the plant journal

SEB Society for Experimental Biology

*The Plant Journal* (2011) **67**, 715–725 doi: 10.1111/j.1365-313X.2011.04619.x

FEATURED ARTICLE

TECHNICAL ADVANCE

# Next-generation mapping of Arabidopsis genes

**Ryan S. Austin[1,†], Danielle Vidaurre[2,†], George Stamatiou[2,†], Robert Breit[2], Nicholas J. Provart[1,2], Dario Bonetta[3], Jianfeng Zhang[2], Pauline Fung[1], Yunchen Gong[1], Pauline W. Wang[1], Peter McCourt[1,2,†] and David S. Guttman[1,2,†,*]**

[1]*Centre for the Analysis of Genome Evolution and Function, University of Toronto, ON, Canada,*

[2]*Department of Cell & Systems Biology, University of Toronto, ON, Canada, and*

[3]*Faculty of Science, University of Ontario Institute of Technology, Toronto, ON, Canada*

Received 23 February 2011; revised 14 April 2011; accepted 19 April 2011; published online 18 July 2011.

[*]For correspondence (fax +1 416 978 5878; e-mail david.guttman@utoronto.ca).

[†]These authors contributed equally to this work.

## SUMMARY

**Next-generation genomic sequencing technologies have made it possible to directly map mutations responsible for phenotypes of interest via direct sequencing. However, most mapping strategies proposed to date require some prior genetic analysis, which can be very time-consuming even in genetically tractable organisms. Here we present a *de novo* method for rapidly and robustly mapping the physical location of EMS mutations by sequencing a small pooled $F_2$ population. This method, called Next Generation Mapping (NGM), uses a chastity statistic to quantify the relative contribution of the parental mutant and mapping lines to each SNP in the pooled $F_2$ population. It then uses this information to objectively localize the candidate mutation based on its exclusive segregation with the mutant parental line. A user-friendly, web-based tool for performing NGM analysis is available at http://bar.utoronto.ca/NGM. We used NGM to identify three genes involved in cell-wall biology in *Arabidopsis thaliana*, and, in a power analysis, demonstrate success in test mappings using as few as ten $F_2$ lines and a single channel of Illumina Genome Analyzer data. This strategy can easily be applied to other model organisms, and we expect that it will also have utility in crops and any other eukaryote with a completed genome sequence.**

**Keywords: *Arabidopsis thaliana*, cell-wall synthesis genes, flupoxam, map-based/positional cloning, mapping population, next-generation genome sequencing.**

## INTRODUCTION

Next-generation sequencing (NGS) technologies provide an unprecedented wealth of high-resolution genotypic information that enables many traditionally difficult, time-consuming and expensive genetic assays to be supplanted by rapid and relatively cheap whole-genome sequencing. An important new application of NGS is as a replacement for traditional map-based or positional cloning of mutations by direct identification via whole-genome sequencing (Sarin *et al.*, 2008; Smith *et al.*, 2008; Srivatsan *et al.*, 2008; Blumenstiel *et al.*, 2009; Irvine *et al.*, 2009; Lister *et al.*, 2009; Schneeberger *et al.*, 2009; Zuryn *et al.*, 2010). Ideally, mutations underlying phenotypes of interest could be identified simply by sequencing the mutant genomes. Unfortunately, this is not possible as numerous unassociated polymorphisms segregate with the causative mutation in a mutagenized population, resulting in a very low signal to noise ratio. For this reason, even NGS mapping approaches require some form of genetic analysis to refine the chromosomal location harboring the causative lesion.

Most NGS mapping approaches overcome the background noise problem by either refining the genomic region of interest through prior genetic analysis, or via bulk analysis of a very large number of mutant lines. For example, Schneeberger *et al.* (2009) developed an approach called SHOREmap that successfully identified a causative Arabidopsis mutation by Illumina Genome Analyzer (GA) sequencing of a population of 500 pooled $F_2$ lines. Unfortunately, it is not clear from this study how robust SHOREmap is when

using fewer $F_2$ lines, which is a very important consideration when mapping mutations with difficult-to-score phenotypes or when the organism is difficult to propagate in large numbers.

In this study, we develop a new NGS mapping approach using Illumina GA data to reliably and easily map candidate causative mutations. Our Next Generation Mapping (NGM) protocol requires no prior mapping information and only a small $F_2$ population. Using NGM, we have identified three new genes that contribute to plant cell-wall composition. These genes were located on various regions and chromosomes, demonstrating that NGM has utility across the Arabidopsis genome. The mapping of one mutant was also replicated using pools of as few as ten $F_2$ lines and a single channel of Illumina GA IIx data (<20 million paired-end reads). Moreover, NGM will identify the non-recombinant, mutation-harboring region of the chromosome regardless of the type of mutation (SNP or insertion/deletion) causing the phenotype. In all experiments, NGM was able to identify a highly restricted genomic region containing very few candidate SNPs, and in one case only a single SNP. These SNPs can then be easily validated using standard reverse genetic techniques. As use of even small mapping populations was successful in identifying candidate genes, we expect that NGM will work well in identifying causative mutations in less experimentally tractable organisms.

## RESULTS

### Cell-wall mutant screen

The plant cell wall is the core constituent in the development of cellulosic biofuel (Somerville and Bonetta, 2001; Pauly and Keegstra, 2010), and understanding the genes that define cell-wall composition therefore has both basic and applied research interest. However, even with fully sequenced genomes, it has been difficult to predict which genes are important for cell-wall assembly or maintenance. Although there has been some success in using predictive approaches, we decided to use a functional approach based on forward genetic screens that can identify key genes in cell-wall biosynthesis and maintenance.

To identify cell wall-related mutants, we developed a screen to identify seedlings that exhibit hypersensitivity to an herbicide that specifically interferes with cell-wall biosynthesis. This strategy relies on the principle that mutations that alter the cell-wall structure will exacerbate the deleterious effects of the herbicide. We used a cellulose biosynthesis inhibitor, flupoxam, whose application leads to pronounced cellular distention due to reduced cellulose content and an easily discerned club root phenotype (Hoffman and Vaughn, 1996; Sabba and Vaughn, 1999; Vaughn and Turley, 2001).

As a proof of concept, we first assayed mutants against flupoxam that were previously known to affect either cell-wall assembly or integrity. In this collection, one mutant, *MURUS11* (*MUR11*), was hypersensitive to flupoxam (Figure 1). Loss-of-function mutations in the *MUR11* gene were originally identified by screening for altered monosaccharide composition of the cell wall by gas chromatography (Reiter *et al.*, 1997). Although *mur11* mutants were identified over 10 years ago, the wild-type gene corresponding to this locus has still not been reported.

