NAVIP: Unraveling the Influence of Neighboring Small Sequence Variants on Functional Impact Prediction

Jan-Simon Baasner[1], Andreas Rempel[2,3], Dakota Howard[4], and Boas Pucker[1,5,]*

1 Genetics and Genomics of Plants, Faculty of Biology & Center for Biotechnology, Bielefeld University, 33615 Bielefeld, Germany

2 Genome Informatics, Faculty of Technology & Center for Biotechnology, Bielefeld University, 33615 Bielefeld, Germany

3 Graduate School "Digital Infrastructure for the Life Sciences" (DILS), Bielefeld Institute for Bioinformatics Infrastructure (BIBI), Bielefeld University, 33615 Bielefeld, Germany

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

5 Plant Biotechnology and Bioinformatics, Institute of Plant Biology & BRICS, TU Braunschweig, 38106 Braunschweig, Germany

* Correspondence: b.pucker@tu-braunschweig.de

Abstract

Once a suitable reference sequence has been generated, intraspecific variation is often assessed by re-sequencing. Variant calling processes can reveal all differences between 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 include neighboring variants which modify each other's functional impact. The Neighborhood-Aware Variant Impact Predictor (NAVIP) considers all variants within a given protein coding sequence when predicting the effect. 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) and accessible through a web server (https://pbb-tools.de).

Author Summary

Intraspecific variation gains increasing relevance as reference genome sequences are available for many investigated (plant) species. Understanding the effects of sequence variants between individuals of a population is a challenge. SnpEff (Cingolani et al., 2012) is the current standard tool for predicting the functional impact of sequence variants, but only considers one sequence variant at a time. We developed NAVIP to properly handle cases in which multiple sequence variants cluster together and influence each other's functional impact. A comparison of two *Arabidopsis thaliana* accessions demonstrates the importance of considering multiple sequence variants simultaneously for the prediction of changes in encoded proteins. NAVIP is universally applicable to any organism for which the relevant sequence information and structural annotation is available. All underlying code is freely available on GitHub and we operate a web server for users' convenience.

Keywords: sequence variants, variant annotation, SNPs, SNVs, InDels, mutations

Introduction

Re-sequencing projects examining many individuals or accessions of a species [1–4], are becoming increasingly important in plant research. Approaches similar to genome-wide association studies (GWAS) which are based on mapping-by-sequencing (MBS) are frequently applied in a wide range of crop species [5–8]. They are boosted by a rapidly increasing availability of high-quality reference genome sequences for crops [9–13], technological advances in long-read sequencing [14], and low sequencing costs [15,16]. *De novo* assemblies are still beneficial for the detection of large structural variants [17–22] and especially to reveal novel sequences [18,19,21,23], but the reliable detection of modifying single nucleotide variants (SNVs) can be achieved based on (short) read mappings. Well established tools for the small sequence variant discovery in plants are BMA MEM and GATK [24–27]. In recent years, long-read sequencing is gaining popularity in studies exploring the intraspecific diversity, as more sequence variants can be detected in previously inaccessible genomic regions [28,29]. One of the most frequently used tools for long read mapping is minimap2 [30] that can handle both relevant technologies, Pacific Biosciences and Oxford Nanopore Technologies, well. Hundreds of dedicated variant calling tools have been developed to harness the specific potential and to cope with challenges that come with long reads. Famous tools for the discovery of SNVs based on long reads are Longshot [31], SVIM-asm [32], and Sniffles2 [33]. One advantage of long reads is the ability to assign small sequence variants to different haplophases.

Once identified, the annotation of sequence variants is performed by predicting their functional implications based on the available gene models (structural annotation). Leading tools such as ANNOVAR [34], VEP [35], and SnpEff [36] currently perform this prediction efficiently by focusing on a single variant at a time. An impact prediction facilitates the identification of targets for post-GWAS analyses and can lead to the identification of small sequence variants that form the molecular basis of commercially relevant phenotypic differences [7,37,38]. Although the effect prediction for single variants is computationally efficient and usually correct, there are challenging cases in which predictions based on a single variant alone cannot be accurate. (1) Multiple InDels could either lead to frameshifts or they could compensate for each other’s effect leaving the sequence with minimal modifications [39–41] and (2) two SNVs occurring in the same codon could lead to a different amino acid substitution compared to the apparent effects resulting from an isolated analysis of each of these SNVs. It is important to note that SNVs and InDels can also influence each other’s effects.

Here we present a computational tool for accurately predicting the combined effect of phased variants on annotated coding sequences. The 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 used 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.

Results

Features of NAVIP

NAVIP predicts the functional impact of sequence variants by considering all sequence variants affecting the coding sequence of a gene simultaneously. Users need to supply a set of sequence variants (VCF), a reference genome sequence (FASTA), and a structural annotation (GFF3). NAVIP returns an annotated VCF file and FASTA files with corrected coding and polypeptide sequences. If phased sequence variants are provided in the VCF file, NAVIP performs separate analyses for the different haplophases.

NAVIP can be retrieved from a GitHub repository (https://github.com/bpucker/NAVIP) and is executable without installation. Additionally, NAVIP is also available free of charge through a web server (https://pbb-tools.de/NAVIP). This makes NAVIP accessible to a wide range of users and applicable to data sets of various sizes. Uploaded files are used only for the intended analysis and are deleted 48 hours after offering the results for download. The web server is able to send notification emails upon completion of a job, which can serve as documentation and facilitate the analysis of large data sets.

