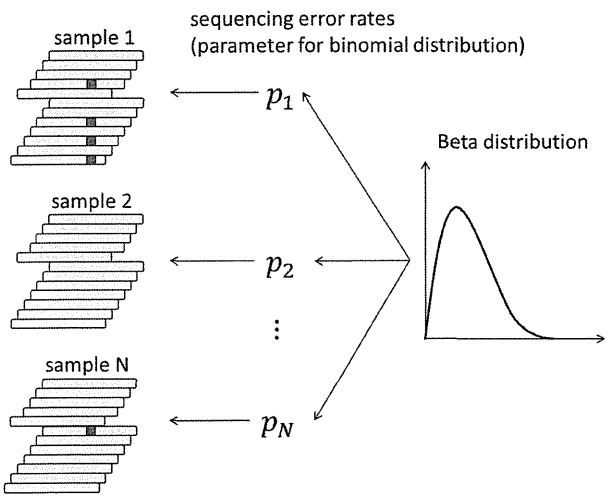


**Figure 3.** A Beta-binomial sequencing error model. First, the error rate for each sample is generated from the Beta distribution. The number of short reads with sequencing errors is then generated according to the binomial distribution using the parameters of the above error rate for each sample. 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 $\Gamma$ 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 $\chi_4^2$ 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 $p^{\text{Fisher}}$ 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.

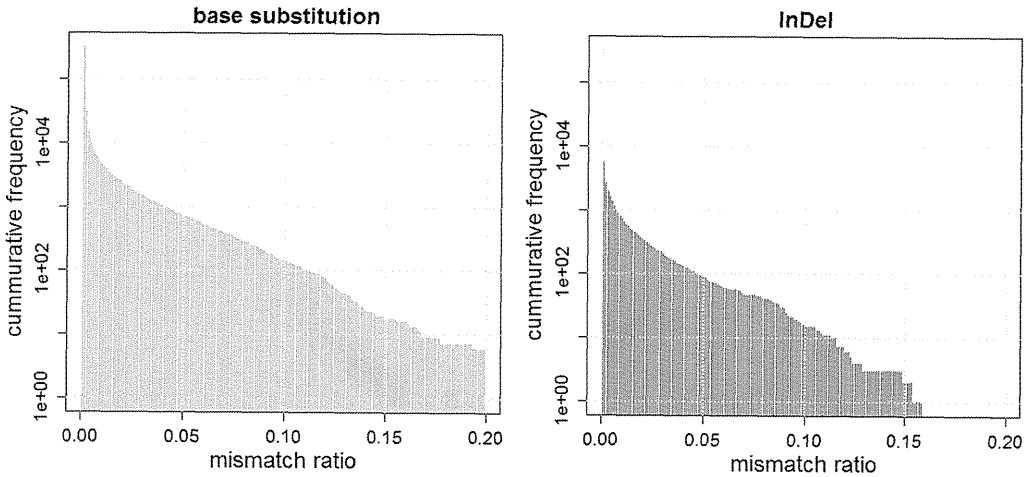


**Figure 4.** Two bar plots showing the numbers of base substitutions and InDels, whose mean mismatch ratios are above the determined threshold values. For instance, the numbers of base substitutions with mean mismatch ratios of more than 0.01, 0.02, and 0.05 are 4472, 2232, and 727, respectively, while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

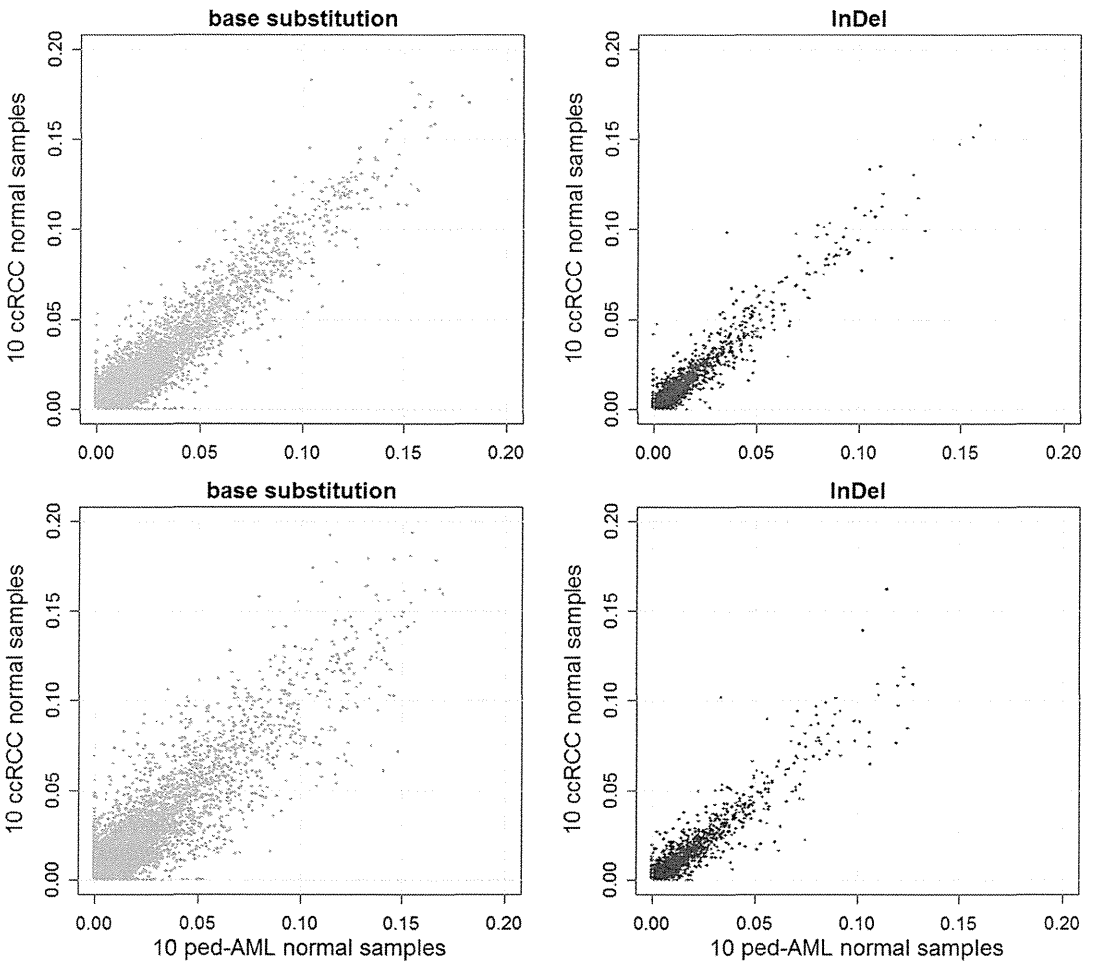


**Figure 5.** A comparison of scatter plots of the mean mismatch ratios of the base substitution and InDels for two sets consisting of 10 ccRCC normal samples each (upper), and 10 ccRCC normal samples and 10 ped-AML normal samples (lower). The correlation coefficients are 0.777, 0.723, 0.943 and 0.917 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. $6.54 \times 10^{-4}$ for *EBCall*, $1.97 \times 10^{-3}$ for *VarScan*, 60 for *SomaticSniper* and $5.85 \times 10^{-3}$ 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.

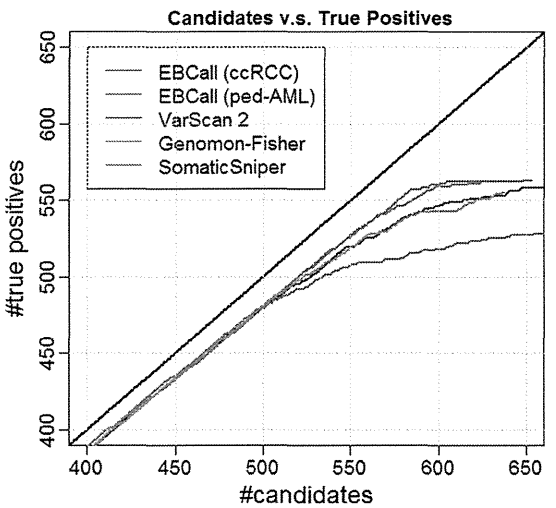


**Figure 6.** Comparative performance for *EBCall* (20 ccRCC or ped-AML normal samples used as normal reference sets), *Genomon-Fisher*, *VarScan 2* and *SomaticSniper*. The horizontal and vertical axes show the number of candidate somatic mutations and true positives (when changing the threshold of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

