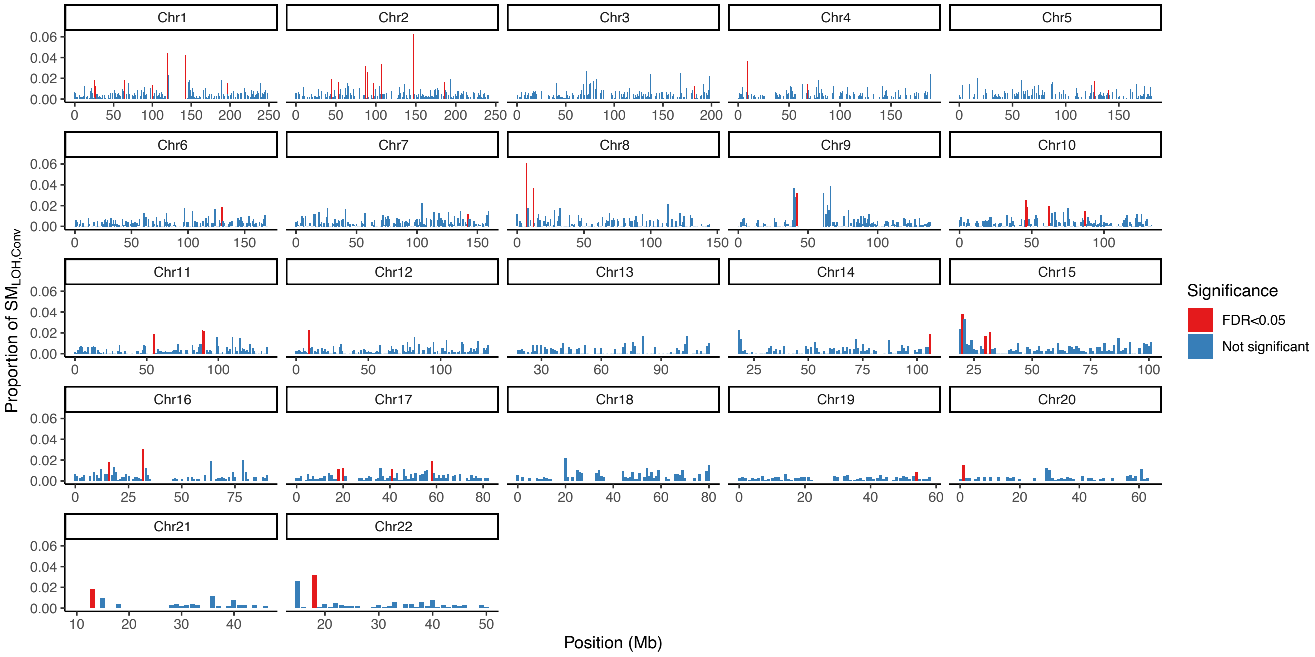
SUPPLEMENTARY INFORMATION: Frequent somatic gene conversion as a mechanism for loss of heterozygosity in tumor suppressor genes

