scieneers
DRIVEN BY DATA

UNIKLINIK RWTH AACHEN
Institut für Humangenetik
und Genommedizin

GenomAIx

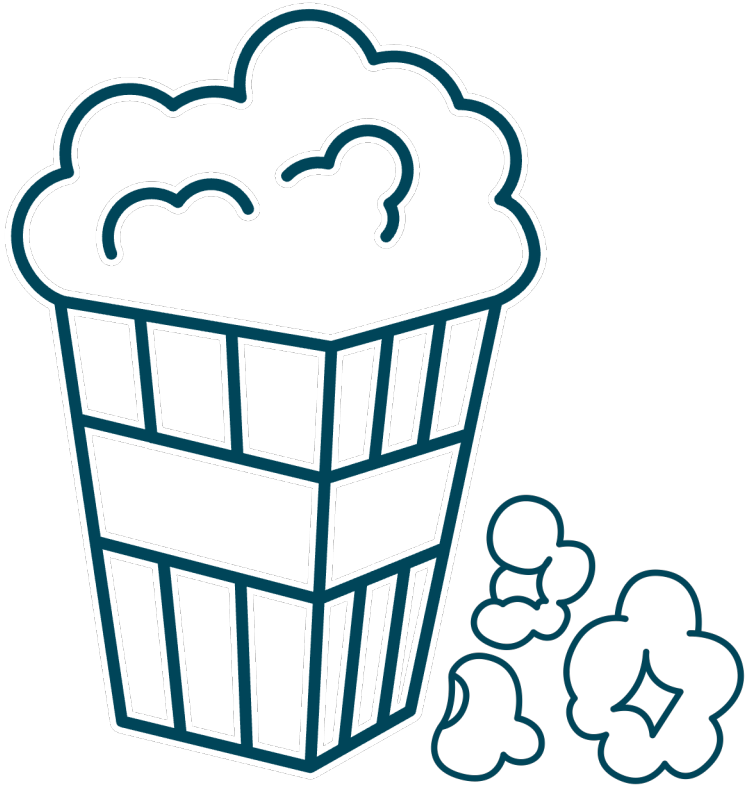
Erforschung des Dark Genome mit KI zur Entwicklung neuartiger Krankheitsinterventionen bei seltenen Erkrankungen

scieneers
DRIVEN BY DATA

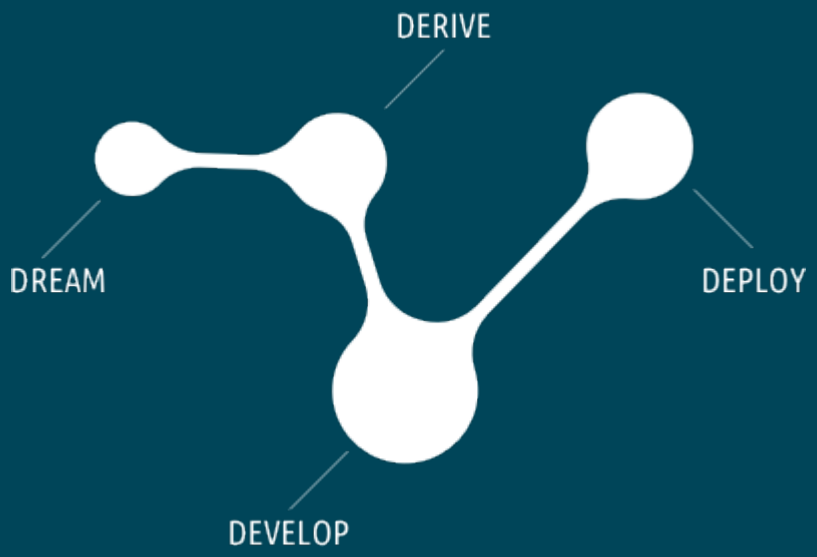
UNIKLINIK RWTH AACHEN
Institut für Humangenetik und Genommedizin

Outline

1. Short Introduction
2. Motivation
3. Workflow & Cloud Architecture
4. What we achieved so far
5. Next Steps
6. Discussion



Kurzvorstellung: scieneers GmbH



DREAM
Wir starten bei der Beratung, greifen Ihre Ideen auf, entwickeln gemeinsam den Business Case,

DERIVE
entwickeln Datenstrukturen und Modelle,

DEVELOP
implementieren die Lösung

DEPLOY
und integrieren diese nahtlos in Ihre operativen IT-Systeme und Geschäftsprozesse

Wir gewinnen Erkenntnisse aus Daten und schaffen damit Werte.
Für unsere Kunden, die Gesellschaft und uns selbst.

