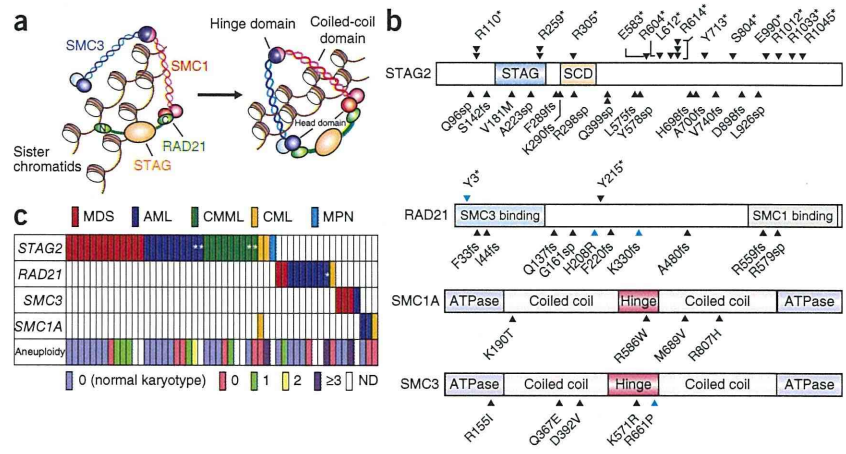


Figure 1 Genetic alterations of the cohesin complex in myeloid neoplasms. (**a**) Cohesin holds chromatin strands within a ring-like structure that is composed of four core components STAG, RAD21, SMC1 and SMC3. (**b**) Mutations in the core components of the cohesin complex found in myeloid malignancies (black arrowheads) and myeloid leukemia-derived cell lines (blue arrowheads). The amino acids in the alterations are referred to using their one-letter abbreviations (for example, R110* represents p.Arg110*). (**c**) Distribution of cohesin mutations and deletions showing a nearly mutually exclusive pattern among different myeloid neoplasms. Gene deletions are indicated by asterisks. The number of numerical chromosome abnormalities in each cohesin-mutated or -deleted case is shown at the bottom. ND, not determined.

and STAG proteins, together with a number of regulatory molecules such as PDS5, NIPBL and ESCO proteins (**Fig. 1a**)[4,5]. Forming a ring-like structure, cohesin is thought to be engaged in the cohesion of sister chromatids during cell division[5], post-replicative DNA repair[6,7] and the regulation of global gene expression through long-range *cis* interactions[8–12]. Germline mutations in cohesin components lead to the congenital multisystem malformation syndromes known as Cornelia de Lange syndrome and Roberts syndrome[13–15].

To investigate a possible role of cohesin mutations in myeloid leukemogenesis, we examined an additional 581 primary specimens of various myeloid neoplasms for mutations in nine cohesin or cohesin-related genes that have been implicated in mitosis[5] using high-throughput sequencing (**Supplementary Table 2**). We also investigated copy-number alterations in cohesin loci in 453 samples using SNP arrays (**Supplementary Table 3**). After excluding known and putative polymorphisms that are registered in the dbSNP or the 1000 Genomes project databases or that were predicted from multiple computational imputations, we identified a total of 60 nonsynonymous mutations involving nine genes in a total of 610 primary samples, which we validated by Sanger sequencing (**Fig. 1b** and **Supplementary Table 4**). After conservative evaluation of the probability of random mutational events across these genes, only four genes remained significantly mutated: *STAG2*, *RAD21*, *SMC1A* and *SMC3* ($P < 0.001$) (**Supplementary Table 5** and Online Methods). In addition, we detected five deletions in *STAG2* ($n = 4$) and *RAD21* ($n = 1$) (**Supplementary Fig. 2a,b** and **Supplementary Table 6**). We also found mutations in these four genes in four of the 34 myeloid leukemia cell lines studied (12%) (**Supplementary Table 7**).

We found mutations and deletions of these four genes in a mostly mutually exclusive manner in a variety of myeloid neoplasms, including acute myeloid leukemia (AML) (19/157), chronic myelomonocytic leukemia (CMML) (9/88), myelodysplastic syndromes (MDS) (18/224) and chronic myelogenous leukemia (CML) (4/64). Mutations were rate in classical myeloproliferative neoplasms (MPN) (1/77) (**Fig. 1c**, **Table 1** and **Supplementary Table 8**). In MDS, mutations were more frequent in refractory cytopenia with multilineage dysplasia and refractory anemia with excess blasts (11.4%) but were rare in refractory anemia, refractory anemia with ring sideroblasts, refractory cytopenia with multilineage dysplasia and ring sideroblasts and MDS with isolated del(5q) (4.2%) ($P = 0.044$). We also evaluated promoter methylation in 33 cases either with ($n = 12$) or without ($n = 21$) cohesin mutations or deletions for which sufficient nonamplified DNA was available using the HumanMethylation450 BeadChip; however, we found no aberrant methylations in cohesin loci, with the exception of hemimethylation of the *SMC1A* promoter that we found in two female cases (**Supplementary Fig. 3**).

We confirmed somatic origins for 17 mutations detected in 16 cases for which matched normal DNA was available (**Supplementary Table 4**). The somatic origins of an additional 23 mutations in *STAG2* or *SMC1A* found in 20 male cases were supported by the presence of reproducible wild-type signals or reads in Sanger and/or deep sequencing of the tumor samples, which were considered to originate from the X chromosome of the residual normal cells (**Supplementary Fig. 4**). In addition, for 20 mutations, the observed allele frequencies determined by pyrosequencing, deep sequencing or digital PCR showed significant deviations from the expected value for polymorphisms in the absence of apparent chromosomal alterations in a SNP array analysis ($P < 0.01$) (**Supplementary Figs. 5** and **6** and **Supplementary Tables 9–12**), suggesting their somatic origins. In addition, 32 of the 33 *STAG2* mutations and all of the nine *RAD21* mutations were either nonsense ($n = 18$), frameshift ($n = 14$) or splice-site ($n = 9$) changes, which were predicted to cause premature truncation of the protein or abnormal exon skipping (**Fig. 1b** and **Supplementary Figs. 7** and **8**). Thus, we considered the majority of the mutations to represent functionally relevant changes, probably of somatic origins (**Supplementary Table 13**).

Most of the cohesin mutations and deletions were heterozygous, except for the *STAG2* and *SMC1A* mutations on the single X chromosome in male cases ($n = 23$). In female samples, the *STAG2* promoter

Table 1 Frequencies of mutations and deletions of cohesin components in 610 myeloid neoplasms

Disease type	*n*	*STAG2*	*RAD21*	*SMC1A*	*SMC3*	Total	Percentage
MDS	224	13	2	0	3	18	8.0
CMML	88	9[a]	0	0	0	9	10.2
AML	157	10	7	2	1	19	12.1
de novo AML	120	8[a]	6	2	1	16	13.3
AML/MRC	37	2[a]	1[a]	0	0	3	8.1
CML	64	2[b]	1	2[b]	0	4	6.3
MPN	77	1	0	0	0	1	1.3
Total	610	35[b]	10	4[b]	4	52	8.5

Diseases are classified according to the World Health Organization 2008 classification. AML/MRC, AML with myelodysplasia-related changes.
[a]Two of the nine cases with *STAG2* alterations in CMML, one of the eight cases with *STAG2* alterations in *de novo* AML, one of the two cases with *STAG2* alterations in AML/MRC cases and one case with *RAD21* alteration in AML/MRC case involved genetic deletions. [b]One CML case having mutations in both *STAG2* and *SMC1A* was counted as a single case. A more detailed list is available in **Supplementary Table 8**.

