G&I Genomics & Informatics

pISSN 1598-866X eISSN 2234-0742
Genomics Inform 2013;11(1):46-51
http://dx.doi.org/10.5808/GI.2013.11.1.46

ORIGINAL ARTICLE

Genetic Risk Prediction for Normal-Karyotype Acute Myeloid Leukemia Using Whole-Exome Sequencing

Seong Gu Heo, Eun Pyo Hong, Ji Wan Park*

Department of Medical Genetics, Hallym University College of Medicine, Chuncheon 200-702, Korea

Normal-karyotype acute myeloid leukemia (NK-AML) is a highly malignant and cytogenetically heterogeneous hematologic cancer. We searched for somatic mutations from 10 pairs of tumor and normal cells by using a highly efficient and reliable analysis workflow for whole-exome sequencing data and performed association tests between the NK-AML and somatic mutations. We identified 21 nonsynonymous single nucleotide variants (SNVs) located in a coding region of 18 genes. Among them, the SNVs of three leukemia-related genes (*MUC4*, *CNTNAP2*, and *GNAS*) reported in previous studies were replicated in this study. We conducted stepwise genetic risk score (GRS) models composed of the NK-AML susceptible variants and evaluated the prediction accuracy of each GRS model by computing the area under the receiver operating characteristic curve (AUC). The GRS model that was composed of five SNVs (rs75156964, rs56213454, rs6604516, rs10888338, and rs2443878) showed 100% prediction accuracy, and the combined effect of the three reported genes was validated in the current study (AUC, 0.98; 95% confidence interval, 0.92 to 1.00). Further study with large sample sizes is warranted to validate the combined effect of these somatic point mutations, and the discovery of novel markers may provide an opportunity to develop novel diagnostic and therapeutic targets for NK-AML.

Keywords: acute myeloid leukemia, DNA sequence analysis, genetic variation, risk assessment

Introduction

Acute myeloid leukemia (AML), which is characterized by the rapid growth of abnormal myeloid cells in the marrow, is a highly malignant and cytogenetically heterogeneous type of cancer [1]. AML is categorized into three risk groups: favorable, intermediate, and unfavorable. AML with a normal karyotype (NK-AML) accounts for approximately 40% to 50% of adult patients and 25% of pediatric patients with AML and is composed of a heterogeneous group with an intermediate prognosis [2].

In the past few years, next-generation sequencing technologies have been developed rapidly in the field of cancer genomics [3, 4] and enabled a direct genome-wide association test for NK-AML at single-base resolution in an unbiased manner [5]. A comprehensive understanding of the genetic lesions is an important basis for developing personalized therapies for the treatment of NK-AML [6].

Ley *et al*. [5] carried out the first whole-genome association study for AML by using single-end whole-genome sequencing data and found 181 single nucleotide variations (SNVs), including 28 indels, nonsynonymous, and splicing sites (*NPM1* and *FLT3*, respectively). Mardis *et al*. [7] performed paired-end whole-genome sequencing for cytogenetically normal AML, found seven nonsynonymous SNVs (nsSNVs), one splice site SNV, two indels in coding regions, and 52 somatic point mutations in conserved or regulatory genomic regions, and observed a 10% recurrence rate in the *IDH1* mutation in 188 additional AML samples. In the following year, Ley *et al*. [5] carried out a case-control study of relapsed AML among patients who were reported in 2008. They found a one-base-pair deletion in the *DNMT3A* gene, which showed a recurrence risk of 22% in AML [8]. De Weer *et al*. [9] detected deletions within a chromosomal region of 7q35-q36, which contains the *CNTNAP2* gene, in both Kasumi-3 and MUTZ-3 cell lines using high-resolution array comparative genomic hybridization (CGH). Particularly, Bejar *et al*. [10] identified novel associations between somatic mutations in two genes, *ETV6* and *GNAS*, and both myelodysplastic syndrome (previously known as preleu-

Received January 23, 2013; Revised February 13, 2013; Accepted February 18, 2013

***Corresponding author:** Tel: +82-33-248-2691, Fax: +82-33-248-2693, E-mail jwpark@hallym.ac.kr

kemia) and AML using high-throughput genotyping technology. In the same year, Link *et al*. [11] identified a novel *TP53* susceptibility mutation in AML patients through the use of whole-genome sequencing. Recent exome-sequencing studies found more than 60 susceptible variants, including indels and splicing site mutations, in AML patients [12]. However, to the best of our knowledge, none of the previous studies attempted to predict the risk of NK-AML by analyzing whole-exome sequencing data.

