

GENOMENAL

N* Новосибирский
государственный
университет
*НАСТОЯЩАЯ НАУКА

NOVEL[®]
Software Systems



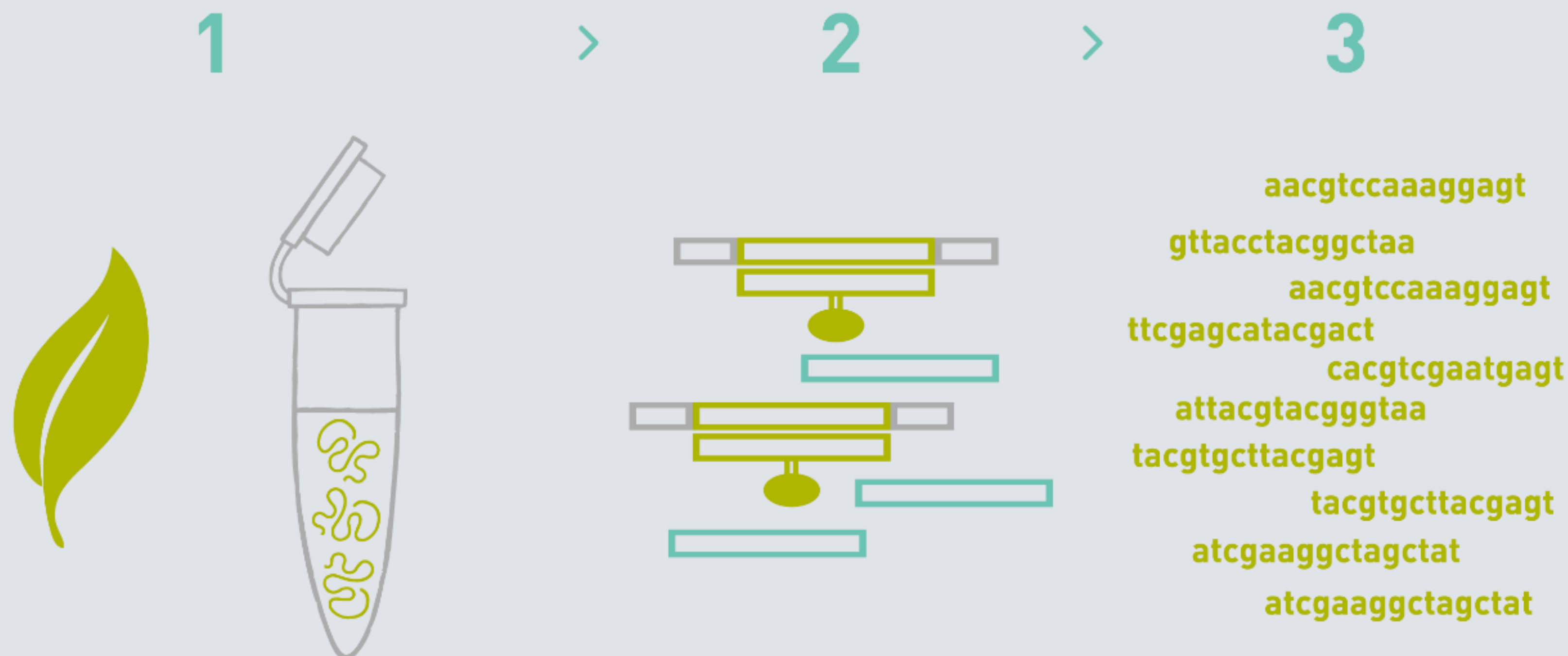
Dmitry Shtokalo, PhD

GENOMENAL platform: ecosystem
solution for digital genomics

Institute of Informatics Systems SB RAS
Novosibirsk State University
Novel Software Systems LLC

Next generation sequencing (NGS)

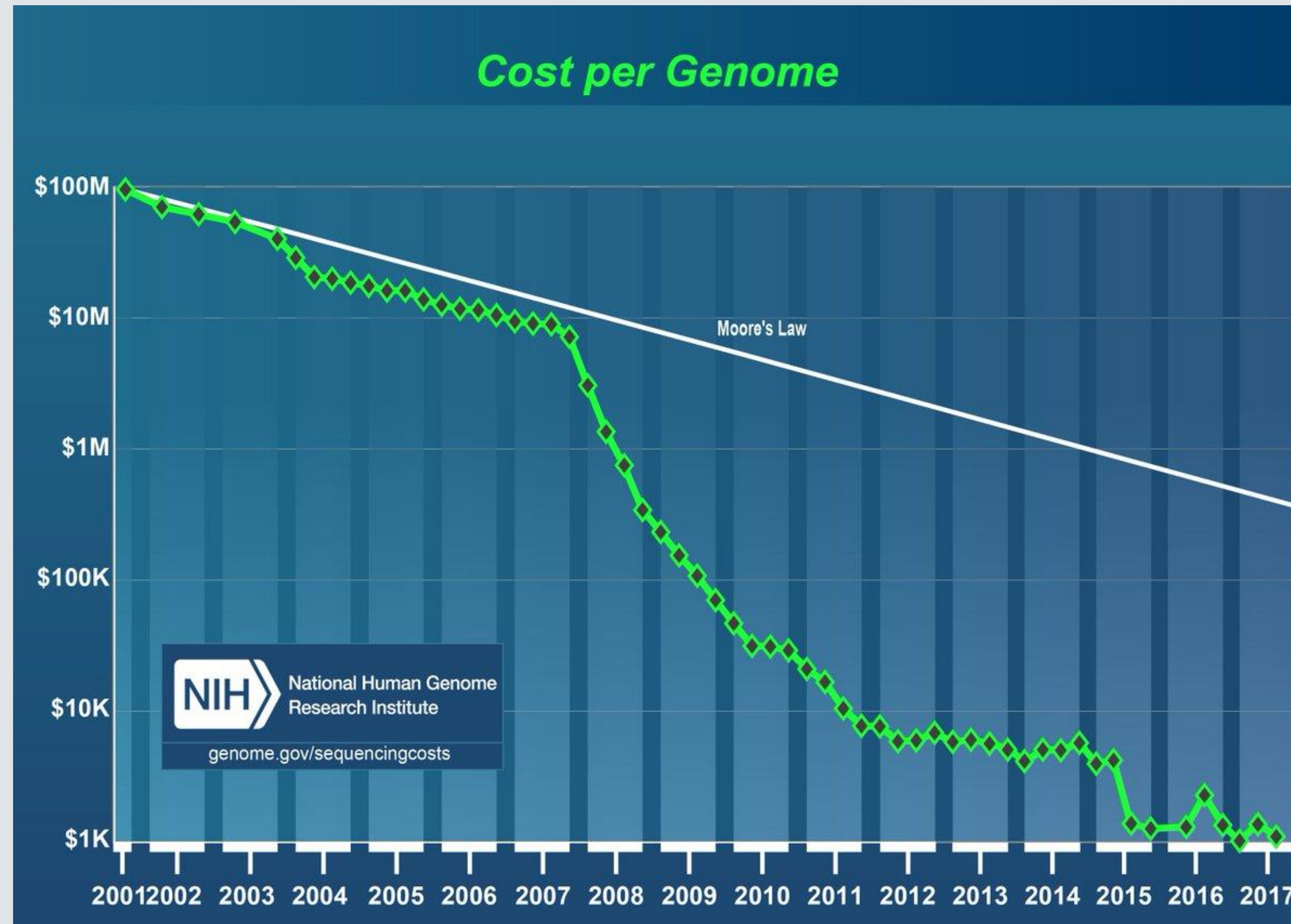
Is a technology of DNA and RNA molecules reading











