MRC Doctoral Training Partnership in Biomedical Sciences

Cover Sheet – Rotation Project Report

Student name	MR TIMIR G R WESTON
Rotation 1, 2 or 3	1 **2** 3
Report title	Assessing the impact of variants in Cardiomyopathies and the role of the obscurin gene
Co-supervisor 1	PROFESSOR FRANCA FRATERNALI
Co-supervisor 2	PROFESSOR MATHIAS GAUTEL
Clinical supervisor (if appropriate)	N/A
Submission date	20/03/19
Word count	2496

Guidance notes:

1. The project report should be written in font size 12, with 1.5-line spacing
2. This cover sheet must be completed and should be the front page of the report
3. The report should be submitted to KEATS via Turnitin as a single PDF file saved in the file name format: FAMILY NAME-FIRST INITIAL_ Report X (1, 2 or 3) and by the set deadline

Please copy the abstract from the project report and paste it into the box below:

Abstract

Giant modular sarcomeric proteins such as obscurin, titin, and myomesin have come under increasing scrutiny for their contributions to disease, most notably cardiomyopathies. Individual missense mutations have been shown to have far-reaching structural consequences that can impact binding affinities and thus the integrity of the sarcomere, which would cause or modulate the disease state; however, owing to the size of these proteins, assessing the impact of all possible SNVs is a daunting task. Obscurin is currently not comprehensively studied, as while several variants have been strongly linked with cardiomyopathies through computational and functional studies, the vast majority are likely as yet undiscovered. Furthermore, obscurin's modular architecture necessitates the use of structural information in addition to sequence data, which is not always available. We propose a centralised resource for information relating to obscurin variants that would enable users to call pertinent data about SNVs of interest, map variants to domains, and visualise these variants on domain structural models, as has been created previously for titin. To this end, we have created a basic proof-of-concept database (available at https://github.com/Fraternalilab/obscurinDB) which allows for simple queries to be conducted and information to be extracted and have illustrated certain uses to which the resource might be put, including the distribution of SNVs across the length of the protein or domains, and the visualisation of disease-associated SNVs on their respective structures. We also discuss the current state of research in this area and what improvements should be made before the resource or anything similar is made available for use.

MRC Doctoral Training Partnership in Biomedical Sciences

Capsule

Background: Obscurins play a role in cardiomyopathies, but information is limited.
Results: We created a basic resource to collate and allow easy navigation of information concerning SNVs in obscurin.
Conclusion: A resource is helpful in classifying obscurin variants.
Significance: A finalised resource will be useful in further study of obscurin-related diseases.

Abstract

Giant modular sarcomeric proteins such as obscurin, titin, and myomesin have come under increasing scrutiny for their contributions to disease, most notably cardiomyopathies. Individual missense mutations have been shown to have far-reaching structural consequences that can impact binding affinities and thus the integrity of the sarcomere, which would cause or modulate the disease state; however, owing to the size of these proteins, assessing the impact of all possible SNVs is a daunting task. Obscurin is currently not comprehensively studied, as while several variants have been strongly linked with cardiomyopathies through computational and functional studies, the vast majority are likely as yet undiscovered. Furthermore, obscurin's modular architecture necessitates the use of structural information in addition to sequence data, which is not always available. We propose a centralised resource for information relating to obscurin variants that would enable users to call pertinent data about SNVs of interest, map variants to domains, and visualise these variants on domain structural models, as has been created previously for titin. To this end, we have created a basic proof-of-concept database (available at https://github.com/Fraternalilab/obscurinDB) which allows for simple queries to be conducted and information to be extracted and have illustrated certain uses to which the resource might be put, including the distribution of SNVs across the length of the protein or domains, and the visualisation of disease-associated SNVs on their respective structures. We also discuss the current state of research in this area and what improvements should be made before the resource or anything similar is made available for use.

Introduction

Obscurins are a family of protein isoforms which range from small proteins of around 40kDa to giant modular proteins of around 800kDa and are responsible for

structural and signalling roles in cardiac and skeletal muscle alike similar giant proteins such as titin(1–4). They are mainly composed of immunoglobulin (Ig) domains; the largest isoform, obscurin-b, also includes a Rho/Rac/Cdc42-like GTPase Guanine nucleotide Exchange Factor (RhoGEF), an IQ calmodulin-binding motif, and two protein kinase domains at the COOH-terminus(5). These kinase domains are absent in the first-discovered, second-largest isoform, obscurin-a (Fig.1). Recent work has implicated obscurin variants in cardiomyopathies, diseases of the myocardium where dysfunction occurs due to issues with the muscle itself, rather than the consequences of other conditions(6). The severity of cardiomyopathies varies greatly, but they can negatively impact quality of life and lead to death or morbidity(7). While the causes of cardiomyopathies can include lifestyle choices, infections, or other diseases, there is substantial evidence that inheritable genetic variants are also associated(8). Many variants in sarcomeric proteins have been linked to hypertrophic cardiomyopathy (HCM), including four missense variants in obscurin(9–11). Other obscurin variants have been associated with other cardiomyopathies such as dilated cardiomyopathy (DCM)(12) and left ventricular non-compaction (LVNC)(13).

Finding and prioritising cardiomyopathy-associated variants in the giant sarcomeric proteins presents several difficulties(14). First, proteins such as obscurin are large enough that on average all individuals will carry some variants, a minority of which will be damaging. Second, because these variants can be found in clinically healthy individuals, rare damaging alleles do not always appear as such when attempting to prioritise variants. Third, given the modular structure of obscurin, single variants can rarely be considered wholly causative of a given condition, and may only have an effect in concert with other mutations, or modify risk or existing phenotypes. Fourth, the frequency of variants can differ greatly between different sub-populations(15), and it is often unclear whether this indicates that a variant is benign or if certain sub-populations are more susceptible to certain cardiomyopathies.

To solve these problems with respect to the giant protein titin, the web tool TITINdb(14) was created, wherein a single, centralised database of data concerning missense variants in titin is made available and accessible to any who require it. Structures for individual domains are included and individual single nucleotide polymorphisms (SNPs) can be visualised on the models, which allows for both sequence- and structure-based predictions of impact on the protein and domain as well as the identification and prioritisation of variants and spatial mutation hotspots. We propose that a similar resource for obscurin would be invaluable for future work elucidating the relationship between individual SNPs in this protein and associated myocardial

conditions. We present a proof-of-concept resource with basic functionality and demonstrations of potential uses, while also reviewing the proposed pathogenic obscurin variants from the existing literature.

