The American Journal of Human Genetics, Volume *108*

Supplemental Data

Mutation-specific pathophysiological mechanisms define different neurodevelopmental disorders associated with SATB1 dysfunction

Joery den Hoed, Elke de Boer, Norine Voisin, Alexander J.M. Dingemans, Nicolas Guex, Laurens Wiel, Christoffer Nellaker, Shivarajan M. Amudhavalli, Siddharth Banka, Frederique S. Bena, Bruria Ben-Zeev, Vincent R. Bonagura, Ange-Line Bruel, Theresa Brunet, Han G. Brunner, Hui B. Chew, Jacqueline Chrast, Loreta Cimbalistienė, Hilary Coon, The DDD Study, Emmanuèlle C. Délot, Florence Démurger, Anne-Sophie Denommé-Pichon, Christel Depienne, Dian Donnai, David A. Dyment, Orly Elpeleg, Laurence Faivre, Christian Gilissen, Leslie Granger, Benjamin Haber, Yasuo Hachiya, Yasmin Hamzavi Abedi, Jennifer Hanebeck, Jayne Y. Hehir-Kwa, Brooke Horist, Toshiyuki Itai, Adam Jackson, Rosalyn Jewell, Kelly L. Jones, Shelagh Joss, Hirofumi Kashii, Mitsuhiro Kato, Anja A. Kattentidt-Mouravieva, Fernando Kok, Urania Kotzaeridou, Vidya Krishnamurthy, Vaidutis Kučinskas, Alma Kuechler, Alinoë Lavillaureix, Pengfei Liu, Linda Manwaring, Naomichi Matsumoto, Benoît Mazel, Kirsty McWalter, Vardiella Meiner, Mohamad A. Mikati, Satoko Miyatake, Takeshi Mizuguchi, Lip H. Moey, Shehla Mohammed, Hagar Mor-Shaked, Hayley Mountford, Ruth Newbury-Ecob, Sylvie Odent, Laura Orec, Matthew Osmond, Timothy B. Palculict, Michael Parker, Andrea K. Petersen, Rolph Pfundt, Eglė Preikšaitienė, Kelly Radtke, Emmanuelle Ranza, Jill A. Rosenfeld, Teresa Santiago-Sim, Caitlin Schwager, Margje Sinnema, Lot Snijders Blok, Rebecca C. Spillmann, Alexander P.A. Stegmann, Isabelle Thiffault, Linh Tran, Adi Vaknin-Dembinsky, Juliana H. Vedovato-dos-Santos, Samantha A. Schrier Vergano, Eric Vilain, Antonio Vitobello, Matias Wagner, Androu Waheeb, Marcia Willing, Britton Zuccarelli, Usha Kini, Dianne F. Newbury, Tjitske Kleefstra, Alexandre Reymond, Simon E. Fisher, and Lisenka E.L.M. Vissers

Supplemental information:

- Fig S1 Pedigrees of (suspected) mosaic families with *SATB1* variants.
- Fig S2 Amino acid sequence alignments of the CUT1, CUT2 and Homeobox domain of SATB1.
- Fig S3 Heterozygous (partial) gene deletions of the *SATB1* gene.
- Fig S4 Clinical evaluation of individuals with SATB1 variants.
- Fig S5 Grouped HPO features based on semantic similarity and clustering results per individual.
- Fig S6 Overexpression of SATB1 missense variants as YFP-fusion proteins.
- Fig S7 MetaDome analysis of the *SATB1* missense variants.
- Fig S8 Functional characterization of the SATB1 p.R410* variant.
- Fig S9 Overexpression of SATB1 NMD-escaping PTVs as YFP-fusion proteins.
- Fig S10 SUMOylation of SATB1 protein truncating variants escaping NMD.
- Fig S11 Clinical evaluation of individuals with *SATB1* variants in three subcohorts.
- Fig S12 The SATB2 p.E396Q missense variant has comparable effects on protein functions as the p.E407G and p.E530K/Q SATB1 variants affecting equivalent positions.
- Fig S13 Missense variants identified in individuals with NDD displayed in an amino acid sequence alignment of SATB2 and SATB1.

- Table S1 Clinical features and variant details of individuals with (*de novo*) *SATB1* variants (as separate .xlsx file)
- Table S2 Splice-AI predictions for missense variants at intron-exon or exon-intron junctions.
- Table S3 Phenotypic information of individuals from the UK10K cohort with rare *SATB1* missense variants.
- Table S4 NMD efficacy predictions for SATB1 truncating variants.
- Table S5 Summary of clinical characteristics associated with *(de novo) SATB1* PTVs and (partial) gene deletions predicted to result in haploinsufficiency and PTVs in the last exon
- Table S6 Summary of clinical, molecular and functional findings of this study per *SATB1* variant (as separate .xlsx file)
- Table S7 List of phenotypic features grouped based on semantic similarity, used for HPO-based clustering analysis (as separate .xlsx file)
- Table S8 Primers for site-directed mutagenesis.
- Table S9 Primers for amplifying and subcloning human UBC9 (NM_194260.2) and SATB1 (NM_001131010.4).
- Table S10 Primers to amplify regions that include the SATB1 NMD escaping truncating variants used for testing for NMD.

- Supplemental Acknowledgements

- Supplemental Materials and Methods

- Detailed descriptions of 3D protein modeling of (*de novo*) SATB1 variants (at the end of the Supplemental information)

- Clinical phenotypic data of individuals with (*de novo*) SATB1 variants in standardized HPO format (as separate zipped .JSON file)

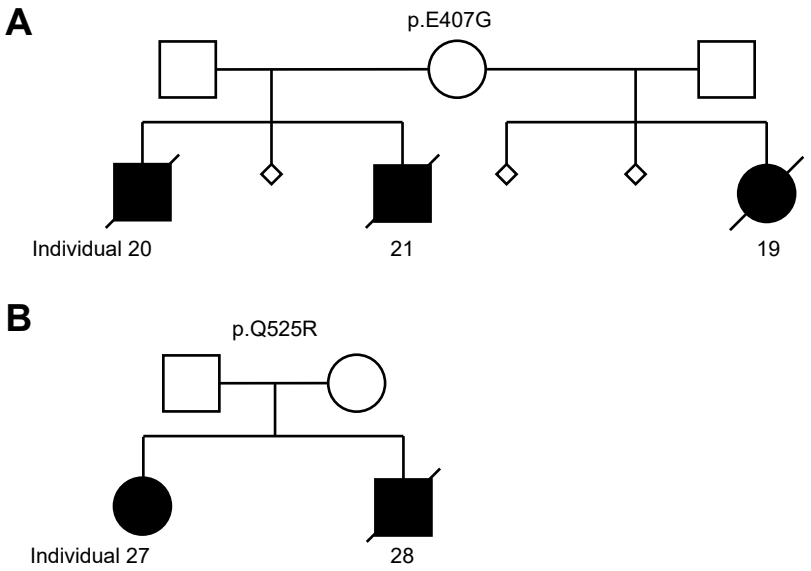


Figure S1. Pedigrees of (suspected) mosaic families with *SATB1* variants. **A**) Pedigree of family with proband and siblings carrying a heterozygous SATB1 p.E407G variant. The mother presents the variant in 1 of 69 reads in whole exome sequencing data, so the estimated percentage is 1.4% in the peripheral blood. Karyotyping was normal. **B**) Pedigree of family with proband and sibling carrying a heterozygous SATB1 p.Q525R variant. Suspected mosaicism in one of the parents could not be confirmed with Sanger sequencing of DNA derived from peripheral blood. **A-B**) In both families, none of the pregnancies resulted in healthy offspring.

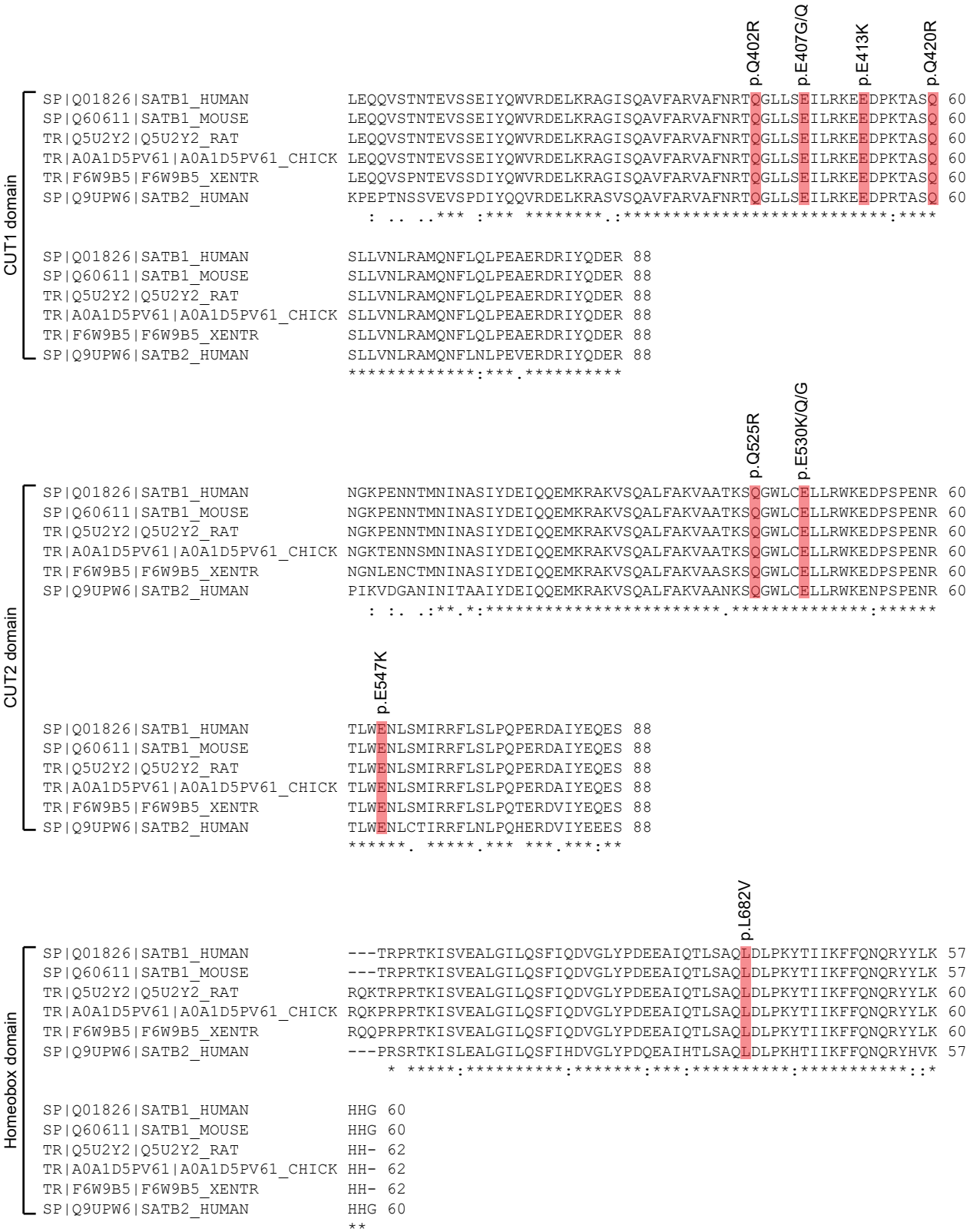
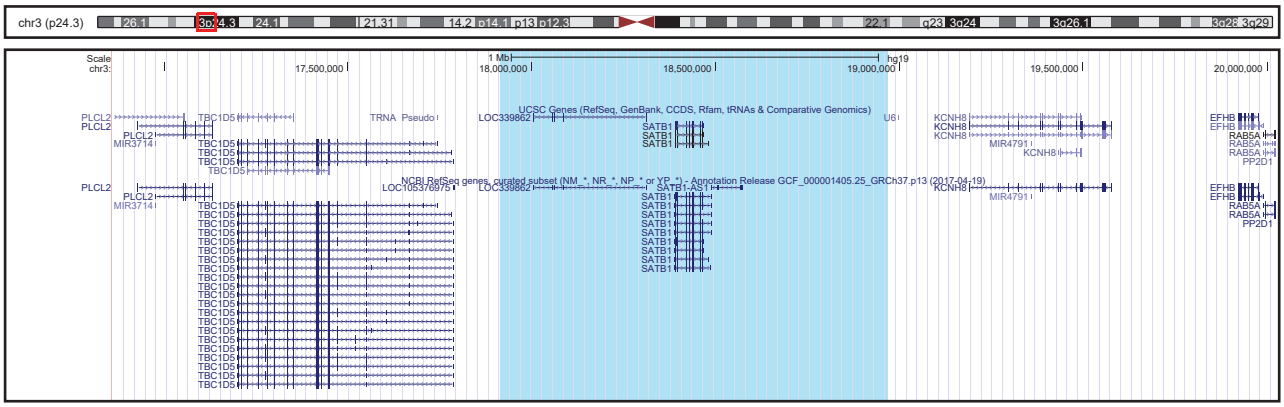


Figure S2. Amino acid sequence alignments of the CUT1, CUT2 and Homeobox domain of SATB1. Amino acid sequences of the CUT1, CUT2 and Homeobox domain of human SATB1 (Q01826, UniProt) aligned to the mouse (Q60611), rat (Q5U2Y2), chicken (A0A1D5PV61) and Xenopus tropicalis (F6W9B5) sequences, and the sequences of the homolog domains in human SATB2 (Q9UPW6). Alignment was performed with Clustal Omega (1.2.4) with default settings using UnitProt alignment tool. Missense variants described in this study and identified in these functional domains are shaded in red.

1.05Mb deletion - chr3:17915162-18968823



55Kb deletion - chr3:18376866-18432504

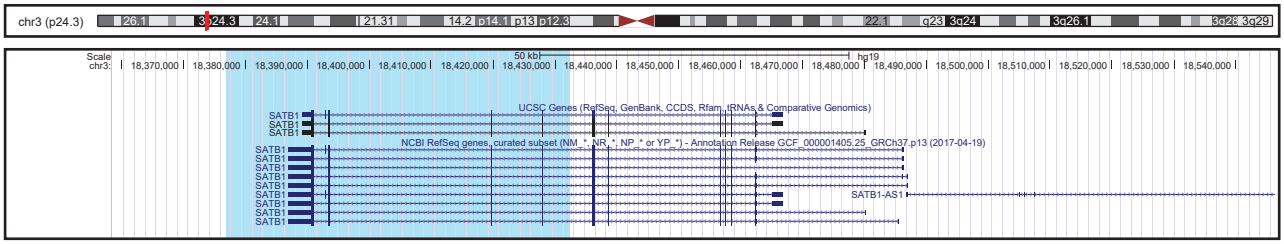
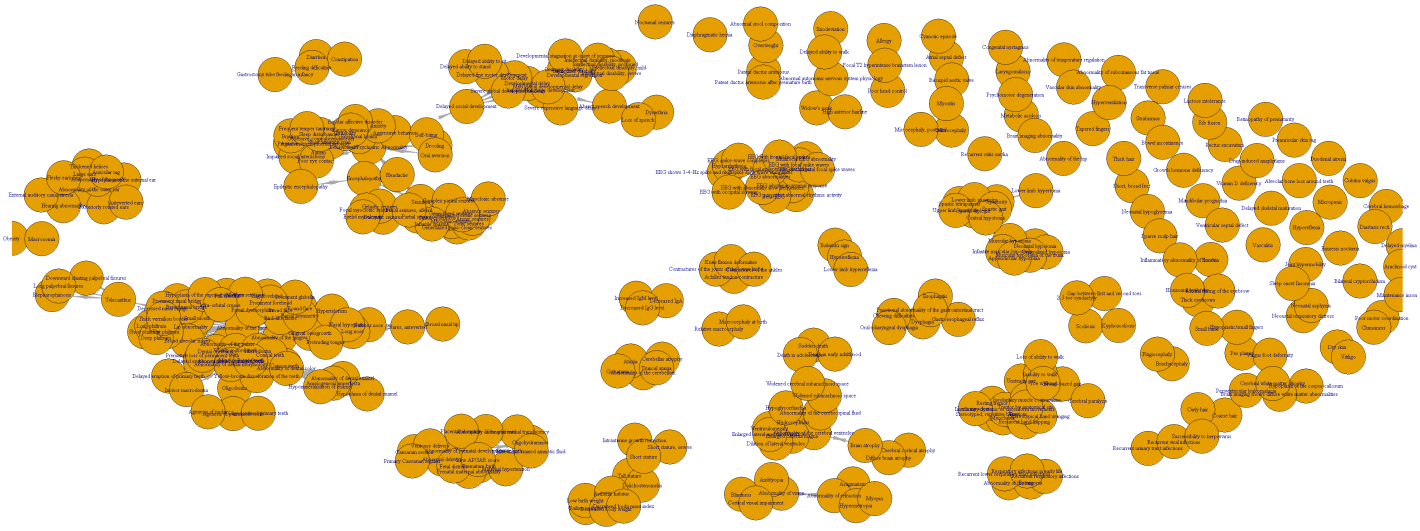


Figure S3. Heterozygous (partial) gene deletions of the *SATB1* gene. Genome overviews of two reported heterozygous deletions that include the *SATB1* gene, generated in the UCSC Genome Browser (assembly Feb. 2009 GRCh37/hg19). The deleted regions are shaded in red in the chromosome ideogram, and in light blue in the genome overview.

Figure S4. Clinical evaluation of individuals with *SATB1* variants. A) Side view photographs, depicting prominent ears (individuals 4, 8, 14, 17, 19, 34, 35), with thickened helices (individuals 8, 14, 17, 19, 33, 34, 35), and retrognathia (individuals 8, 14, 17, 19, 27, 34). **B**) Additional photograph of teeth. No evident enamel or dental positioning problems in individuals 8 and 14, although missing molars (individual 8) and malformed teeth (individual 14) are reported. Lower teeth of individual 28: discoloration, malpositioning and teeth decay. **C**) Photographs of hands and feet. Features include contractures resulting from spasticity (individual 17), tapered fingers (individuals 13, 14, 23, 35), short broad fingers (individuals 13, 14, 23), clinodactyly of 5th finger (individual 9), overlapping 2nd toe (individual 35) or 4th toe (individual 9) and broad feet with short toes and small toe nails (individuals 13, 14, 23).

