Mutation identification by direct comparison of whole-genome sequencing data from mutant and wild-type individuals using *k*-mers

Karl J V Nordström[1,4], Maria C Albani[1,4], Geo Velikkakam James[1], Caroline Gutjahr[2,3], Benjamin Hartwig[1], Franziska Turck[1], Uta Paszkowski[3], George Coupland[1] & Korbinian Schneeberger[1]

Genes underlying mutant phenotypes can be isolated by combining marker discovery, genetic mapping and resequencing, but a more straightforward strategy for mapping mutations would be the direct comparison of mutant and wild-type genomes. Applying such an approach, however, is hampered by the need for reference sequences and by mutational loads that confound the unambiguous identification of causal mutations. Here we introduce NIKS (needle in the *k*-stack), a reference-free algorithm based on comparing *k*-mers in whole-genome sequencing data for precise discovery of homozygous mutations. We applied NIKS to eight mutants induced in nonreference rice cultivars and to two mutants of the nonmodel species *Arabis alpina*. In both species, comparing pooled F_2 individuals selected for mutant phenotypes revealed small sets of mutations including the causal changes. Moreover, comparing M_3 seedlings of two allelic mutants unambiguously identified the causal gene. Thus, for any species amenable to mutagenesis, NIKS enables forward genetics without requiring segregating populations, genetic maps and reference sequences.

Forward genetic screens have been of fundamental importance in elucidating biological mechanisms in model species[1]. Their success, however, has relied on the feasibility of mutant gene isolation. Identification of causal mutations typically begins with genetic mapping, followed by candidate gene sequencing and complementation studies using transformation. Advances in DNA sequencing technologies have tremendously accelerated genetic mapping by combining bulk segregant analysis, that is, pooling recombinant genomes, with whole-genome sequencing, usually referred to as mapping by sequencing[2,3]. This approach is now becoming standard for mutation mapping and identification in many model species[3–12] and has even been applied to decipher quantitative traits with complex genetic architectures[13,14]. Recently, mutagen-induced changes have been used as novel markers, allowing mapping of mutations using isogenic mapping populations[10,15]. Nevertheless, all mapping-by-sequencing methods rely on resequencing, a method for whole-genome reconstruction based on aligning sequences to a reference sequence. Therefore, this requirement restricts the application of the technique to species for which such a reference genome sequence is available.

Many reference-sequence assembly projects are currently in progress, including ones for most of the major crop species and breeding animals. However, even with an existing reference sequence, extending mapping-by-sequencing methods beyond the sequenced reference accessions has proved technically challenging. Mutant alleles of genes that are not present in the reference sequence cannot be identified within resequencing data alone. In particular, fast-evolving genes, such as those involved in disease resistance, might not always be represented in the reference sequence[16,17].

Alternative solutions for mapping-by-sequencing in species without reference sequences have been proposed, such as mapping-by-sequencing based on reference sequences of related species or expressed sequence tag collections[11,18]. However, all of these methods greatly rely on low sequence divergence and high levels of synteny between the mutant genome and alignment target. Recently, methods for direct genome comparison of multiple samples without a reference sequence were introduced, but none has proven to be accurate and precise enough for the identification of mutations[19–21].

NIKS is a method for reference-free genome comparison based solely on the frequencies of short subsequences within whole-genome sequencing data. It is geared toward identifying mutagen-induced, small-scale, homozygous differences between two highly related genomes, independent of their inbred or outbred background, and provides a route to identification of mutations without requiring any prior information about reference sequences or genetic maps.

RESULTS

Principles and performance of NIKS

NIKS relies on the analysis of *k*-mers, which are defined as subsequences of length *k* of a sequencing read. NIKS starts by assessing the frequency of each *k*-mer within the sequencing data of each sample using the *k*-mer–counting software Jellyfish[22]. *K*-mers that overlap with sequencing errors will be of low frequency, as these errors are not present in all reads from the corresponding region, and it is therefore possible to separate them from reads that are error free (**Fig. 1**).

[1]Max Planck Institute for Plant Breeding Research, Cologne, Germany. [2]Institute of Genetics, Ludwig-Maximilians-University, Biocenter Martinsried, Martinsried, Germany. [3]Department of Plant Molecular Biology, University of Lausanne, Biophore, Lausanne, Switzerland. [4]These authors contributed equally to this work. Correspondence should be addressed to K.S. (schneeberger@mpipz.mpg.de).

Received 16 October 2012; accepted 22 January 2013; published online 10 March 2013; doi:10.1038/nbt.2515

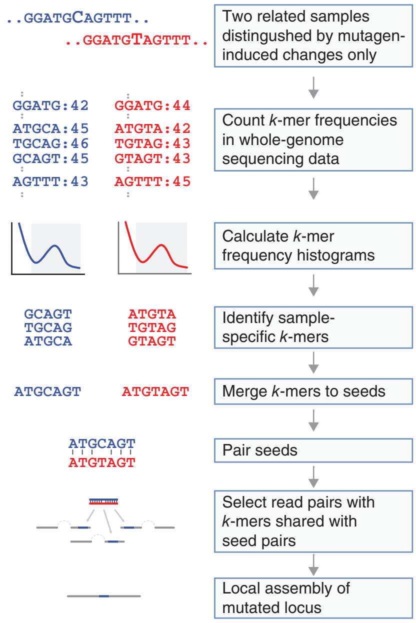
..GGATGCAGTTT..
..GGATGTAGTTT..
Two related samples distingushed by mutagen-induced changes only
GGATG:42
ATGCA:45
TGCAG:46
GCAGT:45
AGTTT:43
GGATG:44
ATGTA:42
TGTAG:43
GTAGT:43
AGTTT:45
Count k-mer frequencies in whole-genome sequencing data
Calculate k-mer frequency histograms
GCAGT
TGCAG
ATGCA
ATGTA
TGTAG
GTAGT
Identify sample-specific k-mers
ATGCAGT
ATGTAGT
Merge k-mers to seeds
ATGCAGT
ATGTAGT
Pair seeds
Select read pairs with k-mers shared with seed pairs
Local assembly of mutated locus

Figure 1 Workflow of NIKS. Whole-genome sequencing data of two related genomes is analyzed for the frequency of all *k*-mers. *K*-mer frequency histograms provide the power to distinguish between native *k*-mers (area highlighted in gray) and *k*-mers overlapping with sequencing errors. Comparing the two sets of *k*-mers of two highly related genomes discloses sample-specific, overlapping *k*-mers that result from subtle differences between the genomes. These sample-specific *k*-mers are then merged to a seed if they can be paired with a homologous, but not identical, seed in the other sample. Read pairs that share at least one *k*-mer with a seed pair will be used for local assemblies. This results in contigs that are centered on the mutated sites.

