nature medicine

LETTERS

https://doi.org/10.1038/s41591-018-0267-4

Corrected: Publisher Correction

Myelodysplastic syndrome progression to acute myeloid leukemia at the stem cell level

Jiahao Chen[1], Yun-Ruei Kao[1], Daqian Sun[2,3], Tihomira I. Todorova[1], David Reynolds[4], Swathi-Rao Narayanagari[2,3], Cristina Montagna[5,6], Britta Will[1,2,7,8], Amit Verma[2,7,8,9]* and Ulrich Steidl[1,2,7,8]*

Myelodysplastic syndromes (MDS) frequently progress to acute myeloid leukemia (AML); however, the cells leading to malignant transformation have not been directly elucidated. As progression of MDS to AML in humans provides a biological system to determine the cellular origins and mechanisms of neoplastic transformation, we studied highly fractionated stem cell populations in longitudinal samples of patients with MDS who progressed to AML. Targeted deep sequencing combined with single-cell sequencing of sorted cell populations revealed that stem cells at the MDS stage, including immunophenotypically and functionally defined pre-MDS stem cells (pre-MDS-SC), had a significantly higher subclonal complexity compared to blast cells and contained a large number of aging-related variants. Single-cell targeted resequencing of highly fractionated stem cells revealed a pattern of nonlinear, parallel clonal evolution, with distinct subclones within pre-MDS-SC and MDS-SC contributing to generation of MDS blasts or progression to AML, respectively. Furthermore, phenotypically aberrant stem cell clones expanded during transformation and stem cell subclones that were not detectable in MDS blasts became dominant upon AML progression. These results reveal a crucial role of diverse stem cell compartments during MDS progression to AML and have implications for current bulk cell–focused precision oncology approaches, both in MDS and possibly other cancers that evolve from premalignant conditions, that may miss pre-existing rare aberrant stem cells that drive disease progression and leukemic transformation.

MDS are malignant, preleukemic, hematologic disorders with poor clinical outcome and a median overall survival of less than 2 years in higher-risk subtypes[1,2]. Delaying progression to secondary AML (sAML) is one of the key challenges in the clinical management of patients with MDS. The clonal origin of MDS and AML has been demonstrated to lie within the phenotypic and functionally defined stem cell compartment[3–11]. Previous seminal studies have investigated bulk tumor cells from patients with MDS, as well as fully transformed bulk cells (blasts) upon progression to sAML[12–14]. However, stem cell compartments, which represent a very small subset of total bone marrow cells, cannot be effectively interrogated by bulk sequencing even when performed at substantial depth. Clonal evolution at the stem cell level, which is crucial for MDS pathogenesis and progression to sAML, has not yet been directly examined.

To obtain direct insights into the pathogenesis of MDS and progression to sAML at the stem cell level, we utilized longitudinal, paired samples from seven patients with MDS who had later progressed to sAML (Supplementary Table 1). For both MDS and paired sAML samples, we utilized multiparameter fluorescence-activated cell sorting (FACS) to fractionate phenotypically defined malignant stem cells (MDS-SC, AML-SC) and premalignant stem cells (pre-MDS-SC, pre-AML-SC) as well as blast populations (MDS blasts, AML blasts) (Fig. 1a and Supplementary Figs. 1 and 2). Specifically, we isolated hematopoietic stem and progenitor cells (HSPC, $Lin^-CD34^+CD38^-$) expressing at least one of the LSC markers (CD45RA, CD123, or IL1RAP) that were previously identified[15–18] to enrich for MDS-SC, AML-SC (Supplementary Fig. 1a). At the same time, we isolated HSPCs that were triple-negative for CD45RA, CD123, and IL1RAP to enrich for premalignant pre-MDS-SC, pre-AML-SC (Supplementary Fig. 1a). We observed significant expansion of the phenotypic malignant stem cell population within the total HSPC population during progression from MDS to sAML; this population increased from 30.3% (MDS) to 66.9% (sAML) on average ($P < 0.001$; Supplementary Fig. 1b,c). Xenotransplantation of phenotypic MDS-SC led to predominantly myeloid engraftment ($CD33^+$) compared to pre-MDS-SC (73.2% versus 11.5%; Supplementary Fig. 3b,c), whereas phenotypic pre-MDS-SC resulted in significantly higher lymphoid engraftment ($CD19^+$) compared to MDS-SC (82.4% versus 18.8%; Supplementary Fig. 3b,c). Similar findings were obtained upon xenotransplantation of sorted pre-AML-SC and AML-SC (Supplementary Fig. 3d–f). Moreover, consistent with previous reports[19,20], we also observed significantly lower clonogenicity (Supplementary Fig. 4a,b) and increased myeloid bias (Supplementary Fig. 4c,d) of sorted MDS-SC and AML-SC compared to pre-MDS-SC and pre-AML-SC, respectively. These data indicate that CD45RA/CD123/IL1RAP-expressing HSPC are indeed enriched for malignant stem cells, and CD45RA/CD123/IL1RAP triple-negative HSPCs are enriched for premalignant stem cells in MDS and AML.

To prospectively analyze clonal evolution at the stem cell level during the progression of MDS to AML, all seven cell populations (pre-MDS-SC, MDS-SC, MDS blasts; pre-AML-SC, AML-SC, AML

[1]Department of Cell Biology, Albert Einstein College of Medicine, Bronx, NY, USA. [2]Ruth L. and David S. Gottesman Institute for Stem Cell Research and Regenerative Medicine, Albert Einstein College of Medicine, Bronx, NY, USA. [3]Stem Cell Isolation and Xenotransplantation Facility, Albert Einstein College of Medicine, Bronx, NY, USA. [4]Genomics Core Facility, Albert Einstein College of Medicine, Bronx, NY, USA. [5]Department of Genetics, Albert Einstein College of Medicine, Bronx, NY, USA. [6]Department of Pathology, Albert Einstein College of Medicine, Bronx, NY, USA. [7]Department of Medicine (Oncology), Albert Einstein College of Medicine-Montefiore Medical Center, Bronx, NY, USA. [8]Albert Einstein Cancer Center, Albert Einstein College of Medicine, Bronx, NY, USA. [9]Department of Developmental & Molecular Biology, Albert Einstein College of Medicine, Bronx, NY, USA.
*e-mail: amit.verma@einstein.yu.edu; ulrich.steidl@einstein.yu.edu

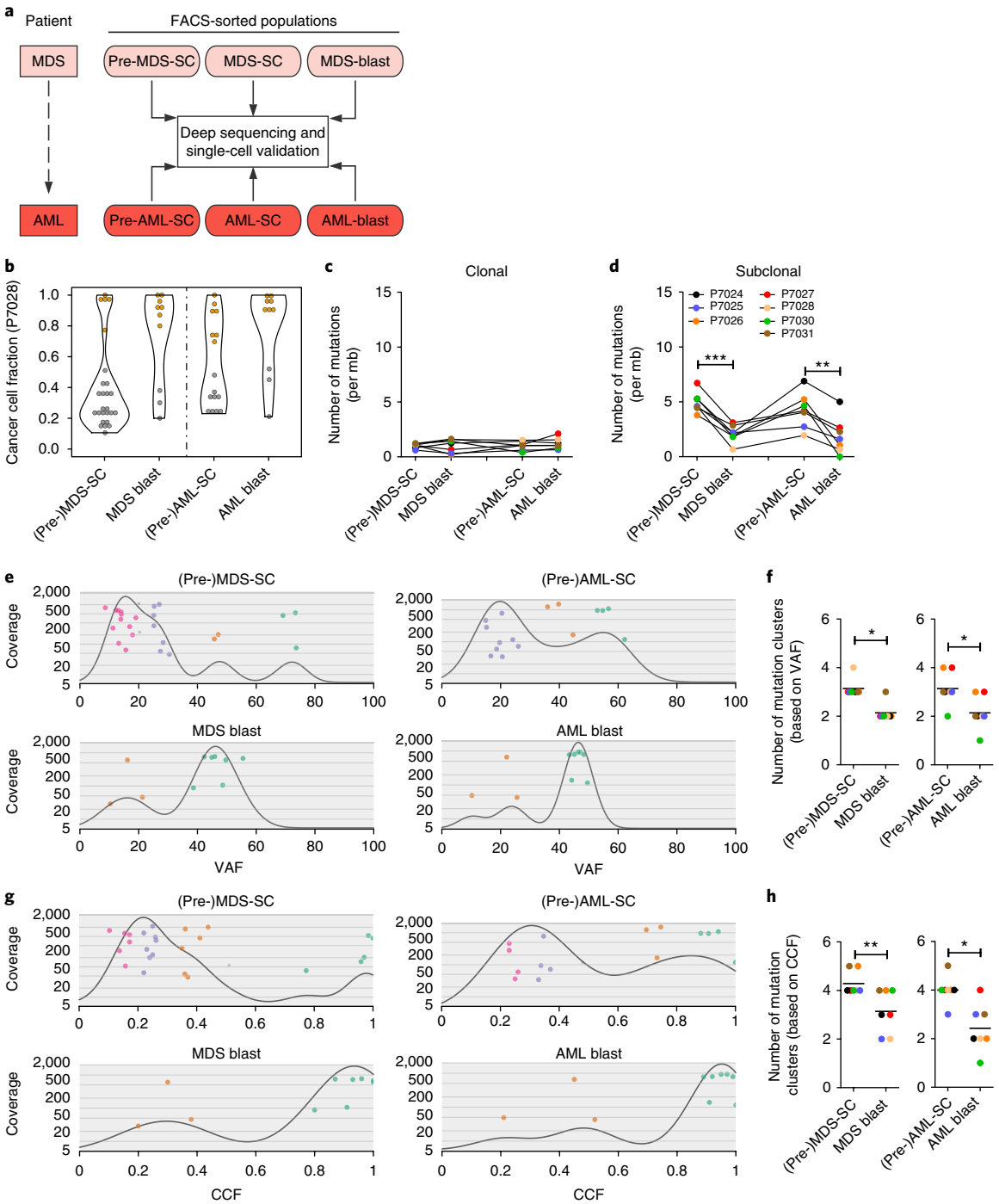

Fig. 1 | Higher subclonal diversity at the stem cell level than in blasts in patients with MDS and sAML. a, Schematics of experimental strategy of deep targeted sequencing and single-cell validation of longitudinal, paired samples from patients with MDS who later progressed to sAML. Multiparameter cell sorting was used to fractionate premalignant stem cells (pre-MDS-SC, pre-AML-SC), malignant stem cells (MDS-SC, AML-SC), and blast populations (MDS blasts, AML blasts). Nonhematopoietic cells (CD45-negative) were used as germline control for detection of somatic mutations and copy number changes. Selected mutations in each population were further examined with single-cell sequencing. **b**, Representative distribution of CCFs in stem cells (pre-MDS-SC and MDS-SC; or pre-AML-SC and AML-SC) and blasts of patient P7028, showing that stem cells had more mutations at a lower frequency than blasts for both the MDS and sAML stages, respectively. The violin plot shows the frequency distribution (kernel density) of clonal mutations (orange) and subclonal mutations (gray). **c**,**d**, Burden of clonal (**c**) and subclonal (**d**) mutations in stem cell and blast populations at the MDS ($P=0.0002$) and AML ($P=0.005$) stages across patients ($n=7$). **e**, Clonal composition of stem cell and blast populations in MDS (upper left, lower left) and sAML (upper right, lower right), respectively, in patient P7028. Based on the VAFs, mutations covered by >30× are clustered as clones and denoted with the same color. Mutation was denoted with gray if the estimated possibility of the mutation to be clustered in the subclone was lower than 0.95. **f**, Number of mutation clusters, as estimated by VAFs of mutations, in stem cells and blasts at the MDS (left, $P=0.013$) and AML (right, $P=0.021$) stages across all patients studied ($n=7$). Black bar represents the mean of clone numbers across the samples. **g**,**h**, Clonal composition of stem cell and blast populations at MDS (left, $P=0.0047$) and AML (right, $P=0.02$) estimated by CCFs of mutations ($n=7$). For **e** and **g**, from bottom to top, the horizontal lines mark sequencing depths of 5×, 10× (not labeled), 20×, 50×, 100× (not labeled), 200×, 500×, and 2,000×. Unless specified otherwise, data are mean ± s.e.m. * $P<0.05$, ** $P<0.01$, *** $P<0.001$ (two-tailed paired Student's *t*-test).