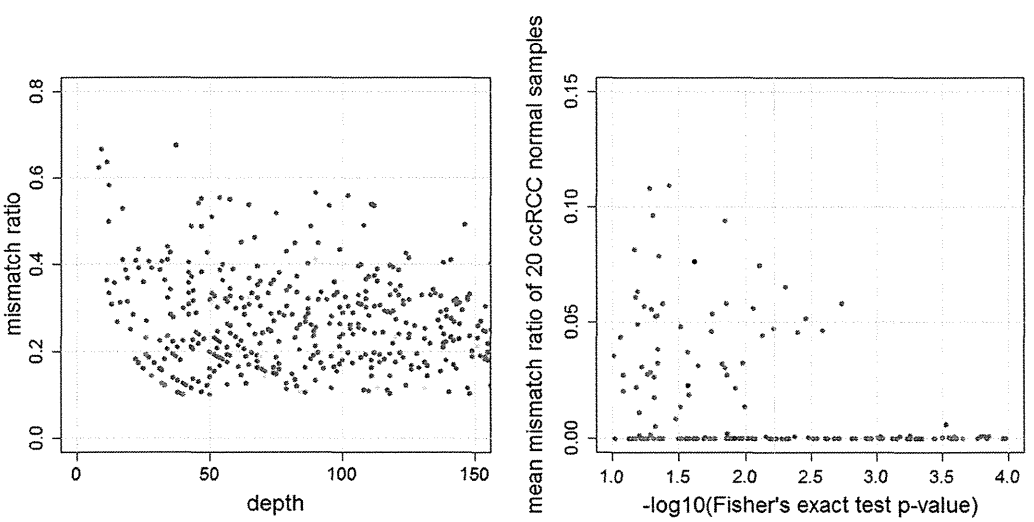


**Figure 7.** (Left) The comparative results between *EBCall* and *Genomon-Fisher*. Each point, in which the sequencing depth and variant allele frequency are indicated, shows the candidate somatic mutations called by both or either of the two methods. The threshold values are determined such that the 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 |

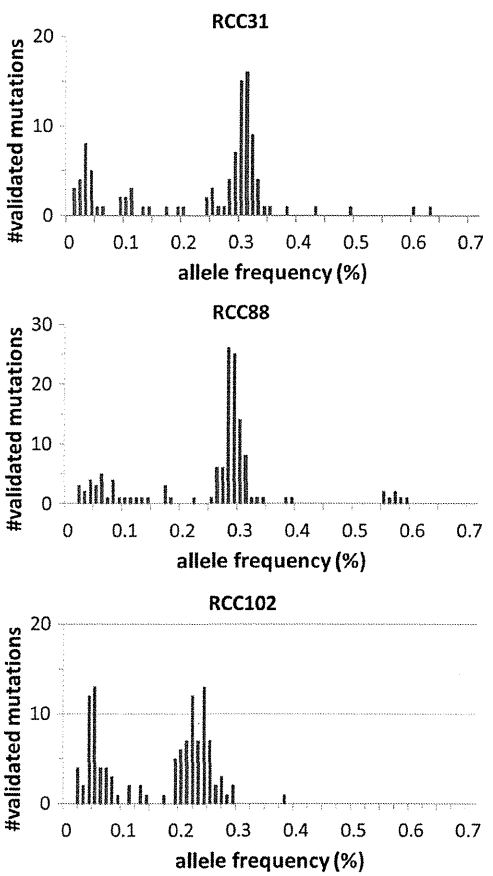


**Figure 8.** Histograms of the allele frequencies of validated mutations for RCC31 (left), RCC88 (centre) and RCC102 (right).

## 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.

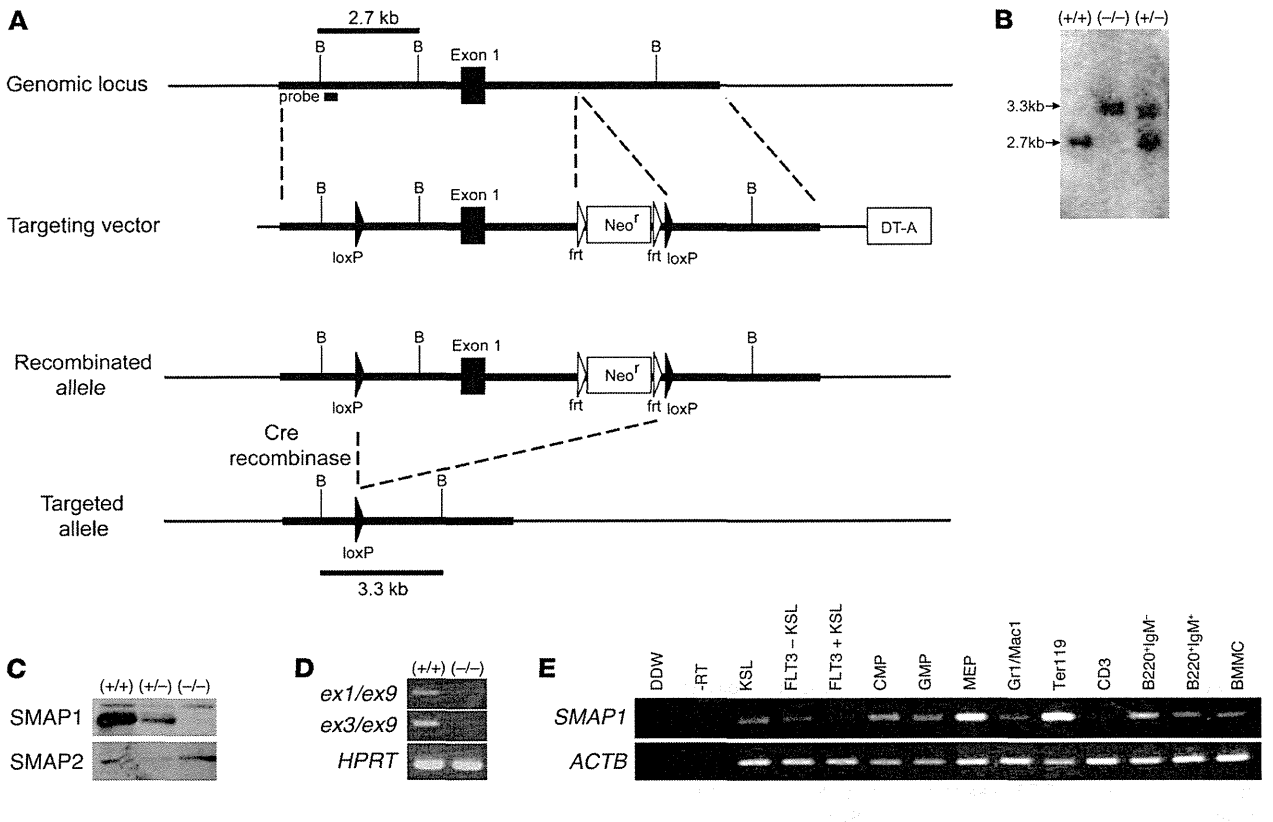


**Figure 1**
Establishment of *Smap1*-targeted mice and SMAP1 expression. (**A**) Physical maps of the *SMAP1* gene locus and its targeting vector. Features of the recombined and targeted *SMAP1* alleles are also depicted. Horizontal lines indicate the genomic sequences. The thick lines indicate the 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

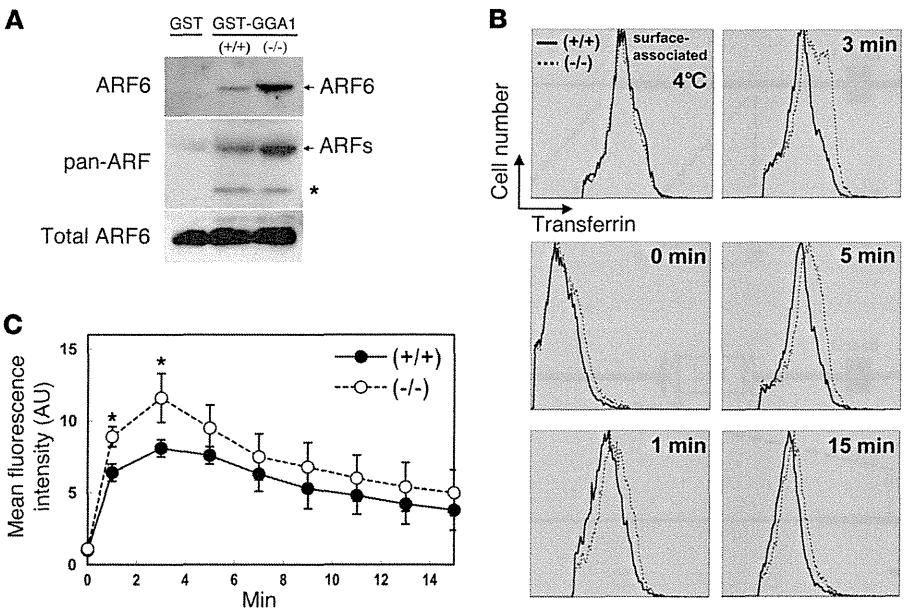


**Figure 2**

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.

