SUPPLEMENTARY DISCUSSION

Additional criteria for data filtering

To gauge the level of systematic noise of our Tn detection method and cross-contamination between samples, we routinely processed specimens from different induced mice in the same LM-PCR/sequencing analysis. A typical result of such an experiment was shown in Extended Data Figure 4a. While many Tn tags were restricted to one mouse, overlaps between tests and controls were frequently observed (shown in the red box in Extended Data Fig. 4a), indicating the presence of cross-contamination during sample preparation. Importantly, most of these commonly detected tags could be definitively assigned to a particular mouse based on their distinct abundance in the individual samples. It was also evident that the cross-contamination level was low in general, and a majority of them occurred with a frequency lower than 50 reads. As these features were repeated observed in different experiments (Data not shown), we empirically used 50-reads as a generic cutoff to retain Tn tags of high confidence. Following this filtering step, we performed manual curation of the resulted list of Tn tags to further remove the rare abundant tags whose frequencies were not ten times higher than the cumulative frequencies of the corresponding tags found in control samples. While these filtering steps will inevitably prevent detection of truly low abundant clones, they will allow us to obtain high quality tags representing the major clones in PB or BM.

Characterization of methodology for Tn insertion tags detection

Sensitivity and reproducibility of our methodology for detection of Tn insertion tags are critical factors for drawing appropriate conclusions from the sequencing data. The LM-PCR method applied here displayed comparable sensitivity as the linear-amplification mediated PCR (LAM-PCR) technique when tested with genomic DNA of uninduced M2/HSB/Tn mice (Extended Data Fig. 4b). To determine the lower-end detection limit of our method under more complex conditions, we assembled polyclonal samples, in which a serial dilutions of ten different HEK293 clones, each bearing an unique Tn insertion tag, were mixed with 10,000 $DsRed^{+}$ PB mononuclear cells from an induced M2/HSB/Tn mouse (Extended Data Fig. 4c, establishment of the ten HEK293 clones is described in Additional Methods). Duplicate samples were prepared for each dilution to assess reproducibility of the method. The inclusion of multiple HEK293 clones with different Tn insertion tags will help to determine if the observed detection characteristics of the method are generic or specific to particular insertion sites; and the use of blood cells from the induced animal will provide a polyclonal population similar to what was being investigated in our study.

Out of the ten input clones, seven were detected at multiple cell dosages (Extended Data Fig. 4c), suggesting that a majority of the insertion tags are detectable when examined in a polyclonal population. One-cell dosage was not sufficient to detect any of these tags, whereas 5- and 25-cells allowed detection

of multiple clones (clones #7, 8, and 9 at 5-cells dosage, and clones #3, 6, 7, 8, 9, and 10 at 25-cells dosage, Extended Data Fig. 4c). Tn tags from all seven detectable clones were consistently present when 100 or more cells were utilized (Extended Data Fig. 4c). Therefore, this serial dilution experiment indicates that the smallest detectable clone is approximately 5-25 cells, which encompasses around 0.05% - 0.25% of the polyclonal sample. The read frequencies of the individual Tn tags were significantly correlated in duplicate samples (Extended Data Fig. 4d), confirming a high reproducibility of the detection method. This reproducibility was further demonstrated with technical repeats of PB granulocyte samples collected from an induced M2/HSB/Tn mouse, whereby over 65% of the Tn tags were commonly detected, and the read frequencies of these tags were also highly correlated (Extended Data Fig. 4e, f).

We further analyzed the quantitative correlation between the input cell numbers and their observed frequencies in sequencing data. While positive correlation was observed for the individual tags (data not shown), significant variation in read frequencies were evident among the HEK293 clones, even when they were examined with the same input cell numbers (Extended Data Fig. 4g). Thus, this result suggests that the read frequency itself does not reflect the absolute abundance of the corresponding clones when different clones are compared. Therefore, it cannot serve as a reliable surrogate for clone size in polyclonal samples. Parenthetically, a similar conclusion was likewise reached in

a recent publication suggesting a non-quantitative nature of the LM-PCR method[1].

Reference

1 Brugman, M. H. *et al.* Evaluating a ligation-mediated PCR and pyrosequencing method for the detection of clonal contribution in polyclonal retrovirally transduced samples. *Hum Gene Ther Methods* **24**, 68-79, doi:10.1089/hgtb.2012.175 (2013).