Development of a versatile high-throughput mutagenesis assay with multiplexed short read NGS using DNA-barcoded *supF* shuttle vector library amplified in non-SOS *E. coli*

Hidehiko Kawai[1,2,*], Ren Iwata[2], Shungo Ebi[2], Ryusei Sugihara[2], Shogo Masuda[1], Chiho Fujiwara[2], Shingo Kimura[3], and Hiroyuki Kamiya[1,2,*]

[1] Graduate School of Biomedical and Health Sciences, Hiroshima University, 1-2-3 Kasumi, Minami-ku, Hiroshima 734-8553, Japan
[2] School of Pharmaceutical Sciences, Hiroshima University, 1-2-3 Kasumi, Minami-ku, Hiroshima 734-8553, Japan
[3] Analysis Center of Life Science, Natural Science Center for Basic Research and Development, Hiroshima University, 1-2-3 Kasumi, Minami-ku, Hiroshima 734-8553, Japan

* To whom correspondence should be addressed. Tel: +81-82-257-5301; Fax: +81-82-257-5334; Email: kawaih@hiroshima-u.ac.jp (H. Kawai) or hirokam@hiroshima-u.ac.jp (H. Kamiya).

ABSTRACT

A forward mutagenesis assay using the *supF* gene has been widely employed for the last several decades in studies addressing mutation frequencies and mutation spectra associated with various intrinsic and environmental mutagens. In this study, by using a *supF* shuttle vector and non-SOS-induced *Escherichia coli* with short read Next Generation Sequencing (NGS) technology, we present an advanced method for the study of mutations, which is simple, versatile, and cost-effective. We demonstrate the performance of our newly developed assay via pilot experiments with UV-irradiation, the results from which emerge more relevant than expected. The NGS data obtained from samples of the indicator *E. coli* grown on titer plates provides mutation frequency and spectrum data, and uncovers obscure mutations that cannot be detected by a conventional *supF* assay. Furthermore, a very small amount of NGS data from selection plates reveals the almost full spectrum of mutations in each specimen and offers us a novel insight into the mechanisms of mutagenesis, despite them being considered already well-known. We believe that the method presented here will contribute to future opportunities for research on mutagenesis, DNA repair, and cancer.

INTRODUCTION

In the last half century, various methods such as the Ames test and the micronucleus test have been employed as simple, sensitive and reliable approaches for evaluation of the mutagenicity of chemicals and drugs (1,2). More recently, the technological advancement in science has allowed us, in addition to evaluation of mutagenicity, to address the distinct spectra of mutations produced by a broad range

of intrinsic and environmental mutagens or anticancer drugs (3). The mutational signatures extracted from these mutation spectra have provided valuable information about the role of endogenous and exogenous risk factors in the development of many human disorders and have been crucial in finding potential targets for treatment (4-6).

Until the present day, a vast amount of diverse studies focused on the mechanisms of mutagenesis and DNA repair from the chemical, biochemical, and biological aspects, have heavily relied on experimental strategies using cloned genes (7-10). One such widely used system is the SV40-based shuttle vector mutagenesis assay, which utilizes vector plasmids containing the *supF* gene of *Escherichia coli* and an indicator bacteria strain (11-15). The *supF* gene codes for an amber suppressor transfer RNA (tRNA) that translates the amber stop codon (UAG) to tyrosine (16). Depending upon the position and type of mutations in the *supF* gene, the *supF* tRNA cannot translate selectable marker genes with an amber mutation in its indicator *E. coli* strains, such as KS40/pOF105 (17,18). Because of its properties, the *supF* gene is particularly useful for acquiring mutation spectra and identifying mutation signatures, and provides reliable data on mutations generated by specific DNA lesions on the leading or the lagging template DNA strand via translesion synthesis (TLS) (8). However, one needs to be aware of the limitation of the phenotypic mutation spectrum of the *supF* gene; in our experiments for example, from the entire base sequence of the *supF* gene only about 30% can be detected for dual-antibiotic selection of indicator *E. coli*, i.e. whether these sites can be identified or not is largely dependent on the experimental conditions, including the scale of the experiment, reporter genes for the selection, the mutagens used, the frequency of mutations, and the efficiency of the detection (18-20). Recently, we reported the development of a new indicator *E. coli* strain – RF01, that is able to identify mutation spectra more efficiently than other strains used for the *supF* assay, even though it still does not allow the elucidation of trinucleotide mutational signatures because of the limited resolution of the mutation spectrum analysis (20).These limitations can be overcome by advanced techniques, namely Next Generation Sequencing (NGS), which over the last decade is enabling us to obtain various kinds of information from nucleic acid sequence data, including mutations and three-dimensional organization of the genomic DNA at a single cell level (21-23). On the other hand, the NGS technology also has its limitations, especially for identification and quantification of low-frequency mutations, as it can be difficult to distinguish true mutations from the baseline errors of NGS. In recent years, in order to detect low-frequency mutations, numerous library construction methods and sequencing methods have been developed, such as maximum depth sequencing or duplex sequencing (24-26). For every scientific objective, trade-offs between accuracy, cost, and speed of the multiple available approaches must be carefully considered. Importantly, there is a scarcity of methods whose technical simplicity and material availability enables any researchers, not only those specialized in NGS, to accumulate, reproduce and compare data across research fields. Certainly, the availability of an easy, low-cost assay for accurate analysis of mutational signatures would benefit many laboratories who are deterred by the complexity of the existing methods.

In this study, we attempt to develop a simple mutagenesis assay using a random 12-nucleotide DNA barcoded *supF* shuttle vector library with non-SOS-induced *E. coli* which can amplify the DNA with "a high-fidelity replication mechanism" as a template for the NGS sample preparation (27). This promotes cost-benefit, high-throughput, and high accuracy with almost no sequencing errors. In the first part, we present data from the multiplexed NGS assay utilizing the barcoded *supF* shuttle vector library and the BGI-platform that uses the rolling circle replication method, and we compare the results for different conditions of sample preparation. In the later part of this study, we present analysis of ultraviolet (UV) radiation-induced mutations in direct *E. coli* experiments or after transfection into mammalian cells, as an example of the performance of our newly developed assay. UV-radiation is a known environmental carcinogen and causes a characteristic mutational pattern via the formation of pyrimidine-pyrimidine dimers (28-30). The C to T transition at C<u>C</u>N and T<u>C</u>N trinucleotide sites (the mutated base is underlined) is a mutational signature highly specific to UV-mediated mutagenesis (mutational COSMIC signatures SBS7a-d) (31). The cytosine in UV-induced pyrimidine-pyrimidine dimers undergoes deamination to uracil, followed by insertion of an A nucleotide opposite the lesion by error-prone TLS polymerases (32-34). The substantial amount of sequence data on UV-induced mutations in mammalian cells obtained in this study allowed us to uncover some insightful information about UV-induced mutagenesis, underscoring the importance of a systematic process for sufficient data collection and analysis. We believe that our new approach using the barcoded *supF* shuttle vector library and NGS technology presents a benefit to future research on mutagenesis, DNA repair, and cancer.

MATERIALS AND METHODS

Plasmids and plasmid constructions

The eight series of *supF* shuttle vector plasmids – pNGS2-K1, -K2, -K3, -K4, -A1, -A2, -A3, -A4, shown in Supplementary Figure S1, were constructed from a pSB189L *supF* shuttle vector (11,12,15) using NEBuilder HiFi DNA Assembly Master Mix (New England Biolabs, USA). Briefly, DNA fragments containing the amber suppressor tRNA gene (*supF*), the TP53-/Rb- binding-deficient mutant SV40 large T antigen (E107K/D402E) with the SV40 replication origin, the pBR327 origin of replication, the M13 intergenic region, or the ampicillin-resistance gene (amp^r) were amplified with the overlap sequences from pSB189L by PCR using high-fidelity KOD One PCR master Mix (TOYOBO, Japan). DNA fragments containing the kanamycin-resistance gene (km^r) were amplified from a MCS1-EF1α-Blasticidin-pA-MCS2-PGK-hsvTK HR targeting vector (System Biosciences, USA). The overlapping oligonucleotide primers were designed by NEBuilder Assembly Tool (New England Biolabs) and were synthesized commercially and purified on a reversed-phase column by FASMAC (Japan). The amplified DNA fragments were assembled to construct the plasmid DNAs by following the manufacturer's instructions (NEBuilder HiFi DNA Assembly Master Mix) and were transformed into *E. coli* HST08 competent cells (TaKaRa Bio Inc., Japan) for cloning. The sequences of the series of pNGS2 plasmid DNAs were confirmed by Sanger sequencing. The plasmid DNAs were then

transformed into *E. coli* JM109 competent cells (TaKaRa Bio Inc.) for amplification, after which they were extracted using either GenElute Plasmid Miniprep or Midiprep plasmid DNA purification kit (Merck, Germany).

Construction of the N_{12} barcoded *supF* shuttle vector library

The pNGS2 plasmids (-A1~A4 or K1~K4) were digested at a unique site by incubation with *Eco*RV-HF restriction enzyme (New England Biolabs) at 37°C for 1 hour. The linearized pNGS2 was purified by agarose gel electrophoresis, followed by extraction using FastGene Gel/PCR Extraction kit (NIPPON Genetics, Japan). A quantity of 12 fmol purified *Eco*RV-digested pNGS2 was combined with 60 fmol of single-stranded oligonucleotides in a volume of 5 μL. The oligonucleotides containing randomized N_{12} (12-nucleotide sequences, 5'-dGGCCTCAGCGAATTGCAAGCTTCTAGAAGGCGATNNNNNNNNNNNNATCGAATTCGGATCCTT TCTCAACGTAA-3', where N = A, T, G or C) were synthesized and purified on a reversed-phase column by FASMAC (Japan). The assembly reaction was performed using NEBuilder HiFi DNA Assembly Master Mix (New England Biolabs) following the manufacturer's instructions. The *E. coli* JM109 electro-competent cells were transformed with 1-2 μL of the assembly reaction mixture by electroporation. Following a 1-hour recovery in SOC medium at 37°C with agitation, the transformants were serially diluted, seeded on LB agar plates containing 25 μg/mL kanamycin, and incubated at 37°C overnight. The grown colonies were counted and harvested by scraping. Approximately 10^2, 10^3, and 10^4 colonies were harvested for each desired complexity of N_{12}-BC sequences (pNGS2-K1-N_{12}-BC library: Lib10^2_#1, Lib10^2_#2, Lib10^3, and Lib10^4, pNGS2-K3-N_{12}-BC library: Lib10^4_#1 and Lib10^4_#2), and pNGS2-N_{12}-BC libraries were extracted from ~1.5 g of harvested *E. coli* JM109 cell pellets by using NucleoBond Xtra Midi (TaKaRa Bio Inc.) and protocol for low copy plasmids. Each prepared library was interrogated for the presence of potential *supF* mutants by using the indicator *E. coli* strain RF01, and a mutation frequency of less than 10^{-4} was considered mutant free.

Preparation of PCR templates for multiplexed NGS libraries

Each pNGS2-N_{12}-BC library was brought to a concentration of 100 ng/μL in sterile distilled water and exposed to UV-C radiation from germicidal lamps at a dose rate of 1.0 $J/m^2 \cdot s$ for 0, 50, 100, and 200 seconds. For experiments in non-SOS induced *E. coli*, immediately after irradiation the N_{12}-BC library (5, 25, 50, and 100 ng for 0, 50, 100, and 200 J/m^2, respectively) was electroporated into the indicator *E. coli* strain RF01 and seeded on either titer LB agar plates containing 25 μg/mL kanamycin and 30 μg/mL chloramphenicol, or selection plates containing 25 μg/mL kanamycin, 30 μg/mL chloramphenicol, 50 μg/mL nalidixic acid, and 100 μg/mL streptomycin. For experiments in mammalian cells, the UV-irradiated pNGS2-N_{12}-BC library was transfected into the human osteosarcoma U2OS cell line (HTB-96™, American Type Culture Collection, USA), with Lipofectamine

2000 (Thermo Fisher Scientific, USA), as we have previously described (15). Briefly, U2OS cells were seeded into 6-well plates at 3×10^5 cells/well the day before transfection, and transfection was carried out with 400 ng of pNGS2-N_{12}-BC library per well following the manufacturer's instructions. After 48 h of culture, the transfected pNGS2-N_{12}-BC library was extracted from the cells (35). To digest unreplicated plasmids in the library, the library was treated with the restriction enzyme *Dpn* I which possesses specificity for methylated 5'-G^{m}ATC-3' sequences. The library was then transformed into RF01 following the already described transformation procedure. For both experiments – in *E. coli* and in mammalian cells, the *supF* mutant frequencies were estimated based on the number of colonies grown on the individual plates. The presumed number of required colonies were harvested as a cell pellet by scraping, and libraries were extracted using GenElute plasmid miniprep kit (Merck).

Preparation of samples with unique indexes for multiplexed NGS

Each sample with a unique index for multiplexed NGS was prepared by PCR using KOD One PCR Master Mix (TOYOBO). For the standard protocol, 10 ng of the library (or from 1 to 100 ng for experiments presented in Figures 2, 3, and 4) extracted from *E. coli* as described above, was used for a 50 μL PCR reaction (30 cycles of amplification at 98°C for 10 sec, 60°C for 5 sec, and 68°C for 1 sec) with a series of primer sets (Supplementary Figure S2) containing a pre-designed 6-nucleotide sequence (N_6) as an index sequence for distinguishing each sample by sequence data analysis. PCR products were purified using NucleoSpin Gel and PCR Clean-up (TaKara Bio Inc.), and then analyzed and quantified by Nanodrop spectrometer (Thermo Fisher Scientific) and agarose gel electrophoresis. The uniquely indexed samples were combined together in the proper proportion according to the expected number of NGS reads for each sample. More than one hundred reads per N_{12}-BC sequence were obtained for all samples in this study (approximately 400 Mb and 16 Mb of sequencing data for the titer plates and the selection plates, respectively).

DNBSeq sequencing and sequencing processing

Preparation of NGS libraries and sequencing were conducted by Seibutsu Giken inc. (Japan). The libraries were subjected to multiplexed deep sequencing in 200-bp paired-end mode on the NGS using a BGISEQ-G400 platform (MGI Tech, China). The concentration of the NGS samples was measured using either Synergy LX (Agilent Technologies, USA) and QuantiFluor dsDNA System (Promega, USA), or Qubit 3.0 Fluorometer (Thermo Fisher Scientific) and dsDNA HS Assay Kit (Thermo Fisher Scientific). The libraries for DNBSeq were prepared using either MGIEasy FS DNA Library Prep Set (MGI Tech) (25 ng of sample, 8 cycles) or MGIEasy PCR-Free DNA Library Prep Set (MGI Tech) (200 ng of sample, enzymatic digestion for 4 min). The quality of the prepared libraries was assessed using an Agilent 2100 bioanalyzer and a High Sensitivity DNA kit (Agilent

Technologies), or Fragment Analyzer and dsDNA 915 Reagent Kit (Agilent Technologies). The DNA nanoball was generated from the libraries using the MGIEasy Circularization Kit (MGI Tech).

NGS data analysis

The FASTQ files generated from each N_6-index sample (sorting was performed by Seibutsu Giken inc.) were used for the following analysis: 1) The reverse-complement sequences were obtained from FASTQ data using SeqKit tools (36); 2) The paired-end reads from FASTQ data including the reverse-complement sequences were merged with a minimum 30-bp overlap region using CASPER (37); 3) The merged sequences were mapped to a reference sequence, and only aligned sequences were converted to BAM files by using minimap2 (38); 4) BAM files were converted to FASTQ files using SAMtools (39,40); 5) The extraction of the most likely true N_{12}-BC sequences with error-corrections was performed based on a whitelist by UMI-tools (41). Two 6-nucleotide sequences, GGCGAT and ATCGAA, were used for the N_{12}-BC extraction, with only one-base differences between N_{12}-BC sequences being allowed, and corrections were implemented using the following regular expression: (".*(GGCGAT){s<=0}(?P<cell_1>.{12})(?P<umi_1>.{0})(ATCGAA){s<=0}.*"); 6) Using VarDict, the sequence data for all N_{12}-BC sequences with more than 9 reads per N_{12}-BC was employed for the alignment to the reference sequence, variant calling and calculation of allele frequencies (42); 7) Spotfire version 7.11.1 (TIBCO, USA) was used for data organization, filtering and visualization.

RESULTS

Construction of barcoded *supF* shuttle vector libraries for NGS assay

Aiming to develop a more advanced and versatile assay system suitable for a wide range of mutagenesis studies, we attempted to apply a short-read NGS platform with multiplex sample preparation to the conventional *supF* mutagenesis assay. At first, we constructed a series of *supF* shuttle vector plasmids, named pNGS2-K1~4 or -A1~4 (Supplementary Figure S1). Each plasmid carries all regions depicted for pNGS-K1 or A1 in Figure 1A, including an antibiotic resistance gene for either kanamycin (km^r) or ampicillin (amp^r). Different orientations of the *supF* gene and the M13 intergenic region were utilized for the purposes of individual experiments, such as the detection of specific types and sites of DNA-damage (15,43,44). To investigate both mutation frequencies and profiles as precisely as possible, *supF* shuttle vector libraries containing random 12-nucleotide (N_{12}) barcode sequences (N_{12}-BC *supF* library) were generated from pNGS2 to be used for NGS analysis. An easy and reliable method to construct N_{12}-BC *supF* libraries with sufficient complexity by using single-stranded DNA was developed in this study, as was described in detail in the MATERIALS AND METHODS section. To prepare the samples for NGS analysis, a region from the N_{12}-BC *supF* library (Figure 1B), was amplified by PCR by using the indicated primer sets (Supplementary Figure S2). We

opted to use PCR-amplification instead of acoustic shearing by ultrasound, in order to avoid inducing oxidative damage to the DNA (45,46). Also, each primer pair for PCR-amplification contains in one of the primers a pre-designed 6-nucleotide index sequence (N_6) for multiplexed NGS assay, which is used for classification of independent samples in every NGS-run. The independently amplified PCR products for each sample were purified and combined in the proper proportion, after which they were subjected to multiplexed deep sequencing in 200-bp paired-end mode on the NGS using a BGISEQ platform (45). We have also carried out several pilot experiments using 150-bp paired-end mode and/or Illumina platform, which confirmed that these modalities can also be successfully employed for the assay. The procedure for the *supF* mutagenesis NGS assay in this study is illustrated in Figure 1C. Depending on the purpose and the design of each experiment, the N_{12}-BC *supF* library or cells can be treated with mutagens, and any or all of the libraries (i) ~ (iv) in Figure 1C can be subjected to NGS analysis.

Analysis for potential bias during NGS-sample preparation

To address the issue of detection-bias, where the number of N_{12} barcode (N_{12}-BC) sequences or variant sequences detected by NGS may be affected by the number of N_{12}-BC sequences per library in different experimental conditions, N_{12}-BC *supF* libraries with different numbers of N_{12}-BC sequences (approximately 1×10^2 in Lib10^2_#1; 1×10^2 in Lib10^2_#2; 1×10^3 in Lib10^3, and 1×10^4 in Lib10^4) were prepared from pNGS2-K1 and analyzed. The background *supF* mutant frequency of each library was first checked by the conventional *supF* assay, and was confirmed to be ~10^{-5}, indicating that there was no *supF* phenotypic mutation in the libraries. After that, different amounts of PCR template (1.0, 3.3, 10, 33, or 100 ng of the N_{12}-BC *supF* library), were used for PCR amplification and analyzed by NGS, as described in the MATERIALS AND METHODS section. The PCR products were purified and mixed in the optimal ratio so that NGS analysis yields approximately 200-250 reads per barcode sequence. The N_{12}-BC sequences for each sample were error-corrected based on a whitelist (41), and classified by the identical N_{12}-BC sequence. The number of unique N_{12}-BC sequences in each library is shown in Figure 2A (left-side graph in each pair of bar plots). The number of identified unique N_{12}-BC sequences was 117-119 for Lib10^2_#1, 97 for Lib10^2_#2, 1421-1426 for Lib10^3, and 13125-13172 for Lib10^4, very close to the estimation based on the number of colonies used to prepare each library. The number of reads was also within the expected range, and comparable at different amounts of PCR template (Figure 2A, right-side graph in each pair of bar plots). This data suggests that most N_{12}-BC sequences could be detected and classified, and that there was almost no overlapping of N_{12}-BC sequences in *E. coli* colonies used for library preparation. The actual detected number of reads for each identical N_{12}-BC sequence was compared in stacked bar graphs in Figure 2B. There was no correlation between the amount of PCR template and the number of reads for each identical N_{12}-BC sequence, confirming that an appropriate PCR procedure can evenly amplify the reads from the plasmid DNA template in the libraries, independently from the template amount.

