Published in final edited form as:

Cell Rep. 2018 August 21; 24(8): 2013–2028.e6. doi:10.1016/j.celrep.2018.07.059.

Bridging Genomics to Phenomics at Atomic Resolution through Variation Spatial Profiling

Chao Wang[1] and **William E. Balch**[1,2,3,*]

[1]Department of Molecular Medicine, The Scripps Research Institute (TSRI), La Jolla, CA 92037, USA

[2]The Skaggs Institute for Chemical Biology, The Scripps Research Institute (TSRI), La Jolla, CA 92037, USA

[3]Lead Contact

SUMMARY

To understand the impact of genome sequence variation (the genotype) responsible for biological diversity and human health (the phenotype) including cystic fibrosis and Alzheimer's disease, we developed a Gaussian-process-based machine learning (ML) approach, variation spatial profiling (VSP). VSP uses a sparse collection of known variants found in the population that perturb the protein fold to define unknown variant function based on the emergent general principle of spatial covariance (SCV). SCV quantitatively captures the role of proximity in genotype-to-phenotype spatial-temporal relationships. Phenotype landscapes generated through SCV provide a platform that can be used to describe the functional properties that drive sequence-to-function-to-structure design of the polypeptide fold at atomic resolution. We provide proof of principle that SCV can enable the use of population-based genomic platforms to define the origins and mechanism of action of genotype-to-phenotype transformations contributing to the health and disease of an individual.

Graphical Abstract

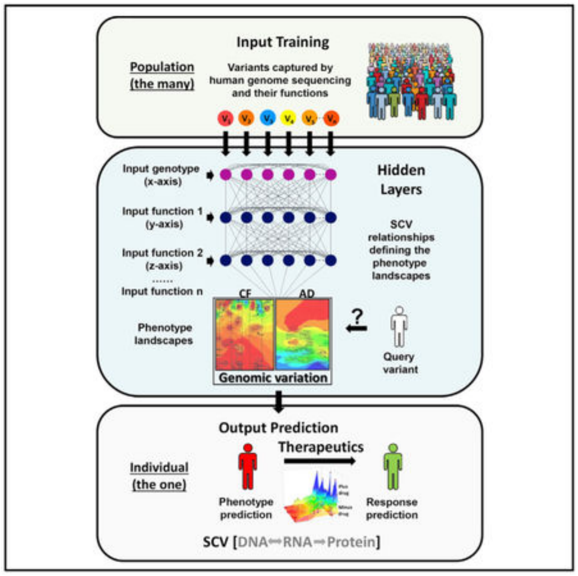
[*]Correspondence: webalch@scripps.edu.

AUTHOR CONTRIBUTIONS
C.W. and W.E.B. developed the methodology and concept, C.W. performed the analysis, and C.W. and W.E.B. wrote the manuscript.

SUPPLEMENTAL INFORMATION
Supplemental Information includes seven figures, one table, and four videos and can be found with this article online athttps://doi.org/10.1016/j.celrep.2018.07.059.

DECLARATION OF INTERESTS
The authors declare no competing interests. The authors declare no advisory, management, or consulting positions. The authors have filed a provisional patent application (serial number 62/685,554).

SUPPORTING CITATIONS
The following references appear in the Supplemental Information: Carter et al. (1992); Masica et al. (2015); Taylor-Cousar et al. (2017); Wainwright et al. (2015).



In Brief

Wang and Balch develop variation spatial profiling (VSP), a machine learning approach to integrate genomics and phenomics of the population to inform on the phenotype of the individual at atomic resolution. VSP is based on the principle of spatial covariance (SCV) that defines central dogma as matrices to track information flow from the genotype-to-phenotype to facilitate high-definition medicine.

INTRODUCTION

Interpreting the impact of familial and somatic variation in the genome on the protein fold and function in diverse physiological contexts (Anfinsen, 1973) is critical for implementation of high-definition medicine (Torkamani et al., 2017). Associated with this concern is the need to link the genotype to the phenotype—a universal challenge in the era of human genome sequencing (Manolio et al., 2017). To assess the impact of genetic diversity on protein function and structure, ancestral approaches can be used to compare residue conservation across evolutionary time to assign evolved chemical and/or physical constraints defining the function of the polypeptide fold (Hopf et al., 2017), whereas deep mutational scanning (DMS) attempts to facilitate interpretation through induced random genetic variation (Starita et al., 2017). These approaches fail to guide an understanding of the impact of genetic diversity on protein function found in the many cell- and tissue-specific environments that are unique to each one of us.

To understand the genotype-to-phenotype transformation contributing to function, we hypothesized that sequence variation in the human population can be used as a collective to generate a platform that quantitatively tracks hidden sequence-to-function-to-structure relationships that contribute to diversity and function in the individual. For this purpose, we developed variation spatial profiling (VSP). VSP uses the fiduciary (trusted) sequence positions (i.e., genotypes) of a sparse collection of inherited disease-associated variants found in the population with known biological functions (i.e., phenotypes) to map their collective spatial relationships, which we define as spatial covariance (SCV). Herein, we first develop and validate the interpretive power of VSP using the recessive, loss-of-function variants of the cystic fibrosis (CF) transmembrane conductance regulator (CFTR) to reveal the sequence-to-function-to-structure relationships contributing to CF disease, a platform useful for application of therapeutics to the individual CF patient. To generalize our SCV-based platform, we use allele frequency to assess the evolutionary impact of variation on CF. To expand the application of the SCV principle, we use function and allele frequency to predict the pathogenicity of dominant gain-of-function amyloid precursor protein (APP) variants responsible for Alzheimer's disease (AD), and to capture the value of the Aβ-42/Aβ-40 ratio to predict age of onset (AO) of dementia. We suggest VSP provides an unanticipated approach to read the genome by interpreting central dogma in the context of genetic diversity of the population through the principle of SCV.

RESULTS

Defining SCV through a Gaussian Process

To address the role of genetic variation in biological diversity and human healthspan, we reasoned that variants found in the population report on conserved (but largely unknown) evolutionary rules that dictate the biophysical, biochemical, and/or biological properties of folding intermediates informing normal protein function. To bridge sequence variation with phenotypic diversity, we developed VSP. VSP is inspired by well-established Gaussian-process (GP)-based regression approaches used in geostatistics (Chilès and Delfiner, 2012) that analyze relationships between datasets based on x axis (latitude) and y axis (longitude) position coordinates (Figure 1A; see STAR Methods). These coordinates are used to build an image of the landscape that predicts the probability of the distribution of, for example, a geologic feature such as a commodity (e.g., oil) (Figure 1A, z axis). This matrix-based map is derived using the spatial relationships between a sparse collection measured positions in the landscape (e.g., "boreholes" used for oil) and the covariance calculated between the values found in these positions (Figure 1A). These known spatial relationships are then used to capture the values for all unmeasured (unknown) positions in the landscape based on the rationale that measured positions closely separated in geophysical space (their proximity) are more correlated to each other than those at more distant spatial locations. Only a sparse collection of positional relationships in the landscape, that is, their proximity-linked measurements, are necessary to define unknown values across the entire landscape with high confidence. In geostatistics, high-confidence (low uncertainty) predictions typically require ~50 or greater sampling positions (Kerry and Oliver, 2007).

We recognized that variants in the population can experimentally serve as fiduciary (trusted) "molecular markers," like geological boreholes, to yield fundamental insights into the value of relationships that link each variant position (its genotype) to its value in the polypeptide chain, that is, its phenotype defined by protein function. For this purpose, we apply a proximity-based biological principle we term SCV. SCV captures the impact of GP-based covariance to map value in sequence-to-function relationships as a continuous landscape image that can be transformed to structure. By linking the linear sequence information found in the genotype of the population to functional states of the polypeptide fold, SCV relationships can be used to predict unknown functional and structural values for every amino acid position in the polypeptide sequence.

Applying SCV to Profile Disease Variants

To test the SCV principle, we first turned our attention to the broad field of human Mendelian disease, where inherited variants have a transformative impact on protein folding, stability, and function. Familial disease provides a robust genotype-to-phenotype differential relative to the normal wild-type (WT) protein to address the role of sequence variation in human physiology. Of the over 10,000 rare diseases cataloged to date (Landrum et al., 2016), CF is a well-studied and prevalent (~100,000 patients worldwide) early-onset, autosomal recessive (loss-of-function) disorder involving variants in the CFTR (Cutting, 2015).

CFTR is a multi-membrane-spanning polypeptide (Figure 1B) belonging to a large and diverse ABC transporter family containing transmembrane domains (TMDs) and regulatory nucleotide-binding domains (NBDs) (Figures 1B and S1A) (Liu et al., 2017; Zhang et al., 2017). At the apical surface, CFTR functions as a key chloride channel that maintains ion balance and hydration in sweat, intestinal, pancreatic, and pulmonary tissues, each providing a unique physiological environment likely differentially contributing to CFTR function (Amaral and Balch, 2015).

Assigning CFTR Landscape Coordinates: Step 1

Of the CFTR variants found in the population with a confirmed CF clinical phenotype (Sosnay et al., 2013), 159 have an allele frequency above 0.01% and encompass ~96% of the patient population. 67 genotypes are missense or deletion variants that result in the expression of a full-length but dysfunctional protein (Figures 1B and S1A). The Phe508 deletion (F508del) variant contributes to ~85% of clinical disease in homozygous (~45%) or heterozygous state with other rare variants. Recent cryoelectron microscopy (cryo-EM) structures of CFTR in the presence or absence of phosphorylation and ATP binding reveal that large conformational changes accompany channel gating and function (Liu et al., 2017; Zhang et al., 2017). The impact of variation (Cutting, 2015) (https://www.cftr2.org/) on these structural states and their contribution to the natural history of disease, risk management, and/or clinical intervention through therapeutics for each individual in the CF patient population remain to be defined.

To generate the input data for our VSP approach, we used 63 experimentally characterized CFTR missense variants (Sosnay et al., 2013) (Figures 1B and S1A). In the first step of VSP (Figure 1C, step 1), we positioned these variants as distance relationships based on the position of their genotype encoded variant amino acid along a linear (1-dimensional [1D]) polypeptide sequence normalized to the full-length WT chain set as a value of 1. Here, we refer to this value as the variant sequence position (VarSeqP) (Figure 1C, step 1, 1D). For the y and z axis coordinates that will contribute to sequence-to-function relationships, we used biologic features associated with each variant. CFTR requires trafficking in the exocytic pathway from the endoplasmic reticulum (ER) through the Golgi to its final destination at the apical cell surface of epithelial cells to achieve biological function. Therefore, as a second dimension (2D) y axis coordinate, we assigned the value of each variant's experimentally determined trafficking to the Golgi, referred to as the trafficking index (TrIdx) (Figure 1C, step 1,2D). The Trldx is the fraction of a CFTR variant exported from the ER relative to the total amount of variant found in the cell, normalized to WT CFTR. The resulting plot (Figure 1C, step 1,2D) links the genotype (x axis) to a phenotype (y axis).

To correlate sequence position (VarSeqP) (x axis) and trafficking (y axis) to a feature to be predicted by VSP, the third dimension (the z coordinate) was defined by the experimentally measured chloride conductance (ClCon) value for each variant normalized to the ClCon value of WT (Figure 1C, step 1, three-dimensional [3D]). The z axis functional feature is equivalent to the measured values recovered from a sparse distribution of geological features (e.g., such as oil found in boreholes; Figure 1A). The ClCon value is spatially defined in the context of its unique x axis (sequence position) and y axis (trafficking) coordinates. The spatial relationships defined by the x axis and functional y and z axes coordinates provide a quantitative framework to assign value and map function across the entire polypeptide sequence through GP regression.

Building the Phenotype Landscape: Step 2

To transform the sparse genotype sequence information encoded by our collection of 63 variants into the phenotype of the entire polypeptide chain, in the second step of VSP (Figure 1C, step 2), we assessed the spatial relationships of each known variant (x axis) and its unique biological features (y and z axes) using a variogram (STAR Methods). The variogram is a GP descriptor that captures biological spatial correlations that are used for ML based on the input sparse collection of variants and their features (Figure 1C, step 1).

Generation of the variogram involves pairwise analysis of the 63 sparse variants to yield all possible 1,953 combinations of spatial relationships as output (Figure 1C, step 2, top). The 2D distance values linking VarSeqP to TrIdx (Figure 1C, step 2, bottom, x axis) were first calculated to report how CFTR trafficking is changed in response to each variant sequence position. The associated 3D spatial relationships with ClCon were then calculated to assess variance of the proximity values of ClCon for all combinations of the VarSeqP coupled TrIdx positions (Figure 1C, step 2, bottom, y axis) to generate the variogram (Figures S1B-S1D). The variogram reports on the SCV relationships of known sequence positions to trafficking to ClCon function to define the unknown SCV relationships as output matrix, just as x and y axis linear coordinates in geostatistics links the positions of boreholes to predict the spatial distribution pattern of commodity values as output (z axis) (Figure 1A).

Our "molecular" variogram quantitates the sequence range where the variants co-vary with each other for a given set of functional relationships, in this case the TrIdx and ClCon values. We find that the spatial variance of ClCon for CFTR increases according to the linked changes in both VarSeqP and TrIdx until it reaches a plateau (Figure 1C, step 2, bottom). The plateau occurs at distance of ~0.14 (Figure 1C, step 2, bottom), a computed feature of the fold we refer to as the molecular range. A molecular range of ~0.14 reveals that the TrIdx and ClCon function of variants are generally dependent on each other only over a short sequence range, a module of function, in this case ~150–200 amino acids.

