nature genetics

Supplementary information

https://doi.org/10.1038/s41588-024-01754-2

Cell-type-specific consequences of mosaic structural variants in hematopoietic stem and progenitor cells

In the format provided by the authors and unedited

Supplementary Figures

Supplementary Methods

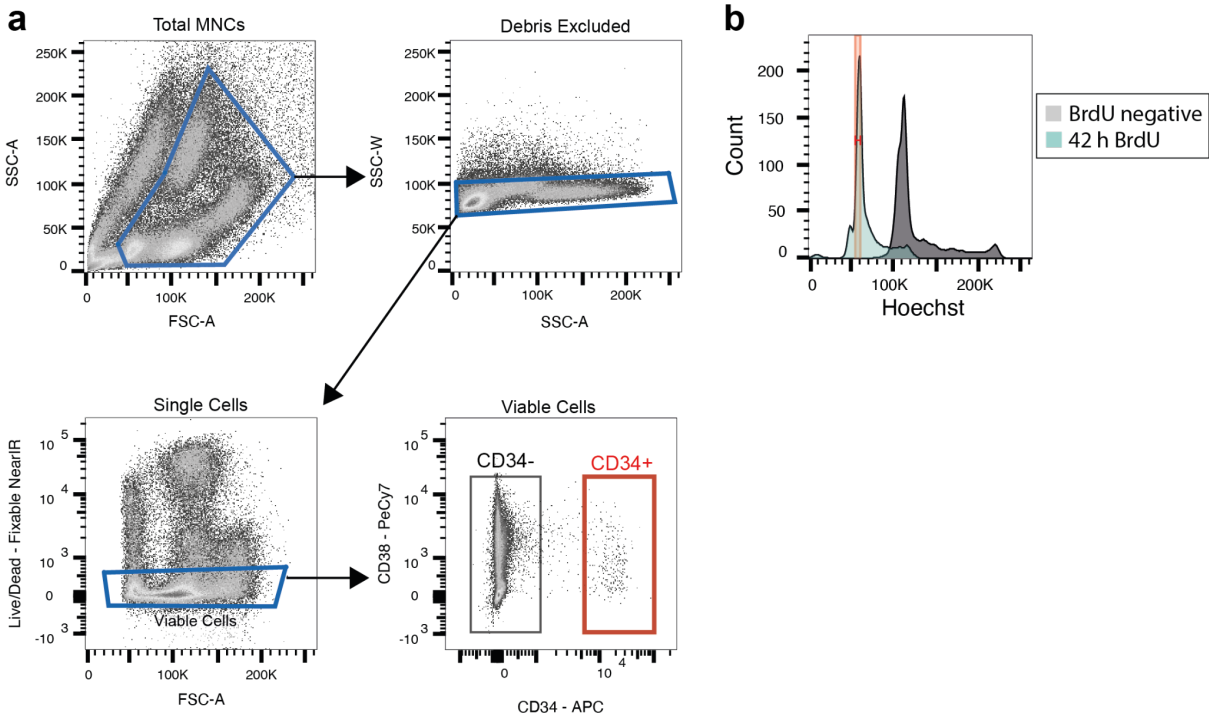
1. Identification of active X chromosome from the female genome
2. Protein-protein interaction (PPI) network analysis using STRING
3. Similarity analysis for over-represented pathways of dysregulated genes in mSV subclones

Supplementary Notes

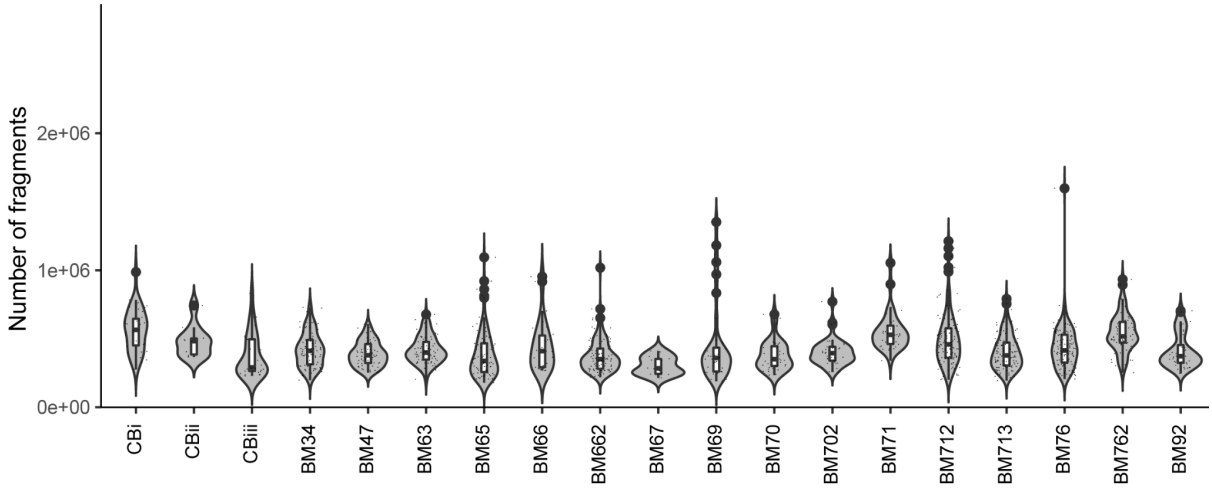
1. Selecting an optimal timepoint for BrdU incorporation in cultured human HSPCs
2. Characteristics of de novo mSVs in HSPCs
3. Comparison with prior surveys of mosaic copy-number alterations and mSVs
4. Investigation of genes associated with local effect of subclonal inversion in BM65
5. Investigation of functional links between dysregulated TFs in the 17p-Del subclone in BM712
6. Potential small deletions at regions of recurrent SCE/mSV formation
7. Analysis of somatic SNVs from the IntoGen Clonal Hematopoiesis Mutation Browser
8. Analysis of somatic SNVs affecting the AR gene in the UK Biobank
9. Singleton CNA discovery in HSPCs in scWGS data
10. Analysis of data release from a CRISPR in vivo knockout (KO) screen
11. Interplay between mSVs and clonal hematopoiesis.

Supplementary References

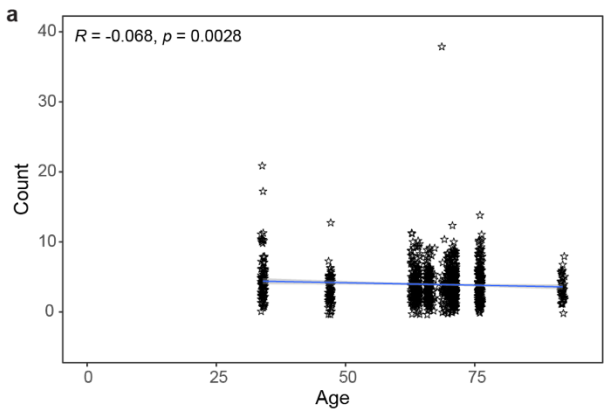
Supplementary Figures



Supplementary Figure 1: Sorting strategies used to generate Strand-seq libraries from HSPCs. **a)** Gating strategy for isolation of viable CD34+ cells from human bone marrow/umbilical cord blood mononuclear cells. **b)** Gating strategy for sorting of single, BrdU-containing nuclei from cultured HSPCs.

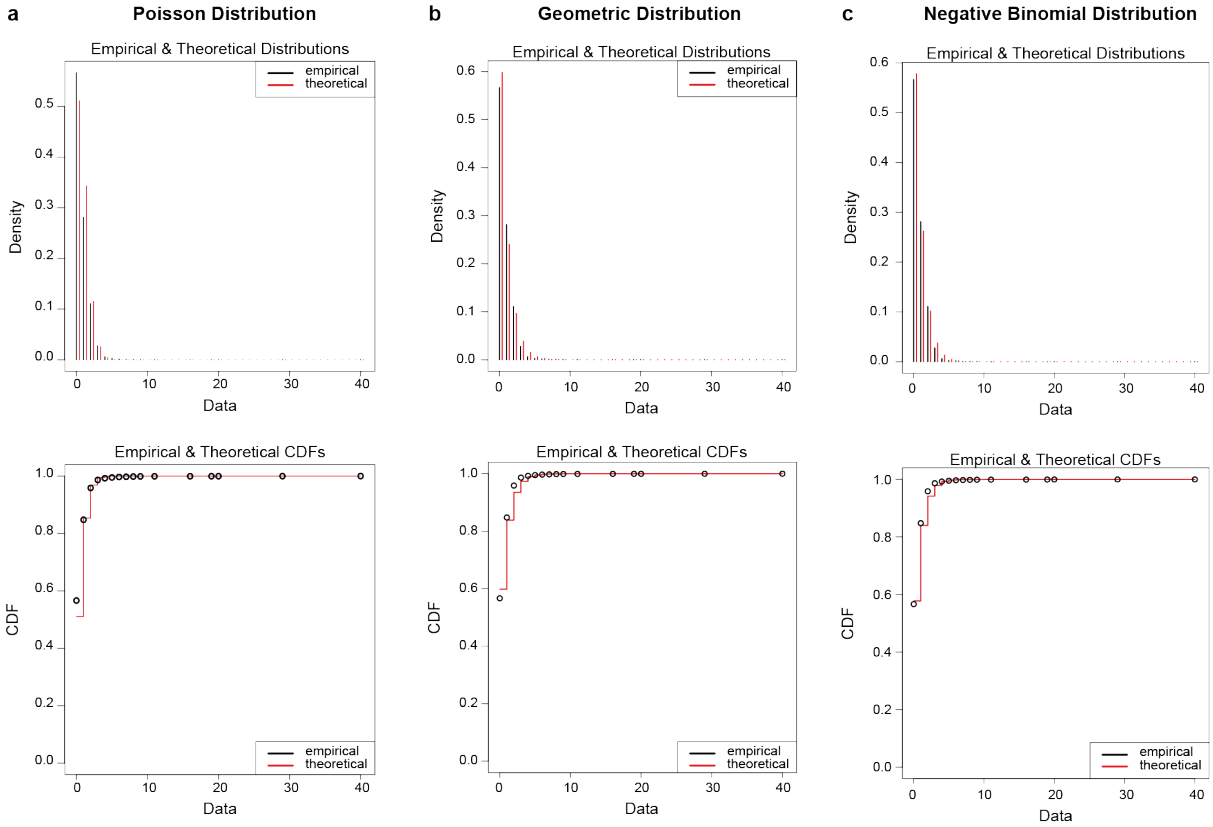


Supplementary Figure 2: Number of uniquely mapped fragments per Strand-seq library per donor across cohort.
The violin plots show the number of uniquely mapped fragments per cell profiled from 19 donors. Labels in the X-axis refer to sample names of samples, in which the tissue of origin (CB; cord blood, BM; bone marrow), and the age of each donor, is indicated. Boxplots were defined by minima = 25th percentile - 1.5X interquartile range (IQR), maxima = 75th percentile + 1.5X IQR, center = median, and bounds of box = 25th and 75th percentile.



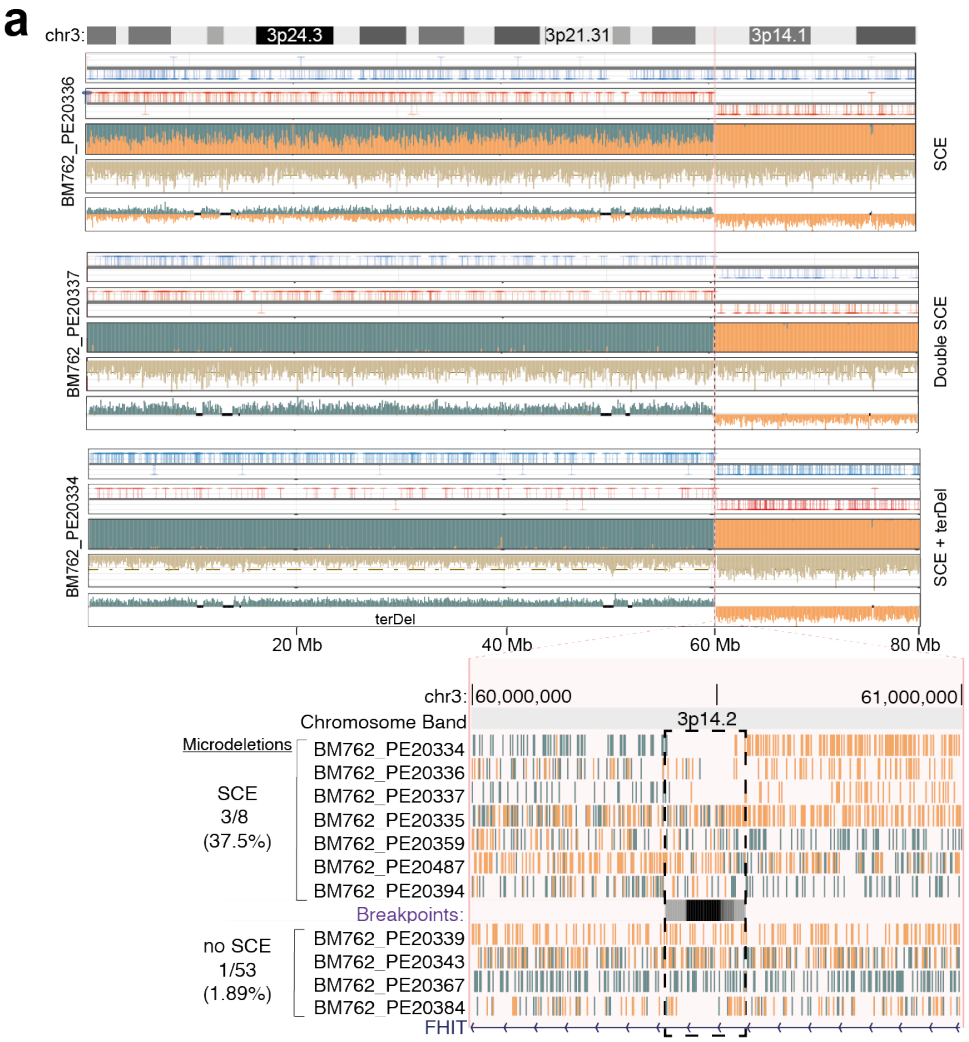
Supplementary Figure 3: Anticorrelation of SCE frequency with age is independent of tissue-of-origin.

a) Whether UCB samples are included (**Fig. 1g**) or excluded (**a**), there remains a weak anticorrelation between the age of a donor and the number of SCEs seen per cell (R=-0.068; *P*=0.0028 excluding UCB). *R*: correlation coefficient calculated from the x and y-axis; p-value (*p*) is based on the two-sided significance test for the Pearson correlation coefficient, testing the hypothesis that it is 0.

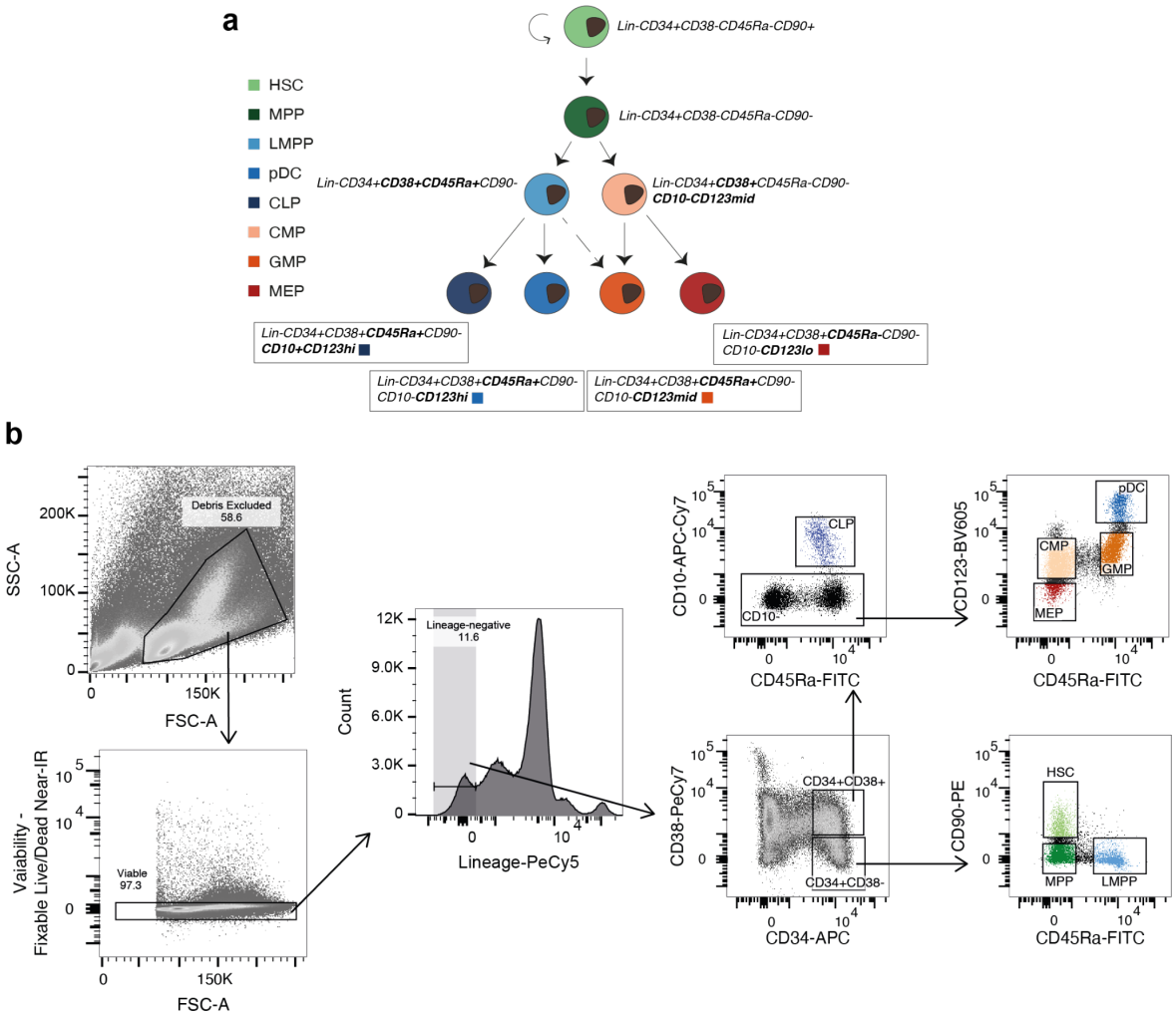


Supplementary Figure 4: Determining the appropriate distribution for permutation testing of SCE occurrence across the genome.

