frontiers in Genetics

METHODS
published: 06 December 2018
doi: 10.3389/fgene.2018.00620

VarQ: A Tool for the Structural and Functional Analysis of Human Protein Variants

Leandro Radusky[1,2], Carlos Modenutti[1,2], Javier Delgado[3], Juan P. Bustamante[1,2,4], Sebastian Vishnopolska[1,2], Christina Kiel[3], Luis Serrano[3,5], Marcelo Marti[1,2] and Adrián Turjanski[1,2]**

[1] Departamento de Química Biológica Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Pabellón II de Ciudad Universitaria, Buenos Aires, Argentina, [2] Instituto de Química Biológica Facultad de Ciencias Exactas y Naturales (IQUIBICEN) CONICET, Pabellón II de Ciudad Universitaria, Buenos Aires, Argentina, [3] EMBL/CRG Systems Biology Research Unit, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology, Barcelona, Spain, [4] Facultad de Ingeniería de la Universidad Nacional de Entre Ríos, Oro Verde, Argentina, [5] Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain

OPEN ACCESS

Edited by:
Shameer Khader,
Northwell Health, United States

Reviewed by:
Harinder Singh,
J. Craig Venter Institute, United States
Sabeena Mustafa,
King Abdullah International Medical Research Center (KAIMRC), Saudi Arabia

***Correspondence:**
Marcelo Marti
marti.marcelo@gmail.com
Adrián Turjanski
aturjans@gmail.com

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: 17 May 2018
Accepted: 23 November 2018
Published: 06 December 2018

Citation:
Radusky L, Modenutti C, Delgado J, Bustamante JP, Vishnopolska S, Kiel C, Serrano L, Marti M and Turjanski A (2018) VarQ: A Tool for the Structural and Functional Analysis of Human Protein Variants. Front. Genet. 9:620. doi: 10.3389/fgene.2018.00620

Understanding the functional effect of Single Amino acid Substitutions (SAS), derived from the occurrence of single nucleotide variants (SNVs), and their relation to disease development is a major issue in clinical genomics. Despite the existence of several bioinformatic algorithms and servers that predict if a SAS is pathogenic or not, they give little or no information at all on the reasons for pathogenicity prediction and on the actual predicted effect of the SAS on the protein function. Moreover, few actual methods take into account structural information when available for automated analysis. Moreover, many of these algorithms are able to predict an effect that no necessarily translates directly into pathogenicity. VarQ is a bioinformatic pipeline that incorporates structural information for the detailed analysis and prediction of SAS effect on protein function. It is an online tool which uses UniProt id and automatically analyzes known and user provided SAS for their effect on protein activity, folding, aggregation and protein interactions, among others. We show that structural information, when available, can improve the SAS pathogenicity diagnosis and more important explain its causes. We show that VarQ is able to correctly reproduce previous analysis of RASopathies related mutations, saving extensive and time consuming manual curation. VarQ assessment was performed over a set of previously manually curated RASopathies (diseases that affects the RAS/MAPK signaling pathway) related variants, showing its ability to correctly predict the phenotypic outcome and its underlying cause. This resource is available online at http://varq.qb.fcen.uba.ar/. Supporting Information & Tutorials may be found in the webpage of the tool.

Keywords: variation diagnosis, bioinformatics, web server, single amino acid substitutions, single amino acid substitutions classification, FoldX

INTRODUCTION

The potential for genomics to contribute to clinical care has long been recognized, and the clinical use of information about a patient's genome is rising. In this sense, precision medicine initiatives for disease treatment and prevention that take into account individual variability in genes are being implemented worldwide. Single nucleotide variants (SNVs) that manifest as protein variants are the best known and most important form of variations in the genome, therefore a critical problem is to understand how single amino acid substitutions (SAS) affect protein function and protein interaction networks (Vidal et al., 2011; Hu et al., 2016) leading to disease.

There are a several SAS effect predictors, most of them perform a bioinformatic analysis and provide a pathogenicity outcome or score, i.e., the propensity of a variant to be involved on the development of a disease, something which is not straightforward to determine. They are usually based on sequence information, focusing on residue conservation and lacking a structural viewpoint which has proven to be critical to identify their effect (Kiel and Serrano, 2014). Moreover, even if they incorporate structural data (Bromberg and Rost, 2007; De Baets et al., 2012), usually they do not provide information that helps the user to understand the molecular mechanisms underlying their prediction, thus preventing a personal assessment. This black box prediction is possibly rooted in the fact that proper (structural) analysis of possible effects of a given variant is time-consuming and requires expert handling of different tools, thus preventing its wide applicability in clinical genomics. As filtering algorithms allow researchers to identify a handful of variants that are likely pathogenic, manual curation and careful analysis are usually needed. Furthermore, even if a variant has been identified, mainly because it has been previously associated with disease, its molecular effect on protein function could still be unknown.

