Genome analysis of myelodysplastic syndromes among atomic bomb survivors in Nagasaki

Masataka Taguchi,[1] Hiroyuki Mishima,[2] Yusuke Shiozawa,[3] Chisa Hayashida,[2] Akira Kinoshita,[2] Yasuhito Nannya,[3] Hideki Makishima,[3] Makiko Horai,[1] Masatoshi Matsuo,[1] Shinya Sato,[1] Hidehiro Itonaga,[1] Takeharu Kato,[1] Hiroaki Taniguchi,[4] Daisuke Imanishi,[1] Yoshitaka Imaizumi,[1] Tomoko Hata,[1] Motoi Takenaka,[5] Yukiyoshi Moriuchi,[4] Yuichi Shiraishi,[6] Satoru Miyano,[6,7] Seishi Ogawa,[3] Koh-ichiro Yoshiura[2] and Yasushi Miyazaki[1]

[1]Department of Hematology, Atomic Bomb Disease Institute, Nagasaki University Graduate School of Biomedical Sciences, Nagasaki; [2]Department of Human Genetics, Atomic Bomb Disease Institute, Nagasaki University Graduate School of Biomedical Sciences, Nagasaki; [3]Department of Pathology and Tumor Biology, Graduate School of Medicine, Kyoto University, Kyoto; [4]Department of Hematology, Sasebo City General Hospital, Sasebo; [5]Department of Dermatology, Nagasaki University Graduate School of Biomedical Sciences, Nagasaki; [6]Laboratory of DNA Information Analysis, Human Genome Center, The Institute of Medical Science, The University of Tokyo, Tokyo and [7]Laboratory of Sequence Analysis, Human Genome Center, The Institute of Medical Science, The University of Tokyo, Tokyo, Japan

 doi:10.3324/haematol.2019.219386

Received: February 12, 2019.

Accepted: May 16, 2019.

Pre-published: May 17, 2019.

Correspondence: *YASUSHI MIYAZAKI* - y-miyaza@nagasaki-u.ac.jp

Supplemental file

[1] Supplemental methods

Patients

This study was approved by the ethics committee of Nagasaki University. Samples and data were collected with written informed consent in accordance with the Declaration of Helsinki. Cryopreserved bone marrow mononuclear cells (BMMCs) and BM smears of patients who had already died before starting this study were analyzed with the approval of the ethics committee of Nagasaki University.

In this study, we analyzed 32 patients having MDS, and 3 patients with idiopathic cytopenia of undetermined significance (ICUS) among A-bomb survivors (Supplemental Table 1), diagnosed at Nagasaki University Hospital and Sasebo City General Hospital, whose samples were collected from 1995 to 2015. The estimated dose of A-bomb radiation at 1, 2, and 2.7 km from the hypocenter was approximately 9.8, 0.14, and 0.005 Gy, respectively, in Nagasaki according to the dosimetry system, DS02 [1], and the survivors exposed within 2.7 km of the hypocenter (exposed to more than 0.005Gy) are usually defined as "proximally exposed survivors" by Radiation Effect Research Foundation [1]. We followed this criterion, and the patients were classified into two groups based on their exposure distance from the hypocenter: patients exposed within 2.7 km of the hypocenter were categorized as the proximally exposed (PE) group, and the distally exposed (DE) group comprised patients exposed between 2.7 and 10 km of the hypocenter and those who entered within a 2 km radius from the hypocenter within two weeks after the bombing (they did not exhibit acute radiation syndrome). There were 18 and 17 patients in the PE and DE groups, respectively. Considering the dose of A-bomb radiation in DE-group (less than 0.005Gy), we compared data between PE and DE group in this study to clearly search the effect of radiation on genome alteration.

Genomic DNA extraction

Genomic DNA (gDNA) were extracted from BMMCs and BM smears using QIAamp DNA Blood Mini Kit (QIAGEN) according to the manufacture's protocol. The standard procedure using proteinase K and phenol-chloroform was applied to dermal fibroblasts.

Next generation sequencing

1) Whole exome sequencing (WES)

Non-amplified, 200 ng of gDNA extracted from BMMCs and dermal fibroblasts were sonicated to a mean of 150-200 bp by the Covaris sonicator E220 (Covaris Inc.). The sonicated gDNA were processed using SureSelect Human All exon v5 Kit (Agilent Technologies) according to the manufacture's protocol to enrich protein-coding regions, and then were subjected to Illumina HiSeq 2500 platform to generate 100 bp paired-end reads. The mean depth of coverage was 102× (range 79–121×) (Supplemental Table 5).

2) Targeted capture sequencing (T-S)

Non-amplified gDNA from BM smears were sonicated by setting to 500 bp using the Covaris sonicator E220 because the gDNA had already been fragmented especially in the old BM smears. Twenty nanogram of sonicated DNA were processed using KAPA Hyper Prep kit (KAPA Biosystems), and then hybridization-based target-capture was performed using the custom SeqCap EZ Choice Library (Roche NimbleGen, Inc.) according to the manufacture's protocol. The target-capture enriched 154 genes which were the known and putative drivers in hematological malignancies [2-5] or the candidates in this study (Supplemental table 3). Enriched DNA were subjected to Illumina HiSeq 2500 platform to generate 100 bp paired-end reads.

We also performed targeted capture sequencing (T-S) using non-amplified gDNA of BMMCs of two patients (patient's ID, U-WES-3 and -5, see main text) in order to compare the power of detecting copy number alterations between SNP-array and T-S (Supplemental figure 2). The mean depth of coverage was 223× (155–285×) (Supplemental Table 5).

3) Whole genome sequencing (WGS)

Non-amplified, 1000 ng of gDNA were sonicated to a mean of 400 bp by the Covaris sonicator (Covaris Inc.) and were subjected to Illumina HiSeq 2000 platform to generate 100 bp paired-end reads according to the manufacture's protocol. The mean depth of coverage was 41× (31–55×) (Supplemental Table 5).

Alignment of sequencing reads

Sequencing reads of WES and WGS were aligned to the human reference genome (hg19) by NovoalignMPI version 3.02.05 (Novocraft Technologies Sdn Bhd) on default setting. Base quality scores were recalibrated during the alignment using SNPs information of the dbSNP138 dataset. Duplicated reads were excluded using Picard MarkDuplicates (http://broadinstitute.github.io/picard/) version 1.112.

Sequencing reads of T-S were aligned to hg19 by Burrows Wheeler Aligner version 0.7.10 [6] on default settings. PCR duplicates were removed using Picard MarkDuplicates version 1.39.