Relevance of NAVIP for prediction of premature stop codons

Running NAVIP on an *A. thaliana* Nd-1 data set with 644,261 SNVs (S1 File, S2 File) took about 5 minutes on a single core with a peak memory usage of about 3 GB RAM. To the best of our knowledge, SnpEff is the most frequently used tool for the annotation of variants and is also universally applicable. Therefore, the NAVIP output was compared with the SnpEff predictions generated for the same data set and structural annotation. The results are largely congruent, but interesting cases for comparison are predictions of premature stop codons, as these may have severe biological consequences. While a single SNV would cause a premature stop codon, the simultaneous presence of two SNVs can result in an amino acid encoding codon (**Figure 1a**). Of 600 premature stop codons predicted by SnpEff, 144 were identified as amino acid substitutions when considering multiple SNVs in the same codon via NAVIP (**Figure 1b**). Given the total of 600 predicted premature stop codons in this Nd-1 data set, 24% were false positive predictions. NAVIP revealed that tyrosine frequently occurs instead of a premature stop codon because the tyrosine codons are very similar to two of the three stop codons. There are also 17 additional premature stop codons predicted by NAVIP, which are the consequence of two sequence variants affecting the same codon. Despite the surprisingly large difference between the SnpEff and NAVIP results when it comes to predicting premature stop codons, the differences in affected genes are smaller. Many genes with a predicted premature stop codon have multiple downstream premature stop codons. While the prediction of an individual premature stop codon might be wrong for a certain position, the gene can still be correctly identified by both tools as harboring premature stop codons if additional ones occur further downstream (S3 File). If a premature stop codon results in a loss-of-function event, the accumulation of additional variants is likely due to a lack of purifying selection. To support the assumption that genes with premature stop codons lost their function, the rate of amino acid changing variants in these genes was compared to all other genes (**Figure 1c, Figure 1d**). The number of variants changing amino acids (aa_N) to those resulting in the same amino acid (aa_S) was calculated for all genes (aa_N/aa_S). A significantly higher proportion of amino acid changing variants was observed in genes with predicted premature stop codons compared to all other genes (Mann-Whitney U test, p-value=10^{-161}). Premature stop codons might frequently appear in genes undergoing pseudogenization that are barely expressed, as purifying selection would be weak or even absent in these cases. Therefore, we investigated the expression of genes with premature stop codons in *A. thaliana*. A comparison of the average expression of genes with a premature stop codon against all other protein encoding genes (**Figure 1e, Figure 1f**) revealed a significantly lower expression of genes with premature stop codons (Mann-Whitney U test, p-value=10^{-70}).

