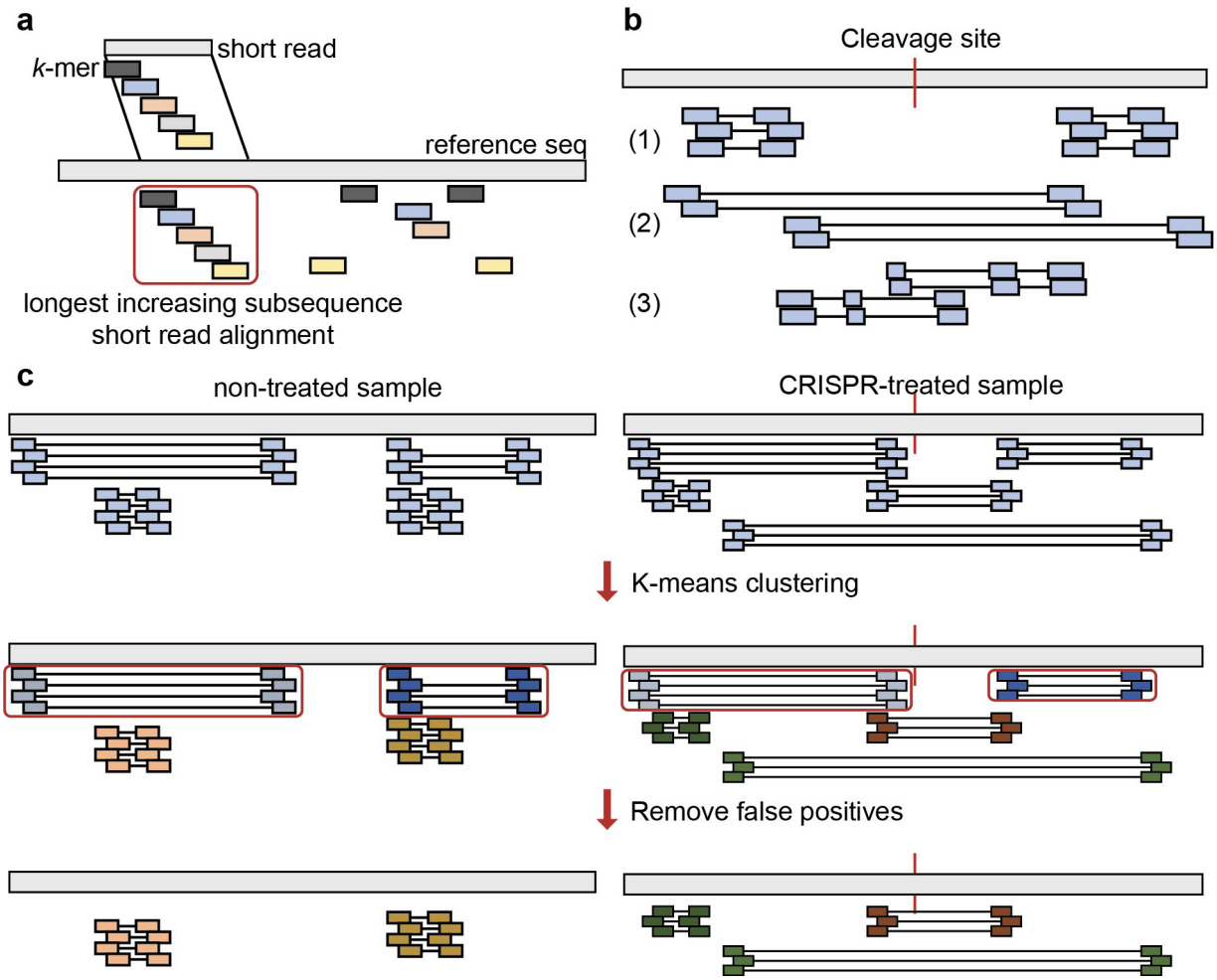
nature biomedical engineering

Supplementary information

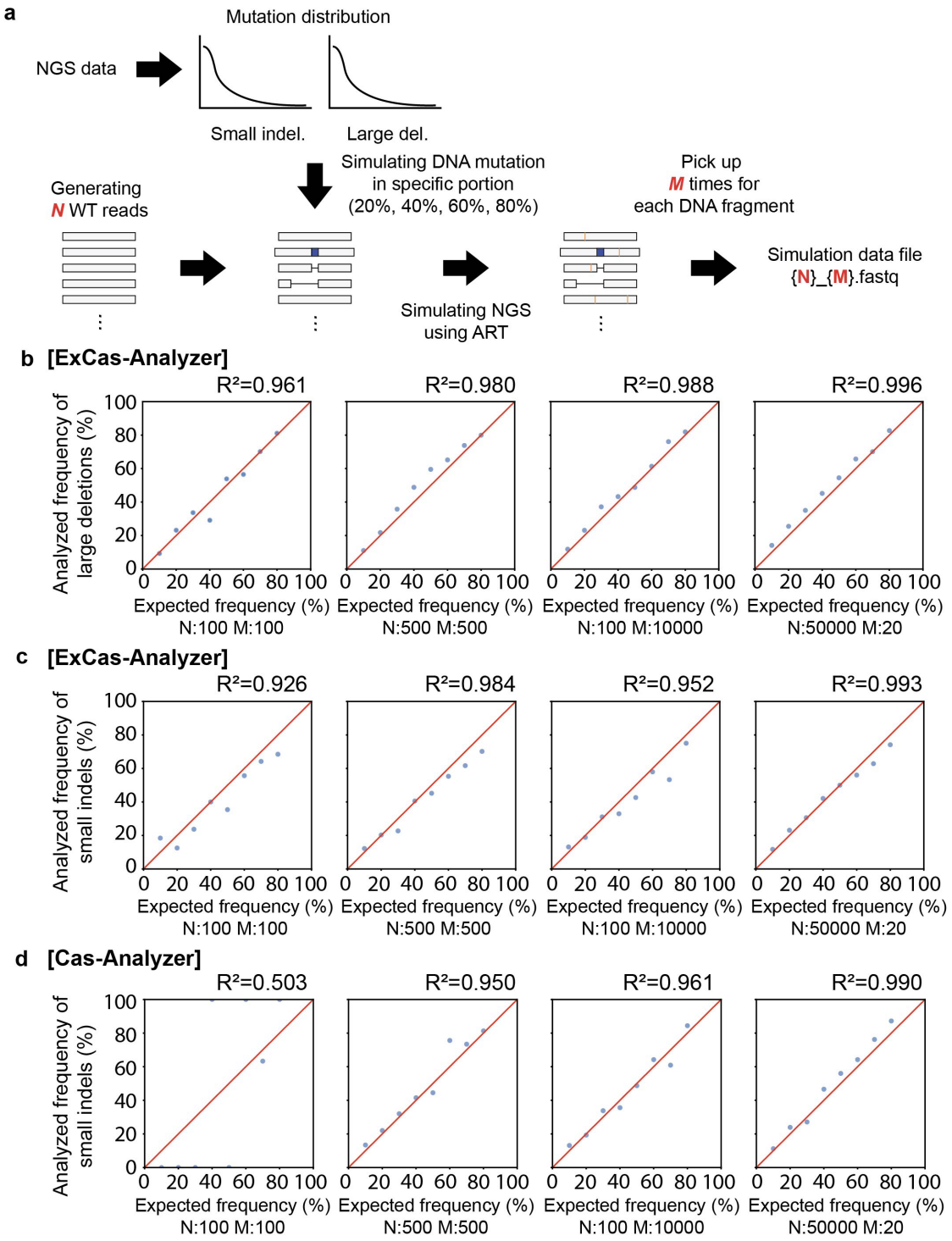
https://doi.org/10.1038/s41551-024-01277-5

Large DNA deletions occur during DNA repair at 20-fold lower frequency for base editors and prime editors than for Cas9 nucleases

