frontiers in Molecular Biosciences

PERSPECTIVE
published: 12 August 2015
doi: 10.3389/fmolb.2015.00047

Anatomy of protein disorder, flexibility and disease-related mutations

Hui-Chun Lu[1], *Sun Sook Chung*[1,2], *Arianna Fornili*[1,3] *and Franca Fraternali*[1]*

[1] Randall Division of Cell and Molecular Biophysics, King's College London, London, UK, [2] Department of Haematological Medicine, King's College London, London, UK, [3] School of Biological and Chemical Sciences, Queen Mary University of London, London, UK

Integration of protein structural information with human genetic variation and pathogenic mutations is essential to understand molecular mechanisms associated with the effects of polymorphisms on protein interactions and cellular processes. We investigate occurrences of non-synonymous SNPs in ordered and disordered protein regions by systematic mapping of common variants and disease-related SNPs onto these regions. We show that common variants accumulate in disordered regions; conversely pathogenic variants are significantly depleted in disordered regions. These different occurrences of pathogenic and common SNPs can be attributed to a negative selection on random mutations in structurally highly constrained regions. New approaches in the study of quantitative effects of pathogenic-related mutations should effectively account for all the possible contexts and relative functional constraints in which the sequence variation occurs.

Keywords: non-synonymous SNPs, protein disorder, order-disorder propensity, disease-related mutations, protein flexibility

OPEN ACCESS

Edited by:
Kris Pauwels,
Vrije Universiteit Brussel, Belgium

Reviewed by:
Elena Papaleo,
University of Copenhagen, Denmark
Sreenivas Chavali,
MRC Laboratory of Molecular Biology, UK

***Correspondence:**
Franca Fraternali,
Randall Division of Cell and Molecular Biophysics, King's College London, New Hunt's House, Guy's Campus, London SE1 1UL, UK
franca.fraternali@kcl.ac.uk

Specialty section:
This article was submitted to Structural Biology, a section of the journal Frontiers in Molecular Biosciences

Received: *31 May 2015*
Accepted: *29 July 2015*
Published: *12 August 2015*

Citation:
Lu H-C, Chung SS, Fornili A and Fraternali F (2015) Anatomy of protein disorder, flexibility and disease-related mutations. Front. Mol. Biosci. 2:47. doi: 10.3389/fmolb.2015.00047

Introduction

Because of the intrinsic complexity of biological systems, reductionist approaches have traditionally been used that concentrate on carefully chosen sub-systems. The availability of complete genome sequences and large (but incomplete) collections of biomolecular structures at atomic resolution favors large-scale computational approaches to investigate multiple components and their interactions (Lu et al., 2013). The undisputed relationship between protein-coding elements and their protein products has dominated the field of genomics/proteomics research in the past and the relationship between structure and function has been widely investigated.

Large-scale studies have been performed on how disease-related mutations may disrupt protein functions and ultimately regulate the function of biological systems (Studer et al., 2013). Mutations are classified as "loss of function," "gain of function," or "neutral" according to their effect on protein function. These effects can be mediated by alterations of the protein stability induced by the mutation (Yue et al., 2005; Studer et al., 2013). The impact of SNPs on protein function and structural stability has been extensively studied at the level of the single protein (Yue et al., 2005; Schuster-Bockler and Bateman, 2008; Wang et al., 2012; Nishi et al., 2013; Studer et al., 2013; Yates and Sternberg, 2013; Scharner et al., 2014) and a number of predictors have been developed to evaluate the impact of SNPs on individual proteins (Thomas and Kejariwal, 2004; Capriotti et al., 2005; Bromberg and Rost, 2007; Adzhubei et al., 2010;

Reva et al., 2011; Al-Numair and Martin, 2013; Shihab et al., 2013; Pires et al., 2014; Yates et al., 2014). With the intention to expand the single protein structure-function paradigm, the interplay between Protein Protein Interactions (PPI) networks, structures, and disease mutations has been explored by several groups (see reviews Lu et al., 2013; Yates and Sternberg, 2013) and reference therein, Kelley et al., 2015; Mosca et al., 2015). Particularly the crucial role of interfaces in modulating the effects of pathogenic variation in binding and signaling (Stefl et al., 2013; Yates and Sternberg, 2013) has been generally accepted. In recent years additional findings have contributed to further expanding classical structure-function approaches: firstly, the widely recognized importance of non-coding elements (Necsulea and Kaessmann, 2014; Ling et al., 2015) (not discussed here) and the enrichment of SNPs in these (Consortium, 2012; Kircher et al., 2014); secondly, the role of unstructured regions, intrinsically disordered elements and flexibility in protein function versatility (Uversky, 2013; Dunker et al., 2015; Wright and Dyson, 2015). Even in the absence of intrinsic disorder, there is growing evidence that conformational flexibility is important in regulating protein-protein interactions (Dobbins et al., 2008; Stefl et al., 2013; Uversky, 2013). This effect has also been shown for proteins that have multiple partners (hubs) and are essential in protein-protein communication and signaling. Hubs' promiscuous binding sites have been demonstrated to display specific dynamical properties, pre-existing in the isolated state of the individual protein, allowing for polyvalent partner binding (Fornili et al., 2013). In any case, quantification of the occurrence of SNPs in disordered and flexible protein regions is a complex task, because different shades of disorder have been identified as playing a role in protein function stability and binding (Uversky et al., 2014; Wright and Dyson, 2015 and references therein). One particularly interesting case is represented by mutations related to disorder-to-order (D-O) transitions; there are often associated to post-translational modifications or with defense mechanisms to protect proteins from toxic aggregation and oxidative stress (Winter et al., 2008) and therefore may result in a stronger impact on the protein functional role. Consequently, order/disorder-sensitive descriptors of the specific chemico-physical environment in the vicinity of the observed variant are needed to evaluate rigorously the relationship between disorder and disease-related mutations.

