

# Convergent somatic evolution commences in utero in a germline ribosomopathy

---

Received: 4 October 2022

---

Accepted: 14 August 2023




---

Published online: 22 August 2023

---

 Check for updates

---

Heather E. Machado<sup>1,13</sup>, Nina F. Øbro<sup>2,3,4,13</sup>, Nicholas Williams<sup>1,13</sup>, Shengjiang Tan<sup>2,3,5,13</sup>, Ahmed Z. Boukerrou<sup>2,3,5</sup>, Megan Davies<sup>2,3</sup>, Miriam Belmonte<sup>2,3</sup>, Emily Mitchell<sup>1,2</sup>, E. Joanna Baxter<sup>3</sup>, Nicole Mende<sup>2,3</sup>, Anna Clay<sup>2,3</sup>, Philip Ancliff<sup>6</sup>, Jutta Köglmeier<sup>6</sup>, Sally B. Killick<sup>7</sup>, Austin Kulasekararaj<sup>8</sup>, Stefan Meyer<sup>9,10,11</sup>, Elisa Laurenti<sup>2,3</sup>, Peter J. Campbell<sup>1</sup>, David G. Kent<sup>2,3,12,14</sup> , Jyoti Nangalia<sup>1,2,3,14</sup>  & Alan J. Warren<sup>2,3,5,14</sup> 


---

Clonal tracking of cells using somatic mutations permits exploration of clonal dynamics in human disease. Here, we perform whole genome sequencing of 323 haematopoietic colonies from 10 individuals with the inherited ribosomopathy Shwachman-Diamond syndrome to reconstruct haematopoietic phylogenies. In ~30% of colonies, we identify mutually exclusive mutations in *TP53*, *EIF6*, *RPL5*, *RPL22*, *PRPF8*, plus chromosome 7 and 15 aberrations that increase *SBDS* and *EFL1* gene dosage, respectively. Target gene mutations commence in utero, resulting in a profusion of clonal expansions, with only a few haematopoietic stem cell lineages (mean 8, range 1-24) contributing ~50% of haematopoietic colonies across 8 individuals (range 4-100% clonality) by young adulthood. Rapid clonal expansion during disease transformation is associated with biallelic *TP53* mutations and increased mutation burden. Our study highlights how convergent somatic mutation of the p53-dependent nucleolar surveillance pathway offsets the deleterious effects of germline ribosomopathy but increases opportunity for *TP53*-mutated cancer evolution.

All cells acquire somatic mutations over time through a range of exogenous and endogenous DNA damaging processes. The tracking of such mutations has enabled the reconstruction of lineage histories of individual haematopoietic stem cells (HSC) to chart clonal dynamics

in healthy and malignant human haematopoiesis over life<sup>1-4</sup>. These studies have shown that some HSCs gain a fitness advantage over others, typically through acquisition of certain somatic mutations, resulting in slow but continuous clonal expansion over a lifetime<sup>3</sup>. By

---

<sup>1</sup>Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, UK. <sup>2</sup>Wellcome MRC Cambridge Stem Cell Institute, University of Cambridge, Cambridge, UK. <sup>3</sup>Department of Haematology, University of Cambridge, Cambridge, UK. <sup>4</sup>Department of Clinical Immunology, Copenhagen University Hospital, Rigshospitalet, Copenhagen, Denmark. <sup>5</sup>Cambridge Institute for Medical Research, Keith Peters Building, Cambridge, UK. <sup>6</sup>Department of Haematology, Great Ormond Street Hospital for Children NHS Foundation Trust, London, UK. <sup>7</sup>University Hospitals Dorset NHS Foundation Trust, The Royal Bournemouth Hospital, Bournemouth, UK. <sup>8</sup>Department of Haematological Medicine, King's College Hospital NHS Foundation Trust and King's College London, London, UK. <sup>9</sup>Division of Cancer Sciences, School of Medical Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester Cancer Research Centre, Wilmslow Road, Manchester, UK. <sup>10</sup>Department of Paediatric Haematology and Oncology, Royal Manchester Children's Hospital, Manchester Foundation Trust, Manchester, Oxford Road, Manchester, UK. <sup>11</sup>Teenage and Adolescent Oncology, The Christie NHS Foundation Trust, Wilmslow Road, Manchester, UK. <sup>12</sup>York Biomedical Research Institute, Department of Biology, University of York, York, UK. <sup>13</sup>These authors contributed equally: Heather E. Machado, Nina F. Øbro, Nicholas Williams, Shengjiang Tan. <sup>14</sup>These authors jointly supervised this work: David G. Kent, Jyoti Nangalia, Alan J. Warren.  e-mail: [david.kent@york.ac.uk](mailto:david.kent@york.ac.uk); [jn5@sanger.ac.uk](mailto:jn5@sanger.ac.uk); [ajw1000@cam.ac.uk](mailto:ajw1000@cam.ac.uk)

the 7th to 8th decade of life, there is a collapse in HSC clonal diversity in blood with many clonal expansions driven by mutations in a range of genes (e.g. *DNMT3A*) and copy number changes (e.g. loss of Y)<sup>3,5,6</sup>. Relatively little, however, is understood about how clonal selection and population dynamics differ in individuals born with germline mutations that compromise haematopoiesis and confer an increased risk of blood cancer.

Shwachman-Diamond syndrome (SDS) is an inherited ribosome assembly disorder caused by compound heterozygous germline mutations in the *SBDS* gene, typically the combination of one null and one hypomorphic allele<sup>7–9</sup>. The wild-type *SBDS* protein cooperates with the GTPase *EFL1* to catalyse release of the anti-association factor *eIF6* from the intersubunit face of the large ribosomal subunit to promote ribosome maturation and recycling<sup>8–12</sup>. The resulting ribosome assembly defect and reduced protein synthesis results in bone marrow failure (BMF), with over one-third of individuals subsequently developing myelodysplasia (MDS) and acute myeloid leukaemia (AML) by the fourth decade of life<sup>13,14</sup>.

A number of recurrent somatic genetic events have been identified in SDS. In individuals with one null and one hypomorphic *SBDS* allele on chromosome (chr) 7q, copy number neutral loss of heterozygosity (LOH) increases the gene dose of the hypomorphic *SBDS* allele c.258+2T>C and replaces the null allele<sup>15,16</sup>. Similarly, uniparental disomy can occur on chr15 to mitigate against the more damaging compound heterozygous *EFL1* mutation combinations in SDS<sup>17</sup>. Chr20q deletion and *EIF6* point mutations also reduce *eIF6* dosage and/or its affinity for the ribosome<sup>14,18–21</sup>. Each of these genetic events likely compensate for defective *SBDS* function in SDS by restoring ribosome homeostasis.

Impaired ribosome assembly stabilises the tumour suppressor protein p53 via the nucleolar surveillance pathway (NSP)<sup>22</sup>. Increased p53 expression is observed in haematopoietic cells from individuals with SDS<sup>23</sup> and targeted disruption of *Sbds* in murine models causes p53-dependent induction of apoptosis in haematopoietic progenitor cells<sup>24,25</sup>. Indeed, *TP53* mutations are recurrent across, and within, individuals with SDS<sup>18,26</sup>. Since *TP53* is the most frequently altered gene in human tumours<sup>27</sup>, with mutations arising both early in tumorigenesis, such as in glioblastoma and ovarian cancers<sup>28–30</sup>, as well as late during cancer progression<sup>28,31</sup>, it is critical to understand the impact of *TP53* mutations on cellular competition. SDS provides a unique window into understanding the earliest stages of *TP53*-mutated clonal selection due to the selective pressure imposed by the germline *SBDS* mutation.

In this study, we use whole-genome sequencing (WGS) of single-cell-derived haematopoietic colonies to interrogate the mutational consequences, selection landscape and clonal dynamics in the germline ribosomopathy, SDS. We show that *TP53* and the p53-dependent nucleolar stress pathway are a frequent target of mutually exclusive, convergent somatic mutation from early life, including in utero. These mutations drive early loss of clonal diversity that while offsetting the deleterious effects of defective ribosome assembly, increases the propensity for *TP53*-mutated cancer evolution.

## Results

### Premature and marked loss of haematopoietic clonal diversity in SDS

We studied ten individuals with SDS aged 4–33 years, who harboured biallelic germline loss-of-function mutations in the *SBDS* gene. We isolated single haematopoietic stem and progenitor cells (HSPC) and mononuclear cells (MNCs) from peripheral blood or bone marrow from individuals, and following whole-genome sequencing of single-cell-derived colonies ( $n = 323$ ), we used the somatic mutations to reconstruct haematopoietic phylogenies, using published methodology<sup>4</sup> (Fig. 1a, b). Individuals had typical clinical features of SDS including neutropenia, pancreatic insufficiency and osteopenia,

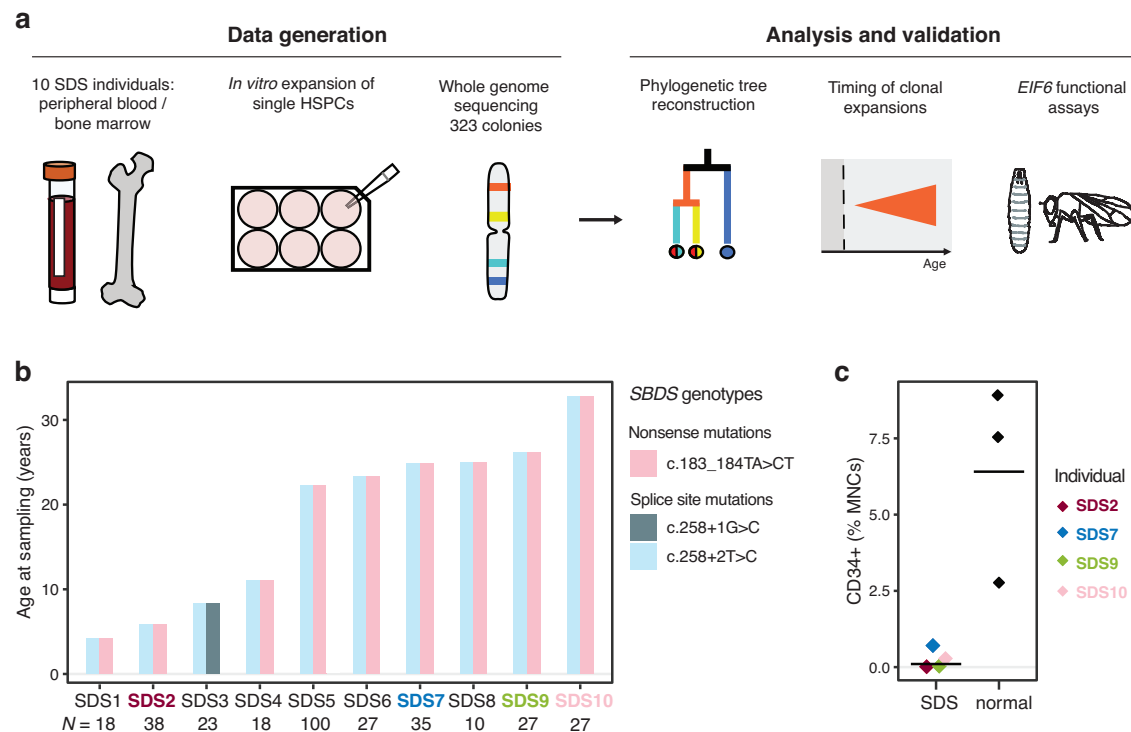
presenting early in life with failure to thrive. Histomorphology revealed bone marrow hypocellularity (range 10–40%), dyserythropoiesis, and decreased granulopoiesis with a reversed myeloid-erythroid ratio (1:3–4). One individual (SDS8) had progressed to MDS with trilineage dysplasia shortly before sampling. Flow cytometric phenotyping of bone marrow (BM) and peripheral blood (PB) mononuclear cells showed a reduced frequency of total CD34<sup>+</sup> progenitors in individuals with SDS (median 0.2%) compared to healthy bone marrow donors (median 7.5%) ( $t$ -test  $p = 0.01$ ; Fig. 1c and Supplementary Fig. 1), consistent with previous studies<sup>32</sup>. We undertook WGS of 323 individual single-cell-derived colonies seeded from haematopoietic stem and progenitor cells (HSPC), to a mean depth of 20x reads, together with matched buccal swab DNA as a germline reference, in all individuals. Somatic mutations together with embryonic variants were identified through a combination of mutation calling using a matched germline reference as well as an unmatched variant calling approach, as detailed in the methods. Haematopoietic colonies were clonally derived with a somatic mutation mean variant allele fraction (VAF) of >0.4, thus representing the somatic mutations present in the single-cell that seeded the colony. In total, we identified 118,564 single nucleotide variants, 6287 small insertions and deletions, 74 structural variants and 5 chromosomal copy number aberrations across the cohort (Supplementary Dataset 1). The number of SNVs per colony per individual ranged from a median of 130 (range 99–163) in the youngest (age 4 years) to a median of 714 (range 593–744) for one of the older individuals (age 25 years) with MDS (SDS8).

