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

Ancient DNA

We 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

National Bioinformatics Infrastructure (NBIS)

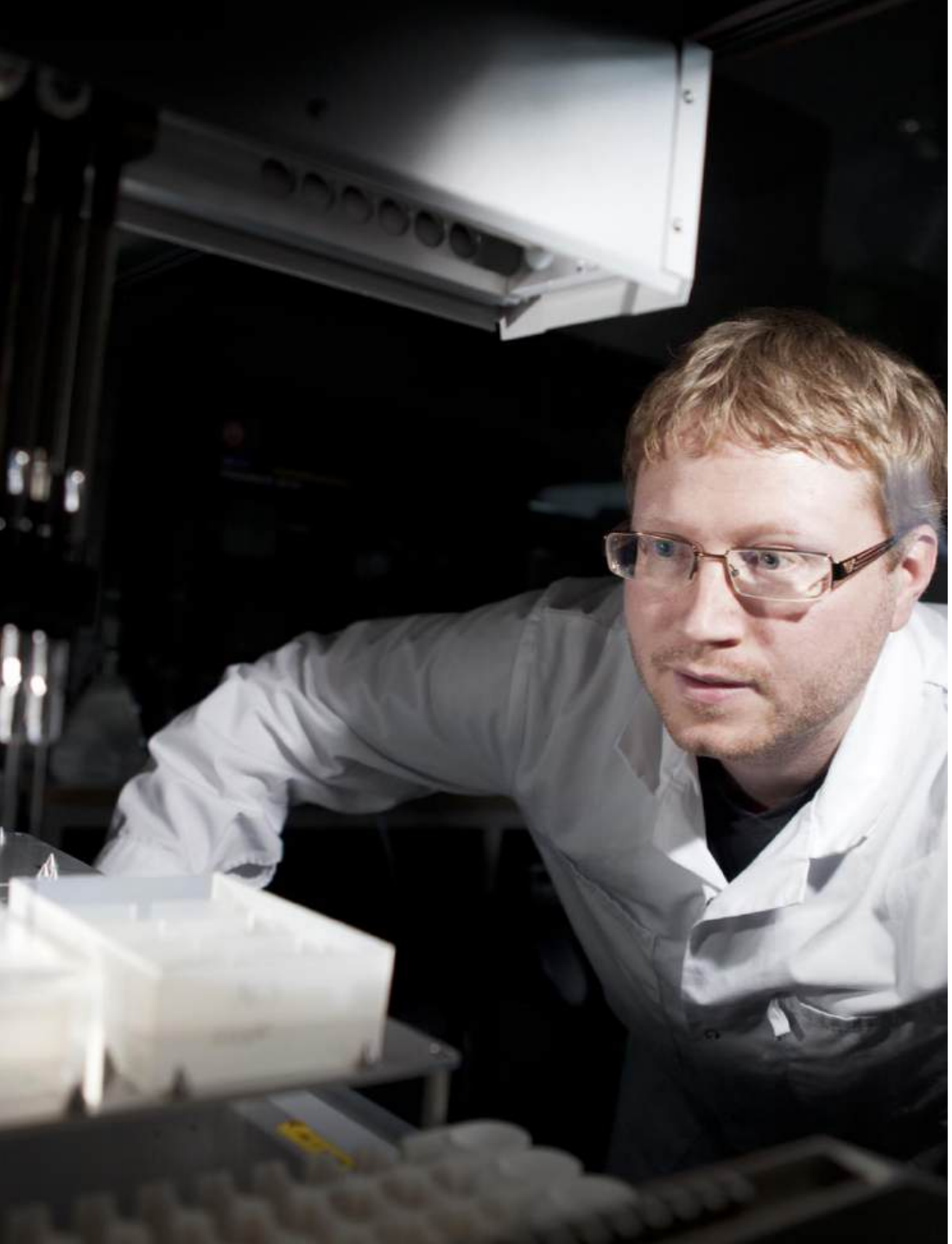
Provides custom-tailored support with analysis of genomics data generated at SciLifeLab or elsewhere, as well as tools and training.

Learn More

National Genomics Infrastructure (NGI)

Provides services for next generation sequencing and SNP genotyping on all scales using a comprehensive range of modern technology.

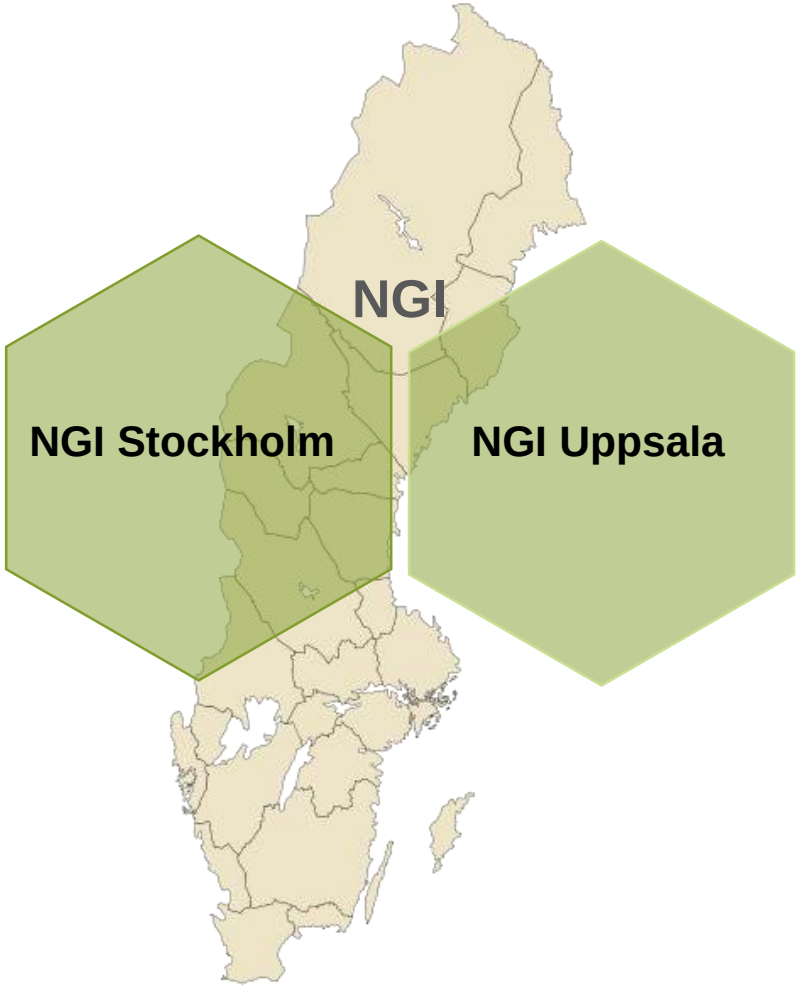
Learn More



What is National Genomics Infrastructure (NGI)?

NGI provides access to technology for next generation sequencing, genotyping, proteomics with NGS read out 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

NGI 2024

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

Amount of sequenced base pairs
- 809 Tbp – short reads
- 24 Tbp – high quality long reads
- 10.9 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

133k

1331

645

833Tbp

250

78

624

72

3407

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 19 different universities, institutes, healthcare and industry companies used NGI services in 2024

Communication tickets
- 42932 ticket updates
- 99% satisfaction score

NGI services

Multi-omics services

Genome Sequencing
De novo, re-seq, targeted...

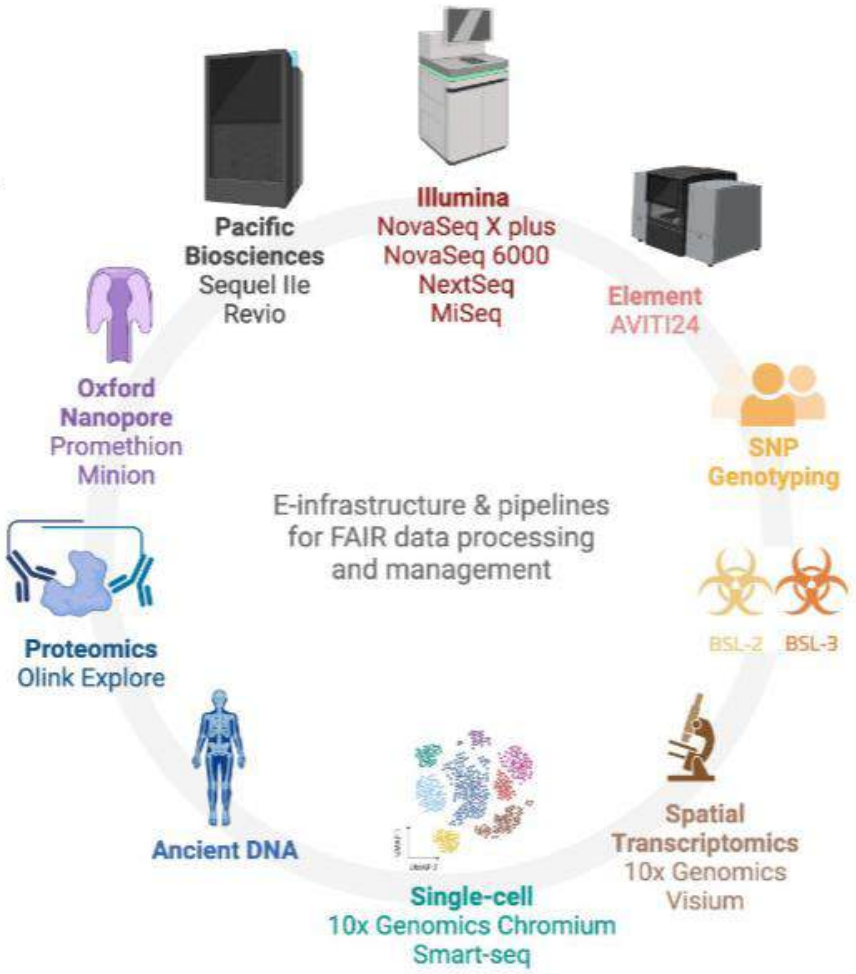
Epigenomics
Methylation, chromatin state, HiC...

Transcriptomics
Short-read, long-read

Proteomics
Olink Explore

Arrays
SNPs, methylation

Source material
Tissues
Cells
Microbes
Plasma
Nucleic acids
Archaeological material
Environmental samples
Read-made libraries



Sequencing instruments at NGI

Short read NGS

High troughput, low cost per base

NovaSeq X Plus

Illumina NovaSeq

Illumina MiSeq

Illumina NextSeq

AVITI, Element Biosiences

Long read NGS

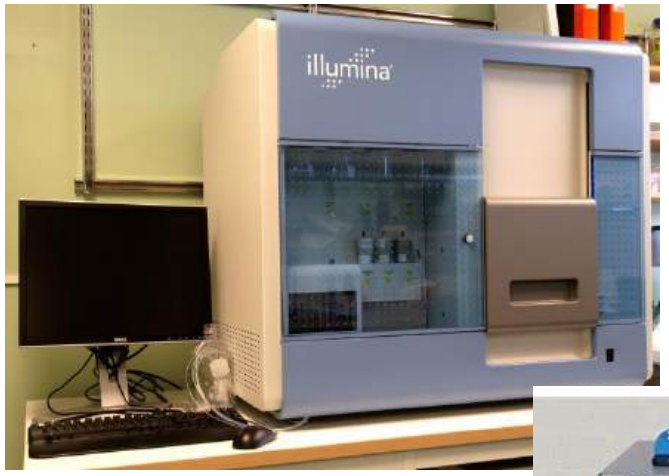
Very long reads, lower troughput

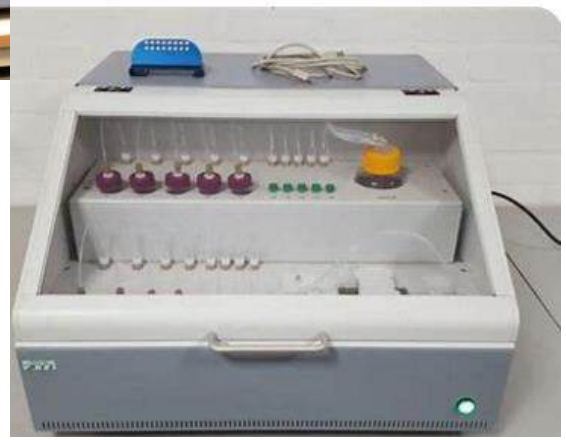
PacBio Revio

Oxford Nanopore-PromethION

15 years of Illumina sequencing at NGI

2007: Installation of Illumina GA

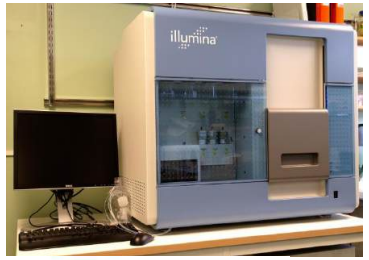




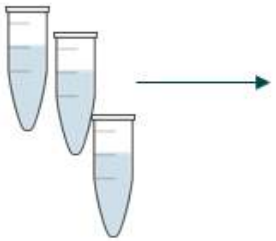
2023: Arrival of NovaSeq X Plus

Instruments over the years

Sequencer	Launched	Runtime	Output	Cost per Base
Illumina Genome Analyzer (GA)	2006	~3 days	~1 Gb per run	~$0.10
HiSeq 2000	2010	~10 days	Up to 600 Gb per run	~$0.01
HiSeq 2500	2012	~1 day in rapid mode	Up to 600 Gb per run	~$0.01
HiSeq X Ten	2014	~3 days per run	~1.8 Tb per run	<$0.01
NovaSeq 6000	2017	~13–44 hours	Up to 6 Tb per run	~$0.005
NovaSeq X / X Plus	2022	~24–48 hours	Up to 20 Tb per run	<$0.004

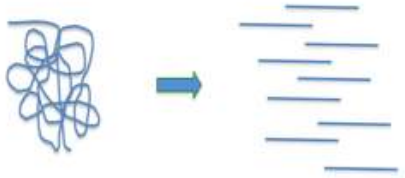


Workflow, Illumina sequencing

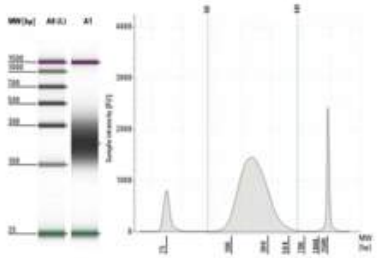


Quality control of template (DNA, RNA, amplified regions)

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

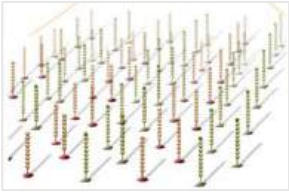
Library preparation

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

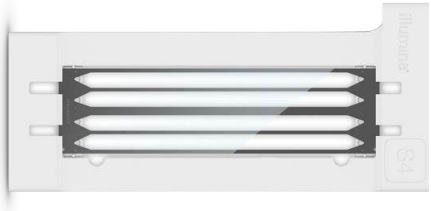
- Approx. 25 protocols for Illumina library prep at NGI

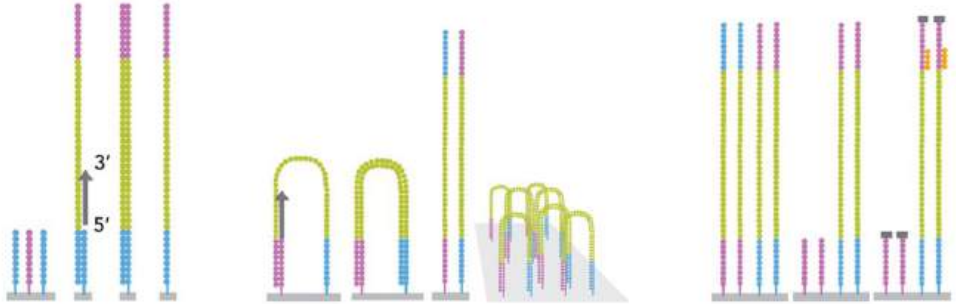
P5 Adapter -- [Index 1] -- [Read 1 Primer Site] -- [Insert DNA] -- [Read 2 Primer Site] -- [Index 2] -- P7 Adapter

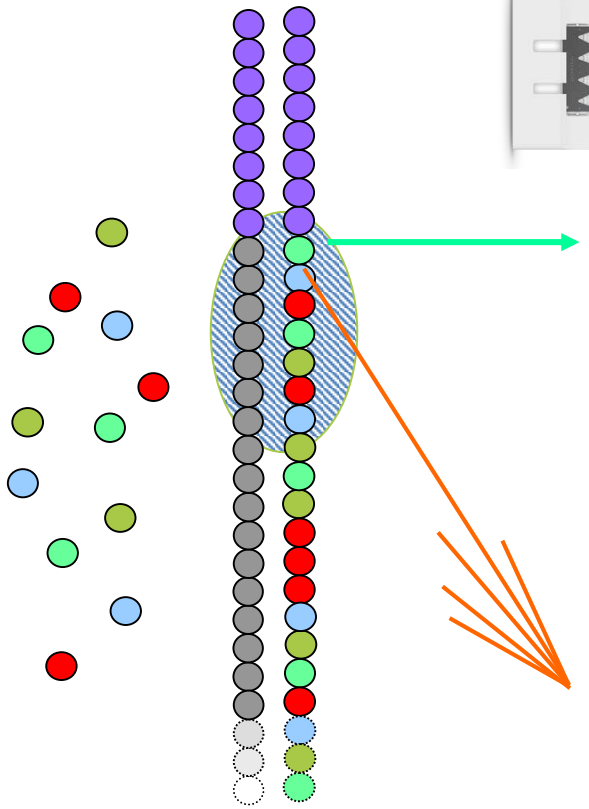
Illumina cluster generation & sequencing



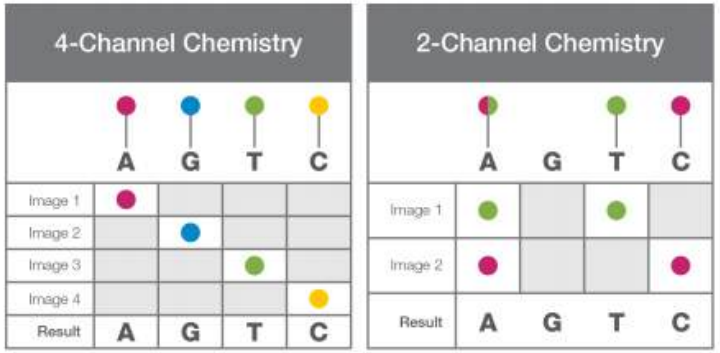
- The sequencing library is hybridized to a flowcell (”cluster generation”)
    - - A flowcell is a slide that is coated with oligos
- Rapid bridge amplification
- Hybidization of sequencing primers
- Sequencing by syntehsis
    - fluorophore labeled nucleotides emitting light

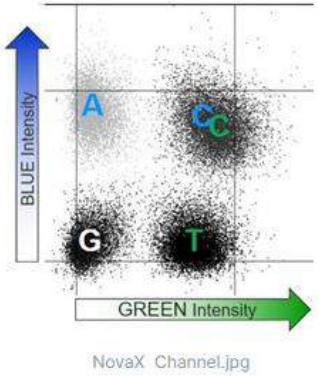






Illumina sequencing by synthesis





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

NovaSeq X Plus – new instrument!

