Whole Genome Sequencing and Analysis
Acute Lymphoblastic Leukemia Phase I (ALL P1)

Illumina genomic plate-based library construction (350-450bp insert size):
2ug of genomic DNA in a 96-well format was fragmented by Covaris E210 sonication for 30 seconds using a "Duty cycle" of 20% and "Intensity" of 5. The paired-end sequencing library was prepared following the BC Cancer Agency's Genome Sciences Centre 96-well Genomic ~350bp-450bp insert Illumina Library Construction protocol on a Biomek FX robot (Beckman-Coulter, USA). Briefly, the DNA was purified in a 96-well microtitre plate using Ampure XP SPRI beads (40-45uL beads per 60uL DNA), and was subject to end-repair, and phosphorylation by T4 DNA polymerase, Klenow DNA Polymerase, and T4 polynucleotide kinase respectively in a single reaction, followed by cleanup using Ampure XP SPRI beads and 3' A-tailing by Klenow fragment (3' to 5' exo minus). After cleanup using Ampure XP SPRI beads, picogreen quantification was performed to determine the amount of Illumina PE adapters used in the next step of adapter ligation reaction. The adapter-ligated products were purified using Ampure XP SPRI beads, then PCR-amplified with Phusion DNA Polymerase (Thermo Fisher Scientific Inc. USA) using Illumina's PE indexed primer set, with cycle conditions: 98˚C for 30sec followed by 6 cycles of 98˚C for 15 sec, 62˚C for 30 sec and 72˚C for 30 sec, and a final extension at 72˚C for 5min. The PCR products were purified using Ampure XP SPRI beads, and checked with Caliper LabChip GX for DNA samples using the High Sensitivity Assay (PerkinElmer, Inc. USA). PCR product of the desired size range was gel purified (8% PAGE or 1.5% Metaphor agarose in an in-house custom built robot), and the DNA quality was assessed and quantified using an Agilent DNA 1000 series II assay and Quant-iT dsDNA HS Assay Kit using Qubit fluorometer (Invitrogen), then diluted to 8nM. The final concentration was confirmed by Quant-iT dsDNA HS Assay prior to generating 100bp paired end reads on the Illumina HiSeq 2000/2500 platform using v3 chemistry.

WGS/hg19 alignment:
Illumina paired-end whole genome sequencing reads were aligned to the hg19 reference using BWA version 0.5.7. This reference contains chromosomes 1-22, X, Y, MT, 20 unlocalized scaffolds and 39 unplaced scaffolds. Multiple lanes of sequences were merged and duplicated reads were marked with Picard Tools.

Structural variant detection
Was performed using ABySS (v1.3.2) and trans-ABySS (v1.4.6). For RNA-seq assembly alternate k-mers from k50-k96 were performed using positive strand and ambiguous stand reads as well as negative strand and ambiguous strand reads. The positive and negative strand assemblies were extended where possible, merged and then concatenated together to produce a meta-assembly contig dataset. The genome (WGS) libraries were assembled in single end mode using k-mer values of k24, and k44. The contigs and reads were then reassembled at k64 in single end mode and then finally at k64 in paired end mode. The meta-assemblies were then used as input to the trans-ABySS analysis pipeline (Robertson et al., 2010).

Large scale rearrangements and gene fusions from RNA-seq libraries were identified from contigs that had high confidence GMAP (v2012-12-20) alignments to two distinct genomic

regions. Evidence for the alignments were provided from aligning reads back to the contigs and from aligning reads to genomic coordinates. Events were then filtered on read thresholds. Large scale rearrangements and gene fusions from WGS libraries were identified in a similar way, but using BWA (v0.6.2-r126) alignments.

Insertions and deletions were identified by gapped alignment of contigs to the human reference using GMAP for RNA-seq and BWA for WGS. Confidence in the event was calculated from the alignment of reads back to the event breakpoint in the contigs. The events were then screened against dbSNP and other variation databases to identify putative novel events.

