### Václav Svatoň, Sebastián Husár, Martin Mokrejš, Jan Martinovič IT4Innovations, VŠB - Technical University of Ostrava **Czech Republic**



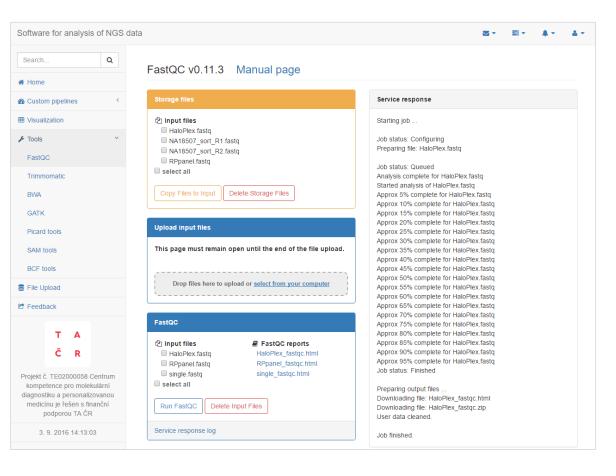
**IT4INNOVATIONS** NATIONAL SUPERCOMPUTING CENTER





The primary objective of the **Center of competence for molecular diagnostics and personalized medicine**, "MOLDIMED" is to achieve critical mass of experts, institutions, and knowledge in research, development, protection of IP, certification, technology transfer, and commercialization of in vitro diagnostics, in order to create market oriented flexible national network in area of diagnostic, prognostic and predictive biomarkers and to enable further development of personalized medicine.

The methods of **massive parallel sequencing** (MPS) have started to play a key role in clinically oriented research and DNA diagnostics of molecular pathologies. Thus, the concept of **personalized medicine** replaces low-throughput classical approaches, which are often methodically time -consuming to cover long DNA regions. MPS methods, especially WES generate huge amount of data, which must be further processed. Therefore the MPS processing platform for the next generation DNA sequencing (NGS) and data processing in detection of hereditary and somatic DNA variants was created.



Specialized platform for the next generation DNA sequencing with custom annotation tool and a number of opensource bioinformatics software was created. Platform is deployed at I4Innovations and also at the Institute of Molecular and Translational Medicine. Both instances are utilizing HEAppE to access the local HPC infrastructure.

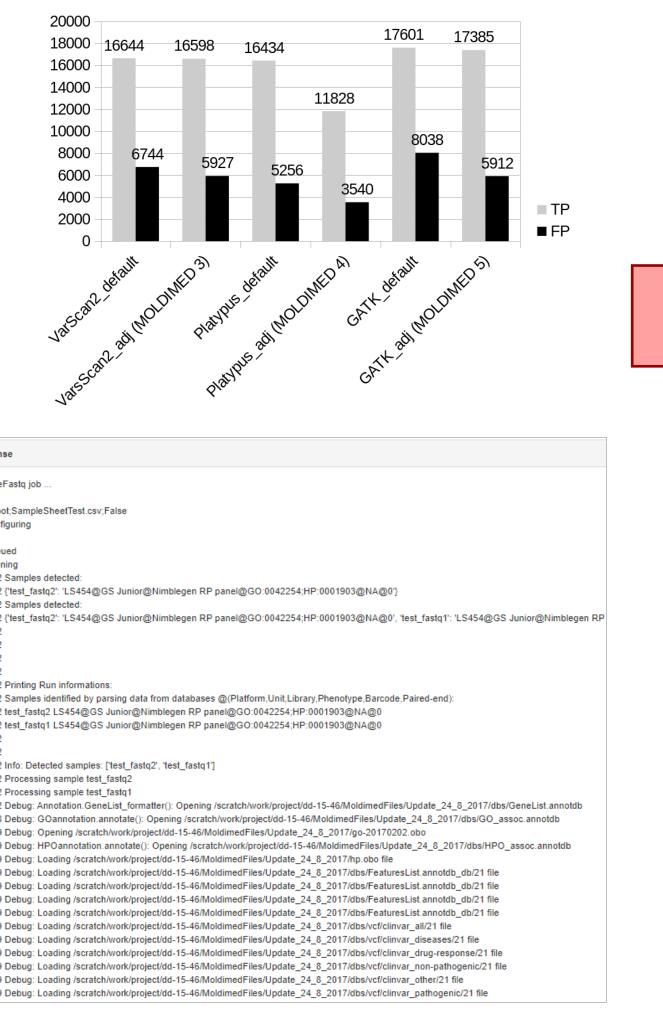
## **Step 1** Analytical Pipeline & Annotation Tool

