This is a preprint of an article under review for publication in Human Molecular Genetics

Proteins linked to autosomal dominant and autosomal recessive disorders harbor characteristic rare missense mutation distribution patterns

Tychele N. Turner[1,2], Christopher Douville[3], Dewey Kim[3], Peter D. Stenson[4], David N. Cooper[4], Aravinda Chakravarti[2], Rachel Karchin[3,5,*]

[1]Predoctoral Training Program in Human Genetics and Molecular Biology, McKusick-Nathans Institute of Genetic Medicine,
Johns Hopkins University School of Medicine, Baltimore, MD 21205,USA.

[2]Center for Complex Disease Genomics,
Johns Hopkins University School of Medicine, Baltimore, MD 21205,USA.

[3]Departments of Biomedical Engineering, Institute for Computational Medicine , Johns Hopkins University, Baltimore, MD 21210, USA and

[4]Institute of Medical Genetics, School of Medicine, Cardiff University, Heath Park, Cardiff CF14 4XN, UK

[5]Department of Oncology, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA.

* Send all correspondence to:

Rachel Karchin, Ph.D.
Departments of Biomedical Engineering and Oncology
Institute of Computational Medicine
Johns Hopkins University
Hackerman Hall 217A
Baltimore, MD 21210
T: 1-410-516-5578
F: 1-410-516-5294
E: karchin@jhu.edu

ABSTRACT

The role of rare missense variants in disease causation remains difficult to interpret. We explore whether the clustering pattern of rare missense variants (MAF<0.01) in a protein is associated with mode of inheritance. Mutations in genes associated with autosomal dominant (AD) conditions are known to result in either loss or gain of function, whereas mutations in genes associated with autosomal recessive (AR) conditions invariably result in loss of function. Loss-of-function mutations tend to be distributed uniformly along protein sequence, while gain-of-function mutations tend to localize to key regions. It has not previously been ascertained whether these patterns hold in general for rare missense mutations. We consider the extent to which rare missense variants are located within annotated protein domains and whether they form clusters, using a new unbiased method called CLUstering by Mutation Position (CLUMP). These approaches quantified a significant difference in clustering between AD and AR diseases. Proteins linked to AD diseases exhibited more clustering of rare missense mutations than those linked to AR diseases (Wilcoxon P=5.7x10^{-4}, permutation P=8.4x10^{-4}). Rare missense mutation in proteins linked to either AD or AR diseases were more clustered than controls (1000G) (Wilcoxon P=2.8x10^{-15} for AD and P=4.5x10^{-4} for AR, permutation P=3.1x10^{-12} for AD and P=0.03 for AR). Differences in clustering patterns persisted even after removal of the most prominent genes. Testing for such non-random patterns may reveal novel aspects of disease etiology in large sample studies.

INTRODUCTION

Hermann Muller was the first geneticist to posit the existence of different classes of functional mutations effective at the protein level, mutations that he termed nullomorphs (complete loss of function), hypomorphs (reduced function), hypermorphs (increased function), antimorphs (antagonistic to wild-type) and neomorphs (new function) (1, 2). These classes of mutation can cause human disease, as well as phenotypic variability in general. Nullomorphs and hypomorphs are generally referred to today as loss-of-function mutations, and there has been speculation that they are not preferentially located at specific amino acid residue positions (2-4). This is because loss-of-function is often caused by destabilization of the hydrophobic protein core (5), or by frameshifts and premature stop codons that lead to the nonsense mediated decay (NMD) of truncated transcripts (6). On the other hand, hypermorphic, antimorphic and neomorphic mutations are generally referred to as gain-of-function mutations and are more likely to occur at specific amino acid residue positions, such as at sites of post-translational modification, ligand binding, or protein-protein interaction (5). To our knowledge, we present the first study to systematically assess and quantify the extent to which these clustering patterns are also applicable to rare missense mutations causing human inherited disease.

Single-gene diseases in which the causal mutations lie in genes residing on the autosomes are generally recognized to display either dominant (1 copy required) or recessive (2 copies) inheritance. These diseases can be caused by mutations in any of the classes mentioned above. There is a unique set of autosomal dominant diseases that are recognized to exhibit mutations in a highly restricted set of amino acid residue positions with very specific effects on protein function. By contrast, with autosomal recessive diseases, mutations are often loss-of-function and result in no or little usable protein product. Examples of specific protein functional effects include the autosomal dominant diseases Cherubism (*SH3BP2* mutations) (7) and Achondroplasia (*FGFR3* mutations) (8). In Cherubism, mutations occur at a binding site required for proper ubiquitylation and subsequent proteolytic degradation of SH3BP2(9, 10). In Achondroplasia, a mutation at residue 380 causes FGFR3 to become constitutively activated (11).

Based on the realization that mutations are often loss-of-function in recessive disease but can be either loss-of-function or gain-of-function in dominant diseases, we hypothesized that: 1) rare missense mutations within autosomal dominant (AD) disease genes might be more clustered than those in autosomal recessive (AR) disease genes; and 2) rare variants in controls might be less clustered than either. In this work, we define clustering, for a given set of mutations, as an event when mutations are closer to each other in primary protein sequence than would be expected by chance. We reasoned that if these mutation patterns generally held true, non-random clustering of rare missense mutations might provide key insights into the molecular mechanisms underlying inherited diseases. The search for new Mendelian disease genes based on whole exome sequencing is often focused on loss-of-function variants and deleterious missense variants (12). By examining non-random clustering, it becomes possible to detect regions that are critical to protein function, regardless of whether the clustered mutations are deleterious or result in gain of function.

To test the first hypothesis, we used data from The Human Gene Mutation Database (HGMD) (13), which comprises a collection of inherited mutations causing human genetic disease. To our knowledge, these data have not been previously assessed for a relationship between patterns of rare missense mutation clustering and mode of disease inheritance. To test the second hypothesis, we compared the rare missense mutations in these AD and AR genes to rare missense variants in these genes found in individuals from the 1000 Genomes Project.

First, we applied a biased approach that considered the fraction of missense mutations (or variants) in a given protein that occurred within annotated protein domains from the Human Protein Reference Database (HPRD) (14) (*domain occupancy score*). However, the assumption that rare missense mutations of large effect will only occur in protein domains, regions of regular secondary structure whose function is known and that occur paralogously in multiple proteins, is potentially problematic. Thus, we developed a new unbiased clustering method to score clustering of missense mutations in protein sequence. The method makes no *a priori* assumptions about the importance of these positions or the number of clusters.

