Table 1 Splicing factor mutations in myeloid neoplasms

Gene	Chromosome	Size (aa)	Frequency	Phenotype association	Mutational hot spots	Functions
SF3B1	*2q33.1*	1304	Common	RARS, RCMD-RS	K700, K666, K662, K622	3′ recognition
SRSF2 (SC35)	*17q25.1*	221	Common	CMML	P95	3′ recognition
U2AF35 (U2AF1)	*21q22.3*	240	Common	MDS/CMML	S34, Q157	3′ recognition
ZRSR2	*Xp22.1*	483	Common	MDS/CMML	Nonsense, frameshift	3′ recognition
SF3A1	*22q12.2*	793	Rare	Unclear	None	3′ recognition
U2AF65 (U2AF2)	*19q13.42*	475	Rare	Unclear	None	3′ recognition
SF1	*11q13.1*	639	Rare	Unclear	None	3′ recognition
PRPF40B	*12q13.12*	871	Rare	Unclear	None	3′ recognition (speculated)
PRPF8	*17p13.3*	2335	Rare	Unclear	None	Aligning 5′ and 3′ exons
LUC7L2	*7q34*	392	Rare	Unclear	None	Recognition of nonconsensus splice sites

changes strongly suggested that they could be associated with some gain of function rather than represented simple loss of functions. In contrast, mutations of *ZRSR2* on X chromosome were distributed along the entire coding region [13]. About two-thirds of mutations were either nonsense or frameshift changes, causing a premature stop codon [13, 17, 23]. The majority of the *ZRSR2* mutated cases were male, in which single mutations resulted in complete loss of functions [13].

Genotype–phenotype association

While compromised 3′ splice site recognition seems to be a common consequence of different splicing factor mutations, there exist strong genotype and phenotype associations for splicing factor mutations. This is most prominent for the association of *SF3B1* mutations with increased ring sideroblasts. *SF3B1* mutations were found in 68–82 % of refractory anemia with ring sideroblasts (RSRS) and 57–76 % of refractory cytopenia with ring sideroblasts (RCMD-RS) [12, 13, 15, 18]. Malcovati et al. [33] reported that regardless of disease type, SF3B1 mutations strongly predicted the presence of increased sideroblasts with 97.7 % positive predictive value, although it did not necessarily satisfy the criteria for RARS or RCMD-RS (i.e., 15 % of all erythroblasts). Less prominently, *SRSF2* mutations were more frequently found in CMML (30.7–47 %) than in other subtypes of myeloid neoplasms [13, 21]. Interestingly, *SF3B1* mutations, but not other splicing factor mutations, have also been reported in 5–15 % of chronic lymphocytic leukemia (CLL), especially in high-risk cases [34–38]. In addition, *SF3B1* are mutated in several solid cancers, including breast, bladder, endometrial and other cancers, although the mutation frequencies were low [12, 39]. These genotype–phenotype associations may reflect gene-specific functions of individual mutations. For example, *SF3B1* was shown to participate in Hox gene regulation through functional interaction with *polycomb* and *trithorax* genes [40]. *SRSF2* has also been implicated in genetic stability, and its defect could lead to hypermutability [41].

Impact of spliceosome mutations on clinical outcome

Several reports have described the clinical impact of splicing factor mutations. However there appear to be discrepancies in this impact among different studies. Initial reports indicated a significantly better overall survival for *SF3B1* mutated cases compared to unmutated cases in MDS [12, 33], while other studies showed no significant impact of the mutations on survival [18, 23, 27]. *SRSF2* mutations were reported to be associated with poor prognosis in univariate analysis, but may not be an independent prognostic predictor [18, 23, 27]. Also, *U2AF35* mutations were associated with poor prognosis or higher risk of progression to AML in univariate analysis in some series [14, 42] but not in others. To elucidate the exact impacts of splicing factor mutations, a well-designed control study should be required taking other common mutations also into account.

Abnormal RNA splicing caused by splicing factor mutations

The high frequency of mutations in different components of the RNA splicing machinery in MDS suggests that

abnormal RNA splicing is the common consequence of these mutations, which is relevant to the pathogenesis of MDS. However, their effects on RNA splicing have been evaluated only in a very limited context. When expressed in HeLa cells, the S34F U2AF35 allele induces global defects of splicing, causing abnormal retention of intronic sequences in a wide variety of mature mRNA species [13]. On the other hand, the same mutation was shown to promote splicing and exon skipping of a minigene reporter in 293T cells [14]. These are observations in highly artificial systems; no information about abnormal splicing is available for other splicing factor mutants. To understand the role of these mutations in MDS pathogenesis, the effects of the mutant alleles on RNA splicing should be evaluated in more physiological conditions using primary hematopoietic cells. It is particularly important to identify the gene targets of possible splicing defects.

Biological impact of mutations

To date, biological consequences of splicing factor mutations have been tested only for *U2AF35* mutants. Unexpectedly, S34F U2AF35 mutant-transduced HeLa cells showed severely suppressed cell growth rather than enhanced cell proliferation, accompanied by apoptosis and G2/M arrest [13]. The effect of the S34F and Q157P/R U2AF35 mutants was also tested in competitive repopulation assays, in which highly purified mouse hematopoietic stem cells ($CD34^{-}c\text{-}Kit^{+}Scal^{+}Lin^{-}$) were transduced with each mutant and together with normal competitors, transplanted into lethally irradiated mice. In this assay, mutant transduced stem cells showed lower repopulating capacity compared to mock- or wild-type *U2AF35*-transduced cells, as determined by the chimerism in peripheral blood 6 weeks after transplantation [13], indicating that both *U2AF35* mutants could disturb normal hematopoiesis. However, the lower repopulation of mutant-transduced stem cells raises a serious difficulty to our understanding of how *U2AF35* mutated cells achieve clonal dominance over the remaining normal hematopoietic cells, although some oncogenes, such as oncogenic RAS, have been shown to induce apoptosis rather than transformation and promotion of cell growth depending on cell contexts [43, 44]. The importance of cooperation with coexisting mutations and/or the effects of abnormal bone marrow environment represent possible explanations.

Concluding remarks

Whole exome sequencing has revealed the otherwise unexpected involvement of multiple components of the RNA splicing machinery by gene mutations that characterize MDS and related myeloid neoplasms and as such, demonstrated the power of massively parallel sequencing technologies in cancer research. This discovery represents a significant advance in the field of MDS research, providing a novel clue to understanding of the pathogenesis of MDS. However, a number of critical issues remain unsolved: what is the molecular mechanism for these mutations to contribute to MDS pathogenesis, to what extent the deregulated RNA splicing could be involved in that process, what is the molecular nature of the predicted gain-of-functions, and what are the targets of these genes. Their impact on clinical parameters should also be clarified. Finally, the question of whether the recently characterized small molecular inhibitor of SF3B1 might have a therapeutic role in the treatment of myeloid neoplasms with splicing factor mutations remains to be addressed [45–47].

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).

Conflict of interest None.

References

1. Corey SJ, Minden MD, Barber DL, Kantarjian H, Wang JC, Schimmer AD. Myelodysplastic syndromes: the complexity of stem-cell diseases. Nat Rev Cancer. 2007;7:118–29.
2. Bejar R, Levine R, Ebert BL. Unraveling the molecular pathophysiology of myelodysplastic syndromes. J Clin Oncol Off J Am Soc Clin Oncol. 2011;29:504–15.
3. Shih AH, Abdel-Wahab O, Patel JP, Levine RL. The role of mutations in epigenetic regulators in myeloid malignancies. Nat Rev Cancer. 2012;12:599–612.
4. Levine RL, Carroll M. A common genetic mechanism in malignant bone marrow diseases. New Engl Journal Med. 2009;360:2355–7.
5. Delhommeau F, Dupont S, Della Valle V, James C, Trannoy S, Masse A, et al. Mutation in TET2 in myeloid cancers. New Engl J Med. 2009;360:2289–301.
6. Ley TJ, Ding L, Walter MJ, McLellan MD, Lamprecht T, Larson DE, et al. DNMT3A mutations in acute myeloid leukemia. New Engl J Med. 2010;363:2424–33.
7. Mardis ER, Ding L, Dooling DJ, Larson DE, McLellan MD, Chen K, et al. Recurring mutations found by sequencing an acute myeloid leukemia genome. New Engl J Med. 2009;361:1058–66.
8. Langemeijer SM, Kuiper RP, Berends M, Knops R, Aslanyan MG, Massop M, et al. Acquired mutations in TET2 are common in myelodysplastic syndromes. Nat Genet. 2009;41:838–42.
9. Yan XJ, Xu J, Gu ZH, Pan CM, Lu G, Shen Y, et al. Exome sequencing identifies somatic mutations of DNA methyltransferase gene DNMT3A in acute monocytic leukemia. Nat Genet. 2011;43:309–15.
10. Bejar R, Stevenson K, Abdel-Wahab O, Galili N, Nilsson B, Garcia-Manero G, et al. Clinical effect of point mutations in myelodysplastic syndromes. New Engl J Med. 2011;364:2496–506.

