G&I Genomics & Informatics

pISSN 1598-866X eISSN 2234-0742
Genomics Inform 2013;11(1):24-33
http://dx.doi.org/10.5808/GI.2013.11.1.24

ORIGINAL ARTICLE

Bioinformatics Interpretation of Exome Sequencing: Blood Cancer

Jiwoong Kim[1†], Yun-Gyeong Lee[1,2†], Namshin Kim[1,2*]

[1]Korean Bioinformation Center, Korea Research Institute of Bioscience and Biotechnology, Daejeon 305-806, Korea,
[2]Department of Bioinformatics, University of Science and Technology, Daejeon 305-806, Korea

We had analyzed 10 exome sequencing data and single nucleotide polymorphism chips for blood cancer provided by the PGM21 (The National Project for Personalized Genomic Medicine) Award program. We had removed sample G06 because the pair is not correct and G10 because of possible contamination. In-house software somatic copy-number and heterozygosity alteration estimation (SCHALE) was used to detect one loss of heterozygosity region in G05. We had discovered 27 functionally important mutations. Network and pathway analyses gave us clues that *NPM1*, *GATA2*, and *CEBPA* were major driver genes. By comparing with previous somatic mutation profiles, we had concluded that the provided data originated from acute myeloid leukemia. Protein structure modeling showed that somatic mutations in *IDH2*, *RASGEF1B*, and *MSH4* can affect protein structures.

Keywords: acute myeloid leukemia, computational biology, exome, mutation

Introduction

Thanks to the next-generation sequencing (NGS) technology, genomics research area has been developed dramatically in recent years. Since 2008, many efforts have been made to find somatic mutations in cancer genomes; hundreds of samples by exome sequencing have been used to develop diagnosis and prognosis markers in various cancers in recent researches. Cheaper price has enabled us to get more exome sequencing data from hundreds of samples, even thousands of genomes.

By applying exome sequencing to cancer genomes, we can discover somatic mutations and somatic copy-number alterations (SCNAs) on a massive and genome-wide scale. Traditionally, array-based comparative genomic hybridization has been used to discover SCNA, but exome sequencing has many advantages, in that one can derive various results, such as somatic mutations, loss of heterozygosity (LOH), and SCNA. It is still under debate because bioinformatics methods could give different results depending on their parameters and software, but major driver genes of somatic mutations show a similar consensus.

We had used an in-house bioinformatics pipeline and databases for this research, especially "somatic copy-number and heterozygosity alteration estimation" (SCHALE) to find LOH and SCNA. Furthermore, driver genes were also inferred from network and pathway analyses. Protein structure prediction was used to predict putative structural changes by somatic mutations and also possibly their functional changes.

Methods

Dataset

Whole-exome sequencing data of tumor/normal paired samples from 10 individuals with blood cancer were obtained in FASTQ format from PGM21 (The National Project for Personalized Genomic Medicine). For 10 of the samples, the genotype information from single nucleotide poymorphism (SNP) chip was additionally obtained (but not the raw data).

Data processing and quality control (QC)

The QC of sequencing reads was performed with NGS QC

Received January 30, 2013; Revised February 20, 2013; Accepted February 22, 2013

***Corresponding author:** Tel: +82-42-879-8540, Fax: +82-42-879-8519, E-mail: deepreds@kribb.re.kr

[†]These two authors contributed equally to this work.

Toolkit, and the high-quality reads were aligned to the human reference genome (hg19, http://genome.ucsc.edu) using BWA (http://bio-bwa.sourceforge.net/) [1] with default parameteres. Picard (http://picard. sourceforge.net) was used to remove PCR duplicates. Base quality recalibration and local realignment around indels were performed using Genome Analysis Toolkit (GATK, http://www. broadinstitute.org/gatk/). Because our previous study showed that the discordance of indel alignments had potential to generate false-positive somatic mutations, we performed the local realignment after the merging of alignments from same patients. Template lengths were calculated using the processed alignments, and the distributions for each samples were plotted to examine the qualities of sequencing libraries. Variant genotyping for each sample was performed with UnifiedGenotyper of GATK, and the variants were filtered as follows: mapping quality zero (MQ0) $\geq$ 4 && MQ0/depth of coverage (DP) $>$ 0.1, variation quality (QUAL) $<$ 200, quality by depth (QD) $<$ 5, Fisher strand (FS) $>$ 60, MQ $<$ 40, DP $<$ 10 (5 if no variant allele), genotype quality (GQ) $<$ 13. The genomic annotation of variants was performed with custom-made scripts and refSeq information.

Identification of somatic mutations

We had developed a bioinformatics pipeline for this award program (Fig. 1). VarScan [2] was used to identify somatic mutations. The output of VarScan was filtered as follows: somatic p-value $\leq$ 0.05, normal allele depth $\leq$ 1, tumor allele depth $\geq$ 5, tumor total depth $\geq$ 10. The mutation set was annotated using a custom-made script with refSeq, SIFT (http://sift.jcvi.org/), MutationAssessor, phyloP46way (vertebrate), dbSNP 137 [3], in-house Korean variation database, etc. We selected the driver mutations that were not in dbSNP 137, excluding clinically associated variants (Flagged track in University of California Santa Cruz [UCSC] Genome Browser) and an in-house Korean variation database, and were deleterious missense, nonsense, read-through, splice-site, or coding indel mutations (NS/SS/I). The deleterious missense mutations were determined if at least 2 out of the following 3 were true: PhyloP score $>$ 1.5, SIFT prediction = deleterious, MutationAssessor Func. Impact = high or medium.

