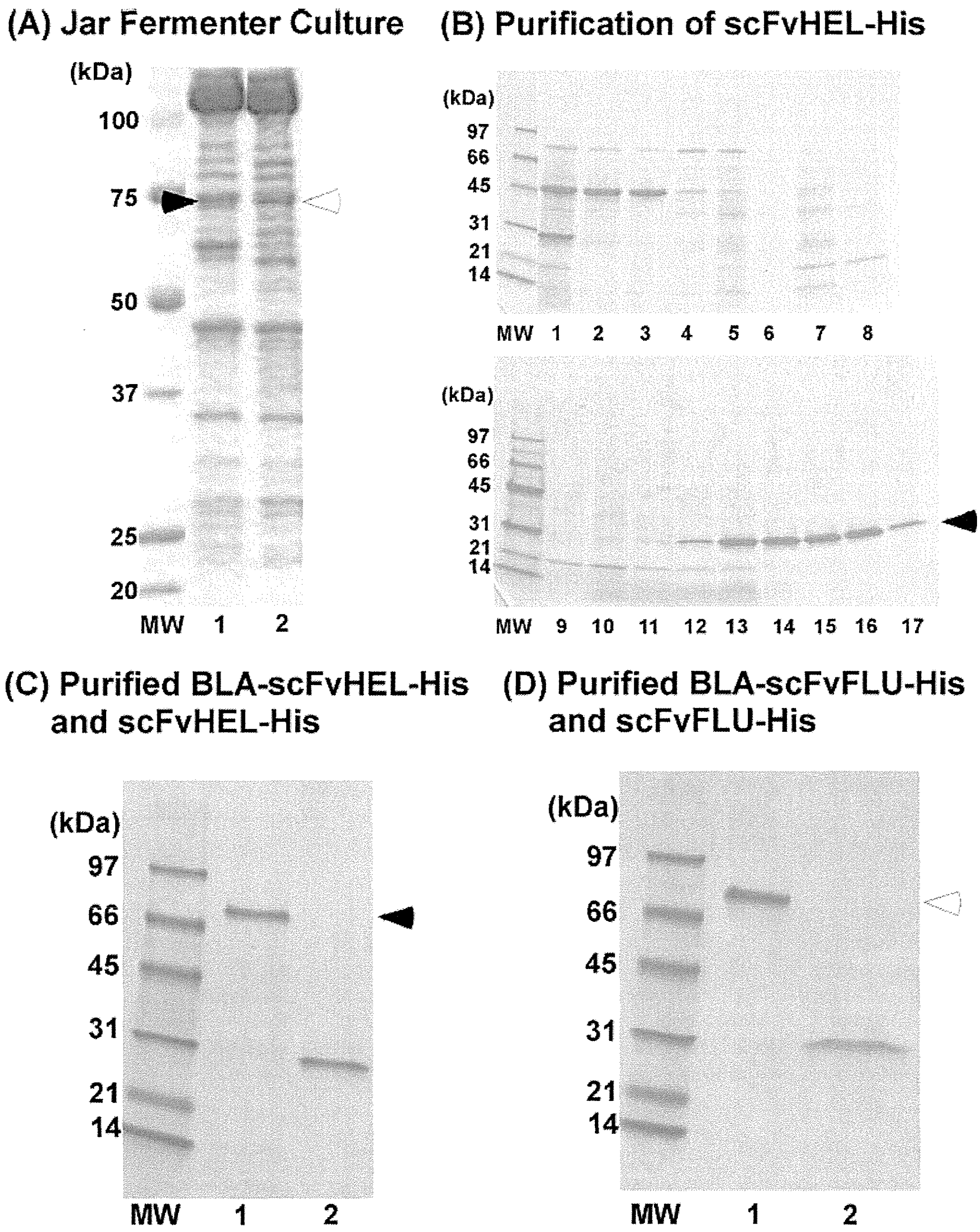
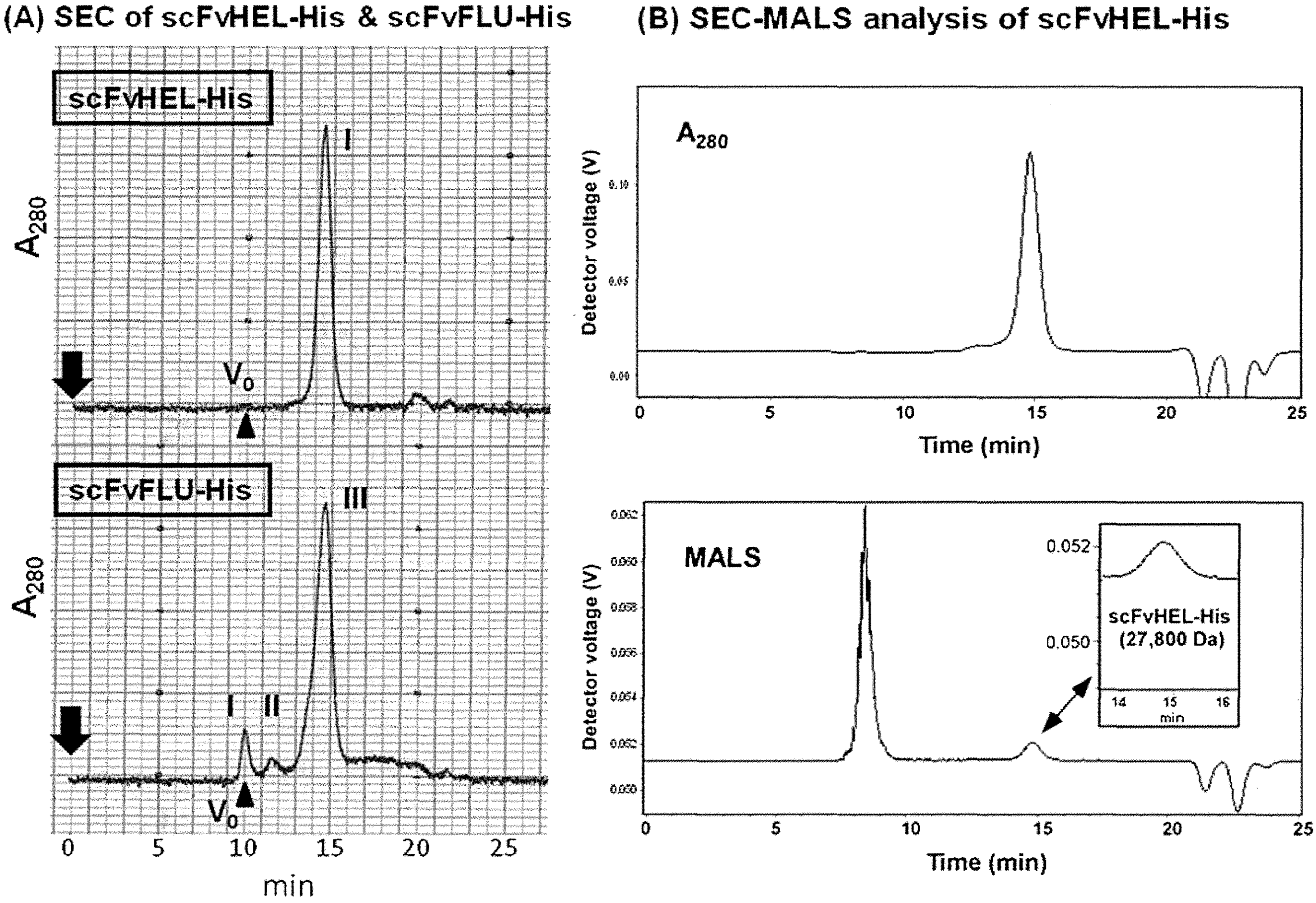

(A) Jar Fermenter Culture
(kDa)
100
75
50
37
25
20
MW 1 2
(B) Purification of scFvHEL-His
(kDa)
97
66
45
31
21
14
MW 1 2 3 4 5 6 7 8
(kDa)
97
66
45
31
21
14
MW 9 10 11 12 13 14 15 16 17
(C) Purified BLA-scFvHEL-His and scFvHEL-His
(kDa)
97
66
45
31
21
14
MW 1 2
(D) Purified BLA-scFvFLU-His and scFvFLU-His
(kDa)
97
66
45
31
21
14
MW 1 2

Fig. 3 Expression and purification of scFv proteins from jar fermenter culture. **a** Expression of BLA-scFvHEL-His (*white arrowhead*, *lane 2*) and BLA-scFvFLU-His (*black arrowhead*, *lane 1*) in culture supernatant of *Brevibacillus* jar fermenter culture. Each 1 μl was applied to SDS–PAGE. **b** Purification of scFvHEL-His protein with Ni-NTA column after digestion of fusion protein. *Black arrowhead*—scFvHEL-His. *Lane 1* thrombin-digested fusion protein before column; *lanes 2–17* represent fractions of Ni-NTA column, flow through (*2*), 20 mM (*3–5*), 50 mM (*6–8*), 100 mM (*9–11*), 200 mM (*12–14*), and 300 mM (*15–17*) imidazole-eluted fractions. **c** Purified BLA-scFvHEL-His (*lane 1*, *black arrowhead*) and scFvHEL-His (*lane 2*) proteins. **d** Purified BLA-scFvFLU-His (*lane 1*, *white arrowhead*) and scFvFLU-His (*lane 2*) proteins

SEC and SEC–MALS analysis

These scFv-His proteins purified to homogeneity (Fig. 3b, lanes 15–16; Fig. 3c and d, lane 2) from the BLA-fusion construct were analyzed by SEC on Superdex 75 using 0.1 M Na–phosphate buffer (pH 6.8)/0.2 M arginine mobile phase. As shown in Fig. 4a, a single peak of scFvHEL-His was observed (upper panel), with no apparent peaks corresponding to aggregated species and consistent with the purity observed in SDS–PAGE (Fig. 3c). Figure 4a (lower panel) shows SEC analysis of scFvFLU-His with a main peak (III), corresponding to scFvFLU-His. A few minor peaks (I and II) were seen in this sample, perhaps corresponding to oligomers. The observed oligomers are consistent with the aggregation tendency of the scFvFLU protein and consequent batch-dependent product quality variation as described above. The retention times for two scFv proteins (peak I in upper panel and peak III in lower panel) were very similar (14.4 min and 14.6 min), consistent with the similar molecular weight.

The monomeric state of the scFvHEL-His in aqueous solution was confirmed by SEC–MALS technique. As shown in Fig. 4b, the elution profile (upper panel, A_{280}) is similar to the result shown in Fig. 4a, showing one major peak with a shoulder before the peak. The lower panel shows the light scattering intensity of the eluted peaks. The molecular mass of the main absorbance peak was determined to be 27,800 Da, consistent with the theoretical molecular mass (26,495 Da), indicating a monomeric structure of scFvHEL-His. A large light scattering peak was also


(A) SEC of scFvHEL-His & scFvFLU-His

(B) SEC-MALS analysis of scFvHEL-His

Fig. 4 Size-exclusion chromatography and light scattering analysis of scFv-His proteins. **a** Superdex 75 SEC analysis of scFvHEL-His (*upper panel*) and scFvFLU-His (*lower panel*). *Arrow* shows start point of chromatography. V_0 shows void volume. Peak I (retention time, 14.4 min) in *upper panel* shows monomer peak of scFvHEL-His. Peak III (retention time, 14.6 min) in *lower panel* shows monomer peak of scFvFLU-His. Peaks I and II show a small amount of aggregations of scFvFLU-His. **b** SEC–MALS analysis of scFvHEL-His protein. *Upper panel* shows A_{280} signal and *lower panel* shows light scattering signal. In the *lower panel*, the peak region corresponding to the scFv monomer was expanded in inserted figure

observed at the void volume (lower panel), where a small drift of A_{280} base line was observed (upper panel). This is simply due to the fact that light scattering intensity is proportional to the molecular weight and hence even small amounts of large aggregates can generate strong light scattering signals. Thus, it can be concluded that some aggregates are present in this scFvHEL-His preparation, but in negligible amounts (as seen in A_{280} profile). A similar retention time of peak III in Fig. 4a (lower panel) to the retention time for scFvHEL-His suggests that a majority of scFvFLU-His preparation is also monomeric in solution.

Antibody activity assay of scFvHEL-His and scFvFLU-His proteins

The effects of scFvHEL-His on lysozyme activity were examined based on inhibition of the enzyme activity and formation of their complex. Figure 5a shows titration curve of lysozyme activity as a function of scFvHEL-His concentration. As the scFvHEL-His concentration was increased at the fixed concentration of lysozyme, the enzyme activity was gradually reduced (from right to left). When the ratio of scFvHEL-His to lysozyme was 0.5, the lysozyme activity was reduced to half and the activity was negligible at the ratio of 1. Thus, binding of scFvHEL-His to lysozyme is nearly stoichiometric and of high affinity. It is evident that scFvHEL-His binds to lysozyme in a manner that causes blocking of enzyme active site. Binding of scFvHEL-His to lysozyme was confirmed by SEC analysis. Figure 5b shows SEC analysis of 1:1 molar ratio mixture of scFvHEL-His and lysozyme. Figure 5b [bottom panel (C)] shows the elution of lysozyme (black arrow), while the middle panel (B) shows the elution position of scFvHEL-His (white arrow). When they were mixed [upper panel (A)], the lysozyme peak (black arrow) disappeared, suggesting that it was incorporated into a complex with scFvHEL-His. The peak position of scFvHEL-His (peak II) was also moved to smaller elution position, indicating binding of lysozyme. Based on fairly small shift of elution position, the

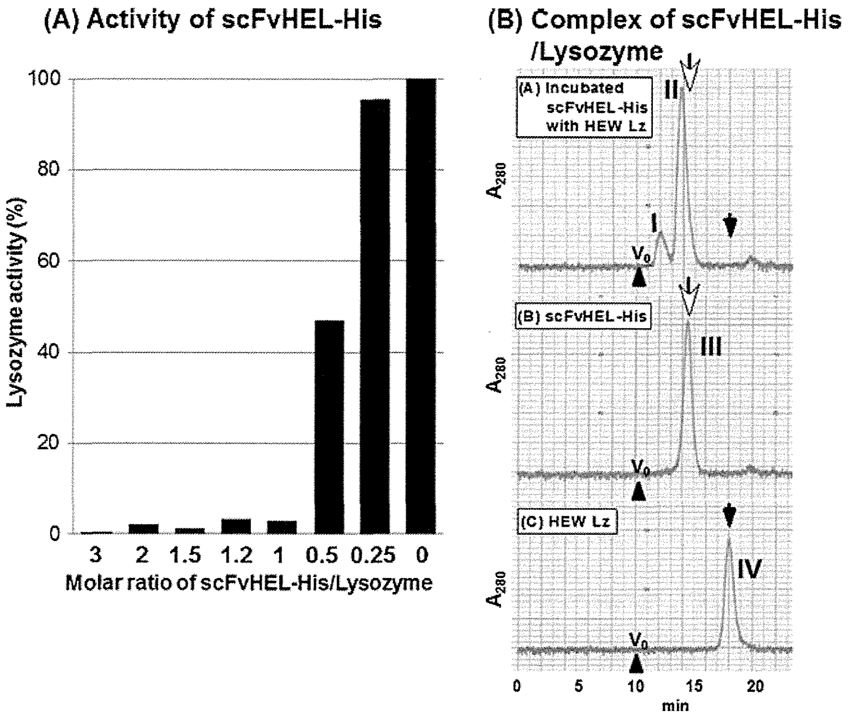


Fig. 5 Antibody activity of scFvHEL-His and detection of complex of scFvHEL-His/hen egg white lysozyme with Superdex 75 SEC analysis. **a** Antibody activity of scFvHEL-His. Hen egg white lysozyme (0.043 nmol) was incubated with various molar ratios (3–0) of scFvHEL-His for 1 h at 25 °C, and then lysozyme activity was measured. Results shown here are average data of three experiments and deviation is less than 5 %. **b** Hen egg white lysozyme (HEW Lz) and scFvHEL-His (both 0.344 nmol) was incubated for 1 h at 25 °C, and subjected to Superdex 75 SEC analysis. *Upper panel* (*A*) shows SEC of complex. Peaks I and II represent dimer and monomer of complex. *White arrow* shows position of scFvHEL-His monomer and *black arrow* shows position of lysozyme. *Middle panel* (*B*) and *lower panel* (*C*) show control experiments: peak III in (*B*) represents monomer position of scFvHEL-His and peak IV in (*C*) represents position of lysozyme

complex formed may be fairly compact and hence have only a slightly increased hydrodynamic size relative to the size of scFvHEL-His (white arrow). A small peak was observed at earlier elution position (peak I), which appeared to be dimers of scFvHEL-His/lysozyme complex.