GENOMENETICS

Today's price of genome sequencing is 700\$

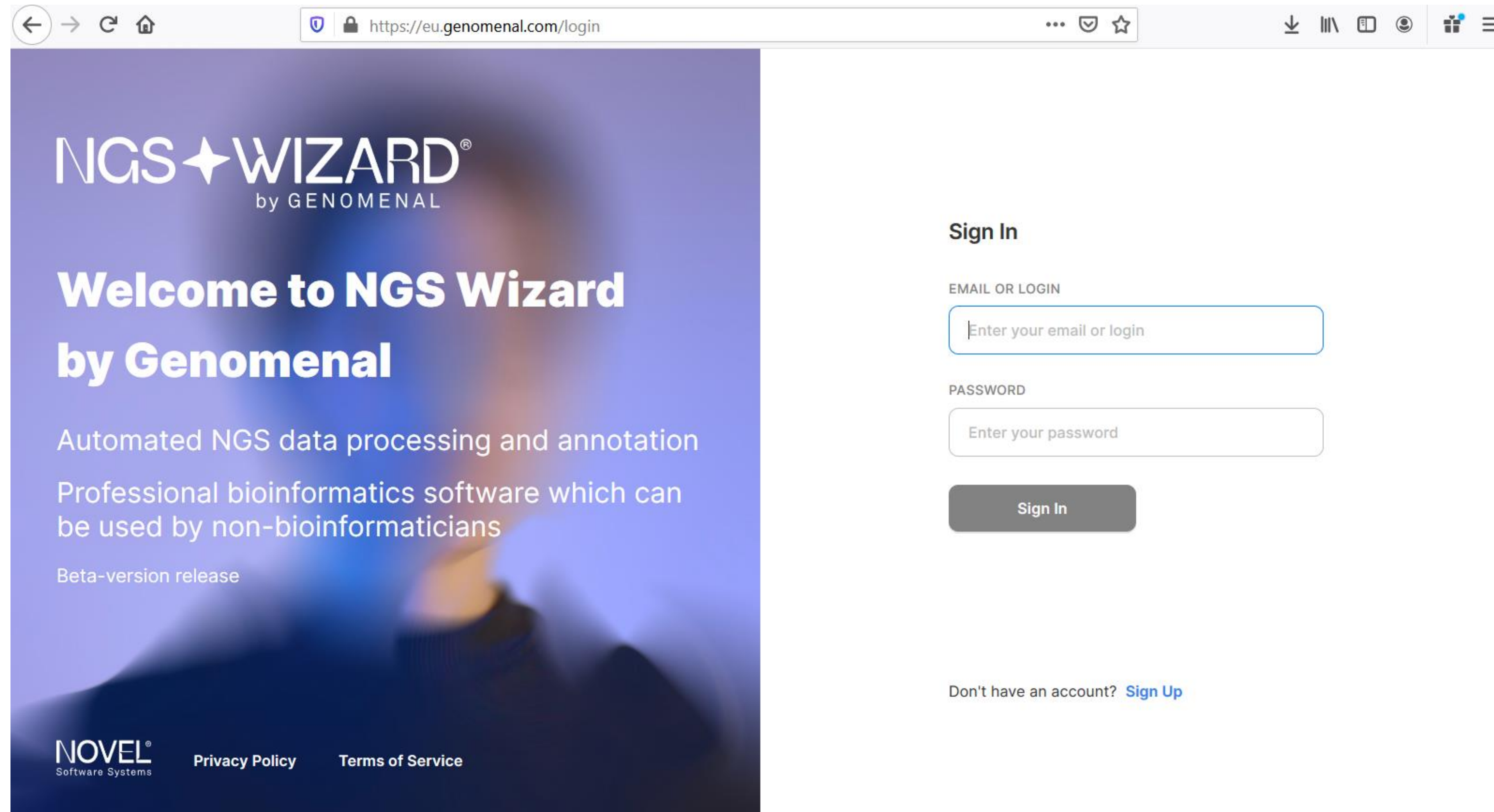
Dynamics of human genome sequencing cost over the 2001-2017 period.



Why to sequence genome?

- | | | | |
|---|--|---|---|
|  | Pharmacogenomics:
drug efficiency |  | Metagenomics:
ecological research, personalized medicine |
|  | Oncology:
screening and diagnostics |  | Fundamental research:
evolutionary science, genetics, archeology |
|  | Inherited diseases:
diagnostics |  | Agriculture:
quality control, highly-efficient production |
|  | Epidemiology:
infectious agents |  | Lifestyle:
to live a better life |

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



The screenshot shows a web browser window with the URL <https://eu.genomenal.com/login>. The page features a large purple banner on the left with the text "NGS WIZARD[®] by GENOMENAL", "Welcome to NGS Wizard by Genomenal", "Automated NGS data processing and annotation", "Professional bioinformatics software which can be used by non-bioinformaticians", and "Beta-version release". At the bottom of the banner are the "NOVEL[®] Software Systems" logo, "Privacy Policy", and "Terms of Service" links. On the right, the "Sign In" section contains two input fields: "EMAIL OR LOGIN" with the placeholder "Enter your email or login" and "PASSWORD" with the placeholder "Enter your password". A "Sign In" button is located below these fields. At the bottom of the sign-in section, it says "Don't have an account? [Sign Up](#)".

NGS WIZARD[®]
by GENOMENAL

Welcome to NGS Wizard by Genomenal

Automated NGS data processing and annotation
Professional bioinformatics software which can
be used by non-bioinformaticians

Beta-version release

NOVEL[®]
Software Systems

[Privacy Policy](#) [Terms of Service](#)

Sign In

EMAIL OR LOGIN

PASSWORD

Sign In

Don't have an account? [Sign Up](#)

What is GENOMENAL NGS WIZARD

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



Upload

Analyse

See result

PATIENTS

All Patients

In Progress

Problems

Archive

COHORT

OB Primary

OB Recurrence

Research 1

Research 2

Research 3

Research 4

Johndoe@gmail.com

Settings

Logout

1. Upload Files

Drag sample here or click to open File Browser

Upload sample with a link

UNCATEGORIZED SAMPLES

SRR8314997_part1.fastq

39%

SRR8314997_part12.fastq

56%

ERRORS

SRR8314997_part1.fastq

0%

Need pair file

2. Compose Set

PATIENT

Francisco Murphy

WORKFLOW

Workflow 1

NORMAL

SRR8314997.fastq

100%

TUMOR

SRR8314997_part1.fastq

39%

SRR8314997_part2.fastq

56%

SRR8314997_part1.fastq

38%

SRR8314997_part2.fastq

56%

Submit samples

3. Check Samples

Lily Nguen

Default

SRR8314997_1.fastq

SRR8314997_2.fastq

SRR8314988_1.fastq

SRR8314988_2.fastq

SRR8314988_1.fastq

SRR8314988_2.fastq

Wade Williamson

Germline Only

SRR8314997_1.fastq

SRR8314997_2.fastq

No Tumor Sample

Caalvin Fox

Default

SRR8314997_1.fastq

SRR8314997_2.fastq

SRR8314988_1.fastq

SRR8314988_2.fastq

SRR8314988_1.fastq

SRR8314988_2.fastq

Cameron Henry

Alignment

SRR8314997_1.fastq

SRR8314997_2.fastq

No Normal Sample

SRR8314988_1.fastq

SRR8314988_2.fastq

SRR8314988_1.fastq

SRR8314988_2.fastq

PATIENTS

All Patients

In Progress

Problems

Archive

COHORT

OB Primary

OB Recurrence

Research 1

Research 2

Research 3

Research 4

Johndoe@gmail.com

Settings

Logout

All Patients / Lily Nguen

Upload Samples

General Report

Sample Sets

3

Uploading...

SRR8314991_1.fastq

SRR8314991_2.fastq

Uploading...

SRR8314992.fastq

Uploading...

SRR8314993_1.fastq

SRR8314993_2.fastq

Uploading...