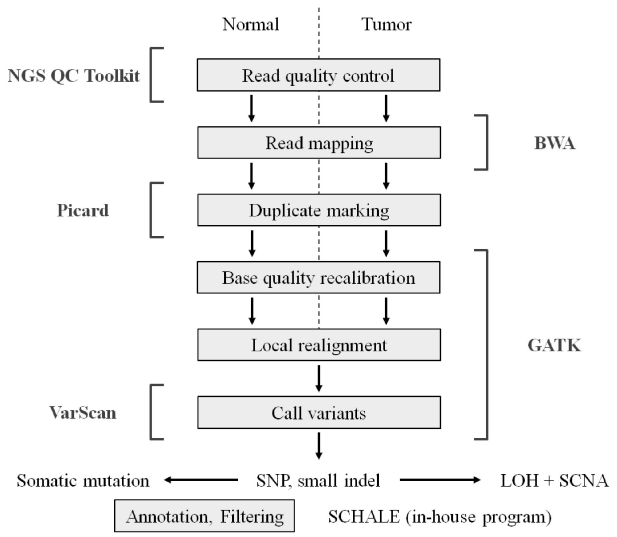
The total numbers of candidates for each filtering step are summarized in Fig. 1. We found somatic mutations in samples G10, G08, and G09, but they were not NS/SS/I.

Comparison with genotypes from SNP chip

To compare with genotypes from SNP chip, the genotyping was performed using UnifiedGenotyper with the intervals, including all the coordinates of SNP chip and the option-output_mode EMIT_ALL_SITES. For each sample with genotype information from SNP chip, overall genotype concordance, non-reference sensitivity, and non-reference discrepancy rate were calculated according to the guide of GATK VariantEval using the genotypes with DP $\geq$ 5.

Genotype comparison of samples

The multi-sample SNP genotyping was performed with UnifiedGenotyper to compare the genotypes on the variant positions where all samples were able to be genotyped. The filtration scheme was the same as above. The distance between two samples was calculated as the sum of genotype difference values (0 if two genotypes are same, 1 if a genotype is 0/1 and another is 0/0 or 1/1, 2 if a genotype is 0/0 and another is 1/1). The distances were divided by the number of the variant positions. Heatmap and similarity matrix were used to visualize the distances.



Patient	Somatic (VarScan 2)	Filtered	Novel or Flagged	NS/SS/I
G01	1,123	118	15	0
G02	1,151	109	26	4
G03	1,217	132	40	5
G04	1,187	156	50	6
G05	1,493	172	35	6
G07	1,187	126	41	7
G08	1,086	102	17	0
G09	1,146	125	16	0

Fig. 1. Bioinformatics pipeline for exome data and numbers of somatic mutation candidates for each filtering step. SNP, single nucleotide polymorphism; LOH, loss of heterozygosity; SCNA, somatic copy-number alterations; SCHALE, somatic copy-number and heterozygosity alteration estimation; NS/SS/I, nonsense, splice-site, or coding indel mutations.

Genome annotation

We used in-house genome annotation pipelines and various databaes, as summarized below.

Genomic annotation: refGene

Human Genome Variation Society (HGVS) nomenclature: nucleotide (non-coding, coding), protein

SIFT - tolerant amino acid substitutions

MutationAssessor

phyloP46way_vertebrate

GERP; GERP_element: rsScore

RepeatMasker

Tandem repeat definition

dbSNP 137; Common, Flagged, Mult

Korean variation database (54 Personal Genomes)

OMIM morbid genes

Oncogenes and tumor suppressors from Ariadne Pathway Studio

Detection of LOH and SCNA by SCHALE

Detection LOH and SCNA was performed with custom-made software SCHALE (Fig. 2). For the positions where the genotypes were heterozygous in normal samples, the values of LOH and SCNA were calculated. The values of LOH were calculated by the equation:

LOH = | RT − 0.5 | − | RN − 0.5 |, where RN is the count of reads supporting a non-reference allele divided by the total depth in the normal sample and RT is in tumor sample. The values of SCNA were calculated by the equation:

SCNA = $\log_2$(tumor depth/normal depth), where the depths were normalized using R library tweeDEseq. The values of LOH and SCNA were segmented together using R library DNAcopy with the parameters alpha = 0.01 and undo.SD = 2. The segments and values were plotted for each chromosome with the idiograms in R library SNPchip.

Network analysis

Ingenuity Pathway Analysis (IPA) was used to perform the network study of the identified somatically mutated genes and the genes in the LOH region.

Protein structure modeling

Before the structural modeling, the protein features related to the somatic mutations were found by BLAST on UniProt (10 most similar proteins, expectation value = 1e-30) and residue mapping. Homology modeling was performed using MODELLER (http://salilab.org/modeller) with the templates 1LWD for IDH2, 3KSY for RASGEF1B,