Variants with spatial relationships extending beyond the module range are generally not correlated and therefore likely to have more extended (direct or indirect) relationships to modulate function, perhaps reflecting flexible intra- or inter-domain interactions found in the full-length protein and/or in their interactions with other proteins in the complex environment of the cell (Pankow et al., 2015). Thus, SCV reports on spatial relationships that coordinate sequence position with function that now enable us to calculate an output matrix, the “phenotype landscape” that captures the unknown.

Using the Phenotype Landscape to Define Function in the Individual: Step 3

Based on the SCV relationships generated in step 2 as input, we apply GP regression to relate our characterized sparse collection of variants (the known) to the uncharacterized amino acids comprising the remainder of the polypeptide chain (the unknown). The resultant matrix-based output phenotype landscape allows us to quantitatively assess all unmeasured ClCon values in the context of the TrIdx for amino acids spanning the entire polypeptide sequence, along with an uncertainty associated with each value (Figure 1C, step 3, ~2,100,000 predictions shown as a color gradient; Video S1). We refer to this 3D landscape (Figure 1C, step 3) as the ClCon-phenotype landscape reflecting its z axis coordinate.

The SCV-based landscape generated from genetic diversity in the population can be used to assess function in the individual harboring a specific variation. For this purpose, the ClCon-phenotype landscape (Figure 1C, step 3) is back-projected to a 2D map with the color scale (a heatmap) representing the z axis ClCon function (Figure 2A). The molecular variogram (Figure 1C, step 2, bottom) used to generate the ClCon-phenotype landscape also defines the confidence or uncertainty for each mapped value. These values can be plotted as a gradient of contour lines (a molecular fingerprint) representing the uncertainty in applying SCV relationships for each uncharacterized amino acid in the CFTR full-length sequence (Figure 2A, gray contours; Figure S1E). For example, a location within the top 25% confidence quartile (Figure 2A, opaque color regions) have input variant values within the top one-third of the molecular range (Figure 1C, step 2, bottom). These SCV relationships are of high confidence and more dependent on one another than locations outside the top 25% confidence quartile (Figure 2A, transparent color regions). The residues in the top 25% contours with similar predicted ClCon values we refer to as clusters. Clusters reveal the contribution of both known and unknown (predicted) amino acids to the overall functional spatial design of the fold.

To validate the output of the ClCon-phenotype landscape (Figure 2A), we used a different dataset of diverse CF variants (Van Goor et al., 2014; Yu et al., 2012) (Figure S1F, inset) not included in the training dataset (Sosnay et al., 2013) (Figure 2A, plus symbols). Validation reveals a strong correlation (Figure S1F; Pearson’s r = 0.81, p value = 2×10^{-4}) between all the experimentally measured values and the newly mapped values that define the output phenotype landscape. These results demonstrate that VSP can incorporate complex sequence and feature-based functional relationships using >50 fiduciary variant markers (Figure S1G), which comprise only 5% of the total CFTR sequence, to generate a continuous landscape view of physiological features spanning the entire CFTR polypeptide. For example, the ClCon-phenotype landscape reveals that for all residues that have a TrIdx value of approximately <0.4–0.5 (Figure 2A, y axis), VSP predicts a nearly complete loss of ClCon, reflecting the impact of SCV states that prioritize cellular location (i.e., ER) relative to ClCon function found at the cell surface (Figure 2A, z axis, red). In contrast, for CFTR variants that have a TrIdx value of approximately >0.4–0.5 (Figure 2A, y axis), VSP predicts substantial sequence-based variability in ClCon (from none to greater than WT), illustrating the sensitivity of the CFTR fold to highly variable endocytic trafficking and channel regulation pathways at the cell surface that have no impact on export from the ER.

Translating the Phenotype Landscape to a Functional Structure

To examine whether phenotype landscapes derived from linear sequence information and associated biological features can provide functional insight into the conformation(s) captured by structural methods, we mapped phenotype landscape values to cryo-EM snapshots of CFTR (Liu et al., 2017; Zhang et al., 2017) in open and closed channel conformations reflecting the response of the channel to ATP-binding (Figure 2B). We assigned the prediction value with highest confidence to uncharacterized residues (Figure 2B, left panels) to link function to conformation where TrIdx (Figure 2B, right panels, ball size), predicted values of ClCon (Figure 2B, right panels, color gradient), as well as their confidence in prediction (Figure 2B, right panels, transparency gradient) provide a complete map of sequence-to-function-to-structure relationships in CFTR (STAR Methods). We refer to this overlay of phenotypic landscape values onto the CFTR structure snapshot as a functional structure.

To illustrate the biological design of CFTR (Figure 2B) revealed by our VSP perspective (Figure 2A), the predicted sequence regions within the high-confidence 25% contour that have low trafficking values (TrIdx < 0.2) allow us to quantitatively assign the role of the ER in the folding and trafficking of CFTR (Figure 2C). For example, NBD1 can be defined by the SCV relationships that form the high-confidence cluster 1 (<25% confidence contour) (Figure 2C, top, SCV cluster 1) that includes the common CF variant F508del and the critical S492 residue central to the molecular dynamics of the NBD1 module controlling trafficking (Proctor et al., 2015). This cluster also contains the diacidic exit code required for ER export (Figure 2C, bottom, circle 1, black arrows; Figure S1H) (Wang et al., 2004). The high-confidence SCV relationships defined by cluster 1 in this subdomain of NBD1 (Figure S1H) illustrate the spatial design of intra-domain functional interactions that coordinate the interaction of NBD1 with COPII for ER export. Moreover, VSP predicts that NBD1 does not operate in isolation from the other modular features of the CFTR fold. Cluster 2 (Figure 2C, top, cluster 2; bottom, bar 2 on functional structure) in TMD2 defines longer-range, inter-domain interactions that tune ER stability and/or export, a conclusion supported by experimental observations (Mendoza et al., 2012; Rabeh et al., 2012). These two major clusters together with several other regions contributing to trafficking in the functional structure provide a mechanism in which two legs of the transmembrane fold (Figure 2C, bottom, TM11-ICH4-NBD1 [leg 1] and TM4-ICH2-NBD2 [leg 2] connected by TM1 and TM3; Figure S1I) that defines the functionality of NBD1 for export. Most of the predicted residues restricting trafficking are neither facing the interior of the gated channel nor involved in ATP binding (Figure 2C, bottom, top view), indicating that ER export is largely uncoupled from features guiding CFTR channel and gating function at the surface.

Consistent with this view, ~30% of CF missense variants show >80% of the WT trafficking value but have deficient ClCon function (<15% of WT ClCon) (Figure 2A). In contrast to residues modulating ER export, when we mapped the sequence clusters in the phenotype landscape with WT-like TrIdx but deficient ClCon function onto the CFTR functional structure (Figure 2D), all of them can be aligned along the channel faces or in ATP-binding regions that do not impact ER export. For example, SCV clusters 3 and 4 found at the interface of the NBD1 and NBD2 are predicted to couple inter-domain interactions to mediate the channel gating (Figure 2D, bottom; Figure S1J). Thus, VSP transforms SCV relationships (i.e., high-confidence clusters) into structural units and links them by their contributions to function that highlights unanticipated modularity of the fold for trafficking and function.

Using Phenotype Landscapes to Assess Value in Therapeutics

To demonstrate that VSP can reveal how the local chemical environment influences the genotype-to-phenotype transformation, we applied VSP to the variant dataset (Van Goor et al., 2014; Yu et al., 2012) that we used for validation of the CFTR ClCon-phenotype landscape (Sosnay et al., 2013). Variants were either untreated or treated with the US Food and Drug Administration (FDA)-approved therapeutic ivacaftor, a channel gating potentiator that increases the open probability of cell-surface-localized CFTR (Van Goor et al., 2014; Yu et al., 2012). While ivacaftor has no effect on export of F508del, it was shown to have a substantial impact on improving ClCon of the G551D variant found in SCV cluster 3 at the NBD1-NBD2 interface (Figure 2D, bottom, #) which traffics normally to the cell surface, but lacks conductance (Figure 3A, left, #).

The variogram (Figure S2B) reveals that ivacaftor has only a minor impact on the molecular range but increases the spatial variance of the plateau value from 0.05 in the absence of ivacaftor to 0.29 in its presence. This unexpected large change suggests that ivacaftor mechanistically increases the overall spatial variance of the fold leading to decreased stringency in gating and/or channel activity to restore function. Consistent with this interpretation, VSP reveals a striking change in the ClCon-phenotype landscape output for a substantial fraction of the polypeptide chain (Figures 3A and S2C). The ivacaftor responsive phenotype landscape demonstrates that variants with a measured or predicted minimum TrIdx value of ~0.3–0.4 (Figures 3A and S2D–S2F; Pearson’s r = 0.6, p value = 4 × 10^{-7}) and a level of post-ER mature glycoform of approximately >0.4–0.5 of that observed for WT CFTR (Figures S2G–S2I; Pearson’s r = 0.73, p value = 8 × 10^{-12}), will be responsive to management by the drug. For example, in addition to G551D (Figure 3A, #), most of the variants that were recently approved by the FDA based on *in vitro* data (Ratner, 2017) (Figure 3A, right, black triangles) are mapped by VSP to be responsive to ivacaftor with the exception of A455E (Figure 3A, right, *) that has a TrIdx of 0.3 and is predicted by VSP to be an ivacaftor nonresponder (Figures S2F and S2I, *), suggesting that this variant is not a good candidate for ivacaftor intervention, as observed in the clinic (McGarry et al., 2017).

To visualize the therapeutic response of ClCon-phenotype landscapes from our functional structure view, the highest-confidence predicted values following ivacaftor treatment for each residue were mapped onto the closed and open CFTR structure snapshots (Figures 3B and S2E). As expected, a before and after comparison of the ER-restricted residues such as SCV cluster 1 in NBD1 domain and cluster 2 in TMD2 failed to show a response to ivacaftor (Figure 3B). In contrast, 63% of CFTR residues within the 25% confidence quartile (Figure 3A, right) are shown to have at least a 20% increase in function relative to that of WT ClCon in response to ivacaftor (Figure 3B, right, 813 residues, yellow to blue balls;Figure S2E ; Table S1). These variants already have a significant TrIdx and are mostly located in the ATP-binding site contributed by SCV clusters 3 and 4 found at the NBD1-NBD2 interface and along the channel region (Figures 3B and S2E; Videos S2 and S3). The integrated results captured by VSP lead us to suggest that ivacaftor unexpectedly serves as a dynamic “SCV agonist” that triggers a ripple effect that either directly or indirectly spans most of polypeptide chain to improve its spatial flexibility to improve its channel function (Figures 3B and S2J). These SCV relationships now provide a platform explain the basis for correction of sequence-to-function-to-structure responses of numerous CFTR variants to ivacaftor. Furthermore, the impact on the variable response to ivacaftor by different cell-based and/or clinical modifier environments, or the response of different variants at the same physical location in the sequence, can be assessed by deep analysis of 3D projections of phenotype landscapes (Figure S3).

Tissue-Specific Phenotype Landscapes

To demonstrate that our VSP strategy can capture SCV relationships defining genotype to phenotype transformations impacting the onset and progression of disease in the clinic, we used TrIdx as the input y axis value with known clinical measures of CF disease as input z axis values (Figure 4A) (Sosnay et al., 2013). Patient measures include sweat chloride (SC), forced expiratory volume in 1 breath (FEV1), *Pseudomonas* burden (PB), and pancreatic insufficiency (PI) (Sosnay et al., 2013). To make all z axis input measures comparable, we normalized their values by setting the F508del value to 0 and that of WT to 1. Here, phenotype landscapes (Figure 4A) and their functional structures (Figure 4B) demonstrate, as expected, that a poor TrIdx predicts not only poor ClCon across all human tissue environments (Figure 4A; ClCon layer, y axis < 0.4 [red to orange]) but also poor FEV1, SC, PB, and PI clinical outcomes (Figure 4A; SC, FEV1, PB, PI layers, y axis < 0.4 [red to orange]). For example, NBD1-based SCV relationships that limit ER export (e.g., Figure 4B, cluster 1 and bar 2) are defective for all phenotypes. Moreover, residues localized to the ATP-binding site managing ClCon (Figure 4B, cluster 3 and 4) are also defective in all tissue environments. These results suggest a conserved role for these residues in managing the CFTR fold for all tissue function.

In contrast to the conserved roles of trafficking and channel gating variants, VSP captures a number of SCV relationships that either under- (Figure 4B, cluster 5) or overestimate (Figure 4B, cluster 6) the potential impact of a variant on a given clinical phenotype relative to the cell-based derived measurement of ClCon. Tissue-specific SCV relationships are best seen by the divergent FEV1 and PI phenotype landscapes (Figure 4A; compare FEV1 to PI layers, arrow) and their functional structures (Figure 4B, compare FEV1 to PI, arrow; Video S4). For example, cluster 7 presents as a severe phenotype for FEV1 and PB but is mild for PI and SC (Figure 4B). In contrast, cluster 8 is mild for FEV1 but severe for PI (and other clinical responses) (Figure 4B). Moreover, cluster 9 is severe for all clinical indications but has only a mild impact on PI (Figure 4B). The differences found in tissue specificity of function may reflect the fact that CFTR manages ClCon and hydration in a non-homeostatic environment in the lung, while CFTR manages bicarbonate secretion that is critical for pancreas function in response to homeostatic environment (Figure S4) (LaRusch et al., 2014).

