Discovery of somatic mutations in the progression of chronic myeloid leukemia by whole-exome sequencing

Y. Huang[1], J. Zheng[2], J.D. Hu[2], Y.A. Wu[1], X.Y. Zheng[2], T.B. Liu[2] and F.L. Chen[1]

[1]Provincial Clinical College, Fujian Medical University, Fuzhou, Fujian, China
[2]Fujian Institute of Hematology, Union Hospital, Fujian Medical University, Fuzhou, Fujian, China

Corresponding author: Y. Huang
E-mail: hyi8070@126.com

ABSTRACT. We performed whole-exome sequencing in samples representing accelerated phase (AP) and blastic crisis (BC) in a subject with chronic myeloid leukemia (CML). A total of 12.74 Gb clean data were generated, achieving a mean depth coverage of 64.45 and 69.53 for AP and BC samples, respectively, of the target region. A total of 148 somatic variants were detected, including 76 insertions and deletions (indels), 64 single-nucleotide variations (SNV), and 8 structural variations (SV). On the basis of annotation and functional prediction analysis, we identified 3 SNVs and 6 SVs that showed a potential association with CML progression. Among the genes that harbor the identified variants, *GATA2* has previously been reported to play important roles in the progression from AP to BC in CML. Identification of these genes will allow us to gain a better understanding of the pathological mechanism of CML and represents a critical advance toward new molecular diagnostic tests for the development of potential therapies for CML.

Key words: Exome sequencing; Chronic myeloid leukemia; Accelerated phase; Blastic crisis

INTRODUCTION

Chronic myeloid leukemia (CML) is characterized by the chromosomal translocation t(9;22)(q34;q11) leading to the Philadelphia chromosome (Ph) and generating the *Bcr-Abl1* fusion gene that encodes a constitutively active protein, tyrosine kinase (Goldman and Melo, 2003). Previous *in vitro* studies have shown that *Bcr-Abl1* is oncogenic in hematopoietic cells, promotes leukemic cell proliferation, and inhibits apoptosis (Lugo et al., 1990; Druker et al., 1996). The discovery that *Bcr-Abl1* mediates the pathogenesis of CML has provided the rationale for designing selective tyrosine kinase inhibitors (TKIs). Moreover, structural studies have facilitated the rational design of therapeutics targeting the tyrosine kinase activity of *Bcr-Abl1*. The impressive results obtained with the first agent of this kind, imatinib mesylate, has spurred the development of targeted therapies in cancer medicine. Clinical use of imatinib started in 1998 and it has shown remarkable effectiveness as a single agent for the treatment of CML patients, with most improvements seen in patients with chronic phase (CP) disease (Hochhaus and La Rosee, 2004).

However, these initial results were tempered by the fact that *Bcr-Abl1* transcripts are readily detectable in most patients receiving imatinib and that responses in the accelerated phase (AP) or the blastic crisis (BC) of the disease, when they occur, are generally short lived (Druker et al., 2001). Moreover, imatinib also inhibits the non-oncogenic C-Abl kinase in normal cells, and long-term administration of the drug can lead to cardiotoxicity (Kerkela et al., 2006). These findings have fueled interest in elucidating the mechanisms of resistance to TKI therapy and in developing novel agents to overcome these limitations.

Cancer genomes are enormously diverse and complex. The karyotypic heterogeneity of cancer genomes led to the proposal by Peter Nowell, in the mid-1970s, that tumorigenesis progresses by a stepwise evolutionary process (Nowell, 1976). Nowell suggested that, following an initiating event that converts a normal cell to a neoplastic cell, cancer progression results from the acquisition of genetic instability, leading to the accumulation of genetic alterations and the continual selective outgrowth of variant subpopulations of tumor cells with a proliferative advantage.

The development of massively parallel sequencing technologies makes it feasible to catalog all classes of somatically acquired mutations in a tumor (Ley et al., 2008; Stratton et al., 2009; Pleasance et al., 2010). However, a major challenge of cancer genome analysis is to identify "driver" mutations (Stratton et al., 2009), and several recent genome studies of leukemia and solid tumors have concentrated their analysis on coding regions (the "exome") to increase the likelihood of identifying these driver mutations (Dalgliesh et al., 2010; Morin et al., 2010). To gain new insight into pathogenesis, especially into the molecular basis underlying the increase in severity from AP to BC in CML, we carried out exome sequencing to follow the exacerbation process in CML.

