The American Journal of Human Genetics, Volume *107*

Supplemental Data

Recessive, Deleterious Variants in *SMG8* Expand the Role of Nonsense-Mediated Decay in Developmental Disorders in Humans

Fatema Alzahrani, Hiroyuki Kuwahara, Yongkang Long, Mohammed Al-Owain, Mohamed Tohary, Moeenaldeen AlSayed, Mohammed Mahnashi, Lana Fathi, Maha Alnemer, Mohamed H. Al-Hamed, Gabrielle Lemire, Kym M. Boycott, Mais Hashem, Wenkai Han, Almundher Al-Maawali, Feisal Al Mahrizi, Khalid Al-Thihli, Xin Gao, and Fowzan S. Alkuraya

Supplemental Clinical Data

Family 1. The index individual (19DG1424) is a 23 months old female who is a product of cesarean section at term due to intrauterine growth restriction (IUGR) and polyhydramnios. She had a short stay in the NICU for low birth weight (2kg) and a heart murmur that was found to be due to a ventricular septal defect (VSD) on echocardiography, which also revealed patent foramen ovale, persistent left superior vena cava and dilated coronary sinus. There is history of global developmental delay, she is unable to sit or roll over and has no clear words. The parents are distant cousins and have a healthy 5-year-old boy. Physical examination revealed poor growth (length -2.7SD, weight -2.8SD) and microcephaly (-4.1SD). She was dysmorphic with hypertelorism, low set posteriorly rotated ears, bulbous nose, long philtrum and micrognathia. She also had facial eczema. Neurological examination revealed axial and appendicular hypotonia. Chest X-ray performed at 5 months of age revealed cardiomegaly with increased pulmonary vascular markings and peribronchial wall thickening, but no pleural effusion or pneumothorax. Chromosomal microarray was normal.

In the course of the study, the mother became pregnant and although initial scan at 18+3wk of gestation was normal her amniocentesis revealed the fetus (MDLREQ2019-6451) was homozygous for the same *SMG8* variant detected in the index individual. A repeat scan at 27+4wk, however, was grossly abnormal: shortened long bones, borderline microcephaly, widened cavum septum pellucidum, prominent 3rd and 4th ventricles, and micrognathia (Figure 1). Fetal echocardiography was suspicious for left pulmonary artery sling vs hypoplasia.

Family 2. The index individual (19DG1391) is a 7-year-old boy, product of consanguineous marriage, born full-term with birth weight of 2500 grams. Pregnancy was complicated by placenta previa necessitating delivery via caesarean section. Apgar scores were 9, 9 and 10 at 0, 1, and 5 minutes, respectively. Developmental delay was evident since infancy in all domains. Later, he displayed autistic behavior, hyperactivity, and was diagnosed with intellectual disability (IQ 66). Physical examination at age 7 years revealed weight and height on the 10th centile and head circumference on the 7th centile (50cm). He was dysmorphic with hypertelorism, bulbous nose, prominent nasolabial folds, thick lips, and large ears (Figure 1). Lab tests showed normal blood indices, normal renal function, normal liver function, normal thyroid function, normal metabolic screen (ammonia, lactate, plasma acylcarnitines, urine organic acids, very long chain fatty acids, and carbohydrate deficient transferrin). Hearing test was normal as was cardiac assessment. Brain MRI revealed a faint high T2 signal of the white matter in the peritrigonal region of posterior periventricular regions with no restricted diffusion.

19DG1396 is the 19-year-old brother of the index individual, a product of a complicated pregnancy with vaginal bleeding and premature labor at 33 weeks gestation. Birth weight was 900 grams and Apgar scores were 5, 6, and 9 at 0, 1, and 10 minutes, respectively. He required active resuscitation and admission to NICU for 6 weeks. He displayed signs of global developmental delay: sat at 11 months, walked at 20 months and had very poor speech development. He was diagnosed with intellectual disability (IQ 45), right-sided mild sensorineural hearing loss and strabismus with refractive error and astigmatism. His behavior is described as autistic. He has history of a ventricular septal defect diagnosed at age 4yrs but has since closed spontaneously. Physical examination showed facial similarities to the index individual. Height and weight were normal, but head circumference was on the 8th centile (53cm). Neurological examination showed spasticity especially in the lower limbs with brisk deep tendon reflexes and toe-walking.

Family 3. The index individual (19DG0152) is an 18-year-old Saudi boy who was referred to us because of microcephaly, dysmorphic facial features and global developmental delay. He was a product of a full-term normal pregnancy and a normal vaginal delivery. He had an uneventful neonatal period, but displayed progressive microcephaly and developmental delay during infancy. He sat at 12m and walked at 24m and still cannot make intelligible single words. His gait is described as wide-based with toe walking. He follows simple commands with some difficulty. There is some impulsivity with self-mutilation noted at hands, but no hyperactivity. He is still not toilet-trained and requires help with self-hygiene and daily activities, although he can feed himself. A bout of urinary retention at age 2 years prompted renal ultrasound, which revealed a dilated right upper renal collecting system, and a trabeculated urinary bladder was noted. His medical history is also notable for hypospadias, xerosis with excoriation-related hyperpigmentation, and cataract operated at age 12yrs although the age of onset is unclear. Physical examination revealed weight of 54.7 kg (-1.59SD), height 166.5 cm (-1.39SD), and head circumference of 51cm (-2.8SD). He appeared microcephalic with a prominent nose, prominent nasolabial folds, large ears and strabismus. Neurological examination revealed generalized hypertonia. Hearing assessment was difficult but there seems to be evidence of hearing loss. The metabolic screen (urine organic acids analysis, plasma acylcarnitines, carbohydrate deficient transferrin, lactate and ammonia) was normal as was FMR1 repeat analysis. Brain MRI was normal.

19DG1342 is the older brother of the index individual, a 28-year old man with unremarkable prenatal, perinatal and postnatal history. He has history of global developmental delay of a very similar degree to the index individual. His medical history is also notable for hypospadias, xerosis with excoriation-related hyperpigmentation, and cataract operated at age 12yrs although the age of onset is unclear. Physical examination revealed normal growth parameters (OFC 54 cm on the 28th centile, weight 79 kg, 73rd centile, height 173 cm, 31st centile). Other salient physical features include toe walking with ataxic gate and skin hyperpigmentation.

19DG0377 is the younger sister of the index individual, a 10yr old girl with intellectual disability and history of global developmental delay. Physical examination revealed microcephaly (OFC 47 cm, -4.1SD) and small body size (weight 24.5 kg, 7th centile, and height 128 cm, 6th centile). Other salient physical findings include a similar gait and skin hyperpigmentation to the other two siblings.

Family 4 (10DF10800): this is an Omani family with two siblings born to consanguineous parents. The index individual (10DF10800-a) was evaluated shortly after birth when found to have bilateral cataract. She was the first child to be born to a 21-year-old primigravida mother following a pregnancy complicated with pregnancy-induced hypertension, oligohydramnios, and intra-uterine growth restriction (IUGR). She was delivered via induced vaginal delivery at 38 weeks of gestation, with a birth weight of 2.3 kg, OFC of 33 cm and length of 48 cm. She required no active resuscitation and had normal APAGR scores. She was noted to have parasternal systolic murmur and echocardiography revealed a small secundum atrial septal defect and a small perimembranous ventricular septal defect with left to right shunting. She had no other clinical concerns immediately after birth. She was admitted for lensectomy and vitrectomy at age 4 weeks. Her weight then was 3.2 kg, OFC was 34 cm, and length was 52 cm. She had normal glucose, and normal liver functions tests. ECG, EEG, and milk scan were all normal. 25-hydroxyvitamin D level was low, and alkaline phosphatase was elevated. She was thought to have nutritional rickets and was thus started on cholecalciferol. At age 6 months, she was admitted with fever and vomiting. Her weight then as 5.7 kg (-2.8 SD) and had evident microcephaly with OFC of 38 cm (-3.6 SD). She was proportionate for growth, and had frontal bossing, with increased furrowing of the forehead more evident upon crying. She also had bitemporal narrowing and low set ears. As she grew older, she evidently had profound global developmental delay, barely being able to sit unsupported and had no words when last seen at the age of 8 years. Her OFC was 49 cm (-2.1 SD), weight was 26 kg (50th percentile) and height of 114 cm (-2.4 SD). Follow up ophthalmology examination revealed pale optic disc with pigmentary retinopathy. Brain stem auditory evoked responses showed evidence of bilateral sensorineural hearing loss. Ammonia and plasma amino acids were normal. Urine organic acids were also normal. TORCH screen was negative. CT brain was normal and MRI brain revealed periventricular white matter changes involving the trigone and frontal lobe with associated mild atrophy of the white matter. Urine was negative for reducing substances while on normal mild and normal infant formula. Galctose-1-phosphate uridyltransferase, galactose metabolites and galactokinase enzyme activities were all normal. Chromosomal microarray did not reveal abnormal copy number variants. Transferrin isoforms were normal by HPLC. Respiratory chain enzymology, pyruvate carboxylase and pyruvate dehydrogenase enzyme activities were all normal in fibroblasts.

While this individual was under follow up, a similarly affected brother (10DF10800-b) was delivered at a different hospital. He was found to have bilateral cataract recognized immediately after birth and had transient neonatal conjugated hyperbilirubinemia with elevated gamma-glutamyltransferase. Ultrasound of the liver showed mildly dilated intrahepatic bile ducts, but no evidence of biliary atresia on HIDA scan. Reportedly there was some pancreatic hypoplasia. Follow up liver function tests were normal with no evidence of liver dysfunction. However, he also experienced similar morbidities to his sister's with congenital cataracts, global developmental delay and microcephaly. He was noted to have genitourinary malformation evident as hypospadias with wide midpenile meatus, and no chordae. He also had mild brain volume reduction and mild periventricular white matter T2 hyper-intensity on brain MRI. Very long chain fatty acids and pristanic acids in plasma were normal. When he was last seen at age 6 years, he was barely able to sit unsupported and he had no identifiable words.

Physical examination at age 6 years showed microcephaly (OFC 48 cm (-2.5 SD)), short stature (104 cm, -2.2 SD) and normal weight (19 kg, 25th centile). He had frontal bossing, and bitemporal narrowing. He had a smooth philtrum with thin upper lip. His teeth were widely spaced, and he had hypodontia with bluish discoloration of the teeth. He had axial hypotonia with increased tone in the lower limbs with pyramidal signs (bilateral hyperreflexia and ankle clonus). Additionally, ophthalmology examination revealed pigmentary retinopathy. Brain stem auditory evoked responses showed evidence of bilateral sensorineural hearing loss.

