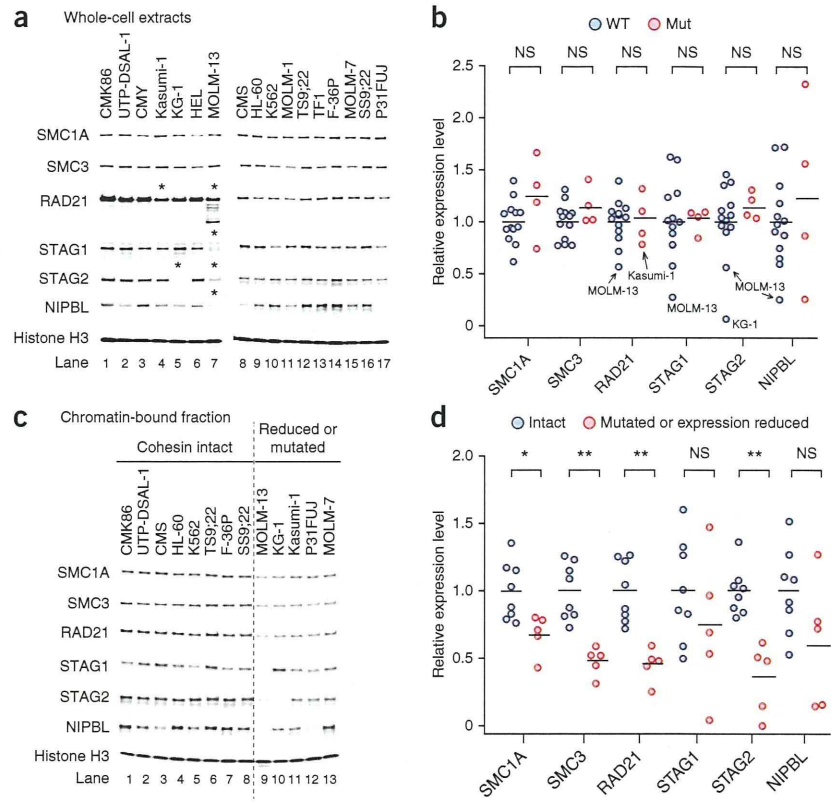
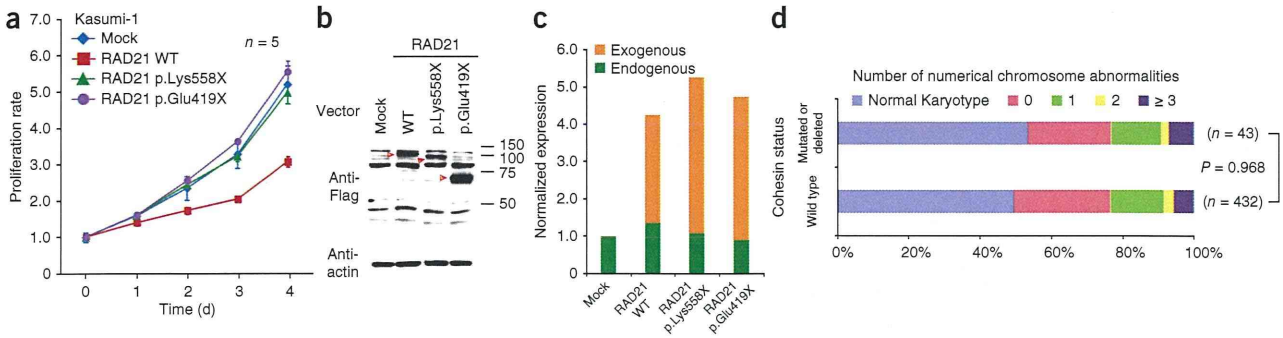


Figure 3 Abnormal cohesin expression and chromatin binding of various cohesin components in myeloid leukemic cell lines. (**a**) Protein blot analysis of the expression of various cohesin components in whole-cell extracts in 17 myeloid leukemia cell lines. Cohesin components showing evaluable reduction 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**).



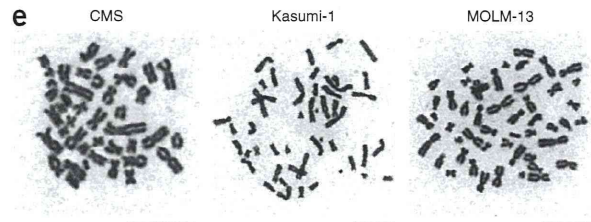


Figure 4 Impact of cohesin mutations on cell proliferation and karyotypes. (**a**) Proliferation of the Kasumi-1 cell line stably transduced with either wild-type RAD21, a truncated allele of *RAD21* (RAD21 p.Lys558X or p.Glu419X) or a mock construct measured by MTT assays ($n = 5$ wells per group). 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 NotI-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).