Somatic mutations from individual colonies were used to reconstruct phylogenetic trees of haematopoiesis (Fig. 2). We identified a profusion of clonal expansions in 7 of 10 individuals (SDS2, SDS4, SDS5, SDS6, SDS7, SDS8 and SDS10), an observation highly uncharacteristic of healthy haematopoiesis in individuals <70 years of age or individuals with blood cancers studied to date<sup>1,3,4</sup>. Given that by birth, blood cells typically have already acquired around 50–65 somatic mutations<sup>3</sup>, we defined post embryonic clonal expansions as any clade comprising  $\geq 2$  colonies, whose common ancestor was observed after 75 mutations from the start of the phylogenetic trees. We identified 18 such clonal expansions of varying sizes across the trees, representing 21% of colonies (Supplementary Fig. 2 and Supplementary Dataset 2).

### Somatic mutations under selection in individuals with SDS

The congenital ribosomopathy SDS provides strong selective pressure for the expansion of HSCs that have accumulated fitness-enhancing somatic mutations<sup>14,15,17–21,26</sup>. We observed several genomic events that directly target *SBDS*, identifying four instances of chr7q LOH, each resulting in an extra copy of the c.258+2T>C donor splice site mutant hypomorphic *SBDS* allele (SDS5, 4, 7, Fig. 2) and one somatically acquired nonsynonymous *SBDS* mutation, also occurring on the hypomorphic allele (SDS10, Fig. 2). We also identified a chr15 event (15q24–26 tetra) that doubles the copy number of the *EFL1* gene located at 15q25.2. These somatic events appear to be directly compensating for the germline defect that impairs the cooperation between *SBDS* and *EFL1* that is required for ribosome maturation<sup>8</sup>.

More commonly, we observed frequent and independently acquired somatic mutations affecting five other genes. Three of these genes have been reported as recurrently mutated in SDS (*PRPF8*, *TP53*, *EIF6*)<sup>18,21,26</sup>. In addition, we identified somatic nonsynonymous mutations under positive selection in the *RPL5* and *RPL22* genes (ratio of normalised nonsynonymous (dN) to normalised synonymous mutations (dS)  $dN:dS > 1$ ,  $q < 0.01$ ) (Figs. 2 and 3a), both encoding protein components of the large ribosomal subunit. Overall, we identified 24 independent missense mutations in *TP53* (Supplementary Fig. 3), by far the most commonly mutated gene in the cohort, and 1 start codon loss, 4 missense, 2 nonsense, 1 frameshift mutation and 5 gene deletions in *EIF6*. The somatic mutations in *RPL22* suggested loss of



**Fig. 1 | Study design and cohort. a** Schematic of experimental design. Single haematopoietic stem and progenitor cells (HSPC) and mononuclear cells (MNCs) from peripheral blood or bone marrow from individuals with SDS were expanded into colonies *in vitro* and each colony underwent whole-genome sequencing (WGS). Somatic mutations were used to reconstruct haematopoietic phylogenies. The timing of acquisition, clonal dynamics and functional consequences were investigated for driver mutations associated with SDS. Inkscape. **b** Age at sampling and *SBDS* genotype for each individual with SDS. The two bi-coloured columns

represent the two parental alleles, with all individuals having biallelic germline mutations in *SBDS*. Highlighted samples (SDS2, SDS7, SDS9 and SDS10) were measured for frequency of CD34+ HSPCs, shown in (c). *N* the number of haematopoietic colonies analysed per individual. **c** Frequency of CD34+ HSPCs in bone marrow samples (expressed as a % of total viable MNCs) was analysed by flow cytometry in four of the individuals with SDS and three healthy individuals (black). Source data are provided as a Source Data file.

function (1 start codon loss, 1 nonsense, 2 splice site, 1 missense and 2 in frame deletions), while *RPL5* and *PRPF8* mutations were missense SNVs ( $n=4$  and 2 respectively). Mutations in *TP53*, *EIF6*, *RPL5* and *RPL22* genes were recurrent within the same individual, with 9 different *TP53* mutations observed in SDS5 and 5 independent *EIF6* mutations occurring in SDS7 (Fig. 2). In addition to recurrent mutations in *PRPF8*, *TP53*, *EIF6*, *RPL5* and *RPL22*, we observed mutations in *DNMT3A*, *ASXL1*, *TET2* and *RUNX1* associated with clonal haematopoiesis (CH), consistent with the study by Kennedy et al.<sup>18</sup>. We term recurrent mutations in either SDS-associated or known genes of CH<sup>33,34</sup>/haematological cancers<sup>35</sup> as driver mutations (see “Methods”). Across all the individuals in this study who had a mean age of only 18 years (4–33 years range), 31% of colonies (101/323) harboured a driver mutation (Fig. 2 and Supplementary Fig. 2).

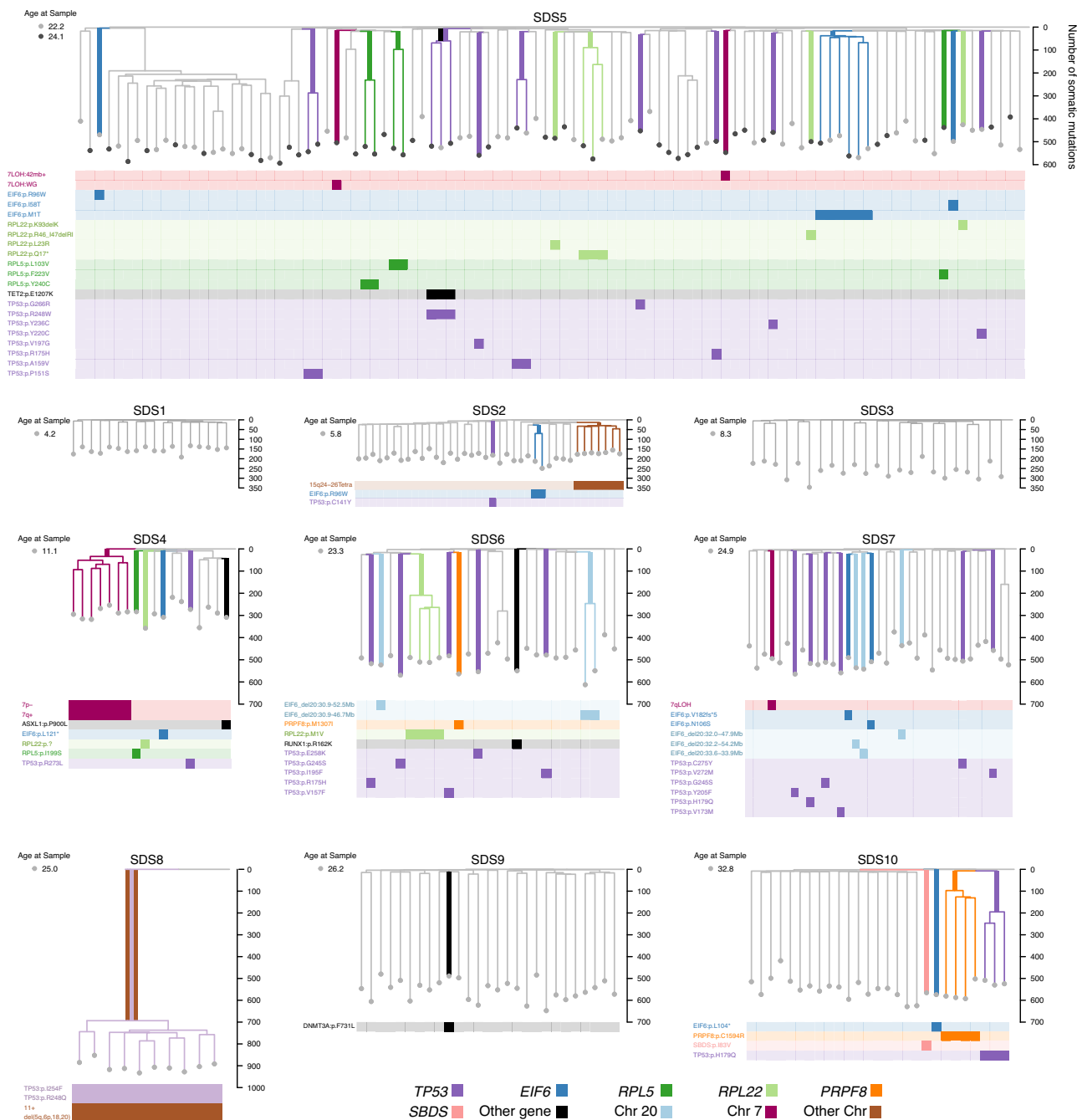
There was heterogeneity in both the frequency of driver mutations and the degree of clonal expansion across individuals. No driver mutations were identified in 67 of 68 haematopoietic colonies from three individuals (SDS1, SDS3 and SDS9). In contrast, 46% of the remaining 255 colonies from 7 individuals harboured driver mutations (median 48%, range: 26–100%; Figs. 2 and 3b). The three individuals with a sparsity of driver mutations also lacked detectable clonal expansions (SDS1, SDS3 and SDS9). We did not observe any correlation between the prevalence of clonal expansions or driver mutations and peripheral blood count cytopenias (Supplementary Table 1).