Here, we present VarQ, a Web Server that automatically analyzes several effects of SAS on protein function, particularly protein folding, activity, protein-protein interactions, and drug or cofactor binding. Analyzed variants are either automatically extracted from clinical databases and/or can be submitted directly by the user. VarQ automatically selects a critical set of representative structural conformations of the desired protein and diagnoses variants effect based on their impact. For this sake several properties are computed for each variant, including involvement in ligand binding, presence in the catalytic site (Porter et al., 2004), presence in protein pockets using the Fpocket software (Schmidtke et al., 2010), the free energy change using FoldX (Schymkowitz et al., 2005), the residue sidechain exposure to solvent, presence in a protein-protein interface as identified in the corresponding protein-protein complex structure and/or in the 3did database (Stein et al., 2009), the conservation of each residue in the Pfam family (Bateman et al., 2004), the switchability propensity using abSwitch software (Diaz et al., 2014) and the aggregability factor using Tango software (Fernandez-Escamilla et al., 2004). VarQ is intuitive, user-friendly, and provides clinicians, biochemists, geneticist and all professionals involved in personalized medicine with a straightforward tool to annotate and analyze protein variants. A detailed description of the programs and databases used is given in the Web Site.

MATERIALS AND METHODS

Implementation of the Method as a Bioinformatic Pipeline

The VarQ tool is built around a collection of structural analysis tools tied together with the help of the workflow system Ruffus (Goodstadt, 2010). Having as unique input a list of UniProt (Magrane and UniProt Consortium, 2011) accession codes, the pipeline performs all the calculation steps described in **Figure 1**, parallelizing the acquisition of independent properties to improve its computational performance. Accession codes already computed are stored in a database to retrieve results without re-computing (this option works only for database stored substitutions and not for new user uploaded mutations).

A Web Server developed with the Bottle Python library[1] allows both to visualize the results and download them for further analysis. Some of the features available are the mapping of the found variants on the protein sequence (**Figure 2**) and structure using JSmol (Hanson et al., 2013).

Structure Mining

When the user specifies a UniProt accession, the VarQ Pipeline searches within all available structures mapped to the that UniProt within the PDB database, that cover different segments of the sequence, preferring those with highest coverage first and those with the best resolution in second place as a tiebreaker. Only those crystals covering at least 20 amino acids of the sequence of the target protein are considered. If there is more than one crystal structure covering the same part of the protein sequence, but coupling with different ligands, then both will be considered. If the same protein is crystallized in complex with another protein they will be considered separately and will appear with the information of the partner protein. The UniProt database is requested online to list all available crystal structures.

If the protein has not resolved structures in the PDB, an error message will be displayed in the web page together with 2 links to aid the user to search for possible solutions to this problem: one is the search of the UniProt accession in the RCSB web page and the second is the search of all crystallized UniProt accession codes that have the same gene name. No decisions are automatically taken at this point since no homology modeling is performed by the tool and sequence differences in homolog non human proteins could lead to unexpected outcomes that have to be considered by the user.

Variation Mining

For each UniProt accession analyzed, VarQ pipeline automatically mines online for known variants. Actually, there are several databases populated with variants coming from different sources: clinical trials (Landrum et al., 2016),