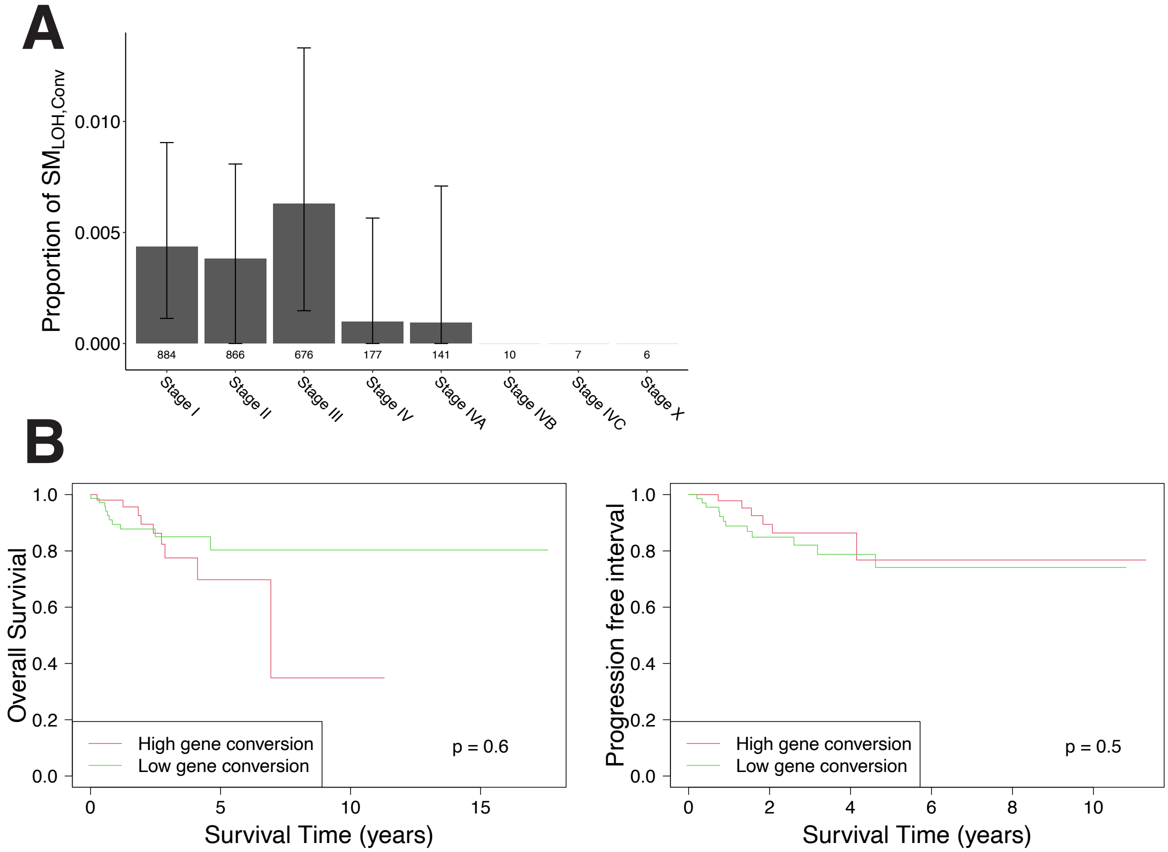
[1], **Kazuki K Takahashi**[1,2,3], **and Hideki Innan**[1,*]

[1]SOKENDAI, The Graduate University for Advanced Studies, Hayama, Kanagawa, 240-0193, Japan.
[2]Laboratory of Plant Genetics, Graduate School of Agriculture, Kyoto University, Kyoto, 606-8502, Japan.
[3]Laboratory of Molecular Medicine, Human Genome Center, The Institute of Medical Science, The University of Tokyo, Tokyo 108–8639, Japan
[*]e-mail: innan_hideki@soken.ac.jp

