RECEPȚIONAT

Agenția Națională pentru

Cercetare și Dezvoltare _____

"___" _____ 2025

AVIZAT

Secția AȘM _____

"___" _____ 2025

RAPORT ȘTIINȚIFIC ANUAL

(pentru etapa 2025)

privind implementarea proiectului din cadrul concursului „Proiecte complexe bilaterale cu Republica Moldova"

Proiectul **„Cercetare genomică colaborativă privind variațiile genetice implicate în sănătatea cardiovasculară în Europa de Est"**

(titlul proiectului)

Cifrul proiectului ***25.80013.8007.02ROMD***

Prioritatea Strategică ***I „Sănătate"***

Rector U.T.M. **dr. hab. Viorel BOSTAN** _____

(numele, prenumele) *(semnătura)*

Președintele Consiliului științific UTM **dr. hab. Vasile TRONCIU** _____

(numele, prenumele) *(semnătura)*

Conducătorul proiectului **dr. Dumitru CIORBĂ** _____

(numele, prenumele) *(semnătura)*

L.Ș.

Chișinău, 2025

CUPRINS:

1. Scopul etapei

Scopul etapei aferente anului 2025 este stabilirea cadrului operațional și analitic al proiectului CardioGen 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[1]-urilor de QC și pre-procesare.

2. Obiectivele etapei

În cadrul etapei au fost definite următoarele obiective strategice:

- Implementarea mecanismelor de management și coordonare necesare derulării eficiente a proiectului.
- Stabilirea și standardizarea infrastructurii HPC și a mediului software pentru analize genomice.
- Compilarea, descărcarea și verificarea datelor genomice publice relevante pentru construirea setului ECD-PD.
- Dezvoltarea și operarea unui pipeline complet de control al calității (QC) și pre-procesare pentru datele brute.
- Evaluarea comparativă a metodelor de pre-procesare printr-un benchmarking aplicat pe un set de test reprezentativ.
- Consolidarea colaborării instituționale prin activități de training, vizite bilaterale și schimb de expertiză între parteneri.

3. Acțiunile planificate pentru realizarea scopului și obiectivelor etapei

Nr.	Acțiune planificată	Partener responsabil
1	Implementarea managementului de proiect și a mecanismelor de coordonare RO-MD	SCUB (RO) și UTM (MD)
2	Configurarea infrastructurii HPC și instalarea mediului software	UTM (MD)
3	Achiziția, compilarea și verificarea setului de date ECD-PD[2]	SCUB (RO) și UTM (MD)
4	Implementarea pipeline-ului QC și documentarea acestuia	UTM (MD)
5	Realizarea benchmarkingului metodelor de pre-procesare	SCUB (RO) și UTM (MD)
6	Consolidarea colaborării RO-MD prin vizite și schimb de experiență	SCUB (RO) și UTM (MD)

[1] Pipeline: ansamblu structurat de pași software prin care datele genomice brute sunt pre-procesate, filtrate, analizate și transformate în rezultate interpretabile

[2] ECD-PD: European cardiac disease public datasets

4. Acțiunile realizate pentru atingerea scopului și obiectivelor etapei

Nr.	Acțiune realizată	Partener responsabil	Statut realizare	Note
1	Mecanismele de management și coordonare RO–MD au fost implementate	SCUB (RO) și UTM (MD)	Finalizat	Ședințe regulate, documente operaționale
2	Infrastructura HPC și mediul software bioinformatic au fost configurate	UTM (MD)	Finalizat	Docker, Nextflow, GitHub, indexare GRCh38
3	Setul de date ECD-PD a fost compilat și metadatele au fost caracterizate	SCUB (RO) și UTM (MD)	În progres	Continuă descărcarea fișierelor brute. În pregătire documentația pentru accesul datelor din bazele de date UK Biobank, dbGaP și gnomAD
4	Pipeline-ul QC a fost implementat în conformitate cu standardele recomandate	UTM (MD)	Finalizat	Integrare FastQC, Cutadapt, Trimmomatic, fastp
5	Benchmarking-ul pentru metodele de pre-procesare a fost efectuat	SCUB (RO) și UTM (MD)	Finalizat	Profiluri comparative între metode, rezultate integrate
6	Colaborarea RO–MD a fost consolidată prin training și schimb de experiență	SCUB (RO) și UTM (MD)	În progres	Activități comune și mobilități scurte

5. Rezultatele obținute

Compilarea setului de date ECD-PD. Pentru crearea setului de date ECD-PD în prima etapă au fost efectuat o căutare în baza de date publice NCBI SRA[3], utilizând cuvinte cheie specifice (`"dilated cardiomyopathy" OR "hypertrophic cardiomyopathy" OR "arrhythmogenic right ventricular cardiomyopathy" OR "restrictive cardiomyopathy" OR "non-compaction cardiomyopathy" OR "long QT syndrome" OR "short QT syndrome" OR "familial atrial fibrillation" OR "aortic valve disease" OR "familial hypercholesterolemia" OR "marfan syndrome" OR "loeys-dietz syndrome" OR "noonan syndrome" OR "brugada syndrome") AND (txid9606[Organism])`) și parametrii adiționali de căutare care au inclus date cu acces public și controlat cu sursă ADN genomic (de tip exom și genom) sau ARN.

