PorphyriaDB - *a cloud-driven genetic database for the acute hepatic porphyrias.*

Micheál Mac Aogáin[1,2], Brendan Lawlor[3], Sarah Savage[1], Elena Walsh[1], Nadia Brazil[1], Erum Rasheed[1], Thomas Cronin[1], Vivion Crowley[1]

[1]Biochemical Genetics Laboratory, Biochemistry Department, Central Pathology Laboratory, St James's Hospital, Dublin 8.

[2]School of Medicine, Trinity College Dublin, St James's Hospital, Dublin 8.

[3]Department of Computer Science, Munster Technological University, Cork, Ireland.

Genetic diagnosis of the acute hepatic porphyria

➢The acute hepatic porphyrias (AHPs) are autosomal dominant disorders that include acute intermittent porphyria (AIP), variegate porphyria (VP) and hereditary coproporphyria (HP) and can clinically manifest with acute neuropathic episodes causing significant morbidity (Balwani and Desnick, 2012).

➢Pathogenesis is related to variants in genes of the haem biosynthetic pathway (Figure 1A), hence molecular genetic analysis is central to confirmatory diagnosis and kindred follow-up for identifying porphyria susceptibility. Genetic services for the acute porphyrias enable the diagnosis of genetic susceptibility to AIP, VP and HCP by assessment of three key loci; *HMBS*, *PPOX* and *CPOX* (Figure 1B).

➢For clinical diagnostics, functional characterisation of rare missense variants can facilitate definitive confirmation of their pathogenic status and risk. However, such characterisation can be difficult to achieve and consequently there is a heavy reliance on *in silico* predictors to determine clinical relevance of these variants, in the absence of functional characterisation or clinical description.

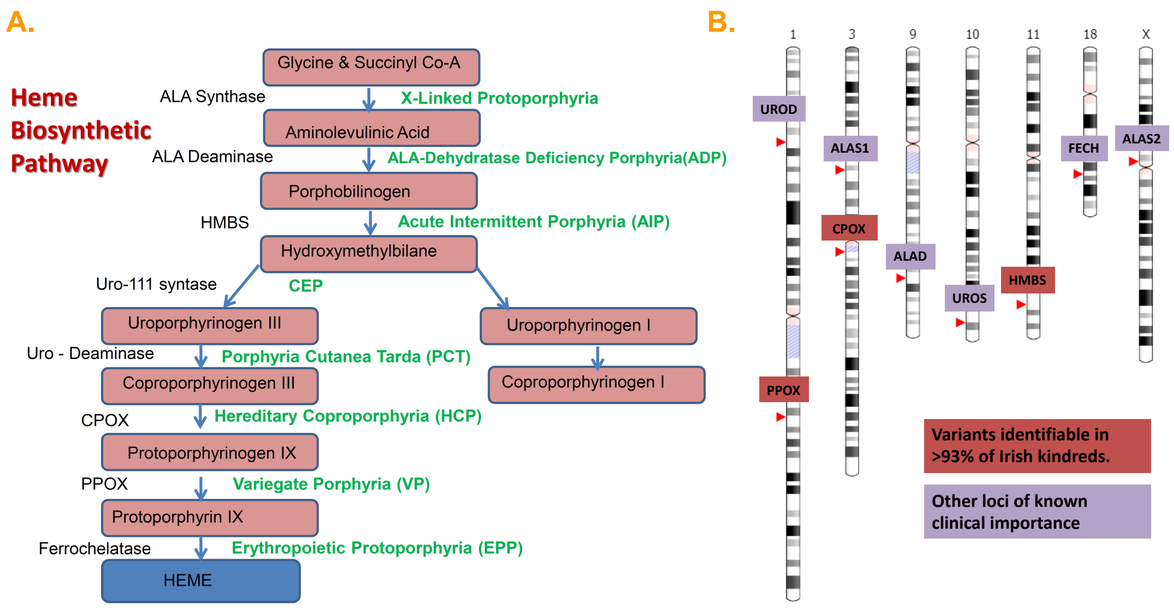


Figure 1. (A) Schematic overview of heme biosynthetic pathway, including key components Hydroxymethylbilane synthase – *HMBS,* Ċoproporphyrinogen oxidase – *CPOX*, and Protoporphyrinogen oxidase – *PPOX*. (B) The chromosomal location of key heme biosynthetic genes *HMBS, CPOX and PPOX* on chromosomes 1, 3 and 11. These three loci harbour variants in >93% of Irish kindreds (Savage *et al.* 2020a) and form the core of our accredited genetic diagnostics service.

