

1 artMAP: a user-friendly tool for mapping EMS-induced mutations in Arabidopsis

2

3 Peter Javorka^{1,3}, Vivek Raxwal^{2,3,*}, Jan Najvarek¹, Karel Riha^{2,*}

4

5 ¹ ARTIN, Bozetechova 19, Brno, Czech Republic

6 ² CEITEC, Masaryk University, Kamenice 753/5, 625 00 Brno, Czech Republic

7 ³ Equal contribution

8 * Correspondence: vivek.raxwal@ceitec.muni.cz, karel.riha@ceitec.muni.cz

9

10 Abstract

11 Mapping-by-sequencing is a rapid method for identifying both natural as well as induced
12 variations in the genome. However, it requires extensive bioinformatics expertise along with
13 the computational infrastructure to analyze the sequencing data and these requirements have
14 limited its widespread adoption. In the current study, we develop an easy to use tool, artMAP,
15 to discover ethyl methanesulfonate (EMS) induced mutations in the Arabidopsis genome. The
16 artMAP pipeline consists of well-established tools including TrimGalore, BWA, BEDTools,
17 SAMtools, and SnpEff which were integrated in a Docker container. artMAP provides a
18 graphical user interface and can be run on a regular laptop and desktop, thereby limiting the
19 bioinformatics expertise required. artMAP can process input sequencing files generated from
20 single or paired-end sequencing. The results of the analysis are presented in interactive
21 graphs which display the annotation details of each mutation. Due to its ease of use, artMAP
22 made the identification of EMS-induced mutations in Arabidopsis possible with only a few
23 mouse click. The source code of artMAP is available on Github
24 (<https://github.com/RihaLab/artMAP>).

25

26

27 Introduction

28 One of the key driving forces of evolution is *de novo* mutations that randomly occur in a
29 genome, which may become fixed through natural selection or genetic drift. Natural genetic
30 diversity can be used to identify genes responsible for phenotypes of interest either by
31 standard and Quantitative Trait Locus (QTL) mapping or through random genome-wide
32 integration of transgenes and transposons. Mutagenesis is then followed by the selection of
33 mutant lines which exhibit the desired phenotype. Induced mutagenesis generates a much
34 wider range of mutations than occur naturally, as many mutations would be selected against in
35 natural populations. The key advantage of forward genetic screens over reverse genetic
36 approaches, such as targeted gene knock-outs, is their ability to link biological functions to
37 unknown genes in an unbiased manner. Furthermore, in contrast to knock-outs, irradiation
38 and chemical mutagenesis produce a broad range of gene variants with different degrees of
39 functionality, which can be instrumental for studying a gene's regulation and its mechanism of
40 action. For these reasons, forward genetic screens have been successfully applied in a
41 number of model organisms to decipher the biological functions of many genes (Forsburg,
42 2001; Patton & Zon, 2001; Casselton & Zolan, 2002; Jorgensen & Mango, 2002; Page &
43 Grossniklaus, 2002; St Johnston, 2002; Shuman & Silvahy, 2003; Grimm, 2004; Kile & Hilton,
44 2005; Candela & Hake, 2008).

45 While forward genetic screens are one of the most effective approaches for gene
46 discovery, they still require a substantial time commitment and non-negligible monetary
47 investments. While screening for a mutant line with a desired phenotype is often tedious,
48 identification of the causative mutation is usually the main limiting factor in terms of resources,
49 manpower, and time. This process usually involves the generation of mapping populations,
50 which are used to associate a genomic region with the phenotype. The induced mutations in
51 the associated region are then identified and causally linked to the phenotype by
52 complementation tests or through the acquisition of independent alleles. In the pre-genomics
53 era, mapping populations were derived from crosses with a genetically divergent strain that
54 provided genetic markers for association mapping. Association mapping was used to identify
55 the broader region of the genome and then followed by sequencing methods such as
56 chromosome walking to identify the causative mutation. This approach is time-consuming and
57 prone to many limitations, including the density of known polymorphisms in the divergent

58 strain, introgression of unlinked phenotypic modulators, and distribution of meiotic crossovers.
59 With the advent of Next Generation Sequencing (NGS) methodologies, many of these
60 limitations were overcome by the direct sequencing of recombinant mapping populations.
61 Because this approach identifies induced mutations genome-wide, mapping populations can
62 be generated by back-crosses with parental strains using the *de novo* mutations as markers
63 for association mapping (Hartwig *et al.*, 2012; Lindner *et al.*, 2012).