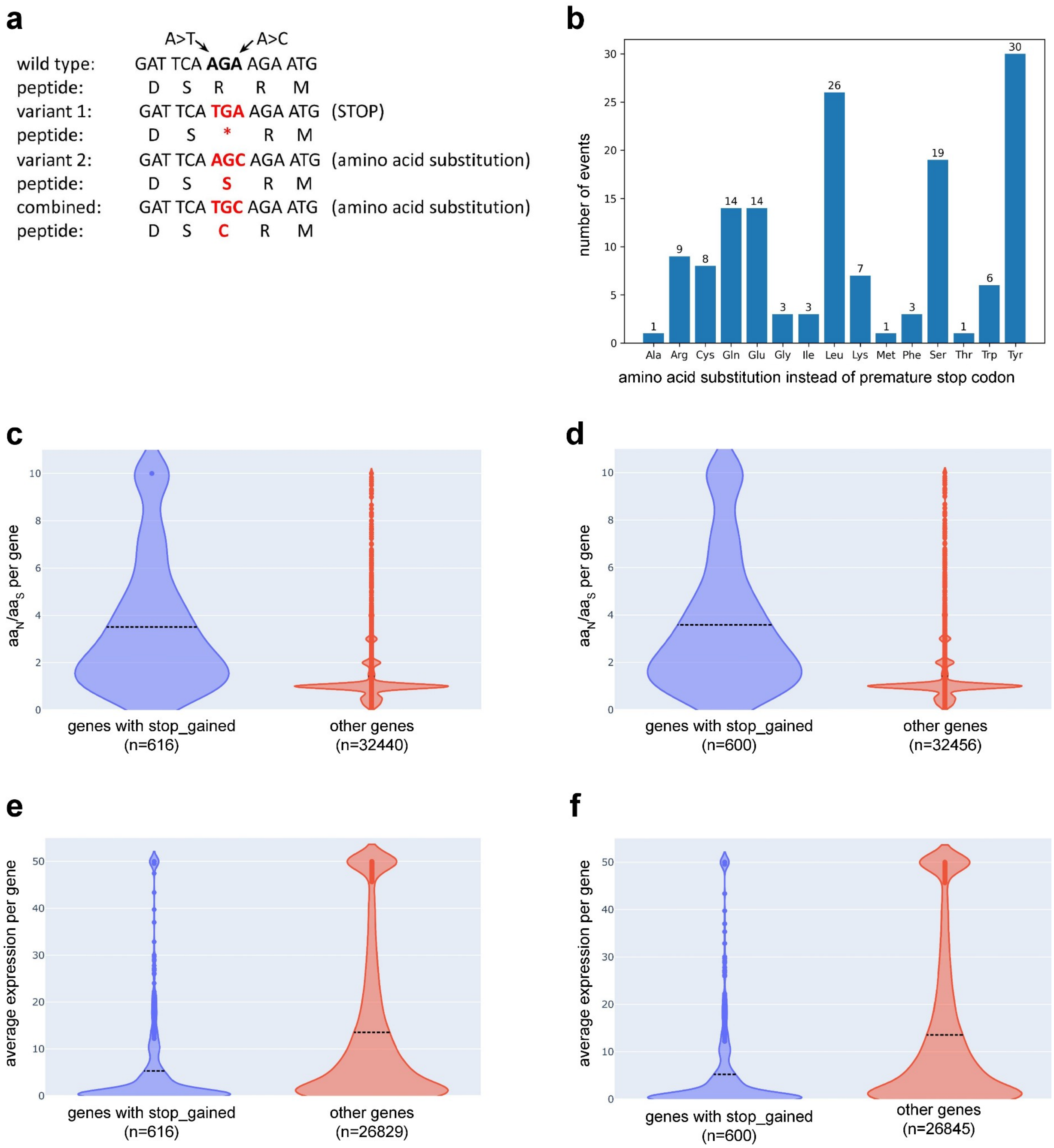

Figure 1: (a) This illustration shows the concept of two SNVs affecting the same codon resulting in different prediction outcomes. (b) Second site variants within the same codon turn premature

stop codons predicted by SnpEff into amino acid substitutions. In 144 cases, premature stop codons are substitutions by the respective amino acids. (c) The proportion of amino acid changing variants is significantly higher in genes with premature stop codons predicted by NAVIP (blue) compared to all other genes (red). aa_N is the number of variants changing an amino acid residue and aa_S is the number of variants resulting in the same amino acid residue. (d) The proportion of amino acid changing variants is significantly higher in genes with premature stop codons predicted by SnpEff (blue) compared to all other genes (red). Data underlying these visualizations are available in S3 File. (e) Comparison of the average expression of genes with a premature stop codon predicted by NAVIP against all other protein encoding genes with available expression data. (f) Comparison of the average expression of genes with a premature stop codon predicted by SnpEff against all other protein encoding genes with available expression data.

To demonstrate the scalability of NAVIP, we processed 200 samples from the 1135 accession comparison study [1]. On average, an accession harbored 498 cases of stop codons predicted by SnpEff were classified as amino acid substitutions by NAVIP (S4 File).

While premature stop codons are probably the most severe changes, we also explored the influence of neighboring SNVs on amino acid substitutions between Col-0 and Nd-1. A total of 50,122 amino acid substitution predictions were analyzed including cases in which one of the annotation tools predicts no change of the amino acid. Predictions of NAVIP and SnpEff were congruent in 46,680 cases (93.1%) and differed in 3442 cases (6.9%) (S5 File).

Role of compensating InDels (cInDels)

InDels can compensate for each others' frameshift when occurring together in the same haplophase (**Figure 2a**). While the first InDel can alter the reading frame, the second one could revert the reading frame back to the original one, thus resulting in only a few altered codons enclosed by the two events. Since premature stop codons can emerge in the novel codons following the first frameshift, the distance between such InDels is expected to be very small. An analysis of the distance distribution of the InDels between Nd-1 and Col-0 (S6 File) revealed that most compensating InDels (cInDels) occur within a very short distance of 2-8 bp (**Figure 2b**). Multiples of three are more frequent than other distances of a similar size, which might be connected to the length of codons. Since *A. thaliana* is considered highly homozygous, we assume that all identified sequence variants are located in the same haplophase.

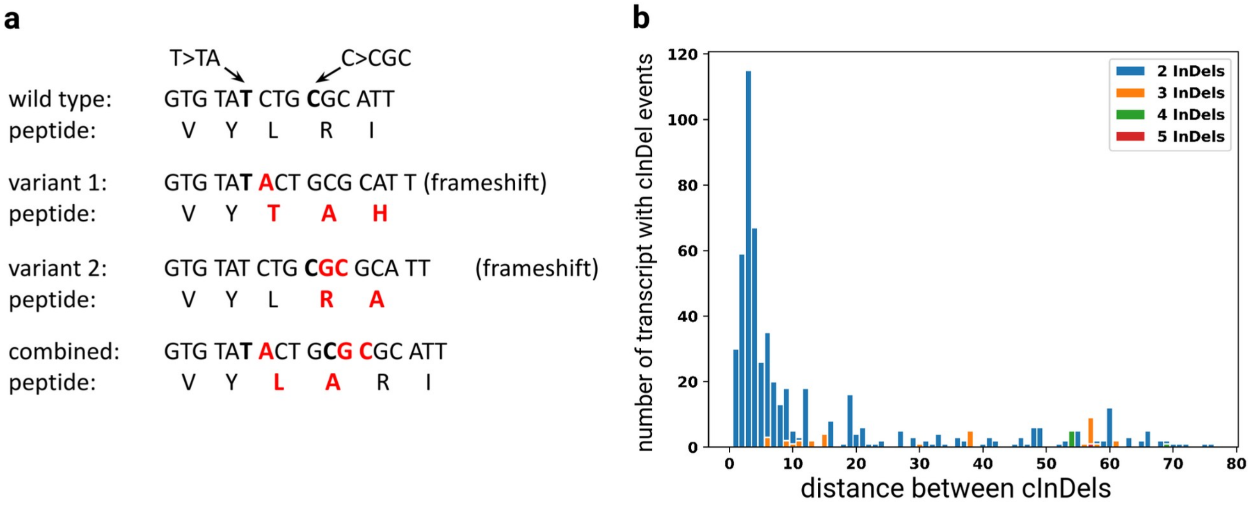

Figure 2: (a) Theoretical concept of two InDels compensating each others' frameshift. The first insertion changes the reading frame, while the second insertion shifts the reading frame back to the original one. While each individual variant would suggest a loss-of-function due to a frameshift mutation, the combination of both results in 'only' two additional amino acids in the gene product. (b) Distribution of distances between compensating InDels (cInDels). As the

second InDel can compensate for the frameshift caused by the upstream InDel, distances between such cInDels are short and frequently multiples of three. In total, 484 genes were identified to contain cInDels in the Nd-1 data set.

