**Figure S1**

### a Unrearranged locus vs. Rearranged locus

- **Unrearranged locus**
  - Concordant read pairs

- **Rearranged locus**
  - Cluster of discordant read pairs, "bundle"
  - Concordant read pairs

### b Physical and sequence coverage

- Cov\textsubscript{physical} = Cov\textsubscript{sequence} * L\textsubscript{insert} / (2 * L\textsubscript{read})
- Cov\textsubscript{physical}(Cov\textsubscript{sequence} = 30, L\textsubscript{insert} = 500) = 75
- Cov\textsubscript{physical}(Cov\textsubscript{sequence} = 30, L\textsubscript{insert} = 3500) = 525

### c Genomic region and physical coverage

- Large-insert mate-pairs (3000-4000 bps)
- Large insert footprints (2000-5000 bps)
- Conventional paired-end reads (300-500 bps)
- Conventional read footprints (500-800 bps)
Sheared genomic DNA (3 - 4 Kb)

Ligation of CAP adapters

Gel size-selection (3-4 kb)

Selective DNA digest (T7 exonuclease and S1 nuclease)

Biotin selection

Circularization

Biotin

Internal adapter

Nick translation (E. coli DNA Pol I)

Ligation of paired-end adapters

PCR amplification

Paired-end sequencing (2x100 bp)

Stripping adapter sequence

Mapping to the reference

3.5 Kb
Deletion
dir(A) = +
dir(B) = +
coord(A) < coord(B)

Small insertion
dir(A) = +
dir(B) = +
coord(A) < coord(B)

Tandem duplication
dir(A) = +
dir(B) = +
coord(A) < coord(B)

Inversion
dir(A) = +
dir(B) = -
coord(A) < coord(B)
dir(C) = -
dir(D) = +
coord(C) < coord(D)
coord(A) = coord(C)
coord(B) = coord(D)

Reciprocal translocation
dir(A) = dir(D)
dir(C) = dir(B)
coord(A) = coord(D)
coord(C) = coord(B)

Duplicative translocation
coord(A) = coord(D)
dir(A) = dir(D)
dir(B) = dir(C)
<table>
<thead>
<tr>
<th>SV_num</th>
<th>Breakpoint_id</th>
<th>STT5989T (fresh frozen NPC5989)</th>
<th>STT5753T (FFPE NPC5989)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SV1</td>
<td>B_8_128-8_128</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>SV2</td>
<td>B_1_198-1_203</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>SV3</td>
<td>B_1_197-1_198</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>SV8</td>
<td>B_1_197-1_203</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>SV4</td>
<td>B_1_154-8_128</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>SV5</td>
<td>B_11_95-11_102</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>SV6</td>
<td>B_11_101-11_102</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>SV7</td>
<td>B_11_95-11_101</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>SV9</td>
<td>B_2_66-2_66</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>SV10</td>
<td>B_1_155-1_155</td>
<td>+</td>
<td>-</td>
</tr>
</tbody>
</table>
### Figure S5