Setul de date obținut conține 2.394 de probe SRA, grupate in 14 categorii distincte de fenotipuri patologice cardiovasculare. În ansamblu datele sunt dominate de cardiomiopatia dilatativă, care reprezintă aproximativ 37.6% din totalul probelor, urmată de hipercolesterolemia

[3] Sequence Read Archive (SRA) - https://www.ncbi.nlm.nih.gov/sra

familială (23.9%) și boala de valvă aortică (23,0%). Împreună, aceste trei grupe reprezintă aproximativ 85% din întregul set de date. Fenotipurile mai puțin frecvente includ cardiomiopatia hipertrofică (aproximativ 5,3%), sindromul Brugada (2,7%), sindromul Noonan (2,2%), sindromul Loeys-Dietz (1,3%) și sindromul Mafan (1,1%). Mai rare sunt condiții patologice precum cardiomiopatia aritmogenă a ventricolului drept, sindromul QT lung, sindromul QT scurt, cardimiopatia restrictivă și fibrilația atrială familială, fiecare contribuind cu aproximativ 0,2-0,8 din totalul probelor (Figura 1A-B).

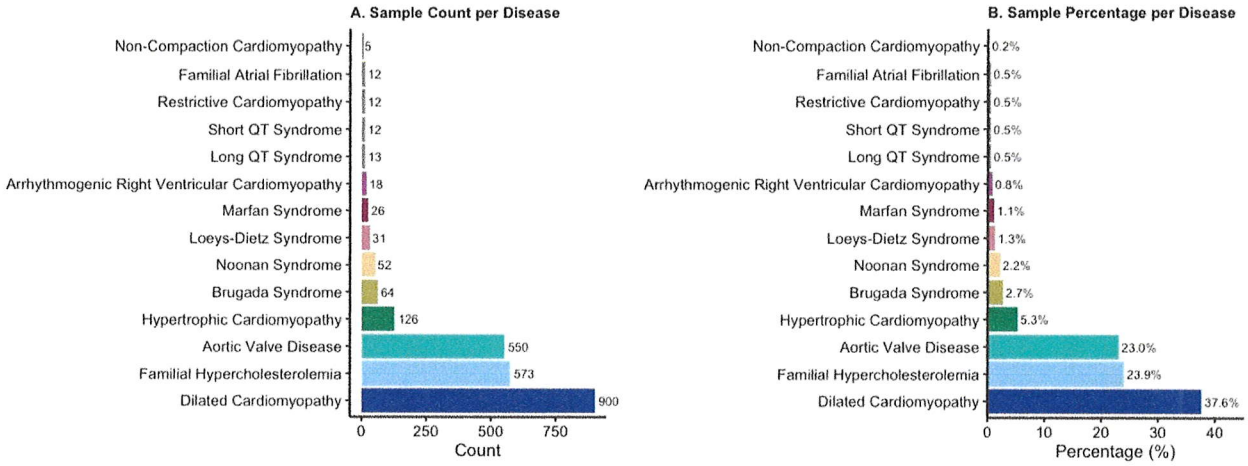

Figura 1. Distribuția probelor pe tipuri de boli cardiace în cohorta analizată, (A) număr absolut de probe, (B) procent relativ din totalul de probe.

Deși setul de date ECD-PD include 2.394 de probe împărțite în 14 categorii fenotipice cardiovasculare, distribuția acestora este puternic dezechilibrată. Trei afecțiuni – cardiomiopatia dilatativă (37,6%), hipercolesterolemia familială (23,9%) și boala de valvă aortică (23,0%) – însumează aproximativ 85% din totalul probelor, în timp ce fenotipurile rare contribuie cu sub 1% fiecare. Această structură neuniformă reduce semnificativ utilitatea setului pentru analize comparative robuste, în special pentru bolile cu prevalență scăzută. În plus, datele publice obținute din SRA prezintă un nivel ridicat de inconsistență și incompletitudine a metadatelor: informații esențiale precum vârsta, sexul, etnia, statutul clinic exact, criteriile diagnostice sau tipul precis de secvențiere sunt absente sau incomplete pentru un procent semnificativ de probe.

Acest lucru limitează capacitatea de a realiza analize statistice controlate și studii genomice comparative riguroase. Din aceste motive, dezechilibrul major al distribuției probelor și insuficiența metadatelor, s-a decis extinderea setului de date prin integrarea unor resurse genomice cu acces controlat, bine curate și standardizate din trei baze de date europene, UK Biobank, dbGaP și gnomAD (Tabelul 1), pentru a asigura o acoperire echilibrată între fenotipuri, a permite acces la metadate complete și standardizate, și a oferi o bază robustă pentru analize comparative ale variantelor genetice în bolile cardiovasculare.

Tabelul 1. Bazele de date planificate a fi utilizate în proiect

BD	Denumirea completă	Scop	URL
UK Biobank	UK Biobank	Cazuri, Controale	https://www.ukbiobank.ac.uk/
dbGap	Database of Genotypes and Phenotypes	Cazuri	https://dbgap.ncbi.nlm.nih.gov/home/
gnomAD	Genome Aggregation Database	Controale	https://gnomad.broadinstitute.org/

Caracterizarea metadatelor setului de date ECD-PD. Analiza metadatelor aferente setului de date ECD-PD evidențiază o structură tehnologică puternic eterogenă, reflectând diversitatea strategiilor de secvențiere utilizate în studiile cardiovasculare disponibile public. Strategia de pregătire a bibliotecilor este dominată de secvențierea exomului (WXS), care reprezintă 34.6% din totalul probelor, urmată de tehnologiile de captare țintită a genelor (18.9%) și de RNA-Seq (17.3%). Probele generate prin metode bazate pe amplificare (Amplicon-Seq) constituie 9.4%, în timp ce secvențierea întregului genom (WGS) este prezentă într-o proporție mai redusă (5.0%). Strategiile mai specializate – precum miRNA-Seq, ChIP-Seq, ATAC-Seq, bisulfite-Seq sau Hi-C – contribuie colectiv cu sub 3% din total (Fig. 2A).