➢Variant interpretation, in particular the analysis of uncharacterised missense variants, is further impeded by the lack of a community-curated Porphyria genetic database (Chen et al. 2019).

A curated genetic database of AHP missense mutations

➢Non-synonymous missense variants are commonly identified in porphyria patients (Figure 2A), which may confound genetic reporting where uncharacterised variants are involved.

➢To facilitate automation of our reporting workflow, we sought to curate an in-house reference database of missense variants in *HMBS*, *PPOX*, and *CPOX*.

➢The database currently contains 309 curated entries, and includes variants present in Clinvar (25%), HGMD (88%) or from other sources (18%).

➢ Several predictors and metapredictors (Polyphen-2, SIFT, Consurf, Mutation assessor, Provean, CONDEL, FATHMM, CADD and REVEL) were evaluated in the assessment of all variant entries.

➢Previous work has established a consensus approach in the classification of non-synonymous missense variants (Figure 2B) (Savage 2020b), using this database.

➢Each database entry contains defined variant information and a list of prediction scores generated by multiple predictors and metapredictors, as well as our SJH consensus *in silico* prediction score.

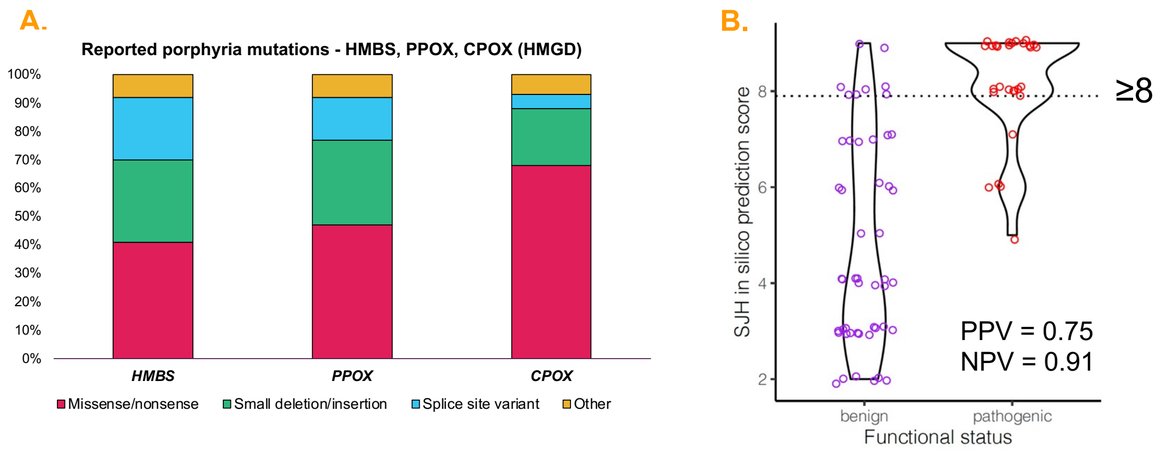


Figure 2. (A) Variant annotation of identified porphyria mutations highlight a significant proportion of missense mutations. (B) Applying a consensus *in silico* prediction approach, assessing the outputs of several predictors and metapredictors benchmarked against functionally characterised mutations (*HMBS*), allows prediction of pathogenic variants with a high negative predictive value (cut-off ≥8)

Deployment of a cloud-native clinical genetic database

➢ In order to codify our analytical decision rules in the form of a reproducible and auditable reference database, a cloud-native framework embodied by an infrastructure as a service (IaaS) approach was engineered using Amazon Web Services (AWS) accessible (Figure 3).