This study aimed to evaluate the effect of nsSNVs in 10 pairs of tumor and normal cells of NK-AML patients with whole-exome sequencing data. Finally, we developed genetic risk prediction models for NK-AML by creating genetic risk scores (GRS).

Methods

Whole-exome sequencing data analysis

A total of 20 DNA samples were obtained from 10 pairs of bone marrow (tumor) and germline (normal) cells collected from 10 NK-AML patients. We illustrated the workflow for whole-exome sequencing data analysis in Fig. 1. The raw reads in FASTQ file format were mapped to the human reference, which was developed by the 1,000 genome project (human_g1k_v 37.fasta), by using the Burrows-Wheeler aligner (BWA v 0.6.1), which generates SAM format files [13]. The SAM files were converted into binary format files (BAM) by samtools v0.2.6, which reduces the file size and improves computing efficiency [14]. The read group information of the sequencing machine was added to the head of the BAM files. The aligned sequences were sorted in the order of chromosomes by Picard v1.79 (http://picard.sourceforge.net) and went through a PCR duplicate marking process, which enables the Genome Analysis Toolkit (GATK) to ignore duplicates in subsequent processing [15]. Finally, the BAM files were indexed by bamtools v2.2.0 [16]. We performed a local realignment prior to recalibration, which gives the most accurate quality scores for each sample. Local realignment with known indel sites (Mills_and_1000G_gold_standard.indels.b37.vcf, 1000G_phase1.indels.b37.vcf) for each individual does not require multiple sample realignments, which demand extreme computational power. However, we carried out a local realignment with the matched tumor and normal samples together to prevent misalignment due to the differences between these two tissue types. Recalibration was performed with multiple known sites (dbSNV_137.b37, Mills_and_1000G_gold_standard.indels.b37.vcf, and 1000G_phase1.indels.b37.vcf), which may increase recalibration accuracy. We reduced the BAM file size to about 1/100 of the original file size by using the GATK tool, which saved variant calling time without losing any essential information. We used the UnifiedGenotyper of GATK for variant calling, followed by variant recalibration with known sites (hapmap_3.3.b37.vcf, 1000G_omni2.5.b37.vcf, dbsnp_137.b37.vcf, and Mills_and_1000G_gold_standard.indels.b37.vcf), and annotated by them using snpEff v2.0.57 [17].

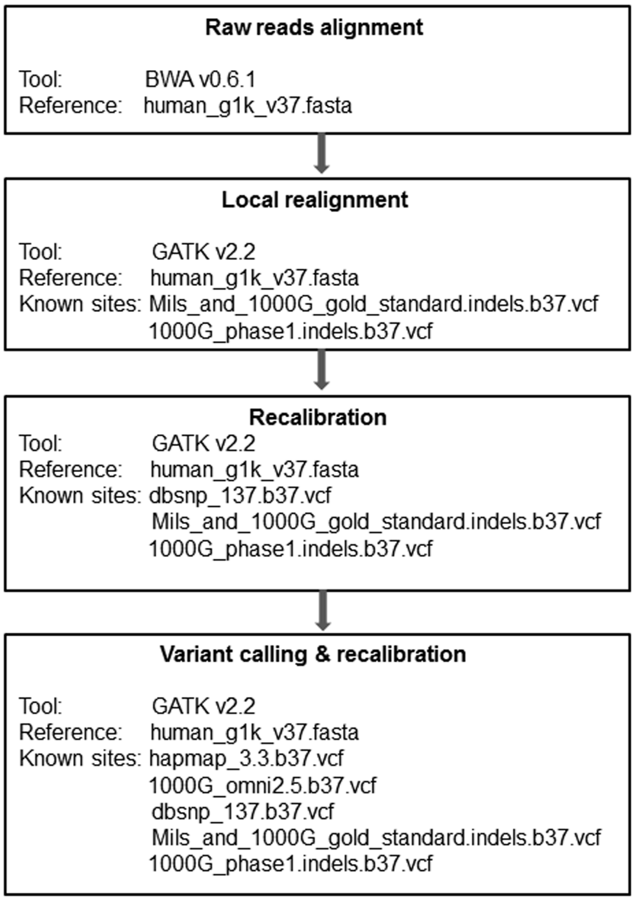
Raw reads alignment
Tool: BWA v0.6.1
Reference: human_g1k_v37.fasta
Local realignment
Tool: GATK v2.2
Reference: human_g1k_v37.fasta
Known sites: Mils_and_1000G_gold_standard.indels.b37.vcf
1000G_phase1.indels.b37.vcf
Recalibration
Tool: GATK v2.2
Reference: human_g1k_v37.fasta
Known sites: dbsnp_137.b37.vcf
Mils_and_1000G_gold_standard.indels.b37.vcf
1000G_phase1.indels.b37.vcf
Variant calling & recalibration
Tool: GATK v2.2
Reference: human_g1k_v37.fasta
Known sites: hapmap_3.3.b37.vcf
1000G_omni2.5.b37.vcf
dbsnp_137.b37.vcf
Mils_and_1000G_gold_standard.indels.b37.vcf
1000G_phase1.indels.b37.vcf

