SciLifeLab

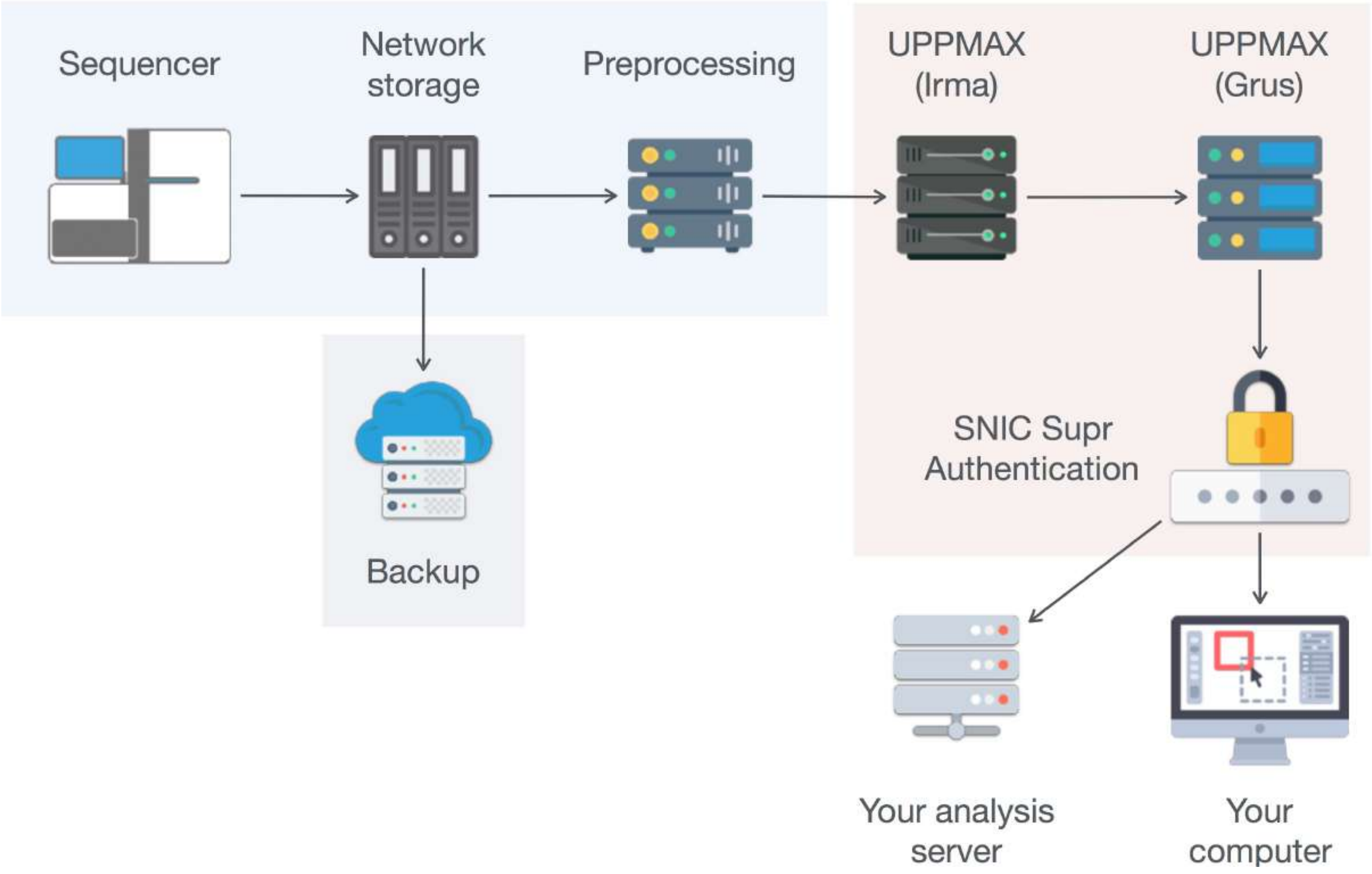
Next Generation Sequencing and Bioinformatics Analysis Pipelines

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

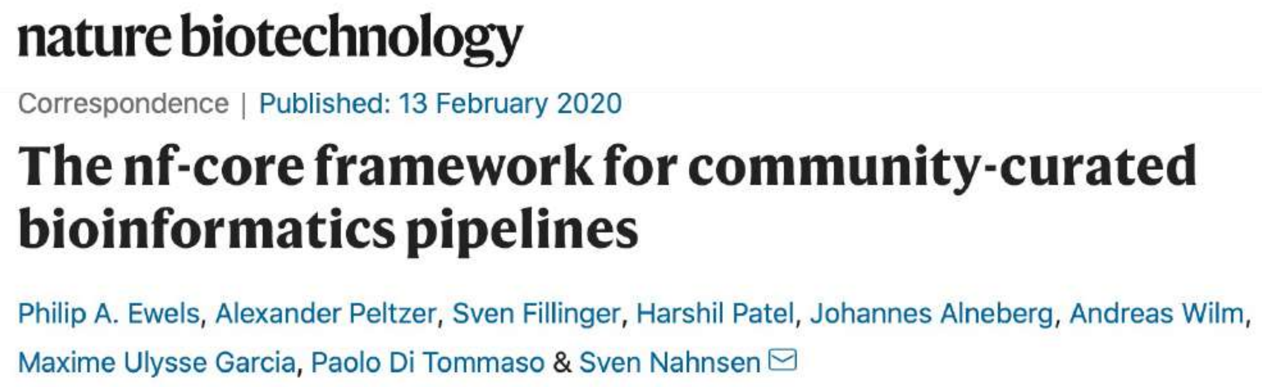
- Initial data analysis for major applications:
  - **Mapping:** Align sequences to a reference genome
  - **SNV calling:** Detect genetic variants
  - **RNA-seq:** Quantify gene expression
  - ***De novo* assembly:** Generate new reference genomes
  - **and more…**
- Analysis requirements: Automated,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

