Table 1. MDS/sAML Patients with *C-CBL* Mutations.

Patient No.	Age (Year)/Sex	WHO Subtype	Marrow Blasts (%)	Cytogenetics at MDS	C-CBL Amino Acid Change		IPSS	Time to AML (Months)	Survival from MDS (Months)
			MDS/sAML		MDS	sAML			
001	70/F	RAEB-1/sAML	5.5/71.7	NA	—	Y371S	>1.0	3.5	3.9
010	48/F	RAEB-2/sAML	17.0/32.2	46,XX,dup(1)(q21q32)[28/28]	—	F418S	2.5	5.3	28.1
032	72/M	RAEB-2/sAML	16.6/79.0	46,XY,-5,-8,-9,add(11)(q25), t(12;18)(p11;p11),-17,+4mar[6/20]	—	L370_Y371 ins L	3.0	2.2	3.3
109	22/M	RCMD/sAML	1.4/35.2	NA	—	L399V	≥0.5	7.0	7.6
119	54/F	RCMD/sAML	5.0/59.0	NA	G415S	G415S	1.0	22.3	25.2
125	64/F	RAEB-1/sAML	8.5/45.8	45,XX,-7[22/26]/46,XX[4/26]	—	C416W	1.5	14.9	23.8+

F indicates female; M, male; NA, not available; WHO, World Health Organization.

of survival were calculated according to the Kaplan-Meier method. Comparisons of estimated survival curves were analyzed by the log-rank test. Statistical analyses were carried out by software SPSS 17.0 (SPSS, Inc, Chicago, IL). In all analyses, *P* values were two-sided and considered statistically significant when values lower than .05.

Results

C-CBL *Mutations in Paired Samples of MDS and sAML*

C-CBL mutation was detected in only 1 of 167 *de novo* high-risk MDS at the initial diagnosis. Eighty-six patients progressed to sAML with a median time of 9.8 months (range = 1.0-143.1 months). Of the 51 paired MDS/sAML samples, 1 patient (no. 119) with RCMD had a *C-CBL* mutation located at C-terminal of the RF domain (G415S) at initial diagnosis; she retained the same *C-CBL* mutation at sAML evolution. The other five patients acquired *C-CBL* mutations during sAML transformation. Patient no. 109 with sAML transformed from RCMD gained a missense mutation at the linker region (L399V). Two sAML patients (nos. 001 and 125) transformed from RAEB-1 gained a mutation at the linker region (Y371S) and the RF domain (C416W), respectively. Of the two RAEB-2 patients who acquired *C-CBL* mutations during sAML transformation, one (no. 010) had a mutation at the RF domain (F418S) and the other (no. 032) had an insertion mutation (L370_Y371 ins L) at the linker region. The frequency of *C-CBL* mutations increased from 0.6% (1/167) in the MDS phase to 11.8% (6/51) at sAML transformation. The clinicohematological features and the *C-CBL* mutation status of the six paired BM samples at both MDS and sAML phases are shown in Table 1.

Figure 1 shows the sequencing electropherograms of the six paired BM MDS/sAML samples carrying *C-CBL* mutations and the CNAG output for the four sAML samples available for SNP array analysis. Patient no. 001 acquired a homozygous mutation (Y371S) during sAML evolution (Figure 1*A*). Patient no. 010 was negative for *C-CBL* mutation at the initial diagnosis of MDS. She gained a small *C-CBL* mutant clone in the follow-up sample 3.5 months later when her disease was still in the MDS phase, and then she had an expansion of *C-CBL* mutant clone at the sAML phase 6 months after the diagnosis of MDS. The SNP array analysis showed the presence of 11q-UPD at the sAML phase; the UPD-positive cells were 33% calculated by the observed difference in allele-specific copy number (ASCN) divided by the expected value (Figure 1*B*), implying that the presence of a homozygous mutation in a fraction of cells in patient no. 010. *C-CBL* mutation was not detected in patient no. 032 at diagnosis of MDS but a small mutant clone was identified at the sAML phase (Figure 1*C*, *right upper panel*). Because the allelic burden of the mutant clone was very low, the PCR product was then cloned into the PCRII-TO PO vector (Invitrogen). Twenty-six clones were subsequently sequenced and six mutant clones were obtained, of which the sequence confirmed the presence of L370_Y371L shown in Figure 1*C*, *left lower panel*. The SNP array analysis revealed a small deletion and an amplification at 11q23.3 in the sAML sample. Patient no. 109 acquired a missense mutation (L399V) only after sAML transformation (Figure 1*D*). The only one patient (no. 119) harboring the identical *C-CBL* mutations at both MDS and sAML phases carried a small subclone of mutant at the MDS phase and progressed to a higher level, which slightly exceeded the wild-type allele. SNP array analysis for the sAML sample did not reveal any abnormality in 11q23.3 (Figure 1*E*). Patient no. 125, negative for *C-CBL* mutation at MDS phase, acquired a missense mutation (C416W) after sAML transformation (Figure 1*F*). SNP array analysis did not show an abnormal finding in 11q23.3 at the sAML phase.

Other Genetic Abnormalities in MDS/sAML Patients Harboring C-CBL *Mutations*

Coexistence of additional gene mutations in *C-CBL* mutated patients was found in four of six patients (Table 2). Patient nos. 010, 119, and 125 acquired N-*RAS* mutation, *JAK2*^{V617F}, and *PTPN11* mutation, respectively, during sAML transformation. We did not find any cooperating mutation involving receptor TKs (RTKs) or the RAS pathway with *C-CBL* gene in patient nos. 001, 032, and 109. In addition, patient no. 125 had evidence of cytogenetic clonal evolution, 45,XX,-7[22/26]/46,XX[4/26] at MDS and 45,XX,-7[20/25]/45,XX,-7,del(16)(q12.1)[2/25]/46,XX[3/25] at sAML.

Clinicohematological Features and Outcome of MDS/sAML with C-CBL *Mutations*

Of the 51 patients, there was no difference in age, sex, hemoglobin level, platelet counts, white blood cell counts, percentage of blasts in BM or peripheral blood, cytogenetics, or IPSS (≤1.5 *vs* ≥2.0) between *C-CBL* mutation–positive and –negative groups at both MDS and sAML. The time to sAML transformation and the survival from the diagnosis of MDS in the six patients who harbored *C-CBL* mutations at sAML phase are shown in Table 1. Because only one MDS patient harbored *C-CBL* mutation at the initial diagnosis in the whole cohort of MDS patients, it precluded a meaningful analysis of the mutation status on the risk and time to sAML transformation or overall survival. No significant difference in overall survival from the diagnosis of sAML was observed regarding *C-CBL* mutation status ($n = 51$, estimated overall survival = 1.1 [95% confidence interval {CI} = 0-3.7] *vs* 5.6 [95%

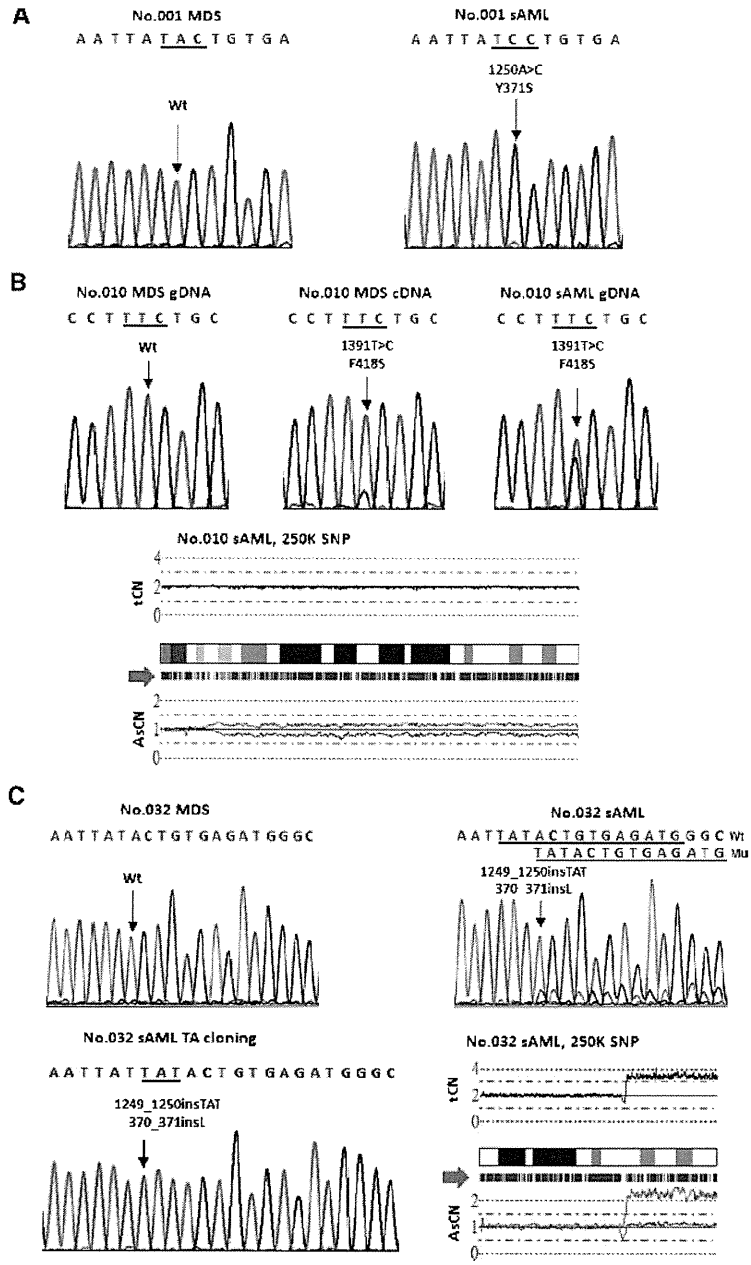

Figure 1. Chromatograms of *C-CBL* mutations in six paired MDS/sAML samples (Wt, wild-type; Mu, mutant) and CNAG output for total copy number (tCN) and allele-specific copy number (AsCNs) in the long arm of chromosome 11 in four patients at the sAML phase. The green bars below each ideogram of 11q indicate the position of heterozygous SNP calls (red arrowhead). Dissociation of AsCN plot indicates the presence of UPD in 11q. (A) Patient no. 001 acquired a homozygous missense mutation (Y371S) after sAML transformation. (B) Patient no. 010 had wild-type *C-CBL* gene at the initial diagnosis of RAEB-1, acquired a small missense mutant clone (C418S) later, and the level of the mutant increased further at sAML phase. SNP array analysis at sAML revealed an 11q-UPD. (C) Patient no. 032, negative for *C-CBL* mutation at the MDS phase, acquired L370_Y371 ins L at sAML phase. The small *C-CBL* mutant clone was confirmed by TA cloning. SNP array analysis showed a small deletion and an amplification at 11q23.3 at the sAML phase. (D) Patient no. 109 acquired a missense mutation (L399V) only after sAML transformation. (E) Patient no. 119 had a small subclone of G415S mutant at the MDS phase, which was more clearly shown in reverse complement and expanded during sAML transformation. SNP array analysis for the sAML sample did not reveal any abnormality at 11q23.3. (F) Patient no. 125, negative for *C-CBL* mutation at MDS phase, acquired a missense mutation (C416W) after sAML transformation. SNP array analysis did not show abnormal findings in 11q23.3 at sAML phase.

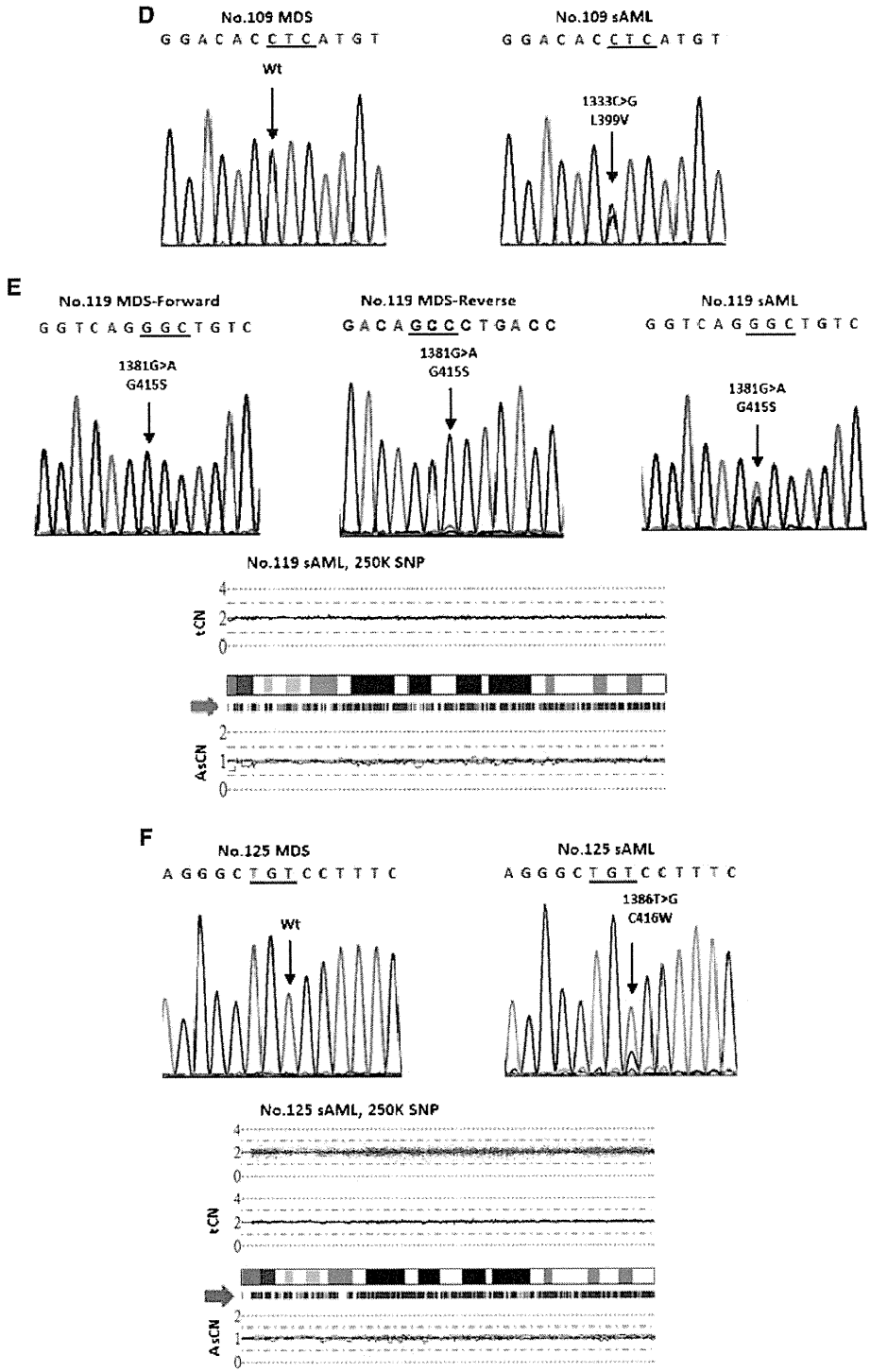

Figure 1. (*continued*).

