health2030 genome center

Data Analysis and Interpretation at the Health 2030 Genome Center: Now and Beyond

health2030 genome center

Agenda

- Introduction
- DNA Sequencing Platform (DSP)
- Data and Governance Platform
- Analysis and Interpretation Platform
- Conclusions
- Q&A

health2030
genome center

HEALTH 2030 GENOME CENTER

Emmanouil Dermitzakis

Director

Health 2030 Genome Center

health2030
genome center

Multi-institutional hub to promote genomic medicine

HUG Hôpitaux Universitaires Genève

EPFL

CHUV

health2030 genome center

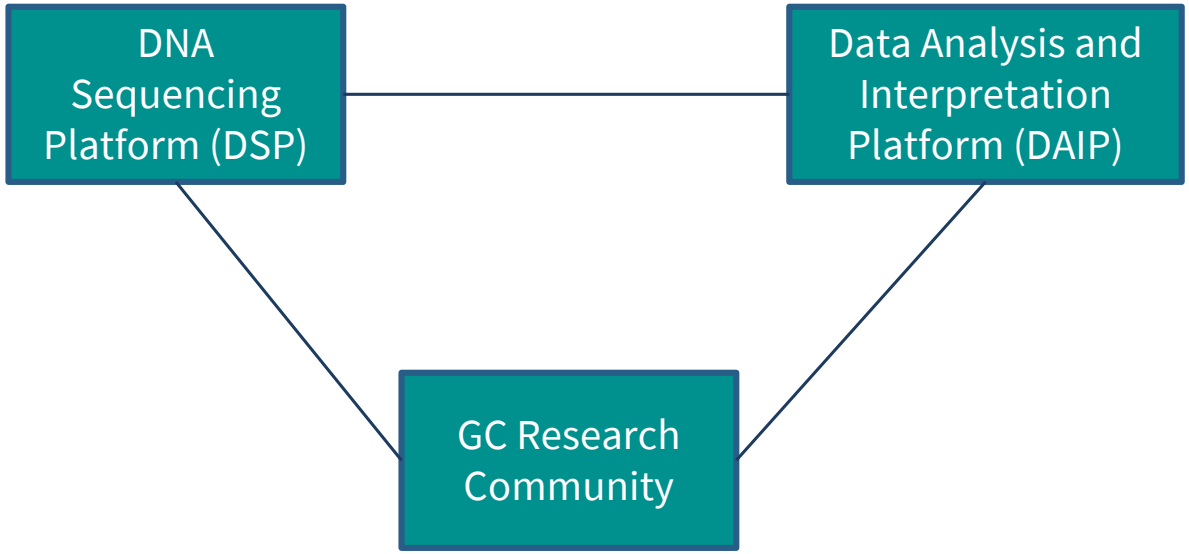
UNIVERSITÉ DE GENÈVE

UNIL | Université de Lausanne

UNIVERSITÄT BERN

INSELSPITAL
UNIVERSITÄTSSPITAL BERN
HOPITAL UNIVERSITAIRE DE BERNE
BERN UNIVERSITY HOSPITAL

Health 2030 Genome Center Multiple Pillars



health2030 genome center

Health 2030 Genome Center DNA Sequencing Platform

Keith Harshman
Chief Operations Officer

health2030 genome center

Health 2030 Genome Center DNA Sequencing Platform (DSP)

Mission:

- Provide DNA sequencing services in support of clinical diagnostics as well as large-scale genetics and genomics research

ISO 15189 Accreditation

- ISO 15189 accreditation of the DNA Sequencing Platform – provides a laboratory environment certified to produce clinical-grade sequencing data
- **All sequencing data** produced by the Platform using the accredited assays will be **clinical-grade**, regardless of whether the sample originates in a diagnostic or research setting

Sequencing services for diagnostic and research applications

Current sequencing services:

- Whole Genome Sequencing (WGS)
- Whole Exome Sequencing (WES)
- RNA-seq
- Sequencing user-prepared libraries
- SARS-CoV-2 WGS

Other application types based on interest and demand, i.e.:

- Small sample quantity RNA-seq
- Microbiome

Sequencing services for diagnostic and research applications

Current sequencing services:

- Whole Genome Sequencing (WGS)
- Whole Exome Sequencing (WES)
- RNA-seq
- Sequencing user-prepared libraries
- SARS-CoV-2 WGS

>> www.health2030genome.ch

Other application types based on interest and demand, i.e.:

- Small sample quantity RNA-seq
- Microbiome

Sequencing services for diagnostic and research applications

Current sequencing services:

- Whole Genome Sequencing (WGS)
- Whole Exome Sequencing (WES)
- RNA-seq
- Sequencing user-prepared libraries
- SARS-CoV-2 WGS

ISO 15189 accredited

Other application types based on interest and demand, i.e.:

- Small sample quantity RNA-seq
- Microbiome

health2030 genome center

Current sequencing infrastructure

- NovaSeq 6000
- 2x HiSeq 4000

- Evaluation of new sequencing technologies as well as complementary genomic technologies to see if/how they can contribute to Genome Center mission

Current sequencing infrastructure

Laboratory Infrastructure

» Wide use of automation for sample quality assessment, library preparation and library quality control provides sufficient capacity to support yearly processing thousands of genomes, exomes and transcriptomes

IT Infrastructure

» High capacity and secure data storage and analysis infrastructure in place

DNA Sequencing Platform staff

- Anthony Blin
- Melyssa Elies
- Henri Pegeot
- Deborah Penet
- Jérôme Thomas
- Elena Torres (Genome Center Administrative Coordinator)