A



B

Identifier	Variant type	2_cluster _pred	2_correct	3_cluster _pred	3_correct
Individual1	PTV_non_last_exon	PTV	CORRECT	PTV_last_exon	INCORRECT
Individual2	PTV_non_last_exon	PTV	CORRECT	PTV_non_last_exon	CORRECT
Individual3	PTV_non_last_exon	PTV	CORRECT	PTV_last_exon	INCORRECT
Individual4	PTV_non_last_exon	PTV	CORRECT	PTV_last_exon	INCORRECT
Individual5	PTV_non_last_exon	Missense	INCORRECT	Missense	INCORRECT
Individual6	PTV_non_last_exon	PTV	CORRECT	PTV_last_exon	INCORRECT
Individual7	PTV_non_last_exon	Missense	INCORRECT	PTV_last_exon	INCORRECT
Individual8	PTV_last_exon	PTV	CORRECT	PTV_last_exon	CORRECT
Individual9	PTV_last_exon	PTV	CORRECT	PTV_last_exon	CORRECT
Individual10	PTV_last_exon	PTV	CORRECT	PTV_last_exon	CORRECT
Individual11	PTV_last_exon	PTV	CORRECT	PTV_non_last_exon	INCORRECT
Individual12	PTV_last_exon	PTV	CORRECT	PTV_last_exon	CORRECT

Identifier	Variant type	2_cluster _pred	2_correct	3_cluster _pred	3_correct
Individual13	Missense	PTV	INCORRECT	PTV_last_exon	INCORRECT
Individual14	Missense	PTV	INCORRECT	PTV_last_exon	INCORRECT
Individual15	Missense	PTV	INCORRECT	PTV_last_exon	INCORRECT
Individual17	Missense	Missense	CORRECT	Missense	CORRECT
Individual18	Missense	PTV	INCORRECT	PTV_last_exon	INCORRECT
Individual19	Missense	Missense	CORRECT	Missense	CORRECT
Individual20	Missense	Missense	CORRECT	Missense	CORRECT
Individual21	Missense	Missense	CORRECT	Missense	CORRECT
Individual23	Missense	PTV	INCORRECT	PTV_last_exon	INCORRECT
Individual24	Missense	Missense	CORRECT	Missense	CORRECT
Individual25	Missense	Missense	CORRECT	PTV_non_last_exon	INCORRECT
Individual26	Missense	Missense	CORRECT	Missense	CORRECT
Individual27	Missense	Missense	CORRECT	Missense	CORRECT
Individual28	Missense	Missense	CORRECT	Missense	CORRECT
Individual29	Missense	Missense	CORRECT	Missense	CORRECT
Individual30	Missense	Missense	CORRECT	Missense	CORRECT
Individual31	Missense	Missense	CORRECT	PTV_non_last_exon	INCORRECT
Individual33	Missense	Missense	CORRECT	PTV_non_last_exon	INCORRECT
Individual34	Missense	Missense	CORRECT	PTV_non_last_exon	INCORRECT
Individual35	Missense	PTV	INCORRECT	PTV_last_exon	INCORRECT
Individual36	Missense	PTV	INCORRECT	PTV_last_exon	INCORRECT
Individual37	Missense	Missense	CORRECT	Missense	CORRECT
Individual38	Missense	Missense	CORRECT	Missense	CORRECT
Individual39	Missense	Missense	CORRECT	PTV_non_last_exon	INCORRECT
Individual40	Missense	PTV	INCORRECT	PTV_last_exon	INCORRECT
Individual42	Missense	PTV	INCORRECT	PTV_last_exon	INCORRECT
Correctly predicted individuals:			**27**		**17**

Figure S5. Grouped HPO features based on semantic similarity and clustering results per individual. A) The semantic similarity between all the HPO terms used in this cohort (356 features) was calculated using the Wang algorithm in the HPOsim package in R. HPO terms with at least a 0.5 similarity score were grouped and a new feature was created as a replacement, which was the sum of the grouped features. **B**) Individual HPO-based phenotypic clustering results for both analyses with two and three clusters.

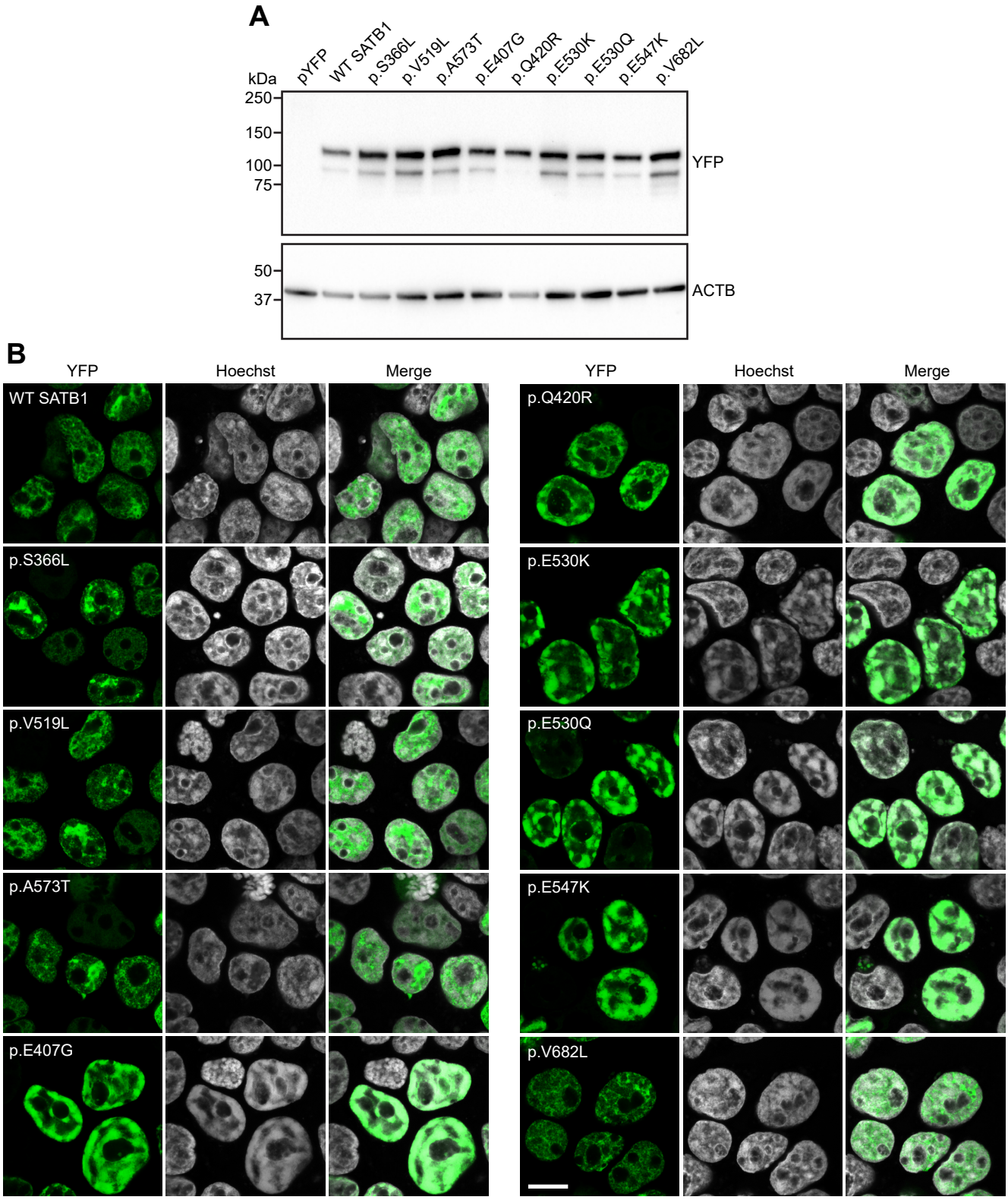
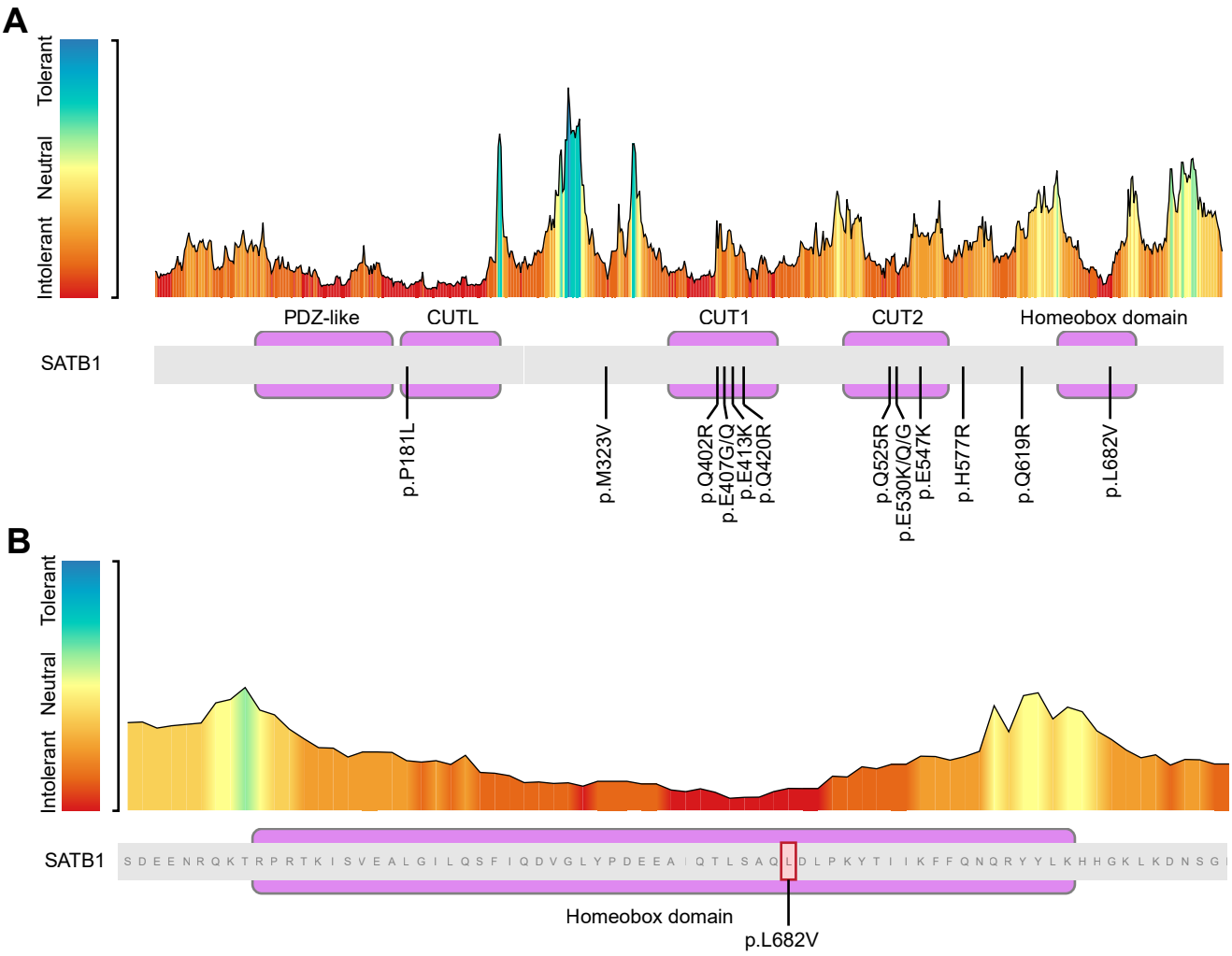


Figure S6. Overexpression of SATB1 missense variants as YFP-fusion proteins. A) Immunoblot of whole-cell lysates expressing YFP-tagged SATB1 variants probed with anti-EGFP antibody. Expected molecular weight for all variants is ~115 kDa. The blot was probed for ACTB to ensure equal protein loading. **B**) Direct fluorescence micrographs of HEK293T/17 cells expressing YFP-SATB1 fusion proteins (green). Nuclei were stained with Hoechst 33342 (white). Scale bar = 10 µm.



C

Gene name	gDNA	cDNA	Residue change	Type	gnomAD allele frequency
HMX3	chr10:124896963C>G	NM_001105574.1:c.790C>G	p.(Leu264Val)	missense	0.000004
HOXC10	chr12:54383114A>G	NM_017409.3:c.913A>G	p.(Ile305Val)	missense	0.000004
HOXC11	chr12:54369087C>G	NM_014212.3:c.805C>G	p.(Leu269Val)	missense	0.000004
ISX	chr22:35478636A>G	NM_001303508.1:c.355A>G	p.(Ile119Val)	missense	0.000004
NANOGNB	chr12:7922891T>G	NM_001145465.1:c.415T>G	p.(Phe139Val)	missense	0.000023
NKX1-2	chr10:126136333G>C	NM_001146340.2:c.598C>G	p.(Leu200Val)	missense	0.000009
NKX2-2	chr20:21492890T>C	NM_002509.3:c.493A>G	p.(Ile165Val)	missense	0.000004
NOBOX	chr7:144097323C>T	NM_001080413.3:c.927G>A	p.(Val309=)	synonymous	0.000004
OTP	chr5:76932672T>C	NM_032109.2:c.421A>G	p.(Ile141Val)	missense	0.00007
PAX3	chr2:223096822G>A	NM_181459.3:c.767C>T	p.(Ala256Val)	missense	0.000004
POU2F2	chr19:42599569G>C	NM_001207025.2:c.1000C>G	p.(Leu334Val)	missense	0.000004
POU6F1	chr12:51584125G>C	NM_001330422.1:c.1741C>G	p.(Leu581Val)	missense	0.000004
ZFHX3	chr16:72828547C>T	NM_006885.3:c.8034G>A	p.(Val2678=)	synonymous	0.000004
ZFHX4	chr8:77767083C>A	NM_024721.4:c.7926C>A	p.(Val2642=)	synonymous	0.000096
ZFHX4	chr8:77767083C>T	NM_024721.4:c.7926C>T	p.(Val2642=)	synonymous	0.000008

Figure S7. MetaDome analysis of the *SATB1* missense variants. A) Overview of the SATB1 protein (transcript NM_001131010.2) tolerance landscape. All missense variants identified in affected individuals are indicated. **B**) Detailed overview of the SATB1 homeobox domain tolerance landscape, with the p.L682V variant indicated. **C**) Table listing all residue changes at positions equivalent to the SATB1 p.L682 position in homolog homeobox domain proteins that change to a valine. The gnomAD allele frequency is indicated.

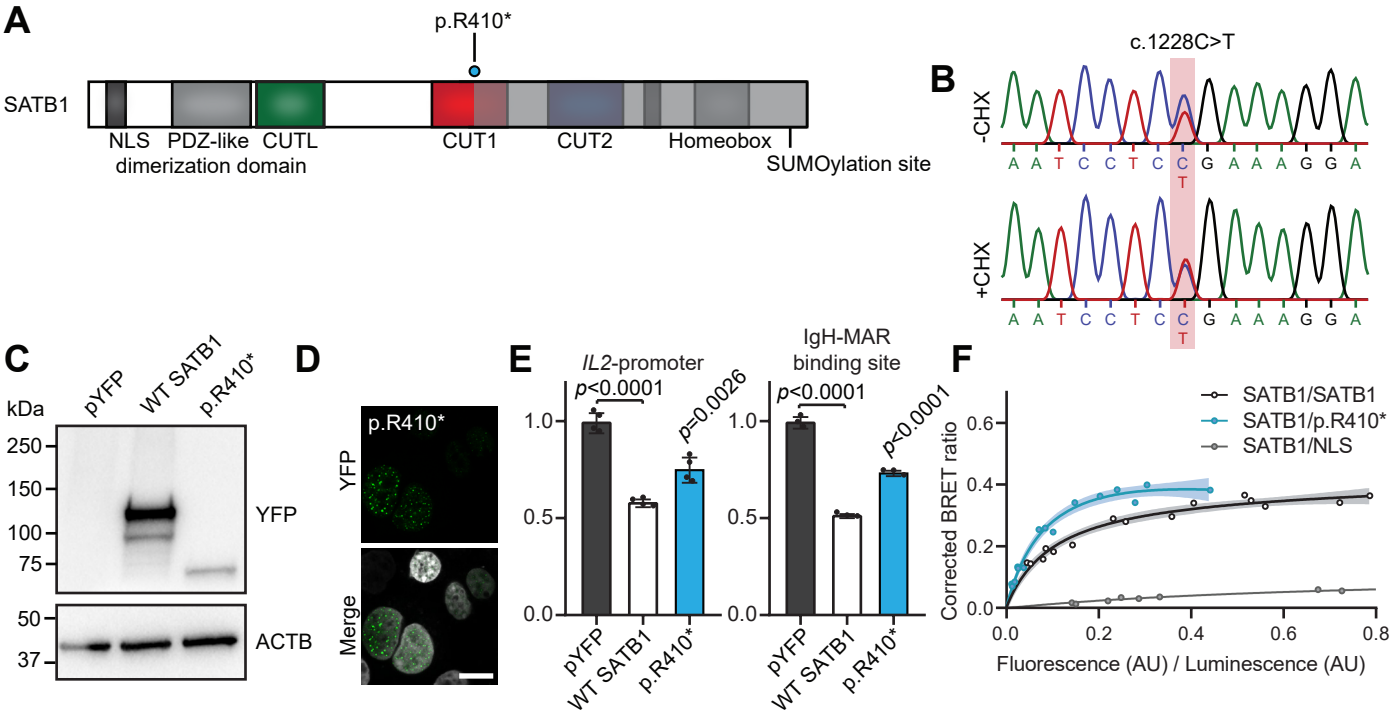


Figure S8. Functional characterization of the SATB1 p.R410* variant. A) Schematic representation of SATB1 with the p.R410* variant labeled in cyan. **B**) Sanger sequencing traces of patient-derived EBV transformed lymphoblastoid cell lines treated with or without cyclohexamide (CHX) to test for NMD. The mutated nucleotides are shaded in red. **C**) Immunoblot of whole-cell lysates expressing YFP-tagged SATB1 and p.R410* probed with anti-EGFP antibody. Expected molecular weight is SATB1: ~115 kDa, p.R410*: ~75kDa. The blot was probed for ACTB to ensure equal protein loading. **D**) Direct fluorescence micrographs of HEK293T/17 cells expressing YFP-SATB1 p.R410* fusion proteins (green). Nuclei were stained with Hoechst 33342 (white). Scale bar = 10 µm. **E**) Luciferase reporter assays using reporter constructs containing the IL2 promoter region and the IgH matrix associated region (MAR) binding site. Values are expressed relative to the control (pYFP; black) and represent the mean ± S.E.M. (n = 4 for IL2-promoter, n = 3 for IgH-MAR binding site, p-values compared to wildtype (WT) SATB1 (white), one-way ANOVA and *post-hoc* Bonferroni test). **F**) BRET assays for SATB1 dimerization in live cells. The plot shows the mean BRET saturation curves ± 95% C.I. fitted using a non-linear regression equation assuming a single binding site (y = BRETmax * x / (BRET50 / x); GraphPad). The corrected BRET ratio is plotted against the ratio of fluorescence/luminescence (AU) to correct for expression level differences between conditions (n = 3).

