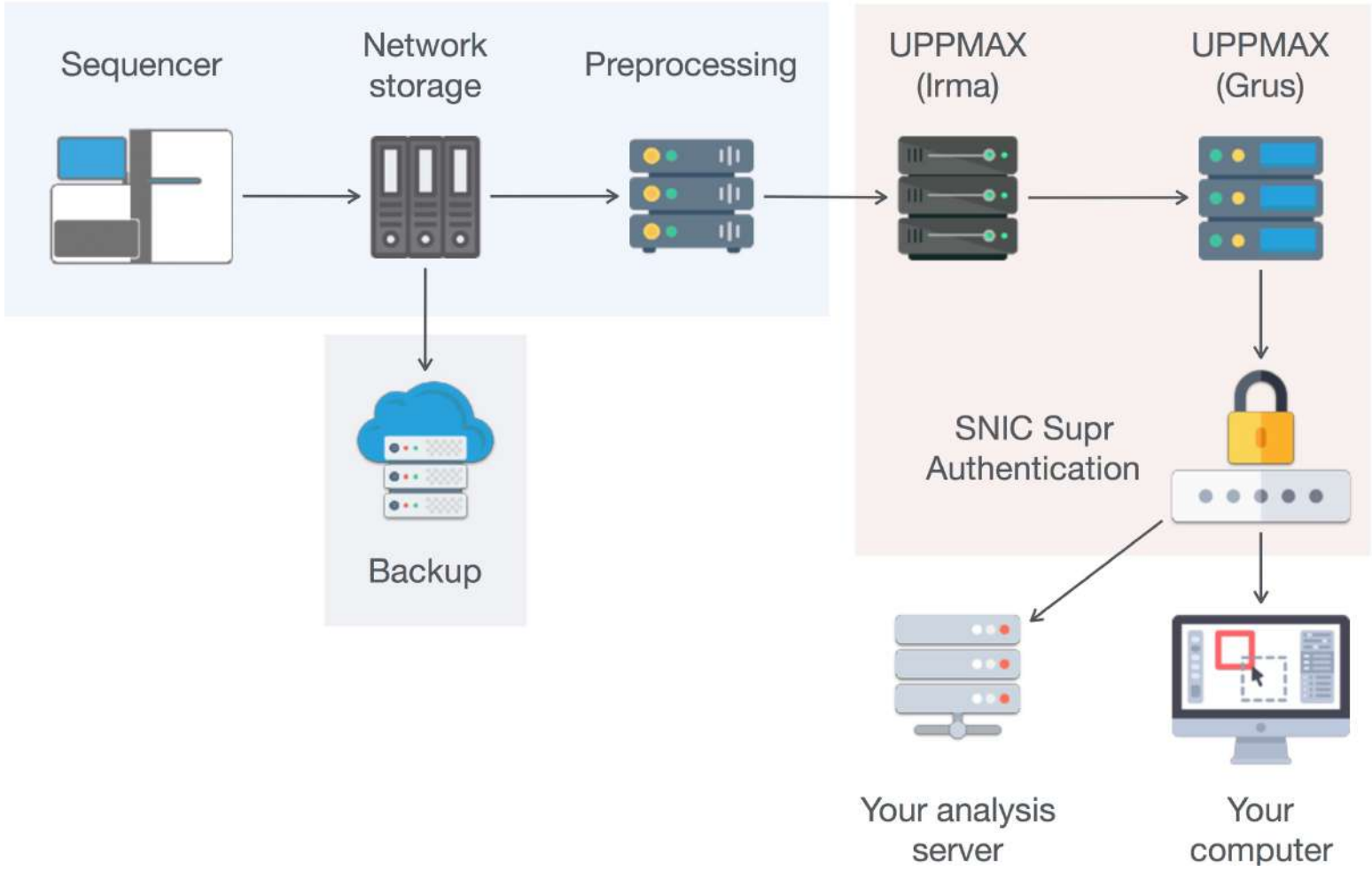
Next Generation Sequencing and Bioinformatics Analysis Pipelines

Adam Ameur
National Genomics Infrastructure
SciLifeLab Uppsala
adam.ameur@igp.uu.se

Today's lecture

- Data analysis and management at NGI/SciLifeLab
- Human whole genome sequencing
- The Earth Biogenome Project
- Other R&D activities at NGI

NGI Data Handling



Analysis pipelines

- Initial data analysis for major applications:
  - **Mapping:** Align sequences to a reference genome
  - **SNV calling:** Detect genetic variants
  - **RNA-seq:** Quantify gene expression
  - ***De novo* assembly:** Generate new reference genomes
  - **and more…**
- Analysis requirements: Automated, reliable, easy to run, reproducible

nf-core

- A community effort to collect a curated set of Nextflow analysis pipelines
- GitHub organisation to collect pipelines in one place
- No institute-specific branding
- Strict set of guideline requirements

nf-core

https://nf-co.re

nature biotechnology

Correspondence | Published: 13 February 2020

The nf-core framework for community-curated bioinformatics pipelines

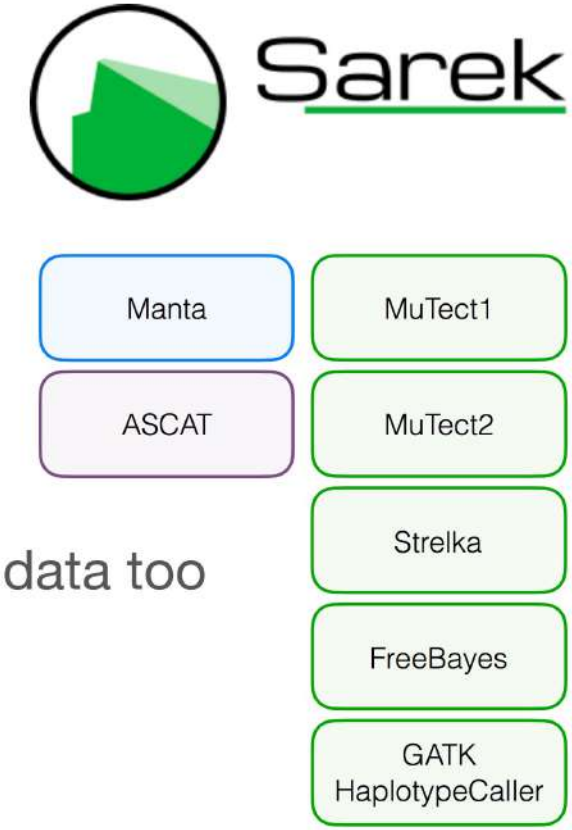
Philip A. Ewels, Alexander Peltzer, Sven Fillinger, Harshil Patel, Johannes Alneberg, Andreas Wilm, Maxime Ulysse Garcia, Paolo Di Tommaso & Sven Nahnsen

Phil Ewels, NGI Sthlm

Example pipeline - Sarek

GitHub https://github.com/SciLifeLab/Sarek

- Tumour/Normal pair WGS analysis based on GATK best practices
  - SNPs, SNVs and indels
  - Structural variants
  - Heterogeneity, ploidy and CNVs
- Works with regular WGS and Exome data too



