NGI Sweden

Next Generation Sequencing at the National Genomics Infrastructure

SciLifeLab

NGI stockholm

Phil Ewels
phil.ewels@scilifelab.se

Introduction to Bioinformatics Using NGS Data
Umeå, 2018-11-14

Overview

National Genomics Infrastructure

Sequencing Technologies

Sequencing Applications

Bioinformatics at the NGI

SciLifeLab

NGI stockholm

The National Genomics Infrastructure

SciLifeLab NGI

SciLifeLab

Research Programs

Technology Platforms

- National Genomics Infrastructure
- Proteomics
- Metabolomics
- Single-Cell Biology
- Cellular & Molecular Imaging
- Molecular Structure
- Chemical Biology
- Genome Engineering
- Diagnostic Development
- Drug Discovery & Development

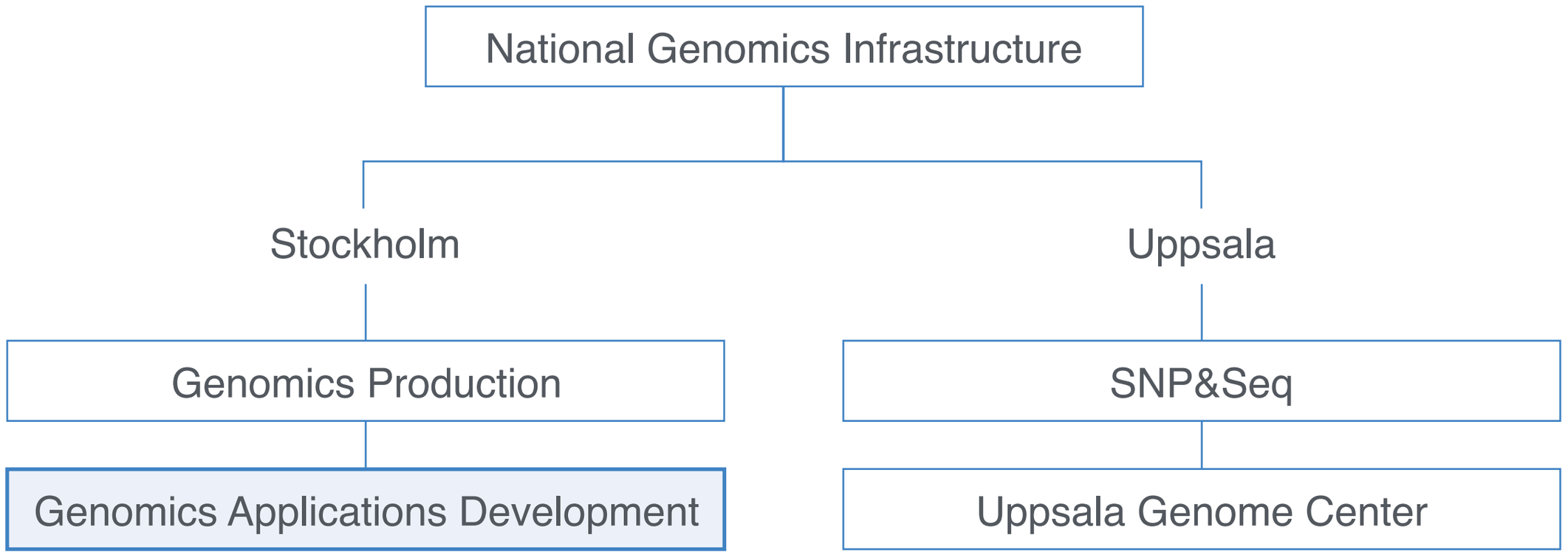
National Bioinformatics Infrastructure

Data Office

SciLifeLab

NGI stockholm

SciLifeLab NGI



SciLifeLab

NGI stockholm

SciLifeLab NGI

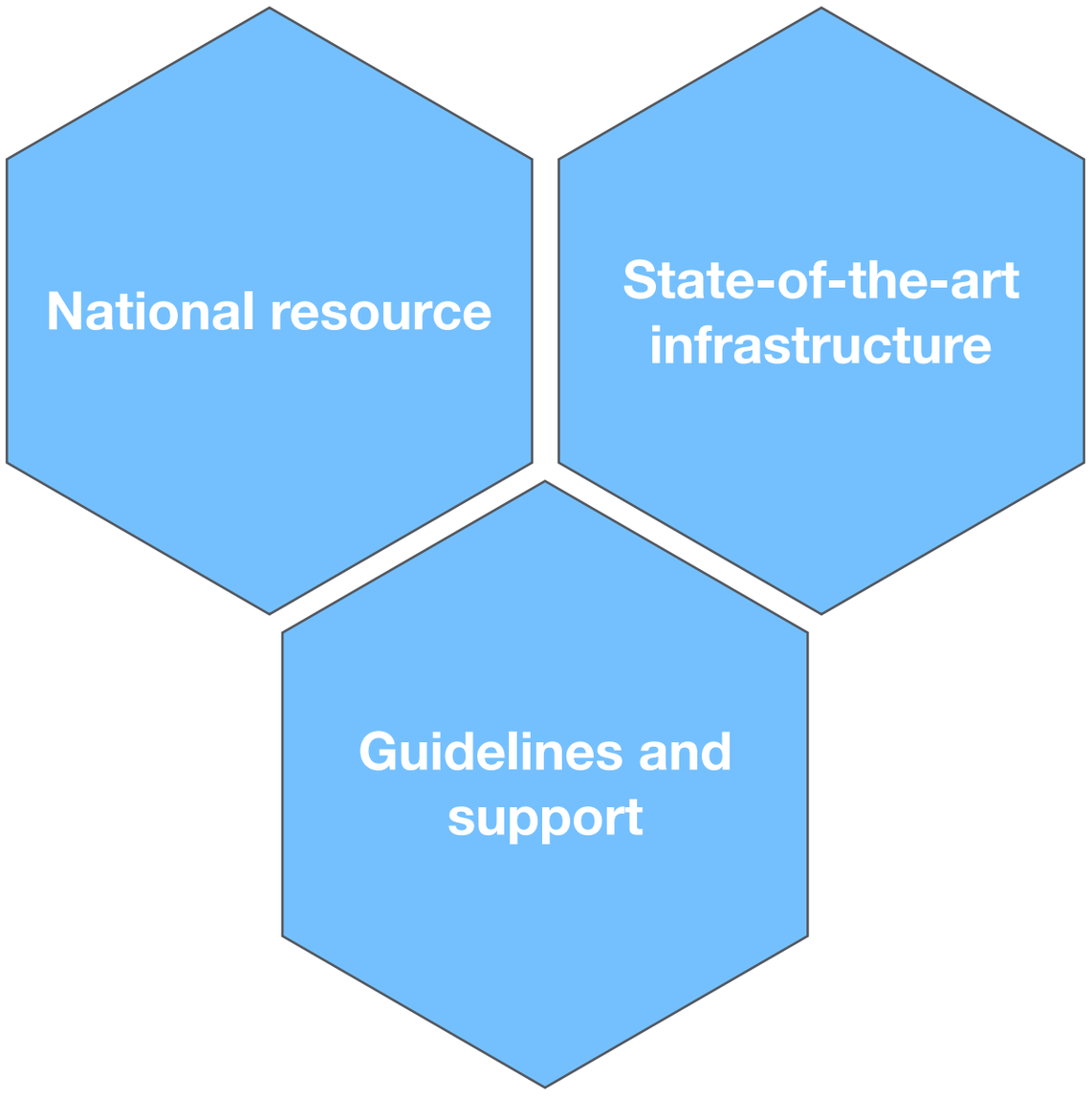
GANATIONALCTAC
ATCAGENOMICSGT
INFRASTRUCTURE

