OPEN ACCESS

Edited by:
Marcos Egea-Cortines,
Universidad Politécnica de Cartagena,
Spain

Reviewed by:
Unai López De Heredia,
Polytechnic University of Madrid,
Spain
Hiroyuki Koga,
The University of Tokyo, Japan

***Correspondence:**
José Luis Micol
jlmicol@umh.es

[†]*These authors have contributed equally to this work*

[‡]***Present address:***
David Wilson-Sánchez,
Department of Comparative Development and Genetics, Max Planck Institute for Plant Breeding Research, Cologne, Germany
Sergio Andreu-Sánchez,
Department of Pediatrics and Genetics, University of Groningen, Groningen, Netherlands

Specialty section:
This article was submitted to Technical Advances in Plant Science, a section of the journal Frontiers in Plant Science

Received: *18 January 2021*
Accepted: *18 March 2021*
Published: *07 May 2021*

Citation:
Lup SD, Wilson-Sánchez D, Andreu-Sánchez S and Micol JL (2021) Easymap: A User-Friendly Software Package for Rapid Mapping-by-Sequencing of Point Mutations and Large Insertions. Front. Plant Sci. 12:655286. doi: 10.3389/fpls.2021.655286

Easymap: A User-Friendly Software Package for Rapid Mapping-by-Sequencing of Point Mutations and Large Insertions

Samuel Daniel Lup[†], David Wilson-Sánchez[†‡], Sergio Andreu-Sánchez[‡] and José Luis Micol*

Instituto de Bioingeniería, Universidad Miguel Hernández de Elche, Elche, Spain

Mapping-by-sequencing strategies combine next-generation sequencing (NGS) with classical linkage analysis, allowing rapid identification of the causal mutations of the phenotypes exhibited by mutants isolated in a genetic screen. Computer programs that analyze NGS data obtained from a mapping population of individuals derived from a mutant of interest to identify a causal mutation are available; however, the installation and usage of such programs requires bioinformatic skills, modifying or combining pieces of existing software, or purchasing licenses. To ease this process, we developed Easymap, an open-source program that simplifies the data analysis workflows from raw NGS reads to candidate mutations. Easymap can perform bulked segregant mapping of point mutations induced by ethyl methanesulfonate (EMS) with DNA-seq or RNA-seq datasets, as well as tagged-sequence mapping for large insertions, such as transposons or T-DNAs. The mapping analyses implemented in Easymap have been validated with experimental and simulated datasets from different plant and animal model species. Easymap was designed to be accessible to all users regardless of their bioinformatics skills by implementing a user-friendly graphical interface, a simple universal installation script, and detailed mapping reports, including informative images and complementary data for assessment of the mapping results. Easymap is available at http://genetics.edu.umh.es/resources/easymap; its Quickstart Installation Guide details the recommended procedure for installation.

Keywords: bioinformatics, NGS, mapping-by-sequencing, candidate mutations, forward genetics, linkage analysis mapping, bulked segregant analysis

Abbreviations: AF, allele frequency; CV, coverage; EMS, ethyl methanesulfonate; NGS, next-generation sequencing; RD, read depth; SNP, single-nucleotide polymorphism.

INTRODUCTION

Forward genetic screens consist of random mutagenesis followed by the isolation of mutants exhibiting a phenotype of interest, and genetic analysis of these mutants to identify the mutations that cause their phenotypes. Two commonly used mutagenesis strategies are the induction of G→A substitutions using the chemical mutagen ethyl methanesulfonate (EMS) (Neuffer and Ficsor, 1963; James and Dooner, 1990; Jansen et al., 1997) and the disruption of genes by insertional mutagens such as transposons or T-DNA (Cooley et al., 1988; Alonso et al., 2003; Frøkjær-Jensen et al., 2014). Linkage analysis of molecular markers in segregant mapping populations is the classically preferred approach to map the point mutations induced by a chemical mutagen, carried by mutants isolated in a genetic screen (Michelmore et al., 1991; Ponce et al., 1999). By contrast, localization of insertional mutations has relied on methods to capture the genomic sequences present at their flanks (Gasch et al., 1992; Medford et al., 1992; Liu et al., 1995; Ponce et al., 1998; O'Malley et al., 2007).

The preliminary identification and subsequent validation of the mutations that cause a phenotype of interest is the most laborious and time-consuming step of a forward genetic screen. Next-generation sequencing (NGS) of DNA has facilitated and revitalized such approaches, through the so-called mapping-by-sequencing methods, which combine NGS with linkage analysis (Schneeberger and Weigel, 2011; Hartwig et al., 2012; James et al., 2013; Candela et al., 2015). Mapping-by-sequencing approaches for the identification of causal mutations are much faster than previous methods but can be hampered by the lack of computing resources and/or accessible software. Currently available programs for mapping-by-sequencing data analysis suffer from one or several of the following issues: they require the purchase of licenses (Smith, 2015); they require a certain level of bioinformatics skills to use (Abe et al., 2012; Fekih et al., 2013; Jiang et al., 2015; Sun and Schneeberger, 2015; Ecovoiu et al., 2016; Wachsman et al., 2017; see also https://sourceforge.net/projects/mimodd/); they only do a part of the computing tasks required for a mapping-by-sequencing experiment (Li et al., 2009; Langmead and Salzberg, 2012; Hill et al., 2013); they are designed for a specific type of mutation or mapping strategy (Gonzalez et al., 2013; Ewing, 2015; Hénaff et al., 2015; Solaimanpour et al., 2015; Sun and Schneeberger, 2015; Wachsman et al., 2017; Klein et al., 2018; Javorka et al., 2019); they are hosted at a public server but usage is limited (Gonzalez et al., 2013; Afgan et al., 2018); or they can no longer be accessed or used (Minevich et al., 2012; Pulido-Tamayo et al., 2016).

Here, we describe Easymap, an accessible graphical-interface program that analyses NGS data from mapping populations derived through a variety of experimental designs from either insertional or EMS-induced mutants. This software package avoids the aforementioned issues, enabling mapping-by-sequencing experiments to be conducted by researchers with minimal bioinformatics experience. Easymap features a web-based graphic interface, a simple installation script, robust mapping analyses for several experimental designs, and thorough user-oriented mapping reports.