NATIONAL GENOMICS INFRASTRUCTURE

NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

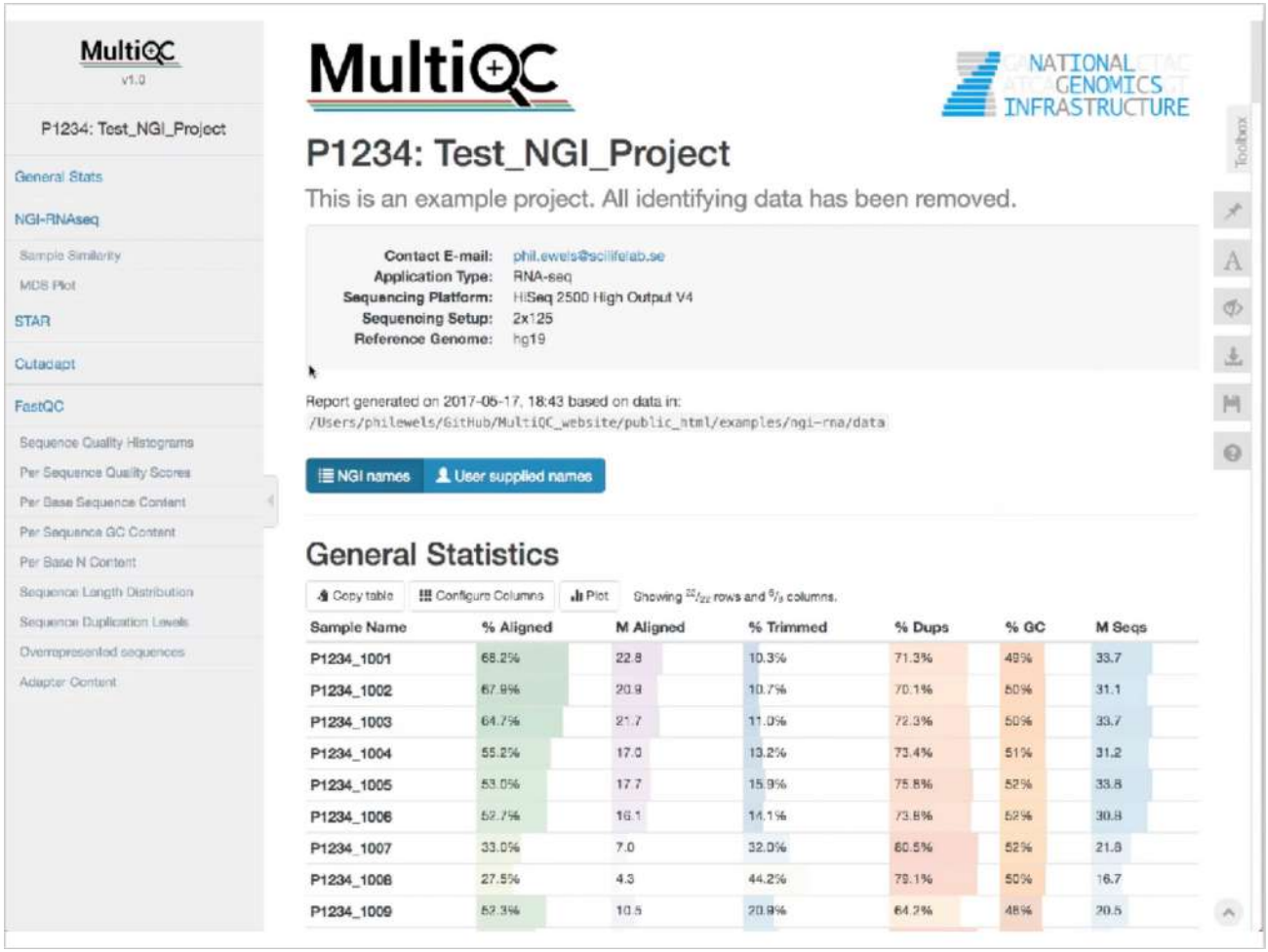
Barntumörbanken

Quality control

- Every project has some level of quality control checks
  - Technical run performance
  - Read length distribution
  - Sequencing quality
- Analysis pipelines give application-specific QC
- Reporting done using MultiQC (Illumina projects)

MultiQC

Multi QC example



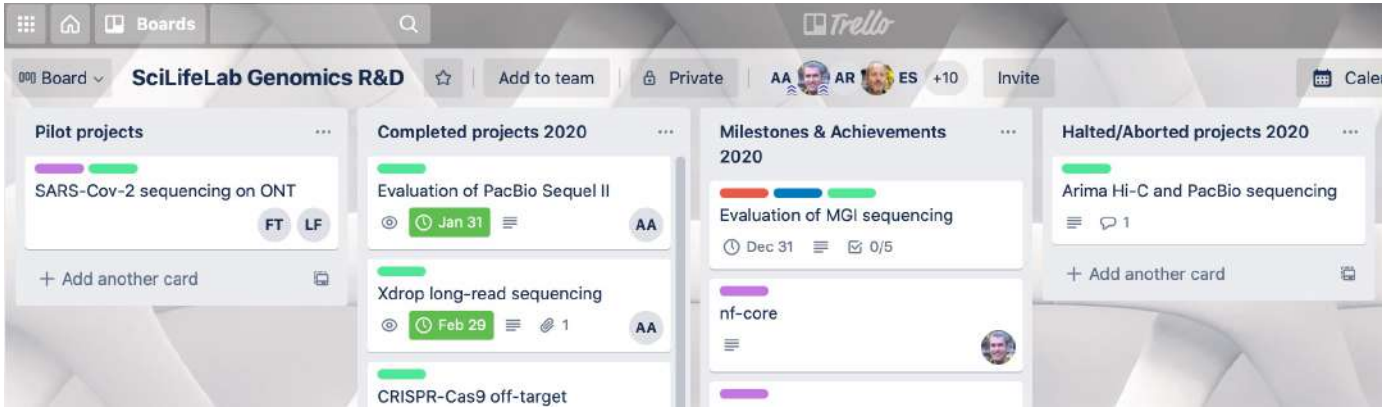
Data delivery via GRUS

- GRUS is an UPPMAX tool for NGI data deliv
  - NGI creates a SNIC Supr “delivery proje NGI sequencing project
  - Project PI and contact person is
  - Email sent with project ID ctions
- Grus is fo hort-term storage only!
  - wo-factor authentication

New delivery system to be launched in 2022!

NGI Research & Development projects

- For some projects, NGI allocates additional resources for development
  - New applications where we see the need to develop a pipeline
  - Construction of reference datasets and resources
  - Strategic collaborative projects



Example: The SweGen project

- A whole-genome resource for researchers and clinical labs

From SweGen release party on Oct 19th 2016

SweGen: 1000 Swedish Whole Genomes

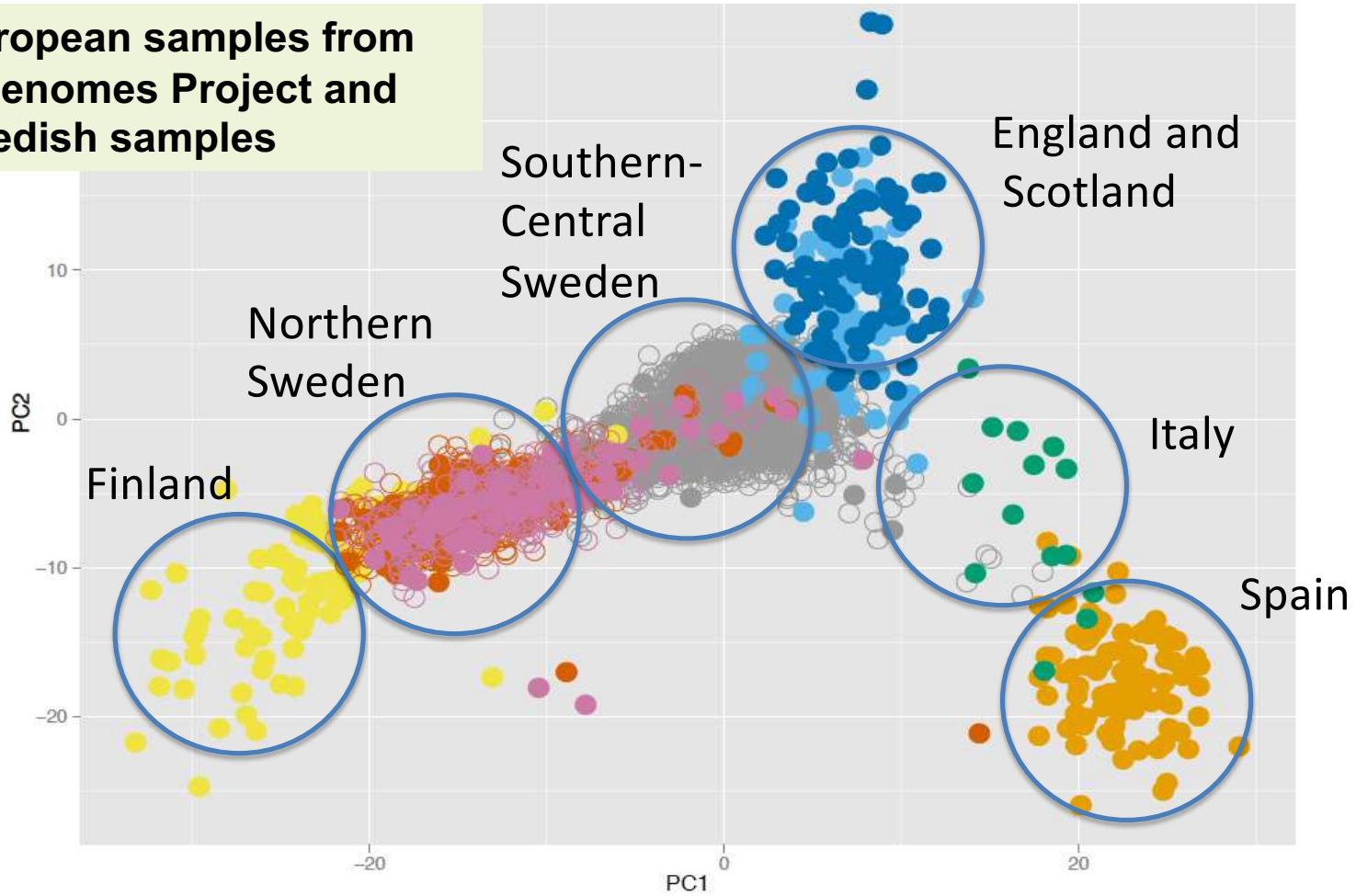
- What can the SweGen dataset be used for?
  - Look up genetic variant frequencies
  - Use as matched controls
  - Study population genetics
  - Study human evolutionary history

High demand for the data from many different groups:

➔ Make the data available as **quickly** and **openly** as possible!