11. Bacher U, Schnittger S, Haferlach T. Molecular genetics in acute myeloid leukemia. Curr Opin Oncol. 2010;22:646–55.
12. Papaemmanuil E, Cazzola M, Boultwood J, Malcovati L, Vyas P, Bowen D, et al. Somatic SF3B1 mutation in myelodysplasia with ring sideroblasts. New Engl J Med. 2011;365:1384–95.
13. Yoshida K, Sanada M, Shiraishi Y, Nowak D, Nagata Y, Yamamoto R, et al. Frequent pathway mutations of splicing machinery in myelodysplasia. Nature. 2011;478:64–9.
14. Graubert TA, Shen D, Ding L, Okeyo-Owuor T, Lunn CL, Shao J, et al. Recurrent mutations in the U2AF1 splicing factor in myelodysplastic syndromes. Nat Genet. 2012;44:53–7.
15. Visconte V, Makishima H, Jankowska A, Szpurka H, Traina F, Jerez A, et al. SF3B1, a splicing factor is frequently mutated in refractory anemia with ring sideroblasts. Leukemia Off J Leuk Soc Am Leuk Res Fund UK. 2012;26:542–5.
16. Cui R, Gale RP, Xu Z, et al. Clinical importance of SF3B1 mutations in Chinese with myelodysplastic syndromes with ring sideroblasts. Leuk Res. 2012;36:1428–33.
17. Damm F, Kosmider O, Gelsi-Boyer V, Renneville A, Carbuccia N, Hidalgo-Curtis C, et al. Mutations affecting mRNA splicing define distinct clinical phenotypes and correlate with patient outcome in myelodysplastic syndromes. Blood. 2012;119:3211–8.
18. Damm F, Thol F, Kosmider O, Kade S, Loffeld P, Dreyfus F, et al. SF3B1 mutations in myelodysplastic syndromes: clinical associations and prognostic implications. Leukemia Off J Leuk Soc Am Leuk Res Fund UK. 2012;26:1137–40.
19. Jeromin S, Haferlach T, Grossmann V, et al. High frequencies of SF3B1 and JAK2 mutations in refractory anemia with ring sideroblasts associated with marked thrombocytosis strengthen the assignment to the category of myelodysplastic/myeloproliferative neoplasms. Haematologica. 2012. doi:10.3324/haematol.2012.072538.
20. Lasho TL, Finke CM, Hanson CA, Jimma T, Knudson RA, Ketterling RP, et al. SF3B1 mutations in primary myelofibrosis: clinical, histopathology and genetic correlates among 155 patients. Leukemia Off J Leuk Soc Am Leuk Res Fund UK. 2012;26:1135–7.
21. Meggendorfer M, Roller A, Haferlach T, et al. SRSF2 mutations in 275 cases with chronic myelomonocytic leukemia (CMML). Blood. 2012. doi:10.1182/blood-2012-01-404863.
22. Patnaik MM, Lasho TL, Hodnefield JM, Knudson RA, Ketterling RP, Garcia-Manero G, et al. SF3B1 mutations are prevalent in myelodysplastic syndromes with ring sideroblasts but do not hold independent prognostic value. Blood. 2012;119:569–72.
23. Thol F, Kade S, Schlarmann C, Loffeld P, Morgan M, Krauter J, et al. Frequency and prognostic impact of mutations in SRSF2, U2AF1, and ZRSR2 in patients with myelodysplastic syndromes. Blood. 2012.
24. Visconte V, Rogers HJ, Singh J, et al. SF3B1 haploinsufficiency leads to formation of ring sideroblasts in myelodysplastic syndromes. Blood. 2012. doi:10.1182/blood-2012-05-430876.
25. Wu SJ, Kuo YY, Hou HA, et al. The clinical implication of SRSF2 mutation in patients with myelodysplastic syndrome and its stability during disease evolution. Blood. 2012. doi:10.1182/blood-2012-02-412296.
26. Zhang SJ, Rampal R, Manshouri T, Patel J, Mensah N, Kayserian A, et al. Genetic analysis of patients with leukemic transformation of myeloproliferative neoplasms shows recurrent SRSF2 mutations that are associated with adverse outcome. Blood. 2012;119:4480–5.
27. Bejar R, Stevenson KE, Caughey BA, et al. Validation of a prognostic model and the impact of mutations in patients with lower-risk myelodysplastic syndromes. J Clin Oncol Off J Am Soc Clin Oncol. 2012;30(27):3376–82.
28. Hirabayashi S, Flotho C, Moetter J, et al. Spliceosomal gene aberrations are rare, coexist with oncogenic mutations, and are unlikely to exert a driver effect in childhood MDS and JMML. Blood. 2012;119(11):e96–99.
29. Takita J, Yoshida K, Sanada M, et al. Novel splicing-factor mutations in juvenile myelomonocytic leukemia. Leukemia Off J Leuk Soc Am Leuk Res Fund UK. 2012;26(8):1879–81.
30. Wahl MC, Will CL, Luhrmann R. The spliceosome: design principles of a dynamic RNP machine. Cell. 2009;136:701–18.
31. Chen M, Manley JL. Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. Nat Rev Mol Cell Biol. 2009;10:741–54.
32. Tronchere H, Wang J, Fu XD. A protein related to splicing factor U2AF35 that interacts with U2AF65 and SR proteins in splicing of pre-mRNA. Nature. 1997;388:397–400.
33. Malcovati L, Papaemmanuil E, Bowen DT, Boultwood J, Della Porta MG, Pascutto C, et al. Clinical significance of SF3B1 mutations in myelodysplastic syndromes and myelodysplastic/myeloproliferative neoplasms. Blood. 2011;118:6239–46.
34. Rossi D, Bruscaggin A, Spina V, Rasi S, Khiabanian H, Messina M, et al. Mutations of the SF3B1 splicing factor in chronic lymphocytic leukemia: association with progression and fludarabine-refractoriness. Blood. 2011;118:6904–8.
35. Wang L, Lawrence MS, Wan Y, Stojanov P, Sougnez C, Stevenson K, et al. SF3B1 and other novel cancer genes in chronic lymphocytic leukemia. New Engl J Med. 2011;365:2497–506.
36. Damm F, Nguyen-Khac F, Fontenay M, Bernard OA. Spliceosome and other novel mutations in chronic lymphocytic leukemia, and myeloid malignancies. Leukemia Off J Leuk Soc Am Leuk Res Fund UK. 2012;26(9):2027–31.
37. Quesada V, Conde L, Villamor N, Ordonez GR, Jares P, Bassaganyas L, et al. Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. Nat Genet. 2012;44:47–52.
38. Rossi D, Rasi S, Spina V, et al. Different impact of NOTCH1 and SF3B1 mutations on the risk of chronic lymphocytic leukemia transformation to Richter syndrome. Br J Haematol. 2012;158(3):426–29.
39. Ellis MJ, Ding L, Shen D, Luo J, Suman VJ, Wallis JW, et al. Whole-genome analysis informs breast cancer response to aromatase inhibition. Nature. 2012;486:353–60.
40. Isono K, Mizutani-Koseki Y, Komori T, Schmidt-Zachmann MS, Koseki H. Mammalian polycomb-mediated repression of Hox genes requires the essential spliceosomal protein Sf3b1. Genes Dev. 2005;19:536–41.
41. Xiao R, Sun Y, Ding JH, Lin S, Rose DW, Rosenfeld MG, et al. Splicing regulator SC35 is essential for genomic stability and cell proliferation during mammalian organogenesis. Mol Cell Biol. 2007;27:5393–402.
42. Makishima H, Visconte V, Sakaguchi H, et al. Mutations in the spliceosome machinery, a novel and ubiquitous pathway in leukemogenesis. Blood. 2012;119(14):3203–10.
43. Tanaka N, Ishihara M, Kitagawa M, Harada H, Kimura T, Matsuyama T, et al. Cellular commitment to oncogene-induced transformation or apoptosis is dependent on the transcription factor IRF-1. Cell. 1994;77:829–39.
44. Serrano M, Lee H, Chin L, Cordon-Cardo C, Beach D, DePinho RA. Role of the INK4a locus in tumor suppression and cell mortality. Cell. 1996;85:27–37.
45. Kaida D, Motoyoshi H, Tashiro E, Nojima T, Hagiwara M, Ishigami K, et al. Spliceostatin A targets SF3b and inhibits both splicing and nuclear retention of pre-mRNA. Nat Chem Biol. 2007;3:576–83.
46. Kotake Y, Sagane K, Owa T, Mimori-Kiyosue Y, Shimizu H, Uesugi M, et al. Splicing factor SF3b as a target of the antitumor natural product pladienolide. Nat Chem Biol. 2007;3:570–5.
47. Webb TR, Joyner AS, Potter PM. The development and application of small molecule modulators of SF3b as therapeutic agents for cancer. Drug Discov Today. 2012 [Epub ahead of print].

Published online 6 March 2013

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

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

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

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

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

ABSTRACT

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

INTRODUCTION

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

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

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

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

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

MATERIALS AND METHODS

Patient samples and sequencing procedures

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

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

Outline of the mutation calling method

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

Alignment of sequencing data

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

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

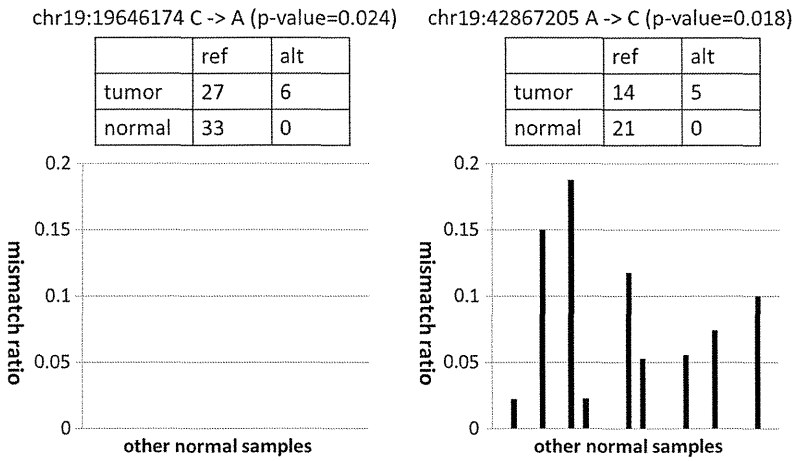

Figure 1. Examples of mismatch ratios of other normal samples for mutation candidates with moderate *P*-values. In both cases, although the mismatch ratios of the target tumour sample were relatively high, the numbers of corresponding supporting variant reads were small. For the candidate on the left,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

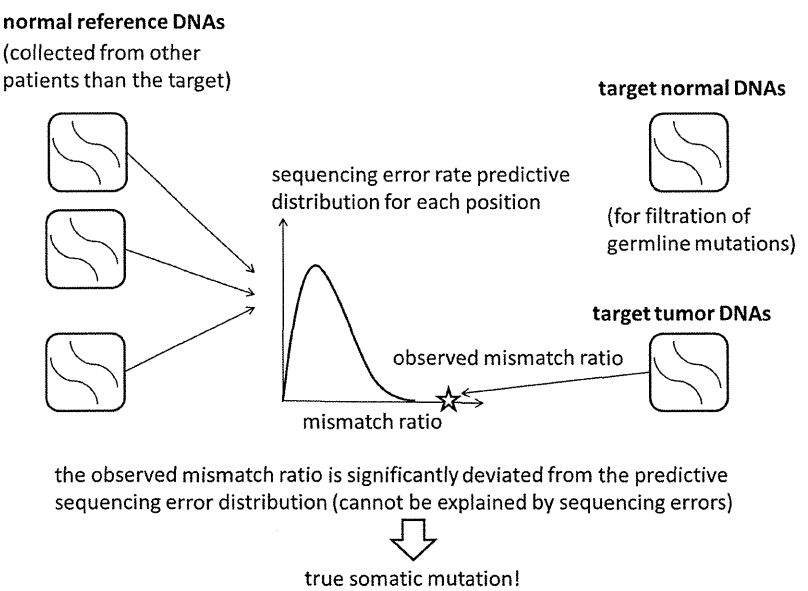

Figure 2. An illustrative description of the proposed method. For each genomic site, the distribution of sequencing errors is estimated using non-paired normal samples from patients other than the target. The mismatch ratio of the target tumour sample is then compared with the distribution. If the mismatch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

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

Definition of variables

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

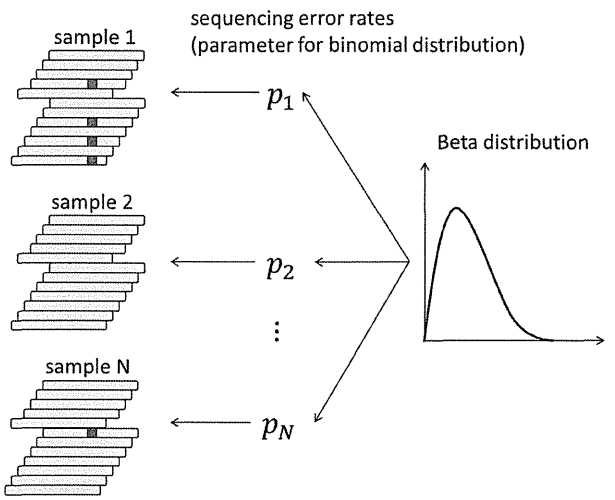

Figure 3. A Beta-binomial sequencing error model. First, the error rate for each sample is generated from the Beta distribution. The number of short reads with sequencing errors is then generated according to the binomial distribution using the parameters of the above error rate for each sample. The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