health2030 genome center

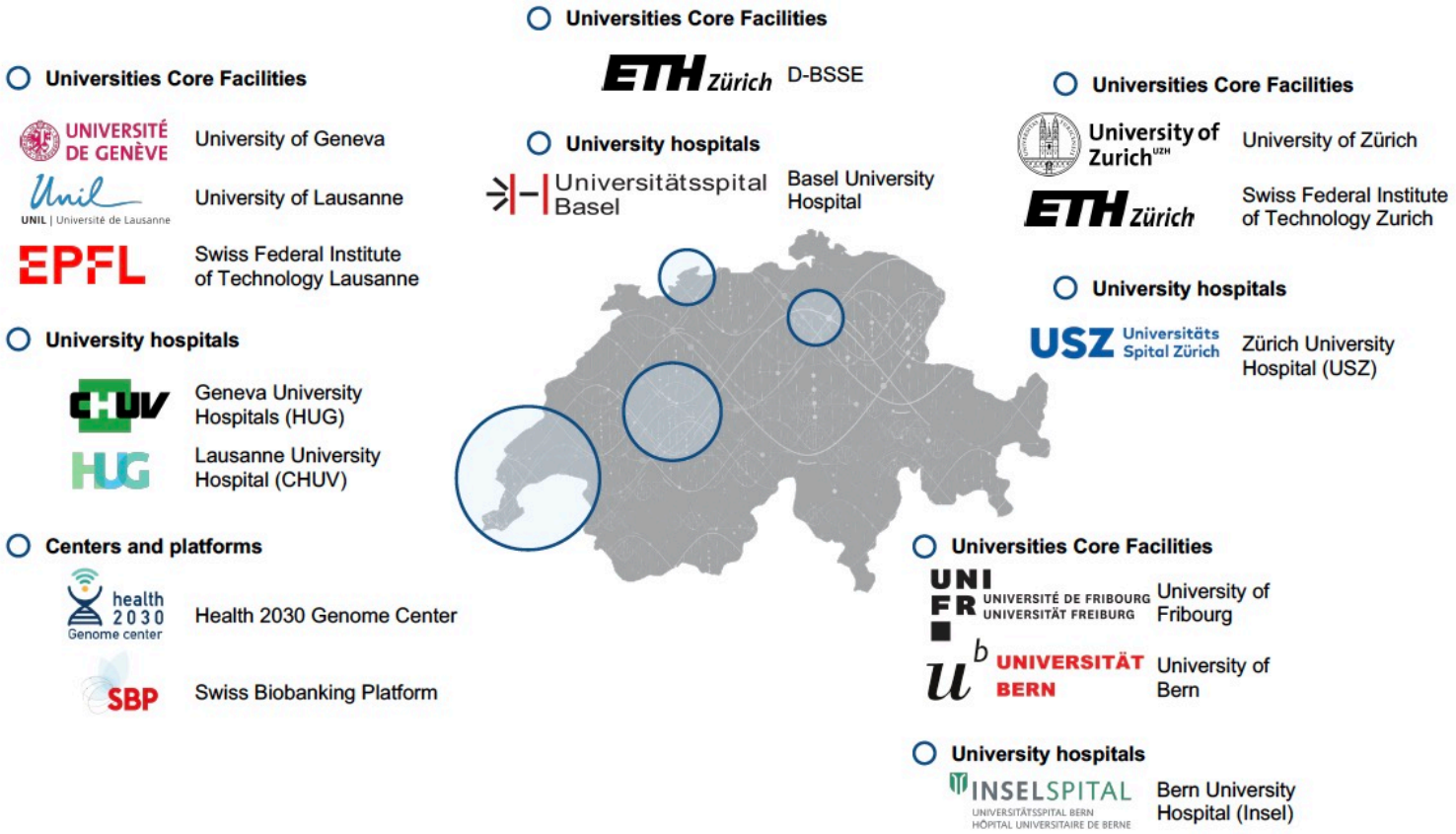
HEALTH 2030 GENOME CENTER
Data and Governance Platform

Katrin Männik

Head of Data Strategy

health2030
genome center

The Swiss genomics landscape



Slide courtesy of Jan Armida & Katrin Crameri, PHI SIB

Multi-institutional hub to promote genomic medicine

HUG Hôpitaux Universitaires Genève

EPFL

CHUV

health2030 genome center

UNIVERSITÉ DE GENÈVE

UNIL | Université de Lausanne

UNIVERSITÄT BERN

INSELSPITAL
UNIVERSITÄTSSPITAL BERN
HOPITAL UNIVERSITAIRE DE BERNE
BERN UNIVERSITY HOSPITAL

Building strategic partnerships in data analysis and governance

Strategic Focus Area
Personalized Health and Related Technologies

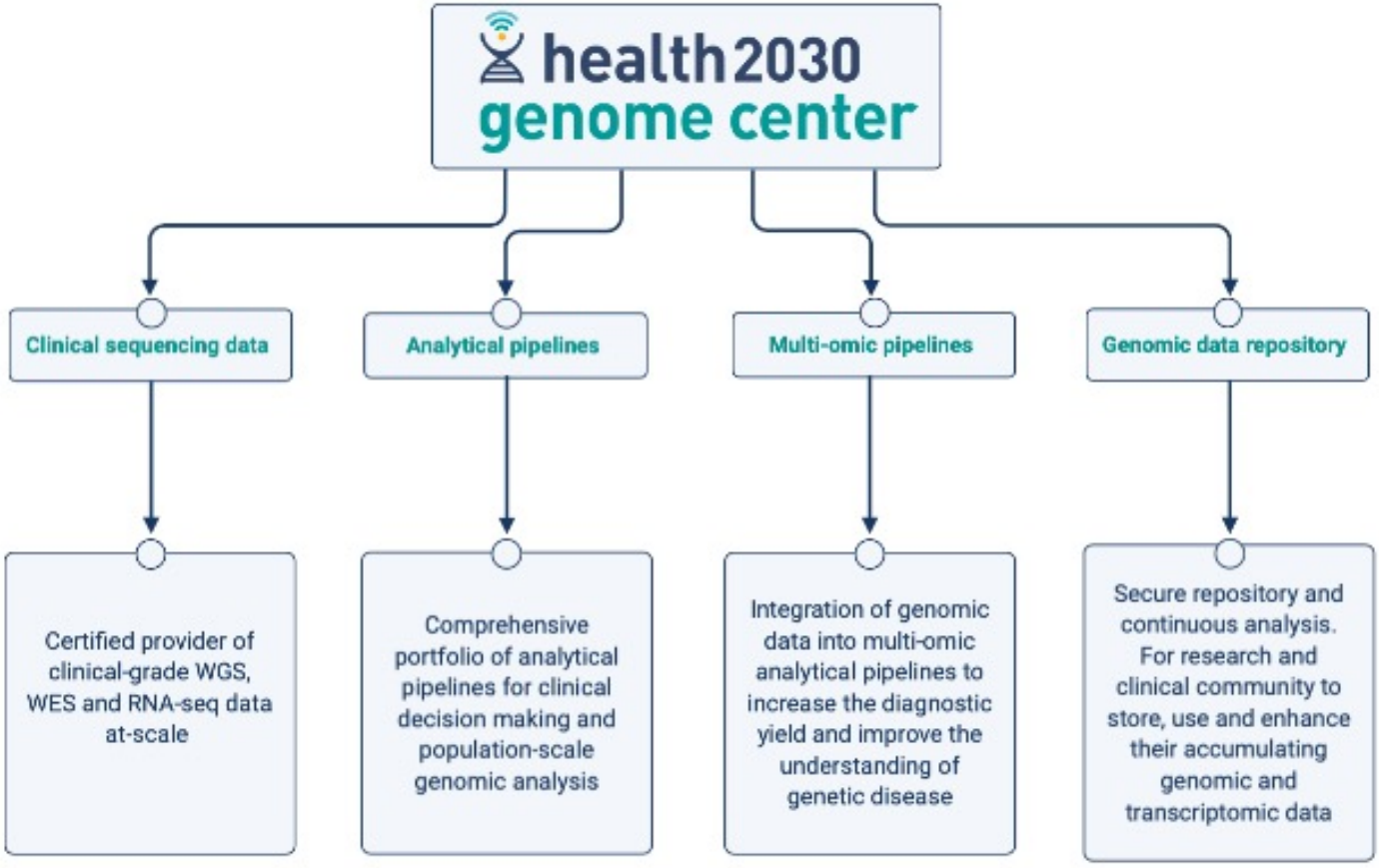
Swiss Personalized Health Network
BioMed IT