We performed statistical testing to assess whether rare missense mutations in AD genes and AR genes exhibit different clustering patterns than in controls and from each other. AD genes were found to exhibit significantly higher protein domain occupancy than AR genes and controls, and both AD and AR genes had significantly higher occupancy than controls. When we removed the domain bias from our analysis by applying an unsupervised clustering algorithm we developed (CLUMP), we found that collectively AD genes exhibited significantly lower CLUMP scores (associated with greater clustering) than AR genes and that AD genes and AR genes had significantly lower CLUMP scores than controls. These trends persisted even after 18 outlier genes with the highest statistical significance were removed from the analysis, supporting the generality of the clustering patterns.

RESULTS

Generation of high quality mutations dataset and AD/AR annotations

By searching the Human Gene Mutation Database (HGMD) and using a customized pipeline (Figure 1) we generated a rare missense mutation dataset for AD genes (6,337 mutations underlying 162 diseases involving 181 genes and AR genes (6,493 mutations underlying 195 diseases involving 159 genes). A rare missense mutation was defined by a minor allele frequency < 0.01 in European controls from the 1000 Genomes Project.

Known disease-causing mutations are more likely to fall in domains

The general trends observed in our domain occupancy analysis are evident in (Figure 2A). The empirical cumulative distribution functions (CDFs) of domain occupancies for AD disease, AR disease, and controls (1000GP) show that the three sets are distinct and that the trend for AR disease lies midway between AD disease and controls. These trends can be further quantified by means of a non-parametric Wilcoxon test. Rare missense mutations associated with AD diseases are significantly more likely to occur within domains than are rare missense variants seen in the 1000 Genomes ($p= 2.8x10^{-15}$, Wilcoxon test, AD median = 55%, AD mean = 55%, 1000G median = 23%, 1000G mean = 31%). Rare missense mutations associated with AR diseases also exhibit this pattern ($p= 4.5x10^{-4}$, Wilcoxon test, AR median = 40%, AR mean = 41%) although significantly less so than those associated with AD diseases ($p= 5.7x10^{-4}$, Wilcoxon test). In addition to these tests of mutations in individual proteins, a global analysis of all mutations shows that rare missense mutations more often reside in domains in AD diseases (total AD mutations in domains = 2,728, total AD mutations =6,337, percent AD mutations in domains = 43.0%) than in AR diseases (total AR mutations in domains = 1,771, total AR mutations=6,493, percent AR mutations in domains = 27.3%) (Fisher one-sided p=9.2x10^{-79}). Generally, as previously documented (15-17) disease mutations (AD union AR) more often reside in domains than in controls (total control mutations in domains = 24,663, total control mutations=113,547, percent control mutations in domains = 21.7%) (Fisher one-sided p=6.7x10^{-233}).

Disease vs. control comparison of domain occupancy reveals proteins with significant differential clustering

Next, we considered whether domain occupancy could be applied to analysis of individual proteins to differentiate clustering patterns of rare missense disease mutations and control variants. We applied Fisher's Exact test to each protein in the AD and AR sets and compared mutation clustering patterns in disease vs. controls (1000G). We identified four genes with a significant number of domain mutations in the autosomal dominant dataset and two genes in the autosomal recessive dataset, and these genes appear as outliers in a quantile-quantile (QQ) plot of raw P-values (Figure 2B). AD genes were *NOTCH3* in cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy (CADASIL, p= 2.77x10^{-3}, Benjamini-Hochberg (BH) correction), *KRT14* in epidermolysis bullosa simplex (p= 4.24x10^{-3}, BH), *TP63* in ankyloblepharon-ectodermal defects-cleft lip/palate (AEC syndrome, p= 6.29x10^{-3}, BH), and *RUNX2* in cleidocranial dysplasia (p=6.57x10^{-3}). AR genes were *EYS* in retinitis pigmentosa (p=3.9x10^{-3}, BH) and *CFTR* in cystic fibrosis (p=0.03, BH) (Figure 2B). The general trends seen in the Wilcoxon test persisted even after these outliers were removed (AD vs 1000G P=9.4x10^{-14}, AR vs 1000G P=1.0x10^{-3}, AD vs AR P=1.0x10^{-3})

CLUMP analysis reveals increased clustering of autosomal dominant disease mutations

Whereas rare missense variants that occur in domains are more likely to have more influence on protein activity than those occurring outside of domains, many proteins do not have complete domain annotations (18). We further considered whether the mutation clustering trends defined by domain occupancy would persist if clustering was defined by an unbiased approach. To this end, we generated CLUMP scores for all proteins in the AD, AR and 1000 Genomes data. The empirical CDFs of CLUMP scores for AD disease, AR disease, and controls (1000G) show a similar trend to the domain occupancy scores, although the three sets are not as well separated across the full range of CLUMP scores (Figure 2C). However, the differences between the three sets remained statistically significant. Proteins with AD mutations exhibited lower scores (more clustering) than 1000 Genomes (P=3.1x10^{-12}) and AR (P=8.4x10^{-4}, Wilcoxon) proteins and AR proteins are themselves more localized than 1000 Genomes (P=0.03, Wilcoxon).

Disease vs. control comparison of CLUMP scores reveals proteins with significant differential mutation clustering

To assess the statistical significance of CLUMP scores, we applied permutation testing to each protein in the AD and AR sets and compared CLUMP scores in disease vs. controls (1000G). This analysis identified 9 genes with significantly lower CLUMP scores (increased clustering) in the autosomal dominant dataset and 5 genes in the autosomal recessive dataset. Two of the AD genes were also identified in the domain occupancy analysis (*TP63* and *RUNX2*). All significant genes appear as outliers in a quantile-quantile (QQ) plot of raw P-values (Figure 2D). AD genes were *RUNX2* in cleidocranial Dysplasia, *SH3BP2* in cherubism, *TP63* in ectrodactyly, ectodermal dysplasia, clefting (EEC) syndrome, *SCN9A* in primary erythermalgia, *NOD2* in Blau syndrome, *CHD7* in CHARGE syndrome, *FBN1* in aortic aneurysm, *APOB* in hypercholesterolaemia, and *GJB2* in keratitis-ichthyosis-deafness syndrome. AR genes were *DYSF* in limb girdle muscular dystrophy, *USH2A* in Usher Syndrome, *CRB1* in Leber congenital amaurosis, *SMARCAL1* in Schimke immuno-osseous dysplasia, and *PAH* in phenylketonuria (Figure 2D). For CLUMP scores, the general trends seen in the Wilcoxon test also persisted after outliers were removed (AD vs. 1000G P = 2.5x10-^{10}, AR vs. 1000G P= 0.06, AD vs. AR P = 2.3x10^{-3}).