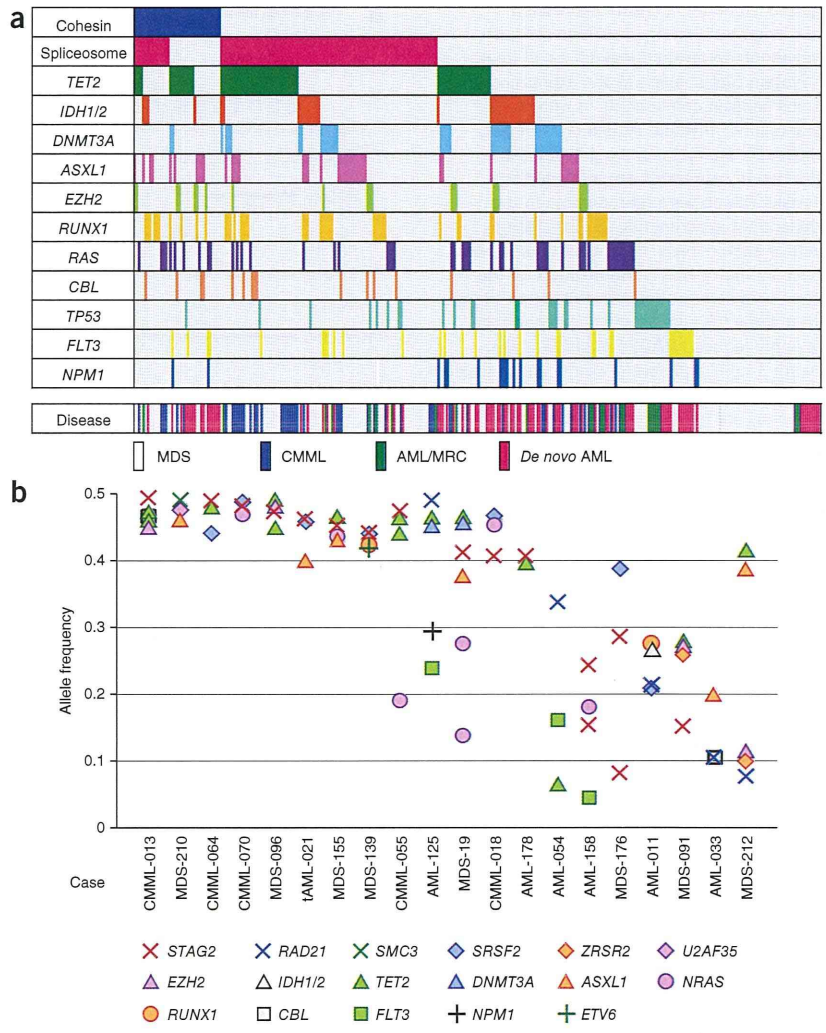


Figure 2 Relationship between cohesin mutations and other common mutations in myeloid malignancies. (**a**) Mutations in the cohesin complex and other common targets in 310 cases with different myeloid neoplasms. The corresponding disease types are shown in the bottom lane. *IDH1/2*, either *IDH1* or *IDH2*. AML/MRC, AML with myelodysplasia-related changes. (**b**) Allele frequencies of mutations in cohesin components and other coexisting mutations in 20 myeloid neoplasms determined by deep sequencing.

was hemimethylated through X inactivation regardless of mutation status (**Supplementary Fig. 3**), and a heterozygous mutation of the unmethylated *STAG2* allele would lead to biallelic *STAG2* inactivation, as has been previously documented in a female case with Ewing's sarcoma[16] and was also confirmed in a single case (CMML-036) in our cohort (**Supplementary Fig. 9**).

Cohesin mutations frequently coexisted with other mutations that are common in myeloid neoplasms and significantly associated with mutations in *TET2* ($P = 0.027$), *ASXL1* ($P = 0.045$) and *EZH2* ($P = 0.011$) (**Fig. 2a**). We performed deep sequencing of the mutant alleles in 20 available samples with cohesin mutations, which allowed for accurate determination of their allele frequencies. The majority of the cohesin mutations (15/20) existed in the major tumor populations, indicating their early origin during leukemogenesis. In the remaining five samples, we found cohesin mutations only in a tumor subpopulation, indicating that the mutations were relatively late events (**Fig. 2b**). Two male cases (MDS-176 and AML-158) harbored two independent subclones with different *STAG2* mutations, indicating that *STAG2* mutation could confer a strong advantage to pre-existing leukemic cells during clonal evolution (**Supplementary Fig. 10**). The number of mutations determined by whole-exome sequencing[3] was significantly higher in four cases with cohesin mutation or deletion compared to cases with no mutation or deletion of cohesin ($P = 0.049$) (**Supplementary Fig. 11**).

Next we investigated the possible impact of mutations on cohesin function. We examined the expression of STAG1, STAG2, RAD21, SMC3, SMC1A and NIPBL in 17 myeloid leukemia cell lines with ($n = 4$) or without ($n = 13$) known cohesin mutations, as well as in the chromatin-bound fractions of 13 cell lines (**Fig. 3a–d** and **Supplementary Table 14**)[14,17–19]. Although we observed an evaluable reduction in RAD21 expression in Kasumi-1 cells that harbored a frameshift alteration in RAD21 (p.Lys330ProfsX6) (**Fig. 3a**), alterations in P31FUJ (RAD21 p.His208Arg), CMY (RAD21 p.Tyr3X) and MOLM-7 (SMC3 p.Arg661Pro) cells were not accompanied by measurable decreases in the corresponding mutated proteins compared to wild-type cell lines. In contrast, we observed severely reduced expression of one or more cohesin components in KG-1 (STAG2)[16] and MOLM-13 (STAG1, STAG2, RAD21 and NIPBL) cells without any accompanying mutations in the relevant genes (**Fig. 3a**). We found no significant differences in protein expression of the cohesin components in cohesin-mutated and non-mutated cell lines in whole-cell extracts (**Fig. 3b**). However, expression of one or more cohesin components, including SMC1, SMC3, RAD21 and STAG2, was significantly reduced in the chromatin-bound fractions of cell lines with mutated or reduced expression of cohesin components, including Kasumi-1, KG-1, P31FUJ, MOLM-7 and MOLM-13 cells, compared with the cell lines with no known cohesin mutations or abnormal cohesin expression ($P < 0.05$), suggesting a substantial loss of cohesin-bound sites on chromatin (**Fig. 3c,d** and **Supplementary Table 14**)[14].

We next examined the effect of forced expression of wild-type cohesin components on the proliferation of a cohesin-mutated cell line (Kasumi-1) or a cell line with reduced expression of cohesin components (MOLM-13). Forced expression of wild-type *RAD21* and/or *STAG2*, but not of a truncated *RAD21* allele, induced significant growth suppression of the Kasumi-1 (with mutated *RAD21*) and MOLM-13 (with severe reduction of RAD21 and STAG2 expression) cell lines but not the K562 and TF1 (with wild-type *RAD21*) cell lines, supporting a leukemogenic role for compromised cohesin functions (**Fig. 4a–c** and **Supplementary Fig. 12a–g**). To explore the effect of forced expression of RAD21 on global gene expression, we performed expression microarray analysis of *RAD21*- and mock-transduced Kasumi-1 cells. In agreement with previous experiments with other cohesin and cohesin-related components, the magnitudes of the