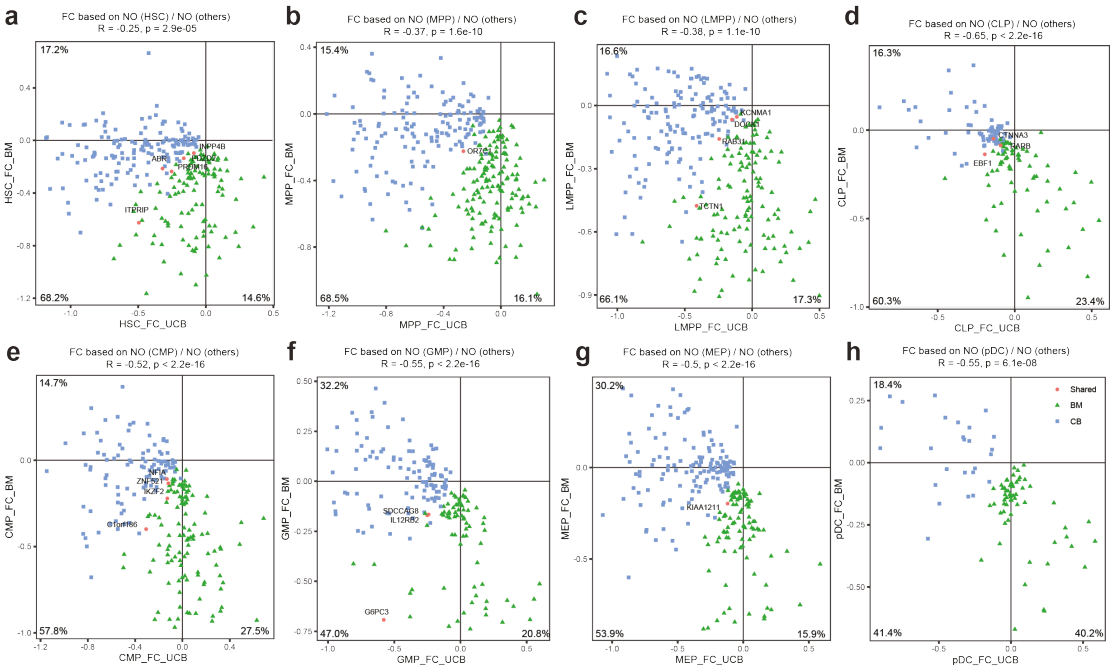
Evaluation of the distribution of empirical and theoretical data of SCE occurrence across the genome genome-wide (hg38, 500kb bins). The following distributions are tested **a**) Poisson distribution, **b**) Geometric distribution, **c**) Negative binomial distribution.



Supplementary Figure 5: *De novo* DNA rearrangement acquisition at SCE hotspots.
a) Upper: Strand-seq data showing recurrent SCE and mSV co-occurrence at the common fragile site (CFS) *FRA3B*, an SCE hotspot. Lower: Zoom in to the breakpoint region at *FRA3B*. Both cells with (top 8) and without (lower 4) SCEs at this CFS show potential deletions, evident at sub-200 kb resolution.

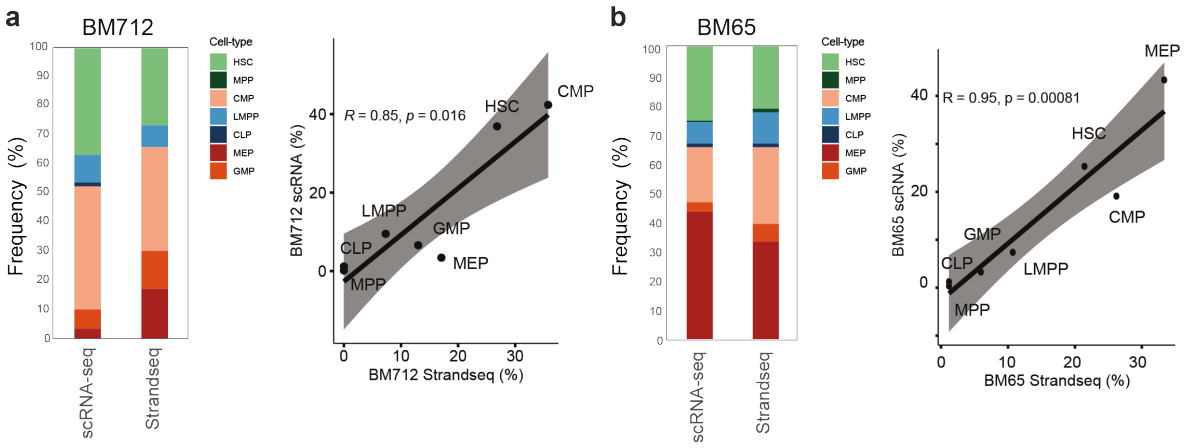


Supplementary Figure 6: Isolation of HSPCs for scMNase-seq reference dataset generation. **a**) Defining immunophenotypes for 8 HSPC cell types, as previously described in [1]. **b**) FACS gating strategy for isolating each of the cell types mentioned in **a**).



Supplementary Figure 7: Comparison of HSPCs from UCB and BM samples.

a-h) Scatter plots comparing UCB (x-axis) and BM (y-axis) in terms of fold change of nucleosome occupancy (NO) by cell-type (HSC, MPP, LMPP, CLP, pDC, CMP, GMP, and MEP). Each dot represents the cell-type classifier genes selected from UCB (blue), BM (green), or both compartments (pink). The Names of classifier genes shared by both compartments (pink dots) are shown in the scatter plots. *PRDM16* shows significantly decreased NO in HSCs from both BM and UCB, in agreement with its previously determined role in HSC generation and maintenance[2]. However, many genes show distinct NO profiles in BM vs UCB, with only 21 genes appearing in both BM- and UCB-derived NO classifiers. *R* : correlation coefficient calculated from the x and y-axis; p-value (*P*) is based on the two-sided significance test for the Pearson correlation coefficient, testing the hypothesis that it is 0.



Supplementary Figure 8: Verification of NO-based cell-type BM classifier using scRNA-seq. **a)** A stacked bar graph (left) depicts the HSPC cell-type composition in BM712 estimated through SingleR cell-type annotations[3] in scRNA-seq data in comparison to NO-based cell-typing of Strand-seq data. The scatter plot (right) shows that cell-type compositions in this donor sample are highly correlated between Strand-seq (X-axis) and scRNA-seq (Y-axis). **b)** Similarly, we find that in BM65 cell-type compositions estimated based on Strand-seq and scRNA-seq are highly correlated. The error band indicates the confidence interval controlling 95% confidence region. *R*: correlation coefficient calculated from the x and y-axis; p-value (*P*) is based on the two-sided significance test for the Pearson correlation coefficient, testing the hypothesis that it is 0.

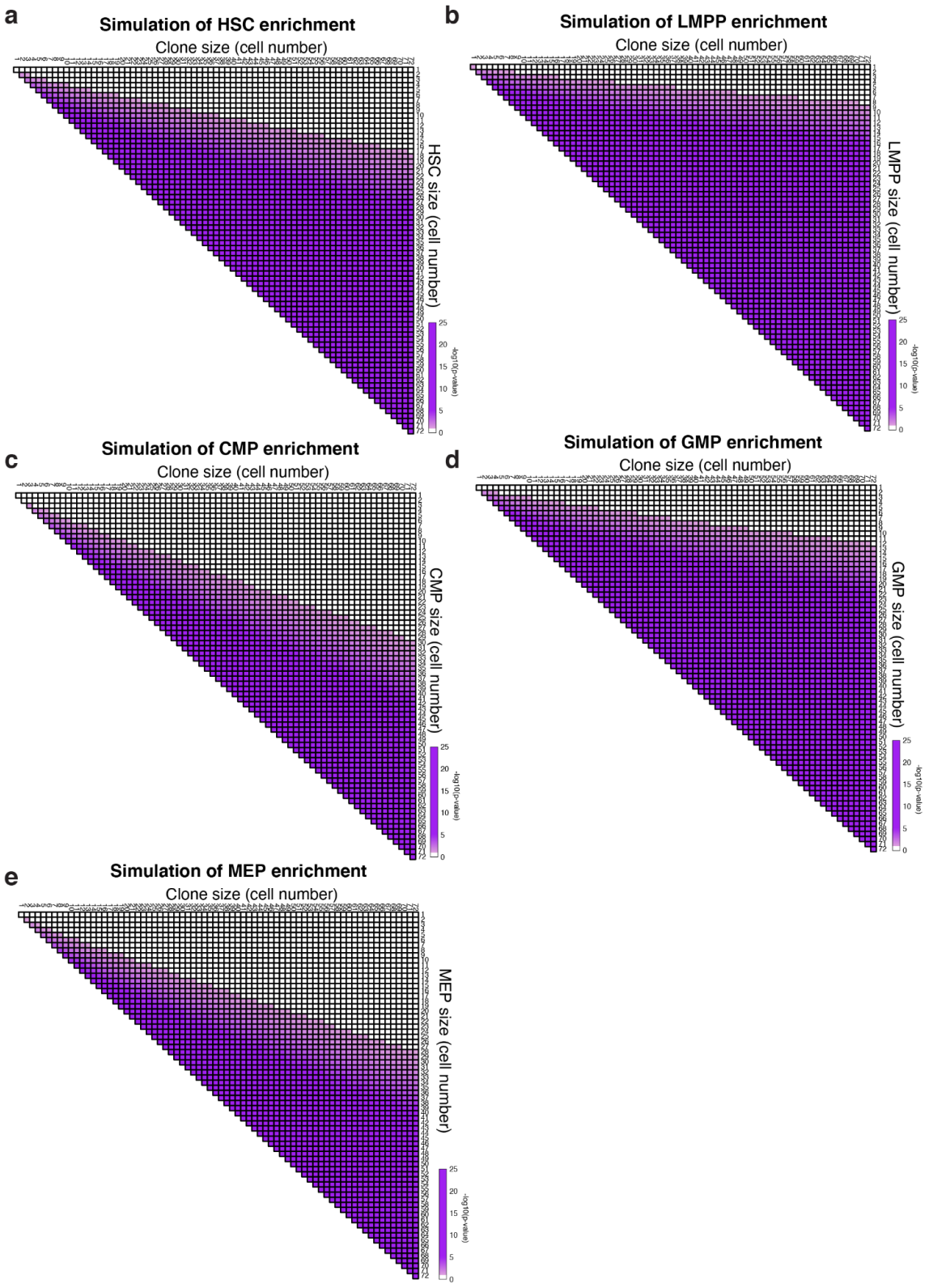


Figure S9: Simulation analysis of cell-type bias for different subclone sizes.
Each pyramid plots show the number of cells in each cell-type (Y-axis) required to statistically show the cell-type bias for given size of mSV clones (X-axis). This simulation analysis was performed for the mSV clone size range between 1 to 72 for five different cell-types we used for the cell-type enrichment analysis in **Fig. 3c-e**.

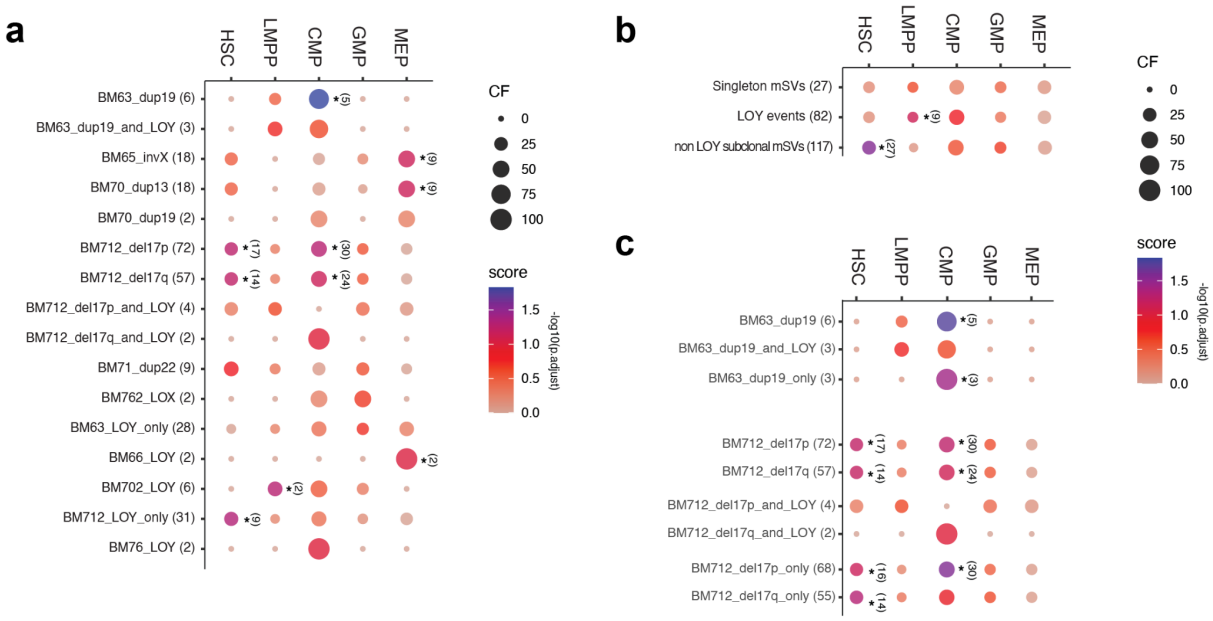
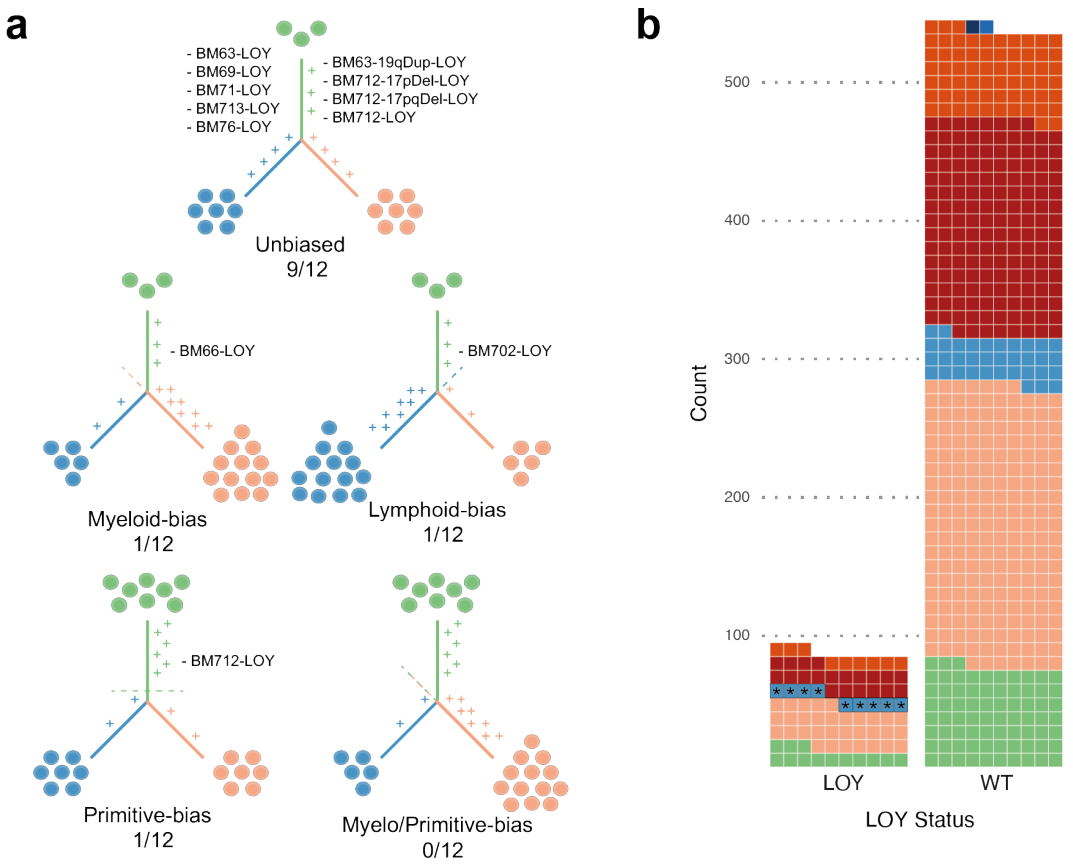
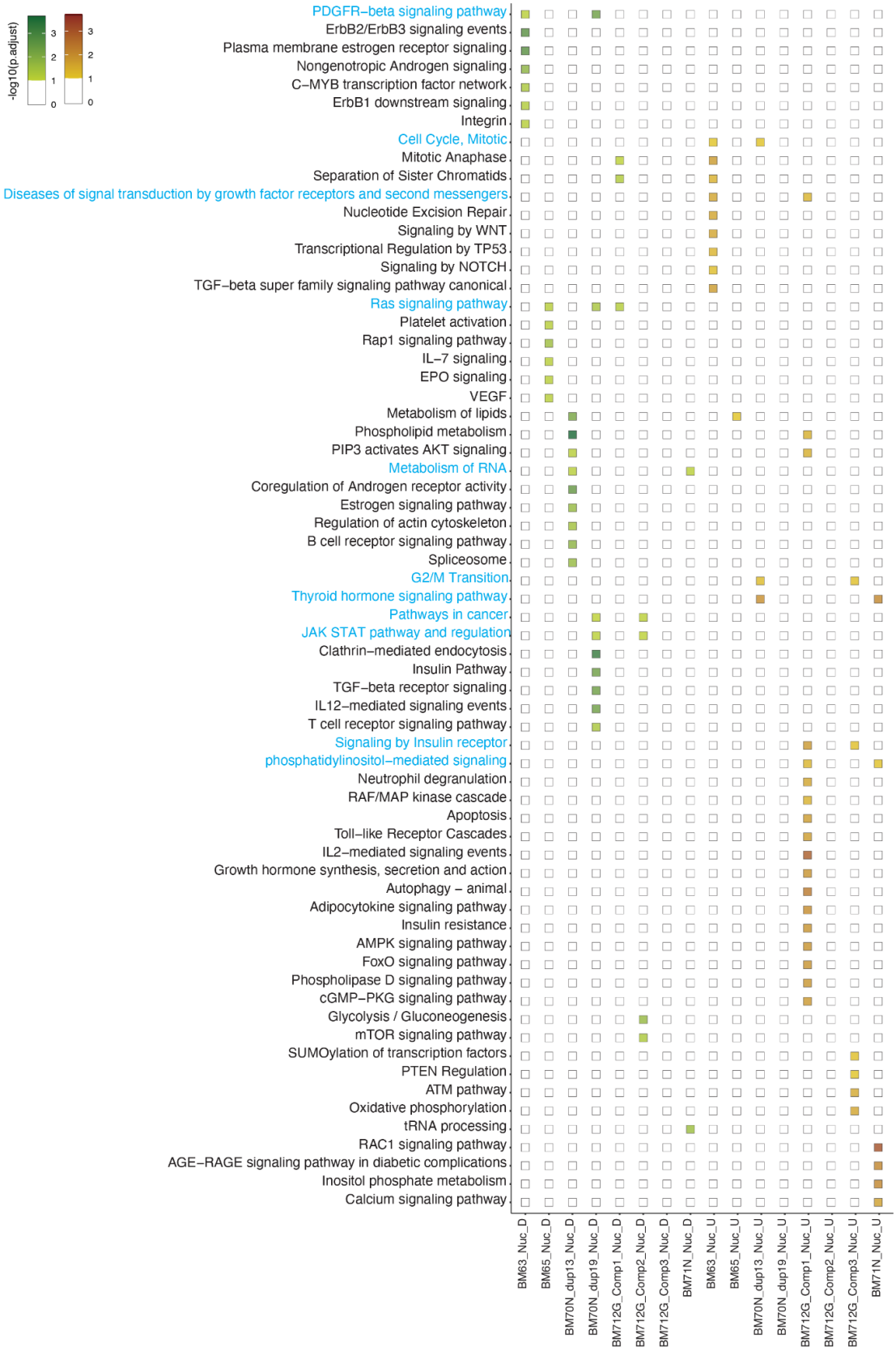