SRR8314990_1.fastq

SRR8314990_2.fastq

15 May 2019

SRR8314710_1.fastq

Need pair sample

Missing file

SRR8314702_1.fastq

SRR8314702_2.fastq

Processing error

27 April 2019

SRR8314997_1.fastq

SRR8314997_1.fastq

Complete

SRR8314997_1.fastq

SRR8314997_1.fastq

Complete

SRR8314997_1.fastq

SRR8314997_1.fastq

Complete

SRR8314997_1.fastq

SRR8314997_1.fastq

Complete

Patient Info

PATIENT ID

123456789

REGISTRATION DATE

Aug 17, 2019

NAME

Lily

MIDDLE NAME

LAST NAME

Nguyen

SEX

Female

DATE OF BIRTH

Nov 25, 1983

AGE

35

GROUP

Room 101

DIAGNOSIS

Glioblastoma is a type of astrocytoma, a cancer that forms from star-shaped cells in the brain called astrocytes. In adults, this cancer usually starts in the cerebrum, the largest part of your brain. Glioblastoma tumors make their own blood supply, which helps them grow.

DATE OF DIAGNOSIS

Aug 17, 2019

COMMENTS

Glioblastoma is a type of astrocytoma, a cancer that forms from star-shaped cells in the brain called astrocytes. In adults, this cancer usually starts in the cerebrum, the largest part of your brain. Glioblastoma tumors make their own blood supply, which helps them grow.

PATIENTS

All Patients

In Progress

Problems

Archive

COHORT

OB Primary

OB Recurrence

Research 1

Research 2

Research 3

Research 4

Johndoe@gmail.com

Settings

Logout

All Patients / Lily Nguen

SRR8314997_1.fastq

SRR8314997_2.fastq

Upload Samples

General Report

Main Info

Bioinformatic Report

Details

Parameters

1. File stats

Reads

SRR8314997_1.fastq

SRR8314997_2.fastq

SRR8314997_1.fastq

SRR8314997_2.fastq

FINAL

19,666,107

19,666,107

19,666,107

19,666,107

START

19,666,107

19,666,107

19,666,107

19,666,107

(%)

(100%)

(100%)

(100%)

(100%)

Start Reads

SRR8314997_1.fastq

SRR8314997_2.fastq

SRR8314997_1.fastq

SRR8314997_2.fastq

MIN LENGTH

36

36

36

36

MAX LENGTH

101

101

101

101

MEDIAN

101

101

101

101

2. Quality Report

Metric

SRR8314997_1.fastq

SRR8314997_2.fastq

SRR8314997_1.fastq

SRR8314997_2.fastq

TOTAL SEQUENCES

✓

✓

✓

✓

LENGTH DISTRIBUTION

✓

✓

✓

✓

TILES SEQUENCE QUALITY

✓

✓

✓

✓

FIRST BASE SEQUENCE QUALITY

✓

✓

✓

✓

MIDDLE SEQUENCE QUALITY

✓

✗

✓

✓

LAST BASE SEQUENCE QUALITY

✓

✓

✓

✓

OVERREPRESENTED SEQUENCES

✓

✓

✓

✓

ADAPTER CONTAMINATED

✓

✓

✗

✓

GENOMENAL NGS Wizard solves bottle neck problems

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

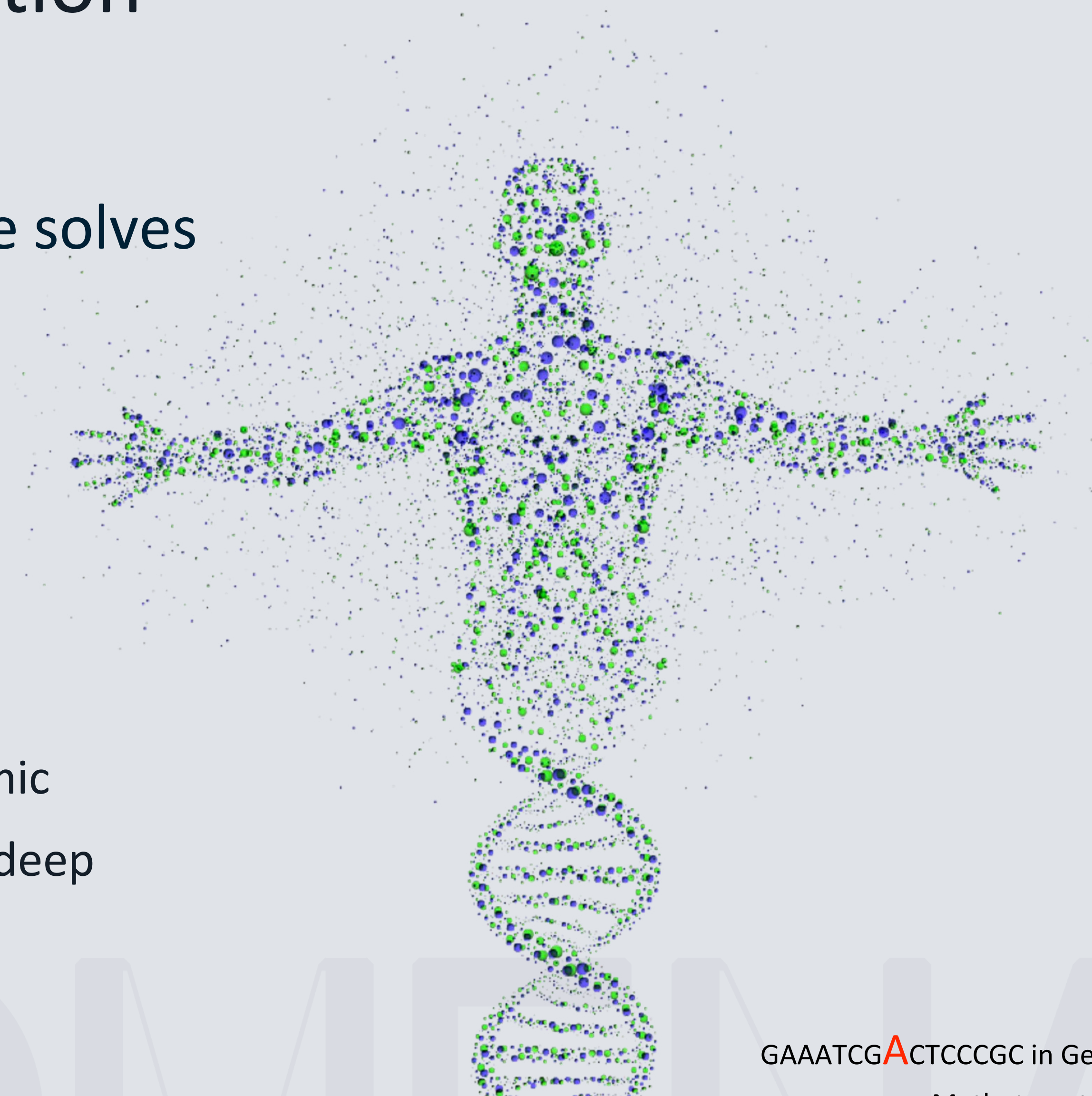
PRACTICAL PROBLEMS

- 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)



GAAATCGA^ACTCCCGC 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

NGS+WIZARD[®]
by GENOMENAL

PATIENTS

All Patients

In Progress

Problems

Archive

COHORT

GB Primary

GB Recurrence

Research 1

Research 2

Research 3

Research 4

Johndoe@gmail.com

Settings

Logout

1. Upload Files

Drag sample here or click to open File Browser

Upload sample with a link

UNCATEGORIZED SAMPLES

SRR8314997_part1.fastq 39%

SRR8314997_part2.fastq 56%

SRR8314997.fastq 100%

SRR8314997.fastq 100%

ERRORS

SRR8314997_part1.fastq 0%

Need pair file

2. Compose Set

PATIENT

Francisco Murphy

WORKFLOW

Workflow 1

NORMAL

SRR8314997.fastq 100%

TUMOR

SRR8314997_part1.fastq 39%

SRR8314997_part2.fastq 56%

SRR8314997_part1.fastq 39%

SRR8314997_part2.fastq 56%

Submit samples

3. Check Samples

Start and Close

Lily Nguen

Default

SRR8314997_1.fastq

SRR8314988_1.fastq

SRR8314997_2.fastq

SRR8314988_2.fastq

Wade Williamson

Germline Only

SRR8314997_1.fastq

SRR8314997_2.fastq

No Tumor Sample

Caalvin Fox

Default

SRR8314997_1.fastq

SRR8314997_2.fastq

SRR8314988_1.fastq

SRR8314988_2.fastq

Cameron Henry

Alignment

No Normal Sample

SRR8314988_1.fastq

SRR8314988_2.fastq

SRR8314988_1.fastq

SRR8314988_2.fastq

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

The screenshot displays the GENOMENAL NGS Wizard interface for a patient named Lily Nguyen. The interface is divided into several sections:

- Header:** Includes a navigation bar with "All Patients / Lily Nguyen" and an "Upload Samples" button.
- Sample Sets:** A list of sample sets with their upload status. The first set, dated 15 May 2019, shows three pairs of files (SRR8314991, SRR8314992, and SRR8314993) all in "Uploading..." status. The second set, dated 27 April 2019, shows three pairs of files, all marked as "Complete".
- Patient Information:** A sidebar on the right containing patient details: PATIENT ID (12345678), NAME (Lily), LAST NAME (Nguyen), DATE OF BIRTH (Nov 25, 19), GROUP (Room 101), DIAGNOSIS (Glioblastoma that forms from cells called astrocytes and starts in the brain. Glioblastoma is a type of brain tumor that can be life-threatening), and DATE OF DIAGNOSIS (Aug 17, 20).
- Comments:** A section at the bottom right for additional notes, currently containing the same text as the diagnosis.

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

IGS+WIZARD

by GENOMENAL

PATIENTS

All Patients

In Progress

Problems

Archive

COHORT

GB Primary

GB Recurrence

Research 1

Research 2

Research 3

Research 4

← All Patients / Lily Nguyen / ● SRR8314997_1.fastq — ● SRR8314997_1.fastq

Upload Samples

General Report

Main Info

Bioinformatic Report

Details

Parameters

1. File stats

Reads	● SRR8314997_1.fastq	● SRR8314997_2.fastq	○ SRR8314997_1.fastq	○ SRR8314997_2.fastq
FINAL	19,666,107	19,666,107	19,666,107	19,666,107
START	19,666,107	19,666,107	19,666,107	19,666,107
(%)	(100%)	(100%)	(100%)	(100%)

Start Reads

● SRR8314997_1.fastq ● SRR8314997_2.fastq ○ SRR8314997_1.fastq ○ SRR8314997_2.fastq

MIN LENGTH	36	36	36	36
MAX LENGTH	101	101	101	101
MEDIAN	101	101	101	101

2. Quality Report

Metric	● SRR8314997_1.fastq	● SRR8314997_2.fastq	○ SRR8314997_1.fastq	○ SRR8314997_2.fastq
TOTAL SEQUENCES	✓	✓	✓	✓
LENGTH DISTRIBUTION	✓	✓	✓	✓
TILES SEQUENCE QUALITY	✓	✓	✓	✓
FIRST BASE SEQUENCE QUALITY	✓	✓	✓	✓
MIDDLE SEQUENCE QUALITY	✓	✗	✓	✓
LAST BASE SEQUENCE QUALITY	✓	✓	✓	✓
OVERREPRESENTED SEQUENCES	✓	✓	✓	✓
ADAPTER CONTAMINATED	✓	✓	✗	✓