[1] bottlepy.org

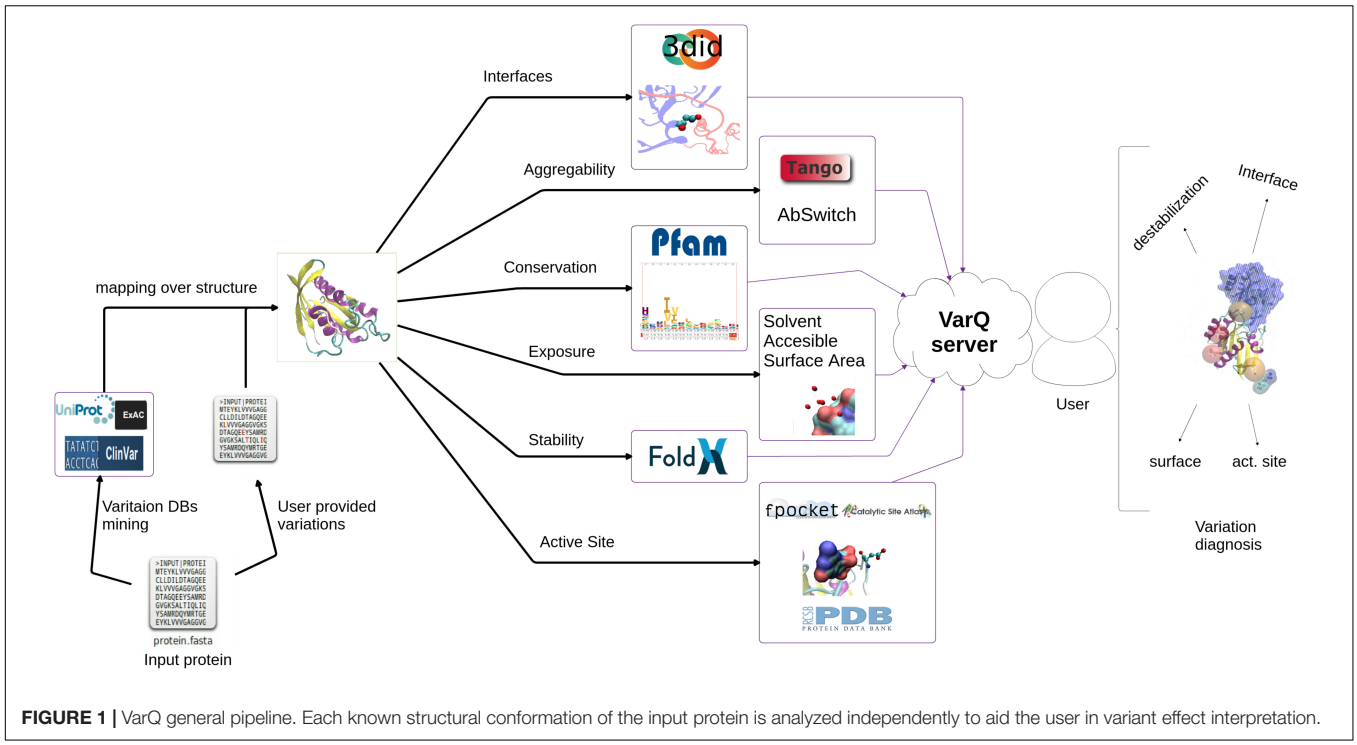

FIGURE 1 | VarQ general pipeline. Each known structural conformation of the input protein is analyzed independently to aid the user in variant effect interpretation.

sequencing information (Forbes et al., 2011), user submission (Day, 2010), etc. In this work, we are considering as a source of variants the following databases: (i) UniProt annotated variants coming from dbSNP and BioMuta databases. (ii) UniProt curated variants for human genes, UniProt database provides a database called humsavar (Famiglietti et al., 2014) that contains qualitative information classifying for all listed variants if they are disease-related or not and which is the disease involved in each mutation. (iii) ClinVar variants provide additional mutations coming from clinical studies. When a protein target is introduced to our pipeline, all this information is downloaded and the variations are kept for subsequent analysis.

Binding and Catalytic Residues

Each structure of the Protein Data Bank (PDB) provides as an annotation for each crystallized ligand (compound, ion, cofactor, etc.) A dataset of solvent molecules were built to consider as binding residues only those interacting with non-solvents as binding residues. The information is parsed directly by the pipeline and the corresponding residues are labeled as involved in binding.

Also, we used the FPocket software for each protein structure considered in the analysis to calculate the protein pockets. We only considered the pockets with Druggability Score greater than 0.5, and/or if a ligand is present in the pocket. All the residues belonging to the pocket (even those not in contact with the ligand) are considered as binding residues.

Catalytic Site Atlas (CSA) is a database documenting the active sites and catalytic residues of enzymes. CSA contains 2 types of entries: original manual-annotated entries, derived from the primary literature and homology-determined entries, based on sequence comparison methods to one of the original entries. All the residues belonging to the CSA database are labeled as catalytic residues in the VarQ pipeline.

