

An integrated computational approach can classify *VHL* missense mutations according to risk of clear cell renal carcinoma

Lucy Gossage^{5,†,*}, Douglas E. V. Pires^{2,†}, Álvaro Olivera-Nappa^{2,3,†}, Juan Asenjo³, Mark Bycroft⁴, Tom L. Blundell² and Tim Eisen¹

¹Department of Oncology, Cambridge University Hospitals NHS Foundation Trust, Box 193 (R4) Addenbrooke's Hospital, Cambridge Biomedical Campus, Hill's Road, Cambridge CB2 0QQ, UK, ²Department of Biochemistry, University of Cambridge, Tennis Court Road, Cambridge CB2 1GA, UK, ³Centre for Biochemical Engineering and Biotechnology, University of Chile, Beauchef 850, Santiago, Chile, ⁴MRC Laboratory of Molecular Biology, Francis Crick Avenue, Cambridge Biomedical Research Centre, Cambridge CB2 0QH, UK and ⁵Cancer Research UK Cambridge Institute, University of Cambridge, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, UK

Received May 25, 2014; Revised May 25, 2014; Accepted June 17, 2014

Mutations in the *von Hippel–Lindau (VHL)* gene are pathogenic in VHL disease, congenital polycythaemia and clear cell renal carcinoma (ccRCC). pVHL forms a ternary complex with elongin C and elongin B, critical for pVHL stability and function, which interacts with Cullin-2 and RING-box protein 1 to target hypoxia-inducible factor for polyubiquitination and proteasomal degradation. We describe a comprehensive database of missense *VHL* mutations linked to experimental and clinical data. We use predictions from *in silico* tools to link the functional effects of missense *VHL* mutations to phenotype. The risk of ccRCC in VHL disease is linked to the degree of destabilization resulting from missense mutations. An optimized binary classification system (symphony), which integrates predictions from five *in silico* methods, can predict the risk of ccRCC associated with *VHL* missense mutations with high sensitivity and specificity. We use symphony to generate predictions for risk of ccRCC for all possible *VHL* missense mutations and present these predictions, in association with clinical and experimental data, in a publically available, searchable web server.

INTRODUCTION

von Hippel–Lindau (VHL) disease is an autosomal dominant syndrome associated with multiple tumours including retinal and central nervous system (CNS) haemangioblastoma, clear cell renal carcinoma (ccRCC) and pheochromocytoma (PCC), which results from mutations in the *VHL* gene (reviewed in 1). Over 1000 *VHL* mutations including >900 *VHL* kindreds are documented (2,3). Fifty-two percent of *VHL* disease mutations are missense (3), which are broadly distributed throughout the gene. In addition, inheritance of certain *VHL* mutations in an autosomal recessive fashion, with either homozygous or compound heterozygous alleles, can lead to congenital polycythaemias (4–12). Germline *VHL* mutations also account for up to

50% of patients with apparently isolated familial PCC and 11% of patients with an apparently sporadic PCC (13,14).

The canonical *VHL* protein product, pVHL isoform 1 (pVHL₃₀), has two structurally different domains: an N-terminal 53 amino acid disordered domain not needed for tumour suppression and a C-terminal ordered domain consisting of an α -helical domain (residues 155–192) and a mainly β -sheet domain (residues 63–154 and 193–204). pVHL forms a ternary complex with the elongin C and elongin B proteins (15–17) (henceforth VCB complex) that is critical for pVHL stability (18) and function. Mutations that affect pVHL-binding residues in elongin C have been described in ccRCC (19), supporting the hypothesis that the tumorigenic effects of *VHL* mutations relate to dysfunction of the VCB complex. Thus, the entire VCB complex should

*To whom correspondence should be addressed. Email: lucy.gossage@doctors.org.uk; lucy.gossage@cruk.cam.ac.uk

†These authors contributed equally.

be considered a single entity when assessing the structural and functional effects of *VHL* mutations. The VCB complex nucleates a complex containing Cullin-2 and RING-box protein 1 (16,20–23), which targets prolyl-hydroxylated hypoxia-inducible factors (HIFs) for polyubiquitination and proteasomal degradation (24,25). pVHL also has less well-characterized HIF-independent functions.

VHL disease is classified into Type 1 or Type 2 depending on the presence of PCC. In Type 1 disease the risk of PCC is low. In Type 2 disease, accounting for up to 20% of *VHL* kindreds is subdivided into: (2A) PCCs and other typical VHL disease manifestations but low risk of ccRCC, (2B) PCCs and other typical VHL disease manifestations including ccRCC and (2C) PCCs only. A major limitation of this classification is that, due to the variability of expression in VHL disease, accurate classification can only be made in large kindreds. Furthermore, its use in assisting clinical management is limited since a family may move from one subtype to another. Most patients with truncating mutations or exon deletions have Type 1 disease while kindreds with Type 2 disease usually have a missense mutation.

Experimental data support diverse effects for missense *VHL* mutations on both HIF-dependent and HIF-independent pVHL functions (Supplementary Material, Table S1). *In vitro* modelling of naturally occurring mutations suggests (1) a correlation between the risk of haemangioblastoma and the ability of a *VHL* mutation to impair HIF regulation and (2) the risk of developing ccRCC in VHL disease is linked to the degree to which HIF activity is compromised (26,27). In contrast, certain Type 2C VHL disease mutations retain their ability to downregulate HIF α (26,27). Nonsense and frameshift mutations may have a higher risk of ccRCC and haemangioblastomas than missense mutations (28–30). Allelic heterogeneity and genetic modifiers may influence the phenotypic variability of VHL disease (31–33).

Somatic biallelic inactivation of *VHL* also occurs in the majority of sporadic ccRCCs (34–37). Nearly 250 different missense mutations (32%) have been described in sporadic ccRCC (38). Numerous studies have investigated with conflicting results whether functional loss of *VHL* or the type of *VHL* mutation may influence prognosis in ccRCCs (reviewed in 39).

Several different computational approaches to study and predict the effects of missense mutations on protein structures have been proposed (40–47). The methods require either sequence or structural information and present different limitations, their performance depending on the impact on structural stability or intermolecular interactions of the mutation; methods are often complementary in approach to others (43) implying that overall prediction quality could be improved by combining different computational methods. pVHL poses several unusual challenges to computational models: (i) it forms part of a multi-subunit complex where folding is concerted with assembly, (ii) the inter-subunit contacts predominantly involve hydrogen bonding and (iii) it has a small hydrophobic core and a significant portion of the stabilization comes from hydrogen-bonding interactions (20).

Here we describe an integrated computational approach, built upon a comprehensive database of missense *VHL* mutations linked to experimental and clinical data. We present an optimized predictive model (named symphony) that integrates predictions from *in silico* models. Predictions both link the functional effects of missense *VHL* mutations to phenotype and classify ccRCC

risk with high sensitivity and specificity. Our observations emphasize: first, the importance of structural knowledge to delineate mechanisms of disease; subtle structural and functional changes resulting from missense mutations through multiple mechanisms can be related to different phenotypes in VHL disease. Secondly, they highlight the success of combining diverse yet complementary computational approaches to obtain a robust disease predictor for complex proteins such as pVHL. We use symphony to generate predictions for risk of ccRCC for all possible *VHL* mutations and present these predictions, in association with clinical and experimental data, in a publically available, searchable website.

RESULTS

Development of an integrated *in silico* workflow

We first constructed a comprehensive database of mutations in *VHL* annotated with the available experimental and phenotypic data (Supplementary Material, Table S2A). We established an *in silico* workflow (i) to predict the quantitative impact of mutations on stability and affinity of the VCB complex and (ii) to correlate mutation effects with risk of ccRCC. To achieve the latter, the predictions together with the compiled database (Supplementary Material, Table S2A) were used as evidence to train and test a binary classifier (which we named symphony) that outputs the predicted risk of ccRCC according to Figure 1.

We developed new computational strategies that predict changes in protein stability and protein–protein affinity. We have previously shown that combining computational methods based on different protein descriptors can lead to a predictor that performs better overall (43). Here we combine five *in silico* methods, which consider different information regarding short-/long-range structural ordering, side-chain interactions and stability, evolutionary conservation of physicochemical properties and protein–protein interactions. Mutation Cutoff Scanning Matrix (mCSM) (43) (<http://structure.bioc.cam.ac.uk/mcsm>) is based on the cutoff scanning matrix (CSM) concept and relies on graph-based structural signatures (50,51). It is a protein structural signature originally proposed and successfully used in protein function prediction and structural classification tasks; it has been recently extended and applied to large-scale receptor-based protein ligand prediction. Site-directed mutator (SDM; <http://mordred.bioc.cam.ac.uk/~sdm/sdm.php>) (44,45) uses knowledge of structures of proteins where amino acid replacements are tolerated within families of homologues over evolutionary time. MOSST predictions (<http://www.biomedcentral.com/1471-2105/12/122/additional>, last accessed on 19 July 2014) are based on evolutionary and functional information obtained from conservation rules of physicochemical properties of amino acids in a protein family (42). PoPMuSiC (40) (<http://http://babylone.ulb.ac.be/popmusic/>) relies on statistical potentials to represent different protein descriptors and elucidate correlations between them. This has been adapted as a predictor of binding free energy changes in protein–protein complexes due to single-point mutations in the BeAtMuSiC method (<http://http://babylone.ulb.ac.be/beatmusic/>) (40,41). In order to generate a consensus prediction exploiting the diversities of each method, we combined the results obtained by each category of method in an optimized predictor using a regression model tree (48). This resulted in two combined output predictions: (i)