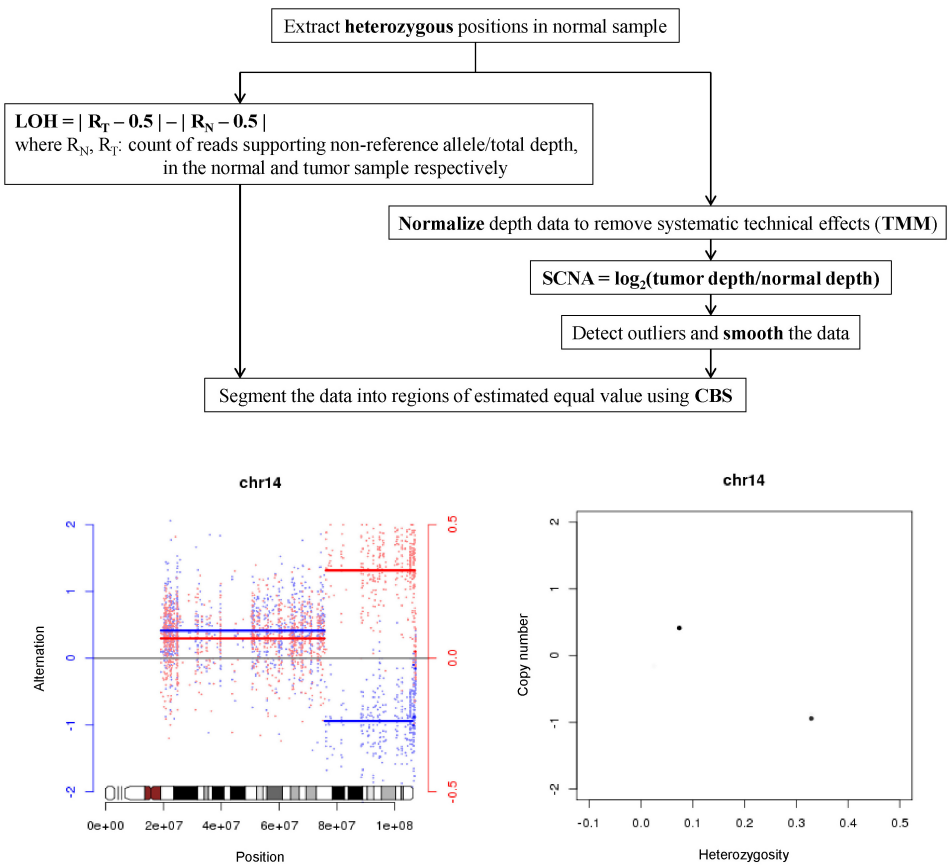


Fig. 2. Somatic copy-number and heterozygosity alteration estimation (SCHALE), developed for detection of loss of heterozygosity (LOH) and somatic copy-number alterations (SCNA). TMM, trimmed mean of M values; CBS, circular binary segmentation.

and 2O8B for MSH4. Molecular dynamics simulation was performed with GROMACS (http://pubs.acs.org/doi/abs/10.1021/ct700301q). The structural images were drawn using PyMOL (http://www.pymol.org/).

Results

QC and results

We had used several methods to check the qualities of sequence reads. First, we used NGS QC Tools [4] to confirm sequence qualities (Supplementary Table 1). Overall, they had a very good quality, according to NGS QC Tools results. Template length (library size) was also evenly distributed for all samples (Supplementary Fig. 1).

Reference genome mapping

About 98% of reads were mapped onto the human reference genome (UCSC Genome Browser, hg19), summarized in Supplementary Table 2. After a duplicate removal step, 77–85% reads were retained; each sample had up to 25% duplicates. Eighty-nine percent of target areas covered at least 10 sequence reads, which is sufficient for discovering somatic mutations. We did not get target enrichment kit information from PGM21, so we had compared coverage information from several kits and concluded that SureSelect All Exon 50Mb was used for exome sequencing.

Table 1. A list of functionally important somatic mutations in 8 samples

Patients	Gene	Mutation	Nucleotide substitution	Amino acid substitution	PhyloP	SIFT prediction	Mutation Assessor Func. Impact	COSMIC (occurrence)
G02, G07	*NPM1*	Frameshift	c.773_776dupTCTG c.860_863dupTCTG	p.Trp259Cysfs*12 p.Trp288Cysfs*12	3.463			Yes (2,126)
G02	*ALDH6A1*	Missense	c.166T>A	p.Trp56Arg	4.875	Deleterious	Medium	No
G02	*IDH2*	Missense	c.419G>T	p.Arg140Leu	4.024	Deleterious	High	Yes (24)
G02	*MYH13*	Missense	c.926C>G	p.Pro309Arg	5.335	Deleterious	High	p.P309T
G03	*RPS6KA1*	Missense	c.2101A>G c.2128A>G	p.Thr701Ala p.Thr710Ala	4.664	Deleterious	Medium	No
G03	*KCNH1*	Missense	c.856C>T	p.Leu286Phe	4.083	Deleterious	High	No
G03	*CACNB4*	Missense	c.1184C>T c.1232C>T c.1286C>T	p.Ala395Val p.Ala411Val p.Ala429Val	5.993	Tolerated	Medium	No
G03	*RAD21*	Nonsense	c.1238T>A	p.Leu413*	5.044			No
G03	*KSR2*	Missense	c.967C>T	p.Pro323Ser	5.180	Tolerated	Medium	No
G04	*CCT7*	Frameshift	c.1011C[4] c.1362C[4] c.1491C[4] c.1623C[4]	p.His339Thrfs*45 p.His456Thrfs*45 p.His499Thrfs*45 p.His543Thrfs*45	5.138			No
G04	*GATA2*	Missense	c.989G>T	p.Arg330Leu	5.914	Deleterious	High	p.R330Q
G04	*TRA2B*	Missense	c.16C>Tc.316C>T	p.Arg6Cys p.Arg106Cys	2.951	Deleterious	Medium	No
G04	*RBMXL2*	Missense	c.727T>A	p.Tyr243Asn	3.247	Deleterious	Medium	No
G04	*KRT9*	Missense	c.1004C>T	p.Ala335Val	0.137	Deleterious	Medium	No
G04	*CEBPA*	In-frame	c.937_939dupAAG	p.Lys313dup	5.464			Yes (46)
G05	*MAST2*	Missense	c.2384G>A	p.Arg795His	5.974	Deleterious	Medium	No
G05	*L1TD1*	Nonsense	c.451G>T	p.Glu151*				No
G05	*MSH4*	Missense	c.1243G>A	p.Glu415Lys	3.441	Deleterious	Low	No
G05	*GATA2*	Missense	c.1033T>G c.1075T>G	p.Leu345Val p.Leu359Val	2.421	Deleterious	High	Yes (16)
G05	*VDAC2*	Missense	c.637C>A c.682C>A	p.Leu213Ile p.Leu228Ile	3.188	Tolerated	Medium	No
G05	*PHF6*	Splicing	c.585+1G>A c.588+1G>A		5.145			No
G07	*RASGEF1B*	Missense	c.781T>A	p.Trp261Arg	4.968	Deleterious	Low	No
G07	*FZD9*	Missense	c.1027G>A	p.Gly343Arg	3.725	Deleterious	Medium	No
G07	*ARHGEF5*	Missense	c.3956G>A	p.Arg1319His	1.817	Deleterious	High	No
G07	*TLL2*	Missense	c.2732A>C	p.Asn911Thr	2.032	Tolerated	Medium	No
G07	*IDH2*	Missense	c.419G>A	p.Arg140Gln	4.024	Deleterious	High	Yes (1,462)
G07	*ZNF317*	Missense	c.184G>T	p.Val62Phe	1.154	Deleterious	High	p.V62V