RESULTS

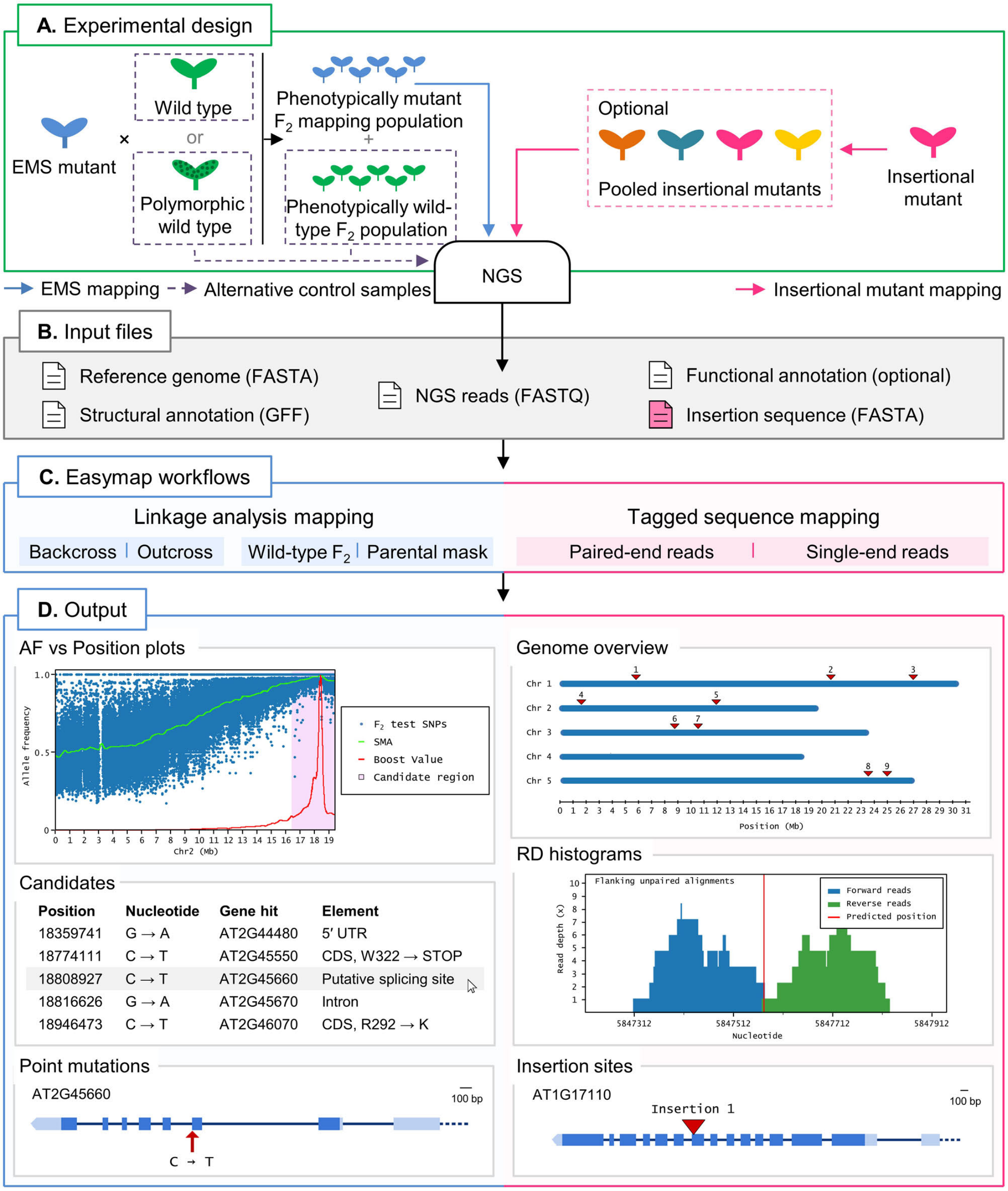
Mapping Strategies for Which Easymap Can Be Used

Easymap offers the user two alternative workflows, to map point and insertional mutations (Workflows 1 and 2; see next section), as well as a set of complementary tasks common to all analyses. **Figure 1** offers an overview of both mapping strategies, from the initial selection of the mutants of interest to the output that Easymap generates. To ease both mapping scenarios, Easymap automates the whole data analysis process requiring no user intervention with alignment, variant-calling, or filtering parameters.

Workflow 1: Linkage Analysis Mapping

Mapping of causal EMS-induced mutations is typically achieved by linkage analysis in a bulked segregant population. The user must obtain a mapping population by phenotyping the M_2 offspring of an M_1 individual or the F_2 offspring of a backcross (in which an M_2 mutant is crossed to an individual genetically identical to the parent subjected to mutagenesis) or an outcross (in which an M_2 mutant is crossed to an individual unrelated and genetically polymorphic to the parent that was subjected to mutagenesis).

Easymap includes the splice-aware aligner HISAT2 (Kim et al., 2019), which is three times faster than commonly used aligners in default conditions with no impact on memory usage or sensitivity. The implementation of HISAT2 allows the user to input RNA-seq and DNA-seq reads indistinctly for point mutation mapping. Easymap requires NGS reads from test and control samples. The test sample consists of NGS reads obtained from a population of individuals exhibiting the mutant phenotype of interest, hence homozygous for a recessive mutation that causes that phenotype. The control sample can be pooled M_2 or F_2 phenotypically wild-type individuals, or individuals genetically identical to the parent that was subjected to mutagenesis, or the strain to which the mutant is outcrossed. A minimum coverage of 25× is recommended for each sample, although since coverage directly correlates to the reliability of the results, higher coverages are encouraged. **Table 1** and the Easymap Documentation (**Supplementary Data Sheet 2**) describe the different experimental designs supported by Easymap, four of which are detailed in **Figure 2**. Easymap first calls the single-nucleotide polymorphisms (SNPs) between the reads obtained by the user and the reference sequence for the genome of the species under study, and identifies high-confidence SNPs that are informative for mapping. The allele frequencies (AFs) of these biallelic markers are then averaged in overlapping sliding windows throughout the genome, and the average AF values are corrected by comparing adjacent windows using a weighted average approach to reduce noise and randomly generated peaks, taking into account up to six adjacent windows for each correction. Easymap then analyses the revised AF values to select the window with the highest value, which is the most likely to contain the



Position	Nucleotide	Gene hit	Element
18359741	G → A	AT2G44480	5′ UTR
18774111	C → T	AT2G45550	CDS, W322 → STOP
18808927	C → T	AT2G45660	Putative splicing site
18816626	G → A	AT2G45670	Intron
18946473	C → T	AT2G46070	CDS, R292 → K

FIGURE 1 | Overview of two typical mapping-by-sequencing experiments with Easymap in *Arabidopsis*. **(A)** Experimental design. For EMS-induced mutants, an outcross or backcross is first performed. The F_1 plants derived from the cross are selfed, and the resulting F_2 is screened for the mutant phenotype to create a phenotypically mutant mapping population. A control sample is required for the mapping analysis, which can be either one of the parental individuals crossed or, alternatively, a pool of phenotypically wild-type F_2 individuals. For mapping of large insertions, the DNA of different insertional mutant lines can be sequenced individually or pooled, and no control sample is required. **(B)** Input files. Easymap takes NGS paired-end or single-end short reads as input. The remaining mandatory input files are available on public databases for each model species. **(C)** Easymap workflows. The user selects the experimental design used for mutation mapping from a variety of options for both EMS mutation mapping (backcross and outcross strategies, alternative control samples) and tagged-sequence mapping (paired-end and single-end reads). **(D)** Output. Easymap produces comprehensive mapping reports with organized tabular data to ease interpretation of the results. As an example of EMS-induced mutations, data from the *Arabidopsis suppressor of overexpression of CONSTANS 1-2* (*soc1-2*) mutant (Sun and Schneeberger, 2015) was used for this figure. Allele frequency (AF) versus position plots are drawn for each chromosome containing the polymorphisms used for the analysis. A candidate region is highlighted in pink; all putative EMS-type mutations contained in this region are regarded as candidates, and their position and relevant information are presented in a table, such as the gene affected by the mutation. For each gene affected by a candidate mutation, a gene plot is made in which the position of the mutation is shown, followed by further information (genotyping primers, flanking sequences, functional annotation, etc.). As an example of large insertion mapping, the figure includes data from an unpublished mapping experiment made in our laboratory (see **Table 2**). A genomic overview is drawn showing the positions of the insertions found. Read depth (RD) histograms are generated for each read cluster pointing to an insertion site showing the information supporting the insertion. Finally, a gene plot is made for each gene interrupted by an insertion.