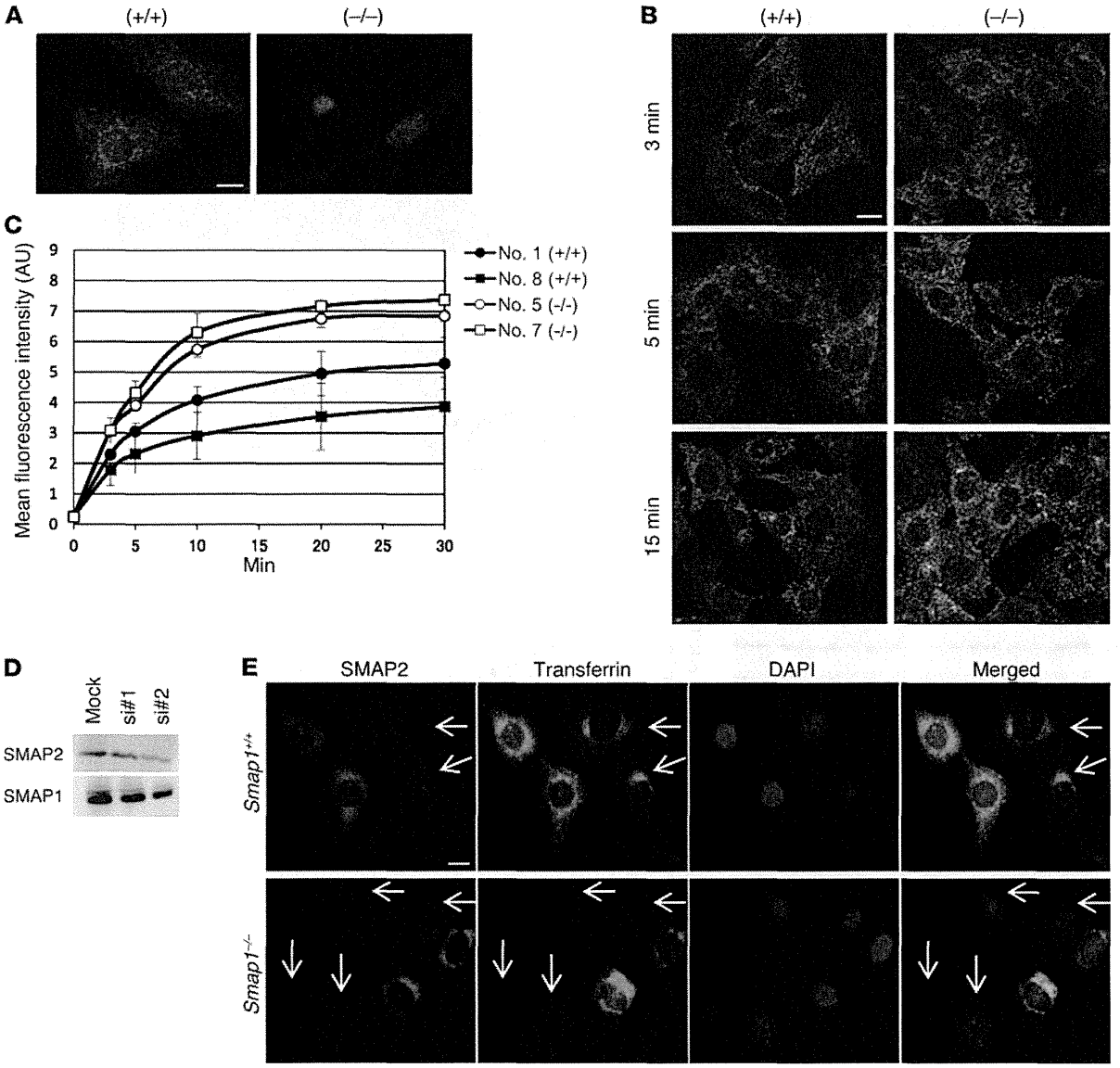


**Figure 3**

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-

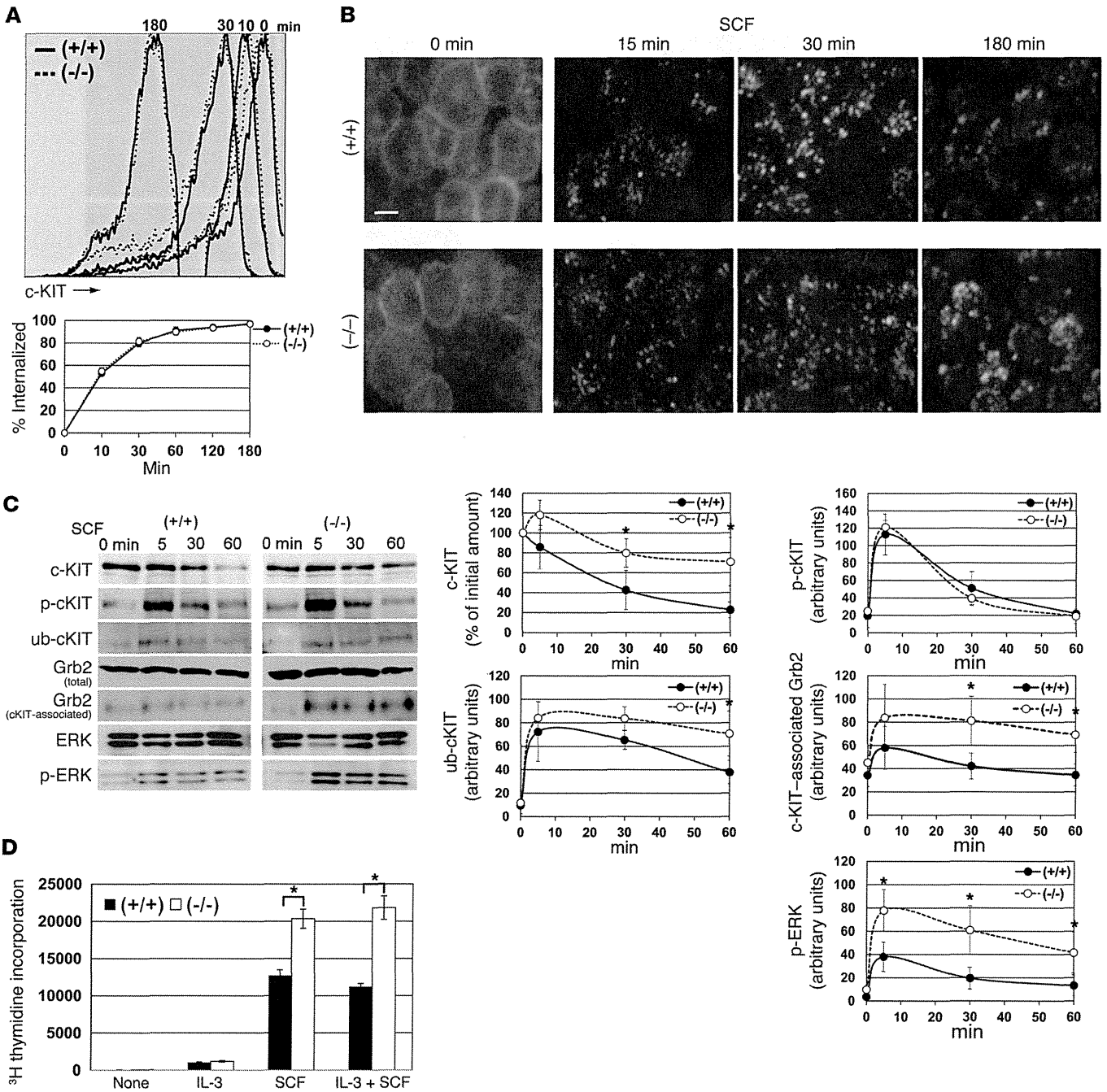


**Figure 4**

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.

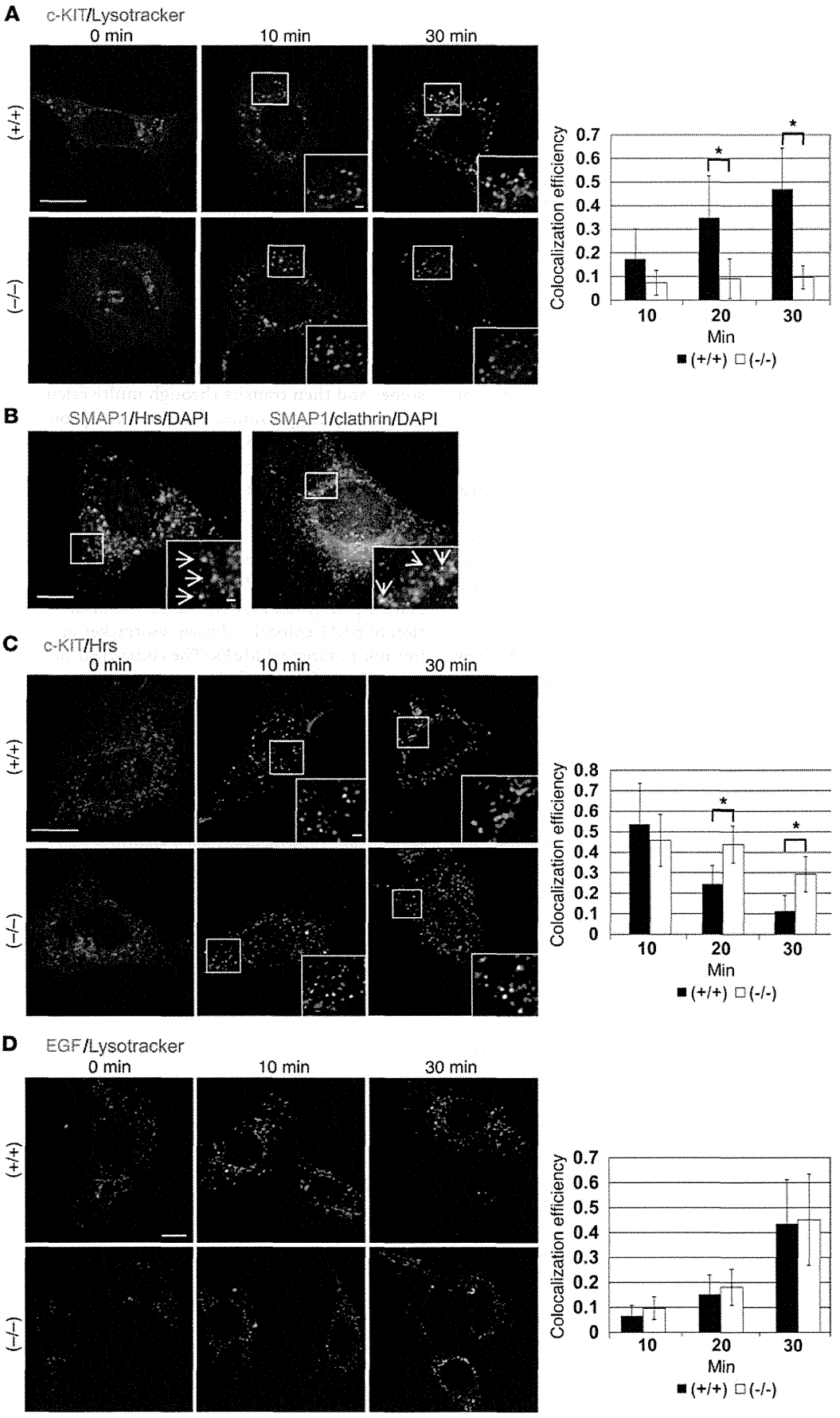
A
c-KIT/Lysotracker
0 min
10 min
30 min
(+/+)
(-/-)
Colocalization efficiency
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
10
20
30
Min
■ (+/+) □ (-/-)
B
SMAP1/Hrs/DAPI
SMAP1/clathrin/DAPI
C
c-KIT/Hrs
0 min
10 min
30 min
(+/+)
(-/-)
Colocalization efficiency
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
10
20
30
Min
■ (+/+) □ (-/-)
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 ($^AP$ < 0.001, $^BP$ < 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.