We aim to contribute to this debate by exploring and quantifying in a systematic way the relationship between order/disorder and the occurrence of common variants (dbSNP: common variations from the 1000 Genomes project, Sherry et al., 2001), disease-related SNPs (OMIM: Mendelian genetic diseases, Hamosh et al., 2005) and cancer-related SNPs (COSMIC, Forbes et al., 2011). To this end we decompose the protein sequences anatomically in folded domain regions, unfolded-disordered (intra-domain) regions and inter-domain disordered regions and calculate the enrichment/depletion of SNPs in each of these regions. These comparisons based on mapping SNPs on static, crystallographic structures, represent a first step in quantifying the different roles played by the two (ordered vs. disordered) environments in which a common or pathogenic mutation may occur. We also explored scenarios of mutual effects of mutations in ordered regions on the disorder content within that domain. We discuss two cases of hubs that are strongly involved in cancer: BRAF (Haling et al., 2014; Thevakumaran et al., 2015) and JAK2 (Bandaranayake et al., 2012), both with phenotypic pathogenic mutations occurring in ordered regions and affecting the disorder content of distal sites in the domain.

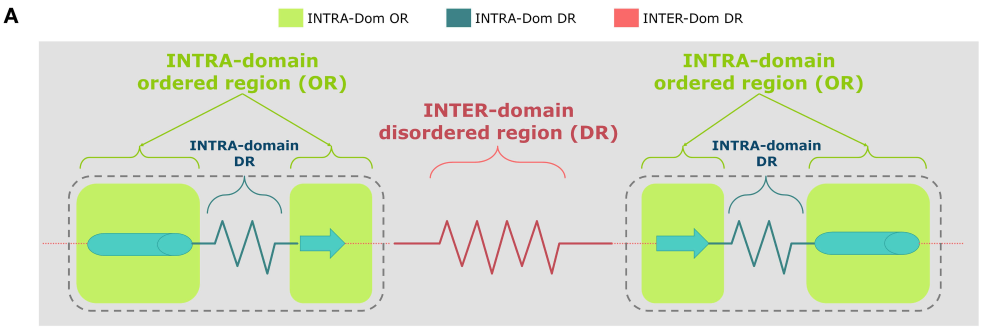
Results

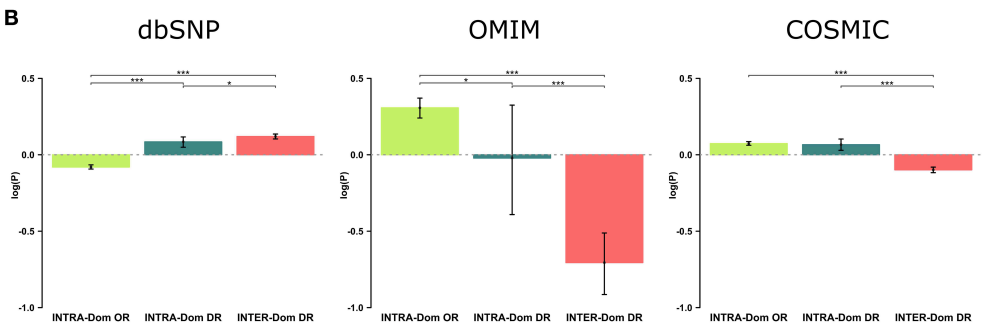
Dearth of Disease-Related SNPs in Inter-domain Disordered Regions

The relative enrichment of SNPs in the dissected disordered regions of the protein have been analyzed by comparing three different classes of SNPs: (a) the common variants from the 1000 Genomes project (dbSNP) (Sherry et al., 2001), (b) the genetic-disease variants from OMIM (Hamosh et al., 2005), and (c) the COSMIC cancer-related SNPs (Forbes et al., 2011). Details on the enrichment/depletion measures are given in the Section "Strategy for the investigation of disordered regions and SNPs occurrence."

The outcome of our analysis is presented **Figure 1B** and the barplots relative to each region are colored according to the scheme in **Figure 1A**. The results for dbSNP data are reported as a comparison of the observed human variation in the analyzed regions vs. the pathogenic mutations observed for OMIM and COSMIC data. To our knowledge, this is the first time that such a comparison is presented. The results have been statistically tested (see Strategy section) and the *p*-values of the comparison tests between the distributions are annotated with stars to show their significance (see **Figure 1** legend for clarification).

The most striking difference amongst all data lies in the opposite behavior observed for INTER-domain disordered regions (INTER-Dom DRs, red) in dbSNP vs. OMIM and COSMIC data (enrichment vs. depletion, respectively). The reasons for such trend can be ascribed to the fact that common variations usually do not occur in structurally and functionally constrained regions but rather accumulate in disordered regions, particularly inter-domain ones. These are usually more flexible to allow the orientation of protein domains and binding multiplicity (Fong and Panchenko, 2010). Conversely, an opposite trend is observed for the INTRA-domain ordered regions (INTRA-Dom OR, light green) of dbSNP vs. the disease-related INTRA-Dom OR plots. For both the disease-related OMIM (**Figure 1B**, center) and COSMIC (**Figure 1B**, right) datasets, there is clear evidence that pathogenic mutations are enriched in ordered domain regions. These are the fragile sites that once mutated can cause a functional impairment of the protein either by destabilizing the fold (Studer et al., 2013), or by affecting structurally important regions for partner binding and consequent signaling activity (Yates and Sternberg, 2013). The enrichment in INTER-Dom disordered regions vs. INTRA-Dom ordered regions is particularly pronounced for the OMIM dataset, but also significantly important for the COSMIC data. The difference in the relative order/disorder populations of the two datasets might be related to the fact that mutations with Mendelian inheritance





