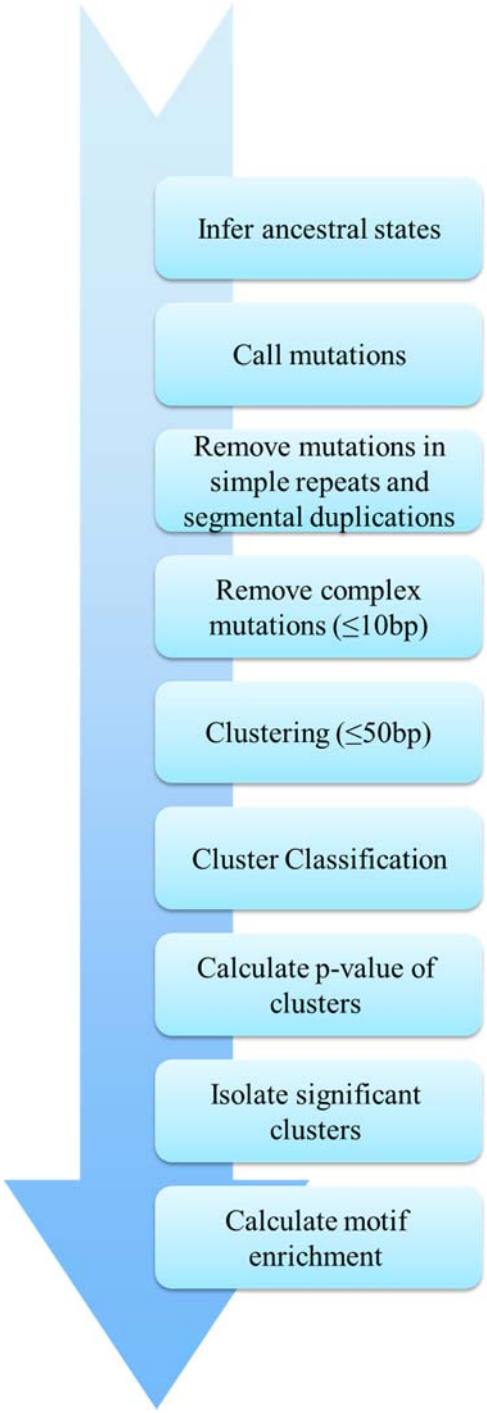
Clustered mutations in hominid genome evolution are consistent with APOBEC3G enzymatic activity

Yishay Pinto, Orshay Gabay, Leonardo Arbiza, Aaron J. Sams, Alon Keinan, Erez Y. Levanon

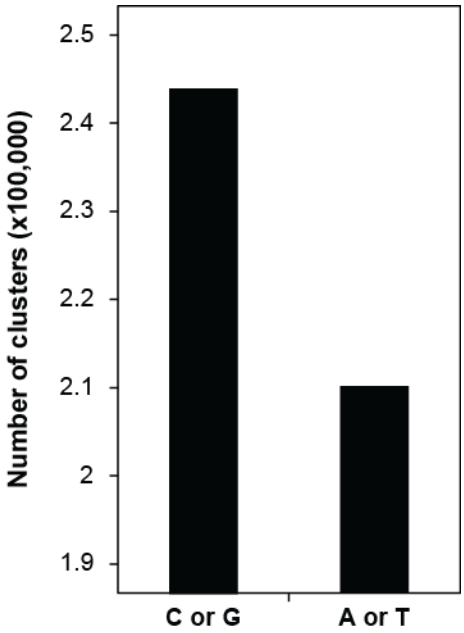
Supplemental Figures (S1-S15)

Supplemental Figure S1:



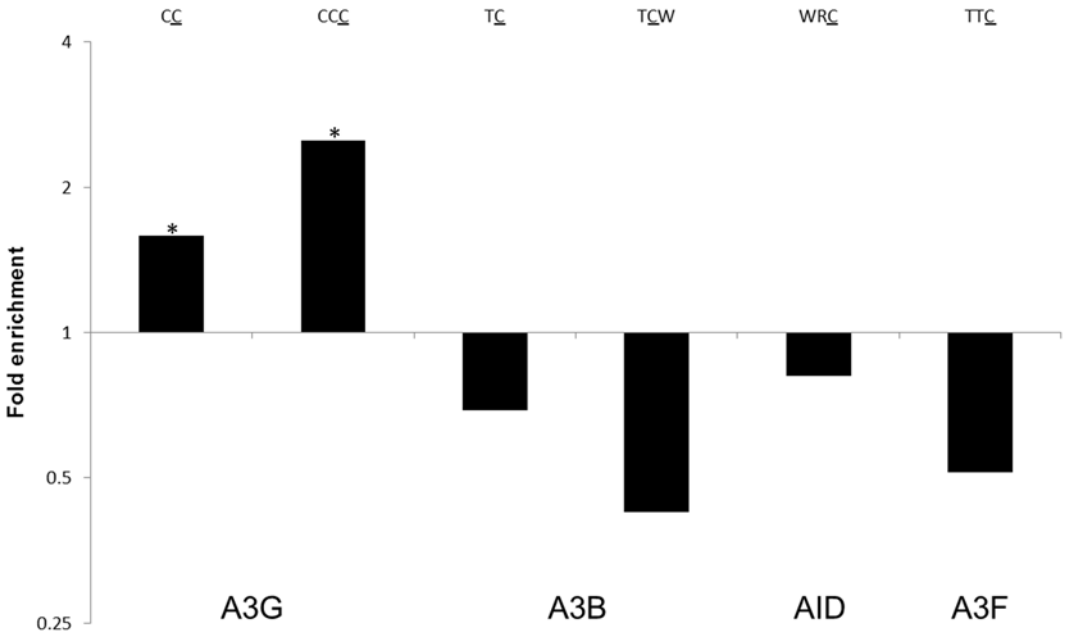
Supplemental Figure S1. Workflow used to compute motif enrichment. The pipeline consists of three major steps: calling of mutations, followed by clustering, and finally the estimation of motif enrichment for different motifs in different clusters types.

Supplemental Figure S2:



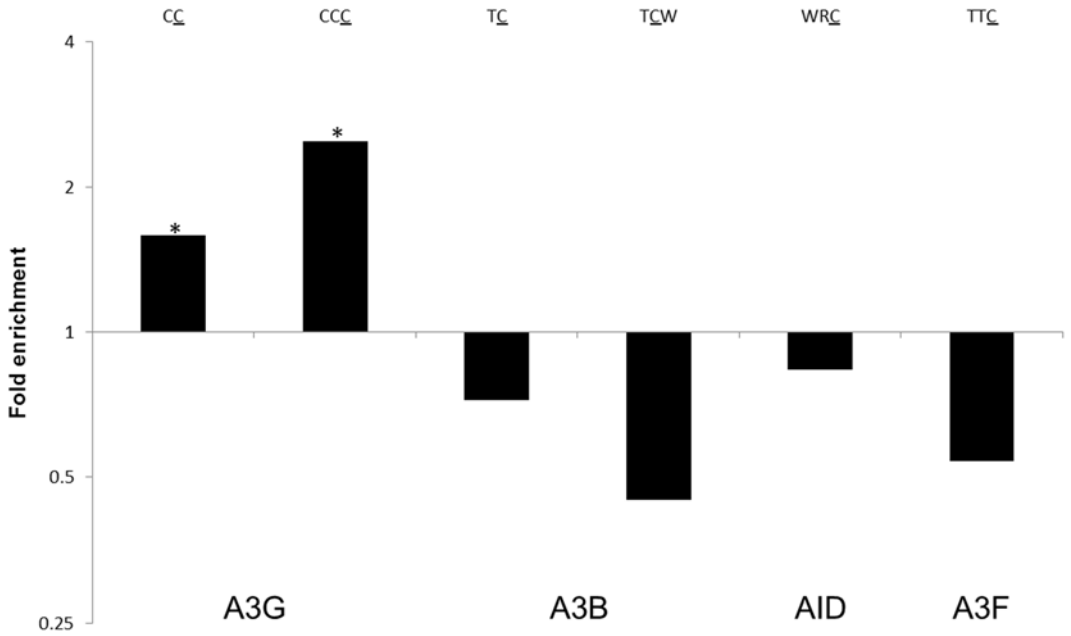
Supplemental Figure S2. The number of C or G clusters is higher than expected. The total count of C- (or G-) coordinated clusters is higher than that of T- (or A-) coordinated clusters (χ^2 test of independence $p = 1.901 \times 10^{-112}$).