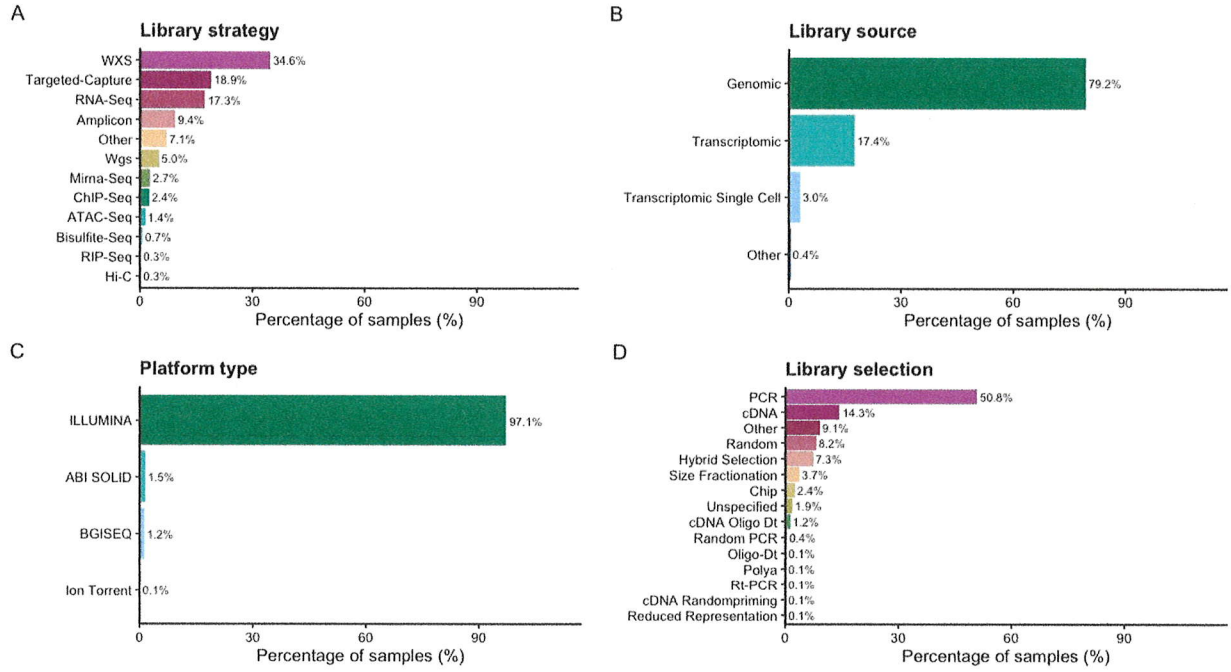

Figura 2. Caracteristicile tehnice ale probelor din setul de date ECD-PD, incluzând strategia de bibliotecă (A), sursa bibliotecii (B), platforme de secvențiere (C) și metoda de selecție a bibliotecii (D).

Din perspectiva sursei bibliotecii, majoritatea covârșitoare a probelor este derivată din ADN genomic (79.2%), în timp ce bibliotecile transcriptomice reprezintă 17.4%, iar secvențierea transcriptomică single-cell un procent limitat (3.0%; Fig. 2B). În mod concordant, tehnologia de secvențiere este dominată aproape exclusiv de platformele Illumina, care însumează 97.1% din toate probele, reflectând standardizarea actuală a studiilor genomice de mare volum; contribuțiile

BGI-SEQ, ABI SOLID şi Ion Torrent sunt marginale (<2%; Fig. 2C). Din perspectiva metodelor de selecţie a bibliotecii, procedurile bazate pe PCR sunt cele mai frecvente (50.8%), urmate de biblioteci cDNA (14.3%) şi selecţii aleatoare (Random, 8.2%), în timp ce alte abordări, precum selecţia hibridă, fracţionarea pe dimensiuni sau tehnicile bazate pe oligonucleotide, au contribuţii sub 8% fiecare (Fig. 2D).

Analiza originii tisulare evidenţiază o predominanţă marcată a probelor sanguine, care reprezintă 50.4% din total, urmate de eşantioane cardiace (în principal ţesut miocardic, valvă aortică sau sept interventricular), fiecare categorie individualizată contribuind cu proporţii reduse, de regulă sub 2% (Fig. 3A). Alte tipuri de ţesut – precum ficatul, splina, rinichiul sau ţesuturi vasculare – apar sporadic, fiecare sub 1% din setul de date. Într-un mod similar, analiza tipului de material biologic utilizat pentru extracţia ADN/ARN confirmă dominaţia probelor de sânge integral sau fracţiuni derivate (PMNC, RBC pellet, buffy coat), care însumează peste 92% din totalul probelor (Fig. 3C). Structura demografică este aproape echilibrată între sexe, cu 48.5% probe provenite de la indivizi de sex masculin şi 47.0% de la indivizi de sex feminin; doar 3.7% dintre eşantioane au sexul neanotat, iar restul (<1%) sunt etichetate drept „not collected" sau „not determined" (Fig. 3B). Această distribuţie uniformă este benefică pentru analize comparative ulterioare, în contextul patologiilor cardiovasculare unde diferenţele dependente de sex sunt documentate.