Linking Bench to Bedside through VSP

Given that VSP is a highly flexible platform that can integrate a common set of sparse variant datasets, we generated phenotype landscapes and the predicted functional structures for all 30 pairwise combinations of y and z axis coordinates reflecting both bench and bedside measurements (Figure 5A). These phenotype landscapes were used to cross-correlate the predicted output of a basic and/or a clinical feature with one another. Using a leave-one-out cross-validation analysis to evaluate the prediction accuracy of each phenotype landscape (Figure 5B), we found significant Pearson r values of 0.52 ($p = 2 \times 10^{-5}$) and 0.77 ($p = 3 \times 10^{-13}$) using the bench-based model to predict either ClCon or TrIdx-phenotype landscapes as the z axis value, respectively (Figure 5B, bottom left quadrants). Moreover, statistically significant SCV correlations were found using FEV1, SC, PB, or PI as a y axis value to predict a different clinical feature as the output z axis value (Figure 5B, top right quadrant). For example, we observed a significant quantitative relationship using FEV1 as the y axis to predict PB as the z axis (Figure 5B; panel 9; Pearson's $r = 0.67$, $p = 3 \times 10^{-9}$) or, conversely, using PB to predict FEV1 (Figure 5B; panel 14; Pearson's $r = 0.64$, $p = 2 \times 10^{-8}$). These results are consistent with the fact that these features are physiologically linked in airway-associated CF disease. In contrast, when using PI as the y axis coordinate to predict FEV1 as the z axis value, we found a substantially lower Pearson's r value (Figure 5B; panel 15; Pearson's $r = 0.32$, $p = 0.01$), consistent with their very different physiologic role(s) in CF clinical progression (LaRusch et al., 2014).

To link bench to bedside, we tested the value of cell-based (bench) measurements as the y axis value to predict clinical measures (bedside) as the z axis value across the entire predicted CF variant population (Figures 5A and 5B, top left quadrant). Such relationships present a fundamental challenge in high definition medicine where most cell-based and animal models fail to predict clinical outcome, leading to substantial loss of time and financial resources. Consistent with this concern, nearly all VSP bench-to-bedside predictions show weak but statistically significant correlations (Figure 5B, top left quadrant). The strongest correlation was seen when we use ClCon as the y axis to predict SC (Figure 5B, panel 17; Pearson's $r = 0.63$, $p = 3 \times 10^{-8}$). Thus, cell-based ClCon measurements largely capture SC responses recovered from the patient population, a prediction validated by clinical observations (Collaco et al., 2016). These results validate the utility of the VSP to serve as a guide to link the value of SCV relationships generated by cell-based models to assess the impact of a therapeutic for a physiologically relevant clinical feature (Figures S5A and S5B).

Generalizing VSP Using Allele Frequency

To generalize the SCV principle, we considered the possibility that allele frequency from the GnomAD database (http://gnomad.broadinstitute.org/; 138,632 individuals) could serve as a universal genome-based y axis coordinate, like the x axis position coordinate, to assess biologically relevant functional SCV relationships (z axis) for all variant genotypes found in the population. Using CFTR variants to first calibrate whether allele frequency as the y axis coordinate can inform on SCV relationships contributing to disease in the CF population, we generated TrIdx- (Figures 5C and S5C; Pearson's r = 0.6, p = 3 × 10^{-7}) and ClCon-phenotype landscapes (Figures 5D and S5D; Pearson's r = 0.67, p = 5 × 10^{-9}) as predicted z axis features. Intriguingly, they are strikingly different. As shown in Figure 5C in the Trldx-landscape molecular fingerprint, variants are found distributed as small clusters throughout the primary sequence, reflected in the very short molecular range found in the variogram (Figure S5E; range = 0.17, ~150 amino acids). These results suggest that allele frequency reports on trafficking through local SCV relationships (Figure S5G). In contrast to the TrIdx phenotype landscape, the ClCon-phenotype landscape molecular fingerprint shows that allele frequency largely correlates with ClCon function across the entire polypeptide (Figure 5D). Here, variants with allele frequency below ~0.02% of the population (Figure 5D) all have deficient ClCon function, while most of variants with allele frequency above ~0.02% have strong ClCon values. The ClCon variogram has an extended molecular range (Figure S5F; range = 2.65, i.e., the full-length protein), indicating that the entire polypeptide operates as a functional unit to determine the evolutionary trajectory of the fold in the health of the individual (Figures 5D and S5H). Exceptions are F508del (NBD1) and L997F (Figure 5D, highlighted by *), possibly due to their beneficial role in partial protection of the population to pathogens such as *V. cholerae* (Thiagarajah et al., 2015). Thus, allele frequency provides an unanticipated y axis feature that can be used to assess SCV relationships in recessive loss-of-function genotype to phenotype transformations.

Using VSP to Assess Onset of AD

To address the ability of allele frequency as general metric to move beyond loss-of-function recessive rare diseases such as CF and provide insight into the pathogenicity of more common age-related gain-of-toxic function such as neurodegenerative diseases, we applied VSP to AD. Whereas combined inherited and somatic forms of AD impact nearly 50 million people worldwide, ~25% of the population has familial AD (FAD), of which ~95% is defined by late-onset AD (LOAD) (age >60–65 years) and 5% is defined by early-onset AD (EOAD) (age <60–65 years), largely in response to variants in APP and presenilin 1 (PS1). APP contributes to 10%–15% and PS1 contributes to ~50% of EOAD (Giri et al., 2016). APP is a single-membrane-spanning protein whose cleavage through the sequential activity of β- and γ-sec-retases (Hunter and Brayne, 2018) is altered in response to inherited and/or sporadic disease, leading to the generation of amyloidogenic peptides referred to as Aβ.

For VSP, we used as input the available 45 missense variants of APP reported in ClinVar (Landrum et al., 2016) and ALZFORUM (https://www.alzforum.org/) databases as x axis values, allele frequency reported in the GnomAD database as y axis values, and pathogenicity as reported in the ClinVar and ALZFORUM databases as z axis values to generate as output the APP pathogenicity (APP^{path})-phenotype landscape (Figures 6A and S6A; Pearson's r = 0.9, p = 2 × 10^{-13}; STAR Methods). VSP achieves 0.98 area under the curve (AUC) in receiver-operating characteristic (ROC) analysis, which is significantly higher than other variant function prediction algorithms, which are all below 0.75 (Figure

S6B), indicating that VSP can consistently capture the biological principle(s) underlying AD from population genomics. As shown in the APP^{path}-phenotype landscape, “benign” or “likely benign” variants of higher frequency in the population are predicted by VSP to be distributed throughout the sequence (Figure 6A, green-yellow). In contrast, nearly all pathogenic variants generate a high-confidence SCV cluster in the C-terminal region of the APP^{path}-phenotype landscape that is absent from GnomAD (STAR Methods), emphasizing their rarity in the population (Figure 6A, *) with an exception of A713T (Figure 6A, **). These residues can be mapped to a partial APP functional structure (Barrett et al., 2012) (Figure 6B, red residues ~667–728). This SCV hotspot contains the nonpathogenic α-secretase cleavage site as well as the β- and γ-secretases cleavage sites that are responsible for the generation of Aβ-40 and the highly pathogenic Aβ-42 peptides found in amyloid plaques (Figure 6B) (Hunter and Brayne, 2018). In addition, VSP based on sparse variants in population predicts a high allele frequency region around the γ-secretase cleavage site (Figure 6A, ** and ***; Figure 6B, large balls), which is validated by plotting all the variants found in GnomAD (Figure S6C), suggesting that the sequence at this region is being continually optimized to (re)balance the composition of different Aβ peptides in human population possibly in response to aging.

To link the SCV hotspot (Figure 6A, *) in APP found in the population to the impact of Aβ fragments in familial disease in the individual, we applied VSP to variants found in presenilin 1 (PS1), the catalytic subunit of the γ-secretase that generates Aβ-42 and Aβ-40 fragments. Each variant has been shown to contribute differentially to levels of Aβ-42 or Aβ-40 (Sun et al., 2017), although no statistically significant correlation was found between either the total absolute amount of Aβ-42 plus Aβ-40 and the mean AO or between the Aβ-42/Aβ-40 ratio and the mean AO using conventional statistical parameters (Sun et al., 2017). Here, the Aβ-42/Aβ-40 ratio relative to that observed for WT PS1 (set as value of 1) was used as the y axis coordinate to predict the mean AO as the z axis coordinate in an AO-phenotype landscape (Figure 6C, left; Figure S6D, Pearson’s $r = 0.37$, $p = 4 \times 10^{-4}$). Using input data from 89 PS1 variants to generate the AO-phenotype landscape (Sun et al., 2017), we found variants that generate ~10-fold-change higher Aβ-42/Aβ-40 ratio than that of WT (Figure 6C, left, y axis > 10) show an early AO (Figure 6C, left, orange to red, AO < ~40). Variants that generate a 1- to 10-fold change Aβ-42/Aβ-40 ratio relative to that of WT (Figure 6C, left, 1 < y axis < 10) show a broad range of AO (Figure 6C, left panel, light blue to orange). In general, the overall impact of variants in this region (Figure 6C, left, 1 < y axis < 10) leads to a later AO compared to variants with y axis value above this range (Figure 6C, y axis > 10) (Figure S6E, $p = 0.02$). Consistent with these results, when the Aβ-42/Aβ-40 ratio is lower than WT (Figure 6C, y axis < 1), SCV reveals a significant delay in AO compared to all other variants (Figure 6C, left, blue; Figure S6E). Using the absolute level of Aβ-42 as the y axis coordinate in the AO-landscape, we found that the delay of onset does not simply reflect Aβ-42 levels (Figures S6F–S6H). Furthermore, neither absolute Aβ-40 nor absolute Aβ-40 plus absolute Aβ-42 as y axis values yield significance in predicting AO (Figures S6I and S6J).

The corresponding functional structure projection (Figure 6C, right) of the Aβ-42/Aβ-40-ratio-based AO-phenotype landscape (Figure 6C, left) onto the structure of PS1 (Bai et al., 2015a, 2015b) reveals the sequence-to-function-to-structure relationships contributing to

AO. Here, an SCV cluster leading to early onset comprises a region of PS1 that comprises TM2, TM3, and the loop between TM1 and TM2 (Figure 6C, left, highlighted by *). This cluster forms a putative APP-binding pocket (Bai et al., 2015a) in the PS1 functional structure (Figure 6C, right, dashed oval *). In contrast, the variant values between 1-fold change and 10-fold change relative to WT (Figure 6C, left, $1 < y$ axis < 10) show diverse AO relationships that highlight different SCV clusters contributing to AO based on the Aβ-42/Aβ-40 ratio. For example, the cluster comprising TM8 (Figure 6C, left, **) has an earlier age of onset compared to other residues with a similar Aβ-42/Aβ-40 ratio. The Pro-Ala-Leu (PAL) motif adjacent to this cluster has been shown to contribute to the catalytic core of PS1 (Figure 6C, right, arrows) (Bai et al., 2015b). In contrast, y axis Aβ-42/Aβ-40 ratio values < WT (Figure 6C, left, ***) contribute to a cluster found at the C terminus that begins at the hydrophilic loop (HL) region (Figure 6C, right, dashed oval, ***) affecting EOAD progression (Nelson et al., 2011). These results reinforce the ability of SCV to capture the importance of the residues impacting the Aβ-42/Aβ-40 ratio as a broadly predictive sensor of onset and progression of disease, a prediction consistent with its biomarker value in cerebrospinal fluid (Baldeiras et al., 2018) and in plasma of the AD population (Nakamura et al., 2018).

DISCUSSION

We have developed a platform that assigns SCV relationships to track as matrices the flow of information from the genotype to the phenotype (Figure 6D). VSP requires only a sparse collection of variants recorded in the genome of the population (Figure 6D, top, Input Training) to serve as fiduciary input reporters of evolution-based rules responsible for the phenotype. Variation can be used to build phenotype landscapes that predict the unknown from the known based on GP (Chilès and Delfiner, 2012; Rasmussen and Williams, 2006) (Figure 6D, middle, Hidden Layers). Using the linear sequence information stored in the genome, VSP captures same spatial relationships used by transcriptional and translational machineries to build flexible design into the protein fold for function in diverse physiological states (Anfinsen, 1973) defined by the y and z axis coordinates. From this perspective, the phenotype landscape creates an image-based view of features that can be used to quantitate and predict at atomic resolution how the physiological state of the fold utilizes SCV to generate function in the individual (Figure 6D, bottom, Output).

Our ability to use SCV-based phenotype landscapes to map the unknown from the known in the context of extant biology and physiology cannot be captured by structure snapshots that are generated out of context of their biological function(s) or by ancestral approaches that rely on evolutionary divergent physiological states. Moreover, SCV-based insight informs new relationships that cannot be defined using PolyPhen-2, SIFT and related predictive algorithms (Glusman et al., 2017) and is able to achieve predictive insights with higher fidelity (Figures S5G, S5H, and S6B). While we used available snapshot structures of CFTR, APP, and PS1 to validate SCV relationships, a structure is not necessary for the generation of the phenotype landscape. On the contrary, it is VSP that provides insight into structure snapshots that lack value without function. Our SCV platform suggests that polypeptides can have numerous diverse and unanticipated spatial relationships reflecting their physiological state based on the y and z function coordinates (Anfinsen, 1973).

