Title: A SIMPLE pipeline for mapping point mutations

running title: Point mutations mapping pipeline

Guy Wachsman[1], Jennifer L. Modliszewski[2], Manuel Valdes[1] and Philip N. Benfey[1,3]

[1]Department of Biology, Duke University, Durham, NC 27708, USA; [2]UNC, Genome Sciences Building 4144, 250 Bell Tower Drive, Chapel Hill, NC, USA; [3]Howard Hughes Medical Institute, Duke University, Durham, NC 27708, USA

Corresponding author: philip.benfey@duke.edu

Keywords: SNP, mapping

Abstract

A forward genetic screen is one of the best methods for revealing the regulatory functions of genes. In plants, this technique is highly efficient since it is relatively easy to grow and screen the phenotypes of hundreds or thousands of individuals. The cost-efficiency and ease of data production afforded by next-generation sequencing techniques have created new opportunities for mapping induced mutations. The principles of genetic mapping remain the same. However, the details have changed, which allows for rapid mapping of causal mutations. Current mapping tools are often not user-friendly and complicated or require extensive preparation steps. To simplify the process of mapping new mutations, we developed a pipeline that takes next generation sequencing fastq files as input, calls on several well-established and freely available genome-analysis tools, and outputs the most likely causal DNA change(s). The pipeline has been validated in Arabidopsis and can be readily applied to other species.

Introduction

Identifying genetic mutations that underlie phenotypic changes is essential for understanding a wide variety of biological processes. A forward genetic screen is one of the most powerful tools for searching for such mutations. Spontaneous and induced mutations have been used to identify genes underlying aberrant phenotypes for over one hundred years (Morgan 1910). In the common case, a mutagen is used to generate a few thousand random mutations in the genome by physical (radiation; (Muller 1927)), chemical (EMS; (Lewis and Bacher 1968)) or biological (transposons; (McClintock 1950)) agents followed by a screen for the desired phenotype caused by one of the mutations. Once a plant with the desired phenotype is isolated, the researcher must identify the causal mutation. This is done by testing for an association between known genetic markers and the phenotype. Such genetic linkage (assessed by co-segregation) indicates that the causal mutation is located in the vicinity of the genetic marker. The introduction of next-generation sequencing (NGS) for mapping purposes has proven to be very promising, as it is possible to quickly identify a small number of potential causal SNPs. However, the currently available tools e.g., SNPtrack (Leshchiner et al. 2012), SHOREmap (Schneeberger et al. 2009) and NGM (Austin et al. 2011) are either inoperable (SNPtrack) or require coding knowledge. We have developed the SIMPLE tool (SImple Mapping PipeLinE), which operates on the input of the NGS fastq files generated from WT and mutant bulked DNA pools, and produces tables and a plot showing the most likely candidate genes and locations, respectively. The tool can be easily downloaded and executed with no prior bioinformatics knowledge and requires only a few simple preparatory steps to initiate. Once the program runs, the user accesses a table with the most

likely candidate genes and a figure file that marks the location(s) of these candidates. Our pipeline has several advantages in comparison to other mapping tools. First, the entire process is user-friendly. It does not require any prior programming knowledge or NGS analysis skills. Second, it is "all-inclusive"; besides a few initial steps such as downloading the fastq files to the right folder and naming them, the user only needs to paste three lines into the terminal application to run the program. These steps are described in the README file. Third, the program can accept as input any paired-end or single-end fastq combination from the WT and mutant bulks. Fourth, this tool accepts, based on our experience, several types of segregating populations such as M2, M3, back-cross and map-cross. The project is currently hosted on GitHub and is available for download at https://github.com/wacguy/Simple.