64 Forward genetic screens have been extensively used in Arabidopsis due to its well-
65 annotated genome, self-pollination, and availability of genetic resources (Clouse *et al.*, 1996;
66 Yin *et al.*, 2002; Manavella *et al.*, 2012; Berardini *et al.*, 2015). A genetic screen in Arabidopsis
67 begins with the mutagenesis of seeds (M_0), usually by EMS, followed by screening self-
68 pollinated M_1 or M_2 plants for the phenotype of interest. Dominant mutations exhibit their
69 phenotype in the M_1 generation, whereas recessive mutations are scored in M_2 (**Figure 1**). For
70 both dominant and recessive mutations, M_2 plants are either crossed with another Arabidopsis
71 ecotype or back-crossed to the parental strain to produce recombinant mapping populations.
72 The pool of plants displaying the desired phenotype is sequenced, providing the location of the
73 associated genomic region and a set of candidate mutations. This approach greatly reduces
74 the time and resources required to identify the causal mutation and also circumvents the
75 dependence on genetic markers.

76 Two major bottlenecks in modern forward genetic screens are the high cost of NGS and
77 the complexity of analyzing high throughput sequencing data. With the price of NGS falling
78 continuously, data analysis remains the major bottleneck. While various pipelines have been
79 developed to analyze sequencing data generated from forward genetic screens (Schneeberger
80 *et al.*, 2009; Austin *et al.*, 2011; Minevich *et al.*, 2012; Wachsman *et al.*, 2017), they all require
81 additional computational infrastructure along with bioinformatics expertise. Recently, SIMPLE
82 was introduced to facilitate the analysis of forward genetic data, but even this method requires
83 a certain level of bioinformatics understanding (Wachsman *et al.*, 2017). To date, there is no
84 open source tool available which can be used by a biologist with no bioinformatics expertise.
85 To fill this void, we developed an easy to use tool with a graphical user interface, artMAP,
86 which can be used without any bioinformatics expertise to map EMS-induced mutations in
87 Arabidopsis and assess their association with the desired phenotype.

88

89 **Result and Discussion**

90 **Description of artMAP**

91 The artMAP pipeline consists of various open sources tools integrated into a docker container
92 (<https://www.docker.com/>) to provide a graphical user interface (GUI) and the ability to run on
93 all the three computer platforms (Windows/Mac/Linux). The pipeline is presented in **Figure 2**.
94 Integrating any sequencing analysis pipeline into a single GUI faces several technical
95 challenges as open source tools differ in their programming language, have multiple
96 dependencies, may produce incompatibility issues when brought together. Moreover, the
97 availability of a wide variety of sequencing platforms and format types increases the
98 complexity. artMAP overcomes these issues by using five open source bioinformatics tools
99 (SAMtools, BEDTools, BWA, Trimgalore, and SnpEff) along with in-house scripts to enable the
100 analysis of all possible sequencing data types for EMS based genetic screens. Briefly, the
101 artMAP pipeline consists of 6 steps: 1) pre-processing of the sequencing read files by
102 Trimgalore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/), 2) alignment of
103 reads to the Arabidopsis genome by BWA (Li & Durbin, 2009), 3) post-processing of aligned
104 reads by SAMtools (Li *et al.*, 2009), 4) identification of single nucleotide polymorphisms (SNPs)
105 specific to mutant samples through the combined use of SAMtools and BEDTools suite
106 (Quinlan & Hall, 2010), 5) Visualization of the SNPs, 6) annotation of SNPs by SnpEff
107 (Cingolani *et al.*, 2012). Finally, artMAP provides a list of SNPs along with their allele
108 frequency, depth, and annotation in a tab separated file.

