Published online 31 January 2023 Nucleic Acids Research, 2023, Vol. 51, No. 7 e37
https://doi.org/10.1093/nar/gkad036

SMAP design: a multiplex PCR amplicon and gRNA design tool to screen for natural and CRISPR-induced genetic variation

Ward Develtere[1,2], Evelien Waegneer[3,4], Kevin Debray[1,2,3], Jonas De Saeger[1,2], Sabine Van Glabeke[3], Steven Maere[1,2], Tom Ruttink[1,3] and Thomas B. Jacobs[1,2,*]

[1]Department of Plant Biotechnology and Bioinformatics, Ghent University, (Technologiepark-Zwijnaarde 71) 9052, Ghent, Belgium, [2]VIB Center for Plant Systems Biology, (Technologiepark-Zwijnaarde 71), 9052, Ghent, Belgium, [3]ILVO, Flanders Research Institute for Agriculture, Fisheries and Food, Plant Sciences Unit, (Caritasstraat 39), 9090, Melle, Belgium and [4]Laboratory for Plant Genetics and Crop Improvement, Division of Crop Biotechnics, Department of Biosystems, Katholieke Universiteit Leuven, Leuven, Belgium

Received July 16, 2022; Revised December 14, 2022; Editorial Decision January 10, 2023; Accepted January 12, 2023

ABSTRACT

Multiplex amplicon sequencing is a versatile method to identify genetic variation in natural or mutagenized populations through eco-tilling or multiplex CRISPR screens. Such genotyping screens require reliable and specific primer designs, combined with simultaneous gRNA design for CRISPR screens. Unfortunately, current tools are unable to combine multiplex gRNA and primer design in a high-throughput and easy-to-use manner with high design flexibility. Here, we report the development of a bioinformatics tool called *SMAP design* to overcome these limitations. We tested *SMAP design* on several plant and non-plant genomes and obtained designs for more than 80–90% of the target genes, depending on the genome and gene family. We validated the designs with Illumina multiplex amplicon sequencing and Sanger sequencing in Arabidopsis, soybean, and maize. We also used *SMAP design* to perform eco-tilling by tilling PCR amplicons across nine candidate genes putatively associated with haploid induction in *Cichorium intybus*. We screened 60 accessions of chicory and witloof and identified thirteen knockout haplotypes and their carriers. *SMAP design* is an easy-to-use command-line tool that generates highly specific gRNA and/or primer designs for any number of loci for CRISPR or natural variation screens and is compatible with other SMAP modules for seamless downstream analysis.

INTRODUCTION

The detection of genetic variation is of great value to medicine and agriculture as it allows researchers to uncover molecular mechanisms, study genetic pathways, and assign gene function. In the frame of breeding, genetic variation can be discovered by screening for beneficial alleles in natural accessions or gene pools. Sequence-based allele mining, also called eco-tilling (1), is a method to screen for naturally occurring mutations in a set of genomic regions (i.e. candidate genes) across a broad collection of genotypes. A versatile and cost-efficient method for such targeted sequencing is highly multiplex (HiPlex) amplicon sequencing, in which multiple target regions (tens to thousands) are amplified in a single PCR reaction and all amplicons are sequenced via Illumina sequencing. Adding sample-specific indices during library preparation allows the pooling of up to hundreds or thousands of samples in a single sequencing run. There are ample examples of studies using this technique in several crops including barley, rice, soy, and wheat (reviewed in (2)).

Carriers of rare, defective alleles often display useful phenotypes (3,4) making their identification important for crop breeding and fundamental research. However, population genetics theory predicts that defective alleles may be maintained at a low frequency in natural populations as they negatively affect plant fitness and are subject to negative selection (5). When such carriers are difficult to find in natural populations, genetic variation can be induced with random mutagens like ethyl methanesulfonate or in a targeted fashion with genome editing technologies like CRISPR. CRISPR has become the staple genome editing tool due to its efficacy and simple design. In its most basic form, a CRISPR-associated endonuclease (Cas) is directed to a target site via a guide RNA (gRNA), where it creates a double-

*To whom correspondence should be addressed. Tel: +32 9 331 38 60; Email: thomas.jacobs@psb.ugent.be

© The Author(s) 2023. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/nar/article/51/7/e37/7016455 by guest on 02 May 2023

stranded DNA break (6,7). In most eukaryotes, imperfect repair typically results in insertions and/or deletions (indels), or infrequently substitutions, around the breakpoint (8,9). In principle, CRISPR can be used to knock out any protein-coding gene by disrupting the reading frame or regulatory regions. Multiplex CRISPR screens go a step further by simultaneously targeting multiple genomic loci with arrays of gRNAs to produce large collections of individuals with unique combinations of induced DNA modifications (10). Therefore, eco-tilling and CRISPR are highly complementary mutation screening approaches.

An essential component of any mutant screen is the genotyping assay to identify the underlying sequence variant. Specific design parameters need to be considered for each assay's respective purpose and constraints. For eco-tilling, complete coverage of the regulatory and coding regions is preferred as it allows one to identify all existing mutations in the set of candidate loci across the gene pool and thereby identify conserved and variable genic regions and carriers of defective alleles. While primer design is typically based on a single representative reference genome sequence, eco-tilling may be subject to amplicon dropout in highly divergent regions due to primer-template mismatches. In addition, HiPlex amplicons cannot overlap within a single multiplex reaction as the smaller amplicons, formed through cross-amplification (Supplementary Figure S1), would dominate the PCR and reduce coverage. Therefore, in multiplex PCR applications, multiple primer mixtures need to be designed to specifically amplify complementary (partially overlapping) regions across the candidate loci in separate reactions.

In contrast, most CRISPR screens do not require complete coverage of the candidate gene sequence. Instead, the design focuses on specific and efficient gRNAs, using prior knowledge of essential regions of the CDS and/or regulatory sequences, and covers those target regions with few amplicons with high primer-binding specificity so that mutations can be evaluated. Furthermore, many CRISPR screens only sequence the gRNA(s) present in the population as a proxy for the mutant genotype and do not reveal the true underlying genotype at the target loci. Therefore, high-throughput genotyping assays that can specifically cover the large number of target loci in a CRISPR screen are desirable. Genotyping assays are relatively simple to design manually for a handful of targets, but it can take months for a combinatorial, multiplex CRISPR screen with hundreds of targets. Designing amplicons for gene families is particularly challenging due to sequence similarity between paralogous genes and the chance of cross-amplification or off-target amplification (here collectively called mispriming; Supplementary Figure S1). In addition, specific design parameters such as amplicon size range need to be adjusted depending on the downstream library preparation and sequencing technology (e.g. paired-end Illumina short reads or Sanger sequencing). While there are several online and command-line tools available for gRNA design (CRISPOR, CHOPCHOP, FlashFry, CRISPRscan, CCTop) and primer design (PrimerMapper, PrimerView, Primer3), none are integrated with genotyping assay design in a high-throughput manner with the flexibility and specificity required for large-scale multiplex experiments (11–18). Thus, combined gRNA and amplicon design is currently one of the limiting factors to perform medium to high-throughput multiplex CRISPR screens on tens to thousands of genes.

Here, we report the development of a bioinformatics tool called *SMAP design* that addresses these limitations and seamlessly fits into the larger SMAP (Stack Mapping Anchor Points) package that analyzes naturally occurring and CRISPR-induced sequence variants (19). *SMAP design* uses Primer3 (18) to create sets of amplicons with localized or global coverage across reference sequences. For CRISPR experiments, amplicon coordinates are intersected with gRNA target sites from algorithms such as CRISPOR (13) or FlashFry (12), based on user-defined positional boundaries. Sets of amplicons/gRNAs can be created within minutes for tens of loci, or up to a few hours for more complex designs with a few hundred genes. We performed *in silico* designs on 80–95 gene families of varying sizes and in different species and implemented several options to increase coverage. We further empirically validated the designs by PCR amplification and Sanger sequencing or HiPlex sequencing on reference or CRISPR materials (Arabidopsis, soybean and maize). We performed eco-tilling in natural accessions of chicory and witloof (*Cichorium intybus* var. *sativum* and *C. intybus* var. *foliosum*) using primer sets made by *SMAP design* targeting nine candidate genes putatively involved in haploid induction and demonstrate strategies to enhance the mutation screening capacity by combining multiplexing and sample pooling.

MATERIALS AND METHODS

SMAP design

SMAP design is a command-line tool written in Python3 and is an addition to the SMAP package (19). The program uses the Primer3-py (https://pypi.org/project/primer3-py, version 0.6.1 or newer), Biopython (https://biopython.org, version 1.77 or newer), Pandas (https://pandas.pydata.org, version 1.1.5 or newer), Numpy (https://numpy.org, version 1.18.5 or newer), and Matplotlib (https://matplotlib.org, version 3.3.3 or newer) packages. All *SMAP design* runs were performed on a computing cluster with Intel Xeon CPUs on one core. *SMAP design* and its source code are available at https://gitlab.com/ilvo/smap-design, and a detailed user manual and guidelines are available at https://ngs-smap.readthedocs.io/en/latest/design/, under the GNU Affero General Public License v3.0. *SMAP design* is also available online as a graphical user interface at usegalaxy.be.

Genome source and extraction of candidate genes. All genome and annotation files were retrieved from PLAZA dicot 4.5, PLAZA monocot 4.5, or PLAZA pico 3.0 (Supplementary Table S1; (20,21)). A novel, in-house assembled and annotated reference genome sequence of *C. intybus* var. *sativum* (unpublished) was used for genome-wide identification of gene family members from the selected gene families. Input files for *SMAP design* were generated with *SMAP target-selection. SMAP target-selection* is a command-line tool written in Python3 that extracts the genomic sequence of candidate genes from the reference genome (optionally with extra upstream or downstream flanking regions),

Downloaded from https://academic.oup.com/nar/article/51/7/e37/7016455 by guest on 02 May 2023

based on a user-provided list of gene IDs (optionally combined with grouping based on gene family, genetic network, or genetic pathway membership; here, PLAZA homology groups were used), and orients the sequences in the reference sequence FASTA file so that the CDS is encoded on the positive strand. A GFF file that contains the relative coordinates of the annotated features (gene, CDS, exon, optionally critical domains) of the extracted genes is also generated for further downstream analysis with *SMAP design* and the other modules in the SMAP package (19).

gRNA design with FlashFry and CRISPOR. FlashFry (12) was used to design gRNAs for all the analyses except for the genome-wide design of *Physcomitrium patens*, for which gRNAs were designed by CRISPOR (13). Both programs were run with the default settings except for the mismatch parameter of CRISPOR, which was set to 3 (*–mm*). SpCas9 with NGG PAM sequence was used for all designs.

SMAP design parameter settings. Primer3 default settings (18) were used with the exception of the user-defined settings in *SMAP design* (Supplementary Table S2) and the minimum distance between adjacent forward and reverse primers (set to 5 bp). Four general designs were performed for this report: (i) $\text{Design}_{\text{HiPlex}}$, the amplicon size was set to 120–150 bp and a distance of 15 bp between the gRNAs and primers to be compatible with the HiPlex sequencing service of Floodlight Genomics LLC. (ii) $\text{Design}_{\text{PE}}$, the amplicon size was set to 220–250 bp and a 15 bp distance between the gRNAs and primers to be compatible with paired-end 150 bp Illumina sequencing. (iii) $\text{Design}_{\text{Sanger}}$, the amplicon size was set to 400–800 bp and a 150 bp distance between the gRNAs and primers as these are the preferred settings for ICE (22) or TIDE (23) analysis of Sanger sequences. (iv) $\text{Design}_{\text{NatVar}}$, the amplicon size was set to 120–150 bp and the maximum number of amplicons for each of the candidate genes was requested. After one design round with $\text{Design}_{\text{NatVar}}$, the selected primer binding sites were encoded as 'N' sequences in the reference sequence input file to exclude those regions from Primer3 design in a second iteration with the same settings. This provides a strategy to iteratively create two or more complementary HiPlex assays with partially overlapping (i.e. tiled) amplicons that together increase coverage of the reference sequence when performed in parallel.