C

	N(Proteins)	N(SNPs)	N(SNPs) INTRA-Dom OR	N(SNPs) INTRA-Dom DR	N(SNPs) INTER-Dom DR
dbSNP subset	5,504	148,828	58,486	9,255	48,402
<L>(region) ± SEM			268±3	36±1	182±4
OMIM subset	700	3138	2,013	213	344
<L>(region) ± SEM			355±13	52±5	167±10
COSMIC subset	5,422	102,972	46,870	6,253	26,458
<L>(region) ± SEM			268±4	36±1	180±4

FIGURE 1 | Analyses of non-synonymous single nucleotide polymorphisms (SNPs) in intra-domain ordered regions, intra-domain disordered regions and inter-domain disordered regions. (A) Scheme of protein regions. A protein contains (intra–)domain regions (dashed boundary line) and inter-domain regions. Domain regions contain ordered regions (INTRA-Dom ORs; light-green squares) and disordered regions (INTRA-Dom DRs; dark green zigzag line). Inter-domain regions are predominantly disordered (INTER-Dom DR; red zigzag line). **(B)** SNP frequency analysis. The propensity of SNPs P(SNP) to occur in each region was calculated using Equation 1. Average propensity values are reported as relative entropies log(P(SNP)). Error bars were estimated using bootstrap re-sampling with 10,000 replicates. Stars denote the alpha levels of the test statistics (*$p < 0.05$; ***$p < 0.001$). **(C)** Number of SNPs mapped onto different protein regions. The number of nsSNPs in each class and the average lengths of the protein regions are listed together with the standard error of the mean (SEM). The column "N(proteins)" contains the number of proteins selected for the study of a SNP class, while column "N(SNPs)" reports the total number of SNPs mapped onto the reference proteins.

are potentially more harmful to the protein than some of the passenger mutations observed in cancer.

Our results support previous studies that compared differences in "natural" mutations from dbSNP and disease-associated OMIM data (De Beer et al., 2013). The difference in order vs. disorder propensities observed in our study is therefore an additional discriminant in evaluating the mutability of proteins.

Examples of Intra-domain Mutations and Effects on Disorder Occurrence

In a number of recent studies it has been reported that disordered regions harbor pathogenic mutations (Iakoucheva et al., 2002; Uversky et al., 2008; Babu et al., 2011; Hu et al., 2011; Pajkos et al., 2012; Vacic and Iakoucheva, 2012; Vacic et al., 2012). Some of these observations referred to SNPs in segments involved in D-O transitions, but as we observed a clear dearth of pathogenic mutations in INTER-domain disordered regions (INTER-Dom DRs), we decided to investigate the occurrence of SNPs in INTRA-Dom DRs in more detail. A particularly interesting case is the mutual effect of intra-domain pathogenic mutations and disorder observed within the domain, even at sites distant from the original mutation. We found such examples in BRAF and JAK2 kinases, which are involved in cancer pathologies (Vogelstein and Kinzler, 2004).

We previously studied the BRAF V600E mutation that destabilizes the inactive conformation of the BRAF kinase and consequently induces ERK activation (Satoh et al., 2012; Lu et al., 2013). The V600 residue is in a cluster of hydrophobic residues with Phe468, therefore the presence of a negative charge (residue E) will be disruptive for this cluster, resulting in destabilization of the inactive conformation. Interestingly, introducing the V600E mutation in the BRAF protein kinase domain increases the INTRA-Dom DRs prediction, as shown in the table (**Figure 2A**) and the plot (**Figure 2B**). By running the DISOPRED2 predictor for the V600E mutant, one can observe an increase in the span of the predicted disordered region found in a distal site (607–611). Notably, the predicted disorder region span was not affected by mutations found within the INTRA-Dom DRs (yellow residues in **Figure 2A** for BRAF). These findings suggest that, besides destabilizing the hydrophobic cluster, the V-E substitution in the kinase domain (Pkinase_Tyr(PF07714)) might also have an effect on the INTRA-Dom disorder content by unwinding the downstream loop, as shown in the wild type 3D structure (4MNE_B) (Haling et al., 2014) (**Figure 2B** and Figure S1). This could in turn affect the ligand-binding region (structure 4WO5_A) (Thevakumaran et al., 2015), with a possible impact on the binding affinity.

The mutation V600E has been studied in detail by sophisticated enhanced sampling methods (Marino et al., 2015) and one of the main consequences of the pathogenic variant highlighted in this study is reflected the enhancement of the active-to-inactive state barrier and the increased flexibility (disorder) of the activation loop (region 602–612). These combined effects result in keeping the kinase in an active state and therefore favor phosphorylation to occur. This study supports the idea that an accurate descriptions of the structure, dynamics, and energetics of the protein and its mutated states are necessary to extract molecular fingerprints that rationalize the impact of pathogenic vs. commonly occurring mutations. Interestingly, in recent times the tendency of BRAF in adopting permanently an active state not detectable by current structure has been highlighted as one of the paradigmatic cases for which the currently adopted strategies for structure-based drug discovery may be ineffective (Holderfield et al., 2014).

Long-range effects of mutations on domain-disorder content are partially observed also for the V617F SNP of the JAK2 kinase, a mutation mostly observed in leukaemias. Our predictions indicate that this mutation leads to an extension of the INTRA-Dom DRs, (Bandaranayake et al., 2012) as shown in the table (Figure S2D) and the plot (Figure S2E).