TABLE 1 | Validation of point-mutation mapping strategies using published, real experimental data.

Genetic background of the mutant	Mapping strategy	Control sample	Mutant	Sample type	Results obtained by Easymap
Same as reference sequence	Backcross/M_2	Wild-type parental strain	*hasty*	DNA	CM[a] is the only candidate[b] within the CI[c]
			ago1-25	DNA	CM and 2 more candidates within the CI
			alp1-3	DNA	CM and one more candidate within the CI
			gl3	RNA	Correct CI, CM undetectable with our methods
		Phenotypically wild-type siblings	*atgr2*	DNA	CM and 3 more candidates within the CI
			shortroot	DNA	CM and 2 more candidates within the CI
			rth6	RNA	Correct CI, CM unknown in the original paper
	Outcross	Wild-type parental strain	*rol-6(su1006)*	DNA	Correct CI, CM unknown in the original paper
		Genetically distant strain crossed to the mutant	*soc1-2*	DNA	CM and 6 more candidates within the CI
			rol-6(su1006)	DNA	Correct CI, CM unknown in the original paper
Different from reference sequence	Backcross/M_2	Phenotypically wild-type siblings	*icu11-1*	DNA	CM is the only candidate within the CI
			jj410	DNA	Correct CI, CM lost due to very low coverage
			gl13	RNA	Correct CI, CM is fine-mapped in later experiments
	Outcross	Wild-type parental strain	*HKT2.4* allele	DNA	Correct CI, CM lost due to experimental design
			ten	DNA	CM and 70 more candidates within the CI

[a] *CM, causal mutation.* [b] *Candidate: only protein-altering mutations are considered.* [c] *CI: candidate interval defined by Easymap. See* ***Supplementary Table 1*** *for more detailed information.*

causal mutation. The center of the window defines the center of the mapping region, and a candidate interval of 4, 10, or 20 Mb is set depending on the size of the input genome. The SNPs in the candidate region are then analyzed and reported as candidates to be the mutation causing the phenotype under study. Although Easymap has been designed to map EMS mutations, its mapping report has a link to a file containing all the mutations found in the genome and their potential impact on genes and the products encoded by these genes. This file includes mutations that are not those typically produced by EMS. The Easymap documentation (**Supplementary Data Sheet 2**) contains more detailed information about the SNP selection and mapping algorithms implemented.

Workflow 2: Tagged Sequence Mapping

Easymap uses a tagged-sequence strategy to map the positions of large DNA insertions of known sequence (**Figure 3**). The user has to obtain paired-end (e.g., Illumina-like) or single-end (e.g., Ion Proton-like) NGS reads from a mutant carrying an insertion of partially or completely known sequence. In previous works, we used simulations to assess that a 5× read depth can be sufficient to map most insertions in a sample using the methods that we implemented in Easymap (Wilson-Sánchez et al., 2019); however, as higher read depths directly translate to more reliable results, a minimum read depth of 10× is advisable. Easymap can also use reads from multiple mutants pooled into a single DNA sample, in which case the minimum read depth recommended is 10× per pooled mutant. The program finds reads that overlap the left and right junctions of a given insertion, as well as unpaired alignments neighboring the insertion site; then it uses them as probes against the whole genome sequence (**Figure 3**). Bowtie2 is used as the alignment tool for this workflow, because local alignment of reads is currently not compatible with HISAT2, which favors spliced alignment of reads. If several hits accumulate around a locus, its physical position is reported as a putative insertion site.

If more than one insertional event is detected in a given mutant, the aforementioned experimental design and analysis workflow cannot discriminate the insertion causing the mutant phenotype. Such a mutant, however, can also be crossed and analyzed as described in Workflow 1. The mapping report from Workflow 2 includes a histogram that shows the distribution of the data supporting each putative insertion (**Figure 1D**). The user can inspect the histograms visually to easily discern false positives [disorganized clusters with very low accumulated read depths (RDs)] from genuine insertions (clusters of organized data with a high number of accumulated RDs).

Complementary Tasks

An Easymap run automatically performs several essential tasks such as input data quality controls, including the verification of the FastQ encoding and quality, and assessment of the RD distribution for each sample. After each analysis, Easymap creates a comprehensive report containing high-resolution images and tabular data to assist the user in interpreting the mapping results (e.g., allele frequency plots for EMS-induced mutations, RD histograms for insertional mutations, and gene plots for each putatively damaged gene; **Figure 1D**), a prediction of the functional effect of the candidate mutations, the flanking sequences of each mutation, and the sequences of oligonucleotide primers to genotype such mutation.

Next-generation sequencing experiment simulations can be helpful for optimizing the design of effective mapping experiments (James et al., 2013; Wilson-Sánchez et al., 2019). Therefore, Easymap includes a built-in experiment simulator that

