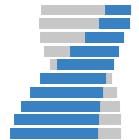


# NGI Sweden

Next Generation Sequencing at the  
National Genomics Infrastructure

**SciLifeLab**

 **NGI** stockholm

Phil Ewels

[phil.ewels@scilifelab.se](mailto:phil.ewels@scilifelab.se)

Introduction to Bioinformatics Using NGS Data

Linköping, 2018-05-23

# — Overview

National Genomics Infrastructure

Sequencing Technologies

Sequencing Applications

Bioinformatics at the NGI

# The National Genomics Infrastructure



# SciLifeLab NGI

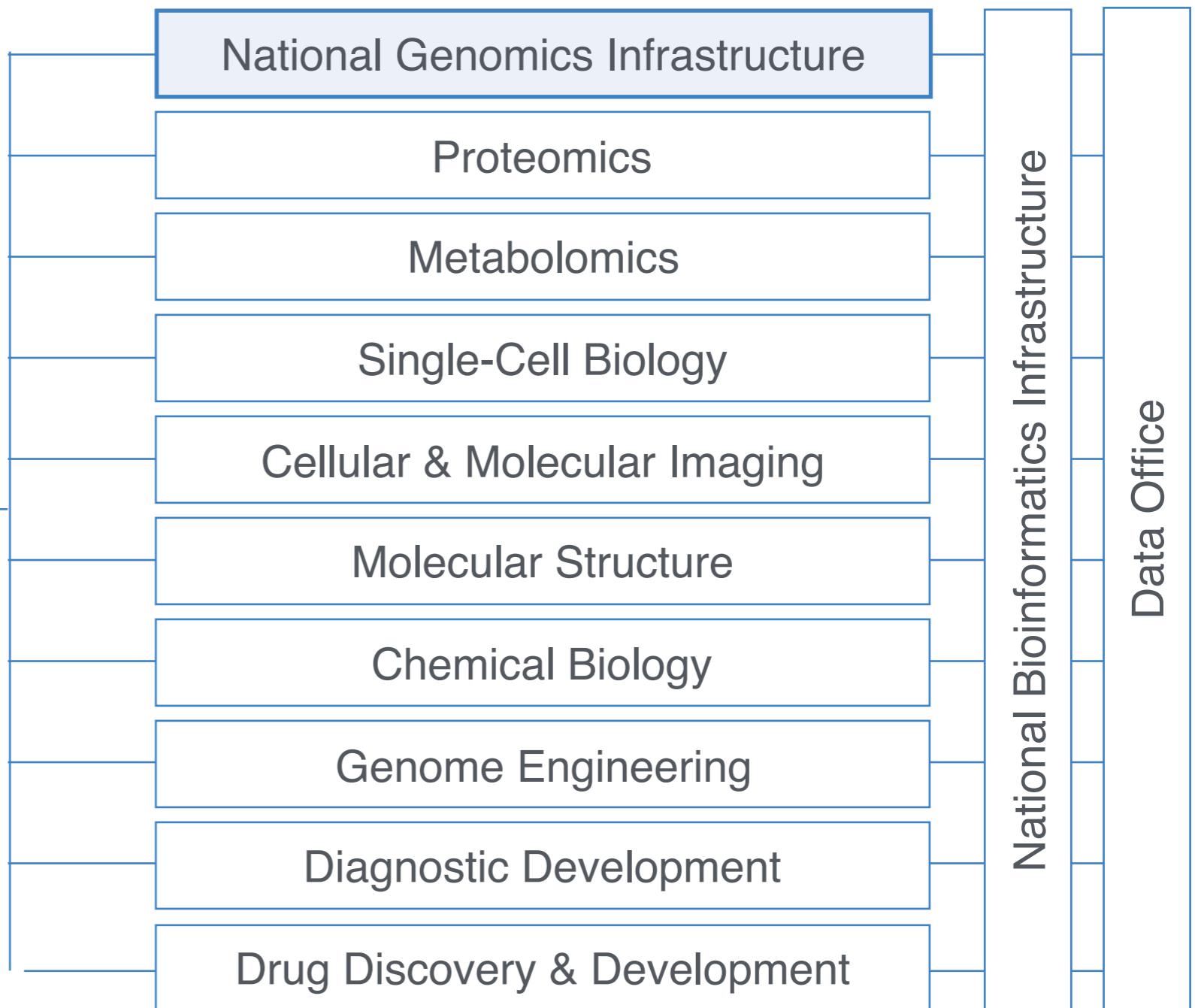


Research Programs

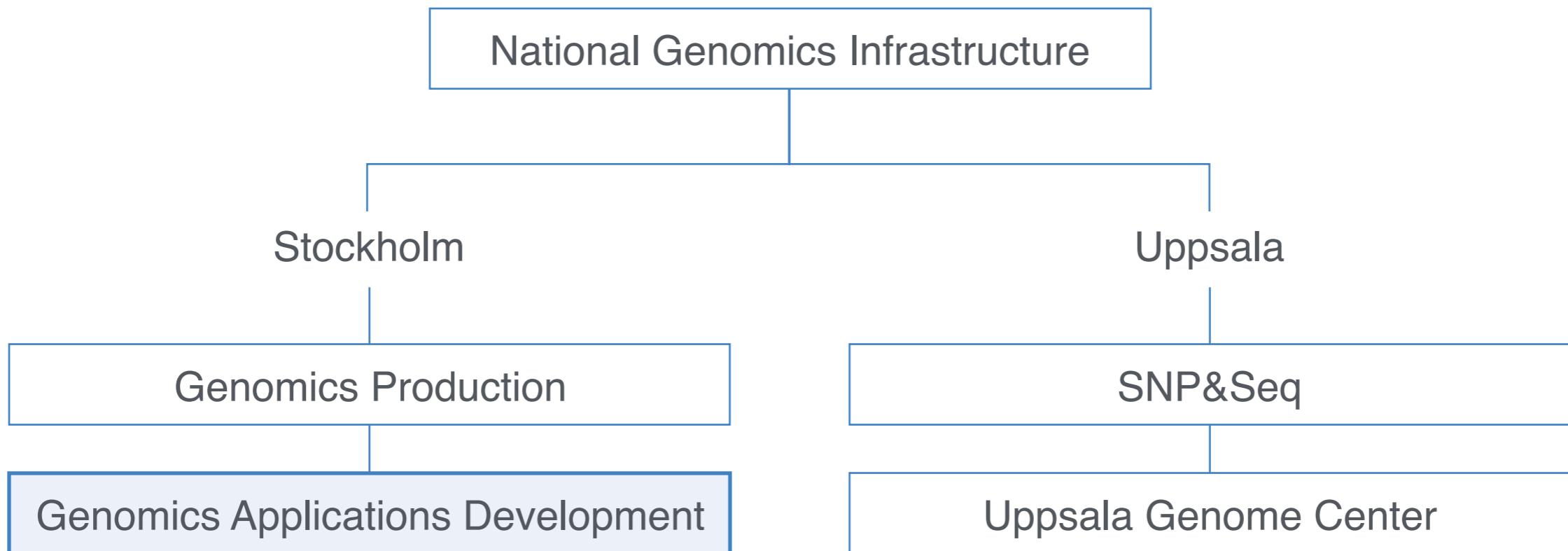
Technology Platforms

SciLifeLab

 NGI stockholm



# SciLifeLab NGI

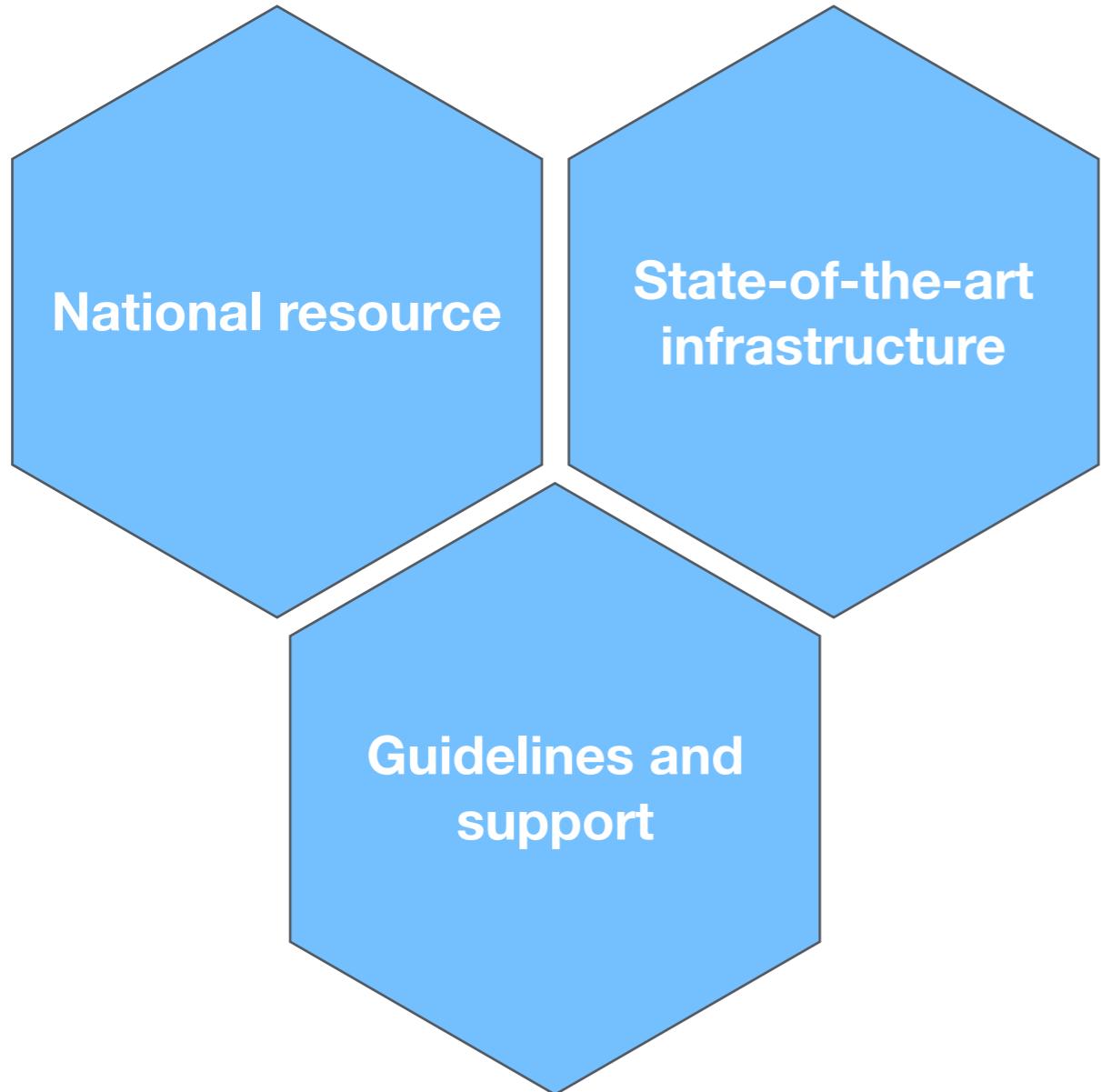


# SciLifeLab NGI



Our mission is to offer a  
**state-of-the-art infrastructure**  
for massively parallel DNA sequencing  
and SNP genotyping, available to  
researchers all over Sweden

# SciLifeLab NGI



We provide  
**guidelines and support**  
for sample collection, study  
design, protocol selection and  
bioinformatics analysis

