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

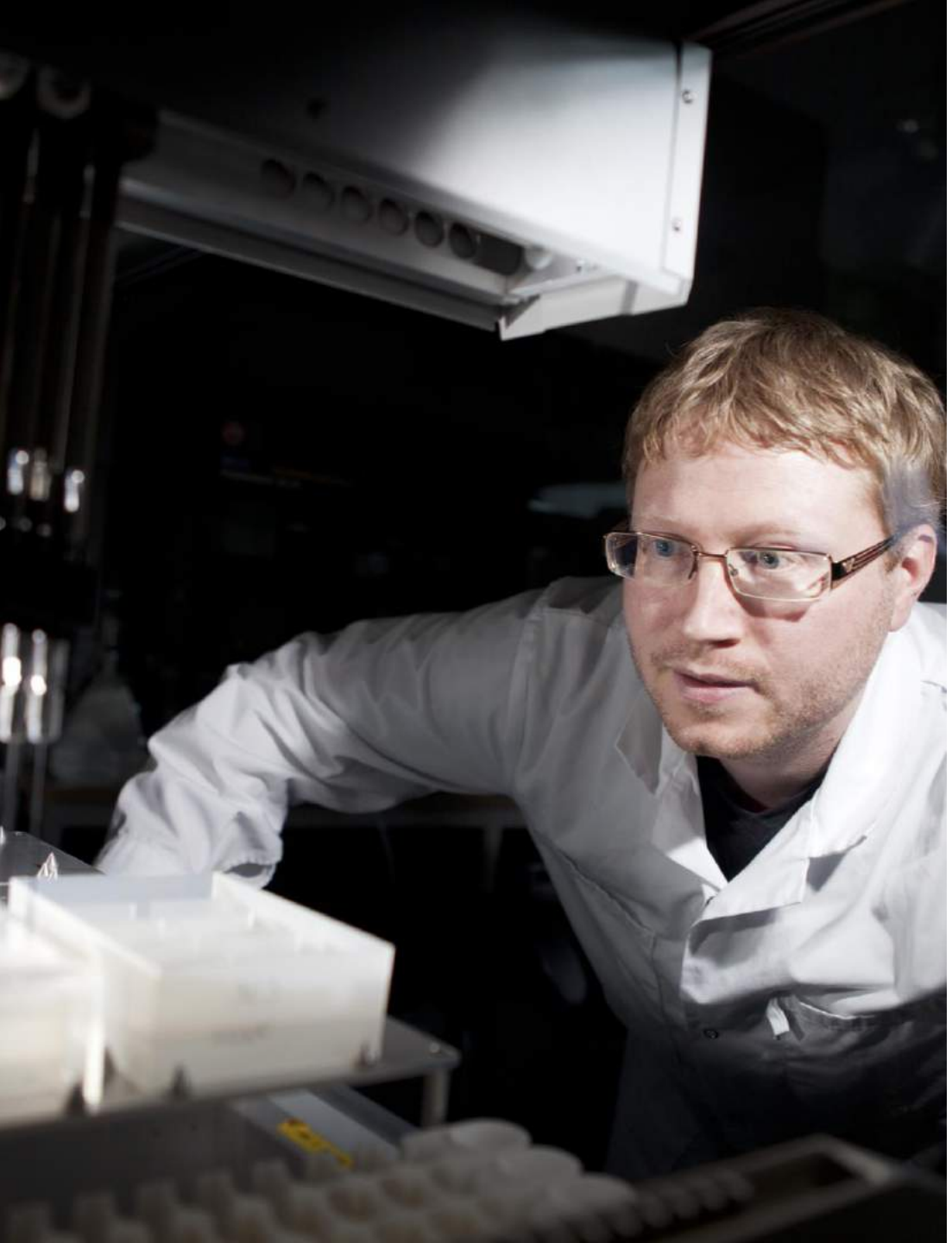
Provides streamlined single-cell sorting, lysis, whole-genome amplification and screening of individual microbial cells, as well as whole genome and targeted gene sequencing (...)

Learn More

National Bioinformatics Infrastructure (NBIS)

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

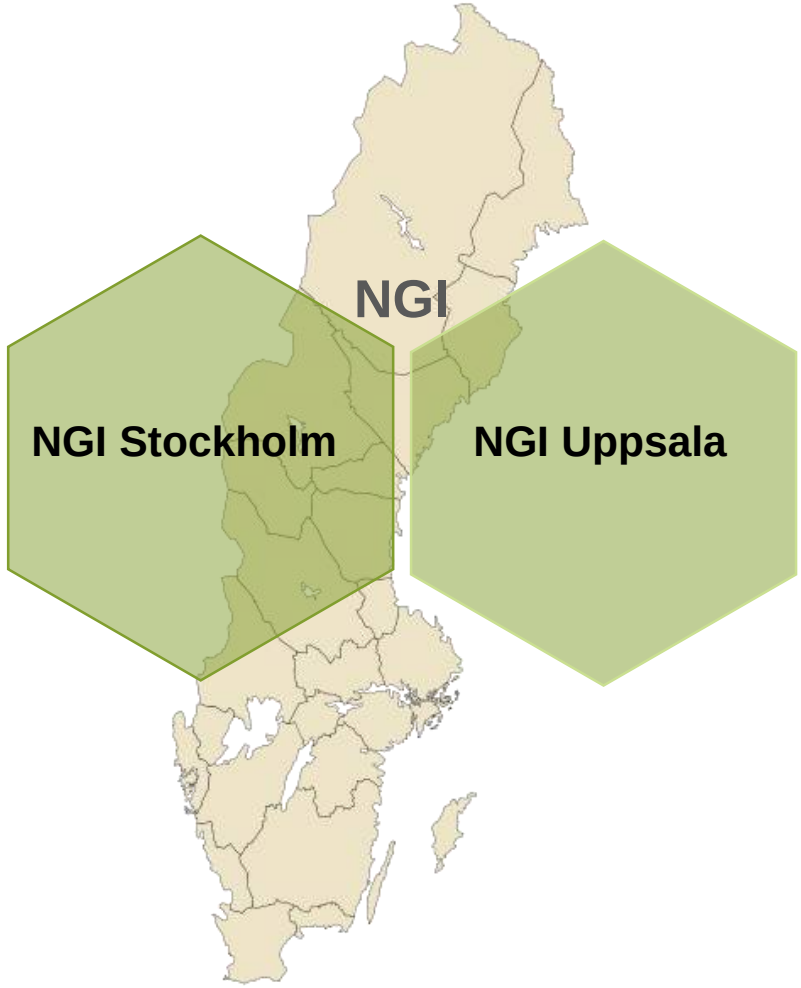
Learn More



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

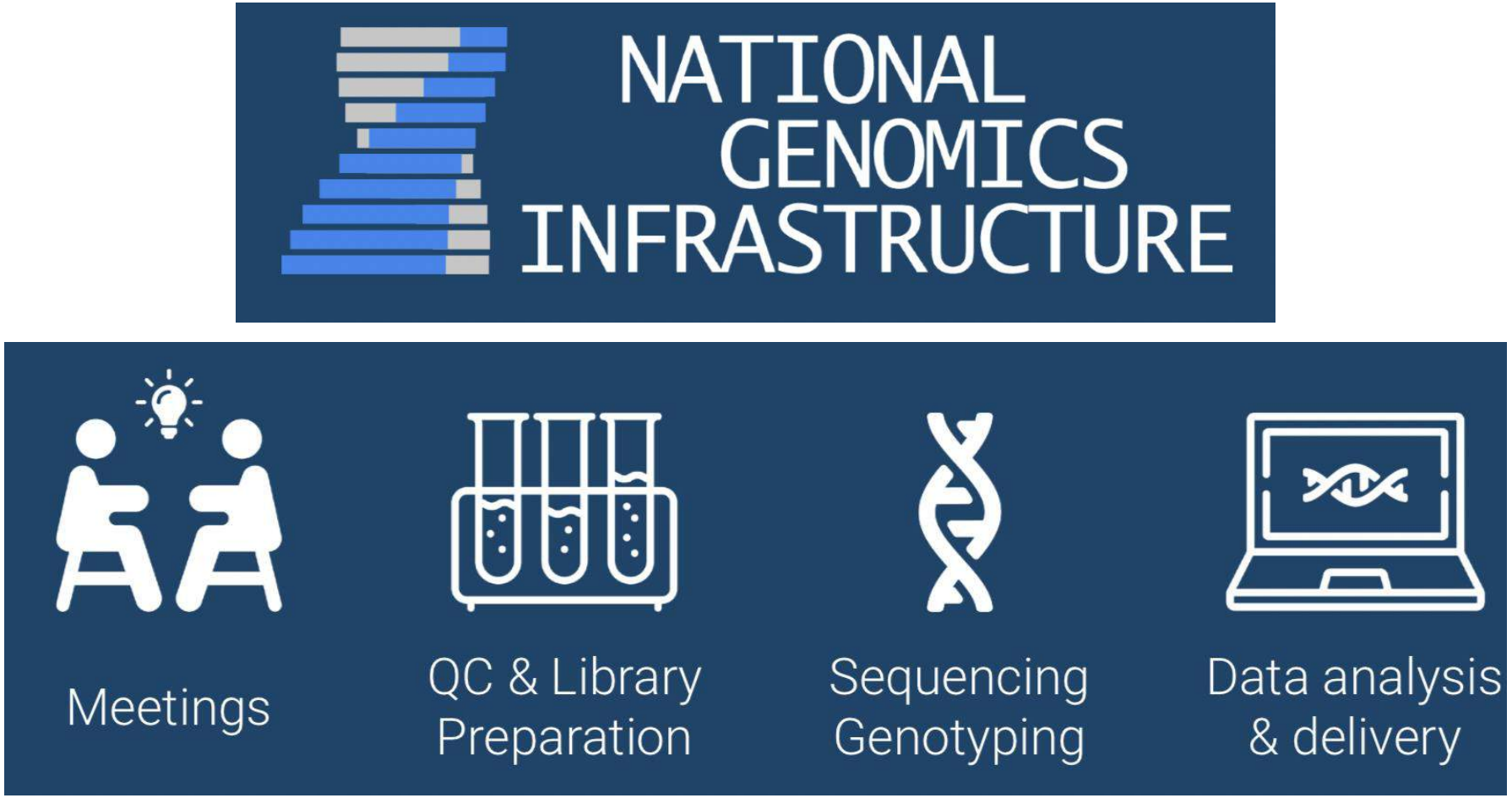
Lars Feuk
Platform Co-Director
Professor UU

NGI-Uppsala SNP&SEQ Technology platform

NGI-Uppsala Uppsala Genome Center

NGI-Stockholm

Project workflow



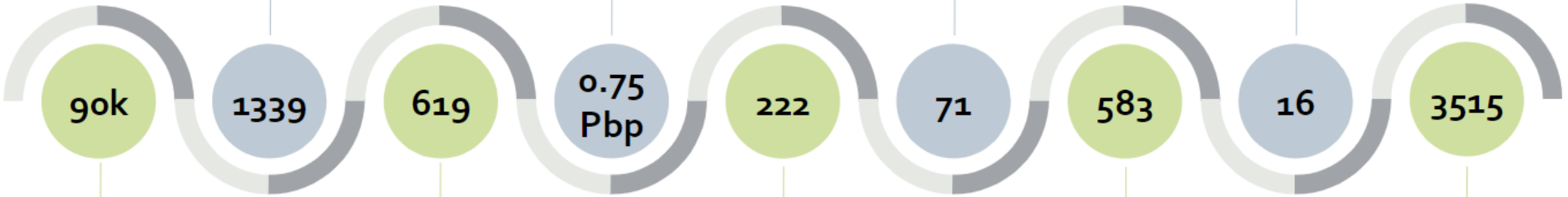
NGI 2023

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

Amount of sequenced base pairs
- 643.2 Tbp – short reads
- 108.5 Tbp – long reads
- 13.1 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



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 2023

Communication tickets
- 44865 ticket updates
- 98.6% 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

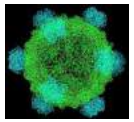
Very long reads, lower troughput

1 x PacBio Revio – New!

1 x PacBio Sequel IIe

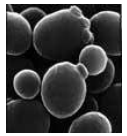
1 x Oxford Nanopore-PromethION

History and current status of sequencing

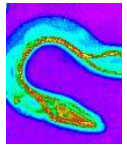


First genome: virus ϕ X 174 - 5 368 bp (1977)

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

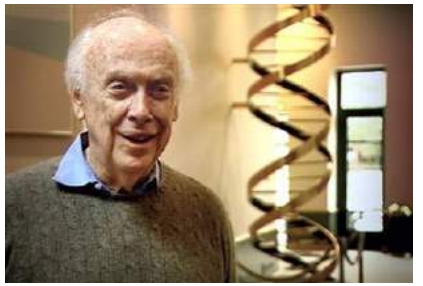
Human genome project (HUGO, 2003)
Sanger Sequencing
2.7 Billion USD

Craig Venter's Genome
Sanger Sequencing
70 Million USD

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

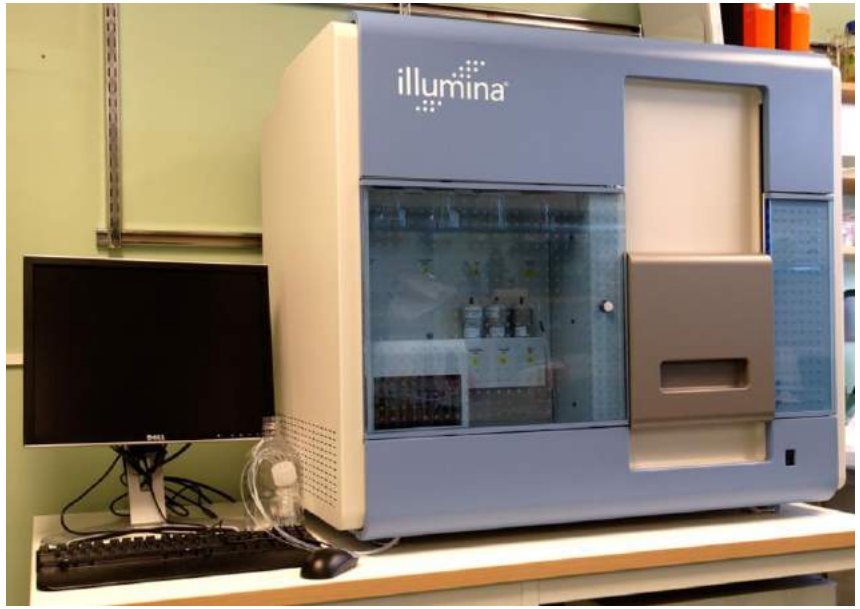
Yesterday's genome
NovaSeq 6000 (Illumina)
~1 000 USD

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



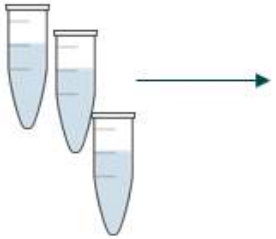
15 years of Illumina sequencing at NGI

2007: Installation of Illumina GA



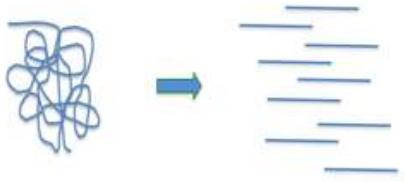
2023: Arrival of NovaSeq X Plus

Workflow, Illumina sequencing

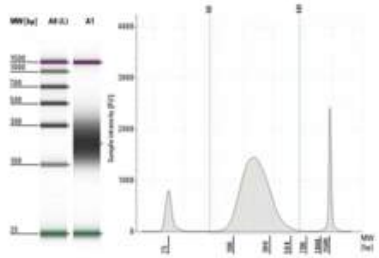


Quality control of template (DNA, RNA, amplified regions)

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

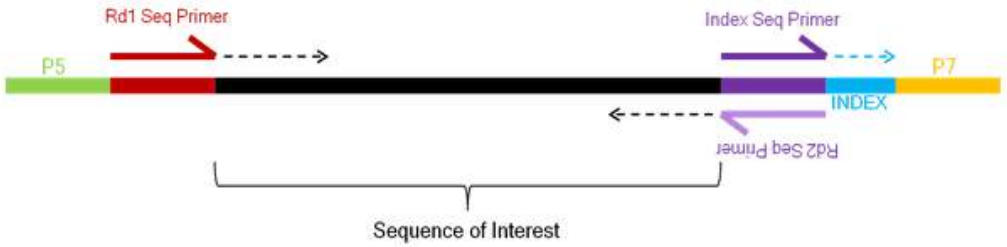
36-300 bp, paired end sequcning
150 Mb-16 Tb per run
12 hours - 3 days

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

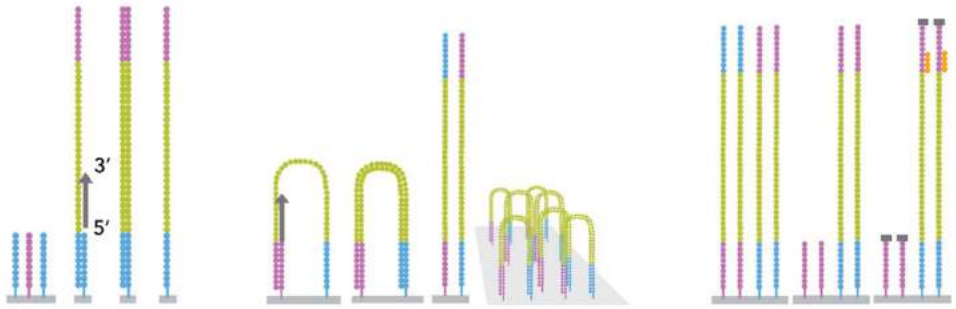
- A sequencing library is a pool of DNA fragments with adapters attached to both ends of the fragments

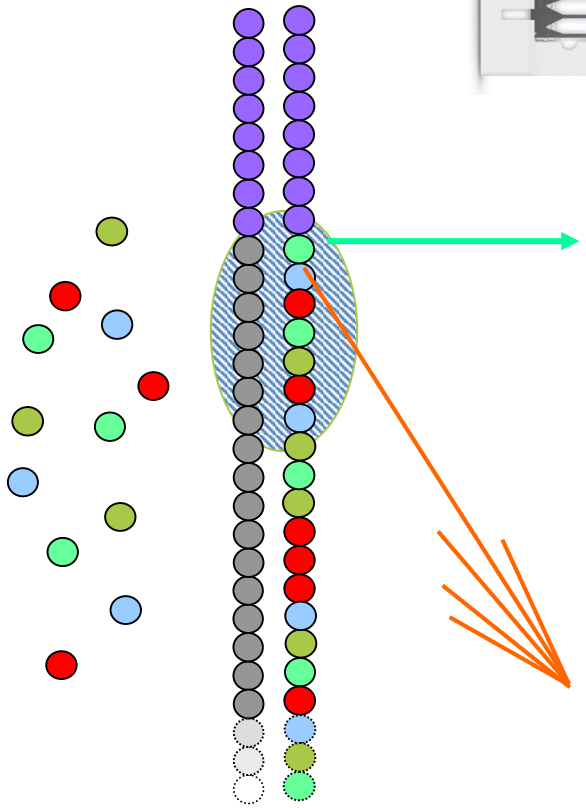
- Approx. 20 protocols for Illumina library prep at NGI

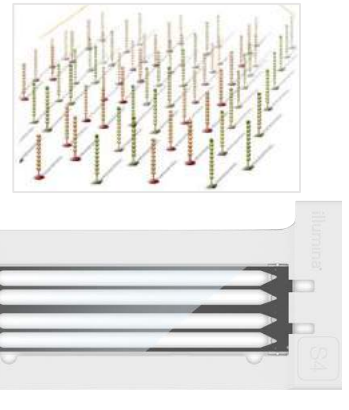


Illumina cluster generation & sequencing

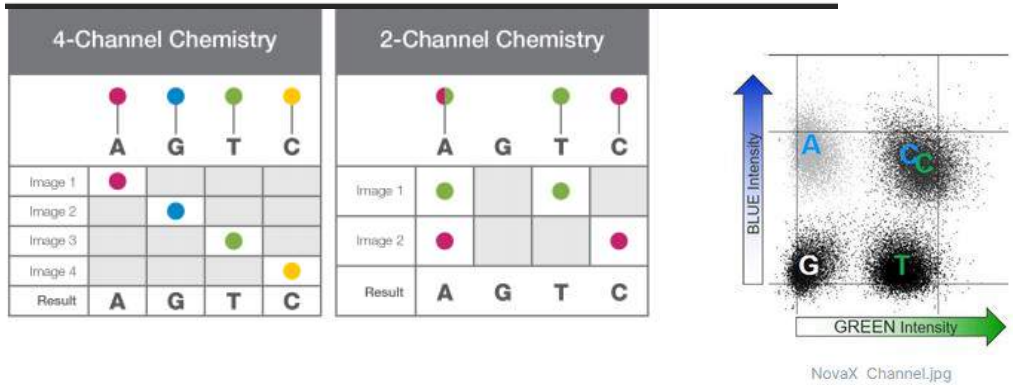
- The sequencing library is hybridized to a flowcell (”cluster generation”)
    - - A flowcell is a slide that is coated with oligos
- Rapid bridge amplification
- Hybidization of sequencing primers
- Sequencing by syntehsis
    - fluorophore labeled nucleotides emitting light