The scFvFUL-His protein was also similarly purified from the BLA fusion (Fig. 3d, lane 2 and Fig.4a, lower panel). The monoclonal antibody, from which scFvFUL-His was derived, has been shown to stoichiometrically inhibit the fluorescence of traditional fluorophore, fluorescein (Kudou et al. 2011). Fluorescence of fluorescein was titrated with the purified scFvFUL-His. Figure 6 plots the fluorescence intensity of 0.6 μM fluorescein as a function of scFvFUL-His concentration. At the ratio of 0.5, the fluorescence intensity was about half, again suggesting nearly stoichiometric binding of scFvFUL-His to fluorescein. However, when the ratio was increased to 1, there were small but significant amounts of fluorescence. The ratio above 1.5 appeared to be required for greater fluorescence suppression, indicating that the binding affinity of scFvFUL-His is not as strong as for scFvHEL-His binding to lysozyme. A similar titration experiment was done with bovine serum albumin, showing no inhibition of fluorescein. Thus, it may be concluded that the observed suppression of fluorescein fluorescence by scFvFUL-His is due to its specific binding.

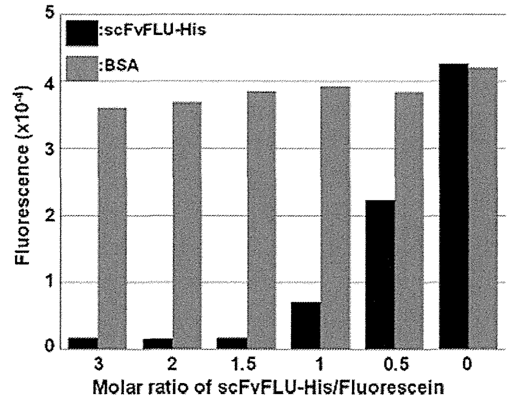


Fig. 6 Antibody activity of scFvFLU-His protein. Fluorescein (0.6 μM) was incubated with various molar ratios of scFvFLU-His for 1 h at 25 °C, and then the fluorescence intensity (*black bar*) was measured at excitation 480 nm and emission 515 nm. *Gray bar* shows control experiment using the same amount of bovine serum albumin (BSA) instead of scFvFLU-His. Results shown here are average data of two experiments and deviation is less than 10 %

Expression of scFv fusion protein with *Brevibacillus* secretory protein P45 as a partner protein

Expression of these scFv with another fusion partner, i.e., P45, in *Brevibacillus* was attempted in a similar manner to the BLA fusions. This P45 is an intrinsic 45-kDa (417 amino acid residues) abundant secretory protein of *Brevibacillus* bacterium, as supposed to a foreign protein of BLA (halophilic origin) and may enhance soluble expression of the fusion protein. Both P45-scFvHEL-His (Fig. 1, construct 4) and P45-scFvFLU-His (construct 5) were clearly expressed in culture supernatant, as indicated by the arrowhead (Fig. 7a, lane 2 and lane 3). Both proteins were purified by His-Trap columns. Figure 7b shows expression of P45-scFvHEL-His in culture supernatant of 200 ml flask culture (lane 1, dot) and purified fraction from the His-Trap column (lane 2). Highly homogeneous preparation of P45-scFvHEL-His was obtained, indicating both soluble expression and stability of this construct. Figure 7c shows purification of P45-scFvFLU-His construct. For this protein as well, highly homogeneous preparation (lane 2) was obtained from crude supernatant (lane 1). However, thrombin cleavage was rather difficult and inconsistent, in particular for scFvFLU fusion construct, most likely due to weak ability of P45 to enhance the solubility of the scFv proteins. Greater aggregation tendency of scFvFLU described above made it more aggregating with P45 fusion. Thus, the analysis of scFv proteins from the P45 constructs was not possible and was done as a fusion protein.

Antibody activity of whole fusion proteins

Expression of both BLA fusion and P45 fusion has resulted in soluble scFv proteins. If these scFv proteins were functional in the fusion form, they could be a useful reagent, in particular when cleavage of the fusion partner is difficult (for example, the P45 fusion). The effects of BLA- and P45-fusion constructs on lysozyme activity and fluorescein fluorescence were thus examined without cleaving the fusion partners. Figure 8a shows titration of lysozyme activity at fixed lysozyme concentration as a function of BLA-scFvHEL-His concentration. Similar to scFvHEL-His, the lysozyme activity was nearly stoichiometrically inhibited by this fusion protein, indicating that BLA does not interfere with binding of scFvHEL to the enzyme. BLA-scFvFLU-His did not inhibit lysozyme activity, indicating that the BLA portion of BLA-scFvHEL-His fusion protein has no influence on lysozyme activity (data not shown). Lysozyme activity was similarly titrated with P45-scFvHEL-His, showing gradual reduction of enzyme activity by this fusion construct (Fig. 8b). However, it appears that the degree of inhibition was slightly less for P45 construct than for the BLA construct and hence that P45 may sterically interfere with binding of scFvHEL to the enzyme, leading to slightly reduced binding strength. Alternatively, the P45-scFvHEL-His preparation may be heterogeneous due to aggregation tendency of P45, resulting in lower concentration of functional monomeric form. Nevertheless, it is evident that P45-scFvHEL-His does bind to lysozyme. As with BLA-scFvFLU-His protein, P45-scFvFLU-His

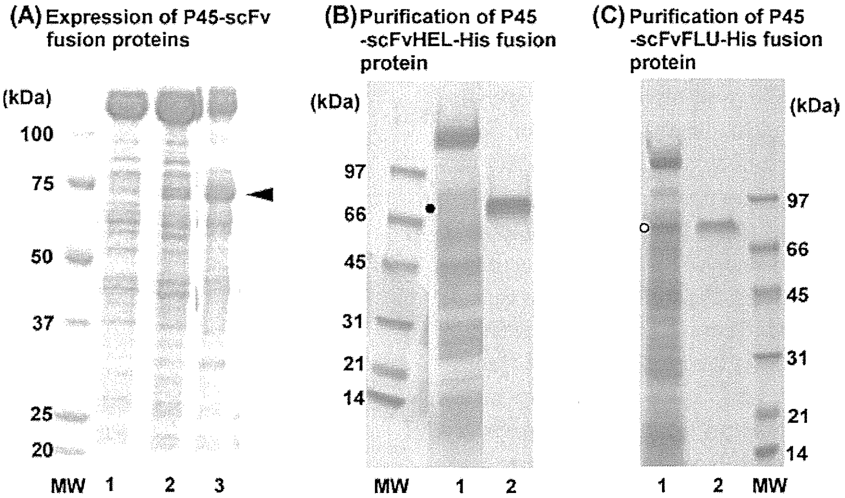

(A) Expression of P45-scFv fusion proteins
(kDa)
100
75
50
37
25
20
MW 1 2 3
(B) Purification of P45 -scFvHEL-His fusion protein
(kDa)
97
66
45
31
21
14
MW 1 2
(C) Purification of P45 -scFvFLU-His fusion protein
(kDa)
97
66
45
31
21
14
1 2 MW

Fig. 7 Expression and purification of P45-scFv-His fusion proteins. **a** Expression of P45-scFvHEL-His (*lane 2*) and P45-scFvFLU-His (*lane 3*) in test tube culture (shown by *arrowhead*). *Lane 1* vector control without fusion gene. Each 4 μl of culture supernatant was applied. **b** Purification of P45-scFvHEL-His fusion protein from 200 ml batch culture. *Lane 1* culture supernatant. *Dot* shows P45-scFvHEL-His fusion protein; *lane 2* P45-scFvHEL-His fusion protein purified with His-Trap column. **c** Purification of P45-scFvFLU-His fusion protein from 200 ml batch culture. *Lane 1* culture supernatant. *White dot* shows P45-scFvFLU-His fusion protein; *lane 2* P45-scFvFLU-His fusion protein purified with His-Trap column

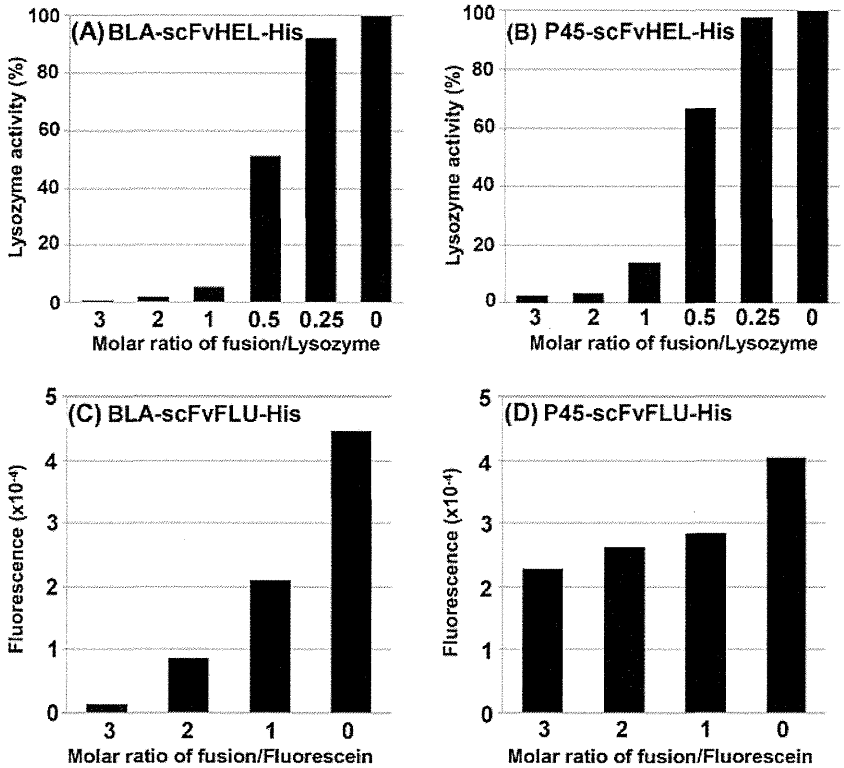


Fig. 8 Antibody activity of BLA- and P45-scFv fusion proteins. Whole BLA-scFv and P45-scFv fusion proteins (without protease digestion) were assayed using their antibody activities. Lysozyme (0.043 nmol) was incubated with various molar ratios of BLA-scFvHEL-His (**a**) or P45-scFvHEL-His (**b**) fusion protein for 1 h at 25 °C, and then lysozyme activity was measured. Fluorescein (0.6 μM) was incubated with various molar ratios of BLA-scFvFLU-His (**c**) or P45-scFvFLU-His (**d**) proteins for 1 h at 25 °C, and then fluorescence intensity was measured

protein had no effects on lysozyme activity, indicating that P45 portion as well did not affect lysozyme activity (not shown).

Titration of fluorescein was also done for BLA- and P45-fusion constructs. When fluorescein at a fixed concentration was titrated, the fluorescence gradually decreased with increasing BLA-scFvFLU-His concentration (Fig.8c, from right to left). However, there was significant fluorescence (about 50 %) at the molar ratio of 1. Nearly complete suppression of fluorescence required a molar ratio of 3, indicating that for this antigen, the BLA moiety does interfere with binding of scFvFLU to fluorescein. Alternatively, the BLA-scFvFLU-His preparation is heterogeneous due to aggregation tendency of scFvFLU, which caused reduced activity of the preparation used. As shown in Fig. 8d, only small reduction of fluorescence was observed by the addition of P45-scFvFLU-His. Even at the molar ratio of 3, the fluorescence intensity was still above 50 % of the original fluorescence (no P45-scFvFLU-His). This can be explained by reduced concentration of functional P45-scFvFLU-His due to aggregation-prone properties of this fusion protein, although possible steric hindrance of P45 portion on scFvFUL binding cannot be excluded.

Discussion

Antibody fragments, in particular scFv, are intensely investigated primarily as an anti-cancer drug (Beck et al. 2010; Demarest and Glaser 2008; Kontermann 2010). Strong interest is based on its small size, which allows more efficient penetration into solid tumors expressing antigen markers, to which the fragments bind. Cytotoxic agents are conjugated to the fragments that deliver the anti-cancer agents to the specific site (Beckman et al. 2007; Ottiger et al. 2009). However, production of antibody fragments, more specifically scFv, is not often straightforward, requiring laborious developmental work both in expression and purification. Refolding may be required when the fragments were expressed insoluble (Fujii et al. 2007; Fursova et al. 2009; Kurucz et al. 1995; Tsumoto et al. 1998). Here we were able to successfully express both scFv molecules, i.e., scFvHEL and scFvFLU, as a fusion protein in the culture media of *B. choshinensis*. Proteolytic cleavage of the fusions resulted in functional scFv proteins.

One of the fusion proteins used is halophilic BLA derived from *Chromohalobacter* sp. 560. We have shown before that BLA is extremely soluble even at high temperatures, at which it is fully unfolded (Arakawa et al. 2010; Tokunaga et al. 2004, 2006a). Such high solubility of heat-denatured structure is the key to its ability to support folding of the target protein, to which the BLA is fused (Tokunaga et al. 2010b). The scFv is an artificial molecule, in which two variable domains of antibody heavy and light chains are connected by an artificial linker, and hence may suffer both folding and aggregation problems during folding process, in particular when the folding is slow. The intermediate structures may aggregate during folding process in cellular environments. High solubility of halophilic BLA may afford the solubility of the fusion protein, allowing a sufficient time for the fused scFv to fold into the native structure. P45 derived from *Brevibacillus* bacterium appears to function in a similar manner by maintaining the solubility of the fusion protein. Although the solubility and folding properties of P45 have not been extensively investigated, its solubilizing effects on scFv, in particular scFvFLU, appeared to be weaker than BLA.

