the plant journal

SEB Society for Experimental Biology

The Plant Journal (2016) **86**, 349–359 doi: 10.1111/tpj.13161

TECHNICAL ADVANCE

Highly efficient *de novo* mutant identification in a *Sorghum bicolor* TILLING population using the ComSeq approach

Habte Nida[1], Shula Blum[1], Dina Zielinski[2], Dhruv A. Srivastava[1,3], Rivka Elbaum[1], Zhanguo Xin[4], Yaniv Erlich[2,5,6], Eyal Fridman[3,*] and Noam Shental[7,*]

[1]*The Robert H. Smith Institute of Plant Sciences and Genetics in Agriculture, Robert H. Smith Faculty of Agriculture, Food and Environment, The Hebrew University of Jerusalem, Rehovot, Israel,*

[2]*New York Genome Center, 101 Avenue of the Americas, New York, NY, USA,*

[3]*Institute of Plant Sciences, Agricultural Research Organization, The Volcani Center, Bet Dagan, Israel,*

[4]*Plant Stress and Germplasm Development Unit, US Department of Agriculture/Agricultural Research Service, Lubbock, TX, USA,*

[5]*Department of Computer Science, Fu Foundation School of Engineering, Columbia University, New York, NY, USA,*

[6]*Center for Computational Biology and Bioinformatics, Columbia University, New York, NY, USA, and*

[7]*Department of Mathematics and Computer Science, The Open University of Israel, Raanana, Israel*

Received 11 January 2016; revised 10 February 2016; accepted 1 March 2016; published online 9 March 2016.

*For correspondence (e-mails fridmane@agri.gov.il or shental@openu.ac.il).

SUMMARY

Screening large populations for carriers of known or *de novo* rare single nucleotide polymorphisms (SNPs) is required both in Targeting induced local lesions in genomes (TILLING) experiments in plants and in screening of human populations. We previously suggested an approach that combines the mathematical field of compressed sensing with next-generation sequencing to allow such large-scale screening. Based on pooled measurements, this method identifies multiple carriers of heterozygous or homozygous rare alleles while using only a small fraction of resources. Its rigorous mathematical foundations allow scalable and robust detection, and provide error correction and resilience to experimental noise. Here we present a large-scale experimental demonstration of our computational approach, in which we targeted a TILLING population of 1024 *Sorghum bicolor* lines to detect carriers of *de novo* SNPs whose frequency was less than 0.1%, using only 48 pools. Subsequent validation confirmed that all detected lines were indeed carriers of the predicted mutations. This novel approach provides a highly cost-effective and robust tool for biologists and breeders to allow identification of novel alleles and subsequent functional analysis.

Keywords: *Sorghum bicolor*, Targeting induced local lesions in genomes, rare alleles, *de novo* SNPs, compressed sensing, large scale screening, technical advance.

INTRODUCTION

Discovering rare alleles and their carriers has broad applications in Targeting induced local lesions in genomes (TILLING) experiments (McCallum *et al*., 2000; Comai *et al*., 2004), as well as in large-scale screening for disease-related single nucleotide polymorphisms (SNPs) and their carriers in humans. Despite continuously decreasing next-generation sequencing (NGS) costs, the cost of library preparation of the thousands of samples required to detect rare SNPs is still prohibitive. Further, despite the advancement of allele editing in crop plants (Belhaj *et al*., 2013), TILLING populations still provide a rich source of null and functional alleles that are indispensable for studying allelic series of genes and incorporating novel alleles in breeding programs without GMO regulatory constraints. Hence there is an unmet need for efficient detection of rarer SNPs and their carriers in increasingly large populations while minimizing library preparation requirements.

In previous work (Erlich *et al*., 2010; Shental *et al*., 2010), we suggested use of a compressed sequencing approach (ComSeq) that bridges this gap and provides a practical method for identifying rare alleles and their carriers by sequencing pooled DNA and computationally detecting the carriers' identities (Figure 1). ComSeq is able to detect any number of carriers (as opposed to detecting a single carrier), and is suitable for both detection of de *novo* SNPs and their carriers, and for large-scale screening of carriers of known rare SNPs, as is often the case in screening human populations. Our former publication (Shental *et al.*, 2010) presented the computational approach and provided *in silico* estimates of its performance, but lacked an explicit experimental validation, which is presented here.

ComSeq combines NGS and the mathematical field of compressed sensing (Candès, 2006; Donoho, 2006) to efficiently identify carriers of rare SNPs. Briefly, samples are pooled according to a pre-defined design whereby each sample is represented in several pools to provide a unique 'signature'. This carefully designed scheme enables recovery of rare allele carriers based on read counts for major and minor alleles in each pool. ComSeq offers a significant saving in resources, as all sample preparation and sequencing is performed over the pools, whose number is logarithmic to the number of samples. Our *in silico* simulations (Shental *et al.*, 2010) showed that the method is not limited to detection of a single rare allele carrier but may be applied in the case of multiple carriers. The method is efficient up to a maximum minor allele frequency of approximately 5%. For higher frequencies, the required number of pools is comparable to the number of samples and hence the naive approach is preferable. In contrast to the naive approach of testing each sample individually, which becomes prohibitive when seeking carriers of rarer alleles, the efficiency of ComSeq increases with decreasing allele frequencies, as pooling provides more efficient detection of rare allele carriers. ComSeq also seamlessly detects carriers of either heterozygous or homozygous rare alleles.

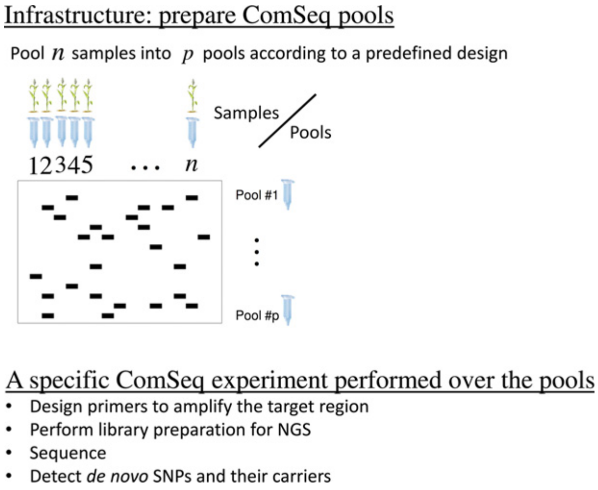
Infrastructure: prepare ComSeq pools
Pool n samples into p pools according to a predefined design
Samples
Pools
12345
...
n
Pool #1
Pool #p
A specific ComSeq experiment performed over the pools
Design primers to amplify the target region
Perform library preparation for NGS
Sequence
Detect *de novo* SNPs and their carriers

Figure 1. Flowchart of the ComSeq approach.
In the infrastructure step, samples, i.e. lines, are pooled according to a predefined compressed sensing design. Here each row corresponds to a pool of several genotypes, and the bars represent genotypes that participate in a given pool. Once created, targeted selection may be applied to the set of pools.

Experimental outline

We present an experimental validation of the ComSeq approach in which we searched for *de novo* SNPs and their carriers in an ethyl methanesulfonate (EMS) TILLING population of 1024 *Sorghum bicolor* lines. Equal amounts of DNA from these lines were combined to create a set of 48 pools, which corresponds to a more than 20-fold (1024/48) reduction in resources compared to the naive approach. Each pool contained DNA from 128 lines, and each line appeared in six pools. One line was replaced by a carrier of a known mutant of the caffeic acid *O*-methyltranferase (*COMT*) gene, to serve as a positive control (Xin *et al.*, 2008).