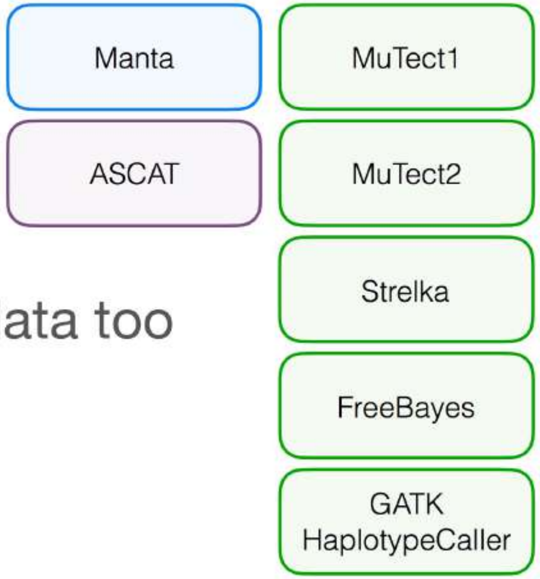
Phil Ewels, NGI Sthlm

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

Sarek



Manta

MuTect1

ASCAT

MuTect2

Strelka

FreeBayes

GATK HaplotypeCaller

NATIONAL GENOMICS INFRASTRUCTURE

NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

Barntumörbanken

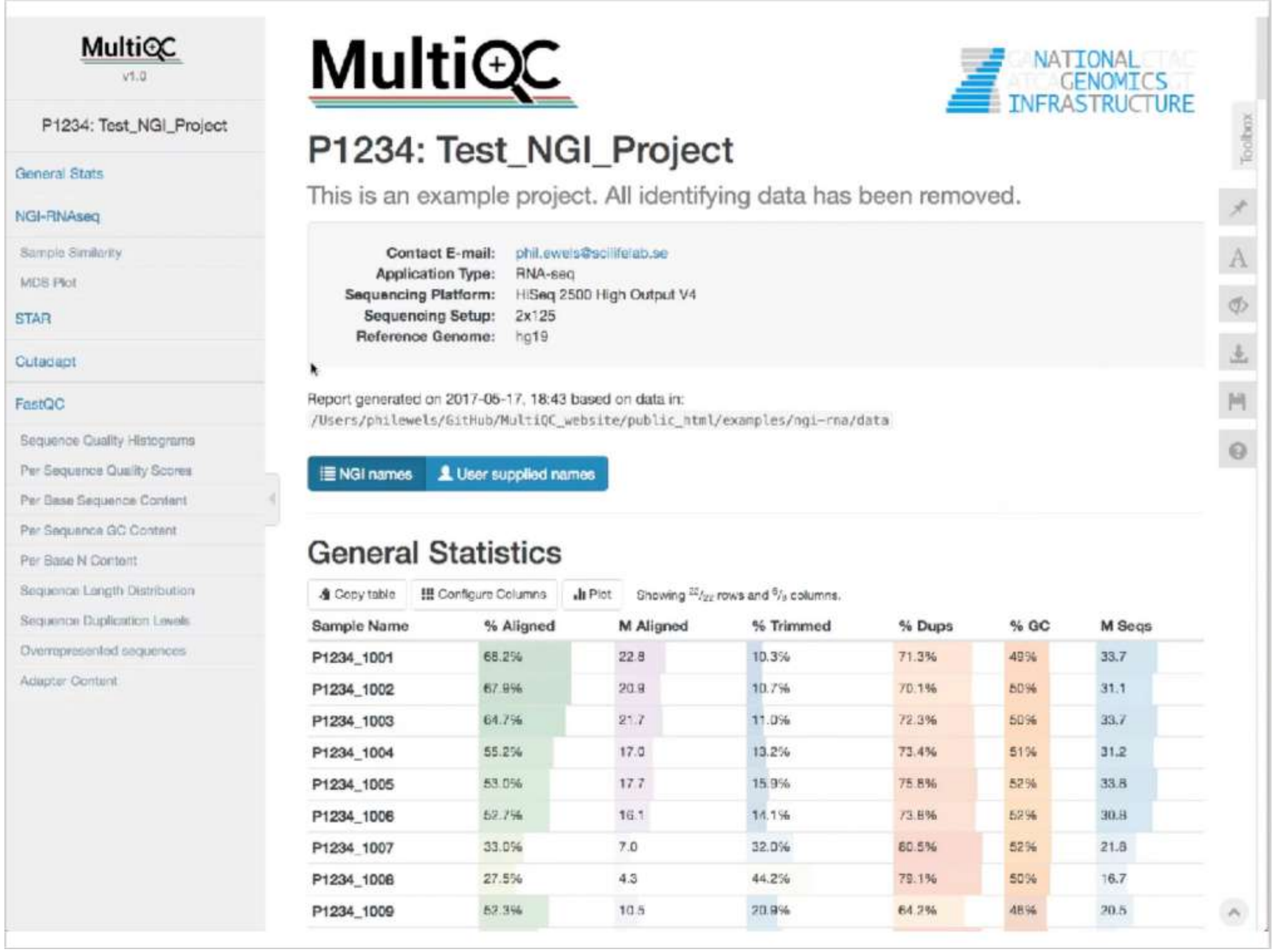
Quality control

- 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)

Multi QC example



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/philewels/GitHub/MultiQC_website/public_html/examples/ngi-rna/data`

General Statistics

Showing 22/22 rows and 6/9 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%	80.5%	52%	21.8
P1234_1008	27.5%	4.3	44.2%	79.1%	50%	16.7
P1234_1009	52.3%	10.5	20.9%	64.2%	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 given access
  - Email sent with project ID and instructions

- Grus is for secure short-term storage only!
  - Requires two-factor authentication

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

NBIS
NATIONAL BIOINFORMATICS
INFRASTRUCTURE SWEDEN

GANATIONALCTAC
ATCAGENOMICSGT
INFRASTRUCTURE

SciLifeLab
Clinical Genomics
PLATFORM

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

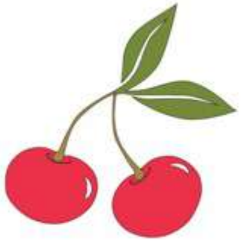
- What can the SweGen dataset be used for?
  - Look up genetic variant frequencies
  - Use as matched controls
  - Study population genetics
  - Study human evolutionary history

High demand for the data from many different groups:

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

Deciding on a cohort to use

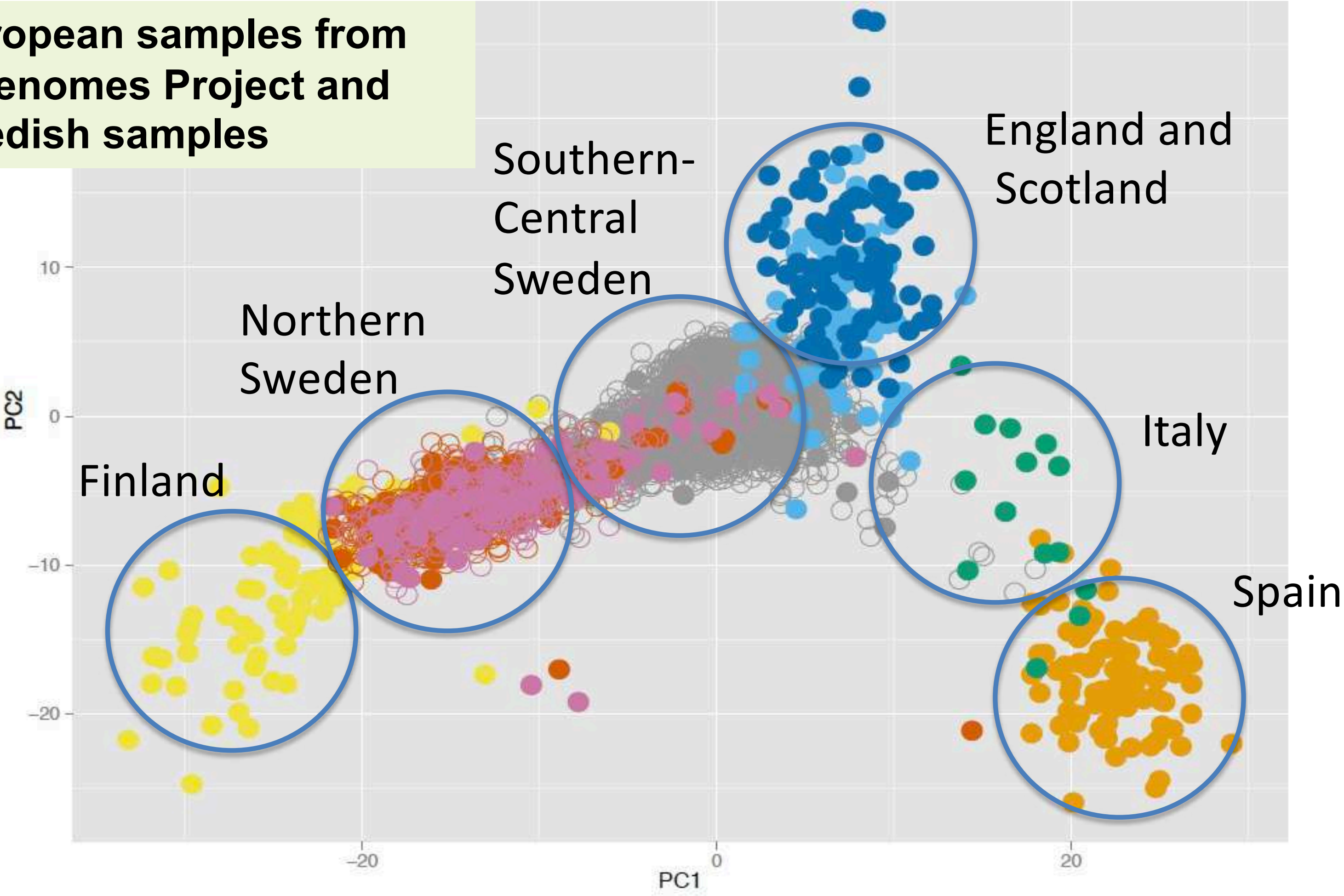
- The Swedish Twin Registry:
  - Inclusion based on twinning
  - Distribution like population density
  - General population-prevalence of disease
  - 10,000 individuals have been analysed with SNP arrays



- **Identify 1,000 individuals based on genetic structure and diversity across Sweden**