We explored clonal expansions lacking driver mutations as well as non-expanded branches for somatic mutations in additional putative target genes. We observed a large embryonic clonal expansion in SDS5 comprising 21 colonies which harboured a chromosomal translocation affecting *GPR137B*, coding for an mTORC1 regulatory protein.

Nonsynonymous somatic variants were also identified in genes involved in translation (*EIF4A1* and *EIF5A1*), RNA metabolism (*DDX23*, *DDX42*, *DDX60*, and *DDX39B*), and ribosomal proteins (*RPS14* and *RPS21*) (Supplementary Dataset 1). Since these mutations were only observed in single colonies, their potential pathogenicity remains unclear.

Apart from one individual (SDS8) with clonal evolution to biallelic *TP53* mutations and MDS transformation, and one individual (SDS5) with a concurrent *TET2* mutation within a *TP53*-mutated clade, we did not observe any instances where more than one driver mutation was present within the same lineage (Fig. 2). Colonies harbouring copy number alterations that compensated for *SBDS* or *EFL1* dosage (SDS2, SDS4, SDS5 and SDS7) were also mutually exclusive with colonies harbouring nucleotide substitutions in driver genes. This suggests that a single heterozygous mutation in one of several target genes is sufficient to provide a fitness advantage in the context of the germline ribosome assembly defect.

In total, 131 of 323 colonies (41%) were either in an expanded lineage and/or harboured a driver mutation (median across individuals 37%, range 0–100%; Supplementary Fig. 2). Excluding the 2 young individuals without clonal expansions or driver mutations (SDS1 and 3), a median of ~50% of haematopoietic colonies in individuals harboured a driver mutation or were part of expanded lineages (range 4–100%). Assuming that single colonies with driver mutations also represent small clonal expansions, on average, 8 expanded HSC lineages (range 1–24 HSC lineage expansions across 8 individuals) were producing half of the haematopoietic cells sampled in these individuals. The oligoclonality in young individuals with SDS is in stark contrast to healthy/non-SDS haematopoiesis,



**Fig. 2 | Recurrent mutations across SDS haematopoietic phylogenies.** Phylogenetic trees of haematopoietic colonies for ten individuals with SDS. Each individual had between 10–100 colonies sequenced and included for phylogenetic analysis. Branches with somatic mutations in driver genes previously reported and/or under positive selection in this study are coloured. Branches with known driver mutations of clonal haematopoiesis are shown in black, and those associated with SDS in other colours. The branch harbouring the driver mutation is shown with a thicker coloured line. The Y-axis shows the total number of somatic mutations

including driver mutations. Rows beneath phylogenetic trees show the specific driver mutations, with colonies harbouring that mutation more densely coloured. Of note, no known driver mutations were detected amongst the somatic mutations found in SDS1 and SDS3. \*SDS8 was diagnosed with transformation to myelodysplasia with trilineage dysplasia 1 month before sampling. All SDS8 colonies had a complex karyotype with many chromosomal (Chr) copy number aberrations (CNA). CNAs confidently shared across SDS8 colonies are shown on the tree.

where such a degree of clonal expansion is not observed until after the age of 70 years<sup>3</sup>.

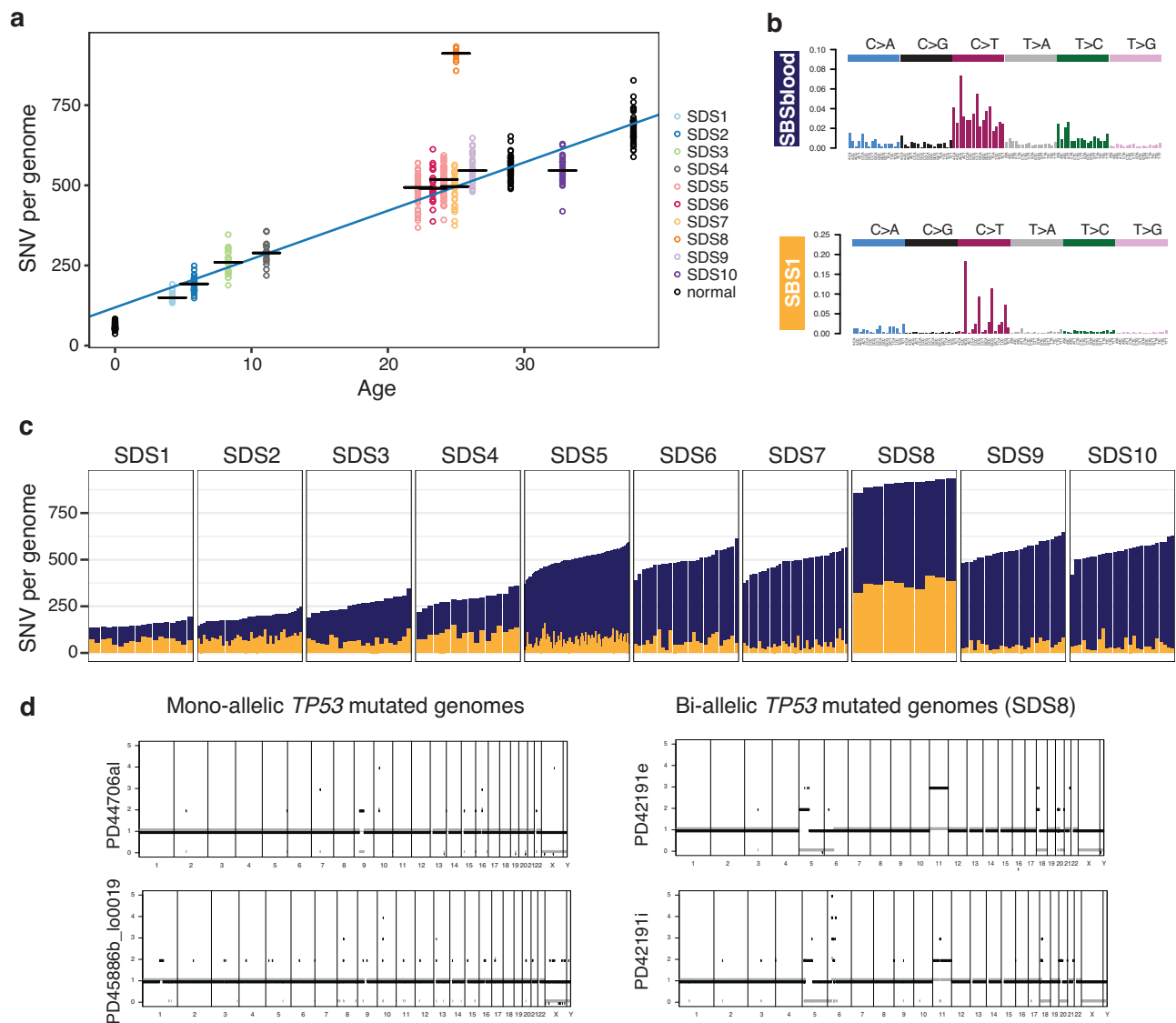
**Timing haematopoietic clonal expansions in individuals with SDS**

Due to the linear acquisition of somatic mutations over time, we can use the phylogenetic trees to estimate when clonal expansions driven by driver mutations commenced in life. This is possible by converting

the number of somatic mutations acquired by the most recent common ancestor of a clonal expansion that shares a driver mutation to chronological age (see methods). We timed the start of 14 different clonal expansions driven by mutations in *TP53*, *RPL5*, *RPL22*, *PRPF8*, *EIF6*, as well as copy number events affecting *SBDS* and *EIF6* (Fig. 3c). They exhibit a range of clonal expansion times over life, commencing from early in utero up to age 12. With the exception of the copy number events affecting *SBDS* and *EFL1* on chromosomes 7 and 15,





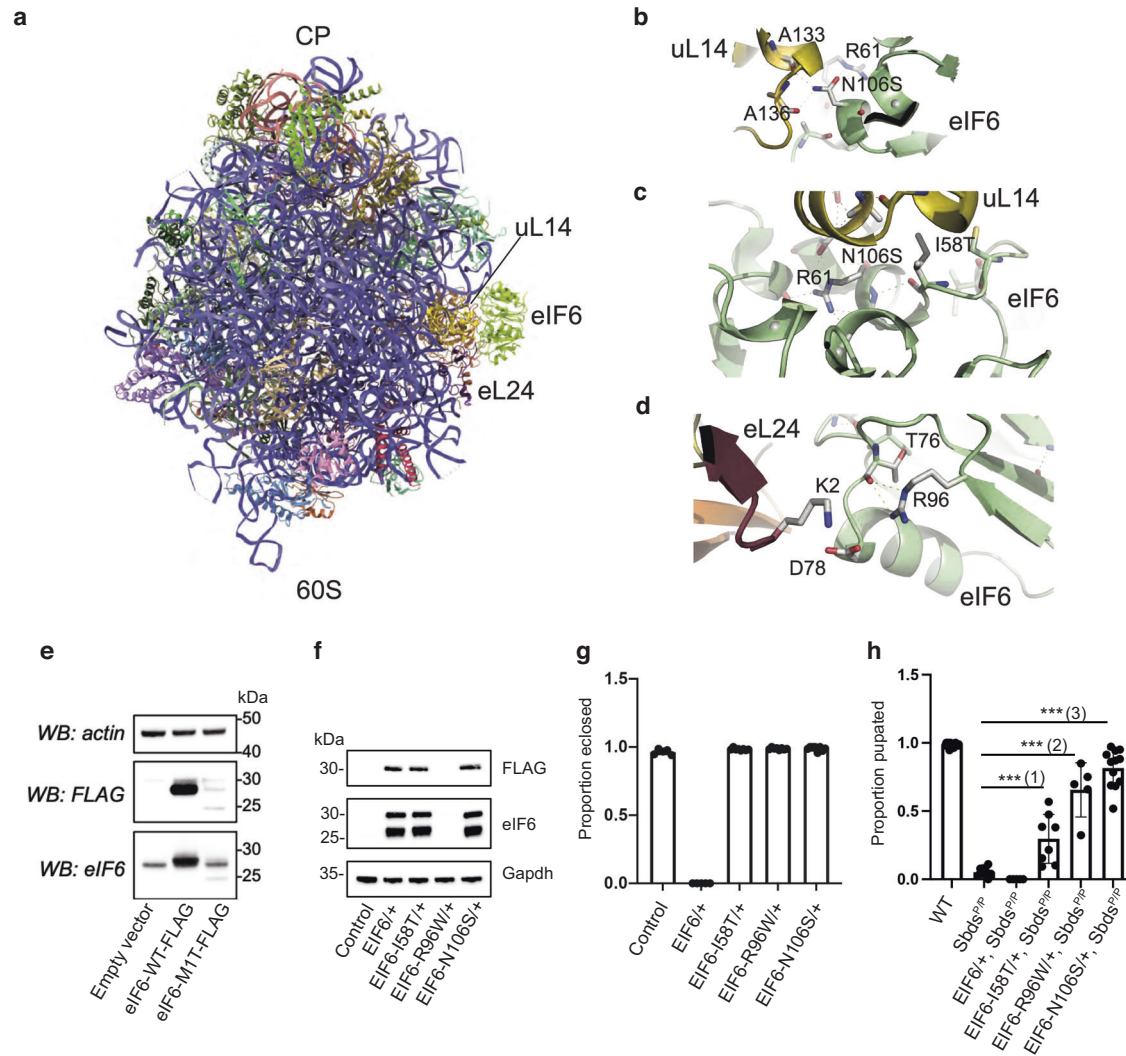


**Fig. 4 | Mutation burden, mutational processes and biallelic *TP53* variants.** **a** Mutation burden (number of SNVs) as a function of age for ten individuals with SDS. Each circle represents one colony genome, with the black horizontal bars representing the median burden per individual. Two timepoints from SDS5 are shown at different ages. Circles coloured black (normal) representing mutation burdens from three haematopoietically healthy (non-SDS) individuals (published data<sup>3</sup>) are shown for comparison. The blue line represents the regression line through the colonies from individuals with SDS. **b** Trinucleotide context of somatic mutations. The two mutational signatures were identified across all genomes. SBS1<sup>42,43</sup> is characterised by spontaneous deamination of cytosines, and the second

mutational signature, termed SBSblood<sup>1,2,41</sup> represents mutations typical of endogenous mutations in HSCs. **c** Number of SNVs attributable to the mutational signatures SBS1 (green) and SBSblood (blue) across each colony from each individual with SDS. Each bar represents the genome from one colony. Note SDS8 has a higher total mutation burden due to increased SBS1 mutations. **d** Copy number variation for two representative colony genomes with heterozygous/mono-allelic *TP53* mutation (from different individuals) and two clonally related colonies from SDS8 with biallelic *TP53* mutations. Ploidy is shown on the y-axis and genomic location on the x-axis, for the two parental alleles (green and red). CNA copy number aberration. Source data are provided as a Source Data file.