Changes in Protein Stability

All mutations that were mapped to any protein structure were analyzed with the FoldX software. The FoldX software predicts the free energy change of a given mutation on the protein stability or on the stability of a protein-protein complex, and has been widely tested as an accurate predictor for point mutations without the need of more time-consuming techniques like Molecular Dynamics or Monte Carlo simulations (Guerois et al., 2002). Mutagenesis was performed using the BuildModel option of FoldX. The stabilities were calculated using the Stability command of FoldX, and $\Delta\Delta G$ values are computed by subtracting the energy of the WT from that of the mutant. The FoldX energy function includes terms that have been found to be important for protein stability. The equation describing the calculation of free energy of unfolding (ΔG) of a target protein is described in detail in the FoldX web server. Briefly, it consists of an empirical force field that estimates the difference in the Van der Waals contributions of the protein with respect to the solvent, the differences in solvation energy for apolar and polar groups, the free energy difference between hydrogen bond formation, the difference in electrostatic interactions and the change in entropy due to fixing of the backbone and the sidechain in the folded state. When protein-protein stability is estimated, the empirical force field also includes a term related to the change in translational and rotational entropy and a term that takes into account the role of electrostatic interactions on the association constant. Those mutations that

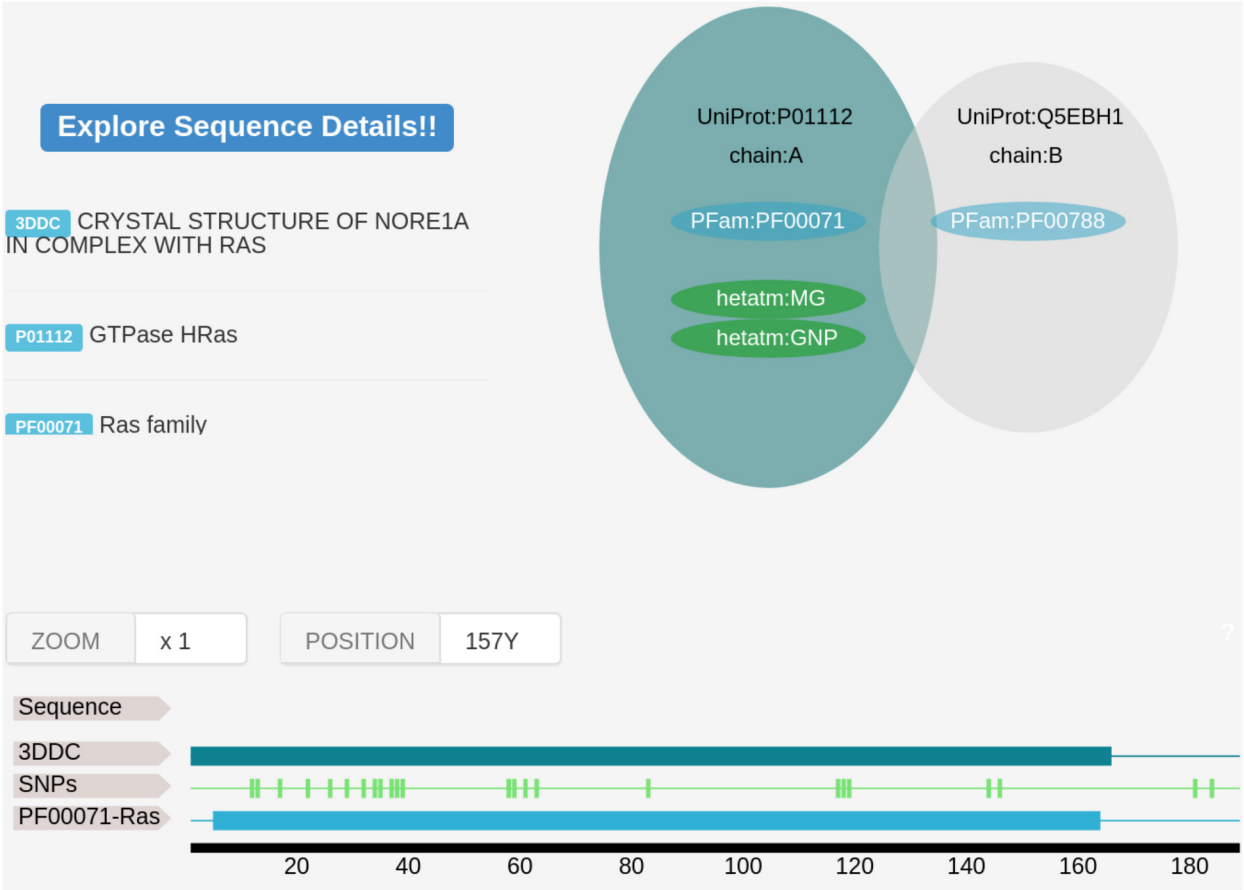

FIGURE 2 | An Example of VarQ output for the HRAS gene (PDB id 3DDC). On top left, we show the structural features derived from crystallographic data. Target gene and possible interacting genes are shown in green and gray respectively. Structural coverage, PFam family assignment and location of the variant are mapped over the UniProt original sequence.

have more than 2 kcal/Mol $\Delta\Delta G$ are labeled as "High $\Delta\Delta G$ variation."