Selecting 1000 individuals based on PCA

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



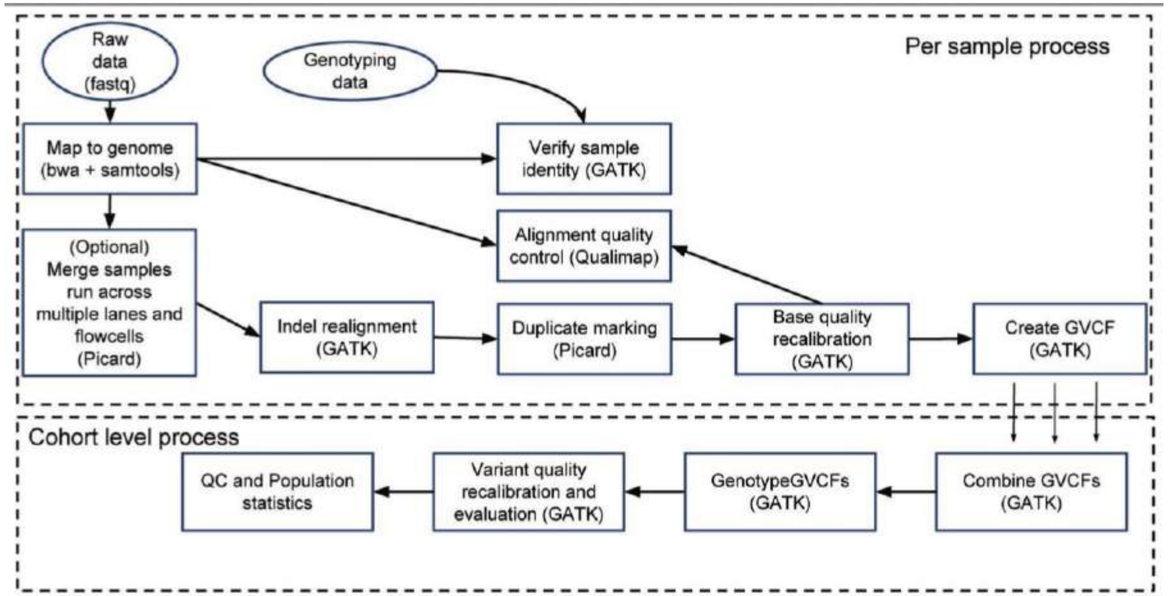
Whole Genome Sequencing

- 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

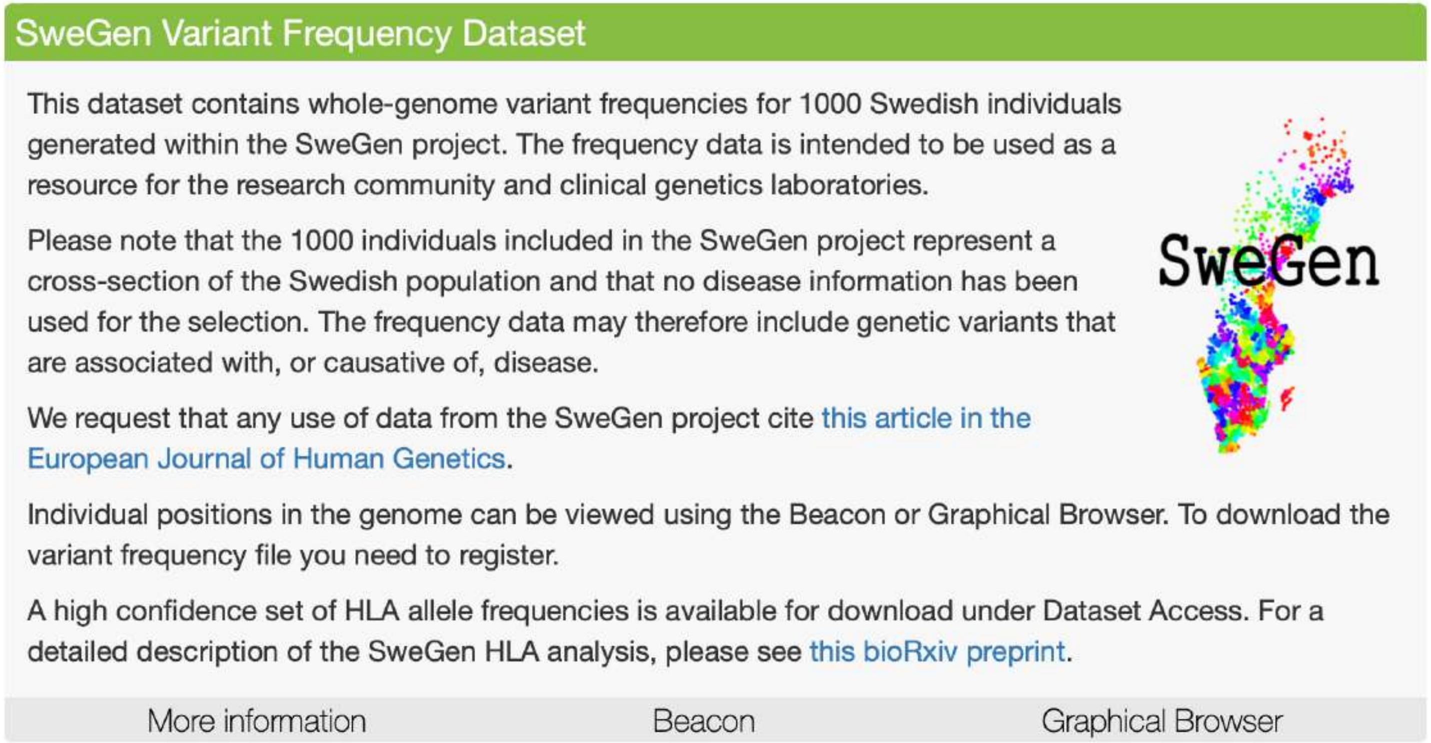
Data analysis pipeline

- NGI pipeline developed for mapping and variant calling


Per sample process
Raw data (fastq)
Genotyping data
Map to genome (bwa + samtools)
Verify sample identity (GATK)
Alignment quality control (Qualimap)
(Optional) Merge samples run across multiple lanes and flowcells (Picard)
Indel realignment (GATK)
Duplicate marking (Picard)
Base quality recalibration (GATK)
Create GVCF (GATK)
Cohort level process
QC and Population statistics
Variant quality recalibration and evaluation (GATK)
GenotypeGVCFs (GATK)
Combine GVCFs (GATK)

- 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.

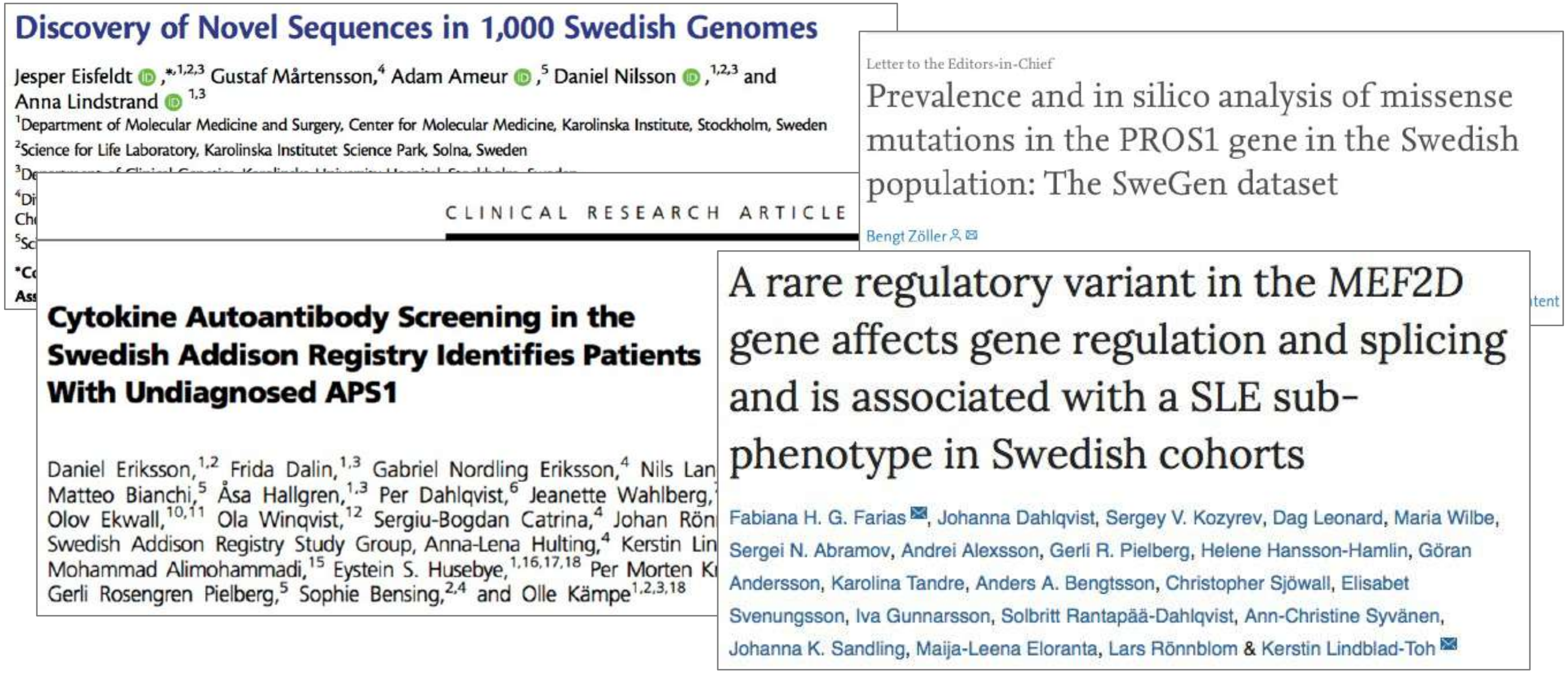
Individual positions in the genome can be viewed using the Beacon or Graphical Browser. To download the variant frequency file you need to register.

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.