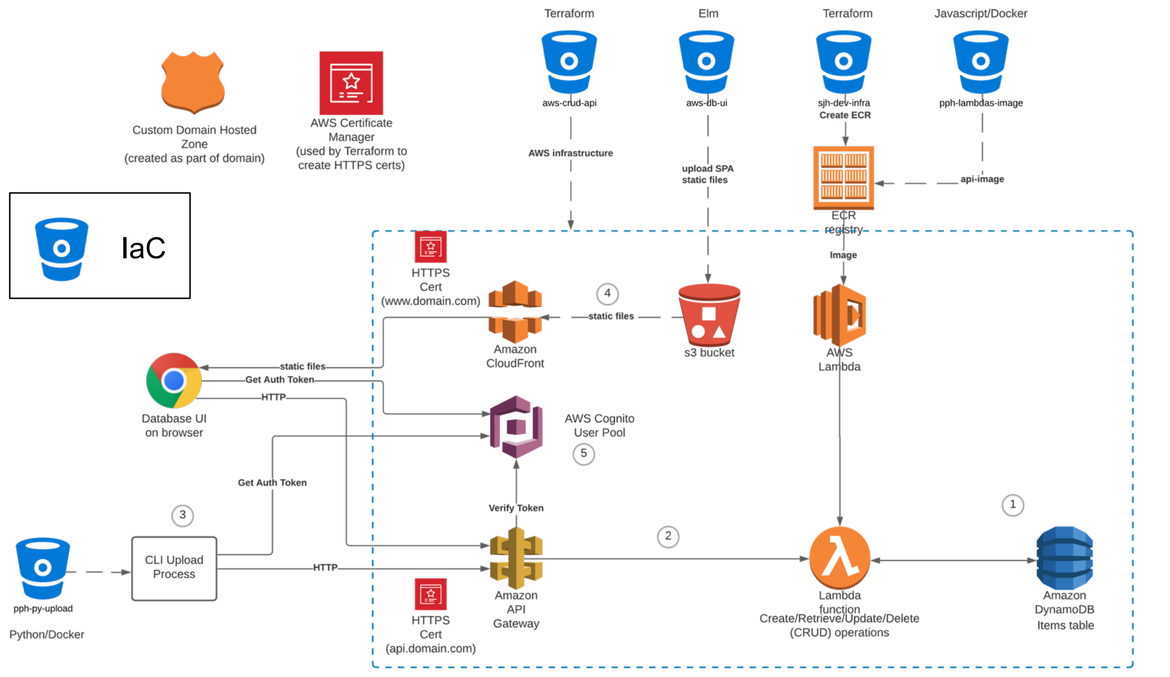


Figure 3. Backend architecture for deployment of a could-driven genetic database for https://www.porphyriadb.com/. This diagram shows the AWS cloud products that combine to create the infrastructure for the database. There are five main components to that infrastructure: 1: The database (DynamoDB) which contains the biologically useful data, 2: The API Gateway which provides programmatic access to that data via a lambda function, 3: The external CLI upload process that runs in the lab to upload/update the data to the database via the API, 4:The web application that gives human users search and access to the data, via the API 5: The user database (Cognito) that manages user registration and login. This diagram also shows the role of the infrastructure as code (IaC) approach (represented by blue bitbucket icons, outside the broken rectangle). Instead of manually assembling the AWS cloud products using the console, the IaC makes this assembly reproducible and transparent. Changes and additions are made by changing the underlying code repositories and running them through a Continuous Deployment pipeline.

