NGS: technologies and challenges

Johanna Lagensjö, Project coordinator & Head of laboratory operations, NGI-Uppsala
Adam Ameur, Associate professor and senior bioinformatician, NGI-Uppsala

Today we will talk about

- Genomics Platforms and sequencing services at NGI, SciLifeLab
- History and current status of technologies for sequencing
- NGS applications and technologies
- NGS challanges and sample requirements
- Data analysis pipelines, R&D and strategic projects

Service areas of SciLifeLab

- Bioinformatics
- Bioimaging and Molecular Structure
- Chemical Biology and Genome Engineering
- Drug Discovery
- Diagnostics
- Genomics
- Metabolomics
- Single Cell Biology
- Spatial Omics
- Proteomics

Across all service areas: dedicated staff scientists that can offer support **throughout the experimental process** – from study design to data handling

SciLifeLab Genomics

RELEVANT UNITS / GENOMICS

National Genomics Infrastructure (NGI)

The National Genomics Infrastructure (NGI) provides services for next generation sequencing and SNP genotyping on all scales using a comprehensive range of modern (...)

Learn More

Ancient DNA

Use cleanroom labs and specialized molecular genetics techniques to extract, make libraries, sequence and analyze DNA in ancient and/or degraded biological material.

Learn More

Clinical Genomics

Develops and provides clinical genetic tests using state-of-the-art genomic methods, such as next-generation sequencing, for translational research and healthcare.

Learn More

Eukaryotic Single Cell Genomics

Provides service for high-throughput single cell genomics analysis

Learn More

Microbial Single Cell Genomics

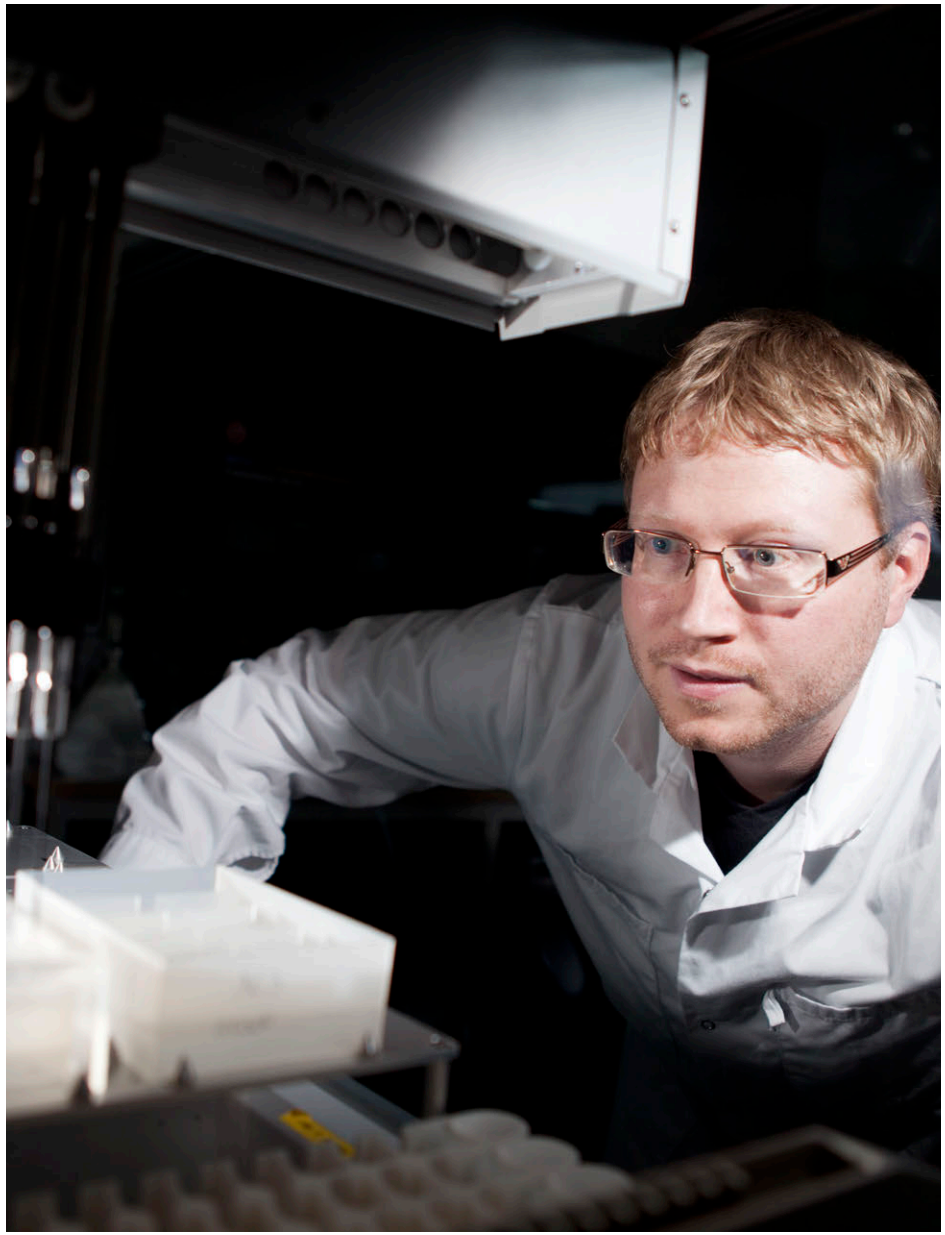
Provides streamlined single-cell sorting, lysis, whole-genome amplification and screening of individual microbial cells, as well as whole genome and targeted gene sequencing (...)

Learn More

National Bioinformatics Infrastructure (NBIS)

Provides custom-tailored support with data analysis, computational tools, systems development and training.

Learn More



What is NGI?

NGI provides access to technology for massively parallel/next generation DNA sequencing, genotyping and associated bioinformatics support

NGI Platform organisation

Tuuli Lappainen
Platform Director
Professor KTH

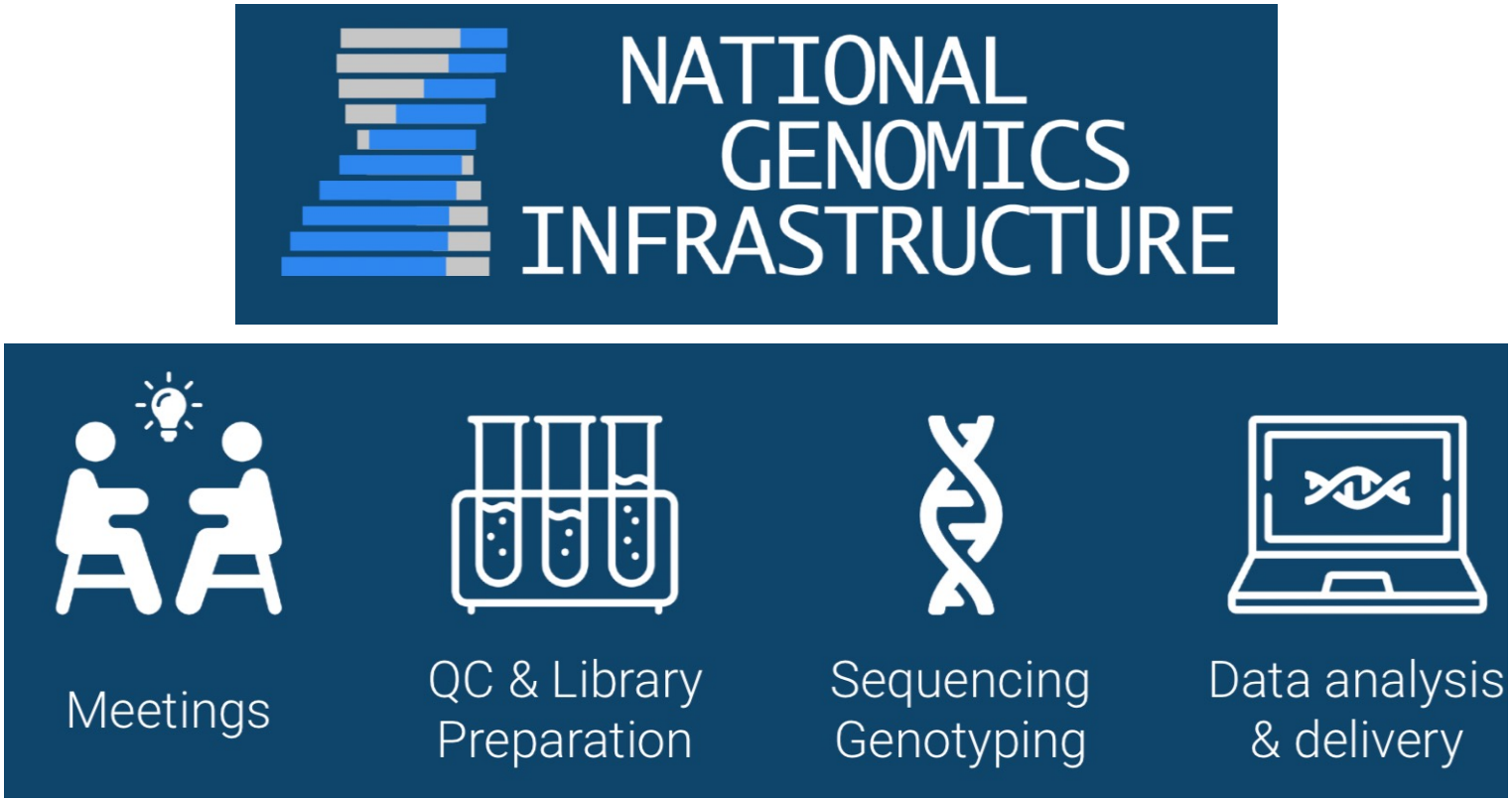
Lars Feuk
Platform Co-Director
Professor UU

NGI-Uppsala SNP&SEQ Technology platform

NGI-Uppsala Uppsala Genome Center

NGI-Stockholm

Project workflow



NGI 2022

Projects
- Assemblies of high-quality reference genomes
- Human genome variation analyses
- Transcriptome profiling
- Single-cell sequencing and much more

Amount of sequenced base pairs
- 802 Tbp – short reads
- 60 Tbp – long reads
- 23.5 B – genotypes

Technology development
- Evaluation of new protocols, applications, bioinformatics tools and sequencing methods
- Methodological developments in spatial and single-cell transcriptomics technologies

Education and Outreach
- Teaching at courses from undergraduate to PhD level
- Participating in national and international conferences
- Webinars, workshops and hackathons

92k | 1030 | 621 | 0.9 Pbp | 208 | 27 | 514 | 19 | 3292

Samples
- All types of sample sources: from environment, lab cultured, biobank, etc
- All types of organisms: microbes, plants, insects, mammals, ...

Support meetings
- Experimental design
- Advising on sample preparations
- Optimizing sequencing setup
- Guidelines for further data analysis

Publications
- Contribution to a number of articles in high impact journals such as Nature, Cell, Science, Nature Biotechnology, Nature Genetics, Nature Neuroscience, etc.

Users
- Unique project PIs from more than 18 different universities, institutes, healthcare and industry companies used NGI services in 2021

Communication tickets
- 42800 ticket updates
- 97% satisfaction score

NGS technologies at NGI

Short-reads

illumina

Short-reads

ion torrent

by life technologies™

Long -reads

PACBIO®

Oxford NANOPORE Technologies

Sequencing instruments at NGI

Short read NGS

High troughput, low cost per base

3x NovaSeq X Plus – **New!**

5 x Illumina NovaSeq

4 x Illumina MiSeq

1 x Illumina NextSeq

1 x Illumina iSeq

1 x Thermo Fisher IonS5

Long read NGS

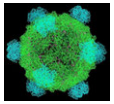
Very long reads, lower troughput

1 x PacBio Revio – New!

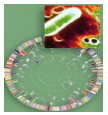
1 x PacBio Sequel IIe

1 x Oxford Nanopore-PromethION

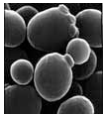
History and current status of sequencing



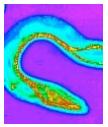
First genome: virus ϕ X 174 - 5 368 bp (1977)



First organism: *Haemophilus influenzae* - 1.5 Mb (1995)



First eukaryote: *Saccharomyces cerevisiae* - 12.4 Mb (1996)



First multicellular organism: *Cenorhabditis elegans* - 100 Mb (1998-2002)

First plant: *Arabidopsis thaliana* - 157 Mb (2000)

First human genome- 3Gb (2003)

An interessting comparison

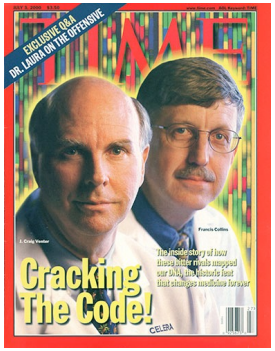
Human genome project (HUGO, 2003)
Sanger Sequencing
2.7 Billion USD

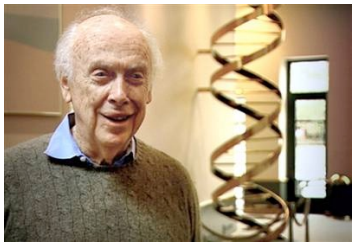
Craig Venter's Genome
Sanger Sequencing
70 Million USD

James Watson's Genome
454 pyro sequencing (Roche)
2 Million USD

Yesterday's genome
NovaSeq 6000 (Illumina)
~1 000 USD

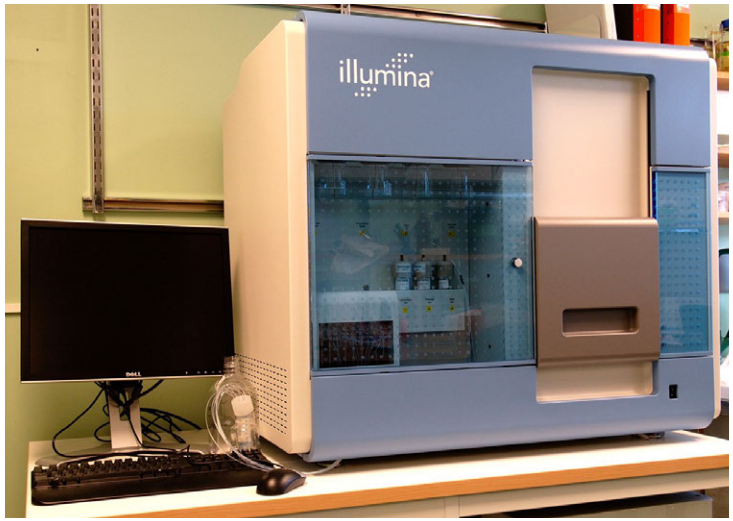
Today's genome
NovaSeq X (Illumina)
~600 USD





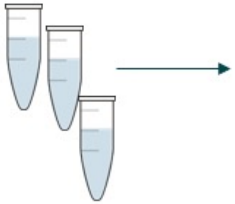
15 years of Illumina sequencing at NGI

2007: Installation of Illumina GA



2023: Arrival of NovaSeq X Plus

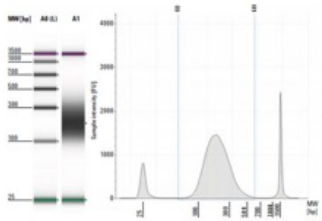
Workflow, Illumina sequencing



Quality control of template (DNA, RNA, amplified regions)

Sequencing library preparation
Diffrent protocols available, hours to many days

Quality control of sequencing library



Sequencing

Data processing
Raw data files converted to a readable format (fastq-files), demultiplexing

Data analysis
Read alignment to a reference
DNA-seq: Do we find mutations?
RNA-seq: Changes in gene expression?
De novo assembly of long reads

Short reads, Illumina sequencing

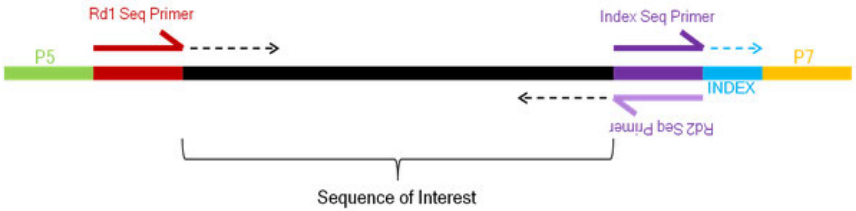
36-300 bp, paired end sequcning
150 Mb-16 Tb per run
12 hours - 3 days

Whole genome sequencing, any size
Whole genome sequencing, human
Exome
Transcriptomes
Target genes and panels
Amplicons (up to 500 bp)
ChIP-sequencing
Methylome
RAD-sequencing
Metagenomes and metatrascriptomes
Ultra-low input samples

Library preparation

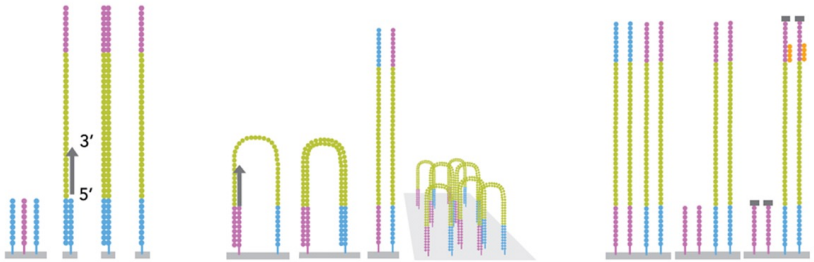
- A sequencing library is a pool of DNA fragments with adapters attached to both ends of the fragments

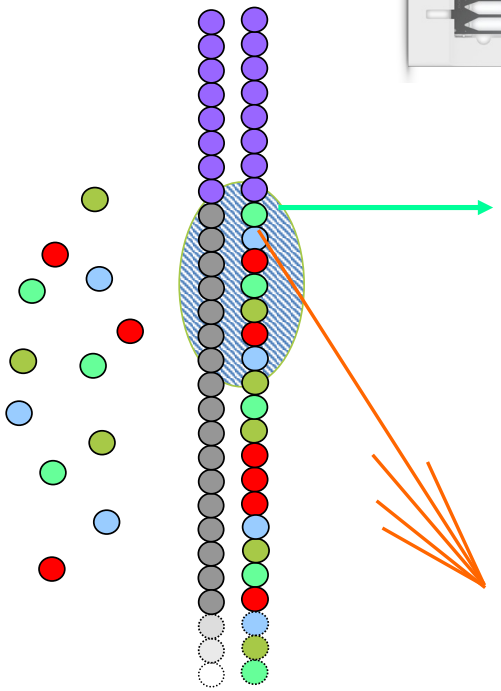
- Approx. 20 protocols for Illumina library prep at NGI

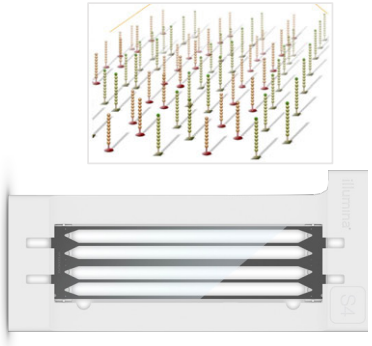


Illumina cluster generation & sequencing

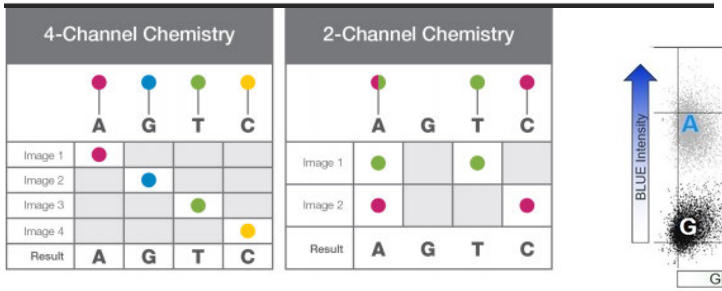
- The sequencing library is hybridized to a flowcell (”cluster generation”)
    - - A flowcell is a slide that is coated with oligos
- Rapid bridge amplification
- Hybidization of sequencing primers
- Sequencing by syntehsis
    - fluorophore labeled nucleotides emitting light

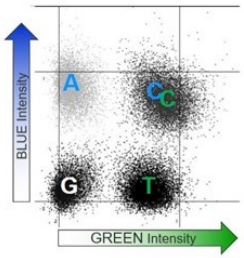






Illumina sequencing by synthesis





NovaX Channel.jpg

Youtube:
https://www.youtube.com/watch?v=fCd6B5HRaZ8

New instrument - NovaSeq X Plus

Flowcell Type	1.5 B	10 B	25 B
Output per flowcell (paired end150 bp)	500 Gb	3 Tb	8 Tb
Number of human genomes per flowcell	~ 4	~ 24	~ 64
Run time (paired end150 bp)	21 h	24 h	48 h