CI = 2.9-8.3] months; $P = .958$; Figure 2) in patients with sAML evolved from MDS.

Discussion

In the present study, an analysis of *C-CBL* mutations in the matched paired BM samples from 51 patients at both phases of MDS and sAML was performed. We found that only one MDS patient had *C-CBL* mutation at initial presentation, and additional five patients acquired *C-CBL* mutations during the disease progression to sAML. The frequency of *C-CBL* mutations at MDS was very low and markedly increased at sAML transformation (1/167 *vs* 6/51). To the best of our knowledge, the present series is the first longitudinal and

Table 2. Additional Gene Mutations in MDS/sAML Patients Harboring *C-CBL* Mutations.

Patient No.	Mutation Status at MDS/sAML Phase								
	C-CBL Amino Acid Change	N-*RAS*	K-*RAS*	*FLT3*-ITD	*FLT3*-TKD (D835)	*JAK2*^{V617F}	*PTPN11*	C-*KIT*	C-*FMS*
001	-/+ (Y371S)	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-
010	-/+/+ (F418S)*	-/+ (Q61H)	-/-	-/-	-/-	-/-	-/-	-/-	-/-
032	-/+ (L370_Y371 ins L)	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-
109	-/+ (L399V)	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-
119	+/+ (G415S)	-/-	-/-	-/-	-/-	-/++	ND/-	-/-	ND/-
125	-/+ (C416W)	-/-	-/-	-/-	-/-	-/-	-/+ (Q510L)	-/-	ND/ND

++ indicates homozygous mutation; ND, not done.
*Mutation status at MDS/MDS/sAML phases.

systematical study that demonstrated the acquisition and/or clonal expansion of *C-CBL* mutations in the progression of MDS to sAML.

Because signaling of RTK-activating mutations, RAS pathways, and *C-CBL* mutations are similar in cell models, we also performed the mutational analyses for these genes. Coexistence of *C-CBL* mutations with other gene mutations involving the RTKs and RAS pathways was common in our patients at the sAML phase. Three patients had additional mutations of N-*RAS*, *JAK2*, or *PTPN11* genes during sAML evolution. The only one MDS patient who retained the identical *C-CBL* mutant clones at sAML transformation acquired *JAK2*^{V617F} during disease progression. These observations suggested that acquisition of *C-CBL* mutation collaborating with other gene mutations played a role in the transformation of sAML from MDS. Furthermore, a clonal cytogenetic evolution was also detected in one patient (no. 125) during sAML progression. For those who did not harbor *C-CBL* mutations at sAML phase, 10 patients acquired the activating mutations in RTKs and/or RAS pathways (data not shown). Our result showed

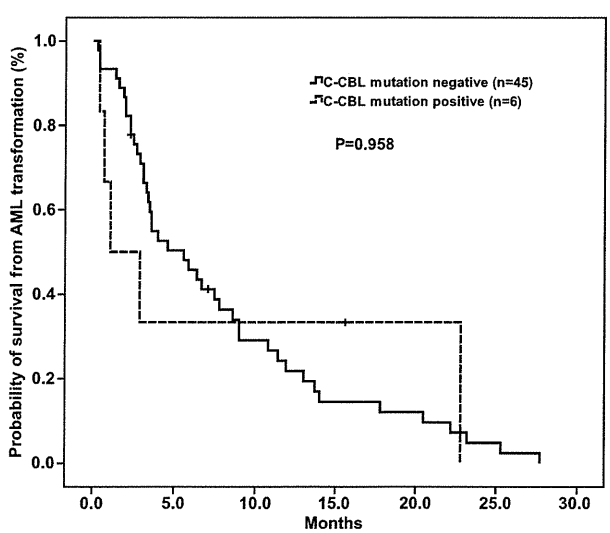

Figure 2. Kaplan-Meier estimates of overall survival in sAML patients according to *C-CBL* mutation status. The survival from the diagnosis of sAML was 1.1 months (95% CI = 0-3.7) for the patient with *C-CBL* mutation compared with the estimated median survival of 5.6 months (95% CI = 2.9-8.3) in *C-CBL* mutation–negative patients (P = .958).

that *C-CBL* mutations constituted one of the accumulated genetic alternations associated with the progression of MDS to sAML. Patient nos. 001, 032, and 109, who acquired *C-CBL* mutations at sAML transformation, had only *C-CBL* mutations detected among the genetic lesions we analyzed, suggesting that *C-CBL* mutations might play a major role in the disease progression or cooperate with other genetic abnormalities not examined in the present study for these patients.

Patient no. 010 with wild-type *C-CBL* gene at initial diagnosis of MDS acquired *C-CBL* mutation later when her disease was still in the MDS phase and the mutant level increased further in the sAML phase. The presence of UPD-positive cells in a subfraction of cells (33%) calculated by a signal ratio attributed to the presence of 11qUPD at the sAML phase in this case without accompanying a homozygous sequencing electropherogram. In patient no. 119 carrying *C-CBL* mutations in both MDS and sAML phases, the very small *C-CBL* mutant clone expanded to a slightly predominant clone during sAML evolution in the absence of UPD in the CNAG output. The observed discrepancy might be explained as an allele measurement on sequencing electropherogram probably not accurately enough to conclude the predominance of the mutant allele based on such a subtype difference of signals. Barresi et al. [14] also described an expansion of *C-CBL* mutated subclone occurred in a case during MDS progression to sAML. This finding indicated that the *C-CBL* mutated subclone conferred a growth advantage when MDS progressed to sAML. Acquisition with expansion of *C-CBL* mutated clones was also reported in one patient during the progression of primary myelofibrosis to sAML [30].

The sequencing analysis showed that five of the six *C-CBL* mutations in MDS or sAML were missense mutations; the remaining one was an insertion mutation. All of the *C-CBL* mutations found in our patients involved the linker region or RF domain that is central to the E3 ubiquitin ligase activity [5,6]. Because we analyzed mutations specifically at exons 7, 8, and 9, *C-CBL* mutations located outside exons 7 to 9 would not be detected in the present study. It is of note that most of the *C-CBL* mutations reported previously was missense mutation. Insertion mutations were only described in one AML patient with ins(SK366) at intron7/exon8 splice site [31]. The mutation character of three-base insertion that leads to L370_Y371 ins L without frameshift at the linker region of C-CBL, which was verified by cloning analysis, had not been described before. L370_Y371 ins L might cause conformational change of the Linker region and result in decreased E3 activity. It has been found that homozygous *C-CBL* mutations were frequently observed in patients with CMML and strongly associated with 11q aUPD [9,13,19]. In the present study, we found that UPD seemed to be less frequent in patients with MDS/sAML.

Whether *C-CBL* mutations confer unique clinical characteristics is not clearly defined. We did not find any association between *C-CBL* mutations and specific clinicohematological features at the initial presentation or at sAML. In the literature review along with the present result, except for one with refractory anemia [19], MDS patients harboring *C-CBL* mutations were mostly of RAEB or RCMD subtypes [9,13–15]. Current evidences suggested that *C-CBL* mutations were associated with more aggressive types of MDS. The impact of *C-CBL* mutations on clinical outcome in MDS is not known because of the rare occurrence of the mutation at the initial diagnosis. The clinical and prognostic relevance of *C-CBL* mutation on MDS remains to be determined by a larger cohort of patients. We did not observe a survival impact of *C-CBL* mutation in sAML patients. Our patients received different treatment options both at MDS and sAML stages; this might have influences on survival analysis. Nevertheless, sAML has a very poor prognosis; the cohort of patients that carrying *C-CBL* mutations in this series would have a poor outcome. It is the fact that *C-CBL* mutation in sAML evolution is of biologic relevance but not necessary of prognostic importance.

To sum up, we analyzed *C-CBL* mutations in matched paired BM samples of patients with high-risk *de novo* MDS at initial presentation and sAML. Our results showed that *C-CBL* mutations were rare in MDS at the presentation, but acquisition and/or expansion of *C-CBL* mutated clones occurred not infrequently during its progression to sAML. The higher occurrence of *C-CBL* mutations at sAML transformation in patients with MDS suggested that *C-CBL* mutation might play a role, either dominantly or cooperatively with other genetic abnormalities, in a subset of MDS patients during sAML evolution.

Acknowledgments

The authors thank Yu-Feng Wang for her secretarial assistance.

References

[1] Tefferi A and Vardiman JW (2009). Mechanisms of disease: myelodysplastic syndromes. *N Engl J Med* **361**, 1872–1885.

[2] Quesnel B, Guillerm G, Vereecque R, Wattel E, Preudhomme C, Bauters F, Vanrumbeke M, and Fenaux P (1998). Methylation of the $p15^{INK4b}$ gene in myelodysplastic syndromes is frequent and acquired during disease progression. *Blood* **91**, 2985–2990.

[3] Shih LY, Huang CF, Wang PN, Wu JH, Lin TL, Dunn P, and Kuo MC (2004). Acquisition of FLT3 or N-*ras* mutations is frequently associated with progression of myelodysplastic syndrome to acute myeloid leukemia. *Leukemia* **18**, 466–475.

[4] Nau MM and Lipkowitz S (2003). Comparative genomic organization of the *cbl* genes. *Gene* **308**, 103–113.

[5] Thien CB and Langdon WY (2001). Cbl: many adaptations to regulate protein tyrosine kinases. *Nat Rev Mol Cell Biol* **2**, 294–307.

[6] Schmidt MH and Dikic I (2005). The Cbl interactome and its functions. *Nat Rev Mol Cell Biol* **6**, 907–918.

[7] Lee PS, Wang Y, Dominguez MG, Yeung YG, Murphy MA, Bowtell DD, and Stanley ER (1999). The Cbl protooncoprotein stimulates CSF-1 receptor multiubiquitination and endocytosis, and attenuates macrophage proliferation. *EMBO J* **18**, 3616–3628.

[8] Murphy MA, Schnall RG, Venter DJ, Barnett L, Bertoncello I, Thien CB, Langdon WY, and Bowtell DD (1998). Tissue hyperplasia and enhanced T-cell signalling via ZAP-70 in c-Cbl–deficient mice. *Mol Cell Biol* **18**, 4872–4882.

[9] Sanada M, Suzuki T, Shih LY, Otsu M, Kato M, Yamazaki S, Tamura A, Honda H, Sakata-Yanagimoto M, Kumano K, et al. (2009). Gain-of-function of mutated C-CBL tumour suppressor in myeloid neoplasms. *Nature* **460**, 904–908.

[10] Zeng S, Xu Z, Lipkowitz S, and Longley JB (2005). Regulation of stem cell factor receptor signaling by Cbl family proteins (Cbl-b/c-Cbl). *Blood* **105**, 226–232.

[11] Mohamedali A, Gäken J, Twine NA, Ingram W, Westwood N, Lea NC, Hayden J, Donaldson N, Aul C, Gattermann N, et al. (2007). Prevalence and prognostic significance of allelic imbalance by single-nucleotide polymorphism analysis in low-risk myelodysplastic syndromes. *Blood* **110**, 3365–3373.

[12] Raghavan M, Lillington DM, Skoulakis S, Debernardi S, Chaplin T, Foot NJ, Lister TA, and Young BD (2005). Genome-wide single nucleotide polymorphism analysis reveals frequent partial uniparental disomy due to somatic recombination in acute myeloid leukemias. *Cancer Res* **65**, 375–378.

[13] Dunbar AJ, Gondek LP, O'Keefe CL, Makishima H, Rataul MS, Szpurka H, Sekeres MA, Wang XF, McDevitt MA, and Maciejewski JP (2008). 250K single nucleotide polymorphism array karyotyping identifies acquired uniparental disomy and homozygous mutations, including novel missense substitutions of c-Cbl, in myeloid malignancies. *Cancer Res* **68**, 10349–10357.

[14] Barresi V, Palumbo GA, Musso N, Consoli C, Capizzi C, Meli CR, Romano A, Di Raimondo F, and Condorelli DF (2010). Clonal selection of 11q CN-LOH and *CBL* gene mutation in a serially studied patient during MDS progression to AML. *Leuk Res* **34**, 1539–1542.

[15] Rocquain J, Carbuccia N, Trouplin V, Raynaud S, Murati A, Nezri M, Tadrist Z, Olschwang S, Vey N, Birnbaum D, et al. (2010). Combined mutations of *ASXL1*, *CBL*, *FLT3*, *IDH1*, *IDH2*, *JAK2*, *KRAS*, *NPM1*, *NRAS*, *RUNX1*, *TET2* and *WT1* genes in myelodysplastic syndromes and acute myeloid leukemias. *BMC Cancer* **10**, 401.

[16] Reindl C, Quentmeier H, Petropoulos K, Greif PA, Benthaus T, Argiropoulos B, Mellert G, Vempati S, Duyster J, Buske C, et al. (2009). CBL exon 8/9 mutants activate the FLT3 pathway and cluster in core binding factor/11q deletion acute myeloid leukemia/myelodysplastic syndrome subtypes. *Clin Cancer Res* **15**, 2238–2247.

[17] Grand FH, Hidalgo-Curtis CE, Ernst T, Zoi K, Zoi C, McGuire C, Kreil S, Jones A, Score J, Metzgeroth G, et al. (2009). Frequent CBL mutations associated with 11q acquired uniparental disomy in myeloproliferative neoplasms. *Blood* **113**, 6182–6192.

