GENOMENAL

Новосибирский государственный университет
*НАСТОЯЩАЯ НАУКА

NOVEL® Software Systems

25 ФОНД СОДЕЙСТВИЯ ИННОВАЦИЯМ

Dmitry Shtokalo, PhD

GENOMENAL platform: ecosystem solution for digital genomics

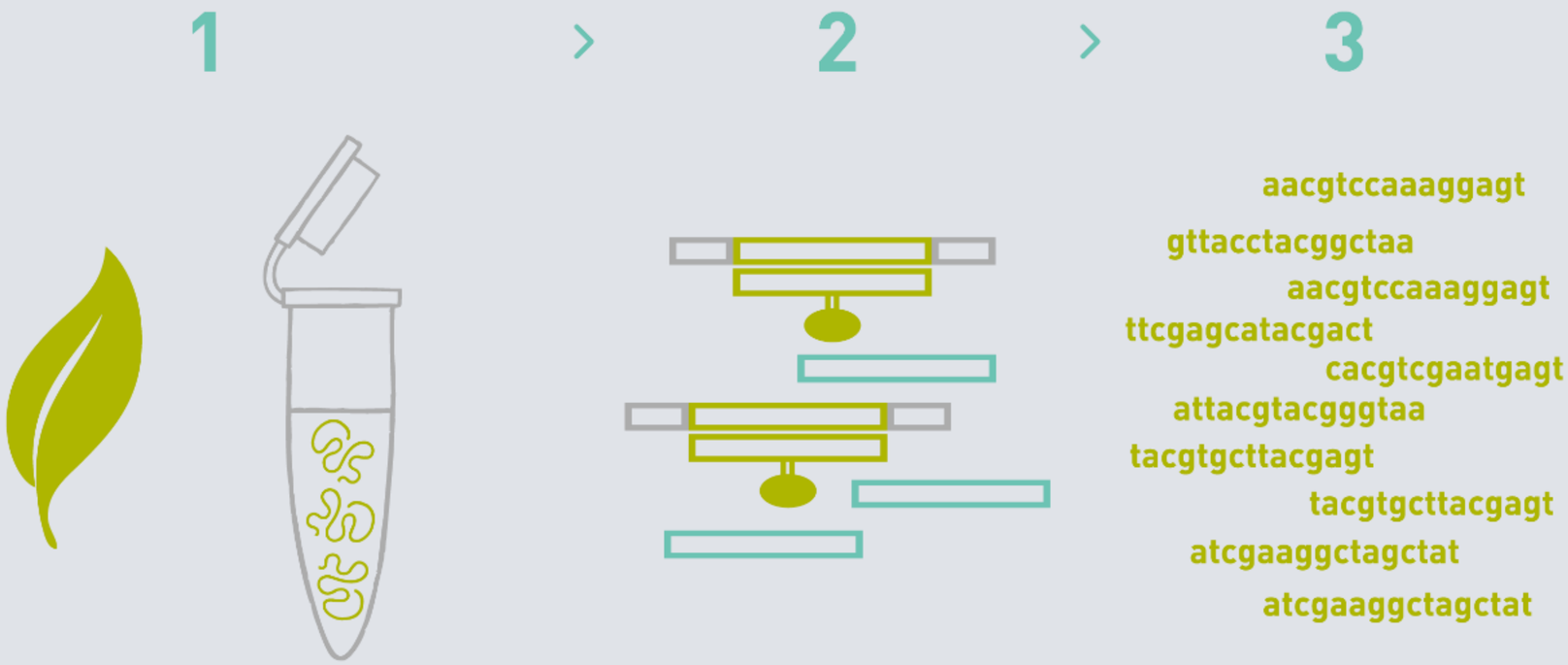
Institute of Informatics Systems SB RAS

Novosibirsk State University

Novel Software Systems LLC

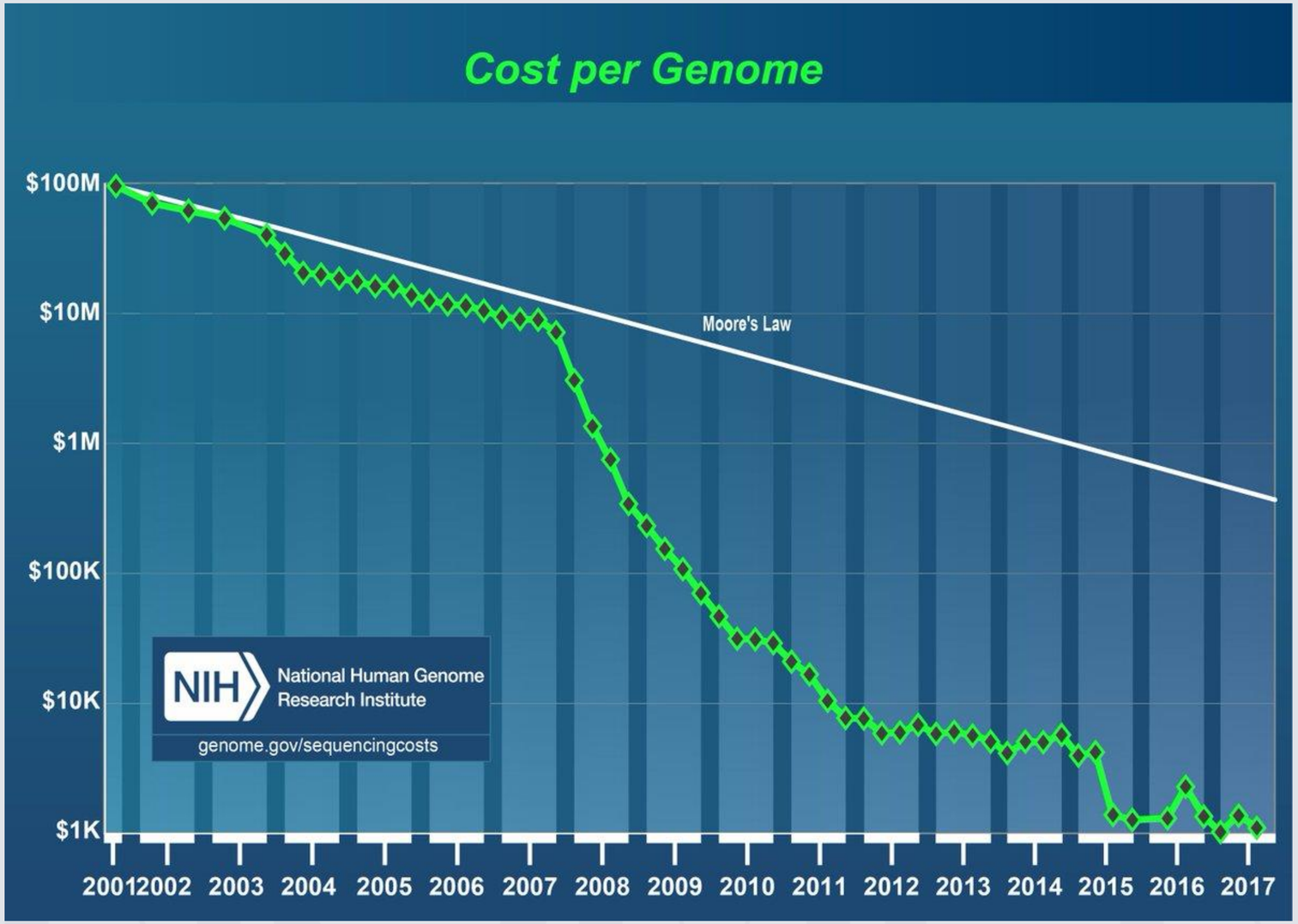
Next generation sequencing (NGS)

Is a technology of DNA and RNA molecules reading



Today’s price of genome sequencing is 700$

Dynamics of human genome sequencing cost over the 2001-2017 period.



Why to sequence genome?

