Biophysical Journal
Article

Biophysical Society

Molecular Mechanisms of Macular Degeneration Associated with the Complement Factor H Y402H Mutation

Reed E. S. Harrison[1] and Dimitrios Morikis[1,*]
[1]Department of Bioengineering, University of California, Riverside, Riverside, California

ABSTRACT A single nucleotide polymorphism, tyrosine at position 402 to histidine (Y402H), within the gene encoding complement factor H (FH) predisposes individuals to acquiring age-related macular degeneration (AMD) after aging. This polymorphism occurs in short consensus repeat (SCR) 7 of FH and results in decreased binding affinity of SCR6-8 for heparin. As FH is responsible for regulating the complement system, decreased affinity for heparin results in decreased regulation on surfaces of self. To understand the involvement of the Y402H polymorphism in AMD, we leverage methods from bioinformatics and computational biophysics to quantify structural and dynamical differences between SCR7 isoforms that contribute to decreased pattern recognition in $SCR7^{H402}$. Our data from molecular and Brownian dynamics simulations suggest a revised mechanism for decreased heparin binding. In this model, transient contacts not observed in structures for SCR7 are predicted to occur in molecular dynamics simulations between coevolved residues Y402 and I412, stabilizing $SCR7^{Y402}$ in a conformation that promotes association with heparin. H402 in the risk isoform is less likely to form a contact with I412 and samples a larger conformational space than Y402. We observe energy minima for sidechains of Y402 and R404 from $SCR7^{Y402}$ that are predicted to associate with heparin at a rate constant faster than energy minima for sidechains of H402 and R404 from $SCR7^{H402}$. As both carbohydrate density and degree of sulfation decrease with age in Bruch's membrane of the macula, the decreased heparin recognition of $SCR7^{H402}$ may contribute to the pathogenesis of AMD.

INTRODUCTION

High-throughput sequencing and genome-wide association studies have identified both genetic hallmarks and risk factors for developing age-related diseases. How exactly genetic variation results in disease is often less well characterized. Such explanations are complex in nature and may involve perturbations in protein structure, disruption of interactions between biological molecules, or other environmental factors that affect expression levels and spatial concentration profiles. Other sequence-based methods can be used to provide further insight concerning how polymorphisms diverge from similar sequences. For example, genetic variation across orthologous protein sequences can be harnessed to identify coevolved pairs that are predictive of intramolecular contacts (1,2).

Within the complement system, a part of innate immunity, such genome-wide association studies have identified a missense mutation in the gene-encoding factor H (FH) and the splice-variant factor H–like protein 1 (FHL-1) that conveys risk for developing age-related macular degeneration (AMD) (3). Specifically, this single nucleotide polymorphism (SNP) mutates tyrosine at position 402 to histidine (Y402H) and is associated with decreased binding affinity of particular FH and FHL-1 domains for heparin, C-reactive protein, and products of oxidative stress (4–7). Although the effect of age on the etiology of macular degeneration is still being studied, preliminary results have suggested that the extracellular matrix changes in terms of heparin density and composition with age (8). Thus, the reduced affinity for heparin that is exhibited by the risk isoform, Y402H, may not result in disease until heparin density in the macula drops below a particular threshold or heparin composition changes significantly.

Structurally, 20 short consensus repeats (SCRs) joined by 19 short linker sequences comprise FH. SCR1-4 interact with complement components to inhibit an immune response, SCR7 recognizes patterns in carbohydrates that are associated with surfaces of self, and SCR19-20 recognize both carbohydrate patterns of self and a fragment of complement component 3 that is a hallmark of complement

Submitted May 4, 2018, and accepted for publication December 7, 2018.
*Correspondence: dmorikis@ucr.edu
Editor: Madan Babu Mohan.
https://doi.org/10.1016/j.bpj.2018.12.007

inactivation (9–11). As a result, the pattern recognition properties of SCR7 and SCR19-20 direct the regulation of SCR1-4 toward surfaces of self preventing an autoimmune response. The Y402H polymorphism occurs within SCR7 and has been shown to decrease the binding affinity of SCR6-8 for heparin, although the SNP in full-length FH does not significantly affect binding affinity (4,5). As FHL-1 only contains the first seven SCR of FH, the regulator lacks the second recognition domain of SCR19-20. Thus, regulation by FHL-1 is more likely to be affected by the SNP. Within the field, it has been hypothesized that FHL-1 may impart more regulation in the macula than FH because the seven SCR of FHL-1 may diffuse more easily through Bruch's membrane than the 20 SCR of FH (12). This effect may be more pronounced as Bruch's membrane thickens during early stages of AMD (13).

Despite the known effects of the Y402H polymorphism on the association of SCR7 from FH and FHL-1 with heparin, no clear mechanism for how the SNP results in reduced affinity has been identified. Experimental studies have characterized the static structures of both SCR7 isoforms (4,14) and investigated interactions between $SCR7^{H402}$ and sucrose octasulfate (14). However, all crystal structures from the Protein Data Bank (PDB) are of $SCR7^{H402}$, and the NMR structure of $SCR7^{Y402}$ does not contain any binding partners such as heparin. Indeed, no structural study of interactions between the nonrisk isoform and heparin currently exists. Some speculation based on static structures hypothesizes that Y402 sterically prohibits interactions with particular sulfation patterns in carbohydrates (14), but such a hypothesis assumes that the conformation of $SCR7^{Y402}$ in complex with heparin will adopt the solution structure that is mostly conserved for $SCR7^{Y402}$ and $SCR7^{H402}$. As protein sidechains are flexible and ligand binding often involves structural changes, such an assumption may not be warranted despite the solution structures of $SCR7^{Y402}$ and $SCR7^{H402}$ being similar (Fig. 1) (4,5). Herein, we hypothesize that the Y402H mutation alters the preferred arrangement of key residues known to be involved in heparin binding from heteronuclear single quantum coherence (HSQC) and mutagenesis experiments (4,5,15). To assess the validity of this hypothesis, we leverage methods from bioinformatics and computational biophysics to assess the relevance and effects of the Y402H polymorphism. Our results show increased backbone fluctuations at position 402 and proximal residues as a result of disrupting a coevolved contact between Y402 and I412 that is observed in molecular dynamics (MD) simulations but not structures of FH SCR7 from the PDB. Moreover, we describe differences in conformational sampling between SCR7 isoforms involving energy minima for sidechain orientations of positions 402 and 404 only exhibited by $SCR7^{Y402}$ where the molecule is able to associate with heparin significantly faster than $SCR7^{H402}$.

METHODS

Direct coupling analysis

The sequence of FH SCR7 was submitted to the jackhmmer (16) web server with default parameters and subject to three iterations. The resulting 44,511 sequences from the uniprotrefprot database were passed through CD-HIT (17,18) with a clustering threshold of 95%, a word size of 5, a 90% alignment coverage relative to the longer sequence, and a length difference cutoff of 90%. MAFFT (19) was used to align the 22,825 sequences with the dpparttree method, 1000 parts, the BLOSUM62 matrix (blocks substitution matrix 62) (20), a gap opening penalty of 1.53, and a gap extension penalty of 0.123. The multiple sequence alignment (MSA) was filtered such that only records with nongap characters at positions corresponding to the four conserved cysteines forming two disulfide bridges, the conserved tryptophan, and Y402 were retained. Residues before the first cysteine of the first disulfide bridge and after the last cysteine of the last disulfide bridge were trimmed from the MSA, and columns with greater than 5% gap characters were removed. The final MSA of 4940 sequences (including the record for FH SCR7) was passed through the direct coupling analysis (DCA) algorithm (2) with a reweighting factor of 90 and regularization factor of 0.01.