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

Evaluation of sequencing errors using a Beta-binomial model

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

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

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

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

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

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

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

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

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

EBCall pipeline

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

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

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

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

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

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

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

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

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

Evaluation of sequencing error susceptibility among multiple samples

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

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

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

Comparison with other mutation calling methods

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

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

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

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

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

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

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

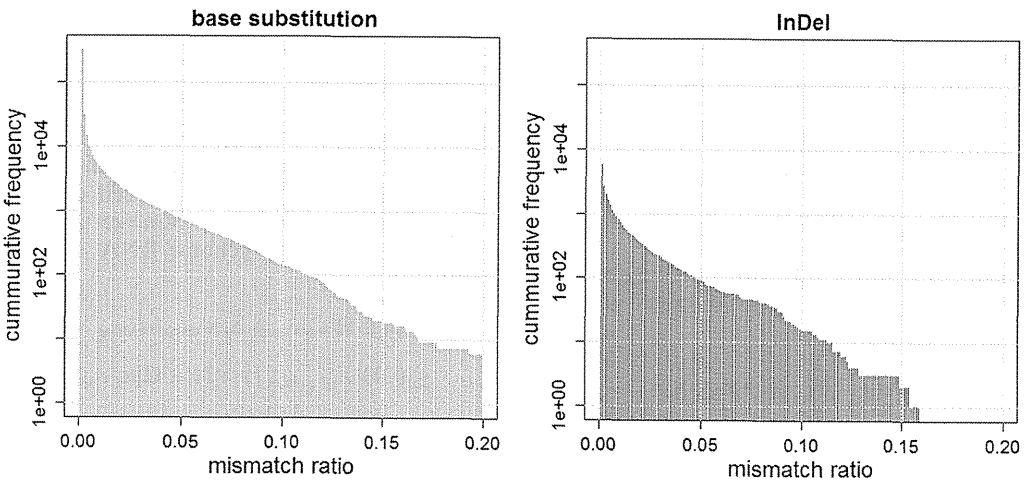

Figure 4. Two bar plots showing the numbers of base substitutions and InDels, whose mean mismatch ratios are above the determined threshold values. For instance, the numbers of base substitutions with mean mismatch ratios of more than 0.01, 0.02, and 0.05 are 4472, 2232, and 727, respectively, while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

RESULTS

Susceptibility to sequencing errors

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

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

Performance comparison with other algorithms for moderate allele frequencies

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

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

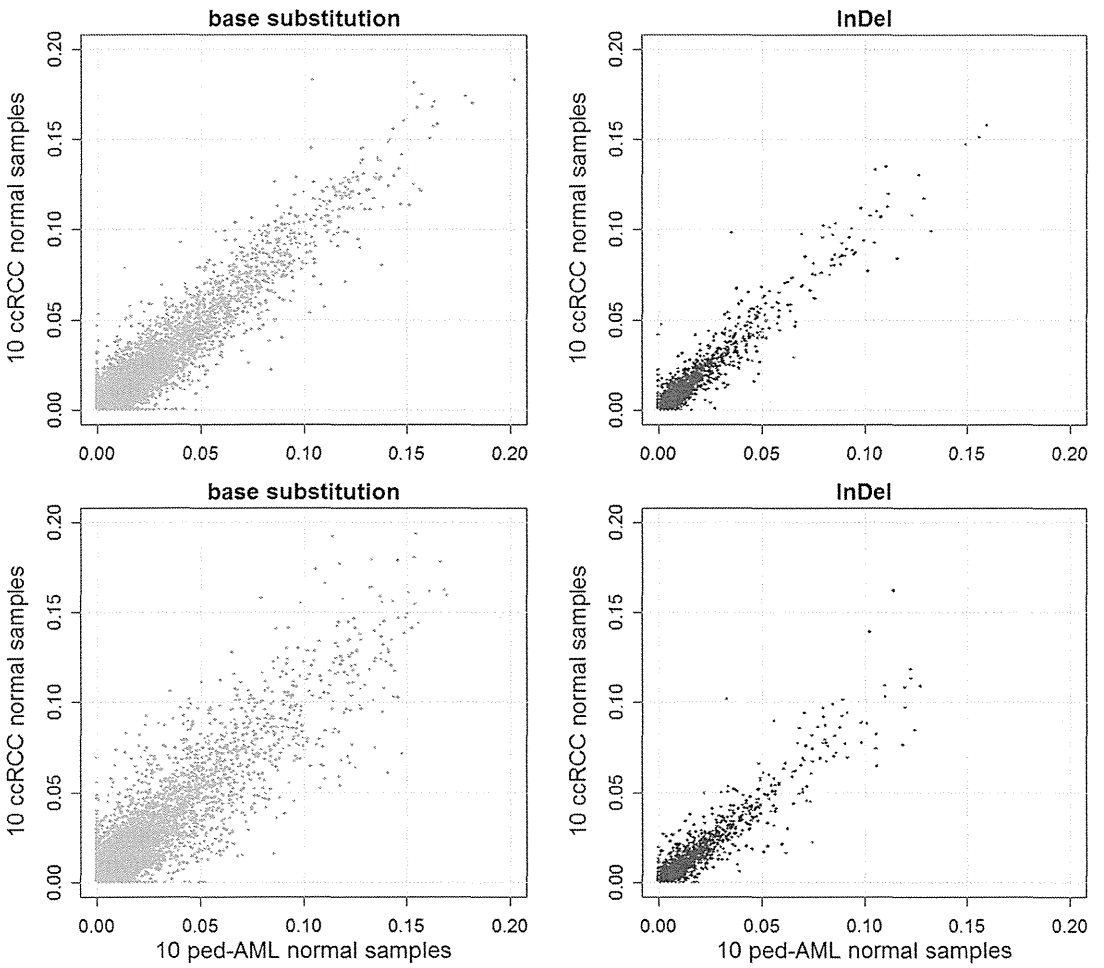

Figure 5. A comparison of scatter plots of the mean mismatch ratios of the base substitution and InDels for two sets consisting of 10 ccRCC normal samples each (upper), and 10 ccRCC normal samples and 10 ped-AML normal samples (lower). The correlation coefficients are 0.777, 0.723, 0.943 and 0.917 for the upper-left, lower-left, upper-right and lower-right panels, respectively.

Next, we investigated the sample-wise sensitivity of each method, in which the threshold value for each method was determined under false positive rates of 0.05, (i.e. 6.54×10^{-4} for *EBCall*, 1.97×10^{-3} for *VarScan*, 60 for *SomaticSniper* and 5.85×10^{-3} for *Genomon-Fisher*). As shown in Supplementary Figure S4, *EBCall* generally outperformed the other calling methods ($P < 0.0074$, Mann–Whitney U test). The improvement in sensitivity varied among the samples may depend on the difference in the mean coverage of the sequencing and tumour contents.

As shown in Figure 7, *EBCall* detected 51 more mutations with six fewer false positives at the cost of nine more false positives as compared with *Genomon-Fisher*. Most of the mutations captured only by *EBCall* showed low sequencing depths or low allele frequencies. Furthermore, *EBCall* detected a number of mutations whose P-value based on Fisher's exact test is moderate (0.1–0.01), maintaining a TPR of 95%. Many candidates with low P-values showed high mean mismatch ratios in other normal samples. These were generally considered to be false positives resulting from sequencing errors that were specific to the target tumour samples at sequencing error-prone sites. To avoid these false positives and maintain a high TPR, a high threshold value had to be set for *Genomon-Fisher*. On the other hand, *EBCall* effectively removed most of these false positives and recovered a number of true somatic mutations. Furthermore, we tested *EBCall* by changing the threshold values for base qualities and mapping qualities and confirmed that the efficiency our method is robust against different parameter values (Supplementary Figure S5).

The processing time of *EBCall* for one sample was 6.5–9.7 h using single core CPU, Intel Quad Core Xeon E5450, 3.0 GHz), whereas those of *VarScan 2*, and *SomaticSniper* were 3.2–6.6 h and 0.7–1.1 h, respectively.

Detection of mutations with low allele frequencies

In total, 557 candidate somatic mutations were called from three tumour samples (RCC31, RCC88 and RCC102) by

EBCall with an additional constraint for the Fisher's *P*-values (see 'Materials and Methods' section). Among these, 395 were evaluable by deep sequencing, of which 349 were successfully confirmed as true mutations. The remaining 162 candidates were not evaluable in deep sequencing owing to either a failure in the design of the PCR primers or low sequencing depths (<5000) for either positive or negative strands. Therefore, they were excluded from the calculation of the true and false positives rates.

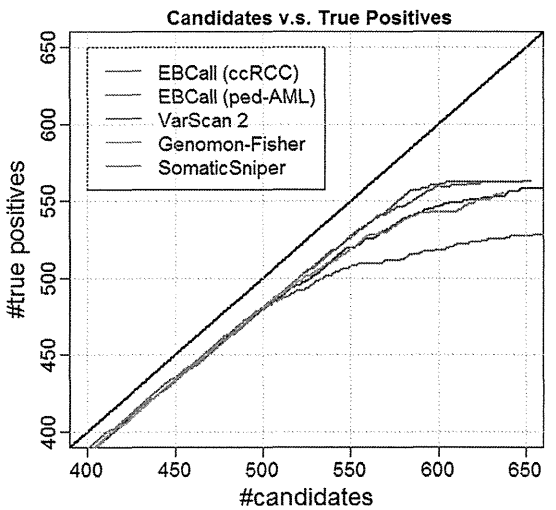

Figure 6. Comparative performance for *EBCall* (20 ccRCC or ped-AML normal samples used as normal reference sets), *Genomon-Fisher*, *VarScan 2* and *SomaticSniper*. The horizontal and vertical axes show the number of candidate somatic mutations and true positives (when changing the threshold of the confidence score for each method) verified by whole genome and whole transcriptome data, respectively.

As shown in Table 1, high TPRs were obtained for candidates with high apparent allele frequencies (>10%): 100, 99.1 and 94.6% for RCC31, RCC88 and RCC102, respectively. For mutations with lower allele frequencies (<10%), TPRs were lower but still showed relatively high values of 79.3, 88.0 and 59.0% for RCC31, RCC88 and RCC102, respectively. Among the 10 candidates called by only *Genomon-Fisher*, only one was successfully validated.

Next, we investigated the causes of false positive results in RCC102. We found that many false positive candidates were supported by reads that were aligned more consistently with the transcriptome than with the genome sequence (Supplementary Figure S6), indicating that small amounts of RNA may have contaminated the exome sequencing library in RCC102, resulting in the calling of several false positives owing to the existence of ambiguous alignments. These false positives were successfully eliminated without affecting the sensitivities by filtering those candidates that have other mutations within 300 bp from the mutation site, through which the TPR increased to 83.6% (Table 2). As the allele frequencies for this kind of false positive were mostly below 10%, RNA contamination may have been problematic only when calling mutations with a low allele frequency.

Finally, the distribution of allele frequencies calculated in deep sequencing for each sample is plotted in Figure 8. The histogram clearly shows the presence of minor tumour subpopulations of cancer cells with <10% allele frequencies in each sample, suggesting that the sensitive detection of somatic mutations with low allele frequencies is effective in capturing intratumoural heterogeneity.