Figure S10: Cell type enrichment across all unique genotypes at the a) single donor/single genotype and b) cross-donor/cross-genotype levels. a) Extended dotplot of results of the cell-type enrichment analysis for each mSVs identified, showing the CF, enrichment and significance of enrichment in cell type per mSV sub-clone vs. an idealised control. This analysis was extended from the result in **Fig. 3c** to investigate the effect of whole chromosome losses (LOY, LOX), and the effect of newly arisen mosaicisms within subclones harboring mSVs. For instance, BM63_dup19_and_LOY denotes the LOY subclone originated from the bigger subclone harboring a duplication on chromosome 19. **b)** Dotplot of combined enrichments for the cells has singletons ('Singleton mSVs'), has LOYs ('LOY events'), or harboring subclonal mSVs that are not LOYs ('non LOY subclonal mSV'). **c**) Comparison of cell-type enrichment between subclones acquired secondary mSVs, and the bigger subclones they originated from. Significant scores of cell-type enrichment (-log10 p.adjust) were visualized as color gradient of dot plots which were calculated by the permutation adjusted *P-values* of binomial test (**Methods**).

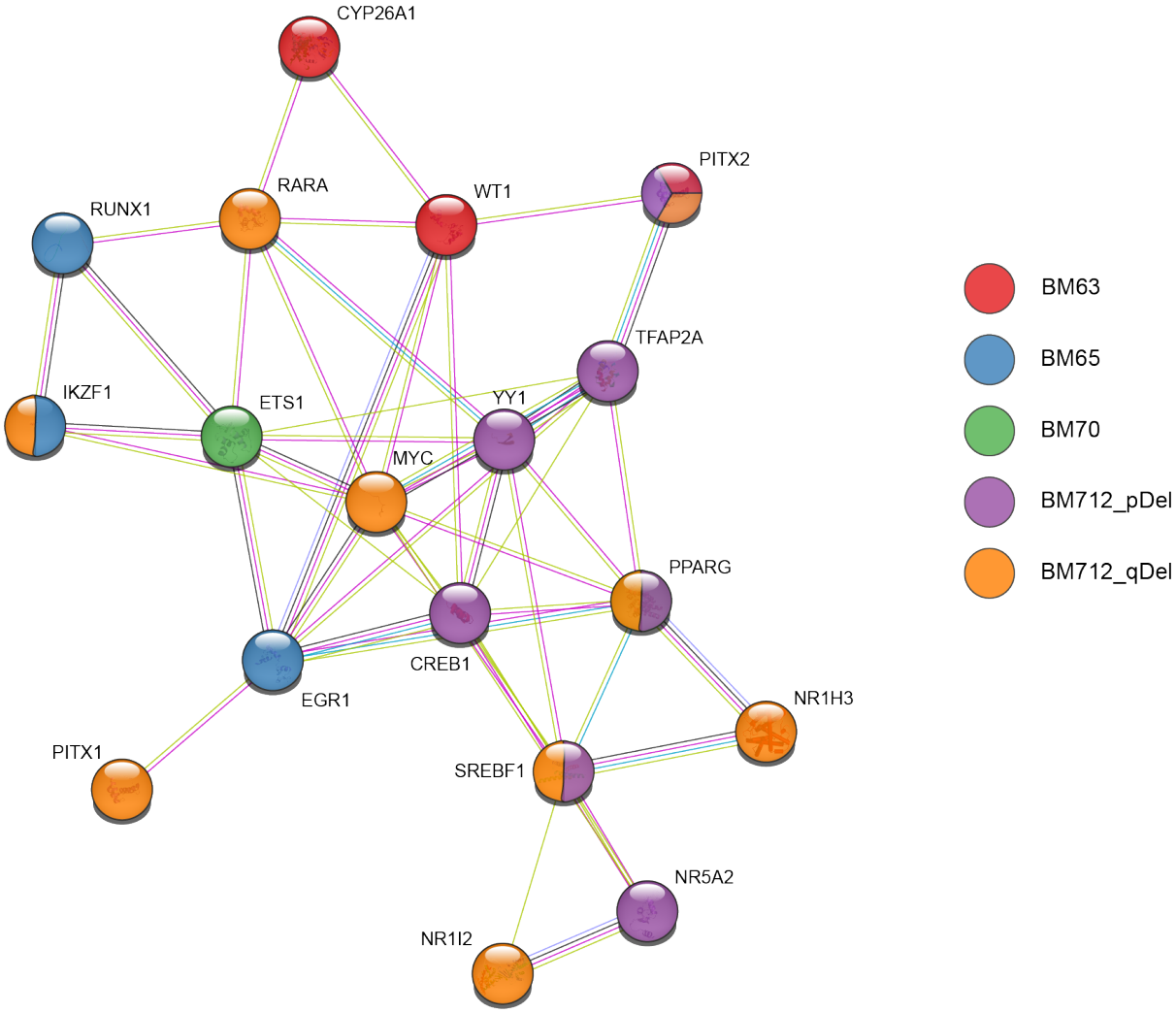


Supplementary Figure 11: Lineage biases observed for LOY.

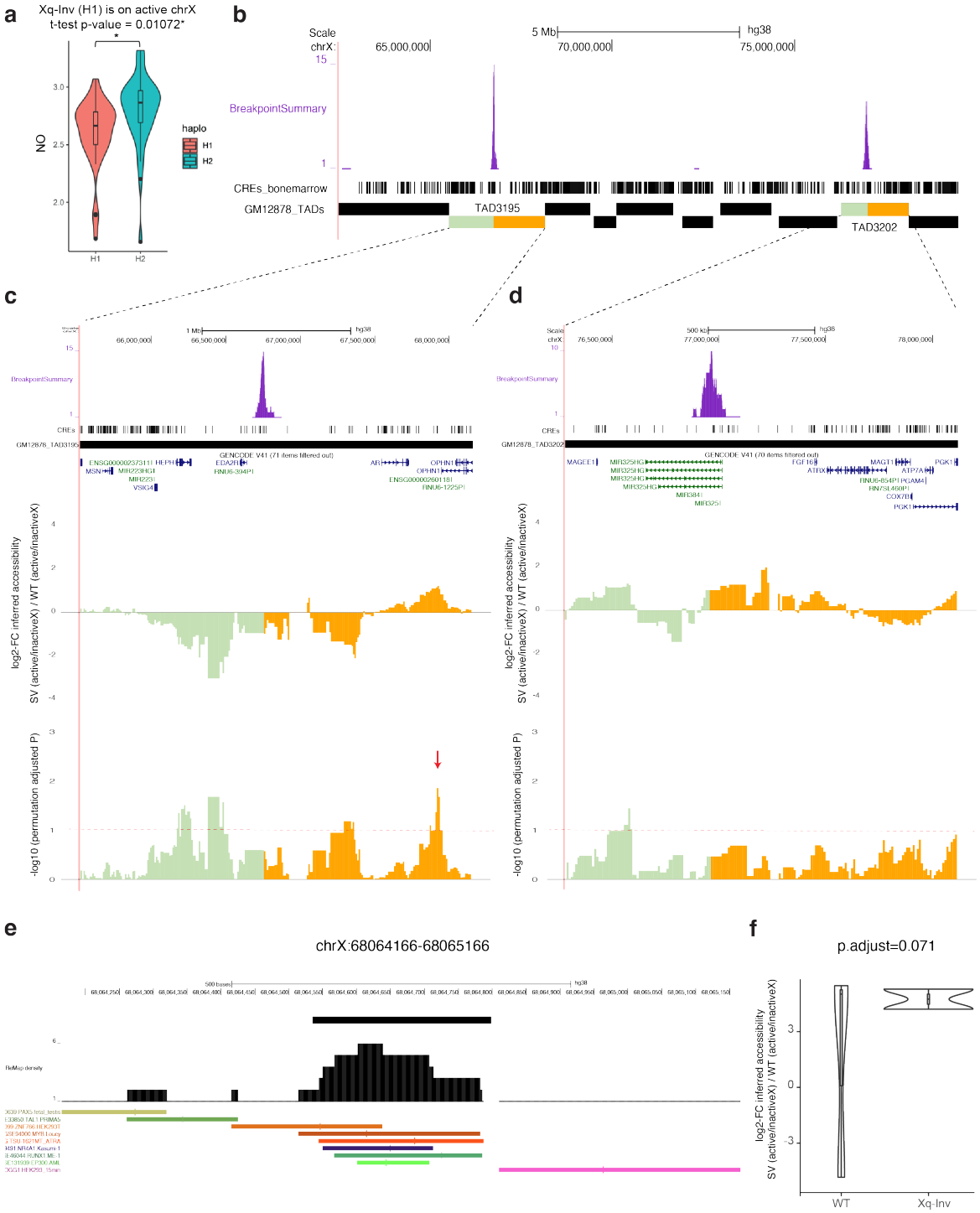
a) Lineage-level biases observed per donor for LOY subclones. **b**) Cohort-level bias observed for LOY. In the waffle plot, each small square represents a single cell with its cell-type identity inferred from Strand-seq (scNOVA), grouped by the presence of LOY. When comparing the cell-type composition of cells harboring LOY and an idealized control population of HSPCs from male donors, we observe a significant enrichment of LOY in the LMPPs (10% FDR; the starred boxes depict the significant enrichment of this cell type in LOY cells).



Supplementary Figure 12: Over-represented pathways of dysregulated genes in mSV subclones. Plot of pathway enrichment analysis for subclonal mSVs across the cohort. Pathways showing enrichment in at least one mSV are shown here. Pathways enriched in 2 or more mSVs in the same direction are highlighted in blue font (those pathways are also depicted in **Fig. 3f**).

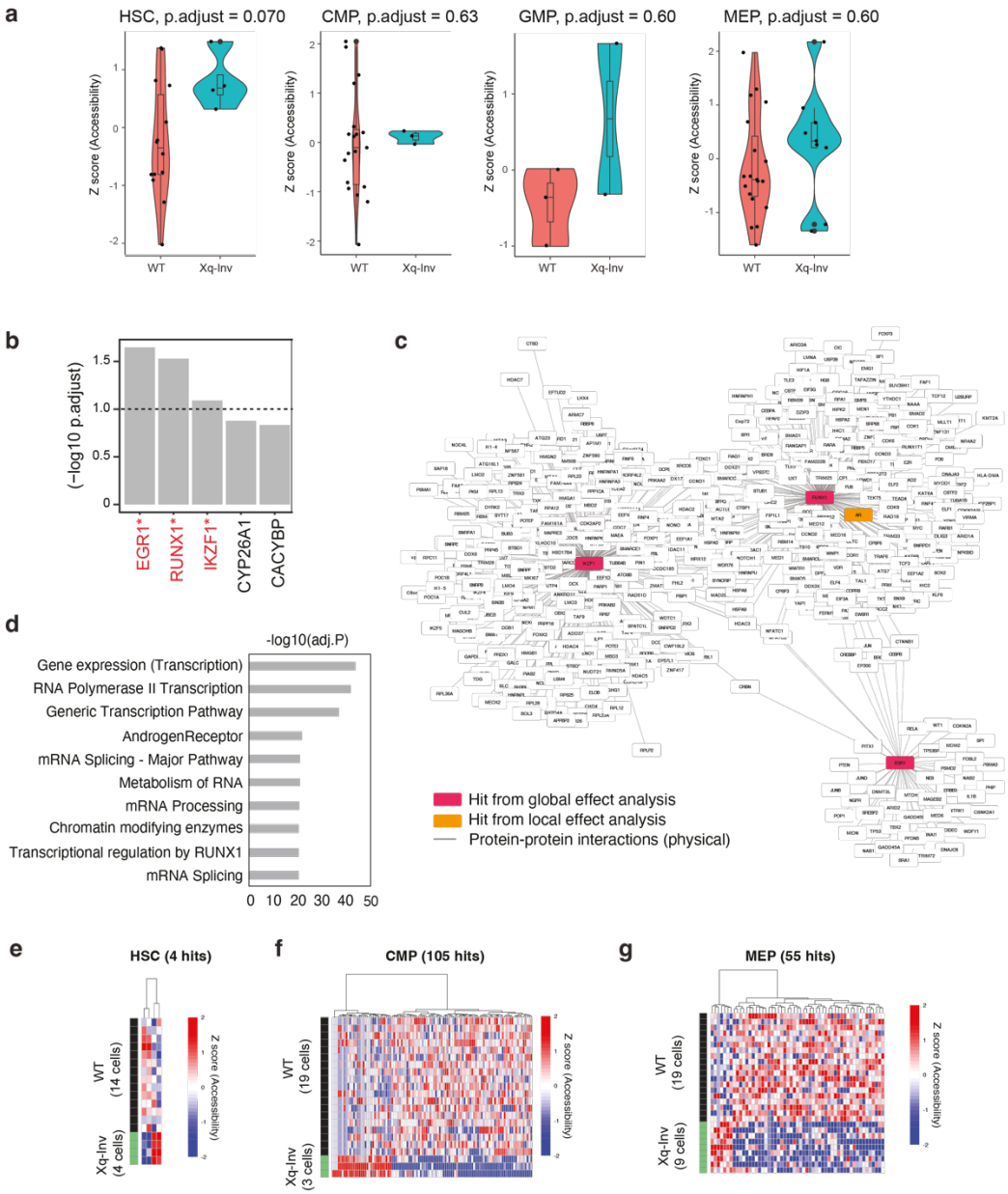


Supplementary Figure 13: Protein-protein interaction network identified from differentially active TFs across all mSV subclones.
STRING network of all differentially active TFs identified in samples exhibiting subclonal mSVs. TFs are coloured by sample and mSV. For significantly enriched gene-sets, see **Supplementary Table 20**.