health2030 genome center

SDSC

SWISS BIOBANKING PLATFORM

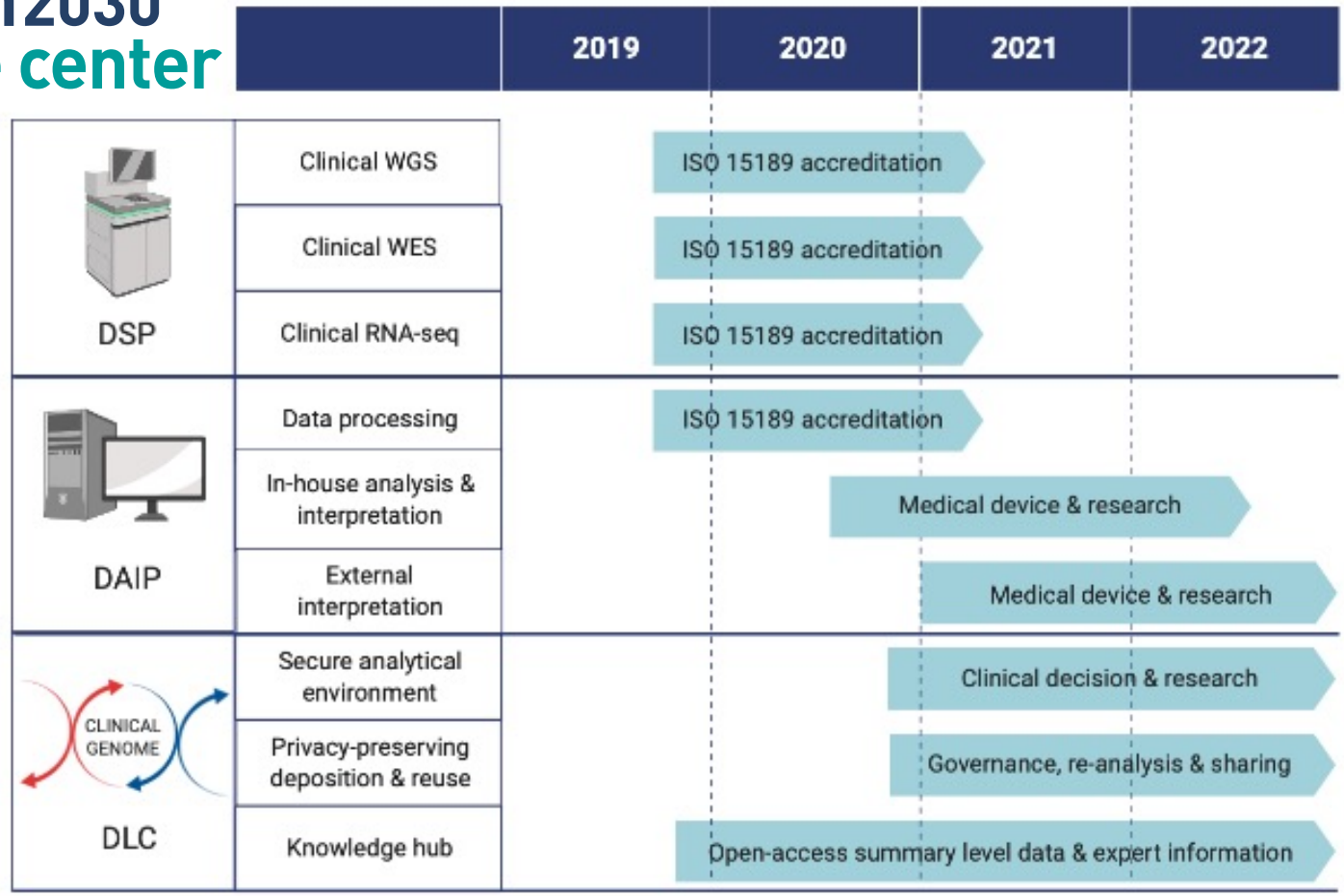
A hub for genomic medicine



Graphics created in BioRender.com

A hub for genomic medicine

health2030 genome center

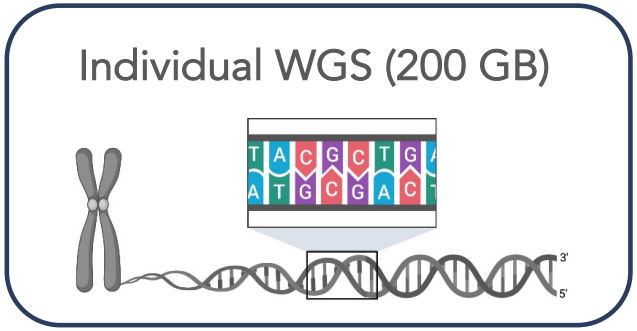


Graphics created in BioRender.com

Key gaps & needs in genomic data governance

Accumulating amounts of sequence data

Keeping data securely stored becomes heavy burden for most hospitals and many research groups



Increasing requirements to share (raw) data

Need for forward-looking solutions to govern and enhance data to increase reproducibility, transparency and benefits

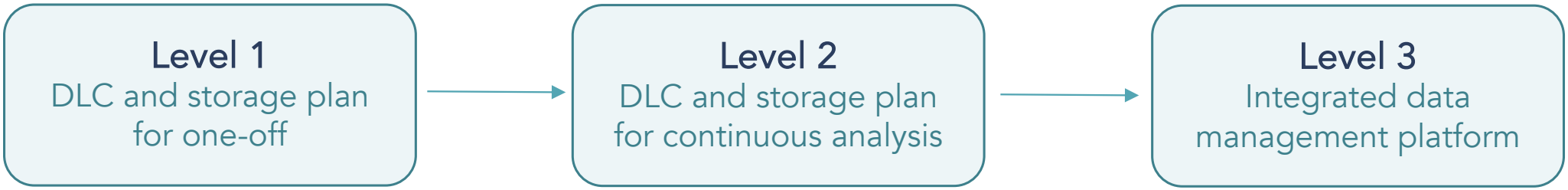
Need for iterative re-analysis

Systematic re-analysis of previously generated genomic data improves diagnostic yield and scientific knowledge

Poorly characterized isolated datasets

Representative cohorts with sufficient sample size are essential for reliability in genomic research and clinical decision making

Progressive genomic data life cycle (DLC) service



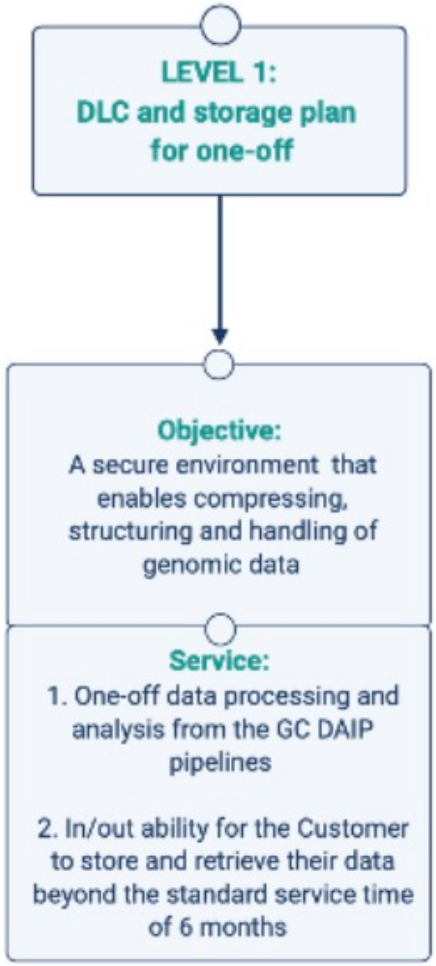
health2030 genome center

The Genome Center is an enabler for genomic medicine

- In all cases, the Customer is the owner and the controller of the provided biological samples and the produced data