Selecting 1000 individuals based on PCA

PCA of European samples from the 1000 Genomes Project and 10,000 Swedish samples



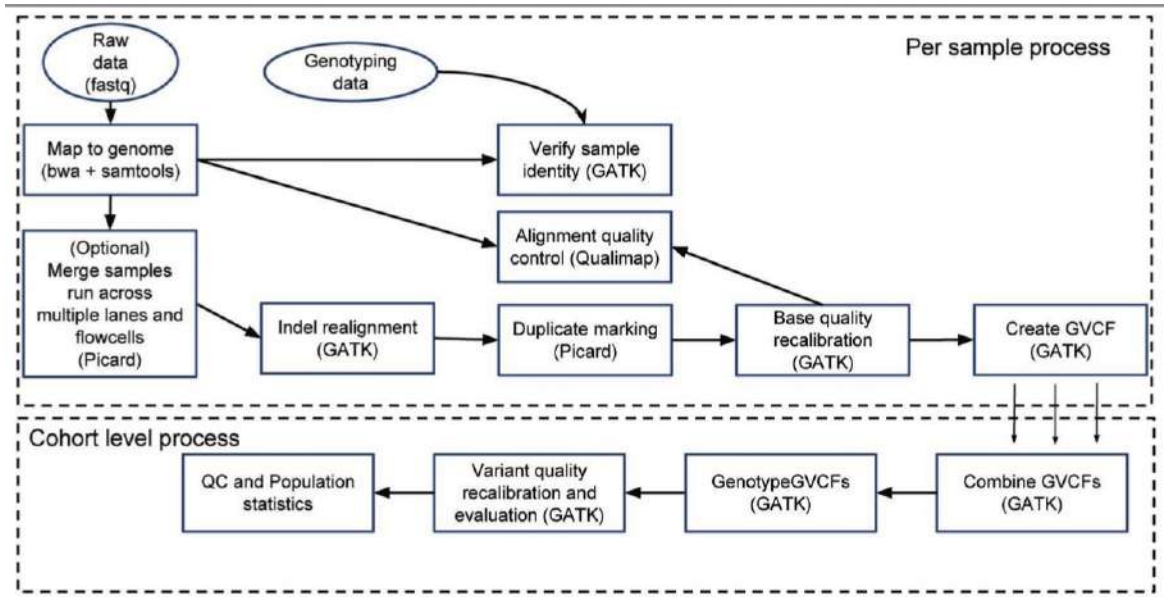
Whole Genome Sequencing

- 30X Illumina WGS generated for all 1,000 individuals

- Sequencing done both at NGI Sthlm and NGI Uppsala
- All 1,000 samples completed in September 2016

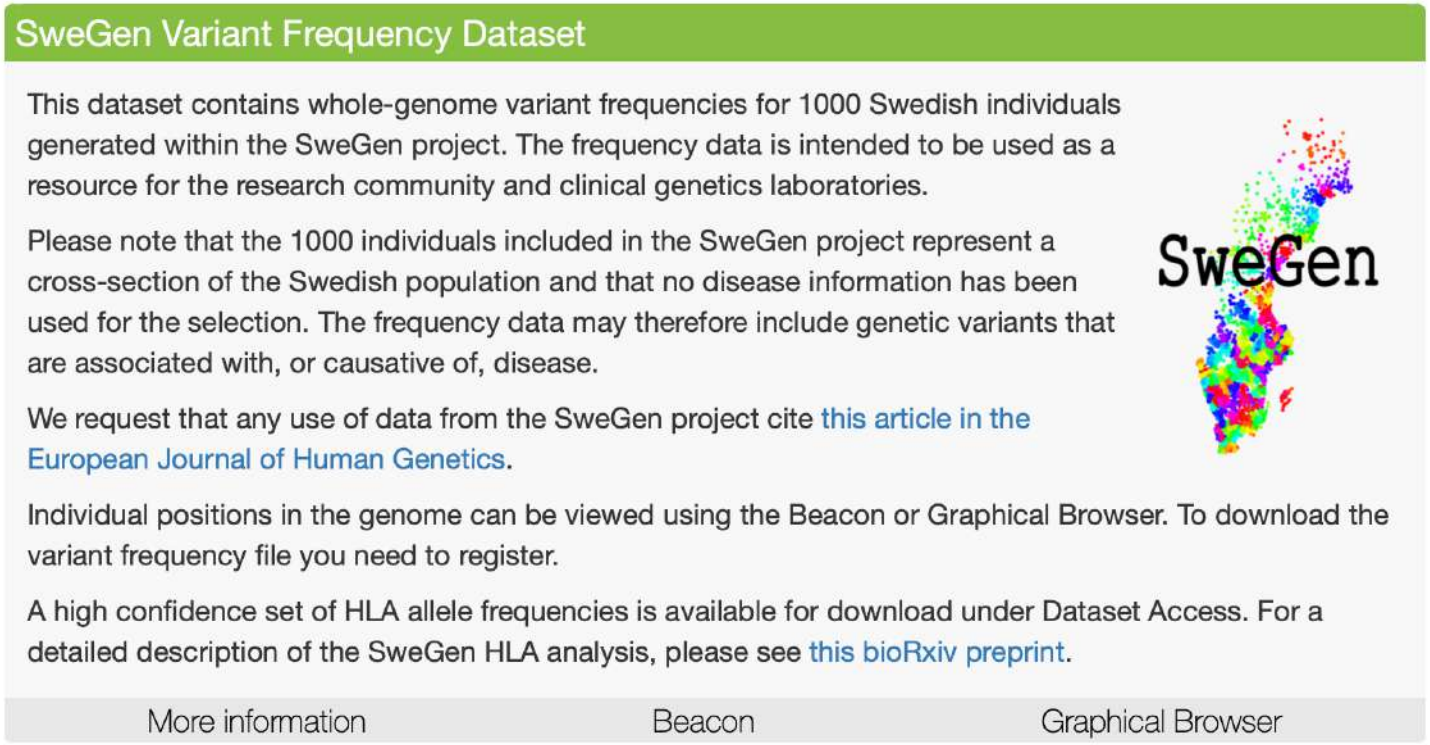
Data analysis pipeline

- NGI pipeline developed for mapping and variant calling



- About 100Gb data generated, and 2 million CPU hours used…
- This pipeline has become standard for all WGS projects at NGI

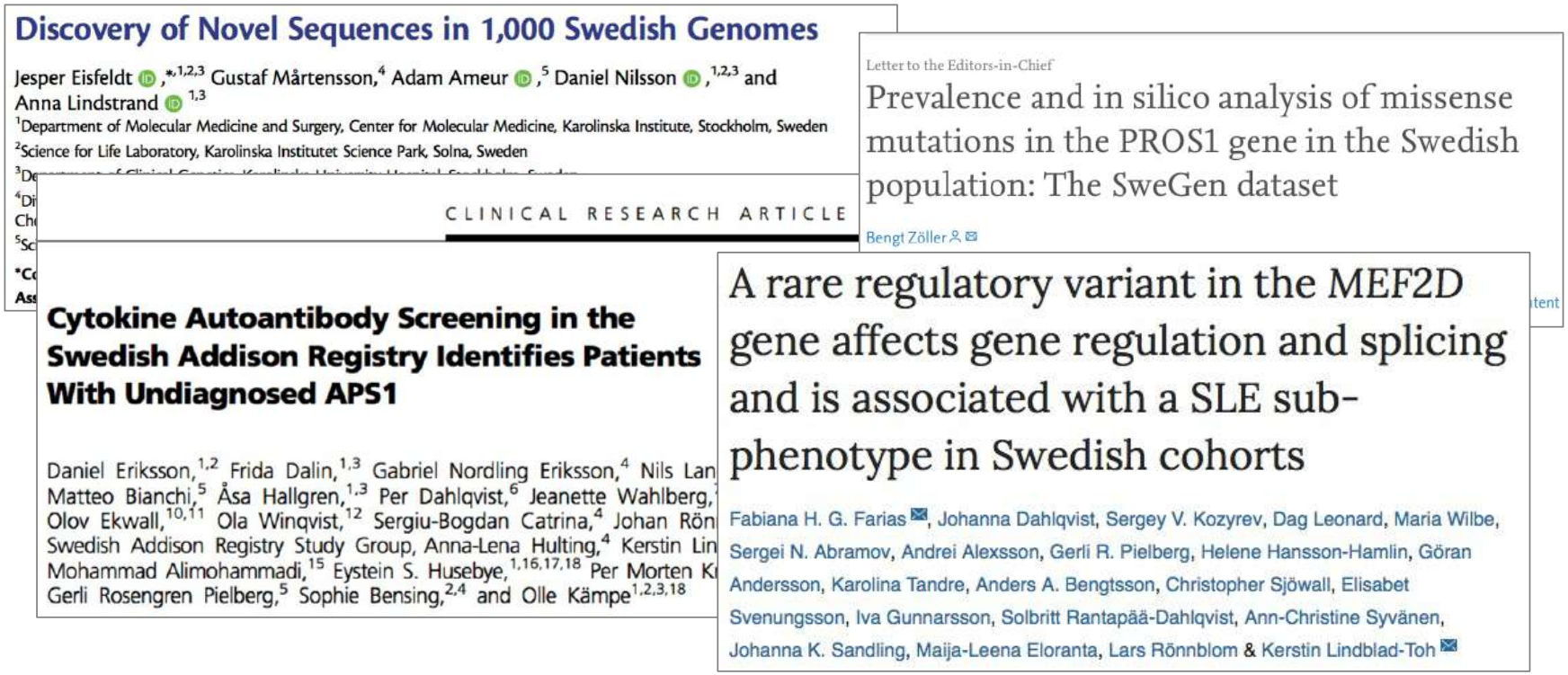
Making data available



- Aggregated frequencies available from: ***swefreq.nbis.se***
- Possible to access individual genotype data through Uppmax/Bianca

SweGen: a resource for collaboration

- Over 100 publications have made use of the SweGen dataset



Discovery of Novel Sequences in 1,000 Swedish Genomes

Jesper Eisfeldt [1,2,3] Gustaf Mårtensson,[4] Adam Ameur [5] Daniel Nilsson [1,2,3] and Anna Lindstrand [1,3]