Fig. 1. Next-generation sequencing data analysis workflow.

Statistical analysis

We performed logistic regression analyses between the somatic mutations and NK-AML using PLINK/SEQ v0.08 (http://atgu.atgu.mgh.harvard.edu/plinkseq), which provides powerful utilities in variant call format (vcf) for analyzing whole- exome and -genome data. Further, we verified the odd ratios and p-values estimated from PLINK/SEQ using Stata, v11.2 (Stata Corp., College Station, TX, USA).

We selected the somatic nsSNVs with complete call rates and evaluated the GRS models composed of the variants associated with NK-AML. The GRS was calculated for each individual by accumulating the number of risk alleles (0, 1, or 2) of the SNVs. We created stepwise GRS models, comprised of the selected SNVs, according to their significance level; if the significance level was equal between two or more SNVs, we selected the SNVs in the order of their chromosomal position. In addition, we evaluated a GRS model that consisted of gene variants reported in previous leukemia studies [9, 10, 18, 19]. We compared the area under the receiver operating characteristic curve (AUC) of each GRS model using the “roctab” and “roccomp” commands in Stata.

Results

The improvement in concordance between the empirical and reported quality scores of SNVs is shown in Fig. 2A. A significant improvement in the accuracy of quality scores after recalibration is shown in Fig. 2B. The association test for NK-AML yielded 42 SNVs that had a p-value of less than

