mRNA Sequencing

Acute Myeloid Leukemia (AML) and AML Induction Failure (AML-IF)

*Protocols Performed at British Columbia Cancer Agency.

RNA-seq lite library construction (AML-IF):
For each sample, approximately 10ng of total RNA was processed using the SMART(TM) cDNA synthesis protocol including SMARTScribe Reverse Transcriptase (Clontech, #639536). This method deploys a modified oligo(dT) primer to prime the first strand synthesis reaction and a template switching mechanism to generate full-length single-stranded cDNAs containing the complete 5' end of the mRNA as well as universal priming sequences for end-to-end amplification during 20 cycles of PCR. The amplified cDNA was subject to Illumina paired-end library construction using NEBNext paired-end DNA sample Prep Kit (NEB, E6000B-25). Libraries were sequenced on Illumina HiSeq2000 instruments.

Strand-specific RNA-seq (plate based) library construction:
Total RNA samples were checked using an Agilent Bioanalyzer RNA nanochip or Caliper GX HT RNA LabChip, and samples passing quality control were arrayed into a 96-well plate. PolyA+ RNA was purified using the 96-well MultiMACS mRNA isolation kit on the MultiMACS 96 separator (Miltenyi Biotec, Germany) from 2ug total RNA with on-column DNaseI-treatment as per the manufacturer's instructions. The eluted PolyA+ RNA was ethanol precipitated and resuspended in 10µL of DEPC treated water with 1:20 SuperaseIN (Life Technologies, USA).

First-stranded cDNA was synthesized from the purified polyA+RNA using the Superscript cDNA Synthesis kit (Life Technologies, USA) and random hexamer primers at a concentration of 5µM along with a final concentration of 1ug/uL Actinomycin D, followed by Ampure XP SPRI beads on a Biomek FX robot (Beckman-Coulter, USA). The second strand cDNA was synthesized following the Superscript cDNA Synthesis protocol by replacing the dTTP with dUTP in dNTP mix, allowing second strand to be digested using UNG (Uracil-N-Glycosylase, Life Technologies, USA) in the post-adapter ligation reaction and thus achieving strand specificity.

The cDNA was quantified in a 96-well format using PicoGreen (Life Technologies, USA) and VICTOR^{3}V Spectrophotometer (PerkinElmer, Inc. USA). The cDNA was fragmented by Covaris E210 sonication for 55 seconds at a "Duty cycle" of 20% and "Intensity" of 5. The paired-end sequencing library was prepared following the BC Cancer Agency Genome Sciences Centre strand-specific, plate-based and paired-end library construction protocol on a Biomek FX robot (Beckman-Coulter, USA). Briefly, the cDNA was purified in 96-well format using Ampure XP SPRI beads, 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 purification using Ampure XP SPRI beads, picogreen quantification was performed to determine the amount of Illumina PE adapters to be used in the next step of adapter ligation reaction. The adapter-ligated products were purified using Ampure XP SPRI beads, and digested with UNG (1U/ul) at 37^{o}C for 30 min followed by deactivation at 95^{o}C for 15 min. The digested cDNA was

purified using Ampure XP SPRI beads, and then PCR-amplified with Phusion DNA Polymerase (Thermo Fisher Scientific Inc. USA) using Illumina's PE primer set, with cycle condition 98˚C 30sec followed by 10-13 cycles of 98˚C 10 sec, 65˚C 30 sec and 72˚C 30 sec, and then 72˚C 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 purified using 8% PAGE, 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 library concentration was double checked and determined by Quant-iT dsDNA HS Assay again for Illumina Sequencing.

RNA-Seq/hg19 read alignment:
Illumina paired-end RNA sequencing reads were aligned to GRCh37-lite genome-plus-junctions reference using BWA version 0.5.7. This reference combined genomic sequences in the GRCh37-lite assembly and exon-exon junction sequences whose corresponding coordinates were defined based on annotations of any transcripts in Ensembl (v59), Refseq and known genes from the UCSC genome browser, which was downloaded on August 19 2010, August 8 2010, and August 19 2010, respectively. Reads that mapped to junction regions were then repositioned back to the genome, and were marked with 'ZJ:Z' tags. BWA is run using default parameters, except that the option (-s) is included to disable Smith-Waterman alignment. Finally, reads failing the Illumina chastity filter are flagged with a custom script, and duplicated reads were flagged 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.

SNV analysis of strand-specific RNA-seq data:
After repositioning, hg19-aligned BAM files were split into positive-fragment and negative-fragment BAM files based on the orientation of the paired-end reads. Unmapped and improperly paired aligned reads were put into the mix-fragment BAM. SNVs were then detected on positive- and negative-split BAMs separately using SNVMix2 (Goya et al., 2010) with parameters Mb and Q30. The SNVs were further filtered to exclude those called based on 1) reference base N; 2) only 1 read supports the variant; 3) probability of heterozygous and homozygous of variant allele smaller than 0.99; 4) a position overlapping with insertions or deletions; 5) read supports from positions no more than 5 bases from read ends; 6) supports from reads only spanning an exon-exon junction; 7) more than 0.5 proportion of supporting reads were improper paired; 8) fewer than 2 proper-paired supporting reads. SNVs located in exons equal or smaller than the read length, 100bp in this case, are a special case, because all their coverage may come from exon-exon junction spanning reads, so we also identified small-exonic SNVs that ware only supported by reads that spanning exon-exon junction but passed all other 7 filtering criteria mentioned above. These SNVs were finally annotated with SnpEff (Cingolani et al., 2012b) (Ensembl 66) and SnpSift (Cingolani et al., 2012a) (dbSNP137 and COSMIC64).

mRNA-Differential expression:
We used SAMseq (samr v2.0, R 2.15.0) two-class unpaired analyses with an FDR threshold of 0.05 to identify genes that were differentially expressed. For each run on a pair of sample groups, we first reduced the number of genes by removing those with median less than 5 RPKM in both groups, and those for which the Wilcoxon BH adjusted P-value between the two groups was greater than 0.05. This subset of genes was submitted to SAMseq. Each run generated a pair of files: genes 'up' and 'down'. We then ranked the genes by a median-based fold change, and generated a figure showing up to 10 of the largest fold changes in each direction.

mRNA-NMF:
For specific mRNA-Seq expression datasets, we first removed genes expressed at or below a noise threshold of ≤ 0.2 reads per kilobase (of gene model) per million mapped reads (RPKM) in at least 75% of samples. We created the NMF input matrix using the top 25% most-variant genes, by ranking expressed genes having a mean RPKM of at least 10 by the coefficient of variation. We generated consensus clustering results with NMF v0.5.02 in R v1.12.0, with the default Brunet algorithm, and 200 iterations for the clustering run. Rank survey profiles for cophenetic and silhouette width suggest a specific cluster solution.