UNIKLINIK RWTH AACHEN

Institut für Humangenetik
und Genommedizin

scieneers
DRIVEN BY DATA

UNIKLINIK RWTH AACHEN
Institut für Humangenetik
und Genommedizin

GenomAIx

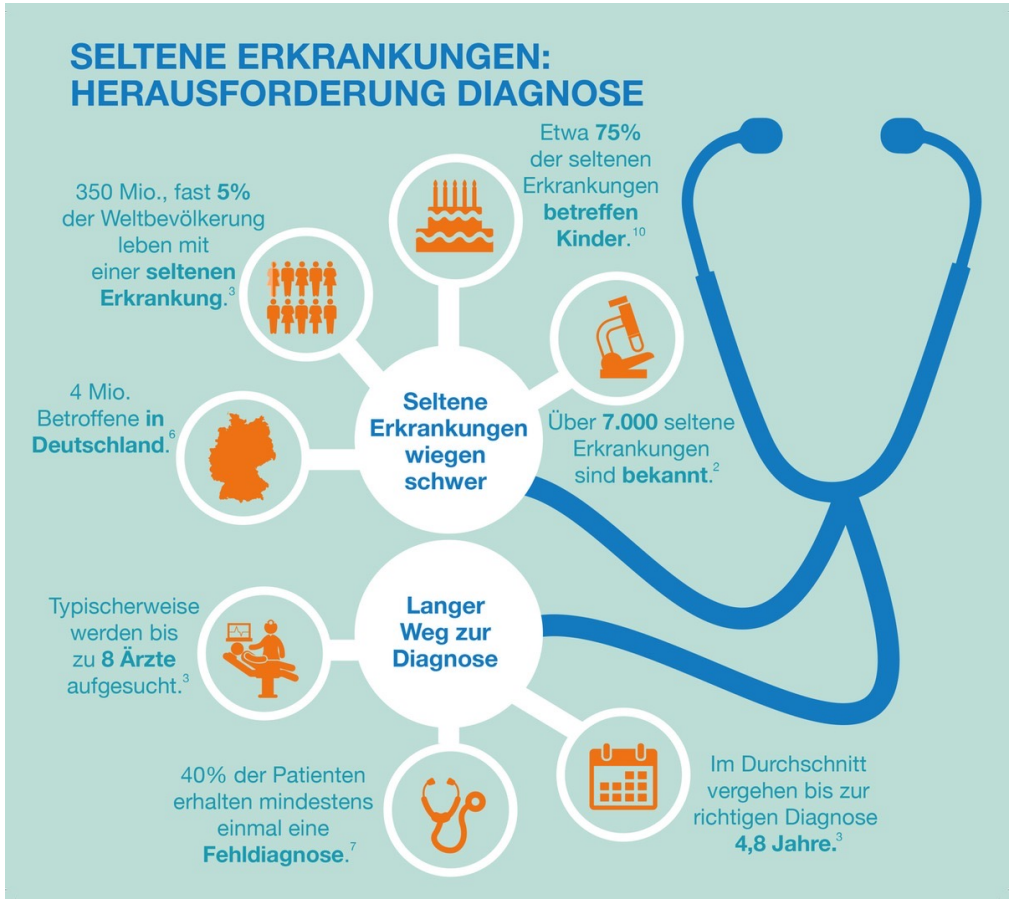
Motivation

scieneers
DRIVEN BY DATA

UNIKLINIK RWTH AACHEN
Institut für Humangenetik und Genommedizin

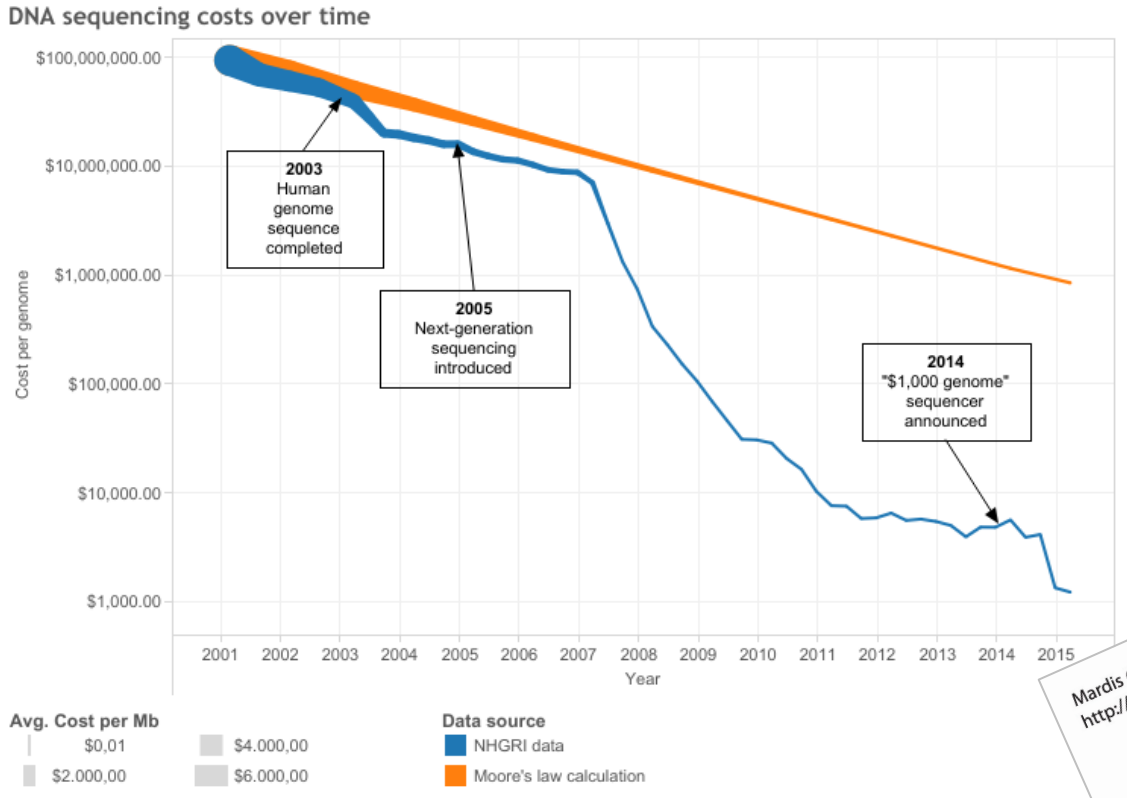
Genomsequenzierung bei seltenen Erkrankungen

Seltene Erkrankungen

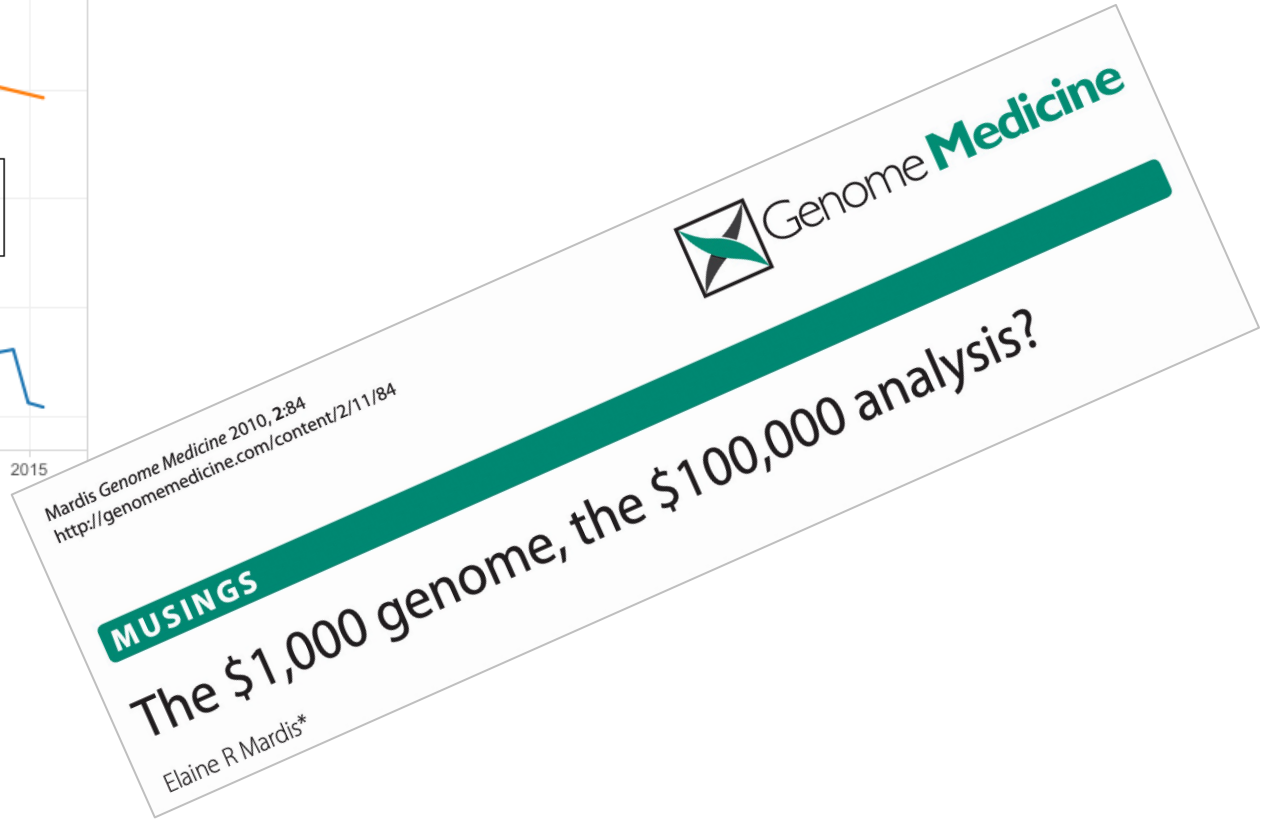


80% dieser Erkrankungen entstehen durch eine einzige genetische Veränderung und können durch eine Genomanalyse diagnostiziert werden.

Kosten für Genome sinken…
Auswertung ist das Bottleneck



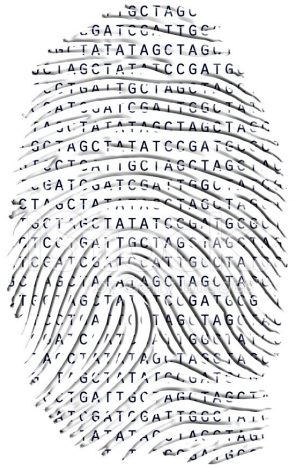
Decline in real costs compared to expected declines based on Moore's Law.
Trend line: Cost per human genome. Line width: Cost per megabase (Mb)
(*Data: NHGRI* https://www.genome.gov/27541954/dna-sequencing-costs-data/)



scieneers
DRIVEN BY DATA

UNIKLINIK RWTH AACHEN
Institut für Humangenetik und Genommedizin

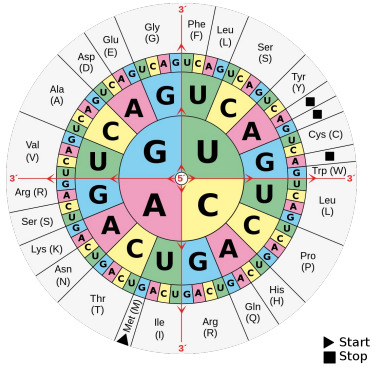
Genomsequenzierung bei seltenen Erkrankungen



Bausteine des Genoms	3.3 Milliarden
Varianten im Genom	3.5 Millionen
Varianten im Exom	70.000
Seltene Varianten	3.000
„Ultraseltene“ Varianten	300

...but what are variants?

Let´s do a short recap on Protein-Synthesis



Codon Wheel

DNA

Transcription

AUG ACU CUA GAG GGU AGC GCC

mRNA

Translation

M (Start) T L Q G S A

Aminoacid Sequence

Protein

... but what if we encounter a so called single nucleotide variant (SNV)

AUG ACU CCA GAG GGU AGC GCC

mRNA

Translation

M T P Q G S A

Aminoacid Sequence

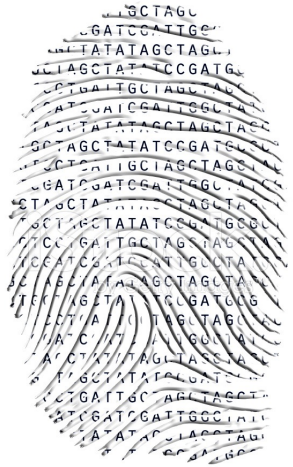
Protein

Note: On the right side a single nucleotide variant is displayed leading to a Missense Mutation.
It is important to mention that there are also other types of mutations like Non-Sense, Synonym and Loss-of-Function Mutations.

scieneers
DRIVEN BY DATA

UNIKLINIK RWTH AACHEN
Institut für Humangenetik und Genommedizin

Genomsequenzierung bei seltenen Erkrankungen



Bausteine des Genoms	3.3 Milliarden
Varianten im Genom	3.5 Millionen
Varianten im Exom	70.000
Seltene Varianten	3.000
„Ultraseltene“ Varianten	300

Auf der Suche nach **DER EINEN (!)** Variante, die die Krankheit ausgelöst hat!
Welche Veränderung ist pathogen? Welche harmlos?

scieneers
DRIVEN BY DATA

UNIKLINIK
RWTH AACHEN

Institut für Humangenetik
und Genommedizin


Oder…

scieneers
DRIVEN BY DATA

UNIKLINIK RWTH AACHEN
Institut für Humangenetik und Genommedizin

Verständnis genomischer Varianten

Wo stehen wir?