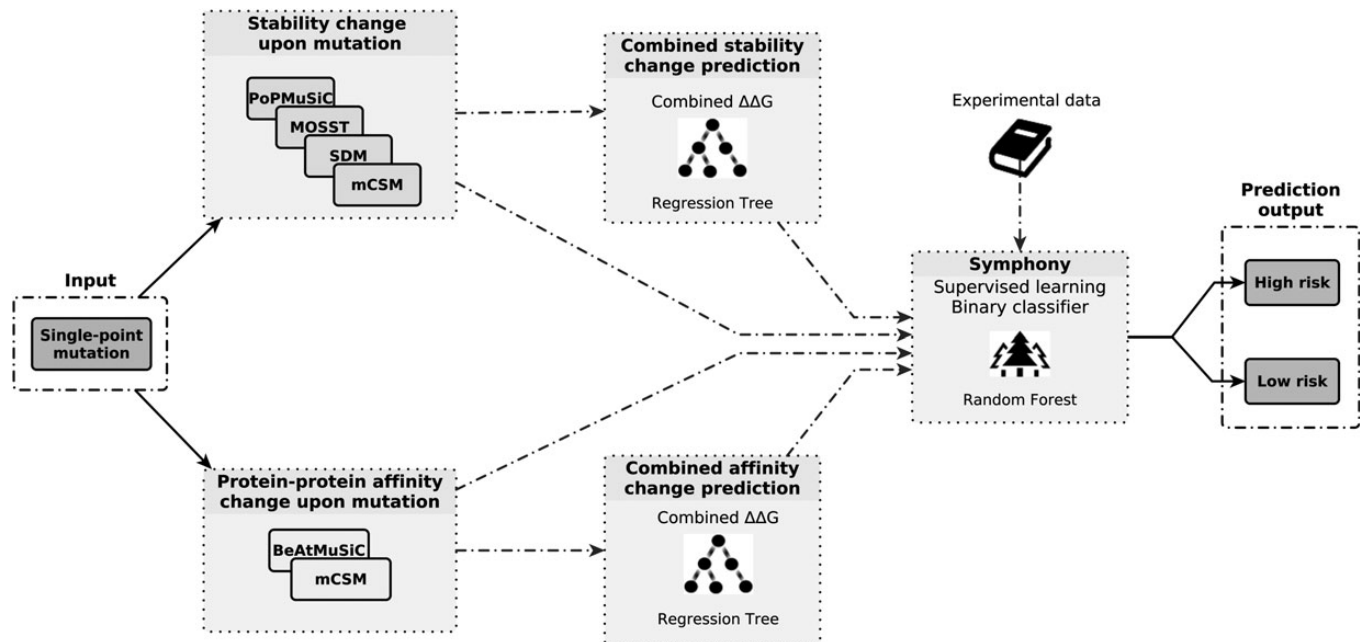


Figure 1. Workflow for predicting ccRCC risk for missense mutations in *VHL*. For a given mutation, computational methods from different paradigms are used to quantitatively assess its effects on protein stability and protein interactions with other proteins or ligands, all of which could affect function. These results are combined in optimized predictors via a regression model tree (using the M5 algorithm) (48) as a way to leverage the best of each method as well as to generate a consensus prediction. Experimental data were used to label each mutation in a training set according to ccRCC risk. The stability and affinity predictions are then used as evidence to train and test a binary classifier, using the ensemble learning method Random Forest (49), that outputs the predicted risk of ccRCC in a binary classification scheme (high or low risk). We named this integrated computational approach symphony.

combined predicted stability change (CPSC) and (ii) protein–protein affinity change (PPAC).

VHL disease ccRCC-associated mutations are significantly more destabilizing than mutations not associated with ccRCC

Mutations of surface residues were less common in ccRCC-associated than non-ccRCC-associated VHL disease (16.1 versus 34.4%, $P = 0.0353$; Supplementary Material, Table S3; Fig. 2). Solvent-exposed mutations would in general be expected to be less destabilizing than buried mutations at interfaces or within the protein core. Consistent with this, predicted CPSCs for ccRCC-associated mutations were significantly more destabilizing than those for non-ccRCC-associated mutations, irrespective of the groups of mutations included in the analysis (Supplementary Material, Table S4).

On subgroup analysis, predicted CPSCs for ccRCC-associated Type 2 mutations were significantly more destabilizing than those for non-ccRCC-associated Type 2 mutations ($P = 0.0043$; Supplementary Material, Table S4). There was no difference in predicted CPSC between Type 1 mutations associated and not associated with ccRCC, though this may simply reflect the small number of mutations in the latter cohort ($n = 7$).

There was no difference in predicted stability change between Type 1 and Type 2 mutations

Although Type 1 mutations were less likely to involve surface residues than Type 2 mutations (10.2 versus 29.6%,

$P = 0.0113$), there was no difference in CPSC ($P = 0.395$) or in the proportion of mutations which involved interface residues in Type 1 versus Type 2 VHL disease ($P = 0.353$, Fig. 2; Supplementary Material, Table S5).

Interface residues (Supplementary Material, Tables S3, S5 and S6)

Type 1 missense mutations were more likely to disrupt the HIF interface than Type 2 mutations ($P = 0.0014$). In contrast to previous findings (52), we found no difference in the proportion of mutations predicted to disrupt elongin C binding between Type 1 (11/49, 22%) and Type 2 missense mutations (23/72, 32%, $P = 0.254$) and found no particular association of ccRCC with specific pVHL regions.

Germline mutations associated only with pheochromocytomas

Seven Type 2C mutations are listed in the literature; one, L188V, is also associated with hereditary polycythaemia. We have not included R64P, R161Q, R167G, G93S and C162Y as Type 2C mutations, since they have been reclassified as Type 2B or Type 1 mutations in different kindreds, suggesting a variability of expression influenced by factors other than genotype. A representative list of germline *VHL* mutations associated with isolated PCCs is shown in Supplementary Material, Table S2B (13,14,53–55). We analysed only those germline mutations that are exclusively associated with PCCs germline mutations associated only with pheochromocytomas (GMEPs; $n = 26$;

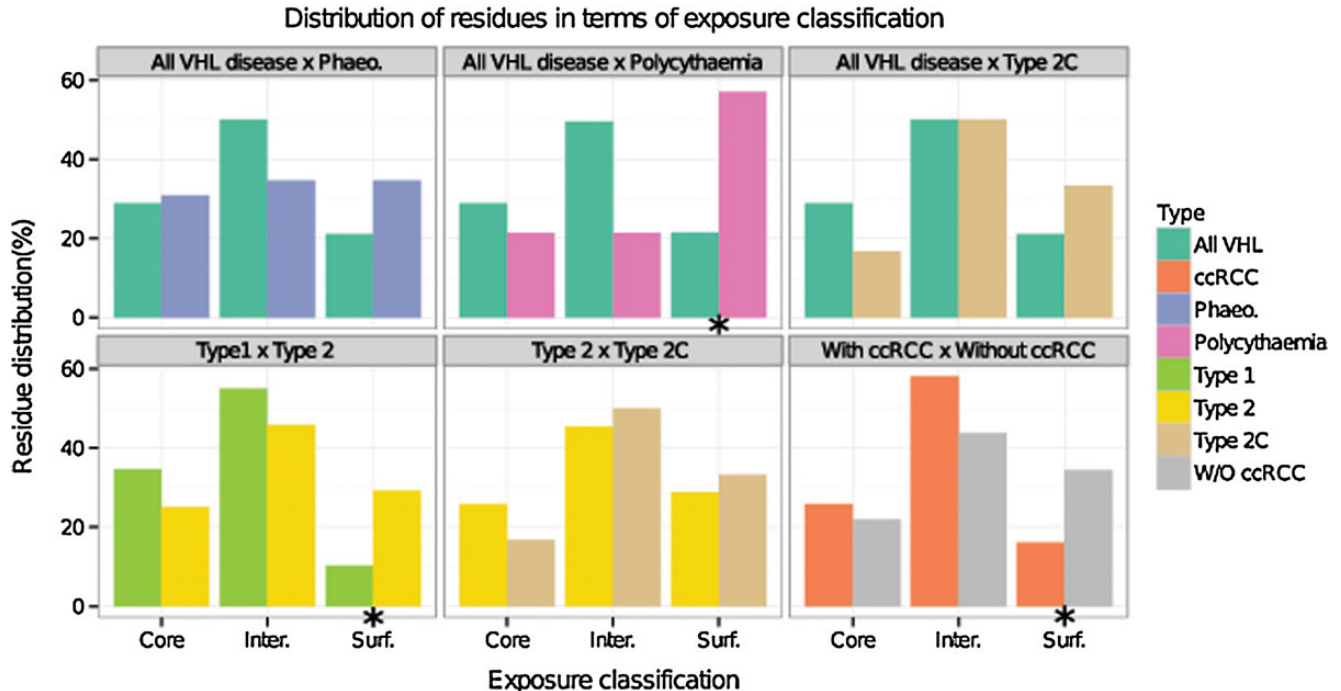


Figure 2. Exposure classification of mutations in different categories of disease. Type 2 mutations are more likely to involve surface residues than Type 1 mutations. ccRCC-associated VHL disease mutations are less likely to involve surface residues than non-ccRCC-associated VHL disease mutations. A high proportion of polycythaemia-associated mutations involves surface residues. Statistical significance ($P < 0.05$) indicated by asterisk. Inter., interface; Surf., surface; W/O, without.

20 PCC-associated germline mutations and six Type 2C VHL disease mutations).

GMEPs form a diverse group in terms of location within the structure of pVHL and predicted effects on the stability of the VCB complex. Predictions range from severely destabilizing to non-destabilizing (CPSC ranged from 1.32 to -4.797). Experimental data support diverse HIF α -regulation functional effects for these mutations (Supplementary Material, Table S7). There was no difference in amino acid exposure classification (Supplementary Material, Tables S3 and S5) or CPSC (Supplementary Material, Table S4) between GMEPs and VHL disease mutations.

Polycythaemia mutations are significantly more stable than all other VHL mutation groups

To date, 18 VHL mutations have been described in patients with hereditary polycythaemias (Supplementary Material, Table S8). Fourteen have never been associated with VHL disease or PCC. However, L188V is a Type 2C mutation and G104V, V130I and Y175C have been described as germline mutations associated with PCC; these mutations were excluded from subsequent analyses. When analysed as a group, CPSCs for hereditary polycythaemia VHL mutations are significantly less destabilizing than any other group of mutations, including non-ccRCC-associated VHL disease mutations (Supplementary Material, Table S4).