Amplification artifacts in sequencing can increase the frequency of error-based *k*-mers and have the potential to introduce error-based *k*-mers into the set of error-free *k*-mers. Thus, filtering for identical reads before running NIKS can reduce the impact of sequencing errors. Differences between genomes give rise to multiple, sample-specific and overlapping *k*-mers. NIKS identifies and then merges all sample-specific *k*-mers into longer sequences (or seeds).

To identify only those differences that were introduced during mutagenesis, NIKS considers only those seeds that feature a homologous, but not identical, seed in the second sample (seed pairing; **Supplementary Fig. 1a**). In fact, the two seeds of such a seed pair represent the wild-type and the mutant allele, respectively, and are distinguished only by the mutagen-induced mutations themselves. Mutations that are spaced by less than $k - 1$ bp and small indels will be combined in one elongated seed (**Supplementary Fig. 1b**,**c**). Larger indels might not result in one contiguous seed for the wild-type allele, but can be too complex to be assembled into one seed. Thus, NIKS additionally screens for other types of seed pairing. Seeds representing deletions can match up to two seeds in the other sample. These two seeds represent the breakpoints of the deletion in the wild-type sample. If both breakpoints of the deletion yield a seed, NIKS will identify a two-sided junction, and if only one breakpoint generates a seed it will produce a one-sided junction (**Supplementary Fig. 1d**,**e**).

Like homozygous mutations, heterozygous mutations and mutations in repetitive regions will introduce novel *k*-mers into the mutant sample. However, unlike for homozygous mutations, both the wild-type and the novel mutant allele will be present in the mutant sample. The presence of the wild-type allele in the mutant sample will then prevent the generation of a seed pair. In detail, NIKS records all *k*-mers that support the wild-type allele in the mutant sample and vice versa (mirror counts). Only mutations that feature less than a total of *k* *k*-mers from the respective other allele will be considered as homozygous.

In a last step, NIKS generates local *de novo* assemblies to extend the sequences associated with the mutated site[23]. Usually, this results in contigs of multiple hundreds of base pairs in length, including the mutated site. These contigs allow for functional annotation to assess the putative effect of mutations either by *de novo* gene predictions or alignments against known gene annotations.

To evaluate the performance of NIKS, we simulated 160 whole-genome sequencing experiments by first introducing ~2,000 random, homozygous, single-base mutations into the mouse and maize genome reference sequences and then sequencing at 17, 25, 35 and 50-fold genome coverage[24,25] (**Supplementary Notes**). The reference genomes are comparable in size—the mouse genome is 2.6 Gb and maize is 2.0 Gb—but differ drastically in their repeat content (**Supplementary Fig. 2**). To distinguish between mutations in unique regions and those in repeats, we classified each position into one of three classes according to the number of overlapping, repetitive *k*-mers (**Supplementary Notes**).

We compared the simulated sequencing data of each mutant to the sequencing data of the reference strain using NIKS (**Supplementary Fig. 3**). Within unique regions NIKS' sensitivity was >90%, and slightly increased with higher coverage levels. Notably, it was almost the same in the mouse and the maize experiments, indicating that sensitivity in unique regions, which typically represent large parts of the complement of genes, is not influenced by the overall repeat content of a genome. As expected, NIKS did not identify any of the mutations within repetitive regions, though it did identify some of the mutations in regions where repetitive and unique reads are present. From 25× genome coverage and up, the percentage of correctly identified mutations among all predictions (positive predictive value) was >98% across all experiments and backgrounds.

In addition to mutation identification, marker development in non-model organisms would profit from reference-independent methods. Applying NIKS to a whole-genome sequencing data set from two natural varieties of *Arabidopsis thaliana*, we were able to identify nearly 300,000 single-nucleotide polymorphisms (**Supplementary Table 1**). However, because of the high density of polymorphisms between natural accessions, dedicated tools, such as Cortex[21], may be more suitable for this task (**Supplementary Notes**).

Testing NIKS on seven previously analyzed rice mutants

Ethyl methanesulfonate (EMS) mutations have been identified by resequencing isogenic bulked, or pooled, recombinants of seven rice mutants[10]. To reduce the high load of mutations introduced by EMS, researchers[10] analyzed DNA pools of backcrossed recombinants rather than single-mutant genomes. The mutations were induced in the elite cultivar background of Hitomebore, which has no assembled reference sequence. By creating a pseudo-reference sequence and resequencing bulk segregant populations at coverage levels of 12–17×, the authors identified 3–11 putative candidate EMS mutations for six of the seven mutants.

We reanalyzed all seven samples with NIKS (**Supplementary Notes** and **Supplementary Fig. 4**). Because NIKS performs a two-sample comparison, each of the seven samples was compared to six other samples separately. After removing mutations identified in only one comparison and keeping only canonical EMS mutations, we retained 7–21 mutations per sample (**Table 1** and **Supplementary Tables 2** and **3**). To compare both analyses, we carried out a functional characterization of the mutations using the reference annotations of rice[26]. For four mutants, NIKS revealed the same candidate genes reported previously[10]. Among those was the experimentally validated mutation in the sample Hit1917-pl1.

