

# eScholarship

**Title**

Quantitating the epigenetic transformation contributing to cholesterol homeostasis using Gaussian process

**Permalink**

<https://escholarship.org/uc/item/16k6n8sd>

**Journal**

Nature Communications, 10(1)

**ISSN**

2041-1723

**Authors**

Wang, Chao  
Scott, Samantha M  
Subramanian, Kanagaraj  
et al.

**Publication Date**

2019

**DOI**

10.1038/s41467-019-12969-x




Peer reviewed

ARTICLE

<https://doi.org/10.1038/s41467-019-12969-x>

OPEN

# Quantitating the epigenetic transformation contributing to cholesterol homeostasis using Gaussian process

Chao Wang <sup>1</sup>, Samantha M. Scott<sup>1</sup>, Kanagaraj Subramanian<sup>1</sup>, Salvatore Loguercio <sup>1</sup>, Pei Zhao<sup>1</sup>, Darren M. Hutt<sup>1</sup>, Nicole Y. Farhat<sup>2</sup>, Forbes D. Porter<sup>2</sup> & William E. Balch <sup>1\*</sup>

To understand the impact of epigenetics on human misfolding disease, we apply Gaussian-process regression (GPR) based machine learning (ML) (GPR-ML) through variation spatial profiling (VSP). VSP generates population-based matrices describing the spatial covariance (SCV) relationships that link genetic diversity to fitness of the individual in response to histone deacetylases inhibitors (HDACi). Niemann-Pick C1 (NPC1) is a Mendelian disorder caused by >300 variants in the NPC1 gene that disrupt cholesterol homeostasis leading to the rapid onset and progression of neurodegenerative disease. We determine the sequence-to-function-to-structure relationships of the NPC1 polypeptide fold required for membrane trafficking and generation of a tunnel that mediates cholesterol flux in late endosomal/lysosomal (LE/Ly) compartments. HDACi treatment reveals unanticipated epigenomic plasticity in SCV relationships that restore NPC1 functionality. GPR-ML based matrices capture the epigenetic processes impacting information flow through central dogma, providing a framework for quantifying the effect of the environment on the healthspan of the individual.

<sup>1</sup>Department of Molecular Medicine, Scripps Research, La Jolla, CA 92037, USA. <sup>2</sup>Section on Molecular Dysmorphology, Division of Translational Medicine, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, MD 20814, USA.

\*email: [webalch@scripps.edu](mailto:webalch@scripps.edu)

A major challenge in the genomic era is to understand the impact of genetic and epigenetic diversity on the protein fold and its function in biology in response to the environment<sup>1–4</sup>. Lysine (Lys) acetylation/deacetylation balance, which is catalyzed by writers (histone acetyltransferases (HATs)) and erasers (histone deacetylases (HDACs)), manage epigenetic information to harmonize numerous cellular processes through readers (acetyl-binding proteins)<sup>5</sup>. Given the strong links to development, metabolism and aging, Lys acetylation/deacetylation has been considered an important epigenetic mechanism that uses cellular and environmental cues to specialize individual phenotypes in response to their unique genomic and environment exposure<sup>5–8</sup>. How the acetylation/deacetylation balance managed by HDAC coordinates the phenotype of genetic variation programmed by central dogma that is responsible for fitness in health and disease is largely unknown.

To generate a more complete understanding of the impact of HDAC on genetic variation in the population and its impact on the individual, we apply variation spatial profiling (VSP)<sup>1</sup> to understand the process triggering onset and progression of Niemann-Pick C1 (NPC1) disease, an early onset neurodegenerative disease caused by accumulation of unesterified cholesterol in the late endosome/lysosome (LE/Ly) compartments of all cells. VSP is a Gaussian-process regression (GPR)-based machine-learning (ML) (GPR-ML) approach<sup>9,10</sup> we have recently developed to characterize spatial covariance (SCV) relationships that predict the genomic-phenomic transformation in the individual in response to genetic diversity in the extant population<sup>1,3,11</sup>. VSP emphasizes sequence-to-function as a prior to structure as it is function that dictates the process of fitness in evolvability in response to the local genetic, biologic and physical environment<sup>1</sup>. Using our VSP strategy to define the impact of HDAC on Niemann-Pick C1 variants found in the world-wide population that disrupt cholesterol homeostasis, we show on a residue-by-residue basis that epigenetic-sensitive SCV relationships in the protein fold can be retuned by HDACi to coordinate protein trafficking with the flow of cholesterol in the LE/Ly. Restoring plasticity to the fold design disrupted by variation in the population we predict will delay the age of neurological disease onset. Our results highlight the power of GPR-ML to quantitatively capture the role of HDAC biology in acetylation balance to improve fitness using a SCV matrix based definition of central dogma (SCV[DNA|RNA|Protein])<sup>1</sup>.

## Results

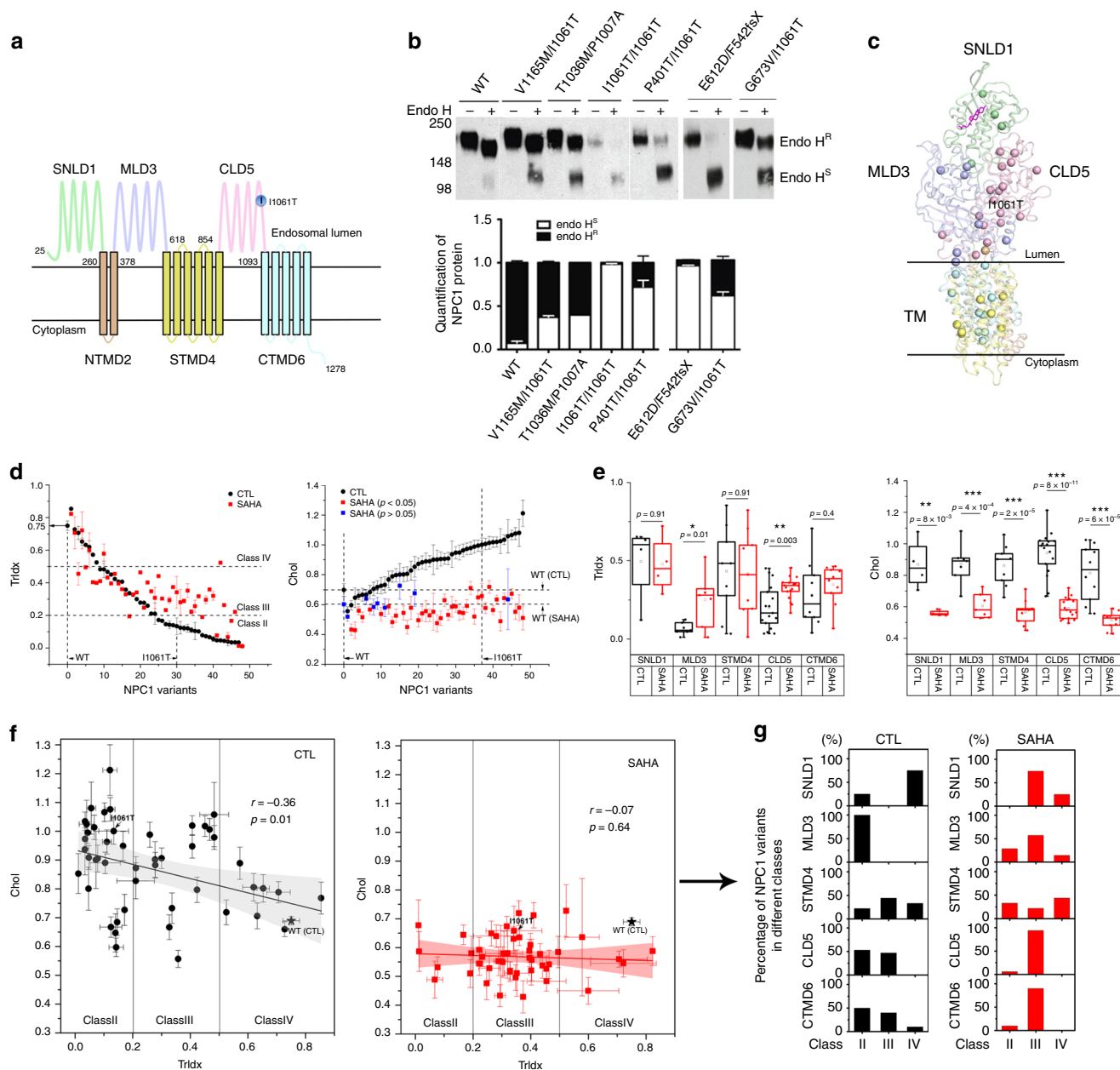
**NPC1 variant impact on trafficking and cholesterol.** To characterize the process by which HDACs impact genetic variation in the human population using GPR-ML<sup>1</sup>, we focused on the rare autosomal recessive childhood disease, Niemann-Pick type C (NPC). NPC disease is triggered by familial genetic diversity within the NPC1 gene (95% of patients) through >300 disease variants<sup>12,13</sup>. Defects in NPC1 function lead to abnormal accumulation of cholesterol and other lipids in the LE/Ly in all cell types, although typically manifested in clinical presentation as neurodegeneration where the age of onset and severity of disease is unique for each patient genotype<sup>13–17</sup>. NPC1 encodes a multi-membrane spanning protein (Fig. 1a) that is translocated and folded in the endoplasmic reticulum (ER) and trafficked through the Golgi to late endosome (LE)/lysosome (Ly) (LE/Ly) compartments where it manages cellular cholesterol homeostasis<sup>18,19</sup>. Immunoblot analysis of patient-derived NPC1 primary fibroblasts harboring different alleles shows substantial heterogeneity in both polypeptide expression and stability compared with fibroblasts expressing the wild-type (WT) NPC1 (Supplementary Fig. 1, Supplementary Table 1).

NPC1 acquires up to 14 N-linked glycans during co-translational translocation into the ER. These ER-localized high mannose glycoforms are sensitive to digestion by endoglycosidase (Endo) H (Endo H<sup>S</sup>) in cell lysates prepared by detergent solubilization. Following delivery from the ER to the Golgi, the N-linked glycans are progressively processed to Endo H resistant (Endo H<sup>R</sup>) glycoforms by the Golgi, leading to slower migration on SDS-PAGE. In primary fibroblasts (Fig. 1b), the WT-NPC1 glycoform was highly resistant to Endo H digestion, indicating efficient transfer to the Golgi whereas heterozygous alleles and the common I1061T homozygous variant fibroblast populations<sup>20</sup> showed differential sensitivity to Endo H with I1061T being restricted to the ER (Fig. 1b). To characterize the impact of each allele on NPC1 features, we silenced NPC1 expression with shRNA to generate stable HeLa and U2OS null cell lines<sup>21</sup>. These null cell lines were transiently transfected with a sparse collection of plasmids that each harbors one of 48 distinct NPC1 disease-associated variants distributed among the various NPC1 domains (Fig. 1c, Supplementary Fig. 2)<sup>21</sup>.

To follow trafficking from the ER, we measured the level of Endo H<sup>S</sup> and Endo H<sup>R</sup> glycoforms for each variant and generated an Endo H<sup>R</sup>/(Endo H<sup>S</sup> + Endo H<sup>R</sup>) ratio that reports on the ER versus post-ER distribution of NPC1 variants in the cell, referred to hereafter as the trafficking index (TrIdx)<sup>1</sup>. We binned this sparse collection of variants into four functional classes (classes I–IV) based on TrIdx (Fig. 1d, left panel, black circles; Supplementary Table 2, Source data are provided as a Source Data file). Class I variants lack polypeptide expression in response to non-sense or splicing (truncation) mutations (null) (Supplementary Table 2); class II missense variants are largely ER retained (defined as  $\sim < 0.2$  TrIdx); class III missense variants show an intermediate level of trafficking (defined as  $\sim 0.2$  to  $\sim 0.5$  TrIdx); and class IV missense variants have a level of TrIdx  $\sim > 0.5$  indicating significant export from the ER (Fig. 1d, left panel; Supplementary Table 2). The distribution of TrIdx of each variant in the two cell lines tested were largely consistent (Supplementary Fig. 3, Pearson's  $r = 0.55$ ;  $p$ -value =  $1 \times 10^{-4}$  (ANOVA test)) with the exceptions where select class II variants in HeLa cells are scored as class III in U2OS cells reflecting increased flow from the ER (Supplementary Fig. 3, Supplementary Table 2). Comparison of the two developmentally distinct cell-based lineages suggests a remarkable general conservation of NPC1 variant trafficking through the endomembrane system to the LE/Ly (Supplementary Fig. 3).