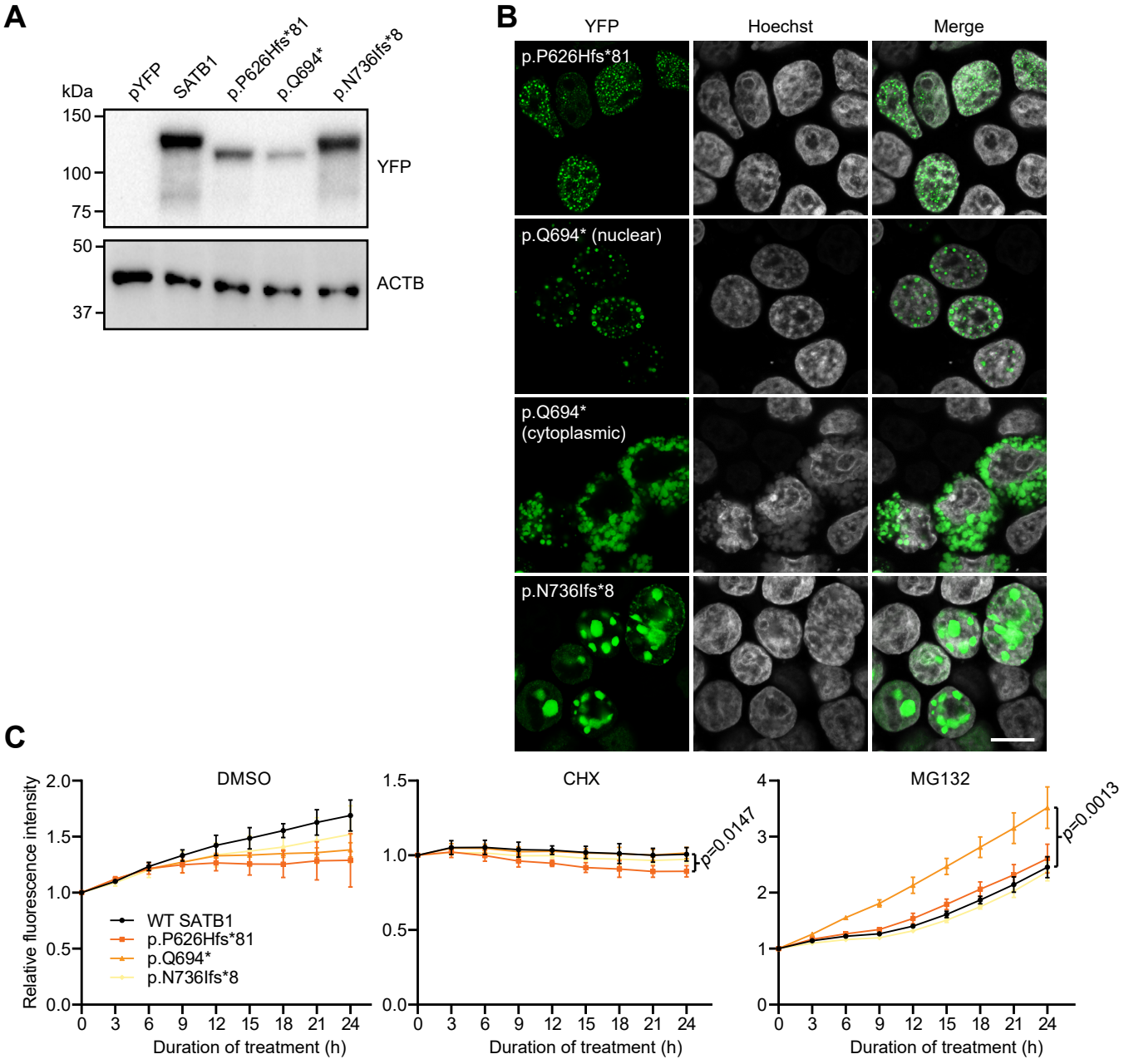
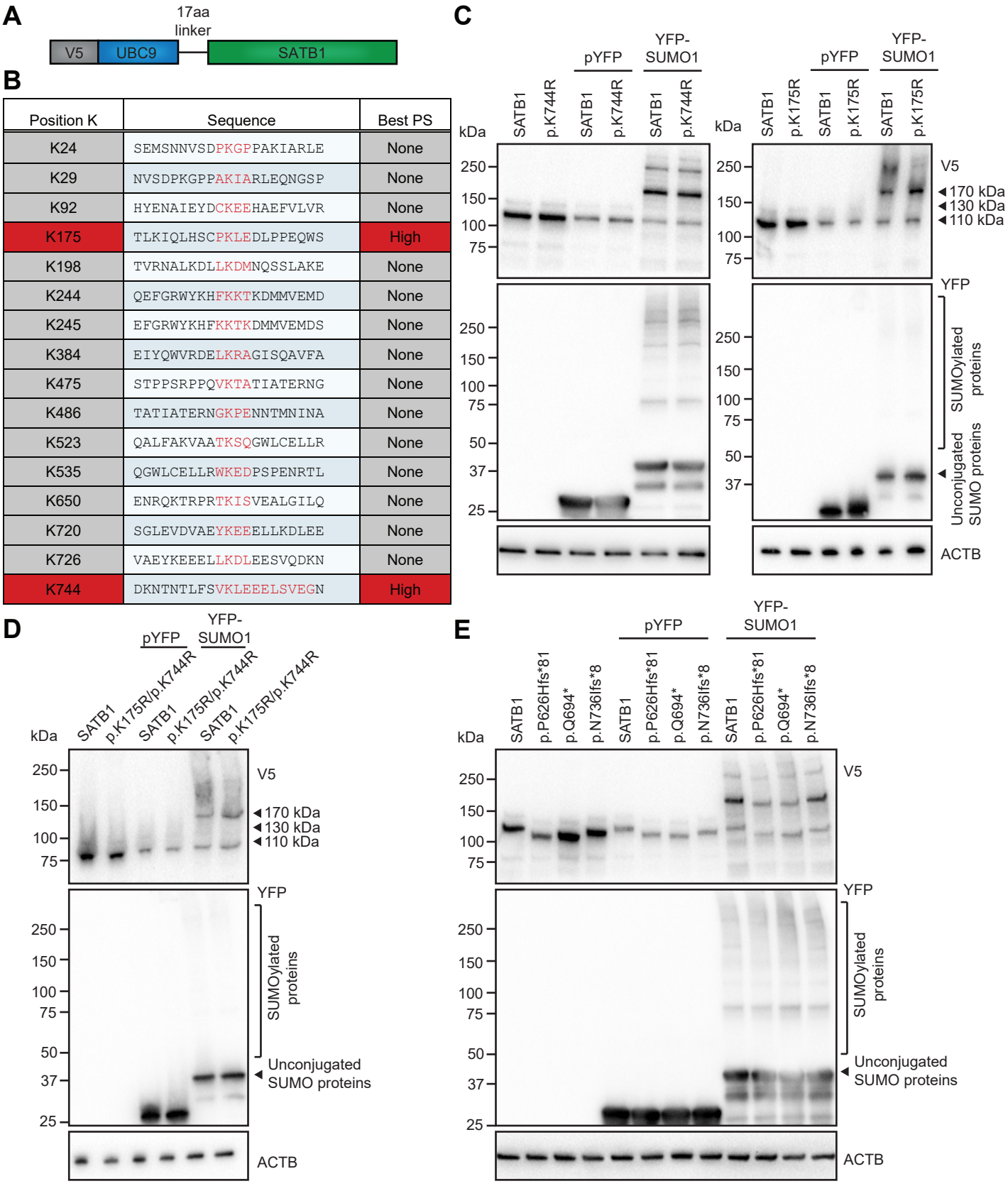


Figure S9. Overexpression of SATB1 NMD-escaping PTVs as YFP-fusion proteins. A) Immunoblot of whole-cell lysates expressing YFP-tagged SATB1 variants probed with anti-EGFP antibody. Expected molecular weight: WT SATB1 = ~115 kDa, p.P626Hfs*81 = ~109 kDa, p.Q694* = ~107 kDa, p.N736Ifs*8 = ~113 kDa. The blot was probed for ACTB to ensure equal protein loading. **B**) Direct fluorescence imaging of HEK293T/17 cells expressing YFP-SATB1 fusion proteins (green). Nuclei were stained with Hoechst 33342 (white). Scale bar = 10 µm. **C**) Results of assay for protein stability of SATB1 NMD-escaping PTVs, using cycloheximide (CHX) to arrest protein synthesis, and MG132 to block protein degradation by the 26S proteasome complex. Values represent the mean protein expression levels of YFP-tagged SATB1 variants ± S.E.M. in live cells as measured by YFP fluorescence and expressed relative to the 0 h time point (n = 3, two-way ANOVA for repeated measures with Geisser-Greenhouse correction, followed by a *post-hoc* Bonferroni test). Although p.P626Hfs*81 showed a slight but significant decrease in relative expression level after treatment with CHX, and p.Q694* showed a significant increase in relative expression level after treatment with MG132 when compared to WT SATB1, none of the variants tested showed both a decrease in levels after CHX treatment and an increase after MG132 treatment, which would be indicative of reduced protein stability.



Position K	Sequence	Best PS
K24	SEMSNNVSDPKGPPAKIARLE	None
K29	NVSDPKGPPAKIARLEQNGSP	None
K92	HYENAIEYDCKEEHAEFVLVR	None
K175	TLKIQLHSCPKLEDLPPEQWS	High
K198	TVRNALKDLLKDMNQSSLAKE	None
K244	QEFGRWYKHFKKTKDMMVEMD	None
K245	EFGRWYKHFKKTKDMMVEMDS	None
K384	EIYQWVRDELKRAGISQAVFA	None
K475	STPPSRPPQVKTATIATERNG	None
K486	TATIATERNGKPENNTMNINA	None
K523	QALFAKVAATKSQGWLCELLR	None
K535	QGWLCELLRWKEDPSPENRTL	None
K650	ENRQKTRPRTKISVEALGILQ	None
K720	SGLEVDVAEYKEEELLKDLEE	None
K726	VAEYKEEELLKDLEESVQDKN	None
K744	DKNTNTLFSVKLEEELSVEGN	High

Figure S10. SUMOylation of SATB1 protein truncating variants escaping NMD. A) Schematic representation of the UBC9-SATB1 fusion protein with an N-terminal V5 epitope tag. **B**) Prediction of putative SATB1 (Uniprot Q01826) SUMOylation sites using Joined Advanced SUMOylation Site and SIM Analyser (JASSA, www.jassa.fr/). JASSA uses a scoring system based on a Position Frequency Matrix derived from the alignment of experimental SUMOylation sites. K175 corresponds to a direct consensus site ([Ψ]-[K]-[x]-[α], with Ψ = A,F,I,L,M,P,V or W; α = D or E) with a high prediction score (PS), and K744 to a negatively charged amino acid-dependent SUMOylation site (NDSM, $[\Psi]\text{-}[K]\text{-}[x]\text{-}[\alpha]\text{-}[x]\text{-}[\alpha]_6$ with Ψ = A,F,I,L,M,P,V or W; 2 out of 6 α must be D or E) with a high PS. **C**) Gel shift assay for SATB1 SUMOylation. UBC9-SATB1 and a p.K175R or p.K744R mutant were expressed in HEK293T/17 cells together with a YFP-fusion of SUMO1. Top panel: western blot probed with anti-V5 antibody to detect UBC9-SATB1. The 110 kDa species is unmodified UBC9-SATB1. The 130 kDa species is UBC9-SATB1 modified with endogenous SUMO1. The 170 kDa species is UBC9-SATB1 modified with YFP-SUMO1. Middle panel: western blot probed with anti-YFP antibody, with unconjugated YFP-SUMO1 indicated with an arrow head. Higher molecular weight species are cellular proteins modified with YFP-SUMO1. Bottom panel: western blot probed with anti-ACTB to confirm equal protein loading. **D**) Gel-shift assay for SUMOylation of a SATB1 p.K175R/p.K744R double-mutant. **E**) Gel-shift assay for SUMOylation of SATB1 NMD escaping protein truncating variants.

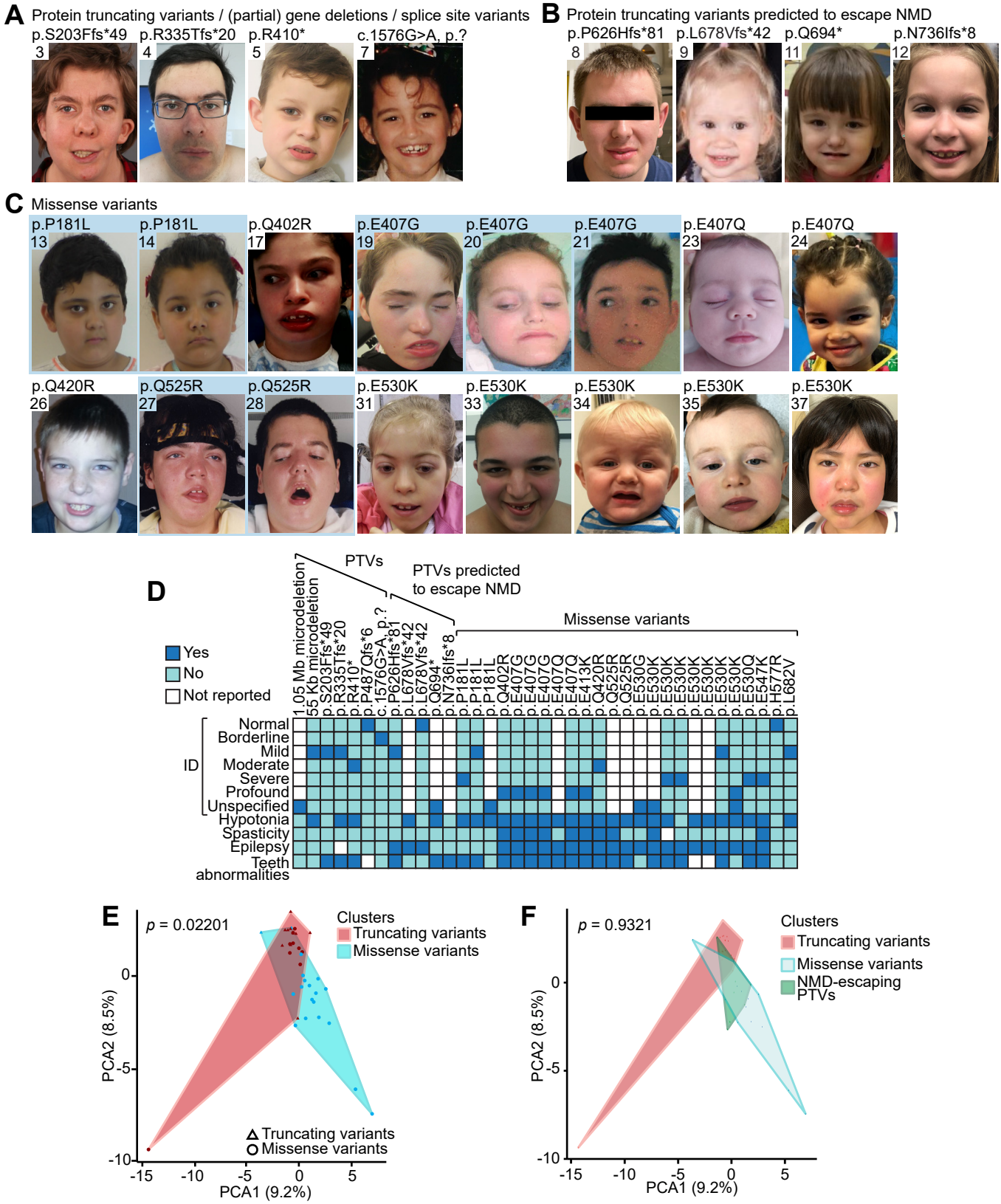
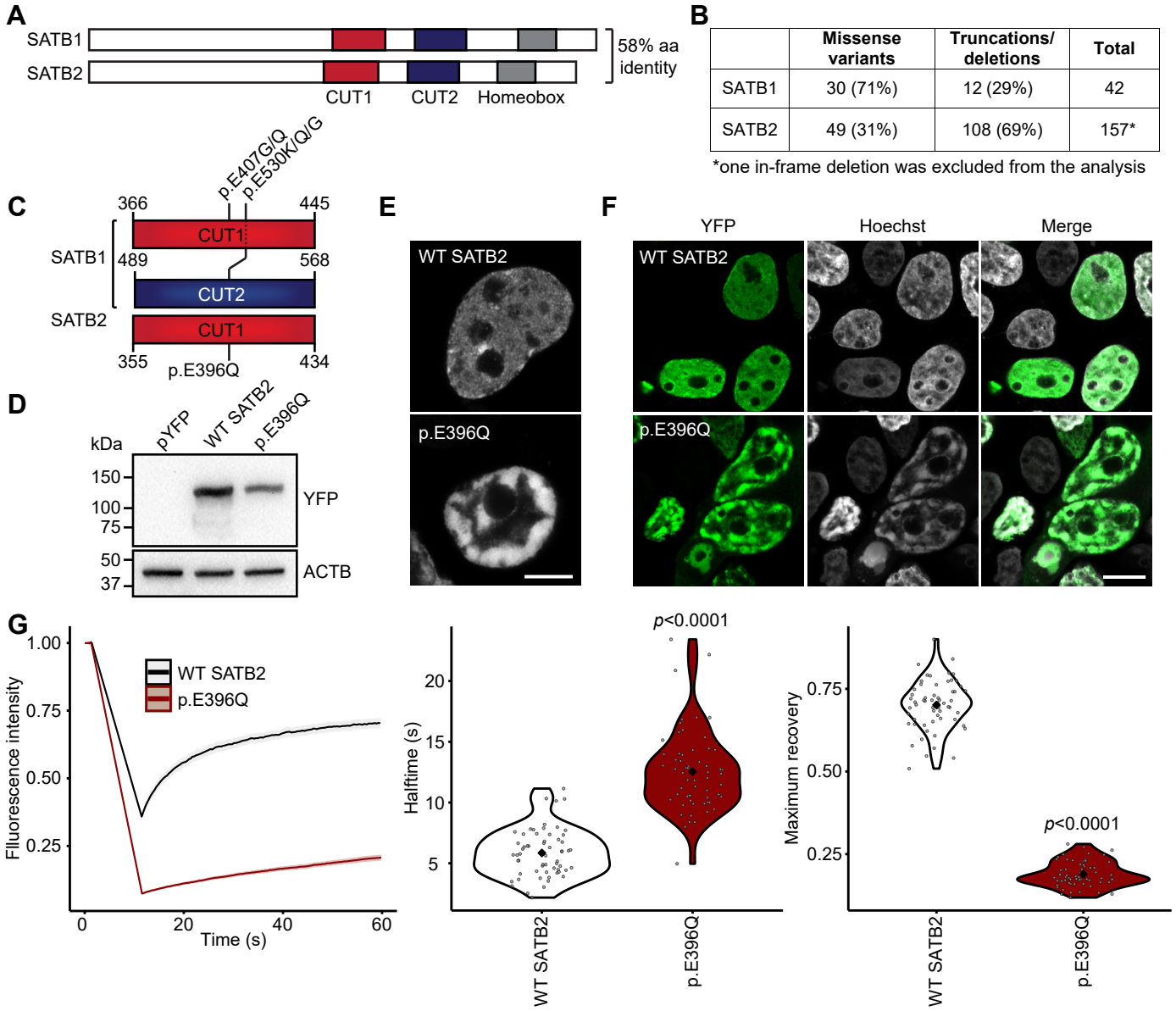


FIgure S11. Clinical evaluation of individuals with *SATB1* variants in three subcohorts. A-C) Facial photographs of individuals with (partial) gene deletions and truncations predicted to result in haploinsufficiency (**A**), of individuals with truncations predicted to escape from NMD and resulting in transcriptionally active proteins (**B**) and of individuals with missense variants (**C**). All depicted individuals show facial dysmorphisms and although overlapping features are seen, no consistent facial phenotype can be observed for the group as a whole. Overlapping facial dysmorphisms include facial asymmetry, high forehead, prominent ears, straight and/or full eyebrows, puffy eyelids, downslant of palpebral fissures, low nasal bridge, full nasal tip and full nasal alae, full lips with absent cupid's bow, prominent cupid's bow or thin upper lip vermilion (Table S1B). Individuals with missense variants are more alike than individuals in the truncating cohorts, and we observed recognizable overlap between several individuals in the missense cohort (individuals 17, 27, 31, 37, the siblings 19, 20 and 21, and to a lesser extent individuals 24 and 35). A recognizable facial overlap between individuals with the other two variant types could not be observed. Related individuals are marked with a blue box. **D**) Mosaic plot presenting a selection of clinical features. Individuals with no or very limited clinical data were omitted (for details, see Supplemental Materials and Methods). **E**) The Partitioning Around Medoids analysis of clustered HPO-standardized clinical data from 38 individuals with truncating (triangle) and missense variants (circle) shows a significant distinction between the clusters of individuals with missense variants (blue) and individuals with PTVs (red). Applying Bonferroni correction, a p-value smaller than 0.025 was considered significant. **F**) Plot of Partitioning Around Medoids clustering analysis on clustered clinical data (HPO) showing no significant distinctions between individuals with missense variants, individuals with truncating variants and deletions, and individuals with NMD-escaping truncating variants



	Missense variants	Truncations/ deletions	Total
SATB1	30 (71%)	12 (29%)	42
SATB2	49 (31%)	108 (69%)	157*

*one in-frame deletion was excluded from the analysis

Figure S12. The SATB2 p.E396Q missense variant has comparable effects on protein functions as the p.E407G and p.E530K/Q SATB1 variants affecting equivalent positions. **A**) SATB1 and SATB2 are highly conserved paralogs. **B**) In SATB1 more missense variants (71%) than truncations/deletions (29%) are observed, while for SATB2 the reverse is reported (31% versus 69% respectively). **C**) Schematic representation of SATB1 and SATB2 CUT DNA binding domains, with variants on equivalent positions indicated. **D**) Immunoblot of whole-cell lysates expressing YFP-tagged SATB2 and p.E396Q probed with anti-EGFP antibody. Expected molecular weight is ~112 kDa. The blot was probed for ACTB to ensure equal protein loading. **E**) Direct fluorescence super-resolution imaging of nuclei of HEK293T/17 cells expressing YFP-SATB2 fusion proteins. Scale bar = 5 µm. **F**) Direct fluorescence imaging of HEK293T/17 cells expressing YFP-SATB2 fusion proteins (green). Nuclei were stained with Hoechst 33342 (white). Scale bar = 10 µm. **G)** FRAP experiments to assess the dynamics of SATB2 chromatin binding in live cells. Left, mean recovery curves ± 95% C.I. recorded in HEK293T/17 cells expressing YFP-SATB2 fusion proteins. Right, violin plots with median of the halftime and maximum recovery values based on single-term exponential curve fitting of individual recordings (n = 60 nuclei from three independent experiments, p-values compared to WT SATB2, unpaired t-test).