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

SweGen: a resource for collaboration

- Over 90 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
[3]De
[4]Di
Ch
[5]Sc
*Co
Ass

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

CLINICAL RESEARCH ARTICLE

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

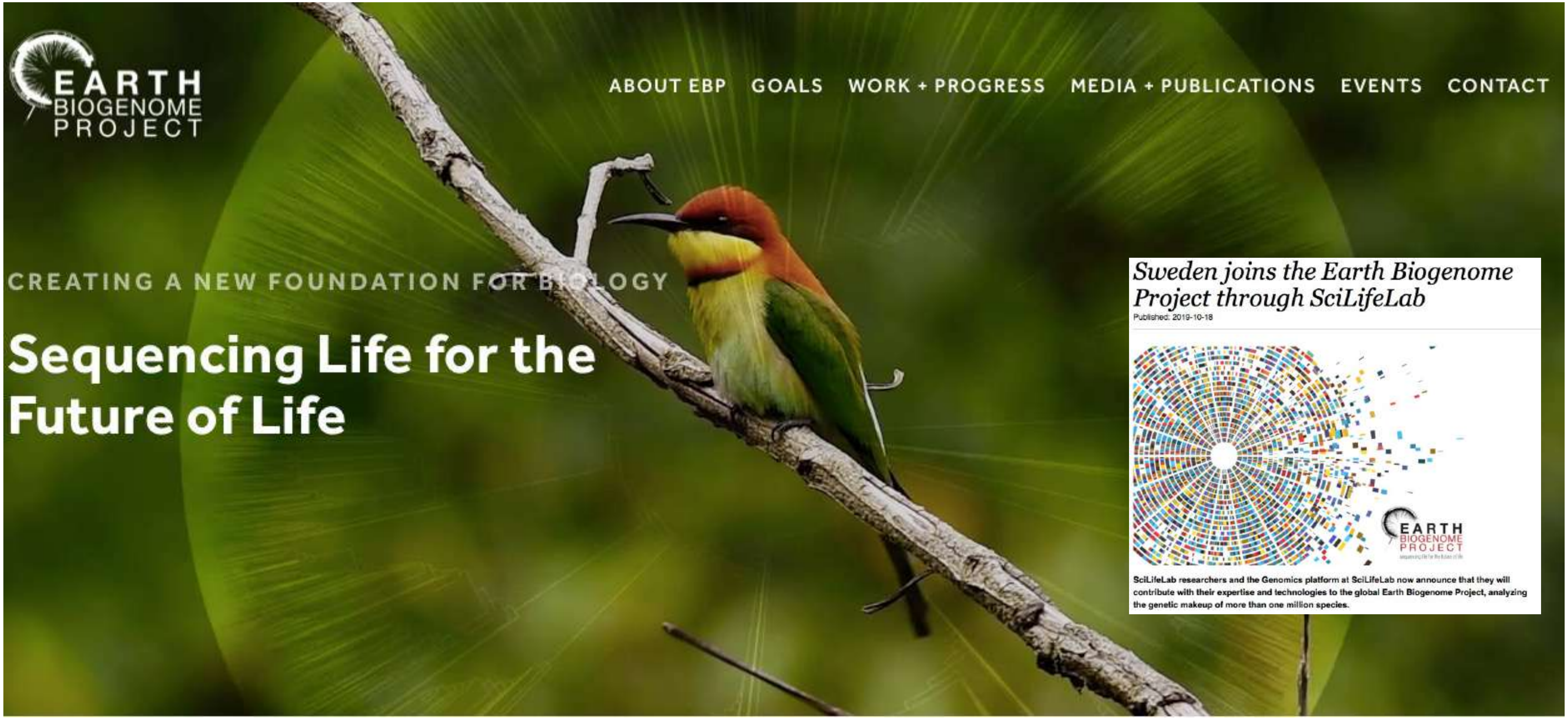
Daniel Eriksson,[1,2] Frida Dalin,[1,3] Gabriel Nordling Eriksson,[4] Nils Lan
Matteo Bianchi,[5] Åsa Hallgren,[1,3] Per Dahlqvist,[6] Jeanette Wahlberg,
Olov Ekwall,[10,11] Ola Winqvist,[12] Sergiu-Bogdan Catrina,[4] Johan Rön
Swedish Addison Registry Study Group, Anna-Lena Hulting,[4] Kerstin Lin
Mohammad 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

Earth Biogenome Project



CREATING A NEW FOUNDATION FOR BIOLOGY

Sequencing Life for the Future of Life

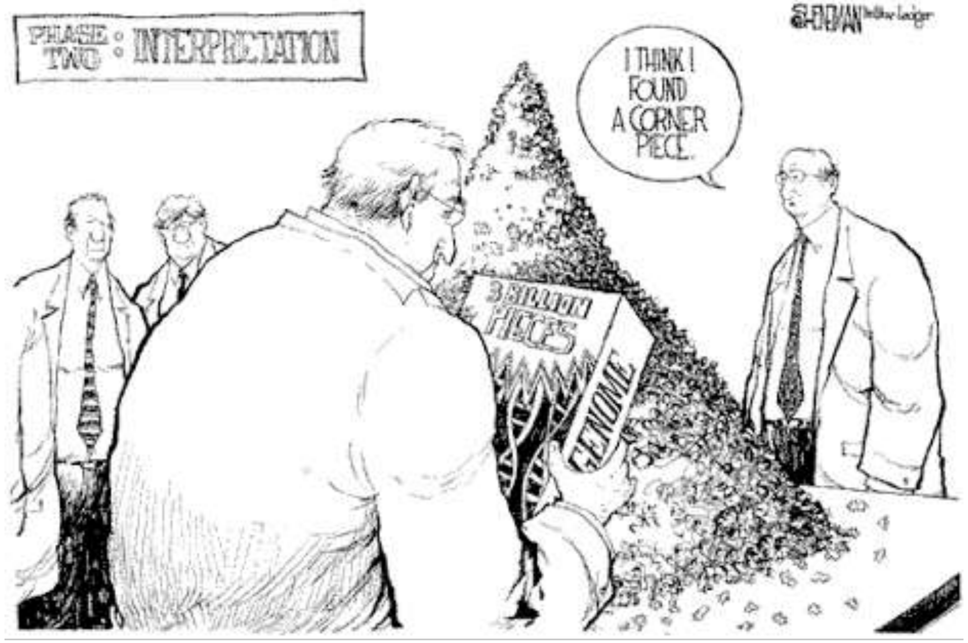
Sweden joins the Earth Biogenome Project through SciLifeLab

Published: 2019-10-18

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

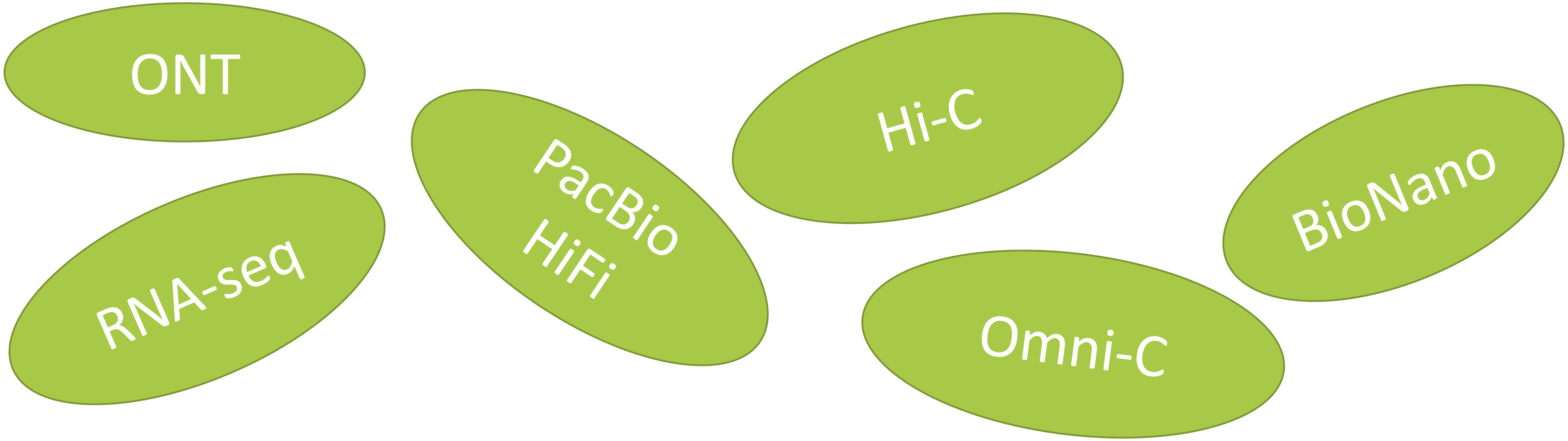
- Over the coming years, many new species will be sequenced
- A combination of different instruments and technologies will be used



- We need good strategies for data analysis and management!