Variant calls from aligning reads to the reference sequence

The variant sequences to a reference sequence were detected by VarDict, and the threshold for allele frequency was set to 0.1 (42). In this study, the allele frequency for each classified N_{12}-BC sequence was referred to as variant frequency (VF). The VF has to be considered and interpreted depending on the experimental conditions and purposes. The read counts for individual N_{12}-BC sequences with variants were graphed on a scatter plot as a function of VFs (Figure S3A). The VFs appeared to be distributed in a bipolar manner, but no correlation was observed between the number of reads and VFs, suggesting that a low number of reads cannot be the cause for lower VFs. For all samples, about 20-40% of N_{12}-BC sequences contained reads with variants (VFs equal to or greater than 0.1). In Figure S3B, the distribution of the number of N_{12}-BC sequences with a variant by VF is presented as bar histograms. Based on this data we decided in the experiments hereafter to divide the N_{12}-BC sequences with variants into two sub-groups along the line of 0.4 of the VF.

Variant frequencies and baseline sequencing errors

In Figure 3A-C, applying a threshold VF of 0.4, the number of N_{12}-BC sequences with variants above and below the threshold were plotted in separate graphs (Figure 3A and 3C, respectively) with respect to their position in the sequencing region shown in Figure 1B. It becomes obvious that almost all N_{12}-BC sequences with variant reads and VFs 0.4-1.0 were located in the region between -80 and -20. This region is the cloning region for the random 12-nucleotide barcoding and does not manifest as *supF* mutant phenotype by indicator *E. coli* in the conventional *supF* assay. These mutations most likely originate from the oligonucleotides used for the N_{12}-BC sequence insertion during preparation of the N_{12}-BC *supF* library. The types of variant sequences in this region are indicated by different colors in Figure 4B. Notably, the proportions of variant calls are practically identical at different template amounts for the same N_{12}-BC *supF* library, which means that the pattern of variant types can provide a fingerprint property for each N_{12}-BC *supF* library and a practical index for the detection of mutations.

On the other hand, the N_{12}-BC sequences with variant reads in the range of VF 0.1-0.4 were almost exclusively located at either or both positions 55 (a SNS from A to C) and 69 (from T to C) of the *supF* gene, as shown in Figures 3C and 3D. Also, almost all N_{12}-BC sequences with variants at these positions had VF < 0.4 (Figure 3D). Therefore, variant sequences at these positions were assumed to be sequencing errors and not true mutations. The percentage of reads with 0.1-0.4 VFs for each position is shown in Figure 4E (the value is the median from all 20 libraries). Positions 55 and 69 contained variant calls in 24.7% and 42.6% of reads, respectively. Taken together, this data suggests that as long as the number of harvested colonies is less than the number of identified N_{12}-BC sequences, VF around 0.4 can be used as the threshold value between baseline sequencing errors and true mutations.

In addition, almost identical results to those shown in Figures 2 and 3 were obtained from a series of N_{12}-BC libraries originating from pNGS2-A1 containing ampicillin-resistance instead of the kanamycin-resistance gene (raw data is available in DDBJ database).

Performance of the NGS mutagenesis assay using N_{12}-BC *supF* libraries

Next, two independently prepared N_{12}-BC 10^4 libraries from pNGS-K3, Lib10^4_#1 (#1) and Lib10^4_#2 (#2), were used to evaluate the performance of the newly developed mutagenesis assay. To produce different numbers of characteristic DNA lesions, the libraries were irradiated with UV-C at 0, 50, 100, 200 J/m^2, and then directly transformed into the indicator *E. coli* strain RF01 which is derived from *E. coli* KS40 under non-SOS-induced conditions (27). The RF01 transformed with irradiated or non-irradiated libraries were diluted and plated on LB plates with or without antibiotics (selection or titer plates, respectively), as for a conventional *supF* assay. As expected, the number of RF01 colonies on the titer plates was obviously decreased by UV-irradiation in a dose-dependent manner (the average number of colonies per ng of the N_{12}-BC 10^4 library, from three experiments, were 5.0×10^6 for 0 J/m^2, 3.6×10^6 for 50 J/m^2, 1.7×10^6 for 100 J/m^2, and 3.5×10^4 for 200 J/m^2), and on the selection plates containing nalidixic acid and streptomycin the mutant frequencies were in the order of 10^{-4} to 10^{-5}. Approximately ten thousand colonies of RF01 transformed with N_{12}-BC *supF* libraries were harvested from one to three titer plates, libraries were extracted, and approximately 400 Mb per sample were analyzed by NGS. The number of N_{12}-BC sequences shown in Figure 4A were identified as unique N_{12}-BC sequences, indicating that even though some N_{12}-BC sequences were overlapping between colonies, the majority were not. The number of N_{12}-BC sequences with a variant and variant frequency 0.4-1.0, as well as their distribution in the analyzed region (positions -67~214) are shown in Figure 4B and Supplementary Figure S4, respectively. The mutations seem to be evenly distributed, except for the region around the 12-nucleotide barcoding site. This data suggests that all of the variants in non-irradiated samples were derived from the barcoding process with the custom-synthesized N_{12} oligonucleotides during the N_{12}-BC library construction, and the number of N_{12}-BC sequences with a variant in the *supF* gene was slightly increased by UV-irradiation in a dose-dependent fashion. When the N_{12}-BC sequences with a variant were classified according to nucleotide position along the line of position -20, near the start site of the *supF* gene (position -15 is the first nucleotide of the promoter), it is evident that there were no variant calls in the region of the *supF* gene in the non-irradiated library (Figures 4C and 4D). The major type of variant sequences identified by variant calls in the irradiated libraries were single nucleotide substitution (SNS) (Figures 4C and 4E). Also, more than 80% of the SNSs consisted of G:C to A:T transitions (largest proportion) and G:C to C:G transversions (Figure 4F). The mutation frequency for the *supF* gene was calculated as the ratio between the number of variant sequences and the total number of analyzed N_{12}-BC sequences (Figure 4G). For comparison, the *supF* mutant frequency was also determined by the conventional *supF* assay, as the number of colonies on selection plates versus that on titer plates (Figure 4H). The results from the two methods exhibit very similar trends of mutation frequency

increasing in a UV-dose-dependent fashion, yet the *supF* mutation frequencies calculated from the NGS data were about twenty times higher than those determined by the conventional *supF* assay. The possible reasons for this discrepancy are the following: 1) a *supF* gene containing a UV-damaged non-transcribed strand and a non-damaged transcribed strand confers a nalidixic acid and streptomycin-sensitive phenotype in individual colonies of RF01 due to the wild-type *supF* tRNA expression; 2) there is a selection bias due to “silent mutations” or “phenotypic differences” in the mutant phenotype of the *supF* gene, depending on either nucleotide positions or the substituted bases (for instance, positions 21-60 of the *supF* gene correspond to the pre-tRNA sequence); 3) a plasmid may contain several mutations at different positions.

In order to investigate in more detail the SNS mutation spectrum, all possible 192 trinucleotide contexts of mutations from the titer plates were compared between the complementary DNA strands. The combined data from libraries Lib10^4_#1 and Lib10^4_#2 is shown in Figure 5, top and bottom plot for the non-transcribed and transcribed DNA strand, respectively. The most common of the SNSs was a change at the trinucleotide sequence T<u>C</u>N to either T<u>T</u>N or T<u>G</u>N. In addition, we subjected samples of UV-irradiated N_{12}-BC libraries to direct PCR amplification without transformation into *E. coli* and analyzed them. In this case, no dose-dependent increase of the variant sequences was observed (raw data is available in DDBJ database), and the data from all irradiated samples were very similar to the data shown in Figures 3 and 4. The G:C to A:T transitions were very rarely detected, indicating that the high fidelity DNA polymerase with no TLS activity used in the PCR process cannot replicate DNA and produce mutations in the presence of lesions generated by UV-irradiation, such as cyclobutane pyrimidine dimers (CPDs) or pyrimidine(6–4)pyrimidone photoproducts (6–4PPs).

Mutation spectra in the UV-irradiated libraries from selection plates

In the case of the conventional *supF* assay, to analyze the sequences of the *supF* gene in the mutants, a number of single colonies on the selection plate were picked up and analyzed by Sanger sequencing. Next, to analyze the sequences of the *supF* gene in the mutants using the newly developed NGS assay, the approximate number of colonies on the selection plate were determined, and the colonies were harvested together. The results are presented in Figure 6A. Only a low number of colonies were grown on the selection plates in the non-irradiated condition. The libraries were purified from the collected pellet of *E. coli* and analyzed as described above for the titer plates (16 Mb per sample were analyzed). Again, the number of detected N_{12}-BC sequences were almost equal to the number of harvested colonies, and a limited number of N_{12}-BC sequences (~2%) contained more than one variant sequence (Figure 6A). The positions of the variant sequences are shown in Figure 6B and Supplementary Figure S5, where the colors indicate the variant types or the variant frequencies, respectively. As expected, in contrast to the titer plates (Figure 4C), here the mutation pattern manifested more specific peaks at certain positions, which could be associated with both the *supF* mutation phenotype and the nucleotide signature. Although the mutation/mutant frequency was increased by UV irradiation in a dose-dependent manner (Figure 4G and 4H), the proportion of

mutation types (Figure 6C) and the sequences of base substitutions (Figure 6D) remained very similar at different irradiation doses, as were also the di-nucleotide substitutions (DNSs) (Figure 6E). The mutation types and sequences were comparable to those found in the samples from titer plates shown in Figure 4. However, it is noteworthy that the distributions of the tri-nucleotide sequences were significantly different between the non-transcribed and transcribed strand of the *supF* gene, contrary to the observations from the titer plates (Figure 7 versus Figure 5). Nevertheless, mutations from both titer and selection plates were almost exclusively C:G to T:A transitions at CCN:NGG to CTN:NCG and TCN:NGA to TTN:NAA.

UV irradiation mutagenesis in mammalian cells

Finally, to evaluate the performance of the newly developed mutagenesis assay in mammalian cells, the UV-irradiated N_{12}-BC 10^4 libraries (Lib10^4_#1 and Lib10^4_#2) were transfected into human osteosarcoma U2OS cells. After 48 hours, the libraries were extracted from the cells, and treated with dam-G^{m}ATC-methylated DNA specific restriction enzyme *Dpn* I to digest un-replicated DNAs that contain UV-photoproducts. The purified libraries were then transfected into the indicator *E. coli* strain RF01, and samples for NGS analysis were prepared from titer and from selection plates, same as described above for the direct RF01 experiments. The data is presented in Figures 8 and 9 analogously to Figures 4 and 5. The trends found from experiments in mammalian cells were comparable to these in *E. coli*, although the data was significantly different in terms of the number of mutations. First, we analyzed the samples from N_{12}-BC *supF* libraries transfected into U2OS cells and amplified in the indicator *E. coli* on titer plates. Approximately, from 4,000 to 7,000 N_{12}-BC sequences were identified per sample from around 10,000 colonies transformed with libraries from U2OS cells (Figure 8A). A significant dose-dependent increase was observed in the number of N_{12}-BC sequences with variant calls, and higher number of variant sequences were obtained from the same number of colonies compared to directly transformed with the libraries *E. coli* (Figure 8B vs 4B). These differences most likely reflect mutagenic events in U2OS cells arising from UV-induced DNA lesions, mainly from their deamination products. The nucleotide positions and types of variant calls are presented in Figure 8C and Supplementary Figure S6. It is evident that only a very low number of variant sequences were detected in non-irradiated samples. The detected mutations were distributed throughout the analyzed region, including around the barcoding sites (Figures 8C and 8D). As shown in Figure 8E, the major type of variant sequences was SNSs, which is the expected UV-induced mutation signature, and this observation was the same as in directly *E. coli* experiments. However, in the U2OS cells experiment, the percentages of both di-nucleotide substitutions and insertions at deletion sites (Del-ins) were increased in a dose-dependent manner. As will be described later, most of the variant sequences detected as Del-in by VarDict were triplet mutations, which are known UV-specific mutations and are defined as multiple base substitutions within a three nucleotide sequence containing at least one dipyrimidine. In contrast, the proportions of SNSs were not significantly changed at different irradiation doses and the majority were G:C to A:T transitions (Figure 8F). Same

as in the direct *E. coli* experiments, the mutation frequency for the *supF* gene was calculated using the ratio between the number of variant sequences and the total number of analyzed N_{12}-BC sequences (Figure 8G). The *supF* mutant frequencies were also determined by the conventional assay, as the number of colonies on the selection plates versus that on the titer plates (Figure 8H). Again, the results exhibit a very similar trend of a dose-dependent increase, and predictably, the *supF* mutation frequencies calculated from NGS data were about fifteen times higher than those determined by the conventional assay. As in the direct *E. coli* experiments (Figure 5), the SNS mutation spectra were compared between the complementary DNA strands (Figure 9). The most common of the SNSs were G:C to A:T transitions, mainly found at TCN:NGA to TTN:NAA.

Next, the samples obtained from N_{12}-BC *supF* libraries transfected into U2OS cells and amplified in the indicator *E. coli* on selection plates were analyzed. Contrary to the direct *E. coli* experiments (Figure 6A), approximately one thousand colonies were yielded for each sample from selection plates, even in the non-irradiated condition. As shown in Figure 10A, in spite of the fact that the samples for all UV doses were prepared from almost the same number of colonies (one thousand), the identified number of N_{12}-BC sequences was less than one thousand and showed an increase in a dose-dependent manner, approaching 1000 only at 200 J/m^2. This result indicates that the classified number of N_{12}-BC sequences with variant sequence(s) were already saturated for the samples irradiated at doses of 0, 50, and 100 J/m^2. Even if more colonies were obtained from a larger number of plates, the same overlapping N_{12}-BC sequences with variant(s) would be identified. In other words, almost all variant sequences were already detected at doses lower than 200 J/m^2, as long as the indicator *E. coli* transformed with the libraries could form a colony and the sample could be amplified by the PCR procedure. Moreover, the dose-dependency of the discrepancy between the numbers of N_{12}-BC sequences and the number of variant sequences indicates that multiple mutations in the same read were produced. The number of mutations per N_{12}-BC sequence is shown by a horizontal stacked bar graph in Figure 10B, and it can be seen that the proportion of N_{12}-BC sequences with multiple mutations was slightly increased in a dose-dependent manner. The positions and types of mutations plotted in Figure 10C and Supplementary Figure S7 reveal that the overall results were very similar to those from *E. coli* directly transformed with N_{12}-BC *supF* libraries (Figure 6B), the reasons for which are associated with the properties of the *supF* gene and UV-induced DNA lesions. Regarding the proportion of mutation types and SNSs, the data from selection plates (Figures 10D and 10E) exhibited a trend similar to the data from titer plates (Figures 8E and 8F). In contrast, as shown in Figure 11, as in the case of *E. coli* directly transformed with libraries and grown on selection plates, the tri-nucleotide mutation patterns obviously differed between the non-transcribed and transcribed strand of the *supF* gene. Although the tri-nucleotide mutation patterns were observed to be biased between the strands, this is assumingly due to the phenotypic variability among the mutation types and mutation sites in the *supF* gene. The combined data of the tri-nucleotide substitution patterns from both strands becomes similar to that from the titer plates. Remarkably, the substantial amount of data on mutation spectra that we obtained had led to a better appreciation of the fact that UV-induced DNA-lesions had resulted in mutation clusters with a significant strand bias. As mentioned above, the

rates of N_{12}-BC sequences with multiple mutations were increased in a UV dose-dependent fashion (Figure 10A and 10B), therefore we next focused on the correlation between the numbers, types, positions, and mutated bases for each mutation. At first, the N_{12}-BC sequences were classified as containing a single mutation versus two or more mutations, and N_{12}-BC sequences in each category were classified by the type of mutations they contained (Figure 12A). The total number of N_{12}-BC sequences with mutations is indicated in brackets. If a N_{12}-BC sequence contained one mutation, the proportion of mutations classified as Del-ins or SNSs from cytosine sites were increased in a dose-dependent manner. On the other hand, if a N_{12}-BC sequence contained more than one mutation, the mutations classified as DNSs and SNSs from guanine sites were increased. Then, the N_{12}-BC sequences were classified as containing1 to 4~ mutations per sequence, and depicted in Supplementary Figure S8, where the mutations are shown according to their nucleotide positions. It can be seen that if a N_{12}-BC sequence contained one mutation, especially in the case of SNSs, the mutations were detected at relatively specific sites that can be considered as sites responsible for *supF* mutation phenotype. On the other hand, if a N_{12}-BC sequence contained more than one mutation, the mutations were relatively sparsely distributed. In both cases, SNSs were predominantly substituted from either C or G. This result may suggest that the DNA lesions produced by UV-irradiation at cytosine sites on the same strand of the DNA molecule produce multiple mutations simultaneously at the first replication cycle.

Next, the SNSs were classified as either one or multiple SNS(s) per N_{12}-BC sequence and shown for the non-transcribed strand of the *supF* gene with all 12 types of base substitutions (Figure 12B). In the case of single SNSs (Figure 12B, graph on top), G to A was the most frequent and C to T was the second most frequent substitution, which may be associated with the number of G bases (18 sites) and C bases (14 sites) in the strand of the *supF* gene responsible for the selection phenotype (refer to Figure 14A). On the other hand, in the case of multiple SNSs per N_{12}-BC sequence (12B, bottom graph), C to T was the most frequent and G to A was the second most frequent, which may be associated with the number of G bases (56 sites) and C bases (70 sites), GA (13 sites) or TC (18 sites) in the non-transcribed strand of the *supF* gene (refer to Figures 14A and 14B). The number of N_{12}-BC sequences with other types of single mutations are shown in Figure 12C, where a clear UV-dose-dependent increase can be seen in the number of 1-nt deletions, DNSs, and Del-ins, but not in the number of large deletions and insertions. Figures 12D and 12E represent the most frequent mutation sequences for DNS and Del-in mutations, respectively. The DNSs were gradually increased in a dose-dependent manner and seem to be saturated at a dose of 200 J/m^2; also, they seem to mainly result from GG:CC to AA:TT transition in at least tetra-nucleotide repeats of C or T at several distinct positions in the gene (refer to Figure 14C). On the other hand, in the case of Del-ins, the mutations increased at higher doses and were the result of GAG:CTC to AAA:TTT transitions, also at distinct positions (refer to Figure 14D). These triplet mutations may be generated from UV-photolesions at TC dipyrimidine sites by error prone TLS polymerases such as polζ and polκ. In addition, in the non-irradiated samples the Del-in mutations were peculiarly different from those in the irradiated samples and were detected as TTGAT to ATCAA. This TTGAT penta-nucleotide sequence

located in the potential hairpin loop of a quasi-palindromic sequence, GGCGCGTCATTTGATATGATGCGCC, may lead to template switching and give rise to complementary cluster mutations (refer to Figure 14E).

Next, we focused on the multiple mutations (more than one mutation in the same N_{12}-BC sequence) in terms of their combinations and mutation spectra. The multiple mutations were classified by their combinations and depicted for each UV-dose in Supplementary Figure S9. The comparison between the spectra of the single and multiple mutations for the sample irradiated with 200 J/m^2 was summarized in Figure 13 (single mutation per N_{12}-BC – Figure 13, bottom part; more than one mutation per N_{12}-BC – Figure 13, top and center-left part) and Supplementary Figure S10. It is obvious that there is a consensus for multiple mutations in a single N_{12}-BC sequence as follows: 1) In the non-irradiated samples the mutations were by a large majority SNSs, all of them either from C or G (Supplementary Figure S9); 2) In the UV-irradiated samples all combinations of more than two mutations in a single N_{12}-BC sequence were increased in a dose-dependent manner, except for those composed only of substitutions from A and T; 3) As for the Del-in and DNS mutations, if they contained C to T substitutions, the other mutation(s) were mainly C to T SNSs, and if they contained G to A substitutions, the other mutations were mainly G to A SNSs (Supplementary Figures S10); 4) The number of N_{12}-BC sequences with multiple mutations including a mutation from C was higher than those including a mutation from G, same as the trend for multiple SNS mutations shown in Figure 12B. 5) The distance between two SNSs in multiple mutations induced by UV irradiation was relatively shorter than the theoretically expected based on the sequence (Supplementary Figures S11 and S12).