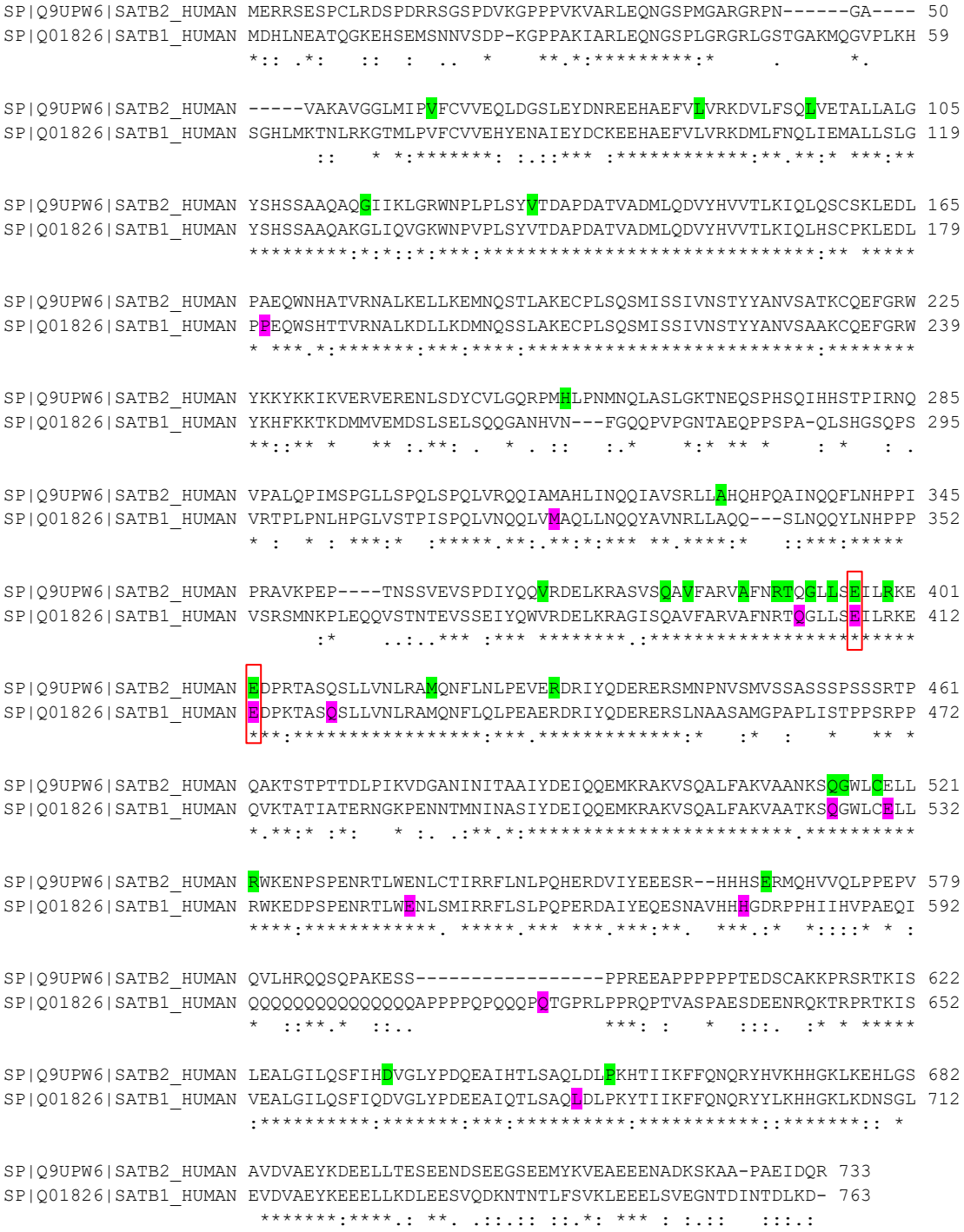


Figure S13. Missense variants identified in individuals with NDD displayed in an amino acid sequence alignment of SATB2 and SATB1. SATB2 (Q9UPW6, UniProt) sequence is aligned to SATB1 (Q01826) sequence. Alignment was performed with Clustal Omega (1.2.4) with default settings using UnitProt alignment tool. Previously reported missense variants in SATB2 (PMID: 31021519) are shaded in green, SATB1 missense variants (this study) are shaded in magenta. Only two missense variants occur at equivalent positions (marked with a red box): SATB2 p.E396Q is equivalent to SATB1 p.E407G/Q, and SATB2 p.E402K is equivalent to SATB1 p.E413K. We functionally characterized SATB2 p.E396Q (Figure S12).

Table S2. Splice-AI predictions for missense variants at intron-exon or exon-intron junctions.

g.DNA-position	c.DNA	Protein effect	spliceAI-G delta score§ - acceptor gain (position*)	spliceAI-G delta score§ - acceptor loss (position*)	spliceAI-G delta score§ -donor gain (position*)	spliceAI-G delta score§ -donor loss (position*)
Chr3:g.18435955T>C	c.1205A>G	p.Q402R¥	0 (-1)	0 (45)	0.0099 (32)	0.2482 (-1)
Chr3:g.18419663T>C	c.1574A>G	p.Q525R£	0 (-1)	0 (19)	0 (20)	0 (-1)
Chr3:g.18393687C>T	c.1576G>A	p.G526R#	0.6666 (-2)	0.0937 (0)	0 (-2)	0 (-17)

*a negative nucleotide position represents positions upstream of the variant, a positive nucleotide position represents positions downstream of the variant.
§cut offs for splice-AI delta score: 0.2 (high recall), 0.5 (recommended), and 0.8 (high precision)

¥p.Q402R:

- Although the variant affects the last amino acid of exon 7, none of the Splice-AI delta scores exceeds the recommended cut-off of >0.5, specifically not the scores for loss or gain of splice donor sites.

£p.Q525R

- Although the variant affects the last amino acid of exon 9, none of the Splice-AI delta scores exceeds the recommended cut-off of >0.5, specifically not the scores for loss or gain of splice donor sites.

#p.G526R:

- The variant affects the first amino acid of exon 10. Splice-AI predicts splice acceptor site gain 2 nucleotides upstream of the variant, resulting in a frameshift.

Table S3. Phenotypic information of individuals from the UK10K cohort with rare *SATB1* missense variants. Wechsler Intelligence Scale for Children (WISC) test scores for individuals from the UK10K cohort, carrying rare *SATB1* missense variants. Standard deviation scores (std score) were calculated by comparing individual scores of carriers to the mean test scores from UK10K non-carriers. Test scores that were lower compared to mean non-carrier scores are shaded in red, while test score that were higher compared to mean non-carrier scores are shaded in green. All carrier test scores were within 2.5 standard deviations compared to the mean non-carrier scores, and thus within normal range.

	UK10K non-carriers (n=1732, ±Std)	UK10K carriers (n=9, ±Std)	UK10K carriers								
Variant	-	-	rs148337599	rs148337599	rs148337599	rs148337599	rs148337599	rs760272331	rs760272331	rs185604711	rs185604711
Residue change (SATB1 NM_001131010.4)	-	-	p.S366L	p.S366L	p.S366L	p.S366L	p.S366L	p.V519L	p.V519L	p.A573T	p.A573T
gnomAD v2.1.1 frequency			6.61e-4 (allele 282848)					8.67e-6 (allele 230660)		1.17e-4 (allele 282890)	
WISC - Verbal IQ: F@8	111.92 (±16.57)	116.89 (±16.07)	103	133	139	103	111	133	128	99	103
VIQ std score	-	0.30 (±0.97)	-0.54	1.27	1.63	-0.54	-0.06	1.27	0.97	-0.78	-0.54
WISC - Performance IQ: F@8	103.49 (±16.79)	109.00 (±15.33)	104	125	115	119	109	115	90	80	124
PIQ std score	-	0.33 (±0.91)	0.03	1.28	0.69	0.92	0.33	0.69	-0.80	-1.40	1.22
WISC - Total IQ: F@8	109.12 (±16.00)	114.89 (±14.73)	104	133	132	111	111	130	111	88	114
IQ std score	-	0.36 (±0.92)	-0.32	1.49	1.43	0.12	0.12	1.30	0.12	-1.32	0.31
WISC - Verbal Comprehension Index: F@8	48.47 (±11.10)	51.79 (±12.51)	38	63	72	44	47	58	63	37	44
VCI std score	-	0.30 (±1.13)	-0.94	1.31	2.12	-0.40	-0.13	0.86	1.31	-1.03	-0.40
WISC - Perceptual Organisation Index: F@8	41.51 (±10.50)	43.80 (±8.37)	41	49	50	53	44	50	30	31	47
POI std score	-	0.23 (±0.80)	-0.05	0.71	0.81	1.09	0.24	0.81	-1.10	-1.00	0.52
WISC - Freedom from Distractability Index: F@8	22.17 (±5.96)	24.11 (±5.42)	27	26	22	22	28	35	19	20	18
FDI std score	-	0.33 (±0.91)	0.81	0.64	-0.03	-0.03	0.98	2.15	-0.53	-0.36	-0.70

≦-1.0 Std score compared to UK10K non-carriers
≦-0.0 Std score compared to UK10K non-carriers
≧1.0 Std score compared to UK10K non-carriers
≧0.0 Std score compared to UK10K non-carriers

Table S4. NMD efficacy predictions for SATB1 truncating variants.

(Hg19/GRCh37)g.DNA-position	g.DNA-position of introduced (downstream) stopcodon	c.DNA-position (NM_001131010.4)	Protein effect**	NMDetective A¥ (1)	NMDetective A¥ (2)	NMDetective B¥ (1)	NMDetective B¥ (2)	Conclusion based on predictions with NMDetectiveA/B	Prediction based on canonical# and non-canonical§ NMD rules
Chr3:g.18456634_18456635delCT	Chr3:g.18436407	c.607_608delAG	p.S203Ffs*49	0.63	0.45	0.65	0.41	Conflicting; NMDetectiveA/B (1): triggers NMD, NMDetectiveA/B (2): intermediate NMD efficacy.	Triggers NMD, none of (non)-canonical NMD rules applicable
Chr3:g.18436155_18436156delTC	Chr3:g.18436098	c.1004_1005delGA	p.R335Tfs*20	0.51	0.52	0.41	0.41	Intermediate NMD efficacy	Might escape from NMD. None of canonical NMD rules applicable, non-canonical long-exon rule applicable (exon 7; 454 nucleotides).
Chr3:g.18428082G>A	Chr3:g.18428082	c.1228C>T	p.R410*	0.6		0.65		Triggers NMD	Triggers NMD, none of (non)-canonical NMD rules applicable
Chr3:g.18419777delG	Chr3:g.18419762	c.1460delC	p.P487Qfs*6	0.62	0.62	0.65	0.65	Triggers NMD	Triggers NMD, none of (non)-canonical NMD rules applicable
Chr3:g.18393687C>T	Chr3:g.18393611	c.1576G>A	p.(?)	0.57	0.6	0.65	0.65	Triggers NMD	Triggers NMD, none of (non)-canonical NMD rules applicable
Chr3:g.18391077delG	Chr3:g.18390837	c.1877delC	p.P626Hfs*81	0.08	0.26	0	0	Conflicting; NMDetectiveA/B (1) and NMDetectiveB (2): escapes NMD; NMDetectiveA (2): intermediate NMD efficacy,	Escapes NMD based on canonical last exon rule
Chr3:g.18390921_18390922delCA	Chr3:g.18390797	c.2032_2033delCT	p.L678Vfs*42	0.18	0.17	0	0	Escapes NMD	Escapes NMD based on canonical last exon rule
Chr3:g.18390874G>A	Chr3:g.18390874	c.2080C>T	p.Q694*	0.2		0		Escapes NMD	Escapes NMD based on canonical last exon rule
Chr3:g.18390747delT	Chr3:g.18390726	c.2207delA	p.N736Ifs*8	0.16	0.16	0	0	Escapes NMD	Escapes NMD based on canonical last exon rule

**For frameshift mutations, scores for NMDetectiveA and NMDetectiveB were assigned both based on the genomic location of the indel (1) and based on the genomic location of the first downstream stopcodon in the new reading frame (2; first nucleotide of introduced stopcodon) (PMID: 31659324). For splice site mutations, NMDetectiveA and NMDetectiveB were assigned based on the effect predicted by spliceAI (PMID: 30661751).

¥NMDetectiveA and NMDetectiveB cut-off scores (v2):
<0.25 predicted to escape NMD
≥0.25 - ≤0.52 predicted intermediate NDM efficacy
>0.52 predicted to trigger NMD (PMID: 31659324)

#Canonical rules of NMD (PMID: 27618451):
NMD is typically not triggered when the location of the protein truncating variant is
1. less than 50 nucleotides upstream of last exon-exon junction; or
2. in the last exon.

§Non-canonical rules of NMD (PMID: 27618451):
NMD is not triggered when the location of the protein truncating variant is
1. in a very long exon (> ±400 nucleotides); or
2. within 150 nucleotides from the start codon.

$ - predicted amino acid sequences of NMD-escaping truncating variants in SATB1

Amino acid sequence of SATB1 (NM_002971.4/NM_001131010.4) in the normal situation
MDHLNEATQGKEHSEMSNNVSDPKGPPAKIARLEQNGSPLGRGRLGSTGAKMQGVPLKHSGHLMKTNLRKGTMLPVFCVVEH
YENAIEYDCKEEHAEFVLVRKDMLFNQLIEMALLSLGYSHSSAAQAKGLIQVGKWNPVPLSYVTDAPDATVADMLQDVYHVVTLK
IQLHSCPKLEDLPPEQWSHTTVRNALKDLLKDMNQSSLAKECPLSQSMISSIVNSTYYANVSAAKCQEFGRWYKHFKKTKDMMV
EMDSLSELSQQGANHVNFGQQPVPGNTAEQPPSPAQLSHGSQPSVRTPLPNLHPGLVSTPISPQLVNQQLVMAQLLNQQYAVN
RLLAQQSLNQQYLNHPPPVSRSMNKPLEQQVSTNTEVSSEIYQWVRDELKRAGISQAVFARVAFNRTQGLLSEILRKEEDPKTAS
QSLLVNLRAMQNFLQLPEAERDRIYQDERERSLNAASAMGPAPLISTPPSRPPQVKTATIATERNGKPENNTMNINASIYDEIQQE
MKRAKVSQALFAKVAATKSQGWLCELLRWKEDPSPENRTLWENLSMIRRFLSLPQPERDAIYEQESNAVHHHGDRPPHIIHVPA
EQIQQQQQQQQQQQQQQQAPPPPQPQQQPQTGPRLPPRQPTVASPAESDEENRQKTRPRTKISVEALGILQSFIQDVGLYPDE
EAIQTLSAQLDLPKYTIIKFFQNQRYYLKHHGKLKDNSGLEVDVAEYKEEELLKDLEESVQDKNTNTLFSVKLEEELSVEGNTDINT
DLKD

Aminoacid sequence of SATB1 (NM_002971.4/NM_001131010.4) in patient 5
*Chr3:g.18428082G>A; c.1228C>T; p.R410**
MDHLNEATQGKEHSEMSNNVSDPKGPPAKIARLEQNGSPLGRGRLGSTGAKMQGVPLKHSGHLMKTNLRKGTMLPVFCVVEH
YENAIEYDCKEEHAEFVLVRKDMLFNQLIEMALLSLGYSHSSAAQAKGLIQVGKWNPVPLSYVTDAPDATVADMLQDVYHVVTLK
IQLHSCPKLEDLPPEQWSHTTVRNALKDLLKDMNQSSLAKECPLSQSMISSIVNSTYYANVSAAKCQEFGRWYKHFKKTKDMMV
EMDSLSELSQQGANHVNFGQQPVPGNTAEQPPSPAQLSHGSQPSVRTPLPNLHPGLVSTPISPQLVNQQLVMAQLLNQQYAVN
RLLAQQSLNQQYLNHPPPVSRSMNKPLEQQVSTNTEVSSEIYQWVRDELKRAGISQAVFARVAFNRTQGLLSEIL*