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%

Advantages and challenges NovaSeqX

+

Cost per base is low

Quick data generation

Easy workflow in the lab

Reagents shipped in RT

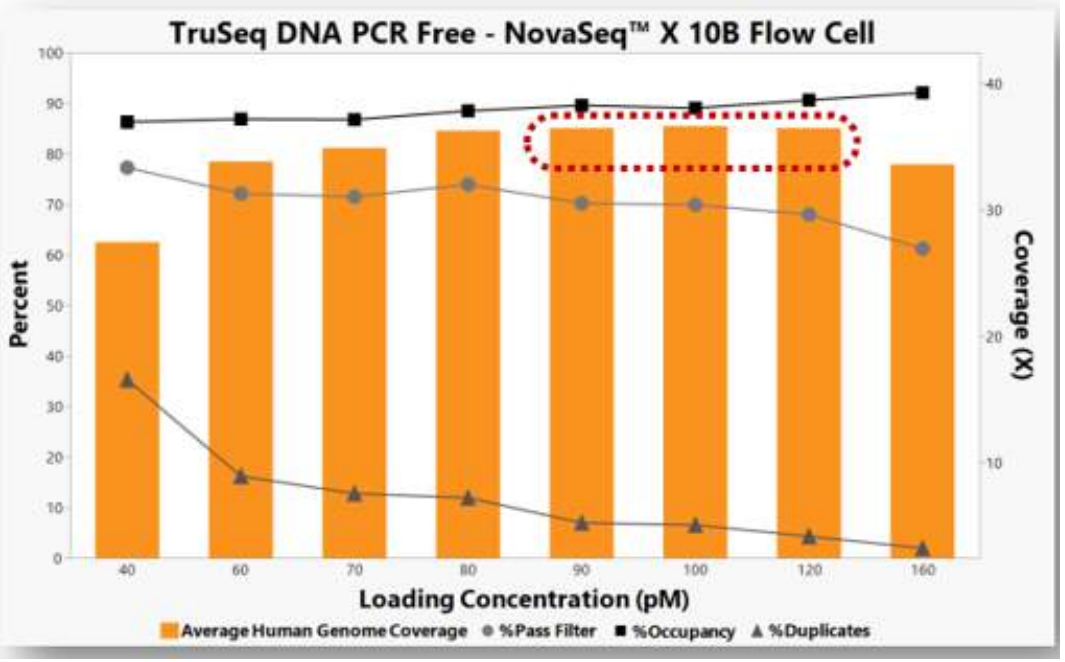
On-instrument analysis

-

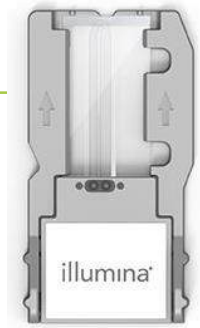
Yield vs duplicates

More sensitive to challenging samples and short inserts

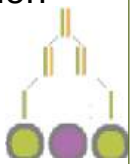
Sensitive to colour balancing (C-A)



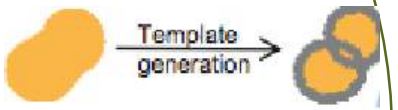
Duplicates, duplicates, duplicates….



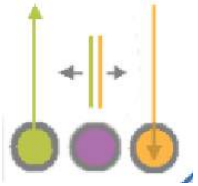
PCR duplicates
Duplicate molecules arising from amplification in library prep



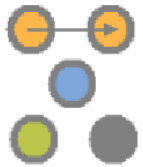
Optical duplicates
A single cluster falsley called as two by RTA



Sister duplicates
Complement strands of the same libary form independent clusters



Clustering duplicates*
Duplicates in nearby wells, a library occupies two nearby wells during cluster generation
”Optical duplicates”



Present on all Illumina Platforms

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

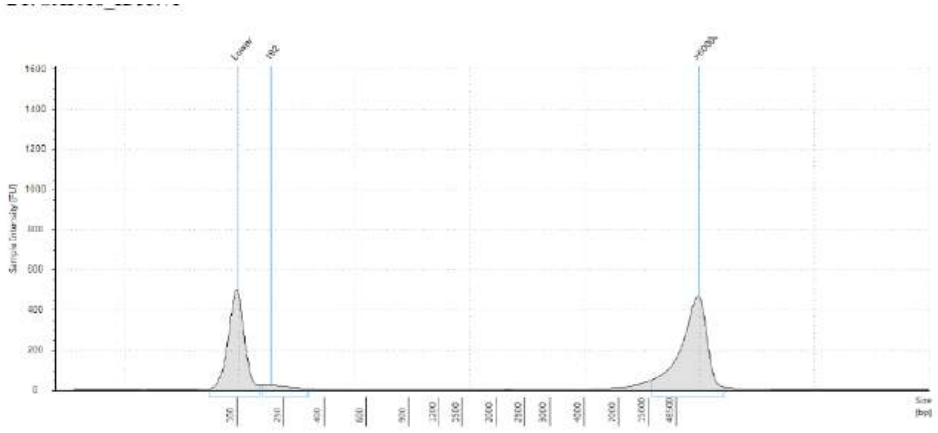
Patterned flowcells only (NextSeq, NovaSeq 6000/X

Quality control of RNA/DNA

DNA

Concentration: QuantIT

Degradation: Fragment Analyzer/TapeStation



Sample Table

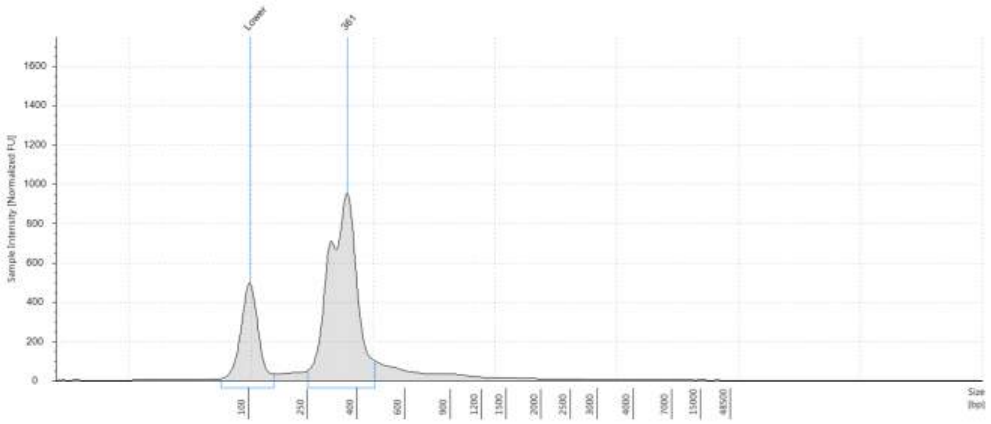
Well	DIN	Conc. [ng/µl]	Sample Description	Alert	Observations
B1	9.6	16.0	SX1018_ID33.v1		

High quality DNA sample

RNA

Concentration + RIN-value:

Fragment Analyzer/TapeStation



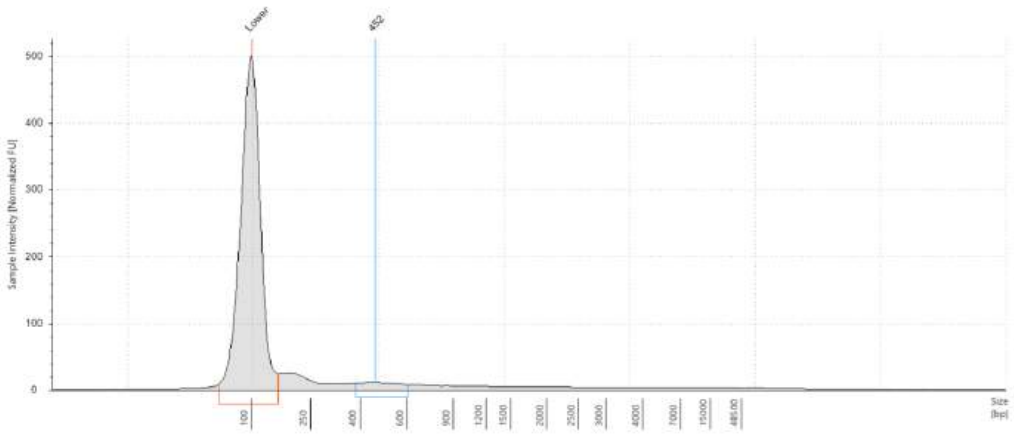
Sample Table

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



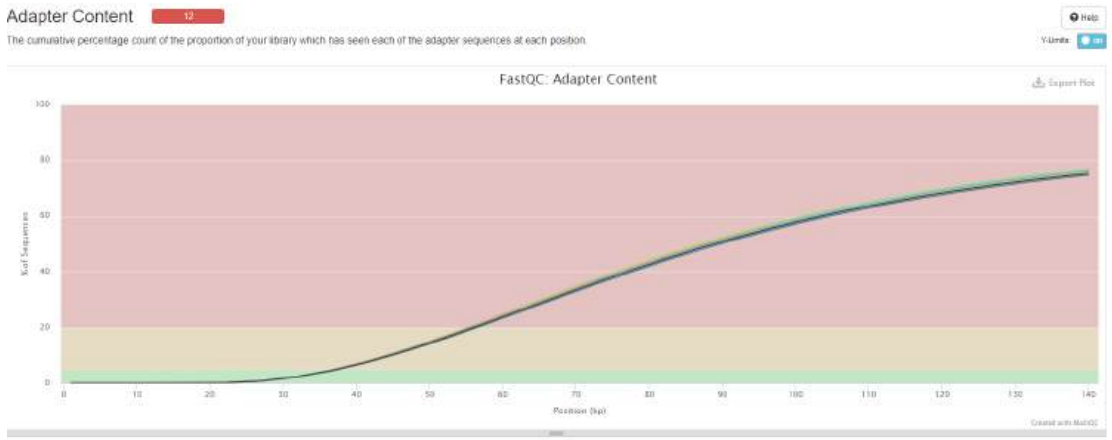
Sample Table

Well	DIN	Conc. [ng/ul]	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

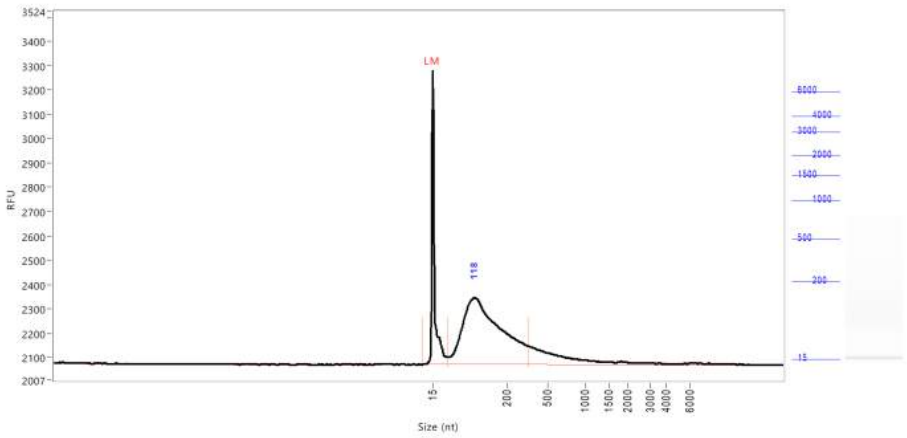


Copy table | Configure Columns | Plot | Showing 7/7 rows and 14/23 columns.

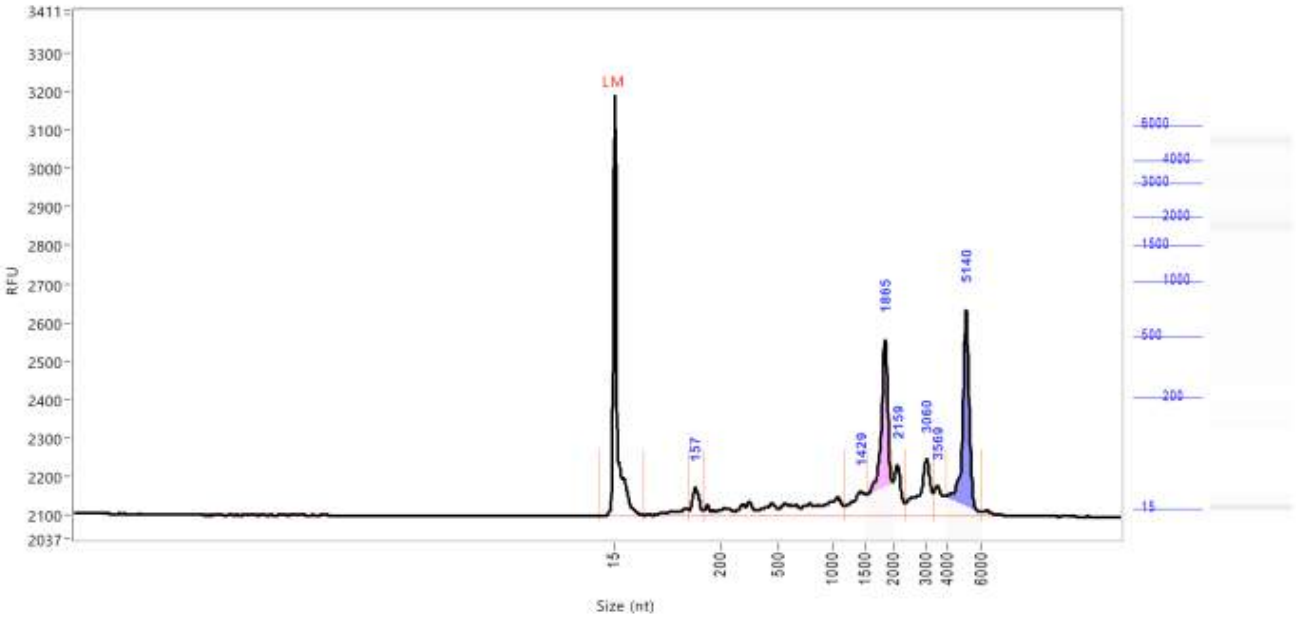
Sample Name	% GC	Ins. size	≥ 30X	Coverage	% Aligned	Change rate	Ts/Tv	M Variants	TiTV ratio (known)	TiTV ratio (novel)	% Dups	% Dups	% GC	M Seqs
S1	46%	55	11.1%	2.0X	98.2%	893	1.645	3.47	2.0	1.6	76.6%			

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

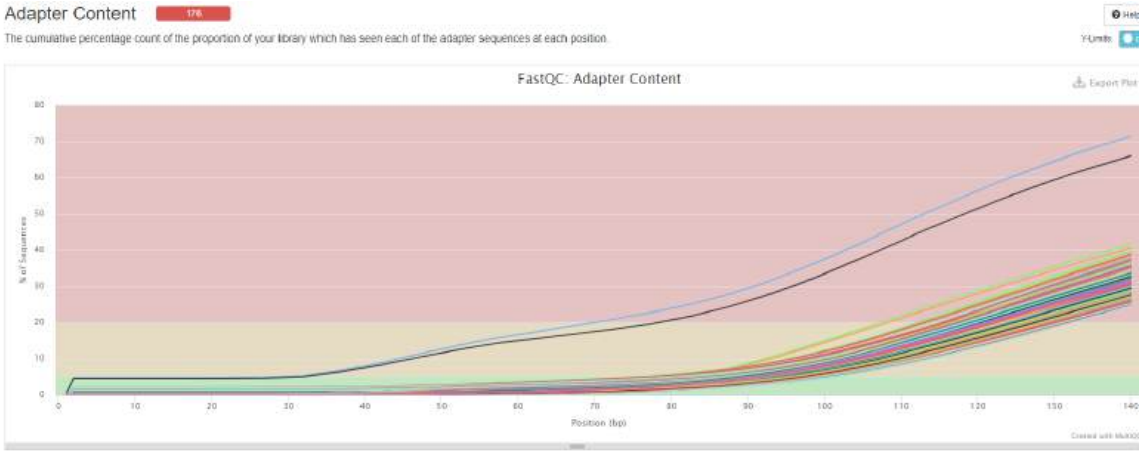
QC-reults RNA-seq

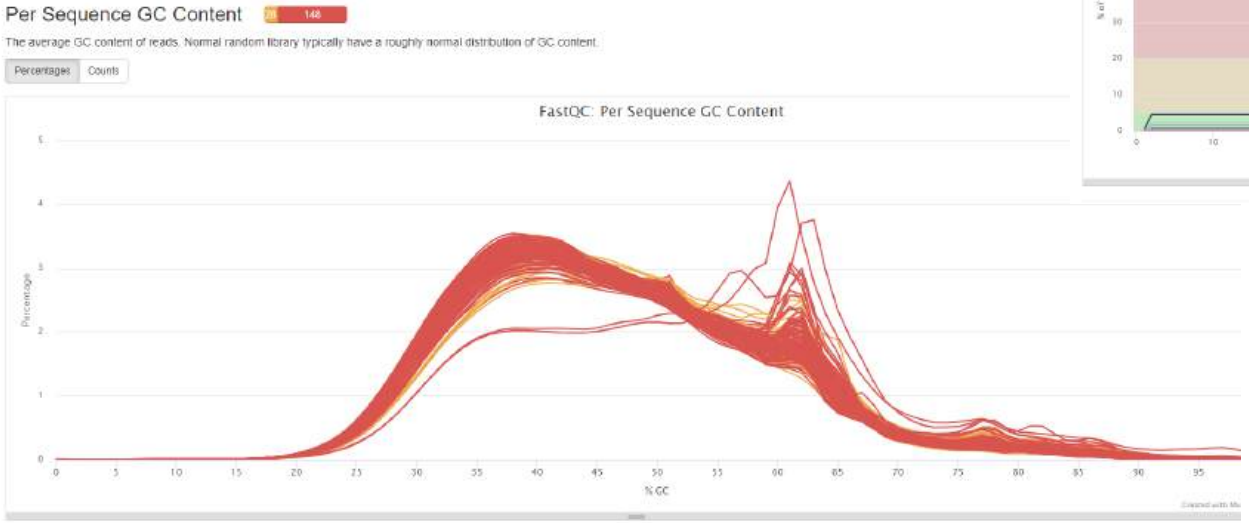
Uneven amounts of data (17-100 M reads per sample)