Results

The concept behind the mapping tool

All methods that aim to map a causal DNA change implement a similar strategy. For bulk-segregant analysis, a segregating F2 population is divided into mutant phenotype and WT phenotype. DNA from each of the two bulked samples is sequenced by next-generation sequencing. For recessive mutations, the basic principle is to find the DNA changes that have only non-reference reads (i.e., locations that differ from the reference genome) in the mutant bulk and ~1:2 mutant to WT ratio of reads in the WT bulk (since the WT bulk is comprised of +/- and +/+ individuals in a 2:1 ratio). For dominant mutations, the concept is similar, but the non-segregating (with a reference linked locus) bulk would be the WT batch (+/+) whereas the mutant bulk is comprised of mixed genotypes of mutant individuals (+/- and -/-). Our pipeline is based

on a short BASH script that calls BWA (Li and Durbin 2009), Samtools (Li et al. 2009), Picard (http://broadinstitute.github.io/picard), GATK (McKenna et al. 2010; DePristo et al. 2011; Van der Auwera et al. 2013), SnpEff (Cingolani et al. 2012) and R (R Core Team 2013) in order to generate two VCFs (variant call files) and a plot. The first VCF lists all candidate genes and the second one contains the entire SNP population found by the GATK HaplotypCaller tool. Generation of the candidate list is based on several criteria. First, and most importantly, the mutation has to segregate in the correct ratio. Additionally, we chose to focus on SNPs that are consistent with a specific set of criteria to identify the most likely causal mutation. However, these criteria can be readily changed by manipulating the simple.sh file. For example, we select only single nucleotide changes, since single nucleotide mutagens such as EMS are commonly used. We also select for mutations that have a significant effect on the protein rather than synonymous mutations or changes in intergenic regions. Figure 1a shows a flowchart of the pipeline concept.

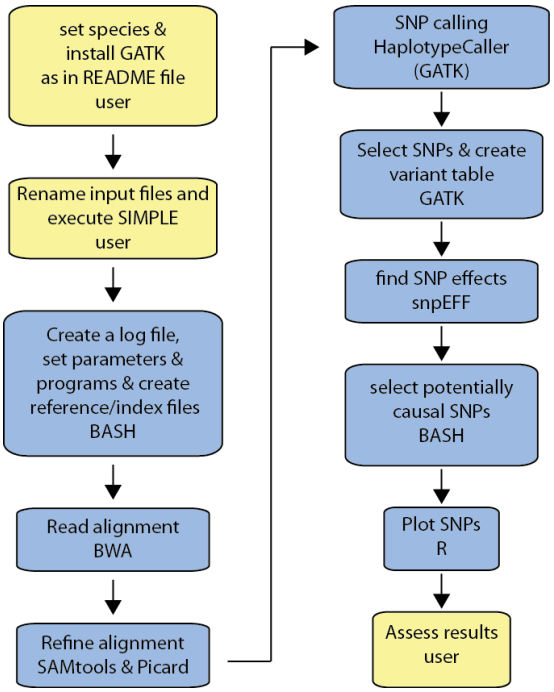


Figure 1. The SIMPLE pipeline workflow

User-required actions are in yellow and SIMPLE actions are in blue. The specific program that executes the action(s) of SIMPLE is indicated at the bottom of each blue box.

To test our pipeline, we sequenced ten populations that were generated in three independent EMS screens. We used four different population types including M2 (a segregating population generated by selfing a heterozygous plant from a mutagenized population), M3 (a segregating population generated by selfing a heterozygous M2 plant), back-cross (an F2 segregating population generated by crossing a mutant plant with the original parental line) and map-cross (an F2 segregating population generated by crossing a mutant plant with another accession, e.g., *L. erecta*). Some of the mutants were sequenced more than once or in successive generations, if they had more than one mapping population or if initial mapping failed (see Table 1 for details).