[18] Kohlmann A, Grossmann V, Klein HU, Schindela S, Weiss T, Kazak B, Dicker F, Schnittger S, Dugas M, Kern W, et al. (2010). Next-generation sequencing technology reveals a characteristic pattern of molecular mutations in 72.8% of chronic myelomonocytic leukemia by detecting frequent alterations in *TET2*, *CBL*, *RAS*, and *RUNX1*. *J Clin Oncol* **28**, 3858–3865.

[19] Makishima H, Cazzolli H, Szpurka H, Dunbar A, Tiu R, Huh J, Muramatsu H, O'Keefe C, Hsi E, Paquette RL, et al. (2009). Mutations of E3 ubiquitin ligase Cbl family members constitute a novel common pathogenic lesion in myeloid malignancies. *J Clin Oncol* **27**, 6109–6116.

[20] Swerdlow SH, Campo E, Harris NL, Jaffe ES, Pileri SA, Stein H, Thiele J, and Vardiman JW (2008). *World Health Organization Classification of Tumours of Haematopoietic and Lymphoid Tissues*. IARC, Lyon, France.

[21] Greenberg P, Cox C, LeBeau MM, Fenaux P, Morel P, Sanz G, Sanz M, Vallespi T, Hamblin T, Oscier D, et al. (1997). International scoring system for evaluating prognosis in myelodysplastic syndromes. *Blood* **89**, 2079–2088.

[22] Shih LY, Huang CF, Wu JH, Lin TL, Dunn P, Wang PN, Kuo MC, Lai CL, and Hsu HC (2002). Internal tandem duplication of FLT3 in relapsed acute myeloid leukemia: a comparative analysis of bone marrow samples from 108 adult patients at diagnosis and relapse. *Blood* **100**, 2387–2392.

[23] Kuo MC, Liang DC, Huang CF, Shih YS, Wu JH, Lin TL, and Shih LY (2009). *RUNX1* mutations are frequent in chronic myelomonocytic leukemia and mutations at the C-terminal region might predict acute myeloid leukemia transformation. *Leukemia* **23**, 1426–1431.

[24] Xiao W and Oefner PJ (2001). Denaturing high-performance liquid chromatography: a review. *Hum Mutat* **17**, 439–474.

[25] Liang DC, Shih LY, Fu JF, Li HY, Wang HI, Hung IJ, Yang CP, Jaing TH, Chen SH, and Liu HC (2006). K-*ras* mutations and N-*ras* mutations in childhood acute leukemias with or without mixed-lineage leukemia gene rearrangements. *Cancer* **106**, 950–956.

[26] Shih LY, Huang CF, Wu JH, Wang PN, Lin TL, Dunn P, Chou MC, Kuo MC, and Tang CC (2004). Heterogeneous patterns of FLT3 ASP(835) mutations in relapsed *de novo* acute myeloid leukemia: a comparative analysis of 120 paired diagnostic and relapse bone marrow samples. *Clin Cancer Res* **10**, 1326–1332.

[27] Shih LY, Liang DC, Huang CF, Chang YT, Lai CL, Lin TH, Yang CP, Hung IJ, Liu HC, Jaing TH, et al. (2008). Cooperating mutations of receptor tyrosine kinases and *Ras* genes in childhood core-binding factor acute myeloid leukemia and a comparative analysis on paired diagnosis and relapse samples. *Leukemia* **22**, 303–307.

[28] Baxter EJ, Scott LM, Campbell PJ, East C, Fourouclas N, Swanton S, Vassiliou GS, Bench AJ, Boyd EM, Curtin N, et al. (2005). Acquired mutation of the tyrosine kinase JAK2 in human myeloproliferative disorders. *Lancet* **365**, 1054–1061.

[29] Tartaglia M, Niemeyer CM, Fragale A, Song X, Buechner J, Jung A, Hählen K, Hasle H, Licht JD, and Gelb BD (2003). Somatic mutations in *PTPN11* in juvenile myelomonocytic leukemia, myelodysplastic syndromes and acute myeloid leukemia. *Nat Genet* **34**, 148–150.

[30] Beer PA, Delhommeau F, LeCouédic JP, Dawson MA, Chen E, Bareford D, Kusec R, McMullin MF, Harrison CN, Vannucchi AM, et al. (2010). Two routes to leukemic transformation after a *JAK2* mutation-positive myeloproliferative neoplasm. *Blood* **115**, 2891–2900.

[31] Fernandes MS, Reddy MM, Croteau NJ, Walz C, Weisbach H, Podar K, Band H, Carroll M, Reiter A, Larson RA, et al. (2010). Novel oncogenic mutations of *CBL* in human acute myeloid leukemia that activate growth and survival pathways depend on increased metabolism. *J Biol Chem* **285**, 32596–32605.

Table W1. Primers for *C-CBL* Mutation Analysis by cDNA PCR Assay.

Primer Name	Primer Sequence 5′→3′	Amplicon Size (bp)	Method
CBL-ex6-F	CTCCAGACAATCCCTCACAATAAA	350	DHPLC
CBL-ex9-R	ACCACGATGGGTTCAGTACCTTTA		
CBL-ex8-9-F	ACTGTGAGATGGGCTCCACATT	374	DHPLC
CBL-ex8-9-R-gc	cggcgggggcGAAGCTTGTGGGGCCATG		
CBL-ex6-F	CTCCAGACAATCCCTCACAATAAA	533	Direct sequencing
CBL-ex8-9-R	GAAGCTTGTGGGGCCATG		

Table W2. Primers for *C-CBL* Mutation Analysis by gDNA PCR Assay.

Primer Name	Primer Sequence 5′→3′	Amplicon Size (bp)	Method
CBL-ex7-F-gc	cgcccgccgccgcccGGCAAATTGGCTTAAATAAAACC	187	DHPLC
CBL-ex7-R	GTGTCCAGTGATATGGTTATCATG		
CBL-ex9-F	CTATCTTTTGCTTCTTCTGCA	323	DHPLC
CBL-ex9-R-gc	gacgggcggcgggcggcggcgggcgTCGTTAAGTGTTTTACGGCTTTAG		
CBL-7F	CTTACACCACGTTGCCCTTT	364	Direct sequencing
CBL-7R	TGGGTCCTATTTTAAGCTCCA		
CBL-8F	AGGACCCAGACTAGATGCTTTC	387	Direct sequencing
CBL-8R	GGCCACCCCTTGTATCAGTA		
CBL-9F	CTGGCTTTTGGGGTTAGGTT	400	Direct sequencing
CBL-9R	TCGTTAAGTGTTTTACGGCTTT		

Genome-wide Analysis of Myelodysplastic Syndromes

Masashi Sanada* and Seishi Ogawa

Cancer Genomics Project, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan

Abstract: Myelodysplastic syndromes (MDS) are heterogeneous hematopoietic neoplasms characterized by ineffective hematopoiesis and a risk for progression to acute myeloid leukemia. A number of cytogenetic changes have been described that are characteristic to MDS and of clinical relevance; the specific gene targets of these alterations were largely unknown. On the other hand, over the past decade, technologies have been dramatically improved to enable high-throughput analysis of entire MDS genomes, leading to identification of frequent copy number neutral events and a number of novel gene targets implicated in the pathogenesis of MDS. In this review, we briefly overview the recent progress in the genetics of MDS, focusing on the newly identified gene targets in MDS.

Keywords: Microarray, SNP array, CNN-LOH, somatic mutation, high-throughput parallel sequencing.

INTRODUCTION

Myelodysplastic syndromes (MDS) are intractable clonal disorders of hematopoietic systems characterized by bone marrow dysplasia, peripheral blood cytopenia due to ineffective hematopoiesis, and a high propensity to acute myeloid leukemia (AML) [1, 2]. One of the prominent features of MDS is the high frequency of unbalanced chromosomal changes that accompany copy number alterations of chromosomal segments. Gains and losses of one or more chromosomal segments are found in approximately 50% of MDS patients in conventional cytogenetics and represent major determinants of the prognosis of MDS [3-5], indicating that these changes could be closely related to the pathogenesis of MDS. Unfortunately, however, most of the common changes typically involve large chromosomal segments, and with the lack of specific positional markers that pinpointed the critical genetic loci, the gene targets of these chromosomal lesions have not been determined until recently. This shows a stark contrast to *de novo* AML, where the breakpoints of disease type-specific translocations provided reliable positional markers to identify the major gene fusions that are relevant to molecular classification and characterization of AML [6,7]. The breakthrough for this situation has been brought about over the past decade, during which there have been dramatic improvements in genome technologies that allowed high-throughput/ resolution analysis of genomes [8], particularly with the development of single nucleotide polymorphism (SNP) array-based technology for copy number analysis.The SNP array-based copy number detection technologies enabled detection of copy-number (CN) alterations as well as allelic imbalances or loss of heterozygosity (LOH) in cancer genomes [9-13] and successfully applied to the analysis of MDS genomes, leading to the identification of a number of novel gene targets, frequently mutated in MDS as well as other myeloid cancers [14-18].Interestingly, many of the newly identified mutational targets are those involved in epigenetic regulation, such as DNA methylation and chromatin modifications, which is in accordance with the clinical observation that demethylating agents (azacitidine and decitabine) have been demonstrated to be effective in the treatment of high-risk MDS patients [19-21]. Thus, the frequent mutations of epigenesis-regulating genes support the possibility that the epigenetic alterations in MDS could be at least partly explained by the primary genetic alterations.

CYTOGENETICS IN MDS

Conventional cytogenetics provides an invaluable clue to the management of MDS, since the types and numbers of chromosomal lesions have been tightly linked to the prognosis of MDS cases. Cytogenetic findings are among the key parameters for the prediction of prognosis in the International Prognostic Scoring System (IPSS), and also adopted for the World Health Organization (WHO) classification-based Prognostic Scoring System (WPSS) [22]. Hasse *et al.* and other researchers also demonstrated that rare but recurrent cytogenetic alterations and specific karyotypic combinations could be used as beneficial markers for determining the prognosis of MDS [4, 23-25]. On the other hand, a potential caveat in conventional cytogenetics is that it absolutely depends on viable cells to obtain metaphases for analysis. Conventional cytogenetics fails to detect any abnormalities in approximately half of the patients with MDS. In fact, using interphase fluorescent in situ hybridization (FISH) analysis with 4 FISH probes, Rigolin *et al.* reported occult cytogenetic alterations in 17.8% of MDS patients with normal karyotype, including deletions of 5q31, 7q31 and 17p13, as well as trisomy8 [26]. Although providing a sensitive method for detecting submicroscopic alterations of known targets that are present in a small fraction of tumor samples without depending on cell divisions, interphase FISH analysis cannot be applied to genome-wide detection of genetic lesions.

ARRAY COMPARATIVE GENOMIC HYBRIDIZATION

Array-based comparative genomic hybridization (aCGH) enables comprehensive genome-wide analysis of genetic aberrations in cancers [8], in which differentially labeled DNAs from both tumor and normal reference samples are comparatively hybridized to a large number of probes on microarray. The ratio of the signal intensity of the test to that of the reference DNA is then calculated for the measurements of genomic copy numbers. The density of probes on microarray has been increased up to 4.2 million (NimbleGen), allowing for detection of smaller, more focal amplifications and deletions [27,28]. In the previous studies of MDS, a number of small, cryptic chromosomal abnormalities were identified using a CGH that could otherwise escape conventional cytogenetic analysis [29-32].

SNP ARRAY ANALYSIS

High density SNP arrays were originally developed for large-scale genotyping that is required for genome-wide association studies (GWAS) [33, 34]. However, the quantitative nature of the preparative whole-genome amplification and array hybridization thereafter allows for accurate estimation of genomic copy numbers at high resolution [35-37]. Furthermore, SNP array analysis also enables genome-wide LOH detection using genotyping data. With these desirable features, SNP arrays are now widely used for genome-wide copy number and LOH analyses in cancer research and the diagnosis of rare congenital disorders [10, 12-14,38,39]. Currently, two SNP array platforms are commercially available, AffymetrixGeneChip SNP Genotyping array [33] and Illumina beads array [40]. A number of software are developed for the analysis of

Address correspondence to this author at the Cancer Genomics Project, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan;
Tel: +81-3-3815-5411; Ext: 35615; Fax: +81-3-5800-9047;
E-mail: sanada-tky@umin.ac.jp

1873-4286/12 $58.00+.00

genomic copy numbers [35, 37, 41, 42], among which CNAG/AsCNAR software [36, 43], is one of the most widely used for this purpose. CNAG/AsCNAR is implements with a series of data compensation algorithms to accurately estimate copy numbers. In addition, by detecting subtle distortions in allele-specific signals caused by allelic imbalance, CNAG/AsCNAR enables sensitive detection of LOH with accurate determination of allele-specific copy numbers even in the face of up to 80% normal cell contamination [43].

Using AffymetrixGeneChip50k or 250k array, we analyzed a total of 222 MDS and myelodysplastic/myeloproliferative neoplasms (MDS/MPN) specimens, 87 of the 137 MDS cases (63.5%) had one or more regions showing allelic imbalances [14] Fig. (**1**). In accordance with previous cytogenetic studies, MDS genomes showed high frequencies of unbalanced genetic changes, including –5/5q–, –7/7q–, +8, 9p+, 12p–, 17p–, 18q+, 19p+, 19q+, 20q–, and 21q+, which were detected with higher sensitivity using SNP arrays. For example, hidden copy number alterations were successfully detected by SNP array-based copy number analysis in 14 out of 55 cases of normal karyotype MDS in our series [14]. However, the major advantage of SNP array analysis is the ability to detect genome-wide copy-number neutral (CNN)-LOH, which is undetectable by conventional cytogenetics, FISH or array CGH.

CNN-LOH IN MDS

CNN-LOH or uniparental disomy (UPD) is a common genetic alteration in cancer genome, majority of LOH in cancer being due to CNN-LOH rather than simple allelic deletion. Although CNN-LOH has been considered to be a common mechanism of inactivation of tumor suppressor genes, the discovery of a gain-of-function mutation of *JAK2* kinase associated with 9pUPD in myeloproliferative neoplasms (MPN)lead to a concept that CNN-LOH could also provide the genetic mechanism for clonal selection of a gain-of-function mutation [44]. CNN-LOH has been documented in 10-25% of MDS cases [14, 45, 46], 10-20% of *de novo* AML [47-52], and over 35%of chronic myelomonocytic leukemia (CMML) cases [14, 45].