Pharmacogenomics:
drug efficiency

Oncology:
screening and diagnostics

Inherited diseases:
diagnostics

Epidemiology:
infectious agents

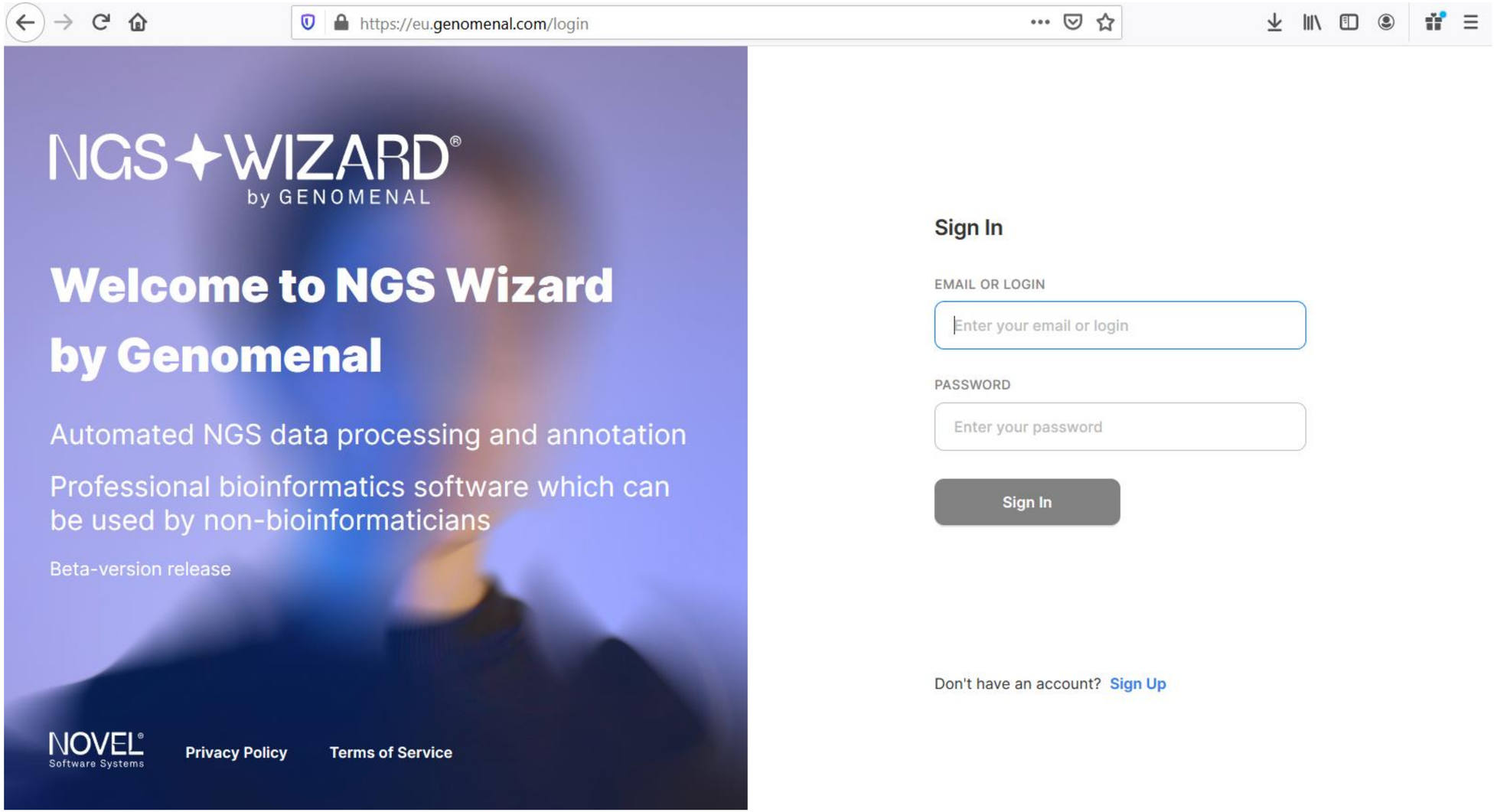
Metagenomics:
ecological research, personalized medicine

Fundamental research:
evolutionary science, genetics, archeology

Agriculture:
quality control, highly-efficient production

Lifestyle:
to live a better life

GENOMENAL NGS WIZARD is accessible at: eu.genomenal.com

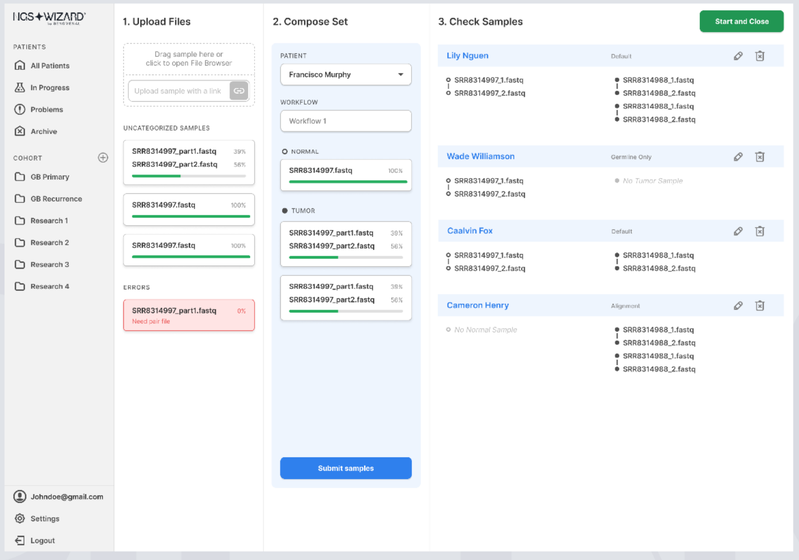


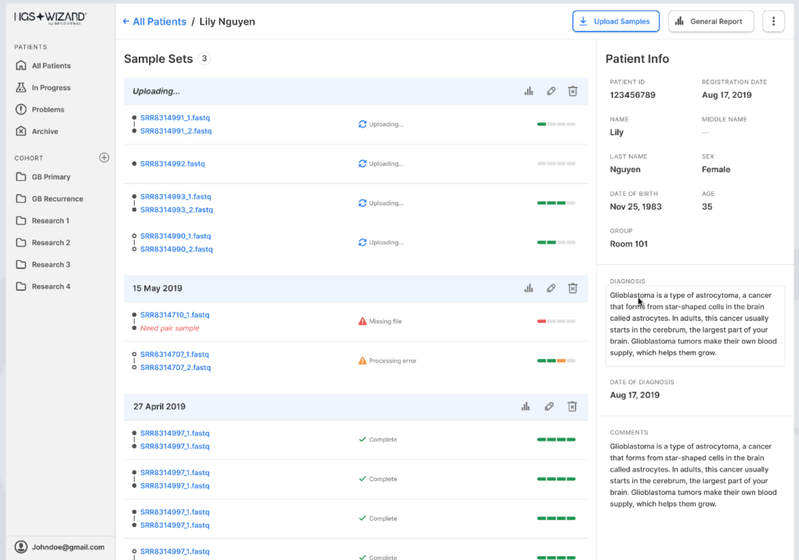
What is GENOMENAL NGS WIZARD

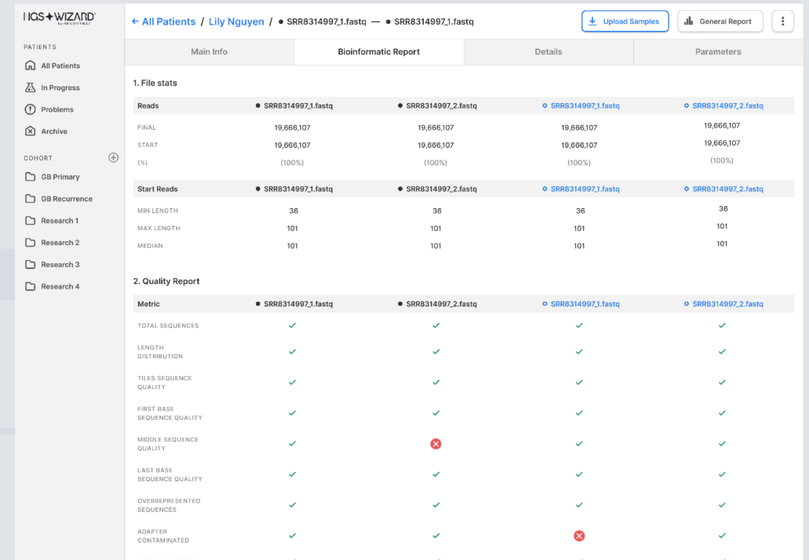
NGS WIZARD is a web application for DNAseq data processing (RNAseq later)

Clinical Whole Genome Sequencing	**Lowcoverage Whole Genome Sequencing**
Whole Exome Sequencing	**Target Gene Sequencing Panels**

Upload → Analyse → See result







GENOMENAL NGS Wizard solves bottle neck problems

Novosibirsk State University *THE REAL SCIENCE

NOVEL® Software Systems

OBJECTIVE PROBLEMS

- Instability. Usage of different models of sequencer, different reagents and different software parameters give different results on the same samples
- Lack of knowledge. Clinical exome provide “too many” mutations which require filtration, prioritization and interpretation
- Low coverage. Whole genome sequencing for clinical purposes is still expensive thus low coverage whole genome sequencing is usually performed