We also performed a small-scale genetic screen to identify *flupoxam hypersensitive* (*fph*) mutants using EMS-mutagenized Col-0 seeds (Figure 1). Two of these

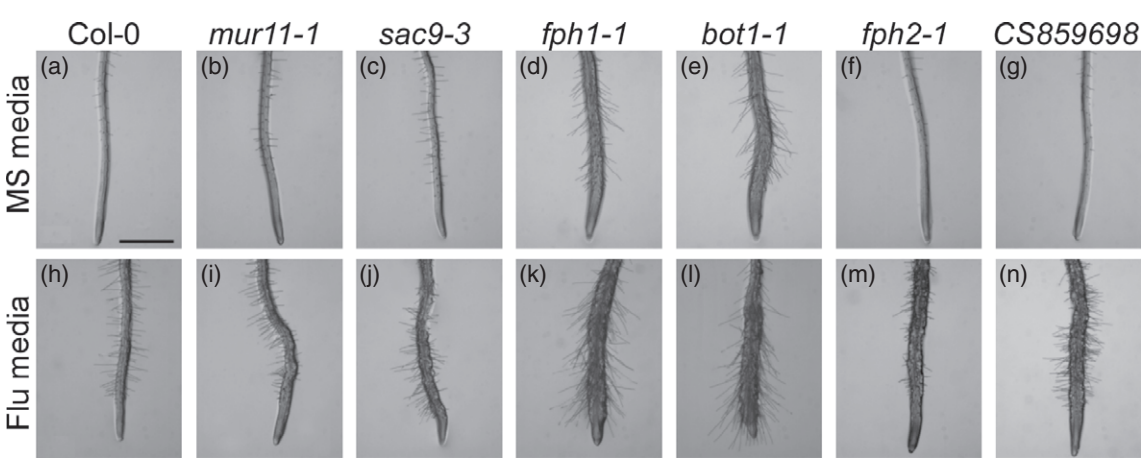



**Figure 1.** Flupoxam hypersensitivity phenotypes of *mur11-1/sac9*, *fph1-1/erh3* and *fph2-1/ost3* mutants.
Seedlings were grown vertically for 7 days on MS medium (a–g) or medium supplemented with 2.5 nM flupoxam (h–n) (see Experimental Procedures). Wild-type seedlings grown on MS (a) have fewer root hairs than when grown on flupoxam (h). *mur11-1* (b) and *sac9-3* (c) roots exhibit a wild-type phenotype on MS medium but form swollen bulges along the root when transferred to flupoxam (i and j, respectively). *fph1-1* (d) and *bot1-1* (e) produce ectopic root hairs on MS medium and show a dramatic increase in radial swelling as well as root hair formation in the presence of flupoxam (k and l, respectively). In contrast, *fph2-1* (f) and its T-DNA insertion allele *fph2-2* (g) exhibit a wild-type phenotype on MS, but are hypersensitive to flupoxam, as indicated by an increase in root hair production and radial swelling (m and n, respectively). Scale bar = 1 mm.

mutants, designated *fph1* and *fph2*, plus the *mur11-1* allele were chosen for NGM studies.

**Illumina GA sequencing and Next Generation Mapping**

Pooled genomic DNA from multiple $F_2$ lines should largely consist of an equal mixture of both the mutant and mapping parental genomes (ecotypes Col-0 and L*er*, respectively). However, $F_2$ individuals exhibiting the recessive mutant phenotype will be homozygous for the mutant genome in the region surrounding the underlying causative mutation due to the genetic linkage of flanking regions with the recessive mutation of interest. The extent of the resulting linkage disequilibrium between the selected polymorphism and flanking neutral polymorphisms depends on the distance from the selected site, the strength of selection, and the rate of recombination. In the context of a mapping population, selection is maximized, as only lines with homozygous recessive mutations are artificially selected, and the rate of recombination is proportional to the number of $F_2$ lines selected. One complication is that, although the Arabidopsis ecotypes used in this study are the same as those previously sequenced, they are still genotypically distinct individuals. Consequently, we expect some discrepancy between our genome sequences and the published genome sequences of these two ecotypes.

We extracted, pooled and sheared genomic DNA from 80 $F_2$ lines generated by crossing the *mur11* mutant line and two independent *fph* mutants to the L*er* mapping line. Each of the three pooled samples were sequenced on seven flowcell channels using 38 nt paired-end sequencing on the Illumina GA IIx platform. Table 1 shows the throughput for each run. Illumina paired-end reads were mapped to the TAIR9 release of the Arabidopsis Col-0 genome using the Maq short-read mapping software suite (Li *et al.*, 2008; Li and Durbin, 2009). SNPs refer specifically to sites that differ from the Col-0 genome sequence. Raw SNP calls returned by the Maq software were then filtered using the depth and quality cut-offs listed in Table S1. Final collections of 253 790, 231 511 and 310 196 SNPs were identified for *mur11-1*, *fph1* and *fph2*, respectively. Although 99 832 SNPs are shared among all three mutants, these were not filtered out during the analysis, in order to maintain a completely *de novo* prediction.

We first assessed the pattern of genome-wide SNPs by plotting the SNP frequencies for each mutant using a bin size of 250 kb (Figure 2). Localized regions lacking in SNPs can clearly be seen near the end of chromosome 3, near the end of chromosome 1, and in the second half of chromosome 1 for *mur11*, *fph1* and *fph2*, respectively. These 'SNP deserts' correspond to expected non-recombinant blocks created by linkage to the recessive mutation.

As these non-recombinant SNP deserts are fairly large, encompassing 3–4 Mb for each mutant, we developed a methodology that could narrow the search window in which the mutation of interest would reside. To qualitatively characterize SNP frequencies, we chose to use a modification of the Illumina chastity statistic, which is used in the Illumina pipeline to measure data ambiguity caused by interference among reads (crowding of clusters) during the sequencing process. In this case, we pegged SNP chastity inversely against the reference genome, in order to measure the proportion of reads at a polymorphic genomic position that differ from the reference genome sequence (Figure S1). Thus, a 'discordant chastity' statistic ($Ch_D$) of approximately zero would be expected at all genomic positions where the reference base is observed. A $Ch_D$ value of approximately 0.5 is expected at all genomic positions that vary between, and have equal representation from, the mutant and mapping parental genotypes. Finally, a $Ch_D$ value of approximately 1.0 is expected for all positions that are homozygous for a base that differs from the Col-0 genome. The non-recombinant SNP desert flanking the causative mutation will, by definition, have very few sites with discordant chastity scores near 0.5, and any SNP in this region with a $Ch_D$ value of approximately 1.0 may either represent the actual causal mutation of interest, a silent EMS-induced mutation carried along due to linkage with the causative mutation, or a difference between our mutant line and the published genome sequence. Consequently, we can identify the region of interest by simply scanning for a genomic block that is deficient in discordant chastity values of approximately 0.5 and enriched in discordant chastity values of approximately 1.0.