Table 1. Mapping populations that were used to test the SIMPLE pipeline

line[1]	generation	mapped gene	At_num	mut/wt seedlings	CDS changes	protein change[2]	number of reads (mut.ref/mut.alt ; wt.ref/wt.alt)	mean coverage (mut/wt)	validation[3]	remarks
474-3	M3	*SHR*	AT4G37650	23/23	664C>T	Arg222*	0/43;38/22	73/75	allelic test & similar phenotype to known allele	
300	M2	*RHD3*		19/19	1751C>T		0/13;10/0	21/25	allelic test & similar phenotype to known allele	
300-4	M3		AT3G13870	50/100	1751C>T	Ser584Phe	0/64;76/42	89/156		
300-7	M3	*SECA1*		50/100	2254C>T		0/19;30/9	23/61	allelic test & similar phenotype to known allele	
300-7	M3		AT4G01800	55/50	2254C>T	Arg752*	1/16;110/41	90/198		
633	M2	*GLUTATHIONE REDUCTASE 2*	AT3G54660	30/30	790G>A	Ala264Thr	1/33;38/9	61/57	similar phenotype to known allele (Yu et al., 2013)	couldn't be mapped with M3 population
M381	F2-map cross on Ler	*MYB36*		50/100			0/21;14/0	32/21	complementation & similar phenotype to known allele	wt plants from Ler parental line
B381	F2-back-cross on Col-0; CASP1::GFP		AT5G57620	50/100	174G>A	Trp58*	0/20;60/1	31/83		wt plants from Col-0;CASP1::GFP parental line
EMS608	M3	*BIN4*	AT5G24630	47/150	971G>A	Gly324Glu	1/20;18/1	46/39	similar phenotype to known allele (Breuer et al., 2007)	
194	M2	*RHD3*	AT3G13870	23/49	1951C>T	Gln651*	10/22;49/6	23/49	similar phenotype to known allele	

Column 1 denotes the line number. Line 300-4 was generated by selfing a heterozygous plant from line 300. [1]Each color represent a different EMS mutagenizing experiment; [2]asterisk indicates a stop codon; [3]see methods for more information.

The pipeline produces three files, cands4.txt, plot4.txt and Rplot.pdf (see details below) to help identify the causal mutation. In most cases, the candidate list output file (cands4.txt files) together with the position to SNP-ratio plot were sufficient to identify the causal mutation (Fig. 2a). In line 194 no genes were found in the candidate list, most likely due to an erroneous inclusion of WT seedlings in the mutant bulk. Nevertheless, the program provides additional information to help identify strong candidates. For example, the plot indicates a linked locus near the center of chromosome 3 (Fig. 2b). Browsing the SNP population in this region identifies a premature stop codon in *RHD3* (*AT3G13870*). Indeed, this line has a short root and wavy root hair phenotype similar to the mapped M2-300/M3-300-4 line and to *rhd3-1* (Wang et al. 1997; 2015).

There are two parameters that appear to strongly influence the success of the pipeline in identifying the causal mutation or at least a short list of potential SNPs. First, incorrect inclusion of WT plants in the mutant bulk leads to reference reads in the mutant fastq file. As a result, the causal SNP and linked mutations are viewed as heterozygous. Therefore, it is essential that all individuals in the mutant bulk be correctly phenotyped as mutant. In cases where the phenotype is difficult to recognize, for example, as often happens with QTLs, it is recommended to work with

smaller but high confidence populations rather than include potentially wild-type individuals in the mutant bulk. Alternatively, mutant plants can be tested based on the segregation of their offspring. A second crucial parameter is sequencing depth. This is important because different genomic regions have variable coverage depth and even when mean coverage appears to be sufficient, some regions may still have very few reads, which renders them almost impossible to genotype. We recommend a minimum of 30x coverage.

We recommend working with an F2/M2 generation rather than an F3/M3 generation for two reasons. First, a segregating F3/M3 population generated from a single F2/M2 heterozygous plant is homozygous for one quarter of the loci as well as being homozygous for the causal SNP whereas an F2/M2 population is homozygous only for the causal mutation and the genetically linked region. Therefore, there is substantial information loss in the F3/M3 generation and mapping is heavily dependent on the segregation of the WT bulk. This bulk is less informative because its 2:1 WT to mutant ratio in the binomial read-distribution has a higher potential to vary from the expected values. In contrast, the mutant bulk is expected to have strictly no reference reads and a few dozen alternate reads. Another reason to prefer the F2/M2 generation is that the following generation goes through a second round of recombination that generates chromosomes and chromosomal regions with more complex haplotypes that are more difficult to interpret.

