

**Γεφυρώνοντας μεγάλα ομικά, γενετικά και ιατρικά
δεδομένα για την ευρεία εφαρμογή της Ιατρικής
Ακριβείας στην Ελλάδα**

ΠΑΡΑΔΟΤΕΟ Π3.2

**«Επαναξιολόγηση παραλλαγών άγνωστης κλινικής σημασίας και
PRS στον κληρονομικό καρκίνο»**

Φορέας	Ερευνητικό Κέντρο Βιοϊατρικών Επιστημών «Αλέξανδρος Φλέμινγκ»
Τύπος Παραδοτέου	Άλλο
Ημερομηνία Υποβολής Παραδοτέου	31 Δεκεμβρίου 2025
Ενότητα Εργασίας	Ενότητα Εργασίας 3 «Κληρονομικός Καρκίνος - Επαναξιολόγηση Δεδομένων Αλληλούχησης Ασθενών»

1. Επαναξιολόγηση παραλλαγών άγνωστης κλινικής σημασίας

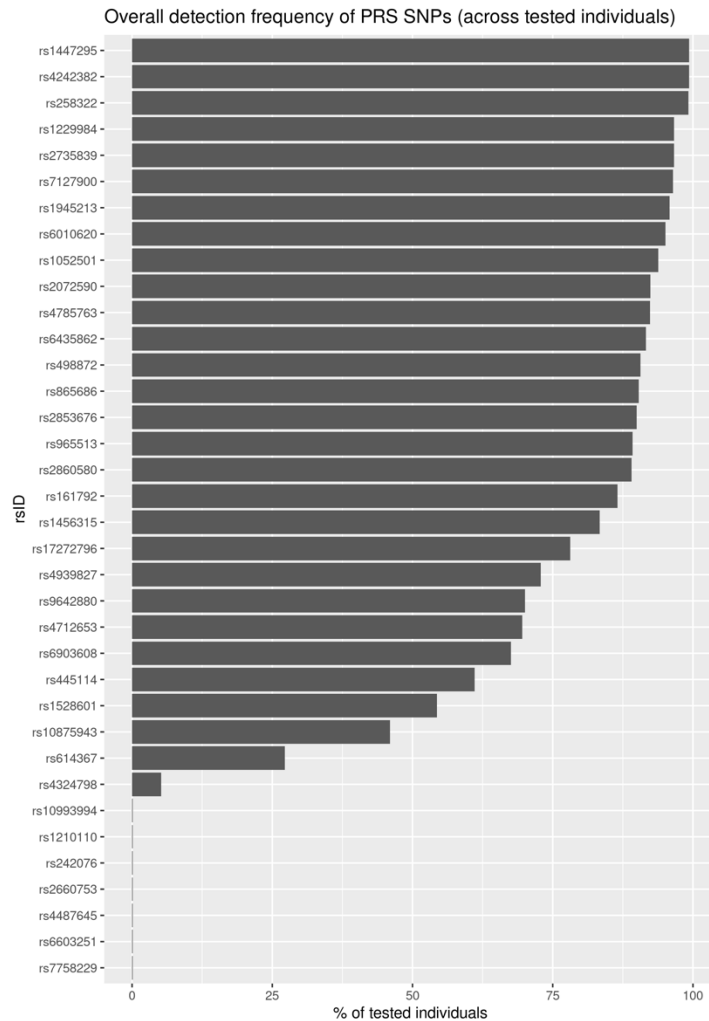
Στο πλαίσιο της Εργασίας 1.1 πραγματοποιήθηκε επαναξιολόγηση όλων των παραλλαγών άγνωστης κλινικής σημασίας (Variants of Uncertain Significance – VUS) που βρίσκονται στη βάση CanVaS με βάση τα κριτήρια του American College of Medical Genetics and Genomics (ACMG). Για την τεκμηρίωση της κλινικής ερμηνείας αξιοποιήθηκαν διαθέσιμες βάσεις δεδομένων και η σχετική βιβλιογραφία, ενώ εφαρμόστηκε συστηματική αναζήτηση/παρακολούθηση νεότερων δεδομένων μέσω του εργαλείου LitVar2, το οποίο υποστηρίζει την ανάκτηση ενημερωμένης βιβλιογραφίας για συγκεκριμένες παραλλαγές σε τακτά χρονικά διαστήματα.