To identify these regions, we made use of probability density estimates that reflect the frequency of polymorphism across the length of each chromosome. Density estimators have been shown to be very useful in representing genomic features derived from high-throughput sequencing data (Boyle *et al.*, 2008). In our approach, we first calculated 80 'chastity threads' across the chromosome, in which each chastity thread is a probability density estimate that reflects the positional frequency of SNPs with discordant chastity values within a specific, discrete interval. The first interval quantified the frequency of SNPs with $Ch_D$

**Table 1** Maq mapping of Illumina reads to the *A. thaliana* genome (TAIR9 release)

| Mutant | Number of reads | Number of reads mapped | Number of paired reads mapped | Mean depth | Genome coverage | Error |
|---|---|---|---|---|---|---|
| *mur11* | 115 610 496 | 96 737 060 | 75 111 880 | 29 | 99.96 | 0.035 |
| *fph1* | 302 978 858 | 265 885 982 | 215 745 005 | 74 | 99.56 | 0.028 |
| *fph2* | 148 509 664 | 126 349 616 | 122 778 391 | 39 | 97.26 | 0.0065 |

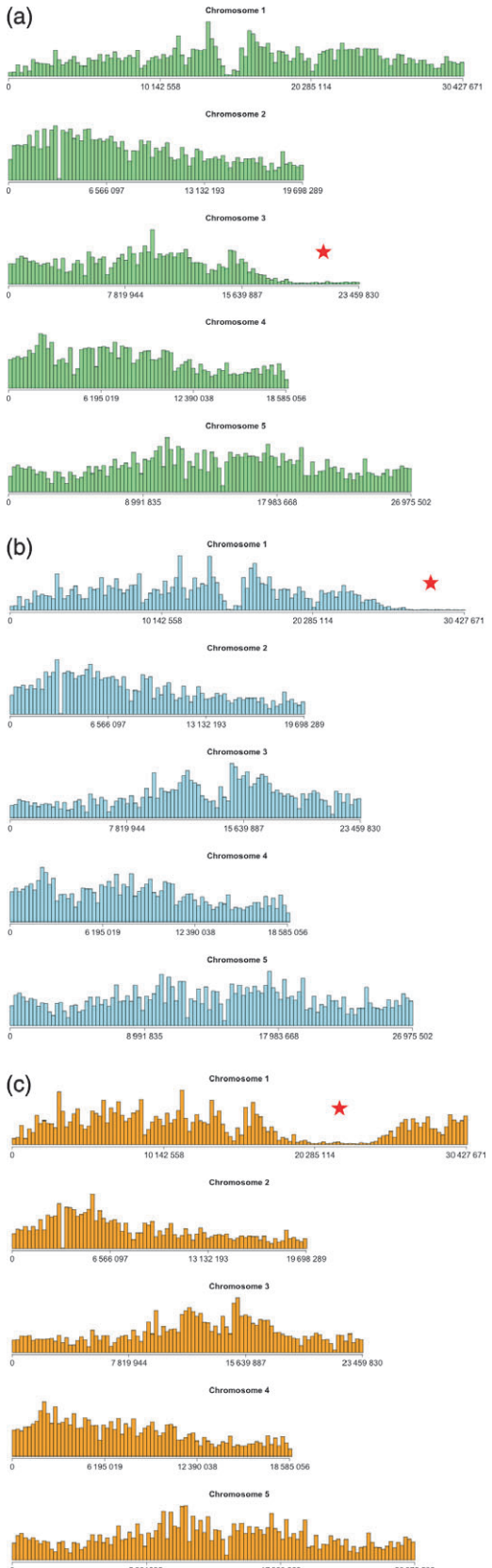



**Figure 2.** Genome-wide SNP frequencies plotted as a function of chromosomal position for (a) MUR11 (At3g59770), (b) FPH1 (At1g80350) and (c) FPH2 (At1g61790).
Filtered SNPs were plotted based on their abundance over each chromosome using a bin size of 250 kb. Non-recombinant regions within the tail of chromosome 3, the tail of chromosome 1 and the right arm of chromosome 1 for mutants MUR11, FPH1 and FPH2, respectively, are indicated by a star.

between 0.20 and 0.30, while each successive scan increased the $Ch_D$ interval by 0.01. We then generated chastity threads by applying a kernel density estimation to the positional SNP frequencies for each interval (Figure 3).

In order to identify the desired homozygous and heterozygous signals, we clustered the 80 chastity threads into a smaller number of distinct 'chastity belts', by performing *k*-means clustering on the pairwise correlations among all chastity threads. We then identified the two chastity belts that encompassed the expected homozygous mutant sites ($Ch_D$ = 1.0) and heterozygous sites ($Ch_D$ = 0.5), and calculated the ratio of homozygous to heterozygous signals. This entire procedure was repeated using several progressively smaller kernel sizes in order to refine the region of greatest divergence between the two signals. Finally, we identified the candidate region and identified all non-synonymous SNPs within this region for further validation. This analysis returned five, one and two candidate SNPs for the *mur11*, *fph1* and *fph2* mutant lines, respectively (Table 2).

We have developed a web-based application called NGM to perform these analyses (http://bar.utoronto.ca/NGM). NGM provides an intuitive interface for quickly localizing a small set of candidate SNPs responsible for a causal phenotype. It proceeds via an interactive series of steps using data generated from any of several popular short-read mapping and SNP calling software packages available in the public domain, and provides a shortlist of annotated SNPs potentially responsible for the observed phenotype. Details are available in Appendix S1.

### Validation of candidate SNPs

One of the candidate SNPs identified by NGM was in the gene *SUPPRESSOR OF ACTIN 9* (*SAC9*, At3g59770). This mutant has been studied extensively, and many of the reported phenotypes of *sac9* appear to be similar to those reported for *mur11* mutants (Williams *et al.*, 2005). Sanger sequencing of *mur11-1* identified a G → A substitution in the coding region of the *SAC9* gene that results in replacement of the highly conserved arginine residue (R274) in domain IV (Zhong and Ye, 2003) by a histidine (H) residue (Figure 1). Comparisons of *mur11-1* with a T-DNA insertion allele (SALK_058870) in SAC9 (*sac9-3*) showed that *mur11-1* has similar but weaker *SAC9* loss-of-function phenotypes (Figure S2). More importantly, the *sac9-3* allele showed good root flupoxam hypersensitivity, strongly suggesting that *MUR11* and *SAC9* are the same gene (Figure 1). The

**Table 2** Annotation of SNPs from the three *A. thaliana* mutants

| Mutation | Chromosome | Position | RefN[a] | SNP | RefP[b] | Amino acid substitution | Depth | $Ch_D$ | Strand | Accession |
|---|---|---|---|---|---|---|---|---|---|---|
| *mur11* | 3 | 20795011 | A | T | D | E | 16 | 1.00 | – | AT3G56040.1 |
| *mur11* | 3 | 20795012 | T | C | D | G | 19 | 1.00 | – | AT3G56040.1 |
| *mur11* | 3 | 22003267 | C | T | W | Stop | 17 | 0.87 | – | AT3G59570.1 |
| *mur11*[c] | 3 | 22084518 | C | T | R | H | 14 | 1.00 | – | AT3G59770.1 |
| *mur11* | 3 | 22752769 | G | A | G | S | 23 | 0.96 | + | AT3G61480.1 |
| *fph1*[c] | 1 | 30206068 | G | A | P | S | 124 | 0.98 | – | AT1G80350.1 |
| *fph2*[c] | 1 | 22814777 | C | T | Q | Stop | 29 | 1.00 | + | AT1G61790.1 |
| *fph2* | 1 | 22926537 | C | T | D | N | 28 | 0.96 | – | AT1G62030.1 |

[a]Reference nucleotide.
[b]Reference amino acid.
[c]Validated mutation.