109 In forward genetic screens, whole genome sequencing read files are generated from
110 two biological samples, one treated with EMS (mutant) and a control (usually the parent). The
111 Illumina sequencing platform is preferred for re-sequencing applications, like mapping
112 mutations, owing to its low error rate. An important parameter in NGS is the read length (in
113 base pairs) and type of sequencing (single-end vs paired-end). Paired-end sequencing and a
114 longer read length is often recommended for accurately mapping mutations, especially from
115 the repetitive region of the genome. In most cases, however, single-end sequencing with high
116 depth also enables mapping mutations with a single nucleotide resolution (James *et al.*, 2013;
117 Wachsman *et al.*, 2017). artMAP can process both paired- and single-end sequencing reads
118 regardless of their length. artMAP requires data in BAM or FASTQ formats, which are usually

119 used by Illumina and other sequencing platforms. The data are first processed by Trimgalore,
120 which remove sequencing adapters as well as bad quality sequences from the reads. Since
121 Trimgalore process only FASTQ files, sequencing reads provided in the BAM format are
122 converted to FASTQ using the bam2fastq function of BEDTools.

123 Next, high-quality sequencing reads from the mutant and parent samples are aligned to
124 the Arabidopsis reference genome by BWA. The major advantage of BWA is the ability to align
125 both short and long reads. artMAP include “BWA aln” as well as “BWA mem” for aligning
126 shorter and longer reads, respectively, and artMAP can choose the appropriate aligner based
127 on the length of the input reads. Also, as BWA requires an index of the genome to run,
128 artMAP includes a pre-built BWA index of the Arabidopsis reference genome, eliminating the
129 need to generate a genome index on every run. BWA not only reports the location of the
130 sequencing reads but also records mismatches present between the genome and sequencing
131 reads. This information is later exploited to identify SNPs. The results of the alignment from
132 both control and mutant files are stored in a user-provided location in the SAM format.
133 However, storing and downstream processing of these files are computationally inefficient as
134 they require a large amount of operational memory (RAM) as well as storage space. To
135 increase the computational efficiency, these aligned files are converted to a binary format
136 (BAM). Handling and processing BAM files are computationally less demanding than SAM
137 files. SAMtools is used to convert SAM to BAM files, which are then further sorted according to
138 chromosome number and the genomic location of aligned reads (Li *et al.*, 2009). artMAP
139 provides an additional option to control the removal of PCR duplicates from the control and
140 mutant BAM files. This step is turned on by default but can be disabled if required. Control and
141 mutant BAM files are indexed to prepare for SNP calling with SAMtools. The BAM alignment
142 files generated in this step can be viewed in genome browsers such as IGV (Robinson *et al.*,
143 2011) for further analysis.

144 SNPs are identified from both the control and mutant BAM files generated in the
145 previous step. These SNPs includes all single nucleotide changes present in the control and
146 mutant samples. As EMS induces G to A or C to T transitions in DNA, artMAP only retains
147 these SNPs. EMS induces a plethora of mutations in the genome, including both causative and
148 non-causative SNPs. In the mapping population, the frequency of causative SNPs along with

149 the surrounding linked SNPs is higher (approaching 100 % for a recessive mutation) compared
150 to non-causative SNPs. artMAP applies two filter criteria to remove background
151 polymorphisms that may occur as technical errors. First, artMAP removes SNPs with a
152 frequency lower than 30%. Second, it applies a depth filter to retain SNPs with a sequencing
153 depth between 10-100X. Since these filters can greatly affect the final outcome of the analysis,
154 they can be changed in additional settings.