Experimental Procedures

Domain Architecture

An initial difficulty in establishing a resource for collecting obscurin variants was the current lack of consensus over domain boundaries. Currently, the most well-defined and referenced domain boundaries are those for the giant obscurin-b isoform; however, this omits several exons and domains present in the *OBSCN* gene. We derived the inferred complete obscurin sequence (hereafter obscurin-IC) from the RefSeq(16) transcript NP_001258152.2 (hereafter RefSeq), along with a consensus architecture of 67 Ig domains and three FnIII domains alongside the functional domains mentioned earlier (Fig.2). We then used protein sequence analysis programs PHMMER(17), NCBI CD-Search(18), and InterPro 72.0(19) to construct domain architectures for obscurin-IC. For InterPro, because different databases gave different domain boundaries for the same domain, the non-overlapping domain boundaries with the smallest E-values were used. These architectures were compared to one another, to domain architectures from bioinformatics resources (16, 20), and to existing experimental structures to develop a consensus architecture. The alignment software T-Coffee(21) was also used to determine the similarity of the Ig domains in the architectures to give an estimate of the accuracy of the domain boundaries.

Sources for Missense Variants

We used the current Ensembl database(22) as our primary source for SNVs, including variants from both dbSNP(23) and COSMIC(24, 25). The information provided includes chromosomal position, protein consequence and consequence type (missense, frameshift etc.), pathogenicity prediction scores, the global minor allele frequency (MAF) where possible, variant ID, and transcript ID. We also retrieved all SNV entries from gnomAD(26) for a separate database for illustrations of applications involving population frequencies, but this was not incorporated into the main resource.

Homology Modelling of Obscurin Domains

As experimental structures do not exist for most obscurin domains, we used a homology modelling approach to generate structures for those domains lacking coverage. We used BLASTP(27) to search the Protein Data Bank (PDB) for structures similar in sequence for individual domains, then used the alignment software T-Coffee(21) to generate an alignment which could be used by the homology modelling web server SWISS-MODEL(28) to generate a structure for that domain. The models were validated using the MolProbity(29) web server.

Implementation

The resource makes use of simple Excel features to make pertinent information from a mass database more easily accessible(30). Users can search for a single SNV for all SNVs within a range or ranges using filters and can access information as desired using the pivot table functions. Included are the SNV, their dbSNP or COSMIC variant ID, chromosomal, genetic and protein sequence positions and consequences, as well as pathogenicity metrics and, where available, minor allele frequency (MAF). Population and sub-population frequencies are not yet included.

As the pathogenicity predictions included in the Ensembl database and retrieved from their Variant Effect Predictor did not appear to align with independent calculations performed using the most recent releases of the same, we felt it necessary to recalculate the pathogenicity scores as well as incorporating additional metrics not previously included. The current resource uses SIFT(31), PolyPhen2(32), CADD(33), PROVEAN(34), and Condel(35). We also included a measure of protein stability change using mCSM(36) for currently modelled domains.

Potential Applications

The primary goal for a resource such as this would be to make it easier to prioritise variants that are more likely to be linked to disease. One such method would be through the identification and characterisation of disease-associated hotspots. While there are at present only 11 pathogenic obscurin SNVs reported in the literature (see Table 1), and thus it is not possible to identify hotspots with any certainty, we can still use them to illustrate some ways in which the resource may be used in future.

To see if there were any immediate differences in the protein locations of damaging and common variants, we took the frequency data for every missense SNV from gnomAD and retrieved those SNVs with a frequency in gnomAD of greater than

100,000. Then, we aligned the obscurin Ig domain sequences using T-Coffee and mapped both the pathogenic SNVs and the most common SNVs separately to a representative Ig domain structure (Ig59) and compared the locations of the SNVs.

Some of the known pathogenic SNVs were partly classified as such owing to significant decreases in binding affinity for obscurin with its sarcomeric interaction partners (10, 37). Thus, we attempted to determine the structural consequences of common SNVs on the obscurin-titin-myomesin ternary complex(38, 39) using the crystal structures of Ig1 alone and in complex with titin M10. We queried all common missense SNVs from Ensembl in both structures using mCSM to determine the predicted protein stability change (ΔG) in kCal/mol for each variant, then mapped these to the respective structures. In addition, we looked at whether there was any correlation between residue relative solvent accessibility (RSA) and ΔG for the two structures using the in-built linear regression functionality in R(40), to determine the relationship between how buried a residue is in structure and how the structure is affected by its mutation.

Results

Domain Architecture

While the domain architectures posited by the various programs were mostly similar, there were minor differences of a few residues at the domain boundaries. PHMMER was discounted for lacking numerous domains present in the other architectures. Of the remaining architectures (see Fig. 2), RefSeq and CD-Search were broadly similar and most of the central Ig domain boundaries were identical, while InterPro favoured larger Ig and FnIII domains with no interdomain space separating them but corresponded with CD-Search's architecture on the functional domains and protein kinases. T-Coffee's alignment software gave the CD-Search Ig domains a higher score (consensus 848) than those of InterPro (791); consequently, we used the CD-Search architecture as a base, but with domain boundaries for Ig7 from RefSeq (as CD-Search differed strongly from the other two). An additional Ig domain (Ig30) from InterPro was included within a region that CD-Search considered disordered, as it showed homology to the other Ig domains, and the IQ motif was specified. The finalised domain boundaries are given in Table 1, alongside a comparison with existing domain boundaries for obscurin-b and obscurin-IC.

Potential Applications

Figure 3a shows the distribution of common and pathogenic SNVs with reference to the final obscurin domain architecture. There do not seem to be any regions in common, except for the Ig58-Ig61 region, which seems to have a clustering of both pathogenic and high-frequency common SNVs; this suggests that the regions targeted by pathogenic and highly common variants are different, and that the pathogenic hotspots may be in regions vital to protein stability or protein-protein affinity. It is worth noting, however, that two of the pathogenic SNVs in this region (Table 2; R4344Q and A4484T) were not considered damaging by most of the pathogenicity prediction algorithms we used, and that both have had their classification called into question for other reasons (see Discussion), so this may not represent a clustering of pathogenic SNVs regardless. Figures 3b and 3c show the structural positions of the Ig-domain pathogenic SNVs and Ig-domain common SNVs with frequency greater than 100,000, respectively, on a representative structure. The pathogenic SNVs seem to appear more towards the terminal positions in secondary structures, mostly the β-sheets, while the common SNVs seem to be more present in the loops and the middle of secondary structures.

Figure 4a and 4b show the size of the protein stability changes associated with all known SNVs in Ig1 mapped to the structures of Ig1 and the Ig1-M10 complex, respectively. Here, in Ig1 alone, the most destabilising SNVs are to be found in the β-sheet secondary structures, whereas in the complex, the same SNVs seem to have a less destabilising effect in most cases, but there is a region of highly destabilising SNVs in the location most spatially close to the M10 complex. To test whether there was correlation between the RSA and the protein stability change in both cases, we fitted a linear regression model to both (Fig. 4c, 4d). A Pearson product-moment correlation test indicates that both models have significant correlation (Table 3). Although the data violates the assumption of homogeneity for linear regression, we have also reported the R^2 value and standard error as an indication.