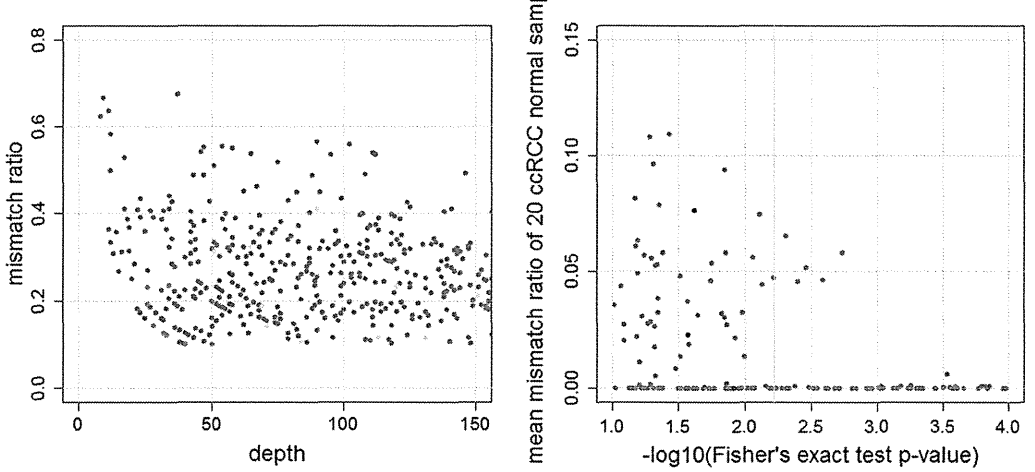

Figure 7. (Left) The comparative results between *EBCall* and *Genomon-Fisher*. Each point, in which the sequencing depth and variant allele frequency are indicated, shows the candidate somatic mutations called by both or either of the two methods. The threshold values are determined such that the false positive rates are 0.05. The green and red points show true positive mutations called by both of the two methods, and only *EBCall*, respectively. The yellow, cyan and magenta points show false positive mutations called by both of the two methods, only *EBCall*, and only *Genomon-Fisher*, respectively. The numbers of green, red, yellow, cyan and magenta points are 506, 51, 20, 9 and 6, respectively. There are no true positive mutations called by *Genomon-Fisher* exclusively. (Right) The *P*-values of Fisher's exact test and the mean mismatch ratio of 20 ccRCC normal samples are plotted. The red and blue points show true positive mutations called and not called by *EBCall*, respectively. On the other hand, the cyan and magenta points show false positive mutations called and not called by *EBCall*, respectively. The yellow vertical line shows the threshold value of the *Genomon-Fisher* determined with false positive rates of 0.05.

Table 1. The numbers of true and false positives for mutations with moderate (above 10%) allele frequencies

Sample	RCC31	RCC88	RCC102	RCC102 (filtered)
No. of true positives	78	109	71	69
No. of false positives	0	1	4	1

Table 2. The numbers of true and false positives for mutations with low (above 2% and below 10%) allele frequencies

Sample	RCC31	RCC88	RCC102	RCC102 (filtered)
No. of true positives	23	22	46	46
No. of false positives	6	3	32	9

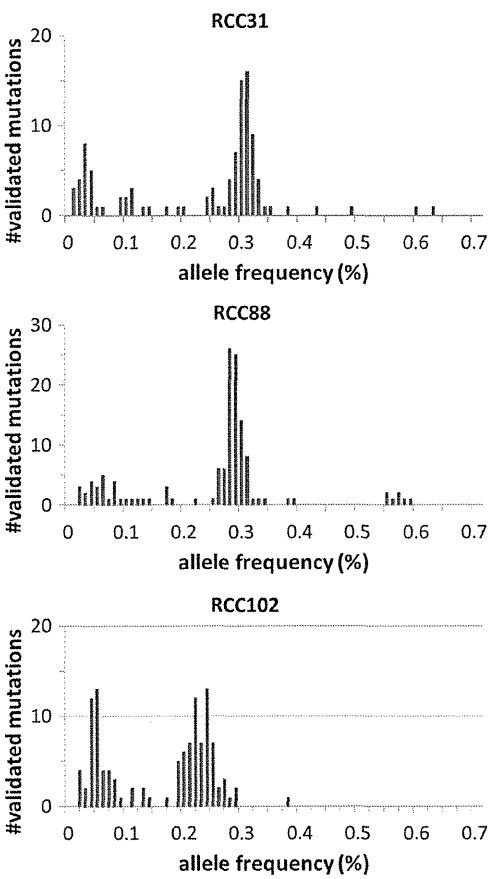

Figure 8. Histograms of the allele frequencies of validated mutations for RCC31 (left), RCC88 (centre) and RCC102 (right).

DISCUSSION

In this article, we have proposed a novel statistical framework, *EBCall*, for detecting somatic mutations using a massively parallel sequencing of the cancer genome. The concept of using data from multiple samples to eliminate sequencing errors is not completely new, but it has been adopted in previous studies (10,16) to discriminate true somatic mutations from errors in the targeted sequencing of much smaller regions. However, most of these approaches filter out somatic mutations with approximately the same common non-reference allele frequencies among multiple tumour samples by regarding them as common sequencing errors. Our approach, on the other hand, uses multiple non-paired normal samples to explicitly estimate the distribution of sequencing errors. Furthermore, we extended this approach to much larger genomic regions (~50 Mb) and accomplished accurate mutation calling from whole-exome sequencing. *EBCall* was not only superior to several existing methods for somatic mutations with moderate-to-high allele frequencies but also effectively detected somatic mutations with low allele frequencies of <10%, which helps in the identification of a clonal architecture within a cancer population. The fact that *EBCall* was robust to the choice of normal reference samples implies that we could improve the accuracy of mutation calling just by using normal samples available in a regular project. Although we confined its application to exome sequencing data in this article, we expect that our approach can improve the accuracy in whole-genome sequencing data with moderate sequencing depths.

A simpler approach for the empirical elimination of sequencing errors would be to identify error-prone genomic positions that satisfy an arbitrary set of criteria (e.g. a 2% mismatch ratio for ≥3 samples among groups of 20 normal samples) and exclude all variants at these positions. However, as the number of sequencing errors has a long-tailed distribution, setting a threshold value for extracting a set of sequencing error prone sites is not a trivial task. The use of overly strict criteria may not remove false positives effectively. On the other hand, when we filter too broad a range of error prone sites, we may miss some true somatic mutations, even when their allele frequencies are considerably higher than the slightly elevated sequencing error rate at that position. Our approach is more flexible in discriminating true mutations from errors because it relies on a rigorous statistical model.

Another approach is to eliminate sequencing errors based on knowledge of the error-prone sequencing features, such as a homo-polymer sequence and specific sequence motifs (11,12). These features can be used to eliminate more sequencing errors and achieve further improvements in accuracy. However, the prediction of error-prone features may not be exhaustively identified or uniformly applied to real sequencing data, regardless of the experimental conditions.

As discussed previously, an understanding of the intratumoural architecture of gene mutations provides an important insight into the clonal evolution of tumour cells, in which the detection of mutations with low allele frequencies is of critical importance. A recent study elegantly approached this issue using deep sequencing (×200) of the whole genome in a breast cancer sample (5). Whole-genome deep sequencing is a powerful approach for detecting sufficient numbers of somatic mutations and reliably identifying tumour subclones. However, the cost of whole-genome deep sequencing for multiple samples

remains expensive. Alternatively, with improved detection of low allele frequency mutations, sequencing data from more targeted regions, such as a whole exome, at a similar depth (e.g. 150–300) can provide an opportunity to capture a sufficient number of repertories of gene mutations within the coding sequences and disclose fine clonal architectures of mutations for multiple samples at acceptable costs.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1 and 2, and Supplementary Figures 1–6.

ACKNOWLEDGEMENT

The super-computing resource was provided by Human Genome Center, Institute of Medical Science, the University of Tokyo. The authors also thank H. Tanaka, Y. Mori and N. Mizota for their technical assistance.

FUNDING

Funding for open access charge: Integrative Systems Understanding of Cancer for Advanced Diagnosis, Therapy and Prevention (Grant-in-Aid for Scientific Research on Innovative Areas from the Ministry of Education, Culture, Sports, Science and Technology, Japan).

Conflict of interest statement. None declared.

REFERENCES

1. Meyerson,M., Gabriel,S. and Getz,G. (2010) Advances in understanding cancer genomes through second-generation sequencing. *Nat. Rev. Genet.*, **11**, 685–696.
2. Shah,S.P., Morin,R.D., Khattra,J., Prentice,L., Pugh,T., Burleigh,A., Delaney,A., Gelmon,K., Guliany,R., Senz,J. *et al.* (2009) Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature*, **461**, 809–813.
3. Ding,L., Ley,T.J., Larson,D.E., Miller,C.A., Koboldt,D.C., Welch,J.S., Ritchey,J.K., Young,M.A., Lamprecht,T., McLellan,M.D. *et al.* (2012) Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature*, **481**, 506–510.
4. Shah,S.P., Roth,A., Goya,R., Oloumi,A., Ha,G., Zhao,Y., Turashvili,G., Ding,J., Tse,K., Haffari,G. *et al.* (2012) The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature*, **486**, 395–399.
5. Nik-Zainal,S., Van Loo,P., Wedge,D.C., Alexandrov,L.B., Greenman,C.D., Lau,K.W., Raine,K., Jones,D., Marshall,J., Ramakrishna,M. *et al.* (2012) The life history of 21 breast cancers. *Cell*, **149**, 994–1007.
6. Larson,D.E., Harris,C.C., Chen,K., Koboldt,D.C., Abbott,T.E., Dooling,D.J., Ley,T.J., Mardis,E.R., Wilson,R.K. and Ding,L. (2012) SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics*, **28**, 311–317.
7. Roth,A., Ding,J., Morin,R., Crisan,A., Ha,G., Giuliany,R., Bashashati,A., Hirst,M., Turashvili,G., Oloumi,A. *et al.* (2012) JointSNVMix: a probabilistic model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data. *Bioinformatics*, **28**, 907–913.
8. Yoshida,K., Sanada,M., Shiraishi,Y., Nowak,D., Nagata,Y., Yamamoto,R., Sato,Y., Sato-Otsubo,A., Kon,A., Nagasaki,M. *et al.* (2011) Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature*, **478**, 64–69.
9. Koboldt,D.C., Zhang,Q., Larson,D.E., Shen,D., McLellan,M.D., Lin,L., Miller,C.A., Mardis,E.R., Ding,L. and Wilson,R.K. (2012) VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.*, **22**, 568–576.
10. Li,M. and Stoneking,M. (2012) A new approach for detecting low-level mutations in next-generation sequence data. *Genome Biol.*, **13**, R34.
11. Dohm,J.C., Lottaz,C., Borodina,T. and Himmelbauer,H. (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.*, **36**, e105.
12. Nakamura,K., Oshima,T., Morimoto,T., Ikeda,S., Yoshikawa,H., Shiwa,Y., Ishikawa,S., Linak,M.C., Hirai,A., Takahashi,H. *et al.* (2011) Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res.*, **39**, e90.
13. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
14. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
15. Kent,W.J. (2002) BLAT–the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
16. Bansal,V. (2010) A statistical method for the detection of variants from next-generation resequencing of DNA pools. *Bioinformatics*, **26**, i318–i324.
17. Wang,K., Li,M. and Hakonarson,H. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.*, **38**, e164.

