Published in final edited form as:
Nat Methods. 2018 February ; 15(2): 107–114. doi:10.1038/nmeth.4540.

Interactome INSIDER: a structural interactome browser for genomic studies

Michael J. Meyer[1,2,3,†], **Juan Felipe Beltrán**[1,2,†], **Siqi Liang**[1,2,†], **Robert Fragoza**[2,4], **Aaron Rumack**[1,2], **Jin Liang**[2], **Xiaomu Wei**[1,5], and **Haiyuan Yu**[1,2,*]

[1]Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, New York, 14853, USA

[2]Weill Institute for Cell and Molecular Biology, Cornell University, Ithaca, New York, 14853, USA

[3]Tri-Institutional Training Program in Computational Biology and Medicine, New York, New York, 10065, USA

[4]Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY 14853, USA

[5]Department of Medicine, Weill Cornell College of Medicine, New York, New York, 10065, USA

Abstract

We present Interactome INSIDER, a tool to link genomic variant information with structural protein-protein interactomes. Underlying this tool is the application of machine learning to predict protein interaction interfaces for 185,957 protein interactions with previously unresolved interfaces, in human and 7 model organisms, including the entire experimentally determined human binary interactome. Predicted interfaces exhibit similar functional properties as known interfaces, including enrichment for disease mutations and recurrent cancer mutations. Through 2,164 *de novo* mutagenesis experiments, we show that mutations of predicted and known interface residues disrupt interactions at a similar rate, and much more frequently than mutations outside of predicted interfaces. To spur functional genomic studies, Interactome INSIDER (http://interactomeinsider.yulab.org) enables users to identify whether variants or disease mutations are enriched in known and predicted interaction interfaces at various resolutions. Users may explore known population variants, disease mutations, and somatic cancer mutations, or upload their own set of mutations for this purpose.

[*]To whom correspondence should be addressed. Tel: 607-255-0259; Fax: 607-255-5961; haiyuan.yu@cornell.edu.
[†]The authors wish it to be known that, in their opinion, the first 3 authors should be regarded as joint First Authors.

Author Contributions
M.J.M, J.F.B, S.L., and H.Y. conceived the study. H.Y. oversaw all aspects of the study. M.J.M., J.F.B., S.L., and A.R performed computational analyses. M.J.M., and J.F.B. designed ECLAIR. J.F.B. designed the web interface. R.F., J.L., and X.W. performed laboratory experiments. M.J.M wrote the manuscript with input from J.F.B., S.L., and H.Y. All authors edited and approved of the final manuscript.

Competing Financial Interests
The authors declare no competing financial interests.

INTRODUCTION

Protein-protein interactions facilitate much of known cellular function. Recent efforts to experimentally determine protein interactomes in human[1] and model organisms[2–4], in addition to literature curation of small-scale interaction assays[5], have dramatically increased the scale of known interactome networks. Studies of these interactomes have allowed researchers to elucidate how modes of evolution affect the functional fates of paralogs[4] and to examine, on a genomic scale, network interconnectivities that determine cellular functions and disease states[6].

While simply knowing which proteins interact with each other provides valuable information to spur functional studies, far more specific hypotheses can be tested if the spatial contacts of interacting proteins are known[7]. In the study of human disease, it has been demonstrated that mutations tend to localize to interaction interfaces and mutations on the same protein may cause clinically distinct diseases by disrupting interactions with different partners[6,8]. However, the binding topologies of interacting proteins can only be determined at atomic resolution through X-ray crystallography, NMR, and more recently cryo-EM[9] experiments, limiting the number of interactions with resolved interaction interfaces.

To study protein function on a genomic scale, especially as it relates to human disease, a large-scale set of protein interaction interfaces is needed. Thus far, computational methods, such as docking[10] and homology modeling[11], have been employed to predict the atomic-level bound conformations of interactions whose experimental structures have not yet been determined. However, But docked models are not yet available on a large scale, and while homology modeling has been used to produce models at scale[12], it is only amenable to interactions with structural templates (<5% of known interactions). Together, co-crystal structures and homology models comprise the currently available pre-calculated sources of structural interactomes, covering only ~6% of all known interactions (Fig. 1a–b).

Here, we present Interactome INSIDER (**<u>IN</u>**tegrated **<u>S</u>**tructural **<u>I</u>**nteractome and genomic **<u>D</u>**ata brows**<u>ER</u>**), a tool for functional exploration of human disease on a genomic scale (http://interactomeinsider.yulab.org). Interactome INSIDER is based on a structurally-resolved, proteome-wide human interactome. We assembled this resource by building an interactome-wide set of protein interaction interfaces at the highest resolution possible for each interaction. We compiled structural interactomes by calculating interfaces in experimental co-crystal structures and homology models, when available. For the remaining ~94% of interactions, we applied a machine learning framework to predict partner-specific interfaces by applying recent advances in co-evolution- and docking-based feature construction[13,14]. Interactome INSIDER combines predicted interaction interfaces for 185,957 previously un-resolved interactions (including the full human interactome and 7 commonly studied model organisms) with disease mutations and functional annotations in an interactive toolbox designed to spur functional genomics research. It allows users to find enrichment of disease mutations at different scales: in protein interaction domains, in residues, and through atomic 3D clustering in protein interfaces.

RESULTS

To build Interactome INSIDER, we first constructed an interactome-wide set of protein interaction interfaces. While there are well-established methods for predicting whether or not two proteins interact[15,16], we focused on interactions that have been experimentally determined, but whose interfaces are unknown (Supplementary Note 1). For this task, there is a rich literature exploring the potential of many structural, evolutionary, and docking-based methods to predict protein interaction interfaces. However, so far, none of these methods have been used to produce a whole-interactome dataset of protein interaction interfaces (Supplementary Note 2).

We used ECLAIR (**E**nsemble **C**lassifier **L**earning **A**lgorithm to predict **I**nterface **R**esidues), a unified machine learning framework, to predict the interface of protein interactions. ECLAIR leverages several complementary and proven classification features, including sequence-based biophysical features, structural features, and recently proposed features for predicting binding partner-specific interfaces, including co-evolutionary[17,18] and docking-based metrics[14] (Supplementary Note 3, Supplementary Figs. 1–2). Unfortunately, many protein-protein interactions have missing features (especially structural features). In fact, this type of non-random missing-feature problem is present in many biological prediction studies, and cannot be adequately resolved by commonly-used imputation methods. To address this issue, ECLAIR is structured as an ensemble of 8 independent classifiers, each covering a common case of feature availability. This unique structure of ECLAIR enables it to be applied to any interaction, while using the most informative subset of available features for that interaction (Supplementary Note 4–5, Supplementary Figs. 3–4).

We comprehensively optimized hyperparameters for ECLAIR using a recently published Bayesian method, the tree-structured Parzen estimator approach (TPE)[19], which allowed us to simultaneously tune up to 8 hyperparameters for each sub-classifier (Supplementary Note 4). We trained and tested each ECLAIR sub-classifier using a set of known protein interaction interfaces, and observed that interfaces can be predicted by the single, top-performing sub-classifier available for each residue (Supplementary Note 4, Supplementary Fig. 5). Sub-classifier performance increases with the number of features used. We observe an area under the ROC curve (AUROC) of 0.64 for our top sequence-only sub-classifier and AUROC of 0.80 for our top sub-classifier using both sequence and structural features. In total, we used ECLAIR to predict the interfaces of 185,957 interactions with previously unknown interfaces, including for 115,576 human interactions (Supplementary Fig. 5). Specifically, residues classified by ECLAIR with a High or Very High interface potential have a precision of 0.69, and >90% of all 115,576 human interactions with predicted interfaces in Interactome INSIDER have 1 or more residues that fall into these categories. We supplemented known structural interfaces from co-crystalized proteins and homology models with our predictions to create structural interactomes at both the atomic and residue levels (Fig. 2a) in 7 model organisms and human (including all 122,647 human experimentally-determined binary interactions reported in major databases; see Online Methods). We used this resource to explore human disease with Interactome INSIDER.

Comprehensive evaluation of predicted interfaces

We established that our predictions are of high quality through both machine learning and biological evaluation. We first evaluated the trade-offs between false positive rate and true positive rate, and between precision and recall for each of the 8 independent sub-classifiers that compose ECLAIR. As expected, we find that as more informative features are added to subsequent classifiers, the areas under the ROC and precision-recall curves increase, justifying the use of classifiers trained on more features for residues where this information is available (Supplementary Fig. 6).

We next compared ECLAIR to several other prediction methods through two independent validations. First, we used several readily available predictors[20–24] to predict interfaces for interactions in our testing set. We find that for the set of interactions for which all classifiers can predict, ECLAIR performs as well or slightly better than these methods by measures of precision, recall, true positive rate and false positive rate (Fig. 2b, Supplementary Fig. 7). Finally, we applied ECLAIR to a standard external benchmark set of protein interaction interfaces[25] which has been used to evaluate the performance of 10 other interface prediction methods[26]. We find that ECLAIR outperforms all benchmarked methods in accuracy, and is comparable to the top performers in all other metrics (Supplementary Table 1). Furthermore, ECLAIR is applicable to any interaction, while methods in this benchmark rely on single-protein structure inputs, making them much less applicable to genome-wide studies. In fact, 86.1% of interactions without structural features contain at least 1 predicted interface residue at a ECLAIR score corresponding to a precision ≥ 0.6.