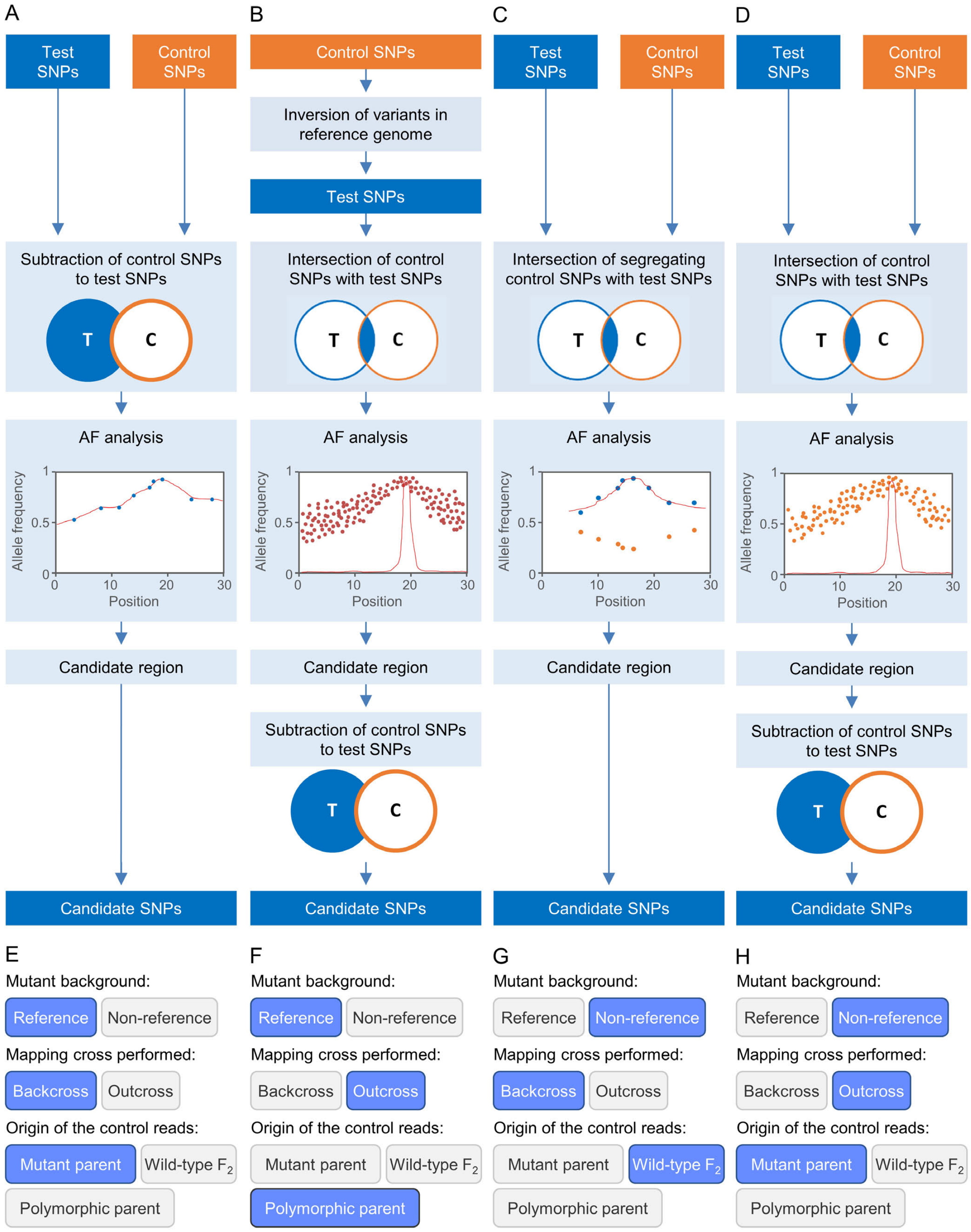

A
Test SNPs
Control SNPs
Subtraction of control SNPs to test SNPs
T
C
AF analysis
Allele frequency
1
0.5
0
0
10
20
30
Position
Candidate region
Candidate SNPs
B
Control SNPs
Inversion of variants in reference genome
Test SNPs
Intersection of control SNPs with test SNPs
T
C
AF analysis
Allele frequency
Position
Candidate region
Subtraction of control SNPs to test SNPs
T
C
Candidate SNPs
C
Test SNPs
Control SNPs
Intersection of segregating control SNPs with test SNPs
T
C
AF analysis
Allele frequency
Position
Candidate region
Candidate SNPs
D
Test SNPs
Control SNPs
Intersection of control SNPs with test SNPs
T
C
AF analysis
Allele frequency
Position
Candidate region
Subtraction of control SNPs to test SNPs
T
C
Candidate SNPs
E
Mutant background:
[✓ Reference] Non-reference
Mapping cross performed:
[✓ Backcross] Outcross
Origin of the control reads:
[✓ Mutant parent] Wild-type F_2 Polymorphic parent
F
Mutant background:
[✓ Reference] Non-reference
Mapping cross performed:
Backcross [✓ Outcross]
Origin of the control reads:
Mutant parent Wild-type F_2 [✓ Polymorphic parent]
G
Mutant background:
Reference [✓ Non-reference]
Mapping cross performed:
[✓ Backcross] Outcross
Origin of the control reads:
Mutant parent [✓ Wild-type F_2] Polymorphic parent
H
Mutant background:
Reference [✓ Non-reference]
Mapping cross performed:
Backcross [✓ Outcross]
Origin of the control reads:
[✓ Mutant parent] Wild-type F_2 Polymorphic parent

FIGURE 2 | Some strategies for EMS-induced mutation mapping implemented in Easymap. **(A–D)** The input reads are processed into control and test SNP lists. The lists are contrasted to determine the SNPs that can be informative for mapping, which are subjected to an allelic frequency (AF) analysis to find the mapping region. A candidate region is defined around the center of the mapping region, and the potentially causal SNPs within the candidate region are collected as candidate SNPs. **(A)** For a mutant strain obtained in the reference genetic background, a backcross is performed to obtain the mapping population and the control sample used is the parental of the mutagenized line. **(B)** For a mutant obtained in the reference genetic background, an outcross is performed to obtain the mapping population, and the control sample is the polymorphic wild-type parent. **(C)** For a mutant obtained in a non-reference strain, a backcross is performed to obtain the mapping population, and the control sample used is a pool of phenotypically wild-type F_2 individuals. **(D)** For a mutant obtained in a non-reference strain, an outcross is performed to obtain the mapping population, and the control sample is the parent of the mutagenized line. **(E–H)** Selection of the experimental design corresponding to panels **(A–D)** in the multiple-choice selectors of the graphic interface of Easymap.

