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

VSB TECHNICAL UNIVERSITY OF OSTRAVA | IT4INNOVATIONS NATIONAL SUPERCOMPUTING CENTER

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

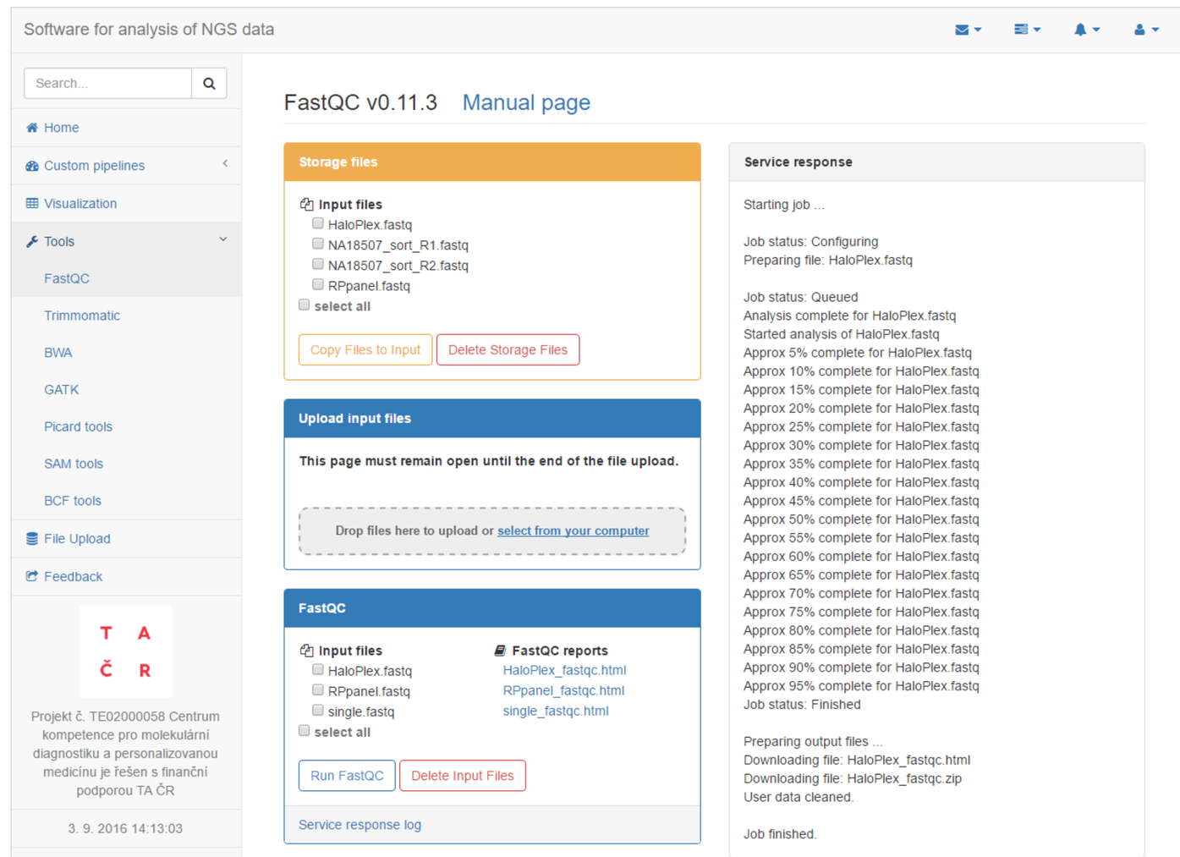
Petr Vojta, Zuzana Macečková, Marián Hajdúch
IMTM, Palacky University Olomouc
Czech Republic

IMTM

MOLDIMED

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.



T A Č R

Specialized platform for the next generation DNA sequencing with custom annotation tool and a number of open-source 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.

