Supporting Information

Breakpoints characterization

We generated a bed file containing all the breakpoint positions from our cohort (Supp. Table S3). We considered that breakpoints separated by less than 1000 base pairs belonged to the same region.

To check if breakpoints are preferentially located in regions containing genes, OMIM genes or repetitive elements, we first generated bed files containing the coordinates of all genes from the GRCh37 genome using the RefGene database, OMIM genes database and RepeatMasker database for repeated regions (https://genome.ucsc.edu/, 28/03/2018). We annotated the breakpoint positions from our cohort using these bed files. We then calculated the proportion of breakpoints falling into a gene, OMIM gene and reapeated region. These proportions were compared to the expected proportions according to the reference files from RefGene, OMIM and RepeatMasker.

Identification of SVs with linkedSV

DNA from patients 1 to 16 was analyzed with linkedSV (https://github.com/WGLab/LinkedSV, Li et al., bioRxiv) using alignments generated by LongRanger as input. LinkedSV uses information from barcodes and reconstructed linked-reads to identify candidate SV regions and quantify the evidence using a novel probabilistic model. Breakpoints are then refined with short-read information such as discordant read pairs and split reads. The mean numbers per sample for candidate deletions, duplications, inversions and translocations identified by linkedSV are 66, 34, 62 and 78 respectively.

Comparison between linkedSV and LongRanger SV calls

We performed the comparison between linkedSV (LK) and LongRanger (LR) SV calls for deletions, duplications and inversions (Supp. Figure S5) with bedtools intersect (Quinlan and Hall 2010; Quinlan 2014). For this, we set parameters of "–r" and "–f 0.7", meaning that regions of compared SVs, defined by start and end positions, have to share a reciprocal (-r) overlap (-f) equal to 70% at least.

Since we focused on large SVs (>=30kb) in the main analysis, we only reported comparison for this range of SVs. LongRanger and linkedSV both extracted a robust set of SVs from the candidate list, and there is a high likelihood that these are genuine SVs. Next, we performed the comparison for candidate and robust calls (Figure S5 A, B, C and D).

When considering all the comparisons, we observed that: i) the number of LR candidate calls was far greater than that of LK candidate calls for duplications (Supp. Figure S5-B-1) and inversions (Supp. Figure S5-C-1), and ii) the number of LK robust calls was greater than that of LR robust calls for all SV types. In addition, most of the LR deletions and inversions were included in the LK deletions and inversions (Supp. Figure S5-A-2 and Supp. Figure S5-C-2). In conclusion, LK provides equal or smaller lists of SVs candidate calls compared to LR, and provides less stringent robust SV call shortlists.

We also compared the linkedSV calls with the expected events for all patients (Supp. Table S4). Despite the fact that most of the events were considered as robust, linkedSV failed to detect all the events detected by the two first strategies.

All the samples seem to have the same pattern of SV calling except for two patients, 4 and 5, which exhibit a small number of LK deletions and LR inversions. In addition, for these two SV types, there are no common calls between LR and LK. The genome of patient 4 is known to contain CCR, which could explain the less efficient SV calling. However, there are no such known chromosomal alterations for B00I2FB.

Quinlan AR. 2014. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Current protocols in bioinformatics / editoral board, Andreas D Baxevanis [et al]* **47**: 11 12 11-11 12 34.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)* **26**: 841-842.