The impact of cholesterol (Chol) storage in the LE/Ly for each of these variants was measured using a well-established automated high content screening image analysis based on filipin staining in the U2OS cell line<sup>21,22</sup> (Fig. 1d, right panel, black circles). The combined results revealed a gradient of cholesterol accumulation in the LE/Ly compartments reflecting the differential impact of a specific variant on trafficking and/or function in the LE/Ly (Fig. 1d, black circles)<sup>21</sup>. Because each of the 48 variants tested contribute to clinical disease, these results suggest that either nascent synthesis, ER stability, trafficking from the ER or function in LE/Ly can differentially contribute to clinical presentation of disease.

**Impact of SAHA on trafficking and cholesterol.** The FDA-approved histone deacetylase inhibitors (HDACi) Vorinostat (SAHA) and Panobinostat (LBH589) can correct the I1061T phenotype in cell-based models by stabilizing the I1061T protein for export to the LE/Ly where it contributes to improved cholesterol homeostasis<sup>21,23,24</sup>. Each of the 48 NPC1 variants were examined for the impact of SAHA on Chol in the LE/Ly in the U2OS cell line to that of trafficking in the HeLa cell line given the



**Fig. 1** Response of NPC1 variants to SAHA. **a** Schematic representation of domains in NPC1 (sterol-binding luminal N-terminal domain 1 (SNLD1), N-terminal transmembrane domain 2 (NTMD2), middle luminal domain 3 (MLD3), sterol-sensing transmembrane domain 4 (STMD4), C-terminal luminal domain 5 (CLD5), C-terminal transmembrane domain 6 (CTMD6)). **b** (Upper panel) Trafficking of NPC1 variants from the ER (Endo H sensitive (Endo H<sup>S</sup>)) to post-ER LE/LY compartments (Endo H resistant (Endo H<sup>R</sup>)) glycoforms in wild-type (WT) and NPC1 patient fibroblasts. (Bottom panel) Fraction of total of NPC1 in Endo H<sup>R</sup> (black) and Endo H<sup>S</sup> (white) glycoforms (mean  $\pm$  s.d.). **c** NPC1 variants are mapped as balls on the C-alpha position in the NPC1 structure<sup>25,29</sup>. The cholesterol molecule is shown in magenta<sup>25,35</sup>. **d** (Left panel) Trafficking index (TrIdx) of 48 NPC1 variants in the absence (black circles as control (CTL)) or presence of SAHA (red squares). WT, I1061T variants are indicated by vertical dash lines and trafficking classes (II, III, IV) by horizontal dash lines. (Right panel) Cholesterol (Chol) homeostasis of 48 NPC1 variants in the absence (black circles as control (CTL)) or presence of SAHA (red squares for  $p$ -value < 0.05; blue squares for  $p$ -value > 0.05; student's two tailed  $t$ -test)<sup>21</sup>. WT and I1061T variants are indicated by vertical dash lines. Chol of WT in the absence or presence of SAHA highlighted by horizontal dashed lines. Data are presented as mean  $\pm$  s.e.m.<sup>21</sup> (Source data are provided as a Source Data file). **e** Domain specific response to SAHA. NPC1 variants are clustered into different domains for TrIdx (left panel) and Chol (right panel) comparison (box and whisker plot: box = 25th and 75th, whiskers extend from minimum to maximum of the data;  $p$ -value (Student's two tailed  $t$ -test)). **f** Correlation between TrIdx and Chol of NPC1 variants in the absence (left panel) or presence (right panel) of SAHA. Pearson's  $r$  and the  $p$ -value (ANOVA test) with null hypothesis as the coefficient equal to zero is indicated. **g** Redistribution of different trafficking classes in response to SAHA. The percentage of variants in different classes in absence (**f**, left panel) or presence (**f**, right panel) of SAHA is shown

larger differential of trafficking features captured for each variant (Fig. 1d; Supplementary Fig. 3; Supplementary Table 2). Similar to NPC1-I1061T fibroblasts<sup>20,21,23</sup>, the majority of NPC1 variants with a TrIdx <0.2 (Fig. 1d, left panel, class II) were corrected by SAHA to at least a class III TrIdx, while variants with >0.2 TrIdx (Fig. 1d, left panel, classes III and IV) showed a more variable impact of SAHA on TrIdx. In contrast, HDACi improved Chol homeostasis for most variants to a level comparable to or greater than that observed in either untreated or SAHA-treated WT-NPC1 cells (Fig. 1d, right panel), a result similar to that observed for LBH589, a HDACi previously shown to correct I1061T at nM concentrations<sup>21</sup> (Supplementary Fig. 4a). The effects of HDACi were NPC1-dependent as restoration of Chol homeostasis by SAHA was not observed in fibroblasts lacking NPC1<sup>23</sup> and in null cell lines lacking NPC1<sup>21</sup>.

NPC1 contains three luminal domains (SNLD1, MLD3, and CLD5) and three transmembrane domains (NTMD2, STMD4, and CTMD6) with 13 transmembrane helices (TM 1–13) (Fig. 1a)<sup>25–29</sup>. When we grouped variants by structural domains (Fig. 1a, c), both the luminal MLD3 and CLD5 variants that have relatively low TrIdx values showed a statistically significant correction of TrIdx in their response to SAHA compared with all other domains (Fig. 1e, left panel). In contrast, cholesterol homeostasis was significantly restored by either SAHA (Fig. 1e, right panel) or LBH589 (Supplementary Fig. 4b, c)<sup>21</sup> for variants in all domains. While Chol homeostasis of NPC1 variants in the vehicle control showed a modest but significant correlation with TrIdx (Fig. 1f, left panel, Pearson's  $r$ -value =  $-0.36$ ,  $p$ -value =  $0.01$  (ANOVA test)), SAHA eliminated this correlation by shifting most variants to a class III phenotype (Fig. 1f, right panel, Pearson's  $r$ -value =  $-0.07$ ,  $p$ -value =  $0.64$  (ANOVA test), Fig. 1g). Moreover, the correlation between the delta ( $\Delta$ ) of Chol and  $\Delta$  of TrIdx in the absence or presence of SAHA, or the  $\Delta$  of Chol and the  $\Delta$  of the mature NPC1 glycoform, indicating delivery to the Golgi in response to SAHA, were not statistically significant (Supplementary Fig. 5). These results suggest that SAHA differentially impacts the management of variants contributing to ER export and those contributing to cholesterol homeostasis in the LE/Ly, illustrating that the HDACi modified environment can significantly and differentially contribute to variant functionality.

**VSP analysis of NPC1 variants.** To understand in depth the complex genotype to phenotype transformation process responsible for NPC1 disease in response to HDACi reported by NPC1 variants, we applied variation spatial profiling (VSP) to quantitate the spatial covariance (SCV) of sequence-to-function-to-structure relationships<sup>1</sup> for the entire NPC1 polypeptide in the absence or presence of HDACi. VSP uses sparse known variants and their related functional features in the population to predict a quantitative function value (referred to as a SCV set-point) for each residue in the polypeptide chain. These set-point values are associated with an uncertainty that assigns the SCV tolerance for all known and unknown variants in predicting any functional feature<sup>1,11</sup>. The combined known and predicted unknown values constitute a phenotype landscape that provides a quantitative interpretation of covarying sequence-to-function-to-structure relationships across the entire NPC1 fold, in this case the impact of epigenetics on the functional rescue of NPC1 variant by HDACi.

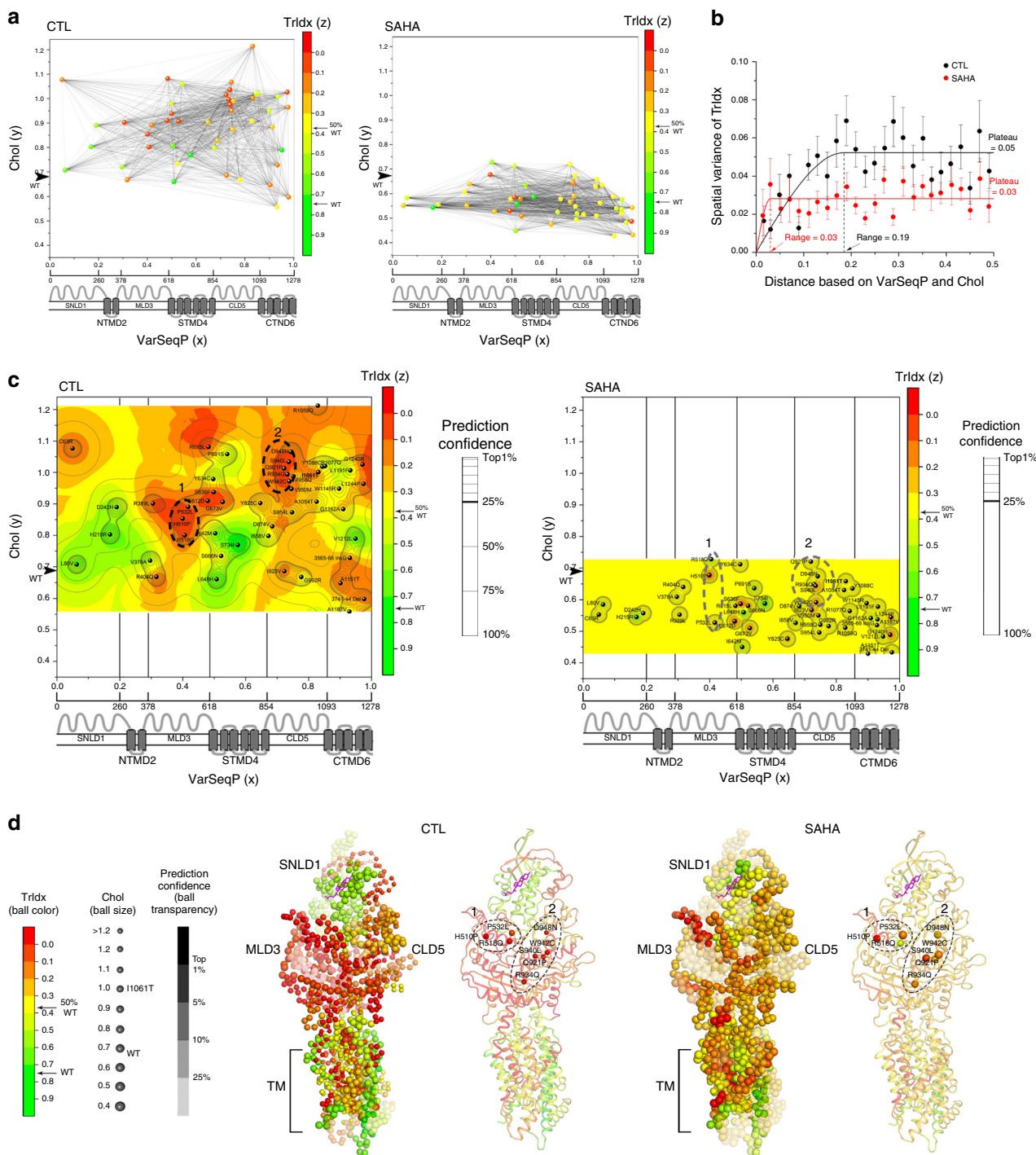
To build phenotype landscapes<sup>1</sup>, the  $x$ -axis is the normalized linear position of each of the 48 variant in the NPC1 polypeptide sequence (referred to as VarSeqP) that provides spatial information across the entire primary sequence as would be normally assigned by transcription and translation machineries (Fig. 2a). To assign the relationship of each genotype to a function, the  $y$ -

axis input was set as the known value of each variant's Chol measurement (Fig. 2a). To address the impact of VarSeqP-to-Chol relationships on the TrIdx, given the spatial importance of cellular location in proper management of cholesterol by NPC1 in LE/Ly<sup>18,30–32</sup>, an input  $z$ -value was set as the known TrIdx value for each variant (Fig. 2a,  $z$ -axis, colored gradient). The spatial relationships defined by the  $x$ -axis genotype and functional  $y$ - and  $z$ -axes coordinates provide a quantitative framework to understand the transformation from the linear genotype to phenotype reflected in function based on the choice of  $y$ - and  $z$ -axis coordinates<sup>1</sup>.