<table>
<thead>
<tr>
<th>SV_num</th>
<th>Breakpoint_id</th>
<th>Mate-pairs</th>
<th>TC (from MPs)</th>
<th>TC (from gel)</th>
<th>TC (from qPCR)</th>
<th>TC (from cnv)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Tandem duplication</strong> SV1</td>
<td>B_8_128-8_128</td>
<td>212</td>
<td>72%</td>
<td>92%, 72%</td>
<td>33%, 37%</td>
<td></td>
</tr>
<tr>
<td><strong>Coupled inversion</strong> SV2</td>
<td>B_1_198-1_203</td>
<td>216</td>
<td>88%, 88%</td>
<td>21%, 56%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SV3</td>
<td>B_1_197-1_198</td>
<td>199</td>
<td>56%</td>
<td>91%</td>
<td>104%, 81%</td>
<td></td>
</tr>
<tr>
<td>SV8</td>
<td>B_1_197-1_203</td>
<td>75</td>
<td>81%, 49%</td>
<td>80%, 39%</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Duplicative translocation</strong> SV4</td>
<td>B_1_154-8_128</td>
<td>204</td>
<td>70%</td>
<td>34%</td>
<td>42%</td>
<td>56%, 64%</td>
</tr>
<tr>
<td><strong>Coupled inversion</strong> SV5</td>
<td>B_11_95-11_102</td>
<td>155</td>
<td>71%, 81%</td>
<td>50%, 25%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SV6</td>
<td>B_11_101-11_102</td>
<td>157</td>
<td>52%</td>
<td>34%, 77%</td>
<td>104%, 44%</td>
<td></td>
</tr>
<tr>
<td>SV7</td>
<td>B_11_95-11_101</td>
<td>141</td>
<td>81%, 44%</td>
<td>32%, 65%</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Deletion</strong> SV9</td>
<td>B_2_66-2_66</td>
<td>184</td>
<td>63%</td>
<td>137%, 141%</td>
<td>66%, 54%</td>
<td></td>
</tr>
<tr>
<td><strong>Deletion</strong> SV10</td>
<td>B_1_155-1_155</td>
<td>145</td>
<td>50%</td>
<td>58%</td>
<td>47%</td>
<td></td>
</tr>
</tbody>
</table>
Figure S6

a

Control primer L  Bp primer2

Breakpoint

Bp primer1  Control primer R

b
**SUPPLEMENTAL FIGURE LEGENDS**

Supplemental Figure 1 (a) Examples of structural variants and discordant read pairs. Unrearranged locus (left part of the figure) has only concordant read pairs, which map within the expected distance of each other. This distance is set by the library insert size distribution. The rearranged locus (right part of the figure) generates both concordant and discordant read pairs. Inserts spanning the breakpoint (position where blue and green genomic regions are fused) produce discordant read pairs. Other inserts from the adjacent regions (blue and green) are concordant since they do not cross the breakpoint. Provided sufficient genome coverage, breakpoints are represented by multiple independent read pairs originating from DNA inserts that span that breakpoint. Multiple discordant read pairs representing a single breakpoint are termed “read bundle”. (b) Relationship between sequence coverage and physical coverage. A genomic region (depicted as a blue bar) is sampled by multiple inserts represented by read pairs that map to that region (depicted as narrow blue bars linked by the dashed purple lines, see the bottom of the panel). Sequence coverage (blue curve) represents the number of reads (sequenced portion of the insert, narrow blue bars) that span any given base pair. Physical coverage (blue curve at top of panel) represents the number of inserts (reads as well as unsequenced part of the insert depicted by the dashed purple line) that cover any given base pair. The mathematical formulas in the right part of the panel provide the dependence between the average genome-wide physical coverage and average sequence coverage. The two examples of the physical coverage calculation (75 and 525) provide an illustration of how physical coverage increases when using larger inserts, even when sequence coverage is unchanged. (c) Illustration of differences associated with using short insert fragment libraries and long insert mate pair libraries. Not only is the physical coverage less with shorter inserts, but the breakpoint-associated read pairs have different distributions as well. In particular, reads from short inserts occupy narrow regions (referred to as ‘footprints’) next to the breakpoint. Reads from the large-insert library have a significantly bigger footprint.

Supplemental Figure 2. Large insert mate-pair library construction. Genomic DNA is first sheared to the desired size range (3-4 Kb), followed by ligation of CAP adapters that lack the 5’ phosphate group at the end of the short oligo. DNA fragments are then circularized via ligation of internal biotin-labeled adaptor. Circular constructs contain single-stranded nicks at the sites of ligation due to missing 5’ phosphate group. Nicks are enzymatically moved in 5’ to 3’ direction inside the DNA insert. T7 exonuclease recognizes the nicks and digests the nicked DNA in the 5’ to 3’ direction. The exposed ssDNA is then digested using S1 nuclease. Undigested biotin-labeled DNA is isolated using magnetic streptavidin beads. Illumina sequencing adapters are ligated onto the ends of the purified fragments on beads and the library DNA is enriched by PCR and sequenced using the paired-end illumina HiSeq protocol (2×101 bp). Sequencing reads are analyzed to strip the internal adapter sequence and the resulting sequencing fragments are mapped to the human reference genome (hg19) and the EBV genome sequence. The resulting
mapped read pairs are spaced by the average size of the library insert (3.5 Kb) and point outwards.