Gedankenexperiment

Wir bilden die gesamte genetische Information eines Menschen auf 52 Karteikarten ab:

Die Informationen auf 51 Karteikarten entziehen sich bislang der Bewertbarkeit!

Das entspricht einem Anteil von **98%** an unserem Genom. Dieser Teil des Genoms wird deshalb auch als Dark Genome oder Non-Coding Region bezeichnet.

Dort sind u.a. die bis zu 300.000 vorhergesagten small Open Reading Frames (sORFs) beheimatet. Diese sind bisher kaum verstanden. Einzelne Varianten in diesen sORFs konnte bereits eine Pathogenität nachgewiesen werden.

Unser bisheriges Verständnis des Genoms kann auf einer Karteikarte abgebildet werden.

Das entspricht einem Anteil von **2%** unseres Genoms und wird als Exom oder Coding-Region bezeichnet

Dort sind u.a. unsere ca. 22.000 Gene beheimatet, deren Varianten vergleichsweise gut verstanden sind.

scieneers
DRIVEN BY DATA

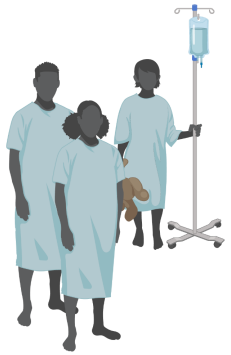
UNIKLINIK RWTH AACHEN
Institut für Humangenetik und Genommedizin

GenomAIx

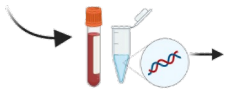
Erforschung des Dark Genome mit KI zur Entwicklung neuartiger Krankheitsinterventionen

Klinischer Bedarf

500.000 Krebs-Neuerkrankungen pro Jahr,
4 Mio Menschen mit Seltenen Erkrankungen

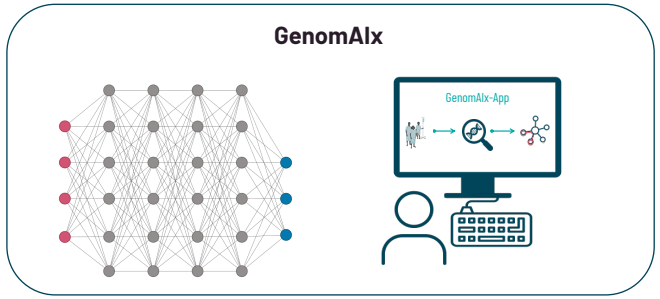


Genomanalyse



Next-Generation Sequencing (NGS)

Genomauswertung / Bioinformatik



- Gesamtzahl zu bewertender Varianten pro Genom: 3,5 Millionen
- Hinsichtlich Krankheitsrelevanz auswertbare Regionen (**grün**)

Aktuelle Software-lösungen

Phase 1
Proof-of-concept: sORFs

Phase 2
Transfer auf andere Genom-abschnitte

Verbessertes Genomverständnis / Verbesserte Krankenversorgung

Verständnis von Genomveränderungen

Therapieentscheidung auf Basis bewertbarer Genomveränderungen

(Personalisierte Medizin)

scieneers
DRIVEN BY DATA

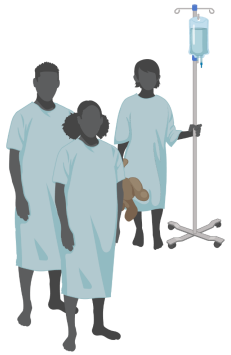
UNIKLINIK RWTH AACHEN
Institut für Humangenetik und Genommedizin

GenomAIx

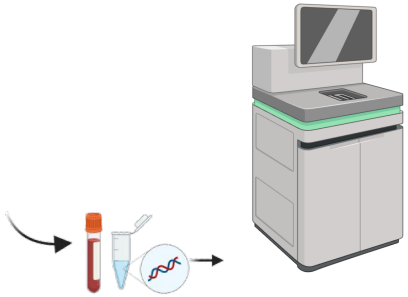
Erforschung des Dark Genome mit KI zur Entwicklung neuartiger Krankheitsinterventionen

Klinischer Bedarf

500.000 Krebs-Neuerkrankungen pro Jahr,
4 Mio Menschen mit Seltenen Erkrankungen

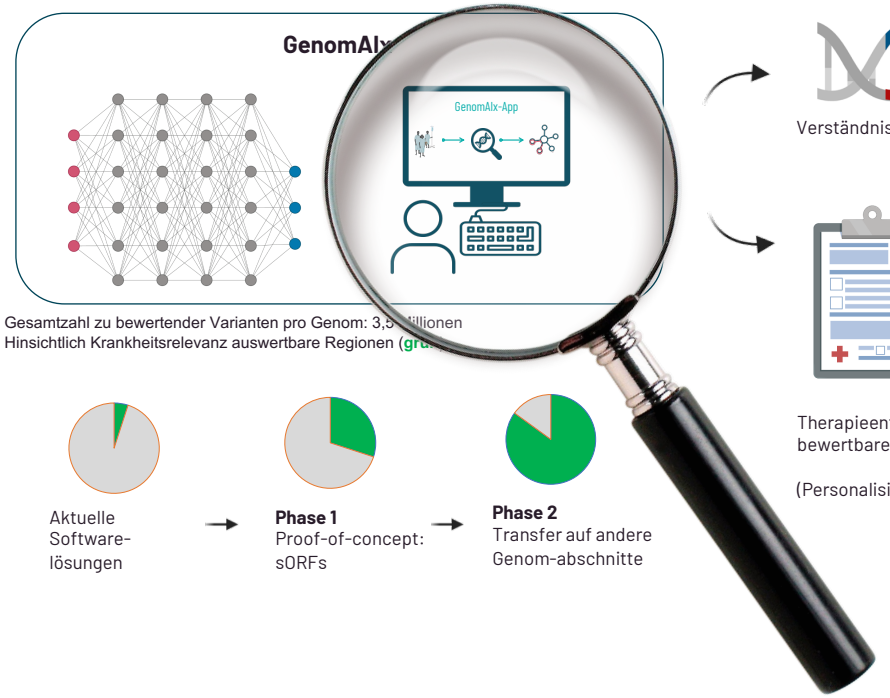


Genomanalyse



Next-Generation Sequencing (NGS)

Genomauswertung / Bioinformatik



- Gesamtzahl zu bewertender Varianten pro Genom: 3,5 Millionen
- Hinsichtlich Krankheitsrelevanz auswertbare Regionen (gr

Aktuelle Software-lösungen

Phase 1
Proof-of-concept: sORFs

Phase 2
Transfer auf andere Genom-abschnitte

Verbessertes Genomverständnis / Verbesserte Krankenversorgung

Verständnis von Genomveränderungen

Therapieentscheidung auf Basis bewertbarer Genomveränderungen

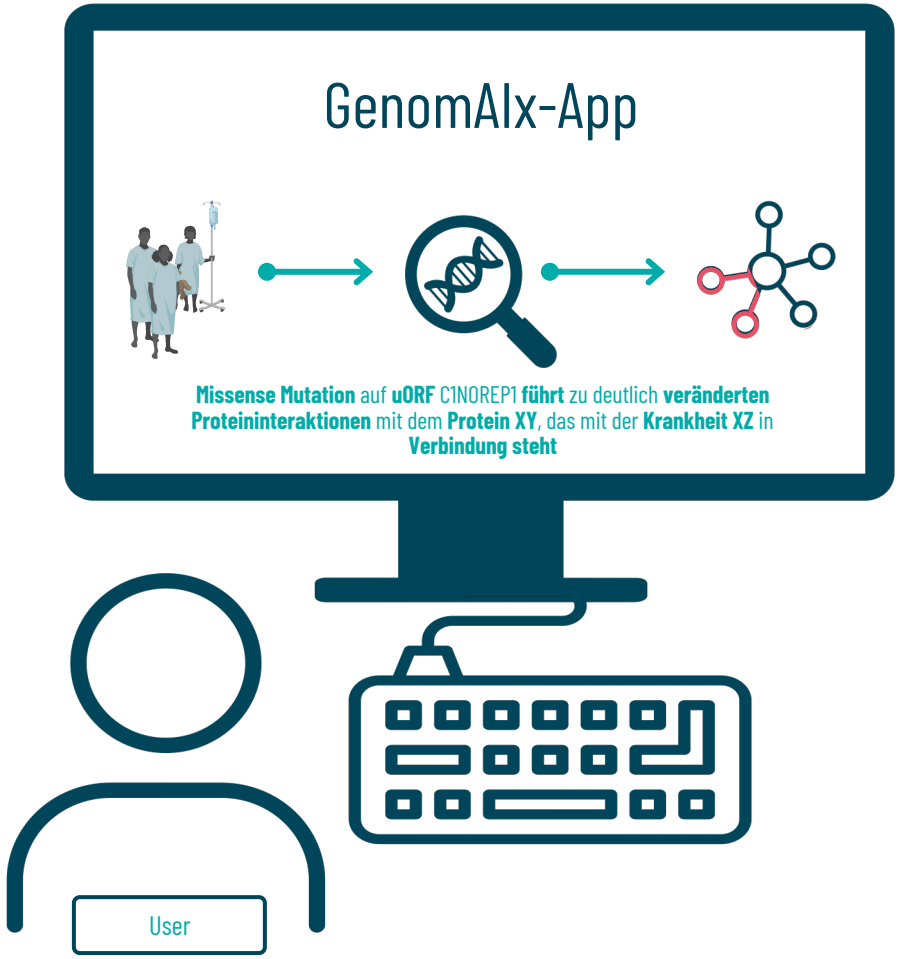
(Personalisierte Medizin)