nature genetics

Integrated molecular analysis of clear-cell renal cell carcinoma

Yusuke Sato[1,2,11], Tetsuichi Yoshizato[1,11], Yuichi Shiraishi[3,11], Shigekatsu Maekawa[1,2,11], Yusuke Okuno[1,11], Takumi Kamura[4], Teppei Shimamura[3], Aiko Sato-Otsubo[1], Genta Nagae[5], Hiromichi Suzuki[1], Yasunobu Nagata[1], Kenichi Yoshida[1], Ayana Kon[1], Yutaka Suzuki[6], Kenichi Chiba[3], Hiroko Tanaka[7], Atsushi Niida[3], Akihiro Fujimoto[8], Tatsuhiko Tsunoda[8], Teppei Morikawa[9], Daichi Maeda[9], Haruki Kume[2], Sumio Sugano[6], Masashi Fukayama[9], Hiroyuki Aburatani[5], Masashi Sanada[1,10], Satoru Miyano[3,7], Yukio Homma[2] & Seishi Ogawa[1,10]

Clear-cell renal cell carcinoma (ccRCC) is the most prevalent kidney cancer and its molecular pathogenesis is incompletely understood. Here we report an integrated molecular study of ccRCC in which ≥100 ccRCC cases were fully analyzed by whole-genome and/or whole-exome and RNA sequencing as well as by array-based gene expression, copy number and/or methylation analyses. We identified a full spectrum of genetic lesions and analyzed gene expression and DNA methylation signatures and determined their impact on tumor behavior. Defective VHL-mediated proteolysis was a common feature of ccRCC, which was caused not only by *VHL* inactivation but also by new hotspot *TCEB1* mutations, which abolished Elongin C–VHL binding, leading to HIF accumulation. Other newly identified pathways and components recurrently mutated in ccRCC included PI3K-AKT-mTOR signaling, the KEAP1-NRF2-CUL3 apparatus, DNA methylation, p53-related pathways and mRNA processing. This integrated molecular analysis unmasked new correlations between DNA methylation, gene mutation and/or gene expression and copy number profiles, enabling the stratification of clinical risks for patients with ccRCC.

Renal cell carcinomas (RCCs) constitute 2–3% of all adult malignancies, with 271,000 new cases and 116,000 related deaths estimated worldwide in 2008 (ref. 1). RCC can be histologically classified into several subtypes, among which ccRCC is the most common, accounting for 70–80% of all kidney cancers[2]. Although immunomodulation using interferon-α and VEGF and/or mTOR inhibition has been applied as systemic therapy for patients with locally advanced or metastatic disease[3], complete surgical resection remains the only curative treatment for ccRCC, except for high-dose interleukin-2, which is used for only limited cases[3]. Genetically, ccRCC is characterized by a very high frequency of biallelic *VHL* inactivation caused by allelic deletion or loss of heterozygosity (LOH) on chromosome 3p (>90%)[4] along with gene mutation (~50%)[5,6] or promoter hypermethylation (5–10%)[7]. In addition, recent whole-exome and targeted sequencing studies have identified frequent recurrent mutations in genes involved in chromatin modification, such as *PBRM1* (ref. 8), *SETD2* (ref. 9), *KDM5C*[9], *KDM6A*[9] and *BAP1* (refs. 10,11), as well as in those involved in the ubiquitin-mediated proteolysis pathway[11]. However, in previous studies, gene mutations were comprehensively investigated for entire coding sequences in only a limited number of cases, and other genetic or epigenetic lesions, including structural abnormalities and DNA methylation, have not been addressed in a comprehensive manner. Thus, knowledge about genetic and/or epigenetic alterations in ccRCC is most likely still incomplete. For example, a subset of ccRCC cases has no detectable *VHL* alterations, and pathogenesis in this subset is poorly characterized compared to that in *VHL*-mutated ccRCC cases in which accumulated hypoxia-inducible factors (HIFs) have a critical role.

Here we performed an integrated molecular study of ccRCC in which ≥100 ccRCC specimens were simultaneously analyzed by whole-genome and/or whole-exome and RNA sequencing in conjunction with microarray-based gene expression, DNA methylation and genomic copy number analyses and immunohistochemistry (Online Methods and **Supplementary Table 1**). An extended cohort of 240 ccRCC specimens, including 106 discovery specimens, was analyzed by SureSelect-based targeted deep sequencing to validate and clarify the effects of major genetic lesions. In addition to previously described common

[1]Cancer Genomics Project, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan. [2]Department of Urology, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan. [3]Laboratory of DNA Information Analysis, Human Genome Center, Institute of Medical Science, The University of Tokyo, Tokyo, Japan. [4]Division of Biological Science, Graduate School of Science, Nagoya University, Nagoya, Japan. [5]Genome Science Division, Research Center for Advanced Science and Technology, The University of Tokyo, Tokyo, Japan. [6]Department of Medical Genome Sciences, Graduate School of Frontier Sciences, The University of Tokyo, Tokyo, Japan. [7]Laboratory of Sequence Analysis, Human Genome Center, Institute of Medical Science, The University of Tokyo, Tokyo, Japan. [8]Center for Genomic Medicine, RIKEN, Yokohama, Japan. [9]Department of Pathology, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan. [10]Department of Pathology and Tumor Biology, Graduate School of Medicine, Kyoto University, Kyoto, Japan. [11]These authors contributed equally to this work. Correspondence should be addressed to S.O. (sogawa-tky@umin.ac.jp).

Received 15 April; accepted 18 June; published online 24 June 2013; doi:10.1038/ng.2699

mutational targets, we identified new mutated genes and pathways that are involved in the pathogenesis of ccRCC, including potentially druggable molecular targets. We also identified unique correlations between mutations, gene expression, DNA methylation and copy number profiles. Our study highlights the role of integrated genome, transcriptome and methylome analyses in clarifying tumor biology and identifying potential therapeutic targets in human cancers.