Supplemental: In silico assessment of SMG8-H208R

MATERIALS AND METHODS

Data preparation

The target protein is human SMG1-SMG8-SMG9 (or SMG1 Complex, SMG1C), a phosphatidylinositol-3-kinase (PI(3)K)-related protein kinase (PIKK) complex central to messenger RNA surveillance. FASTA sequences of human SMG1, SMG8, and SMG9 were retrieved from the UniProt[1] database (UniProt ID: Q96Q15, Q8ND04, Q9H0W8). We obtained the 3.45Å resolution crystal structure of SMG1C (PDB ID: 6L54) directly from a recently published paper about SMG1C cryo-EM structure[2].

Two systems were prepared for structural analysis and molecular dynamics (MD) simulations as follows: the wild type system, and the SMG8-H208R mutants. Since the original structure of SMG1C (6L54) missed too many amino acids, and MD systems usually do not allow losing positions, both the wild type and the mutant were further modeled with SWISS-MODEL[3].

MD simulations

MD simulations were performed on two systems at 300K by the GROMACS 2018[4]. The free protein atoms were soaked in a cubic box of water molecules with a dimension of 10 Å based on the Simple Point Charge (spc216) water model. Charges of the protein were neutralized by adding Na^+ and Cl^- ions. Then the energy of all systems was minimized with 10000 steps of steepest descent. After that, equilibration was conducted in two phases: the first phase is done under the NVT ensemble (constant number of particles, volume, the temperature at 100 ps), temperatures are subsequently raised to 300K during this equilibration. Next, equilibration of pressure was conducted under an NPT ensemble (constant number of practices, pressure, and temperature at 100ps) to stabilize the pressure. After these equilibration phases, the Particle Mesh Ewald method was used for a 1ns MD simulation. The resulting trajectories were analyzed with GROMACS g_energy, g_rms, g_gyrate, g_hbond, g_sas functions, respectively.

Structural analysis

Furthermore, structure analysis studies have been carried out with various methods[5-11]. The stability effect of SMG8-H208R was predicted using advanced computational methods. DynaMut[11] is used to sample protein conformations and assess the impact of mutations on protein dynamics and stability resulting from vibrational entropy changes. SDM[10] works as a computational method that analyzes the variation of amino acids within the family of homologous proteins of known 3D structures. DUET[6], mCSM[7]and mCSM-PPI2[8] are machine learning-based prediction tools designed for stability change prediction. ENCoM[9] calculates the differences in protein stability with a coarse-grained method resulting from changes in vibrational entropy. FoldX[5] predicts the effect of mutations using the biophysical properties of protein molecules.

Residual frustration analysis

The frustratometer[12] server was used to calculate the residual frustration index in the wild type and the mutant structures. It compared the energy distributions of the native state at each residue with respect to a set of structural decoys. A contact was considered as minimally frustrated if the native energy is at the lower end of the distribution, which equals to the Z-score > 0.78. In contrast, it would be defined as highly frustrated if the Z-score < -1. Other values outside the two ranges were considered as neutral[13]. The residual frustration index is quite useful since sites of high local frustration often correlate with functional regions such as binding sites and regions involved in allosteric transitions[14].

RESULTS

Modeling with SWISS-MODEL

We first conducted homology modeling with the template protein 6L54 to fill the missing residues. Fig. S1 shows the filled structure of human SMG1C, with chain A (SMG1) in green, chain B (SMG8) in gray and chain C (SMG9) in pink:

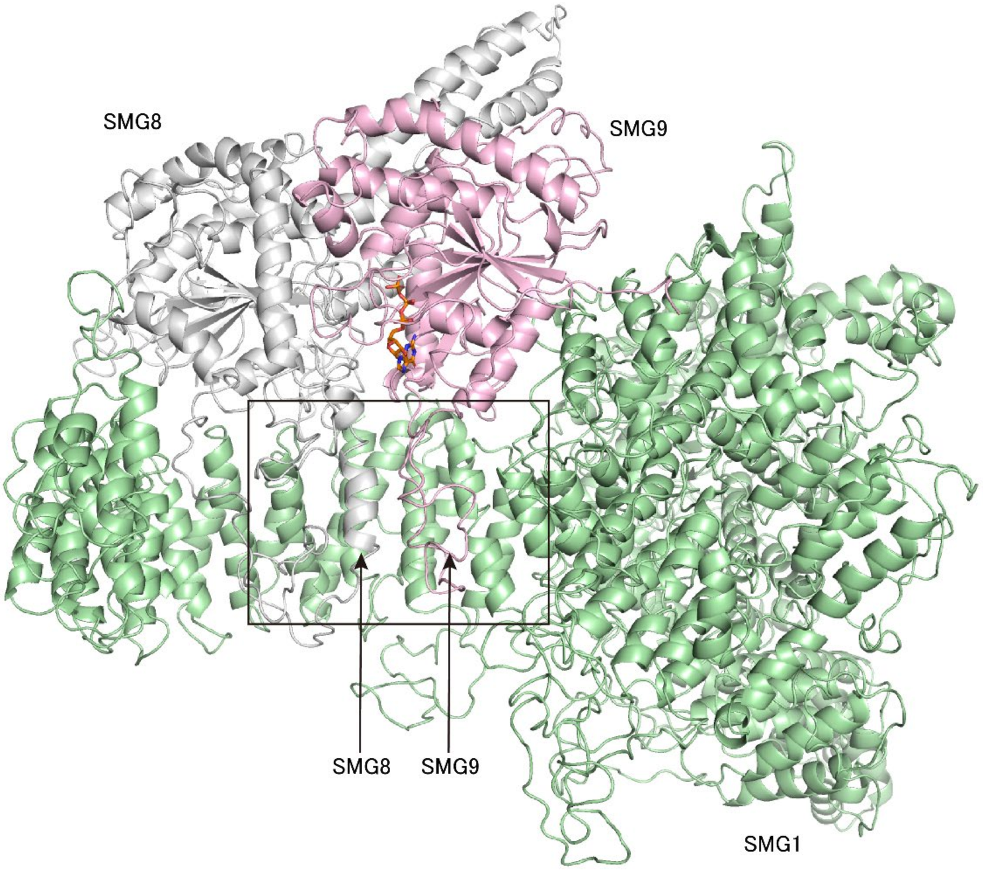


Fig. S1. Ribbon representations of the human SMG1C structural model. SMG1 is colored in green, SMG8 is shown as gray, and SMG9 is represented as pink. The intermolecular interfaces are indicated with a rectangle.

Structural analysis

Protein stability prediction

To quantify the effects of H208R, we employed six different computational methods for predicting the structure stability changes. DynaMut[11], ENCoM[9], mCSM[7], SDM[10], DUET[6], and FoldX[5] provide depth of single-point mutations over protein stability. Here, we define the energy change as the change in the folding free energy of a mutation from a wild-type protein to its mutant ($\Delta\Delta G_{WT\rightarrow MT}$). Thus, $\Delta\Delta G \geq 0$ is defined as stabilizing and $\Delta\Delta G < 0$ as destabilizing.

Furthermore, ENCoM also provides Δ Vibrational entropy energy prediction between the wild type and the mutant, and mCSM-PPI2[8] could predict the impact of mutations on protein interaction binding affinity. We conducted experiments with these tools, and the results are shown in Table S1.

Prediction tools	Task	Predicted ΔΔG	conclusion
Dynamut	Stability	-0.367 kcal/mol	destabilizing
ENCoM	Stability	-0.237 kcal/mol	destabilizing
mCSM	Stability	-1.927 kcal/mol	destabilizing
SDM	Stability	-1.540 kcal/mol	destabilizing
DUET	Stability	-2.009 kcal/mol	destabilizing
FoldX	Stability	-7.051 kcal/mol	destabilizing
ENCoM	Flexibility	0.297 kcal.mol^{-1}.K^{-1}	increasing flexibility
mCSM-PPI2	Binding affinity	-0.326 kcal/mol	decreasing binding affinity

Table S1. Predicted changes in protein stability, flexibility, and protein-protein binding affinity in SMG8-H208R mutant of SMG1C.

Changes in SMG1C stability due to the SMG8-H208R mutation have been captured as ΔΔG in DynaMut, ENCoM, mCSM, SDM, DUET and FoldX. All stability prediction tools suggested that it was a destabilizing effect. Besides, ENCoM and mCSM-PPI2 suggested that the mutation increased molecule flexibility and decreased binding affinity.

Overall, these results clearly suggest that the SMG8-H208R mutation alters the flexibility of the SMG8 molecule, thus decreases its binding affinity with SMG1; as a result, the entire SMG1 complex structure stability decreases. To get a detailed picture of the conclusion, we turned our view into a local region of SMG8, the SMG1-binding region, to see how the mutation influences the binding affinity.

Molecule flexibility prediction

The ENCoM[9] calculates the Δ Vibrational Entropy Energy between the wild type and the mutant, the visual representation is in Fig. S2. Amino acids are colored according to the vibrational entropy change upon mutation. Blue represents a rigidification of the structure, and red shows a gain in flexibility of SMG8. The mutation point is represented as an orange stick. In Fig. S2, we can see that amino acids located at the intermolecular interfaces gain variability after the mutation, which supports the destabilizing prediction results and the decreasing binding affinity result.