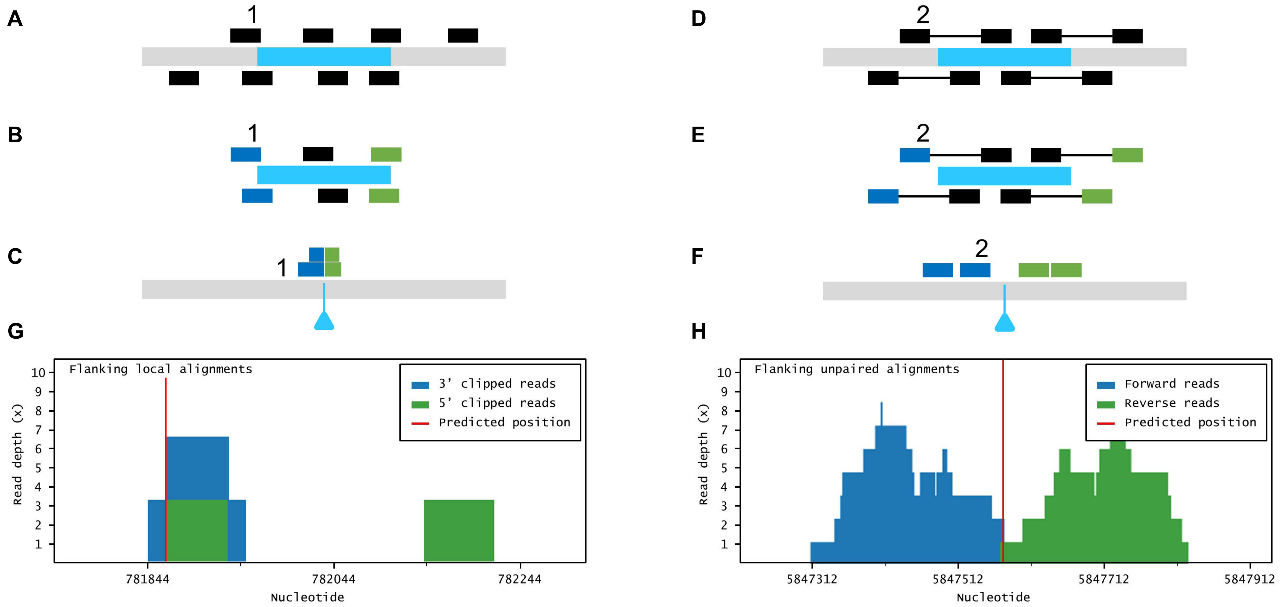


FIGURE 3 | Large insertion mapping with Easymap. **(A–C)** Local alignment analysis. **(A)** The DNA insert appears in blue, over genomic DNA in gray. Individual reads are taken from the mutant genome. **(B)** The reads are aligned to the insertion sequence. Locally aligned reads (e.g., 1) are selected and sorted according to the end that is truncated (in blue and green). **(C)** The selected reads are aligned to the genomic reference sequence. The blue triangle indicates the position of the insertion in the mutant genome. **(D–F)** Paired-read analysis. **(D)** Paired reads are taken from the mutant genome. **(E)** The reads are aligned to the insertion sequence. Unaligned reads with aligned mates (e.g., 2) are selected and sorted according to their position in relation to the insertion (in blue and green). **(F)** The selected reads are aligned to the reference sequence, delimiting a candidate region for the insertion site. **(G,H)** Read depth histograms for examples of local alignment **(G)** and paired-read analyses **(H)**. **(G)** False-positive insertion, characterized by low overall read depths and disorganized data. **(H)** True-positive insertion, characterized by high read depths and organized data.

allows the user to simulate NGS data to test different mapping designs and parameters.

Assessment of Easymap Performance With Simulated and Real Data

We tested Easymap with tens of simulated datasets for each mapping strategy and analysis workflow supported by the program. This allowed us to hard-code appropriate values for analysis variables (e.g., SNP filtering thresholds) under different experimental conditions, saving the user from having to set complex parameters. However, to add more flexibility to the analysis, Workflow 1 allows the user to choose between two levels of stringency for SNP selection. We tested Easymap with data from real mapping-by-sequencing experiments; **Tables 1**, **2** show the results that we obtained when analyzing reads from a range of previously published mutants. We reproduced previously published results, demonstrating the reliability of Easymap even under extreme conditions with average read depths as low as 5× (Obholzer et al., 2012; Wilson-Sánchez et al., 2014). The mapping reports for each of these experiments are available in our preview Easymap installation[1] and additional information about these experiments is provided in **Supplementary Table 1**. Among the data used for testing the Easymap linkage analysis mapping workflows, we employed NGS reads obtained from mutants of *Arabidopsis thaliana* (Morel et al., 2002; Hartwig et al., 2012; Rishmawi et al., 2014; Sun and Schneeberger, 2015; Wachsman et al., 2017; Mateo-Bonmatí et al., 2018), *Zea mays* (Liu et al., 2012; Li et al., 2013; Li L. et al., 2016; Klein et al., 2018), *Caenorhabditis elegans* (Fay and Spencer, 2006), and *Danio rerio* (Obholzer et al., 2012), which included F_2 (Hartwig et al., 2012; Obholzer et al., 2012; Rishmawi et al., 2014; Sun and Schneeberger, 2015; Klein et al., 2018; Mateo-Bonmatí et al., 2018) and M_2/M_3 (Wachsman et al., 2017) mapping populations obtained to identify recessive mutations, as well as dominant mutations mapped in F_2 after an F_3 screening (Fay and Spencer, 2006), and RNA-seq datasets (Liu et al., 2012; Li et al., 2013; Li L. et al., 2016; **Table 1** and **Supplementary Table 1**). Among the mentioned datasets, we included the two experiments used for the validation of the mapping tool SHOREmap v3.0 (Sun and Schneeberger, 2015), a backcross and an outcross approach corresponding to the *hasty* and *soc1* alleles, respectively (**Table 1**). The candidates reported by Easymap are comparable with the ones produced by SHOREmap. With Easymap, we identified two candidate mutations for *hasty*, while SHOREmap reported three (the missing candidate is outside the candidate region delimited by Easymap). Whereas SHOREmap reported two mutations of interest for *soc1-2*, Easymap identified seven mutations within a broader candidate region.

For tagged sequence mapping, we reproduced previous results for *Arabidopsis thaliana* (Wilson-Sánchez et al., 2014; Li W.X. et al., 2016) and *Oryza sativa* (Yang et al., 2013) mutants;

[1] http://atlas.umh.es/genetics/

TABLE 2 | Validation of large-insertion mapping strategies with real experimental data.

References	Sequenced sample	Results obtained by Easymap
Wilson-Sánchez et al. (2014)	Pool of 10 SALK mutants	17 of 19 known insertions were detected by Easymap; the remaining 2 were filtered out as false positives due to very few supporting reads (average RD per sample was 4.5×, instead of the recommended 10×)
Lup and Micol, unpublished	Pool of 6 SALK mutants	9 insertions detected
Yang et al. (2013)	Mutagenized T1c-19 rice	2 of 2 insertions detected; 2 clearly distinguishable false positives
	Mutagenized TT51-1 rice	2 of 2 insertions detected; 47 clearly distinguishable false positives due to the presence of an endogenous sequence in the insertion sequence
Li W.X. et al. (2016)	Mutant T027	2 of 2 insertions detected; 1 false positive common to all lines from this article (omitted below)
	Mutant T182	1 of 1 insertion detected; 1 false positive
	Mutant T204	1 of 1 insertion detected
	Mutant T273	1 of 1 insertion detected split in two clusters due to a large deletion in the mutagenized genome
	Mutant T400	1 of 1 insertion detected

*See **Supplementary Table 1** for more detailed information.*

we also analyzed an unpublished dataset obtained in our laboratory (**Table 2** and **Supplementary Table 1**).

Easymap Architecture

We developed Easymap for UNIX-based operating systems since high-performance machines most commonly run Linux, and several tasks in Easymap are performed by third-party software that has already been extensively tested and used in Linux. These open-source, publicly available programs obtained by previous authors are listed in **Table 3**.

Easymap comprises a software stack consisting of a controller layer, a workflow layer (representing linkage analysis mapping, tagged sequence mapping, and common processes), and a tasks layer (representing custom and third-party programs). The controller exposes a simple application programming interface (API) with which the web and command line scripts interact. This allows Easymap to be installed either locally or remotely while offering simultaneous command-line and graphical interface access (**Supplementary Figure 1**).