Proteolytic cleavage of BLA fusion partner from two scFv was efficient, indicating no apparent aggregation. Repeated experiments showed more reproducible cleavages of BLA-scFvHEL fusion than BLA-scFvFLU fusion. When protein aggregates, proteolytic cleavage of peptide bonds will be normally compromised, thus suggesting greater aggregation tendency for BLA-scFvFLU fusion as described in the "Results" section. Both purified scFvHEL and scFvFLU stoichiometrically bound to their antigens and blocked the activity of the antigens. Due to more ideal solution property of scFvHEL, binding could be directly demonstrated by the SEC analysis of the complex formation between scFvHEL and lysozyme. The additional work on binding equilibrium and kinetics of the scFv proteins to the antigens will be necessary to be compared with each other and to commence further application of these preparations.

ScFv has been seen to suffer proteolytic cleavages during expression, although the exact sites and cleavage mechanism have not been determined. This often resulted in copurification of cleaved scFv fragments. This problem does not appear to occur in the present fusion expression system, as no such fragments were observed in the final product, perhaps due to stabilization of scFv against proteolytic cleavage by the fusion partners.

Both scFv molecules were functional even before proteolytic cleavage, meaning that they were already folded. Thus, these scFv may have potential application as a diagnostic agent. Antigen binding was slightly reduced for BLA-scFvHEL-His and P45-scFvHEL-His, more so for the latter, perhaps due to steric hindrance by the BLA and P45: a larger size of P45 may have caused greater hindrance and thereby reduced binding. On the contrary, binding to fluorescein was greatly reduced by BLA and P45 fusion, more for P45. One possibility is steric hindrance as suggested for lysozyme. Alternatively, the reduced activity of the scFvFLU fusion proteins, in particular P45-scFvFLU, may be due to the strong aggregation tendency of scFvFLU protein as described earlier. It is possible that the preparations used for inhibition assay may be a mixture of various aggregated species. Since both BLA and scFvHEL are highly soluble, their fusion must be soluble and inhibits similarly to the scFvHEL. The observed slight reduction on lysozyme inhibition for the fusion may in fact be due to steric hindrance by BLA for binding of the scFvHEL to the lysozyme. Due to a lower solubility of P45 or scFvFLU, the fusions showed greater reduction in inhibition, more likely due to enhanced aggregation of the P45 fusion proteins. The fusion with both scFvFLU and P45 was the worst as expected from the aggregation tendency of scFvFLU and the weaker ability of P45 to enhance the solubility of the fusion proteins.

Acknowledgments We are grateful to Drs. Daisuke Ejima and Haruna Sato for the assistance of SEC MALS measurements. This work was supported by a Grant in Aid for Science Research (20580372, 23580475) to M.T. from MEXT Japan, and by funding from the Institute for Fermentation, Osaka to M.T.

References

Andersen DC, Reilly DE (2004) Production technologies for monoclonal antibodies and their fragments. Curr Opin Biotechnol 15:456–462

Andersson M, Wittgren B, Wahlund KG (2003) Accuracy in multiangle light scattering measurements for molar mass and radius estimations. Model calculations and experiments. Anal Chem 75:4279–4291

Arakawa T, Tokunaga H, Yamaguchi R, Tokunaga M (2010) High solubility supports efficient refolding of thermally unfolded beta-lactamase. Int J Biol Macromol 47:706–709

Beck A, Wurch T, Bailly C, Corvaia N (2010) Strategies and challenges for the next generation of therapeutic antibodies. Nat Rev Immunol 10:345–352

Beckman RA, Weiner LM, Davis HM (2007) Antibody constructs in cancer therapy: protein engineering strategies to improve exposure in solid tumors. Cancer 109:170–179

Chon JH, Zarbis-Papastoitsis G (2011) Advances in the production and downstream processing of antibodies. N Biotechnol 28:458–463

DasSarma S, Berquist BR, Coker JA, DasSarma P, Muller JA (2006) Post-genomics of the model haloarchaeon *Halobacterium* sp. NRC-1. Saline Systems 2:3

Demarest SJ, Glaser SM (2008) Antibody therapeutics, antibody engineering, and the merits of protein stability. Curr Opin Drug Discov Devel 11:675–687

Ejima D, Yumioka R, Arakawa T, Tsumoto K (2005) Arginine as an effective additive in gel permeation chromatography. J Chromatogr A 1094:49–55

Elcock AH, McCammon JA (1998) Electrostatic contributions to the stability of halophilic proteins. J Mol Biol 280:731–748

Fujii T, Ohkuri T, Onodera R, Ueda T (2007) Stable supply of large amounts of human Fab from the inclusion bodies in *E. coli*. J Biochem 141:699–707

Fursova KK, Laman AG, Melnik BS, Semisotnov GV, Kopylov PK, Kiseleva NV, Nesmeyanov VA, Brovko FA (2009) Refolding of scFv mini-antibodies using size-exclusion chromatography via arginine solution layer. J Chromatogr B Analyt Technol Biomed Life Sci 877:2045–2051

Humphreys DP, Glover DJ (2001) Therapeutic antibody production technologies: molecules, applications, expression and purification. Curr Opin Drug Discov Devel 4:172–185

Kontermann RE (2010) Alternative antibody formats. Curr Opin Mol Ther 12:176–183

Kudou M, Ejima D, Sato H, Yumioka R, Arakawa T, Tsumoto K (2011) Refolding single-chain antibody (scFv) using lauroyl-L-glutamate as a solubilization detergent and arginine as a refolding additive. Protein Expr Purif 77:68–74

Kurucz I, Titus JA, Jost CR, Segal DM (1995) Correct disulfide pairing and efficient refolding of detergent-solubilized single-chain Fv proteins from bacterial inclusion bodies. Mol Immunol 32:1443–1452

Laemmli UK (1970) Cleavage of structural proteins during the assembly of the head of bacteriophage T4. Nature 227:680–685

Lilie H, Schwarz E, Rudolph R (1998) Advances in refolding of proteins produced in *E. coli*. Curr Opin Biotechnol 9:497–501

Mevarech M, Frolow F, Gloss LM (2000) Halophilic enzymes: proteins with a grain of salt. Biophys Chem 86:155–164

Midelfort KS, Hernandez HH, Lippow SM, Tidor B, Drennan CL, Wittrup KD (2004) Substantial energetic improvement with minimal structural perturbation in a high affinity mutant antibody. J Mol Biol 343:685–701

Mizukami M, Hanagata H, Miyauchi A (2010) *Brevibacillus* expression system: host–vector system for efficient production of secretory proteins. Curr Pharm Biotechnol 11:251–258

Ottiger M, Thiel MA, Feige U, Lichtlen P, Urech DM (2009) Efficient intraocular penetration of topical anti-TNF-alpha single-chain antibody (ESBA105) to anterior and posterior segment without penetration enhancer. Invest Ophthalmol Vis Sci 50:779–786

Shukla AA, Thommes J (2010) Recent advances in large-scale production of monoclonal antibodies and related proteins. Trends Biotechnol 28:253–261

Smith PK, Krohn RI, Hermanson GT, Mallia AK, Gartner FH, Provenzano MD, Fujimoto EK, Goeke NM, Olson BJ, Klenk DC (1985) Measurement of protein using bicinchoninic acid. Anal Biochem 150:76–85

Stockwin LH, Holmes S (2003) The role of therapeutic antibodies in drug discovery. Biochem Soc Trans 31:433–436

Takagi H, Kadowaki K, Udaka S (1989) Screening and characterization of protein-hyperproducing bacteria without detectable exoprotease activity. Agric Biol Chem 53:691–699

Tokunaga H, Ishibashi M, Arakawa T, Tokunaga M (2004) Highly efficient renaturation of beta-lactamase isolated from moderately halophilic bacteria. FEBS Lett 558:7–12

Tokunaga H, Arakawa T, Fukada H, Tokunaga M (2006a) Opposing effects of NaCl on reversibility and thermal stability of halophilic beta-lactamase from a moderate halophile, *Chromohalobacter* sp. 560. Biophys Chem 119:316–320

Tokunaga H, Oda Y, Yonezawa Y, Arakawa T, Tokunaga M (2006b) Contribution of halophilic nucleoside diphosphate kinase sequence to the heat stability of chimeric molecule. Protein Pept Lett 13:525–530

Tokunaga H, Arakawa T, Tokunaga M (2008) Engineering of halophilic enzymes: two acidic amino acid residues at the carboxy-terminal region confer halophilic characteristics to *Halomonas* and *Pseudomonas* nucleoside diphosphate kinases. Protein Sci 17:1603–1610

Tokunaga H, Arakawa T, Tokunaga M (2010a) Novel soluble expression technologies derived from unique properties of halophilic proteins. Appl Microbiol Biotechnol 88:1223–1231

Tokunaga H, Saito S, Sakai K, Yamaguchi R, Katsuyama I, Arakawa T, Onozaki K, Tokunaga M (2010b) Halophilic beta-lactamase as a new solubility- and folding-enhancing tag protein: production of native human interleukin 1alpha and human neutrophil alpha-defensin. Appl Microbiol Biotechnol 86:649–658

Tsumoto K, Nakaoki Y, Ueda Y, Ogasahara K, Yutani K, Watanabe K, Kumagai I (1994) Effect of the order of antibody variable regions on the expression of the single-chain HyHEL10 Fv fragment in *E. coli* and the thermodynamic analysis of its antigen-binding properties. Biochem Biophys Res Commun 201:546–551

Tsumoto K, Shinoki K, Kondo H, Uchikawa M, Juji T, Kumagai I (1998) Highly efficient recovery of functional single-chain Fv fragments from inclusion bodies overexpressed in *Escherichia coli* by controlled introduction of oxidizing reagent—application to a human single-chain Fv fragment. J Immunol Methods 219:119–129

Tsumoto K, Ejima D, Kumagai I, Arakawa T (2003) Practical considerations in refolding proteins from inclusion bodies. Protein Expr Purif 28:1–8

Ventosa A, Nieto JJ, Oren A (1998) Biology of moderately halophilic aerobic bacteria. Microbiol Mol Biol Rev 62:504–544

Wörn A, Plückthun A (2001) Stability engineering of antibody single-chain Fv fragments. J Mol Biol 305:989–1010

Yamada T (2011) Therapeutic monoclonal antibodies. Keio J Med 60:37–46

Yamaguchi R, Tokunaga H, Ishibashi M, Arakawa T, Tokunaga M (2011) Salt-dependent thermo-reversible alpha-amylase: cloning and characterization of halophilic alpha-amylase from moderately halophilic bacterium, *Kocuria varians*. Appl Microbiol Biotechnol 89:673–684

Yamaguchi R, Inoue Y, Tokunaga H, Ishibashi M, Arakawa T, Sumitani J, Kawaguchi T, Tokunaga M (2012) Halophilic characterization of starch-binding domain from *Kocuria varians* alpha-amylase. Int J Biol Macromol 50:95–102

Yashiro K, Lowenthal JW, O'Neil TE, Ebisu S, Takagi H (2001) High-level protein production of recombinant chicken interferon-γ by *Brevibacillus choshinensis*. Protein Expr Purif 23:113–120

Yonezawa Y, Izutsu K, Tokunaga H, Maeda H, Arakawa T, Tokunaga M (2007) Dimeric structure of nucleoside diphosphate kinase from moderately halophilic bacterium: contrast to the tetrameric *Pseudomonas* counterpart. FEMS Microbiol Lett 268:52–58

nature genetics

The landscape of somatic mutations in Down syndrome–related myeloid disorders

Kenichi Yoshida[1,2,17], Tsutomu Toki[3,17], Yusuke Okuno[1,17], Rika Kanezaki[3], Yuichi Shiraishi[4], Aiko Sato-Otsubo[1,2], Masashi Sanada[1,2], Myoung-ja Park[5], Kiminori Terui[3], Hiromichi Suzuki[1,2], Ayana Kon[1,2], Yasunobu Nagata[1,2], Yusuke Sato[1,2], RuNan Wang[3], Norio Shiba[5], Kenichi Chiba[4], Hiroko Tanaka[6], Asahito Hama[7], Hideki Muramatsu[7], Daisuke Hasegawa[8], Kazuhiro Nakamura[9], Hirokazu Kanegane[10], Keiko Tsukamoto[11], Souichi Adachi[12], Kiyoshi Kawakami[13], Koji Kato[14], Ryosei Nishimura[15], Shai Izraeli[16], Yasuhide Hayashi[5], Satoru Miyano[4,6], Seiji Kojima[7], Etsuro Ito[3,18] & Seishi Ogawa[1,2,18]

Transient abnormal myelopoiesis (TAM) is a myeloid proliferation resembling acute megakaryoblastic leukemia (AMKL), mostly affecting perinatal infants with Down syndrome. Although self-limiting in a majority of cases, TAM may evolve as non-self-limiting AMKL after spontaneous remission (DS-AMKL). Pathogenesis of these Down syndrome–related myeloid disorders is poorly understood, except for *GATA1* mutations found in most cases. Here we report genomic profiling of 41 TAM, 49 DS-AMKL and 19 non-DS-AMKL samples, including whole-genome and/or whole-exome sequencing of 15 TAM and 14 DS-AMKL samples. TAM appears to be caused by a single *GATA1* mutation and constitutive trisomy 21. Subsequent AMKL evolves from a pre-existing TAM clone through the acquisition of additional mutations, with major mutational targets including multiple cohesin components (53%), *CTCF* (20%), and *EZH2*, *KANSL1* and other epigenetic regulators (45%), as well as common signaling pathways, such as the JAK family kinases, *MPL*, *SH2B3* (*LNK*) and multiple RAS pathway genes (47%).

TAM represents a transient proliferation of immature megakaryoblasts that occurs in 5–10% of perinatal infants with Down syndrome[1,2]. Although morphologically indistinguishable from AMKL, TAM is self-limiting in the majority of cases and usually terminates spontaneously within 3–4 months of birth[1]. Hepatic infiltration of myeloid cells is a common finding and can be severe enough to be fatal, owing to hepatic failure, with liver fibrosis occurring in 5–16% of cases[2–4]. Moreover, even when spontaneous remission is achieved, approximately 20–30% of surviving infants develop DS-AMKL years after remission, although some DS-AMKL cases have no documented history of TAM[4]. In contrast to non–Down syndrome–related AMKL (non-DS-AMKL), which generally shows poor prognosis, individuals with DS-AMKL typically have a favorable prognosis. In molecular pathogenesis of these Down syndrome–related myeloid disorders, *GATA1* mutations are detected in virtually all affected infants, suggesting their central role in Down syndrome–related myeloid proliferation[5,6]. However, it is still open to question whether a *GATA1* mutation is sufficient for the development of TAM in individuals with Down syndrome, what is the cellular origin of the subsequent AMKL, whether additional gene mutations are required for progression to AMKL, and, if so, what are their gene targets, although several genes have been reported to be mutated in occasional cases with DS-AMKL, including *JAK1*, *JAK2* and *JAK3* (refs. 7–10), *TP53* (refs. 10,11), *FLT3* (ref. 8) and *MPL*[12]. We reasoned that identifying a comprehensive registry of gene mutations and tracking them at a clonal level using massively parallel sequencing would provide vital information for addressing these questions.