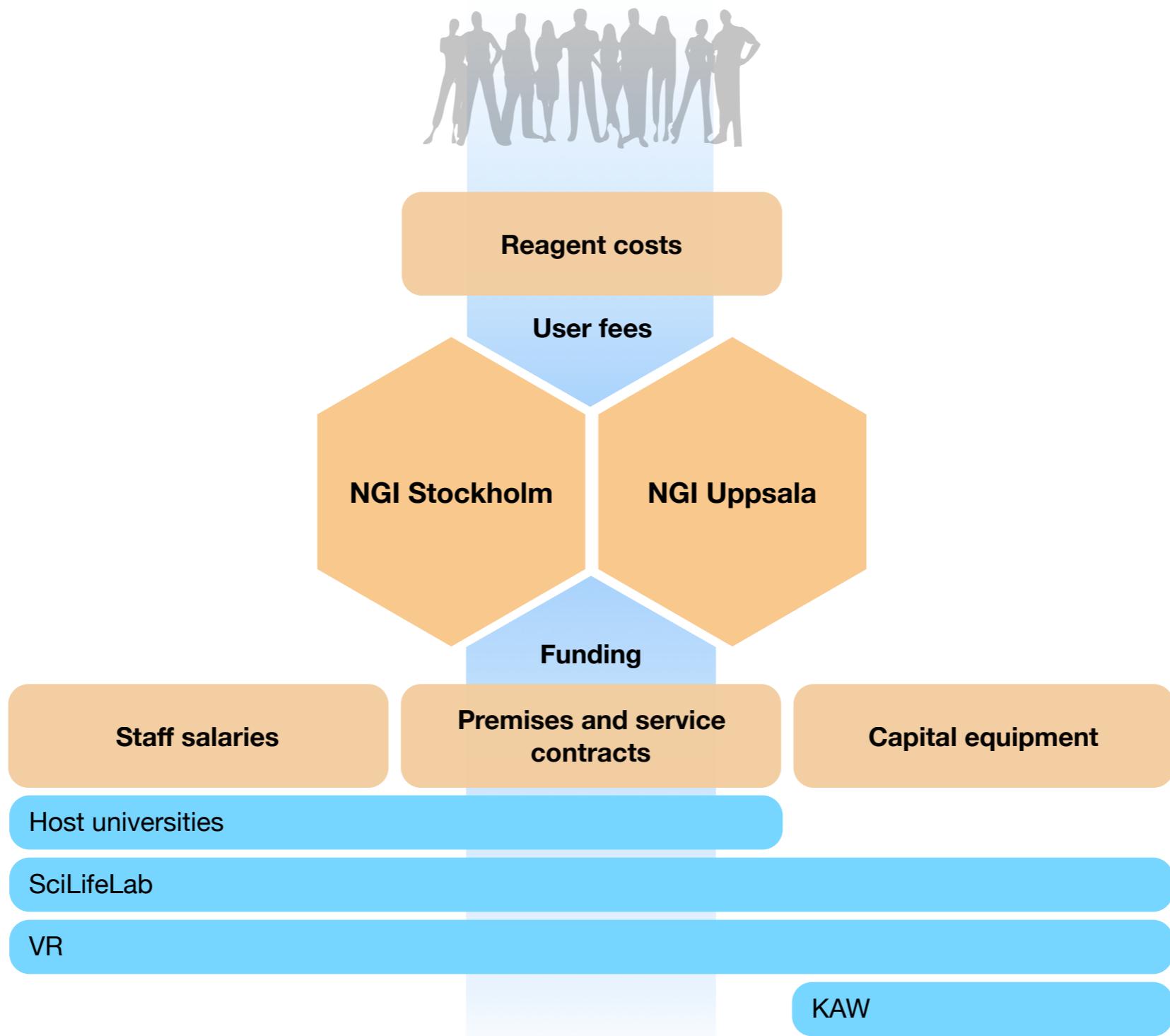
# - NGI Organisation



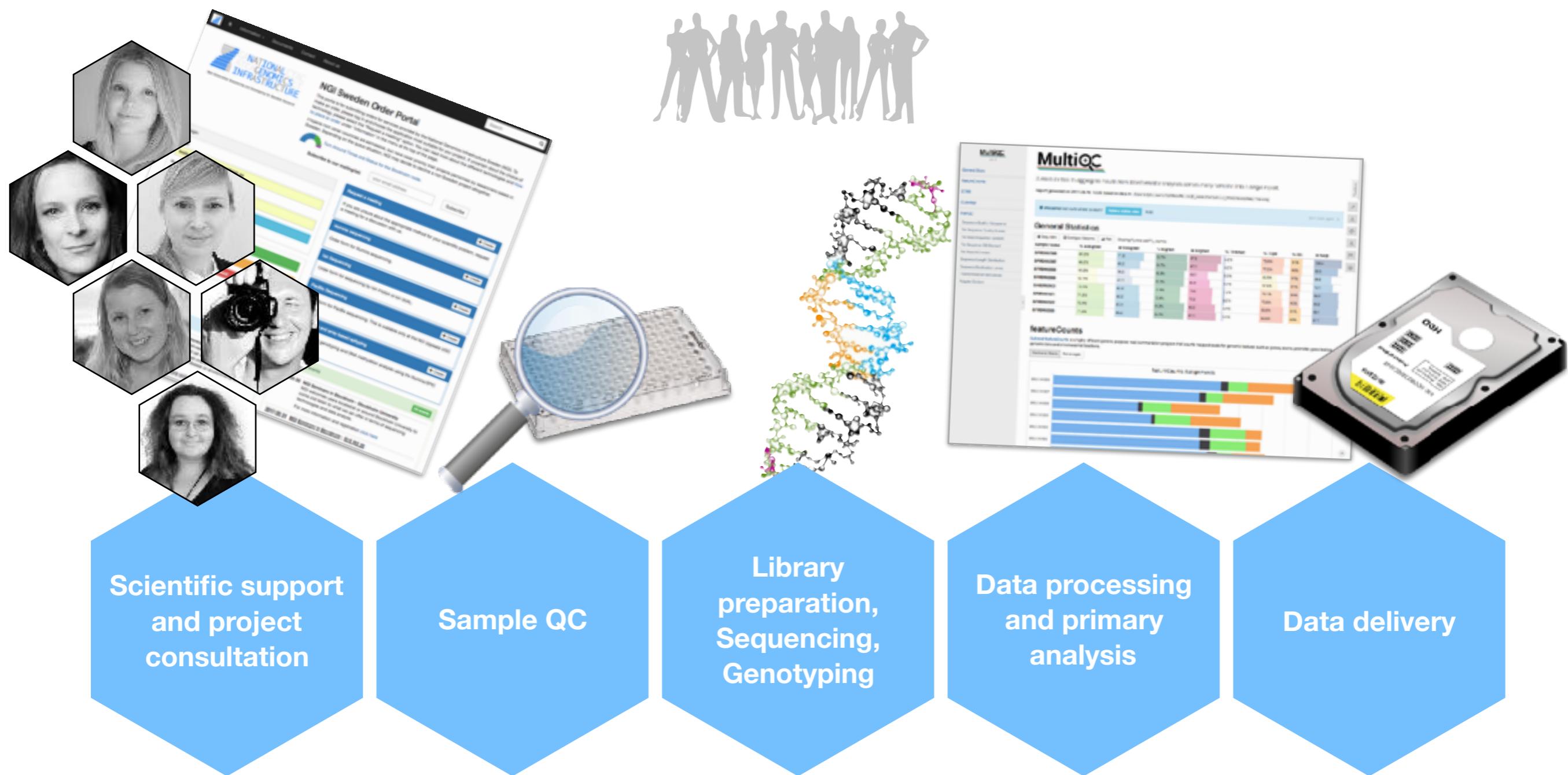
SciLifeLab

 NGI stockholm

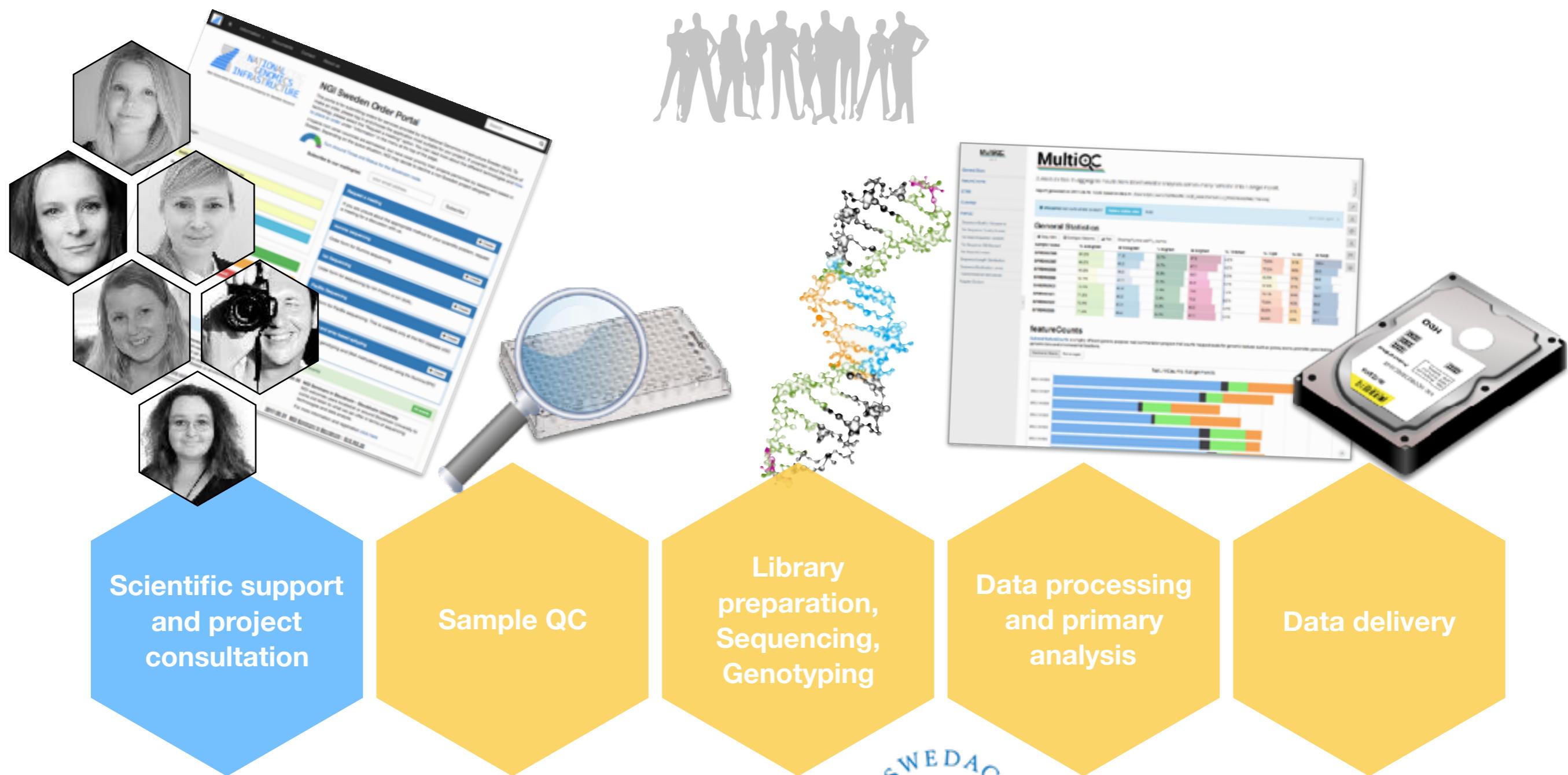
# - NGI Organisation



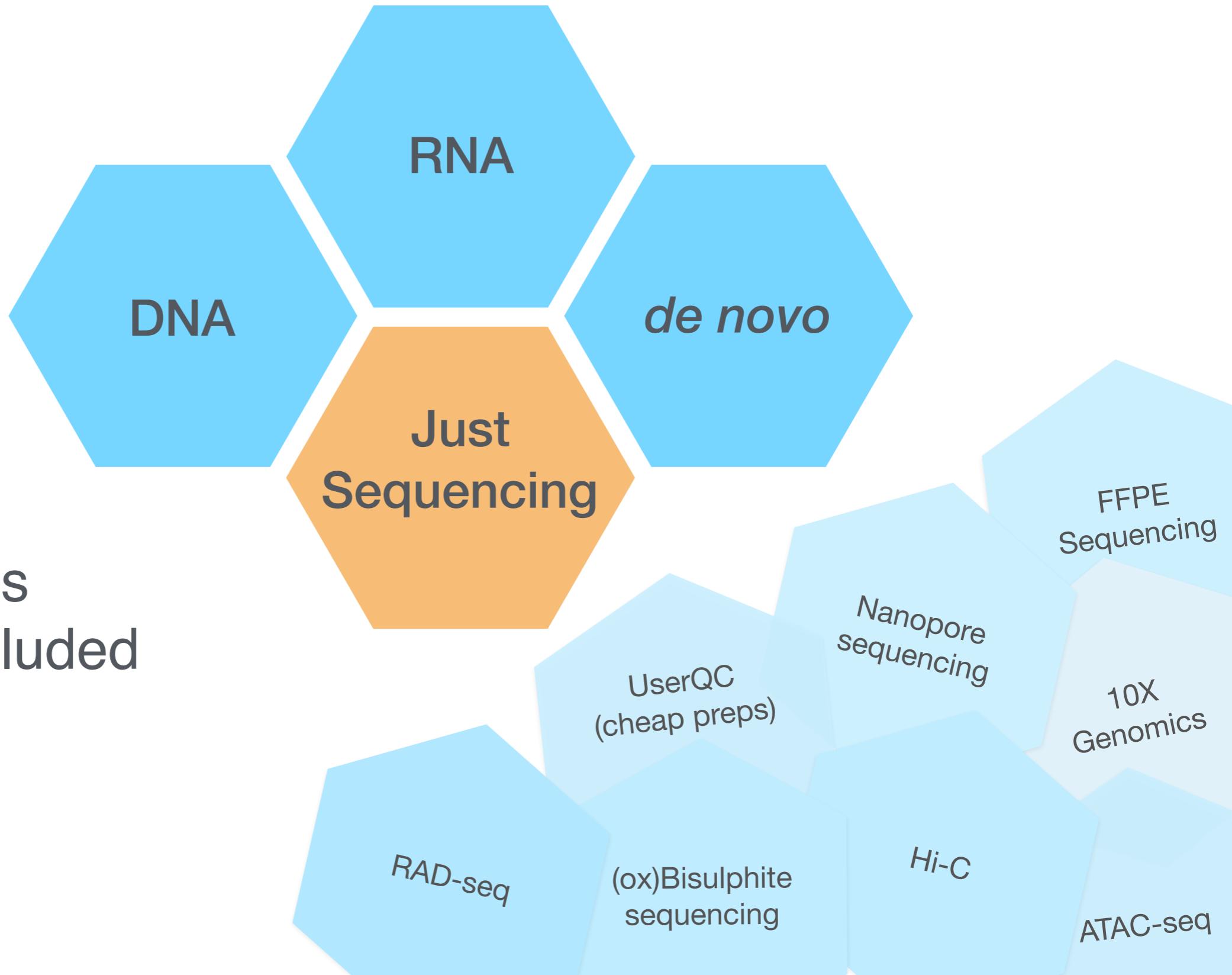
# Project timeline



# Project timeline

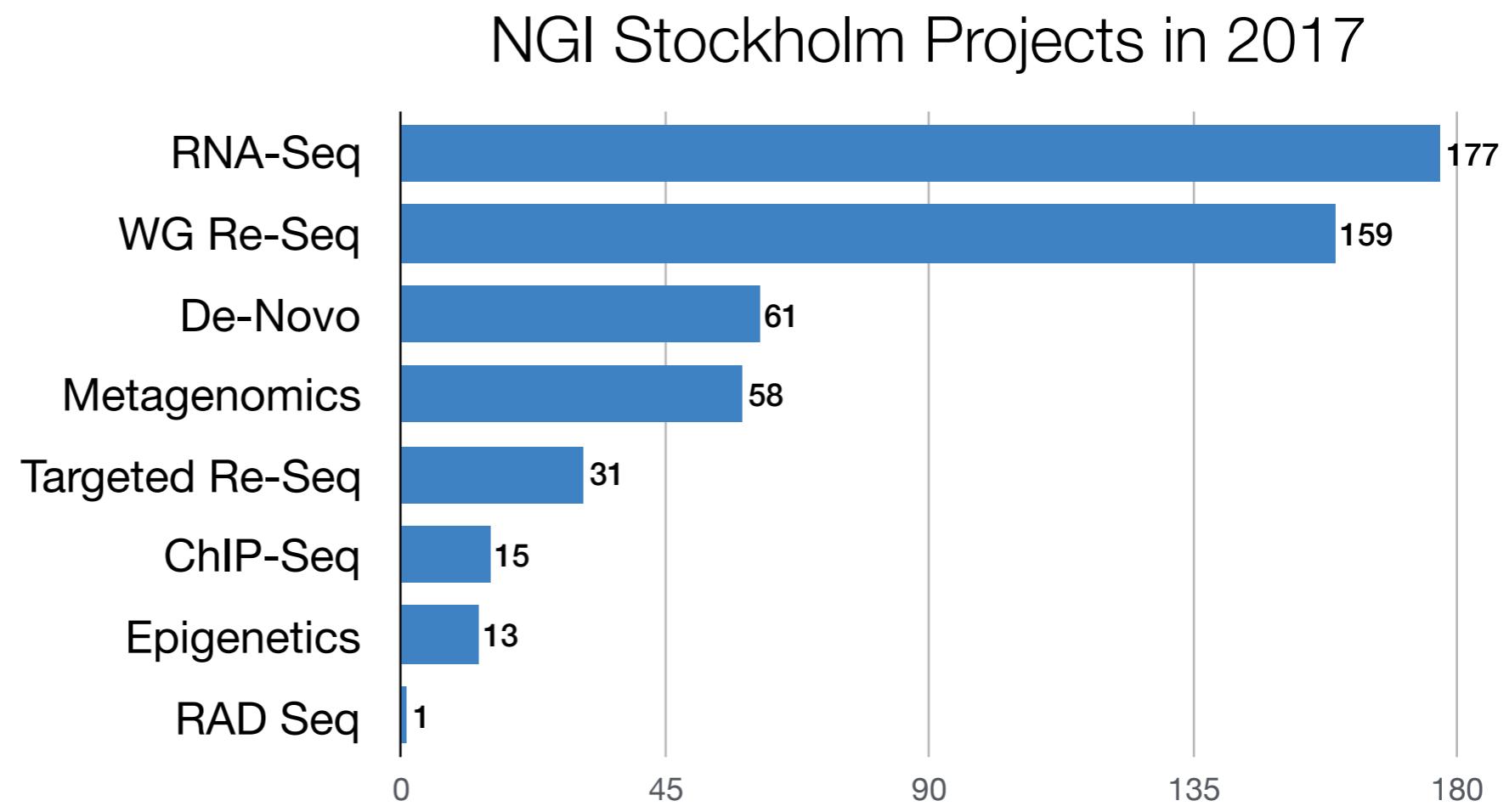


# Methods offered at NGI



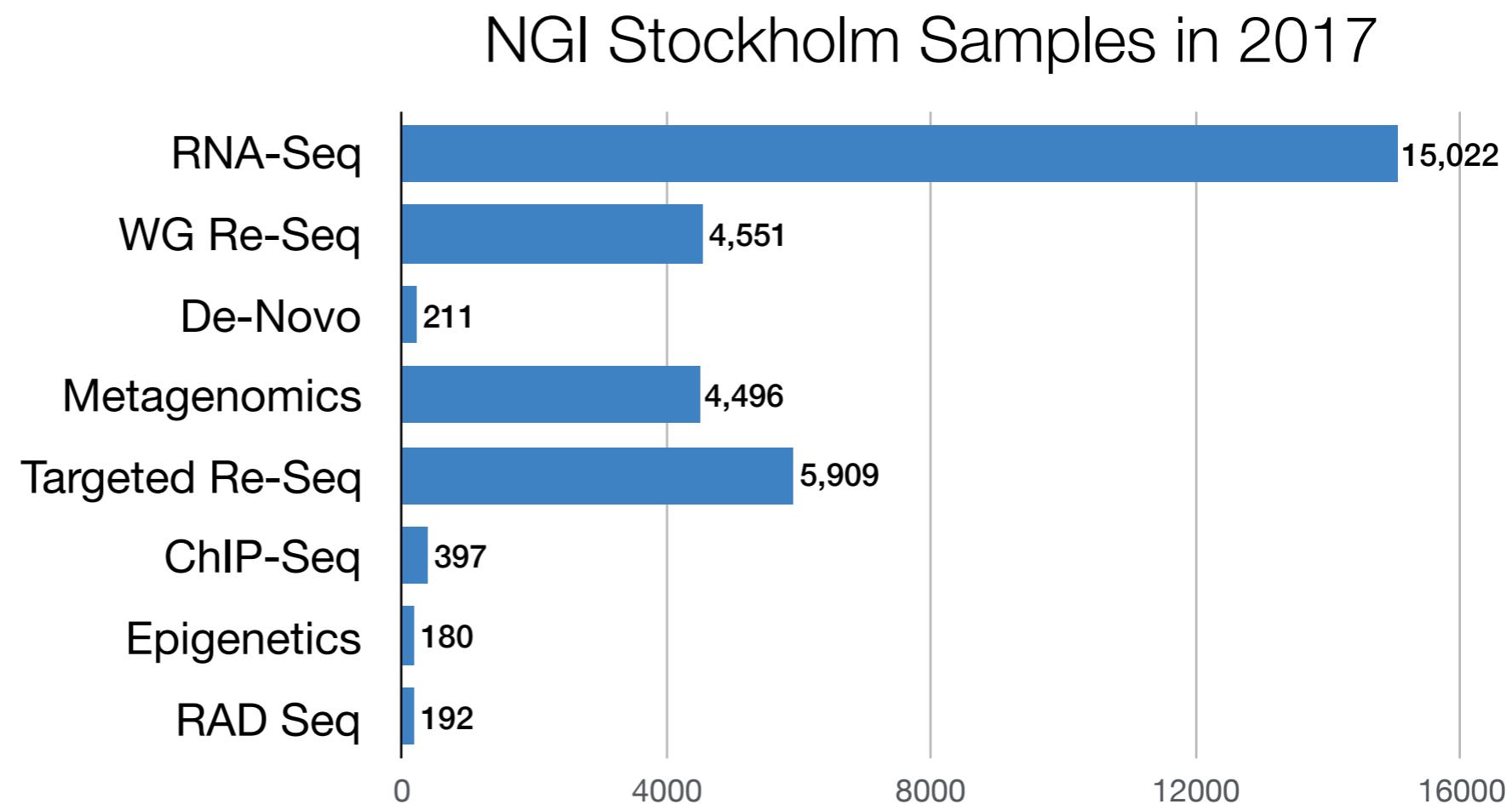
# - NGI Stockholm

- RNA-seq is the most common project type



# NGI Stockholm

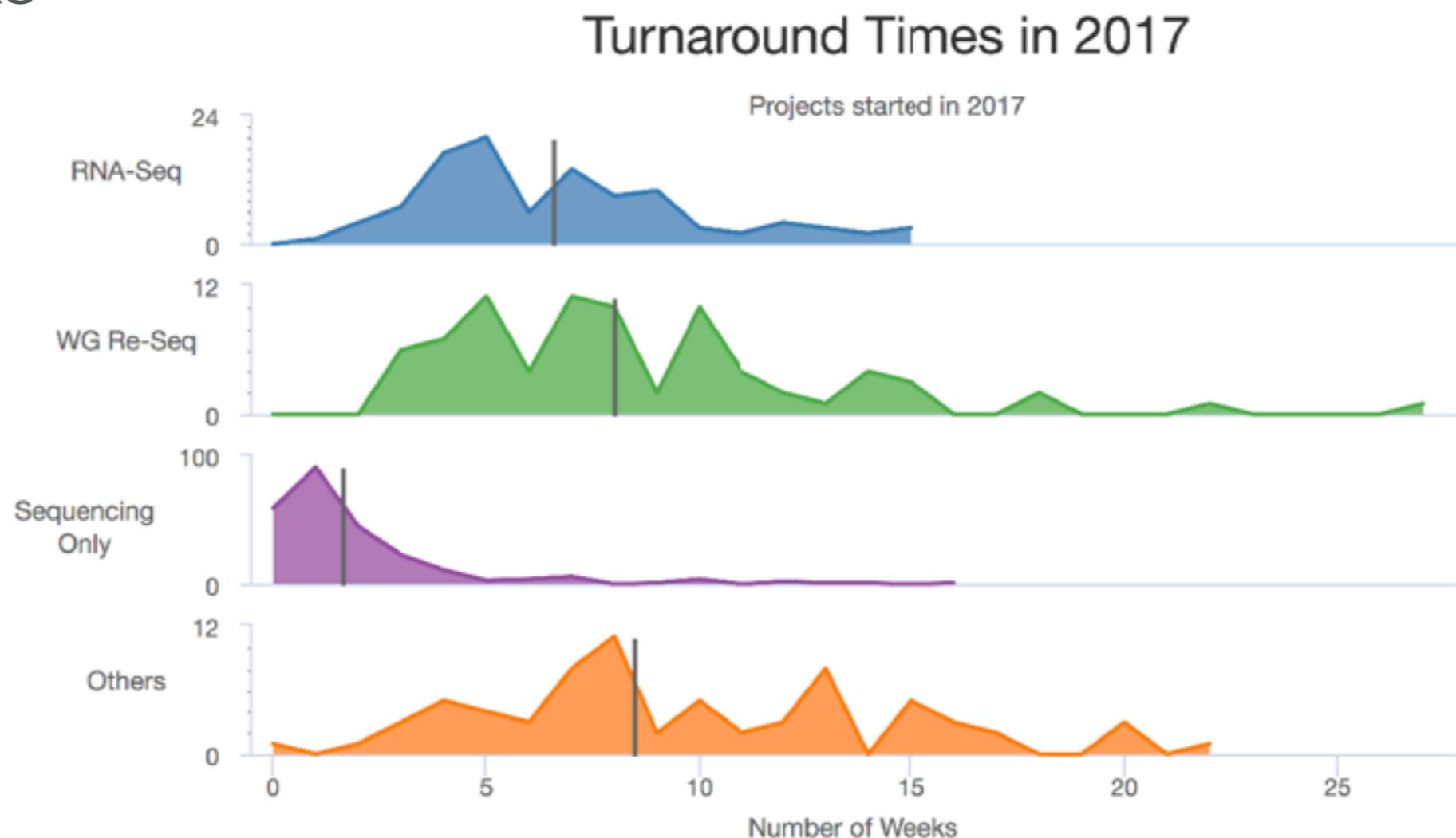
- RNA-seq is the most common project type
- In total, NGI Sweden processed 1068 NGS projects with almost 50 000 samples in 2017



# NGI Stockholm

- Median turn around times from QC passed to data delivered for 2017