in the study, albeit with a mild increase in SBS1 (27% of shared mutations) (Supplementary Fig. 4a, b). This raised the possibility that the transition from normal mutation acquisition or cell division rates in SDS8 to rapid clonal expansion occurred at some point along the shared trunk of the SDS8 phylogeny. In order to estimate when along this shared branch such a transition may have occurred, we decomposed the mutational spectrum of the shared branch into the sum of two mutational profiles—the mutation profile of SDS haematopoiesis observed in other SDS individuals of a similar age (composite age-matched SDS signature) and the mutational spectrum in the private end branches of SDS8 (transformation signature) as described in “Methods”. The combination of these two profiles accurately reconstructed the mutation spectrum of the shared trunk (composite age-matched SDS

signature 0.80, transformation signature 0.20, cosine similarity 0.958, methods, Supplementary Fig. 4c). This provides a rough estimate for when rapid growth may have commenced, assuming this occurred at a single time point historically and suggested a very recent age of transformation of 23.5 years (95% CI 19.2–24.9). Even the lower bound of this age range implies rapid clonal outgrowth. Assuming a very simple model of a single clone expanding at a constant rate to clonal dominance (see “Methods”), it would suggest that this clone was growing by 5200% (150%–15,000%) per year, corresponding to the mutant HSC clone size doubling roughly every 2 months. These data highlight the potential rapidity of transformation to MDS in this individual with SDS and provide a potential explanation for the often abrupt disease progression that may be observed clinically. However, it is important to



**Fig. 5 | Functional consequences of *EIF6* variants.** **a** Atomic model of human eIF6 bound to the 60S ribosomal subunit (PDBID: 7OW7). CP central protuberance. Stabilising interactions formed by eIF6 residues N106 (**b**), I58 (**c**) and R96 (**d**) are predicted to be lost with somatic mutation. Figures were generated using Pymol v1.2. eL24 is coloured salmon; uL14, gold; eIF6, green. **e** Cell extracts from HEK293T cells expressing empty vector, human eIF6-WT-FLAG or eIF6-MIT-FLAG mutant for 24 h were immunoblotted to detect eIF6, FLAG and actin as loading control. **f**, **g** Overexpression of eIF6 variants in WT flies. Genotypes of fly samples are indicated in Supplementary Table 2. **f** Extracts from larvae with the stated

genotypes were immunoblotted to detect the indicated proteins (minimum 3 replicates). Control, da-GAL4 line. **g** Proportion of indicated fly genotypes that enclosed (minimum 5 replicates, minimum  $n = 216$ ; error bars represent mean  $\pm$  SD). **h** Genetic complementation of *Sbds*-deficient *Drosophila* (*Sbds*<sup>P/P</sup>) with SDS-related eIF6 variants. Proportion of indicated genotypes developing to the pupal stage is shown (minimum 5 replicates, minimum  $n = 256$ ; error bars represent mean  $\pm$  SD; \*\*\*two tailed *t*-test,  $p(1) = 0.00060711$ ;  $p(2) = 2.56426E-07$ ;  $p(3) = 2.3141E-13$ . Source data are provided as a Source Data file.

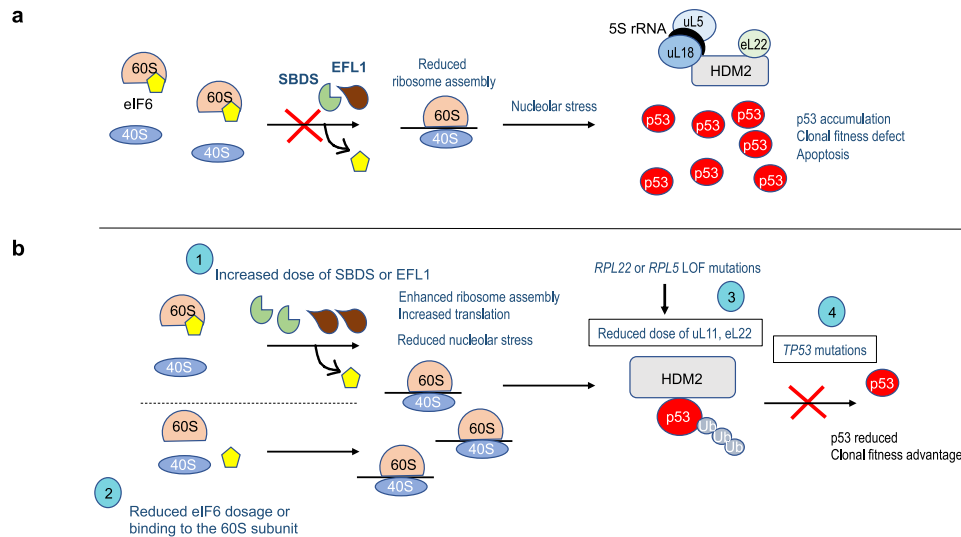
note that we have only characterised transformation to MDS in a single individual. Analysis of further individuals is required to confidently estimate the trajectory to disease transformation in SDS. Of interest, we found no evidence that heterozygous *TP53*-mutated colonies across the cohort increased mutation burden (linear mixed model  $p = 0.22$ , all comparisons with *TP53*: Tukey  $p \geq 0.87$ ) suggesting that the rapid clonal expansion and copy number aberrations observed in SDS8 were driven by biallelic mutation of *TP53* and/or the chromosomal aberrations present.

### Functional impact of *EIF6* mutations

Mutations in the *EIF6* gene confer a fitness advantage to SBDS-deficient cells<sup>8,18,21,44</sup>. Of the 13 *EIF6* mutational events identified, start-codon loss (MIT), missense mutations (I58T, R96W, N106S), nonsense mutations (L121\*, L104\*), frameshift truncating mutations (V182fs\*5), and deletions were observed. While R96W and MIT were associated with clonal expansions, the remaining events affected single colonies. The

differences both in somatic variants and clone size suggest variable functional effects of individual *EIF6* mutations.

To study the functional consequences of eIF6 mutations on ribosome assembly, we mapped residues I58, N106 and R96 to the 2.4 Å cryo-EM structure of the human eIF6-60S complex (PDBID: 7OW7) (Fig. 5a–d). The eIF6 residue N106 lies at the interface between eIF6 and the 60S ribosomal subunit protein uL14, forming hydrogen (H)-bonding interactions with the backbone oxygen atoms of uL14 residues A133 and A136 (Fig. 5b). N106S likely reduces the hydrogen bonding interface between eIF6 and uL14 to aid its dissociation<sup>21</sup>. Similarly, the side chain of residue I58 forms hydrophobic interactions with uL14, while the main chain oxygen of I58 forms an intra-protein H-bond with the main chain nitrogen of R61 which in turn forms a series of intra-protein H-bonds with the side chain and backbone atoms of N106 (Fig. 5c). Replacement of isoleucine with the more polar threonine side chain in the I58T variant may increase solvation and reduce the stability of the eIF6-uL14-interface. Indeed, eIF6 variants I58T and



**Fig. 6 | TP53 and the nucleolar surveillance pathway as targets for convergent somatic mutation. a** Defective germline ribosome assembly in SDS promotes nucleolar stress through inhibitory binding of the 5S RNP complex (consisting of the 5S rRNA, uL5, encoded by *RPL11* and uL18, encoded by *RPL5*) to the nuclear E3 ligase HDM2 (enhanced by eL22, encoded by *RPL22*), promoting p53 accumulation and apoptosis<sup>22</sup>. **b** Convergent evolution of somatic mutations restores ribosome

homeostasis, favouring HDM2-dependent p53 ubiquitination and degradation, through multiple independent somatic genetic rescue events including: (1) increased dose of SBDS or EFL1 proteins (2) reduced eIF6 dosage or eIF6 binding to the 60S subunit; (3) disrupted inhibitory binding of HDM2 to p53 through mutations in *RPL5* and *RPL22*<sup>50</sup>; (4) *TP53* mutations. Ub ubiquitin.

NI06S reduce the affinity of eIF6 for the 60S subunit across yeast, *Dictyostelium* and human cells<sup>8,21</sup>. Similarly, the side chain of R96 stabilises the polar interaction between eIF6 (residue D78) and ribosomal protein eL24 (residue K2) which is likely lost by the replacement of arginine with tryptophan in the R96W variant (Fig. 5d).

To assess expression of the eIF6 MIT variant, we performed immunoblotting of extracts from human HEK293T cells expressing wild-type (WT) or FLAG-tagged eIF6-MIT protein. We confirmed that the start codon loss variant eIF6-MIT significantly reduced eIF6 expression as anticipated (Fig. 5e), further indicating that a subset of *EIF6* missense variants reduces the dose of eIF6<sup>18,21</sup>. Nonsense (L104\*, L121\*) variants, deletion-causing frameshift mutations, or genomic eIF6 deletions, would also be expected to reduce eIF6 dosage due to *EIF6* haploinsufficiency. Immunoblotting of *Drosophila* larval cell extracts revealed that expression of the eIF6-I58T and eIF6-NI06S mutants was comparable to eIF6-WT, but expression of the eIF6-R96W variant was reduced (as detected by anti-FLAG antibody) (Fig. 5f). Total eIF6 expression (FLAG-tagged variant plus endogenous eIF6 protein), as detected by anti-eIF6 antiserum, was comparable for eIF6-WT, eIF6-I58T or eIF6-NI06S variants, but undetectable for eIF6-R96W (Fig. 5f below). These data indicate that transgenic overexpression of eIF6 WT or variants does not significantly induce expression of the endogenous eIF6 protein. Importantly, overexpression of WT eIF6 but not the eIF6 variants (NI06S, R96W and I58T), reduces the viability of WT flies (Fig. 5g), further demonstrating that overexpression of eIF6 variants does not result in functionally significant induction of the endogenous *Drosophila* eIF6 protein.

