

Rezumatul activității și a rezultatelor obținute în proiect în anul 2025

Cifrul proiectului	<u>25.80013.8007.02ROMD</u>
Denumirea proiectului	<u>Cercetare genomică colaborativă privind variațiile genetice implicate în sănătatea cardiovasculară în Europa de Est</u>

Rezumat

Etapa 2025 a proiectului CardioGen a avut ca scop stabilirea cadrului operațional și analitic necesar pentru desfășurarea cercetărilor ulterioare în genomica cardiovasculară, prin implementarea mecanismelor de management, consolidarea infrastructurii computaționale, compilarea și verificarea datelor publice relevante, precum și obținerea unui set de date curate și standardizate prin dezvoltarea și validarea pipeline-urilor de QC și pre-procesare. În acest context au fost definite și atinse obiectivele privind managementul de proiect, setarea infrastructurii de calcul bazat pe centrul HPC a UTM, compilarea datasetului ECD-PD, implementarea pipeline-ului de QC, analiza metodelor de pre-procesare și consolidarea colaborării instituționale RO–MD. Pe plan organizațional, au fost implementate mecanismele de management și coordonare (ședințe periodice, documente operaționale comune), asigurând o guvernanta clară a proiectului. La nivel tehnic, UTM (MD) a configurat infrastructura HPC și mediul software bioinformatic (Docker, Nextflow, GitHub), incluzând indexarea genomului de referință GRCh38 și structurarea unui pipeline complet pentru analize WES/WGS. A fost construit și caracterizat setul de date public ECD-PD (2.394 de probe, 14 fenotipuri cardiovasculare), evidențiindu-se atât potențialul său pentru explorări preliminare, cât și limitările legate de distribuția dezechilibrată a fenotipurilor și de calitatea metadatelor. Aceste constatări au fundamentat decizia de a extinde baza de date prin integrarea unor resurse cu acces controlat, bine gestionate și standardizate (UK Biobank, dbGaP, gnomAD), pentru a asigura o acoperire fenotipică mai echilibrată și metadata complete, potrivite analizelor caz–control și studiilor pe variante rare. La nivel metodologic, a fost dezvoltat și definit un pipeline de QC și pre-procesare bazat pe FastQC, MultiQC, Cutadapt, Trimmomatic și fastp, orchestrat cu Nextflow și containerizat cu Docker, ceea ce conferă întregului flux analitic portabilitate, scalabilitate și reproductibilitate. Performanța acestor metode a fost evaluată printr-un benchmarking sistematic pe 15 probe, demonstrând că fastp și Cutadapt oferă un compromis optim între retenția citirilor și filtrarea contaminanților, în timp ce combinația Cutadapt + Trimmomatic maximizează stringența filtrării, cu costul unui volum de date mai redus. Rezultatele permit selectarea parametrilor optimi de QC pentru etapele viitoare. Etapa a fost însoțită de activități de diseminare și consolidare a capacității științifice, inclusiv prezentarea „*Transparency and reproducibility in computational biomedical research*” la conferința Smart Diaspora 2025 (Cluj-Napoca, România) și inițierea unui review de sinteză privind epigenetica, genele modificatoare și zgomotul molecular în cardiomiopatii. În ansamblu, acțiunile realizate în 2025 au consolidat capacitatea instituțională de analiză bioinformatică, au conturat setul de date necesar și setat infrastructura analitică, plasând proiectul CardioGen pe un fundament tehnologic potrivit pentru investigațiile genomice planificate în etapele următoare.

Summary

The 2025 stage of the CardioGen project aimed to establish the operational and analytical framework needed for subsequent research in cardiovascular genomics by implementing management mechanisms, strengthening the computational infrastructure, compiling and verifying relevant public data, and obtaining a clean and standardized dataset through the development and validation of QC and pre-processing pipelines. In this context, the objectives related to project management, setting up the computing infrastructure based on UTM's HPC center, compilation of the ECD-PD dataset, implementation of the QC pipeline, benchmarking of pre-processing methods, and consolidation of RO-MD institutional collaboration were defined and achieved. At the organizational level, management and coordination mechanisms (regular meetings, shared operational documents) were implemented, ensuring clear governance of the project. At the technical level, UTM (MD) configured the HPC infrastructure and the bioinformatics software environment (Docker, Nextflow, GitHub), including indexing of the GRCh38 reference genome and structuring of a complete pipeline for WES/WGS analyses. The public ECD-PD dataset (2,394 samples, 14 cardiovascular phenotypes) was constructed and characterized, revealing both its potential for preliminary explorations and its limitations related to the unbalanced distribution of phenotypes and variable metadata quality. These findings provided the basis for extending the database by integrating well-curated and standardized controlled-access resources (UK Biobank, dbGaP, gnomAD) to ensure a more balanced phenotypic coverage and complete metadata suitable for case-control analyses and rare-variant studies. At the methodological level, a QC and pre-processing pipeline based on FastQC, MultiQC, Cutadapt, Trimmomatic and fastp was developed and standardized, orchestrated with Nextflow and containerized with Docker, which gives the entire analytical workflow portability, scalability and reproducibility. The performance of these methods was evaluated through systematic benchmarking on 15 samples, showing that fastp and Cutadapt provide an optimal compromise between read retention and contaminant filtering, while the Cutadapt + Trimmomatic combination maximizes filtering stringency at the cost of a smaller data volume. The results allow the selection of optimal QC parameters for future stages. This stage was accompanied by dissemination and capacity-building activities, including the presentation "Transparency and reproducibility in computational biomedical research" at the Smart Diaspora 2025 conference (Cluj-Napoca, Romania) and the initiation of a synthesis review on epigenetics, modifier genes and molecular noise in cardiomyopathies. In general, the actions carried out in 2025 strengthened the institutional capacity for bioinformatic analysis, outlined the required dataset, and set up the analytical infrastructure, placing the CardioGen project on a solid technological foundation for the genomic investigations planned for the next stages.

Conducătorul de proiect Dr. Dumitru CIORBĂ

Data: 01 decembrie 2025