Discussion and Conclusions

The need for a resource through which regions and variants of obscurin associated with disease can be elucidated is highlighted by the present difficulty in doing so; attempts to integrate the information available on obscurin and its SNVs are hampered by the lack, and sometimes confusion, of information. For example, the previously-mentioned difficulty over prioritising variants with greatly different allele frequency in different sub-populations is illustrated by the SNVs R4344Q and A4844T. R4344Q, where its low frequency in the general population was given as part of the

reasoning for its initial classification as a likely pathogenic(10); however, further investigation revealed its significantly higher MAF (15%) in the black American sub-population, which casts doubt on its role as a disease-causing variant(15). Similarly, A4844T was also reported as pathogenic, although there was no phenotypic change in models where A4844T was the only mutation(10); it too has a high MAF in the black American sub-population (2%). Subsequently, some authors have reported R4344Q as pathogenic but A4844T as not based on the literature(41), while others have reported the opposite(15). The current consensus, in addition to the pathogenicity metrics reported here, is that both SNPs are likely benign(9). Thus, a resource in which this information is readily available would be of great benefit in future studies on obscurin-linked cardiomyopathies.

The results reported illustrate some of the problems which the resource should in future be able to help solve. With a greater number of pathogenic variants known, it should be possible not only to identify and characterise pathogenic and non-pathogenic variant hotspots, but also determine whether the variants associated with specific cardiomyopathies also cluster, which would also enable us to get closer to a molecular understanding of what separates the different cardiomyopathies. In addition, we mentioned the ternary complex of Ig1 and Ig3 with titin and myomesin, respectively, at the NH2 terminus; other binding partners, such as titin and ankyrin isoforms at the COOH terminus(42), could also be prioritised in the same way, in order to better prioritise the most damaging SNVs. Finally, while we only briefly investigated solvent accessibility as an indicator of pathogenic variant hotspots, the propensity of buried residues to have large effects on structure with deleterious consequences is known(43), and thus a measure of solvent accessibility should be provided for all residues in the resource, as well as appropriate visualisation, so that this can feed into the prioritisation process.

In future, before this resource or one inspired by it is released to the public domain, there are many improvements that should be made or considered. Compared to TITINdb, the current resource only shows the SNV positions on the obscurin-b or obscurin-IC isoforms, rather than the wider range available for titin; furthermore, it lacks the visualisation capacity that is vital to the end user. First, a pipeline should be established to enable homology modelling for all the remaining domains, as only certain relevant domains are currently modelled; then, the ability to view an SNV on the respective structure should be implemented via a molecule viewer such as Jmol(44), so that the positions of different SNVs linked to conditions can be compared. Furthermore, new pipelines should automate the process of adding new SNVs from databases such as gnomAD, ExAC, and 1000G, as well as computing the relevant metrics for each. For

classifying by cardiomyopathy, we should accommodate newly proposed classification systems such as MOGE(S)(45, 46), which may be necessary for future organisation. Finally, while the current Excel-based system is functional, there is an argument for migration to a web-based server, which would simplify the process of updating, make the resource more readily accessible, and arguably improve the ease of use when more functions and properties are added.

Acknowledgements

The author would like to thank Professor Franca Fraternali for her guidance and support throughout this project, Professor Mathias Gautel for his direction and discussion concerning obscurin, Dr Anna Laddach for her help with various scripts used to query the pathogenicity and protein stability programs, and all the members of the Fraternali Lab, especially Joseph Ng and Irene Marzuoli, for their help.

References

1. Young, P., Ehler, E., and Gautel, M. (2001) Obscurin, a giant sarcomeric Rho guanine nucleotide exchange factor protein involved in sarcomere assembly. *The Journal of Cell Biology*. **154**, 123–136
2. Fukuzawa, A., Lange, S., Holt, M., Vihola, A., Carmignac, V., Ferreiro, A., Udd, B., and Gautel, M. (2008) Interactions with titin and myomesin target obscurin and obscurin-like 1 to the M-band – implications for hereditary myopathies. *Journal of Cell Science*. **121**, 1841–1851
3. Borisov, A. B., Sutter, S. B., Kontrogianni-Konstantopoulos, A., Bloch, R. J., Westfall, M. V., and Russell, M. W. (2005) Essential role of obscurin in cardiac myofibrillogenesis and hypertrophic response: evidence from small interfering RNA-mediated gene silencing. *Histochem Cell Biol*. **125**, 227
4. Bowman, A. L., Kontrogianni-Konstantopoulos, A., Hirsch, S. S., Geisler, S. B., Gonzalez-Serratos, H., Russell, M. W., and Bloch, R. J. (2007) Different obscurin isoforms localize to distinct sites at sarcomeres. *FEBS Lett*. **581**, 1549–1554
5. Fukuzawa, A., Idowu, S., and Gautel, M. (2005) Complete human gene structure of obscurin: implications for isoform generation by differential splicing. *J Muscle Res Cell Motil*. **26**, 427–434
6. Brigden, W. (1957) UNCOMMON MYOCARDIAL DISEASES: THE NON-CORONARY CARDIOMYOPATHIES*. *The Lancet*. **270**, 1243–1249
7. Braunwald Eugene (2017) Cardiomyopathies. *Circulation Research*. **121**, 711–721
8. Garfinkel, A. C., Seidman, J. G., and Seidman, C. E. (2018) Genetic Pathogenesis of Hypertrophic and Dilated Cardiomyopathy. *Heart Failure Clinics*. **14**, 139–146
9. Grogan, A., and Kontrogianni-Konstantopoulos, A. (2018) Unraveling obscurins in heart disease. *Pflugers Arch - Eur J Physiol*. 10.1007/s00424-018-2191-3
10. Arimura, T., Matsumoto, Y., Okazaki, O., Hayashi, T., Takahashi, M., Inagaki, N., Hinohara, K., Ashizawa, N., Yano, K., and Kimura, A. (2007) Structural analysis of obscurin gene in hypertrophic cardiomyopathy. *Biochemical and Biophysical Research Communications*. **362**, 281–287
11. Xu, J., Li, Z., Ren, X., Dong, M., Li, J., Shi, X., Zhang, Y., Xie, W., Sun, Z., Liu, X., and Dai, Q. (2015) Investigation of Pathogenic Genes in Chinese sporadic Hypertrophic Cardiomyopathy Patients by Whole Exome Sequencing. *Sci Rep*. **5**, 16609
12. Marston, S., Montgiraud, C., Munster, A. B., Copeland, O., Choi, O., Dos Remedios, C., Messer, A. E., Ehler, E., and Knöll, R. (2015) OBSCN Mutations Associated with Dilated Cardiomyopathy and Haploinsufficiency. *PLoS ONE*. **10**, e0138568
13. Rowland, T. J., Graw, S. L., Sweet, M. E., Gigli, M., Taylor, M. R. G., and Mestroni, L. (2016) Obscurin Variants in Patients With Left Ventricular Noncompaction. *J Am Coll Cardiol*. **68**, 2237–2238
14. Laddach, A., Gautel, M., and Fraternali, F. (2017) TITINdb—a computational tool to assess titin's role as a disease gene. *Bioinformatics*. **33**, 3482–3485
15. Manrai, A. K., Funke, B. H., Rehm, H. L., Olesen, M. S., Maron, B. A., Szolovits, P., Margulies, D. M., Loscalzo, J., and Kohane, I. S. (2016) Genetic Misdiagnoses and the Potential for Health Disparities. *New England Journal of Medicine*. **375**, 655–665
16. O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C. M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V. S., Kodali, V. K., Li, W., Maglott, D., Masterson, P., McGarvey, K. M., Murphy, M. R., O'Neill, K., Pujar, S., Rangwala, S. H., Rausch, D., Riddick, L. D., Schoch, C., Shkeda, A., Storz, S. S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R. E., Vatsan, A. R., Wallin, C., Webb, D., Wu, W., Landrum, M. J., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T. D., and Pruitt, K. D. (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733-745