- The Genome Center will not analyze, publish, share or otherwise use any generated and/or hosted data unless a specific task is mandated by the owner or the controller of the data

Genomic DLC and long-term storage



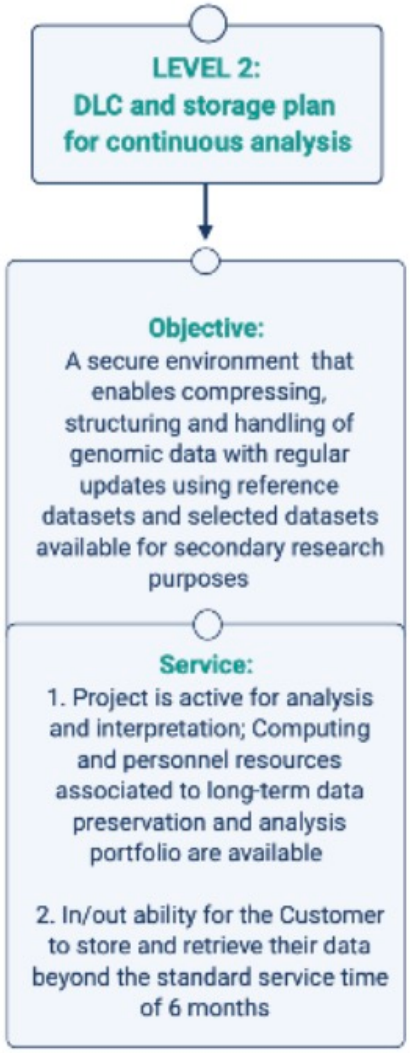
The Customer is the data owner and the controller

Community need and Readiness-level

- Confirmed high community need
- The Genome Center is **ready to offer**

Graphics created in BioRender.com

Genomic DLC and long-term storage



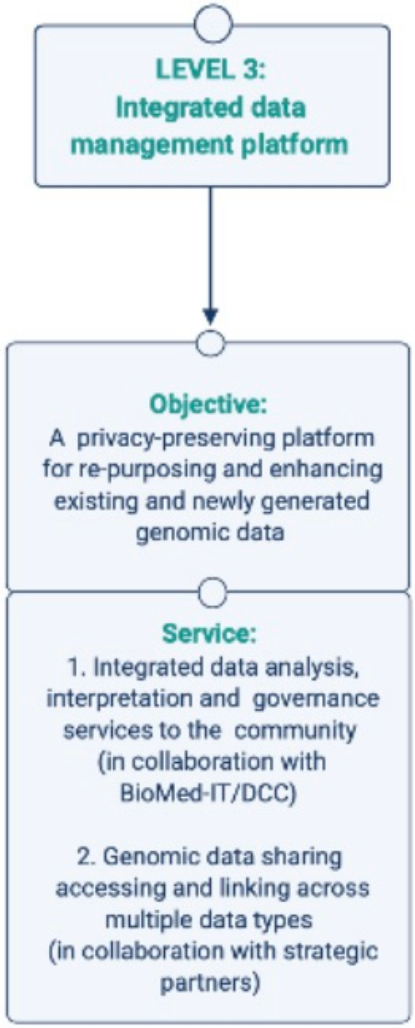
The Customer is the data owner and the controller

Community need and Readiness-level

- Expected high community need, investigation among customers in progress
- Preparation for service in progress, expected launch in Q3 2021

Graphics created in BioRender.com

Integrated platform for genomic data management



The Customer is the data owner and the controller

Community need and Readiness-level

- Community need to be explored, feedback from research and clinical community essential
- Forward-looking and longer-term direction
- The Genome Center has started preparatory work, in collaboration with strategic partners (incl. BioMed-IT)

Graphics created in BioRender.com

health2030 genome center

Data Analysis and Interpretation a few vignettes

health2030 genome center

Data Analysis and Interpretation Platform Staff

- Katrin Männik
- Arnaud Hungler
- Cedric Howald
- Lorenzo Cerutti
- Ilya Kolpakov
- Arkadiy Shevrikuko

health2030 genome center

Analysis and Interpretation at Scale

- Designing for medical diagnostic (CE-IVD)
- Automation of bioinformatic analysis (T2D) :SARS-CoV2 surveillance
- Providing expertises in downstream high throughput analysis and integrative multi-omics analytics

health2030 genome center

health2030 genome center

Ilya Kolpakov

From Software to Hardware-Accelerated Pipelines

health2030 genome center

Genome Production Pipeline
for WGS/WES/Rna-Seq

- Accredited for data processing
  - Essential for internal processes including QC in fastq delivery
- WGS/WES processing based on GATK best practices
  - Significant computation load for WGS samples
- A distributed system running on the internal cluster
  - Robust and highly integrated but not easily extensible
  - Noticeable maintenance/support burden

https://github.com/health2030genomecenter/genomecenter

health2030
genome center

Alternative pipelines :
Hardware-Accelerated

Motivation:

- *Reduction in time-to-delivery and support burden*
- Support for specialized pipelines (e.g. tumor, CNVs)
- Interpretation-readiness (e.g. variant annotations)

We evaluated:

- NVidia Parabricks (GPU-based)
- Illumina Dragen (FPGA-based)

