Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

Cell. Author manuscript; available in PMC 2025 February 01.

Published in final edited form as:

Cell. 2024 February 01; 187(3): 642–658.e19. doi:10.1016/j.cell.2023.12.016.

Inherited blood cancer predisposition through altered transcription elongation

Jiawei Zhao[1,2,10,*,†], **Liam D. Cato**[1,2,*], **Uma P. Arora**[1,2], **Erik L. Bao**[1,2], **Samuel C. Bryant**[2], **Nicholas Williams**[3], **Yuemeng Jia**[4,5], **Seth R. Goldman**[6], **Jyoti Nangalia**[3], **Michael A. Erb**[7], **Seychelle M. Vos**[8,9], **Scott A. Armstrong**[1,2], **Vijay G. Sankaran**[1,2,4,11,†]

[1]Division of Hematology/Oncology, Boston Children's Hospital and Department of Pediatric Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA, USA.

[2]Broad Institute of MIT and Harvard, Cambridge, MA, USA.

[3]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, UK and MRC-Wellcome Cambridge Stem Cell Institute, University of Cambridge, Cambridge, UK.

[4]Harvard Stem Cell Institute, Cambridge, MA, USA

[5]Stem Cell Program, Boston Children's Hospital, Boston, MA, USA; Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, MA, USA

[6]Nascent Transcriptomics Core, Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA, USA.

[7]Department of Chemistry, The Scripps Research Institute, La Jolla, CA, USA.

[8]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA.

[9]Howard Hughes Medical Institute, Massachusetts Institute of Technology, Cambridge, MA, USA

[10]Center for Cancer Immunology, Faculty of Pharmaceutical Sciences, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences (CAS), Shenzhen, China.

[11]Lead Contact

[†] Correspondence: jw.zhao3@siat.ac.cn, sankaran@broadinstitute.org.

[*]These authors contributed equally to this work.

Author Contributions

J.Z., L.D.C., V.G.S. conceptualized the study. J.Z., L.D.C., V.G.S. devised methodology. J.Z., L.D.C., U.P.A., E.L.B., S.C.B., N.W., Y.J., S.R.G., J.N., M.A.E., S.M.V., S.A.A., V.G.S. performed studies. J.Z., L.D.C., U.P.A., Y.J., S.M.V., V.G.S. provided data visualization media. V.G.S. acquired funding for this work and provided overall project oversight. J.Z., L.D.C., V.G.S. wrote the original manuscript, as well as edited the manuscript with input from all authors.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Declaration of Interests

M.A.E. serves as an advisor and has equity in Nexo Therapeutics. M.A.E. is an inventor on patent applications covering SR-0813. S.A.A. serves as an advisor and/or has equity in Neomorph Inc., Imago Biosciences, Cyteir Therapeutics, C4 Therapeutics, Nimbus Therapeutics, and Accent Therapeutics. S.A.A. receives research support from Janssen and Syndax. V.G.S. serves as an advisor to and/or has equity in Branch Biosciences, Ensoma, and Cellarity, all unrelated to the present work. There are no other relevant declarations to report.

Inclusion and Diversity

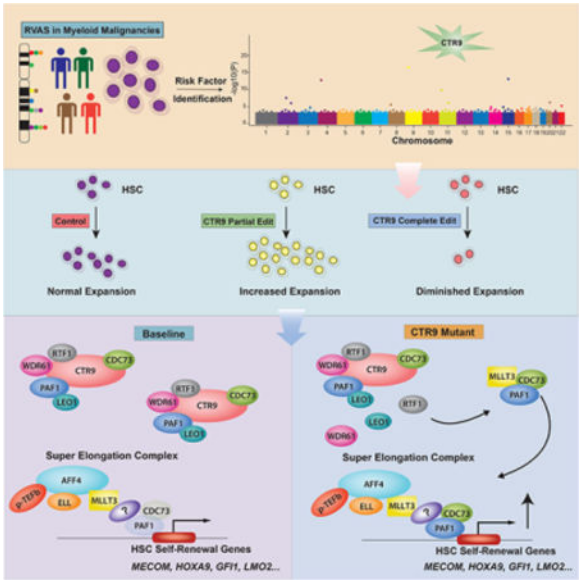
We support inclusive, diverse, and equitable conduct of research.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

SUMMARY

Despite advances in defining diverse somatic mutations that cause the myeloid malignancies, a significant heritable component for these cancers remains largely unexplained. Here, we perform rare variant association studies in a large population cohort to identify inherited predisposition genes for these blood cancers. *CTR9*, which encodes a key component of the PAF1 transcription elongation complex, is among the significant genes identified. The risk variants found in cases cause loss-of-function and result in a ~10-fold increased odds of acquiring a myeloid malignancy. Partial CTR9 loss-of-function expands human hematopoietic stem cells (HSCs) by increased super elongation complex-mediated transcriptional activity, which thereby increases expression of key regulators of HSC self-renewal. By following up on insights from a human genetic study examining inherited predisposition to the myeloid malignancies, we define a previously unknown antagonistic interaction between the PAF1 and super elongation complexes. These insights could enable targeted approaches for blood cancer prevention.

Graphical Abstract



In Brief

Rare inherited *CTR9* loss-of-function variants predispose to myeloid malignancies by altering the balance between the PAF1 and super elongation complexes. Specific subunits of the PAF1 complex then act in concert with the super elongation complex to promote transcription elongation of genes that can drive hematopoietic stem cell self-renewal.

INTRODUCTION

The myeloid malignancies are a group of blood cancers that arise when somatic driver mutations are acquired in the hematopoietic stem cell (HSC) compartment[1,2], and includes the myeloproliferative neoplasms (MPNs), myelodysplastic syndromes (MDS), and acute myeloid leukemias (AMLs). While these cancers are heterogeneous in terms of phenotype

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

and outcome, genomic studies over the past decades have uncovered the drivers underlying most of these diseases and have revealed common molecular pathways impacted, including factors involved in signaling, transcriptional regulation, DNA damage repair, and splicing[3]. More recent studies of these diseases at single-cell resolution have uncovered additional complexity in cell states in these diseases that could underlie some of the variable phenotypes observed, even when similar driver mutations are found in cases[2,4–6]. A great deal of knowledge has emerged by cataloging and studying the acquired mutations that drive these diseases.

However, it is clear that there is a significant heritable component to these cancers that remains incompletely defined and poorly understood mechanistically[7]. Common variant association studies (CVAS) have provided valuable insights[8–13], particularly for the MPNs, but only identify a fraction of the heritable risk with effect estimates that are typically quite small for significant loci (most often with odds ratios of <1.5). Studies of families harboring multiple individuals with myeloid malignancies have revealed highly-penetrant mutations that can strongly predispose individuals to acquire these blood cancers, but that are rare in the population[14–17]. With the increasing availability of large-scale sequencing studies, we reasoned that rare variant association studies (RVAS) could serve as a valuable complementary approach to CVAS and family-based studies to identify additional inherited risk factors for the myeloid malignancies. In particular, this is likely to be the case for inherited variants that occur at somewhat higher frequency, but lower penetrance, than occurs with rare familial syndromes and which might have larger effects on risk than the common variants identified in CVAS. Indeed, recent RVAS in large population cohorts have proven valuable to provide unanticipated biological insights into the risk for different complex diseases[18,19]. In addition to the opportunity to identify new risk variants underlying these disorders, since predisposition to these cancers likely impacts the function or regulation of HSCs, follow up of these genetic studies should provide opportunities to gain new mechanistic insights into hematopoiesis.

RESULTS

CTR9 loss-of-function variants predispose to the development of myeloid malignancies

Given that no prior RVASs for myeloid malignancy predisposition have been reported to date, we conducted such a study in 166,953 participants for whom exome sequencing data was available from the UK Biobank cohort[20]. Following quality control of this data (STAR Methods), we identified 691,460 rare variants under a frequency of 0.1% and of predicted deleteriousness that were included in a gene-based collapsing association analysis (Figure 1A). We combined the myeloid malignancies, including AML, MDS, and MPNs, into one phenotype ($N_{myeloid}$= 793 (N_{MPN}= 578, N_{AML}= 175, N_{MDS}= 135), $N_{controls}$= 166,160), given that many risk factors are shared across these blood cancers and to increase overall power. Significant burden was seen in a number of genes (*JAK2*, *TET2*, *IDH2*, *DNMT3A*, *ASXL1*) known to harbor somatic mutations in clonal hematopoiesis of indeterminate potential (CHIP)[21,22], and served as positive controls for the assessment of deleterious alleles, as the presence of CHIP is known to predispose to the myeloid malignancies. Of note, we also identified significant burden in a number of genes including *CTR9*, *CHEK2*,

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

AOX1, *MCCC1*, and *HBB*, which showed variant allele fractions in the germline range (Figure S1A) and are typically not seen in CHIP nor as somatic mutations from prior studies of cancer genomes[23].

While some of the genes harboring germline variants with a significant burden in myeloid malignancy cases had previously been implicated in blood cancer predisposition, such as *CHEK2*[24], several were unknown. Among this group, *CTR9* was of particular interest, as germline variants in the gene had not been previously implicated in conferring risk for developing myeloid malignancies, but heterozygous loss-of-function mutations in *CTR9* have been associated with familial Wilms' tumor[25], which suggested a broader role for this protein as a tumor suppressor. *CTR9* encodes a key component of the PAF1 complex (PAF1c), which has multiple functions during transcription, including facilitating transcription elongation after promoter-proximal pausing by RNA polymerase II[26,27]. The mammalian PAF1c is composed of five to six core subunits: CTR9, CDC73, PAF1, LEO1, WDR61, and sometimes RTF1[28]. Deleterious mutations in the *CDC73* gene increase risk of parathyroid carcinoma[29] and cause the hyperparathyroidism-jaw tumor syndrome[30]. Interestingly, PheWAS for deleterious *CTR9* variants across 1,027 phenotypes revealed cancer and disease of salivary glands, bone, and hyperparathyroidism as among the most significant, suggesting possible interference with the PAF1c in a manner producing similar phenotypic outcomes to CDC73 heterozygous loss-of-function (Figure S1B).

The majority of the signal from the gene burden testing for *CTR9* was driven by the MPNs, although it should be noted that these were among the most common myeloid malignancies in the UK Biobank cohort (Figure 1B). As expected, rare variant collapsing tests showed much larger odds ratios (OR) than variants identified from CVAS of myeloid malignancies (Figure 1C). Deleterious variants in *CTR9* showed a large OR of 9.6 for myeloid malignancies (95%CI = 4.86–19.04, SKAT-O p-value = 5.47×10^{-7}) (Table S1), but were not significantly associated with risk of detectable CHIP (p=0.17). However, there is low sensitivity to identify CHIP due to *JAK2* mutations and specific other variants in this cohort, as a result of low sequence coverage (STAR Methods). There were eight cases that carried six deleterious mutations in CTR9, found across multiple exons of the gene (Figure 1D). Following our initial discovery association in an earlier release of UK Biobank exomes, we had an opportunity to confirm our associations with the recent release of the entire UK Biobank cohort ($N_{myeloid}$ = 2211 (N_{MPN}= 1608, N_{AML}= 476, N_{MDS}= 430), $N_{controls}$ = 465,273) and identified significant burden in similar sets of genes, including genes known to harbor CHIP associated variants and *CTR9* (p<0.001).

Structural analysis of PAF1c[31,32] revealed how the deleterious variants in CTR9 found in cases often occur at interfaces involving other key components necessary for transcription. For example, residue 498 is located at the binding interface between CDC73 and PAF1, while residues 698 and 701 are located at the binding interface with subunit RPB8 of RNA polymerase II (Figure 2A, B)[31,32]. Consistent with the likely disruptive role of these mutations, when expressed as a tagged protein along with wildtype CTR9, we observed consistently reduced interactions with other PAF1c components and RNA polymerase II (RPB1/POLR2A is the large subunit of this complex), demonstrating loss-of-function by all of these variants (Figure 2C, D). Crucially, we observed similar binding by wildtype CTR9

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

in all of these cases, demonstrating that no mutations were acting in a dominant negative manner, at least in terms of PAF1c assembly (Figure S2).

Partial loss of CTR9 augments HSC maintenance *in vitro*

A greater extent of hematopoietic stem cell (HSC) self-renewal has been shown to increase the risk for acquiring myeloid malignancies[8,33–35]. Given that the risk variants in *CTR9* cause loss-of-function, we hypothesized that these variants might impact HSC or progenitor self-renewal or function. To recreate loss-of-function alleles in an isogenic setting in primary human hematopoietic stem and progenitor cells (HSPCs), we used Cas9 ribonucleoprotein (RNP) delivery with four independent guide RNAs targeting *CTR9* with an editing efficiency of over 80% (Figure S3A). Interestingly, there was an increase in phenotypic long-term repopulating HSCs (LT-HSC)[6,36,37] in the days after editing, but these cells were then depleted (Figure 3A–D). We noted that while CTR9 protein levels were partially reduced (~50%) soon after RNP delivery, more profound depletion (>80%) occurred subsequently in tandem with the reduction in LT-HSCs (Figure S3A–B). To more faithfully model heterozygous loss-of-function, we titrated the amount of Cas9 RNP delivered into HSPCs so that *CTR9* would be edited in a predominantly heterozygous manner, as we have done in the case of other heterozygous mutations we have studied previously[6] (Figure S3C–E). Strikingly, this editing achieved expansion of both the phenotypic LT-HSC and more differentiated short-term HSC (ST-HSC) compartments without any change in the overall number of HSPCs (marked by $CD34^+CD45RA^-$) across multiple time points in culture (Figure 3E–H). As phenotypic markers to assess HSCs can be limited, we also performed single-cell RNA sequencing (scRNA-seq) on 3,335 *AAVS1* edited control and 4,154 *CTR9* edited $CD34^+CD90^+CD45RA^-$ cells and observed an expansion in cells harboring a previously defined HSC molecular signature[8] with *CTR9* editing (Figure 3I–J, Figure S3F–G). In addition to the observed phenotypic and molecular HSC expansion, we also observed increased colony plating capacity of HSPCs with larger colonies noted upon CTR9 editing (Figure 3K–L). Given that prior studies have shown how HOXA family members are both necessary for HSC self-renewal and demonstrate increased expression with altered transcription elongation[38–40] - the major function of PAF1c and other interacting complexes - we examined expression of HOXA family genes and saw consistent increases accompanying the observed HSC expansion (Figure S3H), whereas this expression of HOXA genes was diminished as CTR9 levels were reduced with more profound depletion (Figure S3I).

Partial loss of CTR9 increases HSC self-renewal *in vivo* and collaborates with myeloid malignancy driver mutations

To fully characterize the effect of partial loss of CTR9 on HSC function *in vivo*, we performed transplantation of $CD34^+$ HSPCs with editing targeting the *AAVS1* control locus or *CTR9* in a partial or complete manner into the NOD.Cg-Kit^{W41J}Tyr+ $Prkdc^{scid}Il2rg^{tm1Wjl}$/ThomJ (NBSGW) strain of immunodeficient and *Kit* mutant mice[37,41,42]. Consistent with the increase observed *in vitro* in phenotypic HSCs, partial *CTR9* edited HSPCs showed consistently higher engraftment of human $CD45^+$ cells in the peripheral blood over a period of 20 weeks of transplantation, whereas complete *CTR9* edited HSPCs showed significantly lower engraftment (Figure 4A). Similarly, there

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

is also a slightly higher engraftment observed in the bone marrow and spleen of NBSGW mice transplanted with *CTR9* partially edited HSPCs, which was diminished with complete editing of *CTR9* (Figure 4B, C). Indeed, we observed no significant change of the editing efficiency with partially edited HSPC xenotransplants, but a significant reduction of editing efficiency with completely edited HSPC xenotransplants in peripheral blood, bone marrow, and spleen (Figure S4). Notably, in NBSGW mice transplanted with completely edited *CTR9* HSPCs, the editing efficiency was reduced down to ~30% overall after 20 weeks. The strongest negative selection appeared to occur within the first 8 weeks, which resulted in a population with an editing efficiency of ~50% overall.