To determine compartment specific events the structural variant calls for each patient from all matched genome and RNA-seq samples were concatenated together and screened against matching genome tumour, and where available germline bam files. This resulted in compartment specific structural variant events and where germline was available putative somatic and germline events. The events were further filtered against a compendium of germline structural variants to remove recurrent false positives.

Genomic SNV analyses

SNVs from WGS-seq data were analyzed using all three methods described below:

Mpileup

SNVs were analyzed with SAMtools mpileup v.0.1.17 (Li et al., 2009) either on single or paired libraries. Each chromosome was analyzed separately using the -C50-DSBuf parameters. The resulting vcf files were merged and filtered to remove low quality SNVs by using samtools varFilter (with default parameters) as well as to remove SNVs with a QUAL score of less than 20 (vcf column 6). Finally, SNVs were annotated with gene annotations from ensembl v66 using snpEff (Cingolani et al., 2012b) and the dbSNP v137 db membership assigned using snpSift (Cingolani et al., 2012a).

Strelka

To analyze compartment specific SNVs, samples were analyzed pair wise with the default settings of Strelka v0.4.7 (Saunders et al., 2012). Primary tumor samples and relapse/met were compared against the germline sample. In the absence of a germline sample, the relapse/met samples were compared against the primary tumor sample.

MutationSeq

SNVs were analyzed pair wise with SAMtools mpileup v.0.1.17 (Li et al., 2009). Each chromosome was analyzed separately using the -C50-DSBuf parameters. Before merging the resulting vcf files, they were filtered to remove all indels and low quality SNVs by using samtools varFilter (with default parameters) as well as to remove SNVs with a QUAL score of less than 20 (vcf column 6). The SNVs in the resulting vcf files were further filtered and scored using mutationSeq v1.0.2 and annotated with gene annotations from ensembl v66 using snpEff (Cingolani et al., 2012b) and the dbSNP v137 and cosmic 64 db membership using snpSift (Cingolani et al., 2012a).

Copy number variation (CNV) analysis

The techniques outlined in (Jones et al., 2010) were followed to analyze copy number changes. Sequence quality filtering was used to remove all reads of low mapping quality (Q < 10). Due to the varying amounts of sequence reads from each sample, aligned reference reads were first used to define genomic bins of equal reference coverage to which depths of alignments of sequence from each of the tumor samples were compared. This resulted in a measurement of the relative number of aligned reads from the tumors and reference in bins of variable length along the genome, where bin width is inversely proportional to the number of mapped reference reads. A hidden Markov model (HMM) was used to classify and segment continuous regions of copy number loss, neutrality, or gain using methodology outlined previously (Shah et al., 2006). The five states reported by the HMM were: loss (1), neutral (2), gain (3), amplification (4), and high-level amplification (5).

Amplified and deleted CNV regions are further screened for interspersed repeats, and low complexity DNA sequences, which includes long interspersed nuclear elements (LINE), short interspersed nuclear element (SINE), long terminal repeat elements (LTR), DNA repeat elements (DNA), low complexity repeats, satellite repeats, simple repeats (micro-satellites), and RNA repeats (including RNA, tRNA, rRNA, snRNA, scRNA, srpRNA).

Repeat sequences in the genome pose challenges in the identification of CNVs with next generation sequencing data as the short reads sequenced from repetitive regions cannot be mapped unambiguously. Exclusion or random placement of the reads aligned to multiple regions can either reduce sensitivity of CNV detection or result in the identification of false deletions in repeated regions. Due to the limitations of both alignment and subsequent segmentation algorithms, CNVs called in the regions harboring highly repeated sequences should be carefully scrutinized. Therefore, in addition to focal CNV functional annotation, recurrence among patients, and presence of TransAbyss overlapping events, the number and types of repeats are added to the annotation of candidate CNVs to further narrow down the prioritized list for verification. It is recommended that the candidate CNVs be prioritized based on the presence of genes of interest, high recurrence among patients, presence of overlapping TransAbyss events, and low frequency or absence of repeat nuclear elements.