Supplemental Figure S3:



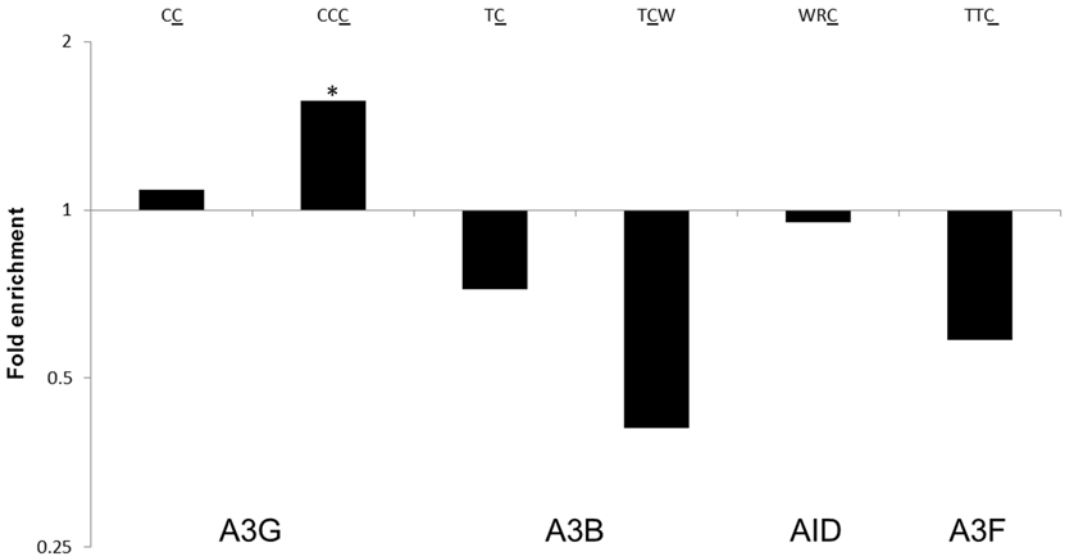
Supplemental Figure S3. Enrichment of motifs relative to 100 random sets of mutations. A3G C<u>C</u> and CC<u>C</u> motifs are significantly enriched within the set of C- (or G-) coordinated clusters with p-value ≤ 0.0001 (* $q < 3.3 \times 10^{-13}$, one-tailed Fisher's exact test after Bonferroni correction). p-values were estimated by comparing the frequency of mutations within and outside of given motifs against those arising from C or G nucleotides in 100 random sets of mutations matched by number and type of mutation). Other than the motifs corresponding to A3G, no other APOBEC-related motifs were found to be enriched within clusters.

Supplemental Figure S4:



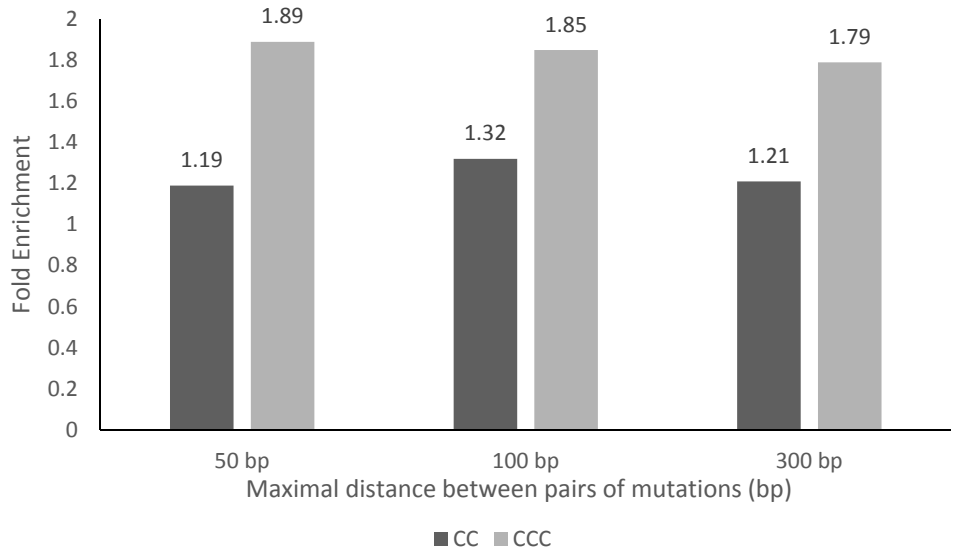
Supplemental Figure S4. Enrichment of motifs relative to 100 random sets of clusters. A3G CC and CCC motifs are significantly enriched within the set of C- (or G-) coordinated clusters with p-value ≤ 0.0001 (* $q < 1.3 \times 10^{-13}$, one-tailed Fisher's exact test after Bonferroni correction). p-values were estimated by comparing the frequency of mutations within and outside of a certain motif against that of C or G nucleotides in 100 random sets of clusters matched by number and type of mutations and type of clusters). Other than the motifs corresponding to A3G, no other APOBEC-related motifs were found to be enriched within clusters.

