



# **Next Generation Sequencing and Bioinformatics Analysis Pipelines**

Adam Ameur  
National Genomics Infrastructure  
SciLifeLab Uppsala  
[adam.ameur@igp.uu.se](mailto: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

illumina®



ThermoFisher  
life **ion torrent**  
Δ \* Δ O X □ + ∞



PACIFIC  
BIOSCIENCES®

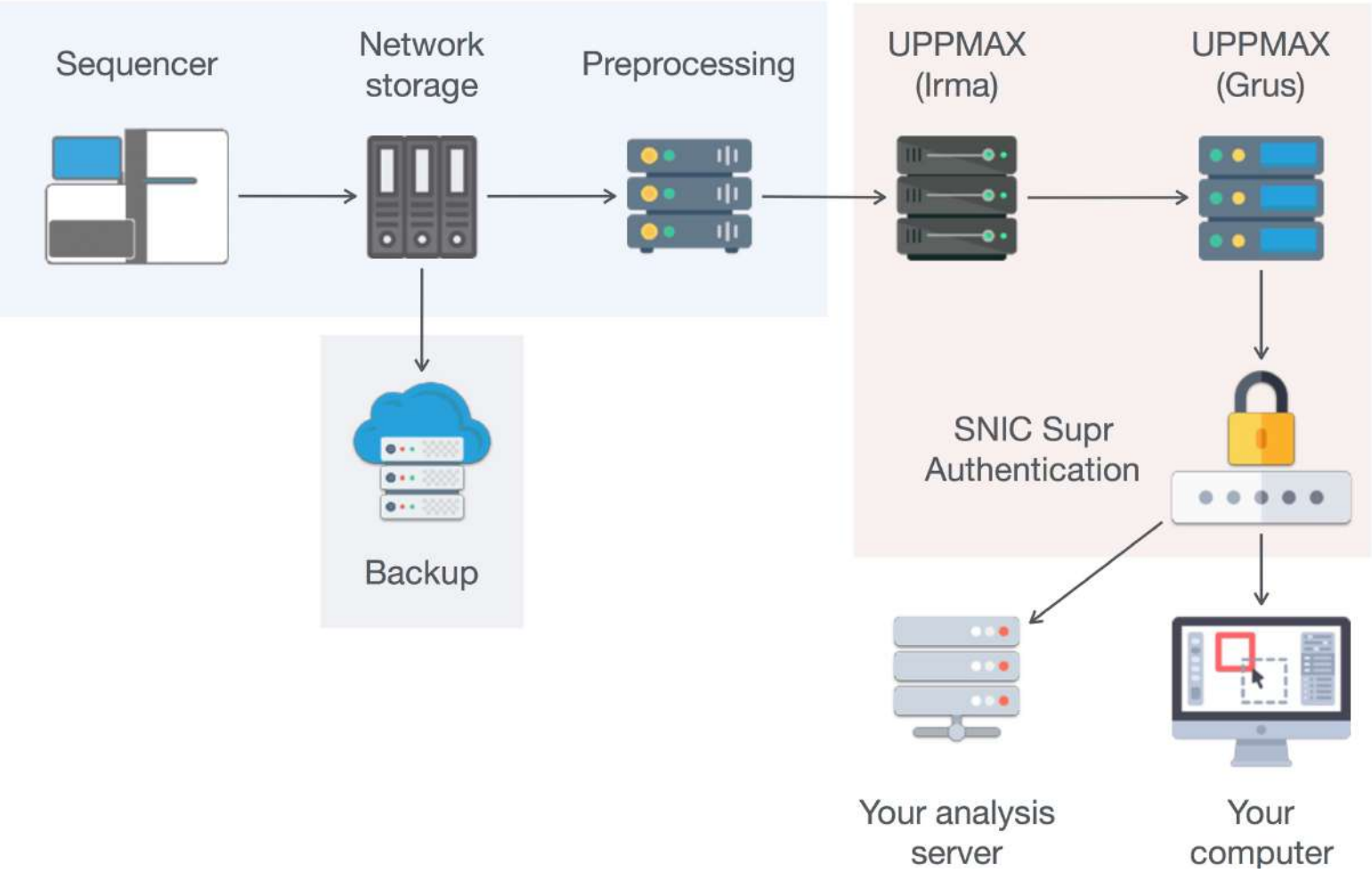


Oxford  
**NANOPORE**  
Technologies





# 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



## 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



**Sarek**

Manta

MuTect1

ASCAT

MuTect2

Strelka

FreeBayes

GATK  
HaplotypeCaller





# 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)





# Multi QC example

**MultiQC** v1.0

P1234: Test\_NGI\_Project

General Stats

NGI-RNAseq

Sample Similarity

MDS Plot

STAR

Cutadapt

FastQC

Sequence Quality Histograms

Per Sequence Quality Scores

Per Base Sequence Content

Per Sequence GC Content

Per Base N Content

Sequence Length Distribution

Sequence Duplication Levels

Overrepresented sequences

Adapter Content

## MultiQC

**P1234: Test\_NGI\_Project**

This is an example project. All identifying data has been removed.

Contact E-mail: phil.ewels@scilifelab.se  
Application Type: RNA-seq  
Sequencing Platform: HiSeq 2500 High Output V4  
Sequencing Setup: 2x125  
Reference Genome: hg19

Report generated on 2017-05-17, 18:43 based on data in:  
/Users/phil.ewels/GitHub/MultiQC\_website/public\_html/examples/ngi-rna/data

NGI names | User supplied names

### General Statistics

Copy table | Configure Columns | Plot | Showing 22/22 rows and 6/6 columns.

Sample Name	% Aligned	M Aligned	% Trimmed	% Dups	% GC	M Seqs
P1234_1001	68.2%	22.8	10.3%	71.3%	49%	33.7
P1234_1002	67.9%	20.9	10.7%	70.1%	50%	31.1
P1234_1003	64.7%	21.7	11.0%	72.3%	50%	33.7
P1234_1004	55.2%	17.0	13.2%	73.4%	51%	31.2
P1234_1005	53.0%	17.7	15.9%	75.8%	52%	33.8
P1234_1006	52.7%	16.1	14.1%	73.8%	52%	30.8
P1234_1007	33.0%	7.0	32.0%	60.5%	52%	21.8
P1234_1008	27.5%	4.3	44.2%	79.1%	50%	16.7
P1234_1009	52.3%	10.5	20.9%	64.7%	48%	20.5



# Data delivery via GRUS

---

- GRUS is an UPPMAX tool for NGI data delivery
  - NGI creates a SNIC Supr “delivery project” for each NGI sequencing project
  - Project PI and contact person is responsible for the project
  - Email sent with project ID and instructions
- Grus is for short-term storage only!
  - Requires two-factor authentication

