Centru Cloud si Big Data pentru Participarea la Cloud-ul European pentru Stiinta Deschisa

Realizarea de servicii si aplicatii informatice pentru suportul activitatii de analiza a datelor de secventiere de noua generatie

George Necula

IFIN-HH

Proiect cofinantat din Fondul European de Dezvoltare Regionala prin Programul Operational Competitivitate 2014-2020
Pentru informatii detaliate despre celelalte programe cofinantate de Uniunea Europeana, va invitam sa vizitati www.fonduri-ue.ro
"*Continutul acestui material nu reprezinta in mod obligatoriu pozitia oficiala a Uniunii Europene sau a Guvernului Romaniei*"

Obiectivele subactivității 2.5

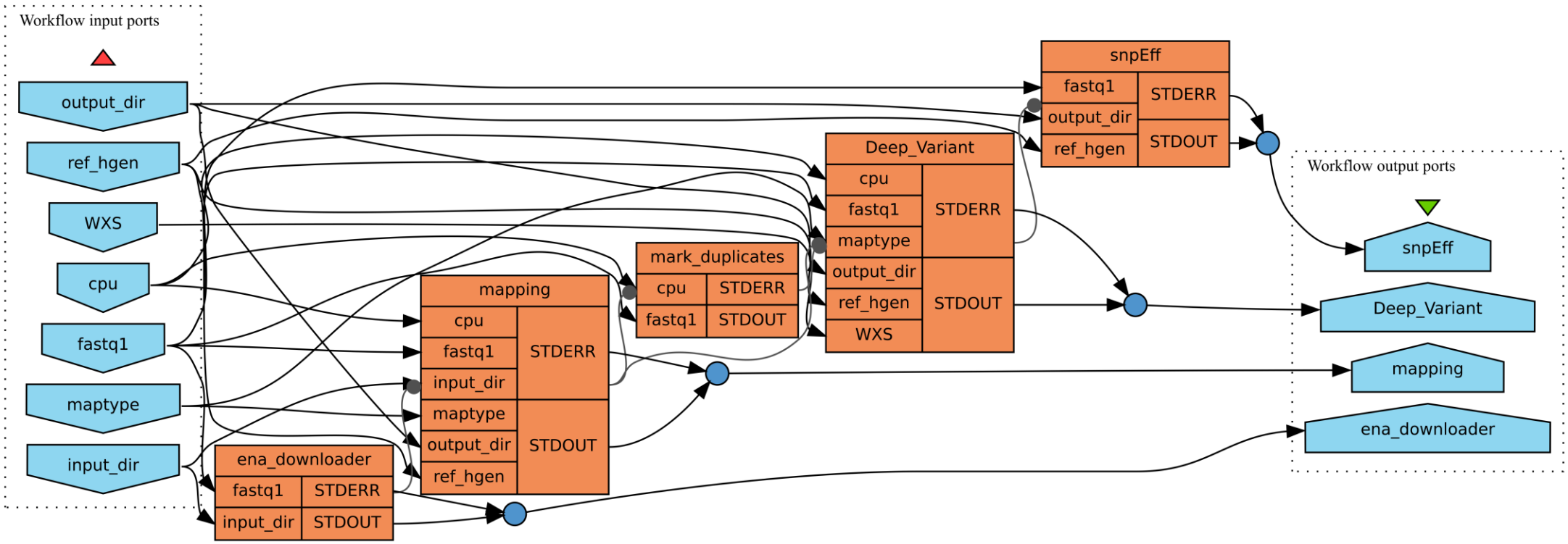
- Dezvoltarea fluxurilor de lucru automatizate pentru analiza bioinformatică a datelor NGS
  - Testarea sistemelor de management ale fluxurilor de lucru (Apache Taverna sau Galaxy) pe clusterul HPC si centrul de resurse Cloud al DFCTI (CLOUDIFIN)
  - Testarea programelor pentru analiza bioinformatică a datelor brute NGS (WGS/WES) pe infrastructura de calcul DFCTI
  - Proiectarea, testarea si optimizarea fluxurilor de lucru pentru analiza bioinformatică a datelor NGS
  - Validarea rezultatelor obținute cu ajutorul fluxurilor de lucru pentru identificarea variantelor germinale/somatice
- Integrarea fluxurilor de lucru în platforma online de analiză a datelor NGS