For some of these AD genes, evidence of specific protein function affected by a mutation cluster has been previously recognized. In cleidocranial Dysplasia, mutations in the transcription factor *RUNX2* cluster in the Runt domain, interfering with DNA binding (19); in EEC syndrome, mutations in the transcription factor *TP63* cluster in the DNA binding domain , disrupting DNA binding (20); and in Blau syndrome, mutations in *NOD2* cluster at its ATP-binding site and within its helical domain, dysregulating hydrolysis and autoinhibition, respectively (21).

Proteins exhibiting increased clustering in Mendelian diseases

Of the genes whose protein products were identified to have significantly increased clustering when compared to controls, there were some that were already known to either localize in domains or cluster in a specific region of the protein. This included *RUNX2* in Clediocranial dysplasia (MIM 119600), the *TP63* gene in the AEC and EEC syndromes (MIM 603273), *SH3BP2* in Cherubism (MIM 118400), and *KRT14* in Epidermolysis bullosa simplex (MIM 148066). Our results also support the presence of a clustering pattern in the first 60 amino acid residues of the Keratitis-ichthyosis-deafness syndrome *GJB2*, which was previously observed in a small study of 10 patients (22).

Autosomal dominant mutations are bioinformatically predicted to be more pathogenic than autosomal recessive.

We have developed and published a bioinformatic variant pathogenicty classifier called the Variant Effect Scoring Tool (VEST), which outperformed SIFT or PolyPhen2 on a carefully curated benchmark set (five-fold gene holdout cross-validation cite) by a small margin (23). VEST scores range from 0 to 1 with the most having a score of 1. When we ran VEST on AD and AR variants we found that AD variants were overall more pathogenic than AR variants (Wilcoxon one-sided $p=4.2x10^{-10}$). In addition, we found the clustered/domain variants to be more pathogenic than non-clustered/non domain variants (Wilcoxon one-sided $p=3.2x10^{-3}$).

DISCUSSION

A very large number of rare missense variants are now being discovered by high throughput sequencing in an assortment of human disease studies. Identifying those that are pathogenic or which contribute to disease remains very challenging. We have previously shown that visualizing the distribution of missense variants in a given protein sequence can be informative in relation to identifying potentially causal variants (24). However, such visualization does not provide quantitative assessment of clustering patterns and it cannot be applied in a high-throughput setting. In this work, we present two methods for the rapid determination of mutation clustering patterns and their statistical significance. The first method is a domain occupancy score, which considers the fraction of variants in a protein that occur within annotated domains. This score is necessarily biased, because it depends on existing knowledge of those protein regions considered to comprise functional domains, and it may miss functionally important regions that occur outside of domains. The second method is the CLUMP score, which performs unsupervised clustering of amino acid residue positions where variants occur, without any prior knowledge of their functional importance. Interestingly, we observed remarkably similar results with both methods: proteins linked to AD diseases harbor significantly more clustering of disease mutations than those linked to AR diseases, and both AD and AR disease proteins exhibit more clustering of these mutations than controls from 1000G. Moreover, these trends are not driven by a few outliers, since they persist even when the 18 genes with the most significant P-values in our Fisher's Exact test and permutation test were removed.

It has been shown in some cases, that loss-of-function mutations (nullomorphs and hypomorphs) exhibit less clustering in protein sequence than hypermorphs and neomorphs (3, 4), but to our knowledge this is the first study to systematically assess these patterns with respect to rare missense mutations causing human inherited disease. The search for new Mendelian genes through whole exome or genome sequencing of patients has generally been focused on loss-of-function mutations (25), which have the advantage of being more readily interpretable. Bioinformatics scoring of missense mutation deleteriousness is also widespread in analysis pipelines, and features such as inter-species evolutionary conservation at a given mutation position implicitly identify amino acid substitutions that are damaging to that protein (26, 27). Often, researchers are faced with multiple rare missense variants in a gene of interest, none of which have been assessed to be damaging by popular bioinformatics tools. Our results support the idea that many of these variants may be important to Mendelian disease, but could be mutations that cause a protein gain of function and are inherited in an autosomal dominant inheritance pattern.

We have confirmed that the clustering patterns of rare missense mutations are systematically associated with mode of inheritance, and this pattern was robust with respect to whether clustering was defined by occurrence in protein domains of known functional importance or by an unbiased clustering approach. Our results are consistent with the notion that autosomal dominant disease genes harbor a mixture of deleterious and gain-of-function rare missense mutations, whereas autosomal recessive disease genes harbor only deleterious rare missense mutations.

Futher, these results suggest that sequencing studies of specific disease genes could benefit by testing for non-random clustering of rare missense variants. These clusters may provide insights into the molecular basis of inherited diseases, and such testing will become more powerful as sample sizes increase.

MATERIALS AND METHODS

Generation of a high quality list of disease mutations and mode of inheritance.

A list of 61,537 missense mutations causing inherited disease (DM) and occurring on autosomes was downloaded from the Human Gene Mutation Database (HGMD) Professional version 2014.2 on June 10, 2014. In this study, we focused on autosomal diseases and not X-linked due to lack of information on sample sex in this dataset. For each mutation, we first parsed all abstracts in PubMed (http://www.ncbi.nlm.nih.gov/pubmed/) to identify the mode of inheritance associated with the gene in which the mutation occurred, using a custom script and BioPython libraries (28). For each entry, we generated a Boolean query of the architecture *geneName AND diseaseName AND autosomal* (example: *CFTR* AND cystic fibrosis AND autosomal). Abstracts that matched the query were then parsed for the keywords "autosomal dominant" and "autosomal recessive." We counted the number of abstracts containing "autosomal dominant", "autosomal recessive" or which did not contain either of these terms. An initial assignment of each entry to the autosomal dominant (AD) class, the autosomal recessive (AR) class, or as "not determined" (ND) was performed by a vote of abstracts matching these keywords, so that

$$e_i = \begin{cases} AD & if\,\#\{AD\} > \#\{AR\} \\ AR & if\,\#\{AR\} > \#\{AD\} \\ ND & if\,\#\{AD\} = \#\{AR\} \end{cases} \qquad \text{(Eq 1)}$$

where e_i is an entry consisting of a gene/disease pair, #{AD} is the number of abstracts that contained the keywords "autosomal dominant", and #{AR} is the number of abstracts that contained the keywords "autosomal recessive". Because our study focuses on Mendelian disease, we filtered out any entries with a cancer disease association (containing the keywords cancer, sarcoma, carcinoma, leukemia, lymphoma, blastoma, glioma, melanoma, myeloma, tumor, tumour, metastasis, adenoma, neoplasia, or cytoma). At this stage, 3539 abstracts remained. To obtain high confidence calls, we further required that an entry's classification (Eq 1) was supported by at least 12 or more abstracts and that the classification was supported by a sizeable majority (75%) of the abstracts. These criteria filtered out 80% of abstracts identified by our initial queries, yielding a high-quality set of 706 abstracts that was tractable for manual inspection. Next, every entry was manually checked for correctness of our class assignment. For each entry, we first checked for confirmation in GeneReviews (GeneTests 1999-2014), followed by OMIM (http://omim.org/), and the primary literature. Manually confirmed entries were retained.