We targeted two genes, Sb07g021400 (a member of the 'no apical meristem' (NAC) transcriptional regulator superfamily, hereafter referred to as NAC7) and Sb04g028020 (an ortholog of 'low Silicon in rice' (*Lsi*), hereafter referred to as *Lsi1*) as described below, and a region of the *COMT* gene that contained the positive control SNP from previous work (Xin *et al.*, 2008). Pools were sequenced on an Illumina HiSeq 2500 sequencing system, generating approximately 22 million reads in total (corresponding to approximately one-fifth of a lane), and subsequently analyzed using ComSeq. The results yielded five predictions, namely five lines that were predicted to carry *de novo* heterozygous SNPs at specific locations. These lines were subsequently tested via Sanger sequencing, and all were found to carry the predicted SNP. In addition, the *COMT* mutant carrier, which served as a positive control, was also correctly detected. A single *de novo* homozygous SNP carrier was also detected; however, neither DNA nor seeds were available for subsequent Sanger sequencing (Figure S2).

Whole-genome sequencing of 256 other lines from the same population indicated that there were an average of 11 SNPs per 1 Mb (Jiao and Burke, 2016), which corresponds to approximately seven SNPs within the total number of sequenced nucleotides across all lines. Based on this estimate, we suggest that there were no false-negative detections, and ComSeq correctly detected all SNPs in the targeted regions (mutation frequency was about 0.1%).

The detected SNPs included a C→T substitution (one case) and a G→A substitution (three cases), which are typical for EMS, which generates transition mutations by alkylating G residues (Greene *et al.*, 2003; Comai and Henikoff, 2006). The additional confirmed *de novo* SNP involved a transversion (A→T), which is not common for EMS mutagenesis and probably resulted from DNA heterogeneity in the original seed stocks used for DNA extraction. Mutagenesis was performed *de novo* in this study and is not representative of the DNA samples described by Xin *et al.* (2008).

DNA was extracted from a pool of 12 sprouting M_4 seeds (Experimental procedures), which were advanced by pooling 10 M_3 panicles, and thus lines are expected to contain more heterozygous mutations than homozygous mutations. Whole-genome sequencing indicated a ratio of more than 4.2:1 heterozygous to homozygous SNPs across the entire genome (Jiao and Burke, 2016), consistent with our findings.

Genetic targets

In this study, we focused on regions in two genes, each localized to a different chromosome and thought to play a role in different physiological pathways.

Lsi1 is a member of a gene family of silicon channels that were initially identified in rice (Ma *et al.*, 2007; Yamaji *et al.*, 2008). Silicon minerals are a major component in soils. Accumulation of silica in sorghum roots and leaves was correlated with increased tolerance to drought by maintenance of a high stomatal conductance, increased efficiency of water use, increased net growth, and more efficient water uptake from the soil. Current efforts in the field of bio-silicification involve linking physiological and anatomical observations to molecular mechanisms to explain the mechanism by which silicification increases stress tolerance of plants. As silicon starvation is not possible for plants grown in silicon-rich soils, mutants that are defective in silicon absorption may be identified by forward genetic approaches involving counter-selection for genotypes unable to uptake the analogous and toxic element germanium (Ma *et al.*, 2007), or through allele mining of candidate genes, as in this study. Indeed, screening a population of fast neutron-radiated *Sorghum* plants allowed us to identify one such knockout mutant in a root silicon transporter, SbLsi1, which enabled us to analyze leaf epidermal silicification, showing the occurrence of active cellular deposition of silicon via activity of this transporter (Markovich *et al.*, 2015). In this study, an amplicon was designed to capture a region encoding a conserved NPA (Asn-Pro-Ala) motif that is known to play a key role in silica transport into cells (Mitani *et al.*, 2009).

The second gene, *NAC7*, is an ortholog of GPC-1, a transcription factor from wheat (*Triticum aestivum*) that has been shown to regulate protein content in grains (Uauy *et al.*, 2006; Avni *et al.*, 2013). We became interested in its function after it was found to be associated with grain nitrogen content of sorghum hybrids (I. Ben-Israel, E. Fridman, Plant Sciences Institute, The Volcani Research Centre, Bet-Dagan, Israel, unpublished results). The targeted SNPs in this gene were located in the promoter region, for which significant association with grain nitrogen content was identified.

Pooling design

The pooling design applied in this experiment was based on a Reed–Solomon (RS) code (Reed and Solomon, 1960), whose error correction properties proved useful in several ways. First, this design allows correct carrier identification even when some pools have low coverage or when entire pools drop out for technical reasons: any four of the total of six pools are sufficient to detect a sample). Second, error correction provides excellent robustness to experimental noise, as a carrier may be detected only if it fits its pre-defined 'signature'. As a result, we achieved perfect detection, overcoming cases of highly variable coverage, complete loss of pools, variability in DNA concentrations, and PCR bias across lines (Kharabian-Masouleh *et al.*, 2011). The applied RS code provides a general framework that allows 'tailoring' of the code according to the number of samples, maximal predicted mutation rate, etc.

Number of reads per pool and ease of detection

The median number of reads per pool varied between 16 000 and 180 000 across amplicons. Pools containing the carrier had a minor (mutated) allele count of approximately 50–400 reads, while the rest of the pools had almost zero reads of the minor allele. As a result, the experimental signal-to-noise ratio was extremely high, and carrier detection was straightforward. We reason that, due to the error correction properties of the code, it is possible to allocate fewer reads per pool while still providing accurate detection. This allows an increase in the number of target intervals to cover the entire gene of interest at the same or lower cost. In addition, it would improve identification of allelic series required to fine-tune studies of gene function, i.e. amending hitherto unknown domains.

Potential PCR bias of pooled samples

Each pool contained DNA from 128 lines, and predicted PCR biases (Marroni *et al.*, 2012) did not play a significant role. A relatively uniform amplification is achieved even when pooling and PCR-amplifying more than 500 samples (Zaboli *et al.*, 2012). Using larger pools increases the efficiency by decreasing the number of pools while increasing the total number of samples screened.

Comparison with other methods

Current methods for detecting carriers when mining TILLING populations are based on 'multi-dimensional pooling' designs (Missirian *et al.*, 2011; Tsai *et al.*, 2011), in which each sample is assigned to *d* pools, where *d* is the 'dimensionality' of the design. For example, a common 2D code for 96 samples is set by creating 12 'column pools' and 8 'row pools' based on their location in a standard 96-well plate. A SNP in a certain sample is detected by observing the minor allele in its corresponding 'column' and 'row' pools.

ComSeq has several advantages over such multi-dimensional designs, which may all be attributed to the rigorous mathematical foundations of compressed sensing. First, ComSeq readily allows detection of many carriers, as opposed to detection of only a single carrier in multi-dimensional pooling. Second, the more elaborate pooling design of ComSeq provides robustness to common experimental problems (e.g. pools that have almost zero read counts, or samples that are mistakenly present or absent from certain pools). Standard multi-dimensional pools are not robust in the face of such experimental problems, and cannot detect a carrier in the case of even a single faulty pool. As a result, multi-dimensional pooling designs require sophisticated SNP calling methods (Missirian *et al.*, 2011), which are unnecessary in the case of ComSeq. To the best of our knowledge, our application of ComSeq used the largest number of lines compared to former multi-dimensional pooling experiments, while displaying a much larger efficiency ratio of lines to pools. However, multi-dimensional pooling requires much lower coverage per pool (30–40x). Also, former multi-dimensional pooling experiments targeted much larger regions and hence found many SNPs and their carriers, a task that was beyond the scope of our proof-of-concept study.

