Characterization of intermediate-sized insertions using whole genome sequencing data and analysis of their functional impacts on gene expression

Saeideh Ashouri[1,2], Jing Hao Wong[1], Hidewaki Nakagawa[3], Mihoko Shimada[1], Katsushi Tokunaga[2], Akihiro Fujimoto[1]

[1] Department of Human Genetics, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan

[2] Toyama Genome Medical Science Project, National Center for Global Health and Medicine, Tokyo, Japan

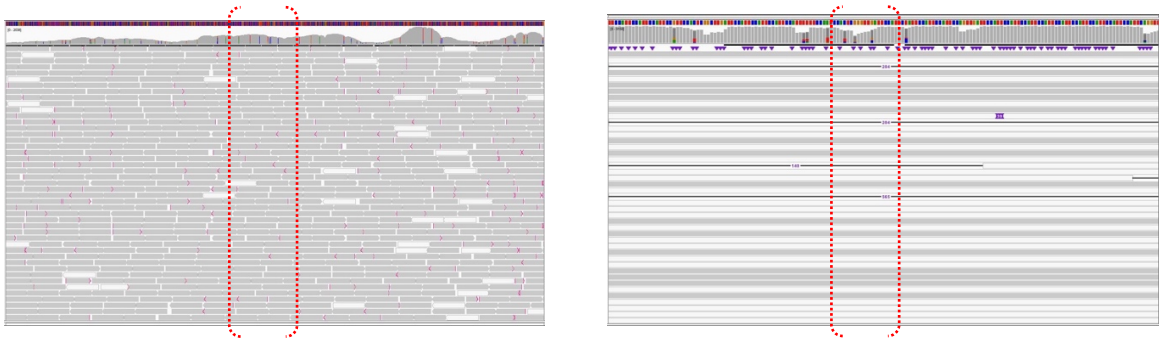
[3] Laboratory for Cancer Genomics, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan

[*]*Correspondence should be addressed to Akihiro Fujimoto (afujimoto@m.u-tokyo.ac.jp)*.

Saeideh Ashouri; sashouri@ri.ncgm.go.jp, Jing Hao Wong; jhwong@m.u-tokyo.ac.jp, Hidewaki Nakagawa; hidewaki@riken.jp, Mihoko Shimada; shimada.mihoko@mail.u-tokyo.ac.jp, Katsushi Tokunaga; katokunaga@ri.ncgm.go.jp, Akihiro Fujimoto; afujimoto@m.u-tokyo.ac.jp

(a) chr4:49119130

(b) chr4:49119130



(c) chrY:11293631

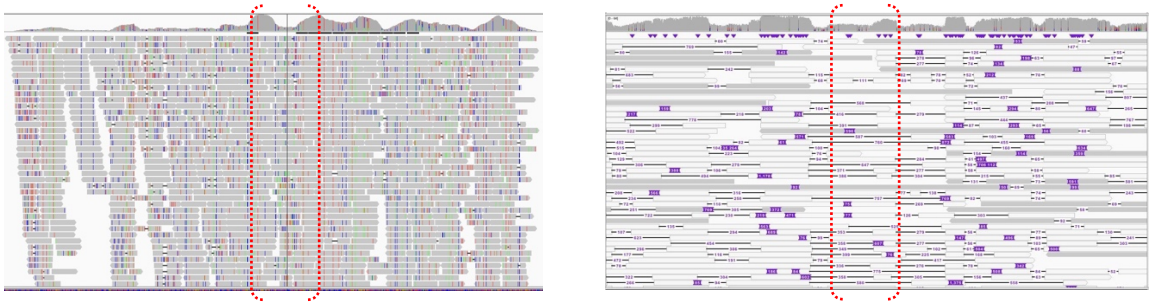


Figure S1: Examples of true positive and false positive insertion candidates detected by IMSindel and Nanopore. Surrounding regions are visualized by IGV (ver2.7.2). Insertion positions are show by dashed brackets. (a) An example of a true positive insertion on chr4:30540793 detected by both IMSindel (on the left) and Nanopore (on the right). For true positive insertions the insertion position is clearly seen in short reads and there is one insertion at the same position in long reads. (b) An example of a false positive insertion in the centromeric region on chr4:49119130. As can be seen from the figure on the left, the insertion position is not clear in short reads, the number of reads is too high. In long reads in the right figure, there is no clear insertion detected. (c) Another example of a false positive insertion in a region full of simple repeats on chrY:11293631. The insertion position is not clear in short reads (left figure) and there are many insertions in long reads which is usually caused by misalignment of reads due to repetitive sequences.