We also performed >2,000 mutagenesis experiments to measure the rate at which population variants in our predicted interfaces disrupt interactions compared to variants within known co-crystal interfaces and non-interfaces (see Material and Methods). Using our high-throughput yeast two-hybrid assay[27], we find that mutations in our predicted interfaces break their corresponding interactions at a significantly higher rate than those known to be away from the interface and at similar rates to mutations in known interfaces. Since it is known that mutations at protein interfaces are more likely to break interactions[6,27], our experimental results indicate that there is rich functional signal in our ECLAIR predictions (Fig. 2c).

Functional annotation of disease mutations in structural interactomes

Interactome INSIDER is a tool for identifying functionally enriched areas of protein interactomes, and for browsing our multi-scale structural interactome networks: 198,503 protein interactions whose interfaces have been either experimentally determined, homology modeled, or predicted using ECLAIR. Interactome INSIDER also includes 56,159 disease mutations from HGMD[28] and ClinVar[29], and 1,300,352 somatic cancer mutations from COSMIC[30] with their per-disease, pre-calculated enrichment in protein interaction interfaces at the residue level, domain level, and through atomic clustering. The site includes information on >600,000 population variants from the Exome Sequencing Project[31], 1000 Genomes Project[32] and more[33] (see Online Methods). Users can search Interactome INSIDER by protein to retrieve all interaction partners and their interfaces, or by disease to retrieve all interaction interfaces that are enriched for mutations of that disease. Additionally, users can upload their own set of mutations to find how they are distributed in the interactome and whether they are enriched in any protein interaction interfaces at the residue, domain, and atomic levels (Fig. 3).

We demonstrate the utility of Interactome INSIDER and the validity of its underlying database by investigating the functional and biological properties of our predicted interaction interfaces. We measured functional properties of our *in silico* predicted interfaces (those without prior experimental evidence) and compared these measurements to those of known interfaces from co-crystal structures. We find that disease mutations preferentially occur in our predicted interfaces at similar rates to known interface residues in PDB co-crystal structures (Fig. 4a), indicating the viability of using predicted interfaces to study molecular disease mechanisms. Furthermore, each higher-confidence bin of predicted interface residues is more likely to contain disease mutations than the previous, showing that ECLAIR prediction scores are correlated with true protein function. We looked at the locations of somatic cancer mutations from COSMIC in our interface-resolved human interactome. We specifically focused on recurrent cancer mutations as these are known to be more likely to be functional drivers[34,35]. We find a marked enrichment of recurrent cancer mutations in our predicted interfaces compared to outside these interfaces (Fig. 4b). The same trend is observed inside and outside of known interfaces from co-crystal structures, suggesting that the functional links between cancer and the potential disruption of protein interactions can be observed within the entire Interactome INSIDER human interface dataset. We also looked at the distribution of population variants, and show that their placement in and out of predicted interfaces matches that of known interfaces, with rarer mutations showing an enrichment in protein interfaces (Fig. 4c). Furthermore, population variants in our predicted interfaces are more likely to be damaging to protein function than variants outside of predicted interfaces, as predicted by PolyPhen-2[36] (Fig. 4d) and EVmutation[37] (Fig. 4e), matching the established trend for experimentally determined interfaces[38]. We validated many of these biological trends for interactions lacking structural features (Supplementary Figs. 8–10, Supplementary Note 6), suggesting the utility of Interactome INSIDER even in feature-poor interactions and across different resolution scales.

We used Interactome INSIDER to search for sub-networks in the human interactome that are enriched for mutations associated with a single disease, by calculating the enrichment of disease mutations in interaction interfaces interactome-wide. This identified the TGF-β/BMP signaling pathway, which is known to be involved in juvenile polyposis syndrome (JPS)[39], and contains multiple proteins harboring JPS mutations (Fig. 5a). We focused on a specific group of mutations in the SMAD4-SMAD8 interface, which can be found using 3D atomic clustering. Using our mutagenesis Y2H assay, we were able to test a JPS mutation (SMAD4 Y353S)[40], which is at the interface of SMAD4-SMAD8, and show that it breaks this interaction, implicating SMAD8 in JPS (Fig. 5a, Supplementary Fig 11). Although SMAD8 (also known as SMAD9) has not been reported to harbor JPS mutations in HGMD[28], its involvement in the disease has been suggested[41], showing the ability of Interactome INSIDER to implicate new proteins in disease. Y353S is not predicted by ECLAIR to be at the interface of SMAD4 and another of its binding partners, RASSF5. Indeed, through our Y2H experiment, Y353S does not break this interaction, demonstrating the functional insight Interactome INSIDER can provide about differential interfaces and how they might be relevant to understanding the molecular mechanisms of disease.

Disease etiology revealed by partner-specific interfaces

Interactome INSIDER enables interrogation of different interfaces for the same protein, dependent upon its binding partner (Fig. 5b). For the study of protein function and disease, this is especially important as a protein may maintain different functional pathways through different interfaces, and disruption of one interface may leave others intact[4,8]. To test this on a large-scale, we looked at pairs of disease mutations in the human interactome that appear at interaction interfaces, as predicted by ECLAIR. Similar to previous reports[8], we observed that mutation pairs in the interface of two interacting proteins are much more likely to cause the same disease than mutation pairs in other interfaces of the same proteins that do not mediate the given interaction (Fig. 5c). We also find that mutation pairs on the same protein, but in separate interfaces with different binding partners tend to cause different diseases (Fig. 5d). This trend is observed in both known and predicted interfaces. These results indicate that Interactome INSIDER can be used to form functional hypotheses about the specificity of mutations to specific interactions and molecular pathways.

We next used Interactome INSIDER to find sub-networks in the human interactome enriched for mutations associated with a single disease. We uncovered a set of interacting proteins known to harbor mutations causal for hypertrophic cardiomyopathy (HCM)[42] and thereby recapitulated the core constituents of a known KEGG pathway related to the same disease (Fig. 6). These proteins were identified by enrichment of disease mutations in their shared interaction interfaces and, in the case of TNNI3-TNNC1, using cross-interface atomic clustering of disease mutation positions in 3D (features available via the Interactome INSDIER website). In addition to identifying known members of the HCM pathway, Interactome INSIDER also identified several additional proteins, including CSRP3, MYOM1, ANKRD and TCAP, which are not part of the known KEGG pathway, but carry HCM mutations enriched at their respective interaction interfaces with members of the pathway. We also identify a protein, TNNT1, which, although it contains no HCM mutations of its own, can be implicated in HCM through its interactions with two proteins TPM1 and TNNC1, which are enriched for HCM mutations at their interfaces with TNNT1. Finally, we note that Interactome INSIDER reveals cases of partner-specific interfaces in this pathway. For instance, the known HCM pathway protein TTN’s interface with ACTA1 is enriched for HCM mutations, and ACTA1 mutations are increasingly linked to HCM[43]. On the other hand, a separate interface of ACTA1 with its binding partner dystrophin is enriched with mutations causing a distinct disorder, actin myopathy[44]. This shows how ACTA1 can play roles in two different diseases through separate interaction interfaces with TTN and dystrophin, and demonstrates Interactome INSIDER’s unique ability to discover such cases of differential function mirroring differential interfaces.

DISCUSSION

We anticipate Interactome INSIDER will help to bridge the divide between genomic-scale datasets and structural proteomic analyses. Now that large-scale sequencing data from many contexts are readily available, for instance from whole-genome/whole-exome population variant studies[31,45] and cancer studies[46,47], researchers have become increasingly interested in ways to assess the potential functional consequences of variants on a genomic scale[48,49]. Recently we and others have developed methods to predict functional cancer driver mutations by finding hotspots of mutations in the structural proteome[35,50]. With the comprehensive map of protein interfaces presented, we can now go a step further to predict specific etiologies of cancer and disease based on induced biophysical effects[51,52] that may break interactions. Because our interface map is partner-specific, it can also be applied to predict pleiotropic effects, wherein several mutations in a single protein may affect different pathways depending upon which binding interfaces are mutated[8]. This could be the basis for designing new therapeutics and for rational drug design to selectively target specific protein functional sites[53].

We have shown that hyperparameter optimization, which is surprisingly lacking in much of the current literature, can drastically improve the performance of classifiers for biological classification studies. The tiered ensemble form of the ECLAIR classifier represents a broadly applicable paradigm in practical machine learning that could be readily applied to solving other problems with large amounts of non-uniformly missing data, which very frequently occur in biology due to study biases.