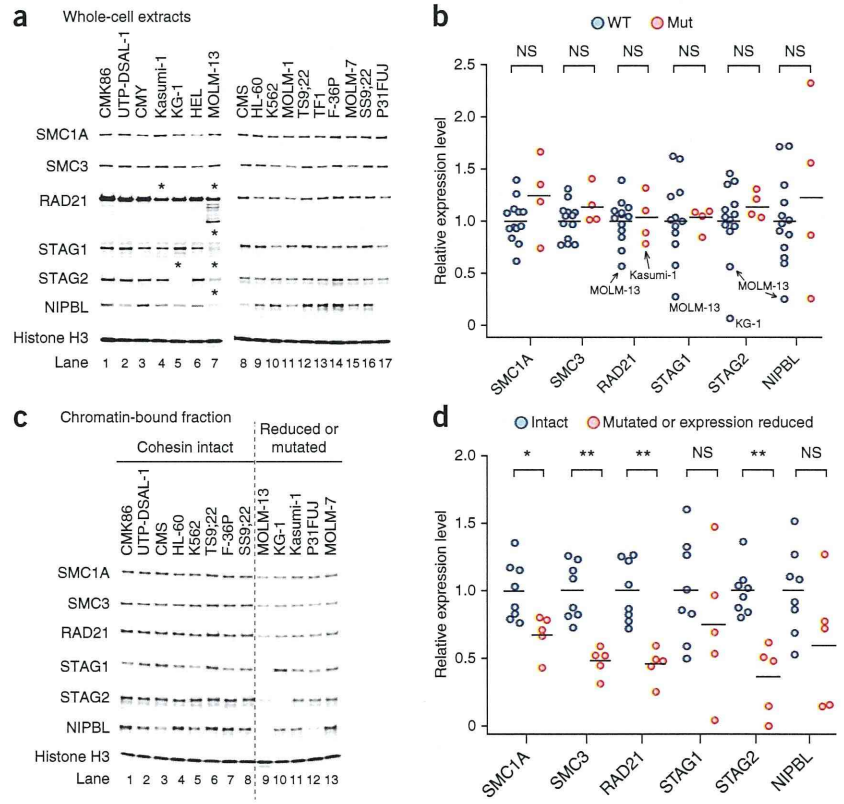
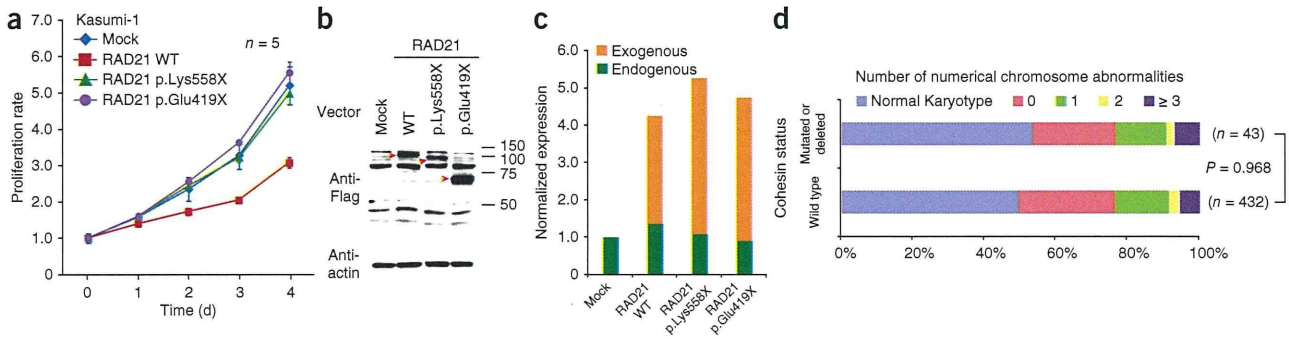


Figure 3 Abnormal cohesin expression and chromatin binding of various cohesin components in myeloid leukemic cell lines. (**a**) Protein blot analysis of the expression of various cohesin components in whole-cell extracts in 17 myeloid leukemia cell lines. Cohesin components showing evaluable reduction in expression are indicated by asterisks, which were reproducible in two independent experiments. (**b**) Expression levels of each cohesin component measured by densitometry after normalization for the mean value across all non-mutated cell lines, with histone H3 signals serving as controls. Evaluably reduced RAD21 expression in Kasumi-1 cells and severely reduced expression of cohesin components in MOLM-13 and KG-1 cells are indicated within the plots. No significant differences (NS) in the expression of the cohesin components were observed between cohesin-mutated and non-mutated cell lines (Mann-Whitney *U* test). Each circle represents a single cell line. (**c**) Protein blot analysis of cohesin components in the chromatin-bound fractions of 13 myeloid leukemia cell lines having intact cohesin (lanes 1–8), cohesin mutations and/or reduced expression of cohesin in whole-cell extracts (lanes 9–13). A representative result of two independent experiments reproducibly showing reduced chromatin-bound cohesin fractions in the cell lines in lanes 9–13 is presented. (**d**) Expression levels of cohesin components in the chromatin-bound fractions measured by densitometry after normalization for the mean value across cell lines with intact cohesin components, with histone H3 signals serving as controls. $*P < 0.05$, $**P < 0.005$ (Mann-Whitney *U* test). Horizontal bars in **b** and **d** indicate the mean values. The densitometric data are presented in **Supplementary Table 14**.

transcriptional changes induced by forced RAD21 expression were generally small[14,16,20]. However, 63 genes reproducibly and significantly showed a more than 1.2-fold increase ($n = 35$) or decrease ($n = 28$) in gene expression ($P < 0.05$), which was validated by quantitative PCR and/or RNA sequencing for 59 of the 63 genes (**Supplementary Fig. 13a–c** and **Supplementary Tables 15** and **16**).



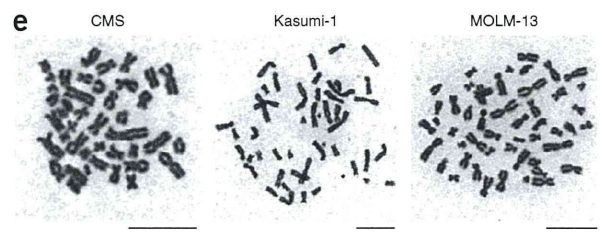


Figure 4 Impact of cohesin mutations on cell proliferation and karyotypes. (**a**) Proliferation of the Kasumi-1 cell line stably transduced with either wild-type RAD21, a truncated allele of *RAD21* (RAD21 p.Lys558X or p.Glu419X) or a mock construct measured by MTT assays ($n = 5$ wells per group). The data are shown as the means ± s.d. of the absorbance at 450 nm relative to the value at day 0. Representative results of three independent experiments are shown. (**b**) Protein blot analysis showing expression of the transduced wild-type and mutant RAD21 alleles. (**c**) Expression of endogenous and exogenous *RAD21* transcripts in Kasumi-1 cells transduced with indicated constructs measured using RNA sequencing by enumerating the corresponding reads. (**d**) The numbers of cases with numerical cytogenetic abnormalities were compared between two groups, those with and those without cohesin mutations or deletions ($P = 0.968$, χ^2). The numbers of numerical chromosome abnormalities are shown at the top. (**e**) Representative metaphases of cell lines with intact (CMS) or abnormal (Kasumi-1 and MOLM-13) cohesin components showing almost normal sister chromatid cohesion. Scale bars, 10 μm.

Mutations in the cohesin complex have recently been reported in a cohort of *de novo* AML and MDS in which four major cohesin components were mutated in 6.0–13.0% of cases[21–25]. Less frequent mutations of cohesin components have been described in other cancers, including *STAG2* mutations in glioblastoma (4/68), melanoma (1/48) and Ewing's sarcoma (1/24)[16]. In primary colon cancer samples, in which impaired cohesion and consequent aneuploidy have been implicated in oncogenesis, mutations in *SMC1A* (4/132), *NIPBL* (4/132), *STAG3* (1/130) and *SMC3* (1/130) have been reported[26]. In contrast, in our cohort of myeloid neoplasms, we found no significant differences in the number of numerical chromosome abnormalities between cohesin-mutated and non-mutated cases, and the 43 cases with cohesin mutations or deletions showed diploid or near-diploid karyotypes, including 23 cases with completely normal karyotypes (**Fig. 4d**). Therefore, in these euploid cases, cohesin-mutated cells were not clonally selected as a result of aneuploidy. Supporting this finding is the observation that expression of *scc1p*, a *RAD21* homolog, at only 13% of its normal level was sufficient for normal cohesion in yeast[27]. Furthermore, Kasumi-1 and MOLM-13 cells showed almost normal cohesion of sister chromatids, even though Kasumi-1 cells have a truncated *RAD21* allele and MOLM-13 cells have substantially reduced expression of multiple cohesin components (**Fig. 4e**).