Analiza bioinformatică a datelor NGS - etape generale

- Controlul calității (FastQC/MultiQC, Fastp, Trimmomatic, etc.)
- Alinierea citirilor la genomul de referință uman i.e. GRCh37/GRCh38 (BWA, SOAP3, Bowtie2, etc.)
- Identificarea variantelor scurte SNP/SNV și indel din linie germinală/somatică
  - Variante germinale: HaplotypeCaller (GATK4), DeepVariant, SpeedSeq
  - Variante somatice: MuTect2 (GATK4), Strelka2, VarScan
- Filtrarea si adnotarea variantelor identificate (snpeff, snpsift, annovar, OpenCRAVAT etc.)
  - Baze de date pentru adnotare: OMIM, ClinVar, COSMIC, Cancer Genome Interpreter, etc.

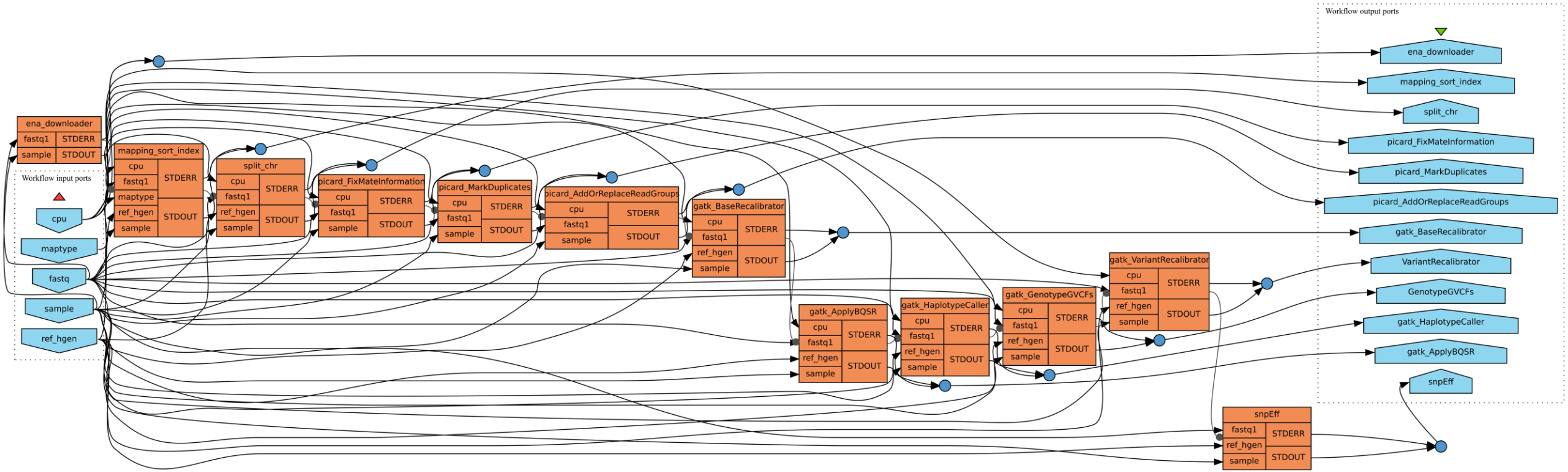
Flux de lucru DeepVariant (DV)

- Identifica variantele SNP si indel germinale cu ajutorul unui algoritm *deep learning* construit pe baza Nvidia TensorFlow
- Clasifică tensori folosind o rețea neuronală convoluțională