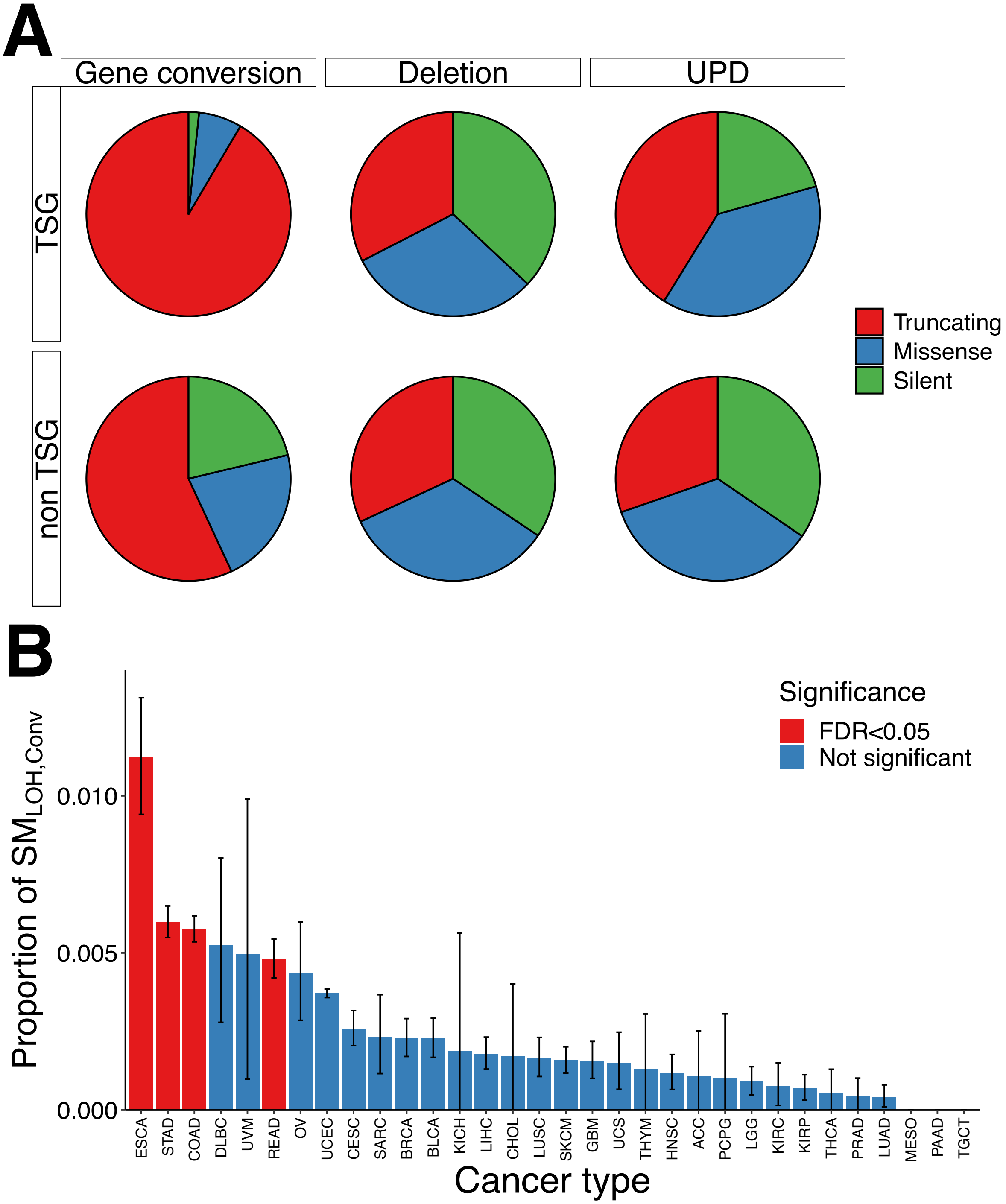
Supplementary Figures



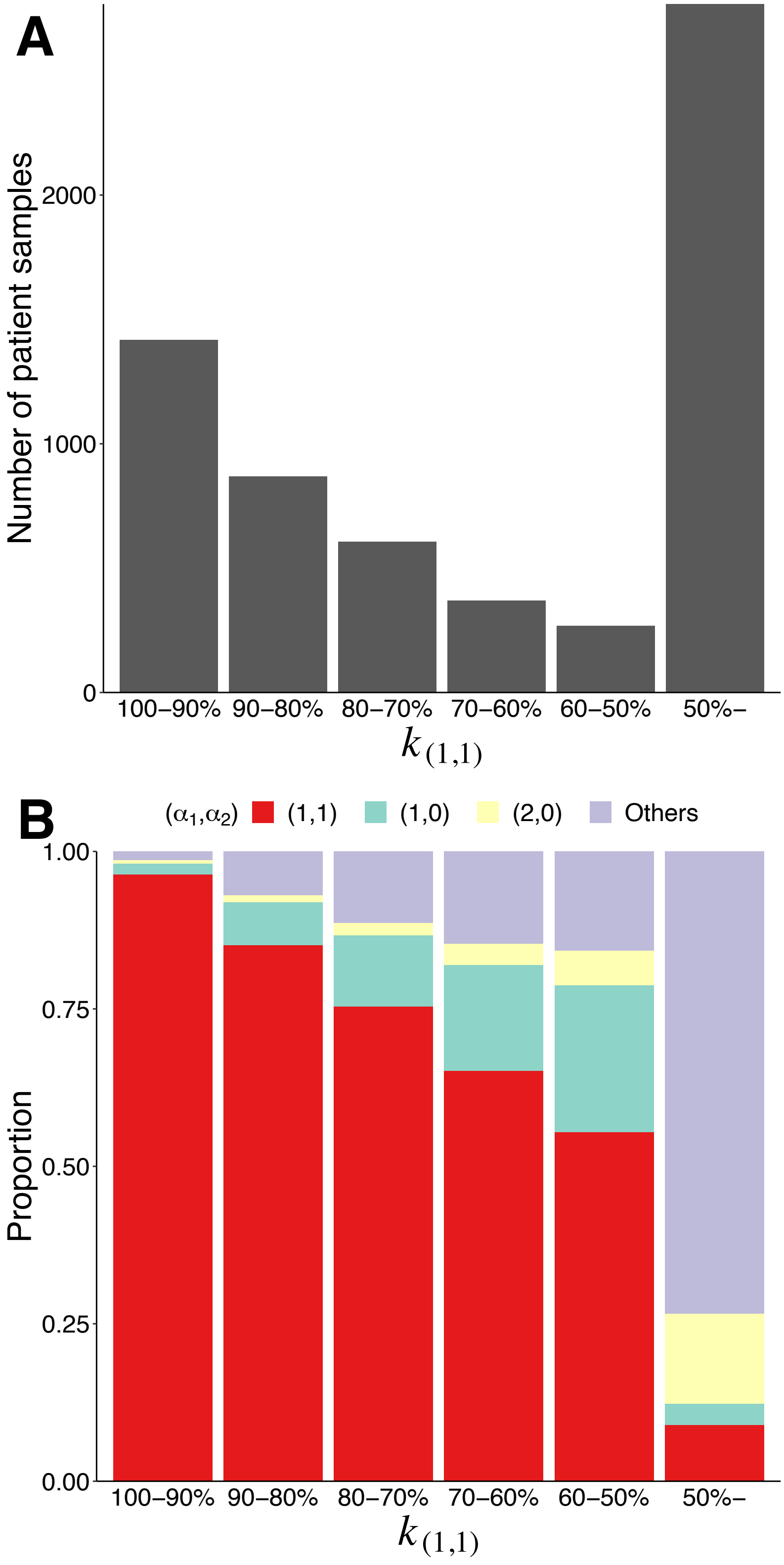
Supplemental Figure S 1. Spatial distribution of the density of $SM_{LOH,Conv}$ across the genome.