Supplementary Figure 14: Investigation of *cis* effects of the mosaic inversion in BM65.
a) The inversion detected in donor BM65 is mapped to homolog 1 (H1) of chromosome X, using the Strand-seq data generated in this sample. We first checked whether the inversion-affected homolog represents the epigenetically-active or -inactive X chromosome, utilising the haplotype-resolved NO profiles. The violin plot shows the average NO for H1, n = 43 cells and haplotype 2 (H2, n = 43 cells) in each single-cell. Using a t-test to compare the two homologs indicates that the inversion harboring homolog (H1) represents the active X chromosome (P<0.01, two-sided t-test). Boxplots were defined by minima = 25th percentile - 1.5X interquartile range (IQR), maxima = 75th percentile + 1.5X IQR, center = median, and bounds of box = 25th and 75th percentile. **b**) Browser track snapshot showing that two TADs (TAD3195, TAD3202) are disrupted as a result of the inversion. **c-d**) Genome browser

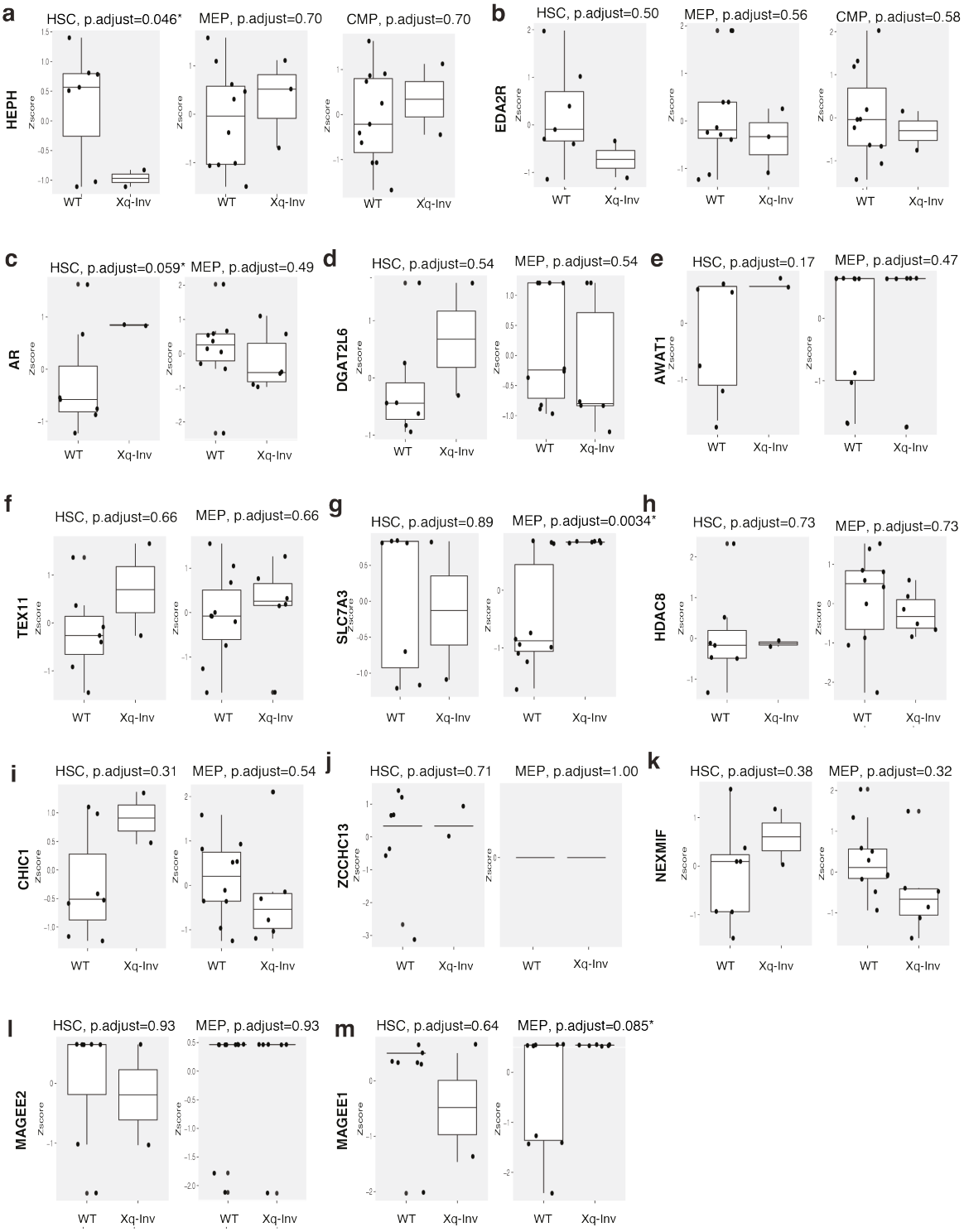
tracks showing log2-fold changes in haplotype-resolved NO (indicative of changes in chromatin accessibility[4]) between the mSV clone and WT cells on the active X chromosome, with permutation adjusted *P*-values (shown for TAD3195 (**c**) and TAD3202 (**d**), respectively). Red arrow indicates the sliding window showing the most significant difference between mSV clone and WT cells. **e**) *Cis*-regulatory element located in the most significantly different sliding window highlighted in panel (**c**). NO of this individual CRE was significantly lower in the mSV subclone compared to WT cells (P_{adj} < 0.071), indicating increased chromatin accessibility. TF motif sequences located in this region are depicted below. **f**) Violin plot showing the single-cell level distribution of log2-fold changes in inferred chromatin accessibility[4] between the mSV subclone and WT cells. *P.adjust value* was calculated from the two-sided Exact test followed by Benjamini Hochberg multiple correction.



Supplementary Figure 15: Identification of global effects of Xq-Inv in BM65 using TF activity analysis.

a) Z-scores of the inferred activity of AR target genes, based on scNOVA, were compared for each cell-type separately (n = 14, and 4 for WT and Xq-Inv for HSC; n = 19, and 3 for WT and Xq-Inv for CMP; n = 3, and 2 for WT and Xq-Inv for GMP; n = 19, and 9 for WT and Xq-Inv for MEP).

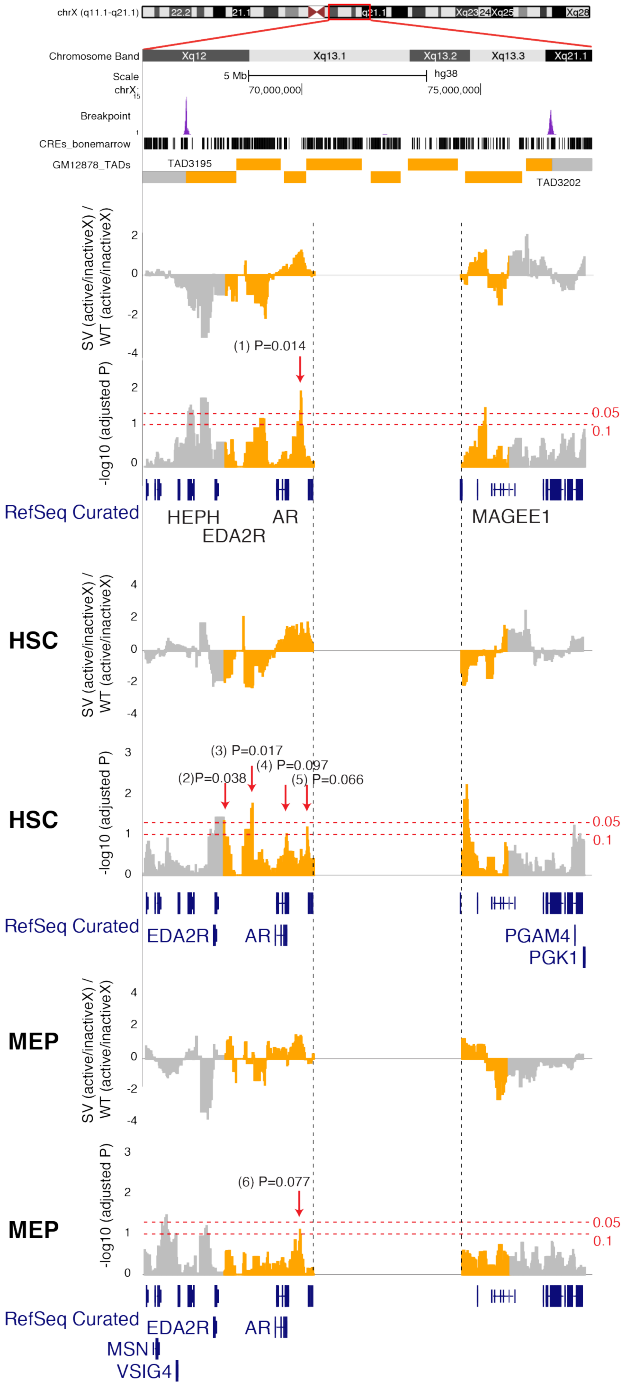
Boxplots were defined by minima = 25th percentile - 1.5X interquartile range (IQR), maxima = 75th percentile + 1.5X IQR, center = median, and bounds of box = 25th and 75th percentile. *P.adjust* values are based on the two-sided likelihood ratio test followed by Benjamini Hochberg multiple correction. (**b**) TFs identified from the TF-target over-representation analysis[4–6] of deregulated genes in the inversion subclone. This analysis reveals 3 TFs with differential activity in Xq-Inv cells: EGR1, RUNX1, and IKZF1 – all of which are linked to AR signaling (FDR 10% based on the hypergeometric test followed by Benjamini Hochberg multiple correction)[7]. (**c**) Protein-protein interaction network representing the first interactors of all three significant TFs identified in (**b**), using an FDR of 10%. Protein-protein interactions were obtained from the NCBI gene resource (https://www.ncbi.nlm.nih.gov/gene/). The 3 TFs identified in (**b**) are highlighted in red. The AR protein, which in this network is a direct interactor of *RUNX1,* is highlighted in orange. (**d)** Pathways over-represented by the genes involved in the network in (**c**) based on the ConsensusPathDB[8]. (**e-g**) Dysregulated genes in the inversion subclone identified by scNOVA in a cell-type specific manner, for HSCs (**e**), CMPs (**f**), and MEPs (**g**). We find that 3/4 differential NO genes of HSCs are AR targets[9], while 23/105 and 12/55 differential NO genes of CMPs and MEPs are AR targets[9], respectively.



Supplementary Figure 16: Inferred haplotype-specific gene activity of genes located within the inverted region and affected TAD boundaries.

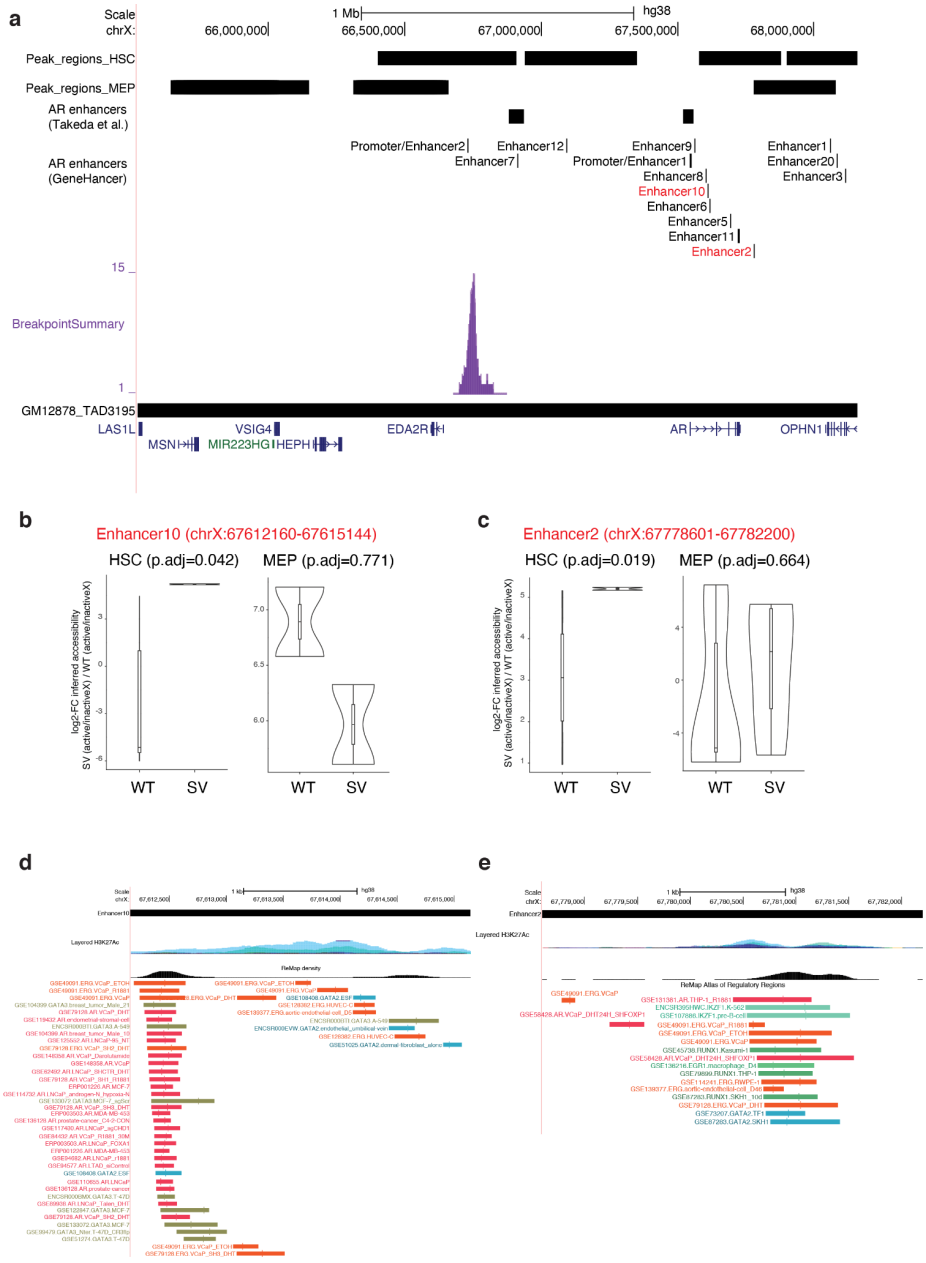
In total 13 genes are found in the TADs affected by inversion breakpoints or within the inverted segments (**Fig. 4b**). For those 13 genes, we compared the NO at gene bodies for the mSV clone and the WT cells in the active X homolog in a cell-type specific manner (**a-m**). Z-scores in the Y-axis indicate gene body-based inferred gene activities, inferred by scNOVA. This analysis requires at least two cells with WC configuration to resolve the haplotype into the active and inactive X homolog. Significance testing was achieved using two-sided t-tests. For the genes that reside in the TADs affected by the inversion breakpoints, but outside of the inverted region, we performed testing in

HSCs, CMPs, and MEPs. For genes residing within the inverted segment, we performed testing in HSCs and MEPs (as the segregation pattern for the inverted segment and the rest of the chromosome is different, explaining differences in our capacity to test for different cell-types). Cell-type and P_{adj} of genes showing significant difference of NO are indicated with asterisks (FDR 10%) (For HEPH and EDA2R in the inverted loci, n = 7, and 2 for WT and Xq-Inv for HSC; n = 11, and 2 for WT and Xq-Inv for CMP; n = 10, and 3 for WT and Xq-Inv for MEP. For the other genes in the rest of the loci, n = 7, and 2 for WT and Xq-Inv for HSC; n = 10, and 6 for WT and Xq-Inv for MEP). Boxplots were defined by minima = 25th percentile - 1.5X interquartile range (IQR), maxima = 75th percentile + 1.5X IQR, center = median, and bounds of box = 25th and 75th percentile throughout this figure.



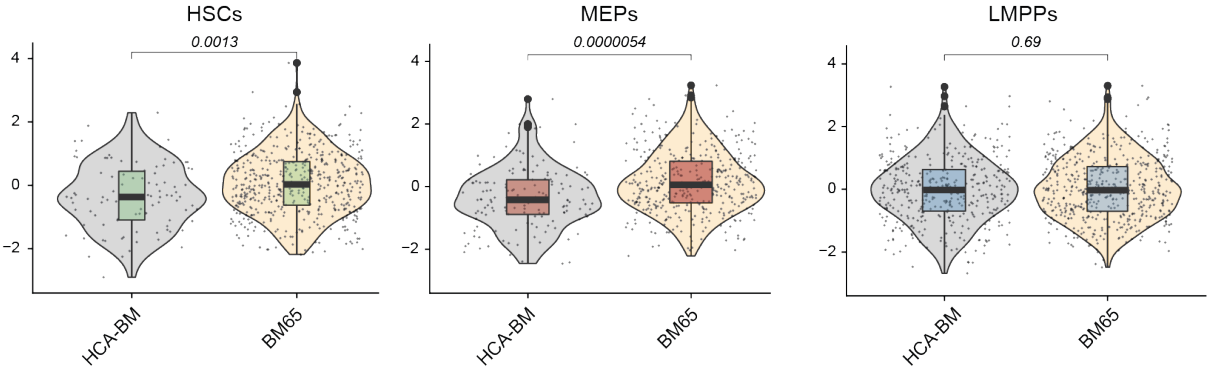
Supplementary Figure 17: Investigation of cell-type resolved local effects of the mosaic inversion on chromatin accessibility.

For the mSV affected TAD boundaries (TAD3195, TAD3202), we compared the NO profiles of the active X homolog between the mSV clone and the WT cells in a cell-type resolved fashion. We performed this analysis for HSCs and MEPs, as the cell count of those cell-types fulfilled the minimum required cell count needed for statistical testing. For each cell type, log2 fold changes of the mSV (active/inactive X) subclone compared to WT (active/inactive X) cells, as well as permutation adjusted p-values of two-sided likelihood ratio test are shown in the browser tracks. Additionally, names of the nearest genes adjacent to significant peaks (FDR 10%) are highlighted in the RefSeq genes track. Genomic coordinates of inferred differentially accessible peak regions are as follows: (1) chrX:67730000-68090000 (2) chrX:66400000-66910000 (3) chrX:66940000-67350000 (4) chrX:67580000-67880000 (5) chrX:67900000-68200000 (6) chrX:65640000-66150000.



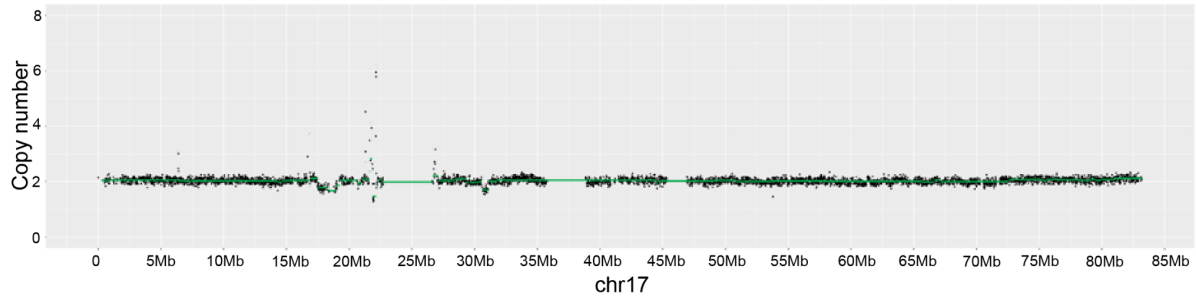
Supplementary Figure 18: Cell-type specific inferred accessibility of individual AR enhancers. **(a)** A browser track shows the genomic locations of peak regions with differential accessibility for HSCs and MEPs based on analyzing Strand-seq data, and the annotated AR enhancers based on

GeneHancer[10] and prior literature[11]. Individual enhancers showing significant differential accessibility inferred by NO are highlighted in red. (**b-c**) Violin plots show the inferred haplotype-resolved accessibility for WT cells (left) and the mSV subclone (right) for enhancer 10 (**b**) and enhancer 2 (**c**). For (**b-c**), a two-sided t-test was performed for each cell-type followed by Benjamini Hochberg multiple correction (n = 14, and 4 for WT and SV for HSC; n = 19, and 9 for WT and SV for MEP). Boxplots were defined by minima = 25th percentile - 1.5X interquartile range (IQR), maxima = 75th percentile + 1.5X IQR, center = median, and bounds of box = 25th and 75th percentile. Significance was established using the Wilcoxon rank-sum test. (**d-e**) Browser track snapshots depicting TF binding sites within enhancer 10 (**d**) and enhancer 2 (**e**) according to the ReMap Atlas.



