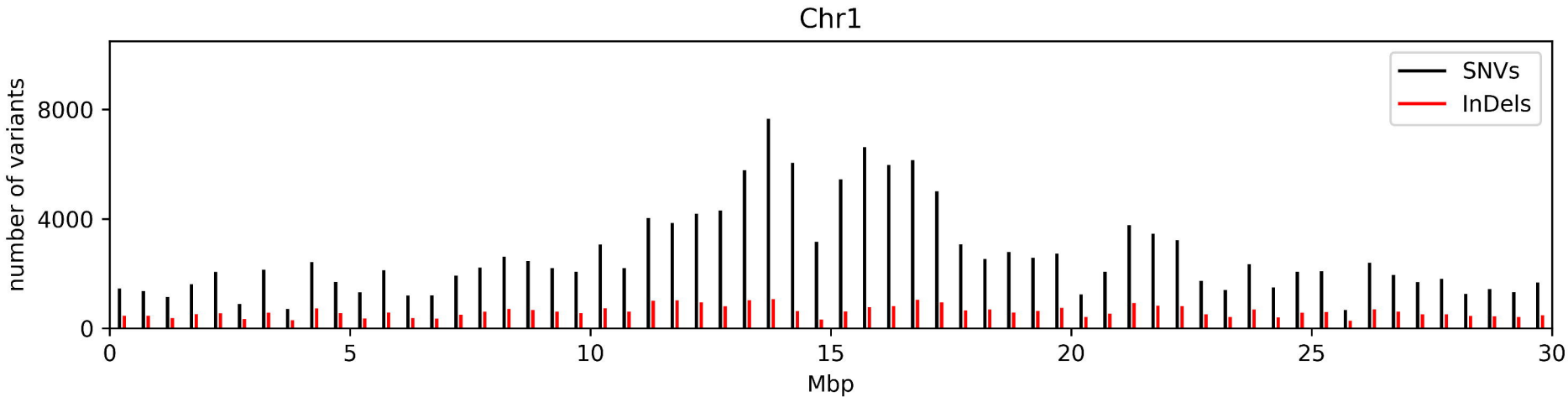
Additional Files

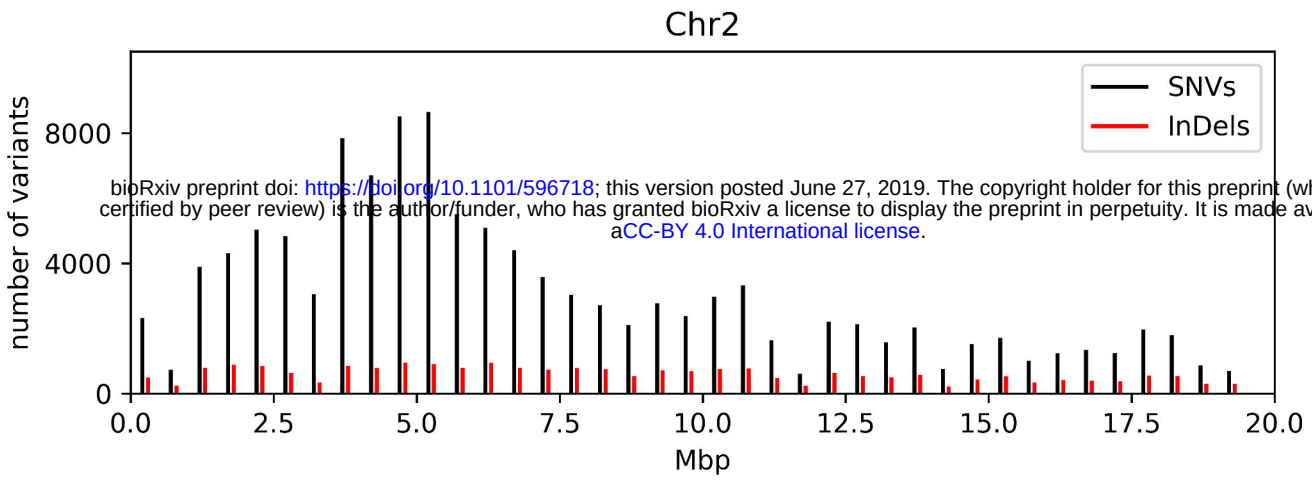
AdditionalFile1: Schematic illustration of the variant validation process.