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 stockholm

SciLifeLab NGI



GANATIONALCTAC
ATCAGENOMICSGT
INFRASTRUCTURE

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

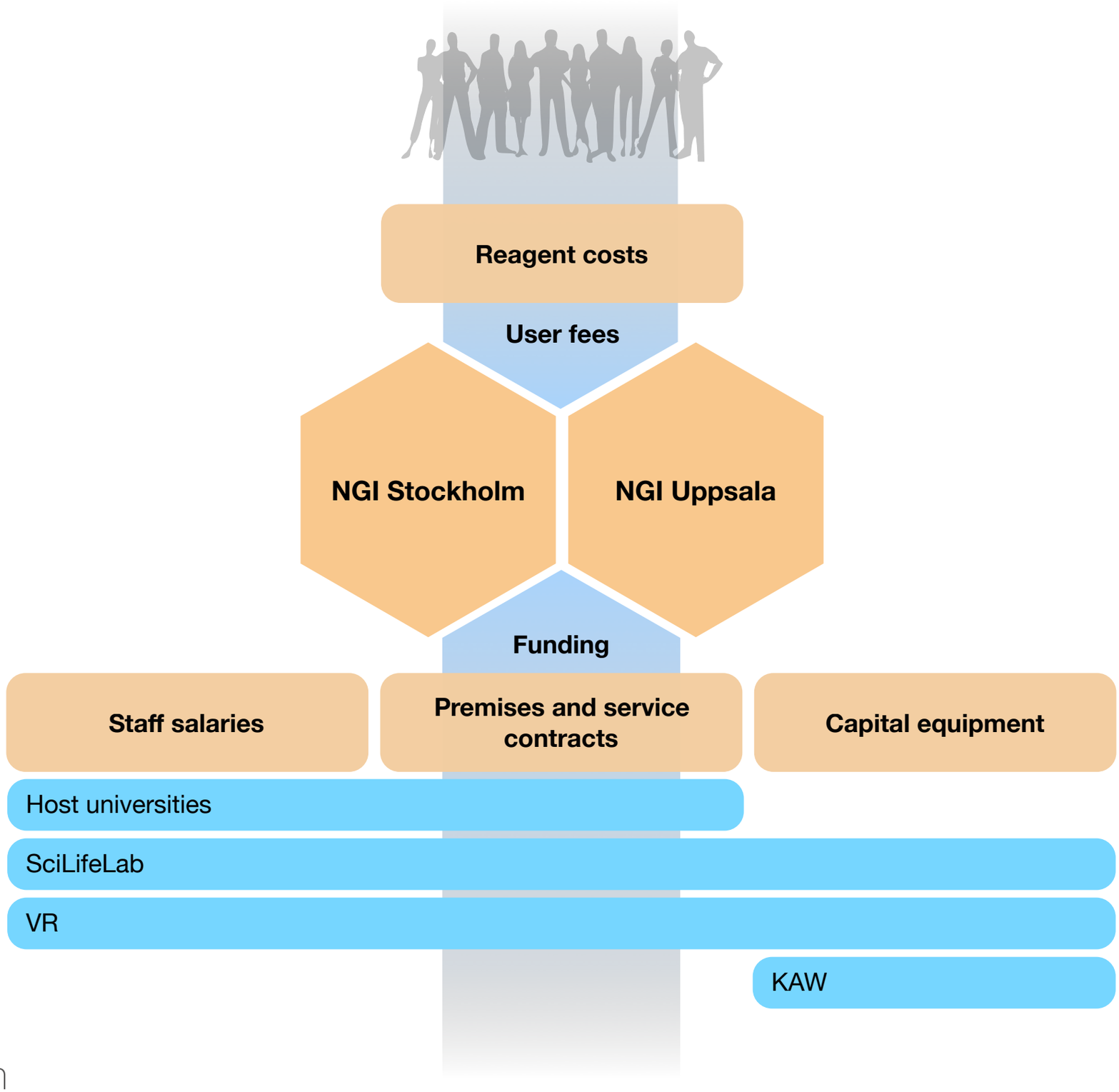
SciLifeLab
NGI stockholm

NGI Organisation

SciLifeLab

NGI stockholm

NGI Organisation



SciLifeLab

NGI stockholm

Project timeline

Scientific support and project consultation

Sample QC

Library preparation, Sequencing, Genotyping

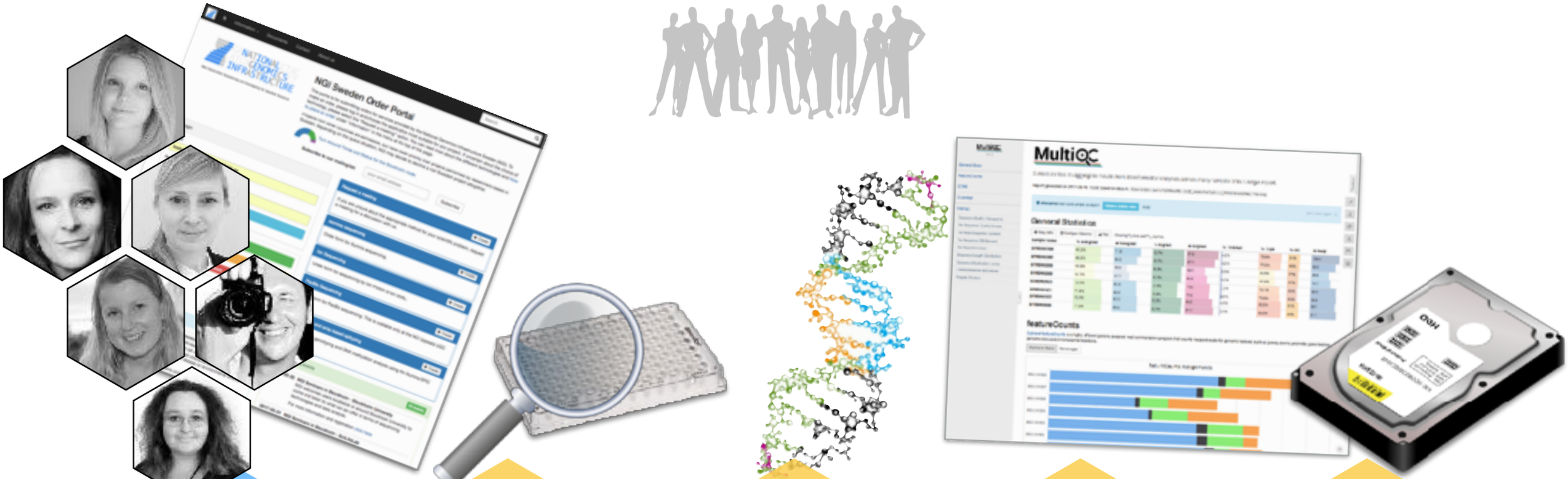
Data processing and primary analysis