  - Sequencing only: 11.5 days
  - RNA: 6.5 weeks
  - WGS: 8 weeks

[https://ngisweden.scilifelab.se/  
file/stockholm\\_dashboard](https://ngisweden.scilifelab.se/file/stockholm_dashboard)



# Sequencing Technologies



# — Sequencing Types

Illumina

PacBio

Oxford Nanopore

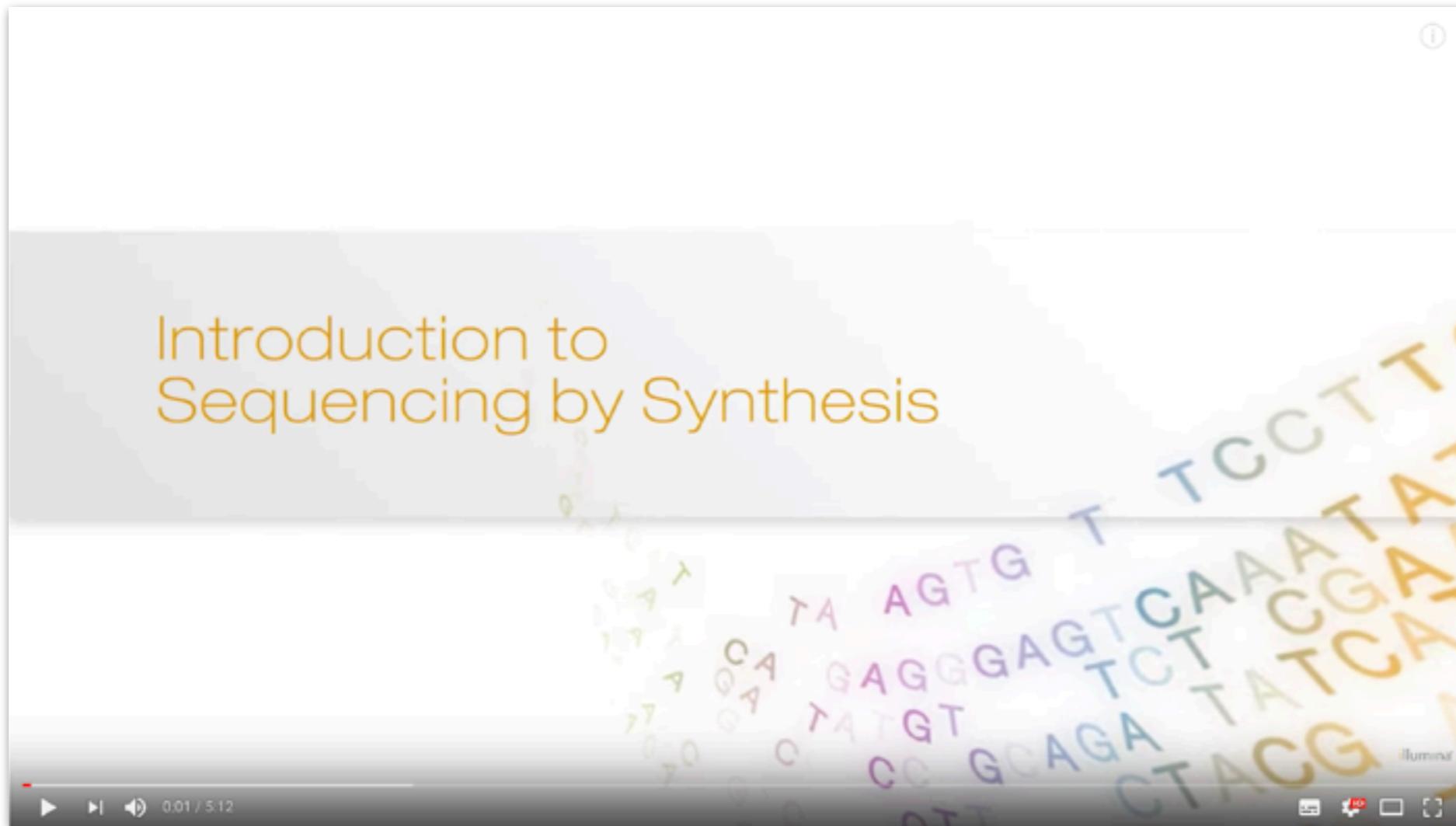
Ion Torrent

illumina®

# Illumina Sequencing

- Largest provider of sequencing technology
- NGS machines use "Sequencing-by-synthesis"
  - Developed at the University of Cambridge in 1990s
  - Spun into a company called Solexa in 1998
  - Solexa acquired by illumina in 2007
- Responsible for vast majority of DNA sequencing experiments worldwide