Supplemental Figure S5:



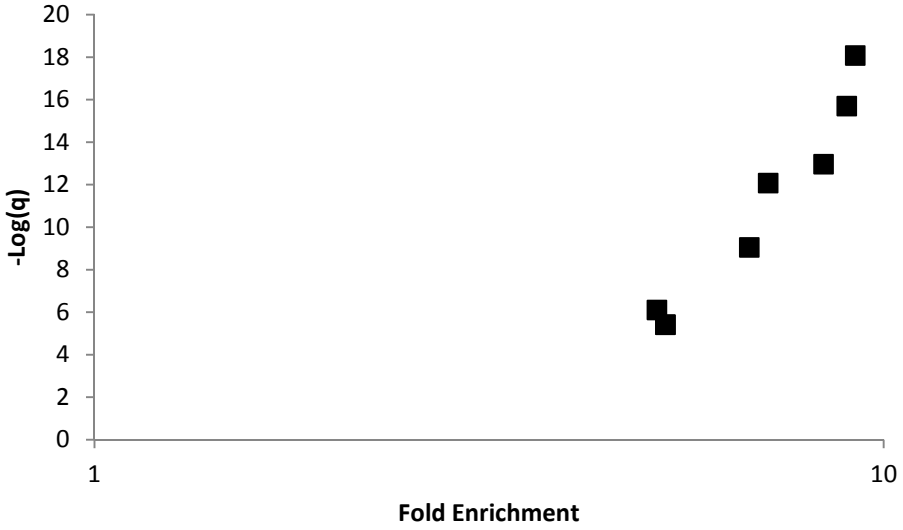
Supplemental Figure S5. Enrichment of motifs relative to local context. A3G CCC motif is significantly enriched within the set of C- (or G-) coordinated clusters with p-value ≤ 0.0001 (* $q < 5.9 \times 10^{-5}$, one-tailed Fisher's exact test after Bonferroni correction). p-values were estimated by comparing the frequency of mutations within and outside of a certain motif against that of C or G nucleotides in 10kb regions encompassing each cluster). Other than CCC, no other APOBEC-related motifs were found to be enriched within clusters.

Supplemental Figure S6:



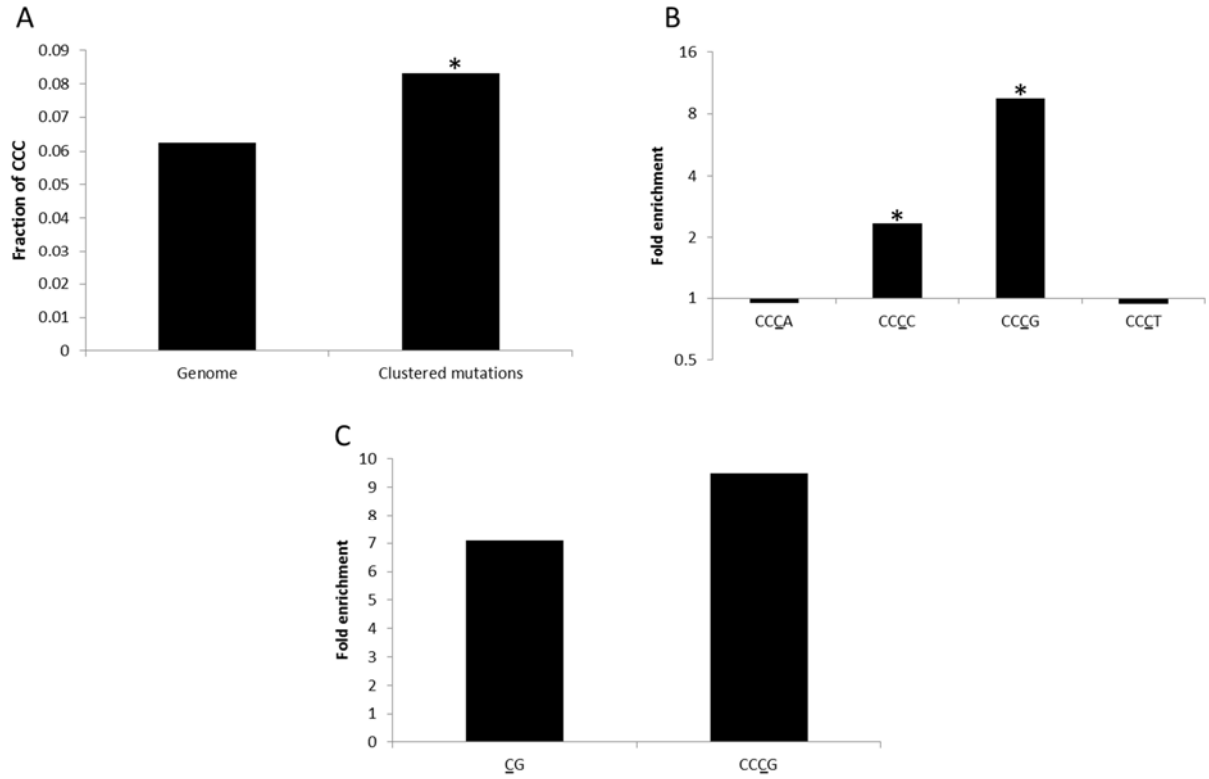
Supplemental Figure S6. Enrichment of C<u>C</u> and CC<u>C</u> motifs in different clustering distances. A3G C<u>C</u> and CC<u>C</u> motifs are significantly enriched within the set of C- (or G-) coordinated clusters with p-value ≤ 0.0001 when clustering with maximal distance of 50, 100 and 300 bp. ($p < 0.05$, one-tailed Fisher's exact test calculated by comparing the frequency of mutations within and outside a given motif while controlling for the frequency of C or G nucleotides within and outside this motif in the genome). For 1000 bp when applying the strict statistical criterion (p-value ≤ 0.0001), no clusters are left for performing an enrichment analysis. For this case we used clusters p-value ≤ 0.1 and fold enrichment values are 1.18 and 1.30 for C<u>C</u> and CC<u>C</u> respectively.