Discussion

This study demonstrates features of NAVIP by utilizing a previously generated set of high confidence sequence variants [26]. There is always a trade-off between sensitivity and specificity in the variant calling process [26,42] (see S1 File for details). The benchmarking of NAVIP is conducted by comparing it with SnpEff, which controls for the quality of the sequence variant dataset to minimize its impact on the results. As an additional validation of the outcome, NAVIP results were analyzed for additional amino acid substitutions in genes with premature stop codons. The frequency of such variants was higher in genes with premature stop codons compared to others suggesting a lack of purification selection in these genes which could point to pseudogenization. The comparison against all other genes also clearly revealed the increased frequency of amino acid substitutions in genes with premature stop codons. Additionally, a low expression of genes with premature stop codons compared to other genes suggests a pseudogenization. In summary, the properties observed for genes with premature stop codons match the expectations thus supporting the biological validity of the data set.

One motivation for the development of NAVIP was to fill a gap that exists between variant calling and variant annotation software. Variant calling involves the identification of genetic variants from raw sequencing data. This process typically features algorithms that analyze read alignments and uses statistical models to detect variants. Variant callers such as GATK [60] produce VCF files containing potential genetic variants. Variant annotation, on the other hand, assigns functional relevance to identified variants. This step requires databases and algorithms to provide additional information about each variant. Annotation tools such as ANNOVAR [34], VEP [35], or SnpEff [36] process VCF files previously generated by callers, rather than performing the variant calling themselves, thus losing access to the original read information. The separation between these two steps is due to technical and conceptual differences and serves several purposes. First, a separation of concerns: Variant calling focuses on the detection of variations, while annotation concentrates on the interpretation of those variants, allowing for specialized optimization of each step without complicating the other. Second, computational efficiency: Calling variants requires processing raw sequencing data, which can be computationally intensive. A streaming application would need to stop processing and accumulate all variants until there is complete gene information before annotating, which can be challenging in terms of memory usage, especially for large genes or when dealing with many samples simultaneously. Thus, separating the annotation step from the initial variant calling allows for a more efficient use of computational resources. Third, data flow and scalability: By separating calling and annotation, researchers can perform these steps independently, allowing for parallel processing and easier scaling of analysis pipelines. The VCF format used in variant calling is optimized for documenting detected variants, while other annotation formats are better suited for downstream analysis.

We developed NAVIP to simultaneously assess the impact of all neighboring sequence variants in protein encoding sequences and to be universally applicable. The described cases in the comparison of two *A. thaliana* accessions demonstrate the necessity to have such a tool at hand. NAVIP revealed the presence of second site mutations that compensate for other variants, e.g. turning a presumed premature stop codon into an amino acid substitution or vice versa. Another example are frameshifts resulting from InDels that are compensated by downstream InDels, which shift the reading frame back to the original pattern. Neglecting these interactions of sequence variants during the functional impact prediction can lead to mis-annotation. While NAVIP was developed to accurately predict changes in the polypeptide sequence based on DNA sequence variants, downstream tools are needed to predict consequences of these changes on the function of proteins. Tools like SIFT [43], PolyPhen-2 [44], or SNAP2 [45] could be applied for this next step. Many computational tools for the assessment of DNA sequence variant impact focus on human data sets [46–49]. The objective is often to identify pathogenic variants [43,50]. Universally applicable tools like SnpEff [36], which are also suitable to analyze plant data sets, predict the impact of isolated sequence variants. The purpose of NAVIP is to offer novel functionalities to the plant science community and other communities working on non-model organisms. NAVIP 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. The consideration of multiple variants during the effect prediction could reveal novel targets in GWAS-like approaches. The availability through a web server enables a large community of scientists without computational skills to benefit from NAVIP.

The remaining challenge is now the reliable detection of sequence variants prior to the application of NAVIP. A range of tools is available for the mapping of short reads and the following identification of sequence variants [26]. There is also rapid progress in the development of long read mapping tools [51,52] and the subsequent variant identification [53–56]. For heterozygous and polyploid species, phasing of these variants is another task that needs to be addressed in the future. Variant callers could directly report multiple SNVs of one haplophase as one MNV by collapsing the individual variants. In contrast to variant callers, variant annotators do not have access to the aligned reads and cannot infer this information.

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. Previous studies demonstrated that sequence variants might only affect individual isoforms in a negative way [50]. NAVIP analyzes all annotated transcript isoforms and would be able to discover such cases. Currently, a major limitation is the lack of isoform-resolved annotation for non-model plant species. Given the rapid progress in long read sequencing [14,57,58], it is likely that highly accurate structural annotation will become available for most plant species in the next few years.