# Illumina Sequencing



<https://youtu.be/fCd6B5HRaZ8>

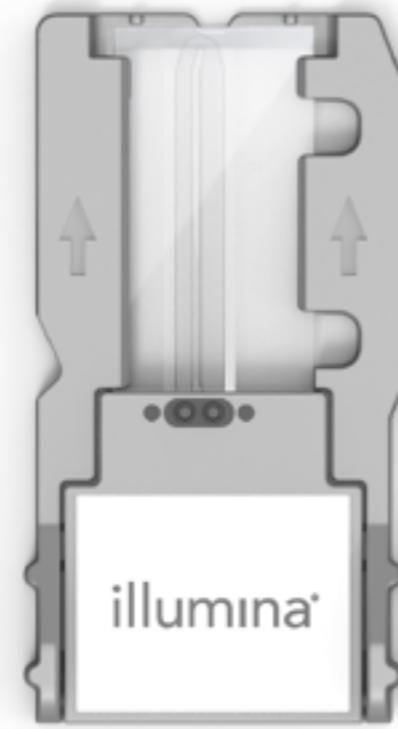
SciLifeLab

NGI stockholm

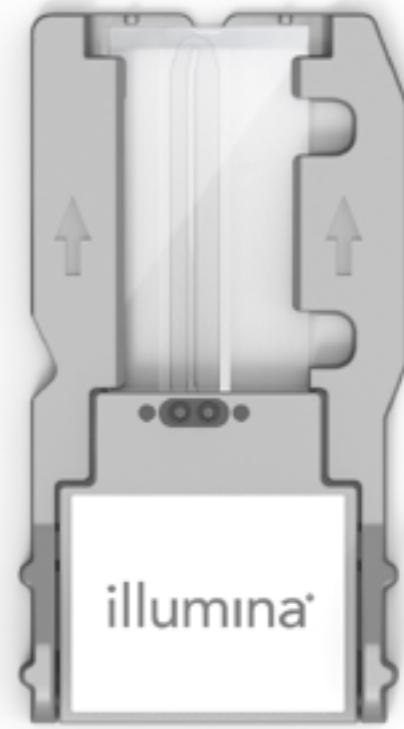
# Illumina iSeq 100



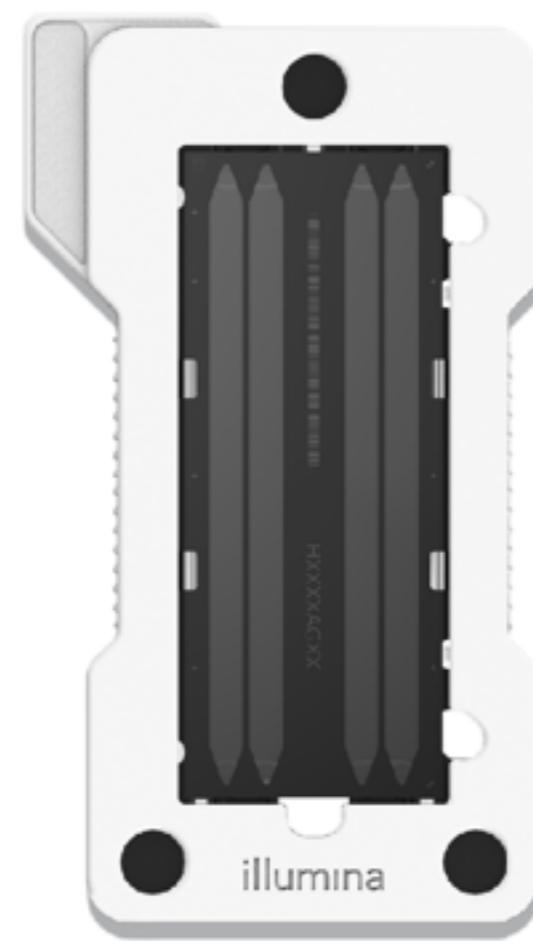
# Illumina MiniSeq 100



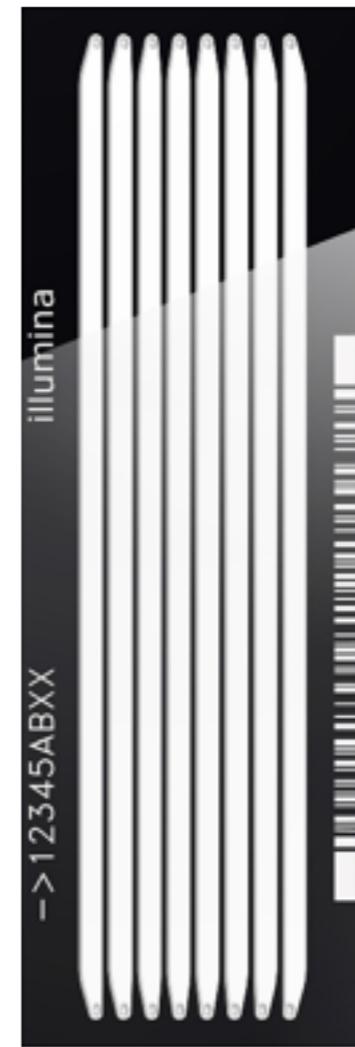
# Illumina MiSeq



# Illumina NextSeq



# Illumina HiSeq 2500



SciLifeLab

NGI stockholm

# Illumina HiSeq 3000



SciLifeLab

NGI stockholm

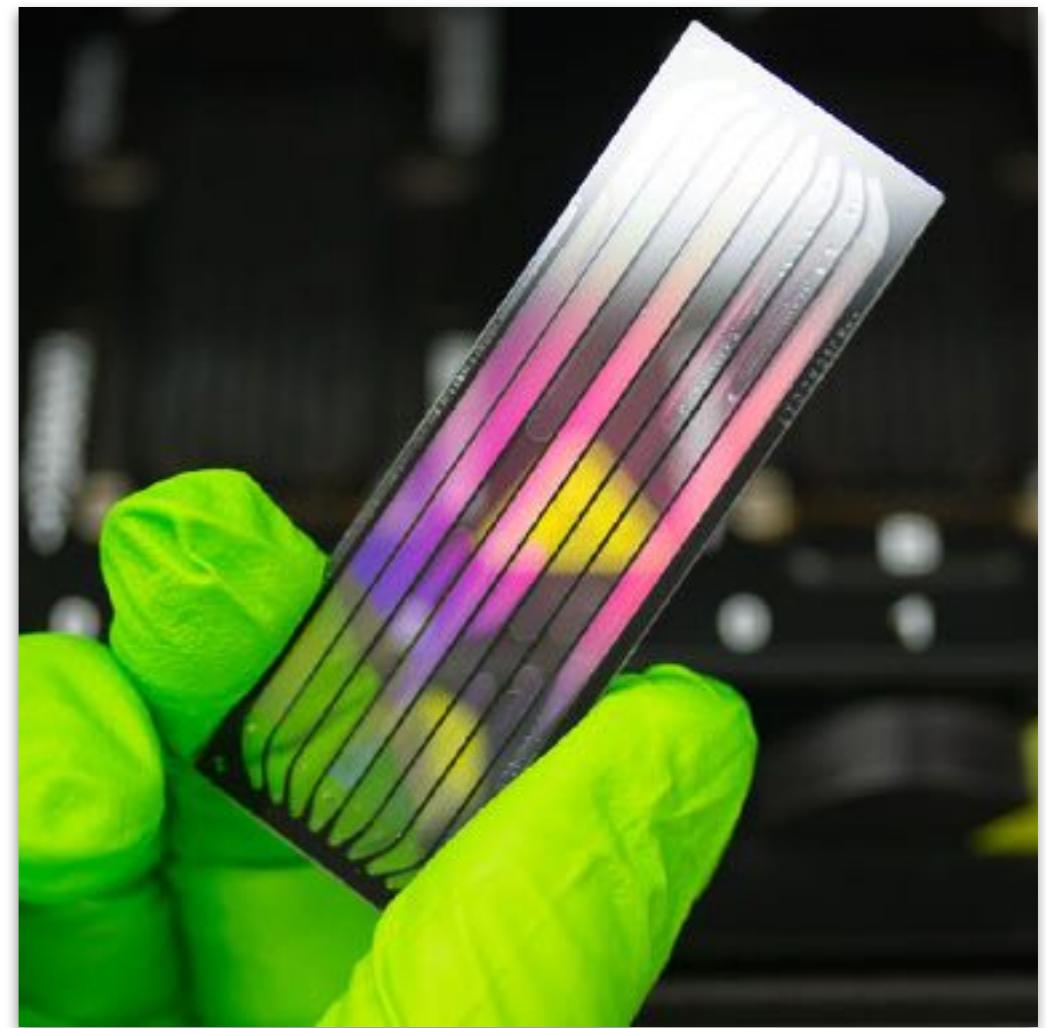
# Illumina HiSeq 4000



SciLifeLab

NGI stockholm

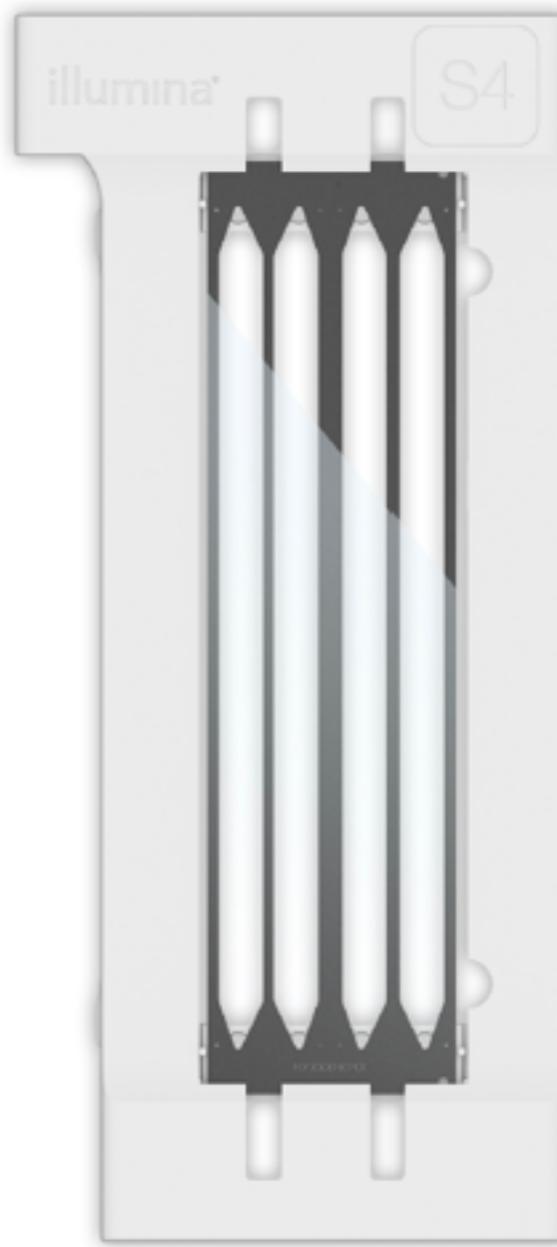
# Illumina HiSeq X



SciLifeLab

 NGI stockholm

# Illumina NovaSeq 6000



SciLifeLab

NGI stockholm

# Illumina at NGI

iSeq 100

Coming soon to NGI Uppsala  
Small cheap runs

MiSeq

Small runs, long reads (2x300bp)

HiSeq 2500

Primary machine for most of NGI's history

HiSeq X

Cheap, high throughput  
Only allowed to run WGS with > 15X coverage

NovaSeq 6000

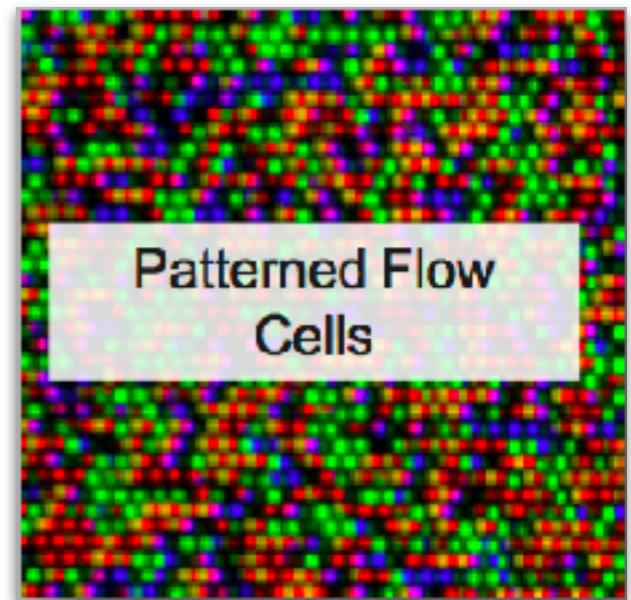
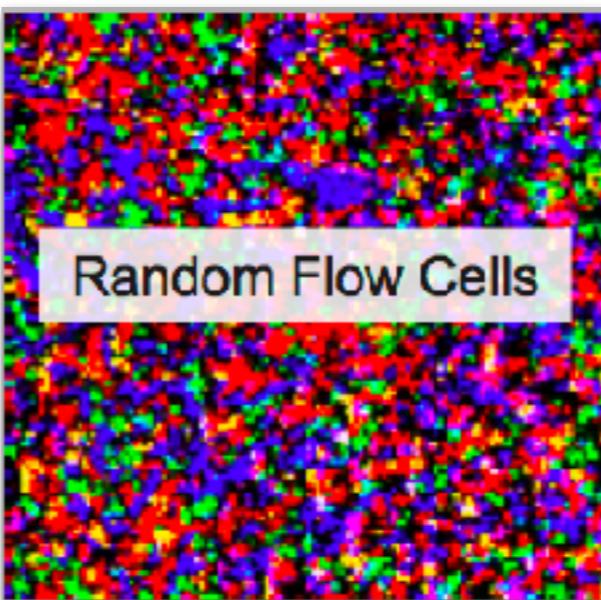
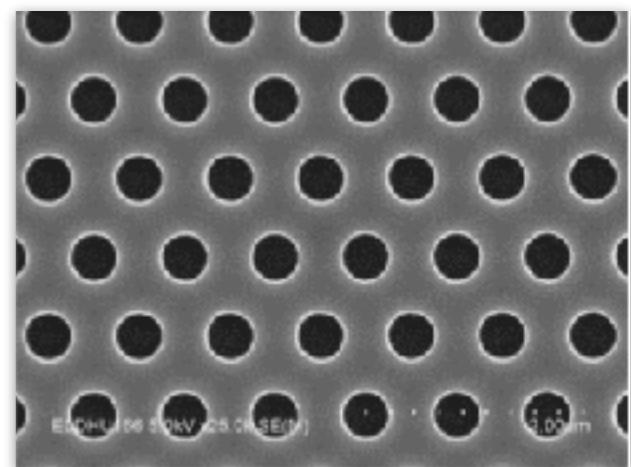
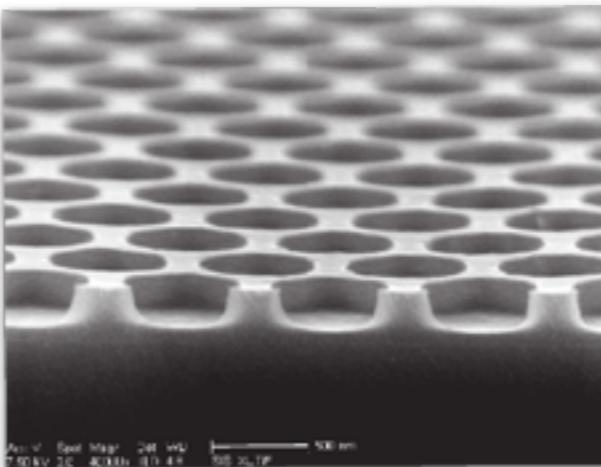
Newest machine, both Stockholm & Uppsala  
Will eventually replace HiSeq 2500

# How to choose

- Number of reads required
  - How many samples, how deeply sequenced?
- Type of reads required
  - Single End / Paired End, length?
- Urgency and cost
  - Sharing flow cells with other users
  - Best price for the project

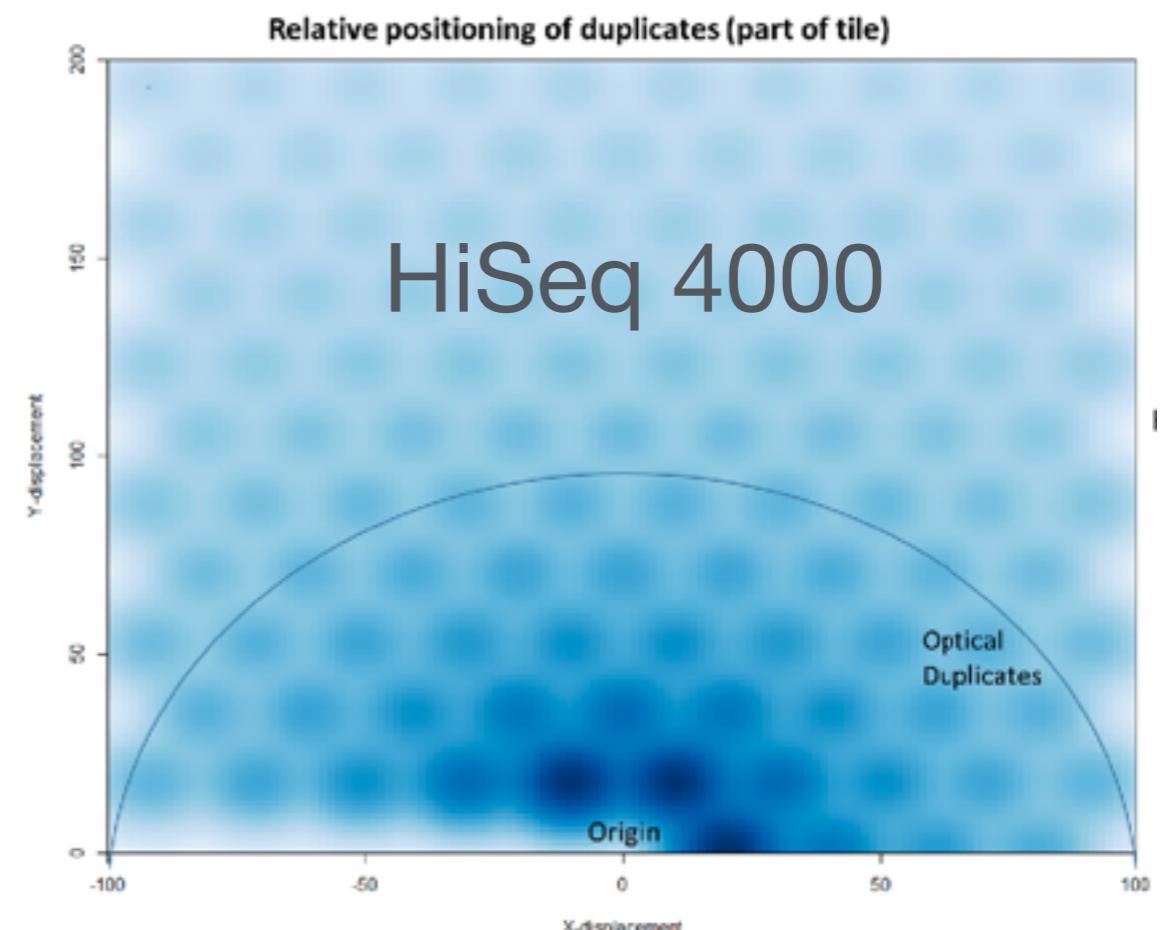
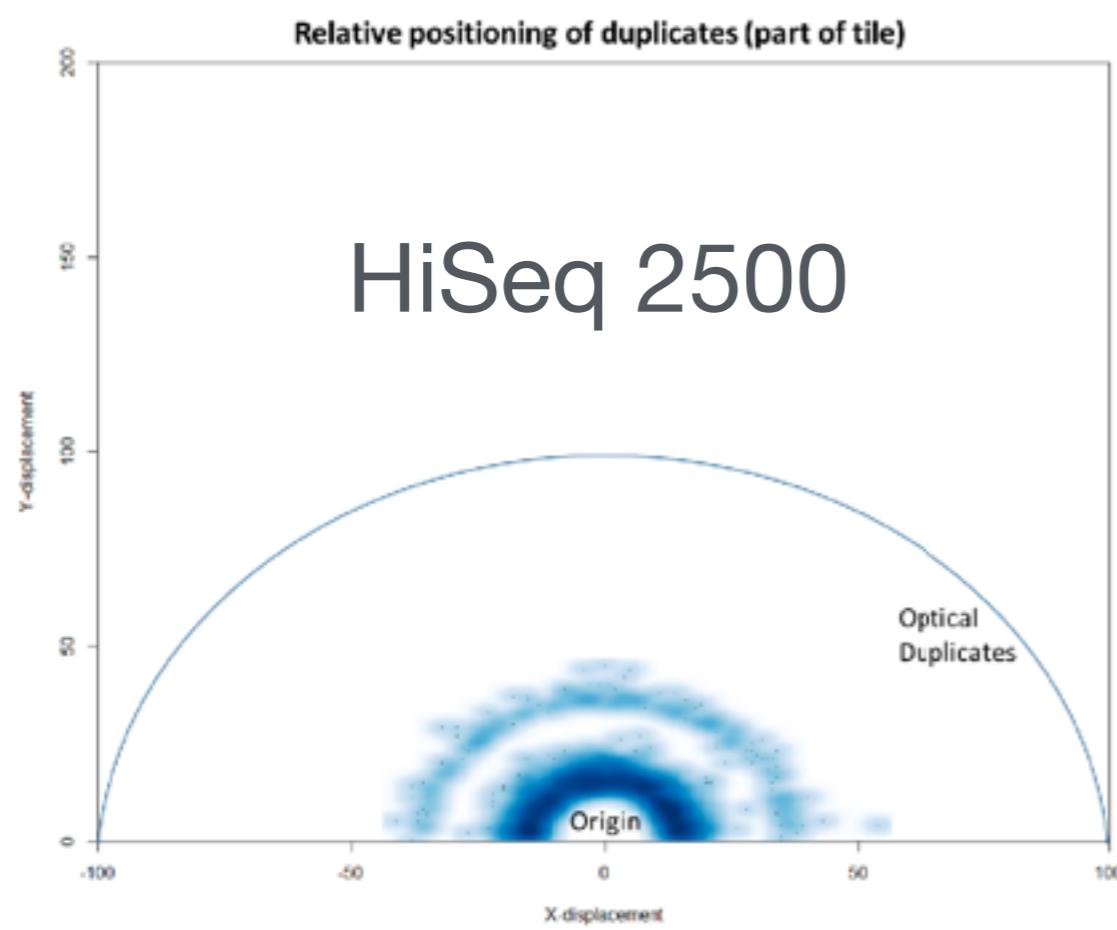
# Patterned flow cells