For the following process of mutation calling of WES and WGS data with matched germline controls, the value of edit distance in NM tag of SAM files generated from Novoalign MPI was corrected using in-house script (https://github.com/misshie/novo-tag-adjust). NovoalignMPI uses the reference sequence containing SNPs information of the dbSNP138 dataset, and the bases at SNP positions are presented as ambiguous bases in the reference sequence. Thus, mismatching positions to ambiguous reference bases are counted as "edit distance" in NM tag even if the aligned bases are matched to the reference bases of hg19, resulting in an increase of the value in NM tag. In order to correct the overestimated value, we counted the number of ambiguous reference bases in MD tag of SAM files, except for the bases in deleted regions, and deducted the number from the value in NM tag using the in-house script.

Mutation Detection Pipeline

1) Whole exome sequencing with matched germline controls

In the data of WES with matched germline controls, single nucleotide variants (SNVs) and insertions and deletions (INDELs) were called by *EBCall* algorithm [7] with the following process:

1) Omitting the variants found in normal samples with variant allele frequencies (VAFs) of > 2%;

2) Omitting the variants of $-\log_{10}(P^{\mathrm{Fisher}}) < 1$, where P^{Fisher} is the *P*-value in Fisher's exact test;

3) Omitting the variants of $-\log_{10}(P^{\mathrm{Comb}}) < 3$, where P^{Comb} is the combined *P*-value for

EBCall.

The called variants were annotated using ANNOVAR [8] with the following databases: GENCODE v19 [9]; Human Genetic Variation Database v1.42 (HGVD, http://www.genome.med.kyoto-u.ac.jp/SnpDB) consisting of WES data of 1208 individuals in Japan; Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP), Seattle, WA, (ESP6500SI, http://evs.gs.washington.edu/EVS/), the 1000Genome project 2014 August release (1000Genome) [10], and Catalogue of Somatic Mutations in Cancer v68 (COSMIC) [11]. Polymorphic variants were excluded if they were found in either of HGVD, ESP6500si, or 1000Genome with minor allele frequencies (MAFs) of > 0.14% [2]. Then, all candidates of somatic mutations were subjected to validation described below.

2) Whole Exome Sequencing without matched germline controls

EBCall is applied only for paired samples. Thus, we used HaplotypeCaller of Genome Analysis Toolkit (GATK) version 3.1-1 on default setting for the data of WES without matched germline controls. Variants were annotated using ANNOVAR with the same databases as described above, and polymorphic variants were excluded with the same way. In order to detect known and putative drivers in hematological malignancies [2-5] and the candidates in this study, we selected 356 genes for the validations (Supplemental table 2).

3) Targeted capture sequencing

In the data of T-S without matched germline controls, SNVs and INDELs were called using the established pipeline [3, 12-14] with the following parameters:

1) Mapping Quality score was ≥ 40;
2) Base Quality score was ≥ 20;
3) Number of SNVs on the same read was < 5
4) Number of INDELs on the same read was < 2

T-S data were also processed by NovoalignMPI version 3.03.02 with the read alignment using HaplotypeCaller of GATK version 3.4-46 on default setting for high confidence of mutation calling. Variants were annotated using ANNOVAR with the following databases: GENCODE v19; HGVD v1.42; Integrative Japanese Genome Variation Database (1KJPN) consisting of WGS data of 1070 Japanese individuals from Tohoku

University Tohoku Medical Megabank Organization cohort [15]; ESP6500SI-V2, the 1000Genome project 2015 August release (1000G15Aug), and COSMIC v70.

Following variants were excluded:

1) Variants in non-coding regions;

2) Polymorphic variants found in either of HGVD, 1KJPN, ESP6500SI-V2, or 1000G15Aug with MAFs of > 0.14% [2];

3) Polymorphic variants or sequencing errors found in in-house SNPs database.

SNVs fulfilling one of the following conditions were selected as candidates of validation:

1) Number of the variant reads was ≥ 5;

2) Number of the variant reads was 2, 3, or 4, and the variants were also called by HaplotypeCaller;

3) Number of the variant reads was 2, 3, or 4, and the variants were found in COSMIC v70 as those in hematopoietic and lymphoid tissues.

For INDELs, we applied the following conditions:

1) Number of variant reads was ≥ 3;

2) Number of variant reads was 2 and the variants were found in COSMIC v70 as those in hematopoietic and lymphoid tissue;

3) Number of variant reads was 2 and the variants had more than 3 bp of insertion or deletion.

Then, final list of candidates was validated using amplicon sequencing and/or Sanger sequencing described below.

4) Whole Genome Sequencing

SNVs were called using the established pipeline [14] with some modifications:

1) Omitting the reads of which mapping quality were < 30;

2) Omitting the reads in which more than four SNVs were found;

3) Omitting the reads of which depth were < 10 in either tumor or normal samples;

4) Adopting SNVs as somatic candidates if more than six of variant reads were found in tumor and zero in normal samples.

The called SNVs were annotated using ANNOVAR with the same databases in WES analysis. Further filtering was performed:

1) Excluding the variants in the regions of segmental duplication and tandem repeat

identified by tandem repeats finder [16];

2) Excluding the variants found in in-house SNPs database with minor allele frequencies (MAFs) of ≥ 0.25%;

3) Excluding the variants which were on non-coding regions and found in dbSNP138;

4) Excluding the polymorphic variants found in either of HGVD, ESP6500SI, or 1000Genome with MAFs of > 0.14% [2].

Then, final list of candidates of somatic mutations was obtained. We randomly picked SNVs on both coding and non-coding regions from the list, and performed validation using amplicon sequencing and/or Sanger sequencing described below.

Validation of candidate mutations

In order to validate the candidate mutations and calculate more accurate VAFs, amplicon-based deep sequencing was performed. Primers were designed with Ion AmpliSeq designer v3.6, v4.0, or v5.4.1 (Thermo Fisher Scientific, Waltham, MA, USA) or Primer3 [17,18]. Multiplex PCR was performed on 10 to 20 ng of non-amplified gDNA with Ampliseq-designed primers using KAPA2G FAST Multiplex Kits (KAPA Biosystems, MA) [19]. The conditions of multiplex PCR were as follows: initial denaturation and enzyme activation for 2 min at 95 °C, followed by 25 to 30 thermal cycles of denaturation for 30 s at 95 °C, annealing for 2 min at 60 °C, and extension for 2 min at 72 °C. After amplification, uracil DNA glycosylase (final concentration 0.05 U/μL) and endonuclease IV (final concentration 0.1 U/μL) were added directly to the reaction mixture and incubated at 37 °C for 30 min to separate uracil-containing sites. PCR was also performed on 10 ng of non-amplified gDNA using KOD FX (TOYOBO) with the primers designed by Primer3 according to the manufacture's protocol. The amplicons, which were yielded with Ampliseq- or Primer3-designed primers, were mixed for each patient, and the mixtures were purified using Agencourt AMPure® XP (Beckman Coulter). Then, DNA libraries were constructed using a KAPA HTP library construction kit for Illumina with adaptor oligonucleotides suitable for the Illumina pair-end index sequencing protocol (KAPA). The libraries were then subjected to Illumina MiSeq platform to generate 150-250 bp paired-end reads according to the manufacture's protocol. Sequencing reads were aligned to the human reference genome (hg19) using NovoalignMPI version 3.02.05 on default settings. Duplicated reads were not omitted. Variants were confirmed by visual inspection with