To summarize, we present a scalable and robust method for detecting rare allele carriers. Following this proof-of-concept experiment, we pooled an additional group of more than 3000 lines, which are all available to the community to further mine rare alleles and their carriers.

RESULTS

We detected six *de novo* SNPs and their carriers (in addition to the carrier of the *COMT* positive control). We describe each of these detections as they reflect different facets of the ComSeq approach. The SNPs were detected as explained in Experimental procedures.

Detecting carriers of SNPs in the *Lsi1* gene

We targeted a single 248 bp region in the Sb04g028020 (*Lsi1*) gene (ComSeq target 18; CST18, Table S1), and present it as an introductory example for detecting carriers via ComSeq.

Example 1: introduction. Figure 2(a) shows the read count for each pool at position 197 along the amplicon. The median number of reads per pool was approximately 27 000, which corresponds to approximately 200x coverage per line, as each pool contains 128 lines. The read count is dominated by the major allele (G in this case). Figure 2(b) shows a region of low read counts. The read count for the minor allele (A) is significantly higher in six pools while being almost zero in all others. The other two possible alleles (C and T) are absent from all pools. Applying ComSeq, namely solving Eqn. (1) (Experimental procedures), resulted in detection of the rare allele carrier line 165. Figure 2(c) shows the Sanger sequencing results for line 165, indicating the two predicted alleles (A and G) at location 197.

The coverage varied significantly across pools, with a standard deviation of approximately 11 000 reads, and pool 32 was unsuccessfully amplified, showing zero reads. The same variation was observed at other locations along the amplicon, and therefore depends on the specific PCR reaction. ComSeq is robust to these variations as it estimates the ratio between the minor allele read count and the total number of reads, and hence works well as long as the number of reads is high enough to determine this ratio. The dotted line in Figure 2(b) indicates the 256th fraction of the total number of reads in a pool, which corresponds to the expected read count in the case of a pool with 128 diploids (256 allele copies) that contains a heterozygous carrier. The 'fingerprint' of six distinct peaks in Figure 2(b) indicates the high signal-to-noise ratio and the ease of detection in this case.

Example 2: overcoming pool dropout. Figure 3 shows the read count among the pools for the 51st position of the CST18 amplicon (Table S1), and low read counts typical of the minor allele, as in Figure 2. The major and minor alleles at this locus are C and T, respectively. Solving Eqn. (1) results in detection of line 206, which was subsequently validated by Sanger sequencing (Figure 3c). A minimum of four pools harboring the minor allele is necessary to identify the carrier. In this case, only five distinct peaks make up the 'fingerprint' for line 206 (Figure 3b), as the sixth pool containing this line (pool 32) had no sequencing coverage.

Detecting carriers in NAC7 targets

We targeted two regions of the NAC7 gene, CST9 and CST11 (Table S1), which yielded two and one independent carrier detections, respectively.

Example 3: the importance of the ratio of minor allele and total read counts. Figure 4 shows the read count across pools for position 58 of the CST9 amplicon, and a magnification of a lower read count region typical of the minor allele, as in Figures 2 and 3. The major and minor alleles at this locus are A and T, respectively. Solving Eqn. (1) in this case results in detection of line 665, which was subsequently validated by Sanger sequencing (Figure 4c). The read count for the minor allele (T) was significant in six pools and almost zero in all other cases, as were the read counts for the two other potential alleles.

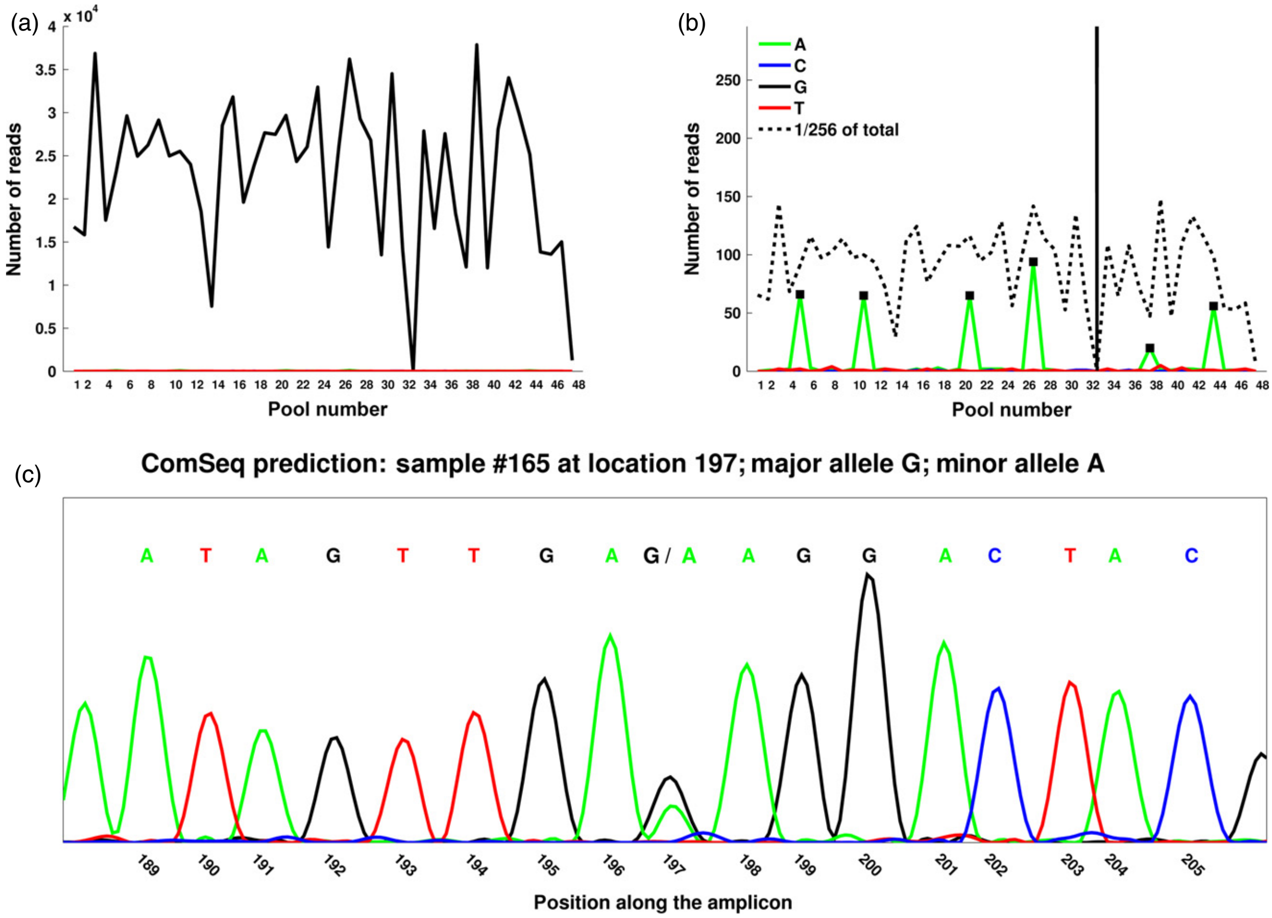

Figure 2. Detecting a *de novo* SNP in the *Lsi1* gene (Sb04g028020) and its carrier.
(a) Read count for all nucleotides at position 197 of the amplicon. The count for the major allele G (black) is higher than the count for all other alleles by at least two orders of magnitude, hence the latter counts are not visible at this scale.
(b) Magnification of a low read count region. The other three alleles may now be observed. Note that the major allele (G) is observed in pool 32 (shown as a vertical black line) as this pool was unsuccessfully amplified and thus showed zero counts during sequencing. Also displayed is a dotted line showing the level of 1/256 of the total number of reads. Only six pools show a distinct read count for the minor allele A (green), which, together with the ComSeq pooling design, readily allows identification of line 165 as a carrier.
(c) Sanger sequencing chromatogram for line 165, confirming the G→A heterozygous SNP at position 197.