Smap1 deficiency perturbs receptor trafficking and predisposes mice to myelodysplasia

Shunsuke Kon,[1] Naoko Minegishi,[2] Kenji Tanabe,[3] Toshio Watanabe,[1] Tomo Funaki,[1] Won Fen Wong,[1] Daisuke Sakamoto,[1] Yudai Higuchi,[1] Hiroshi Kiyonari,[4] Katsutoshi Asano,[5] Yoichiro Iwakura,[6] Manabu Fukumoto,[7] Motomi Osato,[8] Masashi Sanada,[9] Seishi Ogawa,[9] Takuro Nakamura,[10] and Masanobu Satake[1]

[1]Department of Molecular Immunology, Institute of Development, Aging and Cancer, and [2]Tohoku Medical Megabank Organization, Tohoku University, Sendai, Japan. [3]Medical Research Institute, Tokyo Women's Medical University, Tokyo, Japan. [4]Laboratory for Animal Resources and Genetic Engineering, RIKEN Center for Developmental Biology, Kobe, Japan. [5]Nihon Gene Research Laboratories, Sendai, Japan. [6]Institute of Medical Science, The University of Tokyo, Tokyo, Japan. [7]Department of Pathology, Institute of Development, Aging and Cancer, Tohoku University, Sendai, Japan. [8]Cancer Science Institute, National University of Singapore, Singapore. [9]Cancer Genomics Project, Faculty of Medicine, The University of Tokyo, Tokyo, Japan. [10]Division of Carcinogenesis, The Cancer Institute, Japanese Foundation for Cancer Research, Tokyo, Japan.

The formation of clathrin-coated vesicles is essential for intracellular membrane trafficking between subcellular compartments and is triggered by the ARF family of small GTPases. We previously identified SMAP1 as an ARF6 GTPase-activating protein that functions in clathrin-dependent endocytosis. Because abnormalities in clathrin-dependent trafficking are often associated with oncogenesis, we targeted *Smap1* in mice to examine its physiological and pathological significance. *Smap1*-deficent mice exhibited healthy growth, but their erythroblasts showed enhanced transferrin endocytosis. In mast cells cultured in SCF, *Smap1* deficiency did not affect the internalization of c-KIT but impaired the sorting of internalized c-KIT from multivesicular bodies to lysosomes, resulting in intracellular accumulation of undegraded c-KIT that was accompanied by enhanced activation of ERK and increased cell growth. Interestingly, approximately 50% of aged *Smap1*-deficient mice developed anemia associated with morphologically dysplastic cells of erythroid-myeloid lineage, which are hematological abnormalities similar to myelodysplastic syndrome (MDS) in humans. Furthermore, some *Smap1*-deficient mice developed acute myeloid leukemia (AML) of various subtypes. Collectively, to our knowledge these results provide the first evidence in a mouse model that the deregulation of clathrin-dependent membrane trafficking may be involved in the development of MDS and subsequent AML.

Introduction

Intracellular and extracellular homeostasis is maintained by a vesicle transport system that mediates the trafficking of membrane proteins to appropriate organelles. Clathrin-coated vesicles are formed at donor membrane sites in a highly ordered manner, and a number of molecules are involved in this process. Among them, small GTPases of the ARF family play a central role in vesicle formation. An ARF molecule cycles between two conformations, an active GTP-bound form and an inactive GDP-bound form. This cycling is mediated by a guanine nucleotide exchange factor that replaces GDP with GTP and a GTPase-activating protein (GAP) that hydrolyzes GTP to GDP and converts ARF into its inactive form. There are 6 ARFs (ARF1–ARF6) and several ARF-related proteins in mammals (1, 2). ARF6 is an isoform that localizes mainly to the plasma membrane and functions in the endocytosis and recycling of vesicles as well as in actin rearrangement and lipid metabolism (3, 4).

We previously demonstrated that small ARF GAP1 (referred to as SMAP1) is a regulator of clathrin-dependent endocytosis, based on a series of observations (5, 6). First, SMAP1 exhibits GAP activity against ARF6, as assessed by an in vitro GAP assay. Second, SMAP1 localizes to juxta-plasma membrane regions in which ARF6 also exists. Third, SMAP1 binds to the clathrin heavy chain directly via its clathrin-binding box. Fourth, overexpression of SMAP1 abrogates clathrin-dependent internalization of the transferrin receptor and E-cadherin.

Recently, mutations and chromosomal translocations associated with various human cancers and leukemia have been identified in the genes that encode endocytosis-related proteins (7–10). However, the precise molecular mechanisms that underlie the effect of these genetic alterations on membrane trafficking and lead to disorders in cell growth and/or differentiation remain poorly understood. Therefore, the significance of these mutations needs to be clarified. One process that could link membrane traffic to alterations in cell growth/differentiation is the deregulation of receptor tyrosine kinase (RTK) downregulation. Alterations in the endocytosis and/or lysosomal degradation of RTKs result in the persistence of these molecules on the membrane, which leads to the activation of growth and differentiation pathways (11–13).

Several studies have reported the involvement of *SMAP1* in oncogenesis in humans. For example, the *MLL* gene is a frequent target for recurrent chromosomal translocations in acute myeloid leukemia (AML), and more than 50 MLL fusion partners have been identified, including endocytosis-related genes, such as *EPS15*, *CALM*, and *EEN* (10). Interestingly, *SMAP1* was previously identified as one of the fusion partners of *MLL* (14). In colorectal cancers displaying microsatellite instability, mutations causing the truncation of the polypeptide chain have been detected in *SMAP1* (11% homozygous and 73% heterozygous) (15). This finding suggests that *SMAP1* may be acting as a tumor suppressor gene in intestinal cells. Based on these findings, we generated *Smap1*-targeted mice to examine the function of SMAP1 in clathrin-dependent vesicle trafficking and to determine the potential role of SMAP1 in cell growth and differentiation in vivo.

Conflict of interest: The authors have declared that no conflict of interest exists.

Citation for this article: *J Clin Invest.* 2013;123(3):1123–1137. doi:10.1172/JCI63711.

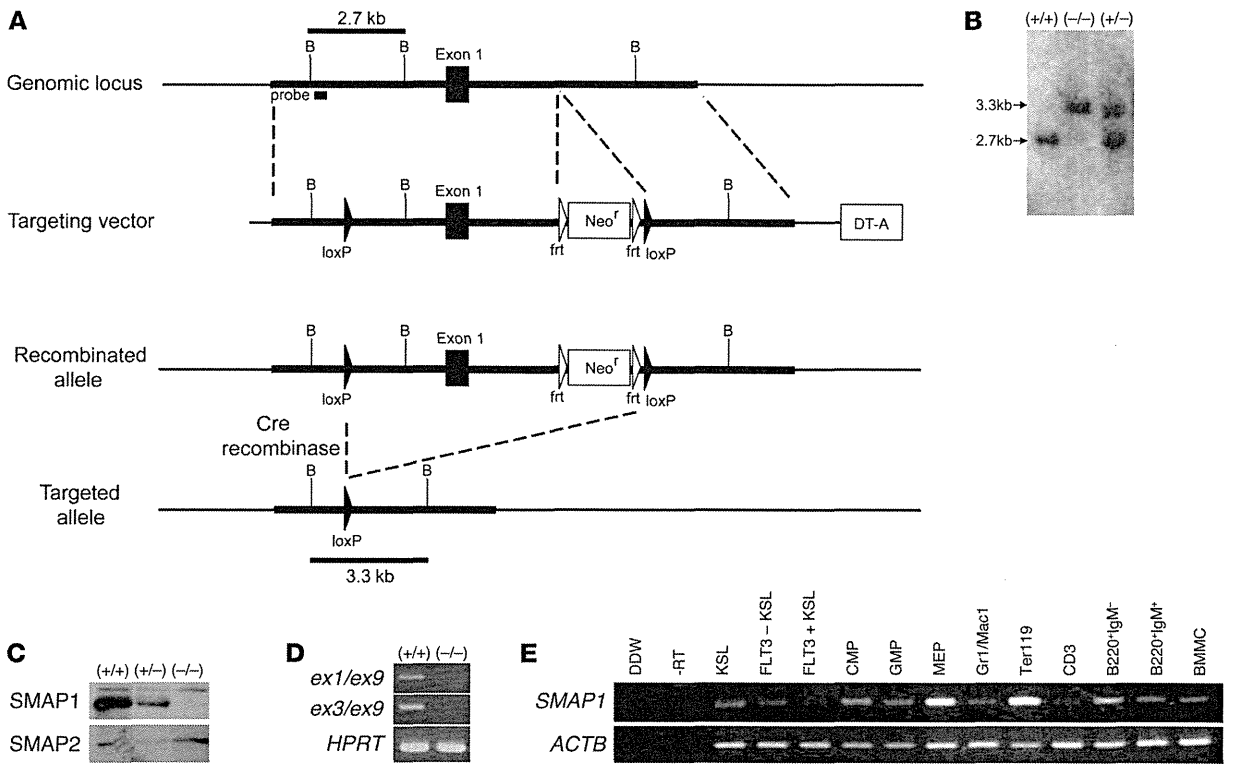

Figure 1

Establishment of *Smap1*-targeted mice and SMAP1 expression. (**A**) Physical maps of the *SMAP1* gene locus and its targeting vector. Features of the recombined and targeted *SMAP1* alleles are also depicted. Horizontal lines indicate the genomic sequences. The thick lines indicate the sequence incorporated into the targeting vector. Exon 1, neomycin resistance gene, and diphtheria toxin subunit A gene are indicated. Black and white arrowheads indicate the loxP and frt sequences, respectively. The small rectangle under the line corresponds to the probe that was used for Southern blot hybridization. B indicates a *BamH*I restriction site. (**B**) Southern blot analysis of genomic DNA prepared from *Smap1*$^{+/+}$ (+/+), *Smap1*$^{+/-}$ (+/–), and *Smap1*$^{-/-}$ (–/–) mice. DNA was digested by *BamH*I and processed for Southern blotting using the hybridization probe shown in **A**. The wild-type and targeted alleles gave rise to 2.7-kb and 3.3-kb bands, respectively. (**C**) Immunoblot analysis of protein lysates prepared from bone marrow cells of *Smap1*$^{+/+}$, *Smap1*$^{+/-}$, and *Smap1*$^{-/-}$ genotypes. The 50-kDa band represents SMAP1, whereas SMAP2 served as a control. Three independent experiments were performed, and one representative reproducible result is shown. (**D**) RT-PCR analyses of *Smap1* transcripts in bone marrow cells from *Smap1*$^{+/+}$ and *Smap1*$^{-/-}$ mice. Primers were set between exons 1 or 3 and exon 9. *HPRT* served as a control. (**E**) *SMAP1* expression in hematopoietic cells. Fractions of various hematopoietic lineages were sorted from bone marrow cells of wild-type mice by flow cytometry, and RNA was prepared from each and processed for semiquantitative RT-PCR analyses. DDW, distilled deionized water; -RT, without reverse transcription.