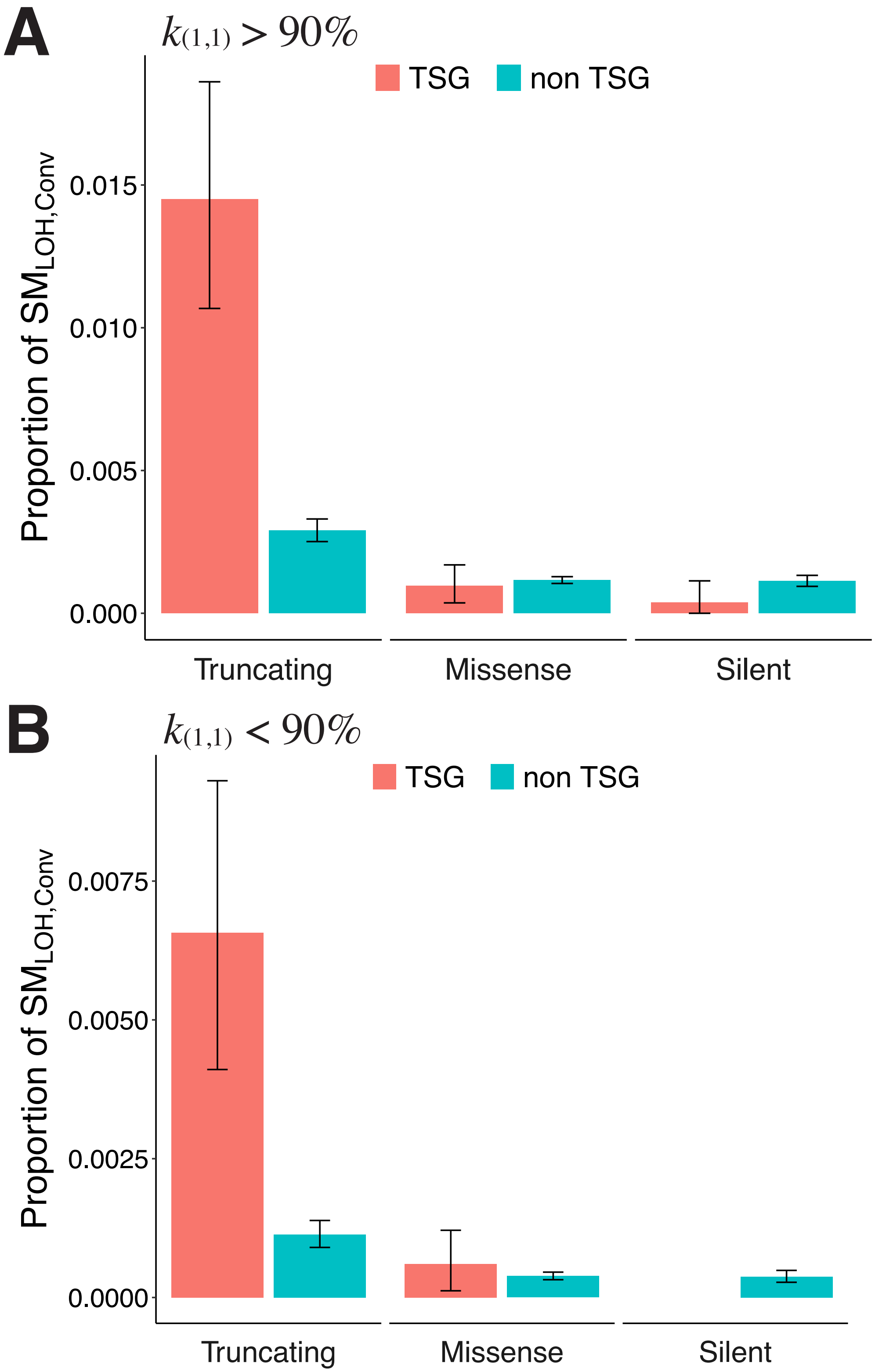
Supplemental Figure S 2. Relationship between the gene conversion rate per patient and clinical factors. (*A*) Proportion of $SM_{LOH,Conv}$ in different cancer stages. The error bars represent the 95% confidence intervals. (*B*) Kaplan-Meier plots of the overall survival and progression free interval. 89 patients with more than 2500 somatic mutations were used, which were classified into two classes with higher ($n = 35$) and lower ($n = 54$) than the average All clinical data were downloaded from Liu et al.[1]