To assess the SCV relationships for all possible 1128 pairwise combinations of known variants (Fig. 2a), a molecular variogram<sup>1</sup> is computed (Fig. 2b, Supplementary Fig. 6a). Here, the molecular variogram shows that the spatial variance of TrIdx in the absence of SAHA (Fig. 2b, black line,  $y$ -axis) increases according to VarSeqP-to-Chol distances (Fig. 2b,  $x$ -axis), referred to as the range<sup>1</sup> (Fig. 2b,  $x$ -axis position 0.19, vertical dashed black line), until it reaches a plateau (Fig. 2b,  $y$ -axis position 0.05, horizontal black line). A range of 0.19 suggests that the TrIdx and Chol function of variants are generally dependent on each other only over a short sequence range of ~250 amino acids (Fig. 2b,  $x$ -axis)<sup>1</sup>, indicating a SCV modular design of the NPC1 polypeptide sequence that relates genotype to the different phenotypes contributing to trafficking and cholesterol management. Strikingly, SAHA significantly reduced the range shown on the  $x$ -axis (Fig. 2b, red line SAHA, range value 0.03 (~40 amino acids)). Moreover, the spatial variance on the  $y$ -axis defines a plateau which describes the stringency of the fold<sup>1</sup>. Here, the plateau of the spatial variance for TrIdx is reduced by ~40% in response to SAHA (Fig. 2b; compare horizontal black line vehicle control (control (CTL), plateau = 0.05) to the horizontal red line in the presence of SAHA (plateau = 0.03); Supplementary Fig. 6a,  $p$ -value =  $7 \times 10^{-23}$  (student's  $t$ -test)). Thus, the plateau spatial variance defined by the molecular variogram suggests that HDACi increases the set-point tolerance of each known variant residue interaction leading to increased plasticity in TrIdx. The outcome is to reduce the loss-of-function impact of the variants, resulting in an improvement in cholesterol homeostasis for most of the disruptive variant features.

### Phenotype landscapes predict the TrIdx-functional structure.

To expand the spatial relationships modeled by the molecular variogram with known variants (Fig. 2b) to all other unknown residues, we apply GPR-ML<sup>1</sup> to generate an output TrIdx-phenotype landscape (Fig. 2c). The TrIdx-phenotype landscape quantitates all SCV relationships between the known and the unknown (predicted) TrIdx values ( $z$ -axis) for each amino acid residue in NPC1 polypeptide chain in response to the VarSeqP ( $x$ -axis) and Chol ( $y$ -axis) coordinates relationships, thereby linking sequence position to cellular location to cholesterol-related function of NPC1 (Fig. 2c). GPR-ML not only generates set-point predictions on a residue-by-residue basis but also assesses the uncertainty of the prediction (Supplementary Fig. 6b) (i.e., SCV tolerance of each set-point), of which the leave-one-out cross-validation achieves a more accurate prediction than multivariate linear regression or decision tree-based regression methods (Supplementary Fig. 6c–e). The uncertainty (or confidence) in using SCV relationships for prediction is defined by contour maps that are embedded in the phenotype landscape (Fig. 2c, Supplementary Fig. 6b) that show the strength of all SCV relationships found in the range of the molecular variogram (Fig. 2b)<sup>1</sup>. The TrIdx-phenotype landscape (Fig. 2c, left) reveals diverse trafficking values where variants with defective trafficking of high confidence are mainly localized in the luminal MLD3 and CLD5 domains (Fig. 2c, left; SCV clusters 1 and 2), indicating that



**Fig. 2** Trldx-phenotype landscape and Trldx-structure in the absence and presence of SAHA. **a** NPC1 variants are positioned by their variant sequence position (VarSeqP) (x-axis) normalized to the full-length sequence and measured Chol values of NPC1 variants (y-axis normalized to I1061T) in the absence (left panel) or presence (right panel) of SAHA. The projected z-axis (color gradient) is defined by the absolute measured Trldx of each variant. The spatial relationships of pairwise comparisons of variants are illustrated as black lines. **b** The molecular variogram illustrating the spatial variance of measured Trldx for distance values defined by VarSeqP and Chol in the absence (black line) or presence (red line) of SAHA (see Methods; Supplementary Fig. 6a). **c** VSP predicted Trldx-phenotype landscape is shown as a heatmap in the absence (left panel) or presence (right panel) of SAHA. Confidence relationships (see Supplementary Fig. 6b) within the molecular variogram range are plotted as contour lines and the top 25% confidence quartile is shown as a bold line. Color-scale values: defective trafficking (red); -50% WT trafficking (yellow); WT trafficking (green). I1061T ER-restricted variant is bolded. Dashed ovals 1 and 2 highlight two SCV clusters found in the top 25% confidence quartiles that have defective trafficking in the absence of SAHA. **d** Mapping the Trldx-phenotype landscape on the NPC1 structure<sup>25,29</sup>. The highest confident prediction of the Trldx value and the corresponding Chol value is assigned for each residue with ball color representing Trldx, ball size representing Chol, and transparency representing prediction confidence. SCV cluster 1 and 2 are selectively highlighted as balls in cartoon presentation of the structure

these sequence regions are critical for export from the ER based on poor acquisition of Endo H<sup>R</sup> glycoforms. Furthermore, TrIdx-phenotype landscapes based on input data from U2OS cells reveal similar regions critical for trafficking (Supplementary Fig. 6g), indicating the conserved role of MLD3 and CLD5 mediating NPC1 trafficking across different cell lines.

Strikingly, the TrIdx-phenotype landscape (Fig. 2c, left panel) undergoes a general compression in the presence of SAHA (Fig. 2c, right panel). These results reveal that the global improvement in cholesterol homeostasis in response to HDACi (Fig. 2c, y-axis) occurs, in part, through new SCV relationships that convert poor class II TrIdx values (Fig. 2c, left panel, orange-red) to improved class III TrIdx values (Fig. 2c, right panel, yellow). For example, SCV cluster 2 in CLD5 (Fig. 2c, left panel, dashed oval 2), that has a severe defect for both TrIdx and Chol, undergoes a coordinated shift towards WT-like Chol function (y-axis) and a class III TrIdx (z-axis) (Fig. 2c, right panel, dashed oval 2) as highlighted in a 3D projection of the TrIdx-phenotype landscape (Supplementary Fig. 6f, cluster 2). In contrast in SCV cluster 1, the lack of significant correction of Chol of H510P<sup>21</sup> is due to the inability of SAHA to improve its TrIdx (Fig. 2c, Supplementary Fig. 6f). Moreover, the large improvement of the TrIdx of R518Q set-point does not significantly improve cholesterol homeostasis reflecting the critical role of this residue in binding of NPC2 for cholesterol transfer (Fig. 2c, Supplementary Fig. 6f)<sup>21</sup>.

VSP generated TrIdx-phenotype landscapes (Fig. 2c) can be directly mapped onto the NPC1 structure<sup>25,27,29</sup> to generate a functional structure<sup>1</sup> that enables function assignment of each amino acid residue in a structure snapshot (Fig. 2d). The functional structure reveals the overlapping and/or distinct contribution of all the NPC1 residues to trafficking (Fig. 2d, ball color) and cholesterol homeostasis (Fig. 2d, ball size) with a prediction confidence (Fig. 2d, ball transparency)<sup>1</sup>. Strikingly, the TrIdx-functional structure reveals a molecular handshake between MLD3 and CLD5 (Fig. 2d, CTL panel, clusters 1 and 2) as a central feature that determines the ER export of NPC1 in HeLa (Fig. 2d, CTL panel, red residues) and U2OS cell lines (Supplementary Fig. 6h) from the ER. Moreover, atomic resolution mapping of the impact of SAHA on the functional structure reveals a significant improvement in cholesterol homeostasis for nearly all residues (Fig. 2d, SAHA panel, increased of ball size), partially by shifting the class II TrIdx (Fig. 2d, CTL panel, red balls) to that of class III indicative of significant export from the ER (Fig. 2d, SAHA panel, orange and yellow balls). Thus, the phenotype landscape predicted from the collective of fiduciary NPC1 variants found in the world-wide patient population using VSP teach us about the SCV modules and their function linked structural features required for normal NPC1 trafficking.

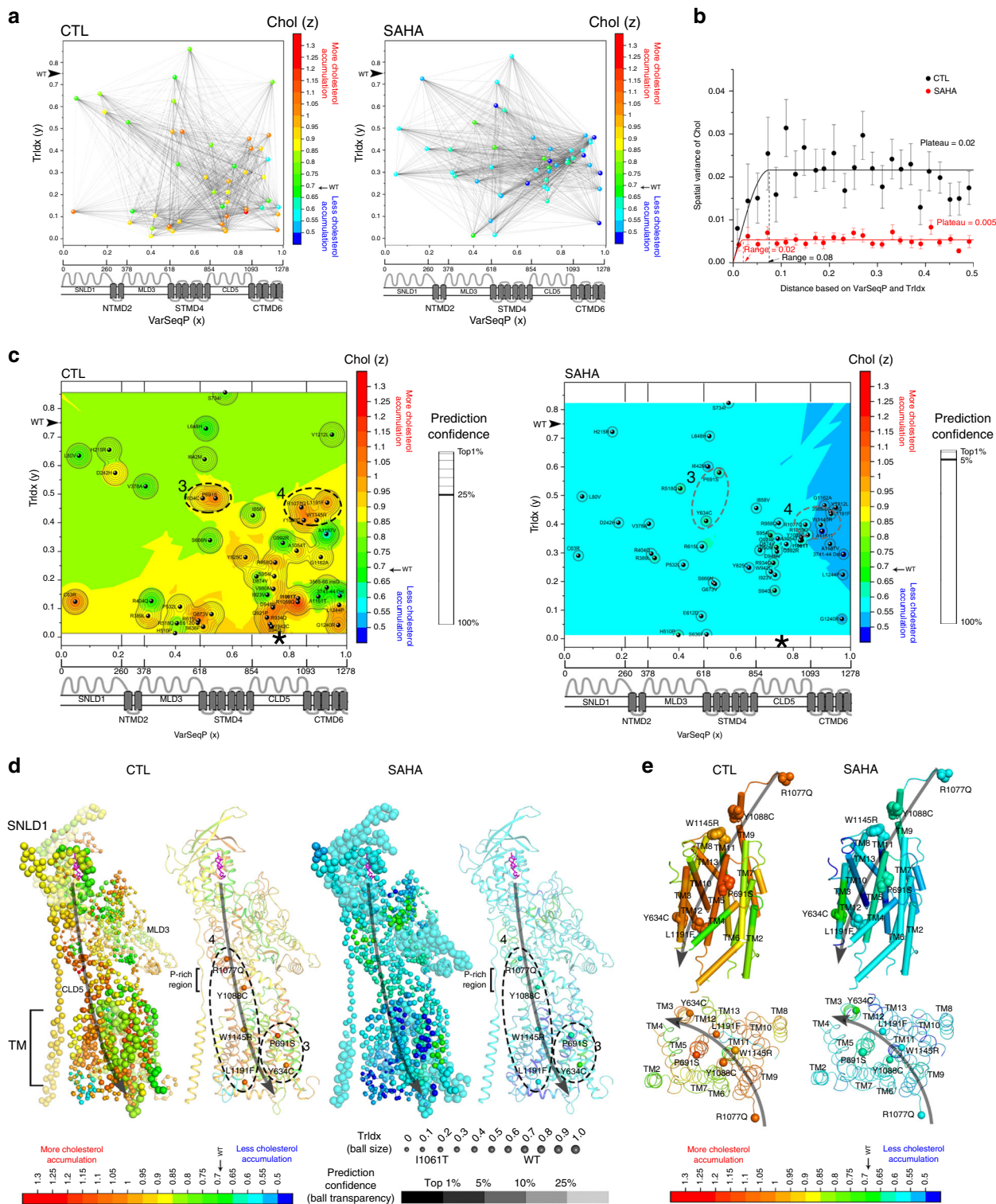
**Phenotype landscapes predict the Chol-functional structure.** To assign a value to cholesterol homeostasis based on the TrIdx response to NPC1 variants, we flipped the biological features used for y-axis and z-axis (Fig. 3a, Supplementary Fig. 7). The molecular variogram modeling of these relationships in the absence of SAHA shows that the spatial variance of Chol (Fig. 3b, y-axis) increases according to the distance defined by the VarSeqP-to-TrIdx spatial relationship (Fig. 3b, x-axis), revealing a range ~0.08 (Fig. 3b, black vertical dash line). A range of ~0.08 suggests that the spatial variance of cholesterol homeostasis (Fig. 3b, black line, y-axis) is dependent on VarSeqP-to-TrIdx over a sequence range of ~100 amino acids (Fig. 3b, black line, x-axis). Strikingly, SAHA decreases the range from 0.08 to 0.02 and significantly reduces the spatial variance of the Chol value by nearly 70% (Fig. 3b; compare

black line plateau = 0.02 with red line plateau = 0.005; Supplementary Fig. 7a,  $p$ -value =  $5 \times 10^{-79}$  (student's  $t$ -test)). The decreased Chol spatial variance of NPC1 variants in response to SAHA (Fig. 3b, 70%) is significantly larger than the decrease observed for TrIdx spatial variance (Fig. 2b, 40%), indicating that SAHA correction in post-ER LE/Ly acidic compartments increases the overall SCV plasticity<sup>1</sup> of the variant fold to improve function to facilitate Chol homeostasis.