With future increases to the scale of biological databases from which we derive features, we expect that Interactome INSIDER will come to encompass even higher confidence predictions for many more interactions, thereby becoming increasingly applicable to functional studies. This may also address some limitations of structural databases today. For instance, the PDB is depleted of disordered proteins[54], and it has been shown that disordered regions can form interfaces[55]. Since ECLAIR has not been trained on disordered interfaces, it is unlikely to predict new disordered interfaces. However, the ensemble classifier structure of ECLAIR uniquely positions it to incorporate all newly-available evidence into interface predictions without sacrificing quality or scale, ensuring a high quality map of interaction interfaces now and in the future. Furthermore, the addition of new variants, especially cancer mutations and population variants from large-scale sequencing studies, will only increase the value of performing systems-level explorations with Interactome INSIDER.

ONLINE METHODS

Interaction datasets

We compiled binary protein interactions available for *H. sapiens*, *D. melanogaster*, *S. cerevisiae*, *C. elegans*, *A. thaliana*, *E. coli*, *S. pombe*, and *M. musculus* from 7 primary interaction databases. These databases include IMEx[56] partners DIP[57], IntAct[58], and MINT[59], IMEx observer BioGRID[60], and additional sources iRefWeb[61], HPRD[62], and MIPS[63]. Furthermore, iRefWeb combines interaction data from BIND[64], CORUM[65], MPact[66], OPHID[67], and MPPI[68]. We filtered these interactions using the PSI-MI[69] evidence codes of assays that can determine experimental binary interactions (Supplementary Table 2), as these are interactions where proteins are known to share a direct binding interface that we can then predict[5]. In total, we curated 198,503 interactions in these 8 species including the full experimentally determined binary interactome in human (122,647 interactions)

(Supplementary Note 1). Those interactions with known interface residues based on available co-crystal structures in the Protein Data Bank (PDB)[70] were set aside for use in training and testing the classifier. Interactions without known interface residues comprise the set for which we make predictions.

Testing and training sets for interface residue prediction

For those interactions with known co-crystal structures in the PDB, we calculate interface residues for their specific binding partners. To identify UniProt protein sequences in the PDB, we use SIFTS[71], which provides a mapping of PDB-indexed residues to UniProt-indexed residues[33]. For each interaction and representative co-crystal structure, interface residues are calculated by assessing the change in solvent accessible surface area of the proteins in complex and apart using NACCESS[72]. Any residue that is at the surface of a protein (≥15% exposed surface) and whose solvent accessible surface area (SASA) decreases by ≥1.0 $Å^2$ in complex is considered to be at the interface. We aggregate interface residues across all available structures in the PDB for a given interaction, wherein a residue is considered to be at the interface of the interaction if it has been calculated to be at the interface in one or more co-crystal structures of that interaction (all other residues are considered to be away from the interface). In building our final training and testing sets, we only consider interactions for which aggregated co-crystal structures have combined to cover at least 50% of UniProt residues for both interacting proteins.

The training and testing sets each include a random selection of 400 interactions with known co-crystal structures, of which 200 are heterodimers and 200 are homodimers (Supplementary Table 3). To ensure an unbiased performance evaluation, we disallowed any homologous interactions (i.e. interactions whose structures could be used as templates for homology modeling) between the training and testing set. We also disallowed repeated proteins between the two sets to avoid simply reporting a remembered shared interface between a protein and multiple binding partners, thereby artificially elevating the performance of our classifier on the testing set.

Hyperparameter optimization with TPE

To train our ensemble of classifiers that comprise ECLAIR, we used the tree-structured Parzen estimator approach (TPE)[19], a Bayesian method for optimizing hyperparameters for machine learning algorithms. TPE models the probability distribution $p(x|y)$ of hyperparameters given evaluated loss from a defined objective function, $L(x)$. We selected the following loss function to minimize based on classical hyperparameter inputs and residue window sizes:

$$L(\theta, w) = 1 - \min_{n \in \{1, 2, 3\}} \left\{ AUROC_{\theta, w, n} \right\}$$

where x is comprised of θ, a set of hyperparameters, and w, a set of residue window sizes. The evaluation metric, $AUROC_n$, is the area under the roc curve for the nth left-out evaluation fold in a three-fold cross-validation scheme. We then used TPE to randomly sample an initial uniform distribution of each of our hyperparameters and window sizes and evaluate the loss function for each random set of inputs. TPE then replaces this initial distribution with a new distribution built on the results from regions of the sampled distribution that minimize *L(x)*:

$$p(x \mid y) = \begin{cases} l(x) & if\ y < y^* \\ g(x) & if\ y \le y^* \end{cases}$$

where y^* is a quantile γ of the observed y values so that $p(y < y^*) = \gamma$. Importantly, y^* is guaranteed to be greater than the minimum observed loss, so that some points are used to build *l(x)*. TPE then chooses candidate hyperparameters to sample as those representing the greatest expected improvement, *EI*, according to the expression:

$$EI_{y^*}(x) = \frac{\gamma y^* l(x) - l(x) \int_{-\infty}^{y^*} y p(y) dy}{\gamma l(x) + (1 - \gamma) g(x)} \propto \left(\gamma + \frac{g(x)}{l(x)} (1 - \gamma) \right)^{-1}$$

To maximize *EI*, the algorithm picks points *x* with high probability under *l(x)* and low probability under *g(x)*. Each iteration of the algorithm returns x^*, the next set of hyperparameters to sample, with the greatest *EI* based on previously sampled points.

Training the classifier

The ECLAIR classifier was trained in three stages, using a custom wrapper of the scikit-learn[73] random forest[74] classifier to allow for use of TPE to search both algorithm hyperparameters and residue window sizes simultaneously. In all cross-validations performed, we allowed TPE to search the following hyperparameters, beginning with uniform distributions of the indicated ranges: (1) minimum samples per leaf (0–1000), (2) maximum fraction of features per tree (0–1), and (3) split criterion (entropy or gini diversity index). The number of estimators (decision trees) in each random forest was fixed at either 200 for training the feature selection classifiers, or 500 for training the full ensemble. We also allowed TPE to search over residue window sizes (± 0–5 residues for a total window of up to 11 residues, centered on the residue of interest). This was achieved by allowing extra features for neighboring residues to be included at the time of algorithm initialization.

In the first stage of training, cross-validation using TPE was performed on classifiers trained using only features from 1 of the 5 feature categories. The feature or set of features from each category with the minimum loss was selected to represent that category in building the ensemble classifier (Supplementary Table 4). In the second stage, the ensemble classifier was built of 8 random forest classifiers, each trained on different subsets of feature categories, and hyperparameters and window sizes were again chosen using cross-validation and TPE (Supplementary Table 5). In the final stage, following performance measurement on the testing set, the 8 sub-classifiers were retrained using the full set of 3,447 interactions with at least 50% UniProt residue coverage in the PDB, using the same hyperparameters and window sizes found in the previous step.

Evaluating the ensemble

After training and optimizing using only the training set, we predicted interface residues in a completely orthogonal testing set. For each sub-classifier of the ensemble, all residues in the testing set that could be predicted (given the full set of necessary features or a superset) were ranked according to their raw prediction scores to produce ROC and precision-recall plots.

Benchmarking against other methods

Interfaces for interactions in our testing set were computed using several popular interface prediction methods[20–24]. We compiled a set of representative protein structures from the PDB for each protein in our testing set, selecting the structure with the highest UniProt residue content based on SIFTS and excluding any PDB structures of interacting protein pairs from our testing set. We then evaluated the precision, recall, and false positive rate for proteins that were able to be classified by all methods. These represent point estimates of these metrics for the external methods with binary prediction scores.

We also compared ECLAIR to 10 popular methods for interface prediction by predicting interfaces in a standard benchmark set of protein complexes[25] (Supplementary Table 1). Here, we followed the experimental procedures laid out by Maheshwari *et al.*[26], and excluded complexes in which the receptor is <50 or >600 amino acids, where the interface is made up of <20 residues, or where multiple interfaces are present.

Predicting new interfaces

We retrained the ensemble using all available co-crystal structures, including those from both testing and training sets, a standard machine learning practice that makes maximal use of labeled data[75]. Using this fully trained ensemble of classifiers, we predicted interface residues for the remaining 185,957 interactions not resolved by either PDB structures or homology models. Sub-classifiers were ordered based on the number and information content of features used in their training. Each residue was then predicted by only the top ranking classifier of the ensemble trained on the full set or a subset of available features for that residue.

Interface enrichment and 3D atomic clustering

Interface domain enrichment, residue enrichment, and 3D atomic clustering can be calculated through the Interactome INSIDER web interface. For enrichments presented in this study, we accessed all disease mutations from the Human Gene Mutation Database (HGMD)[28] and ClinVar[29], recurrent cancer mutations appearing ge; 6 times in COSMIC[30], and population variants from the Exome Sequencing Project[31] to compute the log odds ratio:

$$LOR = ln\left(\frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}}\right)$$

where p_1 is the probability of a mutation or variant being at the interface and p_2 is the probability of any residue being at the interface. We computed the log odds ratio for residues in each of the interface prediction potential categories. We also computed the log odds ratio for interactions with known interfaces from PDB co-crystal structures, defined as all known interface residues from NACCESS calculations and all residues in Pfam[76] domains with ≥ 5 interface residues. For the disease mutation enrichment analysis (Fig. 4a, we used all disease mutations available from HGMD, and the following numbers of mutations occurred in each category: 10,196 Very Low, 10,547 Low, 2,970 Medium, 1,135 High, and 305 Very High. We also computed enrichment of 18,638 mutations in known interfaces and 17,760 mutations in known non-interfaces (from co-crystal structure evidence).