17. Potter, S. C., Luciani, A., Eddy, S. R., Park, Y., Lopez, R., and Finn, R. D. (2018) HMMER web server: 2018 update. *Nucleic Acids Res*. **46**, W200–W204
18. Marchler-Bauer, A., and Bryant, S. H. (2004) CD-Search: protein domain annotations on the fly. *Nucleic Acids Res*. **32**, W327–W331
19. Mitchell, A. L., Attwood, T. K., Babbitt, P. C., Blum, M., Bork, P., Bridge, A., Brown, S. D., Chang, H.-Y., El-Gebali, S., Fraser, M. I., Gough, J., Haft, D. R., Huang, H., Letunic, I., Lopez, R., Luciani, A., Madeira, F., Marchler-Bauer, A., Mi, H., Natale, D. A., Necci, M., Nuka, G., Orengo, C., Pandurangan, A. P., Paysan-Lafosse, T., Pesseat, S., Potter, S. C., Qureshi, M. A., Rawlings, N. D., Redaschi, N., Richardson, L. J., Rivoire, C., Salazar, G. A., Sangrador-Vegas, A., Sigrist, C. J. A., Sillitoe, I., Sutton, G. G., Thanki, N., Thomas, P. D., Tosatto, S. C. E., Yong, S.-Y., and Finn, R. D. (2019) InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res*. **47**, D351–D360
20. UniProt: a worldwide hub of protein knowledge (2019) *Nucleic Acids Res*. **47**, D506–D515
21. Notredame, C., Higgins, D. G., and Heringa, J. (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **302**, 205–217
22. Zerbino, D. R., Achuthan, P., Akanni, W., Amode, M. R., Barrell, D., Bhai, J., Billis, K., Cummins, C., Gall, A., Girón, C. G., Gil, L., Gordon, L., Haggerty, L., Haskell, E., Hourlier, T., Izuogu, O. G., Janacek, S. H., Juettemann, T., To, J. K., Laird, M. R., Lavidas, I., Liu, Z., Loveland, J. E., Maurel, T., McLaren, W., Moore, B., Mudge, J., Murphy, D. N., Newman, V., Nuhn, M., Ogeh, D., Ong, C. K., Parker, A., Patricio, M., Riat, H. S., Schuilenburg, H., Sheppard, D., Sparrow, H., Taylor, K., Thormann, A., Vullo, A., Walts, B., Zadissa, A., Frankish, A., Hunt, S. E., Kostadima, M., Langridge, N., Martin, F. J., Muffato, M., Perry, E., Ruffier, M., Staines, D. M., Trevanion, S. J., Aken, B. L., Cunningham, F., Yates, A., and Flicek, P. (2018) Ensembl 2018. *Nucleic Acids Res*. **46**, D754–D761
23. Sherry, S. T., Ward, M.-H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., and Sirotkin, K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*. **29**, 308–311
24. Cosmic COSMIC - Catalogue of Somatic Mutations in Cancer. [online] https://cancer.sanger.ac.uk/cosmic (Accessed March 17, 2019)
25. Forbes, S. A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate, J., Cole, C. G., Ward, S., Dawson, E., Ponting, L., Stefancsik, R., Harsha, B., Kok, C. Y., Jia, M., Jubb, H., Sondka, Z., Thompson, S., De, T., and Campbell, P. J. (2017) COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res*. **45**, D777–D783
26. Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., Gauthier, L. D., Brand, H., Solomonson, M., Watts, N. A., Rhodes, D., Singer-Berk, M., Seaby, E. G., Kosmicki, J. A., Walters, R. K., Tashman, K., Farjoun, Y., Banks, E., Poterba, T., Wang, A., Seed, C., Whiffin, N., Chong, J. X., Samocha, K. E., Pierce-Hoffman, E., Zappala, Z., O'Donnell-Luria, A. H., Minikel, E. V., Weisburd, B., Lek, M., Ware, J. S., Vittal, C., Armean, I. M., Bergelson, L., Cibulskis, K., Connolly, K. M., Covarrubias, M., Donnelly, S., Ferriera, S., Gabriel, S., Gentry, J., Gupta, N., Jeandet, T., Kaplan, D., Llanwarne, C., Munshi, R., Novod, S., Petrillo, N., Roazen, D., Ruano-Rubio, V., Saltzman, A., Schleicher, M., Soto, J., Tibbetts, K., Tolonen, C., Wade, G., Talkowski, M. E., Consortium, T. G. A. D., Neale, B. M., Daly, M. J., and MacArthur, D. G. (2019) Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv*. 10.1101/531210
27. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410
28. Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F. T., de Beer, T. A. P., Rempfer, C., Bordoli, L., Lepore, R., and Schwede, T. (2018) SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res*. **46**, W296–W303