Sensitive prediction of ccRCC risk for germline VHL mutations

CPSC and PPAC predictions were combined to classify each mutation as 'high risk' or 'low risk' of association with

ccRCC. For the 121 VHL mutations included in the training set, symphony was 100% sensitive and 98% specific at predicting risk of ccRCC (Supplementary Material, Table S9A). We then looked at symphony's predictions for mutations included in both the training set and the test set. The predictions for the 162 germline VHL mutations are shown in Supplementary Material, Tables S2C and S9B and Figure 3. Of the 90 high-risk mutations, 84 (93.3%) had a diagnosis of VHL disease. 14 of 14 germline mutations associated only with hereditary polycythaemia were predicted to be low risk. The binary classifications for the 121 mutations associated with VHL disease are shown in Supplementary Material, Table S10.

For VHL disease mutations, the sensitivity of symphony in terms of predicting risk of ccRCC in VHL disease was 100% (95% CI 94.2–100%) and its specificity was 81.3% (95% CI 63.6–92.8%). Six VHL disease mutations were predicted high risk but have not yet been associated with ccRCC in VHL disease (Supplementary Material, Table S11). Of these, five have been described in sporadic ccRCC and some have experimental data suggesting that the resulting pVHL mutants are defective in HIF α -regulation. We suggest that patients with these germline mutations are at risk of ccRCC.

The 'Sorting Tolerant From Intolerant' (SIFT) algorithm (56) is commonly used to predict whether a single amino acid substitution affects protein function. SIFT predictions for mutations included in our training set are shown in Supplementary Material, Table S9C. The sensitivity (82.3%, 95% CI 70.5–90.8%) and specificity (54.2%, 95% CI 40.8–67.3%) of SIFT was significantly lower than that of symphony. Eleven high-risk mutations were predicted to be tolerated by SIFT.

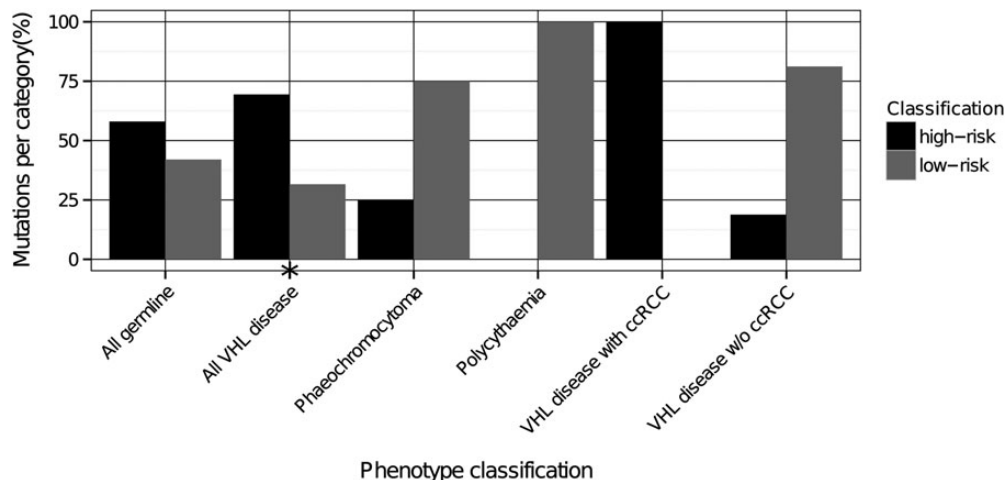


Figure 3. Binary classification of ccRCC risk for germline *VHL* mutations associated with different phenotypes. High-risk mutations were more likely to be associated with VHL disease than low-risk mutations. All ccRCC-associated VHL disease mutations were high risk. All polycythaemia-associated mutations were low risk. Statistical significance ($P < 0.05$) indicated by asterix. w/o, without.

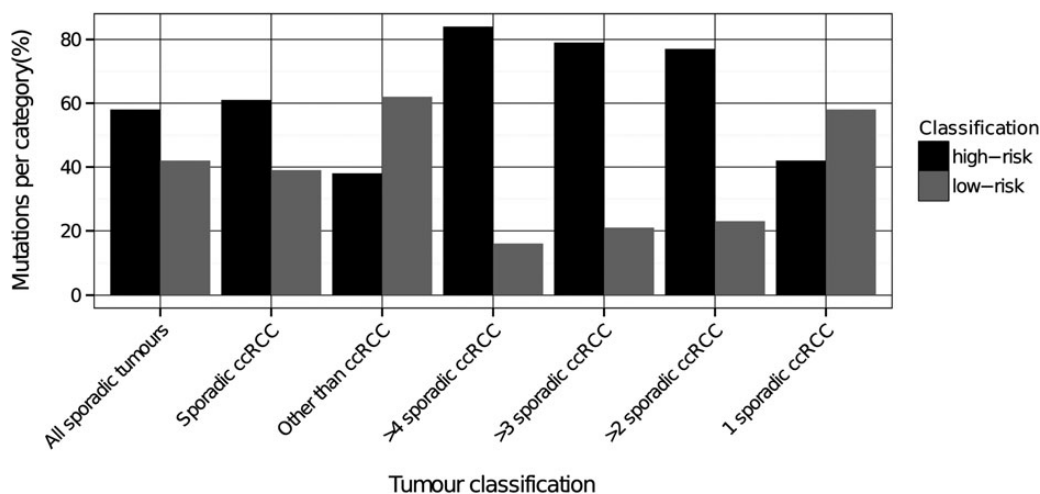


Figure 4. Predictions for risk of ccRCC in sporadic tumours. The proportion of high-risk mutations is significantly higher for mutations which have been described several times in sporadic disease compared with those that have been described only once and is significantly higher in sporadic ccRCC compared with other somatic tumour types.

Predicting ccRCC risk for somatic *VHL* mutations

Two hundred and fifteen somatic *VHL* mutations are listed on COSMIC (38) at the time of writing. Of these, 186 have been described in sporadic ccRCC and 29 have only been described in tumours other than ccRCC (Supplementary Material, Tables S2A and S12). The prediction summary for ccRCC risk in sporadic tumours is shown in Supplementary Material, Table S13 and Figure 4.

Of the 215 somatic mutations, 124 (58%) were predicted high risk and 91 (42%) low risk (Supplementary Material, Table S13). Seventy-three of 186 (39%) of somatic ccRCC missense mutations were predicted to be low risk. The proportion of high-risk mutations is significantly higher for mutations described several times in sporadic disease compared with those described only once and is significantly higher in sporadic

ccRCC compared with other somatic tumour types (61 versus 38%, $P = 0.0207$; Supplementary Material, Table S13; Fig. 4). Fifty-three percent of somatic ccRCC high-risk mutations have also been described in VHL disease (and of these 75% have definitely been associated with ccRCC), compared with only 21% of low-risk mutations (none of which have definitely been associated with ccRCC) ($P < 0.0001$; Supplementary Material, Table S14).

For mutations reported more than once in sporadic ccRCC, 62% of the high-risk mutations have been associated with VHL disease, compared with 26% of the low-risk mutations ($P < 0.0001$; Supplementary Material, Table S14 and Fig. S2). For somatic mutations reported only once in ccRCC or in other tumour types, only 38% of the high-risk mutations have been associated with VHL disease, compared with 18% of the low-risk mutations ($P = 0.0132$; Supplementary Material, Table S14).

Predicted ccRCC risk for *VHL* mutations investigated experimentally

A review of the literature revealed 65 missense *VHL* mutations with measured experimental effects on HIF α regulation. These experimental settings include non-standardized biophysical and biochemical data in addition to cell culture studies using cell lines that may express HIF1 α , HIF2 α or both. This explains to some extent why different studies report different effects for the same mutation. For example, pVHL^{R167Q} has been reported to be defective in HIF α regulation in some studies but similar to wild type (WT) in others. Similarly, some pVHL mutants have been reported to have different effects on HIF1 α and HIF2 α which would not be detected in studies only looking at the effect on one HIF α isoform. The balance of evidence suggests that while HIF2 α is an oncogene in ccRCC pathogenesis, HIF1 α may act as a tumour suppressor (57).

With these caveats, 15 mutations were predicted to be low risk but are reported to be defective in HIF α regulation experimentally (Supplementary Material, Table S18). None of these mutations have been described in VHL disease associated with ccRCC, suggesting that, in many cases, the extent of dysregulation of HIF α seen in experimental systems may not be enough for tumourigenesis. Paradoxically, D121Y and L153P have both been described several times in sporadic ccRCC and are reported to be defective in HIF α regulation, suggesting that in these cases our low-risk prediction may be incorrect. Five mutations were predicted to be high risk but have been reported to regulate HIF α similarly to WT *VHL* in experimental systems. Four of these (S80N, P81S, T157I and I180V) have all been described in VHL disease associated with ccRCC suggesting the mutations are high-risk despite appearing to regulate HIF α normally under certain experimental conditions. In certain situations, our predictions thus seem more sensitive than experimental data regarding HIF α regulation in terms of determining the probable pathogenic effect of a mutation. A finer structure–function analysis of pVHL could shed light on these incongruences between prediction, experiments and disease manifestation.

Development of a publically available web server

We use symphony to generate predictions for risk of ccRCC for all possible *VHL* mutations. We present these predictions, in association with clinical and experimental data, in a publically available, searchable web server which can be freely accessed by research groups worldwide (<http://structure.bioc.cam.ac.uk/symphony>).