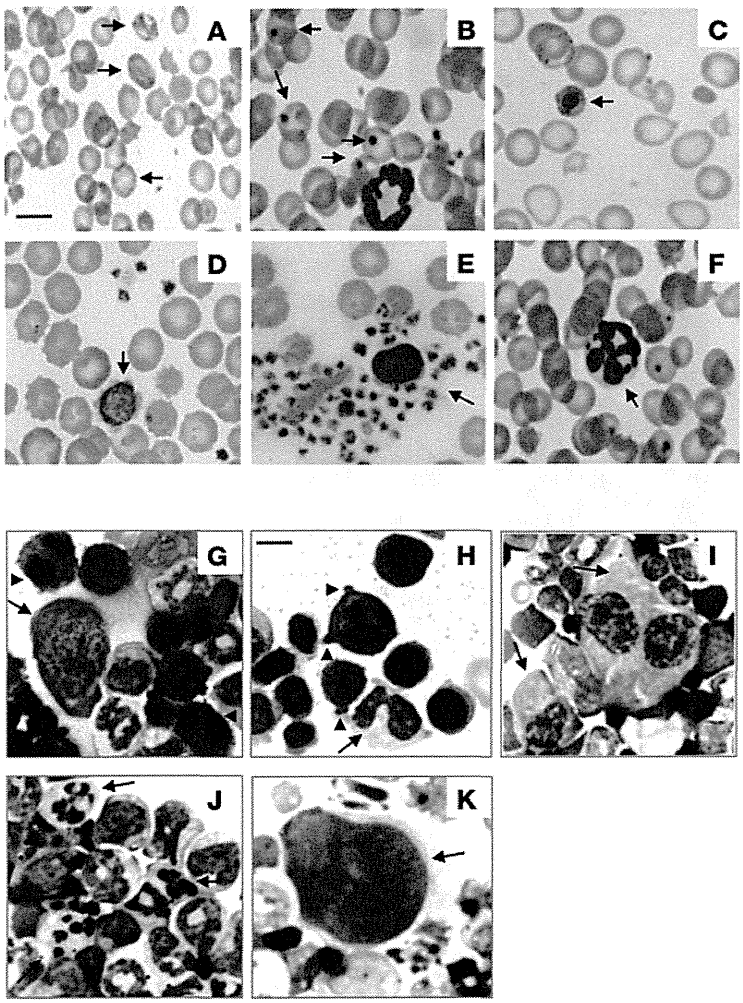


**Figure 6**
Morphology of peripheral blood and bone marrow cells from *Smap1*$^{-/-}$ mice. Smear samples were stained by May-Grunwald Giemsa. (**A**–**F**) In peripheral blood (PB), the cells are (**A**) polychromatic erythrocytes (arrows), (**B**) Howell-Jolly bodies (arrows), (**C**) erythroblasts (arrows), (**D**) giant platelets (arrow), (**E**) micromegakaryocytes (arrow), and (**F**) hypersegmented neutrophils (arrow). (**G**–**K**) In bone marrow cells (BM), the abnormal cells include (**G**) megaloblasts (arrow) and binucleated basophilic erythroblasts (arrowheads), (**H**) basophilic erythroblasts with cytoplasmic blebs (arrowheads) and pseudo-Pelger-Huet neutrophils (arrow), (**I**) multinucleated polychromatic megaloblasts, (**J**) orthochromatic erythroblasts with fragmented nuclei (arrows), and (**K**) megakaryocytes with hypolobulated nuclei. Scale bar: 10 μm.

vations suggest a possible involvement of SMAP1 in the MVB-to-lysosome transport pathway.

To verify this, the fate of internalized c-KIT was monitored in relation to Hrs (Figure 5C). After 10 minutes of SCF stimulation, a substantial proportion of internalized c-KIT was colocalized with Hrs, indicating the localization of c-KIT in MVBs. After 20 and 30 minutes, c-KIT colocalization with Hrs was not detected in wild-type MEFs but was present in *Smap1*$^{-/-}$ MEFs, indicating that the transport of c-KIT from the MVB to the lysosome was impaired in *Smap1*-targeted MEFs and possibly explaining the accumulation of internalized c-KIT in *Smap1*$^{-/-}$ BMMCs. The delay in the exit of c-KIT from late endosomes/MVBs in *Smap1*$^{-/-}$ MEFs was confirmed by another marker, Rab7 (Supplemental Figure 2A). However, the exit of c-KIT from early endosomes was not affected by *SMAP1* targeting, as shown by its colocalization with the early endosome marker Rab5 (Supplemental Figure 2B).

The possible involvement of ARF GTPase itself in the intracellular transport of c-KIT was examined in COS7 cells. Overexpression of the active form of ARF, which can mimic *SMAP1* deficiency, induced the substantial colocalization of c-KIT, Hrs, and ARF (Supplemental Figure 3). This implies the delay or block of c-KIT exit from MVBs. On the other hand, the EGF-induced transport of EGFR, another RTK, to the lysosome was not affected in *Smap1*$^{-/-}$ MEFs (Figure 5D). Therefore, an ARF/SMAP1 system appears to be functioning in the MVB-to-lysosome transport of c-KIT but not that of EGFR.

*Smap1*$^{-/-}$ *aged mice develop phenotypes similar to those of myelodysplastic syndrome in humans.* The results described above (Figures 2–5) were based on MEFs from embryos and bone marrow cells from 2- to 4-month-old mice, and despite the alterations in membrane traffic, *Smap1*$^{-/-}$ mice were healthy up to the age of 12 months old. When Kaplan-Meier curves were plotted from the 35-month observation period, no statistically significant differences were detected between the survival percentages of *Smap1*-targeted and wild-type mice, although homozygously targeted mice showed a tendency to die at a somewhat younger age (Supplemental Figure 4A). Notably, a substantially higher percentage of homozygously targeted mice showed ill-health conditions after 12 months (Supplemental Figure 4B), suggesting the development of age-related diseases.

In fact, certain mice older than 1 year developed hematological disorders, and 33 of these *Smap1*$^{-/-}$ mice older than 1 year were analyzed by measuring the number of peripheral blood cells in each individual mouse (Table 1 and Supplemental Table 1). Based on the number of rbc, *Smap1*$^{-/-}$ mice were categorized into nonanemic and anemic/myelodysplastic syndrome (anemic/MDS) groups (100% and 70% rbc count as compared with wild type, respectively).