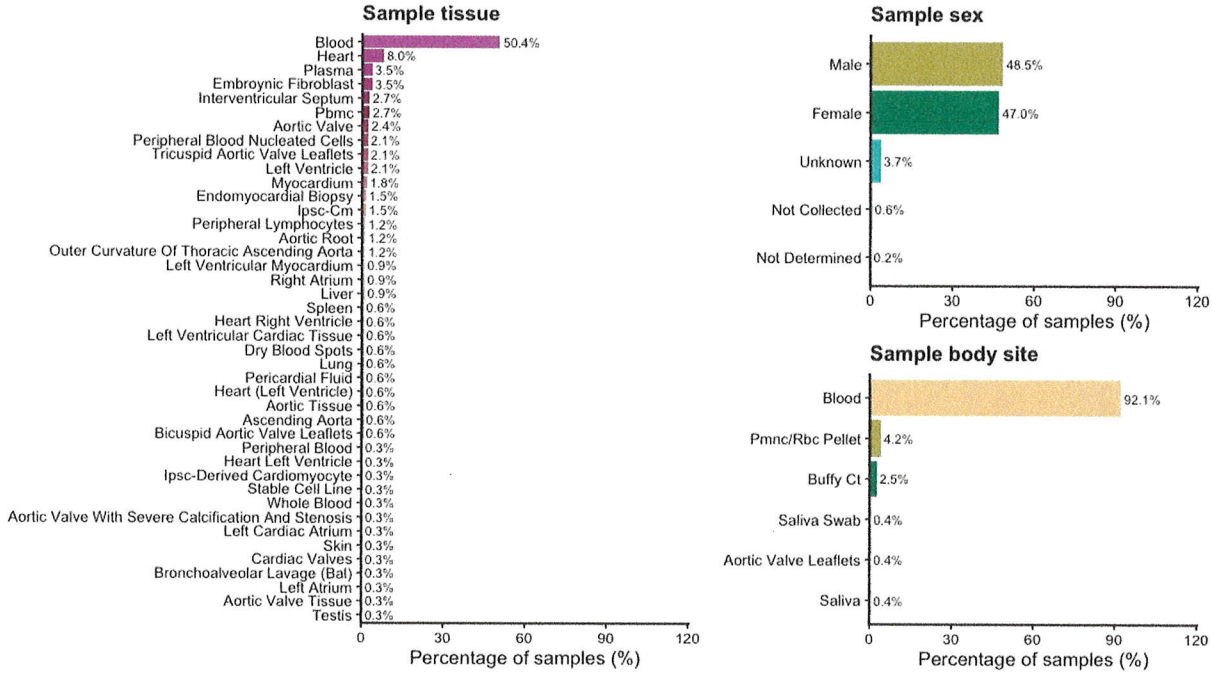

Figura 3. Distribuţia probelor în funcţie de ţesut (A), sexul donatorilor (B) şi sursa biologică a probelor (C) în setul de date ECD-PD.

Astfel, profilul metadatelor indică un set de date robust din perspectivă tehnologică, dar cu o compoziţie biologică puternic centrată pe sânge şi o prevalenţă limitată a ţesuturilor cardiovasculare primare. Aceste caracteristici reflectă pe de o parte tendinţele actuale ale studiilor de genomică populaţională, iar pe de altă parte subliniază necesitatea integrării unor cohorte suplimentare cu distribuţie tisulară mai bogată pentru analize funcţionale profund orientate către mecanismele patogene specifice bolilor cardiace.

Infrastructura computațională. Pentru procesarea datasetului ECD-PD a fost utilizată infrastructura HPC din cadrul TUM, configurată pentru rularea analizelor NGS la scară largă și pentru execuție paralelă pe volume mari de date. Mediul de lucru a fost containerizat integral folosind Docker, fiecare componentă software fiind ambalată în imagini dedicate pentru a asigura reproductibilitatea, izolarea dependențelor și stabilitatea analizelor.

Pentru etapa de QC și preprocesare, au fost integrate containere dedicate pentru FastQC, MultiQC, Cutadapt, Trimmomatic și fastp, acoperind filtrarea, curățarea și evaluarea calității citirilor brute. Etapa de aliniere genomică a utilizat containere separate cu BWA-MEM și Bowtie2, două instrumente standard pentru maparea citirilor NGS.

Genomul de referință utilizat în proiect este GRCh38/hg38, descărcat în forma sa oficială (primary assembly) și pregătit pentru aliniere pe infrastructura HPC–TUM. GRCh38 reprezintă standardul actual în diagnosticul genetic, analizele WES/WGS clinice, GWAS moderne și în ecosistemul bazelor de date majore precum gnomAD v4, ClinVar, dbSNP și Ensembl (v110+). În cadrul infrastructurii, genomul a fost procesat prin generarea indexurilor specifice pentru BWA-MEM (bwa index) și Bowtie2 (bowtie2-build), asigurând compatibilitatea completă cu fluxurile de aliniere și variant calling. Această versiune stabilește un cadru reproducibil și robust pentru toate analizele cardiogenomice și constituie referința „sigură și compatibilă" pentru întregul pipeline al proiectului CardioGen.

Fluxurile de analiză au fost orchestrate prin Nextflow, care permite paralelizarea eficientă a proceselor, gestionarea automată a resurselor și reluarea sigură a execuțiilor în cazul întreruperilor. Prin utilizarea Nextflow și Docker, întregul pipeline a devenit portabil și scalabil, fiind posibilă reproducerea identică a rezultatelor pe orice sistem compatibil cu containere.