Table 1 Number of canonical EMS mutations detected in seven rice mutants

		Without reference sequence (NIKS analysis[a])	With reference sequence (MutMap analysis[b,c])	
Sample	Coverage[c]	Without linkage	With linkage	Without linkage
Hit1917-pl1	~12.5	21 (4)	10 (7)	24 (14)
Hit0813-pl2	~13.7	20 (5)	7 (5)	10 (5)
Hit1917-sd	~16.6	7 (4)	9 (6)	10 (7)
Hit0746-sd	~14.0	16 (1)	3 (4)	13 (5)
Hit5500-sd	~13.2	12 (6)	4 (4)	9 (6)
Hit5814-sd	~14.2	10 (5)	4 (4)	4 (4)
Hit5243-sm	~14.1	19 (13)	11 (11)	21 (17)

[a]Numbers include all mutations that have been identified in more than one comparison. Values in brackets describe the number of mutations that appeared to be completely homozygous. [b]Values describe mutations with 90% mutant allele frequency. Values in brackets describe the number of mutations that appeared to be completely homozygous. [c]Values are taken from reference 10.

For two of the remaining mutants, our analysis revealed a distinct set of causal candidate genes. No candidate mutation had been previously reported for the last mutant. Our analysis revealed a mutation in an intron that is putatively retained in an alternatively spliced isoform, which is expressed in various developmental stages (**Supplementary Notes** and **Supplementary Fig. 5**). In summary, NIKS identified candidate mutations for all seven rice mutants, appearing to be at least as accurate in our hands as the previously published method[10], and using the reference sequence only for functional analysis.

Mutation identification without reference or genetic map

To perform NIKS on a undescribed mutant of a nonmodel species, we selected *A. alpina* mutants from a recent EMS mutagenesis screen[27]. *A. alpina* is a perennial Brassicaceae species with an estimated genome size of 375 Mb for which no reference sequence, annotation or genetic map is available. One of the selected mutants *floral defective 1* (*fde1*) displayed floral homeotic defects (**Fig. 2a**). A second mutant *perpetual flowering 1-1* (*pep1-1*) was previously shown, through a homology-based candidate gene approach, to carry a splice-site lesion in the *PEP1* gene that is responsible for the phenotype[27] (**Fig. 2b**). Comparing the two genomes using NIKS, we aimed to identify the unknown lesion in *fde1* and simultaneously confirm the *pep1-1* mutation.

To reduce the high load of mutations, we compared pooled DNA of backcrossed recombinants rather than single-mutant genomes, as described for the rice mutants[10] (**Supplementary Notes**). After backcrossing to the nonmutagenized progenitor, F_1 individuals were self-pollinated and DNA of 97 and 86 F_2 plants of *pep1-1* and *fde1*, respectively, with the mutant phenotype were pooled for whole-genome sequencing (**Supplementary Table 4**). Applying NIKS to whole-genome sequencing data at a genome fold coverage of 67 (*pep1-1*) and 158 (*fde1*), we found 29 seed pairs that represent homozygous mutagen-induced changes (**Supplementary Notes**, **Supplementary Fig. 6a**,**b**, **Supplementary Fig. 7** and **Supplementary Table 5**). To demonstrate that such high levels of coverage were not necessary to identify these mutations, we repetitively bootstrapped the data at different coverage levels and still retained, on average, >20 at 17× coverage (**Supplementary Notes** and **Supplementary Table 6**). All 29 changes were C:G ↔ T:A mutations, as expected for EMS mutagenesis[28]. Based on the known bias of EMS mutations, we defined all T and A alleles as mutant alleles, thereby assigning 16 mutations to *fde1* and 13 to *pep1-1*, respectively (**Table 2**). We have resequenced all 16 *fde1* mutations by PCR amplification of adjacent regions followed by Sanger sequencing and verified the presence of all predicted changes (**Supplementary Table 7**). Finally, local assemblies of all read pairs, that harbored *k*-mers, shared with the seed pairs reconstructed contigs up to 922 bp in length, surrounding all 29 mutations in both genomes. These were used to unambiguously identify the *fde1* mutation.

FDE1 is the ortholog of *A. thaliana AP2*

We annotated the effect of the 29 mutations using two independent methods. First, all 29 contigs were aligned against a collection of publicly available full-length reference sequences[29] (NCBI Genomic Reference Sequences). Surprisingly, we could not identify any reliable hits for seven of these contigs (e-value, 1e-05), whereas all others provided reliable hits against the reference sequence of *A. thaliana*, as well as against other genomes[30]. For uniformity, we considered only hits against *A. thaliana* (**Supplementary Notes**). Based on these alignments, we inferred the structure of putative genes of *A. alpina* and assigned a putative effect to each of these mutations (**Table 2**). Six of the *pep1-1* mutations overlapped with the annotation of genes. Among those, we identified the known causal mutation, aligned to a splice donor site of the *PEP1* ortholog *FLC*[27,31] (**Fig. 2c**). Six of the *fde1* mutations aligned against genes. One of them altered the putative amino acid sequence of the *A. alpina* homolog *APETALA 2* (*AaAP2*) (**Fig. 2d**). The *A. thaliana ap2* mutant alleles cause a flower deformation very similar to that observed in *fde1* (ref. 32) (**Fig. 2a**).