Integrative Genomics Viewer (IGV) [20,21]. Using deep sequencing data, VAFs were calculated as follows: (variant read count) / (reference read count + variant read count). Sanger sequencing was performed if the candidates could not be evaluated by the amplicon sequencing. For all patients, *FLT3-ITD* mutation was analyzed by Sanger sequencing. For the WGS data, we randomly picked up SNVs on both coding and non-coding regions and validated them using the amplicon-based deep sequencing and/or Sanger sequencing, and the true positive rate was 96.3 % (156 out of 162 variants). Validated SNVs with a read depth of 250 or more were used to infer the clonal architecture of MDS in three patients (U-WES-3, 4, and 7) using the PyClone model [22] with the default settings.

Annotation of variants identified in the sequencing without matched germline controls

In the sequencing without matched germline controls, it is impossible to confirm somatic or germline mutations. Thus, we annotated the validated variants as "oncogenic", "possibly oncogenic", or "unknown significance" according to the following criteria [2]:

1) Oncogenic

- Known oncogenic variants of myeloid malignancies previously reported in the literatures;
- Truncating variants (nonsense mutations, splice site mutations or frameshift INDELs) in genes related to myeloid malignancies.

2) Possibly oncogenic

- Previously unreported variants within ±3 amino acid of known oncogenic variants in COSMIC.

3) Unknown significance

- Variants found outside the cluster of known oncogenic variants;
- Variants in the genes of which role in myeloid malignancies is not yet well established.

In the main text, the phrase "oncogenic mutations" includes the "oncogenic" and "possibly oncogenic" variants.

Copy number analysis

For all the U-WES and B-WES-T cases and one T-S case (T-S-13), the DNA copy number alterations (CNAs) were analyzed with a SNP array (CytoScan HD Array, Affymetrix, Santa Clara, CA) according to the manufacturer's protocol. The resulting CEL files were processed with the Chromosome Analysis Suite (ChAS) software version 2.1. (Affymetrix). CNAs were detected using both this software and visual inspection. Somatic CNAs were confirmed using matched germline controls for the U-WES cases.

Annotation of CNAs in the analysis without matched germline controls

CNAs identified by SNP-array or T-S without matched germline controls were annotated as "oncogenic" if the alterations were fulfilled the following criteria [23-25]:

1) CNAs identified by conventional G-banding;
2) Loss or gain larger than 5 Mb;
3) Loss or gain smaller than 5 Mb and including known driver genes of hematological malignancies;
4) Losses or gains smaller than 5 Mb and identified a lot on the same chromosome
5) Telomeric copy neutral loss of heterozygosity (CN-LOH) larger than 8.7 Mb
6) Interstitial CN-LOH larger than 25 Mb

CNAs in T-S cases were identified in the sequencing data using the CNACS pipeline [26], because of the insufficient quality and quantity of gDNA for SNP array. Although copy number states of whole chromosomes could not be evaluated, frequently affected regions in MDS, such as the long arms of chromosome 5 (5q), 7q, and 20q, were included among the targets. We also generated more probes towards the genes affected by 11q deletion to evaluate the whole arm of 11q because this region was of interest in this study. Apparent changes in copy number could be detected with this method, although low-level mosaicisms could not (Supplemental Figure 2). The oncogenic roles of the CNAs were annotated according to the same criteria used for the SNP array analysis.

Confirmation of low-level mosaicism using droplet digital PCR

Among the identified CNAs, low-level mosaicisms on chromosome 11 in two patients (U-WES-5 and -8) were confirmed using droplet digital PCR (ddPCR, Bio-Rad

Laboratories, Hercules, CA) (Supplemental Figure 1). Primers and probes for *ATM* and *KMT2A* were designed using Primer3 and PrimerQuest (PrimeTime® qPCR Probes, Integrated DNA Technologies, Inc.). *ALB* was used as 2-copy Y internal control (Supplemental table 12). Ten nanogram of gDNA were subjected to ddPCR according to the manufacturer's protocol. The Thermal cycling conditions were 95℃×10 min (1 cycle), 96℃×30 sec and 56.6℃×2 min (40 Cycles), 98℃×10 min (1 cycle), and 4℃ hold. The data were analyzed using QuantaSoft (BioRad).

Statistical methods

Statistical analysis was performed using R version 3.4.1 with Rcmdr version 2.3-3 and EZR package [27]. Dichotomous and continuous variables were compared by Fisher's exact test and Student's T-test, respectively. Survival time was estimated and compared using Kaplan-Meier method with log-rank test. P-values were two-sided, and P-values of < 0.05 were considered as significant.

[2] Legends for supplemental tables

Supplemental Table 1.

Detailed characteristics of patients in this study. U-WES, unbiased whole exome sequencing (WES with germ ine control); B-WES-T, biased WES with target validation of 356 genes (WES without germline control); T-S, targeted apture sequencing of 154 genes.

Supplemental Table 2.

Number of patients in proximally exposed (PE) group and distally exposed (DE) group analyzed with various methods of sequencing and copy number analysis. WGS, whole genome sequencing.

Supplemental Table 3.

Target genes for the validation of whole exome sequencing without matched germline control (validated genes for B-WES-T).

Supplemental Table 4.

Genes for targeted capture sequencing (for T-S cases).

Supplemental Table 5.
Depth of coverage of each sequencing method.

Supplemental Table 6.
Somatic mutations on coding exons in U-WES cases.

Supplemental Table 7-1.
Somatic mutations on the whole genome of U-WES-3.

Supplemental Table 7-2.
Somatic mutations on the whole genome of U-WES-4.

Supplemental Table 7-3.
Somatic mutations on the whole genome of U-WES-7.

Supplemental Table 8.
Validated mutations in B-WES-T cases.

Supplemental Table 9.
Validated mutations in T-S cases.

Supplemental Table 10.
Frequencies of the mutated genes categorized to assumed functional pathways.

Supplemental Table 11.
Details of the copy number alterations in each case.