Supplementary Figure 19. Comparative analysis of HSPCs from BM65 integrated with bone marrow data from the Human Cell Atlas.

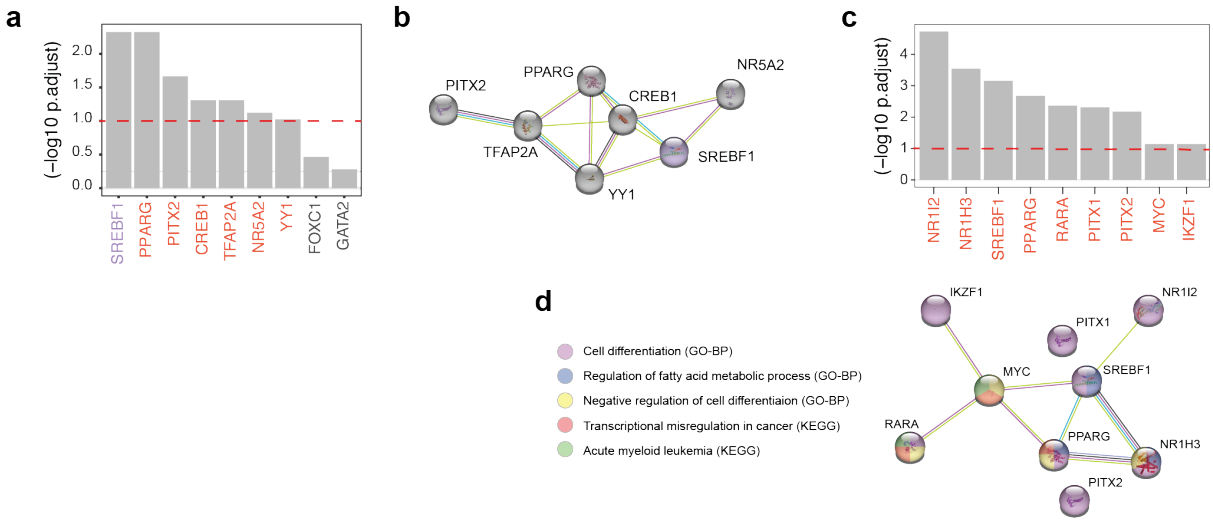
Single-cell RNA-seq data for CD34+ cell from BM65 was integrated with CD34+ cells from the Human Cell Atlas bone marrow dataset [12], and scored for enrichment of pathway activity of the AR signaling pathway using the escape [13] package. Violins show enrichment of the PID AR signaling pathway, indicating the P_{adj} value of enrichment (Benjamini-Hochberg adjusted two-sided FE-test). (n = 94 and 517 for HCA-BM and BM65 for HSCs; n = 145 and 407 for HCA-BM and BM65 for MEPs; n = 391 and 524 for HCA-BM and BM65 for LMPPs). Boxplots were defined by minima = 25th percentile - 1.5X interquartile range (IQR), maxima = 75th percentile + 1.5X IQR, center = median, and bounds of box = 25th and 75th percentile throughout this figure.



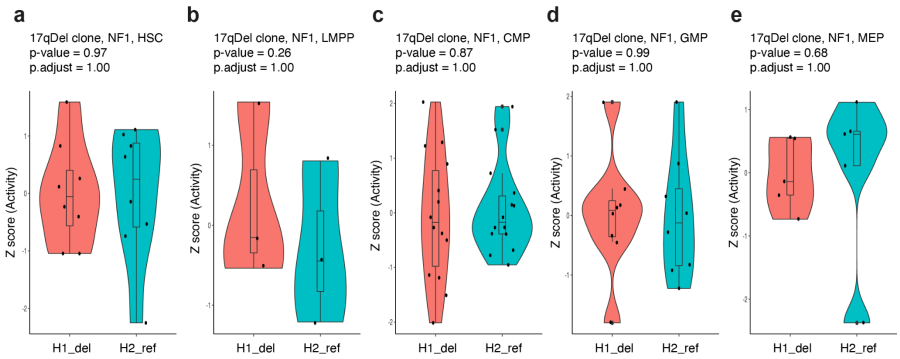
Supplementary Figure 20: Verification of mosaic deletions in BM712 using whole genome sequencing (WGS).

Chromosome 17 read coverage plot of whole genome sequencing (WGS) data obtained from BM712, revealing the presence of the 17p-Del and 17q-Del events in sorted CD34- cells. Split-read and paired-

end analysis using the Delly2 tool[14] identify the breakpoints of the 17q-Del event with base pair resolution (chr17:30602839-31097552), and validate the exon 1 deletion of *NF1*.

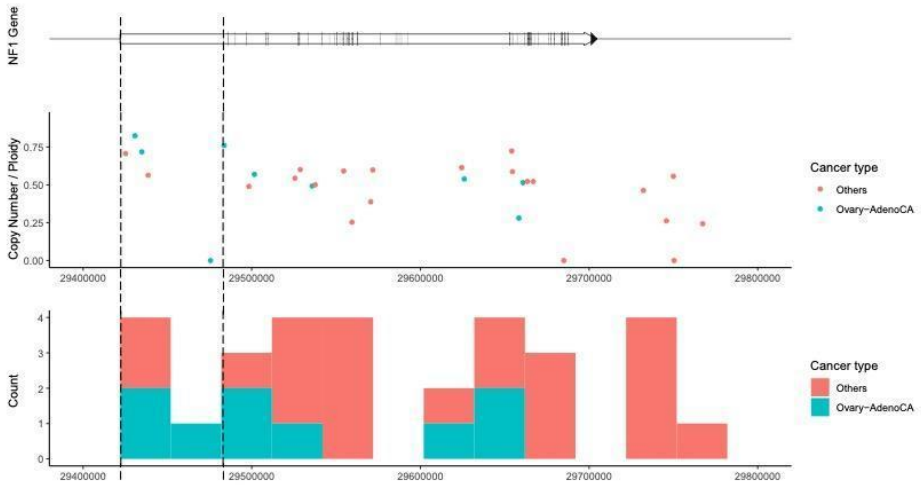


Supplementary Figure 21: TF activity and interaction analysis for the 17p- and 17q-Del subclones.
a,c) TFs whose targets show significantly differential NO in 17p-Del cells (**a**) and 17q-Del cells (**c**) vs. WT cells. Significant TFs are coloured (10% FDR threshold); and genes within the specific deleted locus are coloured in purple. **b**) STRING network of TFs in (**a**) (PPI enrichment *P*=3.57e-08, hypergeometric test[15]). **d**) STRING network of TFs in (**c**) (PPI enrichment *P*=0.00838, hypergeometric test[15]). Proteins are coloured based on their contribution to relevant significant genesets listed in the legend.



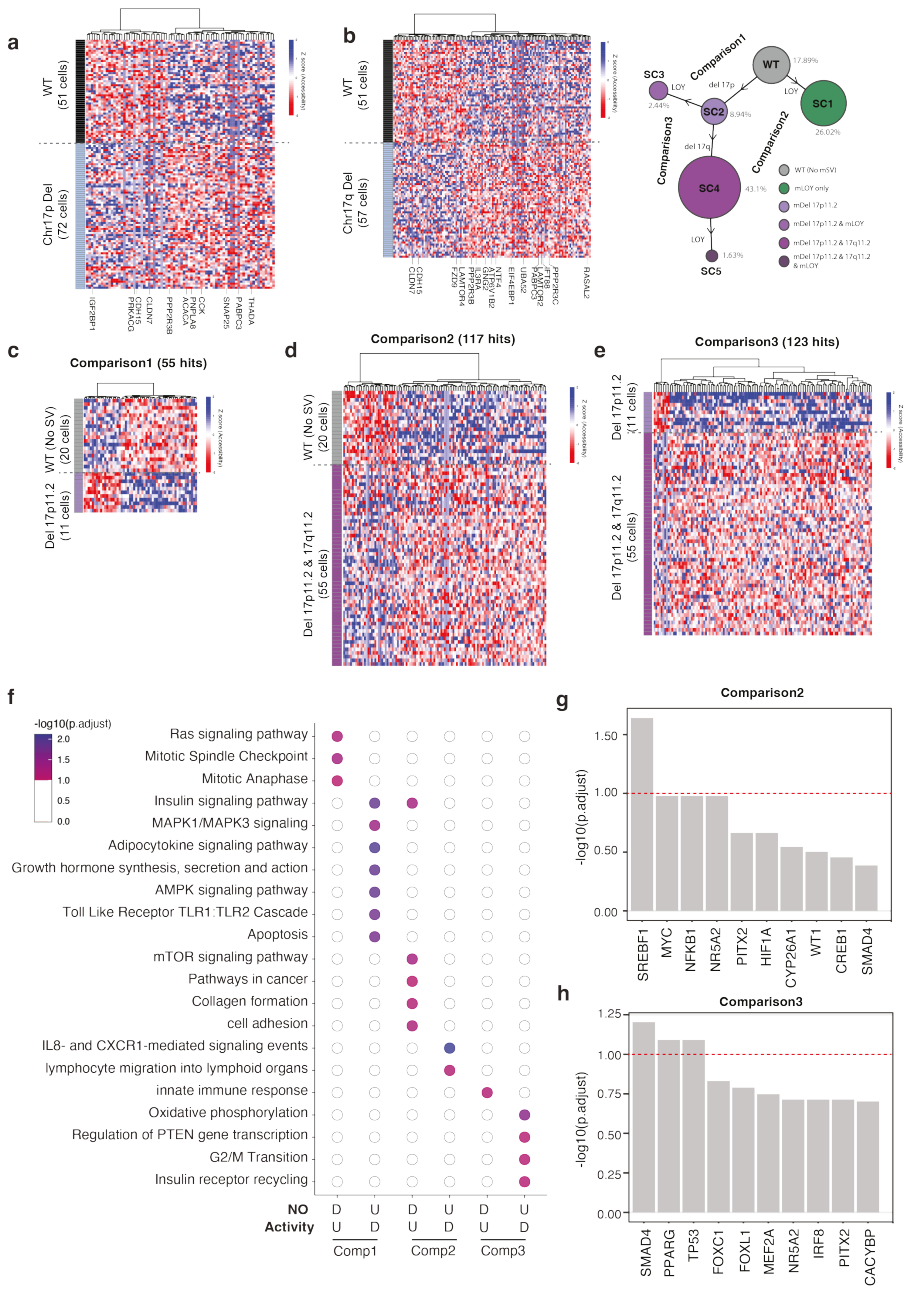
Supplementary Figure 22: Haplotype-resolved *NF1* activity based on NO in specific cell types
Using scNOVA, we extracted the haplotype-resolved NO at the *NF1* gene body for the 17q-Del clone and calculated the Z-score of the inferred gene activity for H1 (the rearranged homolog) and H2 (the WT homolog). The difference between two homologs was evaluated using the two-sided t-test followed by Benjamini-Hochberg multiple testing correction. We performed this analysis in a cell-type specific manner for **(a)** HSC (n = 8 both for H1_del, H2_ref), **(b)** LMPP (n = 3 both for H1_del, H2_ref), **(c)** CMP (n = 14 both for H1_del, H2_ref), **(d)** GMP (n = 8 both for H1_del, H2_ref), and **(e)** MEP (n = 5 both for H1_del, H2_ref). There is no significant difference in NO between the two

haplotypes at the *NF1* gene body, suggesting that the *NF1* gene fragment, which is truncated at its 5`-end (exon 1 deletion), is expressed from the rearranged homolog. Boxplots were defined by minima = 25th percentile - 1.5X interquartile range (IQR), maxima = 75th percentile + 1.5X IQR, center = median, and bounds of box = 25th and 75th percentile throughout this figure.



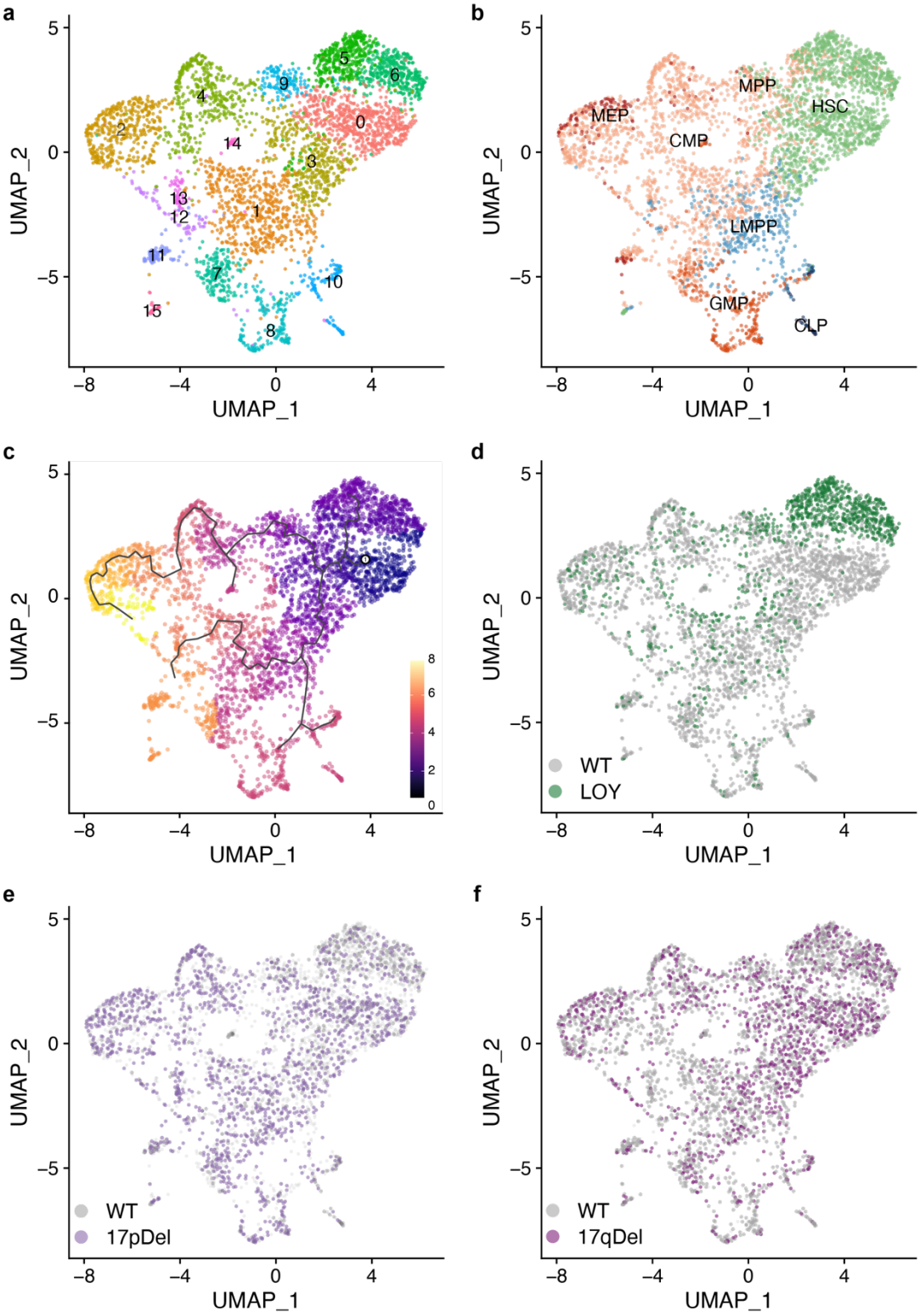
Supplementary Figure 23: Somatic deletion breakpoints in intron 1 of the *NF1* gene in PCAWG samples

Somatic deletion data were downloaded from the PCAWG resource of 2,583 whole cancer genomes[16]. Only deletions fully spanning the first exon of *NF1* are shown. The top panel shows the *NF1* gene structure; the middle panel shows the estimated copy number normalized by ploidy with each point representing a PCAWG donor; the bottom panel shows the distribution of breakpoints. The X-axis depicts the breakpoint coordinates in chromosome 17 (hg19). Notably, 3 out of 5 samples exhibiting an intron 1 deletion breakpoint are ovarian cancer samples – with such deletions seen in 3% of ovarian cancers (3 out of 110 ovarian cancer cases in the whitelisted PCAWG resource).

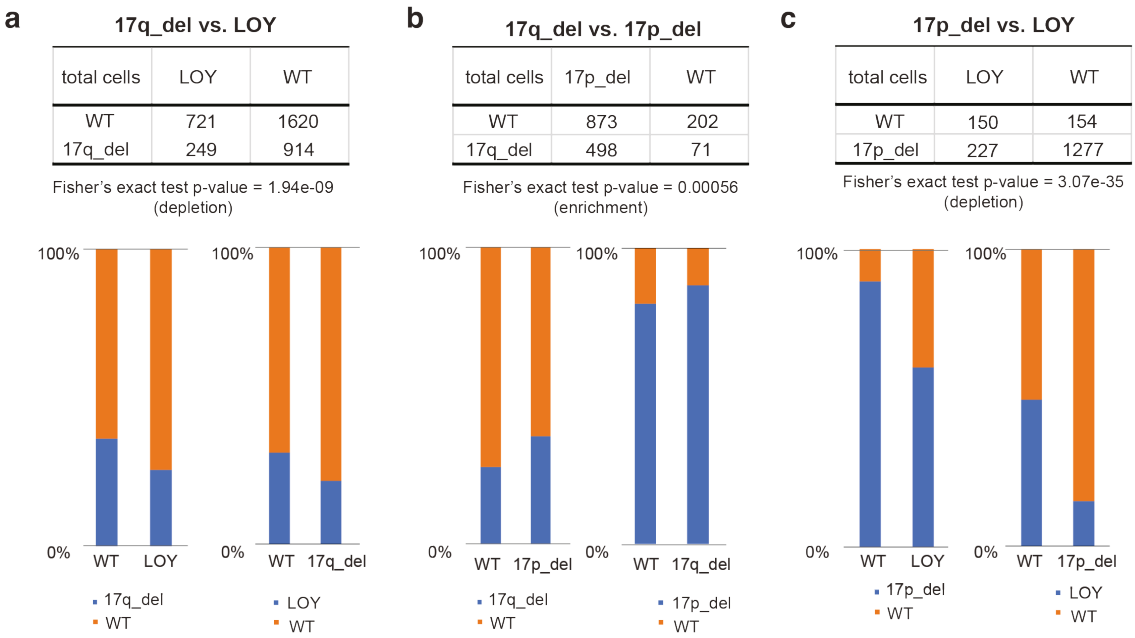


Supplementary Figure 24: Inference of altered gene activity in mSV subclones detected in BM712.

a) Differentially active genes between cells harboring the 17p-Del and the cells without 17p-Del (WT (no 17p-Del)), using scNOVA[4]. **b**) Differentially active genes between cells harboring 17q-Del and WT cells (no 17p-Del). **c-e**) Pairwise comparison of differential gene activity between cells bearing 17p11.2-Del (11 cells), cells bearing 17p11.2-Del and 17q11.2-Del (55 cells), and WT cells (no mSV, 20 cells). **f**) Pathway over-representation analysis using ConsensusPathDB[8] for the genes identified in the pairwise comparisons in (**c-e**). Significant pathways were identified by controlling the FDR at 10%. In the x-axis, U and D indicate inferred Up and Down-regulation, respectively. Comp1: 17p11.2-Del (11 cells) vs. WT (no mSV, 20 cells); Comp2: Del 17p11.2-Del & 17q11.2-Del (55 cells) vs. WT (no mSV, 20 cells); Comp3: 17p11.2-Del & 17q11.2-Del (55 cells) vs. 17p11.2-Del (11 cells). (**g-h**) TF-target over-representation analysis for the dysregulated genes identified in the Comp2 (**g**) and Comp3 (**h**). For Comp1, no significant TF was identified.