Because VSP gains its interpretative power based on only a sparse collection of fiduciary markers found in the extant population, it differs substantially from DMS approaches that rely on large-scale mutagenesis to model disease (Starita et al.,2017). We have found that interpretation of the results generated using DMS can substantially benefit by application of VSP principles (Figures S7A–S7C). Furthermore, by embracing the high dimensionality of the protein physiological state (Anfinsen, 1973), VSP assigns value to structure based on evolved diversity that is highly relevant to the human population. VSP substantially differs from the current focus on prediction of protein structure to function relationships based on the chemical-physical properties of amino acid residues. Combining SCV principles with chemical-physical and/or ancestral alignment measurements as y and z coordinates may enable their use from our functional structure perspective (Figure 6D, middle). Moreover, SCV relationships captured by VSP could be used to prioritize functional diversity of native structural conformations using cryo-EM (Shen, 2018) Consistent with our VSP strategy (Figure 6D, middle), GP-based approaches can be used to evolve protein sequences to improve function (Romero et al., 2013).

VSP currently allows us to read sequence-to-function-to-structure relationships from coding sequence defining <2% of the genome (Figure 6D, middle). By focusing on functional relationships, SCV captures biological features reflecting the spatial organization of the genome impacting gene expression, post-translational modifications that impact both genome and proteome function, the buffering capacity of the proteostasis machinery that manages the protein fold (Balch et al., 2008), and interactions within the variation sensitive proteome that are unique to specific cell and tissue environments. Moreover, SCV suggests that endomembrane compartments play specific roles in the tunable management of sequence-to-function-to-structure relationships. For example, ClCon phenotype landscapes suggests that the ER only utilizes a subset of SCV relationships that can be independent of channel function to promote trafficking, suggesting that it does not operate as a quality control compartment to limit the delivery of functionally defective variants to downstream destinations (Ellgaard and Helenius, 2003). Rather, VSP suggest that the ER utilizes SCV relationships to manage the tolerance of the fold in biology (Wiseman et al., 2007).

As VSP generally requires a minimum of ~50 variants for generation of high-confidence landscapes (Kerry and Oliver, 2007), it can currently be applied to most genes found in public databases such as GnomAD (Lek et al., 2016), ClinVar (Landrum et al., 2016), or specialized databases that annotate the natural history of variant disease that link genotype to phenotype. Proteins for which genotype-linked phenotype information is currently not available is necessarily a limitation for application of VSP. Genetic relationships beyond missense mutations, including somatic variation, heterozygous alleles, epistatic alleles, and variants in the non-coding region of genome, can be captured by SCV when annotated by their functional features in the context of human genome sequencing efforts. VSP can serve as a versatile platform for high-throughput screening (HTS) to capture human phenotypic plasticity early in the therapeutic development pipeline (Figure 6D, bottom).

We now posit by quantifying genetic diversity in the extant population, SCV principles provide a universal basis to use the population to define molecular level spatial relationships and mechanisms contributing to fitness of the individual. In this relative way of thinking of spatial-temporal dependencies found in the population (Figure 6D, top, “the many”), phenotype landscapes help us to appreciate the complex integration of the parts (Figure 6D, middle) to understand the individual (Figure 6D, bottom, “the one”). VSP, being an unprecedented interpolation platform that can embrace multiple dimensions (Figure 6D, middle), suggests that SCV may enable the use of predictive data-rich phenotype landscape images to model human variation in the population (Goodfellow et al., 2016; Rasmussen and Williams, 2006) (Figure 6D, middle) and for management of the patient in the clinic (Figure 6D, bottom; Figure S7D). Defining central dogma as matrices of SCV relationships across the genome and proteome (Figure 6D, bottom, SCV[DNA ↔ RNA ↔ Protein]) suggests a potential role of spatial states for understanding the origins of genetic and phenotypic diversity contributing to natural selection (Darwin, 1859).

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

CONTACT FOR RESOURCE SHARING

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, William E. Balch (webalch@scripps.edu).

METHOD DETAILS

Brief introduction of GP in geostatistics—VSP is based on Gaussian Process (GP), which is widely used in geostatistics to analyze and predict spatially continuous phenomena in complex geophysical landscapes encompassing a wide range of geological, epidemiological, anthropological and environmental features (Chilès and Delfiner, 2012). GP is also widely used for regression problems in supervised machine learning and artificial intelligence (AI) applications (Rasmussen and Williams, 2006). GP used in geostatistics generates unbiased distance-based covariance relationships using measurable features in the context of sparse sampling techniques as a limited ‘known’ knowledge-base to predict the ‘unknown’ value in the geophysical landscape. In GP, a higher weight for prediction is placed on measured positions in closer proximity to the unmeasured locations compared to those found in more distant locations. GP not only provides interpolated values, but also measures of uncertainty for those values (confidence contours), generating a metric for assessing the probability of the prediction. The measurement of uncertainty is critical to informed decision making and risk management, as it provides information on the possible values for each location rather than just one interpolated value. In simple terms, GP in geostatistics embraces the general concept that sparse covariance relationships can be used to predict unknown values and their uncertainty across an entire feature-based landscape (Chilès and Delfiner, 2012).

The specific method of GP in geostatistics we used in this paper is Ordinary Kriging, which has the least assumptions and is the most commonly used GP method in geostatistics to provide optimal unbiased prediction(Chilès and Delfiner, 2012). Ordinary Kriging 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 (Chilès and Delfiner, 2012). There are other geostatistical interpolation techniques, for example, Universal Kriging, Probability Kriging, Co-kriging and Empirical Bayesian Kriging that have additional assumptions that are specialized for particular sets of data and may ultimately prove valuable for our VSP approach.

Rationale for applying GP to biological data—In VSP, we consider each variant as a fiduciary (trusted) reporter of proteostasis-sensitive folding intermediates that can be used to define the hidden evolutionary defined SCV relationships directing the genotype to phenotype transformation. VSP uses a sparse collection of variants spanning the full polypeptide sequence to predict function in a similar way that geostatistics uses sparse sampling measurements (boreholes) to predict unknown values across an entire geophysical landscape. Variants (i.e., variation distributed across the population) are the exceptions to the rules that make the rules. In so doing, they help us to understand the rules as they report on the evolved mechanisms that drive the normal function of the protein fold and multiple challenges by the environment to facilitate survival and fitness required for natural selection. From a practical perspective, SCV relationships captured by VSP can teach us, for example: (1) evolutionary design of protein fold for function, (2) relationship(s) that categorize value of population traits and, as shown herein, the onset and progression of disease in the clinic; (3) cell and tissue specific variables impacting variant polypeptide function or, among others, (4) generate a quantifiable common platform to assess the value of bench, animal and bedside derived features for developing interventional management/therapeutic strategies for any gene where clinically relevant variation in the population is available (e.g., (Landrum et al., 2016; Lek et al., 2016; Manolio et al., 2017)). A flowchart illustrating the application of VSP to human variation is shown in Figure S7D.

Spatial organization of the biological data—To integrate the sparse collection of sequence variation information found in the genome (the genotype) with biological features contributed by spatial relationships with function, we positioned the variants, our ‘molecular borehole/locations’, by their sequence positions in the polypeptide chain on the ‘x’ coordinate and measurements of a biological function on the ‘y’ coordinate to describe and predict another biological function along the ‘z’ coordinate. These relationships are similar to the positioning of boreholes defined by their longitude (x axis) and latitude (y axis) coordinates to predict oil reserves (z axis) in geostatistical analysis.

Variogram analysis—A geostatistics prediction is based on the SCV relationships of the input experimental data. A ‘molecular variogram’ (Figure 1C, Step 2, lower panel; Figure S1D) is used to describe how the ‘spatial variance’ (i.e., the degree of dissimilarity) of ‘z’ changes according to the separation distance (proximity) defined by the ‘x’ and ‘y’ coordinates. The molecular variogram defines a sequence-based ‘molecular range’ where the function of the variants depend on one another. The molecular variogram enables the calculation of SCV relationships in the dataset, forming the basis for prediction. The analysis of SCV relationships are described below:

Suppose the ith (or jth) 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 ith and jth observation is expressed 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 (in this case, 2 different variants), which is also the whole variance of ‘z’ value for one observation at the given separation distance ‘h’. In VSP, we refer to the $\gamma(h)$-variance as ‘spatial variance’ as indicated in the y axis of molecular variogram (Figure 1C, Step 2, lower panel; Figure S1D). Using Equations 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. Linear, spherical, exponential or Gaussian models can be used to fit the data in the molecular variogram, and the choice of model is usually determined by the residual maximum likelihood (REML) and the leave-one-out cross-validation result of the final phenotype landscape model. The distance where the model plateaus is referred to as the molecular range. Sample locations separated by distances within the molecular range are spatially dependent on one another, whereas those outside the molecular range are not. The SCV value at the distance (h) is expressed 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).

Confidence contour maps of SCV relationships—According to the variogram, observations that are close in distance (close proximity) are usually highly correlated and have more weight for prediction. To solve the optimum and unbiased weights of SCV relationships, Ordinary Kriging aims to minimize the variance associated with the prediction of the unknown value at location ‘u’, which is generated according to the expression-

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

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’. ω_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 solve the set of weights associated with input observations, but also provide the minimized 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’, ω_i is the weight for data point ‘*i*’, ‘$C_{i,u}$’ are the SCV between data points ‘*i*’ and ‘*u*. ‘μ’ is the Lagrange Parameter that is used to convert the constrained minimization problem in Equation 3 into an unconstrained one.

The standard deviation of prediction is generated as the square root of the resulting minimized Kriging variance in Equation 5. It provides the uncertainty of predictions that represents the confidence for using the SCV relationships both within the input data points and in relation to the unknown locations to make predictions. The confidence level is tightly linked with the distance range in the molecular variogram and the spatial distribution patterns of measured input points surrounding the unknown location. The shorter the distance between an unknown point to the input data points, the higher confidence for using the SCV relationships for the prediction.

The VSP matrix notation—The minimization of Kriging variance (Equation 3) with the constraint that the sum of the weights is 1 (Equation 4) can now be written in matrix form as

$$\mathrm{C} \cdot \mathrm{W} = \mathrm{D}$$

$$\begin{bmatrix} C_{1,1} & \cdots & C_{1,\mathrm{n}} & 1 \\ \vdots & \ddots & \vdots & \vdots \\ C_{\mathrm{n},1} & \cdots & C_{\mathrm{n},\mathrm{n}} & 1 \\ 1 & \cdots & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} \omega_1 \\ \vdots \\ \omega_\mathrm{n} \\ \mu \end{bmatrix} = \begin{bmatrix} C_{1,\mathrm{u}} \\ \vdots \\ C_{\mathrm{n},\mathrm{u}} \\ 1 \end{bmatrix} \quad (6)$$

where ‘C’ is the covariance matrix of the known data points. ‘W’ is the set of weights assigned to the known data points for generating the predicted phenotype landscape. ‘μ’ is the Lagrange multiplier to convert a constrained minimization problem into an unconstrained one. ‘D’ is the covariance matrix between known data points to the unknown data points. Since ‘W’ is the value we want to solve to generate the phenotype transformation (the phenotype landscape), this equation can be also written as

$$W = \underbrace{\mathrm{C}^{-1}}_{\text{Clustering}} \cdot \underbrace{\mathrm{D}}_{\text{Distance}} \quad (7)$$

where ‘C^{-1}‘ is the inverse form of the ‘C’ matrix.

As a more intuitive explanation of the Kriging matrix notation, herein we simply refer to the VSP matrix that generates the phenotype landscape (‘W’) to be based on the two important computational features used for predicting the unknown function values from the known- (1) the clustering (i.e., clustered sequence values with similar functional properties (C^{-1})) and (2) the distance constraints (D). Here, ‘C^{-1}’ represents the clustering information of the known data points while ‘D’ represents predicted statistical distance between known data points to unknown data points.

Generating the VSP prediction—With the solved weights ‘W’, we can calculate the prediction of all unknown values to generate the complete phenotype landscape by the equation

$$z_u^* = \sum_{i=1}^{n} \omega_i z_i \quad (8)$$

where z_u^* is the prediction value for the unknown data point ‘*u*’, ‘ω_i’ is the weight for the known data point and ‘z_i’ is the measured value for data point ‘*i*’ (Chilès and Delfiner, 2012).

Mapping phenotype landscapes onto structure—Phenotype landscapes built based on a sparse collection of input variants contain experimental or clinical information that predict the full range of values describing function (based on the y- and z axis metrics) for the entire polypeptide sequence (x axis). To map the function predictions onto structure, we assign the prediction value with lowest standard deviation (i.e., 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 at atomic resolution. This collection of all possible functional structure states is referred to as CFTR functional structure. The y axis feature is always depicted as ball size; the z axis feature is depicted as ball color and the prediction confidence (i.e., the contour intervals reflecting standard deviation) is shown as ball transparency. All the atomic resolution structure presentations were produced with the software of PyMOL.