In the format provided by the authors and unedited



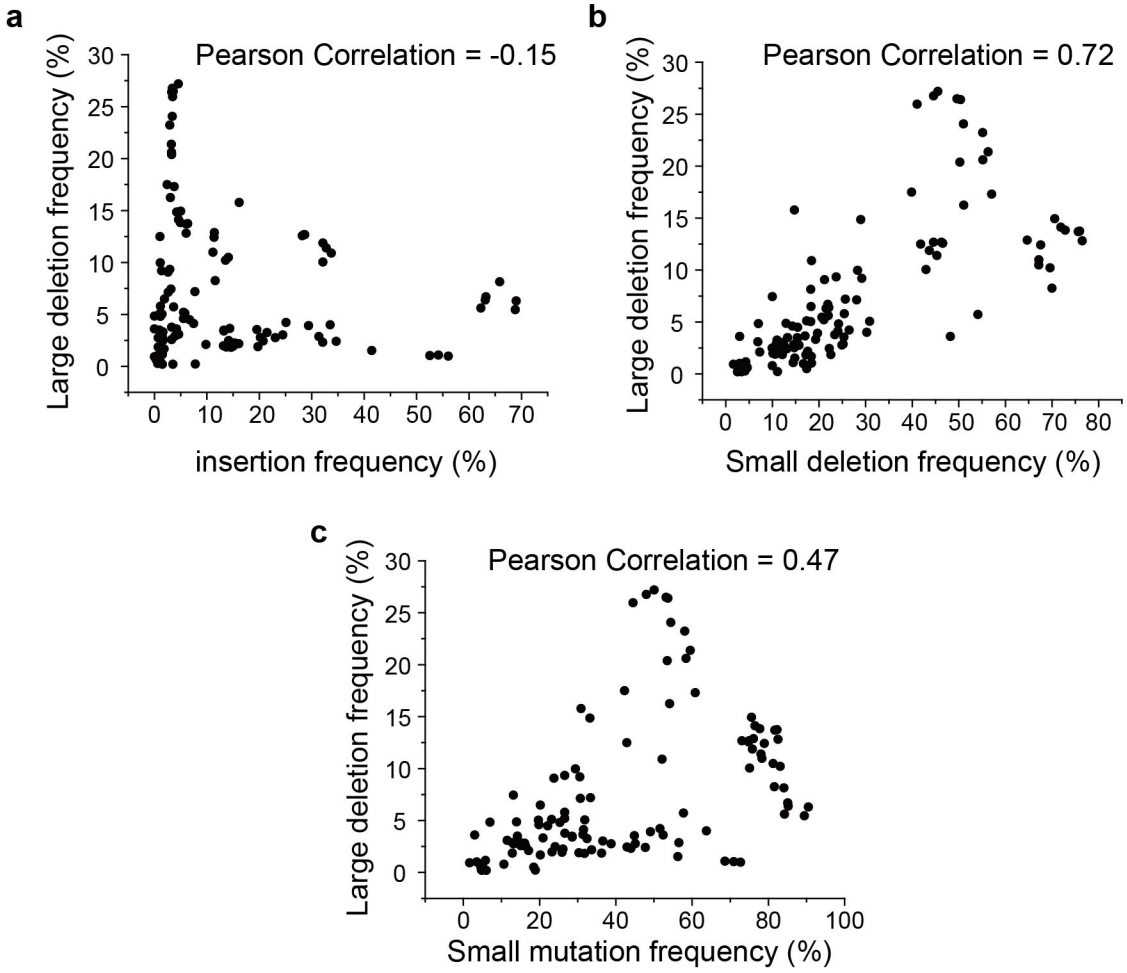
Supplementary Fig. 1 l Visualization of the individual steps within the novel ExCas-Analyzer program. **a,** Schematic representation of short read alignment task. Locate the positions of the *k*-mers within the reference sequence using the *k*-mer hash table derived from the short read. Then, use the longest-incremental sequence (LIS) algorithm to determine the short-read alignment by deriving a LIS of the *k*-mers aligned according to their respective positions within the reference sequence. **b,** Schematic representation of short read classification task. Short read alignment can be categorized into three different cases. In the first case (1), when examining paired-end reads that lack a cleavage site and have a short internal distance, they are referred to as wild type (WT). In the second case (2), when paired-end reads contain a cleavage site and have an extended internal distance, they are classified as candidate variants. The third case (3) is when a single short read is fragmented and mapped, resulting in its classification as a candidate variant. **c,** Schematic representation of removing false positives. The clusters in non-treated sample are considered as false-positive large deletions. To eliminate false-positive large deletions associated with fragment loss during sample preparation or PCR bias, alignment reads are clustered using a K-means clustering algorithm. Then, those clusters that are common to both samples are determined by false-positive large deletions.



