# **Supplementary Note**

# **Data acquisition, whole genome sequencing alignment and sample selection**

Short-read whole-genome sequencing data were gathered from 2,955 tumors and matched blood normals from projects within the International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA). The 2,955 tumors were obtained from 2,834 donors, with some donors contributing multiple samples, usually in the setting of resequencing of recurrent tumors. We used the PCAWG-wide list of preferred samples to select a single sample from each donor (see PCAWG-Tech paper). An additional 141 samples were excluded based on poor sequence quality, as determined by the PCAWG-QC group. Based on the robustness of the SV calls as reviewed by the PCAWG-6 SV working group, we opted to retain the excluded samples from the TCGA LAML cohort. These samples , as they were excluded from SNV analyses because of read direction biases from PCR artifacts, but were not found to have poor rearrangement calls based on poor SNV calls but not poor rearrangement calls. The final number of unique genomes that we analyzed for rearrangements was 2,693 (**Supp. Fig. 1a**). Of the 2,693 unique donors in our cohort, 1,242 were sequenced by RNA-seq (**Supp Fig 1b**). The RNA-seq data was aligned to the reference and filtered as described in the Campbell *et al* PCAWG marker paper..

# **Detection of rearrangements from whole genome sequencing data**

## *SvABA*

SvABA identified rearrangements by performing genome-wide *de novo* local assembly of across every 25 kb bin in the genome[(Wala et al. 2017)](https://paperpile.com/c/Jfp1ce/V6vU). Briefly, assembly within SvABA was performed in each assembly window by SGA3. Only reads with a clipped alignment, an unmapped pair-mate, a gapped alignment, an insert size greater than three standard deviations from the mean, or pair-mate mapped to another chromosome were assembled. The assembled contigs were re-aligned to the reference with BWA-MEM[(Li and Durbin 2009)](https://paperpile.com/c/Jfp1ce/MrNT). Contigs that realigned to the reference with a multi-part alignment were parsed to obtain candidate variants. Contigs with a low mapping quality or which contained only short alignments (< 50% the length of a read) were discarded. Sequencing reads from each assembly window were realigned to the contigs to identify the number of breakpoint-supporting reads from either the tumor or the normal sample.

SvABA detected both the number of bases of microhomology and the presence of novel insertion bases at each rearrangement junction. Junction insertions were obtained from contigs whose multi-part mappings to the reference left unaligned bases between the two mappings. Micro-homology estimates reflected the base pairs at the junctions of the two alignment fragments that were covered by each of the alignments, thus representing bases that could have originated from either end of the breakpoint.

## *BRASS*

The Sanger rearrangement calls were generated using an in-house algorithm Brass, which clusters abnormally paired read pairs into SV calls. The raw SV call set was improved by correcting for skewed insert size distribution, removing fold-back-type read pair artefacts (Tsai et al., 2014, Summarizing specific profiles in Illumina sequencing from whole-genome amplified DNA), removing candidate SVs consistent with mismapped properly paired read pairs (Papaemmanuil E et al., Nat Genet 2014), mismapped bacterial or viral DNA and removing SV calls without a consistent somatic copy number change.

## *DELLY*

DELLY4 v.0.6.8 performed integrated paired-end analysis and split-read detection using paired tumor and normal samples, as previously described5. A high-stringency SV set of germline and somatic calls were derived by requiring at least four supporting pairs, with the additional requirement for split read support for SVs smaller than 500 bp. Somatic SVs were filtered for absence in the paired control and additionally filtering for SVs detected in ≥1% of a set of 1,105 germline samples from healthy individuals belonging to phase I of the 1000 Genomes Project6 and a panel of normal samples, constructed from the collected germline SVs of the PCAWG set. Multi-tumor sample analyses were analyzed jointly together with the paired normal sample to improve genotyping, and subsequently split into individual tumor samples.