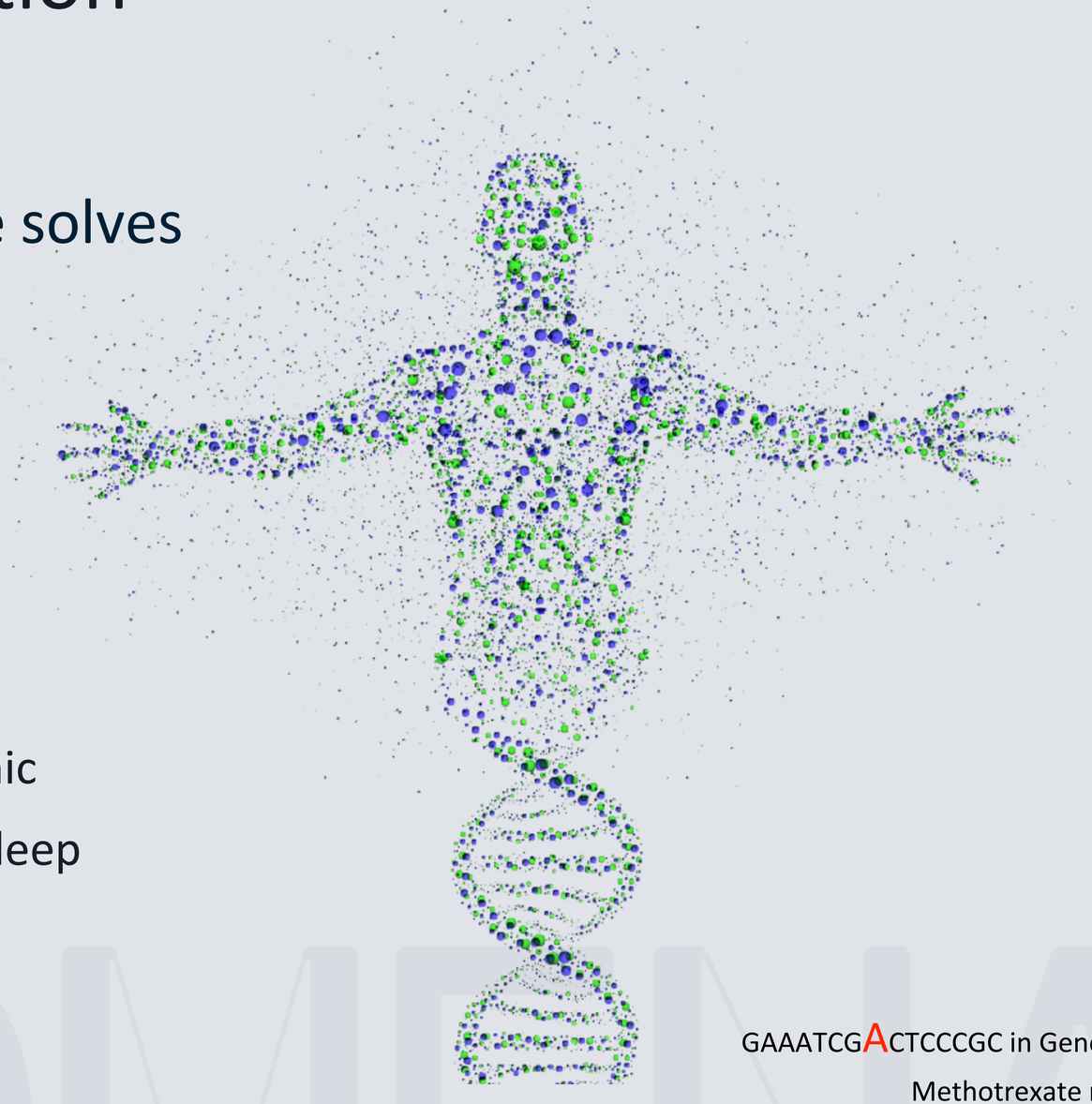
PRACTICAL PROBLEMS

- Lack of qualified bioinformaticians in the team
- Lack of access to knowledge databases

Do you have Next Generation Sequencing data?

GENOMENAL NGS Wizard software solves two problems:

- Lack of qualified bioinformaticians (Medical doctors, laborants, students can use NGS Wizard all alone)
- Deep analysis of genomic and transcriptomic data (Based on knowledge databases and deep learning models)



GAAATCGACTCCCGC in Gene MTHFR

Methotrexate response

https://www.ncbi.nlm.nih.gov/clinvar/variation/3520/

NGS Wizard compensates for deficit of bioinformaticians:

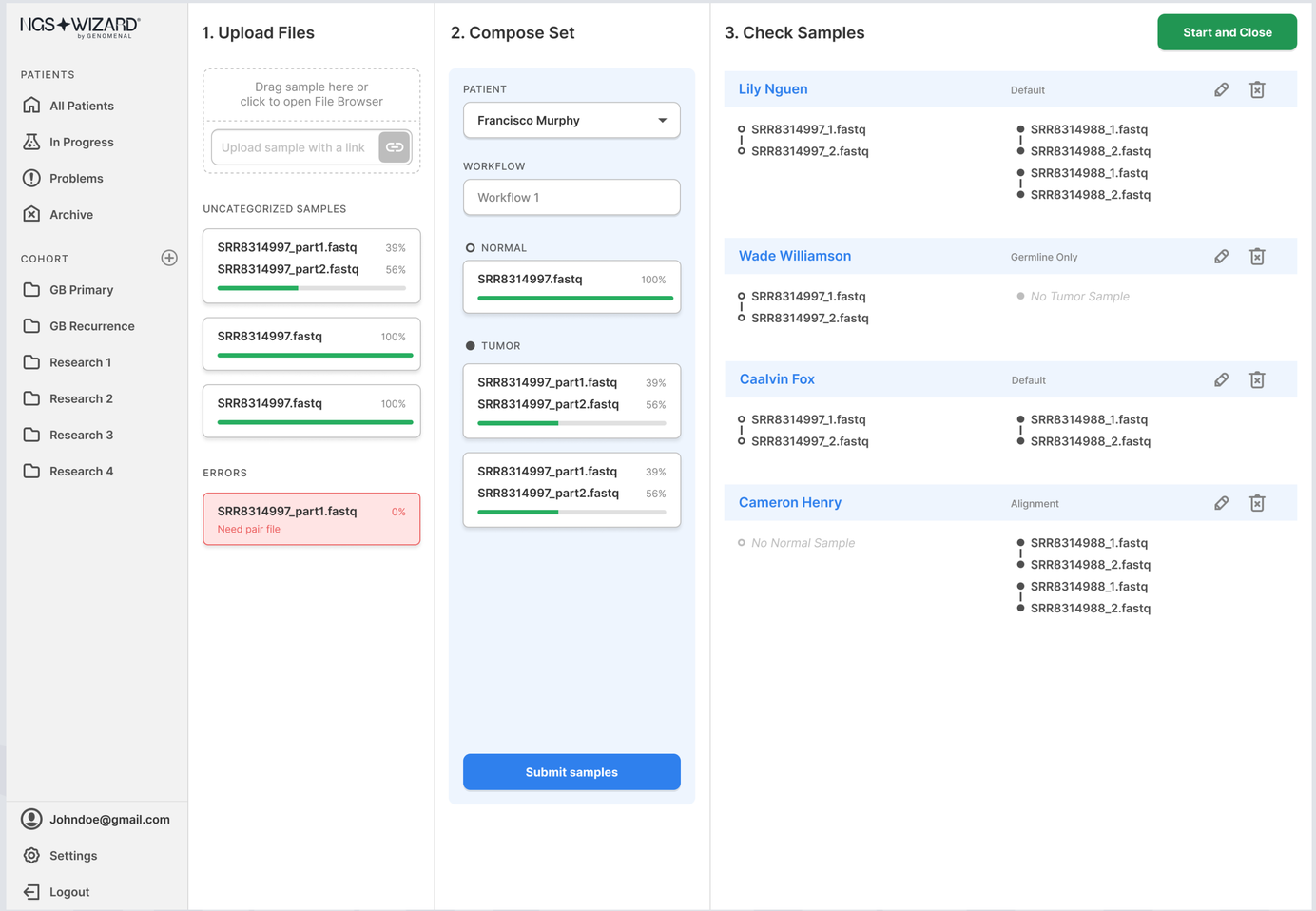
- Allows non-bioinformaticians to get results as if they were middle-qualified bioinformaticians;
- Can be used as instrument for education

GENOMENAL NGS Wizard makes data processing easier:

- Reduces processing time from 1 week to 2-3 days;
- Explains each step of analysis.