We next generated an output Chol-phenotype landscape that predicts cholesterol responses across the entire polypeptide sequence in the context of all VarSeqP-TrIdx spatial relationships (Fig. 3c, Supplementary Fig. 7b–e). Interestingly, the Chol-phenotype landscape in the absence of SAHA reveals two SCV clusters in the top 25% confidence quartile that show class III TrIdx yet have severe cholesterol homeostasis defects (Fig. 3c, left, dashed ovals). One cluster is found in STMD4 (cluster 3) and the other spanning CLD5 and CTMD6 (cluster 4) (Fig. 3c, left panel). These SCV relationships suggest that clusters 3 and 4 are critical in mediating cholesterol management in the LE/Ly. Indeed, P691 in SCV cluster 3 has been shown to be involved in cholesterol binding<sup>26,27,33,34</sup>. Thus, the Chol-phenotype landscape reveals that regions in CLD5, CTMD6 and STMD4 contribute together to tune cholesterol flow (Fig. 3c, left panel).

The Chol-phenotype landscape (Fig. 3c, left panel) undergoes a striking change in the presence of SAHA (Fig. 3c, right panel) highlighting the ability of SAHA to improve cholesterol homeostasis for most of variants (Fig. 3c, red-orange in the left panel shifts to cyan and blue in the right panel). Moreover, the high confidence region (Fig. 3c, left panel, 25%; right panel, 5%, respectively) defined by the range in the molecular variogram (Fig. 3b) are decreased substantially by SAHA indicating a loss of spatial interdependency of variant residues triggering disease. The correction of cholesterol homeostasis (z-axis) of variants found in the regions defined by SCV clusters 3 and 4 even in the absence of improvement of their TrIdx (y-axis) indicates that SAHA can independently adjust the function of NPC1 in the LE/Ly to improve cholesterol homeostasis (Fig. 3c, Supplementary Fig. 7f)<sup>21</sup>.

By projecting Chol-phenotype landscape at atomic resolution onto the structure of NPC1, the variation seen in the NPC1 population can now be used to map a potential path of cholesterol flow in NPC1 (Fig. 3d, e). Based on class III variants (Fig. 3d, CTL panel, medium size balls) that are primarily defective in cholesterol homeostasis (Fig. 3d, CTL panel, orange-red balls), the Chol-functional structure reveals (Fig. 3d, CTL panel) the critical residues for cholesterol export that include SCV cluster 4 residues in CLD5 and CTMD6, as well as SCV cluster 3 residues in STMD4, the later recently proposed to form a second cholesterol binding site<sup>27</sup>. Moreover, the proline-rich linker between SNLD1 and the TM region that has been suggested to facilitate cholesterol transfer<sup>35</sup> is now shown by SCV analysis to have little impact on trafficking yet contribute to the flow of cholesterol in NPC1 (Fig. 3d, CTL panel). The structural mapping of Chol-phenotype landscapes based on trafficking values from the HeLa and U2OS cell lines yields similar pattern (Supplementary Fig. 7g, h), indicating VSP is able to capture a common phenotypic feature using different cell and tissues harboring variant genotypes<sup>1</sup>, suggesting that these are conserved genetic and epigenetic relationships. Based on our SCV analyses, a potential cholesterol flow path is highlighted in the TM region (Fig. 3e). The SCV relationships of variants found along the flow path lead to improved cholesterol homeostasis in response to SAHA (Fig. 3e, SAHA panel, green-cyan-blue), suggesting a role for CLD5 and CTMD6 in export of cholesterol from the LE/Ly by increasing NPC1 plasticity (i.e., a decrease in the SCV dependence) (Fig. 3b).



**HDACi responsive delta ( $\Delta$ )-functional structures.** Given the differences in the basal and HDACi condition (Figs. 2 and 3), we can generate functional structures that directly capture the delta ( $\Delta$ ) values from TrIdx and Chol-phenotype landscapes. The  $\Delta$  TrIdx structural state (Fig. 4a, b) highlights that the trafficking properties conferred by MLD3 and CLD5 residues are corrected by SAHA (Fig. 4a, cyan to blue). In contrast, SNLD1 is largely resistant to SAHA (Fig. 4a, yellow to red in left panel; gray in right panel). The TM region is also largely resistant to SAHA

(Fig. 4a, yellow to red in left panel; gray in right panel) as the trafficking of the variants involved in the predicted path of cholesterol flow are not changed by SAHA (Fig. 4b, yellow balls). In contrast to the  $\Delta$  TrIdx-functional structures, the  $\Delta$  Chol-functional structure highlights the fact that the residues distributed throughout the polypeptide chain can participate in the response of NPC1 to SAHA (Fig. 4c, cyan to blue). These changes are particularly evident in residues involved in the predicted path of cholesterol flow (Fig. 4d). The dramatic improvement of



**Fig. 3** Chol-phenotype landscape and Chol-structure in the absence and presence of SAHA. **a** NPC1 variants are positioned by their variant sequence position (VarSeqP) (*x*-axis) normalized to the full-length sequence and the measured TrIdx values of NPC1 variants (*y*-axis) in the absence (left panel) or presence (right panel) of SAHA. The projected *z*-axis (color gradient) is defined by the measured Chol of each variant normalized to I1061T. The spatial relationships of pairwise comparisons of variants are illustrated as black lines. **b** The molecular variogram illustrating the spatial variance of measured Chol for distance values defined by VarSeqP and TrIdx in the absence (black line) or presence (red line) of SAHA (see Methods; Supplementary Fig. 7a). **c** The VSP predicted Chol-phenotype landscape is shown as a heatmap in the absence (left panel) or presence (right panel) of SAHA. Confidence relationships (see Supplementary Fig. 7b) within the molecular variogram range are plotted as contour lines with the top 25% confidence quartile shown as a bold line in control condition (left panel) and top 5% confidence level shown as bold in SAHA condition (right panel). Color-scale values: severe Chol accumulation (red); WT Chol accumulation (green). I1061T ER-restricted variant is bolded. Dashed ovals 3 and 4 highlight two SCV clusters defined by the top 25% quartile that have good TrIdx but defective Chol. **d** Mapping the Chol-phenotype landscape on the NPC1 structure<sup>25,29</sup>. The highest confident prediction of the Chol value and the corresponding TrIdx value is assigned for each residue with ball color representing Chol, ball size representing TrIdx, and transparency representing prediction confidence. SCV clusters 3 and 4 are selectively highlighted as balls in cartoon presentation of the structure. The proposed cholesterol flow path is shown in gray arrow. **e** Transmembrane region is shown in side view as cylinders (upper panel) or in top view as loops (lower panel) colored by predicted Chol value. Transmembrane helices 1–13 are labeled. The variants in SCV clusters 3 and 4 are labeled and shown in spheres (upper panel) or with C-alpha shown in balls (lower panel)

cholesterol homeostasis for TM3–4 in STMD4 and TM9–13 in CTMD6 (Fig. 4d, cyan and blue regions) is achieved without significant improvement of TrIdx (Fig. 4b, yellow regions), indicating that the dynamics of these TM helices in response to SAHA contribute uniquely to the flow of cholesterol in the LE/Ly.

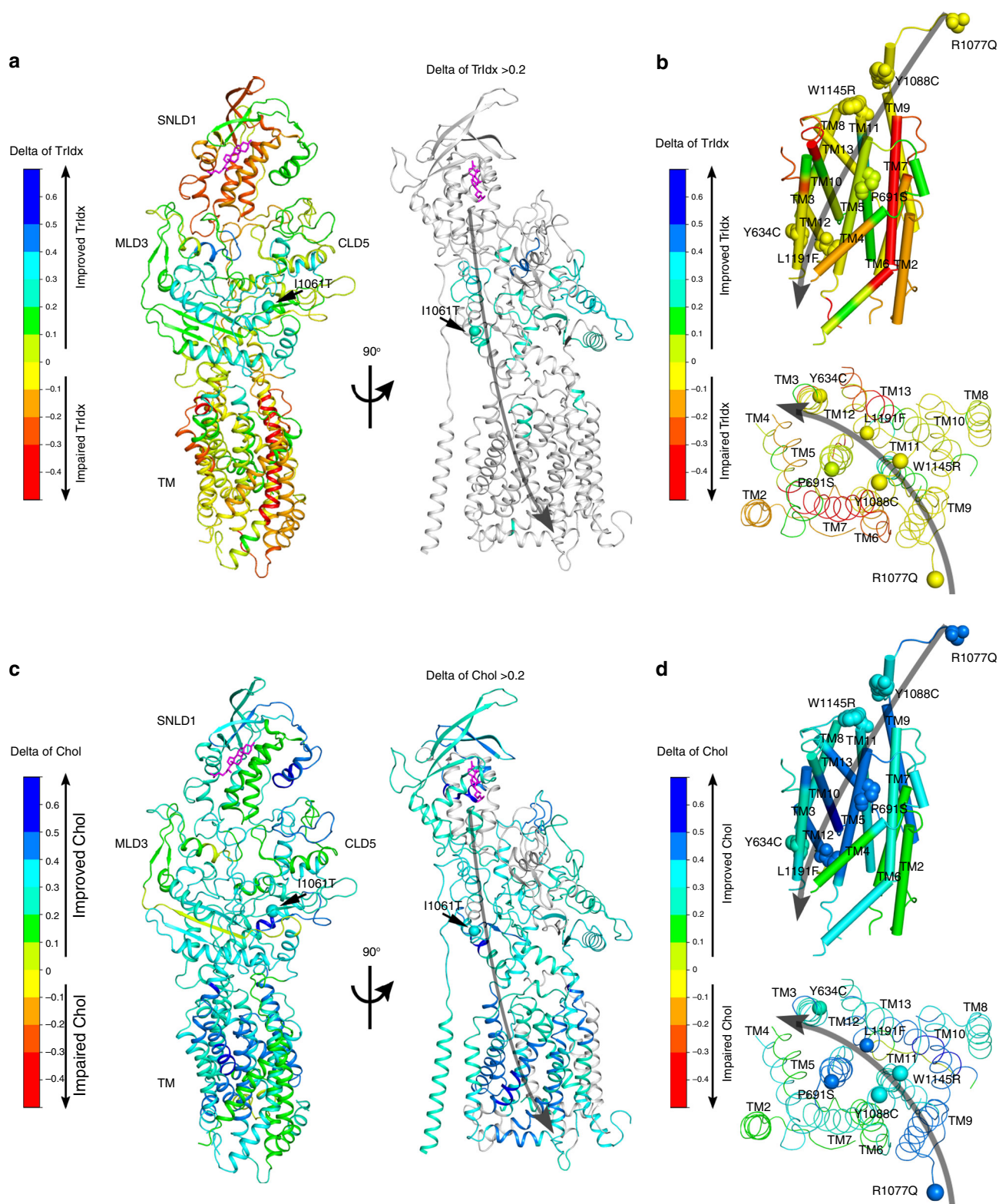
**Mechanism of action of SAHA.** HDACi have been shown to modulate the post-translational acetylation state of numerous cellular proteins<sup>5,8,36</sup>, thereby altering their functional states impacting many cellular pathways including altered stability and targeting for degradation<sup>21,24,37–40</sup>. To examine the impact of HDACi on the acetylation of NPC1, we immunoprecipitated cellular acetylated proteins with an acetylated-lysine (AcK) antibody in the absence or presence of SAHA and probed for the presence of acetylated NPC1. Intriguingly, SAHA promoted hyperacetylation for both WT-NPC1 and I1061T-NPC1 (Fig. 5a), consistent with our recent observations that hyperacetylation of I1061T-NPC1 variant by treatment with the low affinity (mM) HDACi valproic acid (VPA), that targets similar Zn<sup>2+</sup>-dependent HDACs to SAHA, prevents the proteosomal degradation of I1061T-NPC1<sup>41</sup>. To explore this result, we analyzed the heat shock response (HSR) pathway<sup>42,43</sup>, under the control of the heat shock factor 1 (Hsf1) transcription factor. We have previously shown that the chronic expression of misfolded proteins, such as I1061T-NPC1, results in the sustained activation of the HSR termed the maladaptive stress response (MSR)<sup>44</sup>, and that silencing Hsf1 promotes the stability and trafficking of I1061T-NPC1<sup>44</sup>. We observed that SAHA induced a significant reduction in the level of both Hsf1 and Hsf1-P (Fig. 5b) indicating a reduced MSR. Furthermore, the cellular expression of BAGs 1–3, nucleotide exchange factors regulating the activity of heat shock protein 70 (Hsp70) induced by HSR are also significantly reduced by SAHA (Fig. 5b). These results are consistent with the observation that siRNA-mediated silencing of BAGs or chemical disruption of BAG interactions using JG-98, an Hsp70 allosteric regulator<sup>45</sup>, can correct the trafficking defect associated with multiple NPC1 disease-causing variants<sup>11</sup>. Moreover, we found that SAHA treatment results in a significant reduction of HDAC7, a principle HDAC that mediates correction of NPC1<sup>41</sup>, as well as other protein misfolding diseases such as cystic fibrosis<sup>38,46</sup> and alpha-1 antitrypsin deficiency<sup>39,40,47</sup>.