Results

Establishment of Smap1-targeted mice and SMAP1 expression. The functions of *SMAP1* in vivo were analyzed using a gene targeting approach. Figure 1A illustrates the genomic structure of *SMAP1* around exon 1 and the configuration of the targeting vector. Exon 1 was chosen as the targeting site because it harbors the SMAP1-initiating methionine codon. Two independent *Smap1*$^{+/-}$ mouse lines (44 and 64) were established and crossed to each other to generate *Smap1*$^{-/-}$ mice. Genomic DNA was processed for Southern blot analysis (Figure 1B). Based on the size of the detected bands, mouse genotypes were determined as wild-type, heterozygous, or homozygous targeting.

To confirm the expression of SMAP1, protein extracts from bone marrow cells were analyzed by immunoblotting (Figure 1C). SMAP1 was detected in wild-type cells but was substantially reduced in heterozygous cells and not detected in homozygous targeted cells. SMAP2, a homolog of SMAP1 (16), was detected in equal amounts in the 3 cell types. RT-PCR analyses did not detect *SMAP1* transcripts spanning exons 1 or 3 through to exon 9 in the *Smap1*$^{-/-}$ cells (Figure 1D). Thus, homozygous targeting was confirmed to correspond to a *Smap1*-deficient status. *Smap1*$^{+/-}$ mice exhibited no particular abnormality, and *Smap1*$^{-/-}$ mice also grew to adulthood and were apparently healthy. Both male and female *Smap1*$^{-/-}$ mice were fertile, and pups were born following the Mendelian ratio of inheritance.

SMAP1 expression was examined in various hematopoietic lineages isolated from the bone marrow of wild-type mice, as this information is relevant to the phenotypes of targeted mice, as described below. RT-PCR analyses (Figure 1E) showed that a substantial amount of *SMAP1* transcript was detected in the MEP (megakaryo/erythroid progenitor) and Ter119$^+$ fractions, whereas a lower amount was detected in the remaining fractions. This indicates that *SMAP1* is expressed abundantly in the erythroid lineage but is also distributed broadly in the other hematopoietic lineages, including progenitors.

Endocytosis of transferrin is enhanced in Smap1-targeted cells. The identification of SMAP1 as an ARF6 GAP was based on the effects of SMAP1 overexpression on the endocytosis of the transferrin receptor using tissue culture cells (5, 6). Here, we examined whether

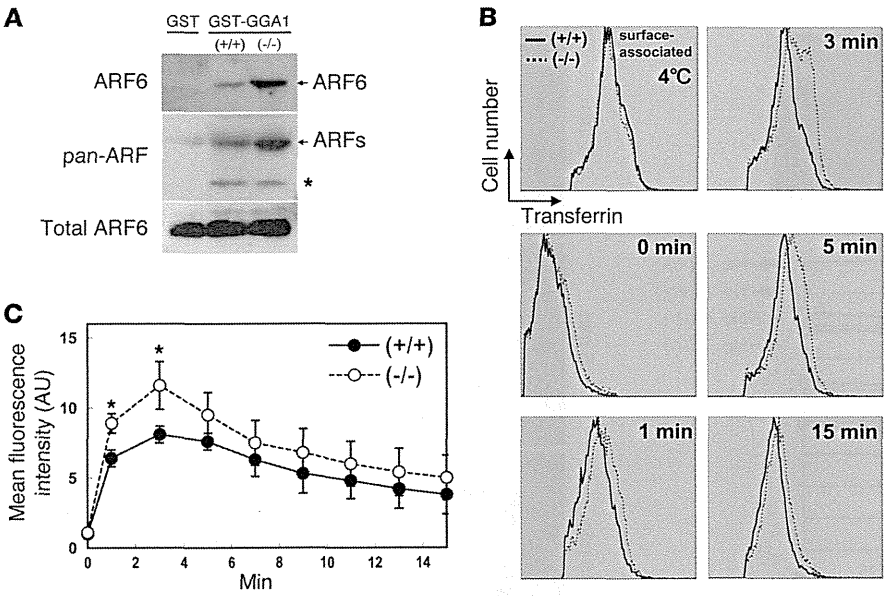

Figure 2

ARF6 activation and transferrin endocytosis in bone marrow cells. (**A**) Protein lysates were prepared from *Smap1*$^{+/+}$ and *Smap1*$^{-/-}$ bone marrow cells and incubated with GST or GST-GGA1 coupled to glutathione-Sepharose. The bound fraction was processed for immunoblot detection by anti-ARF6–specific and anti-panARF antibodies, as indicated. An asterisk represents nonspecific bands. The amounts of ARF6 in each lysate prior to incubation with GST or GST-GGA1 were also evaluated by immunoblotting (see "Total ARF6"). (**B** and **C**) Bone marrow cells were prepared from *Smap1*$^{+/+}$ and *Smap1*$^{-/-}$ mice and labeled with fluorescein-transferrin at 4°C. Excessive transferrin in the medium was washed away (initially bound transferrin at this time is shown as "4°C" as indicated in the top left of **B**), and, after incubation of cells at 37°C for the indicated time, surface-remaining transferrin was stripped off. Cells were labeled with PE-anti-Ter119 and processed for flow cytometry. The Ter119^{+} fraction was gated, and the transferrin-derived fluorescence intensities are displayed. Relative amounts of internalized fluorescein were measured by comparing fluorescence intensities at 0 minutes and each given time. Cells were prepared from 3 independent pairs of *Smap1*$^{+/+}$ and *Smap1*$^{-/-}$ mice and processed for assays. Averages ± SD of internalized transferrin were calculated for each incubation time at 37°C ($n = 3$). *$P < 0.05$.

SMAP1 functions as an ARF6 GAP in mouse tissues. Figure 2A shows the incubation of protein lysates from bone marrow cells with the GST-GGA1 fusion protein. GGA1 is a clathrin-adaptor protein that binds GTP-bound forms but not GDP-bound forms of ARF (17). The GST-GGA1-bound fraction was processed for immunoblot analysis using anti-ARF6 or anti-panARF antibodies, which showed that the amount of GGA1-bound ARF6 was increased by 3.1 fold in *Smap1*$^{-/-}$ cells as compared with that in wild-type cells. Expression of ARF6 itself was not affected by *SMAP1* targeting, as shown by the immunoblotting of lysates prior to the application of GST-GGA1. These results indicate that SMAP1 indeed functions as an ARF6 GAP in vivo.

Erythroblasts, which show highly active transferrin endocytosis, were used to examine the effect of *SMAP1* targeting on ARF6-regulated endocytosis. Bone marrow cells were incubated with transferrin at 4°C, excessive transferrin was washed away, and the cells were incubated at 37°C for various time periods. Then, the remaining surface-bound transferrin was stripped off, leaving only the intracellularly incorporated molecules intact. Figure 2B depicts a time course of transferrin fluorescence intensity that was obtained by gating the Ter119^{+} erythroid cell fraction. The fluorescent intensities were quantified and are shown in Figure 2C. Although no differences were detected during the recycling phase (after 5 minutes), a significant increase in the amount of transferrin was incorporated into *Smap1*$^{-/-}$ cells compared with wild-type cells during the initial uptake at 1 and 3 minutes. Notably, prior to the incubation at 37°C, amounts of the initially cell surface–bound transferrin at 4°C were similar between the 2 genotypes of Ter119^{+} bone marrow cells (Figure 2B, top left).

Transferrin endocytosis in Smap1-targeted cells is mediated by SMAP2. The effect of *SMAP1* deficiency on transferrin endocytosis was investigated in cells of different lineages. Two independent wild-type and *Smap1*$^{-/-}$ mouse embryo fibroblast (MEF) cultures were established. Figure 3A shows the immunofluorescence of endogenous SMAP1 on the cell surface and, as multiple dots in the cytoplasm, of wild-type cells but not targeted cells.

MEFs were incubated with transferrin for various time periods and then washed and fixed. Figure 3B depicts the fluorescence signals derived from internalized transferrin and shows that the intensity of fluorescence is stronger in *Smap1*$^{-/-}$ cells than in wild-type cells. MEFs were recovered as a suspension and processed for flow cytometry. Figure 3C shows the gradual accumulation of transferrin in the cytoplasm. Transferrin accumulation was 1.3- to 1.9-fold more effective in the targeted cells as compared with that in the wild-type cells. When endocytosis and recycling were separately assayed using biotinylated transferrin, internalization was enhanced but recycling was not affected in *Smap1*$^{-/-}$ cells as compared with that in wild-type cells (Supplemental Figure 1; supplemental material available online with this article; doi:10.1172/JCI63711DS1), indicating that the enhanced accumulation of transferrin in *Smap1*$^{-/-}$ MEFs (Figure 3C) is likely due to the enhanced incorporation of molecules (Supplemental Figure 1). Note that the fluorescence intensity of transferrin initially bound to the cell surface was similar in wild-type and *Smap1*$^{-/-}$ MEFs (Supplemental Figure 1A).

We then examined why transferrin endocytosis was not abrogated in *Smap1*$^{-/-}$ cells. The effectiveness of siRNAs against SMAP2 was tested using wild-type MEFs (Figure 3D), and immunoblot analysis showed that siRNA2 worked more efficiently. Figure 3E shows the internalization of transferrin and SMAP2 levels in wild-type and *Smap1*$^{-/-}$ MEFs after siRNA2-mediated silencing of SMAP2. Interestingly, the effects of siRNA2 appeared random and differed among cells, because endogenous SMAP2 remained intact in some cells, whereas it was almost abolished in other cells. Under these conditions, and in the case of *Smap1*$^{+/+}$ MEFs, transferrin was equally incorporated regardless of the levels of SMAP2. In contrast, in *Smap1*$^{-/-}$ MEFs, transferrin was not incorporated in SMAP2-silenced cells. These results suggest that SMAP2 likely compensates for the lack of SMAP1 in *Smap1*$^{-/-}$ MEFs.