A growing body of evidence has suggested that cohesin mediates long-range chromosomal *cis* interactions[28] and regulates global gene expression[11,12]. For example, two cohesin subunits, Rad21 and Smc3, have been implicated in the transcriptional regulation of the hematopoietic transcription factor Runx1 in zebrafish[10]. Furthermore, an up to 80% downregulation of *Nipped-B*, a *NIPBL* homolog in *Drosophila*, does not affect chromosomal segregation but does cause impaired regulation of gene expression[20]. We also previously demonstrated that only mild loss (17–28%) of cohesin binding sites within the genome results in deregulated global gene expression[14,18,19]. These observations suggest the possibility that cohesin mutations participate in leukemogenesis through the deregulated expression of genes that are involved in myeloid development and differentiation.

In conclusion, we report frequent mutations in cohesin components that involve a wide variety of myeloid neoplasms. Genetic evidence suggests that aneuploidy may not be the only leukemogenic mechanism, at least *in vivo*, and that deregulated gene expression and/or other mechanisms, such as DNA hypermutability, might also operate in leukemogenesis. Given the integral functions of cohesin for cell viability, genetic defects in cohesin might be potential targets in myeloid neoplasms[14,29].

URLs. dbSNP, http://www.ncbi.nlm.nih.gov/projects/SNP/; the 1000 Genomes Project, http://www.1000genomes.org/; the UCSC Genome Browser; http://genome.ucsc.edu/cgi-bin/hgGateway/; hg19, http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/; RefSeq genes, http://www.ncbi.nlm.nih.gov/RefSeq/; CNAG/AsCNAR, http://www.genome.umin.jp/; dChip, http://www.dchip.org/; the Integrative Genomics Viewer, http://www.broadinstitute.org/igv/; SIFT, http://sift.jcvi.org/; PolyPhen-2, http://genetics.bwh.harvard.edu/pph2/; Mutation Taster, http://www.mutationtaster.org/.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Whole-exome sequence data have been deposited in the DNA Data Bank of Japan (DDBJ) repository under accession number DRA000433. RNA sequencing data have been deposited in the DDBJ repository under accession number DRA001013. Microarray data have been deposited in the Gene Expression Omnibus under accession number GSE47684.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

This work was supported by Grants-in-Aid from the Ministry of Health, Labor and Welfare of Japan and KAKENHI (23249052, 22134006 and 21790907; S.O.), the Industrial Technology Research Grant Program from the New Energy and Industrial Technology Development Organization (NEDO; S.O.) (08C46598a), NHRI-EX100-10003NI Taiwan (L.-Y.S.), the project for development of innovative research on cancer therapies (p-direct; S.O.) and the Japan Society for the Promotion of Science through the Funding Program for World-Leading Innovative R&D on Science and Technology, initiated by the Council for Science and Technology Policy (CSTP; S.O.). We thank Y. Hayashi (Gunma Children's Medical Centre), R.C. Mulligan (Harvard Medical School), S. Sugano (The University of Tokyo), M. Onodera (National Center for Child Health and Development, Japan) and L. Ström (Karolinska Institute) for providing materials. We thank Y. Yamazaki for cell sorting. We also thank Y. Mori, M. Nakamura, N. Mizota and S. Ichimura for their technical assistance and M. Ueda for encouragement.

AUTHOR CONTRIBUTIONS

A.K., Y.N., K.Y., A.S.-O., Y. Sato and M.S. processed and analyzed genetic materials and performed sequencing and SNP array analysis. Y. Shiraishi, Y.O., R.N., A.S.-O., H.T., T.S., K.C., M.N. and S. Miyano performed bioinformatics analyses of the sequencing data. L.-Y.S. performed pyrosequencing analysis, and A.N. and S.I. performed digital PCR. G.N. and H.A. performed methylation analysis. M.M., M.B. and K.S. performed studies on protein expression of cohesin components. A.K., M.S., T.Y., R.Y., M.O. and H.N. were involved in the functional studies. A.K. and A.S.-O. performed expression microarray experiments and their analyses. L.-Y.S., D.N., T.A., C.H., F.N., W.-K.H., T.H., H.P.K., T.N., H.M., S. Miyawaki, M.S.-Y., K.I., N.O. and S.C. collected specimens and were involved in project planning. A.K., L.-Y.S., M.M., A.S.-O. and S.O. generated figures and tables. S.O. led the entire project, and A.K. and S.O. wrote the manuscript. All authors participated in the discussion and interpretation of the data.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Bejar, R., Levine, R. & Ebert, B.L. Unraveling the molecular pathophysiology of myelodysplastic syndromes. *J. Clin. Oncol.* **29**, 504–515 (2011).
2. Marcucci, G., Haferlach, T. & Dohner, H. Molecular genetics of adult acute myeloid leukemia: prognostic and therapeutic implications. *J. Clin. Oncol.* **29**, 475–486 (2011).
3. Yoshida, K. *et al.* Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature* **478**, 64–69 (2011).
4. Gruber, S., Haering, C.H. & Nasmyth, K. Chromosomal cohesin forms a ring. *Cell* **112**, 765–777 (2003).
5. Nasmyth, K. & Haering, C.H. Cohesin: its roles and mechanisms. *Annu. Rev. Genet.* **43**, 525–558 (2009).
6. Ström, L. *et al.* Postreplicative formation of cohesion is required for repair and induced by a single DNA break. *Science* **317**, 242–245 (2007).
7. Watrin, E. & Peters, J.M. The cohesin complex is required for the DNA damage-induced G2/M checkpoint in mammalian cells. *EMBO J.* **28**, 2625–2635 (2009).
8. Dorsett, D. Cohesin, gene expression and development: lessons from *Drosophila*. *Chromosome Res.* **17**, 185–200 (2009).
9. Dorsett, D. *et al.* Effects of sister chromatid cohesion proteins on cut gene expression during wing development in *Drosophila*. *Development* **132**, 4743–4753 (2005).
10. Horsfield, J.A. *et al.* Cohesin-dependent regulation of Runx genes. *Development* **134**, 2639–2649 (2007).
11. Parelho, V. *et al.* Cohesins functionally associate with CTCF on mammalian chromosome arms. *Cell* **132**, 422–433 (2008).
12. Wendt, K.S. *et al.* Cohesin mediates transcriptional insulation by CCCTC-binding factor. *Nature* **451**, 796–801 (2008).
13. Bose, T. & Gerton, J.L. Cohesinopathies, gene expression, and chromatin organization. *J. Cell Biol.* **189**, 201–210 (2010).
14. Deardorff, M.A. *et al.* HDAC8 mutations in Cornelia de Lange syndrome affect the cohesin acetylation cycle. *Nature* **489**, 313–317 (2012).
15. Deardorff, M.A. *et al.* RAD21 mutations cause a human cohesinopathy. *Am. J. Hum. Genet.* **90**, 1014–1027 (2012).