Supplemental Table 12.
Primer sequence for droplet digital PCR.

[3] Reference for Supplemental Methods

[1] Young R, Kerr G eds. Reassessment of the Atomic Bomb Radiation Dosimetry for Hiroshima and Nagasaki, Dosimetry System 2002, Report of the Joint US-Japan Working Group. Hiroshima: Radiation Effects Research Foundation; 2005

[2] Papaemmanuil E, Gerstung M, Malcovati L, et al. Clinical and biological implications of driver mutations in myelodysplastic syndromes. *Blood*. 2013;122(22):3616–27– quiz 3699.

[3] Haferlach T, Nagata Y, Grossmann V, et al. Landscape of genetic lesions in 944 patients with myelodysplastic syndromes. *Leukemia*. 2014;28(2):241–247.

[4] Cancer Genome Atlas Research Network, Ley TJ, Miller C, et al. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med*. 2013;368(22):2059–2074.

[5] Kihara R, Nagata Y, Kiyoi H, et al. Comprehensive analysis of genetic alterations and their prognostic impacts in adult acute myeloid leukemia patients. *Leukemia*. 2014;28(8):1586–1595.

[6] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25(14):1754–1760.

[7] Shiraishi Y, Sato Y, Chiba K, et al. An empirical Bayesian framework for somatic mutation detection from cancer genome sequencing data. *Nucleic Acids Research*. 2013;41(7):e89–e89.

[8] Wang K, Li M, Hakonarson H. ANNOVAR: Functional annotation of genetic variants from next-generation sequencing data. *Nucleic Acids Research*. 2010 38(16):e164.

[9] Harrow J, Frankish A, Gonzalez JM, et al. GENCODE: The reference human genome annotation for The ENCODE Project. *Gemome Res*. 2012 22(9):1760-1774.

[10] Altshuler DM, Durbin RM, Abecasis GR, et al. An integrated map of genetic variation from 1,092 human genomes. *Nature.* 2012 491(7422):56-65

[11] Forbes SA, Beare, Gunasekaran, et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res*. 2015 43(Database issue):D805-811

[12] Yoshida K, Sanada M, Shiraishi Y, et al. Frequent pathway mutations of splicing

machinery in myelodysplasia. *Nature*. 2011;478(7367):64-69.
[13] Yoshizato T, Dumitriu B, Hosokawa K, et al. Somatic Mutations and Clonal Hematopoiesis in Aplastic Anemia. *N Engl J Med*. 2015;373(1):35-47.
[14] Suzuki H, Aoki K, Chiba K, et al. Mutational landscape and clonal architecture in grade II and III gliomas. *Nat Genet*. 2015;47(5):458-468.
[15] Nagasaki M, Yasuda J, Katsuoka F, et al. Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. *Nat Commun*. 2015;6(1):8018.
[16] Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research*. 1999;27(2):573–580.
[17] Untergasser A, Cutcutache I, Koressaar T, et al. Primer3--new capabilities and interfaces. *Nucleic Acids Research*. 2012;40(15):e115–e115.
[18] Koressaar T, Remm M. Enhancements and modifications of primer design program Primer3. *Bioinformatics*. 2007;23(10):1289–1291.
[19] Horai M, Mishima H, Hayashida C, et al. Detection of de novo single nucleotide variants in offspring of atomic-bomb survivors close to the hypocenter by whole-genome sequencing. *J. Hum. Genet.* 2017;386(11):469.
[20] Robinson JT, Thorvaldsdóttir H, Winckler W, et al. Integrative genomics viewer. *Nat Biotechnol.* 2011;29(1):24–26.
[21] Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics*. 2013;14(2):178–192.
[22] Roth A, Khattra J, Yap D, et al. PyClone: statistical inference of clonal population structure in cancer. *Nat Meth*. 2014;11(4):396–398.
[23] Simons A, Sikkema-Raddatz B, de Leeuw N, et al. Genome-wide arrays in routine diagnostics of hematological malignancies. *Hum. Mutat*. 2012;33(6):941–948.
[24] 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):4552–4560.
[25] Jerez A, Gondek LP, Jankowska AM, et al. Topography, clinical, and genomic correlates of 5q myeloid malignancies revisited. *J. Clin. Oncol*. 2012;30(12):1343–1349.
[26] Yoshizato T, Nannya Y, Atsuta Y, et al. Genetic abnormalities in myelodysplasia

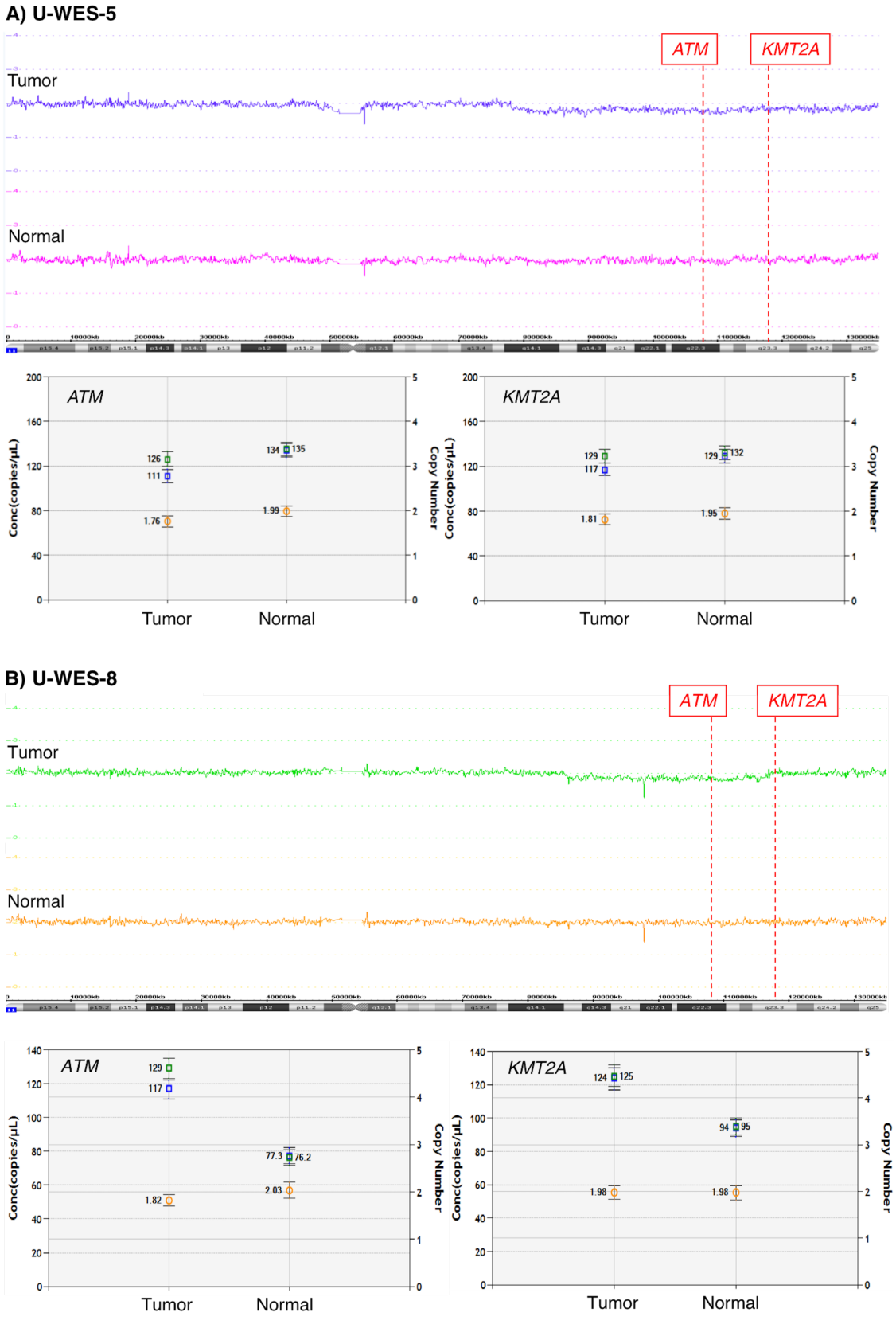
and secondary acute myeloid leukemia: impact on outcome of stem cell transplantation. *Blood*. 2017;129(17):2347–2358.
[27] Kanda Y. Investigation of the freely available easy-to-use software "EZR" for medical statistics. *Bone Marrow Transplant*. 2013;48:452–458.