The average hematocrit and hemoglobin values in the anemic *Smap1*$^{-/-}$ group were lower (70%–75%) than those in the nonanemic *Smap1*$^{-/-}$ group and wild-type mice. MCV values were significantly increased in the anemic group, whereas MCH values did not differ significantly between the 2 groups, indicating the presence of macrocytic and normochromatic anemia in approximately half of *Smap1*$^{-/-}$ mice. A remarkable increase in the number of reticu-

**Table 2**
Summary of the hematological diagnosis seen in mice

| Genotype of mice | No. of mice | Age (mo.) | Diagnosis (no.) | Spleen weight (g) |
|---|---|---|---|---|
| *Smap1*$^{+/+}$ | 24 | 12–25 | Nonanemic (24) | 0.11 ± 0.07 |
| *Smap1*$^{-/-}$ | 33 | 13–25 | Nonanemic (16) | 0.10 ± 0.05 |
| | | | MDS (10) | 0.23 ± 0.16 |
| | | | MPD/MDS (2) | 0.68 ± 0.37 |
| | | | AML (5) | 0.99 ± 0.59 |

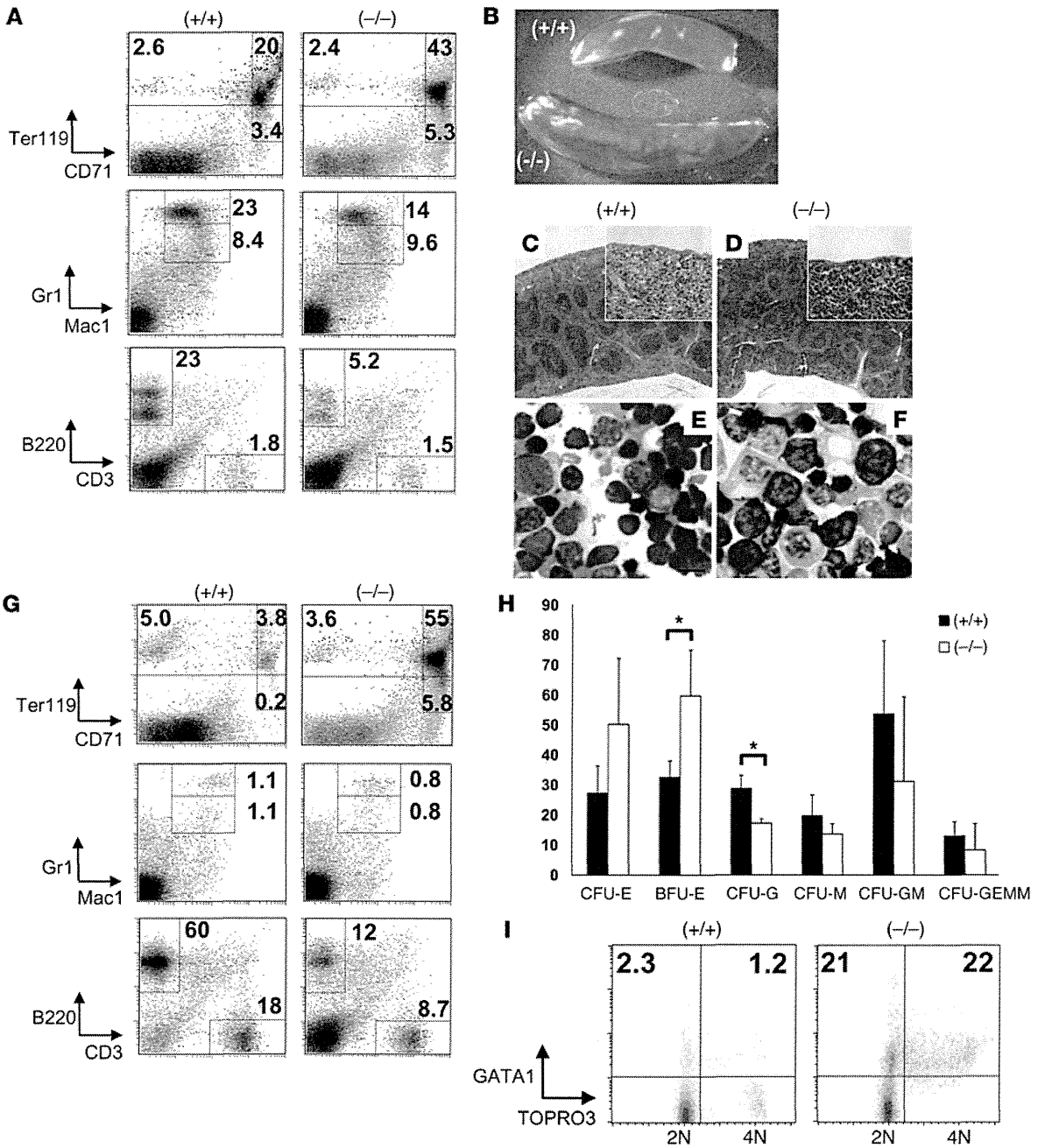


**Figure 7**

Enhanced erythropoiesis in *Smap1*$^{-/-}$ mice with MDS (ID no. 47; see Table 3). (**A** and **G**) Flow cytometry analyses of (**A**) bone marrow cells and (**G**) splenocytes prepared from *Smap1*$^{+/+}$ and *Smap1*$^{-/-}$ mice. The cells were stained for indicated hematopoietic lineage markers. The numbers represent percentages of cells in each gated box. Flow cytometry analyses were performed for all *Smap1*$^{-/-}$ mice with MDS, and reproducible results were obtained. (**B**–**F**) Histology of spleens. (**B**) A macroscopic view of the spleen. Note the enlargement of the *Smap1*$^{-/-}$ spleen. (**C** and **D**) Sections were stained by hematoxylin and eosin. Note the enrichment of cells with densely stained nuclei in the red pulp of the targeted spleen. (**E** and **F**) Smears of Giemsa-stained splenocytes. Erythroblasts with densely stained nuclei are evident in the *Smap1*$^{-/-}$ smear. Scale bar: 10 μm. (**H**) CFU-C assay of bone marrow cells. Cells from wild-type and *Smap1*$^{-/-}$ mice were assayed in vitro for their CFU-C activity. The numbers of colonies were counted under a microscope, and the morphology was classified as shown. Triplicate cultures were prepared from each mouse, and 4 independent pairs of older than 1 year *Smap1*$^{+/+}$ and *Smap1*$^{-/-}$ mice were used. The panel shows the averages ± SD of CFU-C values obtained from 12 cultures from each genotype. *$P < 0.0$. (**I**) Estimation of replicating cells in the spleen. Splenocytes from *Smap1*$^{+/+}$ and *Smap1*$^{-/-}$ mice were processed for flow cytometry analysis of GATA1 and TOPRO3. 2N and 4N represent the diploid and tetraploid status of chromatin.

**Table 3**
Details of the hematological disorder seen in *Smap1*$^{-/-}$ mice

| Mouse ID no. | Founder | Sex | Age (mo.) | Bone marrow | % Blasts in bone marrow cells[A] | Diagnosis |
|---|---|---|---|---|---|---|
| 47 | 44 | F | 13 | Erythroid hyperplasia | 2.9 | MDS |
| 675 | 64 | F | 15 | Erythroid hyperplasia | 0.4 | MDS |
| 834 | 44 | F | 18 | Erythroid hyperplasia | nd | MDS |
| 826 | 64 | F | 18 | Erythroid hyperplasia | 7.4 | MDS |
| 524 | 64 | F | 19 | Erythroid hyperplasia | 4.6 | MDS |
| 231 | 44 | F | 23 | Erythroid hyperplasia | 6.3 | MDS |
| 576 | 64 | M | 18 | Erythroid hyperplasia | 0.9 | MDS |
| 69 | 64 | F | 24 | ns | 3.4 | MDS |
| 72 | 64 | M | 17 | ns | nd | MDS |
| 199 | 64 | M | 25 | ns | 5.0 | MDS |
| 448 | 44 | F | 23 | Myeloproliferative | 12 | MPD/MDS |
| 518 | 64 | F | 17 | Myeloproliferative | 16 | MPD/MDS |
| 192 | 44 | F | 14 | Erythroleukemia | 28 | AML |
| 987 | 64 | F | 20 | Erythroleukemia | 24 | AML |
| 831 | 44 | F | 18 | Monocytic leukemia | 20 | AML |
| 34 | 64 | M | 19 | Monocytic leukemia | 22 | AML |
| 138 | 64 | F | 25 | Monocytic leukemia | 36 | AML |

[A]Note that the average percentage of blasts in bone marrow cells was 0.4 ± 0.3, as counted for 6 *Smap1*$^{+/+}$ mice. Numbers in the "Founder" column indicate the mouse line (the ES number). ns, no significant hyperplasia.

locytes was observed in the *Smap1*$^{-/-}$ anemic group, which likely reflects a mechanism of compensatory erythropoiesis. The number of platelets in *Smap1*$^{-/-}$ mice also decreased to 62%.

Peripheral blood and bone marrow cells from *Smap1*$^{-/-}$ mice were visualized using May-Grunwald-Giemsa staining (Figure 6), which revealed several morphological abnormalities mainly in erythrocytes and erythroid cells but also in megakaryocytic and even myeloid lineages (see the legend Figure 6 for details). Overall, the hematological disorders of *Smap1*$^{-/-}$ mice included (a) macrocytic and normochromatic anemia, (b) thrombocytopenia, and (c) abnormal cellular morphologies observed mainly in the erythroid lineage. These hematological disorders were detected exclusively in aged mice. Based on the Bethesda guidelines (20), these phenotypes correspond to features seen in MDS in aged humans.