Molecular dynamics

To construct simulation systems, structures for the FH $SCR7^{Y402}$ and $SCR7^{H402}$ isoforms were retrieved from the PDB: 2jgx and 2uwn, respectively. PDBFixer (21) was used to add missing heavy atoms to 2uwn and add hydrogens to 2jgx and 2uwn for the most common protonation state at pH 7.4. Structures were trimmed to contain only residues 388 through 444. PROPKA (22,23) was used to verify that protonation states are appropriate for pH 7.4, and we observed no differences in predicted protonation states. For FH $SCR7^{H402}$ in particular, all simulations used a neutrally charged H402 with hydrogen placed on atom $N_{\varepsilon 2}$. Each isoform was then solvated in a water box with a 10 Å buffer region around protein in the positive and negative x, y, and z directions. Water molecules were simulated with the TIP3P (transferable intermolecular potential with three points) model, and sodium and chloride ions were added to each system such that the final ionic strength was 150 mM and the net charge of the system

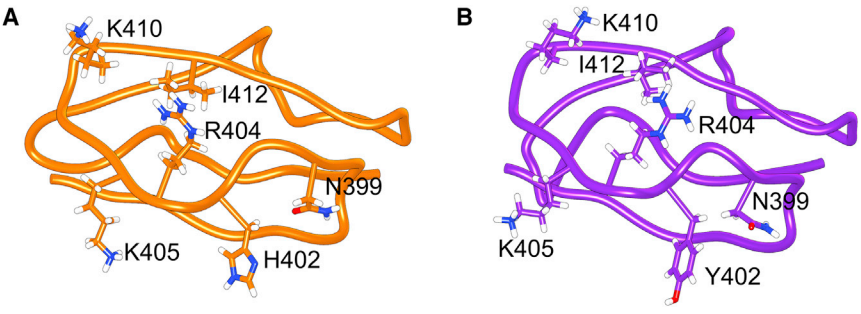

FIGURE 1 Structural representations of SCR7 are shown with key residues labeled. HSQC experiments suggest chemical shift perturbations at labeled residues due to heparin binding. Aside from the Y402H polymorphism, structure is largely conserved between (*A*) $SCR7^{H402}$ from x-ray crystallography and (*B*) $SCR7^{Y402}$ from NMR. For all subsequent molecular graphics, we will display the same amino acid side chains and maintain a color scheme where $SCR7^{H402}$ is orange and $SCR7^{Y402}$ is purple. To see this figure in color, go online.

was neutral. For all MD simulations, the CHARMM36 force field (24) was used.

Each simulation system was subject to steepest-descent minimization in Gromacs (25) with CHARMM36-compatible parameters. Specifically, these parameters require the Verlet cutoff scheme with cutoff distances of 12 Å for van der Waals and Coulombic interactions and a switching distance of 10 Å for van der Waals interactions. Particle mesh Ewald electrostatics (interpolation order of 4, Fourier spacing of 1.6 Å) was enabled while dispersion correction was disabled. These parameters are shared with all subsequent MD simulations. For minimization in particular, an energy minimization step size of 0.1 Å was used, and minimization was terminated when the maximal force was less than 100 kJ/mol per Å.

After minimization, each system was equilibrated with short, 100 ps NVT (constant number of molecules, volume, and temperature) and NPT (constant number of molecules, pressure, and temperature) simulations. NVT equilibration was performed first with a leap-frog integrator and 2 fs time step where initial velocities were randomly assigned from a Maxwell distribution. In this simulation, heavy atoms were restrained and all bonds constrained with the LINCS algorithm (order of four, one iteration). A Berendsen thermostat was used to maintain a system temperature of 300 K with a time constant of 0.1 ps. Next, NPT equilibration was performed with the same integrator and time step, although velocities were preserved from the previous simulation. Similarly, this simulation restrained heavy atoms, constrained bonds with LINCS, and controlled temperature with the Berendsen thermostat. In addition, pressure was controlled with the Parrinello-Rahman barostat using a time constant of 2 ps, a reference pressure of 1 bar, and water compressibility of 4.5×10^{-5} bar^{-1}, and isotropic coupling.

Following equilibration, a primary simulation of 10 μs and 5 secondary simulations of 5 μs were performed for each isoform. Thus, each isoform was simulated for a total of 35 μs. Primary simulations were continued from equilibration runs, whereas secondary simulations branched out from the five most infrequent state transitions observed in the primary simulations. Secondary simulations were initialized with random velocities drawn from a Maxwell distribution. For all primary and secondary MD simulations, a leap-frog integrator with a 2-fs time step was used. Temperature and pressure were controlled with the Berendsen thermostat and Parrinello-Rahman barostat as in NPT equilibration, respectively. No restraints were imposed on any atoms, and only bonds between hydrogens and heavy atoms were constrained by the LINCS algorithm.

Markov chain models for conformational dynamics

Markov chain models that describe conformational sampling of FH $SCR7^{Y402}$ and $SCR7^{H402}$ were based on all data from primary and secondary MD trajectories. Models for each isoform were built separately such that the state space is not shared between SCR7 isoforms; however, we aimed to keep parameters as similar as possible for each isoform to facilitate comparisons. Specifically, Cartesian coordinates for all atoms after every 100 ps were converted to a features set comprising sine and cosine components for the φ, ψ, χ_1, and χ_2 torsion angles. These features were decomposed with time-lagged independent component analysis using a lag time of 100 ps. This lag time was selected because it was observed to maximize the generalized matrix Rayleigh quotient implemented in MSMBuilder (26). Subsequently, the first five independent components were clustered into 100 discrete conformational states using a mini-batch K-means algorithm implemented in MSMBuilder (26). A Bayesian Markov chain, implemented in PyEMMA (27), was built from these discretized trajectories using statistically uncorrelated state transitions and a lag time of 20 ns. This lag time was selected because it was the smallest value where relaxation timescales for the Bayesian Markov chain were not dependent on the selected lag time (Fig. S1). Markov chains were validated with Chapman-Kolmogorov tests (Fig. S2). Networks were constructed from the transition probability matrices of each Markov chain and visualized in Gephi (28) with a force-directed layout (29). Network modularity was calculated in Gephi with a resolution of 1.0 and the randomize and edge weight parameters enabled (30). Representative structures for each of the 100 states were identified by finding the conformation from the MD trajectories for the particular SCR7 isoform with the minimal distance to the cluster center from mini-batch K-means algorithm.

Brownian dynamics

Association of heparin to FH $SCR7^{Y402}$ and $SCR7^{H402}$ was investigated with Brownian dynamics (BD) simulations. For this method, we retrieved the structure of heparin (dp12) from PDB: 1hpn (model 1) (31). In total, 100 representative structures for each FH SCR7 isoform were extracted from the corresponding Markov chain as described previously. In total, 10,000 BD trajectories were acquired for heparin associating with each representative conformation of each FH SCR7 isoform using BrownDye (32). For every trajectory, binding was considered to occur if at least three separate hydrogen bond donor-acceptor pairs between heparin and SCR7 residues 390, 402, 404, 405, 406, 410, 411, 413, or 414 were separated by 4 Å or less during the simulation. No binding was considered to occur if a molecule did not bind within 1×10^6 steps. PDB2PQR (33,34) was used to assign partial charges and van der Waals radii from the CHARMM force field. Glycan reader (35) was used to convert heparin residues and atoms to CHARMM-compatible nomenclature, then partial charges and van der Waals radii were assigned manually from the CHARMM forcefield. The adaptive Poisson-Boltzmann solver (36) was used to estimate a grid of electrostatic potentials for heparin and each SCR7 conformer using a solvent dielectric coefficient of 78.54, a protein dielectric coefficient of 20.0 (37), a solvent with sodium and chloride ions for 150 mM ionic strength, and a Debye length of 8.08 Å in solvent. The average minimal distance moved during a time step was scaled by 0.2 using the minimal-dx parameter of BrownDye (32). Additionally, each simulation included desolvation forces and hydrodynamic interactions.

Electrostatic similarity

Electrostatic similarity for the 100 conformations of each FH SCR7 isoform used in BD simulations were compared using established methods in the Analysis of Electrostatic Structures of Proteins (AESOP) python library (38). All SCR7 conformers were superposed on backbone C_α, and PDB2PQR (33,34) was used to assign partial charges and van der Waals radii to each atom at a pH of 7.4. The adaptive Poisson-Boltzmann solver (36) was used to generate a grid of electrostatic potentials for each SCR7 conformer using a solvent dielectric coefficient of 78.54, a protein dielectric coefficient of 20.0, a 150 mM NaCl solution, and a grid resolution of 1 Å. Because all grids shared common coordinates for vertices, an electrostatic similarity distance (ESD) was calculated pairwise for every pair of SCR7 conformers according to the standard method in AESOP. An ESD of 0 indicates electrostatic potentials are identical. An ESD value of 2 indicates electrostatic potentials that are equal in magnitude but opposite in polarity. Hierarchical clustering was performed on ESD values to generate a dendrogram comparing the electrostatic similarity of $SCR7^{Y402}$ and $SCR7^{H402}$ conformers. Because most representative structures are within an ESD of 0.75, structures seem to have very similar electrostatic properties.