We next tested the ability of the eIF6-I58T, eIF6-R96W and eIF6-NI06S mutants to rescue the larval lethality of SBDS-deficient (*Sbds*<sup>P/P</sup>) *Drosophila*<sup>21</sup> compared to WT eIF6. Homozygous *Sbds*-deficient flies exhibited a severe growth defect, with only 5% of larvae surviving to the early pupal stage (Fig. 5h). While wild-type eIF6 failed to rescue the lethal *Sbds*-deficient phenotype, eIF6-R96W, eIF6-NI06S, and to a lesser extent, eIF6-I58T, rescued a proportion of flies that survived to the late pupal stage (Fig. 5h). Taken together with previous genetic experiments in yeast<sup>5</sup>, our data suggest that different eIF6 variants have significant but variable cellular rescue potency in SDS. We conclude that somatic *EIF6* mutations compensate for the ribosome

maturation defect in SDS either by reducing the *EIF6* gene copy number present in the cell, or by modulating the level of eIF6 protein or its 60S subunit binding function, resulting in at least partial restoration of ribosome homeostasis in SDS.

## Discussion

In this study, we have shown that the strong selective pressure to overcome impaired ribosome biogenesis and avoid p53-mediated cell death in SDS results in convergent somatic mutations in a unique set of target genes, not observed in the context of ageing<sup>45–47</sup> or haematological perturbations such as autoimmunity<sup>48</sup> or chemotherapy<sup>49</sup>. The survival advantage conferred by such mutations results in HSC clonal expansions, often commencing very early in life, even in utero, in individuals with SDS. By young adulthood and even childhood, we estimate that nearly half of haematopoiesis is derived from a small number of expanded HSCs. This is in stark contrast to haematopoiesis in healthy individuals who do not develop comparable oligoclonality until the final decades of life<sup>3,5</sup>.

We posit that the repertoire of target gene mutations under selective pressure in SDS highlights several routes to improved clonal fitness (Fig. 6a, b). Mutations may be directly compensatory by increasing the gene dosage of either *SBDS* or *EFL1* (Fig. 6b, '1'). The increase in *EFL1* gene copy number due to the structural chromosome 15 aberration observed in this study (in the context of germline *SBDS* mutations) is distinct from the somatic *EFL1* copy number changes due to uniparental disomy that have been reported in SDS cases caused by germline *EFL1* mutations<sup>17</sup>. Adaptive somatic mutations that reduce the *EIF6* gene copy number, reduce the level of eIF6 protein or alter its 60S subunit binding activity, may lower the requirement for functional SBDS and EFL1 to release eIF6 from the 60S ribosomal subunit (Fig. 6b, '2'). Each of these routes might be expected to help restore ribosome homeostasis. However, it is important to note that this improved fitness is in the context of the SDS disease state and although somatic *SBDS* gene dosage may be increased, this involves a hypomorphic variant allele that is expected to be only partially restorative. Similarly, alteration of *EIF6* gene copy or alteration of function is restorative, but in context of *SBDS* deficiency. Thus, neither restoration, nor other noted changes would



necessarily rescue all the functional deficiencies of SDS haematopoietic cells.

Defective ribosome assembly activates p53-induced apoptosis via the nucleolar surveillance pathway. Ribosomal proteins uL18 (encoded by *RPL5*) and uL5 (encoded by *RPL11*) bind to the 5S RNA to form the pre-ribosomal 5S ribonucleoprotein (5S-RNP) complex that inhibits the E3 ubiquitin ligase HDM2 to stabilise the p53 protein<sup>22</sup> (Fig. 6a). eL22 (encoded by *RPL22*) may also inhibit the HDM2-p53 circuit<sup>50</sup>. Thus, somatic mutations in regulators of the nucleolar signalling pathway such as *RPL5* and *RPL22* may disrupt nucleolar stress-induced stabilisation of p53 (Fig. 6c, '3') or mutations in *TP53* itself may allow cells to survive despite ongoing impairment of protein synthesis (Fig. 6b, '4'). The identification of recurrent mutations in *RPL5* and *RPL22* in this study may reflect the use of whole-genome sequencing versus the exome sequencing approach used by Kennedy et al.<sup>18</sup>. Although more speculative, mutations in the evolutionarily conserved splicing factor *PRPF8* may also disrupt the splicing of components of the p53-HDM2 axis including uL18, uL5 or p53 itself<sup>51</sup>. The recurrent *CSNK1A1* mutations identified by Kennedy et al.<sup>18</sup> (but not in this study) may potentially reflect the role of casein kinase in 40S ribosomal subunit maturation<sup>52</sup>.

Somatic mutation facilitates driver mutation entry, providing a substrate for clonal selection. Estimates of HSC number in healthy humans suggest that 50,000–200,000 uniquely identifiable and actively contributing HSCs are present in adulthood<sup>13</sup>. With individual HSCs accumulating ~15 somatic mutations/year<sup>13</sup>, one would expect ~1–3 million somatic mutations to enter the HSC pool per year, of which ~10,000–30,000 would be expected to land in coding sequence every year. This number of expected mutations is still greater than the coding footprint of many genes, making it plausible that a non-synonymous somatic mutation could land in a single gene, such as *TP53* (CDS length 1182 bp), in one HSC within the stem cell pool every year. Thus, opportunities for stochastic somatic mutation of genes in HSCs are very likely to be significantly higher than appreciated, resulting in extensive somatic mosaicism within our HSC pool. This may explain the high prevalence and recurrence of driver mutations in individuals with SDS when there is strong selection facilitating their clonal expansion post acquisition. An interesting hypothesis to test is whether, the numbers of somatic mutations observed in SDS cases at young age might, at least in part, reflect the inability of the immune system in SDS to clear rogue cells with driver mutations.

The stochastic nature of somatic driver mutation acquisition, from early life, may also explain the considerable heterogeneity in clinical phenotype, even amongst siblings with SDS with the same germline genotype<sup>13</sup>. Evidence of strong clonal selection was not captured in all individuals with SDS, as three individuals in the cohort did not harbour detectable clonal expansions or driver mutations (Fig. 2). SDS1 was the youngest individual in our cohort (4.2 years) and thus may have had less time to acquire driver mutations and clonal expansions. Alternatively, clonal expansions may have been missed due to a combination of the small sample size ( $n=18$  colonies) and greater HSC clonal diversity expected in younger individuals<sup>3</sup>. SDS3 was the only individual without the *SBDS* c.183\_184TA>CT allele, instead carrying two *SBDS* germline mutations (c.258+2T>C, c.258+1G>C) that disrupt the intron 2 donor splice site<sup>7</sup>. SDS9 was one of the older individuals in our cohort (26.2 years) and similar individuals lacking driver mutations were also observed by Kennedy et al.<sup>18</sup>. In future, it will be interesting to determine whether individuals who progress to marrow aplasia represent a specific subset of SDS disease evolution where driver mutation mediated clonal expansion has been insufficient to mitigate the bone marrow failure phenotype.

Although our study is limited by the small cohort of individuals included and the lack of longitudinal sampling, technologies such as single molecule sequencing<sup>39,53</sup> will bypass the requirement for clonal expansions for the detection of driver mutations which may help

elucidate the complete spectrum of target genes that provide fitness in SDS. Characterisation of different congenital bone marrow failure disorders may also help us understand the nature of the selective advantage provided by driver mutations associated with clonal haematopoiesis occasionally observed in this and an earlier study<sup>18</sup>.

Clinically, individuals with SDS merit close monitoring of emergent clones via regular extended gene sequencing of blood, given the large number of monoallelic *TP53* clonal expansions that many have, each potentially serving as a substrate for clonal evolution to aggressive disease. Our study and others<sup>18</sup> suggest that a key mechanism of transformation in SDS is the acquisition of biallelic mutated *TP53*. Given the very rapid clonal outgrowth and genomic evolution observed, consideration for early therapeutic intervention, such as bone marrow transplantation, may be warranted given the poor prognosis associated with *TP53*-mutated myeloid cancers<sup>35,54–58</sup> and transformed disease in SDS<sup>36</sup>.

## Methods

### SDS samples

Our research complies with all relevant ethical regulations. Individuals with SDS ( $n=10$ ) were prospectively involved in the study following full Research Ethics Committee approval and consent (NHS Research Ethics Committee approvals 07/MRE05/44 (Cambridge South), 11/LO/0512 (London Riverside), 12/EE/0478 (East of England)). Each individual was sampled at one time point, with the exception of SDS5, who was sampled at two time points. Material included peripheral blood and/or bone marrow and buccal swabs for each individual. Sample collections, initial sample processing and sample banking was performed by the Cambridge Blood and Stem Cell Biobank with appropriate NHS Research Ethics committee approval (18/EE/0199 (East of England)). Patients provided written informed consent to use the materials for the research undertaken here and publish the results without compensation.

### HSPC phenotyping

Flow cytometric immunophenotyping of HSPCs was done on stored frozen viable mononuclear cells (MNCs) from PB and/or BM samples obtained from individuals with SDS aged 4–33 years or from healthy/non-SDS donors aged 29–32 years (STEMCELL Technologies). MNCs from healthy (non-SDS) donors were stained with antibodies: CD3-FITC (clone HIT3a, BD #555339; dilution 1:500), CD90-PE (clone 5E10, Biolegend, #328114; 1:33), CD49f-PECy5 (clone GoH3, BD #551129; 1:100), CD38-PECy7 (clone HIT2, Biolegend, #303516; 1:100), CD33-APC (clone WM53, BD #571817; 1:200), CD19-A700 (clone HIB19, Biolegend #302226; 1:300), CD34-APCCy7 (clone 581, Biolegend #343514; 1:100), CD45RA-BV421 (clone HII100, Biolegend #304130; 1:100), and Zombie Aqua (Biolegend #423101; 1:2000). MNCs from individuals with SDS were stained with the following antibodies: CD38-FITC (clone HIT2, BD #555459; 1:12.5), CD34-PE-Cy7 (clone 8G12, BD #348811; 1:33), CD10-BV605 (clone HII10a, Biolegend #312222; 1:33), CD45RA-V450 (clone HI30, BD #560367; 1:100), CD90 APC (Clone 5E10, Biolegend #328110; 1:50), CD3-APC-Cy7 (clone SK7, Biolegend #344818; 1:50), and CD19-APC-Cy7 (clone HIB19, Biolegend #302218, 1:50). After gating for live singlets (7AAD or Zombie negative) and excluding CD3/CD19 positive cells, bulk CD34 positive progenitors were gated (Supplementary Fig. 1).