Data delivery

SciLifeLab

NGI stockholm

Project timeline



Scientific support and project consultation

Sample QC

Library preparation, Sequencing, Genotyping

Data processing and primary analysis

Data delivery

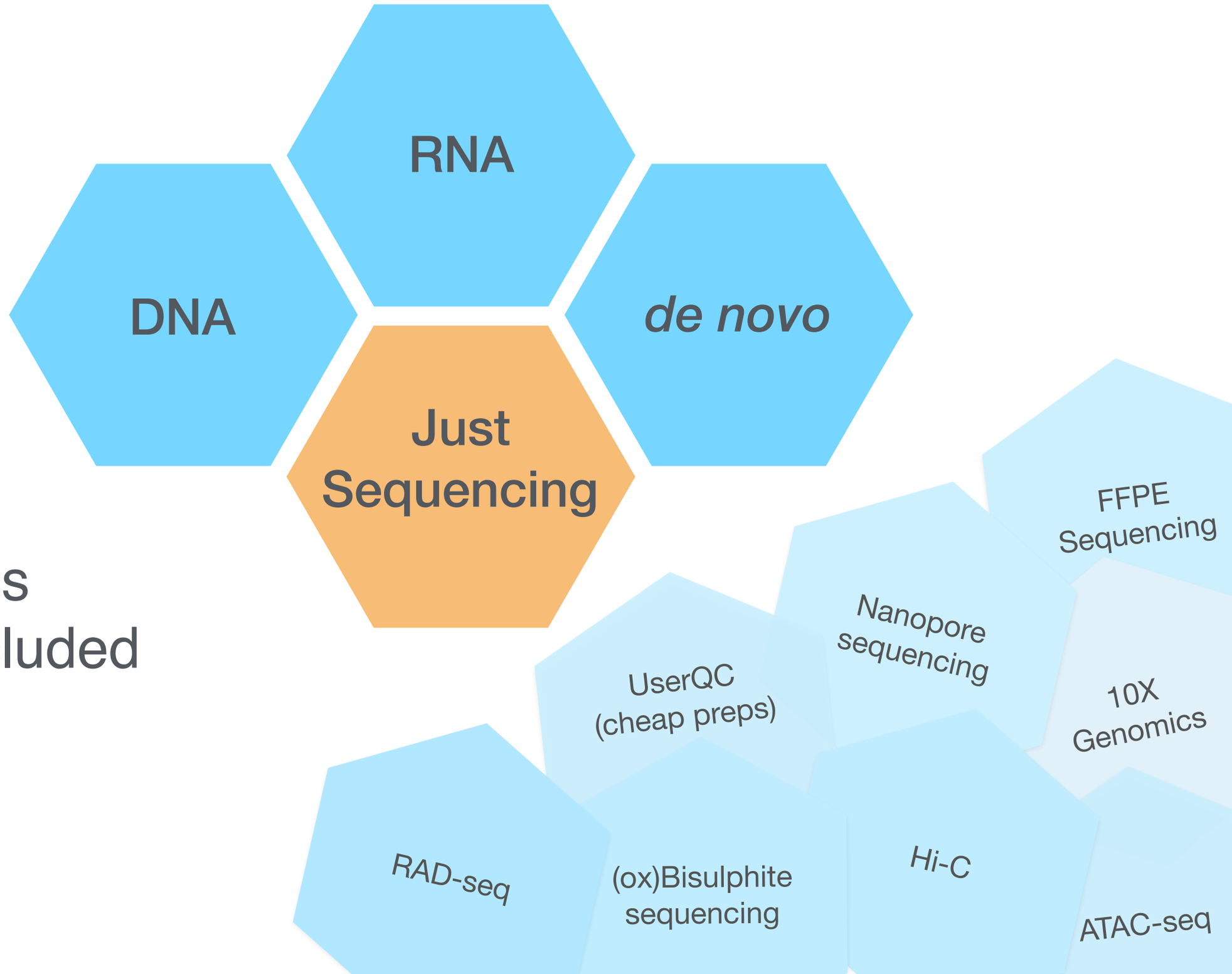
SciLifeLab

NGI stockholm

SWEDAC ACKREDITERING

Ackred. nr 1850
Provning
ISO/IEC 17025

Methods offered at NGI



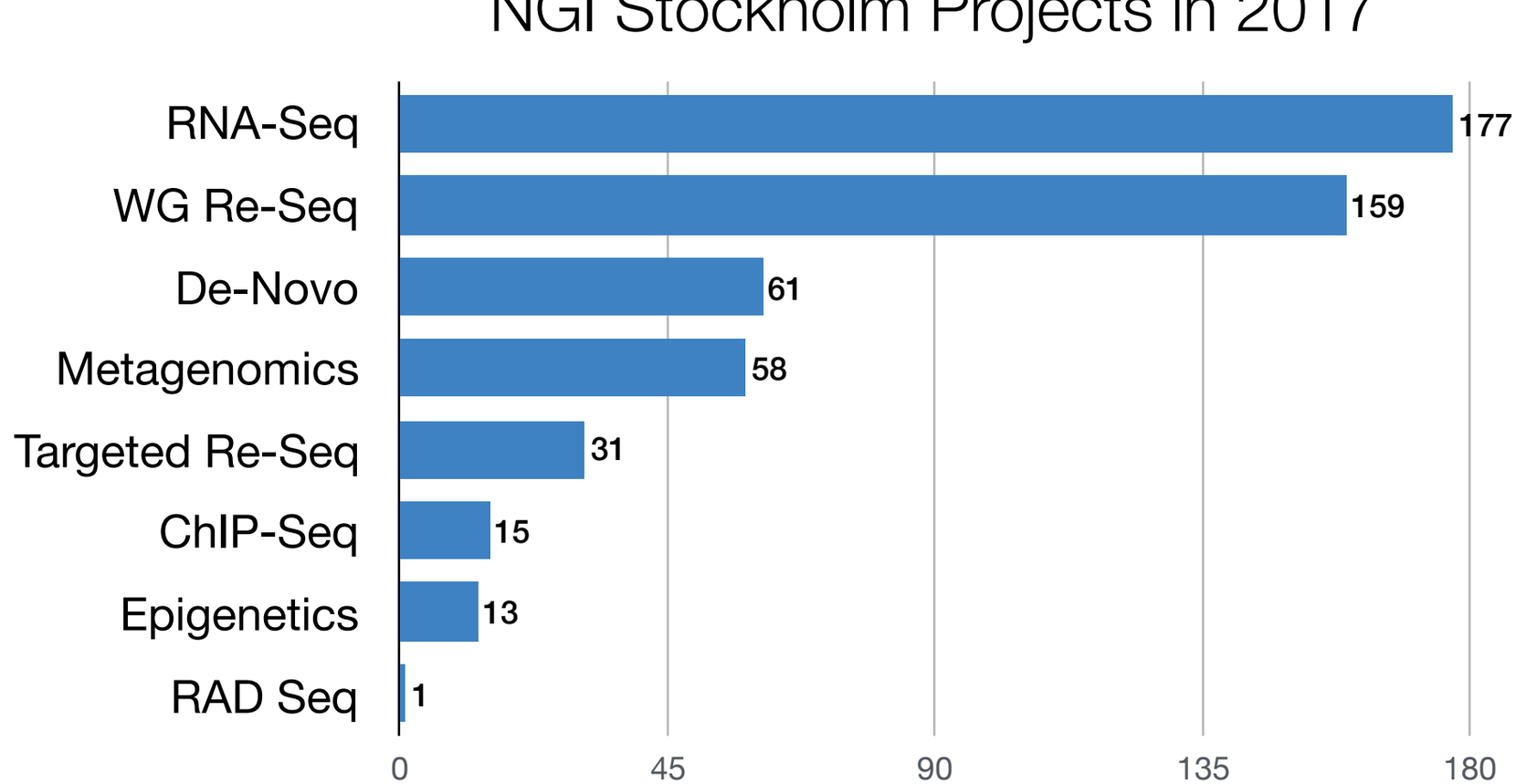
Data analysis pipelines included

SciLifeLab

NGI stockholm

NGI Stockholm

- RNA-seq is the most common project type