Validation of representative SCR7 conformations

Chemical shifts for C_α and C_β atoms were predicted for all 100 representative states of each SCR7 isoform using the SPARTA+ (39) method. Expected per residue chemical shifts were calculated according to the following expectation operator:

$$E[X] = \sum_{i=1}^{k} p_i x_i \ , \qquad (1)$$

where p_i is the probability of state i from the stationary distribution of the Markov chain for the particular SCR7 isoform, k is the number of states (each isoform has 100 states), and x_i is the observable for state i. Root mean-square deviation (RMSD) from experimental chemical shifts is calculated according to

$$RMSD = \sqrt{\sum_{j=1}^{n} \frac{E\left[\left(\delta_{j,\,predicted} - \delta_{j,observed}\right)^2\right]}{n}}, \qquad (2)$$

where j denotes the residue index, n is the total number of residues for which there exists predicted and experimental chemical shifts for the atom of interest, and δ_j is a chemical shift (predicted or observed as indicated by subscript) for residue j. Predicted chemical shifts are values returned from SPARTA+ for either conformations from our MD simulations or reference structures from the PDB, and experimental chemical shifts are values deposited on the Biological Magnetic Resonance Data Bank (BMRB) for SCR7^{Y402} or SCR7^{H402}. The expectation for the squared deviations of chemical shift predictions from experimental values in Eq. 2 is calculated according to Eq. 1 for representative conformers of SCR7. For comparisons between MD and reference structures deposited in the PDB, the expectation operator simplifies to the mean shift for residue j. Note that RMSD values are not comparable across different atom types because chemical shifts for different atom types vary in magnitude.

Custom analysis scripts

Most analyses were performed with custom Python scripts leveraging existing libraries such as MDTraj (40), PDBFixer (21), MSMBuilder (26), PyEMMA (27), SciPy (41), NumPy (42), Biopython (43), AESOP (39), pandas (44), NetworkX (45), matplotlib (46), and statsmodels. Some shell scripts were also used to carry out coevolution methods, calling tools such as CD-HIT (17,18), MAFFT (19), and the DCA algorithm (2) from the Valencia group. To request data sets and analysis scripts, please contact the corresponding author or visit https://biomodel.engr.ucr.edu/.

RESULTS

Residue Y402 has coevolved with I412

Leveraging large databases of protein sequences resulting from high-throughput sequencing methods, we constructed a MSA for homologous SCR domains across 366 species ranging from viruses to humans. From these sequences, we computed coevolutionary couplings between aligned columns using DCA (2). From this method, pairs of positions in the MSA were considered strongly coupled if scores fell within the top 1% of all scores according to a Gaussian distribution (Table S1). Using a Fisher's exact test, we found significant enrichment of true contacts (cutoff distance of 6 Å between heavy atoms) for residue pairs that score in the top 1% than for all possible residue pairs (p-value = 2.5×10^{-5}). Thus, our results agree with previous studies that suggest high scoring coevolved couplings are predictive of contacts in protein structure (Fig. S3) (1,2). We found a total of 17 strongly coupled positions (Table S1) where nine such pairs were separated by three or more residue positions in SCR7. Of these couplings between nonproximal residues, the coevolved pair with the highest score of 1.01 (p-value: 4.0×10^{-8}) occurred between residues V429 and P438 in SCR7. This predicted contact is observed to occur between two β-strands from structures of SCR7 deposited in the PDB (Fig. S4) (4,14). The second coevolved pair with a score of 0.91 (p-value: 8.8×10^{-7}) occurs between residues Y402 and I412 in SCR7. However, no contacts are observed between position 402 and I412 given a minimal distance of 9.3 Å between these positions in the NMR structure of SCR7^{Y402} and the crystal structure for SCR7^{H402} (4,14). Given the high coupling of Y402 with I412, we decided to investigate the conformational dynamics of both the SCR7^{Y402} and SCR7^{H402} isoforms to evaluate if dynamical interactions between Y402 and I412 form and to observe the structural effects of the Y402H mutation.

The disruption of the coevolved pair is associated with increased fluctuations near position 402

To simulate atomic motion in SCR7^{Y402} and SCR7^{H402}, we performed multiple MD simulations for a total of 35 μs per isoform. Interestingly, SCR7^{H402} exhibited an increase of ~1.5 Å in root-mean-square fluctuation (RMSF) of C_α carbons at residue number 402 compared to SCR7^{Y402} (Fig. 2 *A*). We also observed increases in RMSF at positions 406–412, although smaller in magnitude with an average of 0.2 Å. SCR7 isoforms exhibited slightly different RMSFs for the centroids of residue side chains (Fig. 2 *B*). RMSFs were ~2 and 1.2 Å higher in SCR7^{H402} than in SCR7^{Y402} for N401 and [Y/H]402, respectively. These observations

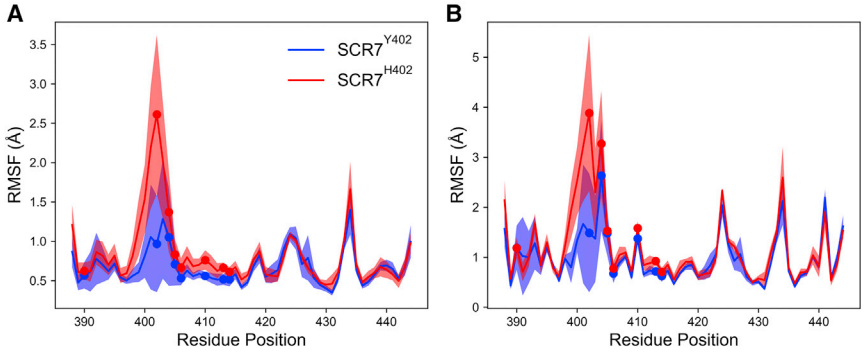

FIGURE 2 Mean RMSF in (*A*) C_α positions and (*B*) side-chain centroids for SCR7^{Y402} (*blue*) and SCR7^{H402} (*red*) are shown for each residue position in SCR7. Shaded regions correspond to one standard deviation above and below the mean RMSF values. Circles are placed in RMSF plots to denote the location of residues Y390, [Y/H]402, R404, K405, F406, K410, D413, V414. Greater differences in RMSF between SCR7 isoforms are observed in C_α positions than in side-chain centroids. Additionally, the Y402H polymorphism is associated with increased RMSF for residues proximal to position 402. To see this figure in color, go online.

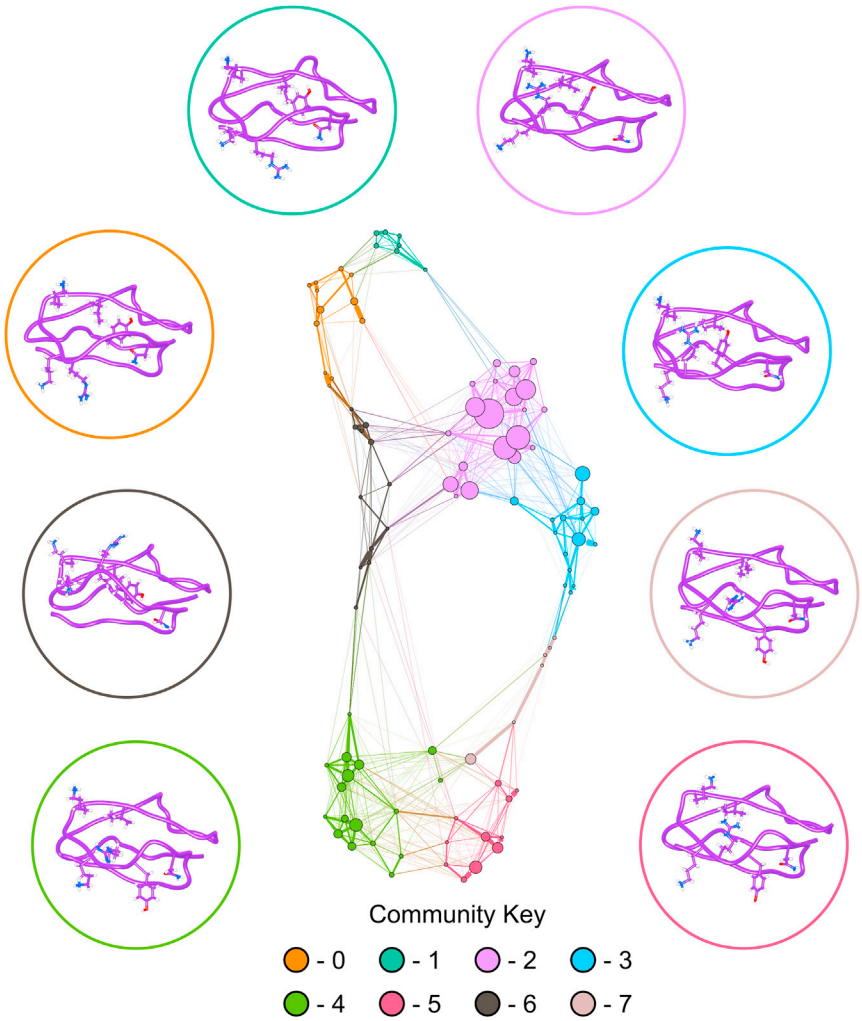

FIGURE 3 Graphical representation of conformational sampling of SCR7^{Y402} is shown with representative structures for each community within the network. Nodes are conformational states from the discretized trajectories that are scaled in size according to the associated probability from the stationary distribution of the Markov chain. Edges represent transitions between conformational states and are scaled according to the associated probability from the Markov chain. Nodes are colored according to community membership, and a representative structure of SCR7^{Y402} is shown for each community. Circles surrounding conformational states denote community membership of each structure, and amino acid side chains for residues N399, Y402, R404, K405, K410, and I412 are displayed. Interestingly, nodes within communities are more connected with other nodes in the same community than with nodes of other communities. This observation suggests a high degree of modularity within the network. To see this figure in color, go online.

