Highly efficient and comprehensive identification of ethyl methanesulfonate-induced mutations in *Nicotiana tabacum* L. by whole-genome and whole-exome sequencing†

H. Ichida,[*1] H. Udagawa,[*2] T. Takeuchi,[*2] T. Abe,[*1] and Y. Takakura[*2]

Tobacco (*Nicotiana tabacum* L.) is one of the most widely cultivated non-food crops. It is a complex allotetraploid ($2n = 4x = 48$) species with a large genome of 4.5 Gb, which possesses a high repetitive element content.[1] Tobacco carries a pair of duplicated genes (referred to as homeologs) from the S- and T-genomes, and owing to this redundancy, loss-of-function mutations in any single homeolog are typically masked by the correspondent, thereby limiting the use of forward genetic phenotypic screens. Owing to its large genome size, whole-genome sequencing is not readily applicable, as it requires a large amount of sequencing reads and computational power and storage. Whole-exome sequencing (WES) is an approach to primarily sequence on genomic regions that encode proteins. Because the protein-coding sequences only considers approximately 1.3% of the tobacco genome, this technique is particularly effective in organisms with large genomes. In the present study, we developed a set of whole-exon capturing probes for tobacco and tested the set by analyzing 19 independent mutant lines produced by ethyl methanesulfonate (EMS) mutagenesis.[2,3] We also investigated the optimum conditions to detect mutations in tobacco.

We employed 19 individual EMS-treated tobacco lines, designated NtEMS-01–19, and a technical replicate of NtEMS-19, called NtEMS-19-rep2, for the sequencing analyses. A commercial target enrichment platform (SureSelect XT Custom kit, Agilent Technologies) was used to establish a whole-exon enrichment method for tobacco. The custom-designed capturing kit contained 517,835 oligonucleotide probes of 120 bases, and it was expected to capture coding sequences with a size of 50.3 Mb spanning 41,038 genes that had at least one coding sequence (CDS) defined in the target genes. After we produced the whole-exome capturing probes, a new version of the K326 reference genome sequence, Nitab-v4.5, which significantly improves coverage and contiguity, was published.[4] Therefore, we translated the target locations to the Nitab-v4.5 sequences and performed all subsequent analysis based on the new reference genome sequence.

The number of read bases varied from 11.6 to 18.9 Gb per sample, which corresponded to 165.3× to 270.8× of the total CDS length and indicated that, as expected, sufficient sequencing data were obtained. More than 98% of the reads were mapped to the reference sequences, resulting in an average read depth of 93.7× in the CDS regions that commonly existed in the new and previous reference genome assemblies. An average of 97.1% of the target CDS bases in the 19 NtEMS lines were covered by at least 30 reads. Among all the mapped read bases, approximately 75% were located on or were adjacent to the target regions, indicating that the whole-exome capturing was successfully accomplished here. In the variant calling, we used two different programs, GATK and BcfTools. Each program detected 60,884 and 50,260 mutations and resulted in a total of 61,146 non-redundant mutations. Almost all (98.8%) of the detected mutations in the 19 NtEMS lines were single-nucleotide variations, while 95.6% of them were C/G to T/A transitions, which is consistent with the known properties of EMS mutagenesis.The numbers of detected mutations varied from 1987 to 4966 in the 19 NtEMS lines. The average numbers of detected mutations in each line were 2715.7 (1987–3857) and 3511.3 (2415–4966) in 0.6% and 0.8% EMS treatments, respectively. The mutation density was largely proportional to the target CDS density in the genome, indicating that the EMS-induced mutations were sufficiently random, at least practically, despite base preferences resulting from the chemical nature of EMS mutagenesis. A computational experiment using down-sampled sequencing reads demonstrated that approximately 80% of the mutations located in the target regions could be detected with 60× sequencing reads.

At 160× of the total CDS length equivalent, a total of 20,960 out of the 35,667 genes located within the targets harbored at least one mutation in the CDS regions of the 19 NtEMS lines defined by WES. This indicates that an average of 1103 genes could be expected to have mutated within their CDS regions in the mutagenized NtEMS population that was partly analyzed in the present study. Based on the probability formula,[5] one can expect at least one mutation from 95 lines (at 95% confidence level) in the NtEMS library, for all of the 35,667 genes in the target regions.

References
1) N. Sierro *et al.*, Nat. Commun. **5**, 3833 (2014).
2) T. Tajima *et al.*, Ann. Phytopathol. Soc. Jpn. **77**, 258 (2011).
3) Y. Takakura *et al.*, Mol. Plant Pathol. **19**, 2124 (2018).
4) K. D. Edwards *et al.*, BMC Genom. **18**, 448 (2017).
5) L. Clarke, J. Carbon, Cell **9**, 91 (1976).

† Condensed from the article in Front. Plant Sci. **12**, 671598 (2021)

*1 RIKEN Nishina Center
*2 Leaf Tobacco Research Center, Japan Tobacco Inc.