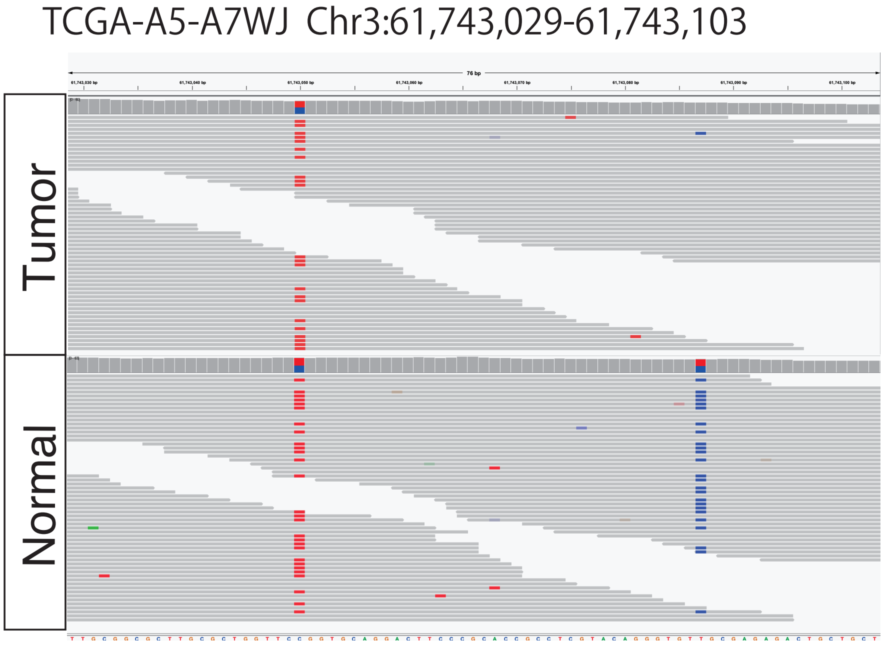
Supplemental Figure S 3. Factors that affect the gene conversion rate. (*A*) Proportion of $SM_{LOH,Conv}$ in each cancer type. (*B*) Proportion of truncating, missense and silent mutations in SM_{LOH} ($SM_{LOH,Conv}$ vs. $SM_{LOH,Del}$ vs. $SM_{LOH,UPD}$). The error bars represent the 95% confidence intervals.