To perform 3D atomic clustering of amino acid loci of interest, we used an established method[35] for clustering and empirical *p*-value calculation and applied it to multi-protein clustering, wherein clusters can occur across an interaction interface. Here, we perform complete-linkage clustering[77] in the shared 3D space of both proteins, and iteratively, and randomly rearrange mutations in each protein to produce an empirical null distribution of cluster sizes.

Mutagenesis validation experiments

We performed mutagenesis experiments in which we introduced random human population variants from the Exome Sequencing Project[31] into known and predicted interfaces. We randomly selected mutations of predicted interface residues in each of the top four ECLAIR categories (Low – Very High). As positive and negative controls, we also selected random mutations of known interface and non-interface residues in co-crystal structures in the PDB. The selected mutations were then introduced into the proteins according to our previously published Clone-seq pipeline[27] and their impact (either disrupting or maintaining the interaction) was assessed using our yeast two-hybrid assay (Supplementary Note 7). In this manner, we tested the impact of 2,164 mutations: 1,664 in our predicted interfaces and 500 in known interface and non-interface residues from co-crystal structures. In Figure 2c, we report the fraction of tested interface residue mutations that caused a disruption of the given interaction for each of the interface residue bins.

Web server

Interactome INSIDER is deployed as an interactive web server (http://interactomeinsider.yulab.org) containing known and predicted interfaces for 198,503 protein interactions in 8 species, as well as variants and functional annotations mapped relative to the residues in the human proteome. For each interaction, the most reliable, high-resolution model is presented, i.e. co-crystal structures are always displayed in lieu of homology models, and all remaining unresolved interactions are predicted by our ECLAIR classifier. Co-crystal structures are derived from the PDB, with extraneous chains removed for each interaction, and homology models are computed by MODELLER[11] and downloaded from Interactome3D[12]. For both types of structural model, we computed all residues at the interface over all available models, and allow users to view any model from which a unique interface residue has been calculated. For predicted interfaces, a non-redundant set of single protein models are shown when available, with locations of predicted interface residues indicated. In total, the resource contains 7,135 interactions with co-crystal structures, 5,411 with homology models, and 185,957 with predicted interfaces.

Interactome INSIDER also includes pre-calculated enrichment of mutations derived from several sources: 56,159 disease mutations from HGMD[28] and ClinVar[29] and 1,300,352 somatic cancer mutations from COSMIC[30]. It also includes 194,396 population variants from the 1000 Genomes Project[32], 425,115 from the Exome Sequencing Project[31], and 54,165 catalogued by UniProt[33]. Predictions of deleteriousness for all variants and any user-submitted variants within the curated interactomes are obtained from PolyPhen-2[36] and SIFT[78], and biophysical property change guides (i.e. polar to non-polar, hydrophobic to hydrophilic) are also displayed for convenience. Mutation and variant enrichment analyses can be triggered by the user for existing variants or for user-submitted sets within interacting protein domains, residues, and 3D clustering using the atomic coordinates of structures when available.

Downloads of known and predicted interface residues on a per-interaction basis are available as plain text and as .bed files that can be visualized alongside other genomic landmarks in the UCSC genome browser[79]. Per-protein visualization tracks, with interface residues of all interaction partners are also available on the “Downloads” page, along with bulk downloads of interfaces for entire species.

Statistics

Statistical analyses were performed using a two-sided Z test or a two-sided Mann-Whitney U test, as indicated in the figure captions. Exact P values are provided for all compared groups, and comparisons with a two-side P value > 0.05 are considered not significant, with all others considered not significant (n.s.).

Code and data availability

Custom code used in this study is freely available at https://github.com/hyulab/ECLAIR and as Supplementary Software. Data produced by this study is available for browsing and bulk download at http://interactomeinsider.yulab.org. Additional information is available in the Life Sciences Reporting Summary.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors would like to thank G. Hooker, D. Bindel, and K. Weinberger for helpful discussions. This work was supported by National Institute of General Medical Sciences grants (R01 GM097358, R01 GM104424, R01 GM124559); National Cancer Institute grant (R01 CA167824); Eunice Kennedy Shriver National Institute of Child Health and Human Development grant (R01 HD082568); National Human Genome Research Institute grant (UM1 HG009393); National Science Foundation grant (DBI-1661380); Simons Foundation Autism Research Initiative grant (367561) to H.Y.

References

1. Rolland T, et al. A proteome-scale map of the human interactome network. Cell. 2014; 159:1212–1226. [PubMed: 25416956]
2. Arabidopsis Interactome Mapping C. Evidence for network evolution in an Arabidopsis interactome map. Science. 2011; 333:601–607. [PubMed: 21798944]
3. Yu H, et al. High-quality binary protein interaction map of the yeast interactome network. Science. 2008; 322:104–110. [PubMed: 18719252]
4. Vo TV, et al. A Proteome-wide Fission Yeast Interactome Reveals Network Evolution Principles from Yeasts to Human. Cell. 2016; 164:310–323. [PubMed: 26771498]
5. Das J, Yu H. HINT: High-quality protein interactomes and their applications in understanding human disease. BMC Syst Biol. 2012; 6:92. [PubMed: 22846459]
6. Sahni N, et al. Widespread macromolecular interaction perturbations in human genetic disorders. Cell. 2015; 161:647–660. [PubMed: 25910212]
7. Kim PM, Lu LJ, Xia Y, Gerstein MB. Relating three-dimensional structures to protein networks provides evolutionary insights. Science. 2006; 314:1938–1941. [PubMed: 17185604]
8. Wang X, et al. Three-dimensional reconstruction of protein networks provides insight into human genetic disease. Nat Biotechnol. 2012; 30:159–164. [PubMed: 22252508]
9. Kuhlbrandt W. Cryo-EM enters a new era. eLife. 2014; 3:e03678. [PubMed: 25122623]
10. Halperin I, Ma B, Wolfson H, Nussinov R. Principles of docking: An overview of search algorithms and a guide to scoring functions. Proteins. 2002; 47:409–443. [PubMed: 12001221]
11. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol. 1993; 234:779–815. [PubMed: 8254673]
12. Mosca R, Céol A, Aloy P. Interactome3D: adding structural details to protein networks. Nature methods. 2013; 10:47–53. [PubMed: 23399932]
13. Hopf TA, et al. Sequence co-evolution gives 3D contacts and structures of protein complexes. eLife. 2014; 3
14. Hwang H, Vreven T, Weng Z. Binding interface prediction by combining protein-protein docking results. Proteins. 2014; 82:57–66.
15. Zhang QC, et al. Structure-based prediction of protein-protein interactions on a genome-wide scale. Nature. 2012; 490:556–560. [PubMed: 23023127]
16. Garzon JI, et al. A computational interactome and functional annotation for the human proteome. eLife. 2016; 5
17. Morcos F, et al. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proceedings of the National Academy of Sciences of the United States of America. 2011; 108:E1293–1301. [PubMed: 22106262]
18. Lockless SW, Ranganathan R. Evolutionarily conserved pathways of energetic connectivity in protein families. Science. 1999; 286:295–299. [PubMed: 10514373]
19. Bergstra JS, Bardenet R, Bengio Y, Kégl B. Algorithms for hyper-parameter optimization. Advances in Neural Information Processing Systems. 2011:2546–2554.
20. Kufareva I, Budagyan L, Raush E, Totrov M, Abagyan R. PIER: protein interface recognition for structural proteomics. Proteins. 2007; 67:400–417. [PubMed: 17299750]
21. Liang S, Zhang C, Liu S, Zhou Y. Protein binding site prediction using an empirical scoring function. Nucleic Acids Res. 2006; 34:3698–3707. [PubMed: 16893954]
22. Porollo A, Meller J. Prediction-based fingerprints of protein-protein interactions. Proteins. 2007; 66:630–645. [PubMed: 17152079]
23. de Vries SJ, Bonvin AM. CPORT: a consensus interface predictor and its performance in prediction-driven docking with HADDOCK. PloS one. 2011; 6:e17695. [PubMed: 21464987]
24. Jordan RA, El-Manzalawy Y, Dobbs D, Honavar V. Predicting protein-protein interface residues using local surface structural similarity. BMC Bioinformatics. 2012; 13:41. [PubMed: 22424103]
25. Hwang H, Vreven T, Janin J, Weng Z. Protein-protein docking benchmark version 4.0. Proteins. 2010; 78:3111–3114. [PubMed: 20806234]

26. Maheshwari S, Brylinski M. Predicting protein interface residues using easily accessible on-line resources. Brief Bioinform. 2015; 16:1025–1034. [PubMed: 25797794]

27. Wei X, et al. A massively parallel pipeline to clone DNA variants and examine molecular phenotypes of human disease mutations. PLoS Genet. 2014; 10:e1004819. [PubMed: 25502805]