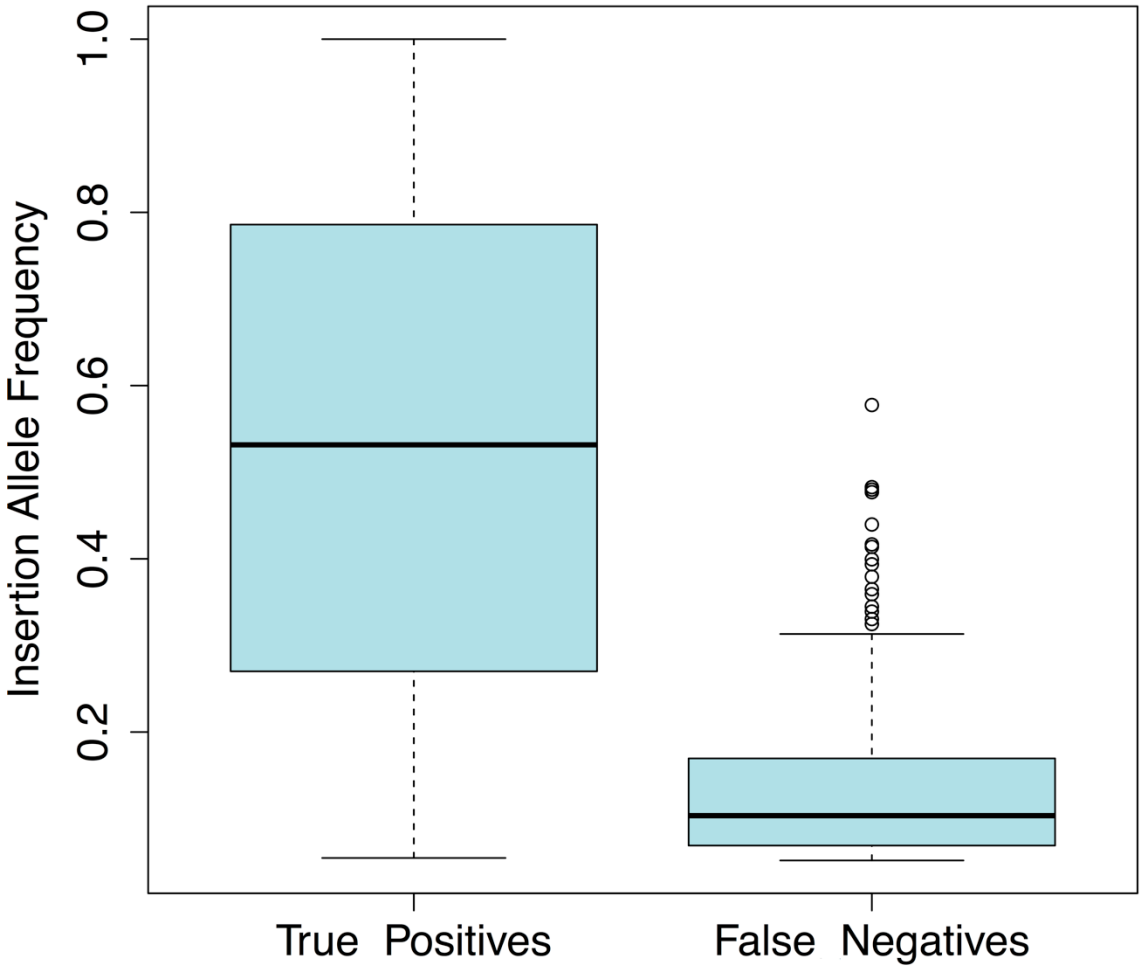


Figure S2: Allele frequency of consistently and inconsistently imputed insertion. The plot shows allele frequencies of insertion positions imputed as homozygous and heterozygous insertion and detected as insertion by Nanopore (True Positives) and that of insertion positions imputed as homozygous for reference allele but detected as insertion by Nanopore (False Negatives). It was observed that False Negatives have generally lower allele frequencies than True Positives.