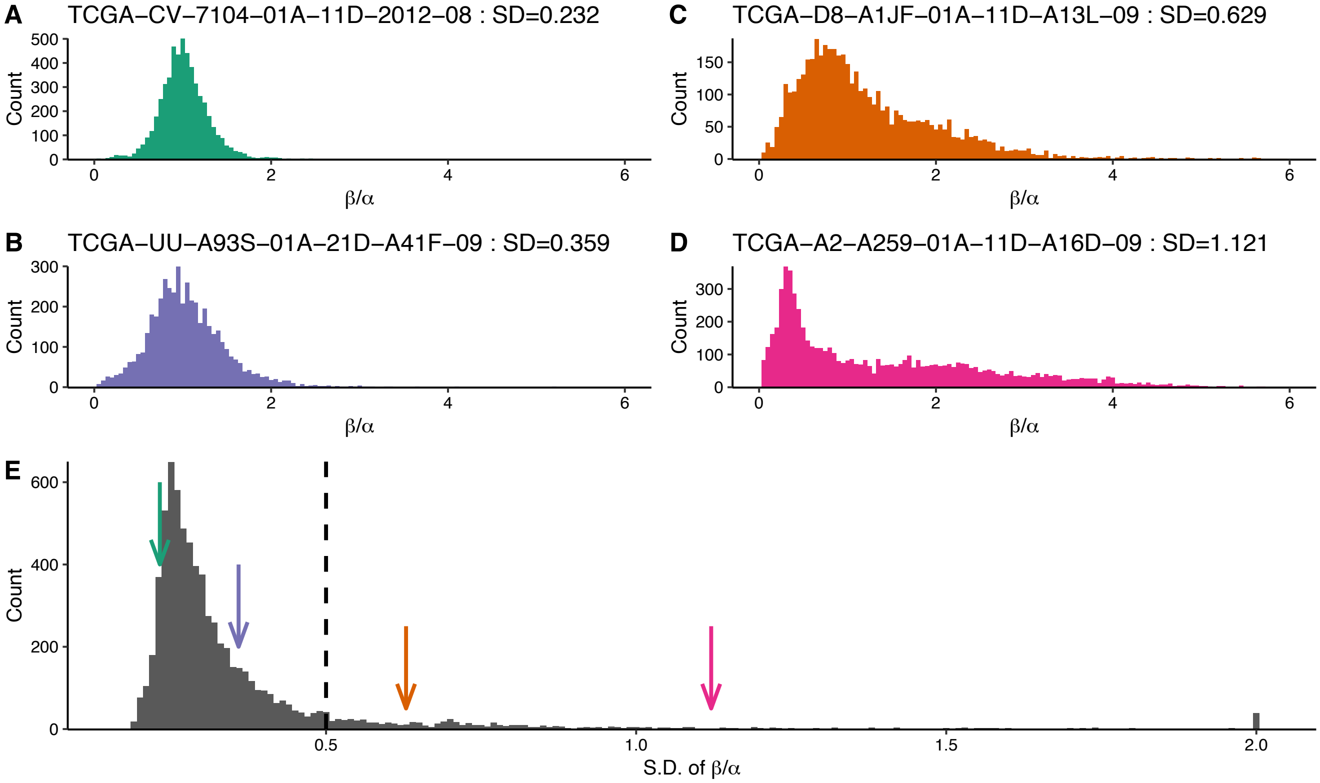
Supplemental Figure S 4. Distribution of $k_{(1,1)}$. (*A*) The density distribution of $k_{(1,1)}$. (*B*) The proportions of the regions of $(\alpha, \beta) = (1,1)$, (1,0), and (2,0) in each class of $k_{(1,1)}$.



Supplemental Figure S 5. Rate of gene conversion in TSG divided by $k_{(1,1)} > 90\%$ (A) and $k_{(1,1)} < 90\%$ (B).



Supplemental Figure S 6. An example of LOH of a germline mutation detected on Chromosome 3 in patient ID: TCGA-A5-A7WJ. The right heterozygote site in the normal cells (in blue) is absent in the tumor cells, indicating this site has experienced somatic gene conversion. There is a tightly linked heterozygote site both in the normal cells and the tumor cells, providing strong evidence for gene conversion. Note that the presence of the red and blue mutations are 100% linked in the normal cells (and also the absence of the two mutations are 100% linked), therefore gene conversion is needed to explain the observed recombinant reads (i.e., red-present and blue-absent reads).

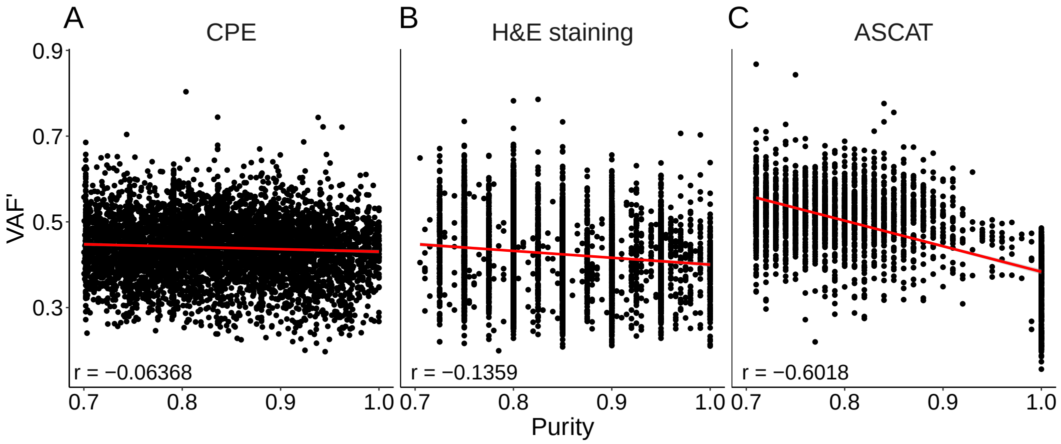


Supplemental Figure S 7. Distributions and SD of β/α. The distributions of β/α for four patient samples (*A*-*D*) and that of SD of β/α (*E*). Two patient samples for a and b represent cases with relatively small SDs, which exhibit normal-like distributions around β/α=1. Two patient samples for c and d represent cases with relatively large SDs, which have wide distributions with long tails. The SD values for the four patient samples are shown in *E* by arrows in the same colors as the distributions in *A*-*D*. In this work, we screened out patient samples with SD>0.5 (dashed line).