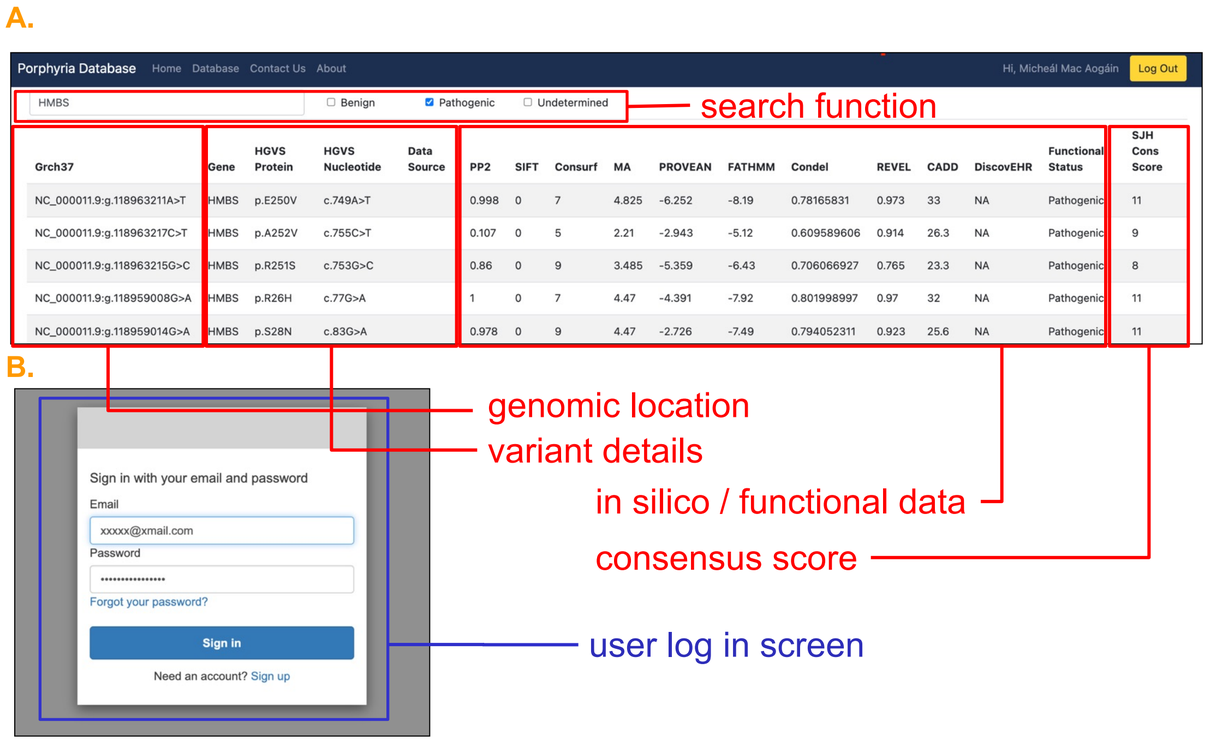


Figure 4. Screen shots from https://www.porphyriadb.com/. (A) Online searchable database indicating genomic location, variant information, *in silico* prediction scores and functional data as well as our in-house consensus score for variant classification, all of which can be navigated via search function. (B) user login screen requiring email and passwords securely managed by AWS Cognito.

Summary

➢*In silico* analysis of clinical variants forms a core component of porphyria genetic diagnostics.

➢Emerging cloud-driven technologies facilitate decentralization of clinical genetic database curation for specific diseases, populations and reference laboratories. This is of particular importance for rare diseases where no dedicated databases exist.

➢Deployment of local cloud-native databases, via an IaC approach, allows for greater reproducibility and enables novel feature implementation (e.g., user variant submission) via stable and fully auditable protocols.

➢This facilitates reproducible implementation and capture of variant data underpinning clinical management of rare diseases such as the acute hepatic porphyrias, with potential to enhance diagnostic accuracy.

References

Balwani and Desnick. The porphyrias: advances in diagnosis and treatment. Blood. 2012 Nov 29;120(23):4496-504.

Savage et al. 2020a. The Molecular Basis of Acute Porphyria and familial Porphyria Cutanea Tarda in the Republic of Ireland – an update. Ulster Med J. 2020 Jan; 89(1): 47–63. OP08.

Chen et al 2019. International Porphyria Molecular Diagnostic Collaborative: an evidence-based database of verified pathogenic and benign variants for the porphyrias. Genet Med. 2019 Nov; 21(11): 2605–2613.

Savage et al. 2020b. *In silico* predictive analysis and evaluation of missense variants in haem biosynthesis genes associated with the acute hepatic porphyrias. Ir J Med Sci. 2021 Feb 20 : 1–16. P16.

ISHG - 2020

In silico predictive analysis and evaluation of missense variants in haem biosynthesis genes associated with the acute hepatic porphyrias.

Sarah Savage[1], Micheál Mac Aogáin[1,2], Elena Walsh[1], Nadia Brazil[1], Erum Rasheed[1], Arsala Rashid[1], Vivion E.F. Crowley[1,2].

[1]Biochemical Genetics Laboratory, Biochemistry Department, Central Pathology Laboratory, St James's Hospital, Dublin 8.