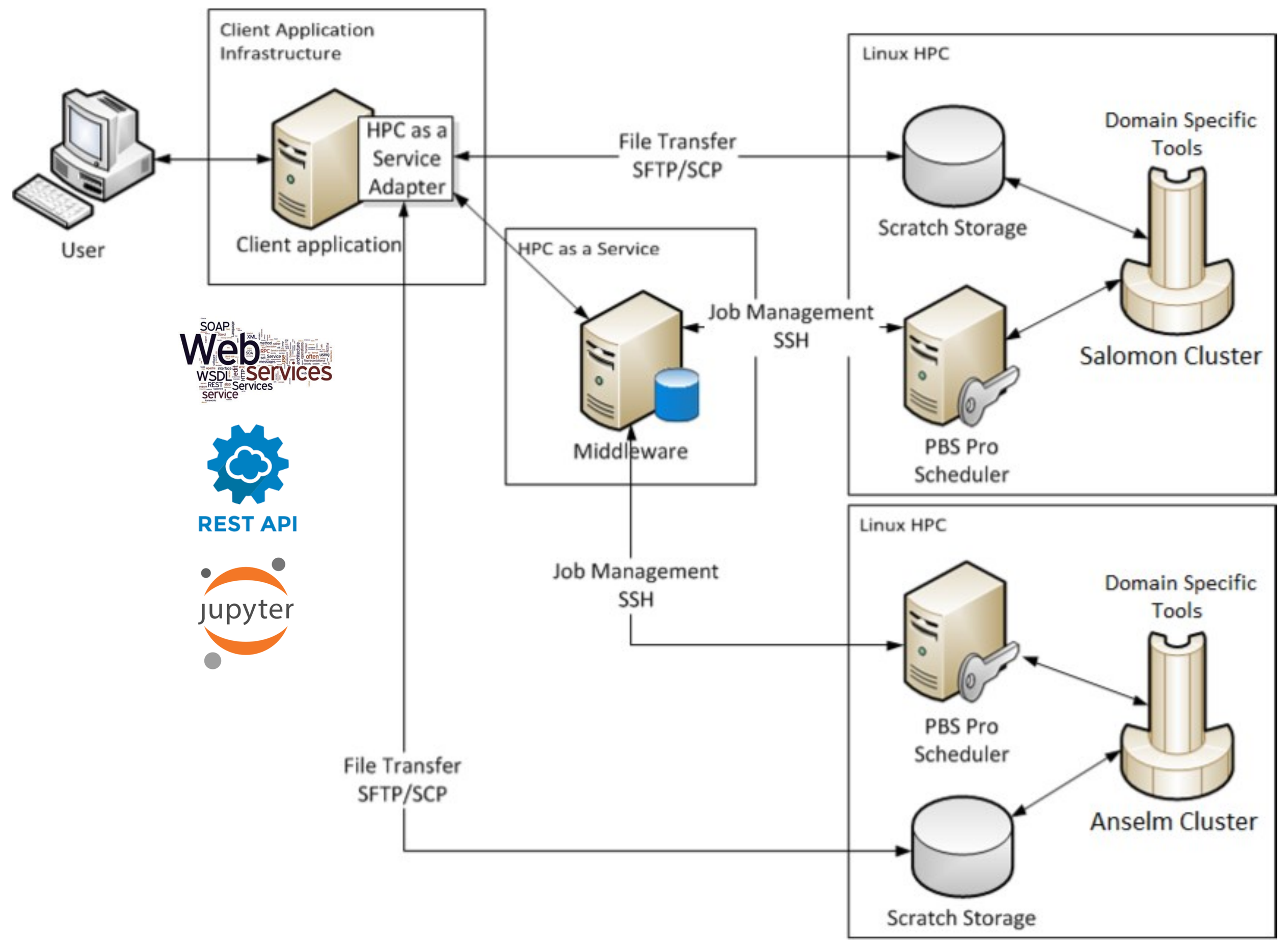
HEAppE: High-End Application Execution Middleware

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**.

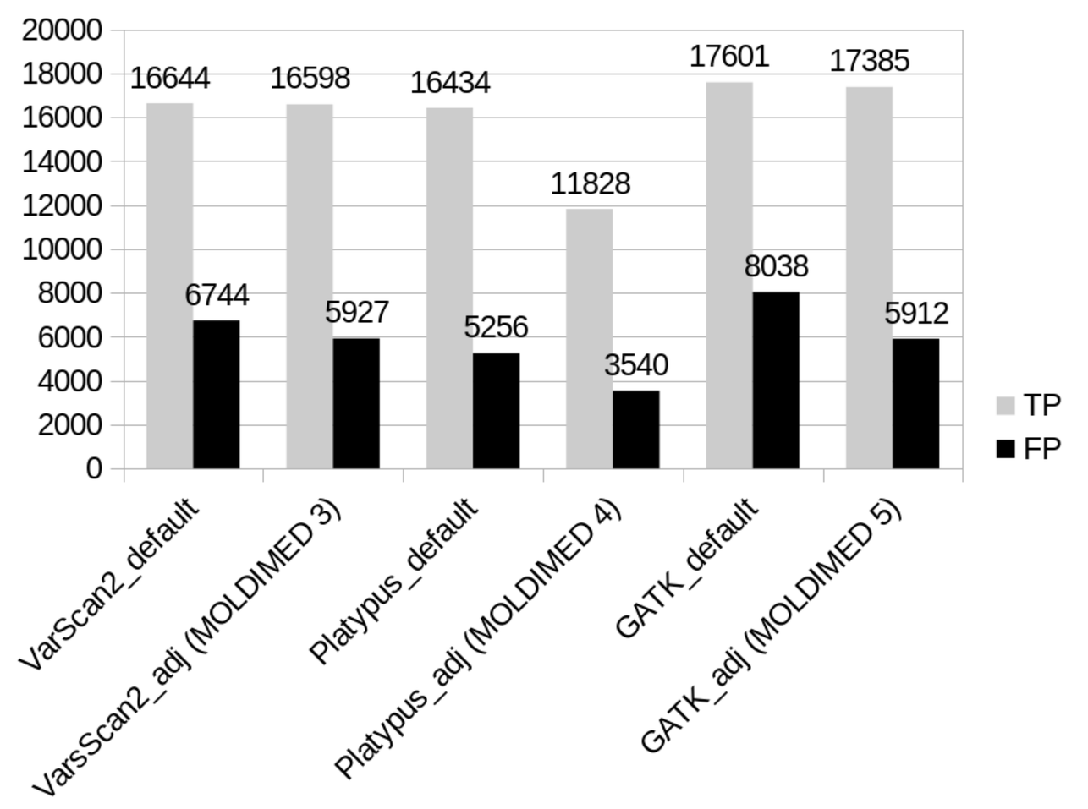
http://www.heappe.eu   support.heappe@it4i.cz

