Supplement for:

"Missense variants in CFTR nucleotide-binding domains predict quantitative phenotypes associated with cystic fibrosis disease severity"

David L. Masica, Patrick R. Sosnay, Karen S. Raraigh, Garry R. Cutting, Rachel Karchin

Supplemental Results

Endophenotype	R^2	Correlation	P-value
Sweat [Cl^-]	0.04	0.20	0.25
P.I.	0.01	-0.08	0.64
Pseudo.	0.01	0.11	0.52
[Cl^-] Cond.	0.06	-0.24	0.15
C/(C + B)	0.09	-0.29	0.08
FEV_1 %pred.	0	0.01	0.95

Table S1: Correlation of POSE score and six endophenotypes for CFTR transmembrane-domain (TMD) variants. *Sweat [Cl^-]* is the mean sweat chloride for patients with the variant; *P.I.* is the percentage of patients displaying pancreatic insufficiency; *Pseudo.* is the percentage of patients with *Pseudomonas aeruginosa* infection; *[Cl^-] Cond.* is the mean chloride conductance for cells expressing the CFTR variant; *C/(C + B)* estimates the fraction of properly processed ("mature") CFTR protein; *FEV_1 %pred* is the mean lung function as a percent of wild type. Increasing *Sweat [Cl^-]*, *P.I.*, and *Psuedo* are each associated with increasing CF severity. For *[Cl^-] Cond.*, *C/(C + B)*, and *FEV1 %pred*, decreasing values are associated with increasing CF severity.

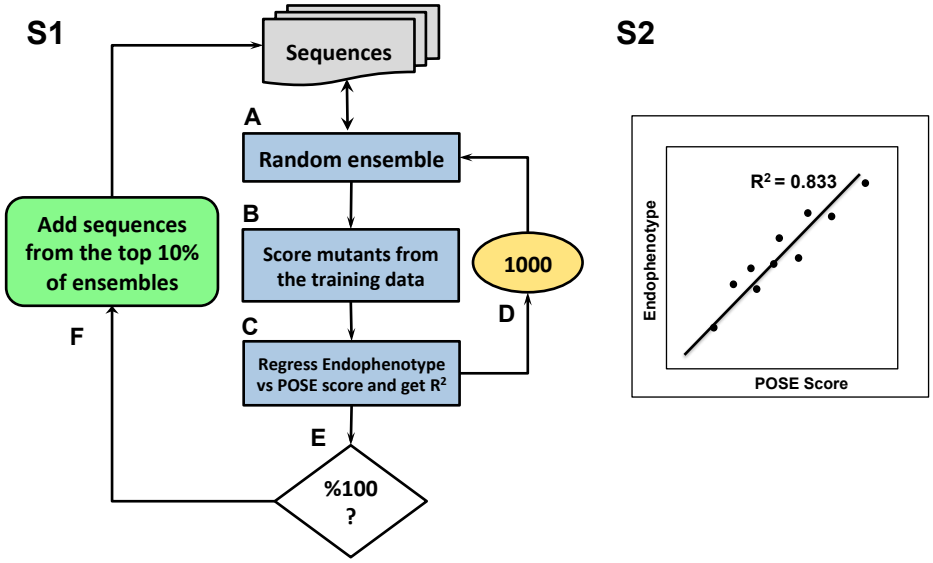
Endophenotype	R^2	Correlation	P-value
Sweat [Cl^-]	0.15 (0.56)	0.38 (0.75)	2.8×10^{-3} (1.5×10^{-4})
P.I.	0.07 (0.41)	0.26 (0.64)	4.9×10^{-2} (2.4×10^{-3})
Pseudo.	0.04 (0.14)	0.20 (0.38)	1.3×10^{-1} (1.0×10^{-1})
[Cl^-] Cond.	0.46 (0.55)	-0.68 (-0.77)	3.7×10^{-9} (1.9×10^{-4})
C/(C + B)	0.13 (0.24)	-0.35 (-0.49)	5.8×10^{-3} (2.7×10^{-2})
FEV_1 %pred.	0.02 (0.10)	-0.14 (-0.31)	2.9×10^{-1} (1.9×10^{-1})

Table S2: Correlation of POSE score and six endophenotypes for all CFTR variants from this study. Values in parentheses were derived using predictions for NBD variants only, but where predictions resulted from training on all 59 variants. See Table S1 for a description of each endophenotypic measurement.