Aminoacid sequence of SATB1 (NM_002971.4/NM_001131010.4) in patient 8
*Chr3:g.18391077del; c.1877del; p.P626Hfs*81*
MDHLNEATQGKEHSEMSNNVSDPKGPPAKIARLEQNGSPLGRGRLGSTGAKMQGVPLKHSGHLMKTNLRKGTMLPVFCVVEH
YENAIEYDCKEEHAEFVLVRKDMLFNQLIEMALLSLGYSHSSAAQAKGLIQVGKWNPVPLSYVTDAPDATVADMLQDVYHVVTLK
IQLHSCPKLEDLPPEQWSHTTVRNALKDLLKDMNQSSLAKECPLSQSMISSIVNSTYYANVSAAKCQEFGRWYKHFKKTKDMMV
EMDSLSELSQQGANHVNFGQQPVPGNTAEQPPSPAQLSHGSQPSVRTPLPNLHPGLVSTPISPQLVNQQLVMAQLLNQQYAVN
RLLAQQSLNQQYLNHPPPVSRSMNKPLEQQVSTNTEVSSEIYQWVRDELKRAGISQAVFARVAFNRTQGLLSEILRKEEDPKTAS
QSLLVNLRAMQNFLQLPEAERDRIYQDERERSLNAASAMGPAPLISTPPSRPPQVKTATIATERNGKPENNTMNINASIYDEIQQE
MKRAKVSQALFAKVAATKSQGWLCELLRWKEDPSPENRTLWENLSMIRRFLSLPQPERDAIYEQESNAVHHHGDRPPHIIHVPA
EQIQQQQQQQQQQQQQQQAPPPPQPQQQPQTGPRLPHGNPRWPLQQSQMRKTDRRPGHEQKFQWKPWESSRVSYKTWA
CTLTKRPSRLCLPSSTFPSTPSSSSFRTSGTISSTTAN*

Aminoacid sequence of SATB1 (NM_002971.4/NM_001131010.4) in patient 9 and 10 *Chr3:g.18390921_18390922del; c.2032_2033del; p.L678Vfs*42*
MDHLNEATQGKEHSEMSNNVSDPKGPPAKIARLEQNGSPLGRGRLGSTGAKMQGVPLKHSGHLMKTNLRKGTMLPVFCVVEH
YENAIEYDCKEEHAEFVLVRKDMLFNQLIEMALLSLGYSHSSAAQAKGLIQVGKWNPVPLSYVTDAPDATVADMLQDVYHVVTLK
IQLHSCPKLEDLPPEQWSHTTVRNALKDLLKDMNQSSLAKECPLSQSMISSIVNSTYYANVSAAKCQEFGRWYKHFKKTKDMMV
EMDSLSELSQQGANHVNFGQQPVPGNTAEQPPSPAQLSHGSQPSVRTPLPNLHPGLVSTPISPQLVNQQLVMAQLLNQQYAVN
RLLAQQSLNQQYLNHPPPVSRSMNKPLEQQVSTNTEVSSEIYQWVRDELKRAGISQAVFARVAFNRTQGLLSEILRKEEDPKTAS
QSLLVNLRAMQNFLQLPEAERDRIYQDERERSLNAASAMGPAPLISTPPSRPPQVKTATIATERNGKPENNTMNINASIYDEIQQE
MKRAKVSQALFAKVAATKSQGWLCELLRWKEDPSPENRTLWENLSMIRRFLSLPQPERDAIYEQESNAVHHHGDRPPHIIHVPA
EQIQQQQQQQQQQQQQQQAPPPPQPQQQPQTGPRLPPRQPTVASPAESDEENRQKTRPRTKISVEALGILQSFIQDVGLYPDE
EAIQTVCPARPSQVHHHQVLSEPAVLSQAPRQTEGQFRFRGRCGRI*

Aminoacid sequence of SATB1 (NM_002971.4/NM_001131010.4) in patient 11
*Chr3:g. 18390874G>A; c.2080C>T; p.Q694**
MDHLNEATQGKEHSEMSNNVSDPKGPPAKIARLEQNGSPLGRGRLGSTGAKMQGVPLKHSGHLMKTNLRKGTMLPVFCVVEH
YENAIEYDCKEEHAEFVLVRKDMLFNQLIEMALLSLGYSHSSAAQAKGLIQVGKWNPVPLSYVTDAPDATVADMLQDVYHVVTLK
IQLHSCPKLEDLPPEQWSHTTVRNALKDLLKDMNQSSLAKECPLSQSMISSIVNSTYYANVSAAKCQEFGRWYKHFKKTKDMMV
EMDSLSELSQQGANHVNFGQQPVPGNTAEQPPSPAQLSHGSQPSVRTPLPNLHPGLVSTPISPQLVNQQLVMAQLLNQQYAVN
RLLAQQSLNQQYLNHPPPVSRSMNKPLEQQVSTNTEVSSEIYQWVRDELKRAGISQAVFARVAFNRTQGLLSEILRKEEDPKTAS
QSLLVNLRAMQNFLQLPEAERDRIYQDERERSLNAASAMGPAPLISTPPSRPPQVKTATIATERNGKPENNTMNINASIYDEIQQE
MKRAKVSQALFAKVAATKSQGWLCELLRWKEDPSPENRTLWENLSMIRRFLSLPQPERDAIYEQESNAVHHHGDRPPHIIHVPA
EQIQQQQQQQQQQQQQQQAPPPPQPQQQPQTGPRLPPRQPTVASPAESDEENRQKTRPRTKISVEALGILQSFIQDVGLYPDE
EAIQTLSAQLDLPKYTIIKFF*

Aminoacid sequence of SATB1 (NM_002971.4/NM_001131010.4) in patient 12
*Chr3:g.18390747del; c.2207del; p.N736Ifs*8*
MDHLNEATQGKEHSEMSNNVSDPKGPPAKIARLEQNGSPLGRGRLGSTGAKMQGVPLKHSGHLMKTNLRKGTMLPVFCVVEH
YENAIEYDCKEEHAEFVLVRKDMLFNQLIEMALLSLGYSHSSAAQAKGLIQVGKWNPVPLSYVTDAPDATVADMLQDVYHVVTLK
IQLHSCPKLEDLPPEQWSHTTVRNALKDLLKDMNQSSLAKECPLSQSMISSIVNSTYYANVSAAKCQEFGRWYKHFKKTKDMMV
EMDSLSELSQQGANHVNFGQQPVPGNTAEQPPSPAQLSHGSQPSVRTPLPNLHPGLVSTPISPQLVNQQLVMAQLLNQQYAVN
RLLAQQSLNQQYLNHPPPVSRSMNKPLEQQVSTNTEVSSEIYQWVRDELKRAGISQAVFARVAFNRTQGLLSEILRKEEDPKTAS
QSLLVNLRAMQNFLQLPEAERDRIYQDERERSLNAASAMGPAPLISTPPSRPPQVKTATIATERNGKPENNTMNINASIYDEIQQE
MKRAKVSQALFAKVAATKSQGWLCELLRWKEDPSPENRTLWENLSMIRRFLSLPQPERDAIYEQESNAVHHHGDRPPHIIHVPA
EQIQQQQQQQQQQQQQQQAPPPPQPQQQPQTGPRLPPRQPTVASPAESDEENRQKTRPRTKISVEALGILQSFIQDVGLYPDE
EAIQTLSAQLDLPKYTIIKFFQNQRYYLKHHGKLKDNSGLEVDVAEYKEEELLKDLEESVQDKILTPFFQ*

Table S5. Summary of clinical characteristics associated with *(de novo) SATB1* PTVs and (partial) gene deletions predicted to result in haploinsufficiency and PTVs in the last exon.

	Individuals with PTVs and (partial) gene deletions predicted to result in haploinsufficiency		Individuals with PTVs in the last exon	
	%	**Present / total assessed**	**%**	**Present / total assessed**
Neurologic				
Intellectual disability	86	6/7	67	2/3
Normal	14	1/7	33	1/3
Borderline	0	0/7	0	0/3
Mild	71	5/7	33	1/3
Moderate	14	1/7	0	0/3
Severe	0	0/7	0	0/3
Profound	0	0/7	0	0/3
Unspecified	0	0/7	33	1/3
Developmental delay	100	7/7	100	5/5
Motor delay	86	6/7	100	5/5
Speech delay	86	6/7	80	4/5
Dysarthria	14	1/7	0	0/4
Epilepsy	0	0/6	40	2/5
EEG abnormalities	0	0/4	67	2/3
Hypotonia	43	3/7	40	2/5
Spasticity	0	0/7	0	0/5
Ataxia	14	1/7	20	1/5
Behavioral disturbances	100	7/7	0	0/5
Sleep disturbances	50	3/6	0	0/5
Abnormal brain imaging	33	1/3	50	2/4
Regression	14	1/7	0	0/5
Growth				
Abnormalities during pregnancy	33	2/6	20	1/5
Abnormalities during delivery	33	2/6	80	4/5
Abnormal term of delivery	0	0/5	20	1/5
Preterm (<37 weeks)	0	0/5	20	1/5
Postterm (>42 weeks)	0	0/5	0	0/5
Abnormal weight at birth	20	1/5	25	1/4
Small for gestational age (<p10)	20	1/5	0	0/4
Large for gestational age (>p90)	0	0/5	25	1/4
Abnormal head circumference at birth	25	1/4	0	0/2
Microcephaly (<p3)	0	0/4	0	0/2
Macrocephaly (>p97)	25	1/4	0	0/2
Abnormal height	14	1/7	0	0/4
Short stature (<p3)	0	0/7	0	0/4
Tall stature (>p97)	14	1/7	0	0/4
Abnormal head circumference	0	0/5	25	1/4
Microcephaly (<p3)	0	0/5	25	1/4
Macrocephaly (>p97)	0	0/5	0	0/4
Abnormal weight	0	0/5	25	1/4
Underweight (<p3)	0	0/5	25	1/4
Overweight (>p97)	0	0/5	0	0/4
Other phenotypic features				
Facial dysmorphisms	67	4/6	60	3/5
Dental/oral abnormalities	50	3/6	60	3/5
Drooling/dysphagia	29	2/7	20	1/5
Hearing abnormalities	17	1/6	20	1/5
Vision abnormalities	67	4/6	80	4/5
Cardiac abnormalities	17	1/6	40	2/5
Skeleton/limb abnormalities	33	2/6	0	0/5
Hypermobility of joints	33	2/6	25	1/4
Gastrointestinal abnormalities	33	2/6	20	1/5
Urogenital abnormalities	0	0/6	0	0/5
Endocrine/metabolic abnormalities	0	0/6	0	0/5
Immunological abnormalities	17	1/6	50	1/2
Skin/hair/nail abnormalities	0	0/6	20	1/5
Neoplasms in medical history	0	0/6	0	0/5

Table S8. Primers for site-directed mutagenesis

SATB1-K175R-F	GGAGGCAAGTCTTCTAGTCGGGGGCAACTGTGTAACTG
SATB1-K175R-R	CAGTTACACAGTTGCCCCCGACTAGAAGACTTGCCTCC
SATB1-S366L-F	TCTGTGTTGGTCAAAACCTGTTGCTCCAAAGGCT
SATB1-S366L-R	AGCCTTTGGAGCAACAGGTTTTGACCAACACAGA
SATB1-E407G-F	CTTCCTTTCGGAGGATTCCTGAAAGCAAGCCCTGA
SATB1-E407G-R	TCAGGGCTTGCTTTCAGGAATCCTCCGAAAGGAAG
SATB1-R410*	GGGGTCCTCTTCCTTTCAGAGGATTTCTGAAAGCA
SATB1-R410*	TGCTTTCAGAAATCCTCTGAAAGGAAGAGGACCCC
SATB1-Q420R-F	GTTTACCAGCAAAGACCGGGATGCAGTCTTGGG
SATB1-Q420R-R	CCCAAGACTGCATCCCGGTCTTTGCTGGTAAAC
SATB1-E530K-F	TCCAGCGTAACAGCTTGCACAACCATCCCTG
SATB1-E530K-R	CAGGGATGGTTGTGCAAGCTGTTACGCTGGA
SATB1-E530Q-F	CCAGCGTAACAGCTGGCACAACCATCCCT
SATB1-E530Q-R	AGGGATGGTTGTGCCAGCTGTTACGCTGG
SATB1-E547K-F	GATCATGGAGAGGTTCTTCCACAGGGTTCTGTTTT
SATB1-E547K-R	AAAACAGAACCCTGTGGAAGAACCTCTCCATGATC
SATB1-V519L-F	GCTTTTGGTTGCTGCAAGCTTTGCAAACAGTGCTT
SATB1-V519L-R	AAGCACTGTTTGCAAAGCTTGCAGCAACCAAAAGC
SATB1-A573T-F	CATGGTGATGCACCGTGTTGCTCTCCTGTTC
SATB1-A573T-R	GAACAGGAGAGCAACACGGTGCATCACCATG
SATB1-P626Hfs*81-F	GTGGGTTGCCGTGGGGGAGCCGAG
SATB1-P626Hfs*81-R	CTCGGCTCCCCCACGGCAACCCAC
SATB1-L682V-F	CTTGGGAAGGTCGACCTGGGCAGACAGAG
SATB1-L682V-R	CTCTGTCTGCCCAGGTCGACCTTCCCAAG
SATB1-Q694*-F	TACCGCTGGTTCTAAAAGAACTTGATGATGGTGTACTTG
SATB1-Q694*-R	CAAGTACACCATCATCAAGTTCTTTTAGAACCAGCGGTA
SATB1-N736I*8-F	AAAAAGGGTGTTAGTATTTTATCTTGGACACTCTCTTCCAAATCCT
SATB1-N736I*8-R	AGGATTTGGAAGAGAGTGTCCAAGATAAAATACTAACACCCTTTTT
SATB1-K744R-F	CACTGACAGCTCTTCTTCTAGTCGCACTGAAAAAAGGGTGTTAGTA
SATB1-K744R-R	TACTAACACCCTTTTTTCAGTGCGACTAGAAGAAGAGCTGTCAGTG
SATB2-E396Q-F	TACGCAGAATCTGAGACAACAATCCCTGTGTGCGG
SATB2-E396Q-R	CCGCACACAGGGATTGTTGTCTCAGATTCTGCGTA

Table S9. Primers for amplifying and subcloning human UBC9 (NM_194260.2) and SATB1 (NM_001131010.4). Sequences of restriction sites are shown in bold, and sequences that were added to extend the linker region between UBC9 and SATB1 are underscored.

UBC9-*BamHI*-F	GAGGGA**GGATCC**TGCTGTCGGGGATCGCCCTCAG
UBC9-*XmaI*-R	TCTAGA**CCCGGG**CAGCGCAAGTGAGGGCGCAAACTTCTTGG
SATB1-*HindIII*-F	CGGTAC**AAGCTT**TTGGCTGTACTGGATCATTTGAACGAGGC
SATB1-*XhoI*-R	CAGTTA**CTCGAG**TCAGTCTTTCAAATCAGTATTAATGTCTG

Table S10. Primers to amplify regions that include the SATB1 NMD-escaping truncating variants used for testing for NMD. The last exon primer set was used for SATB1 p.P626Hfs*81, p.Q694* and p.N736Ifs*8.

SATB1-NMD-R410*-F	CCTGGGCTCGTATCAACACC
SATB1-NMD-R410*-R	CATCCCTGGCTTTTGGTTGC
SATB1-NMD-last_exon-F	GCCATTTATGAACAGGAGAGCA
SATB1-NMD-last exon-R	CAGTATTAATGTCTGTGTTTCCTTCCA

Supplemental Acknowledgements
We wish to thank all the ALSPAC families who took part in this study, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses. The UK Medical Research Council and Wellcome (Grant ref: 102215/2/13/2) and the University of Bristol provide core support for ALSPAC. A comprehensive list of grants funding is available on the ALSPAC website (http://www.bristol.ac.uk/alspac/external/documents/grant-acknowledgements.pdf). Whole-genome sequencing of the ALSPAC samples was performed as part of the UK10K consortium (a full list of investigators who contributed to the generation of the data is available from www.UK10K.org.uk). This research was made possible through access to the data and findings generated by the 100,000 Genomes Project. The 100,000 Genomes Project is managed by Genomics England Limited (a wholly owned company of the Department of Health and Social Care). The 100,000 Genomes Project is funded by the National Institute for Health Research and NHS England. The Wellcome Trust, Cancer Research UK and the Medical Research Council have also funded research infrastructure. The 100,000 Genomes Project uses data provided by patients and collected by the National Health Service as part of their care and support. In addition, in individual 13, 14 and 15, whole-exome sequencing was performed in the framework of the German project "TRANSLATE NAMSE", an initiative from the National Action League for People with Rare Diseases (Nationales Aktionsbündnis für Menschen mit Seltenen Erkrankungen, NAMSE) facilitating innovative genetic diagnostics for individuals with suggested rare diseases. Part of this work (IT) was performed under the Genomic Answers for Kids program funded by generous donors to the Children's Mercy Research Institute.

Supplemental Materials and Methods

Individuals and consent

For all individuals reported in this study, informed consent was obtained to publish unidentifiable data. When applicable, specific consent was obtained for publication of clinical photographs and inclusion of photographs in facial analysis. All consent procedures are in accordance with both the local ethical guidelines of the participating centers, and the Declaration of Helsinki. Individuals with possible (likely) pathogenic *SATB1* variants were identified through international collaborations facilitated by MatchMakerExchange[1], GPAP of RD-connect[2], the Solve-RD consortium, the Decipher Database[3], and through searching literature for cohort-studies for NDD[4; 5]. Clinical characterization was performed by reviewing the medical files and/or revising the phenotype of the individuals in the clinic. All (affected) individuals with a *SATB1* variant are included in Table S1. A summary of clinical characteristics is provided in Table 1, including 38 of 42 individuals: individual 16, 32 and 41 were excluded because no clinical data were available, individual 22 was excluded as she is (low) mosaic for the *SATB1* variant (~1%). In Figure 1G, 37 of 42 individuals were included: in addition to individuals 16, 22, 32, and 41, we also excluded individual 18, for whom only very limited clinical information was available.

Next generation sequencing

For all individuals except individual 1, 2, and 28, *SATB1* variants were identified by whole exome sequencing after variant filtering as previously described[6-12]. Information on inheritance was obtained after parental confirmation, either from parental exome sequencing data or through targeted Sanger sequencing. For individual 1 the *SATB1* variant was identified by array-CGH and for individual 2 an Affymetrix Cytoscan HD array was performed in addition to whole exome sequencing. For individual 28 targeted Sanger sequencing was performed after identification of the variant in his similarly affected sister. To predict deleteriousness of variants, CADD-PHRED V1.4 scores and SpliceAI scores (VCFv4.2; dated 20191004) were obtained for all variants identified in affected individuals[13; 14]. In addition, for all nonsense, frameshift and

splice site variants, NMDetective scores were obtained (v2)[15]. For all missense variants, we analyzed the mutation tolerance of the site of the affected residue using Metadome[16].