health2030
genome center

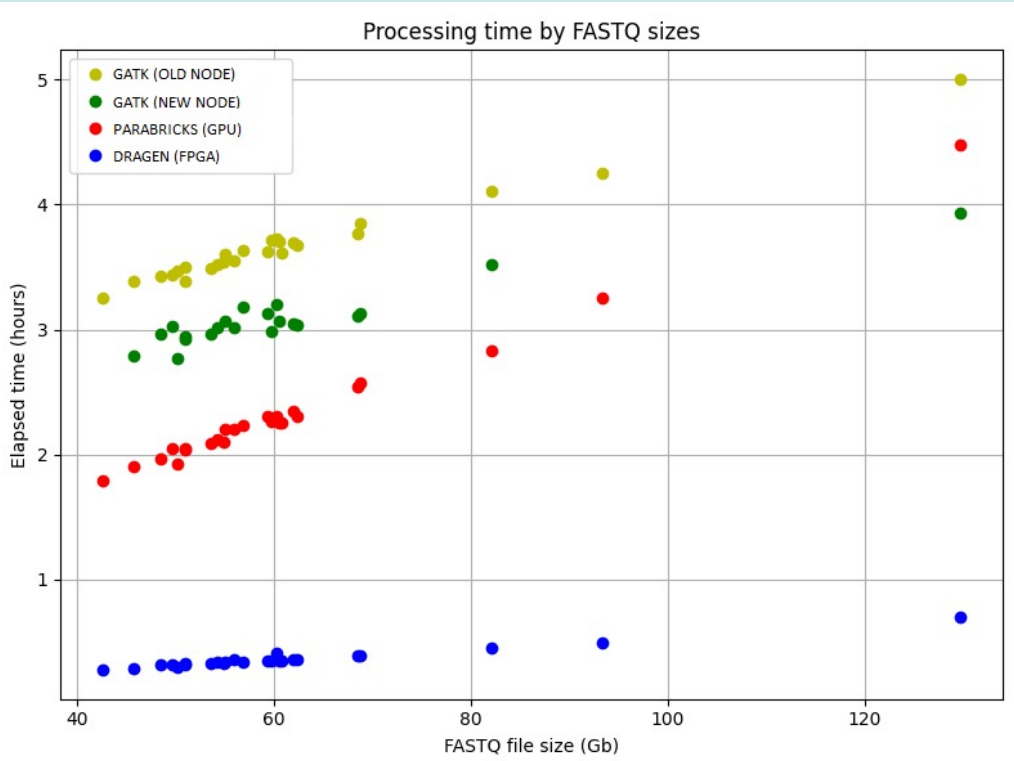
Evaluation: analysis Speed

Small variants on ~32x WGS (51Gb fq.gz)

- Production infrastructure (2016): 3h30
- New generation processor (2019): ~3h (1.11x faster)
- Parabricks (GPU): ~2h (2.5x faster)
- Dragen (FPGA): 0h20 (**15x faster**)

Dragen runtimes increase coverage

- 30x: slightly below 20min
- 90x: around 40 mins



health2030 genome center

Illumina Inc DRAGEN

NVIDIA Clara Parabricks

Accuracy: Small Variants (Dragen vs GATK)

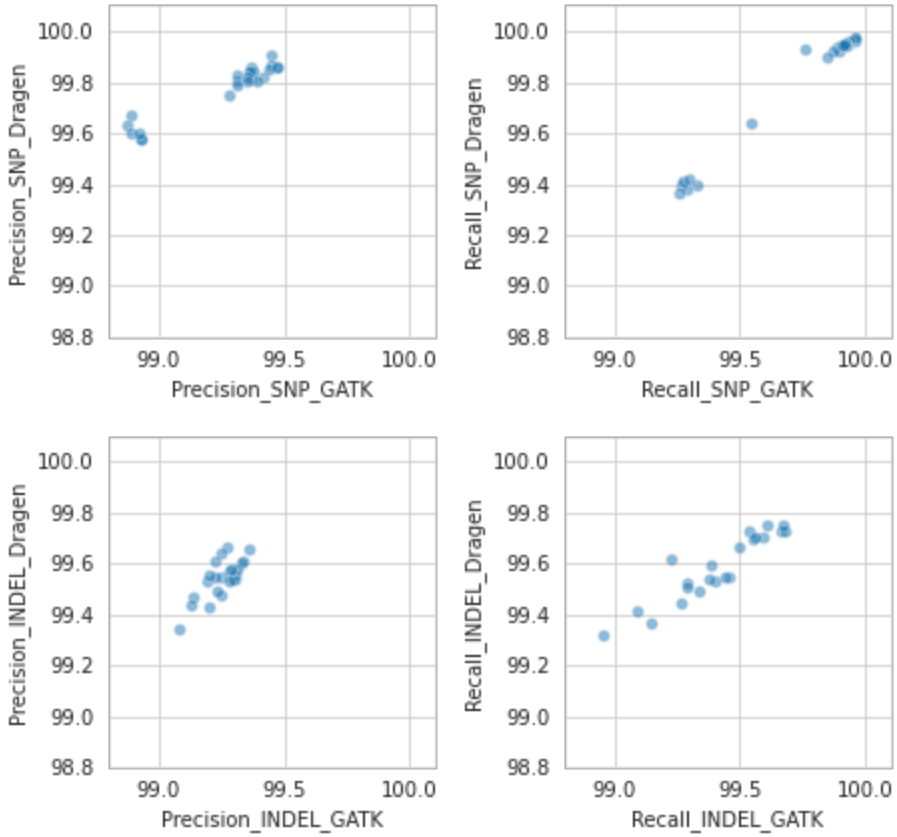
WGS of Genome-in-a-Bottle samples:

- NA12878
- NA24385, NA24149 and NA24143

Findings:

- Higher precision for SNVs
- Slightly higher recall of SNVs
- Higher precision and recall for INDELs

>99% in precision and accuracy



health2030 genome center

Dragen: additional Features

- CNV, SV, SNV including *trio-based* and *tumor-normal*
- Repeat expansions, scRNA, methylation, …
- Lossless Compression (recently added Enancio technology)
- Integration with clinical variant annotation tools

Currently evaluating trio-based capabilities on a set of families selected on CNV variants known to be pathogenic for ASD/ID in collaboration with the HUG.

health2030
genome center

Annotations Example: CNV (GIAB NA24385)

Called variant →

{"chromosome":"chr11", "position":55598706, "svEnd":55684863, "refAllele":"N", "altAlleles":["<DEL>"], "quality":52, "filters":["PASS"], "svLength":86157, "cytogeneticBand":"11q11", "samples":[{"genotype":"0/1", "copyNumber":1, "binCount":78}],

"clingen":[← Source of annotations below

Overlapping annotations

→ {"chromosome":"11", "begin":55265785, "end":56230226, "variantType":"copy_number_gain", "id":"nsv497480", "clinicalInterpretation":"benign", "phenotypes":["Developmental delay AND/OR other significant developmental or morphological phenotypes"], "observedGains":1, "reciprocalOverlap":0.08933, "annotationOverlap":0.08933}, ← % of overlap

→ {"chromosome":"11", "begin":55316535, "end":57539457, "variantType":"copy_number_gain", "id":"nsv932589", "clinicalInterpretation":"uncertain significance", "phenotypes":["Delayed gross motor development", "Delayed speech and language development", "Inguinal hernia", "Intellectual disability", "Muscular hypotonia", "Short stature", "Umbilical hernia"], "phenotypeIds":["HP:0000023", "HP:0000750", "HP:0001249", "HP:0001252", "HP:0001537", "HP:0002194", "HP:0004322", "MedGen:C0019322", "MedGen:C0349588", "MedGen:C1843367", "MedGen:CN000024", "MedGen:CN000706", "MedGen:CN001147", "MedGen:CN001989"], "observedGains":1, "validated":true, "reciprocalOverlap":0.03876, "annotationOverlap":0.03876},

→ {"chromosome":"11", "begin":55316591, "end":55855055, "variantType":"copy_number_gain", "id":"nsv916012", "clinicalInterpretation":"likely benign", "phenotypes":["Autism", "Incoordination"], "phenotypeIds":["HP:0000717", "HP:0002311", "MedGen:C1864113", "MedGen:CN000674"], "observedGains":1, "reciprocalOverlap":0.16, "annotationOverlap":0.16},