- New type of flow cell
  - HiSeq 4000, HiSeq X, NovaSeq
- Single sequence per well
  - Higher density, more data
- What's index-hopping?
  - ExAmp can mix up index pairs in tiny fraction of reads
  - Avoided with dual unique indexes



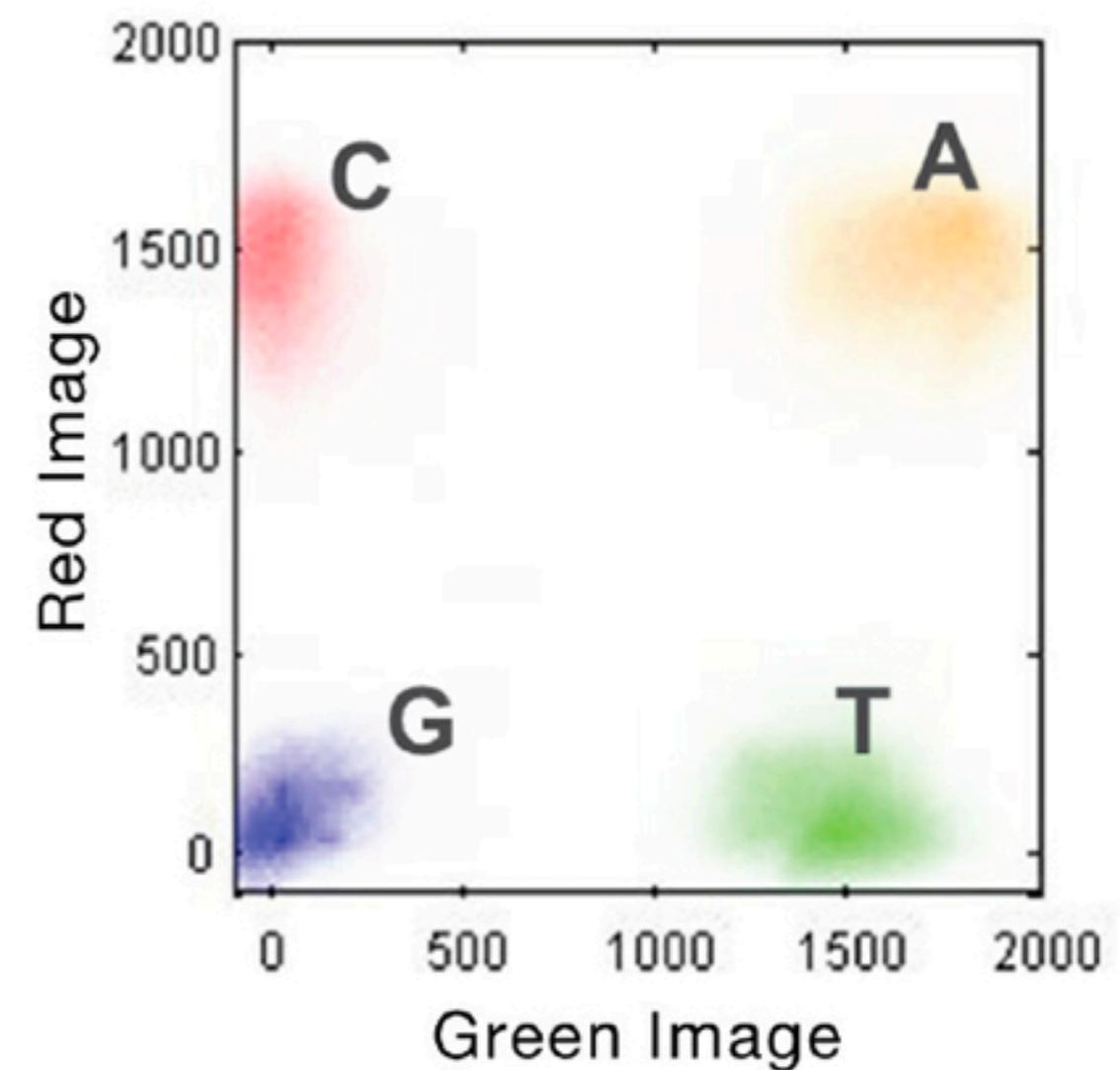
# Patterned flow cells

- Patterned flow cells can give "optical duplicates"
  - <https://sequencing.qcfail.com/articles/illumina-patterned-flow-cells-generate-duplicated-sequences/>
- Can be treated like regular PCR duplicates



# Two colour chemistry

- Older SBS used four different fluorophores
  - One for each nucleotide
- New machines use two
  - Faster and cheaper
  - NextSeq, NovaSeq, iSeq
- No signal = G
  - Can get poly-G if something goes wrong





PACIFIC  
BIOSCIENCES®

# PacBio

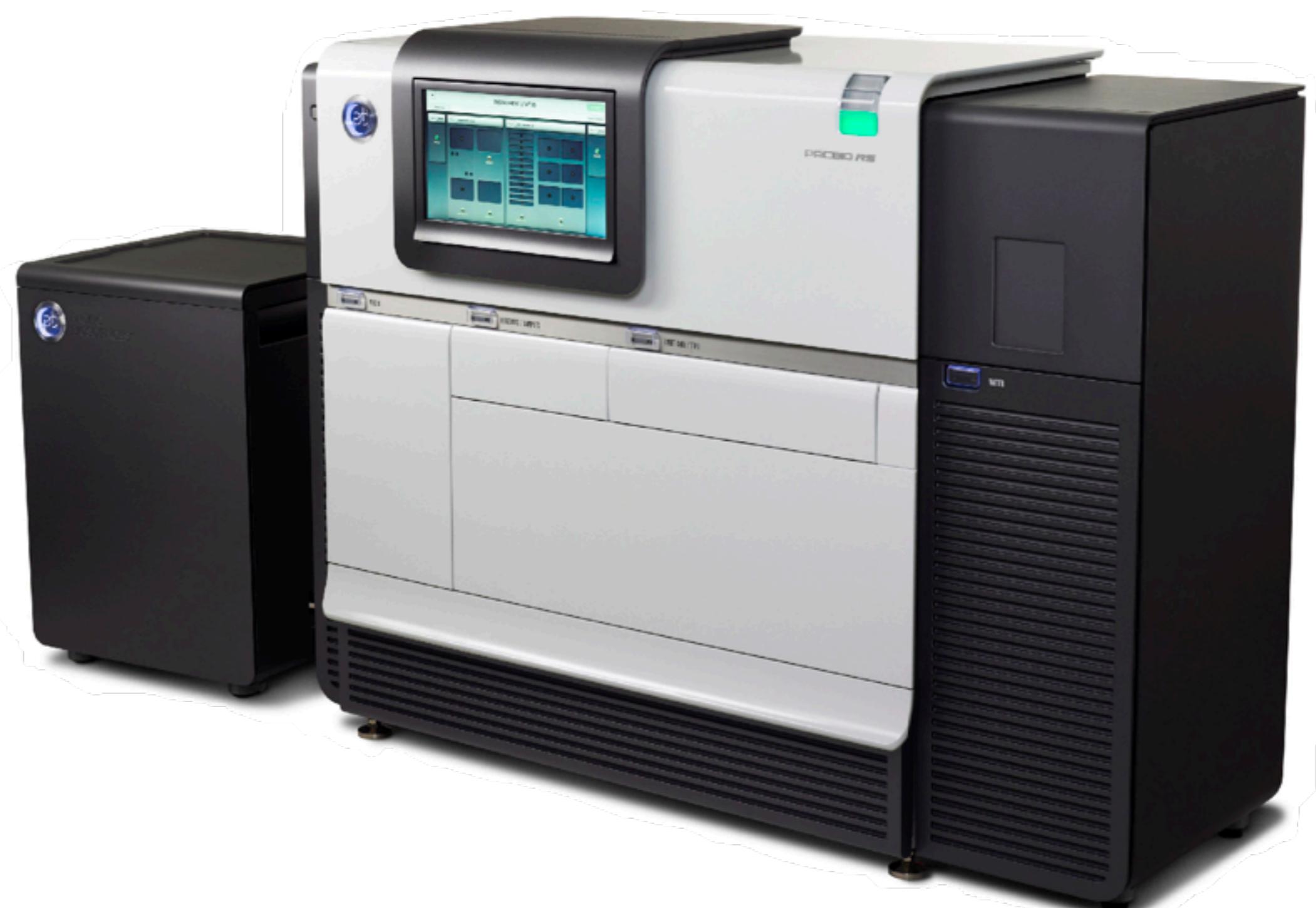
- Pacific Biosciences - specialists in long reads
  - Also uses fluorescent nucleotides
  - Polymerases immobilised at the bottom of tiny wells give off pulses as the nucleotides are incorporated
- Each well is independent, doesn't use sequencing rounds like illumina
- Can work with much longer DNA fragments
  - 250 bp – 60 kb (max ~160 kb)

# PacBio



<https://youtu.be/NHCJ8PtYCFc>

# PacBio RS II



SciLifeLab

NGI stockholm

# PacBio Sequel



SciLifeLab

NGI stockholm

# PacBio Sequencing

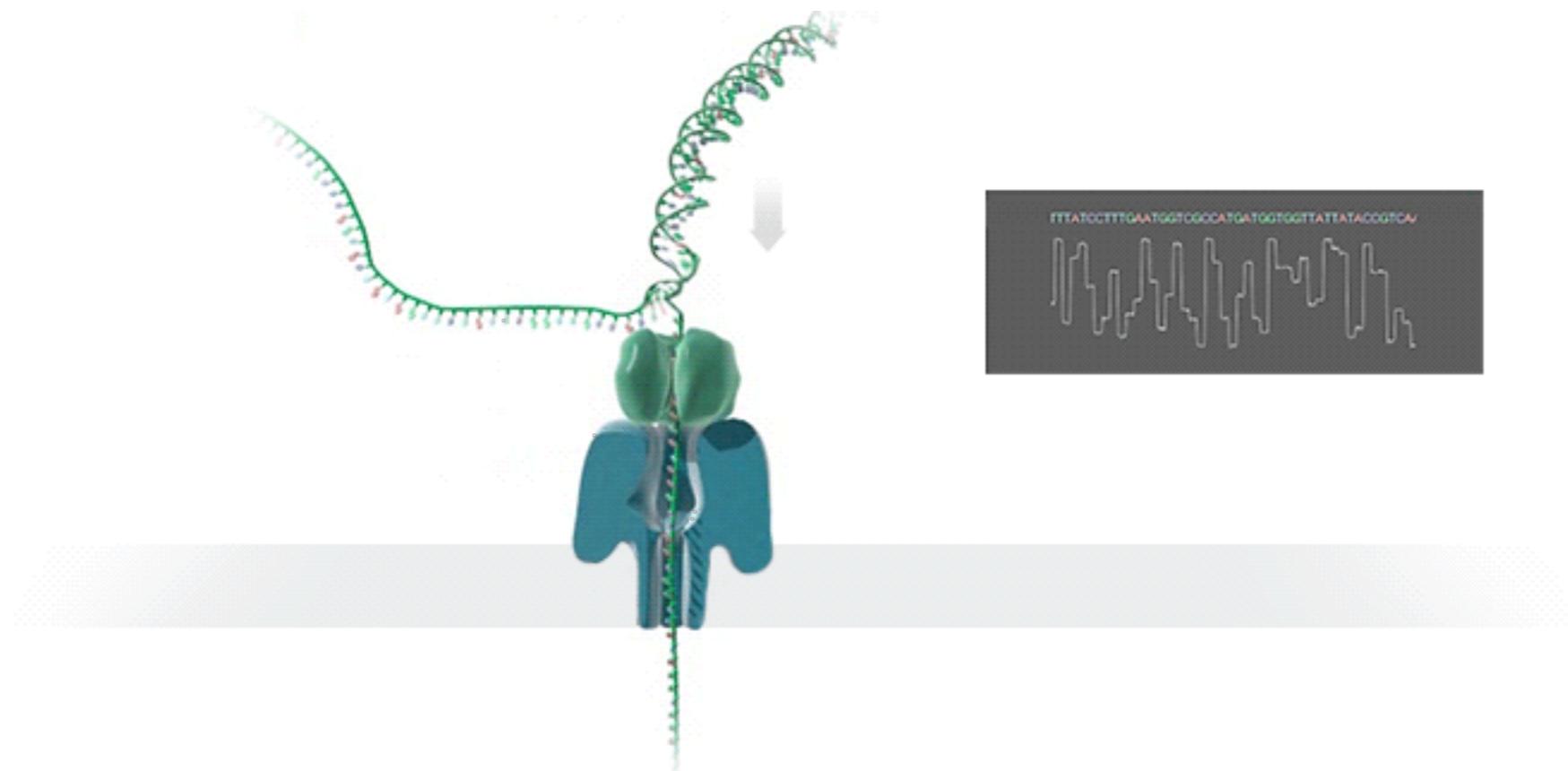
- Long reads are excellent for *de-novo* genome assembly and isoform detection
- Output is expensive compared to illumina, but getting better
  - Small genomes are no problem. Larger genomes are now becoming more feasible.
- New amplification-free enrichment using CRISPR-Cas9



# – Oxford Nanopore

- Newest contender in the sequencing world
  - Lots of hype and taken several years to become a reality
- Still developing very fast
  - Quality, yield and cost changing almost monthly
- High error rates (but better than they used to be)
  - Now 2-13% depending on sequencing type

# Oxford Nanopore



# MinION



SciLifeLab

NGI stockholm

# MinION



SciLifeLab

NGI stockholm

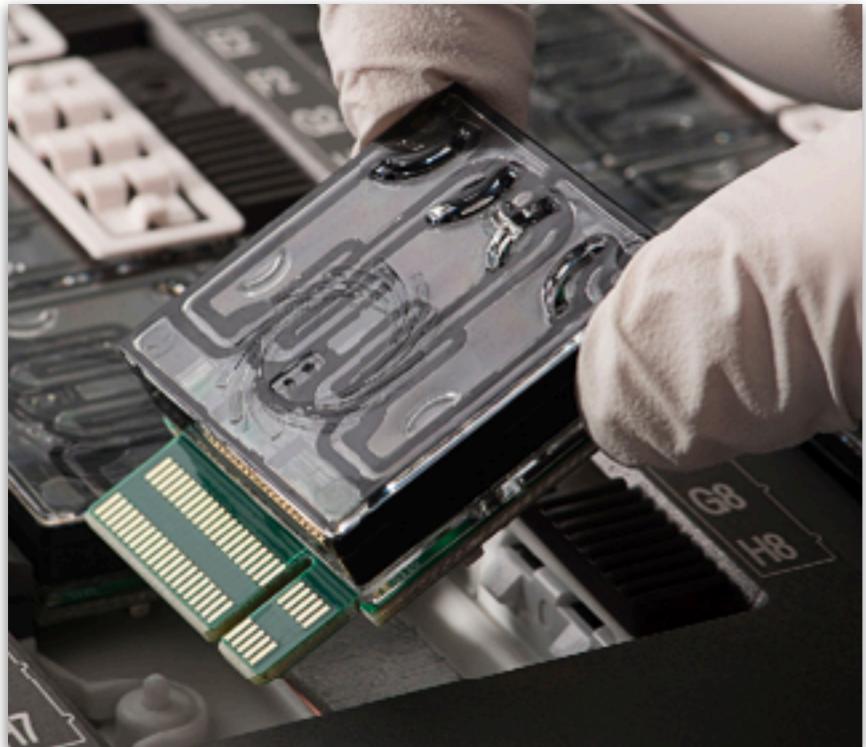
# GridION



SciLifeLab

NGI stockholm

# PromethION



SciLifeLab

NGI stockholm

# SmidgION



(not yet released)

# - Oxford Nanopore

- The best technology available for ultra long reads
  - Twitter users report getting reads over 1 Mbp long
  - "Whale spotting" - finding the longest reads on the end of the distribution curve
- Price dropping rapidly, but still expensive compared to illumina
- NGI has 2x MinIONs, hoping for PromethION soon

# **iontorrent**

**by Thermo Fisher Scientific**

# - Ion Torrent

- Main application
  - Microbial and metagenomic sequencing
  - Targeted re-sequencing (gene panels)
  - Clinical sequencing
- Short, single-end reads
- Fast run times

# - Ion Torrent PGM



- Yield
  - 0.1 - 1 Gbp
- Run time
  - 3 hrs
- Read length
  - 200 - 400 bp

# - Ion Torrent Proton



- Yield
  - 10 Gbp
- Run time
  - 4 hrs
- Read length
  - 200 bp



# - Ion Torrent S5 XL



- Yield
  - 1-13 Gbp
- Run time
  - 3 hrs
- Read length
  - 200 - 600 bp

# — Sequencing Type

- No need to remember all of this
  - Many considerations, changing all the time
  - We are experts - come and speak to us!