Supplementary Note

Purity estimation

For estimates of purity, our analysis mainly relied on the consensus measurement of purity estimations (hereafter, referred to as CPE) estimated by Aran et al.[2]. CPE was obtained by considering various estimates of purity, including an estimate from 'hematoxylin and eosin staining' (H&E staining, provided as 'percent_tumor_nuclei' by TCGA) and an estimate by ASCAT[3]. A caveat is that an estimate of purity could be correlated with VAF' (VAF considering purity. See the Methods). As definition (i.e., VAF'=VAF/purity), VAF' would be small if purity is overestimated, and vice versa. This unappreciated correlation is obvious for the estimates by ASCAT (Supplementary Note Figure 1*C*), while very weak for CPE and estimates from H&E staining. Therefore, we decided to use CPE if available (7,778 patient samples); otherwise, the estimate by H&E staining was used (1,704 samples).



Supplementary Note Figure 1. Correlation between purity and VAF'

The effect of sequence coverage on the detection of $\mathrm{SM_{LOH,Conv}}$

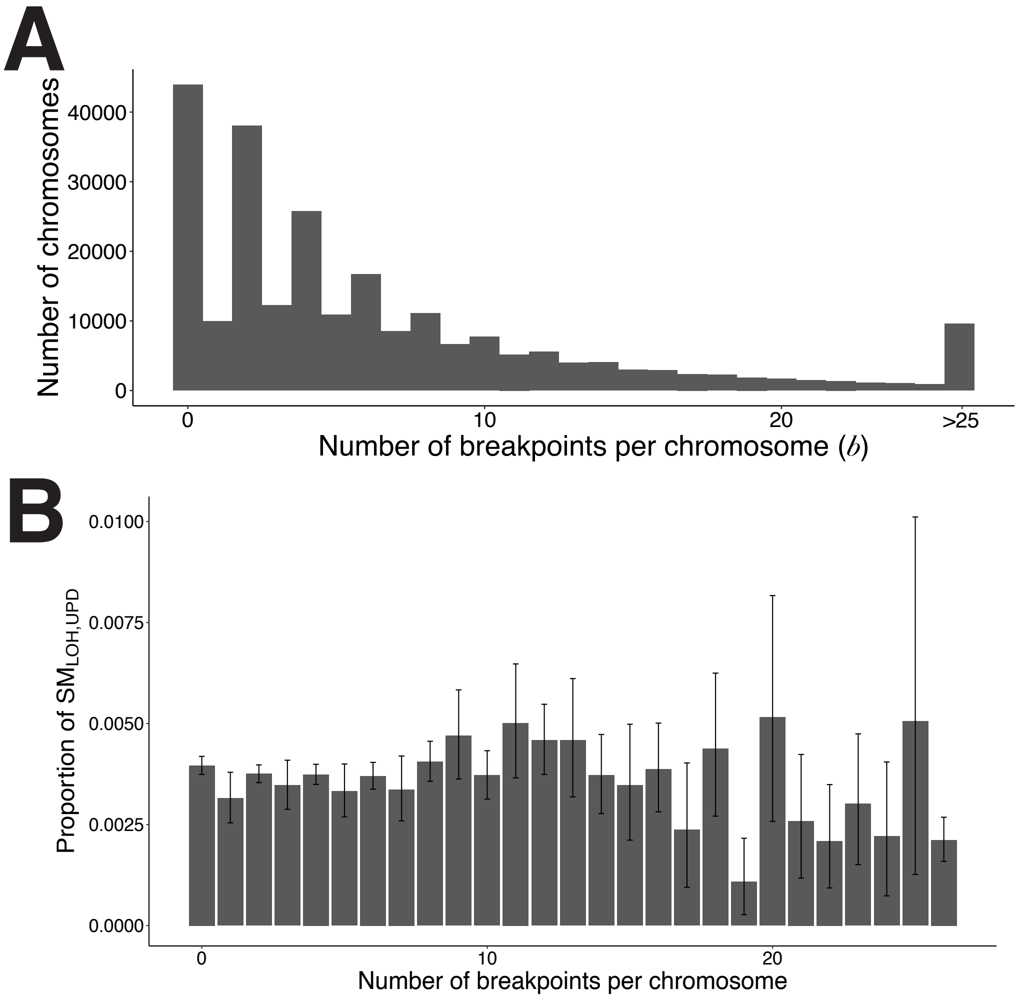
We conducted a binomial test to screen for $\mathrm{SM_{LOH,Conv}}$, in which the statistical power obviously depends on the sequence coverage. As expected, we obtained a relatively small number of $\mathrm{SM_{LOH,Conv}}$ in regions of low coverage, for example, 0.96% conpared with 1.7% when coverage ≤ 10 (see Supplementary Note Table 1). It is demonstrated that the screening process is very severe for mutations with low coverage so that the rate of false positives is minimized.