155 Next, to identify the region of the genome associated with the phenotype of interest,
156 artMAP compares the SNPs present in the control and mutant sample and extracts SNPs
157 exclusive to the mutant. These SNPs are then annotated using SnpEff (Cingolani *et al.*, 2012)
158 to describe their impact on gene structure and amino acid changes. Nonsense mutations
159 producing a stop codon are considered high impact SNPs. Further details regarding the SnpEff
160 annotations can be obtained at <http://snpeff.sourceforge.net/SnpEff.html#intro>. artMAP
161 displays the final results as a graph, where the frequency and position of each SNP are plotted
162 along each Arabidopsis chromosome. These graphs can be zoomed in and can also be saved.
163 Hovering the cursor over SNP reveals key information such as the location, frequency,
164 affected gene, protein and DNA level changes, and the predicted the impact of the SNP. This
165 visualization of the data facilitates a rapid assessment of the results and identification of the
166 region associated with the phenotype.

167 Finally, artMAP provides results as a tab-delimited file with information containing the
168 location of each SNP (chromosome number and position), reference base, mutated base,
169 coverage over the base (depth), frequency, gene identity, and effect on protein change if any.
170 Based on the graph and tab-delimited file, a user can identify the putative candidate gene for
171 further testing. artMAP also produces the raw file at each stage of the pipeline, in case it is
172 required. The detailed description of how to install and run artMAP is provided in
173 supplemented User Manual.

174 **Implementation of artMAP to map EMS-induced mutations**

175 First, we assessed the feasibility of mapping recessive mutations with artMAP. For this,
176 we took data generated from the forward genetic screen for leaf hyponasty mutants (Allen *et al.*, 2013) where the recombinant mapping population (BC₁F₂) was produced by crossing the
177 M2 plant with non-treated parent followed by one round of self-crossing. Since this screen is
178

179 based on bulk segregation analysis of a recessive trait, the causal SNP should be present with
180 a frequency of 100% and surrounded by a collection of linked, high-frequency SNPs. Unlinked
181 SNPs should have a frequency of 50% as expected for the random inheritance of a
182 heterozygous SNP within a population. We unambiguously identified a region linked to the
183 phenotype on chromosome 3 with high-frequency mutations (**Figure 3**). This included a
184 mutation in *HST1* that results in a stop codon at position 451 (Trp451*). This mutation was
185 previously considered the causative mutation in this screen (Allen *et al.*, 2013).

186 Next, we assessed the performance of artMAP compared to a previously published
187 pipeline. For this, we re-analyzed previously published datasets (Wachsman *et al.*, 2017) with
188 SIMPLE (Wachsman *et al.*, 2017) as well as artMAP. This dataset included sequencing reads
189 generated from single- as well as paired-end data. As expected, artMAP was able to
190 accurately map the previously reported causative mutation or those identified by SIMPLE. The
191 list of datasets used and results comparing artMAP and SIMPLE are presented in **Table 1**. It is
192 important to note that while SIMPLE reports the list of likely candidate mutations, artMAP
193 allows the user to interactively browse through graphs displaying the frequencies of individual
194 mutations along the chromosomes, enabling the user to quickly define the linked region and
195 assess whether a mutation may be causative. Also, artMAP displays annotation details and
196 predicted SNP impact directly on the graph. This ability to easily manually assess mutations
197 increases the probability of identifying the actual, causative mutation.

198 **Conclusion**

199 We have developed an interactive tool, artMAP, to map EMS-induced mutations in
200 forward genetic screens in *Arabidopsis thaliana*. artMAP can easily be operated by
201 researchers without any prior expertise in bioinformatics and we demonstrate that the accuracy
202 of artMAP is similar to standard bioinformatics pipelines used to map EMS-induced mutations.
203 It can be run on regular desktops or laptops and does not require extra computational
204 infrastructure. Thus, artMAP greatly facilitates the identification of new mutations in forward
205 genetic screens in *Arabidopsis*, and this tool can easily be adapted for other organisms, if
206 needed.

207

208 **Acknowledgment**

209 The authors are grateful for financial support by the Ministry of Education, Youth and Sports of
210 the Czech Republic, European Regional Development Fund-Project „REMAP“ (No.
211 CZ.02.1.01/0.0/0.0/15_003/0000479) and by the Czech Science Foundation (16-18578S).