16. Solomon, D.A. *et al.* Mutational inactivation of STAG2 causes aneuploidy in human cancer. *Science* **333**, 1039–1043 (2011).
17. Beckouët, F. *et al.* An Smc3 acetylation cycle is essential for establishment of sister chromatid cohesion. *Mol. Cell* **39**, 689–699 (2010).
18. Liu, J. *et al.* Transcriptional dysregulation in NIPBL and cohesin mutant human cells. *PLoS Biol.* **7**, e1000119 (2009).
19. Liu, J. *et al.* Genome-wide DNA methylation analysis in cohesin mutant human cell lines. *Nucleic Acids Res.* **38**, 5657–5671 (2010).
20. Schaaf, C.A. *et al.* Regulation of the *Drosophila* enhancer of split and invected-engrailed gene complexes by sister chromatid cohesion proteins. *PLoS ONE* **4**, e6202 (2009).
21. Ding, L. *et al.* Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature* **481**, 506–510 (2012).
22. Walter, M.J. *et al.* Clonal architecture of secondary acute myeloid leukemia. *N. Engl. J. Med.* **366**, 1090–1098 (2012).
23. Welch, J.S. *et al.* The origin and evolution of mutations in acute myeloid leukemia. *Cell* **150**, 264–278 (2012).
24. The Cancer Genome Atlas Research Network. Genomic and epigenomic landscapes of adult *de novo* acute myeloid leukemia. *N. Engl. J. Med.* **368**, 2059–2074 (2013).
25. Walter, M.J. *et al.* Clonal diversity of recurrently mutated genes in myelodysplastic syndromes. *Leukemia* **27**, 12785–1282 (2013).
26. Barber, T.D. *et al.* Chromatid cohesion defects may underlie chromosome instability in human colorectal cancers. *Proc. Natl. Acad. Sci. USA* **105**, 3443–3448 (2008).
27. Heidinger-Pauli, J.M., Mert, O., Davenport, C., Guacci, V. & Koshland, D. Systematic reduction of cohesin differentially affects chromosome segregation, condensation, and DNA repair. *Curr. Biol.* **20**, 957–963 (2010).
28. Hadjur, S. *et al.* Cohesins form chromosomal *cis*-interactions at the developmentally regulated IFNG locus. *Nature* **460**, 410–413 (2009).
29. Chan, D.A. & Giaccia, A.J. Harnessing synthetic lethal interactions in anticancer drug discovery. *Nat. Rev. Drug Discov.* **10**, 351–364 (2011).

ONLINE METHODS

Patients and samples. Twenty-nine cases analyzed by whole-exome sequencing were described previously[3]. Anonymized genomic DNA from an additional 581 patients with different myeloid neoplasms were collected from collaborating institutes and used for the analyses described below. All the analyses were performed after written informed consent was obtained. This study was approved by the ethics boards of the University of Tokyo, University Hospital Mannheim, University of Tsukuba, the Munich Leukemia Laboratory, Showa University, Tokyo Metropolitan Ohtsuka Hospital and Chang Gung Memorial Hospital.

Cell lines. The CMS, CMY, UTP-DSAL-1, MOLM-1, MOLM-7, HEL, SS9;22 and TS9;22 cell lines were provided by Y. Hayashi. 293gp and 293gpg cells were provided by R.C. Mulligan. P31FUJ and CMK-86 cells were purchased from the Health Science Research Resources Bank (Osaka, Japan). 293T, KG-1, K562 and F-36P cells were obtained from RIKEN BioResource Center Cell Bank (Tsukuba, Japan), and Kasumi-1, HL-60, MOLM-13 and TF-1 cells were from the American Type Culture Collection. Chromosome spreads were performed for the CMS, Kasumi-1 and MOLM-13 cell lines as previously described[14], except that cells were treated with colcemid (100 μg/ml) and hypotonically swollen in 75 mM KCl for 20 min.

Whole-exome sequencing. The whole-exome sequencing of the 29 paired samples of myelodysplasia was previously described[3], through which we identified a total of 497 candidate single-nucleotide variants and insertions/deletions (indels), of which 268 and 167 were determined by Sanger sequencing as true positives and negatives, respectively, with 62 mutations unconfirmed. In the present study, we updated the list of somatic mutations by rigorously validating the remaining 62 unconfirmed mutations by Sanger sequencing and also by deep sequencing (**Supplementary Table 1**).

Mutation analysis of cohesin components. In total, 534 tumor DNA samples from a variety of myeloid neoplasms were analyzed for possible mutations in nine components of the cohesin complex, *STAG1*, *STAG2*, *SMC1A*, *SMC3*, *RAD21*, *PDS5B*, *ESCO1*, *ESCO2* and *NIPBL*, using high-throughput sequencing of pooled exons amplified from pooled genomic DNA samples. In an additional 47 samples, mutations in *STAG2*, *RAD21*, *SMC1A* and *SMC3* were examined by deep sequencing after enrichment for these targets using a SureSelect custom kit (Agilent) designed to capture all of the coding exons from the target genes, performed as previously described with minor modifications in the algorithm for mutation call[30].

For pooled-DNA sequencing, all target exons (n = 232) encompassing 89,323 nucleotides were PCR amplified using a set of primers having common NotI adaptor sequences on their 5′ ends, digested with NotI, ligated using T4 ligase and sonicated to approximately 200-bp fragments using an ultrasonicator (Covaris); these fragments were used for the generation of sequencing libraries according to a modified pair-end protocol from Illumina. The libraries were then sequenced using HiSeq 2000 (Illumina) with a standard 100-bp paired end–reads protocol. On average, 99.5% of the target bases were analyzed at the depth of 12,000 per pool or 1,000 per sample. Data processing and variant calling were performed as previously described[3] with minor modifications. First, each read from a given DNA pool was aligned to the set of target sequences using BLAT[31] with the -fine option. The mapping information in a .psl format was transformed into a .sam format using the my_psl2sam script, which was further converted into the .bam format using SAMtools[32]. Among the successfully mapped reads, reads were removed from further analysis that either mapped to multiple sites, mapped with more than four mismatched bases or had more than ten clipped bases. Next, the Estimation_CRME script was run to eliminate strand-specific errors and exclude PCR-derived errors. Then, a strand-specific mismatch ratio was calculated for each nucleotide variation for both strands using the bases corresponding to 11–50 cycles. By excluding the top five cycles showing the highest mismatch rates, strand-specific mismatch rates were recalculated, and the smaller value between both strands was adopted as the nominal mismatch ratio. In addition, the nucleotide variations that were present across multiple pools were removed based on permutations across different pools using the Permut_Rm_com script because it is probable that such variations result from systemic sequencing errors. Finally, after excluding variations found in the dbSNP database, the database from the 1000 Genomes project or our in-house SNP database, the variants whose mismatch rate exceeded 0.009 were adopted as candidate mutations. Each candidate mutation was validated by Sanger sequencing of the 12 original individual DNAs from the corresponding DNA pools.

The functional impact of each amino acid substitution was evaluated by computer prediction using SIFT[33], PolyPhen-2 (ref. 34) and Mutation Taster[35]. The significance of nonsilent mutations in each cohesin component was evaluated assuming a uniform distribution of the background mutations within the coding regions, which was estimated to be ~0.3 Mb^{-1} on the basis of a previous whole-exome sequencing of myelodysplasia[3].

Determination of variant allele frequencies. Variant allele frequencies were evaluated by deep sequencing of PCR amplicons, pyrosequencing[36,37] and/or digital PCR (Fluidigm CA, US)[38–40] of the variants using nonamplified DNA. For amplicon sequencing, genomic fragments harboring the variants of interest were PCR amplified using Not1-tagged primers. Ninety-two randomly selected SNP loci that do not contain repetitive sequences were amplified using normal genomic DNA as a template, which served as the control. Touch-down PCRs using high-fidelity DNA polymerase KOD-Plus-Neo (TOYOBO, Tokyo) were performed, and an equimolar mixture of all PCR products was prepared for deep sequencing using HiSeq2000 or Miseq (Illumina), as described above, with a 75-bp or 100-bp pair end–read option. To calculate the allele frequency of each variant, all reads were mapped to the target reference sequence using BLAT[31], followed by differential enumeration of the dichotomic variant alleles. For indels, individual reads were first aligned to each of the wild-type and altered sequences and then assigned to the one with better alignment in terms of the number of matched bases.

