Bioinformatics methods

Filtering of sequencing reads. Internally developed software termed SOAPnuke was used to filter reads as follows: Reads with adapters were removed; ii) reads in which unknown bases (N) constituted ≥5% of the sequence were removed; and iii) low quality reads (defined as reads containing >20% adapter sequences, high content of unknown bases and/or low quality bases) were removed. After filtering, the remaining reads were termed 'clean reads' and stored in the FASTQ format (1). Details on the software and parameters: SOAPnuke version 1.5.2; parameters, -l 15 -q 0.2 -n 0.05; available from github.com/BGI-flexlab/SOAPnuke.

Genome mapping. Hierarchical Indexing for Spliced Alignment of Transcripts (HISAT) was used for gene mapping (2). HISAT aligns the DNA or RNA sequencing reads to the human reference genome as well as a single reference genome to predict the genomic or the transcriptomic profile of the sequenced nucleic acid. Details on the software and parameters: HISAT2 version 2.0.4; parameters: -phred64 -sensitive-no-discordant-no-mixed-I 1 -X 1000; available from ccb.jhu.edu/software/hisat.

Gene expression analysis. Clean reads were mapped to the reference genome using Bowtie2 (3), followed by calculating gene expression levels using RSEM (4), which is a software package for estimating gene and isoform expression levels from RNA-Seq data. The software allowed estimation of the transcript's relative abundance of genes and gene isoforms following genome mapping without the need for a reference genome. Bowtie2 software was used to align long sequencing reads (50->1,000 bp), such as the sequencing reads of the mammalian genome. A Pearson's correlation coefficient was then calculated between all samples using the 'cor' function, hierarchical clustering was performed using the 'hclust' function, and the diagrams were drawn using the 'ggplot2' functions in R. Hierarchical cluster analysis was used to gather similar genes together into groups (clusters). RNA sequencing data were clustered according to the expression levels to group the up and the downregulated genes. Details on the software and parameters: Bowtie2 version 2.2.5; parameters, -q -phred64 -sensitive -dpad 0 -gbar 99999999 -mp 1,1 -np 1 -score-min L,0,-0.1 -I 1 -X 1000 -no-mixed -no-discordant -p 1 -k 200; available from http://bowtie-bio.sourceforge.net/Bowtie2/index.shtml. RSEM version 1.2.12; parameters, default; available from deweylab.biostat.wisc.edu/RSEM.

Detection of differentially expressed genes (DEGs). DEGs were detected with PossionDis. PossionDis is based on the possion distribution, which was performed as described by Audic & Claverie (5). The software uses algorithms to screen and distribute the DEGs. Details on the software and parameters: PossionDis; parameters, fold change ≥2.00 and false discovery rate ≤0.001.

References

1. Cock PJ, Fields CJ, Goto N, Heuer ML and Rice PM: The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res 38: 1767-1771, 2010.
2. Kim D, Langmead B and Salzberg SL: HISAT: A fast spliced aligner with low memory requirements. Nat Methods 12: 357-360, 2015.
3. Langmead B and Salzberg SL: Fast gapped-read alignment with Bowtie 2. Nat Methods 9: 357-359, 2012.
4. Li B and Dewey CN: RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12: 323, 2011.
5. Audic S and Claverie JM: The significance of digital gene expression profiles. Genome Res 7: 986-995, 1997.

Figure S1. qPCR analysis of RHAG, SUCNR1 and TM4SF1 expression. Quantitative validation of the RNA sequencing results by qPCR confirmed the apparent gene expression changes observed for RHAG, SUCNR1 and TM4SF1 in the AML case when compared with the corresponding control. β-actin was used as the housekeeping control. Samples were run in triplicate sets and repeated independently twice. qPCR, quantitative PCR; RhAG, Rh associated glycoprotein; SUCNR1, succinate receptor 1; TM4SF1, transmembrane-4 L-six family member-1; AML, acute myeloid leukemia.

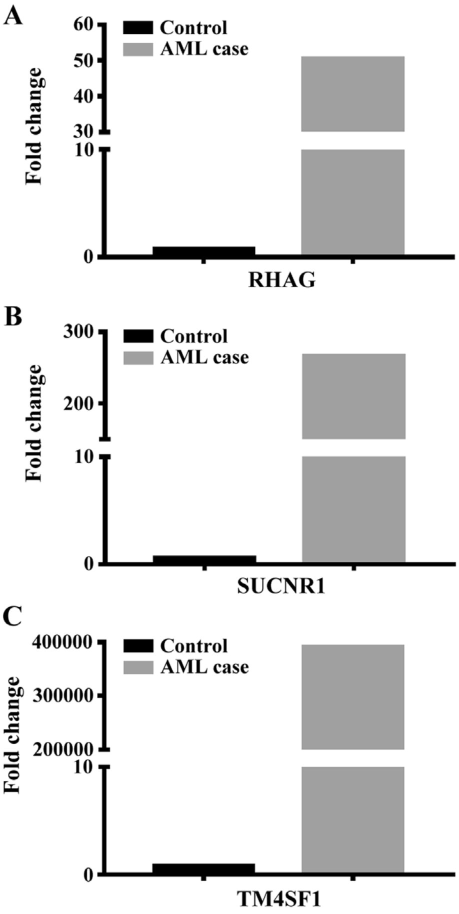

Table S1. Sequences of the primers.

RHAG	Sequence, 5'-3'
Forward	CGAGCAGCTCAACATCACCA
Reverse	TCATGAGGAAGCCAAACCCA
SUCNR1	
Forward	CTCTGCCCCTTGAAAAGCCT
Reverse	GAAGCGATCCTCACATTCCG
TM4SF1	
Forward	GAAAACCACCTCAGCCGCTT
Reverse	TCCTGTTCCAGCCCAATGAA

RhAG, Rh associated glycoprotein; SUCNR1, succinate receptor 1; TM4SF1, transmembrane-4 L-six family member-1.

Table SII. Summary of coding and non-coding transcripts in the patient with acute myeloid leukemia.

Type of transcript	n
Total novel transcripts	10,444
Coding transcripts	9,284
Non-coding transcripts	1,160
Novel splicing variants of known genes	8,033
Novel genes	1,251

Table SIII. Transition and transversion events in the patient with acute myeloid leukemia.

Type of event	Number of events
Total	156,172
Transition	113,011
A-G	56,480
C-T	56,171
Transversions	43,161
A-C	11,517
A-T	7,644
C-G	12,309
G-T	11,691