Toate scripturile, configurațiile Nextflow și metadatele aferente au fost gestionate printr-un depozit GitHub (https://github.com/bioinformatics-lab-utm/cardiogen), asigurând versionare completă, trasabilitate și control riguros al modificărilor. Acest mecanism permite auditarea rapidă a pipeline-ului și facilitează colaborarea între echipele implicate în proiect.

Evaluarea instrumentelor de pre-procesare a datelor. Pentru a evalua în mod sistematic etapele de pre-procesare a datelor, am realizat un benchmarking pe 15 probe provenite din proiectul SRA-PRJNA393768 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA393768), "*Deep-intronic variants in hypertrophic cardiomyopathy*" (Homo sapiens, ADN genomic). Pe acest set de date am comparat patru fluxuri independente de pre-procesare: (i) analiza directă a datelor brute cu FastQC urmată de agregarea rapoartelor prin MultiQC; (ii) filtrare și curățare cu fastp; (iii) filtrare exclusiv cu Cutadapt; (iv) filtrare exclusiv cu Trimmomatic; și (v) un flux combinat Cutadapt + Trimmomatic, în care îndepărtarea adaptorilor a fost urmată de filtrarea pe baza calității citirilor.

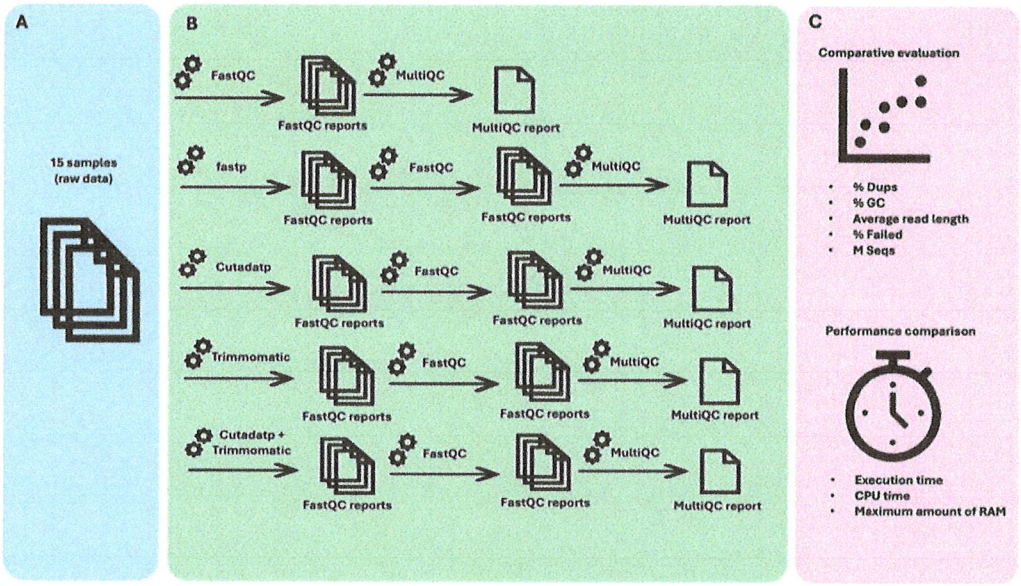

Figura 4. Schemă a designului de benchmarking pentru etapele de pre-procesare: (A) Date brute (proiectul SRA-PRJNA393768); (B) fluxurile paralele de pre-procesare comparate, (C) criteriile de evaluare utilizate, incluzând atât metrici de calitate a datelor și indicatori de performanță computațională.

Toate fluxurile au fost rulate în mod standardizat într-un mediu Docker, folosind praguri comparabile de calitate (Q20–Q30), o lungime minimă a citirilor de ≥35 bp și detecția adaptorilor universali Illumina. Pentru fiecare versiune procesată am generat rapoarte FastQC și un raport integrat MultiQC, pe baza cărora am comparat atât indicatorii de calitate ai datelor (% duplicate, %GC, lungimea medie a citirilor, procentul de citiri eșuate, numărul de citiri), cât și performanța computațională (timp de execuție, timp CPU și memoria RAM maxim utilizată).

Statistici sumare QC a datelor. Pentru cele 15 probe paired-end analizate în cadrul evaluării acurateței metodelor de pre-procesare, setul brut prezintă o medie a duplicărilor de 25.20%, un conținut GC de 44.47%, și lungimi uniforme ale citirilor (97/97 bp). Rata citirilor eșuate este în mod normal 9.30%, însă o probă prezintă valori de până la 18%, indicând variabilitate inițială în calitatea datelor.

Tabelul 1. Sumarul comparativ al performanței metodelor de pre-procesare a datelor NGS.