Genotyping by GATK and comparison with SNP chip

In order to check the genotyping parameters we had used, 10 SNP chip results were compared. Because SNP chip data were genotypes, we could not infer copy number changes; we could only calculate concordance between exome sequencing and SNP chip. We had used a separate bioinformatics pipeline for checking concordance, and GATK [5] was used mainly for genotyping of germline samples. We had discovered 18,500–19,200 coding SNPs. The number of SNPs from the 5' untranslated region (UTR) was close to that of the 3' UTR, so the target enrichment kit was not the latest 3' UTR extension version (Supplementary Table 3). In general, the 3' UTR is much larger than the 5' UTR; a new 3' UTR extension version of the exome kit will give at least a few times more SNPs than 5' UTR.

Two hundred fifty coding SNPs were retained after removal of dbSNP137 [6] and an in-house Korean database comprised of 54 normal individuals. Among them, we can focus on about 160 variants of NS/SS/I, which is believed to be most functionally important (Supplementary Table 4).

Overall concordance was at least 99.6%. In recent personal genome papers, they reported more than 99.95% concordance between whole-genome and SNP chips. Usually, one can achieve higher concordance if he uses strict parameters for their genotyping. Our purpose is to discover somatic mutations, not genotyping of normal germline, so we did not put much effort to get higher concordance values.

Especially, G10.N SNP chip data were totally different from the exome, because only 77% of SNPs were shared (Supplementary Table 5). As one can see in the table, 77% means they are different individuals (false positive for Award program). Distance and similarity matrix were calculated to see the differences and similarity (Supplementary Table 6, Supplementary Fig. 2). We had concluded that sample G06 has a different sample pair from different individuals. G10 also has only 92% similarity and showed a little similarity between G03 and G04. For such reasons, G10 was possibly contaminated and therefore rejected from our bioinformatics analysis.

In conclusion, we had used only 8 sample pairs for discovery of somatic mutations, LOH, and SCNA.

A list of somatic mutations

After removal of SNPs and indels from dbSNP137 [6] and

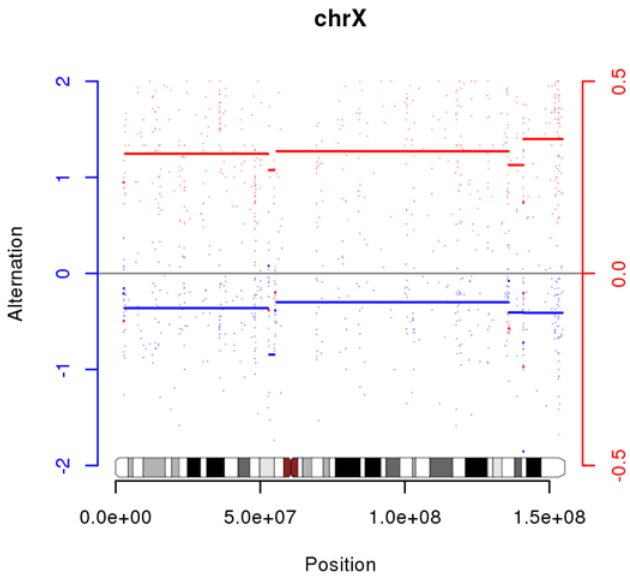
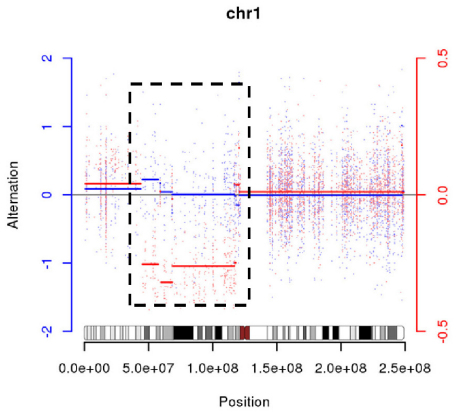


Fig. 3. Changes in loss of heterozygosity and somatic copy-number alteration in sample G10.



LOH in normal sample

Genotypes (N > T)	Total	Region	Percent
0/0 > 0/1	**127**	**39**	**30.7%**
0/0 > 1/1	2	0	0.0%
0/1 > 0/0	26	1	3.8%
0/1 > 0/1	285	1	0.4%
0/1 > 1/1	18	0	0.0%
1/1 > 0/0	1	0	0.0%
1/1 > 0/1	**31**	**26**	**83.9%**
1/1 > 1/1	11	0	0.0%

Chromosome	Start	End	No. of SNPs	LOH	CNA
chr1	45,292,866	57,535,198	61	-0.254	0.224
chr1	68,708,258	100,335,883	109	-0.280	-0.114
chr1	100,336,361	117,076,626	137	-0.245	0.102

Fig. 4. Abnormal region on chr1 in G05. LOH, loss of heterozygosity; SNP, single nucleotide poymorphism; CNA, copy-number alteration.

an in-house Korean database, 27 somatic mutations (NS/SS/I) were discovered from 8 samples. Among them, *NPM1* showed a TCTG internal tandem insertion. *NPM1-ITD* is known to be major driver gene in acute myeloid leukemia (AML) (Table 1, Supplementary Table 7). *NPM1-ITD* and *GATA2* were recurrently found in two samples. We had removed all variants from dbSNP and an in-house Korean database that showed polymorphisms or normal variants, because they can be putatively functionally neutral, even if they are discovered as somatic mutations.