RESULTS

Whole-genome and whole-exome sequencing

The mean coverage by whole-genome sequencing for paired tumor-normal DNA from 14 ccRCC specimens was 47.2× and 33.6× with 95% and 90% of the entire genome analyzed with ≥20 independent reads on average, respectively (**Supplementary Fig. 1**). A total of 71,424 somatic changes, including 68,273 single-nucleotide variants (SNVs) and 3,151 insertion and/or deletion polymorphisms (indels), were detected in 14 cases (1.7 per megabase per sample) with a true positive rate (TPR) of 99% (630 of the 634 tested were confirmed) (**Supplementary Fig. 2a** and **Supplementary Table 2**). The spectrum of SNVs was over-represented by T>C/A>G transitions followed by C>T/G>A transitions and C>A/G>T transversions. C>T/G>A transitions are predominant in most cancer types[12], whereas T>C/A>G transitions and C>A/G>T transversions were characteristic of ccRCC and have also been reported in hepatocellular carcinoma[13–16] (**Supplementary Fig. 2c**). The mean number of structural variations per case was 12 (range of 0–35) with no apparent breakpoint cluster regions (**Supplementary Fig. 3**).

Table 1 Significantly mutated genes in whole-exome analysis of 106 ccRCCs

Gene	Missense mutations	Nonsense, indel or splicing mutations	Total mutations	Samples	Passenger probability (*P* value)	*q* value
VHL	19	23	42	42	1.32×10^{-102}	1.03×10^{-99}
PBRM1	4	24	28	28	2.63×10^{-36}	1.02×10^{-33}
BAP1	3	5	8	8	1.82×10^{-9}	4.71×10^{-7}
TCEB1	5	0	5	5	7.07×10^{-9}	1.37×10^{-6}
SETD2	5	7	12	12	2.06×10^{-8}	3.20×10^{-6}
FPGT	4	1	5	3	1.13×10^{-7}	1.46×10^{-5}
MUDENG	6	1	7	2	3.38×10^{-7}	3.75×10^{-5}
KEAP1	3	2	5	5	5.95×10^{-5}	5.78×10^{-3}
TET2	7	1	8	6	5.59×10^{-5}	4.83×10^{-3}
MUC4	6	0	6	6	1.02×10^{-4}	7.91×10^{-3}
MLLT10	3	0	3	3	2.30×10^{-4}	1.62×10^{-2}
MSGN1	3	0	3	2	2.85×10^{-4}	1.85×10^{-2}
KRT32	3	1	4	4	2.21×10^{-4}	1.32×10^{-2}
M6PR	1	2	3	3	2.77×10^{-4}	1.54×10^{-2}
RPL14	3	0	3	2	3.90×10^{-4}	2.02×10^{-2}
GRB7	4	0	4	4	4.20×10^{-4}	2.04×10^{-2}
TP53	1	2	3	3	3.85×10^{-4}	1.76×10^{-2}
CSMD3	8	1	9	8	7.08×10^{-4}	3.06×10^{-2}
DNHD1	3	1	4	3	6.44×10^{-4}	2.64×10^{-2}
PIK3CA	5	0	5	5	6.90×10^{-4}	2.68×10^{-2}
NLRP12	3	0	3	3	8.93×10^{-4}	3.31×10^{-2}
VMO1	2	0	2	2	9.89×10^{-4}	3.49×10^{-2}
OR4C13	2	1	3	3	1.10×10^{-3}	3.72×10^{-2}
KCNMA1	4	1	5	5	1.24×10^{-3}	4.00×10^{-2}
LMAN2L	1	2	3	2	1.69×10^{-3}	5.24×10^{-2}
MTOR	7	0	7	6	1.44×10^{-3}	4.31×10^{-2}
ZNF536	5	0	5	5	1.63×10^{-3}	4.70×10^{-2}
YIPF3	2	1	3	2	1.57×10^{-3}	4.36×10^{-2}