Control dataset

The 1000 Genomes Project dataset was obtained from ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/ on July 18, 2014. We selected only unrelated individuals of European ancestry from the CEU, FIN, GBR, IBS, and TSI populations.

Statistical tests for clustering of mutations and variants

To ascertain mutation clustering patterns in a gene product we adopted two approaches; the first was designed to look at the fraction of mutations occurring in annotated protein domains from the HPRD (*domain occupancy score*) and the second was the unbiased *CLUMP score*.

For a protein *p,* its *domain occupancy count* is:

$$C_p = \sum_{i=1}^{n} w_i X_i Z_i \qquad \text{(Eq 1)}$$

where X_i is a mutated amino acid residue position, w_i is the count of unique amino acid substitutions at that position in the data of interest, Z_i is binary random variable that is set to 1 when X_i is in an annotated protein domain, and 0 otherwise, and the sum is over the *n* mutated amino acid residue positions in the protein. Likewise,

$$C_p^{Controls} = \sum_{i=1}^{n^{Controls}} w_i^{Controls} X_i^{Controls} Z_i$$

$$C_p^{Disease} = \sum_{i=1}^{n^{Disease}} w_i^{Disease} X_i^{Disease} Z_i \qquad \text{(Eqs 2,3)}$$

and all variables have the same meaning as in (Eq. 1) but are assigned values based only on either variants in the control set or mutations in the disease set.

For a protein *p,* its domain occupancy score (the fraction of mutations occurring in domains) is:

$$D_p = \frac{C_p}{n} \qquad \text{(Eq 4)}$$

and likewise

$$D_p^{Controls} = \frac{C_p^{Controls}}{n^{Controls}}$$

$$D_p^{Disease} = \frac{C_p^{Disease}}{n^{Disease}} \qquad \text{(Eqs 5,6)}$$

We compute $D_p^{Controls}$ for all proteins in the control set and $D_p^{Disease}$ for all proteins in the disease set, and we apply a one-sided Wilcoxon test to ascertain whether the scores of proteins in the disease set are significantly higher than those in the control set. Next, to assess whether domain occupancy is significantly higher in the disease set than in the control set, for each protein we compute a one-tailed Fisher's Exact test, comparing counts of $C_p^{Disease}$, $(n^{Disease} - C_p^{Disease})$, $C_p^{Controls}$, and ($n^{Controls} - C_p^{Controls}$). Multiple testing correction was performed with the Benjamini-Hochberg algorithm and corrected P-values < 0.05 were considered significant.

The CLUMP score applies the partitioning around medoids (PAM) clustering algorithm (29) to a list of (integer-indexed) amino acid residue positions. We use the pamk implementation in the fpc package in R. The number of clusters k is not specified in advance but is estimated by varying k over multiple PAM runs and selecting the $k*$ that yields the maximum average silhouette width. Thus, both the number of clusters and a "medioid" or representative member of each cluster are estimated by the algorithm. Next, for each cluster i , we compute the distance between each member of the cluster and its mediod and take a log sum of these distances over all clusters. The final CLUMP score S_p for a protein p is:

$$S_p = \sum_{i=1}^{k*} \sum_{j=1}^{n_i} \frac{\ln\left(|X_{ij} - m_i| + 1\right)}{n_i} \qquad \text{(Eq 7)}$$

where X_{ij} is the position of mutation j in cluster i, m_i is the position of the mediod of cluster i, n_i is the number of mutations in cluster i, and $k*$ is the total number of clusters in the gene. The maximum clustering possible is when all observed mutations in all clusters occur at the same position as the cluster mediod, yielding a score of 0. In general, a protein with highly localized mutations will have a low score, while a protein with mutations spread across its protein sequence will have a high score.

To assess the statistical significance of S_p (Eq 7), we compute for each gene's protein product p, $S_p^{Controls}$ and $S_p^{Disease}$ as

$$S_p^{Controls} = \sum_{i=1}^{k*^{Controls}} \sum_{j=1}^{n_i^{Controls}} \frac{\ln\left(|X_{ij}^{Controls} - m_i^{Controls}| + 1\right)}{n_i^{Controls}}$$

$$S_p^{Disease} = \sum_{i=1}^{k*^{Disease}} \sum_{j=1}^{n_i^{Disease}} \frac{\ln\left(|X_{ij}^{Disease} - m_i^{Disease}| + 1\right)}{n_i^{Disease}} \qquad \text{(Eqs 8,9)}$$

where all variables have the same meaning as in (Eq. 7) but are assigned values based only on either variants in the control set or mutations in the disease set, *i.e.*, $n_i^{Controls}$ is the total number of variants observed in the protein in the control set, $n_i^{Disease}$ is the total number of mutations observed in the protein in the disease set, *etc.*

We compute $S_p^{Controls}$ for all proteins in the control set and $S_p^{Disease}$ for all proteins in the disease set, and we apply a one-sided Wilcoxon test to determine if the scores of proteins in the control set are significantly higher than those in the disease set. Next, to assess whether $S_p^{Controls}$ is significantly higher than $S_p^{Disease}$ for individual proteins, we use the test statistic $\Delta S_p = S_p^{Controls} - S_p^{Disease}$.

We simulate a null distribution of values $\Delta S_p^{\varnothing}$ that would be expected when the difference between $S_p^{Controls}$ and $S_p^{Disease}$ is due to random chance, by repeatedly sampling with replacement $n_i^{Controls}$ positions in protein *p* (assuming that each position is equally likely under the null hypothesis) and computing $\Delta S_p^{\varnothing\,1}, \Delta S_p^{\varnothing\,2}, \cdots, \Delta S_p^{\varnothing\,N}$, where in this work *N*=10,000. The estimated P-value for ΔS_p is then the fraction of times a value equal to or greater than ΔS_p is seen under the null. Finally we use the Benjamini-Hochberg method (30) to correct for multiple testing.