Supplemental Materials and Methods

POSE implementation for this study

For this study we ran the phenotype-optimized sequence ensemble (POSE) algorithm[1] in endophenotype (ePOSE) mode. POSE calculations are highly customizable, and the parameterizations given below are defaults that can be easily changed with user-defined arguments. The POSE algorithm is freely available for non-profit use and includes a detailed manual with worked examples. To download the POSE algorithm please visit: http://karchinlab.org/apps/appPose.html.



Supplementary Figures S1 and S2: Flow chart of POSE implementation used for this study.

POSE begins by selecting a random ensemble of 150 or fewer sequences from an initial multiple sequence alignment (Figure S1A); see *Materials and Methods*, *CFTR Sequences* in the main text for a description of the multiple sequence alignment (MSA) used for this study. Next, the algorithm scores (see *Score Function*, below) each mutation from a user-provided training set of mutations, where a continuous-valued endophenotypic measurement is provided with each mutation in the training set (Figure S1B). Then, POSE performs a linear regression of POSE scores vs. endophenotypic measurements and computes the R^2 (Figures S1C and S2). This process of randomly selecting sequences, scoring the training set of mutations, and computing R^2 for the linear regression of POSE scores vs. endophenotypes is repeated for 1000 iterations (Figure S1D). After every the 100 iterations, the top-performing 1% of sequences are appended

to the MSA (Figures S1E and F), where top-performing sequences are those belonging to randomly selected sequence ensembles that maximized R^2. Therefore, as the algorithm progresses through many 100s of iterations, the MSA becomes enriched for sequences that optimize the correlation between the POSE scores and endophenotypes. Once 1000 iterations are complete, the top-performing sequence ensemble that was sampled during this iterative optimization process is saved for use of scoring new, holdout mutations. In contrast to classic MSAs used in missense mutation function prediction, these POSEs can contain as many as 100+ copies of a single ortholog or paralog sequence (i.e., the influence of individual sequences on mutation discrimination is differentially weighted).

For this work, the POSE algorithm trained using 19 of 20 CFTR mutations, and prediction was made on the remaining mutation; this process was repeated for each mutation. This leave-one-out cross-validation strategy was applied to each of the six endophenotypes separately. See *Materials and Methods*, *Endophenotypic Data* in the main text for a description of each endophenotype considered in this study.

Score Function

The POSE score function considers both biophysical molecular properties and evolutionary distance of related proteins to a target disease protein. The score (S) for a specific amino acid residue at a specific position in the alignment is calculated as the sum of an amino acid conservation score (S_{aa}), biophysical properties conservation score (S_{prop}), and molecular weight conservation score (S_{mw}; Equation S1). A score can be calculated for any amino acid residue at any alignment column, regardless of whether the amino acid appears in that column.

$$S = S_{aa} + S_{prop} + S_{mw} \quad \text{(Equation S1)}$$

We weight the contribution of each sequence in the alignment by an estimate of its evolutionary distance from the target sequence. Each sequence contributes an amount proportional to its sequence identity, relative to the target. Thus, to calculate the amino acid conservation score (S_{aa}) for a specific amino acid, the algorithm sums the sequence identities (Ω) of all sequences in the alignment that harbor that specific amino acid (n_{aa}), at the position being scored (Equation S2).