Materials and Methods

Implementation of the Neighborhood-Aware Variant Impact Predictor (NAVIP)

The Neighborhood-Aware Variant Impact Predictor (NAVIP) (https://github.com/bpucker/NAVIP) has been implemented in Python3. NAVIP requires a VCF file containing sequence variants, a FASTA file containing the reference sequence, and a GFF3 file containing the structural annotation (gene models) as input. The variants provided must be homozygous or in a phased state to allow an accurate impact prediction per allele. If no information about the phasing is provided, all variants are assumed to be in the same haplophase. Effects on all annotated transcripts are evaluated per gene by taking into account the presence of all given variants simultaneously. NAVIP consists of three modules: VCF preprocessing, the NAVIP main program, and a simple first analysis (SFA) of the generated annotation. The first module is designed to preprocess VCF files line-by-line to check for multiallelic variants, i.e. variants with more than one alternative allele at a given position, split them into two separate entries, and convert them into one of three categories: substitution, insertion, or deletion. This process is crucial, as it allows for a clearer representation, facilitating further analysis and interpretation. The preprocessing also removes conflicting data entries and logs warnings and potential errors, such as identical bases, to ensure that any encountered discrepancies are documented for review. The second module is designed to validate genetic variants against transcript sequences, with a particular focus on insertions and deletions, to ensure that the variants align correctly with the reference and match the corresponding sequences in the transcript. NAVIP generates a new VCF file with an additional annotation field and additional report files. One annotation string in the VCF output file matches the SnpEff result format, but also has a NAVIP-specific string with additional information (see the manual for details: https://github.com/bpucker/NAVIP/wiki). NAVIP also produces FASTA files with sequences harboring all variants. NAVIP enhances the VCF files by incorporating additional information about the variants, including their effects on coding sequences (CDS), codon changes, and amino acid alterations. This allows users to identify variants with a potential impact on protein function, providing researchers with deeper insight into the effects of genetic variation. Frameshift mutations can occur when the number of nucleotides inserted or deleted is not a multiple of three, altering the downstream amino acid sequence. The third module serves as a primary interface for identifying compensating insertions and deletions (cInDels) within a given VCF file, categorizing them based on their effect on the reading frame, and generating output files summarizing the findings. It also includes functionality to visualize the number of InDels across transcripts through bar plots, facilitating interpretation of the results. The automatic assessment of complementing InDels reveals the relevance of simultaneously considering all InDels within a coding sequence when predicting their impact. All NAVIP scripts can be downloaded from the above-mentioned GitHub repository and do not require the installation of any dependencies other than the Python packages. NAVIP is also available through a web server (https://pbb-tools.de/NAVIP) free of charge. Files are kept confidential and will be deleted 48 h after offering the results for download.

Identification and validation of sequence variants

Illumina sequencing reads of *A. thaliana* Nd-1 [17] were mapped to the *A. thaliana* Col-0 reference genome sequence (TAIR10) [59] via BWA MEM v.0.7.13 [24] using the –m option to avoid spurious hits. Variant calling was performed via GATK v3.8 [60] based on the developers' recommendation. This combination of BWA MEM and GATK was previously identified as a reliable approach for this particular data set [26]. All processes were wrapped into 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 from the previous round as the gold standard to avoid the need for hard filtering.

Since a high-quality genome sequence assembly of Nd-1 was previously generated [18], we harnessed this sequence to validate all variants identified by short-read mapping. From the start of each chromosome sequence, variants sorted by genomic position were successively tested 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 (S7 File). Variants were admitted to the following analysis if the assembly supported them. This consecutive inspection of all variants enabled a reliable removal of false positives, leading to a set of high-confidence variants. The genome-wide distribution of the sequence variants was assessed using a previously developed Python script [17].

An independent confirmation of randomly selected sequence variants was performed using Sanger sequencing. *A. thaliana* Nd-1 plants were grown as previously described [17] to extract

DNA from leaf tissue using a cetyltrimethylammonium bromide (CTAB)-based method [61]. Oligonucleotides flanking the regions that harbor the variants of interest were designed manually (S8 File). Amplification via PCR, analysis of PCR products via agarose gel electrophoresis, purification of PCR products, Sanger sequencing, and evaluation of results were following previously established protocols [62].

Comparison of NAVIP and SnpEff stop gain prediction

To the best of our knowledge, SnpEff [36] is the most widely used tool for predicting the effects of sequence variants, thus it was selected for comparison. NAVIP can only provide more accurate effect predictions if multiple sequence variants interfere, e.g. if multiple SNVs are located within the same codon. Otherwise, the predictions of NAVIP and SnpEff would be the same. Consequently, the following comparison focuses only on cases of multiple sequence variants that might interfere with each other.

SnpEff v4.1f [36] was applied with default parameters to the *A. thaliana* Nd-1 variant data set to predict the effects of SNVs based on the Araport11 [63] structural annotation of the TAIR10 genome sequence of *A. thaliana* Columbia-0. NAVIP was also applied to the same data set for benchmarking. Predictions of premature stop codons were compared between NAVIP and SnpEff results, as these cases have the potential to show biologically important differences. This analysis was performed exclusively on SNVs to avoid the influence of frameshifts that would be caused by InDels. Only the most upstream predicted premature stop codon within any gene was considered in this analysis. To support the loss of function of the affected genes, the frequency of amino acid changing variants (aa_N) was compared to the number of variants that did not alter the encoded amino acid (aa_S). This ratio was compared between genes with premature stop codons and all other genes, expecting a higher ratio of variants that change the encoded amino acids if the gene undergoes pseudogenization. The Python package plotly was used to visualize these data distributions in violin plots. A pseudocount was added to both aa_N and aa_S to enable the ratio calculation in case when aa_S would be 0. aa_N/aa_S ratios greater than 10 were set to this maximum value to enable visualization. A Mann-Withney U test was performed using Python to test for significant differences between the two groups. When genes with a premature stop codon undergo pseudogenization, they may show lower than average gene expression. Therefore, a comparison of the expression of genes with a premature stop codon against all other protein-coding genes was performed. A previously compiled count table based on all publicly available paired-end RNA-seq data sets of *A. thaliana* [64] was harnessed for this analysis. Differences were visualized using the Python package plotly as described above, with the expression values clipped at 50 to enable an informative visualization. All Python scripts developed for these analyses are freely available on GitHub (https://github.com/bpucker/variant_calling).