Input files

The user provides the fastq files that are generated by a next-generation sequencing platform such as Illumina 2000 and places them in the fastq folder. Each bulk (mutant or WT) can have

either a single file in the case of single-end sequencing or two files for paired-end. All other dependencies such as reference files and programs are either present or downloaded as part of the pipeline, with two exceptions. The user has to provide the GATK executable due to licensing issues; Java and R should also be pre-installed. A short description of how to run the pipeline is provided in the short README file (Supplementary file 3).

Output files

The program generates more than 30 files, although most of them are not necessary for non-programmers. There are three files that can help identify the causal mutation. The file Rplot.pdf (looks similar to Figure 2a or 2b) shows the chromosomal location of each SNP, plotted against a LOESS-fitted ratio variable. This variable was generated using this equation:

$$ratio = \frac{wt_{ref}}{wt_{ref} + wt_{alt}} - \frac{mut_{ref}}{mut_{ref} + mut_{alt}}$$

where,

wt_{ref} is the number of reads in the WT bulk that are called with the reference genome nucleotide,

wt_{alt} is the number of reads in the WT bulk that are called with a non-reference genome nucleotide,

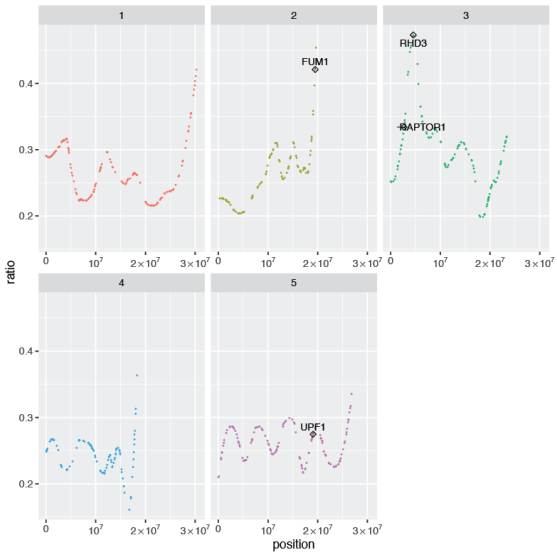
mut_{ref} is the number of reads in the mutant bulk that are called with the reference genome nucleotide

and,

mut_{alt} is the number of reads in the mutant bulk that are called with a non-reference genome nucleotide.

The ratio should be around zero for unlinked SNPs and approximately 0.66 for the causal mutation and genetically linked SNPs (Figure 2). We removed all SNPs with a ratio lower than 0.1 and applied LOESS smoothing with degree=2 and span=0.3.

a



b

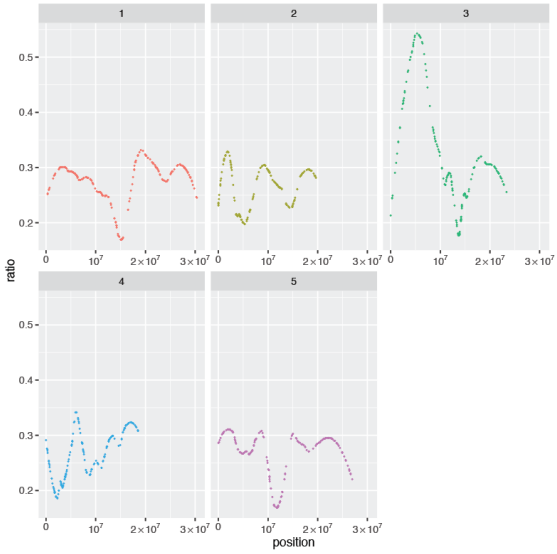


Figure 2. Example of the output plots. x-axis, chromosomal location; y-axis, ratio variable (see equation 1 for details). The number of each chromosome is labeled on top of each panel. a) and b) Line 300 and 194, respectively; see Table 1 and main text for details.