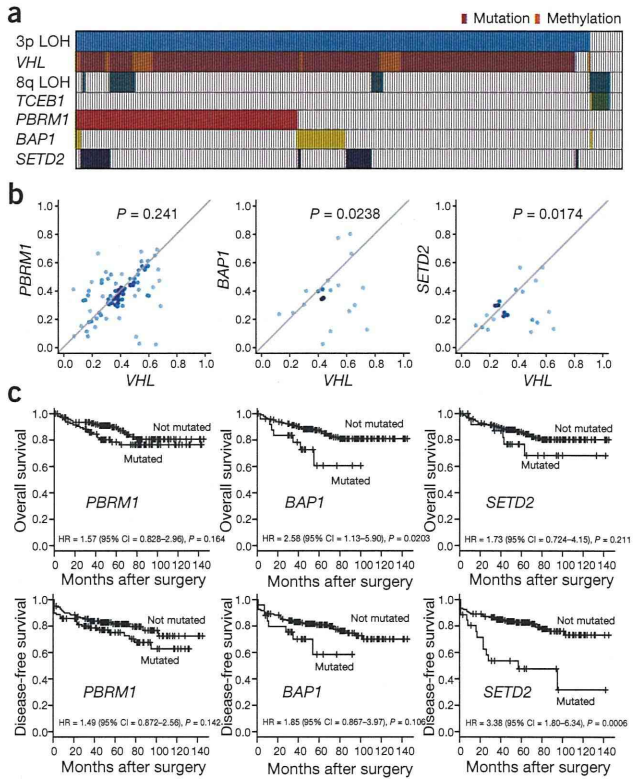


Figure 1 Mutations in 3p target genes and their impact on survival. (**a**) Distribution of common gene mutations and LOH in 240 ccRCC specimens. (**b**) Diagonal plots of observed mutant allele frequencies for *VHL* (*x* axes) and *PBRM1*, *BAP1* and *SETD2* (*y* axes). *P* values were calculated using the paired *t* test (two-sided). (**c**) Effects of common gene mutations on overall survival (top) and disease-free survival (bottom) for 240 ccRCC cases. *P* values were calculated using the log-rank test.

On average, 47 non-silent mutations were identified per case (**Supplementary Fig. 2b**), which accounted for approximately 0.92% of all somatic mutations. The numbers of mutations within coding, intronic (regulatory) and intergenic sequences were roughly proportional across the 14 cases, indicating that mutations were largely random events. To identify the complete spectrum of driver gene targets, we analyzed a total of 106 paired ccRCC specimens by whole-exome sequencing (SureSelect v.4, Agilent Technologies) in which approximately 89% of the target sequences were covered by ≥20 independent reads (**Supplementary Fig. 4**). A total of 5,171 non-silent somatic mutations (48.8 per tumor) were detected with a TPR of 96% (559 of the 582 tested were confirmed). These consisted of 4,234 missense mutations, 232 nonsense mutations, 140 splice-site mutations, 557 indels and 8 read-through changes (**Supplementary Table 3**).

In the 14 specimens that were analyzed by both whole-genome and whole-exome sequencing, 539 of the 839 non-silent mutations (64%) were identified with both platforms. However, reflecting its higher coverage (129×), whole-exome sequencing more efficiently captured the subclonal mutations harbored by a subset of the tumor population, which consequently had lower allele frequencies (**Supplementary Figs. 5** and **6**). Only whole-genome sequencing captured 117 mutations, for which coverage depths were lower in whole-exome than in whole-genome sequencing in most cases (n = 96), even though the mean

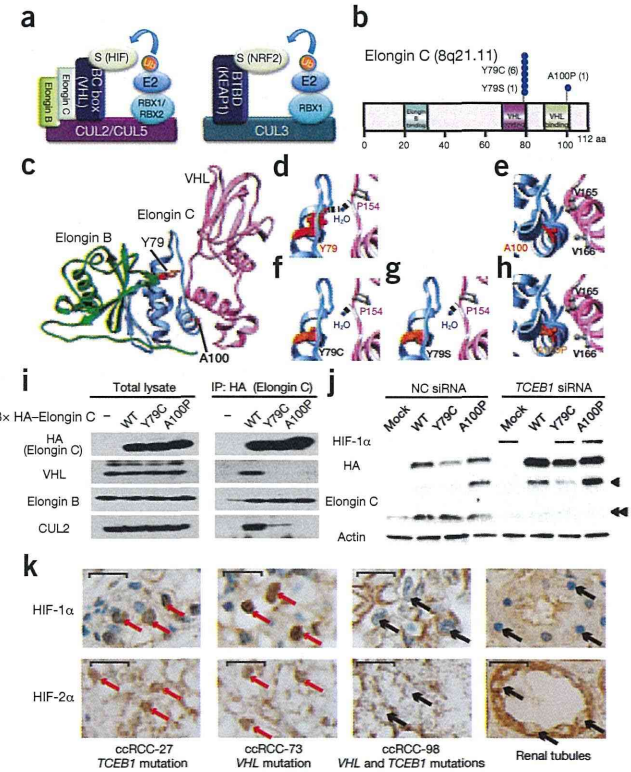


Figure 2 New *TCEB1* mutations and HIF accumulation. (**a**) Two examples of cullin–RING ubiquitin ligase system molecular assemblies using CUL2 or CUL5 (left) and CUL3 (right) that interact with the BC-box protein–Elongin C–Elongin B complex and BTB protein, respectively, to recruit substrate for ubiquitination and subsequent degradation. VHL and KEAP1 are examples of BC-box and BTB proteins, respectively, that recruit HIF and NRF2 proteins for ubiquitin-mediated degradation. (**b**) *TCEB1* mutations (8 of 240 tumors) affect the domains for binding to VHL in Elongin C. (**c**–**e**) Structure of the VHL complex comprising Elongin B, Elongin C and VHL (**c**), with the positions of mutated amino acids Tyr79 (**d**) and Ala100 (**e**) in Elongin C indicated. (**f**,**g**) A critical hydrogen bond between Tyr79 in Elongin C and Pro154 in VHL (**d**) was predicted to be abolished in the Tyr79Cys (**f**) and Tyr79Ser (**g**) Elongin C mutants. (**h**) Hydrophobic binding around Ala100 in Elongin C and Val165 and Val166 in VHL (**g**) could be compromised in the Ala100Pro Elongin C mutant. (**i**) Protein blotting for the indicated components of the VHL-CUL2 complex in total cell lysates (left) and in lysates after immunoprecipitation (IP) with antibody to HA (Elongin C) (right). Lysates were from HEK 293T cells transduced with mock, wild-type (WT) or mutant *TCEB1* constructs (encoding Tyr79Cys and Ala100Pro Elongin C). (**j**) Protein blotting for the effect of *TCEB1* (Elongin C) mutations on HIF accumulation using non-specific siRNA (left) or siRNA specific for endogenous *TCEB1* (right). Endogenous and exogenous Elongin C and HIF-1α were examined. Exogenous 3′ HA–tagged Elongin C was detected with an antibody to HA. Note that antibody to Elongin C could discriminate slower migrating exogenous protein (single arrowhead) from faster migrating endogenous protein (double arrowheads). (**k**) Immunohistochemical analysis of HIF-1α and HIF-2α expression in representative cases from primary ccRCC specimens with *TCEB1* mutations ($n = 5$), *VHL* mutations ($n = 92$) or without *TCEB1* or *VHL* mutations ($n = 9$) as well as in normal kidney tissue ($n = 1$). Red and black arrows indicate positive and negative nuclear immunoreactivity, respectively. Scale bars, 20 μm.