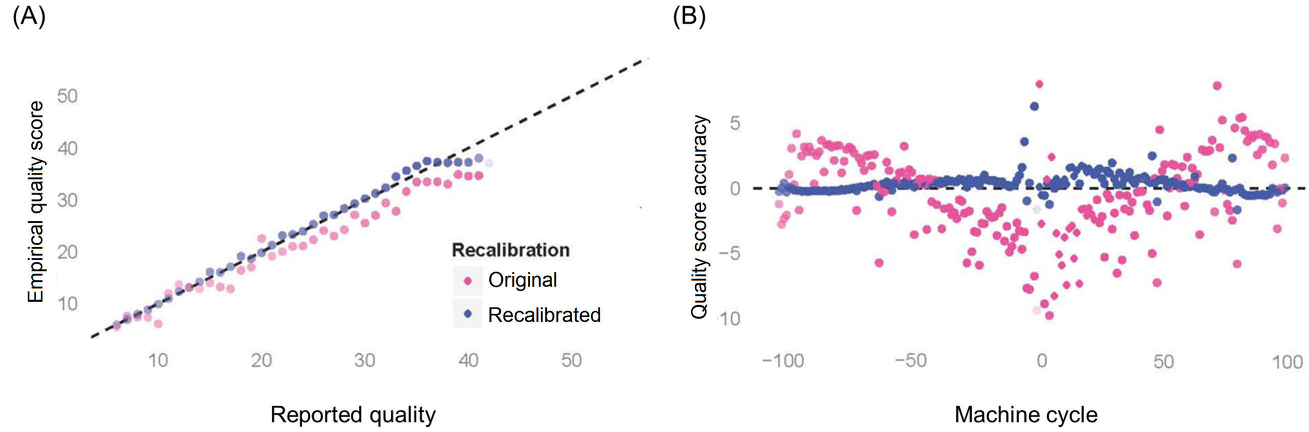

Fig. 2. Base quality score recalibration. Raw reads (pink) and recalibrated (blue) base quality scores for a patient. Panel (A) compares the reported base quality scores to the empirical estimates. Panel (B) shows the difference between the average reported and empirical quality scores for each machine cycle.

Table 1. Associations between 21 SNVs and NK-AML in 10 pairs of tumor and normal samples

Gene	Chr.	SNV	Amino acid change	N/R	RAF case/control	OR (95% CI)	p-value
PPIAL4G	1q21.1	rs6604516	Q24L	A/T	0.95/0.65	21.0 (1.8–248.1)	0.016
LRRN2	1q32.1	rs201287849	A141P	C/G	0.30/0.05	13.5 (1.2–152.2)	0.035
OR2T33	1q44	rs10888338	A169V	A/G	0.40/0.15	9.3 (1.2–73.0)	0.033
ANKRD36	2q11.2	rs79579412	K378R	G/A	0.95/0.70	13.5 (1.2–152.2)	0.035
		rs2443878	K386E	G/A	0.85/0.60	9.3 (1.2–73.0)	0.033
ATP13A3	3q29	rs75156964	V55G	C/A	0.90/0.60	16.0 (1.8–143.2)	0.013
MUC4	3q29	rs75459784[a]	H2306P	G/T	0.85/0.60	9.3 (1.2–73.0)	0.033
		rs55824312	P2192S	A/G	0.80/0.55	13.5 (1.2–152.2)	0.035
TRIML1	4q35.2	rs200129069	S254R	C/G	0.35/0.10	9.3 (1.2–73.0)	0.033
CNTNAP2	7q35	rs200239604	T589P	C/A	0.95/0.70	13.5 (1.2–152.2)	0.035
PABPC1	8q22.2	rs79940439	R475Q	C/T	0.40/0.15	9.3 (1.2–73.0)	0.033
MUC5B	11p15.5	rs202160055	R1097H	G/A	0.30/0.05	13.5 (1.2–152.2)	0.035
TAS2R43	12p13.2	rs201245949	Q210H	C/G	0.40/0.15	9.3 (1.2–73.0)	0.033
ATF7IP	12p13.1	rs199961592	Q883K	C/A	0.45/0.20	13.5 (1.2–152.2)	0.035
KIAA1033	12q24.11	rs199570381	Q381K	C/A	0.30/0.05	13.5 (1.2–152.2)	0.035
ATP8A2	13q12	rs201822155	L624P	G/T	0.35/0.10	9.3 (1.2–73.0)	0.033
L2HGDH	14q21.3	rs201692645	V198G	A/C	0.35/0.10	9.3 (1.2–73.0)	0.033
LYPD5	19q13.31	rs79007092	S151A	A/C	0.35/0.10	9.3 (1.2–73.0)	0.033
GNAS	20q13.3	rs56213454	Q397P	C/A	0.90/0.60	16.0 (1.8–143.2)	0.013
		rs56371919	M404I	T/G	0.90/0.65	9.3 (1.2–73.0)	0.033
SUSD2	22q11	rs399140	A91T	A/G	0.90/0.65	9.3 (1.2–73.0)	0.033

SNV, single nucleotide variation; NK-AML, normal-karyotype acute myeloid leukemia; Chr., chromosome; N/R, nonrisk/risk allele; RAF, risk allele frequency; OR, odds ratio; CI, confidence interval.

[a] The rs75459784 is in complete linkage disequilibrium with the rs74420943.

0.05 and passed the quality threshold of PLINK/SEQ (data not shown). Among 42 SNVs, we excluded 11 SNVs with no rs ID, four SNVs with call rates less than 1, and five synonymous SNVs for subsequent analysis.