To comprehensively assess the role of CTR9 in hematopoietic differentiation, we assessed the reconstitution of various lineages in NBSGW mice transplanted with *AAVS1* control, partial *CTR9* edited, or complete *CTR9* edited HSPCs and found significantly higher reconstitution by HSPCs with partial *CTR9* editing, while complete *CTR9* editing significantly reduced the level of human hematopoietic engraftment (Figure 4D). Importantly, we found that no lineage deviated significantly from the reconstitution of human hematopoietic cells, as assessed by human CD45 expression, suggesting that the effect of partial or complete *CTR9* editing impacted overall hematopoietic reconstitution by HSCs, rather than selectively altering specific lineages (Figure 4E). Given this observed impact on HSC function, we estimated HSC numbers by modeling engraftment dynamics across the 20-week period using maximum likelihood estimation methods (STAR Methods) and found that HSPCs with partial loss of *CTR9* required ~250 functional HSCs, which is ~100 fewer than is seen for control *AAVS1* edited HSPCs, whereas complete *CTR9* edited cells require ~1200 more functional HSCs to achieve a similar level of engraftment (Figure 4F).

To further assess the capacity of long-term self-renewal by HSCs upon partial loss of CTR9, we performed serial replating with long-term engrafted HSCs after xenotransplantation. While colonies for all lineages were exhausted in control edited cells, partial *CTR9* edited cells had robust replating potential with minimal loss over four rounds of replating, providing further support for the observed increase in self-renewal (Figure 4G).

The prior experiments demonstrate how partial loss of CTR9 can alter hematopoiesis and thereby increase the likelihood for any HSC acquiring a somatic driver mutation. However, it is possible that in addition to the impact that would increase the risk for acquiring somatic driver mutations, there might also be cooperativity between these germline variants and drivers of the myeloid malignancies. We examined what happens when we combined partial CTR9 loss with concomitant loss of function mutations in drivers of the myeloid malignancies, TET2 and ASXL1. Strikingly, when CTR9 is partially edited in tandem with perturbation of TET2 or ASXL1, we observed an additive increase in HSC numbers in comparison with individual perturbations (Figure 4H). These observations extend our findings and show that these germline variants might also collaborate with myeloid malignancy drivers in enabling disease initiation and progression. Collectively, our data show that partial loss of CTR9 enables HSC expansion and promotes the development of myeloid malignancies.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Partial loss of CTR9 augments HSC function through enhanced transcription elongation

PAF1c has been suggested to be critical for maintenance of HSCs[43]. Therefore, the observed HSC expansion and increased HOXA family gene expression with heterozygous loss-of-function of *CTR9* appeared paradoxical and raised the question of underlying mechanisms by which blood cancer predisposition could arise. We examined differentially expressed genes in the HSC subpopulation through scRNA-seq. While gene ontology analysis on the differentially expressed genes (both those that are up and downregulated upon *CTR9* editing), revealed a number of pathways including alterations in RNA processing and metabolism, protein translation, processing and membrane targeting, and cellular respiration (Figure S5A), none of these provided clear insights into underlying mechanisms that could explain the observed HSC expansion. Given known interactions between PAF1c and components of the super elongation complex (SEC)[44–46], as well as the HSC expansion phenotype seen with increased expression of SEC component MLLT3 (also known as AF9)[47] that is similar to what we observe with CTR9 heterozygous loss-of-function, we examined the extent to which targets of MLLT3 in human HSCs were upregulated upon CTR9 editing. Remarkably, we observed a significant enrichment of MLLT3 upregulated genes that were increased upon CTR9 editing in HSCs by gene set enrichment analysis (normalized enrichment score = 3.29, p-value < 0.001, Figure 5A, B). These targets included many key HSC self-renewal regulators including mid to posterior HOXA genes and *MECOM*, and similar findings were also observed with an independent guide RNA (Figure S5B, C). Since increased SEC activity should promote transcription elongation, we examined nascent RNAs using precision run-on sequencing (PRO-seq) in edited $CD34^+$ HSPCs (STAR Methods). While there was a slight global increase in transcription elongation, the major contributions appeared to come from known MLLT3 targets in human hematopoiesis and from genes that are upregulated in the scRNA-seq data from HSCs (Figure 5C). Indeed, these genes showed dramatic downregulation in their calculated pausing indices, suggesting that augmented transcription elongation underlies their increased expression (Figure 5D, Figure S5D).

In light of these compelling findings suggesting that SEC activity might be increased in the setting of partial CTR9 loss (as would occur with heterozygous loss-of-function of *CTR9*), we sought to functionally validate these observations. We used a recently described specific and potent inhibitor of the YEATS domains of MLLT3 and MLLT1 (also known as ENL)[48–50]. Although we cannot formally discriminate between targeting of MLLT3 vs. MLLT1, MLLT3 is much more highly expressed in human HSCs than MLLT1 (Figure S5E), suggesting that MLLT3 is the major target in primary human HSCs. Using the YEATS domain inhibitor, SR-0813, we noted suppression of phenotypic LT- and ST-HSC expansion without a major impact on the bulk HSPC population (Figure S5F). Importantly, however, this inhibitor completely suppressed the CTR9-perturbation-mediated HSC expansion and concomitantly downregulated many of the key upregulated target genes (Figure S5G). To ensure that this suppression of MLLT3 was acting through the SEC, we also inhibited CDK9, a key kinase subunit of the p-TEFb subcomplex in the SEC that phosphorylates RNA polymerase II to promote transcription elongation, and observed a similar reduction of HSC expansion that arises from partial CTR9 loss (Figure S5H).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Relocalization of PAF1 complex subunits augments super elongation complex-mediated HSC gene transcription

Given the previously reported antagonism between PAF1c and the SEC[44], as well as given the specific alterations and myeloid malignancy predisposition observed due to CTR9 loss, we investigated the possibility that partial loss of CTR9 results in altered PAF1c complex assembly in primary human HSPCs and this thereby results in the increase in SEC activity observed. When we pulled down PAF1, all binding partners from PAF1c were co-immunoprecipitated as expected, but we also observed MLLT3 co-immunoprecipitation. Importantly, the amount of MLLT3 interacting with PAF1 increases with partial CTR9 loss, suggesting that CTR9 might act as an inhibitor of this interaction (Figure 5E). In contrast, when we immunoprecipitated MLLT3, only the PAF1 and CDC73 subunits of the PAF1c co-immunoprecipitated with MLLT3, and the amount of co-immunoprecipitation also increased with partial CTR9 loss (Figure 5E). Therefore, our results suggest that the increase in HSC self-renewal is a consequence of augmented SEC activity in the setting of reorganized binding by specific PAF1c subunits, including PAF1 and CDC73. To functionally assess the impact of perturbing these other PAF1c components on human HSCs, we edited *PAF1* and *CDC73*. In contrast to editing of *CTR9*, we observed that the phenotypic HSC populations were depleted even at early time points in culture with *PAF1* and *CDC73* editing (Figure S6B–C). Moreover, even Cas9 RNP titration to achieve heterozygous *PAF1* editing resulted in depletion of HSCs and progenitors (Figure S6A, D–E). Interestingly, when we edit another subunit of PAF1c that does not interact with MLLT3, LEO1, we observed an expansion of phenotypic HSCs in culture, as was the case with CTR9 (Figure S6B–C).

To extend our finding in an *in vivo* setting, we transplanted $CD34^+$ HSPCs with partial or complete *PAF1*, *CDC73*, or *LEO1* edits into NBSGW recipients and found that in contrast to *PAF1* and *CDC73* edits, which showed a significant reduction of engraftment in peripheral blood, bone marrow, and spleen, *LEO1* edited cells showed normal to higher engraftment compared to the *AAVS1* edited control, which recapitulated what we observed *in vitro* (Figure 5F, H, Figure S6F, G). Analysis of editing efficiency also showed that *PAF1* and *CDC73* edits were significantly depleted after 20 weeks, whereas *LEO1* edits were maintained at ~50% editing efficiency, similar to what we observed for CTR9 perturbations (Figure 5I). Notably, in contrast to CTR9, which showed an overall reduction of editing efficiency to ~30%−50%, *PAF1* or *CDC73* editing were reduced to negligible levels starting from 8 weeks post-transplantation, which further suggested that normal expression of PAF1 or CDC73 is required for HSC function, whereas LEO1 appears to be tolerant of partial loss, given that editing remained at ~50% while permitting effective reconstitution of long-term hematopoiesis. Further analysis of human hematopoietic reconstitution demonstrated broad impacts on all lineages by *PAF1* or *CDC73* edits, but not *LEO1* (Figure S6H, I). Consistent with this, maximum likelihood estimations (STAR Methods) revealed the need for significantly more HSCs that would be required with *PAF1* or *CDC73* editing to achieve similar reconstitution as seen for controls over 20 weeks (Figure 5J). These findings demonstrate that loss of PAF1 or CDC73 impairs human HSC function.

We next sought to more fully define the interaction between PAF1, CDC73 , and the SEC. With immunoprecipitation of PAF1 or MLLT3, all SEC subunits could be identified, with

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

the exception of AFF2, which is a core subunit of an alternative SEC complex, SEC-L2[51] (Figure S7A). We next sought to examine the nature of potential complexes in cells by performing gel filtration with lysates from human CD34$^+$ HSPCs after either *AAVS1* or partial *CTR9* editing. Indeed, consistent with previous reports[52], we observed that PAF1 and CDC73 are more widely distributed in comparison to other PAF1c subunits (Figure 6A). Moreover, upon partial loss of CTR9, PAF1 and CDC73 appear to redistribute to larger fractions that overlap with the SEC (Figure 6A–B, Figure S6B).

Our data suggests that MLLT3 enables PAF1 and CDC73 to interact with the SEC. To examine the nature of this interaction, we used the YEATS domain inhibitor, SR-0813, targeting MLLT3. Interestingly, inhibiting the MLLT3 YEATS domain does not alter the PAF1-CDC73-SEC interaction (Figure 6C). In contrast, when we deplete MLLT3 in CD34$^+$ HSPCs, the PAF1-CDC73-SEC interaction was largely abrogated (Figure 6D), demonstrating that MLLT3 is required for this interaction. Together, our data suggests that MLLT3 has a key role in the interface between PAF1-CDC73 and the SEC, but this does not require YEATS domain binding. We also overexpressed MLLT3 in CD34$^+$ HSPCs, but we did not observe any increase in interactions with higher MLLT3 expression (Figure 6C). This suggests that while MLLT3 is a critical part of this interaction involving PAF1 and CDC73, it is insufficient to drive this interaction, which suggests that other unknown factors are required for this process (Figure 6E). Finally, to bolster these observations, we could show that we could reverse the partial CTR9-dependent HSC expansion by concomitant perturbation of either PAF1 or CDC73, validating that the PAF1-CDC73-SEC interaction is necessary for the HSC expansion phenotype we observe (Figure 6F).

Given these findings, we sought to define how this alteration in human HSC function could arise mechanistically. We performed chromatin immunoprecipitation by sequencing (ChIP-seq) analyses to examine chromatin occupancy of MLLT3, PAF1, CDC73, LEO1, RNA polymerase II, along with two active transcription histone marks, H3K79me2 and H3K36me3, in human HSC-enriched populations. Consistent with our hypothesis, we observed a significant increase in the distribution of H3K79me2 and RNA polymerase II on the gene body regions of numerous HSC self-renewal genes upon partial CTR9 loss (Figure 7A). These observations are consistent with the augmentation of nascent RNA production upon partial CTR9 loss, which were seen by PRO-seq (Figure 7A). The levels of H3K4me3 across promoter regions showed no difference upon partial CTR9 loss, suggesting that no alterations in transcription initiation were present. Moreover, consistent with our biochemical observations in bulk nuclear extracts, we also observed that the chromatin occupancy of MLLT3, PAF1, CDC73, but not LEO1, on these genes was also increased upon partial CTR9 loss (Figure 7A). This reveals how this complex of select PAF1c components, including PAF1 and CDC73, appear to act with the SEC to augment transcription elongation at key genes necessary for HSC self-renewal. To more systematically assess the global occupancy of these subunits and their relationship with MLLT3, we focused on genomic regions that were previously shown to be regulated by MLLT3[47]. The intensity of MLLT3, PAF1, CDC73, but not LEO1 ChIP-seq signals on these genes were significantly enriched after partial CTR9 loss, which complement the augmented transcription elongation seen with PRO-seq. Moreover, there was also a significant increase in RNA polymerase II occupancy across these gene bodies, suggesting that these MLLT3-

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

regulated genes were more actively transcribed upon partial CTR9 loss (Figure 7B, C, Figure S7C).

Together, these functional experiments in tandem with our biochemical studies validate a model whereby heterozygous loss of CTR9, which confers significant predisposition to myeloid malignancies, frees specific PAF1c subunits, including PAF1 and CDC73, that can then cooperate with components of the SEC, including MLLT3, to augment and promote transcription elongation of genes necessary for HSC self-renewal (Figure 7D).

DISCUSSION

While significant progress has been made in understanding the numerous somatic mutations that can be acquired to give rise to the myeloid malignancies, a significant heritable component for these diseases exists[7,53], yet remains incompletely defined and poorly understood. Family-based studies have identified some large effect risk alleles that are rare in the population, while CVAS have defined common variants that are more frequently observed, yet confer smaller risks for developing these blood cancers. Here, we have employed RVAS to gain new insights into alleles that have comparable effects to those observed in some family-based studies, yet which are more common in the population. Indeed, in the context of CTR9, we estimate that ~0.1% of individuals in the population carry distinct deleterious mutations that confer a ~10-fold increased risk for acquiring a myeloid malignancy. Our study will enable further studies to define the full allelic spectrum of inherited genetic variation that confers risk to acquire these cancers.

Importantly, not only have we been able to define CTR9 and some other risk variants for the myeloid malignancies through our study, but we are also able to define critical new pathways involved in the regulation of transcription elongation. While prior studies had suggested interactions between the PAF1c and SEC in the context of transcription elongation[26,44], our follow up studies of CTR9 loss have uncovered how specific PAF1c components, namely PAF1 and CDC73, can interact with the SEC to augment its activity, at least it the context of human hematopoietic stem and progenitor cells, which are known to rely on SEC activity to modulate self-renewal[47]. While further studies are needed to more fully define how widespread these mechanisms might be and to more fully study how this variable localization of select components can be regulated, these insights pave the way for a deeper understanding of the regulation of transcription elongation.

Given that the incidence of CTR9 loss-of-function variants is far more common than the frequency of individuals who have myeloid malignancies, we have focused our study on how these variants can alter native hematopoiesis to confer an increased risk for acquiring somatic driver mutations that can cause these cancers. The increased self-renewal of human HSCs observed with CTR9 partial loss would increase the likelihood that one of these cells could acquire a mutation that could result in a myeloid malignancy. Importantly, we also observe cooperativity between these germline variants and specific somatic driver mutations, such as mutations disrupting TET2 and ASXL1, and this is an area that warrants further investigation in future studies.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

A critical implication of these findings is that they might suggest broader and more generalizable approaches for mitigating the risk of developing one of these blood cancers. Limiting HSC self-renewal and modulating their function reduces the likelihood of developing blood cancers in model systems[33–35]. Our genetic findings suggest that it might be possible to modulate these processes in at-risk individuals by altering transcription elongation to prevent myeloid malignancy initiation.

Limitations of the study

We would note a few limitations to this study. First, although we have used a large population-based human genetic study to identify rare variants in *CTR9* as conferring risk for acquiring myeloid malignancies, further population studies will help to better delineate the overall disease risk conferred by these variants, particularly in more diverse cohorts. Second, while we would ideally examine hematopoietic stem cells from individuals with germline *CTR9* variants, given the rarity of these variants and limited access to hematopoietic cells from individuals with known genotypes, we needed to rely upon isogenic genome editing to perturb *CTR9* in healthy hematopoietic cells to enable deep mechanistic studies. Third, while we have focused on mechanisms that regulate transcription elongation in human hematopoietic stem and progenitor cells, the applicability of these findings to other cellular contexts remains to be studied.

STAR METHODS

Resource Availability

Lead contact—Further information and reagent requests should be directed to and will be fulfilled by the Lead Contact Vijay G. Sankaran (sankaran@broadinstitute.org)

Materials availability—All unique materials will be available upon request to the lead contact.

Data and Code Availability—The data supporting the findings of this manuscript are reported in the main text, in the figures, or in the supplementary tables. UK Biobank individual genotypic and phenotypic data are available to approved investigators via the UK Biobank study (www.ukbiobank.ac.uk/). Additional information about registration for access to the data are available at www.ukbiobank.ac.uk/register-apply/. Data access for approved applications requires a data transfer agreement between the researcher's institution and UK Biobank. Sequencing data is available under GEO accession GSE186591.