DISCUSSION

Ever since the 1980s, the *supF* forward mutagenesis assay which utilizes a *supF* gene encoding shuttle vector and an indicator *E. coli* has been widely employed in studies on mutagenicity in *E. coli* or mammalian cells (14,16,19,48,49). This conventional *supF* assay is very useful for analysis of both mutation frequencies and mutation spectra. The mutation spectra of single or tandem base-substitutions for inactive *supF* genes identified by using the blue-white colony color assays were comprehensively summarized in an earlier review article, and it was noted that the mutations were detected at 86 sites within a 158-bp region covering the *supF* gene (54%) and at 74 sites within the 85-bp mature tRNA region (87%), thus demonstrating the great sensitivity of the *supF* assay system for analysis of mutation spectra (19). However, obtaining reliable datasets by the conventional *supF* assay requires skill and experience, especially for studies where the mutations of interest are induced with low frequency. The method has been advanced by the construction of indicator bacterial strains with different *supF* reporter genes which allow selection based on survival of bacteria containing mutant *supF* genes. However, the fact that the *supF* phenotypic selection process relies on the structure and function of transfer RNAs that may be differently affected by different mutations means that the improvement of the efficiency of the selection process may cause loss of coverage of the

mutation spectra, as it is under our experimental conditions, where the coverage is about 30% (19,20). Since in recent years NGS has become a routine research tool due to its lowering cost, numerous sophisticated applications of NGS technologies are being developed for research in genomics, epigenomics, and transcriptomics (50,51). In this study, we attempted to combine the *supF* shuttle vector forward mutagenesis assay with sequencing analysis by a short-read NGS platform and consequently developed a novel *supF* NGS assay with potential application in mutagenesis and DNA repair studies. The combination of the DNA-barcoded *supF* shuttle vector library and its amplification via *in vivo* replication in the indicator non-SOS *E. coli* provides us with proper templates for the PCR-facilitated preparation of multiplexed NGS samples. Thus, during the PCR process the amplification and index-tagging of the samples are achieved simultaneously. Designed in such a way, the assay allows us to obtain almost all mutation sequences from a small amount of data from NGS reads, especially in samples from colonies grown on selection plates, where the mutation sequences, although prone to bias, are highly concentrated. This results in the possibility to simultaneously analyze hundreds of samples from different experiments in a single NGS run. Here the cost of mutation spectrum analysis from selection plates was less than $1.00 per sample, even though both the NGS library preparation and the NGS run were outsourced. Further cost reductions can be expected by obtaining more sequencing data in a single run, and also from the continuous improvement of NGS technologies. From this point of view, we believe that we can secure a sufficient number of experiments to improve the accuracy of the analysis and to confirm the reproducibility of the experiments. Furthermore, the data from colonies grown on titer plates provides us, at least in principle, and with the exception of large deletions and insertions, with non-biased mutation spectra for the whole analyzed region. Also, mutation frequencies calculated from titer plate data were fifteen- to twenty-times higher than those obtained by the conventional *supF* assay, demonstrating that our novel assay allows us to analyse more thoroughly the mutation spectrum by using a smaller amount of data and to detect mutations with lower frequencies. The analysis of data from both selection and titer plates is consistent with the desired reliability and low-cost of the mutagenesis assay. Last but not least, by using the *supF* shuttle vector plasmid, mutation spectrum data can be obtained with different experimental designs, such as pre-treatment with metabolic enzymes or using cells with different genetic backgrounds, siRNA transfection, CRISPR-Cas systems, etc. We are currently conducting experiments with an *supF* library containing site-specific single DNA-damages such as 8-oxo-7,8-dihydroguanine (8-oxoG, 8-hydroxyguanine) among others (15,43,44), that will allow us to analyze in detail the repair mechanisms of sequence-specific DNA damage and provide further insights into mutagenesis and DNA repair processes.

The current study posits that when performing the assay, we have to be aware of the following points: 1) The required barcode complexity defined by the number of barcode sequences in the *supF* shuttle vector library depends on the experimental purpose and the mutation detection frequencies. A barcode complexity of 1.0×10^4 for each library is used as the standard in this paper. Even though more complexity may be able to provide better results, especially at lower mutation frequencies, one must keep in mind that the number of colonies to be harvested for the NGS sample preparation, as

well as the platform or setting for data analysis, need to be adjusted depending on the number of barcodes. 2) The number of NGS reads required for each multiplexed sample and for each N_{12}-BC sequence need to be calculated before the preparation of the samples. Therefore, the number of N_{12}-BC sequences and the number of harvested colonies must be optimized first. In addition, the allele frequency given by VarDict, referred to as variant frequency in this study, for each classified N_{12}-BC sequence is the most important determinant of the quality of the variant sequence data and the basis for analyzing overlapping N_{12}-BC sequences. The cost, time, and data quality must be all well-considered and balanced when an experiment is designed. 3) NGS data contain sequencing errors, the rate of which is known to be between 10^{-3} and 10^{-4}, with the majority of the errors being A:T to G:C and G:C to A:T transitions that had occurred during the PCR in the process of NGS library preparation (52,53). In the assay presented here no sonication step is required, thus avoiding DNA oxidation. As already mentioned, the clonal amplification of the *supF* shuttle vector via in vivo replication in the non-SOS *E. coli* provides a proper PCR template for preparation of NGS samples. In addition, the sequence alignment to the reference sequence for all identical N_{12}-BC sequences and the analysis of data obtained from duplicate PCR reactions can almost completely eliminate the sequencing errors resulting from the process of sample preparation. On the other hand, in the case of the amplicon sequencing employed here, the base-calling errors occurred depending on the "per base sequence quality" of the specific sequence. In this study, in addition to the *supF* gene with its total length of 335 bp being sequenced, 200-bp paired-end sequencing was performed with about 30 bp overlapping sequences, and the resulting sequence was merged together. The "per base sequence quality" in the region between read positions 160 and 200 bp is relatively low compared to other positions. Data shown in this study, as well as other experiments conducted in our laboratory, reveals that two positions in particular – 55 and 69 in the *supF* gene, manifest a very high frequency of variant calls – 20-50% of samples, but with low VF (less than 0.3). These two positions are located at 165-167 (depending on the N_0, N_1, N_2-primers) and 181-183 bp from the end of the forward reads, and at 179-181 and 167-169 bp from the end of the reverse reads. The addition of different numbers of random nucleotides – 0-, 1-, or 2-nt at the end of both primers used for sample preparation can contribute to better "per base sequence quality" in clonal amplicon sequencing. However, positions 55 and 69 are located immediately next to CC:GG sequences (position 55 is located in TTACC:GGTAA and position 69 is located in GTTCC:GGAAC), and the result of variant calling was almost exclusively SNS from 55A:T or 69T:A to C:G. Thus, these variant calls appear to be position-specific and sequence-dependent, and for this assay can be regarded as position-specific base-calling errors. Since they had a VF of less than 0.3 and a threshold VF of 0.4 was applied in this study, these base-calling errors do not need to be given consideration. Nevertheless, it must be emphasized that fine-tuning must be done for each individual experiment. 4) Large alterations in the *supF* gene, such as a large deletion or insertion, cannot be analyzed by this assay because of impossibility for PCR amplification or read alignments. This may render our observations on mutation frequency misleading. The frequency of such large alterations is better assessed by colony PCR, especially if the number of colonies grown on the titer plates was different from expected. 5) For the *supF* assay, spontaneous cluster mutations at TC:GA sites were often observed, and it was well illustrated in an earlier study

that a nick in the shuttle vector was a trigger for these asymmetric cluster mutations (54). Therefore, we need to be aware of the quality of each library and how it affects the outcome of each analysis, especially for detection of very low levels of mutations. Depending on the purpose of the experiments, in the preparation of covalently closed circular vector libraries it is essential to eliminate the background level of mutations. In fact, the *in vitro* construction of the library of double-stranded shuttle vectors from single-stranded circular DNA requires the process of treatment with T5 Exonuclease, which drastically decreases background mutations.

In this study, to test the performance of our newly developed *supF* NGS assay, we conducted experiments using N_{12}-BC libraries irradiated with UV-C at different doses. Although these pilot experiments were designed in a relatively simple way, they clearly exemplify the mutation spectrum associated with UV-radiation exposure, and further experiments may provide profound insights into UV-mutagenesis. Here, by using this novel assay we were able to confirm in both *E. coli* and human cells that the SNSs induced by UV were mainly C:G to T:A transitions, which is a known UV-signature, and that the UV-induced DNA lesions were basically dipyrimidines – CPDs at T-T sites, and (6-4)PPs at T-T and T-C sites (28-30,55,56). The SNSs are listed in Figure 10B (presented within pentanucleotide sequences of the *supF* gene) starting from the most frequent. In samples obtained from titer plates, the nucleotide sequence in eight of the top ten mutational hotspots was TTC/GAA, and the sequence of the top four mutational sites was TTCGA/TCGAA. These top four sites represent two palindromic TTCGAA/TTCGAA hexanucleotide sequences located at positions 115-120 and 123-128 in the *supF* gene. All top ten mutations were C to T transitions and occurred at TC sites, and are likely the result of TLS opposite the 3'C of CPDs and (6-4)PPs via specialized DNA polymerases, such as Polη, ι, ζ, κ, θ, λ, and Rev1 (57,58). Notably, the site selection bias observed for the top ten mutation hotspots seems to differ between the titer and the selection plates. In selection plates, three of the top ten hotspot mutations were C to T transitions at the 3'C of CC. These hotspot SNSs, particularly positions 130 and 121, appear to be next to or between hotspot SNSs in the titer plates. However, such SNS mutations, that were only observed in selection plates, were identified in both titer and selection plates as other types of mutations, specifically as hotspot DNS mutations, CC to TT, as shown in Figure 10C (located at 130 and 131 TTCGAATCCTT, at 121 TTCGAAGGTTCGAA, at 5 TTTGATA/TATCAAA, or at 84 and 85 AAGGGAG/CTCCCTT). In addition, as shown in Figures 9B and 9C, the multiple mutations, either C or G mutations at the same base in identical N_{12}-BC sequences, were identified in a dose-dependent manner more frequently from selection plates where TLS polymerases opposite UV-lesions on the same DNA strand can cause clustered mutations. These multiple mutations include at least one phenotypic mutation in the *supF* gene but include also non-phenotypic mutations, which contributes to the specifics of the mutation spectrum in the selection plates. The two-base substitutions within trinucleotides, referred to as Del-in by VarDict, were mainly detected at higher UV doses as GAG:CTC to AAA:TTT at positions 73 to 75. It is unclear whether these mutations arose through unexpected UV-induced lesions at GCTCG sites, or the deamination of either the 5'C or the 3'C has occurred at these specific sites by different mechanisms. Alternatively,

it can be hypothesized that the UV-induced DNA lesions arose on a single-stranded DNA (ssDNA) followed by a blockage of replication, or during DNA repair processes. It is known that ssDNA can be a substrate for a subclass of activation-induced deaminase (AID)/ apolipoprotein B mRNA-editing catalytic polypeptide-like subfamily cytosine deaminases (APOBECs), which cause cluster mutations at TC:GA sites in the same strand of DNA by converting cytosine to uracil (59). It is realistic to assume that the mutations at TC:GA sites had occurred via several different mechanisms. In fact, spontaneous mutations in non-UV-irradiated samples transfected into U2OS cells can also be detected at the order of 10^{-4} per gene. In both UV-irradiated and non-irradiated samples, SNS mutations had basically occurred from C:G at TC:GA sites. Also, the major proportion of multiple mutations were C:G to T:A transitions. However, in non-UV-irradiated samples the major proportion of single mutations were C:G to G:C transversions, which differed from the irradiated condition, where the major proportion was G:C to A:T transition (Figure 9C), It is possible that the oxidatively damaged base 8-oxoG was endogenously generated in the U2OS cells. As we have previously reported, 8-oxoG becomes a trigger for the DNA repair process and generates ssDNA as a substrate for APOBECs (15). Here, the TTC/GAA site has emerged from data from titer plates as a mutational hotspot. According to previous reports, the TTC motif, and more specifically the cytosine in the motif, is a site of preference for APOBEC3A, APOBEC3C, APOBEC3F and APOBEC3G, even though further investigation is needed (60-62).

What might be more, a particular mutation was detected by VarDict in non-UV-irradiated samples as a Del-in mutation – a change from TTGAT:ATCAA to ATCAA:TTGAT at position 3 to 7 within the potential hairpin loop of a quasi-palindromic sequence (Figure 14E), which may be the consequence of template switching events. The mutation was detected with a frequency at the order of 10^{-6} per gene (Figure 12E), and was hardly detected in UV-irradiated samples, where it remains hidden by the UV-introduced mutations. Also, the same mutation was frequently detected as single SNS mutations from T:A to A:T at position 3, from G:C to C:G at position 5, and from T:A to A:T at position 7, or as combinations with other SNS mutations either or both from G:C to A:T at position -3 or/and from T:A to C:G at position 12 (Supplementary Figure S8), which may be dependent on the regions used for template switching. Although the mono- and poly-ubiquitination of PCNA, facilitated by ubiquitin ligases Rad6-Rad18, HLTF or SHPRH, is involved in the mechanism of switching between TLS and template switching in mammalian cells, the details are still ambiguous (63-66). In addition, our recent studies on action-at-a-distance mutations, according to which a site-specific targeted single 8-oxoG or other types of DNA-damage cause cluster mutations at untargeted positions, demonstrated a drastic strand bias for mutations at cytosines, specifically at 5'-TC-3' sites (15,43,44). Data on mutational strand asymmetries from cancer genome sequencing studies have emerged, indicating that APOBEC-associated mutagenesis during replication strongly favors the lagging-strand template (67,68). And most recently, it has been proposed by a study using UPD-seq technique that APOBEC3A targets cytosines in small hairpin loops, within replication forks, in tRNA genes and at transcription start sites (69). In this study, strand biases in base substations from either C or G were clearly observed, consistent with what could be expected from the mechanisms by which UV-

photolesions are processed, such as cytosine deamination and the involvement of TLS. Interestingly, the strand bias switched between single base substitutions and cluster mutations depending on the doses of UV and the number of mutations per N_{12}-BC. The precise reason and mechanisms behind this remain to be elucidated, but the observation brings new insights into the mechanisms of mutagenesis. In addition, the positions of SNSs in the multiple mutations were closer to each other compared to the theoretically expected positions (Supplementary Figures S11 and S12), which may reflect switching events involving TLS polymerases. It should be noted that the presented data for the distance between two SNSs in the multiple mutations was analyzed from the data from selection plates in order to secure a sufficient number of mutations, and therefore, there may be a bias due to hot spots associated with the selection process. However, the results from the limited number of mutations from the titer plates are similar to these from the selection plates. It can be proposed that this assay may also be applied for analysis of TLS polymerases in mammalian cells.

Mutational signatures identified in cancer cells are emerging as valuable markers for cancer diagnosis and therapeutics. Innumerable physical, chemical and biological mutagens, including anticancer drugs, induce characteristic mutations in genomic DNA via specific mutagenic processes. The mutation spectra obtained here by using the presented advanced method were in good agreement with accumulated data from previous papers where the conventional method had been used, with the advantage that our method provided less-biased mutation spectra data. As described above, the datasets presented here highlighted novel mutational signatures and also cluster mutations with a strand-bias, which could be associated with the processes of replication, transcription, or repair of DNA-damage, including a single strand break (a nick). In this study, eight series of *supF* shuttle vector plasmids were constructed, as presented in Supplementary Figure S1; however, the analysis was carried out using N_{12}-BC libraries prepared from either pNGS2-K1 (Figures 1-4) or pNGS2-K3 (Figures 5-10). The pNGS2-K1/-A1/-K4/-A4 and pNGS2-K2/-A2/-K3/-A3 vector series contain an M13 intergenic region with opposite orientations relative to the *supF* gene, which allows us to incorporate specific types of DNA-damage at specific sites in the opposite strand of the vector library. Also, the pNGS2-K1/-A1/-K3/-A3 and pNGS2-K2/-A2/-K4/-A4 vector series contain the SV40 replication origin, which enables bidirectional replication and transcription, at opposite sides of the *supF* gene. Although this is still preliminary data, it is notable that the spontaneously induced mutations for the different vectors in U2OS cells were not significantly different. Therefore, the here presented mutagenesis assay with NGS, by using these series of libraries, can be applied in many different types of experiments to address both quantitative and qualitative features of mutagenesis. It is possible to design series of libraries containing DNA lesions or sequences suitable for the investigation of specific molecular mechanisms, such as TLS, template switching, and asymmetric cluster mutations. We are currently trying to develop a different test system for broader mutation spectrum analysis using the BC library with random oligonucleotide sequences; however, the assay using the *supF* gene presented in this paper has a significant advantage in both efficiency as a mutagenesis assay and cost-effectiveness. We believe that the method presented here will be used in further experiments by us and by other laboratories to advance the research on environmental mutagenesis, carcinogenesis,

DNA repair, and will hopefully contribute to mutagenesis research by promoting accumulation of data for machine learning.

DATA AVAILABILITY

NGS raw data has been uploaded to DDBJ database (http://www.ddbj.nig. ac.jp) under the accession number PRJDB13753 (DRA014420 for Figures 2-4, DRA014421 for Figures 5 and 6, DRA014422 for Figures 7 and 8, DRA014422 for the samples without transformation into *E. coli*)

SUPPLEMENTARY DATA

Supplementary Figures S1-S12 and Supplementary Tables S1-S7.

ACKNOWLEDGEMENT

A part of this study was carried out at the Analysis Center of Life Science, Natural Science Center for Basic Research and Development, and the Joint Usage/Research Center Program of the Research Institute for Radiation Biology and Medicine (RIRBM), Hiroshima University.

FUNDING

This work was supported in part by the Japan Society for the Promotion of Science (JSPS) KAKENHI grant number JP 19K12322 (H Kawai).

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

1. OECD. (2016) Test No. 487: In Vitro Mammalian Cell Micronucleus Test.
2. OECD. (2020) Test No. 471: Bacterial Reverse Mutation Test.
3. David, R. (2020) The promise of toxicogenomics for genetic toxicology: past, present and future. *Mutagenesis*, **35**, 153-159.

4. Nik-Zainal, S., Alexandrov, L.B., Wedge, D.C., Van Loo, P., Greenman, C.D., Raine, K., Jones, D., Hinton, J., Marshall, J., Stebbings, L.A. et al. (2012) Mutational processes molding the genomes of 21 breast cancers. *Cell*, **149**, 979-993.

5. Meier, B. and Gartner, A. (2014) Having a direct look: analysis of DNA damage and repair mechanisms by next generation sequencing. *Exp Cell Res*, **329**, 35-41.

6. Koh, G., Degasperi, A., Zou, X., Momen, S. and Nik-Zainal, S. (2021) Mutational signatures: emerging concepts, caveats and clinical applications. *Nat Rev Cancer*, **21**, 619-637.

7. Razzaque, A., Mizusawa, H. and Seidman, M.M. (1983) Rearrangement and mutagenesis of a shuttle vector plasmid after passage in mammalian cells. *Proc Natl Acad Sci U S A*, **80**, 3010–3014.

8. Drinkwater, N.R. and Klinedinst, D.K. (1986) Chemically induced mutagenesis in a shuttle vector with a low-background mutant frequency. *Proc Natl Acad Sci U S A*, **83**, 3402–3406.

9. Gossen, J.A., de Leeuw, W.J., Tan, C.H., Zwarthoff, E.C., Berends, F., Lohman, P.H., Knook, D.L. and Vijg, J. (1989). Efficient rescue of integrated shuttle vectors from transgenic mice: a model for studying mutations in vivo. *Proc Natl Acad Sci U S A*, **86**, 7971–7975.