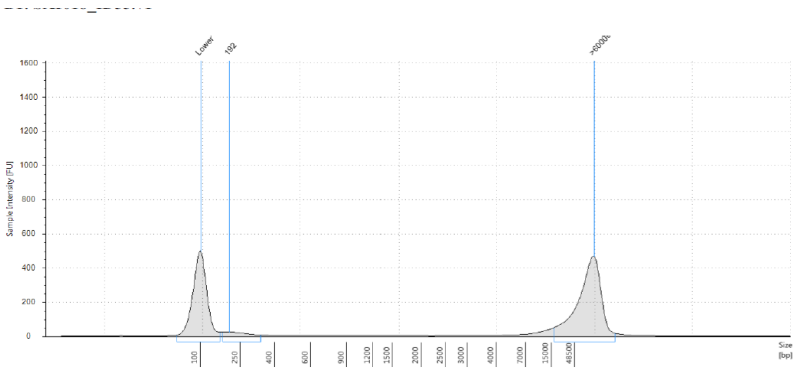
Run ID - Lane	Mb Total Yield	M Total Clusters	% bases ≥ Q30
20230612_LH00179_0005_A2255M2LT3 - L1	295 764.0	979.4	95.4%
20230612_LH00179_0005_A2255M2LT3 - L2	323 896.8	1 072.5	95.3%
20230612_LH00179_0005_A2255M2LT3 - L3	366 557.1	1 213.8	95.6%
20230612_LH00179_0005_A2255M2LT3 - L4	383 028.6	1 268.3	95.0%
20230612_LH00179_0005_A2255M2LT3 - L5	251 454.3	832.6	97.3%
20230612_LH00179_0005_A2255M2LT3 - L6	284 351.5	941.6	97.1%
20230612_LH00179_0005_A2255M2LT3 - L7	388 065.2	1 285.0	94.0%
20230612_LH00179_0005_A2255M2LT3 - L8	363 776.7	1 204.6	95.0%

Quality control of RNA/DNA

DNA

Concentration: QuantIT

Degradation: Fragment Analyzer/TapeStation



Sample Table

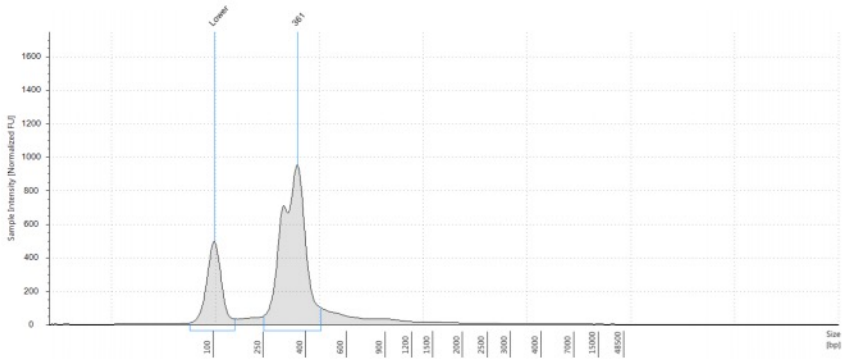
Well	DIN	Conc. [ng/µl]	Sample Description	Alert	Observations
B1	9.6	16.0	SX1018_ID33.v1		

High quality DNA sample

RNA

Concentration + RIN-value:

Fragment Analyzer/TapeStation



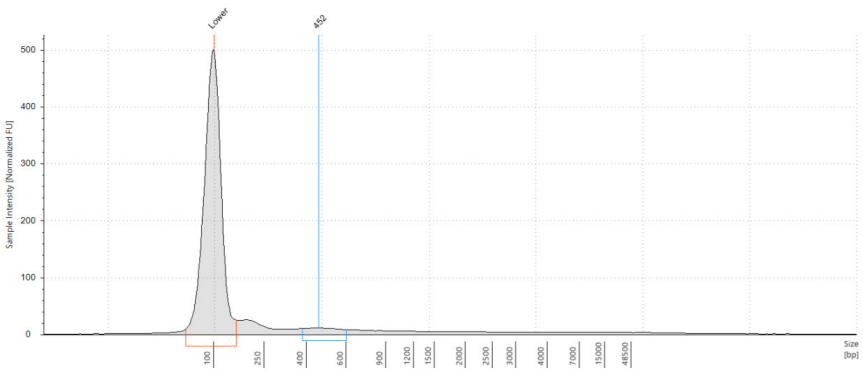
Sample Table

Well	DIN	Conc. [ng/µl]	Sample Description	Alert	Observations
E1	1.0	33.0	92-291039_RJ-1964-pool3		

Degraded DNA sample

Quality of sample/library will affect sequencing result!

DNA-sample: 2.5 ng/ul, DIN-value 0



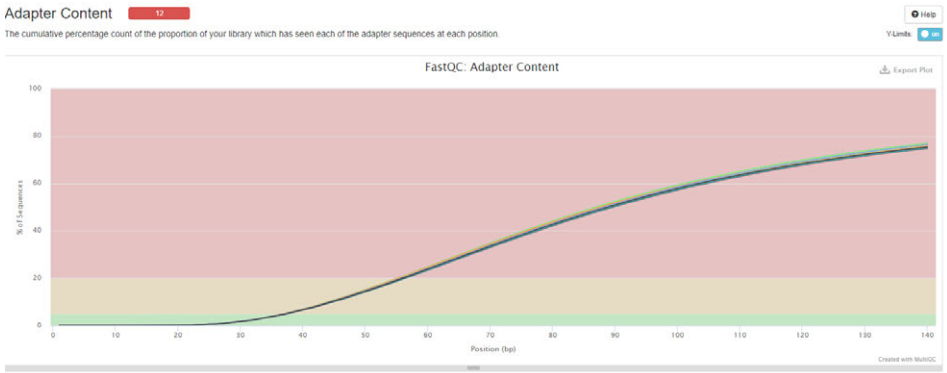
Sample Table

Well	DIN	Conc. [ng/µl]	Sample Description	Alert	Observations
A1	-	2.46	SX1162_S1.v1	▲	Sample concentration outside functional range for DIN

20 ng of DNA, Thurplex Low-input library prep, 3 libraries

Amount of data generated: 800 M read pairs (aiming for ≥60x coverage)

Result: 12x coverage

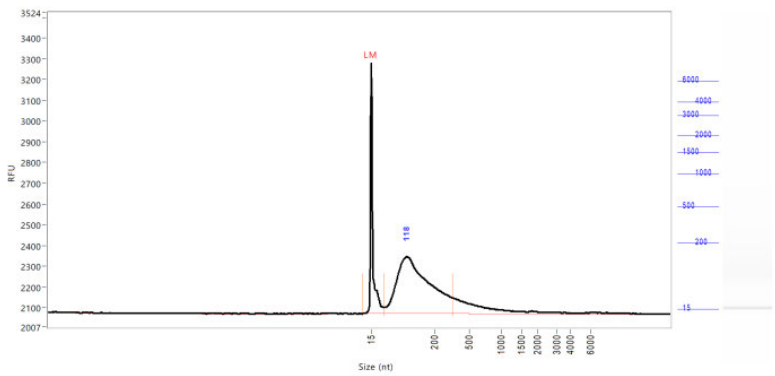


Copy table | Configure Columns | Plot | Showing 7/7 rows and 14/23 columns.

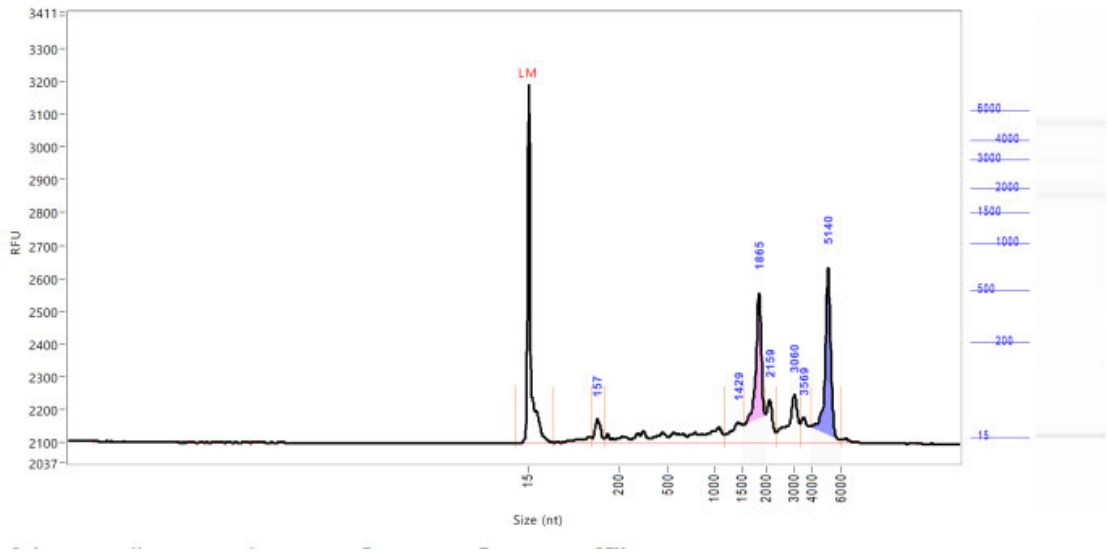
Sample Name	% GC	Ins. size	≥ 30X	Coverage	% Aligned	Change rate	Ts/Tv	M Variants	TiTV ratio (known)	TiTV ratio (novel)	% Dups	% Dups	% GC	M Seqs
S1	46%	55	11.1%	2.0X	98.2%	893	1.645	3.47	2.0	1.6	76.6%			

Quality of sample/library will affect sequencing result!

- RNA samples, RIN-values between 1-9,6
- Library prep Illumina Ligation Ribo-Zero Plus



Peak	Size (nt)	Conc. (ng/uL)	From (nt)	To (nt)	RFU
1	15 (LM)	0.8021	0	54	1209
2	118	27.1086	54	357	272



Results on next page…

Continued…Quality of sample/library will affect sequencing result

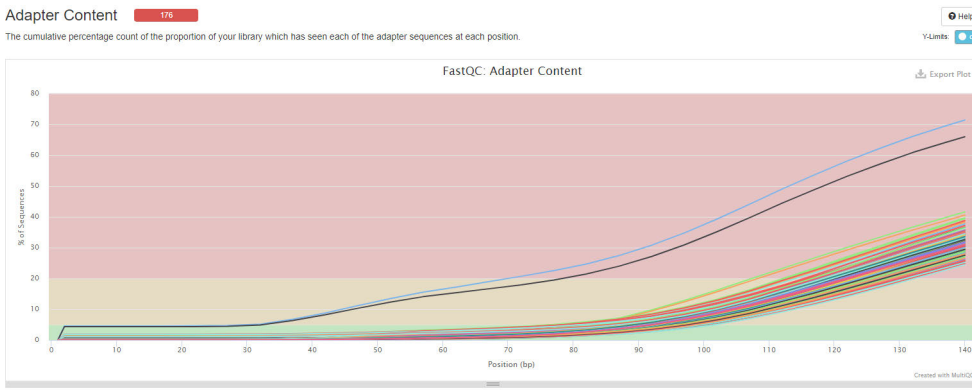
QC-reults RNA-seq

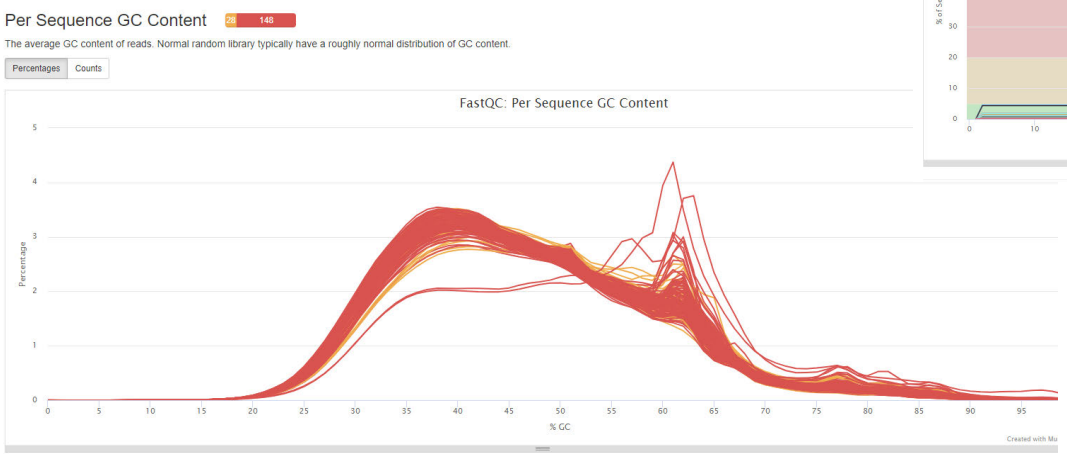
Uneven amounts of data (17-100 M reads per sample)

A lot of duplicates

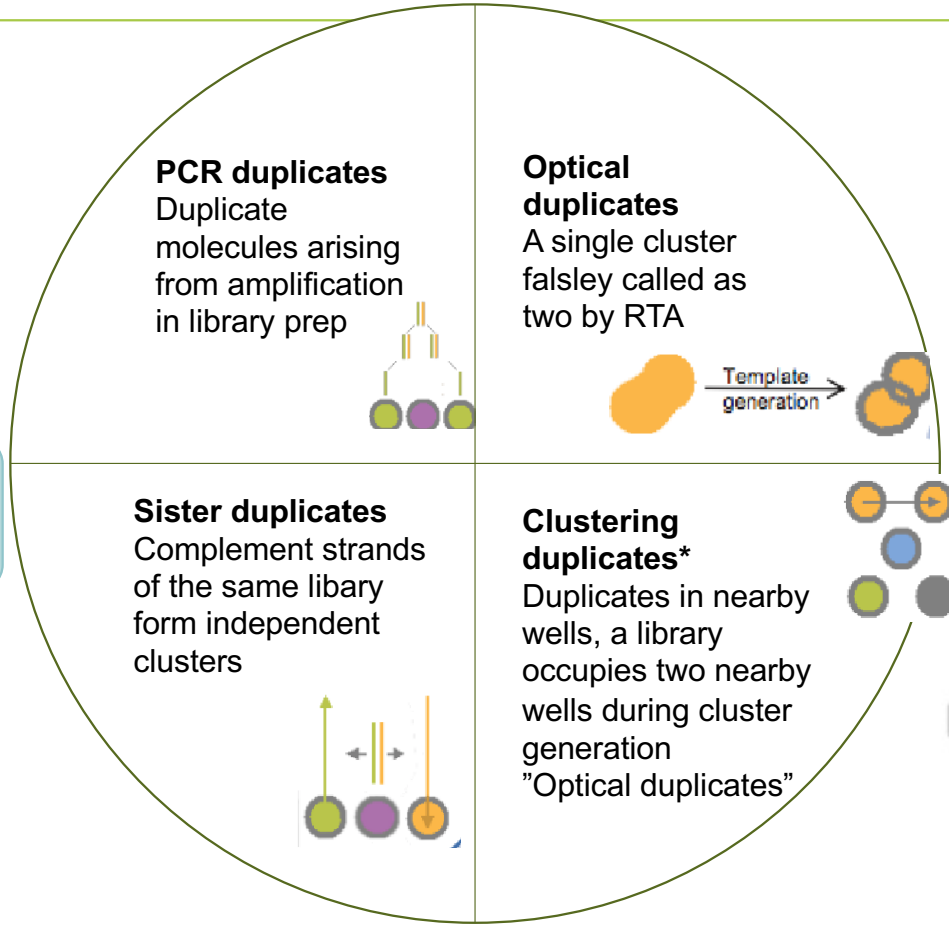
High rRNA content

High adapter content

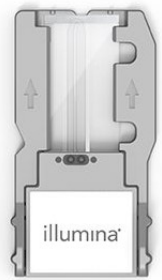




Other challenges – duplicates, duplicates, duplicates….



Present on all Illumina Platforms



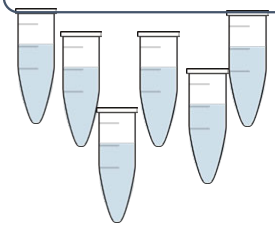
On non-patterned flowcells (MiSeq, HiSeq 2500 etc.)

Patterned flowcells only (NextSeq, NovaSeq 6000/X

Some of the applications offered

Templates:

DNA, RNA, cells, serum or plasma



Whole Genome Sequencing (WGS)

- *De novo* sequencing (PacBio, ONT)
- Re-sequencing (PCR-Free, low input)

Transcriptome Sequencing

- mRNA-Seq (poly-A selection)
- Total RNA-seq (ribosomal depletion)
- miRNA & small RNAs
- Full-length transcriptomes

Targeted re-sequencing

- Exome
- Gene panels
- Amplicons (including bacterial 16S for metagenomics)
- RAD-seq

Epigenetics

- Chromatin (HiC, ATAC-Seq)
- WGBS
- ChIP Sequencing

Ready-made libraries

- User-made libraries
- High throughput
- Fast turn around time

Single-cell applications

- 10x Genomics
- Dolomite Nadia
- Single-cell WGBS

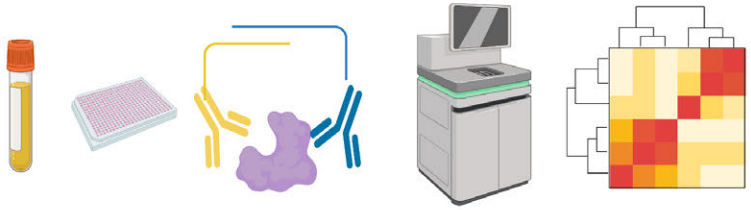
Spatial transcriptomics

- 10x Genomics Visium

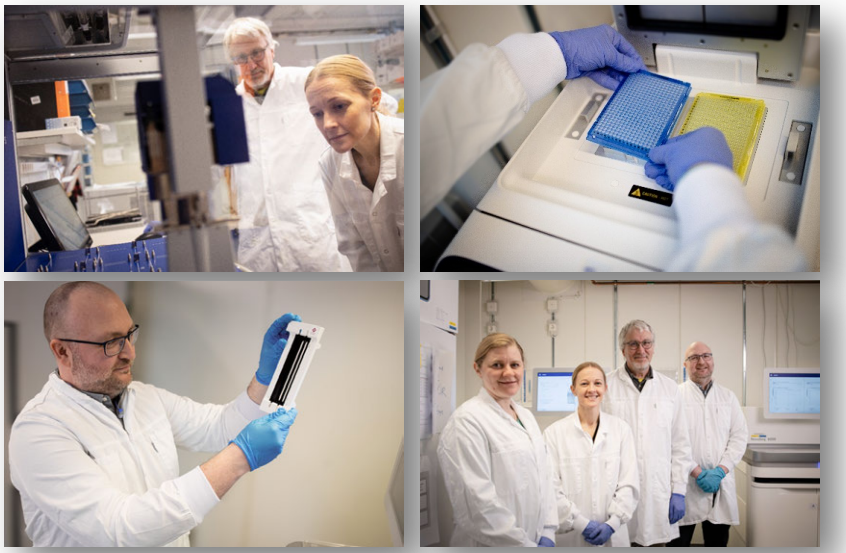
Proteomics with NGS readout

- Olink Explore 1536/3072/5300

Protein analysis, Olink Explore with NGS readout



SciLifeLab Explore Lab: NGI in collaboration with the Affinity Proteomics Uppsala unit and Olink Proteomics AB



- Highly multiplex protein biomarker analysis:
  - Olink Explore 384-5300 protein assays available
    - Cardio-metabolic
    - Inflammation
    - Neurology
    - Oncology
- Stats
  - >25 000 samples analyzed since the method was set up in the spring of 2021

Examples, recent successful projects

Forensic Science International: Genetics 53 (2021) 102525

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsigen

Research paper

Getting the conclusive lead with investigative genetic genealogy – A successful case study of a 16 year old double murder in Sweden

Andreas Tillmar[a,b,*], Siri Aili Fagerholm[c], Jan Staaf[d], Peter Sjölund[e], Ricky Ansell[c,f,**]

[a] Department of Forensic Genetics and Forensic Toxicology, National Board of Forensic Medicine, Linköping, Sweden
[b] Department of Biomedical and Clinical Sciences, Faculty of Medicine and Health Sciences, Linköping University, Linköping, Sweden
[c] National Forensic Centre, Swedish Police Authority, Linköping, Sweden
[d] Polisregion Öst, Swedish Police Authority, Linköping, Sweden
[e] Peter Sjölund AB, Härnösand, Sweden
[f] Department of Physics, Chemistry and Biology, Linköping University, Linköping, Sweden

Article | Published: 17 February 2021

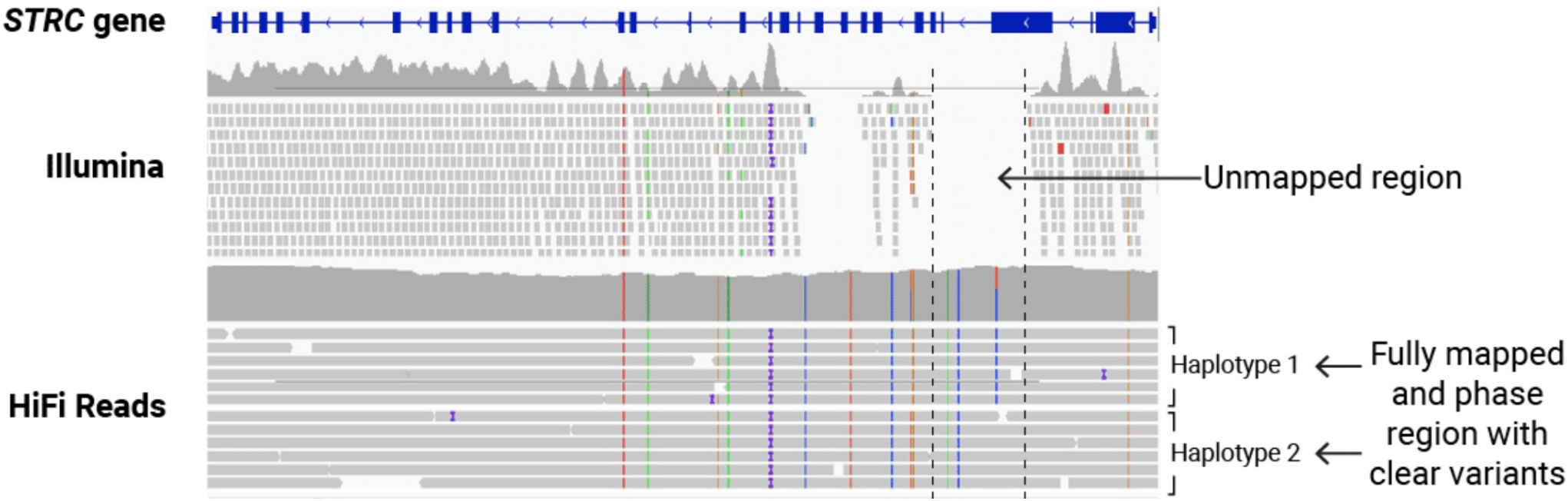
Million-year-old DNA sheds light on the genomic history of mammoths

Tom van der Valk, Patrícia Pečnerová, David Díez-del-Molino, Anders Bergström, Jonas Oppenheimer, Stefanie Hartmann, Georgios Xenikoudakis, Jessica A. Thomas, Marianne Dehasque, Ekin Sağlıcan, Fatma Rabia Fidan, Ian Barnes, Shanlin Liu, Mehmet Somel, Peter D. Heintzman, Pavel Nikolskiy, Beth Shapiro, Pontus Skoglund, Michael Hofreiter, Adrian M. Lister, Anders Götherström & Love Dalén

Nature 591, 265–269 (2021) | Cite this article

Limitations with short reads

- You don’t get complete genome information!