blasts; nonhematopoietic germline control) from the same patient with MDS and sAML were subjected to targeted deep sequencing with a custom panel containing the most frequently altered genes in hematologic malignancies[21] and other recent genes of interest involved in the development of MDS and AML (Fig. 1a and Supplementary Table 2). For each of the target genes, we included all of the exons, 5′ and 3′ UTRs, and the 1,000-base-pair (bp) up- and downstream regions of the gene. Prior to sequencing, we performed whole-genome amplification (WGA) of the sorted cell populations, which was shown not to distort the variant allele frequency (VAF) of mutations (Supplementary Fig. 5a,b). Targeted sequencing achieved consistent coverage across different cell populations in the same patient, ranging from 300× to 1,000× between patients (Supplementary Fig. 5c). To assess mutation patterns across different cell populations, we detected somatic mutations in each of the cell populations through comparison to the germline control (Fig. 1a and Supplementary Table 3), and validated the selected mutations by Sanger sequencing (Supplementary Fig. 5d,e). Interestingly, we found a significantly higher number of mutations, in both exonic and nonexonic regions, in stem cells than in blasts in both MDS and sAML (Supplementary Fig. 5f–h).

Thereafter, we calculated the cancer cell fraction (CCF) within each cell population, considering VAF, purity, and ploidy as previously described[22] (Supplementary Fig. 6a). Mutations were defined as 'clonal' if the 95% confidence interval of the CCF was greater than 0.95; otherwise, they were called 'subclonal'[22]. We found that, while the frequencies of clonal mutations were similar across the cell populations (Fig. 1c and Supplementary Fig. 6), the frequency of subclonal mutations was significantly higher in stem cells than in blast cells in both MDS (4.9 ± 0.92 versus 2.1 ± 0.79 per megabase; $P < 0.001$) and AML (4.2 ± 1.6 versus 1.9 ± 1.6 per megabase; $P < 0.01$) (Fig. 1d). These results indicated that, in both MDS and sAML, stem cells possess higher subclonal complexity than blast cells. Previous studies have found associations of the intrinsic mutational processes in stem cells during life with various cancers, and the burden of mutations in tissue-specific stem cells is highly correlated with age[23,24]. In addition, as several DNA repair pathways are dependent on cell cycling, relative quiescence may render stem cells vulnerable to accumulation of DNA damage during aging[25–27]. Consistent with this idea, we found that mutation patterns in both MDS and sAML stem cells were associated with DNA repair pathways in addition to association with age-related signatures (Supplementary Fig. 7).

To compare the subclonal diversity of stem cells versus blasts, we inferred the clonal architectures of stem and blast cells with Sciclone[28], using VAFs (Fig. 1e,f) as well as CCFs (Fig. 1g,h) of mutations. Interestingly, compared to blast cells, stem cells had a significantly higher total number of inferred mutation clusters (ranging from 2 to 4 versus 1 to 3; $P < 0.05$) at the MDS and sAML stages (Fig. 1e,f). Consistent findings were obtained through clonality analyses using CCFs, in that stem cells had a higher number of mutation clusters compared to the blasts (3 to 5 versus 1 to 4; $P < 0.01$) (Fig. 1g,h and Supplementary Fig. 8a–f). The difference was mainly attributable to a difference in number of non-dominant clones with lower CCFs (Fig. 1g and Supplementary Fig. 8a–f). Taken together, our results indicated that in both MDS and sAML, stem cell compartments have a higher subclonal diversity compared to blasts.

We next examined the patterns of clonal evolution during the progression from MDS to sAML of stem versus blast cell populations. Across all populations, premalignant stem cells, malignant stem cells, and blast cells, we identified shared mutations between MDS and sAML that either had high (clonal) or low (subclonal) CCFs (Supplementary Fig. 9). Interestingly, our results also revealed substantially different patterns of clonal evolution between stem cell compartments and blast cells during MDS progression to sAML in several patients (Supplementary Fig. 9). In addition, we found a somewhat variable extent of clonal evolution of pre-MDS-SC and MDS-SC in individual patients. This may also reflect the phenotypic heterogeneity of putative disease stem cells[29], which will be interesting to study in larger cohorts of patients.

We next compared clonal evolution across all cell populations and during MDS to sAML progression longitudinally. In all the patients studied, we observed one dominant clone that was shared (denoted with orange) in stem cells and blast cells at both MDS and sAML stages (Fig. 2a–g). Within these dominant clones, we found mutations in genes (for example, *TET2*, *EZH2*, *TP53*, *SETBP1*, *U2AF1*, *CSF1R*, and *KRAS*) that are frequently observed in bulk cell sequencing studies of human MDS and AML[30,31], as well as in elderly individuals with clonal hematopoiesis—albeit typically at a low subclonal size[32–34]. Interestingly, both clonal shared mutations (for example *TET2*, *EZH2*, *TP53*, *U2AF1*, *CSF1R*, and *KRAS*) and subclonal shared mutations (for example *KMT2C*, *NOTCH2*, and *FANCD2*) were detectable in T cells (Supplementary Fig. 10), indicating that these shared mutations are acquired early during MDS disease initiation and that the presence of these mutations in stem cells is still compatible with T cell differentiation. This is in line with a recent study that found clonal hematopoiesis-associated mutations, including *DNMT3A*, *TET2*, *TP53*, and *SF3B1* in virtually all hematopoietic populations, including hematopoietic stem cells (HSCs), in elderly individuals[35]. Furthermore, two recent longitudinal studies of healthy individuals who eventually developed AML also detected mutations in some of the shared dominant genes (for example, *TET2*, *TP53*, and *U2AF1*) in peripheral blood DNA many years before the actual diagnosis of AML, and the mutations were associated with increased risk of developing AML[36,37].

In line with the results above (Supplementary Fig. 8), we consistently identified more subclones at the stem cell level compared to blasts in all patients, again revealing distinct subclonal architectures between stem and blast cell compartments. Interestingly, in patient P7026, one subclone (colored red) was well detectable in pre-MDS-SC and MDS-SC, but had a frequency of only 2% in MDS blasts and then expanded to become the dominant clone across all populations upon progression to sAML (Fig. 2c). Moreover, in patients P7024 and P7030, we observed large subclones at the AML stages (colored red; Fig. 2a,f). Most interestingly, these subclones were undetectable in MDS blasts, but were inferred at frequencies of 2–3% in MDS-SC (Fig. 2a,f). Taken together, these results suggested a potential model of nonlinear clonal evolution at the stem cell level during initiation of MDS and progression to sAML: the mutational process would generate a dominant clone as well as distinct subclones at the stem cell level, and only one or a few of these clones would become apparent at the bulk/blast level (Supplementary Fig. 11).

To definitively determine the relationship between different subclones in the same population as well as clonal dynamics across all cell populations, we performed single-cell targeted sequencing of sorted stem and blast populations (Supplementary Fig. 12) with selected mutations from each of the inferred subclones (Fig. 2). We calculated the CCFs of mutations using the single-cell sequencing results and found significant correlation between the CCFs estimated by Hiseq of sorted cell populations and those determined by single-cell sequencing in all patients (Supplementary Fig. 12d–h).

Targeted deep sequencing of sorted populations from patient P7024 had identified that clonal mutations in *EZH2* and subclonal mutations (for example, *KMT2C*) were shared across all stem cell and blast populations (Fig. 3a, left and Supplementary Fig. 13a). By single-cell sequencing, we found that *EZH2* mutations were indeed present in the majority of cells across different populations, whereas *KMT2C* mutations resided in a subclone within *EZH2*-mutated cells (Fig. 3b). Interestingly, mutations in *HDAC4*, *GLI1*, and *RPL22* were present in only small subclones of MDS-SC and were not responsible for MDS blast generation or progression to sAML (Fig. 3a–c).