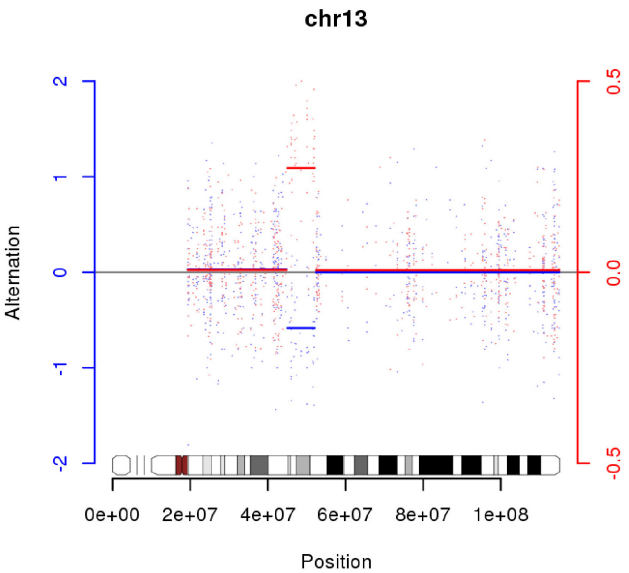


Fig. 5. Loss of heterozygosity on chr13 in G05.

We had downloaded the COSMIC (http://www.sanger.ac.uk/genetics/CGP/cosmic/) database and compared with our somatic mutations. We found the same highly recurrent somatic mutation variants in four genes: *NPM1*, *IDH2*, *CEBPA*, and *GATA2*.

In 26 genes, we can easily see the *NPM1*, *IDH2*, *GATA2*, *CEBPA* genes—known to be highly mutated in AML [3]. Based on the recent somatic mutation profiles from exome sequencing, one can see the following trends of them. In solid tumors, somatic mutation in *TP53* is high. But in melanoma, *BRAF* mutation is high. In hepatocellular carcinoma, *CTNNB1* mutation is high. It means that somatic mutations can play a different role, depending on cancer. In AML, other than the well-known *TP53* mutation, genes, such as *FLT3*, *TET2*, and *NPM1*, show somatic mutations and are the driver genes [7].

Discovery of LOH and SCAN by SCHALE

We had used SCHALE to discover LOH and SCNA. It appears that sample G06 has many LOHs and SCNAs, but an incorrect sample pair can give those results (Supplementary Fig. 3). We also discovered high LOH on chrX in G10 (Fig. 3).

We could not find any somatic mutations in G10. Previously, we had described that it could be contamination, but another possibility is that they are exomes from identical twins with different genders (we do not have any gender information from any samples).

Abnormal regions on chr1 in G05 showed LOH in the germline, but it is quite unnatural (Fig. 4). We are not describing much for this area until experimental validation is

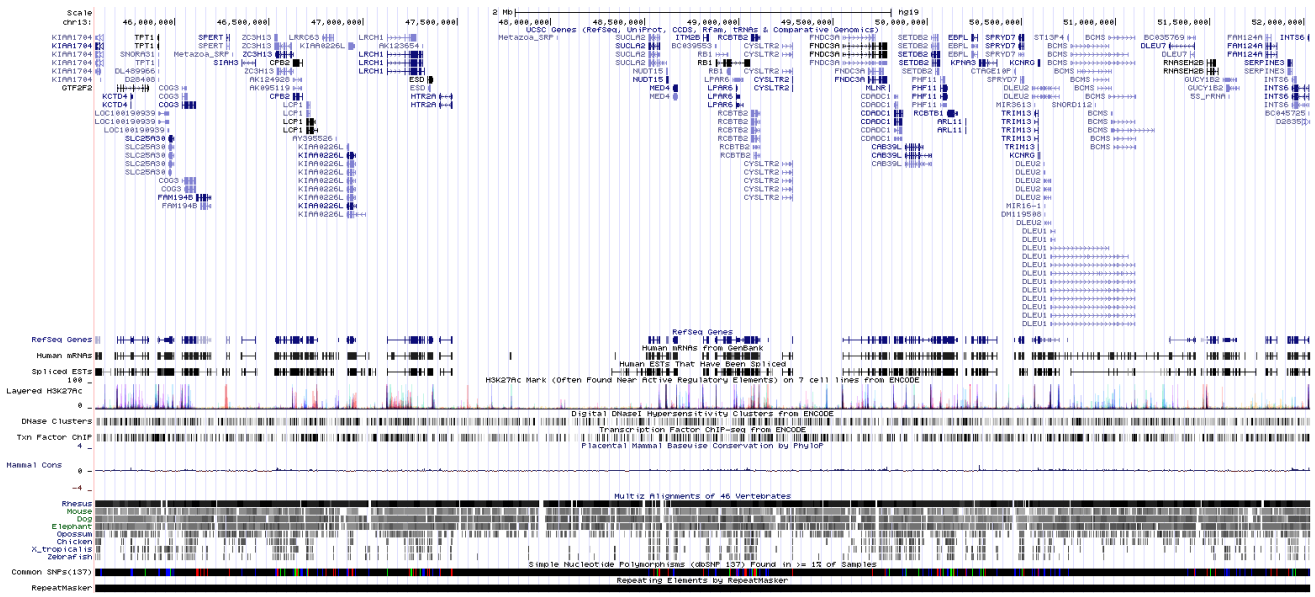


Fig. 6. Genes in a loss of heterozygosity region. Image adopted from University of California Santa Cruz (UCSC) Genome Browser, chr5:45,578,409-52,035,344.