depths of coverage of the entire targeted regions were much higher in whole-exome sequencing (129×) than in whole-genome sequencing (~50×). The coverage in whole-exome sequencing was especially low (<8×) for 45 variants owing to high GC content (for 30 variants) or to no bait designing in targeted exome capture at all (for 15 variants). Subsequent deep sequencing of mutations identified intratumoral heterogeneity in 12 of the 14 ccRCC specimens (**Supplementary Fig. 7**), with the presence of heterogeneity was more explicitly demonstrated in a previous study using whole-exome and targeted deep sequencing combined with extensive multisite sampling from the same tumors and/or metastasized tumor blocks[17].

Recurrent mutations in 3p targets

In whole-genome and/or whole-exome sequencing of the 106 ccRCC specimens, recurrent mutations were observed in 777 genes, of which 28 were considered to be significantly mutated ($q < 0.05$) compared to background mutation rates (**Table 1** and **Supplementary Table 4**). Of the top five significantly mutated genes, *VHL*, *PBRM1*, *BAP1* and *SETD2* were all located within the common site of LOH at 3p between the 3p25 and 3p21 segments (**Supplementary Fig. 8**) and were considered to be the targets of the LOH at 3p found in more than 90% of ccRCC specimens.

Mutations of these common targets were further investigated in detail by deep and/or Sanger sequencing of the relevant genes in combination with assays for DNA methylation status and SNP array–based allele-specific copy number analysis in the extended cohort of paired tumor-normal DNA samples from 240 ccRCC cases (**Fig. 1a** and **Supplementary Fig. 9**). LOH at 3p was found in 226 specimens (94%), which was caused either by simple 3p loss ($n = 175$) or copy-neutral LOH (uniparental disomy, UPD; $n = 51$). There were no significant differences in the mutation rate of 3p target genes in cases with 3p loss and those with 3p UPD. Mutation or promoter hypermethylation was rarely found in cases without LOH at 3p. The vast majority of the 226 cases with LOH at 3p ($n = 221$; 97.8%) had the remaining *VHL* allele affected either by somatic mutation ($n = 197$; 16 nonsense mutations, 70 missense mutations, 100 indels and 11 splice-site mutations) or promoter methylation ($n = 24$). Inactivation of other genes was exclusively caused by gene mutation. Almost all mutations (147/149) involving *PBRM1* (98/98), *SETD2* (25/26) and *BAP1* (24/25) were found in a subset of *VHL*-inactivated cases (**Fig. 1a**) in which *SETD2* and *BAP1* mutations tended to show significantly lower allelic burdens than coexisting *VHL* mutations (**Fig. 1b**). This finding indicates that *SETD2* and *BAP1* mutations are likely to be acquired and selected from within pre-existing *VHL*- and/or *PBRM1*-mutated clones and contribute to tumor progression, as suggested in a recent report.

We next investigated the impact of these mutations on survival and tumor recurrence. In accordance with recent reports[18–20], there was no significant impact of *PBRM1* mutations on overall survival or disease-free survival in univariate analysis (**Fig. 1c**). In contrast, *BAP1* mutations, which were mutually exclusive with *PBRM1* mutations[10] ($P = 2.05 \times 10^{-3}$, Fisher's exact test), were significantly associated with shorter overall survival time (hazards ratio (HR) = 2.58, 95% confidence interval (CI) = 1.13–5.90; $P = 0.0203$), although their impact on relapsed disease was less prominent ($P = 0.106$). This effect could be partly owing to the effects of *SETD2* mutations in *BAP1* mutation–negative cases, as *SETD2*-mutated cases showed a very high relapse rate (HR = 3.38, 95% CI =1.80–6.34; $P = 6.00 \times 10^{-4}$) but did not necessarily have shorter overall survival times. Equivalent results were obtained in multivariate analysis in which mutations in all three 3p target genes were included (**Supplementary Table 5**).