Linking predictions to clinicopathological features in cohort of patients with sporadic ccRCC

Previously, we have presented the results of targeted sequencing of *VHL*, BRCA1-associated protein-1 (*BAP1*), Polybromo 1 (*PBRM1*), SET domain containing 2 (*SETD2*) and lysine (K)-specific demethylase 6A (*KDM6A*) on 132 ccRCCs and matched normal tissues (58). Application of our integrated computational approach to somatic missense *VHL* mutations predicted 26 high-risk (76%) and 8 low-risk (24%) mutations. One mutation (R58W) affects a residue that is not in the *VHL* crystal structure and was excluded. There was no difference in clinicopathological

features between ccRCCs with predicted pathogenic *VHL* alterations (high-risk missense mutations, nonsense or frameshift mutations or promoter methylation) and those without predicted pathogenic *VHL* alterations (including low-risk missense mutations) (Supplementary Material, Table S15).

DISCUSSION

This work demonstrates the application of computational biological approaches to predict the effects of missense *VHL* mutations in VHL disease, sporadic ccRCC and congenital polycythaemia with potential clinical applications. We have created a comprehensive and inclusive database of missense *VHL* mutations linked to experimental data and clinical phenotype. We used this database to train and test an optimized binary classification system (named symphony), which integrates predictions from a variety of *in silico* methods and can predict the risk of ccRCC associated with *VHL* missense mutations with high sensitivity and specificity. We use symphony to generate predictions for risk of ccRCC for all possible *VHL* mutations and present these predictions, in association with clinical and experimental data, in a publically available, searchable web server.

pVHL is an exemplary yet challenging protein to use as the basis for development of an *in silico* predictive model; it forms part of a ternary complex and, despite being small (213 amino acids) we identified 294 unique missense mutations. Experimental data regarding the functional effects of 82 of these mutations enabled us to validate the predictions from early *in silico* models, this information was of particular use during development of our final model and permitted us to identify and learn from mutations which were incompletely assessed by various *in silico* tools used independently. The association between different mutations and distinct phenotypes in VHL disease and *VHL*-associated congenital polycythaemias provided an excellent opportunity to identify the molecular basis of genotype–phenotype correlations using bioinformatics tools. The novel integrated strategy we have developed could easily be adapted for other systems.

Phaeochromocytomas in VHL disease

GMEP mutations are broadly distributed throughout *VHL* and the resulting amino acid changes are predicted to have diverse effects on pVHL and the VCB complex; some, such as F119S, are severely destabilizing hydrophobic core mutations; others, such as D143Q, are minimally destabilizing surface mutations. Experimental data report diverse effects with respect to HIF α regulation for Type 2C VHL disease mutations (Supplementary Material, Table S7). In contrast to previous less comprehensive work (52), we found no evidence that mutations associated with PCCs (Type 2) are more likely to disrupt interactions at the elongin C interface than mutations not associated with PCCs (Type 1). Only 8% of Type 2 mutations were predicted to directly disrupt the pVHL-HIF α interface, compared with 33% of Type 1 mutations, suggesting that a direct disruption of HIF α binding may not be necessary to cause PCCs and that an HIF-independent mechanism underlies the pathogenesis of PCCs (further material in Supplementary Material, Discussion).

pVHL and hereditary polycythaemias

The most common *VHL* polycythaemia mutation is the homozygous C598T mutation, resulting in the amino acid substitution R200W (4). Seventeen additional *VHL* variants (16 missense and 1 nonsense) associated with congenital secondary polycythaemia (CSP) have been described (Supplementary Material, Table S8). Reports of tumour development in patients with *VHL*-associated CSP are extremely rare (59,10), and a knock-in R200W transgenic mouse exhibits polycythaemia without tumour formation (60,61).

The molecular mechanism underlying *VHL*-associated CSPs is debated (Supplementary Material, Table S16) and the lack of tumourigenesis in *VHL*-associated CSPs is notable. Our work demonstrates that mutations associated solely with hereditary polycythaemias are significantly less destabilizing than all other subgroups of disease-associated *VHL* mutations. CSP-associated *VHL* mutations are distributed throughout *VHL* and are not limited to the 3' region of *VHL* exon 3. Along with experimental data summarized in Supplementary Material, Table S16, this suggests that, in the majority of cases of *VHL*-associated CSP, a combination of *VHL*-associated CSP mutations on both *VHL* alleles, each of which independently results in minor impairment of HIF2 α activity, is sufficient to cause CSP but insufficient to cause tumourigenesis.

Risk of ccRCC in VHL disease is linked to the degree of destabilization resulting from missense mutations

There was no difference between the CPSC of Type 1 and Type 2 *VHL* disease mutations, or in the proportion of mutations which involve interface residues, implying no clear functional difference between missense mutations described in Type 1 and Type 2 *VHL* diseases. The description of at least 17 *VHL* missense mutations in both disease types supports this statement (Supplementary Material, Table S2A). The only clear difference between Type 1 and Type 2 mutations was a significantly lower proportion of Type 2 mutations predicted to disrupt the HIF α interface, and, in agreement with previous studies (29), a higher prevalence of surface amino acid substitutions in Type 2 than Type 1 *VHL* disease.

In contrast, we report a clear difference between ccRCC- and non-ccRCC-associated missense mutations; the risk of ccRCC in *VHL* disease is significantly associated with the degree of destabilization resulting from the mutations. This observation is in agreement with experimental data linking risk of ccRCC in *VHL* disease to the degree to which HIF activity is compromised (26,27). Severely destabilizing mutations are expected to dramatically impair the function of pVHL while less destabilizing mutations may allow partial preservation of pVHL's function. Similarly, nonsense and frameshift mutations, which are expected to knock-out most, if not all, of pVHL's functionality, have a higher risk of ccRCC and haemangioblastomas in *VHL* disease than missense mutations (28–30). A small, earlier study also associated ccRCC development in *VHL* disease with a relatively high loss of structural stability in pVHL missense mutants (62).

Here we consider the effects of mutations that might destabilize pVHL itself or its interactions within VCB (Fig. 5). The computational approaches that we have used assess impacts of the

mutation on the conformation and direct interactions of the substituted amino acid on stability of the subunit and its interactions, for example, through SDM, PoPMuSiC and BeAtMuSiC. They also assess the importance of the more extended environment, including depth of the amino acid in the core and its electrostatic environment, which are implicit in mCSM and PoPMuSiC. We suggest that diverse mechanisms of destabilization can result in the same endpoint, namely disruption of pVHL's ability to target HIF α for ubiquitination and degradation, and that the degree of dysfunction is closely associated with the degree of destabilization resulting from a mutation. Alternatively, mutations such as H115Y and S111R, which directly interfere with the HIF α -hydroxyproline-binding site, may disrupt pVHL's ability to target HIF α for ubiquitination and degradation without destabilizing the entire VCB complex; these mutations may be associated with a low risk of PCC. Overall, our data suggest that missense *VHL* mutations, which are drivers in ccRCC pathogenesis, either destabilize the VCB complex as a whole or directly affect the HIF α -hydroxyproline-binding site (Fig. 6). In contrast to previous work, we found no suggestion that disrupted interactions between pVHL and its binding partners correlate with ccRCC-risk in *VHL* disease (52).

This model provides an explanation for the mechanism whereby different mutations at the same position can be associated with different phenotypes. For example, Y98H is a Type 2A *VHL* disease mutation and is associated with a much lower ccRCC risk in *VHL* disease than the Type 2B mutation at the same residue, Y98N. This is reflected by a lower CPSC for Y98N compared with Y98H. Experimental data have previously demonstrated Y98H to exhibit higher stability and greater binding affinities for HIF1 α compared with Y98N (63). Similar findings are seen for Type 2A and Type 2B mutations at positions G93, Y112, A149, R167, V170 and L188 (Supplementary Material, Table S2A).

The data regarding the presence or absence of ccRCC in *VHL* disease relates to kindreds only, rather than figures regarding the proportion of patients with each mutation who developed ccRCC. Thus, we were not able to discriminate between mutations associated with a very high risk of ccRCC and mutations which rarely cause ccRCC. However, our results suggest a gradient effect of *VHL* missense mutations whereby the risk of ccRCC increases roughly in proportion to the destabilizing effect of the mutation.

Development of a binary classification system to predict the risk of ccRCC associated with VHL missense mutations

The disparate relationship between specific missense *VHL* mutations and clinical phenotype in *VHL* disease and congenital polycythaemias provides an excellent opportunity to develop a sensitive and specific classifier to predict the risk of ccRCC in *VHL* disease. The binary classifier we developed (symphony) was trained using a dataset of mutations designated high risk or low risk in terms of ccRCC pathogenesis based on experimental and clinical data. During training, our optimized model was highly sensitive and specific and predicted the association of high- and low-risk mutations with ccRCC with 100% accuracy. Though its specificity was lower (81%) when looking at all *VHL* disease mutations (i.e. including mutations from both the training and test sets) it is possible that the six mutations predicted