To simplify Easymap installation, a single script compiles and/or installs all required software (Python2, Python Imaging Library, Virtualenv, Bowtie2, HISAT2, HTSlib, SAMtools, and BCFtools; **Table 3**) within the Easymap directory. All third-party software is included within the Easymap package so that no dependencies are required. For installations in shared environments, usage (memory and number of concurrent jobs) can be limited by the system administrator through a simple configuration file. The installation script sets up a dedicated HTTP server to run as a background process using the port chosen by the user. Easymap implements chunked file transfers for reliable HTTP transfer of large read files. Further installation setups and usage indications can be found in the Easymap documentation (**Supplementary Data Sheet 2**).

TABLE 3 | Third-party software packages included in Easymap.

Workflow	Software	References	General usage and parameters
Linkage analysis mapping	HISAT2	Kim et al. (2019)	hisat2-build for genome indexing; hisat2 with default options for paired-end or single-end read alignment
	SAMtools	Li et al. (2009)	samtools sort to convert BAM files to SAM; samtools mpileup for first step in variant calling; with arguments: -t DP, ADF, ADR for specific output formatting of VCF file -C50 to fix overestimated mapping qualities (high stringency mode)
	BCFtools	Narasimhan et al. (2016)	bcftools call for second step in variant calling, with argument -mv to report only polymorphic sites
	HTSlib	http://www.htslib.org/	A dependency for BCFtools
Tagged sequence mapping	Bowtie 2	Langmead and Salzberg (2012)	bowtie2-build for genome indexing; bowtie2 with default options for paired-end read alignment; bowtie2 -local for local alignment of paired-end or single-end reads

DISCUSSION

After the advent of NGS, mapping-by-sequencing strategies quickly became the most attractive methods for mutation mapping. Not few software tools have been developed for this purpose, most of which are not easy to manage for researchers without a background in bioinformatics. We attempted to identify the main accessibility issues of such tools and developed Easymap, a program for mapping-by-sequencing that can be used reliably by as many researchers as possible, irrespective of their computer skills, and in as many experimental designs as possible.

The main accessibility features of Easymap are the following: it is free and open source; a single command installs the software and launches the server for the graphical interface; it is easy to use, as it has a graphical interface and workflows that smoothly convert raw data into comprehensive yet simple reports; it is polyvalent, because it can be used for a wide variety of experimental setups (**Table 4**); it is flexible, because it can be

TABLE 4 | Experimental designs supported by different open-source programs used for mapping-by-sequencing.

Mutations under study	Mapping design	Control sample	Software				
			SHOREmap[1] artMAP[2]	**SIMPLE[3]**	**CandiSNP[4]**	**Jitterbug[5] ITIS[6]**	**Easymap[7]**
Point mutations	Backcross	Parental line	D		D		D/R
		Phenotypically wild-type F_2 or M_2		D			D/R
	Outcross	Parental line	D				D/R
		Phenotypically wild-type F_2 or M_2		D			D/R
Large insertions	–	–				D	D

The capabilities of some current mapping tools are compared, D and R indicating an experimental design supported by the software (with DNA-seq or RNA-seq data, respectively). [1]*Sun and Schneeberger (2015).* [2]*Javorka et al. (2019).* [3]*Wachsman et al. (2017).* [4]*Etherington et al. (2014).* [5]*Hénaff et al. (2015).* [6]*Jiang et al. (2015).* [7]*This work.*

installed locally or remotely (on a server) while maintaining its graphical interface (**Supplementary Figures 1, 2**).

The implementation of the HISAT2 aligner allows the use of RNA-seq data for mutation mapping in large genomes, making Easymap the first program of its class to allow performing mapping-by-sequencing with large genomes for which whole genome sequencing may not be affordable. Mapping-by-sequencing approaches with RNA-seq datasets are generally limited to the identification of a candidate region because it is considered likely that the causal mutation of the phenotype under study will not be represented in the sequenced transcriptome, as a consequence of low or null transcription of the mutated gene.

Easymap proved to be reliable under a wide variety of experimental designs, in five different plant (*A. thaliana*, *Z. mays*, and *O. sativa*) and animal (*D. rerio* and *C. elegans*) species and a total of 28 experiments showing unprecedented versatility and adaptability to the input data. While Easymap facilitates mutation mapping, the identification of a causal mutation will inevitably require validation of the candidates via experimental approaches such as allelic complementation tests and phenotypic rescue assays. Furthermore, the linkage analysis workflow will point to the mapping region but will not report any insertions, deletions, or genomic rearrangements. If the causal mutation under study is not a typical point mutation, additional analyses will be required to map such mutations within the candidate region.

In conclusion, here we introduce Easymap, a novel analysis tool for mapping-by-sequencing of large insertions and point mutations, designed to broaden the access to mapping-by-sequencing approaches to a larger number of potential users. Easymap features a web-based graphic interface, a simple installation script, robust mapping analyses for several experimental designs, and thorough user-oriented mapping reports. A preview instance of Easymap is available at http://atlas.umh.es/genetics. This preview contains the mapping reports for all the experiments used during the validation of our program (**Supplementary Table 1**); however, it is not a functional installation of Easymap. The Easymap source code is available for download at https://github.com/MicolLab/easymap and http://genetics.edu.umh.es/resources/easymap. However, we recommend to install the program by following our Quickstart Installation Guide (**Supplementary Data Sheet 1**).

METHODS

Programming Languages and Utilities

Easymap was designed as a modular program, so that each module can be used and modified independently. Modules include custom Python2 scripts and third-party software packages (**Table 3**). Modules are run sequentially by different Bash scripts, or workflows, attending to the user preferences as defined in the web interface or the command line interface.

Python source and libraries are installed within a Virtualenv virtual environment so that previous software installations are not disturbed. The Easymap server is launched using the Python2 CGIHTTPServer function to set up the web interface. The Pillow imaging library is used for the generation of the graphic output.

Testing

We tested Easymap in several operating systems, including different distributions of Linux such as Ubuntu, Fedora, Red Hat, and AMI. Easymap can also run within the Ubuntu apps available in the Windows 10 Microsoft Store. Easymap runs on regular desktop computers and high-performance machines, in local machines and remote instances (e.g., the Amazon Elastic Compute Cloud service), and also in virtual machines running UNIX-based OS within Windows or Mac OS. Appendix D of the Easymap user manual (**Supplementary Data Sheet 2**) includes detailed information for different installation setups.

Easymap's performance obviously depends on the performance of the machine where it is installed, as well as on the size of the genome being analyzed and the size of input reads. A typical Easymap run can take from hours to days, depending on these variables. As a general recommendation, we suggest using machines with a minimum of 4 Gb of RAM and twice the size of all input reads of available disk storage. Analysis of a typical *Arabidopsis* backcross dataset with 50× test and control reads will take around 6 h to complete and use 4 Gb of

RAM memory in a desktop computer. Easymap allows executing multiple projects simultaneously, but this can affect the overall performance of a desktop machine. Easymap does not allow multithreading by default for the sake of simplifying its usage; however, advanced users can find instructions to implement multithreading within the Easymap Documentation.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

JLM obtained funding, provided resources, and supervised the work. DW-S, SDL, and JLM conceived and designed the program and wrote the article. DW-S and SDL developed the program. SA-S contributed a number of Python scripts. SDL tested the software with real datasets. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by grants from the Ministerio de Ciencia e Innovación of Spain [PGC2018-093445-B-I00 (MCI/AEI/FEDER, UE)] and the Generalitat Valenciana (PROMETEO/2019/117), to JLM. SDL held a predoctoral fellowship (ACIF/2018/005) from the Generalitat Valenciana.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2021.655286/full#supplementary-material

Supplementary Figure 1 | Easymap architecture overview.