[1]Department of Molecular Medicine and Surgery, Center for Molecular Medicine, Karolinska Institute, Stockholm, Sweden
[2]Science for Life Laboratory, Karolinska Institutet Science Park, Solna, Sweden

CLINICAL RESEARCH ARTICLE

Letter to the Editors-in-Chief

Prevalence and in silico analysis of missense mutations in the PROS1 gene in the Swedish population: The SweGen dataset

Bengt Zöller

Cytokine Autoantibody Screening in the Swedish Addison Registry Identifies Patients With Undiagnosed APS1

Daniel Eriksson,[1,2] Frida Dalin,[1,3] Gabriel Nordling Eriksson,[4] Nils Lan
Matteo Bianchi,[5] Åsa Hallgren,[1,3] Per Dahlqvist,[6] Jeanette Wahlberg,
Olov Ekwall,[10,11] Ola Winqvist,[12] Sergiu-Bogdan Catrina,[4] Johan Rön
Swedish Addison Registry Study Group, Anna-Lena Hulting,[4] Kerstin Lin
Mohammad Alimohammadi,[15] Eystein S. Husebye,[1,16,17,18] Per Morten K
Gerli Rosengren Pielberg,[5] Sophie Bensing,[2,4] and Olle Kämpe[1,2,3,18]

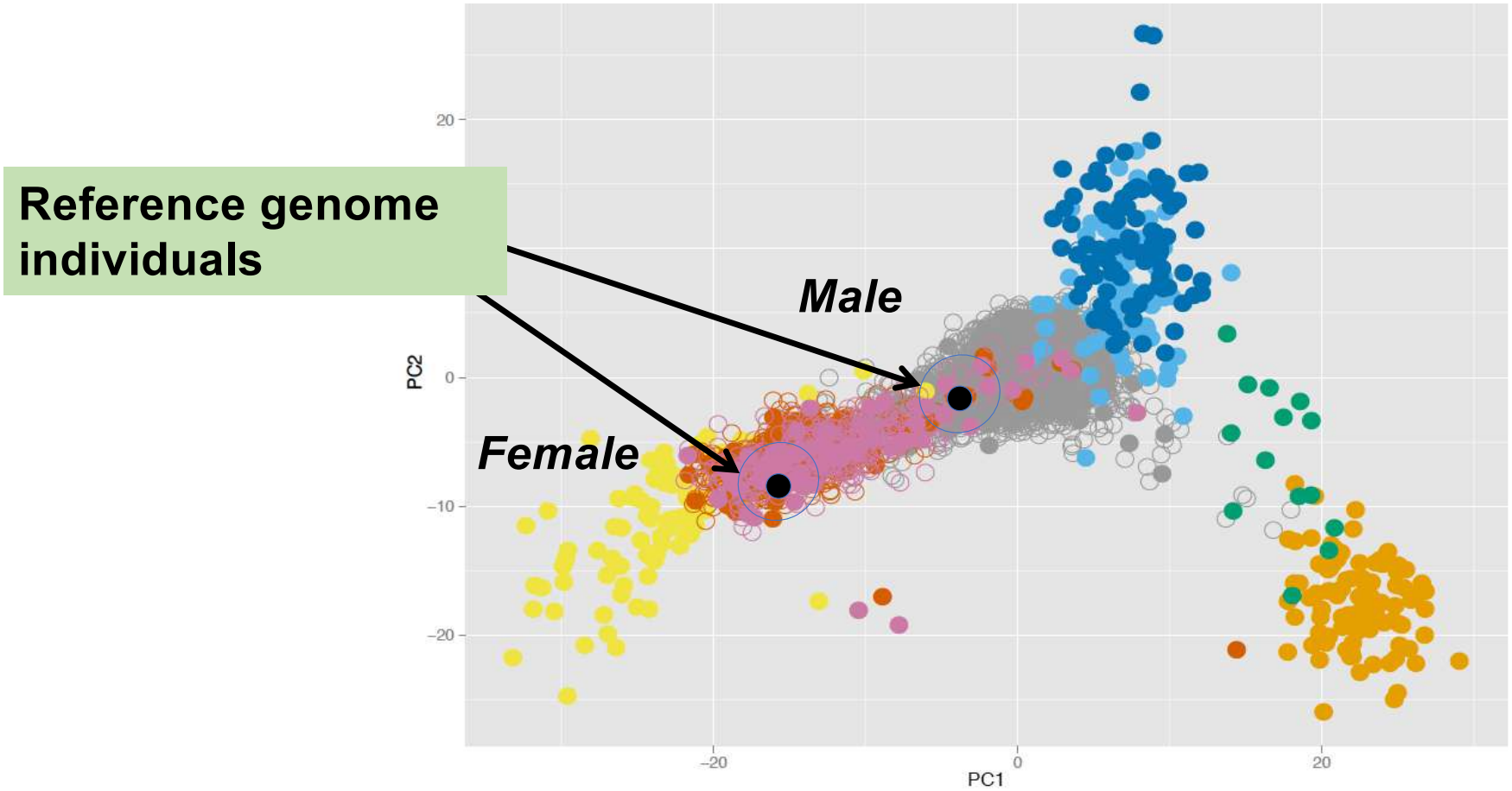
A rare regulatory variant in the MEF2D gene affects gene regulation and splicing and is associated with a SLE sub-phenotype in Swedish cohorts

Fabiana H. G. Farias, Johanna Dahlqvist, Sergey V. Kozyrev, Dag Leonard, Maria Wilbe, Sergei N. Abramov, Andrei Alexsson, Gerli R. Pielberg, Helene Hansson-Hamlin, Göran Andersson, Karolina Tandre, Anders A. Bengtsson, Christopher Sjöwall, Elisabet Svenungsson, Iva Gunnarsson, Solbritt Rantapää-Dahlqvist, Ann-Christine Syvänen, Johanna K. Sandling, Maija-Leena Eloranta, Lars Rönnblom & Kerstin Lindblad-Toh

- … but also, SweGen is used in clinical routine diagnostics

Long-read sequencing of two SweGen individuals

- 2 individuals selected for long-read sequencing



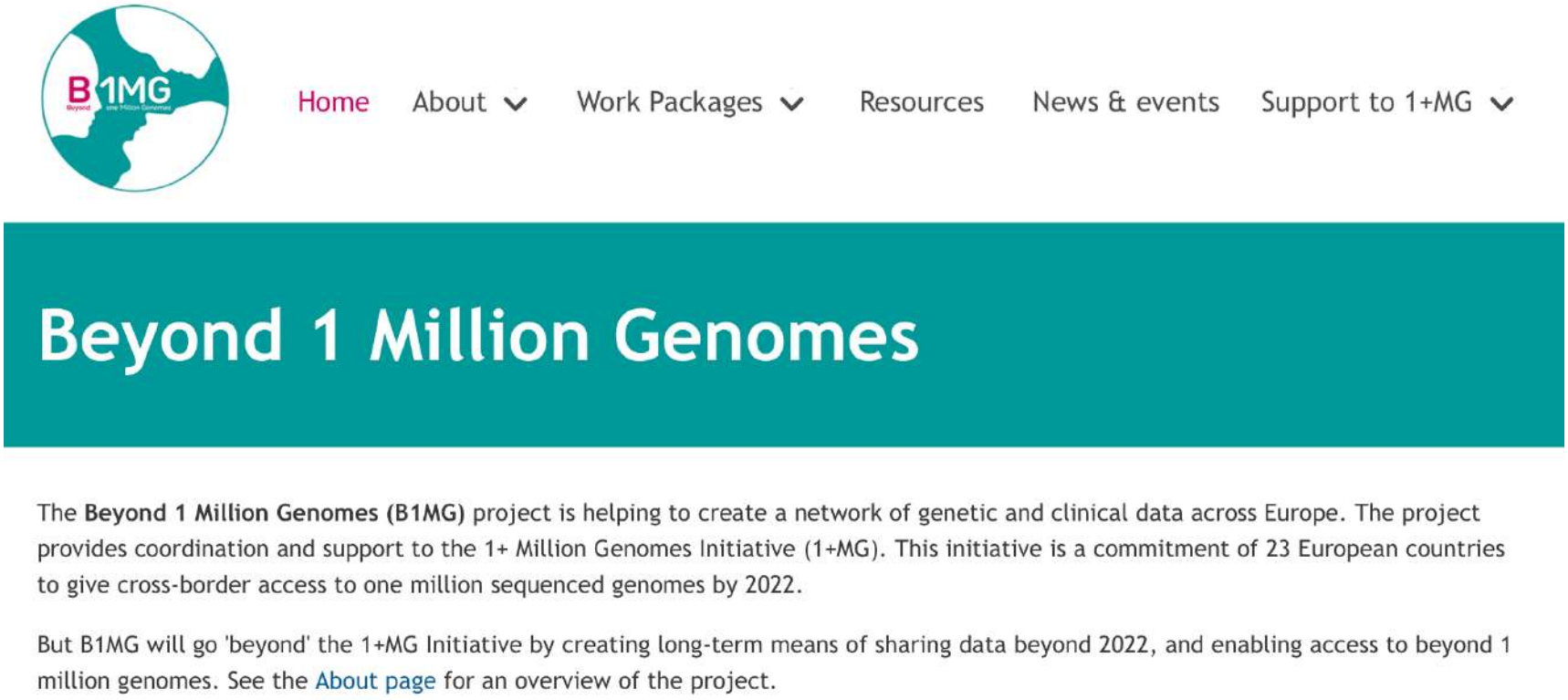
Available data for two reference individuals