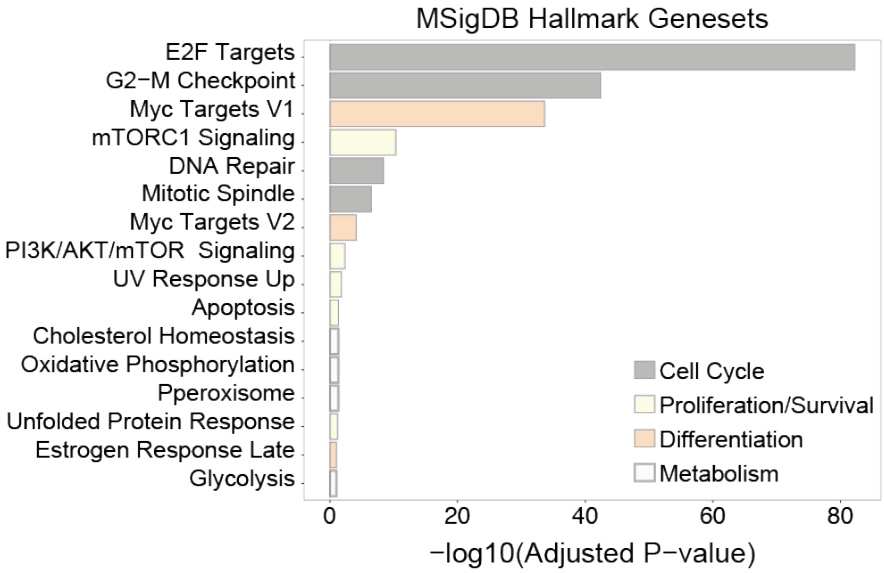


Supplementary Figure 25: Labeled UMAP projections of scRNA-seq data from BM712. **a)** Unsupervised clustering using Seurat[17]. **b)** Referenced-based cell-type labeling using SingleR[3], based on data from[18]. **c)** Pseudotime analysis of single-cell libraries using Monocle3[19]. Cells are ordered along the trajectory based on the decreasing number of HSCs (coloured from dark blue to yellow). **d-f)** Overlay of LOY (**d**), as well as 17p-Del (**e**) and 17q-Del (**f**) re-calls made using CONICSmat.



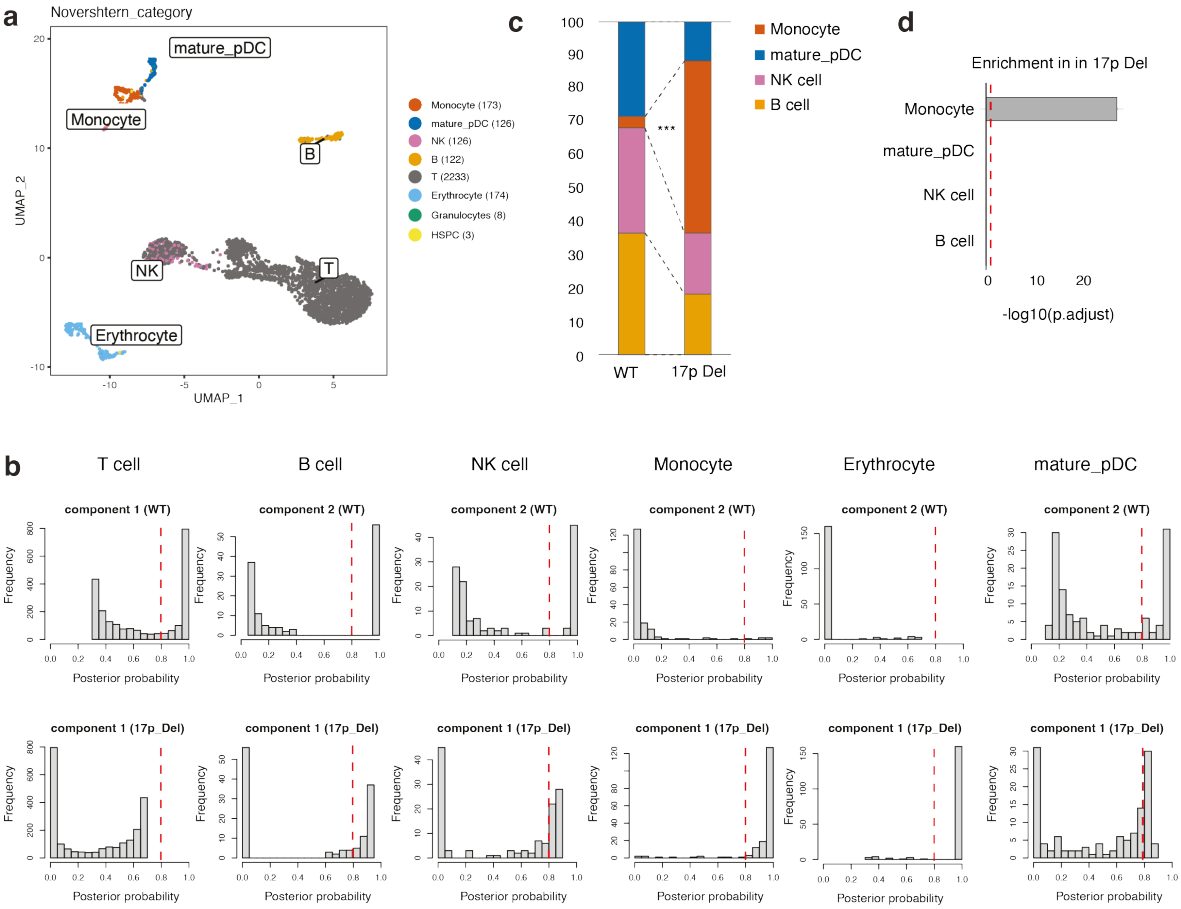
Supplementary Figure 26: Confirmation of Strand-seq-inferred clonal architecture in BM712 by targeted CNA re-calling using CONICSmat.

(a) Contingency table showing the association between 17q-Del and LOY calls generated by targeted CNA recalling[20] in scRNA-seq data from BM712. Bar graphs depicting the proportion of 17q-Del calls within the WT and LOY calls (left panel), and the proportion of LOY calls within the WT and 17q-Del cells (right panel) are shown. Similar analyses were performed for the association between 17q-Del calls and 17p-Del calls **(b)**, as well as 17p-Del calls and LOY calls **(c)**. Whereas the Del events tend to co-occur, we find that LOY is inversely correlated with the presence of each deletion, confirming the BM712 clonal structure inferred based on Strand-seq (**Fig. 5c**). The one-sided FE test was done for each of the comparisons.

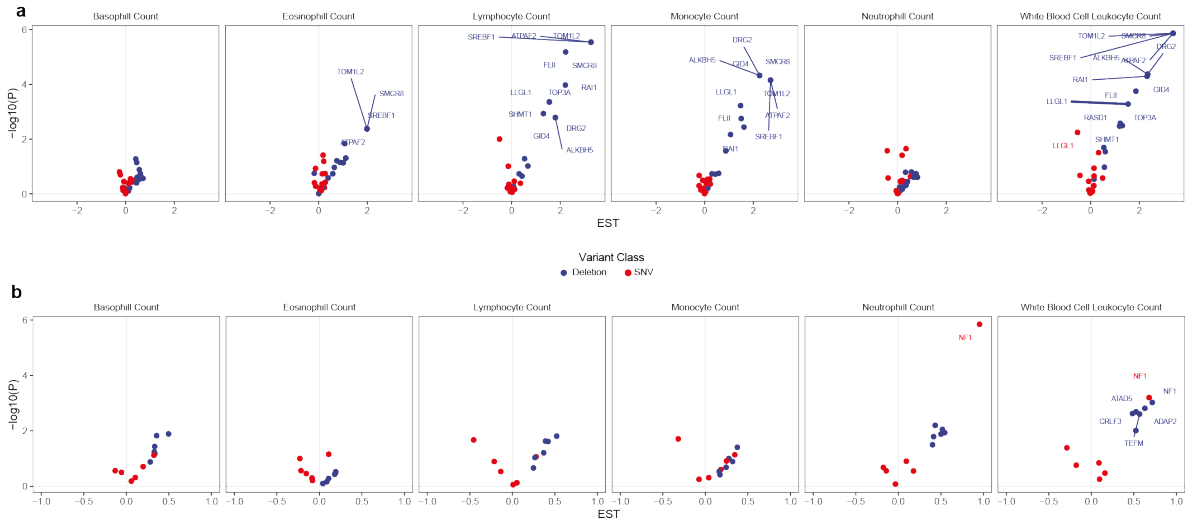


Supplementary Figure 27: Pathway enrichment analysis for 17q-Del cells from BM712.

MSigDB Hallmarks gene set enrichment analysis for differentially expressed genes in 17q-Del cells vs WT cells, identified in scRNA-seq.

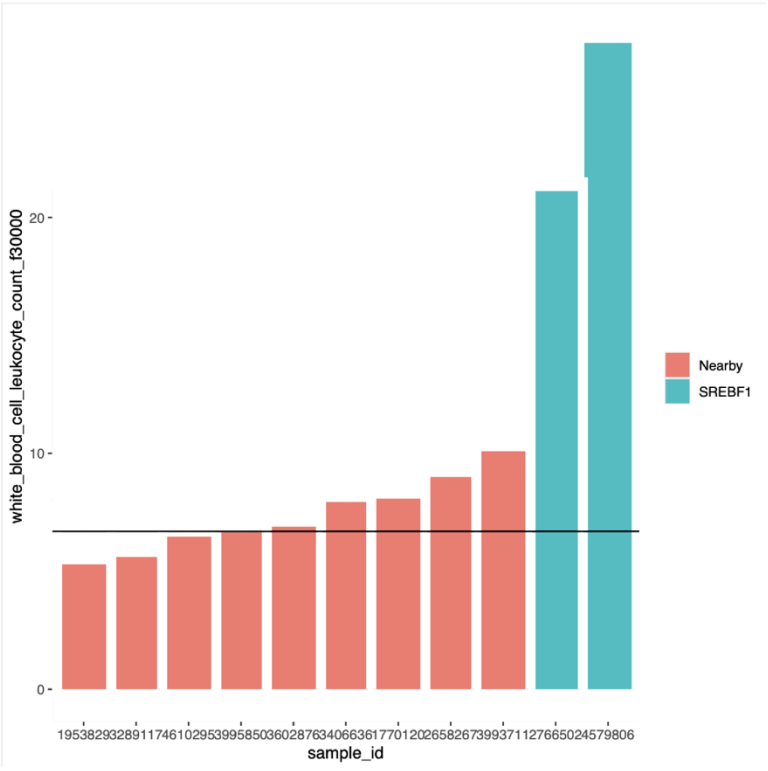


Supplementary Figure 28: Targeted CNA recalling in CD34- cells from BM712.
(A) t-SNE for the scRNA-seq libraries generated from CD34- cells. In total 2965 cells are depicted, and annotated with their cell-type as inferred by SingleR[3] using previously published immune cell-type annotations as a reference profile[18,21]. Inferred cell-types were assigned to higher level groups using a previously reported cell-type hierarchy[18]. Labels of cell-types with at least 10 single-cells are depicted. (B) Targeted CNA recalling was performed for the 17p-Del region using CONICSmat[20]. This analysis provides the posterior probability for cells being WT (upper panel) or 17q-Del cells (lower panel). The dashed red lines represent the cutoff of the posterior probability (> 0.8) used to assign the genotype [20]. Amongst six cell-types tested, confident assignments of WT and 17p-Del cells were made for four cell-types (B cell, NK cell, Plasmacytoid Dendritic Cells [mature pDC], and Monocyte). (C) Cell-type composition for WT cells and 17p-Del cells, depicted using stacked bar graphs. ***: significant cell-type enrichment amongst 17p-Del cells, with $P_{adj} < 0.0001$. Monocytes are significantly enriched in the 17p-Del clone compared to WT cells (P_{adj} = 9.6e-29; one-sided Fisher's exact test followed by Benjamini-Hochberg multiple correction). (D) The Bar graph shows the significance of cell-type enrichment for the 17p-Del clone compared to WT cells based on one-sided Fisher's exact test followed by Benjamini-Hochberg adjustment. The dashed red line indicates the 10% FDR cutoff.



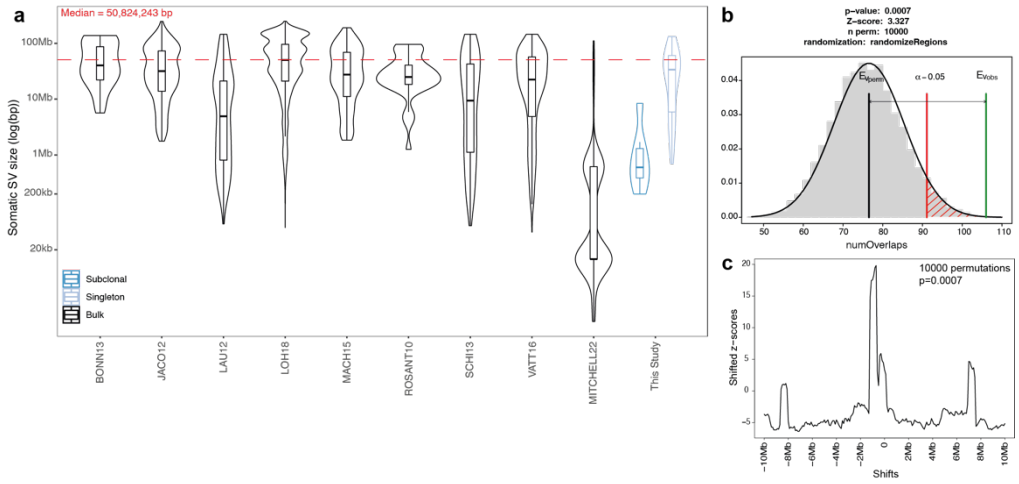
Supplementary Figure 29: Complete blood genotype-phenotype analysis using UK Biobank data for 17p- and 17q-Del regions.

Volcano plots summarize burden test results for 17p-Del (**a**) and 17q-Del (**b**) based on blood count traits and genotype data from the UK Biobank (see also **Fig. 6b, c**).

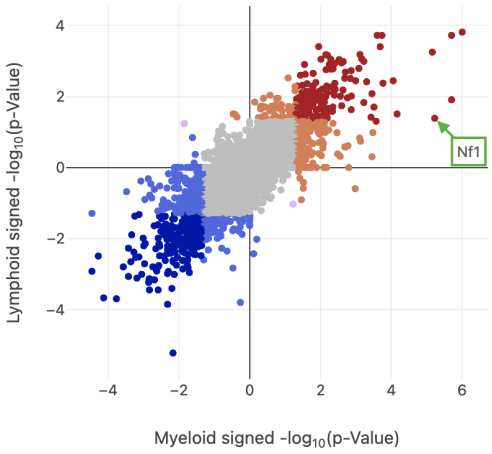


Supplementary Figure 30. Investigation of UK biobank donors with CNAs overlapping *SREBF1*.

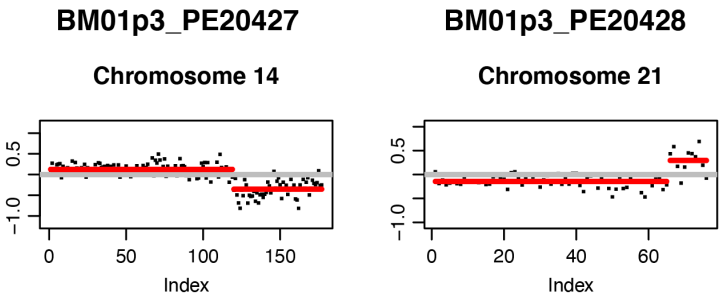
Two UK biobank donors harboring CNAs that overlap *SREBF1* are outliers in terms of blood cell count compared to donors with a CNA nearby.



