HemaSphere

EHA EUROPEAN HEMATOLOGY ASSOCIATION

Letter
Open Access

IthaPhen: An Interactive Database of Genotype-Phenotype Data for Hemoglobinopathies

Maria Xenophontos[1,*], Anna Minaidou[1,*], Coralea Stephanou[1], Stella Tamana[1], Marina Kleanthous[1,**], Petros Kountouris[1,**]

Correspondence: Petros Kountouris (petrosk@cing.ac.cy).

Hemoglobinopathies are the commonest monogenic diseases, with ≈6% of the world's population carrying a pathogenic globin gene variant.[1,2] To date, >2400 disease-causing variants have been reported in the IthaGenes database,[3,4] with various severity and clinical presentation.[5] Nevertheless, most of these variants are rare and their functional impact is unclear. The use of case-level data and subsequent genotype-phenotype correlations can enhance our understanding of the pathophysiological mechanisms of hemoglobinopathies and can assist their diagnosis and clinical management. In fact, the evaluation of case-level data is a critical component of both the general framework for standardized variant classification[6] by the American College of Medical Genetics and Genomics and the Association for Molecular Pathology (ACMG/AMP) and the specified rules currently developed by the ClinGen Hemoglobinopathy Variant Curation Expert Panel (VCEP).[7] However, large amounts of case-level data remain unexploited as they reside in the databases of diagnostic laboratories and/or the publications they produce. The utility of genome-wide databases such as ClinVar,[8] DECIPHER,[9] and the ClinGen Variant Curation Interface[10] in hemoglobinopathies is also limited due to the low number of variants for the globin genes. Notably, hemoglobinopathy-specific databases, such as HbVar[11] and IthaGenes, provide limited phenotypic information for a small fraction of curated variants in tabular or text format, impeding their use for further analysis and variant interpretation. Hence, the development of a comprehensive, curated genotype-phenotype database is an urgent need to bring together available molecular and clinical information published in scientific literature or residing in diagnostic and clinical laboratories.

Herein, we present IthaPhen, a manually curated, searchable, and intuitive database of public anonymized case reports linked to hemoglobinopathies. IthaPhen integrates genotypic data with hematological, biochemical, histological, and clinical data and, through search queries of user-defined genotypes, returns both the primary case-level data and aggregated, visualized reports.

We considered globin gene variants annotated in IthaGenes[3] and collected individual case reports with such variants, through a rigorous literature search. We excluded articles providing only aggregated descriptions of specific cohorts, as IthaPhen does not currently support cohort descriptions. In addition, direct contributions of anonymized, previously unpublished case-level data were accepted through the ITHANET submission process and contributions were acknowledged throughout the ITHANET portal.

We collected demographic information, including age, sex, ethnicity and, if available, the individual's relationship with the proband. Furthermore, a rigorous literature search was performed for the selection of phenotypic parameters that are frequently reported and more relevant to hemoglobinopathies. The phenotypic parameters were expanded accordingly throughout the development period. Currently, IthaPhen collects a total of 86 phenotypic parameters, organized in 13 categories: blood transfusion (2), hematological investigation (12), biochemical investigation (3), biosynthetic ratio (4), hemoglobin profile (9), erythrocyte inclusions (2), hemoglobin properties (7), red blood cell morphology/microscopy (11), anemia (6), diagnostic imaging (2), clinical findings (25), clinical classification (2), and the miscellaneous other test category. The complete list of data parameters and their definitions are available in Suppl. File 1. Moreover, the European Molecular Genetics Quality Network (EMQN) best practice guidelines[12] have been integrated with IthaPhen to provide recommendations about prenatal or preimplantation genetic diagnosis, depending on the user-defined genotype query.

IthaPhen is developed using a relational database management system based on MySQL and PHP and is fully integrated and interconnected with the existing infrastructure of the ITHANET portal.[3,13] The content curation of IthaPhen is included in the existing ITHANET curation strategy with weekly updates, as previously described[3,4] and documented in the ITHANET FAQ page (https://www.ithanet.eu/home/faqs), thus ensuring the sustainability and relevance of the database.