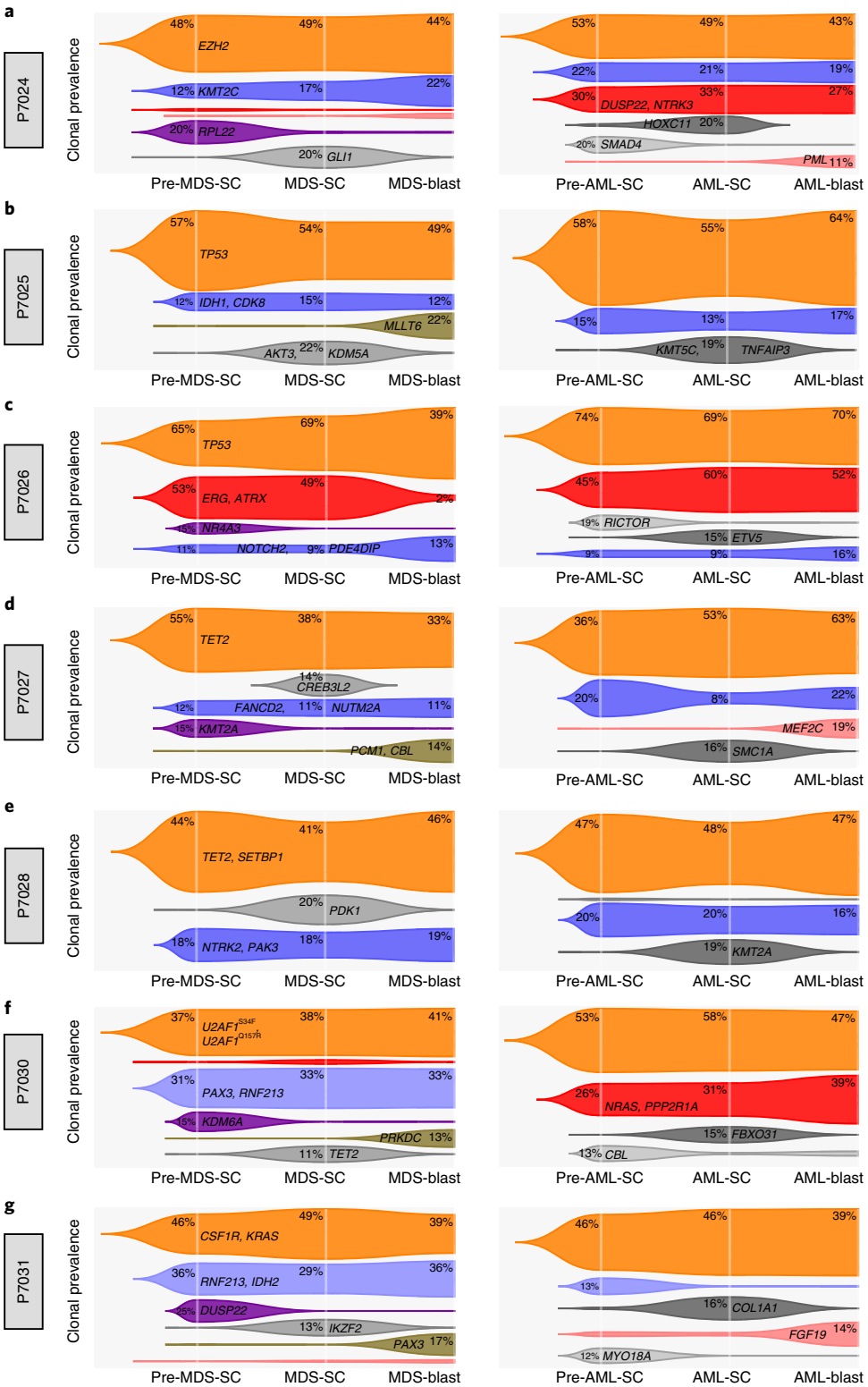

Fig. 2 | Schematic models of subclonal evolution of stem cell and blast populations during progression from MDS to sAML. a–g, Trajectory of individual clones in the different premalignant and malignant stem cell and blast populations at the MDS (left) and sAML (right) stages in individual patients: patient P7024 (**a**), patient P7025 (**b**), patient P7026 (**c**), patient P7027 (**d**), patient P7028 (**e**), patient P7030 (**f**), and patient P7031 (**g**). Clonal prevalence was defined as the mean of VAFs of mutations (as shown) in the clone estimated by SciClone. Relative clonal prevalence within the same cell population is depicted on the *y* axis in the plots. Phylogenetic relationships of different cell populations were inferred by LICHeE and visualized by the Timescape R package. The same clones in MDS and sAML are denoted by the same color within each stem or blast population of the same patient, indicating the dynamics of clonal architecture in different cell populations, as well as longitudinal clonal evolution following progression from MDS to sAML. Clones are shown if the frequency is >1% in at least one of the three populations at the MDS or sAML stages, and representative mutated genes in each clone are indicated.

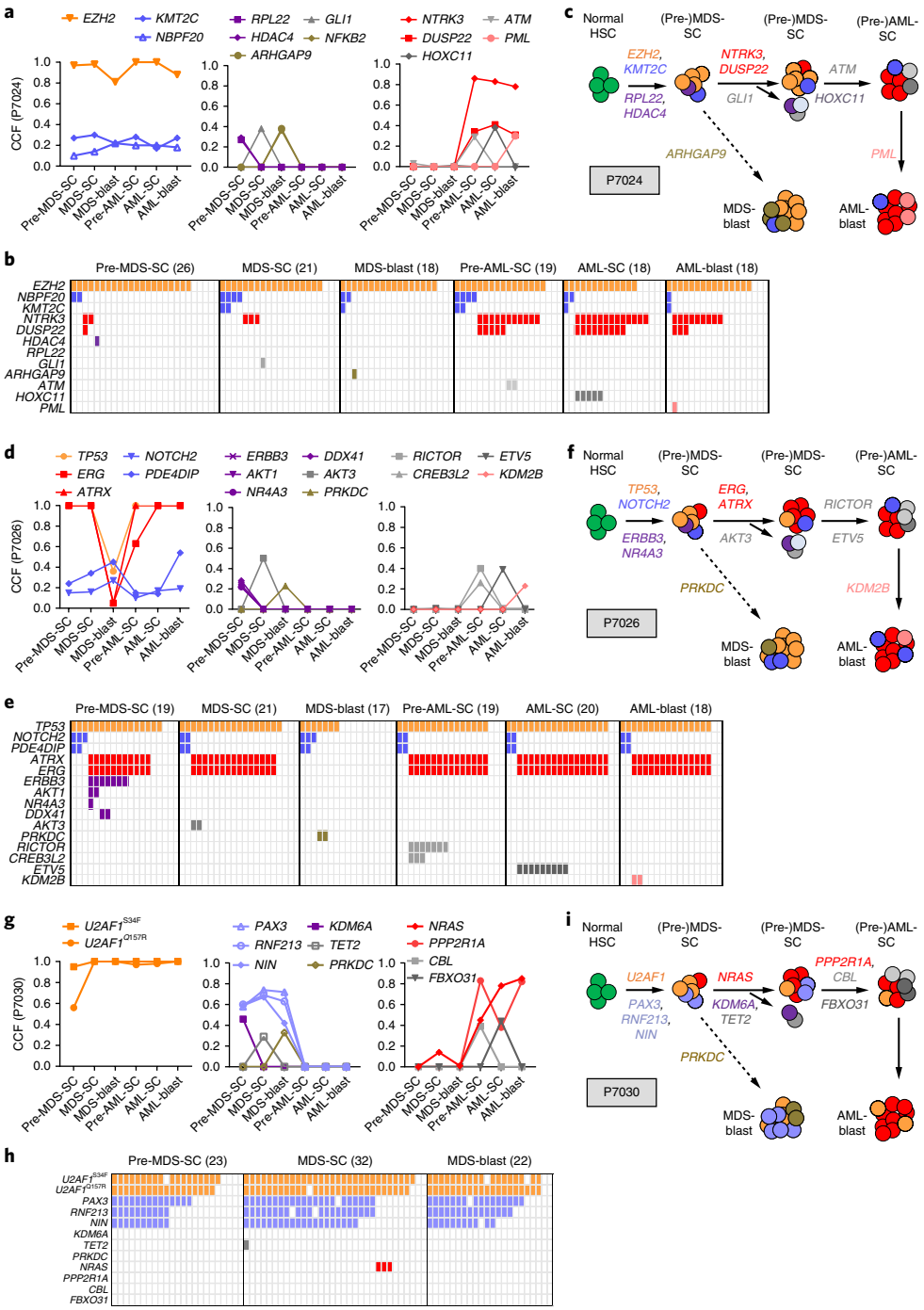

Fig. 3 | Spatiotemporal subclonal evolution during the progression from MDS to sAML determined by single-cell sequencing of sorted stem and blast cells. a, CCFs of shared (left), MDS-specific (middle), and AML-specific (right) mutations across all cell populations in patient P7024. **b**, Single-cell targeted sequencing of mutations across different cell populations in patient P7024. Each column represents the sequencing results of one single cell of the indicated cell population (pre-MDS-SC, MDS-SC, MDS blasts, pre-AML-SC, AML-SC, AML blasts), and the number of single cells tested in each population is shown in parentheses. The occurrence of mutations in a same single clone is indicated by the same color as in **a**. **c**, Schematic model of clonal evolution in different stem and blast cell populations in patient P7024. Mutations in *EZH2* were acquired early in the founding clone at the MDS stage, and acquisition of additional mutations in *NTRK3* and *DUSP22* was associated with progression to sAML, while MDS blasts were characterized by different co-mutations. In this patient, sAML developed from a rare subclone contained within MDS-SC and not through further evolution of MDS blasts. **d**, CCFs of shared (left), MDS-specific (middle), and AML-specific (right) mutations across all cell populations in patient P7026. **e**, Single-cell targeted sequencing of mutations across different cell populations in patient P7026. **f**, Schematic model of clonal evolution in different stem and blast cell populations in patient P7026. Data again indicate that the dominant clone present in sAML stem and blast cells developed from a clone within the MDS-SC that was nearly undetectable in MDS blast, indicating a crucial role of MDS-SC in sAML initiation. **g**, CCFs of shared (left), MDS-specific (middle), and AML-specific (right) mutations in different stem and blast populations at the MDS and sAML stage in patient P7030. **h**, Single-cell targeted sequencing of mutations across different cell populations in patient P7030. **i**, Schematic model of clonal evolution in different stem and blast cell populations in patient P7030. Subclones of MDS-SC with early founding mutations (that is, *U2AF1*) remained present during MDS blast generation as well as AML progression whereas other mutations, for example *PAX3*, *RNF213*, *NIN*, and *KDM6A*, occurred only in MDS but not during progression to sAML. Progression to sAML originated from a subclone of MDS-SC with *NRAS* mutation.