Supplementary Note Table 1. Proportion of low coverage mutations.

coverage	$\leqq 10$	$\leqq 15$	$\leqq 20$
All mutations (after screening)	29196 (1.70%)	112522 (6.56%)	217598 (12.7%)
$\mathrm{SM_{LOH,Conv}}$	48 (0.96%)	248 (4.98%)	603 (12.1%)

The relationship between the number of $\mathrm{SM_{LOH,Conv}}$ and chromothripsis or copy-number unstable regions

Chromothripsis is a mutational phenomenon where many genomic rearrangements are involved simultaneously in a specific chromosomal region. It is suspected that, in a chromosome that underwent chromothripsis, our estimates of copy number alternation may be unreliable, thereby causing a problem in calling $\mathrm{SM_{LOH,Conv}}$. If so, we predict that $\mathrm{SM_{LOH,Conv}}$ could be erroneously enriched in chromosomes with many rearrangement events involved. To test this, we focused on the number of breakpoints of copy-number alterations on each chromosome (b), and investigated if there a correlation between b and the number of $\mathrm{SM_{LOH,Conv}}$. Supplementary Note Figure 2A shows the density distribution of copy-number alteration breakpoints (estimated by ASCAT[3]). Enrichment of even numbers means that copy-number alterations within a chromosome usually have two breakpoints. Supplementary Note Figure 2B shows that $\mathrm{SM_{LOH,Conv}}$ is not particularly related to b. This result suggests that our analyses less likely include $\mathrm{SM_{LOH,Conv}}$ erroneously detected around chromothripsis or copy-number unstable regions.



Supplementary Note Figure 2. (*A*) The density distribution of the number of copy-number alteration breakpoints per chromosome (*b*). (*B*) The proportion of the number of SM$_{\mathrm{LOH,Conv}}$ plotted against the number of copy-number alteration breakpoints per chromosome (*b*). The error bars represent the 95% confidence intervals.

The effect of cutoff values in the screening for $\mathrm{SM}_{\mathrm{LOH}}$

The effect of cutoff values on the number of $\mathrm{SM}_{\mathrm{LOH}}$ was investigated. In the main text, we screened for somatic mutations with purity ≥ 0.7 and VAF' ≥ 0.8. Supplementary Note Table 2 shows the results for other pairs of the cutoff values. We found that the proportion of $\mathrm{SM}_{\mathrm{LOH,Conv}}$ to $\mathrm{SM}_{\mathrm{LOH}}$ is around 15% for all pairs, indicating that our result is very robust to the cutoff values.

Supplementary Note Table 2. Gene conversion contribution to archive LOH in each cutoff.

	VAF'				
purity	$\geqq 0.75$	$\geqq 0.8$	$\geqq 0.85$	$\geqq 0.9$	$\geqq 0.95$
≥ 0.6	8915 / 53218 (16.8%)	7434 / 45991 (16.2%)	6017 / 37965 (15.9%)	4722 / 29851 (15.8%)	3434 / 22140 (15.5%)
≥ 0.7	5699 / 38724 (14.7%)	4978 / 33625 (14.8%)	4177 / 27666 (15.1%)	3355 / 21345 (15.7%)	2442 / 15232 (16.0%)
≥ 0.8	3559 / 23282 (15.3%)	3123 / 20308 (15.4%)	2621 / 16596 (15.8%)	2056 / 12409 (16.6%)	1408 / 8242 (17.1%)

References

1. Liu, J. *et al.* An integrated tcga pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell* **173**, 400–416 (2018).
2. Aran, D., Sirota, M. & Butte, A. J. Systematic pan-cancer analysis of tumour purity. *Nat. Commun.* **6**, 1–12 (2015).
3. Van Loo, P. *et al.* Allele-specific copy number analysis of tumors. *Proc. Natl Acad. Sci. USA* **107**, 16910–16915 (2010).