At the time of writing, IthaPhen contains data for 3997 cases (3853 published and 144 contributed), collected from 764 publications and 28 contributors, while it stores information for 588 families (ie, ≥2 related cases). These cases cover 850 unique globin gene variants, representing over 34% of all variants annotated as causative in IthaGenes. IthaPhen provides a diverse set of case-level data in terms of demographics, genotype, and phenotype. Specifically, Figure 1 illustrates the distribution of

[1]Molecular Genetics Thalassaemia Department, The Cyprus Institute of Neurology and Genetics, Nicosia, Cyprus
*MX and AM shared joint first authorship.
**MK and PK shared joint last authorship.
Supplemental digital content is available for this article.

HemaSphere (2023) 7:7(e922).
http://dx.doi.org/10.1097/HS9.0000000000000922.
Received: November 7, 2022 / Accepted: May 31, 2023

cases according to the reported ethnicity; grouped by the top-level definitions of the Human Ancestry Ontology[14] (Figure 1A), sex and age (Figure 1B), genotypes (Figure 1C), and phenotypic annotation (Figure 1D–1E). Most cases are examined at least with respect to the hemoglobin profile (eg, hemoglobin [Hb] A2) or other hematological aspects (eg, Hb, mean corpuscular volume [MCV], and mean corpuscular hemoglobin [MCH]), but clinical characterization is also prevalent.