29. Williams, C. J., Headd, J. J., Moriarty, N. W., Prisant, M. G., Videau, L. L., Deis, L. N., Verma, V., Keedy, D. A., Hintze, B. J., Chen, V. B., Jain, S., Lewis, S. M., Arendall, W. B., Snoeyink, J., Adams, P. D., Lovell, S. C., Richardson, J. S., and Richardson, D. C. (2018) MolProbity: More and better reference data for improved all-atom structure validation. *Protein Sci.* **27**, 293–315
30. *Database of obscurin single amino-acid variants and mapping to sequence and structural information: Fraternalilab/obscurinDB* (2019) Fraternali Lab, [online] https://github.com/Fraternalilab/obscurinDB (Accessed March 20, 2019)
31. Ng, P. C., and Henikoff, S. (2003) SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res*. **31**, 3812–3814
32. Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., and Sunyaev, S. R. (2010) A method and server for predicting damaging missense mutations. *Nat Methods*. **7**, 248–249
33. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J., and Kircher, M. (2019) CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886–D894
34. Choi, Y., and Chan, A. P. (2015) PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics*. **31**, 2745–2747
35. González-Pérez, A., and López-Bigas, N. (2011) Improving the Assessment of the Outcome of Nonsynonymous SNVs with a Consensus Deleteriousness Score, Condel. *Am J Hum Genet*. **88**, 440–449
36. Pires, D. E. V., Ascher, D. B., and Blundell, T. L. (2014) mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics*. **30**, 335–342
37. Rossi, D., Palmio, J., Evilä, A., Galli, L., Barone, V., Caldwell, T. A., Policke, R. A., Aldkheil, E., Berndsen, C. E., Wright, N. T., Malfatti, E., Brochier, G., Pierantozzi, E., Jordanova, A., Guergueltcheva, V., Romero, N. B., Hackman, P., Eymard, B., Udd, B., and Sorrentino, V. (2017) A novel FLNC frameshift and an OBSCN variant in a family with distal muscular dystrophy. *PLoS One*. 10.1371/journal.pone.0186642
38. Pernigo, S., Fukuzawa, A., Bertz, M., Holt, M., Rief, M., Steiner, R. A., and Gautel, M. (2010) Structural insight into M-band assembly and mechanics from the titin-obscurin-like-1 complex. *PNAS*. **107**, 2908–2913
39. Pernigo, S., Fukuzawa, A., Pandini, A., Holt, M., Kleinjung, J., Gautel, M., and Steiner, R. A. (2015) The Crystal Structure of the Human Titin:Obscurin Complex Reveals a Conserved yet Specific Muscle M-Band Zipper Module. *Journal of Molecular Biology*. **427**, 718–736
40. R: The R Project for Statistical Computing [online] https://www.r-project.org/ (Accessed March 19, 2019)
41. Marston, S. (2017) Obscurin variants and inherited cardiomyopathies. *Biophys Rev*. **9**, 239–243
42. Bagnato, P., Barone, V., Giacomello, E., Rossi, D., and Sorrentino, V. (2003) Binding of an ankyrin-1 isoform to obscurin suggests a molecular link between the sarcoplasmic reticulum and myofibrils in striated muscles. *J Cell Biol*. **160**, 245–253
43. Chen, H., and Zhou, H.-X. (2005) Prediction of solvent accessibility and sites of deleterious mutations from protein sequence. *Nucleic Acids Res*. **33**, 3193–3199
44. Jmol: an open-source browser-based HTML5 viewer and stand-alone Java viewer for chemical structures in 3D [online] http://jmol.sourceforge.net/ (Accessed March 17, 2019)
45. Arbustini, E., Narula, N., Dec, G. W., Reddy, K. S., Greenberg, B., Kushwaha, S., Marwick, T., Pinney, S., Bellazzi, R., Favalli, V., Kramer, C., Roberts, R., Zoghbi, W. A., Bonow, R., Tavazzi, L., Fuster, V., and Narula, J. (2013) The MOGE(S) Classification for a Phenotype–Genotype Nomenclature of Cardiomyopathy: Endorsed by the World Heart Federation. *Journal of the American College of Cardiology*. **62**, 2046–2072
46. Westphal, J. G., Rigopoulos, A. G., Bakogiannis, C., Ludwig, S. E., Mavrogeni, S., Bigalke, B., Doenst, T., Pauschinger, M., Tschöpe, C., Schulze, P. C., and Noutsias,

M. (2017) The MOGE(S) classification for cardiomyopathies: current status and future outlook. *Heart Fail Rev*. **22**, 743–752
47. Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., Cuche, B. A., de Castro, E., Lachaize, C., Langendijk-Genevaux, P. S., and Sigrist, C. J. A. (2008) The 20 years of PROSITE. *Nucleic Acids Res*. **36**, D245–D249
48. Catalano, J., Paynton, B., Kaniper, S., Gerhard, G., and Alvarez, R. (2018) Identification of a Novel Obscurin Protein Variant in Nonischemic Cardiomyopathy. *Journal of the American College of Cardiology*. **71**, A743
49. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000) The Protein Data Bank. *Nucleic Acids Res*. **28**, 235–242
50. *The PyMOL Molecular Graphics System, Version 2.3 Schrödinger, LLC.*

Figures and Tables

Figure 1

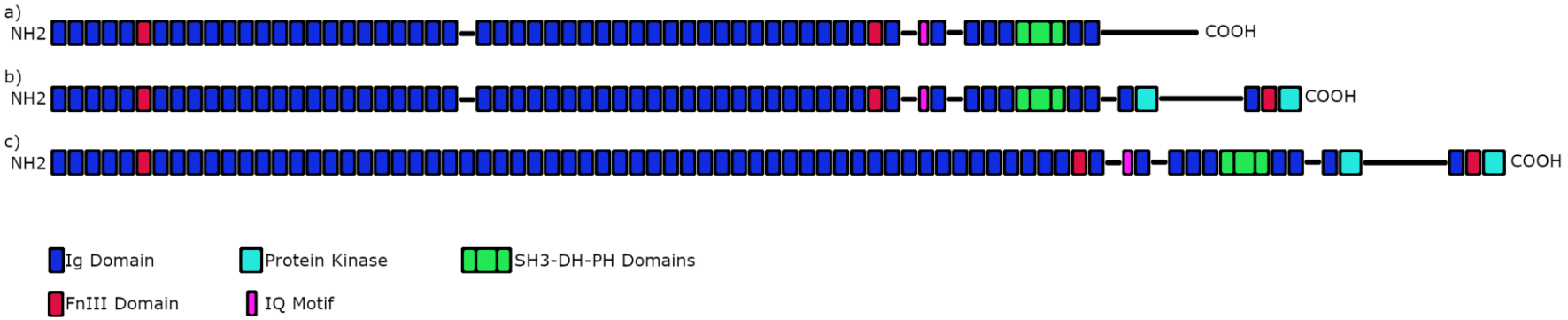

Figure 1 – Current domain architectures for 3 major obscurin isoforms; a) Obscurin-a, also known as isoform 3, was the first discovered isoform of obscurin and has a long disordered region at the COOH terminus where other major proteins such as titin and ankyrin isoforms; b) Obscurin-b, also known as isoform 1, is the current longest known isoform and is almost identical to obscurin-a bar the COOH region, which includes 2 additional Ig and 1 additional FnIII domains, and 2 serine-threonine protein kinase domains; c) Obscurin-IC, the inferred complete isoform of obscurin which includes all exons from the genetic sequence and the conserved domains derived thereof.