Similar to other allelic imbalances, CNN-LOH was not randomly distributed throughout the MDS genomes, but tended to involve particular chromosomal arms in a relatively mutually exclusive manner, including 1p, 1q, 4q, 7q, 11p, 11q, 14q, 17p, and 21q Fig. (**1**). Among these, 7q, 17p, and 21q are also affected by deletions, while LOH in other arms were largely caused by UPD. In contrast, 5q and 20q are frequent targets of deletion in MDS cases, but rarely show CNN-LOH. CNN-LOH in 11p, 13q, 17p and 21q were also seen in *de novo* AML cases, whereas11q CNN-LOH was typically found in cases with MDS/MPN.A significant finding about these recurrent CNN-LOH is that they are frequently associated with homozygous mutations of known gene targets of myeloid neoplasms, including *c–MPL* or *N–RAS* in 1pCNN-LOH [14, 53], *JAK2* in 9pCNN-LOH [43, 44], *FLT3* in 13qCNN-LOH [54], *TP53* in 17pCNN-LOH [14], and *RUNX1* in 21qCNN-LOH [14, 54] (Table **1**). CNN-LOH could result in the duplication of mutated oncogenes after the loss of the normal allele or by inducing deletion of tumor suppressor genes.

MUTATED GENE TARGETS IN MDS (FIG. 2)

1) TET2

The long arm of chromosome 4 has not been reported as a common target of chromosomal abnormalities in myeloid malignancies in conventional cytogenetics [4], but recently turned out to be a recurrent target of CNN-LOH in MDS and CMML in SNP array analysis. Delhommeau *et al.* and Langeimer *et al.* identified

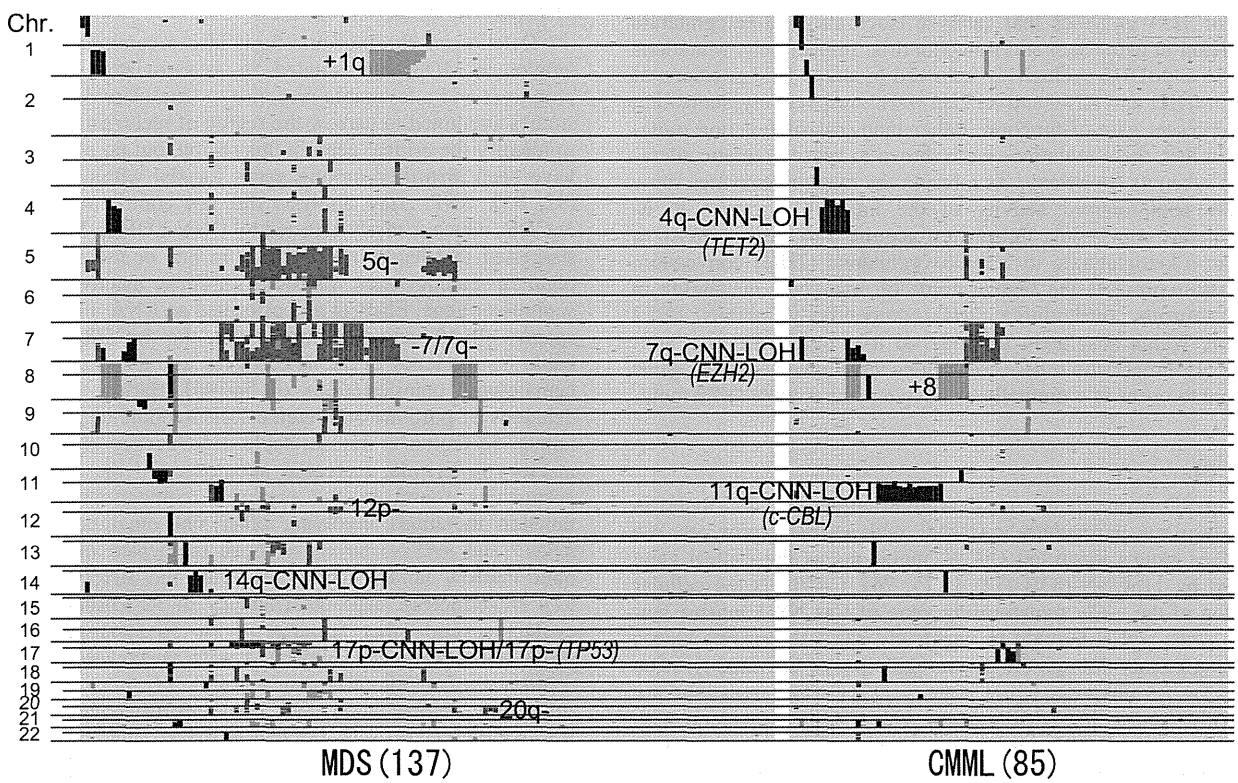

Fig. (1). The genome profile of 222 cases of MDS and related myeloid neoplasms detected by SNP array analysis.
The genetic alterations, including CN gains, losses and CNN-LOH, are color-coded, light gray, gray, and dark gray, respectively. These lesions are plotted vertically in chromosomal order for each sample. Vertical positions of each lesion are proportional to the genetic length and thus the size of the color-coded corresponds to the length of alterations. CNN-LOH, in particular chromosomal arms tends to be found in mutually exclusive cases, enabling clustering based on the site of CNN-LOH, except for 17pLOH, which was frequently accompanied by loss of 5q, loss of chromosome 7 or 7q, and loss of 12p.Common genetic alterations and their target genes are indicated.

loss of function mutations of *TET2* as the target of 4qLOH[15,16], and also mutated frequently in other cases without having 4qLOH. In fact, *TET2*is now shown to represent one of the most frequently mutated genes in MDS (~20%) as well as other myeloid neoplasms [55], including MPN with or without *JAK2*-V617F mutations (~10%), CMML (30-50%), and part of AML(13%) [15, 16, 56, 57]. *TET2* mutations frequently occur during progression of MPN or MDS to secondary AML. The impact of *TET2* mutations on clinical outcomes is still controversial. Some reports demonstrated significantly shorter overall survival in patients with *TET2* mutations [56-58], while others reported favorable or no prognostic impact of *TET2* mutations [16, 55, 59].

TET family proteins (Tet1, Tet2, and Tet3) catalyze the conversion of 5-methyl-cytosine to 5-hydroxymethyl-cytosine (5hmC) [60, 61]. In ES cells, TET1 plays a functional role in maintaining the pluripotent state [61-63]. A recent study demonstrated that 5-hmC generated by TET activity is an intermediate during the process of DNA demethylation [64]. In addition, TET1 directly interacts with Sin3A, a co-repressor protein essential for inhibiting the transcription of a subset of genes [65]. *Tet2* deficiency in mice lead to the progressive enlargement of the hematopoietic stem and progenitor compartment, and also results in abnormalities in mature myeloid and lymphoid cells, and leading to fatal hematopoietic malignancies[66]. Quivoron *et al.* also found that *TET2* mutations were not only seen in myeloid neoplasms but also in various types of B- and T-cell lymphoid tumors in humans.

2) IDH1/IDH2

Mutations of isocitrate dehydrogenase (IDH) 1 and IDH2 are initially identified through comprehensive mutation studies in glioblastoma as well as *de novo* AML in high frequencies [67, 68], and also reported in other myeloid malignancies including secondary AML, MDS and MPN [69-73]. IDH1 and IDH2 are components of TCA enzymes that catalyze isocitrate to α-ketoglutarate conversion in cytoplasm and mitochondria, respectively. Mutations of IDH1 and IDH2 exclusively involved in amino acid positions of R132 in IDH1 and R140 and R172in IDH2, respectively, indicating they represent gain-of-function, rather than loss of function mutations. In fact, these mutations were shown to cause dramatic alteration of substrate specificity. As a result, the mutated enzymes show severely compromised activity of the intrinsic isocitrate to α-ketoglutarate conversion, but in turn acquire a *de novo* activity to catalyze α-ketoglutarate to 2 hydroxyglutarate (2HG) conversion. The 2HG represents the first example of oncogenic metabolite in human cancers. Intriguingly, 2HG competitively inhibits TET2 function, which absolutely depends on α-ketoglutarate as a substrate [74]. In fact, the *IDH1/2* mutations were always heterozygous and tend to occur in a mutually exclusive manner with *TET2* mutations.

3) C-CBL

11qUPD is one of the most common targets of UPD found in myelodysplasia, particularly in CMML with normal karyotypes. We and other groups identified *C-CBL* mutations as the critical gene affected by 11qCNN-LOH [14, 45,75, 76]. *C-CBL* is the cellular homolog of the *v-Cbl* transforming gene of Cas NS-1 murine leukemia virus, and is thought to negatively regulate tyrosine kinase signaling, mainly through the down-regulation of activated tyrosine kinases by E3 ubiquitin ligase activity [77].*C-CBL* mutations are frequently seen in MDS/MPN cases with a tight association with 11q-CNN-LOH. *C-CBL* mutations and other *RAS* pathway mutations (*NRAS*, *KRAS*, *PTPN11,*and *NF1*)occur in a mutually exclusive manner in CMML and juvenile myelomonocytic leukemia (JMML) [76, 78, 79].Interestingly in this regard, similar to other mutations of RAS pathway genes, heterozygous germ-line *C-CBL* mutations may predispose the development of JMML with a Noonan Syndrome-like phenotype [80, 81]. Most *C-CBL* mutations in myeloid malignancies are found in the linker and RING finger domains, which are central to the E3 ubiquitin ligase activity[82].C-CBL mutants show compromised E3 ubiquitin ligase activity, and also inhibit wild type C-CBL and CBLB, leading to prolonged activation of tyrosine kinases following cytokine stimulation [14, 83, 84], leading to hypersensitivity to a wide spectrum of cytokines that underlies the pathogenesis of the myeloproliferative phenotype commonly found in CMML and JMML [82, 84].

4) EZH2

Loss of chromosomes 7 and 7q are one of the most frequent genetic alterations in MDS and known as a reliable predictor of adverse prognosis. Approximately 10% of the patients with MDS carry an abnormality of chromosome 7, either alone or as part of a complex karyotype. This frequency is higher in therapy-related MDS associated with a prior history of treatment with alkylating agents. SNP array analysis has revealed that not only copy number loss but also CNN-LOH is the cause of 7qLOH in MDS and related myeloid neoplasms. Recently, Ernst *et al.* and Nikoloski *et al.* have shown that *EZH2*is mutated in some cases with7q-LOH [17,18], indicating that *EZH2*is one of the gene targets in 7qLOH. *EZH2* encodes a histone methyltransferase that is the catalytic component of the polycomb repressive complex-2 (PRC2), a highly conserved histone H3 at lysine-27 methyl transferase, which functions to initiate epigenetic silencing of genes involved in cell fate decisions [85]. Loss of PRC2 function increases hematopoietic stem cell activity and expansion, which may explain how loss of function mutations of *EZH2* leads myeloid neoplasms [86]. On the other hand, at least three common deleted regions (CDRs) on 7q (7q22, 7q32-33, and 7q35-36) have been identified in myeloid malignanicies [87-89], and therefore, *EZH2*(7q36)does not seem to be the sole target for the deletions of chromosome 7q.

5) Ribosomal Protein

Deletion of chromosome 5q is also a common cytogenetic alteration in MDS, and isolated 5q- is associated with a favorable prognosis and a favorable response to lenalidomide [90, 91]. Many studies attempted to narrow the region of recurrent somatic deletion to identify the critical gene in this region, but no somatic mutations have been identified among genes located within the CDR of 5q [92, 93]. SNP array analysis did not contribute to narrow the 5qCDR, which is rarely affected by CNN-LOH in MDS. It has been suggested that haplo-insufficiency in one or more genes may explain 5q- pathogenesis, instead of bi-allelic inactivation of a tumor suppressor gene. Ebert *et al.* performed an RNA interference screen against all 40 genes located within the 5qCDR and implicated haplo-insufficiency of the *RPS14* gene as a major contributor to the hematologic manifestations of 5q-[94]. Barlow *et al.* generated deletions of portions of syntenic lesion(containing*RPS14*) with the human 5q region in mouse, haplo-insufficiency of this loci caused macrocytic anemia, increased apoptosis and the morphologic abnormalities found in the erythroid compartment [95].Loss-of-function mutations involving other ribosomal components (e.g., *RPS19* and *RPS24*) have also been implicated in rare congenital bone marrow failure syndromes, Diamond–Blackfan anemia [96, 97]. Nevertheless, haploinsufficiency of RPS14does not seem to explain several other features of the 5q–syndrome, which also shows thrombocytosis associated with megakaryocytic dysplasia, neutropenia, and clonal dominance [98, 99]. Interestingly, a recent study has demonstrated that haplo-insufficiency of two micro RNAs within CDR, *miR-145* and *miR-146*, could also contribute to the pathogenesis of 5q- syndrome, supporting a model of haploinsufficiency of multiple gene targets in this syndrome [100].