The changes of disorder probability between the wild type sequences (BRAF and JAK2) and those with the cancer driver SNPs (V600E and V617F, respectively) have been predicted by five different methods which include highly ranked methods in CASP10 (Monastyrskyy et al., 2014) such as DISOPRED3, PrDOS, Biomine_MFDp and a recent method using backbone dynamics, DynaMine (Cilia et al., 2014) (Figures S2, S4). The results do not show a strict consensus in the boundaries and in the absolute differences of disorder content, this can be ascribed to the different algorithms used. However, most of the methods predict an increase of the disorder probability in the mutation distal regions we observe for BRAF that the cancer driver mutation is at the periphery of the kinase binding site and in an ordered region, while the non-driver mutations mostly occur in the disordered regions. The two locations seem to be correlated in the sense that the observed change in the driver mutation alters the disorder content of the other mutation loci. This may be ascribed to correlated dynamical couplings between disordered and ordered regions within the same protein domain that may lead to an enrichment of pathogenic variants in flexible and less structured regions. This long-range coupling is an indirect and probably down-tuned mutational effect on the protein function, which may result in a higher acceptance of the mutation in these regions.

Strategy for the Investigation of Disordered Regions and SNP Frequencies

Data Set Preparation

A data set of human proteins, using UniProt accession identifiers as reference, was generated by mapping SNPs onto experimentally resolved 3D structures of proteins. Native and homologous structures were identified by running NCBI-BLAST (version 2.2.29+) (Camacho et al., 2009) against the PDB sequence library. Homologues were accepted above the 30% sequence identity threshold. Non-synonymous SNPs were retrieved from the dbSNP database (build 141) (Sherry et al., 2001), germ-line disease-related SNPs were extracted from the "Online Mendelian Inheritance in Man" (OMIM) database (version July 2014) (Hamosh et al., 2005) and somatic cancer-related SNPs were taken from the "Catalog of Somatic Mutations in Cancer" (COSMIC) database (version July 2014) (Forbes et al., 2011). Only the proteins having a native/homologous structure and SNP information were selected, yielding a reference data set comprising 5587 proteins.

Definition of Protein Domains and Disordered Regions

The selected proteins were assigned with domain definitions and disordered region predictions. For each protein sequence of the reference data set as query, the HMM sequence aligner HMMER3 (Finn et al., 2011) was used to search against the Pfam domain sequence library (Pfam-A.hmm version 26.0) (Punta et al., 2012) and to assign the matched PFAM domain definition to the query protein, given the alignment *E*-value was smaller than 1e-3. Disordered regions of the selected proteins were predicted using the DISOPRED program (Ward et al., 2004). The combination of domain definitions and disordered region predictions leads to three distinct regional classes (**Figure 1A**): (1) intra-domain ordered region (INTRA-Dom OR), (2) intra-domain disordered region (INTRA-Dom DR), and (3) inter-domain disordered region (INTER-Dom DR).

SNPs Enrichment Analysis

We computed the regional enrichment/depletion of SNPs as propensities P(SNPregion) by normalizing the relative regional

A

Position	BRAF_WT	R444Q	V600E	S607P	H608R	H608L	Q609R	F610S	E611D
600	V . 0.008 (0.018)	V . 0.009 (0.019)	E . 0.008 (0.025)	V . 0.012 (0.020)	V . 0.006 (0.021)	V . 0.013 (0.026)	V . 0.003 (0.015)	V . 0.008 (0.016)	V . 0.007 (0.019)
601	K . 0.012 (0.025)	K . 0.014 (0.026)	K . 0.014 (0.031)	K . 0.017 (0.028)	K . 0.011 (0.026)	K . 0.016 (0.030)	K . 0.006 (0.021)	K . 0.013 (0.022)	K . 0.013 (0.027)
602	S . 0.018 (0.033)	S . 0.020 (0.035)	S . 0.019 (0.037)	S . 0.019 (0.037)	S . 0.016 (0.034)	S . 0.024 (0.038)	S . 0.010 (0.028)	S . 0.017 (0.030)	S . 0.019 (0.037)
603	R . 0.047 (0.038)	R . 0.049 (0.039)	R . 0.083 (0.042)	R . 0.061 (0.040)	R . 0.063 (0.039)	R . 0.070 (0.043)	R . 0.032 (0.032)	R . 0.039 (0.034)	R . 0.040 (0.042)
604	W . 0.043 (0.041)	W . 0.046 (0.042)	W . 0.057 (0.048)	W . 0.060 (0.042)	W . 0.066 (0.045)	W . 0.080 (0.048)	W . 0.033 (0.035)	W . 0.041 (0.037)	W . 0.046 (0.043)
605	S . 0.070 (0.045)	S . 0.076 (0.047)	S * 0.151 (0.055)	S . 0.090 (0.046)	S . 0.133 (0.049)	S . 0.161 (0.051)	S . 0.072 (0.039)	S . 0.067 (0.041)	S . 0.081 (0.047)
606	G . 0.137 (0.050)	G * 0.141 (0.053)	G * 0.235 (0.058)	G * 0.136 (0.052)	G * 0.194 (0.054)	G * 0.214 (0.057)	G . 0.139 (0.045)	G . 0.125 (0.046)	G * 0.141 (0.054)
607	S * 0.187 (0.060)	S * 0.202 (0.062)	S * 0.222 (0.068)	P * 0.208 (0.061)	S * 0.196 (0.063)	S * 0.216 (0.065)	S * 0.167 (0.054)	S * 0.154 (0.054)	S * 0.198 (0.066)
608	H * 0.242 (0.065)	H * 0.259 (0.067)	H * 0.271 (0.066)	H * 0.269 (0.065)	R * 0.223 (0.063)	L * 0.242 (0.065)	H * 0.223 (0.057)	H * 0.218 (0.057)	H * 0.290 (0.069)
609	Q * 0.328 (0.065)	Q * 0.326 (0.067)	Q * 0.267 (0.064)	Q * 0.331 (0.065)	Q * 0.284 (0.061)	Q * 0.305 (0.06)	R * 0.283 (0.055)	Q * 0.287 (0.057)	Q * 0.374 (0.069)
610	F * 0.257 (0.064)	F * 0.261 (0.065)	F * 0.211 (0.062)	F * 0.236 (0.063)	F * 0.219 (0.061)	F * 0.188 (0.061)	F * 0.209 (0.055)	S * 0.223 (0.058)	F * 0.275 (0.069)
611	E * 0.146 (0.062)	E * 0.155 (0.062)	E * 0.127 (0.059)	E * 0.123 (0.058)	E * 0.122 (0.061)	E * 0.099 (0.058)	E * 0.124 (0.053)	E * 0.136 (0.056)	D * 0.142 (0.065)
612	Q . 0.172 (0.051)	Q . 0.166 (0.052)	Q * 0.145 (0.053)	Q . 0.125 (0.046)	Q . 0.155 (0.051)	Q . 0.115 (0.048)	Q . 0.139 (0.044)	Q . 0.162 (0.046)	Q . 0.173 (0.051)
613	L . 0.113 (0.048)	L . 0.113 (0.049)	L . 0.092 (0.049)	L . 0.085 (0.044)	L . 0.100 (0.048)	L . 0.082 (0.045)	L . 0.096 (0.043)	L . 0.106 (0.044)	L . 0.102 (0.048)
614	S . 0.067 (0.049)	S . 0.067 (0.050)	S . 0.049 (0.046)	S . 0.055 (0.046)	S . 0.069 (0.046)	S . 0.048 (0.044)	S . 0.059 (0.043)	S . 0.062 (0.045)	S . 0.059 (0.051)
615	G . 0.042 (0.045)	G . 0.039 (0.045)	G . 0.031 (0.040)	G . 0.033 (0.043)	G . 0.039 (0.041)	G . 0.035 (0.039)	G . 0.036 (0.039)	G . 0.042 (0.041)	G . 0.031 (0.047)