Supplemental Figure S7:



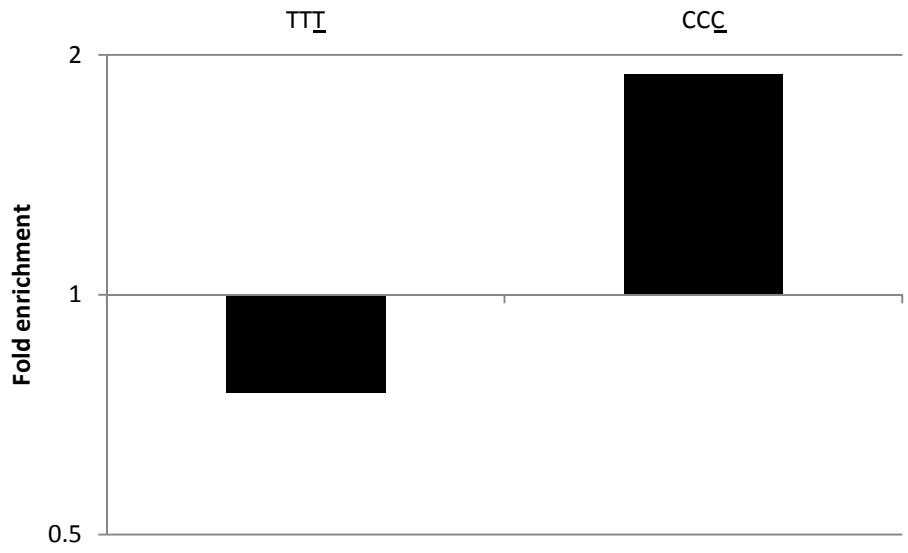
Supplemental Figure S7. Enrichment of triplets containing CpG di-nucleotides. The fold enrichment relative to genomic background of all CpG containing three-nucleotide motifs within C- (or G-) coordinated clusters (x-axis) is shown plotted against Bonferroni-corrected p-values for one-tailed Fisher's exact tests (y-axis).

Supplemental Figure S8:



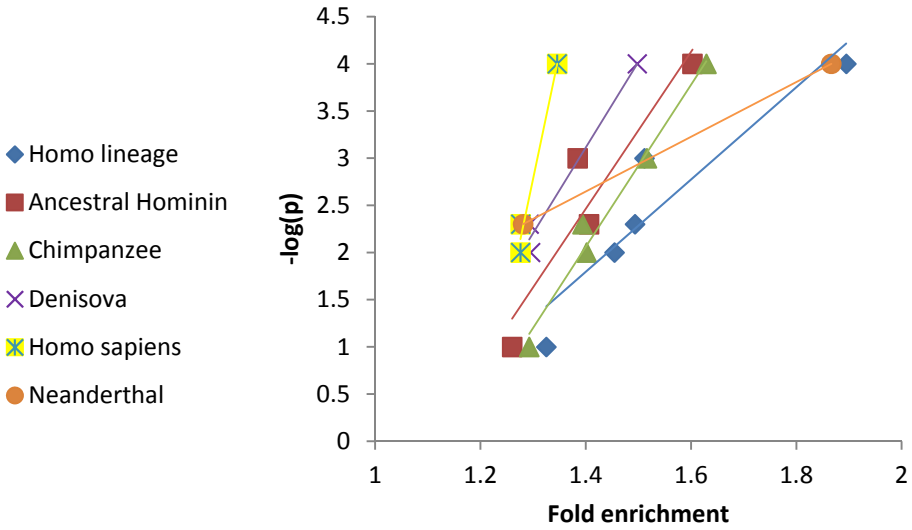
Supplemental Figure S8. Enrichment of CCC motif is not restricted to CpCpCpG. **(A)** Higher fraction of the CCC motif when excluding CpG mutations and masking genomic CpGs relative to the frequency of CCCH (H- A\C\T) in the human genome (* $p < 0.05$). **(B)** CCCC and CCCG motifs are significantly enriched relative to genomic background within the set of C- (or G-) coordinated clusters with a p-value ≤ 0.0001 (* $q < 0.05$, calculated as in Figure 2). It is worthwhile to mention that the high enrichment of the CCCG motif and the insignificant under-representation of the CCCA motif are, in part, devoted to the relatively lower and higher abundance of these tetranucleotides in the genomic background with fold-enrichment of 3.24 and 2.18 relative to 100 random sets of mutations, respectively. **(C)** CCCG motifs have higher enrichment values than CG alone relative to genomic background, suggesting an additive effect between overlapping CCC and CG motifs.