212

213 **Figure legends:**

214 **Figure 1:** Schematic representation of forward genetic screens in Arabidopsis, showing the
215 strategy for mapping dominant and recessive mutants

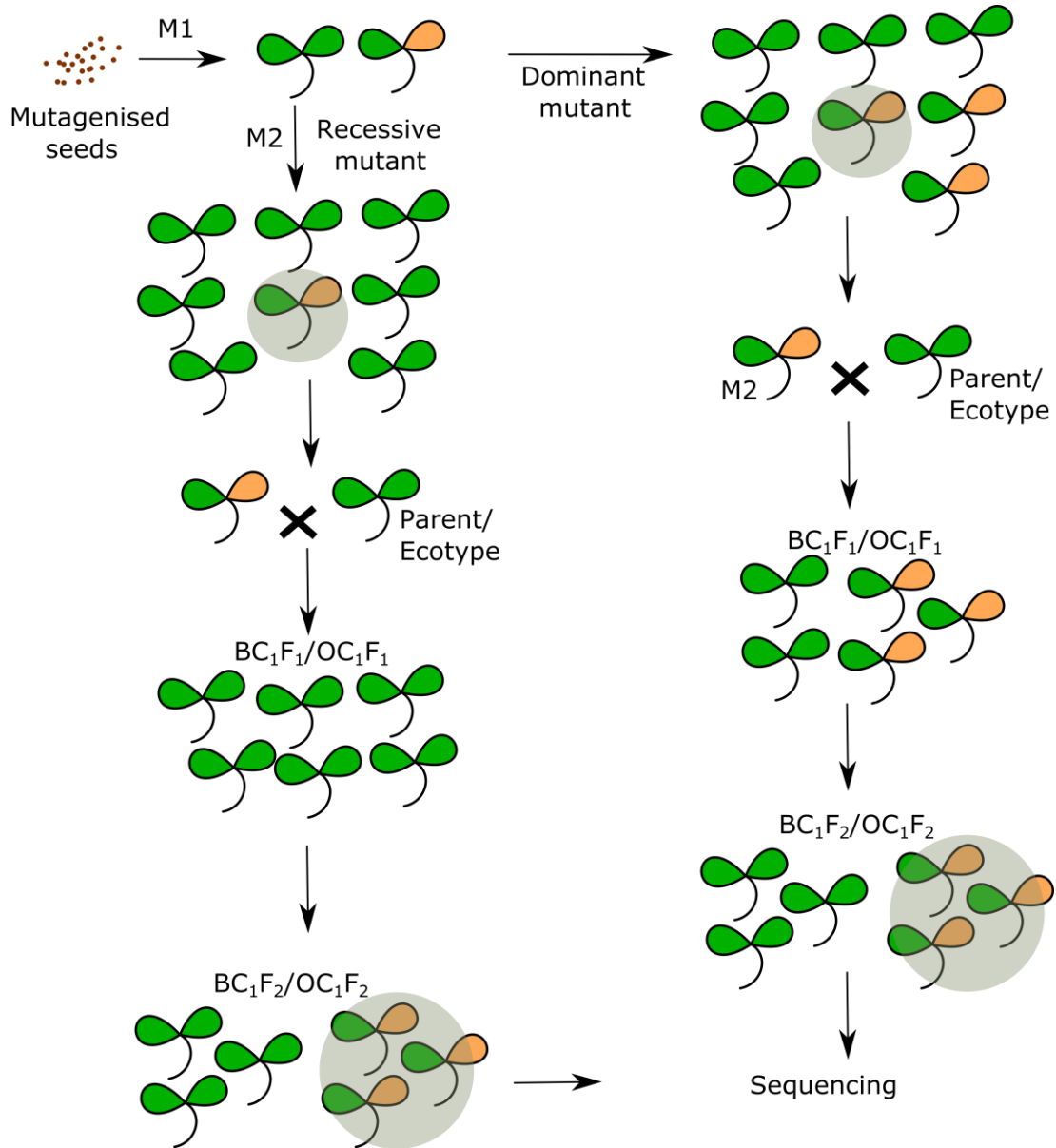
216 **Figure 2:** An outline of the artMAP pipeline showing