B

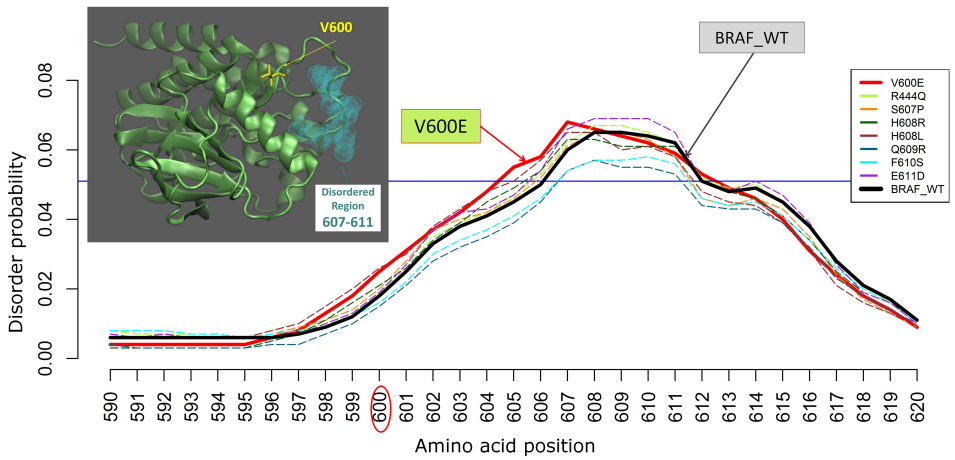


FIGURE 2 | Example of changes in disordered regions (DRs) conferred by SNPs in distant ordered regions. (A) Disorder prediction by DISOPRED2 of wild type (WT) and mutated sequence segments (600–615) of BRAF. Each column is labeled with the specific SNP used for DR prediction and contains the confidence scores of the DISOPRED2 prediction involving raw scores of disorder probability and their filtered scores with parentheses. The residues in DRs are annotated with (*) asterisks and colored in blue. SNPs within the sequence segment 600–615 are colored in yellow. **(B)** Plot of the DISOPRED2 filtered confidence scores of the BRAF WT and mutated sequences. The predicted behavior of V600E (red line) is distinct from that of the BRAF WT sequence (thick black line). The horizontal blue line indicates 5% of filter threshold of the method. The inset shows the 3D structure of the BRAF kinase domain (4MNE_B, cyan cartoon), the location of residue V600 (yellow licorice) and the predicted disordered positions (light green spheres).

frequency with the relative frequency over the total protein length (Equation 1).

$$P\left(SNP_{region}\right) = \frac{(N(SNPs)_{region}/length_{region})}{(N(SNPs)_{protein}/length_{protein})} \quad (1)$$

These propensities are plotted in **Figure 1B** as relative entropies log(P(SNPregion)). A relative entropy of zero indicates a regional frequency equal to the background frequency (denominator), positive values indicate relative enrichment and negative values correspond to relative depletion. All SNPs (from dbSNP, OMIM, and COSMIC) were mapped onto the protein sequences: 5504 of 5587 proteins were mapped with SNPs from dbSNP, 700 of 5587 with SNPs from OMIM and 5422 of 5587 with SNPs from COSMIC. SNPs from each database were further classified into different classes by mapping their positions onto the corresponding protein regions (INTRA-Dom OR SNPs, INTRA-Dom DR SNPs, and INTER-Dom DR SNPs). The number of SNPs in the different classes and the mean lengths of the regions are given in **Figure 1C**.

Statistical Evaluation

To obtain an estimate of the uncertainty associated with the propensity calculations and to reduce biases incurred by the protein selection procedure, we used a bootstrapping method (R function *boot*()) to create random re-sampled subsets of the

reference data set. 10,000 independent subsets were generated of the SNPs propensities within each pre-defined protein regional class (INTRA-Dom OR SNPs, INTRA-Dom DR SNPs, and INTER-Dom DR SNPs) and the mean of each subset was computed. The distributions of the resampled means are by normally distributed, as expected (Figure S4). Confidence intervals at the 95% level were calculated from the bootstrap distributions. The statistical significance of differences between propensity distributions was calculated by Student's *t*-test on the confidence intervals (Wolfe and Hanley, 2002). The statistical analyses were performed using R (R Core Team, 2014).