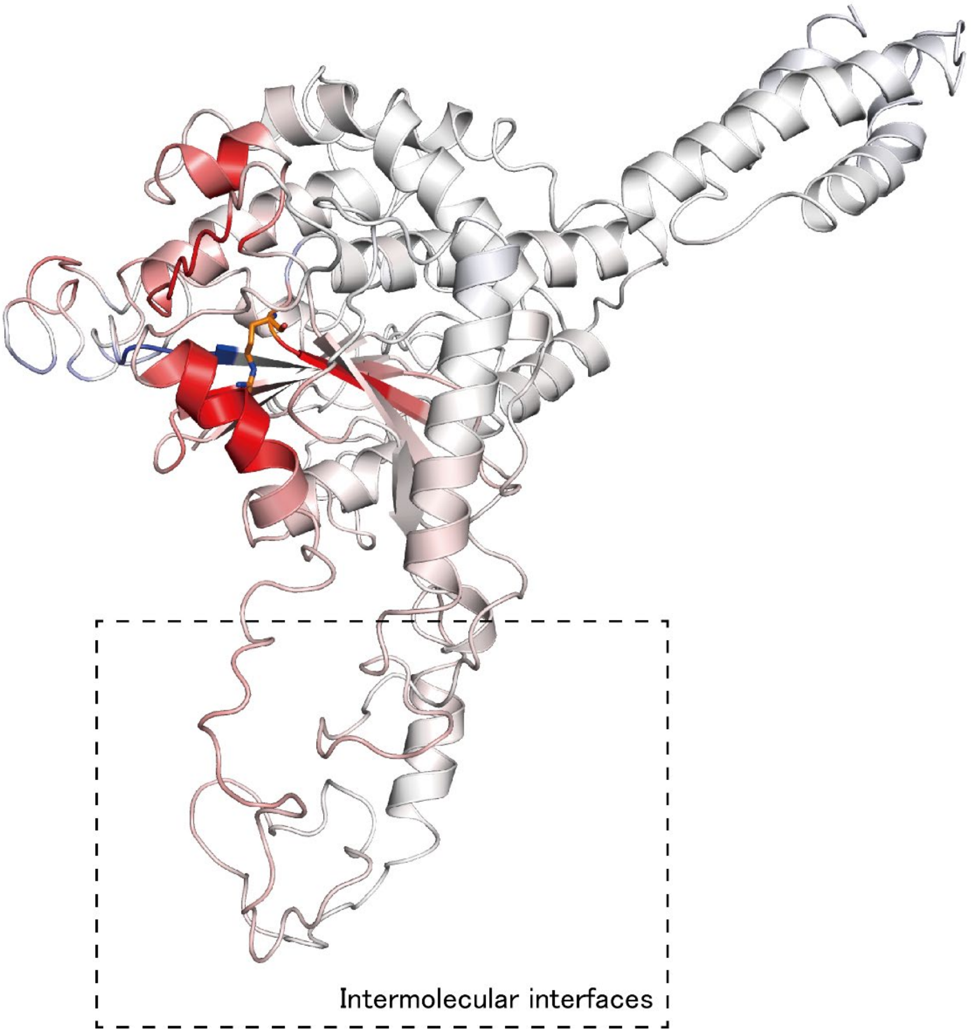


Fig. S2. Effects of SMG8-H208R mutation on SMG8 flexibility. The orange stick represents the mutation point. Red shows an increase in flexibility, while blue indicates a decrease. The SMG1-binding regions of SMG8 are indicated with a dashed rectangle.

Residual frustration analysis

Frustration index shows the relative stability of a given pair of amino acids to all possible contacts at that position. The frustration index of one amino acid is calculated by comparing the native state and a decoy set, which is constructed by randomizing the identity of the particular amino acid while keeping all other interaction parameters and neighboring residues at the native state. A minimally frustrated index means the majority of the other amino acid pairs in that position would be unfavorable, while a highly frustrated contact has an opposite physical meaning.

The comparative analysis between the wild type and the mutant reveals how the mutation of conformational change shifts the frustration state. In the case of the H208R mutant, we analyzed on chain B (SMG8) since the flexibility changes much in the previous study. The frustration indices of some essential amino acids at the intramolecular interfaces[2] are shown in Fig. S3. Our analysis depicts that the H208R mutation largely shifts the frustration pattern at four SMG1-binding residues (Asn61, Glu145, Lys362, Ser400), highlighted as dashed circles in Fig. S3. To avoid the randomness of the modeling methods, we further compared a missense mutation, SMG8-V206I, it's a rare allele that are observed in gnomAD[15], and it may not cause severe dysfunction. Almost all residues located at the intramolecular interfaces of the V206I mutant showed a similar frustration pattern compared with the wild type, which further suggest the functional role of SMG8-H208R mutation.

The shifts in these critical residues of SMG8 might explain the change of flexibility, binding affinity, and the stability of SMG1C. Next, dynamic behavior analysis is essential to get a global look for both the wild type and the mutant.

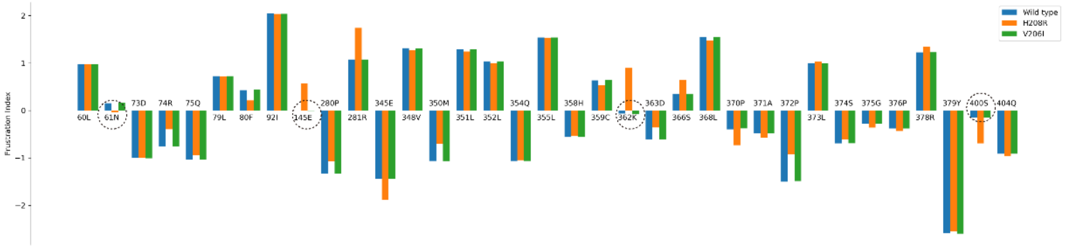


Fig. S3. Frustration indices of SMG1-binding residues on SMG8. Significant changes in frustration are highlighted with a dashed circle.

Molecular dynamics simulation

The decrease in hydrogen bonds for H208R

Hydrogen bonds formed between amino acid residues and water molecules control the protein structure[16]. The hydrogen bonds formed between the protein molecule and water molecules were calculated for SMG1C and its mutants, shown in Fig. S4. The average number of hydrogen bonds of the two systems was 7369 and 7296 for the wild type and the mutant, respectively, along with the simulation trajectories. This suggests that the wild type has a higher number of hydrogen bonds, makes the structure more stable, active, and thus prevents conformational distortion.

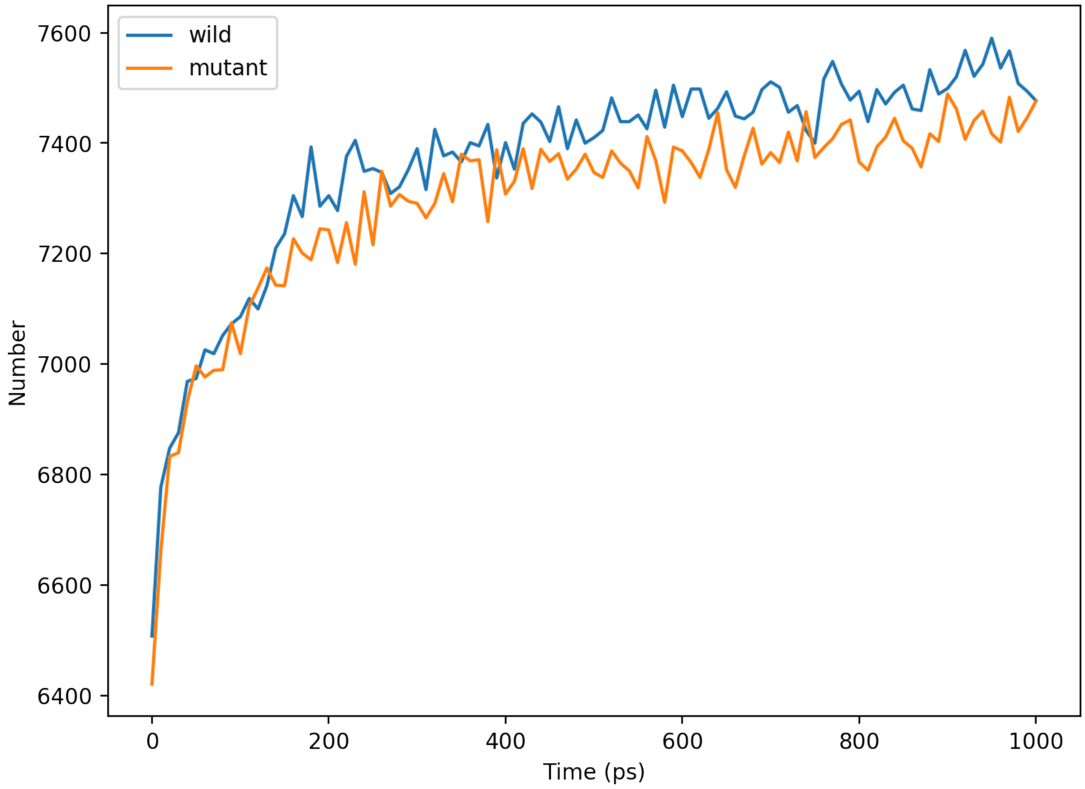


Fig. S4. The average number of hydrogen bonds between SMG1C and water molecules as a function of time.

Compactness influence of H208R

The radius of gyration (Rg) is a parameter linked to the tertiary structural volume of a protein and has been applied to obtain insight into the stability of the protein in a biological system along with the MD simulation (Fig. S5). The average Rg values were found to be 5.16 and 5.19 for the wild type and the mutant, respectively, along with the simulation trajectories. The result suggests that the compactness of SMG1C increases with the introduction of Arginine.

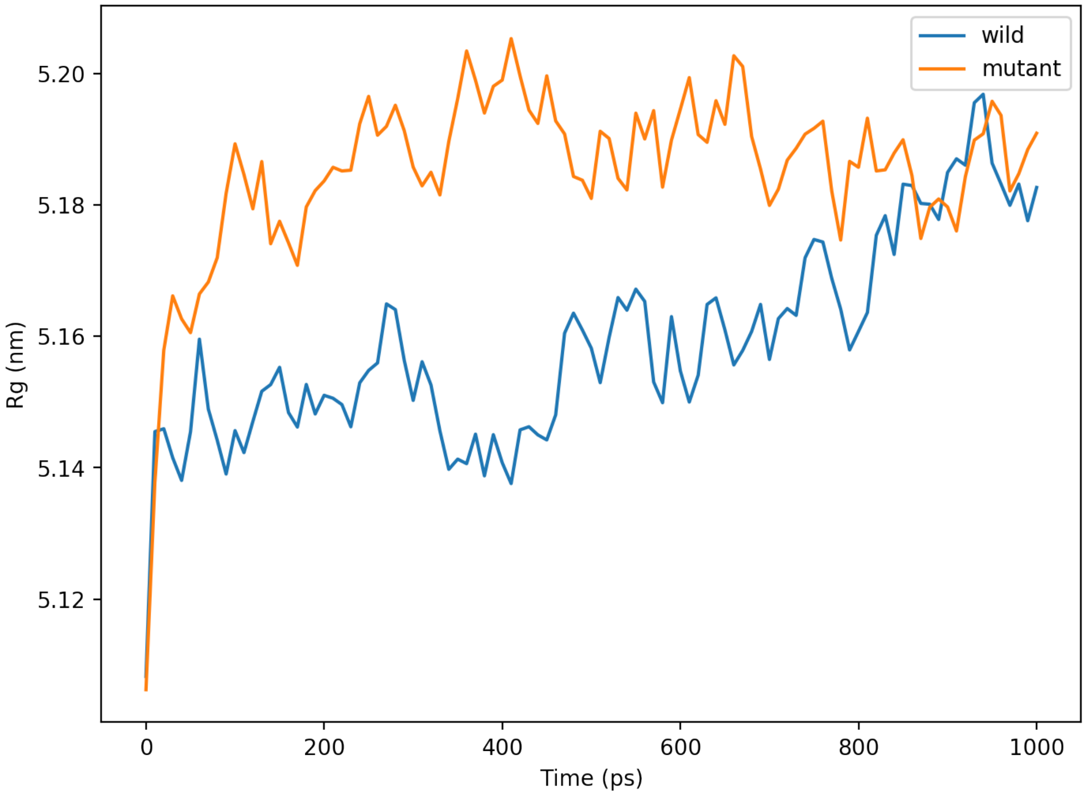


Fig. S5. Plots of the radius of gyration as a function of time.