### In vitro expansion of haematopoietic colonies from mononuclear cells or HSPCs

Previous studies have shown that mutant clonal fractions are equivalent when stem cells or progenitors are sourced from peripheral blood or bone marrow<sup>1,3</sup>. PB or BM samples were collected in Lithium-Heparin (LiHep) tubes, MNCs were isolated by density gradient centrifugation, and RBCs were lysed in NH<sub>4</sub>Cl. Cells from the MNC fraction (or CD34<sup>+</sup> cells for one individual) were plated into MethoCult H4435

(STEMCELL) for in vitro culture and clonal expansion (colony-forming cell (CFC) assay) using a wide range of cell dilutions in order to ensure appropriate seeding density for picking single-cell-derived colonies. After 2–3 weeks in the CFC assay, individual haematopoietic colonies, were picked into PBS or ProteinaseK buffer (Arcturus Picopure DNA Extraction Kit, Applied Biosystems) and stored at  $-20^{\circ}\text{C}$  for subsequent whole-genome DNA sequencing (WGS) (Fig. 1). In two samples, single-cell liquid cultures were initiated with single  $\text{lin}^{-}\text{CD34}^{+}\text{CD38}^{+}\text{CD90}^{-}$  cells using the following antibodies: CD38-FITC (Clone HIT2, BD, San Jose, CA, USA; #555459; 1:12.5), CD34 PerCp-Cy5.5 (Clone 581, Biolegend #343522; 1:33, San Diego, USA, CD90-APC (Clone 5E10, Biolegend #328114; 1:33), after pre-enrichment for  $\text{CD34}^{+}$  cells (EasySep Human CD34 Positive Selection Kit, STEMCELL). Single HSPCs ( $\text{lin}^{-}\text{CD34}^{+}\text{CD38}^{+}\text{CD90}^{-}$ ) were flow-sorted using a BD Influx sorter and cultured in StemSpan supplemented with cytokines and recombinant growth factors SCF, FLT3L, IL3 and IL6 (cc100, STEMCELL).

### DNA extractions

DNA from picked CFC colonies were extracted using the Arcturus Picopure DNA extraction kit (Applied Biosystems). DNAeasy kit (Qiagen) was used for extraction from CFC colonies picked into PBS. DNA from buccal swabs was isolated using QIAmp DNA micro kit (Qiagen Cat. 56304).

### Whole genome sequencing of colonies

Individual colonies underwent whole-genome sequencing to identify somatic single nucleotide variants (SNVs), germ line variants, insertions/deletions and structural variants. We generated 150 bp paired-end sequencing reads using Illumina X Ten machines, resulting in a mean coverage of  $\sim 20\times$  per sample. The sequences were aligned to the human reference genome GRCh37d5 using the BWA-MEM algorithm<sup>59,60</sup>. Following removal of colonies due to low sequencing depth (less than  $6\times$ ) and low clonality (median variant allele frequency of less than 0.4), 323 colonies (range of 10–100 per individual, mean of 32 per individual) were taken forward for subsequent analysis.

### Somatic mutation identification and filtering

Single nucleotide variants (SNV) were identified using CaVEMan<sup>61</sup> for each colony by comparison to an in-silico unmatched sample (PD371s). CaVEMan was run with the “normal contamination of tumour” parameter set to zero, and the tumour/normal copy numbers set to 5/2. In addition to standard filters, reads supporting an SNV had to have a median BWA-MEM alignment Score  $\geq 140$  and less than half of the reads clipped. Filtering designed for quality control following processing through the Sanger low-input sequencing pipeline was also applied<sup>62</sup>. The use of the unmatched normal meant that this process called both somatic and germline SNVs. The removal of germline SNVs and artefacts of sequencing required further filtering. As published<sup>4</sup>, we used pooled information across colonies and read counts from a matched germline WGS buccal sample to ensure that genuine somatic variants that may have been present in the germline sample, either as embryonic variants or due to tumour-in-normal contamination were also identified. Short insertions and deletions were called using cgpPindel<sup>63</sup> with the standard WGS cgpPindel VCF filters applied, except the F018 Pindel filter was disabled as it excludes loci of depth  $< 10$ . Copy-number aberrations (CNA) were identified using ASCAT<sup>64</sup> with comparison to a matched normal sample. The union of colony SNVs and insertion–deletions (indels) was then taken and reads counted across all samples belonging to the individual (colonies and buccal samples) using VAFCorrect.

Structural variants (SVs) were called by BRASS<sup>65</sup>. We removed artefacts from the SV calls using AnnotateBRASS (<https://github.com/MathijsSanders/AnnotateBRASS><sup>66</sup>) with default settings.

### Creating a genotype matrix

The genotype at each locus within each sample was either 1 (present), 0 (absent) or NA (unknown). We inferred the genotype in a depth sensitive manner. We assumed the observed mutant read count for a colony at a given site was  $\text{MTR} \sim \text{Binomial}(n = \text{Depth}, p = \text{expected VAF})$ , if the site was mutant, and  $\text{MTR} \sim \text{Binomial}(n = \text{depth}, p = 0.01)$ , if the site was wild-type. The genotype was set to the most likely of the two possible states provided one of the states was at least 20 times more likely than the other. Otherwise the genotype is set to missing (NA). The expected VAF was usually 0.5 for autosomal sites, but for chromosomes X, Y and CNA sites, it was set to  $1/\text{ploidy}$ . For loss-of-heterozygosity (LOH) sites, the genotype was overridden and set to missing if it was originally 0.

### Phylogenetic tree topology

We constructed phylogenetic tree topologies using maximum parsimony with MPBoot<sup>67</sup>. The inputs for MPBoot were the binary genotype matrices with missing values per individual. Only SNVs were used to infer the topology, but both SNVs and indels were subsequently assigned to the branches of phylogenetic trees.

### Mutation assignment and branch length adjustment

Mutations were then assigned to the tree in a depth sensitive manner using treemut (<https://github.com/nangalialab/treemut><sup>4</sup>) with mutations being hard-assigned to the highest probability branch. Furthermore, branch lengths were adjusted for the branch specific SNV detection sensitivity<sup>4</sup>, where the sensitivity of detection of fully clonal SNV variants was directly estimated from the per colony sensitivity for detecting germline heterozygous SNVs together with a multiplicative correction for the clonality (VAF) of the colonies. In calculating mutation burden and branch lengths copy number regions that are present in any colony in an individual are uniformly masked out in all colonies for that individual and then the overall mutation burden is scaled back up by the reciprocal of 1-expected number of mutations in the masked region.

In addition to SNVs and indels, colonies exhibited a variety of LOH and CNA events. These events were curated as being present or absent in each of the colonies giving an event genotype vector similar to that obtained for SNVs and indels. Once the tree topology was inferred using the SNV genotypes, the branches that exactly matched the event genotype were identified and the event assigned to the corresponding branch.

### Timing branches

Given the linear accumulation of somatic mutations with age, we can infer the time point in life when driver mutations in phylogenetic trees had occurred. Branches at the top of a tree comprise mutations acquired at a young age, with branches lower down representing mutations arising later in life.

We have developed a formal model-based method *rtreefit* (<https://github.com/nangalialab/rtreefit>) for converting trees where branch lengths are expressed in molecular time (i.e. number of mutations) into trees where the branch lengths are expressed in units of time (years)<sup>4</sup>. In brief, the method jointly fits a single constant mutation rate (i.e. number of SNVs accumulated per year) and absolute time branch lengths using a Bayesian per individual tree-based model under the assumption that the number of observed mutations assigned to a branch is Poisson distributed with  $\text{Mean} = \text{Branch Duration} \times \text{Sensitivity} \times \text{Mutation Rate}$ , and subject to the constraint that the root to tip duration is equal to the age at sampling. Additionally, the method accounts for an elevated mutation rate during embryogenesis by assuming an excess mutation rate through development.

The *rtreefit* algorithm was run with 4 chains and 20,000 iterations per chain.

### Detection of driver mutations in WGS data

We searched specifically for hotspot driver mutations, copy number changes and rearrangements in 35 genes known to be associated with haematological malignancy<sup>35</sup> and clonal haematopoiesis<sup>33,34</sup> (*ASXL1*, *BCOR*, *CALR*, *CBL*, *CSF3R*, *CUX1*, *DNMT3A*, *EZH2*, *GATA2*, *GNAS*, *GNB1*, *IDH1*, *IDH2*, *JAK2*, *KIT*, *KRAS*, *MLL3*, *MPL*, *NFI*, *NFE2*, *NRAS*, *PHF6*, *PPM1D*, *PTPN11*, *RBI*, *RUNX1*, *SETBP1*, *SF3B1*, *SRSF2*, *SH2B3*, *STAG2*, *TET2*, *TP53*, *U2AF1*, *ZRSR2*) as well as in the recurrently mutated genes identified in SDS. We identified somatic mutations under positive and negative selection using dNdScv<sup>68</sup>.

### Mutational signature analysis

We characterised mutational profiles present in our dataset by performing signature extraction with *hdp* (<https://github.com/nicolaroberts/hdp>) without any signatures as prior and with no specified grouping of the data. In order to avoid double counting, mutations shared among colonies were randomly assigned to one colony. *hdp* identified the presence of 2 mutational signatures, one with strong similarity to Cosmic signatures SBS1<sup>42</sup> (cosine similarity  $\geq 0.95$ ) and one with strong similarity to SBSblood<sup>1,2,41</sup> (cosine similarity  $\geq 0.91$ ). We then estimated the proportion of SBS1 and SBSblood mutational signatures present in each colony using the programme *sigfit*<sup>69</sup>.

### SDS8 mutation burden and comparison to other individuals

We define the overdispersion in mutation burden as the ratio of the expected burden variance to Poisson variance. We estimate the overdispersion as a function of age using data from healthy/non-SDS blood single-cell-derived colonies reported in Mitchell et al.<sup>3</sup> The within-individual overdispersion at each time point is estimated as the sample mutation burden variance divided by the sample mean mutation burden. The log overdispersion was then modelled using a linear model with age as the explanatory variable. The estimated overdispersion at age 25 is 2.24 (1.85–2.73). For the purposes of assessing the statistical significance of the apparently high SDS8 mutation burden we conservatively account for the very high degree of shared history of the SDS8 colonies by regarding the clade as a single-cell with burden given by the mean burden, and then assess the probability of observing such an extreme mutation burden ( $n = 905$ ) under the null hypothesis that mutations were accrued according to a negative binomial distribution with a mean equal to the expected number of mutations ( $n = 496$ ) and a variance that is 2.24 times the mean.

### SDS8 timing of rapid growth and selection

We assume that there is a single transformation event that switches on a mutational process that is responsible for the distinct signature profile (transformation signature) observed in the expanded clade (Supplementary Fig. 4). We then estimate the timing of this event as the age that corresponds to the number of trunk mutations that can be attributed to the composite signature profile of normal SDS haematopoiesis (SDS6, SDS7 and SDS9) (composite age-matched SDS signature). The number of composite age-matched SDS signature trunk mutations accrued is estimated by using the R package *sigfit*<sup>69</sup> to decompose the trunk SNVs into contributions from the composite age-matched SDS signature and the transformation signature. We then estimate the corresponding age using Approximate Bayesian Computation with the rejection method, requiring that the number of substitutions acquired since birth follows a negative binomial distribution with a mean = age at transformation  $\times$  mutation rate, and variance set to 2.24 times the mean (see section above). The mutation rate itself is drawn from a normal distribution with mean 15.1 and variance of 1. The unconditional age estimate uses a uniform prior age range of 0–100 years, whereas the conditional age estimate uses a uniform prior range of 0–25 years. We estimate an ultrametric tree using *rtreefit*<sup>4</sup> (with age of transformation constrained to the lower bound of the 95% CI for the conditional age estimate (19.3 years). The

phylofit method<sup>3</sup> was then used to estimate the rate of clone growth using the timing and pattern of coalescences. Full details can be found in <https://github.com/nangalialab/ShwachmanDiamond>.