*SAC9* gene encodes a phosphoinositide phosphatase (PI phophatases), and PI phosphatases have been suggested to have roles ranging from signal transduction and actin cytoskeleton organization to vesicle trafficking (Williams *et al.*, 2005; Gong *et al.*, 2006).

NGM identified only a single candidate SNP for the *fph1* mutant. This G → A transition in the *ECTOPIC ROOT HAIR3* (*ERH3*) gene (At1g80350) results in substitution of a proline (P393) amino acid by a serine (S). In addition to being identified by its bushy root phenotype, mutations in *ERH3* have also been identified as *botero1* (*bot1*) mutants, which show reduced hypocotyl cell elongation (Schneider *et al.*, 1997; Bichet *et al.*, 2001). We obtained the *bot1-1* allele and determined this allele was also flupoxam-sensitive. Sequencing of *bot1-1* identified a G → A base pair transition, resulting in conversion of the codon for tryptophan (W57) to a stop codon.

NGM identified two candidate SNPs for the *fph2* mutants. One of these, *OLIGOSACCHARIDE TRANSMEMBRANE TRANSPORTER* (*OST3/OST6*, At1g61790), contained a G → A base pair substitution, resulting in a premature stop codon rather than glutamine at residue 130. The biochemical annotation of this gene as involved in sugar transport suggests that it is a good candidate for modification of cell-wall composition, and we therefore ordered a SALK T-DNA line for the *OST3/OST6* gene. This insertion line showed hypersensitivity to flupoxam, similar to the EMS allele isolated from our screen.

### NGM power analysis and comparison with other methods

NGM was developed in parallel to map-based cloning of the Arabidopsis *MUR11* and *FPH1* genes, providing an excellent opportunity to validate the new approach (Figure S3). We mapped the *mur11-1* mutant to chromosome 3 between markers ciw4 (18.9 Mb) and nga6 (23 Mb) via bulk segregation analysis. Further fine mapping of this region using a population of 665 mutant $F_2$ plants narrowed the interval to a 150 kb region containing 50 putative genes; however, due to lack of recombination, we were not able to further define the region containing the *mur11-1* mutation. Scanning this region for published mutants identified *SUPPRESSOR OF ACTIN 9* (*SAC9*), mutants of which have phenotypes similar to the *mur11* mutants (Williams *et al.*, 2005). Sequencing of the *SAC9* gene in the *mur11-1* background identified a G → A base pair change.

Bulk segregation analysis of *fph1* localized the mutation to chromosome 1 between markers nga111 (27.35 Mb) and the end of the chromosome, while fine-structure mapping of 475 $F_2$ mutant plants narrowed the region to an 86 kb region containing 28 putative genes. All 28 genes were sequenced, and this analysis identified the A → G transition in *ERH3.* In both *MUR11* and *FPH1*, NGM identified the causative mutations using many fewer $F_2$ individuals and in a vastly shorter period of time than our conventional mapping.

To characterize the power of NGM with respect to data requirements, we re-analyzed the *FPH2* data by cumulatively adding one channel of sequence data at a time and re-running NGM using the default parameters (Table 3 and Figure S4). Each channel produced approximately 20 million sequences and increased the read depth by approximately 5.5-fold. Remarkably, even with only one channel of sequence data, we were able to restrict the candidate region to only 4.3 Mb and the number of candidate SNPs to as few as 10. The analysis reached a maximum resolution after only four channels of data; reducing the candidate region to only 615 kb and the number of candidate SNPs to two. Contrary to our expectations, NGM actually performed less well with five and six channels of data compared to four and seven channels, although this was readily resolved by modification of some of the analysis parameters (e.g. kernel size and the number of clusters used in the *k*-means clustering).

We also assessed the power of NGM with respect to the number of $F_2$ lines required for mapping. Instead of pooling 80 $F_2$ lines as was originally done for the *FPH2* mutation, we generated eight pools of ten $F_2$ lines each and sequenced the eight pools independently. Poor DNA shearing during the sample preparation for this run resulted in poorer coverage of the genome than in the previous NGM runs (approxi-

**Table 3** Power analysis relative to the amount of sequence data using mutant FPH2

| Number of channels | Number of reads | Number of reads mapped | Number of paired reads mapped | Mean depth | Percentage genome coverage | Identified region (kb) | Number of candidate SNPs |
|---|---|---|---|---|---|---|---|
| 1 | 22 596 940 | 19 204 310 | 18 635 529 | 6 | 21.75 | 4302 | 10 |
| 2 | 40 342 244 | 34 252 826 | 33 265 026 | 11 | 43.33 | 4302 | 10 |
| 3 | 62 028 258 | 52 720 710 | 51 214 280 | 16 | 59.24 | 3688 | 9 |
| 4 | 83 872 282 | 71 308 750 | 69 279 962 | 22 | 68.48 | 615 | 2 |
| 5 | 105 642 116 | 89 855 042 | 87 308 465 | 28 | 74.33 | 1229 | 4 |
| 6 | 127 066 982 | 108 104 200 | 105 046 033 | 33 | 78.22 | 1229 | 4 |
| 7 | 148 509 664 | 126 349 616 | 122 778 555 | 39 | 81.10 | 615 | 2 |

mately 38% coverage per channel of data versus approximately 80%), but fortunately, this provided us with an additional means to assess the power of NGM (Table S2). Only four of the eight lanes produced reads that mapped to the *fph2* mutation (depths ranged from 2 to 31 reads). Remarkably, even with only one channel of data from ten $F_2$ lines, and twofold coverage of our mutation, NGM was able to identify the causative mutation (Figure S5). Even more encouragingly, when we ran NGM independently on the four lanes that did not produce reads covering the causative mutation, we were still able to define a very restricted region where the mutation should be found due to their characteristic patterns of genetic variation.

We also compared NGM to two other existing high-throughput sequencing methods: (i) the approach proposed recently by Zuryn *et al.* (2010), and (ii) the SHORE-specific mapping tool SHOREmap (Schneeberger *et al.*, 2009). SHOREmap applied in conjunction with the genome mapper SHORE was unable to detect the first two mutants (i.e. *mur11* and *fph1*) due to apparent sensitivity to either the telomeric position of the mutations or possibly the fewer number of $F_2$ lines employed compared to previously reported results (i.e. 80 vs. 500). SHOREmap did identify the *fph2* mutation as the third ranked candidate using 80 $F_2$ lines, once we manually filtered results for non-synonymous substitutions arising from transition mutations. NGM identified *fph2* as the top candidate mutation using the default filters (Table S3).

Although the experimental design for our approach was fundamentally different, we implemented the approach described by Zuryn *et al.* (2010) using the polymorphic data for our three mutants. In their approach, the EMS-induced transition mutation signal is isolated to a specific region by subtracting polymorphic similarities across two or more mutants sequenced in parallel. Unfortunately, we were unable to filter extraneous EMS signal to a level that allowed detection of a finite non-recombinant region (data not shown). This is most certainly due to the lack of back-crossing applied to our mutants. Zuryn *et al.* (2010) recommend five back-crosses to eliminate the EMS-derived polymorphic signal from regions outside the non-recombinant mutation-harboring zone. Although this method is suitable in species with a short generation time, such as *Caenorhabditis elegans*, it is unfortunately restrictive in a plant model, such as Arabidopsis, with a life cycle of several months.