CLINICAL APPLICATION

Given that cytogenetic information provides a valuable clue to the management of MDS as prognostic makers, a more accurate prognosis could be established based on SNP array or other CGH

Table 1. Recurrent Gene Mutations in Myeloid Malignancies

Mutated Gene	Diseases	frequency in MDS	frequency in de novo AML	Associated chromosomal alterations	pathway
TET2	MDS, CMML, MPN	20.0%	13.2%	4qUPD	epigenetic modification
EZH2	MDS, CMML	6.0%	rare	7qUPD	epigenetic modification
ASXL1	AML,MDS,CMML	10-15%	10.8%		epigenetic modification
DNMT3A	AML,MDS	8.0%	22.1%		epigenetic modification
IDH1	AML,MDS	rare-5.2%	6.6-8.5%	normal cytogenetics	epigenetic modification
IDH2	AML,MDS,CMML	4.2%	11-15.4%		epigenetic modification
TP53	AML, MDS	5-10%	<10%	17ploss/UPD, complex karyotype	cell cycle, apoptosis
Nras	MDS, AML, MDS/MPN	3.6-6.3%	10-15%	1pUPD	signal transduction
Kras	MDS, AML	rare	5.0%		signal transduction
cMPL	MPN, RARSt	rare-5%	rare	1pUPD	signal transduction
JAK2	MPN, RARSt	rare-50%	rare	9pUPD	signal transduction
c-CBL	CMML, JMML	rare	rare	11qUPD	signal transduction
FLT3	AML	rare	28-33%(ITD), 5-10%	13qUPD	signal transduction
NF1	JMML	rare	rare	17qUPD	signal transduction
PTPN11	JMML	rare	rare		signal transduction
c-KIT	AML	rare	6-10%		signal transduction
RUNX1	AML, MDS	15-20%	8.6%	21qloss/UPD	transcriptional factor
WT1	AML	rare	10.0%	11pUPD	transcriptional factor
CEBPA	AML	rare	4-9%	19pUPD	transcriptional factor
U2AF35	MDS	11.6%	rare		RNA splicing
SRSF2	MDS, CMML	11.6%	rare		RNA splicing
SF3B1	RARS, MDS	6.5-75.3%	rare		RNA splicing
ZRSR2	MDS	7.7%	rare		RNA splicing
NPM1	AML	rare	25-35%	normal cytogenetics	other

rare, mutations present in <3% of patients
MDS, myelodysplastic syndrome; RARS, refractory anemia with ringed sideroblats; RARSt,RARS and thrombocytosis
MPN, myeloproliferative neoplasm; AML, acute myeloid leukemia; CMML, chronic myelomonocytic leukemia; JMML, juvenile myelomonocytic leukemia

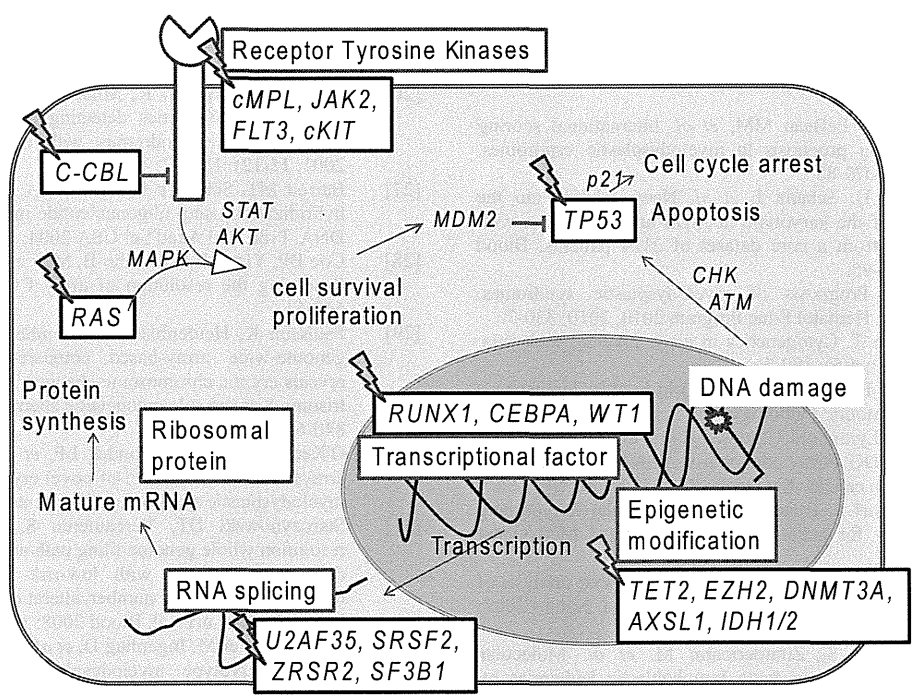

Fig. (2). Molecular pathways of genes affected in MDS.
Mutations of multiple pathways have been indicated in the pathogenesis of MDS. The mutated components areindicated by arrows.

based genomic analysis of MDS. Array-based genome-wide copy number analysis can provide much information on genetic alterations, especially on CNN-LOH, although array-based analysis cannot detect the balanced translocations that are relevant to the management of a large number of hematopoietic malignancies.

Some studies showed that the presence of newly detected alterations by microarray were useful as novel predictors of prognosis [101]. Heinrichs *et al.* and Godek *et al.* showed that 7q-CNN-LOH is a possible marker for poor prognosis [45, 46], although the evidence for the value of each alteration identified with SNP array or aCGH has so far been still incomplete. Clearly, further studies are required to establish the clinical values of array-based karyotyping technologies in MDS. Recently, Bejar *et al.* examined whether the mutation profile of known target genes was associated with the clinical phenotype, and found that mutations in *TP53*, *EZH2*, *ETV6*, *RUNX1* and *ASXL1* are independent predictors of poor prognosis [55]. However, most reported mutations occur infrequently in MDS cases and are also found in the case of AML and other myeloid

neoplasms (Table **1**, Fig. (**2**)). These mutations may explain the limited aspect of pathogenesis of MDS.

CONCLUSION AND RECENT PROGRESS

One of the best targets of SNP-array based genome-wide allele-karyotyping would be MDS and related disorders in which CNN-LOH and unbalanced genetic changes are predominant. Using SNP array, several novel gene mutations, *C-CBL*, *TET2,* and *EZH2*, have been identified in MDS and related myeloid neoplasms. However, as many as 20-30% of primary MDS cases do not show any genetic changes even with SNP array karyotyping or mutation analysis of previously known targets. More problematic is that no gene mutations are specific to MDS but also found in other myeloid cancers, indicating that we still have incomplete knowledge about the molecular pathogenesis of MDS. In this regard, the development of high-throughput parallel sequencing technologies has provided an opportunity to characterize genetic changes across the genome-wide sequences at single nucleotide level [102], and is expected to be successfully applied to the genetic analysis of MDS to reveal more aspects of their pathogenesis in near future. In fact, our recent study using whole exome sequencing has revealed high frequencies (45~85% depending on subtypes of MDS) of pathway mutations involving multiple components of the splicing machinery that are highly specific to myeloid neoplasms showing features of myelodysplasia [103], although more studies are required to elucidate their roles in the pathogenesis of MDS.

REFERENCES

[1] Tefferi A, Vardiman JW. Myelodysplastic syndromes. N Engl J Med 2009; 361(19): 1872-85.

[2] Bejar R, Levine R, Ebert BL. Unraveling the molecular pathophysiology of myelodysplastic syndromes. J Clin Oncol 2011 Feb 10; 29(5): 504-15.

[3] Greenberg P, Cox C, LeBeau MM, *et al.* International scoring system for evaluating prognosis in myelodysplastic syndromes. Blood 1997; 89(6): 2079-88.

[4] Haase D, Germing U, Schanz J, *et al.* New insights into the prognostic impact of the karyotype in MDS and correlation with subtypes: evidence from a core dataset of 2124 patients. Blood 2007; 110(13): 4385-95.

[5] Garcia-Manero G. Prognosis of myelodysplastic syndromes. Hematology Am Soc Hematol Educ Program 2010; 2010: 330-7.

[6] Schoch C, Haferlach T. Cytogenetics in acute myeloid leukemia. Curr Oncol Rep 2002; 4(5): 390-7.

[7] Swerdlow S, Campo E, Harris N, *et al.* World Health Organization Classification of Tumours: Pathology and Genetics of Tumours of Haematopoietic and Lymphoid Tissues. Lyon: IARC; 2008.

[8] Pinkel D, Albertson DG. Array comparative genomic hybridization and its applications in cancer. Nat Genet 2005; 37 Suppl: S11-7.

[9] Rauch A, Ruschendorf F, Huang J, *et al.* Molecular karyotyping using an SNP array for genomewide genotyping. J Med Genet 2004; 41(12): 916-22.

[10] Mullighan CG, Goorha S, Radtke I, *et al.* Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. Nature 2007; 446(7137): 758-64.

[11] Kawamata N, Ogawa S, Zimmermann M, *et al.* Molecular allelokaryotyping of pediatric acute lymphoblastic leukemias by high-resolution single nucleotide polymorphism oligonucleotide genomic microarray. Blood 2008; 111(2): 776-84.

[12] Chen Y, Takita J, Choi YL, *et al.* Oncogenic mutations of ALK kinase in neuroblastoma. Nature 2008; 455(7215): 971-4.

[13] Kato M, Sanada M, Kato I, *et al.* Frequent inactivation of A20 in B-cell lymphomas. Nature 2009; 459(7247): 712-6.

[14] Sanada M, Suzuki T, Shih LY, *et al.* Gain-of-function of mutated C-CBL tumour suppressor in myeloid neoplasms. Nature 2009; 460(7257): 904-8.

[15] Delhommeau F, Dupont S, Della Valle V, *et al.* Mutation in TET2 in myeloid cancers. N Engl J Med 2009; 360(22): 2289-301.

[16] Langemeijer SM, Kuiper RP, Berends M, *et al.* Acquired mutations in TET2 are common in myelodysplastic syndromes. Nat Genet 2009; 41(7): 838-42.

[17] Ernst T, Chase AJ, Score J, *et al.* Inactivating mutations of the histone methyltransferase gene EZH2 in myeloid disorders. Nat Genet 2010; 42(8): 722-6.

[18] Nikoloski G, Langemeijer SM, Kuiper RP, *et al.* Somatic mutations of the histone methyltransferase gene EZH2 in myelodysplastic syndromes. Nat Genet 2010; 42(8): 665-7.

[19] Silverman LR, Demakos EP, Peterson BL, *et al.* Randomized controlled trial of azacitidine in patients with the myelodysplastic syndrome: a study of the cancer and leukemia group B. J Clin Oncol 2002; 20(10): 2429-40.

[20] Silverman LR, McKenzie DR, Peterson BL, *et al.* Further analysis of trials with azacitidine in patients with myelodysplastic syndrome: studies 8421, 8921, and 9221 by the Cancer and Leukemia Group B. J Clin Oncol 2006; 24(24): 3895-903.

[21] Kantarjian H, Oki Y, Garcia-Manero G, *et al.* Results of a randomized study of 3 schedules of low-dose decitabine in higher-risk myelodysplastic syndrome and chronic myelomonocytic leukemia. Blood 2007; 109(1): 52-7.

[22] Malcovati L, Germing U, Kuendgen A, *et al.* Time-dependent prognostic scoring system for predicting survival and leukemic evolution in myelodysplastic syndromes. J Clin Oncol 2007; 25(23): 3503-10.

[23] Sole F, Espinet B, Sanz GF, *et al.* Incidence, characterization and prognostic significance of chromosomal abnormalities in 640 patients with primary myelodysplastic syndromes. Grupo Cooperativo Espanol de Citogenetica Hematologica. Br J Haematol 2000; 108(2): 346-56.

[24] Sole F, Luno E, Sanzo C, *et al.* Identification of novel cytogenetic markers with prognostic significance in a series of 968 patients with primary myelodysplastic syndromes. Haematologica 2005; 90(9): 1168-78.

[25] Bernasconi P, Klersy C, Boni M, *et al.* World Health Organization classification in combination with cytogenetic markers improves the prognostic stratification of patients with de novo primary myelodysplastic syndromes. Br J Haematol 2007; 137(3): 193-205.

[26] Rigolin GM, Bigoni R, Milani R, *et al.* Clinical importance of interphase cytogenetics detecting occult chromosome lesions in myelodysplastic syndromes with normal karyotype. Leukemia 2001; 15(12): 1841-7.

[27] Barrett MT, Scheffer A, Ben-Dor A, *et al.* Comparative genomic hybridization using oligonucleotide microarrays and total genomic DNA. Proc Natl Acad Sci USA 2004; 101(51): 17765-70.

[28] Coe BP, Ylstra B, Carvalho B, Meijer GA, Macaulay C, Lam WL. Resolving the resolution of array CGH. Genomics 2007; 89(5): 647-53.

[29] Paulsson K, Heidenblad M, Strombeck B, *et al.* High-resolution genome-wide array-based comparative genome hybridization reveals cryptic chromosome changes in AML and MDS cases with trisomy 8 as the sole cytogenetic aberration. Leukemia 2006; 20(5): 840-6.

[30] O'Keefe CL, Tiu R, Gondek LP, *et al.* High-resolution genomic arrays facilitate detection of novel cryptic chromosomal lesions in myelodysplastic syndromes. Exp Hematol 2007; 35(2): 240-51.

[31] Starczynowski DT, Vercauteren S, Telenius A, *et al.* High-resolution whole genome tiling path array CGH analysis of CD34+ cells from patients with low-risk myelodysplastic syndromes reveals cryptic copy number alterations and predicts overall and leukemia-free survival. Blood 2008; 112(8): 3412-24.

[32] Thiel A, Beier M, Ingenhag D, *et al.* Comprehensive array CGH of normal karyotype myelodysplastic syndromes reveals hidden recurrent and individual genomic copy number alterations with prognostic relevance. Leukemia 2011; 25(3): 387-99.

[33] Matsuzaki H, Dong S, Loi H, *et al.* Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. Nat Methods 2004; 1(2): 109-11.

[34] Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007; 447(7145): 661-78.

[35] Zhao X, Li C, Paez JG, *et al.* An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. Cancer Res 2004; 64(9): 3060-71.

[36] Nannya Y, Sanada M, Nakazaki K, *et al.* A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. Cancer Res 2005; 65(14): 6071-9.

[37] Van Loo P, Nordgard SH, Lingjaerde OC, *et al.* Allele-specific copy number analysis of tumors. Proc Natl Acad Sci USA 2010; 107(39): 16910-5.

[38] Thoennissen NH, Krug UO, Lee DH, *et al.* Prevalence and prognostic impact of allelic imbalances associated with leukemic transformation of Philadelphia chromosome-negative myeloproliferative neoplasms. Blood 2010; 115(14): 2882-90.

[39] Greenway SC, Pereira AC, Lin JC, *et al.* De novo copy number variants identify new genes and loci in isolated sporadic tetralogy of Fallot. Nat Genet 2009; 41(8): 931-5.

[40] Murray SS, Oliphant A, Shen R, *et al.* A highly informative SNP linkage panel for human genetic studies. Nat Methods 2004; 1(2): 113-7.

[41] Beroukhim R, Lin M, Park Y, *et al.* Inferring loss-of-heterozygosity from unpaired tumors using high-density oligonucleotide SNP arrays. PLoS Comput Biol 2006; 2(5): e41.