scieneers
DRIVEN BY DATA

UNIKLINIK RWTH AACHEN
Institut für Humangenetik und Genommedizin

GenomAIx

eine cloudbasierte SaaS-Anwendung zur umfänglicheren Genomauswertung



scieneers
DRIVEN BY DATA

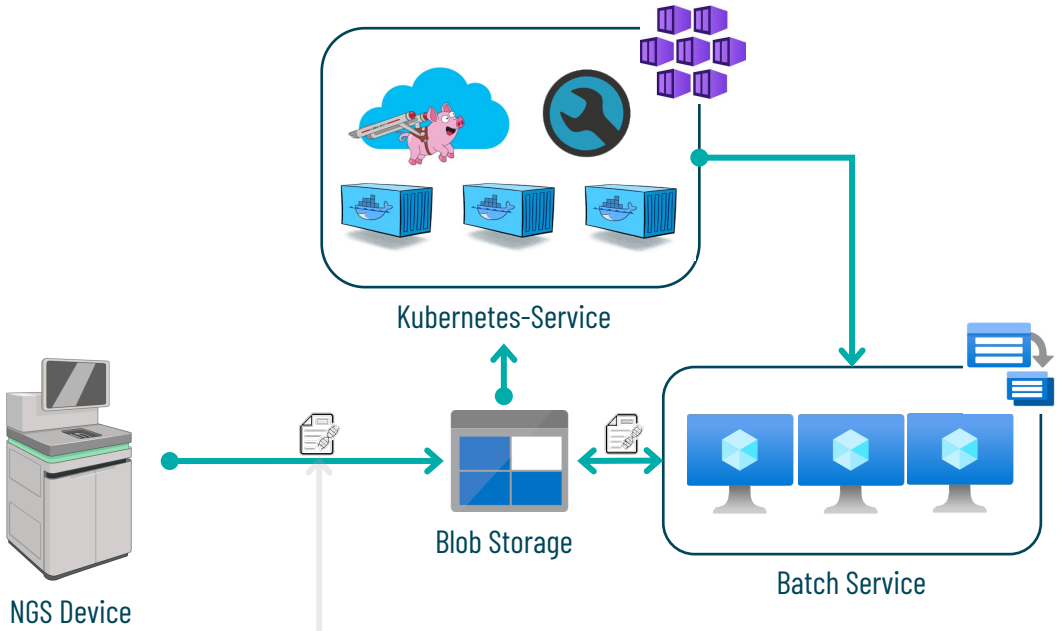
UNIKLINIK RWTH AACHEN
Institut für Humangenetik
und Genommedizin