28. Stenson PD, et al. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Hum Genet. 2014; 133:1–9. [PubMed: 24077912]

29. Landrum MJ, et al. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res. 2016; 44:D862–868. [PubMed: 26582918]

30. Forbes SA, et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res. 2015; 43:D805–811. [PubMed: 25355519]

31. Fu W, et al. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature. 2013; 493:216–220. [PubMed: 23201682]

32. Genomes Project C et al. A global reference for human genetic variation. Nature. 2015; 526:68–74. [PubMed: 26432245]

33. UniProt-Consortium. UniProt: a hub for protein information. Nucleic Acids Res. 2015; 43:D204–212. [PubMed: 25348405]

34. Hodis E, et al. A landscape of driver mutations in melanoma. Cell. 2012; 150:251–263. [PubMed: 22817889]

35. Meyer MJ, et al. mutation3D: Cancer Gene Prediction Through Atomic Clustering of Coding Variants in the Structural Proteome. Hum Mutat. 2016; 37:447–456. [PubMed: 26841357]

36. Adzhubei IA, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010; 7:248–249. [PubMed: 20354512]

37. Hopf TA, et al. Mutation effects predicted from sequence co-variation. Nat Biotechnol. 2017; 35:128–135. [PubMed: 28092658]

38. David A, Razali R, Wass MN, Sternberg MJ. Protein-protein interaction sites are hot spots for disease-associated nonsynonymous SNPs. Hum Mutat. 2012; 33:359–363. [PubMed: 22072597]

39. Wang RN, et al. Bone Morphogenetic Protein (BMP) signaling in development and human diseases. Genes & diseases. 2014; 1:87–105. [PubMed: 25401122]

40. Roth S, et al. SMAD genes in juvenile polyposis. Genes, chromosomes & cancer. 1999; 26:54–61. [PubMed: 10441006]

41. Ngeow J, et al. Exome Sequencing Reveals Germline SMAD9 Mutation That Reduces Phosphatase and Tensin Homolog Expression and Is Associated With Hamartomatous Polyposis and Gastrointestinal Ganglioneuromas. Gastroenterology. 2015; 149:886–889 e885. [PubMed: 26122142]

42. Maron BJ. Hypertrophic cardiomyopathy: a systematic review. Jama. 2002; 287:1308–1320. [PubMed: 11886323]

43. Donkervoort S, et al. Cardiomyopathy in patients with ACTA1-myopathy. Abstracts/ Neuromuscular Disorders. 2015; 25:S184–S316.

44. Sparrow JC, et al. Muscle disease caused by mutations in the skeletal muscle alpha-actin gene (ACTA1). Neuromuscular disorders: NMD. 2003; 13:519–531. [PubMed: 12921789]

45. Lek M, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016; 536:285–291. [PubMed: 27535533]

46. Forbes S, et al. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Research. 2011; 39:50.

47. Kandoth C, et al. Mutational landscape and significance across 12 major cancer types. Nature. 2013; 502:333–339. [PubMed: 24132290]

48. Lawrence M, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013; 499:214–218. [PubMed: 23770567]

49. Tasan M, et al. Selecting causal genes from genome-wide association studies via functionally coherent subnetworks. Nat Methods. 2015; 12:154–159. [PubMed: 25532137]

50. Kamburov A, et al. Comprehensive assessment of cancer missense mutation clustering in protein structures. Proceedings of the National Academy of Sciences of the United States of America. 2015; 112:E5486–5495. [PubMed: 26392535]

51. Kucukkal TG, Petukh M, Li L, Alexov E. Structural and physico-chemical effects of disease and non-disease nsSNPs on proteins. Current opinion in structural biology. 2015; 32:18–24. [PubMed: 25658850]

52. Li M, Petukh M, Alexov E, Panchenko AR. Predicting the Impact of Missense Mutations on Protein-Protein Binding Affinity. J Chem Theory Comput. 2014; 10:1770–1780. [PubMed: 24803870]

53. Lounnas V, et al. Current progress in Structure-Based Rational Drug Design marks a new mindset in drug discovery. Computational and structural biotechnology journal. 2013; 5:e201302011. [PubMed: 24688704]

54. Peng K, Obradovic Z, Vucetic S. Exploring bias in the Protein Data Bank using contrast classifiers. Pac Symp Biocomput. 2004:435–446. [PubMed: 14992523]

55. Dunker AK, et al. The unfoldomics decade: an update on intrinsically disordered proteins. BMC genomics. 2008; 9(Suppl 2):S1.

56. Orchard S, et al. Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. Nat Methods. 2012; 9:345–350. [PubMed: 22453911]

57. Salwinski L, et al. The Database of Interacting Proteins: 2004 update. Nucleic Acids Res. 2004; 32:D449–451. [PubMed: 14681454]

58. Kerrien S, et al. The IntAct molecular interaction database in 2012. Nucleic Acids Res. 2012; 40:D841–846. [PubMed: 22121220]

59. Licata L, et al. MINT, the molecular interaction database: 2012 update. Nucleic Acids Res. 2012; 40:D857–861. [PubMed: 22096227]

60. Chatr-Aryamontri A, et al. The BioGRID interaction database: 2015 update. Nucleic Acids Res. 2015; 43:D470–478. [PubMed: 25428363]

61. Turner B, et al. iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence. Database (Oxford). 2010; 2010:baq023. [PubMed: 20940177]

62. Keshava Prasad TS, et al. Human Protein Reference Database--2009 update. Nucleic Acids Res. 2009; 37:D767–772. [PubMed: 18988627]

63. Mewes HW, et al. MIPS: curated databases and comprehensive secondary data resources in 2010. Nucleic Acids Res. 2011; 39:D220–224. [PubMed: 21109531]

64. Alfarano C, et al. The Biomolecular Interaction Network Database and related tools 2005 update. Nucleic Acids Res. 2005; 33:D418–424. [PubMed: 15608229]

65. Ruepp A, et al. CORUM: the comprehensive resource of mammalian protein complexes--2009. Nucleic Acids Res. 2010; 38:D497–501. [PubMed: 19884131]

66. Guldener U, et al. MPact: the MIPS protein interaction resource on yeast. Nucleic Acids Res. 2006; 34:D436–441. [PubMed: 16381906]

67. Brown KR, Jurisica I. Online predicted human interaction database. Bioinformatics. 2005; 21:2076–2082. [PubMed: 15657099]

68. Pagel P, et al. The MIPS mammalian protein-protein interaction database. Bioinformatics. 2005; 21:832–834. [PubMed: 15531608]

69. Hermjakob H, et al. The HUPO PSI's molecular interaction format--a community standard for the representation of protein interaction data. Nat Biotechnol. 2004; 22:177–183. [PubMed: 14755292]

70. Berman HM. The Protein Data Bank. Nucleic Acids Research. 2000; 28

71. Velankar S, et al. SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. Nucleic Acids Research. 2013; 41:9.

72. Lee B, Richards FM. The interpretation of protein structures: estimation of static accessibility. J Mol Biol. 1971; 55:379–400. [PubMed: 5551392]

73. Pedregosa F, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011; 12:2825–2830.

74. Breiman L. Random Forests. Mach Learn. 2001; 45:5–32.

75. Witten IH, , Frank E, , Hall MA, , Pal CJ. Data Mining: Practical Machine Learning Tools and Techniques Elsevier Science; 2016

76. Punta M, et al. The Pfam protein families database. Nucleic Acids Res. 2012; 40:D290–301. [PubMed: 22127870]

77. Sørensen T. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. Biol Skr. 1948; 5:1–34.

78. Kumar P, Henikoff S, Ng P. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nature protocols. 2009; 4:1073–1081. [PubMed: 19561590]

79. Tyner C, et al. The UCSC Genome Browser database: 2017 update. Nucleic Acids Res. 2017; 45:D626–D634. [PubMed: 27899642]