Amplicon selection and gRNA array cloning for protoplast transfection. We used *SMAP design* to create 117 gRNA and amplicon designs on 67 growth-related maize genes (24). The amplicons were first validated on 20 wild-type maize (B104) leaf samples using HiPlex sequencing. The top 50 amplicons, i.e. amplicons with reads perfectly matching the reference and highest read depth, were selected as Cas9 targets (Supplemental Figure S2). The 50 Golden Gate gRNA entry vectors and 10 expression vectors were constructed as previously described (25). Briefly, oligos were annealed in 48 μl of Milli-Q water followed by an incubation in the thermal cycler with the following program: 5 min at 95°C; 95°C to 85°C at –2°C/s; 85°C to 25°C at –0.1°C/s. The annealed oligos were inserted into gRNA entry vectors with a Golden Gate reaction ((37°C, 5 min; 16°C, 5 min) × 30; 50°C, 5 min; 80°C, 5 min) using the BbsI-HF restriction enzyme (New England Biolabs) and T4 DNA ligase (New England Biolabs) (25). Each gRNA entry vector contains the scaffold and a TaU3 promoter to express the gRNA. The Golden Gate reactions were transformed via heat shock into *ccd*B-sensitive DH5α *Escherichia coli* cells and plated on lysogeny broth + carbenicillin (100 μg/ml.). The entry vectors were isolated (GeneJET Plasmid Miniprep kit, Thermo Fisher Scientific) and validated with Sanger sequencing (Mix2Seq, Eurofins Scientific). Expression vectors were made by combining five gRNA entry vectors and a linker vector pGG-F-linkerII-G (25) into the destination vector pGG-AG-KmR (26) using a Golden Gate reaction with the same parameters as described above, except with BsaI-HFv2 (New England Biolabs) instead of BbsI-HF. Expression vectors were validated using a restriction digest with NheI and SacI (Promega) and whole-plasmid sequencing (Plasmidsaurus). Expression vectors were purified using the plasmid Midiprep kit (Zymo Research) for protoplast transfection. A description of the gRNA arrays and the HiPlex primers used to amplify all target sites can be found in Supplementary Table S3.

Plant material and DNA extraction. DNA was extracted from *Arabidopsis thaliana* Colombia-0 and maize (*Zea mays*) B104 leaves according to Berendzen (27) with the following modifications: extraction buffer was added after the grinding step followed by incubation for 20 min at 60°C and centrifugation for 2 min at 1800 × *g*. After the DNA extraction of the maize leaves, 50 μl per sample was further purified using 50 μl magnetic beads (cleanNGS) according to the manufacturer's instructions. DNA was extracted from soybean (*Glycine max*) Williams 82 leaves and maize (*Zea mays*) B104 protoplasts as previously described (28), with the following modifications: an adapted extraction buffer was used (100 mM Tris–HCl (pH 8.0); 500 mM NaCl; 50 mM EDTA; 0.7% (w/v) SDS) and a 70% (v/v) ethanol washing step was included. For the maize protoplasts, the tissue grinding step was omitted, instead, the extraction buffer was added directly to the samples and the pellets were resuspended in 20 μl TE buffer. For chicory (*C. intybus*), tissue pools were created by sampling a leaf punch of each individual (up to 10 individuals per pool) in the same Eppendorf tube. Pooled leaf material was ground and homogenized and used for DNA extraction with a CTAB extraction protocol (29). Six leaf punches per individual were used for the individual samples. A detailed description of the chicory plant material, including accession names, can be found in Supplementary Table S4.

Protoplast isolation and transfection. Maize protoplast isolation and transfection was performed as previously described (30). Briefly, B104 maize seeds were grown in a growth chamber under long-day conditions (16 h light/8 h dark at 21°C) for 5 days and transferred to 24h dark conditions 8 days prior to protoplast isolation. Per gRNA array, 20 μg of plasmid DNA (12.5 μg of the Cas9 plasmid (35SP-mCherry-N7-NOST-ZmUBIP-AtCas9-Bpstar-G7T; vector ID 20_74; https://gatewayvectors.vib.be) and 7.5 μg of the gRNA array plasmid) was used for transfection of 100 K protoplasts in 100 μl. Each gRNA array was

Downloaded from https://academic.oup.com/nar/article/51/7/e37/7016455 by guest on 02 May 2023

Table 1. *SMAP haplotype-window* settings used to analyze the HiPlex data

	Min. read count	Min. haplotype frequency	Mask frequency	Discrete calls	Dosage filter	Frequency bounds
Arabidopsis	30	5	0	No	NA	NA
Soybean	30	5	0	No	NA	NA
Chicory ind.	30	4	1	Dosage	2	Diploid
Chicory pooled	30	4	1	No	NA	NA
Maize leaf	50	2	1	No	NA	NA
Maize protoplasts	30	1	1	No	NA	NA

Default settings were used to analyze the next-generation sequencing (HiPlex) data with *SMAP haplotype-window* except for the parameters shown here. NA, not applicable.

transfected in triplicate. A negative control consisting of a transfection with only the Cas9 plasmid was performed on six protoplast samples. To each sample, 110 μl of PEG solution (0.2 M mannitol, 100 mM $CaCl_2$, 40% PEG 4000) was added and the protoplasts were incubated in the dark for 15 min after which W5 solution was added to stop the transfection. The protoplasts were centrifuged at 100 × *g* for 2 min and the supernatant was discarded. The protoplasts were resuspended in 1 ml of W5 solution and incubated in 6-well plates in the dark at 25°C (20 rpm). Samples were harvested after two days of incubation and the DNA was extracted as described above.

DNA sequencing. For Sanger sequencing of Arabidopsis amplicons, PCR was performed with Red Taq DNA Polymerase Master Mix (VWR Life Science) according to the manufacturer's instructions and purified with magnetic beads (CleanNGS). The PCR amplified regions were Sanger sequenced via Mix2Seq (Eurofins Genomics).

Sets of HiPlex amplicons were designed for Arabidopsis (40 amplicons), soybean (40 amplicons), maize (117 amplicons), and chicory (two assays with 45 and 49 amplicons, respectively). Genomic DNA for each species was submitted for HiPlex sequencing (Floodlight Genomics LLC). For Arabidopsis and soybean, the sequencing was done on 24 biological replicates of the reference genotypes. For maize, 20 wild-type leaf samples and 36 transfected protoplast samples (10 gRNA arrays × 3 replicates plus 6 Cas9-only negative controls) were sequenced. Six technical replicates of the genotype L9001 (*C. intybus* var. *sativum*) were used as controls for the pooled sequencing run and two technical replicates of the L9001 reference genotype were included in the individual plant sequencing run. Details of the pooling strategy can be found in Supplementary Table S4.

Sequence data analysis. Sanger sequence analysis was performed with Geneious Prime 2022.0.1 (https://www.geneious.com). BWA-MEM 0.7.17 (31) was used for HiPlex read mapping with default parameters using the gene targets as the reference sequence. *SMAP haplotype-window* (19) was used for the analysis of the mapped HiPlex sequencing data with the default parameters except for the settings specified in Table 1. The analysis of the chicory haplotypes revealed that six primer pairs had off-target amplification of pseudogenes and/or genes with similar domains outside the gene family. This issue was resolved by including these (pseudo)genes into the reference sequence FASTA file and repeating the read mapping, thereby allowing the reads to map onto their correct reference sequence.

SMAP effect-prediction (19) was used to predict the effect of the mutations on the encoded protein of all non-reference haplotypes detected in the HiPlex sequencing data of chicory with default parameters. We defined the effect of haplotypes as 'mild effect' if > 50% of the resulting amino acid sequence was identical to the reference protein and as 'strong effect' if less than 50% of the amino acid sequence was identical to the reference protein (*–effect_threshold*). For the maize protoplast samples, *SMAP effect-prediction* was used to discern CRISPR-mediated indels from background and sequencing errors by calling variants only when they appear within 10 bp upstream (*–cut_site_range_upper_bound*) and 10 bp downstream (*–cut_site_range_lower_bound*) of the expected cut site (3 bp upstream of the PAM sequence). *SMAP effect-prediction* was run with the option *–disable_protein_prediction* turned on.

RESULTS

The *SMAP design* workflow

Input for SMAP design. *SMAP design* was created to easily and rapidly design sets of multiplex amplicon sequencing primers and gRNAs for small to large-scale CRISPR screens (Figure 1). Prior to running *SMAP design*, a FASTA file with sets of candidate gene reference sequences (e.g. entire gene families, gene networks, genetic pathways, or any other customized grouping) can be extracted from the reference genome using *SMAP target-selection*. *SMAP target-selection* orients all genes with the CDS on the positive strand for a consistent coordinate system and automatically generates a corresponding GFF file with the relative location of gene features (e.g. CDS or critical domains). The FASTA and GFF files are used as input for *SMAP design* and downstream analyses with other modules of the SMAP package (19). If the user wants amplicons to specifically cover one or more gRNAs, a list of gRNAs is provided as a tab-delimited file in the format as described in the online user manual (output gRNA files from CRISPOR and FlashFry can be directly fed to *SMAP design*).

Amplicon design. *SMAP design* uses the Primer3 module to design, by default, a maximum of 150 amplicons (300 primers) of a user-defined size range for each reference sequence. By default, the specificity of each primer is tested against the entire reference FASTA file to avoid misprim-

Downloaded from https://academic.oup.com/nar/article/51/7/e37/7016455 by guest on 02 May 2023