217 **Figure 3:** A representative figure of the example run showing the output of the artMAP
218 analysis

219

220

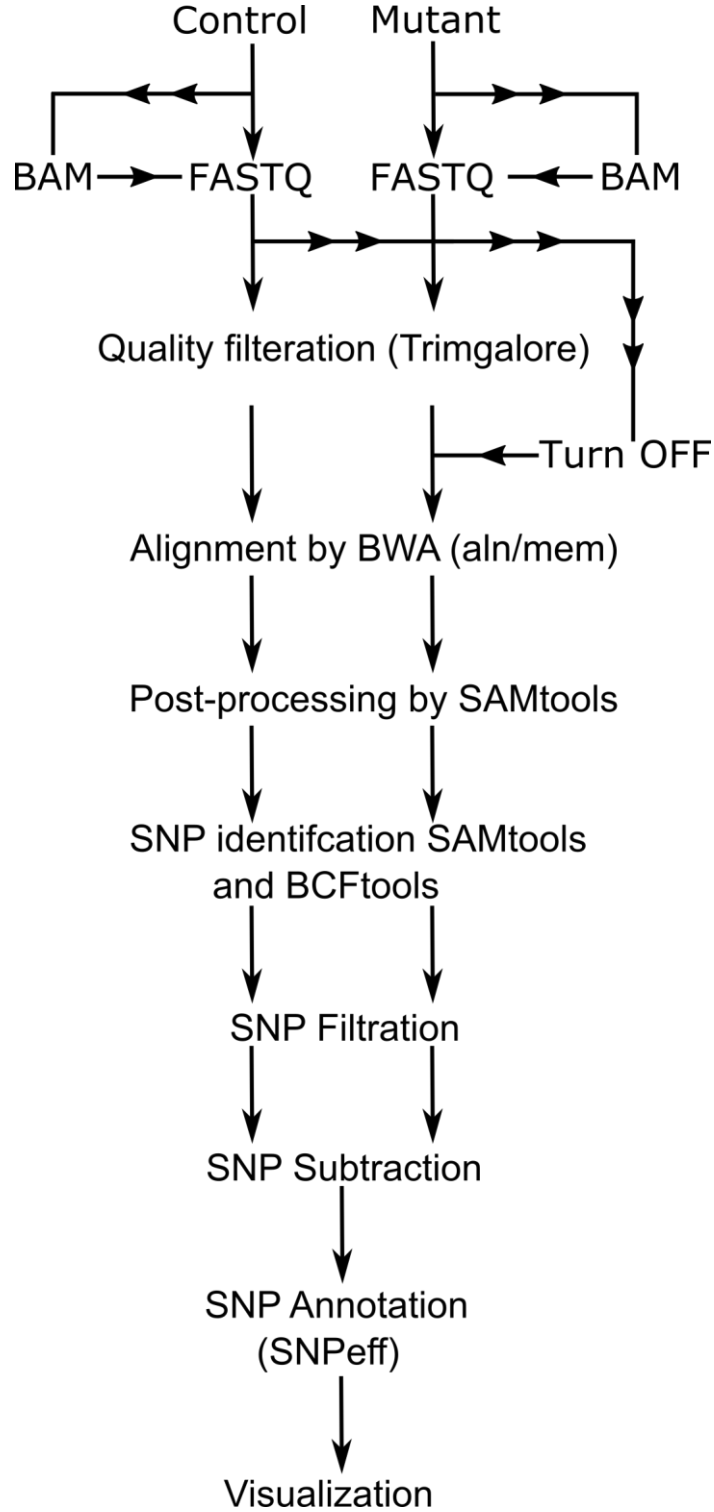
221 **Figure 1:**



222

223

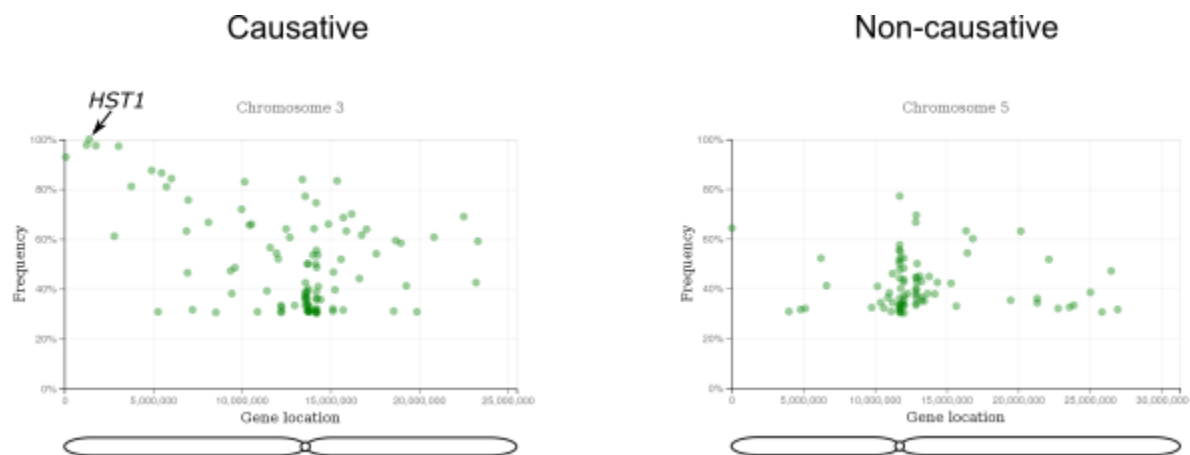
224 **Figure 2:**



225

226

227 **Figure 3:**



228

229

230 **Table 1:** List of datasets used to compare SIMPLE and artMAP

Line	Reported gene	Reported mutation	Mapped by SIMPLE	Mapped by artMAP
300	AT3G13870	Ser584Phe	+	+
300-4	AT3G13870	Ser584Phe	+	+
300-7	AT4G01800	Arg752*	+	+
EMS608	AT5G24630	Gly324Glu	+	+
EMS633	AT3G54660	Ala264Thr	+	+

231

232

233 **References**

- 234 **Allen RS, Nakasugi K, Doran RL, Millar AA, Waterhouse PM. 2013.** Facile mutant
235 identification via a single parental backcross method and application of whole genome
236 sequencing based mapping pipelines. *Frontiers in Plant Science* **4**.
- 237 **Austin RS, Vidaurre D, Stamatiou G, Breit R, Provart NJ, Bonetta D, Zhang J, Fung P,**
238 **Gong Y, Wang PW, et al. 2011.** Next-generation mapping of Arabidopsis genes. *Plant Journal*
239 **67**: 715–725.
- 240 **Berardini TZ, Reiser L, Li D, Mezheritsky Y, Muller R, Strait E, Huala E. 2015.** The
241 arabidopsis information resource: Making and mining the ‘gold standard’ annotated reference