[4] Supplemental figures and legends.

Supplemental figure 1.

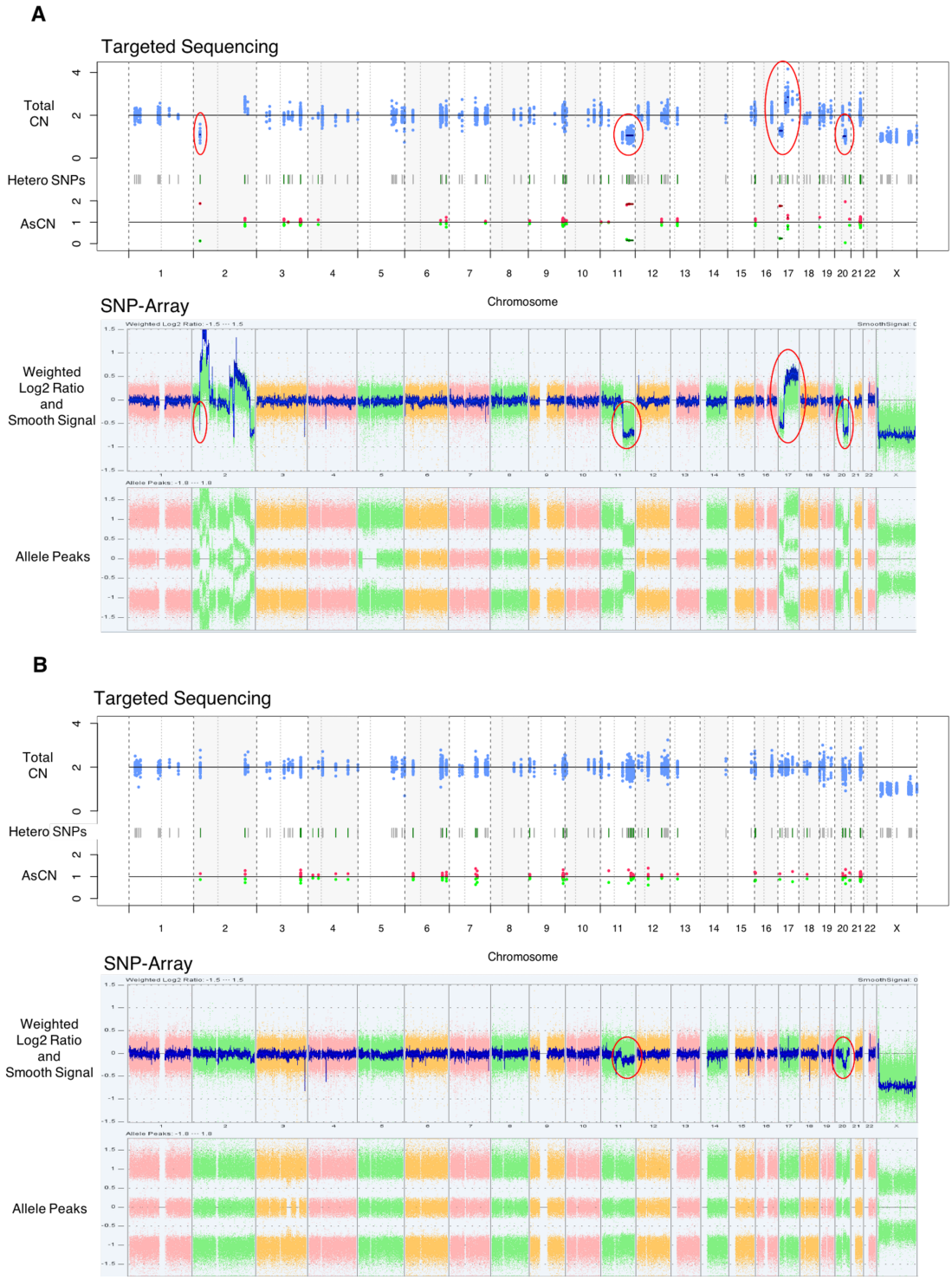
Confirmation of low-level mosaicism of copy number loss on chromosome 11 using ddPCR.

Copy number alterations on *ATM* and *KMT2A* were analyzed. Upper and lower figures were for SNP-array (smooth signal), and ddPCR, respectively. "Tumor" and "Normal" means BMMCs and dermal fibroblasts, respectively. (A) Results of U-WES-5. (B) Results of U-WES-8.



Supplemental figure 2.