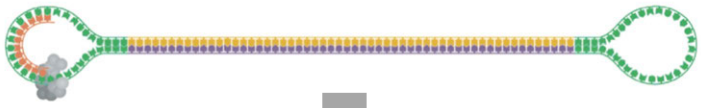


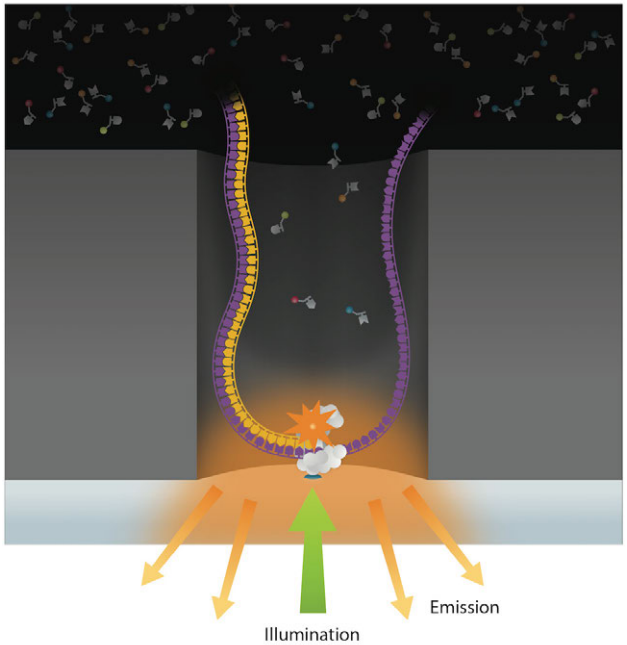
Long-read sequencing

No longer a niche technology!

- Assemble complete genomes
- Find all genetic variants
- Detect epigenetic modifications
- At a “reasonable” cost

PacBio Sequencing

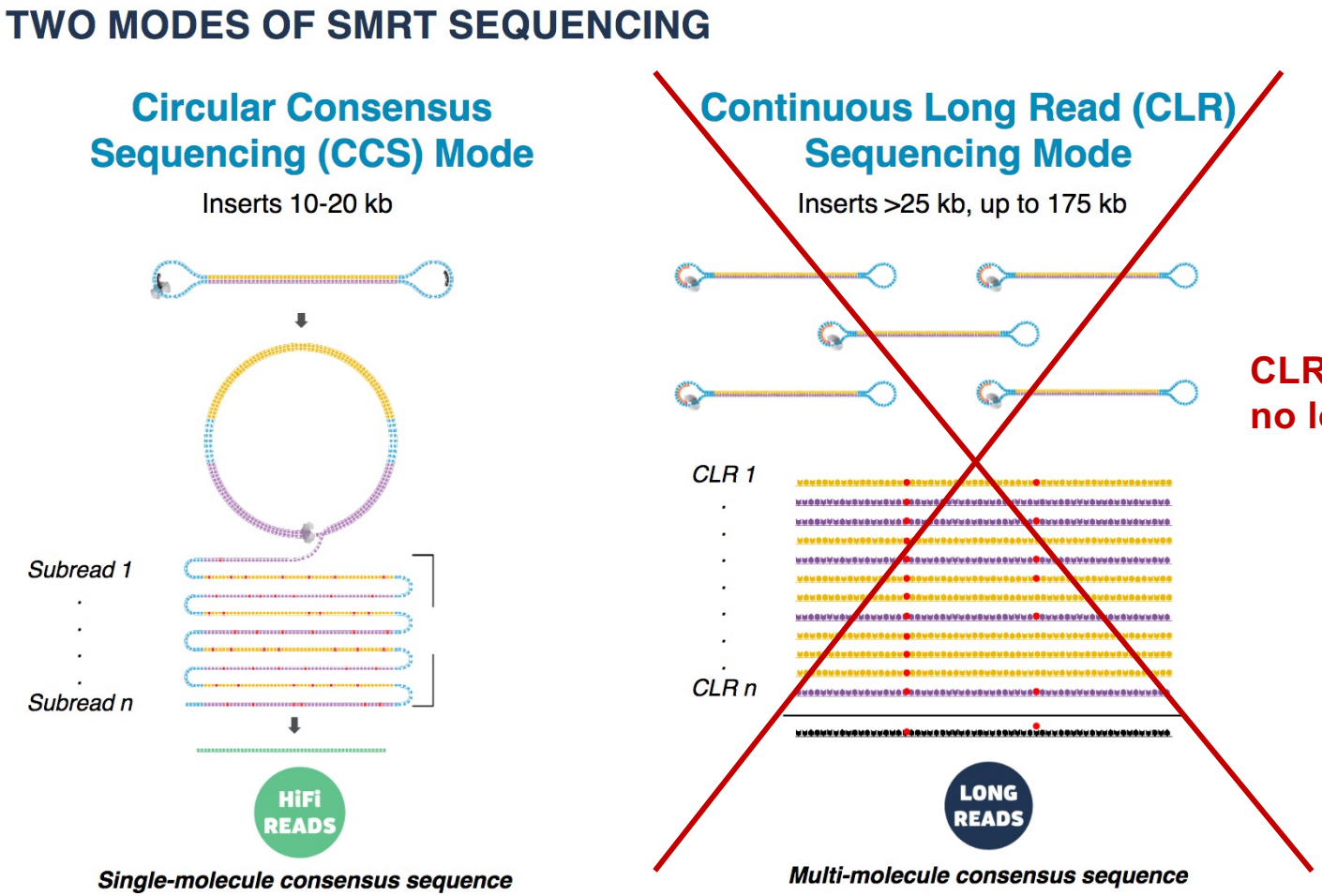




PacBio RSII

PacBio Sequel
(Sequel I & II)

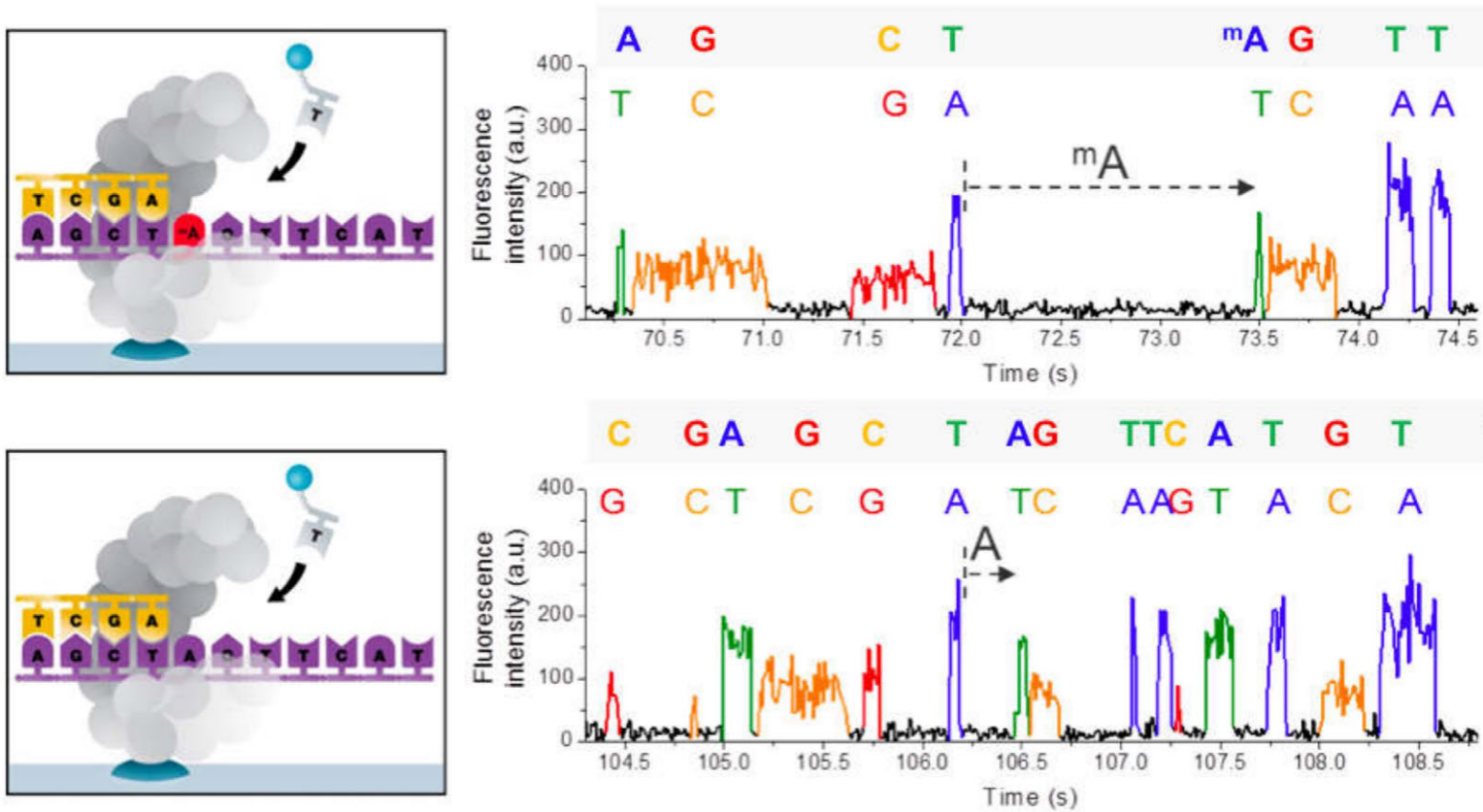
PacBio Sequencing



CLR sequencing no longer supported

PacBio – Methylation detection

- Base modifications on native DNA molecules can be detected!



A decade of PacBio sequencing at NGI

2013: Installation of PacBio RSII

2023: Arrival of PacBio Revio

The PacBio Revio System

- Up to 90Gb data from one SMRT cell
- Read lengths: 15-20kb
- >QV20 quality (>99% read accuracy)
- Can run 1,300 human genomes/year!
- We installed PacBio Revio in March 2023

Revio – results for our first 16 runs

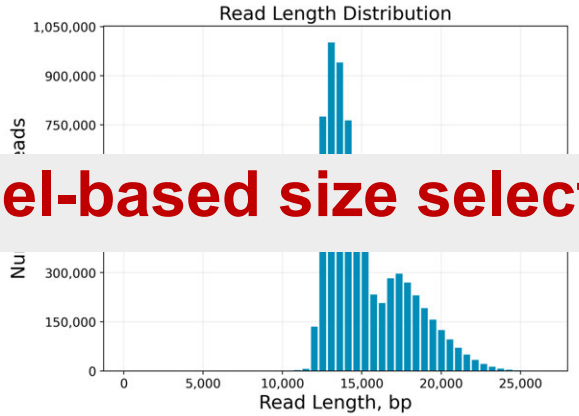
Sample/Species/Proj	Number of reads	Total yield (Gbp)	Average read length (kb)	Size selection method	Comment
Human 1_1	6,873,030	84.7	12.3	Ampure beads	Also Sequel II data
Human 1_2	6,846,419	102.2	15.0	Ampure beads	Also Sequel II data
Human 1_3	7,170,075	90.3	12.6	Ampure beads	Also Sequel II data
Human 1_4	6,015,366	67.6	11.2	Ampure beads	Also Sequel II data
Human 2_1	6,895,775	104.2	15.1	SageELF (2 fract. pooled)	
Human 2_2	5,684,755	100.3	17.6	SageELF (2 fract. pooled)	
Human 2_3	6,022,465	111.5	18.5	SageELF (2 fract. pooled)	
Human 3_1	7,544,871	72.3	9.6	Ampure beads	
Human 3_2	7,857,802	65.6	8.3	Ampure beads	
Human 3_3	7,164,744	102.3	14.3	Ampure beads	
Human 3_4	6,695,524	82.4	12.3	Ampure beads	
Human 3_5	6,541,509	80.4	12.3	Ampure beads	
Plant 1_1	7,683,014	70.1	9.1	Ampure beads	Also Sequel II data
Amphibian 1_1	2,700,447	23.5	8.7	Ampure beads	225 pM loading
Amphibian 1_1	5,219,472	42.3	8.1	Ampure beads	350 pM loading
Bird 1_1	6,812,139	90.2	13.2	Ampure beads	

Revio – results for our first 16 runs

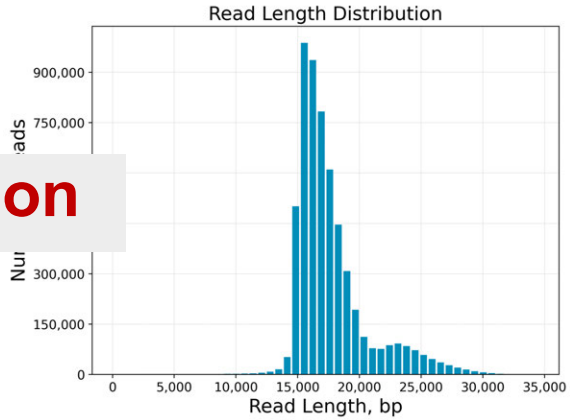
Sample/Species/Proj	Number of reads	Total yield (Gbp)	Average read length (kb)	Size selection method	Comment
Human 1_1	6,873,030	84.7	12.3	Ampure beads	Also Sequel II data
Human 1_2	6,846,419	102.2	15.0	Ampure beads	Also Sequel II data
Human 1_3	7,170,075	90.3	12.6	Ampure beads	Also Sequel II data
Human 1_4	6,015,366	67.6	11.2	Ampure beads	Also Sequel II data
Human 2_1	6,895,775	104.2	15.1	SageELF (2 fract. pooled)	
Human 2_2	5,684,755	100.3	17.6	SageELF (2 fract. pooled)	
Human 2_3	6,022,465	111.5	18.5	SageELF (2 fract. pooled)	
Human 3_1	7,544,871	72.3	9.6	Ampure beads	
Human 3_2	7,857,802	65.6	8.3	Ampure beads	
Human 3_3	7,164,744	102.3	14.3	Ampure beads	
Human 3_4	6,695,524	82.4	12.3	Ampure beads	
Human 3_5	6,541,509	80.4	12.3	Ampure beads	
Plant 1_1	7,683,014	70.1	9.1	Ampure beads	Also Sequel II data
Amphibian 1_1	2,700,447	23.5	8.7	Ampure beads	225 pM loading
Amphibian 1_1	5,219,472	42.3	8.1	Ampure beads	350 pM loading
Bird 1_1	6,812,139	90.2	13.2	Ampure beads	

Size selection method makes a difference!

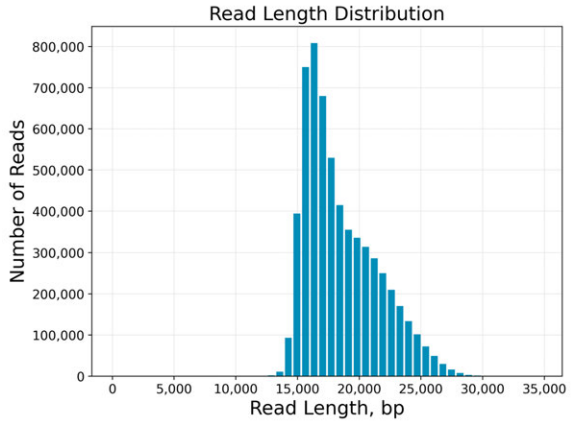
Gel-based size selection



Human 2_1

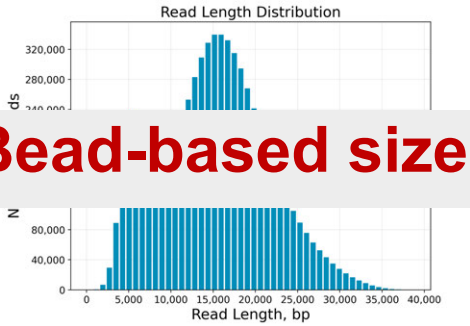


Human 2_2

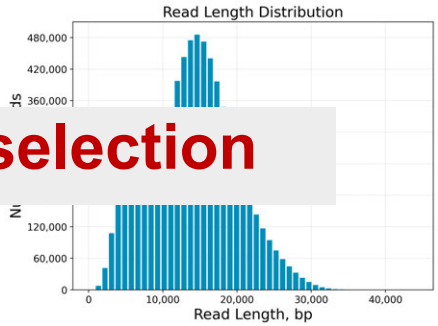


Human 2_3

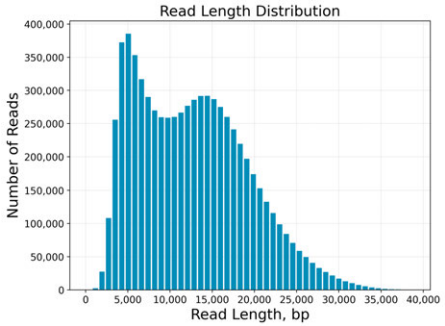
Bead-based size selection



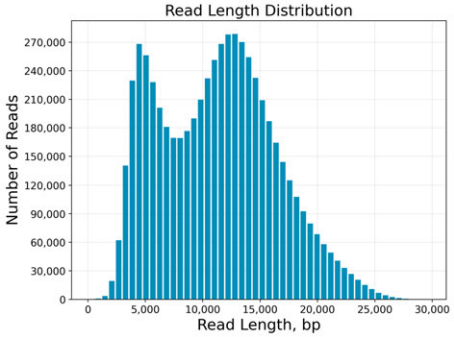
Human 1_1



Human 1_2

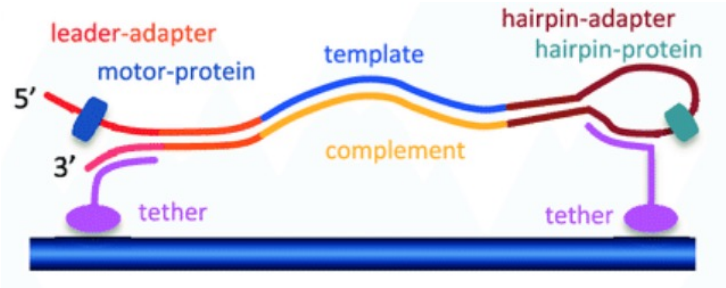


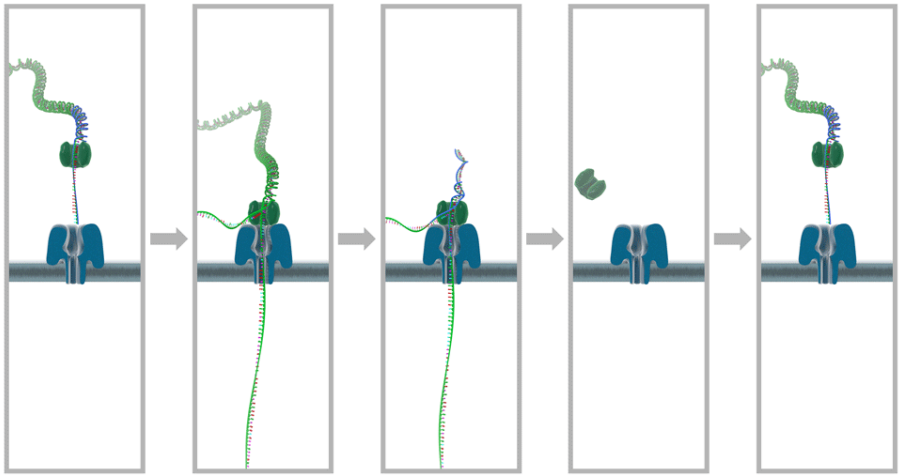
Human 1_3

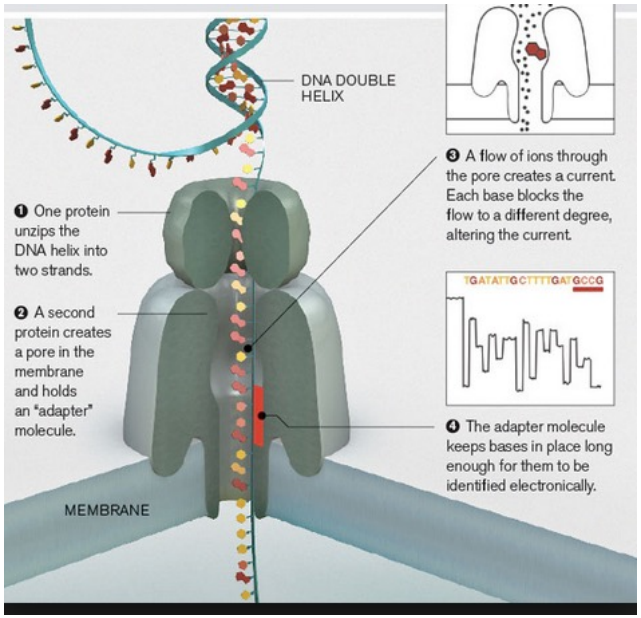


Human 1_4

Oxford Nanopore sequencing







Base modification info is retained

Oxford Nanopore sequencing

Instrument	Run time /FC	Output / FC	Nr of pores	Max read length
Flongle	16 hrs	1 Gb	126	1 Mb
MinION	24 hrs	2-15 Gb	512	1 Mb
GridION	24 hrs	2-15 Gb	512	1 Mb
PromethION	72 hrs	10 – 150 Gb	3 000	2 Mb

ONT - Portability

The International Space Station

In 2016, MinION was used to conduct the first ever DNA sequencing in space. MinION performance was unaffected by the flight to the International Space Station (ISS) or microgravity conditions. The team stated that '*these findings illustrate the potential for sequencing applications including disease diagnosis, environmental monitoring, and elucidating the molecular basis for how organisms respond to spaceflight*.' Further to this, in 2020, an end-to-end sample-to-sequencer workflow conducted entirely aboard the ISS resulted in off-Earth identification of microbes for the first time.