ACKNOWLEDGMENTS

This work was funded by a grant from the NSF (DBI-0845275) to RK. This paper began as a project in the Foundations of Computational Biology and Bioinformatics II course (Spring 2011) at the Johns Hopkins University.

CONFLICT OF INTEREST STATEMENT

We have no conflicts of interest.

REFERENCES

1 Muller, H.J. (1932) Further studies on the nature and causes of gene mutations. *Proc 6th Intl Congr Genet* in press., I:213-255.

2 Hawley, R.S., Walker, M.Y. . (2003) *Advanced Genetic Analysis: Finding Meaning in a Genome*. Blackwell Publishing, Malden, MA.

3 Schindelhauer, D., Weiss, M., Hellebrand, H., Golla, A., Hergersberg, M., Seger, R., Belohradsky, B.H. and Meindl, A. (1996) Wiskott-Aldrich syndrome: no strict genotype-phenotype correlations but clustering of missense mutations in the amino-terminal part of the WASP gene product. *Hum Genet*, 98, 68-76.

4 Bergmann, C., Senderek, J., Sedlacek, B., Pegiazoglou, I., Puglia, P., Eggermann, T., Rudnik-Schoneborn, S., Furu, L., Onuchic, L.F., De Baca, M. *et al.* (2003) Spectrum of mutations in the gene for autosomal recessive polycystic kidney disease (ARPKD/PKHD1). *Journal of the American Society of Nephrology : JASN*, 14, 76-89.

5 Zhe Zhang, M.A.M., Lin Wang, and Emil Alexov. (2012) Analyzing Effects of Naturally Occurring Missense Mutations. *Computational and Mathematical Methods in Medicine*, in press.

6 Frischmeyer, P.A. and Dietz, H.C. (1999) Nonsense-mediated mRNA decay in health and disease. *Hum Mol Genet*, 8, 1893-1900.

7 Ueki, Y., Tiziani, V., Santanna, C., Fukai, N., Maulik, C., Garfinkle, J., Ninomiya, C., doAmaral, C., Peters, H., Habal, M. *et al.* (2001) Mutations in the gene encoding c-Abl-binding protein SH3BP2 cause cherubism. *Nat Genet*, 28, 125-126.

8 Bellus, G.A., Hefferon, T.W., Ortiz de Luna, R.I., Hecht, J.T., Horton, W.A., Machado, M., Kaitila, I., McIntosh, I. and Francomano, C.A. (1995) Achondroplasia is defined by recurrent G380R mutations of FGFR3. *Am J Hum Genet*, 56, 368-373.

9 Levaot, N., Voytyuk, O., Dimitriou, I., Sircoulomb, F., Chandrakumar, A., Deckert, M., Krzyzanowski, P.M., Scotter, A., Gu, S., Janmohamed, S. *et al.* (2011) Loss of Tankyrase-mediated destruction of 3BP2 is the underlying pathogenic mechanism of cherubism. *Cell*, 147, 1324-1339.

10 Guettler, S., LaRose, J., Petsalaki, E., Gish, G., Scotter, A., Pawson, T., Rottapel, R. and Sicheri, F. (2011) Structural basis and sequence rules for substrate recognition by Tankyrase explain the basis for cherubism disease. *Cell*, 147, 1340-1354.

11 Webster, M.K. and Donoghue, D.J. (1996) Constitutive activation of fibroblast growth factor receptor 3 by the transmembrane domain point mutation found in achondroplasia. *Embo j*, 15, 520-527.

12 O'Roak, B.J., Deriziotis, P., Lee, C., Vives, L., Schwartz, J.J., Girirajan, S., Karakoc, E., Mackenzie, A.P., Ng, S.B., Baker, C. *et al.* (2011) Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nat Genet*, 43, 585-589.

13 Stenson, P.D., Mort, M., Ball, E.V., Shaw, K., Phillips, A. and Cooper, D.N. (2014) The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum Genet*, 133, 1-9.

14 Prasad, T.S., Kandasamy, K. and Pandey, A. (2009) Human Protein Reference Database and Human Proteinpedia as discovery tools for systems biology. *Methods Mol Biol*, 577, 67-79.

15 Yue, P., Forrest, W.F., Kaminker, J.S., Lohr, S., Zhang, Z. and Cavet, G. (2010) Inferring the functional effects of mutation through clusters of mutations in homologous proteins. *Human mutation*, 31, 264-271.

16 Peterson, T.A., Park, D. and Kann, M.G. (2013) A protein domain-centric approach for the comparative analysis of human and yeast phenotypically relevant mutations. *BMC genomics*, 14 Suppl 3, S5.

17 Peterson, T.A., Park, D. and Kann, M.G. (2013) Domain landscapes of somatic mutations in cancer. *AMIA Joint Summits on Translational Science proceedings AMIA Summit on Translational Science*, 2013, 136.

18 Fong, J.H. and Marchler-Bauer, A. (2008) Protein subfamily assignment using the Conserved Domain Database. *BMC Res Notes*, 1, 114.

19 Yoshida, T., Kanegane, H., Osato, M., Yanagida, M., Miyawaki, T., Ito, Y. and Shigesada, K. (2003) Functional analysis of RUNX2 mutations in cleidocranial dysplasia: novel insights into genotype-phenotype correlations. *Blood Cells Mol Dis*, 30, 184-193.

20 Brunner, H.G., Hamel, B.C. and Van Bokhoven, H. (2002) The p63 gene in EEC and other syndromes. *J Med Genet*, 39, 377-381.

21 Parkhouse, R., Boyle, J.P. and Monie, T.P. (2014) Blau syndrome polymorphisms in NOD2 identify nucleotide hydrolysis and helical domain 1 as signalling regulators. *FEBS Lett*, 588, 3382-3389.

22 Richard, G., Rouan, F., Willoughby, C.E., Brown, N., Chung, P., Ryynanen, M., Jabs, E.W., Bale, S.J., DiGiovanna, J.J., Uitto, J. *et al.* (2002) Missense mutations in GJB2 encoding connexin-26 cause the ectodermal dysplasia keratitis-ichthyosis-deafness syndrome. *Am J Hum Genet*, 70, 1341-1348.

23 Carter, H., Douville, C., Stenson, P.D., Cooper, D.N. and Karchin, R. (2013) Identifying Mendelian disease genes with the variant effect scoring tool. *BMC genomics*, 14 Suppl 3, S3.