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%

NovaSeq X results for our first 10 runs

Runfolder	Number of reads (B read pairs)	Quality score Q30 (Average % >= Q30)	Error rate Phix (%)
20231019_LH00179_0007_B22CT72LT3	10,16 B/FC (1034-1426 M/lane)	84.6	0.90
20231027_LH00179_0008_A22CT5YLT3	8,84 B/FC (881-1245 M/lane)	87.4	0.40
20231030_LH00179_0009_A22FGLHLT3	11,78 B/FC (1304-1610 M/lane	93.5	0.11
20231110_LH00179_0012_A22FMKMLT3	8,95 B/FC (773-1423 M/lane)	83.8	1.59
20231121_LH00179_0013_B22FMWLLT3	12,05 B/FC (1252-1612 M/lane)	89.5	0.23
20231201_LH00179_0015_A22FY3TLT3	10,9 B/FC (774-1683 M /lane)	92.56	0.32
20231205_LH00179_0016_A22FWYLLT3	12,3 B/FC (1527-1560 M /lane)	92.20	0.19
20231205_LH00179_0017_B22GHKTLT3	11,9 B/FC (1453-1529 M/ lane)	93.87	0.15
20231207_LH00179_0018_A22GHLKLT3	10,7 B/FC (1248-1360 M lane)	92.71	0.20
20231213_LH00179_0019_A22GTJ3LT3	11,7 B/FC (1408-1482 M/lane)	94.5	0.14

Software Upgrade

Advantages and challenges NovaSeqX

+

Cost per base is low

Quick data generation

Easy workflow in the lab

Reagents shipped in RT

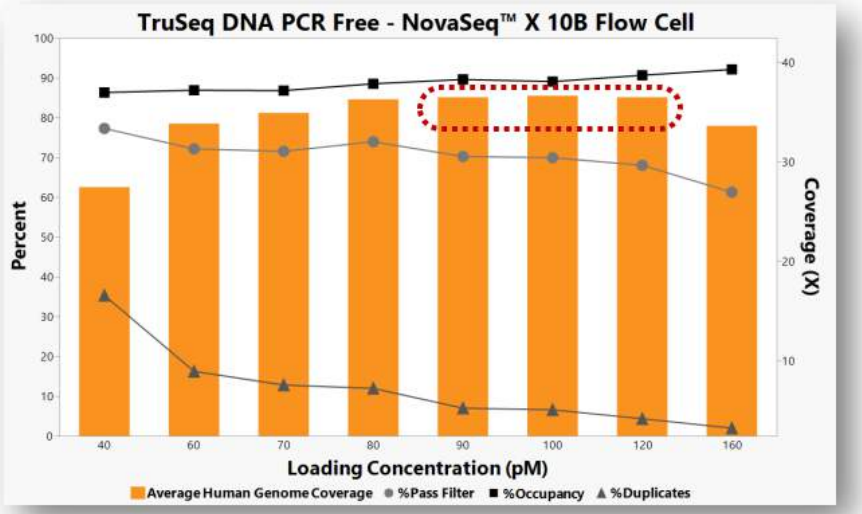
On-instrument analysis

-

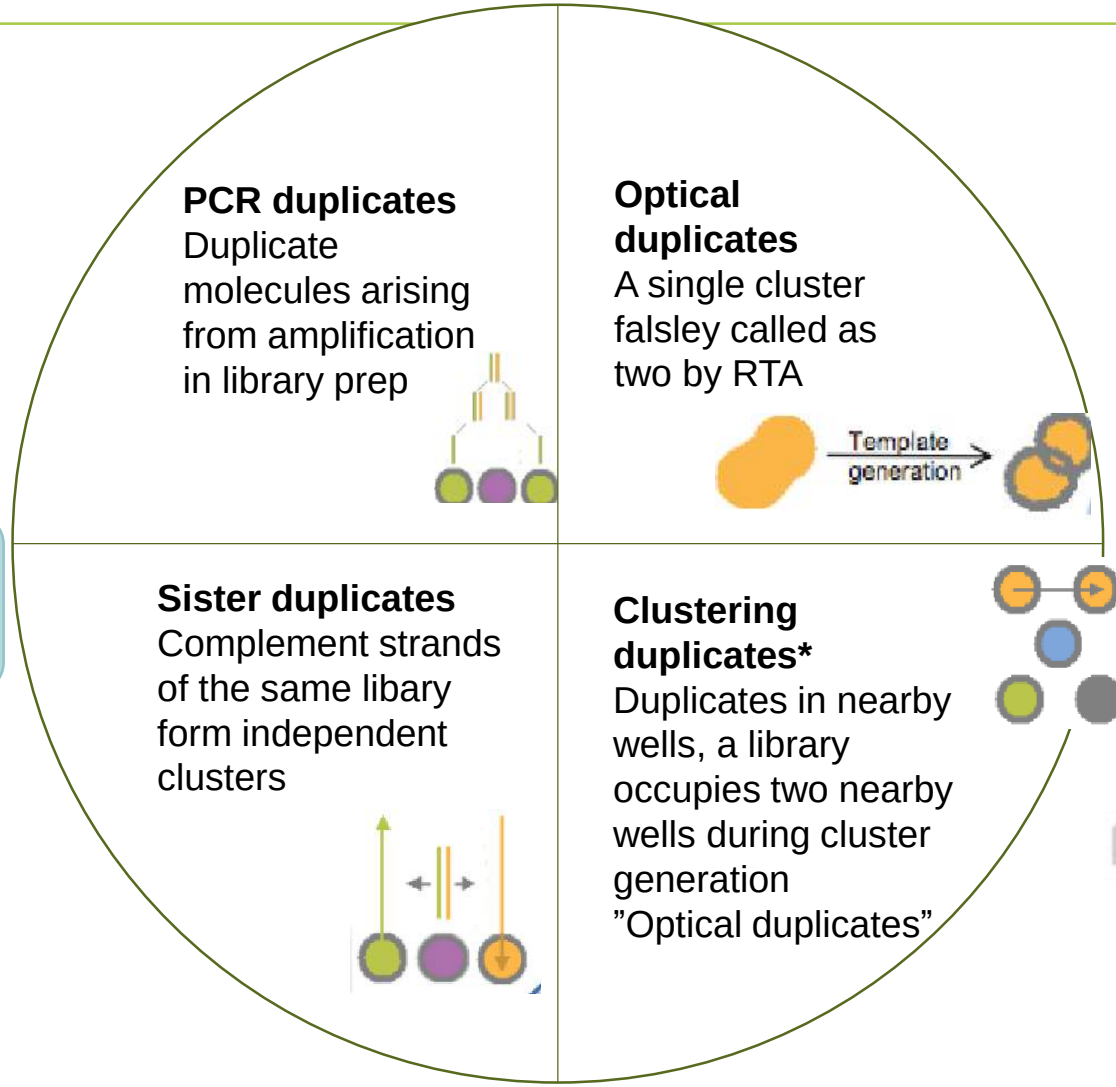
Yield vs duplicates

More sensitive to challenging samples and short inserts

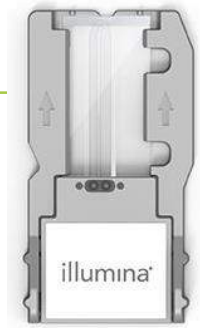
Sensitive to colour balancing (C-A)



Duplicates, duplicates, duplicates….



Present on all Illumina Platforms



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

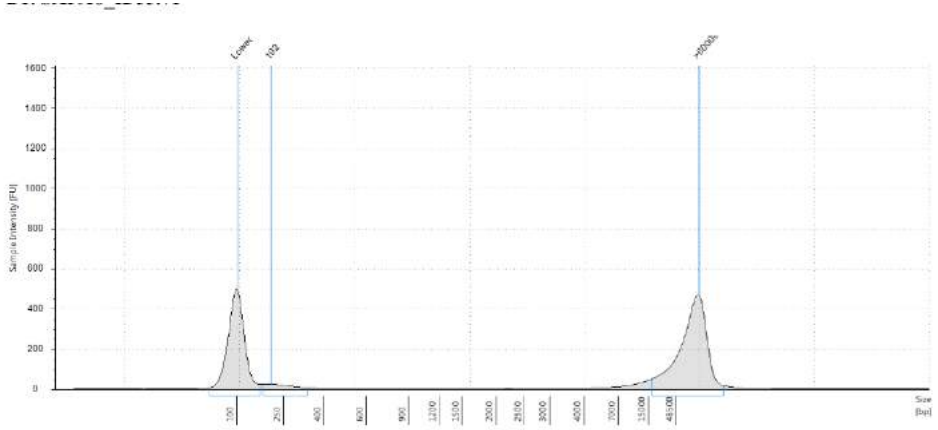
Patterned flowcells only (NextSeq, NovaSeq 6000/X

Quality control of RNA/DNA

DNA

Concentration: QuantIT

Degradation: Fragment Analyzer/TapeStation



Sample Table

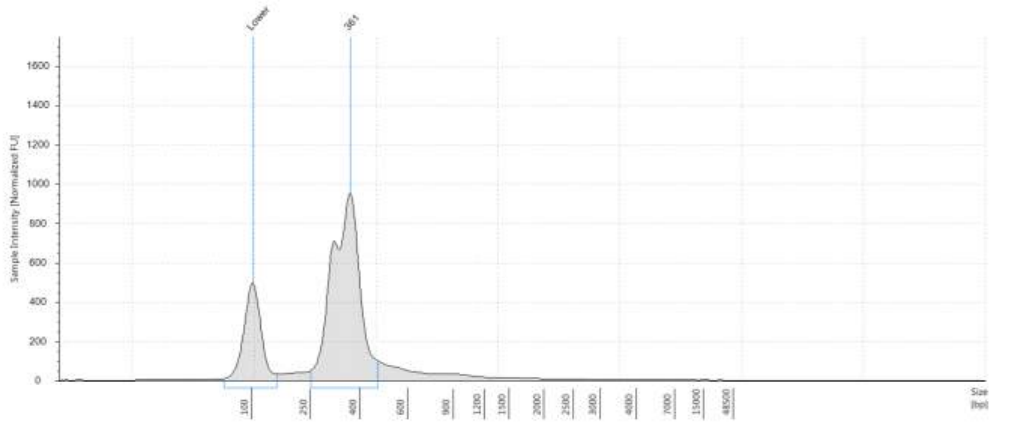
Well	DIN	Conc. [ng/µl]	Sample Description	Alert	Observations
B1	9.6	16.0	SX1018_ID33.v1		

High quality DNA sample

RNA

Concentration + RIN-value:

Fragment Analyzer/TapeStation



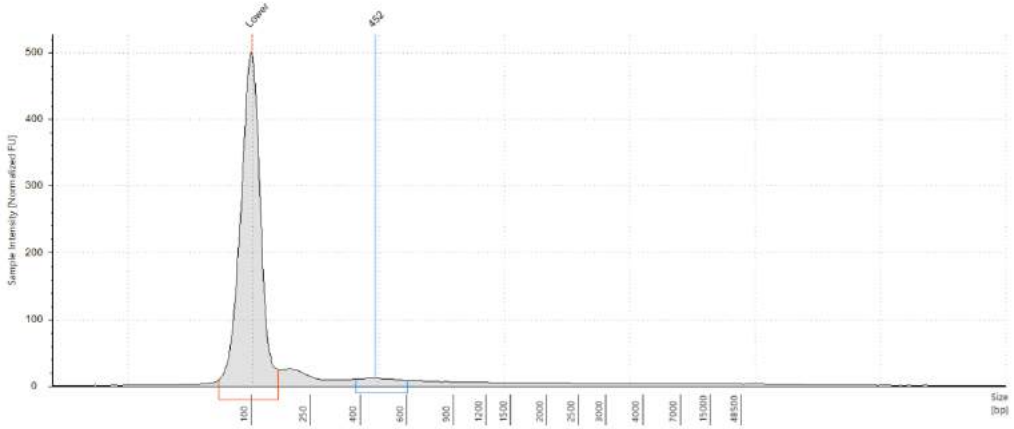
Sample Table

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



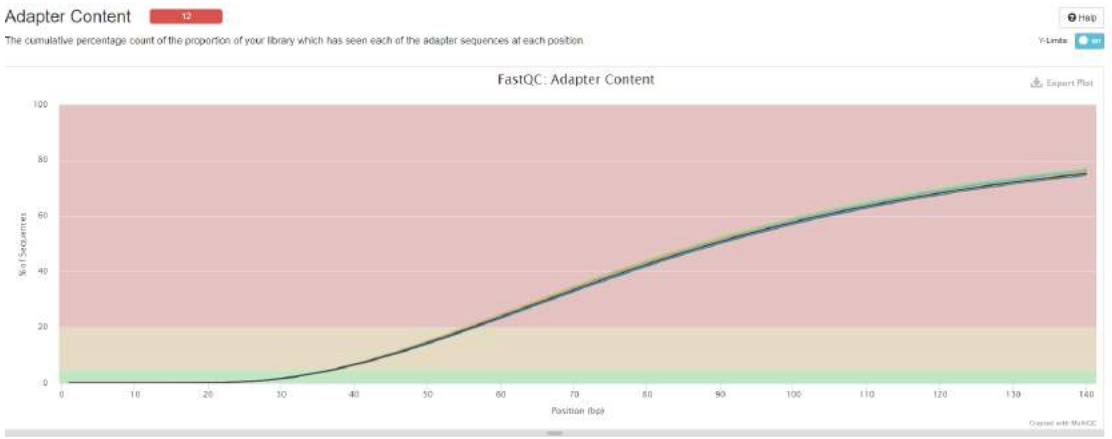
Sample Table

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

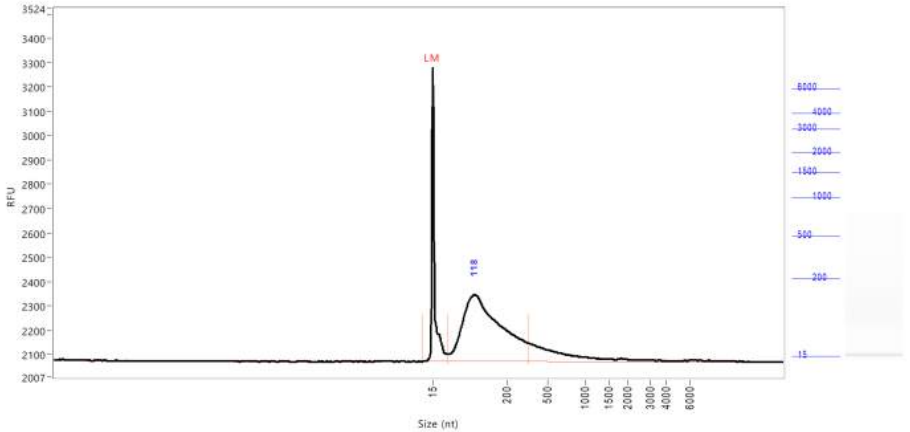


Copy table | Configure Columns | Plot | Showing 7/7 rows and 14/23 columns.

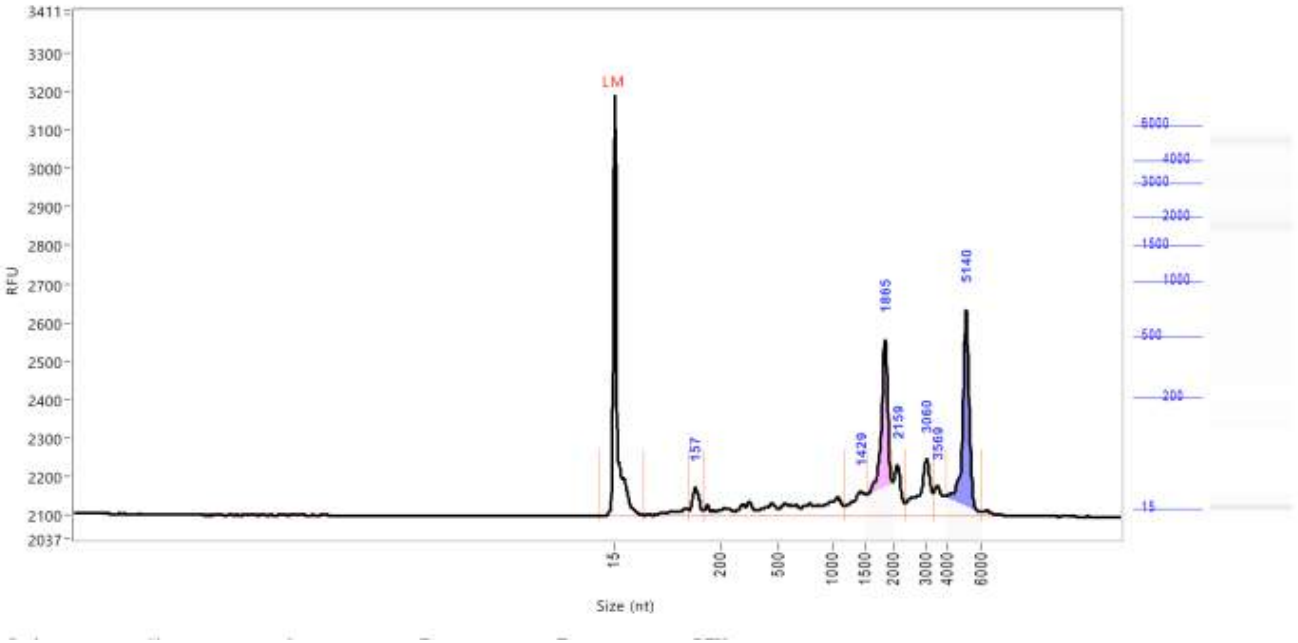
Sample Name	% GC	Ins. size	≥ 30X	Coverage	% Aligned	Change rate	Ts/Tv	M Variants	TiTV ratio (known)	TiTV ratio (novel)	% Dups	% Dups	% GC	M Seqs
S1	46%	55	11.1%	2.0X	98.2%	893	1.645	3.47	2.0	1.6	76.6%			