Array-based copy-number and methylation analyses. Genomic DNA from 453 bone marrow samples with myeloid neoplasms was analyzed using GeneChip SNP genotyping microarrays as previously described using CNAG/AsCNAR software[41,42]. The results of the SNP array karyotyping for 290 of the 453 cases have been previously published[3,41–44]. The promoter methylation of each cohesin component gene was analyzed using the HumanMethylation450 BeadChip (Illumina), as previously described[30,45], in which methylation status was evaluated by calculating the ratio of methylation-specific and demethylation-specific fluorophores (β value) at each CpG site using iScan software (Illumina).

RT-PCR. Complementary DNA synthesis and quantitative RT-PCR analyses were performed as previously described[3]. The primer sequences used are listed in **Supplementary Tables 16** and **17**.

Protein expression of cohesin components in whole-cell extracts and chromatin-enriched fractions. Whole-cell extracts of myeloid cell lines were separated into soluble supernatant and chromatin-containing pellet fractions and analyzed by SDS-PAGE and protein blot analysis for the expression of different cohesin components as previously described[12,14]. Antibodies used for protein blot analysis are described in **Supplementary Table 18**.

Gene expression and cell proliferation assays. A full-length *RAD21* cDNA (BC050381) was provided by S. Sugano. A full-length *STAG2* cDNA was obtained from total cDNA derived from bone marrow cells and cloned into pBluescript. The truncated mutant of *RAD21* was subcloned by PCR. Flag-tagged *RAD21* or *STAG2* cDNAs were constructed into the retrovirus vector pGCDNsamIRESEGFP (provided by M. Onodera)[46] or a tetracycline-inducible lentiviral vector, CS-TRE-Ubc-tTA-IRESPuro. The wild-type *RAD21*, the mutant *RAD21* and/or a mock-induced retroviral vector were generated as previously described[3] and transduced into Kasumi-1, K562 and TF1 cells, which were sorted by GFP marking using a MoFlo FACS cell sorter (Beckman Coulter) or a BD FACSAria cell sorter (BD Biosciences) 48–96 h after retroviral transduction. The wild-type *RAD21*, the wild-type *STAG2* and a mock-induced lentiviral vector were generated as described previously[47], transduced into MOLM-13 cells and selected by 1 μg/ml puromycin. Gene expression was induced by 1 μg/ml doxycycline. For cell growth assays, the cells were inoculated into 96-well culture plates in RPMI 1640 medium supplemented

doi:10.1038/ng.2731

with 5% FCS (and 5 ng/ml GM-CSF for TF1 cells), and cell growth was monitored in three independent experiments by MTT assay using the Cell Counting Kit-8 (Dojindo Co.).

Expression microarray analysis. RNA was extracted from Kasumi-1 cells that were either mock transduced or transduced with wild-type RAD21 and analyzed in triplicate using the Human Genome U133 Plus 2.0 Array (Affymetrix) according to the manufacturer's protocol. For data analysis, raw array signals were first extracted from .CEL files using dChip Software[48]. After background correction and normalization across the six array data sets, the standardized signal value was obtained for each probe set in each of triplicate array experiments, which were compared between mock-transduced and wild-type *RAD21*–transduced cells. Two independent microarray experiments were performed. To identify transcriptionally altered genes, we used the criteria of fold change greater than ±1.2 and $P < 0.05$ (two-tailed paired *t* test) in two independent experiments.

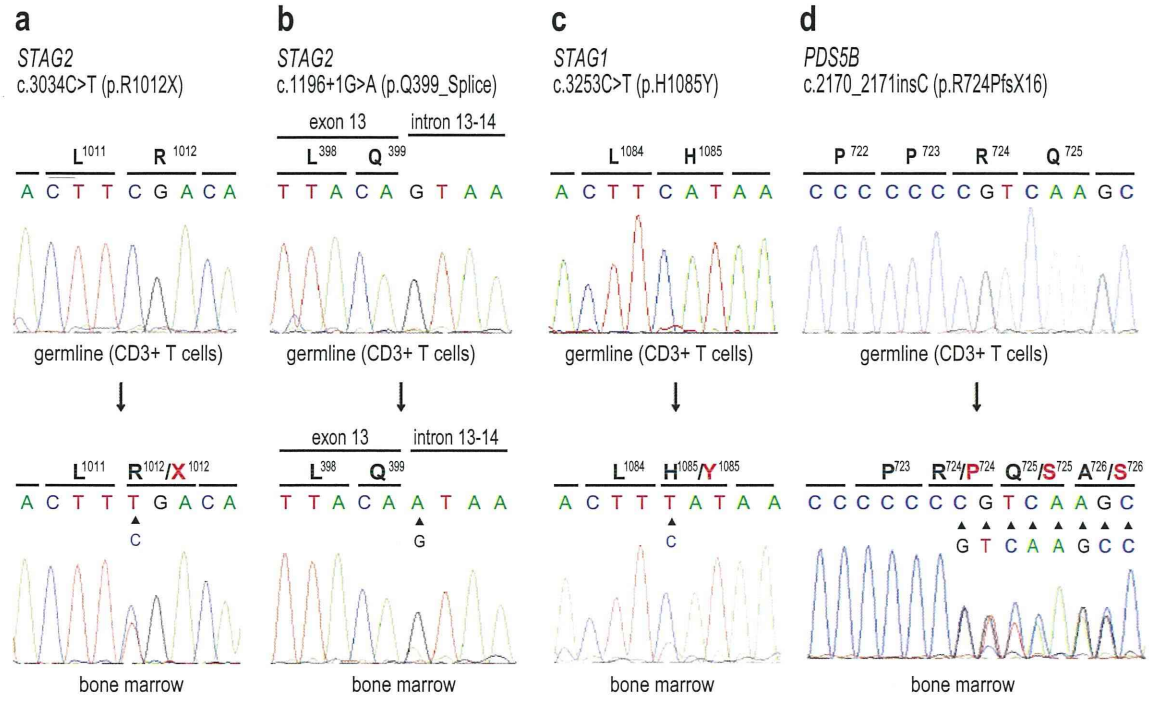
RNA sequencing. RNA sequencing of *RAD21*-transduced Kasumi-1 cells and subsequent data analyses were performed as previously described[3] with minor modifications. For quantifications of expression values from the RNA sequencing data, we used a slightly modified version of RKPM (reads per kb of exon per million mapped reads) measures[49]. After removing the sequencing reads that were inappropriately aligned or that had low mapping quality, the number of bases on each exonic region for each RefSeq gene[50] was counted. Then the number of bases was normalized per kb of exon and per 100 million aligned bases. Finally, the expression value of each gene was determined by taking the maximum values among the RefSeq genes corresponding to the gene symbol.

We measured RAD21 expression by differentially enumerating endogenous and exogenous *RAD21* sequence reads, which were discriminated by the absence and presence of the Flag sequence, respectively. After normalization by the number of total reads for each sample, the raw differential read counts were further calibrated against the read counts containing the stop codon in *RAD21*.

Statistical analyses. The significance of the difference in frequency of cohesin component mutations between disease subtypes was tested by one-tailed Fisher's exact test. The coexistence of mutations was tested by two-tailed Fisher's direct method. The significance of the difference in the total number of somatic mutations between cohesin-mutated or -deleted and non-mutated or -deleted samples was tested by Mann-Whitney *U* test. Differences in the number of numerical abnormalities in cytogenetics between two groups with and without cohesin mutations or deletions was assessed by one-sided χ^2 test.