Experimental model and subject details

Cell culture—HEK293T cells were cultured in DMEM supplemented with 10% FBS and 1X penicillin/streptomycin. Human $CD34^+$ HSPCs were purchased from Fred Hutchinson Cancer Institute and cultured in StemSpan serum free culture medium (StemCell Technologies) supplemented with 1X CC100, 2mM L-glutamine, 100ng/mL TPO, 1X UM171 and 1X penicillin/streptomycin.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Xenotransplantation and animal models—All animal procedures were performed under the protocol approved by Boston Children's Hospital Institutional Animal Care and Use Committee (IACUC). Xenotransplantation was done as previously described[37]. Briefly, cord blood was acquired from Dana-Farber Cancer Institute's Pasquarello tissue core. $CD34^+$ HSPCs were enriched by EasySep Human Cord Blood $CD34^+$ positive selection kit (StemCell Technologies) according to the manufacturer's instructions. Enriched cells were nucleofected with *AAVS1* control, *CTR9*, *PAF1*, *LEO1*, or *CDC73* sgRNAs in a manner identical to what was done for adult $CD34^+$ HSPCs using gRNA/Cas9 RNP delivered through nucleofection by Lonza 4D nucleofector as described in the above section. Two days post-nucleofection, a small fraction of cells will be harvested for ICE analysis of editing efficiency. The remaining cells will be intravenously injected into NOD.Cg-$Kit^{W-41J}Tyr^+Prkdc^{scid}Il2rg^{tm1Wjl}$/ThomJ (NBSGW) immunodeficient and Kit mutant mice (JAX#026622) at 1×10^5 cells per mouse. Autoclaved sulfatrim antibiotic water was given and changed weekly to prevent infections. Peripheral blood was harvested every four weeks, starting from 8 weeks post-transplantation, and bone marrow, and spleen were harvested 20 weeks post-transplantation for analysis of engraftment, editing efficiency of each corresponding edited gene, and lineage-specific markers. Specifically, peripheral blood was harvested through the retro-orbital plexus before sacrifice and through cardiac puncture right after the mice were sacrificed and collected in EDTA coated tubes to prevent coagulation. Red blood cells were depleted by addition of ddH_2O. The cells were resuspended in ddH_2O, which typically takes about 20 seconds for complete resuspension. The osmolarity was balanced immediately after resuspension by adding 10X PBS solution. The cells were centrifuged and pelleted at 470g for 5 minutes. If the pellet appears to be red, a second round of RBC lysis with the same method was used to deplete remaining RBCs. For bone marrow cell harvest, femurs, and tibia were dissected and crushed by mortar and pestle in RPMI 1640 medium supplemented with 10% FBS and 1X penicillin/streptomycin. The cells were passed through 40 μm cell strainer. The cells were washed by ice-cold PBS once and washed cells were frozen down in 90% FBS + 10% DMSO. The majority of RBCs were automatically depleted by cryopreservation. Prior to analysis, the thawed bone marrow cells were further RBC depleted by the above method. For splenocyte isolation, spleens were dissected and minced. The minced spleen was digested by 1 mg/ml collagenase IV for 30 minutes at 37 degrees C. The digestion was stopped by addition of EDTA at a final concentration of 1 mM. Splenocytes were frozen in 90% FBS + 10% DMSO to deplete RBCs and stored in liquid nitrogen for subsequent analysis.

Estimated HSC required for a specific extent of engraftment is calculated by the maximum likelihood estimation method (MLE) using the following formula: $HSC_{required}$ = -log(1 -fraction_stem_cells) * total_cells_transplanted / mean(log(1-donor_chimerism)), where the fraction of stem cell frequency here is assumed to be 0.005 across all groups. While not providing a precise estimate of the number of HSCs in a mix population, as can be achieved with limiting dilution analyses, this approach provides an estimate of how many HSCs are required to achieve a specified level of engraftment and thus provides a metric of HSC numbers based on xenotransplantation[54–57].

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

UKB RVAS—UK Biobank (UKB) whole exome sequencing (WES) data (Category 170, UKB Showcase) was accessed under application 31063. Written consent is in place for all UKB participants. Any participants whom have since withdrawn their consent from the UKB were removed from this analysis. Exome capture was conducted using IDT xGen Exome Research Panel v1.0 including supplemental probes. The population-level variant call format (pVCF) file was created by UKB as described elsewhere[58] (STAR Methods sections: OQFE protocol and OQFE mapping and variant calling). All data was aligned to the GRCh38 reference genome.

Quality Control—The original, unfiltered, pVCF from UKB was subjected to several filtering requirements prior to analysis. Filtering was performed in Hail[59]. Genotypes: Genotype entries were hard filtered using DP > 10 and GQ > 20. Allele balance was required to be within 0.25 and 0.75 if a heterozygous call, or <0.1, >0.9 if a homozygous call. Samples: Samples were excluded if they were not of self-reported White-British ancestry with the same genetic ancestry inference based on a principal components analysis, not included in kinship inference, an excess (>10) of putative third-degree relatives inferred from kinship, an outlier in heterozygosity and missing rates, and putative sex chromosome aneuploidy (generated by UKB). Samples were also filtered out if they had a missing rate of >0.1. Variants: Variants were excluded if they were multiallelic, had a missing rate >0.1, Hardy-Weinberg equilibrium p-value of >1e-15. Variants were required to have an allele frequency of <0.01 (1%) to be classified as rare, anything above this threshold was filtered.

Annotation of variants—Variants in the filtered pVCF files were annotated using VEP[60] (including the LOFTEE[61] plugin), ANNOVAR[62] (using dbNSFP v4.1[63], ClinVar v20200316[64] and ALoFT[65]), and InMeRF[66]. Genes were annotated using RefSeq (via ANNOVAR). Loss-of-function (LoF) was determined by LOFTEE and ALoFT tools, variants were required to be high-confidence (HC), with none of the following warning flags (“SINGLE_EXON”, “NAGNAG_SITE”, “PHYLOCSF_WEAK”, “PHYLOCSF_UNLIKELY_ORF”, “NON_CAN_SPLICE”). Missense variants were determined by an InMeRF score of >0.5. The choice of InMeRF to determine deleterious variants was made from an AUROC performance analysis of several existing tools in dbNSFP. Classifying variants (True Positives) used were ClinVar (Sept-2018 release) “Pathogenic” clinical significance interpretation (CLNSIG) filtered to only include high confidence variants using the CLNREVSTAT field: “multiple submitters”, a “reviewed by expert panel” or using “practice guideline”, and “no conflicts” (N=4460). True negatives were common benign missense variants were obtained from the gnomAD database (in this database there are no known severe Mendelian disorders and as such it is assumed that highly penetrant disease-causing missense variants will be rare in this database (allele frequency < 1%)), these were downsampled for computational efficiency (N=9200). Supplementary Table S2 provides results, InMeRF provided the highest performance (AUROC=0.83), significantly higher than the second highest AUROC (MetaSVM), p=0.007. Importantly, multiple compound combinations of tools were also estimated to have lower performance than InMeRF classification alone.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Phenotype derivation—To gain the largest degree of power and because of the view that these phenotypes share many elements, we combined myeloproliferative neoplasm (MPN), acute myeloid leukemia (AML), and myelodysplastic syndrome (MDS) into a myeloid malignancy mega-phenotype for primary analysis. The individual phenotypes were curated as follows. MPN: polycythemia (ICD10 D45; ICD9 2384), essential thrombocythemia (ICD10 D47.3, D75.2), osteomyelofibrosis (ICD10 D47.4, D75.81), chronic myeloid leukemia (ICD10 C921, C922, C931; ICD9 2051), and chronic myeloproliferative disease (ICD10 D47.1). AML: acute myeloid leukemia (ICD10 C92.0, C92.4, C92.5, C92.6, C92.8; ICD9 2050). MDS: myelodysplastic syndrome (ICD10 D46.0, D46.1, D46.2, D46.3, D46.4, D46.5, D46.6, D46.7, D46.9). Individuals were also classified as cases if they had a self-reported cancer, self-reported illness code, or histology of cancer tumor code for any of the above.

Association testing—When collapsing, only genes for which there was a combined minor allele count (cMAC) of 20 or greater were included in the association analysis. Only deleterious variants were used in the collapsing association test, i.e. classified LoF or disruptive missense. Association in the 200K cohort was tested using SAIGE-GENE v0.44.1[67], which utilizes a robust saddle point approximation (SPA) SKAT-O approach to collapsing tests in a generalized linear mixed model. The genetic relationship matrix (GRM) for step 0 and 1 was formed using common markers (allele frequency >0.01) and was formed using PLINK[68,69]. Association analysis in the 500K cohort was tested using REGENIE v3.2.5 in two steps[70]. First, a whole-genome logistic regression model was used to estimate the fraction of phenotypic variance explained by a subset of SNPs. Then, the predictors from step 1 are used to perform gene based association testing by collapsing SNPs with InMeRF > 0.5 and LoF annotations. The following were included as covariates in the association testing: age, gender, first 10 principal components of ancestry. In addition, the originally sequenced 50K exome samples used a different IDT v1.0 oligo lot so a covariate was included to label this subset[71]. To correct for the multiple genes across the genome being tested, p-values were adjusted using the FDR Benjamini-Hochberg procedure, any gene unit with a q-value less than 0.05 (5% FDR) was considered significant.

Plotting—Results were composed into a Manhattan style plot, by obtaining the RefSeq start coordinates of the genes. Plotting was done in R v4.0.3[72], using ggplot2[73].

Estimation of effect size—Effect size was estimated in a cohort of unrelated samples. The KING kinship coefficient was calculated and used to determine those in 3rd-degree relation or greater, then one of the sample pairs was removed. Priority was given to cases, afterwards exclusion was based on random selection. To allow convergence despite rare events, three methods were used to determine effect size for significant genes, penalized maximum likelihood firth-biased logistic regression (using brglm2[74]), bayesian MCMC logistic regression (using rstanarm[75]), and generalized logistic regression (using glm-base).

CTR9 germline assessment—We undertook several ancillary approaches to robustly classify deleterious *CTR9* variants as likely to be germline. First, as is becoming more common in WGS/WES sequencing analyses, we analysed VAF or percentage alternative

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

reads[18] and found *CTR9* variants consistently had a VAF over 0.3[76] (Figure S1A). Although helpful, VAF does not completely enable separation between somatic clones with large allelic burdens >0.3/0.4, particularly for samples obtained from the blood. Second, we analysed age at blood sampling compared to diagnosis. It is known that somatic clone size is often largest around diagnosis, somatic clones in MPN are detectable, in some cases, years before diagnosis, but are often of small VAF. Years after diagnosis, cytoreductive therapy often reduces clone size and VAF[77]. We found that in 4 of the *CTR9* deleterious variant carriers, blood was sampled years before diagnosis, and that in the remaining 3 blood was sampled at least 5 years after diagnosis (Figure S1D). Both scenarios would suggest a smaller clone size than the average VAF detected in *CTR9* at ~0.45. In addition, none of the deleterious CTR9 carriers progressed to AML which would be the most likely of the myeloid diseases here to possess a clone likely to progress to a large VAF. Indeed, we know currently that *CTR9* does not harbor driver mutations in CHIP or myeloid malignancies from extensive prior studies. Through analysis of COSMIC, a widely used somatic mutation database with just over 336,000 hematopoietic cancer tissue samples, we observed no evidence for *CTR9* variants acting as driver mutations in hematologic malignancies[23]. Of those 336,000 only 4 carry confirmed somatic variants in the CTR9 gene making the probability (4/336,000 = 0.00001) of somatic mutation in *CTR9* very small. Lastly we note in a comprehensive recent review on this topic from Kraft and Godley,[76] that “diagnosis of a hematopoietic malignancy at a younger age than seen in the general population” is a factor leaning towards germline variation. In our analysis, we find *CTR9* deleterious variant carriers have a median age at diagnosis of 59 [50 – 62] yrs old whereas other myeloid malignancy cases that are not carrying deleterious *CTR9* variants are on average 64 [57 – 70] years old, p=0.10 (Wilcoxon rank sum). Although this is a nonsignificant difference, there is certainly a tendency towards younger age in the carriers, and if there were more carriers of these rare deleterious variants in CTR9 then we might see a significant difference in age. We also called somatic variants using Mutect2 in the analyzed samples and found no evidence of somatic calls in the *CTR9* gene. Familial cases are not present to analyze. All of this taken together suggest that the observed deleterious alleles in *CTR9* are likely to be in the germline.

PheWAS of CTR9—Phenotypes were mapped onto ICD10 codes using a Phecodes map file[78]. Only phenotypes with more than n=60 cases were included in the analysis. 1027 phenotypes were tested. SAIGE-GENE was run on the *CTR9* gene region with deleterious variants (as previously defined) individually on each phenotype and results were combined.

Somatic mutation analysis—Somatic calls in UKBB were derived from CRAM files using Mutect2. Quality control included excluding variants seen in a panel of normal samples (n=100) aged less than 40 years old, excluding variants with MMQ<25 and MBQ<15, excluding genotypes with a read depth (DP) <20 and a VAF <2%, as well as the default passing filters employed by Mutect2 and FilterMutectCalls. CHIP was defined in individuals according to previous standards using somatic calls (Table S2 in Bick et. al.)[21]. Prevalence of CHIP in UKBB closely followed that of previous reports (Figure S1C). We found, within the 166,953 filtered samples used for this analysis, that N_{chip}= 4,682.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

CRISPR editing in CD34⁺ HSPCs, and HSC phenotyping—For RNP based CRISPR editing of $CD34^+$ HSPCs, chemically synthesized gRNA and Cas9 protein were mixed and set at room temperature for 10 mins to allow the formation of deliverable RNP. The assembled RNP complexes were then delivered into $CD34^+$ HSPCs by a Lonza 4D Nucleofector with program DZ-100. 48–72hrs after nucleofection, cells were harvested for genomic DNA and the edited region was PCR amplified and the editing efficiency was analyzed by the ICE analysis from Synthego. Edited $CD34^+$ HSPCs were allowed to expand in the culture medium and harvested twice at days 3 and 6 (for complete CTR9 edits) or days 4 and 6 (for titrated CTR9 edits) post-nucleofection. Harvested cells were washed in cold PBS and stained by HSC phenotyping antibody cocktail, which is composed of Brilliant Violet 421 anti-CD34, APC-H7 anti-CD45RA, PE-Cy7 anti-CD90, BB700 anti-CD133, PE anti-EPCR and APC anti-ITGA3 in ice cold FACS buffer (PBS+2%FBS) at 1:60 for 30 minutes in ice[6,37]. Stained cells were then analyzed by BD LSRII flow cytometer. ST-HSC is defined as $CD34^+CD45RA^-CD90^+CD133^+EPCR^+ITGA3^-$ population, whereas LT-HSC is defined as $CD34^+CD45RA^-CD90^+CD133^+EPCR^+ITGA3^+$ cells. For the chemical rescue experiment, CDK9 inhibitors (LDC000067 and AZD4573) and the YEATS domain inhibitor of ENL/MLLT3 (SR-0813) were used at the concentration of IC50.

Molecular Cloning and co-Immunoprecipitation—Human *CTR9* cDNA was amplified from the cDNA library of 293T cells. The amplified *CTR9* cDNA was then TA cloned and sequenced for verification. pLVX-IRES-ZsGreen and pLVX-IRES-mCherry vector, which was modified from pLVX-IRES-ZsGreen by replacing ZsGreen by mCherry, was digested by NotI restriction enzyme at 37C overnight. The verified CTR9 cDNA was amplified as three fragments with FLAG and/or HA tags included in the N terminus and overhangs overlapped each other and were brought into digested vectors by the InFusion cloning kit from ClonTech. Positive colonies were confirmed by Sanger Sequencing. All CTR9 mutation primers were designed by NEBase Changer online software (https://nebasechanger.neb.com/) and the mutagenesis was performed using Q5 mutagenesis kit according to manufacturer's instructions. Sequence verified WT and mutant constructs were co-transfected by Lipofectamine 3000 kit at 1:1 molar ratio. For co-immunoprecipitation, cells were then harvested 48 hrs post-transfection (for HEK293T cells) or 5 days post-nucleofection (for CD34+ HSPCs). The cells were then lysed by EBC buffer (50 mM Tris pH8.0, 120mM NaCl, 0.5%NP40) in presence of protease inhibitor cocktail and PMSF. The extracted protein was then quantified by DC protein quantification assay and 0.5 mg of protein lysate was used for immunoprecipitation. Samples were pre-incubated with target antibody (PAF1, AF9, FLAG, HA or IgG) for 2 hours at 4°C on a head-over-tail rotor, the protein-antibody complex was then incubated with 30 microliter of protein A/G dynabeads for addition 1 hour at 4°C. The supernatant was then removed and the beads were then washed by NETN buffer (20mM Tris pH8.0, 100mM NaCl, 0.5% NP40) for five times at room temperature. The immunoprecipitated proteins were then eluted by 1XSDS sample loading buffer with 55°C heat and run on 5%−20% precast gel along with 5% input for analysis. For normalization and IP efficiency calculations, the band intensity was determined by Biorad Image Lab software. Quantified bands in the IP group were then normalized by the band intensity of the corresponding input.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Colony Forming Unit Assay—Three days post-nucleofection, edited $CD34^+$ HSPCs were plated at a density of 500 cells/mL methylcellulose medium (H4034, Stem Cell Technologies) according to manufacturer's instructions. Plated cells were allowed to grow for 14 days before quantification and analysis. For serial replating colony assays, 50,000 bone marrow cells from long-term xenotransplant recipients were plated in methylcellulose medium. After 7–9 days of growth, the colonies were counted, and then harvested for replating with the same cell numbers.