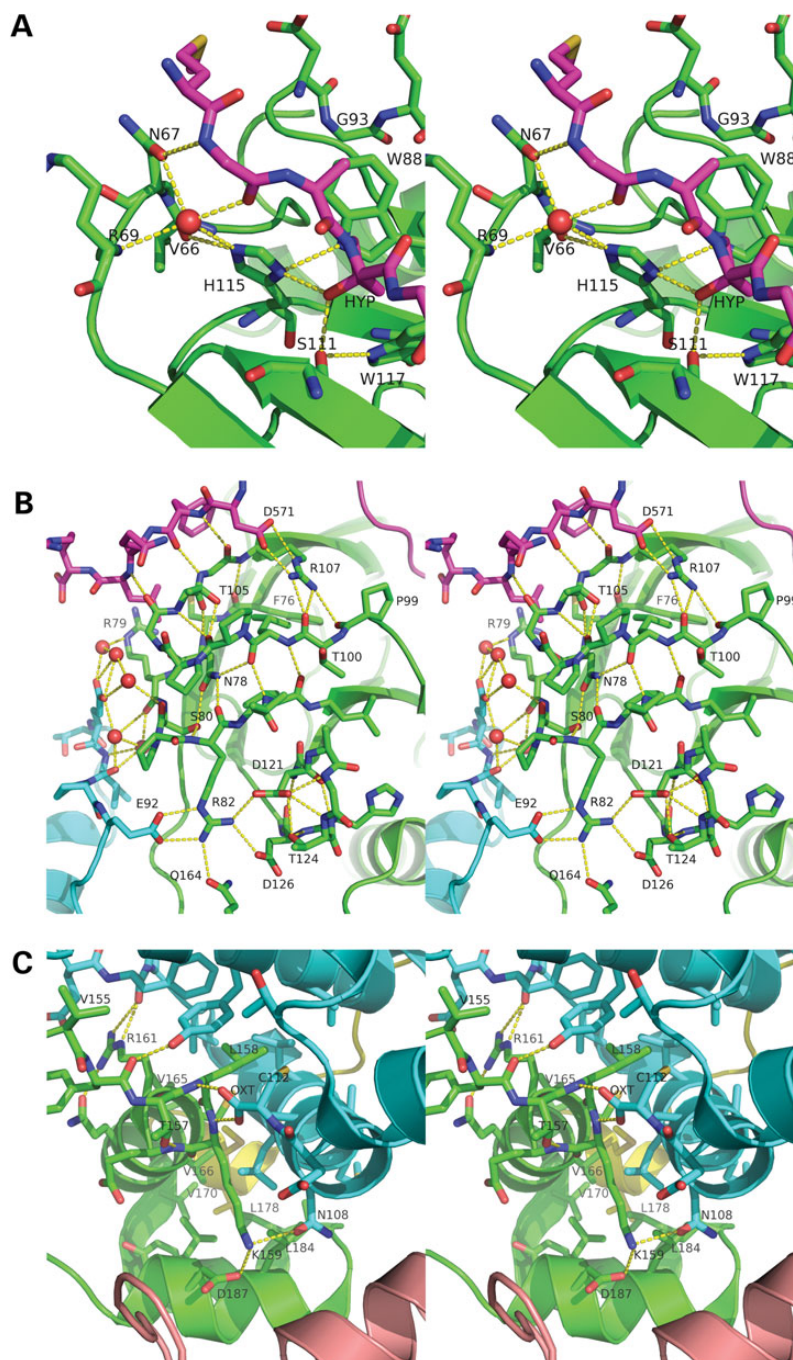


Figure 5. Wall-eye stereograms representing examples of the diverse mechanisms of the effects of *VHL* mutations at the molecular level. (A) Mutations that disrupt the HIF α -hydroxyproline-binding site. Mutations of H115 (A) cause loss of the tetracoordination of a buried water molecule, as well as the loss of an acceptor group for the hydroxyl donor group in the HYP residue in hydroxylated HIF. They also disrupt up to four H-bonds within the buried hydrogen bond network that recognizes the HYP. Mutations at S111 can cause the loss of the hydrogen donor for the HYP hydroxyl, with similar effects as above. Mutations at other depicted residues that participate in the HYP hydrogen-bonding recognition network can have similar outcomes. Mutations at secondary HIF α -binding sites, such as the loop G104-R108 (B) could also impair HIF α binding. Side chains of L562 and Y565 in HIF have not been represented for clarity in A. (B) Mutations that disrupt the hydrophobic core. Mutation in the two VHL hydrophobic cores can disrupt VHL subunit conformation, such as mutations at F76 (beta domain, B) and V170 (alpha domain, C). (C) Mutations that disrupt H-bond networks. Mutation of N78 (B) disrupts a significant buried hydrogen bond network that stabilizes a region of VHL connecting two loops that are key for interaction with HIF and elongin C. Mutation of residues that tightly interact in this network, such as S80 and T105 have the same effect. (D) Mutations that disrupt the conformation of pVHL. Mutations of conserved glycines (e.g. G93, A) and prolines can directly disrupt and destabilize pVHL. (E) Mutations of residues within the elongin C- and elongin B-binding sites. These mutations disrupt interaction with VHL-binding partners through disruption of H-bonding networks, such as mutations at R82 and its neighbouring residues (B) and R161, or through disruption of hydrophobic cores formed by the interacting subunits in the VCB complex, such as mutations at V155, L158, K159, C162, V165, V166, V170, L178 and L184 (C). (F) Mutations that disrupt long-range electrostatic interactions. These mutations can alter the charge complementarity of the subunits in the VCB complex and destabilize protein-protein long- and short-range interaction, such as mutations at R79, R82, R107, D121, D126 (B), K159 and D187 (C). In this figure, VHL is coloured in green, elongin C in cyan, elongin B in yellow, HIF α in magenta, Cullin 2 in pink-orange and water oxygen atoms are represented as red balls.

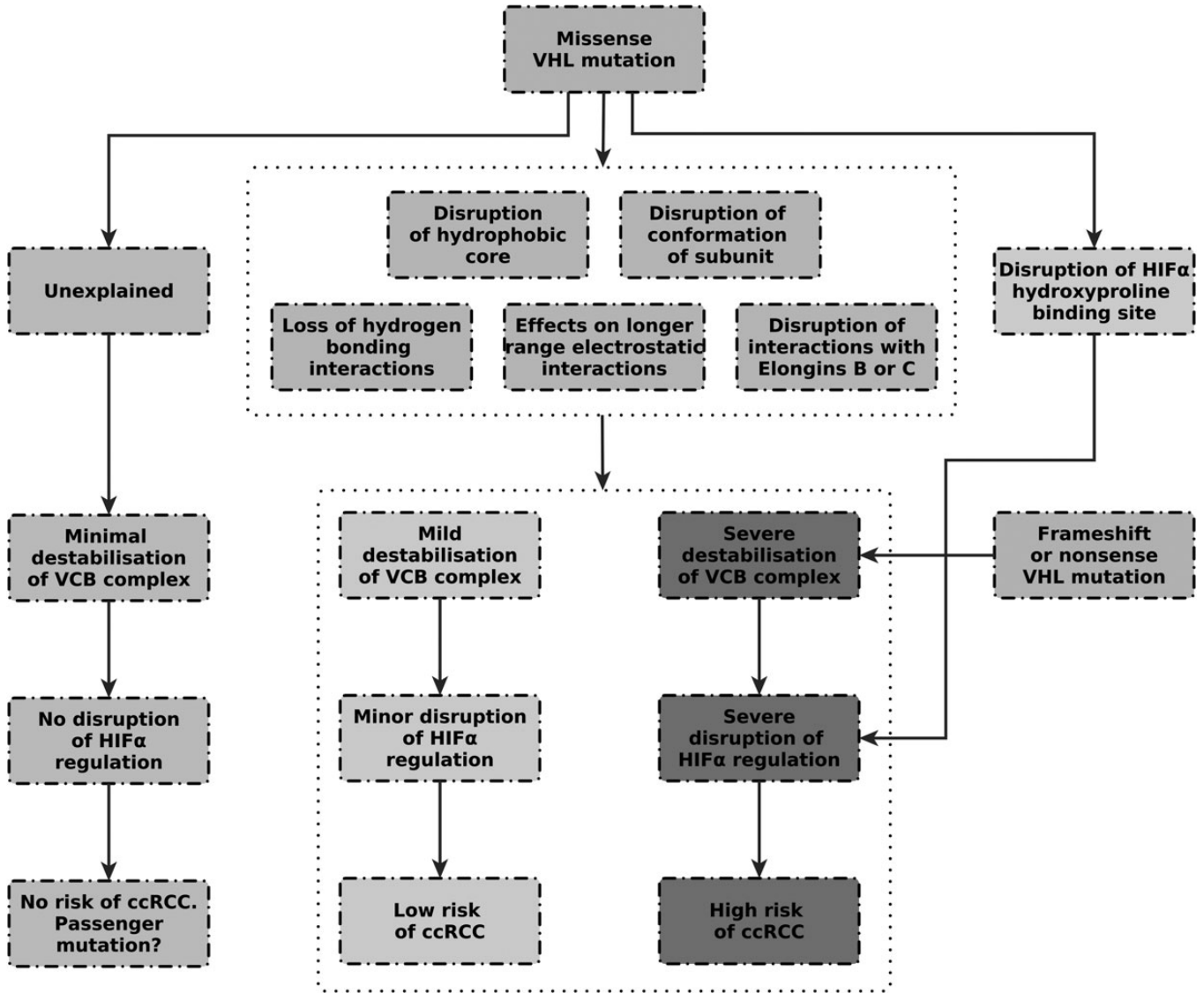


Figure 6. A model to explain diverse phenotypes associated with *VHL* missense mutations. Frameshift/nonsense *VHL* mutations are likely to prevent formation of a functional VCB complex and result in severe disruption of HIF α regulation, thereby explaining the high risk of ccRCC in Type 1 *VHL* disease. Missense *VHL* mutations may destabilize the VCB complex through a variety of mechanisms. Mutations which are severely destabilizing are likely to severely disrupt HIF α regulation and are associated with a high risk of ccRCC. Less severely destabilizing mutations have a milder effect on HIF α regulation resulting in a lower risk of ccRCC. A few mutations do not destabilize the VCB complex as a whole but instead directly disrupt the HIF α -hydroxyproline-binding site, thereby affecting HIF α regulation. These mutations may be associated with a low risk of PCC. Some missense *VHL* mutations are not destabilizing and would not be predicted to affect the HIF α -hydroxyproline-binding site and may represent passenger mutations.

to be high risk that have not yet been associated with ccRCC in *VHL* disease may be in the future.

In a blind test, symphony suggests that 39% of missense mutations described in somatic ccRCC are low risk and may represent passenger changes. Though this figure initially seems high it is supported by several observations. First, the proportion of mutations predicted to be high risk is significantly higher in mutations described several times in sporadic disease compared with those described only once and is significantly higher in sporadic ccRCC compared with other somatic tumour types. Secondly, 53% of somatic ccRCC mutations predicted to be high risk have been described in *VHL* disease (of these 67% have definitely been associated with ccRCC) compared with only 21% of

mutations predicted to be low risk (none of which have definitely been associated with ccRCC). Thirdly, the R200W mutation, which has been clearly demonstrated to be non-tumourigenic in terms of ccRCC pathogenesis, has been identified in two cases of sporadic ccRCC thereby exemplifying the presence of a low-risk *VHL* mutation in sporadic ccRCC. Finally, experimental data for many of the predicted low-risk mutations confirm that they appear to regulate HIF α similarly to WT *VHL*.