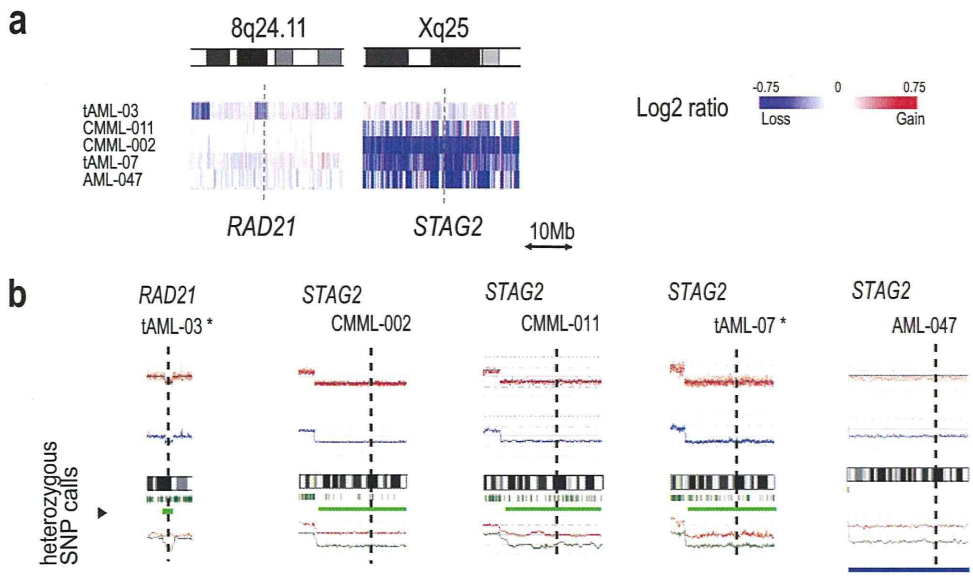
30. Sato, Y. *et al.* Integrated molecular analysis of clear-cell renal cell carcinoma. *Nat. Genet.* doi:10.1038/ng.2699 (24 June 2013).
31. Kent, W.J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
32. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
33. Kumar, P., Henikoff, S. & Ng, P.C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4**, 1073–1081 (2009).
34. Adzhubei, I.A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
35. Schwarz, J.M., Rodelsperger, C., Schuelke, M. & Seelow, D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat. Methods* **7**, 575–576 (2010).
36. Ronaghi, M. Pyrosequencing sheds light on DNA sequencing. *Genome Res.* **11**, 3–11 (2001).
37. Shih, L.Y. *et al.* Emerging kinetics of BCR-ABL1 mutations and their effect on disease outcomes in chronic myeloid leukemia patients with imatinib failure. *Leuk. Res.* **37**, 43–49 (2013).
38. Qin, J., Jones, R.C. & Ramakrishnan, R. Studying copy number variations using a nanofluidic platform. *Nucleic Acids Res.* **36**, e116 (2008).
39. Dube, S., Qin, J. & Ramakrishnan, R. Mathematical analysis of copy number variation in a DNA sample using digital PCR on a nanofluidic device. *PLoS ONE* **3**, e2876 (2008).
40. Totoki, Y. *et al.* High-resolution characterization of a hepatocellular carcinoma genome. *Nat. Genet.* **43**, 464–469 (2011).
41. Nannya, Y. *et al.* A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. *Cancer Res.* **65**, 6071–6079 (2005).
42. Yamamoto, G. *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.* **81**, 114–126 (2007).
43. Hosoya, N. *et al.* Genomewide screening of DNA copy number changes in chronic myelogenous leukemia with the use of high-resolution array-based comparative genomic hybridization. *Genes Chromosom. Cancer* **45**, 482–494 (2006).
44. Sanada, M. *et al.* Gain-of-function of mutated C-CBL tumour suppressor in myeloid neoplasms. *Nature* **460**, 904–908 (2009).
45. Nagae, G. *et al.* Tissue-specific demethylation in CpG-poor promoters during cellular differentiation. *Hum. Mol. Genet.* **20**, 2710–2721 (2011).
46. Nabekura, T., Otsu, M., Nagasawa, T., Nakauchi, H. & Onodera, M. Potent vaccine therapy with dendritic cells genetically modified by the gene-silencing–resistant retroviral vector GCDNsap. *Mol. Ther.* **13**, 301–309 (2006).
47. Agarwal, S. *et al.* Isolation, characterization, and genetic complementation of a cellular mutant resistant to retroviral infection. *Proc. Natl. Acad. Sci. USA* **103**, 15933–15938 (2006).
48. Li, C. & Wong, W.H. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl. Acad. Sci. USA* **98**, 31–36 (2001).
49. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628 (2008).
50. Pruitt, K.D., Tatusova, T., Brown, G.R. & Maglott, D.R. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.* **40**, D130–D135 (2012).

Supplementary Figure 1



Mutations of *STAG1, STAG2* and *PDS5B* found in the discovery samples
Somatic mutations of cohesin components identified by whole exome sequencing were validated by Sanger sequencing in bone marrow samples from cases MDS-19 (*STAG2*) (**a**), MDS-12 (*STAG2*) (**b**), MDS-11 (*STAG1*) (**c**) and tAML-02 (*PDS5B*) (**d**), using CD3+ T cells as germline control.