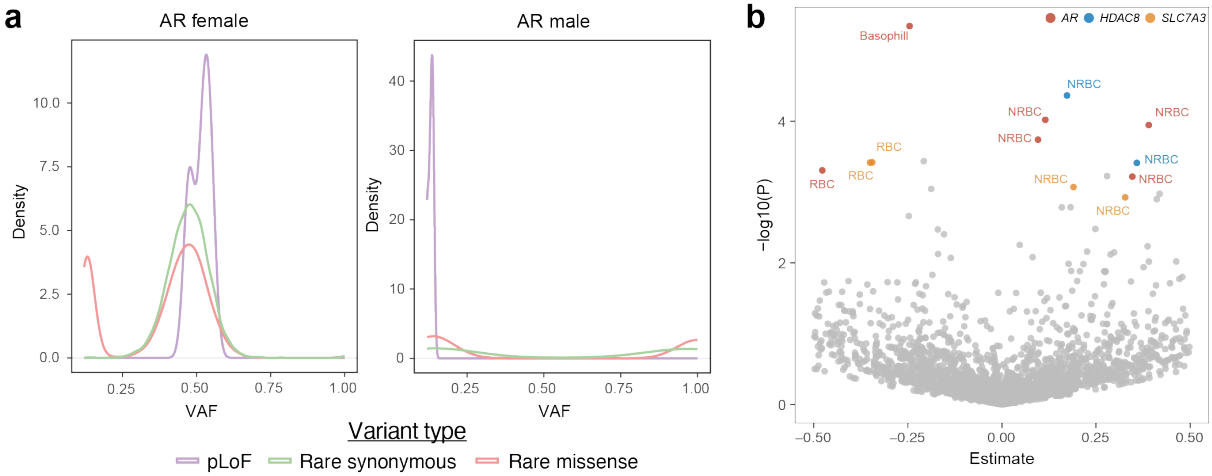
Supplementary Figure 31: Comparing previously published mosaic copy-number alteration (mCNA) data to this study.
a) Violin plots showing the size distribution of mCNAs identified per study[22–30]. (n = 54, 681, 339, 8185, 341, 34, 200, 1141 and 399 for BONN13, JACO12, LAU12, LOH18, MACH15, ROSANT10, SCHI13, VATT16, MITCHELL22, and This Study, respectively; Boxplots were defined by minima = 25th percentile - 1.5X interquartile range (IQR), maxima = 75th percentile + 1.5X IQR, center = median, and bounds of box = 25th and 75th percentile.) **b**) Permutation plot of overlaps between 200 kb regions around interstitial mCNA breakpoints from a), and SCE hotspots from this study. **c**) Local enrichment plot of breakpoints from **a-b**). *P-value* is based on a two-sided permutation test.



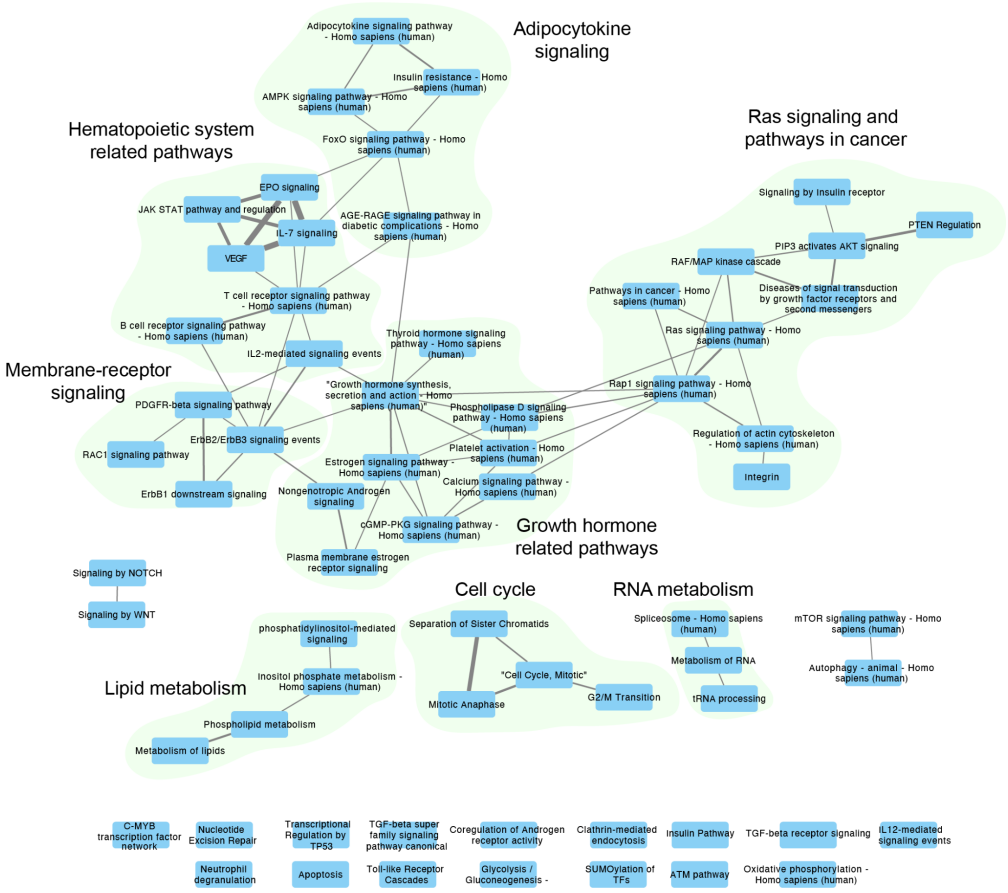
Supplementary Figure 32: Analysis of data release from a CRISPR *in vivo* knockout (KO) screen corroborates effects of *Nf1* loss on myeloid cell bias.
Display item generated using the interactive open access database provided by Haney *et al.*, a study designed to uncover novel regulators of hematopoiesis using *in vivo* KO screen[31] (www.hematopoiesiscrisprscreens.com). The study identified genetic regulators of the lineage trajectory from HSCs to myeloid cells, erythroid cells, T-cells and B-cells. Notably, we find that *NF1* is amongst the most myeloid-biased genes in this KO screen, ranking 4th out of ~7,000 screened genes for myeloid enrichment, based on its enrichment *P*-value estimated by permuting gene-targeting guides[31]. (The color code, based on Haney *et al.*, is as follows: Gray = no significant enrichment. Dark blue = significant depletion in mature cells. Light blue = borderline significant depletion in mature cells. Orange = borderline significant enrichment in mature cells. Red = significant enrichment in mature cells.)



Supplementary Figure 33: Singleton CNA discovery in HSPCs in single-cell whole genome sequencing (scWGS) data.
Single-cell read depth plot for intermediate coverage scWGS data generated from HSPCs (plotted using 500kb-sized genomic bins). We performed copy-number segmentation using the DNAcopy method[32]. In total, we find 20 CNA events in 16 single cells amongst 480 scWGS cells, with each scWGS cell corresponding to a cell sequenced using the scMNase-seq protocol outlined in the **Methods** section [33,34]. Two representative chromosomes with CNAs are depicted.

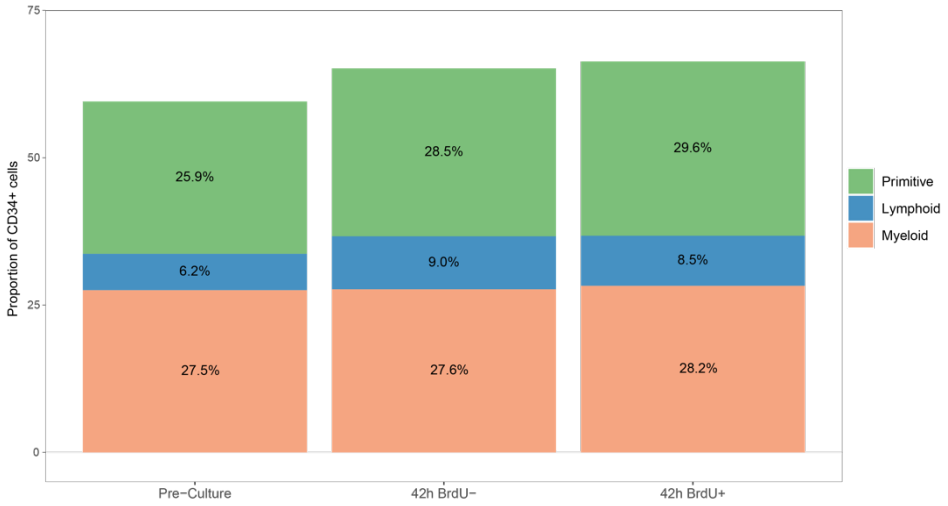


Supplementary Figure 34: Blood genotype-phenotype analysis using UK Biobank data for Xq-Inv region.
a) Variant allele frequency plot for mutations in *AR*, separated by mutation type and sex. **b**) Volcano plot showing association test results of single rare missense variant at the Xq-Inv locus for all 11 blood count traits. The full respective list of missense variants analyzed is available from **Supplementary Table 18**. Variants with P_{adj}<0.05 are colored by gene and labeled by trait: NRBC, nucleated red blood cell count; RBC, red blood cell count; basophil, basophil count. Variants with P_{adj} >= 0.05 are colored in gray. Y-axis depicts nominal *P*-values obtained using a two-sided Wald test (**Methods**).

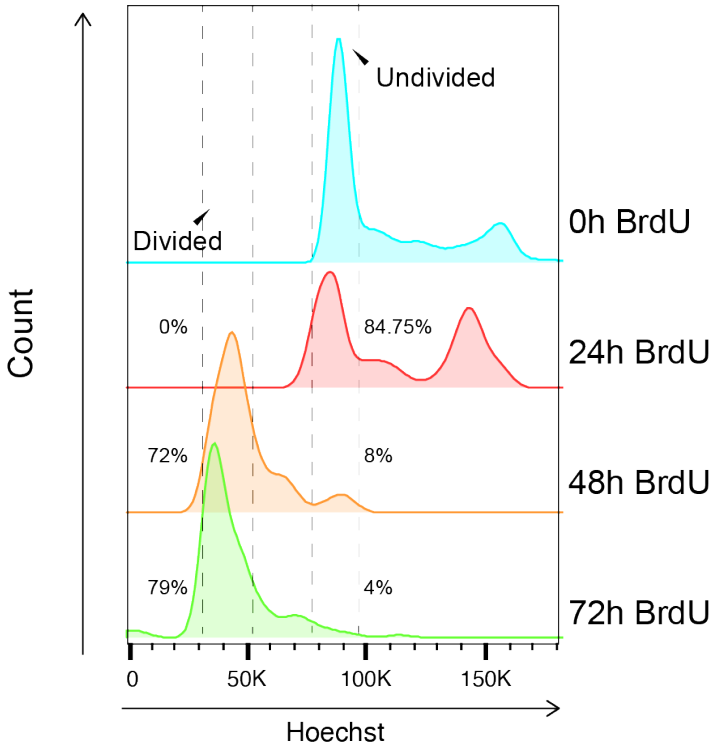


Supplementary Figure 35: Similarity analysis of 66 pathways reported in Supplementary Fig. 12.

In **Supplementary Fig. 12**, over-represented pathways of dysregulated genes in mSV subclones are reported. To reduce the redundancy between the biological pathways and comprehensively understand the underlying hierarchical structure of the over-represented pathways, the similarity between pathways are shown as a network format in this figure. In this network, each node represents one of 66 over-represented pathways in **Supplementary Fig. 12**, and the thickness of edges between the nodes represent the Jaccard similarity between them. Edges are shown when the Jaccard similarity is larger than 0.16. It resulted in the bigger category of pathways as represented in a light green background behind the nodes. We annotated the name of those bigger categories based on the common characteristics of nodes belonging to that category.



Supplementary Figure 36: **Comparison of lineage composition of cells from donor BM92 before culture and after 42 hours of culture without and with BrdU.**
Bars represent the frequency of broad cell-types which could be defined by the antibody panel out of the CD34+ cells (primitive: CD34+CD38-CD45Ra-CD90+/-; lymphoid: CD34+CD38-CD45Ra+CD90-; myeloid: CD34+CD38+CD45Ra-CD90-). We do not detect significant differences in composition between the 3 conditions shown (*Chi*-Square test, *P*=0.95089).



Supplementary Figure 37: Defining an optimum time point for BrdU incorporation into HSPCs.
Flow cytometry histograms showing the Hoechst fluorescence of CD34+ UCB cells cultured for 0, 24, 48 and 72 hours with BrdU. Dashed lines denote the undivided and divided populations, based on the 0 BrdU control. See **Supplemental Notes** for further details.

Supplementary Methods

1. Identification of active X chromosome from the female genome

To identify the transcriptionally active X chromosomal homolog in the cells from the female donor BM65, we performed haplotype-specific nucleosome occupancy analysis using scNOVA, as previously described[4]. First, we analyzed single-cells from the Xq-Inv subclone, and resolved the reads by haplotype for chromosome X in those cells. We then extracted the haplotype-resolved nucleosome occupancy for each gene body and calculated the mean of this value for each cell. Finally, we compared the single-cell nucleosome occupancy from haplotype 1 and haplotype 2 using a *t*-test, to identify the transcriptionally active X chromosome among the two homologous chromosomes.

2. Protein-protein interaction (PPI) network analysis using STRING

To reconstruct the network between dysregulated TFs inferred by scNOVA, we used the STRING[35] multiple protein search algorithm. To evaluate the enrichment of direct interaction between dysregulated transcription factors (TFs), we utilized network statistics from STRING, based on the PPI enrichment *P*-value. For the analysis of BM65, we extended the network of dysregulated TFs by including their first neighbors based on the physical binding partners reported in the NCBI gene pages (Available from: https://www.ncbi.nlm.nih.gov/gene/) (**Supplementary Fig. 13**). To reconstruct the network between dysregulated TFs and their first neighbors, we utilized STRING[35] as mentioned above. Pathways enriched for members of the PPI network were also identified by STRING.

3. Similarity analysis for over-represented pathways of dysregulated genes in mSV subclones

To reduce the redundancy between the biological pathways and comprehensively understand the underlying hierarchical structure of the over-represented pathways, we built a network model showing the links among the 66 pathways enriched by the dysregulated genes in mSV subclones (**Supplementary Fig. 35**). For each pair of the 66 pathways, we computed the Jaccard coefficient and connected the two pathways with the Jaccard coefficient >= 0.16. This analysis resulted in 6 connected subnetworks. Among the connected subnetworks, the biggest subnetworks were further categorized into five network modules based on their connectivity and biological implications.

Supplementary Notes

1. Selecting an optimal timepoint for BrdU incorporation in cultured human HSPCs

In order to obtain a high number of usable, high-quality Strand-seq libraries upon sequencing, a critical step in the experimental setup is to maximize the number of cells which have undergone a single cell division in the presence of BrdU, while ensuring that a second round of DNA synthesis has not begun (as this will incorporate BrdU into the template DNA strands, making cells unusable for Strand-seq). In order to identify this timepoint, we carried out a BrdU timecourse in CD34+ UCB cells, isolating nuclei and profiling their BrdU incorporation by staining with Hoechst at 0h, 24h, 48h and 72h (**Supplementary Fig. 37**; see [36] for more details on BrdU timecourses). Based on these profiles, 48h showed the closest to 100 % of cells having completed a single cell division (*i.e.* a Hoechst fluorescence at ~ 50 % of the 0 BrdU control). However, given the heterogeneity of cells within HSPCs, and the possibility of variability among donors, we settled on a more conservative timepoint of 45 hours for BrdU incorporation in our Strand-seq experiments.

2. Characteristics of *de novo* mSVs in HSPCs

In 1 out of every 43 cells, regardless of donor age, we identify a singleton mSV. We scrutinized these singleton mSVs by utilizing single-cell tri-channel processing (scTRIP), the underlying principle of which is that each structural variant is characterized and discerned by a specific 'diagnostic footprint'[37]. The footprint encapsulates the co-segregation patterns of rearranged DNA segments, identified by sequencing single strands of each chromosome in each cell in a haplotype-resolved manner, using Strand-seq.

Our analysis revealed that singleton mSVs, but not subclonal mSVs, bear the following characteristics indicative for *de novo* DNA rearrangement:

(1) Amongst the 32 singleton mosaicisms we identify in our HSPC single-cell genomic dataset, 21 (66%) display terminal gains or losses confined to a single haplotype. In contrast, subclonal mSVs lack terminal rearrangements entirely and instead exhibit a significant enrichment for interstitial rearrangements (P=0.0004; Fisher's exact test) when compared to singleton mSVs. These terminal rearrangement footprints of singleton mSVs are depicted, amongst all singleton mSVs events, in **Supplemental Data File 1** as well as in **Fig. 1c**. The frequent occurrence of terminal losses and gains in singleton mSVs suggests that the derivative chromosomes emerging from these rearrangements often lack telomeric stabilization events, which may potentially increase the likelihood that more DNA rearrangements accumulate in these cells (**Fig. 1c,f**)[38].

(2) Three singleton mSVs exhibit characteristics of complex chromosomal rearrangements, encompassing mSVs triggered by breakage-fusion-bridge (BFB) cycles[37,39] and an amplification induced by terminal sister chromatid fusion, leading to a sevenfold increase in copy number (**Fig. 1c**). The latter rearrangement event could stem from a BFB process occurring in the absence of telomere stabilization[38].

(3) Our observations, as delineated in **Fig. 1j, Supplementary Fig. 5** and in the main text, show instances of SCEs occurring in the same cell and haplotype directly at the breakpoints of singleton mSVs (we did not, by comparison, observe significant colocalisation with SCEs for the breakpoints of subclonal mSVs). This indicates a link between SCE formation and DNA rearrangement processes resulting in mSV formation[40,41].

(4) On average, singleton mSVs are ~17.6 times larger than subclonal mSVs (mean size of 36.9 and 2.1 Megasbasepairs (Mb), respectively; P=0.0009, Wilcoxon rank-sum test; **Fig. 1d**). Due to their substantial size, these mSVs result in significant autosomal aneuploidy, often tens of megabasepairs in length, which is likely to be detrimental for their clonal expansion given the adverse effects of autosomal aneuploidy in normal cells[42].

We therefore infer that the singleton mSVs identified in our HSPC dataset predominantly represent *de novo* mSV formation events. Considering their distinct characteristics, such as substantial regions of autosomal aneuploidy and likely lacking telomeric functionality on the affected homolog, it is plausible that most newly formed mSVs are incapable of reaching considerable subclonal frequencies in normal HSPCs.