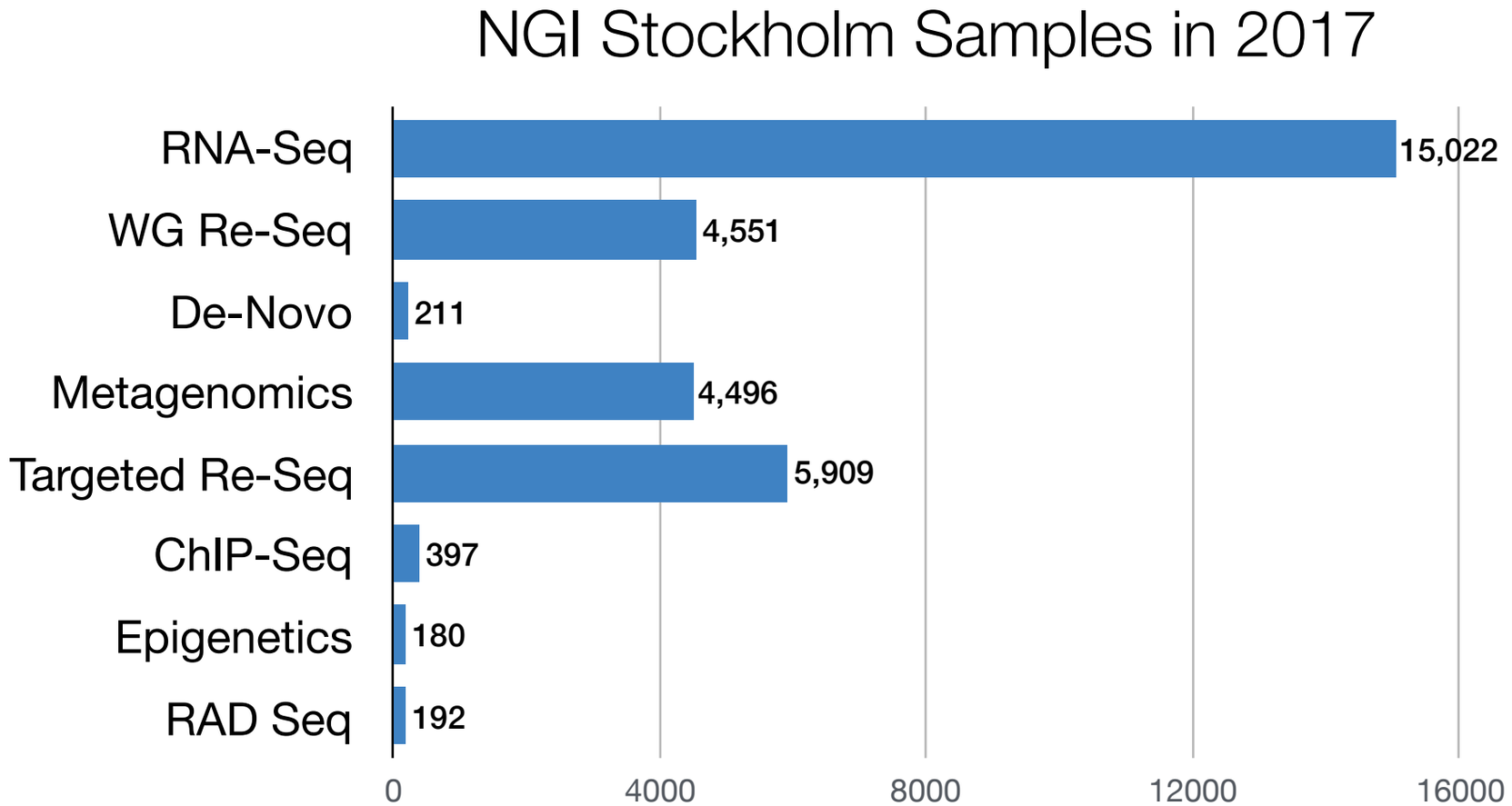


SciLifeLab

NGI stockholm

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



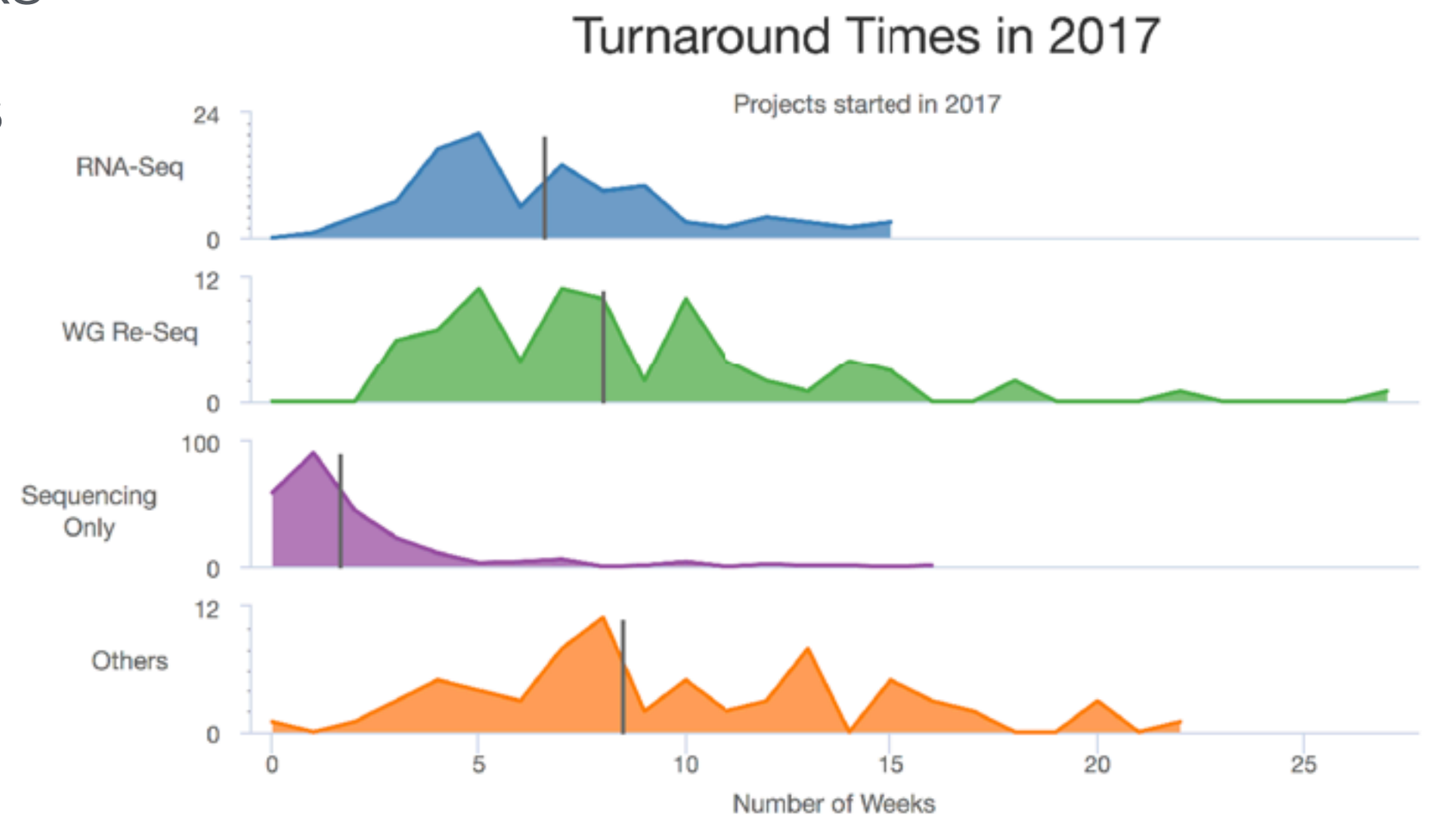
SciLifeLab

NGI stockholm

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



SciLifeLab

NGI stockholm

Sequencing Technologies

Sequencing Types

Illumina

PacBio

Oxford Nanopore

Ion Torrent

SciLifeLab

NGI stockholm


illumina®

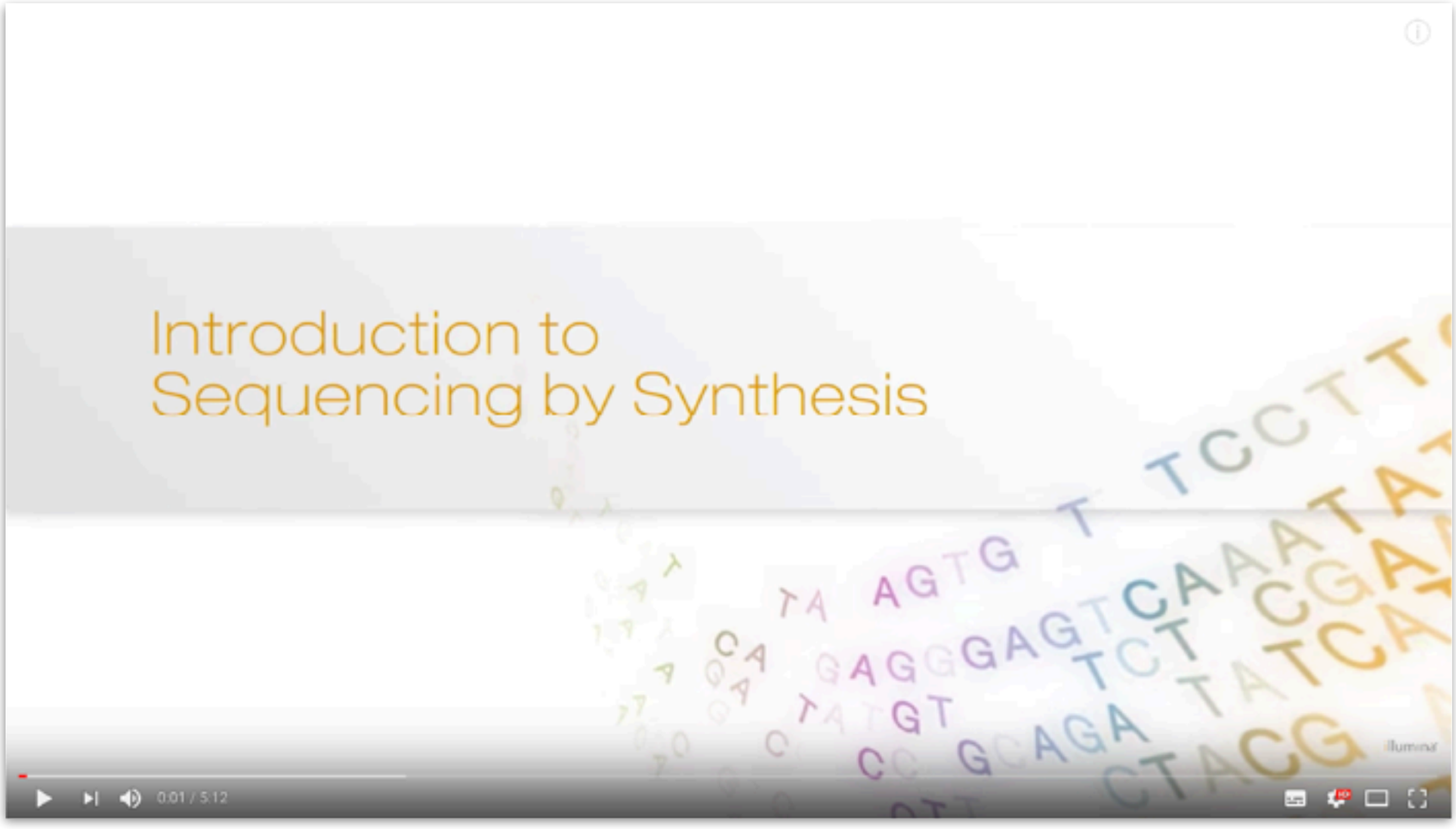
Illumina Sequencing

- 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

SciLifeLab

NGI stockholm