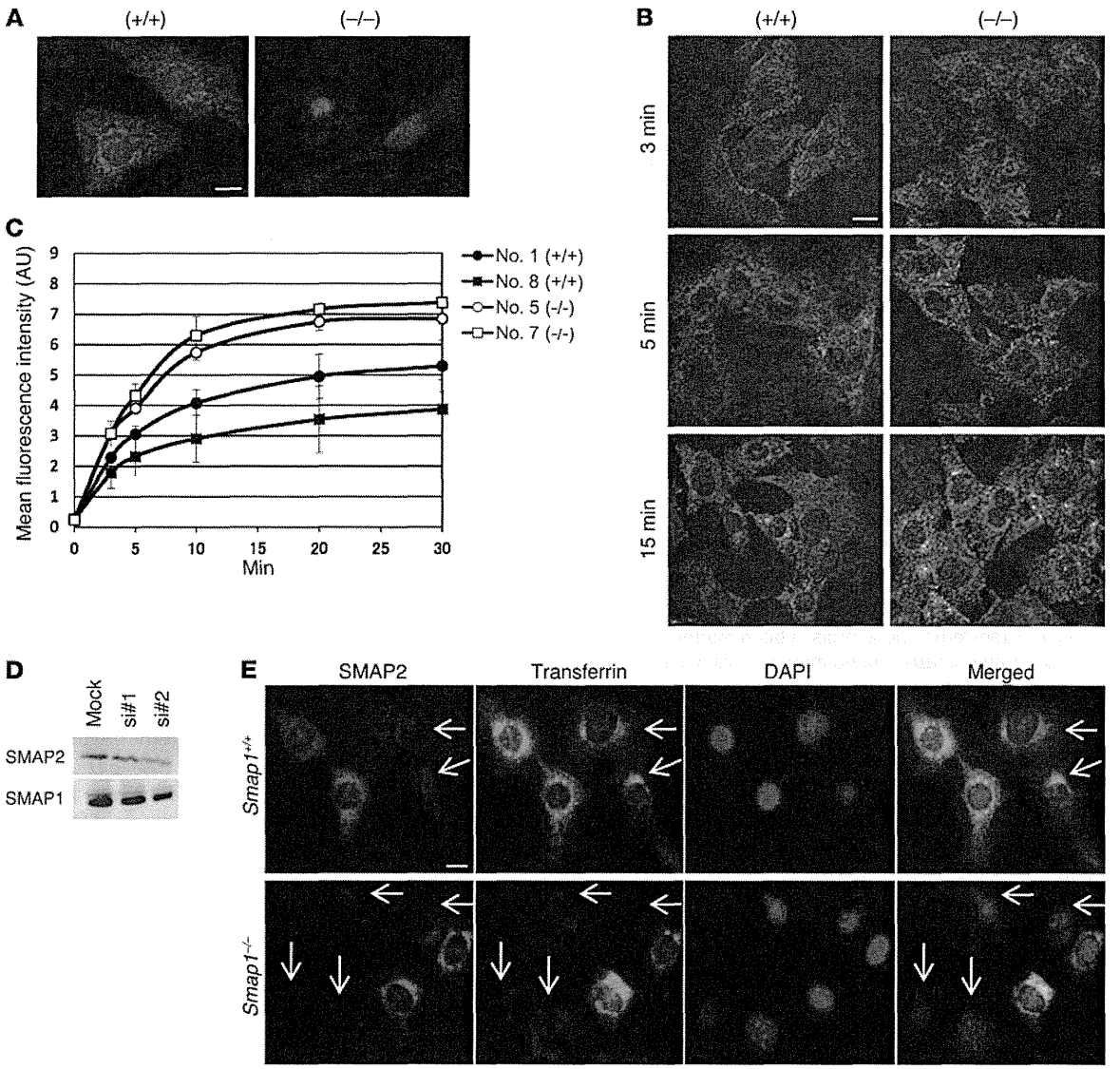

Figure 3

Transferrin transport in MEFs. (**A**) Immunofluorescence detection of endogenous SMAP1 in wild-type and *Smap1*$^{-/-}$ MEFs using an anti-SMAP1 antibody (green). Blue indicates DAPI staining. (**B** and **C**) Internalization of transferrin in wild-type and *Smap1*$^{-/-}$ MEFs. Cells were incubated with fluorescein-transferrin for the indicated times, and then surface-remaining transferrin was stripped off. The cells were then processed for analyses by (**B**) fluorescence microscopy or (**C**) flow cytometry. In **C**, the intensities of intracytoplasmic fluorescence were measured and expressed in relative arbitrary units. Independent cultures were prepared in triplicate from the indicated MEF clones, and averages ± SD are shown (n = 3). "No. 1," "No. 8," "No. 5," and "No. 7" refer to the MEF cell line numbers. (**D**) Effects of siRNA against SMAP2. Wild-type MEFs were treated with or without siRNA against SMAP2 (2 differentially designed siRNAs, siRNA1 and siRNA2, were used). Protein lysates were prepared and processed for immunoblot analyses using anti-SMAP2 and anti-SMAP1 antibodies. (**E**) Effects of SMAP2 knockdown on transferrin incorporation. The *Smap1*$^{+/+}$ and *Smap1*$^{-/-}$ MEFs were incubated with siRNA2 against SMAP2 and then with fluorescent transferrin and processed for immunofluorescence detection using an anti-SMAP2 antibody. The arrows indicate the reduction in fluorescence intensity from SMAP2, whereas DAPI staining indicates the location of cell nuclei. Scale bar: 10 μm.

Accumulation and enhanced signaling of c-KIT in Smap1$^{-/-}$ *cells.* c-KIT is highly expressed in hematopoietic progenitors and mast cells and is internalized through clathrin-coated vesicles. Because *SMAP1* was detected in both types of cells (Figure 1E), the effects of *SMAP1* targeting on c-KIT internalization were examined. Bone marrow–derived mast cells (BMMCs) were prepared and incubated with stem cell factor (SCF), and cell surface–located c-KIT was measured by flow cytometry (Figure 4A). Cycloheximide was added to prevent the de novo synthesis of c-KIT, thereby preventing its expression on the cell surface. The top panel of Figure 4A shows the fluorescence intensity of cell surface c-KIT, as detected by anti–c-KIT, and the bottom panel of Figure 4A shows the percentage of internalized c-KIT. No difference was detected between the 2 genotypes, indicating that the SCF-induced endo-

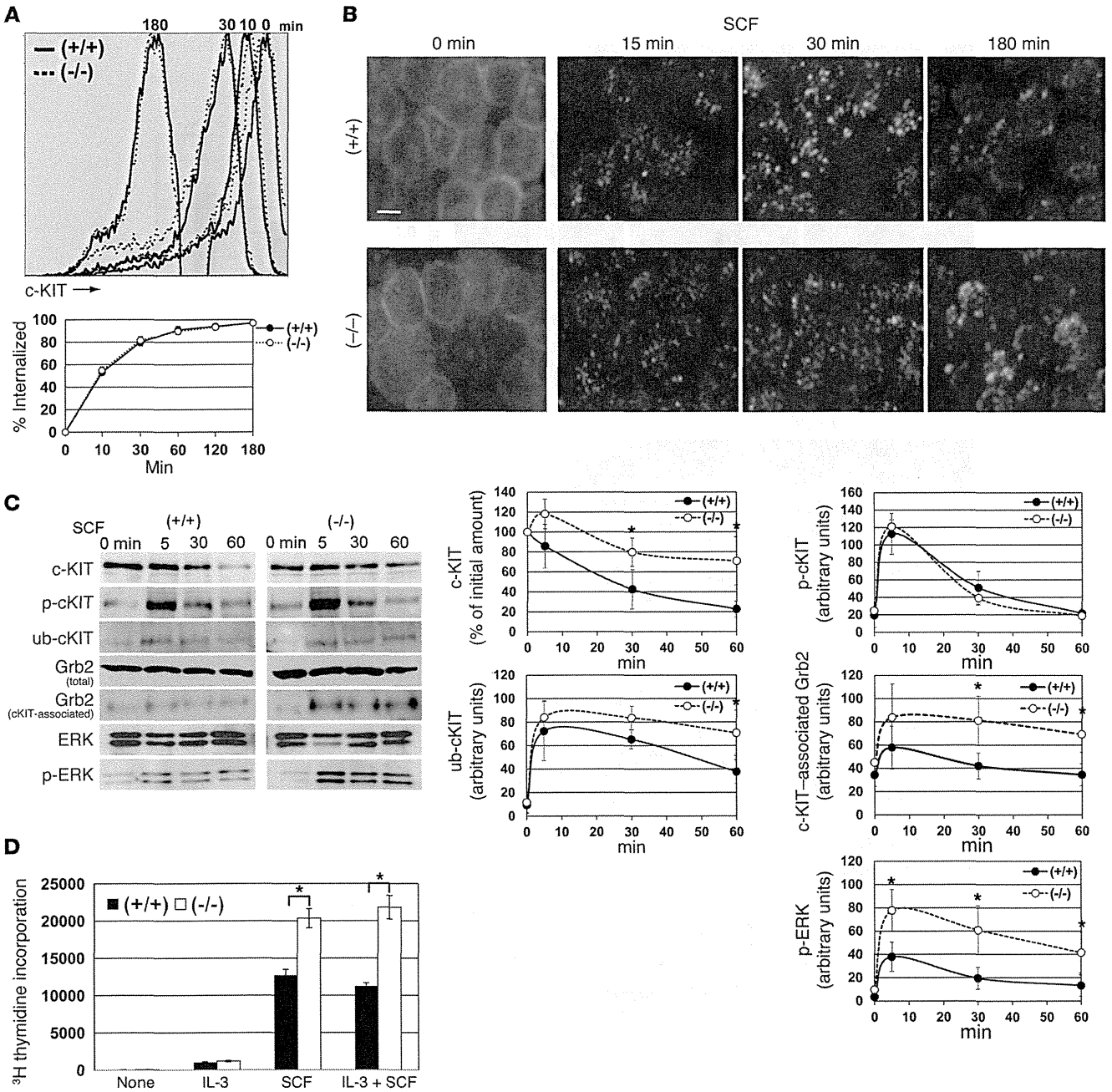

Figure 4

Transport kinetics and c-KIT signaling in BMMCs. (**A**) Endocytosis of c-KIT. *Smap1*$^{+/+}$ and *Smap1*$^{-/-}$ BMMCs were cultured, starved in the presence of cycloheximide, incubated with SCF at 37°C for the indicated times, and processed for flow cytometry analyses. The top panel displays the fluorescence intensity of c-KIT and cell numbers, whereas the bottom panel plots the percentages of internalized c-KIT calculated by considering the initial surface fluorescence to be 100%. BMMCs were prepared from 3 independent pairs of *Smap1*$^{+/+}$ and *Smap1*$^{-/-}$ mice and processed for assays. Averages ± SD of internalized c-KIT were calculated for each incubation time (n = 3). (**B**) Immunofluorescence detection of c-KIT in BMMCs. The *Smap1*$^{+/+}$ and *Smap1*$^{-/-}$ cells were incubated in the presence of SCF for the indicated times and stained for c-KIT. Scale bar: 10 μm. (**C**) Activation status of c-KIT signaling molecules. Wild-type and *Smap1*$^{-/-}$ BMMCs were incubated with SCF for the indicated times, and protein lysates were prepared and processed for immunoprecipitation/immunoblot analyses. Band densities were quantified, and averages ± SD are shown (n = 3). p-c-KIT, phosphorylated form of c-KIT; p-ERK1/2, phosphorylated form of ERK1/2; ub-c-KIT, ubiquitinylated c-KIT; c-KIT-associated Grb2, Grb2 recruited into anti–c-KIT immunoprecipitates. (**D**) DNA synthesis in BMMCs. Triplicate cultures of cells were prepared from each of the wild-type and *Smap1*$^{-/-}$ mice, incubated in the presence of IL-3 and/or SCF for 16 hours, and then treated with ^{3}H-thymidine for 8 hours. The incorporation of ^{3}H-thymidine into acid-insoluble fractions was measured, and averages ± SD are shown (n = 3). *P < 0.05.