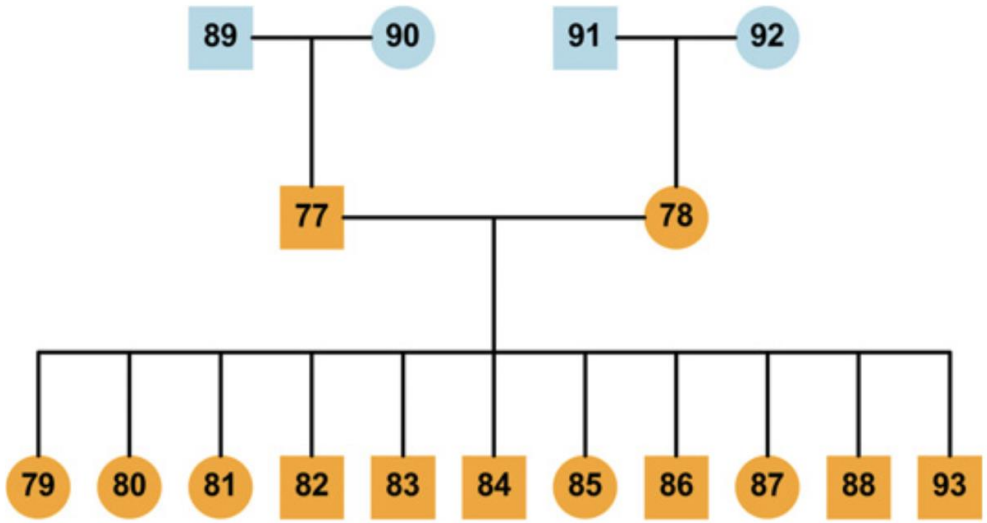
Flux de lucru HaplotypeCaller (GATK4)

- Identifică SNP și indel din linia germinală prin reasamblarea locală a haplotipurilor
- Utilizează un modelul Markov cu aplicabilitate în bioinformatică (Pair Hidden Markov Model - PHMM)