	% Duplicări (medie)	% GC (medie)	Lungime medie/mediană	% Citiri eșuate	Nr. citiri (M, medie)
Date brute	25.20%	44.47%	97/97 bp	9.30%	9.05 M
Fastp	25.53%	44.17%	92.5 /96 bp	9.00%	8.61 M
Cutadapt	25.56%	44.03%	93.5/96 bp	9.00%	8.69 M
Trimmomatic	25.38%	43.03%	92.5/96 bp	9.00%	8.33 M
Cutadapt + trimmomatic	25.38%	43.90%	92.5/96 bp	9.00%	8.33 M

După aplicarea metodelor de QC, fastp menține valori similare ale duplicărilor (25.53%) și ale GC-ului (44.17%), reducând lungimea medie la 92.5 bp (mediană 96 bp) și scăzând numărul total de citiri la 8.61 M, indicând o filtrare moderată și eficientă. Cutadapt produce rezultate foarte apropiate (% duplicări 25.56%, GC 44.03%, lungime 93.5/96 bp), păstrând cel mai mare număr de citiri după QC (8.69 M), ceea ce reflectă eliminarea adaptatorilor cu pierderi minime. Trimmomatic aplică o filtrare mai agresivă, reducând lungimea medie la 92.5/96 bp, scăzând conținutul GC la 43.03%, și păstrând doar 8.33 M citiri. Fluxul combinat Cutadapt + Trimmomatic generează un set de date foarte curat, cu un conținut GC ușor mai ridicat (43.90%) și același număr final de citiri (8.33 M), fiind metoda cu cele mai scurte și mai puține citiri, dar și una dintre cele mai eficiente în eliminarea secvențelor cu calitate slabă. În ansamblu, metodele mențin valori similare pentru %GC și % duplicări față de datele brute, diferențele majore fiind observate în lungimea și numărul citirilor păstrate, în concordanță cu profilul fiecărei metode.

În această etapă a proiectului CardioGen a definit cadrul tehnic și analitic necesar desfășurării fazelor ulterioare, prin integrarea și evaluarea datelor, caracterizarea detaliată a metadatelor, dezvoltarea infrastructurii computaționale și validarea riguroasă a metodelor de pre-procesare.

Principalele realizări includ:

1. Compilarea și evaluarea datasetului ECD-PD prin construirea setului de date european dedicat bolilor cardiace, incluzând 2.394 de probe. Analiza a evidențiat dezechilibre majore în distribuția probelor și lipsuri importante în metadatele clinice și tehnice, justificând extinderea datasetului cu resurse curate și standardizate.
2. Pe baza limitărilor identificate, au fost selectate cohorte complementare cu acces controlat, UK Biobank, dbGaP și gnomAD pentru a completa distribuția fenotipurilor, a obține metadate complete și a susține analize robuste caz–control.
3. Dezvoltarea infrastructurii HPC și a pipeline-ului reproductibil prin implementarea arhitecturii Docker + Nextflow pe genomul GRCh38 care asigură un flux analitic portabil, scalabil și complet reproductibil, compatibil cu standardele internaționale în genomica clinică.
4. Benchmarking-ul metodelor de pre-procesare a arătat că fastp și Cutadapt oferă un compromis optim între retenție și filtrare, în timp ce combinația Cutadapt + Trimmomatic maximizează stringența, cu costul reducerii volumului de date.

6. Diseminarea rezultatelor

Progresul, ipotezele și rezultatele preliminare obținute în cadrul proiectului au fost prezentate la conferința Smart Diasporta 2025 (Conferința Diaspora Științifică Românească), care a avut loc în perioada 4-7 noiembrie 2025, Cluj-Napoca, România, cu prezentarea intitulată "*Transparency and reproducibility in computational biomedical research*", în secțiunea *Perspective in AI driven computational biology*.

Lista publicațiilor din anul 2025

Echipa proiectului colaborează la elaborarea unui articol de sinteză intitulat "*Epigenetics, Modifiers, and Molecular Noise: Rethinking Inheritance in Dilated and Hypertrophic Cardiomyopathies*", care abordează rolul factorilor epigenetici, al genelor modificatoare și al variabilității stocastice în interpretarea modernă a cardiomiopatiilor genetice.

7. Impactul ştiinţific, social şi/sau economic al rezultatelor ştiinţifice obţinute

Acţiunile realizate în această etapă generează un impact direct prin consolidarea capacităţii regionale de analiză bioinformatică a datelor genomice la scară largă, prin definirea fluxurilor bioinformatice tip (flux standard de referinţă) şi prin crearea unui set de date curat şi reproductibil pentru cercetarea variantelor rare. Setarea infrastructurii de calcul bazată pe capacităţile centrul UTM de HPC, dezvoltarea pipeline-urilor containerizate, benchmarking-ul metodelor de pre-procesare şi compilarea datasetului ECD-PD contribuie la creşterea calităţii şi fiabilităţii analizelor ulterioare, reduc riscurile tehnice şi asigură premisele unor rezultate ştiinţifice robuste în etapele viitoare ale proiectului. Aceste acţiuni sporesc autonomia analitică a echipelor RO–MD, optimizează utilizarea resurselor şi creează un cadru metodologic solid pentru dezvoltarea ulterioară a platformei CardioGen.

8. Colaborare la nivel naţional

Colaborarea la nivel naţional în Republica Moldova în cadrul proiectului CardioGen a fost centrată pe activitatea Universităţii Tehnice a Moldovei, prin integrarea expertizei locale în bioinformatică (Laboratorul Bioinformatică, UTM), dezvoltarea infrastructurii HPC (Direcţia Tehnologia Informaţiei şi Comunicaţiilor, UTM) şi coordonarea acţiunilor tehnice necesare procesării datelor genomice. Unele aspecte de prelucrare genetică au fost consultate cu cercetării Institutului de Microbiologie şi Biotehnologie.

Această colaborări a consolidat capacitatea instituţională în analiza bioinformatică a datelor genomice, contribuind la formarea unei baze tehnico-ştiinţifice robuste pentru etapele ulterioare ale proiectului.

9. Colaborare la nivel internaţional

Colaborarea internaţională în cadrul proiectului CardioGen s-a realizat prin parteneriatul direct dintre Universitatea Tehnică a Moldovei (Republica Moldova) şi Spitalul Clinic de Urgenţă din Bucureşti (România), facilitând integrarea expertizei complementare în bioinformatică, genomică şi infrastructuri computaţionale avansate.