TCEB1 mutations in ccRCC

Another highly significant mutational target was *TCEB1*, which encodes Elongin C, a 112-residue protein[21]. Elongin C was originally identified as a subunit of the heterotrimeric RNA polymerase II elongation factor complex (Elongin) that potently induces mRNA elongation but is also known to be a vital component of the VHL complex. In the latter complex,

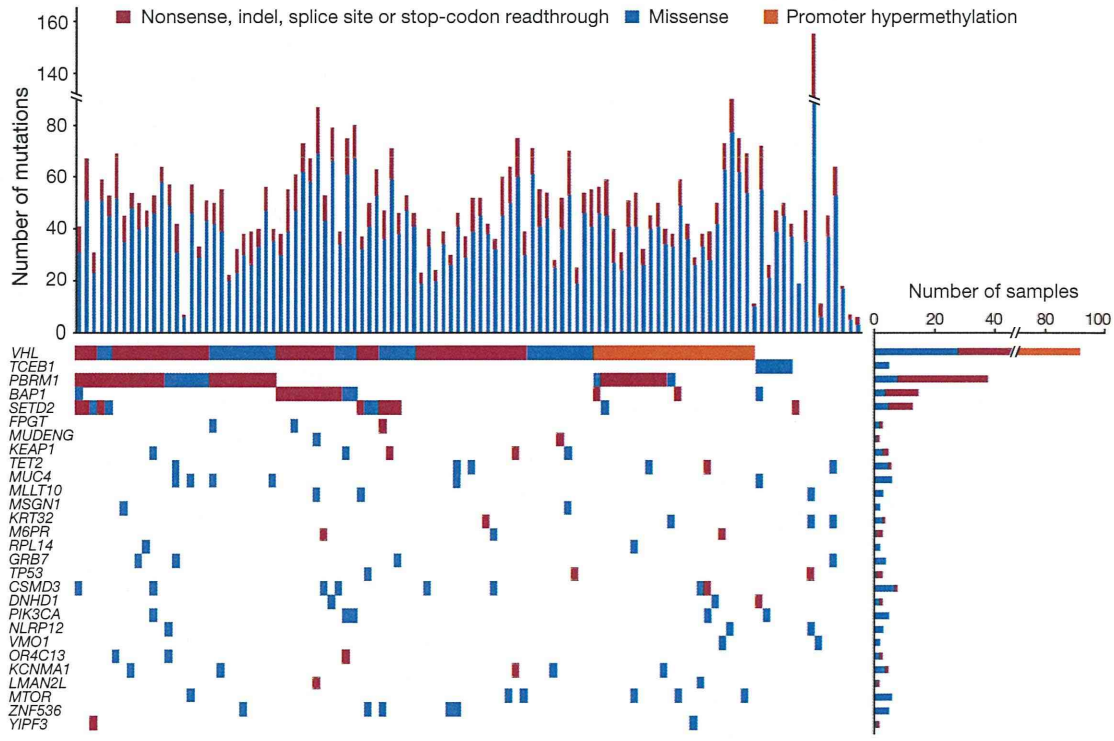


Figure 3 Significantly mutated genes and pathways for 106 ccRCC specimens. The number of somatic mutations in each case (top) and the number of cases that had alterations in significantly mutated genes (bottom right) are shown in a bar plot.

VHL, Elongin B, Elongin C and a catalytic RING subunit (RBX1), which binds ubiquitin-conjugated E2 component, are organized on a cullin scaffold protein (CUL2) to accomplish ubiquitination of VHL-bound HIF proteins (**Fig. 2a,** left)[22]. Targeted deep sequencing and methylation analysis for the entire cohort identified 8 mutations (3.3%) in *TCEB1*. No mutations were detected in other complex components, including *TCEB2* (encoding Elongin B), *CUL2* and *RBX1*. Together with *VHL* lesions, genetic and epigenetic alterations in the VHL complex accounted for 229 of the 240 ccRCC specimens (95.4%)in which *TCEB1* mutations and *VHL* lesions were completely mutually exclusive ($P = 2.50 \times 10^{-14}$, Fisher's exact test), further underscoring the critical role of VHL complex inactivation in the pathogenesis of ccRCC (**Fig. 1a**). There were no significant differences in the clinicopathological characteristics of cases with mutated *VHL* and those with mutated *TCEB1* (**Supplementary Table 6**).

Whereas *VHL* mutations and methylation were closely associated with LOH at 3p, *TCEB1* mutations were always accompanied by loss of chromosome 8 ($P = 3.03 \times 10^{-9}$, Fisher's exact test), leading to complete loss of wild-type *TCEB1* alleles (on 8q21) (**Fig. 1a**). However, in contrast to *VHL* mutations of which nonsense or frameshift alterations frequently result in complete loss of protein function, *TCEB1* mutations exclusively involved two conserved amino acids, Tyr79 ($n = 7$) and Ala100 ($n = 1$), with the former being a mutational hotspot (**Fig. 2b** and **Supplementary Fig. 10**). Notably, these two amino acids are positioned close together within the binding interface for the VHL protein (**Fig. 2c**): Tyr79 mediates a hydrogen bond with Pro154 of VHL via a water molecule (**Fig. 2d**), and Ala100 participates in hydrophobic interactions with Val165 and Val166 of VHL (**Fig. 2e**)[23,24]. Thus, mutations that affect these two amino acids (p.Tyr79Cys, p.Tyr79Ser and p.Ala100Pro) are predicted to abolish the interaction between Elongin C and VHL and to result in compromised ubiquitination and subsequent accumulation of HIF (**Fig. 2f–h**).