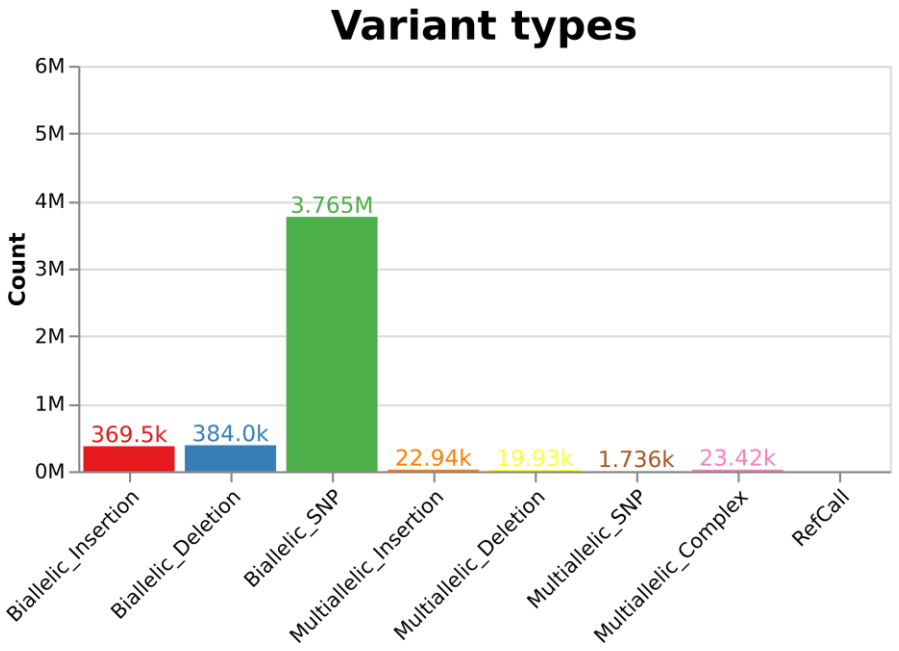
Proba de referință

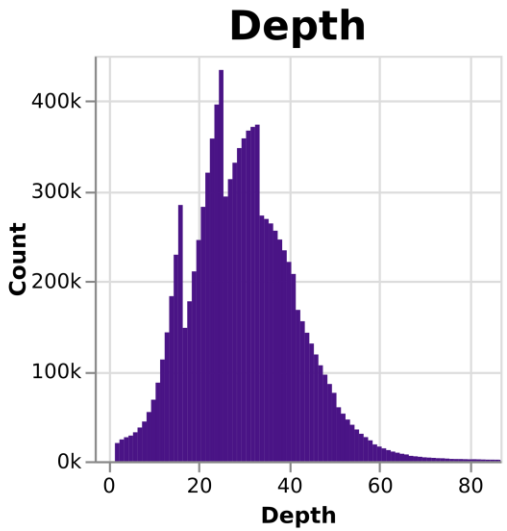
- Consorțiul *Genome in a Bottle* (GiaB) a dezvoltat un set de variante de „aur” (proba NA12878), care este utilizat pe scară largă in etapele de dezvoltare si testare a pipeline-uri de identificare a variantelor
- Milioane de SNP și indel scurte (1-50 pb) au fost trasate de-a lungul a trei generații (membrii pedigriului CEPH/Utah) - pentru a crea un catalog de variante cu un grad ridicat de încredere
- Genomul derivat din linia celulară (limfoblastoide transformate cu EBV din PBL) NA12878/HG001 a devenit unul dintre cele mai studiate și mai bine caracterizate genomuri umane
- Datele brute WGS obținute din acesta linie celulară (cod de acces ENA SRR6794144) cu ajutorul platformei Illumina HiSeq 4000, cu o acoperire de ~37x, au fost folosite pentru validarea ambelor fluxuri de lucru