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

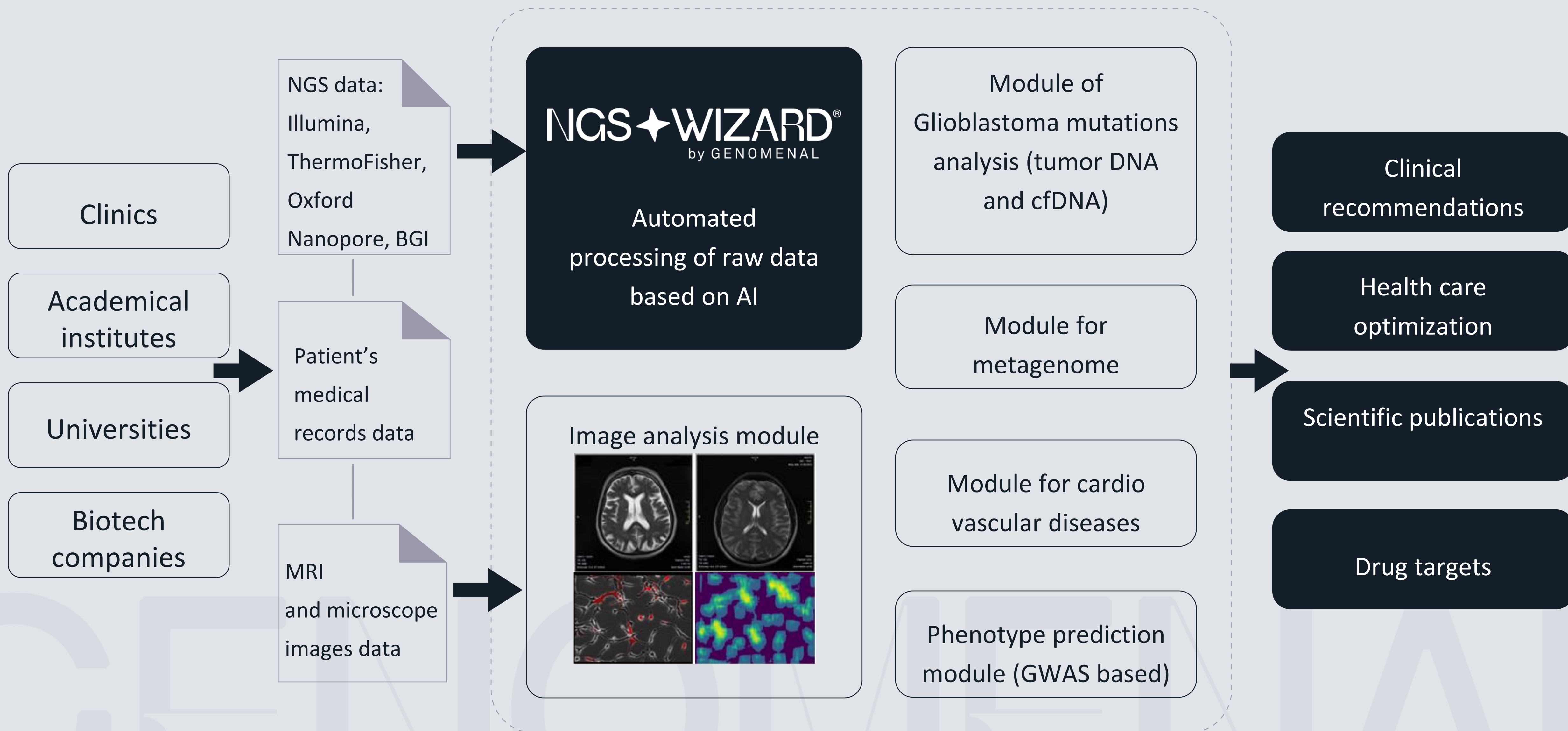
Somatic

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

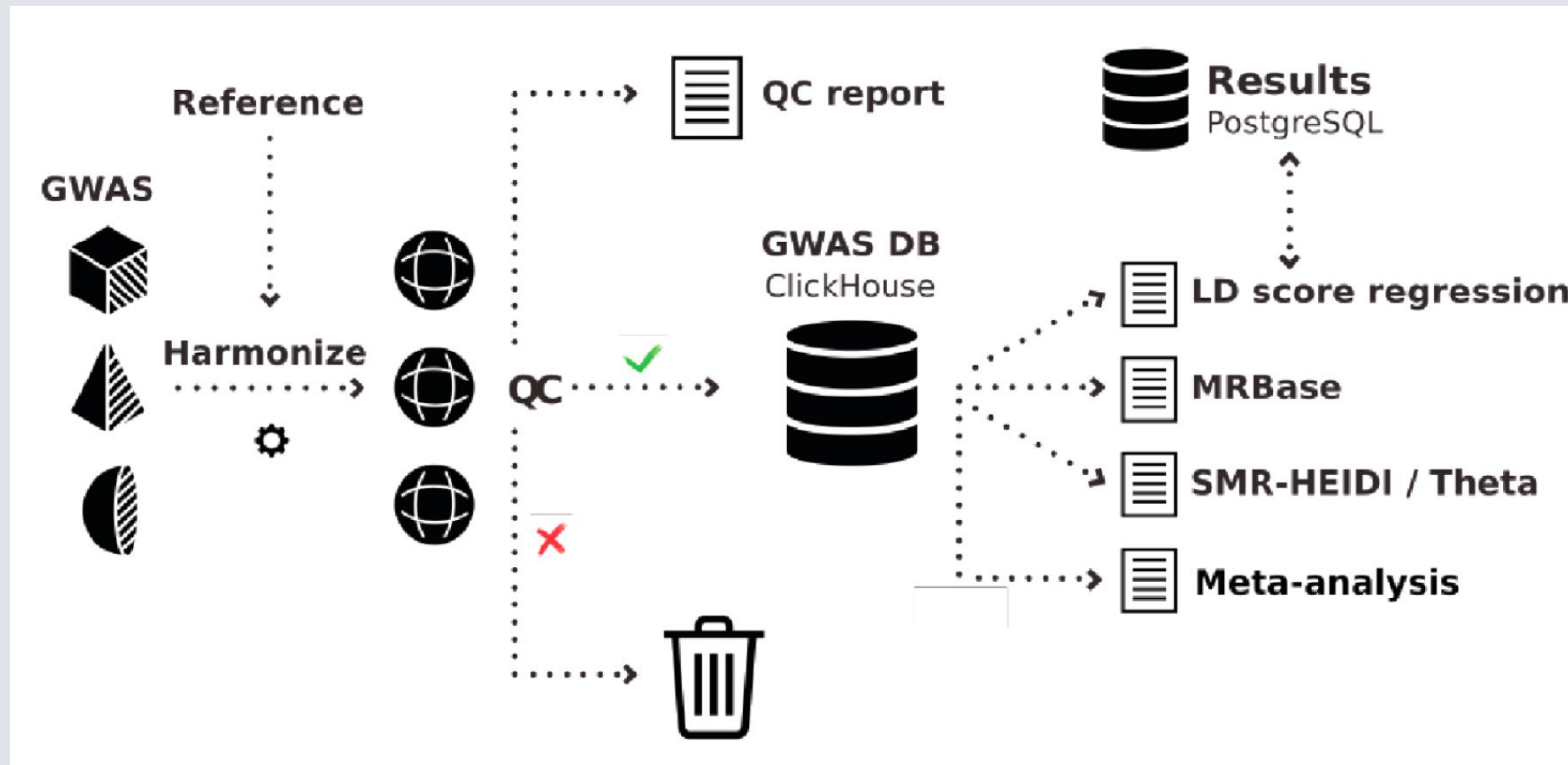
Gene	Variant	Pathogenicity
CDKN2A	NC_000009.12:g.21991924C>T	Pathogenic
KRAS	NC_000012.12:g.25245350C>T	Pathogenic
BRCA2	NC_000017.11:g.43092919G>A	Pathogenic

GENOMENAL platform structure



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

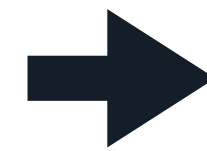
Momozawa Y. et al. IBD risk loci are enriched in multigenic regulatory modules encompassing putative causative genes. Nat Commun. 2018;9: 2427. doi:10.1038/s41467-018-04365-8

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

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

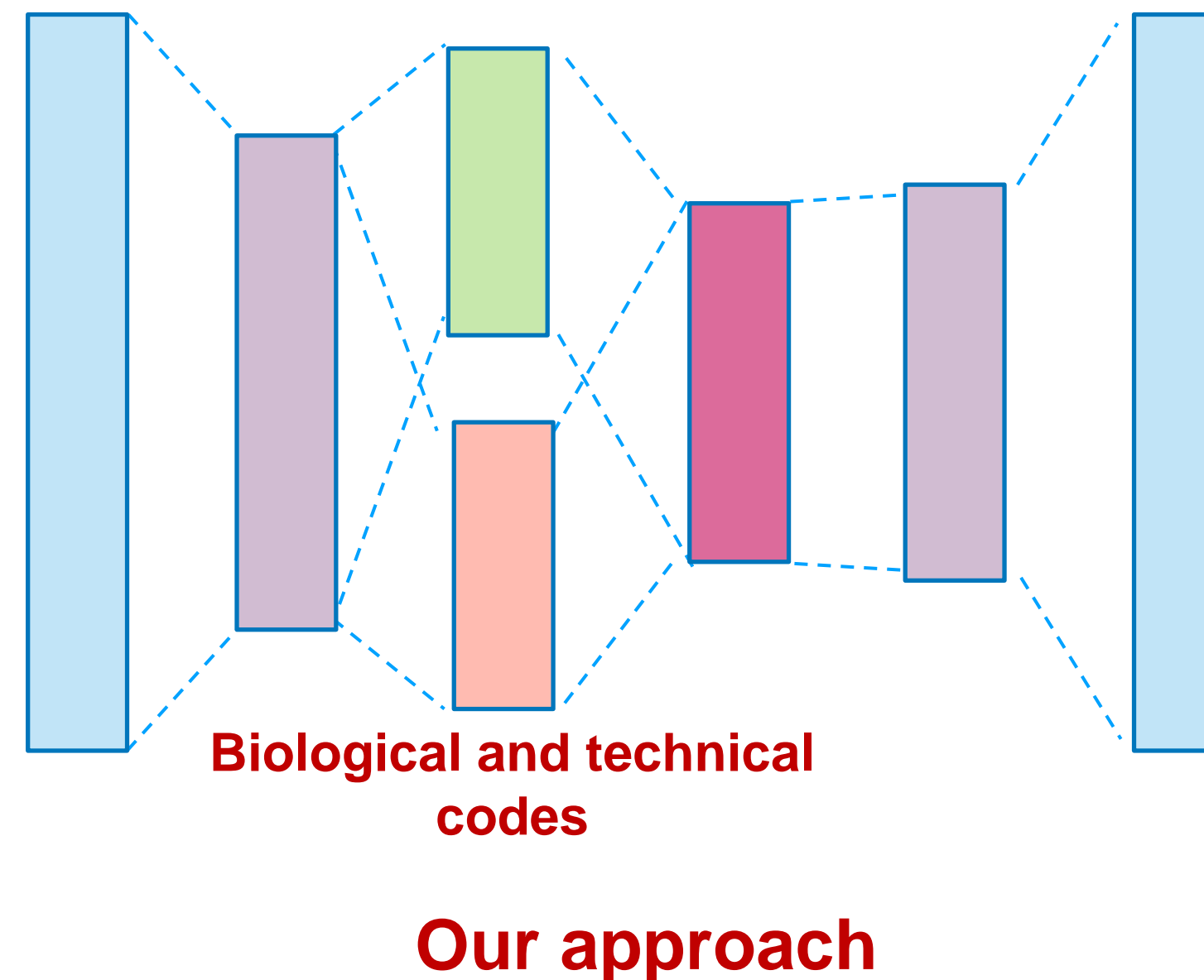
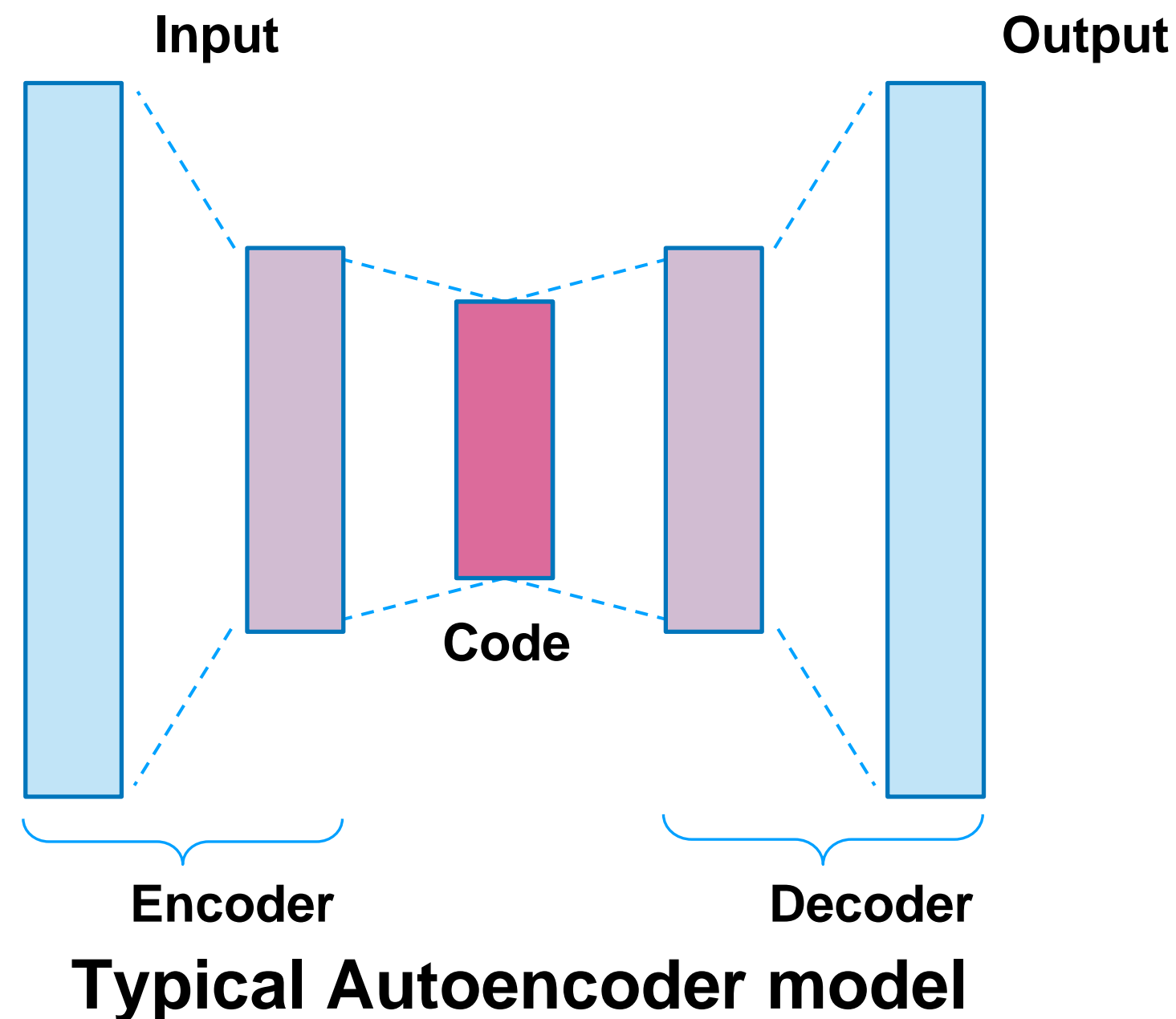
Pain sensitivity (model 1): -20%