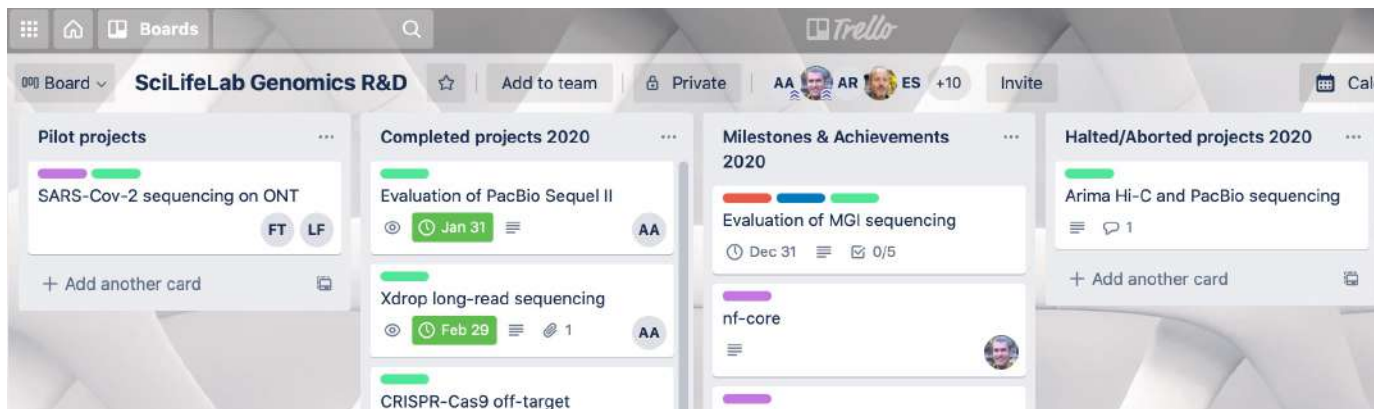
**New delivery system to be launched in 2022!**



# 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 19<sup>th</sup> 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

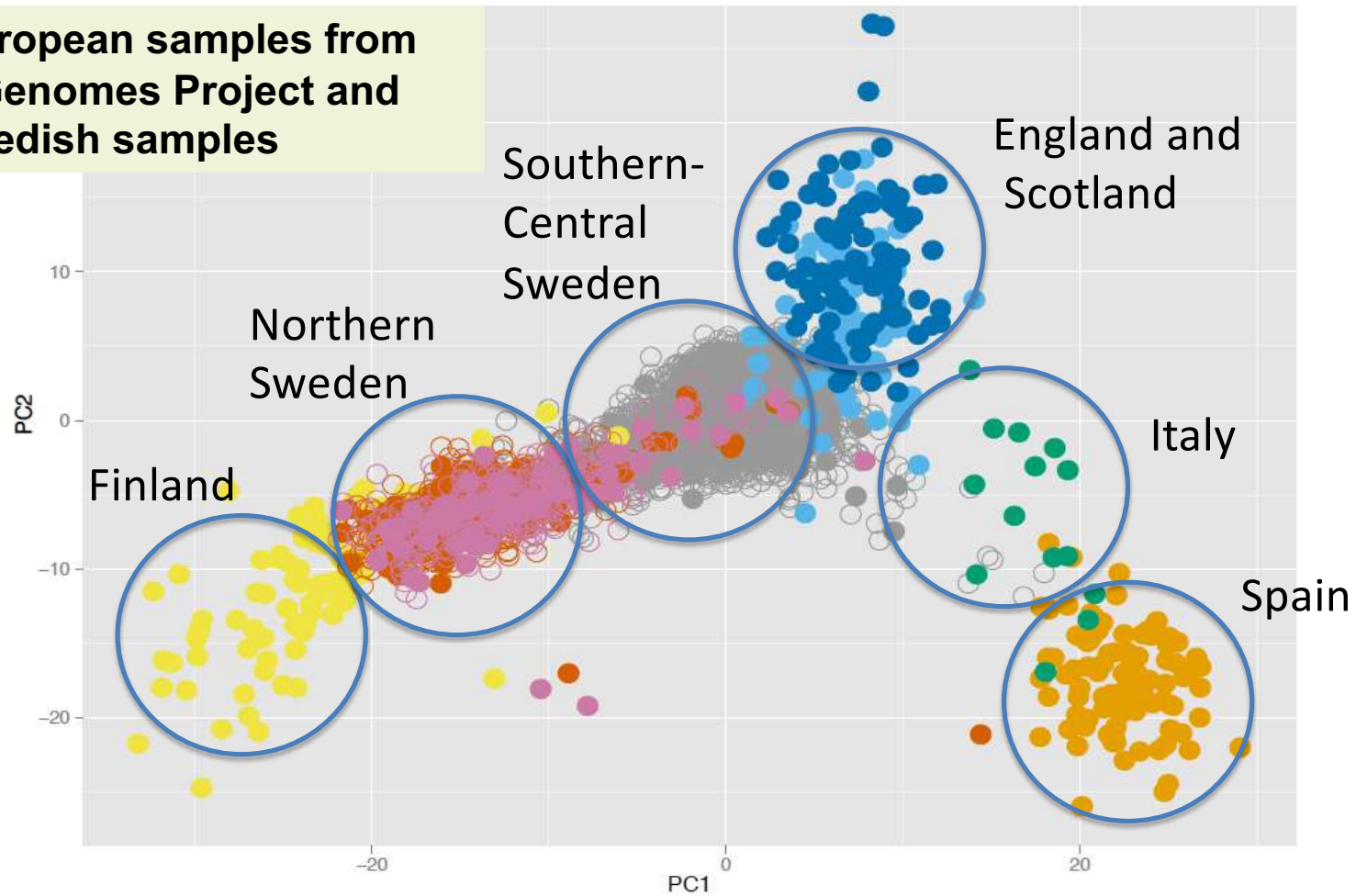
High demand for the data from many different groups:

→ 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

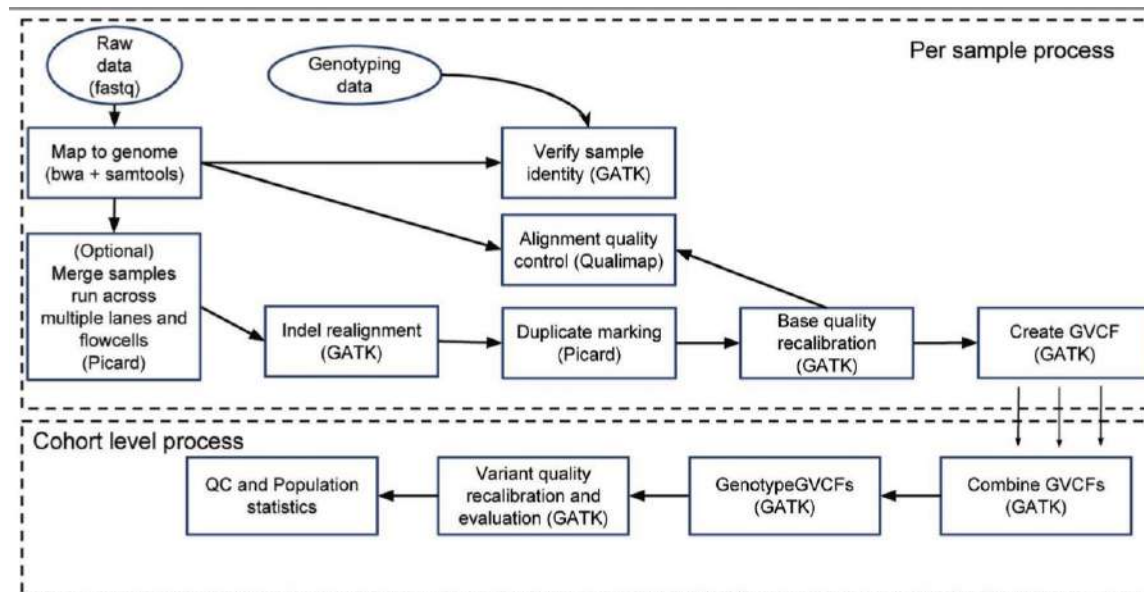


- Sequencing done both at NGI Sthlm and NGI Uppsala
- All 1,000 samples completed in September 2016

# 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



## SweGen Variant Frequency Dataset

This dataset contains whole-genome variant frequencies for 1000 Swedish individuals generated within the SweGen project. The frequency data is intended to be used as a resource for the research community and clinical genetics laboratories.

Please note that the 1000 individuals included in the SweGen project represent a cross-section of the Swedish population and that no disease information has been used for the selection. The frequency data may therefore include genetic variants that are associated with, or causative of, disease.

We request that any use of data from the SweGen project cite [this article in the European Journal of Human Genetics](#).

Individual positions in the genome can be viewed using the Beacon or Graphical Browser. To download the variant frequency file you need to register.

A high confidence set of HLA allele frequencies is available for download under Dataset Access. For a detailed description of the SweGen HLA analysis, please see [this bioRxiv preprint](#).



[More information](#)

[Beacon](#)

[Graphical Browser](#)





- Aggregated frequencies available from: [swefreq.nbis.se](http://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 ,<sup>\*,1,2,3</sup> Gustaf Mårtensson,<sup>4</sup> Adam Ameur ,<sup>5</sup> Daniel Nilsson ,<sup>1,2,3</sup> and Anna Lindstrand <sup>1,3</sup>  
<sup>1</sup>Department of Molecular Medicine and Surgery, Center for Molecular Medicine, Karolinska Institute, Stockholm, Sweden  
<sup>2</sup>Science for Life Laboratory, Karolinska Institutet Science Park, Solna, Sweden  
<sup>3</sup>Department of Clinical Genetics, Karolinska University Hospital, Stockholm, Sweden  
<sup>4</sup>Department of Clinical Chemistry, Karolinska University Hospital, Stockholm, Sweden  
<sup>5</sup>Science for Life Laboratory, Karolinska Institutet Science Park, Solna, Sweden  
\*Correspondence: [jesper.eisfeldt@ki.se](mailto:jesper.eisfeldt@ki.se)  
Assistance: [anna.lindstrand@ki.se](mailto:anna.lindstrand@ki.se)

CLINICAL RESEARCH ARTICLE

**Cytokine Autoantibody Screening in the Swedish Addison Registry Identifies Patients With Undiagnosed APS1**  
Daniel Eriksson,<sup>1,2</sup> Frida Dalin,<sup>1,3</sup> Gabriel Nordling Eriksson,<sup>4</sup> Nils Lan Matteo Bianchi,<sup>5</sup> Åsa Hallgren,<sup>1,3</sup> Per Dahlqvist,<sup>6</sup> Jeanette Wahlberg, Olov Ekwall,<sup>10,11</sup> Ola Winqvist,<sup>12</sup> Sergiu-Bogdan Catrina,<sup>4</sup> Johan Rön Swedish Addison Registry Study Group, Anna-Lena Hulting,<sup>4</sup> Kerstin Lin Mohammad Alimohammadi,<sup>15</sup> Eystein S. Husebye,<sup>1,16,17,18</sup> Per Morten K Gerli Rosengren Pielberg,<sup>5</sup> Sophie Bensing,<sup>2,4</sup> and Olle Kämpe<sup>1,2,3,18</sup>

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  

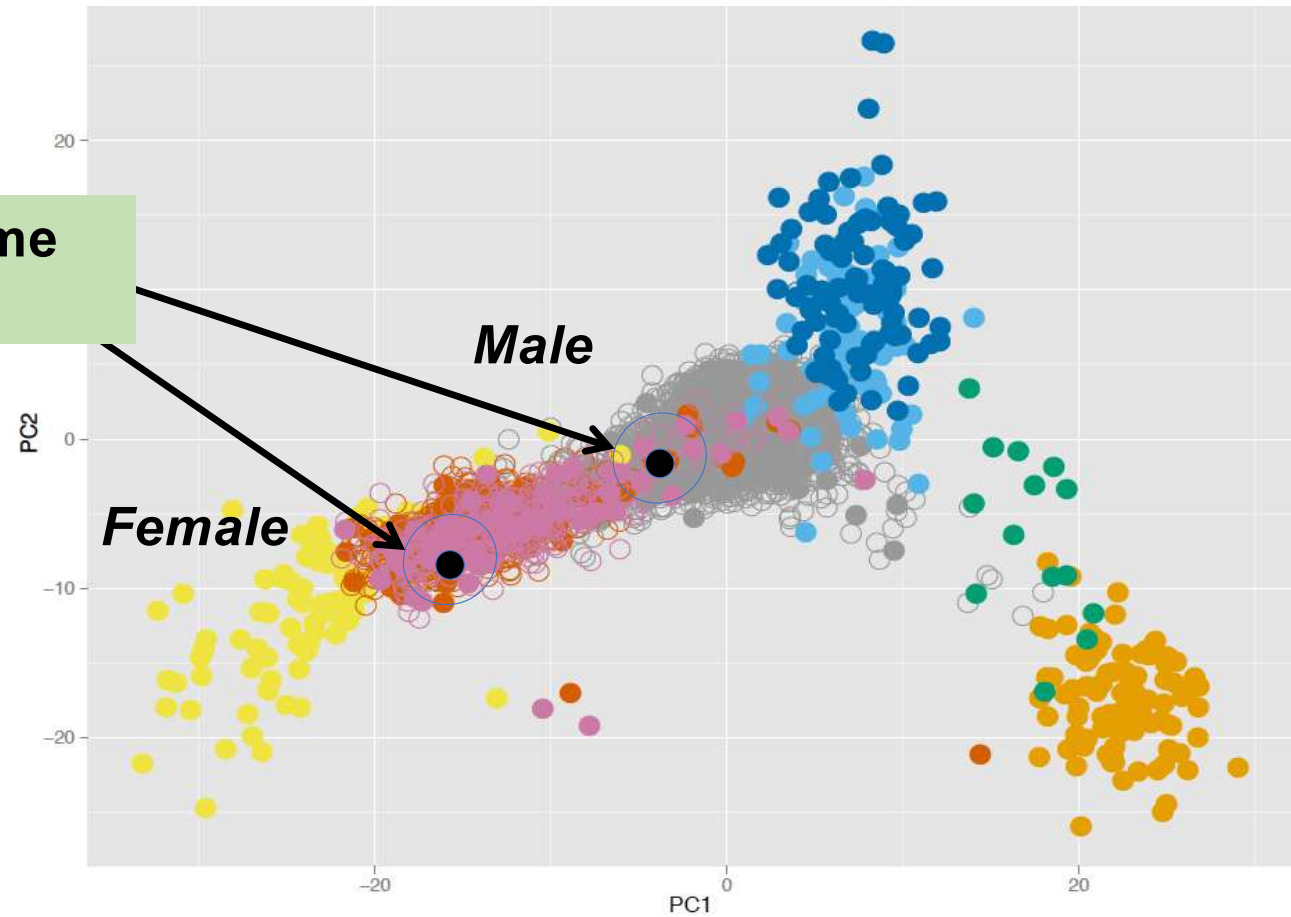
A rare regulatory variant in the MEF2D gene affects gene regulation and splicing and is associated with a SLE sub-phenotype in Swedish cohorts  
Fabiana H. G. Farias , Johanna Dahlqvist, Sergey V. Kozyrev, Dag Leonard, Maria Wilbe, Sergei N. Abramov, Andrei Alexsson, Gerli R. Pielberg, Helene Hansson-Hamlin, Göran Andersson, Karolina Tandre, Anders A. Bengtsson, Christopher Sjöwall, Elisabet Svenungsson, Iva Gunnarsson, Solbritt Rantapää-Dahlqvist, Ann-Christine Syvänen, Johanna K. Sandling, Majja-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