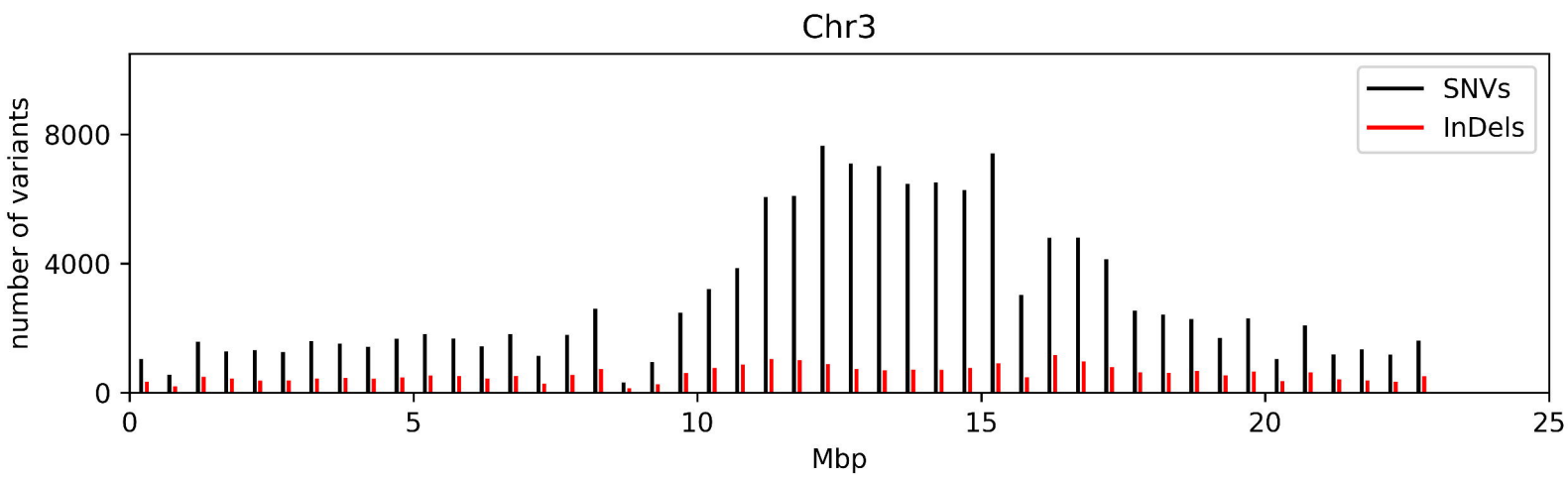
AdditionalFile2: Oligonucleotide sequences used for the validation of randomly selected variants.

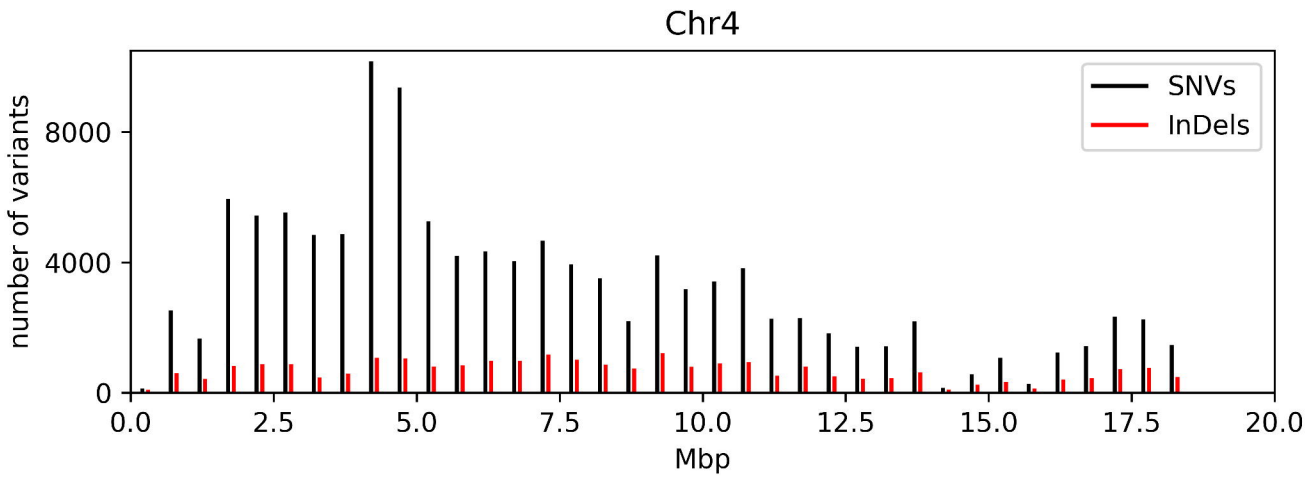
AdditionalFile3: Final set of validated variants.