[42] Laframboise T, Harrington D, Weir BA. PLASQ: a generalized linear model-based procedure to determine allelic dosage in cancer cells from SNP array data. Biostatistics 2007; 8(2): 323-36.

[43] Yamamoto G, Nannya Y, Kato M, *et al.* Highly sensitive method for genomewide detection of allelic composition in nonpaired, primary tumor specimens by use of affymetrix single-nucleotide-polymorphism genotyping microarrays. Am J Hum Genet 2007; 81(1): 114-26.

[44] Kralovics R, Passamonti F, Buser AS, *et al.* A gain-of-function mutation of JAK2 in myeloproliferative disorders. N Engl J Med 2005; 352(17): 1779-90.

[45] Gondek LP, Tiu R, O'Keefe CL, Sekeres MA, Theil KS, Maciejewski JP. Chromosomal lesions and uniparental disomy detected by SNP arrays in MDS, MDS/MPD, and MDS-derived AML. Blood 2008; 111(3): 1534-42.

[46] Heinrichs S, Kulkarni RV, Bueso-Ramos CE, *et al.* Accurate detection of uniparental disomy and microdeletions by SNP array analysis in myelodysplastic syndromes with normal cytogenetics. Leukemia 2009; 23(9): 1605-13.

[47] Raghavan M, Smith LL, Lillington DM, *et al.* Segmental uniparental disomy is a commonly acquired genetic event in relapsed acute myeloid leukemia. Blood 2008; 112(3): 814-21.

[48] Akagi T, Shih LY, Kato M, *et al.* Hidden abnormalities and novel classification of t(15; 17) acute promyelocytic leukemia (APL) based on genomic alterations. Blood 2009; 113(8): 1741-8.

[49] Akagi T, Shih LY, Ogawa S, *et al.* Single nucleotide polymorphism genomic arrays analysis of t(8; 21) acute myeloid leukemia cells. Haematologica 2009; 94(9): 1301-6.

[50] Walter MJ, Payton JE, Ries RE, *et al.* Acquired copy number alterations in adult acute myeloid leukemia genomes. Proc Natl Acad Sci USA 2009; 106(31): 12950-5.

[51] Tiu RV, Gondek LP, O'Keefe CL, *et al.* New lesions detected by single nucleotide polymorphism array-based chromosomal analysis have important clinical impact in acute myeloid leukemia. J Clin Oncol 2009; 27(31): 5219-26.

[52] Bullinger L, Kronke J, Schon C, *et al.* Identification of acquired copy number alterations and uniparental disomies in cytogenetically normal acute myeloid leukemia using high-resolution single-nucleotide polymorphism analysis. Leukemia 2010; 24(2): 438-49.

[53] Dunbar AJ, Gondek LP, O'Keefe CL, *et al.* 250K single nucleotide polymorphism array karyotyping identifies acquired uniparental disomy and homozygous mutations, including novel missense substitutions of c-Cbl, in myeloid malignancies. Cancer Res 2008; 68(24): 10349-57.

[54] Fitzgibbon J, Smith LL, Raghavan M, *et al.* Association between acquired uniparental disomy and homozygous gene mutation in acute myeloid leukemias. Cancer Res 2005; 65(20): 9152-4.

[55] Bejar R, Stevenson K, Abdel-Wahab O, *et al.* Clinical effect of point mutations in myelodysplastic syndromes. N Engl J Med 2011; 364(26): 2496-506.

[56] Kosmider O, Gelsi-Boyer V, Ciudad M, *et al.* TET2 gene mutation is a frequent and adverse event in chronic myelomonocytic leukemia. Haematologica 2009; 94(12): 1676-81.

[57] Abdel-Wahab O, Mullally A, Hedvat C, *et al.* Genetic characterization of TET1, TET2, and TET3 alterations in myeloid malignancies. Blood 2009; 114(1): 144-7.

[58] Metzeler KH, Maharry K, Radmacher MD, *et al.* TET2 mutations improve the new European LeukemiaNet risk classification of acute myeloid leukemia: a Cancer and Leukemia Group B study. J Clin Oncol 2011; 29(10): 1373-81.

[59] Smith AE, Mohamedali AM, Kulasekararaj A, *et al.* Next-generation sequencing of the TET2 gene in 355 MDS and CMML patients reveals low-abundance mutant clones with early origins, but indicates no definite prognostic value. Blood 2010; 116(19): 3923-32.

[60] Tahiliani M, Koh KP, Shen Y, *et al.* Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. Science 2009; 324(5929): 930-5.

[61] Ito S, D'Alessio AC, Taranova OV, *et al.* Hong K, Sowers LC, Zhang Y. Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. Nature 2010; 466(7310): 1129-33.

[62] Wu H, D'Alessio AC, Ito S, *et al.* Dual functions of Tet1 in transcriptional regulation in mouse embryonic stem cells. Nature 2011; 473(7347): 389-93.

[63] Pastor WA, Pape UJ, Huang Y, *et al.* Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells. Nature 2011; 473(7347): 394-7.

[64] Guo JU, Su Y, Zhong C, *et al.* Hydroxylation of 5-methylcytosine by TET1 promotes active DNA demethylation in the adult brain. Cell 2011; 145(3): 423-34.

[65] Williams K, Christensen J, Pedersen MT, *et al.* TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. Nature 2011; 473(7347): 343-8.

[66] Quivoron C, Couronne L, Della Valle V, *et al.* TET2 inactivation results in pleiotropic hematopoietic abnormalities in mouse and is a recurrent event during human lymphomagenesis. Cancer Cell 2011; 20(1): 25-38.

[67] Parsons DW, Jones S, Zhang X, *et al.* An integrated genomic analysis of human glioblastoma multiforme. Science 2008; 321(5897): 1807-12.

[68] Mardis ER, Ding L, Dooling DJ, *et al.* Recurring mutations found by sequencing an acute myeloid leukemia genome. N Engl J Med 2009; 361(11): 1058-66.

[69] Marcucci G, Maharry K, Wu YZ, *et al.* IDH1 and IDH2 gene mutations identify novel molecular subsets within de novo cytogenetically normal acute myeloid leukemia: a Cancer and Leukemia Group B study. J Clin Oncol 2010; 28(14): 2348-55.

[70] Paschka P, Schlenk RF, Gaidzik VI, *et al.* IDH1 and IDH2 mutations are frequent genetic alterations in acute myeloid leukemia and confer adverse prognosis in cytogenetically normal acute myeloid leukemia with NPM1 mutation without FLT3 internal tandem duplication. J Clin Oncol 2010; 28(22): 3636-43.

[71] Yoshida K, Sanada M, Kato M, *et al.* A nonsense mutation of IDH1 in myelodysplastic syndromes and related disorders. Leukemia 2011; 25(1): 184-6.

[72] Tefferi A, Lasho TL, Abdel-Wahab O, *et al.* IDH1 and IDH2 mutation studies in 1473 patients with chronic-, fibrotic- or blast-phase essential thrombocythemia, polycythemia vera or myelofibrosis. Leukemia 2010; 24(7): 1302-9.

[73] Pardanani A, Lasho TL, Finke CM, *et al.* IDH1 and IDH2 mutation analysis in chronic- and blast-phase myeloproliferative neoplasms. Leukemia 2010; 24(6): 1146-51.

[74] Figueroa ME, Abdel-Wahab O, Lu C, *et al.* Leukemic IDH1 and IDH2 mutations result in a hypermethylation phenotype, disrupt TET2 function, and impair hematopoietic differentiation. Cancer Cell 2010; 18(6): 553-67.

[75] Grand FH, Hidalgo-Curtis CE, Ernst T, *et al.* Frequent CBL mutations associated with 11q acquired uniparental disomy in myeloproliferative neoplasms. Blood 2009; 113(24): 6182-92.

[76] Loh ML, Sakai DS, Flotho C, *et al.* Mutations in CBL occur frequently in juvenile myelomonocytic leukemia. Blood 2009; 114(9): 1859-63.

[77] Thien CB, Langdon WY. Cbl: many adaptations to regulate protein tyrosine kinases. Nat Rev Mol Cell Biol 2001; 2(4): 294-307.

[78] Perez B, Kosmider O, Cassinat B, *et al.* Genetic typing of CBL, ASXL1, RUNX1, TET2 and JAK2 in juvenile myelomonocytic leukaemia reveals a genetic profile distinct from chronic myelomonocytic leukaemia. Br J Haematol 2010; 151(5): 460-8.

[79] Kohlmann A, Grossmann V, Klein HU, *et al.* Next-generation sequencing technology reveals a characteristic pattern of molecular mutations in 72.8% of chronic myelomonocytic leukemia by detecting frequent alterations in TET2, CBL, RAS, and RUNX1. J Clin Oncol 2010 ; 28(24): 3858-65.

[80] Niemeyer CM, Kang MW, Shin DH, *et al.* Germline CBL mutations cause developmental abnormalities and predispose to juvenile myelomonocytic leukemia. Nat Genet 2010; 42(9): 794-800.

[81] Martinelli S, De Luca A, Stellacci E, *et al.* Heterozygous germline mutations in the CBL tumor-suppressor gene cause a Noonan syndrome-like phenotype. Am J Hum Genet 2010; 87(2): 250-7.

[82] Ogawa S, Sanada M, Shih LY, *et al.* Gain-of-function c-CBL mutations associated with uniparental disomy of 11q in myeloid neoplasms. Cell Cycle 2010; 9(6): 1051-6.

[83] Sargin B, Choudhary C, Crosetto N, *et al.* Flt3-dependent transformation by inactivating c-Cbl mutations in AML. Blood 2007; 110(3): 1004-12.

[84] Rathinam C, Thien CB, Flavell RA, Langdon WY. Myeloid leukemia development in c-Cbl RING finger mutant mice is dependent on FLT3 signaling. Cancer Cell 2010; 18(4): 341-52.

[85] Valk-Lingbeek ME, Bruggeman SW, van Lohuizen M. Stem cells and cancer; the polycomb connection. Cell 2004; 118(4): 409-18.

[86] Majewski IJ, Ritchie ME, Phipson B, *et al.* Opposing roles of polycomb repressive complexes in hematopoietic stem and progenitor cells. Blood 2010; 116(5): 731-9.

[87] Le Beau MM, Espinosa R, 3rd, Davis EM, *et al.* Eisenbart JD, Larson RA, Green ED. Cytogenetic and molecular delineation of a region of chromosome 7 commonly deleted in malignant myeloid diseases. Blood 1996; 88(6): 1930-5.

[88] Dohner K, Brown J, Hehmann U, *et al.* Molecular cytogenetic characterization of a critical region in bands 7q35-q36 commonly deleted in malignant myeloid disorders. Blood 1998; 92(11): 4031-5.

[89] Wong JC, Zhang Y, Lieuw KH, *et al.* Use of chromosome engineering to model a segmental deletion of chromosome band 7q22 found in myeloid malignancies. Blood 2010; 115(22): 4524-32.

[90] List A, Dewald G, Bennett J, *et al.* Lenalidomide in the myelodysplastic syndrome with chromosome 5q deletion. N Engl J Med 2006; 355(14): 1456-65.

[91] Nimer SD. Clinical management of myelodysplastic syndromes with interstitial deletion of chromosome 5q. J Clin Oncol 2006; 24(16): 2576-82.

[92] Le Beau MM, Espinosa R, 3rd, Neuman WL, *et al.* Cytogenetic and molecular delineation of the smallest commonly deleted region of chromosome 5 in malignant myeloid diseases. Proc Natl Acad Sci USA 1993; 90(12): 5484-8.

[93] Boultwood J, Fidler C, Strickson AJ, *et al.* Narrowing and genomic annotation of the commonly deleted region of the 5q- syndrome. Blood 2002; 99(12): 4638-41.

[94] Ebert BL, Pretz J, Bosco J, *et al.* Identification of RPS14 as a 5q-syndrome gene by RNA interference screen. Nature 2008; 451(7176): 335-9.

[95] Barlow JL, Drynan LF, Hewett DR, *et al.* A p53-dependent mechanism underlies macrocytic anemia in a mouse model of human 5q- syndrome. Nat Med 2010; 16(1): 59-66.

[96] Draptchinskaia N, Gustavsson P, Andersson B, *et al.* The gene encoding ribosomal protein S19 is mutated in Diamond-Blackfan anaemia. Nat Genet 1999; 21(2): 169-75.

[97] Boria I, Garelli E, Gazda HT, *et al.* The ribosomal basis of Diamond-Blackfan Anemia: mutation and database update. Hum Mutat 2010; 31(12): 1269-79.

[98] Van den Berghe H, Cassiman JJ, David G, *et al.* Fryns JP, Michaux JL, Sokal G. Distinct haematological disorder with deletion of long arm of no. 5 chromosome. Nature 1974; 251(5474): 437-8.

[99] Tinegate H, Gaunt L, Hamilton PJ. The 5q-syndrome: an underdiagnosed form of macrocytic anaemia. Br J Haematol 1983; 54(1): 103-10.

[100] Starczynowski DT, Kuchenbauer F, Argiropoulos B, *et al.* Identification of miR-145 and miR-146a as mediators of the 5q-syndrome phenotype. Nat Med 2010; 16(1): 49-58.

[101] Tiu RV, Gondek LP, O'Keefe CL, *et al.* Prognostic impact of SNP array karyotyping in myelodysplastic syndromes and related myeloid malignancies. Blood 2011; 117(17): 4552-60.

[102] Shendure J, Ji H. Next-generation DNA sequencing. Nat Biotechnol 2008; 26(10): 1135-45.

[103] Yoshida K, Sanada M, Shiraishi Y, *et al.* Frequent pathway mutations of splicing machinery in myelodysplasia. Nature 2011; 478(7367):64-9.