Supplementary Figure 2 | Easymap graphic interface to design and run a new project.

Supplementary Table 1 | Results obtained in the validation of Easymap using multiple NGS datasets.

Supplementary Data Sheet 1 | Easymap Quickstart Guide.

Supplementary Data Sheet 2 | Easymap documentation.

REFERENCES

Abe, A., Kosugi, S., Yoshida, K., Natsume, S., Takagi, H., Kanzaki, H., et al. (2012). Genome sequencing reveals agronomically important loci in rice using MutMap. *Nat. Biotechnol.* 30, 174–178. doi: 10.1038/nbt.2095

Afgan, E., Baker, D., Batut, B., Van den Beek, M., Bouvier, D., Čech, M., et al. (2018). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.* 46, W537–W544.

Alonso, J. M., Stepanova, A. N., Leisse, T. J., Kim, C. J., Chen, H., Shinn, P., et al. (2003). Genome-wide insertional mutagenesis of *Arabidopsis thaliana*. *Science* 301, 653–657. doi: 10.1126/science.1086391

Candela, H., Casanova-Sáez, R., and Micol, J. L. (2015). Getting started in mapping-by-sequencing. *J. Integr. Plant Biol.* 57, 606–612. doi: 10.1111/jipb.12305

Cooley, L., Kelley, R., and Spradling, A. (1988). Insertional mutagenesis of the *Drosophila* genome with single P elements. *Science* 239, 1121–1128. doi: 10.1126/science.2830671

Ecovoiu, A. A., Ghionoiu, I. C., Ciuca, A. M., and Ratiu, A. C. (2016). Genome ARTIST: a robust, high-accuracy aligner tool for mapping transposon insertions and self-insertions. *Mob. DNA* 7:3.

Etherington, G. J., Monaghan, J., Zipfel, C., and MacLean, D. (2014). Mapping mutations in plant genomes with the user-friendly web application CandiSNP. *Plant Methods* 10:41.

Ewing, A. D. (2015). Transposable element detection from whole genome sequence data. *Mob. DNA* 6:24.

Fay, D., and Spencer, A. (2006). *Genetic Mapping and Manipulation: Chapter 8-Dominant Mutations (February 17, 2006), WormBook*. The *C. elegans* Research Community. Madison, WI: WormBook, doi: 10.1895/wormbook.1.97.1

Fekih, R., Takagi, H., Tamiru, M., Abe, A., Natsume, S., Yaegashi, H., et al. (2013). MutMap+: genetic mapping and mutant identification without crossing in rice. *PLoS One* 8:e68529. doi: 10.1371/journal.pone.0068529

Frøkjær-Jensen, C., Davis, M. W., Sarov, M., Taylor, J., Flibotte, S., LaBella, M., et al. (2014). Random and targeted transgene insertion in *Caenorhabditis elegans* using a modified Mos1 transposon. *Nat. Methods* 11, 529–534. doi: 10.1038/nmeth.2889

Gasch, A., Aoyama, T., Foster, R., and Chua, N.-H. (1992). "Gene isolation with the polymerase chain reaction," in *Methods in Arabidopsis Research*, ed. C. Koncz (Singapore: World Scientific), 342–356. doi: 10.1142/9789814439701_0014

Gonzalez, M. A., Lebrigio, R. F., Van Booven, D., Ulloa, R. H., Powell, E., Speziani, F., et al. (2013). GEnomes Management application (GEM.app): a new software tool for large-scale collaborative genome analysis. *Hum. Mutat.* 34, 842–846. doi: 10.1002/humu.22305

Hartwig, B., James, G. V., Konrad, K., Schneeberger, K., and Turck, F. (2012). Fast isogenic mapping-by-sequencing of ethyl methanesulfonate-induced mutant bulks. *Plant Physiol.* 160, 591–600. doi: 10.1104/pp.112.200311

Hénaff, E., Zapata, L., Casacuberta, J. M., and Ossowski, S. (2015). Jitterbug: somatic and germline transposon insertion detection at single-nucleotide resolution. *BMC Genomics* 16:768. doi: 10.1186/s12864-015-1975-5

Hill, J. T., Demarest, B. L., Bisgrove, B. W., Gorsi, B., Su, Y. C., and Yost, H. J. (2013). MMAPPR: mutation mapping analysis pipeline for pooled RNA-seq. *Genome Res.* 23, 687–697. doi: 10.1101/gr.146936.112

James, D. W. Jr., and Dooner, H. K. (1990). Isolation of EMS-induced mutants in *Arabidopsis* altered in seed fatty acid composition. *Theor. Appl. Genet.* 80, 241–245. doi: 10.1007/bf00224393

James, G. V., Patel, V., Nordstrom, K. J., Klasen, J. R., Salome, P. A., Weigel, D., et al. (2013). User guide for mapping-by-sequencing in *Arabidopsis*. *Genome Biol.* 14:R61.

Jansen, G., Hazendonk, E., Thijssen, K. L., and Plasterk, R. H. (1997). Reverse genetics by chemical mutagenesis in *Caenorhabditis elegans*. *Nat. Genet.* 17, 119–121. doi: 10.1038/ng0997-119

Javorka, P., Raxwal, V. K., Najvarek, J., and Riha, K. (2019). artMAP: a user-friendly tool for mapping ethyl methanesulfonate-induced mutations in *Arabidopsis*. *Plant Direct* 3:e00146. doi: 10.1002/pld3.146

Jiang, C., Chen, C., Huang, Z., Liu, R., and Verdier, J. (2015). ITIS, a bioinformatics tool for accurate identification of transposon insertion sites using next-generation sequencing data. *BMC Bioinformatics* 16:72. doi: 10.1186/s12859-015-0507-2

Kim, D., Paggi, J. M., Park, C., Bennett, C., and Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* 37, 907–915. doi: 10.1038/s41587-019-0201-4

Klein, H., Xiao, Y., Conklin, P. A., Govindarajulu, R., Kelly, J. A., Scanlon, M. J., et al. (2018). Bulked-segregant analysis coupled to whole genome sequencing (BSA-Seq) for rapid gene cloning in maize. *G3* 8, 3583–3592. doi: 10.1534/g3.118.200499

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Li, L., Hey, S., Liu, S., Liu, Q., McNinch, C., Hu, H. C., et al. (2016). Characterization of maize *roothairless6* which encodes a D-type cellulose synthase and controls the switch from bulge formation to tip growth. *Sci. Rep.* 6:34395.