High structure deviation for H208R

The measure of structural similarity between proteins is a valuable tool for the analysis of protein structures and folding simulations. The average RMSD values were found to be 2.9 Å and 3.4 Å for the wild type and the mutant, respectively, along with the simulation trajectories. During the entire MD simulation process, the mutant H208R shows higher fluctuations compared to the wild type. It suggests that the mutant largely deviates from its native conformation.

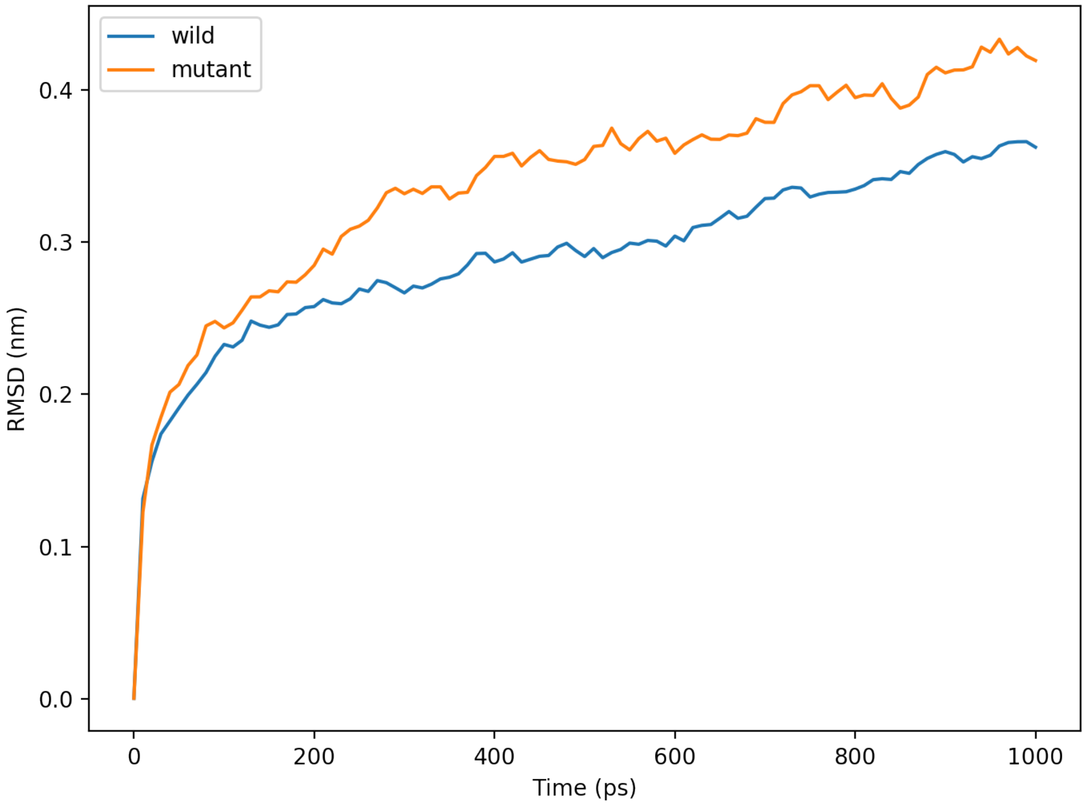


Fig. S6. Plots of RMSD relative to the structure present in the minimized, equilibrated system as a function of time.

Hydrophobic burial

Solvent-accessible surface area (SASA) is defined as the surface area of a protein that interacts with its solvent molecules. The denaturation of the protein causes an increase in SASA. As a result, the hydrophobic regions are gradually exposed to the solvent during the course of unfolding. The calculated average SASA values with respect to backbone were found to be 1934 and 1930 nm^2 for the wild type and the mutant, respectively. According to Fig. S7, the average SASA values of the mutant was lower than that of the wild type. The result can be presumed as the internal residues in the proteins are not exposed after introducing the point mutation.

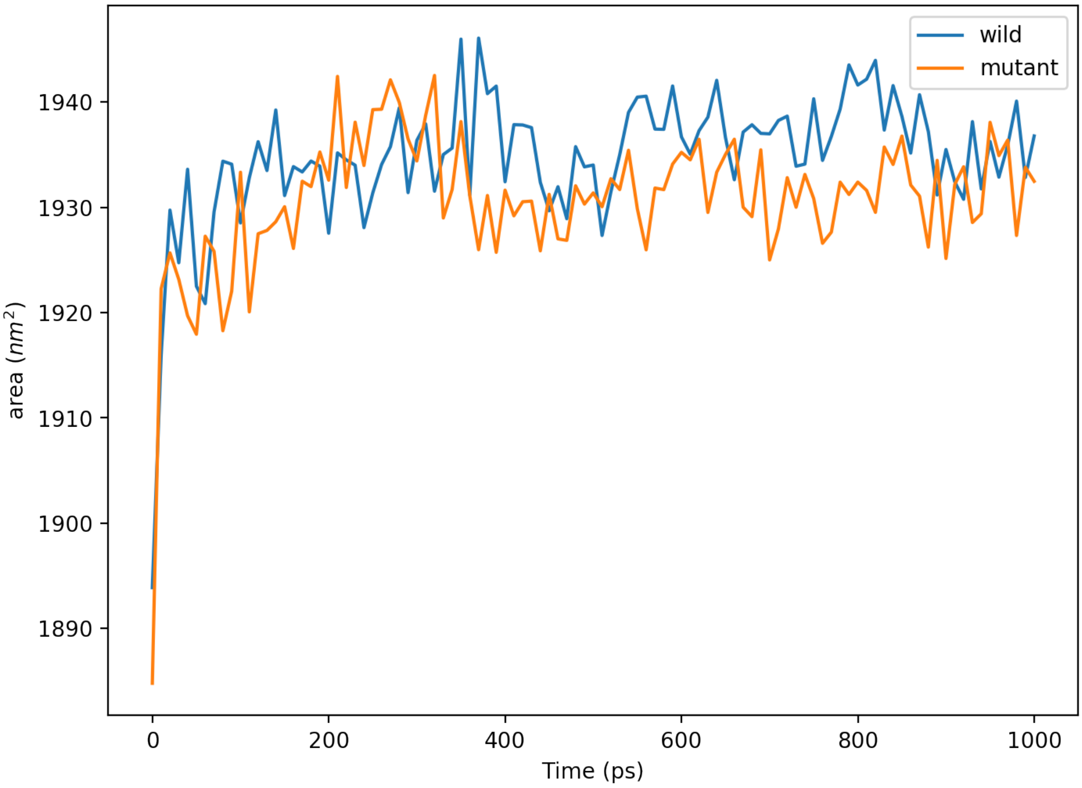


Fig. S7. Plots of SASA as a function of time.

Conclusions

Taken together, our analyses suggest that the SMG8-H208R mutation considerably destabilizes the SMG1 complex. Besides, the mutation prevents the nearby residues from getting exposed to the solvent, hence influences the molecule to interact with the environment. The mutated SMG8 will bind with SMG1 with lower binding affinity, as the changes in residual fluctuation and molecule flexibility are clearly observed at the SMG1-SMG8 interface. The SMG8 and SMG9 activity, in addition to serving as the scaffold for the interaction between SMG1 and UPF1, negatively regulates SMG1 kinase activity and maintains UPF1 in the unphosphorylated form. [17] Therefore, the mutated SMG8 causes a lower binding affinity between SMG8 and SMG1, which, consequently, increases SMG1 phosphorylation of UPF1 and also increases SMG1 kinase activity.

REFERENCES

1 UniProt: the universal protein knowledgebase. *Nucleic acids research* **45**, D158-D169 (2017).

2 Zhu, L., Li, L., Qi, Y., Yu, Z. & Xu, Y. Cryo-EM structure of SMG1–SMG8–SMG9 complex.

Cell Research **29**, 1027-1034, doi:10.1038/s41422-019-0255-3 (2019).

3 Waterhouse, A. *et al.* SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Research* **46**, W296-W303, doi:10.1093/nar/gky427 (2018).

4 Abraham, M. J. *et al.* GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1**, 19-25 (2015).

5 Schymkowitz, J. *et al.* The FoldX web server: an online force field. *Nucleic acids research* **33**, W382-W388 (2005).

6 Pires, D. E., Ascher, D. B. & Blundell, T. L. DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. *Nucleic acids research* **42**, W314-W319 (2014).

7 Pires, D. E., Ascher, D. B. & Blundell, T. L. mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics* **30**, 335-342 (2014).

8 Rodrigues, C. H., Myung, Y., Pires, D. E. & Ascher, D. B. mCSM-PPI2: predicting the effects of mutations on protein–protein interactions. *Nucleic acids research* **47**, W338-W344 (2019).

9 Frappier, V., Chartier, M. & Najmanovich, R. J. ENCoM server: exploring protein conformational space and the effect of mutations on protein function and stability. *Nucleic acids research* **43**, W395-W400 (2015).

10 Worth, C. L., Preissner, R. & Blundell, T. L. SDM—a server for predicting effects of mutations on protein stability and malfunction. *Nucleic acids research* **39**, W215-W222 (2011).

11 Rodrigues, C. H., Pires, D. E. & Ascher, D. B. DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. *Nucleic acids research* **46**, W350-W355 (2018).

12 Parra, R. G. *et al.* Protein Frustratometer 2: a tool to localize energetic frustration in protein molecules, now with electrostatics. *Nucleic acids research* **44**, W356-W360 (2016).

13 Ferreiro, D. U., Hegler, J. A., Komives, E. A. & Wolynes, P. G. Localizing frustration in native proteins and protein assemblies. *Proceedings of the National Academy of Sciences* **104**, 19819-19824 (2007).

14 Wolynes, P. G. Evolution, energy landscapes and the paradoxes of protein folding. *Biochimie* **119**, 218-230 (2015).

15 Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434-443, doi:10.1038/s41586-020-2308-7 (2020).