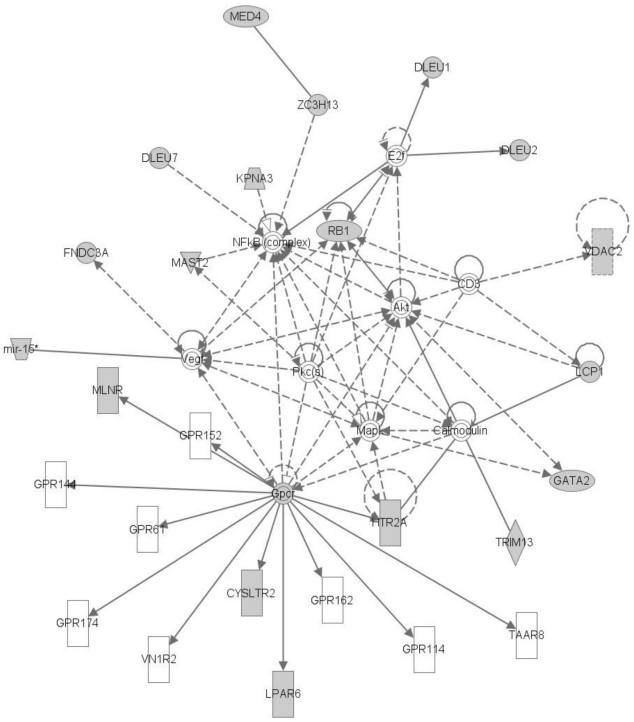


Fig. 7. Ingenity Pathway Analysis (IPA) from genes with loss of heterozygosity and somatic mutation (sample G05). Data were analyzed through the use of IPA (Ingenuity® Systems, http://www.ingenuity.com) [8].

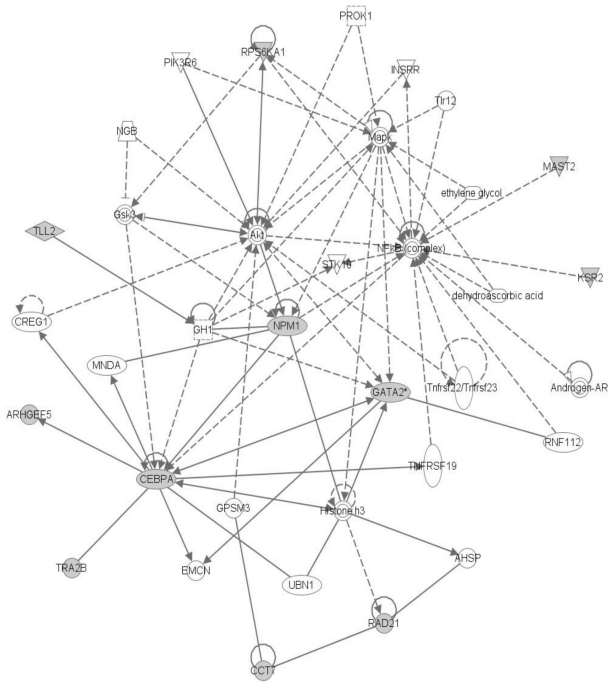


Fig. 8. "Cancer, Hematological Disease, Cell Cycle" Network. Data were analyzed through the use of Ingenity Pathway Analysis (IPA) (Ingenuity® Systems, http://www.ingenuity.com) [8].

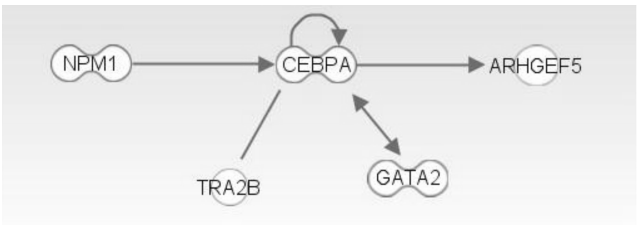


Fig. 9. Major interaction in top network. Data were analyzed through the use of Ingenity Pathway Analysis (IPA) (Ingenuity® Systems, http://www.ingenuity.com) [8].

Table 2. Two top networks inferred from Ingenuity Pathway Analysis

ID	Top networks	
	Associated network functions	Score
1	Cancer, hematological disease, cell cycle	26
2	Cell death and survival, cell cycle, cell-to-cell signaling and interaction, cellular movement	26

performed.

We had found one LOH region on chr13 in G05 (Fig. 5). INTS6 and KIAA1704 are located at both ends and potentially could make a fusion gene (Fig. 6). Experimental validation is also required for this assumption. INTS6 is expressed in blood; many expressed sequence tags (ESTs) were found in INTS6 in UniGene (http://www.ncbi.nlm.nih.gov/UniGene). But, KIAA1704 has a small number of ESTs, so we could not infer expression levels from UniGene. There are 61 genes on chr13:45,578,409-52,035,344, and we had an additional set of network and pathway analyses, including 6 genes (*MAST2*, *L1TD1*, *MSH4*, *GATA2*, *VDAC2*, and *PHF6*) with somatic mutations.

The top network is "Cancer, Hematological Disease, Immunological Disease," comprised of GATA2 and MAST2. Many genes in LOH regions are included in this network. We can infer that those genes with somatic mutations and LOH can interact with each other (Fig. 7).

Network and pathway analyses

Using 26 genes, we had used IPA (http://www.ingenuity.com/) [8] for network and pathway analyses. It can be summarized in two major networks (Table 2).

The top network is "Cancer, Hematologicial Disease, Cell Cycle," and it has *NPM1*, *GATA2*, and *CEBPA* genes (Figs. 8 and 9). These three genes showed a highly connected network and well-known major driver genes in leukemia.

The second network is related to cell death and survival (Fig. 10). *IDH2* and *PHF6* genes were located in this network. But overall, somatic mutation frequency in the top network