Data type	Amount (per individual)
SMRT PacBio	75X coverage
BioNano	2 x 100X coverage
10X Chromium	50X coverage
Illumina WGS	30X coverage
Oxford Nanopore	30X coverage
MGI	30X coverage

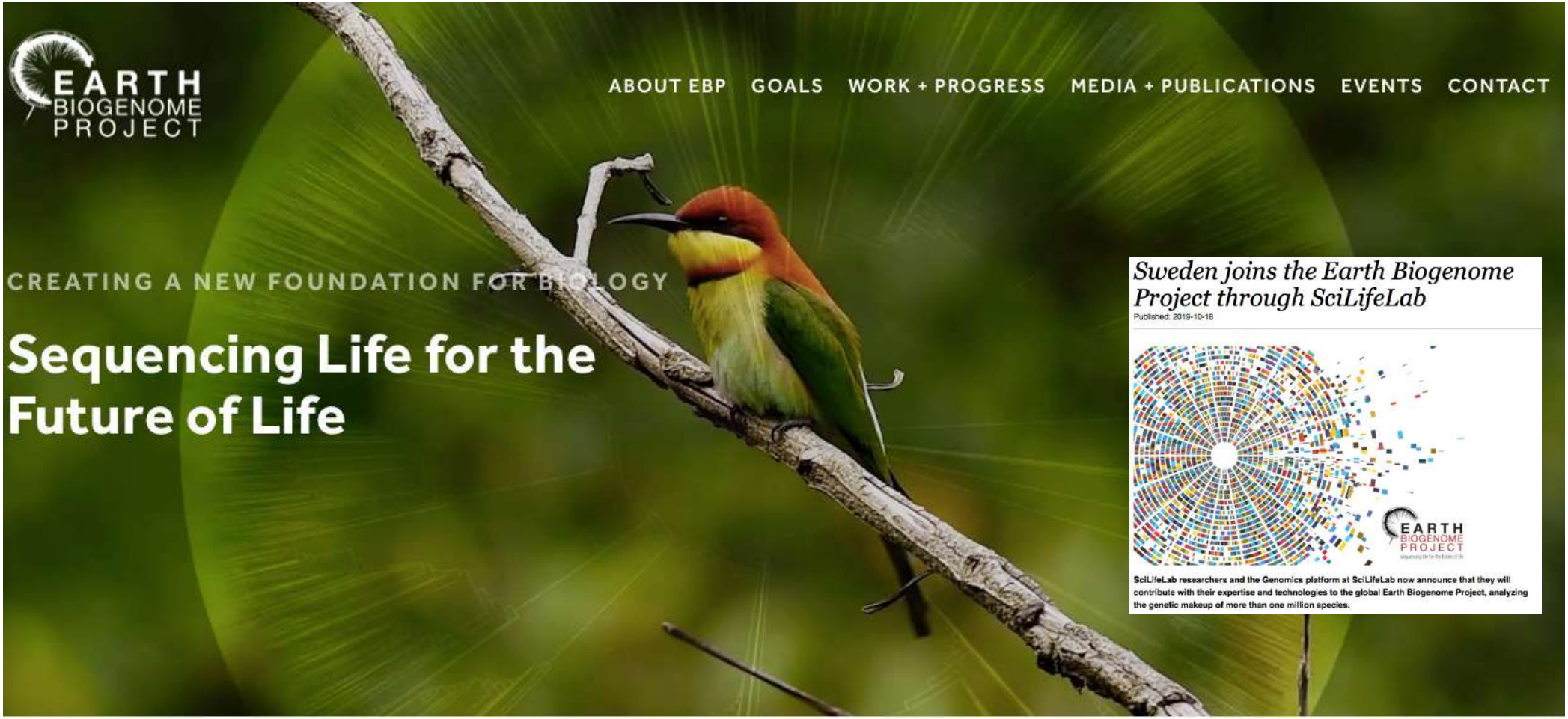
- Aim: use all this data to create high-quality references!

What will happen next?

1+ million genomes to be sequenced across the EU, by 2022(!)



Earth Biogenome Project



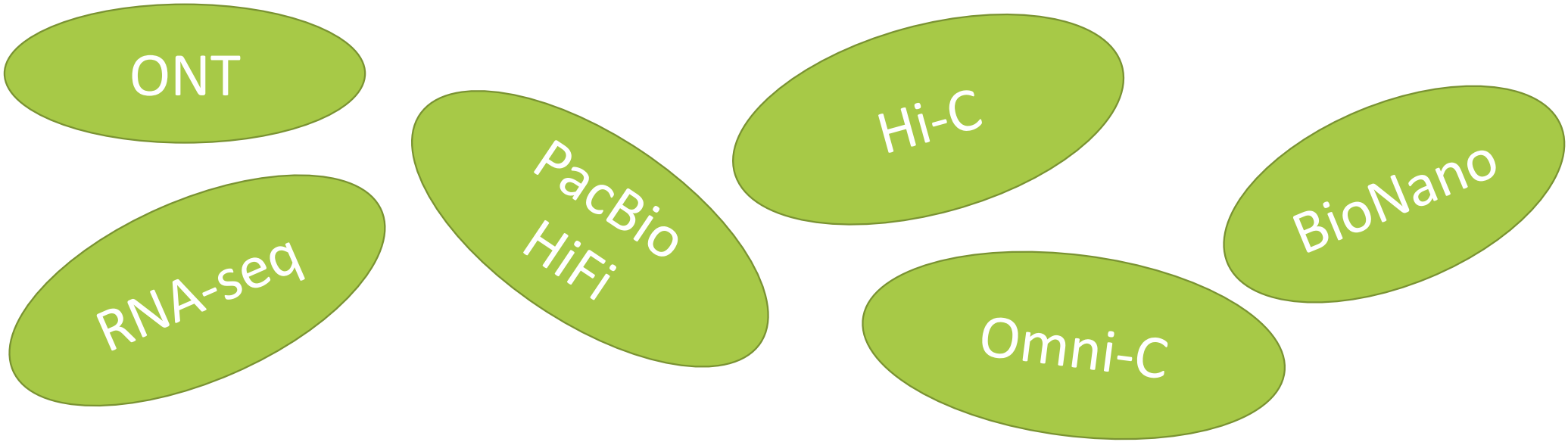
EBP – Data management and analysis

- Over the coming years, many new species will be sequenced
- A combination of different instruments and technologies will be used

- We need good strategies for data analysis and management!

Choice of technology

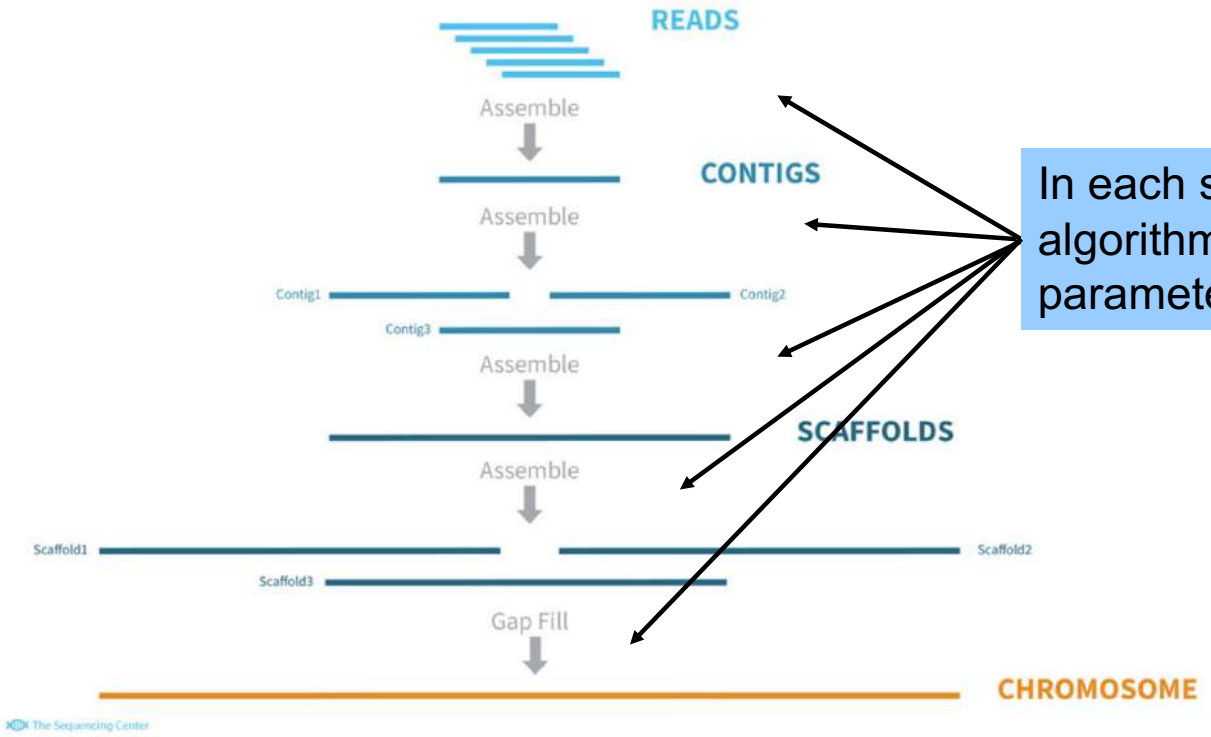
- Make sure sequencing is done using the best technology combination



- This is changing all the time, and lots of different options exist
- The choice will have a big impact on the downstream analysis!