Based on the above results, 10 *Smap1*$^{-/-}$ mice were diagnosed with MDS (Tables 1 and 2). Mice showing moribund conditions were sacrificed, and their bone marrow cells were examined by flow cytometry (Figure 7A). Erythroid lineage differentiation stages include Ter119$^{med}$CD71$^{hi}$ (proerythroblasts), Ter119$^{hi}$CD71$^{hi}$ (basophilic erythroblasts), and Ter119$^{hi}$CD71$^{med/lo}$ (late erythroblasts, including polychromatic/orthochromatic erythroblasts) (21). The Ter119$^{hi}$CD71$^{hi}$ fraction increased substantially in the *Smap1*$^{-/-}$ marrow as compared with the wild-type marrow (43% vs. 20%). In addition, the bone marrow from *Smap1*$^{-/-}$ mice showed hypercellularity, suggesting erythroid hyperplasia. Based on smears and flow cytometry, erythroid hyperplasia was detected in 7 out of 10 MDS-diagnosed *Smap1*$^{-/-}$ bone marrow samples (Table 3).

MDS-bearing *Smap1*$^{-/-}$ mice showed splenomegaly (Table 2), as illustrated in Figure 7B. Stained sections of *Smap1*$^{-/-}$ spleens (Figure 7D) revealed structurally intact red and white pulps, but the red pulp was extensively replaced by cells with densely stained nuclei. Splenocyte smears (Figure 7F) showed that the majority of these densely stained cells were erythroblasts. This was confirmed by flow cytometry analysis (Figure 7G), which revealed an increase in the Ter119$^{hi}$CD71$^{hi}$ fraction from 3.8% in the wild-type splenocytes to 55% in *Smap1*$^{-/-}$ splenocytes. Thus, splenomegaly seen in *Smap1*$^{-/-}$ mice most likely reflects erythroid hyperplasia.

To examine whether erythroid hyperplasia was accompanied by an enhancement of cell growth activity, a CFU assay was performed by in vitro culture of bone marrow cells (Figure 7H). BFU-E activity in *Smap1*$^{-/-}$ marrow cells was 1.8-fold higher than that in wild-type cells. This increase of BFU-E was apparent only in cells from mice older than 1 year of age, as shown in Figure 7H (data from younger mice is shown in Supplemental Figure 5). GATA1 is a master transcription factor of erythroid lineage. Splenocytes were stained for GATA1 expression and TOPRO3 (Figure 7I). In the *Smap1*-targeted spleen, the number of GATA1$^{+}$ cells with a DNA content above the diploid value (>2 N), which represent those undergoing DNA replication, increased to 22%.

The BFU-E assay and GATA1/TOPRO3 staining demonstrated that erythroid cell growth was enhanced in *Smap1*$^{-/-}$ mice. On the other hand, as shown above, these mice were characterized by anemia. Therefore, although erythropoiesis was enhanced in *Smap1*-targeted hematopoietic organs, the presence of morphological abnormalities in erythroid cells was indicative of a dysregulation of this process, resulting in overall ineffective erythropoiesis and eventually anemia.

*AML and myeloproliferative disease in Smap1-targeted mice.* In humans, patients with MDS often develop AML. We therefore assessed the incidence of AML in *Smap1*-targeted mice and found that 5 out of 33 *Smap1*$^{-/-}$ mice developed AML (Tables 1–3). Hematological subtypes of leukemia included erythroid (2 mice) and monocytic (3 mice). Figure 8A shows an example of a monocytic *Smap1*$^{-/-}$ AML mouse in which the liver and spleen were enlarged. Tissue sections of *Smap1*$^{-/-}$ mice (Figure 8B) show the infiltration of leukemic cells in the liver and spleen. Peripheral blood and bone marrow cell smears (Figure 8C) revealed the presence of immature monoblasts with large nuclei in the targeted mice. Figure 8, D and E, shows the results of flow cytometry analyses of bone marrow

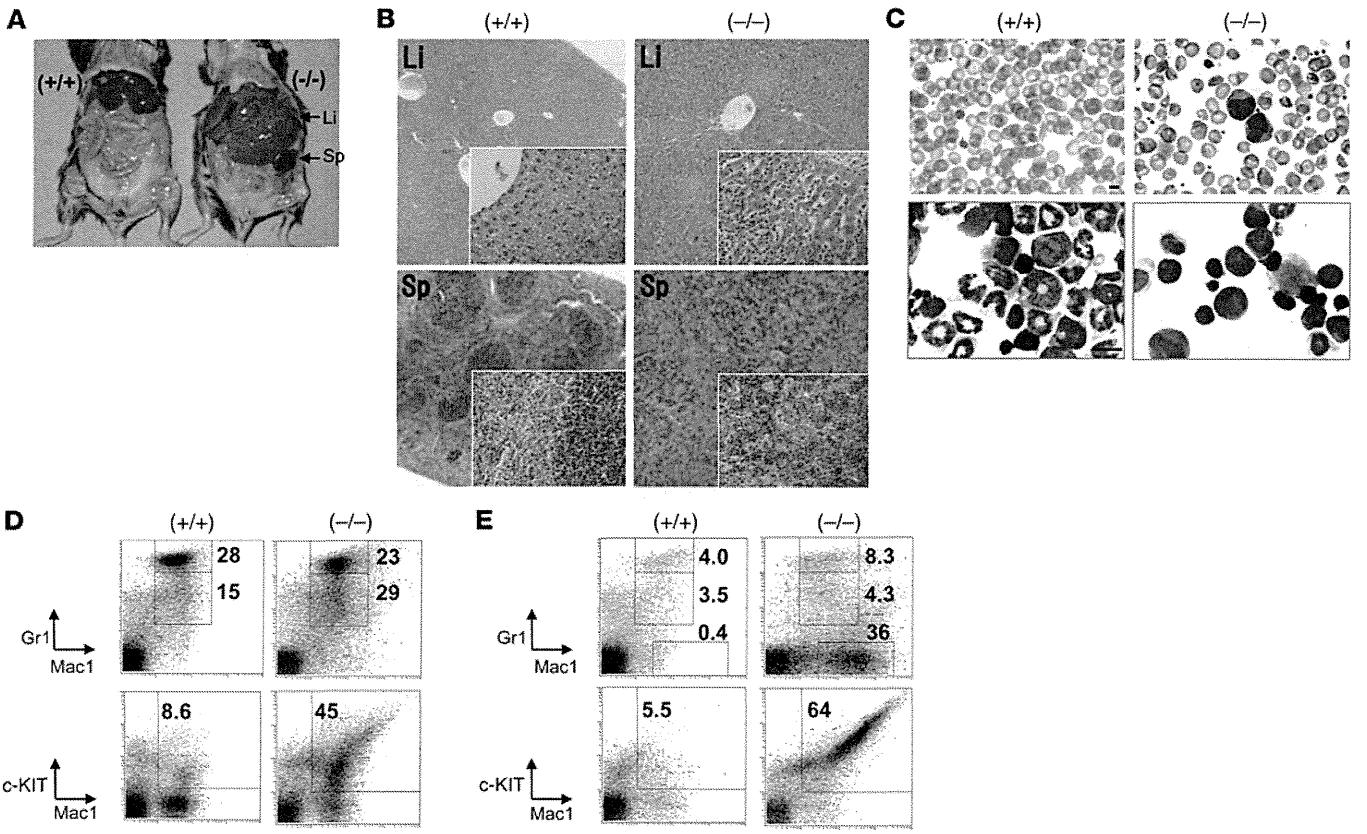


**Figure 8**

Monocytic AML in a *Smap1*$^{-/-}$ mouse (ID no. 831; see Table 3). (**A** and **B**) Macroscopic and histological views of the liver and spleen. Note the remarkably enlarged liver and spleen in the *Smap1*$^{-/-}$ mouse. Sections of liver and spleen were stained by hematoxylin and eosin. Note the massive infiltration of leukemic cells in the *Smap1*$^{-/-}$ tissues. The leukemic cells were relatively small in size and possessed densely stained nuclei. Original magnification, ×40; ×100 (insets). Li, liver; Sp, spleen. (**C**) Peripheral blood (top) and bone marrow cell (bottom) smears. Scale bar: 10 μm. (**D** and **E**) Flow cytometry analyses of (**D**) bone marrow cells and (**E**) splenocytes prepared from *Smap1*$^{+/+}$ and *Smap1*$^{-/-}$ mice. The cells were stained for hematopoietic lineage markers. The numbers indicate the percentage of cells in each gated box.

cells and splenocytes, respectively. *Smap1*$^{-/-}$ cells showed a substantial increase in the c-KIT$^{+}$Mac1$^{+}$ fraction (45% and 64% in bone marrow cells and splenocytes, respectively, as compared with 8.6% and 5.5% in the wild type) and an increase in the Gr1$^{-}$Mac1$^{+}$ fraction in the spleen (36% vs. 0.4% in the wild type). In addition, the immature Ter119$^{hi}$CD71$^{hi}$ fraction increased to 27% in the targeted spleen compared with 4.4% in the wild-type spleen (data not shown). This is likely the effect of enhanced compensatory erythropoiesis in response to anemia, suggesting the occurrence of AML in an MDS background. In fact, dysplastic cells were found in the peripheral blood smear (see *Smap1*$^{-/-}$ in Figure 8C).