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

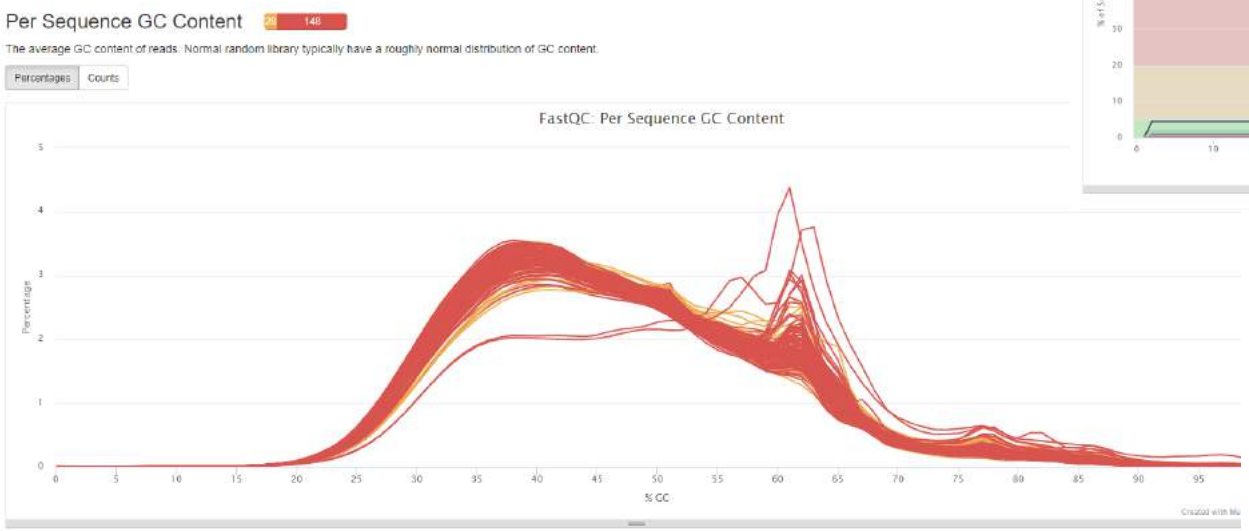
QC-reults RNA-seq

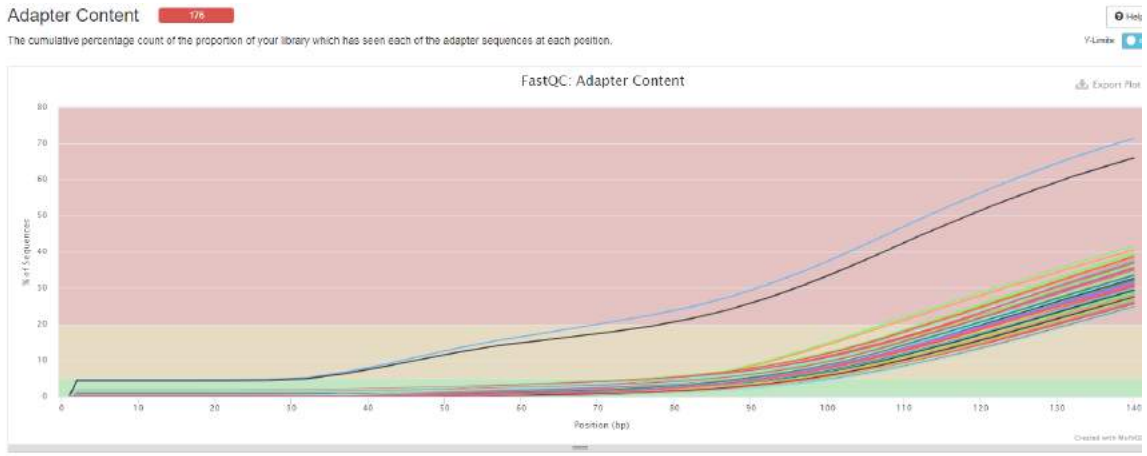
Uneven amounts of data (17-100 M reads per sample)

A lot of duplicates

High rRNA content

High adapter content

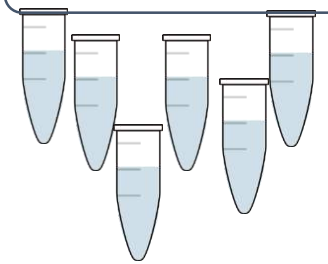




Some of the applications offered

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

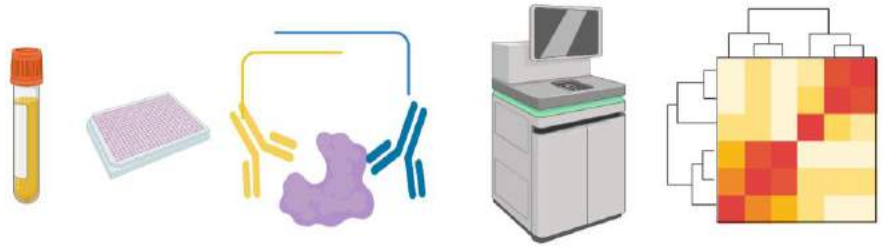
Spatial transcriptomics

- 10x Genomics Visium

Proteomics with NGS readout

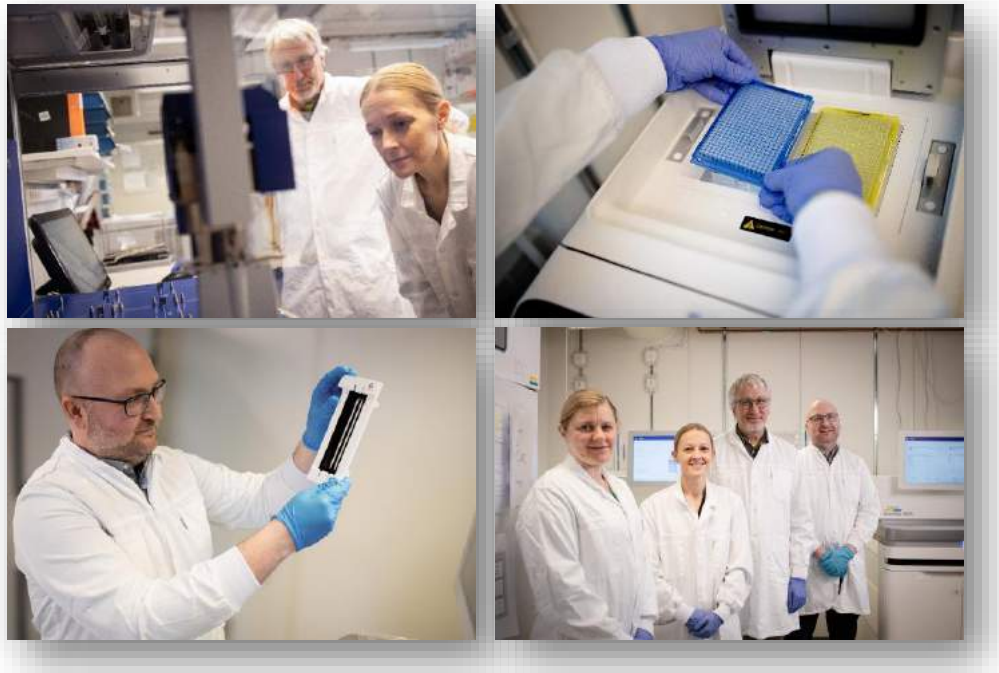
- Olink Explore 1536/3072/5300

Protein analysis, Olink Explore with NGS readout



- 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

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



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

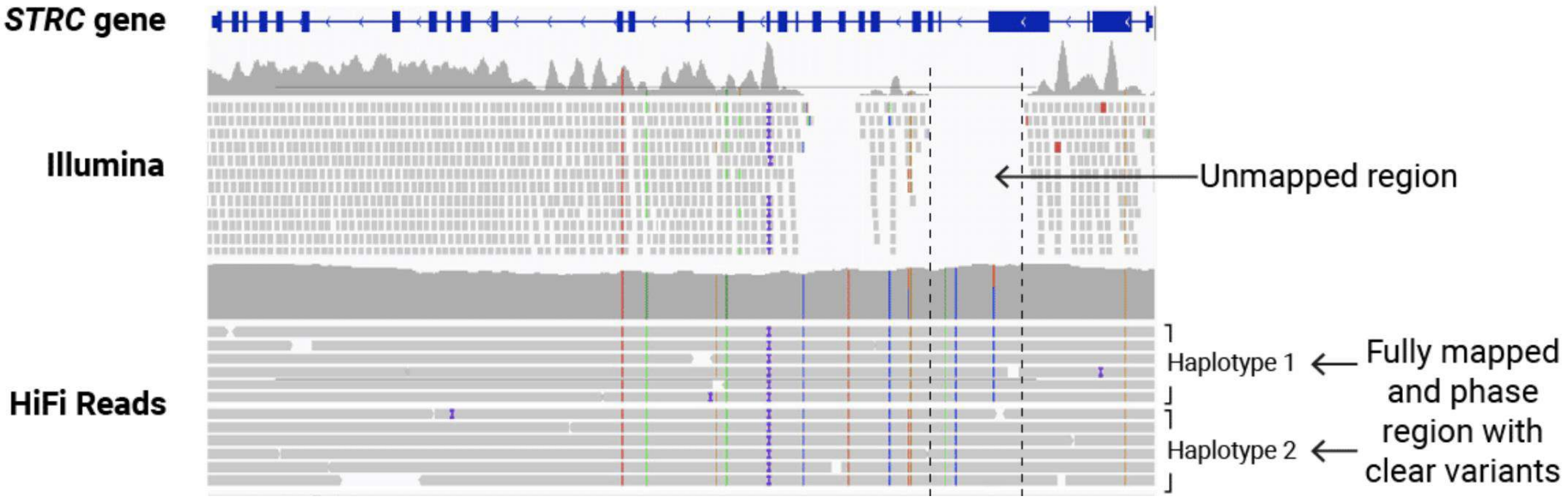
Tom van der Valk, Patrícia Pečnerová, David Díez-del-Molino, Anders Bergström, Jonas Oppenheimer, Stefanie Hartmann, Georgios Xenikoudakis, Jessica A. Thomas, Marianne Dehasque, Ekin Sağlıcan, Fatma Rabia Fidan, Ian Barnes, Shanlin Liu, Mehmet Somel, Peter D. Heintzman, Pavel Nikolskiy, Beth Shapiro,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

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

30k Accesses | 89 Citations | 2528 Altmetric | Metrics

Limitations with short reads

- You don’t get complete genome information!

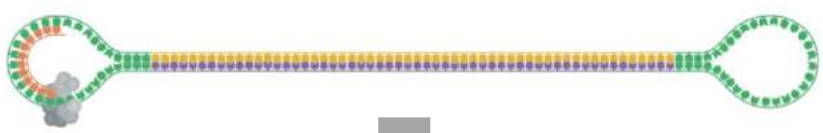


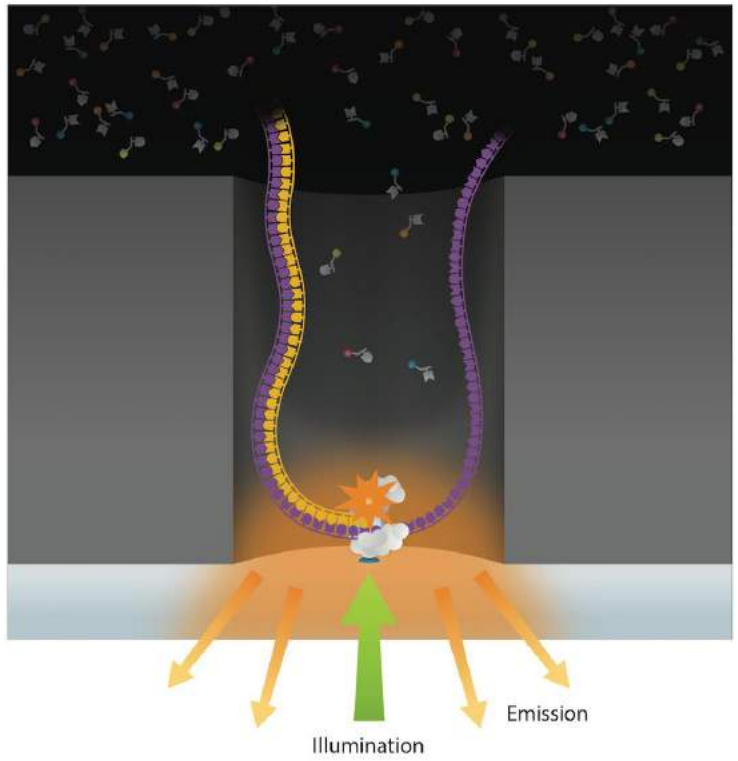
Long-read sequencing

No longer a niche technology!

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

PacBio Sequencing

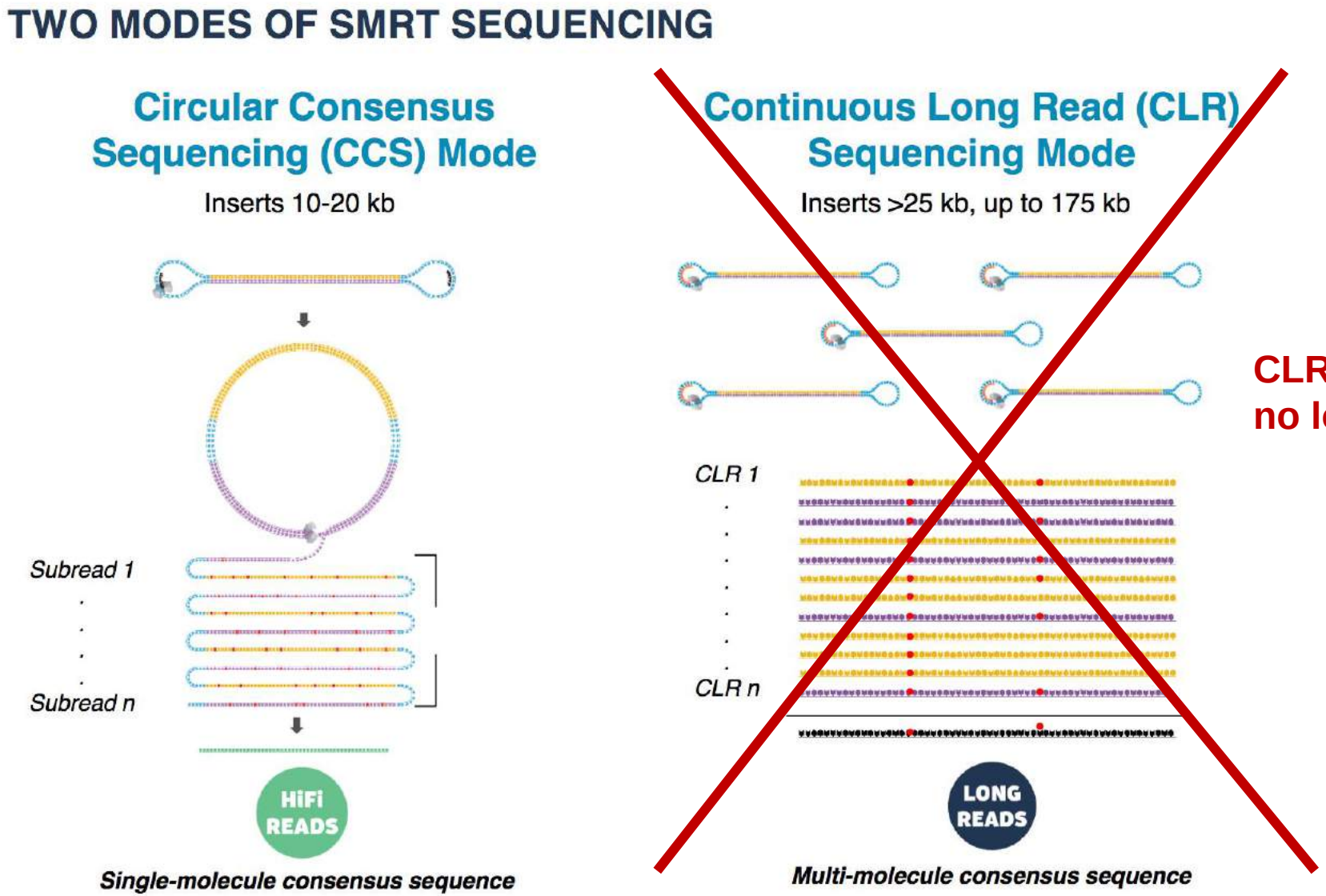




PacBio RSII

PacBio Sequel (Sequel I & II)