Indeed, when expressed in HEK 293T cells, wild-type Elongin C effectively coprecipitated with VHL and CUL2, whereas the interaction with VHL and CUL2 was markedly reduced with mutant Elongin C (Tyr79Cys and Ala100Pro; **Fig. 2i** and **Supplementary Fig. 11a**). We also examined the effects of mutant Elongin C on HIF-1α accumulation by exogenously expressing various Elongin C proteins in HeLa cells. As expected, HIF-1α accumulation was not observed after simply expressing these putative loss-of-function Elongin C mutants or the wild-type protein (**Fig. 2j**, left). However, when endogenous wild-type Elongin C expression was suppressed by small interfering RNA (siRNA) specific for the endogenous *TCEB1* transcripts (Online Methods and **Supplementary Table 7a**), both mock-transduced cells and mutant *TCEB1*–transduced cells but not wild-type *TCEB1*–transduced cells showed recognizable HIF-1α accumulation (**Fig. 2j**, right). These results suggest that the two *TCEB1* mutants (encoding Tyr79Cys and Ala100Pro) actually represent loss-of-function alleles with regard to VHL complex function and that biallelic inactivation is required for HIF-1α accumulation, explaining why *TCEB1* mutations were always accompanied by a deletion of the intact *TCEB1* allele on chromosome 8.

Finally, to confirm the oncogenic role of these *TCEB1* mutations in primary ccRCC, we investigated HIF-1α and HIF-2α expression in primary surgical ccRCC specimens with *TCEB1* mutations by immunohistochemistry. As with *VHL*-mutated tumors, all five tumors with *TCEB1* mutation exhibited increased HIF-1α expression in immunohistochemistry compared to normal kidney samples or tumors that lacked *TCEB1* and *VHL* mutations (**Fig. 2k**, **Supplementary Fig. 12** and **Supplementary Table 8**).

Other recurrent mutations

Other newly identified recurrent mutational targets included *TET2*, *KEAP1* and *MTOR* (**Fig. 3**). *TET2* encodes an α-ketoglutarate–dependent oxigenase frequently inactivated in myeloid malignancies by gene

mutation[25,26]. TET2 catalyzes the conversion of 5-methylcytosine to 5-hydroxymethlcytosine, which is now believed to be a critical step in DNA demethylation. A recent study indicated that TET2 also mediates histone O-GlcNAcylation during gene transcription[27]. No *TET2* mutations have been reported in non-hematopoietic tumors, except for rare mutations in colorectal cancers (5/214 examined; 2.3%)[28]. *TET2* was mutated in 6 of 106 ccRCC cases (5.7%). Except for one frameshift mutation, five were missense mutations, of which four affected the cysteine-rich or catalytic domain (**Supplementary Fig. 13a**). In copy number analysis using SNP arrays (**Fig. 4**), *TET2* was also located within the significantly deleted regions at 4q24 ($n = 11$; 10.4%) (**Fig. 4f**). In combination, *TET2* mutations and deletions accounted for 17 ccRCC cases (16.0%), with no case having biallelic inactivation that indicated a haploinsufficiency effect of *TET2* on the pathogenesis of ccRCC. KEAP1 is a key component of another cullin–RING ubiquitin ligase complex that is involved in oxidative stress responses by regulating the ubiquitination of the KEAP1-bound NRF2 transcription factor (also known as NEF2L2) (**Fig. 2a,** right)[29]. Frequent *KEAP1* and *NRF2* mutations that abrogate their physical interaction were originally reported in squamous cell carcinoma of the lung and in other solid cancers[30–32],