In addition to monocytic leukemia, *Smap1*$^{-/-}$ mice developed erythroleukemia, as shown in Supplemental Figure 6. Overall, in the 5 mice that developed AML, the percentage of bone marrow blasts was above 20%, compared with 0.4%–7.4% in MDS-only mice (Table 2).

Two *Smap1*$^{-/-}$ mice developed a mixture of myeloproliferative disease (MPD) and MDS (Tables 1–3). The diagnosis of MPD was based on the increase in myeloid lineage cells, according to the Bethesda guidelines. Leukocytosis (as well as anemia and the presence of dysplastic cells) was observed in the peripheral blood, whereas in the bone marrow, the percentage of myeloid blasts was 12%–16%. Supplemental Figure 7 shows the results of analyses performed in a MPD *Smap1*$^{-/-}$ mouse.

*Role of c-KIT signaling in the growth of a Smap1*$^{-/-}$ *cell line derived from AML.* Finally, we examined the possible involvement of c-KIT signaling in the growth of *Smap1*$^{-/-}$ cells (Supplemental Figure 8). A cell line was established from the bone marrow of a monocytic AML mouse (ID no. 831; see Table 3). This cell line displayed a macrophage-like morphology and expressed c-KIT$^{lo}$, Mac1$^{hi}$, CD71$^{lo}$, and SCF (Supplemental Figure 8, A and B). Interestingly, as seen in Supplemental Figure 8C, which shows cells cultured without exogenous SCF, c-KIT was detected not only on the cell surface but also in the cytoplasm, and a fraction of cytoplasmic c-KIT colocalized with Hrs. In addition, treatment of cells with imatinib (a tyrosine kinase inhibitor) (22) or ISCK03 (a c-KIT inhibitor) (23) reduced cell viability and ERK1/2 phosphorylation (Supplemental Figure 8, D–F). Therefore, c-KIT signaling appears to have a positive effect on the viability of this AML-derived cell line.

## Discussion

We previously reported that the overexpression of SMAP1 impairs the endocytosis of the transferrin receptor in cultured cells (refs. 5, 6, and see Supplemental Figure 9 for the effects of SMAP1 overexpression on c-KIT transport). In this study, we confirmed that SMAP1 is similarly involved in transferrin endocytosis in mouse tissues but unexpectedly found that *SMAP1*-deleted cells (both erythroblasts

and MEFs) incorporated transferrin more efficiently than wild-type cells. One possible explanation for this enhanced endocytosis of transferrin is the upregulation of the active form of ARF6 in the absence of SMAP1. In this case, SMAP1 could be conferred a negative role in endocytosis. However, a number of reports provided evidence that, for vesicles to be formed properly, ARF has to exert its GTPase activity and itself be converted to an inactive form (24–28). Therefore, as an alternative mechanism of transferrin endocytosis, we hypothesized that endogenous SMAP1 could play a positive role by converting ARF6 to an inactive form and speculated the possible involvement of SMAP2, a SMAP1 homolog. Namely, SMAP1 deficiency might be accompanied by the mobilization of SMAP2 as a compensatory mechanism in *Smap1*-targeted cells (29). Since SMAP2 can function as an ARF1 GAP and exhibit higher GAP activity than SMAP1 (16), the recruitment of SMAP2 might lead to enhancement of transferrin endocytosis. This mechanism was supported by the abolishment of transferrin endocytosis in SMAP2-silenced *Smap1*$^{-/-}$ MEFs. A positive role for SMAP1 is compatible with the function of other ARF GAP proteins whose knockdown impairs vesicle transport. For example, targeting of ARF GAP1 impaired transferrin endocytosis (30), suggesting an ARF GAP1–dependent but SMAP1/2-independent route of transferrin endocytosis. In any case, our results support a positive role of SMAP1 in vesicle formation and may contribute to the discussion on the putative terminator versus effector functions of ARF GAPs (31, 32).

In this study, the loss of SMAP1 from BMMCs had no impact on SCF-mediated endocytosis of c-KIT, indicating that SMAP1 may play a role in the constitutive endocytosis of transferrin but not in the ligand-induced internalization of c-KIT. This result supports the idea that ARF GAP proteins function in distinctive, cargo-dependent pathways of endocytosis (33, 34). Alternatively, the present results may reveal a novel and possibly important function of SMAP1 in the degradation of c-KIT. SCF induces the phosphorylation and monoubiquitination of c-KIT, which are necessary for Grb2 association and endocytosis, respectively. The dephosphorylation of internalized RTKs occurs on early endosomes (7, 35). We observed that the kinetics of c-KIT in the early phases, such as phosphorylation, dephosphorylation, endocytosis from the cell surface, and exit from early endosomes, were not affected by the loss of SMAP1. In the late phases, c-KIT is transported to MVBs and eventually to the lysosomes (18, 19, 36). The endocytic adaptor protein Hrs functions at the MVB together with clathrin to sort ubiquitinated cargo to the lysosome (37, 38). Our results show that endogenous SMAP1 colocalized at least partially with clathrin and Hrs, suggesting that SMAP1 may be involved in the sorting of cargo at the MVB. In line with this notion, transport of c-KIT from MVB to lysosomes was substantially delayed in *Smap1*$^{-/-}$ cells. Importantly, undegraded c-KIT, ubiquitinated c-KIT, and Grb2-associated c-KIT increased in *Smap1*-targeted cells. Collectively, our results suggest that, in *Smap1*$^{-/-}$ cells, c-KIT persisting on MVBs may be in the dephosphorylated form but still capable of associating with Grb2, thereby leading to elevated levels of phosphorylated ERK. This suggests an enhancement of SCF-triggered c-KIT signaling, which was confirmed by an increase in thymidine uptake in these cells as an indicator of enhanced cell growth activity.

The alterations in transferrin and c-KIT transport discussed above were based on the analysis of *Smap1*-targeted cells derived from embryos and/or young mice exhibiting no obvious pathologies. We also revealed that aged *Smap1*-targeted mice exhibited phenotypes that resembled those of hematological disorders in human patients with MDS. Importantly, the signs of MDS were observed in 50% of *Smap1*-targeted aged mice, suggesting that disturbances in membrane transport may act as a predisposing but not a deterministic factor for MDS development. For MDS to occur, a genetic alteration may need to be present in addition to *SMAP1* deficiency.

The MDS-like features observed in *Smap1*-targeted mice included anemia, thrombocytopenia, and the presence of dysplastic blood cells. The majority of these mice showed signs of accelerated erythropoiesis in the bone marrow and spleen, which were confirmed by an increase in the number of erythroid-committed progenitors. Erythroid cells may have acquired enhanced proliferation abilities associated with differentiation defects, resulting in the occurrence of anemia. Enhanced expression of *SMAP1* in the MEPs and Ter119$^{+}$ fractions is in good accordance with the erythroid lineage-specific effects of *SMAP1* targeting.

Patients with MDS are particularly prone to developing AML. In this study, *Smap1*-targeted mice developed a range of AML subtypes, such as erythroid and monocytic leukemia. This is consistent with the notion that, in humans, MDS clones arise in CD34$^{+}$ progenitor cells (39) and with the observation that *SMAP1* expression is detected broadly in various hematopoietic lineages. Several *Smap1*$^{-/-}$ animals that had been diagnosed with MDS in our study subsequently developed AML (data not shown). It must be noted that not all MDS mice progressed to AML. This suggests that an additional genetic alteration(s) might be necessary to fully confer AML phenotypes. Screening and identification of such secondary mutations might provide information on the possible cooperation between *SMAP1* and another gene, resulting in the progression from MDS to AML.

Several clathrin-related molecules have been reported to enhance both transferrin endocytosis and cell growth. Huntingtin-interacting protein 1 (HIP1) is a clathrin-associated protein, and its expression level is frequently elevated in primary human cancers. In addition, overexpression of HIP1 alters the distribution patterns of clathrin and the AP-2 adaptor protein and promotes endocytosis of the transferrin receptor (40). Hrb/AGFG, another ARF GAP, functions positively in transferrin endocytosis in leukemic cells overexpressing Notch (41). It is widely accepted that transferrin receptor expression is increased in malignant tumors, including hematological malignancies, and may promote cell growth (42–44). Therefore, the enhanced endocytosis of transferrin in *Smap1*-targeted mice might facilitate the active iron metabolism and c-KIT–induced growth of erythroblasts and MDS/AML cells.

Furthermore, certain molecules that regulate the endocytosis of RTKs, such as c-KIT, have been associated with several human cancers. *c-Cbl* encodes an E3 ligase that ubiquitinates RTKs and is mutated in some cases of human MDS/AML (45, 46). Tsg101 is a component of the ESCRT-1 complex that functions downstream of Hrs to sort ubiquitinated RTKs to MVBs. In various human cancers, *TSG101* is deleted or its splicing pattern is altered (7, 47). Furthermore, *SMAP1* is frequently mutated in human colon cancer associated with microsatellite instability (15). Base deletion or addition in the $(A)_{10}$ repeat causes a frameshift in the ARF GAP domain, resulting in a loss-of-function type mutation of *SMAP1* (occurrence of colon cancer has not been observed in our *Smap1*$^{-/-}$ mice so far). Although one case of a *MLL-SMAP1* chimeric gene was reported in monocytic AML, the possible involvement of *SMAP1* in human MDS/AML remains to be investigated in the future.