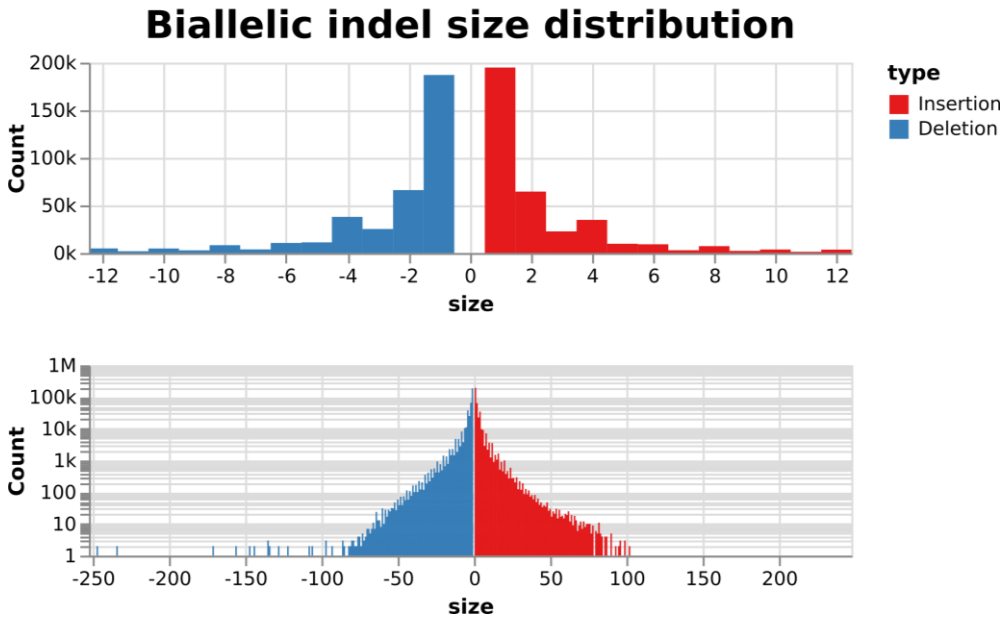


Eberle et al. 2017

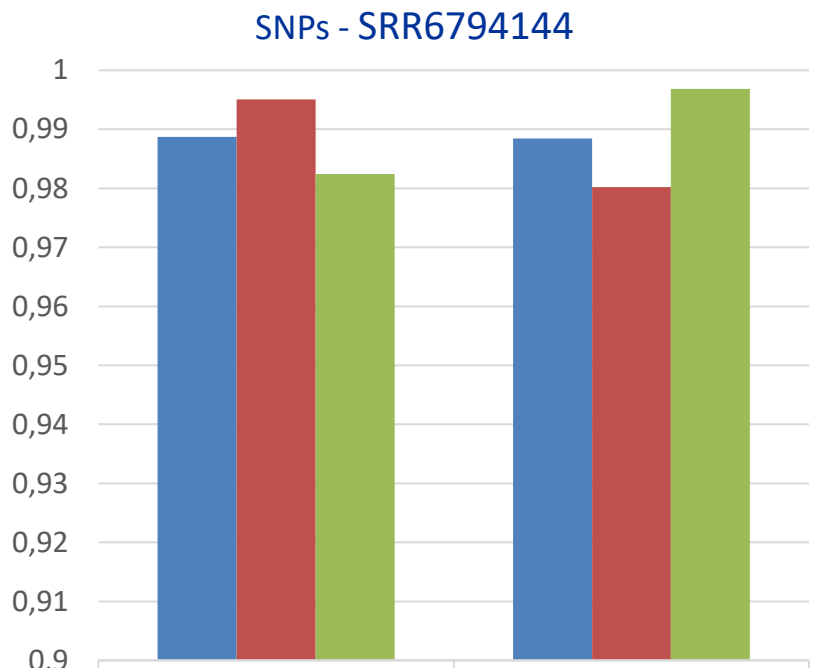
Analiza bioinformatică a setului de date SRR6794144 (DV)



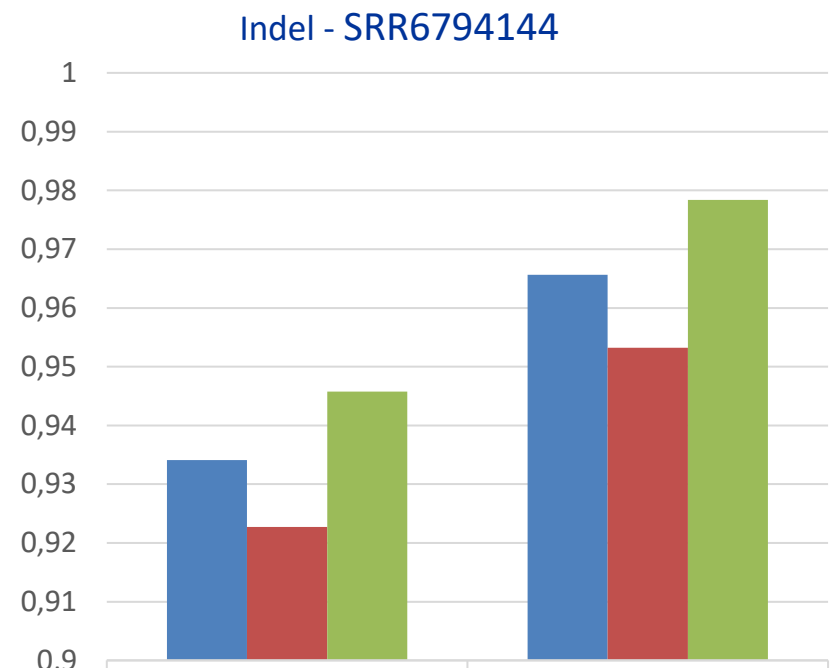




Validarea fluxurilor de lucru pentru identificarea variantelor germinale (WGS)



	GATK4	DeepVariant
F1_Score	0,988697	0,988443
Recall	0,99506	0,98018
Precision	0,982416	0,996846



	GATK4	DeepVariant
F1_Score	0,934092	0,965639
Recall	0,922698	0,953199
Precision	0,945771	0,978409

Validarea fluxului de lucru DeepVariant pentru identificarea variantelor germinale (WES)