GENOMENAL NGS Wizard: just upload data

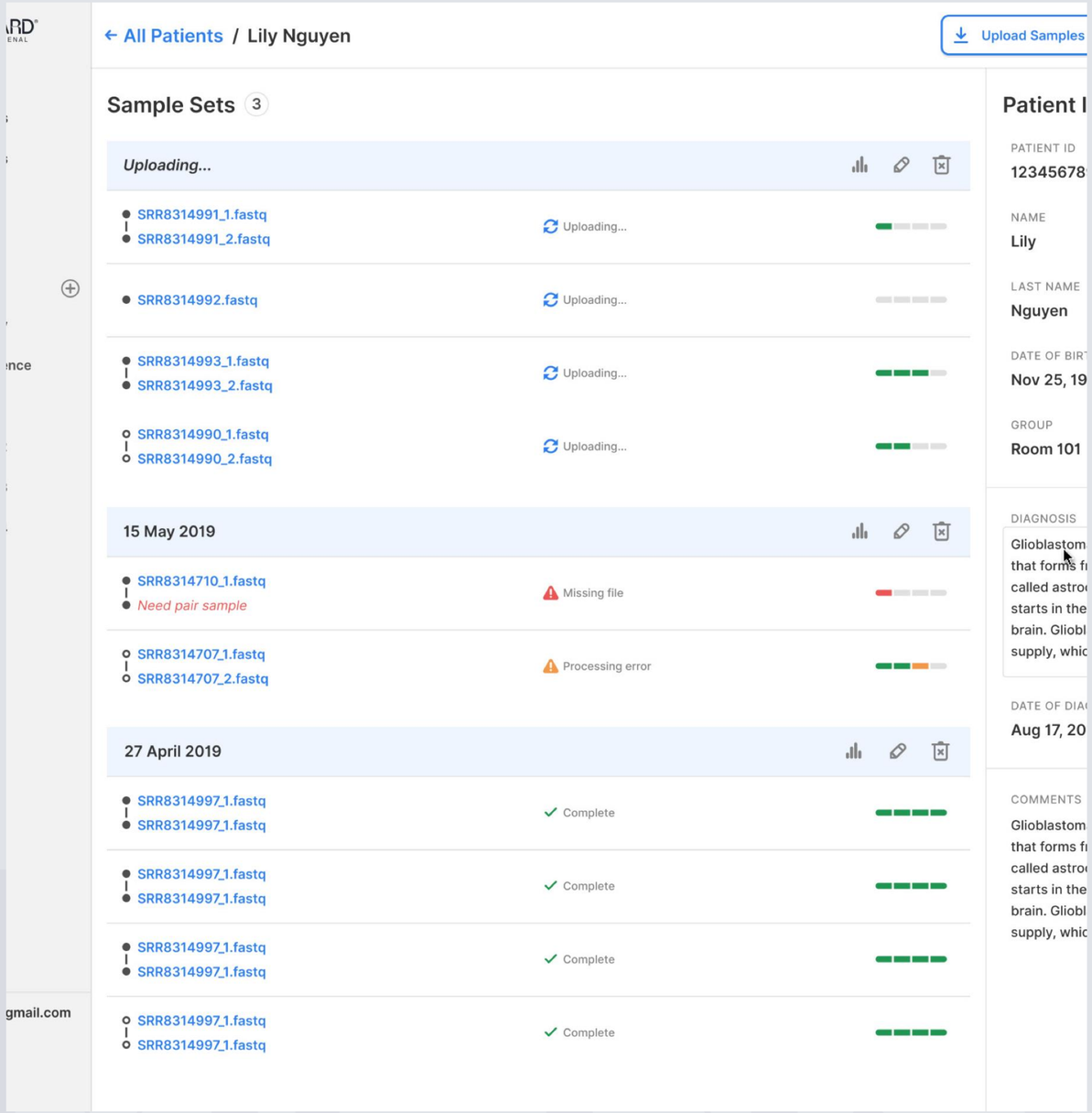
- Upload straight from the biggest online database NCBI SRA, local upload, upload via FTP
- Processing on fly
- Automatic data type recognition



Interface | Upload samples

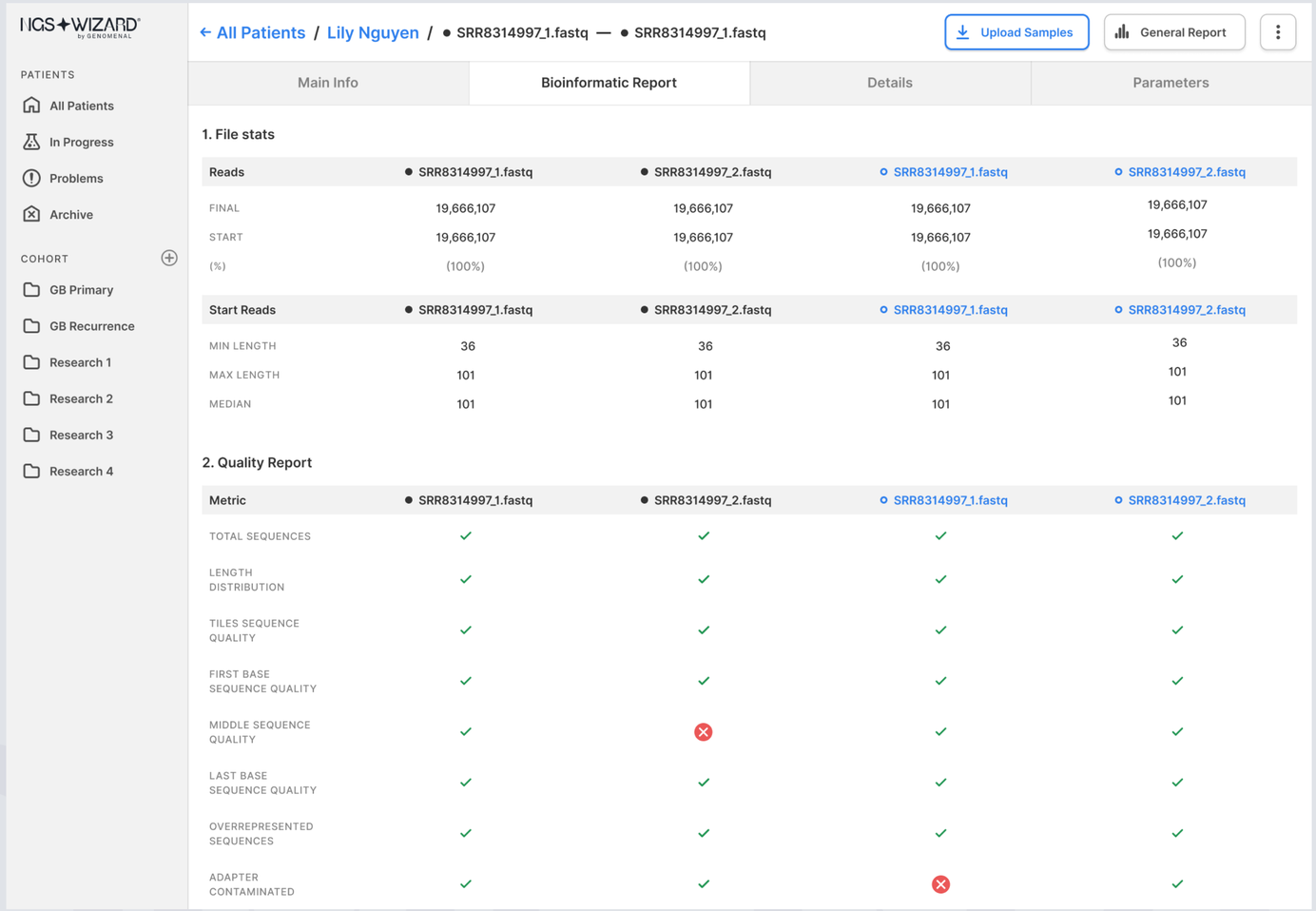
GENOMENAL NGS Wizard will perform:

- NGS data type recognition;
- Quality assessment and filtering;
- Mapping on reference genome;
- Somatic and germline mutations search;
- Evaluation of quality of mutations and their annotation;
- Clinical and technical reports



Data: bioinformatics at your fingertips

- Detailed bioinformatics reports will tell you about data quality and their informativeness
- You always can check how precisely your procedures worked out



Interface | Patient | Sample set | Bioinformatic Report

Clinical Report

- Focuses on specific diseases;
- Provides interpretation for somatic and germline variants;
- Suggests targeted drugs and clinical trials
- Helps to make personalized–treatment decisions