A lot of duplicates

High rRNA content

High adapter content

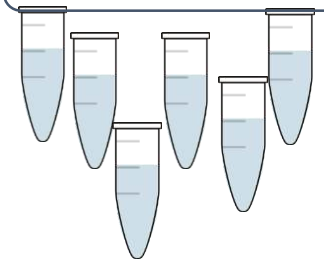




Some of the applications offered

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

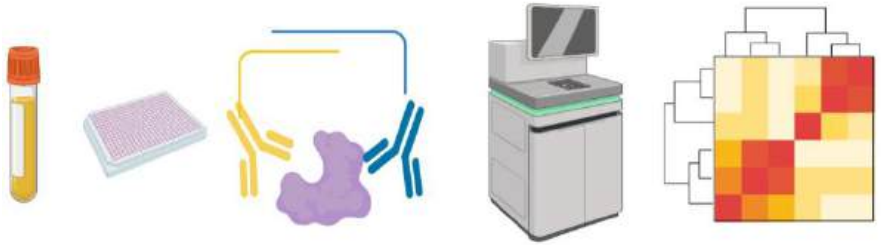
Spatial transcriptomics

- 10x Genomics Visium

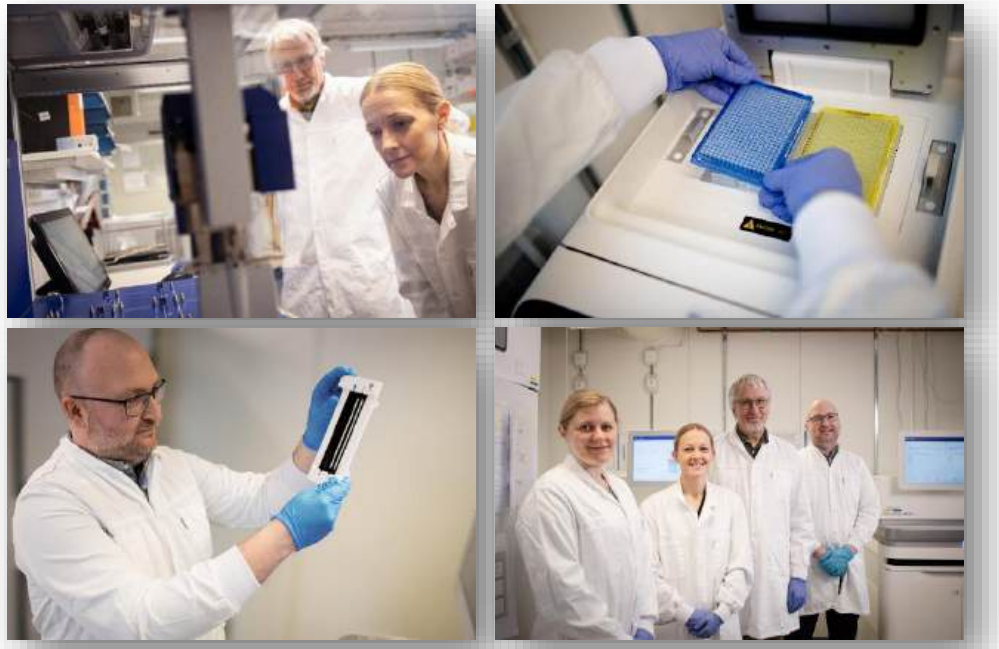
Proteomics with NGS readout

- 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 2021-2023
  - >30 000 samples analyzed in 2024

New instrument – AVITI, Element Biosciences

Category	Specification
Technology	Sequencing by Binding (SBB)
Applications	Genomics, transcriptomics, single-cell sequencing
Read Lengths	Up to 2×150 bp (paired-end reads)
Data Output	Up to 300 Gb per run
Run Time	24-36 hours
Library Compatibility	Standard NGS library preparation protocols
Limitations	Not ideal for ultra-high-throughput projects

Examples, recent successful projects

Massively parallel analysis of single-molecule dynamics on next-generation sequencing chips

J. AGUIRRE RIVERA, G. MAO, A. SABANTSEV, M. PANFILOV, Q. HOU, M. LINDELL, C. CHANEZ, F. RITORT, M. JINEK, AND S. DEINDL

Authors Info & Affiliations

SCIENCE • 22 Aug 2024 • Vol 385, Issue 6711 • pp. 892-898 • DOI: 10.1126/science.adn5371

5,195

nature genetics

Explore content | About the journal | Publish with us

nature > nature genetics > comment > article

Comment | Published: 21 October 2024

Pushing the boundaries of rare disease diagnostics with the help of the first Undiagnosed Hackathon

Angelica Maria Delgado-Vega, Helene Cederroth, Fulya Taylan, Katja Ekholm, Marlene Ek, Håkan Thonberg, Anders Jemt, Daniel Nilsson, Jesper Eisfeldt, Kristine Bilgrav Saether, Ida Höijer, Ozlem Akgun-Dogan, Yui Asano, Tahsin Stefan Barakat, Dominyka Batkovskyte, Gareth Baynam, Olaf Bodamer, Wanna Chetruengchai, Pádraic Corcoran, Madeline Couse, Daniel Danis, German Demidov, Eisuke Dohi, Mattias Erhardsson, ... Ann Nordgren

+ Show authors

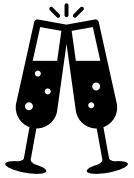
Nature Genetics **56**, 2287–2294 (2024) | Cite this article

NGI OpenLab – opening soon!

New interactive NGS space opening at BMC in Uppsala
January 2025

Launch party at BMC Jan 16, 14.00-16.00

Event Information



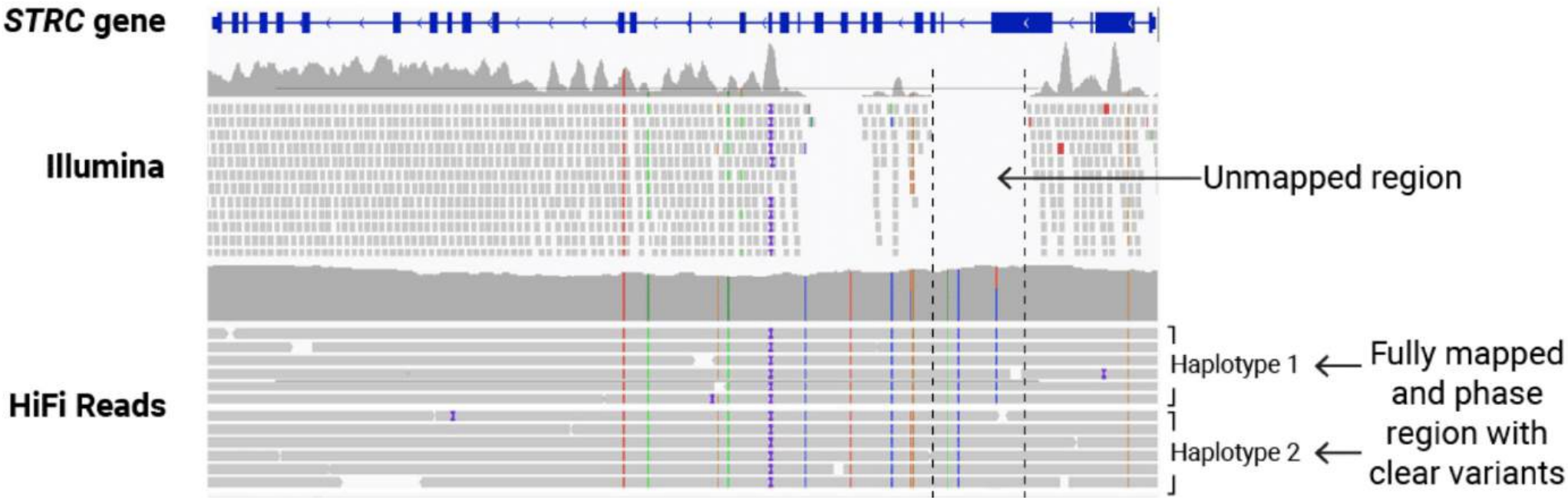
Long-read sequencing, data analysis pipelines, and development projects at NGI

Adam Ameur

National Genomics Infrastructure, SciLifeLab, Uppsala, Sweden

Limitations with short reads

- You don’t get complete genome information!

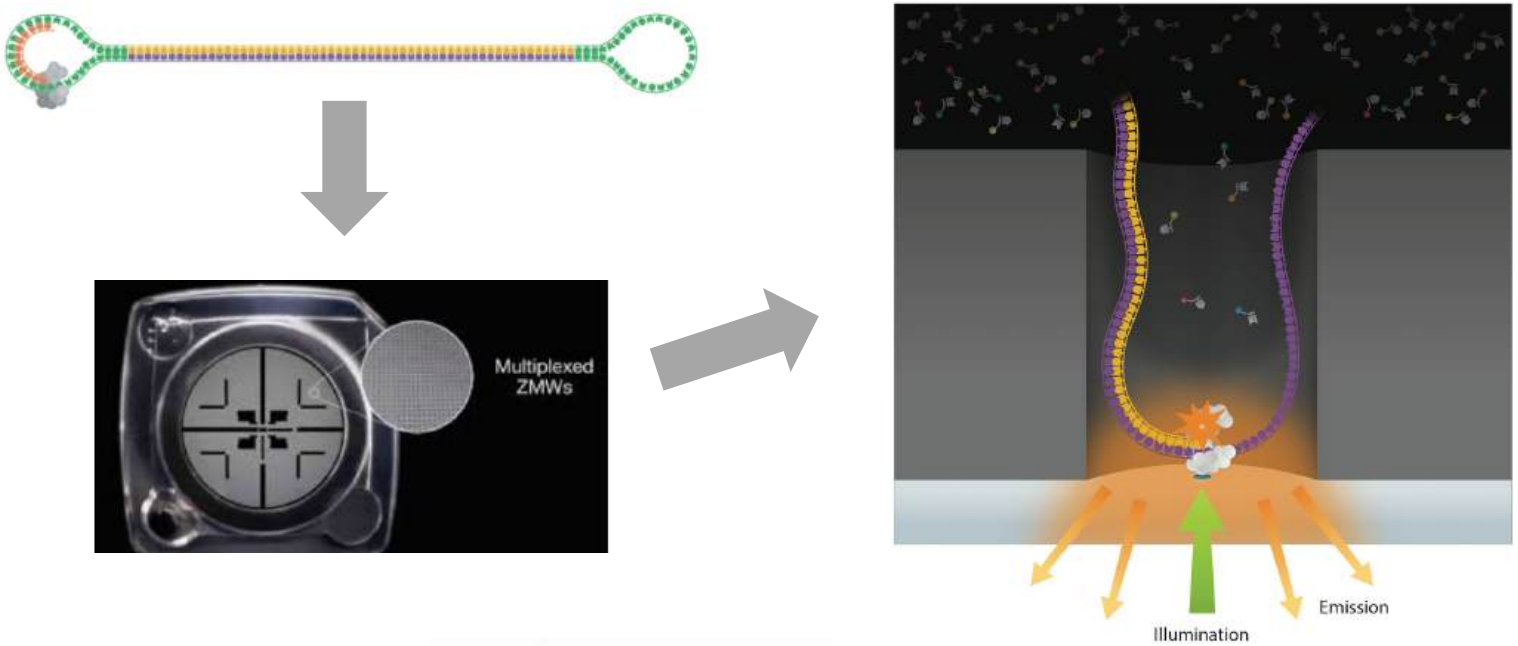


Long-read sequencing

No longer a niche technology!

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

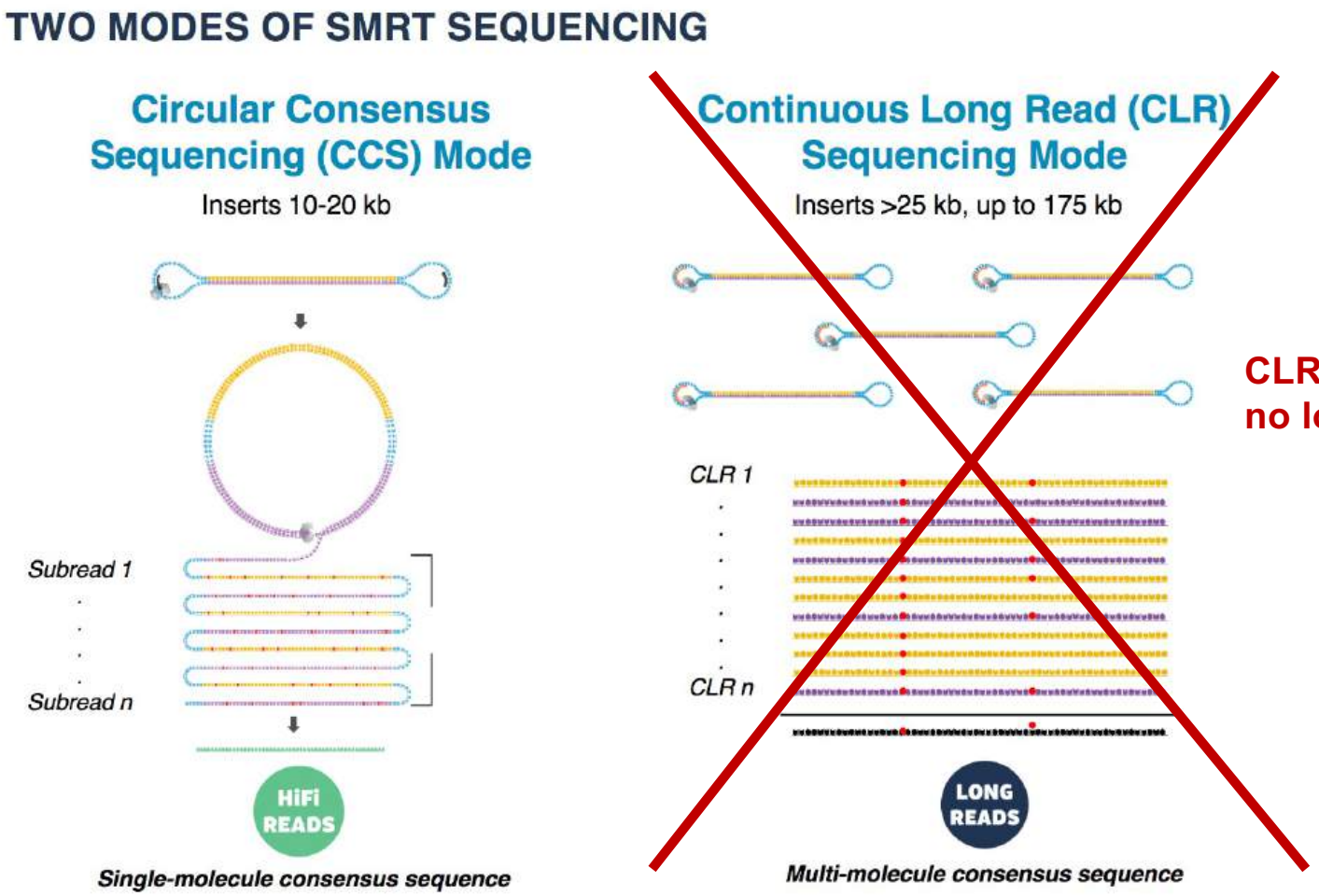
PacBio Sequencing



PacBio RSII

PacBio Sequel
(Sequel I & II)