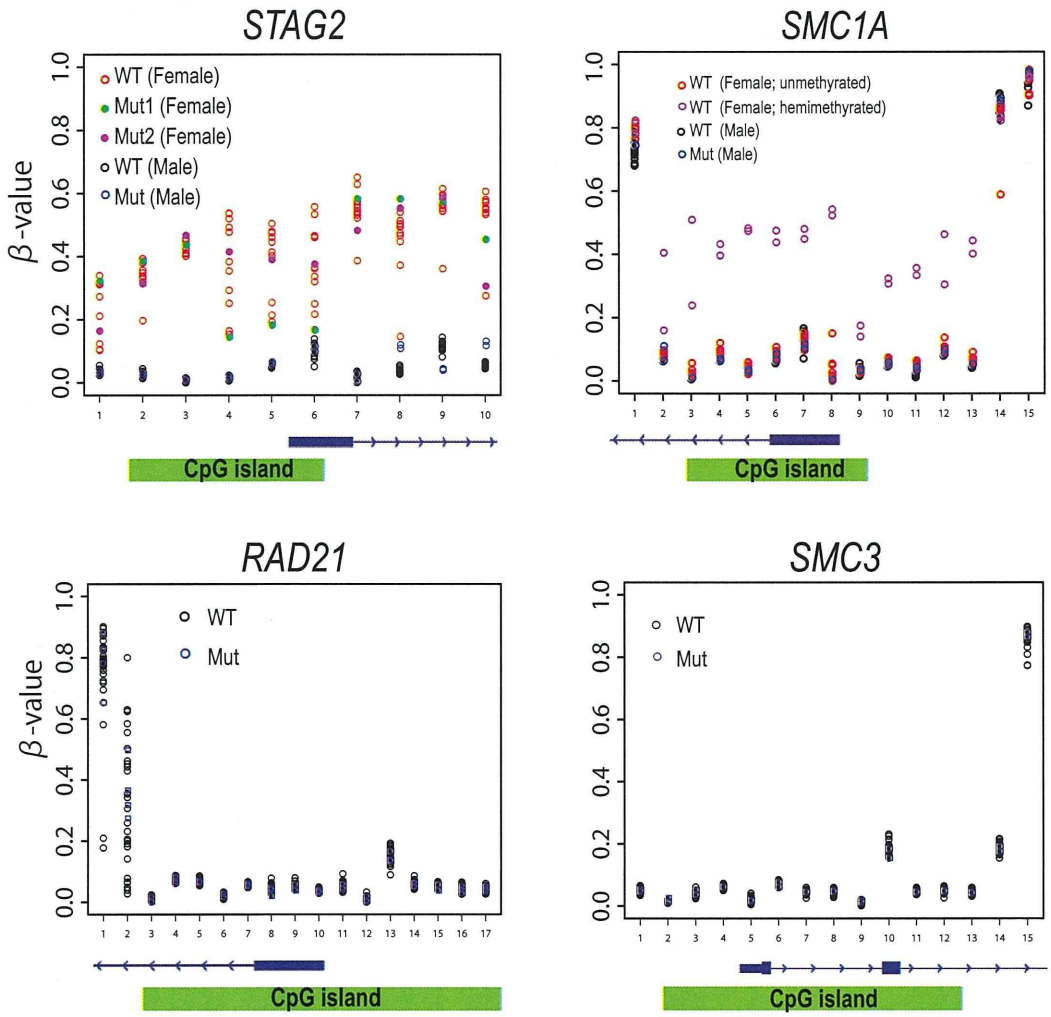
Supplementary Figure 2



Genetic deletions involving the cohesin components
a. Gene deletions involving the *STAG2 and RAD21*, which are indicated by dotted lines. Copy numbers are indicated by color gradients as indicated. **b.** The somatic origins of the deletions were confirmed in 2 cases, in which germline controls were available (asterisks). In 4 cases, the retention of substantial numbers of heterozygous SNP calls within the deletions (green bars) indicated residual normal cells that did not have LOH (i.e. deletions), supporting their somatic origins. In the remaining one case, there was almost complete loss of heterozygous SNP calls within the deletion (blue bar). However, the unusually large size (spanned >150Mbp) of these deletions is unusual for germline events, suggesting its somatic origin (See also **Supplementary Table 6**).

Nature Genetics: doi:10.1038/ng.2731

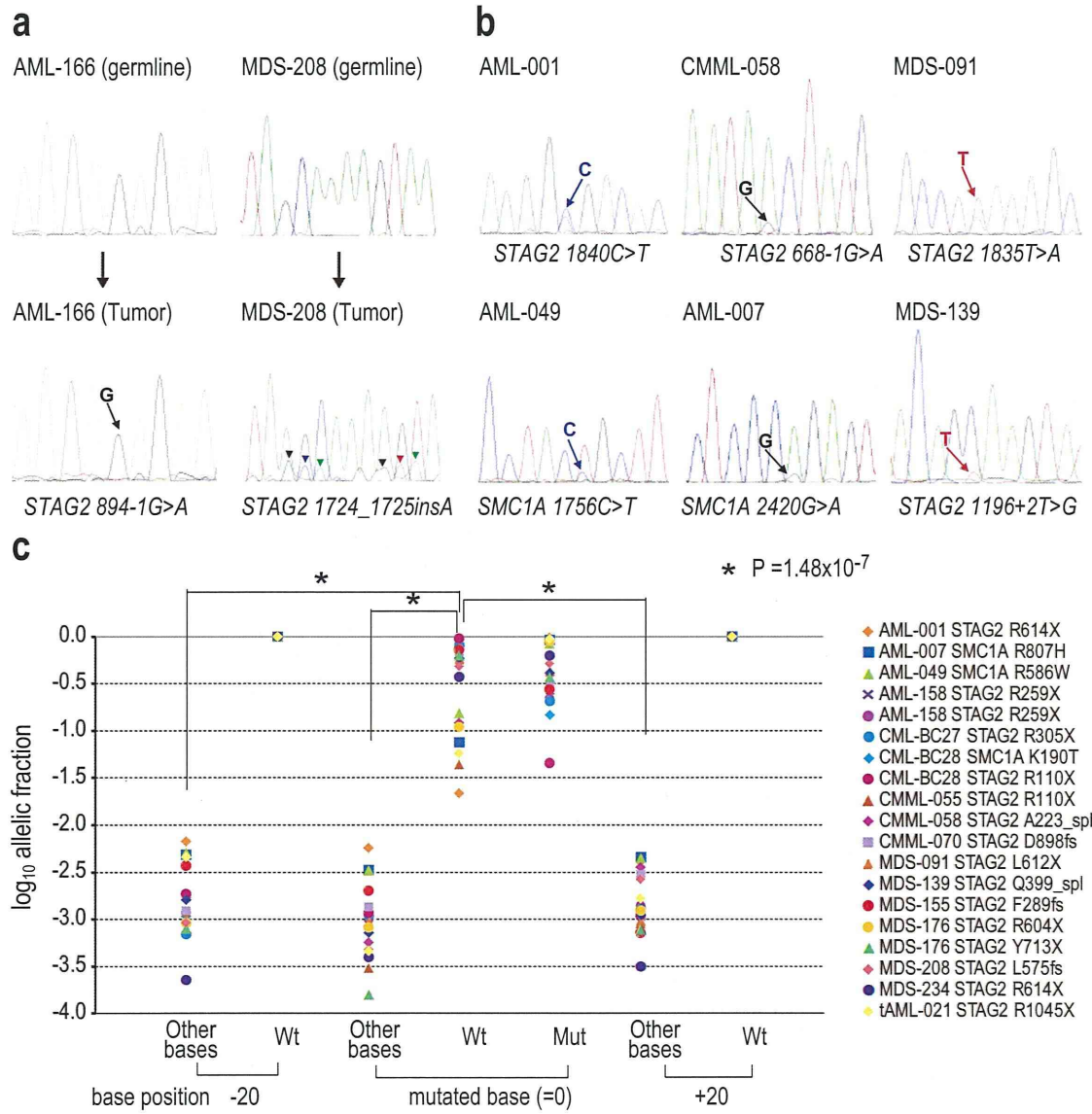
Supplementary Figure 3



Methylation status of promoter regions of 4 cohesin genes
Methylation of multiple sites within and around the promoter regions of 4 cohesin genes were measured using Human Methylation 450 BeadChip (Illumina) and plotted as β-values, where fully methylated, hemi-methylated and non-methylated sites are expected to show values close to 1.0, 0.5 and 0.0, respectively. The positions of CpG islands are indicated in green bars below each panel. In total, 33 samples with (N=12) or without (N=21) mutations of cohesin genes were analyzed. The samples having mutations of corresponding genes (Mut) are discriminated from the wild-type samples (WT). For STAG2 and SMC1A, which are on X chromosome, male and female samples are shown in different colors.

Nature Genetics: doi:10.1038/ng.2731

Supplementary Figure 4



Residual wild-type *STAG2/SMC1A* alleles in male cases with *STAG2/SMC1A* mutations
Since *STAG2* and *SMC1A* are on X chromosome, the presence of the wild type allele within the tumor specimens in male cases should be interpreted as originated from the residual normal cell components and therefore, indicate the somatic origin of the mutations, except for a rare possibility of somatic chimera. This was actually the case with AML-166 and MDS-208 carrying a confirmed somatic mutation **(a)**, in which the wild-type sequences were overlapped with mutated sequences in Sanger sequencing of the tumor specimen (arrow heads). Similarly, the wild-type signals were confirmed in the 11 male cases with an *STAG2* or *SMC1A* mutation, of which 6 mutations are illustrated **(b)**, which are indicated by arrows. **(c).** The presence of the wild-type allele was more sensitively and explicitly detected in deep sequencing, where the genomic region containing each mutation was PCR amplified and subjected to deep sequencing. The number of reads having mutated (Mut) or wild-type (Wt) (or other (Other bases)) base calls at indicated base positions were enumerated and are plotted as logarithm of allelic fractions. Sequencing error rates were estimated at the positions 20bp upstream (-20) and downstream (+20) from each mutation site by enumerating those reads containing the exactly matched 21 bp sequence from -10 to +10 positions around the target position (wild-type reads) and other reads that matched the same sequence except for the target base (error reads). The fraction of the wild-type alleles detected at the mutated positions (position 0) are significantly higher than the background error rates (Other bases) that were evaluated at the mutated position and the position -20 and +20, respectively. (Mann-Whitney's U test).

Nature Genetics: doi:10.1038/ng.2731