→ {"chromosome":"11", "begin":55319519, "end":56212930, "variantType":"copy_number_gain", "id":"nsv1067779", "clinicalInterpretation":"benign", "phenotypes":["Anemia", "Colitis", "Dehydration", "Feeding difficulties in infancy ….

health2030 genome center

Acknowledgements

- Illumina team (Shyamal Mehtalia, Tim Martins, Jennifer Mummery)
- Dell (Nathalie Sers & Anatol Pordes) and NVidia

- Arkadiy Shevrikuko (timing of production and test pipelines)
- Arnaud Hungler (deployment, integration and updates)

- Christelle Borel (for allowing benchmarking experiments)

health2030
genome center

health2030
genome center

Lorenzo Cerutti

Supporting the Swiss Surveillance of SARS-CoV2

led by Laurent Kaiser@HUG

health2030
genome center

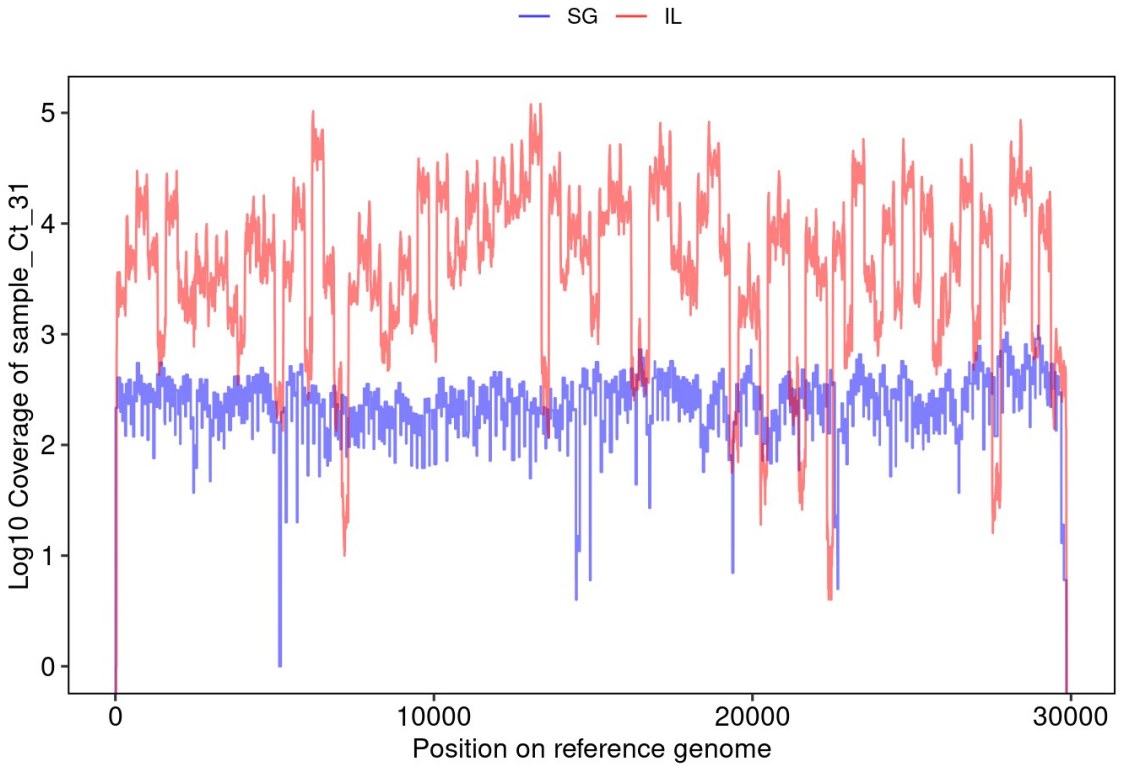
Timeline

- December 2020:
  - Genome Center was asked to provide a solution to sequence the full SARS-CoV-2 genome from large number of patient samples per week for viral surveillance in Switzerland.
  - Setting up sequencing and analysis procedures on a panel of samples from HUG and T. Stadler (Basel).
- Early January 2021:
  - First analysis of clinical samples.
  - Finalization of the sequencing and analysis pipeline.
- Since February 2021:
  - Routinely processed **752** samples per week.
  - The current sequencing and analysis workflow is ready for doubling the number of samples processed each week.
- More than **10’000** clinical samples have been sequenced at the GC since January 2021.

health2030
genome center

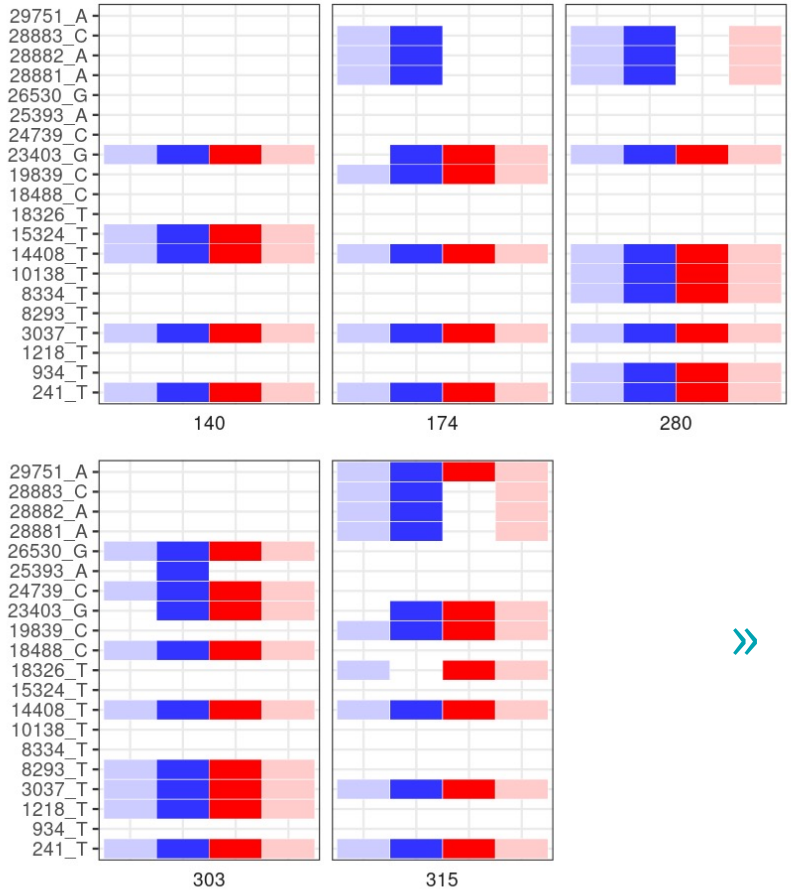
Comparing two sequencing solutions

- **Illumina COVIDSeq Test** and **SOPHiA GENETICS/Paragon Genomics CleanPlex SARS-COV-2**:
  - Illumina higher sequencing depth.
  - SOPHiA more robust at high Ct's .
- Easier scaling-up of Illumina:
  - Time cost
  - Cost