Pools 33 and 38 had a similar number of minor allele counts, although only the latter contains line 665. ComSeq uses the ratio between minor allele read counts and the total number of reads, which is fivefold smaller for pool 33 than for pool 38. This approach rejects pool 33 as containing a carrier at this location, supporting correct identification.

Note that the atypical A→T transversion is probably not caused by EMS (Greene *et al.*, 2003) but rather by natural and rare heterogeneity in the original seed stock used in the TILLING project.

Results for the second heterozygous SNP detected CST9 carrier are shown in Figure S1. ComSeq also detected a carrier of a homozygous SNP for CST9, as shown in Figure S2. We could not validate this prediction, as DNA was not available for the relevant line.

Example 4: robustness of ComSeq in cases of variable coverage across pools. Figure 5(a,b) shows the read count for position 45 of the CST11 amplicon, and a typical region of lower minor allele read counts, in the same format as in Figures 2–4. The major and minor alleles at this locus are G and A, respectively. Solving Eqn. (1) in this case results in detection of line 865, which was subsequently validated by Sanger sequencing (Figure 5c). All six pools are clearly visible in this case. As in the previous example, the variability of read counts for all 48 pools is high, ranging from approximately 20 000 to an order of magnitude more. Nevertheless, in both cases, the compressed sensing analysis allowed identification and validation of rare mutant alleles, highlighting the robustness of the pipeline.

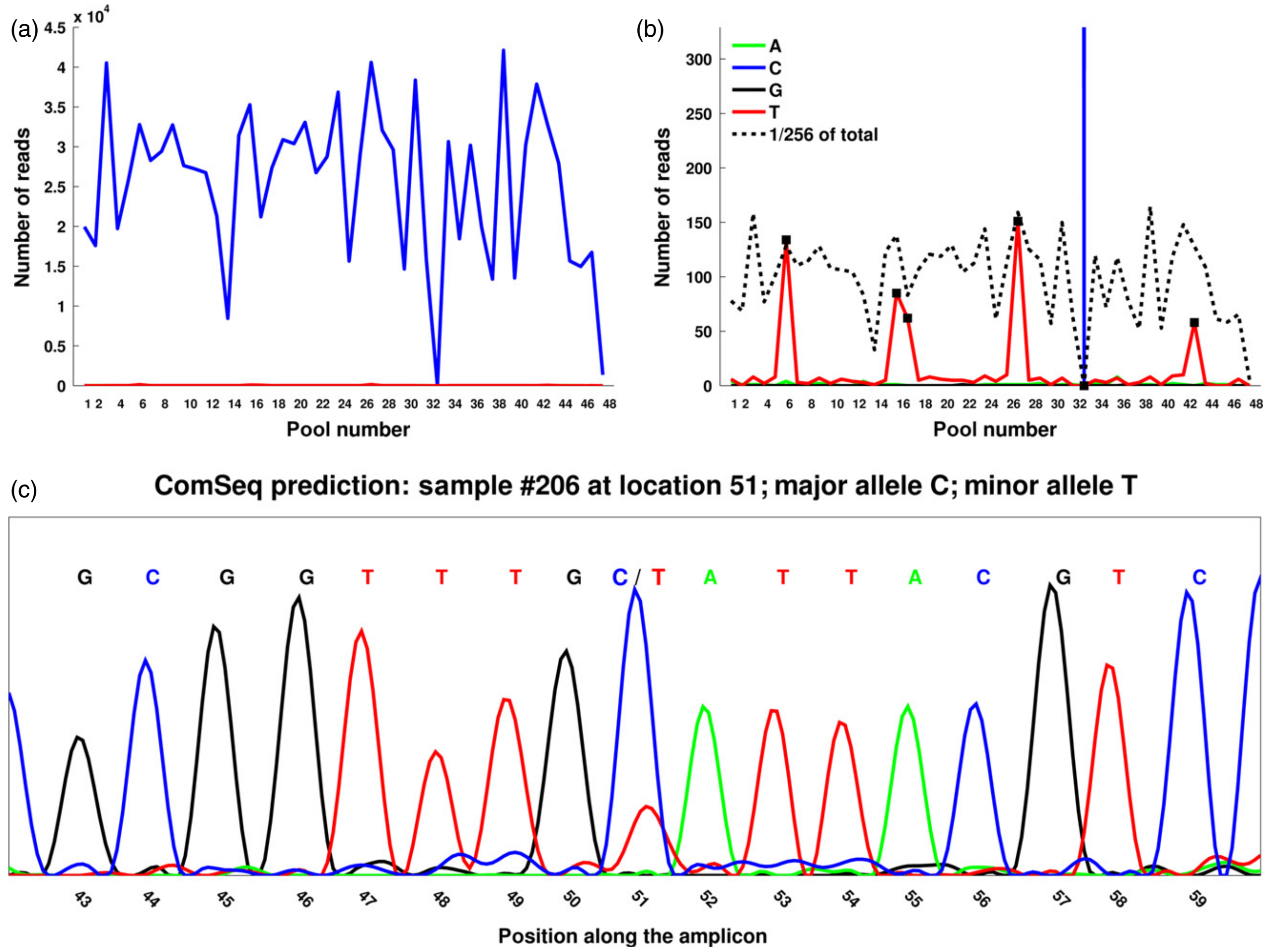

Figure 3. Detecting the second *de novo* SNP in the *Lsi1* gene (Sb04g028020) and its carrier.
(a) Read count for all nucleotides at position 51 of the amplicon. The major allele is C.
(b) Five pools show a distinct read count for the minor allele (T), allowing identification of line 206 as a carrier. Pool 32, which was unsuccessfully sequenced, should have also displayed a high minor allele count for line 206.
(c) Sanger sequencing of line 206, confirming the C→T heterozygous SNP at location 51.

DISCUSSION

This paper presents an initial proof-of-concept experiment for the ComSeq approach. ComSeq detected carriers of *de novo* rare SNPs in a TILLING population of 1024 lines using 48 pools. All detected carriers were validated. Although the number of targeted amplicons in this study was much smaller compared to previous multi-dimensional pooling approaches (Tsai *et al*., 2011; Marroni *et al*., 2012), the number of lines used (1024) and the efficiency factor (>20-fold) were greater than in any of these studies. The mutation rate of carriers was approximately 0.1%, lower than in previous studies. ComSeq has significant advantages over multi-dimensional pooling, allowing detection of several carriers at each locus, and is robust to common experimental noise (e.g. pool dropout). Similar to multi-dimensional pooling, ComSeq is readily scalable to any number of samples, and allows detection of carriers of both heterozygous and homozygous rare alleles. Following this experiment, we pooled an additional group of more than 3000 lines, which are all available to the community to further mine rare alleles and their carriers.

A major limiting factor of any pooling method relates to PCR-based targeting, as the amount of effort scales with the size of the region of interest. Recent developments in whole-genome sequencing have prompted utilization of exome capture for cataloging *de novo* SNPs in TILLING populations (Henry *et al*., 2014). Exome capture may replace PCR amplification to increase the number of targeted regions and provide an inventory of mutations in any gene of interest (Mascher *et al*., 2013).