Photograph: NASA ©

Read more

Uncovering cryptic transmission of Zika virus

The origin and epidemic history of Zika virus (ZIKV) in Brazil and the Americas remained poorly understood despite observed trends in reported microcephaly. Using a mobile genomics lab to conduct genomic surveillance of ZIKV, the team identified the earliest confirmed ZIKV infection in Brazil. Analysis of these genomes estimated that ZIKV is likely to have disseminated from north-east Brazil in 2014, before the first detection in 2015, indicating a period of pre-detection cryptic transmission that would not have been identified without genomic sequencing.

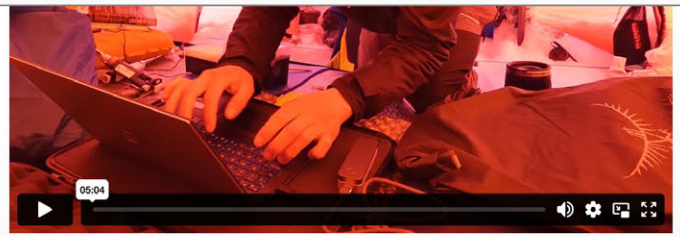
Read more

Credit: Nuno R. Faria

Entirely off-grid, solar-powered sequencing

In 2019, Gowers *et al.* used MinION to demonstrate '*the ability to conduct DNA sequencing in remote locations, far from civilised resources (mechanised transport, external power supply, internet connection, etc.), whilst greatly reducing the time from sample collection to data acquisition*'. The team transported their portable lab for 11 days using only skis and sledges across Europe's largest ice cap (Vatnajökull, Iceland), before carrying out a tent-based study, resulting in 24 hours of sequencing data using solar power alone.

Read more



ONT - Speed

New DNA Sequencing Tech

January 17, 2022

Tweet | Share 1 | Share | Ema

A new ultra-rapid genome sequencing approach
collaborators was used to diagnose rare genetic
unheard of in standard clinical care.

"A few weeks is what most clinicians call 'rapid' w
results," said Euan Ashley, MB, professor of med

Genome sequencing allows scientists to see a p
everything from eye color to inherited diseases.
rooted in their DNA: Once doctors know the spe

Now, a mega-sequencing approach devised by A
diagnostics: Their fastest diagnosis was made in
less time in critical care units, require fewer test

A paper describing the researchers' work is pub
Burnell Professor in Genomics and Precision Health, is the senior author of the paper. Postdoctoral scholar John Gorzynski, DVM, PhD, is the lead author.

nature

Explore content | About the journal | Publish with us

nature > articles > article

Article | Open access | Published: 11 October 2023

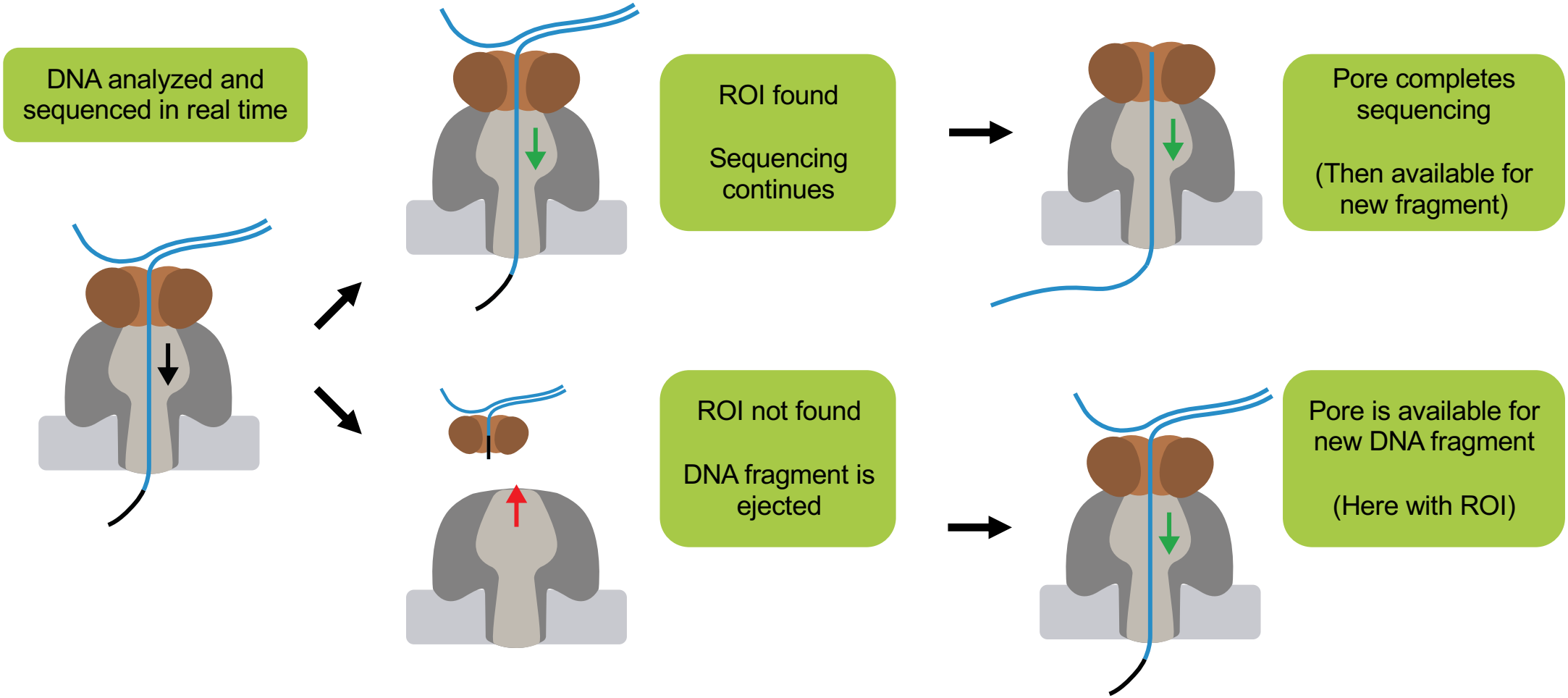
Ultra-fast deep-learned CNS tumour classification during surgery

C. Vermeulen, M. Pagès-Gallego, L. Kester, M. E. G. Kranendonk, P. Wesseling, N. Verburg, P. de Witt Hamer, E. J. Kooi, L. Dankmeijer, J. van der Lugt, K. van Baarsen, E. W. Hoving, B. B. J. Tops & J. de Ridder

Nature **622**, 842–849 (2023) | Cite this article

34k Accesses | **563** Altmetric | Metrics

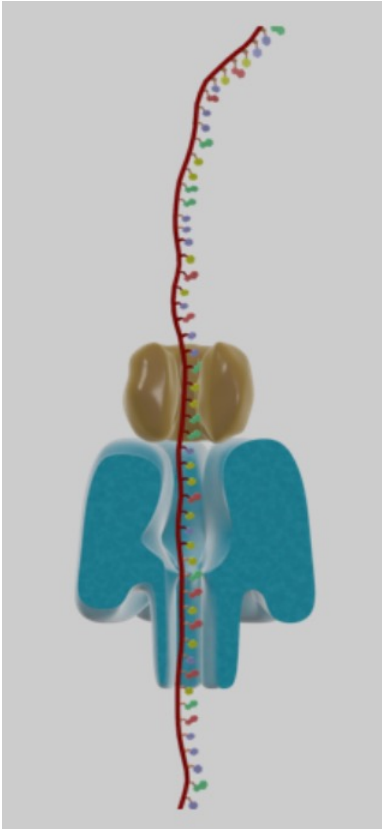
ONT target sequencing - adaptive sampling



ONT direct RNA sequencing

ONT can sequence native RNA molecules!

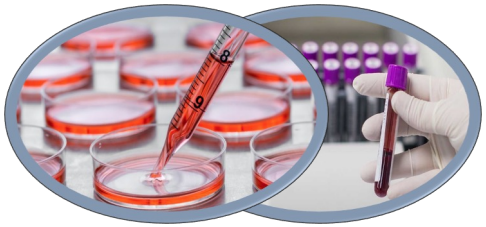
- No bias due to cDNA conversion
- Allows to study RNA modifications
- Higher error rate
- Lower throughput



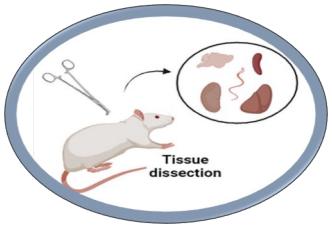
DNA extraction for long-read sequencing

HMW-DNA Extraction – Options at NGI

Cells/Blood
$1x10^6$ - $5x10^6$



Tissue
25-100 mg



Insects/Mollusc/Crustaceans
25-200 mg

Plants
1-3 g

Fungi
100-600 mg

Commercial Kits

MONARCH
High input quality required
Few special protocols

Top choice for high quality samples with low amount of contaminants

NANOBIND
Lower input quality tolerated
Many special protocols
Suplemental buffers for insects

Top choice for most non-standard samples except for low input and high polysaccharide samples

Phenol/Chloroform

SDS Lysis
High polyphenol
High recovery for low input

Top choice for samples high in polyphenols without polysaccharides

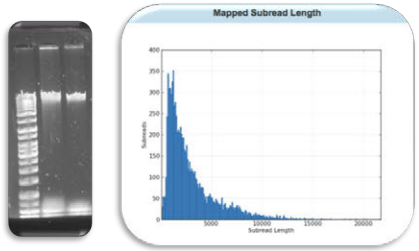
CTAB Lysis
High polysaccharide
Also handles polyphenols

Top choice for plants, fungi, and other samples high in polysaccharides

HMW-DNA Extraction – Contaminants

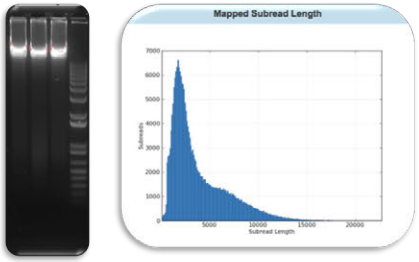
Importance of purity – even for model organisms

Impurities can originate from both host tissue and extraction chemicals.



Polished Contigs	223	Max Contig Length	36,298
N50 Contig Length	2,932	Sum of Contig Lengths	480,087

Same yeast - different extractions!



Polished Contigs	9	Max Contig Length	1,508,929
N50 Contig Length	1,353,702	Sum of Contig Lengths	7,813,244

We extract what we get!

Sequencing of the last supper?

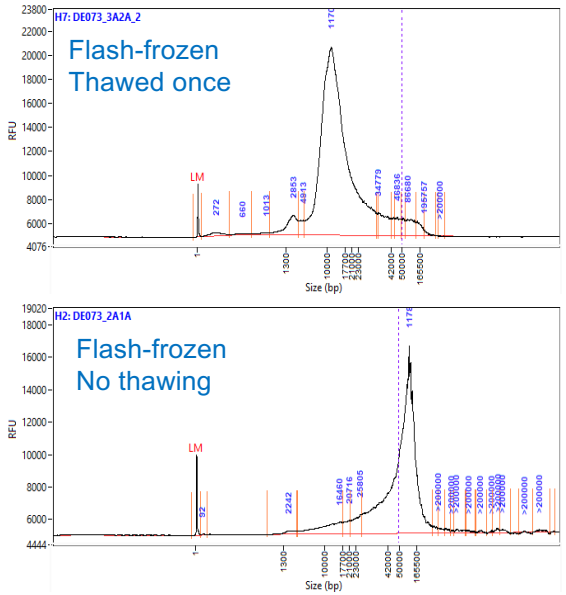
Which would you expect to have less contaminants?

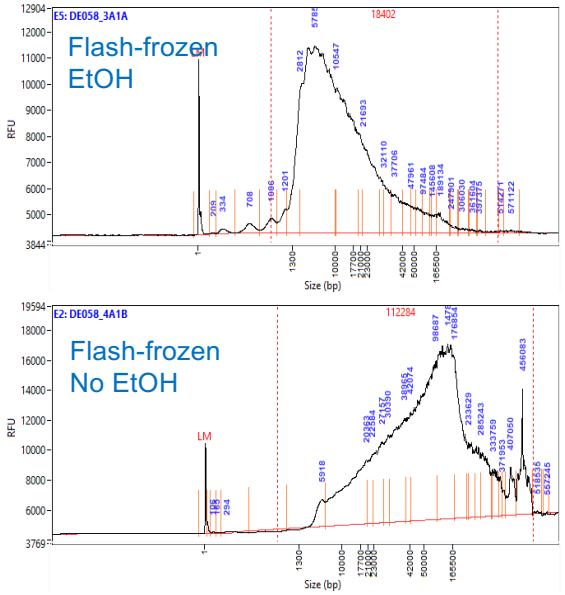
VS

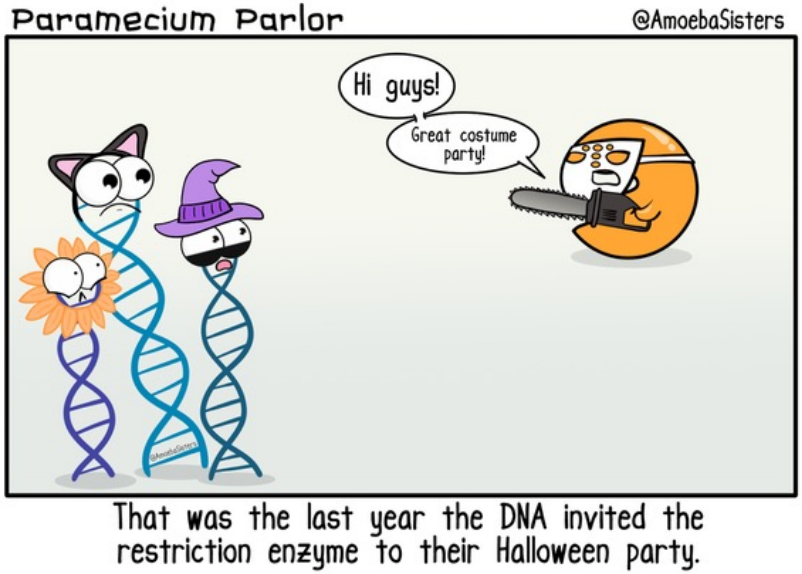
VS

HMW-DNA Extraction - Fragmentation

- Keep cells intact to preserve HMW-DNA
- Dissect pre-freezing to avoid thaw cycles
- Freeze as fast and cold as possible to minimize cell rupt



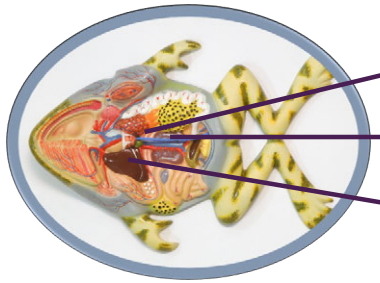




- Ethanol disrupts cells
- Avoid if possible
- Still best option for ambient storage (sample dependent)

HMW-DNA Extraction – Best Options

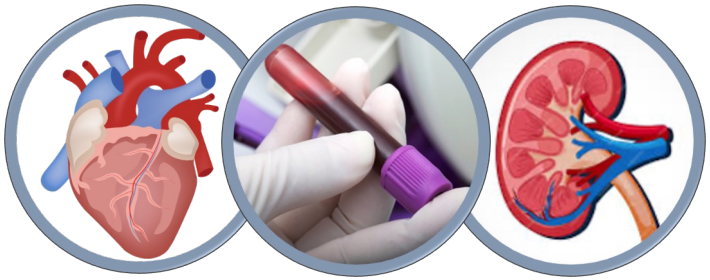
- ❑ Plan ahead and divide according to what you plan to do



DNA?

RNA?

Hi-C?



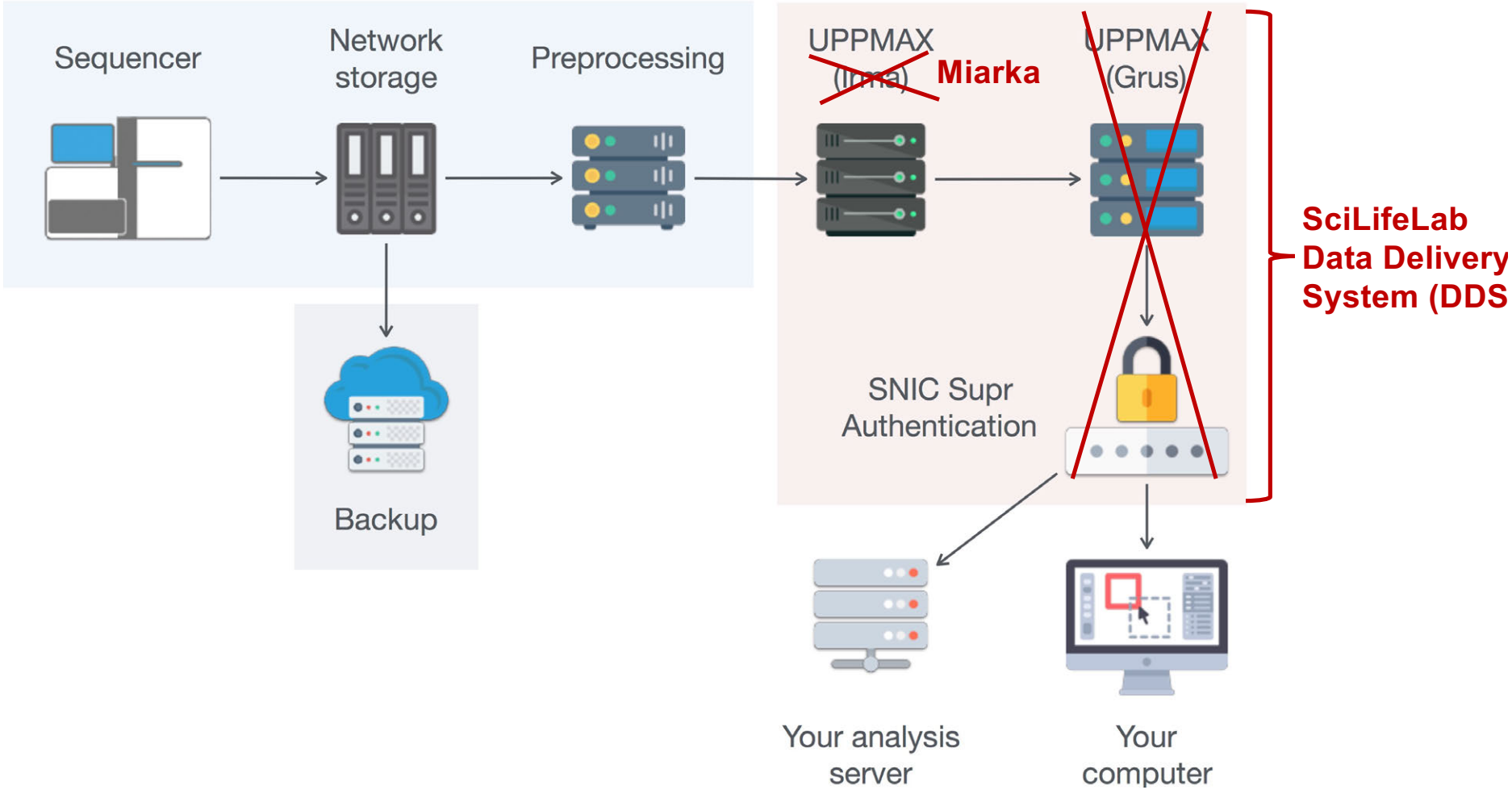
- ❑ Choose tissue high in DNA and low in contaminants when possible

- ❑ Freeze as fast and cold as possible to minimize fragmentation

- ❑ All samples are different – Investigate what are best options for your samples!