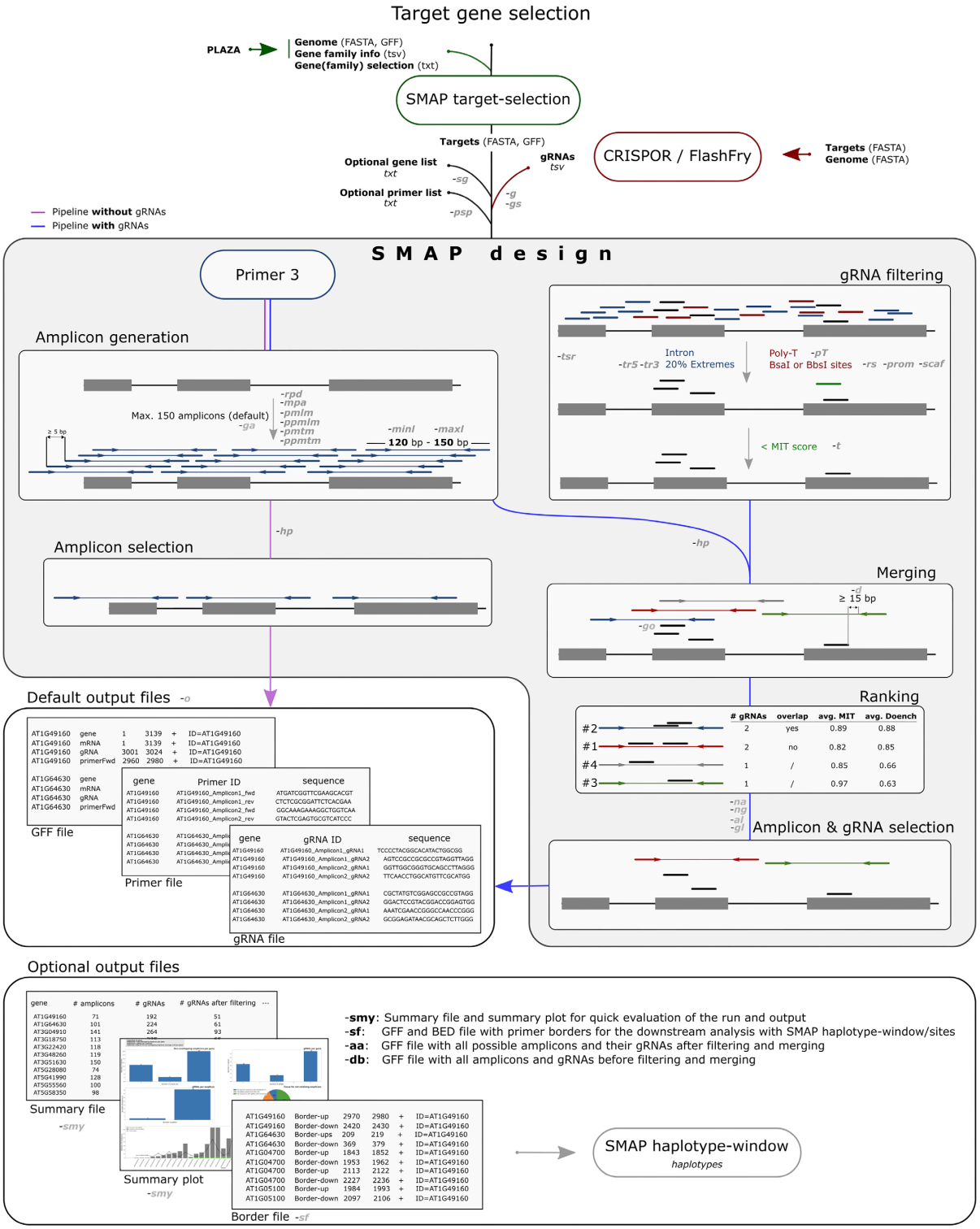

Figure 1. Workflow of *SMAP design*. Users select and extract a set of genes using *SMAP target-selection*. Input files for *SMAP target-selection* can be obtained through PLAZA (or other databases). The FASTA and GFF files are required inputs for *SMAP design*. If no gRNA file is given (purple workflow), *SMAP design* will select only amplicons that do not overlap using amplicons designed by Primer3. If a gRNA file is specified (blue workflow), which can be obtained from third-party software such as CRISPOR or FlashFry, *SMAP design* will filter the gRNAs based on their location in the gene, their sequence, and specificity score. Amplicons designed by Primer3 are merged with the filtered gRNAs and are subsequently ranked based on the gRNAs it overlaps with (number of gRNAs, overlap between gRNAs, and specificity and efficiency scores). Based on the ranking, a maximum (user-defined) number of non-overlapping amplicons per gene are selected. Two or three output files are created by default: a primer and gRNA file with the respective sequences per gene, and a GFF file specifying the location of the primers and gRNAs. Multiple optional files can be generated: a summary file and graph, two debug files, and a border file which is required as input for *SMAP haplotype-window/sites* (for downstream sequence analysis). -xx(x) indicates the abbreviated parameters in *SMAP design*.

Downloaded from https://academic.oup.com/nar/article/51/7/e37/7016455 by guest on 02 May 2023

ing (Supplementary Figure S1). Primer specificity thresholds can be adjusted or switched off. Alternatively, the user can provide a list of gene IDs to limit amplicon design to only that subset, while still using all reference sequences in the FASTA file for primer specificity testing. As Primer3 automatically avoids primer design at ambiguous nucleotides, known polymorphic positions such as SNPs can *a priori* be substituted by N-encoded nucleotides in the reference sequence FASTA to circumvent inefficient primer binding. Primers are spaced by a minimum of 5 bp to spread the amplicons across the target sequences. By default, amplicons with homopolymers (≥ 10 repeated nucleotides) are filtered out because downstream sequencing will likely yield low-quality reads. If no gRNAs are provided by the user (e.g. in case of screening for natural variation), *SMAP design* selects sets of non-overlapping amplicons to maximize reference sequence coverage.

gRNA filtering. *SMAP design* filters the provided gRNAs based on several criteria (Figure 1). By default, gRNAs with a poly-T stretch (≥ 4T, a Pol III termination signal) are removed. Short vector sequences directly flanking the gRNA sequence (i.e. promotor and scaffold) can be provided to simulate vector construction steps and exclude gRNAs with restriction sites (e.g. BsaI or BbsI) that interfere with cloning. To increase the likelihood of making knockout mutations, gRNA selection can be focused on selected domains defined via a particular feature type in the annotation GFF (e.g. kinase domain). The user can also exclude a segment of the 5' and 3' of the CDS to steer gRNA target sites to a part of the CDS. An optional minimum gRNA specificity score (e.g. MIT score (32)) threshold can also be applied.

Ranking and filtering amplicons and gRNAs. Filtered gRNAs are grouped to amplicons by positional overlap. By default, a gRNA is only grouped to an amplicon if the distance between the end of the primer and the gRNA binding site is at least 15 bp. Amplicons are ranked based on the gRNAs they cover according to the following criteria and order: (i) the number of gRNAs (an amplicon with multiple gRNAs will rank higher than an amplicon with a single gRNA); (ii) the positional overlap between gRNAs (amplicons with non-overlapping gRNAs will rank highest); (iii) the average gRNA specificity scores (e.g. MIT score (32)) and (iv) the average gRNA efficiency scores (such as the Doench (33) and out-of-frame scores (34)). If no specificity or efficiency scores are provided in the gRNA file, amplicons are only ranked by the first two criteria. Ultimately, *SMAP design* selects a (user-defined) maximum number of top-ranking, non-overlapping amplicons per gene, each covering a (user-defined) maximum number of gRNAs (Supplementary Table S2).

Output of SMAP design. *SMAP design* generates two files by default: a tab-separated values (TSV) file with the primer sequences sequentially numbered per gene and a GFF file with the primer locations on the target gene reference sequences (and other annotation features that were included in the GFF input file). If a gRNA list is provided, *SMAP design* also generates a TSV file with the selected gRNA sequences per gene (Figure 1). If no design was possible, the underlying reasons are included per gene at the end of the TSV files. Optionally, summary tables and graphs are generated for a quick evaluation of the set of amplicons and gRNAs (Supplementary Figure S3). These graphs show the distributions of the number of gRNAs and non-overlapping amplicons per gene that *SMAP design* generated and indicate the reasons for dropout per gene. For instance, the design may fail because no gRNAs were designed for that gene, none of the gRNAs passed all filters, Primer3 was not capable of designing specific amplicons for the gene, or there was no overlap between the gRNAs and the amplicons. Optionally, a GFF file is created with positions of border sequences required for downstream amplicon analysis by *SMAP haplotype-window* and a BED file required for *SMAP haplotype-sites* (19). In debug mode, an extra GFF output file containing all amplicons and gRNAs prior to filtering is given as a way to visualize the relative positions of all amplicons. After filtering for each gene, an optional GFF file can be generated with all amplicons and their respective gRNAs to visualize and manually select amplicons of interest with a sequence analysis viewer.

In silico testing of *SMAP design* in various species

We evaluated the performance of *SMAP design*, specifically to determine the relationship between successful design and gene family sizes, with the hypothesis that amplicon and/or gRNA design would be more difficult in larger gene families due to sequence homology. Therefore, we tested *SMAP design* on eleven different species representing a broad range of genome sizes and compositions (Arabidopsis, *P. patens*, rice, tomato, potato, maize, soybean, Chlamydomonas, *Saccharomyces cerevisiae*, mouse, and human). Per species, 80–95 gene families (PLAZA homology groups (20,21)) (Supplementary Table S5), containing between 1 and 448 genes per family were selected. Three different design settings ($\text{Design}_{\text{HiPlex}}$, $\text{Design}_{\text{PE}}$, and $\text{Design}_{\text{Sanger}}$, see Material and Methods) for different genotyping approaches were tested on the various genes (Figure 2). We considered at least one amplicon covering at least two gRNAs per gene as the minimum required for a knockout experiment and determined the fraction of genes per family that were 'retained' with these criteria for each design setting (here called 'retention rate').

Overall, the average retention rate per gene family for most tested species is ≥ 80% and there is a clear increase in retention rate with increasing amplicon size (Figure 2). The unicellular species Chlamydomonas and yeast had the highest average retention rates of ≥ 96%. Plant genomes with a lower fraction of recently-duplicated regions such as Arabidopsis, *P. patens*, and rice displayed average retention rates of 90% or higher for all three designs. The average retention rates for tomato, potato and maize ranged from 80% to 90%, with the exception of maize for $\text{Design}_{\text{HiPlex}}$ (65%). Soybean had the lowest average retention rate per gene family with 27%, 44%, 79% for $\text{Design}_{\text{HiPlex}}$, $\text{Design}_{\text{PE}}$ and $\text{Design}_{\text{Sanger}}$, respectively, likely due to the highly duplicated nature of its genome (35,36). These data indicate that genome constitution (e.g. recent genome duplication) affects the retention rate and that increasing the amplicon size can reduce design dropout.

Downloaded from https://academic.oup.com/nar/article/51/7/e37/7016455 by guest on 02 May 2023

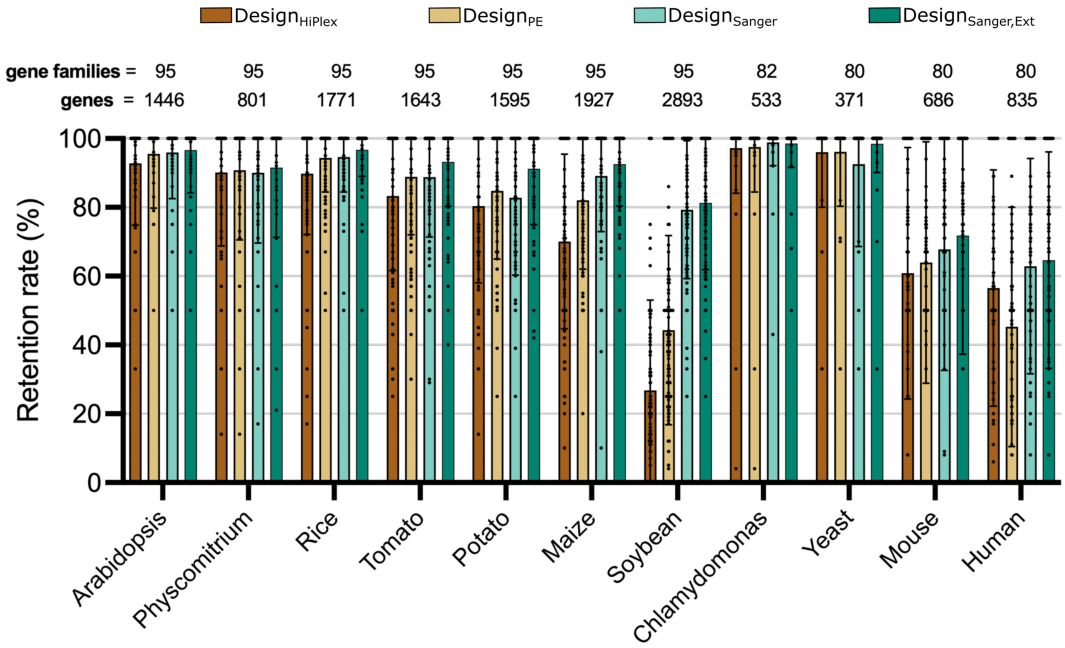

Figure 2. *SMAP Design* average retention rate across settings and species. Four Design settings were tested: Design$_{HiPlex}$ (120–150 bp amplicons), Design$_{PE}$ (220–250 bp), Design$_{Sanger}$ (400–800 bp) Design$_{Sanger,Ext}$ (400–800 bp with gene sequences extended by 500 bp at both ends). The same gene families were tested between settings and between the multicellular plant species and between the non-plant species, respectively. Retention rate is defined as the percentage of genes per gene family that contain at least one amplicon covering a minimum of two gRNAs. The bars show the average (with standard deviation) retention rate across all gene families per species.

During the analysis of Design$_{Sanger}$, we also observed that small genes preferentially dropped out since the gene size was less than the required amplicon length of 400–800 bp. To overcome this limitation, *SMAP target-selection* was set to extract flanking regions 500 bp upstream and downstream of the target genes, resulting in Design$_{Sanger,Ext}$. This increased the number of retained genes from 1399 to 1406 (out of 1446) genes for Arabidopsis (1% improvement), and from 1214 to 1338 (out of 1595) genes for potato (8% improvement; Figure 2). This further shows that primer and gRNA design optimization relies on both accurate selection of target reference sequences with *SMAP target-selection* as well as parameter settings of *SMAP design*.

The average retention rates were much lower for the mouse and human genomes (ranging from 45% to 68%; Figure 2). As Primer3 designs amplicons across the entire sequence (exons and introns) and mammalian genes contain relatively large introns, we reasoned that many of the 150 designed amplicons possibly fell into the introns and were thus filtered out at the merging step. To overcome this, the option *–restrictedPrimerDesign* (*-rpd*) was added to restrict Primer3 to design amplicons near exons (Supplementary Table S2). Running *SMAP design* with *-rpd* increased the average retention rate for Design$_{PE}$ by 10% for both species and the runtime was two to three times faster (Supplementary Figure S4). While ignoring intronic regions results in a clear improvement in retention rate, this result indicates there are additional sequence constraints limiting designs in these genomes and that parameter settings may need to be fine-tuned accordingly.