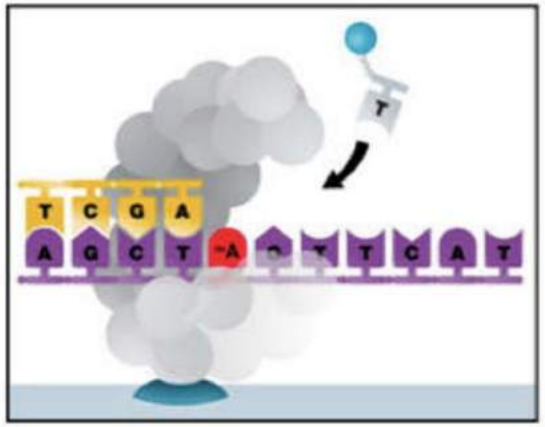
PacBio Sequencing

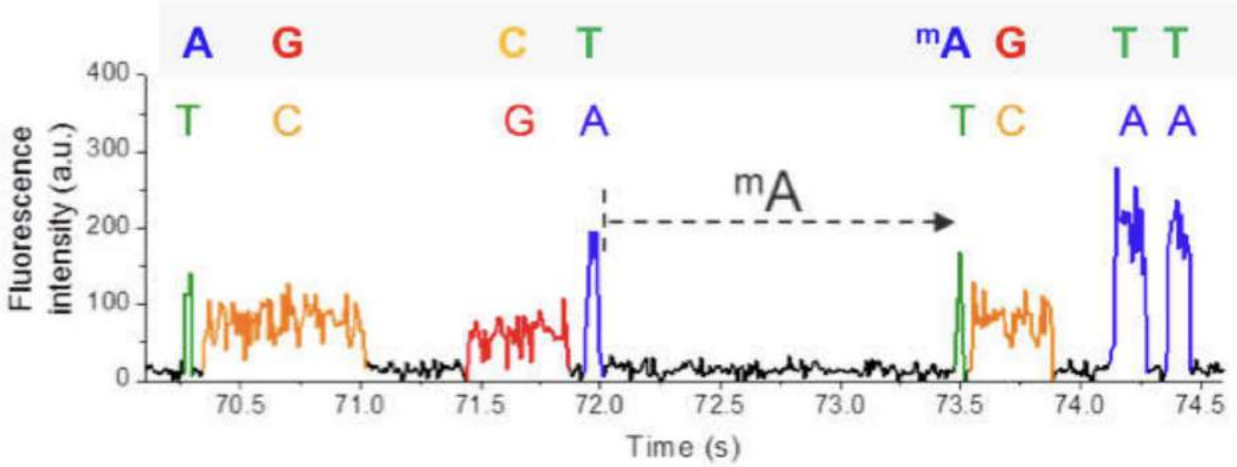


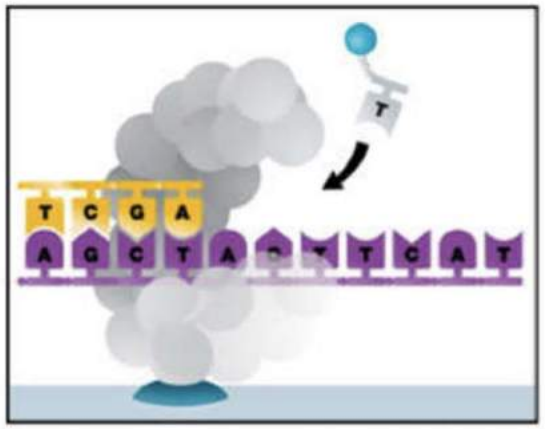
CLR sequencing no longer supported

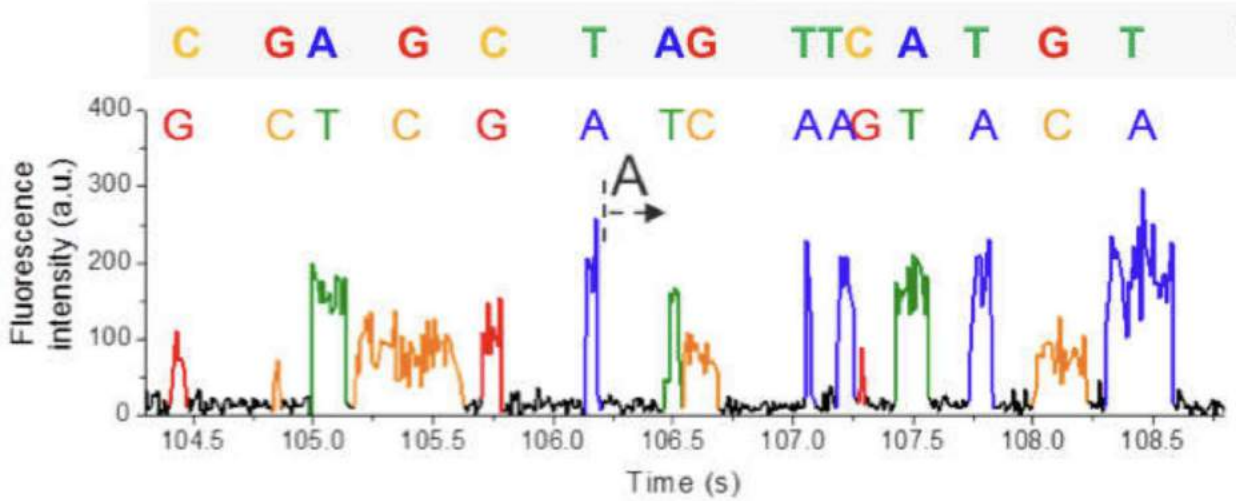
PacBio – Methylation detection

- 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

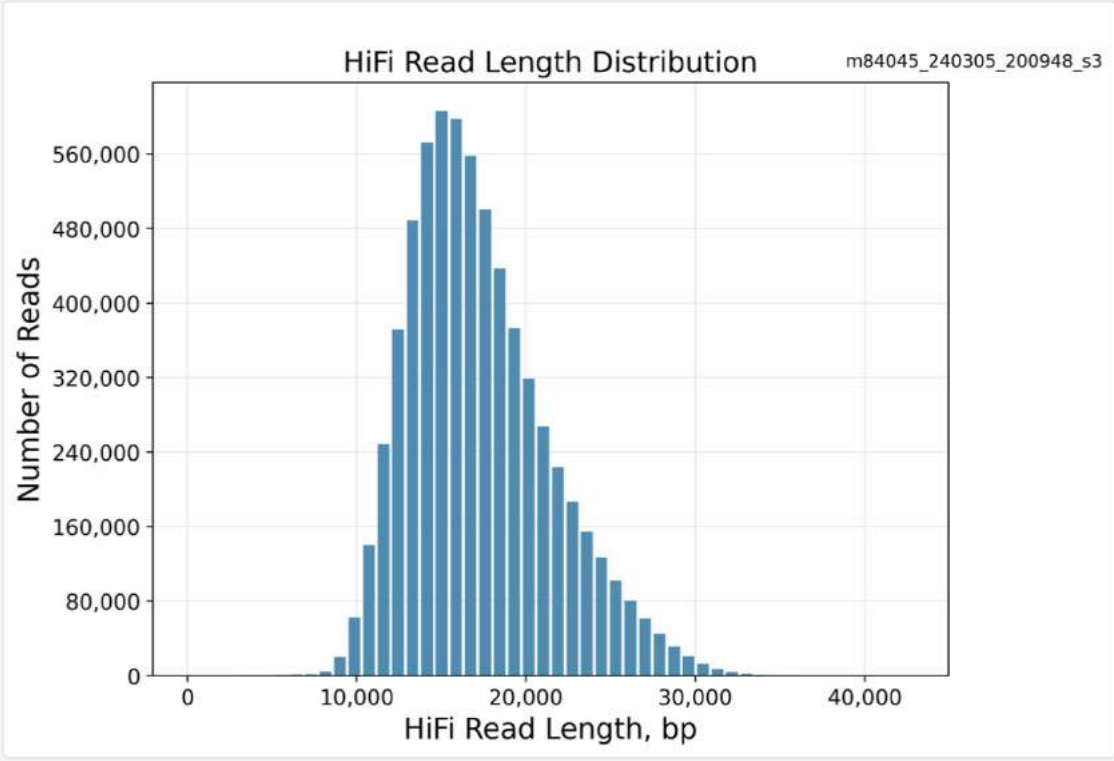
- Up to 90Gb data from one SMRT cell
- Read lengths: 15-20kb
- >QV20 quality (>99% read accuracy)
- Can run 1,300 human genomes/year!
- We installed PacBio Revio in March 2023

Revio – results for our first 16 runs

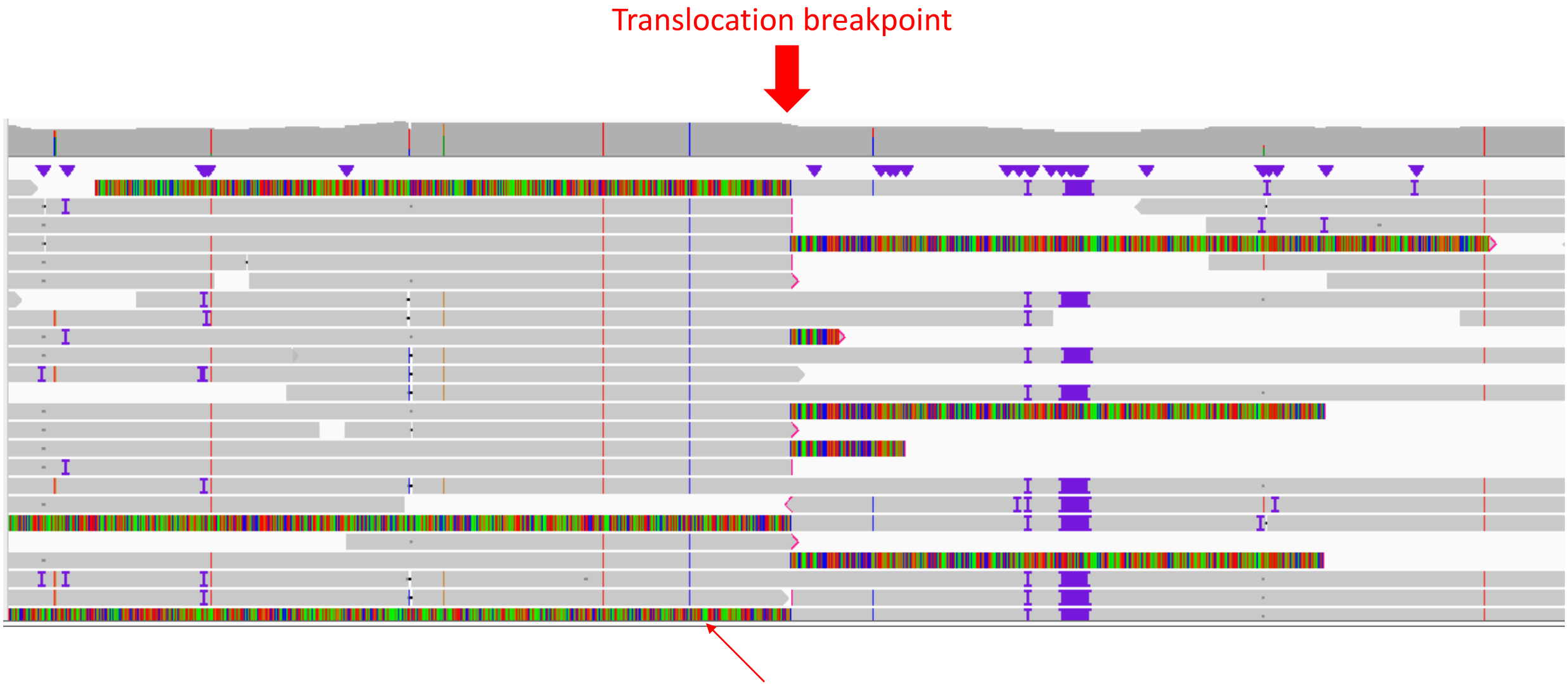
Sample/Species/Proj	Number of reads	Total yield (Gbp)	Average read length (kb)	Size selection method	Comment
Human 1_1	6,873,030	84.7	12.3	Ampure beads	Also Sequel II data
Human 1_2	6,846,419	102.2	15.0	Ampure beads	Also Sequel II data
Human 1_3	7,170,075	90.3	12.6	Ampure beads	Also Sequel II data
Human 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	

Our best run so far > 114 Gb

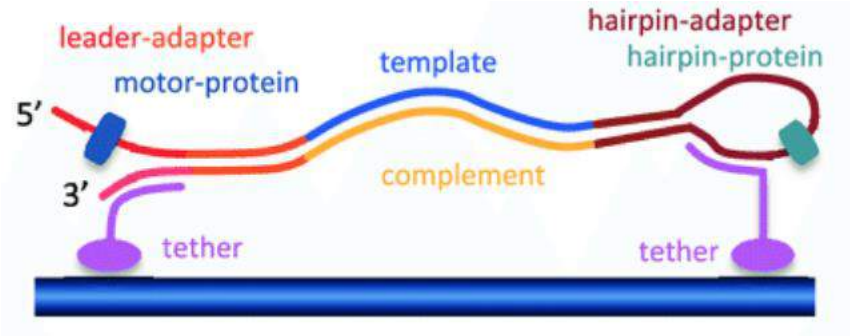
Value	Analysis Metric
6.6 M	HiFi reads
114.17 Gb	HiFi reads yield
17.21 kb	HiFi reads length (mean)
16,564	HiFi reads length (median, bp)
17,585	HiFi Read Length N50 (bp)
Q34	HiFi Read Quality (median)
92.36%	Base Quality ≥Q30 (%)
8	HiFi Number of Passes (mean)

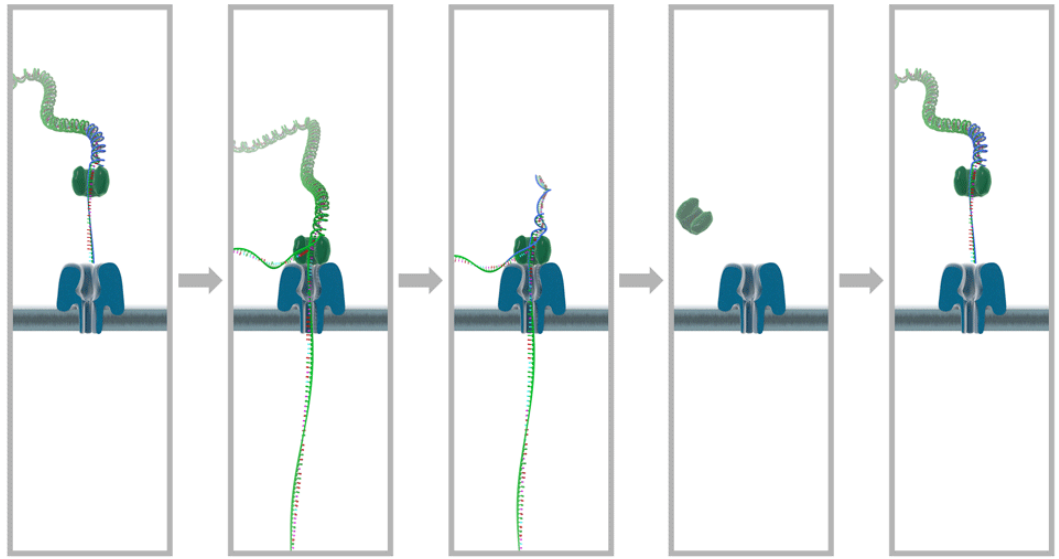


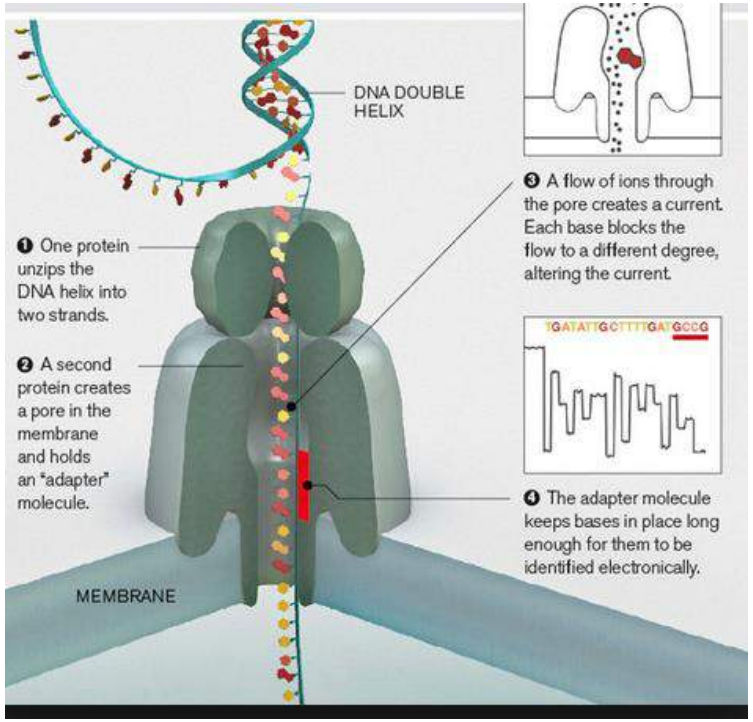
Example: Data at a transclocation site



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

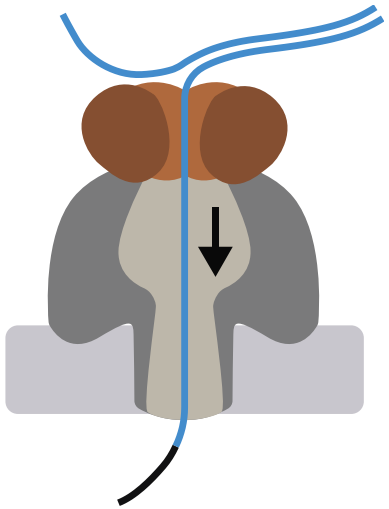
C. Vermeulen, M. Pagès-Gallego, L. Kester, M. E. G. Kranendonk, P. Wesseling, N. Verburg, P. de Witt Hamer, E. J. Kooi, L. Dankmeijer, J. van der Lugt, K. van Baarsen, E. W. Hoving, B. B. J. Tops ✉ & J. de Ridder ✉

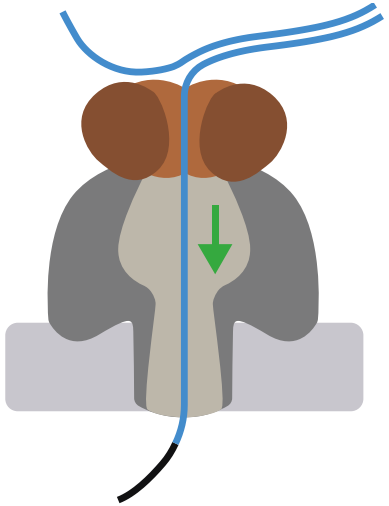
Nature **622**, 842–849 (2023) | Cite this article

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

ONT target sequencing - adaptive sampling

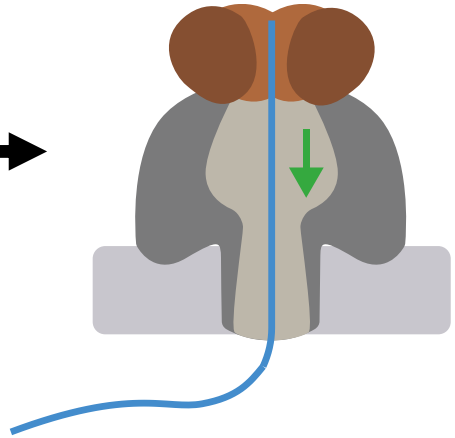
DNA analyzed and sequenced in real time





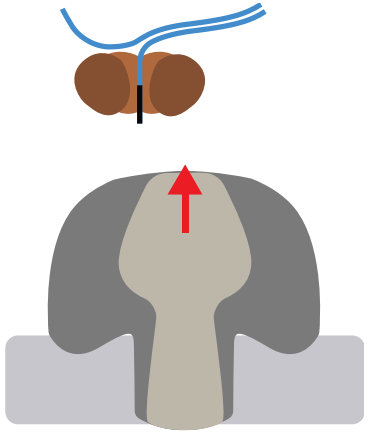
ROI found

Sequencing continues



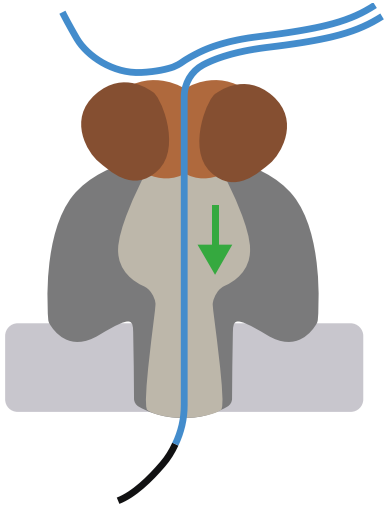
Pore completes sequencing

(Then available for new fragment)



ROI not found

DNA fragment is ejected



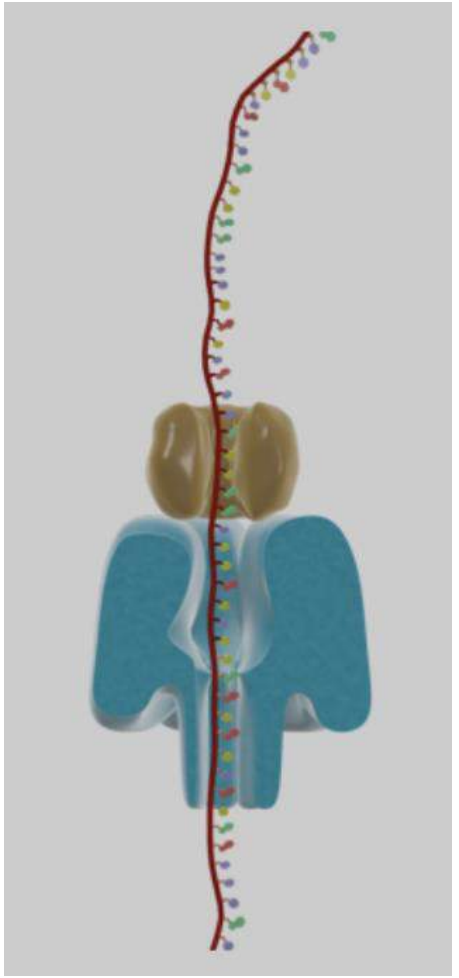
Pore is available for new DNA fragment

(Here with ROI)

ONT direct RNA sequencing

ONT can sequence native RNA molecules!

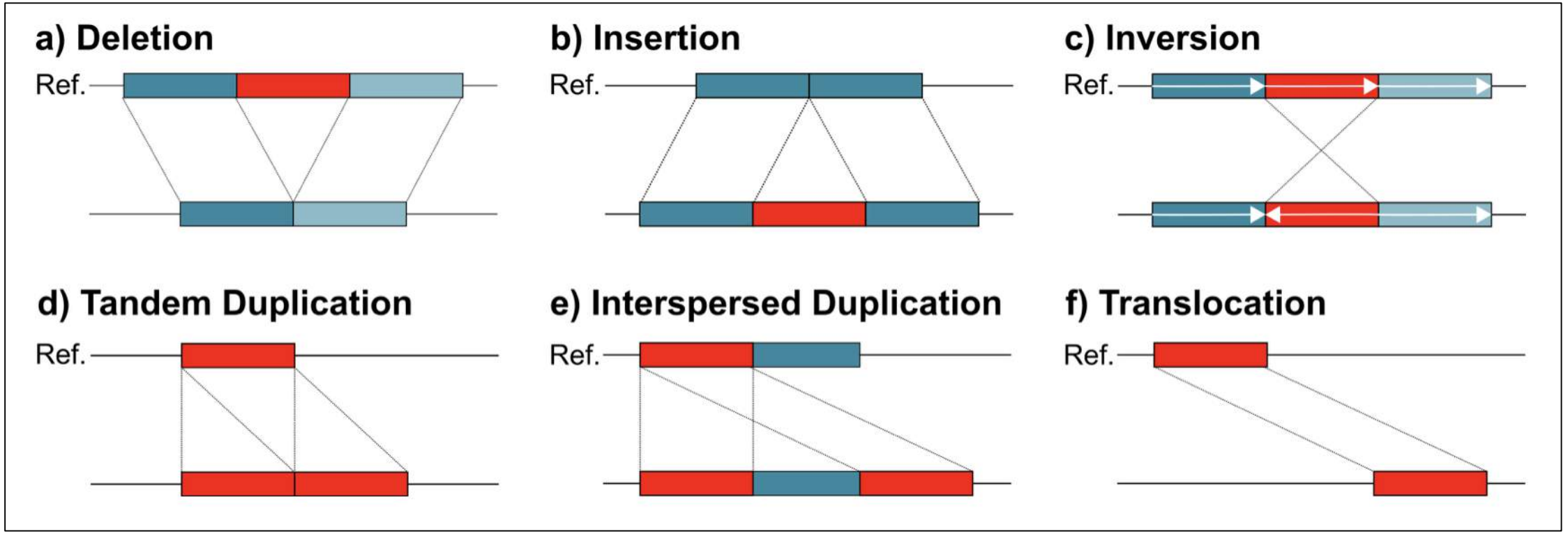
- No bias due to cDNA conversion
- Allows to study RNA modifications
- Higher error rate
- Lower throughput