UK10K controls for functional assays

Genome sequence data from 1,867 ALSPAC[17; 18] individuals in the UK10K[19] dataset were annotated in ANNOVAR[20] and filtered to identify individuals carrying rare coding variants (gnomAD genome_ALL frequency<0.1%) within *SATB1*. In total six rare variants were identified. These variants were carried by 13 individuals, all in a heterozygous state. Three variants (one in the CUT1 domain, one in the CUT2 domain and one outside of critical domains) were selected for functional studies. These variants were carried by nine individuals. Phenotypic data of carriers and non-carriers were available through the ALSPAC cohort, an epidemiological study of pregnant women who were resident in Avon, UK with expected dates of delivery 1st April 1991 to 31st December 1992. This dataset included 13,988 children who were alive at 1 year of age, 1,867 of whom underwent genome sequencing as part of the UK10K project. Of the UK10K individuals, 1,741 children had measures of IQ (WISC) collected at age 8 years providing an indication of cognitive development. The ALSPAC study website contains details of all the data that is available through a fully searchable data dictionary and variable search tool (http://www.bristol.ac.uk/alspac/researchers/our-data/)

Human Phenotype Ontology (HPO)-based phenotype clustering analysis

All clinical data were standardized using HPO terminology[21]. Thirty-eight of 42 individuals were included in analysis: individual 16, 32 and 41 were excluded because no clinical data were available, individual 22 was excluded as she is (low) mosaic for the SATB1 variant (~1%). The semantic similarity between all the HPO terms used in this cohort (356 features) was calculated using the Wang algorithm in the HPOSim package[22; 23] in R. HPO terms with at least a 0.5 similarity score were grouped (Figure S5): a new feature was created as a replacement, which was the sum of the grouped features. For eleven terms, the HPO semantic similarity could not be calculated using HPOSim. Seven of those could be manually assigned to a group, since the feature clearly matched (for instance: nocturnal seizures with the seizure/epilepsy group). For a full list of the grouped features, see Table S7. HPO terms that could not be grouped were added as separate features, as was severity of intellectual disability. This led to 100 features for every individual, instead of the previous 356 separate HPO terms. To quantify the possible genotype/phenotype correlation in the cohort, we used Partitioning Around Medoids (PAM) clustering[24] dividing our cohort into two groups (missense variants versus truncating variants), followed by a permutations test (n=100,000) and relabeling based on variant types, while keeping the original distribution of variant types into account. The same clustering and permutations test was performed when dividing our cohort into three groups. For both analyses, Bonferroni correction for multiple testing was applied and a p-value smaller than 0.025 was considered significant.

Average face analysis

For 24 of 42 individuals facial 2D-photographs were available for facial analysis. As previously described, average faces were generated while allowing for asymmetry preservation and equal representation by individuals[25].

Three-dimensional protein modeling

The crystal structure of the CUT1 domain of SATB1 bound to Matrix Attachment Region DNA (PDB entry 2O4A[26]) was used to contextualize the SATB1 CUT1 variants with respect to DNA using Swiss-PdbViewer[27]. The solution structure of the CUT2 domain of human SATB2 (first NMR model of the PDB entry 2CSF[28]) was used as a template to align the SATB1 residues T491 to H577 (Uniprot entry Q01826), and to build a model using Swiss-PdbViewer[27]. The model of the CUT2 domain was superposed onto the SATB1 CUT1 domain bound to Matrix Attachment Region DNA (PDB entry 2O4A[26] using the “magic fit” option of Swiss-PdbViewer[27]) to contextualize the SATB1 CUT2 variants with respect to DNA. The solution structure of the homeodomain of human SATB2 (second NMR model of the PDB entry 1WI3[29] was used as a

template to align SATB1 residues P647 to G704 (Uniprot entry Q01826), and to build a model using Swiss-PdbViewer[27]. Chains A, C and D of the crystal structure of HNF-6alpha DNA-binding domain in complex with the TTR promoter (PDB entry 2D5V), which has a DNA binding domain similar to the CUT2 domain of SATB1 and a second DNA binding domain similar to the homeobox of SATB1, was used as a template to superpose the model of the SATB2 homeobox domain onto the HNF-6alpha structure using the "magic fit" option of Swiss-PdbViewer[30] to contextualize the SATB1 homeobox variant with respect to DNA.

Spatial clustering analysis of missense variants

Twenty-four of the observed 30 missense variants were included in the spatial clustering analysis. We excluded 6 variants, to correct for familial occurrence. The geometric mean was computed over the locations of observed (*de novo)* missense variants in the cDNA of *SATB1* (NM_001131010.4). This geometric mean was then compared to 1,000,000 permutations, by redistributing the (*de novo)* variant locations over the total size of the coding region of *SATB1* (2,388 bp) and calculating the resulting geometric mean from each of these permutations. The *p*-value was then computed by checking how often the observed geometric mean distance was smaller than the permutated geometric mean distance. This approach was previously used to identify cDNA clusters of variants[7; 31].

DNA expression constructs and site-directed mutagenesis

The cloning of SATB1 (NM_001131010.4), SATB2 (NM_001172509) and SUMO1 (NM_003352.4), has been described previously[32; 33]. Variants in SATB1 and SATB2 were generated using the QuikChange Lightning Site-Directed Mutagenesis Kit (Agilent). The primers used for site-directed mutagenesis are listed in Table S8. cDNAs were subcloned using *BamHI*/*XbaI* (SATB1 and SUMO1) and *BclI*/*XbaI* (SATB2) restriction sites into pRluc and pYFP, created by modification of the pEGFP-C2 vector (Clontech) as described before[34]. To generate a UBC9-SATB1 fusion, the UBC9 (NM_194260.2) and SATB1 coding sequences were amplified using primers listed in Table S9, and subcloned into the pHisV5 vector (a modified pEGFP-C2 vector adding an N-terminal His- and V5-tag) using *BamHI*/*SmaI* (UBC9) and *HindIII*/*XhoI* (SATB1) restriction sites. All constructs were verified by Sanger sequencing.

Cell culture

HEK293T/17 cells (CRL-11268, ATCC) were cultured in DMEM supplemented with 10% fetal bovine serum and 1x penicillin-streptomycin (all Invitrogen) at 37°C with 5% CO_2. Transfections for functional assays were performed using GeneJuice (Millipore) following the manufacturer's protocol. Lympoblastoid cell lines (LCLs) were established by Epstein-Barr virus transformation of peripheral lymphocytes from blood samples collected in heparin tubes, and maintained in RPMI medium (Sigma) supplemented with 15% fetal bovine serum and 5% HEPES (both Invitrogen).

Testing for nonsense mediated decay of truncating variants

Patient-derived LCLs were grown for 4 h with 100 µg/ml cycloheximide (Sigma) to block NMD. After treatment, cell pellets ($10*10^6$ cells) were collected and RNA was extracted using the RNeasy Mini Kit (Qiagen). RT-PCR was performed using SuperScriptIII Reverse Transcriptase (ThermoFisher) with random primers, and regions of interest were amplified from cDNA using primers listed in Table S10.

Fluorescence microscopy

HEK293T/17 cells were grown on coverslips coated with poly-D-lysine (Sigma). Cells were fixed with 4% paraformaldehyde (PFA, Electron Microscopy Sciences) 48 h after transfection with YFP-tagged SATB1 and SATB2 variants. Nuclei were stained with Hoechst 33342 (Invitrogen). Fluorescence images were acquired with a Zeiss LSM880 confocal microscope and ZEN Image Software (Zeiss). For images of single nuclei, the Airyscan unit (Zeiss) was used with a 4.5 zoom factor. All other images were acquired with a 2.0 zoom factor. Intensity profiles were plotted using the 'Plot Profile' tool in Fiji - ImageJ.

FRAP assays
HEK293T/17 cells were transfected in clear-bottomed black 96-well plates with YFP-tagged SATB1 and SATB2 variants. After 48 h, medium was replaced with phenol red-free DMEM supplemented with 10% fetal bovine serum (both Invitrogen), and cells were moved to a temperature-controlled incubation chamber at 37°C. Fluorescent recordings were acquired using a Zeiss LSM880 and Zen Black Image Software, with an alpha Plan-Apochromat 100x/1.46 Oil DIC M27 objective (Zeiss). FRAP experiments were performed by photobleaching an area of 0.98 µm x 0.98 µm within a single nucleus with 488-nm light at 100% laser power for 15 iterations with a pixel dwell time of 32.97 µs, followed by collection of times series of 150 images with a 2.5 zoom factor and an optical section thickness of 1.4 µm (2.0 Airy units). Individual recovery curves were background subtracted and normalized to the pre-bleach values, and mean recovery curves were calculated using EasyFRAP software[35]. Curve fitting was done with the FrapBot application using direct normalization and a single-component exponential model, to calculate the half-time and maximum recovery[36].

Luciferase reporter assays
Luciferase reporter assays were performed with a pIL2-luc reporter construct containing the human *IL2*-promoter region, and a pGL3-basic firefly luciferase reporter plasmid carrying seven repeats of the -TCTTTAATTTCTAATATATTTAGAAttc- MAR sequence identified in an enhancer region 3' of the immunoglobulin heavy chain (IgH) genes (gift from Dr. Kathleen McGuire and Dr. Sanjeev Galande), as described previously[37-39]. HEK293T/17 cells were transfected with firefly luciferase reporter constructs and a Renilla luciferase (Rluc) normalization control (pGL4.74; Promega) in a ratio of 50:1, and with pYFP-SATB1 (WT or variant) or empty control vector (pYFP). After 48 h, firefly luciferase and Rluc activity was measured using the Dual-Luciferase Reporter Assay system (Promega) at the Infinite M Plex Microplate reader (Tecan).

BRET saturation assays
BRET assays were performed as previously described[34]. HEK293T/17 cells were transfected in white clear-bottomed 96-well plates with increasing molar ratios of YFP-fusion proteins and constant amounts of Rluc-fusion proteins (donor/acceptor ratios of 1/0.5, 1/1, 1/2, 1/3, 1/6, 1/9). YFP and Rluc fused to a C-terminal nuclear localization signal were used as control proteins. After 48 h, medium was replaced with phenol red-free DMEM, supplemented with 10% fetal bovine serum (both Invitrogen), containing 60 µM EnduRen Live Cell Substrate (Promega). After incubation for 4 h at 37°C, measurements were taken in live cells with an Infinite M200PRO Microplate reader (Tecan) using the Blue1 and Green1 filters. Corrected BRET ratios were calculated with the following formula: [Green1($_{\text{experimental condition}}$)/Blue1($_{\text{experimental condition}}$)] − [Green1($_{\text{control condition}}$)/Blue1($_{\text{control condition}}$)], with only the Rluc control protein expressed in the control condition. YFP fluorescence was measured separately (Ex: 505 nm, Em: 545 nm) to quantify expression of the YFP-fusion proteins. Curve fitting was done with a non-linear regression equation assuming a single binding site using GraphPad Prism Software, after plotting the corrected BRET ratios against the ratio of total luminescence / total YFP fluorescence.

Immunoblotting and gel-shift assays
Whole-cell lysates were collected by treatment with lysis buffer 48 h post-transfection. For immunoblotting, cells were lysed in 1x RIPA buffer (Cell Signalling) with 1% PMSF and protease inhibitor cocktail (Roche). For gel-shift assays[40], cells were lysed in 1x RIPA buffer with 1% PMSF, protease inhibitor cocktail and 50 µM ubiquitin/ubiquitin-like isopeptidases inhibitor PR-619 (Sigma). Samples were incubated for 20 min at 4°C followed by centrifugation for 30 min at 12,000 rpm at 4°C. Proteins were resolved on 4–15% Mini-PROTEAN TGX Precast Gels (Bio-Rad) and transferred onto polyvinylidene fluoride membranes using a TransBlot Turbo Blotting system (Bio-Rad). Membranes were blocked in 5% milk for 1 h at room temperature and then probed with mouse-anti-EGFP (for pYFP constructs; 1:8000;

Clontech, 632380) or mouse-anti-V5 tag (1:2000; Genetex, GTX42525). Next, membranes were incubated with HRP-conjugated goat-anti-mouse IgG (1:2000; Bio-Rad) for 1 h at room temperature. Bands were visualized with Novex ECL Chemiluminescent Substrate Reagent (Invitrogen) using a ChemiDoc XRS + System (Bio-Rad). Equal protein loading was confirmed by probing with mouse-anti-β-actin antibody (1:10,000; Sigma, A5441).

Fluorescence-based quantification of protein stability

Cells were transfected in triplicate in clear-bottomed black 96-well plates with YFP-tagged SATB1 variants. After 24 h, MG132 (R&D Systems) was added at a final concentration of 10 µM, and cycloheximide (Sigma) at 50 µg/ml. Cells were incubated at 37°C with 5% CO_2 in the Infinite M200PRO microplate reader (Tecan), and the fluorescence intensity of YFP (Ex: 505 nm, Em: 545 nm) was measured over 24 h at 3 h intervals.

Statistical analyses of cell-based functional assays

Statistical analyses for cell-based functional assays were done using a one- or two-way ANOVA followed by a Bonferroni *post-hoc* test, with GraphPad Prism Software. Statistical analyses for FRAP and BRET data were performed on values derived from fitted curves of individual recordings or independent experiments respectively.

Data and Code Availability

Code used in the spatial clustering analysis is available at: https://github.com/laurensvdwiel/SpatialClustering. Codes of HPO-based clustering analysis and computational facial averaging are available on request. All available phenotypic data in HPO is shared as a supplementary file (SATB1_supplementaryJSON.json).

References

1. Sobreira, N., Schiettecatte, F., Valle, D., and Hamosh, A. (2015). GeneMatcher: a matching tool for connecting investigators with an interest in the same gene. Hum Mutat 36, 928-930.
2. Thompson, R., Johnston, L., Taruscio, D., Monaco, L., Beroud, C., Gut, I.G., Hansson, M.G., t Hoen, P.B., Patrinos, G.P., Dawkins, H., et al. (2014). RD-Connect: an integrated platform connecting databases, registries, biobanks and clinical bioinformatics for rare disease research. J Gen Intern Med 29 Suppl 3, S780-787.
3. Firth, H.V., Richards, S.M., Bevan, A.P., Clayton, S., Corpas, M., Rajan, D., Van Vooren, S., Moreau, Y., Pettett, R.M., and Carter, N.P. (2009). DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. Am J Hum Genet 84, 524-533.
4. Satterstrom, F.K., Kosmicki, J.A., Wang, J., Breen, M.S., De Rubeis, S., An, J.Y., Peng, M., Collins, R., Grove, J., Klei, L., et al. (2020). Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. Cell 180, 568-584.e523.
5. Kaplanis, J., Samocha, K.E., Wiel, L., Zhang, Z., Arvai, K.J., Eberhardt, R.Y., Gallone, G., Lelieveld, S.H., Martin, H.C., McRae, J.F., et al. (2020). Evidence for 28 genetic disorders discovered by combining healthcare and research data. Nature.
6. Retterer, K., Juusola, J., Cho, M.T., Vitazka, P., Millan, F., Gibellini, F., Vertino-Bell, A., Smaoui, N., Neidich, J., Monaghan, K.G., et al. (2016). Clinical application of whole-exome sequencing across clinical indications. Genet Med 18, 696-704.
7. Lelieveld, S.H., Reijnders, M.R., Pfundt, R., Yntema, H.G., Kamsteeg, E.J., de Vries, P., de Vries, B.B., Willemsen, M.H., Kleefstra, T., Lohner, K., et al. (2016). Meta-analysis of 2,104 trios provides support for 10 new genes for intellectual disability. Nat Neurosci 19, 1194-1196.
8. de Ligt, J., Willemsen, M.H., van Bon, B.W., Kleefstra, T., Yntema, H.G., Kroes, T., Vulto-van Silfhout, A.T., Koolen, D.A., de Vries, P., Gilissen, C., et al. (2012). Diagnostic

exome sequencing in persons with severe intellectual disability. N Engl J Med 367, 1921-1929.
9. DDD-study. (2015). Large-scale discovery of novel genetic causes of developmental disorders. Nature 519, 223-228.
10. Gueneau, L., Fish, R.J., Shamseldin, H.E., Voisin, N., Tran Mau-Them, F., Preiksaitiene, E., Monroe, G.R., Lai, A., Putoux, A., Allias, F., et al. (2018). KIAA1109 Variants Are Associated with a Severe Disorder of Brain Development and Arthrogryposis. Am J Hum Genet 102, 116-132.
11. Brunet, T., Radivojkov-Blagojevic, M., Lichtner, P., Kraus, V., Meitinger, T., and Wagner, M. (2020). Biallelic loss-of-function variants in RBL2 in siblings with a neurodevelopmental disorder. Ann Clin Transl Neurol 7, 390-396.
12. Yang, Y., Muzny, D.M., Xia, F., Niu, Z., Person, R., Ding, Y., Ward, P., Braxton, A., Wang, M., Buhay, C., et al. (2014). Molecular findings among patients referred for clinical whole-exome sequencing. Jama 312, 1870-1879.
13. Rentzsch, P., Witten, D., Cooper, G.M., Shendure, J., and Kircher, M. (2019). CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res 47, D886-d894.
14. Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., McRae, J.F., Darbandi, S.F., Knowles, D., Li, Y.I., Kosmicki, J.A., Arbelaez, J., Cui, W., Schwartz, G.B., et al. (2019). Predicting Splicing from Primary Sequence with Deep Learning. Cell 176, 535-548.e524.
15. Lindeboom, R.G.H., Vermeulen, M., Lehner, B., and Supek, F. (2019). The impact of nonsense-mediated mRNA decay on genetic disease, gene editing and cancer immunotherapy. Nat Genet 51, 1645-1651.
16. Wiel, L., Baakman, C., Gilissen, D., Veltman, J.A., Vriend, G., and Gilissen, C. (2019). MetaDome: Pathogenicity analysis of genetic variants through aggregation of homologous human protein domains. Hum Mutat 40, 1030-1038.
17. Boyd, A., Golding, J., Macleod, J., Lawlor, D.A., Fraser, A., Henderson, J., Molloy, L., Ness, A., Ring, S., and Davey Smith, G. (2013). Cohort Profile: the 'children of the 90s'--the index offspring of the Avon Longitudinal Study of Parents and Children. Int J Epidemiol 42, 111-127.
18. Fraser, A., Macdonald-Wallis, C., Tilling, K., Boyd, A., Golding, J., Davey Smith, G., Henderson, J., Macleod, J., Molloy, L., Ness, A., et al. (2013). Cohort Profile: the Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. Int J Epidemiol 42, 97-110.
19. Walter, K., Min, J.L., Huang, J., Crooks, L., Memari, Y., McCarthy, S., Perry, J.R., Xu, C., Futema, M., Lawson, D., et al. (2015). The UK10K project identifies rare variants in health and disease. Nature 526, 82-90.
20. Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38, e164.
21. Köhler, S., Carmody, L., Vasilevsky, N., Jacobsen, J.O.B., Danis, D., Gourdine, J.P., Gargano, M., Harris, N.L., Matentzoglu, N., McMurry, J.A., et al. (2019). Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. Nucleic Acids Res 47, D1018-d1027.
22. Wang, J.Z., Du, Z., Payattakool, R., Yu, P.S., and Chen, C.F. (2007). A new method to measure the semantic similarity of GO terms. Bioinformatics 23, 1274-1281.
23. Deng, Y., Gao, L., Wang, B., and Guo, X. (2015). HPOSim: an R package for phenotypic similarity measure and enrichment analysis based on the human phenotype ontology. PLoS One 10, e0115692.
24. Kaufman L., R.P.J. (1987). Clustering by means of medoids https://wis.kuleuven.be/stat/robust/papers/publications-1987/kaufmanrousseeuw-clusteringbymedoids-l1norm-1987.pdf.
25. Reijnders, M.R.F., Miller, K.A., Alvi, M., Goos, J.A.C., Lees, M.M., de Burca, A., Henderson, A., Kraus, A., Mikat, B., de Vries, B.B.A., et al. (2018). De Novo and Inherited Loss-of-Function Variants in TLK2: Clinical and Genotype-Phenotype