Empirical testing of *SMAP design* in various species

To validate the specificity and amplification efficiency of Design$_{HiPlex}$, we designed 40 amplicons on the *MAP3K* gene family in both Arabidopsis and soybean and performed HiPlex sequencing on 24 replicates of the reference genotypes for both species. For both Arabidopsis and soybean, all amplicons were sequenced in all replicates. Read depth was uniform across all amplicons with an average range < 13-fold for 39 of the 40 amplicons for both species (Figure 3). Three amplicons for Arabidopsis and one amplicon for soybean fell below an arbitrary threshold of 1,000 reads across all samples and would likely be removed when establishing the genotyping assay to ensure reliable coverage of all amplicons in a screen. In Arabidopsis, a non-reference haplotype with a 1-bp deletion was found at the AT2G35050 locus with an average relative read depth of 3.8% across all samples. This deletion occurred in a homopolymer of ten adenosines and is therefore likely a sequencing error. Only the reference haplotypes were found for all other loci. The AT3G50730 amplicon displayed the lowest average read depth and failed for the five samples with the lowest overall total reads per library. In soybean, three loci displayed non-reference haplotypes. Two haplotypes contained two mismatches (SNPs) compared to the reference and the third haplotype had a 3-bp deletion compared to the reference in a 'TCC' short sequence repeat. The average relative read depths of the non-reference haplotypes were consistently at or < 5% across all samples and can likely be attributed to low abundance PCR artifacts and/or sequencing errors. Such systematic errors can be removed with the haplotype frequency filters in *SMAP haplotype-*

Downloaded from https://academic.oup.com/nar/article/51/7/e37/7016455 by guest on 02 May 2023

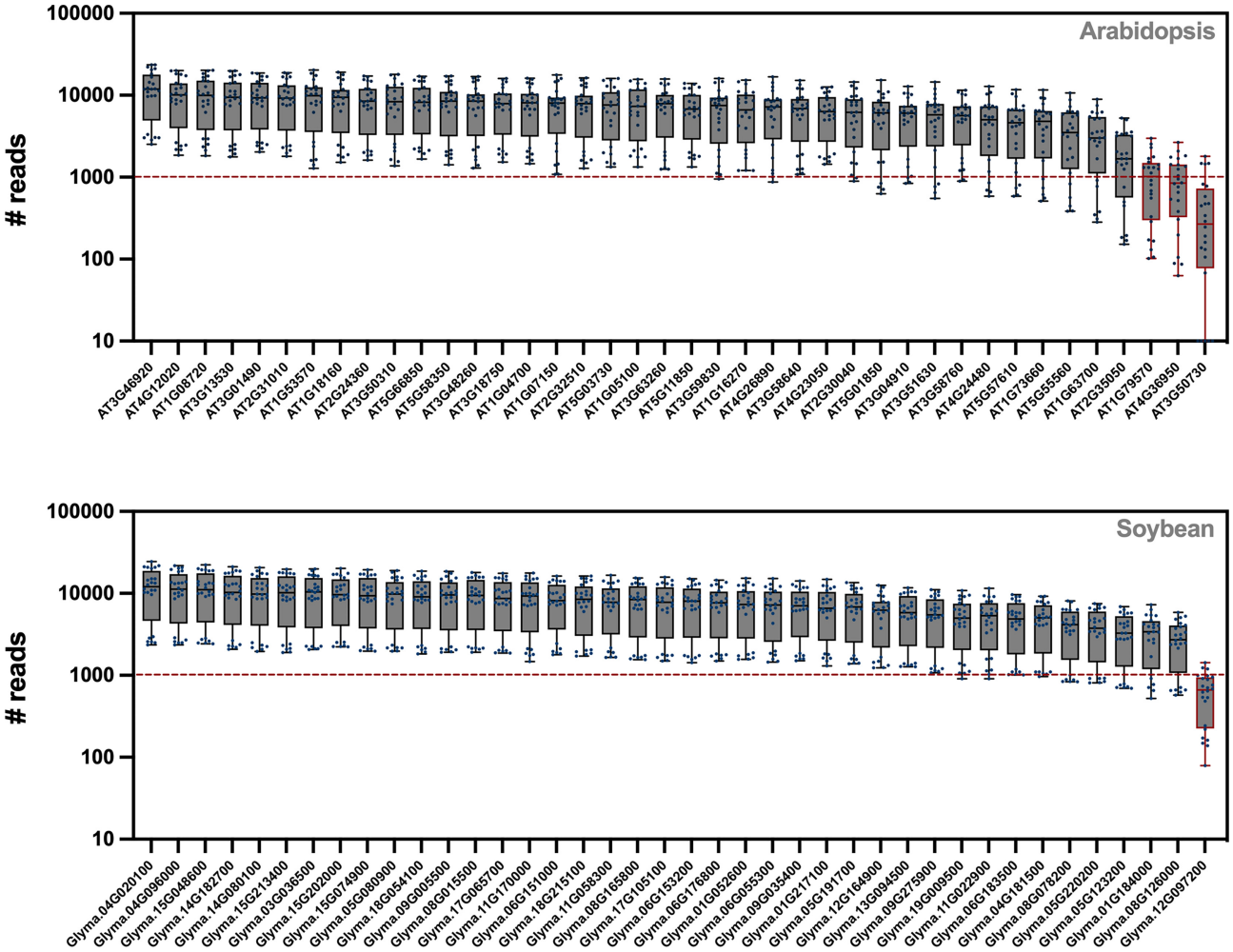

Figure 3. Read count per amplicon. For both the Arabidopsis (top) and soybean (bottom) genome, 40 amplicons were selected (one per gene) and sequenced using HiPlex sequencing on 24 replicate reference samples. The read depth per amplicon is given on the y-axis. The gene identifiers for which gene each amplicon was designed are given on the x-axis. The red dashed line indicates the desired minimum average read depth. Red boxes indicate amplicons with lower average read depth (these amplicons would be discarded and/or re-designed).

window, by the inclusion of the non-reference haplotypes in the FASTA reference sequence used for read mapping, or by filtering for indels around the expected cut site in *SMAP effect-prediction*.

After demonstrating that *SMAP design* could design highly reliable amplicons on wild-type materials, we evaluated the entire pipeline from gRNA and amplicon design to the evaluation and quantification of CRISPR-induced mutations. A total of 117 gRNA and amplicon designs were generated for 67 growth-related maize (B104) genes (24) with the $\text{Design}_{\text{HiPlex}}$ setting and HiPlex sequencing was performed on 20 wild-type leaf samples to validate the amplicons. The top 50 performing amplicons (high read depth and only reads exactly matching the reference sequence) were selected for a CRISPR screen in maize protoplasts (Supplementary Table S3; Supplementary Figure S2). Ten multiplex gRNA arrays were constructed, each containing five gRNAs. gRNAs targeting the same gene were placed on different gRNA arrays to avoid the generation of large deletions that can make the analysis more complex. Each gRNA array was co-transfected in triplicate with a Cas9 plasmid and six transfections with only the Cas9-carrying plasmid were included as negative controls. Two days after transfection, DNA was extracted from the protoplasts and analyzed by HiPlex sequencing using the validated primer mix.

Using this approach, we aimed to sequence all 50 amplicons in all 36 samples. For 46 out of the 50 amplicons a consistent read depth with at least 30 reads per amplicon for > 90% of the samples was observed, while the remaining four amplicons failed to produce any reads (Figure 4). We used *SMAP haplotype-window* and *SMAP effect-prediction* to quantify indels and primarily detected indels at the expected target sites, with frequencies as high as 60% (calculated as the percentage of reads with a non-reference haplotype in the region around the expected cut-site). However, we also detected indels at untargeted sites, though these values were typically < 2% and can be considered background noise. Amplicon 48 was particularly noisy, showing a consistent, but low level of indels for all samples (Figure 4). Overall, 45 of the 46 sequenced gRNA target sites show evidence of indels in at least two replicates with efficien-

Downloaded from https://academic.oup.com/nar/article/51/7/e37/7016455 by guest on 02 May 2023

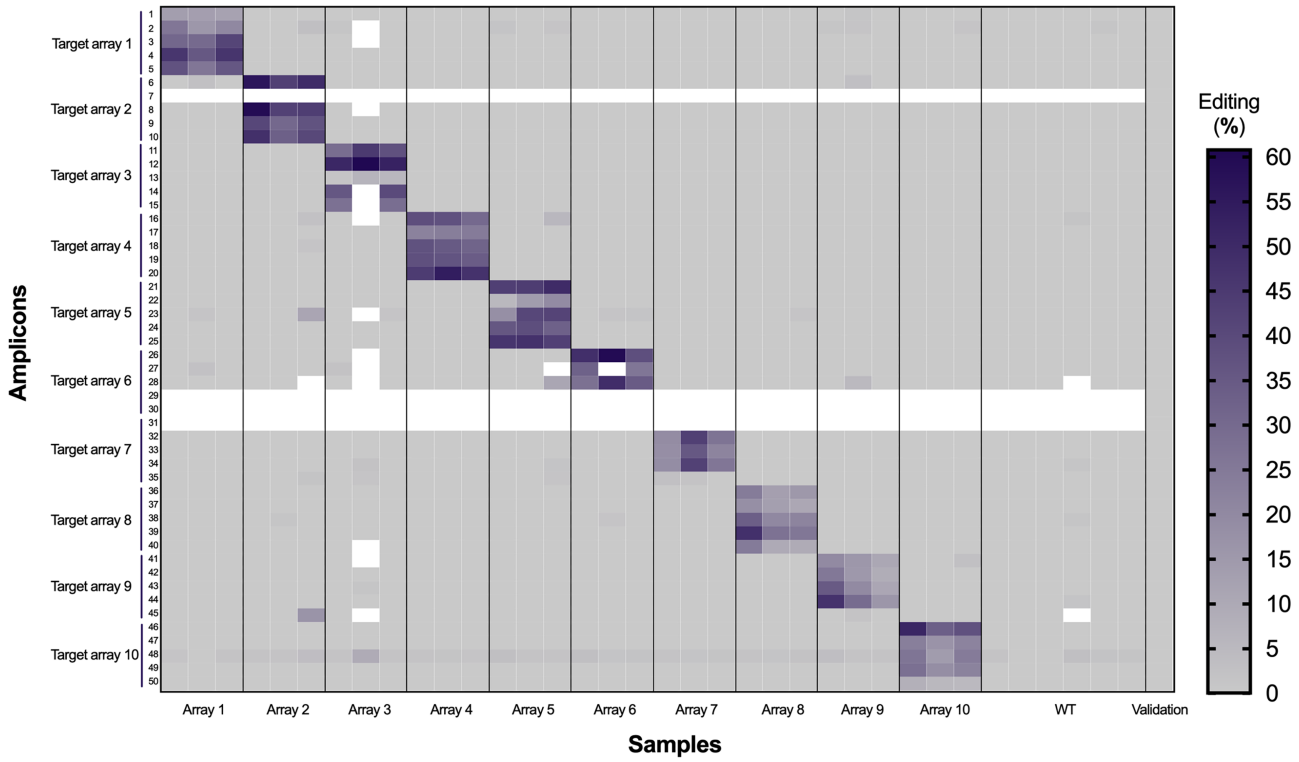

Figure 4. Editing efficiency in maize protoplasts after CRISPR/Cas9 transfection. Protoplasts were transfected in triplicate with a plasmid containing a gRNA array of five gRNAs and a Cas9 plasmid. The Cas9-only negative control was transfected in six replicates. In total, 50 gRNAs targeting 40 genes were transfected in ten separate reactions. For each sample, HiPlex sequencing was performed and analyzed using *SMAP haplotype-window* and *SMAP effect-prediction*. Each row represents the editing efficiency at a particular target site indicated by the amplicon ID (Supplemental Table 3); the columns are the different samples transfected with one of the arrays or without the array (WT). The validation column is the aggregated editing efficiency of the validation run on 20 wild-type leaf samples. Blank cells are samples/loci for which no sequencing data was obtained.

cies > 2%. Approximately 50% of the non-reference haplotypes contain a 1-bp insertion or deletion (Supplementary Table S6). Altogether these data show that *SMAP design* can be used to create reliable amplicon and gRNA designs for CRISPR-based screens.