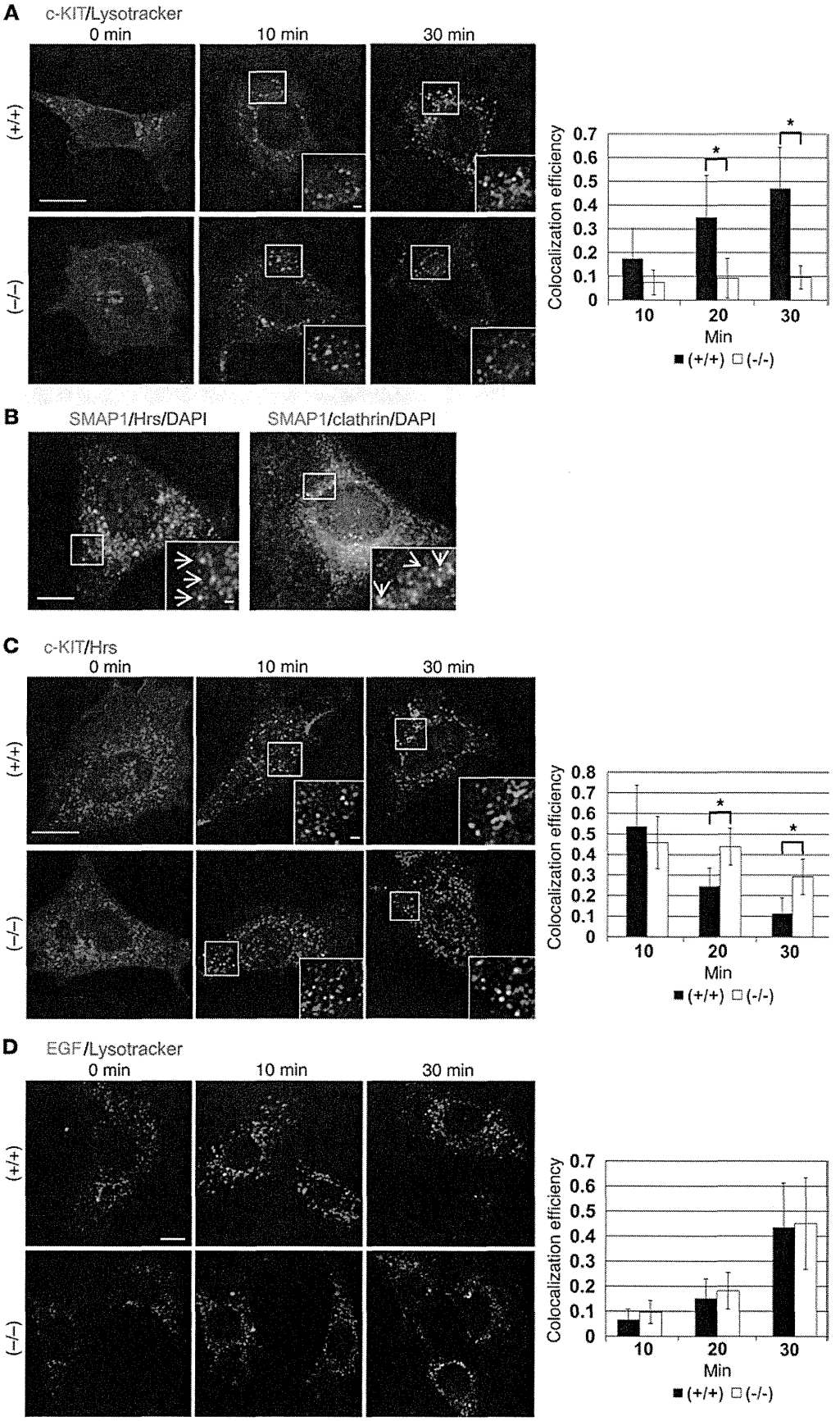
A
c-KIT/Lysotracker
0 min
10 min
30 min
(+/+)
(−/−)
Colocalization efficiency
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
10
20
30
Min
■ (+/+) □ (-/-)
B
SMAP1/Hrs/DAPI
SMAP1/clathrin/DAPI
C
c-KIT/Hrs
0 min
10 min
30 min
(+/+)
(−/−)
Colocalization efficiency
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
10
20
30
Min
■ (+/+) □ (-/-)
D
EGF/Lysotracker
0 min
10 min
30 min
(+/+)
(−/−)
Colocalization efficiency
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
10
20
30
Min
■ (+/+) □ (-/-)

Figure 5
Transport of c-KIT and EGFR in MEFs. (**A** and **C**) Wild-type and *Smap1*$^{-/-}$ MEFs were transfected by EYFP-c-KIT, incubated with SCF for the indicated times, and processed for double-fluorescence detection of (**A**) c-KIT and lysotracker or (**C**) c-KIT and Hrs. (**D**) MEFs were incubated with dye-conjugated EGF for the indicated times and processed for double-fluorescence detection of EGF and lysotracker. In **A**, **C**, and **D**, the colocalization of the 2 molecules was analyzed and plotted as histograms for the indicated incubation period. Data are shown as averages ± SD (n = 50–70). Reproducible results were obtained for 2 independent *Smap1*$^{-/-}$ MEF cultures. *P < 0.05. (**B**) Double-immunofluorescence detection of endogenous SMAP1 and the indicated organelle marker in wild-type MEFs. The nuclei were stained by DAPI. The arrows in insets indicate colocalization of SMAP1 with Hrs or clathrin. Scale bars: 10 μm; 1 μm (insets).

cytosis of cell surface c-KIT was not affected by *SMAP1* targeting, contrary to the effect on transferrin endocytosis.

Immunofluorescence analysis using anti–c-KIT antibodies (Figure 4B) showed that, prior to the addition of SCF, c-KIT was similarly detected on the cell surface in both types of cells. Fifteen and thirty minutes after addition of SCF, internalized c-KIT was detected in a punctate pattern in the cytoplasm. Prolonged incubations for up to 180 minutes resulted in the disappearance of the c-KIT signal in wild-type cells, whereas it was still clearly visible in the targeted cells, suggesting that the downregulation of internalized c-KIT might be delayed in the *Smap1*-targeted BMMCs.

Immunoblot analyses (Figure 4C) confirmed this effect by showing the c-KIT protein at comparable levels in both genotypes before addition of SCF and a significant downregulation of the protein 60 minutes after addition of ligand in the wild-type cells but not the targeted cells. Immunoblotting and immunofluorescence results collectively indicate that although c-KIT endocytosis was not affected by *SMAP1* targeting, the downregulation of internalized c-KIT was delayed in the targeted BMMCs.

Then, we examined whether c-KIT remaining in targeted cells was capable of transmitting growth signals to downstream molecules. SCF binding triggers tyrosine phosphorylation of c-KIT, which is followed by monoubiquitination and Grb2 association. Ubiquitination and Grb2 association are the necessary events leading to endocytosis of c-KIT and signal transmission to ERK1/2, respectively. As seen in Figure 4C, the induction and downregulation of c-KIT phosphorylation and the levels of Grb2 and ERK1/2 did not differ between the 2 genotypes. On the other hand, c-KIT ubiquitination, Grb2 association with c-KIT, and ERK1/2 phosphorylation increased 2 fold in *Smap1*$^{-/-}$ BMMCs as compared with that in the wild-type cells (see the quantification of band densities in Figure 4C, right panels). Figure 4D shows the incorporation of ^{3}H-thymidine into an acid-insoluble fraction and demonstrates that SCF treatment induced DNA synthesis and a 1.6-fold increase in ^{3}H-thymidine incorporation in the targeted cells as compared with that in the wild-type cells. Collectively, the above results indicate that, in the presence of SCF, *Smap1*-targeted BMMCs tend to accumulate c-KIT in the cytoplasm, resulting in enhanced signaling and cell growth activity.

Sorting of c-KIT to lysosomes is delayed in Smap1$^{-/-}$ *MEFs.* Ligand-engaged and internalized c-KIT is transported first to early endosomes and then transits through multivesicular bodies (MVBs) and finally to lysosomes, in which the protein is degraded by digestive enzymes (18, 19). Because the persistent accumulation of c-KIT in the cytoplasm of *Smap1*$^{-/-}$ BMMCs suggests an alteration in the transport pathway, the intracellular trafficking of c-KIT was examined in MEFs.

As shown in Figure 5A, c-KIT was detected on the cell surface prior to SCF stimulation (see 0 minutes) and then internalized into the cytoplasm after 10 minutes of SCF treatment in both *SMAP1* genotypes. However, after 30 minutes, a substantial fraction of c-KIT colocalized with lysotracker in the wild-type MEFs but not in targeted MEFs. The colocalization efficiency of the 2 molecules was quantified, and the result is shown as a histogram (Figure 5A). In *Smap1*$^{-/-}$ cells, although c-KIT was incorporated into the cytoplasm upon SCF addition, its transport to lysosomes appeared impaired.

To identify the specific step in the transport of c-KIT that was affected by *SMAP1* targeting, wild-type MEFs were costained for endogenous SMAP1 and various organelle markers. SMAP1 fluorescence did not overlap with that of EEA1, Rab5, Rab11, and LBPA, and no colocalization with lysotracker was observed (data not shown). However, SMAP1 showed partial colocalization with Hrs, an MVB marker (Figure 5B). Substantial colocalization of SMAP1 and clathrin was as previously reported (5, 6). These obser-

Table 1
Peripheral blood counts in *SMAP1*$^{-/-}$ mice

Genotype	No. of mice	rbc (10^4/μl)	Hematocrit (%)	Hemoglobin (g/dl)	MCV (fl)	MCH (pg)	Reticulocytes (%)	PLT (10^4/μl)	wbc (10^2/μl)
Smap1$^{+/+}$	24	1,002 ± 49	45.2 ± 1.8	14.8 ± 0.6	44.8 ± 1.0	14.7 ± 0.3	4.7 ± 0.9	145.2 ± 21.9	108 ± 27
Smap1$^{-/-}$ (nonanemic)	16	1,012 ± 81	45.8 ± 2.0	15.1 ± 0.7	45.5 ± 2.8	14.9 ± 0.9	4.1 ± 1.2	158.6 ± 47.6	123 ± 59
Smap1$^{-/-}$ (anemic MDS)	10	704 ± 120^{A}	35.8 ± 5.6^{A}	11.0 ± 2.1^{A}	51.4 ± 3.6^{A}	15.3 ± 1.1	16.8 ± 9.0^{B}	90.5 ± 46.9^{B}	101 ± 39
Smap1$^{-/-}$ (MPD/MDS)	2	777 ± 69^{A}	42.1 ± 4.3	14.0 ± 1.6	54.2 ± 0.8^{A}	18.0 ± 0.5^{A}	nd	45.4 ± 43.1^{A}	201 ± 11^{A}
Smap1$^{-/-}$ (AML)	5^{C}	752 ± 93^{A}	36.5 ± 4.0^{B}	12.0 ± 1.1^{A}	48.7 ± 2.9	15.6 ± 1.1	8.0 ± 3.5	150.0 ± 97.5	164 ± 13^{B}

Statistically significant differences were detected between *Smap1*$^{+/+}$ and *Smap1*$^{-/-}$ mice by Student's t test (^{A}P < 0.001, ^{B}P < 0.01). ^{C}Note that, out of 5 AML-suffering mice, 3 mice were examined for their peripheral blood counts (see Supplemental Table 1 as well). MCV, mean corpuscular volume; MCH, mean corpuscular hemoglobin; PLT, platelets; nd, not determined.