Li, L., Li, D., Liu, S., Ma, X., Dietrich, C. R., Hu, H. C., et al. (2013). The maize *glossy13* gene, cloned via BSR-Seq and Seq-walking encodes a putative ABC transporter required for the normal accumulation of epicuticular waxes. *PLoS One* 8:e82333. doi: 10.1371/journal.pone.0082333

Li, W. X., Wu, S. L., Liu, Y. H., Jin, G. L., Zhao, H. J., Fan, L. J., et al. (2016). Genome-wide profiling of genetic variation in *Agrobacterium*-transformed rice plants. *J. Zhejiang Univ. Sci. B* 17, 992–996. doi: 10.1631/jzus.b1600301

Liu, S., Yeh, C. T., Tang, H. M., Nettleton, D., and Schnable, P. S. (2012). Gene mapping via bulked segregant RNA-Seq (BSR-Seq). *PLoS One* 7:e36406. doi: 10.1371/journal.pone.0036406

Liu, Y. G., Mitsukawa, N., Oosumi, T., and Whittier, R. F. (1995). Efficient isolation and mapping of *Arabidopsis thaliana* T-DNA insert junctions by thermal asymmetric interlaced PCR. *Plant J.* 8, 457–463. doi: 10.1046/j.1365-313x.1995.08030457.x

Mateo-Bonmatí, E., Esteve-Bruna, D., Juan-Vicente, L., Nadi, R., Candela, H., Lozano, F. M., et al. (2018). *INCURVATA11* and *CUPULIFORMIS2* are redundant genes that encode epigenetic machinery components in *Arabidopsis*. *Plant Cell* 30, 1596–1616. doi: 10.1105/tpc.18.00300

Medford, J. I., Behringer, F. J., Callos, J. D., and Feldmann, K. A. (1992). Normal and abnormal development in the *Arabidopsis* vegetative shoot apex. *Plant Cell* 4, 631–643. doi: 10.2307/3869522

Michelmore, R. W., Paran, I., and Kesseli, R. V. (1991). Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. *Proc. Natl. Acad. Sci. U.S.A.* 88, 9828–9832. doi: 10.1073/pnas.88.21.9828

Minevich, G., Park, D. S., Blankenberg, D., Poole, R. J., and Hobert, O. (2012). CloudMap: a cloud-based pipeline for analysis of mutant genome sequences. *Genetics* 192, 1249–1269. doi: 10.1534/genetics.112.144204

Morel, J. B., Godon, C., Mourrain, P., Beclin, C., Boutet, S., Feuerbach, F., et al. (2002). Fertile hypomorphic ARGONAUTE (ago1) mutants impaired in post-transcriptional gene silencing and virus resistance. *Plant Cell* 14, 629–639. doi: 10.1105/tpc.010358

Narasimhan, V., Danecek, P., Scally, A., Xue, Y., Tyler-Smith, C., and Durbin, R. (2016). BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. *Bioinformatics* 32, 1749–1751. doi: 10.1093/bioinformatics/btw044

Neuffer, M. G., and Ficsor, G. (1963). Mutagenic action of ethyl methanesulfonate in maize. *Science* 139, 1296–1297. doi: 10.1126/science.139.3561.1296

Obholzer, N., Swinburne, I. A., Schwab, E., Nechiporuk, A. V., Nicolson, T., and Megason, S. G. (2012). Rapid positional cloning of zebrafish mutations by linkage and homozygosity mapping using whole-genome sequencing. *Development* 139, 4280–4290. doi: 10.1242/dev.083931

O'Malley, R. C., Alonso, J. M., Kim, C. J., Leisse, T. J., and Ecker, J. R. (2007). An adapter ligation-mediated PCR method for high-throughput mapping of T-DNA inserts in the *Arabidopsis* genome. *Nat. Protoc.* 2, 2910–2917. doi: 10.1038/nprot.2007.425

Ponce, M. R., Quesada, V., and Micol, J. L. (1998). Rapid discrimination of sequences flanking and within T-DNA insertions in the *Arabidopsis* genome. *Plant J.* 14, 497–501. doi: 10.1046/j.1365-313x.1998.00146.x

Ponce, M. R., Robles, P., and Micol, J. L. (1999). High-throughput genetic mapping in *Arabidopsis thaliana*. *Mol. Gen. Genet.* 261, 408–415. doi: 10.1007/s004380050982

Pulido-Tamayo, S., Duitama, J., and Marchal, K. (2016). EXPLoRA-web: linkage analysis of quantitative trait loci using bulk segregant analysis. *Nucleic Acids Res.* 44, W142–W146.

Rishmawi, L., Sun, H., Schneeberger, K., Hulskamp, M., and Schrader, A. (2014). Rapid identification of a natural knockout allele of *ARMADILLO REPEAT-CONTAINING KINESIN1* that causes root hair branching by mapping-by-sequencing. *Plant Physiol.* 166, 1280–1287. doi: 10.1104/pp.114.244046

Schneeberger, K., and Weigel, D. (2011). Fast-forward genetics enabled by new sequencing technologies. *Trends Plant Sci.* 16, 282–288. doi: 10.1016/j.tplants.2011.02.006

Smith, D. R. (2015). Buying in to bioinformatics: an introduction to commercial sequence analysis software. *Brief. Bioinform.* 16, 700–709. doi: 10.1093/bib/bbu030

Solaimanpour, S., Sarmiento, F., and Mrázek, J. (2015). Tn-seq explorer: a tool for analysis of high-throughput sequencing data of transposon mutant libraries. *PLoS One* 10:e0126070. doi: 10.1371/journal.pone.0126070

Sun, H., and Schneeberger, K. (2015). SHOREmap v3.0: fast and accurate identification of causal mutations from forward genetic screens. *Methods Mol. Biol.* 1284, 381–395. doi: 10.1007/978-1-4939-2444-8_19

Wachsman, G., Modliszewski, J. L., Valdes, M., and Benfey, P. N. (2017). A SIMPLE pipeline for mapping point mutations. *Plant Physiol.* 174, 1307–1313. doi: 10.1104/pp.17.00415

Wilson-Sánchez, D., Lup, S. D., Sarmiento-Mañús, R., Ponce, M. R., and Micol, J. L. (2019). Next-generation forward genetic screens: using simulated data to improve the design of mapping-by-sequencing experiments in *Arabidopsis*. *Nucleic Acids Res.* 47, 140–143.

Wilson-Sánchez, D., Rubio-Díaz, S., Muñoz-Viana, R., Pérez-Pérez, J. M., Jover-Gil, S., Ponce, M. R., et al. (2014). Leaf phenomics: a systematic reverse genetic screen for *Arabidopsis* leaf mutants. *Plant J.* 79, 878–891. doi: 10.1111/tpj.12595

Yang, L., Wang, C., Holst-Jensen, A., Morisset, D., Lin, Y., and Zhang, D. (2013). Characterization of GM events by insert knowledge adapted re-sequencing approaches. *Sci. Rep.* 3:2839.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.