## DISCUSSION

Although reverse genetic approaches have become increasingly popular over the past few decades, they lack the functional appeal of forward genetic screens focused on phenotypes of interest. Additionally, unlike reverse genetics, forward genetic screens have the power to reveal a broad spectrum of mutant alleles. Unfortunately, forward genetic screens are typically limited by the fact that it is often much more difficult to map causative mutations than it is to generate interesting phenotypes. Consequently, as NGS technology has become available, a number of approaches have been proposed to identify causative mutations (Sarin *et al.*, 2008; Smith *et al.*, 2008; Srivatsan *et al.*, 2008; Blumenstiel *et al.*, 2009; Irvine *et al.*, 2009; Lister *et al.*, 2009; Schneeberger *et al.*, 2009; Zuryn *et al.*, 2010). While all of these methods have proven to be extremely powerful, most require some prior knowledge about the location of the causative mutation, back-crossing of the mutant line, generation of a large mapping population of $F_2$ lines and use of existing markers, or present technical hurdles for the average end-user. The variation in reliability of genetic markers, long generation times, and phenotypes that are difficult to generate or score exacerbate these issues. For these reasons, NGS analysis has yet to become a generalized approach for mutant gene identification.

NGM overcomes a number of these obstacles by increasing the power of the analysis while decreasing the difficulty of obtaining the biological material. We have also developed an extremely user-friendly web-based implementation of NGM to reduce the need for specialized computational skills and hardware. The benefits of NGM are apparent when evaluated against the key obstacles hindering mapping by NGS.

### Applicability to other systems

NGM can be readily applied to any organism with a mature genome sequence that can be crossed to a mapping line. The approach should work whenever the causative mutation is Mendelian, recessive, penetrant and not phenocopied by

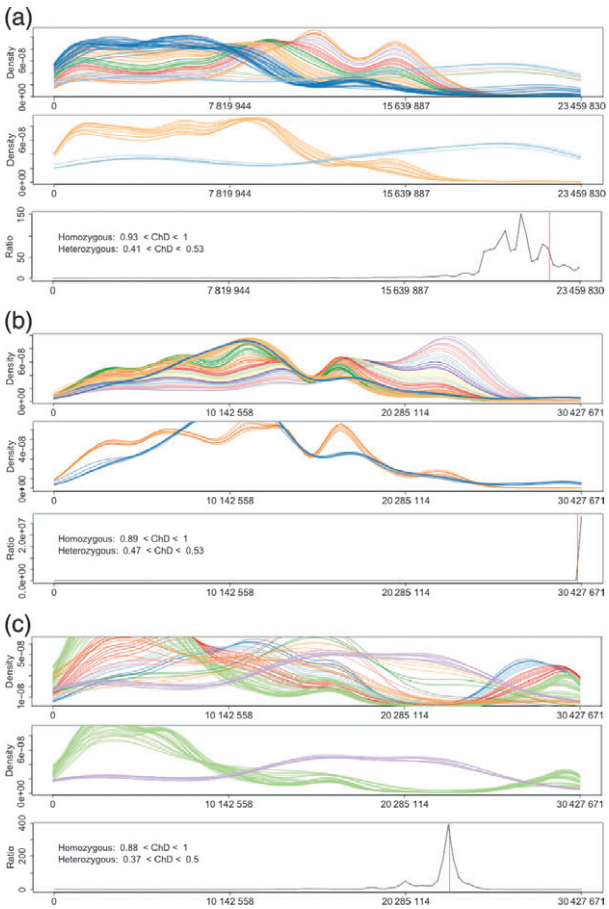



**Figure 3.** Discordant chastity ($Ch_D$) analysis for (a) MUR11 (At3g59770), (b) FPH1 (At1g80350) and (c) FPH2 (At1g61790).
Three sub-panels are shown for each mutant. The top sub-panel shows the distribution of the 100 discordant chastity threads. Each thread shows the frequency of SNPs that fall within a specified discordant chastity interval across the genome. The threads are coded along a color spectrum, with gold indicating intervals in the low-chastity range, and blue indicating intervals in the high-chastity range. The middle sub-panel shows the corresponding heterozygous and homozygous chastity bands, and indicates the chastity intervals used as identified by *k*-means clustering. The bottom sub-panel shows the ratios of homozygous to heterozygous discordant chastity signals with a kernel adjustment of 0.25. The physical position of each EMS mutation is indicated by a vertical red line.

other mutations. We also believe NGM can be used to map mutations in bulked lines and when working with reduced-representation genome sequences. Bulked lines are pools of genotypes distinguished by a phenotype of interest. Reduced-representation sequencing involves sequencing only the genic component of the genome, so that the final genome dataset is made up of thousands of unordered contigs. In this case, NGM should work as long as the contig carrying the causative mutation is large enough to allow identification of the monomorphic signal. We have also had recent success in the application of NGM to a mutant line from a strain where the mapping line possessed the sequenced reference genome (i.e. a L*er* EMS mutant crossed to a Col-0 mapping line). In this case, pre-processing polymorphisms to subtract overlapping SNPs in a manner akin to that described by Zuryn *et al.* (2010) allows use of NGM in mutated species for which a closely related reference is available (data not shown).

### Size of mapping population

Producing large mapping populations can be one of the most time-consuming aspects of traditional map-based cloning. This is particularly true when dealing with a mutant phenotype that is difficult to propagate or score. Traditional map-based cloning approaches commonly use over 1000 $F_2$ lines (Lukowitz *et al.*, 2000; Jander *et al.*, 2002), and an analysis of data from Lukowitz *et al.* (2000) indicates that, in Arabidopsis, approximately 2000 new $F_2$ lines are required for each tenfold reduction in the mapping interval. We have shown that NGM will work in Arabidopsis using as few as ten $F_2$ lines due to the extraordinary increase in the number of usable markers made available by full-genome analysis. Nevertheless, as we do not know how generally applicable these results are, we recommend using at least 50 $F_2$ lines to ensure sufficient recombination among lines.