Residue Exposure

For each residue in each structure, the Solvent Accessible Surface Area (SASA) was computed. Sidechain exposure percentages of each residue are informed, and those with more than 50% of its surface exposed are labeled as exposed, or on the contrary, as buried. For this computation, the PyMol software (DeLano, 2002) is used as a command line tool calling to the get_area feature. The glycine residues are never labeled as buried or exposed since always present a 0% of exposure because of its absence of a sidechain. In the structural window, the value is displayed and a horizontal bar chart is shown based on the total side chain surface.

Protein-Protein Interfaces

In order to detect those mutations that can affect protein-protein interaction, we decided to label the amino acids present in protein-protein interaction surfaces. To label a residue belonging to a protein-protein interface we evaluated the structures that have more than one chain. When an atom of a residue of one chain is at a distance of less than 5Å from an atom of the other chain the residue is labeled as interface residue. We also added an extra criteria; if the residue in the 3did database has been labeled as interacting we labeled the residue as 3did but we did not add the interface label. Despite the fact that the 3did database is not accurate enough for our automatic analysis pipeline, we decided to add this information since we considered it valuable to evaluate possible effects as it expands the database of residues which are involved in protein-protein interfaces, but needs manual inspection.

Other Properties

The B-factor of a residue usually relates to its mobility and we inform this value with respect to the median of the B-factor of all the residues in the protein. We also calculate the Switchability, which informs the tendency of residues to switch from alpha-helix to a beta sheet type of secondary structure; the Aggregability, which is the tendency of a residue to generate aggregation when is mutated and is computed using the Tango Software; the conservation, calculated as the value in bits of the

TABLE 1 | Comparison of mutation-mining of previous hand-curated work against the results given of VarQ Pipeline for benchmark set of RASopathies related proteins.

	Kiel and Serrano, 2014	VarQ
Total variants	956	1109
Mapped onto structure	624	566
Predicted to be neutral	197*	152
Predicted to be affect protein function	427*	414

**In Kiel and Serrano, 2014 variants were classified as neutral or disease causing based on FoldX energy score.*

TABLE 2 | Classification of the previous work (FoldX destabilizing/disease-causing mechanism know category) compared with the automatic classification applied to VarQ results.

Effect in	Kiel and Serrano, 2014	VarQ
Active site + Binding	200 (47%)	203 (49%)
Protein-protein interaction	79 (18%)	74 (18%)
Folding	145 (34%)	137 (33%)
Localization	3 (1%)	–
Total	427	414

Our method do not label any mutation as "localization" so we kept the label used by Kiel & Serrano. Number in parenthesis are the percentages of the total residues with that label.

corresponding letter of the original amino acid in the Pfam family only if it is assigned for the analyzed position. This value is computed by an in-house developed script.

RESULTS

The Web Server is developed to run proteins based on its UniProt ID. To test our developed pipeline we first run a manually curated set of variants derived from our previous work (Kiel and Serrano, 2014), comprising mutations in Ras/MAPK pathway components which can cause either cancer or developmental a group of disorders called RASopathies (Tidyman and Rauen, 2009). This previous case by case analysis required a considerable amount of manual curation and now constitutes an excellent benchmark for our web server capabilities. The results obtained with the newly developed pipeline (**Figure 1**) as implemented in the VarQ web server where all the calculations and decisions run fully automatic are presented in **Table 1**. VarQ automatically retrieved 566 variants, out of the 624 mutations identified manually and structurally mapped by Kiel and Serrano (**Table 1**).

In **Figure 2** an example of the VarQ output for one of the crystallographic configurations available for the HRAS gene (PDB id 3DDC) is shown. It depicts the solved crystallographic unit and the retrieved variations mapped into the protein sequence. PFam family information is also shown. Each structural configuration may produce for each variant a different outcome: for example, bounded-to-ligand and unbounded crystals of the same protein can figure lead to classify one position as "Active site" for only one crystal, leading to another outcome in the other crystal. To classify the variations on the analyzed data set, the outcomes were applied to label each variant following the criteria established in the columns of **Table 3**, from left to right (i.e., one variation labeled as "Active site" in one crystal is considered as Active site even if in other crystals this position is not in bounded with a ligand molecule).

Figure 3 shows the histogram of the $\Delta\Delta G$ energy, estimated using the FoldX software, for neutral and pathogenic SAS. As clearly evidenced, a value higher than 2 kcal/mol is a good indication that the SAS and thus, the underlying mutation is likely to be pathogenic, However, for lower $\Delta\Delta G$ energy values, other properties need to be considered to be able to define potential pathogenicity, and also to understand the underlying molecular mechanism leading to disease.