16 Pace, C. N. *et al.* Contribution of hydrogen bonds to protein stability. *Protein Science* **23**, 652-661 (2014).

17 Yamashita, A. *et al.* SMG-8 and SMG-9, two novel subunits of the SMG-1 complex, regulate remodeling of the mRNA surveillance complex during nonsense-mediated mRNA decay. *Genes & development* **23**, 1091-1105 (2009).

18 Benjamini, Y., & Hochberg, Y.. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, **57**, 289-300 (1995).

Supplemental RNA-seq Analysis

Differential gene expression analysis

We quantified the relative abundance of genes using the gene-level TPM. As described in the main text, we first used Kallisto [Bray et al., *Nat Biotechnol*, 2016] with the reference transcript sequences for hg38 (GENCODE 25) to quantify the abundance of transcripts. Among the annotated genes, we extracted the abundance of transcripts with their functional class being protein-coding (i.e., biotype = 'protein_coding') and considered the abundance of coding transcripts from chromosomes 1-22 and X. To normalize the data with the gene-level TPM, we quantified each coding gene by summing up the length-normalized Kallisto counts of all coding transcripts and scaled the gene-level data to set the total to be one million. To reduce the effects of genes with low abundance levels on our DGE analysis, we added the constant 1 to TPM and used as the normalized count.

Next, we performed sample-level quality control of the normalized count by applying principal components analysis (PCA) and analyzing the distribution of the samples on the PC1 and PC2 coordinates. To perform PCA, we used the prcomp function in R. In the fibroblast dataset, PC1 and PC2 explained 23.09% and 15.35% of the variance, respectively, while in the LCL dataset, they explained 23.24% and 9.83% of the variance, respectively. The visual inspection of these PCA biplots indicated that, while the case samples were clustered together, the control samples were widespread (Supplemental figure S1). This was not surprising since the control samples were from individuals in diverse tribes in Saudi Arabia who were either diagnosed with or suspected of a range of genetic disorders. Thus, although the contribution of the first two PCs were relatively low (in fibroblasts, 38.44%; in LCL, 33.07%), the PCA results provided additional evidence to confirm high levels of heterogeneity among the control samples. To alleviate the high variability in the control group, which could have detrimental effects on the DGE analysis, thus, we filtered out outlier counts from each gene from the normalized count data as mentioned in the main text.

As described in the main text, our DE calling is based on the two criteria: FDR-adjusted p-value < 0.05 and the absolute log fold-changes > 0.5. Prior to our DGE analysis, we computed the z-score of normalized counts for each gene in the control set. We treated samples that have the absolute z-score > 2.0 as outliers and removed them from the control set for each gene. To perform this DGE analysis, we developed custom perl scripts. In the custom scripts, we used the anova method in perl module Statistics::ANOVA with parameters: independent =1; parametric= 1; and ordinal= 0 to perform ANOVA on the datasets. To obtain FDR-adjusted p-values from the ANOVA output, we used perl module Statistics::Multtest. The log fold-change was defined to be the log2-transformed ratio of the mean of the normalized count from the case samples to the mean of the normalized count from the control samples, and it was also implemented in perl.

Quality assessment of the TPM normalization

To assess the quality of the scaling of the TPM normalization on our RNA-seq data, we measured the association between the noise level and the stability level of highly expressed genes in both fibroblasts and LCL datasets. For this analysis, we selected genes with the constraints: mean TPM > 50 and minimum TPM > 0 (LCL, n=2,465; fibroblasts, n=2,596), and we computed the internal control gene-stability measure from geNorm (M) [Vandesompele et al., *Genome Biol*, 2002] of unnormalized counts and the coefficient of variation (CV) of the TPM-normalized counts. The M value measures the average variation of pairwise proportional changes in the gene set. Thus, the lower M indicates higher stability. Because highly stable genes should have lower between-sample variations with the use of a proper scaling factor, an appropriate normalization is expected to reduce CVs for genes with low M values. With this principle, we analyzed the relation between the CV and the M values of the selected genes and found that the two had striking level of rank correlation in both LCL and fibroblasts (LCL, rho = 0.94; fibroblasts, rho = 0.96) (figure S2). This provides evidence that TPM was able to properly normalize our RNA-seq data.

Analysis of the effects of NMD on transcripts with NMD-target PTCs

To analyze the effects of NMD, we set out to compare the abundance of transcripts with NMD-target PTCs (PTC+) and without NMD-target PTCs (PTC-). To this end, we first obtained variant data from whole-exome sequencing for each subject and used VEP [McLaren et al., *Genome Biol*, 2016] to predict the functional effects of each variant. To obtain a list of PTC+, we extracted data for variants whose consequences were predicted to be the emergence of PTCs (i.e., consequence = 'stop_gain'). Next, to filter for transcripts with predicted PTCs, we selected transcripts with PTCs which occur at positions > 50 nt upstream of the last exon-exon splicing junctions for each subject. To have a negative set (i.e., PTC-), we also obtained a list of transcripts without high-impact consequences for each subject (i.e., VEP impact lower than HIGH). To analyze the effects of NMD, we only considered the subjects in the control set since they are assumed to have functional NMD and selected transcripts that have PTC+ instances in at least 2 subjects and PTC- instances in at least 2 subjects. For each in the selected transcripts (LCL, n=476; fibroblasts, n=205), we then computed the mean TPM and log-transformed it with the constant 1 added for both PTC+ and PTC-. We compared the abundance data between PTC+ and PTC- and found that there are no significant differences in their abundances at the global level (Supplemental figure S6). These results suggest that the efficacy of NMD in LCL and fibroblasts is low for specific aberrant transcripts carrying PTCs in our dataset.

Analysis of changes in the relative abundance of PTC+ isoforms between case and control samples

To further check the efficacy of the NMD-mediated PTC+ isoform degradation, we analyzed the relative abundance of PTC+ isoforms between the case and the control groups. To this end, we computed the ratios of the sum of PTC+ isoform sets with the constant 1 to the PTC- counterparts of the case and the control groups. To make the abundance ratios comparable between the two groups, we computed each ratio using the same sets of PTC+ and PTC- isoforms found in a subset of samples in each of the two groups. For each isoform set, we calculated the mean of log-transformed ratio. Not surprisingly, the stringent requirement for the filtering of isoforms made the number of data points relatively small (in the LCL dataset, n=43; in the fibroblasts dataset, n=32). We first performed the pairwise comparison of the ratios between the case and control groups (Supplemental figure S7A) and found that there are no significant differences between the two groups. We also found that a large fraction of the data points are positive, implying that the abundance levels of PTC+ isoforms are higher than the PTC- counterparts regardless of the deficiency in SMG8 or SMG9. We then analyzed the distributions of the ratios between the two groups and found that there are no significant differences (Supplemental figure S7B). Indeed, the two-sample Kolmogorov-Smirnov test, with alternative being the ratios are greater in the case samples than in the control ones, indicated no evidence of differences in the ratios between the two groups ($p > 0.85$ for the LCL and fibroblast datasets). Taken together, consistent with the results from the PTC analysis in the control samples shown in Supplemental figure S6, these suggest low efficacy of the NMD-mediated PTC+ isoform degradation in these datasets.

References

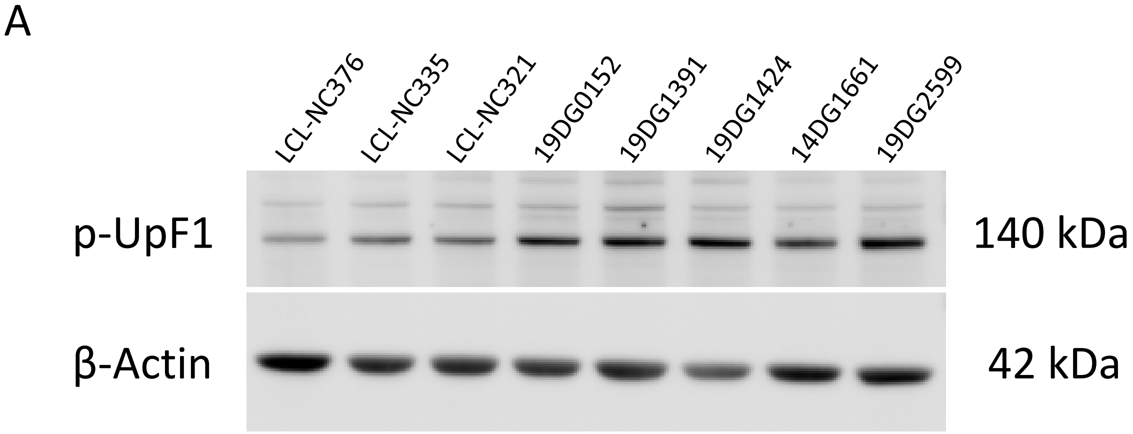
[Vandesompele et al., *Genome Biol*, 2002] Vandesompele, J., De Preter, K., Pattyn, F. *et al.* Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol* 3, research0034.1 (2002). https://doi.org/10.1186/gb-2002-3-7-research0034

[McLaren et al., *Genome Biol*, 2016] McLaren, W., Gil, L., Hunt, S.E. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* 17, 122 (2016). https://doi.org/10.1186/s13059-016-0974-4

[Bray et al., Nat Biotechnol, 2016] Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification [published correction appears in Nat Biotechnol. 2016 Aug 9;34(8):888]. *Nat Biotechnol*. 2016;34(5):525-527. doi:10.1038/nbt.3519

Figure S1. Recessive *SMG8* and *SMG9* variants are associated with increased UPF1 phosphorylation.
A) Western blot analysis on LCLs from affected individuals with *SMG8* variants (19DG1424, 19DG1391 and 19DG0152) from family 1, 2 and 3 respectively, affected individuals with *SMG9* variants (14DG1661 and 19DG2599) and three controls using polyclonal rabbit anti-phospho-Upf1 (Ser1100) (Cat. # 07-1015, EMD Millipore, (1:1,000)) and anti-β-actin (Cat. #mAbcam 8224, Abcam, (1:2000)). B) Quantification of p-UPF1 level relative to controls after correcting for the β-actin level based on six experiments. Please note that the three controls were combined together (NCT).