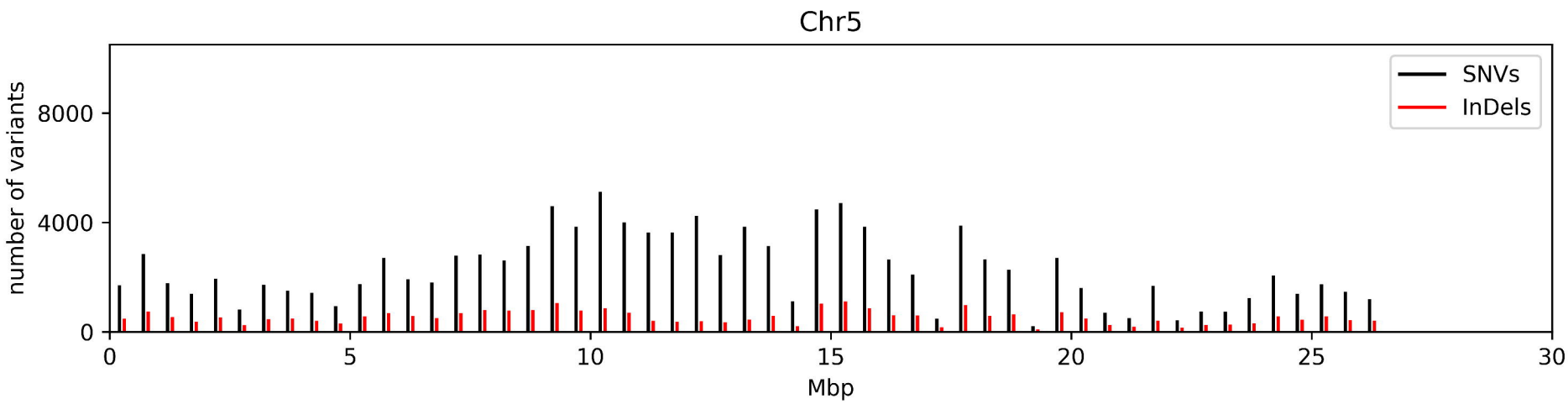
AdditionalFile4: NAVIP annotation of variants between Nd-1 and Col-0.

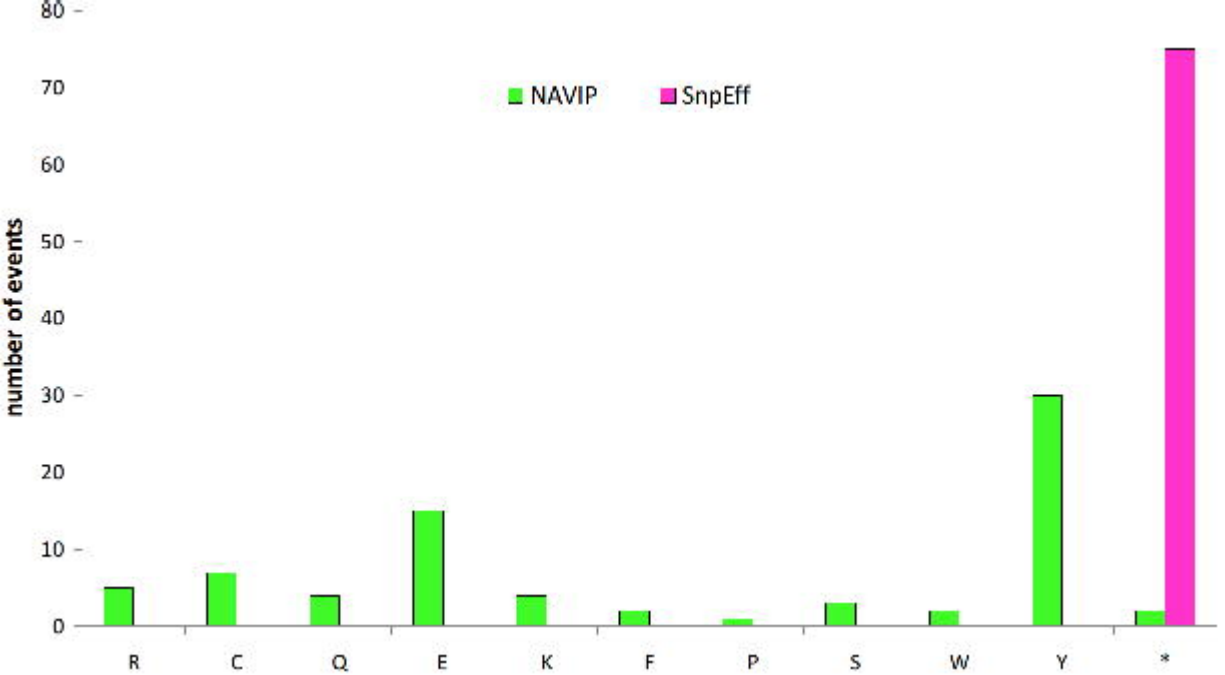


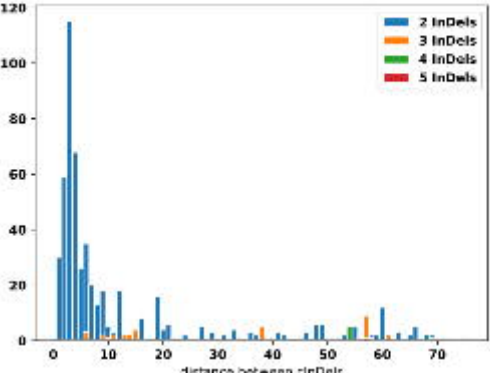









80
70
60
50
40
30
20
10
0
number of events
NAVIP
SnpEff
R
C
Q
E
K
F
P
S
W
Y
*


2 InDels
3 InDels
4 InDels
5 InDels
number of transcripts with cInDel events
distance between cInDels
0
10
20
30
40
50
60
70
0
20
40
60
80
100
120