Choice of technology

- Make sure sequencing is done using the best technology combination



ONT

RNA-seq

PacBio HiFi

Hi-C

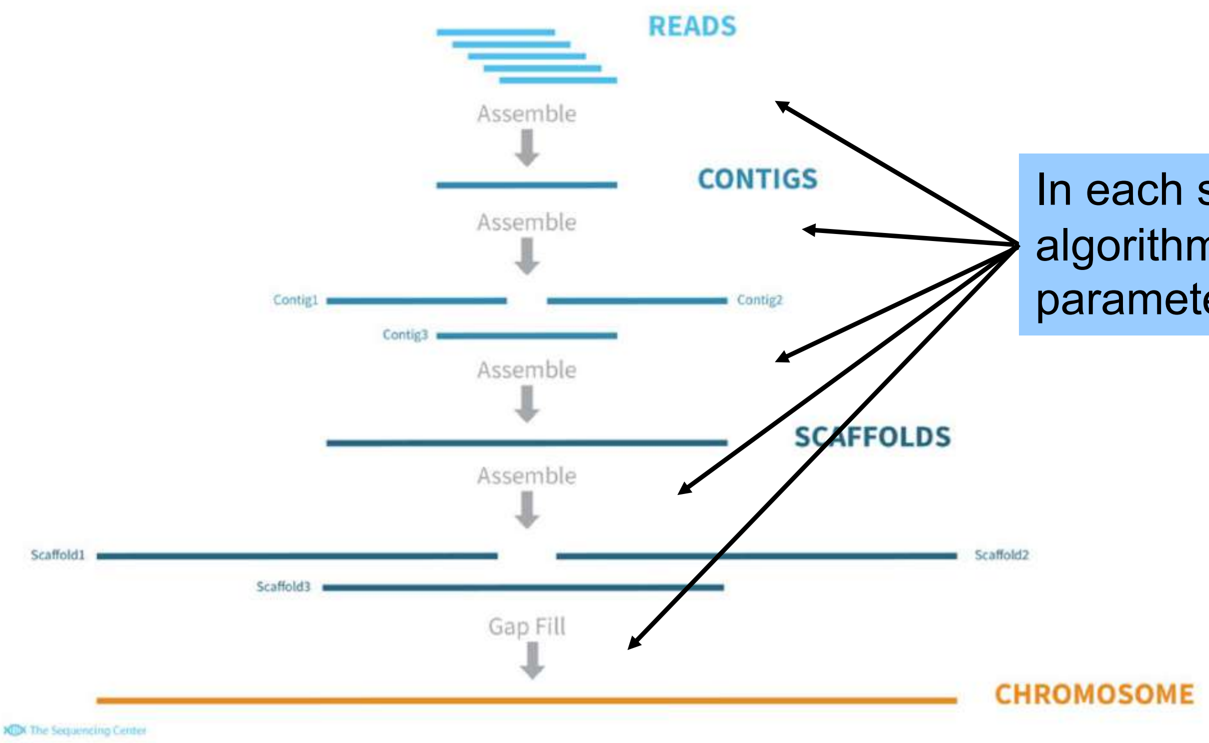
Omni-C

BioNano

- 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



The Sequencing Center

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

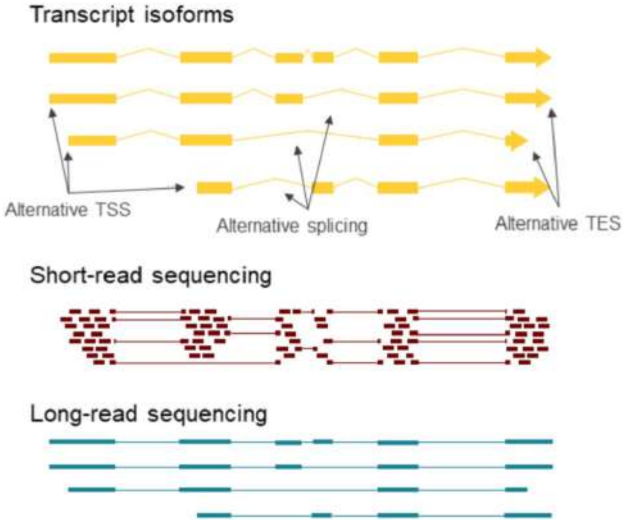
- A challenge for NGI/SciLifeLab is to give best-practice guidelines!

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

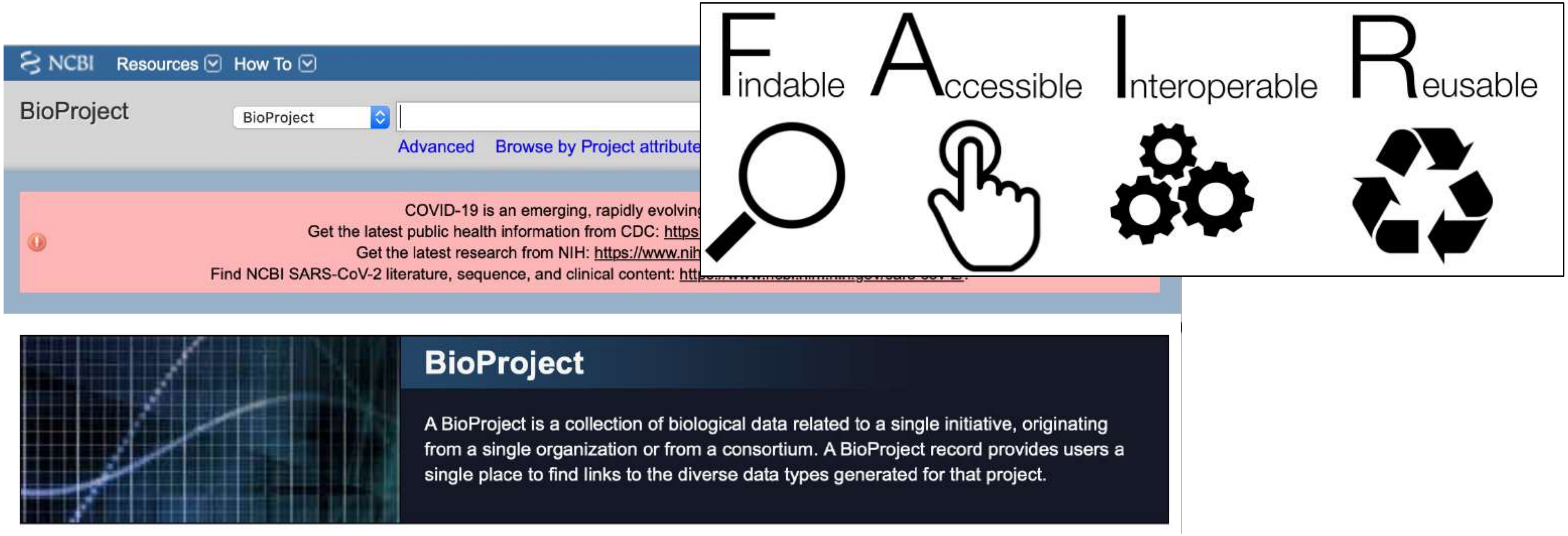
Annotation using RNA-sequencing



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

Data deposition

- Important to deposit the final assembly in public repositories!



COVID-19 is an emerging, rapidly evolvin

Get the latest public health information from CDC: https

Get the latest research from NIH: https://www.nih

Find NCBI SARS-CoV-2 literature, sequence, and clinical content: htt

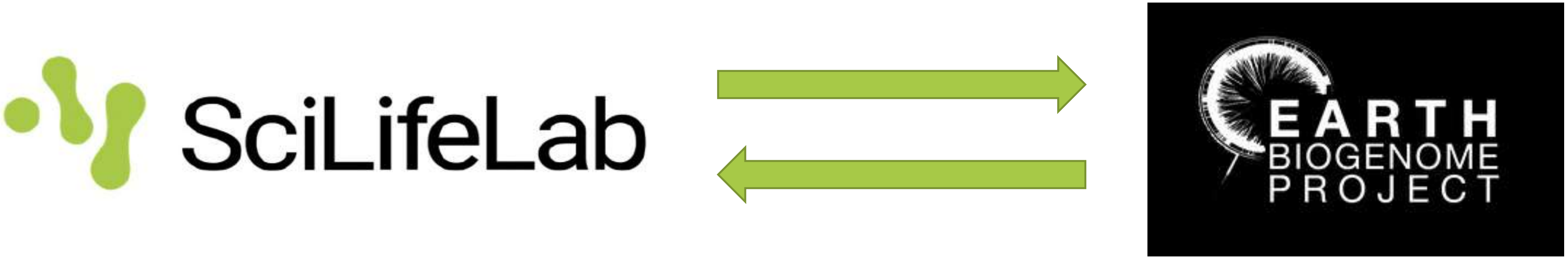
BioProject

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.