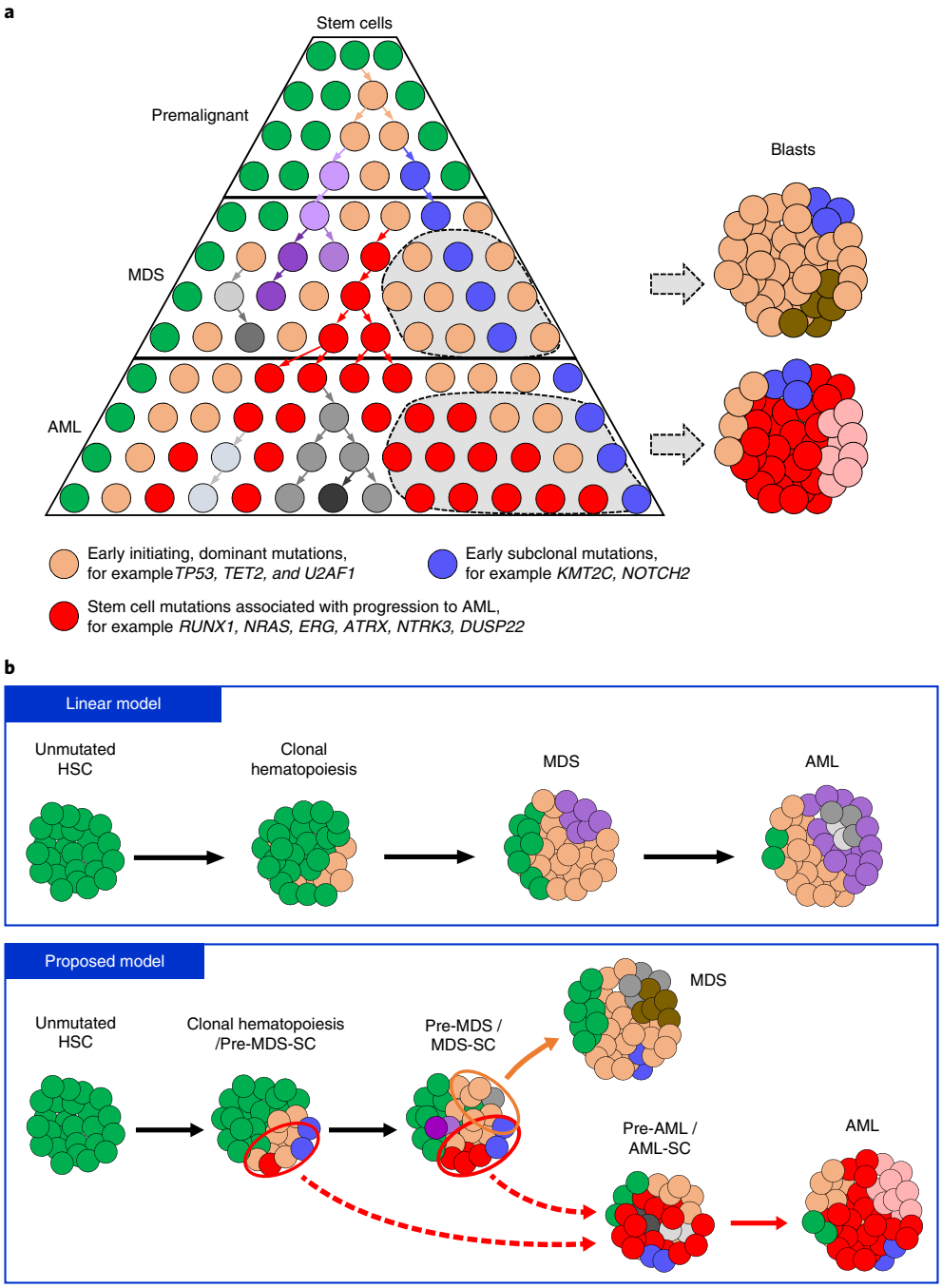

Fig. 4 | Proposed model of subclonal evolution of stem cells during the progression of MDS to sAML. a, Our results suggest a model of nonlinear clonal evolution arising from the stem cell level during development of MDS and progression to sAML. Accumulation of mutations in stem cell compartments gives rise to a highly diverse subclonal architecture (indicated by different colors) in MDS-SC. Certain subclones (orange, for example with *TP53*, *TET2*, or *U2AF1* mutations, 'clonal hematopoiesis') provide a shared basis for both MDS development (MDS blasts) as well as the formation of pre-AML-SC and AML-SC. However, pre-MDS- and MDS-SC acquire different additional mutations that then drive MDS blast formation or progression to sAML, respectively, in a nonlinear and rather parallel manner in all patients studied. In four (P7024, P7026, P7027, and P7030) out of seven cases studied, we identified that the dominant clone at the sAML stage originated from a clone (red, for example with *RUNX1*, *NRAS*, or *ERG* and *ATRX* mutations) that was detectable in pre-MDS and/or MDS-SC, but was undetectable in MDS blast cells. These results indicate that MDS-SC leading to the generation of MDS blasts can be different from those contributing to the progression to sAML, highlighting a crucial role of the entirety of the diverse MDS-SC pool in sAML disease progression, which has implications for current bulk cell–focused diagnostic and therapeutic precision oncology approaches. **b**, Schematics of different models of MDS and sAML development and progression. In comparison to the linear model (top), which suggests serial mutation accumulation during disease progression, our data support a model of parallel clonal evolution at the stem cell level during development of MDS and progression to sAML (bottom). Seven out of seven cases showed a highly diverse pool of (pre-)MDS-SC as the basis of MDS and sAML development; in four out of seven patients, we found very early branching at the (pre-)MDS-SC level towards progression to AML-SC, leading to distinct clonal composition between MDS and AML bulk cells; and three out of seven patients showed a pattern of slightly later branching (dashed red arrows) leading to more similar clonal composition between MDS and AML bulk cells compared to the early-branching cases.

Co-mutations in *NTRK3* and *DUSP22* co-occurred in AML stem and blast cell populations within EZH2-mutated cells, but were not detectable in MDS blast cells; strikingly, however, single-cell sequencing demonstrated small subclones containing these mutations within pre-MDS-SC and MDS-SC stem cell compartments (Fig. 3b,c). In AML populations, we identified mutations of *ATM* and *HOXC11* within the *NTRK3*- and *DUSP22*-mutated stem cells, whereas mutation of *PML* was observed in only a small subclone of *NTRK3*- and *DUSP22*-mutated blast cells (Fig. 3a–c). Taken together, the findings obtained by single-cell sequencing lead to a patient-specific model of clonal evolution across different stem and blast populations in MDS and sAML (Fig. 3b,c). In this patient, mutations in *EZH2* were acquired early in the founding clone at the MDS stage and acquisition of additional mutations in *NTRK3* and *DUSP22* was associated with progression to sAML (Fig. 3c), while MDS blasts were characterized by different co-mutations. Thus, sAML developed from a rare subclone contained within MDS-SC and not through further evolution of MDS blasts (Fig. 3c).

In patient P7026, we detected a shared *TP53* mutation in the majority of single cells across all cell populations (Fig. 3d,e and Supplementary Fig. 13b). We also observed a less frequent but stable subclone with co-mutations of *NOTCH2* and *PDE4DIP* within the *TP53*-mutated cells (Figs. 2b and 3d,e). On the other hand, *ERG* and *ATRX* co-mutations were present in a more frequent (dominant) clone within pre-MDS-SC and MDS-SC (Fig. 3d,e) that was distinct from the subclone with *NOTCH2* and *PDE4DIP* co-mutations. Interestingly, this subclone was nearly undetectable (VAF = 1.95%) in MDS blast bulk cell sequencing and undetectable in MDS blast single-cell sequencing (Figs. 2b and 3d,e) but became dominant in all sAML stem and blast cell populations (Fig. 3d–f), again demonstrating that the subclones contributing to the generation of MDS blasts were different from those contributing to the progression to sAML (Fig. 3e,f). Single-cell sequencing also identified two distinct subclones within the pre-MDS-SC subclone with *ERBB3* mutation, one with co-mutations of *AKT1* and *NR4A3* and another with a mutation of *DDX41* (Fig. 3e). However, none of these specific subclones persisted in MDS blasts or during sAML progression. Taken together, in this patient the dominant clone present in sAML stem and blast cells developed from a clone within the MDS-SC that was, however, undetectable in MDS blasts (Fig. 3f). Mutations of *ERG* are relatively rare in MDS and AML, and mutations of ATRX are also infrequent and found in 0.2–0.8% of the patients with MDS, but they are higher in the MDS subtype associated with α-thalassemia[38,39]. In future studies, it will be interesting to assess whether these mutations play functional roles in promoting the progression of MDS to sAML.

In patient P7030, we identified two clonal mutations in $U2AF1^{Q157R}$ and $U2AF1^{S34F}$ that were shared across all populations (Figs. 2f and 3g,h and Supplementary Fig. 13d). We also identified a relatively large subclone within the *U2AF1*-mutated cells with mutations of *PAX3*, *RNF213*, and *NIN* that was shared in all MDS populations but that did not appear at the sAML stages (Figs. 2f and 3g,h). A mutation in *NRAS* was detectable only in MDS-SC (VAF = 6.5%; Supplementary Fig. 13d) at the MDS stage (and not in MDS blasts) and resided in a subclone within *U2AF1*-mutated cells that was distinct from the *PAX3*-mutated subclone (Fig. 3h). Interestingly, this *NRAS*-mutated MDS-SC subclone then expanded at the sAML stage (Figs. 2f and 3g), accompanied by the acquisition of an additional mutation in *PPP2R1A* (VAF = 0% at MDS-SC; Supplementary Fig. 13d). In this patient, progression to sAML originated from a small subclone of *U2AF1*-mutated MDS-SC bearing the *NRAS* mutation (Fig. 3g–i). Similarly, in patient P7027, we observed that AML progression was associated with a small subclone of MDS-SC with *RUNX1* mutation (Supplementary Fig. 14). Both *NRAS* and *RUNX1* mutations are recurrent in patients with MDS and AML, with markedly higher incidence in high-risk MDS and AML[14,30,31], and *NRAS* mutations are rarely found at initial diagnosis[14,40]. Our results suggest that *NRAS* and *RUNX1* mutations may pre-exist at least in some patients, and they may reside in rare stem cell subclones at a very early disease stage.

Interestingly, in comparison with the patients shown above, we observed slightly more stable clonal evolution at the level of both stem and blast cells in patients P7025 and P7028 (Fig. 2b,e and Supplementary Fig. 15a–d). While most of the clonal mutations were shared between MDS and sAML (for example, *TET2* and *SETBP1* in P7028; *TP53* in P7025), we again observed MDS- and AML-specific mutations, respectively, in particular within MDS-SC and AML-SC (Fig. 2b,e and Supplementary Fig. 15a–d). In patient P7031, we identified clonal mutations on *CSF1R* and *KRAS* that were shared across all cell populations (Fig. 2g and Supplementary Fig. 15e,f). We also observed a larger subclone with mutations in *RNF213*, *RUNX1*, and *IDH2* that were shared in all MDS populations as well as pre-AML-SC but that did not contribute to the generation of AML blasts (Fig. 2g and Supplementary Fig. 15e–g). A $U2AF1^{Q157R}$ mutation was detected in MDS-SC and MDS blast cells with CCFs of 0.26 and 0.17, respectively, and cells with this mutation expanded upon the progression to sAML with CCFs ranging from 0.51 to 0.61 (Supplementary Fig. 15e,f). Overall, compared to patients P7024, P7026, P7027, and P7030 (Fig. 3c,f,i), the results for P7025, P7028, and P7031 revealed a model of slightly later branching of MDS-SC towards progression to sAML (Supplementary Fig. 15b,d,g).

In summary, we chose a strategy of combining rigorous cell sorting with targeted deep sequencing of both stem and blast cells from patients with MDS who progressed to sAML, which resulted in a hitherto unprecedented resolution at the stem cell level (effective depth equivalent to what could only be achieved by 250,000× to 5,000,000× deep bulk sequencing, as a result of ~0.01–0.2% frequency of sorted stem cells and average sequencing depth of approximately 500×). By both ensemble and single-cell sequencing of both stem cell and blast populations of MDS and matched sAML, we found that stem cells at the MDS stage have a significantly higher complexity of subclonal mutations compared to blast cells (Fig. 4a). Subclonal mutations mostly resided within the dominant clone with early mutations (for example, *TET2*, *TP53*, and *U2AF1*) but can dramatically increase in size towards progression to sAML, suggesting that an upfront distinction at the MDS stage of 'dominant' and 'passenger' clones/mutations solely based on clone size may not have disease pathogenetic or predictive relevance. Our findings reveal the crucial role of a diverse stem cell pool regarding full transformation and MDS blast cell generation, as well as progression to sAML, in a nonlinear and rather parallel manner (Fig. 4). These findings have implications for currently employed bulk cell–focused precision oncology approaches and provide a rationale to consider mutational examination of fractionated stem cell populations in patients with MDS, and possibly other cancers arising from premalignant conditions, to more comprehensively assess pharmacologically 'actionable' mutations relevant to later disease progression and development of AML.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability, and associated accession codes are available at https://doi.org/10.1038/s41591-018-0267-4.