Illumina Sequencing



https://youtu.be/fCd6B5HRaZ8

SciLifeLab

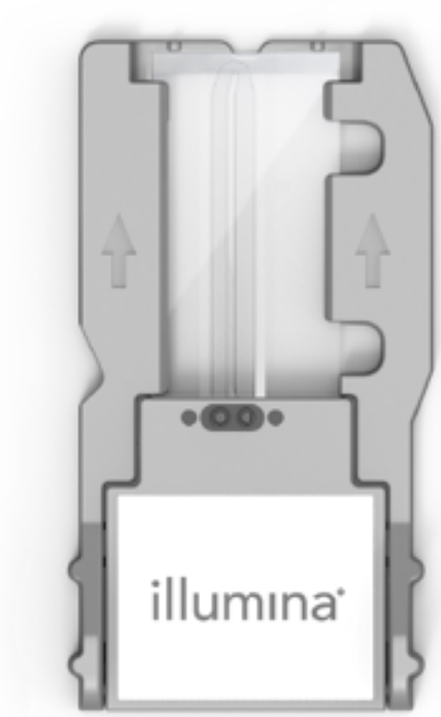
NGI stockholm

Illumina iSeq 100

SciLifeLab

NGI stockholm

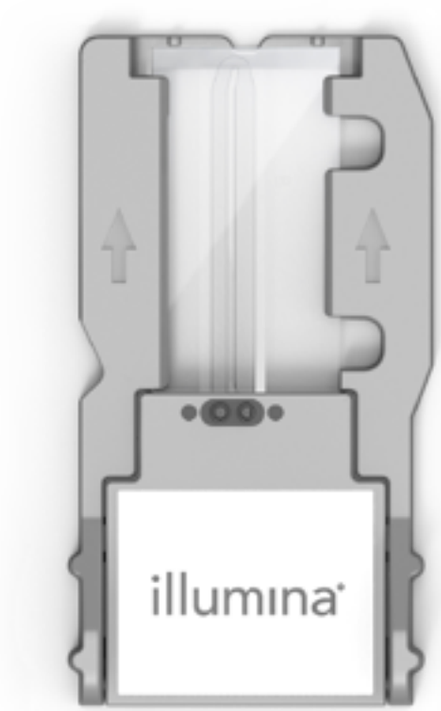
Illumina MiniSeq 100



SciLifeLab

NGI stockholm

Illumina MiSeq



SciLifeLab

NGI stockholm

Illumina NextSeq

SciLifeLab

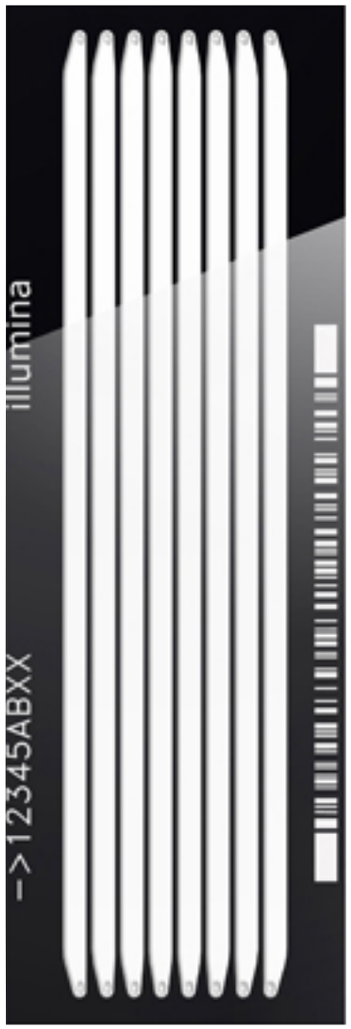
NGI stockholm

Illumina HiSeq 2500

SciLifeLab

NGI stockholm

Illumina HiSeq 3000



SciLifeLab

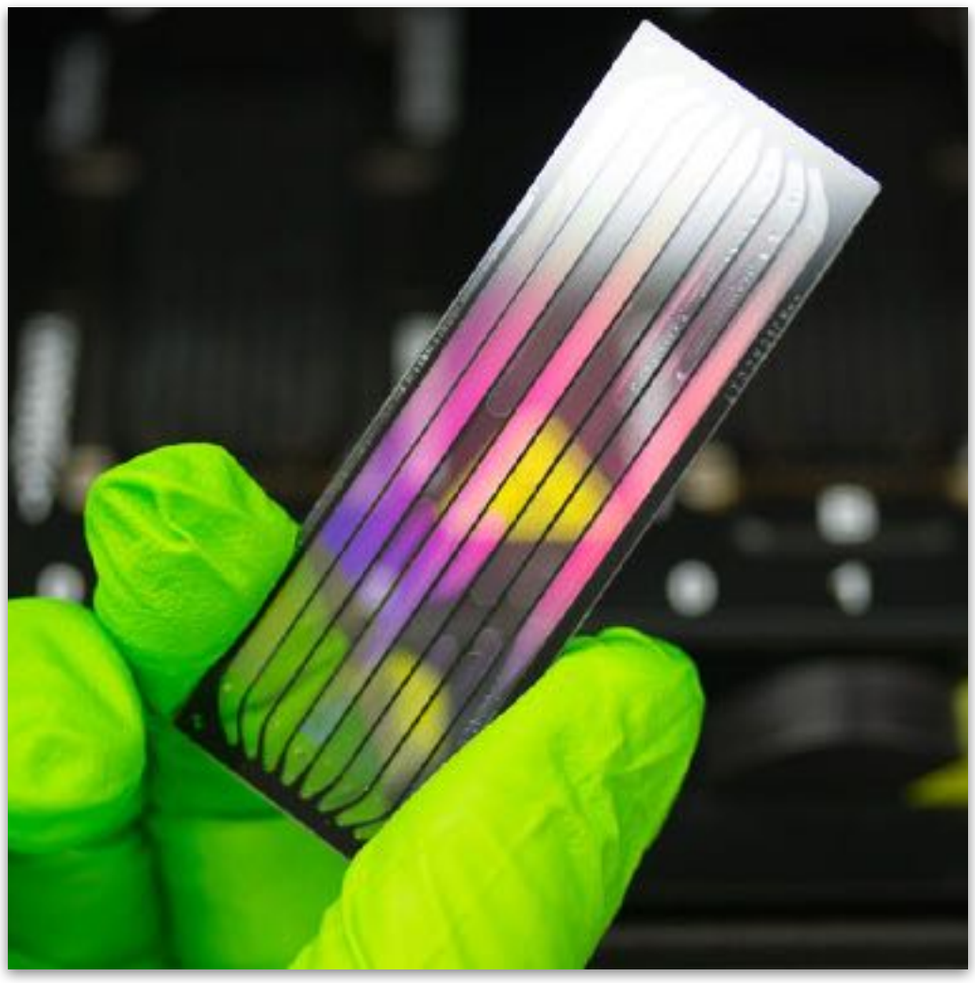
NGI stockholm

Illumina HiSeq 4000