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

EBP – A collaborative project

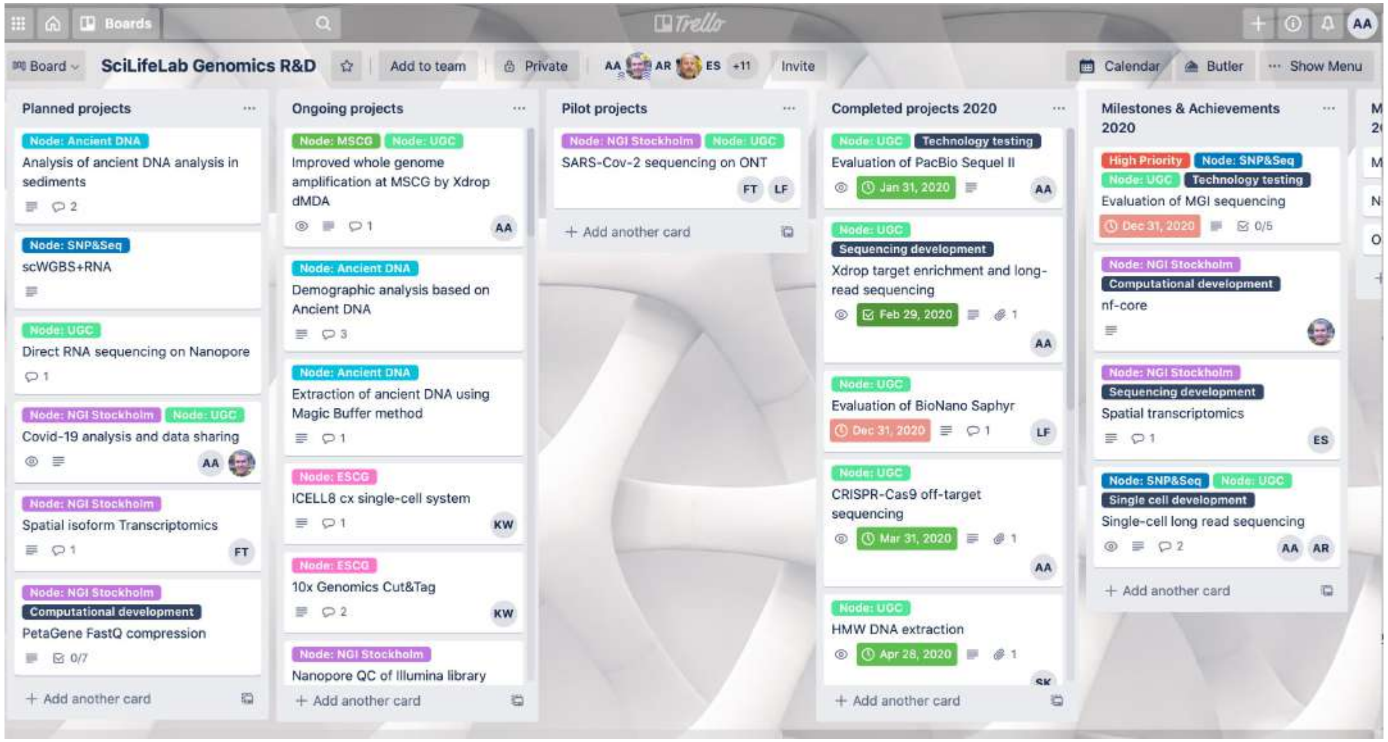
- A lot of challenges ahead of us to establish EBP analyses in Sweden
- … but the good news is that this is a community effort



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

Other R&D activities at NGI

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



SciLifeLab Genomics R&D

Planned projects

- Node: Ancient DNA — Analysis of ancient DNA analysis in sediments — 2
- Node: SNP&Seq — scWGBS+RNA
- Node: UGC — Direct RNA sequencing on Nanopore — 1
- Node: NGI Stockholm, Node: UGC — Covid-19 analysis and data sharing — AA
- Node: NGI Stockholm — Spatial isoform Transcriptomics — 1 — FT
- Node: NGI Stockholm, Computational development — PetaGene FastQ compression — 0/7
- + Add another card

Ongoing projects

- Node: MSCG, Node: UGC — Improved whole genome amplification at MSCG by Xdrop dMDA — 1 — AA
- Node: Ancient DNA — Demographic analysis based on Ancient DNA — 3
- Node: Ancient DNA — Extraction of ancient DNA using Magic Buffer method — 1
- Node: ESCG — ICELL8 cx single-cell system — 1 — KW
- Node: ESCG — 10x Genomics Cut&Tag — 2 — KW
- Node: NGI Stockholm — Nanopore QC of Illumina library
- + Add another card

Pilot projects

- Node: NGI Stockholm, Node: UGC — SARS-Cov-2 sequencing on ONT — FT LF
- + Add another card

Completed projects 2020

- Node: UGC, Technology testing — Evaluation of PacBio Sequel II — Jan 31, 2020 — AA
- Node: UGC, Sequencing development — Xdrop target enrichment and long-read sequencing — Feb 29, 2020 — 1 — AA
- Node: UGC — Evaluation of BioNano Saphyr — Dec 31, 2020 — 1 — LF
- Node: UGC — CRISPR-Cas9 off-target sequencing — Mar 31, 2020 — 1 — AA
- Node: UGC — HMW DNA extraction — Apr 28, 2020 — 1 — SK
- + Add another card

Milestones & Achievements 2020

- High Priority, Node: SNP&Seq, Node: UGC, Technology testing — Evaluation of MGI sequencing — Dec 31, 2020 — 0/5
- Node: NGI Stockholm, Computational development — nf-core
- Node: NGI Stockholm, Sequencing development — Spatial transcriptomics — 1 — ES
- Node: SNP&Seq, Node: UGC, Single cell development — Single-cell long read sequencing — 2 — AA AR
- + Add another card

Examples of technologies evaluated during 2020

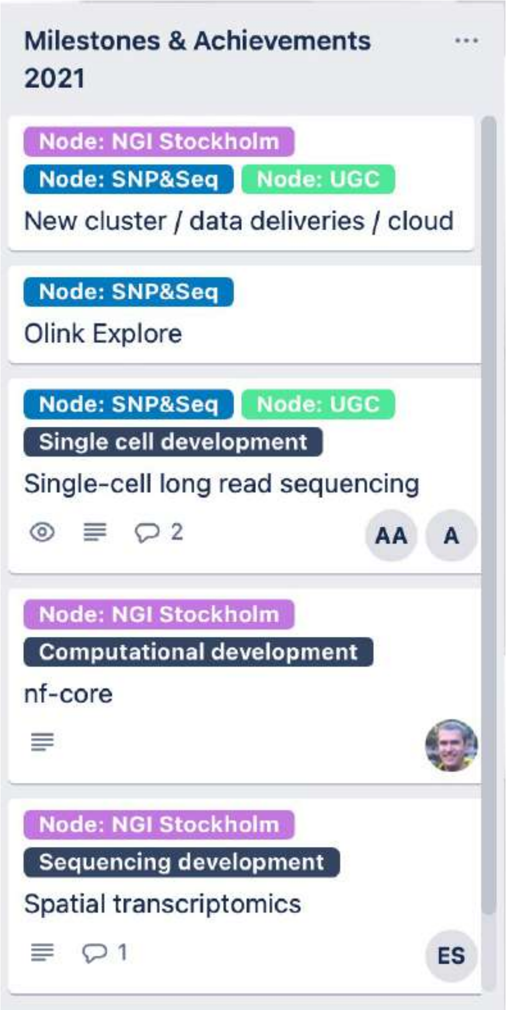
MGI sequencing

Mission Bio – Tapestri

PacBio Sequel IIe

“Milestones and Achievements” 2021

- **New cluster and data delivery system**
- **Olink Explore**
- **Further development of NF-core**
- **Single-cell long read sequencing**
- **Spatial transcriptomics**



Milestones & Achievements 2021

Node: NGI Stockholm
Node: SNP&Seq Node: UGC
New cluster / data deliveries / cloud