**Supplemental Figure 3.** Common types of structural variants. Structural variants can be represented by one or more breakpoints. (a) A somatic deletion is represented by a single breakpoint joining the donor site (A) and acceptor site (B) flanking the deleted region (green bar). The discordant tumor sample read pairs supporting the deletion are represented by arrows joined by dashed lines. The arrow colors represent reference regions to which the reads map. In case of deletion, the observed insert size is significantly larger compared to the expected insert size. (b) Small insertion is an insertion that is smaller than the typical library insert size, and is represented by a single breakpoint. (c) Tandem duplication. (d) Representations of inversion, and the two associated breakpoints. (e) Unbalanced (duplicative) translocation, (f) balanced translocation, and (g) duplication.

**Supplemental Figure 4.** Structural variant comparison within different parts of the NPC5989 tumor tissue. Two parts of the NPC5989 tissue were analyzed using primers against breakpoints detected by SMASH. STT5989T represents DNA from NPC5989 frozen tissue used for whole-genome sequencing and variant discovery. STT5753T represents a different, archival, part of the same tumor tissue and likely represents an earlier stage of the tumor because of a number of structural variants that are not detectable compared to the frozen tissue. However, the coupled inversion that produced the YAP1-MAML2 fusion is detectable in both samples, suggesting that this event is more likely to be a driver mutation compared to other structural variants that emerge later. A 25 Kb Chr2 deletion is also present in both fresh and FFPE NPC5989 samples, suggesting that it also occurred early in the tumor development. This variant deletes uncategorized gene AK131224 and partially deletes uncategorized gene FLJ16124.

**Supplemental Figure 5.** Summary of tumor content assessment. For each of the 10 breakpoints (SV1-10), and corresponding structural variants, SMASH reported a certain number of supporting mate-pair reads. Based on these numbers, and the expected physical coverage, we assessed the tumor content (TC from MPs column). The tumor content for the coupled inversions was obtained by averaging across the three underlying breakpoints. The ‘TC (from gel)’ column represents assessed tumor content based on breakpoint-PCR experiment and measuring the band intensity on the agarose gel (Fig. S6, Supp. Table 5). The ‘TC (from qPCR)’ column represents assessed tumor content based on breakpoint-qPCR experiment (Supp. Table 6). The ‘TC (from CNV)’ column represents assessment of tumor content based on increase of dosage of telomeric end of Chr1 and decrease of telomeric end of Chr8 dosage resulting from the duplicative translocation (Fig. 2).

**Supplemental Figure 6.** Breakpoint-PCR experiment. To assess tumor content using PCR, we designed four primers for each breakpoint (lanes SV1-10). (a) Two
primers were on both sides of the breakpoint, and are expected to produce a 'specific' breakpoint PCR product. Each of the two breakpoint primers were also paired with the control primer, which is expected to produce a control PCR product from the unrearranged allele. (b) The four primers were used together in a competitive PCR reaction, for which we used both the tumor sample (lane ‘T’) and normal sample (lane ‘N’), which we used to distinguish the ‘specific’ band. Lane ‘W’ represents PCR product where we used water instead of sample DNA. All breakpoints SV1-10 were amplified this way, PCR products run on the gel, specific and control bands identified, and their fluorescent intensities were measured using densitometer. The relative amounts of specific and control bands were detected based on the measured band intensities (Supp. Table T5)