Genome-wide *SMAP design*

As the initial *in silico* designs already generated amplicon and gRNA sets for thousands of genes, we decided to precompute amplicons and gRNAs across the entire Arabidopsis genome as a resource for the research community. The Arabidopsis Col-0 reference genome contains 27 655 annotated genes assigned to 9929 gene families containing between 1 and 208 genes (20). As we expected the potential for mispriming to be highest between genes from the same gene family, gene families were kept intact and divided into 28 groups of ±1000 genes to reduce the runtime via parallelization. *SMAP design* was run on each group to generate amplicons with $\text{Design}_{\text{HiPlex}}$, $\text{Design}_{\text{PE}}$ and $\text{Design}_{\text{Sanger,Ext}}$. A maximum of three non-overlapping amplicons covering a maximum of two gRNAs per amplicon were designed per gene. The CPU runtime per group of ±1000 genes ranged from 16 to 78 h (average 47 h) on a server with Intel Xeon CPUs using one core per group. The retention rate (defined here as the fraction of genes within the gene family with at least one amplicon with at least one gRNA per gene) was ~62% (17 186), ~68% (18 886), and ~85% (23 384) of all 27 655 Arabidopsis genes with $\text{Design}_{\text{HiPlex}}$, $\text{Design}_{\text{PE}}$ and $\text{Design}_{\text{Sanger,Ext}}$, respectively.

In an effort to capture a greater fraction of the genes from the genome-wide designs, we checked if there were certain features that led to the dropout of amplicons and/or gRNAs. In particular, we wondered if genes from large gene families were preferentially filtered out in the gRNA/amplicon filtering steps, with the expectation that larger gene families would be overrepresented in the dropout gene set. We compared the relative distribution of the gene family size of the dropout genes to the genome-wide gene family size distribution and found an equal proportion of dropout genes across different gene family sizes (Supplementary Figure S5; two-sided Kolmogorov–Smirnov test P-value = 0,87), suggesting that the retention rate is not biased towards a particular gene family size. We therefore questioned if the dropout was due to random matches of primers to non-gene-family members instead of sequence similarity to gene family members. To test this and increase the coverage of the genome-wide designs, a second run (here termed the dropout-only run) was performed where the dropout genes were run again but grouped only with other members of their gene family to still avoid potential mispriming between gene family members. Adding the designs from the dropout-only run increased the total coverage for $\text{Design}_{\text{HiPlex}}$, $\text{Design}_{\text{PE}}$ and $\text{Design}_{\text{Sanger,Ext}}$ to ~92%, ~94% and ~96%, respectively. Thus, the dropout genes were likely lost due to Primer3-predicted non-specific

Downloaded from https://academic.oup.com/nar/article/51/7/e37/7016455 by guest on 02 May 2023

primer binding onto reference sequences outside of their gene families.

A similar two-step approach was followed for the *P. patens* genome where the 32 926 genes were divided into 33 groups of ±1000 genes, keeping gene family members together (20). The genome consists of 15 604 gene families with 1–273 members. The $\text{Design}_{\text{Sanger,Ext}}$ run yielded amplicons and gRNAs for ∼77% of the genome and was increased to ∼86% by including the designs from the additional dropout-only run. The CPU runtime for a group of ±1000 genes ranged from 26 to 146 h (average 113 h).

To validate the Arabidopsis genome-wide $\text{Design}_{\text{Sanger,Ext}}$ and check for off-target amplification, 48 primer pairs were selected from each of the first and dropout-only runs of the genome-wide design. PCR followed by gel electrophoresis showed that 94 out of 96 amplicons had a single visible band (Supplementary Figure S6) and the two other amplicons (one from the first run and one from the dropout-only run) showed a single, yet less intense band. High-quality Sanger sequencing reads were obtained for 85 out of the 96 PCR products and confirmed all amplicons specifically amplified one single locus. Overall, since no discrepancy was found between the first and dropout-only runs, either through gel electrophoresis or through Sanger sequencing, we conclude that the primers are efficient and specific for Sanger sequencing and our observations suggest that the Primer3 default settings for eliminating non-specific primers were too conservative for Sanger sequencing. We therefore added an option for users to adjust the specificity settings of Primer3 when running *SMAP design*.

The eleven low-quality reads contained stretches with ≥ 10 thymidines or adenosines (homopolymers) resulting in overlapping sequencing peaks which are problematic for Sanger and Illumina-based genotyping. Interestingly, homopolymers with < 10 repeated nucleotides were observed in the sequenced amplicons but did not lead to overlapping sequencing peaks. At least under our Sanger sequencing conditions, there appears to be a threshold of 10 repeated nucleotides. We therefore calculated how many potentially problematic homopolymers were present in the amplicons of the genome-wide design of $\text{Design}_{\text{Sanger,Ext}}$ for Arabidopsis. Out of the 45 189 amplicons, 2889 (6.4%) had at least one homopolymer (≥ 10 nucleotides) with the majority of the homopolymers consisting of poly-A or poly-T (> 99%) (Supplementary Figure S7). A filter was therefore implemented in *SMAP design* to remove amplicons containing homopolymers of a user-defined size (Supplementary Table S2). Running *SMAP design* on $\text{Design}_{\text{Sanger,Ext}}$ with the homopolymer filter (*-hp*) set to 10 nucleotides for the genome of Arabidopsis and *P. patens* yielded a retention rate of ∼95% and ∼85% respectively (including the dropout-only runs). These final genome-wide designs are available in Supplementary data.

Using *SMAP design* to screen for natural variation

Amplicon design and detection rates. To evaluate the use of *SMAP design* to screen for natural variation in a non-model organism (chicory), HiPlex amplicons were designed to sequence nine candidate genes putatively involved in haploid induction (37–39). We aimed to create a catalogue of naturally-occurring sequence variants and ideally find haplotypes affecting the protein sequences as these could be used to generate haploid-inducer lines. We created two complementary HiPlex primer sets with *SMAP design* using $\text{Design}_{\text{NatVar}}$ settings that contain partially overlapping (tiled) amplicons for each of the nine genes for a total of 94 amplicons (Figure 5). We screened 35 chicory (*C. intybus* var. *sativum*) and 25 witloof accessions (*C. intybus* var. *foliosum*) by applying a 1D pooling strategy (Supplementary Figure S8A) in which an equal amount of leaf material from ∼10 individuals was pooled for a single DNA extraction and three independent pools per accession were created ($n \sim 30$ plants). In total, using the two HiPlex assays, 1554 chicory plants were screened in 163 pools, and pools with interesting sequence variants were identified. Individual plants from selected pools were then sequenced to identify carriers of knockout alleles.

We evaluated primer design performance by assessing the absolute read counts per amplicon. We set a 'detection' threshold at a minimum of 30 reads per amplicon per sample and defined the 'detection rate' per amplicon as the percentage of samples with read depth greater than 30 for that amplicon. Six amplicons were removed from further analysis because they either yielded no sequencing reads (CiDMP2_02, CiPLP6_16, and KNL2_03) or had a read depth below 100 reads across all samples (CiCENH3.2_01, CiDMP1_04 and CiDMP2_07). In the pooled sequencing run, 87 of the 88 remaining amplicons were detected in at least 90% of the reference samples (Table 2). The number of amplicons detected in at least 90% of the samples dropped in the individual sequencing run compared to the pooled sequencing run for both chicory and witloof, with chicory having a higher detection rate compared to witloof. For chicory, the detection rate dropped from 97% to 82%, and for witloof from 77% to 73% for pools and individuals, respectively (Table 2), confirming the expectation that amplification becomes less efficient with increasing genetic distance from the chicory reference genome sequence. Overall, we were able to cover 36% to 92% of the CDS per gene, after considering the design and amplification dropouts (Figure 5, Table 2).

Identification of conserved and variable gene regions in a breeding gene pool. We used *SMAP haplotype-window* to list the number of different haplotypes per amplicon per sample and estimated the relative haplotype frequencies in the pooled dataset (Supplementary Table S7). We compared the overall number of haplotypes and haplotypes leading to protein changes between chicory and witloof accessions in the pooled dataset using *SMAP effect-prediction*. In the chicory accessions, 257 different haplotypes were found across the 88 amplified loci, while for the witloof accessions 519 different haplotypes were found across all 88 loci, of which 242 haplotypes are found in both chicory and witloof, illustrating a higher level of sequence variation within these genes in the witloof accessions. In the chicory and witloof accessions, 93 (36%) and 247 (48%) of the haplotypes led to a different protein sequence, respectively. Of the 267 unique haplotypes found leading to protein changes across both chicory and witloof accessions, 253 (95%) were SNPs or in-frame mutations, while only 14 haplotypes (5%) were

Downloaded from https://academic.oup.com/nar/article/51/7/e37/7016455 by guest on 02 May 2023

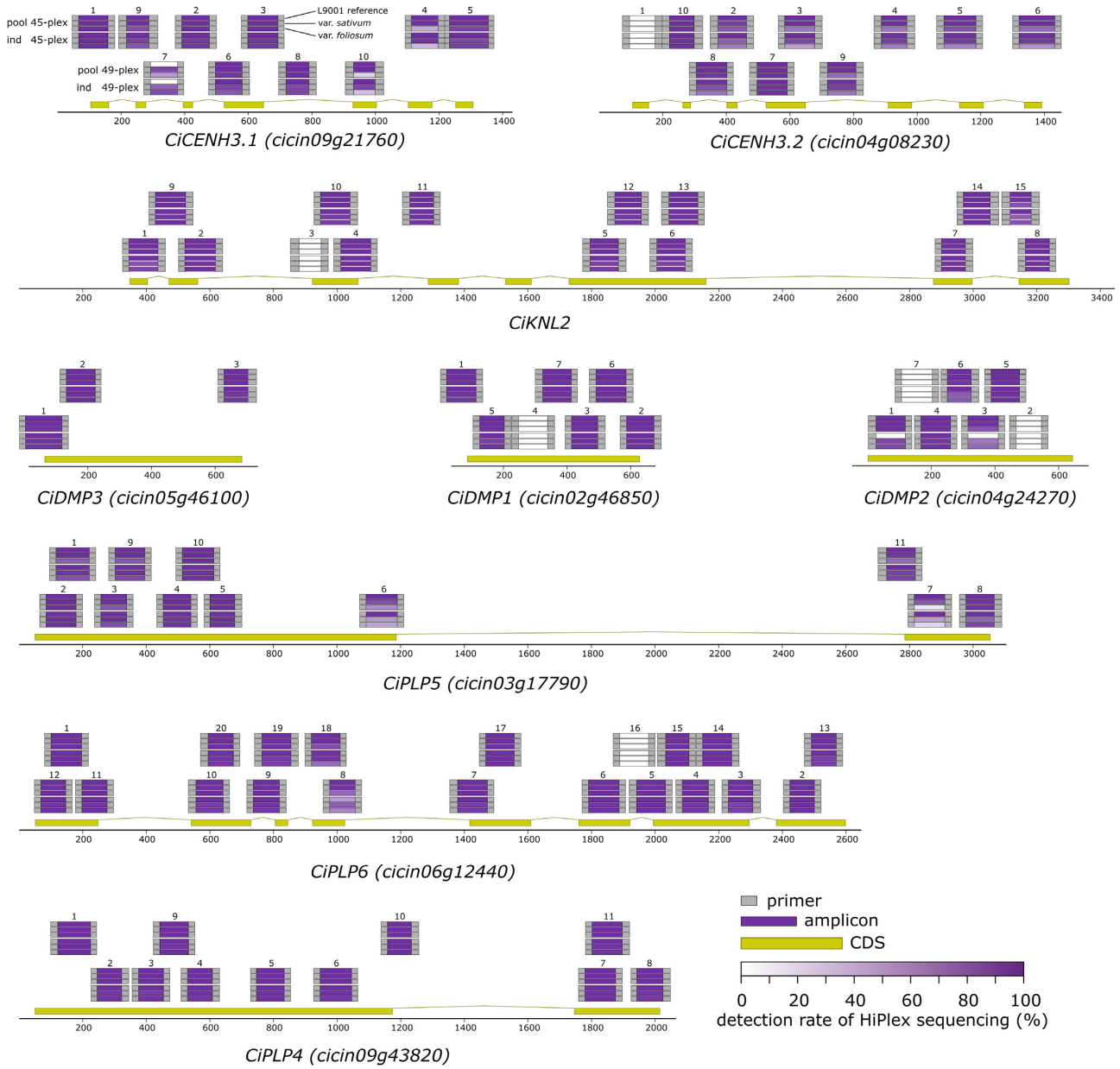

Figure 5. Detection of HiPlex amplification of two (tiled) amplicon designs for nine candidate genes putatively involved in haploid induction. Data from the 45-plex and the 49-plex assay in pooled (pool) and individual (ind) sequencing runs are included. For each amplicon and each run, the detection rate is shown, split by replicates of the L9001 reference genotype (top bar), *C. intybus* var. *sativum* samples (middle bar), and *C. intybus* var. *foliosum* samples (lower bar).

frameshift mutations. A total of 21 amplicons located in exonic regions did not have any haplotypes leading to protein changes and could thus be considered as conserved genic regions. The number of haplotypes per amplicon varied between genes, and across the length of the gene sequences (Figure 6). For instance, in *CiCENH3.1* and *CiCENH3.2,* more haplotypes and haplotypes leading to protein changes were found in the N-terminal region of the genes. The average number of all haplotypes per amplicon per gene ranged from 4.2 (*CiPLP4*) to 7.6 (*CiCENH3.1*), and the average number of haplotypes per amplicon per gene with protein sequence changes ranged from 1.2 (*CiCENH3.2*) to 4.4 (*CiKNL2*).