3. Comparison with prior surveys of mosaic copy-number alterations and mSVs

Placing our findings into the larger scope of research on clonal hematopoiesis and its previously known association with mosaic CNAs (*i.e.* copy-imbalanced mSVs), we find that the subclonal mSVs identified in our study through single-cell genomic sequencing (Strand-seq) are significantly smaller (**Supplementary Fig. 31**) than CNAs detected in surveys based on utilizing blood cells in bulk (primarily pursued using microarray based hybridization)[22–30]. This observation is consistent with the high genomic resolution of Strand-seq, which exceeds bulk approaches for detecting certain subclonal structural variant classes, particularly in the case of sub-Megabase sized somatic variants[37]. By comparison, the subclonal mSVs detected in our study fall into the same size range as for one study, by Mitchell and colleagues[23], who undertook WGS of clones derived from single HSPCs. These data suggest that the mSVs identified in our study as well as by Mitchell et al. may have escaped detection in prior CNA-focused studies using bulk hybridization-based assays. We note that Mitchell et al. and our study are the only two surveys specifically examining HSPCs, rather than the peripheral blood, leaving the possibility that different cell types are differentially impacted by mSVs.

To bolster our findings relating to common fragile sites (CFSs) and their association with mSVs, we performed additional analyses on the aforementioned prior data on mosaic CNAs[22–30]. We permuted breakpoints from previously reported mosaic CNAs against the SCE hotspots identified in our study, identifying a similar trend to that in our own data, whereby mosaic CNA breakpoints demonstrate significant local enrichment at SCE hotspots (**Supplementary Fig. 31**). This observation suggests that genomic loci in HSPCs, predisposed to mSV formation, also denote fragile regions prone to CNAs in peripheral blood.

4. Investigation of genes associated with local effect of subclonal inversion in BM65

Amongst the genes our analysis associates with a local effect of the subclonal inversion in BM65, 3 genes – which include *AR*, the top hit, as well as *HDAC8* and *MAGEE1* – are inferred to be more active in mSV cells, whereas the remaining 10 genes are inferred to be downregulated in mSV cells. 5/13 genes including *AR*, *EDA2R*, *SLC7A3*, *HDAC8*, and *CHIC1* show expression in HSPCs according to previously published bulk RNA-seq data from HSPCs.

5. Investigation of functional links between dysregulated TFs in the 17p-Del subclone in BM712

Prior reports have tied *Srebf1* knockout in murine blood cells to an increase in primitive HSPCs[43], and our findings closely mirror these data, with 17p-Del cells showing enrichment for both HSCs and CMPs (**Fig. 5a**). Protein-protein interaction mapping of the dysregulated TFs using STRING[35] (**Supplementary Methods)** revealed significant functional associations between *SREBF1* and six

additional TFs, which suggests that *SREBF1* cooperates with functionally-related TFs that become dysregulated in association with 17p-Del (*P*=3.57e-08; **Supplementary Fig. 21**). Collectively, these data implicate the somatic hemizygous loss of *SREBF1* as a putative driver of gene dysregulation in 17p-Del-bearing cells. SREBF1 (also known as SREBP1) has been reported to induce *PPARG* expression[44,45]. Furthermore, PPARG activates CREB1 expression by binding its promoter[46,47], in line with the tight functional connections between the TFs identified in the STRING-based PPI analysis, and in support of a possible causal role of *SREBF1* deletion in mediating the molecular phenotype seen in BM712.

6. Potential small deletions at regions of recurrent SCE/mSV formation

We observed marked localized 'fragility' at the *FRA3B* locus in donor BM762 (**Fig. 1j**). Although this donor showed similar SCE counts to the other samples (**Extended Data Fig. 1**), nine single cells exhibit an SCE, an mSV, or both within a 500 kb region of this CFS. One cell shows two SCEs at *FRA3B*, one on each homolog, while another harbors a terminal deletion originating from the same locus (**Fig. 1j, Supplementary Fig. 5**). A closer look at *FRA3B* at sub-Mb resolution reveals potential small deletions (< 200 kb), which are below our mSV discovery resolution[37], aligning with SCEs in the same cells (**Fig. 1j; Supplementary Fig. 5**). Manual inspection shows that these putative small deletions arise in 37.5% (3/8) of cells with an SCE versus 1.89% (1/53) without an SCE (*P*=0.0055; Fisher's exact test).

7. Analysis of somatic SNVs from the IntoGen Clonal Hematopoiesis Mutation Browser

We analysed data from the IntOGen Clonal Hematopoiesis Mutation Browser[48], which at the time of analysis (9th June 2023) reported on 175 mosaic SNVs falling into the *NF1* gene. We find a marked abundance of potentially deleterious variants (27% [47/175]) amongst these SNVs. This includes 26/175 (15%) of predicted loss-of-function (pLoF) SNVs (frameshift variant: 5.1% [9/175]; stop gained: 9.7% [17/175]). Additionally, 21/175 (12%) are predicted to affect splicing and thus potentially resulting in *NF1* transcripts generated from an aberrant open reading frame[49] (splice donor variant: 6.3% [11/175]; splice acceptor variant: 3.4% [6/175]; splice region variant: 2.3% [4/175]). These data suggest that *NF1* hemizygous loss could fuel clonal hematopoiesis.

8. Analysis of somatic SNVs affecting the *AR* gene in the UK Biobank

The *AR* gene has a sex-specific molecular biology[50], and shows highly sex-specifc SNV distributions in the UK Biobank cohort (for example, *AR* pLoF SNVs are seen in females (**Fig. 6d**), whereas they are essentially absent in males from the UK Biobank cohort). Therefore, our comprehensive analysis of presumed mosaic SNVs within the *AR* gene employed a sex-specific approach. We utilized sex as a covariate in the employed multiple linear regression model; in addition, we built two separate models for males and females, respectively. We focused our analysis of the *AR* gene on females, as the mosaic inversion was identified in a female (BM65).

9. Singleton CNA discovery in HSPCs in scWGS data

We analyzed intermediate coverage scWGS data generated followed the fragmentation of single cell genomes with MNase (which cuts the human genome in a highly uniform manner) to explore the frequency of large sized copy-number imbalanced SVs (CNAs) in the absence of cell culturing and BrdU application. We performed copy-number segmentation using DNAcopy[32], with standard parameter settings, to call CNAs among 480 scWGS cells – with each scWGS cell corresponding to a cell sequenced using the scMNase-seq protocol outlined in the Methods section[33,34]. From the 480 single-cell libraries, we find 20 CNAs in 16 cells (3.3%) (see e.g. examples shown in **Supplementary Fig. 30**). In the case of Strand-seq, we find 32 singleton mSVs in 1133 single-cell libraries, a frequency (2.8%) similar to the singleton CNA frequency seen in these intermediate coverage scWGS data. In

conclusion, singleton mSV proportions identified in HSPCs using Strand-seq are in line with findings obtained from scWGS libraries generated without the incorporation of BrdU. Therefore, BrdU incorporation is unlikely to have a significant impact on singleton mSV frequencies seen in HPSCs.

10. Analysis of data release from a CRISPR *in vivo* knockout (KO) screen

During the revision of our manuscript, we reanalyzed CRISPR knockout (KO) screens performed *in vivo* in a mouse model, reported in a manuscript recently posted by Haney and colleagues[31], to bolster observations made with respect to the functional consequences of candidate genes in HSPCs suggested in our study. These screens were conducted using expanded primary mouse hematopoietic stem and progenitor cells; targeting ~7,000 genes from various functional categories including transcription factors, kinases, phosphatases, drug targets, and genes linked to apoptosis and cancer[31].

The KO screen under consideration included the *Nf1* gene but did not encompass knockout of *Srebf1* or activation of the *Ar* gene. Analysis of this screen data release (see www.hematopoiesis_crispr screens.com) suggests that *Nf1* KO promotes HSPC differentiation into mature cells *in vivo* – and preferably those of the myeloid lineage (P<0.00001, effect score=5.6, for myeloid cell enrichment). Remarkably, *Nf1* is amongst the most myeloid-enriched genes detected based on this data release, ranking 4th among ~7,000 screened target genes based on P-value (**Supplementary Fig. 32**). These KO screen data hence further corroborate the data from our Strand-seq experiments (**Fig. 5a**), supporting our conclusions that *Nf1* disruption may bias CD34+ cells towards myeloid lineages.

11. Interplay between mSVs and clonal hematopoiesis.

Another outstanding question regarding mSVs and CH pertains to the potential synergy between mSVs and SNV mosaicism in association with CH[51,52]. Though not our primary focus, we analyzed surplus material from 6 of the 19 donors for common driver SNVs tied to clonal hematopoiesis using high coverage gene panel sequencing (**Supplementary Table 1**). Only one out of these 6 donors (BM762) displayed a detectable SNV (TET2 p.W1291C, CF=39.4%) and also had a low frequency (CF=3.2%) X chromosome loss. In contrast, BM70, which had multiple mSVs, showed no common driver SNVs. These data are consistent with a recent study in 628,388 blood donors showing lack of association between common driver SNVs in CH and mosaic CNAs after adjusting for age, sex, and smoking status[53], and suggest that CH driver SNVs and mSVs emerge as separate events in the blood compartment.

Prior reports have associated mosaic CNAs in blood with CH[22,28,54,55] an age-related phenomenon where HSPCs contribute to genetically distinct blood cell subpopulations. Our findings imply that subclonal mSVs, seen in 36% of donors over 60 years, commonly impact HSPC function by affecting diverse genomic loci, including genes with known or suspected roles in clonal hematopoiesis[48]. The prevalence of this class of mosaicism implies that the cumulative phenotypic impact of mSVs on specific tissues or organs could potentially parallel that of SNVs, a finding that underscores the necessity for future studies in larger cohorts.

Supplementary References

1. Buenrostro, J. D. *et al.* Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation. *Cell* **173**, 1535-1548.e16 (2018).
2. Gudmundsson, K. O. *et al.* Prdm16 is a critical regulator of adult long-term hematopoietic stem cell quiescence. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 31945–31953 (2020).
3. Aran, D. *et al.* Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat. Immunol.* **20**, 163–172 (2019).
4. Jeong, H. *et al.* Functional analysis of structural variants in single cells using Strand-seq. *Nat. Biotechnol.* (2022) doi:10.1038/s41587-022-01551-4.
5. Wingender, E., Dietze, P., Karas, H. & Knüppel, R. TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res.* **24**, 238–241 (1996).
6. Sandelin, A., Alkema, W., Engström, P., Wasserman, W. W. & Lenhard, B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.* **32**, D91-4 (2004).
7. Sharma, N. V. *et al.* Identification of the Transcription Factor Relationships Associated with Androgen Deprivation Therapy Response and Metastatic Progression in Prostate Cancer. *Cancers* **10**, (2018).
8. Kamburov, A. & Herwig, R. ConsensusPathDB 2022: molecular interactions update as a resource for network biology. *Nucleic Acids Res.* **50**, D587–D595 (2022).
9. Lachmann, A. *et al.* ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics* **26**, 2438–2444 (2010).
10. Fishilevich, S. *et al.* GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database* **2017**, (2017).
11. Takeda, D. Y. *et al.* A Somatically Acquired Enhancer of the Androgen Receptor Is a Noncoding Driver in Advanced Prostate Cancer. *Cell* **174**, 422-432.e13 (2018).
12. Hay, S. B., Ferchen, K., Chetal, K., Grimes, H. L. & Salomonis, N. The Human Cell Atlas bone

marrow single-cell interactive web portal. *Exp. Hematol.* **68**, 51–61 (2018).

13. Nick Borcherding, J. A. *Escape*. (Bioconductor, 2020). doi:10.18129/B9.BIOC.ESCAPE.
14. Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339 (2012).
15. Rivals, I., Personnaz, L., Taing, L. & Potier, M.-C. Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics* **23**, 401–407 (2006).
16. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).
17. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
18. Xie, X. *et al.* Single-cell transcriptomic landscape of human blood cells. *Natl Sci Rev* **8**, nwaa180 (2021).
19. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).
20. Müller, S., Cho, A., Liu, S. J., Lim, D. A. & Diaz, A. CONICS integrates scRNA-seq with DNA sequencing to map gene expression to tumor sub-clones. *Bioinformatics* **34**, 3217–3219 (2018).
21. Novershtern, N. *et al.* Densely interconnected transcriptional circuits control cell states in human hematopoiesis. *Cell* **144**, 296–309 (2011).
22. Loh, P.-R. *et al.* Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations. *Nature* **559**, 350–355 (2018).
23. Mitchell, E. *et al.* Clonal dynamics of haematopoiesis across the human lifespan. *Nature* **606**, 343–350 (2022).
24. Vattathil, S. & Scheet, P. Extensive Hidden Genomic Mosaicism Revealed in Normal Tissue. *Am. J. Hum. Genet.* **98**, 571–578 (2016).
25. Machiela, M. J. *et al.* Characterization of large structural genetic mosaicism in human autosomes. *Am. J. Hum. Genet.* **96**, 487–497 (2015).
26. Bonnefond, A. *et al.* Association between large detectable clonal mosaicism and type 2 diabetes

with vascular complications. *Nat. Genet.* **45**, 1040–1043 (2013).

27. Schick, U. M. *et al.* Confirmation of the reported association of clonal chromosomal mosaicism with an increased risk of incident hematologic cancer. *PLoS One* **8**, e59823 (2013).
28. Jacobs, K. B. *et al.* Detectable clonal mosaicism and its relationship to aging and cancer. *Nat. Genet.* **44**, 651–658 (2012).
29. Laurie, C. C. *et al.* Detectable clonal mosaicism from birth to old age and its relationship to cancer. *Nat. Genet.* **44**, 642–650 (2012).
30. Rodríguez-Santiago, B. *et al.* Mosaic uniparental disomies and aneuploidies as large structural variants of the human genome. *Am. J. Hum. Genet.* **87**, 129–138 (2010).
31. Haney, M. S. *et al.* Large-scale in vivo CRISPR screens identify SAGA complex members as a key regulators of HSC lineage commitment and aging. *bioRxiv* 2022.07.22.501030 (2022) doi:10.1101/2022.07.22.501030.
32. Zhao, M., Wang, Q., Wang, Q., Jia, P. & Zhao, Z. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics* **14 Suppl 11**, S1 (2013).
33. Bakker, B. *et al.* Single-cell sequencing reveals karyotype heterogeneity in murine and human malignancies. *Genome Biol.* **17**, 115 (2016).
34. van den Bos, H. *et al.* Single-cell whole genome sequencing reveals no evidence for common aneuploidy in normal and Alzheimer's disease neurons. *Genome Biol.* **17**, 116 (2016).
35. Szklarczyk, D. *et al.* The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.* **49**, D605–D612 (2021).
36. Sanders, A. D., Falconer, E., Hills, M., Spierings, D. C. J. & Lansdorp, P. M. Single-cell template strand sequencing by Strand-seq enables the characterization of individual homologs. *Nat. Protoc.* **12**, 1151–1176 (2017).
37. Sanders, A. D. *et al.* Single-cell analysis of structural variations and complex rearrangements with tri-channel processing. *Nat. Biotechnol.* **38**, 343–354 (2020).
38. Cosenza, M. R., Rodriguez-Martin, B. & Korbel, J. O. Structural Variation in Cancer: Role,

Prevalence, and Mechanisms. *Annu. Rev. Genomics Hum. Genet.* **23**, 123–152 (2022).

39. McClintock, B. The Stability of Broken Ends of Chromosomes in Zea Mays. *Genetics* **26**, 234–282 (1941).
40. Dillon, L. W., Burrow, A. A. & Wang, Y.-H. DNA instability at chromosomal fragile sites in cancer. *Curr. Genomics* **11**, 326–337 (2010).
41. Glover, T. W. & Stein, C. K. Induction of sister chromatid exchanges at common fragile sites. *Am. J. Hum. Genet.* **41**, 882–890 (1987).
42. Tang, Y.-C. & Amon, A. Gene copy-number alterations: a cost-benefit analysis. *Cell* **152**, 394–405 (2013).
43. Lu, Y. *et al.* Srebf1c preserves hematopoietic stem cell function and survival as a switch of mitochondrial metabolism. *Stem Cell Reports* **17**, 599–615 (2022).
44. Teresi, R. E., Planchon, S. M., Waite, K. A. & Eng, C. Regulation of the PTEN promoter by statins and SREBP. *Hum. Mol. Genet.* **17**, 919–928 (2008).
45. Kim, J. B. & Spiegelman, B. M. ADD1/SREBP1 promotes adipocyte differentiation and gene expression linked to fatty acid metabolism. *Genes Dev.* **10**, 1096–1107 (1996).
46. Rudko, O. I., Tretiakov, A. V., Naumova, E. A. & Klimov, E. A. Role of PPARs in Progression of Anxiety: Literature Analysis and Signaling Pathways Reconstruction. *PPAR Res.* **2020**, 8859017 (2020).
47. Mäkelä, J. *et al.* Peroxisome proliferator-activated receptor-γ (PPARγ) agonist is neuroprotective and stimulates PGC-1α expression and CREB phosphorylation in human dopaminergic neurons. *Neuropharmacology* **102**, 266–275 (2016).
48. Pich, O., Reyes-Salazar, I., Gonzalez-Perez, A. & Lopez-Bigas, N. Discovering the drivers of clonal hematopoiesis. *Nat. Commun.* **13**, 4267 (2022).
49. Anna, A. & Monika, G. Splicing mutations in human genetic disorders: examples, detection, and confirmation. *J. Appl. Genet.* **59**, 253–268 (2018).
50. Bennett, N. C., Gardiner, R. A., Hooper, J. D., Johnson, D. W. & Gobe, G. C. Molecular cell biology of androgen receptor signalling. *Int. J. Biochem. Cell Biol.* **42**, 813–827 (2010).
51. Jaiswal, S. *et al.* Age-related clonal hematopoiesis associated with adverse outcomes. *N. Engl. J.*

Med. **371**, 2488–2498 (2014).

52. Genovese, G. *et al.* Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N. Engl. J. Med.* **371**, 2477–2487 (2014).
53. Kessler, M. D. *et al.* Common and rare variant associations with clonal haematopoiesis phenotypes. *Nature* **612**, 301–309 (2022).
54. Forsberg, L. A., Gisselsson, D. & Dumanski, J. P. Mosaicism in health and disease - clones picking up speed. *Nat. Rev. Genet.* **18**, 128–142 (2017).
55. Forsberg, L. A. *et al.* Age-related somatic structural changes in the nuclear genome of human blood cells. *Am. J. Hum. Genet.* **90**, 217–228 (2012).