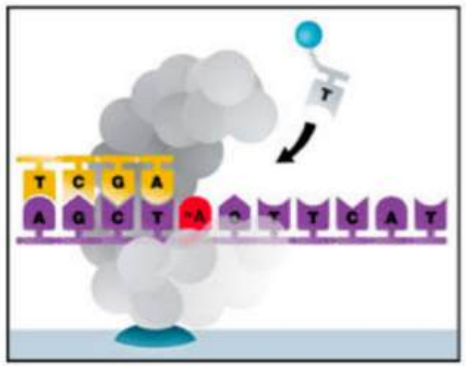
PacBio Sequencing

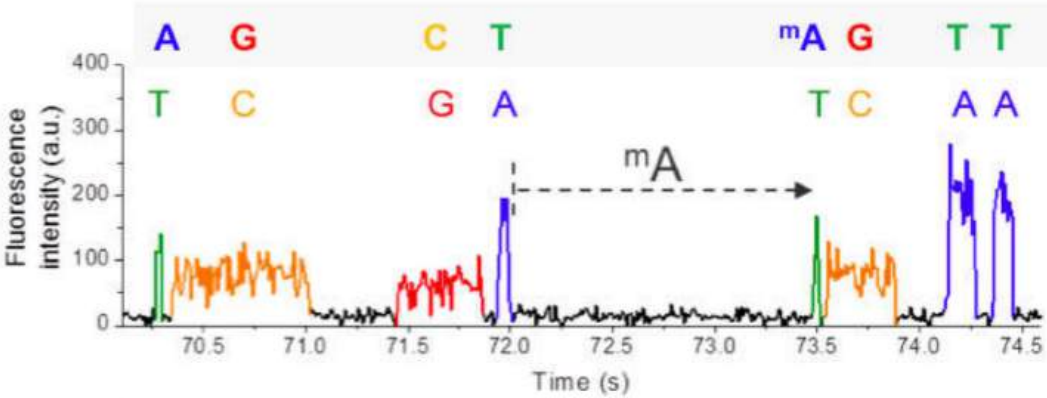


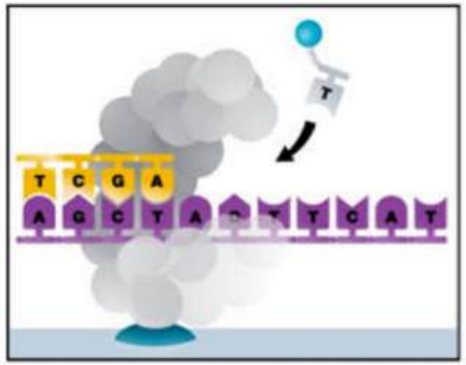
CLR sequencing no longer supported

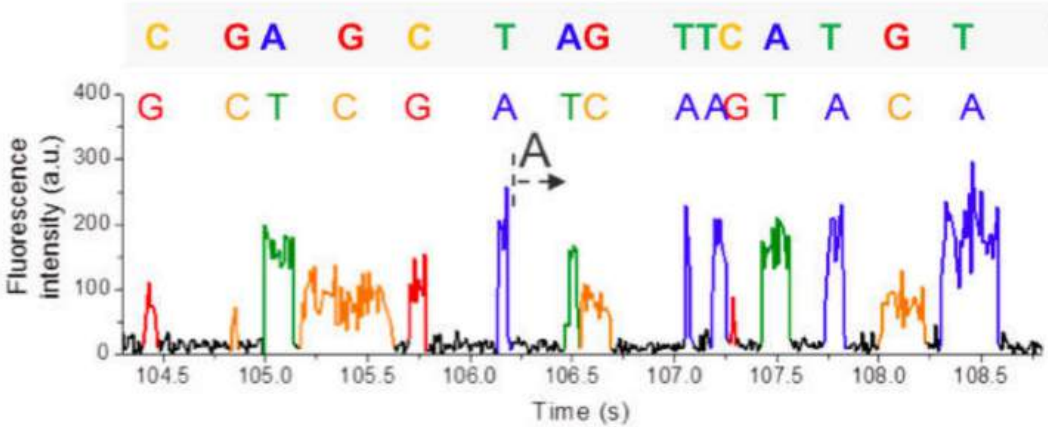
PacBio – Methylation detection

- 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!

The PacBio Revio System – Update 2025

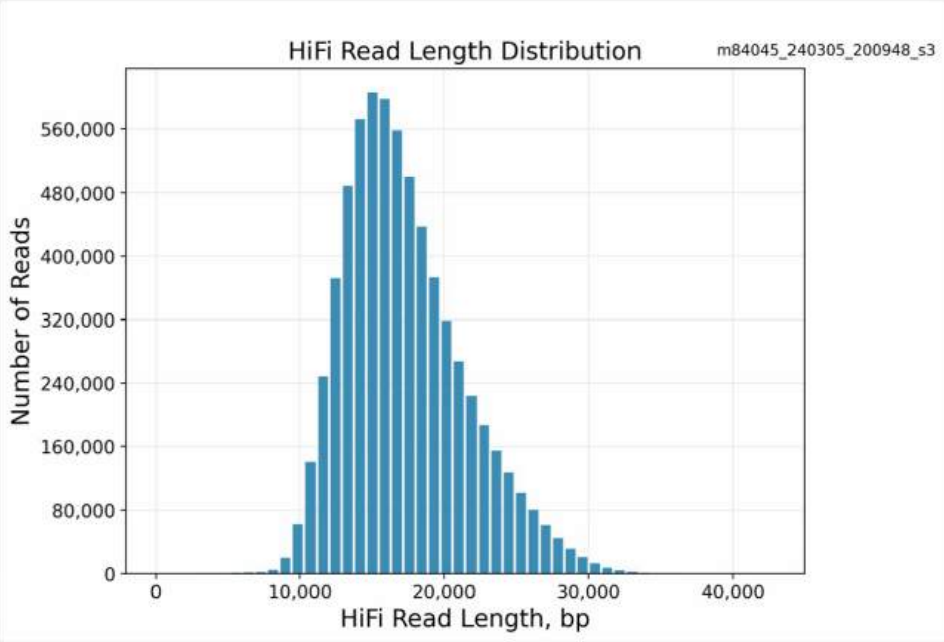
- Up to ~~90Gb~~ 120Gb data from one SMRT cell
- Read lengths: 15-20kb
- >QV20 quality (>99% read accuracy)
- Can run ~~1,300 human genomes~~ 2,500 human genomes/year!

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	

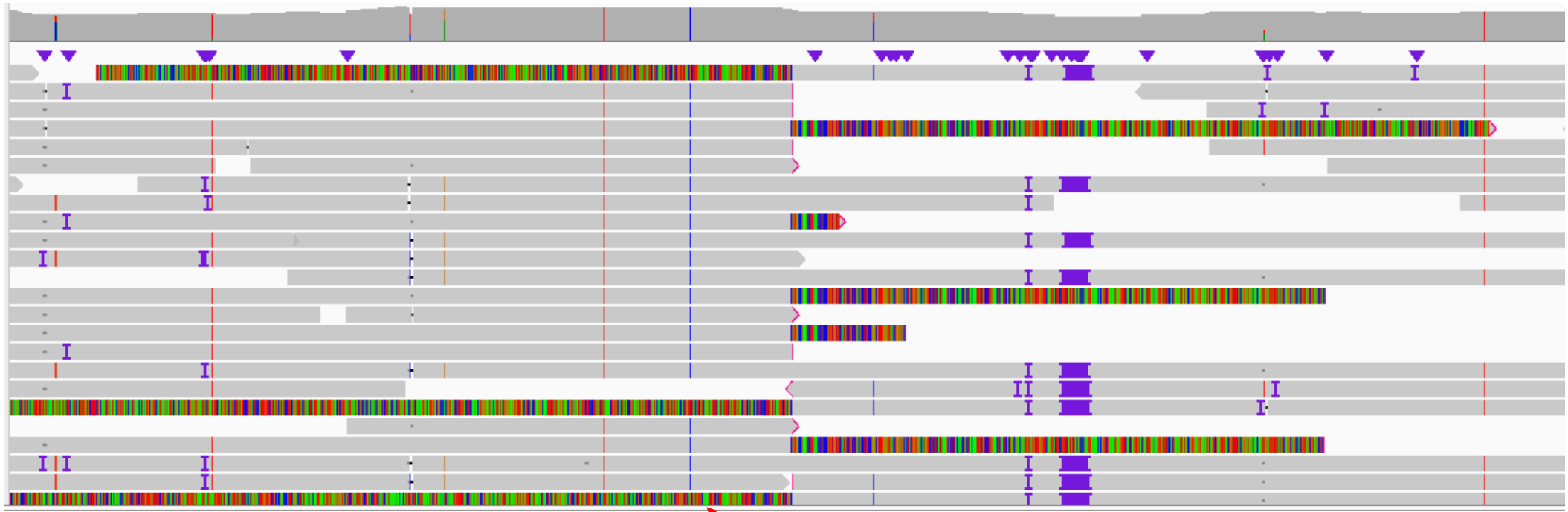
Example of a good run > 114 Gb

Value	Analysis Metric
6.6 M	HiFi reads
114.17 Gb	HiFi reads yield
17.21 kb	HiFi reads length (mean)
16,564	HiFi reads length (median, bp)
17,585	HiFi Read Length N50 (bp)
Q34	HiFi Read Quality (median)
92.36%	Base Quality ≥Q30 (%)
8	HiFi Number of Passes (mean)



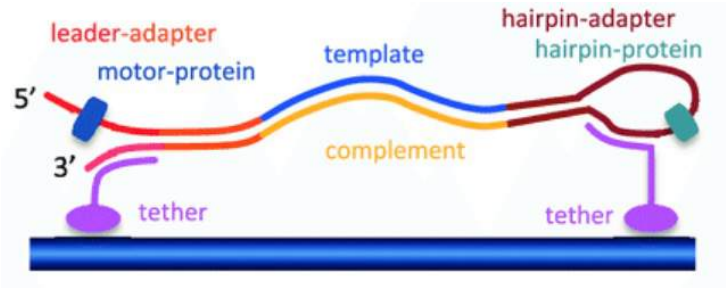
Example: Data at a transclocation site

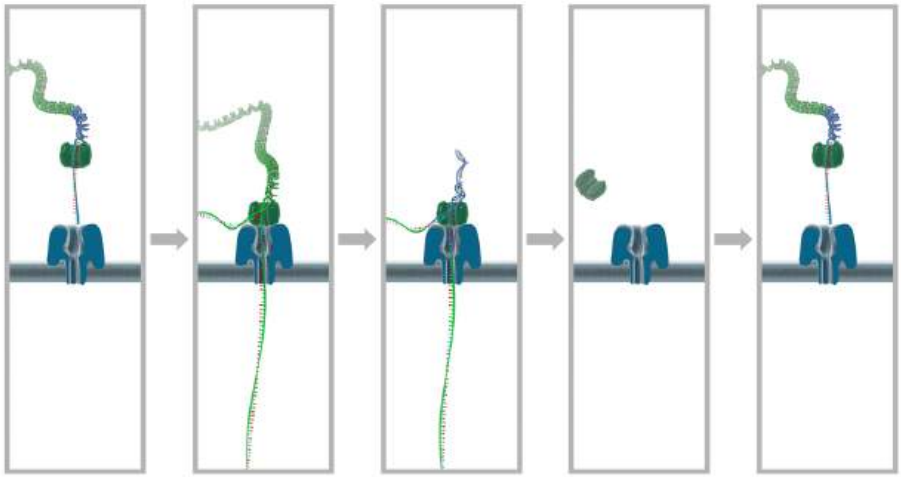
Translocation breakpoint

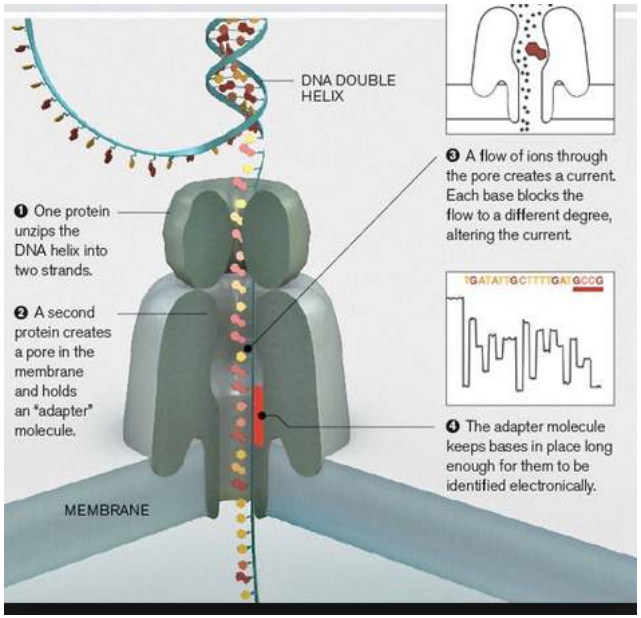


Soft clipped reads, aligning to another chromosome

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

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' v
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

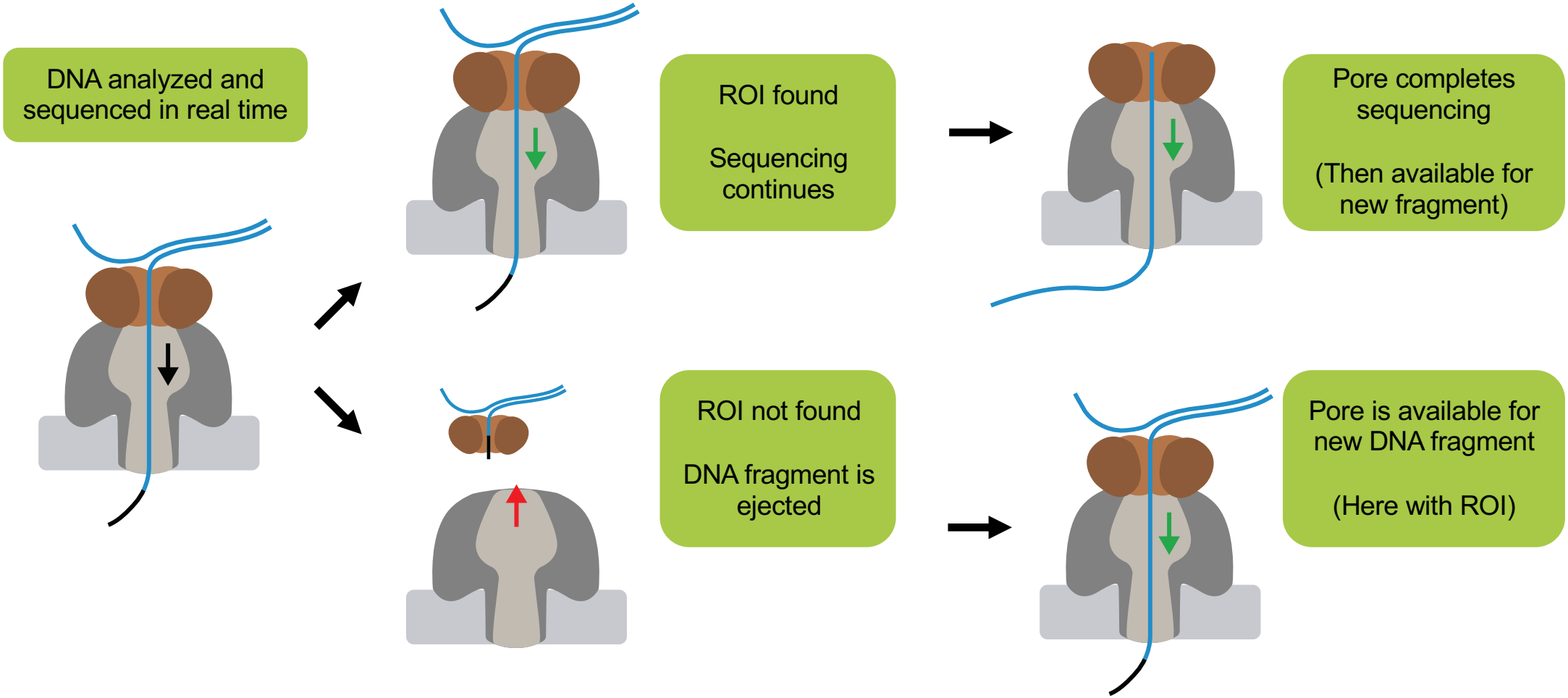
Ultra-fast deep-learned CNS tumour classification during surgery

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

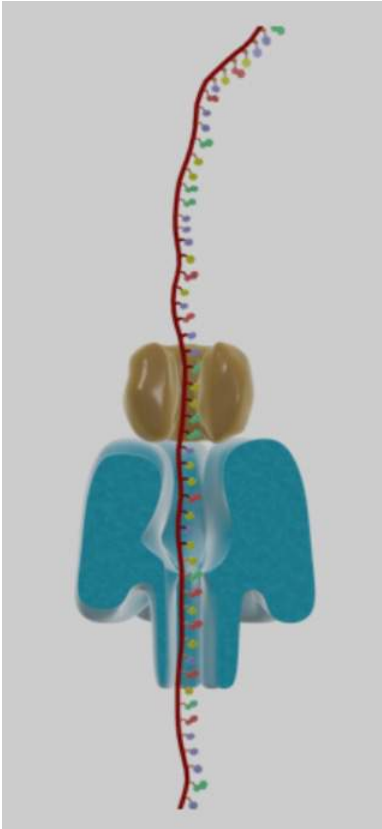
ONT target sequencing - adaptive sampling



ONT direct RNA sequencing

ONT can sequence native RNA molecules!