$$S_{aa} = \frac{1}{L}\sum_{i=1}^{n_{aa}} \Omega_i \quad \text{(Equation S2)}$$

$$L = \sum_{i=1}^{N} \Omega_i \quad \text{(Equation S3)}$$

In Equation S2, L is a normalizing constant, which is the sum of sequence identities for all N sequences in the alignment (Equation S3). Similarly, to calculate the biophysical properties conservation score for a specific amino acid, the algorithm sums the sequence identities of all sequences in the alignment that harbor that specific property, at the position being scored (Equation S4; see below and Table S1); the biophysical properties conservation score is calculated for each property that defines the amino acid being scored (m_{prop}). This score is additionally normalized by the total number of properties defining the amino acid being scored (n_{prop}). The additional normalization constant assures that amino acids with a greater number of properties are not artificially favored relative to amino acids with fewer properties. The molecular-weight conservation score (S_{mw}) is calculated as the probability of observing the molecular weight of the amino acid being scored, given the probability distribution function defined by the identity-weighted molecular weights from all other amino acids in the column (Equation S5). The score for an amino acid substitution (S_{Sub}; Equation S6) is this difference between the wild type (S_{WT}) and mutant amino acids scores (S_{Mut}), where S_{WT} and S_{Mut} are each calculated from Equation 1.

$$S_{prop} = \frac{1}{n_{prop}} \sum_{i=1}^{n_{prop}} \frac{1}{L} \sum_{j=1}^{m_{prop}} \Omega_{ij} \qquad \text{(Equation S4)}$$

$$S_{mw} = \Pr\left[mw_{aa} | PDF(mean_{mw}, std_{mw})\right] \quad \text{(Equation S5), where}$$

mw_{aa} is the molecular weight of the amino acid being scored, and $mean_{mw}$ and std_{mw} are the identity-weighted mean and standard deviation of molecular weights for the column

$$S_{Sub} = S_{WT} - S_{Mut} \quad \text{(Equation S6)}$$

Each component of a scored amino acid (S_{aa}, S_{prop}, and S_{mw}; see Equation S1) can take on a value between 0 and 1.0, where an increasing value means increasing similarity between the amino acid being scored and those in the corresponding column in the alignment. Therefore, 3.0 is the maximum score any amino acid can have for Equation S1. Thus, Equation S6 can vary from -3.0 to 3.0, where a score of -3.0 indicates the most favorable substitution possible (i.e., mutant has a higher score than the wild-type amino acid) and 3.0 is the most disfavored substitution possible (i.e., wild-type amino acid scores higher than the mutant); a score of 0.0 is equivalent to wild type.

When available, 3D coordinates for the protein of interest can also be included for scoring mutations. POSE estimates the normalized degree of burial at each residue in a

protein structure using a novel calculation. This is accomplished by calculating all pairwise residue-residue distances, and for each residue counting the number of neighboring residues within 10 Å; distances are calculated from each residue's center of geometry (based on atomic coordinates). Then, a relative degree of burial is calculated by normalizing all residues relative to the residue with the greatest number of neighbors (i.e., max degree of burial = 1.0, with decreasing values corresponding to decreasing burial). Figure S3 illustrates the normalized residue-burial calculation for the CFTR homology model used in this study. If POSE is provided a 3D protein structure, Equation S7 gives the final POSE score for any given amino acid substitution. Here, γ is the normalized residue burial, and modulates the strength of effect calculated in Equation S6.