VarQ pipeline, further classifies all the amino acids of each protein automatically, in key categories related to their function (being part of the active site, of protein-protein interaction surfaces, etc. see section "Materials and Methods" for details) and presents that information to help decision making (**Table 3**). For example, we were able to identified 94% of the manual curated protein-protein interaction variants; also all variants with high switchability and aggregability, which were correctly classified as folding disruptors.

The active site and binding residues are those either labeled as ligand-binding residues in the PDB file, those belonging

TABLE 3 | Decision tree to propose a possible effect of mutation in an analyzed protein.

ΔΔG	Active site	Interface	Buried	High aggregability	High switchability	Diagnostic	Website Outcome Label
	✓					Protein activity altered	Disrupt protein function
	×	✓				Protein-protein interaction affected	Disrupt protein interface
↑	×	×	✓			Destabilization of domain preventing folding	Disrupt protein folding
	×	×	×		✓	Destabilization of domain preventing folding	Disrupt protein structure
↓	✓					Potential protein activity altered	Potential disruption protein function
	×	✓				Potential protein-protein interaction affected	Potential disruption protein interface
	×	×	✓			Potential destabilization of domain preventing folding	Potential disruption protein folding
	×	×	×		✓	Potential destabilization of domain preventing folding	Potential disruption protein structure
↓	×	×	×	×	×	No effect	Likely non pathogenic mutation

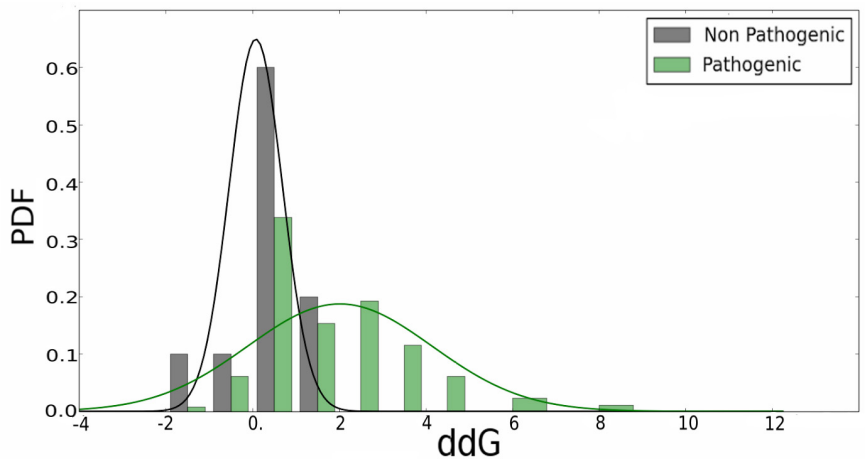

FIGURE 3 | Change in folding energy ($\Delta\Delta G$) computed by the FoldX software for all mined variants in analyzed RASopathies related proteins. Pathogenic and non-pathogenic variants are shown separately.

to the same pocket of these residues as determined by the Fpocket program, and those belonging to the CSA database. For example, there are three HRAS structures, one bound to GNP (i.e., PDB id 3DDC), other to GTP (i.e., PDB id 4K81) and another one with no ligand (i.e., PDB id 6AXG), and reported mutations in the binding pocket have been proposed to modify signaling cascades, which are involved in different diseases (Prior et al., 2012). VarQ properly labeled these residues as "active site residues" in each coupled structure indicating the correct when corresponds, thus allowing the user to rationalize SAS effects in those signaling pathways. For the Glycine 12 the Clinvar database reports several disease associated mutations, each one presenting different energies according to the pose of the ligand, if present, and subsequently to different outcomes that have to be pondered by the user.

Folding residues are those that correspond to neither inter-domain nor Active Site having an accessible surface area percentage lower than 50%. In the particular case of the RASopathies proteins, a high proportion of the residues are marked as interface-involved because these proteins participate in a very complex network with a large set of protein-protein interactions. Localization mutations (a total of 3, see **Table 2**) are not included in our analysis.

To summarize the value of the method, we were able to automatically identify and predict as significantly affecting protein function 414 ($\approx$ 97% efficiency) of the 427 pathogenic variants in the original data set, showing that structural information can enrich the knowledge leading to SAS pathogenicity prediction (see **Figure 4**).