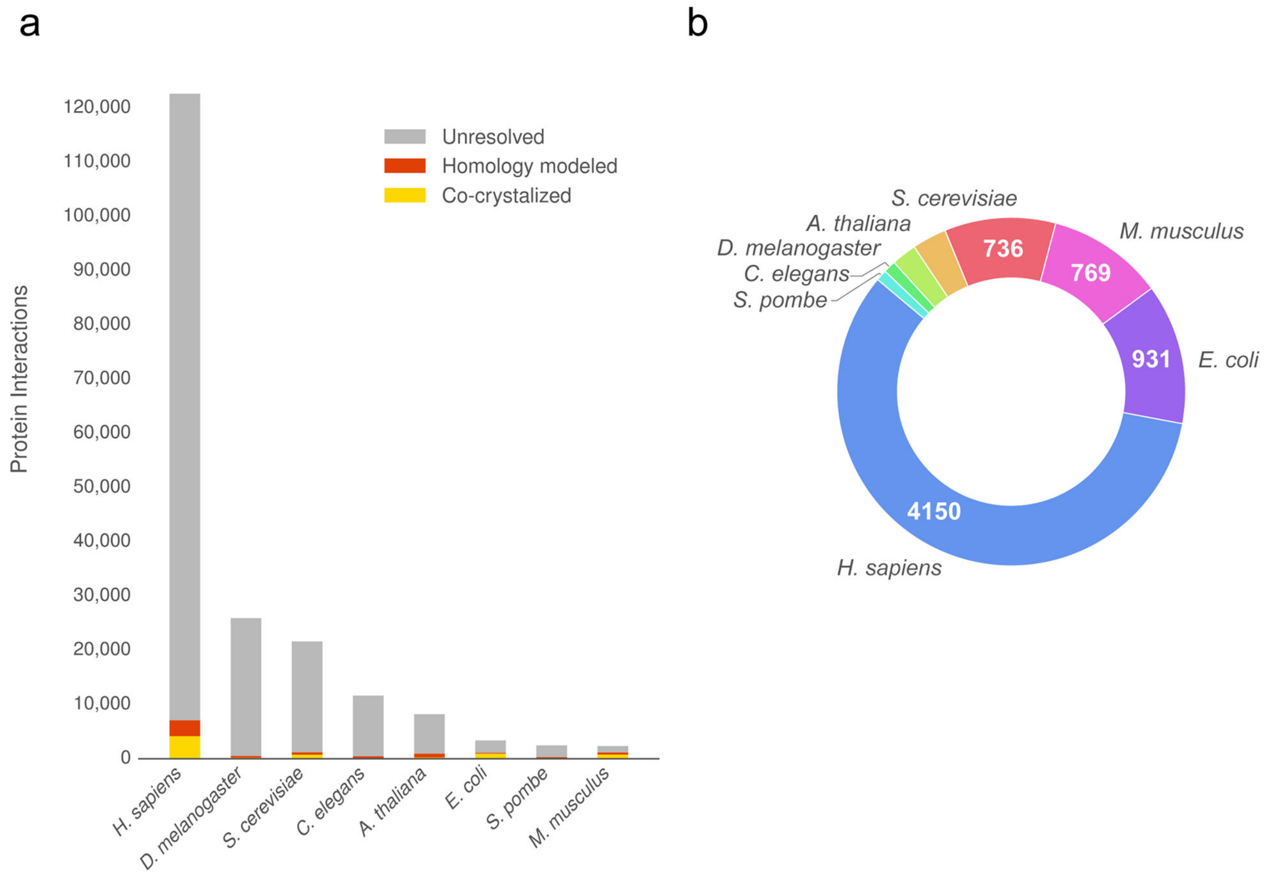


Figure 1.

The current size of structural interactomes. (a) The plot shows the coverage (number of protein interactions) of known high quality binary interactomes with pre-computed co-complexed protein structures. (b) The number of interactions from the 8 largest interactomes with experimentally solved structures.

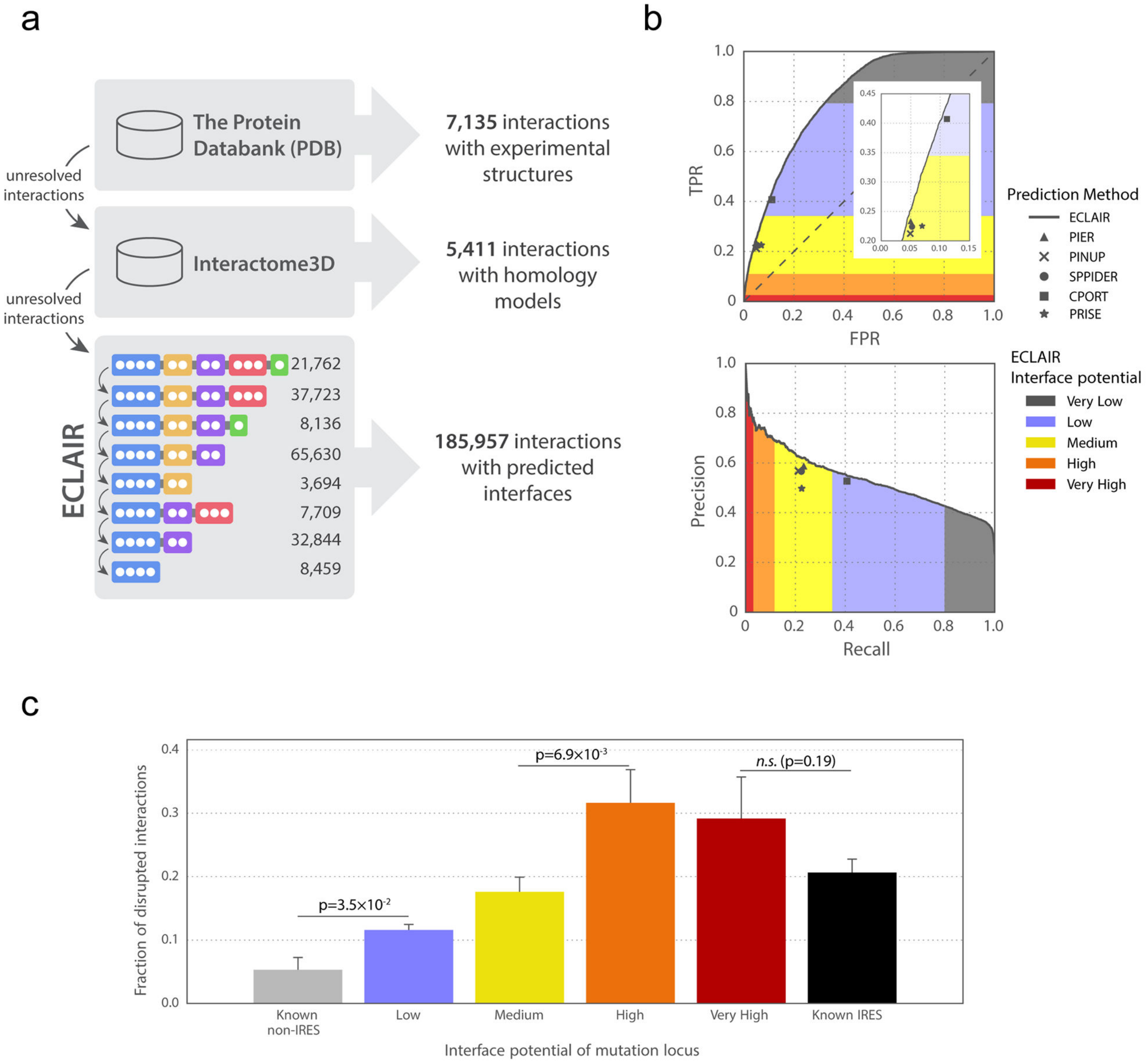
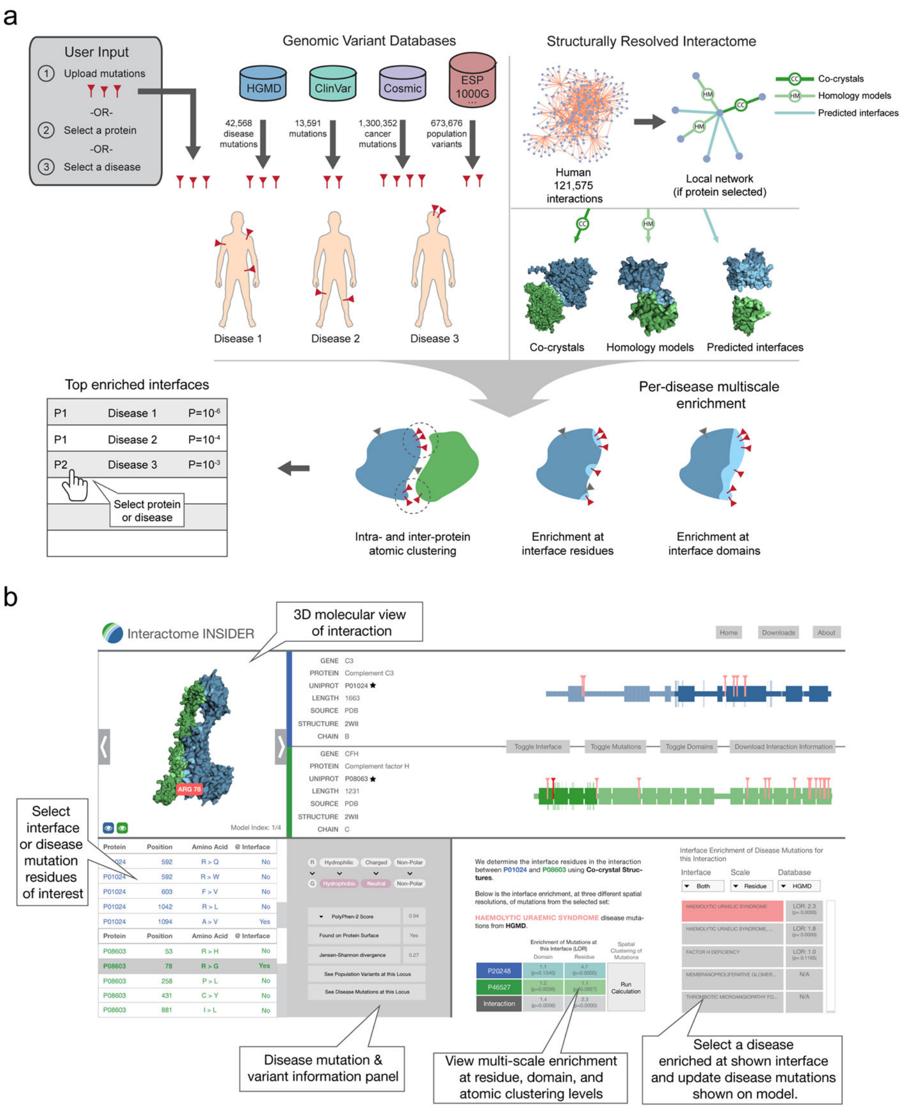


Figure 2.