NGI Data Handling and Analysis Pipelines

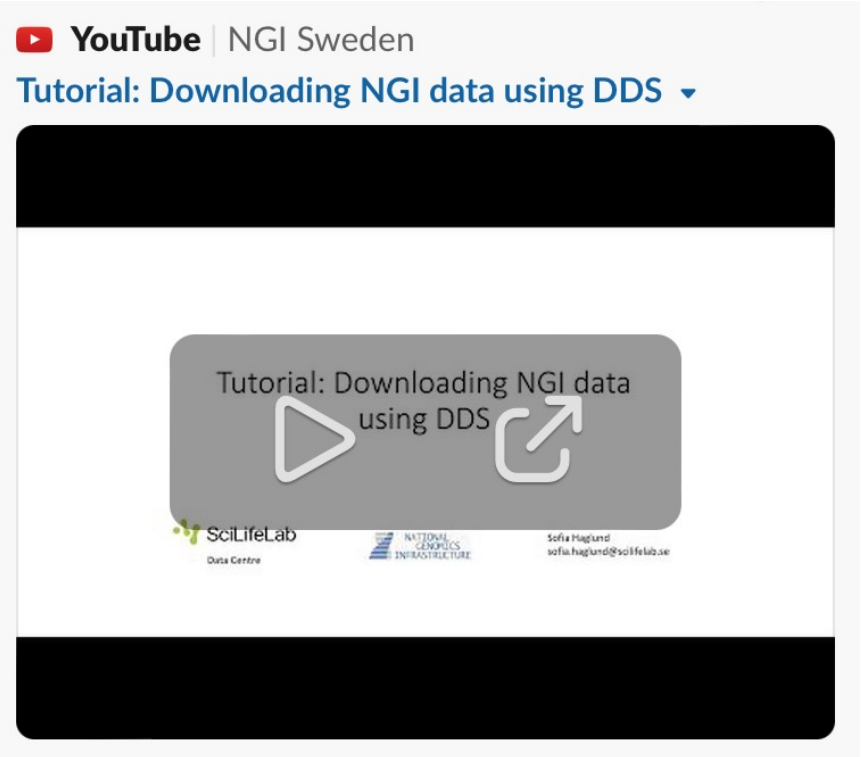
NGI Data Handling



Data delivery via DDS

- DDS is a system for delivery of data from SciLifeLab platforms
  - Cloud-based solution
  - Command line and web interface
  - Can handle also sensitive data

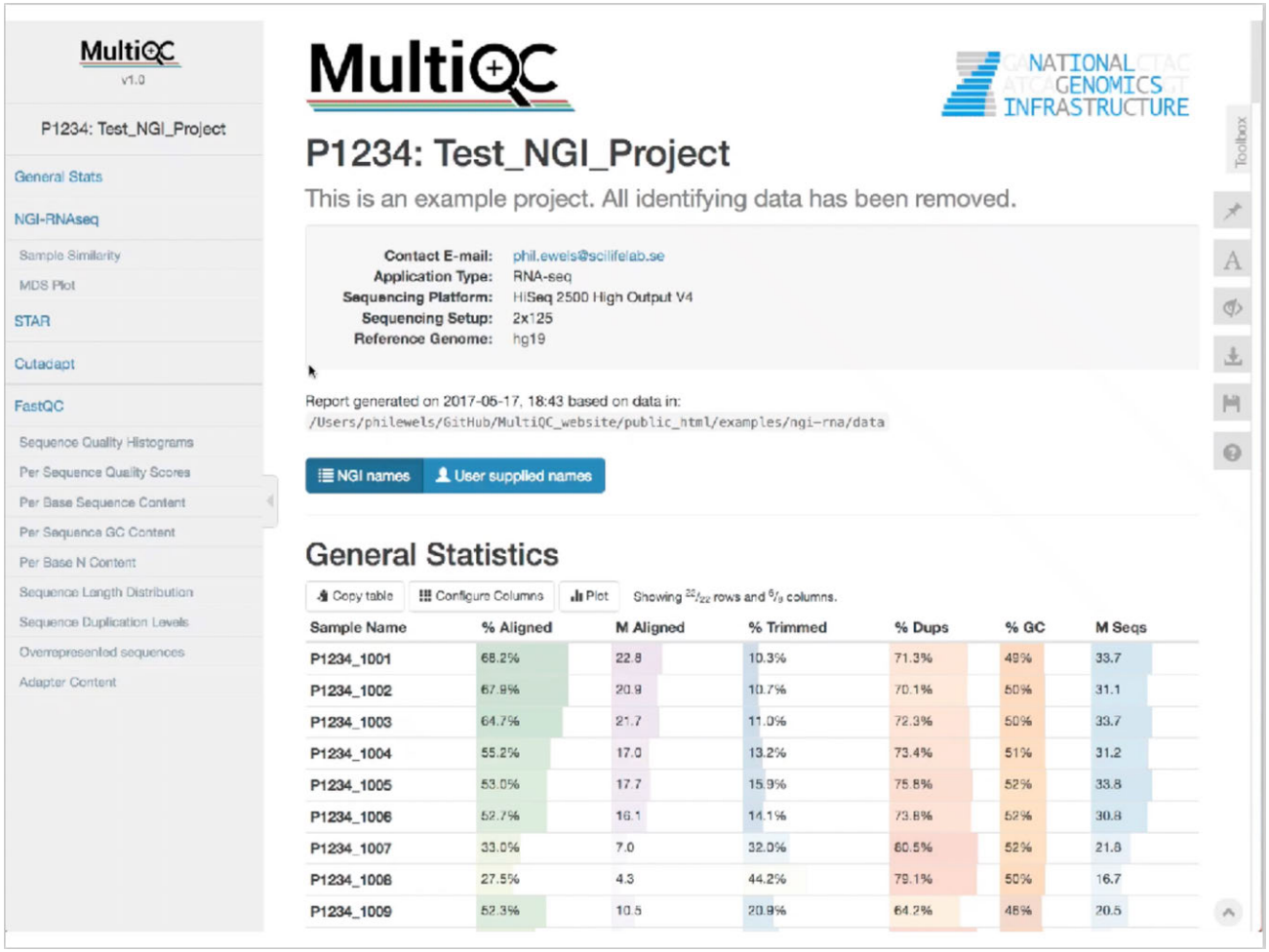
- Instruction video available on Youtube!



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



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 popular pipeline system

- 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 (previously NGI Sthlm

Example pipeline - Sarek

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

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

Manta

ASCAT

MuTect1

MuTect2

Strelka

FreeBayes

GATK HaplotypeCaller

NATIONAL GENOMICS INFRASTRUCTURE

NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

Barntumörbanken

Trend: On-instrument analysis

More and more analyses being done on instrument GPUs

Illumina NovaSeqX

Mapping and variant calling (Dragen)

PacBio Revio

Onboard generation of HiFi reads

➔ Can speed up and streamline the analysis process

NGI Strategic Projects

NGI Strategic 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

Three examples to follow:

1: Swedish human reference dataset

2: Long-read sequencing in Rare Disease

3: Earth Biogenome project

Example I: 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

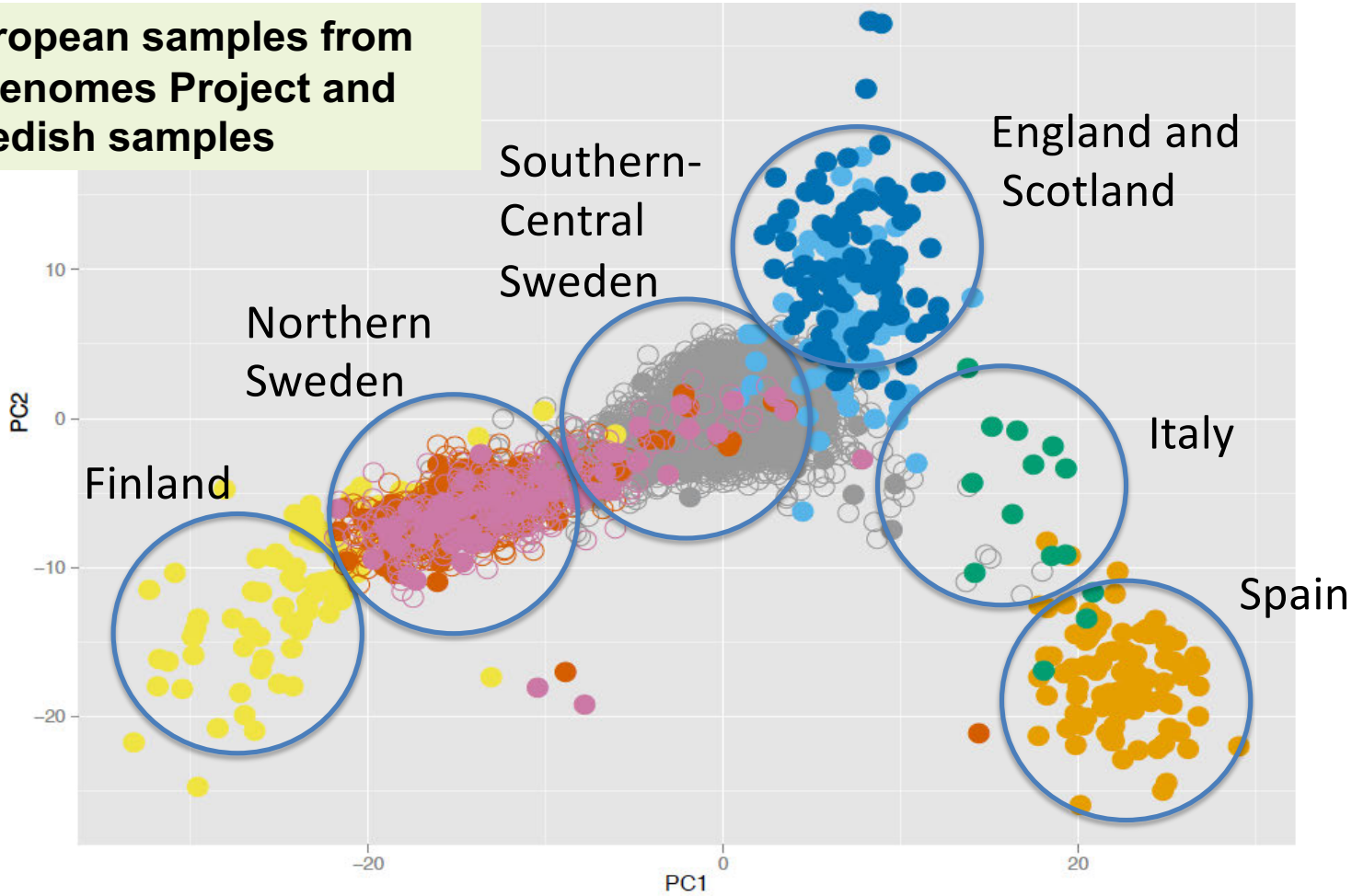
- What can the SweGen dataset be used for?
  - Look up genetic variant frequencies
  - Use as matched controls
  - Study population genetics
  - Study human evolutionary history

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



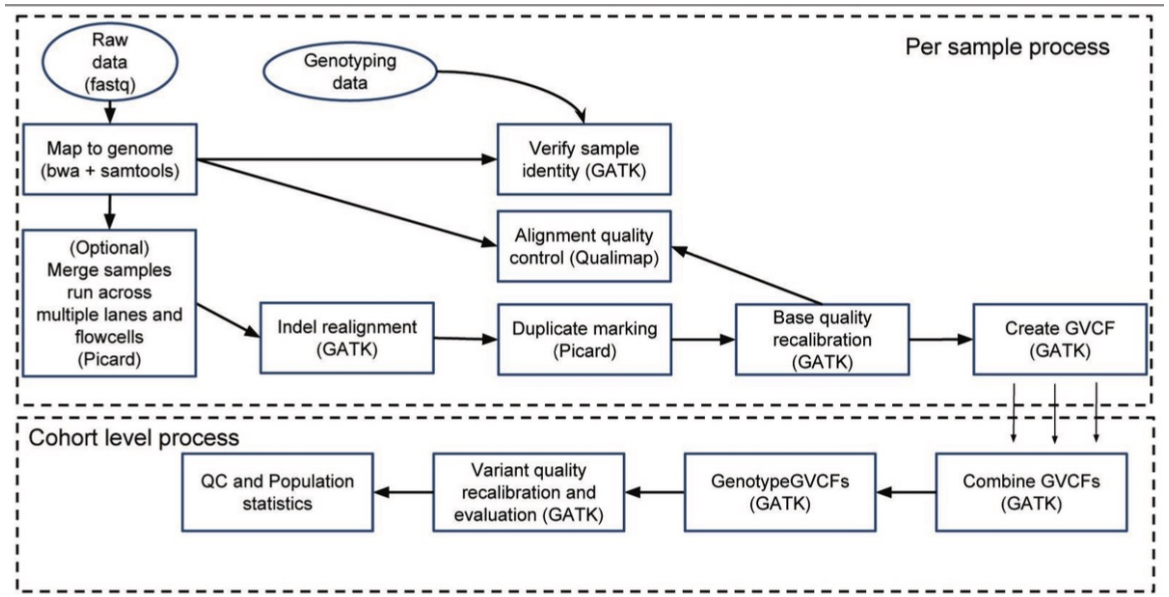
Whole Genome Sequencing

- 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

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***
- 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

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

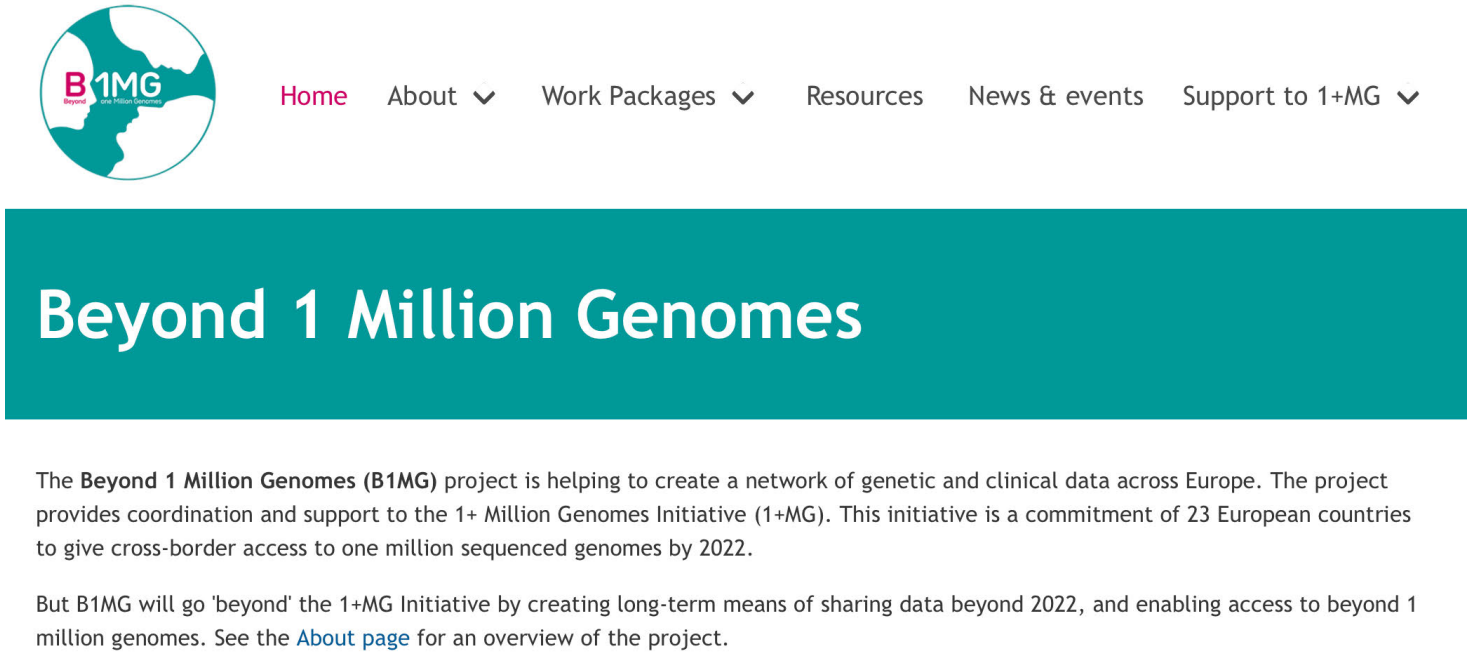
Fabiana H. G. Farias, Johanna Dahlqvist, Sergey V. Kozyrev, Dag Leonard, Maria Wilbe, Sergei N. Abramov, Andrei Alexsson, Gerli R. Pielberg, Helene Hansson-Hamlin, Göran Andersson, Karolina Tandre, Anders A. Bengtsson, Christopher Sjöwall, Elisabet Svenungsson, Iva Gunnarsson, Solbritt Rantapää-Dahlqvist,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

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

What will happen next?

"Genome of Europe" is a new EU initiative within the 1+MG project

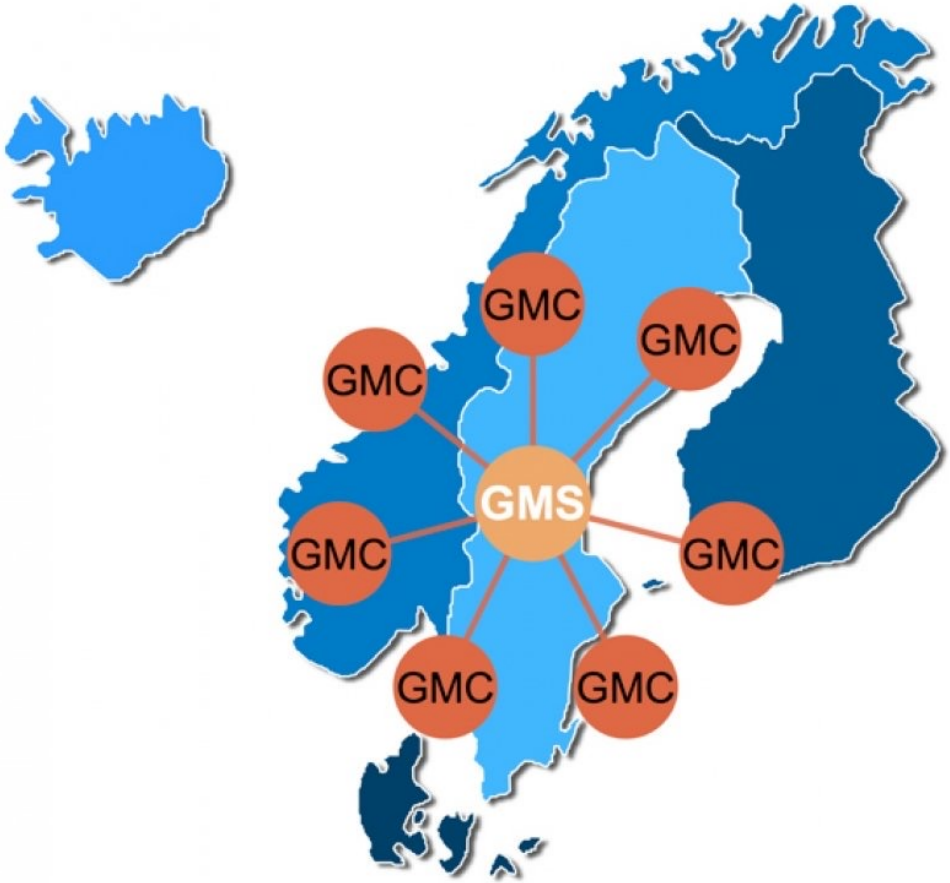
- We will aim to generate a long-read reference dataset for Sweden!