Παράλληλα, βρίσκεται σε εξέλιξη η ανάπτυξη υπολογιστικού μοντέλου βασισμένου σε μεθόδους μηχανικής μάθησης, με στόχο την υποστήριξη της διαδικασίας επαναξιολόγησης και την ιεράρχηση της διαθέσιμης τεκμηρίωσης. Στο πλαίσιο αυτό διερευνάται και η αξιοποίηση προεκπαιδευμένων γλωσσικών μοντέλων (foundation models) για την αυτοματοποιημένη εξαγωγή/σύνοψη πληροφορίας από βιβλιογραφικά δεδομένα, με ιδιαίτερη έμφαση σε παραλλαγές που εμφανίζονται συχνότερα στον ελληνικό πληθυσμό.

2. Υπολογισμός PRS

Στο πλαίσιο της Εργασίας 1.3 πραγματοποιήθηκε διερεύνηση της δυνατότητας εφαρμογής πολυγονιδιακών δεικτών κινδύνου (Polygenic Risk Score – PRS) με βάση κοινές γενετικές παραλλαγές (SNPs), αξιοποιώντας τα διαθέσιμα δεδομένα αλληλούχησης του μητρώου. Ως σύνολο παραλλαγών αναφοράς χρησιμοποιήθηκε η λίστα PRS SNPs που περιλαμβάνονται στο γονιδιακό πάνελ TruSight Hereditary Cancer Panel της Illumina (120 SNPs).

Για τον σκοπό αυτό πραγματοποιήθηκε πιλοτική μελέτη σε αρχεία VCF που αντιστοιχούν σε 1061 άτομα, τα οποία ελέγχθηκαν στο Εργαστήριο Γενετικής του Ανθρώπου του ΕΚΕΦΕ «Δημόκριτος» για κληρονομική προδιάθεση σε καρκίνο μέσω αλληλούχησης νέας γενιάς (Next Generation Sequencing – NGS). Από τα 1061 άτομα, τουλάχιστον ένα από τα 120 PRS SNPs ανιχνεύθηκε σε 1005 ασθενείς (94,7%). Επιπλέον, 30 από τα 120 SNPs ανιχνεύθηκαν σε τουλάχιστον έναν ασθενή στο εξεταζόμενο δείγμα (Εικόνα 1), πιθανώς λόγω χαμηλής συχνότητας ορισμένων SNPs στον υπό εξέταση πληθυσμό και/ή τεχνικών περιορισμών της ανάλυσης (π.χ. κάλυψη και φίλτρα ποιότητας).