The ability to perform site-directed mutagenesis (i.e. gene editing) in plants has advanced dramatically (Cong *et al*., 2013). ComSeq offers several advantages over gene editing in plants by allowing identification of many mutations in a

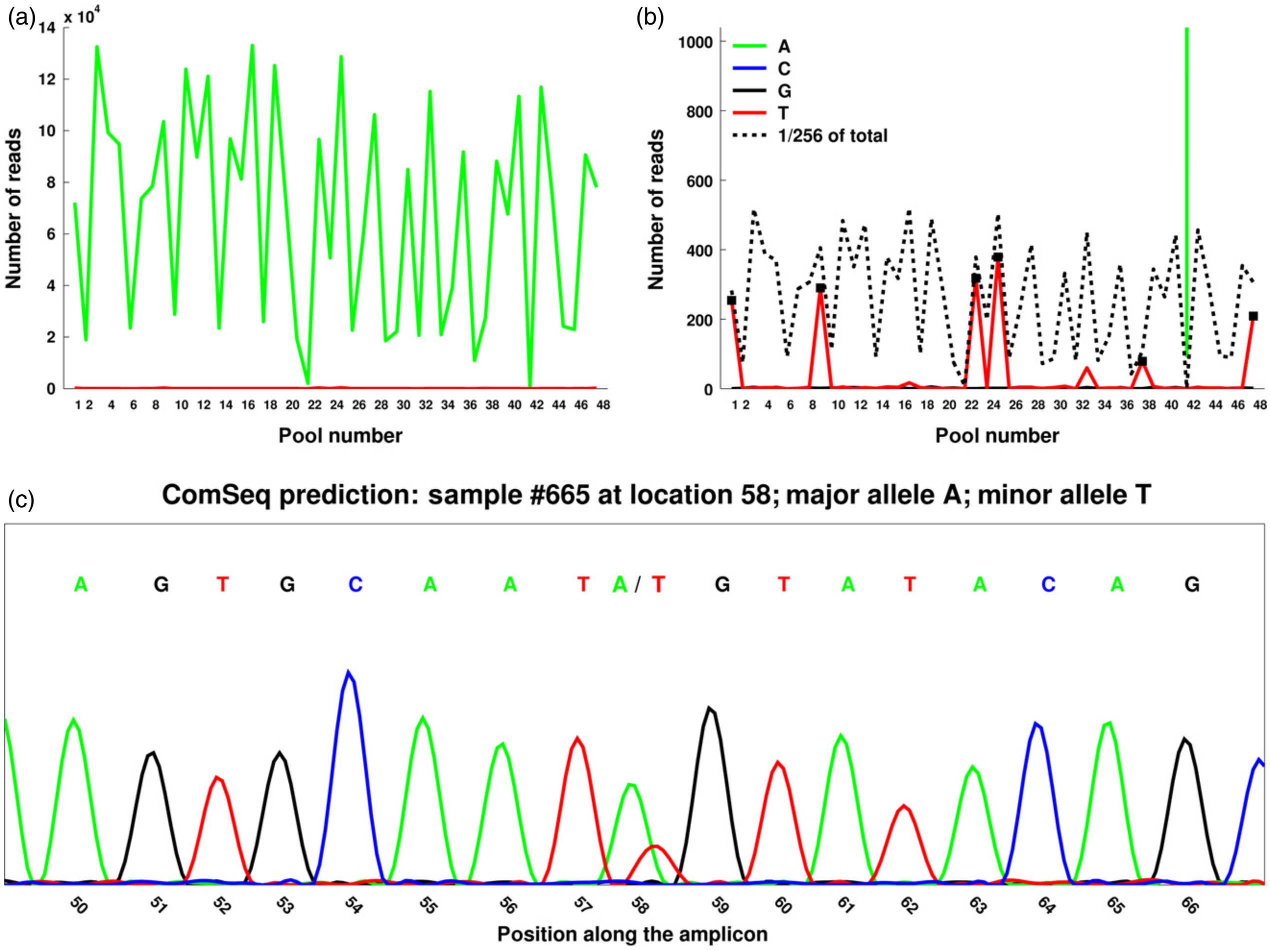

Figure 4. Detection of a *de novo* SNP in the NAC7 gene (Sb07g021400) and its carrier.
(a) Read counts for all nucleotides at position 58 along the amplicon. The major allele is A.
(b) Magnification of a low read count region. All six pools show a distinct read count for the minor allele (T), allowing identification of line 665 as a carrier. Note that although reads containing the minor allele are detected in pool 33, the ratio of T to A is too low to justify identification. Pool 41 was unsuccessfully sequenced for this target.
(c) Sanger sequencing of line 665, confirming the A→T heterozygous mutation at location 58.

gene of interest, with the ability to test them directly in the field without regulatory constraints (non-GMO plants), providing a highly relevant pipeline to bridge functional genomics and breeding. Furthermore, the identification of beneficial alleles via ComSeq provides targets for gene editing in other crops, such as *Saccharum officinarum* and *Miscanthus*, which are closely related to sorghum but have much more complex genomes (de Siqueira Ferreira *et al.*, 2013). Current gene editing methods in crop plants have yet to allow specific editing of genes other than complete knockout. Therefore, efficient mining of TILLING populations via ComSeq is highly relevant to provide functional allelic series for any gene of interest. Although the mutations identified in the regions targeted in this study were mostly synonymous SNPs, the findings demonstrate the ability of ComSeq to identify *de novo* mutations in an efficient manner, which may be applied to larger regions to identify carriers of non-synonymous SNPs. This is imperative for functional analysis, either for testing phenotypic and biochemical consequences of null alleles, or for identifying allelic series with quantitative differences between gene products relevant to naturally occurring DNA variation, e.g. the Brix9-2-5 allelic series for tomato (*Solanum lycopersicum*) (Fridman *et al.*, 2004). From a population genetics and clinical standpoint, ComSeq may prove highly useful for efficient screening of large human populations for carriers of either *de novo* or known rare SNPs.

EXPERIMENTAL PROCEDURES

Detecting rare alleles and their carriers via ComSeq

For completeness, we briefly describe compressed sensing and its application for detection of rare alleles. A detailed description is provided by Shental *et al.* (2010).