suggest that the Y402H polymorphism increases the fluctuations of SCR7^{H402} at residues including position 402 and proximal amino acids. This increased RMSF is also associated with a lower contact probability with I412, as discussed later in the section entitled "Models for conformational dynamics suggest equilibrium differences between SCR7 isoforms in solution." Together, these observations suggest that the Y402H polymorphism destabilizes the structure of SCR7 near position 402 and disrupts a coevolved but transient contact between position 402 and I412.

Conformational dynamics arising from slowest motions in SCR7^{Y402} are more modular than in SCR7^{H402}

To determine how increased fluctuations in SCR7^{H402} result in conformational differences between SCR7^{H402} and SCR7^{Y402}, we constructed Bayesian Markov models to describe conformational dynamics in each isoform (47). These conformations are based on time series of φ, ψ, χ_1, and χ_2 torsion angles that describe protein structure throughout each MD simulation. Time-lagged independent component analysis (48) was used to decompose the SCR7^{Y402} and SCR7^{H402} torsional feature sets into projections along the five slowest eigenvectors for each isoform. As a result, each Markov model describes the five slowest, independent motions for the associated SCR7 isoform. Interestingly, by constructing a network from the probability transition matrix of each Markov model, we noticed a striking difference in modularity (30) between SCR7 isoforms. In SCR7^{Y402}, members of a community are more connected to members of the same community than members of other communities with a modularity of 0.745. SCR7^{H402}, however, has a lower modularity of 0.660. This observation suggests that members of an individual community are more connected to members of another community in SCR7^{H402} than in SCR7^{Y402}. This difference in modularity is clearly visible in Figs. 3 and 4 as a result of the force-directed graph layout (29).

After extracting representative structures from Markov chain states, we compared predicted C_α and C_β chemical shifts to existing experimental data. Each representative

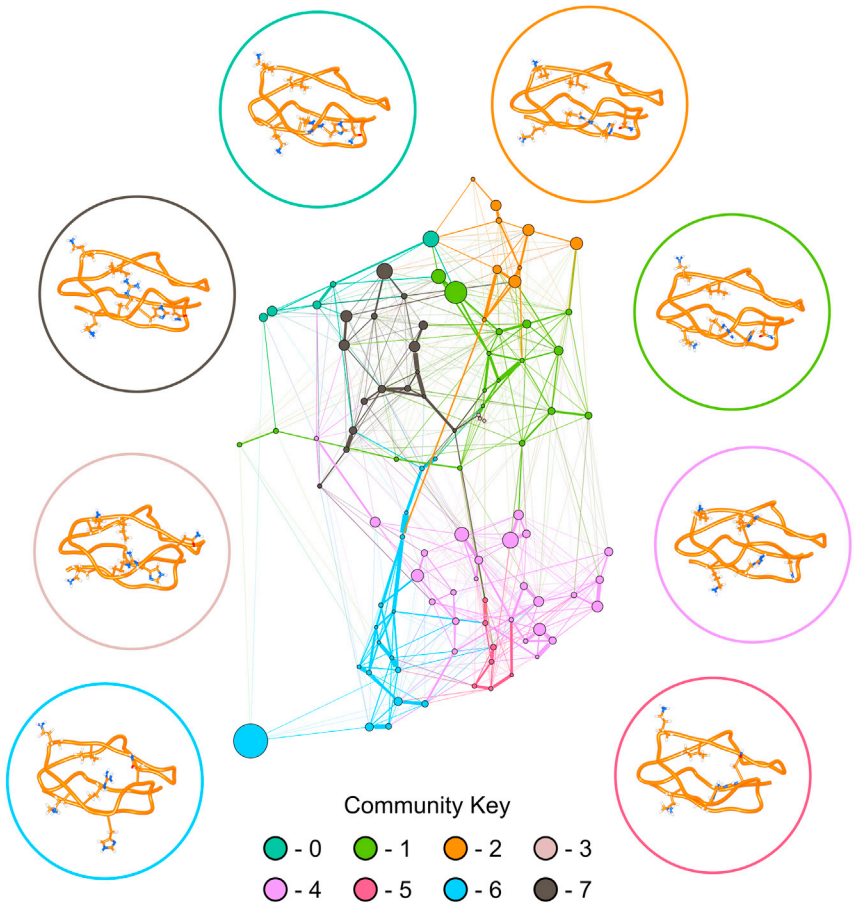

FIGURE 4 Graphical representation of conformational sampling of SCR7^{H402} is shown with representative structures for each community within the network. Nodes are conformational states from the discretized trajectories that are scaled in size according to the associated probability from the stationary distribution of the Markov chain. Edges represent transitions between conformational states and are scaled according to the associated probability from the Markov chain. Nodes are colored according to community membership, and a representative structure of SCR7^{Y402} is shown for each community. Circles surrounding conformational states denote community membership of each structure, and amino acid side chains for residues N399, H402, R404, K405, K410, and I412 are displayed. Compared to the network for SCR7^{Y402} (Fig. 3), the network for SCR7^{H402} appears far less modular because there are more edges between nodes of different communities. To see this figure in color, go online.

structure corresponds to the conformation closest to a cluster center of a single state within the Markov chain as discussed in the Methods. Generally, observed chemical shifts from HSQC experiments deposited on the BMRB fell within the 95% confidence intervals for the mean predicted chemical shifts, although we observed a few discrepancies. When our predictions for representative structures diverged from observed values, we typically observed that chemical shifts predictions on reference structures, the NMR and crystal structures for SCR7 isoforms from the PDB, also diverged. In fact, our MD structures seemed to correct regions with unfavorable energy in the top NMR structure for SCR7^{Y402}, bringing predicted shifts closer to observed chemical shifts. For example, predicted C_α chemical shifts for S437 from the NMR structure of SCR7^{Y402} deviate from experimental values, and the PDB reports one geometric outlier for this residue. Similarly, predicted C_β chemical shifts for residues W436 and S437 in the NMR structure of SCR7^{Y402} deviate from experimental values, and the PDB reports one geometric outlier for each of these residues. Chemical shift predictions for representative structures of SCR7^{Y402} performed better than the NMR reference model for residues W436 and S437. Otherwise, reference predictions typically fell within 95% confidence intervals for mean shifts from representative structures from our Markov chains. To quantify how well representative structures agree with experimental evidence, we calculated RMSD of predicted chemical shifts from observed chemical shifts deposited on the BRMB. This RMSD was calculated for both representative structures from Markov chain models and for reference structures from the PDB (PDB: 2jgx and 2uwn) so that the quality of representative structures could be directly compared to the quality of PDB structures (Figs. S5 and S6). For the C_α chemical shifts from our models, we expect RMSDs of 1.31 and 1.94 parts per million (ppm) for SCR7^{Y402} and SCR7^{H402}, respectively. These values can be compared to reference RMSDs of 1.65 and 1.73 ppm for SCR7^{Y402} and SCR7^{H402}, respectively. Similarly, we expect RMSDs of chemical shifts for C_β of 1.51 and 2.19 ppm for SCR7^{Y402} and SCR7^{H402}, respectively. Reference RMSDs for chemical shifts of C_β are 1.82 and 1.93 for SCR7^{Y402} and SCR7^{H402}, respectively.