Genome assembly

- Apply analysis pipelines to generate high-quality genome assemblies



In each step, many different strategies, algorithms, implementations & parameters can be used!

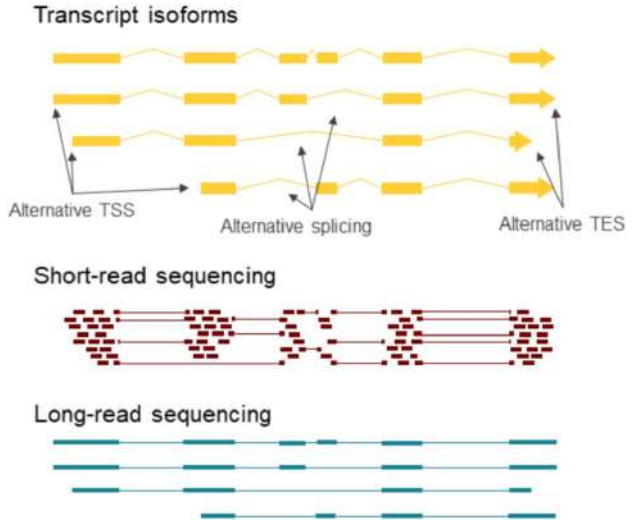
- A challenge for NGI/SciLifeLab is to give best-practice guidelines!

Genome annotation

- Once the assembly is generated, it needs to be annotated!
- Annotation usually means to find out where genes are located

Annotation using computational methods

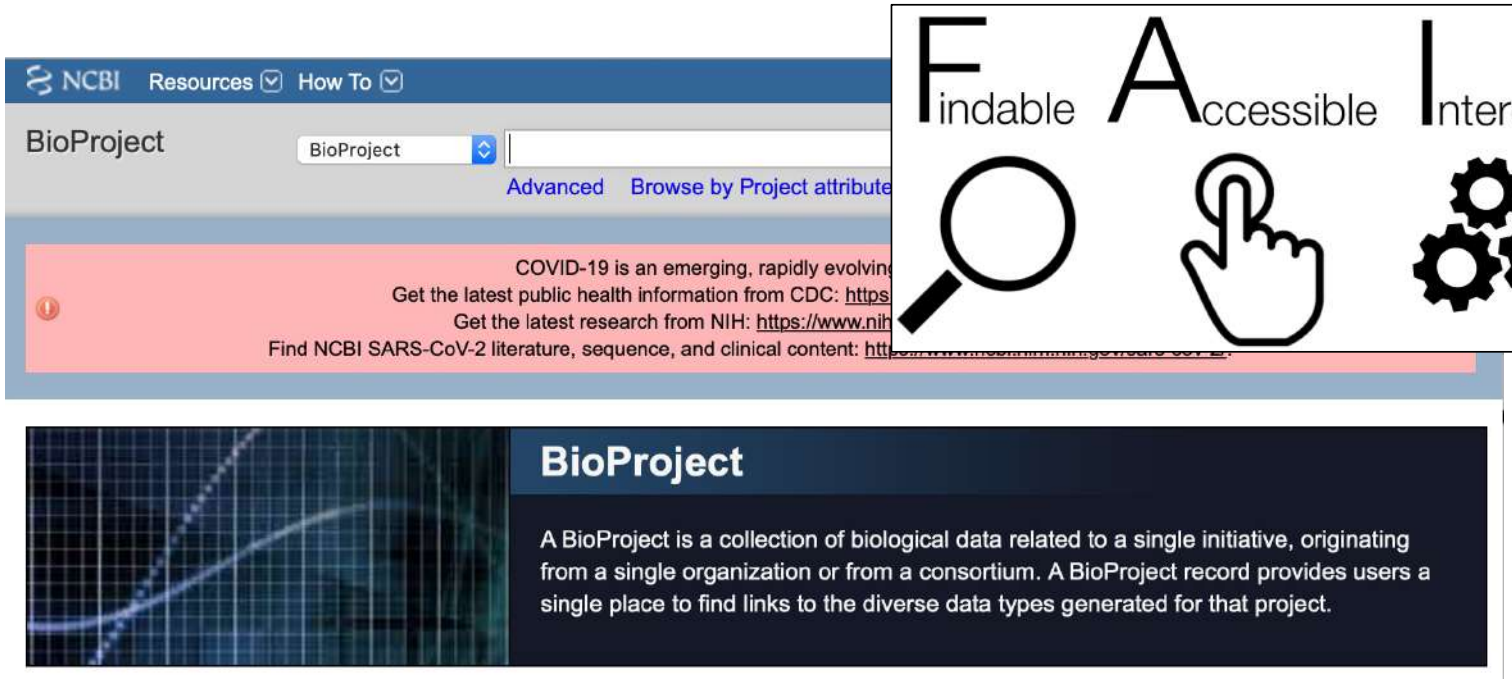
Annotation using RNA-sequencing



- We prefer RNA-sequencing, but still annotation can be challenging!

Data deposition

- Important to deposit the final assembly in public repositories!



- There is a need to develop an interface to international databases

EBP – A collaborative project

- A lot of challenges ahead of us to establish EBP analyses in Sweden
- … but the good news is that this is a community effort

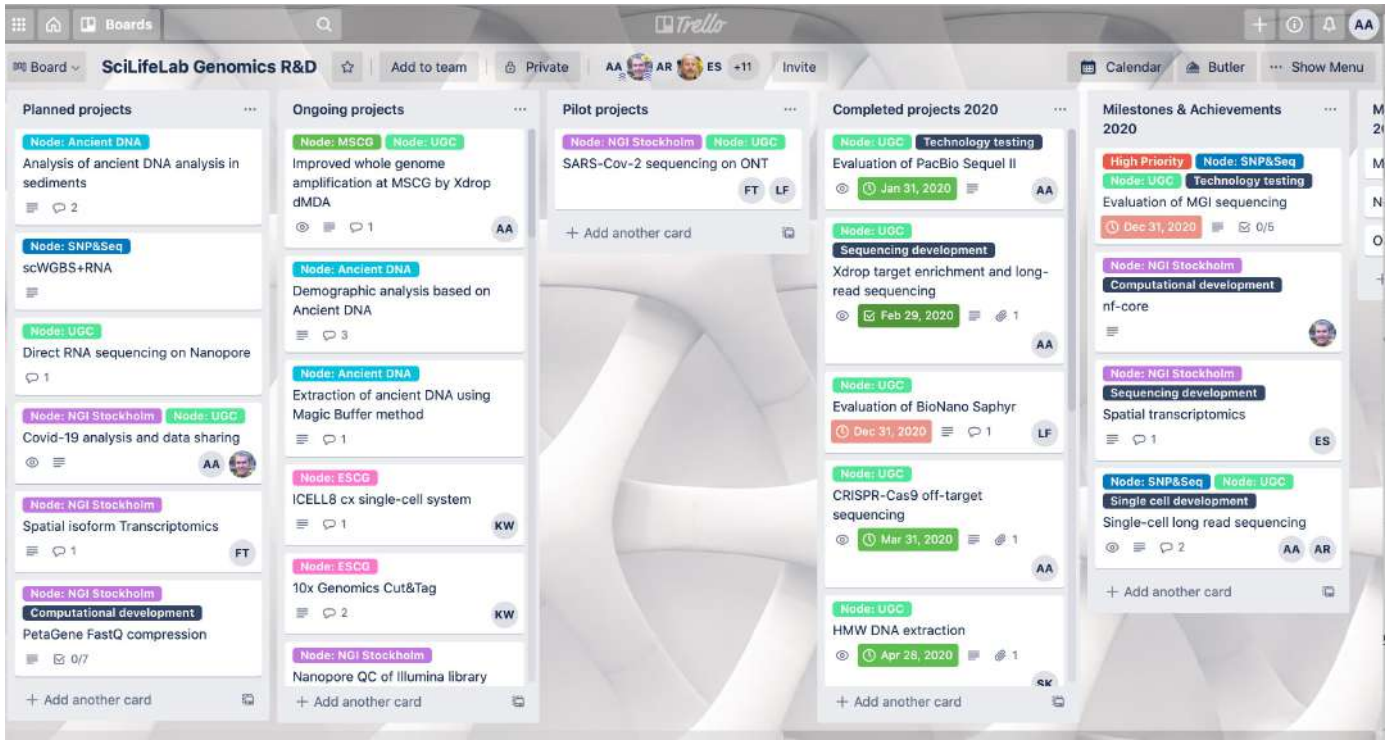
SciLifeLab

EARTH BIOGENOME PROJECT

- There will likely be a lot or opportunities to collaborate!

Research & development at NGI

- We have a joint R&D group for all SciLifeLab genomics facilities
- Aim: to test new applications and possibly offer as service



Research & Development Working Groups

We have started working groups on the following focus areas:

- **E-infrastructure and bioinformatics**
- **Long-read sequencing**
- **Multi-omics**
- **Genome assembly**
- **Epigenetics**
- **Single-cell omics**
- **Library preparation**

Experts from different SciLifeLab platforms take part in these activities!

New solution to measure proteins by NGS!