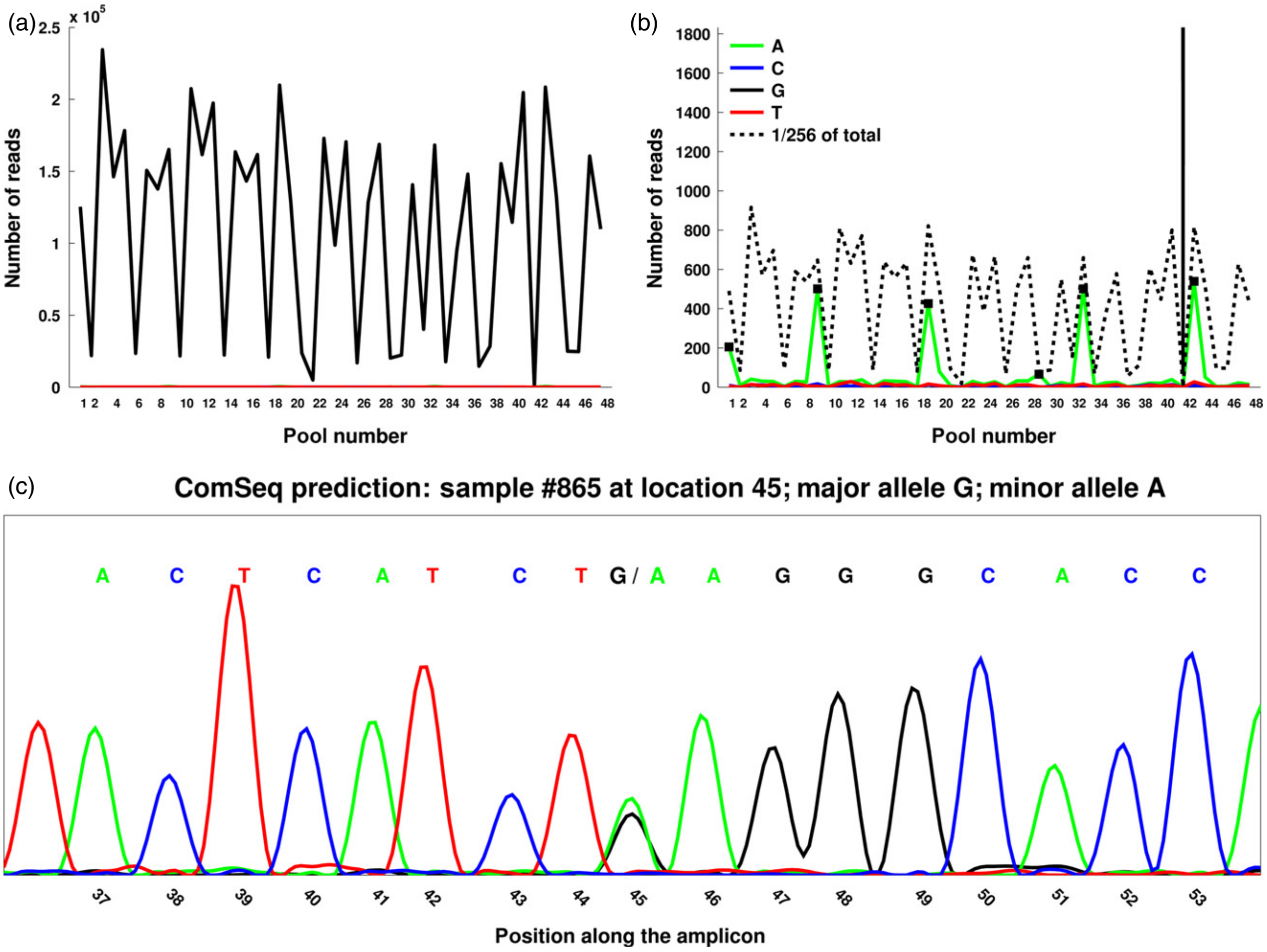

Figure 5. Detection of a second *de novo* SNP in the NAC7 gene (Sb07g021400) and its carrier.
(a) Read count at position 45. The major allele is G.
(b) Magnification of a low read count region. All six pools show a distinct read count for the minor allele (A), allowing identification of line 865 as a carrier. Pool 41 was unsuccessfully sequenced for this target.
(c) Sanger sequencing of line 865, confirming the G→A heterozygous SNP at location 45.

The task in compressed sensing is to efficiently reconstruct the unknown entries of a vector $\boldsymbol{x}$ of length n, assuming that $\boldsymbol{x}$ is sparse, i.e. most its entries are zero. In our case, the entries of $\boldsymbol{x}$ are the alternative allele counts for each sample at a specific locus. The possible entries of $\boldsymbol{x}$ are 0, 1 and 2, but most entries of the n samples are zero, as we seek only rare alleles.

Reconstruction of $\boldsymbol{x}$ is performed using $k \ll n$ basic operations, termed ‘measurements’, where a measurement is simply the output y of the dot product of the (unknown vector) $\boldsymbol{x}$ with a known measurement vector $\boldsymbol{m}$, i.e. $y = \boldsymbol{m}{\cdot}\boldsymbol{x}$. The vector $\boldsymbol{m}$ represents the pool, where $m_{ij} = 1$ if sample j belongs to the ith pool and zero otherwise. When normalizing $\boldsymbol{m}$, namely when dividing each entry of $\boldsymbol{m}$ by twice the number of samples in the pool, the value for y corresponds to the fraction of the number of rare alleles in a pool out of the total number of alleles. The experimental value for y in our case is simply the ratio between the number of alternative allele reads and the total number of reads at the locus of interest. This ratio in all pools is a vector $\boldsymbol{z}$.

It has been rigorously proven (Candès, 2006; Donoho, 2006) that using, the values of these k measurements, and their corresponding $\boldsymbol{m}$ vectors, it is possible to reconstruct the original sparse vector $\boldsymbol{x}$, as the solution of the following optimization problem:

$$\boldsymbol{x}^* = \text{argmin}_{x\in\{0,1,2\}}||\boldsymbol{x}||_1 \quad \text{s. t.} \quad ||\boldsymbol{Mx} - \boldsymbol{z}||_2 < \epsilon \qquad (1)$$

where $\boldsymbol{M}$ is a matrix whose rows are the normalized vectors $\boldsymbol{m}$, and is termed the sensing matrix. The value of $\varepsilon > 0$ corresponds to the maximal level of noise that may be tolerated while still obtaining a sparse solution.

In this study, approximately equal amounts of DNA were taken from each line in a specific pool, and then the mixture was PCR-amplified. The individual pools were barcoded and sequenced on the same lane. The resulting reads were processed to locate positions along the amplicon that may contain a SNP, and the relevant carriers at each locus were detected as a solution to Eqn. (1).

The experimental design consisted of targeting regions in two genes over 1024 *Sorghum bicolor* lines, which were combined into 48 pools. We start by describing the specific sensing matrix and then describe library preparation, sequencing, and subsequent data analysis.

Design of the sensing matrix

The design of the sensing matrix ***M*** in this application was based on an RS error-correcting code (Reed and Solomon, 1960). The code was designed for 1024 samples and 48 pools, whereby each pool contained exactly 128 samples at the same concentration and each sample appeared in six pools (Figure S3). The code was specifically designed to detect a single carrier at each base pair along the amplicon given the expected mutation rate of less than 0.1%.

As opposed to the random matrix applied in our computational paper (Shental *et al.*, 2010), the RS code is deterministic, and offers several advantages over a random matrix, while preserving all compressed sensing prerequisites for a 'good' sensing matrix. First, each RS code has known pre-defined properties, namely a given design is able to correctly detect up to *c* carriers of *n* samples, with up to *l* pools failing. The required values of *n*, *c* and *l* may be tuned according to the experimental requirements, allowing straightforward scalability of the design. Second, an RS code constructs 'uniform' pools, namely a constant number of samples per pool and a constant weight, i.e. number of pools that each sample participates in. This property has proved useful for the robotic system Tecan Freedom EVO (www.lifesciences.tecan.com) that was used to pool samples, but is also important for manual pipetting (Zielinski *et al.*, 2014), during which sample evaporation may occur. A random matrix, as applied previously (Shental *et al.*, 2010) may, in principle, have these properties, but these have to be explicitly refined by hand, by adding or removing pools until uniform pooling is achieved. Such refinement is not required for RS codes.

Our specific design tolerates up to two missing measurements, i.e. a carrier may be detected if it appears in at least four of the six pools in which it participates. If it only appears in three pools, the code will detect two samples. Error correction requires an increase in the number of pools, resulting in a total of 48. This design still displays a more than 20-fold increase in efficiency compared to a naive approach. Including the additional pools needed for error correction also allows carrier detection despite experimental failures in pooling or sequencing that lead to sample or pool dropout.