[2]School of Medicine, Trinity College Dublin, St James's Hospital, Dublin 8.

Background

➢The acute porphyrias are autosomal dominant disorders that include acute intermittent porphyria (AIP), variegate porphyria (VP) and hereditary coproporphyria (HP) and can clinically manifest with acute neuropathic episodes causing significant morbidity.

➢Pathogenesis is related to variants in genes of the haem biosynthetic pathway (Figure 1A). and therefore molecular genetic analysis has become an important component in kindred follow-up for identifying porphyria susceptibility. The Biochemistry Department, St James's Hospital, Dublin, has established a molecular diagnostic service for the acute porphyrias, enabling the diagnosis of genetic susceptibility to AIP, VP and HCP by assessment of the *HMBS*, *PPOX* and *CPOX* loci (Figure 1B).

➢For clinical diagnostic laboratories the functional characterisation of rare genetic missense variants identified during analysis can facilitate definitive confirmation of their pathogenic status. However, such characterisation can be difficult to achieve and consequently there is a heavy reliance on *in-silico* predictors to determine clinical relevance of these variants.

➢Therefore, to optimise diagnostic potential, it is important, where possible, to assess the performance of *in-silico* predictors relative to functionally characterised datasets. This can enable a gene specific selection of *in-silico* tools to enhance the reporting of genetic results.

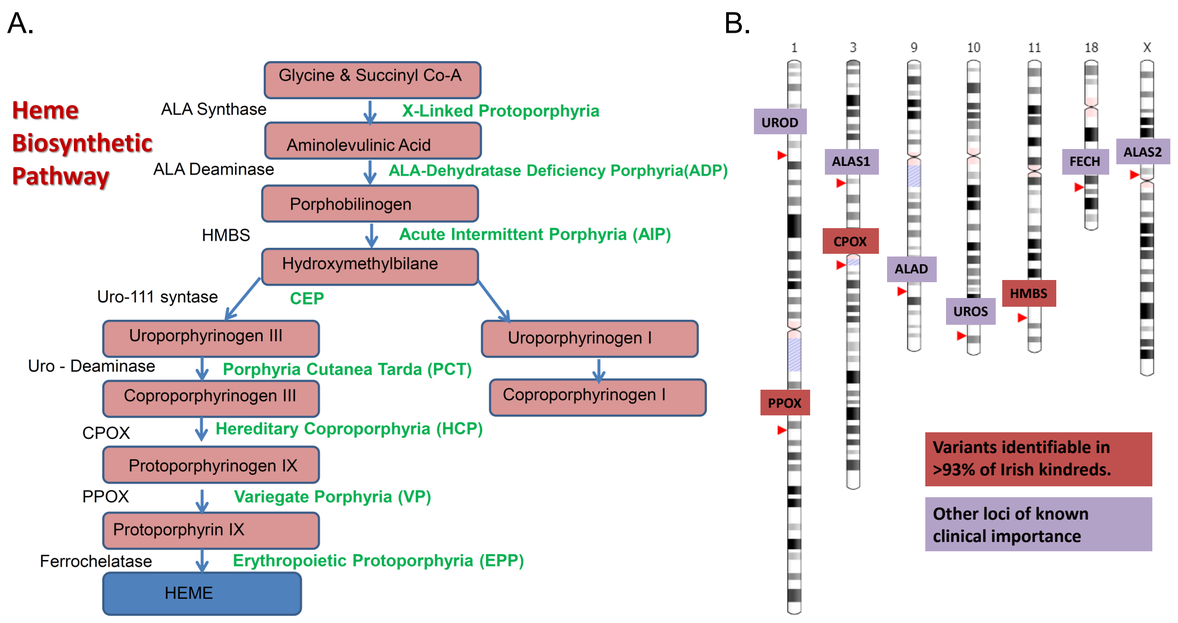


Figure 1. (A) Schematic overview of heme biosynthetic pathway, including key components Hydroxymethylbilane synthase – *HMBS,* Coproporphyrinogen oxidase – *CPOX*, and Protoporphyrinogen oxidase – *PPOX*. (B) The chromosomal location of key heme biosynthetic genes *HMBS, CPOX and PPOX* on chromosomes 1, 3 and 11. These loci harbour variants in >93% of Irish kindreds (Savage *et al.* 2020).