In a second, independent attempt to annotate the effect of mutations, we performed *de novo* gene predictions using the 29 contigs harboring the wild-type allele followed by the annotation of the effect of the respective mutation (**Table 2** and **Supplementary Notes**). Notably, 19 out of 22 predicted effects were identical to the homology-based annotation, including both causal changes. Three of them differed, although one showed homology to a transposable element, which was excluded in the homology-based annotation. In addition,

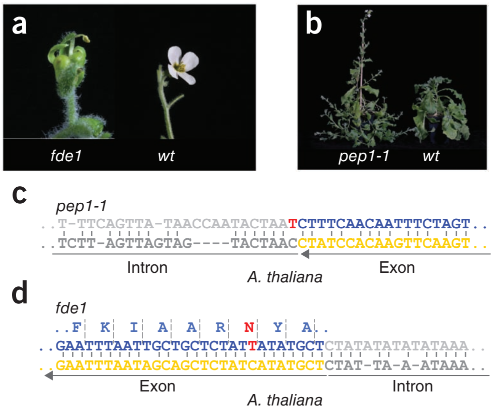
a
fde1
wt
b
pep1-1
wt
c
pep1-1
..T-TTCAGTTA-TAACCAATACTAATCTTTCAACAATTTCTAGT..
..TCTT-AGTTAGTAG----TACTAACCTATCCACAAGTTCAAGT..
Intron
A. thaliana
Exon
d
fde1
..F K I A A R N Y A..
..GAATTTAATTGCTGCTCTATTATATGCTCTATATATATATAAA..
..GAATTTAATAGCAGCTCTATCATATGCTCTAT-TA-A-ATAAA..
Exon
A. thaliana
Intron

Figure 2 *A. alpina* homeotic flower and flowering-time mutants. (**a**) *fde1* mutant flowers are deformed compared to flowers in wild-type plants. (**b**) *pep1-1* mutant plants flower without vernalization compared to wild type, which has an obligate requirement for vernalization to flower. (**c**) Alignment of a NIKS contig of *pep1-1* against the *A. thaliana* reference sequence reveals the splice-site mutation in the *A. alpina* ortholog of *FLOWERING LOCUS C* (*FLC*) previously reported to be responsible for the *pep1-1* phenotype. The mutant base is shown in red, the exonic regions in *A. thaliana* are shown in yellow, the inferred exon of *A. alpina* in blue and all noncoding nucleotides in gray. (**d**) Alignment of a NIKS contig of *fde1* against the *A. thaliana* reference sequence revealing a putative amino acid change likely to alter an aspartic acid to an asparagine within the 4th coding exon of the *A. alpina* ortholog of *AP2*. The frame used to translate the nucleotide sequence to an amino acid sequence was inferred from the annotation of *A. thaliana*. Causal mutations are shown in red. Exonic sequences associated with the mutation are shown in blue and exons of *A. thaliana* in yellow. Only relevant parts of the alignments are shown.

Table 2 Fixed genomic differences between bulked F_2 individuals of *pep1-1* and *fde1*

Allele		Contig assoc. with mutation				Homology-based annotation[a]					
pep1-1	*fde1*	Mutant genome	Length (bp)	Mutation position	Mirror count[c]	Homolog	Chr.	Position[d]	Effect	*De novo* annotation[b]	Agreement of annotations
T	C	*pep1-1*	549	437	0	–	2	5,482,633	Intergenic	None	Yes
T	C	*pep1-1*	625	236	0	AT5G08160	5	2,627,760	Syn. (L > L)	Syn. (L > L)	Yes
T	C	*pep1-1*	807	410	0	AT5G08510	5	2,754,082	Intronic	Intronic	Yes
A	G	*pep1-1*	829	354	0	AT5G09670	5	~2,998,250	Exonic	None	No
A	G	*pep1-1*	889	408	0	AT5G10140	5	3,175,363	Splice-site change	Splice-site change	Yes
T	C	*pep1-1*	812	451	0	–	5	~3,219,708	Intergenic	None	Yes
T	C	*pep1-1*	653	220	17	AT5G10550	5	3,333,724	Syn. (R > R)	Syn. (R > R)	Yes
A	G	*pep1-1*	783	437	0	–	5	3,336,193	Intergenic	None	Yes
T	C	*pep1-1*	780	348	26	AT5G11850	5	3,818,093	Nonsyn. (G > D)	Nonsyn. (G > D)	Yes
A	G	*pep1-1*	882	445	29	–	5	10,116,108	Intergenic	Nonsyn. (*F* > S)	No
A	G	*pep1-1*	732	368	0	AT5G44050	5	17,725,725	Intronic	Intronic	Yes
A	G	*pep1-1*	772	341	21	–	–	–	–	None	NA
A	G	*pep1-1*	850	448	0	–	–	–	–	Nonsyn. (E > K)	NA
G	A	*fde1*	828	410	0	AT4G34320	4	16,422,853	Nonsyn. (Q > STOP)	Nonsyn. (Q > STOP)	Yes
G	A	*fde1*	745	361	0	AT4G35230	4	16,756,818	Intronic	Intronic	Yes
G	A	*fde1*	637	261	0	–	4	~17,051,245	Intergenic	None	Yes
G	A	*fde1*	806	388	0	–	4	17,135,887	Intergenic	None	Yes
C	T	*fde1*	819	388	1	AT4G36360	4	17,178,292	Nonsyn. (G > E)	Nonsyn. (G > E)	Yes
C	T	*fde1*	863	427	0	AT4G36660	4	~17,286,500	Intronic	Nonsyn. (E > K)	No
C	T	*fde1*	764	313	10	–	4	17,357,762	Intergenic	None	Yes
C	T	*fde1*	880	454	0	AT4G36920	4	17,401,794	Nonsyn (D > N)	Nonsyn (D > N)	Yes
G	A	*fde1*	798	385	4	–	4	17,460,182	Intergenic	None	Yes
C	T	*fde1*	789	353	0	AT4G37080	4	17,475,571	Nonsyn. (A > T)	Nonsyn. (A > T)	Yes
G	A	*fde1*	863	429	0	–	4	~17,729,980	Intergenic	None	Yes
G	A	*fde1*	745	381	0	–	–	–	–	None	NA
C	T	*fde1*	903	476	0	–	–	–	–	None	NA
G	A	*fde1*	503	79	0	–	–	–	–	None	NA
G	A	*fde1*	630	254	0	–	–	–	–	None	NA
G	A	*fde1*	922	455	30	–	–	–	–	None	NA

Chr., chromosome identifier; syn., synonymous base substitution; nonsyn., nonsynonymous base substitution; NA, not applicable.