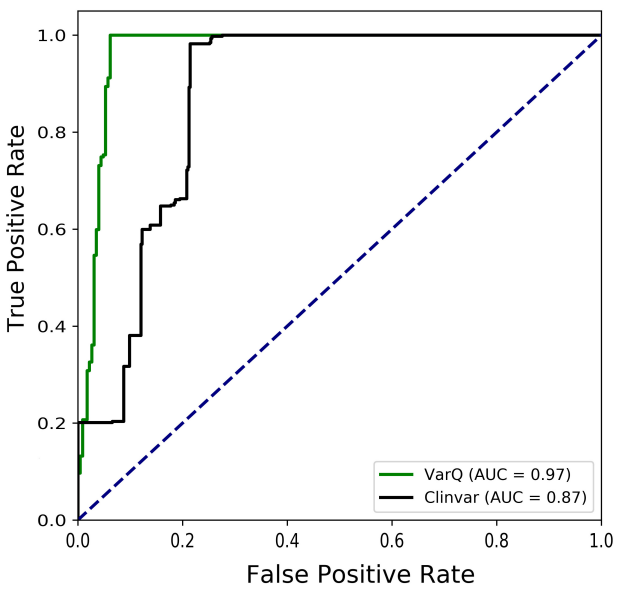

FIGURE 4 | Receiver Operating Characteristic Curves for the diagnosis of Kiel & Serrano manually curated pathogenic variants that were mapped into structure. Pathogenic mutations are considered true positives in Clinvar (black curve) data set when they are labeled as "pathogenic" or "likely pathogenic." For VarQ (green curve) true positives are considered for all the labels except "no effect".

DISCUSSION

In this work we introduce VarQ, a novel online tool that offers an user friendly way to evaluate the effect of protein variants that arise from human genomics projects. We have shown that VarQ is able to correctly reproduce previous analysis of RASopathies related mutations avoiding extensive and time consuming manual curation. Also, we were able to assign 153 novel mutations that cannot be compared with the previous study, a detailed outcome of this mutations is provided in **Supplementary Table S1**.

Users can apply VarQ to either analyze mutations that are already available in clinical databases or to analyze novel unreported variants. To assist variation diagnosis each analyzed

mutation is labeled according to several computed properties. For different conformations of the same protein (i.e., active and inactive determined structures) the same mutation could lead to different labeling, leaving to the user the final assessment of the effect caused by the variation. VarQ displays its full potential on human proteins with known structure(s) but can also be used with any protein. Usually clinical researchers, biochemists and geneticists do not have the bioinformatic resources to massively analyze variants so they use web servers or easy to use softwares to classify the variants. On the other hand, bioinformatic softwares usually are aimed at predicting only pathogenicity but do not give information for the users to be able to gain insight and finally decide the possible/potential effect of the mutation. Structural information is often not available for reported mutations: for the RASopathies analyzed proteins, a highly represented set on the PDB, only 20.3% of the Clinvar mutations for this proteins could be mapped into structure, 2103 vs. 427. When available, this information contributes to our understanding of the molecular basis of disease (see **Figure 4**). This information is of paramount importance when pathogenic prediction is challenging. In this sense, for example a mutation that disrupts a protein-protein interaction may be pathogenic or could be benign depending on the protein function and the biology of the interaction. To the best of our knowledge this is the first application that provides this information in an automatic, simple and intuitive way.

DATA AVAILABILITY STATEMENT

The datasets generated and analyzed for this study can be found in the web page of VarQ (http://varq.qb.fcen.uba.ar/).

AUTHOR CONTRIBUTIONS

CK, LS, MM, and AT conceived the work. LR designed and programmed the pipeline with the support of CM, JD, JB, and SV. LR programmed the website with the support of CM, JD, JB, and SV. LR, CM, and CK analyzed the data. LR, CM, MM, and AT wrote the manuscript with support of JD, CK, and LS. All authors approved the final manuscript.

FUNDING

The research leading to these results has received funding from the European Union Seventh Framework Program (FP7/2007-2013) under grant agreement Nr. PRIMES_278568. This work was supported by the Spanish Ministerio de Economía y Competitividad, Plan Nacional BIO2012-39754 and the European Fund for Regional Development. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. This work has been supported by grant PIP1220110100850 awarded to MM, and by PICT-2010-2805 awarded to AT. This project has received funding from the Marie Curie International Research Staff Exchange Scheme within the 7th European Community Framework Program under grant agreement no. 612583-DEANN. CABANA is funded by the BBSRC under the The Global Challenges Research Fund (GCRF) Growing Research Capability call, contract number BB/P027849/1 (www.ukri.org/research/global-challenges-research-fund/funded-projects/).

ACKNOWLEDGMENTS

A preprint version of this article has been deposited in bioRxiv (Radusky et al., 2018).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2018.00620/full#supplementary-material

TABLE S1 | Novel mutations found by VarQ not analyzed by Kiel and Serrano (2014).