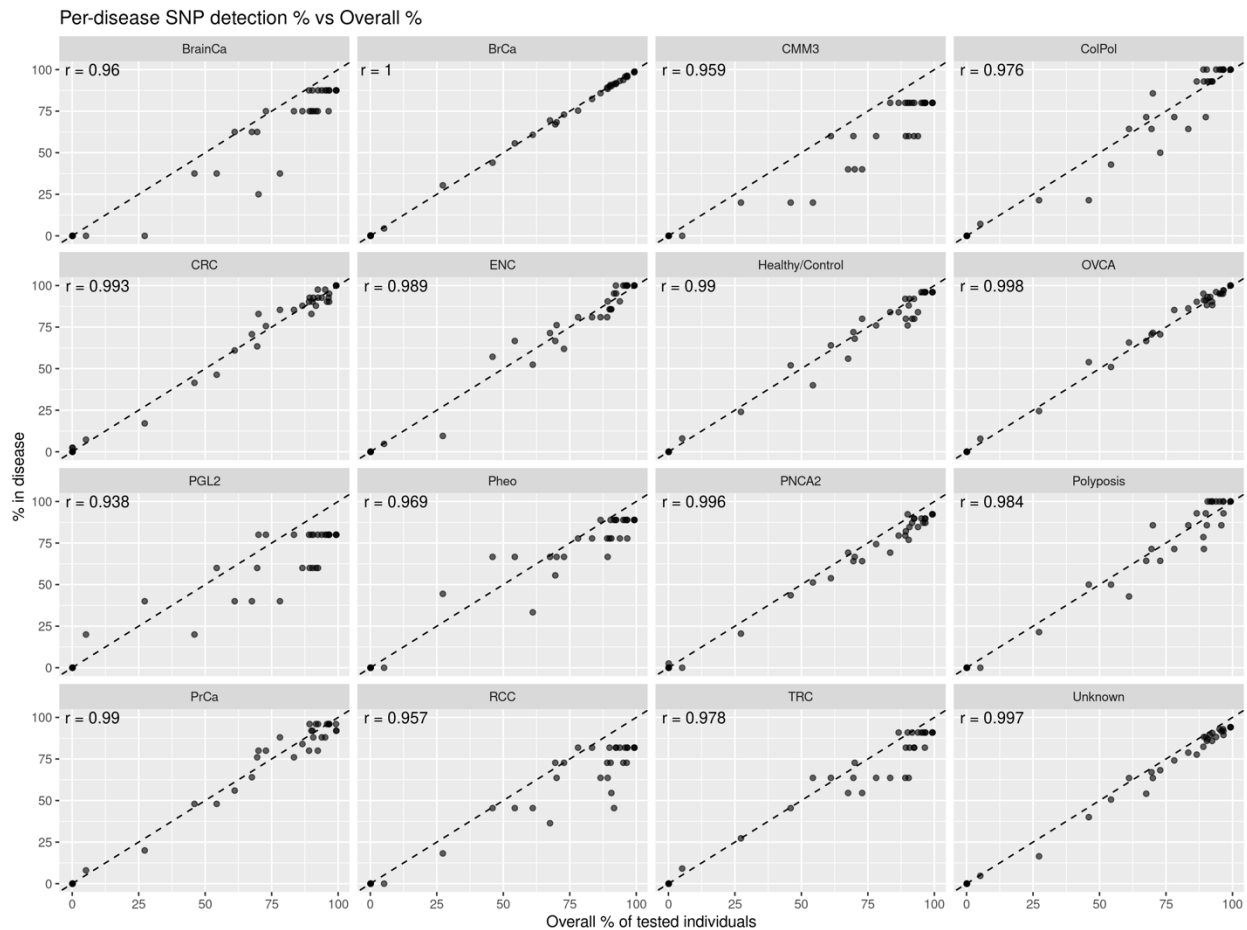
Εικόνα 1: Συχνότητα ανίχνευσης SNPs που έχουν ανιχνευθεί σε έναν ή περισσότερους ασθενείς.

Η ανάλυση των PRS SNPs στο σύνολο των εξετασμένων ατόμων κατέδειξε ότι υποσύνολο των SNPs εμφανίζει υψηλή συχνότητα ανίχνευσης στην υπό εξέταση ομάδα (Εικόνα 2), ενώ για ορισμένες κατηγορίες νόσου (καρκίνος του εγκεφάλου – BrainCa, μελάνωμα – CMM3, καρκίνος των νεφρών – RCC, καρκίνος του θυρεοειδούς – TRC) παρατηρείται σχετικά χαμηλότερη συχνότητα ανίχνευσης των SNPs σε σχέση με το συνολικό πρότυπο, γεγονός που ενδέχεται να

σχετίζεται με το ότι τα συγκεκριμένα PRS SNPs έχουν επιλεγεί κυρίως/μελετηθεί εκτενέστερα για άλλους τύπους καρκίνου.

Παράλληλα, αναπτύχθηκε κώδικας/υπολογιστικό pipeline για την αυτοματοποιημένη εξαγωγή των PRS SNPs από αρχεία VCF και τον υπολογισμό PRS ανά άτομο, με δυνατότητα εφαρμογής σε ευρύτερο σύνολο δεδομένων.