Genetic analysis of *HMBS, CPOX* and *PPOX* in Irish porphyria kindreds

➢To date 25 different mutations have been characterised among 31 different porphyria kindreds, including families who have immigrated to Ireland, with mutations spanning exonic and intronic regions of *HMBS, CPOX* and *PPOX*. The spectrum of mutations for all three porphyrias are indicated in Table 1 (Savage *et al.* 2020).

Table 1. Mutations identified in Irish porphyria kindreds.

Missense (40%)	Gene	Nucleotide substitution	Reference
R26H	HMBS	c.77 G>A	Llewellyn *et al.*
R26C	HMBS	c.76 C>T	Kauppinen *et al.*
G111R (*)	HMBS	c.331 G>A	Gu XF *et al.*
L137R	**HMBS**	**c.410 T>G**	**Novel mutation, SJH**
C247R	HMBS	c.739 T>C	Mgone CS *et al.*
R332Q	**CPOX**	**c.995 G>A**	**Gorman *et al.* SJH**
R332W	**CPOX**	**c.994 C>T**	**Novel Mutation, SJH**
R59W (^)	PPOX	c.175 C>T	Meissner *et al.*
A150D	**PPOX**	**c.449 C>A**	**Novel Mutation, SJH**
R152C	PPOX	c.454 C>T	Frank *et al.*
Nonsense (16%)	**Gene**	**Nucleotide substitution**	**Reference**
W283X	HMBS	c.847 G>A	Mgone CS *et al.*
Q375X (*)	PPOX	c.1123 C>T	Corrigall *et al.*
W427X	PPOX	c.1281 G>A	Whatley *et al.*
Q435X	PPOX	c.1303 C>T	Whatley *et al.*
Small Indels (24%)	**Gene**	**Nucleotide substitution**	**Reference**
c.131insCAGCG	CPOX	c.131insCAGCG	Lamoril *et al.*
c.398delC	**CPOX**	**c.398delC**	**Novel mutation, SJH**
c.1291_2insTG (^)	**CPOX**	**c.1291_2insTG**	**Novel mutation, SJH**
c.199delC	**PPOX**	**c.199delC**	**Novel mutation, SJH**
c.157_160delATCT	PPOX	c.157_160delATCT	Whatley *et al.*
c.1337_8 delTG	**PPOX**	**c.1337_8 delTG**	**Novel mutation, SJH**
Splicing Defects (16%)	**Splice site**	**Nucleotide substitution**	**Reference**
c.33 + 666 A>G	**Intron 1**	**c.33 + 666 A>G**	**Novel Mutation, SJH**
c.160 + 1 G>A	Exon 3 donor	c.160 + 1 G>A	Whatley *et al.*
c.267 - 1 G>C	**Exon 6 acceptor**	**c.267 - 1 G>C**	**Novel Mutation, SJH**
c.339 - 1 G>A	Exon 5 acceptor	c.339 - 1 G>A	Von und zu Fraunberg
Large Deletion (4%)	**Gene**	**Nucleotide substitution**	**Reference**
c.33+1356 to c.619+702	HMBS/3-15	c.33+1356 to c.619+702	Whatley et al. (2009)
	H2AFX		
	DPAGT1/9-5		

(*) identified in 2 kindreds, (^) identified in 3 kindreds. Novel variants highlighted in **red typeface.**

In silico predictive analysis

➢ Several predictors and metapredictors (Polyphen-2, SIFT, Consurf, Mutation assessor, Provean, CONDEL, FATHMM, CADD and REVEL) were evaluated in the assessment of known porphyria mutations in *HMBS, PPOX* and *CPOX*.