SciLifeLab

NGI stockholm

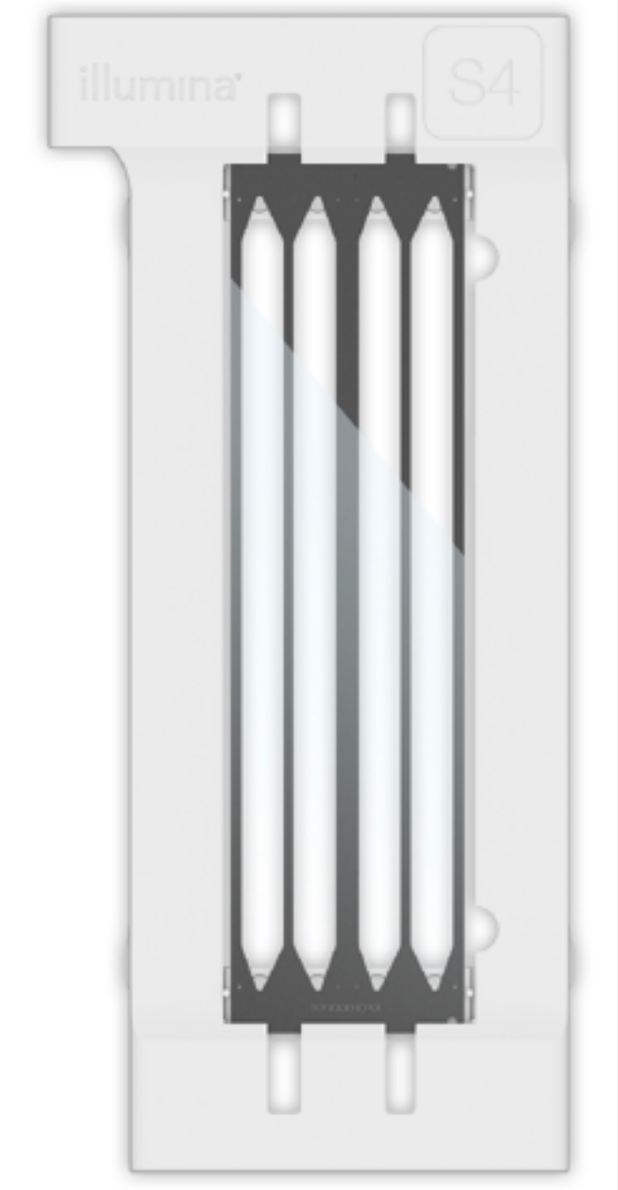
Illumina HiSeq X



SciLifeLab

NGI stockholm

Illumina NovaSeq 6000



SciLifeLab

NGI stockholm

Illumina at NGI

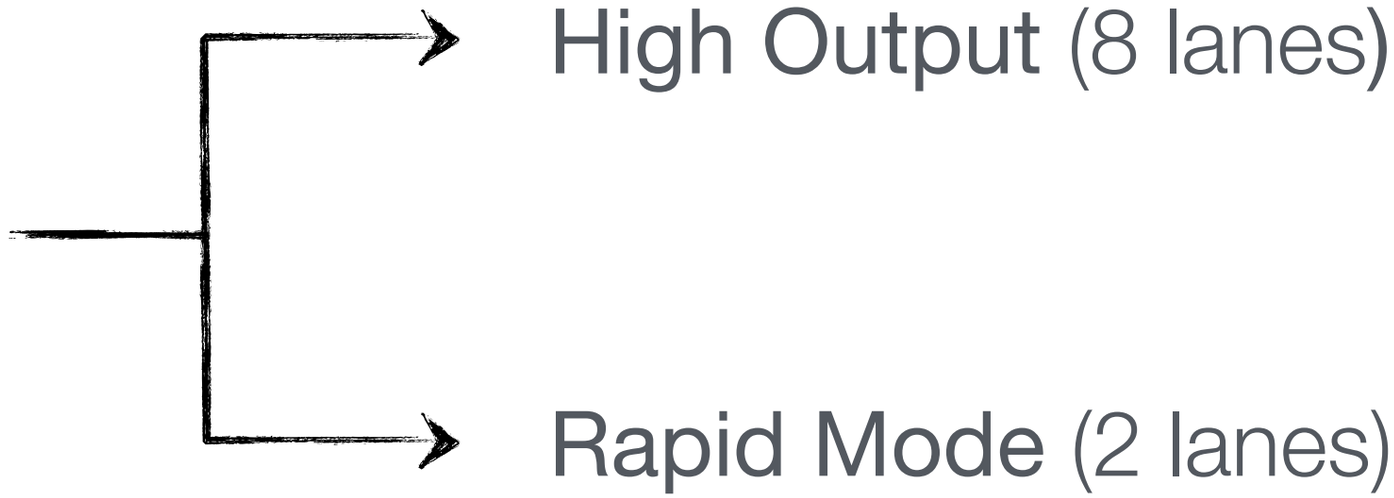
iSeq 100	Coming soon to NGI Uppsala Small cheap runs
MiSeq	Small runs, long reads (2x300bp)
HiSeq 2500	Primary machine for most of NGI's history
HiSeq X	Cheap, high throughput Only allowed to run WGS with > 15X coverage
NovaSeq 6000	Newest machine, both Stockholm & Uppsala Will eventually replace HiSeq 2500