[a]Contigs were aligned against NCBI Genomic Reference Sequences. All contigs that had a reliable hit against one of the genomes also featured a reliable hit against *A. thaliana*. For uniformity we report only the hits against the *A. thaliana* reference sequence. [b]*De novo* gene annotation was performed on the contigs using Augustus annotation tool, mutations were annotated after their effect on this gene structure. [c]Mirror count describes the number of wild-type *k-mers* that were identified within the mutant samples while generating the respective seed. It indicates the presence of low fractions of wild-type alleles and can be used for prioritizing mutations. [d]Marked positions indicate contig alignments for which only regions flanking the mutation mapped to *A. thaliana*. Hence, these inferred positions are estimates.

for one of the seven contigs without a significant homology, a gene model was predicted.

To confirm that the mutation in *AaAP2* caused the flower phenotype of the *fde1* mutant, we further characterized three additional *A. alpina* mutants, designated *fde1-2*, *fde1-3* and *fde1-4*, displaying the same flower deformation as *fde1* (**Supplementary Fig. 8a–c**). The assembled contig sequence surrounding the *AaAP2* mutation was used to design primers to amplify *AP2* cDNA from the additional mutants. *fde1-2* contained a mutation that causes an amino acid change; *fde1-3*, a mutation that causes an early stop codon; and *fde1-4*, a splice site mutation, which was close to the original *fde1* mutation (**Supplementary Notes**). The presence of mutations in *AaAP2* in four of the isolated mutants corroborated that we had identified the causal mutation (**Supplementary Fig. 8d**).

Thus, NIKS correctly identified both causal mutations among a small set of candidate mutations, which could be functionally annotated through public databases and independently annotated *de novo* (**Table 2**).

NIKS identifies footprints of a 169-kb deletion

In contrast to chemically induced mutations, those induced by fast neutron radiation (FNR) can be more complex, such as deletions that span multiple kilobases[33]. To determine NIKS' power to identify complex changes, we applied it to the genome of the rice mutant *hebiba*, which was isolated from an FNR-mutagenized population of *Oryza sativa* ssp. *japonica* cv. Nihonmasari and displayed perturbed seedling photomorphogenesis[34]. Sixty-two *hebiba* mutants, identified from a segregating population, and a pool of 100 wild-type genomes

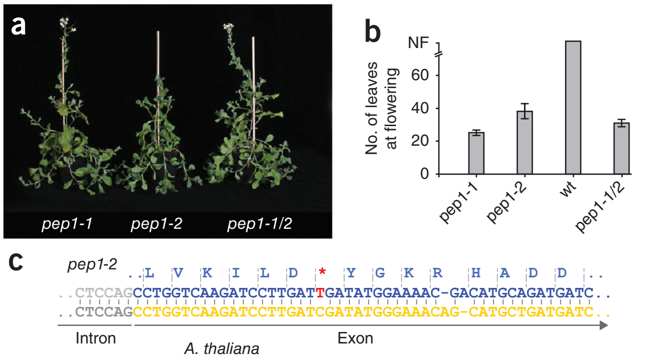

Figure 3 The *pep1-2* mutant is allelic to *pep1-1*. (**a**) *pep1-1* and *pep1-2* flower without vernalization. *pep1-2* does not complement the early flowering phenotype of *pep1-1*. F_1 (*pep1-1/2*) plants, resulted from the cross of *pep1-1* with *pep1-2*, flower in long-day conditions suggesting that they carry independent lesions within the same gene. (**b**) Flowering time analysis measured as number of leaves at flowering. *pep1-1*, *pep1-2* and their F_1 hybrid flowered whereas wild-type plants never flower in long days. Error bars, mean ± s.d, $n = 12$. (**c**) Alignment of a NIKS contig of *pep1-2* against the *A. thaliana* reference sequence identifies a premature stop codon introduced in the second exon of *PEP1*. The frame used to annotate the nucleotide sequence was inferred through alignment to *A. thaliana*. The causal mutation is shown in red. Exonic sequence associated with the mutation is shown in blue and the homologous exon of *A. thaliana* in yellow.

Table 3 Genes with independent lesions in *pep1-1* and *pep1-2*

Allele		Contig assoc. with mutation				Homology-based annotation[a]			
pep1-1	*fde1*	Mutant genome	Length (bp)	Mutation position	Mirror count[b]	Homolog	Chr.	Position	Effect
T	C	*pep1-2*	404	261	0	AT2G24680	2	10,495,120	Intronic
C	T	*pep1-1*	640	267	0	AT2G24680	2	10,496,039	Intronic
A	G	*pep1-1*	552	280	0	AT4G11670	4	7,048,540	Intronic
G	A	*pep1-2*	441	161	0	AT4G11670	4	7,050,257	Intronic
T	C	*pep1-1*	532	267	0	AT5G10140	5	3,175,363	Splice-site change
C	T	*pep1-2*	660	320	0	AT5G10140	5	3,175,642	Nonsyn. (R > STOP)

Chr., chromosome identifier; nonsyn., nonsynonymous base substitution.

[a]Contigs were aligned against NCBI Genomic Reference Sequences. All contigs that had a reliable hit against one of the genomes also featured a reliable hit against *A. thaliana*. For uniformity we report only the hits against the *A. thaliana* reference sequence. [b]Mirror count describes the number of wild-type *k*-mers that were identified within the mutant samples while generating the respective seed. It indicates the presence of low fraction of wild-type alleles and can be used for prioritizing mutations.

were sequenced up to 80.7- and 99.5-fold coverage, respectively, and compared against each other (**Supplementary Notes**, **Supplementary Fig. 9** and **Supplementary Tables 4** and **5**). In addition to 92 small-scale changes, NIKS identified footprints for three large-scale changes, represented by three contigs assembled in the mutant sample (**Supplementary Notes** and **Supplementary Table 8**). Two of these recapitulated single-base mutations in repetitive regions. To annotate a potential large-scale disruption identified by the remaining contig, we aligned the contig against the reference sequence using BLAST, which revealed a 169-kb deletion that was subsequently confirmed by PCR (**Supplementary Notes** and **Supplementary Fig. 10**). Simultaneously with our work, it has been shown that the *hebiba* phenotype is, in fact, induced by this deletion[35].