Received: December 2, 2011 **Accepted: January 28, 2012**

Published online 6 March 2013

Nucleic Acids Research, 2013, Vol. 41, No. 7 ***e89***
doi:10.1093/nar/gkt126

An empirical Bayesian framework for somatic mutation detection from cancer genome sequencing data

Yuichi Shiraishi[1,*], Yusuke Sato[2,3], Kenichi Chiba[1], Yusuke Okuno[2], Yasunobu Nagata[2], Kenichi Yoshida[2], Norio Shiba[2,4], Yasuhide Hayashi[4], Haruki Kume[3], Yukio Homma[3], Masashi Sanada[2], Seishi Ogawa[2,*] and Satoru Miyano[1,*]

[1]Laboratory of DNA Information Analysis, Human Genome Center, Institute of Medical Science, The University of Tokyo, 4-6-1, Shirokanedai, Minato-ku, Tokyo 108-8639, Japan, [2]Cancer Genomics Project, Graduate School of Medicine, The University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo 113-8655, Japan, [3]Department of Urology, Graduate School of Medicine, The University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo 113-8655, Japan and [4]Department of Hematology/Oncology, Gunma Children's Medical Center, 779, Shimohakoda, Hokkitsumachi, Shibukawa, Gunma 377-0061, Japan

Received October 14, 2012; Revised January 25, 2013; Accepted February 10, 2013

ABSTRACT

Recent advances in high-throughput sequencing technologies have enabled a comprehensive dissection of the cancer genome clarifying a large number of somatic mutations in a wide variety of cancer types. A number of methods have been proposed for mutation calling based on a large amount of sequencing data, which is accomplished in most cases by statistically evaluating the difference in the observed allele frequencies of possible single nucleotide variants between tumours and paired normal samples. However, an accurate detection of mutations remains a challenge under low sequencing depths or tumour contents. To overcome this problem, we propose a novel method, Empirical Bayesian mutation Calling (https://github.com/friend1ws/EBCall), for detecting somatic mutations. Unlike previous methods, the proposed method discriminates somatic mutations from sequencing errors based on an empirical Bayesian framework, where the model parameters are estimated using sequencing data from multiple non-paired normal samples. Using 13 whole-exome sequencing data with 87.5–206.3 mean sequencing depths, we demonstrate that our method not only outperforms several existing methods in the calling of mutations with moderate allele frequencies but also enables accurate calling of mutations with low allele frequencies (≤10%) harboured within a minor tumour subpopulation, thus allowing for the deciphering of fine substructures within a tumour specimen.

INTRODUCTION

Cancer is caused by genetic alterations in which acquired or somatic gene mutations, together with germline factors, play definitive roles in cancer development. As such, comprehensive knowledge regarding somatic mutations in the cancer genome is indispensable for the ultimate understanding of cancer pathogenesis. In this regard, the recent advances in massively parallel sequencing technologies have provided an unprecedented opportunity to decipher a full registry of somatic events in the cancer genome at a single nucleotide resolution (1). However, accurate detection of somatic mutations from high-throughput sequencing data may not always be a straightforward task because ambiguities in short read alignment and sequencing errors are inevitably introduced during sample preparation and signal processing, making it difficult to discriminate true somatic mutations from sequencing errors, especially for those mutations with low sequencing depths or allele frequencies. The detection of low allele frequency mutations is not only required for specimens with low tumour contents but is also important for capturing minor tumour subclones to understand the heterogeneity of cancer (2–5) and the underlying causes of tumour recurrence and therapeutic resistance.

*To whom correspondence should be addressed. Tel: +81 3 5449 5615; Fax: +81 3 5449 5442; Email: yshira@hgc.jp
Correspondence may also be addressed to Seishi Ogawa. Tel: +81 3 5800 9045; Fax: +81 3 5800 9047; Email: sogawa-tky@umin.ac.jp
Correspondence may also be addressed to Satoru Miyano. Tel: +81 3 5449 5615; Fax: +81 3 5449 5442; Email: miyano@hgc.jp

For calling somatic mutations, each candidate has to be discriminated from germline variants and artifacts appearing from sequencing errors. Although germline variants can be effectively detected by relying on the base calls in paired normal samples, the elimination of sequencing errors may be a more complex task because of uncertain allele frequencies and tumour contents. Most existing approaches have adopted variants whose allele frequencies in tumour samples are significantly higher than those in normal samples, excluding variants whose allele frequencies are high enough to indicate that they are putative germline variants. Sequencing errors can be eliminated to some extent by testing the differences in allele frequencies, as they are expected to occur with equal probability between tumour and normal samples. To measure the significance of the difference in allele frequencies, *SomaticSniper* (6) and *jointSNVmix* (7) estimate the Bayesian posterior probability that tumour and normal samples have different genotypes, whereas our previous approach (8) and *VarScan 2* (9) both rely on the *P*-values from Fisher's exact test.

Although a direct comparison between tumour and normal samples has achieved a measure of success, a more efficient approach to discriminate between sequencing errors and genuine somatic mutations is possible when prior information on sequencing errors is given. In fact, the susceptibility to sequencing errors in each genomic position is not uniform, but there are many common sequencing error-prone sites across different experiments, as shown by several previous studies (10–12) as well as our current study. This implies that, by inferring the susceptibility to sequencing errors at each genomic site, we can achieve greater sensitivity in the detection of somatic mutations at sites with no sequencing errors while efficiently filtering false positives at sequencing error-prone sites (Figure 1).

In this article, we propose a novel statistical approach for the detection of somatic mutations, which explicitly takes into account prior information of sequencing errors. By introducing a Bayesian statistical model, we propose a framework for empirically estimating the distribution of sequencing errors by using a set of non-paired normal samples. Using this approach, we can directly evaluate the discrepancy between the observed allele frequencies and the expected scope of sequencing errors. The proposed approach, which we call Empirical Bayesian mutation Calling (*EBCall*), is superior to several existing methods in calling somatic mutations with moderate allele frequencies. In addition, we demonstrate that *EBCall* can effectively detect a series of somatic mutations that have allele frequencies of <10% with a high degree of accuracy, thereby identifying subclonal structures of cancer cells that cannot otherwise be found.

MATERIALS AND METHODS

Patient samples and sequencing procedures

After receiving informed consent, paired tumour-normal samples were obtained from 20 patients with clear cell renal cell carcinoma (ccRCC) by sampling their specimens during surgical operations. Of the samples obtained, 13 paired tumour-normal samples were used for a performance evaluation of the mutation detection, and all 20 of the normal samples were used for estimating the sequencing errors as non-paired normal reference samples. In addition, to compare the choice of normal reference samples, 20 normal samples collected from patients with paediatric acute myeloid leukemia (ped-AML) were also used; the informed consent for these sample collections were obtained from the patients' parents. This study was approved by the ethics committees of the University of Tokyo and Gunma Children's Medical Center.

Genomic DNA and total RNA were extracted from the samples using QIAamp DNA Investigator kit (Qiagen) and the RNAeasy Total RNA kit (Qiagen) with DNase treatment, respectively, according to the manufacturers' protocols. For whole-exome sequencing, SureSelect-enriched exon fragments were subjected to sequencing using HiSeq 2000, as previously described (8). The ccRCC samples were sequenced from October 2011 to February 2012, whereas the ped-AML samples were sequenced from April 2012 to June 2012. For 10 ccRCC samples, whole-genome sequencing and RNA sequencing were performed using HiSeq 2000, according to standard protocols recommended by Illumina. The mean sequencing depth for each sample was 65.9–223.0 (Supplementary Table S1 and S2).

Outline of the mutation calling method

The outline of *EBCall* is shown in Figure 2. The key concept in *EBCall* is that sequencing data of multiple non-paired normal samples are used to estimate possible sequencing errors at each genomic site. For this purpose, we modelled the sequencing errors that follow a Beta-binomial distribution, the parameters of which were estimated using the sequencing data from multiple non-paired normal samples (Figure 3). The allele frequencies of the observed variants in the tumour DNA were then compared with the inferred sequencing error distribution at the corresponding genomic positions to exclude sequencing errors. Germline Single Nucleotide Polymorphism (SNPs) were eliminated using sequencing data from the paired normal DNA.

Alignment of sequencing data

The sequencing reads were aligned to NCBI Human Reference Genome Build 37 using Burrows-Wheeler Aligner, version 0.5.8 (13) with the default parameter settings. Polymerase chain reaction (PCR) duplications were eliminated using Picard (http://picard.sourceforge.net/). Low-quality reads showing >5 mismatches with the reference genome or those whose mapping quality was <30 were excluded from further analysis as we did in (8).

For RNA sequencing data, a two-step alignment strategy adopted in *Genomon-fusion* (under submission) was used, in which all sequence reads were first aligned to the known transcript sequences (UCSC known genes)

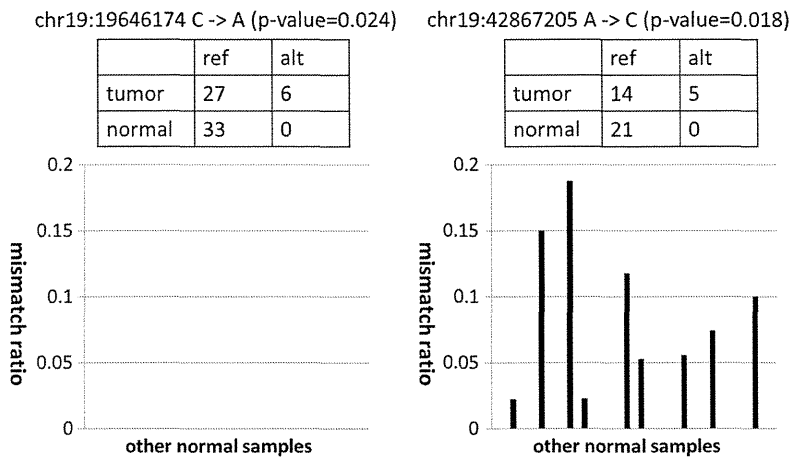

Figure 1. Examples of mismatch ratios of other normal samples for mutation candidates with moderate *P*-values. In both cases, although the mismatch ratios of the target tumour sample were relatively high, the numbers of corresponding supporting variant reads were small. For the candidate on the left, the frequencies of non-reference alleles for other normal samples were consistently zero. Therefore, this supports the prediction that the observed variant reads in the target tumour sample came from a true somatic mutation and not from sequencing errors. On the other hand, for the candidate on the right, we often observed high frequencies of non-reference alleles for several different normal samples. Therefore, the observed variant reads in the target tumour sample likely came from sequencing errors, and it was just by chance that there was no variant read in the target normal sample.

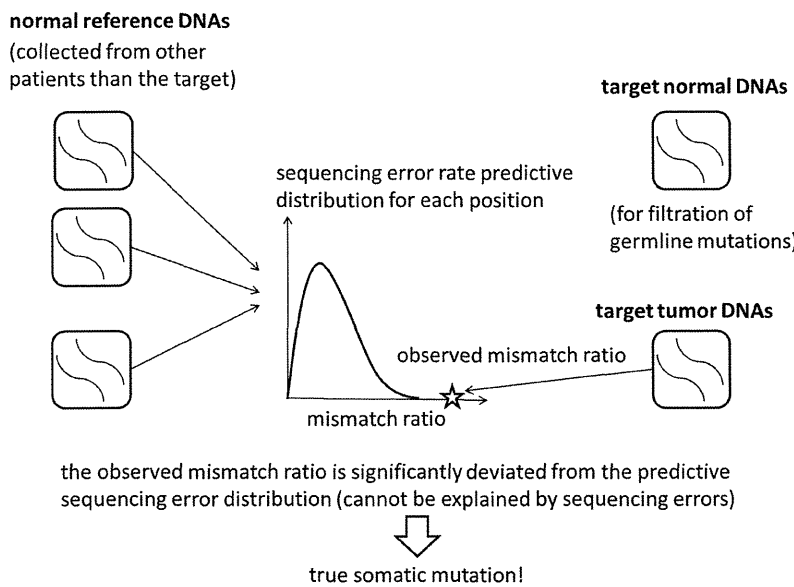

Figure 2. An illustrative description of the proposed method. For each genomic site, the distribution of sequencing errors is estimated using non-paired normal samples from patients other than the target. The mismatch ratio of the target tumour sample is then compared with the distribution. If the mismatch ratio deviates significantly from the distribution, the corresponding variant is then extracted as a somatic mutation candidate. The target normal sample is used for filtering germline mutations.

using bowtie (14), and the non-aligned reads were then aligned to the genome sequences using blat (15). For the whole-genome sequencing data, all reads were aligned using blat.

Definition of variables

Let Ω be an entire set of possible nucleotide variations consisting of combinations of genomic positions and types of nucleotide changes (e.g. chr1:5, C > A or chr20:10 000, A > AAG). Because sequencing errors are often biased to one strand (6,9,16), the number of total (d) and variant reads (x) for a given variant, $v \in \Omega$, were enumerated for each strand separately to distinguish between short reads aligned with the positive ($x_{a,v,+}$, $d_{a,v,+}$) and negative ($x_{a,v,-}$, $d_{a,v,-}$) strands, respectively, where a denotes the type of sample, which is either

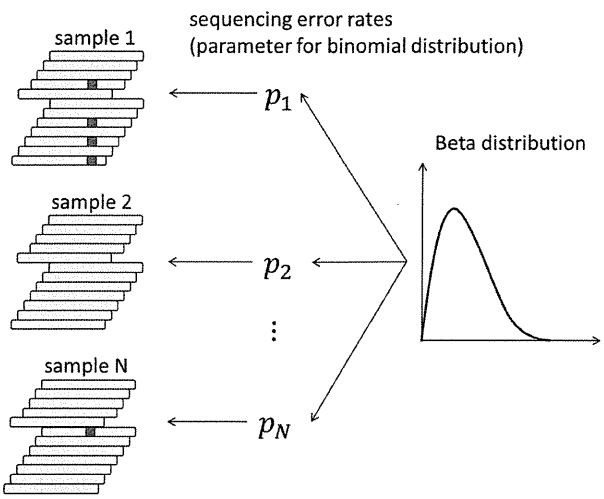

Figure 3. A Beta-binomial sequencing error model. First, the error rate for each sample is generated from the Beta distribution. The number of short reads with sequencing errors is then generated according to the binomial distribution using the parameters of the above error rate for each sample. The parameters of the Beta distribution, which determine the shape of the distribution, are given for each possible variant.

tumour (T), paired normal (N) or non-paired normal reference sample ($R_i, i = 1, 2, \cdots, I$).