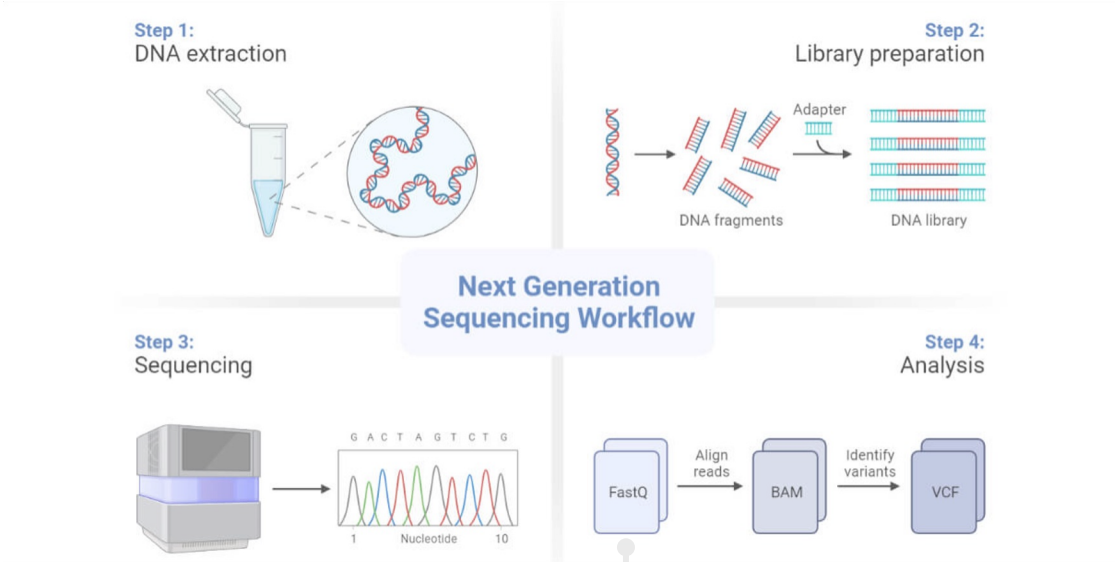
GenomAIx

Workflow & Cloud Architecture

GenomAIx Workflow - Step 1

Variant Calling and Variant Effect Prediction



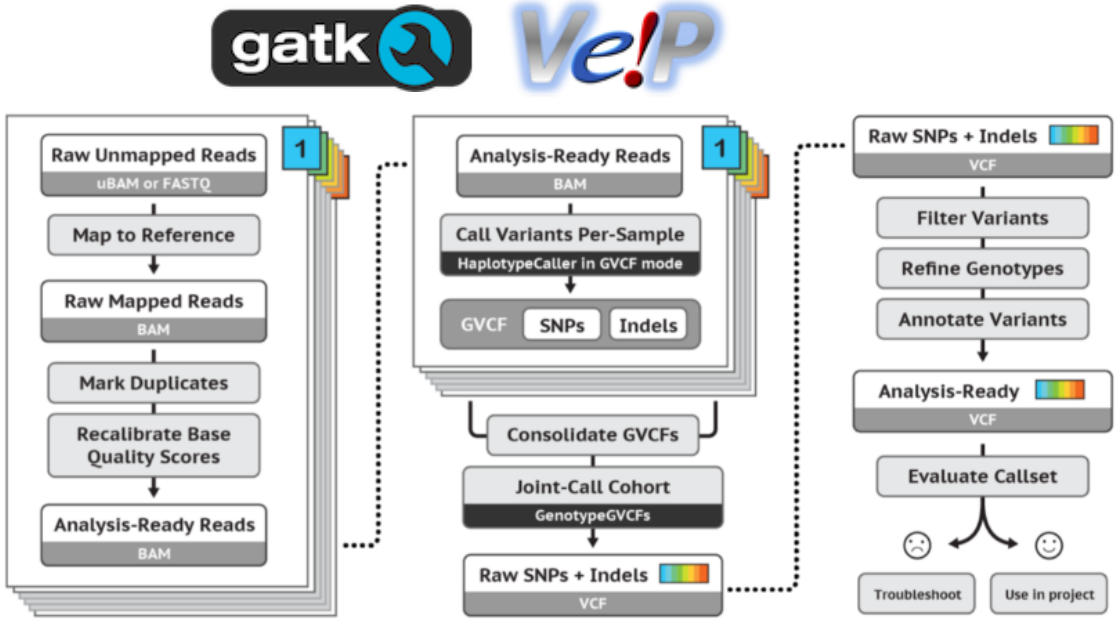


https://app.biorender.com/biorender-templates/t-62ccc36459dc10d0850c42df

Entry Point

GenomAIx Workflow - Step 1

GATK and VEP for Variant Calling and Variant Effect Prediction



http://www.broadinstitute.org/gatk/guide/best-practices

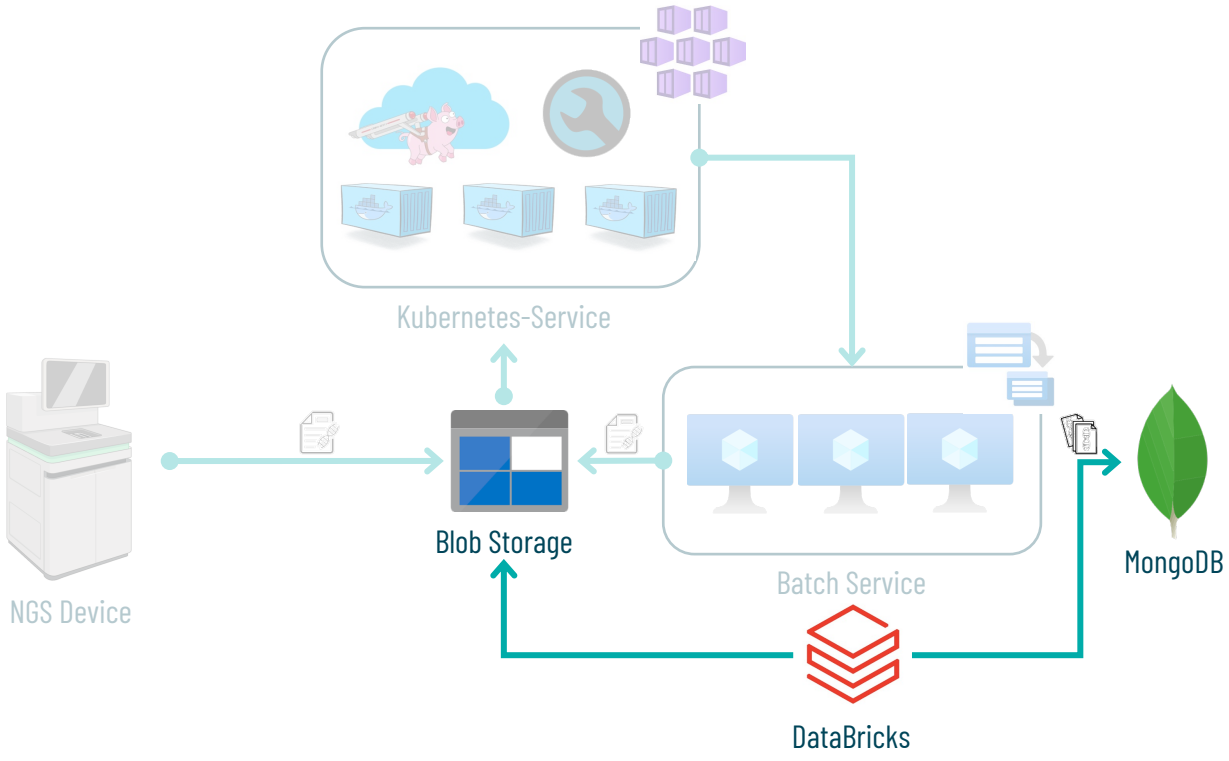
```
1 Ensembl gene        1000  5000  . + . gene_id "gene1"; gene_name "GENE1";
1 Ensembl transcript  1100  4900  . + . gene_id "gene1"; transcript_id "transcript1"; gene_name "GENE1"; transcript_name "GENE1-001"; transcript_biotype "protein_coding";
1 Ensembl exon        1200  1300  . + . gene_id "gene1"; transcript_id "transcript1"; exon_number "exon1"; exon_id "GENE1-001_1";
1 Ensembl exon        1500  3000  . + . gene_id "gene1"; transcript_id "transcript1"; exon_number "exon2"; exon_id "GENE1-001_2";
1 Ensembl exon        3500  4000  . + . gene_id "gene1"; transcript_id "transcript1"; exon_number "exon3"; exon_id "GENE1-001_2";
1 Ensembl CDS         1300  3800  . + . gene_id "gene1"; transcript_id "transcript1"; exon_number "exon2"; ccds_id "CDS0001";
```

scieneers
DRIVEN BY DATA

UNIKLINIK RWTH AACHEN
Institut für Humangenetik und Genommedizin

GenomAIx Workflow - Step 2

Spark based ETL-Pipelines for Constellation Calling and Variant Filtering



scieneers
DRIVEN BY DATA

UNIKLINIK RWTH AACHEN
Institut für Humangenetik und Genommedizin

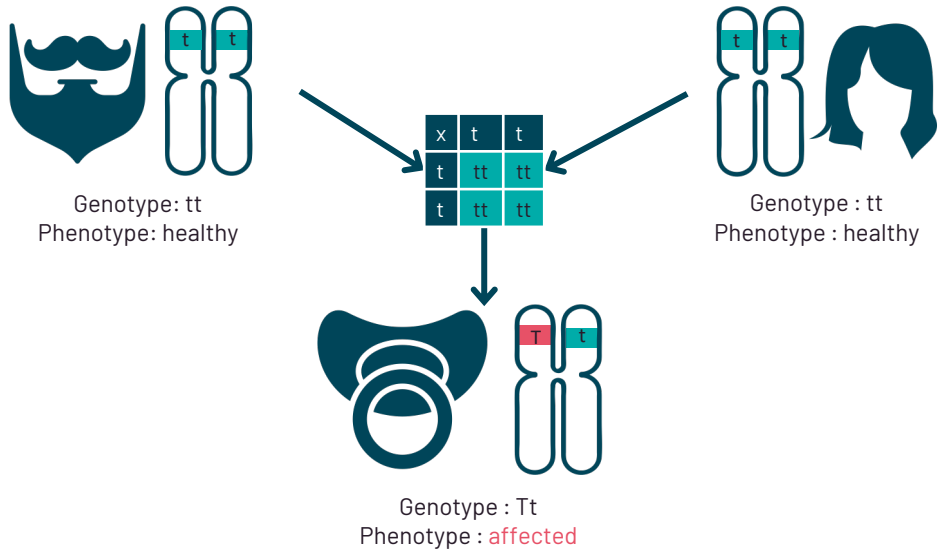
GenomAIx Workflow - Step 2

ETL-Pipelines for Constellation Calling e.g. De-Novo Calling for Trio-Datasets

Input

On average 3.3 Mio variants per individual
(10 Mio per Trio-Dataset)

De-Novo Calling



Output

On average 0 to 5 variants

scieneers
DRIVEN BY DATA

UNIKLINIK RWTH AACHEN
Institut für Humangenetik und Genommedizin

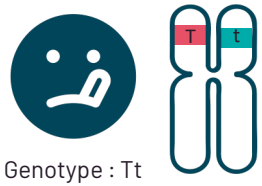
GenomAIx Workflow - Step 2

ETL-Pipelines for Calling Rare Variants by means of Population Genomics

Input

On average 3.3 Mio variants

Custom Filters from Population Genomics



Genotype : Tt
Phenotype : affected

sORF specific metrics derived from Population Analysis using Gnomad & Clinvar

Output

On average 0 to 10 variants for affected individual

GenomAIx Workflow - Step 2

Population Genomics e.g. PLI – Score

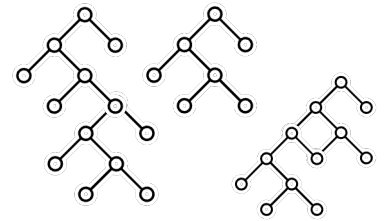
Expected count: A depth-corrected probability prediction model that takes into account sequence context, coverage, and methylation to predict expected variant counts

Observed count: The number of unique single-nucleotide variants in each transcript/gene oberserved in a sample population e.g. *Gnomad with 76.210* whole genome sequences

Observed/Expected ratio (O/E): When a gene has a low O/E value, it is under stronger selection for that class of variation than a gene with a higher O/E value.

pLI-score: probability of being loss-of-function (LoF) intolerant (pLI)