Models for conformational dynamics suggest equilibrium differences between SCR7 isoforms in solution

Network representations of conformational sampling in both SCR7 isoforms highlight structural differences. In

$SCR7^{Y402}$ communities 0–3 and 6, Y402 prefers to interact with I412, a residue with a side chain that is more buried in the structure of SCR7 (Fig. 3). Communities 4–5 and 7, however, have Y402 in a conformation that is more solvent exposed. Thus, Y402 switches between a conformation observed in the NMR structure and a contact predicted by a coevolved pair. In $SCR7^{H402}$ H402 appears to interact with I412 less frequently, preferring to interact with N399 or adopt a solvent exposed conformation (Fig. 4 and 5). According to the Markov models for each isoform, the expected distance between position 402 and I412 is 4.0 Å in $SCR7^{Y402}$ and 7.0 Å in $SCR7^{H402}$. With a cutoff distance of 4 Å, the corresponding contact probabilities are 69.7 and 14.0% for $SCR7^{Y402}$ and $SCR7^{H402}$, respectively.

More broadly, the Markov chains for conformational sampling in each SCR7 isoform suggest expected differences between torsion angles. As shown in Fig. 6, the largest differences involve changes in backbone dihedrals around A425, K424, and L422, although it is unclear how relevant these differences are for etiology of AMD. Other primary differences in backbone structure appear to occur at N402, N399, G403, Q400, and R404. Structural changes here could contribute to differences in binding affinity for heparin. Prior HSQC NMR experiments have identified a number of chemical shift perturbations that result when SCR7 binds heparin. These residues include Y390, Y402, R404, K405, F406, S411, I412, D413, and V414 (4). Other mutagenesis studies suggest the importance of R404, K405, and K410 for binding heparin (5,15). Thus, disruptions to structure between residues N399 and V414 could disrupt the topology of interactions with heparin, leading to a lower binding affinity.

SCR7 isoforms associate with heparin at different rates despite similar electrostatic potentials

In order to evaluate how structural changes may affect association of SCR7 with heparin (Fig. 7), we performed BD simulations with representative structures from each state within the Markov models for both $SCR7^{Y402}$ and $SCR7^{H402}$. $SCR7^{Y402}$ was found to associate with an expected rate constant of 6.89 ($\pm$0.06) $\times$ 10^9 $M^{-1}s^{-1}$, whereas $SCR7^{H402}$ was found to associate with an expected rate constant of 6.75 ($\pm$0.06) $\times$ 10^9 $M^{-1}s^{-1}$. Although these rates are of the same order of magnitude, $SCR7^{Y402}$ was found to associate significantly faster than $SCR7^{H402}$ (p-value < 0.001). Given the relatively small ESD between SCR7 isoforms that suggest similar electrostatic grid potentials (Fig. S7), topological differences in hydrogen bond donors and acceptors on the surface of SCR7 are likely the primary cause of differences in association rates from BD simulations.

Energy minima for side-chain orientations of [Y/H]402 and R404 explain differences in association rates

Given the differences in association rate constants between SCR7 isoforms, we investigated how side-chain conformations contribute to these differences. For each key residue identified from prior experimental studies (Y390, Y402, R404, K405, F406, K410, S411, D413, and V414), we calculated the first two principal components for the coordinates of a representative atom after superimposing all conformations on C_α atoms. In this way we constructed a free energy landscape for the position vector of each amino

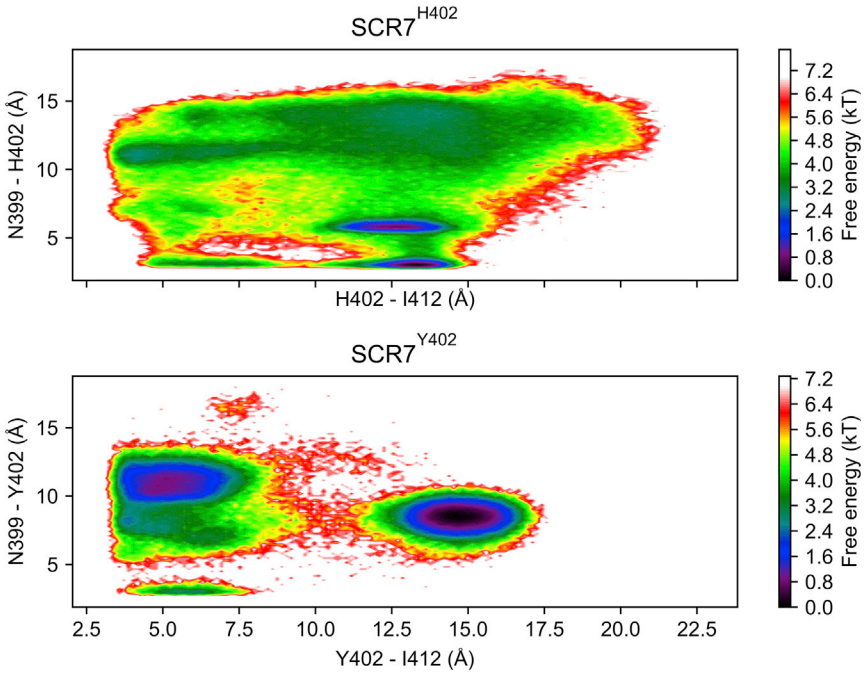

FIGURE 5 Free energy landscapes for distances between residue pairs [Y/H]402 – I412 and N399 – [Y/H]402 are displayed for each SCR7 isoform. The distance between [Y/H]402 and I412 is calculated from the positions of atom CE1 in Y402 or H402 and atom CD1 in I412, whereas the distance between N399 and [Y/H]402 is calculated from atom ND2 in N399 and atom ND1 or OH in H402 or Y402, respectively. Free energies are indicated by inset color bar. States with lower free energy are expected to be observed with higher frequency. Note how the free energy landscape for $SCR7^{Y402}$ is more constrained to two minima than $SCR7^{H402}$ that samples a larger area of the landscape. Also note how $SCR7^{Y402}$ is more likely to form a contact between Y402 and I412 with a cutoff distance of 4 Å than $SCR7^{H402}$, while $SCR7^{H402}$ is more likely to form a contact between N399 and H402 with a cutoff distance of 4 Å than $SCR7^{Y402}$. To see this figure in color, go online.

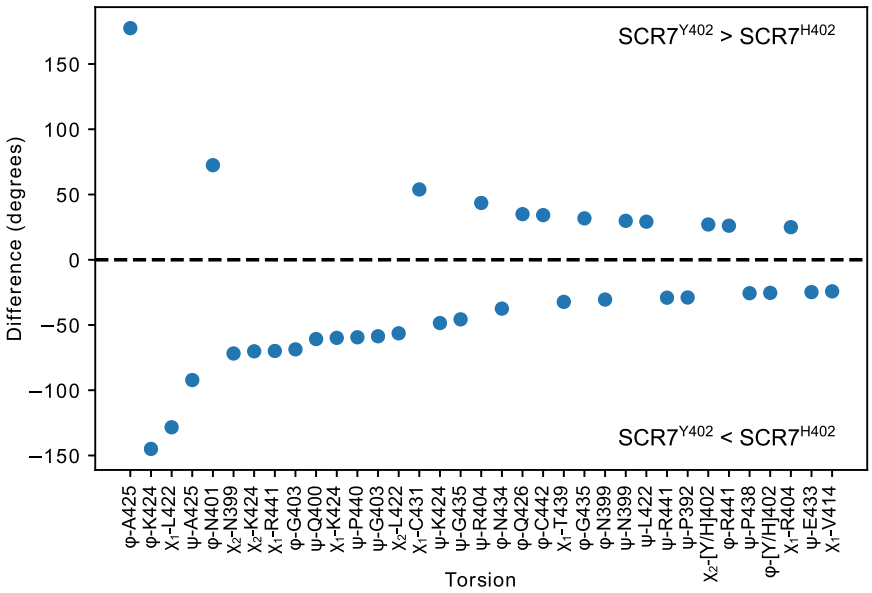

FIGURE 6 Expected differences of torsion angles between Markov chains for SCR7 isoforms are displayed. Torsions larger in SCR7^{Y402} are located above the dashed line, whereas torsions larger in SCR7^{H402} are located below the dashed line. Torsion types and residue membership are labeled along the x axis in order of decreasing magnitude difference between isoforms. Expected torsions appear to be different between SCR7 isoforms at residues L422 and A425, though these differences result in small side-chain rearrangements. Expected torsions also differ at residues N399–R404, resulting in more appreciable differences in sidechain arrangements. To see this figure in color, go online.