Single Cell RNA-seq Analysis—Edited $CD34^+$ HSPCs were sorted and enriched for $CD34^+CD45RA^-CD90^+$ population. The library was prepared using the 10X Genomics protocol according to the manufacturer's instructions. The sequencing results were pre-analyzed by the CellRanger pipeline to generate the matrix file, which was brought to downstream analysis. Normalization, scaling and cell clustering was done in R by the Seurat v4.0.2 package. The top 1000 most highly variable genes were used to calculate the top 50 principal components. The HSC cluster was highlighted by averaging the z-score normalized expression of *CD34, HLF, CRHBP* for each cell. The HSC population, defined by the averaged z score of more than 0.5 for the HSC signature, was used for DEG calling by DESeq2.

Real Time PCR Analysis and Immunoblotting—RNA from HSPCs was harvested by Qiagen RNeasy kit and quantified by nanodrop. 1 microgram of total RNA was then reverse transcribed using Biorad iScript cDNA synthesis kit following manufacturer's instructions. The reverse transcribed cDNA was then diluted (1:20) and real time PCR was run using Biorad iQ SYBR green supermix. Data was normalized by loading control and presented as fold change compared to control samples. For immunoblotting, total protein lysate was extracted by RIPA buffer in presence of protease inhibitor cocktail and PMSF on ice for 10 minutes followed by a 5-minute incubation with MNase at 37C. The total lysate was then linearized by 1X SDS loading buffer and heated at 55C for 10 minutes. The lysate was run on 5–20% gradient gel and then transferred onto PVDF membrane. The membrane was then blocked by 5% nonfat dry milk in TBS/T. The membrane was then incubated with primary antibodies at 1:1000 dilution at 4C overnight. HRP conjugated secondary antibodies were then incubated for 1hr at room temperature. Membrane was developed using ECL and imaged in the Biorad gel imaging system.

PRO-seq library construction—Aliquots of frozen (−80°C) permeabilized $CD34^+$ HSPCs were thawed on ice and pipetted gently to fully resuspend. Aliquots were removed and permeabilized cells were counted using a Luna II, Logos Biosystems instrument. For each sample, 1 million permeabilized cells were used for nuclear run-on, with 50,000 permeabilized Drosophila S2 cells added to each sample for normalization. Nuclear run on assays and library preparation were performed essentially as described in Reimer et al.[79] with modifications noted: 2X nuclear run-on buffer consisted of (10 mM Tris (pH 8), 10 mM MgCl2, 1 mM DTT, 300mM KCl, 20uM/ea biotin-11-NTPs (Perkin Elmer), 0.8U/uL SuperaseIN (Thermo), 1% sarkosyl). Run-on reactions were performed at 37°C. Adenylated 3' adapter was prepared using the 5' DNA adenylation kit (NEB) and ligated using T4 RNA ligase 2, truncated KQ (NEB, per manufacturers instructions with 15%

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

PEG-8000 final) and incubated at 16°C overnight. 180uL of betaine blocking buffer (1.42g of betaine brought to 10mL with binding buffer supplemented to 0.6 uM blocking oligo (TCCGACGATCCCACGTTCCCGTGG/3InvdT/)) was mixed with ligations and incubated 5 min at 65°C and 2 min on ice prior to addition of streptavidin beads. After T4 polynucleotide kinase (NEB) treatment, beads were washed once each with high salt, low salt, and blocking oligo wash (0.25X T4 RNA ligase buffer (NEB), 0.3uM blocking oligo) solutions and resuspended in 5' adapter mix (10 pmol 5' adapter, 30 pmol blocking oligo, water). 5' adapter ligation was per Reimer but with 15% PEG-8000 final. Eluted cDNA was amplified 5-cycles (NEBNext Ultra II Q5 master mix (NEB) with Illumina TruSeq PCR primers RP-1 and RPI-X) following the manufacturer's suggested cycling protocol for library construction. A portion of preCR was serially diluted and for test amplification to determine optimal amplification of final libraries. Pooled libraries were sequenced using the Illumina NovaSeq platform.

PRO-seq data analysis—All custom scripts described herein are available on the AdelmanLab Github (https://github.com/AdelmanLab/NIH_scripts). Using a custom script (trim_and_filter_PE.pl), FASTQ read pairs were trimmed to 41bp per mate, and read pairs with a minimum average base quality score of 20 were retained. Read pairs were further trimmed using cutadapt 1.14 to remove adapter sequences and low-quality 3' bases (--match-read-wildcards -m 20 -q 10). R1 reads, corresponding to RNA 3' ends, were then aligned to the spiked in Drosophila genome index (dm6) using BWA, with those reads not mapping to the spike genome serving as input to the primary genome alignment step. Reads mapping to the hg38 reference genome were then sorted, via samtools 1.3.1 (-n), and subsequently converted to bam files. The bam files are converted to bigwig files by bamCoverage of deepTools-3.5. For metagene plots, bigwig files of three replicates of each group and combined and averaged using WiggleTools. The pausing indices were calculated using NRSA v2 packages[80].

Gel filtration (size exclusion chromatography)—Gel filtration similar to a prior report[81]. Briefly, modified human $CD34^+$ HSPCs were lysed to obtain nuclear extracts. The nuclear extract was then dialyzed with gel filtration buffer, which contains 20 mM 4-(2-hydroxyethyl)-1-piperazine-ethanesulfonic acid, 10 mM KCl, 1.5 mM MgCl2, 1 mM ethylenediaminetetraacetic acid (EDTA), 1 mM (ethylenebis(oxonitrilo))tetra-acetate, 0.02% 3-((3-cholamidopropyl)dimethylammonio)propane-1-sulfonic acid (CHAPS), pH 7.5 supplemented with supplemented with 200 mM NaCl and 1 mM DTT, concentrated to 0.5 ml using the Amicon ultracentrifugal filter and loaded on to column loaded with Superose-6 resin (Cytiva) pre-equilibrated with gel-filtration buffer. Proteins were eluted at a flow rate of 0.5 ml min^{-1} and collected at 0.2 ml per fraction. Eluted protein fractions were then analyzed by Western blots.

ChIP-seq analysis

Native ChIP (N-ChIP)—For histone mark ChIP (i.e. H3K4me3, H3K36me3 and H3K79me2), sorted $CD34^+CD45RA^-CD90^+$ cells were lysed in hypotonic buffer (10 mmol/L HEPES pH7.9, 10 mmol/L KCl, 1.5 mmol/L $MgCl_2$, 0.34 mol/L sucrose, 10% glycerol, and 1 mmol/L DTT) in presence of protease inhibitor cocktail[82], the lysed

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

cells were then digested by MNase until the majority of chromatin was digested to mononucleosome. The rest of the digested genomic DNA was precleared by rabbit IgG and then immunoprecipitated by H3K4me3, H3K36me3 and H3K79me2 antibody bound to protein A/G dynabeads in RIPA buffer overnight at 4°C. The complex was then washed by RIPA buffer twice, high salt RIPA buffer (RIPA buffer + 300 mmol/L NaCl) twice, LiCl wash buffer (250 mmol/L LiCl, 0.5% NP-40, 0.5% sodium deoxycholate) twice and 1XTE buffer + 50 mmol/L NaCl. The DNA was then eluted in SDS elution buffer (50 mmol/L Tris pH 8.0, 10 mmol/L EDTA, 1% SDS) at 65°C for at least 6 hours and treated with RNase and proteinase K. The eluted ChIPed DNA was then purified by NEB Monarch PCR purification kit.

Crosslinked ChIP (X-ChIP)—MLLT3, PAF1, LEO1, CDC73 and RNA polymerase II ChIP was done with Active Motif ChIP-IT Express Enzymatic kit according to the manufacturer's instructions. Briefly, the sorted $CD34^{+}CD45RA^{-}CD90^{+}$ cells were fixed in 1% formaldehyde with methanol for 10 minutes at room temperature. The fixation was terminated by addition of 1X glycine stop solution after a brief wash with ice-cold PBS to remove residual fixative buffer. The cells were then sheared by diluted shearing enzyme cocktail at 37 degrees celsius for 10 minutes. The shearing reaction was stopped by addition of EDTA. Before immunoprecipitation, the protein A/G beads were pre-cleared by rabbit or mouse IgG, depending on the source of primary antibody being used. Up to 25 μg of sheared chromatin were used for ChIP with 3 μg of target antibody. Incubation of primary antibody and protein A/G beads were performed at 4 degrees celsius overnight in presence of protease inhibitor cocktail. The complexes were then washed by ChIP buffer 1 for once and ChIP buffer 2 for twice. For DNA elution, the ChIPed complexes were resuspended with 50 μl of AM2 elution buffer and incubated for 15 minutes at room temperature. The DNA was reverse crosslinked by addition of 50 μl of reverse crosslink buffer. The reverse crosslinked DNA was then digested by protease K and RNase A for 4 hours at room temperature. DNA was purified by the NEB Monarch PCR purification kit.

Library preparation—The ChIP-seq library was prepared by NEBnext Ultra II DNA kit according to the manufacturer's instructions. Briefly, the ChIPed DNA was end-repaired and A tailed by the end repairing and A-tailing module. The end-repaired and A tailed DNA was then ligated by the adaptor, followed by a 15-minute treatment of USER enzyme for removal of the U base at 37 degrees celsius. The adaptor-ligated DNA was then proceeded to library amplification.

ChIP-seq data analysis—The raw fastq files were QCed by fastqc software for detection of adaptor sequences. The fastq files were then trimmed for adaptor sequences by cutadapt software. The trimmed fastq files were then aligned to hg38 genome by Bowtie2 software by default setting. The aligned bam files were used for broad peak calling by MACS2 with input as the control sample. The bam files were converted to bigwig files by DeepTools software with the bamCoverage command with bin size of 2. The bigwig files of three replicates of each ChIP-seq experiment were then merged by the mean using the UCSC toolbox. Chromatin occupancy profiles or heatmap plots were plotted by DeepTools software by "computeMatrix" followed by "plotProfile" or "plotHeatmap" commands.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Quantification and Statistical Analysis

All experiments were done with at least three biological replicates. For $CD34^+$ HSPC experiments, biological replicates are defined by use of samples from different donors. For xenotransplantation experiments, it is done with at least five biological replicates. In all experiments, data were presented as mean±standard deviation. When comparing two samples, two-tailed Student's t test was used for testing statistical significance. When comparing three or more samples, Levene's test was firstly used to test equality of variance. If the variance across all samples were tested insignificantly differed, one way or two way ANOVA with Dunnett's test (for multiple comparison where no reference group is defined) or Tukey's test (for multiple comparison where reference group is defined) as post-hoc analysis was used. If the variance across samples were tested to be significantly different, Kruskal-Wallis test was used instead of ANOVA, with the Dunn test as the post-hoc multiple comparison test used. All statistical tests were performed in Graphpad software or in R, when statistical tests were not available through Graphpad.

Supplementary Tables

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

The authors would like to thank D. Nathan, S. Orkin, A. Muntean, A. Mullally, P. van Galen, K. Adelman, and members of the Sankaran laboratory for valuable comments and suggestions. We also thank the research participants and staff of the UK Biobank. We are grateful to T. DiCesare for assistance in making illustrations. This work was funded by the New York Stem Cell Foundation (V.G.S.); the MPN Research Foundation (V.G.S.); the Edward P. Evans Foundation (V.G.S.), a gift from the Lodish Family to Boston Children's Hospital (V.G.S.); Alex's Lemonade Stand Foundation Crazy 8 Grant (S.M.V.); Leukemia & Lymphoma Society (S.A.A.); and National Institutes of Health Grants R01 CA265726 (V.G.S.), R01 DK103794 (V.G.S.), and R01 HL146500 (V.G.S.). S.M.V. is a Howard Hughes Medical Institute Freeman Hrabowski Scholar. V.G.S. is a Robertson Investigator of the New York Stem Cell Foundation.

REFERENCES

1. Arber DA, Orazi A, Hasserjian RP, Borowitz MJ, Calvo KR, Kvasnicka H-M, Wang SA, Bagg A, Barbui T, Branford S, et al. (2022). International Consensus Classification of Myeloid Neoplasms and Acute Leukemias: integrating morphologic, clinical, and genomic data. Blood 140, 1200–1228. 10.1182/blood.2022015850. [PubMed: 35767897]
2. Miles LA, Bowman RL, Merlinsky TR, Csete IS, Ooi AT, Durruthy-Durruthy R, Bowman M, Famulare C, Patel MA, Mendez P, et al. (2020). Single-cell mutation analysis of clonal evolution in myeloid malignancies. Nature 587, 477–482. 10.1038/s41586-020-2864-x. [PubMed: 33116311]
3. Sperling AS, Gibson CJ, and Ebert BL (2017). The genetics of myelodysplastic syndrome: from clonal haematopoiesis to secondary leukaemia. Nat. Rev. Cancer 17, 5–19. 10.1038/nrc.2016.112. [PubMed: 27834397]
4. van Galen P, Hovestadt V, Wadsworth MH Ii, Hughes TK, Griffin GK, Battaglia S, Verga JA, Stephansky J, Pastika TJ, Lombardi Story J, et al. (2019). Single-Cell RNA-Seq Reveals AML Hierarchies Relevant to Disease Progression and Immunity. Cell 176, 1265–1281.e24. 10.1016/j.cell.2019.01.031. [PubMed: 30827681]
5. Zeng AGX, Bansal S, Jin L, Mitchell A, Chen WC, Abbas HA, Chan-Seng-Yue M, Voisin V, van Galen P, Tierens A, et al. (2022). A cellular hierarchy framework for understanding heterogeneity

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

and predicting drug response in acute myeloid leukemia. Nat. Med 28, 1212–1223. 10.1038/s41591-022-01819-x. [PubMed: 35618837]