Clinical report

NGS WIZARD by GENOMENAL

John Doe

Patient ID	Date of birth	Sex	Registration date	Diagnosis
123456789	Dec 01, 1980	Female	Aug 10, 2019	Glioblastoma

Genomic variants

Somatic

Gene	Variant	Pathogenicity
CDKN2A	NC_000009.12:g.21991924C>T	Pathogenic
KRAS	NC_000012.12:g.25245350C>T	Pathogenic
BRCA2	NC_000017.11:g.43092919G>A	Pathogenic

Germline

Gene	Variant	Pathogenicity
CDKN2A	NC_000009.12:g.21991924C>T	Pathogenic
KRAS	NC_000012.12:g.25245350C>T	Pathogenic
BRCA2	NC_000017.11:g.43092919G>A	Pathogenic

GENOMENAL platform structure

Clinics

Academical institutes

Universities

Biotech companies

NGS data: Illumina, ThermoFisher, Oxford Nanopore, BGI

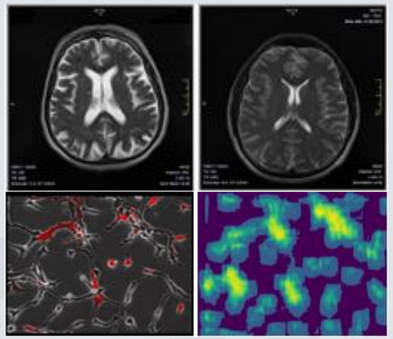
Patient's medical records data

MRI and microscope images data

NGS WIZARD® by GENOMENAL

Automated processing of raw data based on AI

Image analysis module



Module of Glioblastoma mutations analysis (tumor DNA and cfDNA)

Module for metagenome

Module for cardio vascular diseases

Phenotype prediction module (GWAS based)

Clinical recommendations

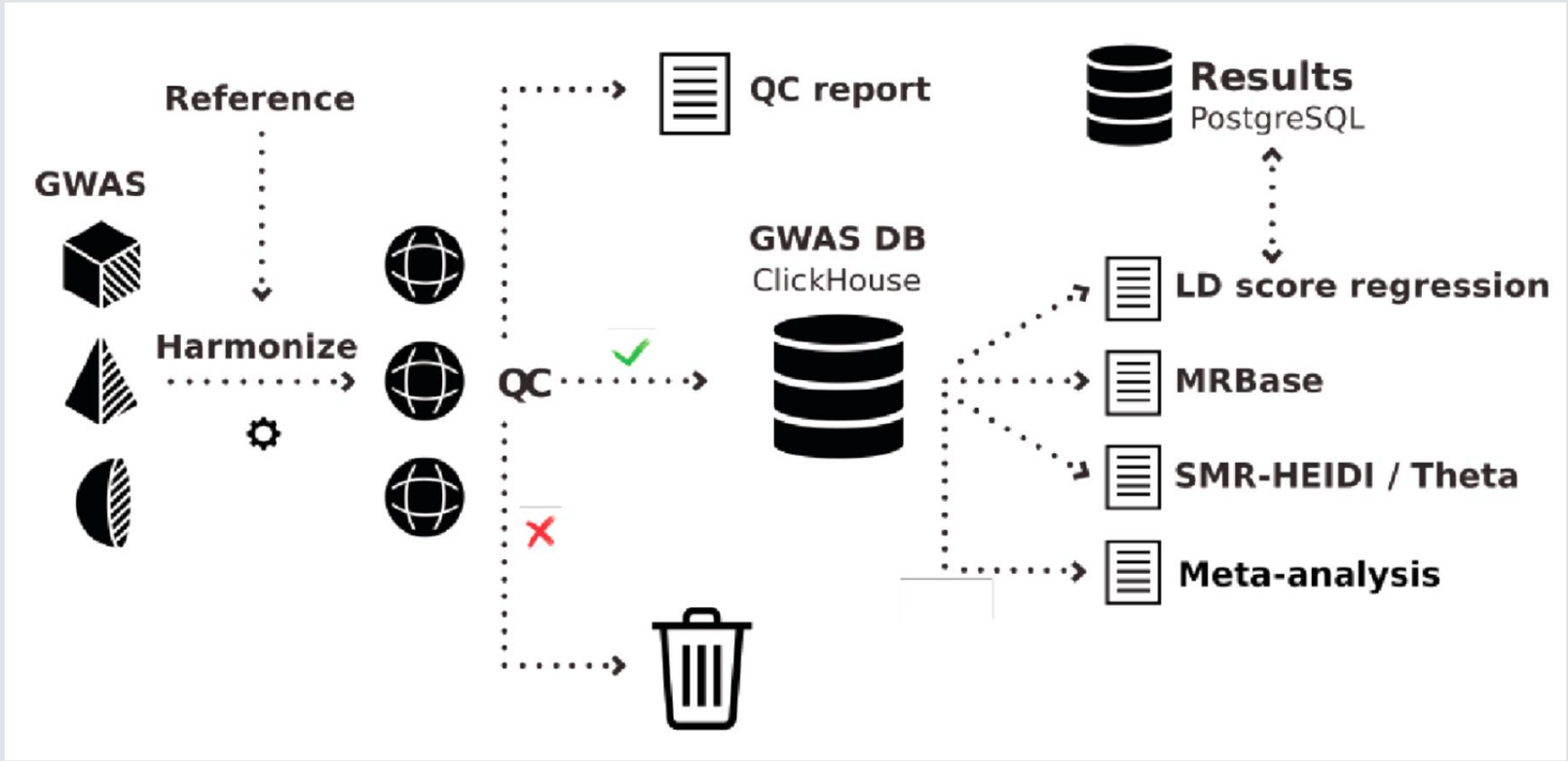
Health care optimization

Scientific publications

Drug targets

Phenotype prediction module (GWAS based)

52 blns of genotype-phenotype associations at the moment



Svishcheva, G. et al. Rapid variance components–based method for whole-genome association analysis. Nat Genet 44, 1166–1170 (2012) doi:10.1038/ng.2410
Momozawa Y. et al. IBD risk loci are enriched in multigenic regulatory modules encompassing putative causative genes. Nat Commun. 2018;9: 2427. doi:10.1038/s41467-018-04365-8
Zenin A. et al. Identification of 12 genetic loci associated with human healthspan. Commun Biol. 2019;2: 41. doi:10.1038/s42003-019-0290-0

Physiological traits analysis on human genome with Phenotype prediction module

Novosibirsk State University
*THE REAL SCIENCE

NOVEL®
Software Systems

Different kinds of input data

- Clinical Whole Genome Sequencing
- Lowcoverage Whole Genome Sequencing
- Whole Exome Sequencing
- Target Gene Sequencing Panels

Traits prediction based on GWAS-MAP

Ischemia: +15%

Height: -10%

Breast cancer: -5%

Eyes: blue

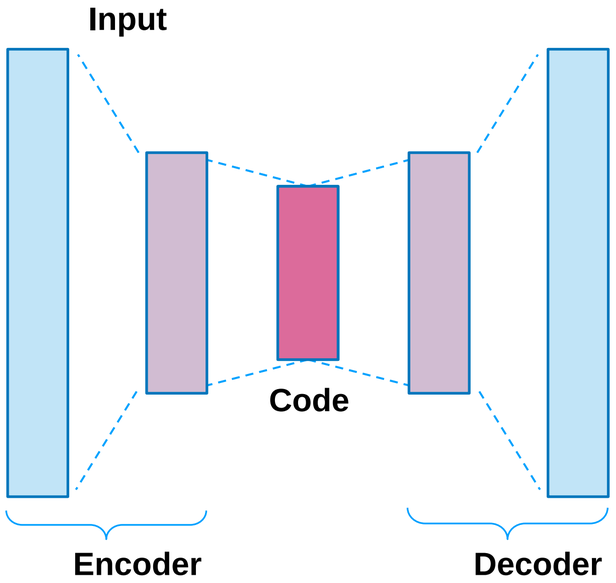
Pain sensitivity (model 1): -20%

Pain sensitivity (model 2): +10%

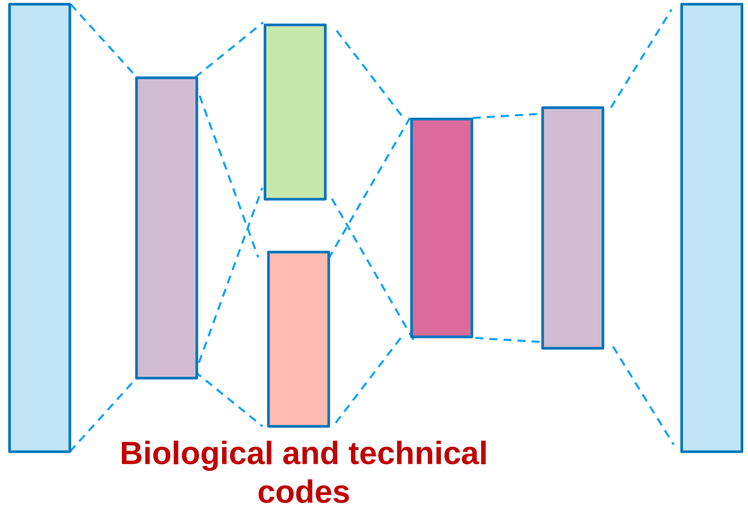
RNAseq processing harmonization problem and possible solution