Figure 2

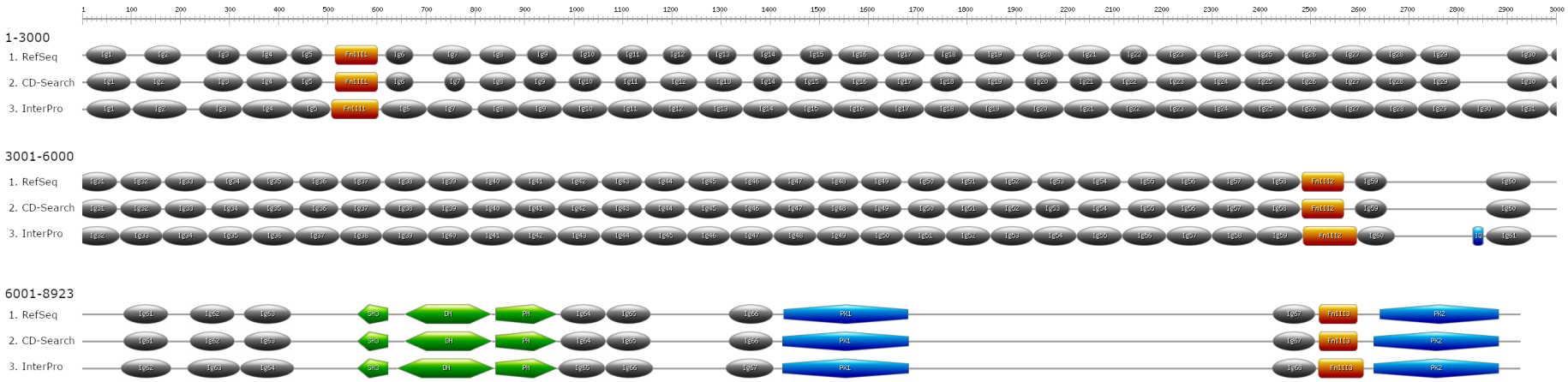

Figure 2 – Proposed domain architectures for obscurin-IC from RefSeq (RefSeq transcript NP_001258152.2), CD-Search (conserved domains from NCBI CD-Search), InterPro (conserved domains from InterPro 72.0). Architectures are largely similar but InterPro returns larger domains with very little or no interdomain space in between, particularly for the Ig domains, whereas RefSeq and CD-Search return slightly smaller domains with larger interdomain space. This image was created using PROSITE(47) MyDomains Image Creator.

Table 1

Obscurin-B (UniProt Q5VST9)			Obscurin-IC (RefSeq NP_001258152)			New Domain Architecture		
Domain	Start	End	Domain	Start	End	Domain	Start	End
Ig1	10	100	Ig1	9	91	Ig1	10	99
Ig2	110	202	Ig2	127	201	Ig2	110	201
Ig3	236	322	Ig3	252	321	Ig3	248	328
Ig4	331	414	Ig4	335	418	Ig4	335	418
Ig5	420	508	Ig5	425	489	Ig5	425	489
FnIII 1	515	612	FnIII 1	515	602	FnIII 1	515	602
Ig6	619	698	Ig6	617	673	Ig6	617	673
Ig7	701	790	Ig7	714	792	Ig7	714	792
Ig8	798	884	Ig8	809	883	Ig8	809	883
			Ig9	905	966	Ig9	898	964
Ig9	886	977	Ig10	997	1056	Ig10	990	1056
Ig10	978	1066	Ig11	1089	1150	Ig11	1084	1148
Ig11	1070	1161	Ig12	1181	1240	Ig12	1176	1251
Ig12	1162	1252	Ig13	1273	1332	Ig13	1268	1343
Ig13	1254	1345	Ig14	1365	1424	Ig14	1365	1424
Ig14	1346	1432	Ig15	1460	1527	Ig15	1450	1516
Ig15	1438	1524	Ig16	1539	1622	Ig16	1542	1619
			Ig17	1631	1714	Ig17	1634	1711
Ig16	1530	1621	Ig18	1733	1792	Ig18	1725	1792
Ig17	1622	1719	Ig19	1815	1898	Ig19	1818	1895
			Ig20	1914	1997	Ig20	1917	1983
			Ig21	2006	2091	Ig21	2009	2075
FnIII 2	1731	1808	Ig22	2111	2168	Ig22	2091	2168
Ig18	1809	1894	Ig23	2185	2268	Ig23	2185	2268
Ig19	1896	1982	Ig24	2275	2358	Ig24	2275	2358
Ig20	1987	2071	Ig25	2365	2447	Ig25	2365	2447
Ig21	2077	2162	Ig26	2453	2536	Ig26	2453	2536
Ig22	2165	2249	Ig27	2542	2625	Ig27	2542	2625
			Ig28	2631	2714	Ig28	2631	2714
Ig23	2289	2380	Ig29	2722	2803	Ig29	2722	2803
						Ig30	2806	2896
Ig24	2468	2559	Ig30	2899	2982	Ig31	2899	2982
Ig25	2564	2643	Ig31	2989	3071	Ig32	2989	3071
Ig26	2646	2730	Ig32	3077	3160	Ig33	3077	3160
Ig27	2736	2823	Ig33	3167	3251	Ig34	3167	3251
Ig28	2826	2908	Ig34	3267	3342	Ig35	3263	3340
Ig29	2920	2999	Ig35	3347	3429	Ig36	3347	3429
Ig30	3003	3092	Ig36	3441	3520	Ig37	3441	3520
Ig31	3095	3183	Ig37	3526	3606	Ig38	3526	3606
Ig32	3184	3268	Ig38	3615	3698	Ig39	3615	3698
Ig33	3273	3356	Ig39	3704	3786	Ig40	3704	3786
Ig34	3359	3444	Ig40	3792	3874	Ig41	3792	3874
Ig35	3449	3532	Ig41	3880	3962	Ig42	3880	3962
Ig36	3537	3620	Ig42	3968	4050	Ig43	3968	4050
Ig37	3625	3708	Ig43	4056	4138	Ig44	4056	4138
Ig38	3713	3796	Ig44	4144	4226	Ig45	4144	4226
Ig39	3801	3884	Ig45	4232	4314	Ig46	4232	4314
			Ig46	4320	4402	Ig47	4320	4402
			Ig47	4408	4490	Ig48	4408	4490
			Ig48	4496	4578	Ig49	4496	4578
			Ig49	4584	4666	Ig50	4584	4666
			Ig50	4679	4754	Ig51	4679	4754
			Ig51	4760	4842	Ig52	4760	4842
Ig40	3890	3973	Ig52	4847	4931	Ig53	4847	4931
Ig41	3978	4062	Ig53	4943	5020	Ig54	4937	5007
Ig42	4068	4160	Ig54	5025	5111	Ig55	5025	5111
Ig43	4171	4239	Ig55	5126	5202	Ig56	5126	5202
Ig44	4248	4337	Ig56	5205	5293	Ig57	5205	5293
Ig45	4340	4427	Ig57	5299	5384	Ig58	5299	5384
Ig46	4430	4518	Ig58	5390	5476	Ig59	5390	5476
FnIII 3	4525	4619	FnIII 2	5480	5565	FnIII 2	5480	5565
Ig47	4624	4714	Ig59	5587	5652	Ig60	5587	5652
IQ	4872	4901				IQ	5828	5850
Ig48	4898	4989	Ig60	5855	5945	Ig61	5855	5945
Ig49	5126	5215	Ig61	6083	6173	Ig62	6083	6173
Ig50	5260	5349	Ig62	6217	6307	Ig63	6217	6307
Ig51	5371	5467	Ig63	6328	6423	Ig64	6328	6423
SH3	5600	5667	SH3	6559	6621	SH3	6559	6621
DH	5693	5877	DH	6654	6830	DH	6654	6832
PH	5895	6004	PH	6839	6963	PH	6839	6963
Ig52	6014	6097	Ig64	6971	7063	Ig65	6971	7063
Ig53	6108	6200	Ig65	7065	7154	Ig66	7065	7154
Ig54	6357	6445	Ig66	7314	7403	Ig67	7314	7403
Protein kinase 1	6468	6721	Protein Kinase 1	7425	7678	Protein Kinase 1	7422	7678
Ig55	7463	7552	Ig67	8420	8505	Ig68	8420	8505
FnIII 4	7557	7649	FnIII 3	8514	8592	FnIII 3	8514	8592
Protein kinase 2	7672	7924	Protein Kinase 2	8637	8879	Protein Kinase 2	8625	8879