SciLifeLab

NGI stockholm

Illumina at NGI

- iSeq 100
- MiSeq
- HiSeq 2500
- HiSeq X
- NovaSeq 6000



SciLifeLab

NGI stockholm

Illumina at NGI

- iSeq 100
- MiSeq
- HiSeq 2500
- HiSeq X
- NovaSeq 6000

SciLifeLab

NGI stockholm

How to choose

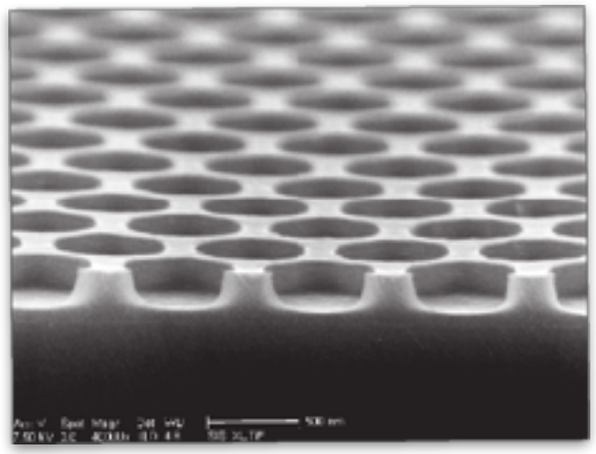
- Number of reads required
  - How many samples, how deeply sequenced?
- Type of reads required
  - Single End / Paired End, length?
- Urgency and cost
  - Sharing flow cells with other users
  - Best price for the project

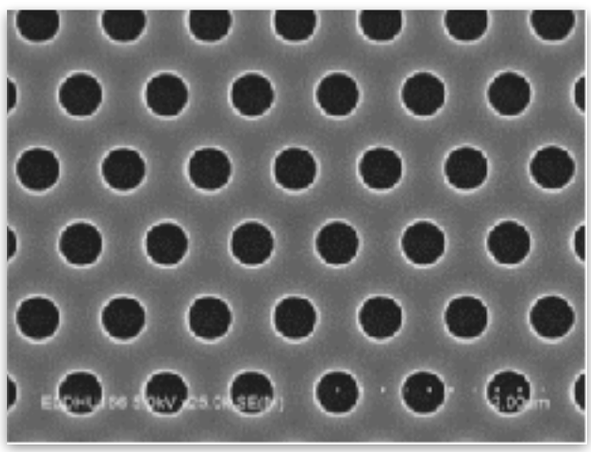
SciLifeLab

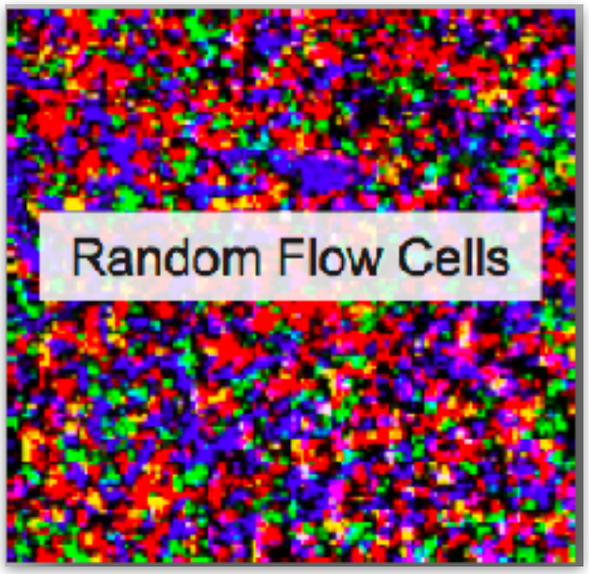
NGI stockholm

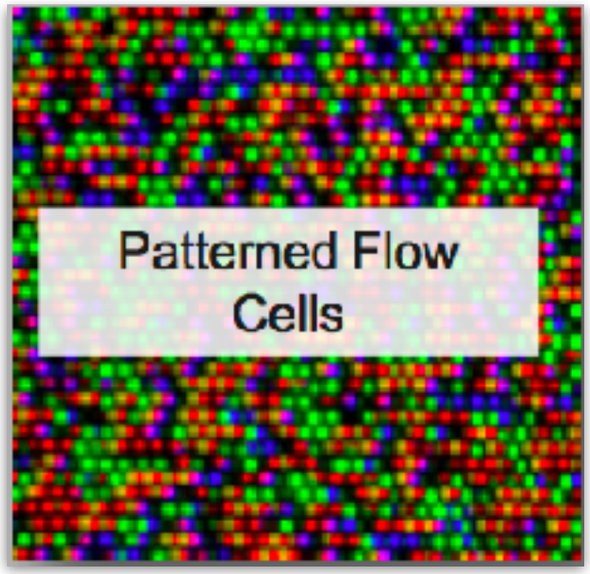
Patterned flow cells

- New type of flow cell
  - HiSeq 4000, HiSeq X, NovaSeq
- Single sequence per well
  - Higher density, more data
- Different side effects
  - Index hopping
  - Duplicate reads