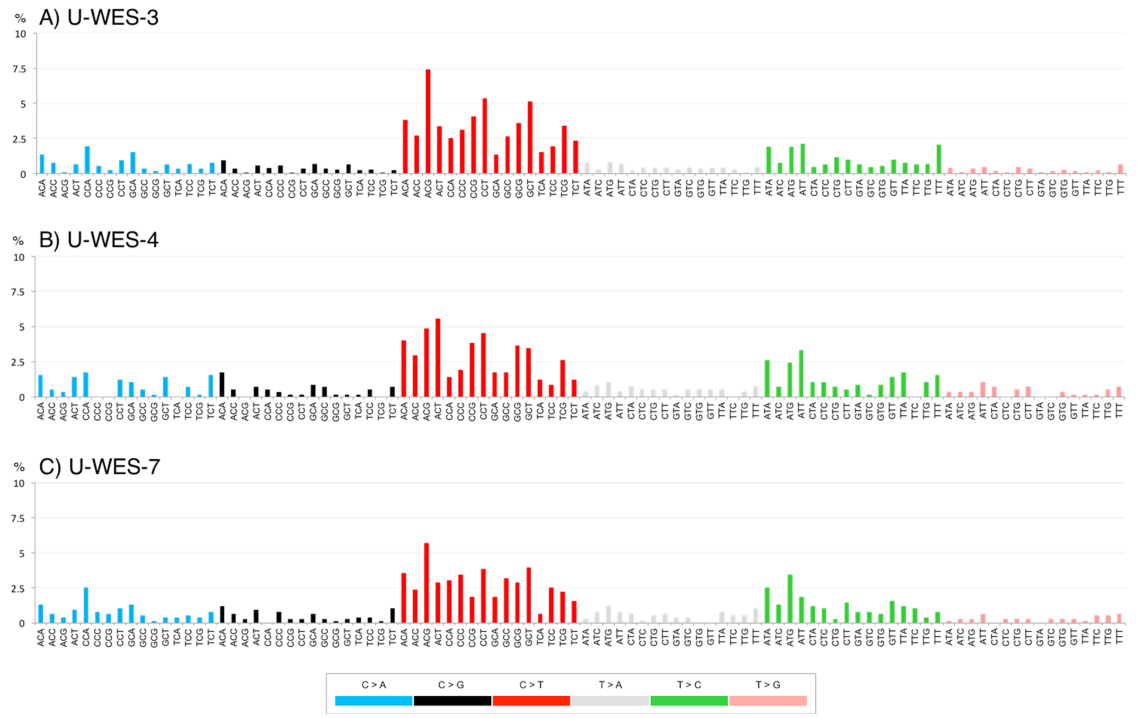
Examples of the copy number analysis using Target sequencing (T-S) and SNP array. Upper and lower figures show the results of T-S and SNP-array, respectively. CN, copy number; AsCN, allele specific copy number. (A) Copy number changes (red circles) in U-WES-3 were identified by both methods. (B) Low-level mosaicism of copy number loss (red circles) in U-WES-5 was not identified by T-S, but SNP array.



Supplemental figure 3.

Mutational spectrums of three patients in PE-group.

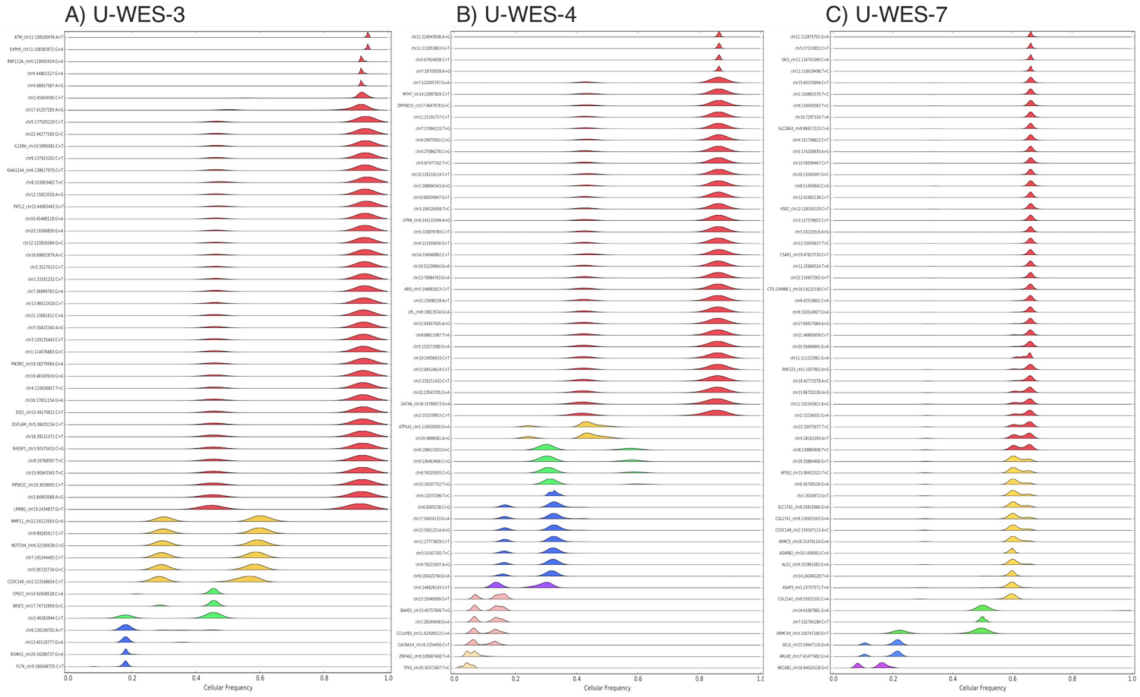
Single nucleotide variants in the whole genome of three patients were classified into 96 classes of substitutions according to the information of the mutated bases (C>A, C>G, C>T, T>A, T>C, and T>G) and the bases immediately 5' and 3' to each mutated base. The mutation types and the percentage of each substitution are shown on horizontal and vertical axis, respectively. (A) U-WES-3 (RAEB-2). (B) U-WES-4 (RAEB-1). (C) U-WES-7 (RAEB-2).



Supplemental figure 4.

Clonal heterogeneities of MDS in three patients of PE-group.