24 Turner, T. (2013) Plot protein: visualization of mutations. *J Clin Bioinforma*, 3, 14.

25 Sobreira, N.L., Cirulli, E.T., Avramopoulos, D., Wohler, E., Oswald, G.L., Stevens, E.L., Ge, D., Shianna, K.V., Smith, J.P., Maia, J.M. *et al.* (2010) Whole-genome sequencing of a single proband together with linkage analysis identifies a Mendelian disease gene. *PLoS Genet*, 6, e1000991.

26 Adzhubei, I., Jordan, D.M. and Sunyaev, S.R. (2013) Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet*, Chapter 7, Unit7.20.

27 Kumar, P., Henikoff, S. and Ng, P.C. (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc*, 4, 1073-1081.

28 Cock, P.J., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B. *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25, 1422-1423.

29 Caliński, T. and Harabasz, J. (1974) A dendrite method for cluster analysis. *Communications in Statistics*, 3, 1-27.

30 Benjamini, Y., and Hochberg, Y. . (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series*, B57, 289-300.

LEGENDS TO FIGURES

Figure 1**: Workflow of this study.** Included are details on the generation of high quality inheritance datasets for all missense variants in autosomal dominant (AD) and autosomal recessive (AR) diseases. Also depicted are our two main approaches to assess mutation clustering within proteins.

***Figure 2:* Statistical test of rare missense variant or mutation clustering within proteins.** (A) Empirical cumulative distribution function of proportion of mutations residing in a domain per protein. (B) Quantile-quantile (QQ) plot of raw P-values for Fisher Exact testing to examine enrichment of mutations within domains in disease versus in controls. (C) Empirical cumulative distribution function of CLUMP scores per protein. (D) QQ plot of raw P-values for permutation testing to examine lower CLUMP scores in disease versus controls. Genes listed are those that attained a level of significance after Benjamini Hochberg correction.

Figure 3: Mutations in the *SH3BP2* gene in cherubism show significant clustering. (A) Shown are all mutations in 1000 Genomes controls and in cherubism. (B) Zoom in of the region where the majority of mutations reside as well as the number of different amino acid changes at each position. The cherubism mutations are significantly more clustered than the control data ($p < 1x10^{-4}$).

TABLES

Table 1: **Proteins with significant enrichment of autosomal dominant and recessive rare, missense mutations in domains.** Shown are counts in annotated HPRD domains or not in domains of rare (minor allele frequency < 0.01 based on controls) missense variants. The control data are from the 1000 Genomes European ancestry data. (★=autosomal dominant, ✚=autosomal recessive).

Protein	**Gene**	**Total Control Mutations (% in domain)**	**Total Disease Mutations (% in domain)**	**Disease**	**p-value (BH corrected p-value)**
NP_000426.2	*NOTCH3*	20 (25%)	209 (99%)	CADASIL★	$5.78x10^{-5}$ $(2.77x10^{-3})$
NP_000517.2	*KRT14*	5 (0%)	24 (92%)	Epidermolysis bullosa simplex★	$1.77x10^{-4}$ $(4.24x10^{-3})$
NP_003713.3	*TP63*	5 (0%)	25 (88%)	AEC syndrome★	$3.93x10^{-4}$ $(6.29x10^{-3})$
NP_001019801.3	*RUNX2*	6 (17%)	52 (88%)	Cleidocranial dysplasia★	$5.48x10^{-4}$ $(6.57x10^{-3})$
NP_001136272.1	*EYS*	17 (18%)	20 (85%)	Retinitis pigmentosa[a]	$5.56x10^{-5}$ $(3.89x10^{-3})$
NP_000483.3	*CFTR*	31 (48%)	533 (77%)	Cystic fibrosis✚	$9.68x10^{-4}$ (0.03)

***Table 2:* Proteins with significantly lower CLUMP scores in autosomal dominant and recessive rare, missense mutations than in controls.** Shown are differential CLUMP scores between controls and disease variants of rare (minor allele frequency < 0.01 based on controls) missense variants. The control data are from the 1000 Genomes European ancestry data. (★ =autosomal dominant, ✚=autosomal recessive).

Protein	**Gene**	**Differential CLUMP Score**	**Dataset**	**p-value (BH corrected p-value)**
NP_001139328.1	*SH3BP2*	2.86	Cherubism★	<1x10^{-4} (<1x10^{-4})
NP_001019801.3	*RUNX2*	2.38	Cleidocranial dysplasia★	<1x10^{-4} (<1x10^{-4})
NP_003713.3	*TP63*	1.72	EEC syndrome★	<1x10^{-4} (<1x10^{-4})
NP_002968.1	*SCN9A*	3.5	Erythermalgia, primary★	3.00x10^{-4} (4.73x10^{-3})
NP_071445.1	*NOD2*	3.62	Blau syndrome★	4.00x10^{-4} (5.04x10^{-3})
NP_060250.2	*CHD7*	2.6	CHARGE syndrome★	1.70x10^{-3} (0.02)
NP_000129.3	*FBN1*	-0.4	Aortic aneurysm★	2.10x10^{-3} (0.02)
NP_000375.2	*APOB*	3.72	Hypercholesterolaemia★	2.60x10^{-3} (0.02)
NP_003995.2	*GJB2*	1.52	Keratitis-ichthyosis-deafness syndrome★	5.40x10^{-3} (0.04)
NP_996816.2	*USH2A*	3.81	Usher syndrome✚	<1x10^{-4} (<1x10^{-4})
NP_001124459.1	*DYSF*	3.05	Muscular dystrophy, limb girdle✚	<1x10^{-4} (<1x10^{-4})
NP_957705.1	*CRB1*	1.44	Leber congenital amaurosis✚	1.10x10^{-3} (0.03)
NP_001120679.1	*SMARCAL1*	2.3	Schimke immuno-osseous dysplasia✚	1.20x10^{-3} (0.03)
NP_000268.1	*PAH*	-0.32	Phenylketonuria✚	2.00x10^{-3} (0.03)

Figure 1.