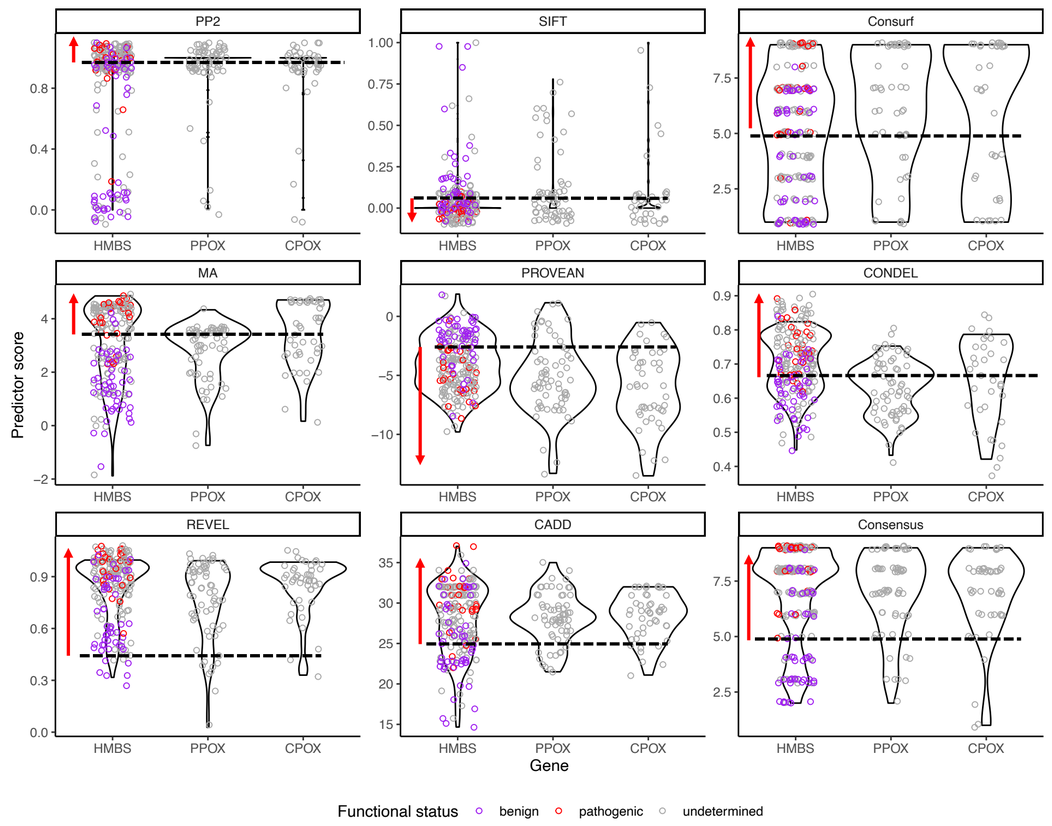
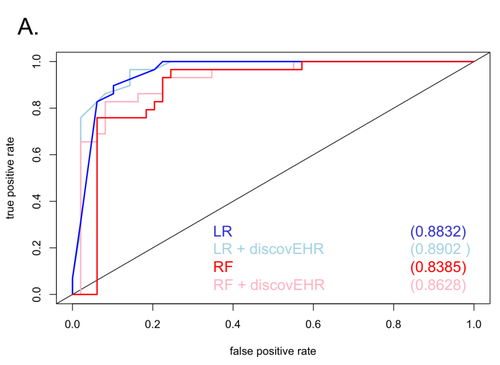


Figure 2. Distribution of *in silico* prediction scores for 329 reported missense mutations in *HMBS*, *PPOX* and *CPOX* derived using predictor and metapredictor algorithms. Functional data on a subset of 78 functionally characterised variants in *HMBS* (Chen *et al.* 2016, Lenglet *et al.* 2018) are colour indicated (benign – purple, pathogenic – red, undetermined – grey). Broken lines indicate applied prediction score classification thresholds while red arrows denote directionality of the pathogenicity call.

➢Using a subset of functionally characterised *HMBS* genetic variants (n=78), predictive accuracy of individual and combined (consensus) *in-silico* scoring approaches was further assessed using binary predictions, logistic regression and random forest models. (Figure 3).



B.