Owing to the complexity and length of the deleted sequence, NIKS could not assemble the complete 169-kb sequence of the wild-type allele, and a reference sequence was required for the identification of the deletion. However, NIKS did correctly assemble the newly generated junction at the deletion site into a 587-bp contig.

Mutation identification by analyzing two independent alleles

Backcrossing to wild-type progenitors has been successfully used to reduce the high number of noncausal EMS mutations[10]. However, this requires at least two more generations and is not practical for species with long generation times. Recently, we proposed that the most straightforward approach for the identification of mutations would be direct sequencing of two or more allelic mutants[36]. Although each mutant genome would feature many mutations, only genes with lesions in all mutants need to be considered as candidates (**Supplementary Fig. 11**). The expected random overlap of disrupted genes between two independently mutagenized genomes is extremely small and effectively disappears when more than two alleles are considered (**Supplementary Notes** and **Supplementary Fig. 12**).

As a proof-of-concept demonstration, we compared the genomes of *pep1-1* and *pep1-2*, which were isolated in the same EMS screen[27] and flower without vernalization. Both mutants share lesions in the same gene and form a complementation group; F_1 generation plants resulting from a cross between *pep1-1* and *pep1-2* also flower without vernalization. The mutation in *pep1-2* was confirmed by targeted sequencing of *PEP1*, revealing an early stop codon in the second exon (**Supplementary Notes** and **Fig. 3**).

We generated 51- and 105-fold whole-genome coverage sequencing data from pools of 35 M_3plants, which were derived from seeds after self-pollination of the original mutant plants, of *pep1-1* and *pep1-2*, respectively. NIKS identified 779 seed pairs, each of them revealing one genomic difference between *pep1-1* and *pep1-2* (**Supplementary Table 9** and **Supplementary Fig. 6c,b**). Of these 779 changes, only five lacked the canonical EMS change. All others could be assigned to either of the mutants and were annotated by aligning the mutation-associated contigs against the reference sequence of *A. thaliana*. Of the mutant contigs of *pep1-1* and *pep1-2*, 165 and 94, respectively, aligned to genes. However, only three of these genes were common for both mutants (**Table 3**). Functional annotation revealed only one gene, for which the mutations in both genomes affected the integrity of the gene. This gene was *PEP1*, the *A. alpina* ortholog of *FLC* and causal for the phenotype of *pep1-1* and *pep1-2* (ref. 27).

DISCUSSION

We introduced strategies for identification of mutagen-induced homozygous changes in unique regions in both of two highly related genomes, such as mutant and wild-type genomes, without requiring a reference genome. For this purpose, we present a whole-genome comparison method, NIKS, which accurately predicts subtle differences based on whole-genome sequencing data alone.

However, mutants may contain hundreds of mutagen-induced changes that hamper the direct readout of causal mutations. One way to reduce the number of mutations is backcrossing to the wild-type, followed by sequencing of pooled genomes of F_2 individuals[10,15]. With this strategy we identified an unknown mutation in *A. alpina*. To investigate whether this would have been possible with conventional means, we performed short-read alignment against the reference sequence of *A. thaliana* and *de novo* assembly followed by whole-genome comparisons, but with neither of these attempts was it possible to unambiguously identify all mutagen-induced changes (**Supplementary Notes**).

To classify genomic differences as mutant or wild-type allele, we used mutagen-specific biases for the type of mutation they introduce. Of all changes identified in our *A. alpina* data, 98.8% were canonical EMS changes. However, even if the mutagen does not introduce a bias in the mutational spectrum, analyzing multiple mutants or the wild-type genome unambiguously identifies mutant and wild-type alleles. In the current implementation, NIKS supports the comparison of two samples at a time, which conceptually could be expanded to an unlimited number of samples, similar to the approaches implemented in Cortex, a sophisticated usage of colored de Bruijn graphs[21].

Cortex and similar tools facilitate simultaneous whole-genome assembly and variation identification[19–21]. Whole-genome assembly structures can leverage the identification of complex differences between the samples, which is a nontrivial task when considering genome structures such as repeats and low-complexity regions. Such tools can additionally integrate a reference sequence into the assembly structure and thereby further compensate for difficulties while assembling long deletions *de novo*. Even without a reference sequence, NIKS identified the footprints of one large-scale deletion within a rice mutant, but it assembled only the mutant allele correctly. The reconstruction of the respective wild-type allele still required a reference sequence.

If one wants to pinpoint the causal mutation among a set of candidate mutations, it is helpful to annotate the effect of mutations on genes, either by imputing public gene annotation[10] or *de novo* gene annotation. *De novo* gene predictions turned out to be very similar to homology-based gene predictions, indicating that annotation of the effect of mutations can be done independently of prior knowledge.

Another solution to reduce the number of candidate mutations emerges by direct sequencing of genomes carrying two or more independent alleles, followed by a subsequent search for genes that harbor unique mutations in the same gene in multiple genomes. In a proof-of-principle experiment, we screened the M_3 genomes of two allelic *A. alpina* mutants for common genes carrying mutations, and we were able to unambiguously identify the causal gene. This analysis required prior identification of different alleles, which might not be straightforward for all species and phenotypes, but on the other hand, bypasses the generation of mapping populations. To our knowledge, this describes the first successful report of mutation identification using whole-genome sequencing of mutant genomes of a complementation group without relying on any kind of recombination.

Our method opens up many possibilities of forward genetics supported by whole-genome sequencing for any species that is amenable to mutagen screens. Thus, the approach has the potential to ease the access to all those species without reference sequences whose complex genomes defy current assembly methods. However, the largest impact might come from the use of NIKS to clone genes from nonmodel species that exhibit important traits[36].