Reference genome individuals



# Available data for two reference individuals

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



[Home](#) [About](#) [Work Packages](#) [Resources](#) [News & events](#) [Support to 1+MG](#)

## Beyond 1 Million Genomes

The **Beyond 1 Million Genomes (B1MG)** project is helping to create a network of genetic and clinical data across Europe. The project provides coordination and support to the 1+ Million Genomes Initiative (1+MG). This initiative is a commitment of 23 European countries to give cross-border access to one million sequenced genomes by 2022.

But B1MG will go 'beyond' the 1+MG Initiative by creating long-term means of sharing data beyond 2022, and enabling access to beyond 1 million genomes. See the [About page](#) for an overview of the project.



# Earth Biogenome Project



**EARTH BIOGENOME PROJECT**

ABOUT EBP GOALS WORK + PROGRESS MEDIA + PUBLICATIONS EVENTS CONTACT

CREATING A NEW FOUNDATION FOR BIOLOGY

## Sequencing Life for the Future of Life

*Sweden joins the Earth Biogenome Project through SciLifeLab*

Published: 2019-10-18

**EARTH BIOGENOME PROJECT**

SciLifeLab researchers and the Genomics platform at SciLifeLab now announce that they will contribute with their expertise and technologies to the global Earth Biogenome Project, analyzing the genetic makeup of more than one million species.

# 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



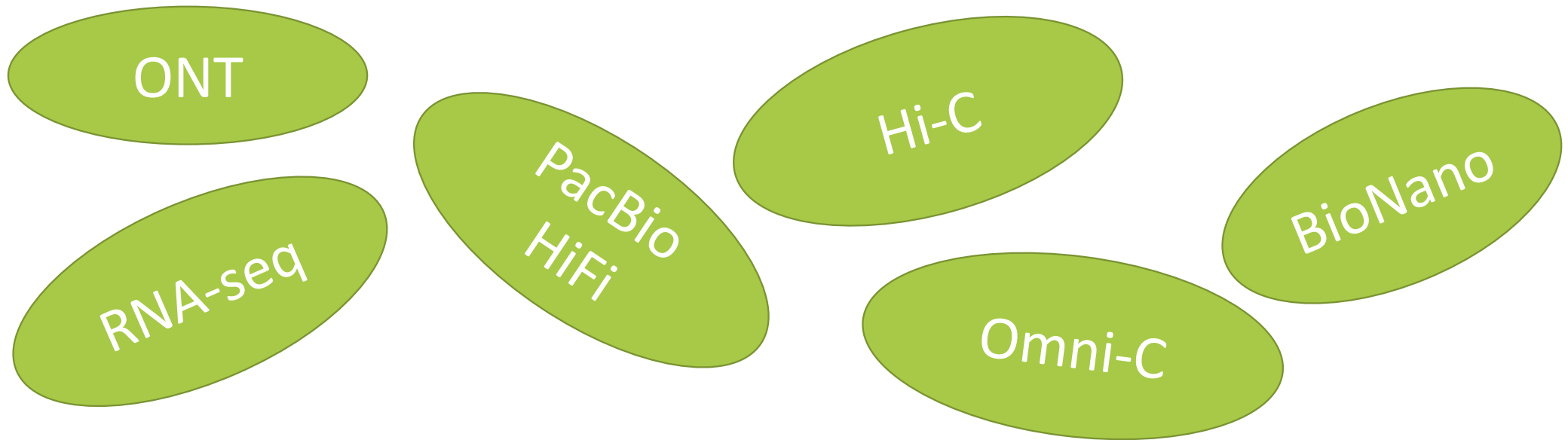
- We need good strategies for data analysis and management!