The ability to assess the risk of ccRCC associated with germline *VHL* missense mutations in *VHL* disease may be clinically useful, particularly since ccRCC is a significant cause of morbidity and mortality (64,65). In sporadic ccRCC, as yet no clear association between *VHL* mutation status and clinicopathological

features has been identified (reviewed in 39). Sensitive and specific identification of passenger mutations which do not drive tumour formation may allow identification, in large datasets, of genotype–phenotype correlations for high-risk mutations that have previously been concealed by the inclusion of passenger mutations in analyses. Inactivation of *VHL* alone is not sufficient to cause ccRCC (66,67) and recently, genomic sequence analysis has identified several genes that are frequently mutated in ccRCC. These include *PBRM1*, *SETD2* and *BAP1*, all of which lie on a relatively small, 43 Mb region of chromosome 3p and are, therefore, potentially deleted alongside *VHL* in tumours with 3p loss. It is tempting to speculate that there may be an association between the presence or absence of high-risk *VHL* alterations and mutations in other driver genes (such as *PBRM1* and *BAP1*); assessment of these factors in combination may be useful in predicting response to targeted therapies. This concept could be investigated using the symphony web server which presents predictions for all possible *VHL* mutations.

CONCLUSIONS

We have combined a variety of bioinformatics tools, each of which uses a different methodology to independently predict the effects of missense mutations with moderate efficacy, to produce a combined model which can predict the risk of ccRCC associated with missense *VHL* mutations with high sensitivity and specificity. This study represents the most comprehensive analysis of *VHL* missense mutations to date. The methodology we have developed is generic and transparent and could easily be adapted for the study of different proteins in other types of cancer. We have generated predictions for risk of ccRCC for all possible *VHL* mutations, presented in a publically available, searchable web server. This resource could easily be utilized in analyses of sequencing data from large patient cohorts, particularly from clinical trials of ccRCC patients.

MATERIALS AND METHODS

Database of *VHL* missense mutations

We compiled a comprehensive table of germline and somatic *VHL* mutations (Supplementary Material, Table S2A) obtained from numerous sources, including original articles, the Universal Mutation Database (UMD; <http://www.umd.be/VHL/>, last accessed on 19 July 2014) (2) and the review article by Nordstrom-O'Brien *et al.* (3). Details of mutations not included in this review article were obtained from the original reference. A list of somatic *VHL* mutations associated with sporadic tumours was obtained from COSMIC (38). A representative list of germline mutations described in non-syndromic PCC was identified (13,14,53–55).

Accurate phenotype data are not publically available for all familial mutations, with many simply being classified as Type 1 or Type 2 with no further details. Furthermore, a single mutation may be classified differently in different kindreds, highlighting the differential expression of *VHL* mutations between individuals. For the purpose of this study, mutations were subgrouped depending on whether they have definitively been associated with ccRCC or not. Mutations that have been associated with

ccRCC in at least one patient were documented as ccRCC associated. If the clinical data associated with a mutation were incomplete the association with ccRCC was documented as ‘unknown’. Mutations reported as both Type 1 and Type 2 were classified as Type 2 for the purpose of this study, since, by definition, they have been associated with PCC in at least one patient. Germline mutations associated with PCCs and no other tumour types were only classed as Type 2C mutations if they have been clearly associated with PCCs across more than one generation. Germline mutations associated with PCCs without a family history were classified as PCC-associated germline mutations.

Experimentally defined functional effects of missense *VHL* mutations were identified using the search terms ‘*VHL*’ and ‘Mutation’ on PubMed.

Annotated datasets for machine learning

The primary aim of this work was to identify *VHL* mutations likely to be pathogenic in ccRCC. We therefore compiled a ‘training’ set of 121 mutations: 62 ‘high-risk’ mutations identified as *VHL* disease mutations clearly associated with ccRCC; and 59 so-called low-risk mutations, comprising (i) 6 mutations described less than or equal to once in sporadic ccRCC with experimental data suggesting no functional effect resulting from the mutation, (ii) 17 germline mutations described in association with hereditary polycythaemia, (iii) 7 single-nucleotide polymorphisms not associated with sporadic or familial disease of any kind as listed on NCBI (68) and (iv) 29 germline *VHL* disease mutations with good quality clinical data documenting no association with ccRCC. The test set of mutations compiled 173 mutations. These comprised: (i) 39 germline mutations associated with *VHL* disease (all types), (ii) 1 germline mutation associated with hereditary polycythaemia, (iii) 1 germline mutation associated with CNS haemangioblastoma, (iv) 13 germline mutations associated with PCCs, (v) 112 somatic mutations associated with sporadic tumours (either ccRCC or other tumour types) as listed on COSMIC (38) and (vi) 7 additional mutations referenced in the literature without associated clinical data. Details of all mutations are listed in the Supplementary Material, Table S2A.

Predicting protein stability and PPAC upon mutation

Five computational methods were used to predict the effects of missense mutations: (i) mCSM (43) (<http://structure.bioc.cam.ac.uk/mcsm>), (ii) SDM (<http://mordred.bioc.cam.ac.uk/~sdm/sdm.php>) (44,45), (iii) MOSST, (iv) PoPMuSiC (40) (<http://babylone.ulb.ac.be/popmusic/>) and (v) BeAtMuSiC (40,41) (<http://babylone.ulb.ac.be/beatmusic/>). In order to improve overall accuracy and obtain a consensus prediction from the several computational methods used, we combined their results using regression trees, via an implementation of the M5 model tree algorithm (48). Supplementary Material, Figure S1 shows the obtained regression tree for the CPSC predictor. For PPAC the model tree obtained for the combined predictor only had one node that describes the following linear model: $\Delta\Delta G = 0.758 \times \text{mCSM} + 0.432 \times \text{BeAtMuSiC} - 0.035$. The regression trees were trained using a diverse dataset of 350 mutations with experimental thermodynamic

data derived from the ProTherm (69) and SKEMPI (70) databases and used in a blind test in a previous study (43). Supplementary Material, Table S17 presents the Pearson's correlation coefficient obtained for each method as well as for the combination of them via regression trees.

Predicting risk of ccRCC in VHL disease

We developed a machine learning strategy to link the effects of VHL missense mutations to phenotype. Statistical analysis of the CPSCs and combined predicted PPACs associated with missense mutations, linked to collated experimental data regarding their functional effects and clinical phenotype, facilitated development of a binary classifier that aims to relate the effects of missense mutations to risk of ccRCC; this was based on the finding that mutations that are associated with ccRCC in VHL disease tend to be more destabilizing than those that are not. The classifier uses CPSC and PPAC predictions as evidence to train the predictive model using the Random Forest algorithm (49), and outputs the predicted risk of ccRCC in a binary classification scheme (high or low risk).

Statistical analysis

All statistical analyses were performed using SPSS Statistics 20.0. Associations between a mutation group and predicted $\Delta\Delta G$ s were determined using unpaired Student's *t*-test. Association between a mutation group and exposure classification was determined using: χ^2 -test for categorical variables if $>80\%$ of the expected counts are >5 ; Fisher's exact test for categorical variables if $>20\%$ of the expected counts are <5 . Unless indicated *P*-values are two sided without adjustment for multiple comparisons.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at *HMG* online.

ACKNOWLEDGEMENTS

We acknowledge the CRUK Cambridge Institute (part of the Cambridge Biomedical Research Centre), the University of Cambridge and Hutchison Whampoa Limited. The authors thank Harry Jubb who kindly provided the accessibility calculations to define interface residues in the VHL complex.

Conflict of Interest statement. L.G., D.E.V.P., A.O.-N., J.A. have no conflicts of interest. T.E. owns shares with Astra Zeneca and has attended advisory boards for Bayer, Pfizer, Roche, GSK and AVEO. He has corporate-sponsored research from Astra Zeneca, GSK, Pfizer and Bayer and received consultation fees from Roche, Bayer, Pfizer, GSK and AVEO. T.B. is Deputy Chair of the Institute of Cancer Research. He owns shares in GSK. He is a founder of the oncology structure-guided drug company, Astex Technology/Therapeutics Ltd., and subsequent to its purchase by Otsuka, now sits on the board of the UK branch, Astex Therapeutics Ltd. He has received science advisory fees from Pfizer, UCB, SKB and Astex.

FUNDING

This work was supported by Cancer Research UK Hales Fellowship (L.G.), the Conselho Nacional de Desenvolvimento Científico e Tecnológico, Brazil (D.E.V.P.), the Institute for Cell Dynamics and Biotechnology (ICM project # P05-001-F) and the Centre for Biotechnology and Bioengineering, University of Chile (CeBiB, project FB0001) and Fondecyt Project No. 1141311. Funding to pay the Open Access publication charges for this article was provided by the Cambridge Biomedical Research Centre.