Figure S1


A
LCL-NC376
LCL-NC335
LCL-NC321
19DG0152
19DG1391
19DG1424
14DG1661
19DG2599
p-UpF1
140 kDa
β-Actin
42 kDa

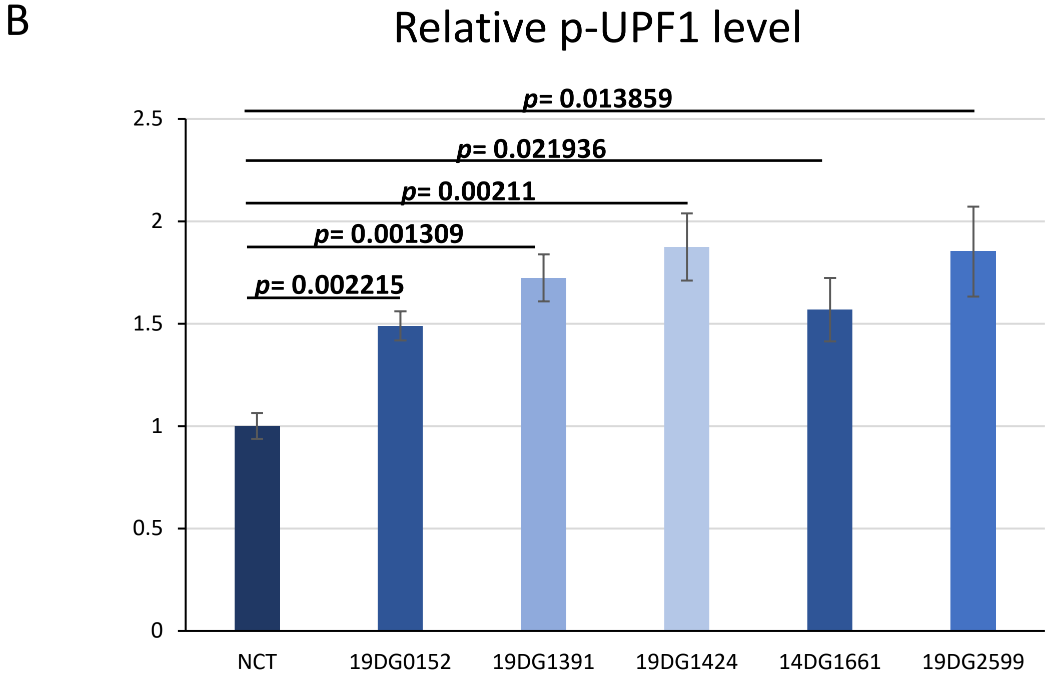

B
Relative p-UPF1 level
p= 0.013859
p= 0.021936
p= 0.00211
p= 0.001309
p= 0.002215
2.5
2
1.5
1
0.5
0
NCT
19DG0152
19DG1391
19DG1424
14DG1661
19DG2599

Figure S2. Assessment of the TPM normalization of the RNA-seq data.

The x-axis is the internal control gene-stability measure (M), while the y-axis is the between-sample coefficient of variation (CV) of the normalized count. Genes with the mean TPM > 50 are used for this assessment. The rho value represents Spearman's rank correlation coefficient between M and CV. P represents the p-value with the null hypothesis being no correlation between M and CV. (A) The fibroblast dataset. (B) The LCL dataset.

Figure S2

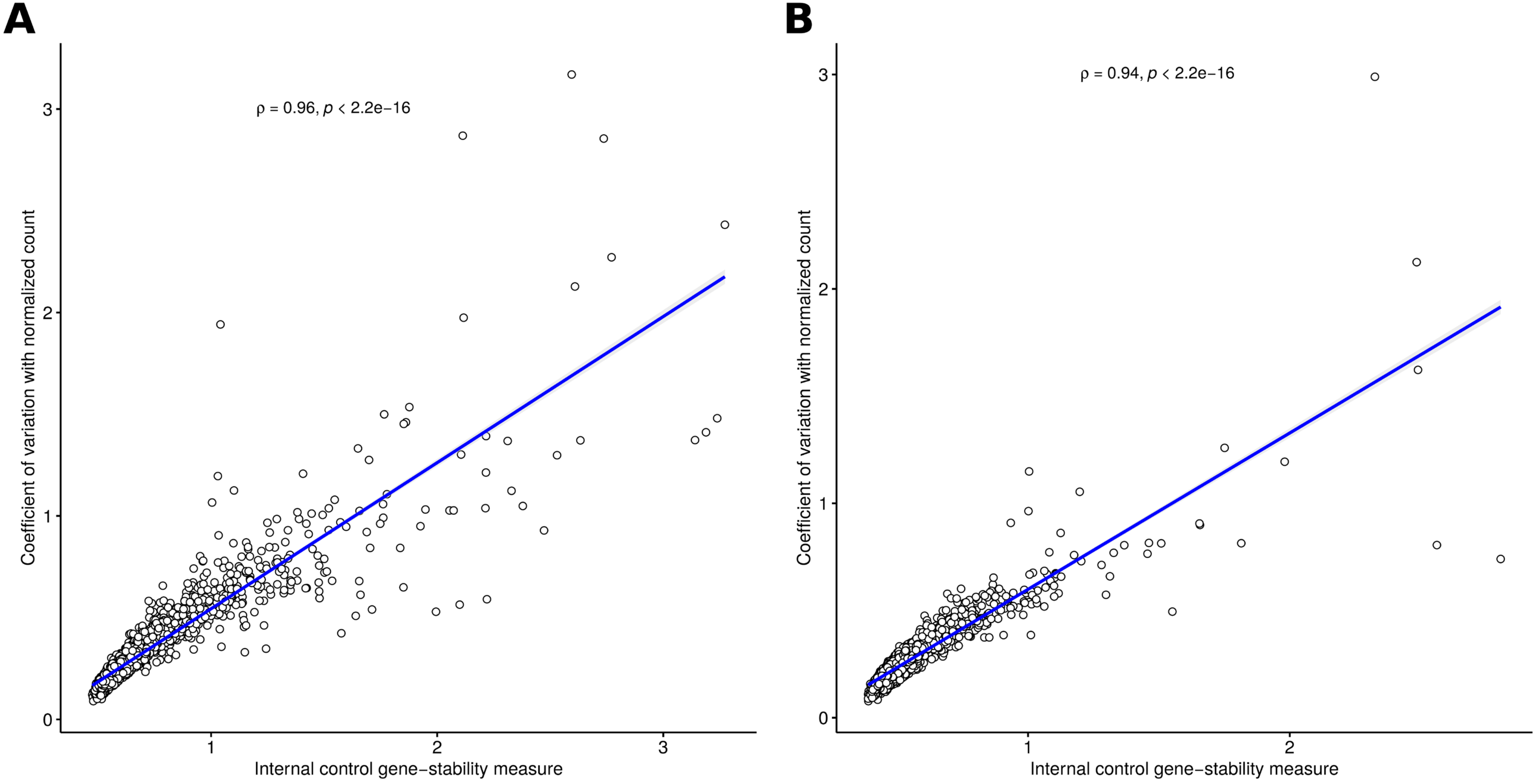

A
ρ = 0.96, p < 2.2e−16
Coefficient of variation with normalized count
0
1
2
3
Internal control gene−stability measure
1
2
3
B
ρ = 0.94, p < 2.2e−16
Coefficient of variation with normalized count
0
1
2
3
Internal control gene−stability measure
1
2

Figure S3. PCA applied on (A) the fibroblast cell type and (B) the LCL type.

Figure S3

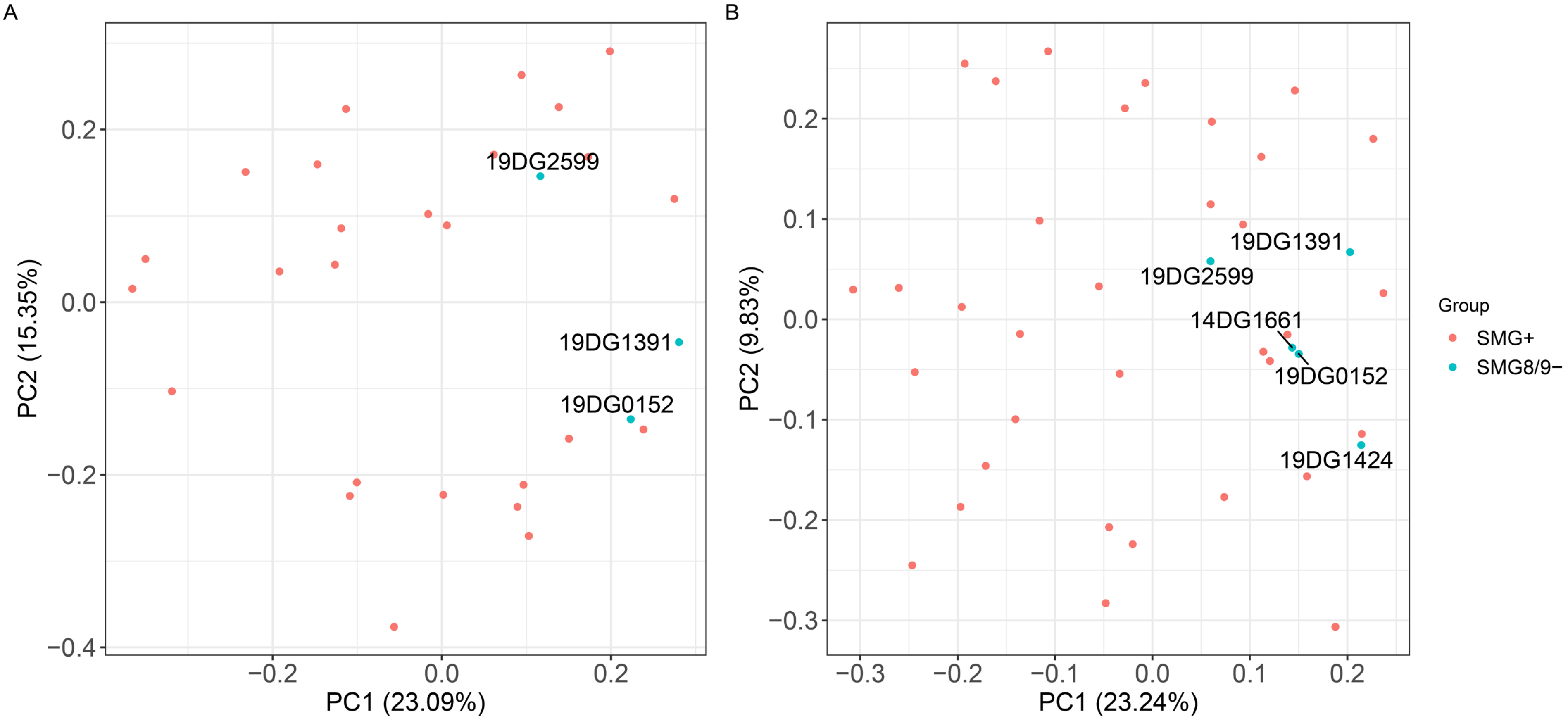

A
B
0.2
0.0
−0.2
−0.4
PC2 (15.35%)
−0.2
0.0
0.2
PC1 (23.09%)
19DG2599
19DG1391
19DG0152
0.2
0.1
0.0
−0.1
−0.2
−0.3
PC2 (9.83%)
−0.3
−0.2
−0.1
0.0
0.1
0.2
PC1 (23.24%)
19DG1391
19DG2599
14DG1661
19DG0152
19DG1424
Group
● SMG+
● SMG8/9−

Figure S4. Core NMD substrates expression profiles in *SMG8/9*-mutated (A) fibroblast cells and (B) LCL.