A total of 21 nsSNVs located in 18 candidate genes were selected and included in the GRS models after considering pairwise linkage disequilibrium (Table 1). This study replicated three somatic mutations of the *MUC4*, *CNTNAP2*, and *GNAS* genes, which were reported in previous studies on leukemia [9, 10, 18, 19]. In addition, we identified novel point mutations in 15 other genes. The SNVs rs75156964 (*ATP13A3*; odds ratio [OR], 9.33; p = 0.013) and rs56213454 (*GNAS*; OR, 9.33; p = 0.013) showed the most significant evidence for association, and rs6604516 (*PPIAL4G*; OR, 21.00; p = 0.016) showed the strongest effect size among the 21 SNVs. Other SNVs also showed strong effects in the risk of NK-AML (OR, 9.33 to 13.5).

Table 2. The area under the ROC curves for genetic risk score models consisting of SNVs

Model[a]	No. of SNVs	SNV	AUC (95% CI)[b]
1	1	rs75156964	0.80 (0.62–0.98)
2	2	Model1 + rs56213454	0.86 (0.67–1.00)
3	3	Model2 + rs6604516	0.92 (0.81–1.00)
4	4	Model3 + rs10888338	0.97 (0.92–1.00)
5	5	Model4 + rs2443878	1.00 (1.00–1.00)

Note: SNVs were included in the model according to 1) the p-values and, 2) in case of the same significance levels, the chromosome in which the SNVs were located.
ROC, receiver operating characteristic; SNV, single nucleotide variation; AUC, area under the ROC curve; CI, confidence interval.
[a]Genetic risk score model sums the numbers of risk alleles for each SNV (0, 1, and 2); [b]AUCs and 95% CIs were estimated from ROC analyses.

We compared the prediction accuracy of 21 stepwise GRS models. The predictive power for NK-AML reached 100% in model 5, which consisted of five nsSNVs: rs75156964 (*ATP13A3*, 3q29), rs56213454 (*GNAS*, 20q13.3), rs6604516 (*PPIAL4G*, 1q21.1), rs10888338 (*OR2T33*, 1q44), and rs2443878 (*ANKRD36*, 2q11.2) (Table 2, Fig. 3A). The five SNVs (rs75459784, rs55824312, rs200239604, rs56213454, and rs56371919) in three previously reported genes—*MUC4*, *CNTNAP2*, and *GNAS* —also showed a high AUC value (AUC, 0.98; 95% confidence interval [CI], 0.92 to 1.00).

Discussion

We evaluated the effects of 21 nsSNVs of 18 candidate genes for NK-AML in this study. Among 18 genes, we replicated previously reported associations of two genes (*CNTNAP2* and *GNAS*) with AML [9, 10]. The mutation in the *MUC4* gene has been reported to be associated with acute lymphoblastic leukemia [18, 19]. The *MUC4* gene encodes a mucin protein and a high-molecular-weight glycoprotein in humans. This integral membrane glycoprotein, which is observed on the cell surface, plays various roles in tumor progression [20]. Particularly, *MUC4*, complexed with