Supplemental Figure S9:



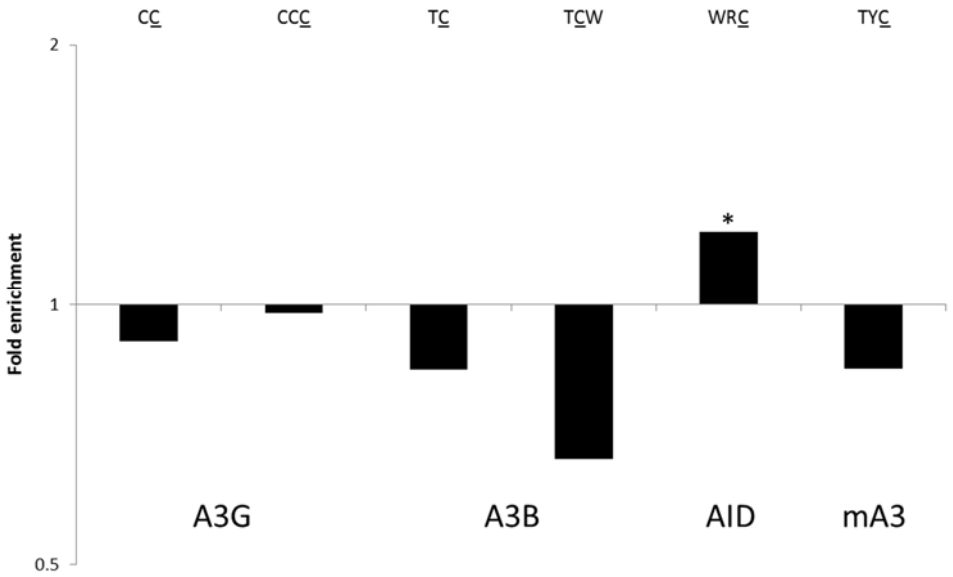
Supplemental Figure S9. Enrichment in CCC is not a result of increased mutations in homotypic trinucleotides. No enrichment relative to genomic background was found for TTT (or the AAA) motif, within the set of T- (or A-) coordinated clusters with p-value ≤ 0.0001.

Supplemental Figure S10:



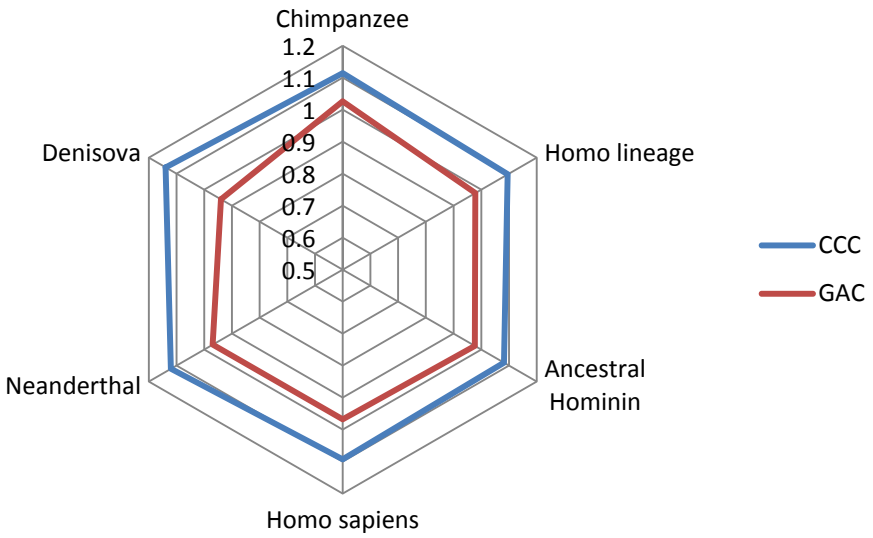
Supplemental Figure S10. Positive correlation between cluster reliability and motif enrichment in all branches. Pearson's correlation coefficient was calculated for each branch separately (average r = 0.96). The Pearson's correlation coefficient estimated jointly across all branches was 0.72 with a p-value of 0.0001. The enrichment values were calculated relative to genomic background.