XGBoost

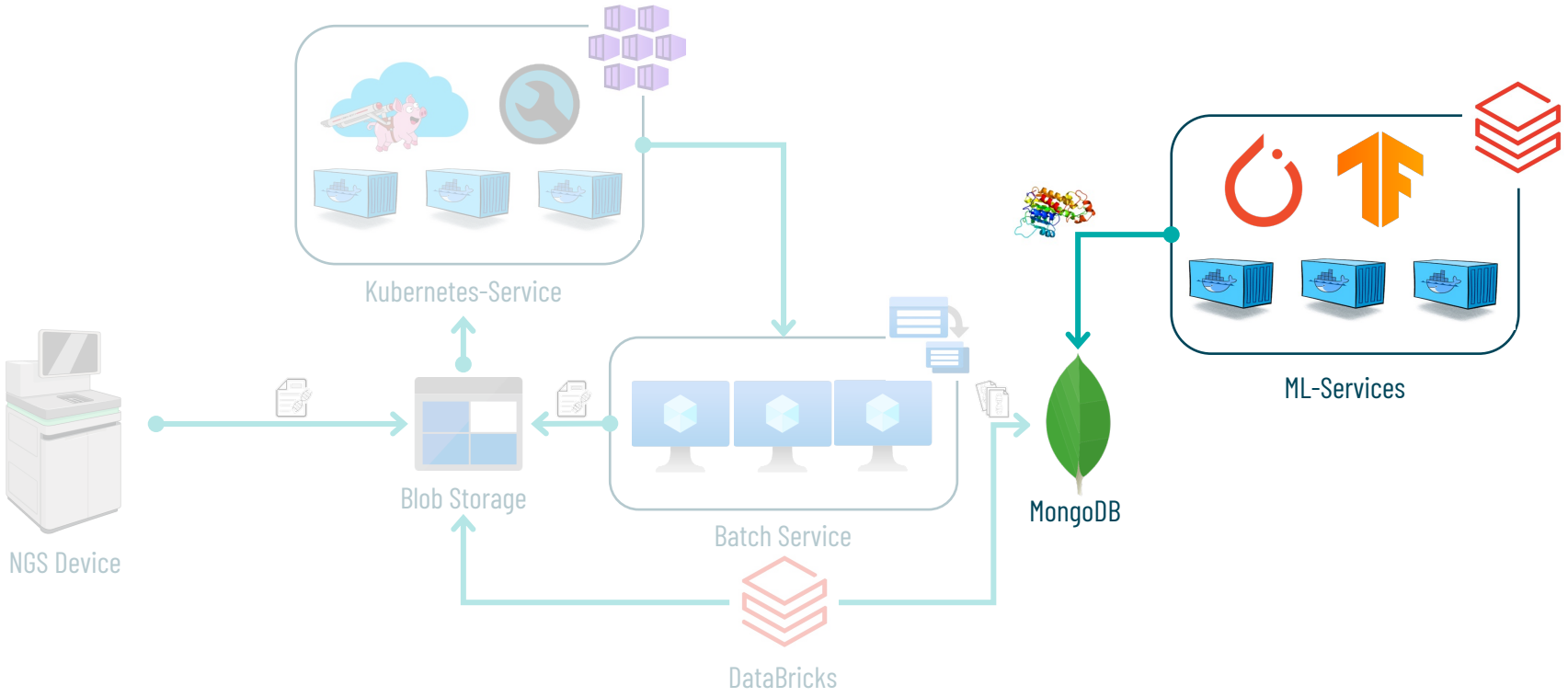


scieneers
DRIVEN BY DATA

UNIKLINIK RWTH AACHEN
Institut für Humangenetik und Genommedizin

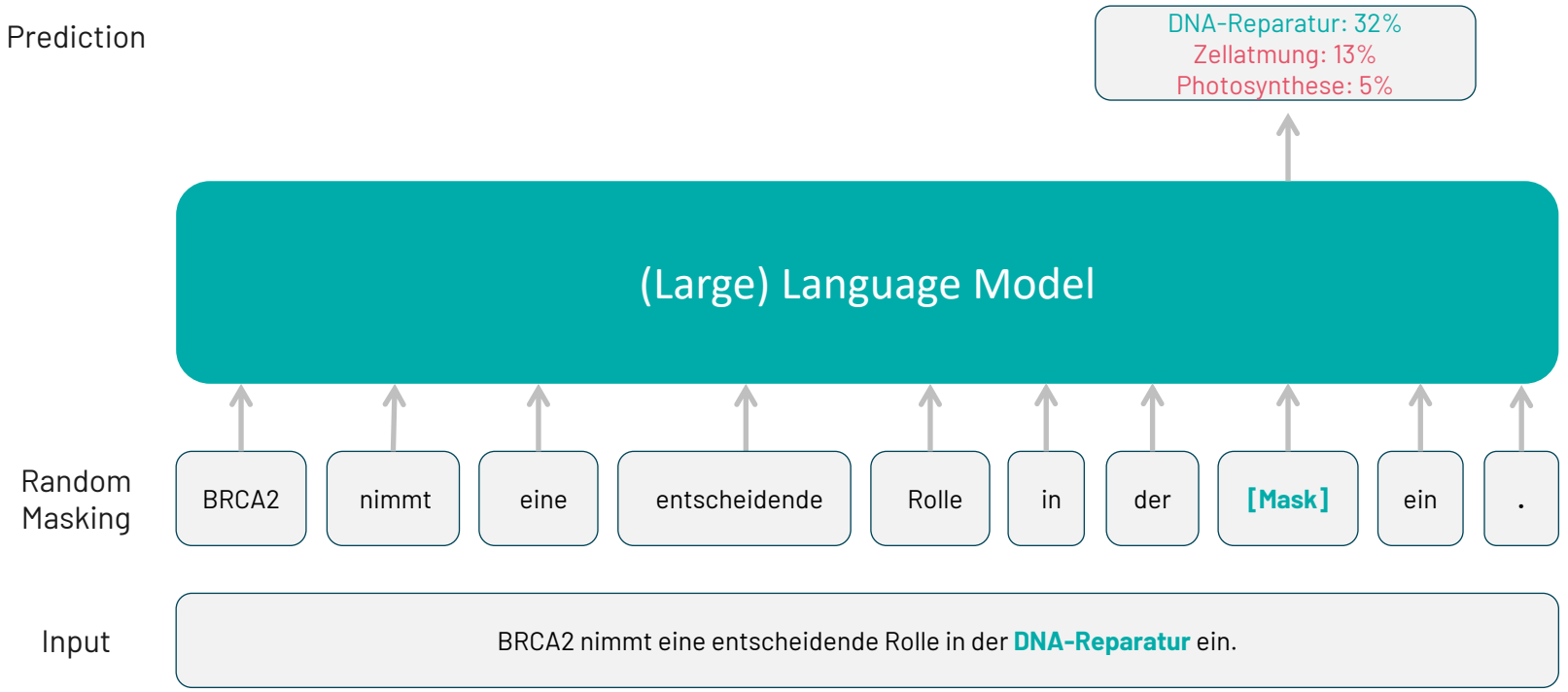
GenomAIx Workflow – Step 3

Data Enrichment using ML- Services



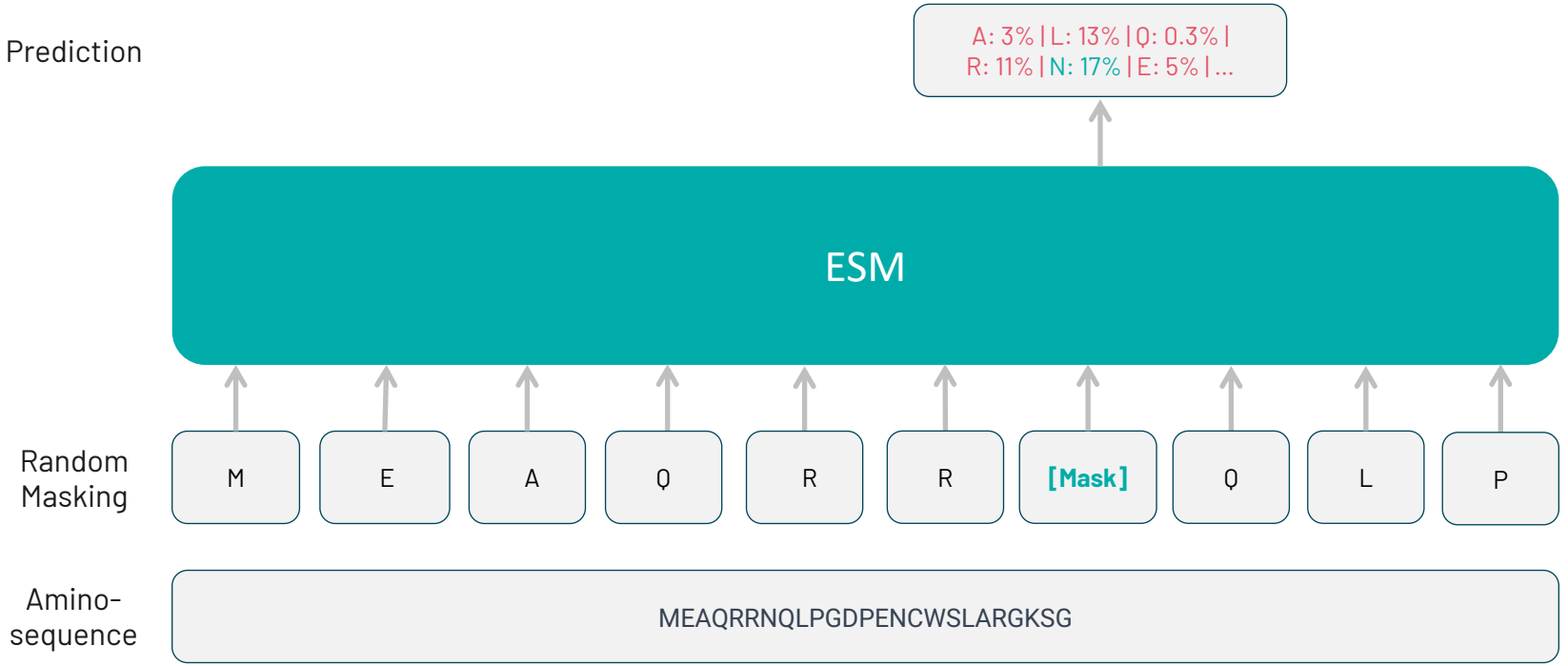
GenomAIx Workflow – Step 3

From LLMs towards Protein-Language Models via Self-Supervised Learning



GenomAIx Workflow – Step 3

From LLMs towards Protein-Language Models via Self-Supervised Learning



There are many similarities while comparing Aminosequences of proteins to language. Nevertheless, proteins do not have clear-cut multi-letter building blocks (such as words and sentences). They are more variable in length than sentences, and show many interactions between distant positions

ProteinBERT: A universal deep-learning model of protein sequence and function, Brandes et al. 2022

GenomAIx Workflow – Step 3

ESM-Fold: Protein Structure Prediction using Language Models

Evolutionary-scale prediction of atomic-level protein structure with a language model