# Choice of technology

---



- Make sure sequencing is done using the best technology combination

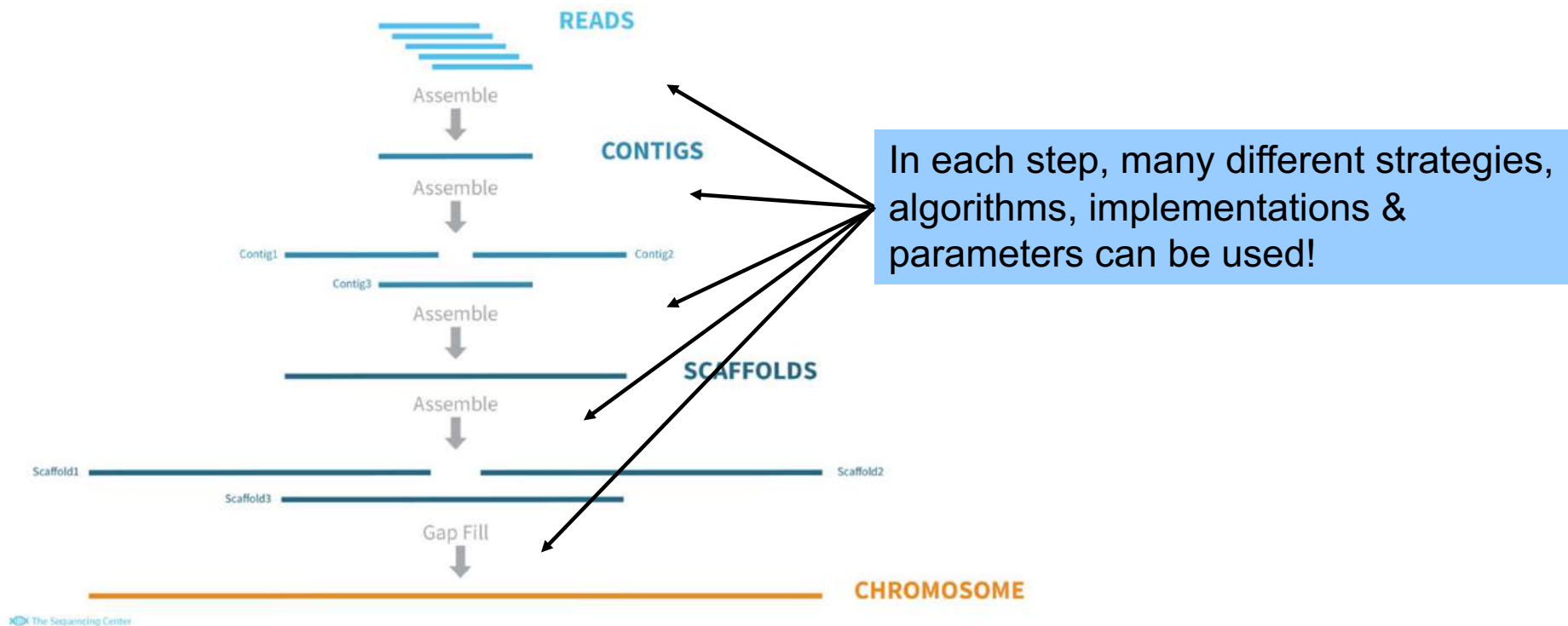


- This is changing all the time, and lots of different options exist
- The choice will have a big impact on the downstream analysis!

# Genome assembly



- Apply analysis pipelines to generate high-quality genome assemblies



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



# Genome annotation

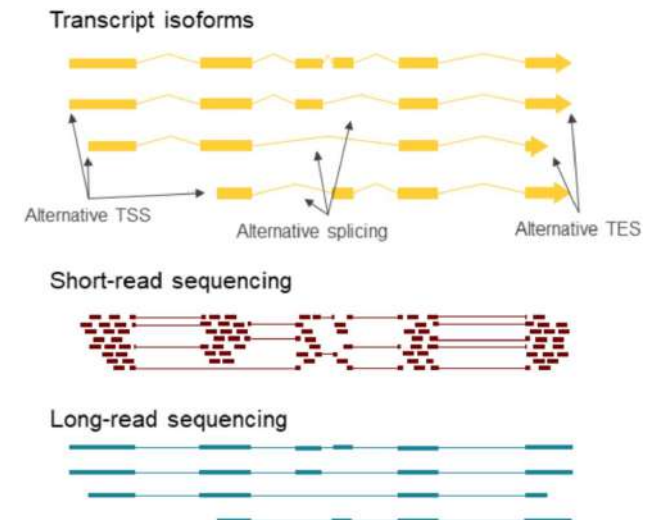


- Once the assembly is generated, it needs to be annotated!
- Annotation usually means to find out where genes are located

## 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!

The image shows a screenshot of the NCBI BioProject website. The top navigation bar includes 'NCBI Resources' and 'How To'. The main content area features a search bar with 'BioProject' entered and a dropdown menu. Below the search bar, there are links for 'Advanced' and 'Browse by Project attribute'. A prominent red banner contains information about COVID-19, including links to CDC and NIH resources. A dark blue box at the bottom left contains the text: 'BioProject A BioProject is a collection of biological data related to a single initiative, originating from a single organization or from a consortium. A BioProject record provides users a single place to find links to the diverse data types generated for that project.'

Overlaid on the right side of the screenshot is a white box containing the FAIR principles: **F**indable, **A**ccessible, **I**nteroperable, and **R**eusable. Each principle is accompanied by a black icon: a magnifying glass for Findable, a hand pointing for Accessible, three interlocking gears for Interoperable, and a recycling symbol for Reusable.

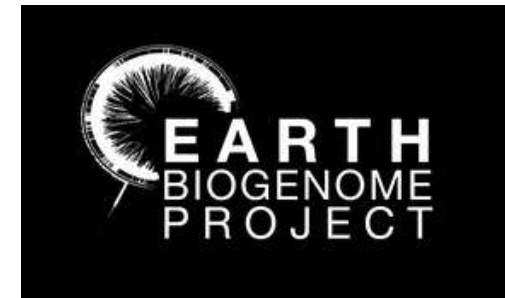
- There is a need to develop an interface to international databases