Problem	Small training set and high dimension of DGE (Digital Gene Expression) data in the biomarker search task
Illustration of the problem	In a typical experiment, a comparison of 100 patients against 100 "healthy" patients in order to find differences in the expression of 200 thousand genes leads to irreproducible results
Solution	Autoencoder neural network, trained on 700 thousand samples, learned the features of 10 models of sequencers, 1000 tissues, cell lines and health conditions, 20 sample preparation protocols and other
Result	Reproducibility of biomarkers predicting even on a small training set

GENOMENAL solution for DGE harmonization



Typical Autoencoder model



Our approach

Applications

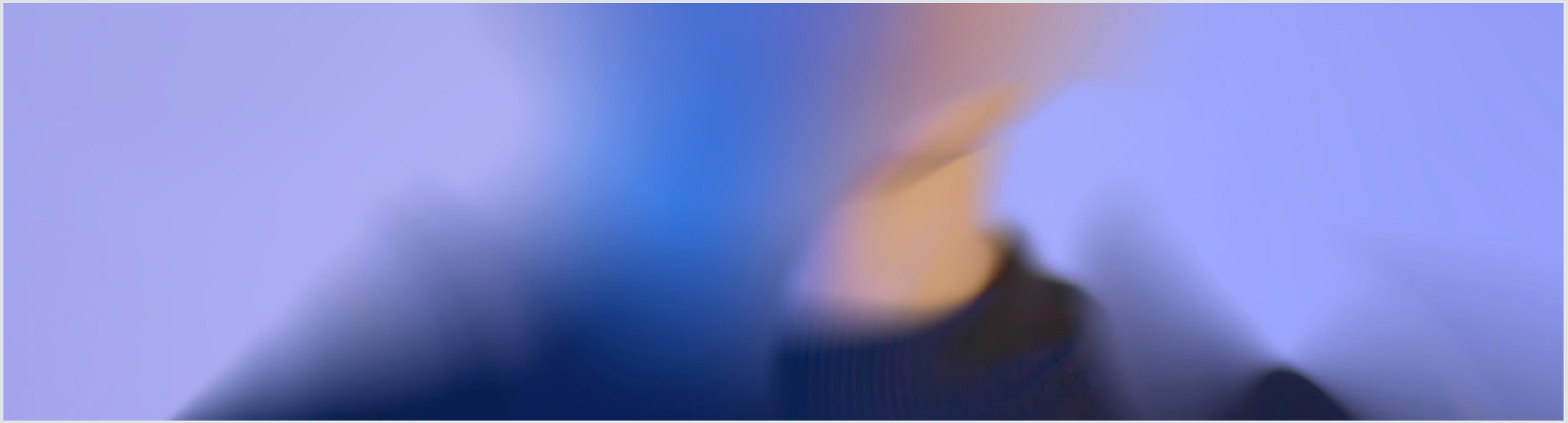
- Dimensionality reduction
- Missing data imputation
- Data harmonization
- Data augmentation
- Virtual screening
- Biomarker research

The main problem is to disentangle biological and technical sources of variation

Requirements to solve this task:

- Uniformly processed datasets *(partially solved by DEE2, ARCHS4, Recount2 etc.)*
- Very detailed metadata, describing different aspects of technical procedures (sample collection, storage, nucleic acid extraction, library preparation and sequencing) and biological characteristics (gender, age, health status, sample type – organ/tissue etc.)
- The metadata should be standardized to make it readily usable for machine learning experiments

Style transfer with variational autoencoders is a promising approach to RNA-Seq data harmonization and analysis (https://www.biorxiv.org/content/10.1101/791962v1)



GENOMENAL™

THANK YOU FOR ATTENTION!

Dmitry Shtokalo

CEO, PhD
dmitry@novel-soft.com
+7 913 392 43 75

Institute of Informatics Systems SB RAS

Novosibirsk State University

Novel Software Systems LLC

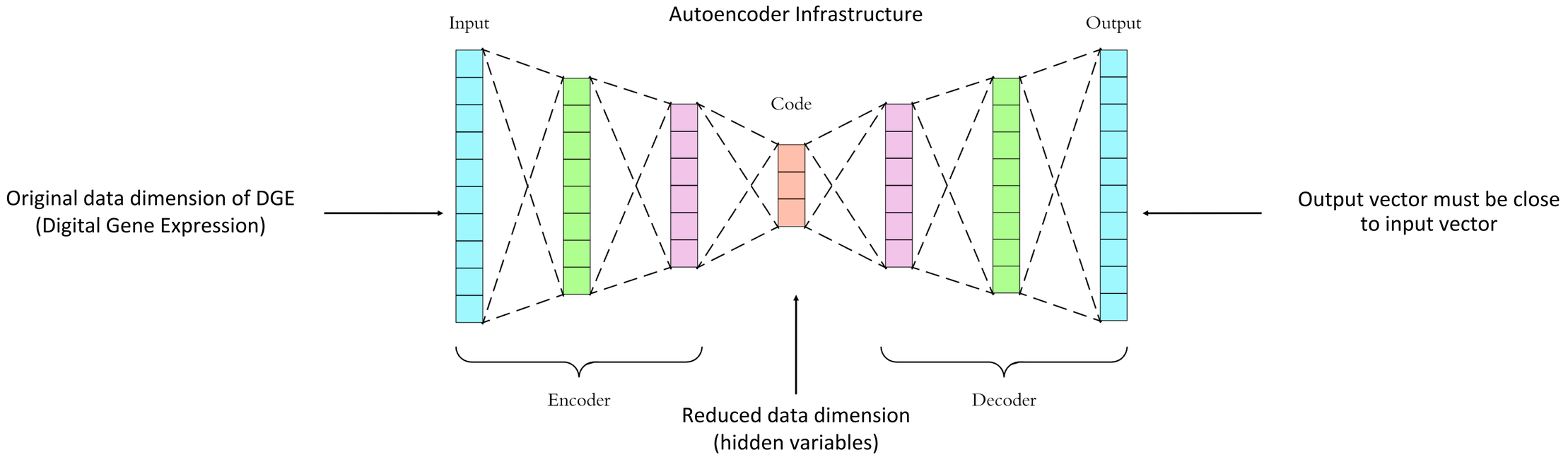
- Typical gene expression experiment contains a very limited number of samples
- Gene expression data is wide
- Data from different experiments suffer from severe batch effects and couldn’t be naively combined

Currently this problems are recognized and several consortia work to collect and harmonize the publicly available gene expression data

The classical approaches can’t harmonize heterogenous data.