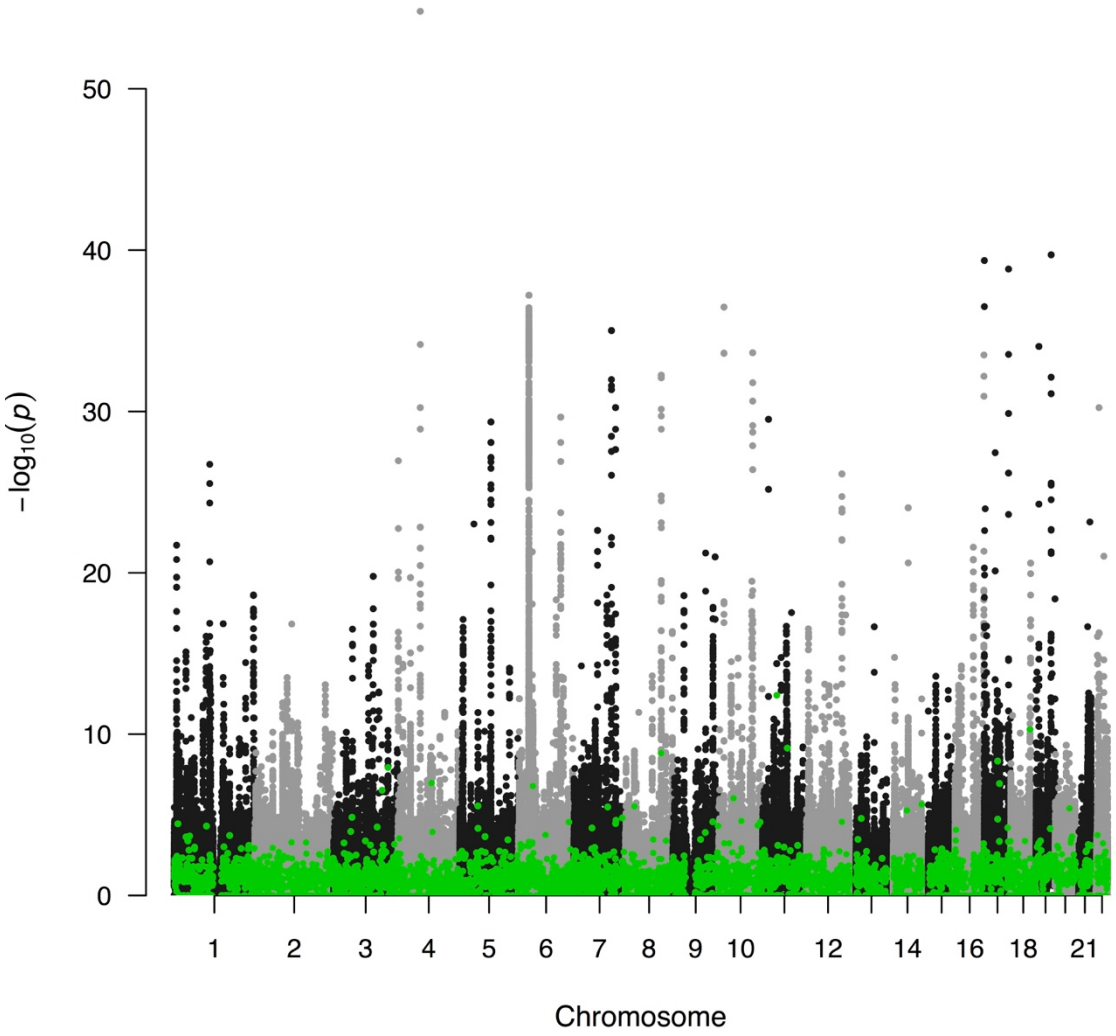


Figure S3: Manhattan plot of eQTL association results of all variants and identified intermediate-sized insertions (highlighted in green). The x-axis shows the chromosomes on which the variants are located and the y-axis is the –log10 of the lowest p-value for each variant in the eQTL association results.