Conclusions and Perspectives

We performed a large-scale statistical analysis of the relationship between protein disorder and disease-related mutations. We report that both genetic-disease variants from OMIM and cancer-related SNPs from COSMIC are depleted in disordered regions compared to common human variation. This is in line with the fact that mutations in highly constrained regions of the protein are more likely to be disruptive or deleterious. This is why mutations in ordered states of proteins (domains, ligand-binding sites, PPI sites) have been investigated quite in detail in the last years.

We offer here a starting and objective point to discriminate between completely ordered regions, disordered regions occurring in ordered domains, and inter domain predicted disordered segments. We observe and quantify the result of the mapping of available SNPs data onto a large set of human proteins and their close homologs. From this study a number of interesting cases can be extracted for functional validation and close investigation of the dynamical role played by the disorder content.

New perspectives in the field can be explored from this starting point, as the more complicate cases in which flexibility and/or disorder play a direct role in the protein function have not yet been fully elucidated. Particularly complex are the cases where flexible residues modulate protein binding and promiscuity (Fornili et al., 2013) and disorder-to-order causing mutations (Vacic and Iakoucheva, 2012; Dunker et al., 2015). These more "dynamically" driven processes are difficult to parametrise and the restraints playing a role in selecting the actual functional states are not always quantifiable. At this purpose, systematic studies collecting critical examples of experimentally proved correlations between flexibility, presence of disordered states, D-O transitions, and functional studies are needed in the field for the benchmarking and validation of predictive tools for the impact of pathogenic variation on proteins and their partners. Most recent development in disorder prediction methods exploits successfully the mutual interplay between backbone and side-chain dynamics (Cilia et al., 2014; Kosciolek and Jones, 2015).

Nevertheless, more sophisticated methods are needed to quantify these observations, like large-scale molecular simulations, [so far performed for isolated cases Vacic et al., 2012; Marino et al., 2015] and measurements of conformational signal transduction within protein structures (Pandini et al., 2012). The correlated dynamical couplings between disordered and ordered regions may be exploited in the design of drugs targeting distal sites from the dominant mutation, and by fine-tuning the effects on the overall protein function. Additionally, the possibility to predict the "allosteric" modulation of mutations occurring in regions with a different level of order/disorder and possibly correlated with the same or different pathogenic manifestation can open new avenues to investigate the underlying molecular mechanisms and rectify current strategies for drug-discovery.

Acknowledgments

We thank Dr. Jens Kleinjung and Dr. Nicholas Shaun B. Thomas for their critical reading of the manuscript. This research was supported by the Biotechnology and Biological Sciences Research Council (BB/H018409/1 to FF), the British Heart Foundation (FS/12/41/29724 to AF and FF) and the Leukaemia & Lymphoma Research (to FF). SSC is funded by a Leukaemia & Lymphoma Research Gordon Piller PhD Studentship.

Supplementary Material

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fmolb.2015.00047

References

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., et al. (2010). A method and server for predicting damaging missense mutations. *Nat. Methods* 7, 248–249. doi: 10.1038/nmeth0410-248

Al-Numair, N. S., and Martin, A. C. (2013). The SAAP pipeline and database: tools to analyze the impact and predict the pathogenicity of mutations. *BMC Genomics* 14(Suppl. 3):S4. doi: 10.1186/1471-2164-14-S3-S4

Babu, M. M., van der Lee, R., de Groot, N. S., and Gsponer, J. (2011). Intrinsically disordered proteins: regulation and disease. *Curr. Opin. Struct. Biol.* 21, 432–440. doi: 10.1016/j.sbi.2011.03.011

Bandaranayake, R. M., Ungureanu, D., Shan, Y., Shaw, D. E., Silvennoinen, O., and Hubbard, S. R. (2012). Crystal structures of the JAK2 pseudokinase domain and the pathogenic mutant V617F. *Nat. Struct. Mol. Biol.* 19, 754–759. doi: 10.1038/nsmb.2348

Bromberg, Y., and Rost, B. (2007). SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Res.* 35, 3823–3835. doi: 10.1093/nar/gkm238

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: architecture and applications. *BMC Bioinformatics* 10:421. doi: 10.1186/1471-2105-10-421

Capriotti, E., Fariselli, P., and Casadio, R. (2005). I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res.* 33, W306–W310. doi: 10.1093/nar/gki375

Cilia, E., Pancsa, R., Tompa, P., Lenaerts, T., and Vranken, W. F. (2014). The DynaMine webserver: predicting protein dynamics from sequence. *Nucleic Acids Res.* 42, W264–W270. doi: 10.1093/nar/gku270

Consortium, E. P. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74. doi: 10.1038/nature11247

de Beer, T. A., Laskowski, R. A., Parks, S. L., Sipos, B., Goldman, N., and Thornton, J. M. (2013). Amino acid changes in disease-associated variants differ radically from variants observed in the 1000 genomes project dataset. *PLoS Comput. Biol.* 9:e1003382. doi: 10.1371/journal.pcbi.1003382

Dobbins, S. E., Lesk, V. I., and Sternberg, M. J. (2008). Insights into protein flexibility: the relationship between normal modes and conformational change upon protein-protein docking. *Proc. Natl. Acad. Sci. U.S.A.* 105, 10390–10395. doi: 10.1073/pnas.0802496105

Dunker, A. K., Bondos, S. E., Huang, F., and Oldfield, C. J. (2015). Intrinsically disordered proteins and multicellular organisms. *Semin. Cell Dev. Biol.* 37, 44–55. doi: 10.1016/j.semcdb.2014.09.025