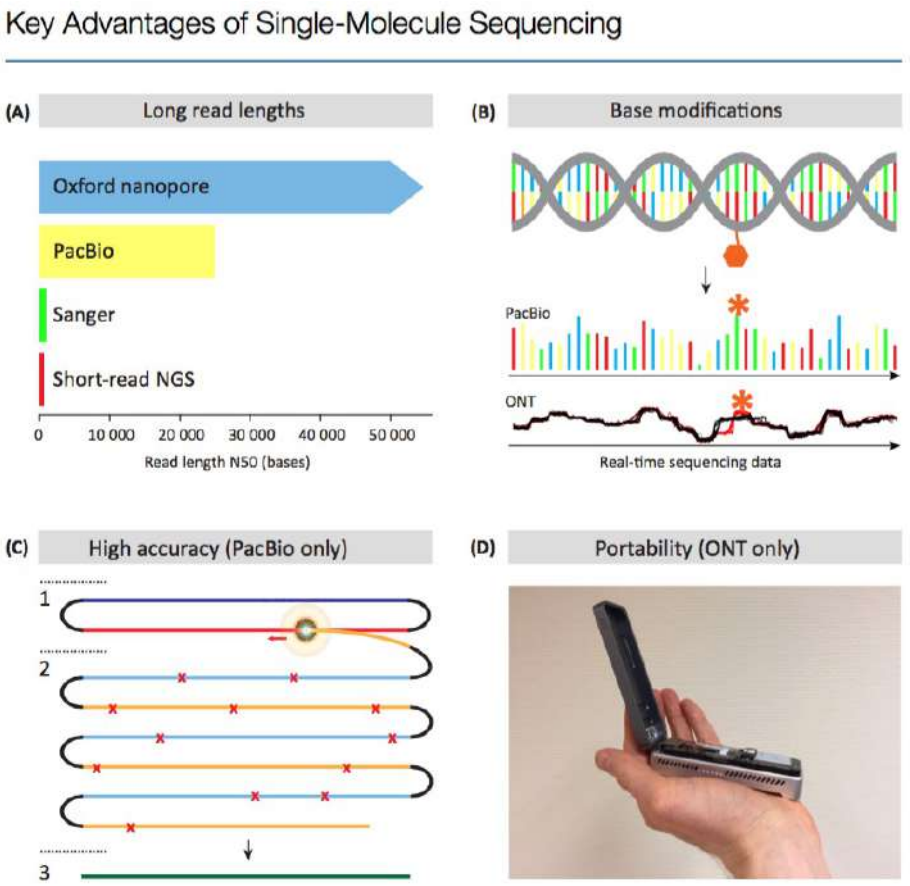
What is Olink Explore?

Olink® Explore 1536/384 is a high-multiplex, high-throughput protein biomarker platform that uses Proximity Extension Assay (PEA) technology coupled to an innovative new readout methodology based on Next Generation Sequencing (NGS) using the following Illumina® instruments:

- NovaSeq 6000
- NextSeq 550
- NextSeq 2000

NGI is the first service provider in Europe for this application!