10. White, P.A., Luijten, M., Mishima, M., Cox, J.A., Hanna, J.N., Maertens, R.M. and Zwart, E.P. (2019) In vitro mammalian cell mutation assays based on transgenic reporters: A report of the International Workshop on Genotoxicity Testing (IWGT). *Mutat Res Genet Toxicol Environ Mutagen*, **847**, 403039.

11. Parris, C.N. and Seidman, M.M. (1992) A signature element distinguishes sibling and independent mutations in a shuttle vector plasmid. *Gene*, **117**, 1-5.

12. Parris, C.N., Levy, D.D., Jessee, J. and Seidman, M.M. (1994) Proximal and distal effects of sequence context on ultraviolet mutational hotspots in a shuttle vector replicated in xeroderma cells. *J Mol Biol*, **236**, 491-502.

13. Sarkar, S., Dasgupta, U.B. and Summers, W.C. (1984) Error-prone mutagenesis detected in mammalian cells by a shuttle vector containing the supF gene of Escherichia coli. *Mol Cell Biol*, **4**, 2227–2230.

14. Yoon, J.H., McArthur, M.J., Park, J., Basu, D., Wakamiya, M., Prakash, L. and Prakash, S. (2019) Error-Prone Replication through UV Lesions by DNA Polymerase theta Protects against Skin Cancers. *Cell*, **176**, 1295-1309 e1215.

15. Suzuki, T., Zaima, Y., Fujikawa, Y., Fukushima, R. and Kamiya, H. (2022) Paradoxical role of the major DNA repair protein, OGG1, in action-at-a-distance mutation induction by 8-oxo-7,8-dihydroguanine. *DNA Repair (Amst)*, **111**, 103276.

16. Camakaris, H. and Pittard, J. (1973). Regulation of tyrosine and phenylalanine biosynthesis in Escherichia coli K-12: properties of the tyrR gene product. *J Bacteriol*, **115**, 1135–1144.

17. Normanly, J., Ogden, R.C., Horvath, S.J. and Abelson, J. (1986) Changing the identity of a transfer RNA. *Nature*, **321**, 213-219.

18. Obata, F., Nunoshiba, T., Hashimoto-Gotoh, T. and Yamamoto, K. (1998) An improved system for selection of forward mutations in an Escherichia coli supF gene carried by plasmids. *J Radiat Res*, **39**, 263-270.

19. Kraemer, K.H. and Seidman, M.M. (1989). Use of supF, an Escherichia coli tyrosine suppressor tRNA gene, as a mutagenic target in shuttle-vector plasmids. *Mutat Res*, **220**, 61-72.

20. Fukushima, R., Suzuki, T. and Kamiya, H. (2020) New indicator Escherichia coli strain for rapid and accurate detection of supF mutations. *Genes Environ*, 42, 28.

21. Shendure, J., Balasubramanian, S., Church, G.M., Gilbert, W., Rogers, J., Schloss, J.A. and Waterston, R.H. (2017) DNA sequencing at 40: past, present and future. *Nature*, **550**, 345-353.

22. Dutta, A.K., Alberge, J.B., Sklavenitis-Pistofidis, R., Lightbody, E.D., Getz, G. and Ghobrial, I.M. (2022) Single-cell profiling of tumour evolution in multiple myeloma - opportunities for precision medicine. *Nat Rev Clin Oncol*, **19**, 223-236.

23. Arrastia, M.V., Jachowicz, J.W., Ollikainen, N., Curtis, M.S., Lai, C., Quinodoz, S.A., Selck, D.A., Ismagilov, R.F. and Guttman, M. (2022) Single-cell measurement of higher-order 3D genome organization with scSPRITE. *Nat Biotechnol*, **40**, 64-73.

24. Schmitt, M.W., Kennedy, S.R., Salk, J.J., Fox, E.J., Hiatt, J.B. and Loeb, L.A. (2012) Detection of ultra-rare mutations by next-generation sequencing. *Proc Natl Acad Sci U S A*, **109**, 14508-14513.

25. Jee, J., Rasouly, A., Shamovsky, I., Akivis, Y., R. Steinman, S., Mishra, B. and Nudler, E. (2016) Rates and mechanisms of bacterial mutagenesis from maximum-depth sequencing. *Nature*, **534**, 693-696.

26. Sloan, D.B., Broz, A.K., Sharbrough, J. and Wu, Z. (2018) Detecting Rare Mutations and DNA Damage with Sequencing-Based Methods. *Trends in Biotechnology*, **36**, 729-740.

27. Napolitano, R.L., Lambert, I.B. and Fuchs, R.P. (1997) SOS factors involved in translesion synthesis. *Proc Natl Acad Sci U S A*, **94**, 5733-5738.

28. Varghese, A.J., and Wang, S.Y. (1967) Ultraviolet irradiation of DNA in vitro and in vivo produces a 3d thymine-derived product. *Science*, **19**, 955–957.

29. Mao, P., Smerdon, M.J., Roberts, S.A. and Wyrick, J.J. (2016) Chromosomal landscape of UV damage formation and repair at single-nucleotide resolution. *Proc Natl Acad Sci U S A*, **113**, 9057-9062.

30. Lindberg, M., Bostrom, M., Elliott, K. and Larsson, E. (2019) Intragenomic variability and extended sequence patterns in the mutational signature of ultraviolet light. *Proc Natl Acad Sci U S A*, **116**, 20411-20417.

31. Alexandrov, L.B., Kim, J., Haradhvala, N.J., Huang, M.N., Tian Ng, A.W., Wu, Y., Boot, A., Covington, K.R., Gordenin, D.A., Bergstrom, E.N. et al. (2020) The repertoire of mutational signatures in human cancer. *Nature*, **578**, 94-101.

32. Pfeifer, G.P., You, Y.H. and Besaratinia, A. (2005) Mutations induced by ultraviolet light. *Mutat Res*, **571**, 19-31.

33. Ikehata, H. and Ono, T. (2011) The mechanisms of UV mutagenesis. *J Radiat Res*, **52**, 115-125.

34. Roberts, S.A. and Gordenin, D.A. (2014) Hypermutation in human cancer genomes: footprints and mechanisms. *Nat Rev Cancer*, **14**, 786-800.

35. Stary, A. and Sarasin, A. (1992) Simian virus 40 (SV40) large T antigen-dependent amplification of an Epstein-Barr virus-SV40 hybrid shuttle vector integrated into the human HeLa cell genome. *Journal of General Virology*, **73**, 1679-1685.

36. Zou, Q., Shen, W., Le, S., Li, Y. and Hu, F. (2016) SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *Plos One*, **11**, e0163962.

37. Kwon, S., Lee, B. and Yoon, S. (2014) CASPER: context-aware scheme for paired-end reads from high-throughput amplicon sequencing. BMC Bioinformatics, 15, S10.

38. Li, H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**, 3094-3100.

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

40. Li, H. (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, **27**, 2987-2993.

41. Smith, T., Heger, A. and Sudbery, I. (2017) UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res*, **27**, 491-499.

42. Lai, Z., Markovets, A., Ahdesmaki, M., Chapman, B., Hofmann, O., McEwen, R., Johnson, J., Dougherty, B., Barrett, J.C. and Dry, J.R. (2016) VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res*, **44**, e108.

43. Suzuki, T., Kuramoto, Y. and Kamiya, H. (2018) Reduction of Werner Syndrome Protein Enhances G:C → A:T Transition by O^6-Methylguanine in Human Cells. *Chem Res Toxicol*, **31**, 319-324.

44. Suzuki, T., Katayama, Y., Komatsu, Y. and Kamiya, H. (2021) Similar frequency and signature of untargeted substitutions induced by abasic site analog under reduced human APE1 conditions. *J. Toxicol. Sci.*, **46**, 283–288.

45. Foox, J., Tighe, S.W., Nicolet, C.M., Zook, J.M., Byrska-Bishop, M., Clarke, W.E., Khayat, M.M., Mahmoud, M., Laaguiby, P.K., Herbert, Z.T. et al. (2021) Performance assessment of DNA sequencing platforms in the ABRF Next-Generation Sequencing Study. *Nat Biotechnol*, **39**, 1129-1140.

46. Pugh, T.J., Morozova, O., Attiyeh, E.F., Asgharzadeh, S., Wei, J.S., Auclair, D., Carter, S.L., Cibulskis, K., Hanna, M., Kiezun, A. et al. (2013) The genetic landscape of high-risk neuroblastoma. *Nat Genet*, **45**, 279-284.

47. Ma, X., Shao, Y., Tian, L., Flasch, D.A., Mulder, H.L., Edmonson, M.N., Liu, Y., Chen, X., Newman, S., Nakitandwe, J. et al. (2019) Analysis of error profiles in deep next-generation sequencing data. *Genome Biol*, **20**, 50.

48. Seidman, M.M., Dixon, K., Razzaque, A., Zagursky, R.J. and Berman, M.L. (1985) A shuttle vector plasmid for studying carcinogen-induced point mutations in mammalian cells. *Gene*, **38**, 233–237.

49. Yoon, J.H., Johnson, R.E., Prakash, L. and Prakash, S. (2021) Implications of inhibition of Rev1 interaction with Y family DNA polymerases for cisplatin chemotherapy. *Genes Dev*, **35**, 1256-1270.

50. McCombie, W.R., McPherson, J.D. and Mardis, E.R. (2019) Next-generation sequencing technologies. *Cold Spring Harbor Perspectives in Medicine*, **9**, a036798.

51. Wang, K., Ma, X., Zhang, X., Wu, D., Sun, C., Sun, Y., Lu, X., Wu, C.I., Guo, C. and Ruan, J. (2016) Using ultra-sensitive next generation sequencing to dissect DNA damage-induced mutagenesis. *Sci Rep*, **6**, 25310.

52. Brodin, J., Mild, M., Hedskog, C., Sherwood, E., Leitner, T., Andersson, B. and Albert, J. (2013) PCR-Induced Transitions Are the Major Source of Error in Cleaned Ultra-Deep Pyrosequencing Data. *PLOS ONE*, **8**, e70388.

53. Costello, M., Pugh, T.J., Fennell, T.J., Stewart, C., Lichtenstein, L., Meldrim, J.C., Fostel, J.L., Friedrich, D.C., Perrin, D., Dionne, D. et al. (2013) Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Research*, **41**, e67-e67.

54. Seidman, M.M., Bredberg, A., Seetharam, S. and Kraemer, K.H. (1987) Multiple point mutations in a shuttle vector propagated in human cells: evidence for an error-prone DNA polymerase activity. *Proc Natl Acad Sci U S A*, **84**, 4944-4948.

55. Kamiya, H., Murata, N., Murata, T., Iwai, S., Matsukage, A., Masutani, C., Hanaoka, F. and Ohtsuka, E. (1993) Cyclobutane thymine dimers in a ras proto-oncogene hot spot activate the gene by point mutation. *Nucleic Acids Res*, **21**, 2355-2361.

56. Kamiya, H., Iwai, S. and Kasai, H. (1998) The (6-4) photoproduct of thymine-thymine induces targeted substitution mutations in mammalian cells. *Nucleic Acids Res*, **26**, 2611-2617.

57. Achar, Y.J. and Foiani, M. (2019) An Error-Prone Polymerase in the Fight against Cancer. *Cell*, **176**, 1241-1243.

58. Yoon, J.H., Basu, D., Sellamuthu, K., Johnson, R.E., Prakash, S. and Prakash, L. (2021) A novel role of DNA polymerase lambda in translesion synthesis in conjunction with DNA polymerase zeta. *Life Sci Alliance*, **4**, e202000900.

59. Pecori, R., Di Giorgio, S., Paulo Lorenzo, J. and Nina Papavasiliou, F. (2022) Functions and consequences of AID/APOBEC-mediated DNA and RNA deamination. *Nat Rev Genet*. *in press*.

60. Lindberg, M., Bostrom, M., Elliott, K. and Larsson, E. (2019) Intragenomic variability and extended sequence patterns in the mutational signature of ultraviolet light. *Proc Natl Acad Sci U S A*, **116**, 20411-20417.

61. Love, R.P., Xu, H. and Chelico, L. (2012) Biochemical analysis of hypermutation by the deoxycytidine deaminase APOBEC3A. *J Biol Chem*, **287**, 30812-30822.

62. Adolph, M.B., Love, R.P. and Chelico, L. (2018) Biochemical Basis of APOBEC3 Deoxycytidine Deaminase Activity on Diverse DNA Substrates. *ACS Infect Dis*, **4**, 224-238.

63. Chiu, R.K., Brun, J., Ramaekers, C., Theys, J., Weng, L., Lambin, P., Gray, D.A. and Wouters, B.G. (2006) Lysine 63-polyubiquitination guards against translesion synthesis-induced mutations. *PLoS Genet*, **2**, e116.

64. Lin, J.R., Zeman, M.K., Chen, J.Y., Yee, M.C. and Cimprich, K.A. (2011) SHPRH and HLTF act in a damage-specific manner to coordinate different forms of postreplication repair and prevent mutagenesis. *Mol Cell*, **42**, 237-249.

65. Masuda, Y., Suzuki, M., Kawai, H., Hishiki, A., Hashimoto, H., Masutani, C., Hishida, T., Suzuki, F. and Kamiya, K. (2012) En bloc transfer of polyubiquitin chains to PCNA in vitro is mediated by two different human E2–E3 pairs. *Nucleic Acids Research*, **40**, 10394-10407.

66. Wong, R.P., Petriukov, K. and Ulrich, H.D. (2021) Daughter-strand gaps in DNA replication - substrates of lesion processing and initiators of distress signalling. *DNA Repair (Amst)*, **105**, 103163.

67. Seplyarskiy, V.B., Soldatov, R.A., Popadin, K.Y., Antonarakis, S.E., Bazykin, G.A. and Nikolaev, S.I. (2016) APOBEC-induced mutations in human cancers are strongly enriched on the lagging DNA strand during replication. *Genome Res*, **26**, 174-182.

68. Haradhvala, Nicholas J., Polak, P., Stojanov, P., Covington, Kyle R., Shinbrot, E., Hess, J.M., Rheinbay, E., Kim, J., Maruvka, Y.E., Braunstein, Lior Z. et al. (2016) Mutational Strand Asymmetries in Cancer Genomes Reveal Mechanisms of DNA Damage and Repair. *Cell*, **164**, 538-549.

69. Sakhtemani, R., Perera, M.L.W., Hübschmann, D., Siebert, R., Lawrence, M.S. and Bhagwat, A.S. (2022) Human activation-induced deaminase lacks strong replicative strand bias or preference for cytosines in hairpin loops. *Nucleic Acids Research*, **50**, 5145-5157.

FIGURES LEGENDS

Figure 1. Strategy for the *supF* mutagenicity NGS assay

(A) Schematic map of the shuttle vector pNGS2-K1 or -A1. The shuttle vector plasmid DNA encodes the amber suppressor tRNA (supF) gene, the TP53-/Rb-binding-deficient mutant SV40 large T antigen (E107K/D402E) gene, the SV40 replication origin, the pBR327 origin of replication, the M13 intergenic region, the kanamycin- or ampicillin-resistance gene (km^r and amp^r), and the randomized 12-nucleotide barcode DNA sequence (N_{12}-BC). (B) Nucleotide sequence of the non-transcribed strand (sense strand) of the DNA fragment from N_{12}-BC libraries amplified by PCR in preparation for NGS. The underlined letters indicate the primer set for PCR amplification (Supplementary Figure S2). All primer sets contain 0, 1, or 2 random nucleotides at the 5' end to ensure proper signal detection for amplicon sequencing with a short-read NGS platform. A pre-designed 6-nucleotide sequence (N_6) in one of the primers serves as an index sequence for distinguishing multiplex samples in a single sequencing run. The N_{12} (green background color) represents the N_{12}-BC DNA sequence used for deep-sequencing data analysis to identify variants from identical template shuttle vectors. The sequence with red background indicates the *supF* gene starting at position 1. (C) Schematic illustration of the mutagenicity NGS assay procedure. One or more of the N_{12}-BC libraries (i) ~ (iv) can be applied for the sequencing analysis.

Figure 2. Number of classified unique N_{12}-BC sequences in the multiplexed NGS libraries prepared by PCR

(A) In the caption of each pair of graphs, the "10^n" (Lib10^2_#1, Lib10^2_#2, Lib10^3, and Lib10^4) indicates the approximate number of colonies used for the N_{12}-BC library preparation from pNGS2-K1, and the "n M/sample" indicates expected "n" million reads per individual sample in the multiplexed NGS. The left-side graph in each of the four pairs represents the number of classified N_{12}-BC sequences in indexed samples prepared by PCR with different amounts of template (1.0, 3.3, 10, 33, and 100 ng, indicated by graded blue colors). The right side graph shows the observed number of reads for each indexed sample. Source data is available in Supplementary Tables S1A-S1D. (B) Coverage distributions for each unique N_{12}-BC sequence for the samples in Fig 2A.

Figure 3. Discrimination between true mutations and sequencing errors at the nucleotide positions

(A) Number of N_{12}-BC sequences with a variant exceeding 0.4 VF according to their nucleotide positions. The bar colors reflect the variant frequency, as indicated by the heatmap at the top of panel (A). Source data is available in Supplementary Table S2. (B) The same dataset as in panel (A), with the bar colors reflecting different variant types. The variant calls are categorized into 6 types: one-nucleotide deletion (1-Del), deletions larger than one nucleotide (Del), deletions with insertions (Del-in), dinucleotide substitutions (DNS), insertions (INS), and single nucleotide substitutions (SNS). The data for the BC10^4 library is outlined with a dashed line and enlarged. (C) Number of N_{12}-BC sequences with a variant in the 0.1-0.4 VF range according to their nucleotide positions. (D) Histograms of the distribution of the number of N_{12}-BC sequences by variant frequency. The data for the 0.3-1.0 variant frequency range are enlarged on the right side. Different colors indicate the positions of the variant sequence: position 55 (violet), 69 (orange) or other (blue). (E) Comparison of the possible sequencing errors at different nucleotide positions in the *supF* gene, presented as percentage of BC sequences with variants below VF 0.4. The value in the right-hand side column of the table is the median from all analyzed samples (n=20).

Figure 4. Performance of the *supF* mutagenicity NGS assay in bacterial cells (data from titer plates)

(A) Number of classified N_{12}-BC sequences analyzed from approximately 10^4 colonies from titer plates (see Fig 1C (iii)). Lib10^4_#1 (#1) and Lib10^4_#2 (#2) represent two independently prepared BC10^4 libraries from pNGS2-K3 irradiated with UV-C at doses of 0, 50, 100, and 200 J/m^2 and used for transformation of the indicator *E. coli* strain RF01. Source data is available in Supplementary

Table S3A. (B) Number of N_{12}-BC sequences with a variant sequence exceeding VF 0.4. The number in brackets refers to the number of unique BCs with variant sequences. (C) Number of N_{12}-BC sequences with a variant exceeding VF 0.4 at different UV-C doses and their distribution by nucleotide position. The variant types are shown in different colors as indicated in the legend. (D) The same dataset shown in Figure 4B is now separated into two graphs: the graph on the left side represents positions from -67 to -20, and the graph on the right side – positions from -19 to 214 containing the *supF* gene. (E) Pie charts of the proportions of mutation types: 1-Del, Del, DNS, Del-in, INS, SNS, categorized as in Figure 3B. Source data is available in Supplementary Table S3B. (F) Mutation spectrum shown as pie charts of the proportions of different base-substitutions. Source data is available in Supplementary Table S3C. (G) The *supF* mutation frequencies determined by NGS. The mutation frequency is determined by dividing the number of N_{12}-BC sequences with a variant sequence at positions from -19 to 214 by the total number of N_{12}-BC sequences in the sample (data shown in Figure 4D and 4A, respectively). (H) The *supF* mutant frequencies determined by the conventional method, i.e. by dividing the number of colonies grown on the selection plate by the number of colonies grown on the titer plate.

Figure 5. The 192 trinucleotide contexts of UV-radiation induced SNSs in bacterial cells (data from titer plates)

UV-radiation induced SNSs shown in 192 trinucleotide contexts for the non-transcribed (top panel) and transcribed (bottom panel) strand of the *supF* gene (nucleotide positions from -19 to 214). The data is combined from $Lib10^4$_#1 and $Lib10^4$_#2. The substituted bases are indicated in colors corresponding to the notation on the horizontal axis; the UV doses are shown on the right side of the graph; the reference sequence for the analysis is denoted as Ref. Source data is available in Supplementary Table S3D.