[support@ngisweden.se](mailto:support@ngisweden.se)

<https://ngisweden.scilifelab.se/>

# Sequencing Applications

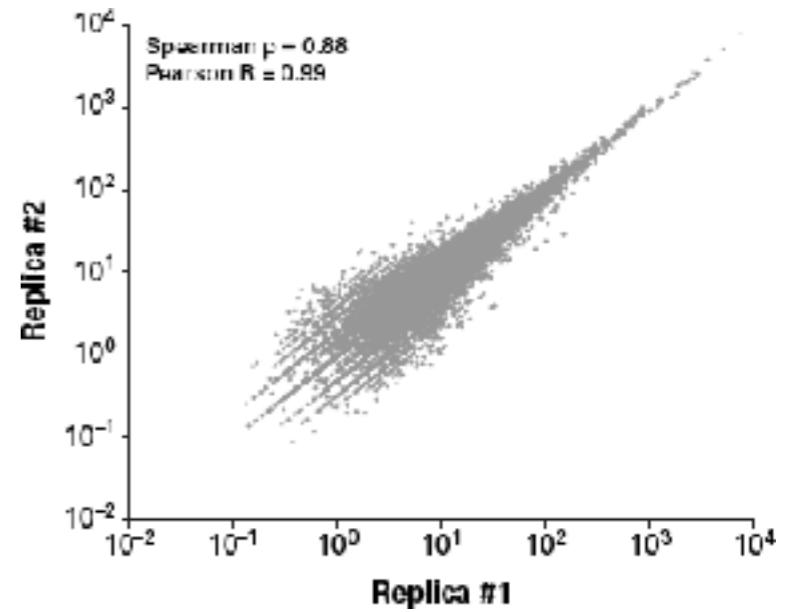


# Library Preparation

- All high throughput sequencing requires some kind of library preparation
  - Add adapters for sequencing chemistry
  - Adjust DNA fragment lengths
  - Incorporate biological signal into sequence
  - Add required enzymes
- Different library preps enable different applications

# RNA Sequencing

- Choose a type of RNA
  - Protein coding mRNA (poly-A)
  - All RNA (rRNA depletion)
  - Small RNA
- Define your limitations
  - Low-input material
  - Low quality material (eg. FFPE)
- Choose your question
  - Differential gene expression
  - Differential isoform detection & quantification
  - Fusion gene detection

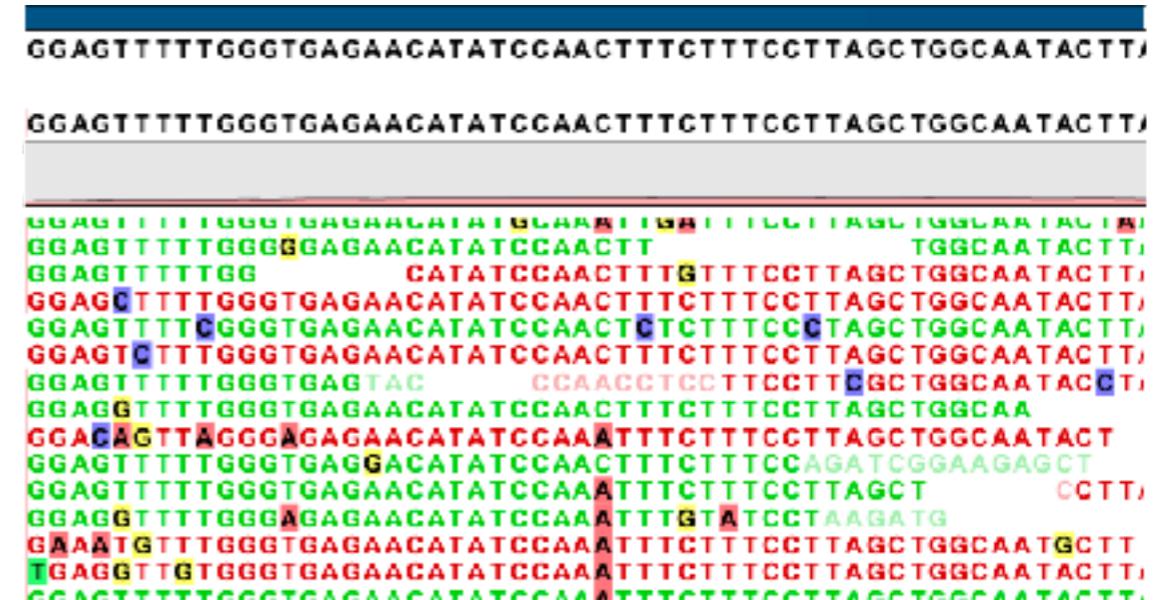


# RNA Sequencing

- Illumina sequencing RNA library prep kits
  - Illumina TruSeq RNA *Protein-coding poly-A*
  - Illumina RiboZero *rRNA depletion*
  - Illumina TruSeq RNA Exome *FFPE / low quality*
  - Clontech SMARTER Pico *low input*
  - Illumina TruSeq Small RNA *small RNA*
- Oxford Nanopore, PacBio, IonTorrent

# DNA Sequencing

- Choose your question
  - SNP, SNV, indel calling
  - Structural variant detection
  - *De-novo* genome assembly
- Choose your priorities
  - Sequencing accuracy
  - Sequencing depth
  - Ultra-long reads
- Define your requirements
  - Low-input material
  - Low quality material (eg. FFPE)



# - DNA Sequencing

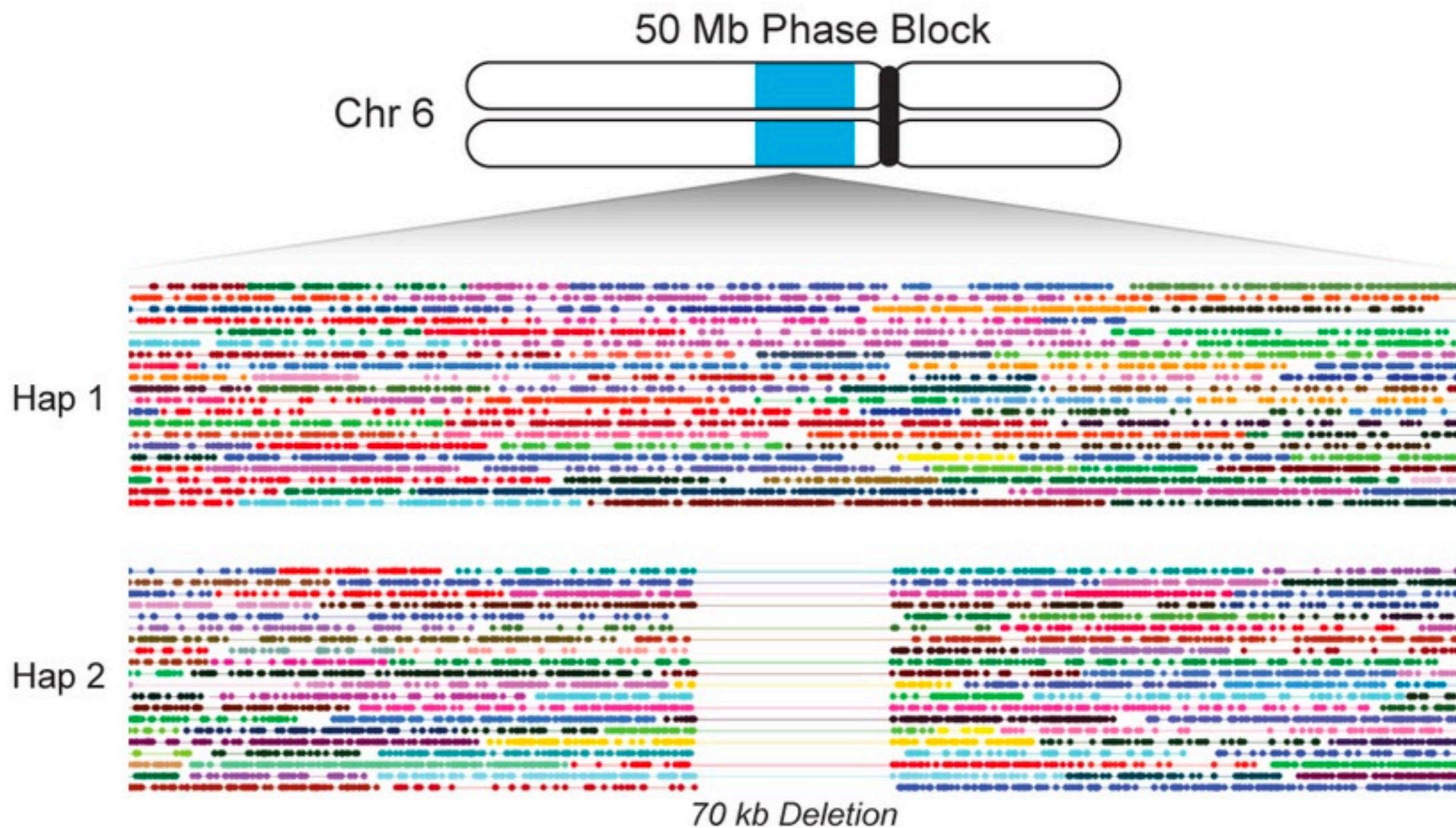
- Illumina sequencing DNA library prep kits
  - Illumina TruSeq DNA PCR Free *Best quality*
  - Rubicon ThruPLEX *Low input*
  - Illumina Nextera XT *Cheap (plate format)*
  - Illumina Nextera Flex *Fast and simple*
  - 10X Genomics *Linked reads*
- Oxford Nanopore, PacBio, IonTorrent

# - 10X Genomics

- Chromium instrument uses droplet emulsion technology for nanoliter reaction volumes
- Linked-read sequencing
  - Large molecules fragmented in droplets and barcoded
  - Normal short-read illumina sequencing used
  - Long fragments (20-100+ Kbp) reassembled from barcodes
- Regular illumina sequencing libraries produced

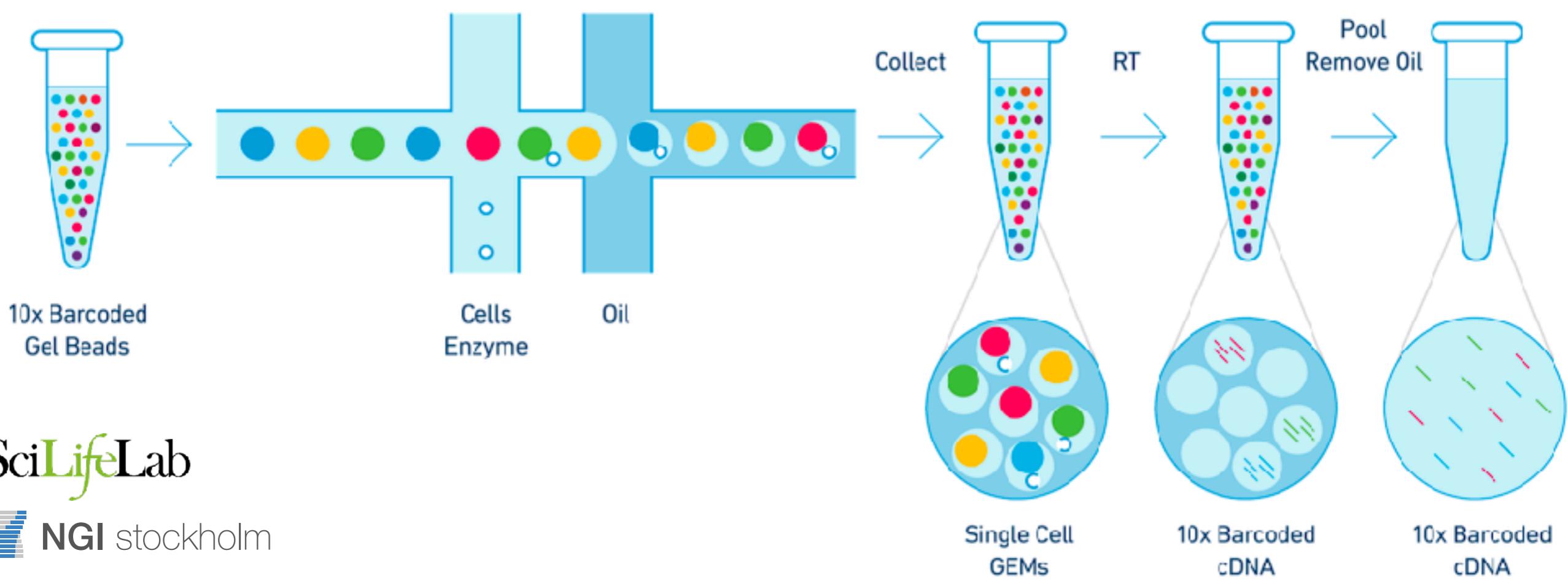


# 10X Genomics



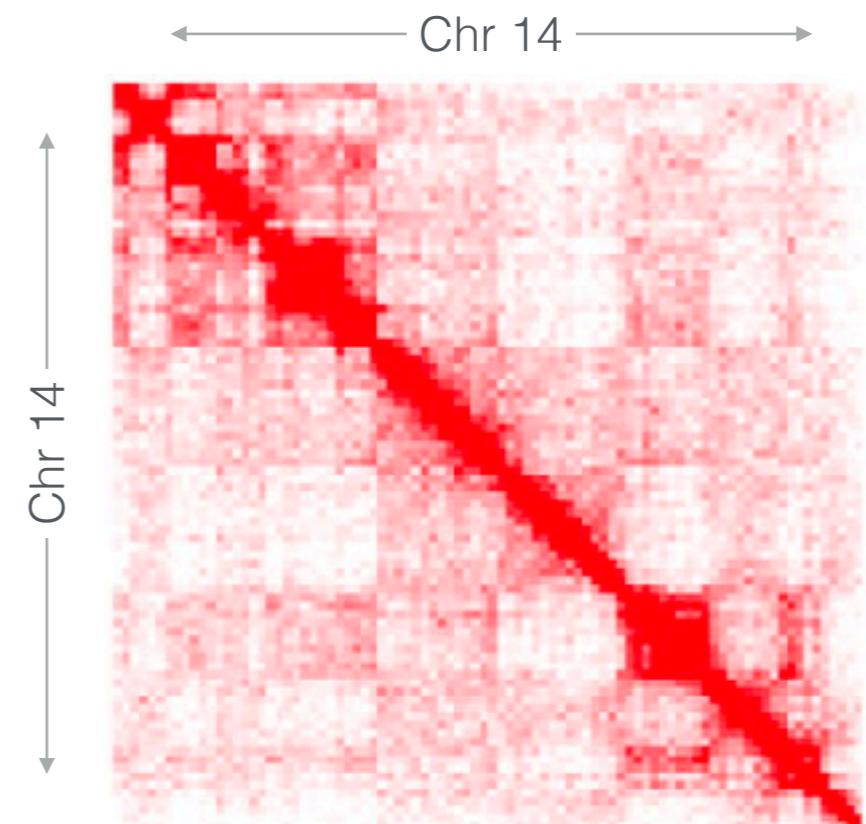
# 10X Genomics

- Single cell RNA sequencing
  - Thousands of cells captured in droplets
  - Each RNA molecule tagged with droplet barcode