Received: 7 August 2018; Accepted: 23 October 2018;
Published online: 3 December 2018

References

1. Greenberg, P. L. et al. Revised international prognostic scoring system for myelodysplastic syndromes. *Blood* **120**, 2454–2465 (2012).
2. Ades, L., Itzykson, R. & Fenaux, P. Myelodysplastic syndromes. *Lancet* **383**, 2239–2252 (2014).

3. Fialkow, P. J. et al. Clonal development, stem-cell differentiation, and clinical remissions in acute nonlymphocytic leukemia. *N. Engl. J.Med.* **317**, 468–473 (1987).
4. Nilsson, L. et al. Involvement and functional impairment of the $CD34^{+}CD38^{-}$ Thy-1^{+} hematopoietic stem cell pool in myelodysplastic syndromes with trisomy 8. *Blood* **100**, 259–267 (2002).
5. Steidl, U. et al. Essential role of Jun family transcription factors in PU.1 knockdown-induced leukemic stem cells. *Nat. Genet.* **38**, 1269–1277 (2006).
6. Will, B. et al. Stem and progenitor cells in myelodysplastic syndromes show aberrant stage-specific expansion and harbor genetic and epigenetic alterations. *Blood* **120**, 2076–2086 (2012).
7. Jan, M. et al. Clonal evolution of preleukemic hematopoietic stem cells precedes human acute myeloid leukemia. *Sci. Transl. Med.* **4**, 149ra118 (2012).
8. Pang, W. W. et al. Hematopoietic stem cell and progenitor cell mechanisms in myelodysplastic syndromes. *Proc. Natl Acad. Sci. USA* **110**, 3011–3016 (2013).
9. Corces-Zimmerman, M. R., Hong, W. J., Weissman, I. L., Medeiros, B. C. & Majeti, R. Preleukemic mutations in human acute myeloid leukemia affect epigenetic regulators and persist in remission. *Proc. Natl Acad. Sci. USA* **111**, 2548–2553 (2014).
10. Shlush, L. I. et al. Identification of pre-leukaemic haematopoietic stem cells in acute leukaemia. *Nature* **506**, 328–333 (2014).
11. Will, B. et al. Minimal PU.1 reduction induces a preleukemic state and promotes development of acute myeloid leukemia. *Nat. Med.* **21**, 1172–1181 (2015).
12. Walter, M. J. et al. Clonal architecture of secondary acute myeloid leukemia. *N. Engl. J. Med.* **366**, 1090–1098 (2012).
13. Walter, M. J. et al. Clonal diversity of recurrently mutated genes in myelodysplastic syndromes. *Leukemia* **27**, 1275–1282 (2013).
14. Makishima, H. et al. Dynamics of clonal evolution in myelodysplastic syndromes. *Nat. Genet.* **49**, 204–212 (2017).
15. Goardon, N. et al. Coexistence of LMPP-like and GMP-like leukemia stem cells in acute myeloid leukemia. *Cancer Cell* **19**, 138–152 (2011).
16. Jordan, C. et al. The interleukin-3 receptor alpha chain is a unique marker for human acute myelogenous leukemia stem cells. *Leukemia* **14**, 1777 (2000).
17. Barreyro, L. et al. Overexpression of IL-1 receptor accessory protein in stem and progenitor cells and outcome correlation in AML and MDS. *Blood* **120**, 1290–1298 (2012).
18. Mitchell, K. et al. IL1RAP potentiates multiple oncogenic signaling pathways in AML. *J. Exp. Med.* **215**, 1709–1727 (2018).
19. Jan, M. et al. Prospective separation of normal and leukemic stem cells based on differential expression of TIM3, a human acute myeloid leukemia stem cell marker. *Proc. Natl Acad. Sci. USA* **108**, 5009–5014 (2011).
20. Chung, S. S. et al. CD99 is a therapeutic target on disease stem cells in myeloid malignancies. *Sci.Transl. Med.* **9**, eaaj2025 (2017).
21. He, J. et al. Integrated genomic DNA/RNA profiling of hematologic malignancies in the clinical setting. *Blood* **127**, 3004–3014 (2016).
22. McGranahan, N. et al. Clonal status of actionable driver events and the timing of mutational processes in cancer evolution. *Sci. Transl. Med.* **7**, 283ra254 (2015).
23. Blokzijl, F. et al. Tissue-specific mutation accumulation in human adult stem cells during life. *Nature* **538**, 260–264 (2016).
24. Adams, P. D., Jasper, H. & Rudolph, K. L. Aging-induced stem cell mutations as drivers for disease and cancer. *Cell. Stem. Cell.* **16**, 601–612 (2015).
25. Rossi, D. J. et al. Deficiencies in DNA damage repair limit the function of haematopoietic stem cells with age. *Nature* **447**, 725–729 (2007).
26. Mandal, P. K., Blanpain, C. & Rossi, D. J. DNA damage response in adult stem cells: pathways and consequences. *Nat. Rev. Mol. Cell Biol.* **12**, 198–202 (2011).
27. Mohrin, M. et al. Hematopoietic stem cell quiescence promotes error-prone DNA repair and mutagenesis. *Cell. Stem. Cell.* **7**, 174–185 (2010).
28. Miller, C. A. et al. SciClone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. *PLoS Comput. Biol.* **10**, e1003665 (2014).
29. Yanagisawa, B., Ghiaur, G., Smith, B. D. & Jones, R. J. Translating leukemia stem cells into the clinical setting: harmonizing the heterogeneity. *Exp. Hematol.* **44**, 1130–1137 (2016).
30. Haferlach, T. et al. Landscape of genetic lesions in 944 patients with myelodysplastic syndromes. *Leukemia* **28**, 241–247 (2014).
31. Cancer Genome Atlas Research, N. et al. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N. Engl. J. Med.* **368**, 2059–2074 (2013).
32. Xie, M. et al. Age-related mutations associated with clonal hematopoietic expansion and malignancies. *Nat. Med.* **20**, 1472–1478 (2014).
33. Genovese, G. et al. Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N. Engl. J. Med.* **371**, 2477–2487 (2014).
34. Jaiswal, S. et al. Age-related clonal hematopoiesis associated with adverse outcomes. *N. Engl. J. Med.* **371**, 2488–2498 (2014).
35. Arends, C. M. et al. Hematopoietic lineage distribution and evolutionary dynamics of clonal hematopoiesis. *Leukemia* **32**, 1908–1919 (2018).
36. Abelson, S. et al. Prediction of acute myeloid leukaemia risk in healthy individuals. *Nature* **559**, 400–404 (2018).
37. Desai, P. et al. Somatic mutations precede acute myeloid leukemia years before diagnosis. *Nat. Med.* **24**, 1015 (2018).
38. Herbaux, C. et al. Incidence of ATRX mutations in myelodysplastic syndromes, the value of microcytosis. *Am. J. Hematol.* **90**, 737–738 (2015).
39. Steensma, D. P., Higgs, D. R., Fisher, C. A. & Gibbons, R. J. Acquired somatic ATRX mutations in myelodysplastic syndrome associated with alpha thalassemia (ATMDS) convey a more severe hematologic phenotype than germline ATRX mutations. *Blood* **103**, 2019–2026 (2004).
40. Bacher, U., Haferlach, T., Kern, W., Haferlach, C. & Schnittger, S. A comparative study of molecular mutations in 381 patients with myelodysplastic syndrome and in 4130 patients with acute myeloid leukemia. *Haematologica* **92**, 744–752 (2007).

Acknowledgements

We thank P. Schultes from the Department of Cell Biology for expert technical assistance. We thank A. Fiallo from the Einstein Genomics Core Facility for technical assistance in single-cell targeted sequencing, and S. Maqbool and S. Mi from Einstein Epigenomics Core Facility for assistance in targeted sequencing with the HiSeq platform. We thank V. Thiruthuvanathan from the Department of Cell Biology for assistance in processing the patient samples. We also thank W. Li for advice regarding whole-genome amplification, and F. C. Chan, C. Steidl, and H. Steidl for helpful discussion. This work was supported by NIH grants no. R01CA166429, no. R01CA217092 (to U.S.), no. R01HL139487, no. R01DK103961 (to A.V.), and no. K01DK105134 (to B.W.); Translational Research Program grants from the Leukemia & Lymphoma Society (to U.S. and A.V., respectively); a research grant from the Taub Foundation for MDS Research (to U.S.); and a research grant from the Evans Foundation (to A.V.). J.C. was supported by The Einstein Training Program in Stem Cell Research from the Empire State Stem Cell Fund through New York State Department of Health Contract (no. C30292GG). U.S. is a Research Scholar of the Leukemia and Lymphoma Society and the Diane and Arthur B. Belfer Faculty Scholar in Cancer Research of the Albert Einstein College of Medicine. This work was supported through the Albert Einstein Cancer Center core support grant (no. P30CA013330).

Author contributions

J.C., U.S., and A.V. designed the study and analyzed and interpreted data. J.C., Y.K., and T.I.T. collected and analyzed clinical samples. J.C., Y.K., D.S., S.N., and B.W. performed the FACS experiments. J.C. and S.N. performed the xenotransplantation assays. J.C. performed the methylcellulose assay and TCR sequencing. J.C. and D.R. performed single-cell targeted sequencing. C.M., A.V., and U.S. designed the targeted capture panel. J.C. analyzed the sequencing data. J.C., A.V., and U.S. wrote the manuscript. All authors reviewed and approved the final version of the manuscript.