# 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

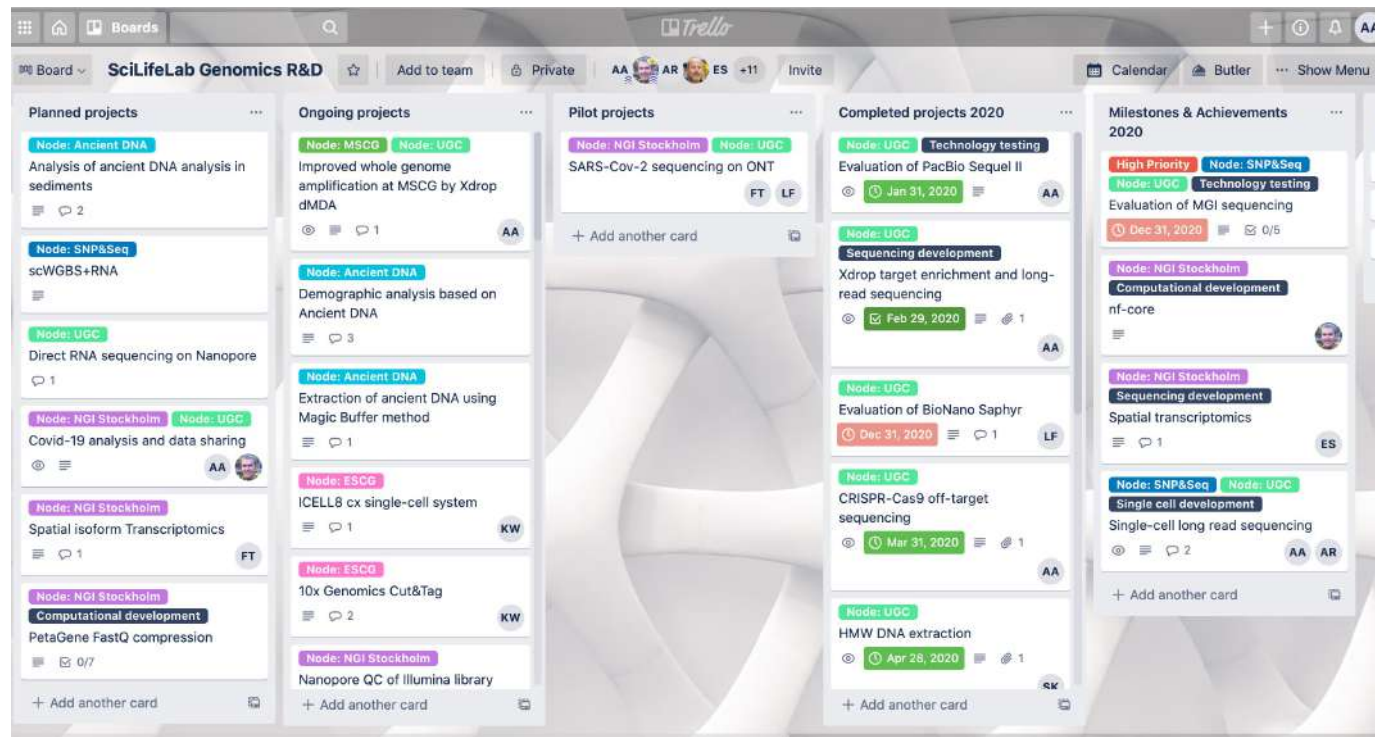


- There will likely be a lot of opportunities to collaborate!

# 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**

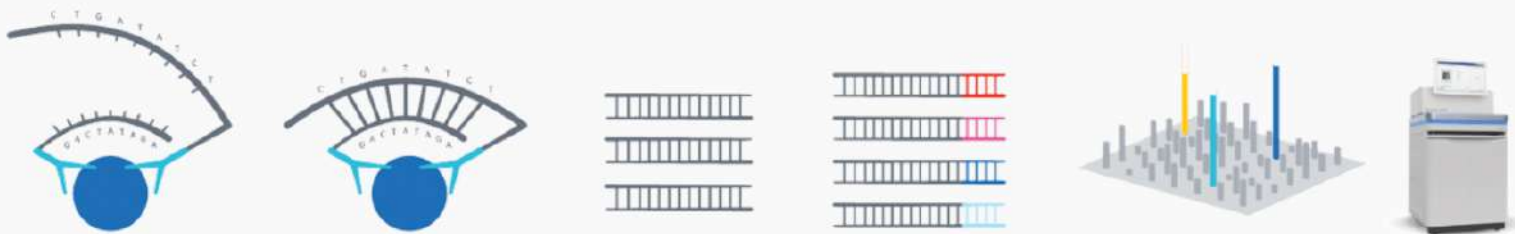
Experts from different SciLifeLab platforms take part in these activities!

# 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



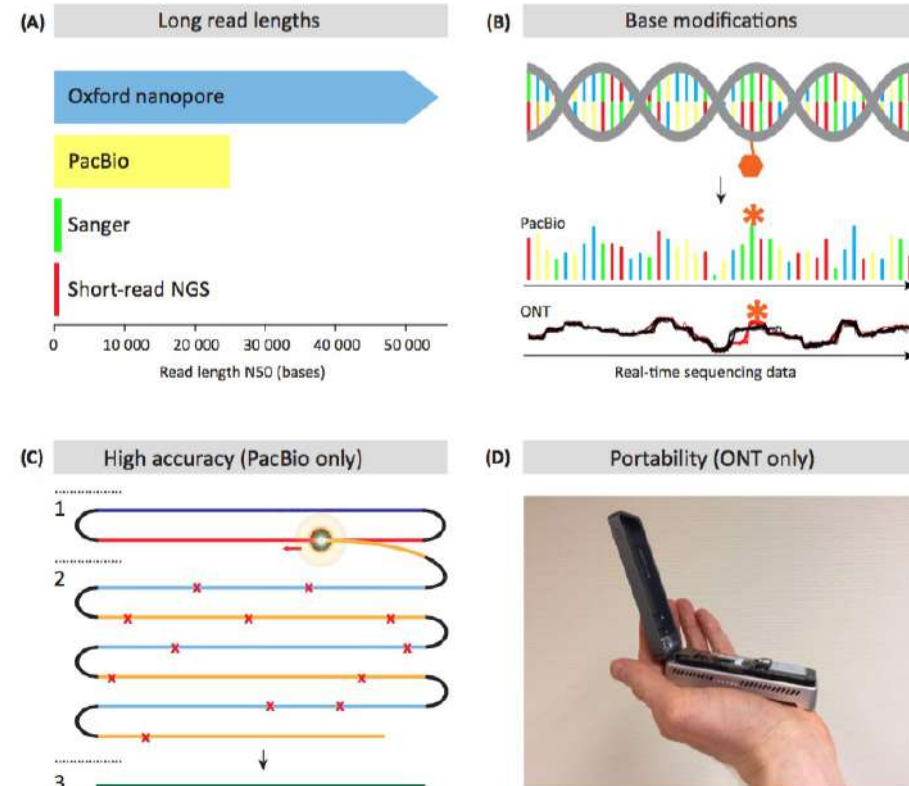
NGI is the first service provider in Europe for this application!



# Long-read sequencing in the clinic



## Key Advantages of Single-Molecule Sequencing



Ameur, Kloosterman & Hestand. *Trends In Biotechnology* 2019

# CRISPR-Cas9 genome editing

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



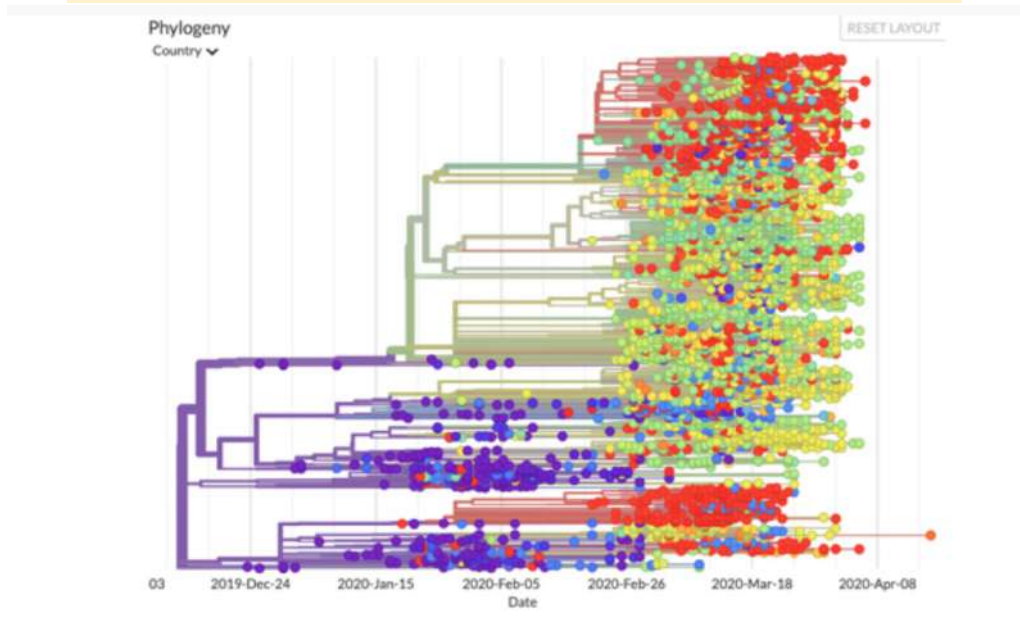
*Höijer et al, Genome Biology 2020; Nature Communications 2022*