**SCV relationships driving clinical presentation of disease.** To understand the role of epigenetic-sensitive SCV relationships defined by our bench-based experimental measurements to those observed in the clinic, we correlated the variant TrIdx value based on acquisition of Endo H<sup>R</sup> with the natural history of 27 NPC1

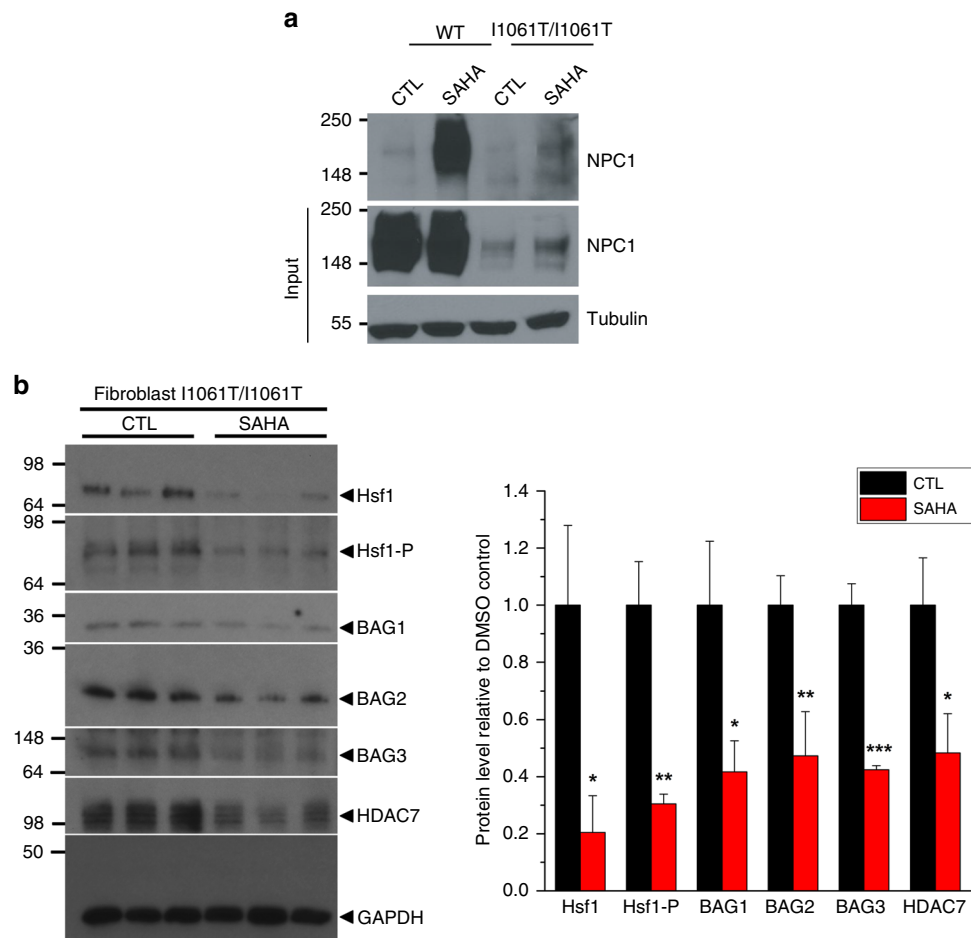
patients that overlap with the input variant dataset (Supplementary Fig. 8; Supplementary Table 3). While there is no significant correlation between trafficking and severity or age-adjusted severity of disease presentation by all NPC1 patients (Supplementary Fig. 8a, b), we found that class III variants containing patients have a significant correlation with a late age of neurological onset (ANO) when compared with all other patients (Supplementary Fig. 8c, middle panel). To map these epigenetic-sensitive SCV relationships that direct ANO, we used the TrIdx as the *y*-axis coordinate to predict the phenotype landscape for the ANO (*z*-axis) (Fig. 6a; Supplementary Fig. 8d). Strikingly, we found in the ANO-phenotype landscape a prominent region defined by a SCV cluster in CLD5 with class III TrIdx properties that shows a significant late ANO (Fig. 6a, b, dashed oval, cyan-blue), likely due to the ability of these residues to maintain a higher ratio (TrIdx) of post-ER functional protein. The known variants that contributed to this predicted age-dependent SCV cluster (V950M, S954L, P1007A, and T1036M) are highly responsive to SAHA treatment. These results illustrate how VSP can be used to quantitate SCV set-point tolerance on a residue-by-residue basis to predict strong candidates for clinical trials (Fig. 6c, ~70% percentile). Moreover, because SAHA also improves nearly all class II CLD5 variants to a class III phenotype with the resultant improvement in cholesterol management (Fig. 6d, arrow), VSP predicts that improving even class II CLD5 variants (e.g., I1061T) to a class III trafficking by a folding corrector, as has been found for CFTR<sup>48</sup>, through either a small molecule chemical chaperone or a proteostasis interventional strategy, may significantly reduce disease progression and reduce the impact of disease in early onset patients.

## Discussion

We have proposed<sup>1</sup> that natural variants distributed across the world-wide population comprise a library of evolutionary tuned information that can be used through GPR-ML based SCV to define the functional properties of the protein fold crucial for human health. We have now shown that epigenetic modulation of NPC1 variants reveal distinct regions of SCV relationships that clearly separate residues involved in trafficking from those responsible for cholesterol homeostasis in downstream LE/Ly. SCV relationships revealed that CLD5 is likely a pivotal module where it forms a biological handshake with MLD3 to direct trafficking from the ER. Given that residues in the STMD4–CTMD6 module have less impact on trafficking from the ER, the genome based phenotype landscapes and corresponding functional structures lead us to suggest that the CLD5 may largely contribute to the organization of residues in the transmembrane spanning STMD4–CTMD6 modules to serve as a tunnel to



**Fig. 4** The SAHA responsive structures for Trlrx and Chol. NPC1 structure is colored by the predicted delta value of Trlrx (**a, b**) or delta value of Chol (**c, d**) in response to SAHA. Color-scale values: impair of Trlrx or Chol homeostasis (red-orange), no change (yellow), improve of Trlrx or Chol homeostasis (cyan-blue). In the right panel of (**a**) and (**c**), regions with delta value < 0.2 are shown in gray to highlight the regions corrected by HDACi (i.e., delta value > 0.2). Transmembrane helices 1-13 are labeled in panels (**b**) and (**d**). Variants critical for predicted cholesterol flow path are highlighted and labeled in panels (**b**) and (**d**)



**Fig. 5** Hyperacetylation of NPC1 and proteostasis impact by SAHA. **a** SAHA leads hyperacetylation of both WT and I1061T-NPC1. Cell lysate of WT or I1061T/I1061T patient fibroblast was immunoprecipitated by acetylated-lysine (AcK) antibody and then recognized by NPC1 antibody through western blot. **b** SAHA impacts the expression of proteostasis components. Immunoblot analysis of SAHA-treated (10  $\mu$ M for 48 h) I1061T/I1061T homozygous fibroblast (left panel) and quantification of total Hsf1, Hsf1 phosphorylated (Hsf1-P), BAG1, BAG2, BAG3, and HDAC7 (right panel) are shown. GAPDH was used as Western blot loading control. Data is presented as fold change to DMSO treatment (mean  $\pm$  s.d.,  $n = 3$ ).  $P$ -values are indicated using student's two tailed  $t$ -test (\* $p < 0.05$ , \*\* $p < 0.01$ , \*\*\* $p < 0.001$ )

channel the transfer of cholesterol from SNLD1 through the bilayer following transfer to the LE/Ly<sup>25,27,49</sup>. A molecular tunnel channeling cholesterol is supported by recent cryo-EM structures of a NPC1 homologous protein, Hedgehog Receptor Patched<sup>50–53</sup> and a molecular dynamics simulation study of NPC1<sup>54</sup>.

Strikingly, many variants contributing to NPC1 disease do not exhibit severe trafficking defects yet are critical for function in downstream LE/Ly, consistent with the notion that export from the ER is not about quality control—rather about governing the SCV tolerance value of a set-point assign to each residue in the fold that we now show can be sensitive to epigenetic control. This conclusion is consistent with the many observations that ~30% of regions contributing to the protein fold are not rigidly ordered, rather are highly dynamic in design, in many cases requiring a cyclic interaction with multiple downstream partners or ligands binding of cholesterol in an acidic LE/Ly in the case NPC1, to achieve function. Mechanistically, we suggest that GPR-ML reveals the potential for amino acid residues in the polypeptide chain to individually or as a collective, contribute to fold management for function through a weighted epigenetic-sensitive SCV matrix that can be discovered using GPR-ML<sup>1</sup>. This matrix response is likely sensitive to proteostasis buffering folding pathways, to UPS or autophagic pathways promoting

degradation<sup>55</sup>, or to HSR management of post-ER stability and function reflecting each residue's role in NPC1 function.

Mechanistically, whether HDACi sensitive events are due to direct alteration of the acetylation balance of Lys residues in the NPC1 polypeptide chain<sup>41,56</sup>, more indirectly through transcriptional and/or post-translational mechanisms affecting HDAC-sensitive proteostasis pathways as shown herein<sup>24,43,57,58</sup>, and/or other HDAC-sensitive events facilitating endomembrane trafficking (ER-Golgi-LE/Ly) compartment function<sup>24,59</sup>, remains to be determined. Recent controversial results<sup>60–64</sup> of HDACi effects in mouse models emphasize the need for use of the human genome sequence and patient related models as captured herein, to point the way<sup>1</sup>, particularly in response to epigenetic modifications that are sensitive to the evolutionary trajectory and the local environment. Consistent with this prediction, from a natural history perspective, we have shown that epigenetic-sensitive SCV relationships can predict residues that are more likely to be responsive to HDACi in human clinical disease.

The sensitivity of SCV relationships dictating the NPC1 fold in disease to HDAC is generalizable given that HDACi correct a broad range of misfolding disorders that are triggered by genetic variation in the population, including cystic fibrosis<sup>38</sup>, alpha-1-antitrypsin deficiency<sup>39,40,47</sup>, Gaucher's disease<sup>65</sup>, and neurodegenerative



**Fig. 6** VSP links bench to bedside. **a** 19 variants from 27 patients with natural history information (Supplementary Table 3) are used as input value for VSP to generate the ANO-phenotype landscape that uses variant position information (x-axis) and cell-based TrIdx measurement (y-axis) to predict the ANO in the clinic (z-axis, color scale). The SCV cluster with late ANO (cyan-blue) in top 25% quartile of prediction confidence is highlighted by dashed oval. **b** The predicted ANO with highest confidence for each residue is mapped on the structure with the color code the same as panel (a). Variants with late ANO are shown in balls and labeled. **c** Distribution of the delta value of cholesterol homeostasis in response to SAHA for all the variants tested<sup>21</sup>. The variants with late ANO in the SCV cluster are shown in blue and labeled. **d** The Chol-phenotype landscape of CLD5 (Fig. 3c, asterisks) are shown with the TrIdx classes highlighted by dash lines. Black arrow indicates the TrIdx and Chol shift of class II CLD5 variants to class III in response to SAHA. **e** HDACi remodels the SCV relationships to manage genotype to phenotype transformation. Addition of HDACi to the same collection of variants from NPC1 population creates a new SCV matrix linking genotype to phenotype. VSP captures the impact of these epigenetic changes through SCV matrices to interpret and predict the dynamics of protein fold for its function in the context of local epigenetic environment. The epigenetic modulation of SCV (epi-SCV in the figure) provides a quantitative platform to characterize the ability of the environment to manage the information flow through central dogma (epi-SCV[DNA|RNA|Protein])

are controlled by the histone deacetylase activity. As such, VSP provides us with a quantitative platform to not only mechanistically understand the impact of genetic variation on function and structure<sup>1</sup>, but as shown herein, its epigenetic plasticity. The striking remodeling effect of HDACi on SCV relationships linking genomics to phenomics in the NPC1 disease population reveals a potential role for the acetylation sensitive epigenome and epiproteome to (re)tune global flow of information from the genome to the functional proteome (Fig. 6e). We have previously proposed<sup>1</sup> that information flow through central dogma can be best captured by changing our current linear (DNA → RNA → Protein) perspective to a SCV-based matrix view (Fig. 6e, SCV [DNA|RNA|Protein] where the vertical lines are defined by GPR-ML matrices that capture the SCV-based information flow through central dogma that generates biology. We now suggest that GPR-ML describes the epigenetic-sensitive SCV matrices (Fig. 6e, epi-SCV[DNA|RNA|Protein]) to reveal the environment sensitive complexity in the individual that make each of us who we are in health and disease<sup>1</sup> over a lifespan (Fig. 6e).