### Amount of sequence data needed

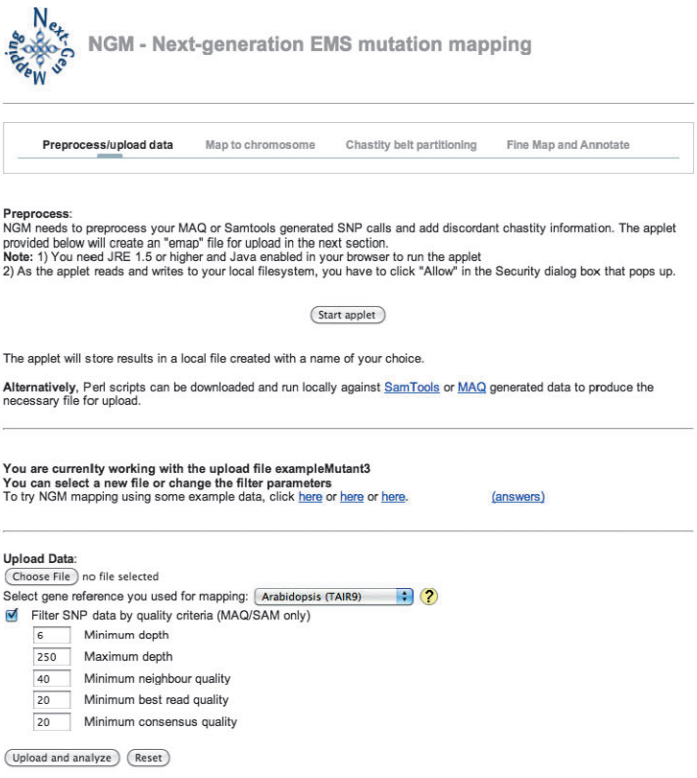
We found that the current generation of the Illumina GA IIx platform producing 20–40 million quality paired-end reads provides adequate throughput to identify an Arabidopsis EMS mutation using a single flowcell channel (Arabidopsis has a genome size of approximately 120 Mb). We also find that 38 nt paired-end reads are more than adequate for reliable reference assembly, although longer reads may be beneficial in other organisms depending on the level of complexity of the genome.

### Robustness with respect to type and location of mutation

We have demonstrated the power of NGM on three independent mutants, and have obtained excellent results even when the mutation was located in difficult genomic regions such as near the telomere. The most challenging mutation was *mur11-1*, which is located in a low-recombination region near the telomere of chromosome 1. Despite these complications, we were able to identify a very manageable shortlist of candidate mutations, and manual annotation of this list revealed only three likely candidate SNPs, one of which proved to be the causative mutation. This same effect was observed in *fph1*; however, in this situation, the mutation was close enough to the end of chromosome 3 to reveal a distinct SNP desert containing polymorphisms highly discordant with the reference genome. Moreover, the reference mapping quality obtained from the *mur11-1* and *fph1* sequences was quite low (see Table 1). Despite this, NGM was still able to effectively identify the mutations in question.

Generally speaking, we found that identification and segregation of the homozygous and heterozygous signals

(a)




(b)

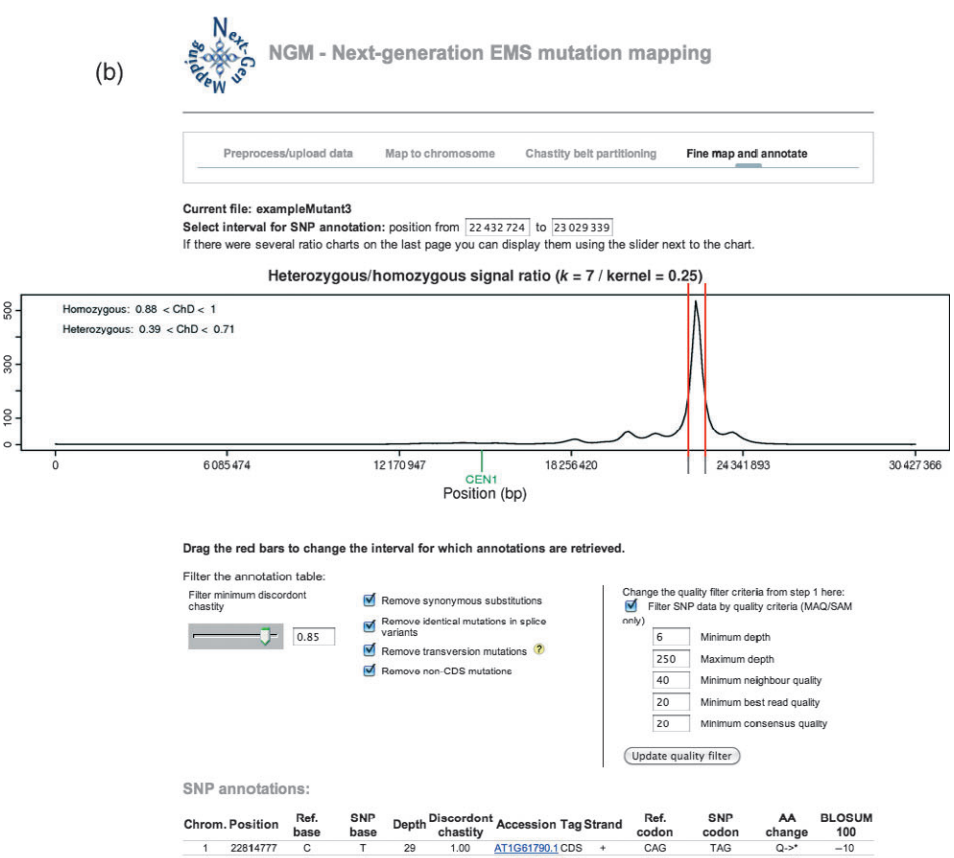


**Figure 4.** The NGM web application. (a) Screenshot of the upload/pre-processing stage of the NGM application. A Java applet is provided for pre-processing large data files on the user's computer before upload to the NGM server. The genomic reference used for mapping is selected from a drop-down box, and a set of default filters are available for filtering SNP data derived from Maq or SAM sources prior to analysis. SNP filters include options to filter on read depth, the quality of the reads that constitute an SNP, the quality of reads surrounding the SNP and the overall consensus quality of the SNP itself.
(b) Screenshot of the final stage of region selection and SNP annotation. The user chooses a region around the observed peak by adjusting the slider bars (red). SNPs are dynamically displayed and filtered using a variety of criteria, including their discordant chastity score, likelihood that they were caused by EMS, and whether they produce a non-synonymous amino acid change within the coding sequence. The ability to filter the initial raw SNP data using new filters is also possible. The NGM tool is available at http://bar.utoronto.ca/NGM.

using chastity belts provided greatly enhanced power to narrow the candidate regions and thereby reduce the number of candidate SNPs. In the case of *fph1* and *fph2*, the approach was able to narrow the regions to approximately 650 and 500 kbp, respectively. The analysis relies on setting a few key parameters, including the kernel size used for smoothing the chastity threads and the number of clusters used in the *k*-means clustering.

Appropriate selection of the kernel size for the Gaussian kernel density estimation strongly influenced our ability to narrow down the candidate region. A larger kernel causes greater smoothing of the chastity threads, while a small kernel improves resolution, but at the cost of an increased potential for over-fitting error. We used a default kernel of 0.25, but found that decreasing it to as low as 0.01 often dramatically improved the analysis.

The choice of *k* for the *k*-means clustering of the chastity threads into chastity belts requires a value large enough to cleanly separate the homozygous and heterozygous signals into distinctive chastity belts. Larger *k* values generally increased the homozygous to heterozygous ratio, but this did not always have a commensurate decrease on the size of the candidate region. We used a default *k* value of 9, but found that a varying *k* from 5 to 11 often improved the analysis. The online NGM tool therefore provides a means to perform NGM runs using multiple parameter settings.

A strength of NGM is that we can map the general location of causative mutations regardless of the nature of the underlying mutation. Linkage disequilibrium between a selected mutation and the surrounding region will necessarily result in locally reduced heterozygosity in the appropriate genomic region of the $F_2$ population regardless of the nature of the mutation. As NGM identifies the reduced-heterozygosity region by measuring a local ratio of the homozygous to heterozygous signals, it can identify a selected region regardless of whether the causative mutation is a non-synonymous SNP, a synonymous SNP, a mutation in a regulatory region, an insertion/deletion, or even an integration of a mobile element such as a T-DNA.