The second important file, cands4.txt lists the strongest candidate genes. (Supplementary file 1 as an example). These candidates are selected based on the following criteria. 1) The SNP has to be homozygous-alternate (namely, only reads that are different from the reference genome) in

the mutant bulk and with approximately 2:1 reference to alternate read ration in the WT bulk. Second, the SNP must be a single nucleotide change rather than an indel. Third, the mutation should have a significant effect on the protein, i.e., splice acceptor, splice donor, start lost, stop gained or missense variant.

The third important file is plot4.txt (Supplementary file 2 as an example). This file lists all the SNPs with their locations, the change in the coding sequence, the effect on the protein and the number of reads for each allele in each bulk. This file is important in case the cands4.txt did not yield any candidate genes. For example, in some cases, the phenotype of the sampled mutants is difficult to distinguish from WT, such as the case when mapping QTLs or mutants with subtle phenotypes. In these scenarios, WT plants might be included in the mutant bulk and the causal SNP is interpreted as heterozygous for both bulks. In such cases, the user can manually browse the plot4.txt file to identify: 1) SNPs that are nearly homozygous-alternate in the mutant bulk and approximately 2:1 reference to alternate allelic ratio in the WT bulk and 2) SNPs with a significant effect on the coding region.

Discussion

Approximately half of the Arabidopsis genes have unknown molecular functions (http://www.arabidopsis.org/portals/genAnnotation/genome_snapshot.jsp) which creates a large opportunity for discovering novel gene functionalities through relatively simple forward-genetic screens. The introduction of next-generation sequencing technologies offers new and rapid opportunities for identification of the genes affected by such screens. The main advantage of genome sequencing over traditional map-based cloning methods that use molecular markers

such as RFLP and AFLP is that NGS provides single-nucleotide resolution. While pre-NGS methods could identify a genomic region, this had to be further mined for potential mutation(s) in candidate genes. By contrast, sequencing a mutagenized genome reveals the entire population of genetic changes. The causal mutation is then precisely identified using bioinformatics/computational tools. Another important benefit of using NGS for mapping is that a small population (a few dozen individuals) is usually sufficient for mapping. Since the mapping resolution is as high as a single nucleotide, the importance of each and every recombinant for reducing the region of the causal mutation ("chromosome walking") is less important. In other words, the size of the region bounded by recombination events in which the causal mutation lies is not important as the researcher is able to visualize each and every SNP, evaluate it, and choose the one that has the highest likelihood of being causal. Mapping by NGS is especially fruitful when a large mapping population can be easily generated and screened which is the case for many plant species, nematodes and yeast. We have developed an easy to use tool that allows mapping of single-nucleotide induced mutations, even by researchers that have very little experience with bioinformatics tools. Our pipeline will take the fasq files as input and will identify causal SNPs with no pre-processing steps required. The output tables and plots can be readily used to identify the most likely mutation. Even in the case where no candidate gene is present, the list of SNPs with their read calls number in the mutant and WT bulks and their effect, will most likely point towards the correct gene. In theory, the SIMPLE pipeline can be used with any diploid species that has bulked mutant and WT mapping populations. Although not many species are currently covered, it is easy to add any species of interest to the pipeline (see README file). The program runs on Mac OSX version 10.11.6 and Linux release 6.7 (GNOME 2.28.2) with Java 1.7 installed

(see README file for specification) which is a commonly used platform in many labs. The SIMPLE project is hosted on GitHub (https://github.com/wacguy/Simple) and includes a quick-start README file.