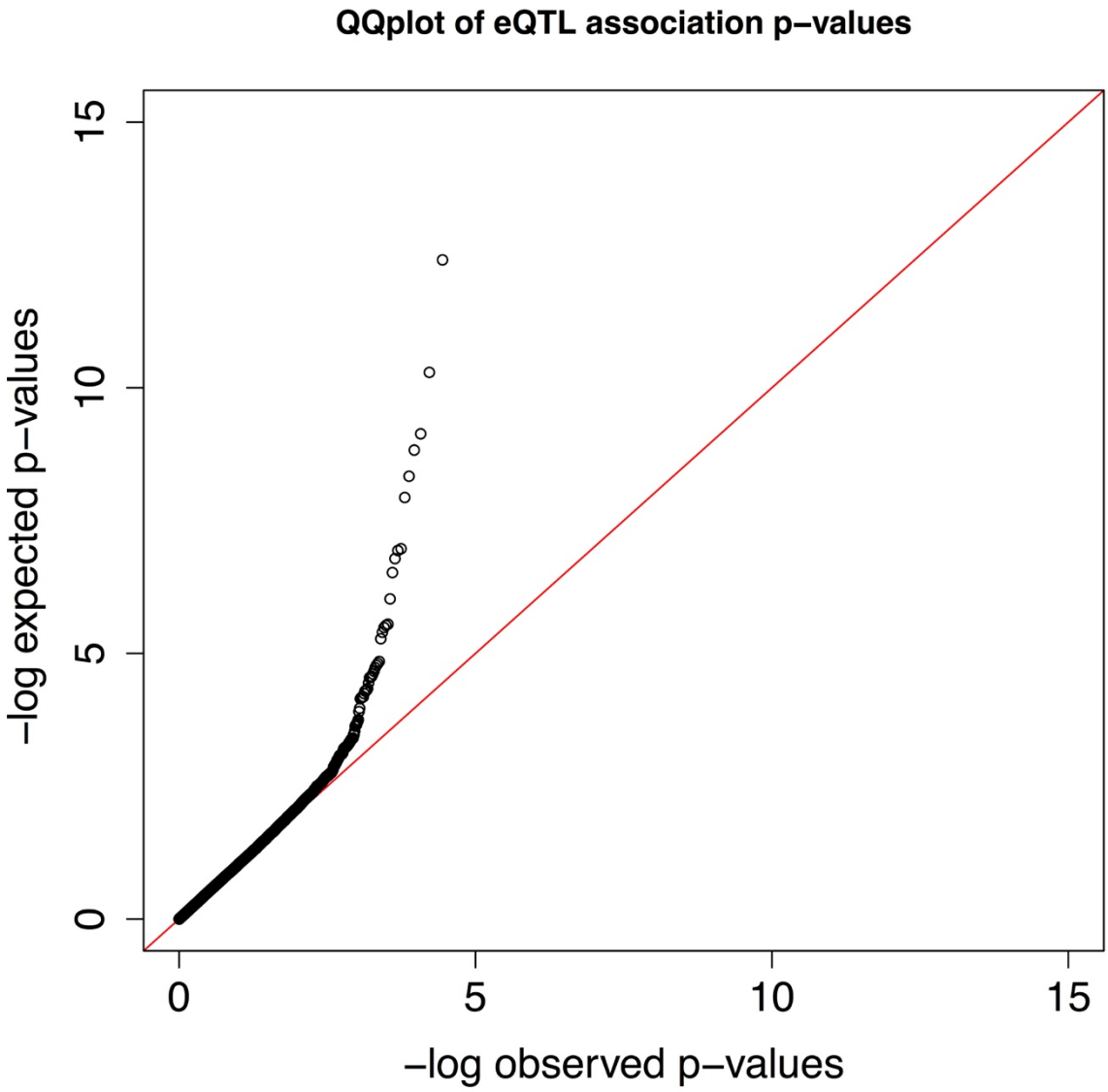
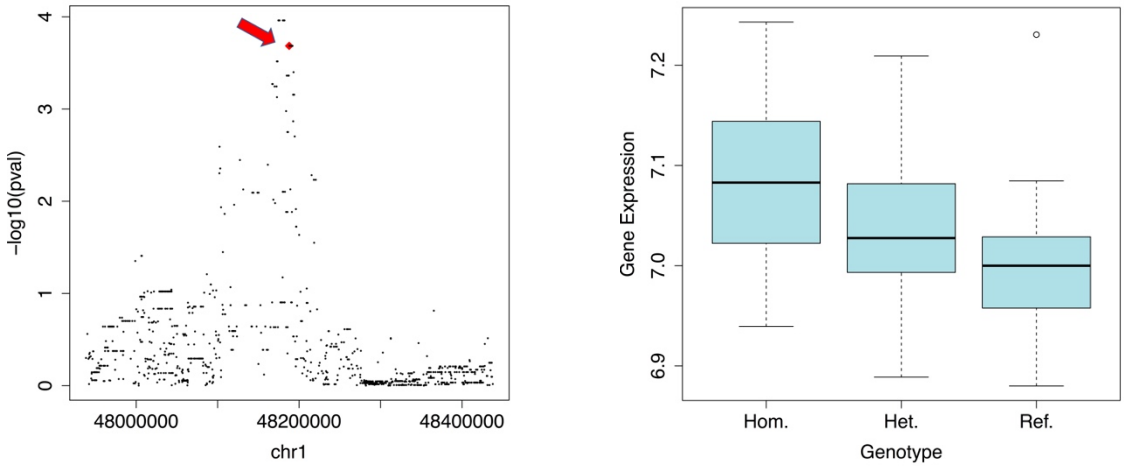
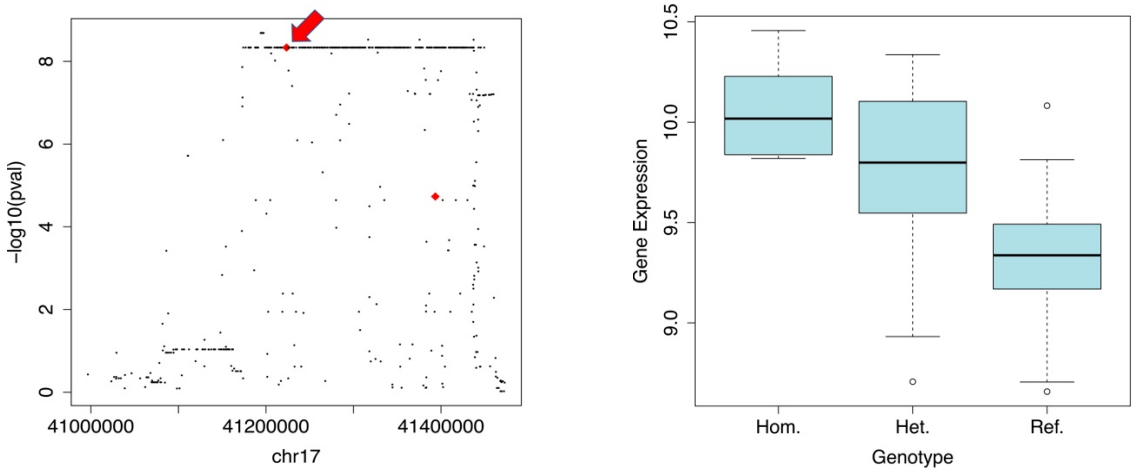