Σημειώνεται ότι τα παραπάνω αποτελέσματα αποτελούν αρχική αποτύπωση της διαθεσιμότητας και κατανομής των PRS SNPs στο διαθέσιμο υποσύνολο δεδομένων και απαιτείται περαιτέρω τεκμηρίωση σε μεγαλύτερο δείγμα, ιδιαίτερα για κατηγορίες νόσου με μικρό αριθμό εξετασμένων ατόμων, ώστε να εξαχθούν ισχυρότερα συμπεράσματα.



Εικόνα 2: Συσχέτιση της συχνότητας ανίχνευσης των PRS SNPs στο σύνολο των εξετασμένων ασθενών και ανά κατηγορία νόσου. BrCa, καρκίνος μαστού. CMM3, μελάνομα. ColPol, πολύποδες παχέος εντέρου. CRC, καρκίνος παχέος εντέρου. ENC, καρκίνος ενδομητρίου. OVCA, καρκίνος ωοθηκών. PGL2, paraganglioma. Pheo, φαιοχρωμοκίττωμα. PrCa, καρκίνος προστάτη. RCC, καρκίνος νεφρού. TRC, καρκίνος θυροειδούς.



A neural network architecture approach for variant prioritization and annotation using CanVas, a population-specific cancer patient database, as a training set

Control/Tracking Number: 2025-A-1617-ESHG

Activity: ESHG Abstract

Current Date/Time: 1/29/2025 3:06:23 PM

A neural network architecture approach for variant prioritization and annotation using CanVas, a population-specific cancer patient database, as a training set

Despoina Kalfakakou^{1,2}, Constantinos Bampos³, Athanasios Papathanasiou², Florentia Fostira², Paraskevi Apostolou², Irene Konstantopoulou², Georgios A. Pavlopoulos⁴, Vasileios Megalooikonomou³, Drakoulis Yannoukakos².

¹*Division of Precision Medicine, Department of Medicine, New York University Grossman School of Medicine, New York, NY, USA*

²*Human Molecular Genetics Laboratory, INRaSTES, NCSR Demokritos, Athens, Greece*

³*Multidimensional Data Analysis and Knowledge Management Laboratory, Computer Engineering and Informatics Department, School of Engineering, University of Patras, Patras, Greece*

⁴*Bioinformatics and Integrative Biology Laboratory Institute for Fundamental Biomedical Research, BSRC Alexander Fleming, Athens, Greece.*

Abstract:

Background: *The vast number of variants with unknown significance in extensive datasets like ClinVar presents significant challenges for automated annotation tools. Rare variants add an additional layer of complexity in accurately assessing clinical significance. National genetic variation registries, such as CanVaS (Cancer Variation reSource), provide valuable insights into population-specific genetic diversity, enhancing automated annotation approaches.*

Material and Methods: *We propose a framework leveraging machine learning and deep learning to improve genetic variant annotation using ClinVar and CanVaS as complementary datasets. CanVaS integrates germline genetic data from ~11,000 Greek cancer patients, analyzed for up to 100 cancer susceptibility genes, with ~7,000 variants manually and thoroughly annotated. Detailed data include allele frequencies, clinical significance,*

segregation, and phenotypic traits. First, we will train tree-based algorithms and Support Vector Machines on BRCA1/BRCA2 variants from ClinVar, chosen due to their extensive study and critical role in cancer. These models will be validated on CanVaS to assess population-specific applicability. Next, we will broaden the approach to annotate missense variants from other cancer genes, using convolutional neural networks to analyze missense variants. To address rare variants, self-supervised learning techniques like contrastive learning will be applied to ClinVar's unlabeled variants, fine-tuned on BRCA1/BRCA2, and validated on CanVaS.

Results: We anticipate high accuracy in pathogenicity predictions, with effective generalization to the Greek population and scalability to broader missense variants.

Conclusion: By integrating advanced algorithms with population-specific data, this framework offers robust tools for variant annotation, enhancing personalized risk assessment and clinical decision-making, providing impactful insights on rare and uncertain variants.

Author Disclosure Information:

D. Kalfakakou: None. **C. Bampos:** None. **A. Papathanasiou:** None. **F. Fostira:** None. **P. Apostolou:** None. **I. Konstantopoulou:** None. **G.A. Pavlopoulos:** None. **V. Megalooikonomou:** None. **D. Yannoukakos:** None.