Given that sequencing costs continue to decline, we anticipate that the genomes of all mutants of a forward genetic screen can be sequenced. By knowing all genomic and phenotypic differences, researchers will be able to group mutants into putative *in silico* allelic groups. With this information at hand, a small number of targeted complementation tests will serve as a first line of mutation validation for all those genes that feature multiple alleles in the screen. Depending on the size of the screen and on the number of genes that contribute to the phenotype of interest, this can be a powerful way to identify multiple mutants simultaneously. For example, simulations of forward genetic screens show that the analysis of 100 mutant genomes would be enough to identify over 27 allelic groups, assuming a screen size of 40,000 individuals and a phenotype that includes 75 genes (**Supplementary Fig. 13**). NIKS, thus, has the potential to reduce the workload of mutant identification to nothing more than whole-genome sequencing and comparison.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession code. Short read data have been deposited in the EBI short read archive under accession number ERP002265.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

The authors would like to thank D. Weigel for helpful discussions and suggestions on the manuscript, G. Marcais for support in using the Jellyfish software, P. Nick for providing *hebiba* mutant seeds, M. Riemann for support during the *hebiba* mutant analysis and the Max Planck Genome Center Cologne for their help on mutant sequencing. This work has been supported by grant DFG SPP1530 ("Flowering time control: from natural variation to crop improvement") (M.C.A. and K.S.).

AUTHORS CONTRIBUTIONS

K.S. and G.C. designed the study. K.J.V.N. designed and implemented NIKS. M.C.A., C.G., B.H., U.P. and F.T. provided biological material and performed experiments. K.J.V.N., G.V.J. and K.S. performed data analysis. K.S. wrote the manuscript with contributions from all authors.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Page, D.R. & Grossniklaus, U. The art and design of genetic screens: *Arabidopsis thaliana*. *Nat. Rev. Genet.* **3**, 124–136 (2002).
2. Michelmore, R.W., Paran, I. & Kesseli, R.V. Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. *Proc. Natl. Acad. Sci. USA* **88**, 9828–9832 (1991).
3. Schneeberger, K. *et al.* SHOREmap: simultaneous mapping and mutation identification by deep sequencing. *Nat. Methods* **6**, 550–551 (2009).
4. Blumenstiel, J.P. *et al.* Identification of EMS-induced mutations in *Drosophila melanogaster* by whole-genome sequencing. *Genetics* **182**, 25–32 (2009).
5. Cuperus, J.T. *et al.* Identification of MIR390a precursor processing-defective mutants in *Arabidopsis* by direct genome sequencing. *Proc. Natl. Acad. Sci. USA* **107**, 466–471 (2010).
6. Birkeland, S.R. *et al.* Discovery of mutations in *Saccharomyces cerevisiae* by pooled linkage analysis and whole-genome sequencing. *Genetics* **186**, 1127–1137 (2010).
7. Zuryn, S., Le Gras, S., Jamet, K. & Jarriault, S. A strategy for direct mapping and identification of mutations by whole genome sequencing. *Genetics* **186**, 427–430 (2010).
8. Laitinen, R.A.E., Schneeberger, K., Jelly, N.S., Ossowski, S. & Weigel, D. Identification of a spontaneous frame shift mutation in a nonreference *Arabidopsis* accession using whole genome sequencing. *Plant Physiol.* **153**, 652–654 (2010).
9. Austin, R.S. *et al.* Next-generation mapping of *Arabidopsis* genes. *Plant J.* **67**, 715–725 (2011).
10. Abe, A. *et al.* Genome sequencing reveals agronomically important loci in rice using MutMap. *Nat. Biotechnol.* **30**, 174–178 (2012).
11. Galvão, V.C. *et al.* Synteny-based mapping-by-sequencing enabled by targeted enrichment. *Plant J.* **71**, 517–526 (2012).
12. Leshchiner, I. *et al.* Mutation mapping and identification by whole genome sequencing. *Genome Res.* **22**, 1541–1548 (2012).
13. Ehrenreich, I.M. *et al.* Dissection of genetically complex traits with extremely large pools of yeast segregants. *Nature* **464**, 1039–1042 (2010).
14. Turner, T.L., Stewart, A.D., Fields, A.T., Rice, W.R. & Tarone, A.M. Population-based resequencing of experimentally evolved populations reveals the genetic basis of body size variation in *Drosophila melanogaster. PLoS Genet.* **7**, e1001336 (2011).
15. Hartwig, B., James, G.V., Konrad, K., Schneeberger, K. & Turck, F. Fast Isogenic mapping-by-sequencing of ethyl methanesulfonate-induced mutant bulks. *Plant Physiol.* **160**, 591–600 (2012).
16. Cai, D. *et al.* Positional cloning of a gene for nematode resistance in sugar beet. *Science* **275**, 832–834 (1997).
17. Song, J. *et al.* Gene RB cloned from Solanum bulbocastanum confers broad spectrum resistance to potato late blight. *Proc. Natl. Acad. Sci. USA* **100**, 9128–9133 (2003).
18. Wurtzel, O., Dori-Bachash, M., Pietrokovski, S., Jurkevitch, E. & Sorek, R. Mutation detection with next-generation resequencing through a mediator genome. *PLoS ONE* **5**, e15628 (2010).
19. Peterlongo, P., Schnel, N., Pisanti, N., Sagot, M.F. & Lacroix,, V. Identifying SNPs without a reference genome by comparing raw reads. in *String Processing and Information Retrieval* (eds. Chavez, E. & Lonardi, S.) 147–158 (Springer, 2010).
20. Ratan, A., Zhang, Y., Hayes, V.M., Schuster, S.C. & Miller, W. Calling SNPs without a reference sequence. *BMC Bioinformatics* **11**, 130 (2010).
21. Iqbal, Z., Caccamo, M., Turner, I., Flicek, P. & McVean, G. *De novo* assembly and genotyping of variants using colored de Bruijn graphs. *Nat. Genet.* **44**, 226–232 (2012).
22. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
23. Zerbino, D.R. & Birney, E. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821–829 (2008).
24. Church, D.M. *et al.* Lineage-specific biology revealed by a finished genome assembly of the mouse. *PLoS Biol.* **7**, e1000112 (2009).
25. Schnable, P.S. *et al.* The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**, 1112–1115 (2009).
26. International Rice Genome Sequencing Project. The map-based sequence of the rice genome. *Nature* **436**, 793–800 (2005).
27. Wang, R. *et al.* PEP1 regulates perennial flowering in Arabis alpina. *Nature* **459**, 423–427 (2009).
28. Greene, E.A. *et al.* Spectrum of chemically induced mutations from a large-scale reverse-genetic screen in *Arabidopsis. Genetics* **164**, 731–740 (2003).
29. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
30. The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis* thaliana. *Nature* **408**, 796–815 (2000).
31. Wang, R. *et al.* Aa TFL1 confers an age-dependent response to vernalization in perennial Arabis alpina. *Plant Cell* **23**, 1307–1321 (2011).
32. Bowman, J.L., Smyth, D.R. & Meyerowitz, E.M. Genes directing flower development in *Arabidopsis. Plant Cell* **1**, 37–52 (1989).
33. Li, X. *et al.* A fast neutron deletion mutagenesis-based reverse genetics system for plants. *Plant J.* **27**, 235–242 (2001).
34. Riemann, M. *et al.* Impaired induction of the jasmonate pathway in the rice mutant hebiba. *Plant Physiol.* **133**, 1820–1830 (2003).
35. Riemann, M. *et al.* Identification of rice ALLENE OXIDE CYCLASE mutants and the function of jasmonate for defence against Magnaporthe oryzae. *Plant J.* (accepted) http://dx.doi.org/10.1111/tpj.12115 (24 January 2013).
36. Schneeberger, K. & Weigel, D. Fast-forward genetics enabled by new sequencing technologies. *Trends Plant Sci.* **16**, 282–288 (2011).