		Performance statistics						Log. Reg.	R.f.
		TP	TN	FP	FN	PPV	NPV	(P>|z|)	(Increased node purity)
Model	Log. Reg.	25	45	4	4	0.862	0.918	-	-
	R.f.	24	45	4	5	0.857	0.900	-	-
Predictors	Mutation assessor	22	46	3	7	0.880	0.868	0.008	4.708
	Provean	28	36	13	1	0.683	0.973	0.996	3.419
	SIFT	29	25	24	0	0.547	1.000	0.995	1.265
	Polyphen-2	26	29	20	3	0.565	0.906	0.626	0.993
	CADD	26	25	24	3	0.520	0.893	0.996	0.492
	Consurf	19	28	21	10	0.475	0.737	0.807	0.356
	REVEL	29	16	33	0	0.468	1.000	0.997	0.265
	CONDEL	29	1	48	1	0.377	1.000	0.999	0.000
Consensus score	Cons.	28	30	19	1	0.596	0.968	-	-

Figure 3. Accuracy of *in-silico* predictions for functionally characterised missense variants in *HMBS* (n=78). (A) Receiver operating characteristic (ROC) curves for logistic regression (LR - blue) and random forest (RF - red) analysis assessing combined performance of 8 *in-silico* prediction scores in the classification of functionally characterised *HMBS* variants. Additional ROC curves incorporating variant frequency data from the DiscovEHR cohort (LR + discovEHR – light blue, RF + discovEHR – pink) are also indicated with balanced accuracy indicated in parenthesis. (B) Summary table of model performance statistics. TP – true positives, TN – true negative, FP – false positives, FN – false negatives, PPV – positive predictive value, NPV – negative predictive value. The performance of a bespoke consensus score (Cons.) is also evaluated.

➢Mutation assessor, Provean, SIFT and Polyphen-2 appeared to perform well with various trade offs in PPV and NPV. Despite representing the most significant predictor in both regression and random forest models, Mutation Assessor misclassified one functionally characterised clinical variant (p.G111R, medium functional impact) when applied to our clinically observed variants highlighting the importance of multiple prediction comparisons (Figure 4).

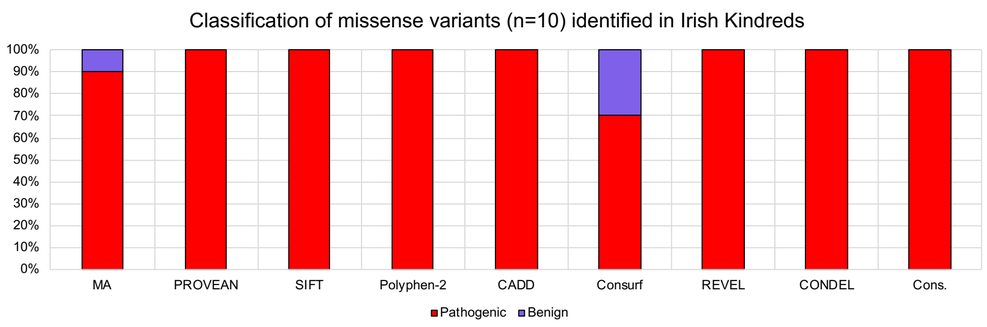


Figure 4. Classification of missense variants observed in Irish kindreds based on individual *in silico* prediction scores.

Summary

•*In-silico* analysis of clinical variants forms a core component of porphyria genetic diagnostics. Benchmarking their performance relative to functionally characterised data sets can determine overall ability to provide accurate variant classification.

•The data presented confirms that multiple prediction scoring systems improve accuracy compared to reliance on individual predictor scores, although absolute rule-out or rule-in thresholds for pathogenic or non-pathogenic variants may be defined using single predictors e.g. SIFT

•Furthermore, integration of large genomic datasets (i.e. DiscoverEHR) demonstrates the potential for improved classification accuracy.

References

1.Savage et al. 2020. The Molecular Basis of Acute Porphyria and familial Porphyria Cutanea Tarda in the Republic of Ireland – an update. Ulster Med J. 2020 Jan; 89(1): 47–63. OP08.

2.Chen et. al. 2016. Acute Intermittent Porphyria: Predicted Pathogenicity of HMBS Variants Indicates Extremely Low Penetrance of the Autosomal Dominant Disease. Hum Mutat. 2016 Nov;37(11):1215-1222.

3.Lenglet et al. 2018. From a dominant to an oligogenic model of inheritance with environmental modifiers in acute intermittent porphyria. Hum Mol Genet . 2018 Apr 1;27(7):1164-1173.