RESULTS

Genomic landscape of Down syndrome–related myeloid neoplasms

We performed whole-genome sequencing of 4 trios consisting of samples from TAM, AMKL and complete remission phases (**Supplementary Figs. 1** and **2** and **Supplementary Table 1**). In total,

[1]Cancer Genomics Project, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan. [2]Department of Pathology and Tumor Biology, Graduate School of Medicine, Kyoto University, Kyoto, Japan. [3]Department of Pediatrics, Hirosaki University Graduate School of Medicine, Hirosaki, Japan. [4]Laboratory of DNA Information Analysis, Human Genome Center, Institute of Medical Science, The University of Tokyo, Tokyo, Japan. [5]Department of Hematology/Oncology, Gunma Children's Medical Center, Shibukawa, Japan. [6]Laboratory of Sequence Analysis, Human Genome Center, Institute of Medical Science, The University of Tokyo, Tokyo, Japan. [7]Department of Pediatrics, Nagoya University Graduate School of Medicine, Nagoya, Japan. [8]Department of Pediatrics, St. Luke's International Hospital, Tokyo, Japan. [9]Department of Pediatrics, Hiroshima University Graduate School of Biomedical Sciences, Hiroshima, Japan. [10]Department of Pediatrics, Graduate School of Medicine, University of Toyama, Toyama, Japan. [11]Division of Neonatology, National Center for Child Health and Development, Tokyo, Japan. [12]Human Health Sciences, Graduate School of Medicine, Kyoto University, Kyoto, Japan. [13]Department of Pediatrics, Kagoshima City Hospital, Kagoshima, Japan. [14]Department of Hematology and Oncology, Children's Medical Center, Japanese Red Cross Nagoya First Hospital, Nagoya, Japan. [15]Department of Pediatrics, School of Medicine, Institute of Medical, Pharmaceutical and Health Sciences, Kanazawa University, Kanazawa, Japan. [16]Functional Genomics, Cancer Research Center, Sheba Medical Center, Tel Hashomer and Tel Aviv University, Tel Aviv, Israel. [17]These authors contributed equally to this work. [18]These authors jointly directed this work. Correspondence should be addressed to S.O. (sogawa-tky@umin.ac.jp) or E.I. (eturou@cc.hirosaki-u.ac.jp).

Received 3 May; accepted 19 August; published online 22 September 2013; corrected after print 30 October 2013; doi:10.1038/ng.2759

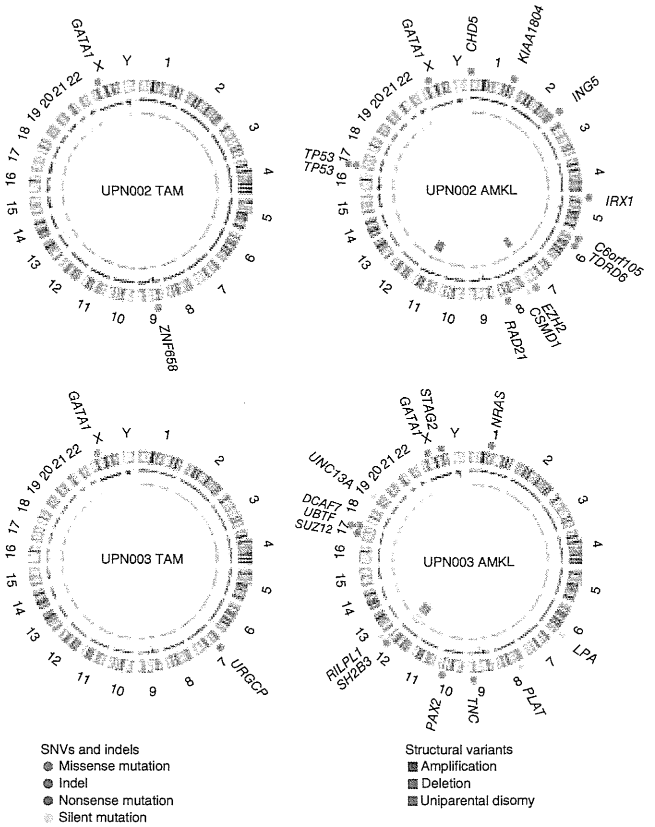


Figure 1 Representative Circos plots of paired TAM and DS-AMKL cases. Locations of somatic mutations, including of missense, frameshift, nonsense and silent mutations (colored circles), are indicated. Total (black) and allele-specific (red and green for alleles showing relatively larger and smaller copy numbers, respectively) genomic copy numbers, as well as somatic structural variants (colored bars), are indicated in the inner circle. Sample IDs are shown within each plot; plots were created with Circus[53].

we confirmed 411 single-nucleotide variants (SNVs) and 17 small nucleotide insertions and deletions (indels) by Sanger sequencing and/or deep resequencing (**Supplementary Fig. 1** and **Supplementary Table 2**). We detected only a few structural variants, including deletion, amplification and uniparental disomy, in the TAM and DS-AMKL genomes (**Fig. 1** and **Supplementary Fig. 3**). The mean number of validated somatic mutations in DS-AMKL samples (71 or 0.023 mutations/Mb) was twice the number observed in TAM samples (36 or 0.012 mutations/Mb) (**Supplementary Fig. 1a**). Mutation numbers in samples from both phases were substantially lower than in most other cancers (**Supplementary Fig. 4**), although differences in mutation rates could partly be affected by different definitions and algorithms for mutation calling. The spectrum of mutations was over-represented by C-to-T and G-to-A transitions in both TAM and DS-AMKL samples, resembling the mutational spectra in gastric and colorectal cancers[13] and in other blood cancers (**Supplementary Fig. 1b**)[14,15]. We unmasked the details of clonal evolution and expansion leading to AMKL through the use of deep sequencing of individual mutations detected by combined whole-genome and whole-exome sequencing (**Fig. 2** and **Supplementary Table 2**). Intratumoral heterogeneity was evident at initial diagnosis with TAM and in the AMKL phase in all cases (**Supplementary Fig. 5**). In UPN001, UPN002 and UPN004, AMKL evolved from one of the major subclones in the TAM phase with a shared *GATA1* mutation, as reported previously in relapsed acute myeloid leukemia (AML) in adults (**Fig. 2a,b,d**)[15]. In contrast, UPN003 showed a unique pattern of clonal evolution, in which AMKL originated from a minor subclone in the TAM phase that was totally unrelated to the predominant clone in terms of somatic mutations, with no mutation shared by both phases, and carried an independent *GATA1* mutation (**Fig. 2c**). In both scenarios, progression to AMKL seemed to be accompanied by many additional mutations, including common driver mutations that were absent in the original TAM population, indicating a multistep process of leukemogenesis.

Exome sequencing

We further investigated non-silent mutations by whole-exome sequencing of additional samples to generate a full registry of driver mutations that are relevant to the development of TAM and subsequent progression to AMKL (**Supplementary Fig. 6** and **Supplementary Table 1**). We detected *GATA1* mutations in all TAM and DS-AMKL cases, indicating sufficient sensitivity in our whole-exome analysis. In total, we confirmed 26 and 81 non-silent somatic mutations identified in the exome analysis of 15 TAM and 14 DS-AMKL samples, respectively, with 3 *GATA1* mutations common to both phases (**Supplementary Table 3**). The mean number of non-silent mutations was significantly higher in DS-AMKL samples (5.8; range of 1–11) than in TAM samples (1.7; range of 1–5) ($P = 0.0002$) (**Fig. 3a**). Of the 107 mutations, 84 were single-nucleotide substitutions that were mostly within coding sequences, except for 4 splice-site mutations. We also observed predominantly C-to-T and G-to-A transitions for non-silent substitutions (**Supplementary Fig. 7**). The remaining mutations were frameshift ($n = 21$) or non-frameshift ($n = 2$) indels, most frequently involving *GATA1* ($n = 13$). One individual with DS-AMKL (UPN004) had no SNVs or indels (**Fig. 3a**), but copy number analysis identified a large deletion at 16q involving the *CTCF* locus (**Supplementary Fig. 3**), suggesting that the alteration of *CTCF* could be a driver event in this case. Therefore, at least one additional genetic lesion other than *GATA1* mutation was detected in our whole-exome sequencing, despite the low frequency of leukemic cells appearing to show the morphology of immature megakaryoblasts (blast percentage) in many cases, which is a known characteristic of DS-AMKL samples[16,17]. Whole-exome sequencing results suggested the presence of intratumoral heterogeneity in the majority of DS-AMKL cases (**Fig. 3b**).

Spectrum of recurrent mutations in DS-AMKL

Recurrently affected genes are of primary interest in identifying driver mutations. Whereas *GATA1* was the only recurrent mutational target in TAM samples, an additional eight genes were recurrently mutated in the DS-AMKL samples, including *RAD21*, *STAG2*, *NRAS*, *CTCF*, *DCAF7*, *EZH2*, *KANSL1* and *TP53* (**Table 1**). These genes are expressed in a wide variety of hematopoietic compartments, including in both myeloid and lymphoid cells, except for *EZH2*, whose expression is largely confined to CD34$^+$ cells[18] (**Supplementary Fig. 8**). We also found that these genes were expressed in DS-AMKL cells at similar levels to common hematopoietic genes[19], although we did not observe significant difference in their expression levels in DS-AMKL and non-DS-AMKL cells (**Supplementary Fig. 9**).

We then performed targeted deep sequencing of these 8 genes in an extended set of 109 samples (including 29 samples in 25 discovery cases) consisting of 41 TAM, 49 DS-AMKL and 19 non-DS-AMKL samples (**Supplementary Tables 1** and **4**). We also included additional genes in targeted sequencing that were either functionally related to the above eight genes or were mutated only in single cases but had been previously reported to be mutated in DS-AMKL (*JAK3*) or other myeloid neoplasms (*SH2B3*, *SUZ12*, *SRSF2* and *WT1*), together with other common mutational targets in adult myeloid malignancies

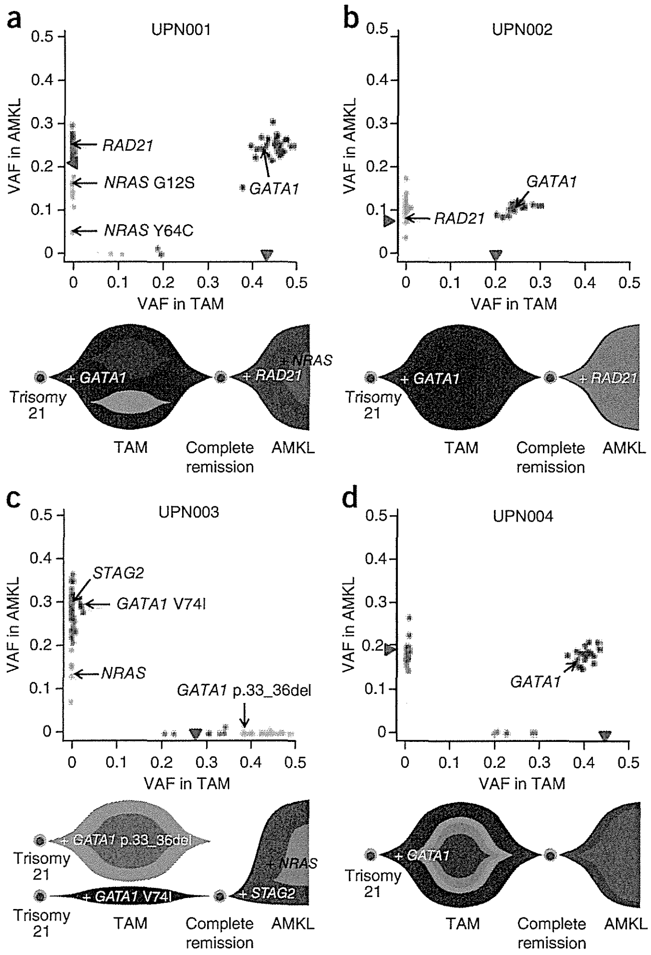


Figure 2 Clonal evolution of Down syndrome–related myeloid disorders. (**a**–**d**) Observed VAFs of validated mutations listed in **Supplementary Table 2** in both TAM and AMKL phases are shown in diagonal plots (top) for UPN001 (**a**), UPN002 (**b**), UPN003 (**c**) and UPN004 (**d**), where VAFs of genes on the X chromosome in male cases or in regions of uniparental disomy were halved. Half the value of the blast percentage, which corresponds to the allele frequency of a heterozygous mutation distributed in all tumor cells, is also shown by a red arrowhead, except for UPN003 AMKL, for which clinical data were not available. Driver mutations including in *GATA1*, *STAG2*, *RAD21* and *NRAS* are indicated by black arrows. Predicted chronological behaviors of different leukemia subclones are depicted below each diagonal plot. Distinct mutation clusters are indicated by color. In UPN001, UPN002 and UPN004, founding clones of TAM shown in blue became dominant in the AMKL samples, in which some subsequent subclones evolved through the serial acquisition of SNVs. In contrast, in UPN003, a subclone in the TAM phase (blue) and not the founding clone of TAM (aqua) became dominant in the AMKL sample. VAFs of some mutations were higher than for *GATA1* but seem to be actually equivalent to it given the error range of PCR-based deep sequencing.

(**Supplementary Fig. 10** and **Supplementary Tables 5** and **6**). We also analyzed by RT-PCR two recurrent fusion genes previously reported in non-DS-AMKL cases, *RBM15-MKL1* (*OTT-MAL*)[20,21] and *CBFA2T3-GLIS2* (refs. 22,23).

Mutations of cohesin and associated molecules

Major components of the cohesin complex, including *RAD21* and *STAG2*, were frequent targets of gene mutations in DS-AMKL (**Table 1**). Including an additional mutation in *NIPBL*, 8 of the 14 discovery DS-AMKL cases (57%) had a mutated cohesin or associated component (**Supplementary Table 3**). Cohesin is a multiprotein complex consisting of 4 core components, including the SMC1, SMC3, RAD21 and STAG proteins[24,25]. In concert with several functionally associated proteins, such as the NIPBL and ESCO proteins, cohesin is engaged in the cohesion of newly replicated sister chromatids by forming a ring-like structure[25], preventing their premature separation before late anaphase. Cohesin has also been implicated in post-replicative DNA repair and long-range regulation of gene expression[26–30]. Targeted deep sequencing confirmed recurrent mutations and deletions in all core cohesin components (*STAG2*, *RAD21*, *SMC3* and *SMC1A*) and in *NIPBL* in 26 of 49 DS-AMKL cases (53%) but in none of the 41 TAM cases, although 2 non-DS-AMKL cases (11%) had *STAG2* mutations (**Fig. 4a,b** and **Supplementary Tables 7** and **8**). Strikingly, all mutations and deletions in different cohesin components were completely mutually exclusive, suggesting that cohesin function was the common target of these mutations. All but one *STAG2* mutation (encoding a p.Arg370Gln substitution) was either a nonsense, frameshift or splice-site change (**Fig. 4a,b**, **Supplementary Figs. 11** and **12a**, and **Supplementary Table 7**). Similarly, 6 of 9 *RAD21* mutations were heterozygous nonsense or frameshift alterations. Four of the five mutations in *NIPBL*, *SMC1A* and *SMC3* were also nonsense or splice-site changes causing abnormal exon skipping (**Fig. 4a** and **Supplementary Table 7**). Thus, most of these mutations were thought to result in premature truncation, leading to loss of cohesin function. The leukemogenic mechanism of mutated cohesin components is still elusive; some studies have implicated aneuploidy caused by cohesin dysfunction in oncogenic actions[31]. However, DS-AMKL cases have been characterized by a largely normal karyotype[32]. We found no significant difference in the frequency of aneuploidy between cases with mutated and wild-type cohesin in the current DS-AMKL cohort. Many cases with mutated cohesin had completely normal karyotypes, except for constitutive trisomy 21, arguing against the hypothesis that aneuploidy has a major role in the pathogenesis of cohesin-mutated DS-AMKL (**Fig. 5a**).