MATERIAL AND METHODS

Sample description

This study was approved by the Institutional Review Boards of all involved institutions. The subject was diagnosed with CML (AP) at 42 years and was a male from Fujian Province, China. Signed informed consent documents were obtained from the patient before he was enrolled in the study. Self-report about family history and medical history was also acquired.

DNA extraction and exome sequencing

Genomic DNA was isolated from peripheral blood of the subject at 2 different phases by using a phenol-chloroform method. Following the manufacturer standard protocol, enrichment of coding exons and flanking intron regions was performed using a solution hybrid selection method with the SureSelect® Human All Exon 38M Kit (Agilent Technologies). The Truseq™ DNA Sample Prep Kit (Illumina) was used to prepare DNA libraries. Briefly, DNA was sheared by acoustic fragmentation (Bioruptor NGS, Diagenode, Liège, Belgium) and fragments about 325 bp long were extracted from electrophoresis gels by using the MinElute Gel Extraction Kit (Qiagen). The ends of DNA fragment were repaired and a single A nucleotide was added to the 3'-ends of the blunt fragments. A pair of adaptors was ligated to both ends of the fragments, followed by purification using AMPure XP beads (Beckman Coulter). The quality of the fragmentation and purification was assessed with a 2100 Bioanalyzer (Agilent). The fragments were then amplified by PCR and captured by hybridization to biotinylated RNA library baits. The resulting DNA library was quantified with the 2100 Bioanalyzer to check the size distribution of the libraries before paired-end sequencing with read length of 116 bp on the Illumina Genome Analyzer IIx platform.

Sequence read alignments

Sequence reads were mapped to the reference human genome (Ensembl GRCh37. p5/hg19) using the Burrows-Wheeler Aligner (Li and Durbin, 2009). First, the reference was indexed with the command "index"; 2nd, suffix array coordinates of good hits for each read were identified with the command "aln"; 3rd, suffix array coordinates were converted to chromosomal coordinates and pair reads using the command "sampe". Sequence Alignment/Map-formatted mapping result files were imported to Samtools (Li et al., 2009) for sorting and merging; the "rmdup" command was used to remove potential PCR duplicates. GATK (McKenna et al., 2010) IndelRealigner was used to minimize the number of mismatching bases caused by differences in insertions or deletions (indels) between the genomes of the subject and the reference. FixMateInformation in Picard, which was developed by the Sequencing Platform at the Broad Institute, was used to produce fixed, coordinate-sorted results for the realigned results.

Variant detection

We focused on somatic variants, and VarScan (v2.2.8) (Koboldt et al., 2012) was used to detect single-nucleotide variations (SNVs) and indels by comparing the samples from AP and BC. Pindel (v0.2.4) (Ye et al., 2009), Crest (v1.0) (Wang et al., 2011), and ExomeCNV (v1.3) (Sathirapongsasuti et al., 2011) were used for somatic structural variant (SV) detection. For Pindel, SVs were detected for both samples with the parameters set as: "-c ALL --min_size 21 --max_size 10,000 --min_supporting_reads 1". SVs that existed only in the BC sample were left for further analysis. For ExomeCNV, the "GATK coverage suite" was used to obtain the average sequencing depth of each exon and the $\log_2$ ratio was calculated to predicted gain or loss of CNVs. Parameters for Crest were set as default.

To concentrate on the variants potentially involved in the progression of CML, we

filtered for known SNVs or indels, which are annotated as unrelated with any phenotype in the Ensembl database (release 64). In addition, SNVs or indels existing in 3 or more persons according to the information from 39 samples from OpenGenome (http://opengenome.net) and the Han Chinese individual (YH) (Wang et al., 2008) were also screened. For SVs, those listed in the dbVAR database (overlap between an identified SV in this study and an SV in dbVAR accounting for more than 50% for both SVs was considered as the same SV) were filtered.

Annotation and functional prediction

All remaining variants were annotated on the basis of the gene information in the Ensembl database (release 64). For SNVs and indels, variants that caused frame shifts in coding sequences, disrupted splice sites, or nonsynonymously altered the amino acid sequence were selected as candidate risk sites associated with CML progression. For further confirmation on the impact of these selected variants on the structure and function of related proteins, MutationTaster (Schwarz et al., 2010) was used for functional prediction of indels and splicing sites, and other variants were analyzed with MutationTaster, PolyPhen-2 (Adzhubei et al., 2010), and SIFT (Kumar et al., 2009). SVs that overlapped with genes were considered as risk sites since they are larger than SNVs and indels, and their location within genes might result in functional damage of the encoded proteins. Genes with variants confirmed as risk candidates were mapped to the KEGG pathway databases (http://www.genome.jp/kegg/) to identify functional regulation mechanisms.