Supplemental Figure S11:



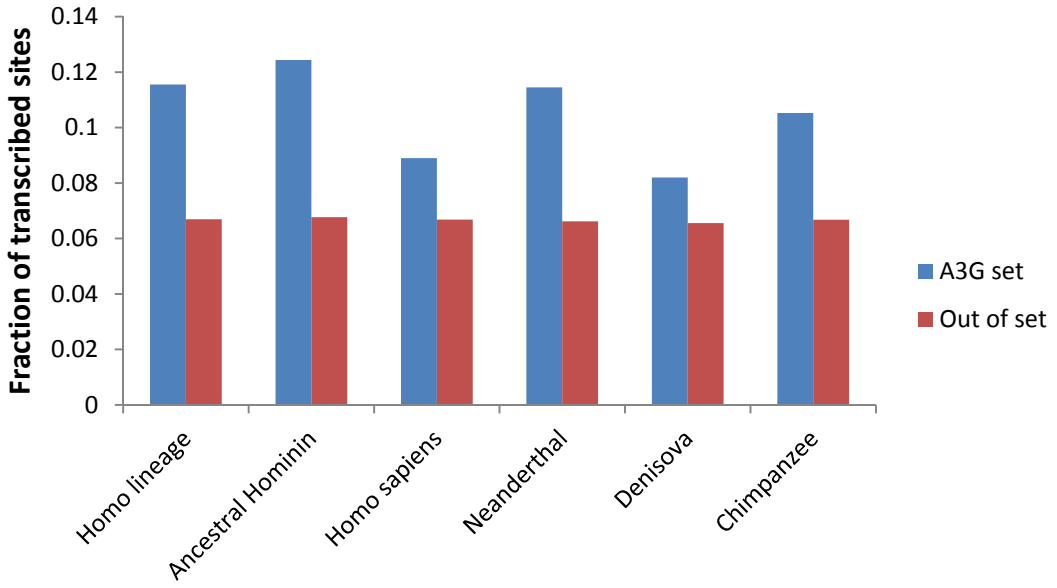
Supplemental Figure S11. No enrichment was found for mA3 or A3G motifs between different mouse strains. Applying the exact same method as in Figure 2A, we calculated motif enrichments for mutations occurring in C57BL/6J mice since the divergence from NZO/HlLtJ. As expected in this negative control, with lineages lacking the enzyme, no enrichment for A3G-related motifs (CC or CCC) was observed. Moreover, no enrichment was found for mA3-related TYC motif, suggesting that APOBEC mutagenesis is a primate-specific phenomenon. (* $q < 0.05$ one-tailed Fisher's exact test after Bonferroni correction. p-value was calculated using the frequency of mutations compared against the frequency of C or G nucleotides within or outside the motifs in the mouse C57BL/6J reference genome (mm9)).

Supplemental Figure S12:



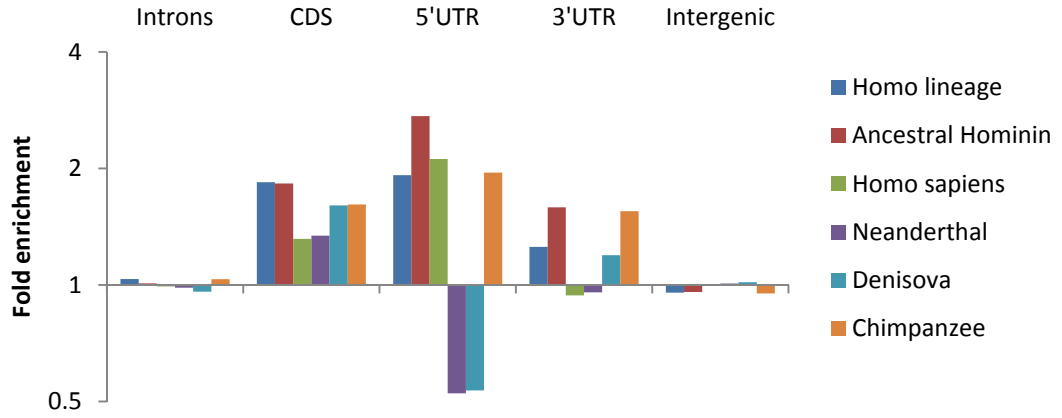
Supplemental Figure S12. **Higher fraction of A3G clusters than expected**. An A3G cluster is defined as a coordinated cluster with at least one CCC mutation. We calculated the expected number of clusters using a binomial variate with probability equal to the chance of mutation in a specific motif. The inert GAC motif (not subject to APOBEC-induced mutations) was used as a control.