Figure 6. Performance of the *supF* mutagenicity NGS assay in bacteria cells (data from selection plates)

(A) Number of classified N_{12}-BC sequences from colonies in selection plates (see Figure 1C (iv)). The approximate numbers of harvested colonies are shown on top of the graph in a rectangle denoted as “Harvested colonies #”. The number above each bar represents the number of classified N_{12}-BC sequences, and the number in the brackets refers to the number of unique BCs with variant sequences. Source data is available in Supplementary Table S4A. (B) Number of N_{12}-BC sequences with a variant exceeding VF 0.4 at different UV doses and their distribution by nucleotide position. The variant types are shown in different colors as indicated in the legend. (C) Pie charts of the percentage distribution of different mutation types. Source data is available in Supplementary Table S4B. (D) Pie charts of the percentage distribution of different SNSs. Source data is available in Supplementary

Table S4C. (E) Pie charts of the percentage distribution of different DNSs. Source data is available in Supplementary Table S4D.

Figure 7. The 192 trinucleotide contexts of UV-radiation induced SNSs in bacterial cells (data from selection plates)

SNSs induced by UV-irradiation in the *supF* gene in 192 trinucleotide contexts. The strand-specific number of BCs are shown for each SNSs, analogous to Figure 5I, but the data are independently plotted for libraries Lib10^4_#1 and Lib10^4_#2. Source data is available in Supplementary Table S4E.

Figure 8. Performance of the *supF* mutagenicity NGS assay in mammalian cells (data from titer plates)

The layout of the figure is analogous to Figure 4, this time for RF01 transformed with libraries extracted from transfected U2OS cells. (A) Numbers of classified N_{12}-BC sequences analyzed from approximately 10^4 colonies from titer plates. Source data is available in Supplementary Table S5A. (B) Number of N_{12}-BC sequences with a variant sequence exceeding VF 0.4 (nucleotide positions from -67 to 214). (C) Number of N_{12}-BC sequences with a variant sequence exceeding VF 0.4 and their distribution by nucleotide position. (D) The dataset shown in panel B is now separated into two graphs according to nucleotide position relative to the *supF* gene. The graph on the left side represents positions from -67 to -20, and the graph on the right side represents positions from -19 to 214. (E) Pie charts of the percentage distribution of different mutation types. Source data is available in Supplementary Table S5B. (F) Pie charts of the percentage distribution of different SNSs. Source data is available in Supplementary Table S5C. (G) The *supF* mutation frequencies determined by the NGS assay. (H) The *supF* mutant frequencies determined by the conventional method.

Figure 9. The 192 trinucleotide contexts of UV-radiation induced SNSs in mammalian cells (data from titer plates)

UV-induced SNSs in 192 trinucleotide contexts for the non-transcribed (top panel) and transcribed (bottom panel) strand of the *supF* gene. Source data is available in Supplementary Table S5D.

Figure 10. Performance of the *supF* mutagenicity NGS assay in mammalian cells (data from selection plates)

The data represents RF01 transformed with libraries extracted from transfected U2OS cells. For identical N_{12}-BC sequences with variants the data is obtained from duplicate PCR samples. (A)

Number of classified N_{12}-BC sequences detected from colonies in selection plates (see Figure 1C (iv)). Approximately 1×10^3 colonies were harvested and analyzed. The number above each bar represents the number of classified N_{12}-BC sequences, and the number in the brackets refers to the number of unique BCs with variant sequences. Source data is available in Supplementary Table S6A. (B) Number of N_{12}-BC sequences with single or multiple (1~6) mutations. The number of BCs with a single mutation is indicated inside each bar, with their percentage of the total N_{12}-BC sequences added in brackets; the total number of mutations is indicated outside the bars and corresponds to the "# of unique BCs" in panel (A). Source data is available in Supplementary Table S6B. (C) Number of N_{12}-BC sequences with a variant exceeding VF 0.4 according to their nucleotide position. (D) Pie charts of the percentage distribution of mutation types. Source data is available in Supplementary Table S6B. (E) Pie charts of the percentage distribution of different SNSs. Source data is available in Supplementary Table S6D.

Figure 11. The 192 trinucleotide contexts of UV-radiation induced SNSs in mammalian cells (data from selection plates)

SNSs induced by UV-irradiation in the *supF* gene shown in 192 trinucleotide contexts. The data for the non-transcribed (top panel) and transcribed (bottom panel) strand of the *supF* gene is shown. Source data is available in Supplementary Table S6E.

Figure 12. UV-induced mutations in the *supF* gene in mammalian cells (data from selection plates)

The mutation sequence on the non-transcribed strand of the *supF* gene is analyzed. (A) Proportions of mutation types in BCs with a single variant sequence (Single mt/BC) or more than one variant sequence (2~ mt/BC). Source data is available in Supplementary Table S7A. (B) Number of SNSs induced by UV-irradiation – one SNS per BC (graph on top), and multiple SNSs (2~) per BC (bottom graph). The nucleotide sequences of SNSs are indicated by colors as shown in the legend. Source data is available in Supplementary Table S7B. (C) Number of mutations induced by UV-irradiation classified by mutation type, except for SNSs. The mutation types are shown in different colors as indicated in the legend. For each color, the lighter shade indicates single mutation per BC, and the darker shade indicates multiple mutations per BC. Source data is available in Supplementary Table S7C. (D) Number of DNSs and their mutation sequences. The mutations have been listed if more than 2 identical mutation sequences among all samples have been detected. Source data is available in Supplementary Table S7D. (E) Number of Del-in mutations and their sequences. The mutations have been listed if more than 2 identical mutation sequences among all samples have been detected. Source data is available in Supplementary Table S7E.

Figure 13. Numbers and combinations of detected mutations per N_{12}-BC in mammalian cells (data from selection plates, 200 J/m^2)

The numbers of detected mutations for each single mutation per N_{12}-BC (lower part) and for combinations of more than one mutations per N_{12}-BC (upper and middle part) are shown by stacked thin horizontal bars. The type of mutation is indicated by a different color (legend on top) and by its abbreviation on the horizontal axis. The number in brackets and the height of the stacked horizontal bars represent the number of BCs with the mutation(s).

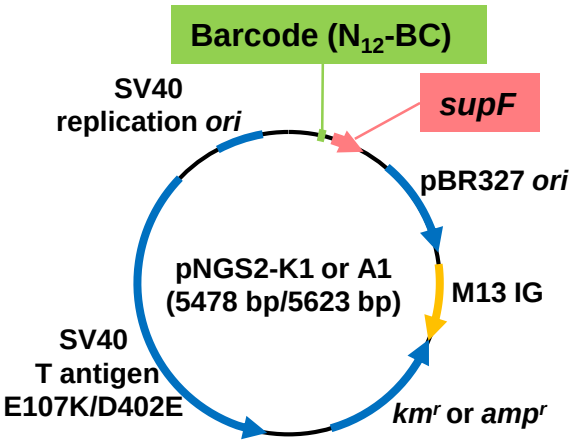
Figure 14. Summary of results for UV-induced mutagenesis of the *supF* gene in mammalian cells

(A) The nucleotide sequence of the non-transcribed strand (sense strand) is analyzed (233 bp, positions -19 to 214). The nucleotides with an asterisk are associated with the *supF* mutant phenotype resulting from one SNS according to our previous experimental data. The 5'-TC-3' sites are shown in green color, and the 5'-GA-3' sites are shown in red color. The nucleotide positions with a number in a circle correspond to the sequences listed in panel (B). A summary of the numbers of indicated 5' to 3' sequences from the non-transcribed strand and their complementary sequences from the transcribed strand are shown in the text box. The first number in the brackets represents the total number of the indicated nucleotide sequences, and the number with an asterisk represents the number of nucleotides associated with the *supF* mutant phenotype resulting from SNSs at the site of the underlined nucleotide. (B) The UV-induced SNSs at individual positions are sorted by the frequency of their detection (number of BCs) from most to least. The rank of the frequency is shown by the numbers in either a white circle referring to data from titer plates (corresponding to Figures 8 and 9) or a black circle referring to data from selection plates (corresponding to Figures 10 and 11). The SNS is denoted as an underlined nucleotide in a pentanucleotide sequence (including two nucleotides upstream and downstream of the mutation, and the sequence of the non-transcribed strand/transcribed strand). The number next to the rank is the position of each mutation, corresponding to the nucleotide sequence in panel (A). The number in brackets represents the number of BCs with the mutation. (C) Data for double nucleotide substitutions, analogous to panel (B). The majority of converted two-nucleotide sequences for each underlined two nucleotides in a hexanucleotide sequence is shown by arrows. The number of the tandem two-nucleotide substitutions is shown before the slash in the brackets, and the total number of the two-nucleotide substitutions – after the slash. (D) Data for deletion/insertion mutations, analogous to panels (B) and (C). (E) The potential hairpin loop of the quasi-palindromic sequence in the reference sequence (on the left side), and the mutated sequence detected from the selection plate of the non-irradiated sample (on the right side). The numbers represent the nucleotide position in the non-transcribed strand sequence of the *supF* gene. The potential Watson-Crick hydrogen bonded base pair is shown by an asterisk. The

nucleotide sequences detected as SNS mutations at positions -3 and 12 are shown in orange color, and for deletion/insertion at positions 3 to 7 – in purple color.

A)

supF shuttle vector N_{12}-BC library



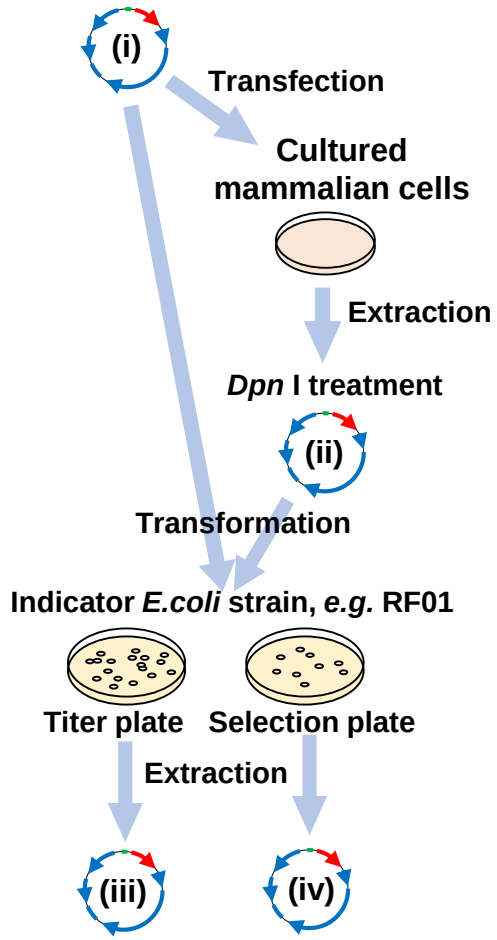
B)

```
-100 (N0~2)ATGC(N6)TTGTAAAGGCCTCAGCGAATTG -67
-66  CAAGCTTCTAGAAGGCGAT(N12)ATCGAATTCGGA -36
-35  TCCTTTCTCAACGTAACACTTTACAGCGGCGCGTC  -1
1    ATTTGATATGATGCGCCCCGCTTCCCGATAAGGGA  35
36   GCAGGCCAGTAAAAGCATTACCTGTGGTGGGGTTC  70
71   CCGAGCGGCCAAAGGGAGCAGACTCTAAATCTGCC  105
106  GTCATCGACTTCGAAGGTTCGAATCCTTCCCCCAC  140
141  CACCATCACTTTCAAAAGTCCGAAAAGATCCGGCG  175
176  GCGCAGCACCTCATCCTCAGCGATATGGCAGCTTA  210
211  TAATCTGAAAGAGGAACTTGGTTAG(N0~2)      235
```

335 bp + 12 bp (N_{12}-BC) + 0, 2, 4 bp ($N_{0\sim2}$ x 2)

C)

N_{12}-BC library
(BC#:10^2 or 10^3 or 10^4)



1. Preparation of N_{12}-BC library

   1-a. No treatment ---- (i)
   1-b. Mutagen treatment ---- (i)
   1-c. *in vitro* construction for site-specific DNA damage ---- (i)

2. Transfection of N_{12}-BC library into cultured mammalian cells

   2-a. No transfection
   2-b. Mutagen treatment
   2-c. Genetically modified
   → Extraction and *Dpn* I treatment ---- (ii)

3. Transformation of indicator non-SOS E. coli. strain with N_{12}-BC library for in vivo amplification

   3-a. Titer plates ---- (iii)
   3-b. Selection plates ---- (iv)

4. The index (N_6) tagging for multiplexed NGS (i) ~ (iv) by PCR

5. MGI-Tech or Illumina sequencing platforms

6. Sequencing analysis (White list/VarDict)

Fig. 1

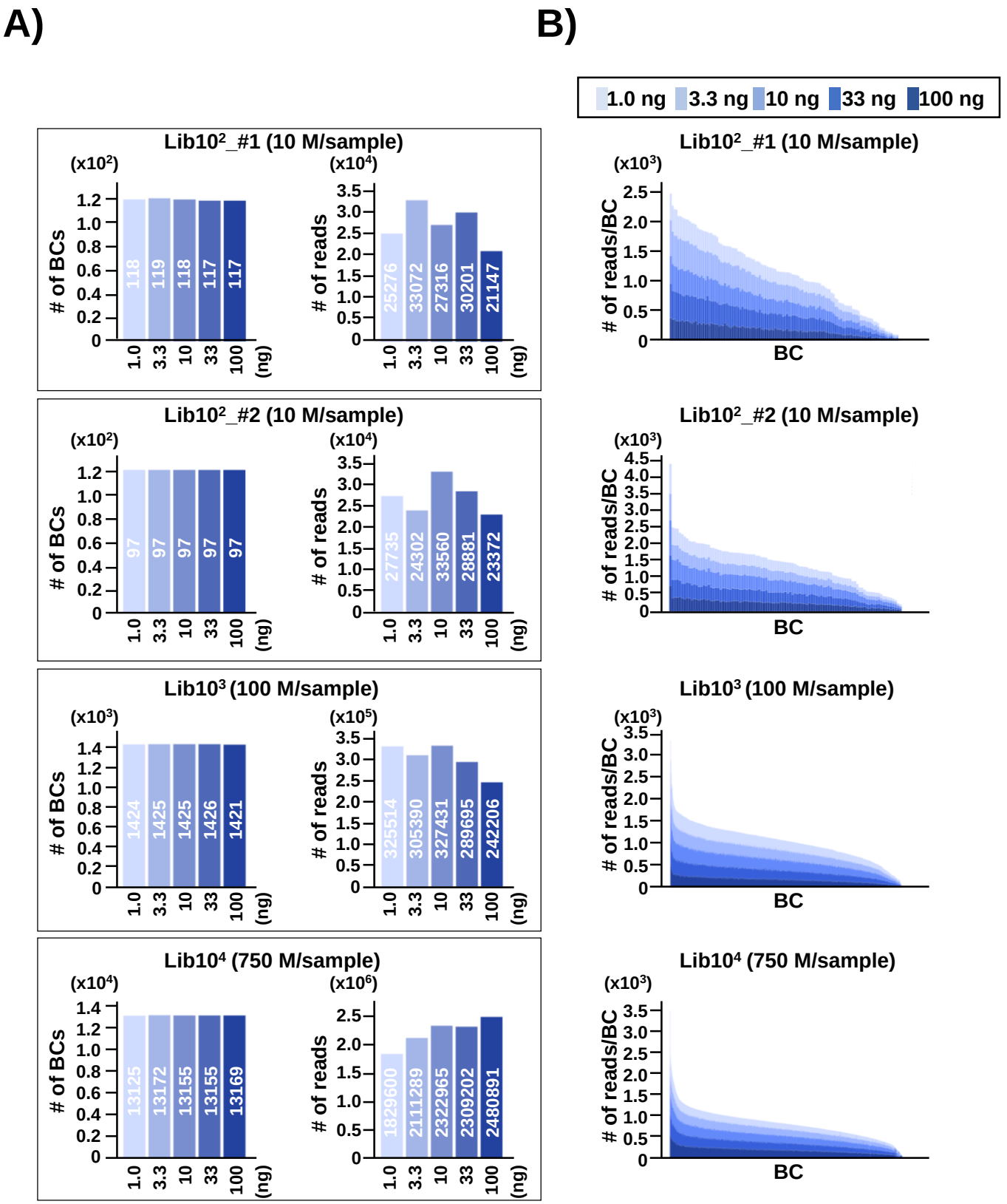
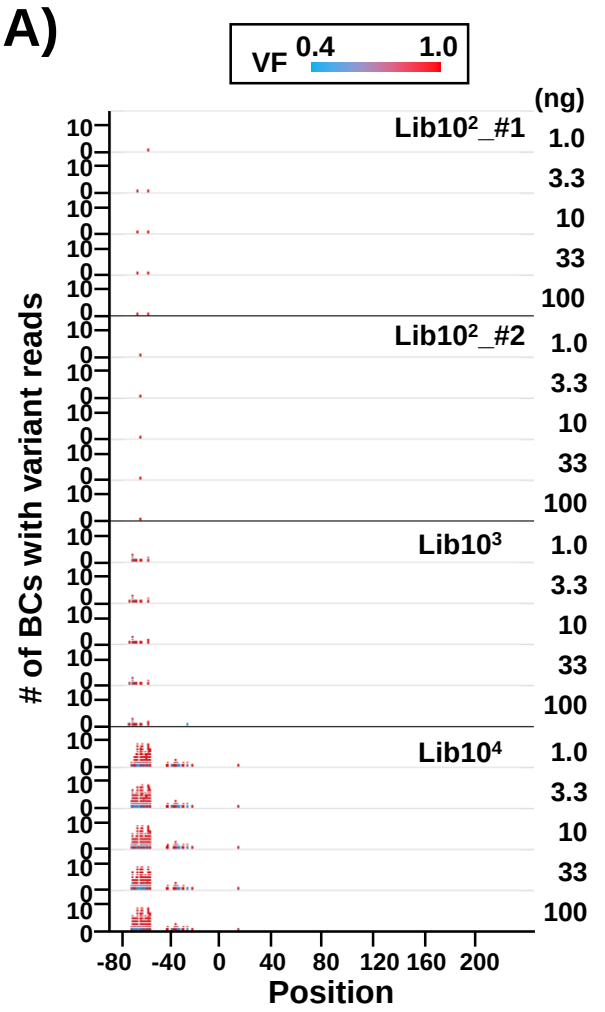
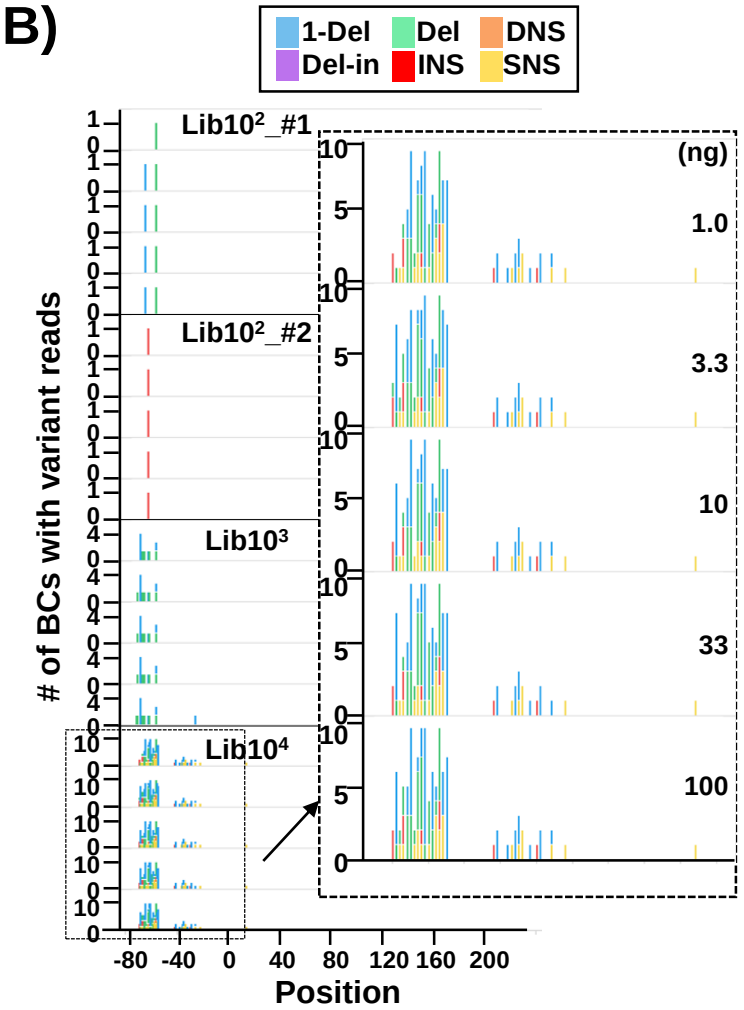
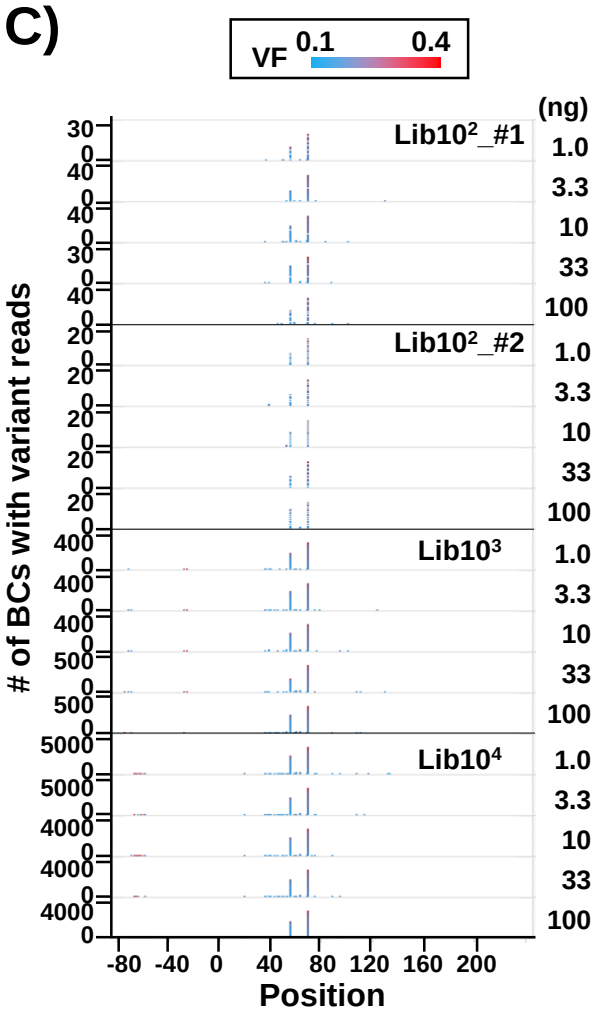
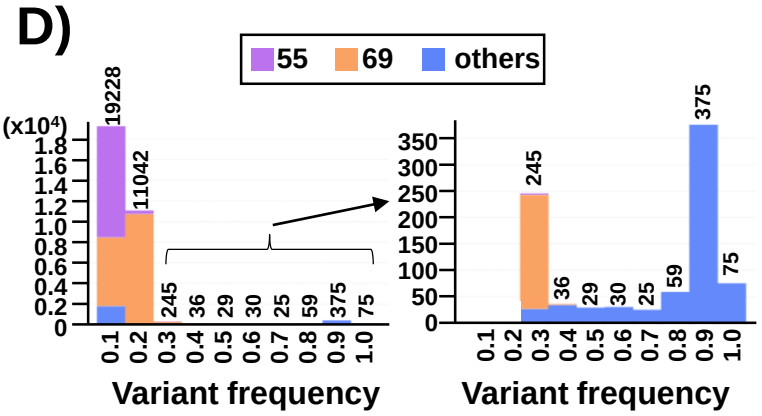