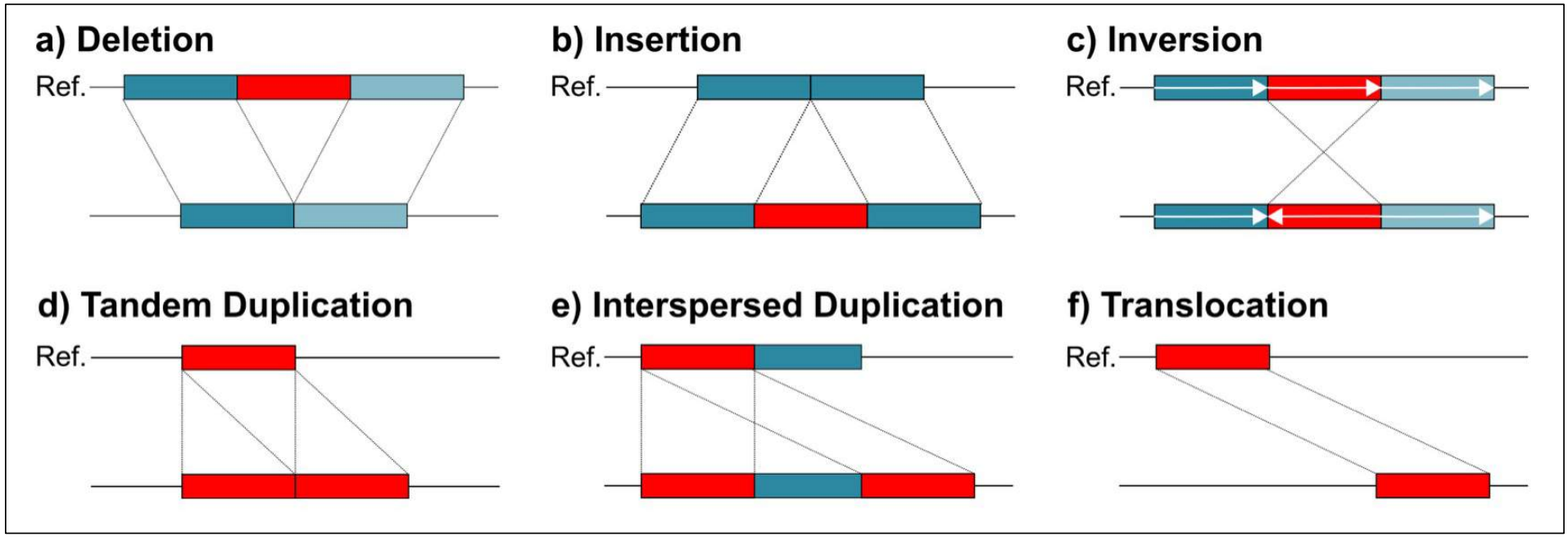
- No bias due to cDNA conversion
- Allows to study RNA modifications
- Higher error rate
- Lower throughput



What people are using long reads for…

Example 1: Detect all genetic variants

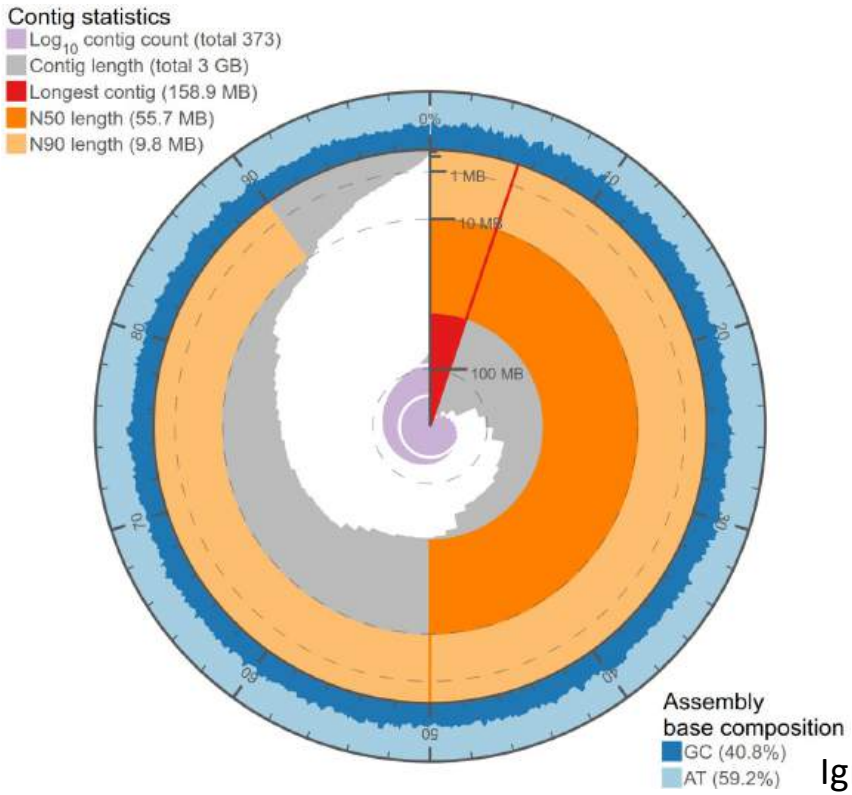
Long-read sequencing can detect more genetic variants than with short reads:



Example 2: Assemble complete genomes

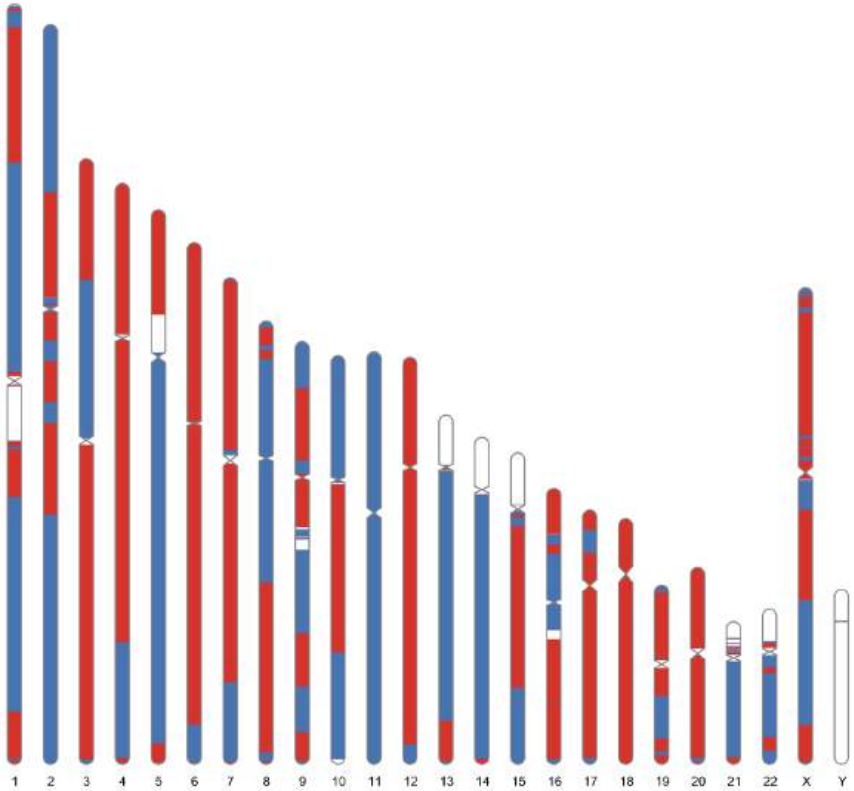
It took just **3.5 h** on a **96** core compute node for *de novo* assembly of a human sample!

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



Ignas Bunikis

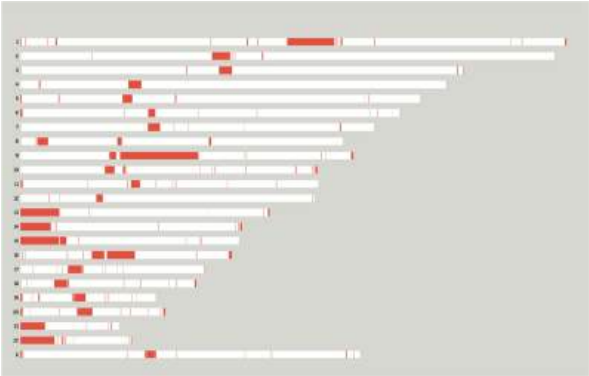
De novo assembly mapped to GRCh38



Colour change represents adjacent contigs

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

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

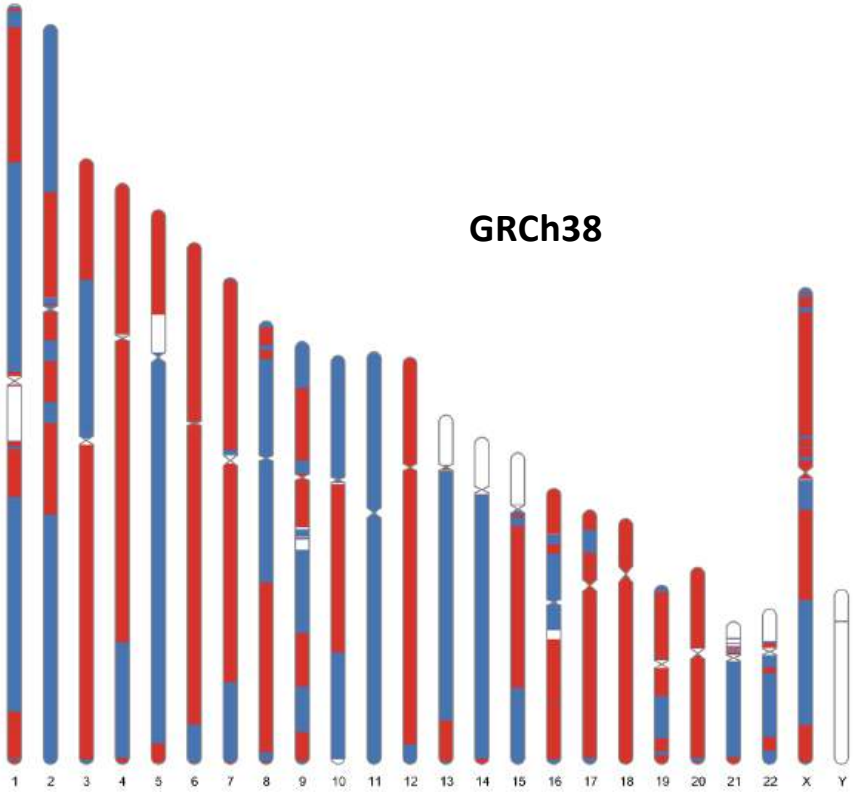


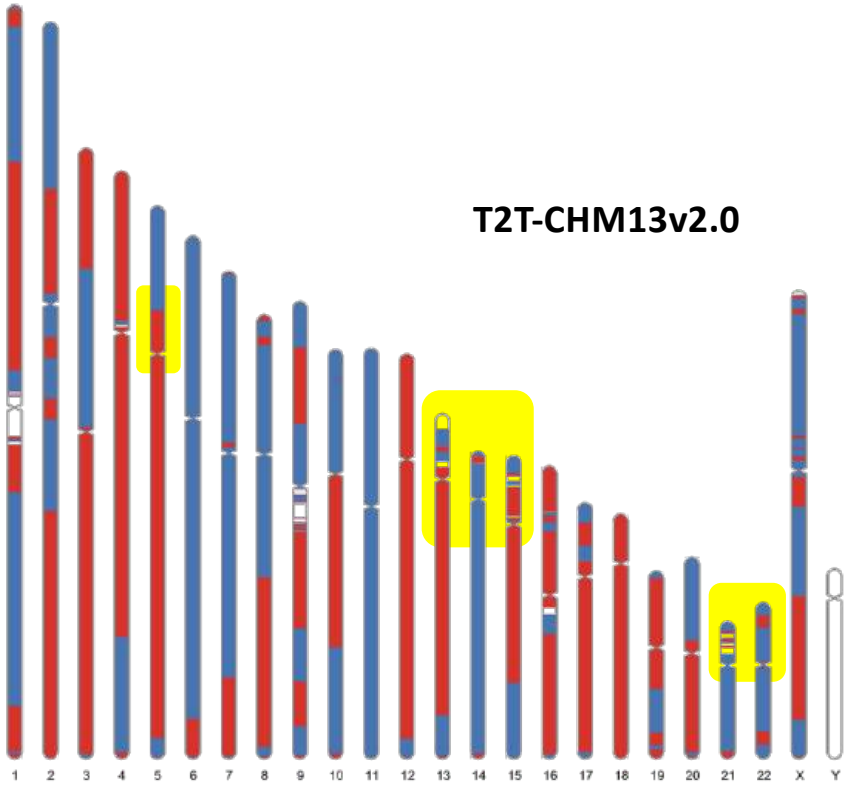
Red segments resolved by T2T Consortium

DOI: 10.1126/science.abp8653

Ignas Bunikis

De novo assembly mapped to T2T



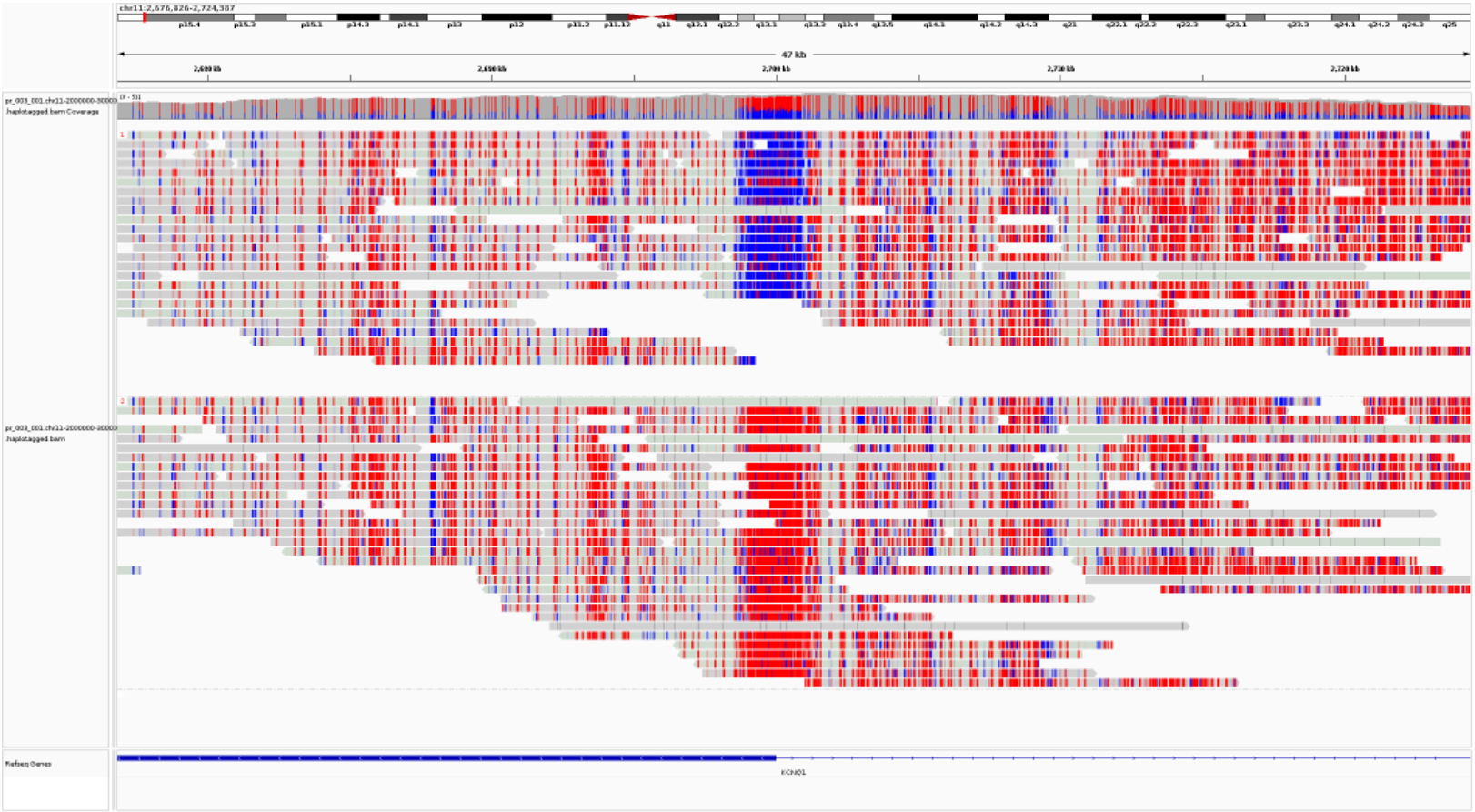


Colour change represents adjacent contigs

Ignas Bunikis

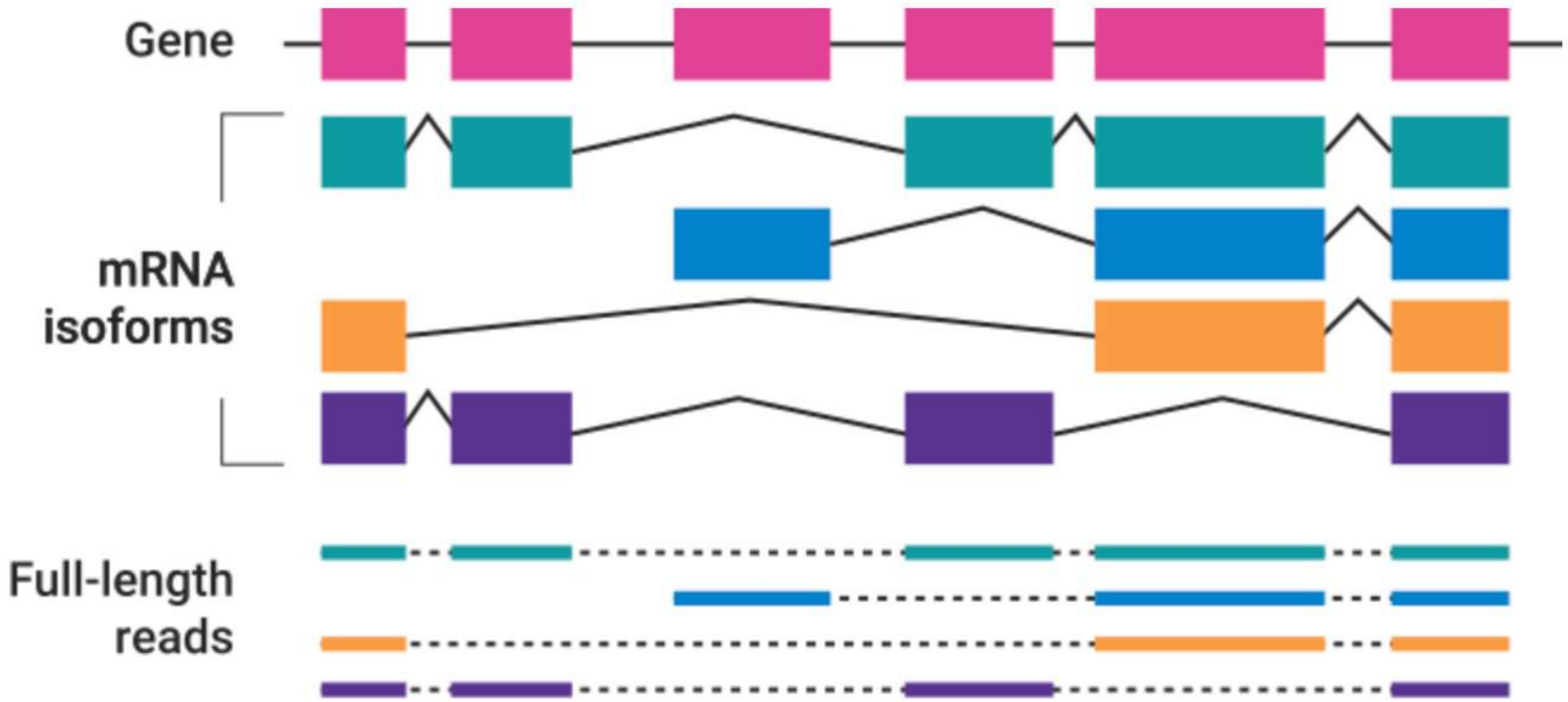
Example 3: Investigate methylation

Obtain methylation patterns, phased with haplotypes (example for imprinted region)