Supplemental Figure S13:



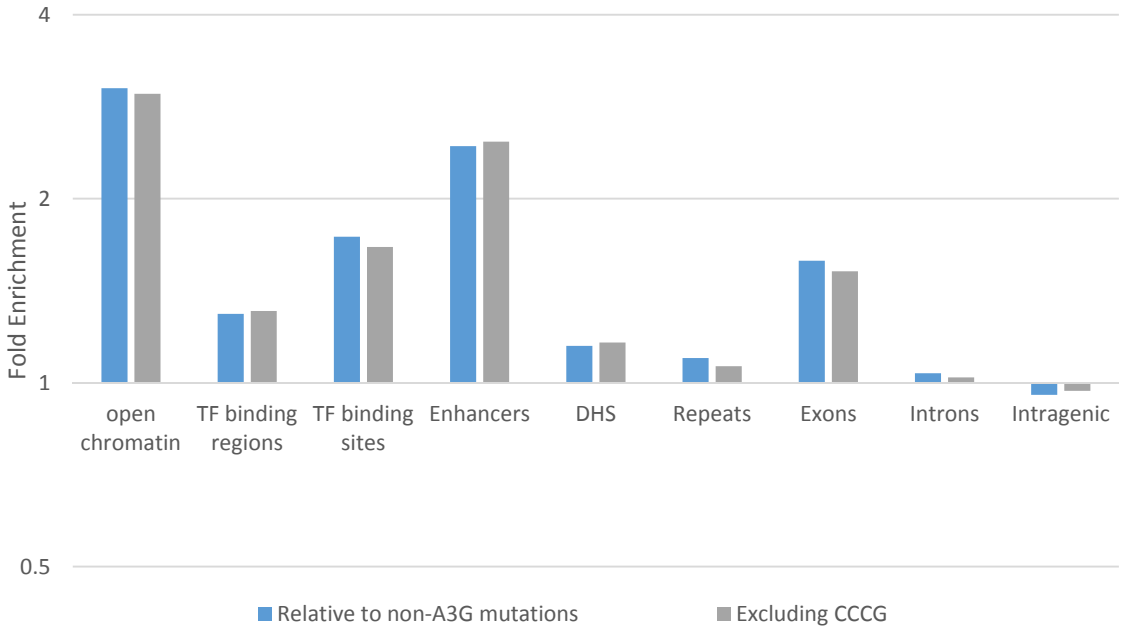
Supplemental Figure S13. A3G mutations have a higher tendency to be located within transcribed regions. A3G-induced clustered mutations tend to be generated within transcribed regions, relative to all other mutations. This phenomenon is common to all lineages. ($q < 0.05$, chi-squared test corrected by FDR).

Supplemental Figure S14:



Supplemental Figure S14. Proportion of APOBEC3G-induced mutations within exonic regions, introns, and intergenic regions. We calculated the proportion of mutations within coding sequences, 5'UTR and 3'UTR, introns and intergenic regions from all mutations in the different A3G sets. Fold enrichment was estimated as the ratio between the proportions of each region in the A3G set relative to that of all other mutations. A two-tailed Fisher's exact test after Bonferroni correction shows $q < 0.05$ for all regions in the *Homo*, chimpanzee and ancestral Hominin lineages, with the exception of introns in the last lineage.

Supplemental Figure S15:



Supplemental Figure S15. Proportion of APOBEC3G-induced mutations within different genomic regions. We calculated the proportion of mutations within the indicated genomic regions including or excluding mutations falling within CCCG motif from all mutations in the *Homo* lineage. The comparable results rule out significant contribution of non-enzymatic deamination in CpG dinucleotide. Fold enrichment was estimated as the ratio between the proportions of each region in the A3G set relative to that of all other mutations. A two-tailed Fisher's exact test after Bonferroni correction shows $q < 0.05$ for all regions in the *Homo* lineage with the exception of introns and intragenic regions.