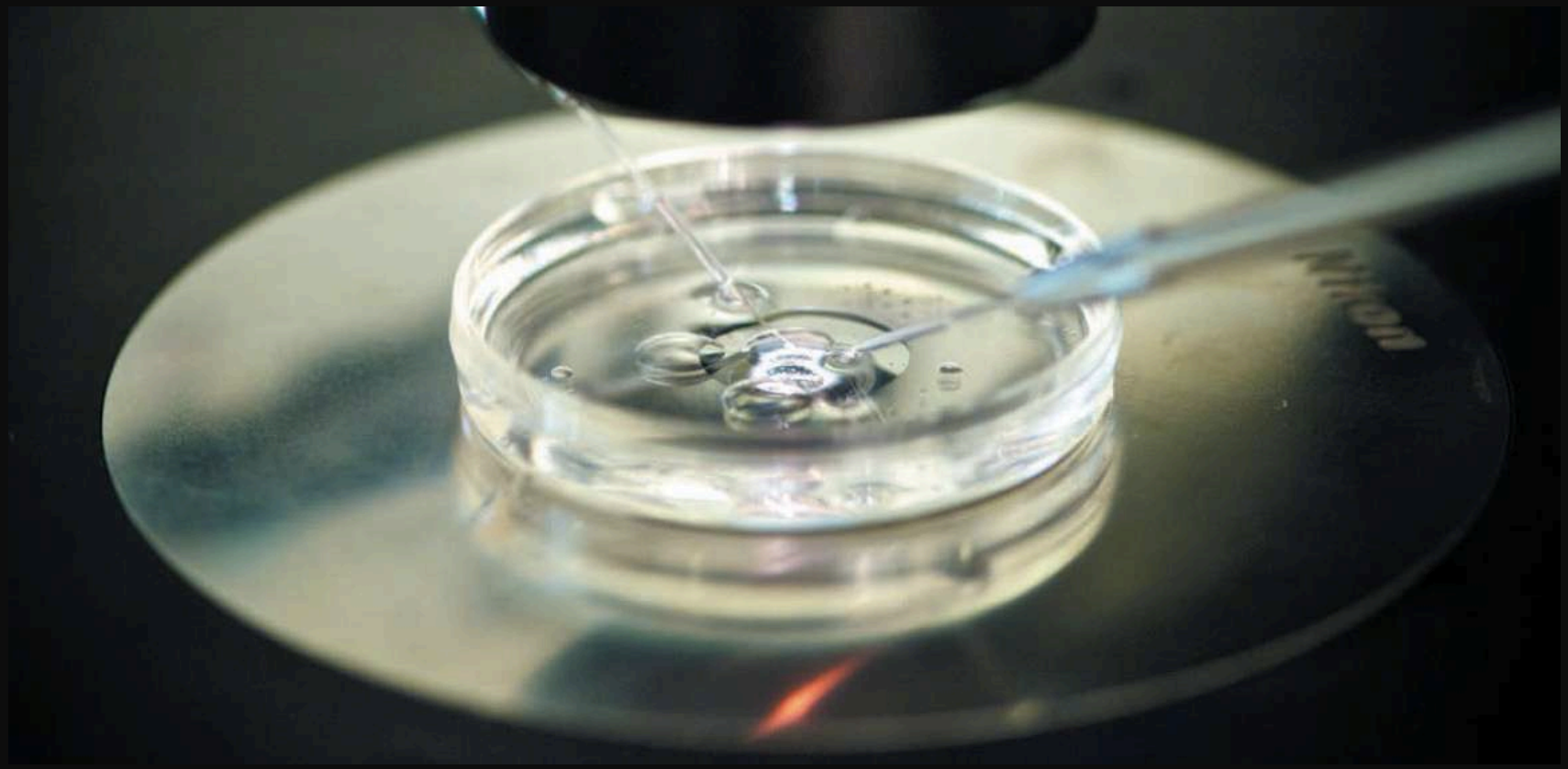
Long-read sequencing in the clinic



Ameur, Kloosterman & Hestand. Trends In Biotechnology 2019

CRISPR-Cas9 genome editing

We are developing new long-read tools to detect “off-target” mutations

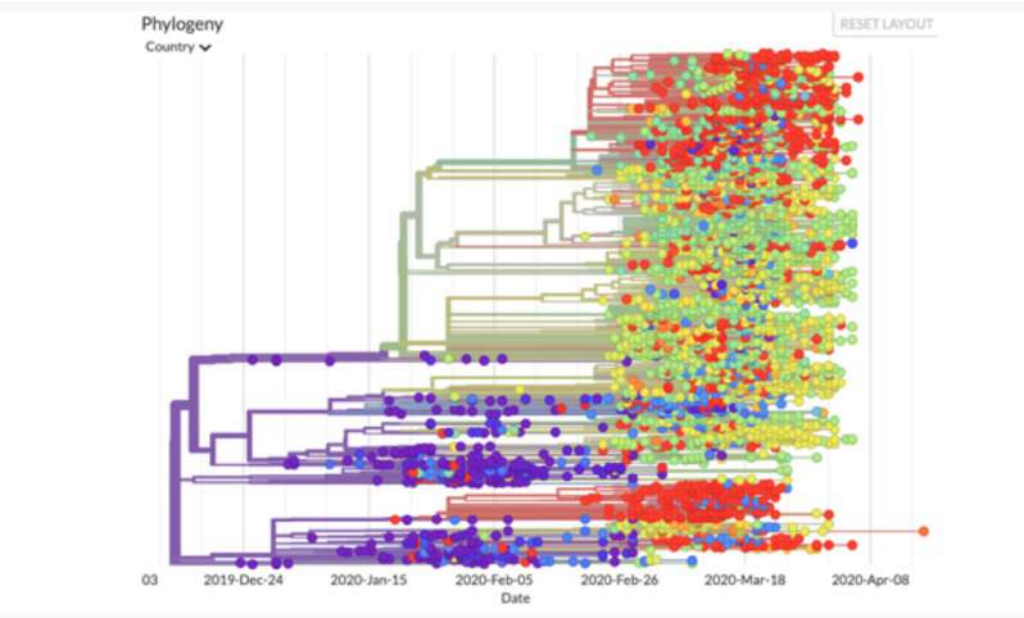


Höijer et al, Genome Biology 2020; Nature Communications 2022

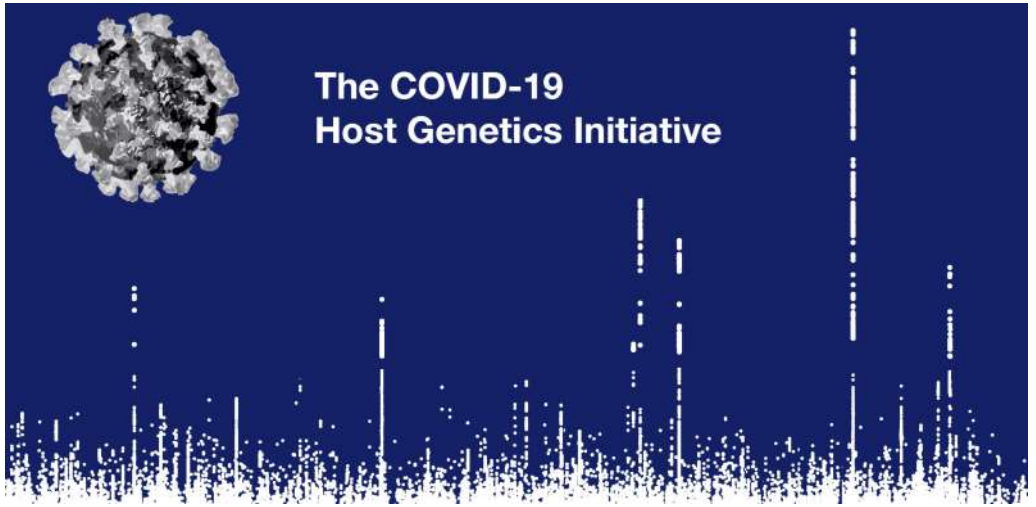
COVID-19 sequencing

- We have been evaluating different methods and sequencing protocols…

SARS-Cov-2 sequencing (mutations)



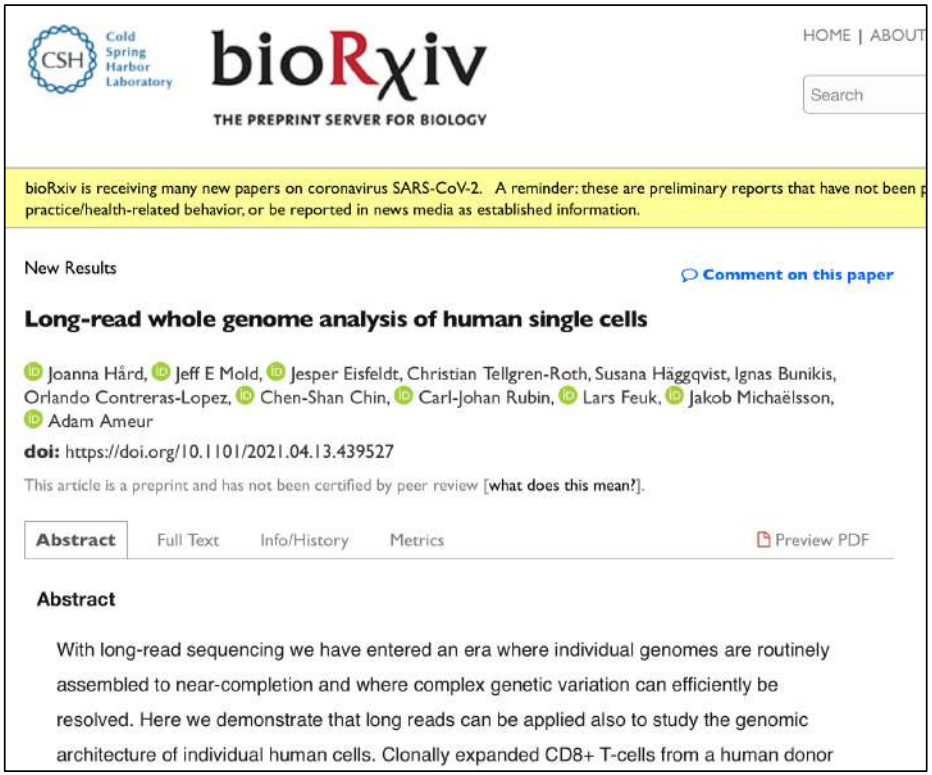
Human re-sequencing (host response)



- We are involved in monitoring variants in Swedish wastewater

Long-read single-cell sequencing

- The first reported WGS of a human single cell with long reads!



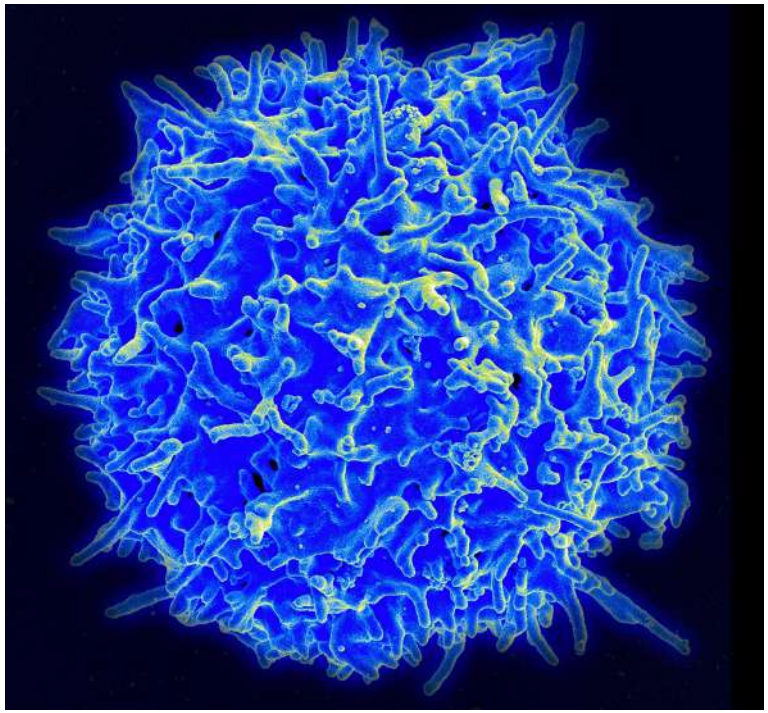
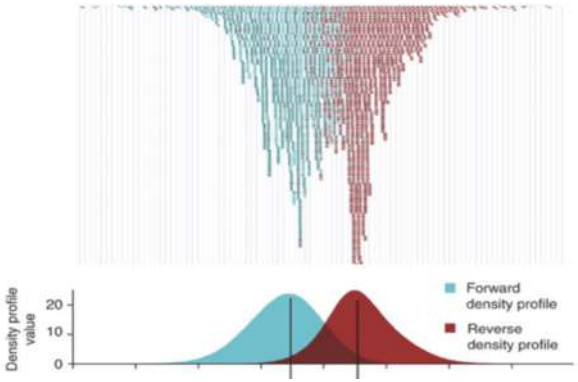


Image from https://en.wikipedia.org/wiki/T_cell

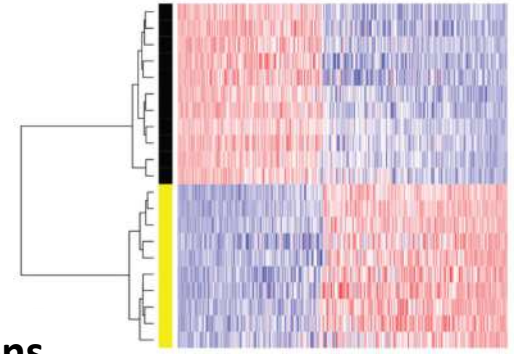
- We are also testing new protocols for long-read single-cell RNA sequencing

Many topics I have not covered…

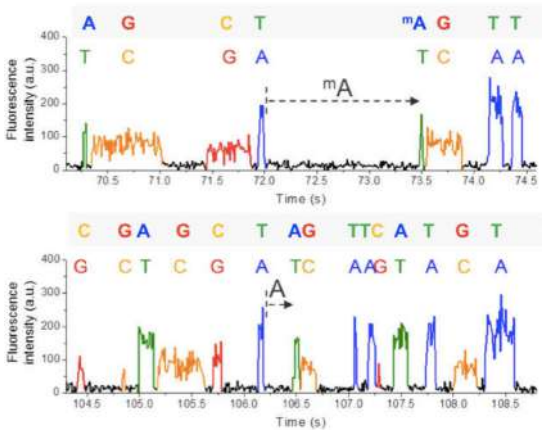
ChIP-seq analysis



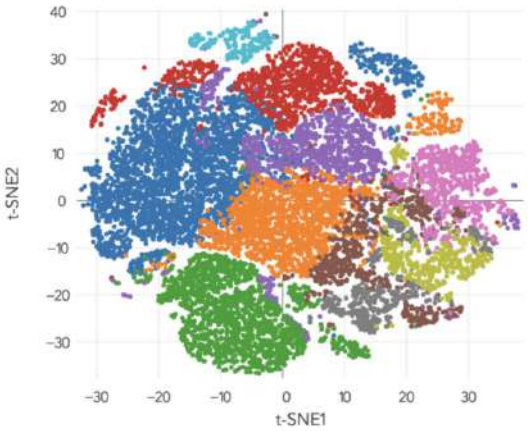
RNA-Seq



DNA base modifications



Single cell RNA-Seq



- Simply too much to talk about in just one lecture…

Long-read Uppsala Meeting 2022!

https://www.lrua2022.se

Thanks for your attention!