Node: SNP&Seq
Olink Explore

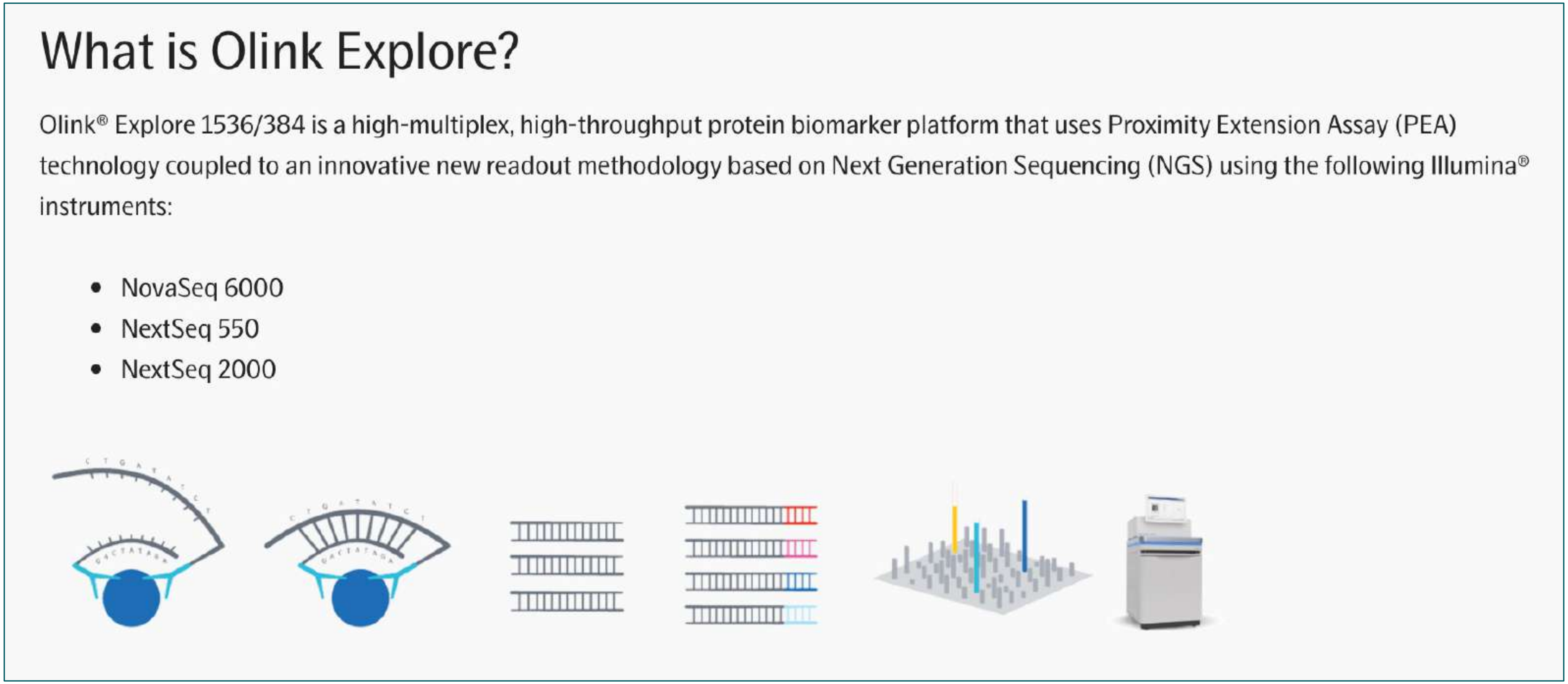
Node: SNP&Seq Node: UGC
Single cell development
Single-cell long read sequencing

Node: NGI Stockholm
Computational development
nf-core

Node: NGI Stockholm
Sequencing development
Spatial transcriptomics

If you have other R&D ideas then let us know!

Olink Explore: measuring proteins by NGS



What is Olink Explore?

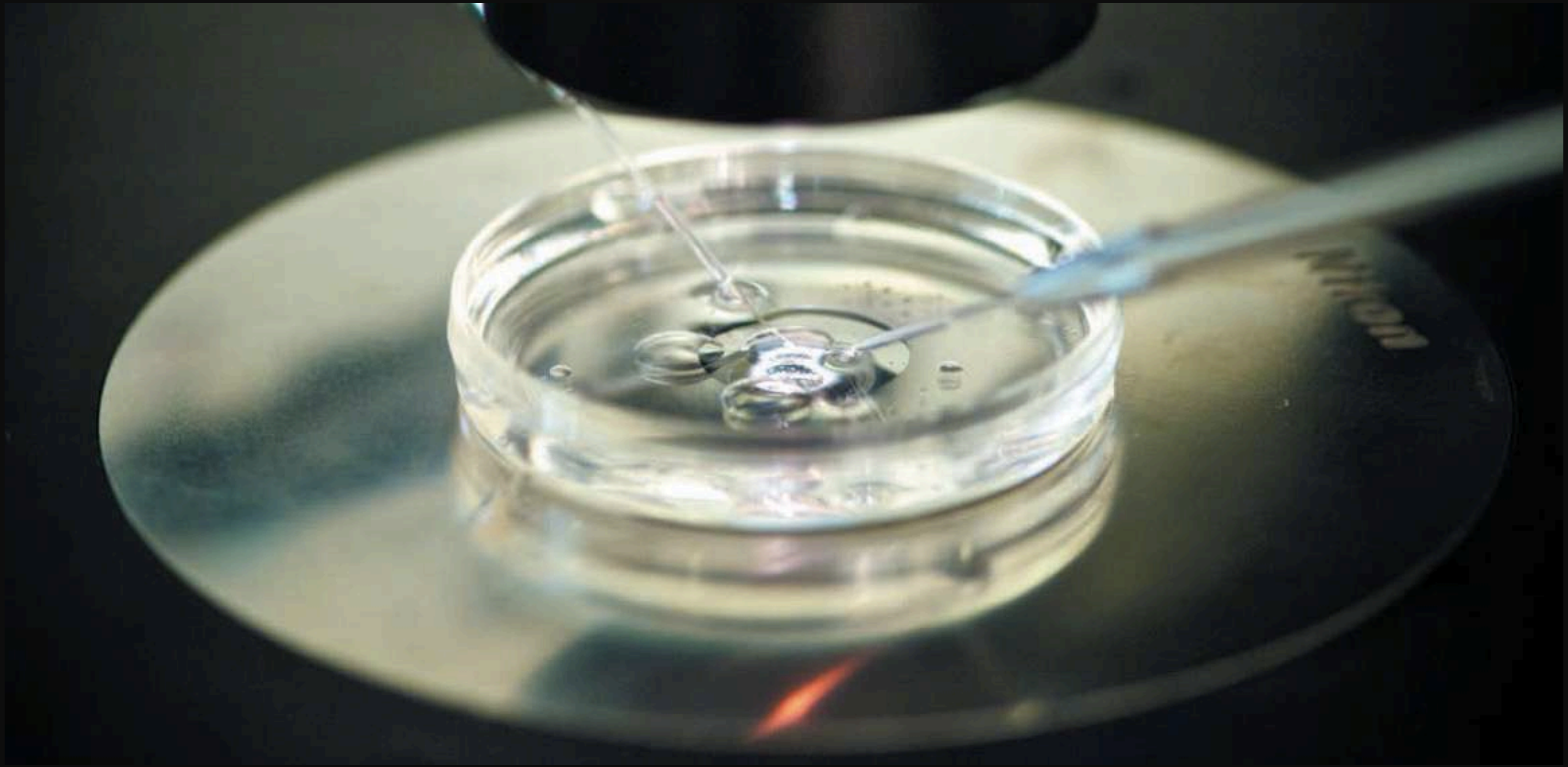
Olink® Explore 1536/384 is a high-multiplex, high-throughput protein biomarker platform that uses Proximity Extension Assay (PEA) technology coupled to an innovative new readout methodology based on Next Generation Sequencing (NGS) using the following Illumina® instruments:

- NovaSeq 6000
- NextSeq 550
- NextSeq 2000

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

CRISPR-Cas9 genome editing

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

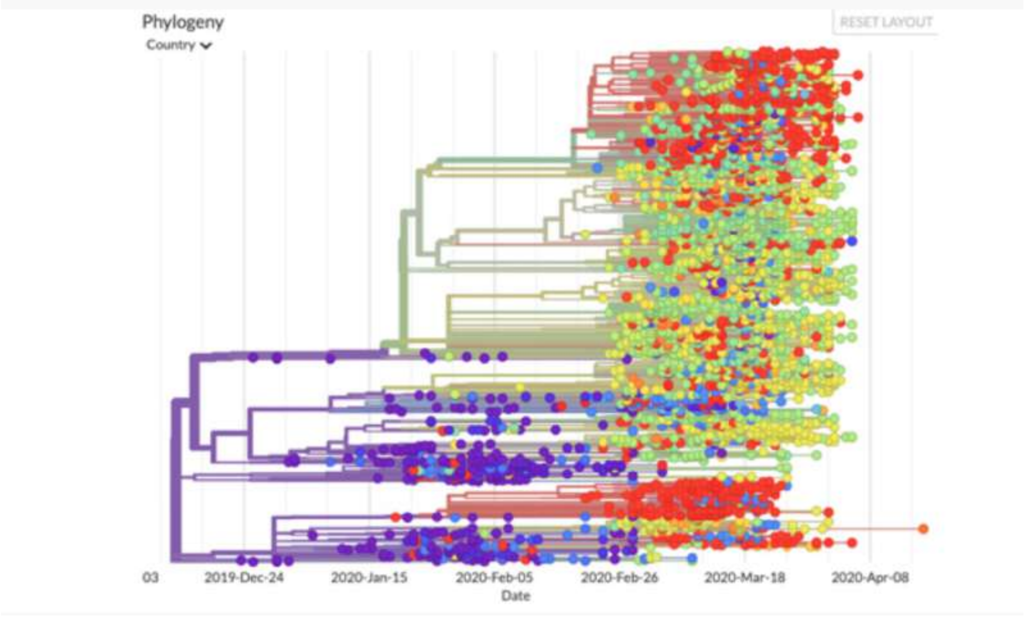


Höijer et al, Genome Biology 2020

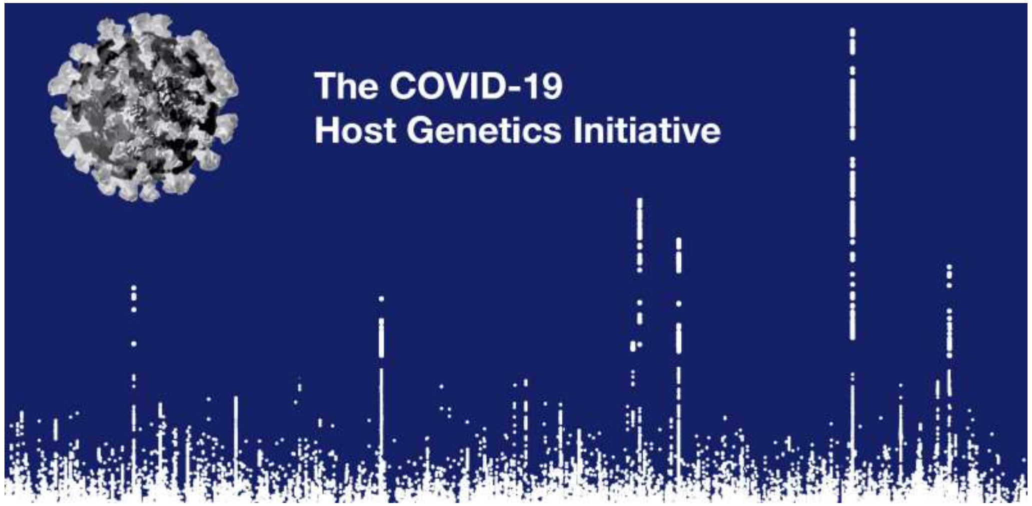
COVID-19 sequencing

- We are evaluating different methods and sequencing protocols…

SARS-Cov-2 sequencing (mutations)



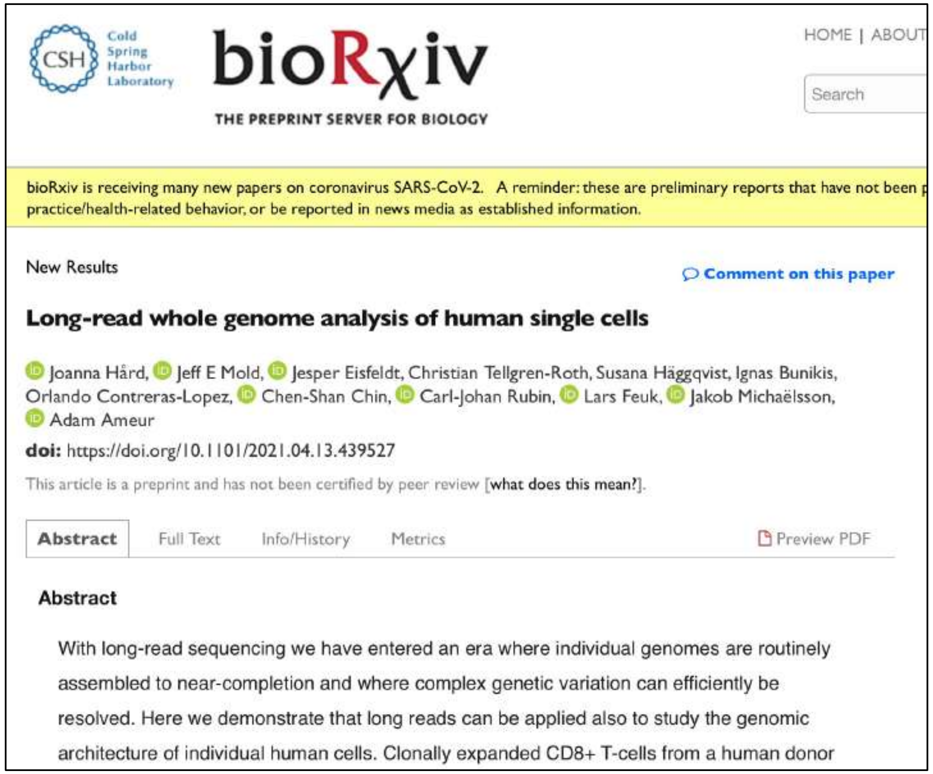
Human re-sequencing (host response)



- If you are interested to discuss COVID-19 projects, then let us know!

Long-read single-cell sequencing

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


Cold Spring Harbor Laboratory
CSH
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
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 Ameur

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

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

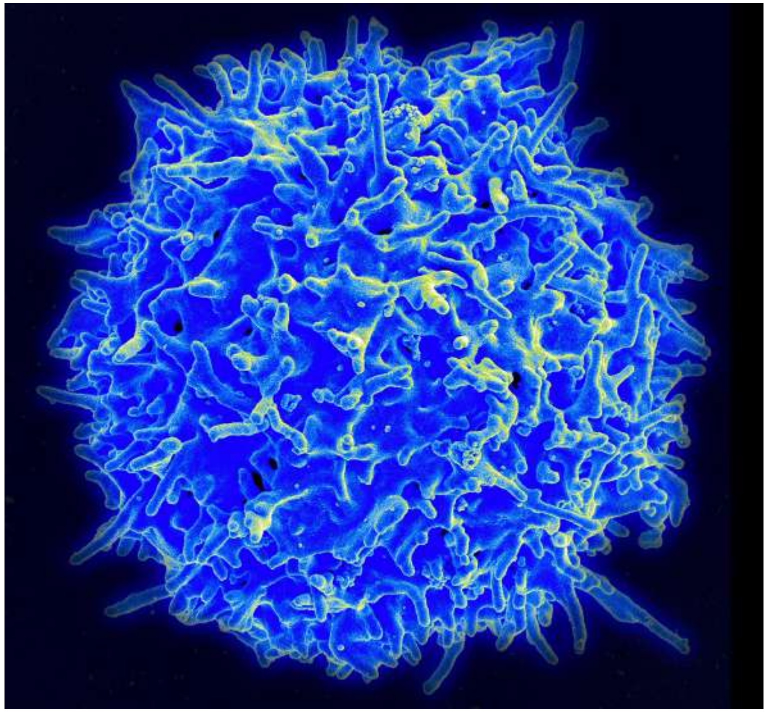
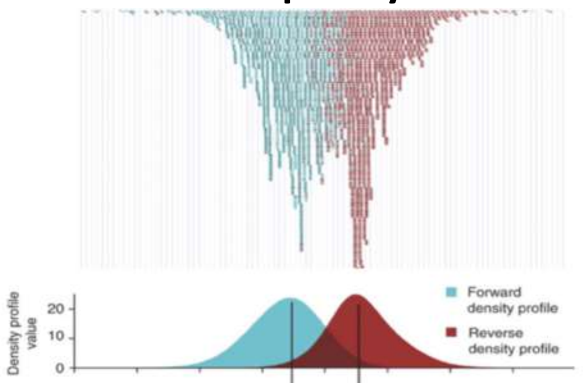


Image from https://en.wikipedia.org/wiki/T_cell

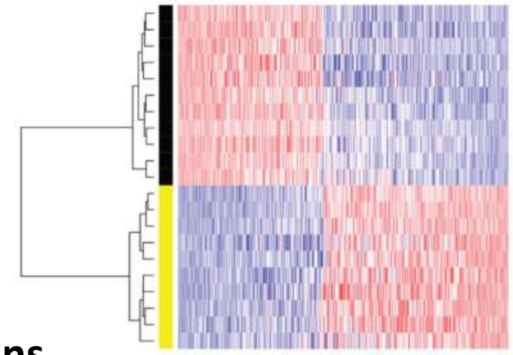
- We are also testing new protocols for long-read single-cell RNA sequencing

Many topics I have not covered…

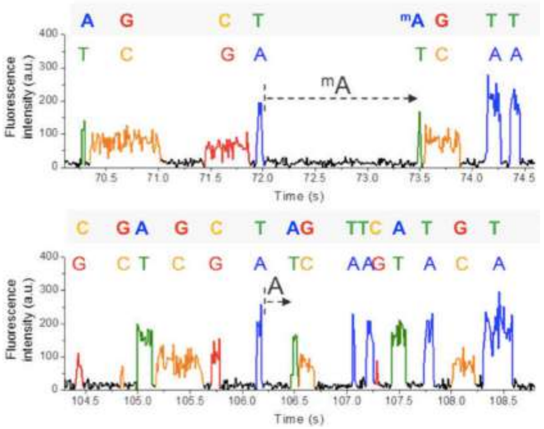
ChIP-seq analysis



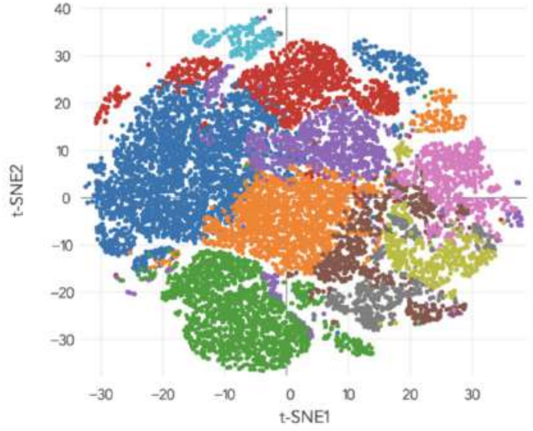
RNA-Seq



DNA base modifications



Single cell RNA-Seq



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

Thanks for your attention!