Pipeline integrates custom developed **annotation tool for DNA variants**. The annotation program is designed for human genetic variants annotation, but the functionality was successfully tested on other types of genomes with the different ploidy. One of the advantages over existing annotation programs is the effective phenotypic prioritization of variants on the basis of ontological relationships allowing the effective annotation of genetic variants in the broad range of human diseases.

- Custom **annotation tool** for DNA variants
- Designed for **human genetic variants** annotation
- Tested on other types of **genomes with the different ploidy**
- Effective **phenotypic prioritization** of variants
- Effective annotation of genetic variants

te or edit SampleSheet

• Applicable for the **broad range of human diseases** 



a type tring		v					Job sta
umn n nter n dd col	ame of column	nn					Job sta Oct 02 Oct 02 Oct 02 Oct 02 Oct 02
+	Sample_name	Phenotype	Platform	Flowcell_name	Platform_unit	Library	Oct 02 Oct 02
7							Oct 02 Oct 02
۳ ش	CARDIO01	g	Illumina		D00183	NexteraTrusight Cardio	Oct 02 Oct 02
Ŵ	CARDIO02	g	Illumina		D00183	NexteraTrusight_Cardio	Oct 02 Oct 02
۳ ش	CARDIO03	g	Illumina		D00183	NexteraTrusight_Cardio	Oct 02 Oct 02
Ŵ	CARDIO04	g	Illumina		D00183	NexteraTrusight_Cardio	Oct 02 Oct 02
Ì	CARDIO05	g	Illumina		D00183	NexteraTrusight_Cardio	Oct 02 Oct 02 Oct 02
Û	CARDIO06	g	Illumina		D00183	NexteraTrusight_Cardio	Oct 02 Oct 02 Oct 02
۵	CARDIO07	a	Illumina		D00183	NexteraTrusight Cardio	Oct 02
jes: ' ave fil	1 1 of 1 e Create folders						Oct 02 Oct 02 Oct 02 Oct 02 Oct 02 Oct 02 Oct 02



### Acknowledgements

This work was supported by The Ministry of Education, Youth and Sports from the IT4Innovations infrastructure which is supported from the Large Infrastructures for Research, Experimental Development and Innovations project "IT4Innovations National Supercomputing Center - LM2015070" and project number CZ.02.1.01/0.0/0.0/16\_013/0001791 within the Operational Programme Research, Development and Education and by TAČR grant TE02000058.

non-reference homozygous

# Web based graphical user interface for data analysis tools for high capacity DNA sequencing and human genome variants annotation



HPC-as-a-Service is a well known term in the area of high performance computing. It enables users to access an HPC infrastructure without a need to buy and manage their own physical servers or data center infrastructure. Through this service academia and industry can take advantage of the technology without an upfront investment in the hardware. This approach further lowers the entry barrier for users who are interested in utilizing massive parallel computers but often do not have the necessary level of expertise in the area of parallel computing.

To provide this simple and intuitive access to the supercomputing infrastructure an in-house application framework called HEAppE has been developed. HEAppE's universally designed software architecture enables unified access to different HPC systems through a simple object-oriented client-server interface using standard web services, REST API or Jupyter notebooks. Thus providing HPC capabilities to the users but without the necessity to manage the running jobs form the command-line interface of the HPC scheduler directly on the cluster.

The IT4Innovations national supercomputing center operates four supercomputers: Anselm (94 TFLOP/s, installed in 2013), Salomon (2 PFLOP/s, installed 2015), Barbora (849 TFLOP/s, installed 2019) and a special system for AI computation, DGX-2 (2 PFlop/s in AI, installed in 2019). The supercomputers are available to academic community within the Czech Republic and Europe and industrial community worldwide. All supercomputers are available to users via HEAppE Middleware.