RESULTS

May-Giemsa staining of smear specimens of the bone marrow is shown in Figure 1. Overall, after filtering out reads of low quality, reads with a total length of 12.74 Gb were generated from the two samples. Approximately 4.82 Gb bases were aligned to the enriched target regions, achieving a mean coverage depth of 64.45 and 69.53 for AP and CP, respectively (Table 1).
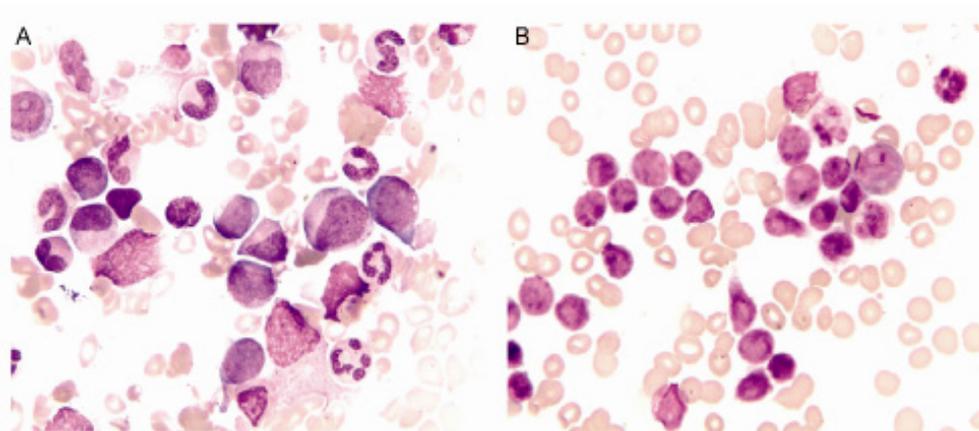

Figure 1. May-Giemsa staining of the smear specimens of the marrows. **A.** Accelerated phase; **B.** blastic crisis.

Table 1. Data production for the two samples.

Samples	Accelerated phase	Blastic phase
Clean data (Mb)	5.985	6.759
Data mapped to genome (Mb)	5.765	6.518
Data mapped to target region (Mb)	2.318	2.510
Target regions with coverage ≥10 x (Mb)	30.56	31.15
Mean depth of target region (x)	64.45	69.53
Coverage of target region (%)	95.31	95.85

The variant detection using VarScan (Koboldt et al., 2012) identified 140 somatic variants (76 indels and 64 SNVs). After the filtering process (see Methods) based on information obtained from the Ensembl database (release 64), OpenGenome (http://opengenome. net) and the Han Chinese individual (YH) (Wang et al., 2008), 121 variants (72 indels and 49 SNVs) potentially associate with CML progression were recovered and annotated (Table 2). Eight SVs were detected and 6 were obtained after the filtering process.

Table 2. Statistics of annotated variants identified by VarScan after filtering process.

Variants	SNV	Indels
Non-synonymous	4	1
Disrupted splice sites	1	3
Reading frame shift	N/A	0
Synonymous	1	N/A
UTR	2	4
Intronic	27	55
Downstream	16	29
Upstream	17	16
Regulatory region	6	10
Intergenic	6	0

Variants are annotated based on transcripts, that is, one variant may be assigned as different types according to different transcripts. Therefore, overlap exists in different cells from the same column in this table. N/A = no reading frame shift detected.