Figure S4: The Quantile-Quantile (QQ) plot of eQTL association analysis of intermediate-sized insertion candidates. The genomic inflation factor (λ_{GC}) was 1.045 which shows the inflation in association results is negligible.

(a) Chr1:48187815 (insertion)



(b) Chr17:41223538 (insertion)



(c) Chr18:59852898

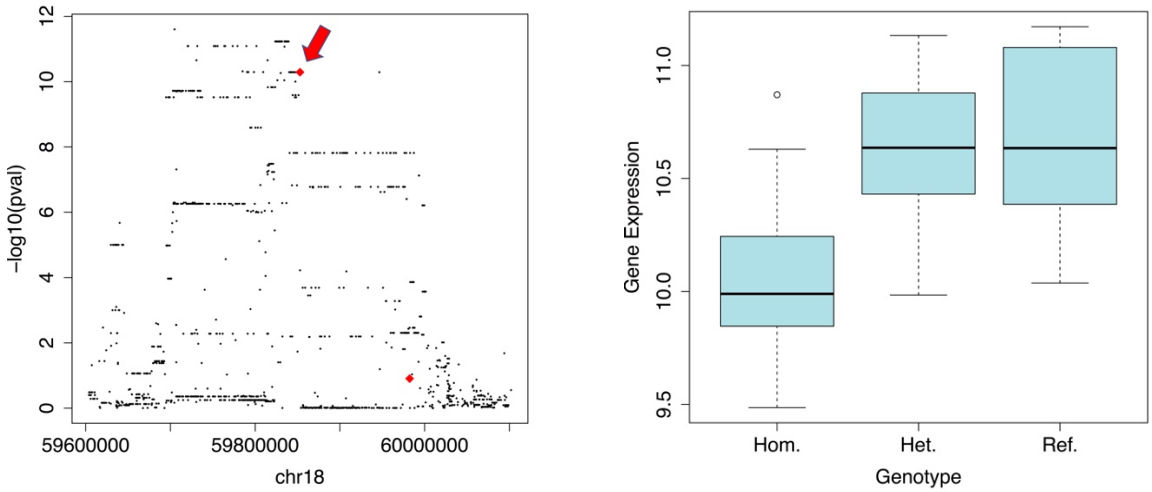


Figure S5: Examples of eQTL association p-value plots and gene expression boxplots of three causal insertion candidates shown with red diamonds with red arrows

indicating their positions. (a) Plots for insertion at chr1:48187815. (b) Plots for insertion at chr17:41223538. (c) Plots for insertion at chr18:59852898.