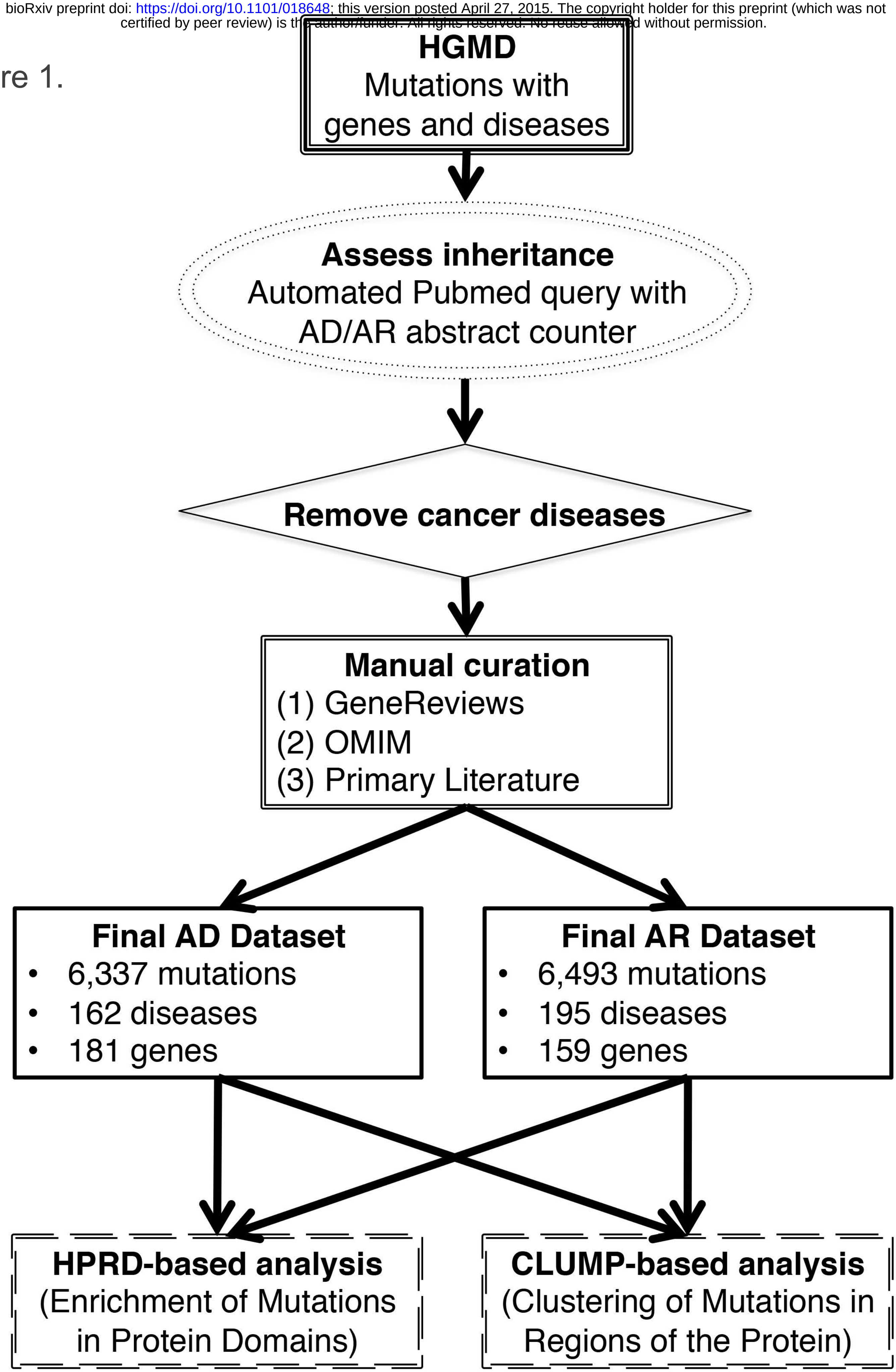

HGMD
Mutations with genes and diseases
Assess inheritance
Automated Pubmed query with AD/AR abstract counter
Remove cancer diseases
Manual curation
(1) GeneReviews
(2) OMIM
(3) Primary Literature
Final AD Dataset
• 6,337 mutations
• 162 diseases
• 181 genes
Final AR Dataset
• 6,493 mutations
• 195 diseases
• 159 genes
HPRD-based analysis
(Enrichment of Mutations in Protein Domains)
CLUMP-based analysis
(Clustering of Mutations in Regions of the Protein)

Figure 2.