Data requirements for VSP—Sampling data input required for reliable Kriging or GP prediction not only depends on the sample size (number of boreholes/locations) but also depends on the spatial distribution of the samples. Thus, there are a number of considerations in deciding the number of variants and their associated function features required to generate a high confidence value molecular range in the variogram to carry out the phenotype landscape prediction using VSP. The number of datapoints in conventional Kriging have ranged from as little as 20-30 in some geophysical applications to analyses predicting a requirement for 150 datapoints. A rule of thumb in Kriging to allow statistical testing is to have a sample size above > 50 (Kerry and Oliver, 2007), although this number can be impacted based on the method of variogram generation (method of moment (MoM) or residual maximum likelihood (REML). In the case of Ordinary Kriging REML is used, requiring fewer datapoints (Kerry and Oliver, 2007).

Validation of VSP based on the 63 variant dataset used in the current study suggests that that number is sufficient to predict with high confidence (top 25%) values within the molecular range that span the entire polypeptide sequence for a large protein such as CFTR. By using K-fold validation (Figure S1G), we found the prediction accuracy keeps stable until the number of training data points drops below ~50, consistent with the empirical rule of ~50 data points and above recommended in geostatistical studies (Kerry and Oliver, 2007).

Furthermore, when we applied the VSP approach to variation in the BRCA1 RING domain to functional readouts using either 1747 deep scanning generated variants or 62 human variants observed in the general population and patient tumor samples (Starita et al., 2015), we found that the VSP model based on the 62 human variants (Figure S7B, Pearson's r = 0.57) more effectively captures the predictive power in a leave-one-out cross validation when compared to input of data from thousands (1747) arbitrary variants (Figure S7A, Pearson's r = 0.46). When predicting the E3 ligase activity of human BRCA1 variants, the output of VSP, using either 1747 DMS variants (Pearson's r = 0.61) or 62 human variants (Pearson's r = 0.57) as input data in a leave-one-out cross-validation, are significantly better than other prediction tools, such as PolyPhen-2 (Adzhubei et al., 2010) (Pearson's r = 0.15), SIFT (Kumar et al., 2009) (Pearson's r = 0.28) and CADD (Kircher et al., 2014) (Pearson's r = 0.26), as well as Envision (Gray et al., 2018) (Pearson's r = 0.38) that is trained with the DMS datasets together with sequence and/or structural properties (Figure S7C).

Given that most disease genes annotated to date have > 50 missense variants (Landrum et al., 2016), many of which are captured in the GenomAD database (Lek et al., 2016), the VSP method should be valid across many disease states- the limitation being the availability of function datasets for the y- and z axis coordinates. The latter issue has been discussed recently (Manolio et al., 2017; Starita et al., 2017) highlighting the need for a change in bench and clinical experimental design from a unidimensional protocols focusing on a single sequence variant to multidimensional (multiplexed) protocols driven by assays using > 50 variants combined with open access to clinical data such as ClinVar (Landrum et al., 2016) using standardized formats (Manolio et al., 2017; Starita et al., 2017) to invoke lessons learned from the population.

QUANTIFICATION AND STATISTICAL ANALYSIS

VSP prediction validation—The statistical validation methods to assess the performance of the VSP strategy used in this study include a leave-one-out cross-validation, k-fold validation and validation by an external dataset. The default validation method is leave-one-out cross-validation because of small sample size modeling. In the leave-one-out cross validation (Figures 5B, S2F, S2I, S5C, S5D, S5G, S5H, S6A, S6D, S6G, S6I, and S6J), all data are initially used to build the molecular variogram and geostatistical models. 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 to the measured value to generate the Pearson's r-value and its associated p value (ANOVA test).

For the k-fold cross-validation (Figure S1G), samples are randomly partitioned into $k = 63$, 20, 10, 5, 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 Figure S1G. The cross-validation process is repeated *k* times and every set is used as validation once. The prediction of each sample is collected. For $k < 63$, the partition process is repeated 5 times and the averaged Pearson's r and p value of the correlation between predicted value and actual value is reported.

For the external dataset validation of ClCon prediction (Figure S1F), we considered the results of 16 CF variants from a separate study (Van Goor et al., 2014; Yu et al., 2012) that were not used for training (Sosnay et al., 2013). Predicted z values were generated by feeding the model with x- and y- values, and subsequently compared to the observed values by Pearson's correlation analysis and p value calculation (ANOVA test).

For the external dataset validation of FEV1 and SC response to Ivacaftor (Figures S5A and S5B), we fed ClCon measurements determined by cell-based assays in the absence or presence of Ivacaftor (Van Goor et al., 2014; Yu et al., 2012) into the FEV1 or SC (z axis) phenotype landscapes (Figure 5A, upper left quadrant, panels 12 and 17). Although these phenotype landscapes were built on the input variant's phenotypes in basal state, the diverse phenotype relationships for the whole collection of fiduciary variants, when interpreted by VSP, can report the dynamic response range of the phenotype value for each variant as an output. Using as input ClCon values measured in absence or presence of Ivacaftor, the projected output clinical values predicted in response to Ivacaftor are subsequently compared to the observed response for the patients from clinical trial datasets (De Boeck et al., 2014; McGarry et al., 2017; Moss et al., 2015; Ramsey et al., 2011) (Figures S5A and S5B). 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: $\omega_i = \left(1 \Big/ \sigma_i^2\right)$ where $\boldsymbol{\sigma_i}$ is the error for *i*. All quantitative correlation analyses and p value calculations were performed using the software Originpro 2016. A p value < 0.05 was considered to indicate statistical significance

DATA AND SOFTWARE AVAILABILITY

Key input datasets can be downloaded from Mendeley Data at:https://data.mendeley.com/datasets/8d7w8963rb/3.

CFTR—The datasets comprising trafficking and chloride conductance measurements of 63 CF variants used to build the phenotype landscape in Figure 2 is from the reference (Sosnay et al., 2013). The dataset used in Figure 3 is from different references (Van Goor et al., 2014; Yu et al., 2012) given the need for the Ivacaftor input data. The clinical data presented in Figure 4 and Figures 5A and5B are from reference (Sosnay et al., 2013). Sweat Chloride (SC) and Forced Expiratory Volume in 1 s (FEV1) values are the average value for all the patients carrying the variant in *trans* with a known CF-causing variant previously shown to have minimal residual function as indicated in reference (Sosnay et al., 2013). Pseudomonas burden (PB) and pancreatic insufficiency (PI) are percentage of patients that are pancreatic insufficient (insulin deficient) or Pseudomonas infected, respectively (Sosnay et al., 2013). All the function or clinical values in Figure 4 and Figures 5A and 5B are normalized to F508del (set as 0) or WT (set as 1) to make them comparable. For the clinical trial results used in Figures S3D, S3E, S5A, and S5B, the FEV1 and SC measurements of patients with G178R, S549N, S549R, G970R, G1244E, G1349D, S1251N and S1255P after Ivacaftor treatment are from reference (De Boeck et al., 2014). The clinical trial results for G551D are from reference (Ramsey et al., 2011). The clinical trial results for R117H are from reference (Moss et al., 2015). The values for R334W, G85E and A455E are from reference (McGarry et al., 2017). This study did not report the exact measurements of FEV1 but stated that none of the subjects showed significant change in FEV1 measurement (McGarry et al., 2017), so we set the FEV1 change of the three variants as ‘0’. The exact SC values for these patients were reported in this study and are used in Figures S3D and S5B.

APP—The clinical classification of APP variants is obtained from ClinVar and ALZFORUM (https://www.alzforum.org/). ClinVar and ALZFORUM classify the variants as ‘Not pathogenic or benign’, ‘likely benign’, ‘likely pathogenic’, ‘pathogenic’ and ‘Variants of uncertain significance (VUS)’. Here, 45 APP variants with clear clinical classification were used as input data. We set ‘Not pathogenic or benign’ as 1, likely benign’ as 0.66, ‘likely pathogenic’ as 0.33, and ‘pathogenic’ as 0 to generate the output ‘APP pathogenicity’ (APP^{path})-phenotype landscape.

PS1—The level of Aβ-40 and Aβ-42 generated by PS1 variants and the AOs of FAD patients associated with each PS1 variant were obtained from (Sun et al., 2017). Among 138 characterized PS1 variants, 42 variants could not be used to generate the Aβ-42/Aβ-40 ratio due to undetectable levels of Aβ-42 and/or Aβ-40; six variants do not have reported AO; one variant (DE9) lacks exon 9. The remaining 89 missense PS1 variants were used as input data in the VSP analysis. For Aβ-42/Aβ-40 ratio value, we used $\log_{10}$ transformation as input data format.

Allele frequency—The allele frequency for CFTR and APP is obtained from GnomAD database (http://gnomad.broadinstitute.org/). If a patient variant is not found in GnomAD, to include the variant in VSP analysis, we assigned the allele count for that variant as 0.5 in the context of total 277,264 allele counts to date found in the 138,632 individuals in GnomAD. The corresponding allele frequency value for these variants is 0.00018%. The $\log_{10}$ value of allele frequency is used as input data format.

BRCA1—When applying VSP to the BRCA1 RING domain the deep mutational scanning data (DMS) was from reference (Starita et al., 2015). We used 1747 missense variants that have both BARD binding score and E3 ligase activity measurements. Among them, the data for 62 variants observed in patients, general population and tumor samples listed in reference (Starita et al., 2015) was extracted for separate VSP modeling and evaluation.

Geostatistical software used in this study—Given the practical value of geostatistics in geological, epidemiological, and anthropological efforts, there are many open-source R packages and GUI (Graphical User Interface)-based software for performing analyses. We used R package such as gstat (https://cran.r-project.org/web/packages/gstat/index.html) and

GUI-based software packages such as Gamma Design Software (https://geostatistics.com/), yielding identical results when using the Ordinary Kriging module.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

We thank M. Pique (TSRI) and A.J. Olson (TSRI) for advice on modeling of structures and S. Loguercio (TSRI) for advice on computational approaches. Support was provided by NIH grants HL095524, DK051870, and AG049665 to W.E.B., the Tobacco-Related Disease Research Program (TRDRP) grant TRDRP 23RT-0012 (to W.E.B.), and by Cystic Fibrosis Foundation Therapeutics (CFFT) (to C.W.). W.E.B. is indebted to Ralph Wolfe and Carl Woese who encouraged stepping through the Looking Glass into the world of diversity.

REFERENCES

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, and Sunyaev SR (2010). A method and server for predicting damaging missense mutations. Nat. Methods 7, 248–249. [PubMed: 20354512]

Amaral MD, and Balch WE (2015). Hallmarks of therapeutic management of the cystic fibrosis functional landscape. J. Cyst. Fibros. 14, 687–699. [PubMed: 26526359]

Anfinsen CB (1973). Principles that govern the folding of protein chains. Science 181, 223–230. [PubMed: 4124164]

Bai XC, Rajendra E, Yang G, Shi Y, and Scheres SH (2015a). Sampling the conformational space of the catalytic subunit of human γ-secretase. eLife 4, e11182. [PubMed: 26623517]

Bai XC, Yan C, Yang G, Lu P, Ma D, Sun L, Zhou R, Scheres SHW, and Shi Y (2015b). An atomic structure of human γ-secretase. Nature 525, 212–217. [PubMed: 26280335]

Balch WE, Morimoto RI, Dillin A, and Kelly JW (2008). Adapting proteo-stasis for disease intervention. Science 319, 916–919. [PubMed: 18276881]

Baldeiras I, Santana I, Leitão MJ, Gens H, Pascoal R, Tábuas-Pereira M, Beato-Coelho J, Duro D, Almeida MR, and Oliveira CR (2018). Addition of the Aβ42/40 ratio to the cerebrospinal fluid biomarker profile increases the predictive value for underlying Alzheimer's disease dementia in mild cognitive impairment. Alzheimers Res. Ther. 10, 33. [PubMed: 29558986]

Barrett PJ, Song Y, Van Horn WD, Hustedt EJ, Schafer JM, Hadzise-limovic A, Beel AJ, and Sanders CR (2012).The amyloid precursor protein has a flexible transmembrane domain and binds cholesterol. Science 336, 1168–1171. [PubMed: 22654059]

Carter DA, Desmarais E, Bellis M, Campion D, Clerget-Darpoux F, Brice A, Agid Y, Jaillard-Serradt A, and Mallet J (1992). More missense in amyloid gene. Nat. Genet. 2, 255–256. [PubMed: 1303275]

Chilès J-P, and Delfiner P (2012). Geostatistics: Modeling Spatial Uncertainty, Second Edition (John Wiley &Sons).

Collaco JM, Blackman SM, Raraigh KS, Corvol H, Rommens JM, Pace RG, Boelle PY, McGready J, Sosnay PR, Strug LJ, et al. (2016). Sources of variation in sweat chloride measurements in cystic fibrosis. Am. J. Respir. Crit. Care Med. 194, 1375–1382. [PubMed: 27258095]

Cutting GR (2015). Cystic fibrosis genetics: from molecular understanding to clinical application. Nat. Rev. Genet. 16, 45–56. [PubMed: 25404111]

Darwin C (1859). On the Origin of Species by Means of Natural Selection (J. Murray).