Această colaborare transfrontalieră RO–MD consolidează capacitatea regiunii de a desfăşura cercetări genomice cardiovasculare la standarde europene şi creează condiţii favorabile activităţilor ştiinţifice comune în etapele următoare ale proiectului.

În această etapă, cooperarea RO–MD induce extinderea către platforme internaţionale majore din Regatul Unit (UK Biobank) şi Statele Unite (dbGaP). În acest context se vor realiza cereri de acces şi utilizare a seturilor de date (cu acces controlat/limitat) necesare proiectului.

10. Dificultăţile în realizarea proiectului

- În implementarea proiectului au existat anumite dificultăţi legate de mobilităţile necesare. Disproporţionalitatea fondurilor alocate pentru cheltuielile de deplasare a generat constrângeri logistice. Cu toate acestea, echipa a depus toate eforturile pentru a asigura continuitatea activităţilor şi atingerea obiectivelor stabilite.

- Proiectul fiind internațional, documentația inițială a fost elaborată în limba engleză. Raportarea în limba română reprezintă o schimbare față de limba de lucru inițială, ceea ce poate afecta continuitatea comunicării.
- Datele WGS sunt foarte voluminoase; cele 15 probe depășesc capacitatea de pre-procesare disponibilă (10 TB), ceea ce a generat limitări tehnice în derularea activităților.

Conducătorul de proiect Dr. Dumitru CIORBĂ

Data: 01 decembrie 2025

LȘ

Anexa 1

Rezumatul activității și a rezultatelor obținute în proiect în anul 2025

Cifrul proiectului	*25.80013.8007.02ROMD*
Denumirea proiectului	*Cercetare genomică colaborativă privind variațiile genetice implicate în sănătatea cardiovasculară în Europa de Est*

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 guvernanță 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 metadate 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

LȘ

Direcția Cercetări Științifice

Anexa 2

Lista lucrărilor ştiinţifice, ştiinţifico-metodice şi didactice publicate în anul 2025 în cadrul proiectului

Cercetare genomică colaborativă privind variaţiile genetice implicate în sănătatea cardiovasculară în Europa de Est

1. **Monografii** (recomandate spre editare de consiliul ştiinţific/senatul organizaţiei din domeniile cercetării şi inovării)

1.1. monografii internaţionale

1.2. monografii naţionale

2. Capitole în monografii naţionale/internaţionale

3. Editor culegere de articole, materiale ale conferinţelor naţionale/internaţionale

4. Articole în reviste ştiinţifice

4.1. în reviste din bazele de date Web of Science şi SCOPUS (cu indicarea factorului de impact IF)

4.2. în alte reviste din străinătate recunoscute

4.3. în reviste din Registrul National al revistelor de profil, cu indicarea categoriei

4.4. în alte reviste naţionale

5. Articole în culegeri ştiinţifice naţionale/internaţionale

5.1. culegeri de lucrări ştiinţifice editate peste hotare

5.2 culegeri de lucrări ştiinţifice editate în Republica Moldova

6. Articole în materiale ale conferinţelor ştiinţifice

6.1. în lucrările conferinţelor ştiinţifice internaţionale (peste hotare)

6.2. în lucrările conferinţelor ştiinţifice internaţionale (Republica Moldova)

6.3. în lucrările conferinţelor ştiinţifice naţionale cu participare internaţională

6.4. în lucrările conferinţelor ştiinţifice naţionale

7. Teze ale conferinţelor ştiinţifice

7.1. în lucrările conferinţelor ştiinţifice internaţionale (peste hotare)

7.2. în lucrările conferinţelor ştiinţifice internaţionale (Republica Moldova)

7.3. în lucrările conferinţelor ştiinţifice naţionale cu participare internaţională

7.4. în lucrările conferinţelor ştiinţifice naţionale

Notă: *vor fi considerate teze şi nu articole materialele care au un volum de până la 0,25 c.a.*

8. Alte lucrări ştiinţifice (recomandate spre editare de o instituţie acreditată în domeniu)

8.1. cărţi (cu caracter informativ)

8.2. enciclopedii, dicţionare

8.3. atlase, hărţi, albume, cataloage, tabele etc. (ca produse ale cercetării ştiinţifice)

9. Brevete de invenţii şi alte obiecte de proprietate intelectuală, materiale la saloanele de invenţii

10. Lucrări ştiinţifico-metodice şi didactice

10.1. manuale pentru învăţământul preuniversitar (aprobate de ministerul de resort)

10.2. manuale pentru învăţământul universitar (aprobate de consiliul ştiinţific /senatul instituţiei)

10.3. alte lucrări ştiinţifico-metodice şi didactice

11. Recomandări, propuneri.

Anexa 3

Executarea devizului de cheltuieli,

conform anexei nr. 2.3 din contractul de finanțare pentru anul 2025

Cifrul proiectului ***25.80013.8007.02ROMD***

Cheltuieli, lei				
Denumirea	**Cod**		**Anul de gestiune**	
	Eco (k6)	Aprobat	Modificat +/-	Precizat
Deplasări de serviciu în interiorul țării	222710			
Deplasări de serviciu peste hotare	222720			
Servicii medicale	222810			
Servicii de editare	222910			
Servicii de protocol	222920			
Servicii de cercetări ştiinţifice contractate	222930	**100 080,0**		**100 080,0**
Servicii neatribuite altor aliniate	222999			
Alte cheltuieli în bază de contracte cu persoane fizice	281600			
Cheltuieli curente neatribuite la alte categorii	281900			
Procurarea maşinilor şi utilajelor	314110			
Procurarea activelor nemateriale	317110			
Procurarea combustibilului, carburanților și lubrifianților	331110			
Procurarea produselor alimentare	333110			
Procurarea materialelor pentru scopuri didactice, ştiinţifice şi alte scopuri	335110			
Procurarea materialelor de uz gospodăresc și rechizite de birou	336110			
Procurarea altor materiale	339110			
TOTAL		**100 080,0**		**100 080,0**