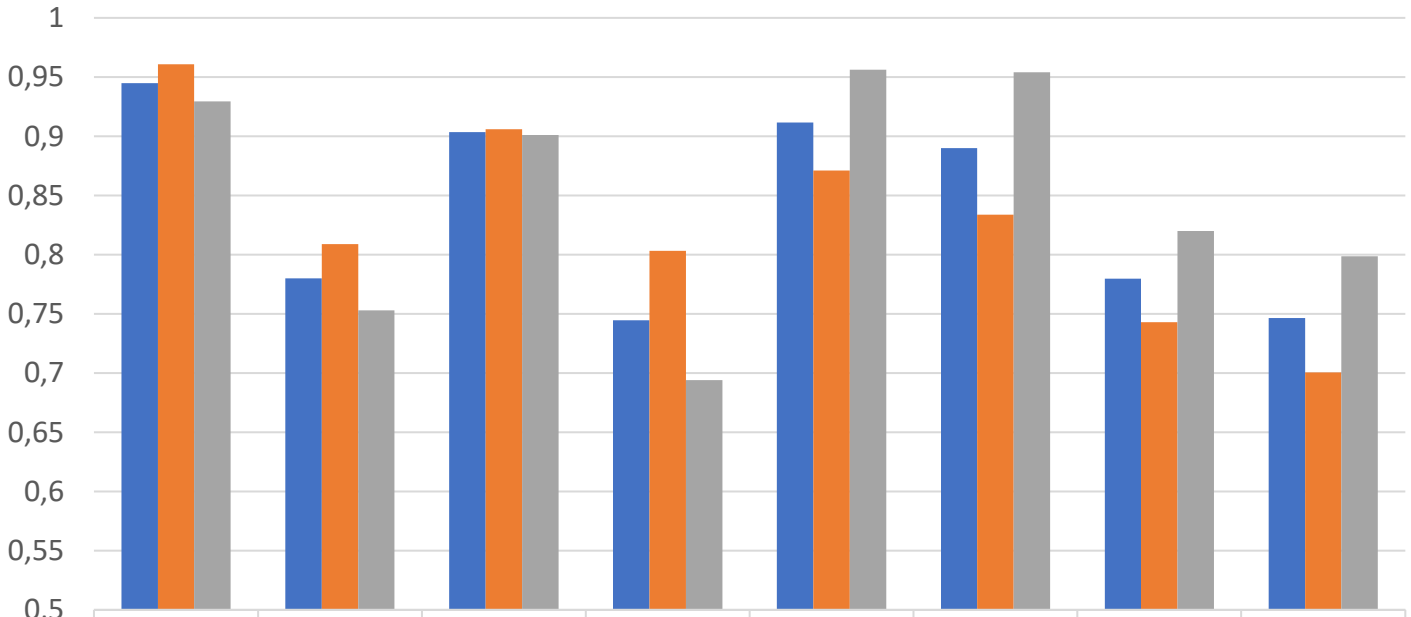
Sample	Variant type	Recall	Precision	F1 Score
HG001	SNP	0.993206	0.998413	0.995803
HG001	Indel	0.992191	0.997548	0.994862
HG002	SNP	0.99436	0.998878	0.996614
HG002	Indel	0.983444	0.99618	0.989771
HG003	SNP	0.993699	0.998115	0.995902
HG003	Indel	0.981077	0.992658	0.986834
HG004	SNP	0.994963	0.998994	0.996974
HG004	Indel	0.989831	0.997079	0.993441
HG005	SNP	0.99653	0.999321	0.997924
HG005	Indel	0.989389	0.995939	0.992653
HG006	SNP	0.98545	0.995834	0.990615
HG006	Indel	0.957627	0.991266	0.974156
HG007	SNP	0.984973	0.995764	0.990339
HG007	Indel	0.963675	0.99345	0.978336