CTCF mutations

Given the high frequency of cohesin mutations, new recurrent *CTCF* mutations were of particular interest because the functional interaction of cohesin and CTCF proteins has been of emerging interest in the long-range regulation of gene expression[26,30,33,34]. CTCF is a zinc-finger protein implicated in diverse regulatory functions, including transcriptional activation and/or repression, insulation, formation of chromatin barrier, imprinting and X-chromosome inactivation[35]. CTCF binds to target sequence elements and blocks the interaction of enhancers and promoters through DNA loop formation (insulator activity)[36], and several lines of evidence suggest that cohesin occupies CTCF-binding sites to contribute to the long-range regulation of gene expression by participating in the formation and stabilization of a repressive loop[26,37]. *CTCF* was mutated or deleted in ten DS-AMKL cases (20%), one TAM case (2%) and four non-DS-AMKL cases (21%), with seven mutations representing nonsense, frameshift or splice-site changes and an additional six alterations representing deletions resulting in the loss of protein function (**Fig. 4a,b**, **Supplementary Figs. 11** and **12b**, and **Supplementary Tables 7** and **8**). To our knowledge, this is the first report of frequent recurrent *CTCF* mutations in cancer, although rare mutations (occurring in approximately 2% of cases) have recently been reported in breast cancer sequencing[38].

Mutations in epigenetic regulators

EZH2, which encodes a catalytic subunit of the Polycomb repressive complex 2 (PRC2) that is responsible for di- and trimethylation of histone H3 lysine 27 (H3K27)[39], is another recurrent mutational target in DS-AMKL (**Table 1**). Inactivating mutations in *EZH2* have

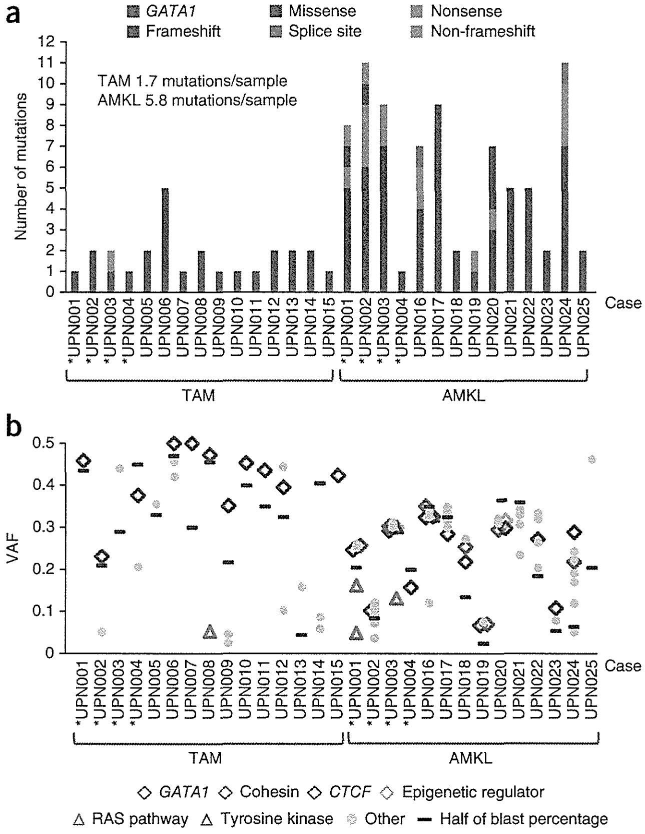


Figure 3 Somatic mutations detected by whole-exome sequencing of Down syndrome–related myeloid disorders. (**a**) Number of validated somatic mutations in 25 individuals with TAM and DS-AMKL identified by whole-exome sequencing. Paired samples are indicated by asterisks. The mutation rates per phase are given. (**b**) VAFs of individual mutations determined by deep sequencing, with VAFs adjusted for genomic copy numbers. Long indels of >3 bp were excluded from the analysis because their VAFs were difficult to accurately estimate. The VAF for each sample estimated on the basis of blast percentage is indicated by a purple horizontal bar.

been reported in up to 13% of myelodysplastic syndromes and related chronic myeloid neoplasms[40]. Although rarely mutated in adult AML[41], *EZH2* represents one of the most frequently mutated and deleted genes in childhood AMKL, as we identified mutations or deletions in 16 of 49 DS-AMKL cases (33%) and in 3 of 19 non-DS-AMKL cases (16%) (**Fig. 4a,b**, **Supplementary Fig. 12c** and **Supplementary Tables** 7 and 8). No other PRC2 components were mutated, except for *SUZ12*, which was mutated in a single DS-AMKL case (**Fig. 4a** and **Supplementary Table** 7). Although frequent mutations in other epigenetic regulators, including in *TET2*, *IDH1* or *IDH2*, *DNMT3A* and *ASXL1*, are cardinal features of myeloid neoplasms in adults, we rarely found these mutations in DS-AMKL and non-DS-AMKL cases, only identifying occasional *DNMT3A* ($n = 1$), *ASXL1* ($n = 1$) and *BCOR* ($n = 2$) mutations in DS-AMKL (**Fig. 4a**).

KANSL1 (encoding KAT8 regulatory NSL complex subunit 1; also known as MSL1V1 or NSL1) represents a new recurrent mutational target in human cancer (**Table 1**), although haploinsufficiency of *KANSL1* through germline deletions or mutations has been implicated in a congenital disease known as 17q21.31 microdeletion syndrome (MIM 610443)[42,43]. We found heterozygous mutations in *KANSL1* in three DS-AMKL and three non-DS-AMKL cases, and most of these mutations were nonsense or frameshifts, leading to loss of protein function (**Fig. 4a** and **Supplementary Table 7**). KANSL1 protein is necessary and sufficient for the activity of the KAT8 (MOF) histone acetyltransferase complex, which is engaged in the acetylation of histone H4 lysine 16 (H4K16), leading to transcriptional activation. Loss of acetylation of H4K16 has been reported to be a common hallmark of human cancer, and other histone acetyltransferases for H4K16 have been reported to form recurrent fusion partners in leukemia, including MOZ and MORF[44], suggesting a role for compromised H4K16 acetylation by *KANSL1* mutations in leukemogenesis. Of interest, KANSL1 is also responsible for the acetylation of the TP53 tumor suppressor that is important for TP53-dependent transcriptional activation[45]. KAT8 also interacts with a histone H3 lysine 4 (H3K4) methyltransferase, MLL, and the interaction of MLL and KAT8 complexes facilitates the cooperative recruitment of both complexes to gene promoters and enhances transcription initiation at target genes[45]. Thus, impaired TP53 function and/or deregulated expression of MLL gene targets could also contribute to leukemogenesis by *KANSL1* mutations.

Table 1 Recurrently mutated genes other than *GATA1* in DS-AMKL samples in whole-exome sequencing

Gene	Mutation type	RefSeq	Amino acid change	Nucleotide change	Sample (UPN) number
CTCF	Splice site	NM_006565	p.Gly318_ splice	c.953–2A>G	016
CTCF	Frameshift	NM_006565	p.Asn314fs	c.940_941insAC	020
DCAF7	Missense	NM_005828	p.Leu340Phe	c.1018C>T	001
DCAF7	Missense	NM_005828	p.Leu340Phe	c.1018C>T	003
EZH2	Frameshift	NM_004456	p.710_716del	c.2129_2148delATCACAGGA TAGGTATTTTT	001
EZH2	Missense	NM_004456	p.Arg25Gln	c.74G>A	002
KANSL1	Frameshift	NM_001193466	p.Arg720fs	c.2159_2160insCG	020
KANSL1	Nonsense	NM_001193466	p.Arg462*	c.1384C>T	024
NRAS	Missense	NM_002524	p.Gly12Ser	c.34G>A	001
NRAS	Missense	NM_002524	p.Tyr64Cys	c.191A>G	001
NRAS	Missense	NM_002524	p.Gly12Ala	c.35G>C	003
RAD21	Nonsense	NM_006265	p.Arg139*	c.415A>T	001
RAD21	Frameshift	NM_006265	p.374_375del	c.1120_1124delTCTTT	002
RAD21	Missense	NM_006265	p.Leu611Arg	c.1832T>G	018
RAD21	Nonsense	NM_006265	p.Arg65*	c.193C>T	024
STAG2	Nonsense	NM_001042750	p.Arg604*	c.1810C>T	003
STAG2	Nonsense	NM_001042750	p.Arg216*	c.646C>T	019
STAG2	Frameshift	NM_001042750	p.Asn863fs	c.2588_2589insT	020
TP53	Nonsense	NM_000546	p.Glu68*	c.202G>T	002
TP53	Non-frameshift	NM_000546	p.157_162del	c.469_486delGTCCGCGCCA TGGCCATC	002

Other mutations in DS-AMKL

RAS pathway mutations are common in hematopoietic malignancies and other human cancers but have not to our knowledge been described in DS-AMKL. In the current cohort, we identified RAS pathway

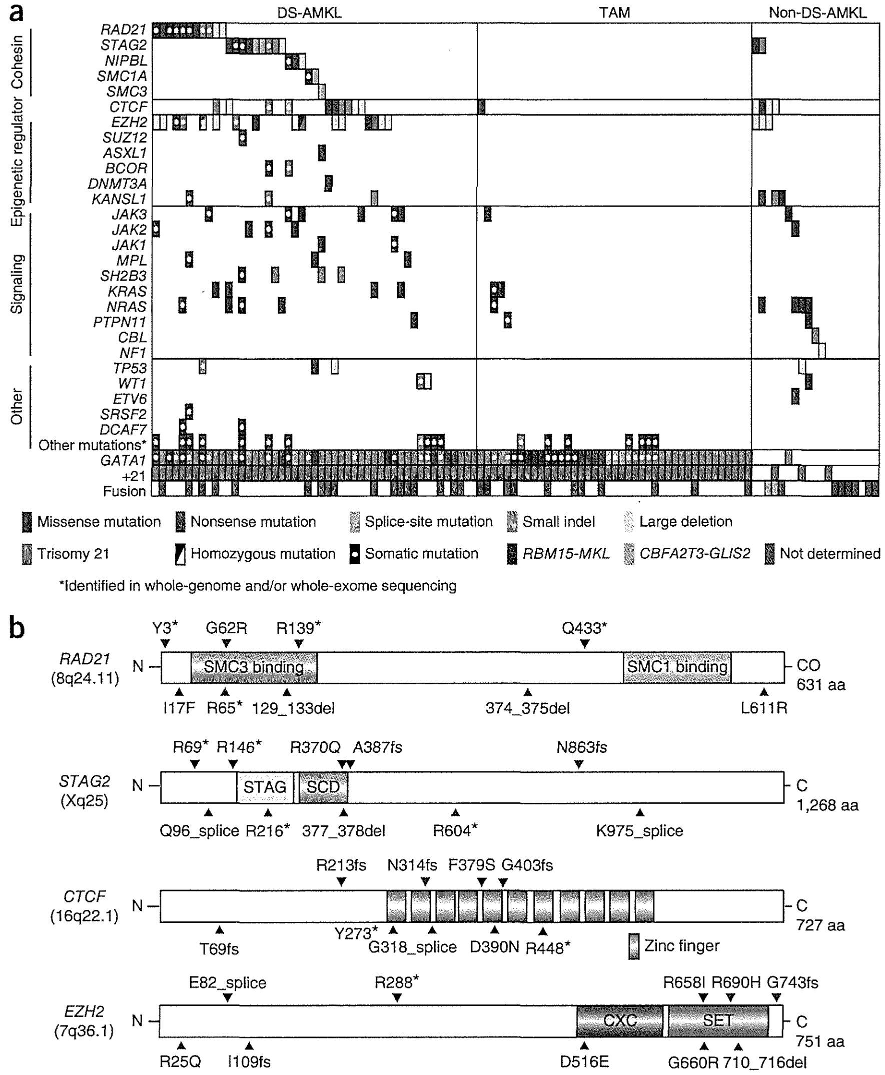


Figure 4 Driver mutations in Down syndrome–related myeloid disorders and non-DS-AMKL. (**a**) Driver mutations in 109 samples of 49 DS-AMKL, 41 TAM and 19 non-DS-AMKL cases. Types of mutations are distinguished by color. Each sample is also described in **Supplementary Table 12**. (**b**) Distribution of RAD21, STAG2, CTCF and EZH2 alterations. Alterations encoded by confirmed somatic mutations are indicated by red arrowheads.

mutations in the *NRAS*, *KRAS*, *PTPN11*, *NF1* and *CBL* genes in 8 DS-AMKL cases (16%) and 6 non-DS-AMKL cases (32%), but these mutations were rarely found in TAM cases (n = 3; 7%) (**Fig. 4a**). Tyrosine kinase and cytokine receptor mutations were also common in DS-AMKL. We found mutations in *JAK1*, *JAK2*, *JAK3*, *MPL* or *SH2B3* (*LNK*) in 17 DS-AMKL cases (35%) but rarely in TAM (n = 1) and non-DS-AMKL (n = 2) cases. We found no *FLT3* mutations in our cohort. The identified mutations were largely mutually exclusive. We found *JAK2* mutations in 4 DS-AMKL cases and 1 non-DS-AMKL case, including mutations encoding p.Val617Phe (n = 2), p.Leu611Ser (n = 1), p.Arg683Ser (n = 1) and p.Arg867Gln (n = 1); of these, *JAK2* mutations encoding p.Arg683Ser and p.Arg867Gln substitutions have been reported in acute lymphoblastic leukemia (ALL)[46,47] but not in myeloid malignancies[8,46]. Thus, we re-evaluated the diagnosis of AMKL in both UPN097 (p.Arg683Ser) and UPN023 (p.Arg867Gln), in whom the initial diagnosis of AMKL was strongly supported by typical surface marker expression of CD41, CD41b, CD117, CD13, CD33, CD34 and CD36 in UPN097 and of CD7, CD13, CD34, CD41a and CD42b in UPN023, together with characteristic cytomorphology. Similarly, the mutation encoding p.Leu611Ser was reported in both ALL[48] and polycythemia vera[49]. Thus, it seems that some *JAK2* mutations are involved in both myeloid and lymphoid leukemogenesis. As reported previously[10,11], *TP53* mutations were found in approximately 10% of DS-AMKL cases. Two identical somatic mutations found in the *DCAF7* gene (encoding p.Leu340Phe) might be interesting because the DCAF7 protein interacts with the DYRK1a kinase encoded within the Down syndrome critical region on chromosome 21 (ref. 50). DCAF7 has been shown to interact with DYRK1a through its N-terminal or C-terminal region, and the p.Leu340Phe substitution identified in our study was also located in the C-terminal domain. However, no additional mutation was detected in the extended cohort; therefore, the relevance of *DCAF7* remains to be determined.