We focused on haplotypes with changes in the predicted protein sequences to identify individuals with potential knockout alleles, defined here as a protein sequence similarity of < 50% compared to the reference protein sequence. Using *SMAP effect-prediction*, we predicted the effect of haplotypes on the protein function as 'mild effect' if > 50% of the resulting amino acid sequence was identical to the reference protein and as 'strong effect' if at most 50% of the amino acid sequence was identical to the encoded reference protein. Out of 267 haplotypes, 255 were classified as a mild effect on protein function (48% of all haplotypes) and 12 haplotypes were classified as strong effect (2% of all haplotypes; Table 3, Supplementary Table S7). In total,

Downloaded from https://academic.oup.com/nar/article/51/7/e37/7016455 by guest on 02 May 2023

Table 2. Primer performance in pooled and individual sequencing of nine chicory genes across accessions of *Cichorium intybus* var. *sativum* and *C. intybus* var. *foliosum*. The *C. intybus* var. *sativum* genotype L9001 was used to create the reference genome sequence and the primer design

Dataset	Species	# samples	# amplicons	# amplicons > 90% of samples successful	# amplicons < 90% and ≥ 50% of samples successful	# amplicons < 50% of samples successful
Pool	Reference	6		87 (99%)	0 (0%)	1 (1%)
	var. *sativum*	104	88	86 (98%)	2 (2.3%)	0 (0%)
	var. *foliosum*	55		68 (77%)	17 (19%)	3 (3.4%)
Individual	Reference	2		84 (95%)	1 (1.6%)	3 (3.4%)
	var. *sativum*	156	88	82 (82%)	14 (16%)	1 (1%)
	var. *foliosum*	140		65 (73%)	18 (20%)	5 (5.7%)

320 plants (from 36 initial pools, 15 from chicory, 21 from witloof) with strong-effect haplotypes were sequenced individually. Due to the loss of plants between sampling of the pools and individuals, data were obtained for 294 individuals (91%). We recovered a total of 13 strong-effect haplotypes in *CiPLP6, CiPLP5, CiDMP1* and *CiDMP2*. *CiPLP6* was the most variable, with 10 strong-effect haplotypes, which were found 84 times in a heterozygous state and eight times in a homozygous state across 46 individuals and were often combined in a single individual (Table 3). Strong protein effect haplotypes in *CiPLP5* (2 individuals), *CiDMP1* (2 individuals) and *CiDMP2* (1 individual) were all present in a heterozygous state (Table 3).

Accuracy and sensitivity of haplotype detection in pool-seq. With the pooled and individual sequencing data, we assessed the sensitivity of pooled sequencing combined with HiPlex sequencing to identify haplotypes across a range of candidate genes in parallel. We analyzed the data from 13 pools of 10 individuals with complete sequencing data in all 130 individuals (ground truth) and calculated the 'expected' pooled haplotype frequency based on the discrete genotype calls of the 10 constituent diploid individuals. For instance, a single heterozygous individual in a pool of 10 diploid plants (i.e. 20 alleles) corresponds to a 5% relative haplotype frequency in HiPlex pooled sequencing data. A strong correlation (R^2 value: 0.8487, $P < 0.001$) was found between the observed haplotype frequencies in pools and the haplotype frequencies in individuals (Figure 7A). Additionally, 1826 of all 1933 (94.5%) haplotypes detected in individuals were also detected in their respective pools (true positives in pools), and 107 haplotypes (5.5%) were only detected in the respective individuals (false negatives in pools). About half (59/107) of the false-negative haplotypes displayed an 'expected' haplotype frequency of 5% or 10% in the individual sequencing data of the respective pool, indicating that pooled sampling effectively detects almost all haplotype diversity, with a weak bias against very low-frequency haplotypes (Figure 7B). Conversely, 106 of 1932 (5.5%) haplotypes detected across all pools were not detected in their respective constituent individuals (false positive in pools). The observed haplotype frequencies of 89 of 106 (84.0%) of the false positives were in the range of 1–5%, characteristic of low-frequency read errors (Figure 7B). Taken together, these data show that HiPlex pooled sequencing ($n = 10$) accurately quantifies the relative frequency of haplotypes within a pool and is sensitive enough to detect nearly all low-frequency haplotypes, including rare defective alleles.

DISCUSSION

There is an increasing need for scalable PCR amplicon design for genotyping as the scale of eco-tilling and multiplex CRISPR experiments continues to increase. This is particularly the case when conducting CRISPR screens with at least ~50 genes as manual designs can take several weeks, if not months, to perform. The currently available CRISPR design tools can generate lists of genome-wide gRNAs, but they lack the ability to combine this with genotyping primer design in a flexible and customizable manner. Tools such as CHOPCHOP (14), CRISPOR (13) or CROPSR (40) allow gRNA and associated primer design, but the user is very limited in the ability to customize it (e.g. number of gRNAs per amplicon, number of amplicons per gene, relative position of gRNAs along the gene, etc.). In addition, these tools provide the user with tens to hundreds of designs per gene, leaving them to sort through the designs that are compatible with each other in a multiplex format. *SMAP design* overcomes these limitations by designing highly specific amplicons at gRNA target sites for any number of user-selected genes and presents the user with compatible, ready-to-order designs. *SMAP design* greatly simplifies larger CRISPR experiments (e.g. multiplex and combinatorial CRISPR screens (41)) by reducing the design step from weeks or months to just hours of CPU time. The pre-designed genome-wide amplicons presented here effectively eliminate the design step altogether as users just need to select their gene identifier in the list and order the associated primers and/or gRNAs. Based on our empirical tests using HiPlex and Sanger sequencing of wild-type and mutant materials, these pre-computed gRNA and amplicon designs are reliable and amplify the target loci with high specificity. The resulting Sanger or NGS data can then be seamlessly analyzed with ICE (22), TIDE (23), or other tools from the SMAP package (19) using the output files from *SMAP target-selection* and *SMAP design* as input (Figure 1).

Similarly, *SMAP design* can be used to screen for naturally-occurring sequence variants with high specificity and sensitivity. For eco-tilling applications, amplicon design relies on gene-specific primers, limited off-target amplification, and covering as much of the gene sequence as possible. Current amplicon design methods consist of primer design by Primer3, followed by a BLAST and/or a preliminary PCR to check for mispriming (42). This can become laborious and time-consuming for large numbers of target genes. *SMAP design* automates this amplicon design and avoids mispriming. Amplicon design for the nine candidate genes in chicory proved to be reliable and gene-specific and

Downloaded from https://academic.oup.com/nar/article/51/7/e37/7016455 by guest on 02 May 2023

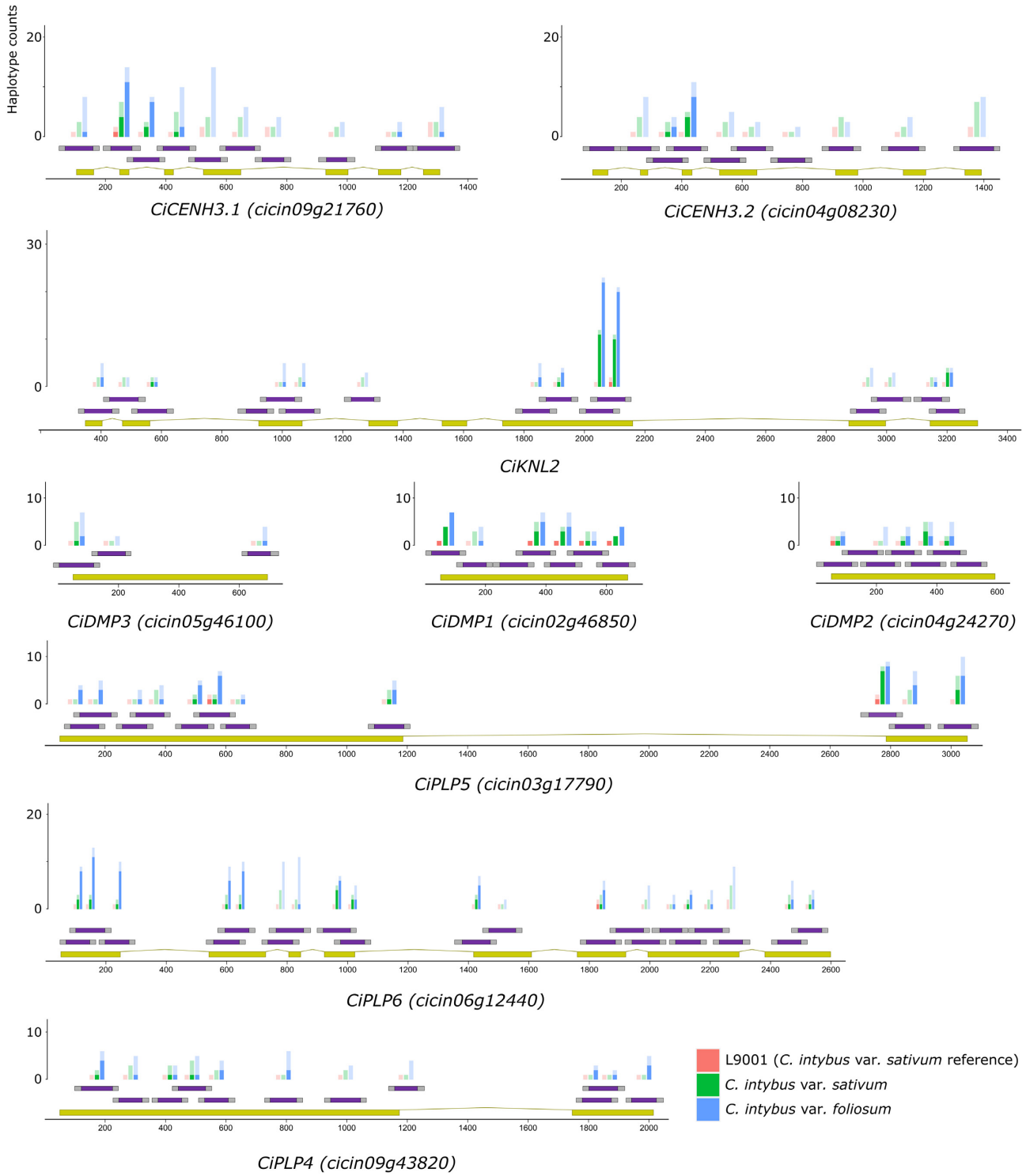

Figure 6. Observed haplotypes in *Cichorium intybus* pooled samples. The total number of haplotypes (transparent) and the number of haplotypes with protein sequence changes (opaque) per amplicon and per accession are indicated as bars. For the L9001 reference genotype, 6 replicates were included. For 35 accessions of *C. intybus* var. *sativum*, a total of 1039 plants were screened in 104 pools. For 25 accessions of *C. intybus* var. *foliosum*, a total of 465 plants were screened in 54 pools.