REFERENCES

1. Maher, E.R., Neumann, H.P. and Richard, S. (2011) von Hippel-Lindau disease: a clinical and scientific review. *Eur. J. Hum. Genet.*, **19**, 617–623.
2. Beroud, C., Joly, D., Gallou, C., Staroz, F., Orfanelli, M.T. and Junien, C. (1998) Software and database for the analysis of mutations in the VHL gene. *Nucleic Acids Res.*, **26**, 256–258.
3. Nordstrom-O'Brien, M., van der Looij, R.B., van Rooijen, E., van den Ouweland, A.M., Majoor-Krakauer, D.F., Lolkema, M.P., van Brussel, A., Voest, E.E. and Giles, R.H. (2010) Genetic analysis of von Hippel-Lindau disease. *Hum. Mutat.*, **31**, 521–537.
4. Ang, S.O., Chen, H., Hirota, K., Gordeuk, V.R., Jelinek, J., Guan, Y., Liu, E., Sergueeva, A.I., Miasnikova, G.Y., Mole, D. *et al.* (2002) Disruption of oxygen homeostasis underlies congenital Chuvash polycythemia. *Nat. Genet.*, **32**, 614–621.
5. Pastore, Y.D., Jelinek, J., Ang, S., Guan, Y., Liu, E., Jedlickova, K., Krishnamurti, L. and Prchal, J.T. (2003) Mutations in the VHL gene in sporadic apparently congenital polycythemia. *Blood*, **101**, 1591–1595.
6. Bento, C., Almeida, H., Maia, T.M., Relvas, L., Oliveira, A.C., Rossi, C., Girodon, F., Fernandez-Lago, C., Aguado-Diaz, A., Fraga, C. *et al.* (2013) Molecular study of congenital erythrocytosis in 70 unrelated patients revealed a potential causal mutation in less than half of the cases (Where is/are the missing gene(s)?). *Eur. J. Haematol.*, **91**, 361–368.
7. Lanikova, L., Lorenzo, F., Yang, C., Vankayalapati, H., Drachtman, R., Divoky, V. and Prchal, J.T. (2013) Novel homozygous VHL mutation in exon 2 is associated with congenital polycythemia but not with cancer. *Blood*, **121**, 3918–3924.
8. Tomasic, N.L., Piterkova, L., Huff, C., Bilic, E., Yoon, D., Miasnikova, G.Y., Sergueeva, A.I., Niu, X., Nekhai, S., Gordeuk, V. *et al.* (2013) The phenotype of polycythemia due to Croatian homozygous VHL (571C>G:H191D) mutation is different from that of Chuvash polycythemia (VHL 598C>T:R200W). *Haematologica*, **98**, 560–567.
9. Bond, J., Gale, D.P., Connor, T., Adams, S., de Boer, J., Gascoyne, D.M., Williams, O., Maxwell, P.H. and Ancliff, P.J. (2011) Dysregulation of the HIF pathway due to VHL mutation causing severe erythrocytosis and pulmonary arterial hypertension. *Blood*, **117**, 3699–3701.
10. Capodimonti, S., Teofili, L., Martini, M., Cenci, T., Iachinimoto, M.G., Nuzzolo, E.R., Bianchi, M., Murdolo, M., Leone, G. and Larocca, L.M. (2012) Von Hippel-Lindau disease and erythrocytosis. *J. Clin. Oncol.*, **30**, e137–e139.
11. Lorenzo, F.R., Yang, C., Lanikova, L., Butros, L., Zhuang, Z. and Prchal, J.T. (2013) Novel compound VHL heterozygosity (VHL T124A/L188V) associated with congenital polycythaemia. *Br. J. Haematol.*, **162**, 851–853.
12. Randi, M.L., Murgia, A., Putti, M.C., Martella, M., Casarin, A., Opocher, G. and Fabris, F. (2005) Low frequency of VHL gene mutations in young individuals with polycythemia and high serum erythropoietin. *Haematologica*, **90**, 689–691.
13. Cascon, A., Pita, G., Burnichon, N., Landa, I., Lopez-Jimenez, E., Montero-Conde, C., Leskela, S., Leandro-Garcia, L.J., Leton, R., Rodriguez-Antona, C. *et al.* (2009) Genetics of pheochromocytoma and paraganglioma in Spanish patients. *J. Clin. Endocrinol. Metab.*, **94**, 1701–1705.
14. Neumann, H.P., Bausch, B., McWhinney, S.R., Bender, B.U., Gimm, O., Franke, G., Schipper, J., Klisch, J., Althoefer, C., Zerres, K. *et al.* (2002) Germ-line mutations in nonsyndromic pheochromocytoma. *N. Engl. J. Med.*, **346**, 1459–1466.
15. Duan, D.R., Pause, A., Burgess, W.H., Aso, T., Chen, D.Y., Garrett, K.P., Conaway, R.C., Conaway, J.W., Linehan, W.M. and Klausner, R.D. (1995)

- Inhibition of transcription elongation by the VHL tumor suppressor protein. *Science*, **269**, 1402–1406.
16. Kibel, A., Iliopoulos, O., DeCaprio, J.A. and Kaelin, W.G. Jr. (1995) Binding of the von Hippel-Lindau tumor suppressor protein to Elongin B and C. *Science*, **269**, 1444–1446.
 17. Kishida, T., Stackhouse, T.M., Chen, F., Lerman, M.I. and Zbar, B. (1995) Cellular proteins that bind the von Hippel-Lindau disease gene product: mapping of binding domains and the effect of missense mutations. *Cancer Res.*, **55**, 4544–4548.
 18. Schoenfeld, A.R., Davidowitz, E.J. and Burk, R.D. (2000) Elongin BC complex prevents degradation of von Hippel-Lindau tumor suppressor gene products. *Proc. Natl. Acad. Sci. USA*, **97**, 8507–8512.
 19. Sato, Y., Yoshizato, T., Shiraiishi, Y., Maekawa, S., Okuno, Y., Kamura, T., Shimamura, T., Sato-Otsubo, A., Nagae, G., Suzuki, H. *et al.* (2013) Integrated molecular analysis of clear-cell renal cell carcinoma. *Nat. Genet.*, **45**, 860–867.
 20. Stebbins, C.E., Kaelin, W.G. Jr. and Pavletich, N.P. (1999) Structure of the VHL-ElonginC-ElonginB complex: implications for VHL tumor suppressor function. *Science*, **284**, 455–461.
 21. Duan, D.R., Humphrey, J.S., Chen, D.Y., Weng, Y., Sukegawa, J., Lee, S., Gnarr, J.R., Linehan, W.M. and Klausner, R.D. (1995) Characterization of the VHL tumor suppressor gene product: localization, complex formation, and the effect of natural inactivating mutations. *Proc. Natl. Acad. Sci. USA*, **92**, 6459–6463.
 22. Kamura, T., Koepf, D.M., Conrad, M.N., Skowrya, D., Moreland, R.J., Iliopoulos, O., Lane, W.S., Kaelin, W.G. Jr., Elledge, S.J., Conaway, R.C. *et al.* (1999) Rbx1, a component of the VHL tumor suppressor complex and SCF ubiquitin ligase. *Science*, **284**, 657–661.
 23. Lonergan, K.M., Iliopoulos, O., Ohh, M., Kamura, T., Conaway, R.C., Conaway, J.W. and Kaelin, W.G. Jr. (1998) Regulation of hypoxia-inducible mRNAs by the von Hippel-Lindau tumor suppressor protein requires binding to complexes containing elongins B/C and Cul2. *Mol. Cell. Biol.*, **18**, 732–741.
 24. Kaelin, W.G. (2005) The von Hippel-Lindau tumor suppressor protein: roles in cancer and oxygen sensing. *Cold Spring Harb. Symp. Quant. Biol.*, **70**, 159–166.
 25. Kaelin, W.G. (2007) Von Hippel-Lindau disease. *Annu. Rev. Pathol.*, **2**, 145–173.
 26. Clifford, S.C., Cockman, M.E., Smallwood, A.C., Mole, D.R., Woodward, E.R., Maxwell, P.H., Ratcliffe, P.J. and Maher, E.R. (2001) Contrasting effects on HIF-1 α regulation by disease-causing pVHL mutations correlate with patterns of tumorigenesis in von Hippel-Lindau disease. *Hum. Mol. Genet.*, **10**, 1029–1038.
 27. Hoffman, M.A., Ohh, M., Yang, H., Klco, J.M., Ivan, M. and Kaelin, W.G. Jr. (2001) von Hippel-Lindau protein mutants linked to type 2C VHL disease preserve the ability to downregulate HIF. *Hum. Mol. Genet.*, **10**, 1019–1027.
 28. Gallou, C., Chauveau, D., Richard, S., Joly, D., Giraud, S., Olschwang, S., Martin, N., Saquet, C., Chretien, Y., Mejean, A. *et al.* (2004) Genotype-phenotype correlation in von Hippel-Lindau families with renal lesions. *Hum. Mutat.*, **24**, 215–224.
 29. Ong, K.R., Woodward, E.R., Killick, P., Lim, C., Macdonald, F. and Maher, E.R. (2007) Genotype-phenotype correlations in von Hippel-Lindau disease. *Hum. Mutat.*, **28**, 143–149.
 30. Gallou, C., Joly, D., Mejean, A., Staroz, F., Martin, N., Tarlet, G., Orfanelli, M.T., Bouvier, R., Droz, D., Chretien, Y. *et al.* (1999) Mutations of the VHL gene in sporadic renal cell carcinoma: definition of a risk factor for VHL patients to develop an RCC. *Hum. Mutat.*, **13**, 464–475.
 31. Zatyka, M., da Silva, N.F., Clifford, S.C., Morris, M.R., Wiesener, M.S., Eckardt, K.U., Houlston, R.S., Richards, F.M., Latif, F. and Maher, E.R. (2002) Identification of cyclin D1 and other novel targets for the von Hippel-Lindau tumor suppressor gene by expression array analysis and investigation of cyclin D1 genotype as a modifier in von Hippel-Lindau disease. *Cancer Res.*, **62**, 3803–3811.
 32. Ricketts, C., Zeegers, M.P., Lubinski, J. and Maher, E.R. (2009) Analysis of germline variants in CDH1, IGFBP3, MMP1, MMP3, STK15 and VEGF in familial and sporadic renal cell carcinoma. *PLoS ONE*, **4**, e6037.
 33. Webster, A.R., Richards, F.M., MacDonald, F.E., Moore, A.T. and Maher, E.R. (1998) An analysis of phenotypic variation in the familial cancer syndrome von Hippel-Lindau disease: evidence for modifier effects. *Am. J. Hum. Genet.*, **63**, 1025–1035.
 34. Foster, K., Prowse, A., van den Berg, A., Fleming, S., Hulsbeek, M.M., Crossey, P.A., Richards, F.M., Cairns, P., Affara, N.A., Ferguson-Smith, M.A. *et al.* (1994) Somatic mutations of the von Hippel-Lindau disease tumour suppressor gene in non-familial clear cell renal carcinoma. *Hum. Mol. Genet.*, **3**, 2169–2173.
 35. Gnarr, J.R., Tory, K., Weng, Y., Schmidt, L., Wei, M.H., Li, H., Latif, F., Liu, S., Chen, F., Duh, F.M. *et al.* (1994) Mutations of the VHL tumour suppressor gene in renal carcinoma. *Nat. Genet.*, **7**, 85–90.
 36. Shuin, T., Kondo, K., Kaneko, S., Sakai, N., Yao, M., Hosaka, M., Kanno, H., Ito, S. and Yamamoto, I. (1995) [Results of mutation analyses of von Hippel-Lindau disease gene in Japanese patients: comparison with results in United States and United Kingdom]. *Hinyokika Kyo*, **41**, 703–707.
 37. Whaley, J.M., Naglich, J., Gelbert, L., Hsia, Y.E., Lamiell, J.M., Green, J.S., Collins, D., Neumann, H.P., Laidlaw, J., Li, F.P. *et al.* (1994) Germ-line mutations in the von Hippel-Lindau tumor-suppressor gene are similar to somatic von Hippel-Lindau aberrations in sporadic renal cell carcinoma. *Am. J. Hum. Genet.*, **55**, 1092–1102.
 38. COSMIC. Catalogue of Somatic Mutations in Cancer. *Catalogue of Somatic Mutations in Cancer*. <http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/>.
 39. Gossage, L. and Eisen, T. (2010) Alterations in VHL as potential biomarkers in renal-cell carcinoma. *Nat. Rev. Clin. Oncol.*, **7**, 277–288.
 40. Dehouck, Y., Grosfils, A., Folch, B., Gilis, D., Bogaerts, P. and Rooman, M. (2009) Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. *Bioinformatics*, **25**, 2537–2543.
 41. Dehouck, Y., Kwasigroch, J.M., Rooman, M. and Gilis, D. (2013) BeAtMuSiC: prediction of changes in protein-protein binding affinity on mutations. *Nucleic Acids Res.*, **41**, W333–W339.
 42. Olivera-Nappa, A., Andrews, B.A. and Asenjo, J.A. (2011) Mutagenesis Objective Search and Selection Tool (MOSST): an algorithm to predict structure-function related mutations in proteins. *BMC Bioinformatics*, **12**, 122.
 43. Pires, D.E., Ascher, D.B. and Blundell, T.L. (2013) mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics*, **30**, 335–342.
 44. Topham, C.M., Srinivasan, N. and Blundell, T.L. (1997) Prediction of the stability of protein mutants based on structural environment-dependent amino acid substitution and propensity tables. *Protein Eng.*, **10**, 7–21.
 45. Worth, C.L., Preissner, R. and Blundell, T.L. (2011) SDM – a server for predicting effects of mutations on protein stability and malfunction. *Nucleic Acids Res.*, **39**, W215–W222.
 46. Frousios, K., Iliopoulos, C.S., Schlitt, T. and Simpson, M.A. (2013) Predicting the functional consequences of non-synonymous DNA sequence variants – evaluation of bioinformatics tools and development of a consensus strategy. *Genomics*, **102**, 223–228.
 47. Kumar, A., Rajendran, V., Sethumadhavan, R., Shukla, P., Tiwari, S. and Purohit, R. (2013) Computational SNP analysis: current approaches and future prospects. *Cell Biochem. Biophys.*, **68**, 233–239.
 48. Quinlan, J.R. (1992) Learning with continuous classes. *Artificial Intelligence '92: Proceedings of the 5th Australian joint Conference on Artificial Intelligence*.
 49. Breiman, L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.
 50. Pires, D.E., de Melo-Minardi, R.C., dos Santos, M.A., da Silveira, C.H., Santoro, M.M. and Meira, W. Jr. (2011) Cutoff Scanning Matrix (CSM): structural classification and function prediction by protein inter-residue distance patterns. *BMC Genomics*, **12**(Suppl. 4), S12.
 51. Pires, D.E., de Melo-Minardi, R.C., da Silveira, C.H., Campos, F.F. and Meira, W. Jr. (2013) aCSM: noise-free graph-based signatures to large-scale receptor-based ligand prediction. *Bioinformatics*, **29**, 855–861.
 52. Forman, J.R., Worth, C.L., Bickerton, G.R., Eisen, T.G. and Blundell, T.L. (2009) Structural bioinformatics mutation analysis reveals genotype-phenotype correlations in von Hippel-Lindau disease and suggests molecular mechanisms of tumorigenesis. *Proteins*, **77**, 84–96.
 53. Kim, J., Seong, M.W., Lee, K., Choi, H., Ku, E., Bae, J., Park, S., Choi, S., Kim, S. and Shin, C. (2013) Germline mutations and genotype-phenotype correlations in patients with apparently sporadic pheochromocytoma/paraganglioma in Korea. *Clin. Genet*. doi: /10.1111/cge.12304/.
 54. Sjuursen, W., Halvorsen, H., Hofslie, E., Bachke, S., Berge, A., Engebretsen, L.F., Falkmer, S.E., Falkmer, U.G. and Varhaug, J.E. (2013) Mutation screening in a Norwegian cohort with pheochromocytoma. *Fam. Cancer*, **12**, 529–535.
 55. D'Elia, A.V., Grimaldi, F., Pizzolitto, S., De Maglio, G., Bregant, E., Passon, N., Franzoni, A., Verrienti, A., Tamburrano, G., Durante, C. *et al.* (2013) A new germline VHL gene mutation in three patients with apparently sporadic pheochromocytoma. *Clin. Endocrinol. (Oxf.)*, **78**, 391–397.