# Hi-C

- Now testing Hi-C in NGI Stockholm
  - Proximity ligation assay to detect physical colocation of DNA fragments within cell nuclei
- Multiple applications for data
  - Epigenetics
  - De-novo genome assembly
  - Structural variation detection



# Methylation Sequencing

- Bisulphite sequencing detects Cytosine methylation in genomic DNA
  - Unmethylated Cs converted to Uracil by bisulfites and sequenced as T
  - Methylated Cs are protected and sequenced as C
- Oxidative bisulphite informs about hydroxy-methylation
  - Current under development at NGI Stockholm
- PacBio and Oxford Nanopore able to detect some native base modifications

# RAD Sequencing

- Restriction-site Associated DNA sequencing, also known as GBS (Genotyping By Sequencing)
  - Genome fragmented using a restriction enzyme
  - Narrow size range purified - same regions of genome for all individuals
- Allows cheap high-depth variant calling for large numbers of samples, without a reference genome
  - Excellent for population genomics and ecology

# Bioinformatics at the NGI

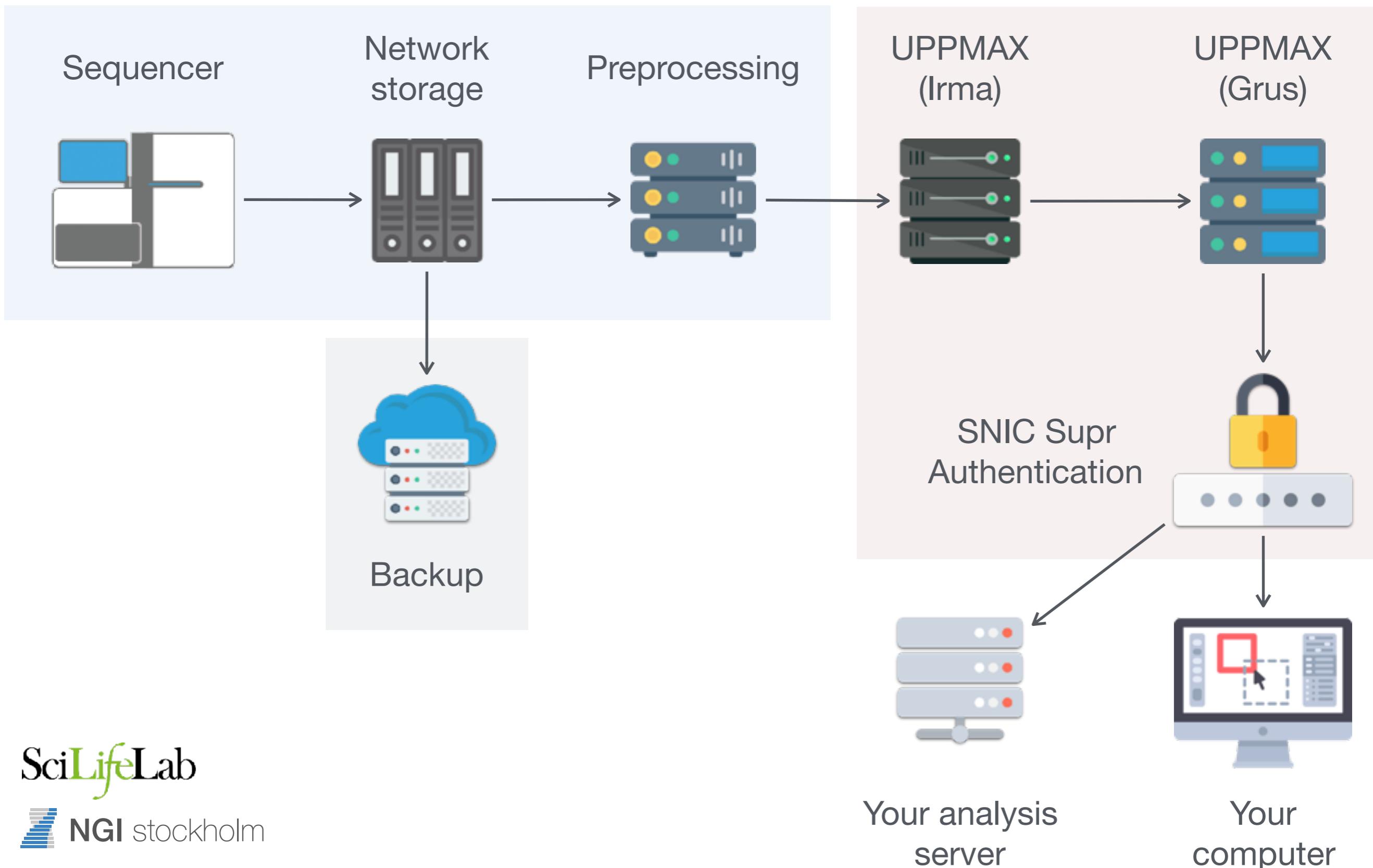


# Bioinformatics at NGI

- Raw sequencing data management
  - Demultiplexing, data transfers, backups, delivery
- Quality control
  - Every project is checked against quality criteria
- Automated analysis pipelines
  - Standardised pipelines give reproducible results
- Software development



# NGI Data Handling



# — Grus Deliveries

- UPPMAX tool for NGI data deliveries
  - NGI creates a SNIC Supr "delivery project" for each NGI sequencing project
  - Project PI and contact person given access, according to what was put on the order form
  - Email sent with project ID and instructions
- Grus is for secure short term storage only
  - Requires two-factor authentication



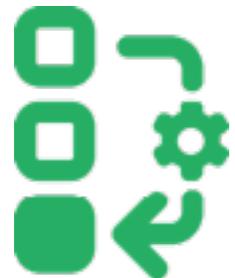
# Analysis Pipelines

- Initial data analysis for major protocols
- Internal QC and standardised starting point for users
- All software open source and on GitHub
  - <http://opensource.scilifelab.se/>
  - <http://github.com/SciLifeLab/>
- Accredited facility



Ackred. nr 1850  
Provning  
ISO/IEC 17025

# - Analysis Requirements



Automated



Reliable

**nextflow**

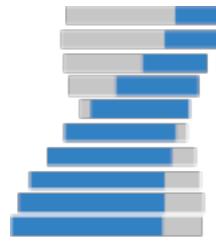


Easy for others to run

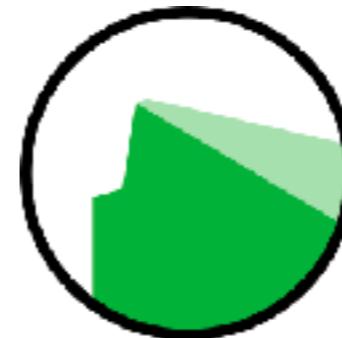


Reproducible results

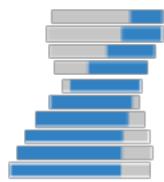
# Analysis Pipelines



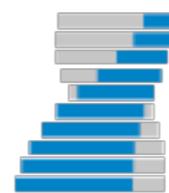
**NGI**-RNAseq



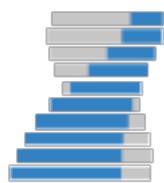
**Sarek**



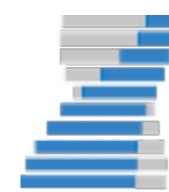
**NGI**-MethylSeq



**NGI**-ExoSeq



**NGI**-smRNAseq



**NGI**-RNAfusion



**NGI**-ChIPseq



**NGI**-NeutronStar

SciLifeLab

NGI stockholm

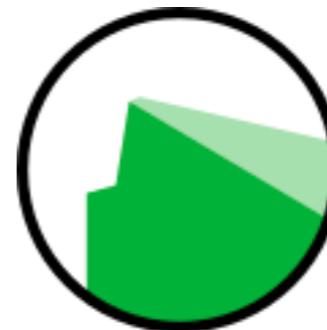
NouGAT (*de-novo*)

# Sarek Somatic

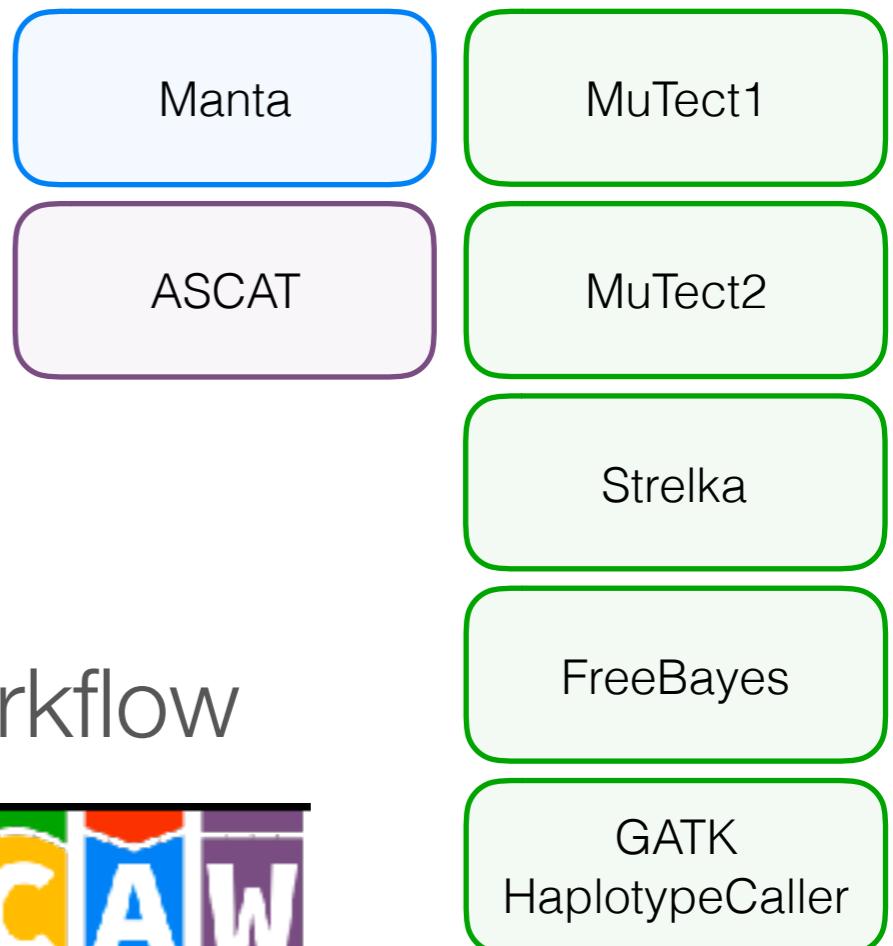


<https://github.com/SciLifeLab/Sarek>

- Tumour/Normal pair WGS analysis based on GATK best practices
  - SNPs, SNVs and indels
  - Structural variants
  - Heterogeneity, ploidy and CNVs
- Germline and/or Somatic analysis
  - Formerly called Cancer Analysis Workflow

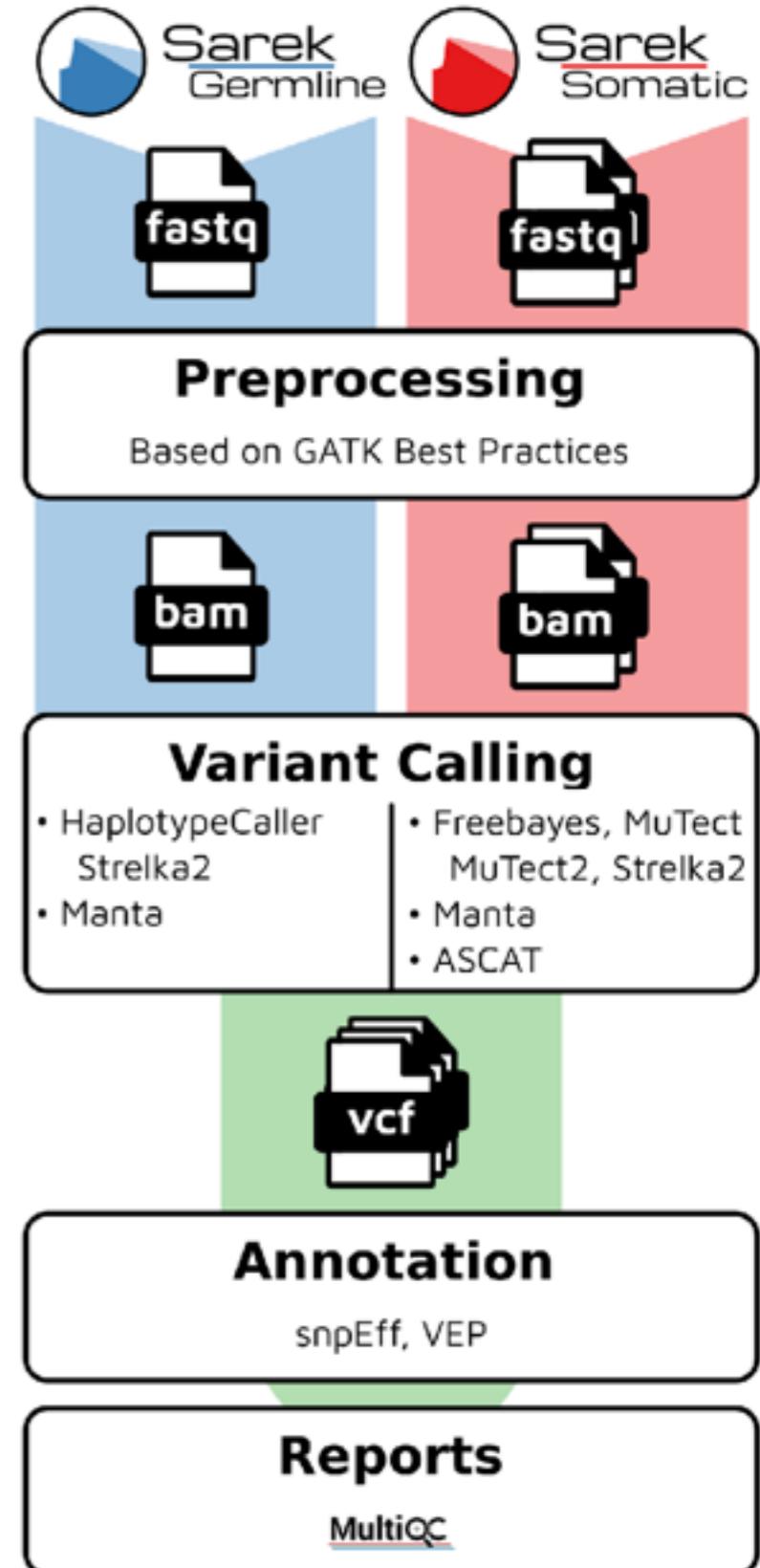


**Sarek**



# Sarek

- Tool split into sub-workflows
- Bash wrapper script runs whole workflow
- Manuscript submitted this week, preprint available on bioRxiv
  - <https://www.biorxiv.org/content/early/2018/05/09/316976>