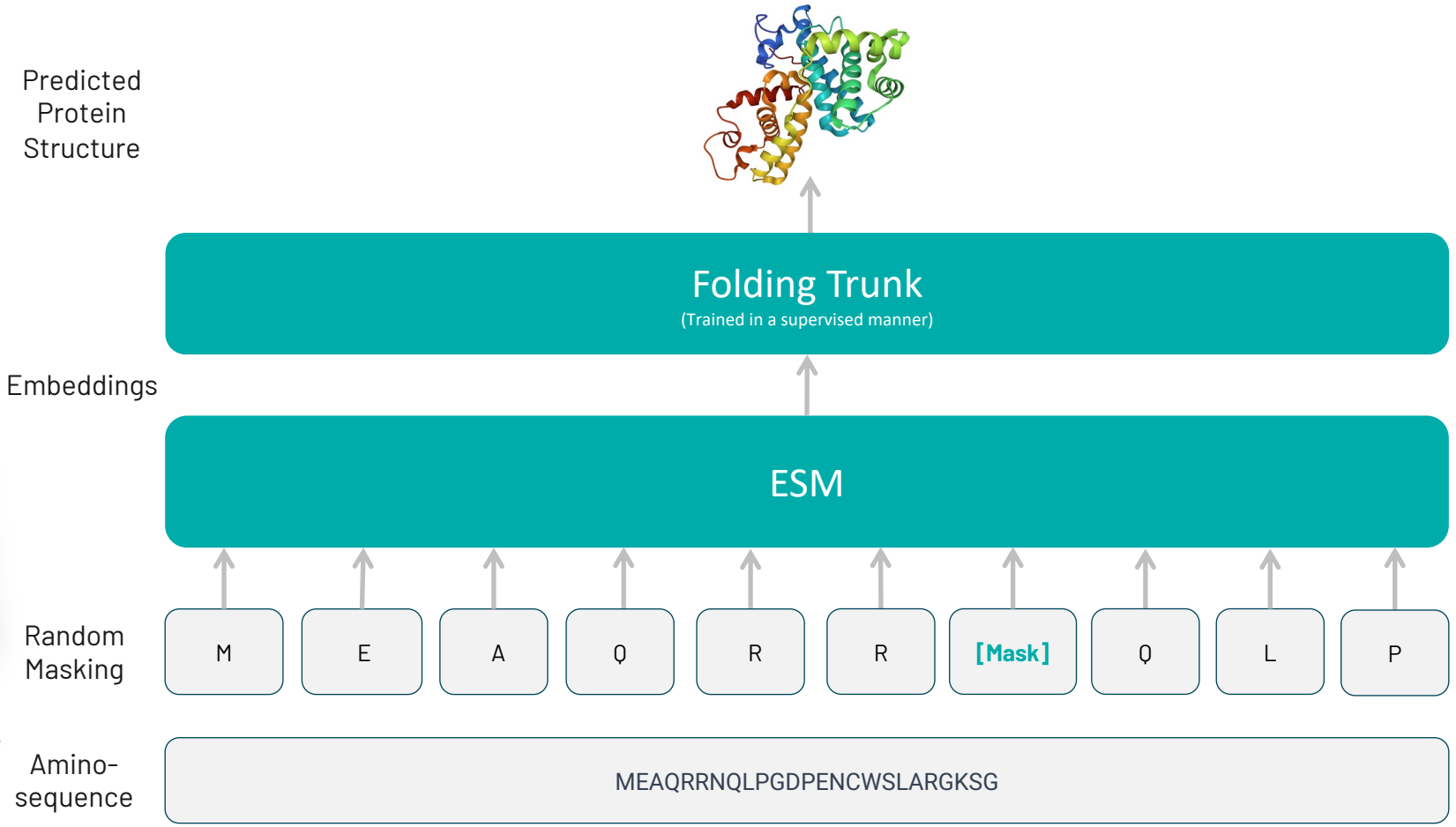
Speedy structures from single sequences

Machine learning methods for protein structure prediction have taken advantage of the evolutionary information present in multiple sequence alignments to derive accurate structural information, but predicting structure accurately from a single sequence is much more difficult. Lin *et al.* trained transformer protein language models with up to 15 billion parameters on experimental and high-quality predicted structures and found that information about atomic-level structure emerged in the model as it was scaled up. They created ESMFold, a sequence-to-structure predictor that is nearly as accurate as alignment-based methods and considerably faster. The increased speed permitted the generation of a database, the ESM Metagenomic Atlas, containing more than 600 million metagenomic proteins. — MAF

Abstract

Recent advances in machine learning have leveraged evolutionary information in multiple sequence alignments to predict protein structure. We demonstrate direct inference of full atomic-level protein structure from primary sequence using a large language model. As language models of protein sequences are scaled up to 15 billion parameters, an atomic-resolution picture of protein structure emerges in the learned representations. This results in an order-of-magnitude acceleration of high-resolution structure prediction, which enables large-scale structural characterization of metagenomic proteins. We apply this capability to construct the ESM Metagenomic Atlas by predicting structures for >617 million metagenomic protein sequences, including >225 million that are predicted with high confidence, which gives a view into the vast breadth and diversity of natural proteins.

https://www.science.org/doi/abs/10.1126/science.ade2574



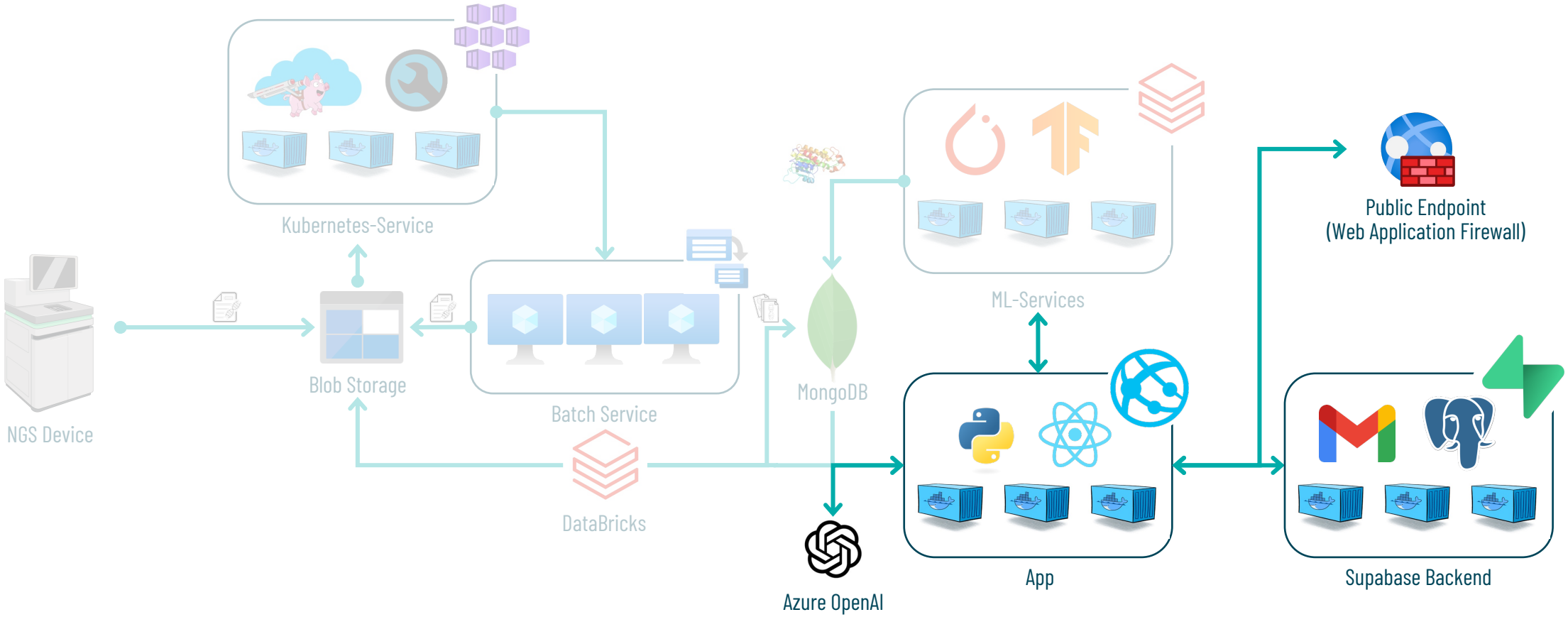
Note: Based on metrics such as perplexity ESM – Model performed worse on small aminoacid sequences (13.54) compared to larger aminoacid sequences (11.27).
Since here we are focussing on proteins consisting of aminoacid sequences with a length < 150 we fully retrained ESM – Model on 12 Mio. sequences with a length < 150 AS.
With that we were able to achieve better performance on small sequences with a perplexity of 12.2 for the ESM– Part of ESMFold.