A)
B)
1.0 ng
3.3 ng
10 ng
33 ng
100 ng
Lib10² _#1 (10 M/sample)
(x10²)
# of BCs
118
119
118
117
117
(x10⁴)
# of reads
25276
33072
27316
30201
21147
1.0
3.3
10
33
100
(ng)
Lib10² _#2 (10 M/sample)
(x10²)
# of BCs
97
97
97
97
97
(x10⁴)
# of reads
27735
24302
33560
28881
23372
Lib10³ (100 M/sample)
(x10³)
# of BCs
1424
1425
1425
1426
1421
(x10⁵)
# of reads
325514
305390
327431
289695
242206
Lib10⁴ (750 M/sample)
(x10⁴)
# of BCs
13125
13172
13155
13155
13169
(x10⁶)
# of reads
1829600
2111289
2322965
2309202
2480891
(x10³)
# of reads/BC
BC

Fig. 2









E)

Position	The median of % of reads with the variant call for all samples (VF 0.1-0.4)
35	0.27
38	0.13
45	0.02
50	0.05
52	0.42
55	24.72
58	0.14
60	0.42
63	1.36
68	0.28
69	42.65

Fig. 3

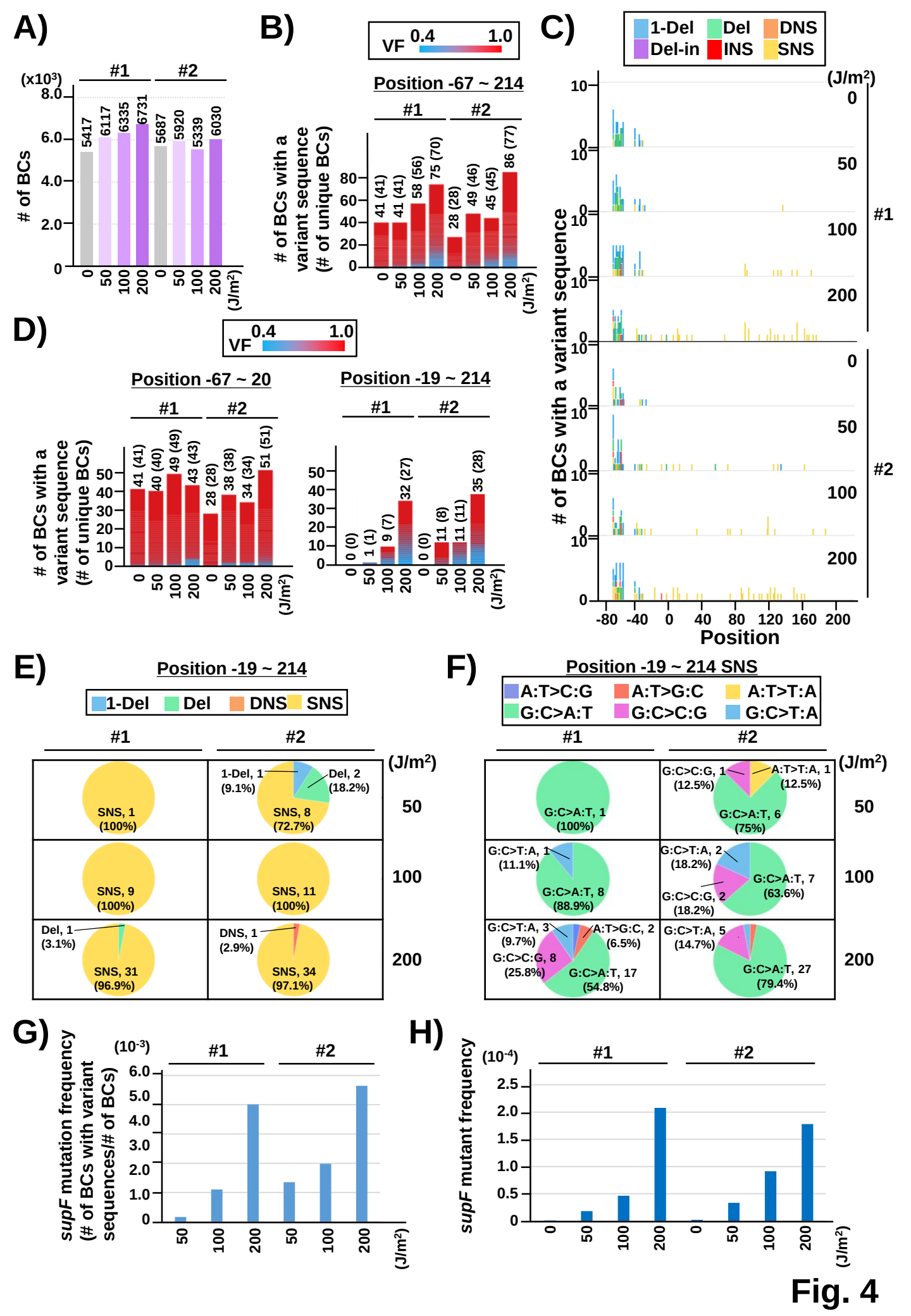

A)
(x10³)
#1
#2
# of BCs
5417
6117
6335
6731
5687
5920
5339
6030
(J/m²)
B)
VF 0.4 1.0
Position -67 ~ 214
# of BCs with a variant sequence (# of unique BCs)
41 (41)
41 (41)
58 (56)
75 (70)
28 (28)
49 (46)
45 (45)
86 (77)
C)
1-Del
Del
DNS
Del-in
INS
SNS
(J/m²)
# of BCs with a variant sequence
Position
D)
VF 0.4 1.0
Position -67 ~ 20
Position -19 ~ 214
41 (41)
40 (40)
49 (49)
43 (43)
28 (28)
38 (38)
34 (34)
51 (51)
0 (0)
1 (1)
9 (7)
32 (27)
0 (0)
11 (8)
11 (11)
35 (28)
E)
Position -19 ~ 214
1-Del
Del
DNS
SNS
SNS, 1 (100%)
1-Del, 1 (9.1%)
Del, 2 (18.2%)
SNS, 8 (72.7%)
SNS, 9 (100%)
SNS, 11 (100%)
Del, 1 (3.1%)
SNS, 31 (96.9%)
DNS, 1 (2.9%)
SNS, 34 (97.1%)
F)
Position -19 ~ 214 SNS
A:T>C:G
A:T>G:C
A:T>T:A
G:C>A:T
G:C>C:G
G:C>T:A
G:C>A:T, 1 (100%)
G:C>C:G, 1 (12.5%)
A:T>T:A, 1 (12.5%)
G:C>A:T, 6 (75%)
G:C>T:A, 1 (11.1%)
G:C>A:T, 8 (88.9%)
G:C>T:A, 2 (18.2%)
G:C>A:T, 7 (63.6%)
G:C>C:G, 2 (18.2%)
G:C>T:A, 3 (9.7%)
A:T>G:C, 2 (6.5%)
G:C>C:G, 8 (25.8%)
G:C>A:T, 17 (54.8%)
G:C>T:A, 5 (14.7%)
G:C>A:T, 27 (79.4%)
G)
(10⁻³)
supF mutation frequency (# of BCs with variant sequences/# of BCs)
H)
(10⁻⁴)
supF mutant frequency

Fig. 4

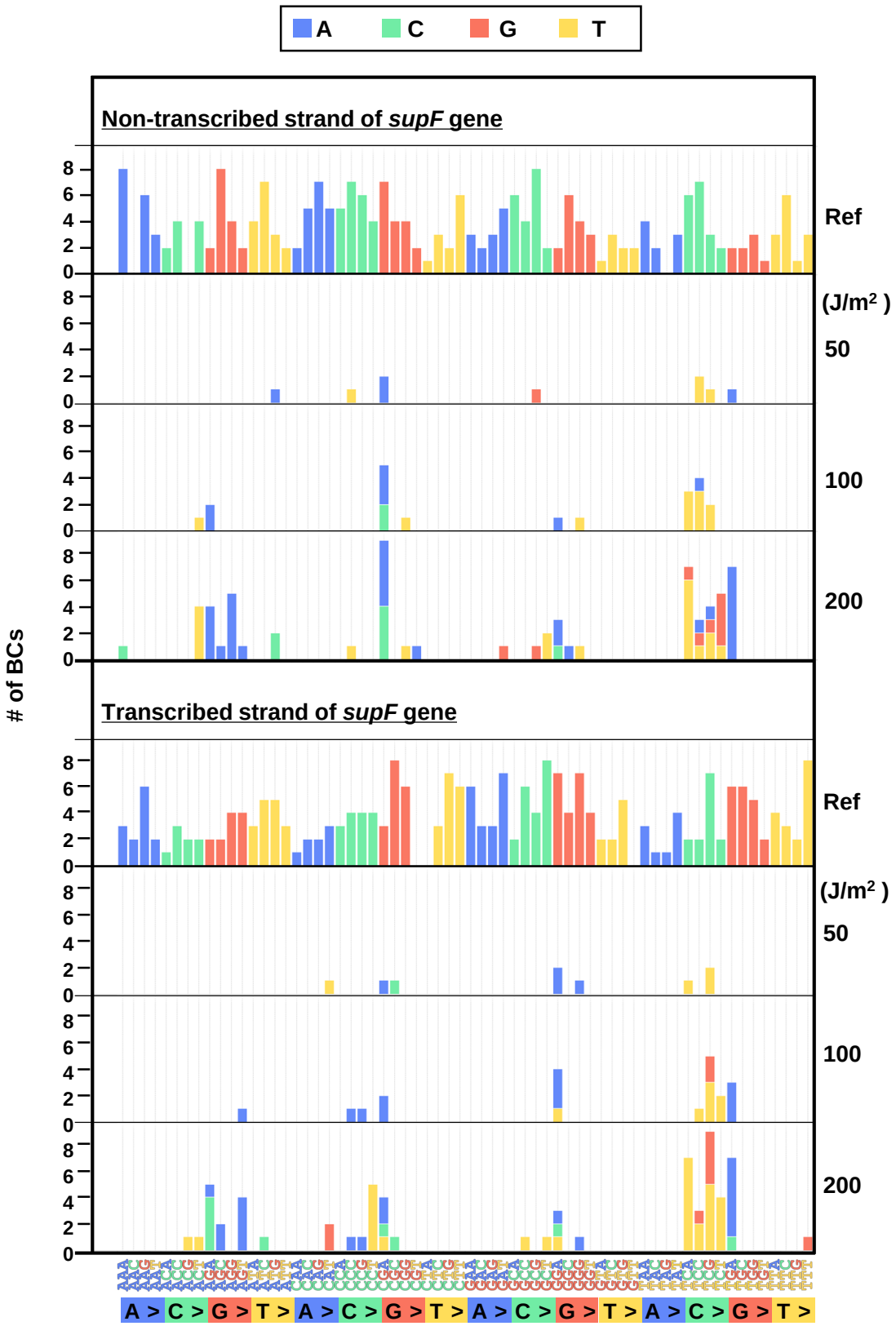

A
C
G
T
Non-transcribed strand of supF gene
Transcribed strand of supF gene
# of BCs
Ref
(J/m2)
50
100
200
A > C > G > T > A > C > G > T > A > C > G > T > A > C > G > T >

Fig. 5

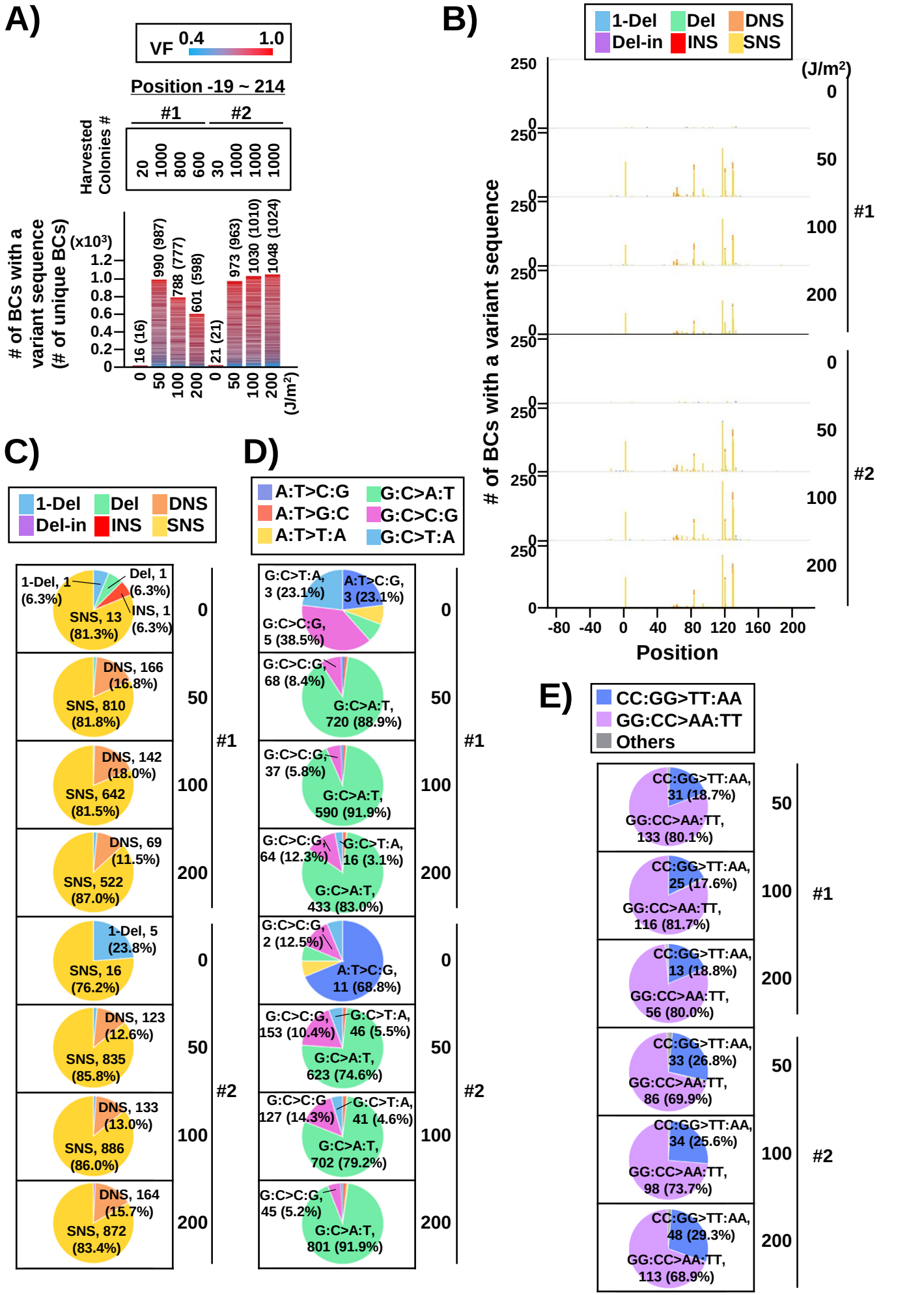

A)
VF 0.4 1.0
Position -19 ~ 214
#1
#2
Harvested Colonies #
20 1000 800 600 30 1000 1000 1000
# of BCs with a variant sequence (# of unique BCs)
(x10³)
1.2 1.0 0.8 0.6 0.4 0.2 0
16 (16)
990 (987)
788 (777)
601 (598)
21 (21)
973 (963)
1030 (1010)
1048 (1024)
0 50 100 200 0 50 100 200
(J/m²)
B)
1-Del Del DNS
Del-in INS SNS
(J/m²)
# of BCs with a variant sequence
250 0
0 50 100 200
#1
#2
-80 -40 0 40 80 120 160 200
Position
C)
1-Del Del DNS
Del-in INS SNS
1-Del, 1 (6.3%)
Del, 1 (6.3%)
INS, 1 (6.3%)
SNS, 13 (81.3%)
DNS, 166 (16.8%)
SNS, 810 (81.8%)
DNS, 142 (18.0%)
SNS, 642 (81.5%)
DNS, 69 (11.5%)
SNS, 522 (87.0%)
1-Del, 5 (23.8%)
SNS, 16 (76.2%)
DNS, 123 (12.6%)
SNS, 835 (85.8%)
DNS, 133 (13.0%)
SNS, 886 (86.0%)
DNS, 164 (15.7%)
SNS, 872 (83.4%)
0 50 100 200
#1
0 50 100 200
#2
D)
A:T>C:G A:T>G:C A:T>T:A
G:C>A:T G:C>C:G G:C>T:A
G:C>T:A, 3 (23.1%)
A:T>C:G, 3 (23.1%)
G:C>C:G, 5 (38.5%)
G:C>C:G, 68 (8.4%)
G:C>A:T, 720 (88.9%)
G:C>C:G, 37 (5.8%)
G:C>A:T, 590 (91.9%)
G:C>C:G, 64 (12.3%)
G:C>T:A, 16 (3.1%)
G:C>A:T, 433 (83.0%)
G:C>C:G, 2 (12.5%)
A:T>C:G, 11 (68.8%)
G:C>C:G, 153 (10.4%)
G:C>T:A, 46 (5.5%)
G:C>A:T, 623 (74.6%)
G:C>C:G 127 (14.3%)
G:C>T:A, 41 (4.6%)
G:C>A:T, 702 (79.2%)
G:C>C:G, 45 (5.2%)
G:C>A:T, 801 (91.9%)
0 50 100 200
#1
0 50 100 200
#2
E)
CC:GG>TT:AA
GG:CC>AA:TT
Others
CC:GG>TT:AA, 31 (18.7%)
GG:CC>AA:TT, 133 (80.1%)
CC:GG>TT:AA, 25 (17.6%)
GG:CC>AA:TT, 116 (81.7%)
CC:GG>TT:AA, 13 (18.8%)
GG:CC>AA:TT, 56 (80.0%)
CC:GG>TT:AA, 33 (26.8%)
GG:CC>AA:TT, 86 (69.9%)
CC:GG>TT:AA, 34 (25.6%)
GG:CC>AA:TT, 98 (73.7%)
CC:GG>TT:AA, 48 (29.3%)
GG:CC>AA:TT, 113 (68.9%)
50 100 200
#1
50 100 200
#2