According to the annotation results, 4 indels and 5 SNVs, annotated as disrupting splice sites or altering amino acid sequence (Table 3) and 6 SVs (Table 4) detected only in BC were considered as candidate risk sites correlated with CML progression. Further functional prediction results from MutationTaster (Schwarz et al., 2010) for indels and splicing sites did not show any support for their impact on the structure and function of related proteins. However, for nonsynonymous SNVs, the prediction results supported by at least two software packages showed that 3 SNVs, located on gene *ANO5* (anoctamin 5, chromosome 11), *GATA2* (GATA-binding protein 2, chromosome 3), and *MAX* (MYC-associated factor X, chromosome 14) were predicted to be deleterious mutations, suggesting that they may be functionally or structurally associated with CML progression from AP to BC. The SNV of the *MAX* gene was previously identified and is named rs80206158. Another SNV, present in *ANGEL1* (angel homolog 1 on chromosome 14), was also predicted to cause disease, supported only by MutationTaster. Only nonsynonymous SNVs that were supported by 2 or more functional prediction results were included in further analysis to avoid false-positive SNVs. The 6 SVs overlapped with the *ENO1*, *COL3A1*, *AFP*, *SERPINA1*, *MGAT5B*, and *ZNF208*.

Table 3. Candidate variants related with CML progression.

Chr	Position (bp)	Reference	Variant	Genotype 1	Genotype 2	Gene	Mutation type	PolyPhen-2	SIFT	Mutation taster
1	92979093	TATATATATATATATA TATATATATATATATG	-	0/0	0/1	EVI5	Indel, splice site	-	-	-
2	37599620	TG	-	0/0	0/1	QPCT	Indel, splice site	-	-	-
3	128200726	C	A	0/0	0/1	GATA2	Non-synonymous	Possibly damaging	Deleterious	Disease causing
7	75651314	A	-	0/0	0/1	STYXL1	Indel, splice site	-	-	-
10	118351922	T	G	0/0	0/1	PNLIPRP1	Splice site	-	-	-
11	22284549	G	A	0/0	0/1	ANO5	Non-synonymous	Possibly damaging	Deleterious	Disease causing
14	65543318	T	G	0/0	0/1	MAX	Non-synonymous	Possibly damaging	Deleterious	-
14	77275753	C	A	0/0	0/1	ANGEL1	Non-synonymous	-	-	Disease causing
22	29885576	-	TGAGAAGGC CAAGTCCCC	0/0	0/1	NEFH	Non-synonymous	-	-	Disease causing

Chr = chromosome; Position (bp) is according to Ensembl GRCh37.p5/hg19; Reference = reference allele according to Ensembl GRCh37.p5/hg19; (–) means where deletion happens; Variant = nucleotide variation found by exome sequencing, (-) means where deletion happens; Genotype 1 = genotype of the patients at accelerated phase, the values are 0 for reference allele and 1 for the variant allele. Genotype 2 = genotype of the patient at blast crisis, the values mean the same as Genotype 1. Gene = official symbol provided by HGNC; the last three columns are analyses of the predicted effect of the substitution on protein function from PolyPhen-2, SIFT and Mutation Taster, respectively, (-) means no prediction available.

Table 4. Candidate structure variations related with CML progression.

Chr	Source	Type	Start	End	Gene
1	Crest	Deletion	8924153	8925344	ENO1 (8921061-8938749)
2	Crest	Deletion	189873948	189874903	COL3A1 (189839046-189877472)
4	Crest	Deletion	74318343	74319482	AFP (74301880-74321891)
14	Crest	Deletion	94847474	94848925	SERPINA1 (94843084-94857029)
17	Pindel	Deletion	74901901	74901925	MGAT5B (74864538-74946459)
19	ExomeCNV	Deletion	22153966	22154926	ZNF208 (22153485-22193745)

Chr = chromosome; Source = the software package used for structural variation (SV) detection; Start = start point (bp) of the SV; End = end point (bp) of the SV; Gene = genes overlapped with SVs, physical positions of the genes are indicated in parentheses.

According to the KEGG pathway mapping analysis, *MAX* is involved in cancer-related pathways, including pathways annotated in KEGG with the following IDs: hsa05200 (pathways in cancer), hsa05202 (transcriptional misregulation in cancer) and hsa05222 (small cell lung cancer). *SERPINA1* is involved in the pathway of complement and coagulation cascades (KEGG ID, hsa04610). *MGAT5B* is involved in glycan biosynthesis and metabolic pathways (hsa00510, hsa00514 and hsa01100). No pathways associated with the other genes were identified.