acid side chain in both SCR7 isoforms (Fig. 8). Next, we measured mean association rate constants in each energy minimum from the free energy landscape using a threshold energy of 2 kT. This was achieved by identifying energy minimum membership for each representative structure from BD simulations. SCR7^{H402} and SCR7^{Y402} were found to have two distinct energy minima for the sidechain orientation of position 402 (Fig. 8). SCR7^{Y402} minimum 1 corresponds to the NMR orientation, whereas minimum 2 corresponds to an orientation that forms a contact with I412 (Fig. S8). In SCR7^{H402} neither minimum 1 nor minimum 2 corresponds to the orientation in the crystal structure or to the formation of a contact with I412 (Fig. S8). Although SCR7^{Y402} minimum 2 is observed to have the highest mean association rate constant, the predicted heparin association rate constant for this minimum is only significantly higher than SCR7^{H402} minimum 2 according to one-way analysis of variance followed by Tukey's Honest Significant Difference hypothesis test ($\alpha = 0.05$). For the side-chain orientation of R404, SCR7^{H402} has three energy minima, whereas SCR7^{Y402} has four energy minima (Fig. 8). Minima 1–3 of SCR7^{Y402} overlap with minimum 3 of SCR7^{H402}. These overlapping minima correspond to the orientation of R404 in structures of SCR7 from the PDB (Fig. S8). Notably, the unique minimum 4 of SCR7^{Y402} exhibits the highest mean predicted heparin association rate constant that is significantly higher than mean rate constants for all minima of SCR7^{H402} according to one-way analysis of variance followed by Tukey's Honest Significant Difference hypothesis test ($\alpha = 0.05$). The mean rate constant for SCR7^{Y402} minimum 1 is also significantly higher than all SCR7^{H402} minima, whereas SCR7^{Y402} minimum 3 is only significantly higher than SCR7^{H402} minimum 2. No significant differences in rate constants could be discerned between energy minima within a single isoform. Free energy minimum 2 for the side-chain orientation of R404 in SCR7^{Y402} was not included in statistical analyses because only one representative structure belonged to this minimum. Overall, regions of energy landscapes with high association rate constants are more probable for SCR7^{Y402} than SCR7^{H402}. Free energy landscapes for side-chain conformations of other key residues do not vary between SCR7 isoforms (Fig. S9).

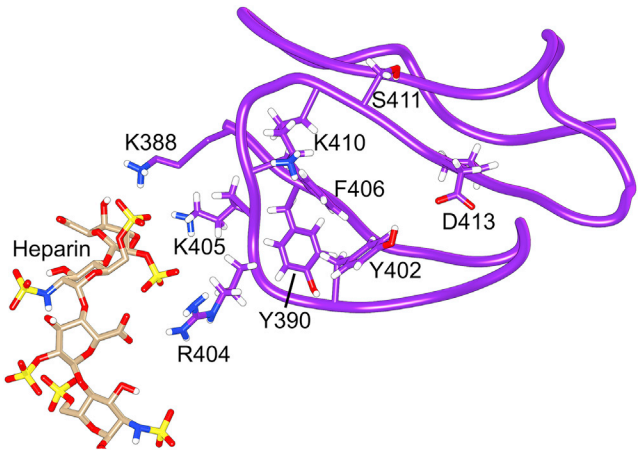

FIGURE 7 Example encounter complex between SCR7^{Y402} and heparin (dp12) from a BD trajectory where binding occurs. All annotated residues on SCR7 except K388 have been implicated in binding heparin by HSQC NMR and mutagenesis experiments. To see this figure in color, go online.

DISCUSSION

Despite structural similarities between SCR7^{Y402} and SCR7^{H402} (Fig. 1), the Y402H mutation in FH SCR7 is known to reduce the binding affinity of FH SCR6-8 for heparin. As Y402 appears to have coevolved with I412, the hydrophobic contribution of Y402 is likely important for promoting a transient, coevolved contact between Y402

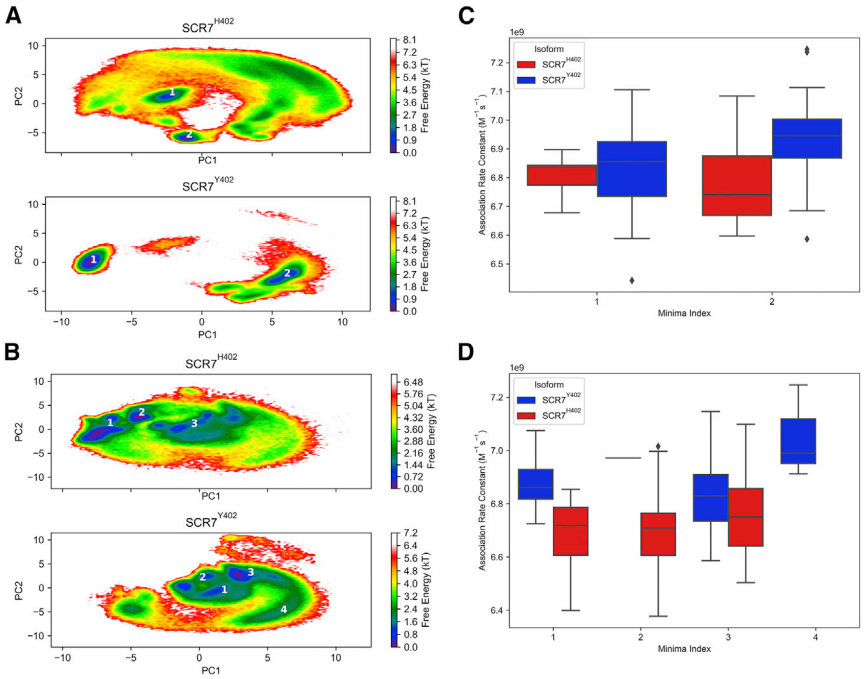

FIGURE 8 Free energy landscapes for a two-dimensional representation of side-chain orientations for residues (*A*) [Y/H]402 and (*B*) R404 are shown for SCR7 isoforms. To describe the orientation of side chains, position vectors are found from coordinates of heavy atoms NE2 in H402, OH in Y402, and CZ in R404 and decomposed into two dimensions with principal component analysis. Regions of the landscape are annotated with numbers that correspond to energy minima, with free energy values of 2 kT or below. Free energies are indicated by the inset color bar. Box plots display predicted heparin association rate constants for each free energy minimum in SCR7 isoforms for (*C*) the side-chain orientation of [Y/H]402 and (*D*) the side-chain orientation of R404. Regions with high average association rate constants are more closely associated with energy minima of $SCR7^{Y402}$. To see this figure in color, go online.

and I412. The Y402H mutation appears to disrupt this contact and increase flexibility at position 402. The existence of a contact between Y402 and I412 in $SCR7^{Y402}$ seems reasonable because within the predicted ensemble of structures determined from NMR experiments for $SCR7^{Y402}$ there exists a structure where the Y402 side chain is oriented toward the core of the protein instead of toward the solvent (although no contact is observed with I412 in particular). In the case of the Y402H polymorphism, H402 rarely interacts with I412. Instead, H402 samples a wide range of conformations. One very probable conformation involves contacts (including hydrogen bonding) between H402 and N399 (Fig. 5). Thus, it is possible that the additional number of hydrogen bond donors and acceptors and the decreased hydrophobicity of H402 destabilize the topology of the heparin-binding site of SCR7.

Although our simulations suggest that residues fluctuate more near position 402 in $SCR7^{H402}$ than in $SCR7^{Y402}$, the conformational space between SCR7 isoforms overlaps. In fact, tertiary structure of SCR7 is largely conserved between isoforms. The two conserved disulfide bridges within SCRs cause any deviations in terms of tertiary structure to be unlikely. Side chain fluctuations and rearrangements in flexible loops, however, are still possible and may contribute to differences in binding affinity. As chemical shifts for $SCR7^{Y402}$ and $SCR7^{H402}$ are deposited on the BMRB, we compared predicted chemical shifts for our representative conformations from MD simulations to predicted chemical shifts for solved structures of PDB: SCR7. Representative structures from the Markov model for $SCR7^{Y402}$ exhibited lower RMSD from experimentally determined chemical shifts than the NMR structure of $SCR7^{Y402}$ (Figs. S5 and S6). $SCR7^{H402}$, however, exhibited slightly higher RMSD from experimentally determined chemical shifts than the crystal structure of $SCR7^{H402}$, although predicted shifts of each residue for the crystal structure generally fell within the 95% confidence interval for predicted chemical shifts. Thus, MD simulations and the associated Markov chains appear to agree with experimental chemical shifts, even correcting regions with unfavorable energies from the NMR structure of $SCR7^{Y402}$ at W436 and S437.