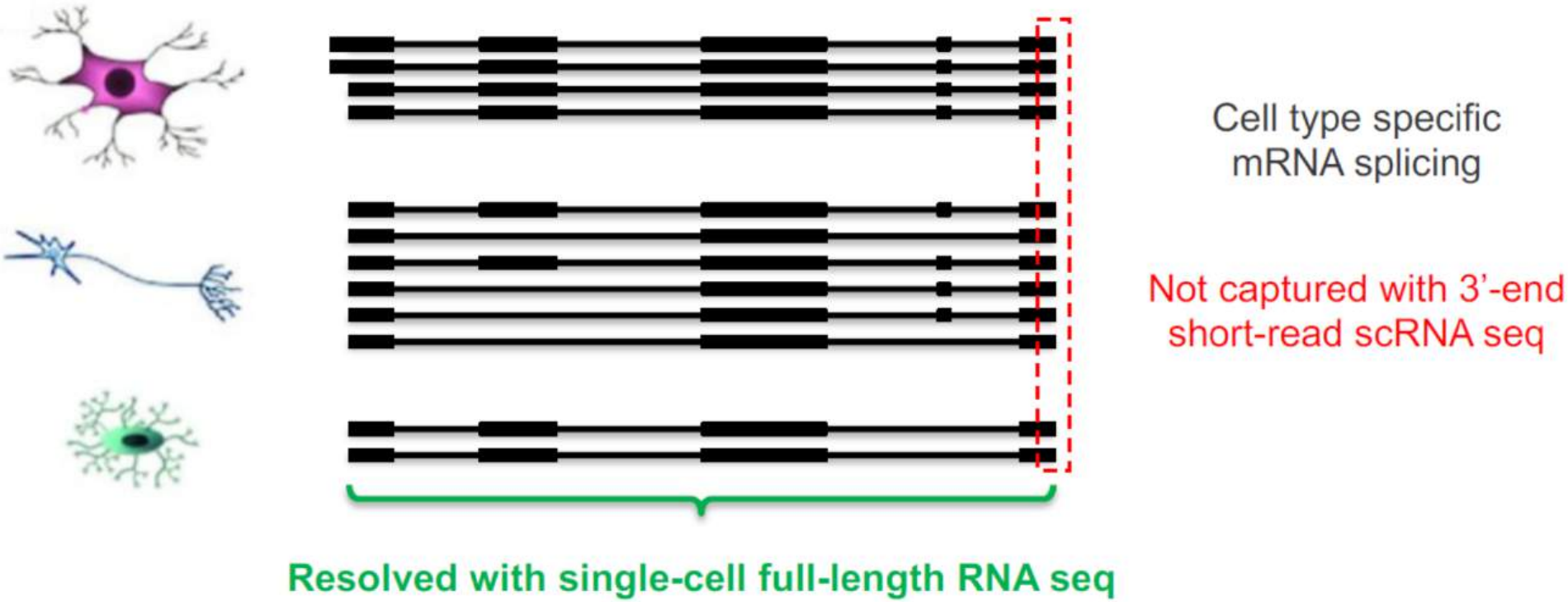
Example 4: Full-length RNA sequencing

- Get complete information about RNA molecules!



Example 5: Single-cell long-read RNA

- It is possible to study RNA isoforms even in single cells!

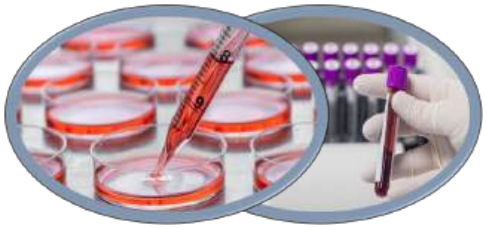


Challenge: good sample quality required!

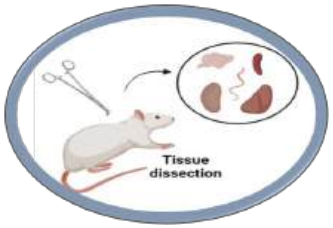
https://www.qiagen.com/ja-us/applications/molecular-biology-research/hmw-dna

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

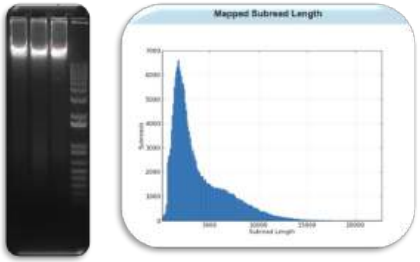
Top choice for plants, fungi, and other samples high in polysaccharides

HMW-DNA Extraction – Contaminants

Importance of purity – even for model organisms

Impurities can originate from both host tissue and extraction chemicals.

Same yeast - different extractions!



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

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?

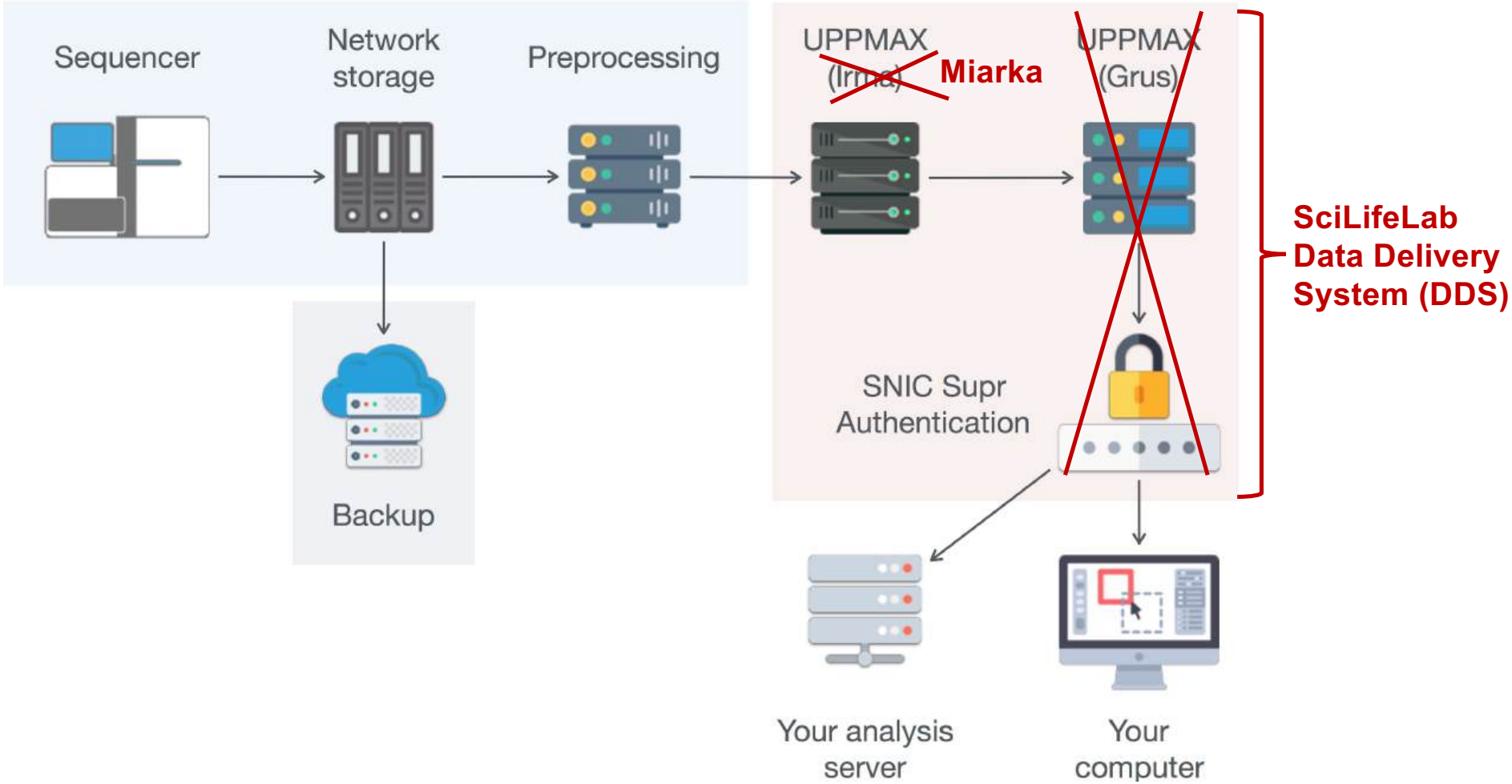
Which would you expect to have less contaminants?

VS

VS

NGI Data Handling and Analysis Pipelines

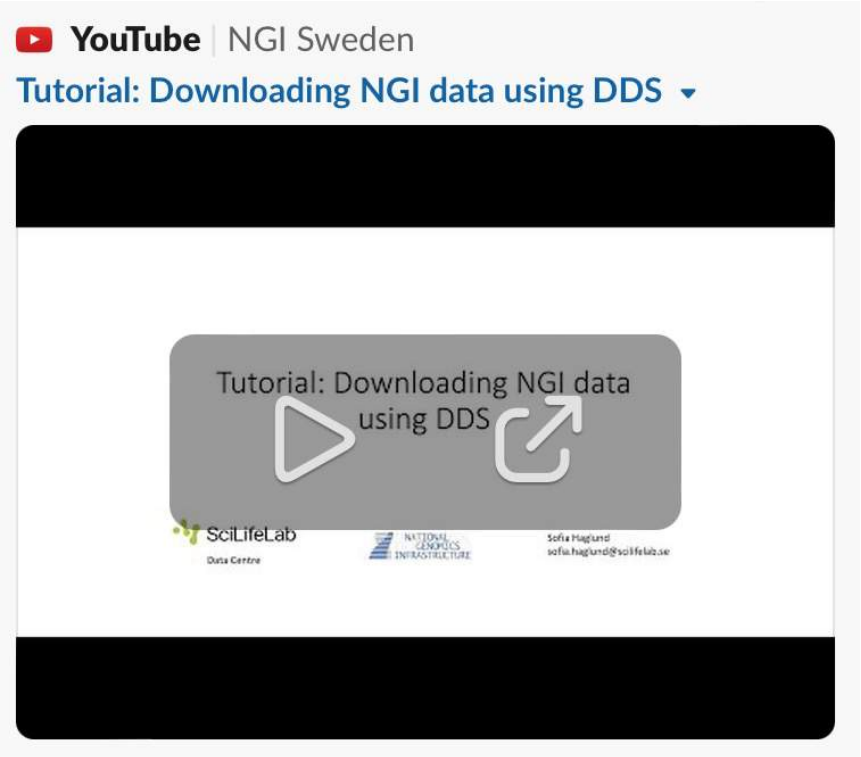
NGI Data Handling



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

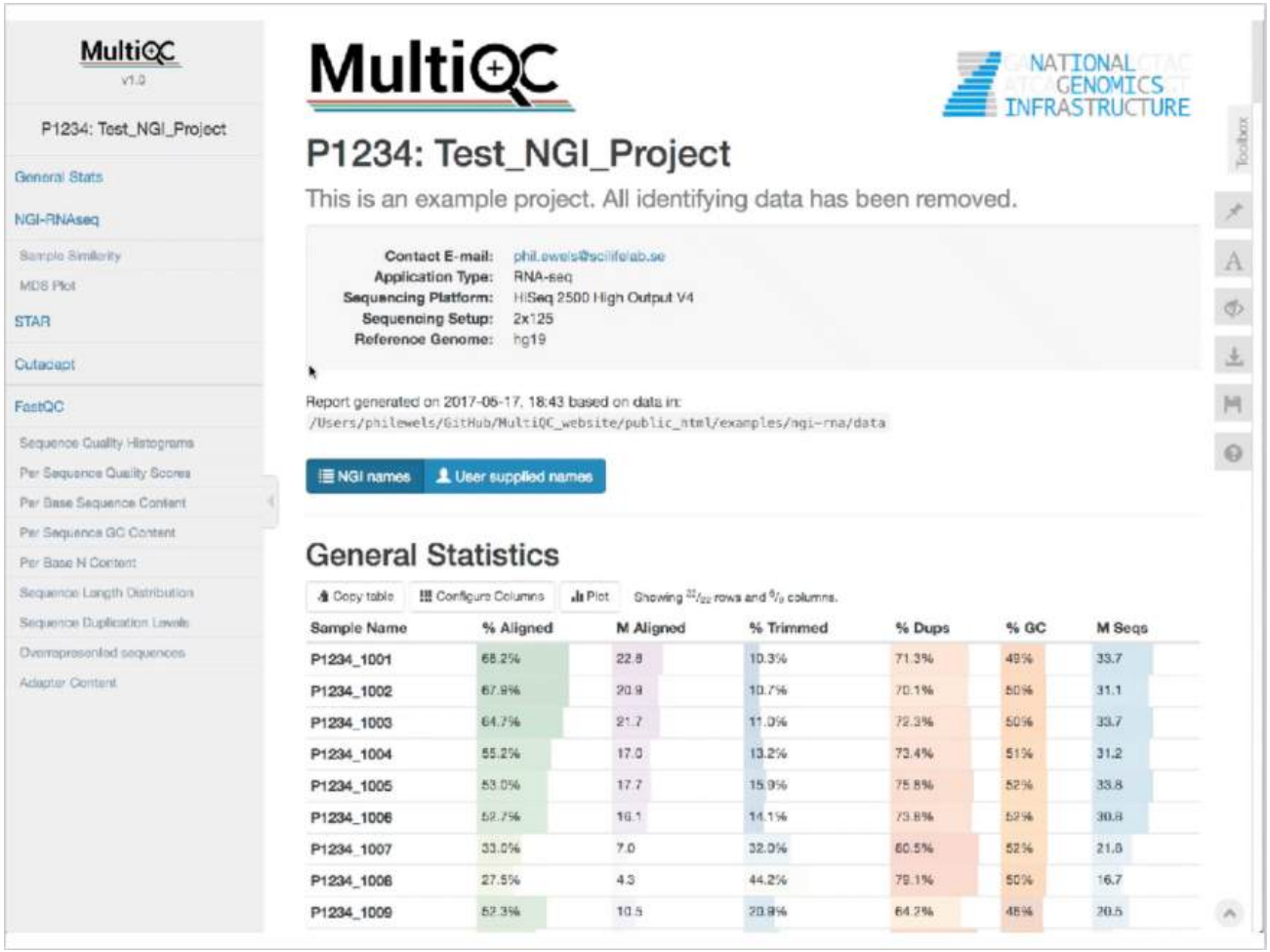
- Instruction video available on Youtube!



Quality control

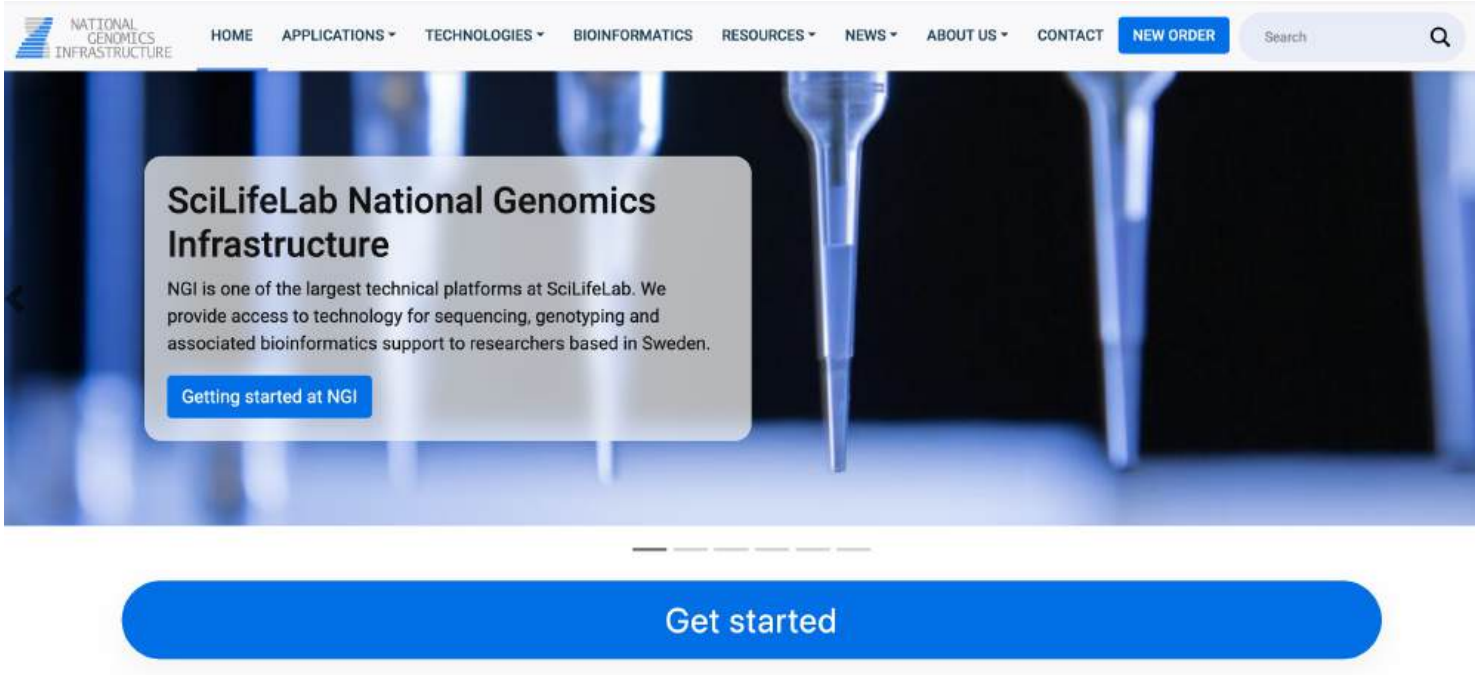
- Every project has some level of quality control checks
  - Technical run performance
  - Read length distribution
  - Sequencing quality
- Analysis pipelines give application-specific QC
- Reporting done using MultiQC (Illumina projects)

Multi QC example



Analysis pipelines

- NGI provides data analysis for most applications
- 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

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

nf-core

https://nf-co.re

Available pipelines at NGI

- All information available on our website: https://ngisweden.scilifelab.se

WES and WGS analysis

WGS and WES germline / somatic analysis

> Runs with illumina DNA-sequencing data, WGS or targeted sequencing e.g. WES. Aligns to the reference genome, gives QC metrics, does variant-calling and finishes with annotation.

nf-core/sarek (paper) is an analysis pipeline for WGS and targeted sequencing data e.g WES. Previously known as the Cancer Analysis Workflow (CAW), Sarek can handle regular samples or tumour/normal pairs, including relapse samples if required. Sarek was co-developed by NGI.