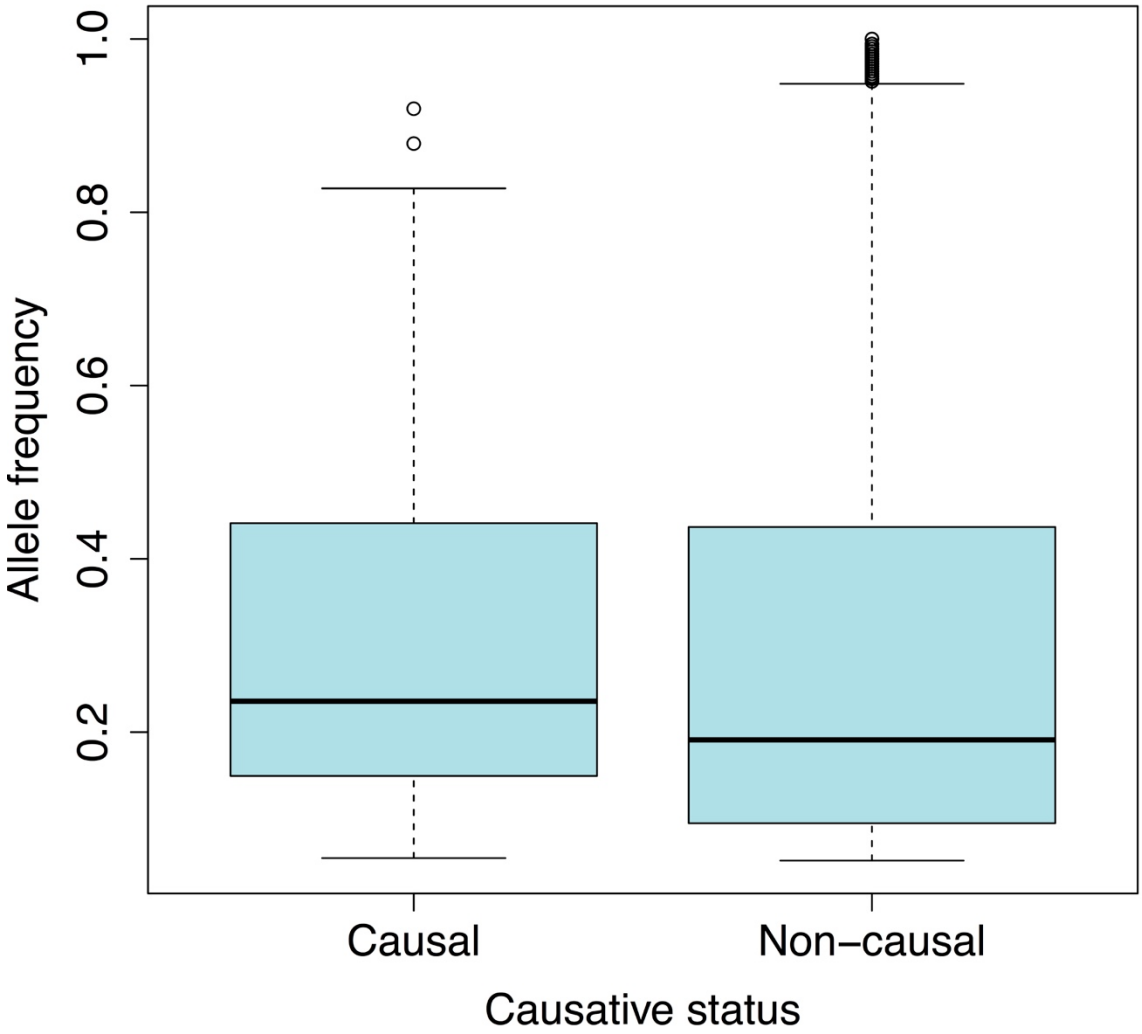


Figure S6: Comparison of allele frequency between causal versus other insertions. This plot shows higher allele frequency of causal insertions compared with other insertions (Wilcoxon P-value=0.015).

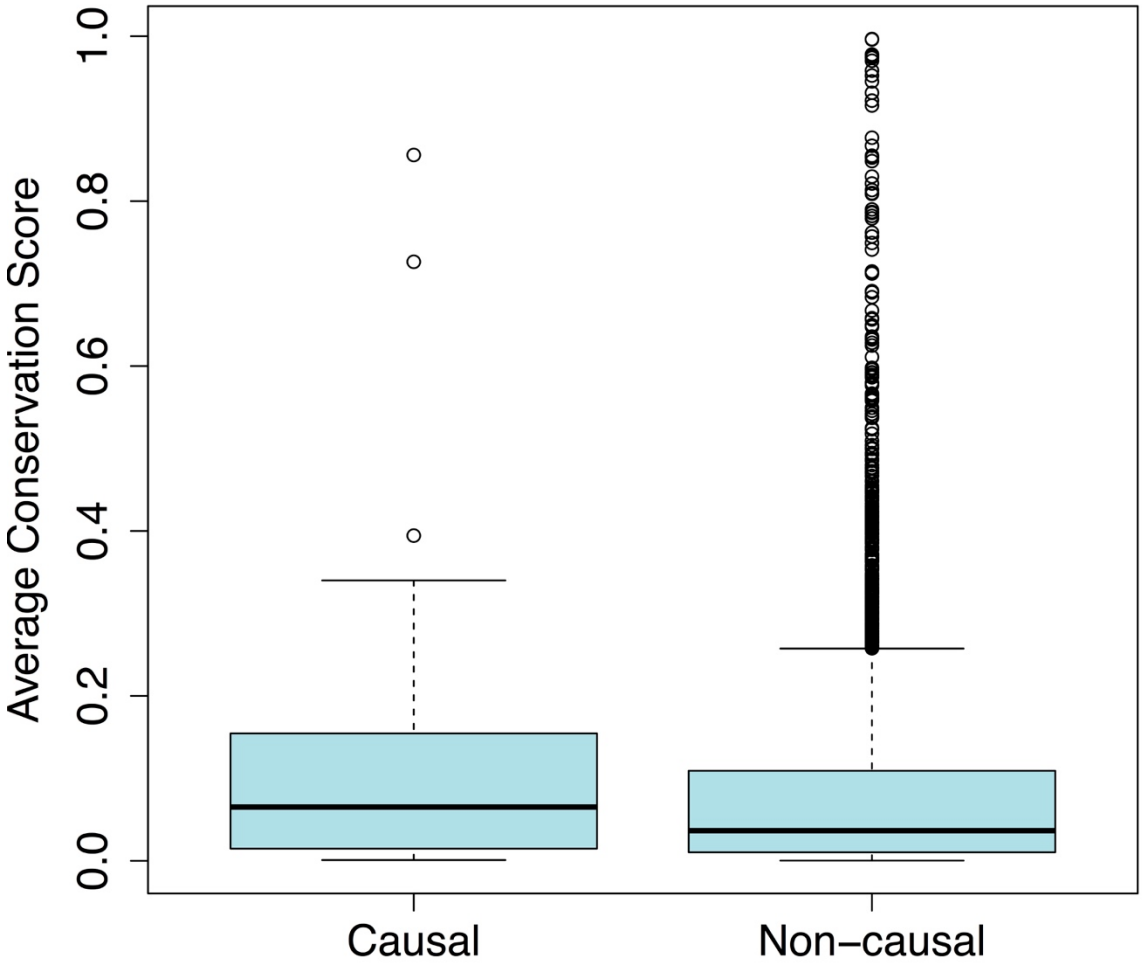


Figure S7: Average genome conservation score between causal and non-causal insertion candidates. The plot shows higher average genome conservation scores of 100bp region flanking the position of causal insertion compared with those of other insertions (Wilcoxon rank-sum test p-value = 3.2×10^{-3}).