What people are using long reads for…

Example 1: Detect all genetic variants

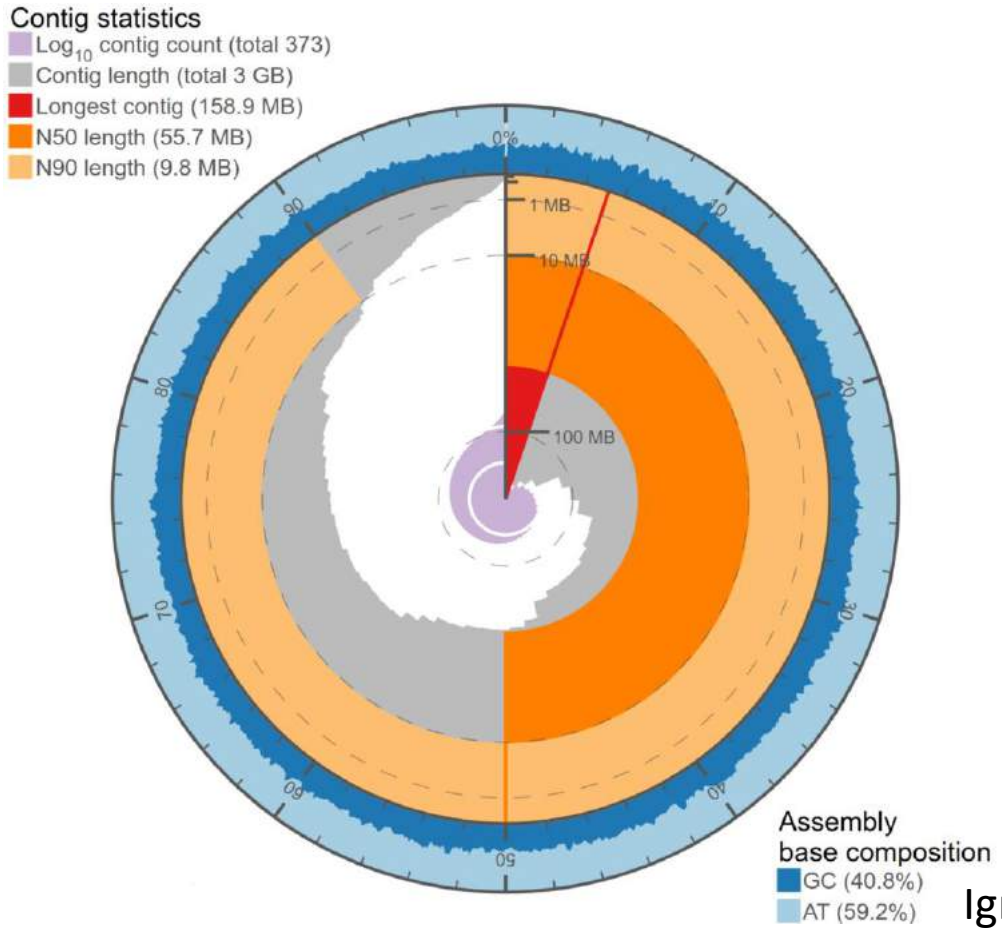
Long-read sequencing can detect more genetic variants than with short reads:



Example 2: Assemble complete genomes

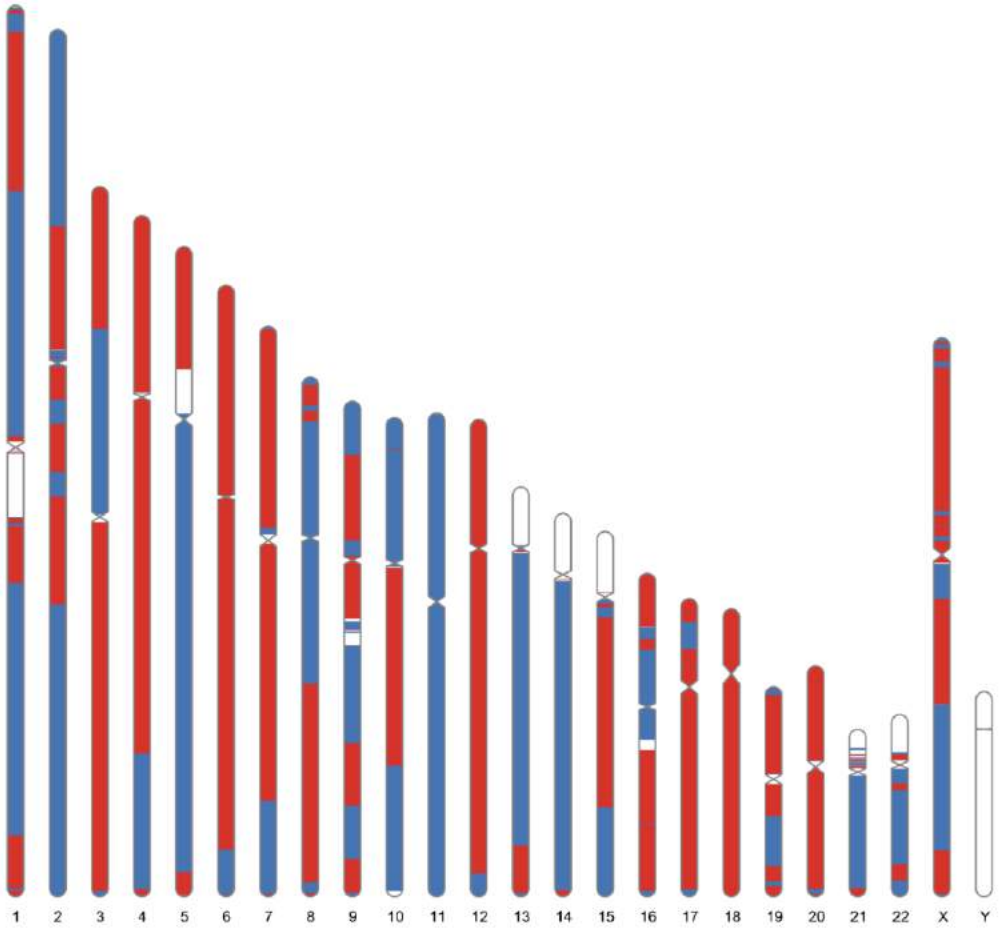
It took just **3.5 h** on a **96** core compute node for *de novo* assembly of a human sample!