A PacBio Revio pilot for rare disease

Project plan:

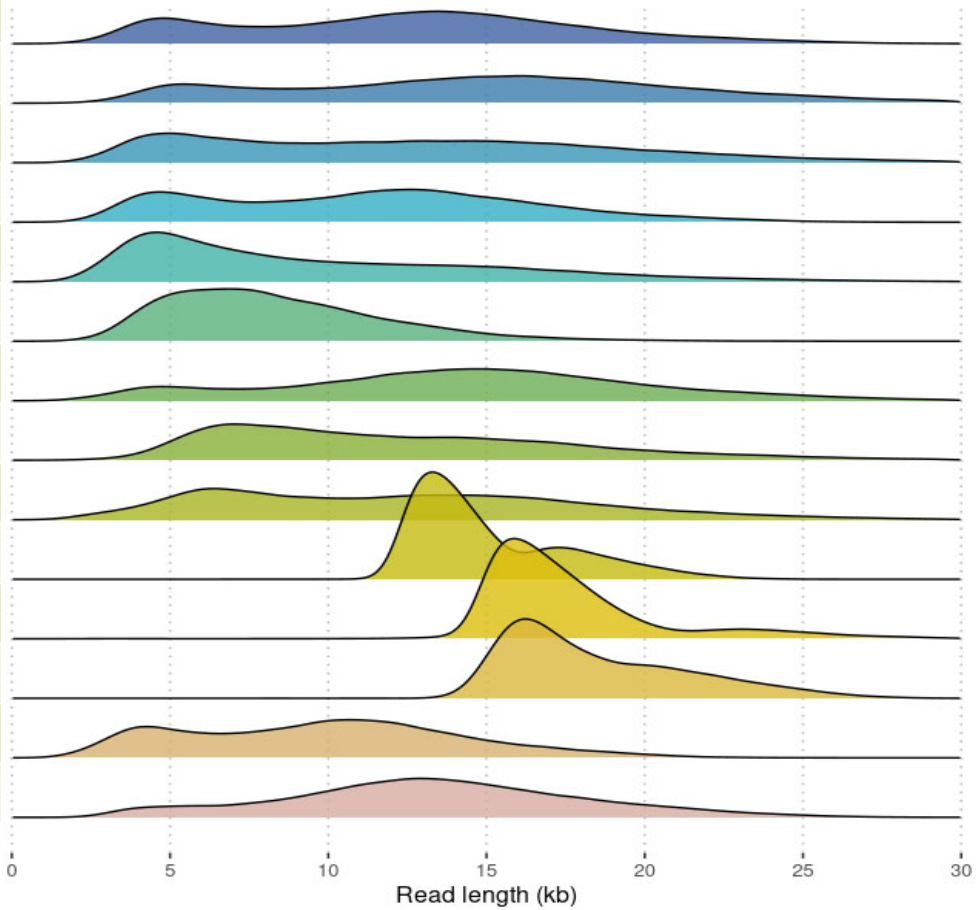
- 15-20 clinical cases
- from 6 Swedish hospital regions
- DNA extracted by regular methods
- Complex SVs suspected
- Other genomics data available (short reads, arrays etc)



Each sample sequenced on one SMRTcell!

Amount of HiFi Revio data

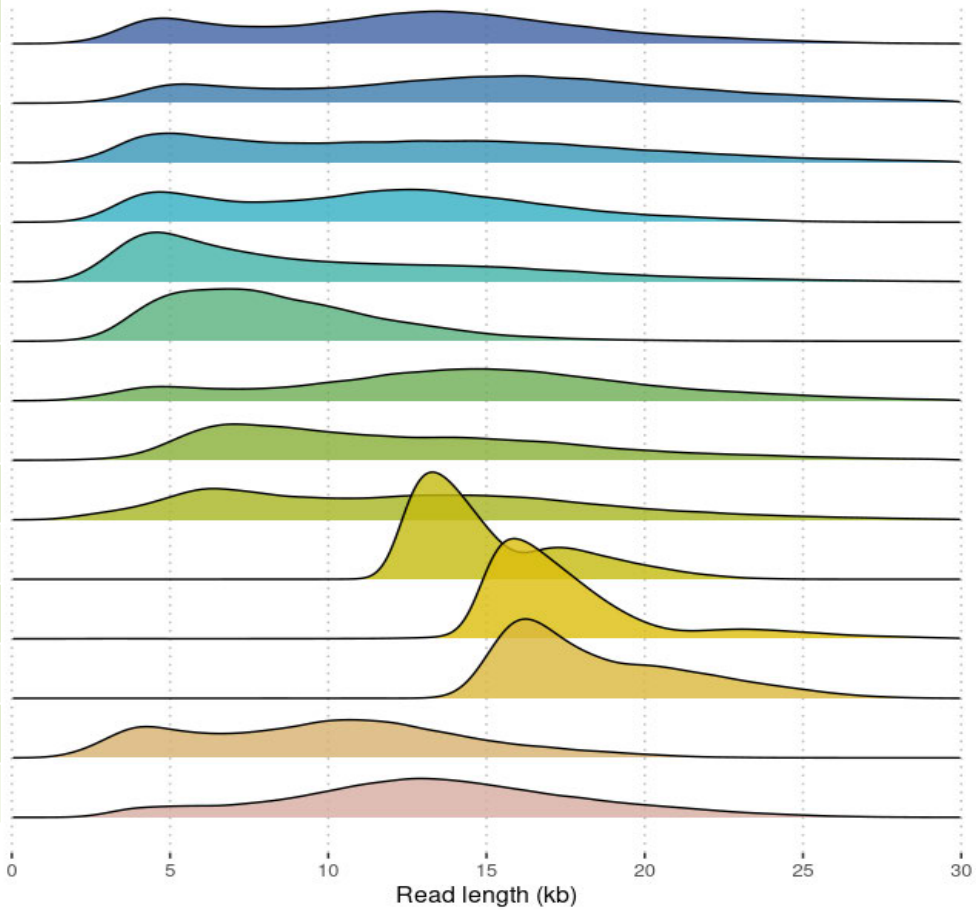
ID	Nr reads	Total yield (Gb)	Avg read length
01	6,710,753	82.59	12,307
02	6,639,606	89.91	14,896
03	6,830,887	85.82	12,564
04	5,785,024	65.99	11,233
05	7,409,630	70.89	9,568
06	7,454,136	62.19	8,343
07	6,934,803	98.93	14,265
08	6,402,650	78.61	12,278
09	6,400,855	78.63	12,284
10	6,622,021	100.0	15,105
11	5,479,327	96.66	17,642
12	5,743,921	106.3	18,506
13	6,359,980	62.64	9,850
14	6,455,409	85.76	13,285



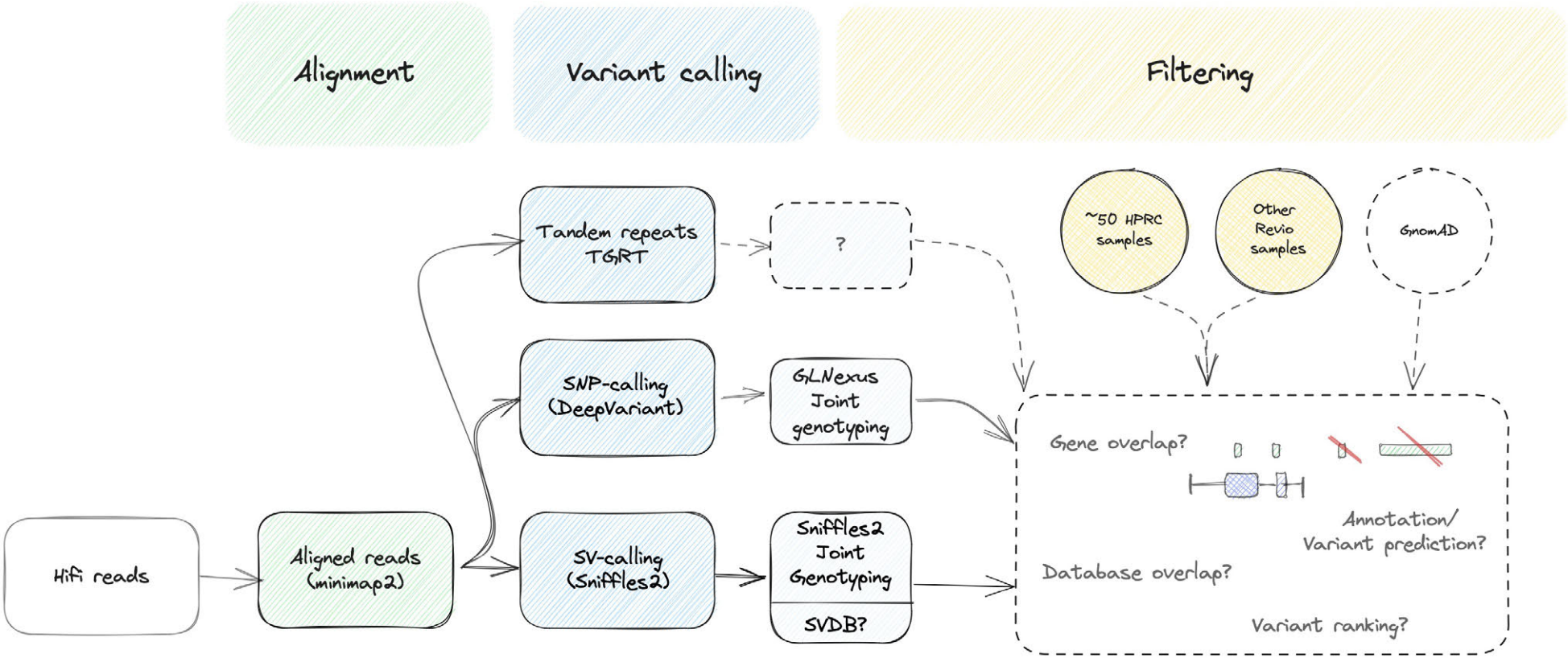
Amount of HiFi Revio data

ID	Nr reads	Total yield (Gb)	Avg read length
01	6,710,753	82.59	12,307
02	6,639,606	89.91	14,896
03	6,830,887	85.82	12,564
04	5,785,024	65.99	11,233
05	7,409,630	70.89	9,568
06	7,454,136	62.19	8,343
07	6,934,803	98.93	14,265
08	6,402,650	78.61	12,278
09	6,400,855	78.63	12,284
10	**6,622,021**	**100.0**	**15,105**
11	**5,479,327**	**96.66**	**17,642**
12	**5,743,921**	**106.3**	**18,506**
13	6,359,980	62.64	9,850
14	6,455,409	85.76	13,285

High quality HMW DNA samples. Size selected on gel



Pipeline for human Revio data



Felix Lenner
github.com/fellen31/skierfe

Results: Variant calling

ID	SNVs (DeepVariant)			SVs > 50bp (Sniffles2)		
	SNPs	Insertions	Deletions	Insertions	Deletions	INV/DUP/BND
01	4.341M	411.9k	412.2k	12,920	9,543	172
02	4.409M	416.9k	426.9k	13,182	9,517	156
03	4.369M	413.1k	423.9k	13,041	9,633	177
04	4.322M	407.9k	396.1k	12,846	9,320	188
05	4.341M	412.4k	405.5k	12,891	9,425	212
06	4.356M	405.4k	414.8k	12,794	9,576	268
				13,331	9,543	181
				13,094	9,595	195
09	4.381M	414.7k	408.7k	13,131	9,478	187
10	4.422M	418.1k	427.0k	13,071	9,444	163
11	4.420M	415.3k	422.1k	13,135	9,488	145
12	4.409M	415.1k	427.5k	13,083	9,535	139
13	4.358M	408.4k	397.7k	12,801	9,481	209
14	4.406M	411.8k	421.6k	12,940	9,474	179
Average	**4.377M**	**412.7k**	**416.3k**	**13,019**	**9,504**	**184**

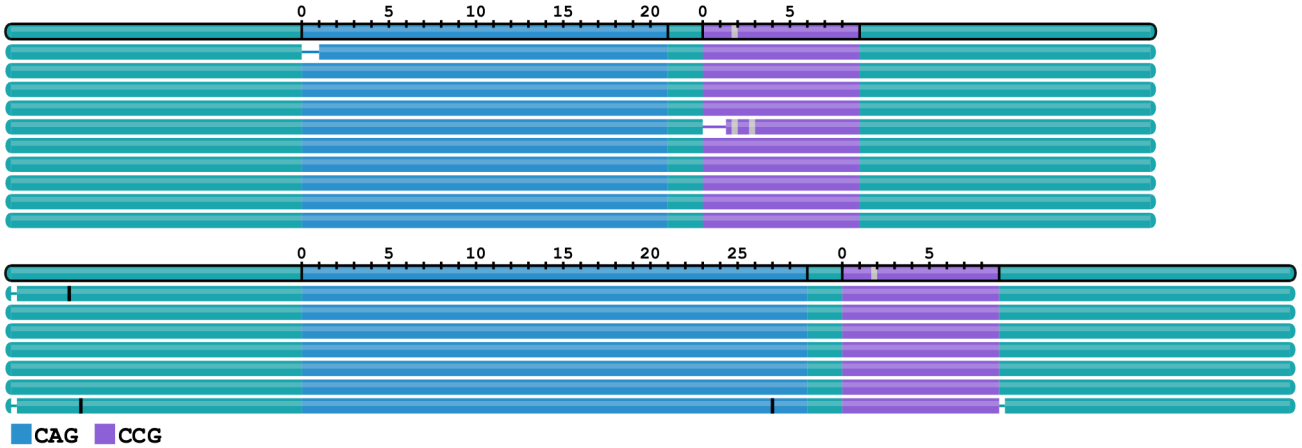
Preliminary result:
~96% of SNVs detected also with short-read WGS

Tandem repeats

HTT

Haplotype 1

Haplotype 2

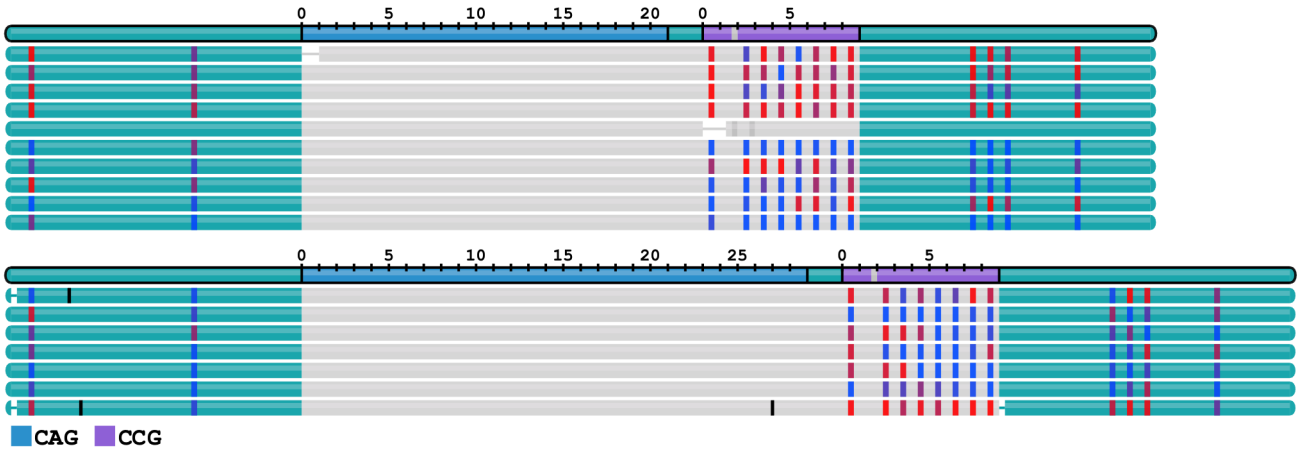


Tandem repeats

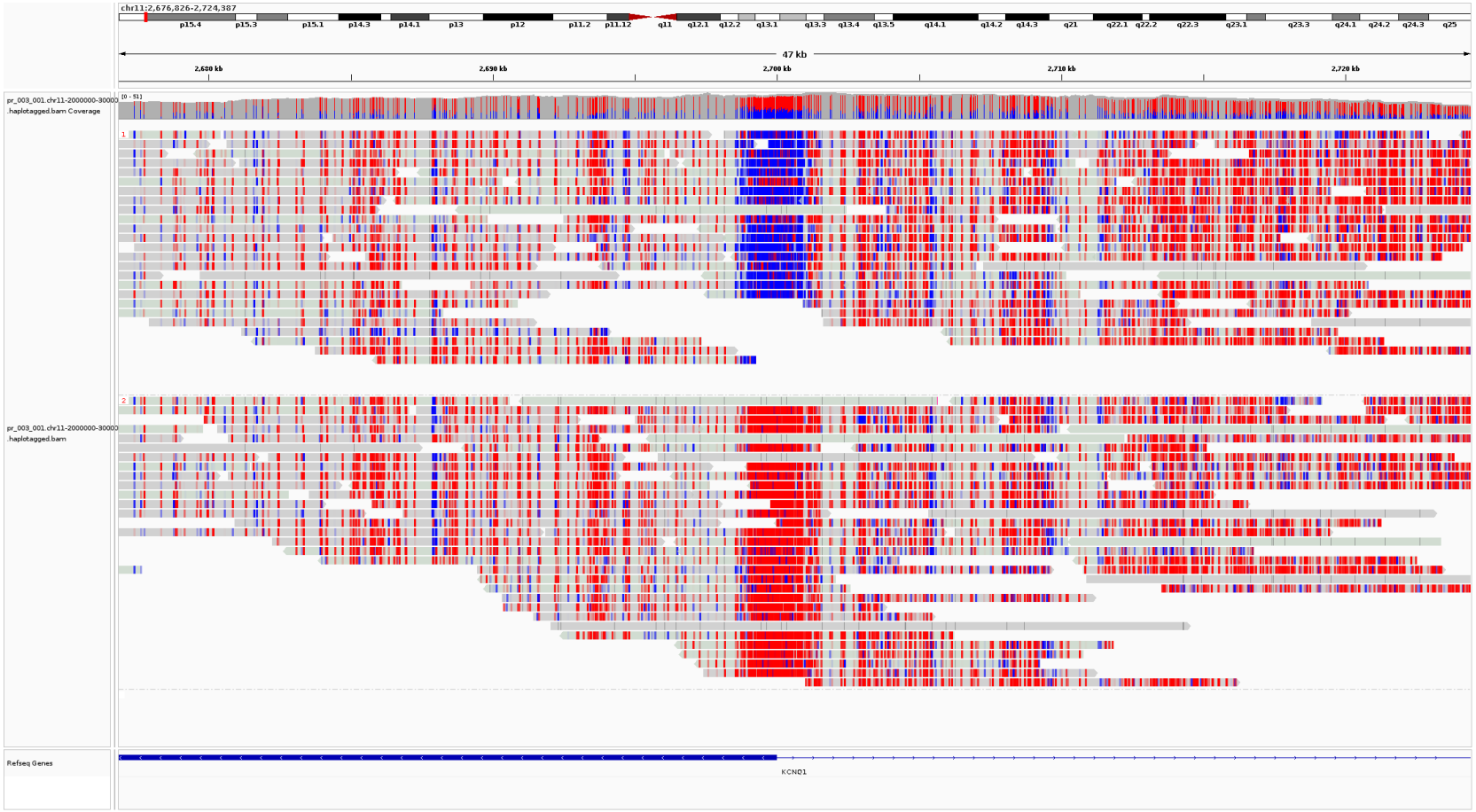
HTT – with methylation

Haplotype 1

Haplotype 2



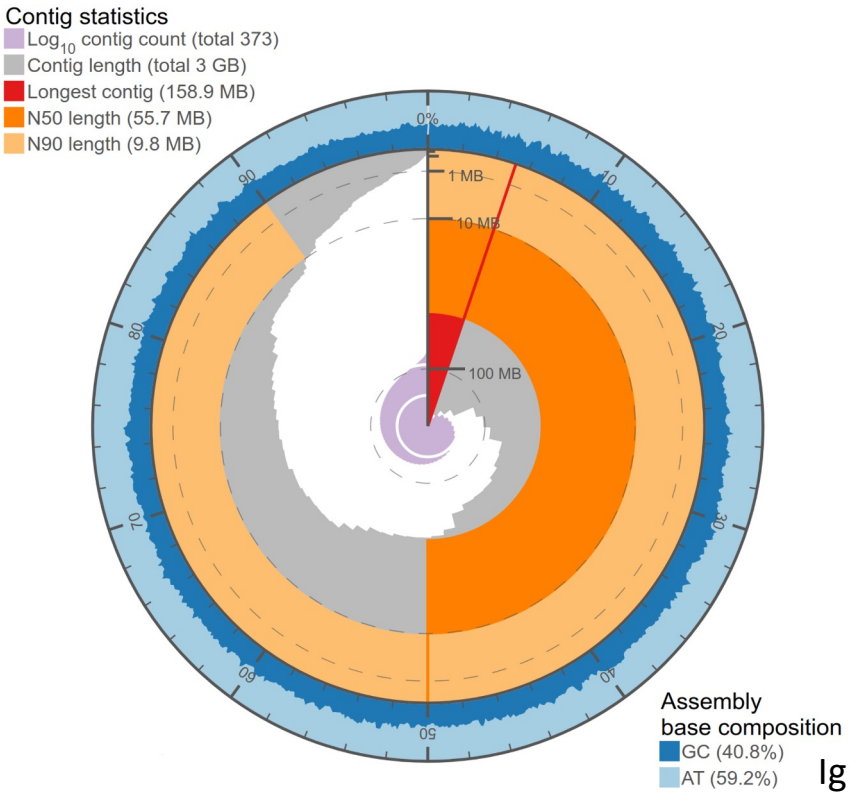
Methylation – known imprinted region



De novo assembly results

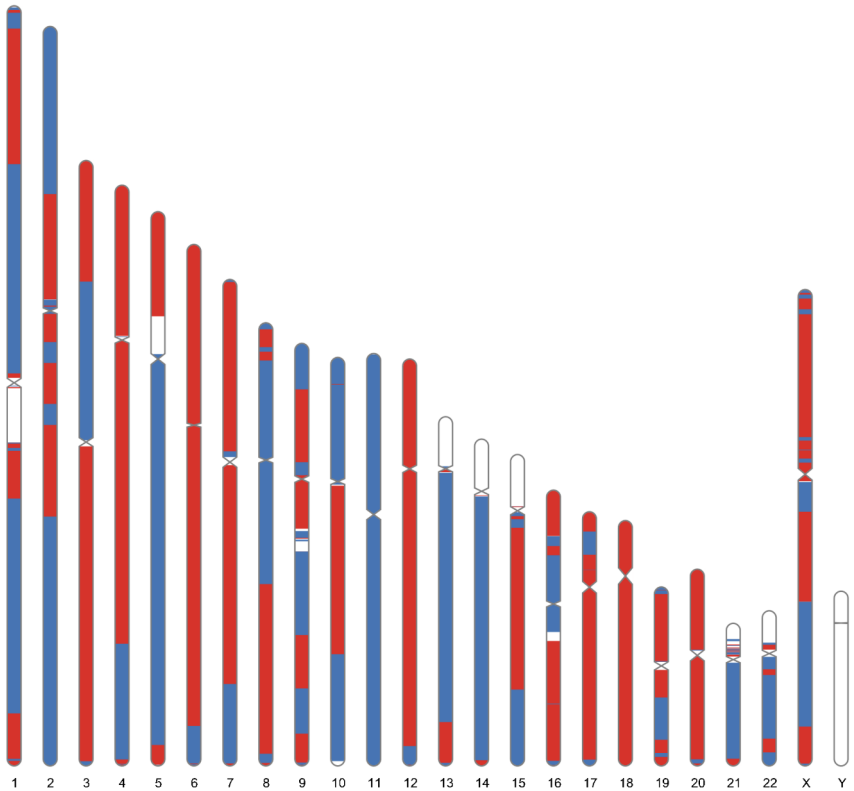
It took just **3.5 h** on a **96** core compute node for *de novo* assembly of a sample with **hifiasm**!

span (Gbp)	3.1
GC (%)	40.84
AT (%)	59.16
longest contig (**Mbp**)	**159**
contig count	373
contig N50 length (**Mbp**)	**56**
contig N50 length count	17
contig N90 length (**Mbp**)	**10**
contig N90 count	59



Ignas Bunikis

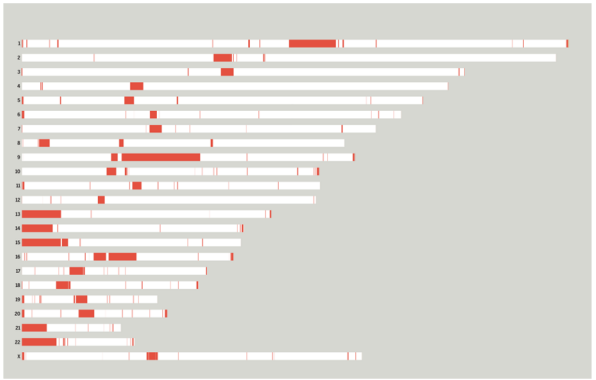
De novo assembly mapped to GRCh38



Colour change represents adjacent contigs

Chromosomes **11** and **18** were assembled in single contigs

...but GRCh38 is missing ~200Mbp of genetic information...