DNA extraction

Twelve sprouting seeds from each of 6400 TILLING lines (generation M_4) (Xin *et al.*, 2008) were transferred into 2 ml tubes and arranged in a 96-well rack. The first 1024 DNA lines were used in this study. The seeds were freeze-dried using a Labconco FreeZone freeze dry system (www.labconco.com), and tissue was ground twice at 26 strokes per second for 90 sec using a TissueLyser (www.qiagen.com). Then 750 μl elution buffer [100 mM Tris, pH 8.0, 20 mM EDTA, 2% cetyltrimethyl ammonium bromide (CTAB) and 1.2 M NaCl] was added to the ground tissue and mixed in two orientations with elution buffer by grinding at 18 Hz for 1 min using a TissueLyser. After incubation for 1 h at 60°C with occasional mixing, 750 μl of chloroform/isoamylalcohol (24:1) was added, and samples were centrifuged at 2000 ***g*** at room temperature for 15 min. Then 500 μl of the top aqueous layer was transferred to a 96-well plate to which 1 ml CTAB dilution buffer (100 mM Tris pH 8.0, 20 mM EDTA and 2% CTAB) was added, and mixed three times using a 96-channel automatic pipette system (Apricot Designs (www.apricotdesigns.com/)). The plate was loosely capped, and incubated for 30 min at 60°C, followed by centrifugation at 2000 *g* at room temperature for 15 min to pellet the DNA. After the supernatant was removed, 500 μl wash buffer (70% EtOH, 10 mM TE) was added. Plates were capped and inverted gently to wash CTAB from the walls, followed by 30 min incubation at room temperature and centrifugation at 2000 *g* at room temperature for 15 min. Ethanol was removed and the plate was left to dry for 5 min at room temperature. The DNA pellet was then re-suspended in 195 μl TE (10 mM Tris, pH 8.0, 2 mM EDTA, 1 M NaCl) and 5μl Rnase A (stock of 10 mg/ml). The resulting suspension was incubated for 30 min at 60°C, followed by addition of 5 μl magnetic beads to each well and vortexing. Then 250 μl of 100% EtOH was added, and samples were incubated at room temperature for 5 min. Using a magnetic plate, the washing liquid was poured off, followed by four additional washes with 400 μl of 70% EtOH at room temperature to remove salts and contaminants. The liquid was removed, and samples were dried in a hood for 5 min. DNA was eluted by adding 200 μl TE to the beads and incubating for 5 min at 60°C. After removing the beads from solution using a magnet, DNA was transferred to a 1.2 ml collection microtube (Qiagen). Then 50 μl DNA was transferred to a UV plate (Costar (www.daigger.com)) to quantify DNA on a Tecan Infinite M200 plate reader using Magellan6 software (lifesciences.tecan.com) . DNA was diluted to 20 ng/μl prior to pooling and PCR.

Pooling of DNA samples

Equal quantities of 1024 DNA samples were combined into 48 pools using a Tecan Freedom EVO four-tip liquid handling arm. The COMT carrier (Xin *et al.*, 2008) replaced the last line in the pooling design to serve as a positive control.

The sample map spreadsheets were entered into a web application (pooling.teamerlich.org) that we created to generate pooling instructions for use with robotic or manual pipetting systems. The resulting CSV file was separated into four files with # pipetting steps % 6 = 0 to allow for pooling in 6–9 h stages (approximately 32 h total), and each file was sequentially imported into Tecan EVOware. The 96-well chimney plates containing 1024 DNA samples were sealed with foil to allow aspiration while preventing sample evaporation, and attached to the worktable in the configuration indicated by the pooling program. Pooling was performed using 50 μl filter tips (Tecan), and samples were dispensed into a single 96-well plate.

Genetic targets

Table S1 describes the targeted amplicons, including genomic coordinates and primers used for library preparation.

Sample library preparation for NGS

PCR was performed in two steps prior to paired-end sequencing on an Illumina HiSeq 2500 system (www.illumina.com).

Primary PCR.

Primers were designed to amplify a 200–250 bp amplicon around the target of interest. The primers were synthesized using the following adaptor sequences: 5′-ACACTCTTTCCCTACACGACGCTCTTCCGATCT-3′ (forward) and 5′-TGCTGTTGAC AGTGAGCG-3′ (reverse). Of the two pairs of primers designed for each target, the one with best specificity was chosen. For each pool, 50 ng DNA was input to a primary PCR reaction for each of the four loci tested. A total of two PCR plates were run for the 48 pools. The PCR reaction included 1 unit of Phusion Green Hot Start II High Fidelity DNA polymerase (Thermo Scientific (www.thermoscientific.com)), 0.2 mM of each dNTP, and 10 μM each of the forward and reverse primers. Each 50 μl reaction was heated to 98°C for 30 sec, followed by 21 cycles of denaturation at 98°C for 10 sec, annealing at 60°C for 30 sec and extension at 72°C for 20 sec, followed by a 10 min final extension at 72°C.

Clean-up. After primary PCR, product mixes were created for each original DNA pool by combining the four loci. Mixes were purified using Agencourt AMpure XP magnetic beads (Beckman Coulter (www.beckmancoulter.com)), according to the manufacturer's instructions. A ratio of 0.8:1 for magnetic beads to PCR product was used to remove primer dimers.

Secondary PCR and barcoding. Each clean mix was used in the second PCR reaction to barcode each pool. The primers used were as follows: 5′-CAAGCAGAAGACGGCATACGAGATCGGTCTC GGCATTCCTGCTGAACCGCTCTTCCGATCTXXXXXXXXtgctgttgacagtgagcg-3′ (forward), where the sequence of Xs represents the barcode sequence (see Table S2) and lower case represent the targeted amplicon, and 5′-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3′ (reverse). The PCR reactions (50 μl) were heated to 98°C for 30 sec, followed by 10 cycles of denaturation at 98°C for 10 sec, annealing at 55°C for 20 sec and extension at 72°C for 40 sec, followed by a final extension for 10 min at 72°C.

The final library was constructed by combining 10 μl of each of the 48 barcode reactions. Following quantification by gel electrophoresis, a 100 μl aliquot of this mix was purified using magnetic beads as previously described. PCR was performed using each of the four internal primers to ensure that all targets were present in the final library. A high-sensitivity DNA assay was performed using a 2100 Bioanalyzer (Agilent) (www.agilent.com) to assess the quality of the final library before NGS. The primer pairs generated primary PCR products of 302-419 bp (Table S1).

Sequencing

Libraries were sequenced on an Illumina HiSeq 2500 sequencing system at the Whitehead Institute (Cambridge, MA), yielding approximately 20 million 100 bp paired-end reads from a single lane (libraries were multiplexed with other unrelated samples). The total numbers of reads that were correctly assigned a barcode and aligned to each amplicon were 670 000, 2 500 000, 3 750 000 and 7 500 000 for *Lsi1*, *COMT*, NAC7 (CST9) and NAC7 (CST11), respectively.

Analysis

Data processing included a filtering step in which reads were assigned to amplicons and pools, and application of ComSeq to the potential loci to detect *de novo* SNPs and their carriers, as described below.

Filtering reads. We excluded reads that did not match any barcode or whose Phred quality score was less than 30 (error probability of 10^{-3}) anywhere along the eight-nucleotide barcode sequence. Reads were also discarded if more than 20% of its nucleotides had a Phred quality score lower than 30. Finally, for every position *i* along the read, we counted the number of appearances of each nucleotide, while only considering reads that match the wild-type until position i.

Detecting de novo SNPs. Using the counts for each nucleotide at each position along the read, we defined the minor (alternate) allele for each position, and kept sites for which the minor allele count was at least 1/512 (twice the number of alleles in each pool) of the total number of reads at that location in at least four pools (four of six pools are sufficient to detect carriers using our RS code). We then applied Eqn. (1) to detect the carrier at each of these sites.