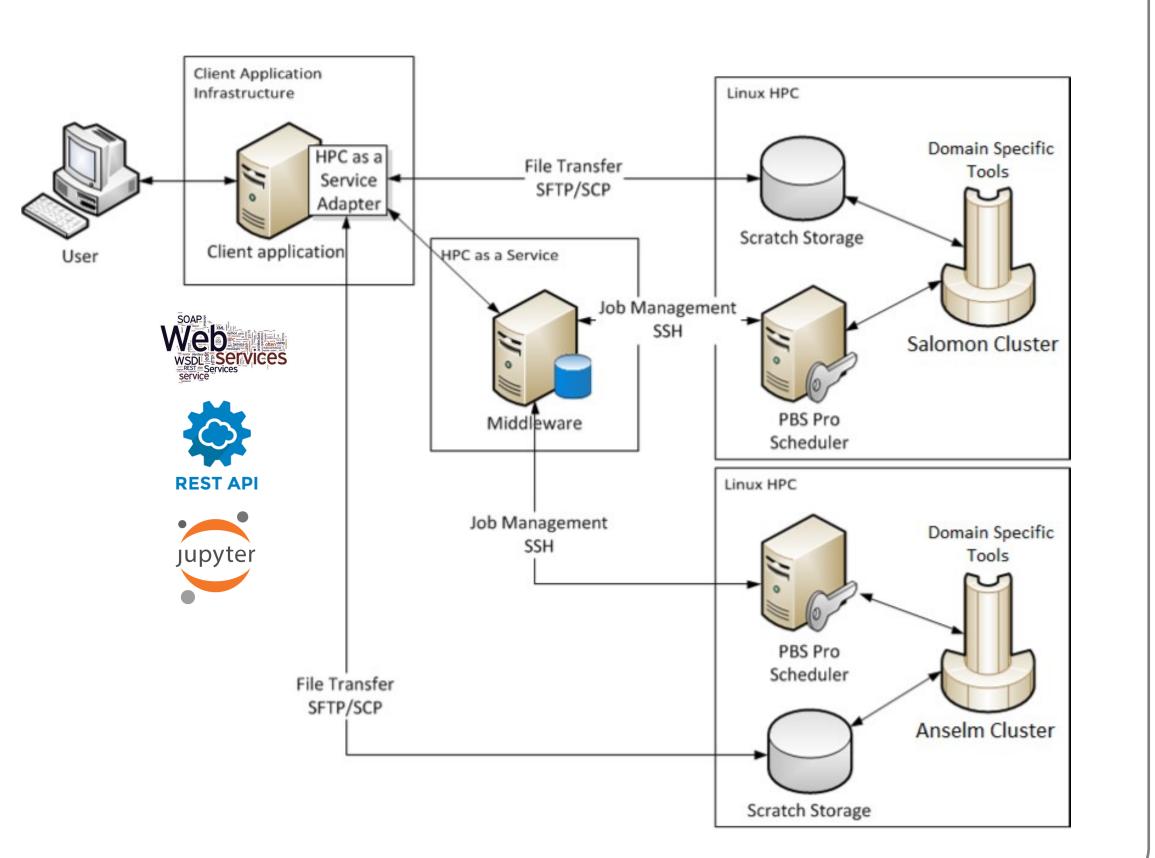
HEAppE Middleware is licensed under the GNU General Public License v3.0 http://www.heappe.eu support.heappe@it4i.cz