56. Kumar, P., Henikoff, S. and Ng, P.C. (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.*, **4**, 1073–1081.
57. Shen, C. and Kaelin, W.G. Jr (2013) The VHL/HIF axis in clear cell renal carcinoma. *Semin. Cancer Biol.*, **23**, 18–25.
58. Gossage, L., Murtaza, M., Slatter, A.F., Lichtenstein, C.P., Warren, A., Haynes, B., Marass, F., Roberts, I., Shanahan, S.J., Claas, A. *et al.* (2014) Clinical and pathological impact of VHL, PBRM1, BAP1, SETD2, KDM6A, and JARID1c in clear cell renal cell carcinoma. *Genes Chromosomes Cancer*, **53**, 38–51.
59. Woodward, S.H., Kaloupek, D.G., Streeter, C.C., Kimble, M.O., Reiss, A.L., Eliez, S., Wald, L.L., Renshaw, P.F., Frederick, B.B., Lane, B. *et al.* (2007) Brain, skull, and cerebrospinal fluid volumes in adult posttraumatic stress disorder. *J. Trauma. Stress*, **20**, 763–774.
60. Hickey, M.M., Lam, J.C., Bezman, N.A., Rathmell, W.K. and Simon, M.C. (2007) von Hippel-Lindau mutation in mice recapitulates Chuvash polycythemia via hypoxia-inducible factor-2alpha signaling and splenic erythropoiesis. *J. Clin. Invest.*, **117**, 3879–3889.
61. van Rooijen, E., Voest, E.E., Logister, I., Korving, J., Schwerte, T., Schulte-Merker, S., Giles, R.H. and van Eeden, F.J. (2009) Zebrafish mutants in the von Hippel-Lindau tumor suppressor display a hypoxic response and recapitulate key aspects of Chuvash polycythemia. *Blood*, **113**, 6449–6460.
62. Ruiz-Llorente, S., Bravo, J., Cebrian, A., Cascon, A., Pollan, M., Telleria, D., Leton, R., Urioste, M., Rodriguez-Lopez, R., de Campos, J.M. *et al.* (2004) Genetic characterization and structural analysis of VHL Spanish families to define genotype-phenotype correlations. *Hum. Mutat.*, **23**, 160–169.
63. Knauth, K., Bex, C., Jemth, P. and Buchberger, A. (2006) Renal cell carcinoma risk in type 2 von Hippel-Lindau disease correlates with defects in pVHL stability and HIF-1alpha interactions. *Oncogene*, **25**, 370–377.
64. Maher, E.R., Yates, J.R., Harries, R., Benjamin, C., Harris, R., Moore, A.T. and Ferguson-Smith, M.A. (1990) Clinical features and natural history of von Hippel-Lindau disease. *Q. J. Med.*, **77**, 1151–1163.
65. Lonser, R.R., Glenn, G.M., Walther, M., Chew, E.Y., Libutti, S.K., Linehan, W.M. and Oldfield, E.H. (2003) von Hippel-Lindau disease. *Lancet*, **361**, 2059–2067.
66. Mandriota, S.J., Turner, K.J., Davies, D.R., Murray, P.G., Morgan, N.V., Sowter, H.M., Wykoff, C.C., Maher, E.R., Harris, A.L., Ratcliffe, P.J. *et al.* (2002) HIF activation identifies early lesions in VHL kidneys: evidence for site-specific tumor suppressor function in the nephron. *Cancer Cell*, **1**, 459–468.
67. Rankin, E.B., Tomaszewski, J.E. and Haase, V.H. (2006) Renal cyst development in mice with conditional inactivation of the von Hippel-Lindau tumor suppressor. *Cancer Res.*, **66**, 2576–2583.
68. National Center for Biotechnology Information, d. <http://www.ncbi.nlm.nih.gov/snp> (accessed on 3 September 2013).
69. Kumar, M.D., Bava, K.A., Gromiha, M.M., Prabakaran, P., Kitajima, K., Uedaira, H. and Sarai, A. (2006) ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. *Nucleic Acids Res.*, **34**, D204–D206.
70. Moal, I.H. and Fernandez-Recio, J. (2012) SKEMPI: a structural kinetic and energetic database of mutant protein interactions and its use in empirical models. *Bioinformatics*, **28**, 2600–2607.