6. Voit RA, Tao L, Yu F, Cato LD, Cohen B, Fleming TJ, Antoszewski M, Liao X, Fiorini C, Nandakumar SK, et al. (2022). A genetic disorder reveals a hematopoietic stem cell regulatory network co-opted in leukemia. Nat. Immunol, 1–15. 10.1038/s41590-022-01370-4. [PubMed: 34789862]
7. Klco JM, and Mullighan CG (2021). Advances in germline predisposition to acute leukaemias and myeloid neoplasms. Nat. Rev. Cancer 21, 122–137. 10.1038/s41568-020-00315-z. [PubMed: 33328584]
8. Bao EL, Nandakumar SK, Liao X, Bick AG, Karjalainen J, Tabaka M, Gan OI, Havulinna AS, Kiiskinen TTJ, Lareau CA, et al. (2020). Inherited myeloproliferative neoplasm risk affects haematopoietic stem cells. Nature 586, 769–775. 10.1038/s41586-020-2786-7. [PubMed: 33057200]
9. Hinds DA, Barnholt KE, Mesa RA, Kiefer AK, Do CB, Eriksson N, Mountain JL, Francke U, Tung JY, Nguyen HM, et al. (2016). Germ line variants predispose to both JAK2 V617F clonal hematopoiesis and myeloproliferative neoplasms. Blood 128, 1121–1128. 10.1182/blood-2015-06-652941. [PubMed: 27365426]
10. Tapper W, Jones AV, Kralovics R, Harutyunyan AS, Zoi K, Leung W, Godfrey AL, Guglielmelli P, Callaway A, Ward D, et al. (2015). Genetic variation at MECOM, TERT, JAK2 and HBS1L-MYB predisposes to myeloproliferative neoplasms. Nat. Commun 6, 6691. 10.1038/ncomms7691. [PubMed: 25849990]
11. McGraw KL, Cheng C-H, Chen YA, Hou H-A, Nilsson B, Genovese G, Cluzeau T, Pellagatti A, Przychodzen BP, Mallo M, et al. (2019). Non-del(5q) myelodysplastic syndromes-associated loci detected by SNP-array genome-wide association meta-analysis. Blood Adv 3, 3579–3589. 10.1182/bloodadvances.2019000922. [PubMed: 31738830]
12. Walker CJ, Oakes CC, Genutis LK, Giacopelli B, Liyanarachchi S, Nicolet D, Eisfeld A-K, Scholz M, Brock P, Kohlschmidt J, et al. (2019). Genome-wide association study identifies an acute myeloid leukemia susceptibility locus near BICRA. Leukemia 33, 771–775. 10.1038/s41375-018-0281-z. [PubMed: 30291333]
13. Wang J, Clay-Gilmour AI, Karaesmen E, Rizvi A, Zhu Q, Yan L, Preus L, Liu S, Wang Y, Griffiths E, et al. (2021). Genome-Wide Association Analyses Identify Variants in IRF4 Associated With Acute Myeloid Leukemia and Myelodysplastic Syndrome Susceptibility. Front. Genet 12, 554948. 10.3389/fgene.2021.554948. [PubMed: 34220922]
14. Homan CC, Scott HS, and Brown AL (2023). Hereditary platelet disorders associated with germ line variants in RUNX1, ETV6, and ANKRD26. Blood 141, 1533–1543. 10.1182/blood.2022017735. [PubMed: 36626254]
15. Makishima H, Bowman TV, and Godley LA (2023). DDX41-associated susceptibility to myeloid neoplasms. Blood 141, 1544–1552. 10.1182/blood.2022017715. [PubMed: 36455200]
16. Calvo KR, and Hickstein DD (2023). The spectrum of GATA2 deficiency syndrome. Blood 141, 1524–1532. 10.1182/blood.2022017764. [PubMed: 36455197]
17. Reilly CR, and Shimamura A (2023). Predisposition to myeloid malignancies in Shwachman-Diamond syndrome: biological insights and clinical advances. Blood 141, 1513–1523. 10.1182/blood.2022017739. [PubMed: 36542827]
18. Wang Q, Dhindsa RS, Carss K, Harper AR, Nag A, Tachmazidou I, Vitsios D, Deevi SVV, Mackay A, Muthas D, et al. (2021). Rare variant contribution to human disease in 281,104 UK Biobank exomes. Nature 10.1038/s41586-021-03855-y.
19. Backman JD, Li AH, Marcketta A, Sun D, Mbatchou J, Kessler MD, Benner C, Liu D, Locke AE, Balasubramanian S, et al. (2021). Exome sequencing and analysis of 454,787 UK Biobank participants. Nature 599, 628–634. 10.1038/s41586-021-04103-z. [PubMed: 34662886]
20. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, Motyer A, Vukcevic D, Delaneau O, O'Connell J, et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature 562, 203–209. 10.1038/s41586-018-0579-z. [PubMed: 30305743]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

21. Bick AG, Weinstock JS, Nandakumar SK, Fulco CP, Bao EL, Zekavat SM, Szeto MD, Liao X, Leventhal MJ, Nasser J, et al. (2020). Inherited causes of clonal haematopoiesis in 97,691 whole genomes. Nature 586, 763–768. 10.1038/s41586-020-2819-2. [PubMed: 33057201]
22. Bowman RL, Busque L, and Levine RL (2018). Clonal Hematopoiesis and Evolution to Hematopoietic Malignancies. Cell Stem Cell 22, 157–170. 10.1016/j.stem.2018.01.011. [PubMed: 29395053]
23. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, Boutselakis H, Cole CG, Creatore C, Dawson E, et al. (2019). COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Res 47, D941–D947. 10.1093/nar/gky1015. [PubMed: 30371878]
24. Stubbins RJ, Korotev S, and Godley LA (2022). Germline CHEK2 and ATM Variants in Myeloid and Other Hematopoietic Malignancies. Curr. Hematol. Malig. Rep 17, 94–104. 10.1007/s11899-022-00663-7. [PubMed: 35674998]
25. Hanks S, Perdeaux ER, Seal S, Ruark E, Mahamdallie SS, Murray A, Ramsay E, Del Vecchio Duarte S, Zachariou A, de Souza B, et al. (2014). Germline mutations in the PAF1 complex gene CTR9 predispose to Wilms tumour. Nat. Commun 5, 4398. 10.1038/ncomms5398. [PubMed: 25099282]
26. Yu M, Yang W, Ni T, Tang Z, Nakadai T, Zhu J, and Roeder RG (2015). RNA polymerase II-associated factor 1 regulates the release and phosphorylation of paused RNA polymerase II. Science 350, 1383–1386. 10.1126/science.aad2338. [PubMed: 26659056]
27. Chen FX, Xie P, Collings CK, Cao K, Aoi Y, Marshall SA, Rendleman EJ, Ugarenko M, Ozark PA, Zhang A, et al. (2017). PAF1 regulation of promoter-proximal pause release via enhancer activation. Science 357, 1294–1298. 10.1126/science.aan3269. [PubMed: 28860207]
28. Van Oss SB, Cucinotta CE, and Arndt KM (2017). Emerging Insights into the Roles of the Paf1 Complex in Gene Regulation. Trends Biochem. Sci 42, 788–798. 10.1016/j.tibs.2017.08.003. [PubMed: 28870425]
29. Li Y, Zhang J, Adikaram PR, Welch J, Guan B, Weinstein LS, Chen H, and Simonds WF (2020). Genotype of CDC73 germline mutation determines risk of parathyroid cancer. Endocr. Relat. Cancer 27, 483–494. 10.1530/ERC-20-0149. [PubMed: 32590342]
30. Carpten JD, Robbins CM, Villablanca A, Forsberg L, Presciuttini S, Bailey-Wilson J, Simonds WF, Gillanders EM, Kennedy AM, Chen JD, et al. (2002). HRPT2, encoding parafibromin, is mutated in hyperparathyroidism-jaw tumor syndrome. Nat. Genet 32, 676–680. 10.1038/ng1048. [PubMed: 12434154]
31. Vos SM, Farnung L, Boehning M, Wigge C, Linden A, Urlaub H, and Cramer P (2018). Structure of activated transcription complex Pol II-DSIF-PAF-SPT6. Nature 560, 607–612. 10.1038/s41586-018-0440-4. [PubMed: 30135578]
32. Vos SM, Farnung L, Linden A, Urlaub H, and Cramer P (2020). Structure of complete Pol II-DSIF-PAF-SPT6 transcription complex reveals RTF1 allosteric activation. Nat. Struct. Mol. Biol 27, 668–677. 10.1038/s41594-020-0437-1. [PubMed: 32541898]
33. Yamashita M, Dellorusso PV, Olson OC, and Passegué E (2020). Dysregulated haematopoietic stem cell behaviour in myeloid leukaemogenesis. Nat. Rev. Cancer 20, 365–382. 10.1038/s41568-020-0260-3. [PubMed: 32415283]
34. Cimmino L, Dolgalev I, Wang Y, Yoshimi A, Martin GH, Wang J, Ng V, Xia B, Witkowski MT, Mitchell-Flack M, et al. (2017). Restoration of TET2 Function Blocks Aberrant Self-Renewal and Leukemia Progression. Cell 170, 1079–1095.e20. 10.1016/j.cell.2017.07.032. [PubMed: 28823558]
35. Agathocleous M, Meacham CE, Burgess RJ, Piskounova E, Zhao Z, Crane GM, Cowin BL, Bruner E, Murphy MM, Chen W, et al. (2017). Ascorbate regulates haematopoietic stem cell function and leukaemogenesis. Nature 549, 476–481. 10.1038/nature23876. [PubMed: 28825709]
36. Tomellini E, Fares I, Lehnertz B, Chagraoui J, Mayotte N, MacRae T, Bordeleau M-È, Corneau S, Bisaillon R, and Sauvageau G (2019). Integrin-α3 Is a Functional Marker of Ex Vivo Expanded Human Long-Term Hematopoietic Stem Cells. Cell Rep 28, 1063–1073.e5. 10.1016/j.celrep.2019.06.084. [PubMed: 31340144]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

37. Zhao J, Jia Y, Mahmut D, Deik AA, Jeanfavre S, Clish CB, and Sankaran VG (2023). Human hematopoietic stem cell vulnerability to ferroptosis. Cell 186, 732–747.e16. 10.1016/j.cell.2023.01.020. [PubMed: 36803603]
38. Wan L, Chong S, Xuan F, Liang A, Cui X, Gates L, Carroll TS, Li Y, Feng L, Chen G, et al. (2020). Impaired cell fate through gain-of-function mutations in a chromatin reader. Nature 577, 121–126. 10.1038/s41586-019-1842-7. [PubMed: 31853060]
39. Dou DR, Calvanese V, Sierra MI, Nguyen AT, Minasian A, Saarikoski P, Sasidharan R, Ramirez CM, Zack JA, Crooks GM, et al. (2016). Medial HOXA genes demarcate haematopoietic stem cell fate during human development. Nat. Cell Biol 18, 595–606. 10.1038/ncb3354. [PubMed: 27183470]
40. Lebert-Ghali C-É, Fournier M, Kettyle L, Thompson A, Sauvageau G, and Bijl JJ (2016). Hoxa cluster genes determine the proliferative activity of adult mouse hematopoietic stem and progenitor cells. Blood 127, 87–90. 10.1182/blood-2015-02-626390. [PubMed: 26585953]
41. McIntosh BE, Brown ME, Duffin BM, Maufort JP, Vereide DT, Slukvin II, and Thomson JA (2015). Nonirradiated NOD,B6.SCID Il2rγ−/− KitW41/W41 (NBSGW) Mice Support Multilineage Engraftment of Human Hematopoietic Cells. Stem Cell Reports 4, 171–180. 10.1016/j.stemcr.2014.12.005. [PubMed: 25601207]
42. Fiorini C, Abdulhay NJ, McFarland SK, Munschauer M, Ulirsch JC, Chiarle R, and Sankaran VG (2017). Developmentally-faithful and effective human erythropoiesis in immunodeficient and Kit mutant mice. Am. J. Hematol 92, E513–E519. 10.1002/ajh.24805. [PubMed: 28568895]
43. Saha N, Ropa J, Chen L, Hu H, Mysliwski M, Friedman A, Maillard I, and Muntean AG (2019). The PAF1c Subunit CDC73 Is Required for Mouse Hematopoietic Stem Cell Maintenance but Displays Leukemia-Specific Gene Regulation. Stem Cell Reports 12, 1069–1083. 10.1016/j.stemcr.2019.03.010. [PubMed: 31031188]
44. Chen FX, Woodfin AR, Gardini A, Rickels RA, Marshall SA, Smith ER, Shiekhattar R, and Shilatifard A (2015). PAF1, a Molecular Regulator of Promoter-Proximal Pausing by RNA Polymerase II. Cell 162, 1003–1015. 10.1016/j.cell.2015.07.042. [PubMed: 26279188]
45. Hetzner K, Garcia-Cuellar M-P, Büttner C, and Slany RK (2018). The interaction of ENL with PAF1 mitigates polycomb silencing and facilitates murine leukemogenesis. Blood 131, 662–673. 10.1182/blood-2017-11-815035. [PubMed: 29217648]
46. He N, Chan CK, Sobhian B, Chou S, Xue Y, Liu M, Alber T, Benkirane M, and Zhou Q (2011). Human Polymerase-Associated Factor complex (PAFc) connects the Super Elongation Complex (SEC) to RNA polymerase II on chromatin. Proc. Natl. Acad. Sci. U. S. A 108, E636–E645. 10.1073/pnas.1107107108. [PubMed: 21873227]
47. Calvanese V, Nguyen AT, Bolan TJ, Vavilina A, Su T, Lee LK, Wang Y, Lay FD, Magnusson M, Crooks GM, et al. (2019). MLLT3 governs human haematopoietic stem-cell self-renewal and engraftment. Nature 576, 281–286. 10.1038/s41586-019-1790-2. [PubMed: 31776511]
48. Erb MA, Scott TG, Li BE, Xie H, Paulk J, Seo H-S, Souza A, Roberts JM, Dastjerdi S, Buckley DL, et al. (2017). Transcription control by the ENL YEATS domain in acute leukaemia. Nature 543, 270–274. 10.1038/nature21688. [PubMed: 28241139]
49. Garnar-Wortzel L, Bishop TR, Kitamura S, Milosevich N, Asiaban JN, Zhang X, Zheng Q, Chen E, Ramos AR, Ackerman CJ, et al. (2021). Chemical Inhibition of ENL/AF9 YEATS Domains in Acute Leukemia. ACS Cent Sci 7, 815–830. 10.1021/acscentsci.0c01550. [PubMed: 34079898]
50. Wan L, Wen H, Li Y, Lyu J, Xi Y, Hoshii T, Joseph JK, Wang X, Loh Y-HE, Erb MA, et al. (2017). ENL links histone acetylation to oncogenic gene expression in acute myeloid leukaemia. Nature 543, 265–269. 10.1038/nature21687. [PubMed: 28241141]
51. Luo Z, Lin C, and Shilatifard A (2012). The super elongation complex (SEC) family in transcriptional control. Nat. Rev. Mol. Cell Biol 13, 543–547. 10.1038/nrm3417. [PubMed: 22895430]
52. Kim J, Guermah M, and Roeder RG (2010). The human PAF1 complex acts in chromatin transcription elongation both independently and cooperatively with SII/TFIIS. Cell 140, 491–503. 10.1016/j.cell.2009.12.050. [PubMed: 20178742]
53. Sud A, Chattopadhyay S, Thomsen H, Sundquist K, Sundquist J, Houlston RS, and Hemminki K (2018). Familial risks of acute myeloid leukemia, myelodysplastic syndromes, and

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

myeloproliferative neoplasms. Blood 132, 973–976. 10.1182/blood-2018-06-858597. [PubMed: 29991558]

54. Goyal S, Kim S, Chen ISY, and Chou T (2015). Mechanisms of blood homeostasis: lineage tracking and a neutral model of cell populations in rhesus macaques. BMC Biol 13, 85. 10.1186/s12915-015-0191-8. [PubMed: 26486451]
55. Kumaravelu P, Hook L, Morrison AM, Ure J, Zhao S, Zuyev S, Ansell J, and Medvinsky A (2002). Quantitative developmental anatomy of definitive haematopoietic stem cells/long-term repopulating units (HSC/RUs): role of the aorta-gonad-mesonephros (AGM) region and the yolk sac in colonisation of the mouse embryonic liver. Development 129, 4891–4899. 10.1242/dev.129.21.4891. [PubMed: 12397098]
56. Berger CN, and Sturm KS (1996). Estimation of the number of hematopoietic precursor cells during fetal mouse development by covariance analysis. Blood 88, 2502–2509. [PubMed: 8839842]
57. Payne RW (1993). Genstat 5 Release 3 Reference Manual (Clarendon Press).
58. Szustakowski J, Balasubramanian S, Sasson A, Khalid S, Bronson PG, Kvikstad E, Wong E, Liu D, Davis JW, Haefliger C, et al. (2020). Advancing Human Genetics Research and Drug Discovery through Exome Sequencing of the UK Biobank. medRxiv, 2020.11.02.20222232. 10.1101/2020.11.02.20222232.
59. Hail Team. (2023). Hail (0.2.126). Zenodo. 10.5281/zenodo.10055486
60. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, Flicek P, and Cunningham F (2016). The Ensembl Variant Effect Predictor. Genome Biol 17, 122. 10.1186/s13059-016-0974-4. [PubMed: 27268795]
61. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 581, 434–443. 10.1038/s41586-020-2308-7. [PubMed: 32461654]
62. Wang K, Li M, and Hakonarson H (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38, e164. 10.1093/nar/gkq603. [PubMed: 20601685]
63. Liu X, Li C, Mou C, Dong Y, and Tu Y (2020). dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. Genome Med 12, 103. 10.1186/s13073-020-00803-9. [PubMed: 33261662]
64. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Jang W, et al. (2018). ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res 46, D1062–D1067. 10.1093/nar/gkx1153. [PubMed: 29165669]
65. Balasubramanian S, Fu Y, Pawashe M, McGillivray P, Jin M, Liu J, Karczewski KJ, MacArthur DG, and Gerstein M (2017). Using ALoFT to determine the impact of putative loss-of-function variants in protein-coding genes. Nat. Commun 8, 382. 10.1038/s41467-017-00443-5. [PubMed: 28851873]
66. Takeda J-I, Nanatsue K, Yamagishi R, Ito M, Haga N, Hirata H, Ogi T, and Ohno K (2020). InMeRF: prediction of pathogenicity of missense variants by individual modeling for each amino acid substitution. NAR Genom Bioinform 2, lqaa038. 10.1093/nargab/lqaa038. [PubMed: 33543123]
67. Zhou W, Zhao Z, Nielsen JB, Fritsche LG, LeFaive J, Gagliano Taliun SA, Bi W, Gabrielsen ME, Daly MJ, Neale BM, et al. (2020). Scalable generalized linear mixed model for region-based association tests in large biobanks and cohorts. bioRxiv, 583278. 10.1101/583278.
68. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, and Lee JJ (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 4, 7. 10.1186/s13742-015-0047-8. [PubMed: 25722852]
69. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet 81, 559–575. 10.1086/519795. [PubMed: 17701901]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