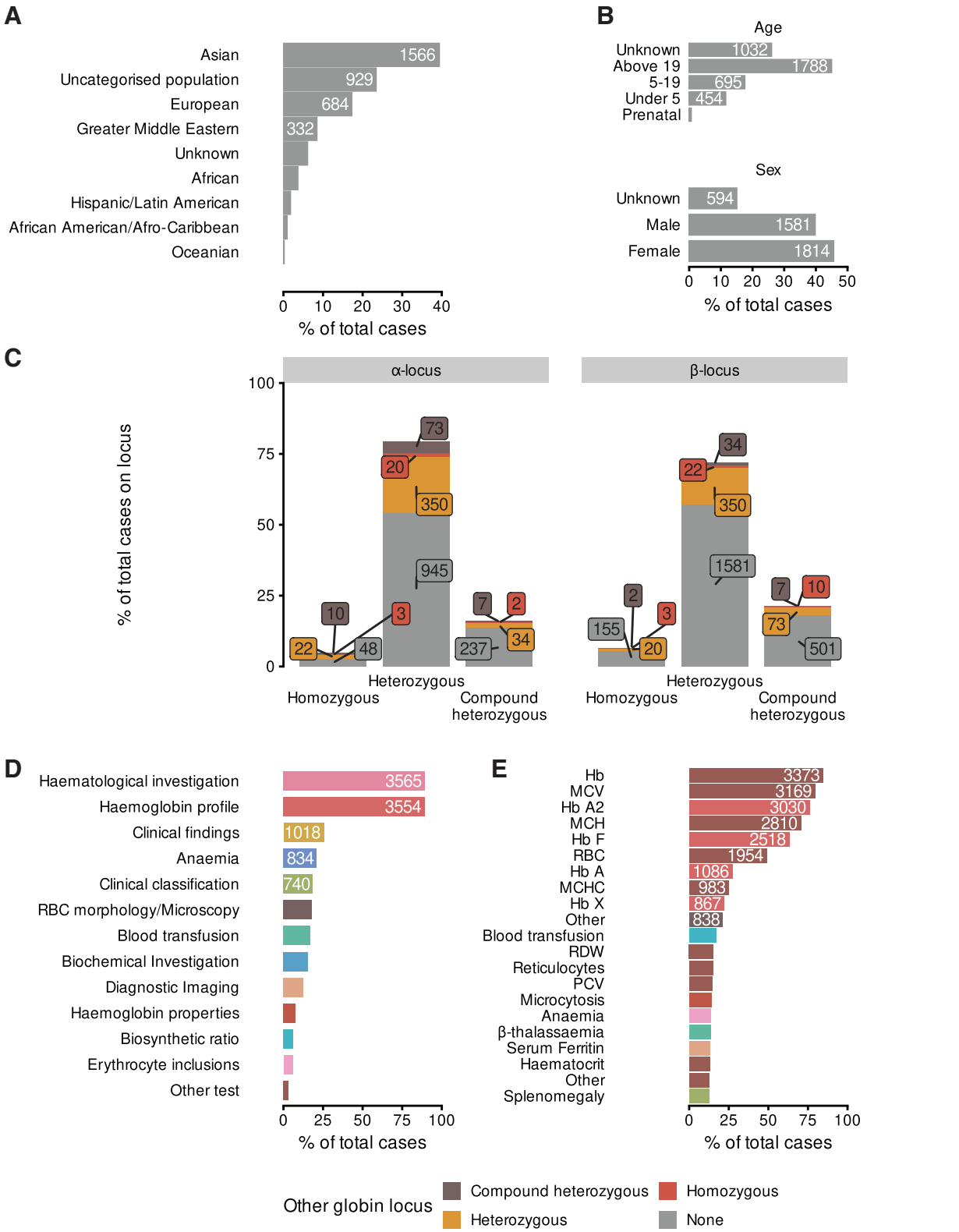

Figure 1. Distribution of IthaPhen entries. (A) Characterization of the origin of individuals reported in IthaPhen according to the Human Ancestry Ontology. Individuals whose origin was not specified by the curated reports, are shown here under the category Unknown. (B) Age and sex distribution of cases reported on IthaPhen. (C) Percentage of homozygous, heterozygous, and compound heterozygous cases per globin loci. (D–E) The phenotypic parameters collected in IthaPhen and their grouping into categories.

The visualization interface is highly intuitive, user-friendly, and informative, and is illustrated in Suppl. Figure S1 for an example query (--^{SEA}/αα). IthaPhen supports queries of any genotype combination, including heterozygous, compound heterozygous, or homozygous individuals, as well as combinations of genotypes in both globin loci. Filtering based on genotype gives the option to select a specific variant or allele phenotype (eg, β^+, α^0) for each allele (Suppl. Figure S1A). Apart from the minimally required genotype filter, IthaPhen also allows further filtering by age and sex. The query results include summary statistics of the phenotype of matching cases presented as tables and graphs (Suppl. Figure S1B and S1C, respectively), but also the raw case data. All summary data produced per queried information can be downloaded as a report, while all raw data for each search can also be downloaded in the Unified Dataset section. Along with the query results, IthaPhen additionally delivers links to relevant queries involving variants in the user's primary search (Suppl. Figure S1D). Importantly, all relevant EMQN best practice guidelines are also reported,[12] as to provide recommendations on prenatal or preimplantation genetic diagnosis, for genotypes that may occur in the offspring of individuals with the queried genotype (Suppl. Figure S1E). Finally, the result section also lists all sources used in curating the individual cases from publications or direct contributions.

The IthaPhen dataset is freely available for noncommercial use and further analysis. Herein, we demonstrate its use with