### Ease of use

NGM analysis simply requires the SNP calls returned by a short-read reference mapping program such as Maq (maq.sourceforge.net) or SAMtools (Li *et al.*, 2009). Mapping and NGM can be completed within approximately 1 day after obtaining the Illumina GA data. The NGM procedure itself takes minutes to complete. As discussed, we have developed a web-based application of NGM that automates all of these analyses, and allows very easy and intuitive manipulation of the relevant parameters and filters (Figure 4).

## CONCLUSION

As the throughput of NGS platforms increases, and the price per gigabase to produce such data decreases, the broad adoption of these approaches will only increase. NGM has proved to be a robust and extremely efficient method for mapping causative mutations. In defining gene and allelic function, the use of NGM will dramatically reduce the cost and time required to map causative mutations, as well as help researchers make informed decisions early in the experimental process. For example, our small flupoxam-hypersensitive screen identified over 20 independent and phenotypically indistinguishable lines; thus, no particular mutant has obvious priority with respect to further cellular or biochemical analysis. Further, the existing annotations for these loci provided very limited useful information. Neither *MUR11/SAC9* nor *ERH3/FPH1* encode cell-wall biosynthetic enzymes, and *OST3/OST6/FPH2* appears to be involved in processing *N*-linked oligosaccharides (Reiter *et al.*, 1997; Schneider *et al.*, 1997; Williams *et al.*, 2005; Gong *et al.*, 2006; Saint-Jore-Dupas *et al.*, 2006). Consequently, the genes identified to date suggest that screening for flupoxam hypersensitivity does not select for biosynthetic enzymes involved in *de novo* synthesis of cell walls, but rather for regulators of cell-wall composition.

It typically takes a skilled graduate student, postdoc or technician 1–2 years to map mutations such as those discussed above using traditional approaches. Next-generation sequencing platforms and NGM can reduce this to a matter of days. Although the cost of NGS is not trivial, by using NGM this cost is rapidly offset, not only by savings in salaries and consumables, but even more so by the benefits of accelerated research and discovery in forward genetic analyses.

## EXPERIMENTAL PROCEDURES

### Plant materials and growth conditions

*Arabidopsis thaliana* $M_2$ ecotype Columbia seeds mutagenized using ethyl methane sulfonate (EMS) were purchased from Lehle Seeds (http://www.arabidopsis.com). EMS alleles *mur11-1* and *bot1-1* and T-DNA insertions for *sac9-3* (SALK_058870) and *fph2-2*

(SALK_067271) were in the Columbia background, and were provided by the Arabidopsis Biological Resource Center (Ohio State University, Columbus, OH, USA). Seeds were surface-sterilized in 50% bleach, 0.01% Tween-20 for 5 min, rinsed five times with sterile water, and stored in the dark at 4°C for 4 days to stratify them and synchronize germination. Seeds were plated on 0.5× strength Murashige and Skoog (MS) agar plates, supplemented with 2.5 nM flupoxam (dissolved in DMSO) as indicated, and sealed with surgical tape, and kept under continuous light at room temperature.

### Mutant screen

Seeds were chilled for 4 days and sown onto 0.5× MS plates on strips of sterilized Whatman paper. The plates were placed vertically under continuous light conditions at room temperature. After 4 days, the filter paper containing approximately 20–30 seeds was aseptically transferred to fresh 0.5× MS plates containing 2.5 nM flupoxam. Seedlings were scored for swelling of the root or altered root growth after 7–9 days. We screened approximately 22 000 seedlings from 32 pools, and isolated 26 flupoxam-sensitive mutants. Hypersensitive mutants were re-tested in the $M_3$ generation.

### Genetic and physical mapping of mutants

Genetic mapping was accomplished using an $F_2$ population derived from a cross between the *mur11-1*, *fph1-1* and *fph2-1* mutants (from the Columbia genotype, Col-0) and Landsberg *erecta* (L*er*). $F_2$ seedlings were scored for sensitivity to flupoxam as indicated by swollen roots and/or short root growth. Genomic DNA was isolated from individual $F_2$ plants from a mapping population possessing the mutant phenotype, and assigned to a chromosome using published SSLP markers. New molecular markers were developed using the Monsanto Col-0 and L*er* polymorphism database (http://www.arabidopsis.org/Cereon).

### Sequencing of candidate genes

The *SAC9*, *ERH3* and *OST3* genes were amplified by PCR using X-Taq DNA polymerase with proofreading activity (Takara, http://www.takara-bio.com/). Sequencing reactions were performed using standard ABI3730 chemistry at the Centre for the Analysis of Genome Evolution and Function (CAGEF) at the University of Toronto. $F_2$ mutants from two independent crosses were used for sequencing and verifying lesions.

### DNA preparation

DNA was extracted from plant tissue using a Puregene DNA purification system (Qiagen, http://www.qiagen.com/), according to the manufacturer's instructions.

### Illumina GA sequencing and data preprocessing

Pooled sheared genomic DNA was loaded onto eight lanes of an Illumina GA IIx flowcell, and run for 38 cycles using the paired-end module. Data were then processed using Genome Analyzer pipeline 1.4, and the results were converted to Sanger-format Fastq data using a customized Perl script (available from the authors).

### Sequence assembly and SNP annotation

Illumina GA IIx reads from all three mutants were mapped against the TAIR9 release of the Arabidopsis genome (TAIR9_chr_all.fas) using the Maq short-read mapping software with default settings for paired-end analysis (Li *et al.*, 2008; Li and Durbin, 2009). Genome-wide SNP positions and pileup information were then collected using the 'maq cns2snp' and 'maq pileup' functions. SNPs were filtered using a subset of the recommended filters employed by the Maq Perl script SNPfilter using an *awk* command (Table S1). Base frequencies for each SNP were then extracted from the 'pileup' information, and discordant chastity scores were calculated and appended to the SNP data for NGM processing.

### Next-Generation Mapping (NGM)

We have developed a web-based application to permit the rapid and robust mapping of EMS mutations via NGM. NGM proceeds in four stages: (i) pre-processing, uploading and filtering of SNP data obtained from the sequenced $F_2$ mapping population, (ii) localization of the mutation to a specific chromosome through identification of the non-recombinant mutation-harboring region, (iii) partitioning of SNP frequencies based on their level of disagreement with the reference genome, and (iv) localization of the causal mutation and annotation of candidate SNPs within a finely mapped region (Figures S6-S8). Each stage allows user intervention and dynamic adjustments, with the ability to return to previous stages. Detailed descriptions and explanations are provided in Appendix S1 and at http://bar.utoronto.ca/NGM.

### Availability and requirements

The NGM server uses standard web-development languages, with AJAX and a MySQL database providing dynamic content access. The NGM algorithmic components for performing the mutation mapping are implemented in the R statistical programming language. As the discordant chastity statistic utilized by the NGM algorithm must be determined prior to analysis and often requires processing of very large data files, a Java applet is provided for preprocessing user data files on a local machine prior to uploading to the server.