Currently there is a big interest to deep learning techniques and, in particular, to autoencoders

RNA seq data analysis with GENOMENAL harmonization

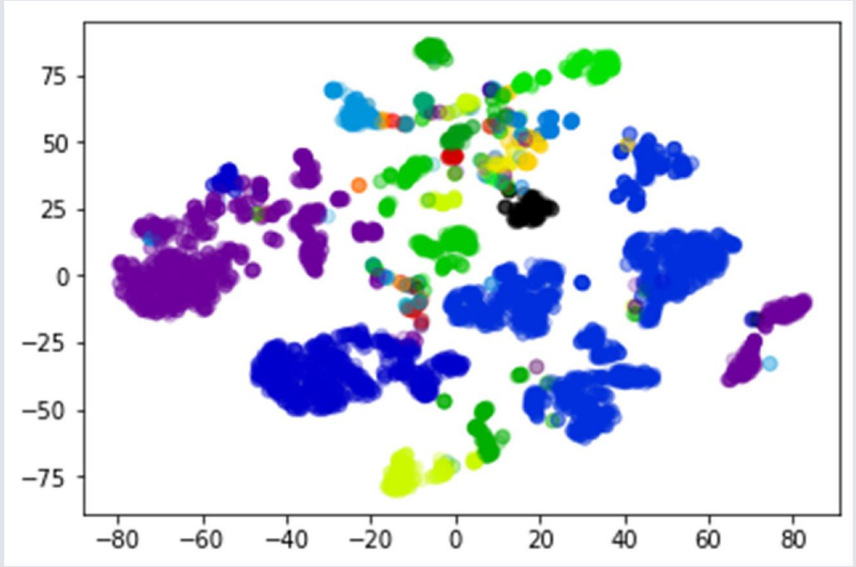


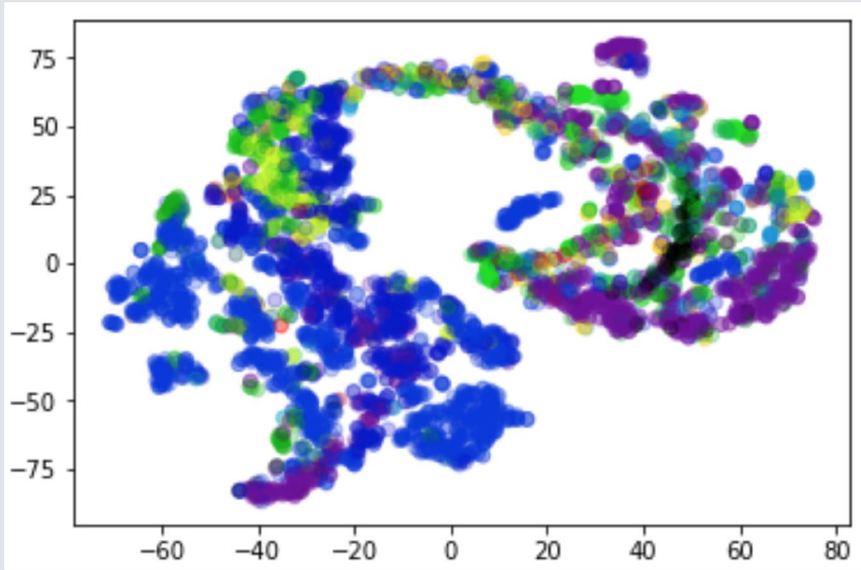
Autoencoder is used for

- Extraction of a small number of highly informative feature
- Certification of samples according to technical characteristics of quality and determination of biological status (tissue type, disease status)
- Integration of small data set with big one

Biomarkers discovery in RNAseq data

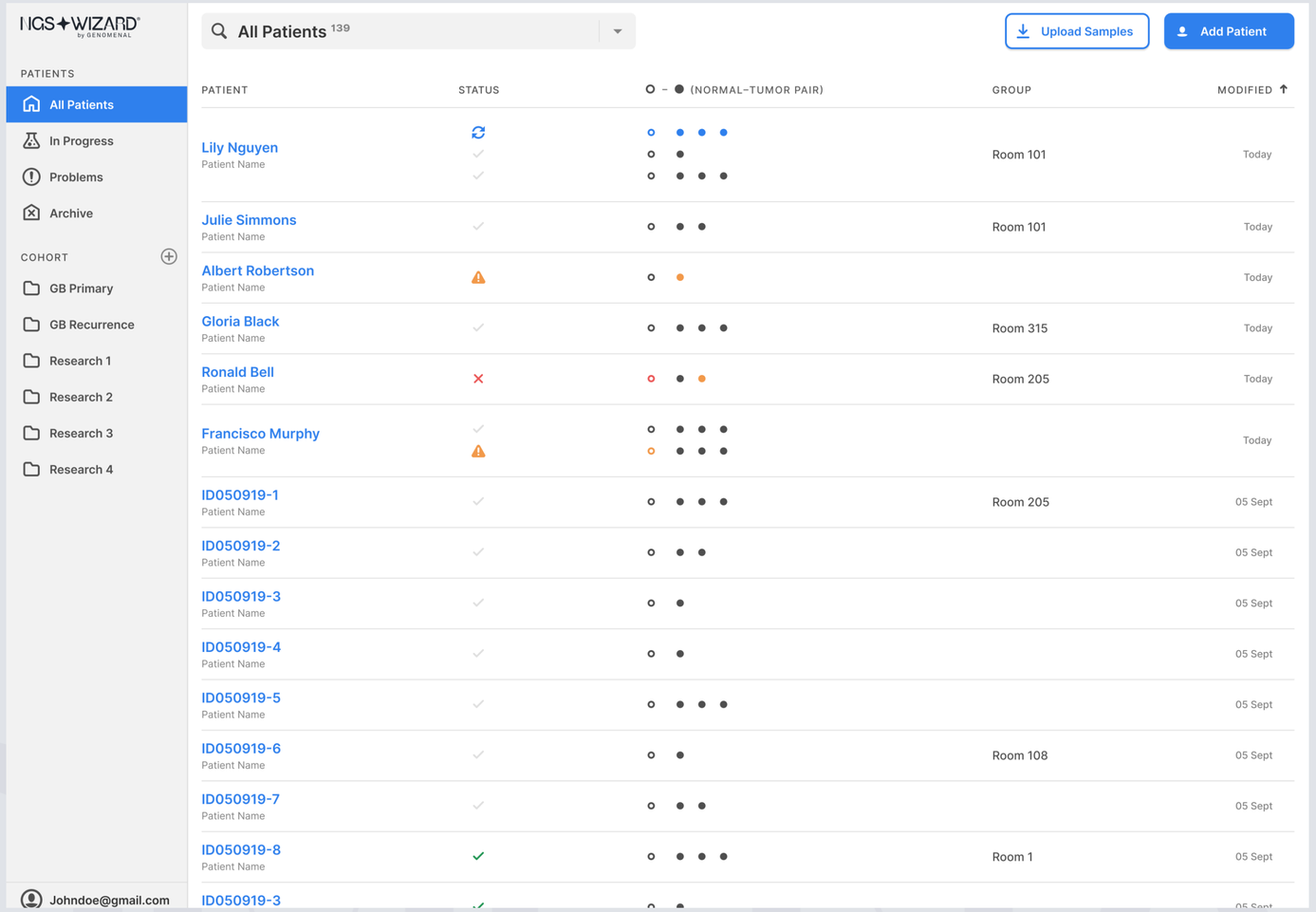
- Data passportization to assess quality
- Molecular pathways activity analysis
- Deep learning models trained on
  200 000 high quality samples





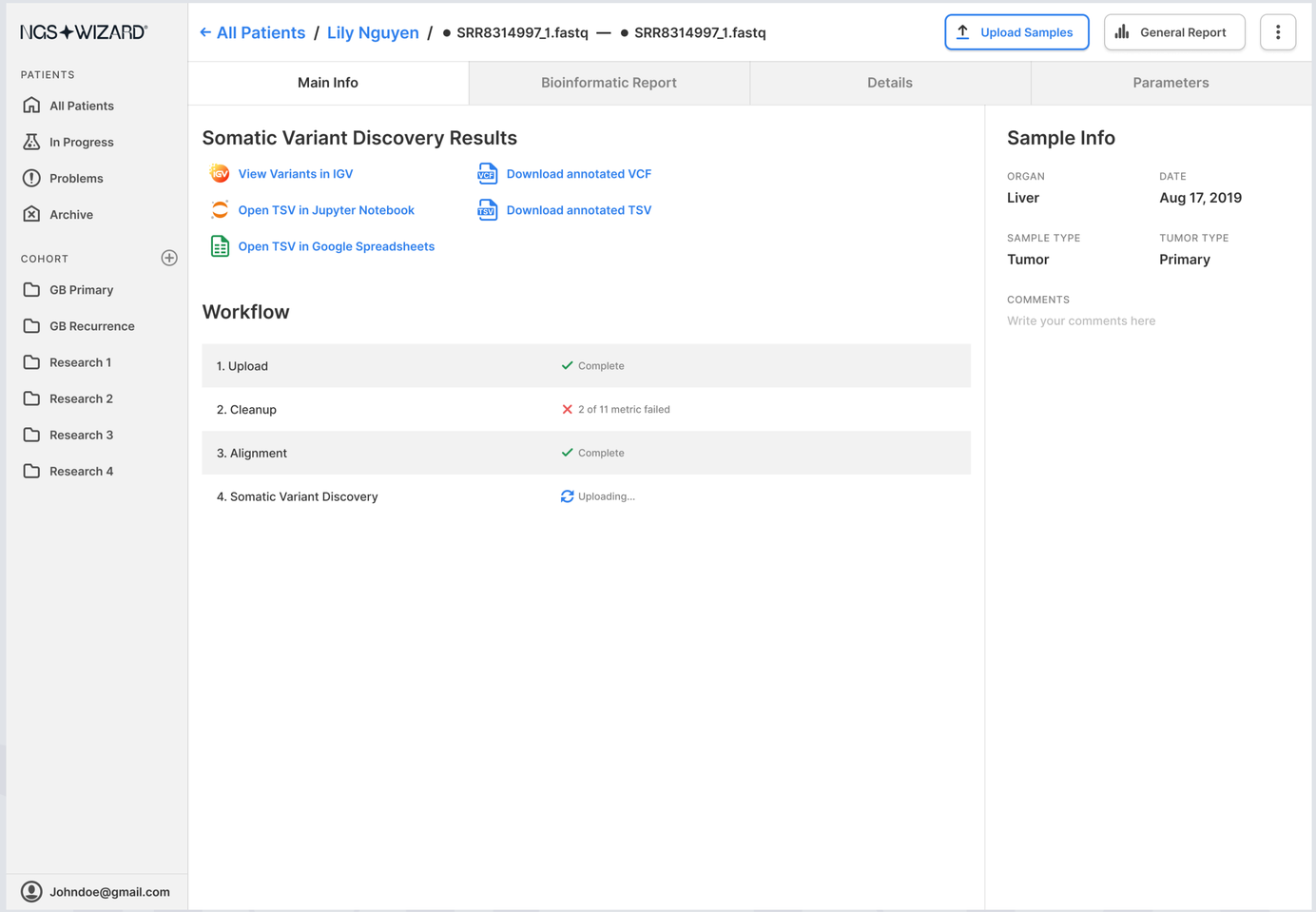
Patients and samples: full control of experiment