Figure S4

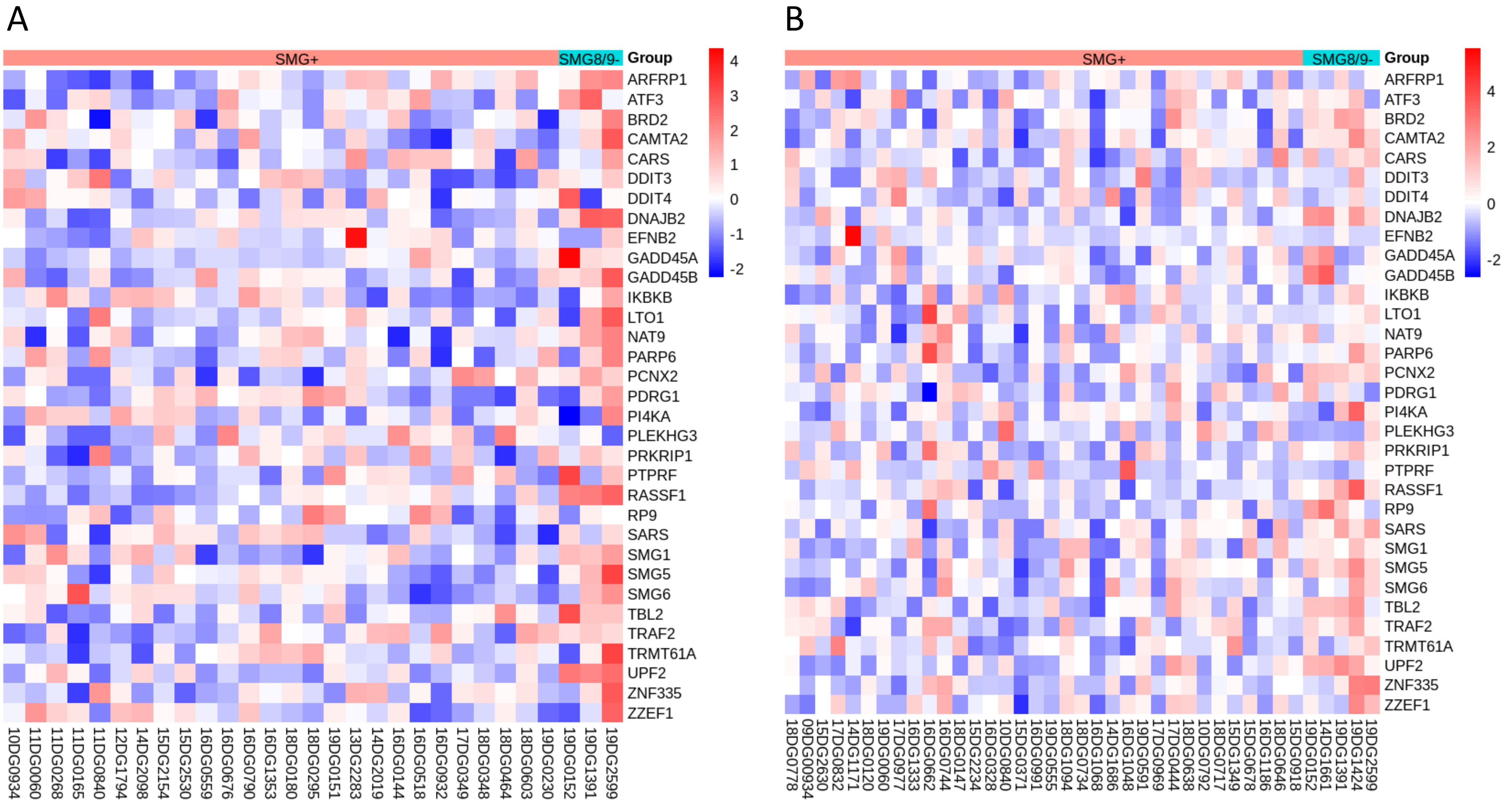

A
SMG+
SMG8/9-
Group
4
3
2
1
0
-1
-2
ARFRP1
ATF3
BRD2
CAMTA2
CARS
DDIT3
DDIT4
DNAJB2
EFNB2
GADD45A
GADD45B
IKBKB
LTO1
NAT9
PARP6
PCNX2
PDRG1
PI4KA
PLEKHG3
PRKRIP1
PTPRF
RASSF1
RP9
SARS
SMG1
SMG5
SMG6
TBL2
TRAF2
TRMT61A
UPF2
ZNF335
ZZEF1
10DG0934
11DG0060
11DG0268
11DG0165
11DG0840
12DG1794
14DG2098
15DG2154
15DG2530
16DG0559
16DG0676
16DG0790
16DG1353
18DG0180
18DG0295
19DG0151
13DG2283
14DG2019
16DG0144
16DG0518
16DG0932
17DG0349
18DG0348
18DG0464
18DG0603
19DG0230
19DG0152
19DG1391
19DG2599
B
SMG+
SMG8/9-
Group
4
2
0
-2
ARFRP1
ATF3
BRD2
CAMTA2
CARS
DDIT3
DDIT4
DNAJB2
EFNB2
GADD45A
GADD45B
IKBKB
LTO1
NAT9
PARP6
PCNX2
PDRG1
PI4KA
PLEKHG3
PRKRIP1
PTPRF
RASSF1
RP9
SARS
SMG1
SMG5
SMG6
TBL2
TRAF2
TRMT61A
UPF2
ZNF335
ZZEF1
18DG0778
09DG00934
15DG2630
17DG0832
14DG1171
18DG0120
19DG0060
17DG0977
16DG1333
16DG0662
16DG0744
18DG0147
15DG2234
16DG0328
10DG0840
15DG0371
16DG0991
19DG0555
18DG1094
18DG0734
16DG1068
14DG1686
16DG1048
19DG0591
17DG0969
17DG0444
18DG0638
10DG0792
18DG0717
15DG1349
15DG0678
16DG1186
18DG0646
15DG0918
19DG0152
14DG1661
19DG1391
19DG1424
19DG2599

Figure S5. Expression profiles of high-confidence NMD substrates upregulated in siUPF1-treated hESCs (Lou, et al) in *SMG8*- and *SMG9*-mutated (A) fibroblasts and (B) LCL.

Figure S5

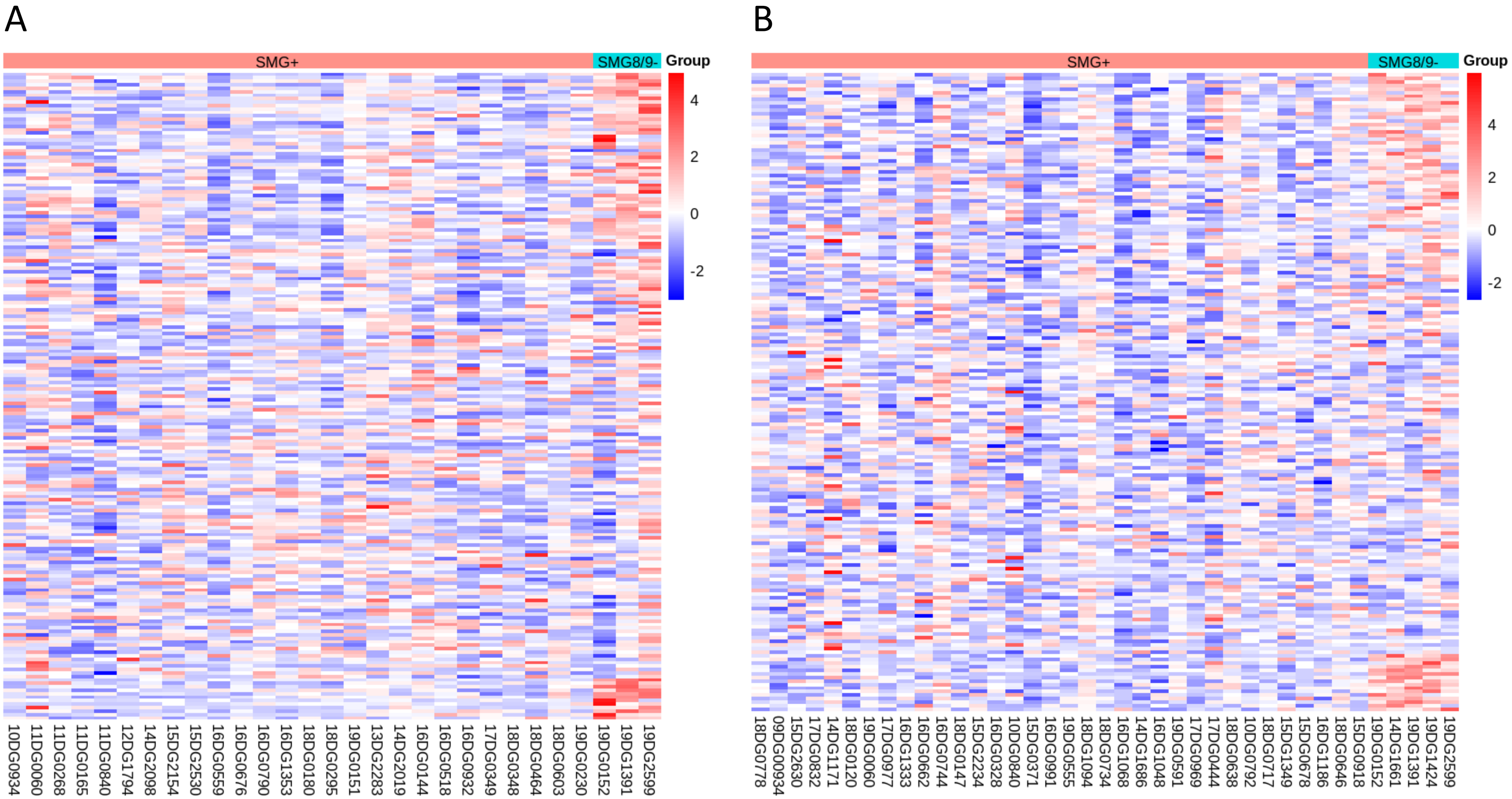

A
SMG+
SMG8/9-
Group
4
2
0
-2
10DG0934
11DG0060
11DG0268
11DG0165
11DG0840
12DG1794
14DG2098
15DG2154
15DG2530
16DG0559
16DG0676
16DG0790
16DG1353
18DG0180
18DG0295
19DG0151
13DG2283
14DG2019
16DG0144
16DG0518
16DG0932
17DG0349
18DG0348
18DG0464
18DG0603
19DG0230
19DG0152
19DG1391
19DG2599
B
SMG+
SMG8/9-
Group
4
2
0
-2
18DG0778
09DG00934
15DG2630
17DG0832
14DG1171
18DG0120
19DG0060
17DG0977
16DG1333
16DG0662
16DG0744
18DG0147
15DG2234
16DG0328
10DG0840
15DG0371
16DG0991
19DG0555
18DG1094
18DG0734
16DG1068
14DG1686
16DG1048
19DG0591
17DG0969
17DG0444
18DG0638
10DG0792
18DG0717
15DG1349
15DG0678
16DG1186
18DG0646
15DG0918
19DG0152
14DG1661
19DG1391
19DG1424
19DG2599

Figure S6. Analysis of the effects of NMD-target PTCs on the abundance of transcripts using the subjects in the control set (i.e., subjects with known functional NMD).