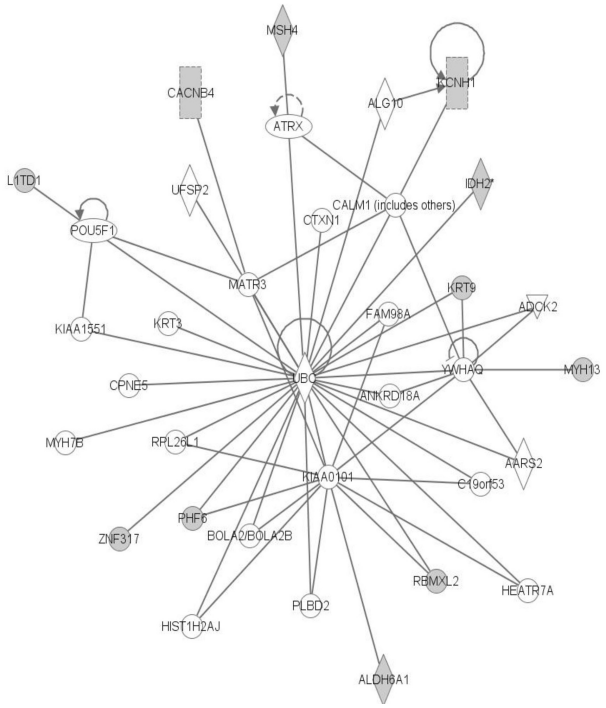


Fig. 10. "Cell Death and Survival, Cell Cycle, Cell-to-Cell Signaling and Interaction, Cellular Movement" Network. Data were analyzed through the use of Ingenity Pathway Analysis (IPA) (Ingenuity® Systems, http://www. ingenuity.com) [8].

is much higher than the second network above.

IPA analysis reports are summarized in Supplementary Table 8.

Protein structure modeling

We have examined protein features related to residues substituted by the identified 27 somatic mutations and found the features of 19 mutations (Table 3). The mutation of *CEBPA* generates a duplication of an amino acid that affects the original helical secondary structure. The residue corresponding to the mutation of *IDH2* has a function to bind with its substrate, so the substitutions of the residue decrease the binding affinity with the substrate. The protein region frame-shifted by the mutation of NPM1 is related to the function of nuclear localization, important for maintaining the structure. The structural modeling of RASGEF1B shows that the mutated residue Arg261 generates a new hydrogen bond with Asn303, affecting the stretch of neighbor helical structures. The protein structure of MSH4 shows that the salt bridge between Glu415 and Arg474 is essential for the folding and stability of the protein structure but is broken by the mutation (Table 4, Fig. 11).

Discussion

We have analyzed 10 exome sequencing samples and SNP chip data using a variety of bioinformatics methods. According to the concordance results, comparing with SNP chip

Table 3. Protein residues in which somatic mutation is located

Gene	Amino acid substitution	Uniprot features of corresponding residues in 10 similar proteins (expectation value = 1e-30)
ALDH6A1	p.Trp56Arg	Transit peptide (mitochondrion)
ARHGEF5	p.Arg1319His	DH domain, required for RHOG activation and mediates interaction with EPHA2
CEBPA	p.Lys313dup	Helix
FZD9	p.Gly343Arg	Cytoplasmic domain
GATA2	p.Leu359Val	GATA-type 2 zinc finger region
IDH2	p.Arg140Leu, p.Arg140Gln	Substrate binding, helix
KCNH1	p.Leu286Phe	Cytoplasmic domain
KRT9	p.Ala335Val	Coil 2, Rod, Necessary for interaction with PNN
MAST2	p.Arg795His	AGC-kinase C-terminal domain
MYH13	p.Pro309Arg	Myosin head-like domain
NPM1	p.Trp259Cysfs*12, p.Trp288Cysfs*12	Required for nucleolar localization, helix, mutagenesis site (destabilization of the structure)
RAD21	p.Leu413*	Cleavage by ESPL1
RASGEF1B	p.Trp261Arg	Ras-GEF domain
RBMXL2	p.Tyr243Asn	Arg/Gly/Pro-rich
RPS6KA1	p.Thr701Ala, p.Thr710Ala	Helix
TLL2	p.Asn911Thr	Disulfide bond, CUB 5 domain
TRA2B	p.Arg6Cys	Arg/Ser-rich (RS1 domain)
VDAC2	p.Leu213Ile	Transmembrane region (beta stranded)
ZNF317	p.Val62Phe	KRAB domain

Table 4. Protein structure modeling

Gene	RefSeq protein	Template PDB	Template PDB title	Coverage (%)
IDH2	NP_002159.2	1LWD	Porcine mitochondrial NADP+-dependent isocitrate dehydrogenase complexed with Mn2+ and isocitrate	91.4
RASGEF1B	NP_689758.1	3KSY	Histone domain, DH-PH unit, and catalytic unit of the Ras activator Son of Sevenless	77.4
MSH4	NP_002431.2	2O8B	Human MutSalpha (MSH2/MSH6) bound to ADP and a G T mispair	74.0
CACNB4	NP_001005747.1	1T3L	Voltage-dependent calcium channel beta subunit functional core complexed with the alpha 1 interaction domain	63.0

PDB, Protein Data Bank.

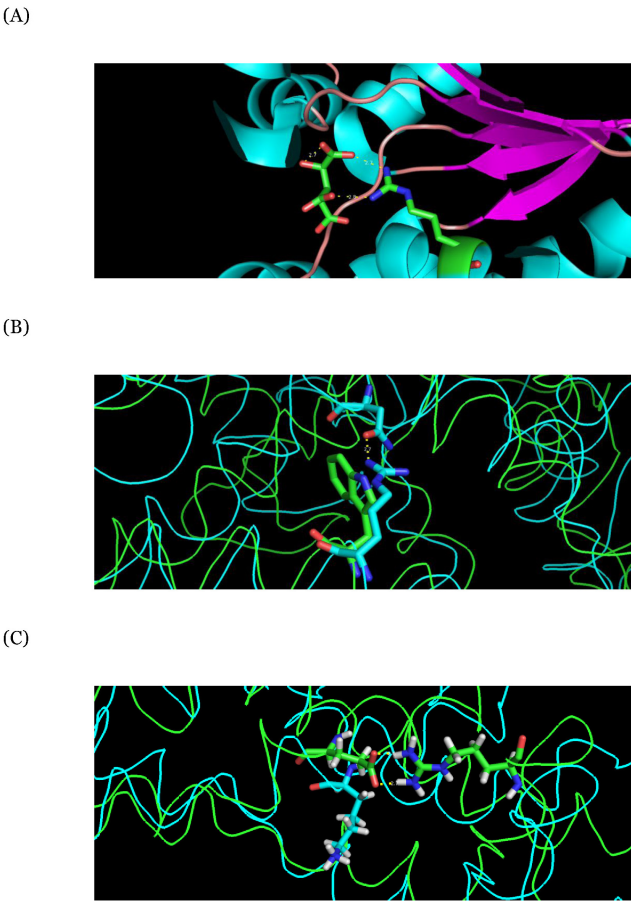


Fig. 11. Location of somatic mutation in protein structure. (A) IDH2. (B) RASGEF1B. (C) MSH4.