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

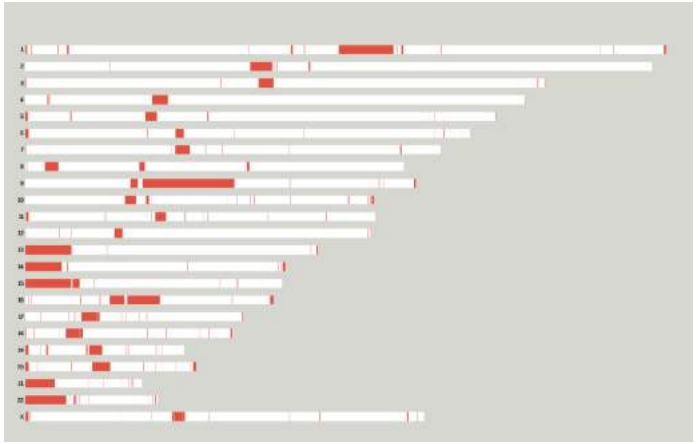
De novo assembly mapped to GRCh38



Colour change represents adjacent contigs

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

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

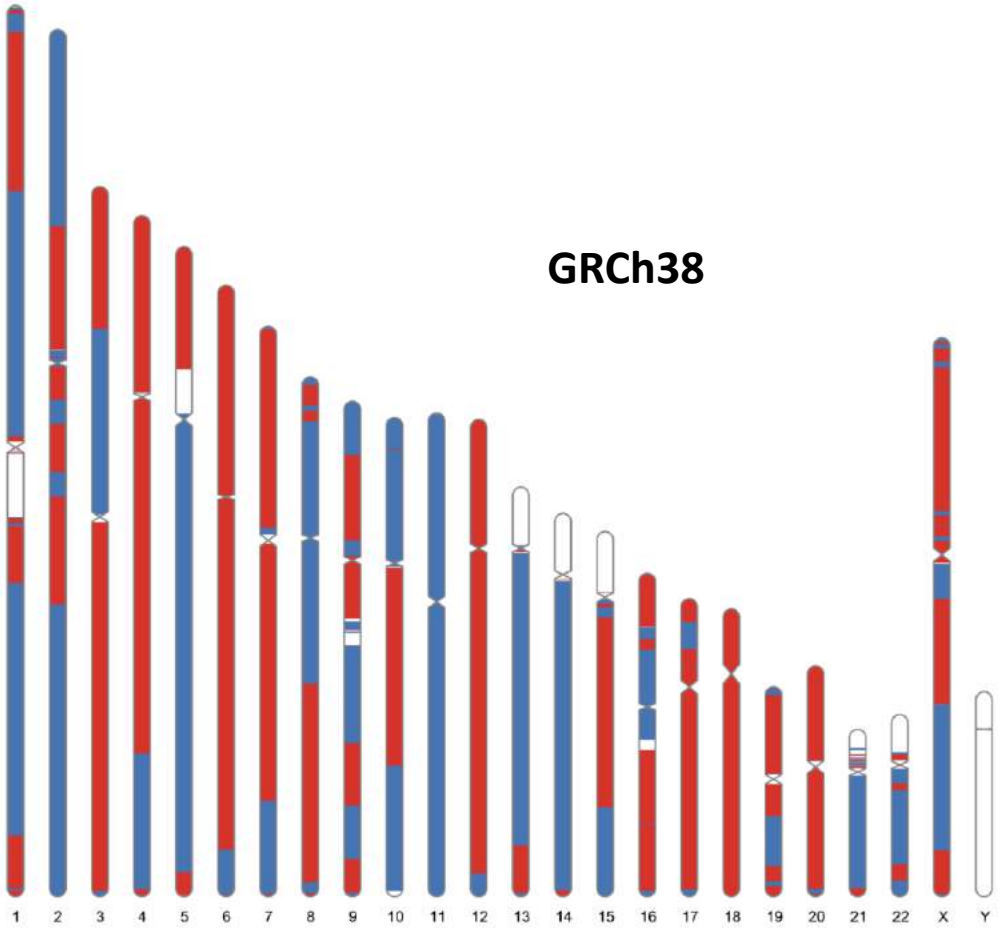


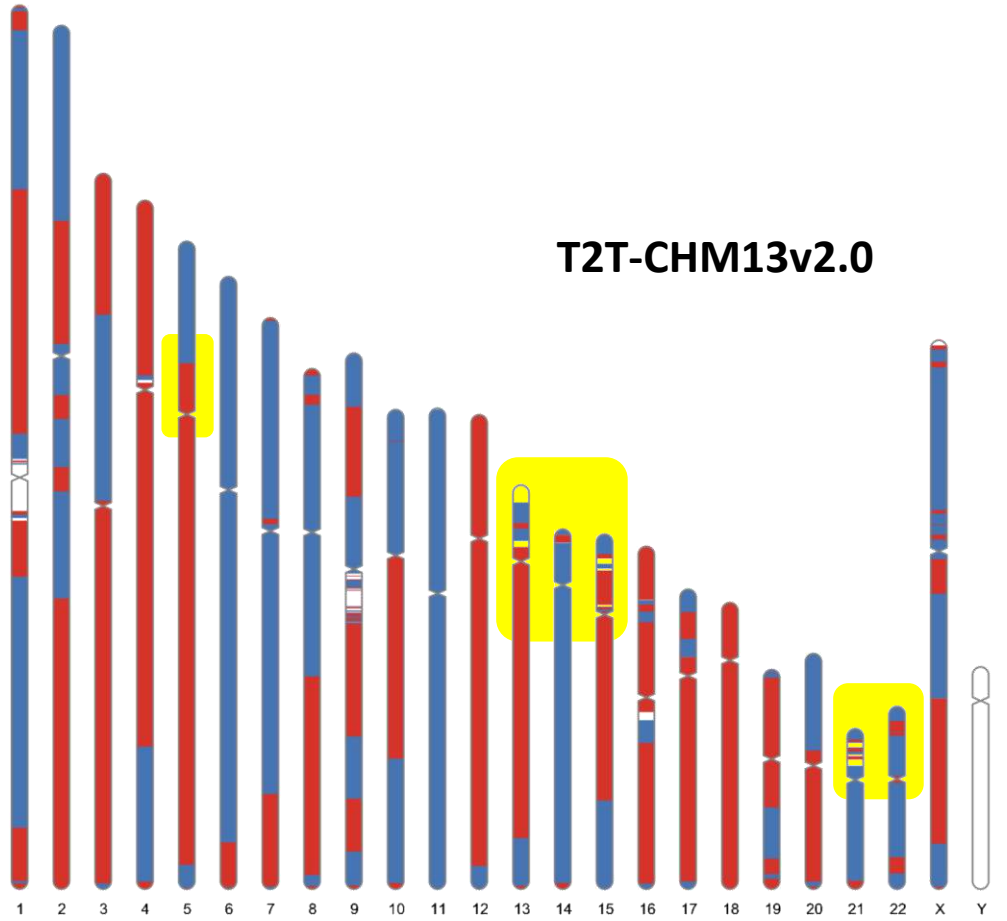
Red segments resolved by T2T Consortium

DOI: 10.1126/science.abp8653

Ignas Bunikis

De novo assembly mapped to T2T



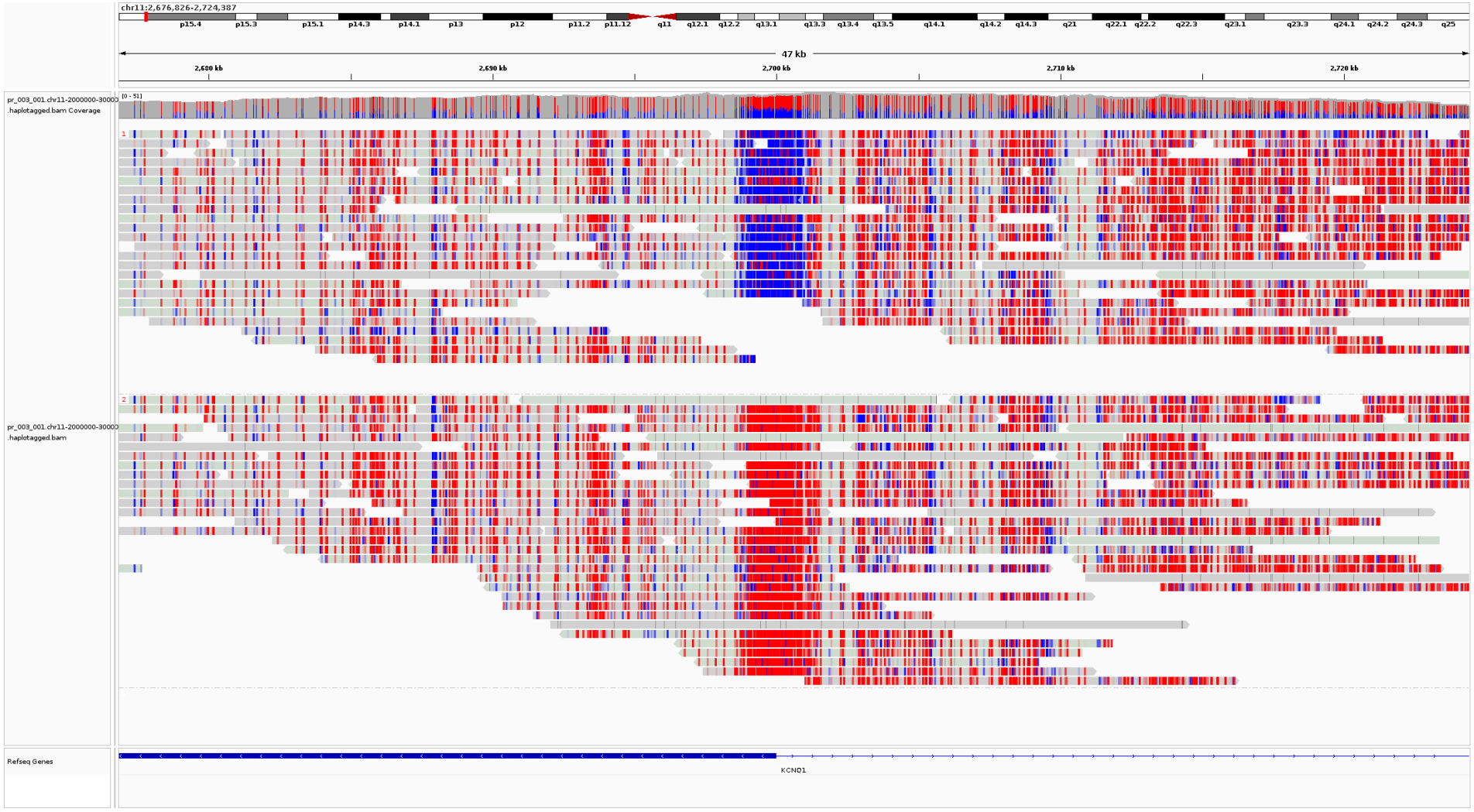


Colour change represents adjacent contigs

Ignas Bunikis

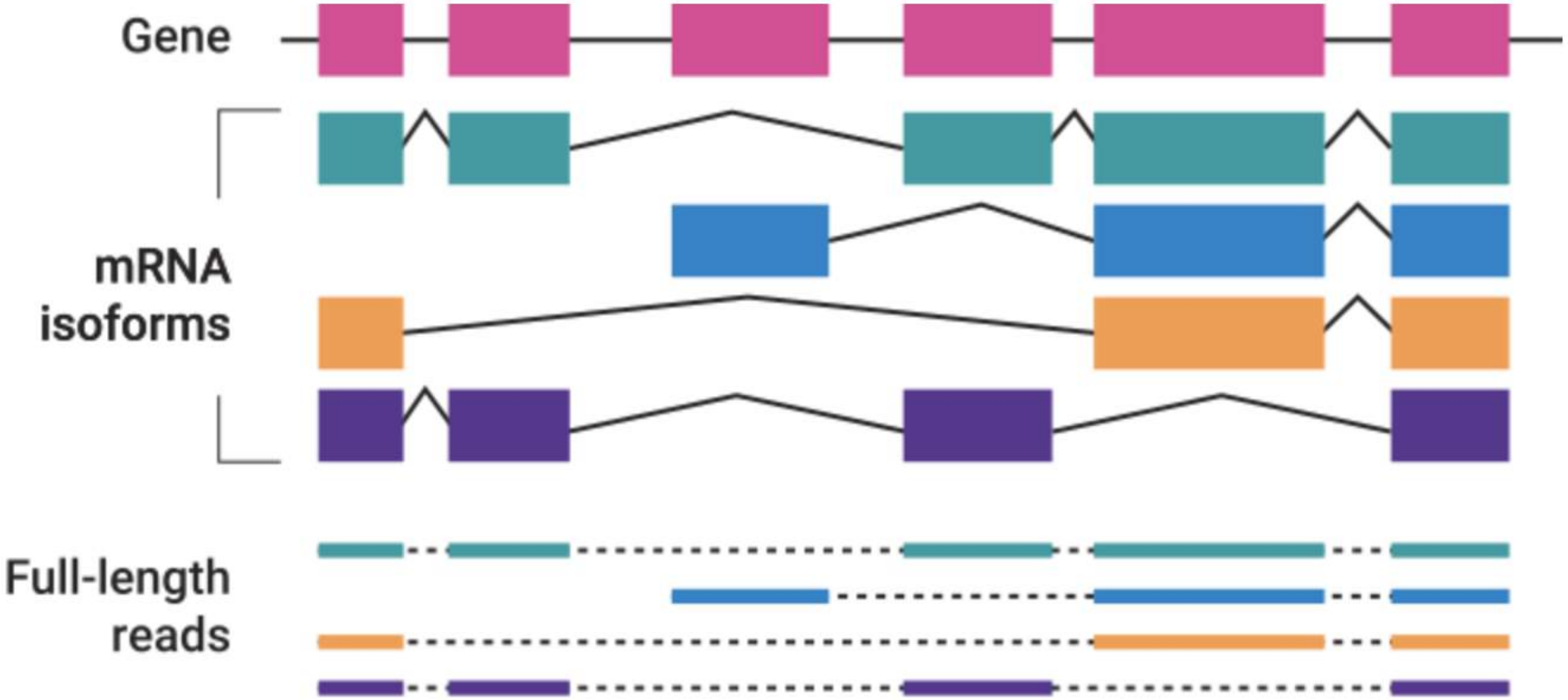
Example 3: Investigate methylation

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



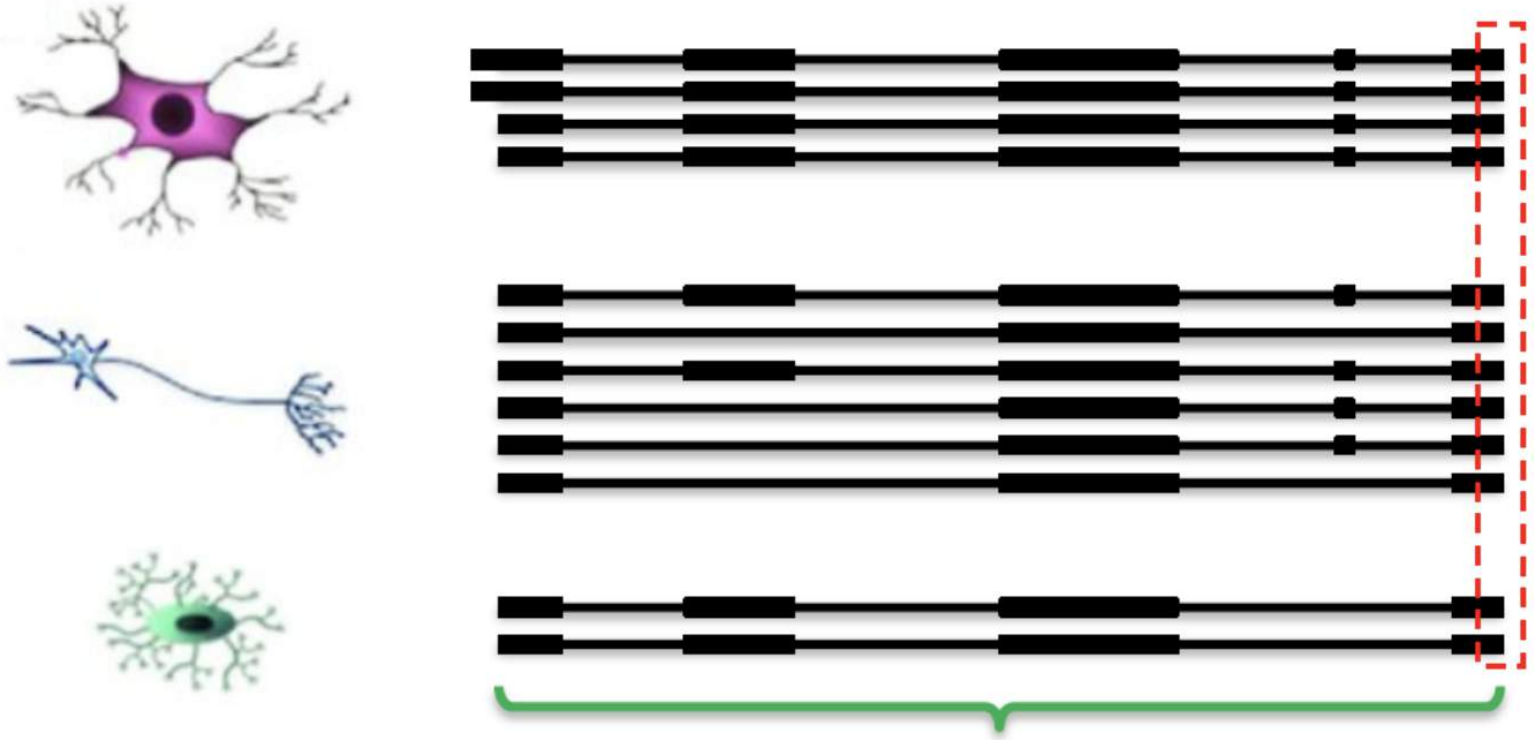
Example 4: Full-length RNA sequencing

- Get complete information about RNA molecules!

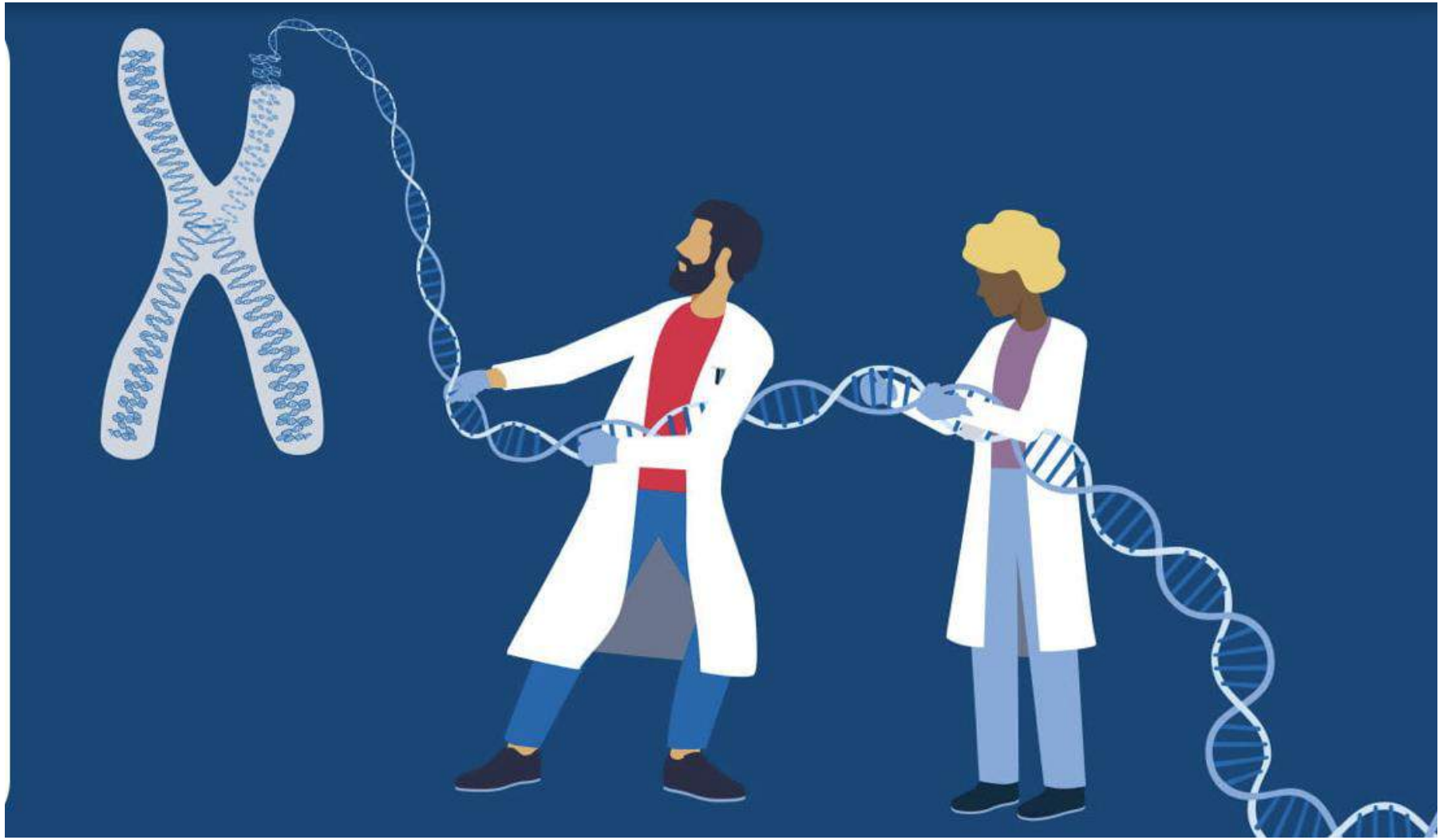


Example 5: Single-cell long-read RNA

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



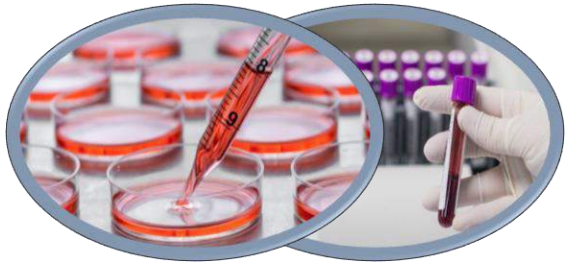
Challenge: good sample quality required!



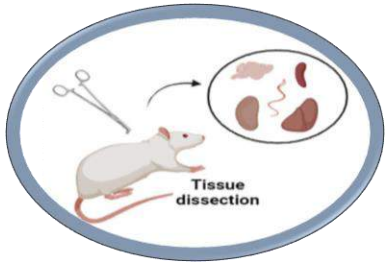
https://www.qiagen.com/ja-us/applications/molecular-biology-research/hmw-dna

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

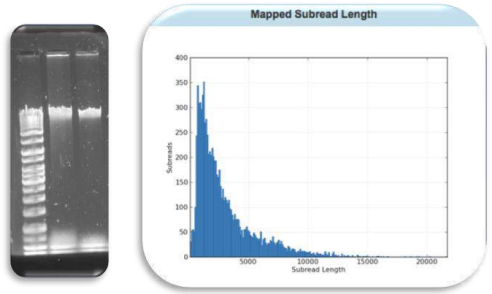
CTAB Lysis
High polysaccharide
Also handles polyphenols

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

HMW-DNA Extraction – Contaminants

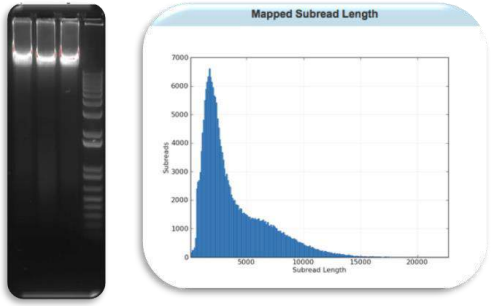
Importance of purity – even for model organisms

Impurities can originate from both host tissue and extraction chemicals.



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?

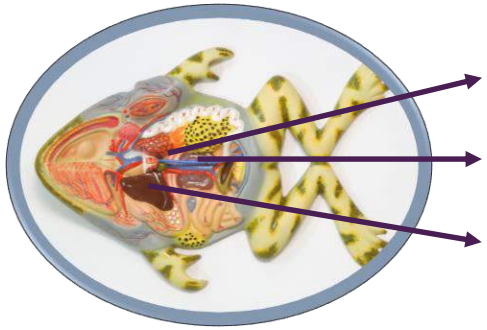
Which would you expect to have less contaminants?

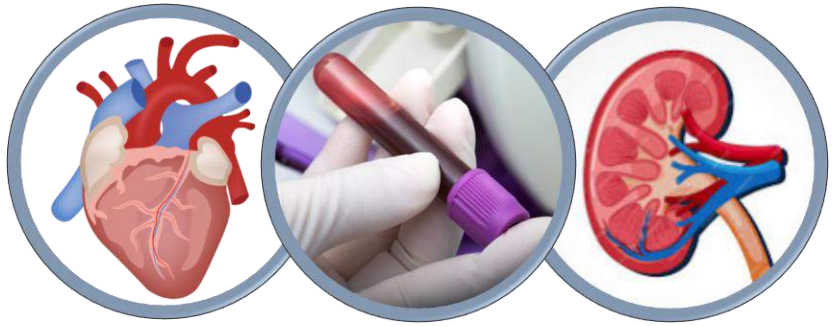
VS

VS

HMW-DNA Extraction – Best Options

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





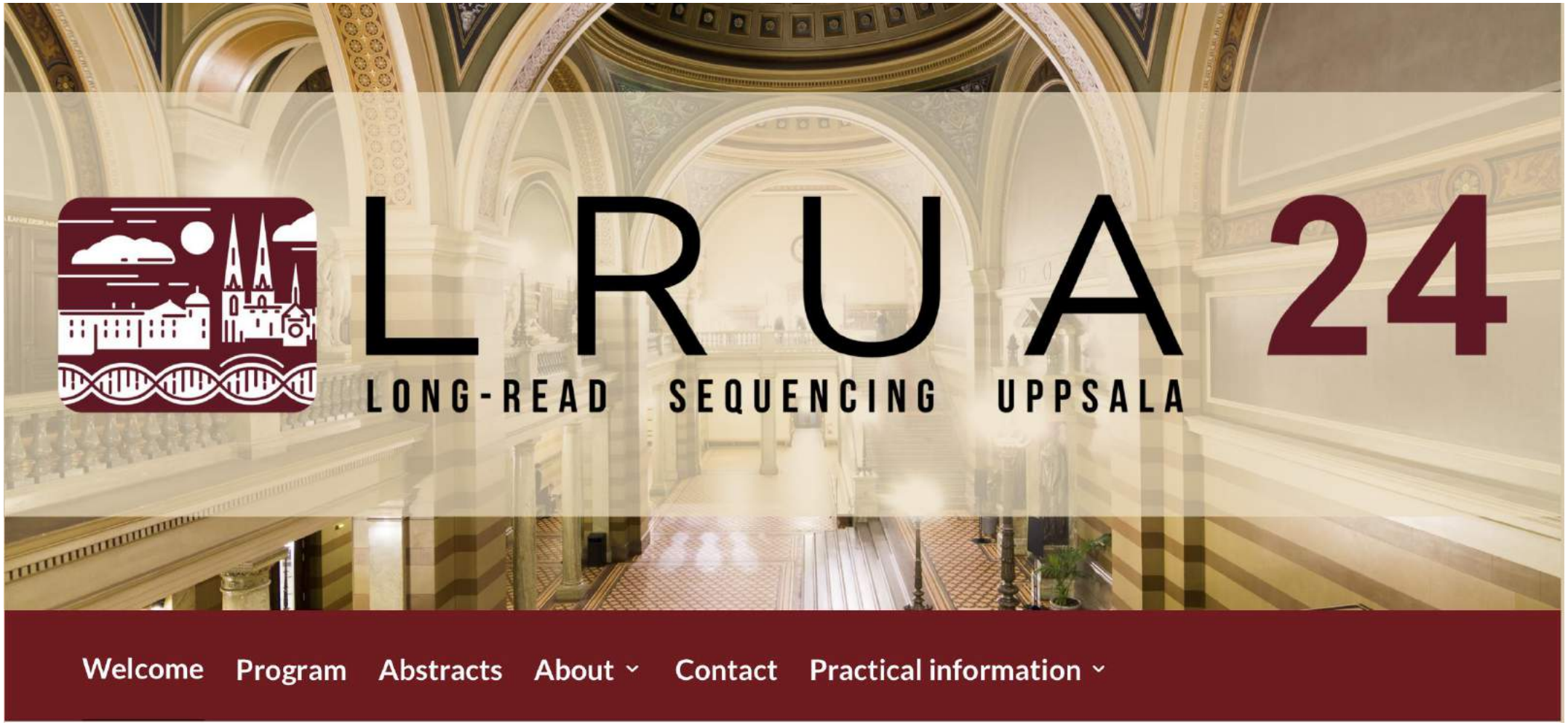
- Choose tissue high in DNA and low in contaminants when possible

- Freeze as fast and cold as possible to minimize fragmentation

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

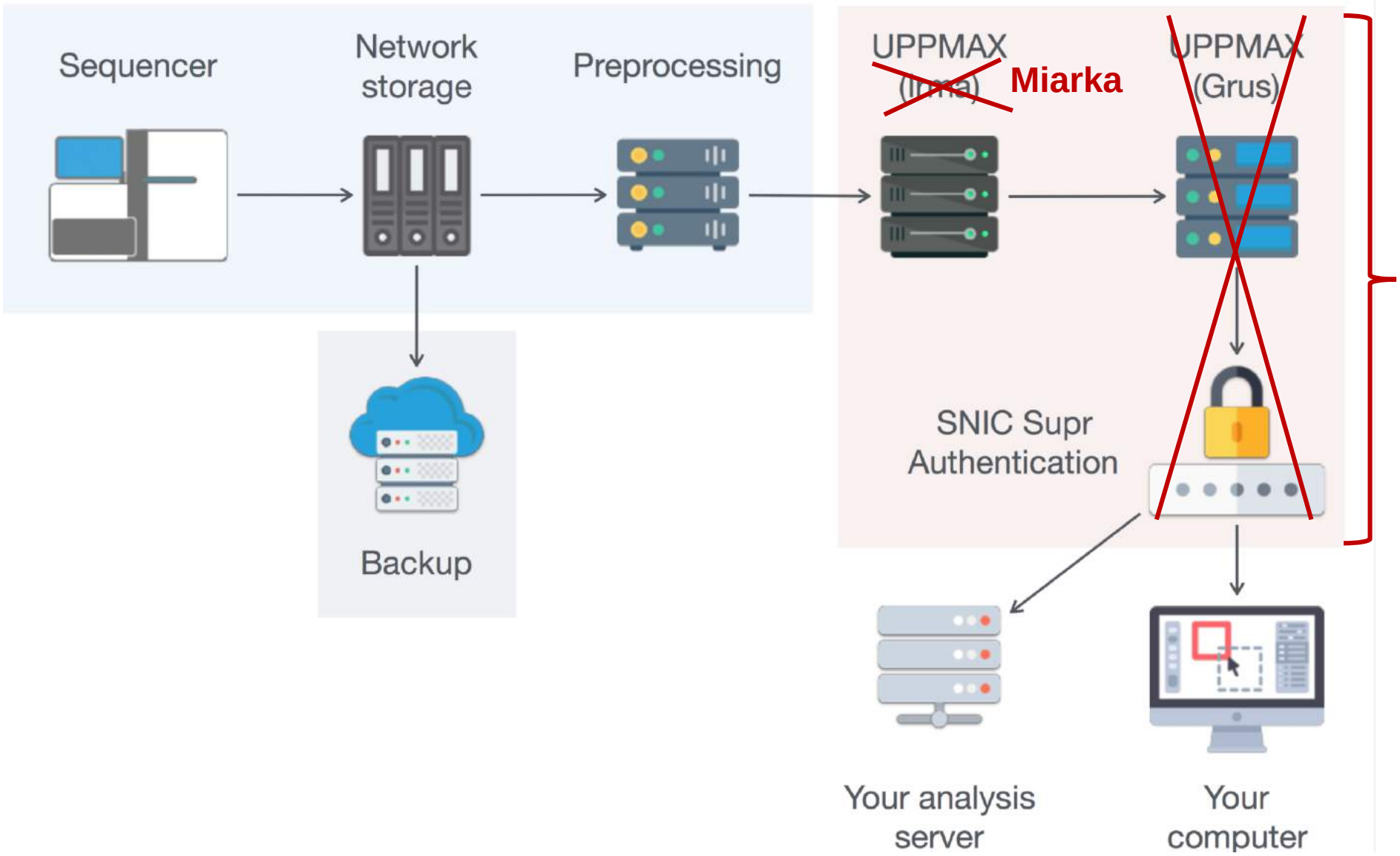
Long-read Uppsala Meeting 2024!