ECLAIR prediction results. (a) Workflow for classifying interfaces for all interactions in 8 species. Interactions without experimentally determined or homology modeled interfaces are classified by ECLAIR. (b) ROC and precision-recall curves comparing ECLAIR with the indicated interface residue prediction methods. (c) Fraction of interactions disrupted by the introduction of random population variants in known and predicted interfaces. (Significance determined by two-sided Z-test; *n.s.* denotes not significant)


a
User Input
1 Upload mutations
-OR-
2 Select a protein
-OR-
3 Select a disease
Genomic Variant Databases
HGMD
ClinVar
Cosmic
ESP 1000G
42,568 disease mutations
13,591 mutations
1,300,352 cancer mutations
673,676 population variants
Disease 1
Disease 2
Disease 3
Structurally Resolved Interactome
Human 121,575 interactions
Local network (if protein selected)
CC Co-crystals
HM Homology models
Predicted interfaces
Co-crystals
Homology models
Predicted interfaces
Top enriched interfaces
P1 Disease 1 P=10-6
P1 Disease 2 P=10-4
P2 Disease 3 P=10-3
Select protein or disease
Per-disease multiscale enrichment
Intra- and inter-protein atomic clustering
Enrichment at interface residues
Enrichment at interface domains
b
Interactome INSIDER
Home
Downloads
About
3D molecular view of interaction
ARG 78
Model Index: 1/4
Toggle Interface
Toggle Mutations
Toggle Domains
Download Interaction Information
Select interface or disease mutation residues of interest
R Hydrophilic Charged Non-Polar
G Hydrophobic Neutral Non-Polar
PolyPhen-2 Score 0.94
Found on Protein Surface Yes
Jensen-Shannon divergence 0.27
See Population Variants at this Locus
See Disease Mutations at this Locus
Run Calculation
Disease mutation & variant information panel
View multi-scale enrichment at residue, domain, and atomic clustering levels
Select a disease enriched at shown interface and update disease mutations shown on model.

GENE C3
PROTEIN Complement C3
UNIPROT P01024 ★
LENGTH 1663
SOURCE PDB
STRUCTURE 2WII
CHAIN B

GENE CFH
PROTEIN Complement factor H
UNIPROT P08063 ★
LENGTH 1231
SOURCE PDB
STRUCTURE 2WII
CHAIN C

Protein	Position	Amino Acid	@ Interface
P01024	592	R > Q	No
P01024	592	R > W	No
P01024	603	F > V	No
P01024	1042	R > L	No
P01024	1094	A > V	Yes
Protein	**Position**	**Amino Acid**	**@ Interface**
P08603	53	R > H	No
P08603	78	R > G	Yes
P08603	258	P > L	No
P08603	431	C > Y	No
P08603	881	I > L	No

We determine the interface residues in the interaction between P01024 and P08603 using Co-crystal Structures.

Below is the interface enrichment, at three different spatial resolutions, of mutations from the selected set:

HAEMOLYTIC URAEMIC SYNDROME disease mutations from HGMD.

	Enrichment of Mutations at this Interface (LOR) Domain	Residue	Spatial Clustering of Mutations
P20248	1.1 (p=0.1340)	4.7 (p=0.0000)	Run Calculation
P46527	1.5 (p=0.0026)	1.1 (p=0.0927)	
Interaction	1.4 (p=0.0006)	2.3 (p=0.0000)	

Interface Enrichment of Disease Mutations for this Interaction

Interface: Both; Scale: Residue; Database: HGMD

HAEMOLYTIC URAELIC SYNDROME	LOR: 2.3 (p= 0.0000)
HAEMOLYTIC URAELIC SYNDROME, ...	LOR: 1.8 (p= 0.0000)
FACTOR H DEFICIENCY	LOR: 1.0 (p= 0.1165)
MEMBRANOPROLIFERATIVE GLOMER...	N/A
THROMBOTIC MICROANGIOPATHY FO...	N/A

Figure 3.

Workflow for calculating mutation and variant enrichment using Interactome INSIDER (http://interactomeinsider.yulab.org). Users may submit mutations or select sets of known disease and cancer mutations to assess their enrichment in interface domains and residues, or may compute 3D atomic clusters of mutations in proteins and across interfaces.

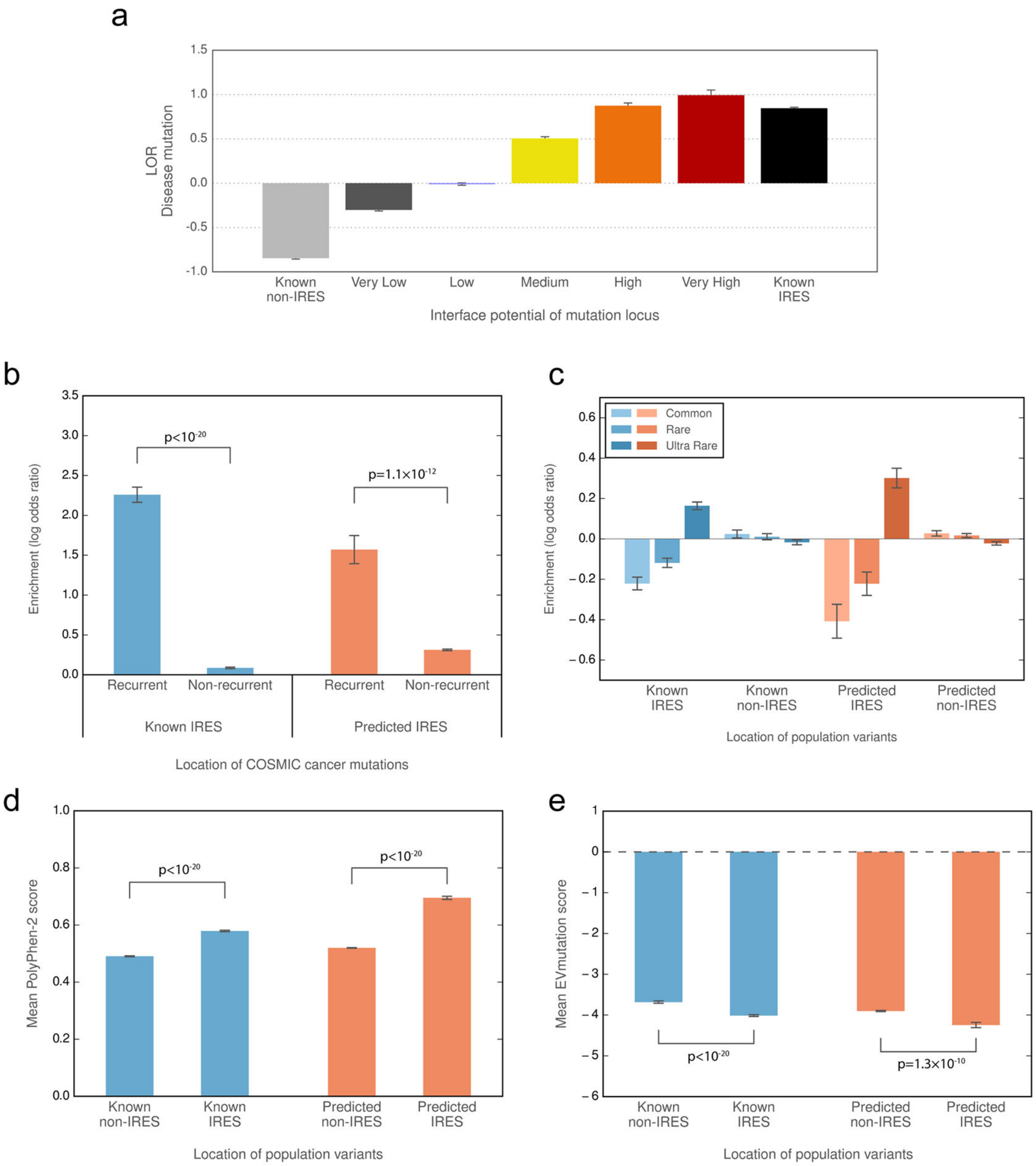


Figure 4.