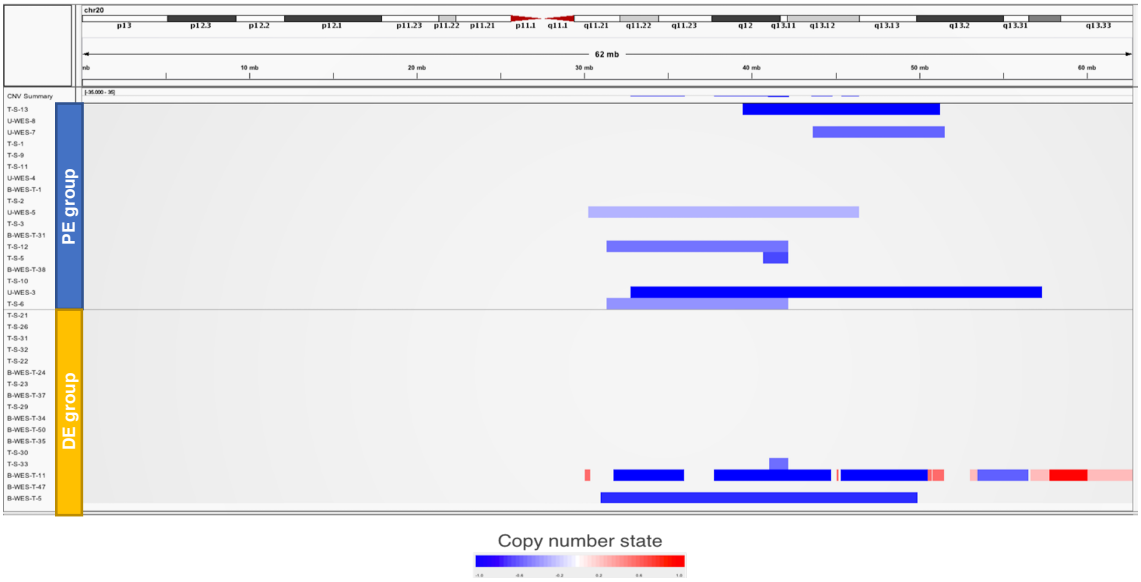
Each color corresponds to each mutational clone. Cellular frequencies of the clones are represented on horizontal axis. Vertical axis shows somatic SNVs used for “PyClone” analysis. (A) U-WES-3 (MDS subtype: RAEB-2). (B) U-WES-4 (RAEB-1). (C) U-WES-7 (RAEB-2).



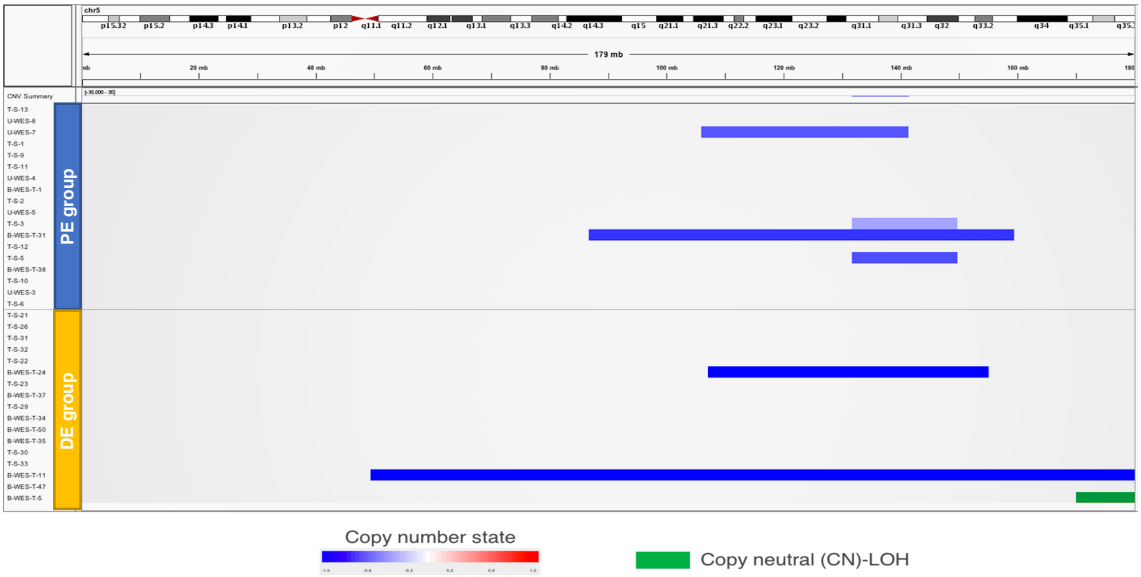
Supplemental figure 5.

Results of copy number alterations.