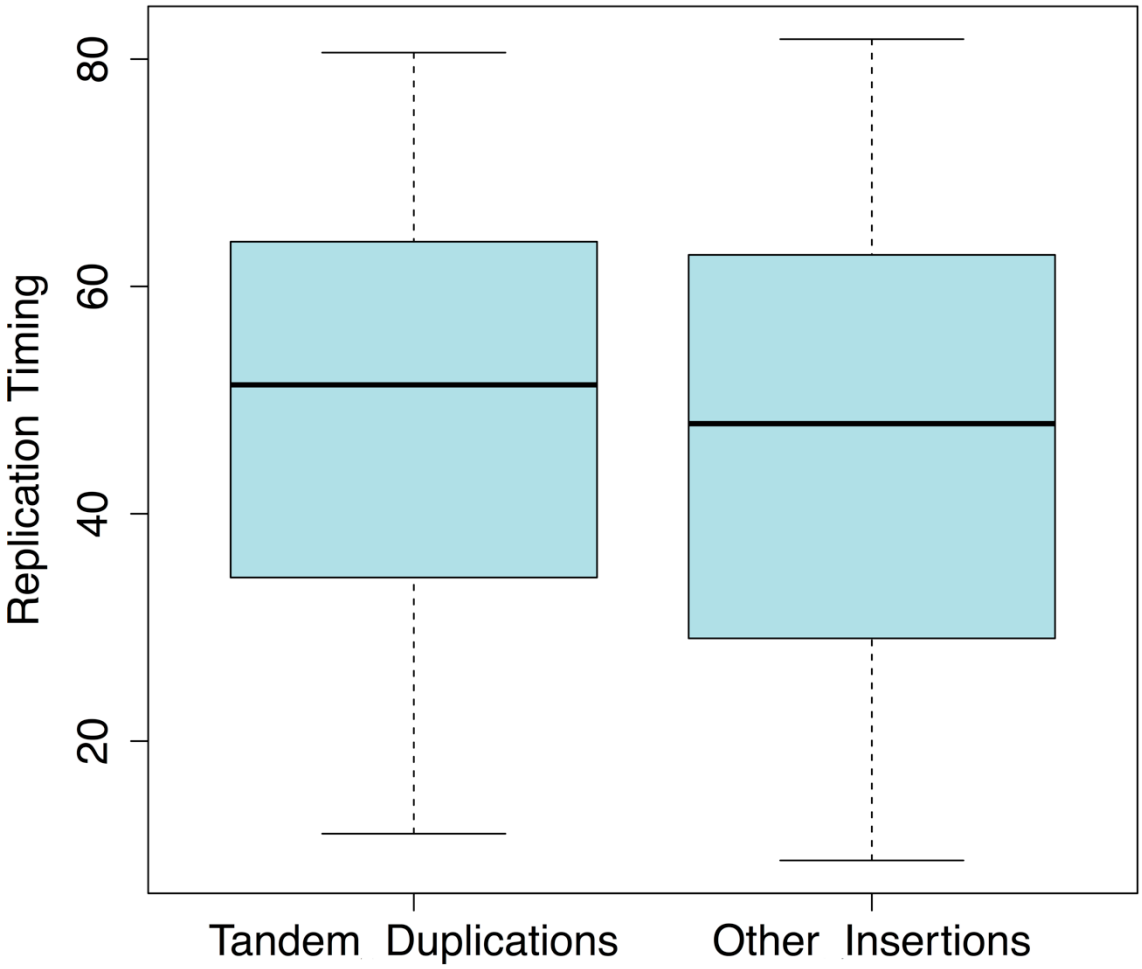


Figure S8: Average replication timing of between tandem duplications and other insertion candidates. The plot shows higher average replication timing for tandem duplications compared with that of other insertions (Wilcoxon rank-sum test p-value = 0.0034).