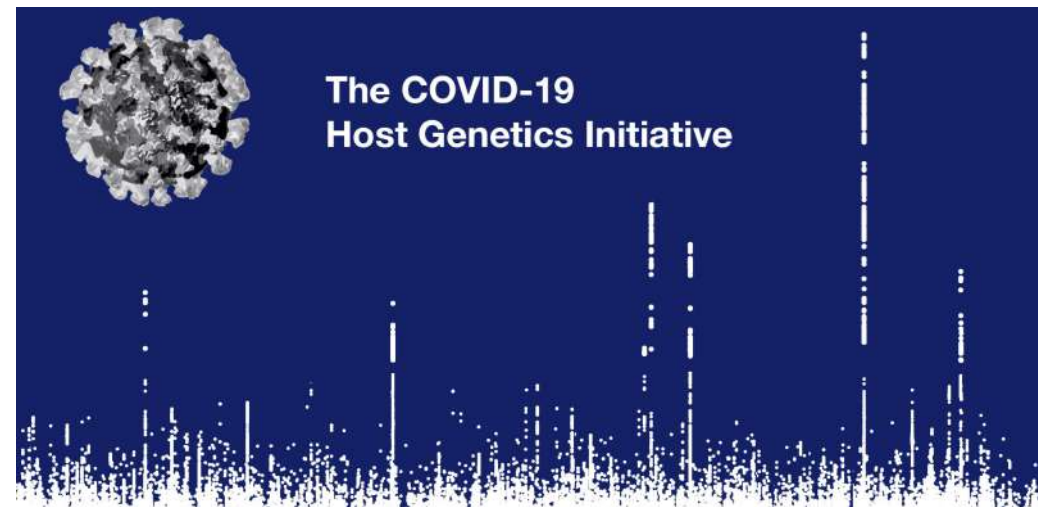
# COVID-19 sequencing

- 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!

CSH Cold Spring Harbor Laboratory **bioRxiv** THE PREPRINT SERVER FOR BIOLOGY HOME | ABOUT Search

bioRxiv is receiving many new papers on coronavirus SARS-CoV-2. A reminder: these are preliminary reports that have not been peer reviewed, and should not be used to guide practice/health-related behavior, or be reported in news media as established information.

New Results [Comment on this paper](#)

### Long-read whole genome analysis of human single cells

Joanna Hård, Jeff E Mold, Jesper Eisfeldt, Christian Tellgren-Roth, Susana Häggqvist, Ignas Bunikis, Orlando Contreras-Lopez, Chen-Shan Chin, Carl-Johan Rubin, Lars Feuk, Jakob Michaëlsson, Adam Ameer

doi: <https://doi.org/10.1101/2021.04.13.439527>

This article is a preprint and has not been certified by peer review [what does this mean?].

**Abstract** Full Text Info/History Metrics [Preview PDF](#)

**Abstract**

With long-read sequencing we have entered an era where individual genomes are routinely assembled to near-completion and where complex genetic variation can efficiently be resolved. Here we demonstrate that long reads can be applied also to study the genomic architecture of individual human cells. Clonally expanded CD8+ T-cells from a human donor

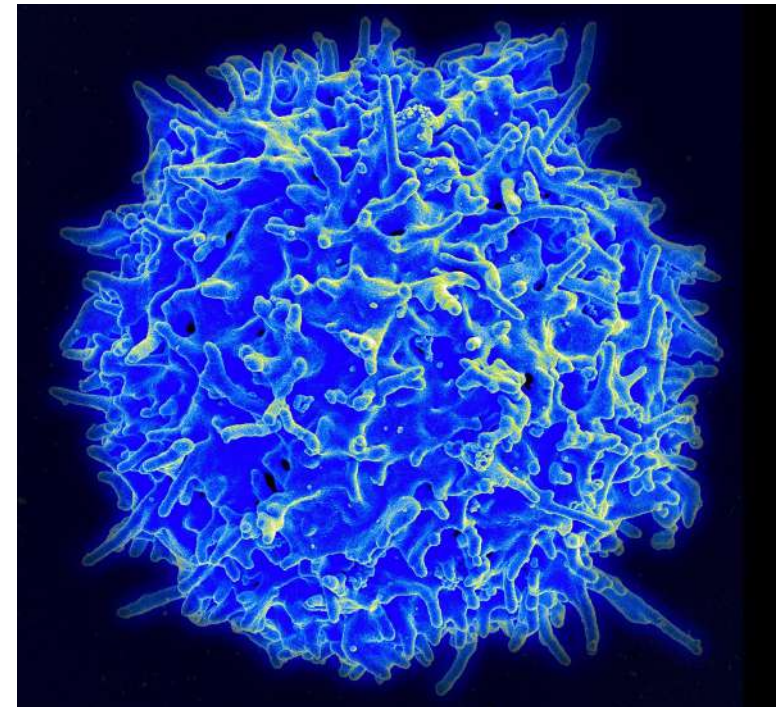


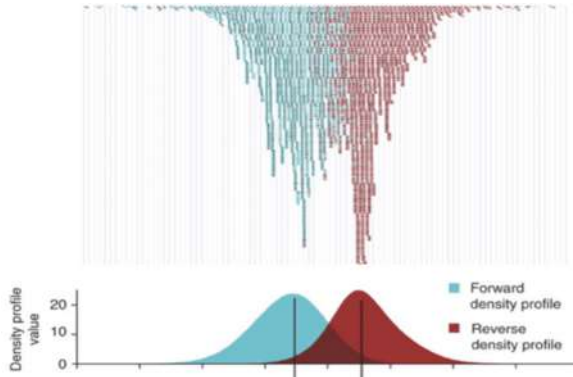
Image from [https://en.wikipedia.org/wiki/T\\_cell](https://en.wikipedia.org/wiki/T_cell)

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

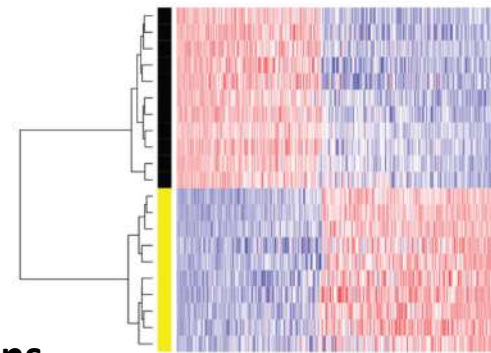
# Many topics I have not covered...