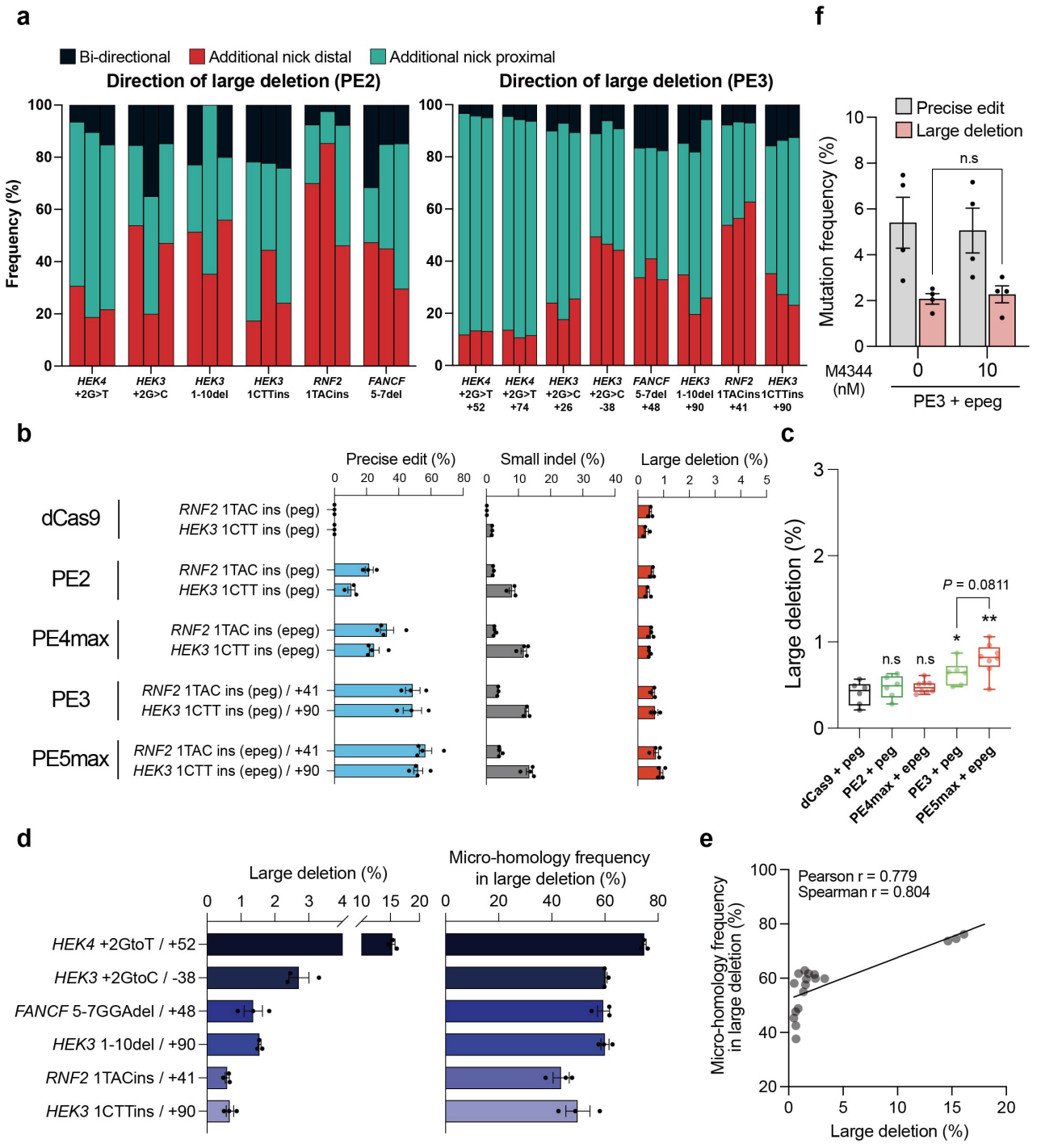
e

Program	Run time (sec)	Maximum usage of Memory (MB)
ExCas-Analyzer	2.93	17
bwa mem	10.63	90
CRISPResso2	843.73	6000

Supplementary Fig. 2 | The simulation for comparing ExCas-Analyzer with other methods. a, Schematic representation of generating simulation datasets. **b-c,** Validation of ExCas-Analyzer using simulation data files: The analysis of large deletions (b), and small indels (c). **d,** Validation of Cas-Analyzer with simulation data files analyzing small indels. The red line is expected correlation. **e,** Comparison of analysis speed and memory usage among ExCas-Analyzer, bwa mem, and CRISPResso2. The simulation file (N:500 and M:500) was employed for this comparison.



Supplementary Fig. 3 l Correlation between large deletions and other mutations. Scatter plots display the relationship between the frequency of large deletions and the frequency of other mutations: **a,** insertions, **b,** small deletions, **c.** small indel mutations.



Supplementary Fig. 4 l The characterization of large deletions mediated by prime editor. a, Direction of large deletion mediated by PE2 or PE3. The patterns are divided into three groups (Bi-directional, additional nick distal and additional nick proximal). **b,** Editing outcomes of original PE system (PE2 protein and pegRNA) and enhanced PE system (PE4max protein and epegRNA) in *RNF2* and *HEK3* (*n*=3, mean ± SEM). For the PE3 and PE5max, the position of additional nicking gRNA is described as number (+41 for *RNF2*, +90 for *HEK3*). **c,** Comparison of large deletion frequency between four PE strategies (PE2 with pegRNA, PE4max with epegRNA, PE3 with pegRNA, and PE5max with epegRNA). **d,** Microhomology frequency in large deletions. The 2-8bp length was determined as microhomology in this analysis. **e,** The correlation between microhomology frequency and large deletion frequency. **f,** The effect of M4344 inhibitor to the large deletion mediated by PE3 approach. *HEK3* +2GtoC epegRNA with -38 ngRNA and *FANCF* 5-7del epegRNA with +48 ngRNA were used with HeLa cell line. **P* < 0.05 and ***P* < 0.005, compared with dCas9 by student's t-test (*P* = 0.2695 for PE2, *P* = 0.1389 for PE4max, *P* = 0.0391 for PE3 and *P* = 0.0071 for PE5max).