Table 5. Somatic mutation profile of three leukemia studies

AML	CLL	BC-CML
FLT3	*p53*	*p16/INK4A*
c-KIT	*TP53*	*p53*
NRAS	*SF3B1*	*RB1*
KRAS	*MYD88*	*RUNX1*
AML1	*FBXW7*	*ASXL1*
C/EBPalpha	*NOTCH1*	*IKZF1*
PU.1	*ZMYM3*	*WT1*
NPM1	*DDX3X*	*TET2*
CEBPA	*MAPK1*	*IDH1*
ASXL1	*POT1*	*NRAS*
DNMT3A	*CHD2*	*KRAS*
IDH1	*LRP1B*	*CBL*
IDH2	*Fas* (absent)	*TP53*
MLL	*Bcl-2* (overexpression)	*BCR-ABL* (translocation)
PHF6	*ATM* (germline and somatic mutation)	*EVI1* (overexpression)
TET2		*AML1* (translocation)
WT1		
ANKRD26		
GATA2		
RUNX1		
TP53		
PML-RARalpha (translocation)		
AML1-ETO (translocation)		
PLZF-RARalpha (translocation)		

AML, acute myeloid leukemia; CLL, chronic lymphocytic leukemia; BC-CML, Blast Crisis chronic myeloid leukemia.

data, we concluded that the G10.N exome sample is not the correct pair. Comparing SNPs between normal and tumor pairs revealed that the normal-tumor pair of sample G06 was not matched, and the G10 exome showed abnormal LOH in chrX but had no somatic mutation. So, we had removed G06 and G10 for discovering somatic mutations.

After removing G06 and G10, with 8 paired-end samples, we discovered 27 functionally important somatic mutations with significance. They included well-known genes in AML, such as *NPM1*, *IDH2*, *GATA2*, and *CEBPA*. Especially, *NPM1* and *GATA2* were found in 2 out of 8 samples.

In sample G05, we found one abnormal region at chr1 and an LOH region at chr13. We could conclude that the LOH region could potentially make a fusion gene by INTS6 and KIAA1704, and we did further network analysis on that. The results of IPA analysis gave "Cell Death and Survival, Cell Cycle" network as the top network, and it includes *GATA2*

and MAST2 (somatic mutation).

We had done IPA analysis with 26 genes, giving the "Cancer, Hematological Disease, Cell Cycle Network" as the top network, and it includes *NPM1*, *GATA2*, and *CEBPA*, which are well-known driver genes in AML [3]. Since those 3 genes have many edges in the result, we could conclude that those genes are indirect evidence of driver genes.

The second network is related with "Cell Death and Survival," and it includes *IDH2* and *PHF6* genes, the same as G05's top network. So far, our conclusion is that 8 samples are related with the "Cancer, Hematological Disease, Cell Cycle Network" but sample G05 is related with the "Cell Death and Survival Network."

We had searched previous works of genome-wide somatic mutation discovery and concluded that 8 exome samples might be AML. Somatic mutation profiles of blood cancer are summarized in Table 5 [3, 9, 10]. But, we could not find any somatic mutations in the FLT3 or RAS gene. Chronic lymphocytic leukemia and Blast Crisis chronic myeloid leukemia did not have somatic mutations in *NPM1* or *CEBPA*.

Seven out of 8 samples showed neither LOH nor SCNA, so we could conclude that those patients had a normal karyotype with no translocation or inversion. If possible, additional RNA-Seq experiments or another technique can provide enough evidence of karyotypes of those samples.

In *CEBPA*, *IDH2*, and *NPM1*, we saw somatic mutations located in helix regions. By using protein structure modeling, we found that *IDH2*, *RASGEF1B*, and *MSH4* could have structural variations, affecting protein functions. Somatic mutations in *IDH2* could break hydrogen bonds.

Supplementary materials

Supplementary data including eight tables and three figures can be found with this article online at http://www.genominfo.org/src/sm/gni-11-24-s001.pdf.

Acknowledgments

This work was supported by a National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (nos. 20110030770 and 20120009215), a grant from the Next-Generation BioGreen 21 Program (nos. PJ008019 and PJ008068), Rural Development Administration, and a grant from KRIBB Research Initiative Program, Republic of Korea.

References

1. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009;25:1754-1760.
2. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, *et al*. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* 2012;22:568-576.
3. Godley LA. Profiles in leukemia. *N Engl J Med* 2012;366:1152-1153.
4. Patel RK, Jain M. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PLoS One* 2012;7:e30619.
5. 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.
6. Liu X, Jian X, Boerwinkle E. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum Mutat* 2011;32:894-899.
7. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, *et al*. A census of human cancer genes. *Nat Rev Cancer* 2004;4:177-183.
8. Ingenuity Systems. Redwood City: Ingenuity Systems. Accessed 2013 Jan 1. Available from: http://www.ingenuity.com/.
9. Grossmann V, Kohlmann A, Zenger M, Schindela S, Eder C, Weissmann S, *et al*. A deep-sequencing study of chronic myeloid leukemia patients in blast crisis (BC-CML) detects mutations in 76.9% of cases. *Leukemia* 2011;25:557-560.
10. Wang L, Lawrence MS, Wan Y, Stojanov P, Sougnez C, Stevenson K, *et al*. *SF3B1* and other novel cancer genes in chronic lymphocytic leukemia. *N Engl J Med* 2011;365:2497-2506.