Table 1 – Comparison of relative positions of domains in the different isoforms and architectures with the new architecture used in our resource.

MRC Doctoral Training Partnership in Biomedical Sciences

Table 2

Variant (obscurin-b)	Variant (obscurin-IC)	Domain (obscurin-b)	Domain (obscurin-IC)	dbSNP ref	Reference	Disease	CADD	PolyPhen-2	PROVEAN	SIFT	mCSM ΔG (kCal/mol)	SDM ΔG (kCal/mol)	DUET ΔG (kCal/mol)
E963K	E1055K	Ig9	Ig10	rs563579474	Marston *et al.* 2015(12)	Dilated Cardiomyopathy	22.6	0.999	-1.999	0.55	-0.128	-0.43	0.232
V2161D	V2536D	Ig21	Ig26	rs376888216	Marston *et al.* 2015	Dilated Cardiomyopathy	24.2	0.999	-4.795	0.00	-1.668	-0.32	-1.408
F2809V	F3238V	Ig27	Ig34	rs189169421	Marston *et al.* 2015	Dilated Cardiomyopathy	24.4	0.996	-5.111	0.86	-1.254	-1.65	-1.458
R4344Q	R5301Q	Ig45	Ig58	rs79023478	Arimura *et al.* 2007(10)	Hypertrophic Cardiomyopathy	21.2	0.037	-0.777	0.11	-0.212	-0.64	-0.215
R4444W	R5401W	Ig46	Ig59	rs369758958	Rossi *et al.* 2017(37)	Distal Muscular Dystrophy	22.7	0.985	-3.126	0.03	-0.366	0.2	-0.5
A4484T	A5441T	Ig46	Ig59	rs116557268	Arimura *et al.* 2007	Hypertrophic Cardiomyopathy	22.6	0.263	-1.544	0.85	-0.721	-0.98	-0.454
R4856H	R5813H	Ig47-IQ	Ig60-IQ	rs200599475	Marston *et al.* 2015	Dilated Cardiomyopathy	1.9	0.002	0.438	0.39			
R5215H	R6172H	Ig49	Ig62	rs375864261	Xu *et al.* 2015(11)	Hypertrophic Cardiomyopathy	28.3	1.000	-2.928	0.16	-0.757	0.2	-0.549
D5966N	D6923N	PH	PH	.	Marston *et al.* 2015	Dilated Cardiomyopathy	33.0	1.000	-3.865	0.00	0.078	-0.22	0.167
G7500R	G8457R	Ig55	Ig68	rs377038292	Xu *et al.* 2015	Hypertrophic Cardiomyopathy	26.6	0.997	-3.651	0.28	-0.232	-3.63	-0.639
W7910R	W8865R	Kin2	Kin2	.	Catalano *et al.* 2018(48)	Systolic Heart Failure / Hereditary Ataxia	34.0	1.000	-7.344	0.85	0.203	-0.09	0.136

Table 2 – Damaging mutations in obscurin from existing literature. Cut-off values for probably deleterious (red) are as follows: CADD > 30, PolyPhen-2 > 0.95, PROVEAN < -2.5, SIFT < 0.05. Values are otherwise colour-coded for distance from cut-off (orange, yellow, green in increasing certainty of benign state). Negative values for mCSM, SDM, DUET protein stability are destabilising; ΔG < -2.0 is considered highly destabilising (red). R5813H is in an interdomain region so lacks structure, and thus also lacks protein stability metrics.