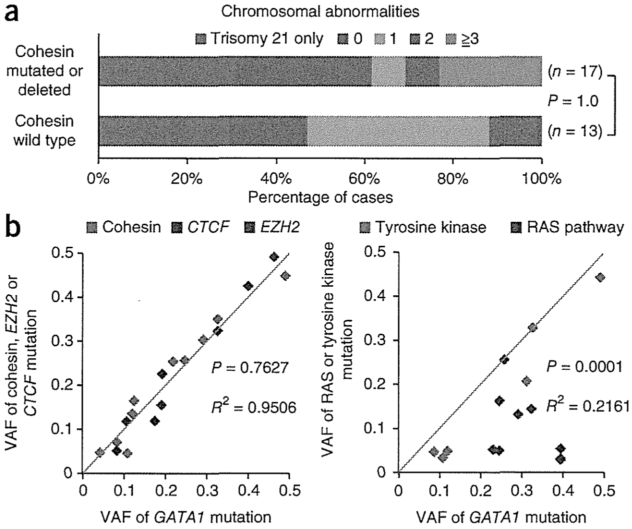
Allelic burden of major recurrent mutations relative to *GATA1* mutations

We assessed intratumoral heterogeneity and the clonal origin of mutations by calculating the variant allele frequency (VAF) of each mutation relative to that of the *GATA1* mutation using deep sequencing. Mutations in cohesin components, *CTCF* and *EZH2* showed comparable VAFs to *GATA1* mutations (**Fig. 5b**), suggesting their role in the early stage of DS-AMKL development. In contrast, RAS pathway and other tyrosine kinases and cytokine receptor mutations showed significantly lower VAFs than corresponding *GATA1* mutations (P = 0.0001) (**Fig. 5b**), indicating that they are more likely to represent subclonal mutations, which were typically preceded by mutations in cohesin components, *CTCF* and *EZH2* and were involved in the evolution of multiple DS-AMKL subclones. Although RAS and JAK pathways activated by gene mutations represent potentially druggable targets and several promising compounds are currently available, this observation may largely preclude the efficient use of such compounds in eradicating founding DS-AMKL clones.

Distinct genetic features of Down syndrome– and non–Down syndrome–related AMKL

Despite their morphological similarities, both forms of AMKL in childhood are characterized by distinctive genetic features. According to the current study and a recent report of integrated analysis of non-DS-AMKL[22], *GATA1* mutations and trisomy 21 are less common in non-DS-AMKL than in DS-AMKL cases (**Fig. 4a** and **Supplementary Table 9**). In our series, DS-AMKL was characterized by high frequencies of mutations in the cohesin complex, *EZH2* and other epigenetic regulators, as well as in JAK family kinases, which were less

Figure 5 Relationship of cohesin mutations with karyotypes and comparison of mutation loads between major gene targets in DS-AMKL and *GATA1*. (**a**) The number of chromosomal abnormalities is compared between cases with and without cohesin mutations or deletions for DS-AMKL cases. Zero signifies chromosomal abnormalities without change in chromosome count, such as partial amplification or deletion of the chromosomal region or balanced translocation. (**b**) Diagonal plots of copy number–adjusted VAFs comparing coexisting *GATA1* and other pathway mutations, including cohesin, *CTCF*, *EZH2*, tyrosine kinase and the RAS pathway mutations, as indicated by color.



common mutational targets in non-DS-AMKL. Previous studies identified recurrent *CBFA2T3-GLIS2* and *RBM15-MKL* gene fusions in non-DS-AMKL, which were found in 27% and 15.2% of non-DS-AMKL cases, respectively[22,51], whereas these fusions were not detected in DS-AMKL cases in another report (n = 10 cases)[23]. Similarly, in the current cohort, RT-PCR analysis identified 2 *CBFA2T3-GLIS2* and 3 *RBM15-MKL* fusion genes in 19 non-DS-AMKL cases but not in TAM and DS-AMKL cases (**Fig. 4a** and **Supplementary Table 10**), illustrating the genetic differences between DS-AMKL and non-DS-AMKL. In addition, our RNA sequencing of the current cases (n = 17) (**Supplementary Table 11**) also showed no *CBFA2T3-GLIS2* and *RBM15-MKL* fusions.

DISCUSSION

Whole-genome and/or whole-exome analyses and follow-up targeted sequencing identified several new aspects of the pathogenesis of Down syndrome–related myeloid proliferation. First, the initial TAM phase was characterized by a paucity of somatic mutations. The mean number of non-silent mutations per sample (1.7; range of 1–5) was surprisingly small compared with that reported in other human cancers (**Supplementary Fig. 13**), in line with a recent report that identified 1.2 (range of 1–2) mutations per sample by whole-exome sequencing in 5 TAM samples[52]. In addition to reporting a low somatic mutation frequency in their initial TAM phase, Nikolaev *et al.*[52] also reported accumulation of somatic mutations (including single cases of *SMC3* and *EZH2* mutation) during progression from TAM to DS-AMKL. Excluding common *GATA1* mutations, we identified no other recurrent mutations, with only 0.7 non-silent mutations per case, indicating that TAM could be caused by a single acquired *GATA1* mutation in addition to constitutive trisomy 21.

Intratumoral heterogeneity was evident not only in the DS-AMKL phase but also at the initial diagnosis of TAM, and subsequent DS-AMKL originated from one of the multiple subclones present in the TAM phase, usually representing the progeny of the largest subpopulation. In most cases, the DS-AMKL clone was accompanied by newly acquired driver mutations not shared by the original TAM population, generating a unique landscape of gene mutations in DS-AMKL, which was characterized by high mutational frequencies in cohesin or *CTCF* (65%), other epigenetic regulators (45%), and RAS or signal-transducing molecules (47%) (**Fig. 4a**). Tumor recurrence or evolution has not to our knowledge been characterized by the distinct gene mutations in greater detail than in the present study. In total, 44 of the 49 DS-AMKL cases had additional mutations beyond those in *GATA1* (**Fig. 4a**), even though there was a clear limitation on capturing mutations using the targeted sequencing approach.

The very high frequency of cohesin (53%) and *EZH2* (33%) mutations and deletions in DS-AMKL but not in TAM or non-DS-AMKL cases was noteworthy because the reported mutation rates of cohesin and *EZH2* in adult AML and other human cancers remain approximately 10% (refs. 14,40,41), underscoring a major role for these mutations in the pathogenesis of DS-AMKL. The leukemogenic mechanism of mutated cohesin remains elusive, and frequent *CTCF* mutations also need further evaluation to characterize their possible cooperative role with cohesin mutations[26,30,33,34]. To our knowledge, *KANSL1* mutations have not been reported previously and represent a new recurrent mutational target in human cancer, although their functional impact on AMKL development remains unknown. Evaluation of the allelic burden of these mutations by deep sequencing disclosed a clonal hierarchy among different driver mutations in which clonal mutations in cohesin, *CTCF* and epigenetic regulators frequently preceded subclonal mutations in RAS and signal transduction molecules.

In conclusion, Down syndrome–related myeloid proliferation is shaped by multiple rounds of acquisition of new mutations and clonal selection, which are initiated by a *GATA1* mutation in the TAM phase and further driven by mutation in cohesin or *CTCF*, *EZH2* or other epigenetic regulators, and RAS or signal-transducing molecules, leading to AMKL. DS-AMKL and non-DS-AMKL showed similar phenotypes but had distinct genetic features, which may underlie their different clinical characteristics.

URLs. European Genome-phenome Archive (EGA), https://www.ebi.ac.uk/ega/; EBCall, https://github.com/friend1ws/EBCall; Catalogue of Somatic Mutations in Cancer (COSMIC), http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/; PubMed, http://www.ncbi.nlm.nih.gov/pubmed/; UCSC Genome Browser, http://genome.ucsc.edu/; Integrative Genomics Viewer, http://www.broadinstitute.org/igv/; DNACopy, http://biostatistics.oxfordjournals.org/content/5/4/557.full.pdf; Genomon-fusion (in Japanese), http://genomon.hgc.jp/rna/.

METHODS

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

Accession codes. Sequencing data have been deposited in the European Genome-phenome Archive (EGA) under accession EGAS00001000546.

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

ACKNOWLEDGMENTS

We thank Y. Mori, M. Nakamura, O. Hagiwara and N. Mizota for their technical assistance. This work was supported by the Research on Measures for Intractable

Diseases Project and Health and Labor Sciences Research grants (Research on Intractable Diseases) from the Ministry of Health, Labour and Welfare, by Grants-in-Aid from the Ministry of Health, Labor and Welfare of Japan and KAKENHI (22134006, 23249052, 23118501, 23390266 and 25461579) and by the Japan Society for the Promotion of Science (JSPS) through the Funding Program for World-Leading Innovative Research and Development on Science and Technology (FIRST Program), initiated by the Council for Science and Technology Policy (CSTP) and research grants from the Japan Science and Technology Agency CREST.

AUTHOR CONTRIBUTIONS

Y.O., Y. Shiraishi, A.S.-O., K.C., H.T. and S.M. performed bioinformatics analyses of the resequencing data. M.S., A.S.-O., Y. Sato, A.H. and H.M. performed microarray experiments and analyses. R.K. and A.H. performed RT-PCR analyses. M.P., K. Terui, R.W., D.H., K.N., H.K., K. Tsukamoto, S.A., K. Kawakami, K. Kato, R.N., S.I., Y.H., S.K. and E.I. collected specimens and were involved in planning the project. K.Y., T.T., H.S., Y.N. and N.S. processed and analyzed genetic materials, prepared the library and performed sequencing. K.Y., T.T., Y.O., A.K. and S.O. generated figures and tables. E.I. and S.O. led the entire project. K.Y. and S.O. wrote the manuscript. All authors participated in discussions and interpretation of the data and results.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Khan, I., Malinge, S. & Crispino, J. Myeloid leukemia in Down syndrome. *Crit. Rev. Oncog.* **16**, 25–36 (2011).
2. Massey, G.V. *et al.* A prospective study of the natural history of transient leukemia (TL) in neonates with Down syndrome (DS): Children's Oncology Group (COG) study POG-9481. *Blood* **107**, 4606–4613 (2006).
3. Muramatsu, H. *et al.* Risk factors for early death in neonates with Down syndrome and transient leukaemia. *Br. J. Haematol.* **142**, 610–615 (2008).
4. Klusmann, J.H. *et al.* Treatment and prognostic impact of transient leukemia in neonates with Down syndrome. *Blood* **111**, 2991–2998 (2008).
5. Xu, G. *et al.* Frequent mutations in the *GATA-1* gene in the transient myeloproliferative disorder of Down syndrome. *Blood* **102**, 2960–2968 (2003).
6. Wechsler, J. *et al.* Acquired mutations in *GATA1* in the megakaryoblastic leukemia of Down syndrome. *Nat. Genet.* **32**, 148–152 (2002).
7. Walters, D.K. *et al.* Activating alleles of *JAK3* in acute megakaryoblastic leukemia. *Cancer Cell* **10**, 65–75 (2006).
8. Malinge, S. *et al.* Activating mutations in human acute megakaryoblastic leukemia. *Blood* **112**, 4220–4226 (2008).
9. Blink, M. *et al.* Frequency and prognostic implications of *JAK 1–3* aberrations in Down syndrome acute lymphoblastic and myeloid leukemia. *Leukemia* **25**, 1365–1368 (2011).
10. Hama, A. *et al.* Molecular lesions in childhood and adult acute megakaryoblastic leukaemia. *Br. J. Haematol.* **156**, 316–325 (2012).
11. Malkin, D., Brown, E.J. & Zipursky, A. The role of p53 in megakaryocyte differentiation and the megakaryocytic leukemias of Down syndrome. *Cancer Genet. Cytogenet.* **116**, 1–5 (2000).
12. Hussein, K. *et al.* MPL^{W515L} mutation in acute megakaryoblastic leukaemia. *Leukemia* **23**, 852–855 (2009).
13. Greenman, C. *et al.* Patterns of somatic mutation in human cancer genomes. *Nature* **446**, 153–158 (2007).
14. Welch, J.S. *et al.* The origin and evolution of mutations in acute myeloid leukemia. *Cell* **150**, 264–278 (2012).
15. Ding, L. *et al.* Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature* **481**, 506–510 (2012).
16. Creutzig, U. *et al.* Diagnosis and management of acute myeloid leukemia in children and adolescents: recommendations from an international expert panel. *Blood* **120**, 3187–3205 (2012).
17. Swerdlow, S.H., Jaffe, E.S. & International Agency for Research on Cancer & World Health Organization *WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues* (International Agency for Research on Cancer, Lyon, France, 2008).
18. Wu, C. *et al.* BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biol.* **10**, R130 (2009).
19. Bourquin, J.P. *et al.* Identification of distinct molecular phenotypes in acute megakaryoblastic leukemia by gene expression profiling. *Proc. Natl. Acad. Sci. USA* **103**, 3339–3344 (2006).
20. Mercher, T. *et al.* Involvement of a human gene related to the *Drosophila spen* gene in the recurrent t(1;22) translocation of acute megakaryocytic leukemia. *Proc. Natl. Acad. Sci. USA* **98**, 5776–5779 (2001).
21. Ma, Z. *et al.* Fusion of two novel genes, *RBM15* and *MKL1*, in the t(1;22)(p13;q13) of acute megakaryoblastic leukemia. *Nat. Genet.* **28**, 220–221 (2001).
22. Gruber, T.A. *et al.* An inv(16)(p13.3q24.3)-encoded CBFA2T3-GLIS2 fusion protein defines an aggressive subtype of pediatric acute megakaryoblastic leukemia. *Cancer Cell* **22**, 683–697 (2012).
23. Thiollier, C. *et al.* Characterization of novel genomic alterations and therapeutic approaches using acute megakaryoblastic leukemia xenograft models. *J. Exp. Med.* **209**, 2017–2031 (2012).
24. Gruber, S., Haering, C.H. & Nasmyth, K. Chromosomal cohesin forms a ring. *Cell* **112**, 765–777 (2003).
25. Nasmyth, K. & Haering, C.H. Cohesin: its roles and mechanisms. *Annu. Rev. Genet.* **43**, 525–558 (2009).
26. Wendt, K.S. *et al.* Cohesin mediates transcriptional insulation by CCCTC-binding factor. *Nature* **451**, 796–801 (2008).
27. Ström, L. *et al.* Postreplicative formation of cohesion is required for repair and induced by a single DNA break. *Science* **317**, 242–245 (2007).
28. Watrin, E. & Peters, J.M. The cohesin complex is required for the DNA damage-induced G2/M checkpoint in mammalian cells. *EMBO J.* **28**, 2625–2635 (2009).
29. Dorsett, D. *et al.* Effects of sister chromatid cohesion proteins on *cut* gene expression during wing development in *Drosophila*. *Development* **132**, 4743–4753 (2005).
30. Parelho, V. *et al.* Cohesins functionally associate with CTCF on mammalian chromosome arms. *Cell* **132**, 422–433 (2008).
31. Solomon, D.A. *et al.* Mutational inactivation of *STAG2* causes aneuploidy in human cancer. *Science* **333**, 1039–1043 (2011).
32. Forestier, E. *et al.* Cytogenetic features of acute lymphoblastic and myeloid leukemias in pediatric patients with Down syndrome: an iBFM-SG study. *Blood* **111**, 1575–1583 (2008).
33. Rubio, E.D. *et al.* CTCF physically links cohesin to chromatin. *Proc. Natl. Acad. Sci. USA* **105**, 8309–8314 (2008).
34. Stedman, W. *et al.* Cohesins localize with CTCF at the KSHV latency control region and at cellular c-myc and H19/Igf2 insulators. *EMBO J.* **27**, 654–666 (2008).
35. Ohlsson, R., Bartkuhn, M. & Renkawitz, R. CTCF shapes chromatin by multiple mechanisms: the impact of 20 years of CTCF research on understanding the workings of chromatin. *Chromosoma* **119**, 351–360 (2010).
36. Phillips, J.E. & Corces, V.G. CTCF: master weaver of the genome. *Cell* **137**, 1194–1211 (2009).
37. Wendt, K.S. & Peters, J.M. How cohesin and CTCF cooperate in regulating gene expression. *Chromosome Res.* **17**, 201–214 (2009).
38. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
39. Cao, R. *et al.* Role of histone H3 lysine 27 methylation in Polycomb-group silencing. *Science* **298**, 1039–1043 (2002).
40. Ernst, T. *et al.* Inactivating mutations of the histone methyltransferase gene *EZH2* in myeloid disorders. *Nat. Genet.* **42**, 722–726 (2010).
41. Patel, J.P. *et al.* Prognostic relevance of integrated genetic profiling in acute myeloid leukemia. *N. Engl. J. Med.* **366**, 1079–1089 (2012).
42. Koolen, D.A. *et al.* Mutations in the chromatin modifier gene *KANSL1* cause the 17q21.31 microdeletion syndrome. *Nat. Genet.* **44**, 639–641 (2012).
43. Zollino, M. *et al.* Mutations in *KANSL1* cause the 17q21.31 microdeletion syndrome phenotype. *Nat. Genet.* **44**, 636–638 (2012).
44. Yang, X.J. The diverse superfamily of lysine acetyltransferases and their roles in leukemia and other diseases. *Nucleic Acids Res.* **32**, 959–976 (2004).
45. Li, X., Wu, L., Corsa, C.A., Kunkel, S. & Dou, Y. Two mammalian MOF complexes regulate transcription activation by distinct mechanisms. *Mol. Cell* **36**, 290–301 (2009).
46. Bercovich, D. *et al.* Mutations of *JAK2* in acute lymphoblastic leukaemias associated with Down's syndrome. *Lancet* **372**, 1484–1492 (2008).
47. Mullighan, C.G. *et al.* JAK mutations in high-risk childhood acute lymphoblastic leukemia. *Proc. Natl. Acad. Sci. USA* **106**, 9414–9418 (2009).
48. Kratz, C.P. *et al.* Mutational screen reveals a novel JAK2 mutation, L611S, in a child with acute lymphoblastic leukemia. *Leukemia* **20**, 381–383 (2006).
49. Nussenzveig, R.H. *et al.* Detection of *JAK2* mutations in paraffin marrow biopsies by high resolution melting analysis: identification of L611S alone and in *cis* with V617F in polycythemia vera. *Leuk. Lymphoma* **53**, 2479–2486 (2012).
50. Miyata, Y. & Nishida, E. DYRK1A binds to an evolutionarily conserved WD40-repeat protein WDR68 and induces its nuclear translocation. *Biochim. Biophys. Acta* **1813**, 1728–1739 (2011).
51. de Rooij, J.D. *et al.* *NUP98/JARID1A* is a novel recurrent abnormality in pediatric acute megakaryoblastic leukemia with a distinct *HOX* gene expression pattern. *Leukemia* doi:10.1038/leu.2013.87 (27 March 2013).
52. Nikolaev, S.I. *et al.* Exome sequencing identifies putative drivers of progression of transient myeloproliferative disorder to AMKL in infants with Down Syndrome. *Blood* **122**, 554–561 (2013).
53. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).