Although the structure containing $SCR7^{H402}$ in complex with sucrose octasulfate suggests favorable contacts between H402 and sulfates (14), no structure for interactions between $SCR7^{Y402}$ and similar carbohydrates exists. Moreover, sucrose octasulfate contains more sulfates than does heparin for molecules with equal degree of polymerization. As a result, many hypotheses that attempt to explain the reduced binding affinity of the Y402H polymorphism are based on the structure of the risk isoform $SCR7^{H402}$. From experiments, however, there is a wealth of information on the key residues that mediate interactions of SCR7 with heparin. HSQC spectra suggest the involvement of Y390, Y402, R404, K405, F406, S411, D413, and V414 in heparin binding (4), whereas mutagenesis suggests the relevance of R404, K405, and K410 (5,15).

By informing BD simulations with potential interacting residues and an ensemble of representative SCR7 conformations from MD, we were able to observe a lower heparin association rate constant for $SCR7^{H402}$ than for $SCR7^{Y402}$. Moreover, we were able to observe associations between free energy minima for side-chain orientations and heparin association rate constants that reveal a mechanism for the decreased pattern recognition resulting from the Y402H

SNP (Fig. 8). Specifically, we observed that differences in association rates between SCR7 isoforms are best explained by the side-chain orientation of R404 where $SCR7^{Y402}$ free energy minimum 4 exhibits the highest mean heparin association rate constant. Distance from this energy minimum (in principal component space) is negatively correlated with the predicted association rate constant (Fig. S10). We also observe that the side-chain orientation of [Y/H]402 explains some differences between heparin association rate constants of SCR7 isoforms because $SCR7^{Y402}$ minimum 2 is observed to have the highest mean association rate constant, although this is only significantly higher from $SCR7^{H402}$ minimum 2. Likely the orientation of R404 is directly involved in heparin binding, whereas the orientation of [Y/H]402 indirectly influences the orientation of R404. This is supported by dynamic cross-correlation analyses of MD trajectories that observe correlated motion between Y402 and R404 but anti-correlated motion between H402 and R404 (Fig. S11). Thus, transient formation of contacts between Y402 and I412 in $SCR7^{Y402}$ may indirectly promote an orientation of the R404 sidechain with a high predicted association rate constant.

Disruption of the Y402–I412 coevolved contact in the risk-isoform SCR^{H402} is observed to change the dynamical structure of SCR7. For example, the network of transitions between conformational states decreases in modularity. Moreover, the decrease in modularity is associated with an increase in square fluctuations proximal to position 402. As a result, side-chain arrangements of H402 and R404 sample a broader conformational space compared to $SCR7^{Y402}$, with H402 interacting frequently with N399 in particular (Fig. 5). As a result, the orientation of R404 is observed to be anticorrelated with the orientation of H402 (Fig. S11), and the side-chain orientation of R404 in $SCR7^{H402}$ poorly samples regions with high predicted heparin association rate constants ($SCR7^{Y402}$ minima 1 and 4). Thus, dynamical behavior of FH SCR7 isoforms provide a mechanism for the decreased binding affinity that results from the Y402H SNP. This mechanism is not apparent from the SCR7 structures deposited on the PDB but agrees with experimental chemical shifts from NMR experiments.

Although FH SCR6-8 are known to be involved in heparin binding, it has been previously shown that SCR7 alone can be studied to understand the role of the Y402H polymorphism (4). Even still, interactions between residues of separate SCRs can in theory change the dynamical behavior of the system. However, we previously observed the formation of a coevolved contact between Y402 and I412 in simulations of SCR6-8 (Fig. S12), 300 ns total per SCR6-8 isoform. By truncating the system to SCR7 only, we were able to achieve significantly longer MD trajectories than if we had to simulate SCR6-8, facilitating more robust dynamical models.

Because protein structures from NMR experiments are models estimated from structural restraints that are heavily dependent on interproton distances, it is possible that structural models from NMR may lose important structural features, including contacts between coevolved pairs, depending on the observation of interproton contacts related to efficiency of magnetization transfer and spectral overlap. MD, however, can be used to identify dynamical changes in protein structure and extract new representative structures from energy minima. A Markovian approach characterizing conformational transitions of $SCR7^{Y402}$ was able to generate representative structures that are closer to expected conformational shifts than the corresponding NMR structure. These changes involve the transient formation of contacts between Y402 and I412. In comparison to the crystal structure of $SCR7^{H402}$, $SCR7^{Y402}$ residues Y402 and R404 are observed to sample different energy minima that correspond to faster association rates with heparin. $SCR7^{H402}$, on the other hand, samples a wide variety of conformations, rarely sampling conformations in regions of the free energy landscape that correspond to fast association with heparin. Guided by existing experimental data and principles from bioinformatics, MD and BD can describe behavior in SCR7 that is difficult to achieve with experimental methods, resulting in a molecular mechanism for decreased heparin-binding affinity associated with the Y402H polymorphism.

CONCLUSIONS

Side chains of [Y/H]402 and R404 prefer different orientations in $SCR7^{Y402}$ and $SCR7^{H402}$, leading to topological differences that contribute to a decreased rate constant for association with heparin in $SCR7^{H402}$. We also observe that $SCR7^{Y402}$ is more likely to preserve the coevolved contact between [Y/H]402 and I412 than $SCR7^{H402}$. Thus, the Y402H polymorphism appears to destabilize a portion of the SCR7 surface responsible for recognition of heparin. Likely, the increased number of possible dipole-dipole interactions and the decreased hydrophobicity of H402 compared to Y402 result in increased fluctuations at H402 and proximal residues in $SCR7^{H402}$. These fluctuations manifest in conformational sampling that is less modular and an energy landscape for the side chain orientation of H402 that is less constrained than Y402 from $SCR7^{Y402}$. Additionally, the Y402H polymorphism appears to perturb the preferred side-chain orientation of R404. Because R404 is involved in heparin binding, this perturbation is associated with lower heparin association rates in $SCR7^{H402}$ than in $SCR7^{Y402}$.

Although conventional hypotheses based on interactions between H402 and sucrose octasulfate from $SCR7^{H402}$ state that Y402 may sterically prevent association with some carbohydrates that are not sterically hindered by H402, our data suggest a revised mechanism for decreased heparin binding. Specifically, the transient contacts between Y402 and I412 in $SCR7^{Y402}$ stabilize the molecule in a conformation that promotes association with heparin by

indirectly affecting the side-chain orientation of R404. This observation is in agreement with experiments that show SCR6-8^{H402} eluting from a heparin column faster than SCR6-8^{Y402} (4).

With increasing age, both carbohydrate density and degree of sulfation have been shown to decrease in Bruch's membrane of the macula (8). Additionally, the FH^{Y402} has been shown to bind a wide range of sulfated carbohydrates, whereas FH^{H402} has been shown to have more specificity and to require a high degree of sulfation in carbohydrates (8). The two energy minima for the orientation of Y402 in SCR7^{Y402} could impart robust recognition of a variety of sulfated carbohydrates by indirectly affecting the orientation of the R404 sidechain.

To investigate structural effects of missense SNPs, it is not always possible to acquire crystal structures for pertinent isoforms. In the absence of complete structural information, our approach suggests that methods from bioinformatics and computational biophysics can be leveraged to explore these structural effects and provide structural insights from timescales that are not accessible from static NMR or crystallographic structures. Notably, discretization of conformations from MD using independent component analysis appears to extract representative structures that are predicted to agree with experimental chemical shifts.

SUPPORTING MATERIAL

Supporting Materials and Methods, twelve figures, and one table are available at http://www.biophysj.org/biophysj/supplemental/S0006-3495(18)34502-8.

AUTHOR CONTRIBUTIONS

R.E.S.H. and D.M. wrote the manuscript. R.E.S.H. performed all methods and analyzed the results. D.M. supervised the study.

ACKNOWLEDGMENTS

We thank Juan Rodriguez, David de Juan, Simone Marsili, and Alfonso Valencia for helpful discussions concerning coevolution methods and for sharing their implementation of the DCA algorithm. We also thank Gary Huber for helpful discussions concerning the usage of BrownDye.

R.E.S.H. has received funding from the University of California President's Dissertation-Year Fellowship, the National Science Foundation Integrative Graduate Education and Research Traineeship, and a Whitaker Foundation Summer grant in support of this study. This work was partially supported by NIH grant R01EY027440.