- Option to put patients in cohorts
- Creation of patient’s medical record
- A full picture of experiment is always before your eyes



Patients and samples: full control of experiment

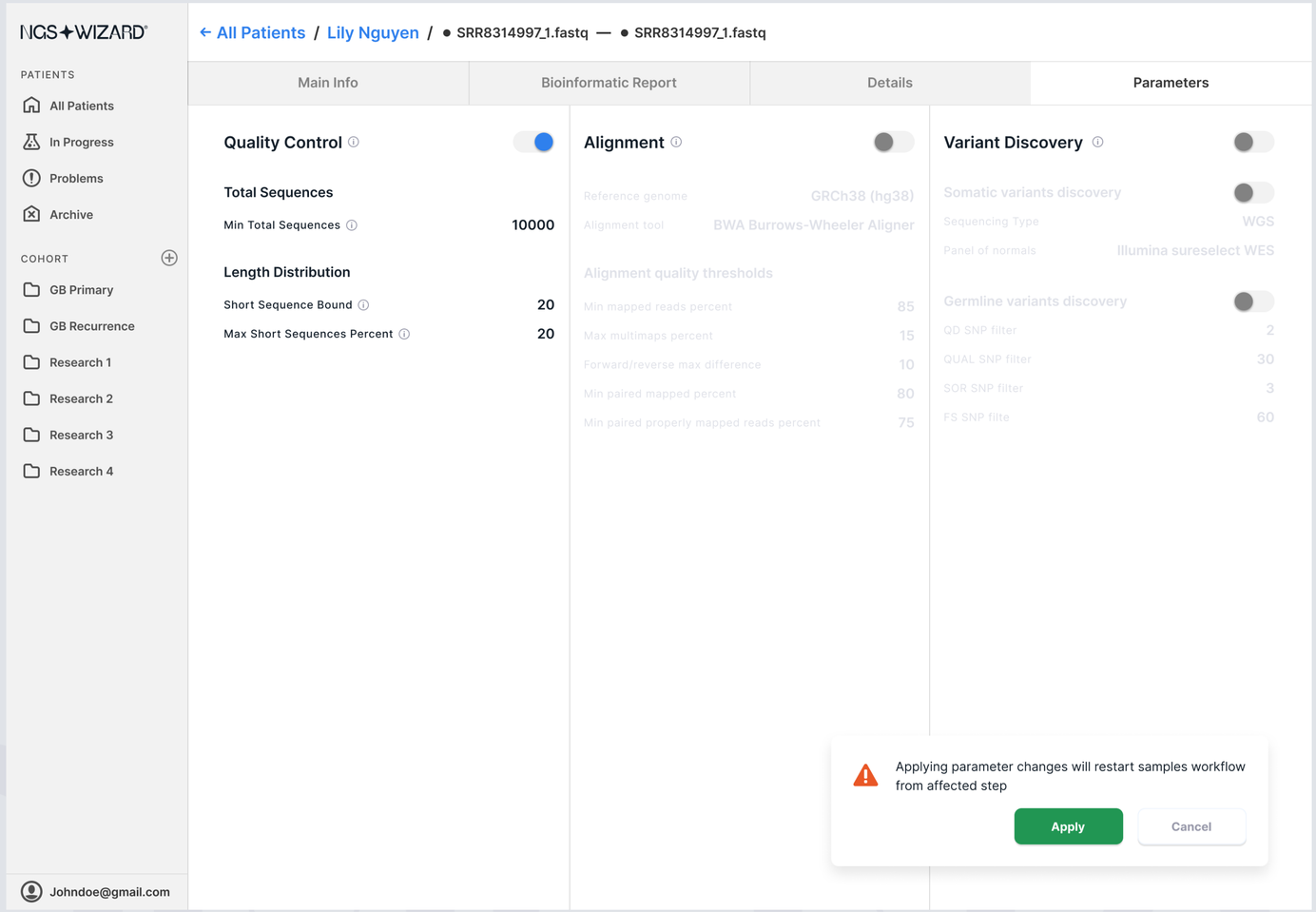
- Processing status: monitor progress
- Sample data is always in order
- Integration with IGV, Google Sheets, Jupyter Notebook allows to use the results at all levels



Interface | Patient | Sample set | More info

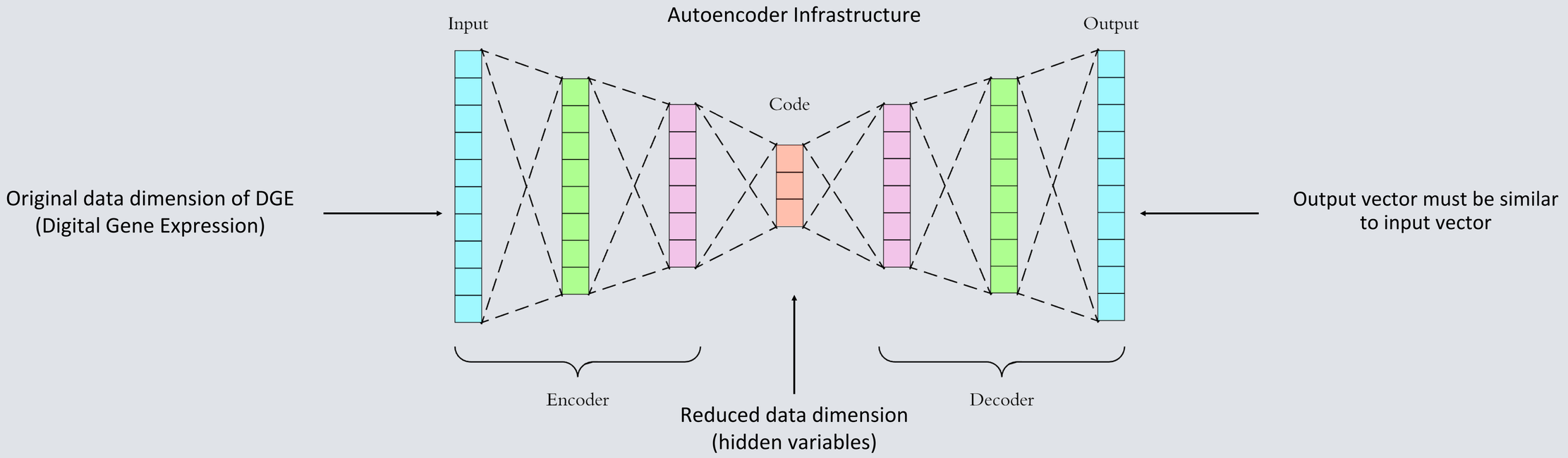
Do you know more? Control more

- We’ve worked hard to optimize the configuration, but you always can try to make it better
- Change the parameters of data preparation, mutation calling, compare the results and learn



Interface | Patient | Sample set | Parameters

RNA seq data analysis with GENOMENAL harmonization



Autoencoder is used for

- Extraction of a small number of highly informative feature
- Certification of samples according to technical characteristics of quality and determination of biological status (tissue type, disease status)
- Integration of small data set with big one