Methods

Growth condition and screening

Arabidopsis seedlings were grown on MS medium ((Murashige and Skoog 1962) containing 1% sucrose. EMS screen was performed according to (Weigel and Glazebrook 2006). DR5::Luc seedling were screened using the Lumazone system according to (Moreno-Risueno et al. 2010) with lines 474-3, 300, 300-4, 300-7, 633 and 194, with pCASP2::GFP expression for lines B381 and M381 (Liberman et al. 2015) and for the cortex marker (Brady et al. 2007) expression in the EMS608 line. The GFP screens for CASP2 and the cortex marker were performed using Leica fluorescence stereo dissecting scope.

Sequencing

Paired or single-end libraries were prepared using NexteraXT or KAPA Hyper-prep Kit according to manufacturer instruction and sequenced on Illumina 2000/2500 instrument with the High or Rapid throughput mode at the Duke Center for Genomic and Computational biology.

Mutant validation

Line 474-3 has a single ground tissue layer and a short root phenotype, similar to other *shr* alleles (Helariutta et al. 2000). The phenotype of F1 offspring in a cross between 474-3 as a male donor

and *shr-2* (Helariutta et al. 2000) has the phenotype of the parental lines suggesting that line 474-3 is allelic to *SHR*.

Line 300 (and 300-4, a segregating line of 300) was crossed with *rhd3-1* (Wang et al. 1997; 2015) and the F1 progenies have short root and wavy phenotype similar to the parental lines suggesting that these lines are allelic. Line 300-7 was crossed with SALK line 063371 (Liu et al. 2010) and the F1 progenies phenocopied both parental lines. *myb36-1* (line 381 M and B) was previously described (Liberman et al. 2015). Line 194 has the same phenotype as line 300-4 and *rhd3-1* allele as described above.

All F1 crosses were sequenced for the male donor allele (see Table 2 for primers).

Table 2

line	primers name	forword	reverse	remarks
474-3	cw49/cw50	TCTCCATACCTCAAACTCCTCC	TTGCCTCTCCGTCTACTGC	these primers were used to genotype the shr-2 allele
300/300-4	seq-RHD3-muts-f/r	cagagctttctgattaaacaaacttc	CAAGTGCTTGAGGCAAGTGA	
300-7	M3-300-7-f/r	CACTGATGAAGAAAGGAAGAAGg	CATCTTGGATTCGATCGGTAA	

Primers that were used to amplify and sequence the male allele in F1 seedlings.

Data access

Whole genome sequencing data have been deposited in the Sequence Read Archive (SRA) with the accession number PRJNA353239.

Acknowledgments

We would like to thank people in the Benfey lab for comments on the manuscript and helpful experimental suggestions and to Carmen Wilson for technical assistant. We also thank the Duke Center for Genomic and Computational Biology for preparing and sequencing the DNA libraries.

This research was funded by a grant to PNB from the NIH (R01-GM043778), and by the Howard Hughes Medical Institute and the Gordon and Betty Moore Foundation (GBMF3405).

Disclosure declaration

The authors declare no conflict of interest.

Supplementary file1

An example of cands4.txt file from line 633.

Supplementary file2

An example of plot4.txt file from line 633.

Supplementary file3

Instruction file for running the pipeline. This file is also available on the SIMPLE GitHub repository.

References

Austin RS, Vidaurre D, Stamatiou G, Breit R, Provart NJ, Bonetta D, Zhang J, Fung P, Gong Y, Wang PW, et al. 2011. Next-generation mapping of Arabidopsis genes. *The Plant Journal* **67**: 715–725.

Brady SM, Orlando DA, Lee JY, Wang JY, Koch J, Dinneny JR, Mace D, Ohler U, Benfey PN. 2007. A High-Resolution Root Spatiotemporal Map Reveals Dominant Expression Patterns. *Science* **318**: 801–806.

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Ruden DM, Lu X. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w(1118); iso-2; iso-3. *Fly (Austin)* **6**: 80–92.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**: 491–498.