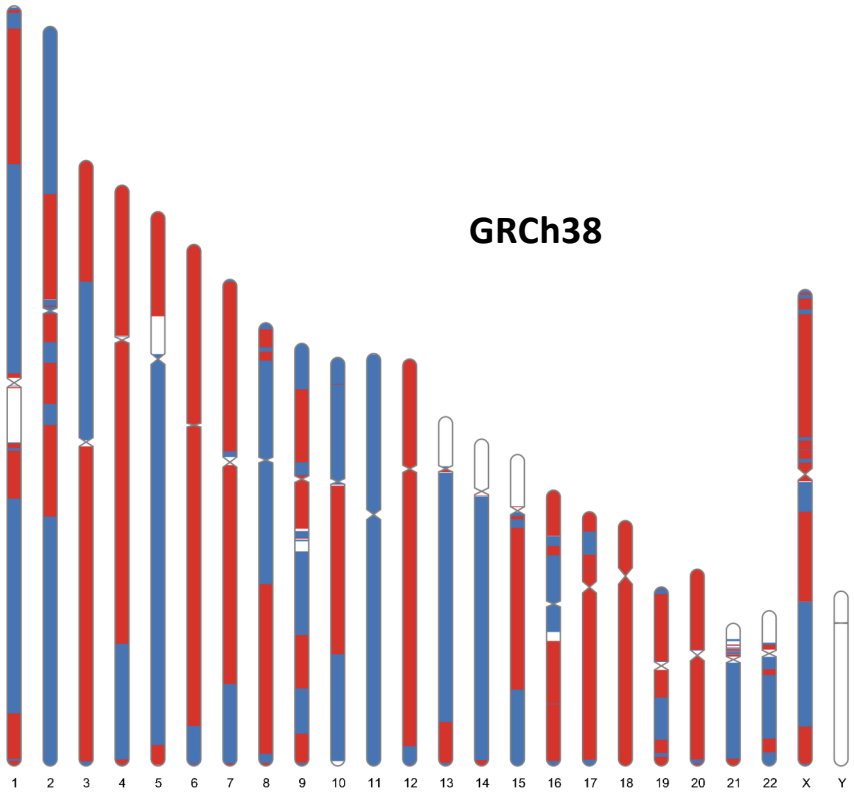


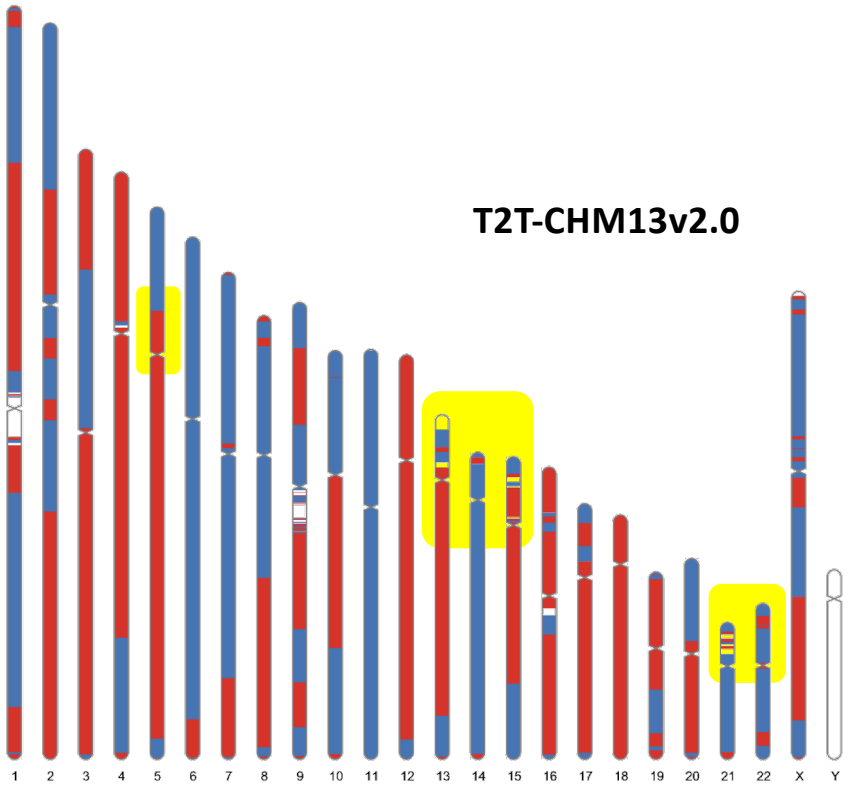
Red segments resolved by T2T Consortium

DOI: 10.1126/science.abp8653

Ignas Bunikis

De novo assembly mapped to T2T



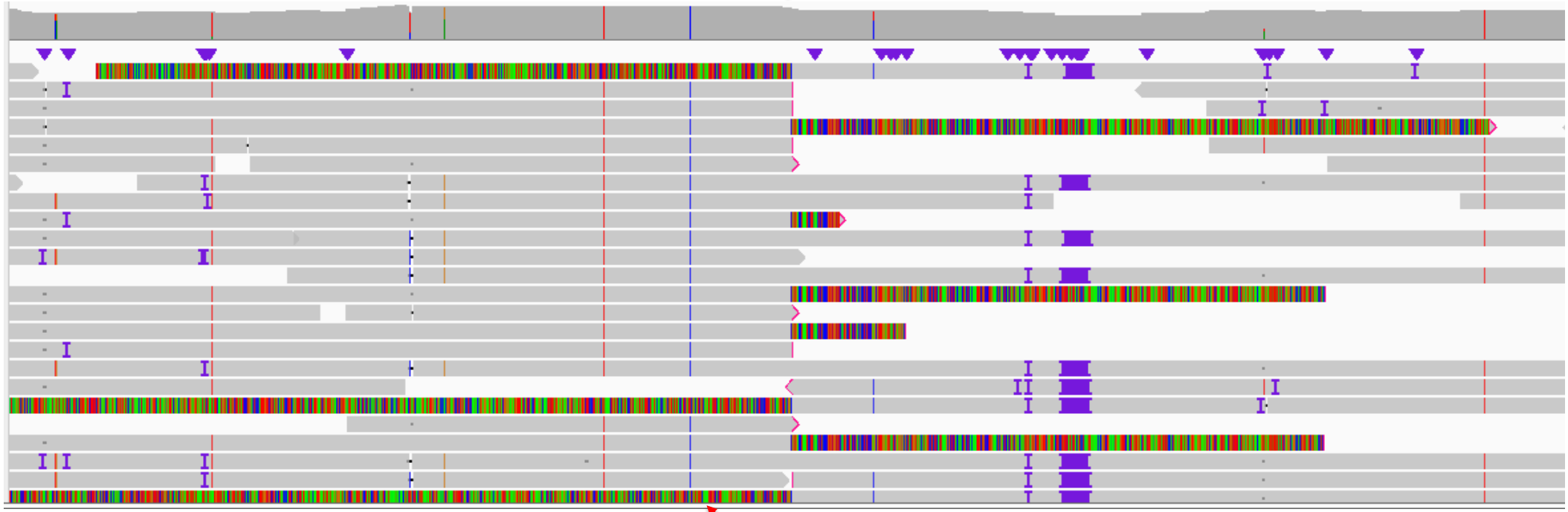


Colour change represents adjacent contigs

Ignas Bunikis

Example of a causative SV breakpoint

Translocation breakpoint

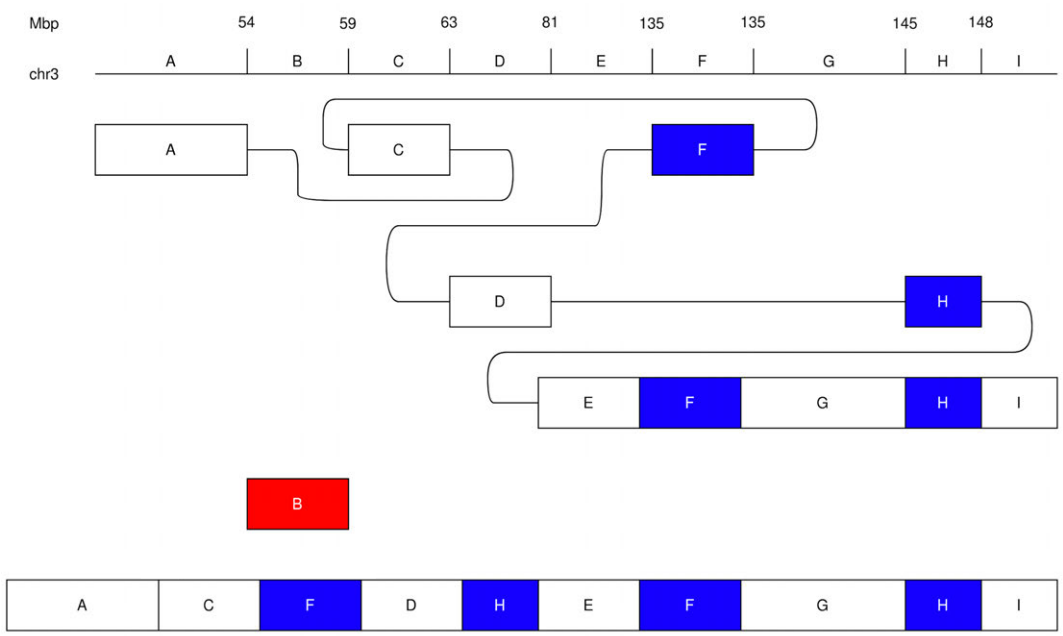


Soft clipped reads, aligning to another chromosome

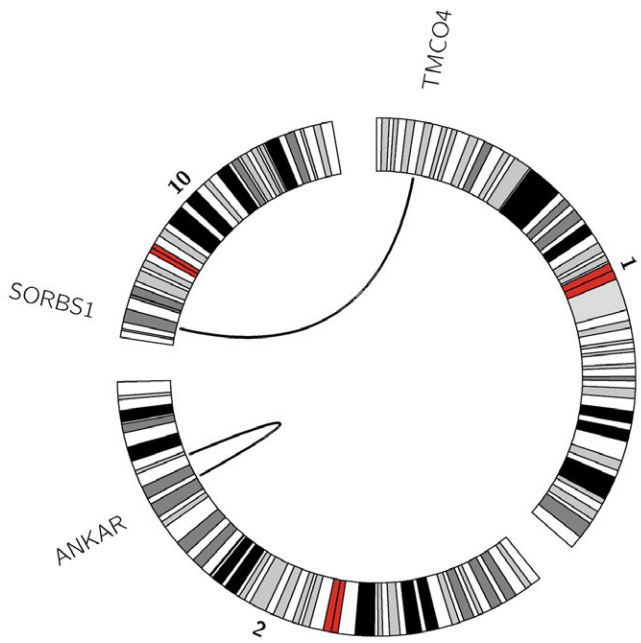
Jesper Eisfeldt

More informative ways to visualize SVs

Subway plots
for complex SV regions

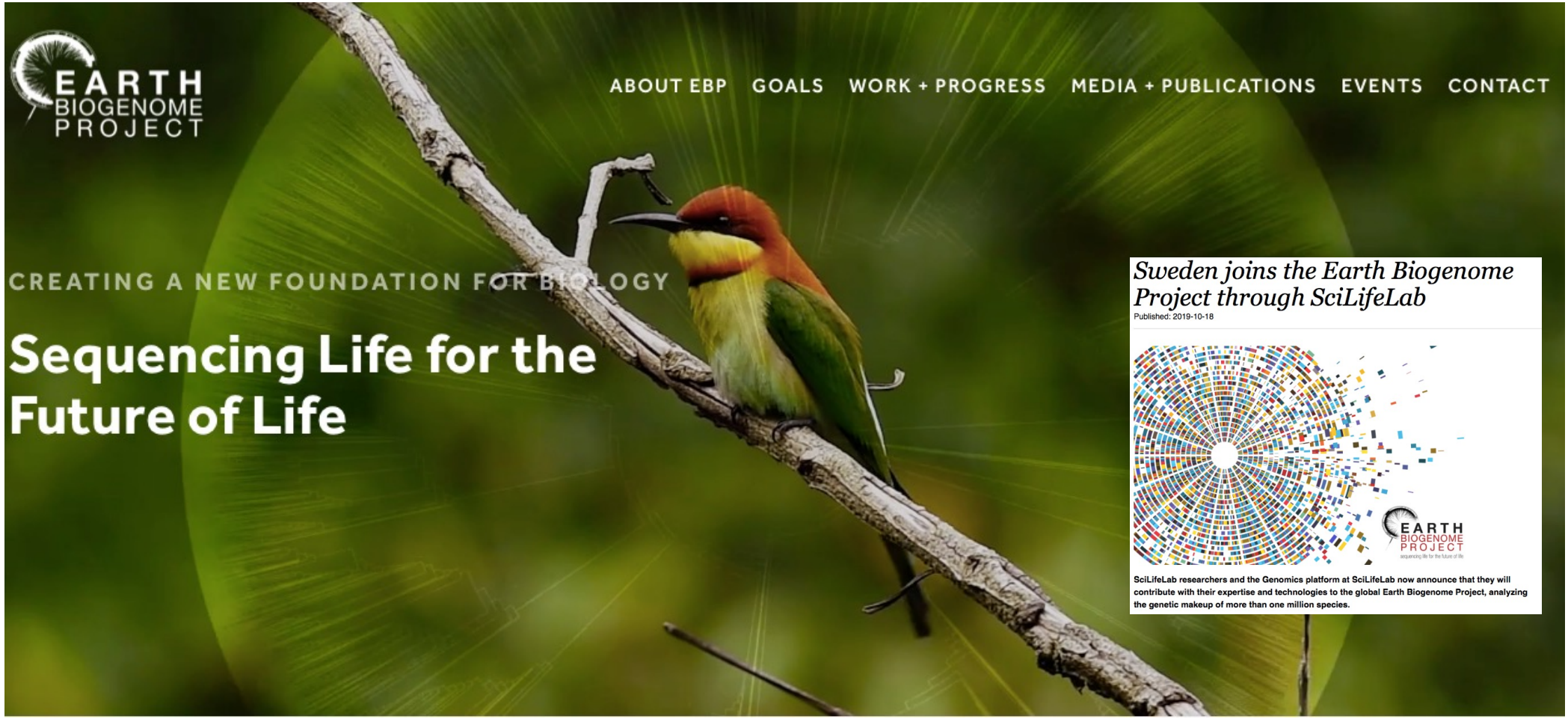


Circos plots
for large-scale SVs and translocations



Jesper Eisfeldt

Example III: Earth Biogenome Project



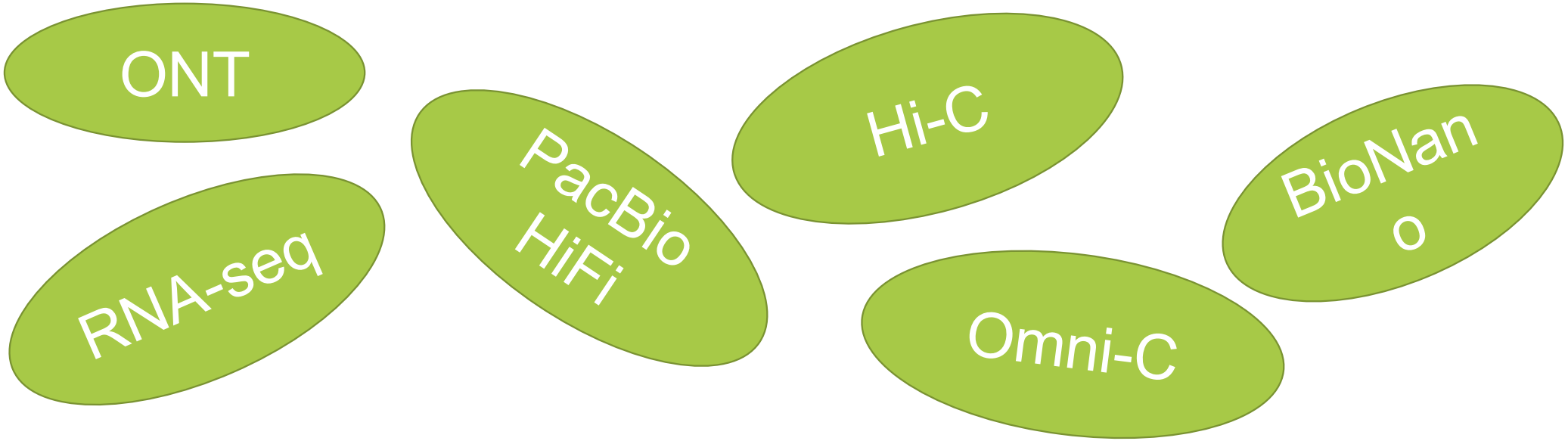
EBP – Data management and analysis

- 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

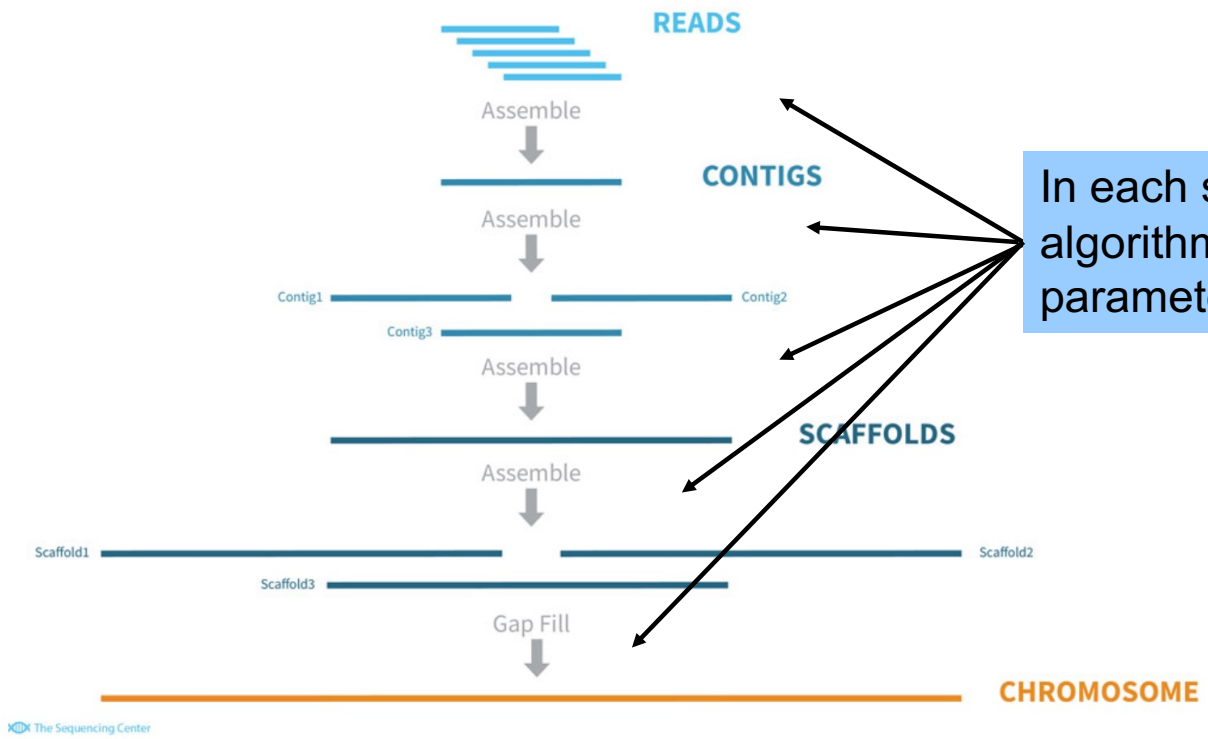
- Make sure sequencing is done using the best technology combination



- 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!