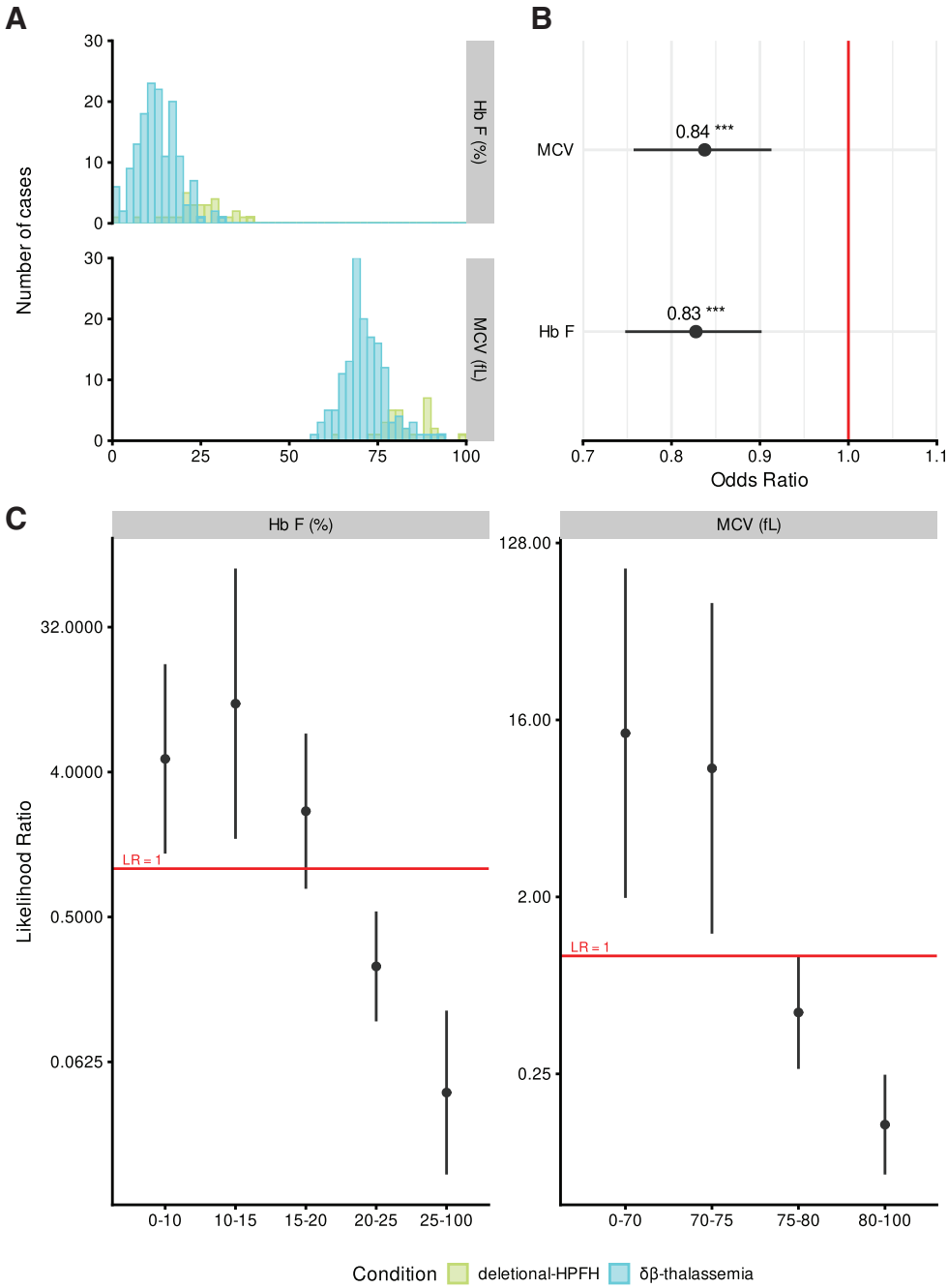

Figure 2. Genotype-phenotype correlation for δβ-thalassemia and deletional HPFH. (A) Distribution of Hb F and MCV measurements in δβ-thalassemia and deletional HPFH. (B) Odds ratio for parameters identified as significant predictors in the logistic regression model. *P* values <0.001 are annotated with 3 asterisks. (C) Likelihood ratio estimates for δβ-thalassemia per hematological indices class. Hb F = fetal hemoglobin; HPFH = hereditary persistence of fetal hemoglobin.

2 examples. First, we investigate phenotypic differences in deletional hereditary persistence of fetal hemoglobin (HPFH) and δβ-thalassemia, two molecularly similar conditions characterized by elevated fetal hemoglobin (Hb F) levels. In heterozygotes, their diagnosis is challenging as it strongly depends on hematological indices and empirical differences in the Hb F levels.[15] Using case-level data stored in IthaPhen, we performed a series of logistic regression models (Suppl. File 2) to demonstrate a quantitative approach in the diagnosis of the 2 conditions. Hb F and MCV are differently distributed between δβ-thalassemia and deletional HPFH (Figure 1A) and were found to be predictive of δβ-thalassemia diagnosis, with odds ratio 0.83 (95% confidence interval [CI], 0.748-0.902) and 0.84 (95% CI, 0.57-0.913), respectively (Figure 2B). Likelihood ratios (LRs) are often employed to create a clinically applicable assessment of a patient's clinical image and aid in selecting further appropriate diagnostic test, while LRs have been widely used in variant classification.[16–18] Thus, we also assessed the LR estimates for these predictors of δβ-thalassemia diagnosis (Suppl. File 2). Significantly higher LRs are observed for low MCV values (<75fL) and HbF <15%, indicating δβ-thalassemia diagnosis, while significantly lower LRs are observed for normal/borderline MCV values and HbF >20% (Figure 2C).