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



Applications

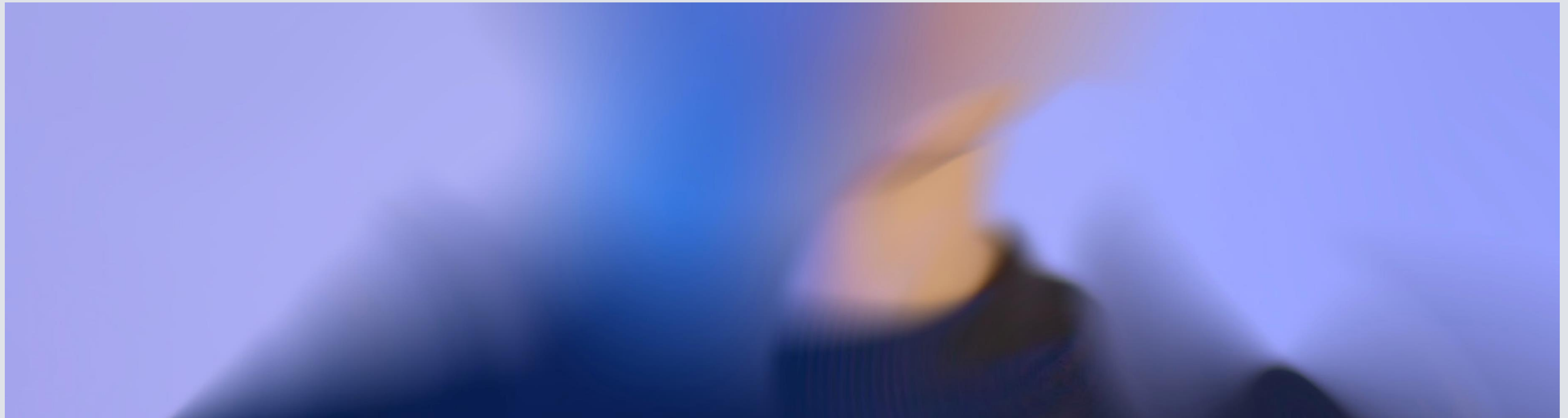
- Dimensionality reduction
- Missing data imputation
- **Data harmonization**
- **Data augmentation**
- **Virtual screening**
- **Biomarker research**

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

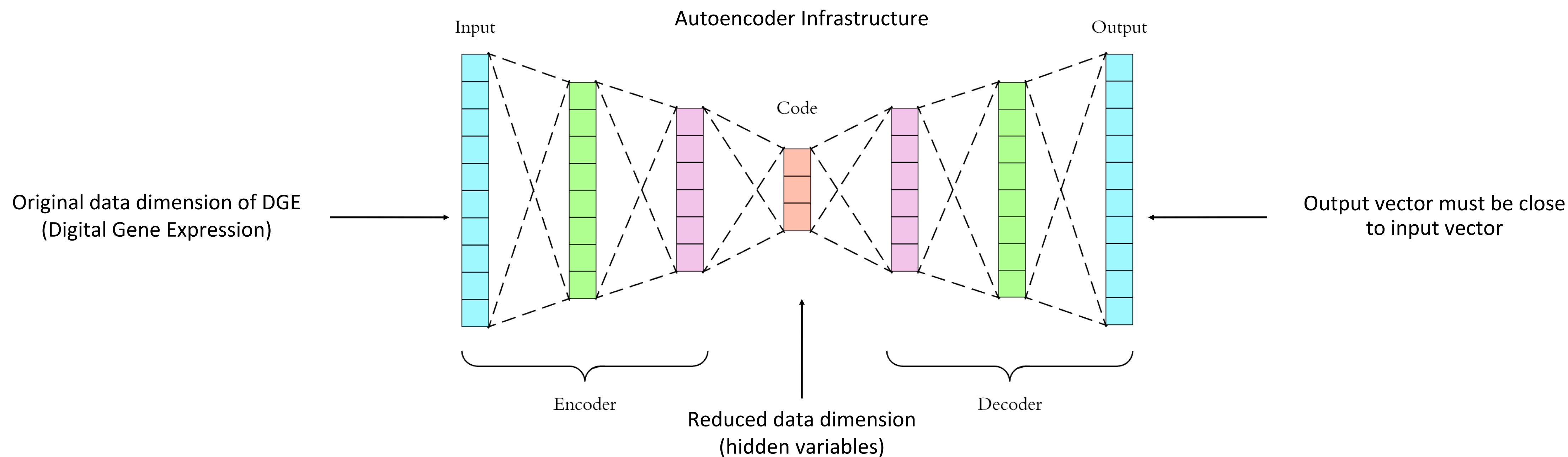
Currently this problems are recognized and several consortia work to collect and harmonize the publicly available gene expression data