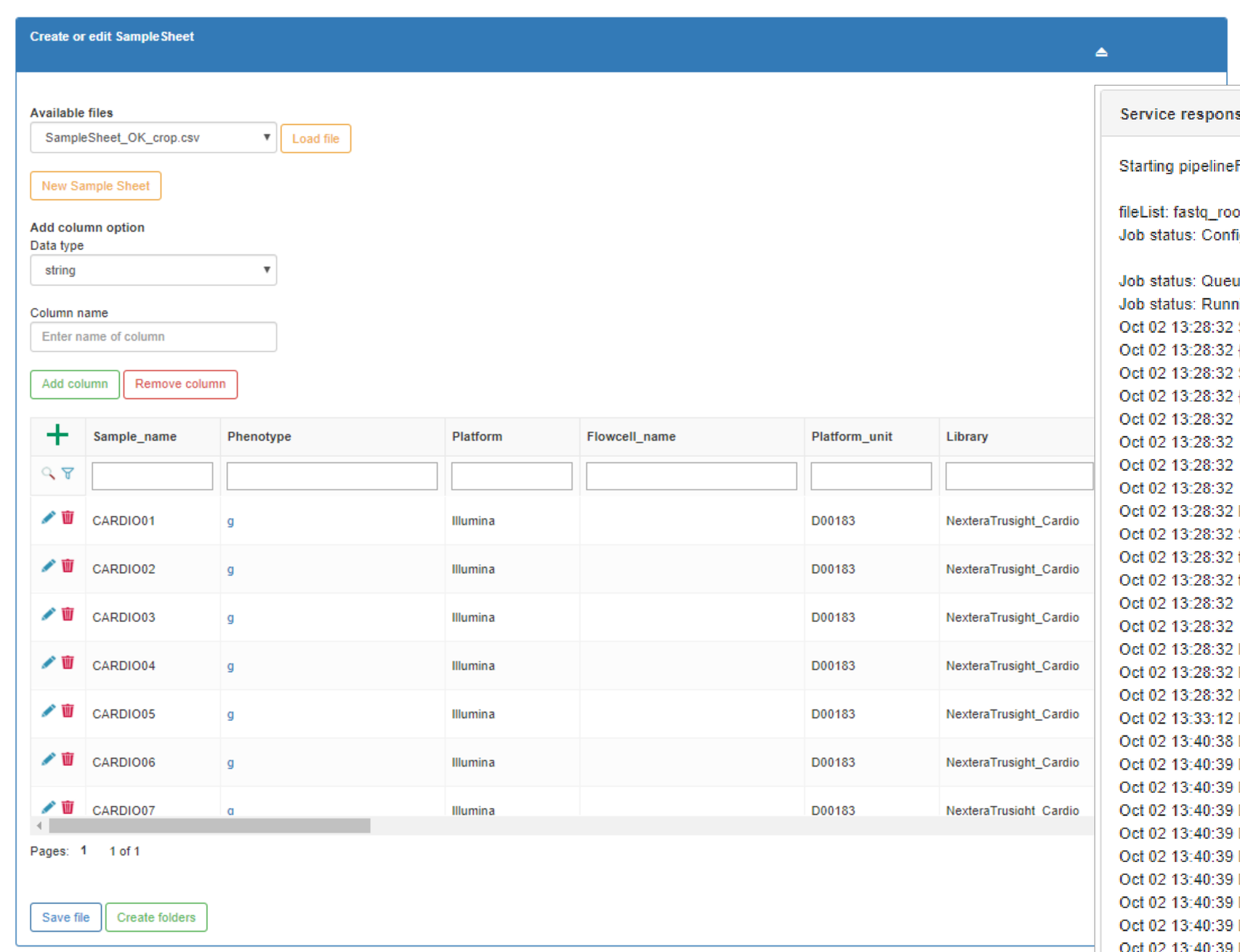


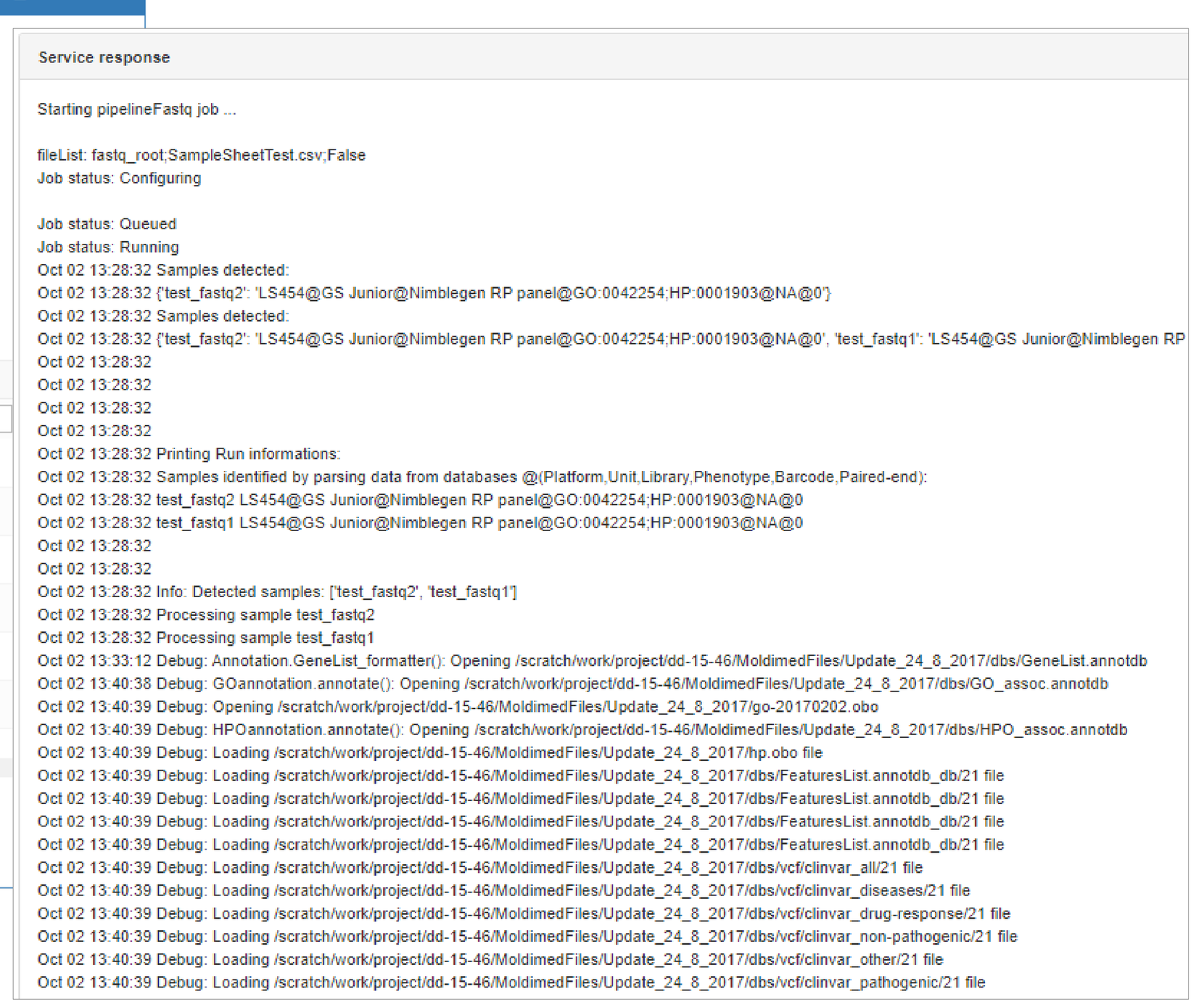
Step 1 Analytical Pipeline & Annotation Tool