$$S_{Sub} = \gamma(S_{WT} - S_{Mut}) \quad \text{(Equation S7)}$$

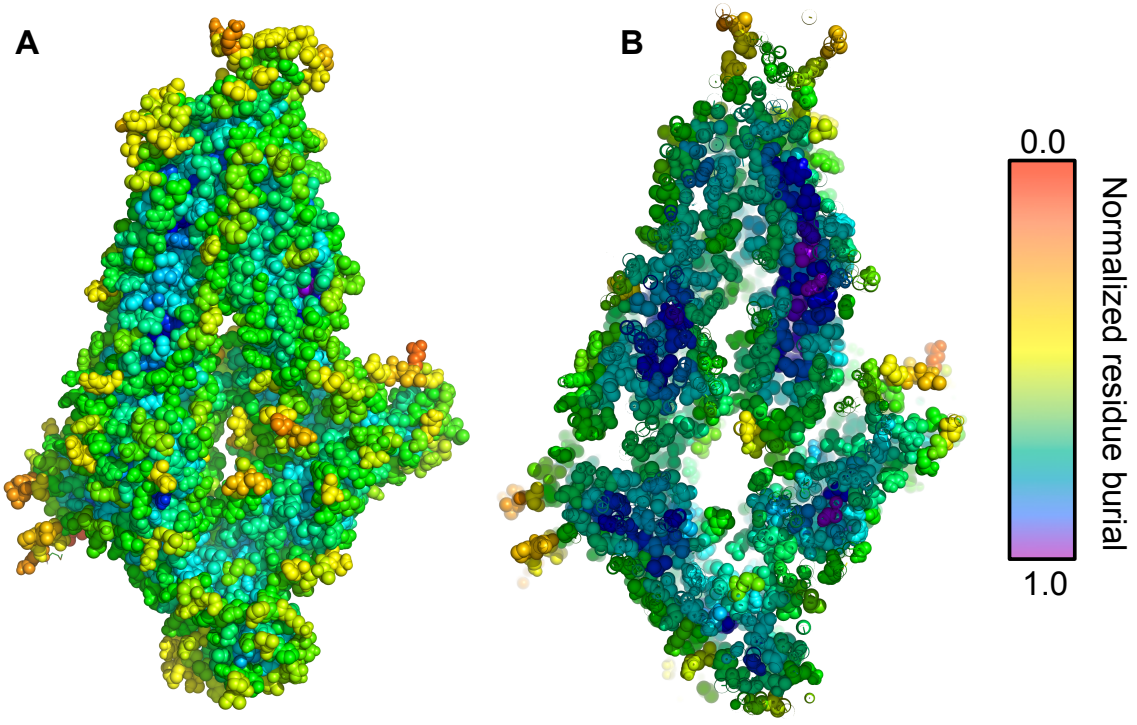


Figure S3: POSE-calculated normalized residue burial for the CFTR homology model. In *A*, POSE normalized residue burial is superimposed on a 3D homology model of CFTR. *B* shows the same structure "clipped" so that the core residues are visible, showing the expected increasing burial in the core of the protein. The CFTR homology was published in Mornon *et al*.[2]

Amino acid	Chemistry	Weight
ALA	Aliphatic	89.0
CYS	Cysteine	121.0
ASP	Acceptor, Negative	132.0
GLU	Acceptor, Negative	147.0
PHE	Aromatic	165.0
GLY	Glycine	75.0
HIS	Donor, Positive	155.0
ILE	Aliphatic	131.0
LYS	Donor, Positive	146.0
LEU	Aliphatic	131.0
MET	Aliphatic	149.0
ASN	Donor, Acceptor	132.0
PRO	Proline	115.0
GLN	Donor, Acceptor	146.0
ARG	Donor, Positive	174.0
SER	Donor	105.0
THR	Donor	119.0
VAL	Aliphatic	117.0
TRP	Aromatic	204.0
TYR	Aromatic, Donor	181.0

Table S1: Complete list of attributes used for scoring. The score function considers *Amino acid* conservation, the conservation of specific biochemical *Properties*, and conservation of molecular *Weight* (g/mol).

Supplementary Table S1 shows the amino acid properties and molecular weights used by the score function. The properties include standard biophysical properties associated with the side chains of each of the 20 naturally occurring amino acids (hydrogen bond donating or accepting, negatively or positively charged, etc.). Histidine is unique because the side chain titrates at physiological pH, resulting in an environment-dependent charge and hydrogen bond capacity; therefore, we define histidine uniquely (see Table S1). Similarly, cysteine is unique in its ability to form inter-side chain covalent bonds and glycine is unique because it lacks a side chain; therefore, we define cysteine and glycine uniquely. And, proline mutation has a unique capacity to perturb well-defined protein secondary structure because it lacks a backbone amide proton and has a restricted backbone Φ torsion angle; therefore, we define proline uniquely. Molecular weights are in units of g/mol.

Supplemental references

1) Masica, David L., et al. "Phenotype-optimized sequence ensembles substantially improve prediction of disease-causing mutation in cystic fibrosis." *Human mutation* 33.8 (2012): 1267-1274.

2) Mornon, Jean-Paul, Pierre Lehn, and Isabelle Callebaut. "Molecular models of the open and closed states of the whole human CFTR protein." *Cellular and Molecular Life Sciences* 66.21 (2009): 3469-3486.