Sarek analysis can be divided into two different use cases: germline analysis and somatic analysis. These two use cases share the same main steps: mapping, variant calling and annotation.

nf-core/sarek

https://github.com/nf-core/sarek

Analysis pipeline to detect germline or somatic variants (pre-processing, variant calling and annotation) from WGS / targeted sequencing - https://nf-co.re/sarek

324 | 324 | 354

When we run analysis

We routinely run Sarek germline analysis upon request for human WGS and WES projects. For the Sarek somatic analysis, the decision to run the analysis is made on a case by case basis. If you're interested, please get in touch with us and mention that you would like us to run this analysis.

The analysis currently works with the human reference genomes available in AWS-iGenomes (GRCh37/GRCh38). If in doubt, please ask whether we can run the pipeline for you.

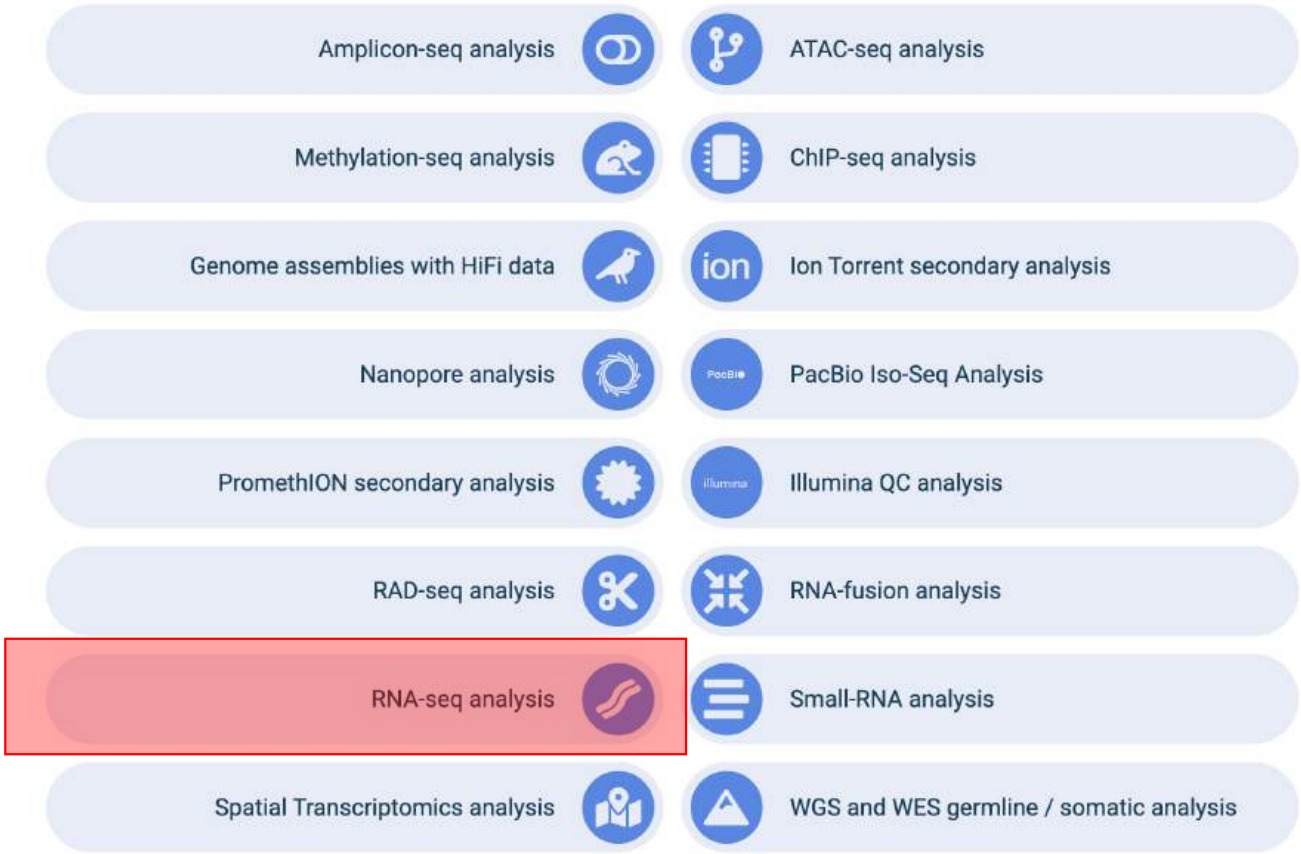
Input data

Sarek can start from the unprocessed demultiplexed FastQ files from the sequencer together with a small bit of contextual data in the form of a TSV-file. For each sample, the TSV-file should denote the sex of the subject and whether the sample is tumour or normal. In most cases, this information needs to be submitted to NGI by the user.

Results

The pipeline generates BAM alignment files and variant-calling VCF files, along with numerous quality control metrics. For more information, please see the official documentation.

Available pipelines at NGI



Example: RNA-seq analysis

RNA-seq analysis

> Runs with illumina total RNA-sequencing data. Aligns to the reference genome, gives QC metrics and finishes with gene count matrices.

RNA-Seq is a bioinformatics analysis pipeline used for RNA sequencing data. The pipeline is built using Nextflow, a workflow tool to run tasks across multiple compute infrastructures in a very portable manner. It processes raw data from FastQ inputs, aligns the reads, generates counts relative to genes or transcripts and performs extensive quality control on the results.

nf-core/rnaseq

https://github.com/nf-core/rnaseq

RNA sequencing analysis pipeline using STAR, RSEM, HISAT2 or Salmon with gene/isoform counts and extensive quality control. - https://nf-co.re/rnaseq

746 746 643

When we run analysis

We run this analysis routinely for all RNA-seq projects where we have prepared the sequencing library in-house. If you have prepared a library yourself and we are just sequencing, please get in touch and mention that you would like us to run this analysis.

The analysis works with any of the species that have a reference genome available in AWS-iGenomes. If in doubt, please ask whether we can run the pipeline for you.

Input data

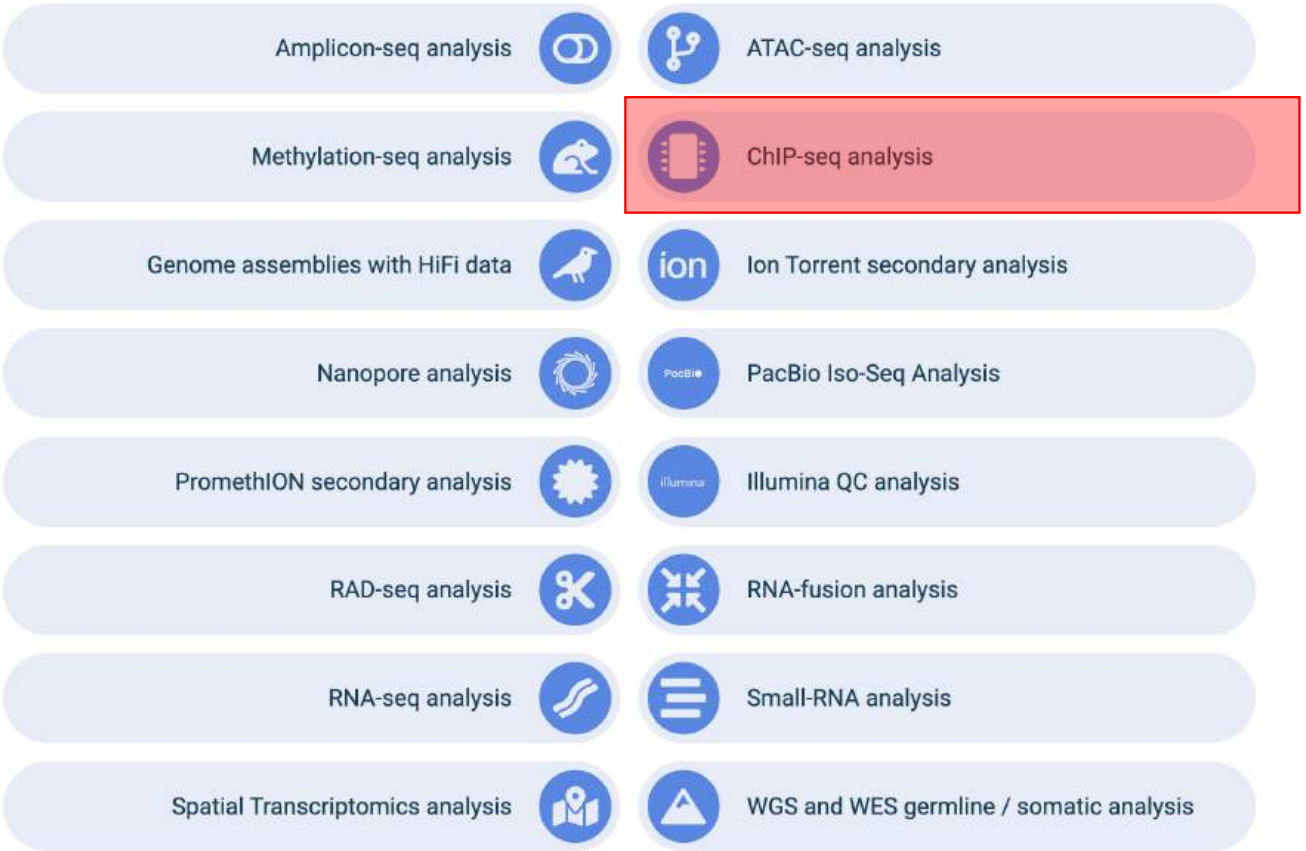
bcl2fastq demultiplexed FastQ files and a genome reference.

Results

The pipeline generates aligned BAM-files, gene count matrices and FPKM metrics for genes and transcripts, along with numerous quality control metrics. For more information, please see https://nf-co.re/rnaseq/[release]/docs/output

Last Updated: 18th October 2023

Available pipelines at NGI



ChIP-seq analysis

ChIP-seq analysis

Runs with ChIP sequencing data. Pre-processes raw data from FastQ inputs, aligns the reads and performs peak calling and extensive quality-control on the results.

ChIP-Seq is a bioinformatics best-practice analysis pipeline used for chromatin immunoprecipitation (ChIP-seq) data analysis. The pipeline uses Nextflow, a bioinformatics workflow tool. It pre-processes raw data from FastQ inputs, aligns the reads and performs peak calling and extensive quality-control on the results.

nf-core/chipseq

https://github.com/nf-core/chipseq

ChIP-seq peak-calling, QC and differential analysis pipeline. - https://nf-co.re/chipseq

161 161 130

When we run analysis

We run this analysis routinely for all ChIP-seq projects where we have prepared the sequencing library in-house. If you have prepared a library yourself and we are just sequencing, please get in touch and mention that you would like us to run this analysis.

The analysis works with any of the species that have a reference genome available in AWS-iGenomes. If in doubt, please ask whether we can run the pipeline for you.

Input data

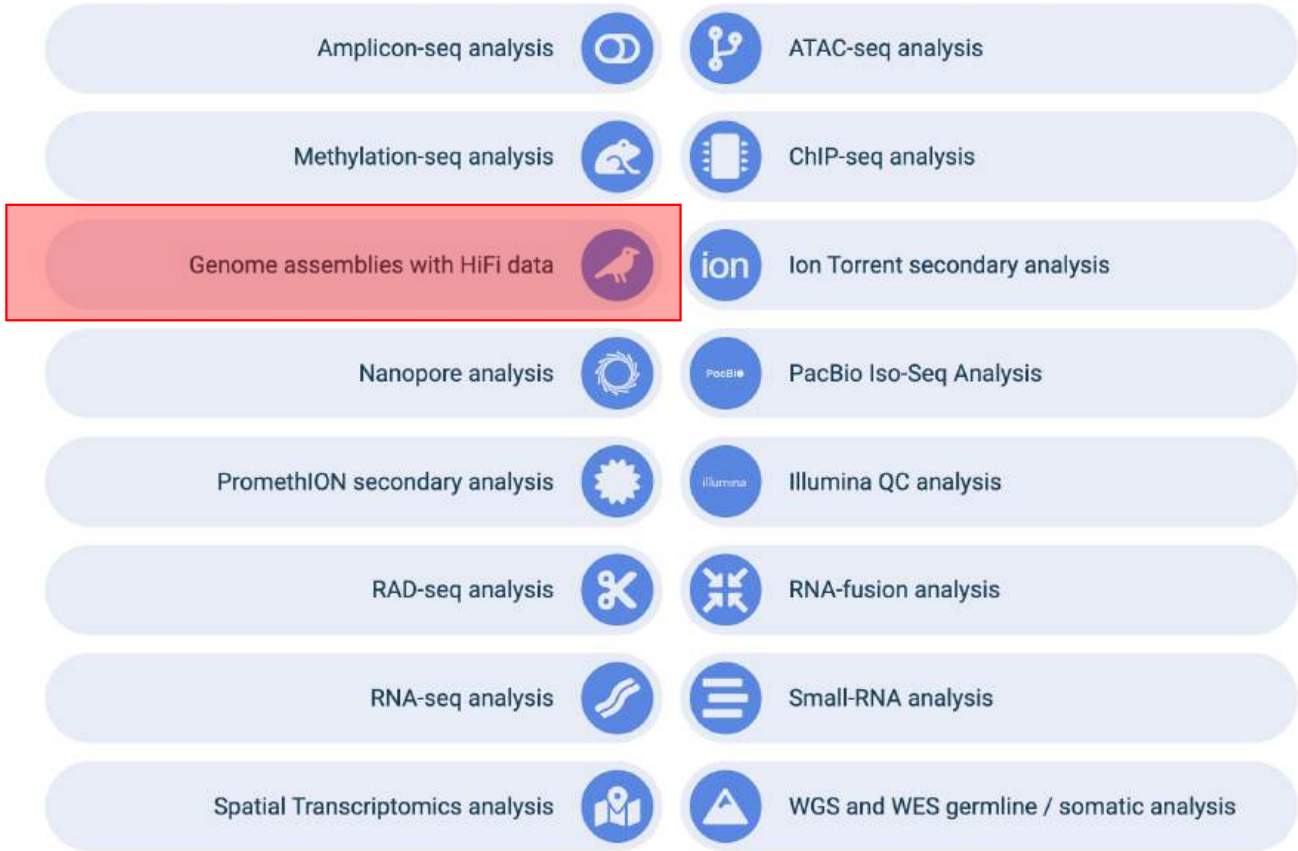
bcl2fastq demultiplexed FastQ files and a genome reference.

Results

The pipeline generates aligned BAM-files, files with information about called peaks, along with numerous quality control metrics. For more information, please see https://nf-co.re/chipseq/docs/output.

Last Updated: 14th July 2023

Available pipelines at NGI



Genome assembly with HiFi data

Genome assemblies with HiFi data

NGI can generate high quality assemblies using IPA and hifiasm assemblers

hifiasm & IPA

Improved Phased Assembler (IPA) is the official PacBio software for **HiFi** genome assembly. IPA was designed to utilize the accuracy of PacBio HiFi reads to produce high-quality phased genome assemblies.

Hifiasm is a fast haplotype-resolved *de novo* assembler for PacBio HiFi reads. It emits partially phased assemblies of quality competitive with the best assemblers. Given parental short reads or Hi-C data, it produces arguably the best haplotype-resolved assemblies so far.

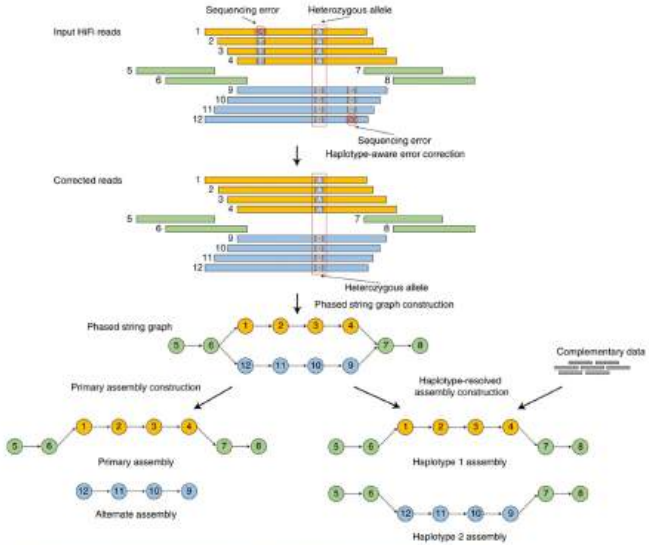


Image: Nat Methods **18**, 170–175 (2021). https://doi.org/10.1038/s41592-020-01056-5

Not yet implemented as a nf-core pipeline!