- October 21-23 2024, more information at: www.lrua2024.se



NGI Data Handling and Analysis Pipelines

NGI Data Handling



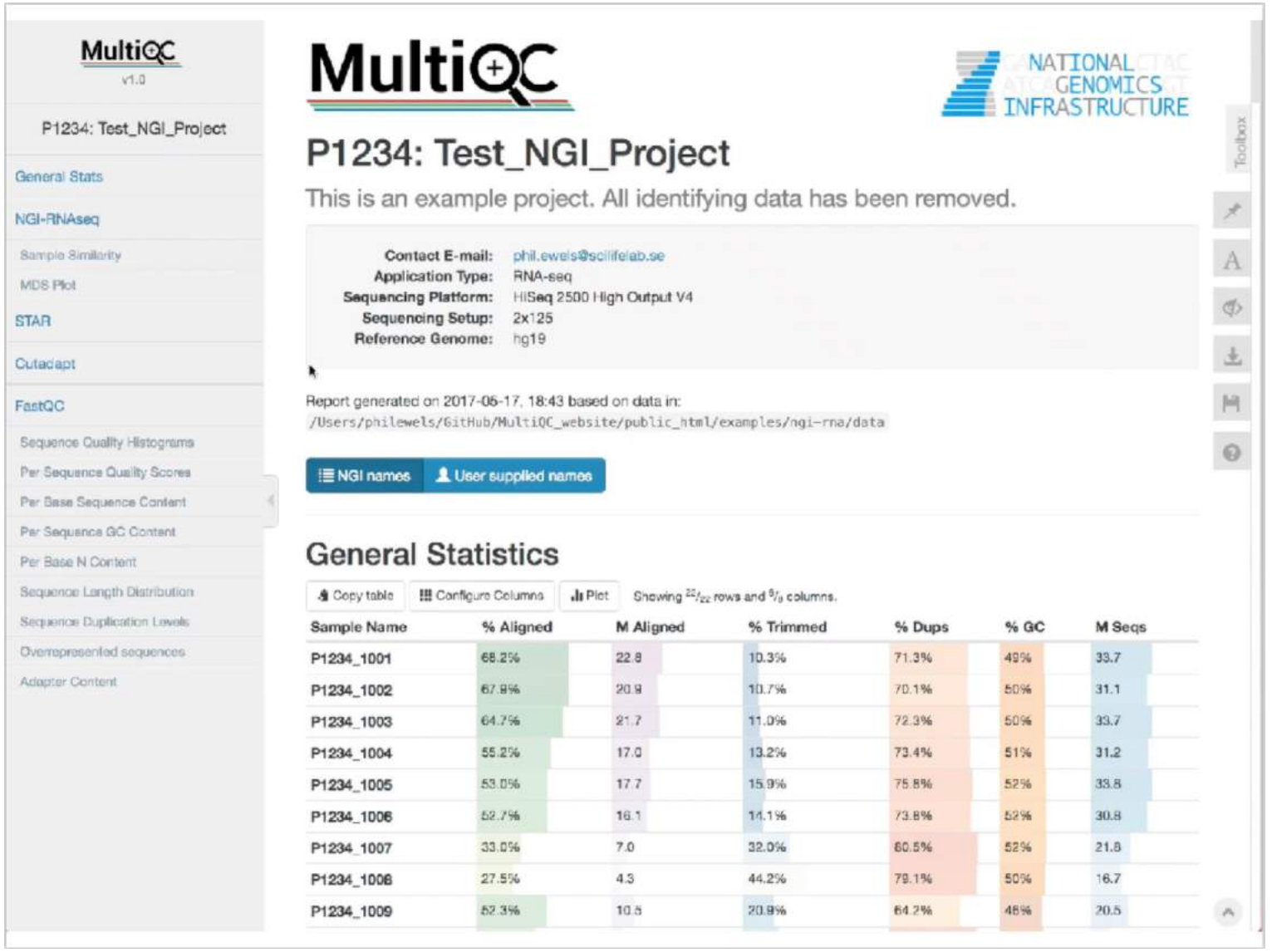
Data delivery via DDS

- 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

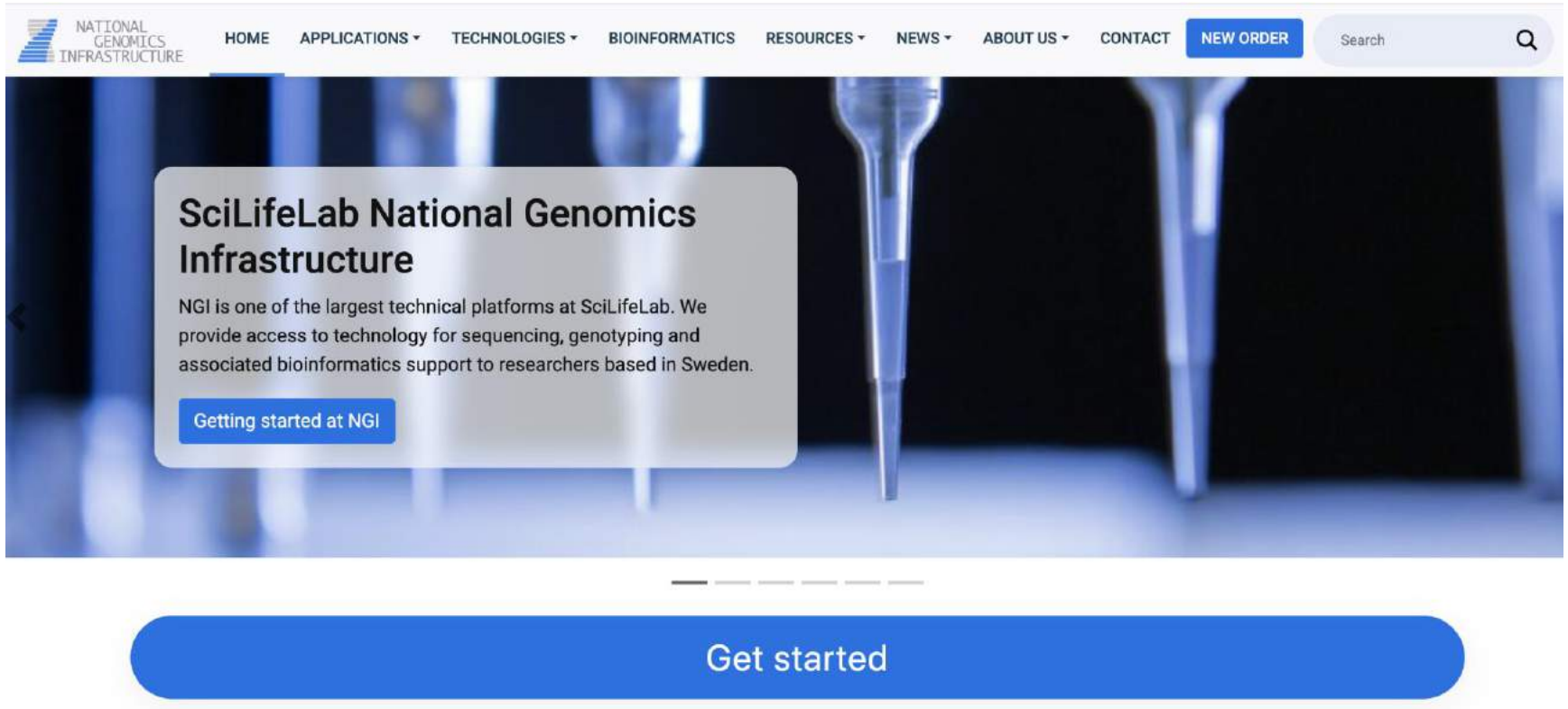
- Every project has some level of quality control checks
  - Technical run performance
  - Read length distribution
  - Sequencing quality
- Analysis pipelines give application-specific QC
- Reporting done using MultiQC (Illumina projects)

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

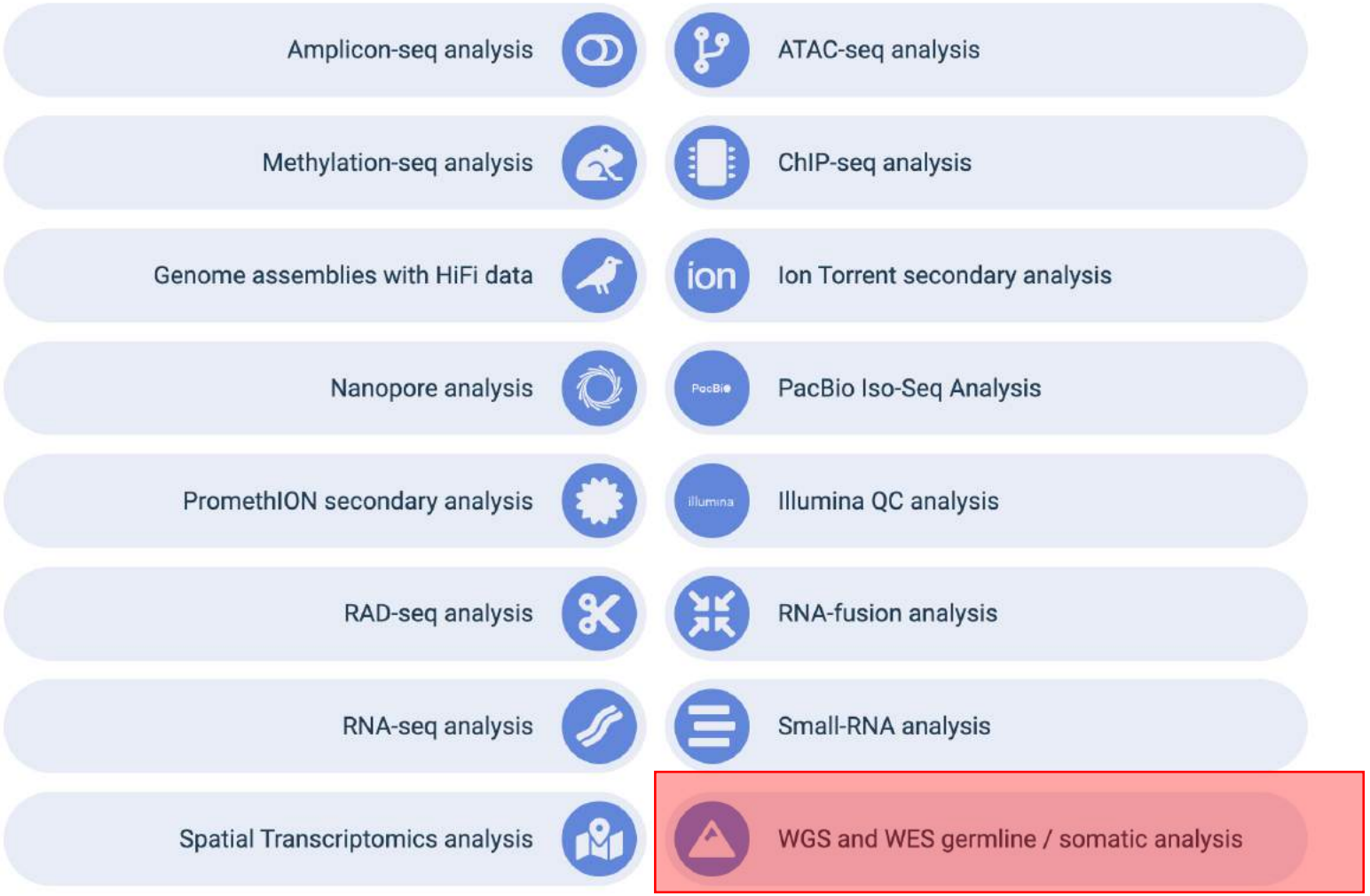
Philip A. Ewels, Alexander Peltzer, Sven Fillinger, Harshil Patel, Johannes Alneberg, Andreas Wilm, Maxime Ulysse Garcia, Paolo Di Tommaso & Sven Nahnsen ✉

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.

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.

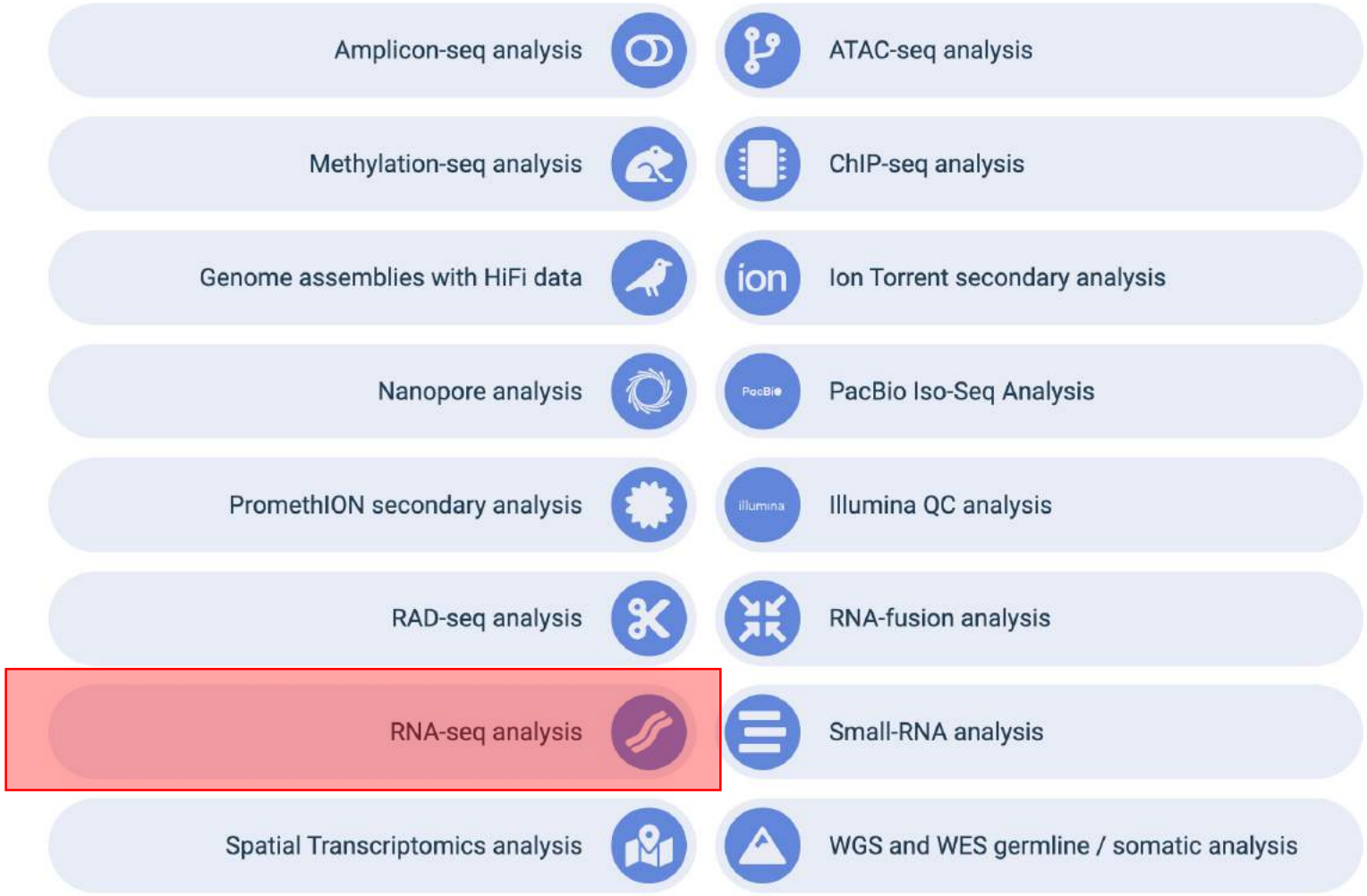
Input data

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.

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

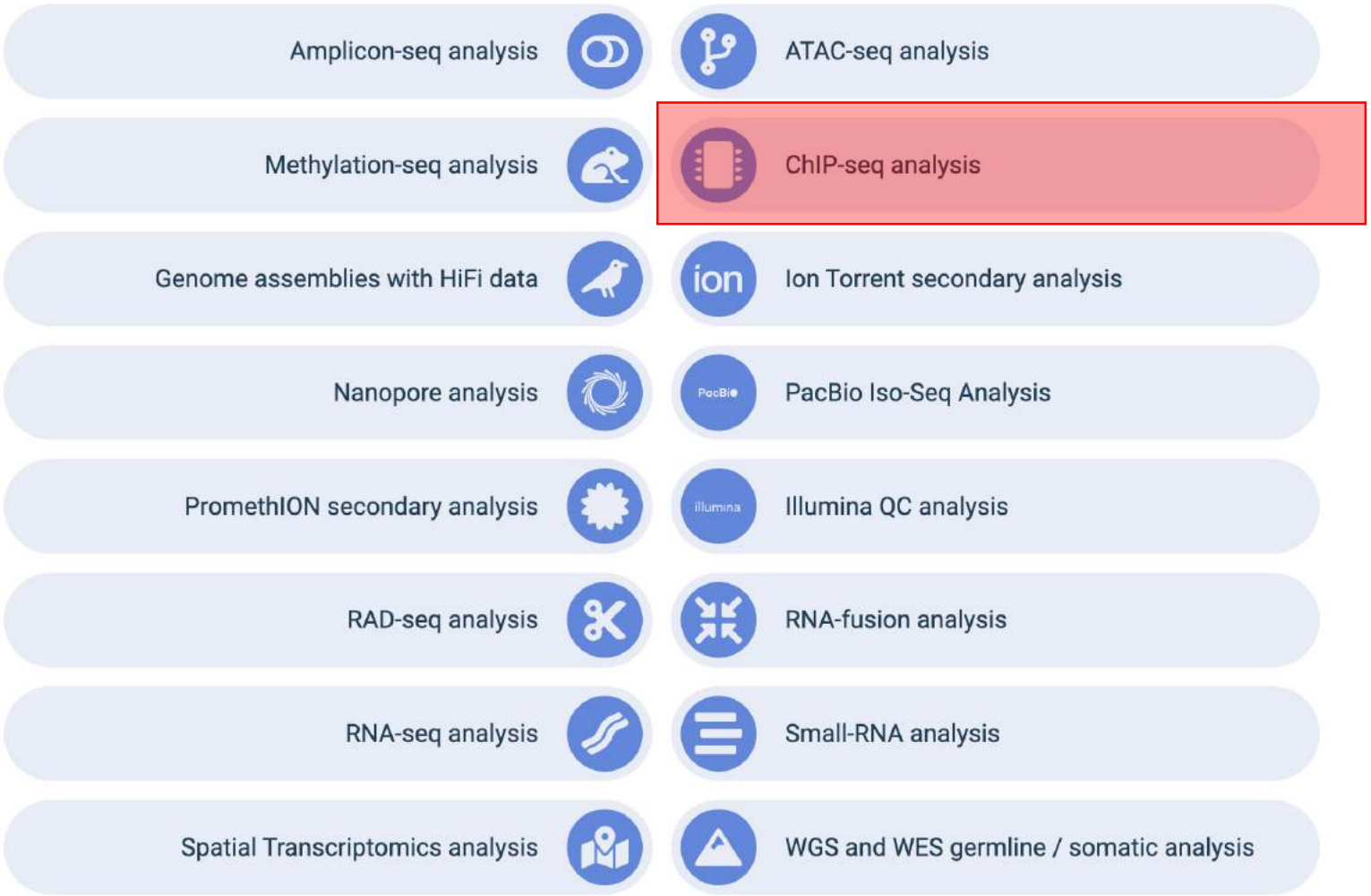
bcl2fastq demultiplexed FastQ files and a genome reference.

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.

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

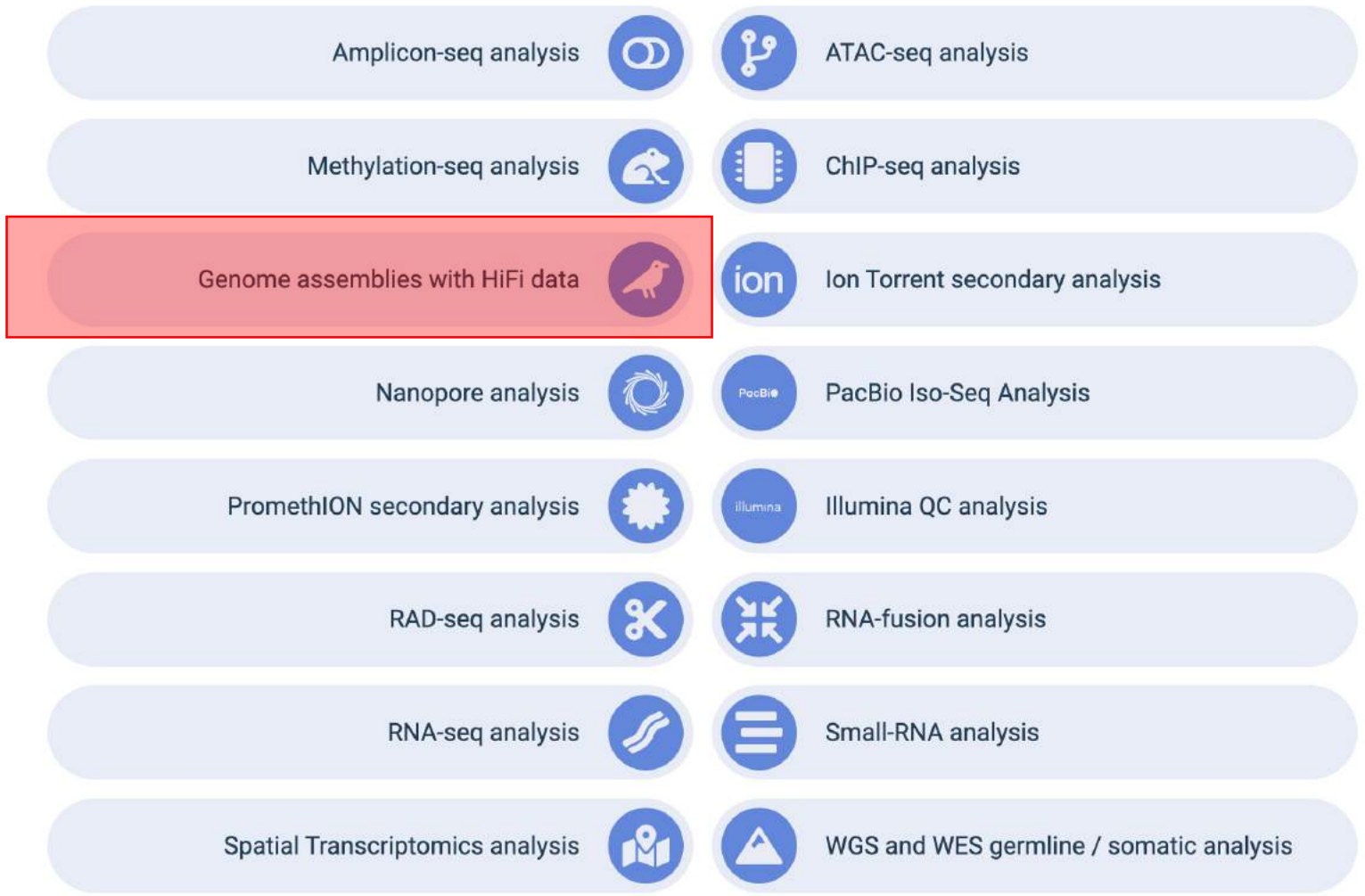
bcl2fastq demultiplexed FastQ files and a genome reference.