Competing interests

U.S. has received research funding from GlaxoSmithKline, Bayer Healthcare, Aileron Therapeutics, and Novartis; has received compensation for consultancy services and for serving on scientific advisory boards from GlaxoSmithKline, Bayer Healthcare, Celgene, Aileron Therapeutics, Stelexis Therapeutics, and Pieris Pharmaceuticals; and has equity ownership in and is serving on the board of directors of Stelexis Therapeutics. A.V. has received research funding from GlaxoSmithKline, Incyte, MedPacto, Novartis, and Eli Lilly and Company, has received compensation as a scientific advisor to Novartis, Stelexis Therapeutics, Acceleron Pharma, and Celgene, and has equity ownership in Stelexis Therapeutics. B.W. has received research support from Novartis Pharmaceuticals.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41591-018-0267-4.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to A.V. or U.S.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Multiparameter high-speed FACS of stem and blast cells from patient samples. Bone marrow (BM) samples from seven patients with MDS and matched sAML were obtained, after written informed consent, from Montefiore Medical Center/Albert Einstein Cancer Center (Institutional Review Board no. 11-02-060E; for patient characteristics see Supplementary Table 1). All studied patients received treatment with hypomethylating agents between MDS and AML progression. Frozen BM aspirates were thawed in a water bath at 37 °C and resuspended in Iscove's modified Dulbecco's medium (IMDM) supplemented with 2% fetal bovine serum (FBS). After repeated washes with IMDM 2% FBS, cells were resuspended in MACS buffer (phosphate buffered saline supplemented with 0.5% bovine serum albumin and 2 mM EDTA, pH 7.2). Thereafter, $CD34^+$ were immunomagnetically separated with Miltenyi MACS technology (130-046-702, Miltenyi Biotec) according to the manufacturer's protocol. $CD34^+$-enriched cells were stained for 30 min on ice with antibodies: PE-Cy5 (Tri-Color)-conjugated lineage markers (CD2, CD3, CD4, CD7, CD8, CD10, CD11b, CD14, CD19, CD20, CD56, Glycophorin A), APC-conjugated blast marker CD33, and hematopoietic stem and progenitor markers (Pacific blue CD34, PE-Cy7 CD38, FITC CD45RA, Alexa Fluor 700 CD123, and PE IL1RAP). A list of antibodies is provided in Supplementary Table 4. We used Lin^-CD34^+ $CD38^-CD45RA^-CD123^-IL1RAP^-$ to enrich for pre-MDS-SC or pre-AML-SC and Lin^-CD34^+ $CD38^-$ ($CD45RA^+$ and/or $CD123^+$ and/or $IL1RAP^+$) to enrich for MDS-SC or AML-SC. Cells were also stained with PE CD45, APC CD33, and pacific orange CD4 to isolate blast cells ($CD45^+$ $CD33^+$), T cells ($CD45^+$ $CD4^+$), and nonhematopoietic cells ($CD45^-$) as germline control for somatic variant calling. Interpatient heterogeneity in the profile of surface markers for disease-relevant stem cells has been observed in patients with MDS and AML[41,42], suggesting that there is a need to utilize a combination of surface markers. In addition, the coexistence of residual normal HSC and numerous subclones of partially transformed pre-MDS-SC as well as fully transformed MDS-SC makes their distinction challenging based on phenotypic markers in individual patients. Isolation of cell populations based on phenotypic markers remains a relative enrichment strategy, which requires functional and genetic validation. Xenografting experiments with the respective populations (Supplementary Fig. 3) demonstrated functionality consistent with pre-MDS-SC versus MDS-SC properties. In addition, the fact that the sorting strategy described here was able to detect relevant mutations in pre-MDS-SC and MDS-SC indicates the validity of the strategy, at least in this cohort of patients. It will be interesting to further validate this sorting scheme for pre-MDS-SC in larger patient cohorts in the future.

Methylcellulose assay. To assess the differentiation potential of phenotypic premalignant stem cells (pre-MDS/AML-SC) and malignant stem cells (MDS/AML-SC), cells were FACS-sorted from additional patients with the same strategy (Supplementary Fig. 1a) and plated in HSC003 methylcellulose medium according to the manufacturer's recommendation (R&D Systems). Colonies of different hematopoietic lineages were scored 2 weeks after plating using an Inverted Infinity and Phase Contrast Microscope (Fisher Scientific). In addition, to examine the expression of lineage makers, methylcellulose medium was dissolved in PBS to dissociate the colonies into a single-cell suspension. Cells were stained with antibodies against CD14, CD15, and CD235a on ice for 30 min and then analyzed on a BD FACSAria II system.

Xenotransplantation assays. Bone marrow samples from additional patients with MDS or AML (unpaired) were processed and stained for surface markers for premalignant stem cells (pre-MDS/AML-SC) and malignant stem cells (MDS/AML-SC), as described above (Supplementary Fig. 1a). Thereafter, 30,000–100,000 sorted cells were washed with and resuspended in Hank's Balanced Salt Solution (HBSS) and transplanted into nonirradiated NOD,B6.SCID $Il2r\gamma^{-/-}Kit^{W41/W41}$ (NBSGW) immunocompromised mice (aged 6–8 weeks) via retro-orbital injection[43]. All experiments conducted on mice were approved by the Institutional Animal Care and Use Committee at Albert Einstein College of Medicine (protocol no. 2016-0103). Engraftment analysis of patient-derived cells was performed from 12 weeks after transplantation. Mouse bone marrow cells were incubated with ammonium chloride potassium buffer for 1 min on ice, and then stained for surface markers for mouse leukocytes, including CD45.1, and markers for human leukocytes, including CD45, CD19, and CD33. The stained cells were then analyzed on a BD FACSAria II system. While several studies have found some remaining lymphoid reconstitution of MDS/AML-SC in irradiated recipient mice in a subset of patients[44,45], many others found an exclusively myeloid output of MDS/AML-SC[8,15]. The partially lymphoid engraftment observed in our study could be due to the nonirradiated NBSGW xenograft model we utilized[43], as myeloid-biased engraftment of stem cells seems to be most pronounced in irradiation-conditioned transplantation assays[46,47].

Whole-genome amplification. WGA was performed with REPLI-g kit (Qiagen), which utilizes the proofreading enzyme Phi 29 polymerase to achieve high-fidelity amplification of genomic DNA[48,49]. For sorted samples with yield cell number larger than 1,000, cells were washed with PBS and then resuspended with 5 µl of sterile PBS. The REPLI-g mini kit was used according to the manufacturer's protocol. For sorted samples with fewer than 1,000 cells or for single-cell analysis, cells were sorted into 5 µl PBS, and REPLI-g single-cell kit (Qiagen) was used for WGA according to the manufacturer's protocol. For DNA samples, we used 1–10 ng DNA as input and REPLI-g mini kit (Qiagen) was used for WGA. All the products of WGA were purified with Agencourt AMPure XP beads (Beckman Coulter) to remove residual dNTP, primers, and random products <100 bp.

Targeted sequencing with HiSeq 2500. From the same patient, seven cell populations (pre-MDS-SC, MDS-SC, MDS blasts; pre-AML-SC, AML-SC, AML blasts; nonhematopoietic germline control) were subjected to targeted sequencing of a 504-gene customized panel containing all the genes in the FoundationOne Heme panel[21] as well as other genes of interest involved in the development of MDS and AML (full list of genes is provided in Supplementary Table 2). For each of the target genes, we included all the exons, 5′ and 3′ UTRs, as well as the 1,000-bp up- and downstream regions of the gene. For targeted sequencing, 500 ng of DNA was used as input for sequencing with an Illumina HiSeq 2500 system (Illumina). In brief, DNA was fragmented by a Covaris ultrasonicator (Covaris) with a target size of ~200 bp, followed by end repair and A-tailing with KAPA LTP library preparation kit for Illumina platforms (Kapa Biosystems) according to the manufacturer's instructions. Thereafter, we linked the DNA products with Illumina TrueSeq sequencing adapters and performed size selection with dual-SPRI beads (Beckman Coulter). Next, we performed eight cycles of pre-capture ligation mediated (LM)-PCR with the adapter-ligated DNAs according to the user's guide for NimbleGen SeqCap EZ Library (Roche NimbleGen). Afterwards, LM-PCR products of different cell populations from the same patient were incubated together for 72 h with NimbleGen SeqCap EZ probes (Roche NimbleGen). Hybridization products were then incubated with capture beads at 47 °C for 45 min, followed by washing and elution with PCR-grade water according to the manufacturer's protocol. Captured DNAs were then amplified with eight cycles of post-capture LM-PCR according to the user's guide for NimbleGen SeqCap EZ Library (Roche NimbleGen). Finally, DNA products were purified with Agencourt AMPure XP beads (Beckman Coulter) and then subjected to massively parallel sequencing (100 bp paired-end) on the HiSeq 2500 platform according to the manufacturer's instructions.

Analysis of sequencing data. We assessed the quality of the raw sequencing data from HiSeq with FastQC v0.11.4 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Reads contaminated with sequencing adapter and those of low quality were removed by Trim Galore 0.4.1 using the default parameters (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). Thereafter, we performed genome alignment (hg19) using Bowtie2 v2.2.9 (ref. [50]). Alignment results were processed as described in GATK best practice for detection of somatic mutation recommended by the Broad institute[51]. Briefly, duplicated reads were marked with a Picard toolkit (http://broadinstitute.github.io/picard/). Thereafter, indel realignment and base recalibration were performed for each of the individual samples with GATK v3.7 (ref. [51]). Moreover, we performed a second run of indel realignment with merged samples from the same human patient to remove false-positive mutations caused by alignment artifacts. After preprocessing of the reads, sequencing coverage of each sample was calculated with the *DepthOfCoverage* module of GATK. For detection of somatic mutations, we used Mutect2 of GATK v3.7 comparing each of the cell populations to the matched germline control with the default parameters[52]. We then merged all the Mutect2 results passing the filter from the same human patient to generate a combined set of mutations for each of the patients. FreeBayes v0.9.20 was used to perform joint variant calling with all samples from the same human patient[53], using the parameters of *-m 1 -q 3 -F 0.05 -C 2 -U 3 –read-indel-limit 2 –min-coverage 20*. We also excluded the variants from FreeBayes results with quality score <10. Thereafter, high-confidence mutations consistently detected by both Mutect2 and FreeBayes were used for downstream analysis. In addition, to address potential false-negatives due to tumor cell contamination of germline controls, we also included somatic mutations reported in MDS or AML by more than two groups in the COSMIC database (http://cancer.sanger.ac.uk/cosmic). Thereafter, we excluded the mutations that were: (i) covered less than 20× in germline control or test cells; (ii) supported by <3 reads or 5% of the reads in test samples; (iii) reported in dbSNP database (SNPs v147), 1000 genome phase 3, or ExAC database 1.0 with population frequency >0.5%. To further remove mutation artifacts caused by sequencing context of low complexity, we excluded mutations that were: (i) located within 10 bp of an indel; (ii) within 20 bp of another single-nucleotide variant (SNV); (iii) less than 5 bp to microsatellite or simple repeats of the UCSC database (https://genome.ucsc.edu); (iv) less than 5 bp to homopolymer (>5 bp). Thereafter, mutations were annotated using the hg19 database by SnpEff v4.1k (ref. [54]).

For analysis of mutation signatures, we combined the somatic mutations in each cell population from the five patients sequenced and examined the pattern of mutation signatures with deconstructSigs 1.8 with the signatures defined previously[55]. The weight of each signature was normalized by the number of times each trinucleotide context was observed in the targeted regions.