ONLINE METHODS

NIKS implementation. NIKS is a Java and bash-based pipeline and implements all analysis steps performed in this study, including *k*-mer counting and selection, seed generation and pairing, and mutation-centric contig assembly. NIKS can be downloaded from https://sourceforge.net/projects/niks/.

***k*-mer counting and selection.** NIKS starts by assessing the frequency of all *k*-mers within whole-genome sequencing data of each sample using the efficient *k*-mer–counting software Jellyfish[22]. For $k > 31$ NIKS implements simple counting and sorting algorithms. Here we used $k = 31$ to make *k* as large as possible, while still being able to utilize the advantages of Jellyfish. In general, *k* needs to be smaller than read size, but large enough to allow for unique assignments to the genome. *K*-mers that overlap with sequencing errors will be of no use, which implies a *k* value significantly smaller than read length. Thus, the optimal choice of *k* depends on read quality and genome complexity, and cannot be generalized. In practical applications, however, it might be advantageous to run NIKS with multiple different values for *k*.

Defining two-thirds of the first local minimum in a histogram of *k*-mer frequencies as an intrinsic cut-off, NIKS can distinguish between native and sequencing error–based *k*-mers (**Fig. 1**). Insufficient coverage will hamper identification of this local minimum and make NIKS consider true all *k*-mers that occur at least twice. This allows reliable querying of samples for the presence or absence of *k*-mers. NIKS then extracts all *k*-mers present in one sample, but not in the other (sample-specific *k*-mers).

Seed generation and pairing. Single point mutations give rise to *k* sample-specific, overlapping *k*-mers. Joining *k*-mers that overlap by $k - 1$ bases merges sample-specific *k*-mers to seeds. For this, NIKS selects all *k*-mers of a sample that share at least one *m*-mer with the sample-specific *k*-mers of the other sample, where *m* is smaller than *k* (here, $k = 31$ and $m = 25$). Thereby, NIKS can preselect seeds that will yield valid seed pairs while allowing for disruptions at each end of a seed (**Supplementary Fig. 1**). This set of *k*-mers is then screened for *k*-mers that do not overlap with other sample-specific *k*-mers on one side, in order to identify putative start or end points of seeds (end *k*-mers). Starting with each end *k*-mer, an exhaustive walk, by combining *k*-mers that overlap by $k - 1$ bases, is conducted without adding the same *k*-mer twice and stopping if another end *k*-mer is reached. To avoid repetitive regions and to reduce the computational load, seeds with the summed coverage of the combined *k*-mers larger than 10,000 bases are discarded.

In the case of point mutations, optimal seeds are $2*k - 1$ bp long and centered on the mutated sites, but in rare cases some seeds do not extend completely, resulting in seeds that are shorter than expected. Multiple closely linked mutations introduce $(k - 1 + s)$ sample-specific *k*-mers, where *s* refers to the length of the mutated region. Long indels also give rise to sample-specific overlapping *k*-mers. In particular, at the breakpoint of deletions, novel sequence is introduced. If this sequence is unique in comparison to the contrasting genome it will generate unique *k*-mers, which will give rise to sample-specific seeds. NIKS identifies these seeds by their partial homology to the breakpoint in the other sample (**Supplementary Fig. 1c–e**). Seeds featuring sequence similarity to multiple other seeds' ends in the other sample are excluded.

Wrongly scored phenotypes or undetected sequencing errors can lead to the presence of *k*-mers, which represent the nonmutagenized allele and can mask real mutations. By default, NIKS discards seeds with support from more than *k k*-mers for the second allele in the other sample. This discards seeds with at least two contradicting reads covering the mutation. As seed pairing requires consistent support from both samples, it introduces high levels of specificity into NIKS.

Mutation-centric assemblies. To extend all valid seeds, NIKS extracts all read pairs sharing at least one of the *k*-mers with a seed. Each such set of read pairs is assembled with Velvet, applying standard parameters, except for the sample specific insert length and automated estimation of the coverage cut-off[23]. Assembled contigs that are longer than the respective seed, but perfectly match the seed, replace the seed.

doi:10.1038/nbt.2515