Fig. 6

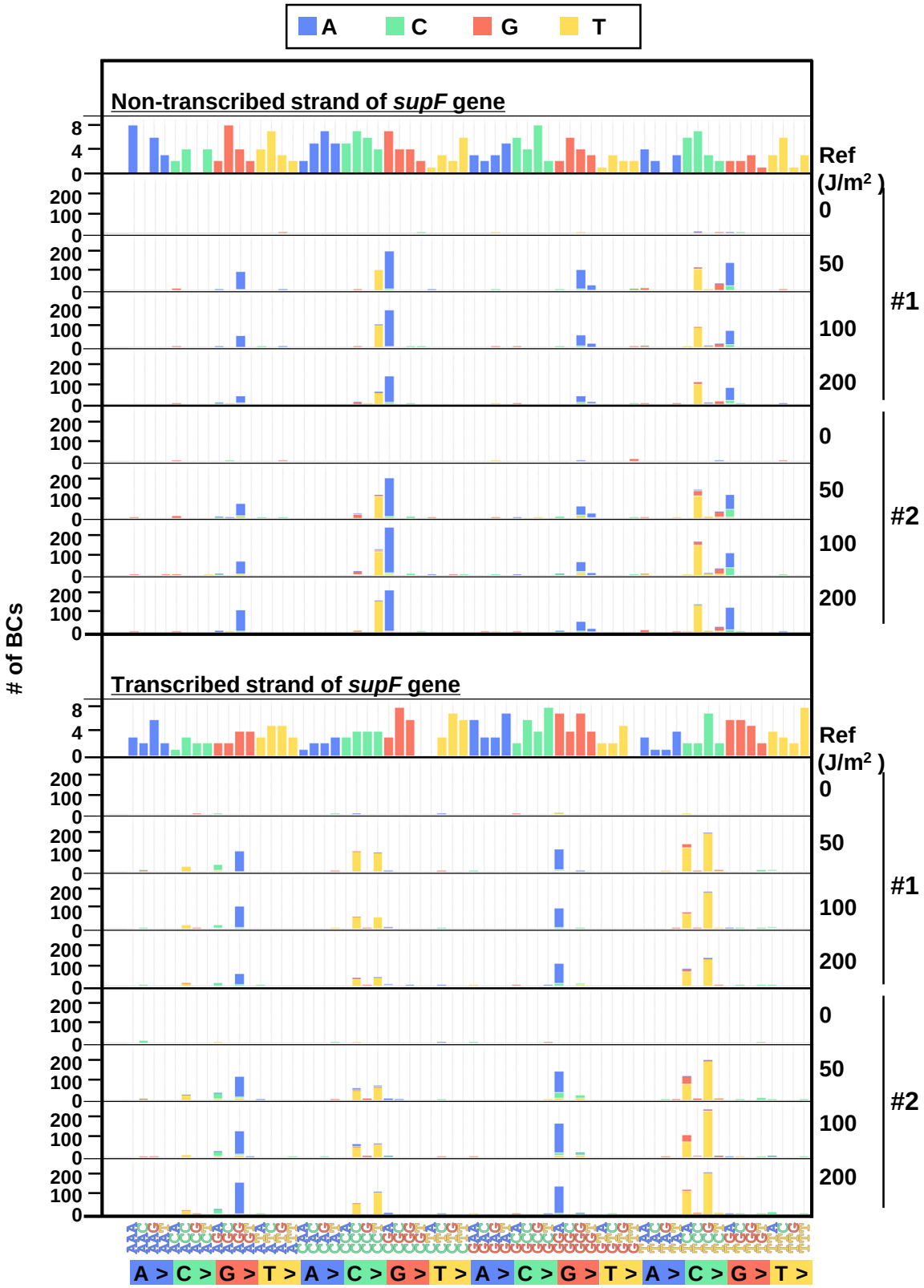

A
C
G
T
Non-transcribed strand of supF gene
Transcribed strand of supF gene
# of BCs
Ref
(J/m2)
0
50
100
200
#1
#2
A > C > G > T > A > C > G > T > A > C > G > T > A > C > G > T >

Fig. 7

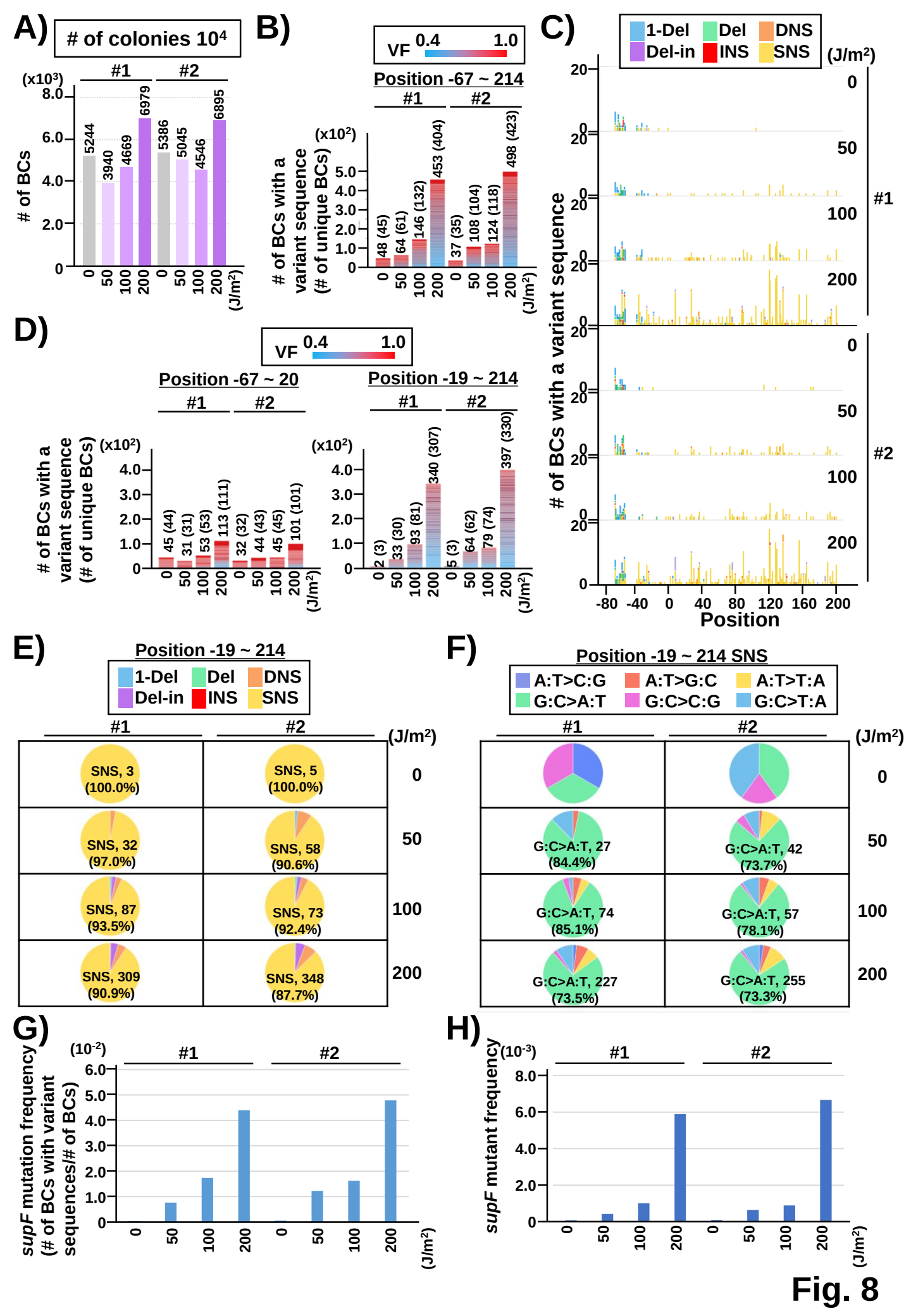

A)
# of colonies 10⁴
(x10³)
#1
#2
# of BCs
5244
3940
4669
6979
5386
5045
4546
6895
(J/m²)
B)
VF 0.4 1.0
Position -67 ~ 214
(x10²)
# of BCs with a variant sequence (# of unique BCs)
48 (45)
64 (61)
146 (132)
453 (404)
37 (35)
108 (104)
124 (118)
498 (423)
C)
1-Del
Del
DNS
Del-in
INS
SNS
# of BCs with a variant sequence
Position
D)
Position -67 ~ 20
Position -19 ~ 214
45 (44)
31 (31)
53 (53)
113 (111)
32 (32)
44 (43)
45 (45)
101 (101)
2 (3)
33 (30)
93 (81)
340 (307)
5 (3)
64 (62)
79 (74)
397 (330)
E)
Position -19 ~ 214
SNS, 3 (100.0%)
SNS, 5 (100.0%)
SNS, 32 (97.0%)
SNS, 58 (90.6%)
SNS, 87 (93.5%)
SNS, 73 (92.4%)
SNS, 309 (90.9%)
SNS, 348 (87.7%)
F)
Position -19 ~ 214 SNS
A:T>C:G
A:T>G:C
A:T>T:A
G:C>A:T
G:C>C:G
G:C>T:A
G:C>A:T, 27 (84.4%)
G:C>A:T, 42 (73.7%)
G:C>A:T, 74 (85.1%)
G:C>A:T, 57 (78.1%)
G:C>A:T, 227 (73.5%)
G:C>A:T, 255 (73.3%)
G)
(10⁻²)
supF mutation frequency (# of BCs with variant sequences/# of BCs)
H)
(10⁻³)
supF mutant frequency

Fig. 8

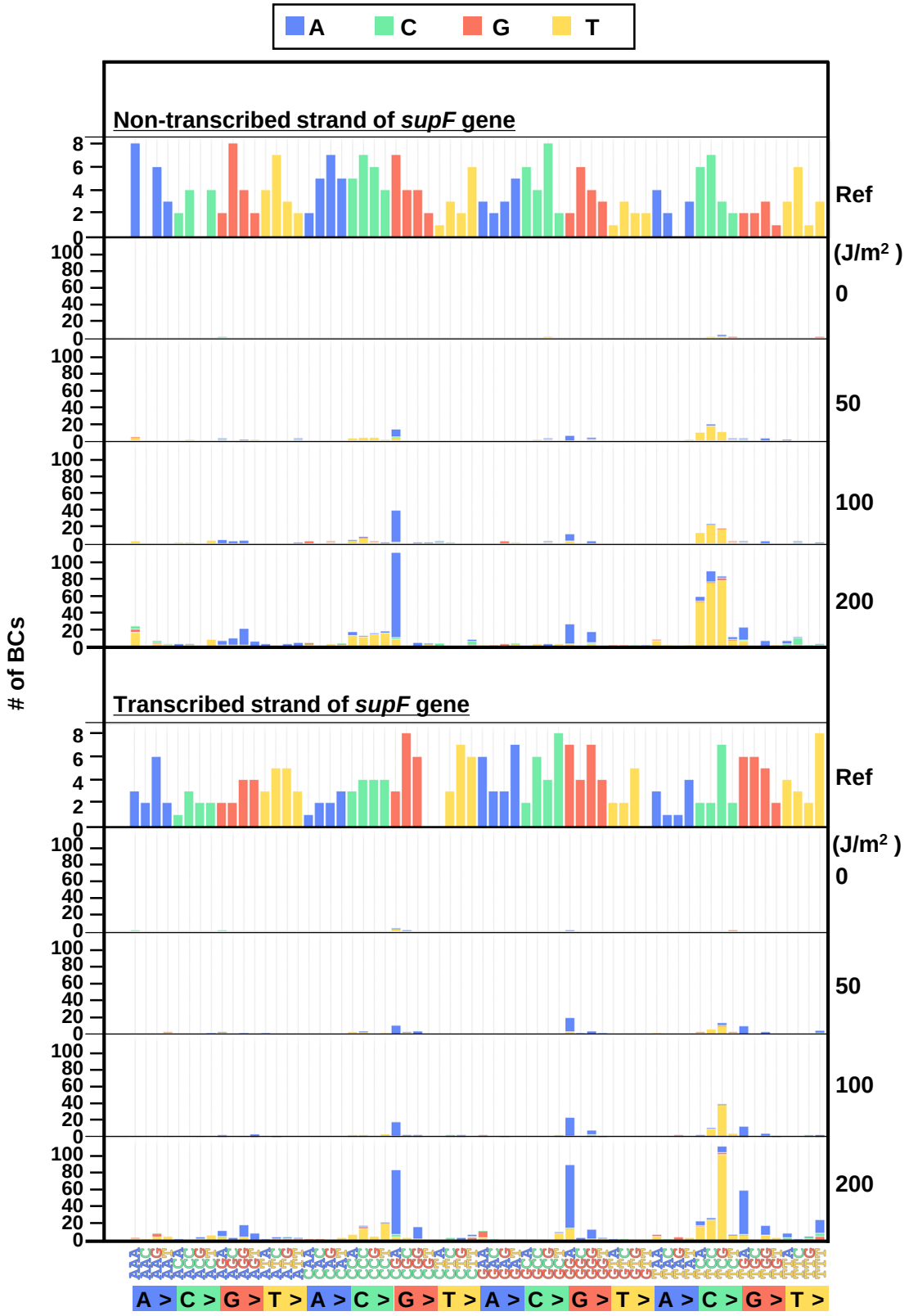

A
C
G
T
Non-transcribed strand of supF gene
Transcribed strand of supF gene
# of BCs
Ref
(J/m2)
0
50
100
200
A > C > G > T > A > C > G > T > A > C > G > T > A > C > G > T >

Fig. 9

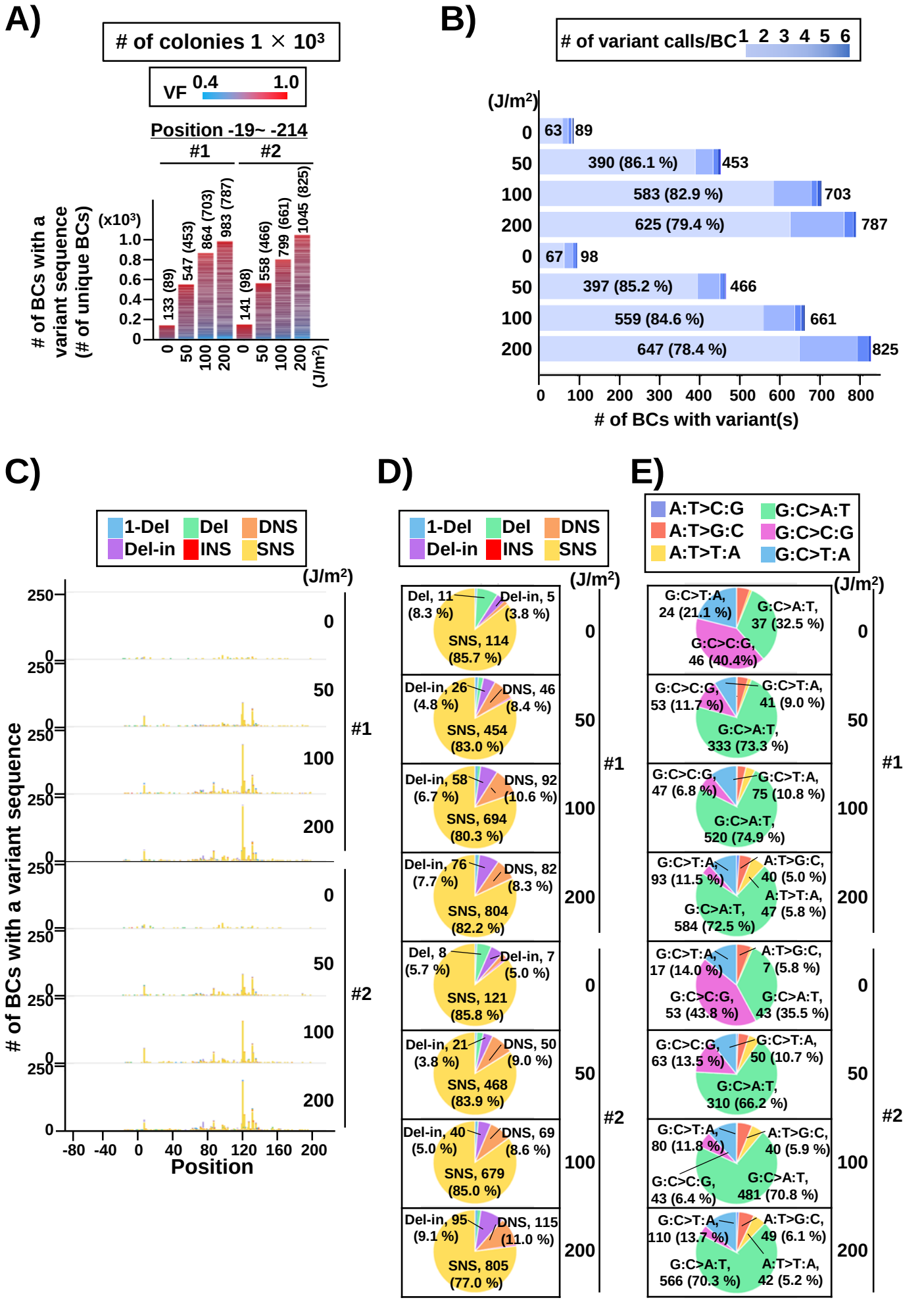

A)
# of colonies 1 × 10³
VF 0.4 1.0
Position -19~ -214
#1
#2
# of BCs with a variant sequence (# of unique BCs)
(x10³)
133 (89)
547 (453)
864 (703)
983 (787)
141 (98)
558 (466)
799 (661)
1045 (825)
(J/m²)
B)
# of variant calls/BC 1 2 3 4 5 6
(J/m²)
63 89
390 (86.1 %) 453
583 (82.9 %) 703
625 (79.4 %) 787
67 98
397 (85.2 %) 466
559 (84.6 %) 661
647 (78.4 %) 825
# of BCs with variant(s)
C)
1-Del Del DNS
Del-in INS SNS
# of BCs with a variant sequence
Position
D)
1-Del Del DNS
Del-in INS SNS
Del, 11 (8.3 %)
Del-in, 5 (3.8 %)
SNS, 114 (85.7 %)
Del-in, 26 (4.8 %)
DNS, 46 (8.4 %)
SNS, 454 (83.0 %)
Del-in, 58 (6.7 %)
DNS, 92 (10.6 %)
SNS, 694 (80.3 %)
Del-in, 76 (7.7 %)
DNS, 82 (8.3 %)
SNS, 804 (82.2 %)
Del, 8 (5.7 %)
Del-in, 7 (5.0 %)
SNS, 121 (85.8 %)
Del-in, 21 (3.8 %)
DNS, 50 (9.0 %)
SNS, 468 (83.9 %)
Del-in, 40 (5.0 %)
DNS, 69 (8.6 %)
SNS, 679 (85.0 %)
Del-in, 95 (9.1 %)
DNS, 115 (11.0 %)
SNS, 805 (77.0 %)
E)
A:T>C:G G:C>A:T
A:T>G:C G:C>C:G
A:T>T:A G:C>T:A
G:C>T:A, 24 (21.1 %)
G:C>A:T, 37 (32.5 %)
G:C>C:G, 46 (40.4%)
G:C>C:G, 53 (11.7 %)
G:C>T:A, 41 (9.0 %)
G:C>A:T, 333 (73.3 %)
G:C>C:G, 47 (6.8 %)
G:C>T:A, 75 (10.8 %)
G:C>A:T, 520 (74.9 %)
G:C>T:A, 93 (11.5 %)
A:T>G:C, 40 (5.0 %)
A:T>T:A, 47 (5.8 %)
G:C>A:T, 584 (72.5 %)
G:C>T:A, 17 (14.0 %)
A:T>G:C, 7 (5.8 %)
G:C>C:G, 53 (43.8 %)
G:C>A:T, 43 (35.5 %)
G:C>C:G, 63 (13.5 %)
G:C>T:A, 50 (10.7 %)
G:C>A:T, 310 (66.2 %)
G:C>T:A, 80 (11.8 %)
A:T>G:C, 40 (5.9 %)
G:C>C:G, 43 (6.4 %)
G:C>A:T, 481 (70.8 %)
G:C>T:A, 110 (13.7 %)
A:T>G:C, 49 (6.1 %)
G:C>A:T, 566 (70.3 %)
A:T>T:A, 42 (5.2 %)