Evaluation of a Distinct Neurodevelopmental Disorder. Am J Hum Genet 102, 1195-1203.
26. Yamasaki, K., Akiba, T., Yamasaki, T., and Harata, K. (2007). Structural basis for recognition of the matrix attachment region of DNA by transcription factor SATB1. Nucleic Acids Res 35, 5073-5084.
27. Guex, N., and Peitsch, M.C. (1997). SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis 18, 2714-2723.
28. Inoue, K., Hayashi, F., Yokoyama, S., RIKEN Structural Genomics/Proteomics Initiative (RSGI). (2005). Solution structure of the second CUT domain of human SATB2. In. (
29. Izumi, K., Yoshida, M., Hayashi, F., Hatta, R., Yokoyama, S., RIKEN Structural Genomics/Proteomics Initiative (RSGI). (2004). Solution structure of the homeodomain of KIAA1034 protein https://www.rcsb.org/structure/1wi3.
30. Iyaguchi, D., Yao, M., Watanabe, N., Nishihira, J., and Tanaka, I. (2007). DNA recognition mechanism of the ONECUT homeodomain of transcription factor HNF-6. Structure 15, 75-83.
31. Lelieveld, S.H., Wiel, L., Venselaar, H., Pfundt, R., Vriend, G., Veltman, J.A., Brunner, H.G., Vissers, L., and Gilissen, C. (2017). Spatial Clustering of de Novo Missense Mutations Identifies Candidate Neurodevelopmental Disorder-Associated Genes. Am J Hum Genet 101, 478-484.
32. Estruch, S.B., Graham, S.A., Quevedo, M., Vino, A., Dekkers, D.H.W., Deriziotis, P., Sollis, E., Demmers, J., Poot, R.A., and Fisher, S.E. (2018). Proteomic analysis of FOXP proteins reveals interactions between cortical transcription factors associated with neurodevelopmental disorders. Hum Mol Genet 27, 1212-1227.
33. Estruch, S.B., Graham, S.A., Deriziotis, P., and Fisher, S.E. (2016). The language-related transcription factor FOXP2 is post-translationally modified with small ubiquitin-like modifiers. Sci Rep 6, 20911.
34. Deriziotis, P., Graham, S.A., Estruch, S.B., and Fisher, S.E. (2014). Investigating protein-protein interactions in live cells using bioluminescence resonance energy transfer. J Vis Exp.
35. Koulouras, G., Panagopoulos, A., Rapsomaniki, M.A., Giakoumakis, N.N., Taraviras, S., and Lygerou, Z. (2018). EasyFRAP-web: a web-based tool for the analysis of fluorescence recovery after photobleaching data. Nucleic Acids Res 46, W467-w472.
36. Kohze, R., Dieteren, C.E.J., Koopman, W.J.H., Brock, R., and Schmidt, S. (2017). Frapbot: An open-source application for FRAP data. Cytometry A 91, 810-814.
37. Siebenlist, U., Durand, D.B., Bressler, P., Holbrook, N.J., Norris, C.A., Kamoun, M., Kant, J.A., and Crabtree, G.R. (1986). Promoter region of interleukin-2 gene undergoes chromatin structure changes and confers inducibility on chloramphenicol acetyltransferase gene during activation of T cells. Mol Cell Biol 6, 3042-3049.
38. Kumar, P.P., Purbey, P.K., Ravi, D.S., Mitra, D., and Galande, S. (2005). Displacement of SATB1-bound histone deacetylase 1 corepressor by the human immunodeficiency virus type 1 transactivator induces expression of interleukin-2 and its receptor in T cells. Mol Cell Biol 25, 1620-1633.
39. Pavan Kumar, P., Purbey, P.K., Sinha, C.K., Notani, D., Limaye, A., Jayani, R.S., and Galande, S. (2006). Phosphorylation of SATB1, a global gene regulator, acts as a molecular switch regulating its transcriptional activity in vivo. Mol Cell 22, 231-243.
40. Jakobs, A., Koehnke, J., Himstedt, F., Funk, M., Korn, B., Gaestel, M., and Niedenthal, R. (2007). Ubc9 fusion-directed SUMOylation (UFDS): a method to analyze function of protein SUMOylation. Nat Methods 4, 245-250.

3D protein modeling

Method for modeling CUTL variants

PDB entry 4Q2J[1] was used to contextualize the p.P181L variant. PDB entry 2O49[2] was superposed onto PDB entry 4Q2J using Swiss-PdbViewer[3] to highlight the relative orientation of DNA with respect to the SATB1 CUTL domain.

Method for modeling CUT1 variants

The crystal structure of the N-terminal CUT Domain of SATB1 Bound to Matrix Attachment Region DNA (PDB entry 2O4A[2]), and the ONECUT homeodomain of transcription factor HNF-6[4] were used to contextualize the various mutations with respect to DNA, using Swiss-PdbViewer[3].

Method for modeling CUT2 variants

The first NMR model of the PDB entry 2CSF [DOI:10.2210/pdb2CSF/pdb] was used as a template to align residues T491 to H577 of the SATB1 human protein (uniprot entry Q01826), and build a model using Swiss-PdbViewer[3]. The resulting model has been superposed onto the CUT1 domain of pdb entry 2O4A[2] using the “magic fit” option of Swiss-PdbViewer to highlight the position of the variants with respect to DNA.

Method for modeling homeobox domain variants

The Solution structure of the homeodomain of human SATB2 (second NMR model of the PDB entry 1WI3 [DOI:10.2210/pdb1wi3/pdb]) was used as a template to align residues P647 to G704 of the SATB1 human protein (uniprot entry Q01826), and build a model using Swiss-PdbViewer[3]. Chains A, C and D of the crystal structure of HNF-6alpha DNA-binding domain in complex with the TTR promoter (PDB entry 2D5V[4]), which has a DNA binding domain similar to the CUT2 domain of SATB1 and a second DNA binding domain similar to the homeobox of SATB1, was used as a template to superpose the model of the SATB1 homeobox domain onto the HNF-6alpha structure using the “magic fit” option of Swiss-PdbViewer.

Modeling

p.P181L

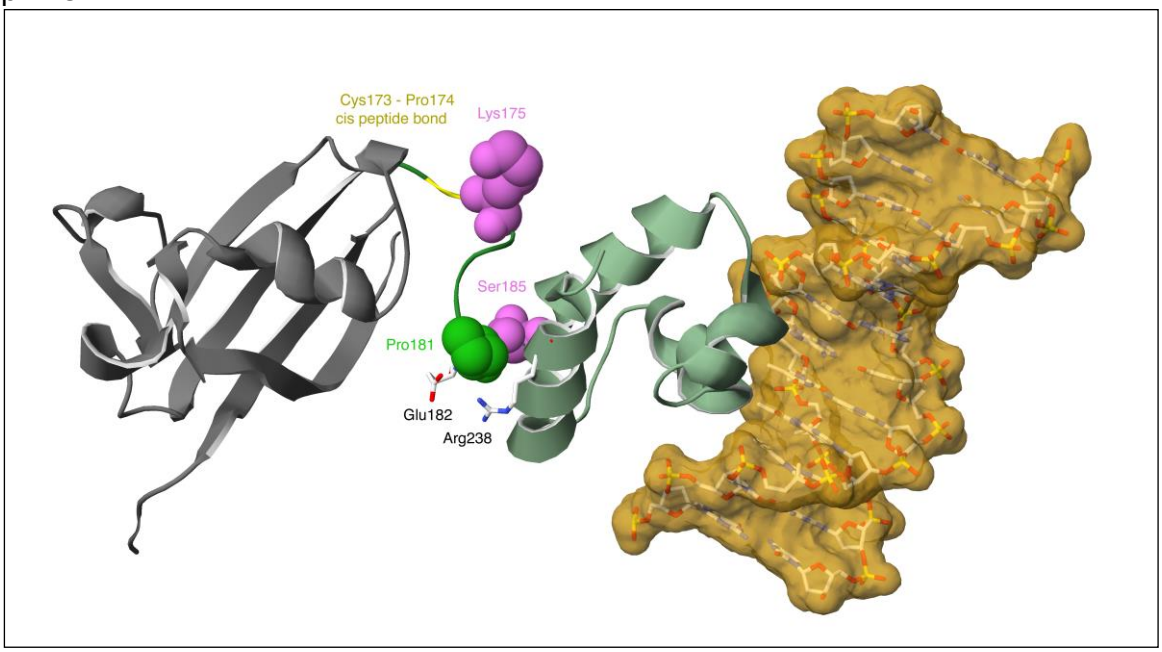


Figure 1. Highlight of the P181 position (green spacefill) with respect to the ubiquitin-like domain (ULD; grey) and the CUT repeat-like (CUTL) domain (dim green). The position of the C173-P174 cis peptide bond is highlighted in yellow. K175 and S185 which can be respectively acetylated and phosphorylated are shown in pink spacefill (top and bottom, respectively).

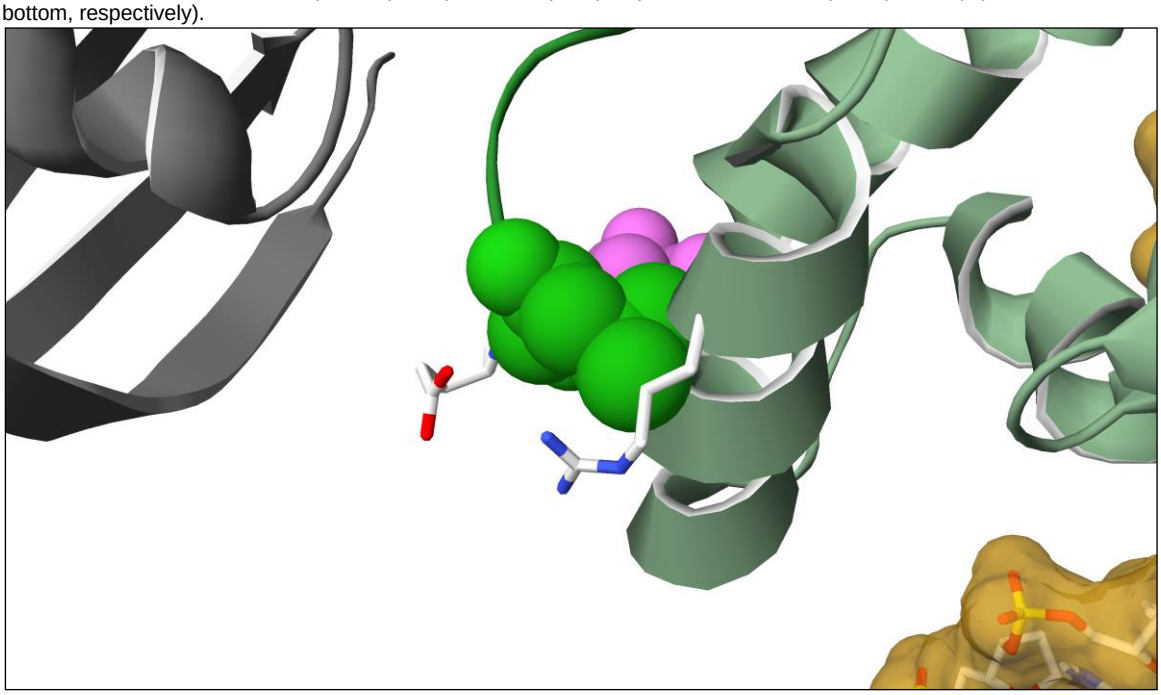


Figure 2. P181L sidechain (green spacefill) clashes into an alpha-helix (A230-K241) of the CUTL domain (dim green), in particular the backbone of residues G237 and R238, as well as in the sidechain of the latter.

The variant P181L variant sits in a linker region between the ubiquitin-like domain (ULD; grey) and a CUT repeat-like (CUTL) domain (dim green). P181 is preceded by another proline, which confers some rigidity and restricts the range of possible relative orientation of the CUTL domain with respect to the UBL domain. There is a third proline in the linker (Pro174), which is preceded by Cys173 and makes a cis peptide bond (highlighted in yellow in Figure 1). Cis-peptide bonds are quite rare (about 0.3% of peptide bonds, although they occur in about 6% of residues followed

by a Proline[5], which shows the importance of the conformation of the linker region. Furthermore, Lys175 and Ser185 (in pink) can be respectively acetylated and phosphorylated and influence the DNA binding capability of SATB1[1]. Sidechains of Glu 182 (from the linker bottom left) and Arg 238 (from the CUTL domain bottom right), positioned just below Pro181 further lock the linker region and the CUTL domain through electrostatic interaction. The relative orientation of these domains cannot be maintained with the P181L mutation, because a leucine sidechain at this position would severely clash into the CUTL domain (backbone of residues Gly237 and Arg238), forcing the linker to adopt a different conformation (Figure 2), which may also potentially affect the ability of K175 to be acetylated.

p.Q402R

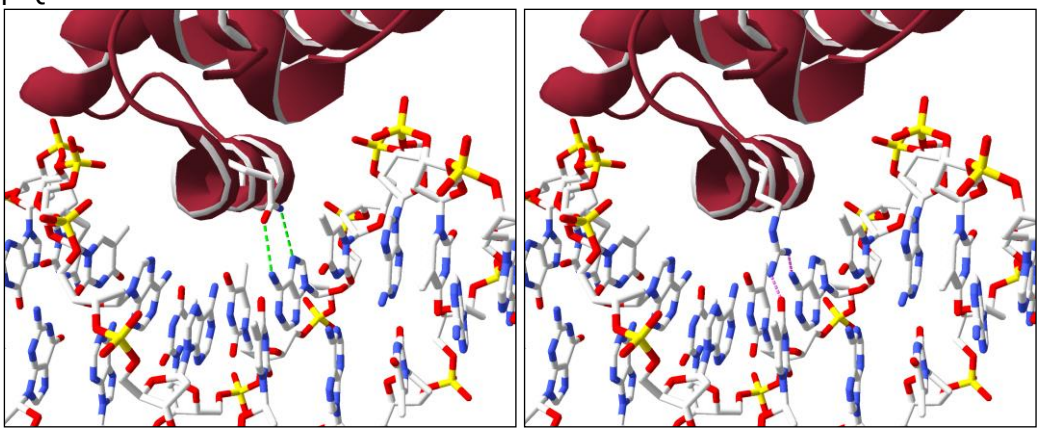


Figure 3. Closeup of the Q402 – DNA interaction (pdb structure 2O4A) highlighting the native residue (Gln, left panel) which makes nice hydrogen bonds to the base (green dotted lines), whereas the longer Arg sidechain (right panel) might collide into the DNA (purple dotted lines) and be forced to adopt a conformation less favorable with respect to binding its cognate DNA.

Q402 is located in the CUT1 domain alpha-helix that binds the major groove of the DNA and is the equivalent of CUT2 domain Q525. Since its sidechain makes direct contact with a nucleotide, a mutation to an arginine, which has a longer sidechain, would need to adopt a conformation less favorable to DNA binding to avoid colliding into the DNA, hence affecting the DNA binding affinity at the cognate sites (Figure 3).

p.E407G

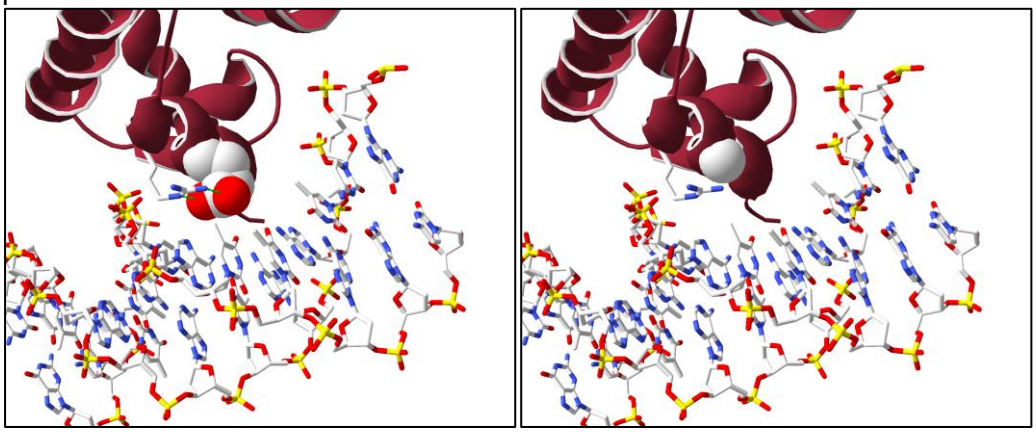


Figure 4. Closeup of the E407 - DNA binding interaction (pdb structure 2O4A) highlighting the native residue (Glu, spacefilled, left panel), which locks in place the sidechain of Arg410 through hydrogen bonds (green dotted lines) and the hole left by the mutation (Gly, spacefilled right panel).

E407 is located in the middle of the CUT1 domain alpha-helix that binds the major groove of the DNA and is the equivalent of CUT2 domain E530. Since its sidechain help maintain the sidechain of Arg410 in place via hydrogen bonds and that both residue make direct contact with the nucleotides, a mutation to a glycine, which bears no sidechain and is not favored in alpha-helices will likely disrupt the local conformation and alter the DNA binding affinity at the cognate sites (Figure 4).

p.E413K

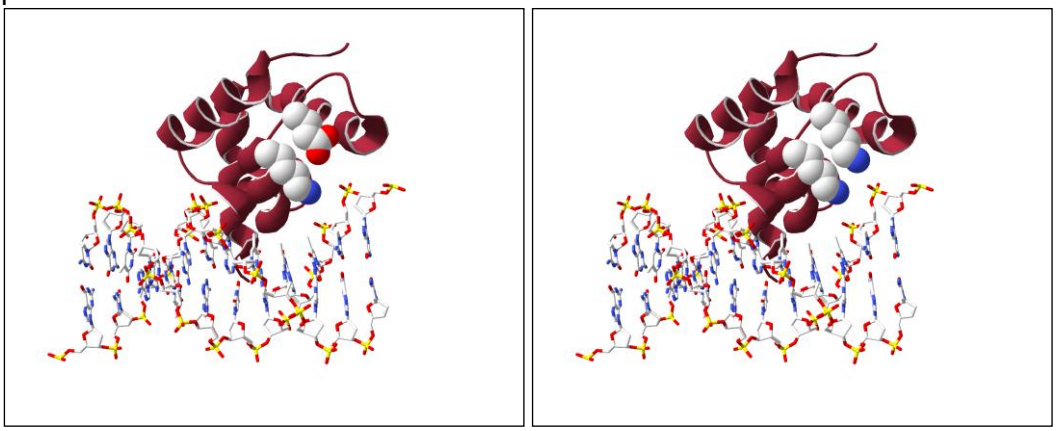


Figure 5. Closeup of E413, solvent exposed in a loop, along Lys411 (left panel). E413 does not make direct DNA contact, and there is enough space to accommodate the E413K mutation (right panel).