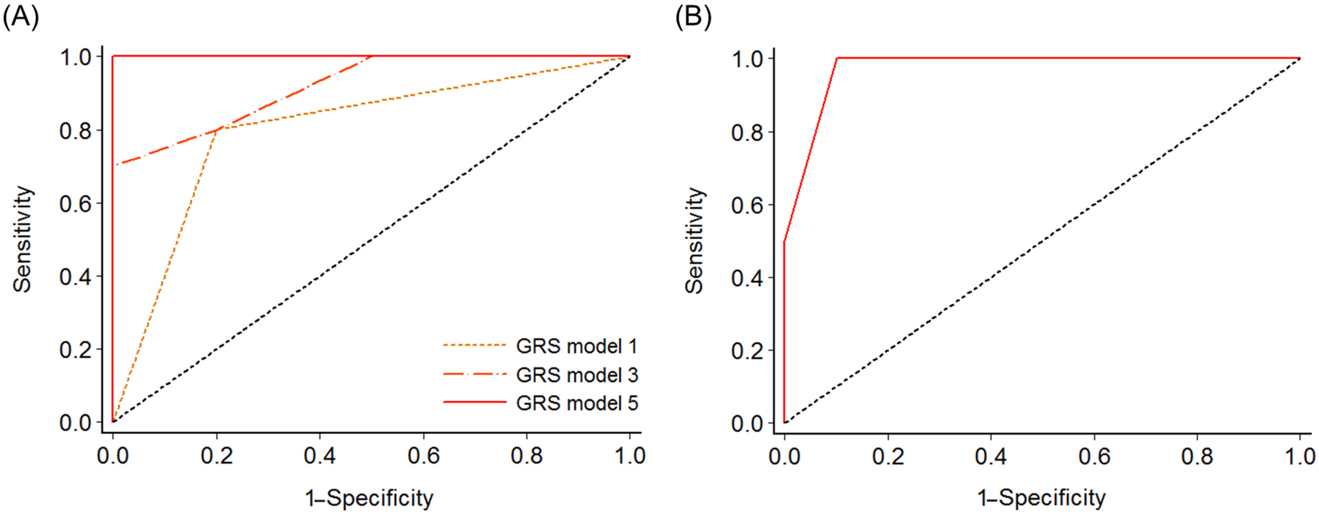

Fig. 3. Receiver operating characteristic analysis to measure the discriminatory power of each genetic risk score model in predicting normal-karyotype acute myeloid leukemia. (A) Model 1 (rs75156964), model 3 (rs75156964, rs56213454, and rs6604516), and model 5 (rs75156964, rs56213454, rs6604516, rs10888338, and rs2443878). (B) A model comprising five candidate single nucleotide variations (rs75459784, rs55824312, rs200239604, rs56213454, and rs56371919) in *MUC4*, *CNTNAP2*, and *GNAS*, reported in previous studies (area under the receiver operating characteristic curve, 0.98; 95% confidence interval, 0.92 to 1.00). GRS, genetic risk score.

v-erb-b2 erythroblastic leukemia viral oncogene homolog 2 (*ERBB2*), results in the repression of apoptosis and the stimulation of proliferation; in addition, the overexpression of *MUC4* in carcinoma cells promotes loss of cellular polarity by ERBB2-mediated disruption [21]. The *CNTNAP2* gene is one of the largest genes in the entire human genome and occupies approximately 1.5% of chromosome 7. This gene encodes a neurexin family member that functions in the vertebrate nervous system. The *GNAS* gene has a highly complex imprinted expression pattern. This gene provides instructions for making one component, the stimulatory alpha subunit, of a protein complex called a guanine nucleotide-binding protein (G-protein). G-protein alpha subunit regulates the cyclic AMP (cAMP) pathway. *GNAS* gene mutations are known to be associated with high cAMP signaling [22]. The novel associations between 15 somatic mutations (*PPIAL4G, LRRN2, OR2T33, ANKRD36, ATP13A3, TRIML1, PABPC1, MUC5B, TAS2R43, ATF7IP, KIAA1033, ATP8A2 , L2HGDH, LYPD5,* and *SUSD2*) and NK-AML need to be replicated, and their functional mechanisms in the AML should be investigated in future studies.

The GRS models that were comprised of the somatic nsSNVs showed extremely high predictive accuracy for the risk of NK-AML (86% to 100%). The combination of five nsSNVs of previously reported genes (*MUC4, CNTNAP2,* and *GNAS*) had a predictive ability of 98% for the risk of NK-AML in this study. One of the limitations in this study was the small number of study subjects (n = 10), which resulted in substantial variations in the 95% CIs of the effects (ORs) of each SNV.

In conclusion, we have highlighted 21 susceptibility nsSNVs located in 18 genes. The GRS model that comprises five candidate SNVs is highly informative in predicting the risk for NK-AML. The discovery of novel markers may provide an opportunity to develop novel diagnostic and therapeutic targets for NK-AML. Further study with a larger sample size is necessary to validate the AML-related gene mutations and will provide an opportunity to develop a powerful genetic risk prediction model for NK-AML.

Acknowledgments

The sequence data for this study were provided by the National Project for Personalized Genomic Medicine (A111218-GM07) (PGM21 Project, A111218), Ministry for Health and Welfare, Republic of Korea.