70. Mbatchou J, Barnard L, Backman J, Marcketta A, Kosmicki JA, Ziyatdinov A, Benner C, O'Dushlaine C, Barber M, Boutkov B, et al. (2021). Computationally efficient whole-genome regression for quantitative and binary traits. Nat. Genet 53, 1097–1103. 10.1038/s41588-021-00870-7. [PubMed: 34017140]
71. Van Hout CV, Tachmazidou I, Backman JD, Hoffman JD, Liu D, Pandey AK, Gonzaga-Jauregui C, Khalid S, Ye B, Banerjee N, et al. (2020). Exome sequencing and characterization of 49,960 individuals in the UK Biobank. Nature 586, 749–756. 10.1038/s41586-020-2853-0. [PubMed: 33087929]
72. R Core Team (2020). R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing).
73. Wickham H (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. ISBN 978–3-319–24277-4, https://ggplot2.tidyverse.org
74. Kosmidis I, Kenne Pagui EC, and Sartori N (2020). Mean and median bias reduction in generalized linear models. Stat. Comput 30, 43–59. 10.1007/s11222-019-09860-6.
75. Brilleman S, Crowther M, Moreno-Betancur M, Buros Novik J, Wolfe R (2018). "Joint longitudinal and time-to-event models via Stan." StanCon 2018. 10–12. Pacific Grove, CA, USA., https://github.com/stan-dev/stancon_talks/.
76. Kraft IL, and Godley LA (2020). Identifying potential germline variants from sequencing hematopoietic malignancies. Blood 136, 2498–2506. 10.1182/blood.2020006910. [PubMed: 33236764]
77. Vannucchi AM, Pieri L, and Guglielmelli P (2011). JAK2 Allele Burden in the Myeloproliferative Neoplasms: Effects on Phenotype, Prognosis and Change with Treatment. Ther. Adv. Hematol 2, 21–32. 10.1177/2040620710394474. [PubMed: 23556073]
78. Wu P, Gifford A, Meng X, Li X, Campbell H, Varley T, Zhao J, Bastarache L, Denny JC, Theodoratou E, et al. (2018). Developing and Evaluating Mappings of ICD-10 and ICD-10-CM codes to Phecodes. bioRxiv, 462077. 10.1101/462077.
79. Reimer KA, Mimoso CA, Adelman K, and Neugebauer KM (2021). Co-transcriptional splicing regulates 3' end cleavage during mammalian erythropoiesis. Mol. Cell 81, 998–1012.e7. 10.1016/j.molcel.2020.12.018. [PubMed: 33440169]
80. Wang J, Zhao Y, Zhou X, Hiebert SW, Liu Q, and Shyr Y (2018). Nascent RNA sequencing analysis provides insights into enhancer-mediated gene regulation. BMC Genomics 19, 633. 10.1186/s12864-018-5016-z. [PubMed: 30139328]
81. Wang Z, Chen K, Jia Y, Chuang J-C, Sun X, Lin Y-H, Celen C, Li L, Huang F, Liu X, et al. (2020). Dual ARID1A/ARID1B loss leads to rapid carcinogenesis and disruptive redistribution of BAF complexes. Nat Cancer 1, 909–922. 10.1038/s43018-020-00109-0. [PubMed: 34386776]
82. Zhao J, Jia Y, Shen S, Kim J, Wang X, Lee E, Brownell I, Cho-Vega JH, Lewis C, Homsi J, et al. (2020). Merkel Cell Polyomavirus Small T Antigen Activates Noncanonical NF-κB Signaling to Promote Tumorigenesis. Mol. Cancer Res 18, 1623–1637. 10.1158/1541-7786.MCR-20-0587. [PubMed: 32753470]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Highlights

- Inherited *CTR9* loss-of-function variants predispose to the myeloid malignancies.
- Partial, but not complete, loss of CTR9 expands human HSCs.
- Partial CTR9 loss expands HSCs through increased transcription elongation.
- Select PAF1 complex subunits interact with and activate the super elongation complex.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

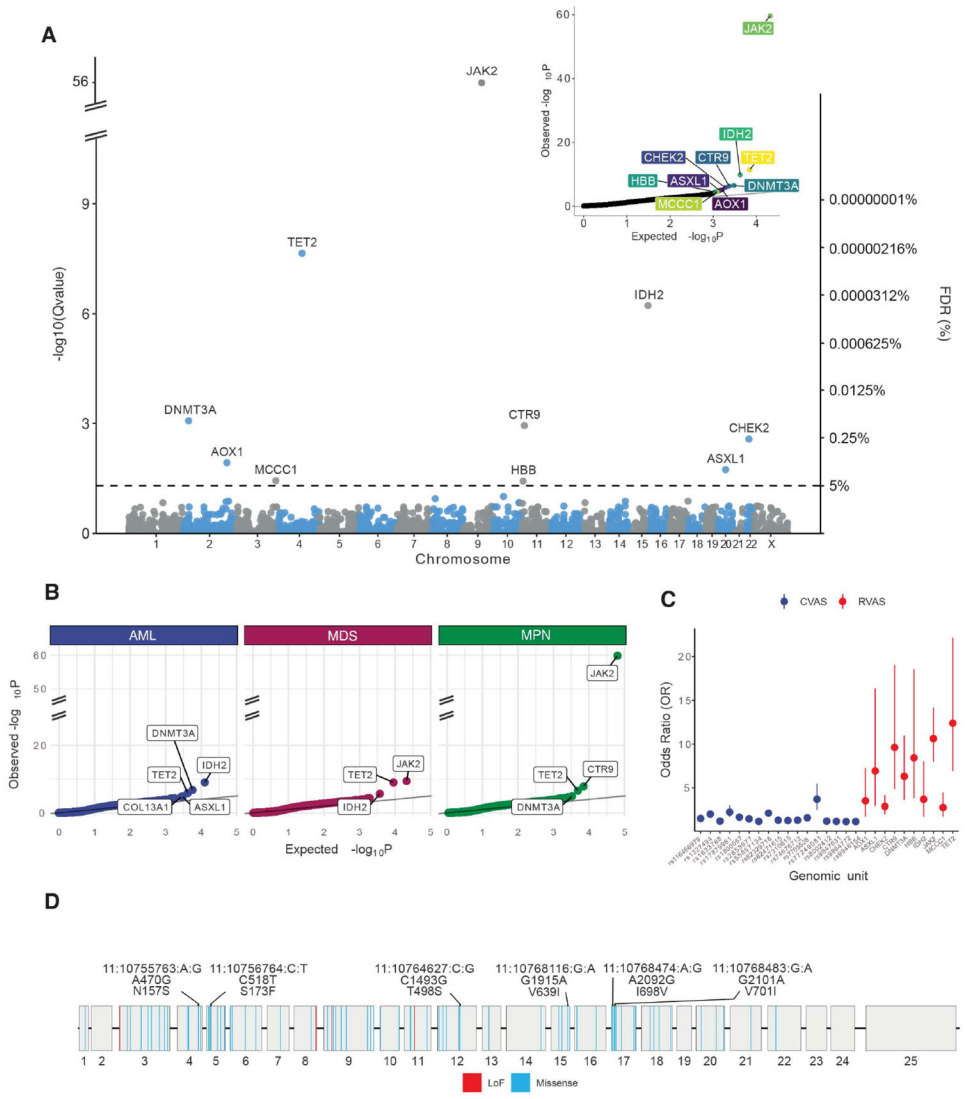


Figure 1. Rare variant association results from UK Biobank show CTR9 among several significant genes.

(A) Rare variant association results, collapsing of predicted loss-of-function and nonsynonymous variants that are grouped by gene. Dashed line indicates the FDR threshold of significance. The inset of this panel shows a Q-Q plot demonstrating the -log10 p-values expected for each gene and the observed values. Higher than expected values that deviate from the best fit line indicate significant enrichments. (B) Dividing myeloid malignancies into constituent diseases, most signal for CTR9 arises from MPNs. (C) CTR9 shows an odds ratio of 9.6 (95% CI=4.86–19.04, SKAT-O p-value=5.47×10–7). Odds ratios for collapsing tests on genes (red) are generally much larger than single variant odds on disease risk (blue; showing GWAS significant common variant markers for MPN disease2). (D) 25 coding exomes of CTR9 are shown with identified LoF variants (red), and predicted nonsynonymous variants (blue). Variants found in myeloid malignancy cases are labelled.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

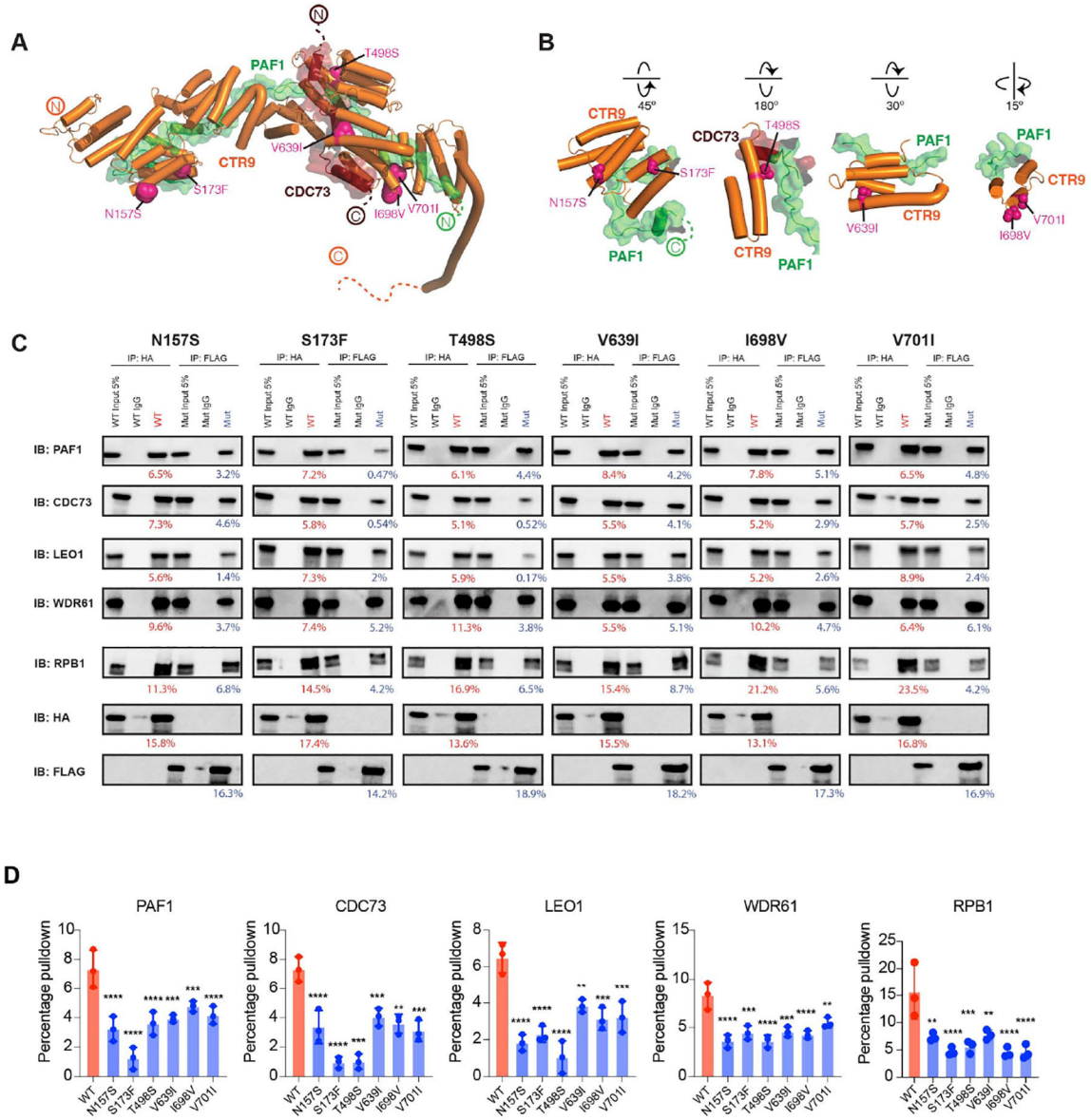


Figure 2. CTR9 variants result in loss of PAF1 complex assembly.

(A) Structure of CTR9-PAF1-CDC73 subcomplex with loss-of-function variants highlighted (hot pink spheres) (PDB ID: 6TED). (B) Zoom in structures of CTR9 deleterious variants. Rotations relative to view in panel A. (C) Co-immunoprecipitation results (with quantification) of six CTR9 variants from myeloid malignancy cases in the discovery cohort. (D) Quantification of immunoprecipitation of (C). All data are presented as mean ± standard deviations (*p < 0.05; **p < 0.01; ***p < 0.001, ****p < 0.0001).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

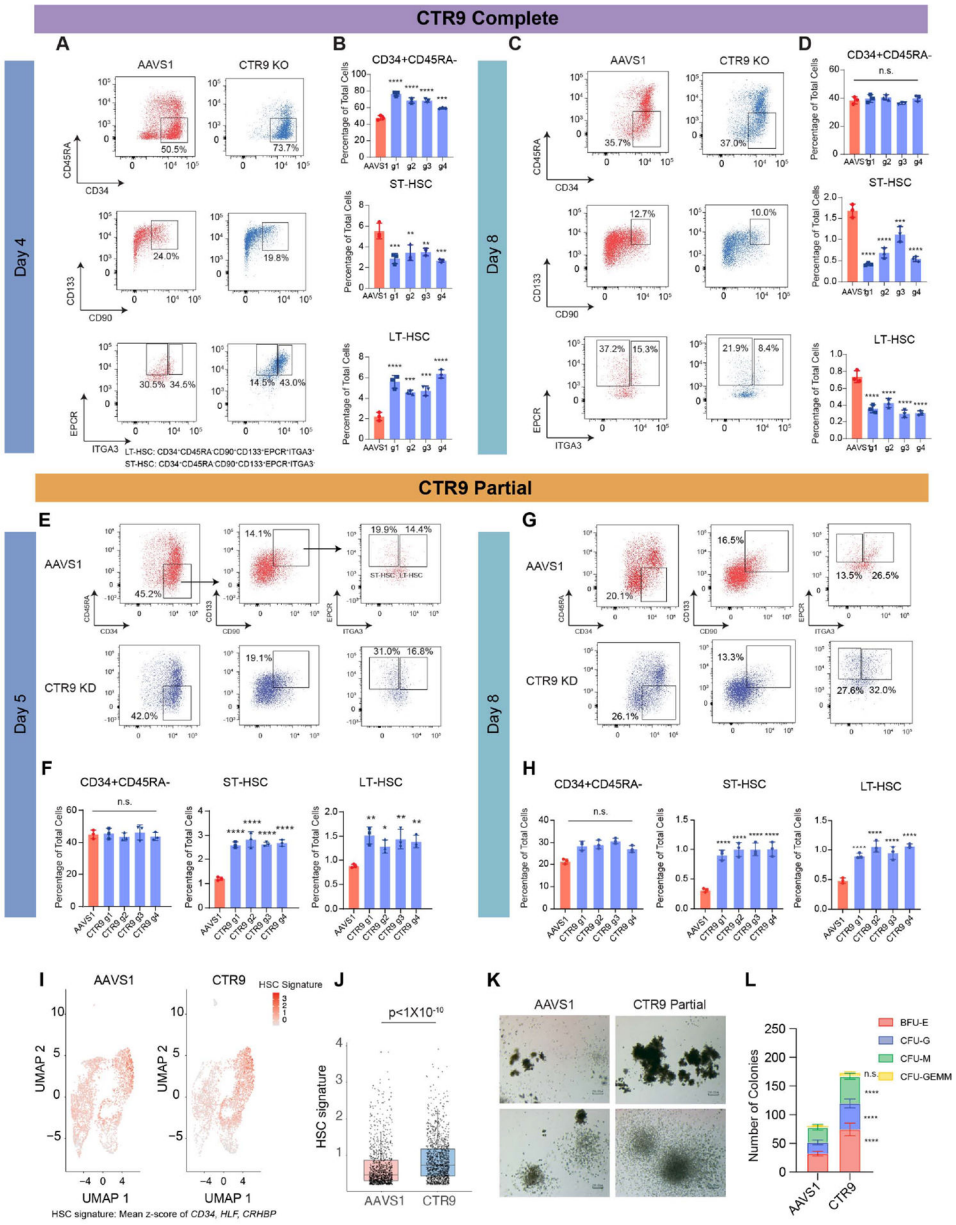


Figure 3. Partial loss of CTR9 results in phenotypic HSC expansion.