Topic (Complete): 17. Bioinformatics, Machine Learning and Statistical Methods

Keyword (Complete): hereditary cancer ; variant annotation ; population-specific database

Presentation Preference (Complete): E-Poster

Select one of the options below: Regular Abstract

Please specify all grants & references related to your abstract: : This work was funded by the EU NextGenEU through the General Secretariat for Research and Innovation of the Hellenic Ministry of Development as part of the project "Bridging big omic, genetic and medical data for the wide application of Precision Medicine in Greece" (TAEDR-0539180).

Fellowships & Awards (Complete):

Do you wish to apply for any of the Fellowships or Awards?: No

I apply for the Conference Fellowship for European Countries: No

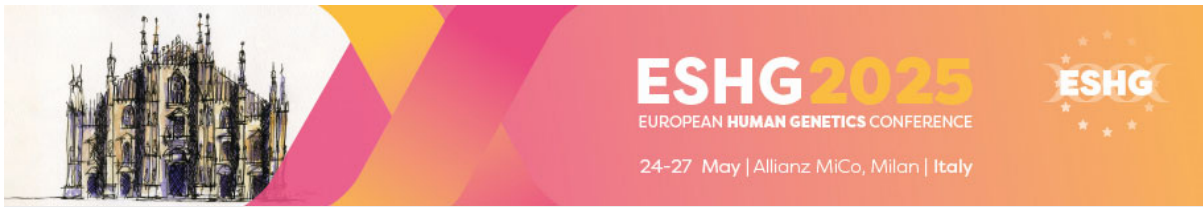
I apply for the Fellowship for Non-European Countries: No

I apply for the Fellowship of Excellence: No

I apply for the Early Career Award: No

Attached Files: No Files Attached

Status: Complete



CERTIFICATE OF PRESENTATION

This is to certify that the abstract

A neural network architecture approach for variant prioritization and annotation using CanVas, a population-specific cancer patient database, as a training set

with the presentation number *EP17.012* was presented in the session

EP | E-Posters at the

European Human Genetics Conference 2025


taking place from May 24-27, 2025 in Milan, Italy and online.

Authors:

*Despoina Kalfakakou^{1,2}, Constantinos Bampos³, Athanasios Papathanasiou², Florentia Fostira², Paraskevi Apostolou², **Irene Konstantopoulou²**, Georgios A. Pavlopoulos⁴, Vasileios Megalooikonomou³, Drakoulis Yannoukakos².*

¹Division of Precision Medicine, Department of Medicine, New York University Grossman School of Medicine, New York, NY, USA, ²Human Molecular Genetics Laboratory, INRaSTES, NCSR Demokritos, Athens, Greece, ³Multidimensional Data Analysis and Knowledge Management Laboratory, Computer Engineering and Informatics Department, School of Engineering, University of Patras, Patras, Greece, ⁴Bioinformatics and Integrative Biology Laboratory Institute for Fundamental Biomedical Research, BSRC Alexander Fleming, Athens, Greece.

***presenting author**


For the organiser
Vienna Medical Academy GmbH
Alser Strasse 4
1090 Vienna, Austria

EUROPEAN SOCIETY OF HUMAN GENETICS | www.eshg.org | [@eshgsociety](https://twitter.com/eshgsociety) | [#eshg2025](https://hashtage.com/eshg2025)

Contact: ESHG 2025 | Vienna Medical Academy GmbH | Alser Str. 4, 1090 Vienna, Austria

T: 0043 (0) 1 405 13 83 35 | E: conference@eshg.org | W: 2025.eshg.org

 Feedback

Powered by [eOASIS](#), The Online Abstract Submission and Invitation System SM
© 1996 - 2025 [CTI Meeting Technology](#). All rights reserved.

A neural network architecture approach for variant prioritization and annotation using CanVaS, a population-specific cancer patient database, as a training set

Despoina Kalfakakou^{1,2}, Constantinos Bampos³, Athanasios Papathanasiou², Florentia Fostira², Paraskevi Apostolou², **Irene Konstantopoulou²**, Georgios A. Pavlopoulos⁴, Vasileios Megalooikonomou³, Drakoulis Yannoukakos².