Functional properties of predicted interfaces. (a) Enrichment of disease mutations in predicted and known interfaces. In a–c, enrichment (log odds ratio) is the odds of mutations and variants to appear in and outside of predicted and known interfaces compared to the odds of any residues to exist in these categories. (b) Enrichment of recurrent cancer mutations in predicted and known interfaces. (c) Enrichment of rare and common population variants in predicted and known interfaces. (d, e) Predicted deleteriousness of population variants in known and predicted interfaces using PolyPhen-2 (d) or EVmutation (e). (In b, significance determined by two-sided Z-test. In d-e, significance determined by a two-sided U-test. IRES=interface residues)

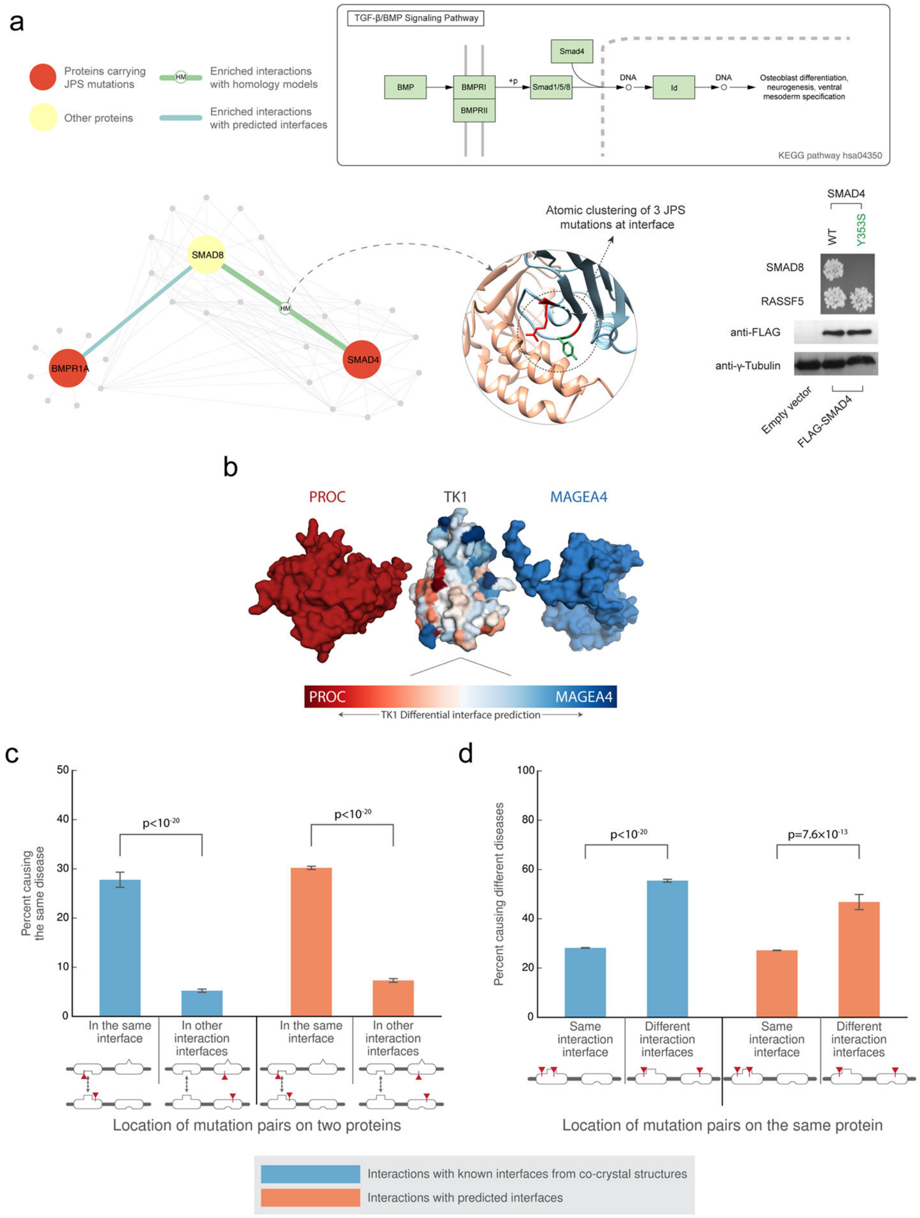


Figure 5.

Interaction partner-specific interface prediction. (a) The top schematic depicts the TGF-β/BMP signaling pathway. The bottom schematic illustrates that atomic clustering reveals a mutation hotspot for juvenile polyposis syndrome at the interface of SMAD8 and SMAD4. At right, yeast two-hybrid experiments test the interactions of one of the SMAD4 mutations (Y353S) with SMAD8 and RASSF5. The mutation is not predicted by ECLAIR to be at the SMAD4-RASS5 interface. (b) Superimposed docking results of two different interaction partners with TK1. The differentially predicted interfaces of TK1 with each of its partners correspond with the orientation of the docked poses. (c) The plot shows the fraction of disease mutation pairs in known (blue) or predicted (orange) interfaces that cause the same disease when mutations are within a given interaction interface compared to when mutations are not within an interaction interface. (d) The plot shows the fraction of disease mutation pairs in known (blue) or predicted (orange) interfaces that cause different diseases when mutations are in the same interaction interface compared to in different interaction interfaces (interaction with other proteins is not shown). (Significance determined by two-sided Z-test)

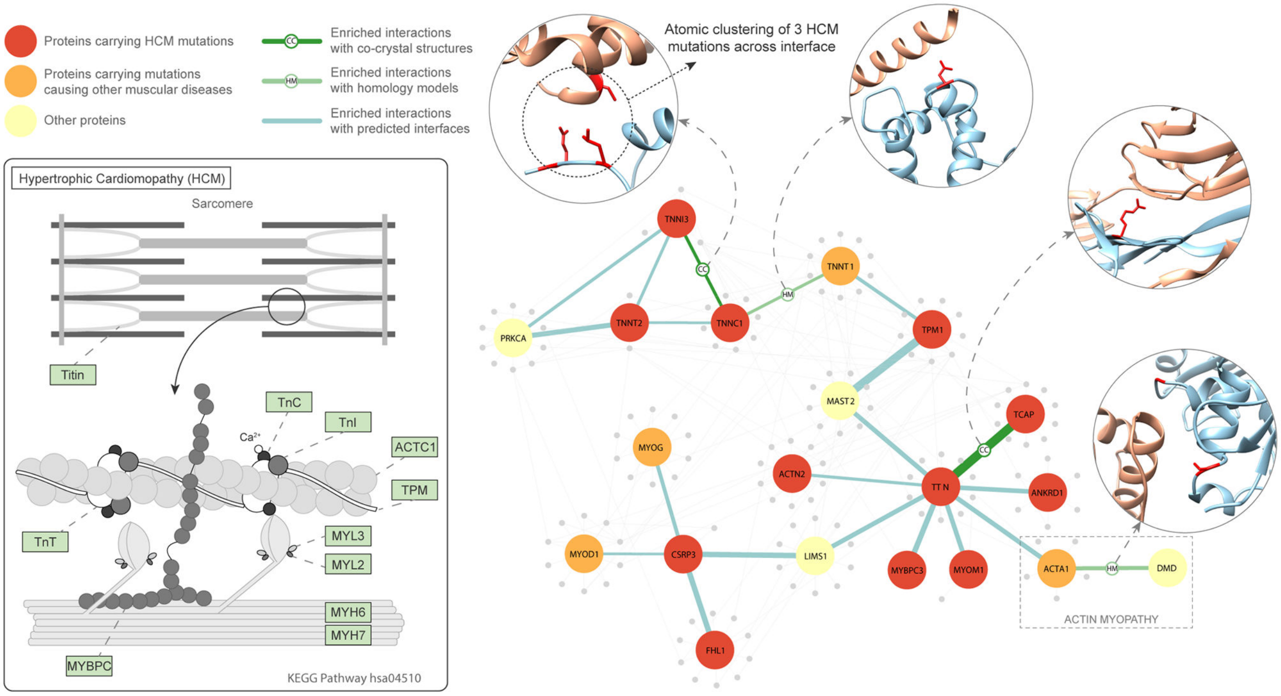

Proteins carrying HCM mutations
Proteins carrying mutations causing other muscular diseases
Other proteins
Enriched interactions with co-crystal structures
Enriched interactions with homology models
Enriched interactions with predicted interfaces
Atomic clustering of 3 HCM mutations across interface
Hypertrophic Cardiomyopathy (HCM)
Sarcomere
Titin
TnC
TnI
Ca2+
ACTC1
TPM
TnT
MYL3
MYL2
MYH6
MYH7
MYBPC
KEGG Pathway hsa04510
TNNI3
TNNT1
PRKCA
TNNT2
TNNC1
TPM1
MAST2
TCAP
MYOG
ACTN2
TTN
ANKRD1
MYOD1
CSRP3
LIMS1
MYBPC3
MYOM1
ACTA1
DMD
FHL1
CC
HM
ACTIN MYOPATHY

Figure 6.

The hypertrophic cardiomyopathy (HCM) pathway. The schematic on the left shows the interaction of proteins in the HGM KEGG pathway (hsa04510). On the right is shown a network of KEGG pathway proteins and their structurally-resolved interactions from Interactome INSIDER. Proteins that harbor HCM mutations are colored in red. Interfaces are noted for their enrichment of HCM mutations.