## Methods

**Reagents.** Modified Eagle's medium (MEM), fetal bovine serum (FBS), penicillin-streptomycin (P/S), and geneticin (G418) were purchased from Invitrogen Life Technologies Corporation (Carlsbad, CA). Effectene Transfection Reagent Kit and DNA purification kit were purchased from QIAGEN Inc. (Valencia, CA). Histone deacetylase inhibitors (SAHA) were purchased from Cayman Chemical Co. SAHA was stocked at 5 mM in DMSO and stored at -20 °C. Endoglycosidase H was purchased from New England Biolabs. A rat monoclonal anti-hNPC1 antibody was developed against the C-terminal peptide (amino acids 1261–1278). Other antibodies used in the study are acetylated-lysine (AcK) antibody (AB193, Abcam), HDAC7 (Ab12174, Abcam), HSF1 (Ab52757, Abcam), HSF1 phosphorylated (Ab76076, Abcam), Bag1 (3.10G3E2, Santa Cruz Biotechnology), Bag2 (Ab79406, Abcam), Bag3 (Ab5898, Abcam), and GAPDH (AM4300, Invitrogen). All other chemicals were purchased from Sigma Chemical (St. Louis, MO).

**Human fibroblast cells.** Human NPC1 fibroblasts were obtained from Coriell Institute (Camden, NJ) (Supplementary Table 1). All human skin fibroblasts were maintained in modified Eagle medium containing FBS.

**Stable silencing of NPC1.** NPC1-deficient stable HeLa-shNPC1 and U2OS-shNPC1 cells were generated by silencing the endogenous npc1 gene with 3'-UTR shNPC1 lentivirus. Mission shRNA clone was purchased from Sigma against 3' UTR (TRCN00000542) of human NPC1 (pLKO-shNPC1) for knockdown of npc1 gene in human cell lines. Stable clones of npc1-deficient cells were selected with 5 µg/ml puromycin for 3 weeks. Stable HeLa-shNPC1 cells were cultured in DMEM medium supplemented with 10% FBS and 5 µg/ml puromycin. Stable U2OS-shNPC1 cells were cultured in McCoy's 5A medium supplemented with 10% FBS, 50-units/ml penicillin, 50 µg/ml streptomycin, 5 µg/ml puromycin, and 1 mg/ml G418.

**NPC1 variant transient expression vectors.** cDNA encoding human ΔU3hNPC1-WT construct was kindly provided by Dan Ory (Washington University, St. Louis, MO). The NPC1 gene was subcloned into bicistronic retroviral plasmid, pMIEG3 using EcoRI and NotI restriction enzyme sites. pMIEG3 vector was generated from pMSCVneo vector (Clontech) in which the murine phosphoglycerate kinase (PKG) promoter and the Neomycin resistance (Neo<sup>r</sup>) genes were replaced by IRES (Internal Ribosome Entry Site) and EGFP (enhanced green fluorescent protein) genes. ΔU3hNPC1-WT construct has four substitutions

compared with a standard reference NPC1 sequence (NM\_000271.4, Uniprot ID: O15118), so this WT sequence is named as WT-V<sup>21</sup>. The substitutions are 387T>C (Y129Y), 1415T>C (L472P), 1925T>C (M642T), and 2587T>C (S863P). ΔU3hNPC1-WT has been used as a WT construct in precious publications<sup>21,75</sup>. It was derived from pSV-SPORT/NPC1 that has been used as the WT control protein in previous publications<sup>56,76,77</sup>. The site-mutagenesis was generated by Quick-Change XL Site-directed Mutagenesis Kit (Stratagene, La Jolla, CA) using pMIEG3-NPC1 as template, and all the variants used in this study contain the four background variants derived from human ΔU3hNPC1-WT (i.e., WT-V) construct.

**Endoglycosidase H (Endo H) assay.** In HeLa cells, NPC1 variants were transiently expressed for 72 h using Fugene transfection reagents. Cells were washed with 1× PBS twice and lysed the cells with RIPA lysis buffer (10 mM Tris pH 8.0, 140 mM NaCl, 1% NP-40, 0.1% sodium-deoxycholate, 0.1% SDS, 1× protease inhibitor cocktail (PIC), 1 mM PMSF) on ice for 30 min. Cell lysates were obtained by 16,000 × g centrifugation at 4 °C. Protein concentration was measured using the standard BCA assay. Around ~300–500 mg of cell lysates were incubated with 1–2 mg of rabbit polyclonal anti-NPC1 antibody or rat monoclonal anti-NPC1 antibody for overnight at 4 °C. In total, 20–30 µl of protein A/G-Sepharose or GammaBind G-Sepharose beads (GE Healthcare) were added to the cell lysates to capture the antibody-bound NPC1 for 1 h at 4 °C. Beads were washed with RIPA lysis buffer 3× with final wash with 1× PBS. Immunoprecipitated NPC1 was eluted with 1× denaturation buffer (0.5% SDS, 40 mM DTT) at 95 °C for 10 min. Eluted NPC1 was divided into equal parts and incubated with absence (–) or absence (+) of 1000 units Endo H enzyme for overnight at 37 °C. Endo H digested samples were subjected to 4–12% Bis-Tris gradient (Invitrogen) or 4–20% gradient (Bio-Rad) SDS-PAGE and immunoblotted with rat monoclonal anti-NPC1 antibody (1:3000). Endo H sensitive (Endo H<sup>S</sup>) and Endo H resistant (Endo H<sup>R</sup>) were identified by their unique migration on SDS-PAGE gels. In U2OS cells, cells were seeded at 2.0 × 10<sup>5</sup> cells/ml in 12-well plates and incubated overnight. Transfections were performed using the reagent FuGENE6 from PROMEGA (Madison, WI) with a 1:4 ratio, 2 µg DNA:8 µl FuGENE6 in Opti-MEM (1×) + Hepes + sodium bicarbonate + L-glutamine + 5% FBS. Samples were then incubated for 5 h at 37 °C and the medium was replaced with normal growth medium. Cells were then incubated for another 48 h and visualized under the fluorescent microscope for GFP-positive transfected cells. These variants were all transiently expressed in the U2OS-SRA-shNPC1 cells. After 72 h from initial plating, cells were lysed with 50 µl/well of 1× RIPA (150 mM NaCl, 1.0% IGEPAL Ca-360, 0.5% Na-deoxychlorate, 0.1% SDS, 50 mM Tris pH 8.0, 1× protease inhibitors, 1× benzonase) on ice for 30 min. Samples were collected and incubated at 37 °C for 20 min, centrifuged at 20,817 × g at 4 °C for 20 min. Supernatant was collected and the BCA assay performed. Endo H was performed by addition of 1× glycoprotein denaturing buffer (New England Biolabs) to 20 µg of lysate and incubated at 50 °C for 30 min. For each sample, 10,000 U/ml Endo H and 1× Glycobuffer (New England Biolabs) were added and incubated overnight at 37 °C. Protein samples were separated on 4–12% Bis-Tris BOLT protein gel (Invitrogen) and transferred to nitrocellulose membrane. Membranes were probed with NPC1 1:1000 and tubulin 1:25,000 diluted in 5% milk solution.

**SAHA treatment on NPC1 variants.** Human NPC1 variants were transiently expressed in HeLa-shNPC1 cells. After 24 h posttransfection, cells were treated with control DMSO or 10 µM SAHA for 72 h. Cell lysates were processed and Endo H treatment of the NPC1 variants were performed as described above.

**Probing proteostasis components in response to SAHA.** Patient I1061T/I1061T fibroblasts were seeded in a 6-well plate and grown until confluent. Once confluent, cells were treated with DMSO or SAHA at 10 µM for 48 h at 37 °C in normal growth media. Cells were then washed twice with 1× PBS and then lysed with 75 µl/well of 1× RIPA (150 mM NaCl, 1.0% IGEPAL Ca-360, 0.5% Na-deoxychlorate, 0.1% SDS, 50 mM Tris pH 8.0, 1× protease inhibitors, 1× benzonase) on ice for 30 min. Samples were collected and then incubated at 37 °C for 20 min. Samples were then centrifuged at 20,817 × g at 4 °C for 20 min. Supernatant was

collected and BCA assay was performed. Once samples were prepared, 40 µg of total protein was separated on a 10% SDS-PAGE. Proteins were transferred to nitrocellulose and probed with antibodies of HSF1 (1:10,000 dilution), HSF1 phosphorylated (1:5000 dilution), Bag1 (1:1000 dilution), Bag2 (1:1000 dilution), and Bag3 (1:1000 dilution), and HDAC7 (1:1000 dilution). GAPDH is probed by GAPDH antibody (1:100,000 dilution) as loading control. Each was further probed with their specific secondary antibodies and detected by chemiluminescence.

**Natural history analysis.** NPC1 subjects were evaluated at the National Institutes of Health Clinical Center in Bethesda, Maryland in an observational/natural history protocol (NCT00344331; <https://clinicaltrials.gov/ct2/show/NCT00344331?term=niemann+pick&rank=11>). This study was approved by the Eunice Kennedy Shriver National Institute of Child Health and Human Development Institutional Review Board. Written consent, and when feasible assent, were obtained. Clinical information related to disease onset was collected from guardians and record review. Neurological severity scores were determined by clinical evaluation of nine major and eight minor clinical areas, such as eye movement, ambulation, speech, swallow, fine motor skills, cognition, hearing, memory, seizures and so on<sup>78</sup>. Age-adjusted severity scores were determined by dividing the neurological severity scores by the age of the subject at the time of evaluation. The age of neurologic onset of each variant is calculated by averaging the age of neurologic onset of the patients (Supplementary Table 3) who contain that variant.

**VSP methods.** VSP uses a biological adaptation of the general principles of geostatistics<sup>1,9</sup>, a well-established Gaussian-process regression (GPR)-based machine-learning (ML) approach that provides a means to predict the distribution of a large range of geological, epidemiological, anthropological and environmental measures in complex geologic landscapes using sparse sampling methods. Below we described how we apply classical GPR-based ML concepts through VSP to understand the role of variation in human disease and its application to precision medicine.

The initial step for VSP is molecular variogram analysis. Assume we have a variable  $z$  ( $z$ -axis value) which is positioned by  $x$  and  $y$  coordinates ( $x$ - and  $y$ -axis values) that describe the phenotype landscape. A molecular variogram is first used to describe how the spatial variance (i.e., the degree of dissimilarity) of  $z$  changes according to the separation distance defined by the  $x$  and  $y$  coordinates which enables the calculation of the spatial covariance (SCV) relationships in the dataset. Suppose the  $i$ th (or  $j$ th) observation in a dataset consists of a value  $z_i$  (or  $z_j$ ) at coordinates  $x_i$  (or  $x_j$ ) and  $y_i$  (or  $y_j$ ). The distance  $h$  between the  $i$ th and  $j$ th observation is calculated by

$$h_{(i,j)} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \quad (1)$$

and the  $\gamma(h)$ -variance for a given distance ( $h$ ) is defined by

$$\gamma(h) = \frac{1}{2}(z_i - z_j)^2 \quad (2)$$

$\gamma(h)$ -variance is the semivariance of  $z$  value between the two observations, which is also the whole variance of  $z$  value for one observation at the given separation distance  $h$ . Here in the paper, we refer  $\gamma(h)$ -variance as spatial variance as indicated in the  $y$ -axis of molecular variogram (main text Figs. 2b and 3b). By Eqs. (1) and (2), the distance ( $h$ ) and  $\gamma(h)$ -variance for all the data pairs are generated. Then, the average values of  $\gamma(h)$ -variance for different distance intervals are calculated to plot  $\gamma(h)$  versus  $h$  used in the molecular variogram. The distance where the model first flattens out is known as the range. Sample locations separated by distances closer than the range are spatially correlated, whereas locations farther apart than the range are not. The spatial covariance (SCV) at the distance ( $h$ ) is calculated by  $C(h) = C(0) - \gamma(h)$ , where  $C(0)$  is the covariance at zero distance representing the global variance of the data points under consideration (i.e., the plateau of the variogram).