### Plasmid generation

cDNA for human eIF6 WT carrying a C-terminal FLAG-tag was generated by PCR using the Phusion High-Fidelity PCR kit (NEB). The PCR product was then inserted into pcDNA3.1 (Thermo Fisher Scientific) using the BamHI/XhoI sites, generating plasmid pEIF6-WT-FLAG). Site-directed mutagenesis was performed to generate the eIF6 MIT mutant (plasmid pEIF6-MIT-FLAG) using the Phusion High-Fidelity PCR kit (Thermo Fisher Scientific). Primers were as follows (5' to 3'):

```
eIF6-WT-Fw,TACTGGATCCATGGCGGTCGAGCTTCGTTCC
eIF6-WT-FLAG-Rev,AGTACTCGAGTCACTTGTCGT-
CATCGTCTTTGTAGTCGGTGAGGCTGTCAATGAGGGAATC eIF6-MIT-
Fw,CGGATCCACGGCGGTCGAGCTTCGTTCCGAGAACA
eIF6-MIT-Rev,
GGACCGCCGTGGATCCGAGCTCGGTACCAAGCTTAA
```

### Immunoblotting of human and *Drosophila* cell extracts

HEK293T cells (Sigma, 12022001) were grown in a 12-well dish to ~80% confluence followed by plasmid transfection using lipofectamine 2000 (Thermo Fisher Scientific) for 24 h. The cells were washed in 1x PBS and lysed in 0.5% NP-40 for 30 min on ice. The lysates were centrifuged at 21,130  $\times g$  for 10 min and the supernatant mixed with 50 mM DTT and 4x NuPAGE LDS sample buffer (Thermo Fisher Scientific) to 1x. Samples were incubated at 70 °C for 10 min. The proteins were separated in a NuPAGE 4–12% Bis-Tris gel (Thermo Fisher Scientific) in 1x MES running buffer (Formedium) prior to transfer to nitrocellulose membrane using the iBlot 2 (Thermo Fisher Scientific) system. The membrane was blocked with 5% (w/v) milk dissolved in PBST buffer (1x PBS with 0.1% [v/v] Tween 20) for 1 h. Human proteins were visualised using anti-FLAG (Sigma, #F7425, 1:5000 dilution), anti-eIF6 (GenTex, #GTX117971) and anti-actin antibodies (Sigma, #A2066), both at 1:1000 dilution. Anti-rabbit IgG HRP-linked antibody (Cell Signalling, #7074; 1:5000) was used as the secondary antibody. Blots were developed with the Western Chemiluminescent HRP substrate (Immobilon) and visualised using the Chemidoc™ MP (Bio-Rad) imaging system. Analysis was performed using Image Lab software v6.0.1 (Bio-Rad).

*Drosophila* third instar larvae (typically 15 larvae) were collected, washed with PBS, homogenised in lysis buffer (20 mM HEPES pH 7.4, 50 mM KCl, 2.5 mM MgCl<sub>2</sub>, 0.5% (v/v) IGEPAL® CA-630 (Sigma, #I8896), 0.5% (w/v) Sodium deoxycholate (Sigma, #30970) with complete EDTA-free protease inhibitors (Roche) and incubated for 15 min on ice. Lysates were cleared in a microcentrifuge at 20,000  $\times g$  for 10 min at 4 °C. Equal amounts (typically 10  $\mu g$ ) of total protein were loaded and separated using SDS-PAGE for immunoblotting. *Drosophila* proteins were visualised using anti-Gapdh (Merck #G9545, 1:20,000 dilution), anti-eIF6 (GeneTex, #GTX117971, 1:1000 dilution) and anti-FLAG antibodies (Abcam, 1:20,000 dilution). Secondary antibodies were all used at 1:10,000 dilution: anti-mouse IgG, HRP-conjugated (Sigma-A5287), anti-rabbit IgG, HRP-conjugated (Cell Signalling 7074), anti-goat IgG, HRP-conjugated (Santa Cruz, sc-2020) antibody. Blots were developed with the SuperSignal™ West Pico PLUS Chemiluminescent Substrate (Thermo Fisher, #34580), and visualised using a Chemidoc™ MP (Bio-Rad) imaging system. Analysis was performed using Image Lab software v6.0.1 (Bio-Rad).

### *Drosophila melanogaster* strains and genetics

Flies were maintained using standard culture techniques. All crosses were performed at 25 °C. Fly strains and genotypes are described in Supplementary Tables 2 and 3. The *Drosophila* lines *Sbds*<sup>9</sup>, *UAS-EIF6-FLAG*, *UAS-EIF6-R96W-FLAG*, *UAS-EIF6-N106S-FLAG* are described in Supplementary Table 3. To generate the *UAS-EIF6-IS8T-FLAG*



transgenic line, the coding sequence for *Drosophila EIF6* (NM\_145105) was amplified by PCR from *Drosophila* larval cDNA<sup>18</sup>. The variant *EIF6*<sup>I58T</sup> was generated by PCR site-directed mutagenesis and subcloned into pTWF (The *Drosophila* Gateway vector collection) to generate plasmid pUAS-EIF6-I58T-FLAG for microinjection. The transgenic *pUAS-EIF6-I58T-FLAG* line was generated by P element-mediated germline transformation into a *w*<sup>1118</sup> strain by BestGene Inc. Oligonucleotide primers used to generate the *Drosophila* strains were as follows (5' to 3'): EIF6-F: CACCATGGCTCTACGCGTCC; EIF6-R: GGACATGTCCTCGATGAGGGC; EIF6-I58T-F: CTGCCGACAATCGGCCGCC; EIF6-I58T-R: GCCGATTGTCGGCAGCCG. The da-GAL4 line was used to induce ubiquitous expression of FLAG-tagged eIF6 variants under the UAS promoter.

### Reporting summary

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

### Data availability

Raw sequencing data for all whole genomes are available at the European Genome-Phenome Archive (accession number [EGAD00001009061](https://www.ebi.ac.uk/ena/browser/view/EGAD00001009061)). Access to this human data hosted in the EGA is restricted due to sensitivity, and therefore it is managed in line with the Wellcome Sanger Data Sharing Policy. Researchers interested in accessing the data should submit a data access application to Sanger via eDAM2 (<https://edam.sanger.ac.uk/>). Further details can be found at <https://www.sanger.ac.uk/about/edam2-guide/#02-04>. When an application is received, a variety of checks are conducted by the data access team, e.g. the applicant's identity as a bona-fide researcher, their affiliation and the project they describe in their application is in line with any usage restrictions associated with the dataset(s) they have requested. There is no time limit on data access; the data access agreements are perpetual and run until terminated. However, the data access would be associated with (1) a specific project, so can only be used for that project for as long as it runs and (2) the researcher's institutional email address, so if they change affiliation, they would lose access to the data and would need to re-apply for data access under their new affiliation. Source data are provided with this paper.

### Code availability

A repository of the code for conducting all analyses can be found at [https://github.com/machadoheather/somatic\\_evolution\\_SDS](https://github.com/machadoheather/somatic_evolution_SDS) (<https://doi.org/10.5281/zenodo.8172028>)<sup>70</sup> and <https://github.com/nangalia/lab/ShwachmanDiamond> (<https://doi.org/10.5281/zenodo.8172581>)<sup>71</sup>.

### References

- Lee-Six, H. et al. Population dynamics of normal human blood inferred from somatic mutations. *Nature* **561**, 473–478 (2018).
- Osorio, F. G. et al. Somatic mutations reveal lineage relationships and age-related mutagenesis in human hematopoiesis. *Cell Rep.* **25**, 2308–2316.e4 (2018).
- Mitchell, E. et al. Clonal dynamics of haematopoiesis across the human lifespan. *Nature* **606**, 343–350 (2022).
- Williams, N. et al. Life histories of myeloproliferative neoplasms inferred from phylogenies. *Nature* **602**, 162–168 (2022).
- Fabre, M. A. et al. The longitudinal dynamics and natural history of clonal haematopoiesis. *Nature* **606**, 335–342 (2022).
- Zink, F. et al. Clonal hematopoiesis, with and without candidate driver mutations, is common in the elderly. *Blood* **130**, 742–752 (2017).
- Boocock, G. R. B. et al. Mutations in SBDS are associated with Shwachman–Diamond syndrome. *Nat. Genet.* **33**, 97–101 (2003).
- Menne, T. F. et al. The Shwachman–Bodian–Diamond syndrome protein mediates translational activation of ribosomes in yeast. *Nat. Genet.* **39**, 486–495 (2007).
- Warren, A. J. Molecular basis of the human ribosomopathy Shwachman–Diamond syndrome. *Adv. Biol. Regul.* **67**, 109–127 (2018).
- Finch, A. J. et al. Uncoupling of GTP hydrolysis from eIF6 release on the ribosome causes Shwachman–Diamond syndrome. *Genes Dev.* **25**, 917–929 (2011).
- Weis, F. et al. Mechanism of eIF6 release from the nascent 60S ribosomal subunit. *Nat. Struct. Mol. Biol.* **22**, 914–919 (2015).
- Jaako, P. et al. eIF6 rebinding dynamically couples ribosome maturation and translation. *Nat. Commun.* **13**, 1562 (2022).
- Donadieu, J. et al. Classification of and risk factors for hematologic complications in a French national cohort of 102 patients with Shwachman–Diamond syndrome. *Haematologica* **97**, 1312–1319 (2012).
- Dror, Y. et al. Clonal evolution in marrows of patients with Shwachman–Diamond syndrome: a prospective 5-year follow-up study. *Exp. Hematol.* **30**, 659–669 (2002).
- Parikh, S. et al. Acquired copy number neutral loss of heterozygosity of chromosome 7 associated with clonal haematopoiesis in a patient with Shwachman–Diamond syndrome. *Br. J. Haematol.* **159**, 480–482 (2012).
- Minelli, A. et al. The isochromosome i(7)(q10) carrying c.258+2t>c mutation of the SBDS gene does not promote development of myeloid malignancies in patients with Shwachman syndrome. *Leukemia* **23**, 708–711 (2009).
- Lee, S. et al. Somatic uniparental disomy mitigates the most damaging EFL1 allele combination in Shwachman–Diamond syndrome. *Blood* **138**, 2117–2128 (2021).
- Kennedy, A. L. et al. Distinct genetic pathways define pre-malignant versus compensatory clonal hematopoiesis in Shwachman–Diamond syndrome. *Nat. Commun.* **12**, 1334 (2021).
- Pressato, B. et al. Deletion of chromosome 20 in bone marrow of patients with Shwachman–Diamond syndrome, loss of the EIF6 gene and benign prognosis. *Br. J. Haematol.* **157**, 503–505 (2012).
- Valli, R. et al. Different loss of material in recurrent chromosome 20 interstitial deletions in Shwachman–Diamond syndrome and in myeloid neoplasms. *Mol. Cytogenet.* **6**, 56 (2013).
- Tan, S. et al. Somatic genetic rescue of a germline ribosome assembly defect. *Nat. Commun.* **12**, 5044 (2021).
- Sloan, K. E., Bohnsack, M. T. & Watkins, N. J. The 5S RNP couples p53 homeostasis to ribosome biogenesis and nucleolar stress. *Cell Rep.* **5**, 237–247 (2013).
- Elghetany, M. T. & Alter, B. P. p53 protein overexpression in bone marrow biopsies of patients with Shwachman–Diamond syndrome has a prevalence similar to that of patients with refractory anemia. *Arch. Pathol. Lab. Med.* **126**, 452–455 (2002).
- Tourlakis, M. E. et al. In vivo senescence in the Sbdbs-deficient murine pancreas: cell-type specific consequences of translation insufficiency. *PLoS Genet.* **11**, e1005288 (2015).
- Zambetti, N. A. et al. Deficiency of the ribosome biogenesis gene Sbdbs in hematopoietic stem and progenitor cells causes neutropenia in mice by attenuating lineage progression in myelocytes. *Haematologica* **100**, 1285–1293 (2015).
- Xia, J. et al. Somatic mutations and clonal hematopoiesis in congenital neutropenia. *Blood* **131**, 408–416 (2018).
- Kandoth, C. et al. Mutational landscape and significance across 12 major cancer types. *Nature* **502**, 333–339 (2013).
- Gerstung, M. et al. The evolutionary history of 2,658 cancers. *Nature* **578**, 122–128 (2020).
- Kuhn, E. et al. TP53 mutations in serous tubal intraepithelial carcinoma and concurrent pelvic high-grade serous carcinoma—evidence supporting the clonal relationship of the two lesions. *J. Pathol.* **226**, 421–426 (2012).