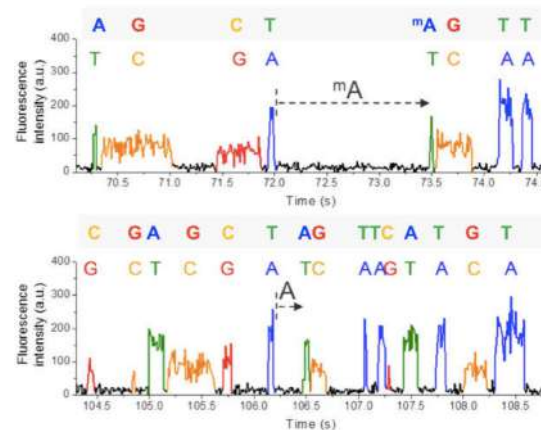
### ChIP-seq analysis



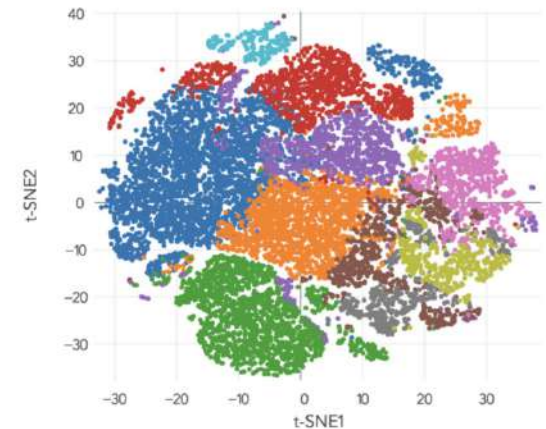
### RNA-Seq



### DNA base modifications



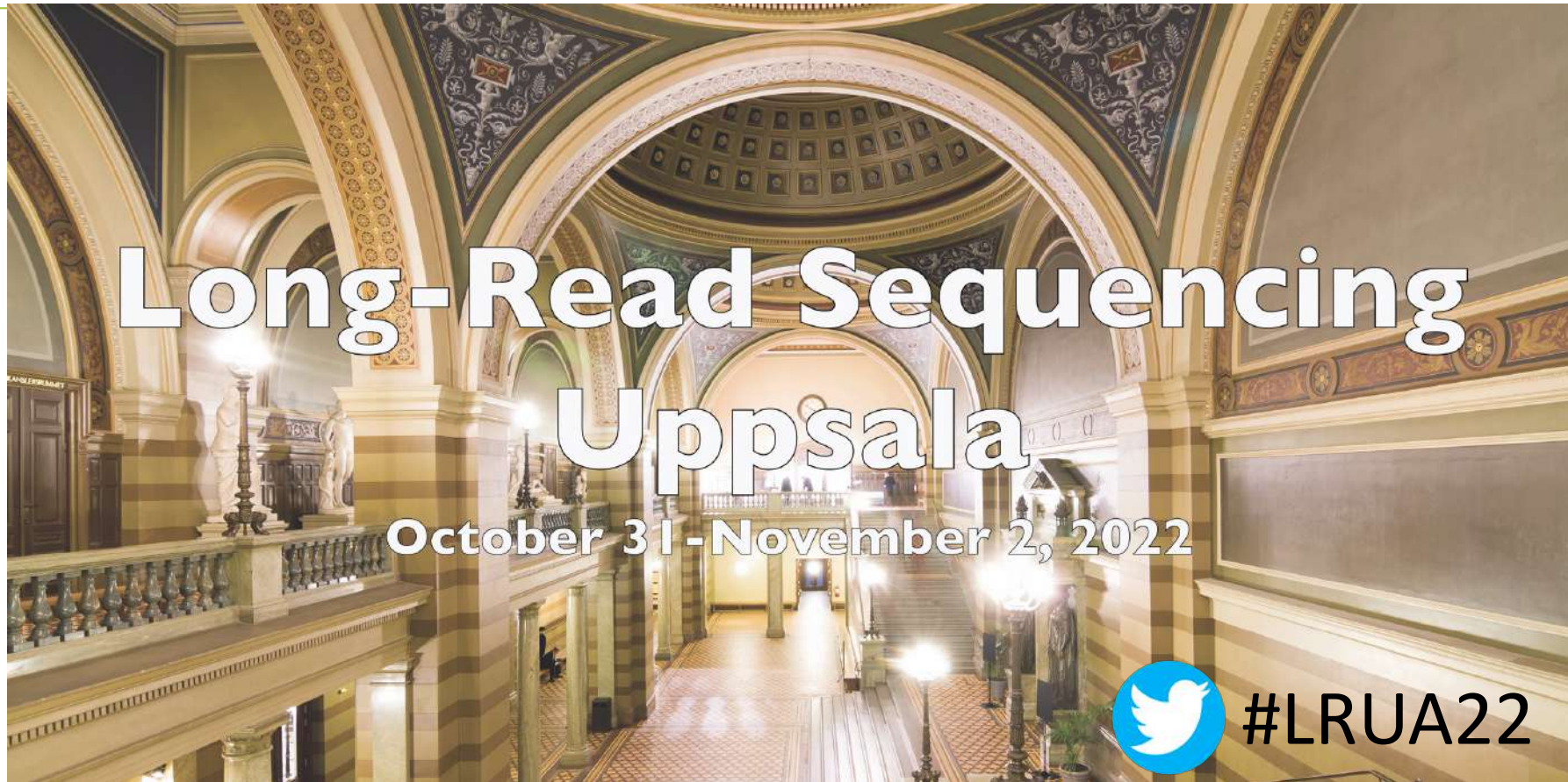
### Single cell RNA-Seq



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



# Long-read Uppsala Meeting 2022!



<https://www.lrua2022.se>

# Thanks for your attention!



Diabetes  
 Alzheimer's disease  
 Whole-genome sequencing  
 Gene therapy  
 Infection screen  
 Whole-transcriptome sequencing  
 Target sequencing  
 Cancer prognosis  
 Gene regulation  
 Crohn's disease  
 Genomics of ageing  
 Exome sequencing  
 Schizophrenia  
 Cancer diagnostics  
 Organ donor matching  
 Gut microflora  
 Gene fusions  
 RNA editing  
 HIV  
 HPV  
 HCV  
 Scoliosis  
 Immune response  
 Monogenic disorders  
 Sudden infant death  
 Cervical cancer  
 Lynch syndrom  
 Leukemia  
 Scoliosis  
 HLA typing  
 Dyslexia  
 MRSA / BRSA screen  
 Sudden cardiac arrest  
 Transcriptional regulation  
 Prenatal diagnostics  
 Muscle dystrophy  
 Individualised cancer therapy  
 and much more...