The second example is the use of IthaPhen data for variant interpretation through the evaluation of case-level and segregation data. Here, we demonstrate how such data collected on IthaPhen can be used under the ACMG/AMP framework for globin genes, specified by the ClinGen Hemoglobinopathy VCEP,[7] to assess variant pathogenicity. As shown in Suppl. File 2: (Suppl. Table S2), 3 globin gene variants currently classified as variants of uncertain significance in ClinVar have been enriched with case-level and segregation evidence organized in IthaPhen, specifically for ACMG/AMP-specified criteria PM3, PM6, PS4, and PP1, resulting in a more informative interpretation.

IthaPhen is a novel disease-specific genotype-phenotype database that aims to uncover the case-level evidence that is fragmented in scientific literature or remains unexploited in diagnostic laboratories, particularly for rare hemoglobinopathy cases. It collects anonymized published and contributed genotype-phenotype data and brings together available molecular, laboratory, and clinical information. IthaPhen demonstrates the correlations between genotype and phenotype and is a unique and powerful tool for clinicians and molecular geneticists.

The integration of IthaPhen with other databases of the constantly updated ITHANET portal provides the community with a valuable expanding dataset of case-level data with rich phenotypic annotation and ensures its sustainability. In addition, it will help generate critical evidence for the interpretation of globin gene variants based on the ACMG/AMP framework, an approach also followed by other VCEPs.[19] Furthermore, the integration of EMQN guidelines into IthaPhen facilitates, but does not replace, genetic diagnosis and genetic counseling for hemoglobinopathies, particularly in countries with limited experience and less exposure to hemoglobinopathies. Notably, we demonstrated the utility of IthaPhen with 2 examples in (a) genotype-phenotype correlation and (b) standardized variant classification, illustrating how IthaPhen can contribute to data-driven and quantitative approaches in the field of hemoglobinopathies.

IthaPhen is currently focused on the annotation of case-level and segregation evidence particularly for less common variants that, due to their frequency, represent the most challenging cases for diagnostic laboratories. However, in the future, we plan to also incorporate data from studies reporting aggregated cohort descriptions and larger amounts of disease-modifying variants.[20,21] This will potentially contribute to the personalized diagnosis and management of hemoglobinopathies.

ACKNOWLEDGMENTS

We thank the Cyprus Institute of Neurology and Genetics for computer equipment and for hosting ITHANET. We owe particular thanks to Kyriaki Michailidou for the many fruitful discussions and advisory on the statistical analyses presented herein.

AUTHOR CONTRIBUTIONS

MX, AM, PK, and MK conceived and designed the study. MX designed and developed the database and tool interface. AM defined the data parameters to be collected. MX analyzed the data and produced the figures. AM, CS, and ST curated and validated data. ST designed the FAQ section. MX, AM, and PK wrote the article. All authors have read and approved the final version of the article.

DATA AVAILABILITY STATEMENT

IthaPhen and all associated data are available via a web interface at https://ithanet.eu/db/ithaphen. We make the source code for running the analyses and generating the figures presented herein, freely available at https://github.com/cing-mgt/ithaphen-db-analysis.

DISCLOSURES

The authors have no conflicts of interest to disclose.

SOURCES OF FUNDING

This work is cofunded by the European Regional Development Fund and the Republic of Cyprus through the Research and Innovation Foundation (Project: EXCELLENCE/1216/256).