The NGM tool is available free for academic use at http://bar.utoronto.ca/NGM. At present, the server provides genomic mapping for *Arabidopsis thaliana*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Saccharomyces pombe* and *Saccharomyces cerevisiae*. However, theoretically it can be applied to any organism with an annotated genome for which an $F_2$ mapping population can be obtained.

## ACKNOWLEDGEMENTS

This work was supported by the Agriculture and Agri-food Canada Applied Bioproducts Innovation Program.

## SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

**Figure S1.** Graphical explanation of chastity and the discordant chastity statistic.
**Figure S2.** Phenotypes of *mur11*.
**Figure S3.** Map-based cloning of *MUR11*, *FPH1* and *FPH2*.
**Figure S4.** Power analysis relative to the amount of sequence data using mutant FPH2.
**Figure S5.** Power analysis relative to the number of $F_2$ lines using mutant FPH2.
**Figure S6.** Screenshot of the NGM chromosome selection stage.
**Figure S7.** Screenshot of the chastity belt partitioning stage of NGM.
**Figure S8.** Screenshot of output from the NGM multi-view option.
**Table S1.** SNP filtering statistics applied to Maq SNP calls.
**Table S2.** Power analysis relative to the number of $F_2$ lines using mutant FPH2.
**Table S3.** Comparison of NGM with SHOREmap.
**Appendix S1.** Next-Generation Mapping (NGM) methodology.

Please note: As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.

## REFERENCES


**Bichet, A., Desnos, T., Turner, S., Grandjean, O. and Hofte, H.** (2001) BOTERO1 is required for normal orientation of cortical microtubules and anisotropic cell expansion in Arabidopsis. *Plant J.* **25**, 137–148.

**Blumenstiel, J.P., Noll, A.C., Griffiths, J.A., Perera, A.G., Walton, K.N., Gilliland, W.D., Hawley, R.S. and Staehling-Hampton, K.** (2009) Identification of EMS-induced mutations in *Drosophila melanogaster* by whole-genome sequencing. *Genetics*, **182**, 25–32.

**Boyle, A.P., Guinney, J., Crawford, G.E. and Furey, T.S.** (2008) F-Seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics*, **24**, 2537–2538.

**Gong, P., Wu, G.S. and Ort, D.R.** (2006) Slow dark deactivation of *Arabidopsis* chloroplast ATP synthase caused by a mutation in a nonplastidic SAC domain protein. *Photosynth. Res.* **88**, 133–142.

**Hoffman, J.C. and Vaughn, K.C.** (1996) Flupoxam induces classic club root morphology but is not a mitotic disrupter herbicide. *Pestic. Biochem. Physiol.* **55**, 49–53.

**Irvine, D.V., Goto, D.B., Vaughn, M.W., Nakaseko, Y., McCombie, W.R., Yanagida, M. and Martienssen, R.** (2009) Mapping epigenetic mutations in fission yeast using whole-genome next-generation sequencing. *Genome Res.* **19**, 1077–1083.

**Jander, G., Norris, S.R., Rounsley, S.D., Bush, D.F., Levin, I.M. and Last, R.L.** (2002) Arabidopsis map-based cloning in the post-genome era. *Plant Physiol.* **129**, 440–450.

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

**Li, H., Ruan, J. and Durbin, R.** (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18**, 1851–1858.

**Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R.** (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

**Lister, R., Gregory, B.D. and Ecker, J.R.** (2009) Next is now: new technologies for sequencing of genomes, transcriptomes, and beyond. *Curr. Opin. Plant Biol.* **12**, 107–118.

**Lukowitz, W., Gillmor, C.S. and Scheible, W.R.** (2000) Positional cloning in Arabidopsis. Why it feels good to have a genome initiative working for you. *Plant Physiol.* **123**, 795–805.

**Pauly, M. and Keegstra, K.** (2010) Plant cell wall polymers as precursors for biofuels. *Curr. Opin. Plant Biol.* **13**, 305–312.

**Reiter, W.D., Chapple, C. and Somerville, C.R.** (1997) Mutants of *Arabidopsis thaliana* with altered cell wall polysaccharide composition. *Plant J.* **12**, 335–345.

**Sabba, R.P. and Vaughn, K.C.** (1999) Herbicides that inhibit cellulose biosynthesis. *Weed Sci.* **47**, 757–763.

**Saint-Jore-Dupas, C., Nebenfuhr, A., Boulaflous, A., Follet-Gueye, M.L., Plasson, C., Hawes, C., Driouich, A., Faye, L. and Gomord, V.** (2006) Plant N-glycan processing enzymes employ different targeting mechanisms for their spatial arrangement along the secretory pathway. *Plant Cell*, **18**, 3182–3200.

**Sarin, S., Prabhu, S., O'Meara, M.M., Pe'er, I. and Hobert, O.** (2008) *Caenorhabditis elegans* mutant allele identification by whole-genome sequencing. *Nat. Methods*, **5**, 865–867.

**Schneeberger, K., Ossowski, S., Lanz, C., Juul, T., Petersen, A.H., Nielsen, K.L., Jorgensen, J.E., Weigel, D. and Andersen, S.U.** (2009) SHOREmap: simultaneous mapping and mutation identification by deep sequencing. *Nat. Methods*, **6**, 550–551.

**Schneider, K., Wells, B., Dolan, L. and Roberts, K.** (1997) Structural and genetic analysis of epidermal cell differentiation in *Arabidopsis* primary roots. *Development*, **124**, 1789–1798.

**Smith, D.R., Quinlan, A.R., Peckham, H.E. *et al.*** (2008) Rapid whole-genome mutational profiling using next-generation sequencing technologies. *Genome Res.* **18**, 1638–1642.

**Somerville, C.R. and Bonetta, D.** (2001) Plants as factories for technical materials. *Plant Physiol.* **125**, 168–171.

**Srivatsan, A., Han, Y., Peng, J., Tehranchi, A.K., Gibbs, R., Wang, J.D. and Chen, R.** (2008) High-precision, whole-genome sequencing of laboratory strains facilitates genetic studies. *PLoS Genet.* **4**, e1000139.

**Vaughn, K.C. and Turley, R.B.** (2001) Ultrastructural effects of cellulose biosynthesis inhibitor herbicides on developing cotton fibers. *Protoplasma*, **216**, 80–93.

**Williams, M.E., Torabinejad, J., Cohick, E., Parker, K., Drake, E.J., Thompson, J.E., Hortter, M. and Dewald, D.B.** (2005) Mutations in the Arabidopsis phosphoinositide phosphatase gene *SAC9* lead to overaccumulation of Ptdlns(4,5)$P_2$ and constitutive expression of the stress-response pathway. *Plant Physiol.* **138**, 686–700.

**Zhong, R.Q. and Ye, Z.H.** (2003) The SAC domain-containing protein gene family in Arabidopsis. *Plant Physiol.* **132**, 544–555.

**Zuryn, S., Le Gras, S., Jamet, K. and Jarriault, S.** (2010) A strategy for direct mapping and identification of mutations by whole-genome sequencing. *Genetics*, **186**, 427–430.