REFERENCES

1. de Juan, D., F. Pazos, and A. Valencia. 2013. Emerging methods in protein co-evolution. *Nat. Rev. Genet.* 14:249–261.
2. Sutto, L., S. Marsili, …, F. L. Gervasio. 2015. From residue coevolution to protein conformational ensembles and functional dynamics. *Proc. Natl. Acad. Sci. USA.* 112:13567–13572.
3. Liszewski, M. K., and J. P. Atkinson. 2015. Complement regulators in human disease: lessons from modern genetics. *J. Intern. Med.* 277:294–305.
4. Herbert, A. P., J. A. Deakin, …, P. N. Barlow. 2007. Structure shows that a glycosaminoglycan and protein recognition site in factor H is perturbed by age-related macular degeneration-linked single nucleotide polymorphism. *J. Biol. Chem.* 282:18960–18968.
5. Clark, S. J., V. A. Higman, …, A. J. Day. 2006. His-384 allotypic variant of factor H associated with age-related macular degeneration has different heparin binding properties from the non-disease-associated form. *J. Biol. Chem.* 281:24713–24720.
6. Weismann, D., K. Hartvigsen, …, C. J. Binder. 2011. Complement factor H binds malondialdehyde epitopes and protects from oxidative stress. *Nature.* 478:76–81.
7. Sjöberg, A. P., L. A. Trouw, …, A. M. Blom. 2007. The factor H variant associated with age-related macular degeneration (His-384) and the non-disease-associated form bind differentially to C-reactive protein, fibromodulin, DNA, and necrotic cells. *J. Biol. Chem.* 282:10894–10900.
8. Keenan, T. D., C. E. Pickford, …, P. N. Bishop. 2014. Age-dependent changes in heparan sulfate in human Bruch's membrane: implications for age-related macular degeneration. *Invest. Ophthalmol. Vis. Sci.* 55:5370–5379.
9. Kieslich, C. A., H. Vazquez, …, D. Morikis. 2011. The effect of electrostatics on factor H function and related pathologies. *J. Mol. Graph. Model.* 29:1047–1055.
10. Makou, E., A. P. Herbert, and P. N. Barlow. 2013. Functional anatomy of complement factor H. *Biochemistry.* 52:3949–3962.
11. E S Harrison, R., R. D. Gorham, Jr., and D. Morikis. 2015. Energetic evaluation of binding modes in the C3d and Factor H (CCP 19-20) complex. *Protein Sci.* 24:789–802.
12. Schmidt, C. Q., J. D. Lambris, and D. Ricklin. 2016. Protection of host cells by complement regulators. *Immunol. Rev.* 274:152–171.
13. Ambati, J., J. P. Atkinson, and B. D. Gelfand. 2013. Immunology of age-related macular degeneration. *Nat. Rev. Immunol.* 13:438–451.
14. Prosser, B. E., S. Johnson, …, S. M. Lea. 2007. Structural basis for complement factor H linked age-related macular degeneration. *J. Exp. Med.* 204:2277–2283.
15. Giannakis, E., T. S. Jokiranta, …, D. L. Gordon. 2003. A common site within factor H SCR 7 responsible for binding heparin, C-reactive protein and streptococcal M protein. *Eur. J. Immunol.* 33:962–969.
16. Johnson, L. S., S. R. Eddy, and E. Portugaly. 2010. Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics.* 11:431.
17. Fu, L., B. Niu, …, W. Li. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics.* 28:3150–3152.
18. Li, W., and A. Godzik. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics.* 22:1658–1659.
19. Yamada, K. D., K. Tomii, and K. Katoh. 2016. Application of the MAFFT sequence alignment program to large data-reexamination of the usefulness of chained guide trees. *Bioinformatics.* 32:3246–3251.
20. Henikoff, S., and J. G. Henikoff. 1992. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA.* 89:10915–10919.
21. Eastman, P. 2016. PDBFixer. Stanford University. https://github.com/pandegroup/pdbfixer.git.
22. Søndergaard, C. R., M. H. Olsson, …, J. H. Jensen. 2011. Improved treatment of ligands and coupling effects in empirical calculation and rationalization of pKa values. *J. Chem. Theory Comput.* 7:2284–2295.
23. Olsson, M. H. M., C. R. Søndergaard, …, J. H. Jensen. 2011. PROPKA3: consistent treatment of internal and surface residues in empirical pKa predictions. *J. Chem. Theory Comput.* 7:525–537.
24. Best, R. B., X. Zhu, …, A. D. Mackerell, Jr. 2012. Optimization of the additive CHARMM all-atom protein force field targeting improved

sampling of the backbone φ, ψ and side-chain χ(1) and χ(2) dihedral angles. *J. Chem. Theory Comput.* 8:3257–3273.

25. Abraham, M. J., T. Murtola, …, E. Lindah. 2015. Gromacs: high performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX.* 1–2:19–25.
26. Harrigan, M. P., M. M. Sultan, …, V. S. Pande. 2017. MSMBuilder: statistical models for biomolecular dynamics. *Biophys. J.* 112:10–15.
27. Scherer, M. K., B. Trendelkamp-Schroer, …, F. Noé. 2015. PyEMMA 2: a software package for estimation, validation, and analysis of Markov models. *J. Chem. Theory Comput.* 11:5525–5542.
28. Bastian, M., S. Heymann, and M. Jacomy. 2009. Gephi: an open source software for exploring and manipulating networks. *In* Third International AAAI Conference Weblogs Social Media, pp. 361–362.
29. Jacomy, M., T. Venturini, …, M. Bastian. 2014. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLoS One.* 9:e98679.
30. Blondel, V. D., J. L. Guillaume, …, E. Lefebvre. 2008. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* 2008:P10008.
31. Khan, S., J. Gor, …, S. J. Perkins. 2010. Semi-rigid solution structures of heparin by constrained X-ray scattering modelling: new insight into heparin-protein complexes. *J. Mol. Biol.* 395:504–521.
32. Huber, G. A., and J. A. McCammon. 2010. Browndye: a software package for Brownian dynamics. *Comput. Phys. Commun.* 181:1896–1905.
33. Dolinsky, T. J., P. Czodrowski, …, N. A. Baker. 2007. PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic Acids Res.* 35:W522–W525.
34. Dolinsky, T. J., J. E. Nielsen, …, N. A. Baker. 2004. PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucleic Acids Res.* 32:W665–W667.
35. Jo, S., K. C. Song, …, W. Im. 2011. Glycan Reader: automated sugar identification and simulation preparation for carbohydrates and glycoproteins. *J. Comput. Chem.* 32:3135–3141.
36. Baker, N. A., D. Sept, …, J. A. McCammon. 2001. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. USA.* 98:10037–10041.
37. Gorham, R. D., Jr., C. A. Kieslich, …, D. Morikis. 2011. An evaluation of Poisson-Boltzmann electrostatic free energy calculations through comparison with experimental mutagenesis data. *Biopolymers.* 95:746–754.
38. Harrison, R. E. S., R. R. Mohan, …, D. Morikis. 2017. AESOP: a Python library for investigating electrostatics in protein interactions. *Biophys. J.* 112:1761–1766.
39. Shen, Y., and A. Bax. 2010. SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network. *J. Biomol. NMR.* 48:13–22.
40. McGibbon, R. T., K. A. Beauchamp, …, V. S. Pande. 2015. MDTraj: a modern open library for the analysis of molecular dynamics trajectories. *Biophys. J.* 109:1528–1532.
41. Eric, J. O., and P. Travis. 2001. SciPy : open source scientific tools for Python. http://www.scipy.org/.
42. van der Walt, S., S. C. Colbert, and G. Varoquaux. 2011. The NumPy array: a struture for efficient numerical computation. *Comput. Sci. Eng.* 13:22–30.
43. Cock, P. J., T. Antao, …, M. J. de Hoon. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics.* 25:1422–1423.
44. McKinney, W. 2010. Data structures for statistical computing in Python. Proc. 9th Python Sci. Conf. 1697900. pp. 51–56.
45. Hagberg, A. A., D. A. Schult, and P. J. Swart. 2008. Exploring network structure, dynamics, and function using NetworkX. *7th Python Sci. Conf.* 836:11–15.
46. Hunter, J. D. 2007. Matplotlib: a 2D graphics environment. *Comput. Sci. Eng.* 9:99–104.
47. Trendelkamp-Schroer, B., H. Wu, …, F. Noé. 2015. Estimation and uncertainty of reversible Markov models. *J. Chem. Phys.* 143:174101.
48. Pérez-Hernández, G., F. Paul, …, F. Noé. 2013. Identification of slow molecular order parameters for Markov model construction. *J. Chem. Phys.* 139:015102.