Figure 3

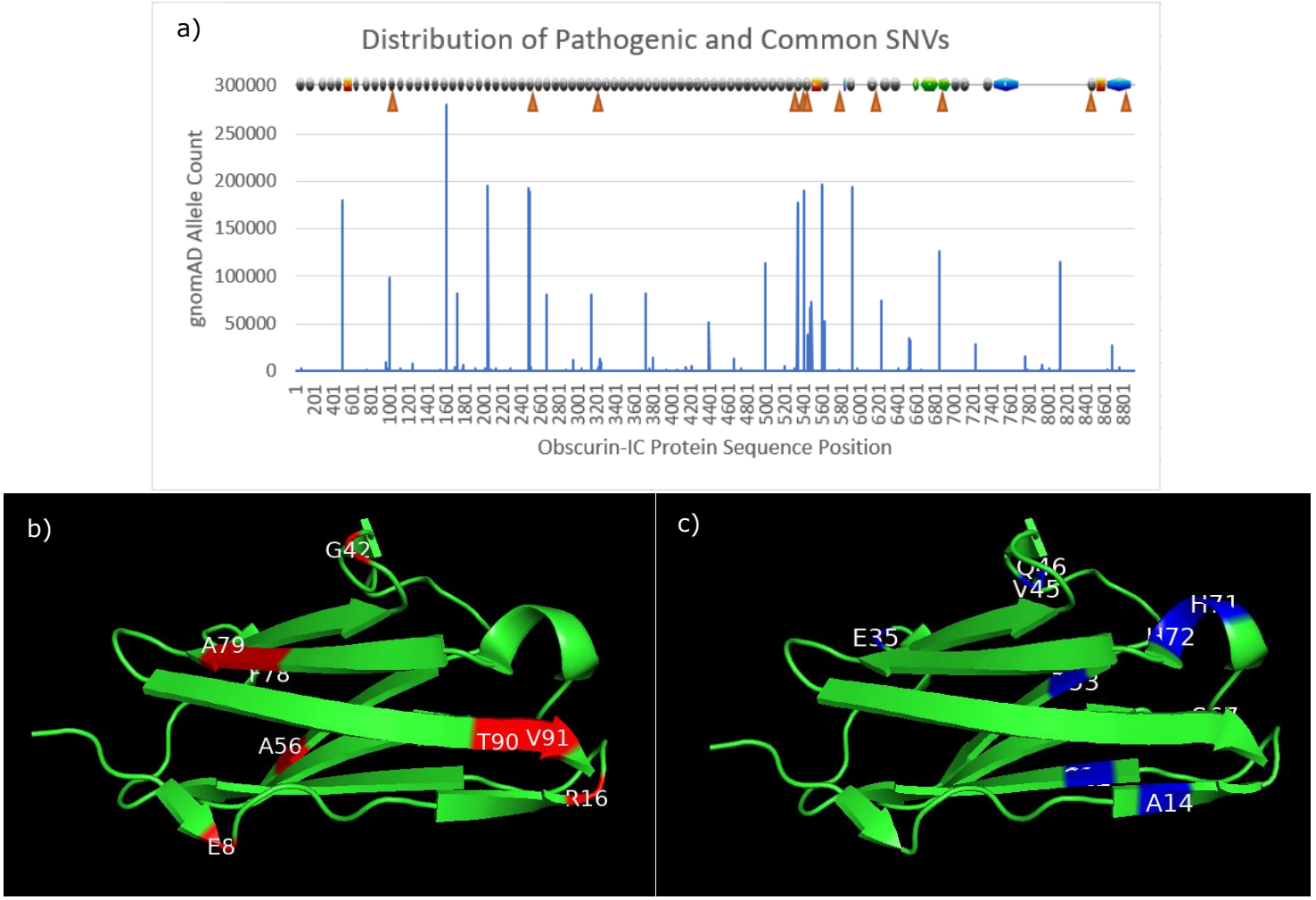

Figure 3 – Representations of the distribution of SNVs in obscurin; a) frequency of common SNVs from gnomAD database versus their position in obscurin-IC sequence, with positions of domains and known pathogenic SNVs for comparison; b) structural positions of all pathogenic SNVs on a representative structure (Ig59; PDB (49) accession 5TZM) in PyMol(50); c) structural positions of all SNVs with allele count > 100,000 on 5TZM in PyMol.

Figure 4

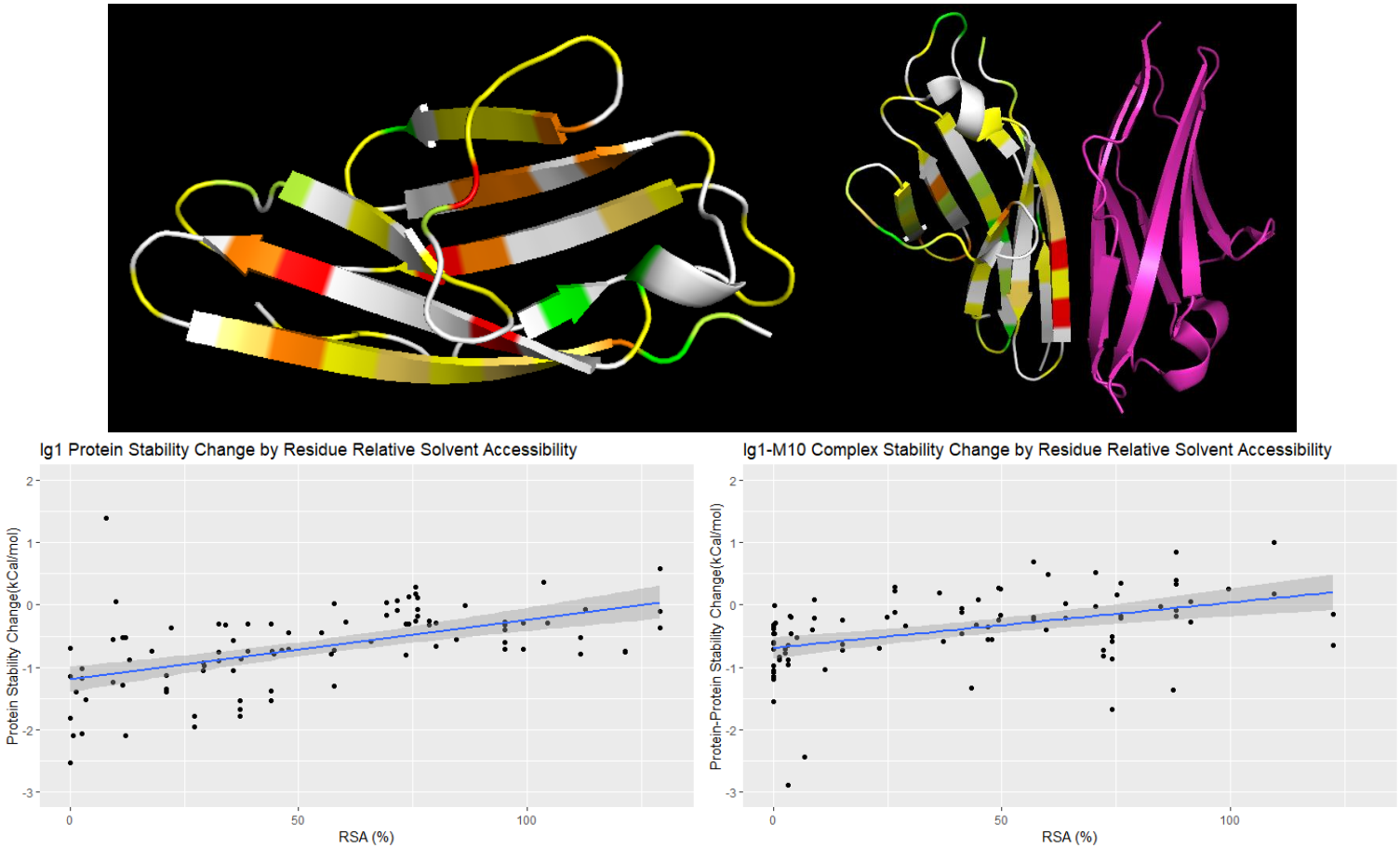

Figure 4 – Protein stability and solvent accessibility computation for Ig1; a) Ig1 (PDB accession 4C4K) annotated to indicate relative protein stability change of mutation at that locus in PyMol (green = stabilising, yellow = destabilising, red = highly destabilising, white = no data); b) Ig1-M10 complex (PDB accession 4C4K) annotated to indicate relative protein stability change of mutation at that locus in PyMol (magenta = titin M10); c) scatter plot of residue relative solvent accessibility against protein stability change for Ig1 variants and linear regression model; c) scatter plot of residue relative solvent accessibility against protein stability change for Ig1-M10 complex variants and linear regression model.

Table 3

	t	df	p	Standard Error	R^2
Ig1	6.0194	91	3.63e-08	4.920	0.2848
Ig1-M10	4.5456	91	1.681e-05	5.593	0.185

Table 3 – Statistics for linear regression models for Ig1 and Ig1-M10 complex. Correlation is considered significant where $p < 0.05$.