scieneers
DRIVEN BY DATA

UNIKLINIK RWTH AACHEN
Institut für Humangenetik und Genommedizin

GenomAIx Workflow – Step 4

Genomaix App as an Interface for Users and Human-In-The-Loop

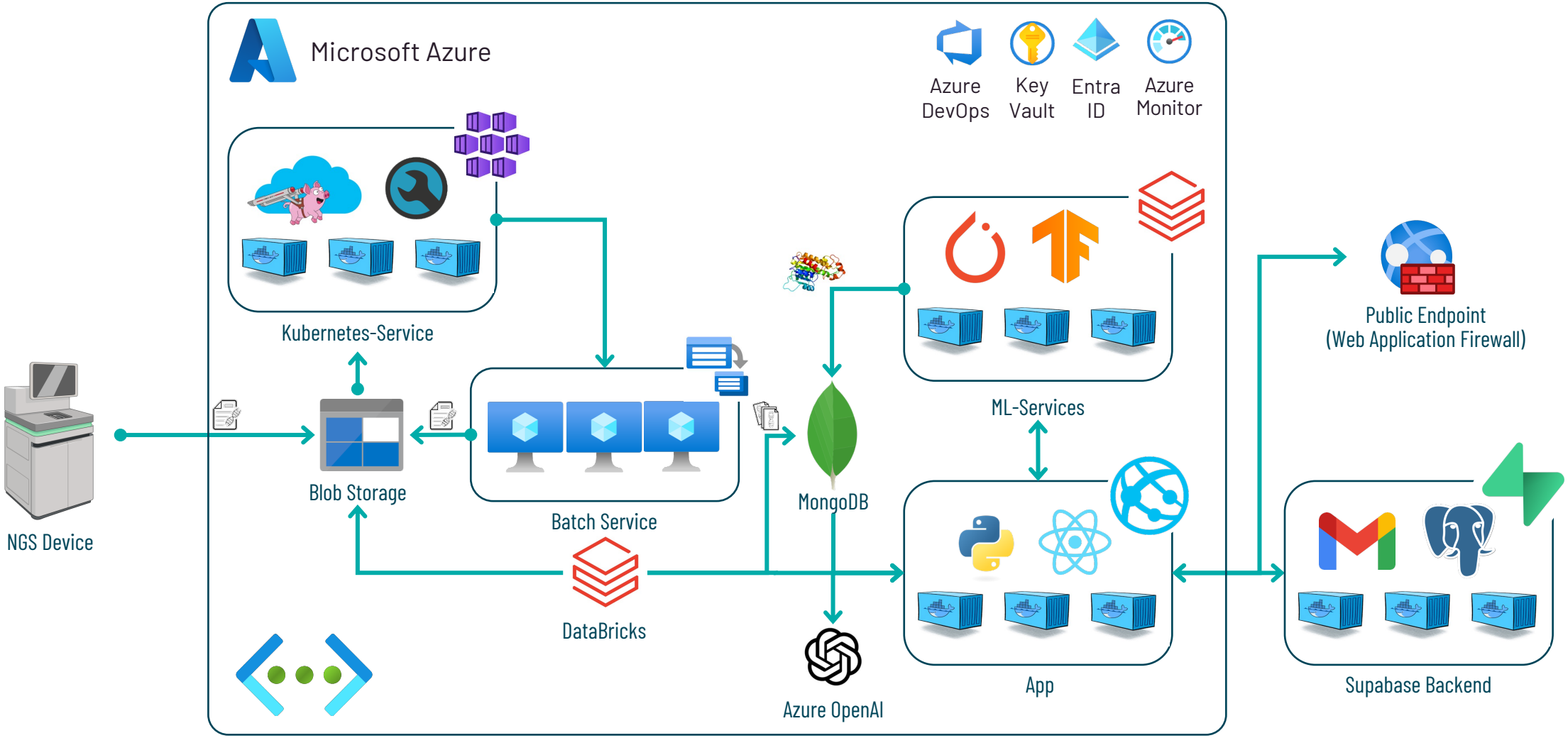


scieneers
DRIVEN BY DATA

UNIKLINIK RWTH AACHEN
Institut für Humangenetik und Genommedizin

GenomAIx Big Picture

Cloud-based Vnet secured Architecture



scieneers
DRIVEN BY DATA

UNIKLINIK RWTH AACHEN
Institut für Humangenetik
und Genommedizin

GenomAIx

What we achieved so far

scieneers
DRIVEN BY DATA

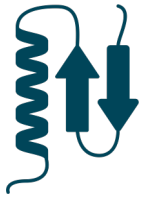
UNIKLINIK RWTH AACHEN
Institut für Humangenetik und Genommedizin

GenomAIx

What we achieved so far

Data of 377 families processed

Data of 431 affected individuals processed



47.500 Protein Structures of sORF Variants

6 high interest variants identified
1 is almost certain to be disease causing

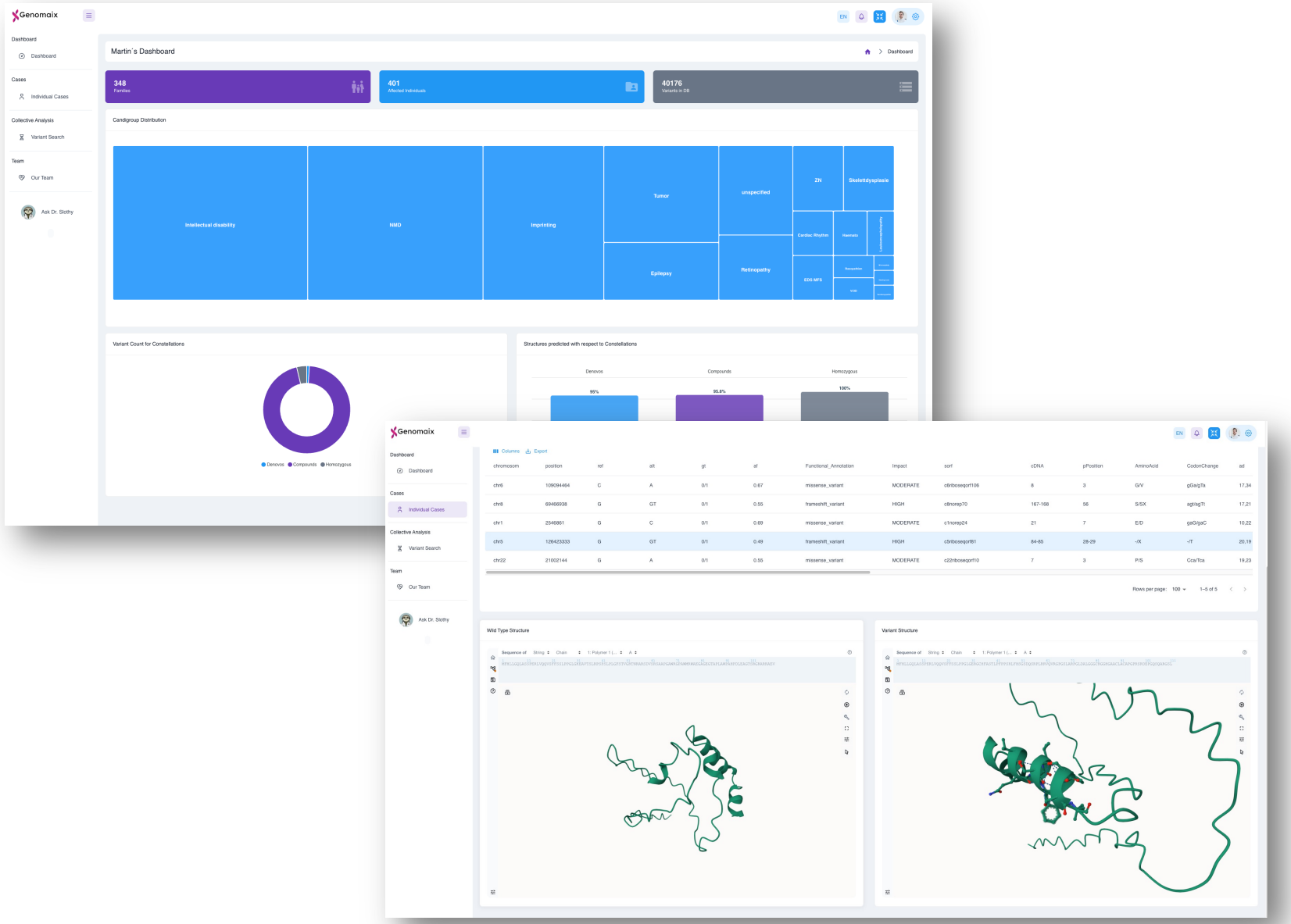
scieneers
DRIVEN BY DATA

UNIKLINIK RWTH AACHEN
Institut für Humangenetik
und Genommedizin

GenomAIx

von der Vision zur Applikation

(Research only)



Get in Contact!

Wir freuen uns über den Austausch.

Dr. Lars Perchalla
Direktor Data Science
scieneers GmbH
lars.perchalla@scieneers.de
Mobil +49 151 551 52 553

Martin Danner
Data Scientist / PhD
scieneers GmbH / Uniklinik RWTH Aachen
martin.danner@scieneers.de
Mobil +49 151 551 52 568

Dr. Jeremias Krause
Assistenzarzt
Uniklinik RWTH Aachen
jkrause@ukaachen.de
Tel.: + 49 241 8087012

Prof. Dr. Ingo Kurth
Direktor Institut
Uniklinik RWTH Aachen
ikurth@ukaachen.de
Tel.: +49 241 8080178

Prof. Dr. Miriam Elbracht
Leitung Klinische Genomik
Uniklinik RWTH Aachen
mielbracht@ukaachen.de
Tel.: +49 241 8088013

Dr. Matthias Begemann
Leiter NGS Diagnostikzentrum
Uniklinik RWTH Aachen
mbegemann@ukaachen.de
Tel.: + 49 241 8080036

More about us on www.scieneers.de & www.ukaachen.de

Diskussion