ONLINE METHODS

Subjects and samples. Genomic DNA from 84 individuals with Down syndrome–related myeloid disorders (41 samples from the TAM phase and 49 from the AMKL phase) and 19 with non-DS-AMKL were analyzed by whole-genome and/or whole-exome and/or targeted deep sequencing. In six cases with Down syndrome–related myeloid disorders, samples were collected from both the TAM and AMKL phases. RNA sequencing was also performed for 12 of the 49 DS-AMKL cases and for 5 additional DS-AMKL cases. RNA samples were also available for RT-PCR analysis from 30 cases with TAM, 32 cases with DS-AMKL and 15 cases with non-DS-AMKL. Written informed consent was obtained from each subject's parents before sample collection (**Supplementary Note**). This study was approved by the Ethics Committees of the University of Tokyo according to the Helsinki convention. *GATA1* mutations were detected by Sanger sequencing of all TAM and DS-AMKL samples according to the previously described procedure[5]. Detailed information on subjects and samples is provided in **Supplementary Tables 1, 4, 11** and **12**. Tumor DNA was extracted from bone marrow– or peripheral blood–derived mononuclear cells at diagnosis. Genomic DNA samples from peripheral blood from subjects in remission or from nail tissues at diagnosis were used as germline controls. Genomic DNA was extracted using a QIAamp DNA Blood Mini kit and a QIAamp DNA Investigator kit (Qiagen). Total RNA was extracted using the RNeasy kit (Qiagen) with RNase-free DNase (Qiagen).

Whole-genome sequencing. DNA samples were processed for whole-exome sequencing using NEBNext DNA sample Prep Reagent (New England Biolabs) according to the modified Illumina protocol. Sequence data were generated on the Illumina HiSeq 2000 platform in 100-bp paired-end reads. Data processing and variant calling were performed as described previously[54]. All candidate variants were validated by deep sequencing.

Validation and quantitative measurements of the frequencies of mutant alleles by deep sequencing. Individual mutation sites were amplified by genomic PCR using primers tagged with NotI cleavage sites and subjected to high-throughput sequencing as described previously[55], except that target DNA was not pooled. Deep sequencing was performed using the MiSeq or HiSeq 2000 platform. Data processing was performed according to the previously described method with minor modifications[55]. Briefly, each read was aligned to a set of PCR-amplified target sequences using BLAT[56], and dichotomic variant alleles were differentially enumerated. For indels, individual reads were first aligned to each of the wild-type and indel sequences and then assigned to the one to which better alignment was obtained in terms of the number of matched bases. Each SNV and indel whose VAF in the tumor sample was equal to or greater than 2.0% and significantly higher than the frequency in the germline sample was adopted as a somatic mutation. The error size for estimated VAFs was evaluated by assuming binomial distributions in deep sequencing, which were confirmed by observed allele frequencies at heterozygous SNPs in normal DNA samples (**Supplementary Fig. 14a**), in which the variance (σ^2) ranged from 4.0–11.0 × 10^{-4} (**Supplementary Fig. 14b**).

Clustering analysis of mutations. To identify the chronological behavior of the structure of the tumor subpopulation for the TAM and AMKL phases, somatic mutations detected in both phases by whole-genome sequencing were clustered according to their VAFs as measured by deep sequencing. Copy number–adjusted deep sequencing data, in which the VAFs of genes on the X chromosome in male cases or in regions of uniparental disomy were halved, were subjected to unsupervised clustering. Six mutations located in amplified or deleted genomic regions were excluded from the analysis. Long indels of >3 bp, except for those affecting key genes such as *GATA1* and *RAD21*, and mutations in repetitive regions were excluded from the analysis because their VAFs could tend to be underestimated.

All validated mutations were grouped into three categories according to the following criteria: (i) mutations found only in TAM (VAF in AMKL < 0.02), (ii) mutations found only in AMKL (VAF in TAM < 0.02) and (iii) mutations found in both TAM and AMKL (VAF in TAM > 0.02 and VAF in AMKL > 0.02). Clustering of mutations in each category was performed using Mclust, provided as an R package, on the basis of the VAFs of the mutations in the TAM and AMKL phases, where one-dimensional clustering of mutations in categories (i) and (ii) was performed on the basis of the homoscedastic model and two-dimensional clustering was performed for mutations in category (iii) on the basis of the ellipsoidal model. The most appropriate number of clusters was determined by using the Bayesian information criterion (BIC) score. Singleton points identified by this algorithm were regarded as outliers. Clonal subpopulations within tumors were also evaluated by kernel density analysis (**Supplementary Fig. 5**), where we drew kernel density estimate plots for the VAFs of validated variants using the density function in R.

Whole-exome sequencing and detection of somatic mutations. Exome capture was performed using SureSelect Human All Exon V3 or V4 (Agilent Technologies) or the TruSeq Exome Enrichment kit (Illumina). Enriched exome fragments were then subjected to massively parallel sequencing using the Genome Analyzer IIx or HiSeq 2000 platform (Illumina). Candidate somatic mutations were detected using our in-house pipeline EBCall (Empirical Bayesian mutation Calling; see URLs)[57]. All candidates were validated by Sanger sequencing or independent deep sequencing.

PCR-based targeted deep sequencing. Deep sequencing of *DCAF7*, *EED*, *JAK1*, *JAK3*, *KANSL1*, *SH2B3*, and *SUZ12* was performed using the primers tagged with NotI cleavage sites whose sequences are listed in **Supplementary Table 6**. Data processing and variant calling were performed as described previously[58]. All candidate variants were validated by Sanger sequencing or independent deep sequencing using non-amplified DNA.

Targeted deep sequencing. In total, 39 gene targets were exhaustively examined for mutations in all 109 cases using deep sequencing (**Supplementary Table 5**). Genomic DNA (1–1.5 µg) from bone marrow–derived mononuclear cells or peripheral blood was enriched for target exons using a SureSelect custom kit (Agilent Technologies) designed to capture all of the coding exons from the 39 target genes, and high-throughput sequencing was performed on the enriched targets using the HiSeq 2000 platform with a standard 100-bp paired-end read protocol. Sequencing reads were aligned to hg19 using Burrows-Wheeler Aligner (BWA) version 0.5.8 with default parameters. The allele frequencies of SNVs and indels were calculated at each genomic position by enumerating the relevant reads with SAMtools[59]. Initially, all variants showing VAF > 0.02 were extracted and annotated using ANNOVAR[60] for further consideration if they were found in >6 reads out of >10 total reads and appeared in both plus- and minus-strand reads. For the cases for which no germline DNA was available, relevant somatic mutations were called by eliminating the following entries, unless they were registered in the Catalogue of Somatic Mutations in Cancer (COSMIC) v60 (ref. 61) or reported as somatic mutations in PubMed: (i) synonymous variants and those having ambiguous (unknown) annotations, (ii) known SNPs in public and private databases, including dbSNP131, the 1000 Genomes Project as of 23 November 2010 and our in-house database, (iii) sequencing or mapping errors, (iv) all missense SNVs with allele frequencies of 0.45–0.55 and (v) variants localized to duplicated regions found in SegDups of the UCSC Genome Browser. To eliminate sequencing errors in category (iii), we excluded all variants found in 31 normal Japanese samples at, on average, allele frequency > 0.25. Mapping errors were removed by visual inspection with the Integrative Genomics Viewer browser[62]. All candidate variants were validated by Sanger sequencing or independent deep sequencing.

Calculation of copy numbers for target exons. Letting $d_i^{j,s}$ be the sequencing depth at the *i*th nucleotide of the *j*th exon in sample *s*, the standardized depth of the *j*th exon is calculated as

$$D_j^s = k_s \sum_i d_i^{j,s}$$

where k_s is determined to satisfy

$$k_0 = \sum_j D_j^s$$

for a fixed constant k_0 (for example, $k_0 = 1$). The correlation coefficient ($R = R^{s,t}$) between two vectors D_i^s and D_i^t was calculated, where D_i^s and D_i^t represent the depth for a given sample (sample *s*) and each of the 443 samples (sample t), analyzed for other projects, with completely normal copy numbers in array–comparative genomic hybridization (aCGH; $t = 1, 2, 3,\ldots, 443$), respectively, through which a total of m_0 (= 12) control samples showing the largest R values were selected (T_m; $m = 1, 2, 3,\ldots, m_0$) and used for copy number calculation. The copy number of the ith target exon of sample s (Cn_i^s) was calculated as

$$Cn_i^s = D_i^s / \hat{D}_i^s$$

where $\hat{D}_i^s$ was calculated by averaging m_0 samples by

$$\hat{D}_i^s = \sum_{m=1}^{m_0} D_i^{T_m} / m_0$$

Copy numbers were calculated for exons with mean depth of >500. Circular binary segmentation was also used to identify discrete copy number segments using DNACopy (see URLs); segmented copy number ($\widehat{Cn_i^s}$) was defined for the ith exon of sample s. The distribution of $\widehat{Cn_i^s}$ was calculated for all samples, and exons showing $|\widehat{Cn_i^s} - E(\widehat{Cn_i^s})| > 4$ s.d. were considered to have copy number losses or gains.

Screening for *CBFA2T3-GLIS2* and *RBM15-MKL1* fusion genes. *CBFA2T3-GLIS2* and *RBM15-MKL1* fusion genes were screened by RT-PCR[22,63]. Primer sequences are given in **Supplementary Table 13**. PCR amplification was performed by 40 cycles at 94 °C for 2 min, 60 °C for 30 s and 68 °C for 1 min, followed by denaturation at 94 °C for 2 min and extension at 68 °C for 7 min.

SNP array analyses. All tumor samples subjected to whole-exome sequencing were also analyzed for copy number alterations using SNP arrays (Affymetrix GeneChip Human Mapping 250K NspI Array or Genome-Wide Human SNP Array 6.0) as described previously[10,64,65].

RT-PCR analysis of *STAG2* and *CTCF* transcripts. To confirm abnormal splicing of *CTCF* in UPN016 and UPN071 and that of *STAG2* in UPN067, RT-PCR were performed using cDNA derived from each subject, with cDNA from CMK11-5 (DS-AMKL–derived cell line with no known mutations in both genes) used as a control (**Supplementary Fig. 11**). Primer sequences are given in **Supplementary Table 14**. Total RNA (1 μg) was subjected to reverse transcription using M-MLV reverse transcriptase (Invitrogen) according to the manufacturer's instructions. Electrophoresis was performed using Experion (Bio-Rad).

RNA sequencing. Detailed information on samples is provided in **Supplementary Table 11**. Library preparation and sequencing were performed as described previously[54]. Fusion transcripts were detected using Genomon-fusion.

Gene expression analysis of recurrently mutated genes. Expression data for the recurrently mutated genes in whole-exome sequencing were retrieved from the BioGPS database[18] for normal hematopoietic cells, including whole bone marrow, CD33$^+$ myeloid cells, CD34$^+$ cells, CD19$^+$ B cells and CD4$^+$ T cells, and from published data[19] and our RNA sequencing data for DS-AMKL samples.