30. Folkins, A. K. et al. A candidate precursor to pelvic serous cancer (p53 signature) and its prevalence in ovaries and fallopian tubes from women with BRCA mutations. *Gynecol. Oncol.* **109**, 168–173 (2008).
31. Fearon, E. R. & Vogelstein, B. A genetic model for colorectal tumorigenesis. *Cell* **61**, 759–767 (1990).
32. Mercuri, A. et al. Immunophenotypic analysis of hematopoiesis in patients suffering from Shwachman–Bodian–Diamond syndrome. *Eur. J. Haematol.* **95**, 308–315 (2015).
33. Bick, A. G. et al. Inherited causes of clonal haematopoiesis in 97,691 whole genomes. *Nature* **586**, 763–768 (2020).
34. Jaiswal, S. et al. Clonal hematopoiesis and risk of atherosclerotic cardiovascular disease. *N. Engl. J. Med.* **377**, 111–121 (2017).
35. Grinfeld, J. et al. Classification and personalized prognosis in myeloproliferative neoplasms. *N. Engl. J. Med.* **379**, 1416–1430 (2018).
36. Myers, K. C. et al. Myelodysplastic syndrome and acute myeloid leukemia in patients with Shwachman Diamond syndrome: a multicentre, retrospective, cohort study. *Lancet Haematol.* **7**, e238–e246 (2020).
37. Steele, C. D. et al. Signatures of copy number alterations in human cancer. *Nature* **606**, 984–991 (2022).
38. Light, N. et al. Germline TP53 mutations undergo copy number gain years prior to tumor diagnosis. *Nat. Commun.* **14**, 77 (2023).
39. Abascal, F. et al. Somatic mutation landscapes at single-molecule resolution. *Nature* **593**, 405–410 (2021).
40. Watson, C. J. et al. The evolutionary dynamics and fitness landscape of clonal hematopoiesis. *Science* **367**, 1449–1454 (2020).
41. Machado, H. E. et al. Diverse mutational landscapes in human lymphocytes. *Nature* **608**, 724–732 (2022).
42. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
43. Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
44. Wong, C. C., Traynor, D., Basse, N., Kay, R. R. & Warren, A. J. Defective ribosome assembly in Shwachman–Diamond syndrome. *Blood* **118**, 4305–4312 (2011).
45. McKerrell, T. et al. Leukemia-associated somatic mutations drive distinct patterns of age-related clonal hemopoiesis. *Cell Rep.* **10**, 1239–1245 (2015).
46. Jaiswal, S. et al. Age-related clonal hematopoiesis associated with adverse outcomes. *N. Engl. J. Med.* **371**, 2488–2498 (2014).
47. van Zeventer, I. A. et al. Evolutionary landscape of clonal hematopoiesis in 3,359 individuals from the general population. *Cancer Cell* **41**, 1017–1031.e4 (2023).
48. Hecker, J. S. et al. CHIP and hips: clonal hematopoiesis is common in patients undergoing hip arthroplasty and is associated with autoimmune disease. *Blood* **138**, 1727–1732 (2021).
49. Mayerhofer, C. et al. Clonal hematopoiesis in older patients with breast cancer receiving chemotherapy. *J. Natl Cancer Inst.* **115**, 981–988 (2023).
50. Cao, B. et al. Cancer-mutated ribosome protein L22 (RPL22/eL22) suppresses cancer cell survival by blocking p53-MDM2 circuit. *Oncotarget* **8**, 90651–90661 (2017).
51. Arzalluz-Luque, Á. et al. Mutant PRPF8 causes widespread splicing changes in spliceosome components in retinitis pigmentosa patient iPSC-derived RPE cells. *Front. Neurosci.* **15**, 636969 (2021).
52. Zemp, I. et al. CK1δ and CK1ε are components of human 40S subunit precursors required for cytoplasmic 40S maturation. *J. Cell Sci.* **127**, 1242–1253 (2014).
53. Ameer, A., Kloosterman, W. P. & Hestand, M. S. Single-molecule sequencing: towards clinical applications. *Trends Biotechnol.* **37**, 72–85 (2019).
54. Bernard, E. et al. Molecular international prognostic scoring system for myelodysplastic syndromes. *NEJM Evid.* **1**, 1–14 (2022).
55. Lindsley, R. C. et al. Prognostic mutations in myelodysplastic syndrome after stem-cell transplantation. *N. Engl. J. Med.* **376**, 536–547 (2017).
56. Papaemmanuil, E. et al. Genomic classification and prognosis in acute myeloid leukemia. *N. Engl. J. Med.* **374**, 2209–2221 (2016).
57. Gerstung, M. et al. Precision oncology for acute myeloid leukemia using a knowledge bank approach. *Nat. Genet.* **49**, 332–340 (2017).
58. Bernard, E. et al. Implications of TP53 allelic state for genome stability, clinical presentation and outcomes in myelodysplastic syndromes. *Nat. Med.* **26**, 1549–1556 (2020).
59. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. <https://doi.org/10.48550/arXiv.1303.3997> (2013).
60. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
61. Jones, D. et al. cgpaVEManWrapper: simple execution of CaVEMan in order to detect somatic single nucleotide variants in NGS data. *Curr. Protoc. Bioinforma.* **56**, 15.10.1–15.10.18 (2016).
62. Ellis, P. et al. Reliable detection of somatic mutations in solid tissues by laser-capture microdissection and low-input DNA sequencing. *Nat. Protoc.* **16**, 841–871 (2021).
63. Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865–2871 (2009).
64. Loo, P. V. et al. Allele-specific copy number analysis of tumors. *PNAS* **107**, 16910–16915 (2010).
65. Nik-Zainal, S. et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54 (2016).
66. Moore, L. et al. The mutational landscape of normal human endometrial epithelium. *Nature* **580**, 640–646 (2020).
67. Hoang, D. T. et al. MPBoot: fast phylogenetic maximum parsimony tree inference and bootstrap approximation. *BMC Evol. Biol.* **18**, 11 (2018).
68. Martincorena, I. et al. Universal patterns of selection in cancer and somatic tissues. *Cell* **171**, 1029–1041.e21 (2017).
69. Gori, K. & Baez-Ortega, A. sigfit: flexible Bayesian inference of mutational signatures. *bioRxiv* 372896 <https://doi.org/10.1101/372896> (2020).
70. Machado, H. E. Convergent somatic evolution commences in utero in a germline ribosomopathy. [https://github.com/machadoheather/somatic\\_evolution\\_SDS](https://github.com/machadoheather/somatic_evolution_SDS), <https://doi.org/10.5281/zenodo.8172028> (2023).
71. Machado, H. E. Convergent somatic evolution commences in utero in a germline ribosomopathy. <https://github.com/nangalialab/ShwachmanDiamond>, <https://doi.org/10.5281/zenodo.8172581> (2023).

## Acknowledgements

J.N. is supported by a Cancer Research UK Fellowship and work in the J.N. lab is supported by the Wellcome Trust, Cancer Research UK, Alborada Trust, Rosetrees Trust and the MPN Research Foundation. Work in the D.G.K. laboratory is supported by a European Research Council Starting Grant (ERC-2016-STG-715371), the Bill and Melinda Gates Foundation (INV-002189 and INV-038816) and a Cancer Research UK Programme Foundation Award (DCRPGF100008). A.J.W. was supported by a Blood Cancer UK Programme Continuity Grant (21002 to A.J.W.), the UK Medical Research Council (MR/T012412/1), the Kay Kendall Leukaemia Fund, Rosetrees Trust, the SDS Foundation, the Shwachman–Diamond Project, the Butterfly Guild, SDS UK, the Connor Wright Project, the Cambridge National Institute for Health Research Biomedical Research Centre and the European Cooperation in Science and Technology (COST) Action CA18233 “European Network for Innovative Diagnosis and treatment of Chronic Neutropenias, EuNet

INNOCHRON" and CA21154, "Translational control in Cancer European Network, TRANSLACORE". Samples were provided by the Cambridge Blood and Stem Cell Biobank, which is supported by the Cambridge NIHR Biomedical Research Centre, Wellcome Trust-MRC Stem Cell Institute and the Cambridge Experimental Cancer Medicine Centre, UK. The authors would also like to thank the individuals for donating the samples that have been used in this study.

### Author contributions

H.E.M. and N.W. performed genomic analyses; N.F.Ø. developed clonal assays and performed phenotypic analyses with assistance from M.D., M.B. and A.C.; S.T. and A.Z.B. performed functional eIF6 assays; E.J.B., M.D. and A.C. assisted with sample generation; E.M., M.D., A.C., N.M. and E.L. shared WGS data from normal individuals; P.A., J.K., S.B.K., A.K., S.M. and A.J.W. provided samples; P.J.C., D.G.K., J.N. and A.J.W. supervised the study. H.E.M., J.N., D.G.K. and A.J.W. wrote the manuscript. All authors reviewed the manuscript.

### Competing interests

A.J.W. and S.T. are consultants for SDS Therapeutics. P.J.C. is a cofounder and shareholder of FL86 Inc. The remaining authors declare no competing interests.

### Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41467-023-40896-5>.

**Correspondence** and requests for materials should be addressed to David G. Kent, Jyoti Nangalia or Alan J. Warren.

**Peer review information** *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>

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

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023