Clonal analysis. VAF for each mutation was calculated by the number of reads supporting the variant divided by total reads, using the FreeBayes output. Moreover, sample purity and local copy number variation (CNV) were estimated by the FACETS v0.5.6 package of R v3.2.3 (ref. [56]), which utilizes the read counts of both heterozygous and homozygous single-nucleotide polymorphism (SNP)

loci. In brief, for each of the samples, we first extracted the read counts of reference and alternative alleles of each SNP reported in dbSNP (Common SNPs v147) or 1000 genome SNP phase 3 database with population frequency larger than 5%. Thereafter, the read count information of the SNP loci covered by at least 20× in the targeted sequencing of each sample was subjected to FACETS as input to estimate the purity and CNV using the default parameters. Thereafter, the CCF of each mutation was estimated using the VAF, purity, and local CNV of the mutation as described before[22]. Mutations were defined as 'clonal' if the 95% confidence interval of CCF overlapped with 0.95, otherwise being defined as 'subclonal.' To investigate the clonal architecture, both VAFs and CCFs of mutations covering >30× were subjected to SciClone v1.1.0 allowing a maximum cluster number of 10 (ref. [28]). When comparing the clonal architecture of different cell populations of the same patient, we first generated a combined list of mutations that covered at least 20× in all samples, then subjected the VAFs of mutations in different populations to SciClone analysis. We excluded the mutations in the cluster if the estimated possibility of the mutation being clustered in the subclone was lower than 0.95. In addition, to examine the clonal relationship between different cell populations in the same samples, we performed phylogenetic reconstruction by LICHeE v1.0 using VAFs of the mutations and the prevalence of each subclone in the samples estimated by SciClone, with the standard parameters (*-maxVAFAbsent* 0.005 *-minVAFPresent* 0.005 *-n* 0) recommended by the LICHeE instructions[57]. Thereafter, the results of phylogenetic relationships determined by LICHeE were visualized by the TimeScape v1.0.0 package[58].

Single-cell targeted sequencing. After staining of surface markers, single cells were directly deposited, using a MoFlo Astrios EQ system (Beckman Coulter), into a 96-well PCR plate containing 5 µl of sterile PBS per well. Thereafter, WGA was performed using a Repli-g single-cell kit (Qiagen) according to the manufacturer's protocol. WGA products were purified with Agencourt AMPure XP beads (Beckman Coulter). For targeted sequencing, we designed primers for each mutation target using Primer 3, with product sizes <200 bp (Supplementary Table 5). Target-specific primers were linked with the Fluidigm forward (5′-ACACTGACGACATGGTTCTACA-3′) and reverse (5′-TACGGTAGCAGAGACTTGGTCT-3′) common sequence (CS) tag for downstream barcoding. To preamplify the DNA of target regions, we first performed specific target amplification (STA) of WGA products using FastStart Taq DNA Polymerase (Roche). In brief, all CS-tagged primers for the same sample were pooled and diluted to make a final concentration of 1 µM for each primer. The amplification mix for each sample was prepared as follows: 0.5 µl of 10× reaction buffer with $MgCl_2$, 0.5 µl $MgCl_2$, DMSO, 10 mM nucleotide mix, 0.2 µl FastStart polymerase, 1 µl 1 µM primer pool, and 10 ng DNA. Next, PCR amplification was performed as follows: 95 °C for 10 min; 2 cycles of 95 °C for 15 s and 60 °C for 4 min; 10 cycles of 95 °C for 15 s and 72 °C for 4 min. As a negative control, we included a no-template control (NTC) in the STA experiment. Thereafter, 10 µl of each STA product diluted to 100 ng μl^{-1} was transferred to half of a new 96-well plate (47 single-cell samples plus one NTC per plate), and treated with ExoSAP-IT (Affymetrix) for purification. For primer preparation, each primer pair was diluted to 1 µM in the 96-well plate with Fluidigm Access Array loading reagent (Fluidigm). Thereafter, plates of STA products and primer pairs were loaded onto 48.48 integrated fluidic circuits (IFC) in a Biomark HD system (Fluidigm). Each of the STA products was mixed with each primer pair, and PCR amplification was performed in the IFC array according to the manufacturer's protocol. Thereafter, PCR products of the same sample were pooled together, and sample barcoding PCR was performed with primers containing the barcode sequence (Fluidigm) and Illumina sequencing adapter (Illumina). We assessed the quality of the barcoded samples with a 2100 Bioanalyzer (Agilent), then all samples were pooled at equal ratios and subjected to sequencing with the MiSeq (150 bp paired-end) system according to the manufacturer's protocol (Illumina).

For analysis of the MiSeq data, we trimmed reads with CS tag and reads contaminated with the sequencing adapter, and we also removed reads of low quality by Trim Galore using the default parameters. Thereafter, we performed genome alignment to hg10 with BWA-MEM v0.7.15 (ref. [59]), and then variant calling with FreeBayes. We also manually confirmed each target mutation with the Integrative Genomics Viewer, with mutations with >20% supporting reads (covering at least 5×) being considered positive.

T cell receptor sequencing. To assess the diversity of the T cell receptor (TCR) repertoire, we extracted total RNA from T cells isolated from the patient samples, as well as cord blood samples as healthy controls, using an RNeasy Micro Kit (Qiagen) according to the manufacturer's protocol. We used 50 ng of total RNA as input for first-strand cDNA synthesis with the reagents supplied in the SMARTer Human TCR a/b Profiling Kit (Takara Bio USA) according to the manufacturer's protocol. Thereafter, a first round of PCR (PCR 1) was performed with SMART Primer 1 to link the Illumina Read 2 sequence to the cDNA, and TCRα and TCRβ primers, to specifically amplify the variable regions and constant regions of TCRα and TCRβ cDNA. The PCR 1 reaction was performed for 21 cycles with a preheated thermal cycler (C1000, Bio-Rad) according to the manufacturer's protocol. Next, 1 µl of PCR1 product was subjected to second-round PCR (PCR 2), which was performed with TCRα and TCRβ Human Primer 2 Reverse HT Index primers (D501) to link the Illumina Read 1 sequence and P5-i5 index sequences. In addition, for different samples, we used different TCR Primer 2 Forward HT Index primers for the linkage of Illumina P7-i7 index sequences. The PCR 2 reaction was performed for 20 cycles with a preheated thermal cycler according to the manufacturer's protocol. Lastly, the products of PCR 2 were purified using Agencourt AMPure XP beads (Beckman Coulter) with a double-size selection approach according to the manufacturer's instructions. The quality and quantity of the purified products (sequencing-ready libraries) were assessed with a 2100 Bioanalyzer (Agilent) and Qubit 2.0 Fluorometer, respectively. Sequencing was performed on an Illumina MiSeq sequencer with paired-end, 300-bp reads. For analyses of the sequencing data, the first 30 bp of read 2, which includes the SMART primer sequence, was trimmed with Trim Galore. The trimmed data was then analyzed with LymAnalyzer 1.2.2 separately for TCRα and TCRβ genes[60]. We then calculated the frequency of each Vα or Vβ gene segment relative to the total sequences mapped to the Vα or Vβ genes.

Statistical analysis. Data are presented as mean ± s.d. if not otherwise specified. Student's *t*-test was performed with GraphPad Prism 7.0, as indicated. Pearson correlation coefficient *r* and statistical significance *P* values were calculated with the built-in *cor.test* function of R, and data were visualized with the ggplot2 package of R.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The high-throughput DNA sequencing data have been deposited in the database of Genotypes and Phenotypes (dbGaP) .

References

41. Shastri, A., Will, B., Steidl, U. & Verma, A. Stem and progenitor cell alterations in myelodysplastic syndromes. *Blood* **129**, 1586–1594 (2017).
42. Thomas, D. & Majeti, R. Biology and relevance of human acute myeloid leukemia stem cells. *Blood* **129**, 1577–1585 (2017).
43. McIntosh, B. E. et al. Nonirradiated NOD,B6.SCID Il2rγ$^{-/-}$ Kit(W41/W41) (NBSGW) mice support multilineage engraftment of human hematopoietic cells. *Stem Cell Reports* **4**, 171–180 (2015).
44. Woll, P. S. et al. Myelodysplastic syndromes are propagated by rare and distinct human cancer stem cells in vivo. *Cancer Cell.* **25**, 794–808 (2014).
45. Terwijn, M. et al. Leukemic stem cell frequency: a strong biomarker for clinical outcome in acute myeloid leukemia. *PLoS ONE* **9**, e107587 (2014).
46. Wang, C. et al. Non-lethal ionizing radiation promotes aging-like phenotypic changes of human hematopoietic stem and progenitor cells in humanized mice. *PLoS ONE* **10**, e0132041 (2015).
47. Lu, R., Czechowicz, A., Seita, J., Jiang, D. & Weissman, I. L. Clonal level lineage commitment pathways of hematopoietic stem cells in vivo. Preprint at https://doi.org/10.1101/262774 (2018).
48. Hosono, S. et al. Unbiased whole-genome amplification directly from clinical samples. *Genome Res.* **13**, 954–964 (2003).
49. de Bourcy, C. F. et al. A quantitative comparison of single-cell whole genome amplification methods. *PLoS ONE* **9**, e105585 (2014).
50. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
51. McKenna, A. et al. The genome analysis toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
52. Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).
53. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. Preprint at https://arxiv.org/abs/1207.3907 (2012).
54. Cingolani, P. et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w(1118); iso-2; iso-3. *Fly* **6**, 80–92 (2012).
55. Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B. S. & Swanton, C. DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome. Biol.* **17**, 31 (2016).
56. Shen, R. & Seshan, V. E. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. *Nucleic Acids Res.* **44**, e131 (2016).
57. Popic, V. et al. Fast and scalable inference of multi-sample cancer lineages. *Genome. Biol.* **16**, 91 (2015).
58. Smith, M.A. et al. E-scape: interactive visualization of single-cell phylogenetics and cancer evolution. *Nat. Methods* **14**, 549–550 (2017).
59. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://arxiv.org/abs/1303.3997 (2013).
60. Yu, Y., Ceredig, R. & Seoighe, C. LymAnalyzer: a tool for comprehensive analysis of next generation sequencing data of T cell receptors and immunoglobulins. *Nucleic Acids Res.* **44**, e31 (2016).

nature research

Corresponding author(s): Ulrich Steidl

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒ Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	Data of flow cytometry was collected with FACSDiva 8 or Summit v62 for the FACS experiments on BD FACSAria II system or MoFlo Astrios EQ system, respectively. And FACS plots were generated by FlowJo v10
Data analysis	We used the following softwares/versions: FastQC v0.11.4; Trim Galore v0.4.1; Bowtie v2.2.9; Picard v1.138; GATK v3.7; Mutect2 v3.7; FreeBayes v0.9.20; SnpEff v4.1k; DeconstructSigs v1.8; FACETS v0.5.6; R v3.2.3; SciClone v1.1.0; LICHeE v1.0; TimeScape v1.0.0; BWA-MEM v0.7.15; Integrative Genomics Viewer v2.3.80; ggplot2 v2.0.0; GraphPad Prism 7.0; FlowJo v10. Further details can be found in the methods section.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The high-throughput DNA sequencing data will be deposited in the database of Genotypes and Phenotypes (dbGaP) , accession code will be available before publication

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample sizes were not based on formal power calculations to detect pre-specified effect sizes for this proof-of-concept study. Sample size was based on availability of 7 individual patients with longitudinal samples both at the MDS stage and later progression to AML (progression time between 9 months and 3 years). This sample size is adequate for the here presented type of proof-of-concept study and is in line with sample sizes that have been used in other prominent studies in this field.
Data exclusions	No data was excluded. All data that passed quality control parameters using standard sequencing and analysis algorithms (described in detail in the methods) were included in the analysis.
Replication	Due to limited amount of patient samples, we were unable to perform technical replicates for each patient in the targeted capture sequencing. We analyzed several primary patients' samples in this study and relevant findings were made in multiple samples/patients analyzed. Mutations were only scored for regions with a pre-defined minimal read support and variant read number following standard practice in the field (see methods for details). Results from targeted sequencing were corroborated by independent processing and sequencing of multiple single cells of the same patients. Each single cell was sequenced once. We performed biological replicates (different patient samples) in the xenotransplantation and colony forming assays. And in xenotransplantation, sorted cells from patient samples were transplanted into multiple recipient mice unless there were not enough cells. The sample size was described in the figure legend. For the T cell receptor profiling assay, both the healthy controls and patient group had at least two biological replicates, and each sample was assessed once.
Randomization	This is not relevant for the present study (no group allocation was performed).
Blinding	This is not relevant for the present study (no group allocation was performed).