### **Step 2** Result Interpretation & Visualization **Step 3** Quality Control & Kits • Create **custom kit definitions** for quality control checks NIH U.S. National Library of Medicine National Center for Biotechnology Information Quality Regior • Import .vcf files or manually create **positive/negative** 🔲 gene 🔲 non-genic 🔲 chr not included = • Enter Quality dbSNP sh Genetic Variations Structural variant controls Coverage SNV 🔲 INDEL = • Enter Coverage • Insert created controls into a specific **sample** Welcome to the Reference SNP (rs) Report Genotype All alleles are reported in the Forward 🔲 non-reference heterozygous 📄 non-reference homozygous 📄 reference heterozygous GO or HPO number HGVS names are in the Aliases ta Enter GO or HPO ni • Include selected samples to a specific kit definition Transcript eference SNP (rs) Repor no duplicates Common db SNP ← Switch to classic site • Share your kit definition with other users All ORS only Without RS Feature NULL UTR5 intron exon CDS spliceSite UTR3 rs115173026 Gene names (Import gene names from csv • Run quality control check for your .vcf files against file and use semicolon (;) as a separator) Organism Amino acid effect Homo sapien: NULL INDEL synonymous nonsynonymous chr1:1020217 (GRCh38.p12) a selected kit definition G>C / G>T Choose File No file chosen /ariation Type 49,877,730 bp 49,877,740 bp 49,077,750 bp 49,077,760 bp Parse CSV file Frequency T=0.29206 (36674/125568, TOPMED del(1) G -</t T=0.3428 (17342/50582, GnomAD\_exome) Γ=0.3274 (9952/30396, GnomAD) (<u>+ 6 more</u>) Filter results Save filter Reset Filters Genomic Placements Variant Details Sequence name IGS kit definition **Clinical Significance** AGRN RefSegGene (LRG 198) Frequency 街 Input files AGRN RefSegGene (LRG 198) test.vcf Aliases GRCh37.p13 chr Search Delete Files GRCh37.p13.chr Cubmissis Available assemblies Stop Genotype Start hg18 Available sample names reference heterozygous 1672877 1672877 reference heterozygous 672877 167287 Available kit names Kit for testing mutations in it4i 1672877 1672877 reference heterozygou Positive 🗹 672877 167287 reference heterozygou Parse input file Parse input file to filled table reference heterozygous 1673060 167306 CosmicID NGS quality check result 1673060 1673060 reference heterozygous reference heterozygous 1673060 1673060 💉 🔟 hg18 955597 rs115173026 non-reference homozygous 1735814 1735814 Yes) 💉 🔟 977330 rs2799066 hg18 non-reference homozygous 1735814 1735814 1 non-reference homozygous 1735814 1735814 977445 hq18 C 1735814 1735814 non-reference homozygous 💉 🔟 hq18 977453 Α non-reference homozygous 1735932 1735932 💉 🔟 ha18 977495 non-reference homozygous 1735932 1735932 🖍 🔟 🛛 hg18 977502 non-reference homozygous 1735932 1735932 non-reference homozygous 1735932 173593 Pages: 1 2 3 4 5 6 7 8 9 10 ... Next Last 1 of 15 non-reference homozygous 1735932 1735932 Fullscreen view non-reference homozygous 6184749 6184749 \_\_\_\_\_ Save to database 6184804 6184804

- **Visualization** of pipeline results
- Dynamic **filtering**
- Saved favorite filters
- **Full-text** search functionality
- Gene names CSV file import for searching

Region				
Chromosome				
VCF_position	ie 🌩	VCF_position	GenotypeConformation $\Rightarrow$	StructuralVaria
VCF_reference		1672876	ТС/Т	INDEL
VCF_alteration		1672876	тс/т	INDEL
GenotypeConformation	-			
StructuralVariant		1672876	ТС/Т	INDEL
Genotype		1672876	ТС/Т	INDEL
<ul><li>✓ Start</li><li>✓ Stop</li></ul>		1673060	G/A	SNV
MinRep_reference		1673060	G/A	SNV
MinRep_alteration		1673060	G/A	SNV
HGV		1673060	G/A	SNV
GQ DP		1735814	C/C	SNV
TranscriptName		1735814	C/C	SNV
Strand		1735814	C/C	SNV
ExFeature		1735814	C/C	SNV
Sense		1735932	G/G	SNV
AminoAcidChange		1735932	G/G	SNV
GeneName clinvar_Benign		1735932	G/G	SNV
clinvar_benign		1735932	G/G	SNV
clinvar_drug_response				
clinvar_Likely_benign	_	1735932	G/G	SNV
clinvar_Likely_pathogenic		6184749	C/C	SNV
clinvar_non_pathogenic		6184804	G/G	SNV

## **HEAppE: High-End Application Execution Middleware**





Petr Vojta, Zuzana Macečková, Marián Hajdúch IMTM, Palacky University Olomouc **Czech Republic** 