Fluxuri de lucru pentru identificarea variantelor SNP si indel somatice

- Pentru identificarea variantelor somatice SNP, indel si SV (variante structurale) se realizează analiza comparativă a două probe NGS: una obținută din celule normale si cealaltă din celule tumorale
- Nu se pot utiliza date NGS obținute de la pacienți pentru validarea metodelor de identificare atât a variantelor scurte somatice sau a variantelor structurale (SV), datorită faptului ca nu se pot determina variante adevărate a priori
- Identificarea variantelor somatice la un nivel ridicat de precizie si sensibilitate este foarte complicată, în principal datorită eterogenității tumorale (i.e. populații celulare subclonale), artefactelor de secvențiere si aliniere, dar si din cauza contaminării cu celule normale
- Prin intermediul inițiativei ICGC-TCGA DREAM Somatic Mutation Calling Challenge s-a încercat simularea realista a unor probe tumorale si stabilirea unui consens intre rezultatele diferitelor metode de analiză a variantelor somatice
- S-au folosit seturile de date brute DREAM Nr. 1-4 pentru a valida fluxurile de lucru MuTect2 si Stelka2
  - DREAM #1 - HCC1143 BL, SNV, VAF: n/a
  - DREAM #2 - HCC1143 BL, SNV, VAF: n/a
  - DREAM #3 - HCC1143 BL, SNV & INDEL, VAF: 50%, 33%, 20%
  - DREAM #4 - CPCG0102R, SNV & INDEL, VAF: 30%, 15%

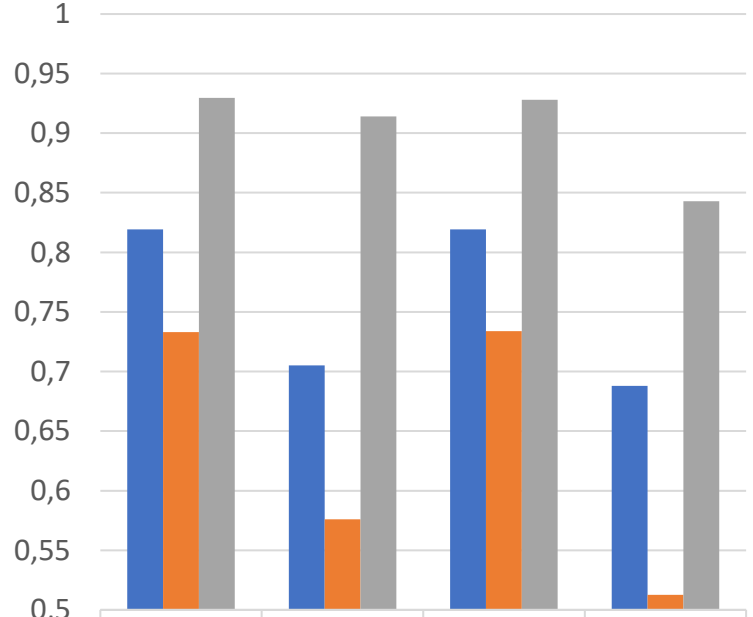
Validarea fluxurilor de lucru pentru identificarea variantelor SNP si Indel din linie somatică

SNP - DREAM Challenge No. 1-4



	#1 Mutect2	#1 Strelka2	#2 Mutect2	#2 Strelka2	#3 Mutect2	#3 Strelka2	#4 Mutect2	#4 Strelka2
F1 score	0,9448	0,77988	0,90354	0,74456	0,91167	0,88986	0,77967	0,74638
Recall	0,9607	0,80887	0,90604	0,80309	0,87118	0,83373	0,74305	0,70049
Precision	0,92943	0,75289	0,90105	0,69399	0,95611	0,95409	0,8201	0,79869

Indel - DREAM Challenge No. 3-4



	#3 Mutect2	#3 Strelka2	#4 Mutect2	#4 Strelka2
F1 score	0,819355	0,70522	0,81927	0,687855
Recall	0,732925	0,57588	0,733885	0,512535
Precision	0,92965	0,914135	0,92799	0,842855

UNIUNEA EUROPEANĂ

Instrumente Structurale 2014-2020

CECBID-EOSC

https://cecbid-eosc.ifin.ro

VA MULTUMESC PENTRU ATENTIE !

Conferinta de incheiere,
28.07.2023

IFIN-HH