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

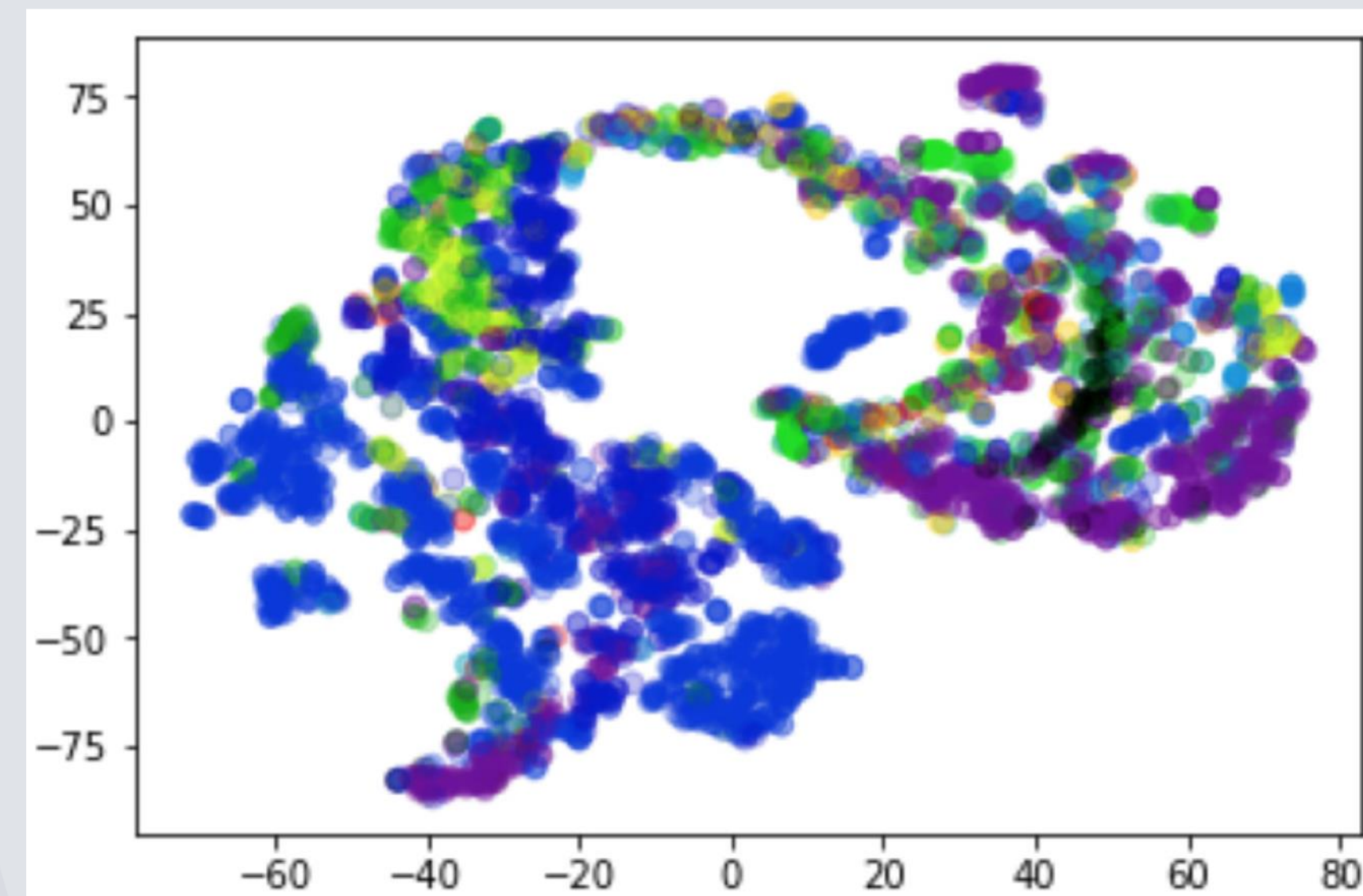
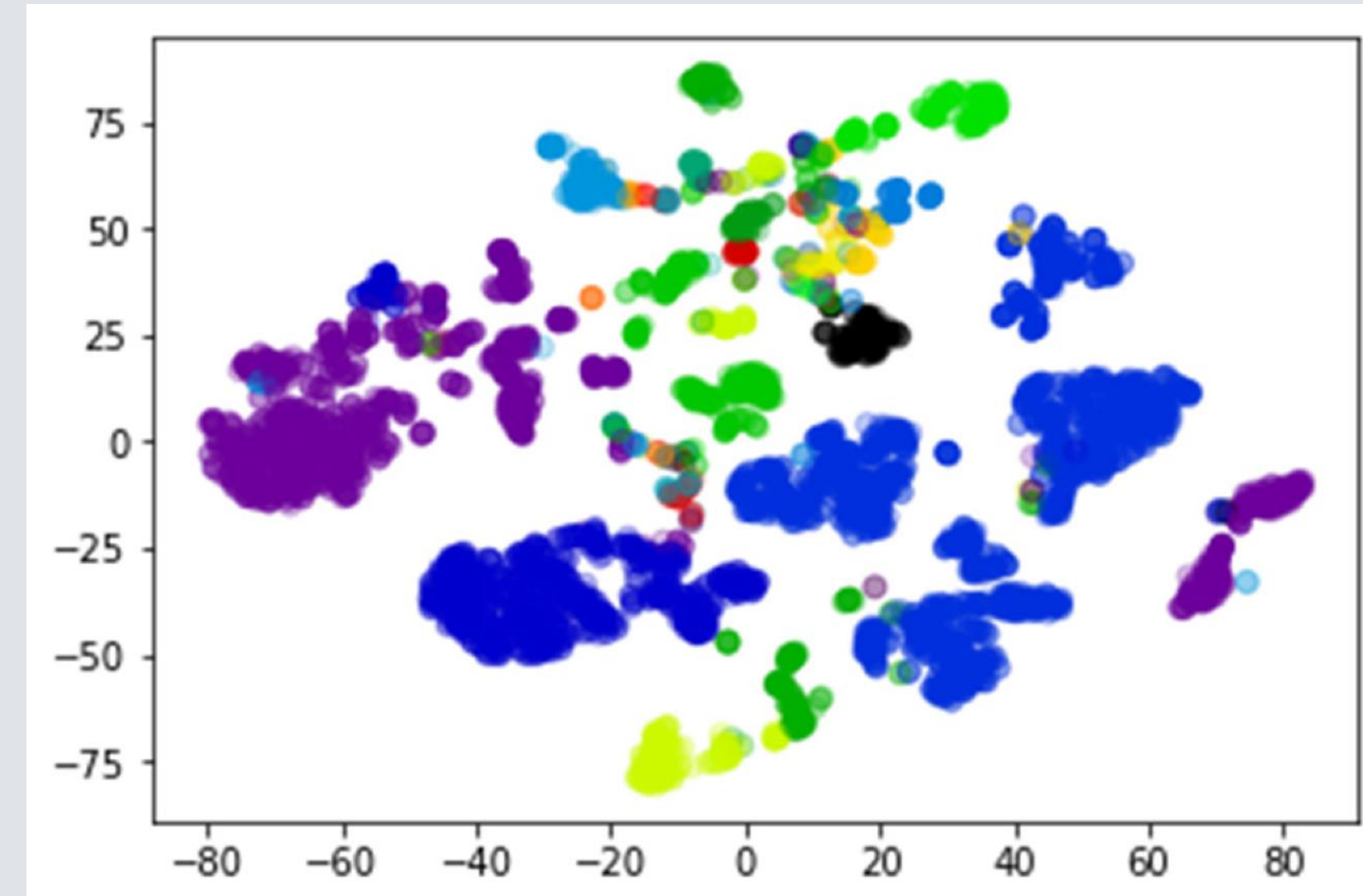


