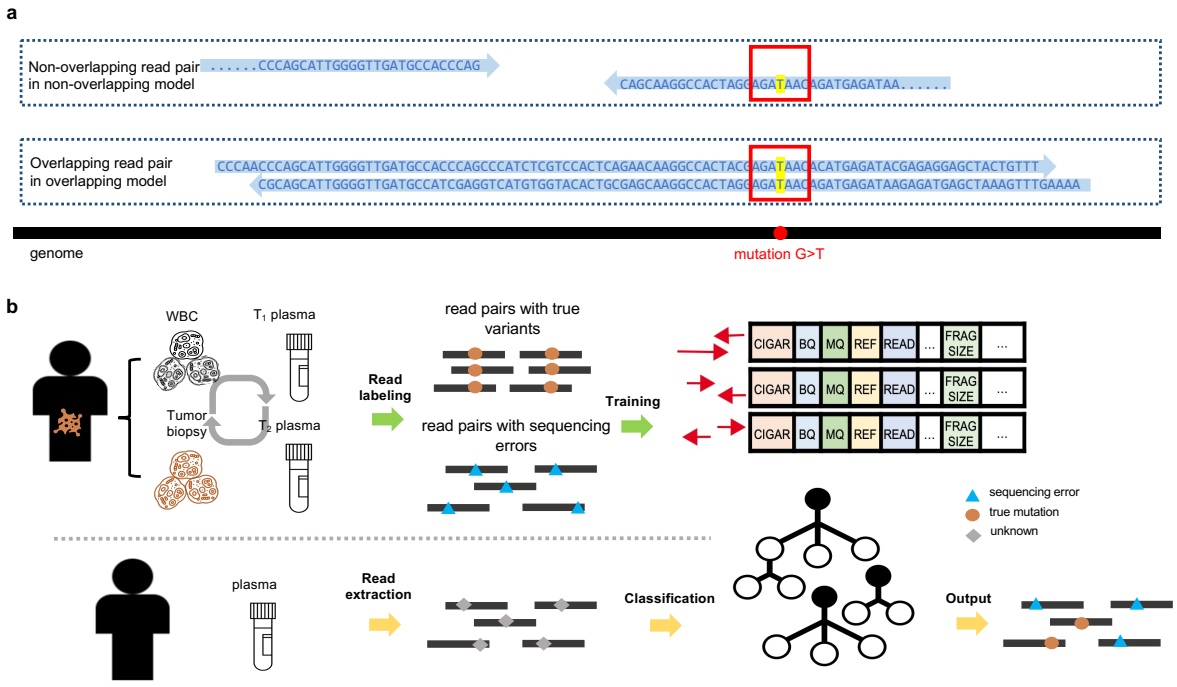

Figure S3. Machine learning model for suppressing sequencing errors. (a) Read pairs in the overlapping and non-overlapping random forest models. Regarding to a mutation site, there are two types of read pairs: one (non-overlapping read pair) doesn't overlap at the site; the other (overlapping read pair) overlaps at the site. The overlapping read pair naturally contains more information at the mutation site than the non-overlapping read pair. Therefore, two independent random forest model was trained for the overlapping read pair and the non-overlapping read pair. **(b)** Training data extraction and utilization of the random forest model for suppressing sequencing errors at the read level. The upper panel shows the training data extraction workflow (for details, see Methods). True variant positions (somatic and germline mutations) and sequencing error positions are identified by comparing the WBC sample, the tumor biopsy sample and two plasma samples from the same patient. Read pairs with nonreference bases at these identified positions are extracted and labeled "true variants" and "sequencing errors", respectively. Then, various features are extracted from each read pair, and these data are used as training and testing data for the random forest model. The lower panel shows the utilization of the random forest model. Given a post-treatment sample, the features from the read pairs at given loci are extracted from the sequencing data and classified as containing a "sequencing error" or a "true variant".