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

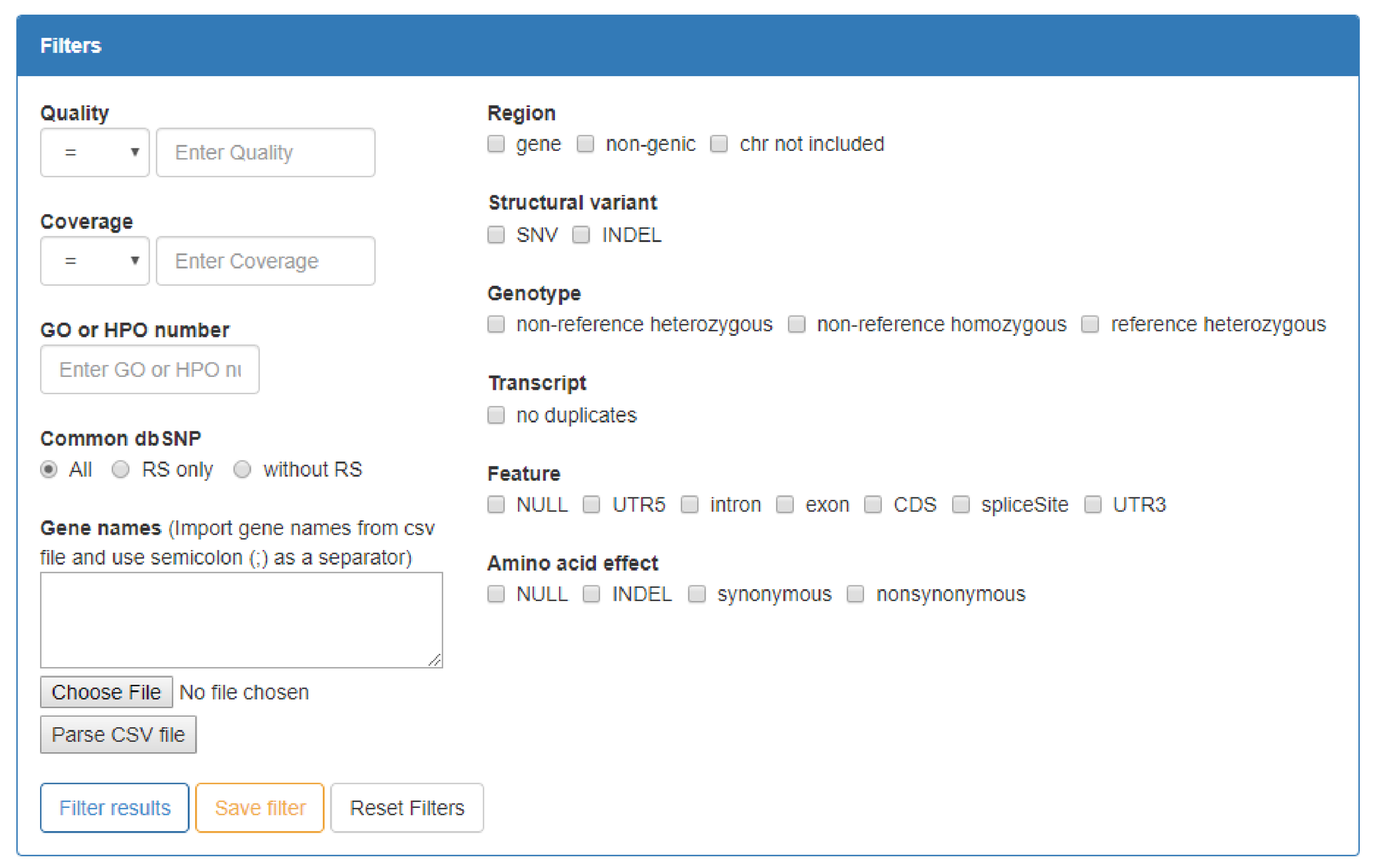


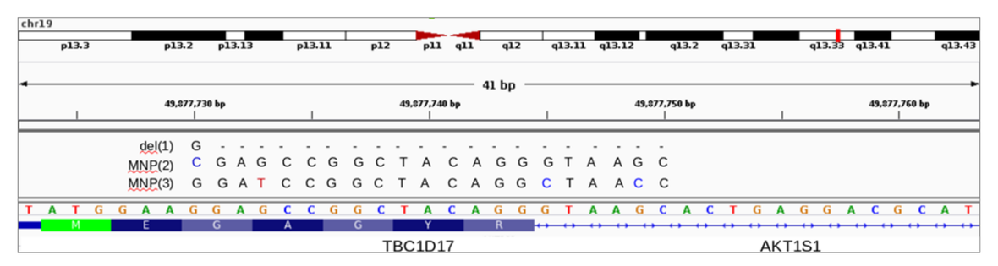


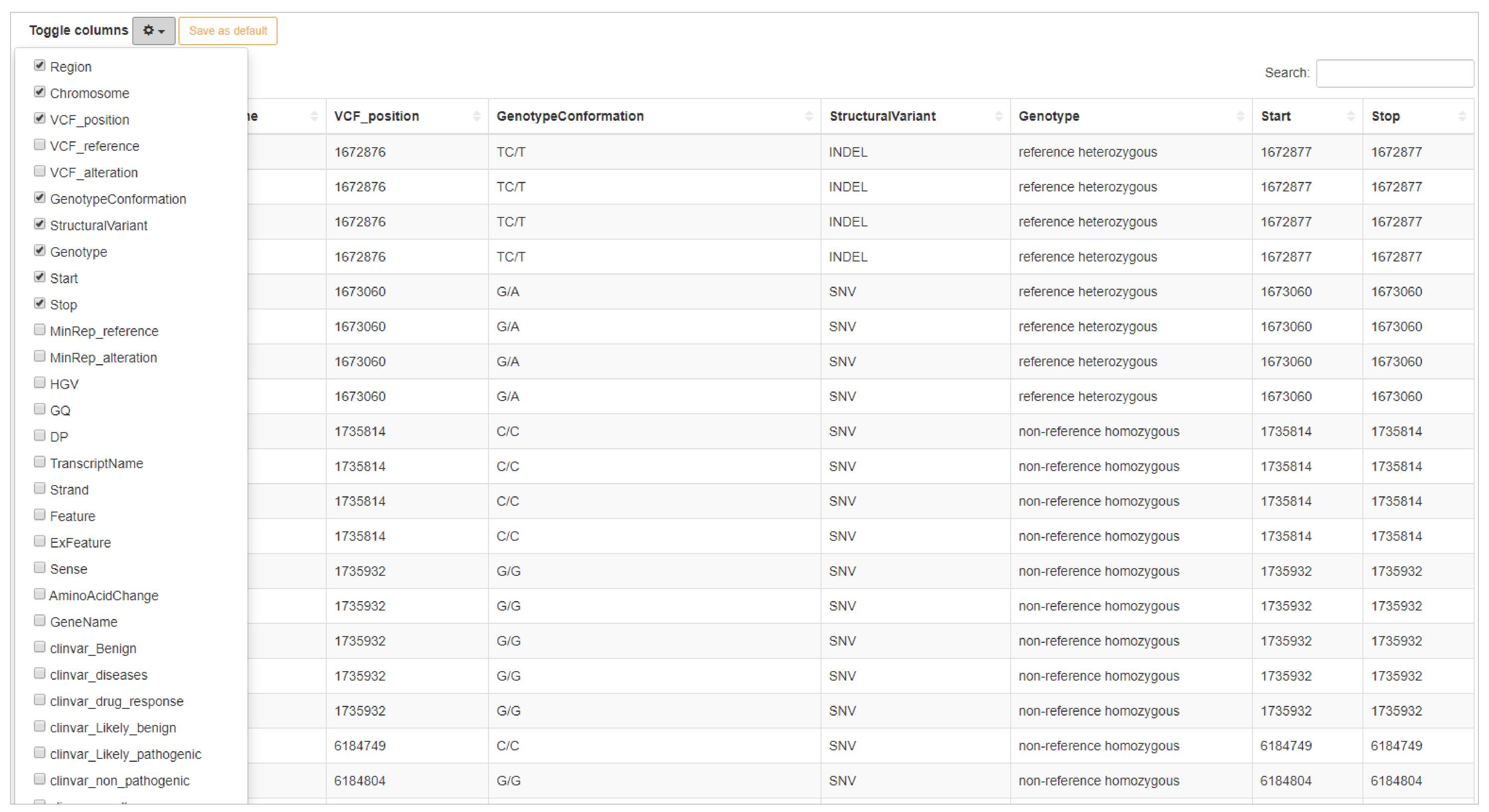


Step 2 Result Interpretation & Visualization

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