Trend: On-instrument analysis

More and more analyses being done on instrument GPUs

Illumina NovaSeqX

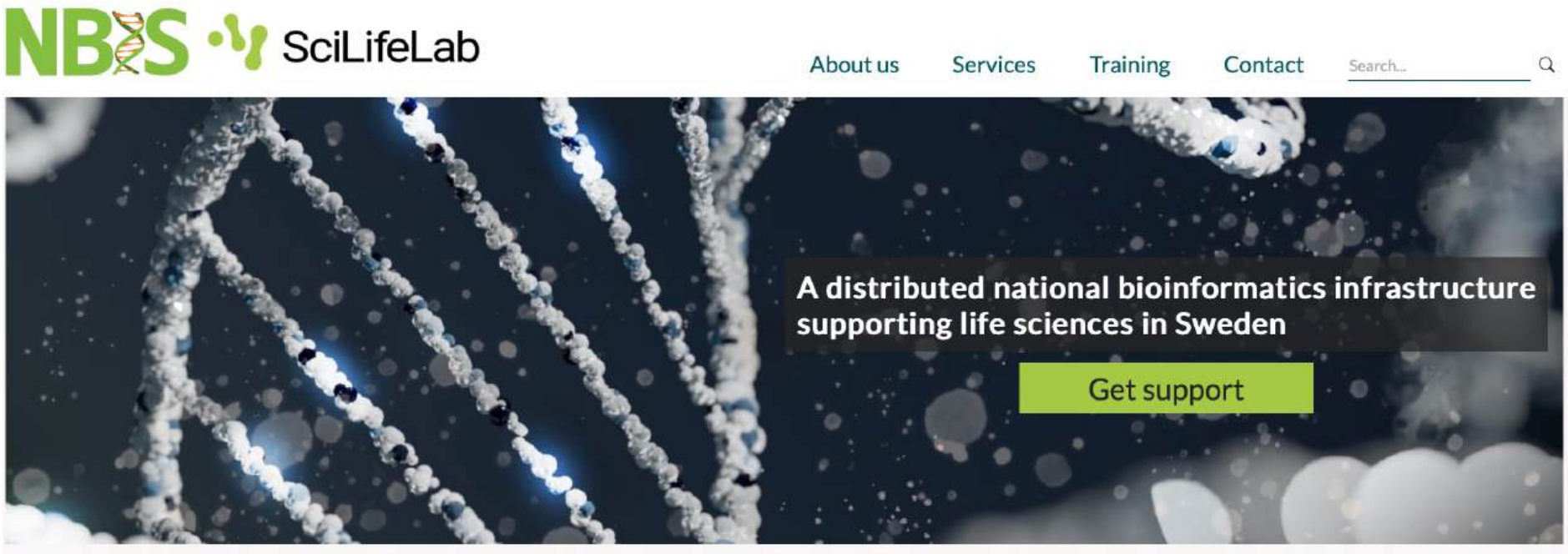
Mapping and variant calling (Dragen)

PacBio Revio

Onboard generation of HiFi reads

➔ Can speed up and streamline the analysis process…

You can also get help from NBIS!



- All solutions are not available from NGI, but NBIS has lots of experts!

Some tips for data analysis…

Think about analysis early on – already when planning the project!

- Which tools should be used?
- Can I run the analysis myself, or do I need assistance?
- Where should the analysis be run?
- Do I have enough storage space?
- Where should the data eventually be archived?

NGI strategic projects and collaborations

We are involved in some larger national and international projects...

Biodiversity genomics

Reference genomes of **any** organism - a very challenging endeavour

Large genomes (18-22Gb)

Tiny organisms

Tiny organisms
with large genomes

And many, many, ..., many other uncooperative organisms

Reference genome sequencing

NGI & NBIS can help out with:

- DNA/RNA extractions
- Long-read sequencing
- Hi-C Illumina sequencing
- RNA sequencing
- De novo assembly
- Genome annotation

https://www.vecteezy.com

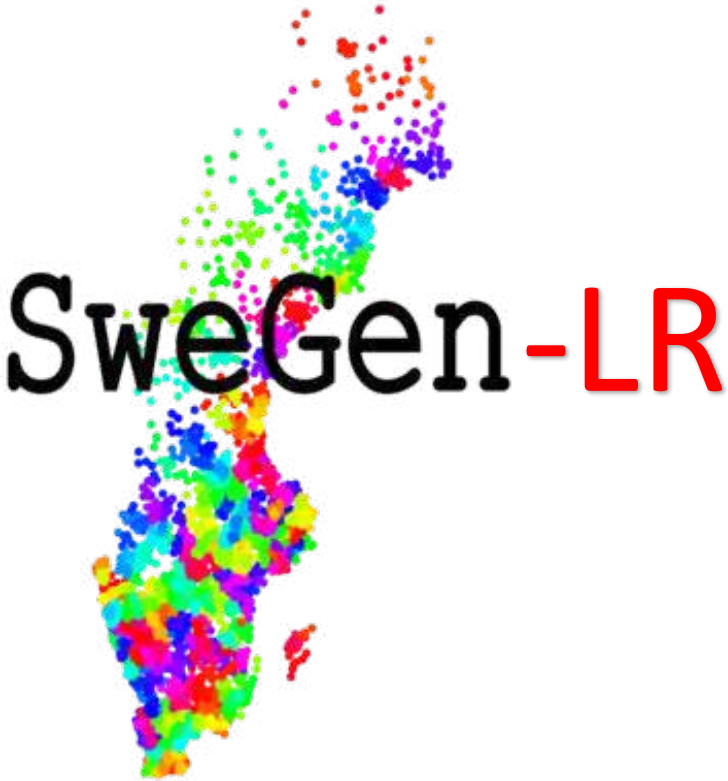
Human genome analysis

Photo: SVT

Swedish WGS reference dataset

2017: 1,000 genomes sequenced on Illumina

2024: Time to do it again, with long reads!



Ameur et al, Eur. J. Hum. Genet. 2017

How to build a long-read reference dataset?

- Planning started 2023, with a wishlist for a new Swedish population cohort

	Description	Priority
Consent for data sharing	It must be possible to share individual-level variant information (VCF files) on national level and ideally also internationally	Crucial
Amount and quality of DNA	At least 5ug of high-quality DNA per individual, ideally from fresh samples extracted for long-read sequencing	Crucial
Phenotype information	Detailed phenotype information available, that can be used for specific research projects (after approval)	Important
A cross-section of Sweden	The individuals should not be enriched for a specific disease or phenotype, and reflect the genetic variation in Sweden (ideally including ethnic minorities)	Important
Additional OMICS data	Possibility to perform other OMICs studies (RNA, protein, etc) on samples from the same individuals	Important
Available SNP array data	Data from SNP arrays, that can be used to infer the genetic background and select representative individuals for sequencing	Beneficial
Funding and resources	Possibility to get additional local funding and resources (for re-consent, sample collection, DNA extraction, etc.)	Beneficial

How to get funding for sequencing?

“Genome of Europe” - A 40M Euro project within the 1+Million Genomes Initiative!

GENOME OF EUROPE

The Genome of Europe

Aim: Construct a european reference WGS dataset

Sweden's representatives - Adam Ameur, Anna Lindstrand, Bengt Persson, Anna Hagwall

- 100,000 individuals, from 27 countries
- Representative of Europe's population
- International data sharing possible
- Swedens contribution: <u>2,600 WGS</u>

#participants in Genome of Europe	AUSTRIA	BELGIUM	BULGARIA	CROATIA	REPUBLIC OF CYPRUS	CZECH REPUBLIC	DENMARK	ESTONIA	FINLAND	FRANCE	GERMANY	GREECE	HUNGARY	IRELAND	ITALY	LATVIA	LITHUANIA	LUXEMBOURG	MALTA	NETHERLANDS	POLAND	PORTUGAL	ROMANIA	SLOVENIA	SPAIN	SWEDEN		NORWAY	SERBIA	TURKEY
Ancestry	2,113	2,903	1,853	1,257	310	2,946	1,414	368	1,462	17,041	20,235	2,924	2,660	1,262	15,762	476	816	200	200	4,510	50	2,904	0	491	12,125	2,591	100,222	1,346	0	0
Albania	0	0	0	0	0	0	0	0	0	0	50	100	0	0	100	0	0	0	0	0	0	0	0	0	0	0	250	0	0	0
Algeria	0	0	0	0	0	0	0	0	0	224	10	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	244	0	0	0
Austria	1,893	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1,893	0	0	0
Belgium	0	2,783	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2,783	0	0	0
Bosnia and Herz	50	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	10	0	0	0	0	0	50	210	0	0	0
Bulgaria	0	0	1,653	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1,653	0	0	0
China	0	0	0	0	0	0	0	0	0	50	50	0	0	0	100	0	0	0	0	50	0	0	0	0	50	10	310	0	0	0
Croatia	0	0	0	1,207	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1,207	0	0	0
Cyprus	0	0	0	0	200	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	200	0	0	0
Czechia	0	0	0	0	0	2,896	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2,896	0	0	0
Denmark	0	0	0	0	0	0	1,384	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1,384	0	0	0
Estonia	0	0	0	0	0	0	0	268	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	268	0	0	0
Finland	0	0	0	0	0	0	0	0	1,462	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1,462	0	0	0
France	0	0	0	0	0	0	0	0	0	16,298	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16,298	0	0	0
Germany	0	0	0	0	0	0	0	0	0	0	17,266	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	17,266	0	0	0
Greece	0	0	0	0	0	0	0	0	0	0	0	2,764	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2,764	0	0	0
Hungary	0	0	0	0	0	0	0	0	0	0	0	0	2,560	0	0	0	0	0	0	0	0	0	0	0	0	0	2,560	0	0	0
Ireland	0	0	0	0	0	0	0	0	0	0	0	0	0	1,162	0	0	0	0	0	0	0	0	0	0	0	0	1,162	0	0	0
Italy	0	0	0	0	0	0	0	0	0	0	0	0	0	0	15,062	0	0	0	0	0	0	0	0	0	0	0	15,062	0	0	0
Kazakhstan	0	0	0	0	0	0	0	0	0	0	312	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	312	0	0	0
Latvia	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	356	0	0	0	0	0	0	0	0	0	0	356	0	0	0
Lithuania	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	706	0	0	0	0	0	0	0	0	0	706	0	0	0
Luxembourg	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	200	0	0	0	0	0	0	0	0	200	0	0	0
Malta	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	200	0	0	0	0	0	0	0	200	0	0	0
Morocco	0	50	0	0	0	0	0	0	0	209	50	0	0	0	100	0	0	0	0	100	0	0	0	0	233	0	743	0	0	0
Netherlands	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4,090	0	0	0	0	0	0	4,090	0	0	0
Norway	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1,266	1,266	0	0
Poland	10	10	0	0	0	0	10	0	0	50	567	0	0	50	50	10	50	0	0	50	50	0	0	0	10	50	967	50	0	0
Portugal	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2,904	0	0	0	0	2,904	0	0	0
Roma (Gypsy)	0	0	100	0	0	0	0	0	0	0	0	50	100	0	0	0	0	0	0	0	0	0	0	0	0	0	250	0	0	0
Romania	50	50	0	0	50	0	0	0	0	50	262	0	0	0	291	0	0	0	0	10	0	0	0	0	100	10	873	0	0	0
Russia	0	0	0	0	10	0	0	100	0	10	337	10	0	0	10	100	50	0	0	0	0	0	0	0	50	0	677	0	0	0
Serbia	50	0	0	50	0	0	0	0	0	50	100	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	260	0	0	0
Slovenia	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	491	0	0	0	491	0	0	0
Spain	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11,522	0	11,522	0	0	0
Sweden	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2,371	2,371	0	0	0
Syria	10	0	0	0	0	0	10	0	0	0	281	0	0	0	0	0	0	0	0	50	0	0	0	0	0	50	411	10	0	0
Turkey	50	10	100	0	0	0	10	0	0	50	750	0	0	0	0	0	0	0	0	100	0	0	0	0	0	50	1,120	0	0	0
Ukraine	0	0	0	0	0	50	0	0	0	0	100	0	0	0	50	10	10	0	0	0	0	0	0	0	50	0	270	0	0	0
United Kingdom	0	0	0	0	50	0	0	0	0	50	50	0	0	50	0	0	0	0	0	50	0	0	0	0	100	0	350	0	0	0

What is the best sample collection for GoE?

- SCAPIS: A prospective population study for heart- and lung disease
- Over 30,000 participants, collected at 6 sites (from Lund to Umeå)
- We are planning to analyze at least 1000 individuals

Collaborations on Rare Disease

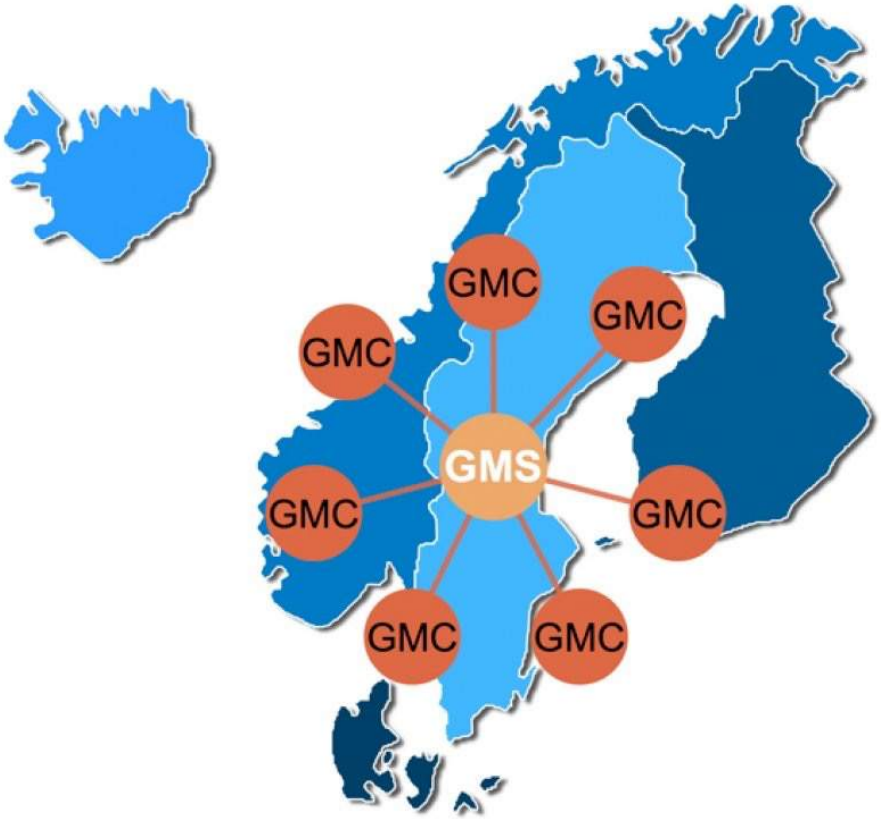
We are collaborating with Genomic Medicine Sweden - Rare Disease Group

Long-Read Whole Genome Sequencing

- Improve diagnostics of rare disease patients
- Resolve complex SVs and other variants
- PacBio Revio and ONT PromethION

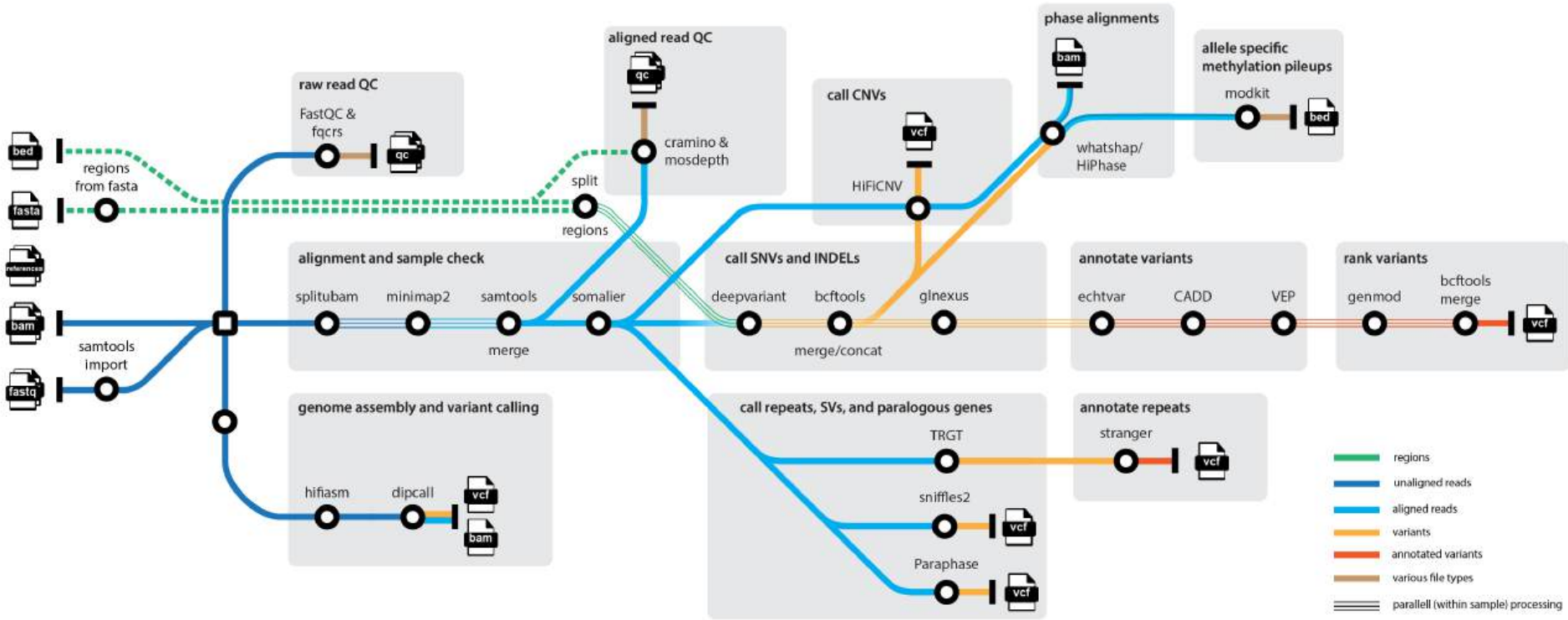
Long-Read Targeted Sequencing

- Develop clinical assays for repeat expansions
- Cas9-based capture or adaptive sampling
- Aim: implementation at different hospital nodes



How to analyze human long-read data?

Nallo: a Nextflow analysis pipeline for patients and controls



Felix Lenner, et al

Thanks for your attention!