Assessment of compensating InDels (cInDels)

An independent analysis of insertions/deletions (InDels) was performed by NAVIP to understand the relevance of considering all InDels within a CDS simultaneously. Transcripts with predicted frameshifts were analyzed to identify downstream insertions/deletions which are compensating each other's effect, i.e. the second frameshift is reverting an upstream frameshift. The distance between these events was analyzed by NAVIP and is included in the standard output. This analysis is not restricted to pairs of cInDels, but can also handle multiple InDels compensating each other's frameshifts.

Availability and requirements

Project name: NAVIP

Project homepage: https://github.com/bpucker/NAVIP

Operating system(s): Linux (website is platform independent)

Programming language: Python3

Other requirements: Python3

License: GNU General Public License v3.0

RRID: SCR_024838

Data availability

The data sets supporting the results of this article are publicly available or 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.10613052) and https://github.com/bpucker/variant_calling (https://doi.org/10.5281/zenodo.10613055).

Declarations

Authors' contributions

BP designed the research. JSB wrote the NAVIP code. AR updated the NAVIP code and added NAVIP to the pbb-tools web server. JSB, DH, and BP conducted bioinformatics analyses. DH and BP performed the experimental validation. JSB, AR, and BP wrote the manuscript. All authors read and approved the final version of the manuscript and agreed to its submission.

Financial Disclosure Statement

The authors received no specific funding for this work.

Competing Interests

JSB, AR, and DH have no competing interests. BP is head of the technology transfer center Plant Genomics and Applied Bioinformatics at iTUBS. This does not alter our adherence to PLOS policies on sharing data and materials.

Acknowledgments

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. We thank the Center for Biotechnology (CeBiTec) at Bielefeld University for providing an environment to perform the computational analyses. Many thanks to the German network for bioinformatics infrastructure (de.NBI, grant 031A533A) and the Bioinformatics Resource Facility (BRF) at the Center for Biotechnology (CeBiTec) at Bielefeld University for providing an environment to perform the computational analyses. We acknowledge support by the Open Access Publication Funds of Technische Universität Braunschweig.

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: 481–491. doi:10.1016/j.cell.2016.05.063

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. doi:10.1038/s41467-017-00336-7

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: 170068. doi:10.3835/plantgenome2017.08.0068
4. Valliyodan B, Brown AV, Wang J, Patil G, Liu Y, Otyama PI, et al. Genetic variation among 481 diverse soybean accessions, inferred from genomic re-sequencing. Sci Data. 2021;8: 50. doi:10.1038/s41597-021-00834-w
5. James GV, Patel V, Nordström KJV, Klasen JR, Salomé PA, Weigel D, et al. User guide for mapping-by-sequencing in Arabidopsis. Genome Biol. 2013;14: R61. doi:10.1186/gb-2013-14-6-r61
6. 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. doi:10.1186/gb-2014-15-6-r78
7. Schilbert HM, Pucker B, Ries D, Viehöver P, Micic Z, Dreyer F, et al. Mapping-by-Sequencing Reveals Genomic Regions Associated with Seed Quality Parameters in Brassica napus. Genes. 2022;13: 1131. doi:10.3390/genes13071131
8. Sielemann K, Pucker B, Orsini E, Elashry A, Schulte L, Viehöver P, et al. Genomic characterization of a nematode tolerance locus in sugar beet. BMC Genomics. 2023;24: 748. doi:10.1186/s12864-023-09823-2
9. 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: 546–549. doi:10.1038/nature12817
10. Jarvis DE, Ho YS, Lightfoot DJ, Schmöckel SM, Li B, Borm TJA, et al. The genome of Chenopodium quinoa. Nature. 2017;542: 307–312. doi:10.1038/nature21370
11. Lightfoot DJ, Jarvis DE, Ramaraj T, Lee R, Jellen EN, Maughan PJ. Single-molecule sequencing and Hi-C-based proximity-guided assembly of amaranth (Amaranthus hypochondriacus) chromosomes provide insights into genome evolution. BMC Biol. 2017;15: 74. doi:10.1186/s12915-017-0412-4
12. Siadjeu C, Pucker B, Viehöver P, Albach DC, Weisshaar B. High Contiguity de novo Genome Sequence Assembly of Trifoliate Yam (Dioscorea dumetorum) Using Long Read Sequencing. Genes. 2020;11: 274. doi:10.3390/genes11030274
13. Marks RA, Hotaling S, Frandsen PB, VanBuren R. Representation and participation across 20 years of plant genome sequencing. Nat Plants. 2021;7: 1571–1578. doi:10.1038/s41477-021-01031-8
14. Pucker B, Irisarri I, Vries J de, Xu B. Plant genome sequence assembly in the era of long reads: Progress, challenges and future directions. Quant Plant Biol. 2022;3: e5. doi:10.1017/qpb.2021.18
15. Stein LD. The case for cloud computing in genome informatics. Genome Biol. 2010;11: 207. doi:10.1186/gb-2010-11-5-207
16. Christensen KD, Dukhovny D, Siebert U, Green RC. Assessing the Costs and Cost-Effectiveness of Genomic Sequencing. J Pers Med. 2015;5: 470–486. doi:10.3390/jpm5040470
17. Pucker B, Holtgräwe D, Sörensen TR, Stracke R, Viehöver P, Weisshaar B. A De Novo Genome Sequence Assembly of the Arabidopsis thaliana Accession Niederzenz-1 Displays Presence/Absence Variation and Strong Synteny. PLOS ONE. 2016;11: e0164321. doi:10.1371/journal.pone.0164321
18. Pucker B, Holtgräwe D, Stadermann KB, Frey K, Huettel B, Reinhardt R, et al. A chromosome-level sequence assembly reveals the structure of the Arabidopsis thaliana Nd-1 genome and its gene set. PLOS ONE. 2019;14: e0216233.