REFERENCES

1. Modell B, Darlison M. Global epidemiology of haemoglobin disorders and derived service indicators. *Bull World Health Organ*. 2008;86:480–487.
2. GBD 2019 Diseases and Injuries Collaborators. Global burden of 369 diseases and injuries in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet*. 2020;396:1204–1222.
3. Kountouris P, Lederer CW, Fanis P, et al. IthaGenes: an interactive database for haemoglobin variations and epidemiology. *PLoS One*. 2014;9:e103020.
4. Kountouris P, Stephanou C, Bento C, et al. ITHANET: information and database community portal for haemoglobinopathies. *bioRxiv*. 2017:209361.
5. Kountouris P, Michailidou K, Christou S, et al. Effect of HBB genotype on survival in a cohort of transfusion-dependent thalassemia patients in Cyprus. *Haematologica*. 2021;106:2458–2468.
6. Richards S, Aziz N, Bale S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med*. 2015;17:405–424.
7. Kountouris P, Stephanou C, Lederer CW, et al. Adapting the ACMG/AMP variant classification framework: a perspective from the clingen hemoglobinopathy variant curation expert panel. *Hum Mutat*. 2022;43:1089–1096.
8. Landrum MJ, Lee JM, Riley GR, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res*. 2014;42:D980–D985.
9. Firth HV, Richards SM, Bevan AP, et al. DECIPHER: database of chromosomal imbalance and phenotype in humans using Ensembl resources. *Am J Hum Genet*. 2009;84:524–533.
10. Preston CG, Wright MW, Madhavrao R, et al. ClinGen variant curation interface: a variant classification platform for the application of evidence criteria from ACMG/AMP guidelines. *Genome Med*. 2022;14:6.
11. Giardine BM, Joly P, Pissard S, et al. Clinically relevant updates of the HbVar database of human hemoglobin variants and thalassemia mutations. *Nucleic Acids Res*. 2021;49(D1):D1192–D1196.
12. Traeger-Synodinos J, Harteveld CL, Old JM, et al. EMQN best practice guidelines for molecular and haematology methods for carrier identification and prenatal diagnosis of the haemoglobinopathies. *Eur J Hum Genet*. 2015;23:426–437.

13. Lederer CW, Basak AN, Aydinok Y, et al. An electronic infrastructure for research and treatment of the thalassemias and other hemoglobinopathies: the Euro-mediterranean ITHANET project. *Hemoglobin*. 2009;33:163–176.
14. Morales J, Welter D, Bowler EH, et al. A standardized framework for representation of ancestry data in genomics studies, with application to the NHGRI-EBI GWAS Catalog. *Genome Biol*. 2018;19:21.
15. Minaidou A, Tamana S, Stephanou C, et al. A novel tool for the analysis and detection of copy number variants associated with haemoglobinopathies. *Int J Mol Sci*. 2022;23:15920.
16. Spurdle AB, Couch FJ, Parsons MT, et al. Refined histopathological predictors of BRCA1 and BRCA2 mutation status: a large-scale analysis of breast cancer characteristics from the BCAC, CIMBA, and ENIGMA consortia. *Breast Cancer Res*. 2014;16:3419.
17. Tavtigian SV, Greenblatt MS, Harrison SM, et al. Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework. *Genet Med*. 2018;20:1054–1060.
18. Tamana S, Xenophontos M, Minaidou A, et al. Evaluation of in silico predictors on short nucleotide variants in HBA1, HBA2, and HBB associated with haemoglobinopathies. *Elife*. 2022;11:e79713.
19. Salehipour D, Farncombe KM, Andric V, et al. Developing a disease-specific annotation protocol for VHL gene curation using Hypothes.is. *Database*. 2023;2023:baac109.
20. Stephanou C, Tamana S, Minaidou A, et al. Genetic modifiers at the crossroads of personalised medicine for haemoglobinopathies. *J Clin Med*. 2019;8:1927.
21. Kountouris P, Stephanou C, Archer N, et al. The International Hemoglobinopathy Research Network (INHERENT): An international initiative to study the role of genetic modifiers in hemoglobinopathies. *Am J Hematol*. 2021;96:E416–E420.