Statistical analysis. The number of non-silent mutations identified by whole-exome sequencing in TAM and DS-AMKL samples (**Fig. 2a**) and the number of chromosome abnormalities in DS-AMKL cases with and without cohesin mutations or deletions (**Fig. 5a**) were compared using the Mann-Whitney U test. The difference in VAF between two mutations (**Fig. 5b**) was tested by Wilcoxon signed-rank test.

54. Sato, Y. *et al.* Integrated molecular analysis of clear-cell renal cell carcinoma. *Nat. Genet.* **45**, 860–867 (2013).
55. Yoshida, K. *et al.* Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature* **478**, 64–69 (2011).
56. Kent, W.J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
57. Shiraishi, Y. *et al.* An empirical Bayesian framework for somatic mutation detection from cancer genome sequencing data. *Nucleic Acids Res.* **41**, e89 (2013).
58. Sakaguchi, H. *et al.* Exome sequencing identifies secondary mutations of *SETBP1* and *JAK3* in juvenile myelomonocytic leukemia. *Nat. Genet.* **45**, 937–941 (2013).
59. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
60. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
61. Forbes, S.A. *et al.* COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.* **39**, D945–D950 (2011).
62. Robinson, J.T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
63. Torres, L. *et al.* Acute megakaryoblastic leukemia with a four-way variant translocation originating the *RBM15-MKL1* fusion gene. *Pediatr. Blood Cancer* **56**, 846–849 (2011).
64. Nannya, Y. *et al.* A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. *Cancer Res.* **65**, 6071–6079 (2005).
65. Yamamoto, G. *et al.* Highly sensitive method for genomewide detection of allelic composition in nonpaired, primary tumor specimens by use of Affymetrix single-nucleotide-polymorphism genotyping microarrays. *Am. J. Hum. Genet.* **81**, 114–126 (2007).

Corrigendum: The landscape of somatic mutations in Down syndrome–related myeloid disorders

Kenichi Yoshida, Tsutomu Toki, Yusuke Okuno, Rika Kanezaki, Yuichi Shiraishi, Aiko Sato-Otsubo, Masashi Sanada, Myoung-ja Park, Kiminori Terui, Hiromichi Suzuki, Ayana Kon, Yasunobu Nagata, Yusuke Sato, RuNan Wang, Norio Shiba, Kenichi Chiba, Hiroko Tanaka, Asahito Hama, Hideki Muramatsu, Daisuke Hasegawa, Kazuhiro Nakamura, Hirokazu Kanegane, Keiko Tsukamoto, Souichi Adachi, Kiyoshi Kawakami, Koji Kato, Ryosei Nishimura, Shai Izraeli, Yasuhide Hayashi, Satoru Miyano, Seiji Kojima, Etsuro Ito & Seishi Ogawa

***Nat. Genet.* 45, 1293–1299 (2013); published online 22 September 2013; corrected after print 30 October 2013**

In the version of this article initially published, the discussion of cited reference 52 should also have noted that the work "reported accumulation of additional somatic mutations (including single cases of *SMC3* and *EZH2* mutation) during progression from TAM to DS-AMKL." The error has been corrected in the HTML and PDF versions of the article.

nature genetics

Recurrent mutations in multiple components of the cohesin complex in myeloid neoplasms

Ayana Kon[1], Lee-Yung Shih[2], Masashi Minamino[3], Masashi Sanada[1,4], Yuichi Shiraishi[5], Yasunobu Nagata[1], Kenichi Yoshida[1], Yusuke Okuno[1], Masashige Bando[3], Ryuichiro Nakato[3], Shumpei Ishikawa[6,7], Aiko Sato-Otsubo[1], Genta Nagae[8], Aiko Nishimoto[6], Claudia Haferlach[9], Daniel Nowak[10], Yusuke Sato[1], Tamara Alpermann[9], Masao Nagasaki[11], Teppei Shimamura[5], Hiroko Tanaka[12], Kenichi Chiba[5], Ryo Yamamoto[13], Tomoyuki Yamaguchi[13,14], Makoto Otsu[15], Naoshi Obara[16], Mamiko Sakata-Yanagimoto[16], Tsuyoshi Nakamaki[17], Ken Ishiyama[18], Florian Nolte[10], Wolf-Karsten Hofmann[10], Shuichi Miyawaki[18], Shigeru Chiba[16], Hiraku Mori[17], Hiromitsu Nakauchi[13,14], H Phillip Koeffler[19,20], Hiroyuki Aburatani[8], Torsten Haferlach[9], Katsuhiko Shirahige[3], Satoru Miyano[5,12] & Seishi Ogawa[1,4]

Cohesin is a multimeric protein complex that is involved in the cohesion of sister chromatids, post-replicative DNA repair and transcriptional regulation. Here we report recurrent mutations and deletions involving multiple components of the cohesin complex, including *STAG2*, *RAD21*, *SMC1A* and *SMC3*, in different myeloid neoplasms. These mutations and deletions were mostly mutually exclusive and occurred in 12.1% (19/157) of acute myeloid leukemia, 8.0% (18/224) of myelodysplastic syndromes, 10.2% (9/88) of chronic myelomonocytic leukemia, 6.3% (4/64) of chronic myelogenous leukemia and 1.3% (1/77) of classical myeloproliferative neoplasms. Cohesin-mutated leukemic cells showed reduced amounts of chromatin-bound cohesin components, suggesting a substantial loss of cohesin binding sites on chromatin. The growth of leukemic cell lines harboring a mutation in *RAD21* (Kasumi-1 cells) or having severely reduced expression of RAD21 and STAG2 (MOLM-13 cells) was suppressed by forced expression of wild-type RAD21 and wild-type RAD21 and STAG2, respectively. These findings suggest a role for compromised cohesin functions in myeloid leukemogenesis.

Recent genetic studies have led to the discovery of a number of new mutational targets in myeloid malignancies, unmasking unexpected roles for deregulated histone modification and DNA methylation in both acute and chronic myeloid neoplasms[1,2]. However, knowledge of the spectrum of gene mutations in myeloid neoplasms remains incomplete. We previously reported a whole-exome sequencing study of 29 paired tumor and normal samples of myeloid neoplasms with myelodysplastic features[3]. Although our major discovery was that frequent spliceosome mutations are uniquely associated with myelodysplasia phenotypes, we also identified hundreds of previously unreported gene mutations[3]. Most of those mutations affected single individuals only and are probably passenger changes. Therefore, their importance in leukemogenesis remains undetermined. However, through closer inspection of an updated list of mutations, including newly validated single-nucleotide variants, we identified additional recurrent mutations involving *STAG2*, a core component of the cohesin complex (Online Methods and **Supplementary Table 1**). In addition, we found that two other functionally related cohesin components, *STAG1* and *PDS5B*, were mutated in single specimens (**Supplementary Fig. 1**).

Cohesin is a multimeric protein complex that is conserved across species and is composed of four core subunits, SMC1, SMC3, RAD21

[1]Cancer Genomics Project, Graduate School of Medicine, The University of Tokyo, Bunkyo-ku, Tokyo, Japan. [2]Division of Hematology-Oncology, Department of Internal Medicine, Chang Gung Memorial Hospital, Chang Gung University, Taipei, Taiwan. [3]Research Center for Epigenetic Disease, Institute of Molecular and Cellular Biosciences, The University of Tokyo, Bunkyo-ku, Tokyo, Japan. [4]Department of Pathology and Tumor Biology, Graduate School of Medicine, Kyoto University, Yoshida-Konoe-cho, Kyoto-shi Sakyo-ku, Kyoto, Japan. [5]Laboratory of DNA Information Analysis, Institute of Medical Science, The University of Tokyo, Minato-ku, Tokyo, Japan. [6]Department of Pathology, The University of Tokyo, Bunkyo-ku, Tokyo, Japan. [7]Department of Genomic Pathology, Medical Research Institute, Tokyo Medical and Dental University, Bunkyo-ku, Tokyo, Japan. [8]Genome Science Division, Research Center for Advanced Science and Technology, The University of Tokyo, Meguro-ku, Tokyo, Japan. [9]Munich Leukemia Laboratory, Munich, Germany. [10]Department of Hematology and Oncology, University Hospital Mannheim, Mannheim, Germany. [11]Division of Biomedical Information Analysis, Department of Integrative Genomics, Tohoku Medical Megabank Organization, Tohoku University, Aoba-ku, Sendai, Japan. [12]Laboratory of Sequence Data Analysis, Human Genome Center, Institute of Medical Science, The University of Tokyo, Minato-ku, Tokyo, Japan. [13]Division of Stem Cell Therapy, Institute of Medical Science, The University of Tokyo, Minato-ku, Tokyo, Japan. [14]Stem Cell and Organ Regeneration Project, Exploratory Research for Advanced Technology (ERATO), Japan Science and Technology Agency (JST), Chiyoda-ku, Tokyo, Japan. [15]Stem Cell Bank, Center for Stem Cell Biology and Regenerative Medicine, Institute of Medical Science, The University of Tokyo, Minato-ku, Tokyo, Japan. [16]Department of Hematology, Faculty of Medicine, University of Tsukuba, Tsukuba, Ibaraki, Japan. [17]Division of Hematology, Department of Medicine, Showa University School of Medicine, Shinagawa-ku, Tokyo, Japan. [18]Division of Hematology, Tokyo Metropolitan Ohtsuka Hospital, Toshima-ku, Tokyo, Japan. [19]Hematology/Oncology, Cedars-Sinai Medical Center, Los Angeles, California, USA. [20]National University of Singapore, Cancer Science Institute of Singapore, Singapore. Correspondence should be addressed to S.O. (sogawa-tky@umin.ac.jp).

Received 15 February; accepted 24 July; published online 18 August 2013; doi:10.1038/ng.2731

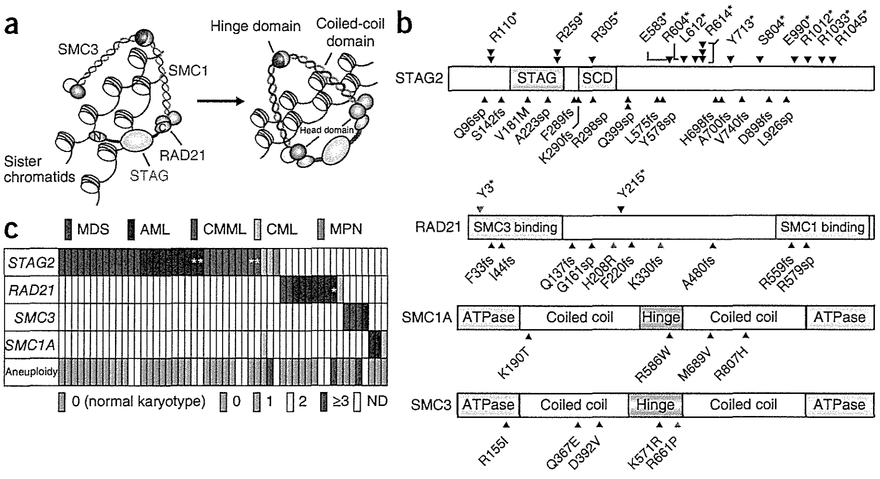


Figure 1 Genetic alterations of the cohesin complex in myeloid neoplasms. (**a**) Cohesin holds chromatin strands within a ring-like structure that is composed of four core components STAG, RAD21, SMC1 and SMC3. (**b**) Mutations in the core components of the cohesin complex found in myeloid malignancies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

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

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

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

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

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

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

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

Diseases are classified according to the World Health Organization 2008 classification. AML/MRC, AML with myelodysplasia-related changes.

[a] Two of the nine cases with *STAG2* alterations in CMML, one of the eight cases with *STAG2* alterations in *de novo* AML, one of the two cases with *STAG2* alterations in AML/MRC cases and one case with *RAD21* alteration in AML/MRC case involved genetic deletions. [b] One CML case having mutations in both *STAG2* and *SMC1A* was counted as a single case. A more detailed list is available in **Supplementary Table 8**.

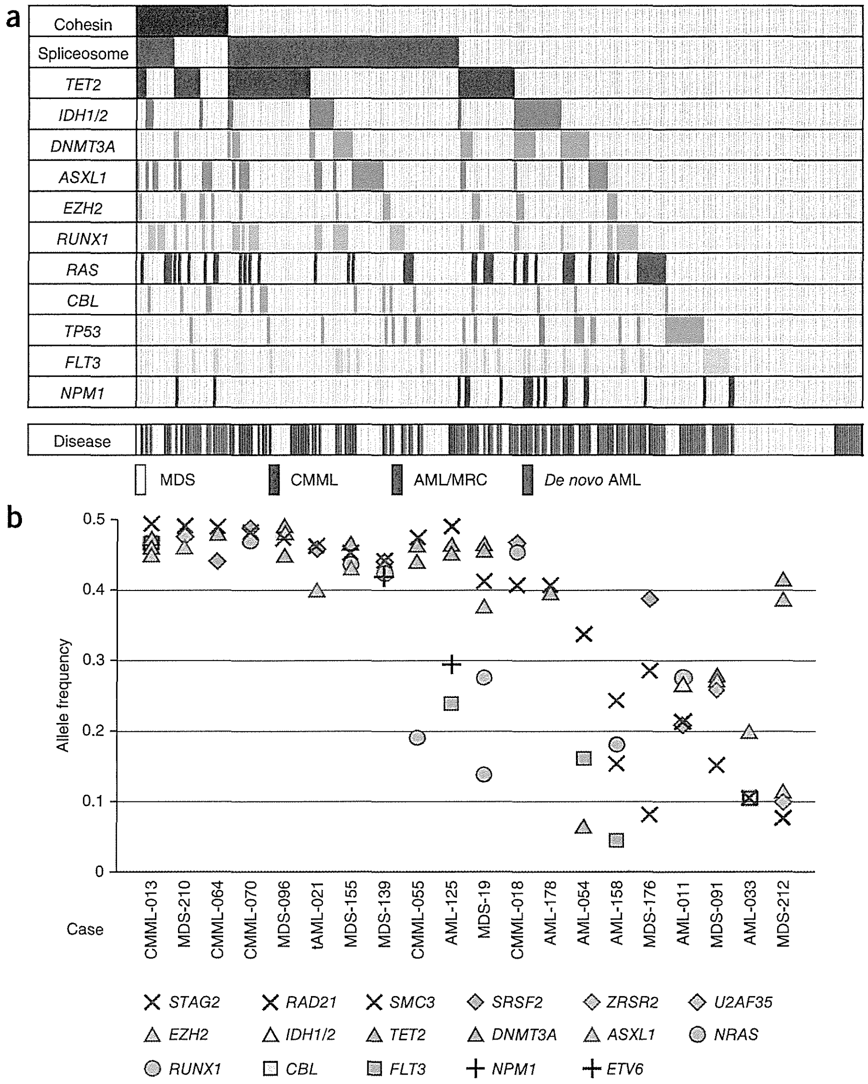


Figure 2 Relationship between cohesin mutations and other common mutations in myeloid malignancies. (**a**) Mutations in the cohesin complex and other common targets in 310 cases with different myeloid neoplasms. The corresponding disease types are shown in the bottom lane. *IDH1/2*, either *IDH1* or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

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

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

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

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