Fig. 10

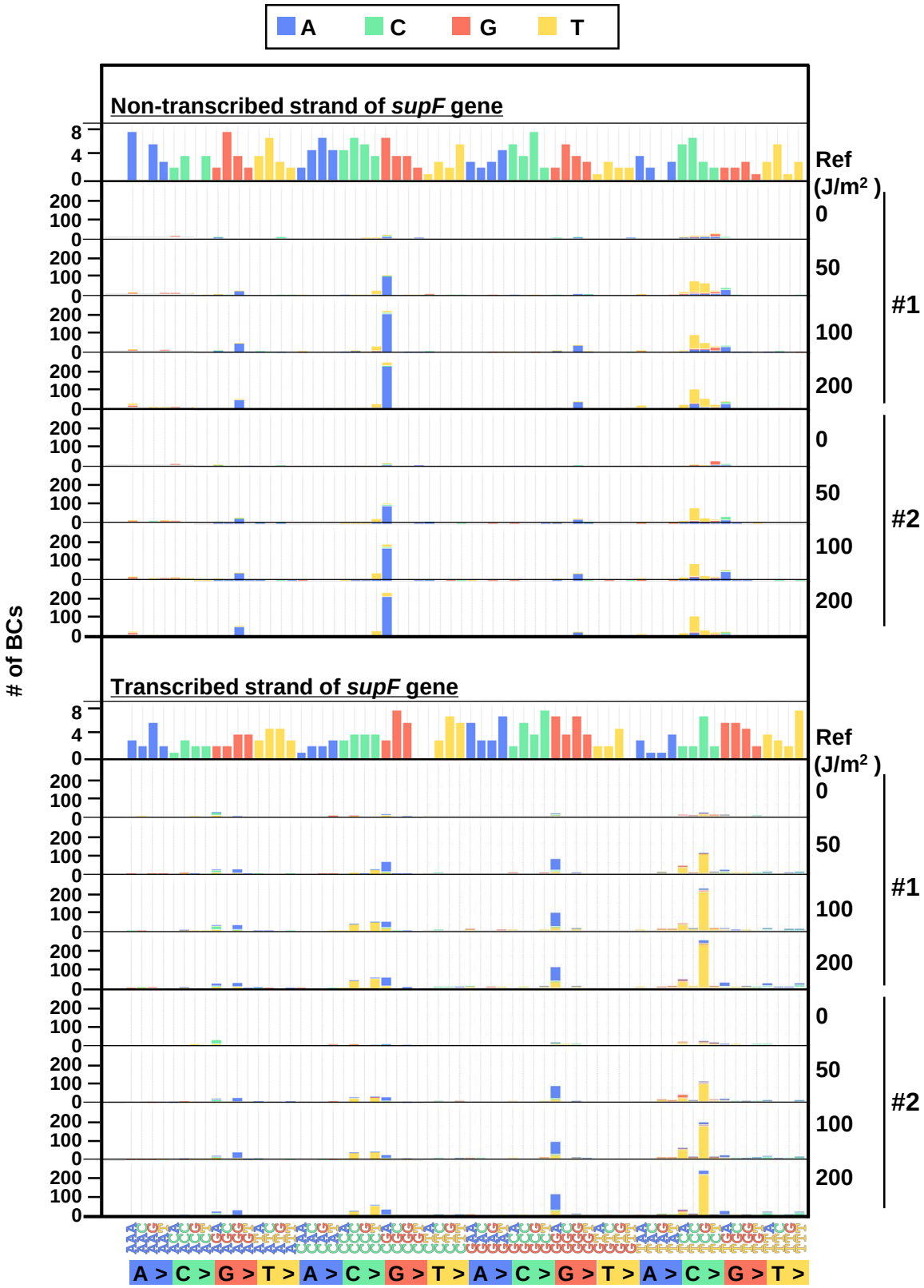

A
C
G
T
Non-transcribed strand of supF gene
Transcribed strand of supF gene
# of BCs
Ref
(J/m2)
0
50
100
200
#1
#2
A > C > G > T > A > C > G > T > A > C > G > T > A > C > G > T >

Fig. 11

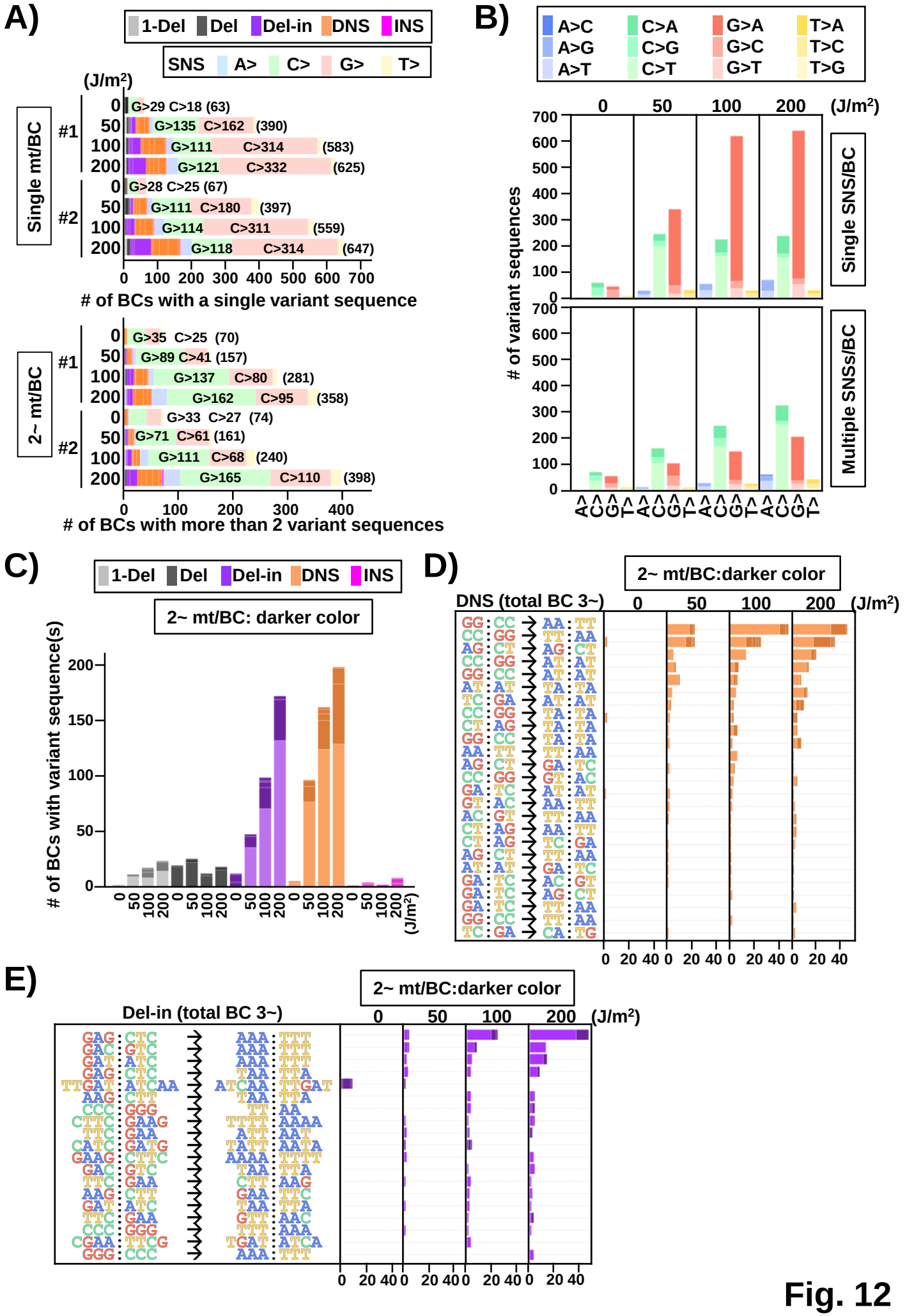

A)
1-Del
Del
Del-in
DNS
INS
SNS
A>
C>
G>
T>
(J/m²)
Single mt/BC
#1
#2
0
50
100
200
G>29 C>18 (63)
G>135 C>162 (390)
G>111 C>314 (583)
G>121 C>332 (625)
G>28 C>25 (67)
G>111 C>180 (397)
G>114 C>311 (559)
G>118 C>314 (647)
0 100 200 300 400 500 600 700
# of BCs with a single variant sequence
2~ mt/BC
G>35 C>25 (70)
G>89 C>41 (157)
G>137 C>80 (281)
G>162 C>95 (358)
G>33 C>27 (74)
G>71 C>61 (161)
G>111 C>68 (240)
G>165 C>110 (398)
0 100 200 300 400
# of BCs with more than 2 variant sequences
B)
A>C
A>G
A>T
C>A
C>G
C>T
G>A
G>C
G>T
T>A
T>C
T>G
0 50 100 200 (J/m²)
# of variant sequences
Single SNS/BC
Multiple SNSs/BC
C)
1-Del
Del
Del-in
DNS
INS
2~ mt/BC: darker color
# of BCs with variant sequence(s)
0 50 100 200 (J/m²)
D)
2~ mt/BC:darker color
DNS (total BC 3~)
0 50 100 200 (J/m²)
0 20 40
E)
2~ mt/BC:darker color
Del-in (total BC 3~)
0 50 100 200 (J/m²)
0 20 40

Fig. 12

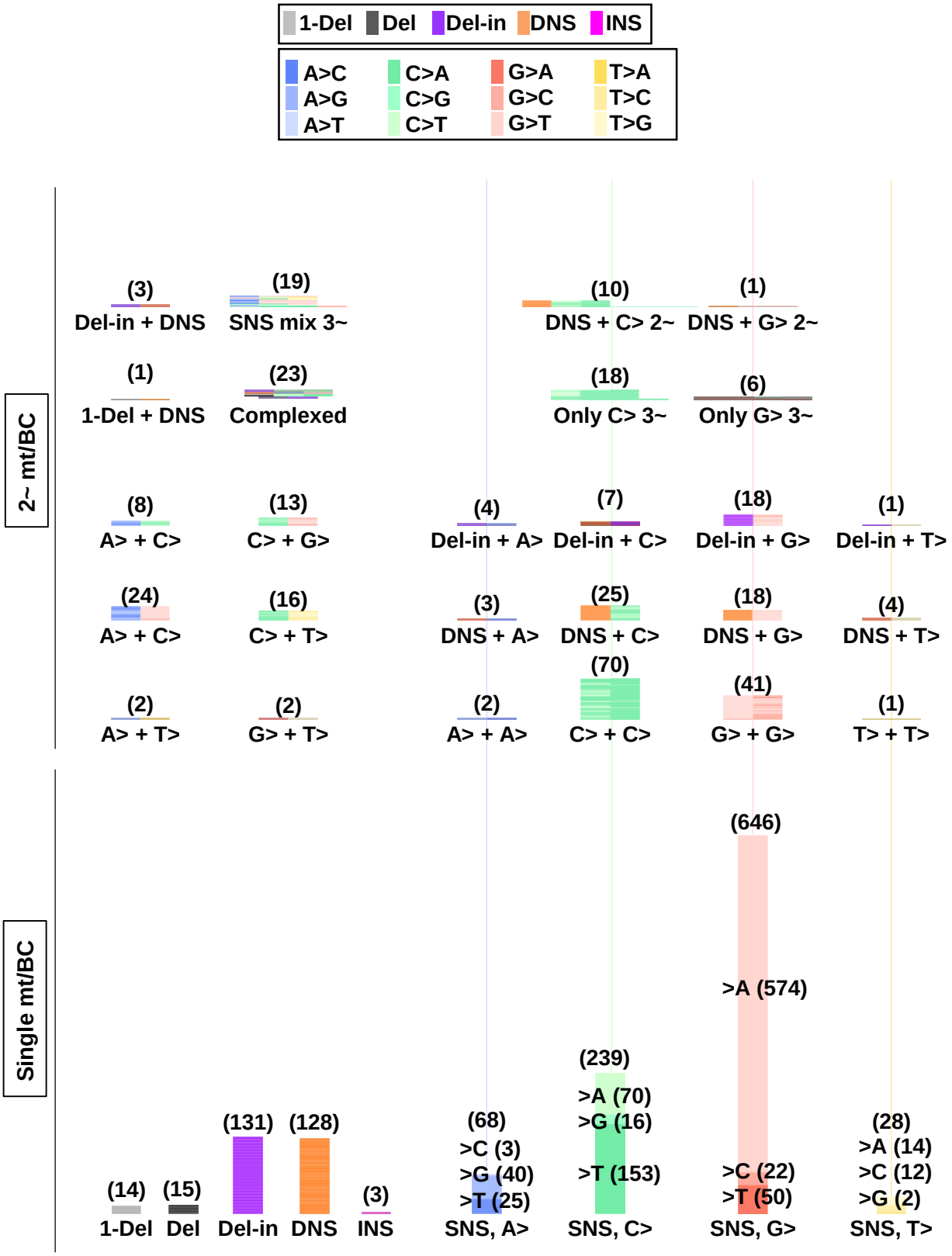

1-Del
Del
Del-in
DNS
INS
A>C
A>G
A>T
C>A
C>G
C>T
G>A
G>C
G>T
T>A
T>C
T>G
2~ mt/BC
(3)
Del-in + DNS
(19)
SNS mix 3~
(10)
DNS + C> 2~
(1)
DNS + G> 2~
(1)
1-Del + DNS
(23)
Complexed
(18)
Only C> 3~
(6)
Only G> 3~
(8)
A> + C>
(13)
C> + G>
(4)
Del-in + A>
(7)
Del-in + C>
(18)
Del-in + G>
(1)
Del-in + T>
(24)
A> + C>
(16)
C> + T>
(3)
DNS + A>
(25)
DNS + C>
(18)
DNS + G>
(4)
DNS + T>
(2)
A> + T>
(2)
G> + T>
(2)
A> + A>
(70)
C> + C>
(41)
G> + G>
(1)
T> + T>
Single mt/BC
(646)
>A (574)
(239)
>A (70)
>G (16)
>T (153)
(68)
>C (3)
>G (40)
>T (25)
(28)
>A (14)
>C (12)
>G (2)
>C (22)
>T (50)
(14)
1-Del
(15)
Del
(131)
Del-in
(128)
DNS
(3)
INS
SNS, A>
SNS, C>
SNS, G>
SNS, T>

Fig. 13

A)

```
-19                                                  CACTTTACA GCGGCGCGTC -1

  1 ATTTGATATG ATGCGCCCCG CTTCCCGATA AGGGAGCAGG CCAGTAAAAG 50
       ❹                     ⑦

 51 CATTACCTGT GGTGGGGTTC CCGAGCGGCC AAAGGGAGCA GACTCTAAAT 100
                                                ⑧
                                                ❼        ❾

101 CTGCCGTCAT CGACTTCGAA GGTTCGAATC CTTCCCCCAC CACCATCACT 150
                    ①②        ④③        ④
                    ❻❶   ❸    ❿   ❷❺   ❽

151 TTCAAAAGTC CGAAAAGATC CGGCGGCGCA GCACCTCATC CTCAGCGATA 200
     ⑥          ⑨                                      ⑩

201 TGGCAGCTTA TAAT                                          214
```

```
A:T(58,*7) C:G(70,*14) G:C(56,*18) T:A(49,*8)

TC:GA(18,*6) GA:TC(13,*6) GG:CC(15,*6) CC:GG(20,*6) TT:AA(13,*4) AA:TT(17,*6)

TCA:TGA(6, 0) TCC:GGA(7,*2) TCG:CGA(3,*2) TCT:AGA(2,*2)
TGA:TCA(2,*1) CGA:TCG(7,*3) AGA:TCT(2,*1) GGA:TCC(2,*1)

ATC:GAT(6,*2) CTC:GAG(3,*1) GTC:GAC(3, 0) TTC:GAA(6,*3)
GAT:ATC(5,*1) GAG:CTC(3,*1) GAC:GTC(2,*1) GAA:TTC(3,*3)
```

B)

```
① 117 TTCGA/TCGAA (55)
② 118 TCGAA/TTCGA (45)
③ 126 TCGAA/TTCGA (44)
④ 125 TTCGA/TCGAA (41)
④ 134 TTCCC/GGGAA (41)
⑥ 153 TTCAA/TTGAA (34)
⑦  24 TTCCC/GGGAA (31)
⑧  86 GGGAG/CTCCC (28)
⑨ 162 CCGAA/TTCGG (25)
⑩ 197 GCGAT/ATCGC (22)

❶ 118 TCGAA/TTCGA (971)(②)
❷ 130 ATCCT/AGGAT (414)(⑫)
❸ 121 AAGGT/ACCTT (261)(㉓)
❹   5 TTGAT/ATCAA (247)(⑭)
❺ 131 TCCTT/AAGGA (186)(㉖)
❻ 117 TTCGA/TCGAA (182)(①)
❼  85 AGGGA/TCCCT (170)(㉔)
❽ 134 TTCCC/GGGAA (102)(⑤)
❾  95 CTCTA/TAGAG  (98)(-)
❿ 126 TCGAA/TTCGA  (94)(③)
```

C)

```
① 134 TTCCCC/GGGGAA →TT/AA(2/4)
②  31 TAAGGG/CCCTTA →GA/TC(2/3)
②  85 AGGGAG/CTCCCT →AA/TT(3/3)
② 130 ATCCTT/AAGGAT →TT/AA(3/3)
② 137 CCCCAC/GTGGGG →AT/AT(2/3)

❶ 130 ATCCTT/AAGGAT →TT/AA(37/50)(②)
❷  84 AAGGGA/TCCCTT →AA/TT(22/27)(-)
❷  85 AGGGAG/CTCCCT →AA/TT(20/27)(②)
❹ 121 GAGGTT/AACCTC →AA/TT(18/22)(-)
❺  70 TTCCCG/CGGGAA →TT/AA(14/21)(-)
```

D)

```
①   5 TTGATAT/ATATCAA →AAA/TTT(6)
②  73 CCGAGCG/CGCTCGG →AAA/TTT(5)
③ 184 CACCTCAT/ATGAGGTG →ACTT/AAGT(3)

❶  73 CCGAGCG/CGCTCGG →AAA/TTT(73/84)
❷ 112 TCGACTT/AAGTCGA →AAA/TTT(25/30)
❸   5 TTGATAT/ATATCAA →AAA/TTT(20/27)
```

E)

```
-8    -3     3 5 7    12    17
GGCGCGTCATTTGATATGATGCGCC
***** **** * * **** *****
CCGCGTAGTATAGTTTACTGCGCGG
```

→

```
-8    -3     3 5 7    12    17
GGCGCATCATATCAAATGACGCGCC
*************************
CCGCGTAGTATAGTTTACTGCGCGG
```

Fig. 14