Autoencoder is used for

- ▶ 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

NOVEL

WIZARD

by GENOMENAL

PATIENTS

All Patients

In Progress

Problems

Archive

COHORT

GB Primary

GB Recurrence

Research 1

Research 2

Research 3

Research 4

JohnDoe@gmail.com

All Patients139

Upload Samples

Add Patient

PATIENT	STATUS	○ - ● (NORMAL-TUMOR PAIR)	GROUP	MODIFIED ↑
Lily Nguyen Patient Name	<div><div>↺</div><div>✓</div><div>✓</div></div>	<div><div>○</div><div>●</div><div>●</div><div>●</div><div>○</div><div>●</div><div>●</div><div>●</div></div>	Room 101	Today
Julie Simmons Patient Name	<div><div>✓</div></div>	<div><div>○</div><div>●</div><div>●</div></div>	Room 101	Today
Albert Robertson Patient Name	<div><div>⚠</div></div>	<div><div>○</div><div>●</div></div>		Today
Gloria Black Patient Name	<div><div>✓</div></div>	<div><div>○</div><div>●</div><div>●</div><div>●</div></div>	Room 315	Today
Ronald Bell Patient Name	<div><div>✗</div></div>	<div><div>○</div><div>●</div><div>●</div><div>○</div></div>	Room 205	Today
Francisco Murphy Patient Name	<div><div>✓</div><div>⚠</div></div>	<div><div>○</div><div>●</div><div>●</div><div>●</div><div>○</div><div>●</div><div>●</div><div>●</div></div>		Today
ID050919-1 Patient Name	<div><div>✓</div></div>	<div><div>○</div><div>●</div><div>●</div><div>●</div></div>	Room 205	05 Sept
ID050919-2 Patient Name	<div><div>✓</div></div>	<div><div>○</div><div>●</div><div>●</div></div>		05 Sept
ID050919-3 Patient Name	<div><div>✓</div></div>	<div><div>○</div><div>●</div></div>		05 Sept
ID050919-4 Patient Name	<div><div>✓</div></div>	<div><div>○</div><div>●</div></div>		05 Sept
ID050919-5 Patient Name	<div><div>✓</div></div>	<div><div>○</div><div>●</div><div>●</div><div>●</div></div>		05 Sept
ID050919-6 Patient Name	<div><div>✓</div></div>	<div><div>○</div><div>●</div></div>	Room 108	05 Sept
ID050919-7 Patient Name	<div><div>✓</div></div>	<div><div>○</div><div>●</div><div>●</div></div>		05 Sept
ID050919-8 Patient Name	<div><div>✓</div></div>	<div><div>○</div><div>●</div><div>●</div><div>●</div></div>	Room 1	05 Sept
ID050919-3 Patient Name	<div><div>✓</div></div>	<div><div>○</div><div>●</div></div>		05 Sept

Interface | All patients

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

NCS WIZARD®

PATIENTS

- All Patients
- In Progress
- Problems
- Archive

COHORT

- GB Primary
- GB Recurrence
- Research 1
- Research 2
- Research 3
- Research 4

Johndoe@gmail.com

← All Patients / Lily Nguyen / • SRR8314997_1.fastq — • SRR8314997_1.fastq

Upload Samples

General Report

Main Info

Bioinformatic Report

Details

Parameters

Somatic Variant Discovery Results

View Variants in IGV

Download annotated VCF

Open TSV in Jupyter Notebook

Download annotated TSV

Open TSV in Google Sheets

Workflow

1. Upload	Complete
2. Cleanup	2 of 11 metric failed
3. Alignment	Complete
4. Somatic Variant Discovery	Uploading...

Sample Info

ORGAN: Liver

DATE: Aug 17, 2019

SAMPLE TYPE: Tumor

TUMOR TYPE: Primary

COMMENTS: Write your comments here

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

NGS WIZARD

PATIENTS

All Patients

In Progress

Problems

Archive

COHORT

GB Primary

GB Recurrence

Research 1

Research 2

Research 3

Research 4

Johndoe@gmail.com

← All Patients / Lily Nguyen / • SRR8314997_1.fastq — • SRR8314997_1.fastq

Main Info

Bioinformatic Report

Details

Parameters

Quality Control ⓘ

Total Sequences

Min Total Sequences ⓘ

Length Distribution

Short Sequence Bound ⓘ

Max Short Sequences Percent ⓘ

10000

20

20

Alignment ⓘ

Reference genome

Alignment tool

Alignment quality thresholds

Min mapped reads percent

Max multimaps percent

Forward/reverse max difference

Min paired mapped percent

Min paired properly mapped reads percent

GRCh38 (hg38)

BWA Burrows-Wheeler Aligner

85

15

10

80

75

Variant Discovery ⓘ

Somatic variants discovery

Sequencing Type

Panel of normals

Germline variants discovery

QD SNP filter

QUAL SNP filter

SOR SNP filter

FS SNP filter

WGS

Illumina sureselect WES

2

30

3

60

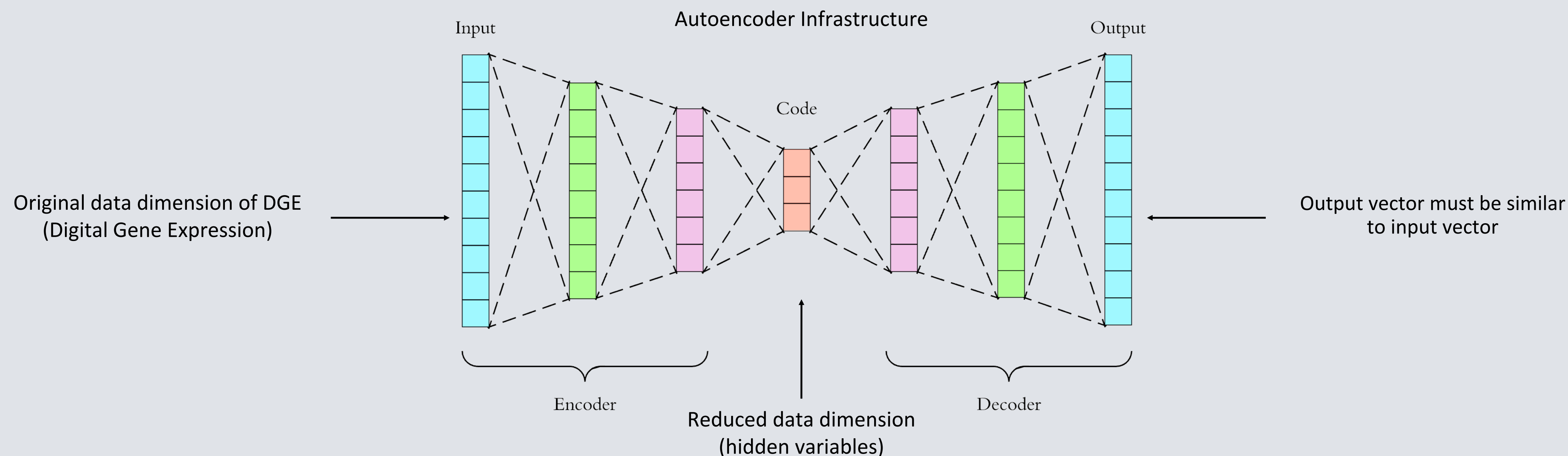
Applying parameter changes will restart samples workflow from affected step

Apply

Cancel

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