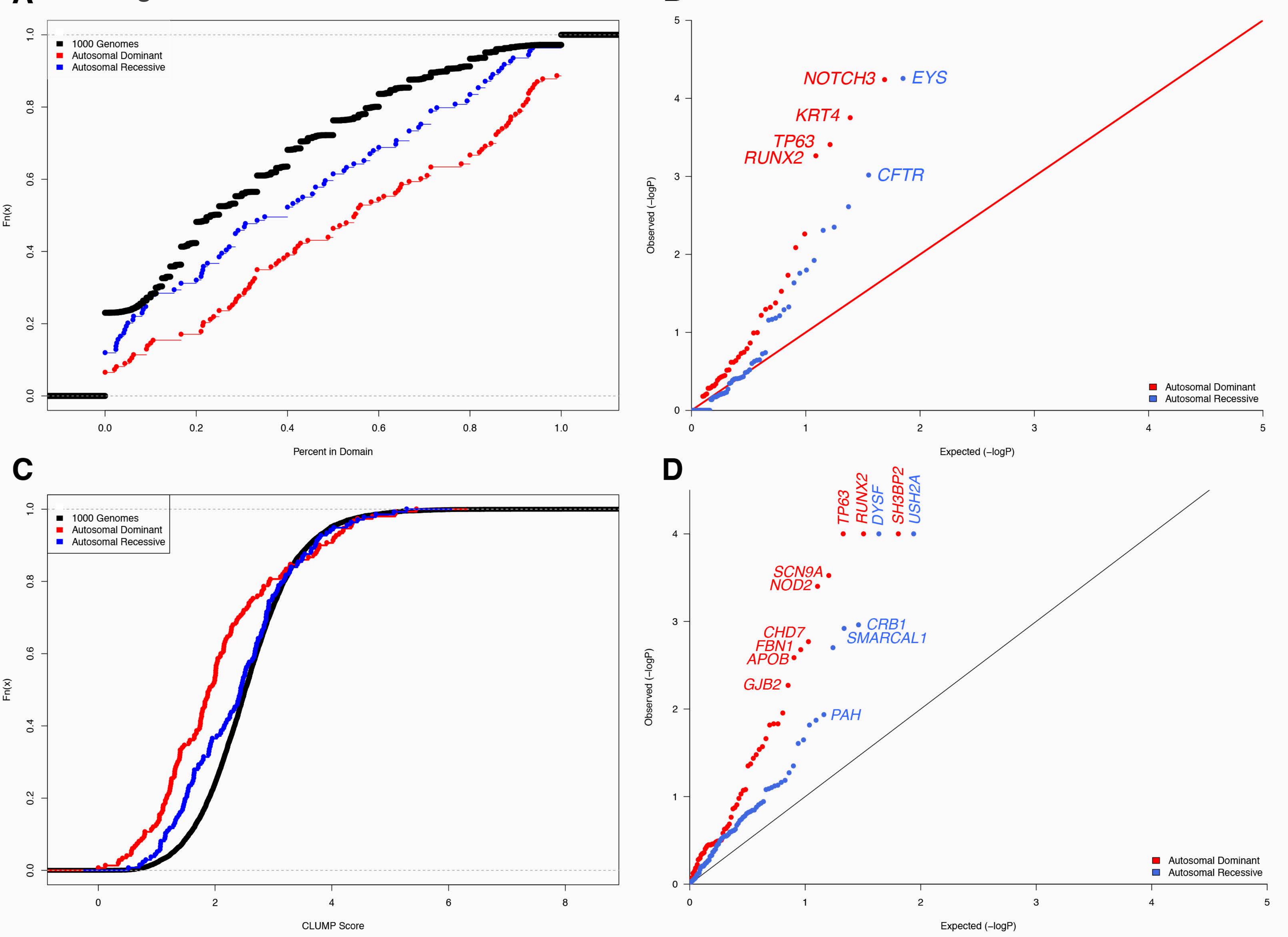
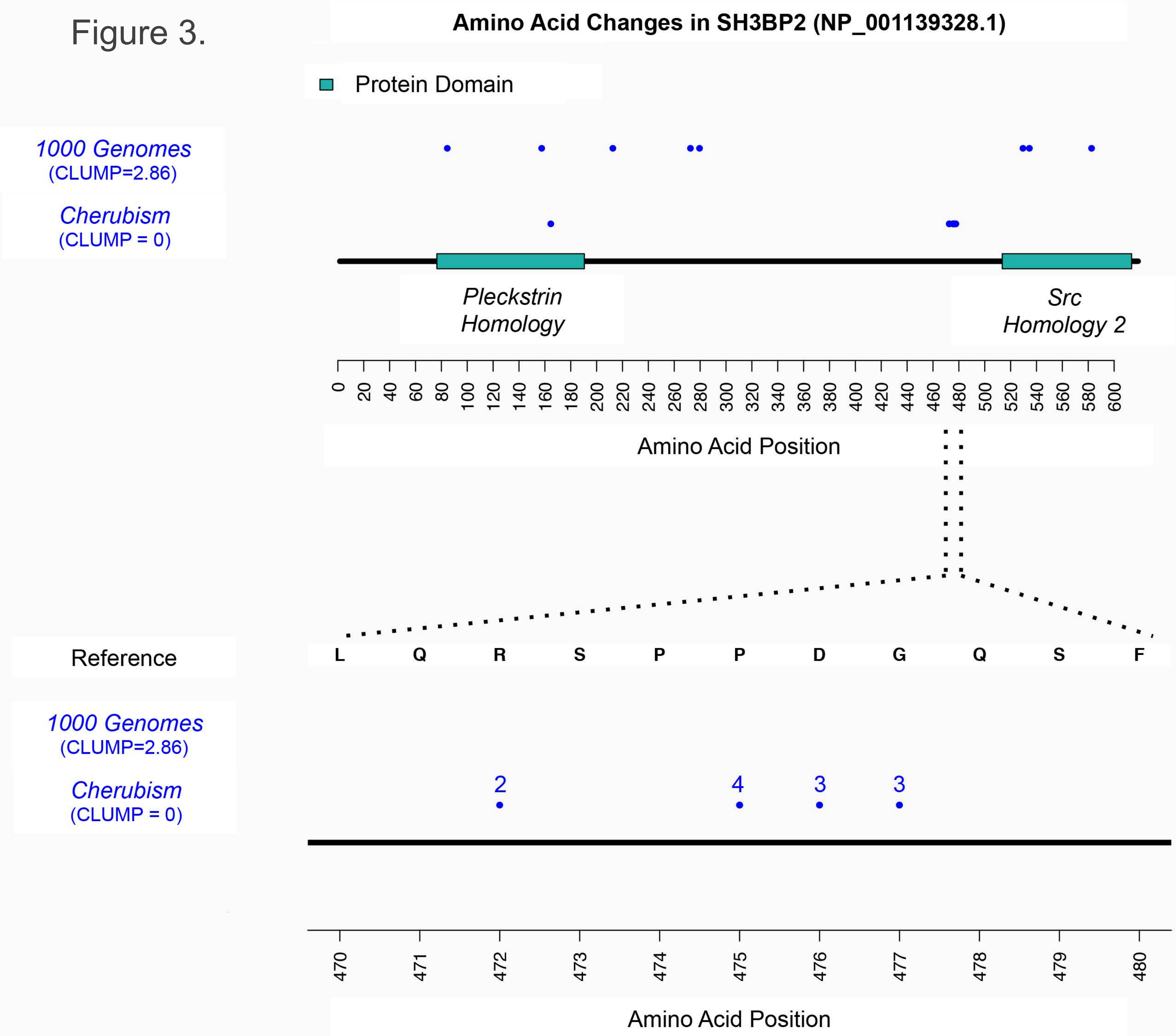

A
1000 Genomes
Autosomal Dominant
Autosomal Recessive
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
Percent in Domain
B
NOTCH3
EYS
KRT4
TP63
RUNX2
CFTR
Autosomal Dominant
Autosomal Recessive
Observed (−logP)
Expected (−logP)
0
1
2
3
4
5
C
1000 Genomes
Autosomal Dominant
Autosomal Recessive
Fn(x)
CLUMP Score
0
2
4
6
8
D
TP63
RUNX2
DYSF
SH3BP2
USH2A
SCN9A
NOD2
CHD7
FBN1
APOB
CRB1
SMARCAL1
GJB2
PAH
Autosomal Dominant
Autosomal Recessive
Observed (−logP)
Expected (−logP)

Figure 3.


Amino Acid Changes in SH3BP2 (NP_001139328.1)
Protein Domain
1000 Genomes
(CLUMP=2.86)
Cherubism
(CLUMP = 0)
Pleckstrin
Homology
Src
Homology 2
0
20
40
60
80
100
120
140
160
180
200
220
240
260
280
300
320
340
360
380
400
420
440
460
480
500
520
540
560
580
600
Amino Acid Position
Reference
L
Q
R
S
P
P
D
G
Q
S
F
1000 Genomes
(CLUMP=2.86)
Cherubism
(CLUMP = 0)
2
4
3
3
470
471
472
473
474
475
476
477
478
479
480
Amino Acid Position