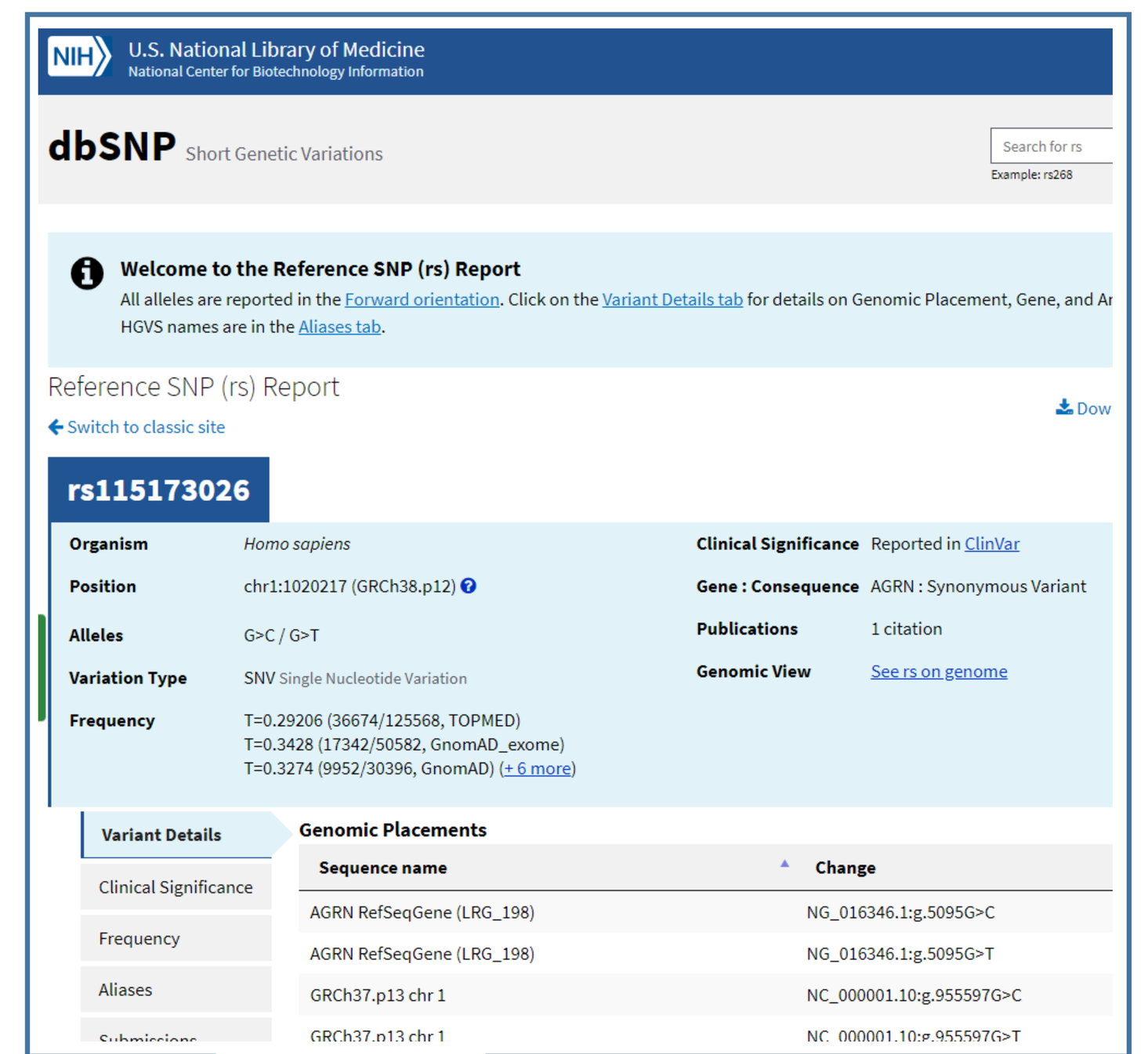


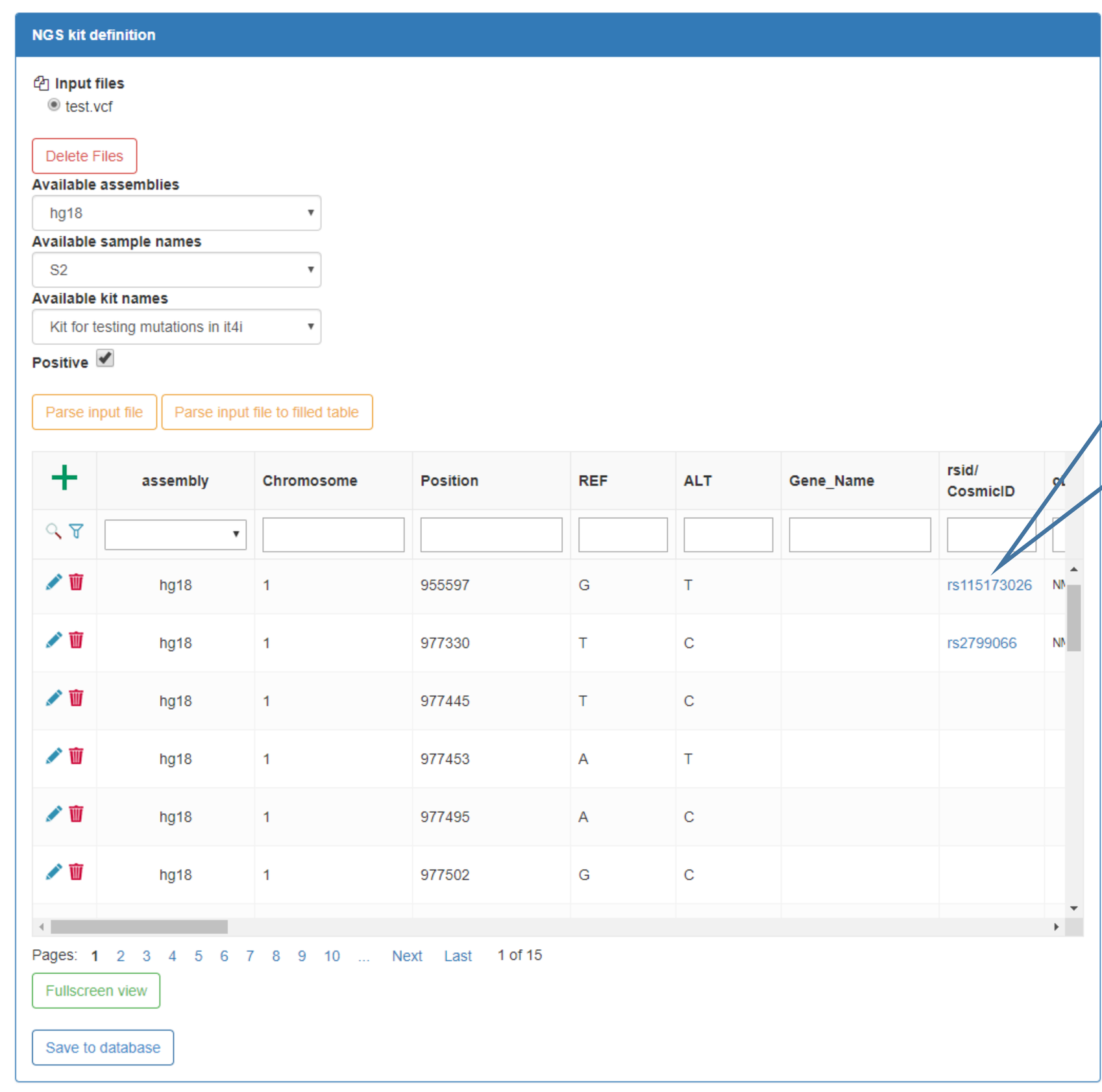


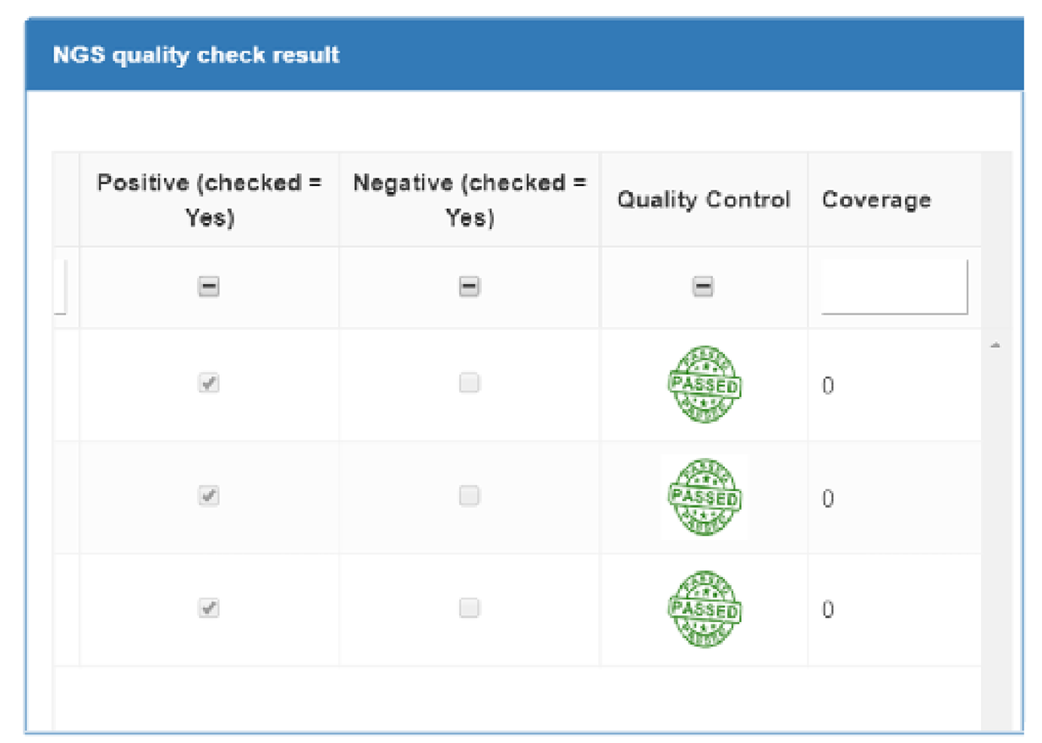


Step 3 Quality Control & Kits

- Create **custom kit definitions** for quality control checks
- Import .vcf files or manually create **positive/negative controls**
- Insert created controls into a specific **sample**
- Include selected samples to a specific **kit definition**
- **Share your kit** definition with other users
- Run **quality control check** for your .vcf files against a selected kit definition







Acknowledgements

This work was supported by The Ministry of Education, Youth and Sports from the National Programme of Sustainability (NPS II) project "IT4Innovations excellence in science - LQ1602", by 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.

MINISTRY OF EDUCATION, YOUTH AND SPORTS