DISCUSSION

This study used whole-exome sequencing of a subject with CML developing from AP to BC to identify genetic variants associated with CML progression. To uniquely identify the causative variants, we used information in public databases to remove variants known not be associated with any phenotype. Among the remaining SNVs and indels, we selected nonsynonymous (including nonsense variants) and splice-site variants as candidates for further analysis. Functional prediction analysis helped to further confirm the impact on the progression of CML of these candidate variants. Finally, we identified 3 SNVs that were associated with CML progression located on the *ANO5*, *GATA2*, and *MAX* genes. All of these 3 SNVs are suggested to cause disease according to the functional prediction analysis. The pathway-mapping analysis indicated that *MAX* is involved in cancer-related pathways, including those annotated as hsa05200, hsa05202 and hsa05222. Six candidate SVs implicated to be involved in CML progression and that overlapped with the *ENO1*, *COL3A1*, *AFP*, *SERPINA1*, *MGAT5B*, and *ZNF208* were identified. *SERPINA1* is involved in the pathway of complement and coagulation cascades and *MGAT5B* is involved in glycan biosynthesis and other metabolic pathways.

GATA2, which is expressed abundantly in hematopoietic stem cell and progenitor cells, is a transcription factor that plays an essential role in the development and proliferation of hematopoietic and endocrine cell lineages (Ostergaard et al., 2011). A previous report (Zhang et al., 2008) showed that the *GATA2* L359V mutation causes a block in myelo-monoblastic differentiation and plays an important role in myeloid BC of CML, implicating *GATA2* in the progression from AP to BC in CML. The SNV discovered in this study has not been previously reported; its mutation from C to A causes deleterious effects on the structure and function of the GATA2 protein, which may contribute to the transformation from AP to BC.

The MAX protein forms homo- and heterodimers with other family members, including c-Myc, an oncoprotein contributing to the genesis of many human cancers (Dang, 2012). The heterodimers bind to a common DNA target site (the E box) in the regulatory regions of

target genes that are involved in cell cycle control, energetic metabolism, protein or ribosome synthesis, and other miscellaneous functions (Vaque et al., 2005). The results of the pathway analysis suggested that Max is involved in cancer-related pathways through its interaction with Myc. In fact, interactions between c-Myc and Bcr, which is the fusion partner of Abl, have been observed by Mahon et al. (2003). However, according to Mahon et al., Bcr cannot interact with Max, or with the c-Myc•Max heterodimer. Thus, the association between Max and CML progression may not be due to the interaction between Bcr and Max or the c-Myc•Max heterodimer. As previously described (Zada et al., 2006), Max is a protein interacting with the transcription factor CCAAT/enhancer-binding protein α (C/EBPα), which plays an important role in the regulation of granulopoiesis and is disrupted in human acute myeloid leukemia (AML). The BC stage of CML behaves like AML, with rapid progression and short survival, suggesting the potential for a role of the Max-C/EBPα pathway in the progression from AP to BC.

Besides, interruption of *ENO1* by SVs may contribute to the tumor exacerbation because alternative splicing of *ENO1* creates a shorter isoform that binds to the c-Myc promoter and functions as a tumor suppressor (Feo et al., 2000). Relationships of other genes with the pathogenesis of leukemia have not yet been reported. These genes are located on different chromosomes, reflecting the fact that cancer genomes are complex and the progression or exacerbation of cancer results from the acquisition and accumulation of various genetic alterations.

In summary, we performed whole-exome sequencing on samples from AP and BC in CML and identified 3 SNVs and 6 SVs that were associated with CML progression. Our observations suggest that mutations in 9 genes may play important roles in exacerbation progress from AP to BC in CML, supported by previous independent studies on *GATA2*. Moreover, new CML-associated variations identified in this study will help improve the understanding of the molecular mechanisms underlying the progression from AP to BC in CML, and facilitate the discovery of potential drug targets.

ACKNOWLEDGMENTS

Conflicts of interest

The authors declare no conflict of interest.

REFERENCES

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, et al. (2010). A method and server for predicting damaging missense mutations. *Nat. Methods* 7: 248-249.

Dalgliesh GL, Furge K, Greenman C, Chen L, et al. (2010). Systematic sequencing of renal carcinoma reveals inactivation of histone modifying genes. *Nature* 463: 360-363.

Dang CV (2012). MYC on the path to cancer. *Cell* 149: 22-35.

Druker BJ, Tamura S, Buchdunger E, Ohno S, et al. (1996). Effects of a selective inhibitor of the Abl tyrosine kinase on the growth of Bcr-Abl positive cells. *Nat. Med.* 2: 561-566.