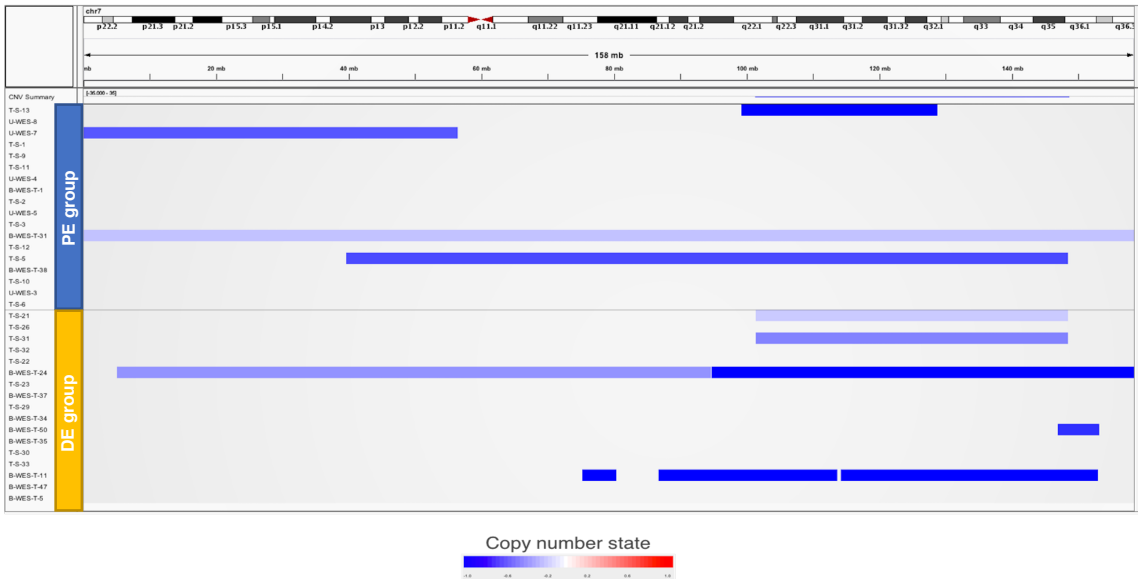
A) Chromosome 20



B) Chromosome 5



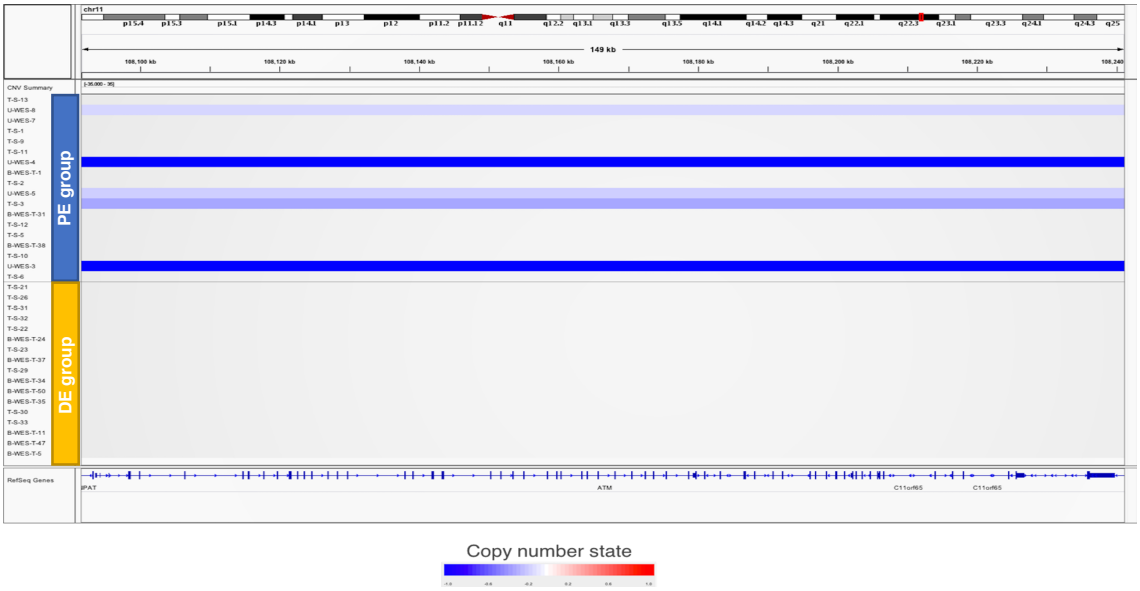
C) Chromosome 7



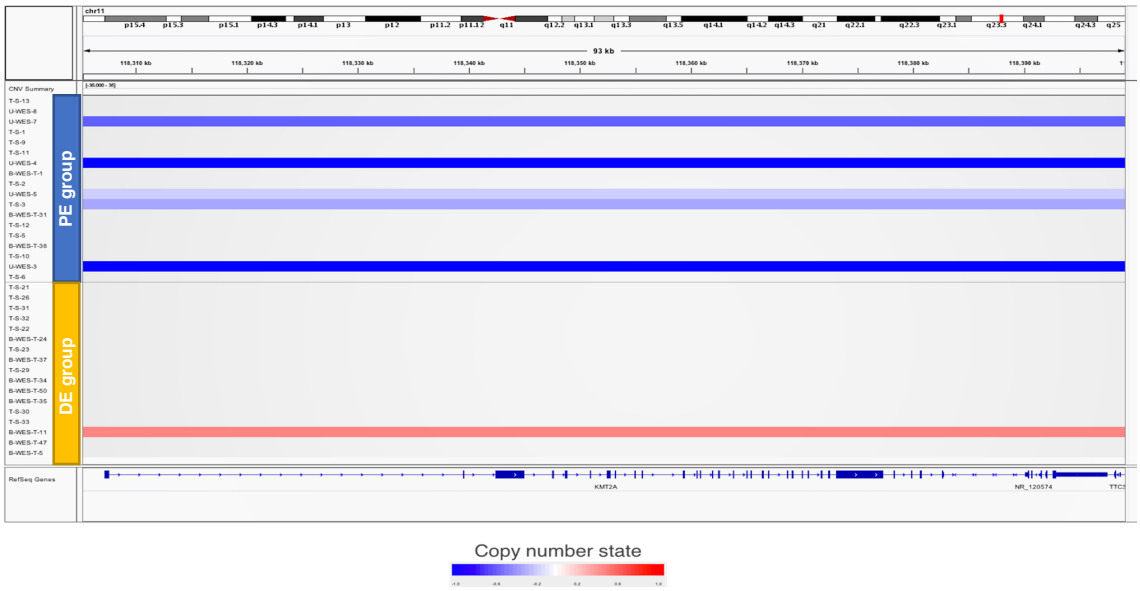
Supplemental figure 6.

Copy number alterations on candidate genes in chromosome 11q.

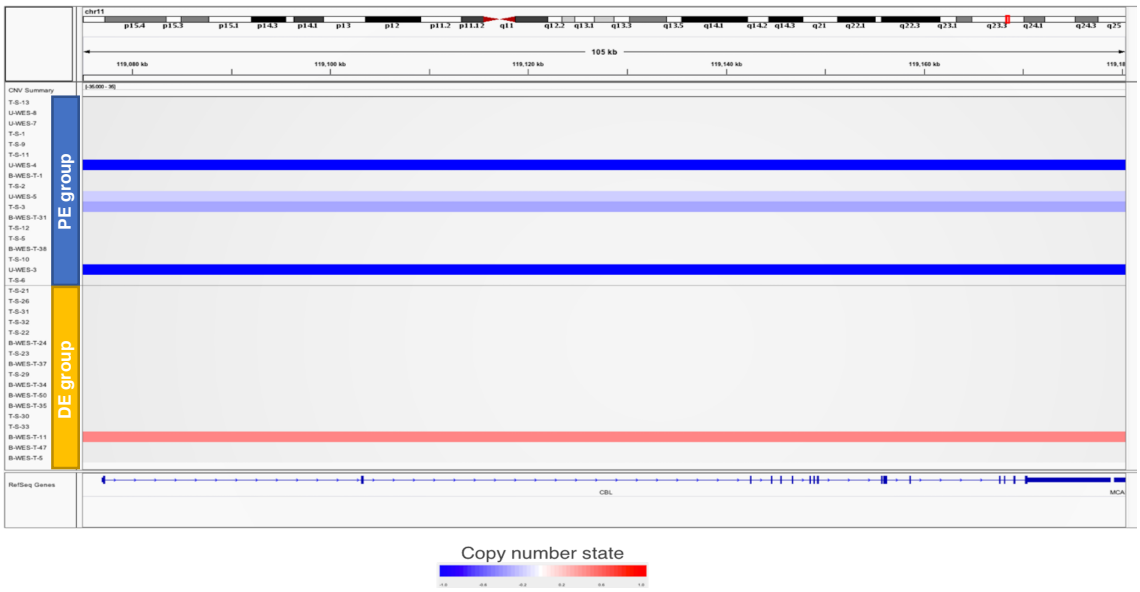
A) *ATM*



B) *KMT2A*



C) *CBL*



Supplemental figure 7.

Overall survival of the patients in proximally and distally exposed group.

Overall survival was analyzed by Kaplan-Meier method and log-rank test. One and two patients in proximally (PE) and distally (DE) exposed group, respectively, were excluded in this analysis because of insufficient information on survival.