References

1. Lowenberg B, Downing JR, Burnett A. Acute myeloid leukemia. *N Engl J Med* 1999;341:1051-1062.
2. Mrozek K, Marcucci G, Paschka P, Whitman SP, Bloomfield CD. Clinical relevance of mutations and gene-expression changes in adult acute myeloid leukemia with normal cytogenetics: are we ready for a prognostically prioritized molecular classification? *Blood* 2007;109:431-448.
3. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. *Nature* 2009;458:719-724.
4. Kwon SM, Cho H, Choi JH, Jee BA, Jo Y, Woo HG. Perspectives of integrative cancer genomics in next generation sequencing era. *Genomics Inform* 2012;10:69-73.
5. Ley TJ, Mardis ER, Ding L, Fulton B, McLellan MD, Chen K, *et al*. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* 2008;456:66-72.
6. Riva L, Luzi L, Pelicci PG. Genomics of acute myeloid leukemia: the next generation. *Front Oncol* 2012;2:40.
7. Mardis ER, Ding L, Dooling DJ, Larson DE, McLellan MD, Chen K, *et al*. Recurring mutations found by sequencing an acute myeloid leukemia genome. *N Engl J Med* 2009;361: 1058-1066.
8. Ley TJ, Ding L, Walter MJ, McLellan MD, Lamprecht T, Larson DE, *et al*. *DNMT3A* mutations in acute myeloid leukemia. *N Engl J Med* 2010;363:2424-2433.
9. De Weer A, Poppe B, Vergult S, Van Vlierberghe P, Petrick M, De Bock R, *et al*. Identification of two critically deleted regions within chromosome segment 7q35-q36 in *EVI1* deregulated myeloid leukemia cell lines. *PLoS One* 2010;5:e8676.
10. Bejar R, Stevenson K, Abdel-Wahab O, Galili N, Nilsson B, Garcia-Manero G, *et al*. Clinical effect of point mutations in myelodysplastic syndromes. *N Engl J Med* 2011;364:2496-2506.
11. Link DC, Schuettpelz LG, Shen D, Wang J, Walter MJ, Kulkarni S, *et al*. Identification of a novel TP53 cancer susceptibility mutation through whole-genome sequencing of a patient with therapy-related AML. *JAMA* 2011;305:1568-1576.
12. Yan XJ, Xu J, Gu ZH, Pan CM, Lu G, Shen Y, *et al*. Exome sequencing identifies somatic mutations of DNA methyltransferase gene *DNMT3A* in acute monocytic leukemia. *Nat Genet* 2011;43:309-315.
13. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 2010;26:589-595.
14. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;25:2078-2079.
15. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, *et al*. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;20:1297-1303.
16. Barnett DW, Garrison EK, Quinlan AR, Stromberg MP, Marth GT. BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics* 2011;27:1691-1692.
17. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, *et al*. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* 2012;6:80-92.
18. Harvey RC, Mullighan CG, Wang X, Dobbin KK, Davidson GS, Bedrick EJ, *et al*. Identification of novel cluster groups in pe-

diatric high-risk B-precursor acute lymphoblastic leukemia with gene expression profiling: correlation with genome-wide DNA copy number alterations, clinical characteristics, and outcome. *Blood* 2010;116:4874-4884.

19. Zhang J, Mullighan CG, Harvey RC, Wu G, Chen X, Edmonson M, *et al*. Key pathways are frequently mutated in high-risk childhood acute lymphoblastic leukemia: a report from the Children's Oncology Group. *Blood* 2011;118:3080-3087.
20. Srivastava SK, Bhardwaj A, Singh S, Arora S, Wang B, Grizzle WE, *et al*. MicroRNA-150 directly targets *MUC4* and suppresses growth and malignant behavior of pancreatic cancer cells. *Carcinogenesis* 2011;32:1832-1839.
21. Kufe DW. Mucins in cancer: function, prognosis and therapy. *Nat Rev Cancer* 2009;9:874-885.
22. Horvath A, Boikos S, Giatzakis C, Robinson-White A, Groussin L, Griffin KJ, *et al*. A genome-wide scan identifies mutations in the gene encoding phosphodiesterase 11A4 (*PDE11A*) in individuals with adrenocortical hyperplasia. *Nat Genet* 2006;38:794-800.