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 ASSEMBLY

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.

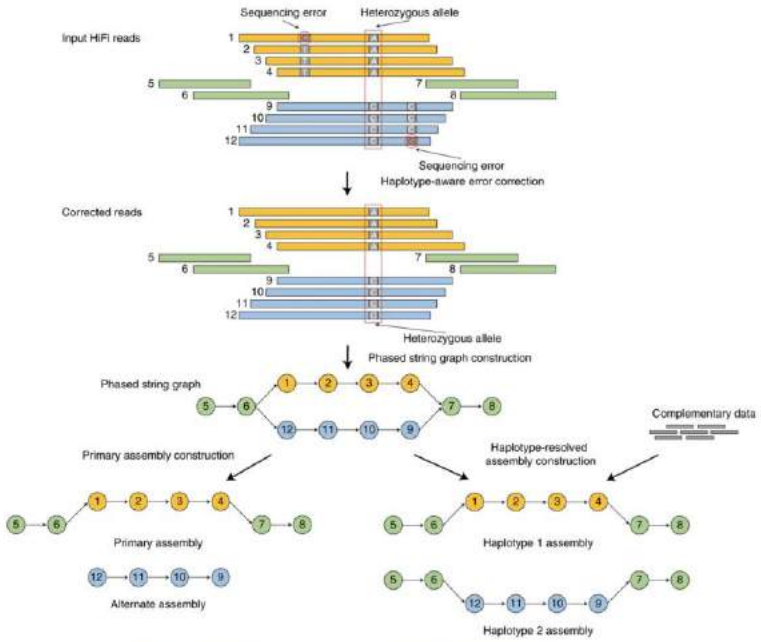


Image: Nat Methods **18,** 170–175 (2021). https://doi.org/10.1038/s41592-020-01056-5

Not yet implemented as a nf-core pipeline!

Trend: On-instrument analysis

More and more analyses being done on instrument GPUs

Illumina NovaSeqX

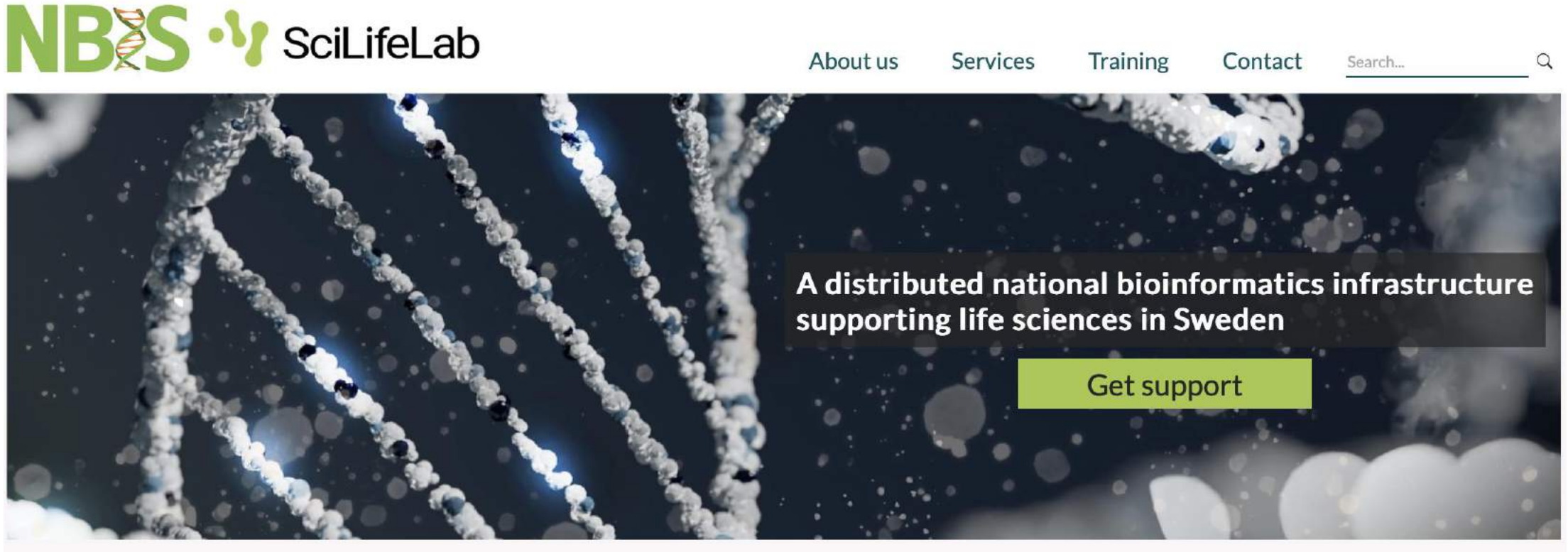
Mapping and variant calling (Dragen)

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!

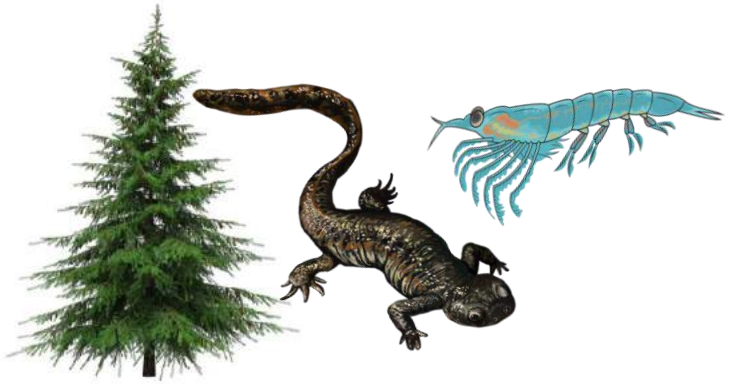
- Which tools should be used?
- Can I run the analysis myself, or do I need assistance?
- Where should the analysis be run?
- Do I have enough storage space?
- Where should the data eventually be archived?

NGI strategic projects and collaborations

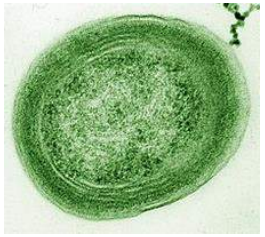
We are involved in some larger national and international projects...

Biodiversity genomics

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



Large genomes (18-22Gb)



Tiny organismsTiny organisms
with large genomes

EARTH BIOGENOME PROJECT
sequencing life for the future of life

eRGA
EUROPEAN REFERENCE GENOME ATLAS

Biodiversity Genomics Europe

Darwin TREE of LIFE

wellcome sanger institute

MAX-PLANCK-GESELLSCHAFT

UNIVERSITAS OSLOENSIS MDCCCXI

UiO : University of Oslo

FLORENTINA STUDIORUM UNIVERSITAS

GENOSCOPE
Centre National de Séquençage

cnag
centre nacional d'anàlisi genòmica
centro nacional de análisis genómico

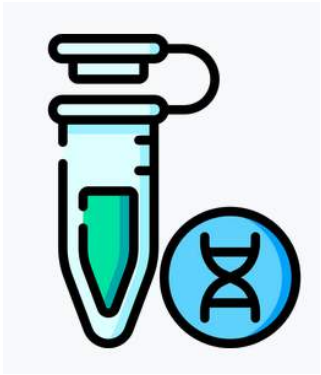
CRG
Centre for Genomic Regulation

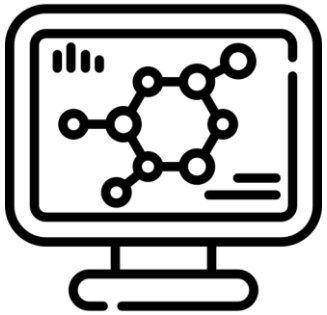
And many, many, …, many other uncooperative

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

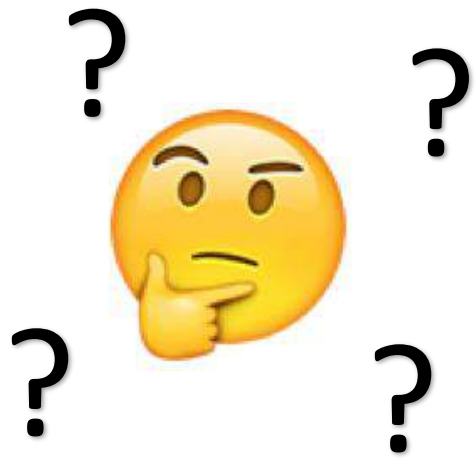
Photo: SVT

Human population-level sequencing

2017: 1,000 genomes sequenced on Illumina

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

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



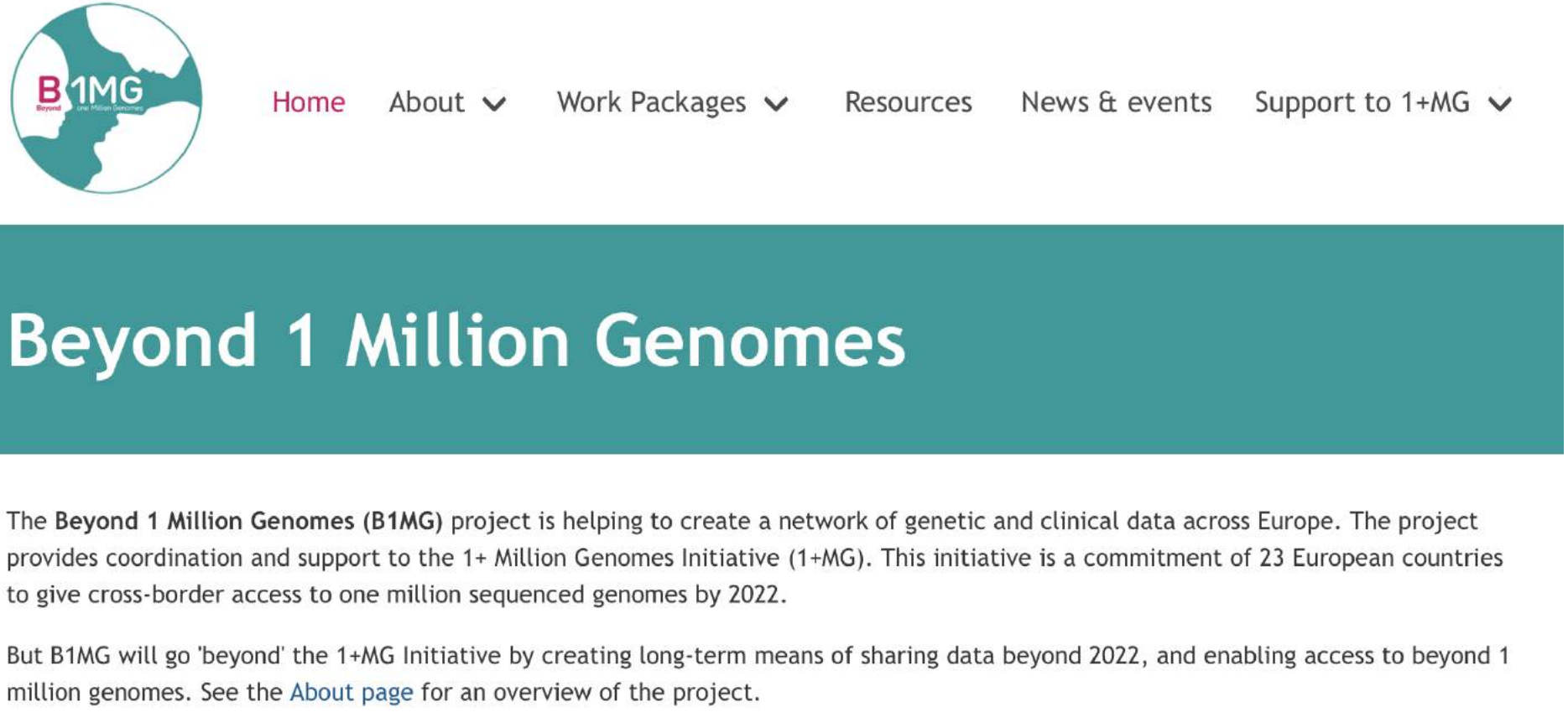
How to build a long-read reference dataset

- Wishlist for a new Swedish population cohort (still in early planning)

	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

We hope to do this as part of a EU project

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



Collaborations on Rare Disease

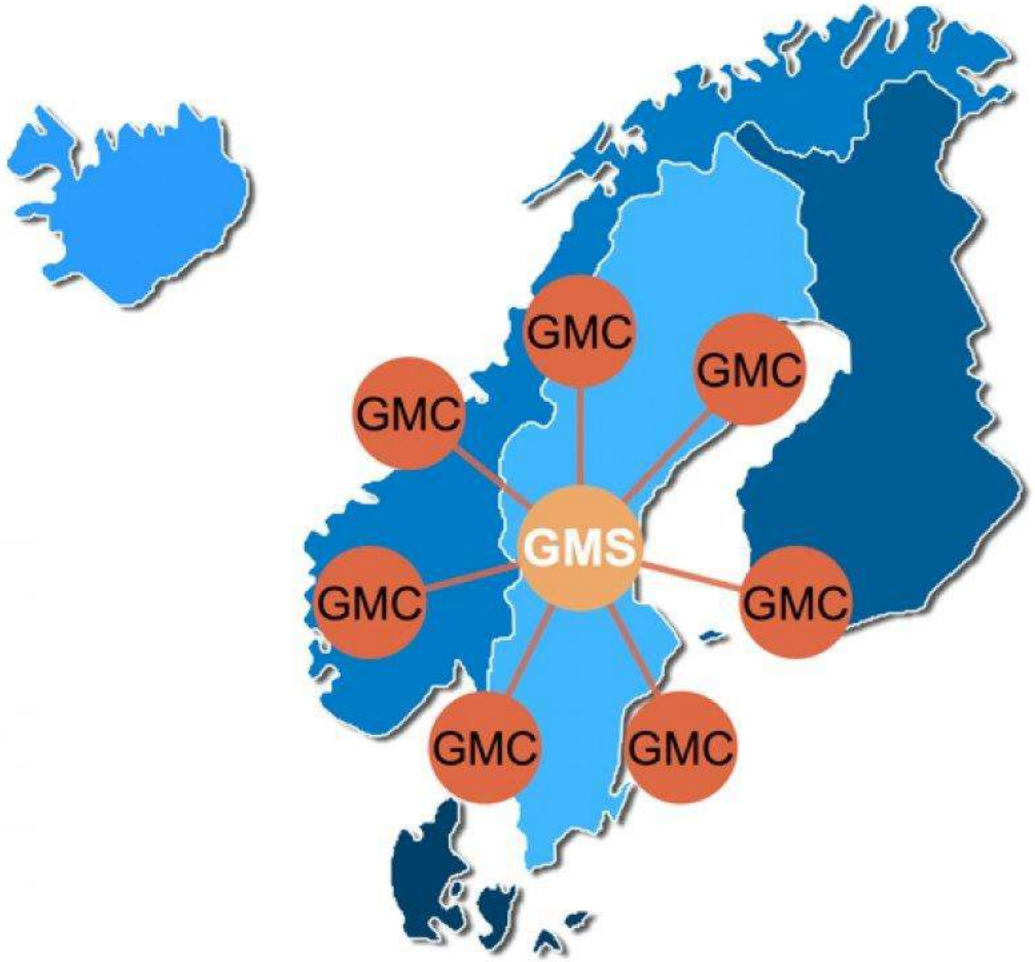
We are collaborating with Genomic Medicine Sweden - Rare Disease Group

Long-Read Whole Genome Sequencing

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

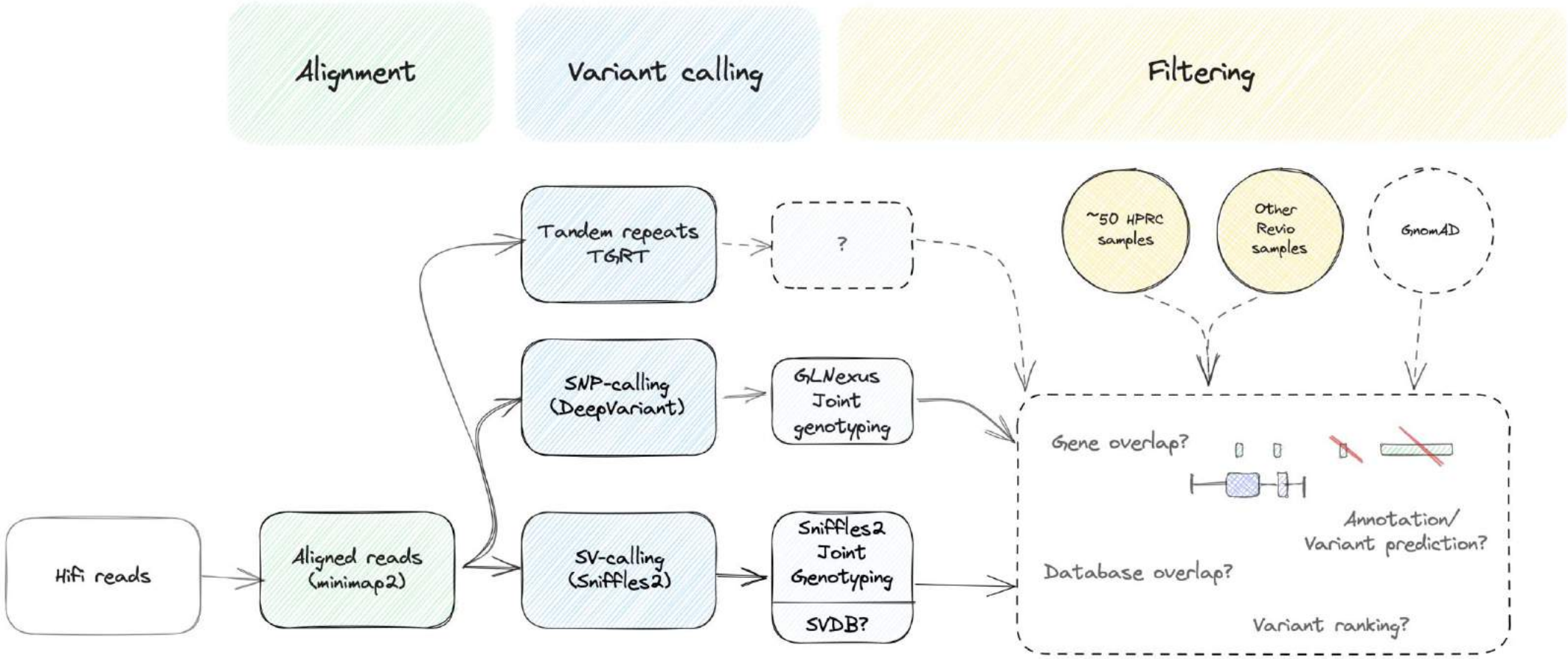
Long-Read Targeted Sequencing

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



Analysis pipelines human WGS

- New pipelines needed for human long-read sequencing



Felix Lenner
github.com/fellen31/skierfe

Thanks for your attention!