Evaluation of sequencing errors using a Beta-binomial model

The number of sequencing errors at a given position in multiple samples is assumed to follow a binomial distribution characterized by a pre-determined parameter, P. Here, we take a Bayesian approach in which the sequencing error rate is a random variable following the Beta distribution, a conjugate prior distribution of the binomial distribution (Figure 3). We adopted a Bayesian approach for the following two reasons. First, although we have discussed that the proneness of sequencing errors is common across multiple experiments to some extent, subtle differences in various factors such as reagents and DNA status can influence the sequencing error rates. Hence, it is inappropriate to assume a homogeneous value for the sequencing error parameters for all experiments. Second, as biological experiments tend to generate a number of outliers, considerably robust inference should be performed. Bayesian modelling, which usually covers a broader range than simple exponential family distributions, serves this purpose.

Given an observed $v \in \Omega$, caused by a sequencing error, the numbers of variant reads, $(x_{R_i,v,\pm})$, in both strands in a normal sample, R_i, are binomially distributed as

$$x_{R_i,v,\pm} \sim \mathrm{Bin}(d_{R_i,v,\pm}, p_{R_i,v,\pm}),\ (i = 1, \cdots, I),$$

where the sequencing error rate $(p_{R_i,v,\pm})$ follows a Beta distribution:

$$p_{R_i,v,\pm} \sim \mathrm{Beta}(\alpha_{v,\pm}, \beta_{v,\pm}).$$

Under these assumptions, a predictive distribution of the number of variant reads, called a Beta-binomial distribution, can be described by the following formula:

$$\Pr(x_{R_i,v,\pm}|d_{R_i,v,\pm}, \alpha_{v,\pm}, \beta_{v,\pm}) = \frac{\Gamma(d_{R_i,v,\pm}+1)}{\Gamma(x_{R_i,v,\pm}+1)\Gamma(d_{R_i,v,\pm}+x_{R_i,v,\pm}+1)} \cdot \frac{\Gamma(x_{R_i,v,\pm}+\alpha_{v,\pm})\Gamma(d_{R_i,v,\pm}-x_{R_i,v,\pm}+\beta_{v,\pm})}{\Gamma(d_{R_i,v,\pm}+\alpha_{v,\pm}+\beta_{v,\pm})} \frac{\Gamma(\alpha_{v,\pm}+\beta_{v,\pm})}{\Gamma(\alpha_{v,\pm})\Gamma(\beta_{v,\pm})},$$

where Γ is the Gamma function. Each Beta distribution is regarded as a prior distribution, and its parameters, $\alpha_{v,\pm}$ and $\beta_{v,\pm}$, are estimated from the observed data of non-paired normal reference samples using a maximum likelihood method, in which the parameter space was restricted to $\alpha_{v,\pm} \geq 0.1$ to avoid over-fitting:

$$\left(\hat{\alpha}_{v,\pm}, \hat{\beta}_{v,\pm}\right) = \arg\max_{\alpha_{v,\pm} \geq 0.1} \sum_{i=1,\cdots,I} \log \Pr(x_{R_i,v,\pm}|d_{R_i,v,\pm}, \alpha_{v,\pm}, \beta_{v,\pm}).$$

EBCall pipeline

In *EBCall* pipeline, somatic mutations were detected using three major steps: the exclusion of less informative variants (step 1) and possible germline variants (step 2), and the sequencing of errors (step 3).

(i) To reduce the computational burden, only variants satisfying all the following conditions are tested in the following steps:
  (a) The total numbers of reads at the relevant position in each strand should be >7 in both the tumour and paired reference:
  $$d_{T,v} = d_{T,v,+} + d_{T,v,-} > 7,$$
  $$d_{N,v} = d_{N,v,+} + d_{N,v,-} > 7;$$
  (b) The mismatch ratio in the tumour sample should be >0.1:
  $$x_{T,v}/d_{T,v} > 0.1,\ x_{T,v} = x_{T,v,+} + x_{T,v,-};$$
  (c) The variant should be supported by >3 reads:
  $$x_{T,v} > 3.$$

(ii) The following are excluded as putative germline polymorphisms/variants:
  (a) Those with a mismatch ratio of >0.02 in the paired normal sample:
  $$x_{N,v}/d_{N,v} > 0.02,\ x_{N,v} = x_{N,v,+} + x_{N,v,-};$$
  (b) Those for which the number of observed variant reads, $x_{N,v}$, is within the 99% confidential interval of the expected read number, under the assumption of a binominal distribution of $\mathrm{Bin}(d_{N,v}, 0.5)$ for dichotomous germline polymorphisms; and

(c) Those registered in either dbSNP131, the 1000 genomes project, or our internal SNP database.

(iii) For each of the remaining variants, the cumulative probabilities for the observed $x_{T,v,+}$ and $x_{T,v,-}$ under the null hypothesis, H_0: the variant is from sequencing errors, are provided by

$$P_{\pm}(v) = \sum_{x \geq x_{T,v,\pm}} \Pr(x | d_{T,v,\pm}, \hat{\alpha}_{v,\pm}, \hat{\beta}_{v,\pm}).$$

The combined P-value, $P(v)$, corresponding to two independent strands, $P_+(v)$ and $P_-(v)$, is obtained according to Fisher's method:

$$P(v) = \Pr(\chi_4^2 \geq P_+(v) + P_-(v)),$$

where χ_4^2 is a random variable distributed from the chi-square distribution with four degrees of freedom. H_0 is then tested with a type I error, (=0.001 by default), for mutation calling. For base substitution mutations, we only used reads with a base quality of ≥15 at the corresponding positions for counting sequencing depths and variant reads. Each threshold value used above can be changed according to the purpose.

Evaluation of sequencing error susceptibility among multiple samples

To examine how many error-prone sites exist and how much they correlate among different experiments, we evaluated the sequencing error proneness by using normal samples of 20 ccRCC and 20 ped-AML patients. For an accurate evaluation of sequencing errors, we included only variants whose sequencing depths of positive and negative strands are >20 for all samples. Furthermore, we removed putative germline variants satisfying the following conditions at least for one sample:

(i) Sequencing depths are >20;
(ii) The non-reference allele frequency is >0.2; and
(iii) At least one variant read is observed in both positive and negative strands.

Furthermore, for base substitutions, we only used reads with a base quality of ≥15 at the corresponding positions for counting sequencing depths and variant reads, as variants with low quality bases are often filtered in actual mutation callings.

Comparison with other mutation calling methods

We evaluated the performance of *EBCall* for calling somatic mutations with moderate allele frequencies (>0.1) through a comparison with other publically available methods, along with our own previous approach (designated as *Genomon-Fisher*) (8), which is obtained by replacing step 3 in *EBCall* with Fisher's exact test for measuring the difference in the allele frequencies of the variants between the tumour and paired normal samples. The default setting was applied for running both *Genomon-Fisher* and *VarScan*. For *SomaticSniper*, the -q 30 -Q 15 option was used. In all cases, low-quality reads with >5 mismatches or a mapping quality of <30 were excluded in advance, as mentioned earlier in the text for *EBCall*. Furthermore, the same filtering procedures as the step 1 and 2 in *EBCall* were applied to all the method to equalize the conditions of sequencing depths and allele frequencies. For the comparison, somatic mutations were detected for whole-exome sequencing data from 10 clear cell carcinoma samples, for which a set of true positive mutations, Φ, was defined using whole genome/RNA sequencing data as follows:

$$\Phi = \{v \in \Omega | d_{N^G,v} \geq 8, x_{N^G,v}/d_{N^G,v} \leq 0.03, n_{N^G,v} \leq 1\} \cap \{\{v \in \Omega | n_{T^G,v} \geq 4, x_{T^G,v}/d_{T^G,v}, \geq 0.08\} \cup \{v \in \Omega | x_{T^R,v} \geq 4, x_{T^R,v}/d_{T^R,v} \geq 0.08\}\}$$

where N^G and T^G/T^R denote whole genome/RNA sequencing data from normal and tumour samples, respectively. Herein, we did not count mutation candidates that do not satisfy $d_{N^G,v} \geq 8$ for either true or false positives, as they may be germline mutations. Mutations in non-coding regions excluding splice-sites were removed, where the gene annotations were performed using ANNOVAR (17). In addition, as *SomaticSniper* does not call InDels, we mainly concentrated substitutions for this comparison.

Validation of somatic mutations with low allele frequencies (<0.1)

We evaluated the performance of *EBCall* for calling somatic mutations with low allele frequencies (≤0.1) by changing the threshold value for the mismatch ratio in the tumour sample to $x_{T,v}/d_{T,v} > 0.02$. For somatic mutations with low allele frequencies to be accurately called, we further imposed that a somatic mutation satisfy $-\log_{10}(p^{\text{Fisher}}) > 0.8$, where p^{Fisher} is the P-value in Fisher's exact test. Furthermore, we stipulated that the number of read pairs with the variant is greater than 3 so as to avoid double counting of a variant located in both the two reads of single read pair with a small insert size. Herein, we included all the mutations including those in the non-coding regions to increase the number of mutations from various clonal populations. All candidate somatic mutations were validated by deep sequencings of the PCR products of the relevant loci using HiSeq 2000, as previously described (8). A candidate variant is thought to be validated if and only if all the following conditions are satisfied:

(i) The sequencing depth is >5000 for both positive and negative strands;
(ii) The mismatch ratio in the paired normal samples is <0.5%; and
(iii) The mismatch ratio in the tumour sample is 5 times larger than that of the normal sample.

To compare the performances of *EBCall* and *Genomon-Fisher*, we also validated several candidates that were not called from *EBCall* but were called from *Genomon-Fisher* from the top in terms of the P-values.

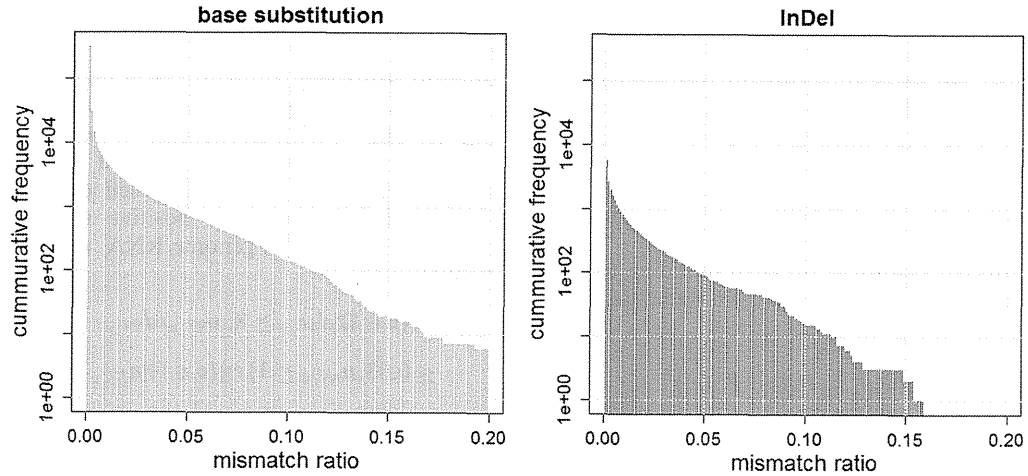

Figure 4. Two bar plots showing the numbers of base substitutions and InDels, whose mean mismatch ratios are above the determined threshold values. For instance, the numbers of base substitutions with mean mismatch ratios of more than 0.01, 0.02, and 0.05 are 4472, 2232, and 727, respectively, while those of InDels are 717, 350, and 89, respectively.

RESULTS

Susceptibility to sequencing errors

The distribution of mean sequencing error rates is shown in Figure 4. Although the error rates were calculated using high-quality sequencing reads (with a mapping quality of ≥30) and high-quality bases (with a base quality of ≥15) for substitution errors, there were many sites with relatively high sequencing error rates, indicating the existence of many sequencing error-prone sites. The higher rate of sequencing errors causes the more harm. When both the tumour and normal samples have a 2% sequencing error rate, the probability that the *P*-value of Fisher's exact test is below 0.05 is ~0.5% for the positions with a sequencing depth of 80 for tumour and normal samples. On the other hand, when the sequencing error rate is 5%, this probability increases to ~2.2%. As there are 2582 sites with >2% mean sequencing error rate, we will obtain at least 13 false positives at the same threshold for data with a mean sequencing depth of 80. Furthermore, a subtle difference in the sequencing error rates between the tumour and normal samples caused by inconsistencies in the experimental conditions will generate an even higher rate of false positives under real situations. Although not a small proportion of sequencing errors was strand specific, there were still many variants prone to bi-directional sequencing errors (Supplementary Figure S1).

We next examined the consistency of sequencing error rates across different sets of samples (Figure 5). The sequencing error rates were highly correlated between the two sets of 10 ccRCC samples. The sequencing error rates were less consistent between the sets of 10 ccRCC samples and 10 ped-AML samples, indicating that it is better to use normal samples collected under conditions as similar as possible to predict sequencing errors. The correlations for InDels were stronger compared with the base substitutions, implying that the sequencing errors found in InDels are more systematic.

Performance comparison with other algorithms for moderate allele frequencies

To compare the performance of different mutation calling algorithms, we first sorted the candidate mutations according to the accompanying confidence score for each method (the combined *P*-value for *EBCall*, the *P*-value of Fisher's exact test for *Genomon-Fisher* and *VarScan 2* and a somatic score for *SomaticSniper*) and checked the relationships between the number of candidates and the number of true positives (Figure 6). For mutations with high confidence values, there was no clear difference among the different calling methods used. However, for low confidence values (i.e. after the 500th confident mutation), *EBCall* showed higher true positive results than the other methods, as indicated by the upward deviation of the plot in Figure 6. The true positive rates (TPR) of *SomaticSniper* decreased more rapidly than those of other methods, whereas *VarScan 2* and *Genomon-Fisher* show comparable plots probably reflecting the fact that both methods are based on Fisher's exact test. For InDels, *EBCall* showed at least similar efficiency to *VarScan 2* and *Genomon-Fisher* (Supplementary Figure S2).

When using 20 ped-AML normal samples as non-paired normal reference samples, the performance of *EBCall* slightly worsened, which is reasonable considering the lower correlation of sequencing errors between the ccRCC samples and ped-AML samples. However, the TPR was still higher than in the other existing approaches, indicating that the proposed approach is robust to the choice of normal reference samples to a certain extent. To examine the required number of normal reference samples, the performance of *EBCall* for different numbers of normal reference samples was measured. As shown in Supplementary Figure S3, it took 15–17 samples for a performance saturation for both the ccRCC and ped-AML reference samples.