health2030 genome center

Variants survey



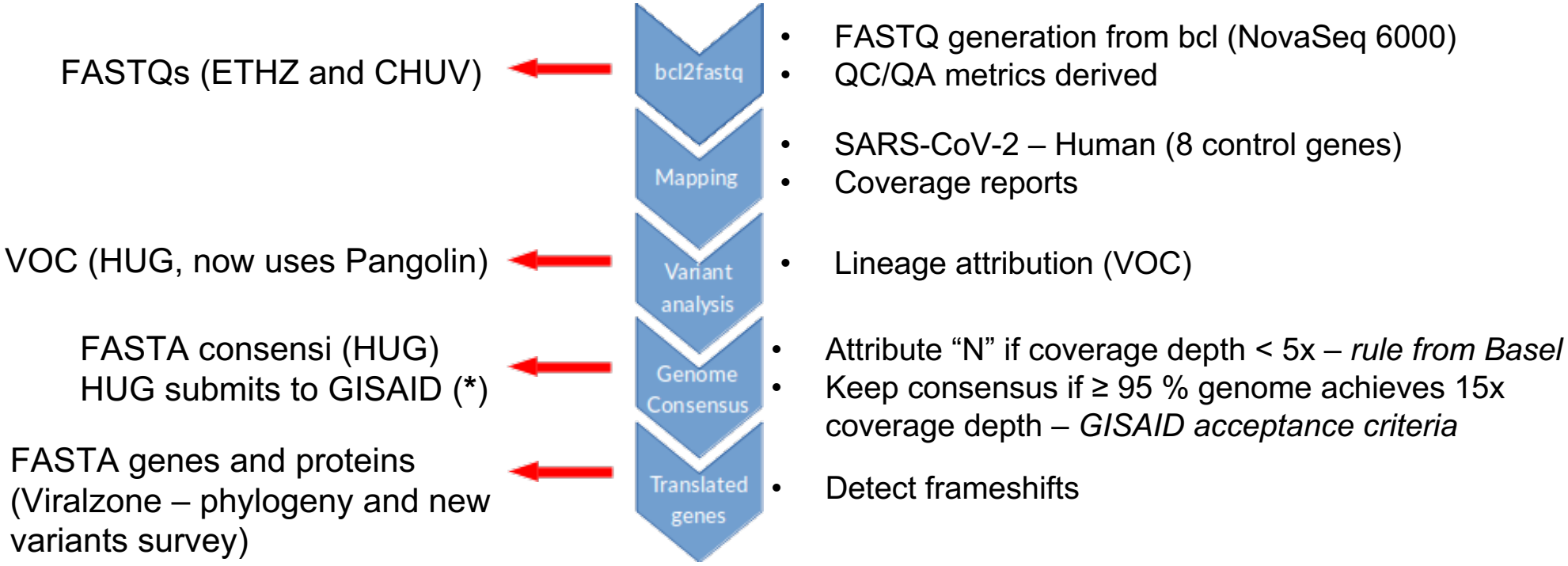
» Reads produced using Illumina together with our *in house* analysis pipeline reconciles the observed differences.

SARS-CoV-2 sequencing @GC

	Batch of 752 samples
Monday	
Tuesday (pm)	Samples reception and Preparation
Wednesday	RNA to cDNA
Thursday	Library preparation/quantification/pooling and overnight Sequencing
Friday (am)	Analytical processing and Data delivery

» Constraint: **max 12 days** between sample collection and GISAID submission. E.g. HUG collects samples on week 1. The batch is sequenced, and data delivered on week 2.

SARS-CoV-2 analysis pipeline @GC



~ 4h

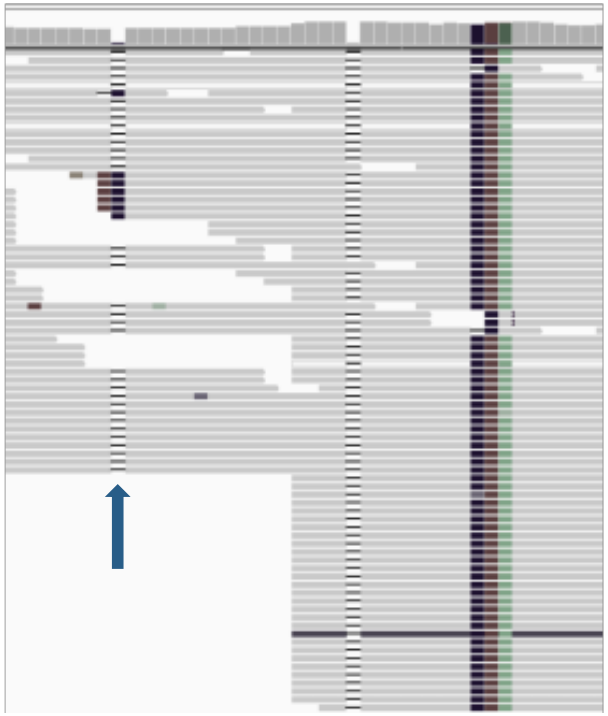
(*) Manual validation of consensi rejected by GISAID. Biocuration @GC.

health2030 genome center

https://github.com/health2030genomecenter/SARS-CoV-2_pipeline

Examples of biocuration @GC

1 del refpos 28254 → confirmed



22 del refpos 28221 → confirmed

2 del refpos 27486 → not confirmed

health2030
genome center

People

- All members of the DNA Sequencing Platform (DSP) and Data Analytics and Interpretation Platform (DAIP) at the H2030 Genome Center.
- Samuel Cordey (Kaiser's group, HUG).
- Philippe Le Mercier (Viralzone, SIB).
- Richard Neher (Univ Basel).
- Ivan Topolsky (Beerenwinkel's group, ETHZ).
- Chaoran Chen (Stadler's group, ETHZ).
- During the evaluation period Illumina and SOPHiA GENETICS.
- For the surveillance period Illumina (COVIDSeq)

health2030
genome center

Providing the right analytics and beyond

- Variant(s) annotation for clinical use (WES/WGS/RNASeq)
- Multiomics analysis method proteo-metabo-genomics
- Compressive genomics and privacy preserving distribution
- Complex analysis on-premise at the Genome Center

health2030
genome center

Variant annotation and clinical interpretation support
A multi prone approach

CHUV SIB

VarApp

- Deployed on-premises
- Variant interpretation

https://www.biorxiv.org/content/10.1101/060806v1

health2030 genome center

SAPHETOR varsomeclinical

- External partner
- Genome annotation and interpretation

Illumina Inc DRAGEN

- Diverse annotations filtering

Congenica
Enabling genomic medicine

- Automated Clinical Decision Support for genomic analysis

SOPHiA™

- External partner
- Genome annotation and interpretation

Kinder Spital : PI Matthias Baumgartner project

Strategic Focus Area
Personalized Health
and Related Technologies

- MMA – Methyl Malonic Acidurea
- Metabolic disease
- Disrupted amino acid metabolism