De Boeck K, Munck A, Walker S, Faro A, Hiatt P, Gilmartin G, and Higgins M (2014). Efficacy and safety of ivacaftor in patients with cystic fibrosis and a non-G551D gating mutation. J. Cyst. Fibros. 13, 674–680. [PubMed: 25266159]

Ellgaard L, and Helenius A (2003). Quality control in the endoplasmic reticulum. Nat. Rev. Mol. Cell Biol. 4, 181–191. [PubMed: 12612637]

Giri M, Zhang M, and Lü Y (2016). Genes associated with Alzheimer's disease: an overview and current status. Clin. Interv. Aging 11, 665–681. [PubMed: 27274215]

Glusman G, Rose PW, Prlić A, Dougherty J, Duarte JM, Hoffman AS, Barton GJ, Bendixen E, Bergquist T, Bock C, et al. (2017). Mapping genetic variations to three-dimensional protein structures to enhance variant interpretation: a proposed framework. Genome Med. 9, 113. [PubMed: 29254494]

Goodfellow I, Bengio Y, and Courville A (2016). Deep Learning (MIT Press).

Gray VE, Hause RJ, Luebeck J, Shendure J, and Fowler DM (2018). Quantitative missense variant effect prediction using large-scale mutagenesis data. Cell Syst. 6, 116–124.e113. [PubMed: 29226803]

Hopf TA, Ingraham JB, Poelwijk FJ, Schärfe CP, Springer M, Sander C, and Marks DS (2017). Mutation effects predicted from sequence co-variation. Nat. Biotechnol. 35, 128–135. [PubMed: 28092658]

Hunter S, and Brayne C (2018). Understanding the roles of mutations in the amyloid precursor protein in Alzheimer disease. Mol. Psychiatry 23, 81–93. [PubMed: 29112196]

Kerry R, and Oliver MA (2007). Comparing smapling needs for variograms of soil properites computed by the mehtod of moents and residual maximum likelihood. Geoderma 140, 383–396.

Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, and Shendure J (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet. 46, 310–315. [PubMed: 24487276]

Kumar P, Henikoff S, and Ng PC (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat. Protoc. 4, 1073–1081. [PubMed: 19561590]

Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Hoover J, et al. (2016). ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res. 44 (D1), D862–D868. [PubMed: 26582918]

LaRusch J, Jung J, General IJ, Lewis MD, Park HW, Brand RE, Gelrud A, Anderson MA, Banks PA, Conwell D, et al.; North American Pancreatitis Study Group (2014). Mechanisms of CFTR functional variants that impair regulated bicarbonate permeation and increase risk for pancreatitis but not for cystic fibrosis. PLoS Genet. 10, e1004376. [PubMed: 25033378]

Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, et al.; Exome Aggregation Consortium (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature 536, 285–291. [PubMed: 27535533]

Liu F, Zhang Z, Csanady L, Gadsby DC, and Chen J (2017). Molecular structure of the human CFTR ion channel. Cell 169, 85–95.e88. [PubMed: 28340353]

Manolio TA, Fowler DM, Starita LM, Haendel MA, MacArthur DG, Biesecker LG, Worthey E, Chisholm RL, Green ED, Jacob HJ, et al. (2017). Bedside back to bench: building bridges between basic and clinical genomic research. Cell 169, 6–12. [PubMed: 28340351]

Masica DL, Sosnay PR, Raraigh KS, Cutting GR, and Karchin R (2015). Missense variants in CFTR nucleotide-binding domains predict quantitative phenotypes associated with cystic fibrosis disease severity. Hum. Mol. Genet. 24, 1908–1917. [PubMed: 25489051]

McGarry ME, Illek B, Ly NP, Zlock L, Olshansky S, Moreno C, Finkbeiner WE, and Nielson DW (2017). In vivo and in vitro ivacaftor response in cystic fibrosis patients with residual CFTR function: N-of-1 studies. Pediatr. Pulmonol. 52, 472–479. [PubMed: 28068001]

Mendoza JL, Schmidt A, Li Q, Nuvaga E, Barrett T, Bridges RJ, Feranchak AP, Brautigam CA, and Thomas PJ (2012). Requirements for efficient correction of ΔF508 CFTR revealed by analyses of evolved sequences. Cell 148, 164–174. [PubMed: 22265409]

Moss RB, Flume PA, Elborn JS, Cooke J, Rowe SM, McColley SA, Rubenstein RC, and Higgins M; VX11–770-110 (KONDUCT) Study Group (2015). Efficacy and safety of ivacaftor in patients with cystic fibrosis who have an Arg117His-CFTR mutation: a double-blind, randomised controlled trial. Lancet Respir. Med. 3, 524–533. [PubMed: 26070913]

Nakamura A, Kaneko N, Villemagne VL, Kato T, Doecke J, Doré V, Fowler C, Li QX, Martins R, Rowe C, et al. (2018). High performance plasma amyloid-β biomarkers for Alzheimer's disease. Nature 554, 249–254. [PubMed: 29420472]

Nelson O, Supnet C, Tolia A, Horré K, De Strooper B, and Bezprozvanny I (2011). Mutagenesis mapping of the presenilin 1 calcium leak conductance pore. J. Biol. Chem. 286, 22339–22347. [PubMed: 21531718]

Pankow S, Bamberger C, Calzolari D, Martínez-Bartolomé S, Lavallíe-Adam M, Balch WE, and Yates JR, 3rd. (2015). ΔF508 CFTR interactome remodelling promotes rescue of cystic fibrosis. Nature 528, 510–516. [PubMed: 26618866]

Proctor EA, Kota P, Aleksandrov AA, He L, Riordan JR, and Dokholyan NV (2015). Rational coupled dynamics network manipulation rescues disease-relevant mutant cystic fibrosis transmembrane conductance regulator Chem. Sci. (Camb) 6, 1237–1246. [PubMed: 25685315]

Rabeh WM, Bossard F, Xu H, Okiyoneda T, Bagdany M, Mulvihill CM, Du K, di Bernardo S, Liu Y, Konermann L, et al. (2012). Correction of both NBD1 energetics and domain interface is required to restore ΔF508 CFTR folding and function. Cell 148, 150–163. [PubMed: 22265408]

Ramsey BW, Davies J, McElvaney NG, Tullis E, Bell SC, Dřevínek P, Griese M, McKone EF, Wainwright CE, Konstan MW, et al.; VX08-770-102 Study Group (2011). A CFTR potentiator in patients with cystic fibrosis and the G551D mutation. N. Engl. J. Med. 365, 1663–1672. [PubMed: 22047557]

Rasmussen CE, and Williams CKI (2006). Gaussian Processes for Machine Learning (MIT Press).

Ratner M (2017). FDA deems in vitro data on mutations sufficient to expand cystic fibrosis drug label. Nat. Biotechnol. 35, 606. [PubMed: 28700545]

Romero PA, Krause A, and Arnold FH (2013). Navigating the protein fitness landscape with Gaussian processes. Proc. Natl. Acad. Sci. USA 110, E193–E201. [PubMed: 23277561]

Shen PS (2018). The 2017 Nobel Prize in Chemistry: cryo-EM comes of age. Anal. Bioanal. Chem. 410, 2053–2057. [PubMed: 29423601]

Sosnay PR, Siklosi KR, Van Goor F, Kaniecki K, Yu H, Sharma N, Ramalho AS, Amaral MD, Dorfman R, Zielenski J, et al. (2013). Defining the disease liability of variants in the cystic fibrosis transmembrane conductance regulator gene. Nat. Genet. 45, 1160–1167. [PubMed: 23974870]

Starita LM, Young DL, Islam M, Kitzman JO, Gullingsrud J, Hause RJ, Fowler DM, Parvin JD, Shendure J, and Fields S (2015). Massively parallel functional analysis of BRCA1 RING domain variants. Genetics 200, 413–422. [PubMed: 25823446]

Starita LM, Ahituv N, Dunham MJ, Kitzman JO, Roth FP, Seelig G, Shendure J, and Fowler DM (2017). Variant interpretation: functional assays to the rescue. Am. J. Hum. Genet. 101, 315–325. [PubMed: 28886340]

Sun L, Zhou R, Yang G, and Shi Y (2017). Analysis of 138 pathogenic mutations in presenilin-1 on the in vitro production of Aβ42 and Aβ40 peptides by γ-secretase. Proc. Natl. Acad. Sci. USA 114, E476–E485. [PubMed: 27930341]

Taylor-Cousar JL, Munck A, McKone EF, van der Ent CK, Moeller A, Simard C, Wang LT, Ingenito EP, McKee C, Lu Y, et al. (2017). Tezacaftor-ivacaftor in patients with cystic fibrosis homozygous for Phe508del. N. Engl. J. Med. 377, 2013–2023. [PubMed: 29099344]

Thiagarajah JR, Donowitz M, and Verkman AS (2015). Secretory diarrhoea: mechanisms and emerging therapies. Nat. Rev. Gastroenterol. Hepatol. 12, 446–457. [PubMed: 26122478]

Torkamani A, Andersen KG, Steinhubl SR, and Topol EJ (2017). High-definition medicine. Cell 170, 828–843. [PubMed: 28841416]

Van Goor F, Yu H, Burton B, and Hoffman BJ (2014). Effect of ivacaftor on CFTR forms with missense mutations associated with defects in protein processing or function. J. Cyst. Fibros. 13, 29–36. [PubMed: 23891399]

Wainwright CE, Elborn JS, Ramsey BW, Marigowda G, Huang X, Cipolli M, Colombo C, Davies JC, De Boeck K, Flume PA, et al.; TRAFFIC Study Group; TRANSPORT Study Group (2015). Lumacaftor-ivacaftor in patients with cystic fibrosis homozygous for Phe508del CFTR. N. Engl. J. Med. 878, 220–231.

Wang X, Matteson J, An Y, Moyer B, Yoo JS, Bannykh S, Wilson IA, Riordan JR, and Balch WE (2004). COPII-dependent export of cystic fibrosis transmembrane conductance regulator from the ER uses a di-acidic exit code. J. Cell Biol. 167, 65–74. [PubMed: 15479737]

Wiseman RL, Powers ET, Buxbaum JN, Kelly JW, and Balch WE (2007). An adaptable standard for protein export from the endoplasmic reticulum. Cell 131, 809–821. [PubMed: 18022373]

Yu H, Burton B, Huang CJ, Worley J, Cao D, Johnson JP, Jr., Urrutia A, Joubran J, Seepersaud S, Sussky K, et al. (2012). Ivacaftor potentiation of multiple CFTR channels with gating mutations. J. Cyst. Fibros. 11, 237–245. [PubMed: 22293084]

Zhang Z, Liu F, and Chen J (2017). Conformational changes of CFTR upon phosphorylation and ATP binding. Cell 170, 483–491.e488. [PubMed: 28735752]

Highlights

- We develop VSP, a Gaussian-process-based approach to interpret genomic diversity
- VSP is based on spatial covariance (SCV) in the genotype-to-phenotype transformation
- SCV uses population genomics to inform individualized phenotypes at atomic resolution
- Phenotype landscapes generated through SCV enable high-definition medicine

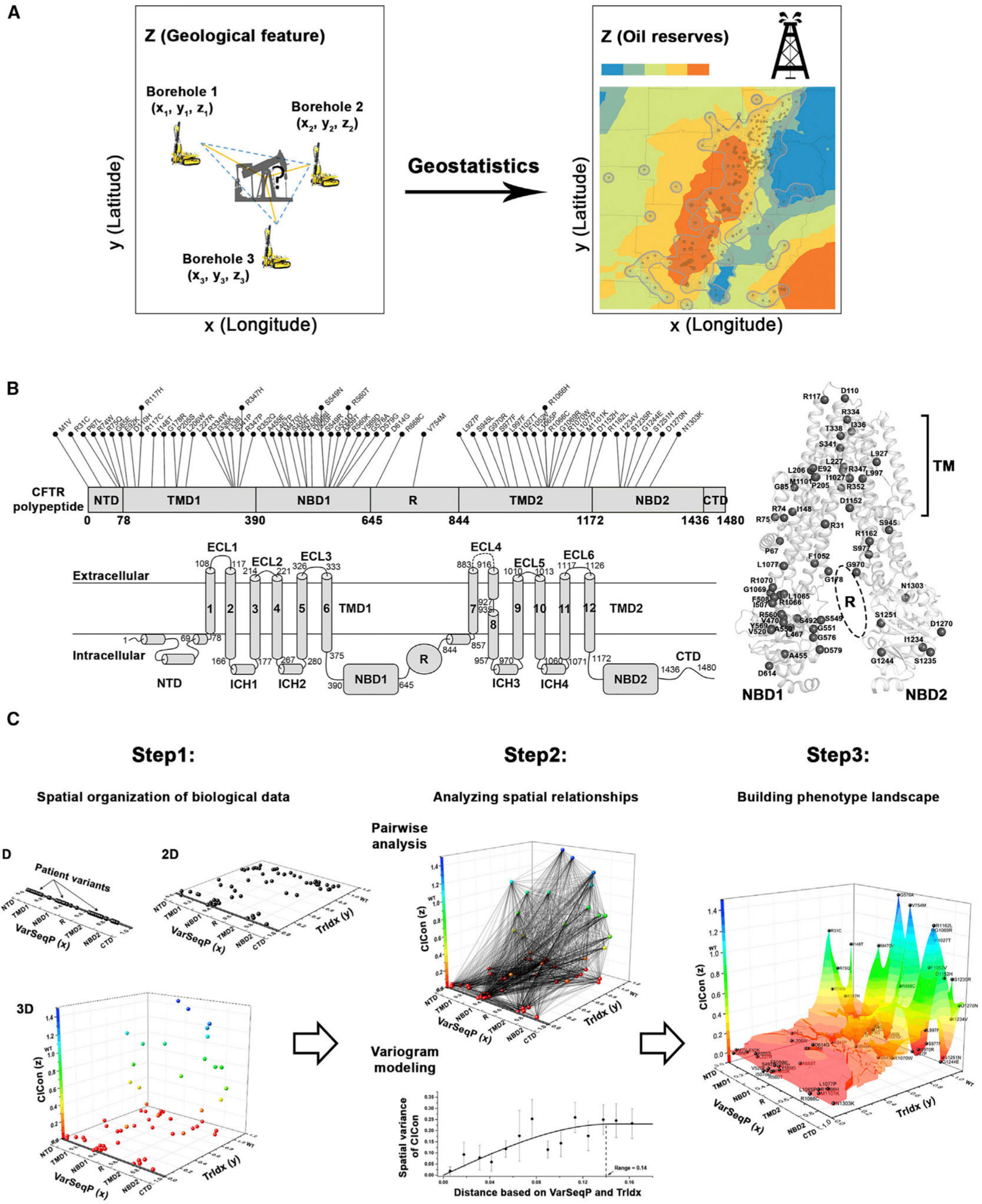


Figure 1. Building Phenotype Landscape Through VSP

(A) A schematic illustrating application of Gaussian-process (GP)-based geostatistics for oil exploration in a geophysical landscape.