(A) Representative FACS plot of HSC phenotyping of CTR9 editing three days post-nucleofection. (B) Quantification of $CD34^{+}CD45RA^{-}$, ST-HSC, and LT-HSC populations three days post-nucleofection. Data are presented as mean±SD for three biological replicates. (C) Representative FACS plot of HSC phenotyping of CTR9 editing at seven days post-nucleofection. ST-HSC is defined as the $CD34^{+}CD45RA^{-}CD90^{+}CD133^{+}EPCR^{+}ITGA3^{-}$ population, whereas LT-HSC is defined as the $CD34^{+}CD45RA^{-}CD90^{+}CD133^{+}EPCR^{+}ITGA3^{+}$ population. (D) Quantification of $CD34^{+}CD45RA^{-}$, ST-HSC, and LT-HSC populations seven days post-nucleofection. (E) Representative FACS plot and gating strategy of HSC in AAVS1 edited control and heterozygous edited CTR9 $CD34^{+}$ HSPCs three days post-nucleofection. (F) Quantification

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

of $CD34^{+}CD45RA^{-}$, ST-HSC, and LT-HSC populations in AAVS1 control and CTR9 edited groups. (G) Representative FACS plot of HSC phenotyping of CTR9 editing six days post-nucleofection. (H) Quantification of $CD34^{+}CD45RA^{-}$, ST-HSC, and LT-HSC populations six days post-nucleofection. (I) Uniform manifold approximation and projection (UMAP) plots of 3,335 (AAVS1) and 4,154 (CTR9) $CD34^{+}CD45RA^{-}CD90^{+}$ cells, highlighted according to average z-score normalized HSC gene signature. The HSC signature is defined as the averaged z-score of *CD34, HLF* and *CRHBP* expression in scRNA-seq. (J) Box plot of cells with z-score normalized HSC signature expression>0.2 of AAVS1 and CTR9 edit groups. (K) Representative light microscopy of colony forming unit assay of AAVS1 and CTR9 edit groups. Scale bar = 200μm. (L) Quantification of total colonies of BFU-E, CFU-G, CFU-M, and CFU-GEMM in AAVS1 and CTR9 edited groups. BFU-E: Burst-forming unit-erythroid, CFU-G: Colony-forming unit-granulocyte, CFU-M: Colony-forming unit-Monocyte, CFU-GEMM: Colony-forming unit-granulocyte/monocyte/erythroid/megakaryocyte. All data are presented as mean ± standard deviations (*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$, ****$p < 0.0001$).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

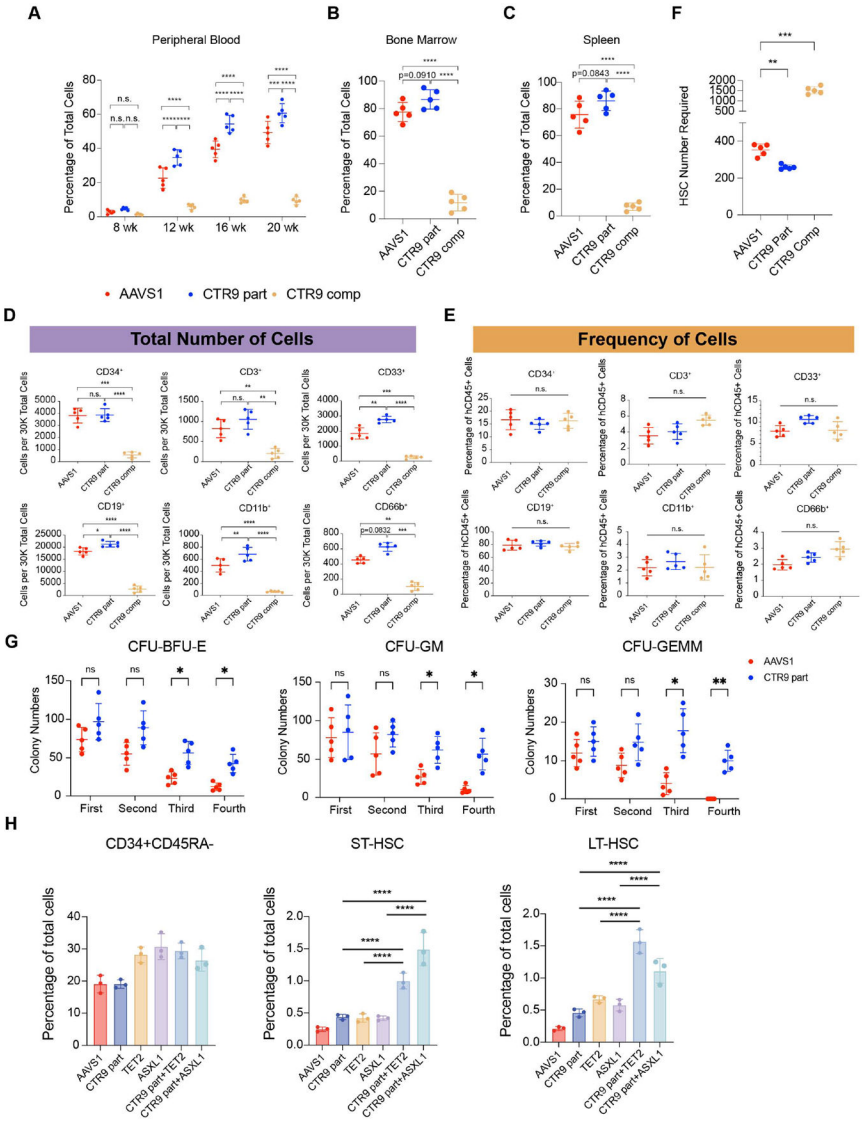


Figure 4. Phenotypically expanded HSCs due to partial loss of CTR9 are functional.

(A) Percentage of total human CD45+ cells after RBC depletion from peripheral blood of NBSGW mice xenotransplanted with AAVS1, partial or complete CTR9 edited (g1) cord blood $CD34^+$ HSPCs at indicated times post-transplantation. (B) Percentage of total human CD45+ cells after RBC depletion from the bone marrow of NBSGW mice xenotransplanted with AAVS1, partial or complete CTR9 edited cord blood $CD34^+$ HSPCs 20 weeks post-transplantation. (C) Percentage of total human CD45+ cells after RBC depletion from the spleen of NBSGW mice xenotransplanted with AAVS1, partial or complete CTR9 edited cord blood $CD34^+$ HSPCs 20 weeks post-transplantation. (D) Cell numbers per 30,000 RBC depleted bone marrow cells of progenitor cells ($CD34^+$), T cells ($CD3^+$), myeloid ($CD33^+$, $CD11b^+$), B cells ($CD19^+$), and granulocytes ($CD66b^+$). (E) Frequency of human CD45+ RBC depleted bone marrow cells of progenitor cells ($CD34^+$), T cells ($CD3^+$), myeloid ($CD33^+$, $CD11b^+$), B cells ($CD19^+$), and granulocytes ($CD66b^+$). (F) Estimated required HSC numbers for the observed reconstitution dynamics for AAVS1, partial or complete

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

CTR9 edited cord blood $CD34^+$ HSPCs in (A). All data are presented as mean ± standard deviations (*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$, ****$p < 0.0001$). (G) Serial replating assays using CTR9 or AAVS1 edited long-term xenotransplanted HSCs over four rounds, with 7 days of growth in each round. (H) Quantification of $CD34^+CD45RA^-$, ST-HSC, and LT-HSC populations of cells edited with AAVS1, CTR9 partial, TET2, ASXL1, or co-edits of CTR9 partial and TET2 or ASXL1 seven days post-nucleofection.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

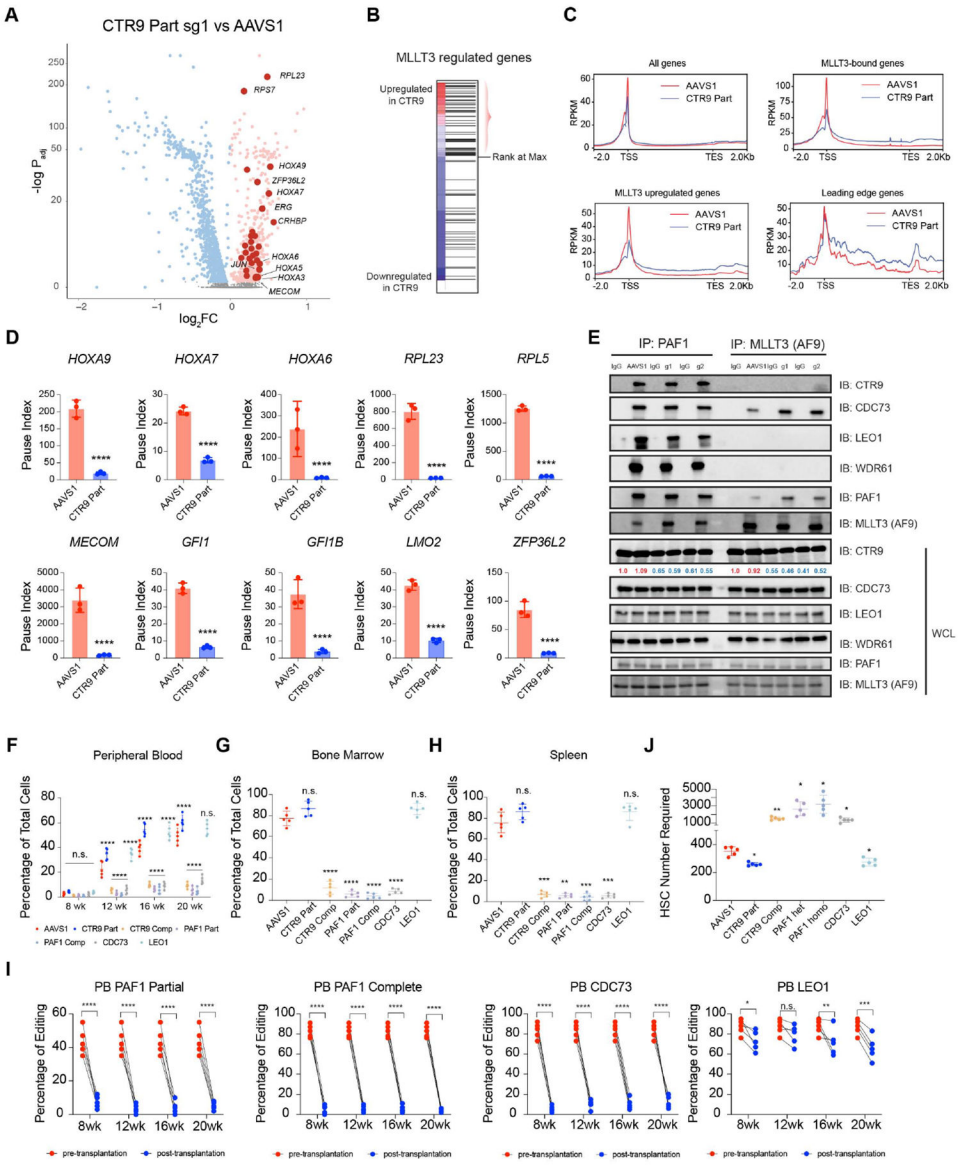


Figure 5. Partial loss of CTR9 increases super elongation complex activity to drive HSC self-renewal gene expression.

(A) Volcano plot of gene expression of HSCs edited by CTR9 vs. AAVS1, highlighting genes that are MLLT3 regulated at the leading edge by gene set enrichment analysis (bold red dots), y-axis for p-values is on a custom scale (Methods). (B) Gene set enrichment analysis of CTR9 vs. AAVS1 edited HSCs for the MLLT3 regulated HSC gene set. Leading edge genes are highlighted. (C) PRO-seq metagene plots for all, MLLT3 bound, MLLT3 regulated and leading edge genes of CTR9 (g1) vs AAVS1 edited cells. (D). Quantification of pausing indices of key leading edge genes in AAVS1 control vs. CTR9 KO groups. (E) Co-immunoprecipitation of CD34^{+} HSPCs by PAF1 and MLLT3 (AF9) antibodies with AAVS1 and CTR9 edited by two guide RNAs (CTR9 g1 and g2). (F) Percentage of total human CD45+ cells after RBC depletion from peripheral blood of NBSGW mice xenotransplanted with AAVS1, partial or complete CTR9 edited, partial or

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

complete PAF1 edited, LEO1 or CDC73 edited cord blood $CD34^+$ HSPCs at indicated time of post-transplantation. The data points for AAVS1, partial or complete CTR9 edited groups are the same as Figure 4A because they are from the experiments done at the same time. (G) Percentage of total human CD45+ cells after RBC depletion from the bone marrow of NBSGW mice xenotransplanted with AAVS1, partial or complete CTR9 edited, partial or complete PAF1 edited, LEO1 or CDC73 edited cord blood $CD34^+$ HSPCs 20 weeks post-transplantation. The data points for AAVS1, partial or complete CTR9 edited groups are the same as Figure 4B because they are from the experiments done at the same time. (H) Percentage of total human $CD45^+$ cells after RBC depletion from the spleen of NBSGW mice xenotransplanted with AAVS1, partial or complete CTR9 edited, partial or complete PAF1 edited, LEO1 or CDC73 edited cord blood $CD34^+$ HSPCs 20 weeks post-transplantation. The data points for AAVS1, partial or complete CTR9 edited groups are the same as Figure 4C because they are from the experiments done at the same time. (I) Editing efficiency analysis of cord blood $CD34^+$ HSPCs before transplantation and peripheral blood cells post-transplantation at indicated time point of post-transplantation for NBSGW mice transplanted with partial or complete PAF1 edited, CDC73 or LEO1 edited $CD34^+$ HSPCs. Lines indicate samples from $CD34^+$ HSPCs and the corresponding transplanted mouse. (J) Estimated required HSC numbers for the observed reconstitution dynamics for AAVS1, partial or complete CTR9 edited, partial or complete PAF1 edited, LEO1 or CDC73 edited cord blood $CD34^+$ HSPCs in (A). The data points for AAVS1, partial or complete CTR9 edited groups are the same as Figure 4F because they are from the experiments done at the same time. All data are presented as mean ± standard deviations (*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$, ****$p < 0.0001$).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

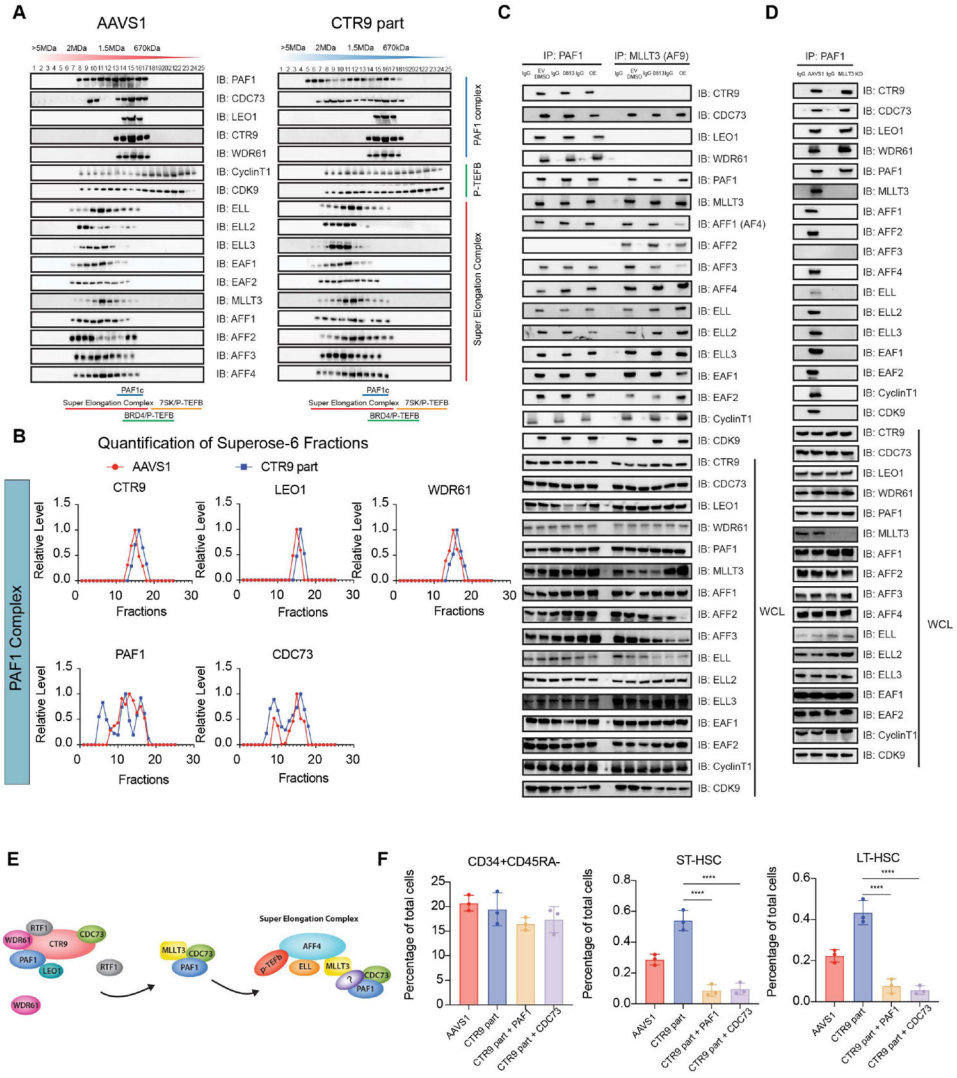


Figure 6. Partial loss of CTR9 promotes PAF1-CDC73-SEC interaction dependent on MLLT3.