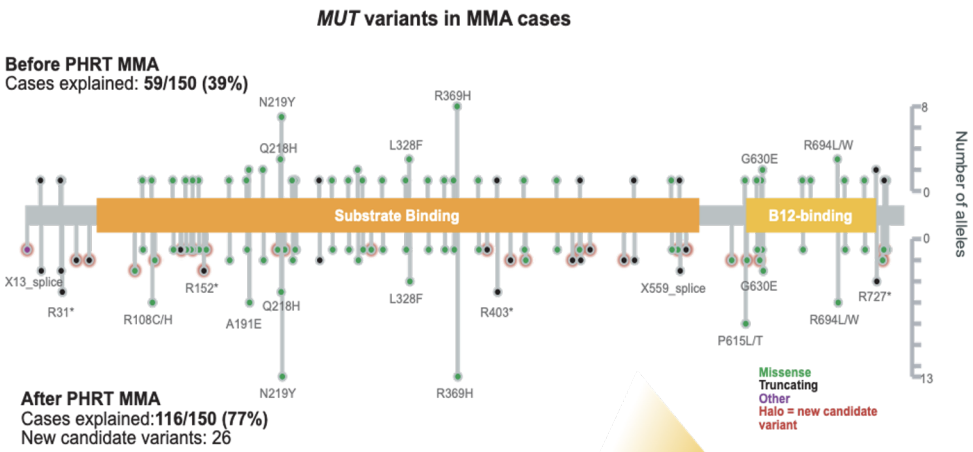
- **250** samples from around Europe and Switzerland

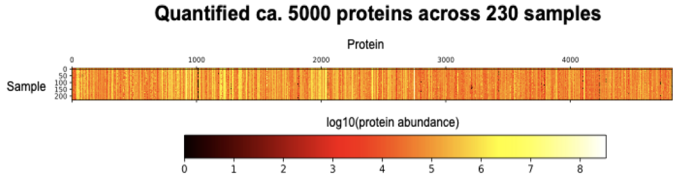
- Apply WGS and RNA-Seq to each sample

Time to deliver the data from beginning of Jan to end of Feb 2018

health2030
genome center

First pass analysis : 2-4 weeks after data being generated





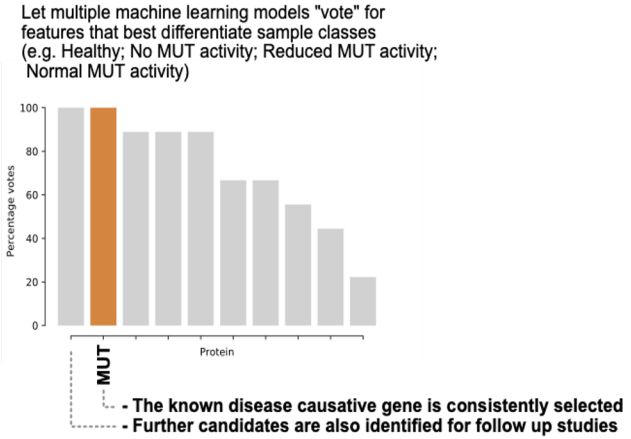
Genomics

Proteomics

Recapitulate existing knowledge and provide future leads to better explain disease mechanisms

Transcriptomics

eQTL analysis identified thousand of eGenes



David Lamparter (now @Roche) /Cedric Howald Health 2030
Ximena Bonilla (G. Rasch lab ETHZ)
Wenguang Shao-Patrick Pedrioli (CPAC ETHZ)
Patrick Forny-Sean Froese (KISPI)

health2030 genome center

Compressive Genomics

- Applying innovative techniques to genomic compression (enabled random access to the data)
  - Genomsys (MPEG-G ISO 23092)
  - Enancio (Illumina technology)

- Data footprint and random access
- Tested on 800 Alzheimer WGS (30x) : ADNI

ADNI
Alzheimer's Disease Neuroimaging Initiative

health2030 genome center

Thanks to

For an early access to the MPEG-G technology

Privacy preserving technology

MedCO

- Deployed on our premises (benchmarking in progress)
- For clinical applications: secure privacy-preserving exploring and sharing
- Queries to obtain and analyze summary data, variant-level data back to the clinician

Jean-Louis Raisaro

CHUV

https://medco.epfl.ch

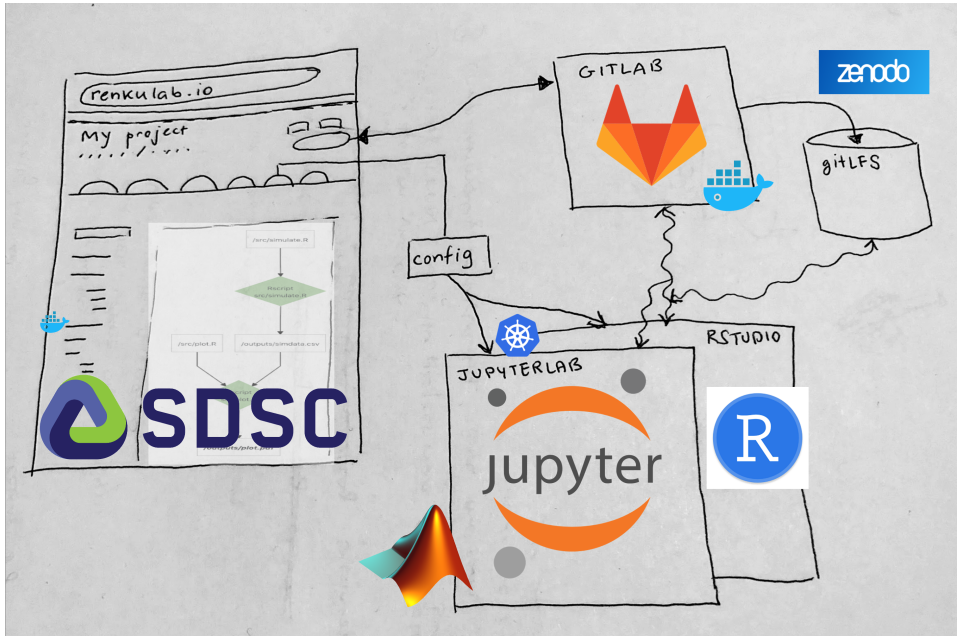
Strategic Focus Area
Personalized Health and Related Technologies

health2030 genome center

Providing the platform for on-premise analytic with the Genome Center staff

SDSC RENKU

- Deployed on our premises
- For research purposes: data analysis and reuse
- Applications in the area of continuous evaluation and assessment of prediction quality (e.g. GIAB)



https://renkulab.io

health2030 genome center

Implementing FPGA-based CNV/SV – Collaborating with HES-SO (Yverdon)

- Collaborate with expert in the field of FPGA
- Interface and teach PhD/Master student in the art of genomic analysis
- Develop innovative technology for CNV/SV

Contribute to the green deal

health2030 genome center

Rick Wertenbroek (PhD student)

Olivier Delaneau (DBC UNIL)

Yann Thoma (HES-SO Yverdon)

The Health 2030 Genome Center
aims to have a set of talents and expertise
and maintain/expand
these over time and be open to
clinical research and diagnostic groups in
Switzerland

Come to work virtually or physically with us

Email: Genome@health2030.ch

@swissgenomics

health2030
genome center

Q&A

health2030
genome center