(B) CFTR linear, secondary, and 3D structure with CF variants indicated.

(C) Steps for generating the phenotype landscape through VSP (see Video S1).

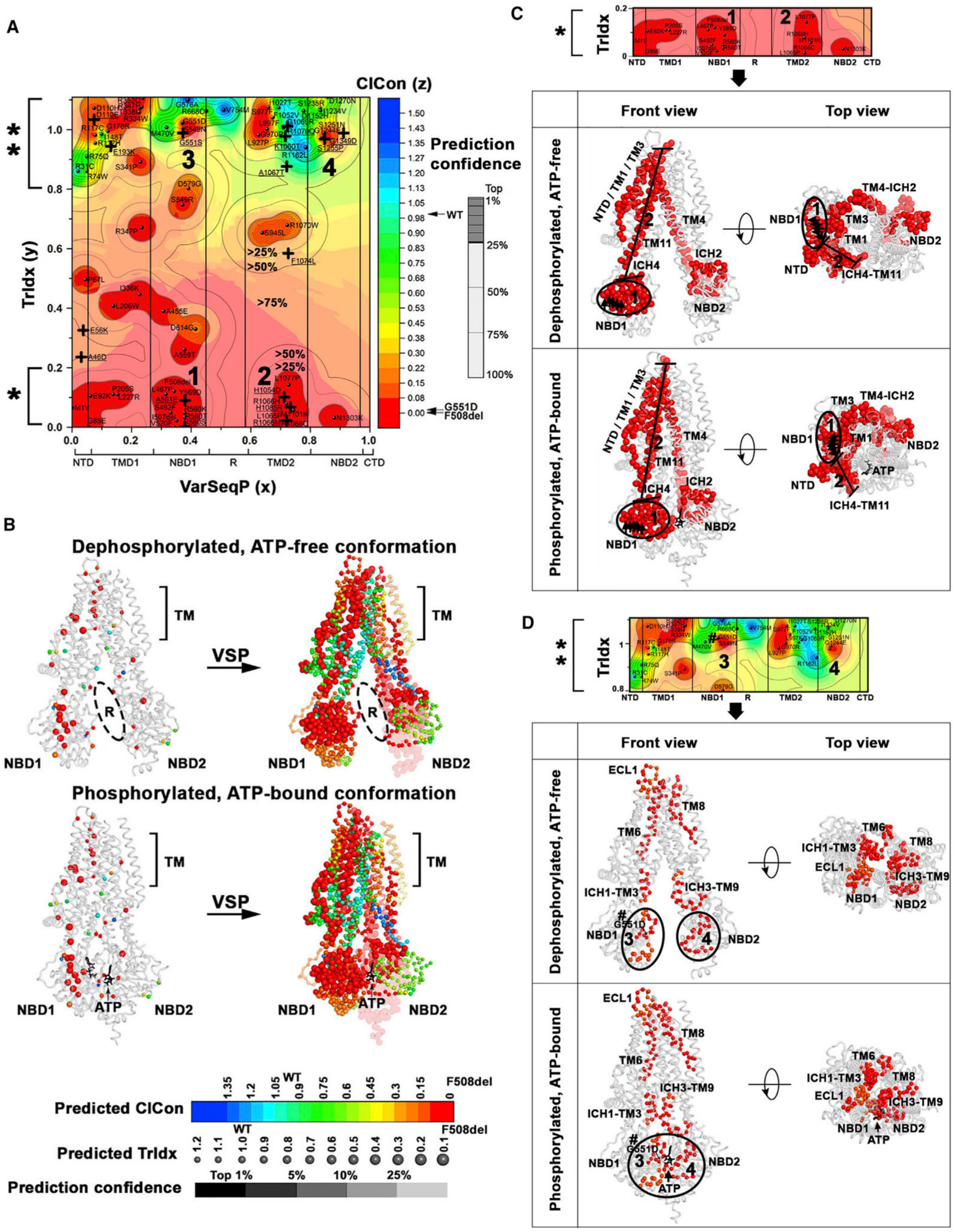


Figure 2. Phenotype Landscape Informs Functional Structure

(A) The VSP-predicted values of ClCon (z axis) relating to TrIdx (y axis) across the entire VarSeqP (x axis) in this Wang-Balch plot is shown as a phenotype landscape overlaid with the confidence contour intervals.

(B) Phenotype landscape is mapped to CFTR structure snapshots (PDB: 5UAK, 5W81) to generate functional structures.

(C) The residues in the functional structure (B) with predicted variants that define low trafficking (TrIdx < 0.2) values in the landscape (top; highlighted by one asterisk in A) are shown as balls in the structural snapshots (bottom). The di-acidic ER exit code of CFTR

(YKDAD) in NBD1 domain is highlighted by black arrows. ATP at the consensus site is shown as black sticks.

(D) The residues in the functional structure (B) with predicted variants that locate on the cell membrane (TrIdx > 0.8) (top; highlighted by two asterisks in A) but with deficient ClCon function (ClCon < 0.15) are shown as balls on the structure snapshots (bottom). The position of G551D is denoted by a number sign.

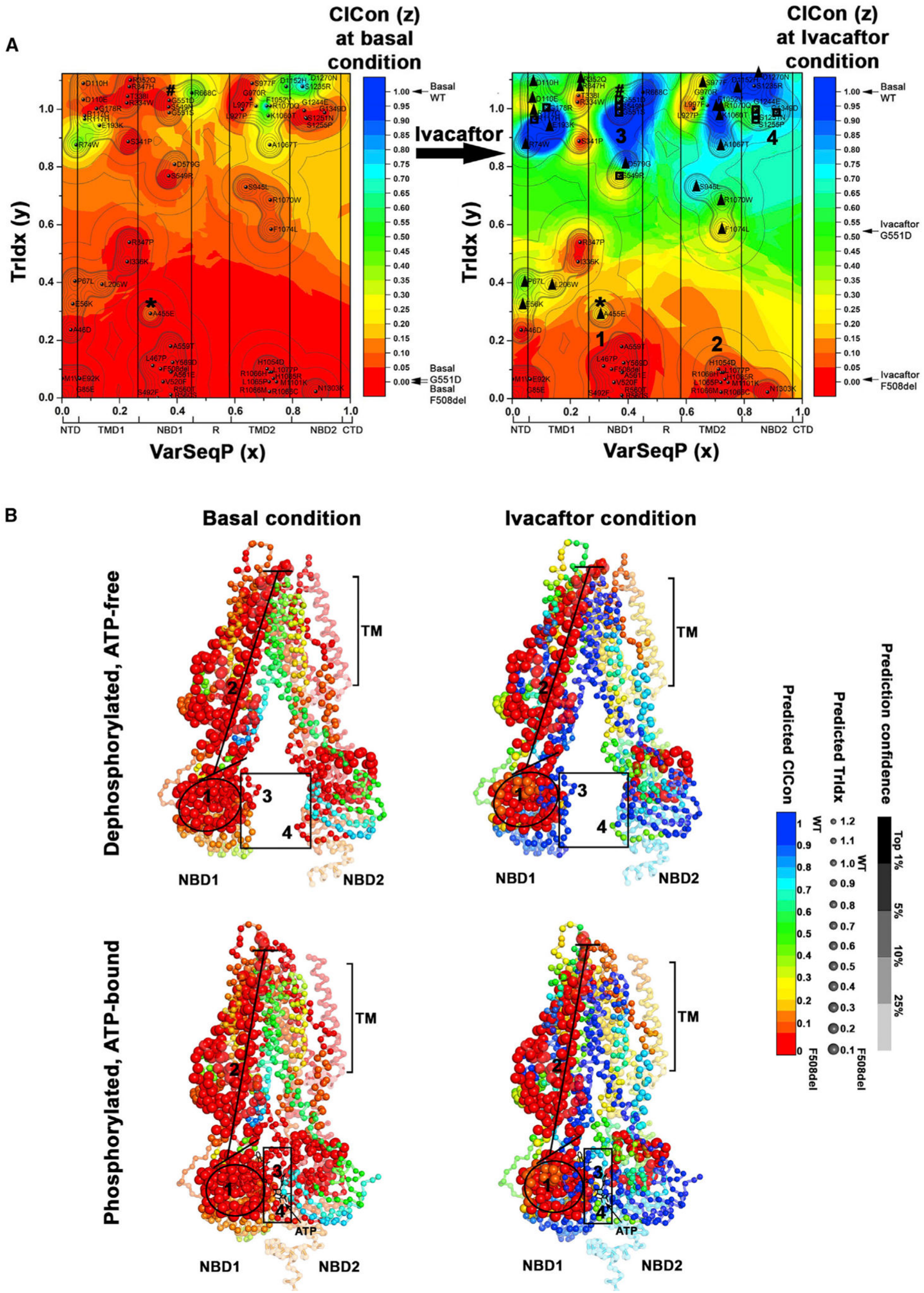


Figure 3. Ivacaftor-Responsive Phenotype Landscapes

(A) The predicted ClCon values (z axis) in the absence (left) or presence of Ivacaftor (right) are shown as phenotype landscapes (see Videos S2 and S3). Top 25% quartile confidence interval of prediction is highlighted by bold contour line. FDA-approved variants for treatment with ivacaftor are highlighted by the square boxes. Variants recently approved based on *in vitro* cell-based data (Ratner, 2017) are highlighted by black triangles. Among them, A455E is highlighted by one asterisk.

(B) Mapping the predicted ClCon on human CFTR structure snapshots to generate therapeutic responsive view of the fold.

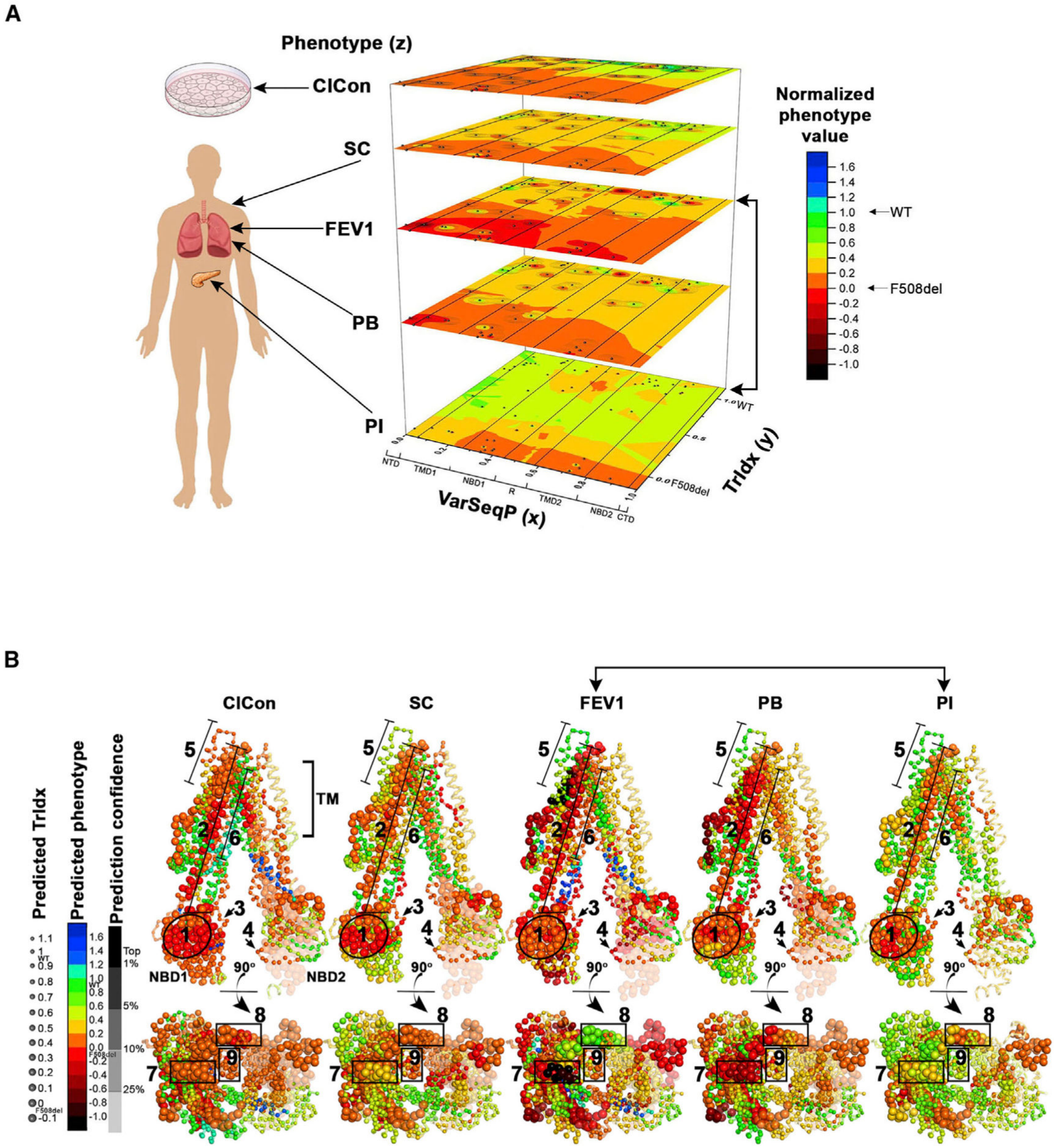


Figure 4. Applying VSP to Clinical Phenotypes

(A) Phenotype landscapes relating the sequence position of variant (x axis) and its cell-based TrIdx (y axis) to the indicated features (z axis): cell-based chloride conductance (ClCon), clinical sweat chloride (SC), clinical forced expiratory volume 1 (FEV1), clinical *Pseudomonas* burden (PB), and clinical pancreatic insufficiency (PI).