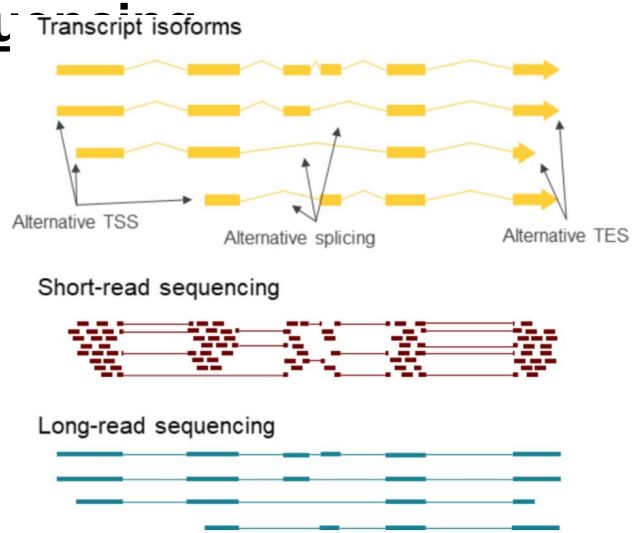
- A challenge for NGI/SciLifeLab is to give best-practice guidelines!

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

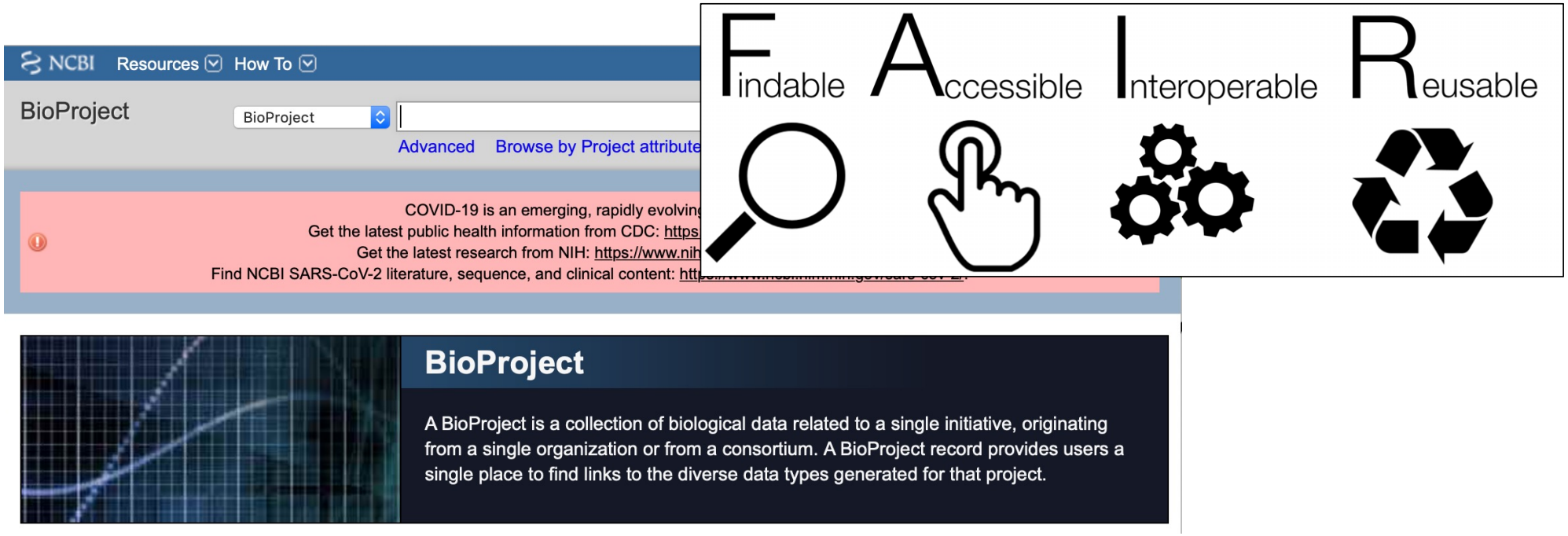
Annotation using RNA-sequencing



- 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

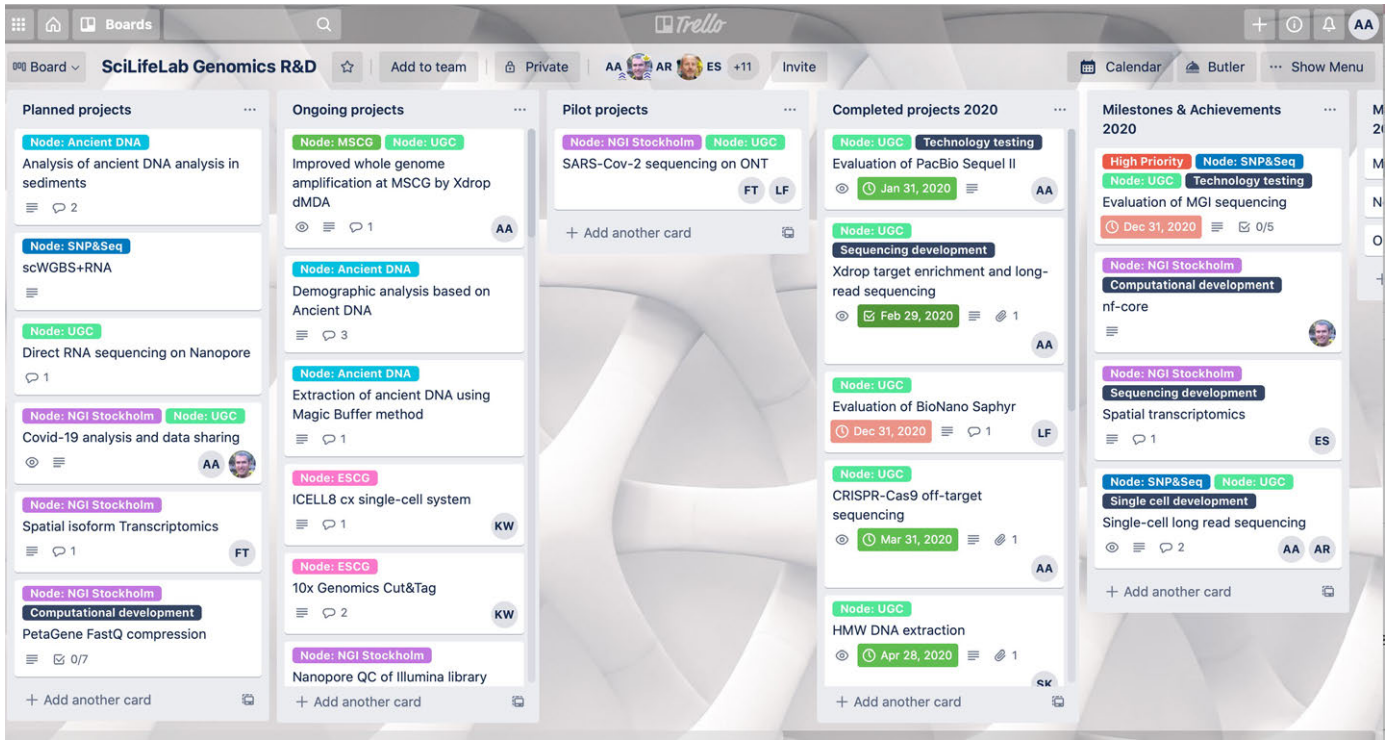
SciLifeLab ⇄ EARTH BIOGENOME PROJECT

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

Internal R&D projects

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



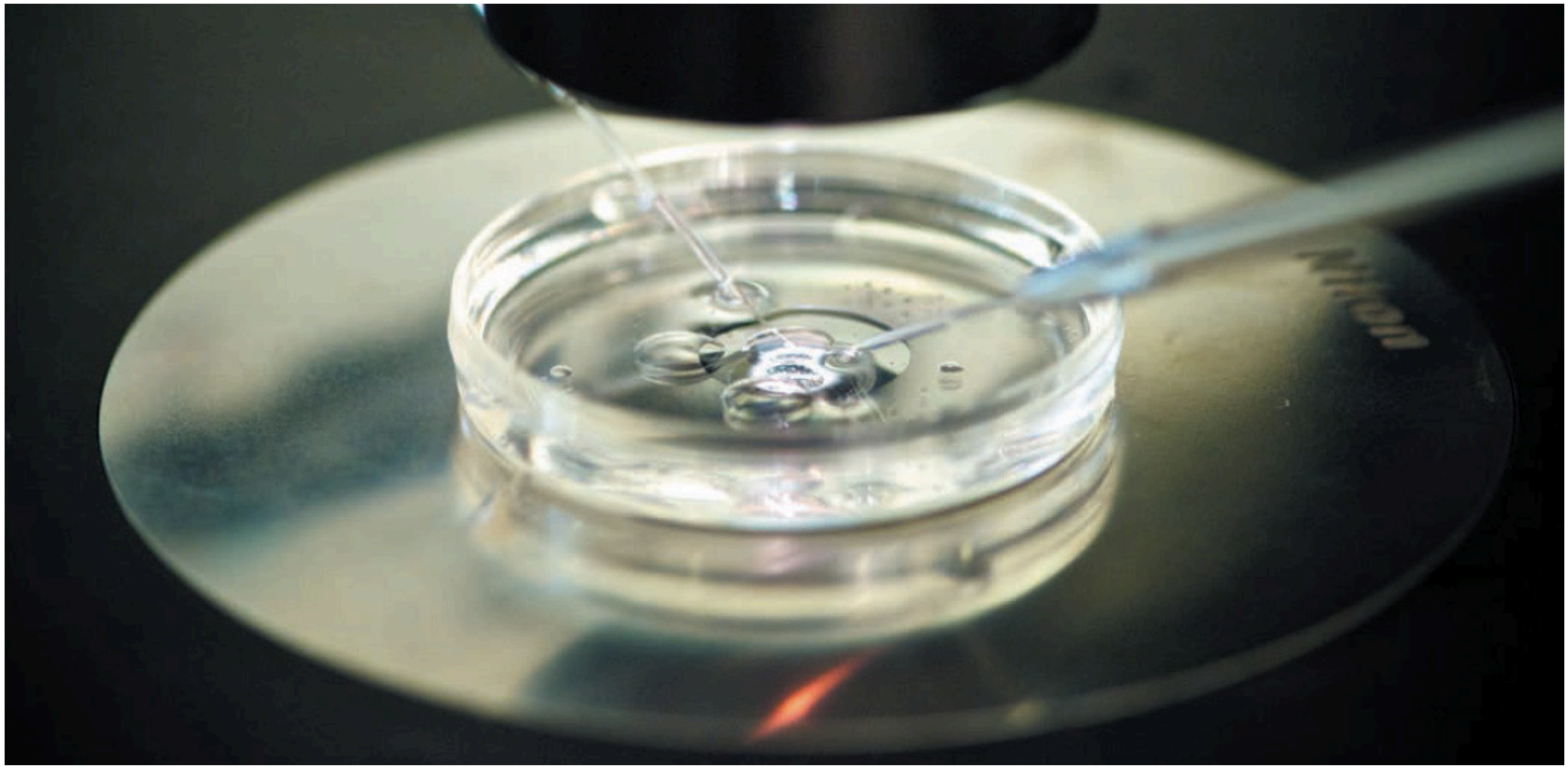
SciLifeLab

Karolinska Institutet

KTH VETENSKAP OCH KONST

Stockholms universitet

UPPSALA UNIVERSITET

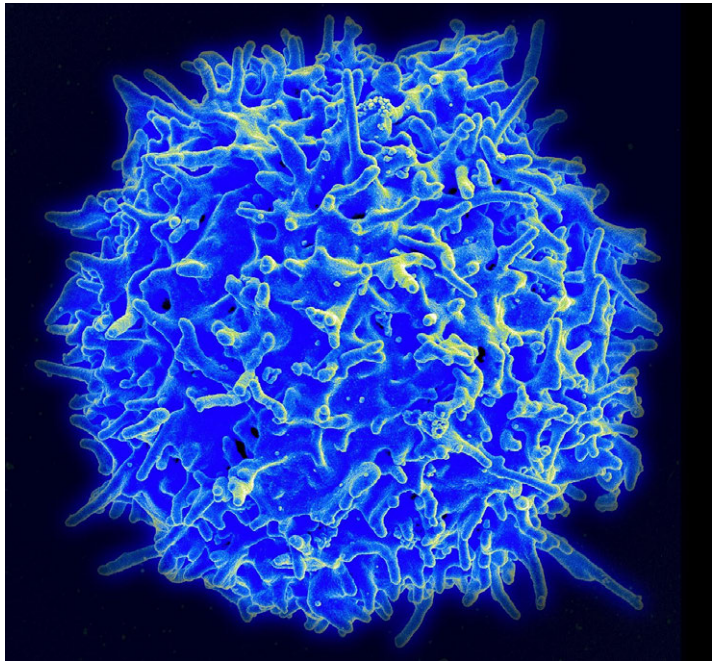


Long-read single cell sequencing

Single-cell transcriptome

Study isoforms in single cells

Single-cell whole genome

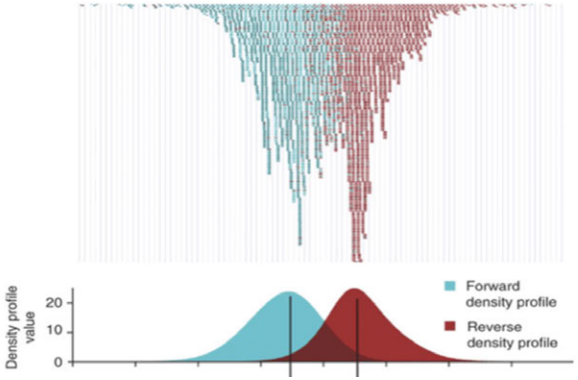


Hård et al, Nature Communications 2023

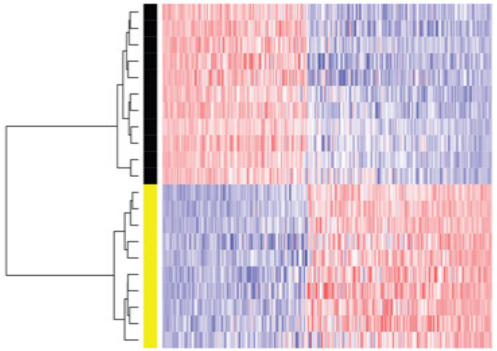
Study structural variation in single-cells

Many topics that have not been covered…

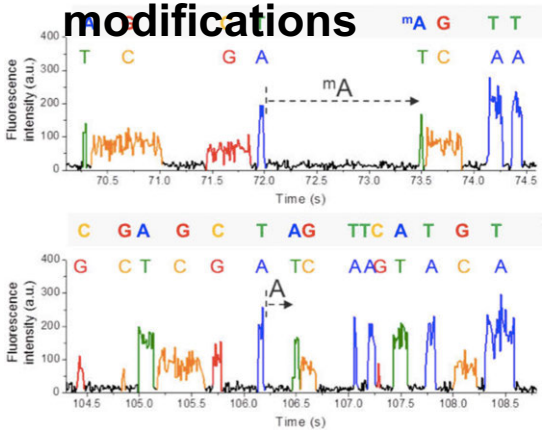
ChIP-seq analysis



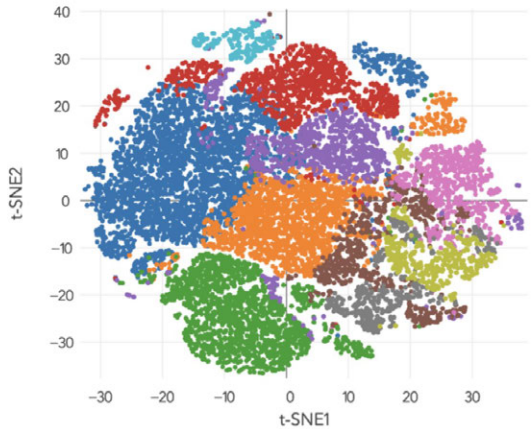
RNA-Seq



DNA base modifications



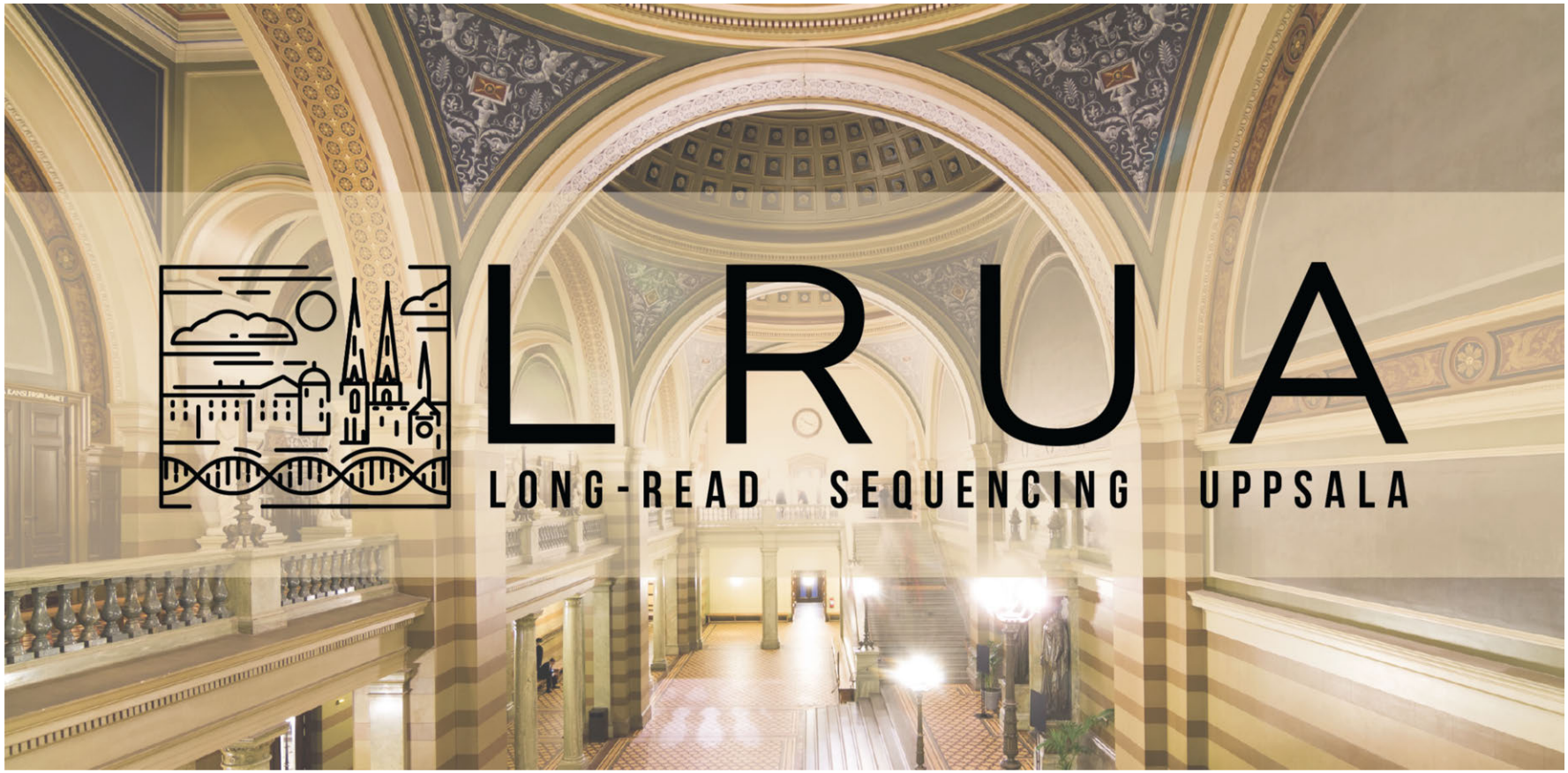
Single cell RNA-Seq



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

Long-read Uppsala Meeting 2024!

- October 21-23 2024, more information soon…



Thanks for your attention!