Downloaded from https://academic.oup.com/nar/article/51/7/e37/7016455 by guest on 02 May 2023

Table 3. Effects of mutations on predicted protein sequence from haplotypes detected in pools and individual (ind) sequencing. Haplotypes with an effect on the protein sequence were defined as 'mild' effect if more than 50% of the resulting protein sequence was identical to the reference protein, or as 'strong' effect if at most 50% of the protein sequence was identical. Abbreviations: HE = heterozygous, HO = homozygous. NA* The gene identifier for this gene is missing in the latest annotation of the genome

Gene	GeneID	CDS coverage (%)	Total number of haplotypes (pools/ind)	# haplotypes mild effect (pools/ind)	# haplotypes strong effect (pools/ind)	# carriers identified HE/HO (pools)
CiCENH3.1	*cicin09g21760*	92.27	76/71	23/21	0/0	0/0 (0)
CiCENH3.2	*cicin04g08230*	79.73	51/43	11/5	0/0	0/0 (0)
KNL2	NA*	65.40	96/65	60/39	1/0	0/0 (0)
CiDMP1	*cicin02g46850*	79.23	32/27	27/18	3/1	2/0 (2)
CiDMP2	*cicin04g24270*	70.17	22/20	11/6	0/1	1/0 (1)
CiDMP3	*cicin05g46100*	35.66	13/11	3/3	0/0	0/0 (0)
CiPLP5	*cicin03g17790*	57.29	62/48	40/28	1/1	2/0 (1)
CiPLP6	*cicin06g12440*	80.53	136/134	63/59	7/10	84/8 (16)
CiPLP4	*cicin09g43820*	58.64	46/46	17/17	0/0	0/0 (0)
Total			534/465	255/196	12/13	89/8 (20)

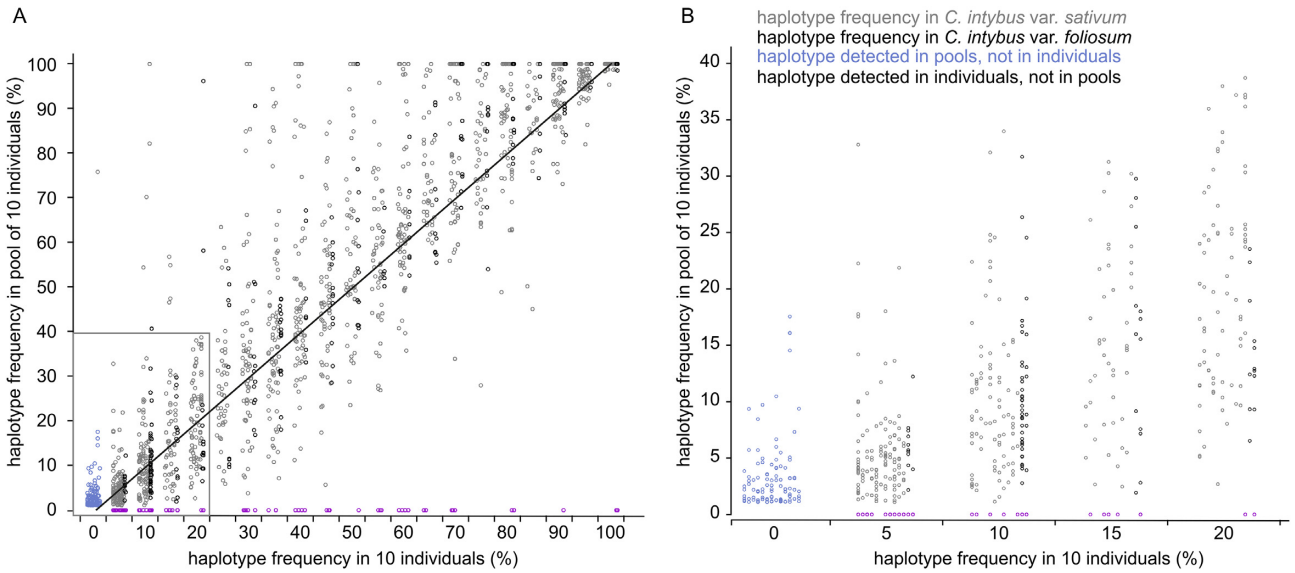

Figure 7. Comparison of haplotype frequency in pools to haplotype frequency in their 10 constituent individuals. (**A**) The diagonal line shows the expected ratio where the observed PoolSeq haplotype frequency is equal to the expected haplotype frequency based on individual sequencing. (**B**) Enlargement of the small panel inset in (A).

amplification was robust in different genotypes and accessions, thus capturing a broad range of sequence variation across the breeding gene pool.

While the number of genes that *SMAP design* can handle is theoretically unlimited, it is not always possible to design gene-specific amplicons or gRNAs for all genes. As shown here, genomes with relatively recent whole genome duplications or polyploid genomes suffer from lower retention rates, most likely due to the primer specificity checks implemented in Primer3. The retention rate depends on the genome and gene family but is generally higher than 80% for most of the tested species. Generating designs in species with highly duplicated genomes, such as soybean, can be more challenging. We show that increasing amplicon size, restricting primer design to exons, and splitting up gene families or groups can increase the retention rate. Furthermore, relaxing the default primer specificity settings in Primer3 would likely increase the retention rate as well. The settings that can be changed are of course dependent on the type of screen that will be performed; if simplex PCR will be used (e.g. for Sanger sequencing), cross-amplification is of no concern so relaxing the primer-specificity filters can be tolerated, but cross-amplification would be problematic for highly multiplex PCR where mixtures of primer pairs are used. As there are a wide range of settings, options and variables (genotyping assay, genome, and target genes), we suggest the practical approach is to first empirically validate all genotyping assays by sequencing reference genotypes and eliminate any primers or amplicons that do not amplify efficiently or are non-specific as we demonstrate in the maize CRISPR screen (Figure 4). This will ensure a smooth genotyping workflow once the mutant materials are generated and need to be characterized. This also ensures that there are no sequence variants in the gRNA targets between the reference sequence used for the design and experimental genotypes, and if there are, corrections can be made before cloning is initiated. Since the composition of the reference sequence influences the specificity of primer and gRNA

Downloaded from https://academic.oup.com/nar/article/51/7/e37/7016455 by guest on 02 May 2023

designs (and thus the overall design retention rate), *SMAP target-selection* is an important utility tool to streamline the construction of alternative reference gene sets and customize input parameter settings for optimal retention rate and reference sequence coverage, while maintaining target specificity.

Some limitations to designing gRNAs and amplicons in a high-throughput fashion remain. For instance, it is not yet possible to design a single gRNA that simultaneously targets multiple genes together with the corresponding gene-specific amplicons. Programs such as CRISPys (43) or MultiTargeter (44) can design gRNAs that allow the targeting of multiple loci by exploiting the capability of gRNAs with mismatches to the target sequences to still be functional. A gRNA list from such a program could be fed to *SMAP design,* however, amplicons will only be designed for the targets with an identical gRNA sequence and primers binding multiple regions will be filtered out. Furthermore, the on-target efficiency scores such as Doench (33) and Out-of-Frame (13) have not translated well to plants (45). Therefore, it is not guaranteed that a gRNA in the output of *SMAP design* will result in a knockout even though it might have a predicted high efficiency. Efficiency scores trained on plant data are thus highly desirable.

We also illustrated the capacity for HiPlex amplicon sequencing to perform eco-tilling by detecting low-frequency alleles (5%) in a pool of 10 individuals. Indeed, to sequence very large numbers of individuals, 1D pooling of 10 plants per pool substantially reduces the time, effort, and cost for DNA extraction, library preparation and sequencing, while retaining detection accuracy and sensitivity for low-frequency haplotypes (Supplementary Figure S8A). Screening efficiency can be further increased with 2D or 3D pool sequencing approaches, routinely used in tilling by sequencing applications (42). For example, a 2D pooling scheme based on a pool size of 10 plants divides 100 plants into $2 \times 10 = 20$ pools, each containing 10 plants in which each plant becomes part of 2 pools (with $X_{1\text{-}10}$ and $Y_{1\text{-}10}$ coordinates; Supplementary Figure S8B). This leads to a 5-fold reduction in the number of PCRs (20 PCRs on pools instead of 100 PCRs on individuals). Given that we observe a minimum detection threshold of ~5% allele frequency per pool, this allows us to detect a single heterozygous mutation in a pool of 10 plants. If a particular mutation occurs only once in the set of 100 (diploid) plants, the plant carrying the mutation can be identified in the corresponding X- and Y-pools. If the same mutation occurs more than once in the set of 100 plants (> 1% population frequency), a second round of screening at the individual plant level is required at each of the intersecting X- and Y-pool coordinates. As knockout alleles are quite rare, such a 2D or even 3D pooling approach can be used to quickly identify such alleles and their carriers in a cost-effective manner. Combined with HiPlex amplicon sequencing, which can screen multiple loci and genes at once, this allows for rapid screening of many genes in large populations. While we demonstrated the versatility of pooled sequencing to screen for natural variation, it is clearly a useful strategy for screening large collections of CRISPR mutants, as each unique type of mutation (defined by the haplotype sequence) is identified independently.

Ultimately, we envision a reverse-genetics approach where a researcher would use *SMAP design* to first screen their gene pool material for natural knockout alleles, as demonstrated for *CiDMP1*, *CiDMP2*, *CiPLP5* and *CiPLP6*. The identified carriers of the alleles could then be utilized for functional analysis and/or breeding. Alternatively, if no genetic variation is found, the researcher then resorts to induced genetic mutations where CRISPR is a highly tractable option for transformable species. For example, we did not observe any strong-effect haplotypes in *CiCENH3.1*, *CiCENH3.2*, *CiKNL2*, *CiDMP3* or *CiPLP4*. Therefore, the straightforward way to continue investigating these genes for haploid induction is to utilize CRISPR mutagenesis of chicory (46). The target sequences have been confirmed in our gene pool material via HiPlex sequencing and the overlapping gRNAs can directly be cloned into Cas9/gRNA expression vectors. Overall, *SMAP design* will be a useful tool to perform high-throughput genetic screens using both natural and induced variation.

DATA AVAILABILITY

All tools within the SMAP package (*SMAP target-selection, SMAP design, SMAP haplotype-window, SMAP effect-prediction*) and pre-computed designs are available in the GitLab repository https://gitlab.com/ilvo/smap-design and https://gitlab.com/truttink/smap. Manuals can be found at https://ngs-smap.readthedocs.io/.

FlashFry and CRISPOR for gRNA design can be found at https://github.com/mckennalab/FlashFry and https://github.com/maximilianh/crisporWebsite respectively.

PLAZA (https://bioinformatics.psb.ugent.be/plaza/) was used to retrieve genomes and annotations files.

BWA-MEM (https://github.com/lh3/bwa) was used to map reads to the reference sequences.

SMAP target-selection and *SMAP design* are also available as a graphical user-interface via Galaxy at https://usegalaxy.be/.

Amplicon NGS files were deposited at SRA (https://www.ncbi.nlm.nih.gov/sra) under the accession numbers PRJNA848638 (Arabidopsis, soybean, and maize) and PRJNA855321 (chicory).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank the ELIXIR Belgium team (supported by Research Foundation-Flanders, project I002819N) for the assistance in making *SMAP design* available in Galaxy. We thank Dries Schaumont for the guidance with the NGS analysis.

FUNDING

Fonds Wetenschappelijk Onderzoek (FWO) [1SE2521N to W.D.]; European Research Council (ERC) [833866 – BREEDIT to K.D.]; E.W. is supported by a scholarship from Chicoline, a division of Cosucra Groupe Warcoing S.A., Belgium. Funding for open access charge: Host institute.

Conflict of interest statement. None declared.