E413 is located in a loop right after the end of the CUT1 domain alpha-helix that binds the major groove of the DNA. Although it does not directly bind to DNA, it is in relatively close proximity (within 10 angstroms) to the negatively charged DNA backbone, and in an extended conformation along Lys411. The mutation E413K would replace a negatively charged residue by a positively charged one and may potentially affect the DNA binding affinity of the CUT1 domain through long range electrostatic interactions (Figure 5).

p.Q420R

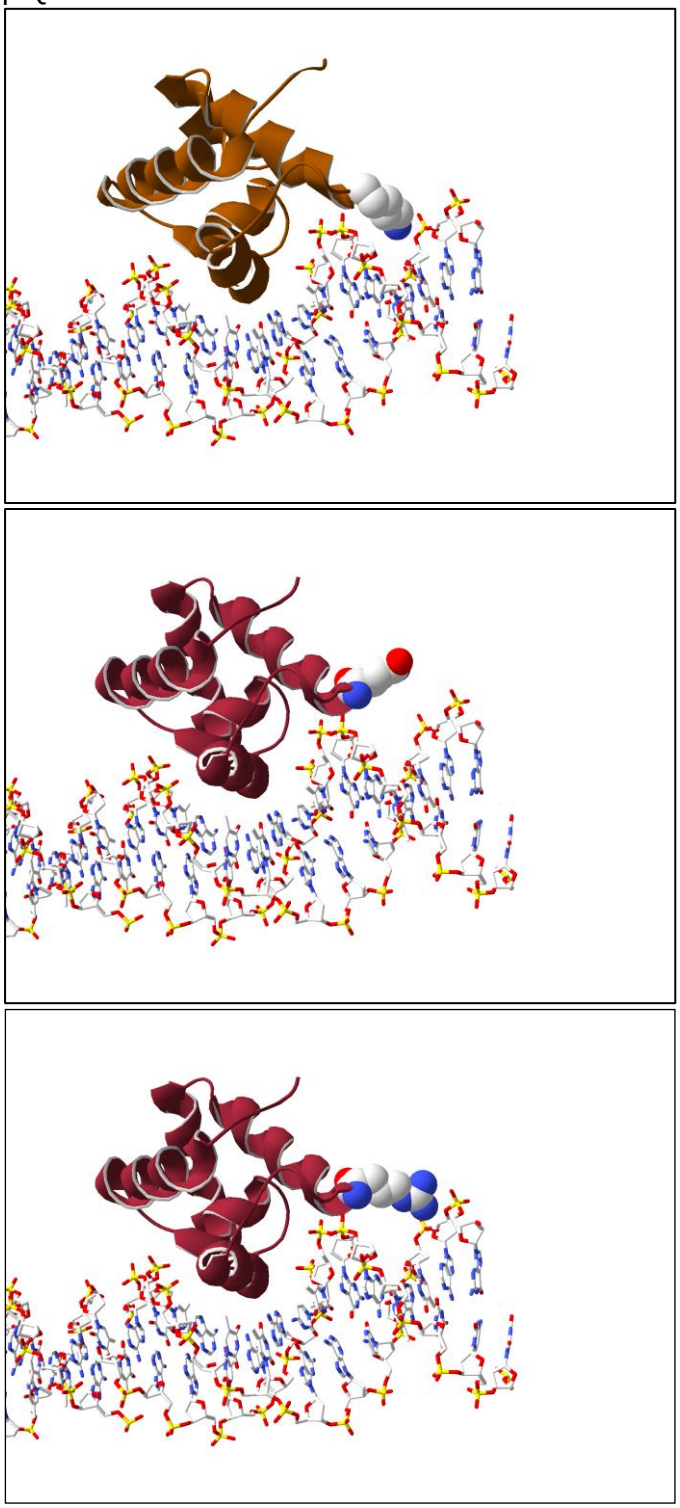


Figure 6. Highlight of the Q420R mutation after superposition of the SATB1 CUT1 domain (pdb entry 2O4A) onto the HNF6alpha DNA binding domain bound to DNA (pdb entry 2D5V) showing its close proximity to DNA backbone. Top: HNFa, middle: SATB1 WT, bottom: SATB1 mutant.

Q420 is located at the surface of the CUT1 domain, not in direct contact with DNA. An arginine at this position could easily be accommodated, but since it is bulkier and positively charged, it may affect the binding of CUT1 to other domains. Of note, the superposition of the CUT1 domain onto the DNA binding domain of rat HNF6 alpha bound to the TTR promoter (pdb entry 2D5V, chain A) reveals that Q420R would be

roughly in the same position as HNF6alpha K53, which points in the minor groove of the DNA and makes indirect contact to the DNA backbone via structural water molecules (Figure 6). This mutation may likely affect the overall affinity of the structural complex.

p.Q525R

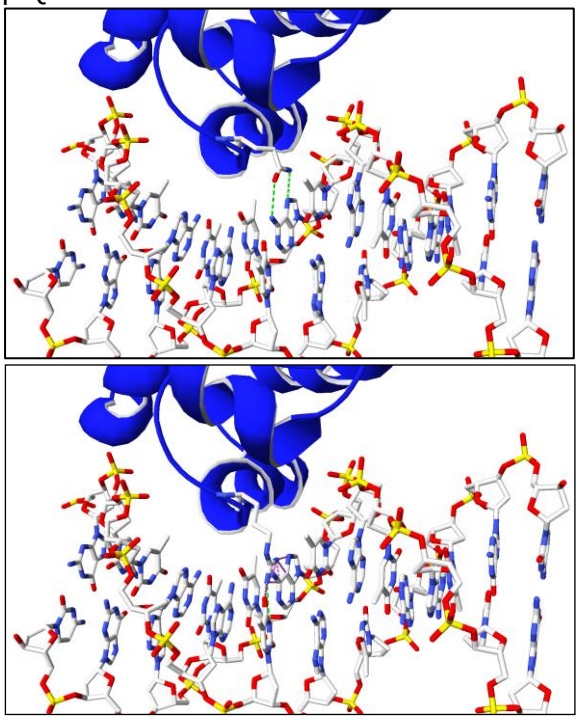


Figure 7. Closeup of the Q525 – DNA interaction highlighting the native residue (Gln, left panel) which could make hydrogen bonds to the base (green dotted lines), whereas the longer Arg sidechain (right panel) might collide into the DNA (purple dotted lines) and be forced to adopt a conformation less favorable with respect to binding its cognate DNA.

Q525 is located in the CUT2 domain alpha-helix that binds the major groove of the DNA, and is the equivalent of CUT1 domain Q402. Since its sidechain makes direct contact with a nucleotide, a mutation to an arginine, which has a longer sidechain, would need to adopt a conformation less favorable to DNA binding to avoid colliding into the DNA, hence affecting the DNA binding affinity at the cognate sites (Figure 7).

p.E530G

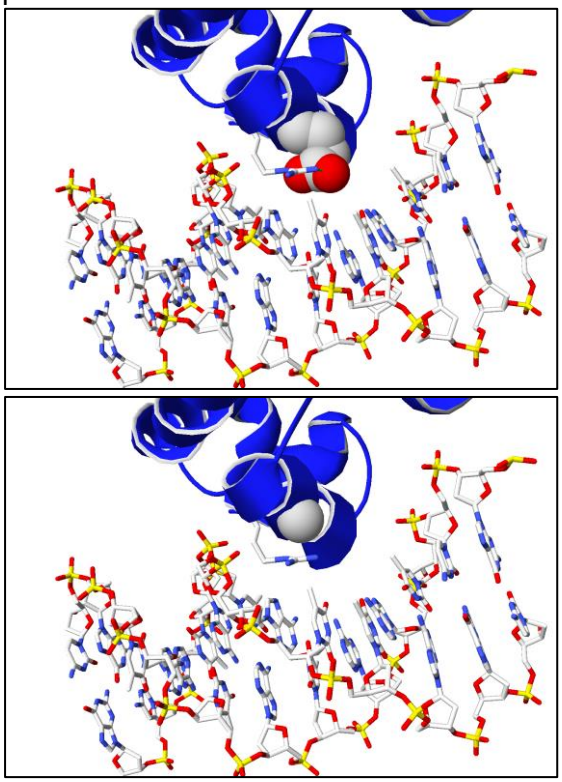


Figure 8. Closeup of the E530 - DNA binding interaction (pdb structure 2O4A) highlighting the native residue (Glu, spacefilled, left panel), which locks in place the sidechain of Arg533 through hydrogen bonds (green dotted lines) and the hole left by the mutation (Gly, spacefilled right panel).

E530 is located in the middle of the CUT2 domain alpha-helix that binds the major groove of the DNA and is the equivalent of CUT1 domain E407. Since its sidechain help maintain the sidechain of Arg533 in place via hydrogen bonds and that both residues make direct contact with the nucleotides, a mutation to a glycine, which bears no sidechain and is not favored in alpha-helices will likely disrupt the local conformation and alter the DNA binding affinity at the cognate sites (Figure 8).

p.E530K

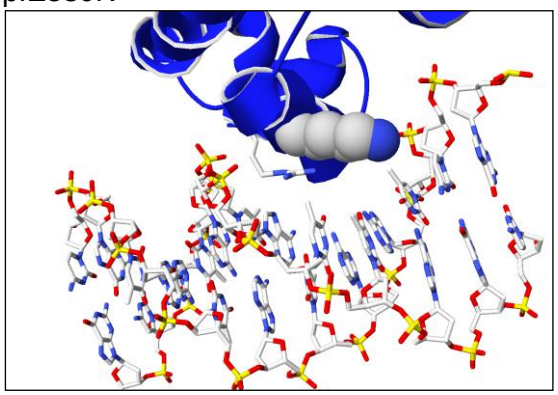


Figure 9. Closeup of the E530 – a conformation that could be adopted by a lysine at this position.

E530 is located in the middle of the CUT2 domain alpha-helix that binds the major groove of the DNA and is the equivalent of CUT1 domain E407. Since its sidechain help maintain the sidechain of Arg533 in place via hydrogen bonds and that both residues make direct contact with the nucleotides. A mutation to a Lysine, which is very flexible and can be accommodated from a steric point of view will likely induce a

rearrangement of these two positively charged sidechains, both in close proximity to DNA bases, and result in a change of affinity at the cognate sites (Figure 9).

p.E530Q

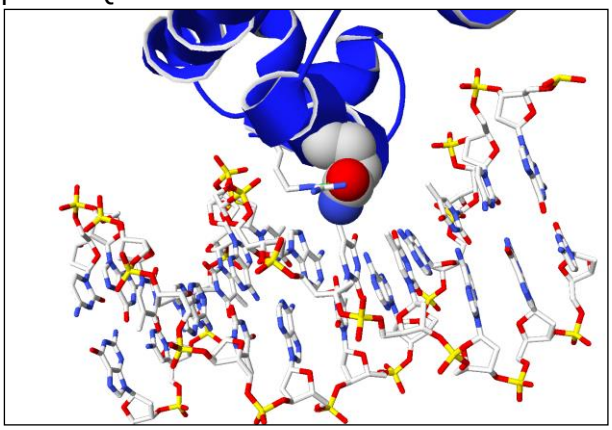


Figure 10. Closeup of the E530 – a conformation that could be adopted by a glutamine at this position.

E530 is located in the middle of the CUT2 domain alpha-helix that binds the major groove of the DNA and is the equivalent of CUT1 domain E407. Since its sidechain help maintain the sidechain of Arg533 in place via hydrogen bonds and that both residues make direct contact with the nucleotides. A mutation to a Glutamine can probably be accommodated from a steric point of view but will induce a rearrangement of these two residues, both in close proximity to DNA bases, and probably result in a change of affinity at the cognate sites (Figure 10).

p.E547K

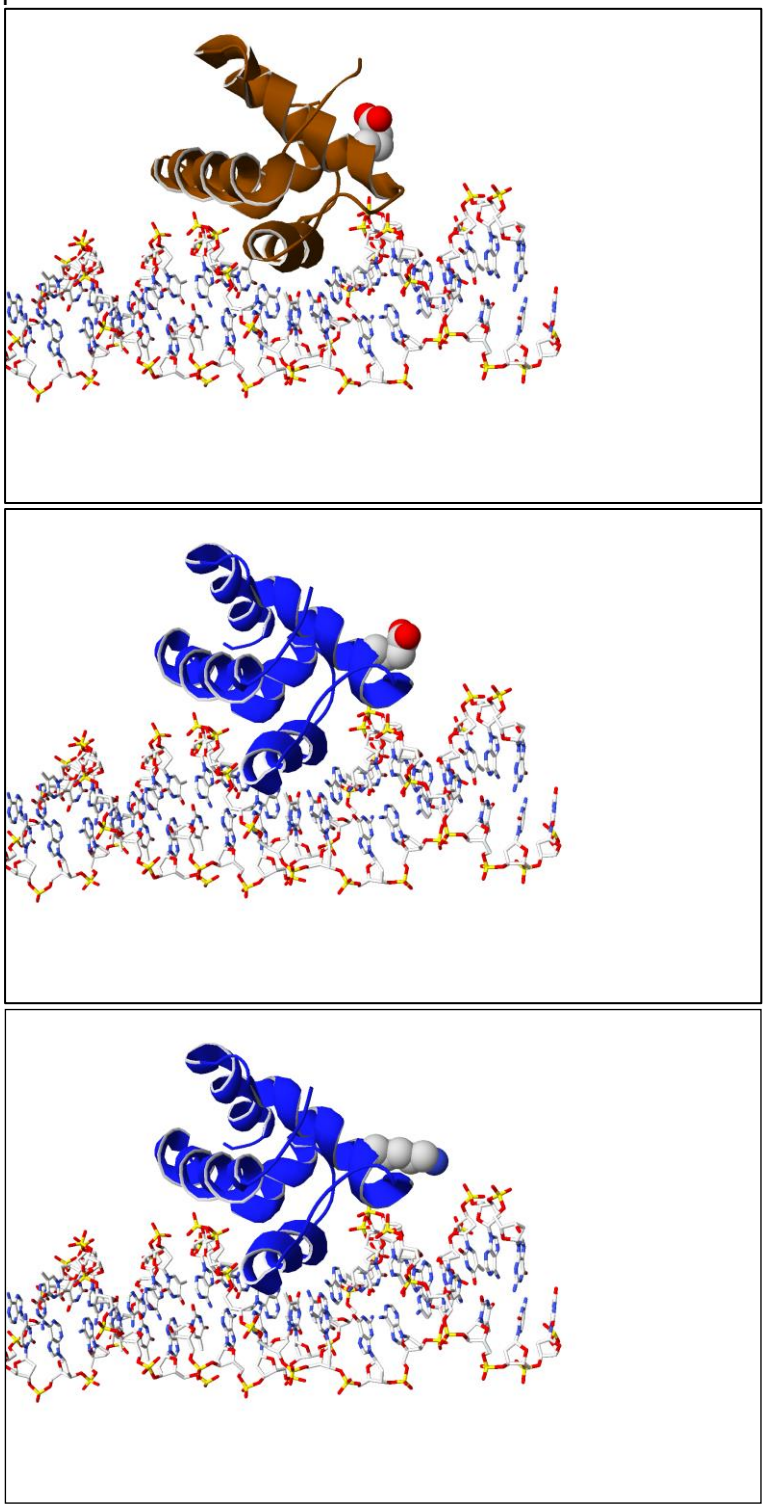


Figure 11. Highlight of the E547K mutation after superposition of the SATB1 CUT2 domain model onto the HNF6alpha DNA binding domain bound to DNA (pdb entry 2D5V) showing its close proximity to DNA backbone. Top: HNFa, middle: SATB1 WT, bottom: SATB1 mutant.

E547 is located at the surface of the CUT2 domain, not in direct contact with DNA. A lysine at this position could easily be accommodated, but since it substitutes a negative charge with a positive one, it may affect the binding of CUT2 to other domains. Of note, the superposition of the CUT2 domain onto the DNA binding domain of rat HNF6 alpha bound to the TTR promoter (pdb entry 2D5V, chain A[4])

reveals that E547K would be roughly in the same position as HNF6alpha E57, which is solvent exposed. Interestingly, it is also in a position close to the CUT1 domain variant Q420R, just one turn of alpha-helix away. This mutation will likely affect the overall binding affinity of other domains to the CUT2 domain.

p.L682V

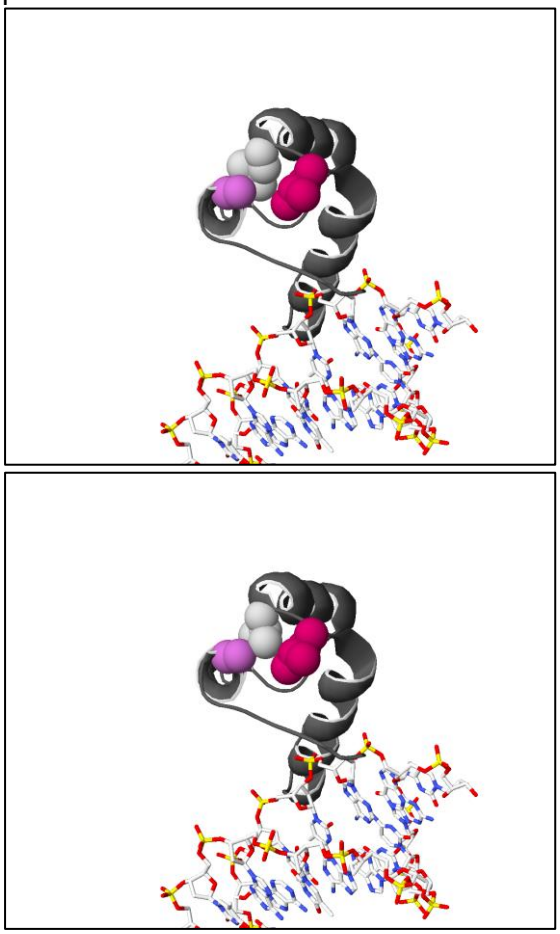


Figure 12. Closeup of the L682V mutation. Left: L682 sidechain (white) is tightly packed with A655 (pink) and L684 (strawberry). Right: V682 sidechain slightly bumps into A655 and L684.

L682 is not proximal to DNA. It is located at the end of the alpha-helix E672-L682, just before a loop, neither of which are either in contact with the DNA. It is buried and probably contributes to maintain the homeobox domain fold. The valine mutant will have a less optimal packing of this region, and its branched sidechain is predicted to moderately clash with Ala 655 and Leu 684 sidechains and is expected to induce a small conformational change in this region. This in turn might subtly affect the binding affinity of other protein domains of the whole complex.

References

1. Wang, Z., Yang, X., Guo, S., Yang, Y., Su, X.C., Shen, Y., and Long, J. (2014). Crystal structure of the ubiquitin-like domain-CUT repeat-like tandem of special AT-rich sequence binding protein 1 (SATB1) reveals a coordinating DNA-binding mechanism. The Journal of biological chemistry 289, 27376-27385.
2. Yamasaki, K., Akiba, T., Yamasaki, T., and Harata, K. (2007). Structural basis for recognition of the matrix attachment region of DNA by transcription factor SATB1. Nucleic acids research 35, 5073-5084.
3. Johansson, M.U., Zoete, V., Michielin, O., and Guex, N. (2012). Defining and searching for structural motifs using DeepView/Swiss-PdbViewer. BMC bioinformatics 13, 173.
4. Iyaguchi, D., Yao, M., Watanabe, N., Nishihira, J., and Tanaka, I. (2007). DNA recognition mechanism of the ONECUT homeodomain of transcription factor HNF-6. Structure (London, England : 1993) 15, 75-83.
5. Weiss, M.S., Jabs, A., and Hilgenfeld, R. (1998). Peptide bonds revisited. Nature Structural Biology 5, 676-676.