## *dRanger*

Somatic rearrangements are detected by the dRanger+Breakpointer tool[(Drier et al. 2013)](https://paperpile.com/c/Jfp1ce/Y7yo). The initial dRanger step requires a cluster of at least two high mapping quality supporting read pair fragments in the tumor sample. dRanger also used a Panel of Normals of 318 WGS samples to reject sites of common artifacts. The list of candidate rearrangements is then passed to Breakpointer, which scans for additional supporting split read evidence to confirm breakpoint junctions in the tumor sample and to reject candidate somatic rearrangements with evidence in of the rearrangement appearing in matched normal samples. We required a combined total of at least 4 supporting reads, either read pairs or split reads, following breakpointer. (Chip)

# **Filtering and merging of somatic rearrangements**

## *Merging rearrangements from SvABA, DELLY, BRASS and dRanger*

Somatic SVs from SvABA, DELLY, BRASS and dRanger were combined into a union set. For each tumor/normal pair, high confident SVs calls from each caller were pairwise joined, allowing 200 bp of slop at the breakpoints, and requiring same strandedness. Calls were merged using a graph structure, by inserting an edge in the graph for each joined pair. High confidence merged calls were derived by requiring at least two out of the four callers to support an SV. For each merged SV, consensus breakpoint positions were chosen based on proximity to the consensus breakpoint.

A total of 292,253 rearrangements were included in the final call set, with the majority of rearrangements coming from breast and ovarian adenocarcinoma and liver hepatocellular carcinoma (**Supp. Fig. 2a**). For calls made by either SvABA or DELLY, the breakpoints were further annotated by their degree of breakpoint microhomology. The microhomology calls from SvABA were highly correlated (R2: 0.923) with the calls from DELLY (**Supp. Fig. 2b).** The vast majority of calls (77%) were identified by three or more variant detection tools (**Supp. Fig 2c**). The total number of rearrangements per genome may be confounded by under-calling rearrangements in tumors with low purity. However, controlling for histological subtype, we observed no significant correlation between the level of tumor cellularity and the number of rearrangements identified (**Supp. Fig. 2d**).

The most highly rearranged single genome, a leiomyosarcoma , contained 2,064 rearrangements. Of the 2,963 genomes, 171 contained no rearrangements. Samples with no rearrangement calls were highly enriched for hematological and brain tumors and pancreatic endocrine tumors and thyroid adenocarcinomas. Acute myeloid leukemias exhibited the highest rate of rearrangement-silent genomes, with 21 of the 45 (46.7%) containing no rearrangements, followed by thyroid adenocarcinoma (37.5%) and pancreatic endocrine tumors (17.6%).

## *Filtering of germline copy number variants and class switch recombination rearrangements*

Germline copy number variants (CNVs) can be misclassified as somatic in cases where there is low sequencing coverage in the paired normal or due to stochastic sampling effects of heterozygous germline variants. To reduce the impact of this, we compared the merged somatic rearrangement calls from each tumor with germline rearrangement calls from SvABA from all normal samples in the dataset. Any putative somatic rearrangement which overlapped on both ends (+/- 8 bp) with four or more germline samples from other patients was reclassified as germline (**Supp. Fig. 3a)**. The germline filter removed 187 rearrangements (median size: 3,308 bp), representing less than 0.058% of the somatic callset.

 Somatic rearrangements contained within the three immunoglobulin loci (IGH, IGK, IGL) reflect physiological class-switch recombination rearrangements in lymphoid tissue rather than cancer rearrangements. We filtered these rearrangements from our analyses to focus only on tumor-specific rearrangements, and to prevent physiologic somatic rearrangements from rising to the top of the cancer-specific analysis of significantly recurrent breakpoints. We obtained the coordinates for the immunoglobulin loci from NCBI and added 1 Mbp padding to each to be certain to capture all class switch recombination events. We observed 503 somatic rearrangements (0.17% of rearrangements) where both breakpoints were contained within the three immunoglobulin loci. These were excluded from further analysis, but were not removed from the consensus somatic rearrangement VCF files (**Supp Fig. 3b**). Among the excluded rearrangements, 75.3% were from a cancer of lymphoid origin. This procedure retained rearrangements where one breakpoint was contained in an immunoglobulin locus, but the partner breakpoint was elsewhere (e.g. *IGH-BCL2*).