doi:10.1371/journal.pone.0216233

19. 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–E4060. doi:10.1073/pnas.1607532113

20. Fan X, Chaisson M, Nakhleh L, Chen K. HySA: a Hybrid Structural variant Assembly approach using next-generation and single-molecule sequencing technologies. Genome Res. 2017;27: 793–800. doi:10.1101/gr.214767.116

21. Michael TP, Jupe F, Bemm F, Motley ST, Sandoval JP, Lanz C, et al. High contiguity Arabidopsis thaliana genome assembly with a single nanopore flow cell. Nat Commun. 2018;9: 541. doi:10.1038/s41467-018-03016-2

22. Wala JA, Bandopadhayay P, Greenwald NF, O'Rourke R, Sharpe T, Stewart C, et al. SvABA: genome-wide detection of structural variants and indels by local assembly. Genome Res. 2018;28: 581–591. doi:10.1101/gr.221028.117

23. Zhou Y, Massonnet M, Sanjak JS, Cantu D, Gaut BS. Evolutionary genomics of grape (Vitis vinifera ssp. vinifera) domestication. Proc Natl Acad Sci U S A. 2017;114: 11715–11720. doi:10.1073/pnas.1709257114

24. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv13033997 Q-Bio. 2013 [cited 20 Oct 2020]. Available: http://arxiv.org/abs/1303.3997

25. Van der Auwera G, O'Connor B. Genomics in the Cloud: Using Docker, GATK, and WDL in Terra. 2020 [cited 24 Jan 2024]. Available: https://www.oreilly.com/library/view/genomics-in-the/9781491975183/

26. Schilbert HM, Rempel A, Pucker B. Comparison of Read Mapping and Variant Calling Tools for the Analysis of Plant NGS Data. Plants. 2020;9: 439. doi:10.3390/plants9040439

27. Yao Z, You FM, N'Diaye A, Knox RE, McCartney C, Hiebert CW, et al. Evaluation of variant calling tools for large plant genome re-sequencing. BMC Bioinformatics. 2020;21: 360. doi:10.1186/s12859-020-03704-1

28. De Coster W, Weissensteiner MH, Sedlazeck FJ. Towards population-scale long-read sequencing. Nat Rev Genet. 2021;22: 572–587. doi:10.1038/s41576-021-00367-3

29. Olson ND, Wagner J, McDaniel J, Stephens SH, Westreich ST, Prasanna AG, et al. PrecisionFDA Truth Challenge V2: Calling variants from short and long reads in difficult-to-map regions. Cell Genomics. 2022;2: 100129. doi:10.1016/j.xgen.2022.100129

30. Li H. New strategies to improve minimap2 alignment accuracy. Bioinformatics. 2021;37: 4572–4574. doi:10.1093/bioinformatics/btab705

31. Edge P, Bansal V. Longshot enables accurate variant calling in diploid genomes from single-molecule long read sequencing. Nat Commun. 2019;10: 4660. doi:10.1038/s41467-019-12493-y

32. Heller D, Vingron M. SVIM-asm: structural variant detection from haploid and diploid genome assemblies. Bioinformatics. 2021;36: 5519–5521. doi:10.1093/bioinformatics/btaa1034

33. Smolka M, Paulin LF, Grochowski CM, Horner DW, Mahmoud M, Behera S, et al. Detection of mosaic and population-level structural variants with Sniffles2. Nat Biotechnol. 2024; 1–10. doi:10.1038/s41587-023-02024-y

34. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38: e164. doi:10.1093/nar/gkq603

35. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016;17: 122. doi:10.1186/s13059-016-0974-4

36. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly (Austin). 2012;6: 80–92. doi:10.4161/fly.19695