GPR-based ML is an interpolation/regression method providing optimal unbiased prediction based on the modeled SCV relationships generated by the molecular variogram<sup>9</sup>. In our study, we used a GPR-based regression approach referred to as ordinary Kriging that has the least assumptions and is the form most commonly used<sup>79</sup>. Essentially, SCV can predict the unknown value by local weighted averaging the surrounding known values, where the weight associated with the known value is determined according to their positions both in relation to the unknown point and to one another.

According to variogram, observations within close distance are usually highly correlated and have more weight for prediction. To solve the optimum and unbiased weights of SCV relationships, molecular geostatistics aims to minimize the variance associated with the prediction of the unknown value at location  $u$ , which is generated according to the expression

$$\sigma_u^2 = E[(z_u^* - z_u)^2] = \sum_{i=1}^n \sum_{j=1}^n \omega_i \omega_j C_{i,j} - 2 \sum_{i=1}^n \omega_i C_{i,u} + C_{u,u} \quad (3)$$

where  $z_u^*$  is the prediction value while  $z_u$  is the true but unknown value,  $C_{i,j}$  and  $C_{i,u}$  are SCV between data points  $i$  and  $j$ , and data points  $i$  and  $u$ , respectively, and  $C_{u,u}$  is the SCV within location  $u$ .  $\omega_i$  is the weight for data point  $i$ . The SCV is obtained from the above molecular variogram analysis. To ensure an unbiased result, the

sum of weight is set as one.

$$\sum_{i=1}^n \omega_i = 1 \quad (4)$$

Equations (3) and (4) not only solved the set of weights associated with input observations, but also provided the minimized ‘molecular Kriging variance’ at location  $u$  which can be expressed as

$$\sigma_u^2 = C_{u,u} - \left( \sum_{i=1}^n \omega_i C_{i,u} + \mu \right) \quad (5)$$

where  $C_{u,u}$  is the SCV within location  $u$ ,  $\omega_i$  is the weight for data point  $i$ , and  $C_{i,u}$  are SCV between data points  $i$  and  $u$ .  $\mu$  is the Lagrange Parameter that is used to convert the constrained minimization problem in Eq. (3) into an unconstrained one. The resulting minimized Kriging variance provides a weighted SCV score that represents the confidence for using the SCV relationships both within the input data points and in relation to the unknown locations. The confidence level is tightly linked with the distance range in the variogram. The shorter distance between the unknown point to the input data points, the higher confidence for using the SCV relationships for prediction.

With the solved weights  $W$ , we can calculate the prediction of all unknown values to generate the complete phenotype landscape by the equation

$$z_u^* = \sum_{i=1}^n \omega_i z_i \quad (6)$$

where  $z_u^*$  is the prediction value for the unknown data point  $u$ ,  $\omega_i$  is the weight for the known data point, and  $z_i$  is the measured value for data point  $i$ <sup>9</sup>.

The parameters, for example, range and sill (i.e., plateau) of the variogram in this study are indicated in Fig. 2b and Fig. 3b, respectively. They were chosen by minimizing the residual sums of squares of the fitting. Spherical model was chosen for variogram modeling because it yielded the best leave-one-out cross-validation result for the datasets when compared with linear, exponential and gaussian models. In the final prediction, for each unknown position, minimal 5 nearby variants and maximal 19–25 nearby variants were included in the weighted averaging calculation (i.e., Eq. (6)). All the procedures were performed in the ordinary Kriging module in GS+, version 10 from gammadesign software (<https://geostatistics.com/>). We also tried the ordinary Kriging module in *gstat* package<sup>80</sup>, which gave identical results. The source code for running *gstat* in R is uploaded to <https://doi.org/10.17632/ycw667nv5f.1> (see Code availability). The multivariate linear regression and random forest regression were performed in R by using basic *lm* function or *randomForest* package, respectively.

We used leave-one-out and  $k$ -fold cross-validation to assess the performance of VSP strategy. In the leave-one-out cross-validation all data are initially used to build the molecular variogram and Kriging models. Then, we remove each data point, one at a time and use the rest of the data points to predict the missing value. We repeat the prediction for all data points and compare the prediction results with the measured value to generate the Pearson’s  $r$ -value and its associated  $p$ -value (ANOVA test). For the  $k$ -fold cross-validation, samples are randomly partitioned into  $k = 48, 24, 12, 8, 4, 3$ , or 2 sets. Of the  $k$  sets, a single set is used as validation data and the remaining  $k-1$  sets are used as training data. The size of training and validation subsamples are indicated for each  $k$ -fold in the figures. The cross-validation process is repeated  $k$  times and every set is used as validation once. The prediction of each sample is collected. For  $k < 48$ , the partition process is repeated five times and the averaged Pearson’s  $r$ - and  $p$ -value of the correlation between predicted value and actual value is reported. The error bars associated with each prediction is the prediction confidence. In the correlation analyses, we took the confidence level into account as weight. A prediction with small uncertainty will have a larger weight because it is more precise than prediction with larger uncertainty. The weight is calculated as:  $\omega_i = \frac{1}{\sigma_i^2}$ , where  $\sigma_i^2$  is the variance for  $i$ . Quantitative correlation analyses and  $p$ -value calculations were performed using the software Originpro 2016 or R. A  $p$ -value  $< 0.05$  was considered to indicate statistical significance.

Phenotype landscapes predicted based on a sparse collection of variants contain experimental information that comprises the full range of values on function  $y$ - and  $z$ -axis for the entire polypeptide sequence ( $x$ -axis). To map the function predictions onto structure, we assign the prediction value with highest confidence to each residue to generate a functional structure that illustrates all values interpolated from the sparse collection variants used to generate the phenotype landscape displaying at atomic resolution. PDB:3JD8 and PDB: 5U73 are used for NPC1 structure mapping<sup>25,29</sup>. All the structural presentations were produced by the software of PyMOL.

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

**Code availability**

We included a demo input data, the R-codes for generating phenotype landscape and instructions as a Zip file named ‘‘ExampleCode.Zip’’, which is uploaded to <https://doi.org/10.17632/ycw667nv5f.1>. Additional modified scripts can be accessed upon request.

**Data availability**

The source data underlying Figs. 1–4 are provided as Source Data file. The uncropped blots of Fig. 5b are shown in Supplementary Fig. 9. The patients data in Fig. 6a, b are provided as Supplementary Table 3. The output phenotype landscapes for Figs. 2c, 3c and 6a are uploaded as output files in <https://doi.org/10.17632/ycw667nv5f1>. All other relevant data are available from the corresponding author on request.

Received: 7 November 2018; Accepted: 10 October 2019;

Published online: 07 November 2019

**References**

- Wang, C. & Balch, W. E. Bridging genomics to phenomics at atomic resolution through variation spatial profiling. *Cell Rep.* **24**, 2013–2028 e6 (2018).
- Hindorf, L. A. et al. Prioritizing diversity in human genomics research. *Nat. Rev. Genet.* **19**, 175–185 (2018).
- Torkamani, A. Drilling for insight: forecasting phenotype from genotype. *Trends Genet.* **34**, 821–822 (2018).
- Balch, W. E., Morimoto, R. I., Dillin, A. & Kelly, J. W. Adapting proteostasis for disease intervention. *Science* **319**, 916–919 (2008).
- Ali, I., Conrad, R. J., Verdin, E. & Ott, M. Lysine acetylation goes global: from epigenetics to metabolism and therapeutics. *Chem. Rev.* **118**, 1216–1252 (2018).
- Benayoun, B. A., Pollina, E. A. & Brunet, A. Epigenetic regulation of ageing: linking environmental inputs to genomic stability. *Nat. Rev. Mol. Cell Biol.* **16**, 593–610 (2015).
- Peleg, S., Feller, C., Ladurner, A. G. & Imhof, A. The metabolic impact on histone acetylation and transcription in ageing. *Trends Biochem. Sci.* **41**, 700–711 (2016).
- Li, X., Egervari, G., Wang, Y., Berger, S. L. & Lu, Z. Regulation of chromatin and gene expression by metabolic enzymes and metabolites. *Nat. Rev. Mol. Cell Biol.* **19**, 563 (2018).
- Chilès, J.-P. & Delfiner, P. *Geostatistics: Modeling Spatial Uncertainty*, xv, 699 (John Wiley & Sons, Hoboken, NJ, 2012).
- Rasmussen, C. E. & Williams, C. K. I. *Gaussian Processes for Machine Learning* (MIT Press, 2006).
- Wang, C. et al. Individualized Management of Genetic Diversity in Niemann-Pick C1 through Modulation of the Hsp70 Chaperone System. *Hum. Mol. Genet.* <https://doi.org/10.1093/hmg/ddz215> (2019).
- Vanier, M. T. Complex lipid trafficking in Niemann-Pick disease type C. *J. Inherit. Metab. Dis.* **38**, 187–199 (2015).
- NPC-db2. NPC-db2: Niemann-Pick Type C Database. <https://medgen.medizin.uni-tuebingen.de/NPC-db2/index.php> (2017).
- Bianconi, S. E. et al. Evaluation of age of death in Niemann-Pick disease, type C: Utility of disease support group websites to understand natural history. *Mol. Genet. Metab.* **126**, 466–469 (2019).
- Wassif, C. A. et al. High incidence of unrecognized visceral/neurological late-onset Niemann-Pick disease, type C1, predicted by analysis of massively parallel sequencing data sets. *Genet. Med.* **18**, 41–48 (2016).
- Stampfer, M. et al. Niemann-Pick disease type C clinical database: cognitive and coordination deficits are early disease indicators. *Orphanet. J. Rare. Dis.* **8**, 35 (2013).
- Tangemo, C., Weber, D., Theiss, S., Mengel, E. & Runz, H. Niemann-Pick Type C disease: characterizing lipid levels in patients with variant lysosomal cholesterol storage. *J. Lipid Res.* **52**, 813–825 (2011).
- Vanier, M. T. Niemann-Pick diseases. *Handb. Clin. Neurol.* **113**, 1717–1721 (2013).
- Platt, F. M. et al. Disorders of cholesterol metabolism and their unanticipated convergent mechanisms of disease. *Annu. Rev. Genomics Hum. Genet.* **15**, 173–194 (2014).
- Gelsthorpe, M. E. et al. Niemann-Pick type C1 I1061T mutant encodes a functional protein that is selected for endoplasmic reticulum-associated degradation due to protein misfolding. *J. Biol. Chem.* **283**, 8229–8236 (2008).
- Pipalia, N. H. et al. Histone deacetylase inhibitors correct the cholesterol storage defect in most Niemann-Pick C1 mutant cells. *J. Lipid Res.* **58**, 695–708 (2017).
- Pipalia, N. H., Huang, A., Ralph, H., Rujoi, M. & Maxfield, F. R. Automated microscopy screening for compounds that partially revert cholesterol accumulation in Niemann-Pick C cells. *J. Lipid Res.* **47**, 284–301 (2006).
- Pipalia, N. H. et al. Histone deacetylase inhibitor treatment dramatically reduces cholesterol accumulation in Niemann-Pick type C1 mutant human fibroblasts. *Proc. Natl Acad. Sci. USA* **108**, 5620–5625 (2011).
- Subramanian, K., Rauniyar, N., Lavalley-Adam, M., Yates, J. R. 3rd & Balch, W. E. Quantitative analysis of the proteome response to the histone deacetylase inhibitor (HDACi) vorinostat in Niemann-Pick Type C1 disease. *Mol. Cell. Proteom.* **16**, 1938–1957 (2017).
- Gong, X. et al. Structural insights into the Niemann-Pick C1 (NPC1)-mediated cholesterol transfer and Ebola infection. *Cell* **165**, 1467–1478 (2016).
- Li, X., Saha, P., Li, J., Blobel, G. & Pfeffer, S. R. Clues to the mechanism of cholesterol transfer from the structure of NPC1 middle luminal domain bound to NPC2. *Proc. Natl Acad. Sci. USA* **113**, 10079–10084 (2016).
- Li, X. et al. Structure of human Niemann-Pick C1 protein. *Proc. Natl. Acad. Sci. USA* **113**, 8212–8217 (2016).
- Li, J. & Pfeffer, S. R. Lysosomal membrane glycoproteins bind cholesterol and contribute to lysosomal cholesterol export. *Elife* **5** (2016).
- Li, X. et al. 3.3 Å structure of Niemann-Pick C1 protein reveals insights into the function of the C-terminal luminal domain in cholesterol transport. *Proc. Natl Acad. Sci. USA* **114**, 9116–9121 (2017).
- Lopez, M. E. & Scott, M. P. Genetic dissection of a cell-autonomous neurodegenerative disorder: lessons learned from mouse models of Niemann-Pick disease type C. *Dis. Model Mech.* **6**, 1089–1100 (2013).
- King, G. & Sharom, F. J. Proteins that bind and move lipids: MsbA and NPC1. *Crit. Rev. Biochem. Mol. Biol.* **47**, 75–95 (2012).
- Rosenbaum, A. I. & Maxfield, F. R. Niemann-Pick type C disease: molecular mechanisms and potential therapeutic approaches. *J. Neurochem.* **116**, 789–795 (2011).
- Ohgami, N. et al. Binding between the Niemann-Pick C1 protein and a photoactivatable cholesterol analog requires a functional sterol-sensing domain. *Proc. Natl Acad. Sci. USA* **101**, 12473–12478 (2004).
- Lu, F. et al. Identification of NPC1 as the target of U18666A, an inhibitor of lysosomal cholesterol export and Ebola infection. *Elife* **4** (2015).
- Kwon, H. J. et al. Structure of N-terminal domain of NPC1 reveals distinct subdomains for binding and transfer of cholesterol. *Cell* **137**, 1213–1224 (2009).
- Choudhary, C. et al. Lysine acetylation targets protein complexes and co-regulates major cellular functions. *Science* **325**, 834–840 (2009).
- Min, S. W. et al. Acetylation of tau inhibits its degradation and contributes to tauopathy. *Neuron* **67**, 953–966 (2010).
- Hutt, D. M. et al. Reduced histone deacetylase 7 activity restores function to misfolded CFTR in cystic fibrosis. *Nat. Chem. Biol.* **6**, 25–33 (2010).
- Bouchecareilh, M., Hutt, D. M., Szajner, P., Flotte, T. R. & Balch, W. E. Histone deacetylase inhibitor (HDACi) suberoylanilide hydroxamic acid (SAHA)-mediated correction of alpha-1 antitrypsin deficiency. *J. Biol. Chem.* **287**, 38265–38278 (2012).
- Wang, C., Bouchecareilh, M. & Balch, W. E. Measuring the effect of histone deacetylase inhibitors (HDACi) on the secretion and activity of alpha-1 antitrypsin. *Methods Mol. Biol.* **1639**, 185–193 (2017).
- Subramanian, K., Hutt, D. M., Gupta, V., Mao, S. & Balch, W. E. Correction of Niemann-Pick type C1 disease with the histone deacetylase inhibitor valproic acid. Preprint at *bioRxiv*. <https://doi.org/10.1101/724187> (2019).
- Li, J., Labbadia, J. & Morimoto, R. I. Rethinking HSF1 in stress, development, and organismal health. *Trends Cell Biol.* **27**, 895–905 (2017).
- Westerheide, S. D., Anckar, J., Stevens, S. M. Jr., Sistonen, L. & Morimoto, R. I. Stress-inducible regulation of heat shock factor 1 by the deacetylase SIRT1. *Science* **323**, 1063–1066 (2009).
- Roth, D. M. et al. Modulation of the maladaptive stress response to manage diseases of protein folding. *PLoS Biol.* **12**, e1001998 (2014).
- Shao, H. et al. Exploration of benzothiazole rhodacyanines as allosteric inhibitors of protein-protein interactions with heat shock protein 70 (Hsp70). *J. Med. Chem.* **61**, 6163–6177 (2018).
- Angles, F., Hutt, D. M. & Balch, W. E. HDAC inhibitors rescue multiple disease-causing CFTR variants. *Hum. Mol. Genet.* **28**, 1982–2000 (2019).
- Wang, C. & Balch, W. E. in *Alpha-1 Antitrypsin. Respiratory Medicine* (eds Wanner, A. & Sandhaus, R.) (Humana Press, Cham, 2016).
- Rowe, S. M. & Verkman, A. S. Cystic fibrosis transmembrane regulator correctors and potentiators. *Cold Spring Harb. Perspect. Med.* **3**, a009761 (2013).
- Zhao, Y., Ren, J., Harlos, K. & Stuart, D. I. Structure of glycosylated NPC1 luminal domain C reveals insights into NPC2 and Ebola virus interactions. *FEBS Lett.* **590**, 605–612 (2016).
- Zhang, Y. et al. Structural basis for cholesterol transport-like activity of the hedgehog receptor patched. *Cell* **175**, 1352–1364 e14 (2018).
- Qi, X., Schmiege, P., Coutavas, E. & Li, X. Two patched molecules engage distinct sites on Hedgehog yielding a signaling-competent complex. *Science* **362**, eaas8843 (2018).
- Gong, X. et al. Structural basis for the recognition of Sonic Hedgehog by human Patched1. *Science* **361**, eaas8935 (2018).
- Qian, H. et al. Inhibition of tetrameric Patched1 by Sonic Hedgehog through an asymmetric paradigm. *Nat. Commun.* **10**, 2320 (2019).