The above procedure was independently performed for both paired-end reads. Six loci passed the filter (five heterozygous SNPs and a single homozygous SNP). Sequencing data, together with the MATLAB code used for analysis, are available at https://github.com/NoamShental/ComSeq_Sorghum_TILLING.

Sanger validation

DNA for Sanger sequencing was either extracted from the original lines or obtained from our DNA extractions. PCR amplicons were amplified using the primers used in the first step of library preparation (Table S1). Gel electrophoresis was performed for verification using 5 μl of PCR product. Excess primers and nucleotides were removed using an enzymatic PCR clean-up kit (Takara (www.clontech.com)): exonuclease 1 and shrimp alkaline phosphatase were mixed in equal amounts, and incubated with the PCR product (1:1:5 ratio) at 37°C for 20 min, followed by enzyme inactivation at 85°C for 15 min. Cleaned PCR products were Sanger sequenced at Hy-Labs (Rehovot, Israel) using forward and reverse primers.

ACKNOWLEDGMENTS

This research was supported by research grant award number IS-4447-11F from the US/Israel Binational Agricultural Research and Development Fund, and by research grant award number 837-0127-11 from the Israeli Ministry of Agriculture and Rural Development. H.N. was supported by the Pears Foundation (UK). We would like to thank the Whitehead Institute Genome Technology Core for their assistance in sequencing.

SUPPORTING INFORMATION

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

Figure S1. Detection of the second carrier of the SNP in the NAC7 gene (CST9).

Figure S2. Detection of the homozygous carrier of the SNP in the NAC7 gene.

Figure S3. Application of the RS code as a sensing matrix.

Table S1. Details of ComSeq target genes, primers and their genomic positions.

Table S2. Barcode sequences.

REFERENCES

Avni, R., Zhao, R., Pearce, S. *et al.* (2013) Functional characterization of GPC-1 genes in hexaploid wheat. *Planta*, **239**, 313–324.

Belhaj, K., Chaparro-Garcia, A., Kamoun, S. and Nekrasov, V. (2013) Plant genome editing made easy: targeted mutagenesis in model and crop plants using the CRISPR/Cas system. *Plant Methods*, **9**, 39.

Candès, E.J. (2006) Compressive sampling. In *Proceedings of the International Congress of Mathematicians*, **3**, 1433–1452.

Comai, L. and Henikoff, S. (2006) TILLING: practical single-nucleotide mutation discovery. *Plant J.* **45**, 684–694.

Comai, L., Young, K., Till, B.J. *et al.* (2004) Efficient discovery of DNA polymorphisms in natural populations by Ecotilling. *Plant J.* **37**, 778–786.

Cong, L., Ran, F.A., Cox, D. *et al.* (2013) Multiplex genome engineering using CRISPR/Cas systems. *Science*, **339**, 819–823.

Donoho, D.L. (2006) Compressed sensing. *IEEE Trans. Inf. Theory*, **52**, 1289–1306.

Erlich, Y., Gordon, A., Brand, M., Hannon, G.J. and Mitra, P.P. (2010) Compressed genotyping. *IEEE Trans. Inf. Theory*, **56**, 706–723.

Fridman, E., Carrari, F., Liu, Y.-S., Fernie, A.R. and Zamir, D. (2004) Zooming in on a quantitative trait for tomato yield using interspecific introgressions. *Science*, **305**, 1786–1789.

Greene, E.A., Codomo, C.A., Taylor, N.E. *et al.* (2003) Spectrum of chemically induced mutations from a large-scale reverse-genetic screen in Arabidopsis. *Genetics*, **164**, 731–740.

Henry, I.M., Nagalakshmi, U., Lieberman, M.C. *et al.* (2014) Efficient genome-wide detection and cataloging of EMS-induced mutations using exome capture and next-generation sequencing. *Plant Cell*, **26**, 1382–1397.

Jiao, X.Y., Burke, J., Chopra, R. *et al.* 2016. A *Sorghum* mutant resource as an efficient platform for gene discovery in grasses.

Kharabian-Masouleh, A., Waters, D.L.E., Reinke, R.F. and Henry, R.J. (2011) Discovery of polymorphisms in starch-related genes in rice germplasm by amplification of pooled DNA and deeply parallel sequencing. *Plant Biotechnol. J.* **9**, 1074–1085.

Ma, J.F., Yamaji, N., Mitani, N., Tamai, K., Konishi, S., Fujiwara, T., Katsuhara, M. and Yano, M. (2007) An efflux transporter of silicon in rice. *Nature*, **448**, 209–212.

Markovich, O., Kumar, S., Cohen, D., Addadi, S., Fridman, E. and Elbaum, R. (2015) Silicification in leaves of *Sorghum* mutant with low silicon accumulation. *Silicon*, **7**, 1–7.

Marroni, F., Pinosio, S. and Morgante, M. (2012) The quest for rare variants: pooled multiplexed next generation sequencing in plants. *Front. Plant Sci.* **3**, 1–9.

Mascher, M., Richmond, T.A., Gerhardt, D.J. *et al.* (2013) Barley whole exome capture: a tool for genomic research in the genus *Hordeum* and beyond. *Plant J.* **76**, 494–505.

McCallum, C.M., Comai, L., Greene, E.A. and Henikoff, S. (2000) Targeting Induced Local Lesions IN Genomes (TILLING) for plant functional genomics. *Plant Physiol.* **123**, 439–442.

Missirian, V., Comai, L. and Filkov, V. (2011) Statistical mutation calling from sequenced overlapping DNA pools in TILLING experiments. *BMC Bioinformatics*, **12**, 287.

Mitani, N., Yamaji, N. and Ma, J.F. (2009) Identification of maize silicon influx transporters. *Plant Cell Physiol.* **50**, 5–12.

Reed, I.S. and Solomon, G. (1960) Polynomial codes over certain finite fields. *J. Soc. Ind. Appl. Math.* **8**, 300–304.

Shental, N., Amir, A. and Zuk, O. (2010) Identification of rare alleles and their carriers using compressed se(que)nsing. *Nucleic Acids Res.* **38**, e179.

de Siqueira Ferreira, S., Nishiyama, M.Y., Paterson, A.H. and Souza, G.M. (2013) Biofuel and energy crops: high-yield Saccharinae take center stage in the post-genomics era. *Genome Biol.* **14**, 210.

Tsai, H., Howell, T., Nitcher, R. *et al.* (2011) Discovery of rare mutations in populations: TILLING by sequencing. *Plant Physiol.* **156**, 1257–1268.

Uauy, C., Distelfeld, A., Fahima, T., Blechl, A. and Dubcovsky, J. (2006) A NAC gene regulating senescence improves grain protein, zinc, and iron content in wheat. *Science*, **314**, 1298–1301.

Xin, Z., Wang, M.L., Barkley, N.A., Burow, G., Franks, C., Pederson, G. and Burke, J. (2008) Applying genotyping (TILLING) and phenotyping analyses to elucidate gene function in a chemically induced sorghum mutant population. *BMC Plant Biol.* **8**, 103.

Yamaji, N., Mitatni, N. and Ma, J.F. (2008) A transporter regulating silicon distribution in rice shoots. *Plant Cell*, **20**, 1381–1389.

Zaboli, G., Ameur, A., Igl, W. *et al.* (2012) Sequencing of high-complexity DNA pools for identification of nucleotide and structural variants in regions associated with complex traits. *Eur. J. Hum. Genet.* **20**, 77–83.

Zielinski, D., Gordon, A., Zaks, B.L. and Erlich, Y. (2014) iPipet: sample handling using a tablet. *Nat. Methods*, **11**, 784–785.