In summary, this study revealed the predisposing role of alterations in clathrin-dependent protein trafficking in the development of MDS (and subsequent AML). To the best of our knowledge, this is the first report describing this mechanism using a mouse model.

## Methods

*Mice.* The 4.8-kb and 6.4-kb fragments corresponding to the 5′ and 3′ sequences, respectively, of exon 1 of *SMAP1* were obtained from the corresponding BAC clones by appropriate restriction digestion. The genomic fragments, as well as the loxP site, *frt*-flanked neomycin resistance cassette, and diphtheria toxin subunit A gene, were inserted into the targeting vector. The resulting plasmid DNA was linearized and electrophoretically transfected into TT2 ES cells, which were derived from an F1 mouse of a C57BL/6 and CBA mating (48). Positive and negative selection and PCR genotyping yielded 14 colonies. Recombinant alleles were detected by PCR screening using the forward and reverse primers (5′-CTGACCGCTTCCTCGTGCTTTACG-3′ and 5′-AATACACATGGCCTAGATATTAACCTATAG-3′) derived from the neomycin resistance cassette and 3′ external region. Recombination was verified by Southern blot analysis, and 3 independent clones were each injected into 8-cell stage embryos of CD-1 mice. Two clones, 44 and 64, were successfully transmitted through the germ line, and *Smap1*-heterozygous mice were backcrossed to C57BL/6 mice for more than 10 generations. During these matings, heterozygous mice were crossed with *E2A-Cre* transgenic mice, causing the deletion of exon 1 in all tissues. *Smap1*$^{+/-}$ mice (acc. no. CDB0427K; http://www.cdb.riken.jp/arg/mutant%20mice%20list.html) were thus established. For PCR genotyping of mice, the reverse primer was 5′-CCTCTGCTAACTCTACTCAG-3′, and the forward primers were 5′-GTCATCCTGGTTAGCCTCAGTCTTG-3′ for the wild-type alleles, 5′-CGCCTTCTATCGCCTTCTTGACG-3′ for the floxed alleles, and 5′-CCTGCCCTTACCCAGACTGTCTTAG-3′ for the targeted alleles. The expected sizes of PCR products were 330 bp, 554 bp, and 480 bp for wild-type, floxed, and targeted alleles, respectively. Mice were maintained in the Animal Facility of the Institute of Development, Aging, and Cancer, Tohoku University, an environmentally controlled and specific pathogen–free facility.

*Cultures of BMMCs, the AML-derived cell line, and MEFs.* Femoral bone marrow cells were cultured in RPMI1640 supplemented with 10% (v/v) FBS, 2 mM L-glutamine, 50 μM 2-mercaptoethanol, 10 mM HEPES (pH 7.4), 0.2 mM nonessential amino acids, 1 mM sodium pyruvate, 100 U/ml penicillin, and 100 μg/ml streptomycin. The cytokines added were 20 ng/ml IL-3 and 10 ng/ml SCF for the first 2 weeks followed by 20 ng/ml IL-3 only for the next 2 weeks. The purity of BMMCs reached over 95%, as assessed by the surface expression of c-KIT and FcεRI. A cell line was established from the bone marrow of an AML-bearing *Smap1*$^{-/-}$ mouse (ID no. 831; see Table 3) and cultured (deposited as TKG 0661 at the Cell Resource Center for Biomedical Research, Institute of Development, Aging and Cancer, Tohoku University). The medium was the same as that used for BMMCs but without any cytokine added. Cell viability was measured using a Cell Counting Kit-8 (Dojindo). In the indicated cases, cells were treated with imatinib mesylate (Santa Cruz Biotechnology Inc.) or ISCK03 (Sigma-Aldrich).

MEFs were prepared according to a previously published procedure (49). Briefly, 16.5-day-old embryos were isolated from pregnant mice. Single cell suspensions were prepared by trypsin digestion, cultured in a monolayer in DMEM supplemented with serum for 3 days, and immortalized by transfection with the SV40 large T antigen. Two independent cultures were established for wild-type and *Smap1*$^{-/-}$ genotypes. Two siRNAs against SMAP2, siRNA1 and siRNA2, were purchased from Japan BioServices Co. LTD. and used to downregulate endogenous SMAP2 in MEFs. The sequences of siRNA1 and siRNA2 were 5′-GGAUUUAUUCGAGAUAAAUTT-3′ and 5′-CCUGUUGUUUUUGAGAAAGTT-3′, respectively.

*Hematological and histological examination.* Peripheral blood was collected from the tail veins of mice, and hematological parameters were measured using an automated cell counter (XT-4000i, Sysmex). Smears of peripheral blood and bone marrow cells were prepared on glass slides and stained with May-Grunwald-Giemsa (Wako). Tissues, such as those derived from the spleen and liver, were fixed in 4% (w/v) paraformaldehyde in PBS for 18 hours at 4°C and embedded in paraffin. Microsections of each specimen were fixed on glass slides, deparaffinized, and stained with hematoxylin and eosin (Wako).

*Flow cytometry analyses.* Single cell suspensions were prepared from the bone marrow and spleen and incubated with CD16/32 mAb (BD Pharmingen) for 15 minutes, followed by incubation with an appropriately diluted, fluorescein-conjugated mAb on ice for 30 minutes. The mAbs used were PE-anti-CD71 (eBioscience) and FITC-anti-c-KIT, PE-anti-Ter119, APC-anti-Ter119, PE-anti-Gr1, APC-anti-Mac1, APC-anti-B220, and FITC-anti-CD3ε (all from BD Pharmingen). For DNA labeling, cells were fixed in 4% (w/v) paraformaldehyde in PBS and permeabilized with 0.25% (v/v) Triton X-100 and 5% (w/v) DMSO in PBS. After blocking with 10% (v/v) goat serum in PBS, cells were incubated with anti-GATA1 mAb (Santa Cruz Biotechnology Inc.) and TOPRO3 (Molecular Probes), followed by a secondary antibody reaction. The labeled cells were separated with an analytical flow cytometer (Beckman Coulter), and the data were analyzed with EXPO32 software. Various hematopoietic progenitor fractions were identified using the appropriate antibodies. These populations were designated as follows: KSL, Lin$^-$/c-KIT$^+$/Sca-1$^+$; FLT3-KSL, Lin$^-$/c-KIT$^+$/Sca-1$^+$/FLT3$^-$; FLT3+KSL, Lin$^-$/c-KIT$^+$/Sca-1$^+$/FLT3$^+$; CMP, Lin$^-$/c-KIT$^+$/Sca-1$^-$/CD34$^{hi}$/FcγR$^{lo}$; GMP, Lin$^-$/c-KIT$^+$/Sca-1$^-$/CD34$^{hi}$/FcγR$^{hi}$; and MEP, Lin$^-$/c-KIT$^+$/Sca-1$^-$/CD34$^{lo}$/FcγR$^{lo}$. Cell sorting was performed using a FACSAria (Becton Dickinson).

*Colony formation and $^3$H-thymidine incorporation assays.* Single cell suspensions were prepared from the femoral bone marrow and plated at a density of 1 × 10$^4$ cells per ml of methylcellulose (M3234, Stem Cell Technologies) in a 3.5-cm–diameter dish. The culture medium contained 50 ng/ml rmSCF (Kirin Brewery Company Ltd.), 10 ng/ml rmIL-3 (Wako), 10 ng/ml rmIL-6 (Wako), and 3 U/ml rhEPO (Peprotech). Colonies formed were observed through a phase-contrast microscope, and their numbers were counted on the third day for CFU-E and twelfth day for BFU-E, CFU-G, CFU-M, CFU-GM, and CFU-GEMM. In certain cases, 20 kBq of $^3$H-thymidine (GE Healthcare) was added to the culture of 2 × 10$^4$ BMMCs for 8 hours, and its incorporation into an acid-insoluble fraction was measured by a beta-counter, Matrix 9600 (Packard), according to the described method (50).

*Transport assay.* Intracellular uptake and recycling of transferrin were evaluated using an erythroblast-containing fraction. Bone marrow cells were incubated in serum-free medium at 37°C for 2 hours and then in RPMI1640 containing 50 μg/ml Alexa Fluor 488–conjugated transferrin (Molecular Probes), 20 mM HEPES (pH 7.4), and 1% (w/v) BSA on ice for 30 minutes. After washing 3 times, transferrin internalization was induced by incubating cells in RPMI1640 containing 10% (v/v) FBS at 37°C for the indicated times. Transferrin remaining on the plasma membrane was removed by incubating cells in a prechilled buffer consisting of 20 mM MES (pH 5), 130 mM NaCl, 50 μM deferoxamine, 2 mM $CaCl_2$, and 0.1% (w/v) BSA on ice for 20 minutes. After washing 3 times, cells were labeled with PE-anti-Ter119, and the fluorescence intensity of internalized transferrin in each Ter119$^+$ fraction was quantified by flow cytometry.

For measuring the internalization of c-KIT, BMMCs were first serum starved and then incubated in RPMI1640 supplemented with 100 ng/ml SCF and 0.1% (w/v) BSA in the presence of 100 μg/ml cycloheximide (Sigma-Aldrich), and the c-KIT remaining on the cell surface was measured by flow cytometry.