Druker BJ, Sawyers CL, Kantarjian H, Resta DJ, et al. (2001). Activity of a specific inhibitor of the BCR-ABL tyrosine

kinase in the blast crisis of chronic myeloid leukemia and acute lymphoblastic leukemia with the Philadelphia chromosome. *N. Engl. J. Med.* 344: 1038-1042.

Feo S, Arcuri D, Piddini E, Passantino R, et al. (2000). ENO1 gene product binds to the c-myc promoter and acts as a transcriptional repressor: relationship with Myc promoter-binding protein 1 (MBP-1). *FEBS Lett.* 473: 47-52.

Goldman JM and Melo JV (2003). Chronic myeloid leukemia - advances in biology and new approaches to treatment. *N. Engl. J. Med.* 349: 1451-1464.

Hochhaus A and La Rosee P (2004). Imatinib therapy in chronic myelogenous leukemia: strategies to avoid and overcome resistance. *Leukemia* 18: 1321-1331.

Kerkela R, Grazette L, Yacobi R, Iliescu C, et al. (2006). Cardiotoxicity of the cancer therapeutic agent imatinib mesylate. *Nat. Med.* 12: 908-916.

Koboldt DC, Zhang Q, Larson DE, Shen D, et al. (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* 22: 568-576.

Kumar P, Henikoff S and Ng PC (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* 4: 1073-1081.

Ley TJ, Mardis ER, Ding L, Fulton B, et al. (2008). DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* 456: 66-72.

Li H and Durbin R (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25: 1754-1760.

Li H, Handsaker B, Wysoker A, Fennell T, et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25: 2078-2079.

Lugo TG, Pendergast AM, Muller AJ and Witte ON (1990). Tyrosine kinase activity and transformation potency of bcr-abl oncogene products. *Science* 247: 1079-1082.

Mahon GM, Wang Y, Korus M, Kostenko E, et al. (2003). The c-Myc oncoprotein interacts with Bcr. *Curr. Biol.* 13: 437-441.

McKenna A, Hanna M, Banks E, Sivachenko A, et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20: 1297-1303.

Morin RD, Johnson NA, Severson TM, Mungall AJ, et al. (2010). Somatic mutations altering EZH2 (Tyr641) in follicular and diffuse large B-cell lymphomas of germinal-center origin. *Nat. Genet.* 42: 181-185.

Nowell PC (1976). The clonal evolution of tumor cell populations. *Science* 194: 23-28.

Ostergaard P, Simpson MA, Connell FC, Steward CG, et al. (2011). Mutations in GATA2 cause primary lymphedema associated with a predisposition to acute myeloid leukemia (Emberger syndrome). *Nat. Genet.* 43: 929-931.

Pleasance ED, Stephens PJ, O'Meara S, McBride DJ, et al. (2010). A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* 463: 184-190.

Sathirapongsasuti JF, Lee H, Horst BA, Brunner G, et al. (2011). Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. *Bioinformatics* 27: 2648-2654.

Schwarz JM, Rodelsperger C, Schuelke M and Seelow D (2010). MutationTaster evaluates disease-causing potential of sequence alterations. *Nat. Methods* 7: 575-576.

Stratton MR, Campbell PJ and Futreal PA (2009). The cancer genome. *Nature* 458: 719-724.

Vaque JP, Navascues J, Shiio Y, Laiho M, et al. (2005). Myc antagonizes Ras-mediated growth arrest in leukemia cells through the inhibition of the Ras-ERK-p21Cip1 pathway. *J. Biol. Chem.* 280: 1112-1122.

Wang J, Wang W, Li R, Li Y, et al. (2008). The diploid genome sequence of an Asian individual. *Nature* 456: 60-65.

Wang J, Mullighan CG, Easton J, Roberts S, et al. (2011). CREST maps somatic structural variation in cancer genomes with base-pair resolution. *Nat. Methods* 8: 652-654.

Ye K, Schulz MH, Long Q, Apweiler R, et al. (2009). Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* 25: 2865-2871.

Zada AA, Pulikkan JA, Bararia D, Geletu M, et al. (2006). Proteomic discovery of Max as a novel interacting partner of C/EBPalpha: a Myc/Max/Mad link. *Leukemia* 20: 2137-2146.

Zhang SJ, Ma LY, Huang QH, Li G, et al. (2008). Gain-of-function mutation of GATA-2 in acute myeloid transformation of chronic myeloid leukemia. *Proc. Natl. Acad. Sci. U. S. A.* 105: 2076-2081.