The x-axis is the abundance of transcripts without NMD-target PTCs (PTC-), while the y-axis is the abundance of transcripts with NMD-target PTCs (PTC+). Each point in the scatter plot shows a pair of the mean TPM values of PTC- and PTC+ with the constant 1 added. The blue line indicates the linear regression line with its slope value indicated at the top-right corner. The left pane shows the results from the fibroblast dataset, while the right pane shows the results from the LCL dataset.

Figure S6

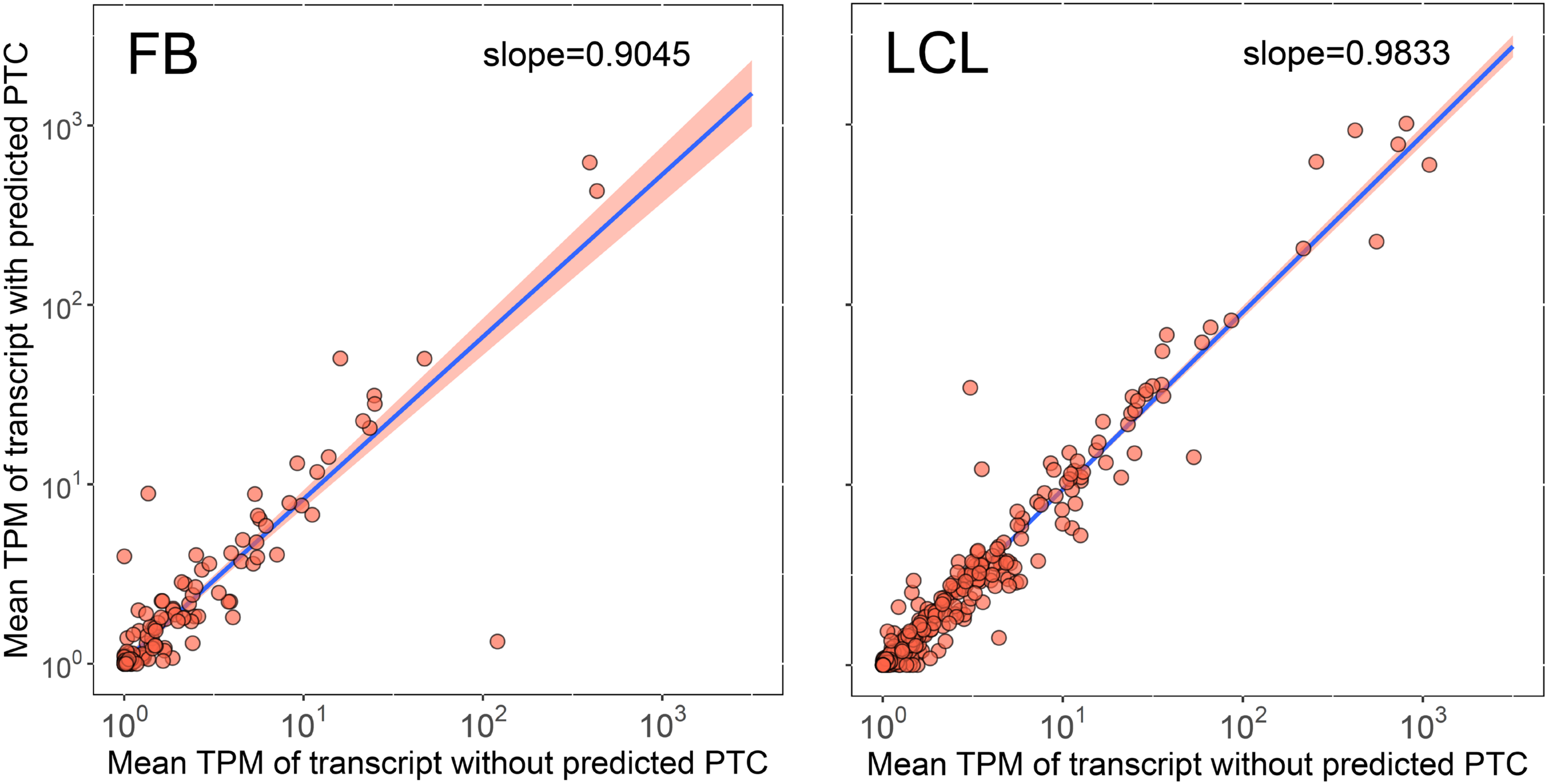
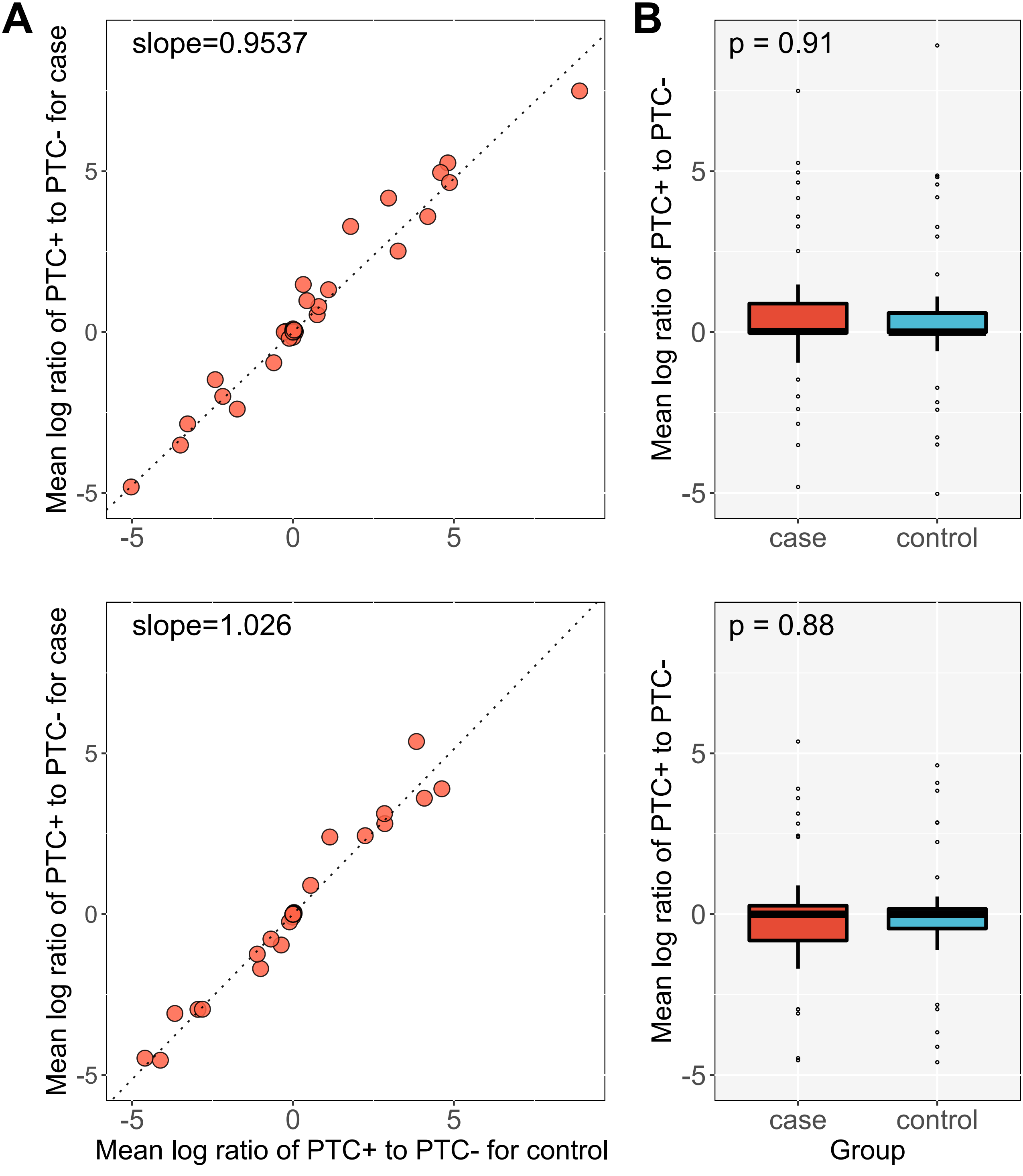

FB
slope=0.9045
LCL
slope=0.9833
Mean TPM of transcript with predicted PTC
Mean TPM of transcript without predicted PTC
Mean TPM of transcript without predicted PTC
10^0
10^1
10^2
10^3

Figure S7.

Comparison of the relative abundance of PTC+ and PTC- isoforms between the case and control groups. (A) Scatterplots comparing the means of log2-transformed abundance ratios of PTC+ and PTC- isoforms between case (y-axis) and control (x-axis) groups. The slope indicated at the upper-left corner is the slope of the linear regression line. (B) Boxplots showing the distribution of the means of log-transformed ratios. The upper panes show the results from the LCL dataset, while the bottom panes show those from the fibroblasts dataset. The p-value at the upper-left corner is from the Kolmogorov-Smirnov test with alternative being the the distribution of the cases greater than that of the controls.


A
slope=0.9537
Mean log ratio of PTC+ to PTC- for case
5
0
-5
-5
0
5
B
p = 0.91
Mean log ratio of PTC+ to PTC-
5
0
-5
case
control
slope=1.026
Mean log ratio of PTC+ to PTC- for case
5
0
-5
-5
0
5
Mean log ratio of PTC+ to PTC- for control
p = 0.88
Mean log ratio of PTC+ to PTC-
5
0
-5
case
control
Group