Fukaki H, Tameda S, Masuda H, Tasaka M. 2002. Lateral root formation is blocked by a gain-of-function mutation in the SOLITARY-ROOT/IAA14 gene of Arabidopsis. *Plant J* **29**: 153–168.

Helariutta Y, Fukaki H, Wysocka-Diller J, Nakajima K, Jung J, Sena G, Hauser MT, Benfey PN. 2000. The SHORT-ROOT gene controls radial patterning of the Arabidopsis root through radial signaling. *Cell* **101**: 555–567.

Leshchiner I, Alexa K, Kelsey P, Adzhubei I, Austin-Tse CA, Cooney JD, Anderson H, King MJ, Stottmann RW, Garnaas MK, et al. 2012. Mutation mapping and identification by whole-genome sequencing. *Genome Research* **22**: 1541–1548.

Lewis EB, Bacher F. 1968. *Method of feeding ethyl methane sulfonate (EMS) to Drosophila males*. Dros. Inf. Serv.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754–1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

Liberman LM, Sparks EE, Moreno-Risueno MA, Petricka JJ, Benfey PN. 2015. MYB36 regulates the transition from proliferation to differentiation in the Arabidopsisroot. *Proceedings of the National Academy of Sciences* **112**: 12099–12104.

Liu D, Gong Q, Ma Y, Li P, Li J, Yang S, Yuan L, Yu Y, Pan D, Xu F, et al. 2010. cpSecA, a thylakoid protein translocase subunit, is essential for photosynthetic development in Arabidopsis. *Journal of Experimental Botany* **61**: 1655–1669.

McClintock B. 1950. The Origin and Behavior of Mutable Loci in Maize. *Proc Natl Acad Sci USA* **36**: 344–355.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**: 1297–1303.

Moreno-Risueno MA, Van Norman JM, Moreno A, Zhang J, Ahnert SE, Benfey PN. 2010. Oscillating Gene Expression Determines Competence for Periodic Arabidopsis Root Branching. *Science* **329**: 1306–1311.

Morgan TH. 1910. Sex limited inheritance in Drosophila. *Science* **32**: 120–122.

Muller HJ. 1927. Artificial Transmutation of the Gene. *Science* **66**: 84–87.

Murashige T, Skoog F. 1962. A Revised Medium for Rapid Growth and Bio Assays with Tobacco Tissue Cultures. *Physiol Plant* **15**: 473–497.

Okumura K-I, Goh T, Toyokura K, Kasahara H, Takebayashi Y, Mimura T, Kamiya Y, Fukaki H. 2013. GNOM/FEWER ROOTS is required for the establishment of an auxin response maximum for arabidopsis lateral root initiation. *Plant and Cell Physiology* **54**: 406–417.

Schneeberger K, Ossowski S, Lanz C, Juul T, Petersen AH, Nielsen KL, Jørgensen J-E, Weigel D, Andersen SU. 2009. SHOREmap: simultaneous mapping and mutation identification by deep sequencing. *Nat Meth* **6**: 550–551.

R Core Team. 2013. R: A language and environment for statistical computing.

Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, et al. 2013. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* **43**: 11.10.1–33.

Wang HY, Lockwood SK, Hoeltzel MF, Schiefelbein JW. 1997. The ROOT HAIR DEFECTIVE3 gene encodes an evolutionarily conserved protein with GTP-binding motifs and is required for regulated cell enlargement in Arabidopsis. *Genes & Development* **11**: 799–811.

Wang J, Wang Y, Yang J, Ma C, Zhang Y, Ge T, Qi Z, Kang Y. 2015. Arabidopsis ROOT HAIR DEFECTIVE3is involved in nitrogen starvation-induced anthocyanin accumulation. *J Integr Plant Biol* **57**: 708–721.

Weigel D, Glazebrook J. 2006. EMS Mutagenesis of Arabidopsis Seed. *CSH Protoc* **2006**: pdb.prot4621. http://cshprotocols.cshlp.org/content/2006/5/pdb.prot4621.abstract.