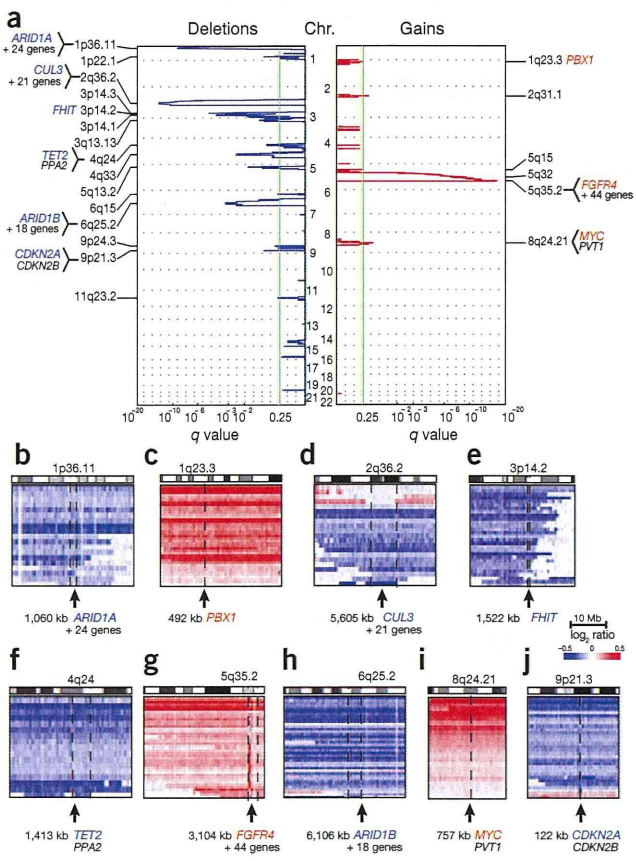


Figure 4 Significant copy number alterations in 240 ccRCC specimens. (**a**) Regions showing statistically significant increase or decrease in genomic copy number were detected using the GISTIC algorithm based on SNP array analysis. For each *q*-value peak, putative gene targets are listed. A dashed line represents the centromere of each chromosome. Red and blue lines indicate *q* value for gains and deletions, respectively. (**b–j**) Log-ratio copy number heatmaps are shown for gene targets at 1p36.11 (**b**), 1q23.3 (**c**), 2q36.2 (**d**), 3p14.2 (**e**), 4q24 (**f**), 5q35.2 (**g**), 6q25.2 (**h**), 8q24.21 (**i**) and 9p21.3 (**j**).

with KEAP1-mediated NRF2 degradation compromised, resulting in deregulated transcriptional activity of the abnormally accumulated NRF2. Of note, compromised KEAP1-mediated NRF2 degradation is also caused by abnormally accumulated fumarate in congenital fumarate hydratase deficiency[33,34], which predisposes to type 2 papillary renal cell carcinoma (pRCC2), and also by somatic mutations in *NRF2* and *CUL3* in sporadic cases with pRCC2 (ref. 35). The current study confirmed that mutually exclusive mutations in *KEAP1* ($n = 5$), *NRF2* ($n = 1$) and *CUL3* ($n = 1$) are also found in the clear-cell subtypes of RCC (6.6%) (**Supplementary Fig. 14**), together with deletions in the *CUL3* locus at 2q36 ($n = 11$; 10.4%) (**Fig. 4d**), with no case having biallelic inactivation in this pathway. *MTOR* was also a newly identified recurrent mutational target and was mutated in 6 of 106 ccRCC cases (5.7%), although a single case with an activating *MTOR* mutation was previously reported[17]. Together with mutations in *PTEN* ($n = 2$), *PIK3CA* ($n = 5$), *PIK3CG* ($n = 2$), *RPS6KA2* ($n = 3$), *TSC1* and *TSC2* (ref. 36) ($n = 2$), and other genes, a total of 28 cases (26%) had mutations that involved phosphoinositide 3-kinase (PI3K)-AKT-mTOR signaling. Except for 3 known tumor suppressor genes—*PTEN*, *TSC1* and *TSC2*—27 mutations were found in 13 genes that are thought to functionally act as oncogenes. In fact, none of the 27 mutations were nonsense, frameshift or splice-site changes, which was highly unexpected from the observed overall frequencies of these types of mutations in ccRCC ($P = 0.00946$, Fisher's exact test), suggesting that these mutations largely act as oncogenes. These mutations were mutually exclusive, except for in two cases that had both *PTEN* and *AKT2* mutations. *FGFR4* was within the significantly amplified region at 5q35 ($n = 69$), and, in total, 81 cases (76%) had genetic alterations in this pathway (**Fig. 5a**). These findings provide additional rationale for the effectiveness of mTOR inhibitors in ccRCC.

Copy number lesions and significantly affected pathways

We also performed SNP array–based copy number analysis for the 240 ccRCC specimens to identify candidate target genes involved in pathogenesis. Most copy number lesions involved large chromosomal segments, as found in LOH at 3p (94%), gain of 5q (65%), gain of 7q (41%), loss of 8p with or without loss of 8q (20%), LOH at 9p (15%), LOH at 14q (27%) and LOH at 18q (11%) (**Supplementary Fig. 15a**). Even though there was a close correlation between monosomy 8 and *TCEB1* mutation, total or partial loss of chromosome 8 was also found in cases with wild-type *TCEB1*, in which common loss of 8p seemed to be relevant to ccRCC pathogenesis. Hyperploid tumors (defined by ploidy of >2.5) accounted for 17.5% ($n = 42$) of the cases and had a significantly higher rate of metastasis ($P = 2.98 \times 10^{-5}$, Cox proportional hazards model) and poor prognosis ($P = 3.93 \times 10^{-2}$, Cox proportional hazards model). In hyperploid tumors, copy numbers in LOH involving 3p, 9p and 14q were largely neutral, suggesting that these tumors had evolved from diploid tumors with typical deletions of the relevant chromosome segments (**Supplementary Fig. 15b,c**). Supporting this notion was the fact that mutations in the 3p target genes in cases with UPD at 3p showed higher allele frequencies than those for mutations within 2*N* regions (**Supplementary Fig. 16**). Using GISTIC 2.0 analysis, significant focal gains and deletions ($q < 0.25$) were found at 20 loci (5 gains and 15 losses) that involved known tumor suppressors and oncogenes, including *ARID1A* (1p36.11), *CUL3* (2q36.2), *FHIT* (3p14.2), *TET2* (4q24), *ARID1B* (6q25.2), *CDKN2A* and *CDKN2B* (9p21.3), *PBX1* (1q23.3), *FGFR4* (5q35.2) and *MYC* (8q24) (**Fig. 4**).

Significantly affected pathways in ccRCC were further investigated by searching for statistically overrepresented gene families that were somatically mutated and expressed and/or showed copy number abnormalities (Online Methods). In addition to PI3K-AKT-mTOR signaling, significantly affected pathways (false discovery rate