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	*Briefly describe the study type including whether data are quantitative, qualitative, or mixed-methods (e.g. qualitative cross-sectional, quantitative experimental, mixed-methods case study).*
Research sample	*State the research sample (e.g. Harvard university undergraduates, villagers in rural India) and provide relevant demographic information (e.g. age, sex) and indicate whether the sample is representative. Provide a rationale for the study sample chosen. For studies involving existing datasets, please describe the dataset and source.*
Sampling strategy	*Describe the sampling procedure (e.g. random, snowball, stratified, convenience). Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient. For qualitative data, please indicate whether data saturation was considered, and what criteria were used to decide that no further sampling was needed.*
Data collection	*Provide details about the data collection procedure, including the instruments or devices used to record the data (e.g. pen and paper,*

Data collection	*computer, eye tracker, video or audio equipment) whether anyone was present besides the participant(s) and the researcher, and whether the researcher was blind to experimental condition and/or the study hypothesis during data collection.*
Timing	*Indicate the start and stop dates of data collection. If there is a gap between collection periods, state the dates for each sample cohort.*
Data exclusions	*If no data were excluded from the analyses, state so OR if data were excluded, provide the exact number of exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.*
Non-participation	*State how many participants dropped out/declined participation and the reason(s) given OR provide response rate OR state that no participants dropped out/declined participation.*
Randomization	*If participants were not allocated into experimental groups, state so OR describe how participants were allocated to groups, and if allocation was not random, describe how covariates were controlled.*

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	*Briefly describe the study. For quantitative data include treatment factors and interactions, design structure (e.g. factorial, nested, hierarchical), nature and number of experimental units and replicates.*
Research sample	*Describe the research sample (e.g. a group of tagged Passer domesticus, all Stenocereus thurberi within Organ Pipe Cactus National Monument), and provide a rationale for the sample choice. When relevant, describe the organism taxa, source, sex, age range and any manipulations. State what population the sample is meant to represent when applicable. For studies involving existing datasets, describe the data and its source.*
Sampling strategy	*Note the sampling procedure. Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient.*
Data collection	*Describe the data collection procedure, including who recorded the data and how.*
Timing and spatial scale	*Indicate the start and stop dates of data collection, noting the frequency and periodicity of sampling and providing a rationale for these choices. If there is a gap between collection periods, state the dates for each sample cohort. Specify the spatial scale from which the data are taken*
Data exclusions	*If no data were excluded from the analyses, state so OR if data were excluded, describe the exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.*
Reproducibility	*Describe the measures taken to verify the reproducibility of experimental findings. For each experiment, note whether any attempts to repeat the experiment failed OR state that all attempts to repeat the experiment were successful.*
Randomization	*Describe how samples/organisms/participants were allocated into groups. If allocation was not random, describe how covariates were controlled. If this is not relevant to your study, explain why.*
Blinding	*Describe the extent of blinding used during data acquisition and analysis. If blinding was not possible, describe why OR explain why blinding was not relevant to your study.*

Did the study involve field work? ☐ Yes ☐ No

Field work, collection and transport

Field conditions	*Describe the study conditions for field work, providing relevant parameters (e.g. temperature, rainfall).*
Location	*State the location of the sampling or experiment, providing relevant parameters (e.g. latitude and longitude, elevation, water depth).*
Access and import/export	*Describe the efforts you have made to access habitats and to collect and import/export your samples in a responsible manner and in compliance with local, national and international laws, noting any permits that were obtained (give the name of the issuing authority, the date of issue, and any identifying information).*
Disturbance	*Describe any disturbance caused by the study and how it was minimized.*

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☐	☒ Unique biological materials
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☐	☒ Animals and other organisms
☐	☒ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☐	☒ Flow cytometry
☒	☐ MRI-based neuroimaging

Unique biological materials

Policy information about availability of materials

Obtaining unique materials

Unique materials used in this study include longitudinal primary samples from patients at the MDS and AML stages, respectively. This is a highly precious resource and some limited additional aliquots/vials of these primary samples are available for some of the samples.

Antibodies

Antibodies used

We used the following antibodies (antibody,conjugate,clone,company, microliter of antibody used per 100ul reaction):
CD2, PE-Cy5, RPA-2.10, eBioscience, 1.2
CD3, PE-Cy5, UCHT1, eBioscience, 1.2
CD4, Tri-color, S3.5, Invitrogen, 1.2
CD7, Tri-color, 6B7, Invitrogen, 1.2
CD8, Tri-color, 3B5, Invitrogen, 1.2
CD10, PE-Cy5, eBioCB-CALLA, eBioscience, 1.2
CD11b, Tri-color, M1/70, Invitrogen, 1.2
CD14, Tri-color, TueK4, Invitrogen, 1.2
CD19, PE-Cy5, HIB19, eBioscience, 1.2
CD20, PE-Cy5, 2H7, eBioscience, 1.2
CD235a , PE-Cy5, GA-R2, BD Pharmingen, 1.2
CD56 , Tri-color, MEM-188, Invitrogen, 1.2
CD33, APC, WM-53, Molecular Probes, 2.4
CD34 , Pacific Blue, 581, Biolegend, 2.4
CD38, PE-Cy7, HIT2, eBioscience, 2.4
CD45RA, FITC, MEM-56, Invitrogen, 2.4
CD123, Alexa Fluor 700, 32703, R&D, 2.4
IL1RAP, PE, 89412, R&D, 2.4
CD45 , PE, HI30, BD Pharmingen, 2.4
CD4, Pacific Orange, S3.5, Invitrogen, 2.4
CD3, FITC, UCHT1, eBioscience, 2.4
hCD45, PE, HI30, BD Pharmingen, 0.5
mCD45.1, PE-Cy7, A20, Invitrogen, 1
CD19, PE-Cy5, HIB19, eBioscience, 0.5
CD33, APC, WM-53, Molecular Probes, 0.5
CD14, Tri-color, TueK4, Invitrogen, 1
CD15, Brilliant Violet 510, W6D3, Biolegend, 1
CD235a, FITC, GA-R2, BD Pharmingen, 1

Validation

Standard FACS antibodies obtained from widely used commercial providers were used in this study. Catalog numbers and clones are given in the Supplementary Table 4. All antibodies were validated through positive and negative control stainings, as well as isotype control antibodies.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals

For transplantation assay, we used NOD,B6.SCID Il2rγ−/−KitW41/W41 (NBSGW) immunocompromised female mice at the age of 6-8 weeks

Wild animals

No wild animals were used in this study.

Field-collected samples

No field-collected samples were used in this study

Human research participants

Policy information about studies involving human research participants

Population characteristics	Information on the patients whose longitudinal samples at the MDS and AML stages, respectively, were included in this study is given in Supplementary Table 1.
Recruitment	Patients with MDS and matched secondary AML were obtained after written informed consent, from Montefiore Medical Center / Albert Einstein Cancer Center following the protocol (IRB# 11-02-060E). Please note that these samples were obtained from a tissue repository, they were deidentified, and not specifically collected for this study. Therefore, technically this is not "human subject research" as defined by NIH, but regulated under so-called exemption #4.

ChIP-seq

Data deposition

☒ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☒ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links *May remain private before publication.*	*For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.*
Files in database submission	*Provide a list of all files available in the database submission.*
Genome browser session (e.g. UCSC)	*Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.*

Methodology

Replicates	*Describe the experimental replicates, specifying number, type and replicate agreement.*
Sequencing depth	*Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.*
Antibodies	*Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.*
Peak calling parameters	*Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.*
Data quality	*Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.*
Software	*Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.*

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	Sample origin and preparation for FACS are described in detail in a separate section in the methods section of the manuscript (paragraph: "Multiparameter high-speed FACS of stem and blast cells from patient samples")
Instrument	Cell sorting and analysis were performed on a Beckman MoFlo Astrios EQ system and a BD Bioscience FACSAria II instrument.
Software	The flow cytometry data was collected with FACSDiva v8.0 of BD FACSAria II, or Summit v62 of MoFlo Astrios EQ during flow cytometry experiments. Thereafter, the data was analyzed with FlowJo v10.
Cell population abundance	Purity of sorted fractions was determined by re-analysis and ranged between 98%-99.9%

Gating strategy

The gating strategy used followed established markers and schemes for the identification of (pre)LSC and bulk populations and is described and shown in detail in the methods section and Supplementary Figure 1. Negativity for any marker was defined as the threshold defined by staining with the respective isotype control antibodies.

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type

Indicate task or resting state; event-related or block design.

Design specifications

Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.

Behavioral performance measures

State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects).

Acquisition

Imaging type(s)

Specify: functional, structural, diffusion, perfusion.

Field strength

Specify in Tesla

Sequence & imaging parameters

Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle.

Area of acquisition

State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined.

Diffusion MRI ☐ Used ☐ Not used

Preprocessing

Preprocessing software

Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).

Normalization

If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.

Normalization template

Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.

Noise and artifact removal

Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).

Volume censoring

Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.

Statistical modeling & inference

Model type and settings

Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).

Effect(s) tested

Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.

Specify type of analysis: ☐ Whole brain ☐ ROI-based ☐ Both

Statistic type for inference (See Eklund et al. 2016)

Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.

Correction

Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).

Models & analysis

n/a	Involved in the study
☐	☐ Functional and/or effective connectivity
☐	☐ Graph analysis
☐	☐ Multivariate modeling or predictive analysis

Functional and/or effective connectivity

Report the measures of dependence used and the model details (e.g. Pearson correlation, partial

Functional and/or effective connectivity	*correlation, mutual information).*
Graph analysis	*Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).*
Multivariate modeling and predictive analysis	*Specify independent variables, features extraction and dimension reduction, model, training and evaluation metrics.*