REFERENCES

Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., et al. (2004). The Pfam protein families database. *Nucleic Acids Res.* 32, D138–D141.

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

Day, I. N. M. (2010). dbSNP in the detail and copy number complexities. *Hum. Mutat.* 31, 2–4. doi: 10.1002/humu.21149

De Baets, G., Van Durme, J., Reumers, J., Maurer-Stroh, S., Vanhee, P., Dopazo, J., et al. (2012). SNPeffect 4.0: on-line prediction of molecular and structural effects of protein-coding variants. *Nucleic Acids Res.* 40, D935–D939. doi: 10.1093/nar/gkr996

DeLano, L. W. (2002). *The PyMOL Molecular Graphics System*. Available at: http://pymol.org

Diaz, C., Corentin, H., Thierry, V., Chantal, A., Tanguy, B., David, S., et al. (2014). Virtual screening on an α-helix to β-strand switchable region of the FGFR2 extracellular domain revealed positive and negative modulators. *Proteins* 82, 2982–2997. doi: 10.1002/prot.24657

Famiglietti, M. L., Estreicher, A., Gos, A., Bolleman, J., Géhant, S., Breuza, L., et al. (2014). Genetic variations and diseases in UniProtKB/swiss-prot: the ins and outs of expert manual curation. *Hum. Mutat.* 35, 927–935. doi: 10.1002/humu.22594

Fernandez-Escamilla, A.-M., Rousseau, F., Schymkowitz, J., and Serrano, L. (2004). Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat. Biotechnol.* 22, 1302–1306.

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

Goodstadt, L. (2010). Ruffus: a lightweight python library for computational pipelines. *Bioinformatics* 26, 2778–2779. doi: 10.1093/bioinformatics/btq524

Guerois, R., Nielsen, J. E., and Serrano, L. (2002). Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.* 320, 369–387.

Hanson, R. M., Prilusky, J., Renjian, Z., Nakane, T., and Sussman, J. L. (2013). JSmol and the next-generation web-based representation of 3D Molecular structure as applied to proteopedia. *Isr. J. Chem.* 53, 207–216.

Hu, J. X., Thomas, C. E., and Brunak, S. (2016). Network biology concepts in complex disease comorbidities. *Nat. Rev. Genet.* 17, 615–629. doi: 10.1038/nrg.2016.87

Kiel, C., and Serrano, L. (2014). Structure-energy-based predictions and network modelling of RASopathy and cancer missense mutations. *Mol. Syst. Biol.* 10:727. doi: 10.1002/msb.20145092

Landrum, M. J., Lee, J. M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., et al. (2016). ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* 44, D862–D868. doi: 10.1093/nar/gkv1222

Magrane, M., and UniProt Consortium. (2011). UniProt Knowledgebase: a hub of integrated protein data. *Database* 2011:bar009. doi: 10.1093/database/bar009

Porter, C. T., Bartlett, G. J., and Thornton, J. M. (2004). The Catalytic site atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res.* 32, D129–D133.

Prior, I. A., Lewis, P. D., and Mattos, C. (2012). A comprehensive survey of Ras mutations in cancer. *Cancer Res.* 72, 2457–2467. doi: 10.1158/0008-5472.CAN-11-2612

Radusky, L., Modenutti, C. P., Delgado, J., Bustamante, J. P., Vishnopolska, S., Kiel, C., et al. (2018). VarQ: a tool for the structural analysis of human protein variants. *bioRxiv* [Preprint]. doi: 10.1101/277491

Schmidtke, P., Le Guilloux, V., Maupetit, J., and Tufféry, P. (2010). fpocket: online tools for protein ensemble pocket detection and tracking. *Nucleic Acids Res.* 38, W582–W589. doi: 10.1093/nar/gkq383

Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., and Serrano, L. (2005). The FoldX web server: an online force field. *Nucleic Acids Res.* 33, W382–W388.

Stein, A., Panjkovich, A., and Aloy, P. (2009). 3did Update: domain-domain and peptide-mediated interactions of known 3D structure. *Nucleic Acids Res.* 37, D300–D304. doi: 10.1093/nar/gkn690

Tidyman, W. E., and Rauen, K. A. (2009). The RASopathies: developmental syndromes of Ras/MAPK pathway dysregulation. *Curr. Opin. Genet. Dev.* 19, 230–236. doi: 10.1016/j.gde.2009.04.001

Vidal, M., Cusick, M. E., and Barabási, A.-L. (2011). Interactome networks and human disease. *Cell* 144, 986–998.

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.