¹Division of Precision Medicine, Department of Medicine, New York University Grossman School of Medicine, New York, NY, USA

²Human Molecular Genetics Laboratory, INRaSTES, NCSR Demokritos, Athens, Greece

³Multidimensional Data Analysis and Knowledge Management Laboratory, Computer Engineering and Informatics Department, School of Engineering, University of Patras, Patras, Greece

⁴Bioinformatics and Integrative Biology Laboratory Institute for Fundamental Biomedical Research, BSRC Alexander Fleming, Athens, Greece



CanVaS database:
<http://ithaka.rpp.demokritos.gr/CanVaS/genes>

EP17.012



Background

The vast number of variants with unknown significance in extensive datasets like ClinVar presents significant challenges for automated annotation tools. Rare variants add an additional layer of complexity in accurately assessing clinical significance. National genetic variation registries, such as CanVaS (Cancer Variation reSource), provide valuable insights into population-specific genetic diversity, enhancing automated annotation approaches.

Results

We anticipate high accuracy in pathogenicity predictions, with effective generalization to the Greek population and scalability to broader missense variants.

Conclusion

By integrating advanced algorithms with population-specific data, this framework offers robust tools for variant annotation, enhancing personalized risk assessment and clinical decision-making, providing impactful insights on rare and uncertain variants.

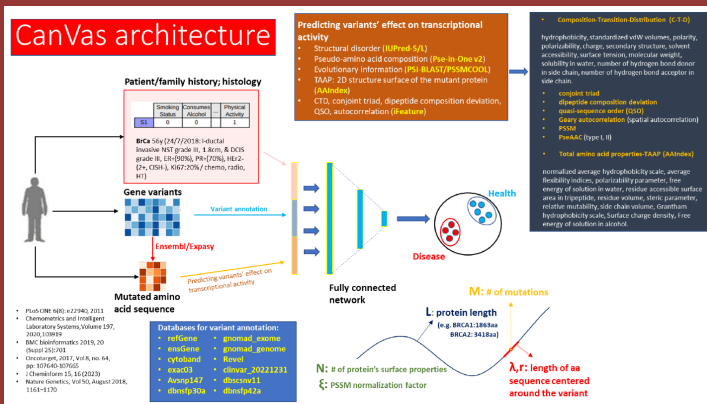


Figure 1. Based on the CanVaS database, we have deployed two main architectures. The first is used to classify variants, while the second is designed to classify phenotypes. Although there are differences between these architectures, they both share a common primary input: variant and amino acid sequence properties. To ensure comprehensive variant annotation, we interrogated 12 different databases. Additionally, we employed various tools and servers to predict the impact of variants on the corresponding transcriptional activity. The main features extracted for our analysis included Composition-Transition-Distribution (CTD) properties and total amino acid properties.

Material and Methods

We propose a framework leveraging machine learning and deep learning to improve genetic variant annotation using ClinVar and CanVaS as complementary datasets. CanVaS integrates germline genetic data from ~11,000 Greek cancer patients, analyzed for up to 100 cancer susceptibility genes, with ~7,000 variants manually and thoroughly annotated. Detailed data include allele frequencies, clinical significance, segregation, and phenotypic traits. First, we will train tree-based algorithms and Support Vector Machines on BRCA1/BRCA2 variants from ClinVar, chosen due to their extensive study and critical role in cancer. These models will be validated on CanVaS to assess population-specific applicability. Next, we will broaden the approach to annotate missense variants from other cancer genes, using convolutional neural networks to analyze missense variants. To address rare variants, self-supervised learning techniques like contrastive learning will be applied to ClinVar's unlabeled variants, fine-tuned on BRCA1/BRCA2, and validated on CanVaS.

The authors have nothing to disclose

This work was funded by the EU NextGenEU through the General Secretariat for Research and Innovation of the Hellenic Ministry of Development as part of the project "Bridging big omic, genetic and medical data for the wide application of Precision Medicine in Greece" (TAEDR-0539180).