Finn, R. D., Clements, J., and Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* 39, W29–W37. doi: 10.1093/nar/gkr367

Fong, J. H., and Panchenko, A. R. (2010). Intrinsic disorder and protein multibinding in domain, terminal, and linker regions. *Mol. Biosyst.* 6, 1821–1828. doi: 10.1039/c005144f

Forbes, S. A., Bindal, N., Bamford, S., Cole, C., Kok, C. Y., Beare, D., et al. (2011). COSMIC: mining complete cancer genomes in the catalogue of somatic mutations in cancer. *Nucleic Acids Res.* 39, D945–D950. doi: 10.1093/nar/gkq929

Fornili, A., Pandini, A., Lu, H.-C., and Fraternali, F. (2013). Specialized dynamical properties of promiscuous residues revealed by simulated conformational ensembles. *J. Chem. Theory Comput.* 9, 5127–5147. doi: 10.1021/ct400486p

Haling, J. R., Sudhamsu, J., Yen, I., Sideris, S., Sandoval, W., Phung, W., et al. (2014). Structure of the BRAF-MEK complex reveals a kinase activity independent role for BRAF in MAPK signaling. *Cancer Cell* 26, 402–413. doi: 10.1016/j.ccr.2014.07.007

Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., and McKusick, V. A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* 33, D514–D517. doi: 10.1093/nar/gki033

Holderfield, M., Deuker, M. M., McCormick, F., and McMahon, M. (2014). Targeting RAF kinases for cancer therapy: BRAF-mutated melanoma and beyond. *Nat. Rev. Cancer* 14, 455–467. doi: 10.1038/nrc3760

Hu, Y., Liu, Y., Jung, J., Dunker, A. K., and Wang, Y. (2011). Changes in predicted protein disorder tendency may contribute to disease risk. *BMC Genomics* 12(Suppl. 5):S2. doi: 10.1186/1471-2164-12-S5-S2

Iakoucheva, L. M., Brown, C. J., Lawson, J. D., Obradovic, Z., and Dunker, A. K. (2002). Intrinsic disorder in cell-signaling and cancer-associated proteins. *J. Mol. Biol.* 323, 573–584. doi: 10.1016/S0022-2836(02)00969-5

Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N., and Sternberg, M. J. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* 10, 845–858. doi: 10.1038/nprot.2015.053

Kircher, M., Witten, D. M., Jain, P., O'roak, B. J., Cooper, G. M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* 46, 310–315. doi: 10.1038/ng.2892

Kosciolek, T., and Jones, D. T. (2015). "Investigations of structural ensembles and disorder-to-order transitions in intrinsically disordered proteins," in *3DSIG An ISMB Satellite Meeting: 3DSIG Structural Bioinformatics and Computational Biophysics* (Dublin), 30.

Ling, H., Vincent, K., Pichler, M., Fodde, R., Berindan-Neagoe, I., Slack, F. J., et al. (2015). Junk DNA and the long non-coding RNA twist in cancer genetics. *Oncogene*. doi: 10.1038/onc.2014.456. [Epub ahead of print].

Lu, H. C., Fornili, A., and Fraternali, F. (2013). Protein-protein interaction networks studies and importance of 3D structure knowledge. *Expert Rev. Proteomics* 10, 511–520. doi: 10.1586/14789450.2013.856764

Marino, K. A., Sutto, L., and Gervasio, F. L. (2015). The effect of a widespread cancer-causing mutation on the inactive to active dynamics of the B-Raf kinase. *J. Am. Chem. Soc.* 137, 5280–5283. doi: 10.1021/jacs.5b01421

Monastyrskyy, B., Kryshtafovych, A., Moult, J., Tramontano, A., and Fidelis, K. (2014). Assessment of protein disorder region predictions in CASP10. *Proteins* 82(Suppl. 2), 127–137. doi: 10.1002/prot.24391

Mosca, R., Tenorio-Laranga, J., Olivella, R., Alcalde, V., Céol, A., Soler-López, M., et al. (2015). dSysMap: exploring the edgetic role of disease mutations. *Nat. Methods* 12, 167–168. doi: 10.1038/nmeth.3289

Necsulea, A., and Kaessmann, H. (2014). Evolutionary dynamics of coding and non-coding transcriptomes. *Nat. Rev. Genet.* 15, 734–748. doi: 10.1038/nrg3802

Nishi, H., Tyagi, M., Teng, S., Shoemaker, B. A., Hashimoto, K., Alexov, E., et al. (2013). Cancer missense mutations alter binding properties of proteins and their interaction networks. *PLoS ONE* 8:e66273. doi: 10.1371/journal.pone.0066273

Pajkos, M., Meszaros, B., Simon, I., and Dosztanyi, Z. (2012). Is there a biological cost of protein disorder? Analysis of cancer-associated mutations. *Mol. Biosyst.* 8, 296–307. doi: 10.1039/C1MB05246B

Pandini, A., Fornili, A., Fraternali, F., and Kleinjung, J. (2012). Detection of allosteric signal transmission by information-theoretic analysis of protein dynamics. *FASEB J.* 26, 868–881. doi: 10.1096/fj.11-190868

Pires, D. E., Ascher, D. B., and Blundell, T. L. (2014). mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics* 30, 335–342. doi: 10.1093/bioinformatics/btt691

Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., et al. (2012). The Pfam protein families database. *Nucleic Acids Res.* 40, D290–D301. doi: 10.1093/nar/gkr1065

R Core Team. (2014). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Reva, B., Antipin, Y., and Sander, C. (2011). Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.* 39, e118. doi: 10.1093/nar/gkr407