54. Elghobashi-Meinhardt, N. Computational tools unravel putative sterol binding sites in the lysosomal NPC1 protein. *J. Chem. Inf. Model.* **59**, 2432–2441 (2019).
55. Schultz, M. L. et al. Coordinate regulation of mutant NPC1 degradation by selective ER autophagy and MARCH6-dependent ERAD. *Nat. Commun.* **9**, 3671 (2018).
56. Nakasone, N. et al. Endoplasmic reticulum-associated degradation of Niemann-Pick C1: evidence for the role of heat shock proteins and identification of lysine residues that accept ubiquitin. *J. Biol. Chem.* **289**, 19714–19725 (2014).
57. Bouchecareilh, M. & Balch, W. E. Proteostasis, an emerging therapeutic paradigm for managing inflammatory airway stress disease. *Curr. Mol. Med.* **12**, 815–826 (2012).
58. McDonnell, E., Peterson, B. S., Bomze, H. M. & Hirschey, M. D. SIRT3 regulates progression and development of diseases of aging. *Trends Endocrinol. Metab.* **26**, 486–492 (2015).
59. Zhang, J. et al. Importance of TFEB acetylation in control of its transcriptional activity and lysosomal function in response to histone deacetylase inhibitors. *Autophagy* **14**, 1043–1059 (2018).
60. Praggastis, M. et al. A murine Niemann-Pick C1 I1061T knock-in model recapitulates the pathological features of the most prevalent human disease allele. *J. Neurosci.* **35**, 8091–8106 (2015).
61. Munkacsy, A. B. et al. Normalization of hepatic homeostasis in the Npc1 (nmf164) mouse model of Niemann-Pick Type C disease treated with the histone deacetylase inhibitor vorinostat. *J. Biol. Chem.* **292**, 4395–4410 (2017).
62. Alam, M. S., Cooper, B., Farris, J. D. & Haldar, K. Tolerance of chronic HDACi treatment for neurological, visceral and lung Niemann-Pick Type C disease in mice. *Sci. Rep.* **8**, 3875 (2018).
63. Alam, M. S., Getz, M. & Haldar, K. Chronic administration of an HDAC inhibitor treats both neurological and systemic Niemann-Pick type C disease in a mouse model. *Sci. Transl. Med.* **8**, 326ra23 (2016).
64. Davidson, J. et al. 2-Hydroxypropyl-beta-cyclodextrin is the active component in a triple combination formulation for treatment of Niemann-Pick C1 disease. *Biochim Biophys. Acta Mol. Cell Biol. Lipids* **1864**, 1545–1561 (2019).
65. Lu, J. et al. Histone deacetylase inhibitors prevent the degradation and restore the activity of glucocerebrosidase in Gaucher disease. *Proc. Natl Acad. Sci. USA* **108**, 21200–21205 (2011).
66. Soragni, E. & Gottesfeld, J. M. Translating HDAC inhibitors in Friedreich's ataxia. *Expert Opin. Orphan Drugs* **4**, 961–970 (2016).
67. Coppede, F. The potential of epigenetic therapies in neurodegenerative diseases. *Front. Genet.* **5**, 220 (2014).
68. Jia, H. et al. Histone deacetylase (HDAC) inhibitors targeting HDAC3 and HDAC1 ameliorate polyglutamine-elicited phenotypes in model systems of Huntington's disease. *Neurobiol. Dis.* **46**, 351–361 (2012).
69. Etchegaray, J. P. et al. The histone deacetylase SIRT6 controls embryonic stem cell fate via TET-mediated production of 5-hydroxymethylcytosine. *Nat. Cell Biol.* **17**, 545–557 (2015).
70. Ferrer, C. M. et al. An inactivating mutation in the histone deacetylase SIRT6 causes human perinatal lethality. *Genes Dev.* **32**, 373–388 (2018).
71. Zhang, W. et al. SIRT6 deficiency results in developmental retardation in cynomolgus monkeys. *Nature* **560**, 661–665 (2018).
72. Kugel, S. et al. SIRT6 suppresses pancreatic cancer through control of Lin28b. *Cell* **165**, 1401–1415 (2016).
73. Booth, L. N. & Brunet, A. The aging epigenome. *Mol. Cell* **62**, 728–744 (2016).
74. Ren, R., Ocampo, A., Liu, G. H. & Izpisua Belmonte, J. C. Regulation of stem cell aging by metabolism and epigenetics. *Cell. Metab.* **26**, 460–474 (2017).
75. Millard, E. E., Srivastava, K., Traub, L. M., Schaffer, J. E. & Ory, D. S. Niemann-pick type C1 (NPC1) overexpression alters cellular cholesterol homeostasis. *J. Biol. Chem.* **275**, 38445–38451 (2000).
76. Watari, H. et al. Niemann-Pick C1 protein: obligatory roles for N-terminal domains and lysosomal targeting in cholesterol mobilization. *Proc. Natl Acad. Sci. USA* **96**, 805–810 (1999).
77. Zhang, M. et al. Cessation of rapid late endosomal tubulovesicular trafficking in Niemann-Pick type C1 disease. *Proc. Natl Acad. Sci. USA* **98**, 4466–4471 (2001).
78. Yanjanin, N. M. et al. Linear clinical progression, independent of age of onset, in Niemann-Pick disease, type C. *Am. J. Med. Genet. B. Neuropsychiatr. Genet.* **153B**, 132–140 (2010).
79. Oliver, M. A. & Webster, R. *Basic Steps in Geostatistics: the Variogram and Kriging*. 1–99 (Springer, 2015).
80. Pebesma, E. J. Multivariable geostatistics in S: the gstat package. *Comput. Geosci.* **30**, 683–691 (2004).

## Acknowledgements

We thank D. Ory (University Washington St. Louis) for the wild-type V NPC1 vector and F. Maxfield (Weill Cornell Medical School) for performing the cholesterol analyses<sup>21</sup>. M. McGuckian (TSRI) for helpful discussions. Support was provided by the National Institutes of Health Grants HL095524 and DK051870, the intramural research program of the Eunice Kennedy Shriver National Institute of Child Health and Human Disease and by the Ara Parseghian Medical Research Foundation (APMRF) Notre Dame to K.S. and P.Z., and Support for Accelerated Research (SOARs) in Niemann-Pick disease to S. M.S. We would also like to express appreciation to the guardians and 27 participants enrolled in our natural history study that provided the clinical profile of each variant that is used for VSP<sup>1</sup> (see Supplementary Table 3).

## Author contributions

C.W. and W.E.B. developed the epigenetic-SCV concept. C.W. developed and performed the VSP analysis. S.M.S., D.H., and P.Z. contributed to the trafficking analysis of NPC1 variants in U2OS cells and probed the proteostasis components under SAHA treatment. K.S. performed the trafficking analysis of NPC1 variants in HeLa cells in response to SAHA and analyzed the hyperacetylation of NPC1 in response to SAHA. S.L. contributed to the programming in R-language. N.Y.F. and F.D.P. collected and analyzed the natural history of NPC1 patients. C.W. and W.E.B. wrote the paper.

## Competing interests

The authors declare no competing interests. The authors declare no advisory, management, or consulting positions. C.W. and W.E.B. have filed PCT Application Serial No. PCT/US2019/046028.

## Additional information


Supplementary information is available for this paper at <https://doi.org/10.1038/s41467-019-12969-x>.

Correspondence and requests for materials should be addressed to W.E.B.

Peer review information *Nature Communications* thanks Emidio Capriotti and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

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

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

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

© The Author(s) 2019