Downloaded from https://academic.oup.com/nar/article/51/7/e37/7016455 by guest on 02 May 2023

REFERENCES

1. Till,B.J., Zerr,T., Comai,L. and Henikoff,S. (2006) A protocol for TILLING and Ecotilling in plants and animals. *Nat. Protoc.*, **1**, 2465–2477.
2. Kumar,G.R., Sakthivel,K., Sundaram,R.M., Neeraja,C.N., Balachandran,S.M., Rani,N.S., Viraktamath,B.C. and Madhav,M.S. (2010) Allele mining in crops: prospects and potentials. *Biotechnol. Adv.*, **28**, 451–461.
3. Marroni,F., Pinosio,S., Di Centa,E., Jurman,I., Boerjan,W., Felice,N., Cattonaro,F. and Morgante,M. (2011) Large-scale detection of rare variants via pooled multiplexed next-generation sequencing: towards next-generation Ecotilling. *Plant J.*, **67**, 736–745.
4. Vanholme,B., Cesarino,I., Goeminne,G., Kim,H., Marroni,F., Van Acker,R., Vanholme,R., Morreel,K., Ivens,B., Pinosio,S. *et al.* (2013) Breeding with rare defective alleles (BRDA): a natural Populus nigra HCT mutant with modified lignin as a case study. *New Phytol.*, **198**, 765–776.
5. Manolio,T.A., Collins,F.S., Cox,N.J., Goldstein,D.B., Hindorff,L.A., Hunter,D.J., McCarthy,M.I., Ramos,E.M., Cardon,L.R., Chakravarti,A. *et al.* (2009) Finding the missing heritability of complex diseases. *Nature*, **461**, 747–753.
6. Wang,H., La Russa,M. and Qi,L.S. (2016) CRISPR/Cas9 in genome editing and beyond. *Annu. Rev. Biochem.*, **85**, 227–264.
7. Zhang,F., Wen,Y. and Guo,X. (2014) CRISPR/Cas9 for genome editing: progress, implications and challenges. *Hum. Mol. Genet.*, **23**, R40–R46.
8. Rodgers,K. and Mcvey,M. (2016) Error-prone repair of DNA double-strand breaks. *J. Cell. Physiol.*, **231**, 15.
9. Zhang,N., Roberts,H.M., Van Eck,J. and Martin,G.B. (2020) Generation and molecular characterization of CRISPR/Cas9-induced mutations in 63 immunity-associated genes in tomato reveals specificity and a range of gene modifications. *Front. Plant Sci.*, **11**, 10.
10. Wong,A.S.L., Choi,G.C.G., Cui,C.H., Pregernig,G., Milani,P., Adam,M., Perli,S.D., Kazer,S.W., Gaillard,A., Hermann,M. *et al.* (2016) Multiplexed barcoded CRISPR-Cas9 screening enabled by CombiGEM. *Proc. Natl. Acad. Sci. U.S.A.*, **113**, 2544–2549.
11. Stemmer,M., Thumberger,T., Del Sol Keyer,M., Wittbrodt,J. and Mateo,J.L. (2015) CCTop: an intuitive, flexible and reliable CRISPR/Cas9 target prediction tool. *PLoS One*, **12**, e0124633.
12. McKenna,A. and Shendure,J. (2018) FlashFry: a fast and flexible tool for large-scale CRISPR target design. *BMC Biol.*, **16**, 74.
13. Concordet,J.P. and Haeussler,M. (2018) CRISPOR: intuitive guide selection for CRISPR/Cas9 genome editing experiments and screens. *Nucleic Acids Res.*, **46**, W242–W245.
14. Labun,K., Montague,T.G., Krause,M., Torres Cleuren,Y.N., Tjeldnes,H. and Valen,E. (2019) CHOPCHOP v3: expanding the CRISPR web toolbox beyond genome editing. *Nucleic Acids Res.*, **47**, W171–W174.
15. Moreno-Mateos,M.A., Vejnar,C.E., Beaudoin,J.D., Fernandez,J.P., Mis,E.K., Khokha,M.K. and Giraldez,A.J. (2015) CRISPRscan: designing highly efficient sgRNAs for CRISPR-Cas9 targeting in vivo. *Nat. Methods*, **12**, 982–988.
16. O'Halloran,D.M. (2016) PrimerMapper: high throughput primer design and graphical assembly for PCR and SNP detection. *Sci. Rep.*, **6**, 20631.
17. O'Halloran,D.M. (2015) PrimerView: high-throughput primer design and visualization. *Source Code Biol. Med.*, **10**, 8.
18. Untergasser,A., Cutcutache,I., Koressaar,T., Ye,J., Faircloth,B.C., Remm,M. and Rozen,S.G. (2012) Primer3-new capabilities and interfaces. *Nucleic Acids Res.*, **40**, e115.
19. Schaumont,D., Veeckman,E., van der Jeugt,F., Haegeman,A., van Glabeke,S., Bawin,Y., Lukasiewicz,J., Blugeon,S., Barre,P., de La,O, Leyva-Pérez,M. *et al.* (2022) Stack Mapping Anchor Points (SMAP): a versatile suite of tools for read-backed haplotyping. bioRxiv doi: https://doi.org/10.1101/2022.03.10.483555, 13 March 2022, preprint: not peer reviewed.
20. Van Bel,M., Diels,T., Vancaester,E., Kreft,L., Botzki,A., Van De Peer,Y., Coppens,F. and Vandepoele,K. (2018) PLAZA 4.0: an integrative resource for functional, evolutionary and comparative plant genomics. *Nucleic Acids Res.*, **46**, D1190–D1196.
21. Vandepoele,K., Van Bel,M., Richard,G., Van Landeghem,S., Verhelst,B., Moreau,H., Van de Peer,Y., Grimsley,N. and Piganeau,G. (2013) pico-PLAZA, a genome database of microbial photosynthetic eukaryotes. *Environ. Microbiol.*, **15**, 2147–2153.
22. Conant,D., Hsiau,T., Rossi,N., Oki,J., Maures,T., Waite,K., Yang,J., Joshi,S., Kelso,R., Holden,K. *et al.* (2022) Inference of CRISPR edits from Sanger trace data. *CRISPR J.*, **5**, 123–130.
23. Brinkman,E.K., Kousholt,A.N., Harmsen,T., Leemans,C., Chen,T., Jonkers,J. and van Steensel,B. (2018) Easy quantification of template-directed CRISPR/Cas9 editing. *Nucleic Acids Res.*, **46**, e58.
24. Lorenzo,C.D., Debray,K., Herwegh,D., Develtere,W., Impens,L., Schaumont,D., Vandeputte,W., Aesaert,S., Coussens,G., de Boe,Y. *et al.* (2023) BREEDIT: a multiplex genome editing strategy to improve complex quantitative traits in maize. *Plant Cell*, **35**, 218–238.
25. Decaestecker,W., Buono,R.A., Pfeiffer,M.L., Vangheluwe,N., Jourquin,J., Karimi,M., van Isterdael,G., Beeckman,T., Nowack,M.K. and Jacobs,T.B. (2019) CRISPR-TSKO: a technique for efficient mutagenesis in specific cell types, tissues, or organs in Arabidopsis. *Plant Cell*, **31**, 2868–2887.
26. Blomme,J., Liu,X., Jacobs,T.B. and de Clerck,O. (2021) A molecular toolkit for the green seaweed Ulva mutabilis. *Plant Physiol.*, **186**, 1442–1454.
27. Berendzen,K., Searle,I., Ravenscroft,D., Koncz,C., Batschauer,A., Coupland,G., Somssich,I.E. and Ülker,B. (2005) A rapid and versatile combined DNA/RNA extraction protocol and its application to the analysis of a novel DNA marker set polymorphic between Arabidopsis thaliana ecotypes Col-0 and Landsberg erecta. *Plant Methods*, **1**, 4.
28. Edwards,K., Johnstone,C. and Thompson,C. (1991) A simple and rapid method for the preparation of plant genomic DNA for PCR analysis. *Nucleic Acids Res.*, **19**, 1349.
29. Doyle,J.J. and Doyle,J.L. (1987) A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem. Bull.*, **19**, 11–15.
30. Gaillochet,C., Fernandez,A.P., Goossens,V., D'Halluin,K., Drozdzecki,A., Shafie,M., Duyse,J.van, Isterdael,G.van, Gonzalez,C., Vermeersch,M. *et al.* (2023) Systematic optimization of Cas12a base editors in wheat and maize using the ITER platform. *Genome Biol.*, **24**, 6.
31. Li,H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv doi: https://arxiv.org/abs/1303.3997, 26 May 2013, preprint: not peer reviewed.
32. Hsu,P.D., Scott,D.A., Weinstein,J.A., Ran,F.A., Konermann,S., Agarwala,V., Li,Y., Fine,E.J., Wu,X., Shalem,O. *et al.* (2013) DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.*, **31**, 827–832.
33. Doench,J.G., Fusi,N., Sullender,M., Hegde,M., Vaimberg,E.W., Donovan,K.F., Smith,I., Tothova,Z., Wilen,C., Orchard,R. *et al.* (2016) Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat. Biotechnol.*, **34**, 184–191.
34. Bae,S., Kweon,J., Kim,H.S. and Kim,J.S. (2014) Microhomology-based choice of Cas9 nuclease target sites. *Nat. Methods*, **11**, 705–706.
35. Cannon,S.B. and Shoemaker,R.C. (2012) Evolutionary and comparative analyses of the soybean genome. *Breed Sci*, **61**, 437.
36. Schmutz,J., Cannon,S.B., Schlueter,J., Ma,J., Mitros,T., Nelson,W., Hyten,D.L., Song,Q., Thelen,J.J., Cheng,J. *et al.* (2010) Genome sequence of the palaeopolyploid soybean. *Nature*, **463**, 178–183.
37. Jacquier,N.M.A., Gilles,L.M., Pyott,D.E., Martinant,J.P., Rogowsky,P.M. and Widiez,T. (2020) Puzzling out plant reproduction by haploid induction for innovations in plant breeding. *Nat. Plants*, **6**, 610–619.
38. Kalinowska,K., Chamas,S., Unkel,K., Demidov,D., Lermontova,I., Dresselhaus,T., Kumlehn,J., Dunemann,F. and Houben,A. (2019) State-of-the-art and novel developments of in vivo haploid technologies. *Theor. Appl. Genet.*, **132**, 593–605.
39. Inna,L. (2017) Generation of haploid plants based on KNL2. WO2017067714A1
40. Müller Paul,H., Istanto,D.D., Heldenbrand,J. and Hudson,M.E. (2022) CROPSR: an automated platform for complex genome-wide CRISPR gRNA design and validation. *BMC Bioinformatics*, **23**, 74.
41. van Huffel,K., Stock,M., Ruttink,T. and de Baets,B. (2022) Covering the combinatorial design space of multiplex CRISPR/Cas experiments in plants. *Front. Plant Sci.*, **13**, 1870.
42. Tsai,H., Ngo,K., Lieberman,M., Missirian,V. and Comai,L. (2015) Tilling by Sequencing. *Methods Mol. Biol.*, **1284**, 359–380.

Downloaded from https://academic.oup.com/nar/article/51/7/e37/7016455 by guest on 02 May 2023

43. Hyams,G., Abadi,S., Lahav,S., Avni,A., Halperin,E., Shani,E. and Mayrose,I. (2018) CRISPys: optimal sgRNA design for editing multiple members of a gene family using the CRISPR system. *J. Mol. Biol.*, **430**, 2184–2195.
44. Prykhozhij,S.V., Rajan,V., Gaston,D. and Berman,J.N. (2015) CRISPR MultiTargeter: a web tool to find common and unique CRISPR single guide RNA targets in a set of similar sequences. *PLoS One*, **10**, e0119372.
45. Naim,F., Shand,K., Hayashi,S., O'Brien,M., McGree,J., Johnson,A.A.T., Dugdale,B. and Waterhouse,P.M. (2020) Are the current gRNA ranking prediction algorithms useful for genome editing in plants?*PLoS One*, **15**, e0227994.
46. De Bruyn,C., Ruttink,T., Eeckhaut,T., Jacobs,T., De Keyser,E., Goossens,A. and Van Laere,K. (2020) Establishment of CRISPR/Cas9 genome editing in Witloof (*Cichorium intybus* var. *foliosum*). *Front. Genome Ed.*, **2**, 604876.

Downloaded from https://academic.oup.com/nar/article/51/7/e37/7016455 by guest on 02 May 2023