Satoh, T., Smith, A., Sarde, A., Lu, H. C., Mian, S., Trouillet, C., et al. (2012). B-RAF mutant alleles associated with Langerhans cell histiocytosis, a granulomatous pediatric disease. *PLoS ONE* 7:e33891. doi: 10.1371/annotation/74a67f4e-a536-4b3f-a350-9a4c1e6bebbd

Scharner, J., Lu, H. C., Fraternali, F., Ellis, J. A., and Zammit, P. S. (2014). Mapping disease-related missense mutations in the immunoglobulin-like fold domain of lamin A/C reveals novel genotype-phenotype associations for laminopathies. *Proteins* 82, 904–915. doi: 10.1002/prot.24465

Schuster-Böckler, B., and Bateman, A. (2008). Protein interactions in human genetic diseases. *Genome Biol.* 9:R9. doi: 10.1186/gb-2008-9-1-r9

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., et al. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 29, 308–311. doi: 10.1093/nar/29.1.308

Shihab, H. A., Gough, J., Cooper, D. N., Stenson, P. D., Barker, G. L., Edwards, K. J., et al. (2013). Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum. Mutat.* 34, 57–65. doi: 10.1002/humu.22225

Stefl, S., Nishi, H., Petukh, M., Panchenko, A. R., and Alexov, E. (2013). Molecular mechanisms of disease-causing missense mutations. *J. Mol. Biol.* 425, 3919–3936. doi: 10.1016/j.jmb.2013.07.014

Studer, R. A., Dessailly, B. H., and Orengo, C. A. (2013). Residue mutations and their impact on protein structure and function: detecting beneficial and pathogenic changes. *Biochem. J.* 449, 581–594. doi: 10.1042/BJ20121221

Thevakumaran, N., Lavoie, H., Critton, D. A., Tebben, A., Marinier, A., Sicheri, F., et al. (2015). Crystal structure of a BRAF kinase domain monomer explains basis for allosteric regulation. *Nat. Struct. Mol. Biol.* 22, 37–43. doi: 10.1038/nsmb.2924

Thomas, P. D., and Kejariwal, A. (2004). Coding single-nucleotide polymorphisms associated with complex vs. Mendelian disease: evolutionary evidence for differences in molecular effects. *Proc. Natl. Acad. Sci. U.S.A.* 101, 15398–15403. doi: 10.1073/pnas.0404380101

Uversky, V. N. (2013). Under-folded proteins: conformational ensembles and their roles in protein folding, function, and pathogenesis. *Biopolymers* 99, 870–887. doi: 10.1002/bip.22298

Uversky, V. N., Dave, V., Iakoucheva, L. M., Malaney, P., Metallo, S. J., Pathak, R. R., et al. (2014). Pathological unfoldomics of uncontrolled chaos: intrinsically disordered proteins and human diseases. *Chem. Rev.* 114, 6844–6879. doi: 10.1021/cr400713r

Uversky, V. N., Oldfield, C. J., and Dunker, A. K. (2008). Intrinsically disordered proteins in human diseases: introducing the D2 concept. *Annu. Rev. Biophys.* 37, 215–246. doi: 10.1146/annurev.biophys.37.032807.125924

Vacic, V., and Iakoucheva, L. M. (2012). Disease mutations in disordered regions–exception to the rule? *Mol. Biosyst.* 8, 27–32. doi: 10.1039/C1MB05251A

Vacic, V., Markwick, P. R., Oldfield, C. J., Zhao, X., Haynes, C., Uversky, V. N., et al. (2012). Disease-associated mutations disrupt functionally important regions of intrinsic protein disorder. *PLoS Comput. Biol.* 8:e1002709. doi: 10.1371/journal.pcbi.1002709

Vogelstein, B., and Kinzler, K. W. (2004). Cancer genes and the pathways they control. *Nat. Med.* 10, 789–799. doi: 10.1038/nm1087

Wang, X., Wei, X., Thijssen, B., Das, J., Lipkin, S. M., and Yu, H. (2012). Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nat. Biotechnol.* 30, 159–164. doi: 10.1038/nbt.2106

Ward, J. J., McGuffin, L. J., Bryson, K., Buxton, B. F., and Jones, D. T. (2004). The DISOPRED server for the prediction of protein disorder. *Bioinformatics* 20, 2138–2139. doi: 10.1093/bioinformatics/bth195

Winter, J., Ilbert, M., Graf, P. C., Ozcelik, D., and Jakob, U. (2008). Bleach activates a redox-regulated chaperone by oxidative protein unfolding. *Cell* 135, 691–701. doi: 10.1016/j.cell.2008.09.024

Wolfe, R., and Hanley, J. (2002). If we're so different, why do we keep overlapping? When 1 plus 1 doesn't make 2. *CMAJ* 166, 65–66. Available online at: http://www.cmaj.ca/content/166/1/65.full

Wright, P. E., and Dyson, H. J. (2015). Intrinsically disordered proteins in cellular signalling and regulation. *Nat. Rev. Mol. Cell Biol.* 16, 18–29. doi: 10.1038/nrm3920

Yates, C. M., Filippis, I., Kelley, L. A., and Sternberg, M. J. (2014). SuSPect: enhanced prediction of single amino acid variant (SAV) phenotype using network features. *J. Mol. Biol.* 426, 2692–2701. doi: 10.1016/j.jmb.2014.04.026

Yates, C. M., and Sternberg, M. J. (2013). The effects of non-synonymous single nucleotide polymorphisms (nsSNPs) on protein-protein interactions. *J. Mol. Biol.* 425, 3949–3963. doi: 10.1016/j.jmb.2013.07.012

Yue, P., Li, Z., and Moult, J. (2005). Loss of protein structure stability as a major causative factor in monogenic disease. *J. Mol. Biol.* 353, 459–473. doi: 10.1016/j.jmb.2005.08.020

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.