# NGI-RNAseq



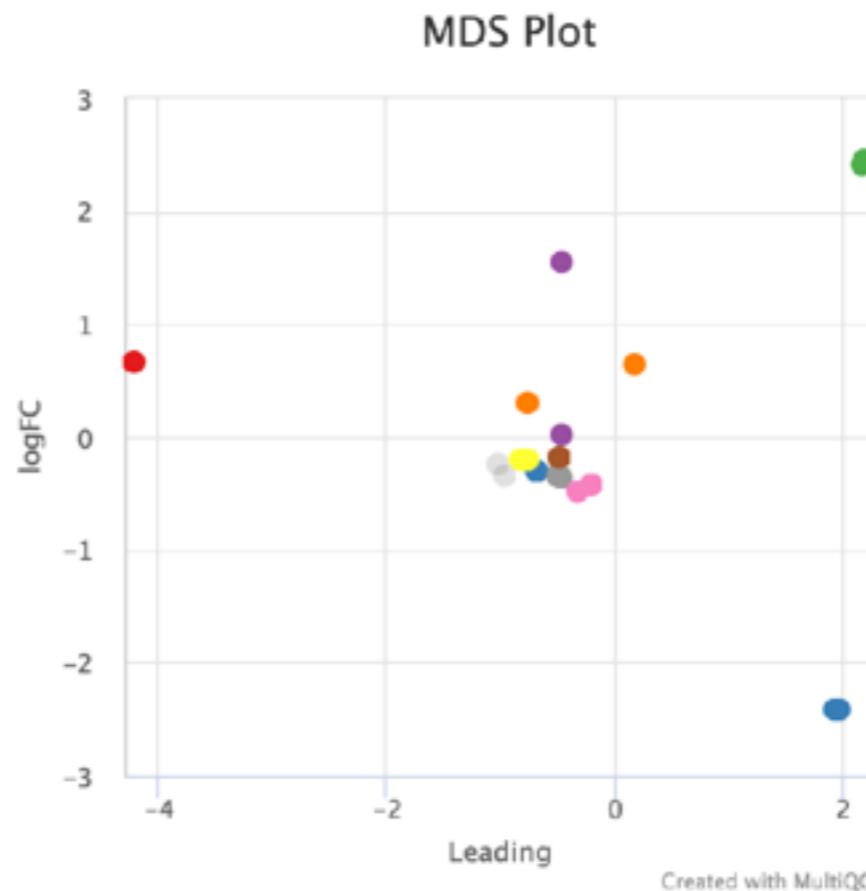
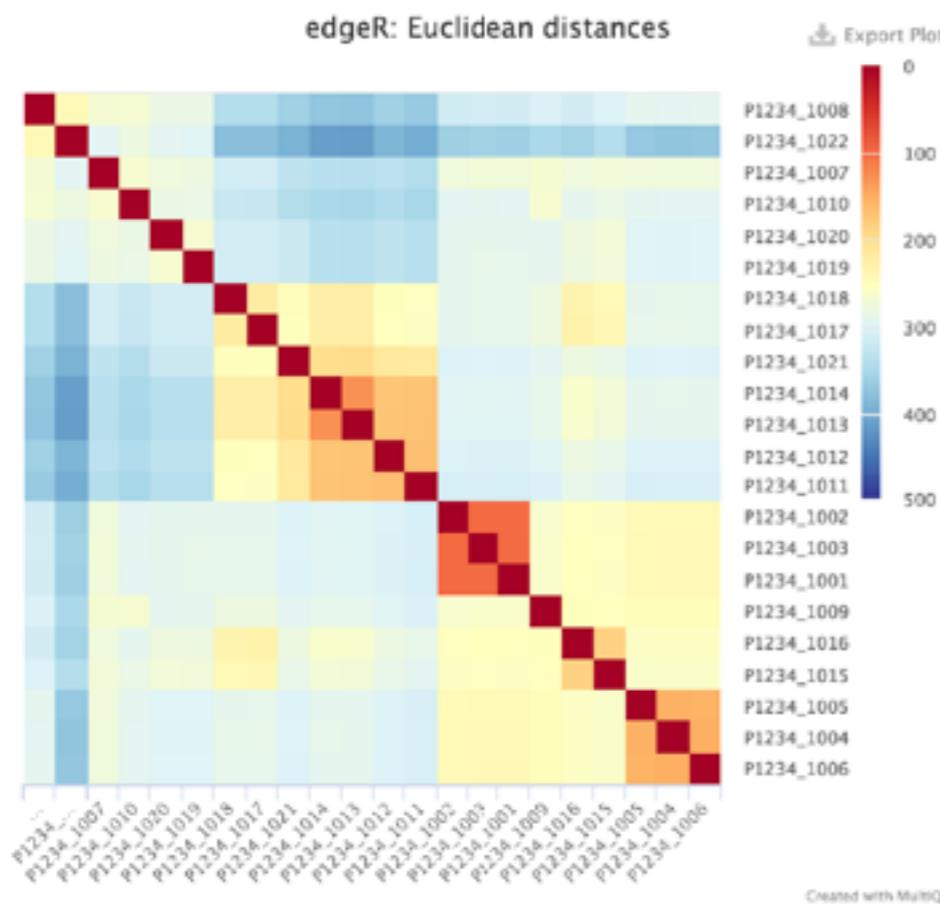
<https://github.com/SciLifeLab/NGI-RNAseq>



NGI, April – December 2017

**10,227** samples processed

**131** user projects



Raw data

Read alignment

Gene counts

Quality Control

Reporting



MIT Licence

# NGI-RNAseq



<https://github.com/SciLifeLab/NGI-RNAseq>



## SciLifeLab



Karolinska  
Institutet



UPPSALA  
UNIVERSITET



LUND'S  
UNIVERSITET



UNIVERSITY OF  
GOTHENBURG



Quantitative Biology Center  
Tübingen, Germany



How to cite your work? PMID =? #21

zhenylsong opened this issue 7 months ago

question

We have recently completed a simple project on human next-generation data analysis. And in the pipeline, we used the small RNA/ mRNA/ChIP Q pipelines from your Github account. I googled to SciLifeLab projects, there seems no published work. How can I show my respect to your work?

Vladimir Kiselev  
@wikiselev

Following

Just had a error-free run of @tallphil's RNA-Seq pipeline using @PaoloDiTommaso's @nextflowio on our LSF cluster! Started yesterday, came back today, all done! Thanks a lot guys for your great work!

A screenshot of a GitHub issue and a tweet from Vladimir Kiselev. The issue is about how to cite work from the NGI-RNAseq repository. The tweet is a success story about running the pipeline on an LSF cluster using Nextflow.

Now running using

slurm  
workload manager

openstack.

aws

kubernetes

QBiC

docker

A screenshot of a resource usage report showing the distribution of resource usage for each process over time. The report includes a timeline at the top and several charts below showing CPU usage, memory usage, and other metrics.

# nf-core

- A community effort to collect a curated set of Nextflow analysis pipelines
  - GitHub organisation to collect pipelines in one place
  - No institute-specific branding
  - Strict set of guideline requirements
  - Automated testing for code style and function

**nf-core**   
<https://nf-co.re>

# nf-core

The screenshot shows the nf-core website homepage. At the top, there's a navigation bar with links for Home, Pipelines, Tools, Docs, and About. Below the header, the nf-core logo is displayed with a small green icon. A sub-header reads: "A community effort to collect a curated set of analysis pipelines built using Nextflow." A large green button labeled "VIEW PIPELINES" is prominent. The main content area is divided into three columns:

- For facilities**: Highly optimised pipelines with excellent reporting. Validated releases ensure reproducibility.
- For users**: Portable, documented and easy to use workflows. Pipelines that you can trust.
- For developers**: Companion templates and tools help to validate your code and simplify common tasks.

Below these columns, a section states: "Nextflow is an incredibly powerful and flexible workflow language. nf-core pipelines adhere to strict guidelines - if one works, they all will."

At the bottom, there are six cards:

- Documentation**: Extensive documentation covering installation, usage and description of output files ensures that you won't be left in the dark. (Icon: clipboard)
- CI Testing**: Every time a change is made to the pipeline code, nf-core pipelines use continuous-integration testing to ensure that nothing has broken. (Icon: Travis CI)
- Stable Releases**: nf-core pipelines use GitHub releases to tag stable versions of the code and software, making pipeline runs totally reproducible. (Icon: checkmark)
- Docker**: Software dependencies are always available in a bundled docker container, which Nextflow can automatically download from dockerhub. (Icon: Docker logo)
- Singularity**: If you're not able to use Docker, built-in support for Singularity can solve your HPC container problems. These are built from the docker containers. (Icon: Singularity logo)
- Bioconda**: Where possible, pipelines come with a bioconda environment file, allowing you to set up a new environment for the pipeline in a single command. (Icon: Bioconda logo)



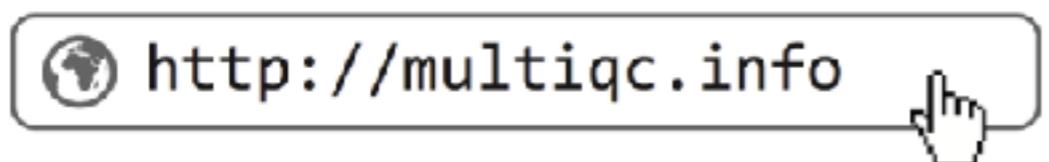
- Easy to run pipelines
- Helpful community
- Super reproducible results

# Quality Control

- Every project has some level of quality control checks
  - Sequencing quality
  - FastQC, FastQ Screen
- Analysis pipelines give application-specific QC
  - Qualimap, RSeQC
- Reporting is done using MultiQC

# – MultiQC

- Reporting tool, parses logs from completed analysis
- Creates single HTML report for all samples & steps in a project
- Interactive plots for data exploration
- Current version now has 61 supported tools
- Works with anything from tens → thousands of samples
- Highly customisable



P1234: Test\_NGI\_Project

General Stats

NGI-RNAseq

Sample Similarity

MDS Plot

STAR

Cutadapt

FastQC

Sequence Quality Histograms

Per Sequence Quality Scores

Per Base Sequence Content

Per Sequence GC Content

Per Base N Content

Sequence Length Distribution

Sequence Duplication Levels

Overrepresented sequences

Adapter Content

## P1234: Test\_NGI\_Project

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

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

Report generated on 2017-05-17, 18:43 based on data in:

/Users/philewels/GitHub/MultiQC\_website/public\_html/examples/ngi-rna/data

**NGI names****User supplied names**

### General Statistics

**Copy table****Configure Columns****Plot**

Showing 22/22 rows and 6/6 columns.

Sample Name	% Aligned	M Aligned	% Trimmed	% Dups	% GC	M Seqs
P1234_1001	68.2%	22.8	10.3%	71.3%	49%	33.7
P1234_1002	67.9%	20.9	10.7%	70.1%	50%	31.1
P1234_1003	64.7%	21.7	11.0%	72.3%	50%	33.7
P1234_1004	55.2%	17.0	13.2%	73.4%	51%	31.2
P1234_1005	53.0%	17.7	15.9%	75.8%	52%	33.8
P1234_1006	52.7%	16.1	14.1%	73.8%	52%	30.8
P1234_1007	33.0%	7.0	32.0%	60.5%	52%	21.8
P1234_1008	27.5%	4.3	44.2%	79.1%	50%	16.7
P1234_1009	52.3%	10.5	20.9%	64.2%	48%	20.5

Toolbox



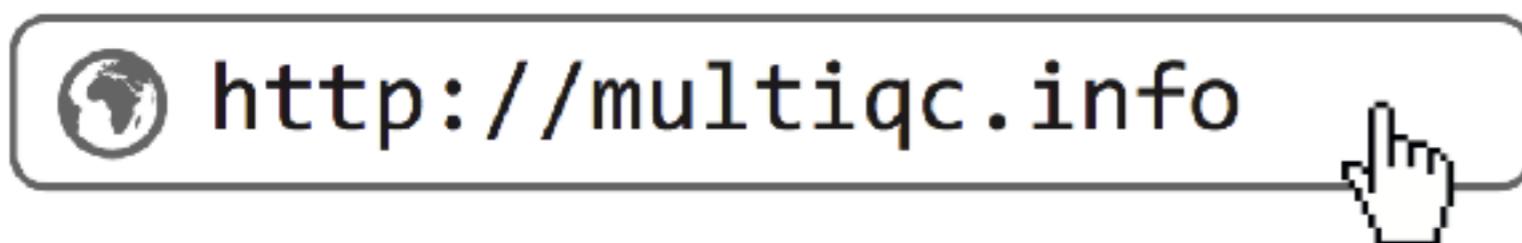
A



# Getting MultiQC



BIOCONDA



The screenshot shows the MultiQC website at <http://multiqc.info>. The page has a dark background with a hexagonal grid pattern. On the left, there's a large "MultiQC" logo with a magnifying glass icon. Below it, a tagline reads: "Aggregate results from bioinformatics analyses across many samples into a single report". A paragraph explains: "MultiQC searches a given directory for analysis logs and compiles a HTML report. It's a general use tool, perfect for summarising the output from numerous bioinformatics tools." On the right, there's a navigation bar with links for "Home", "Docs", "Example Reports", and "GitHub". Below the navigation, there are four blue buttons: "Documentation", "View on PyPI", "View on GitHub", and "Quick Install". The "Quick Install" button contains the command: `pip install multiqc # Install  
multiqc . # Run`. At the bottom, a note says: "Python or pip not installed? See the full installation instructions."

# Conclusions



# If you have a project

- Visit our order portal
  - Create projects
  - Request meetings
- Send us an email

Information • Documents • Contact • About us

**NGI Sweden Order Portal**

This portal is for submitting orders for services provided by the National Genomics Infrastructure Sweden (NGI). To make an order, please log in and choose the application most suitable for your project. If uncertain about the choice of technology, please select the "Request a meeting" option. You can read more about the different technologies and [How to place an order](#) under "Information" in the menu at the top of the page.

Projects from other countries are admissible, but have lower priority than projects performed by researchers based in Sweden. Depending on the queue situation, NGI may decide to decline a non-Swedish project altogether.

Turn Around Times and Status for the Stockholm node.

Subscribe to our mailing list:

**Login**

Email:

Password:

**Request a meeting**

If you are unsure about the appropriate method for your scientific problem, request a meeting for a discussion with us.

**Illumina Sequencing**

Order form for Illumina sequencing.

**Ion Sequencing**

Order form for sequencing by Ion Proton or Ion 550L.

**PacBio Sequencing**

Order form for PacBio sequencing. This is available only at the NGI Uppsala UGC node.

**Genotyping and array-based epityping**

Order form for genotyping and DNA methylation analysis using the Illumina EPIC beadchip.

All news

<https://ngisweden.scilifelab.se>

[support@ngisweden.se](mailto:support@ngisweden.se)

SciLifeLab

NGI stockholm

# Find our tools

- View our open-source software
- All code available on GitHub

<http://opensource.scilifelab.se>

The screenshot shows a GitHub repository page for the SciLifeLab open-source tools. The header features the SciLifeLab logo with the tagline "Open Source is in our DNA". Below the header, there is a list of tools, each with a star rating, icon, name, and description.

Star Rating	Icon	Tool Name	Description
★ 19	AWS-iGenomes	AWS-iGenomes	Reference genomes on AWS S3
★ 41	Chanjo	Chanjo	Coverage analysis for clinical sequencing
★ 16	CheckQC	CheckQC	Quick quality control of Illumina runs
★ 64	Cluster Flow	Cluster Flow	Simple pipelines for bioinformatics
CONCOCT Clustering metagenomic contigs			
cutadapt removes adapters from your reads			
★ 13	FRC	FRC	De Novo Assembly Evaluation Tool
★ 40	genmod	genmod	Inheritance in family studies
★ 348	MultiQC	MultiQC	Summarise results across samples
★ 6	NGI-ChIPseq	NGI-ChIPseq	ChIP-seq analysis pipeline

# Acknowledgements

## Phil Ewels

✉ phil.ewels@scilifelab.se

⌚ ewels

🐦 tallphil

### Thanks to:

Max Käller

Olga Vinnere Pettersson

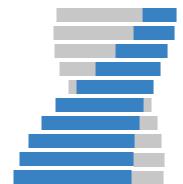
NGI Sweden

support@ngisweden.se

<http://ngisweden.scilifelab.se>

<http://opensource.scilifelab.se>

**SciLifeLab**



**NGI** Stockholm