Rector U.T.M. _____ (semnătura) **dr. hab. Viorel BOSTAN** (numele, prenumele)

Contabil (economist) _____ (semnătura) **Victoria IOVU** (numele, prenumele)

Conducătorul de proiect _____ (semnătura) **dr. Dumitru CIORBĂ** (numele, prenumele)

Data: 01 decembrie '25

L.Ş.

Anexa 4

Componența echipei conform contractului de finanțare 2025

Cifrul proiectului ***25.80013.8007.02ROMD***

Echipa proiectului conform contractului de finanțare (la semnarea contractului) pentru 2025						
Nr	**Nume, prenume (conform contractului de finanțare)**	**Anul nașterii**	**Titlul științific**	**Norma de muncă sau nr. de ore conform contractului**	**Data angajării**	**Data eliberării**
1.	Ciorbă Dumitru	1978	dr.	13.00	01.09.2025	31.12.2025
2.	Leahu Alexei	1942	dr.hab.	12.25	01.09.2025	31.12.2025
3.	Munteanu Viorel	1984	*f-grad*	13.00	01.09.2025	31.12.2025
4.	Catlabuga Eugeniu	1980	*f-grad*	13.25	01.09.2025	31.12.2025

Modificări în componența echipei pe parcursul anului 2025					
Nr	**Nume, prenume**	**Anul nașterii**	**Titlul științific**	**Norma de muncă sau nr. de ore conform contractului**	**Data angajării**
1.					
2.					
3.					
4.					

Rector U.T.M. _____ **dr. hab. Viorel BOSTAN**
(semnătura) *(numele, prenumele)*

Contabil (economist) _____ **Victoria IOVU**
(semnătura) *(numele, prenumele)*

Conducătorul de proiect _____ **dr. Dumitru CIORBĂ**
(semnătura) *(numele, prenumele)*

Data: 01 decembrie '25

L.Ș.

MINISTERUL EDUCAȚIEI
ȘI CERCETĂRII
AL REPUBLICII MOLDOVA

UNIVERSITATEA TEHNICĂ
A MOLDOVEI

MINISTRY OF EDUCATION
AND RESEARCH
OF THE REPUBLIC OF MOLDOVA

TECHNICAL UNIVERSITY
OF MOLDOVA

MD-2004, CHIȘINĂU, BD. ȘTEFAN CEL MARE ȘI SFÂNT, 168, TEL: 022 23-78-61 | FAX: 022 23-54-41, www.utm.md CONSILIUL ȘTIINȚIFIC

EXTRAS
din Procesul Verbal
al ședinței Consiliului Științific UTM
din 03 decembrie 2025

Prezenți: 14 membri ai Consiliului științific al UTM – Vasile Tronciu, *Prorector pentru cercetare, prof. univ., dr. hab.*; Bostan Ion, *Academician AȘM, prof. univ., dr. hab.*; Bostan Viorel, *Rector UTM, prof. univ., dr. hab.*; Siminiuc Rodica, *Directoare a ȘD UTM, conf. univ, dr.*; Sturza Rodica, *Membru cor. AȘM, prof. univ., dr. hab.*; Ghendov-Moșanu Aliona, *conf. univ., dr. hab.*; Caisîn Larisa, *prof. univ., dr. hab.*; Cepoi Liliana, *Director, Institutul de Microbiologie și Biotehnologie al UTM, conf.univ., dr.*; Gheorghiță Maria, *prof. univ., dr.*; Monaico Eduard; *dr., conf. cercet.*; Țurcanu Dinu, *dr., conf. univ.*; Tîrșu Mihai; *Director Institutul de Energetică UTM, conf. univ., dr.*; Popovici Mihail, *conf. univ., dr.*; Muntean Viorel, *Doctorand UTM*

S-A DISCUTAT: audierea rezultatelor științifice obținute pe parcursul anului 2025 al proiectului din cadrul Concursului „Proiecte complexe bilaterale cu Republica Moldova" pentru anii 2025-2026: ***25.80013.8007.02ROMD „Cercetarea genomică colaborativă privind varietațiile genetice implicate în sănătatea cardiovasculară în Europa de Est"***, Conducător de proiect: ***dr. Dumitru CIORBĂ.***

S-A DECIS: aprobarea rezultatelor științifice obținute pe parcursul anului 2025 al proiectului din cadrul Concursului „Proiecte complexe bilaterale cu Republica Moldova" pentru anii 2025-2026: ***25.80013.8007.02ROMD „Cercetarea genomică colaborativă privind varietațiile genetice implicate în sănătatea cardiovasculară în Europa de Est"***, Conducător de proiect: *dr. Dumitru CIORBĂ.*

Președinte al CȘ UTM,
Vasile TRONCIU, dr. hab., prof. univ.

Secretar al CȘ UTM,
Liliana CEPOI, dr. hab.