242 plant genome. *Genesis* **53**: 474–485.
- 243 **Candela H, Hake S. 2008.** The art and design of genetic screens: Maize. *Nature Reviews*
244 *Genetics* **9**: 192–203.
- 245 **Casselton L, Zolan M. 2002.** The art and design of genetic screens: Filamentous fungi.
246 *Nature Reviews Genetics* **3**: 683–697.
- 247 **Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM.**
248 **2012.** A program for annotating and predicting the effects of single nucleotide polymorphisms,
249 SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly* **6**: 80–
250 92.
- 251 **Clouse SD, Langford M, McMorris TC. 1996.** A Brassinosteroid-Insensitive Mutant in
252 *Arabidopsis thaliana* Exhibits Multiple Defects in Growth and Development. *Plant Physiology*
253 **111**: 671–678.
- 254 **Forsburg SL. 2001.** The art and design of genetic screens: yeast. *Nature reviews. Genetics* **2**:
255 659–668.
- 256 **Grimm S. 2004.** The art and design of genetic screens: Mammalian culture cells. *Nature*
257 *Reviews Genetics* **5**: 179–189.
- 258 **Hartwig B, James G V., Konrad K, Schneeberger K, Turck F. 2012.** Fast Isogenic Mapping-
259 by-Sequencing of Ethyl Methanesulfonate-Induced Mutant Bulks. *PLANT PHYSIOLOGY* **160**:

- 260 591–600.
- 261 **James GV, Patel V, Nordström KJV, Klasen JR, Salomé PA, Weigel D, Schneeberger K.**
262 **2013.** User guide for mapping-by-sequencing in Arabidopsis. *Genome Biology* **14**.
- 263 **Jorgensen EM, Mango SE. 2002.** The art and design of genetic screens: Caenorhabditis
264 elegans. *Nature Reviews Genetics* **3**: 356–369.
- 265 **Kile BT, Hilton DJ. 2005.** The art and design of genetic screens: Mouse. *Nature Reviews*
266 *Genetics* **6**: 557–567.
- 267 **Li H, Durbin R. 2009.** Fast and accurate short read alignment with Burrows-Wheeler
268 transform. *Bioinformatics* **25**: 1754–1760.
- 269 **Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin**
270 **R. 2009.** The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079.
- 271 **Lindner H, Raissig MT, Sailer C, Shimosato-Asano H, Bruggmann R, Grossniklaus U.**
272 **2012.** SNP-ratio mapping (SRM): Identifying lethal alleles and mutations in complex genetic
273 backgrounds by next-generation sequencing. *Genetics* **191**: 1381–1386.
- 274 **Manavella PA, Hagmann J, Ott F, Laubinger S, Franz M, MacEk B, Weigel D. 2012.** Fast-
275 forward genetics identifies plant CPL phosphatases as regulators of miRNA processing factor
276 HYL1. *Cell* **151**: 859–870.
- 277 **Minevich G, Park DS, Blankenberg D, Poole RJ, Hobert O. 2012.** CloudMap: A cloud-based
278 pipeline for analysis of mutant genome sequences. *Genetics* **192**: 1249–1269.
- 279 **Page DR, Grossniklaus U. 2002.** The art and design of genetic screens: Arabidopsis thaliana.
280 *Nature Reviews Genetics* **3**: 124–136.
- 281 **Patton EE, Zon LI. 2001.** The art and design of genetic screens: zebrafish. *Nature reviews.*
282 *Genetics* **2**: 956–66.
- 283 **Quinlan AR, Hall IM. 2010.** BEDTools: A flexible suite of utilities for comparing genomic
284 features. *Bioinformatics* **26**: 841–842.
- 285 **Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov**
286 **JP. 2011.** Integrative genomics viewer. *Nature Biotechnology* **29**: 24–26.

- 287 **Schneeberger K, Ossowski S, Lanz C, Juul T, Petersen AH, Nielsen KL, Jørgensen JE,**
288 **Weigel D, Andersen SU. 2009.** SHOREmap: Simultaneous mapping and mutation
289 identification by deep sequencing. *Nature Methods* **6**: 550–551.
- 290 **Shuman H, Silvahy T. 2003.** The art and design of genetic screens: Escherichia coli. *Nature*
291 *reviews. Genetics* **3**: 176–88.
- 292 **St Johnston D. 2002.** The art and design of genetic screens: Drosophila melanogaster.
293 *Nature Reviews Genetics* **3**: 176–188.
- 294 **Wachsman G, Modliszewski JL, Valdes M, Benfey PN. 2017.** A SIMPLE Pipeline for
295 Mapping Point Mutations. *Plant Physiology* **174**: 1307–1313.
- 296 **Yin Y, Wang ZY, Mora-Garcia S, Li J, Yoshida S, Asami T, Chory J. 2002.** BES1
297 accumulates in the nucleus in response to brassinosteroids to regulate gene expression and
298 promote stem elongation. *Cell* **109**: 181–191.
- 299