(A) Gel filtration chromatography of AAVS1 or partial CTR9 edited CD34^{+} HSPCs. (B) Quantification of band intensity of PAF1 complex subunits across all fractions. (C) Co-immunoprecipitation in CD34^{+} HSPCs by PAF1 and MLLT3 (AF9) antibodies with YEATS inhibitor SR-0813 or MLLT3 overexpression. (D) Co-immunoprecipitation of MLLT3 KO only (CTR9 wildtype) CD34^{+} HSPCs by the PAF1 antibody. (E) Schematic diagram showing the increased PAF1-CDC73-SEC interaction upon partial loss of CTR9. (F) Quantification of CD34^{+}CD45RA^{-}, ST-HSC, and LT-HSC populations of partial CTR9 edited cells rescued by concomitant editing of PAF1 (g1) or CDC73 (g1) guides.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

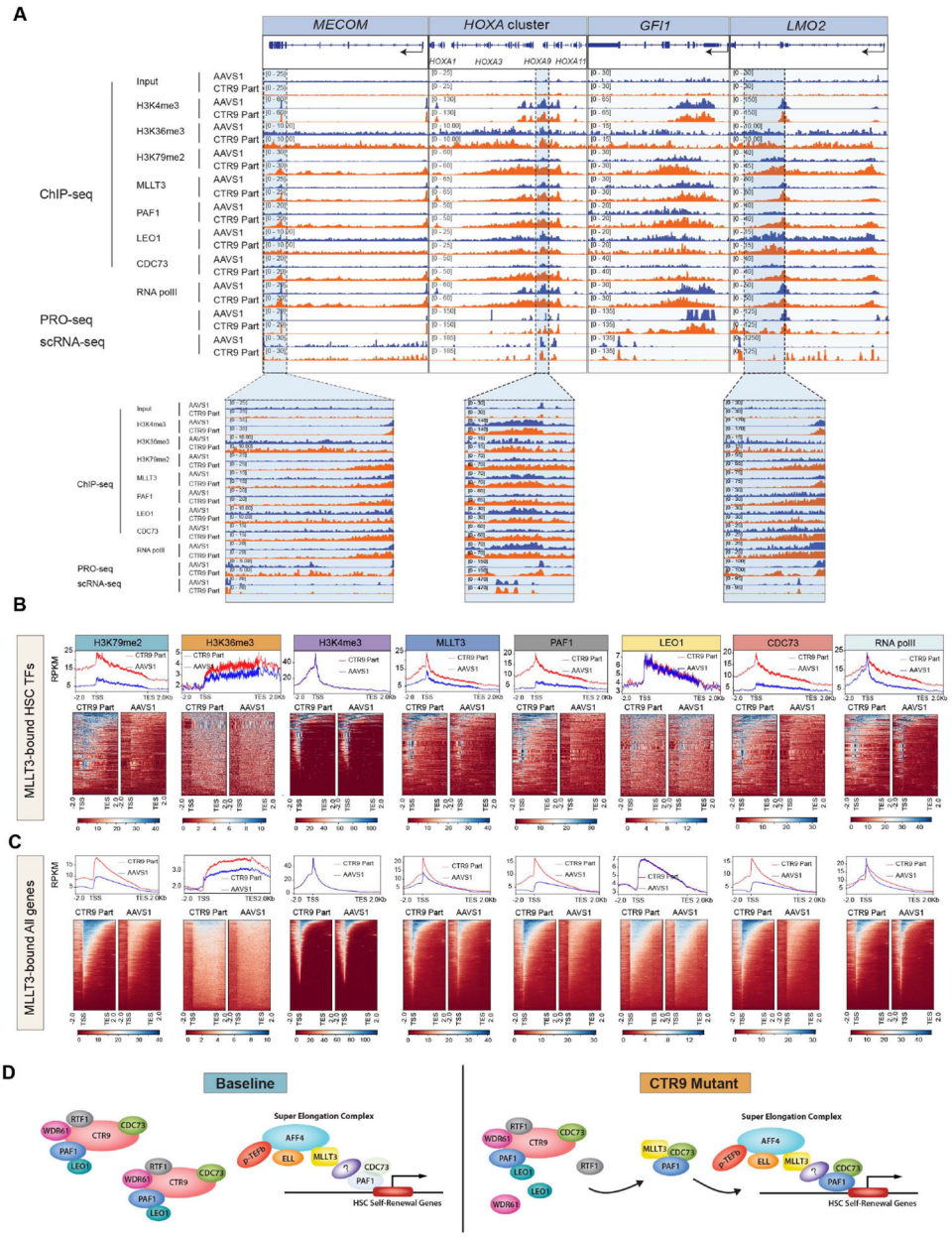


Figure 7. Augmented SEC activity is due to relocalization of PAF1 complex subunits.

(A) Representative BigWig tracks of input, H3K4me3, H3K36me3, H3K79me2, MLLT3, PAF1, LEO1, CDC73 and RNA polymerase II ChIP-seq, PRO-seq and scRNA-seq data of AAVS1 or partial CTR9 edited (g1) $CD34^{+}CD45RA^{-}CD90^{+}$ cells at *MECOM*, *HOXA* cluster, *GFI1*, and *LMO2* loci. Zoomed tracks correspond to the gene body region of *MECOM*, *HOXA9* and *LMO2*, respectively. (B) ChIP-seq metaplots and heatmap plots for H3K4me3, H3K36me3, H3K79me2, MLLT3, PAF1, LEO1, CDC73 and RNA polymerase II ChIP-seq data of AAVS1 or partial CTR9 edited $CD34^{+}CD45RA^{-}CD90^{+}$ cells at MLLT3 bound HSC transcription factor loci. (C) ChIP-seq metaplots and heatmap plots for H3K4me3, H3K36me3, H3K79me2, MLLT3, PAF1, LEO1, CDC73 and RNA polymerase II ChIP-seq data of AAVS1 or partial CTR9 edited $CD34^{+}CD45RA^{-}CD90^{+}$ cells at all MLLT3

bound gene loci. (D) Schematic of the mechanistic model by which partial loss of CTR9 results in HSC expansion.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Key resources table

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Anti-CTR9 antibody	Cell Signaling Technology	12619S
LEO1 Polyclonal antibody	ProteinTech	12281-1-AP
RTF1 Polyclonal antibody	ProteinTech	12170-1-AP
Rpb1 CTD Antibody Sampler Kit	Cell signaling Technology	54020
Brilliant Violet 421™ anti-human CD34 Antibody	BD Biosciences	562577
PE Mouse Anti-Human CD66b	BD biosciences	561650
PE Mouse Anti-Human CD33	BD biosciences	555450
PE Mouse Anti-Human CD19	BD biosciences	555413
CDC73 Monoclonal antibody	ProteinTech	66490-1-Ig
PAF1 antibody	Bethyl laboratories	A300-172A-T
AF9 Antibody	Bethyl laboratories	A300-596A-T
anti-CD90-PECy7	BD biosciences	561558
APC-H7 Mouse Anti-Human CD45RA	BD biosciences	560674
AF4 antibody	Bethyl Laboratories	A302-344A-T
WDR61 Polyclonal antibody 20uL	Proteintech	22536-1-AP
RNA pol II antibody (mAb)	active motif	61667
AF9 antibody ChIP grade	GeneTex	GTX102835
Anti-PAF1/PD2 antibody ChIP	Abcam	ab137519
HRPT2, CDC73 Polyclonal antibody 150ul	proteintech	12310-1-AP
Histone H3K36me3 antibody (pAb)	Active motif	61902
Anti-Histone H3 (di methyl K79) antibody - ChIP Grade	Abcam	ab3594
APC anti-human CD49c (integrin α3) Antibody	Biolegend	343808
Histone H3K4me3 antibody (mAb)	Active motif	61979
EAF1 Polyclonal antibody	Proteintech	13787-1-AP
EAF2 Polyclonal antibody	Proteintech	11172-1-AP
ELL Polyclonal antibody	Proteintech	51044-1-AP
ELL2 Polyclonal antibody	Proteintech	12727-1-AP
AFF4 Polyclonal antibody	Proteintech	14662-1-AP
CDK9 Polyclonal antibody	Proteintech	11705-1-AP
Cyclin T1 Polyclonal antibody	GeneTex	GTX133413
Polyclonal Rabbit anti- Human AFF2 / OX19 Antibody (WB) LS- C670175	LSbio	LS-C670175-50
Polyclonal Rabbit anti- Human LAF4 / AFF3 Antibody (WB) LS- C178457	LSbio	LS-C178457-100
ELL3 Polyclonal Antibody	Invitrogen	PA5-70583
BB700 Mouse Anti-Human CD133	BD biosciences	747638
rat anti-human anti-EPCR-PE	Biolegend	351904
PE anti-human CD34 Antibody	Biolegend	343506

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

REAGENT or RESOURCE	SOURCE	IDENTIFIER
PE anti-human CD3 Antibody	Biolegend	317308
PE anti-human CD11b Antibody	Biolegend	301306
Monoclonal ANTI-FLAG® M2 antibody produced in mouse	Sigma Aldrich	F1804
GAPDH Antibody	Santa Cruz	sc-32233
Chemicals, peptides, and recombinant proteins		
AZD4573	Selleck Chem	S8719
LDC000067	Selleck Chem	S7461
Pierce™ Protein A/G Magnetic Beads	Thermo Scientific	88802
Q5® High-Fidelity 2X Master Mix	NEB	M0492S
Methocult H4034 Optimum	Stem Cell Tech	4034
Normal Rabbit IgG	Cell signaling technology	2729S
Normal mouse IgG	Santa Cruz	sc-2025
USER® Enzyme	NEB	M5505S
SR-0813	Gift from Michael Erb	
Polyethylenimine, Linear, MW 25000, Transfection Grade (PEI 25K™)	PolySciences	23966-100
CC100	StemCell Technologies	2690
Methanol	Sigma Aldrich	179337
RPMI1640 medium	Life Technologies	11875-119
Opti-MEM	Gibco	31985-062
Ficoll-Paque Premium sterile solution	GE Healthcare	17-5442-02
0.25% Trypsin (1X)	Gibco	15050-057
Blotting nonfat dry milk	BioRad	1706404XTU
Cas9 protein, recombinant	IDT	1081061
Penicillin-Streptomycin 10,000u/ml	Life Technologies	15140-122
Recombinant Human TPO	Peprotech	300-18-100UG
Superose 6 Prep Grade	Cytiva	17048901
L-Glutamine 200mM for cell culture	GIBCO	25-030-081
DMEM High Glucose	Life Technologies	11965-118
Critical commercial assays		
Kapa Library Quantification Kit - Complete Kit (Universal)	Roche	7960140001
Deposited data		
Single cell RNA seq	This paper	GSE186591
Pro-seq	This paper	GSE186591
ChIP-seq	This paper	GSE186591
Experimental models: Cell lines		
CD34+ HSPCs	Fred Hutch	
HEK293T	ATCC	CRL-3216
Experimental models: Organisms/strains		
NBSGW	JAX Laboratory	026622

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Stellar Competent cells	Takara/Clontech	636763
One Shot™ Stbl3™ Chemically Competent E. coli	Invitrogen	C737303
Oligonucleotides		
Alt-R® Cas9 Electroporation Enhancer	IDT	1075916
qPCR primers	Genewiz	Table S3
Chemically synthesized sgRNA	Synthego	Table S4
Recombinant DNA		
MGC Fully Sequenced Human MLLT3 cDNA	Horizon Discovery	MHS6278-202809020
MGC Human CTR9 Sequence-Verified cDNA	Horizon Discovery	MHS6278-202760315
Software and algorithms		
Graphpad Prism v.9.0.1	Graphpad	RRID:SCR_002798
Biorender	Biorender	RRID:SCR_018361
Flowjo	Flowjo	RRID:SCR_008520
ImageJ v1.52a	Wayne Rasband	RRID:SCR_003070
R	The R Foundation	
STAR v2.7.10a		
Bowtie2 v2.3.3		
MACS v2.2.9		
IGV v2.11.9		
deepTools v3.5.4		
Cutadapt v4.1		
CellRanger v7.0.0	10xGenomics	
Seurat v4.0.2		
Python 3.10		
Other		
In-Fusion® HD Cloning Plus	Takara/Clontech	638909
Q5 site directed mutagenesis kit	New England Biolabs	E0554S
iScript™ cDNA Synthesis Kit	BioRad	1708891
Monarch® DNA Gel Extraction Kit	New England Biolabs	T1020L
iQ™ SYBR® Green Supermix	BioRad	1708882
P3 Primary Cell 4D Nucelofector X Kit S	Lonza	V4XP-3032
SuperScript™ III Reverse Transcriptase	Invitrogen	18080044
Chromium Next GEM Single Cell 3’ Kit v3.1, 16 rxns	10XGenomics	PN-1000268
Chromium Next GEM Chip G Single Cell Kit, 48 rxns	10XGenomics	PN-1000120
Dual Index Kit TT Set A, 96 rxns	10XGenomics	PN-1000215
Lipofectamine™ 3000 Transfection Reagent	Thermo Scientific	L3000015
ChIP-IT® Express Enzymatic 25rxn	active motif	53009
StemSpan™ SFEM II	StemCell Technologies	09655
Fetal Bovine Serum - Premium Select G21162	Bio-Techne	S11550

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

REAGENT or RESOURCE	SOURCE	IDENTIFIER
SUPERase•In™ RNase Inhibitor (20 U/μL)	Invitrogen	AM2694
EasySep Human Cord Blood CD34 Positive Selection Kit II	StemCell Technologies	17896
Immobilon-P PVDF Membrane	Millipore Sigma	IPVH00010
SPRIselect	Beckman Coulter	B23317
Qubit™ 1X dsDNA HS Assay Kit	ThermoFisher	Q33231
Zero Blunt™ TOPO™ PCR Cloning Kit for Sequencing, without competent cells 25rxn	Invitrogen	450031
NEBNext® Ultra™ II DNA Library Prep with Sample Purification Beads	NEB	E7103S
4–20% Criterion™ TGX™ Precast Midi Protein Gel, 26 well, 15 μl #5671095	Biorad	5671095