37. Hou L, Zhao H. A review of post-GWAS prioritization approaches. Front Genet. 2013;4: 280. doi:10.3389/fgene.2013.00280
38. Ries D, Holtgräwe D, Viehöver P, Weisshaar B. Rapid gene identification in sugar beet using deep sequencing of DNA from phenotypic pools selected from breeding panels. BMC Genomics. 2016;17: 236. doi:10.1186/s12864-016-2566-9
39. 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: 10249–10254. doi:10.1073/pnas.1107739108
40. 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: 956–963. doi:10.1038/ng.911
41. 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: 419–423. doi:10.1038/nature10414
42. 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: 235. doi:10.3389/fgene.2015.00235
43. Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003;31: 3812–3814. doi:10.1093/nar/gkg509
44. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7: 248–249. doi:10.1038/nmeth0410-248
45. Hecht M, Bromberg Y, Rost B. Better prediction of functional effects for sequence variants. BMC Genomics. 2015;16: S1. doi:10.1186/1471-2164-16-S8-S1
46. Holcomb D, Hamasaki-Katagiri N, Laurie K, Katneni U, Kames J, Alexaki A, et al. New approaches to predict the effect of co-occurring variants on protein characteristics. Am J Hum Genet. 2021;108: 1502–1511. doi:10.1016/j.ajhg.2021.06.011
47. Liu Y, Yeung WSB, Chiu PCN, Cao D. Computational approaches for predicting variant impact: An overview from resources, principles to applications. Front Genet. 2022;13. Available: https://www.frontiersin.org/articles/10.3389/fgene.2022.981005
48. Wang D, Li J, Wang Y, Wang E. A comparison on predicting functional impact of genomic variants. NAR Genomics Bioinforma. 2022;4: lqab122. doi:10.1093/nargab/lqab122
49. Katsonis P, Wilhelm K, Williams A, Lichtarge O. Genome interpretation using in silico predictors of variant impact. Hum Genet. 2022;141: 1549–1577. doi:10.1007/s00439-022-02457-6
50. Brandes N, Goldman G, Wang CH, Ye CJ, Ntranos V. Genome-wide prediction of disease variant effects with a deep protein language model. Nat Genet. 2023;55: 1512–1522. doi:10.1038/s41588-023-01465-0
51. Amarasinghe SL, Ritchie ME, Gouil Q. long-read-tools.org: an interactive catalogue of analysis methods for long-read sequencing data. GigaScience. 2021;10: giab003. doi:10.1093/gigascience/giab003
52. Sahlin K, Baudeau T, Cazaux B, Marchet C. A survey of mapping algorithms in the long-reads era. Genome Biol. 2023;24: 133. doi:10.1186/s13059-023-02972-3
53. Ahsan MU, Liu Q, Fang L, Wang K. NanoCaller for accurate detection of SNPs and indels in difficult-to-map regions from long-read sequencing by haplotype-aware deep neural networks. Genome Biol. 2021;22: 261. doi:10.1186/s13059-021-02472-2
54. Shafin K, Pesout T, Chang P-C, Nattestad M, Kolesnikov A, Goel S, et al. Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads. Nat Methods. 2021;18: 1322–1332. doi:10.1038/s41592-021-01299-w
55. Cleal K, Baird DM. Dysgu: efficient structural variant calling using short or long reads.

Nucleic Acids Res. 2022;50: e53. doi:10.1093/nar/gkac039

56. Huang N, Xu M, Nie F, Ni P, Xiao C-L, Luo F, et al. NanoSNP: a progressive and haplotype-aware SNP caller on low-coverage nanopore sequencing data. Bioinformatics. 2023;39: btac824. doi:10.1093/bioinformatics/btac824

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

58. Al-Dossary O, Furtado A, KharabianMasouleh A, Alsubaie B, Al-Mssallem I, Henry RJ. Long read sequencing to reveal the full complexity of a plant transcriptome by targeting both standard and long workflows. Plant Methods. 2023;19: 112. doi:10.1186/s13007-023-01091-1

59. Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, et al. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res. 2012;40: D1202–D1210. doi:10.1093/nar/gkr1090

60. 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 Bioinforma Ed Board Andreas Baxevanis Al. 2013;11: 11.10.1-11.10.33. doi:10.1002/0471250953.bi1110s43

61. Rosso MG, Li Y, Strizhov N, Reiss B, Dekker K, Weisshaar B. An Arabidopsis thaliana T-DNA mutagenized population (GABI-Kat) for flanking sequence tag-based reverse genetics. Plant Mol Biol. 2003;53: 247–259. doi:10.1023/B:PLAN.0000009297.37235.4a

62. 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: 667. doi:10.1186/s13104-017-2985-y

63. Cheng C-Y, Krishnakumar V, Chan AP, Thibaud-Nissen F, Schobel S, Town CD. Araport11: a complete reannotation of the Arabidopsis thaliana reference genome. Plant J. 2017;89: 789–804. doi:10.1111/tpj.13415

64. Choudhary N, Pucker B. Conserved amino acid residues and gene expression patterns associated with the substrate preferences of the competing enzymes FLS and DFR. PLOS ONE. 2024;19: e0305837. doi:10.1371/journal.pone.0305837

Supporting Information

S1 File: Detailed description of the variant calling process, the validation process, and the resulting sequence variant data set.

S2 File: VCF file containing SNVs between Nd-1 and Col-0.

S3 File: Detailed information about premature stop codons predicted by NAVIP and/or SnpEff.

S4 File: Differences in the effect prediction between SnpEff and NAVIP for 200 accessions of the 1,135 *Arabidopsis thaliana* accession resequencing project.

S5 File: Comparison of SnpEff and NAVIP prediction differences between Col-0 and Nd-1. The table lists matches and differences for each possible amino acid substitution type.

S6 File: VCF file containing InDels between Nd-1 and Col-0.

S7 File: Schematic illustration of the variant validation process.

S8 File: FASTA file containing oligonucleotide sequences used for the generation and sequencing of amplicons to validate randomly selected sequence variants.