(B) Mapping clinical phenotype landscapes on CFTR structure snapshots.

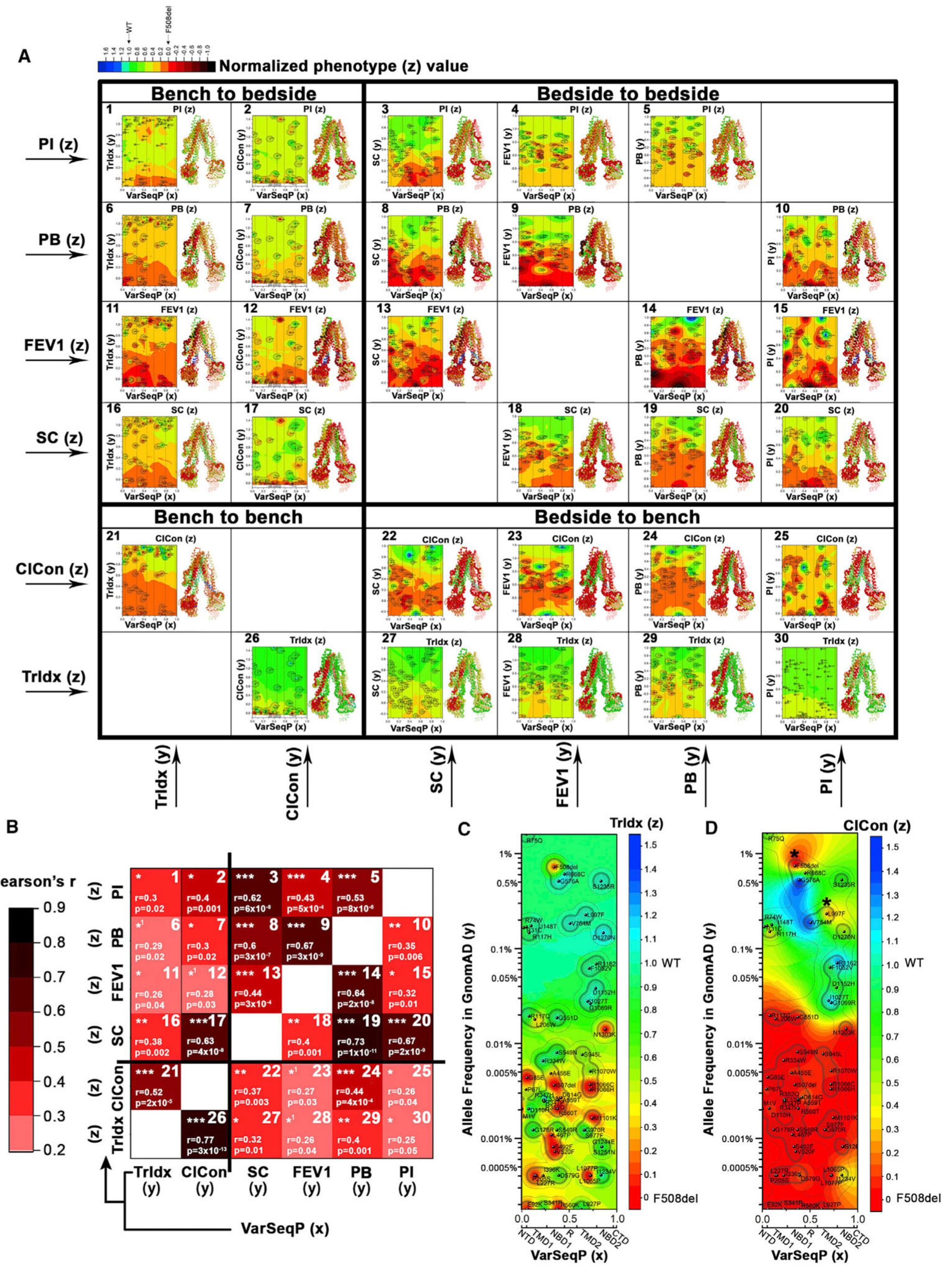


Figure 5. Phenotype Landscapes Linking Bench, Bedside, and Population Genomics

(A) Predicted phenotype landscapes and functional structures that use any two combinations of the indicated cell-based or clinical features as y axis and z axis values.

(B) Leave-one-out cross-validation of phenotype landscapes shown in (A). Pearson's r value is indicated by the pink to dark red color scale; p value is indicated by asterisks (0.01 < *p < 0.05; 0.001 < **p < 0.01; ***p < 0.001; 0.01 <*[1]p<0.05, where V754M is set as an outlier for validation given its variability in phenotype landscapes; Figure S4).

(C and D) Phenotype landscapes relating CFTR variants (x axis) and the allele frequency in GnomAD (y axis) to TrIdx (C) or ClCon (D).

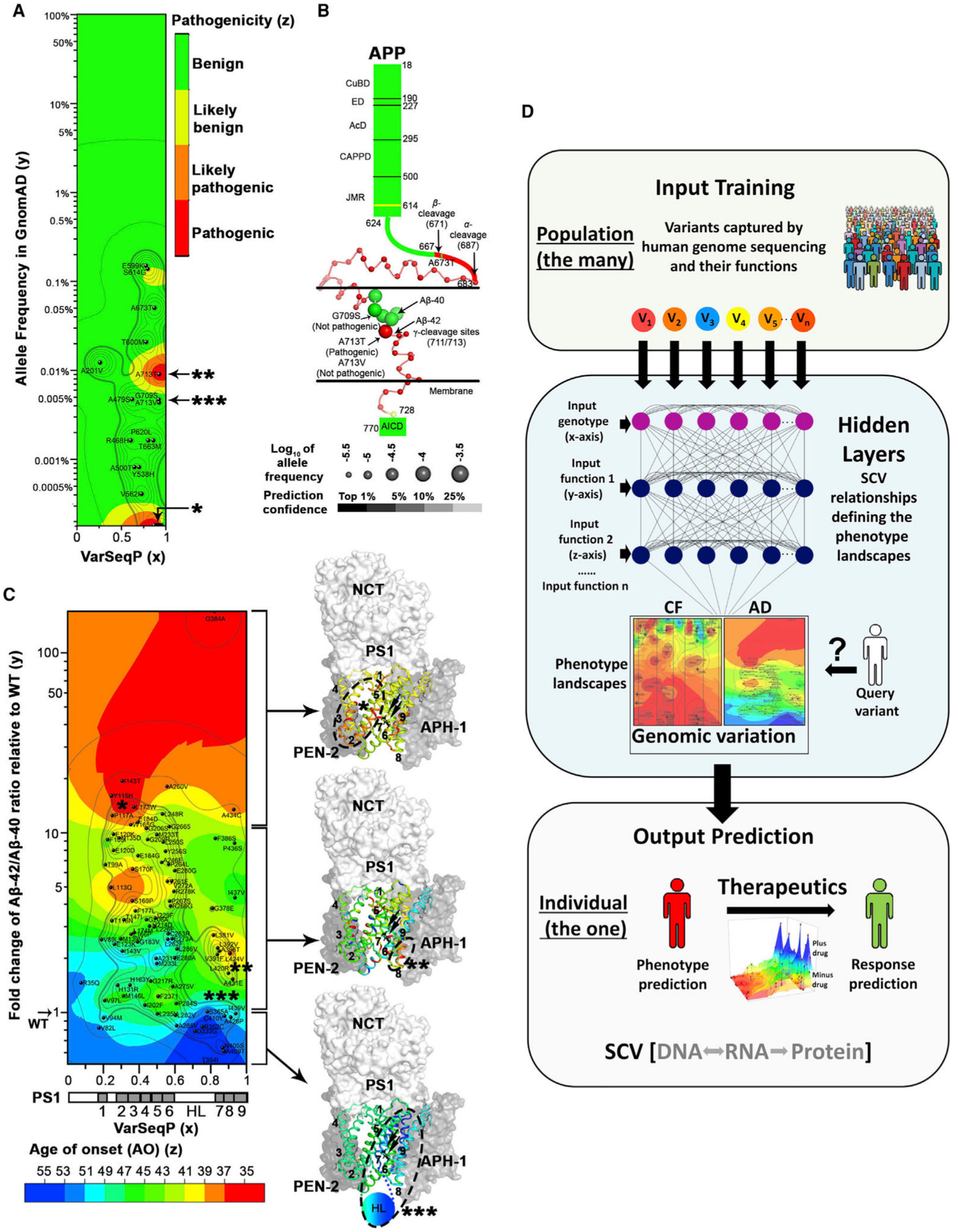


Figure 6. Applying VSP to APP and PS1

(A) Phenotype landscape relating APP variants (x axis) and the GnomAD allele frequency (y axis) to the clinical presentation of Alzheimer's disease (z axis).

(B)The highest confidence prediction of the phenotype generated by VSP is assigned to each residue and mapped on APP schematic structure with atom resolution in the region of (683–728) (PDB: 2LP1). For position of G713, only the clinical value of G713T is assigned for structural presentation, while the clinical values of both G713T and G713V can be captured in the landscape (A).

(C) Phenotype landscape relating PS1 variants (x axis) and the Aβ-42/Aβ-40 ratio relative to WT (y axis) to mean age of onset (AO) from FAD patients (z axis). The landscape is divided into 3 sections (brackets) based on y axis thresholds: below 1, $1 < y < 10$, and $y > 10$. Each landscape section is mapped on the structural snapshot of γ-secretase complex (PDB: 5FN2) separately by assigning predicted AO with highest confidence to each residue of PS1. The SCV clusters (25% confidence level) in each section of the landscape with close sequence-to-function-to-structure relationships are highlighted by one asterisk, two asterisks, and three asterisks, respectively, and the corresponding functional structure projections are highlighted by dashed ovals. The TMs are numerically labeled in the structure and the two catalytic aspartate residues in TM 6 and 7 are shown as black sticks and highlighted by arrows.

(D) Cartoon illustrating VSP. GP-based SCV relationships suggest a matrix-based flow of information in central dogma facilitates the genotype to phenotype transformation (lower panel, SCV[DNA<->RNA->Protein]) where the genome tells the proteome how to shape; the proteome tells the genome how to evolve.

KEY RESOURCES TABLE

RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
TrIdx and ClCon measurements of CF variants (Input data for Figures 1 and 2)	Sosnay et al., 2013	https://data.mendeley.com/datasets/8d7w8963rb/3
Ivacaftor response of CF variants (Input data for Figure 3)	Van Goor et al., 2014; Yu et al., 2012	https://data.mendeley.com/datasets/8d7w8963rb/3
Tissue specific measurements of CF variants (Input data for Figures 4, 5A, and 5B)	Sosnay et al., 2013	https://data.mendeley.com/datasets/8d7w8963rb/3
Allele frequency and TrIdx/ClCon measurements of CF variants (Input data for Figures 5C and 5D)	GnomAD; Sosnay et al., 2013	https://data.mendeley.com/datasets/8d7w8963rb/3
Allele frequency and pathogenicity of APP variants (Input data for Figures 6A and6B)	GnomAD; ClinVar; ALZFORUM	https://data.mendeley.com/datasets/8d7w8963rb/3
Measurements of PS1 variants (Input data for Figure 6C)	Sun et al., 2017	https://data.mendeley.com/datasets/8d7w8963rb/3
Software and Algorithms		
GS+ (Version 10)	Gammadesign software	https://geostatistics.com/index.aspx
Gstat (1.1-6)	R-package	https://CRAN.R-project.org/package=gstat
Originpro 2016	Originlab	https://www.originlab.com; RRID: SCR_015636
Pymol 1.8.6.0	Schrodinger, LLC	https://pymol.org/2/; RRID: SCR_000305