SciLifeLab

NGI stockholm

QCFAIL.com

Articles about common next-generation sequencing problems

Phil Ewels
Simon Andrews

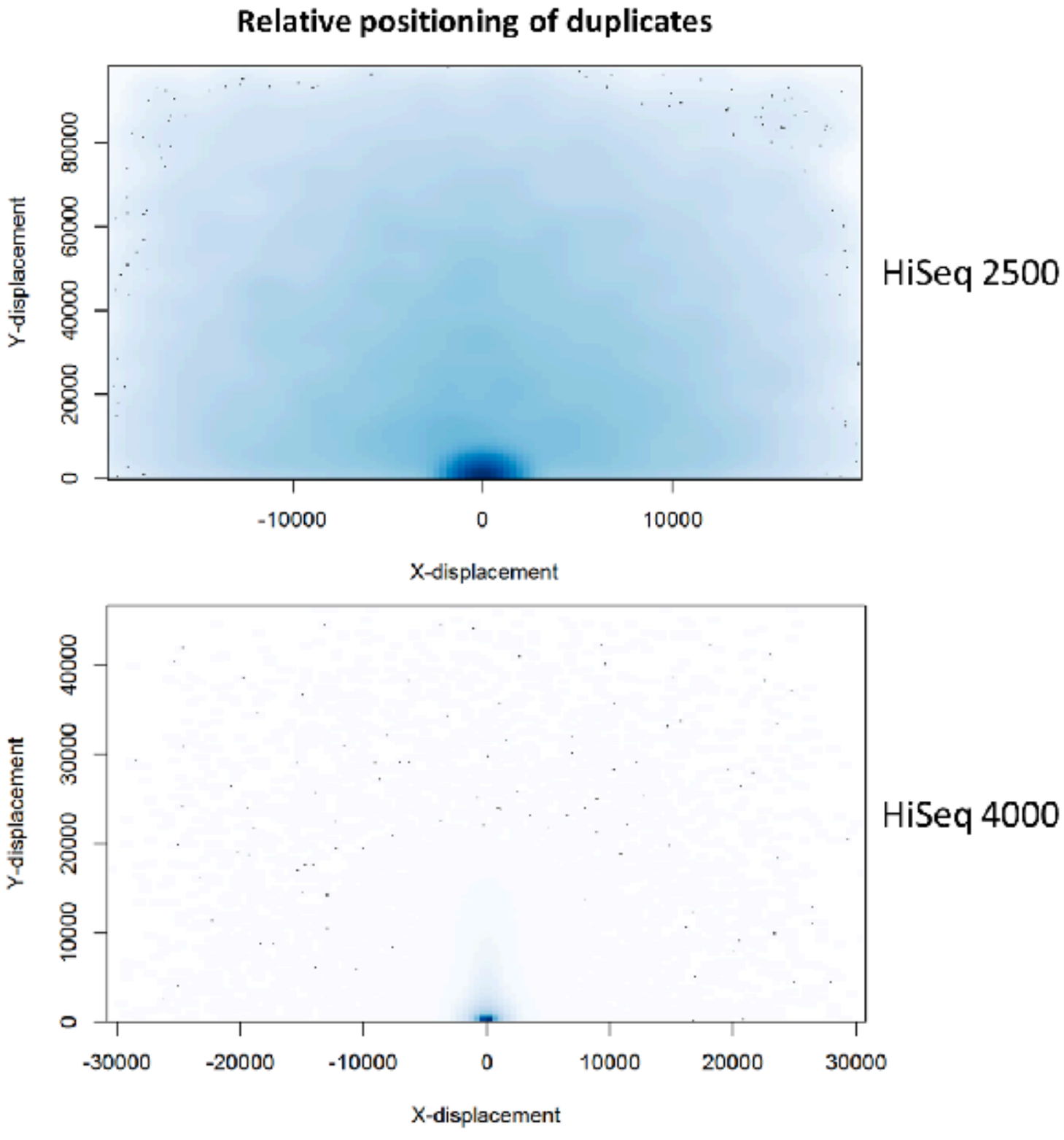
QCFAIL.com

Illumina Patterned Flow Cells Generate Duplicated Sequences

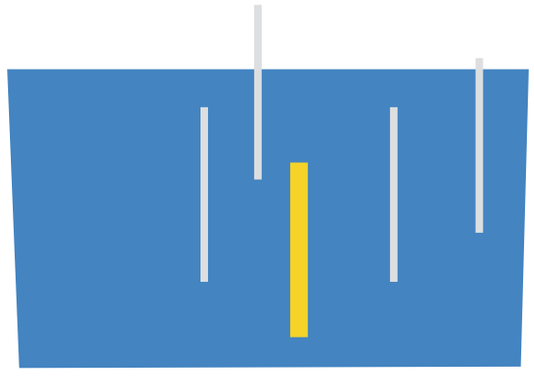
Steven Wingett

https://sequencing.qcfail.com/articles/illumina-patterned-flow-cells-generate-duplicated-sequences/

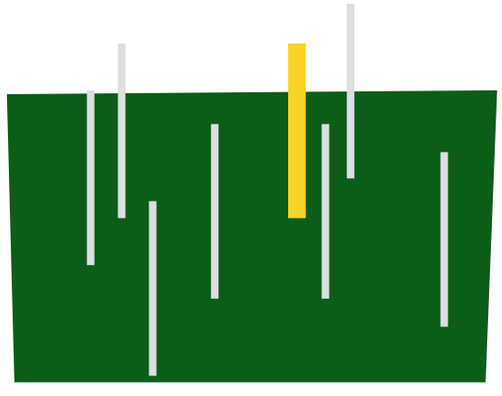
Patterned duplicates



Duplicates on different tiles

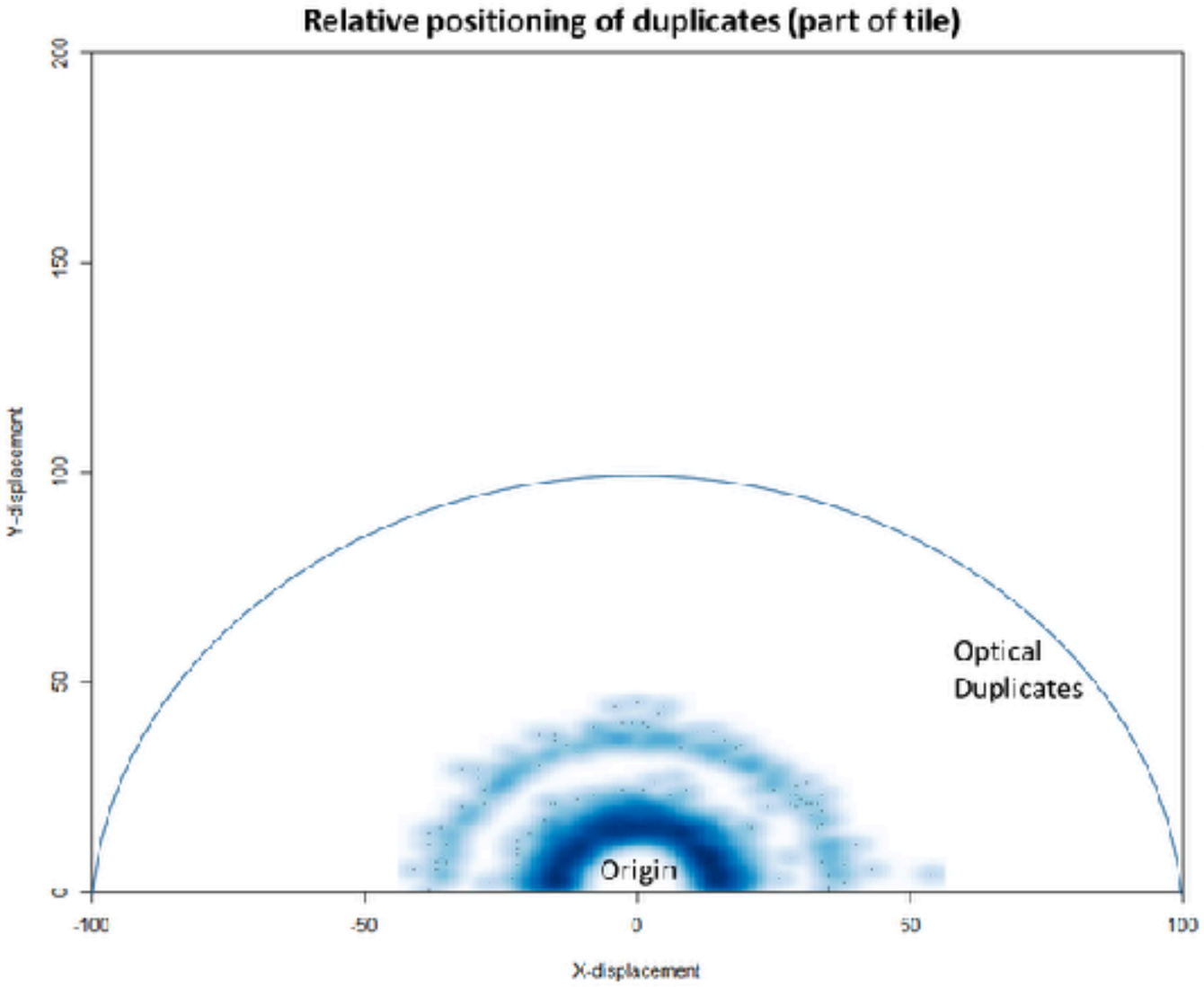


Tile A



Tile B

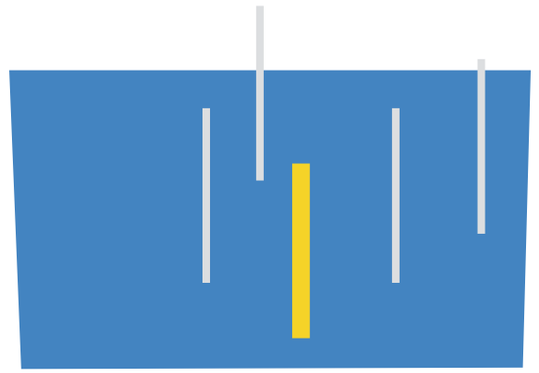
Patterned duplicates



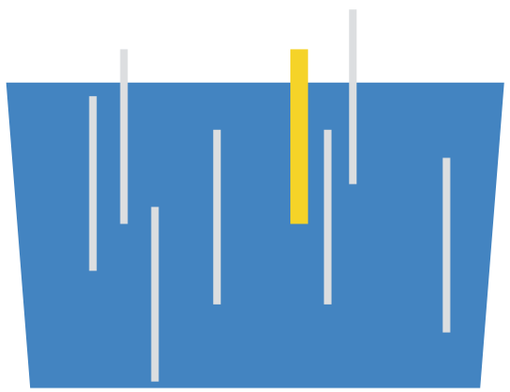
Unpatterned flow cell

Duplicates on the same tile

HiSeq 2500

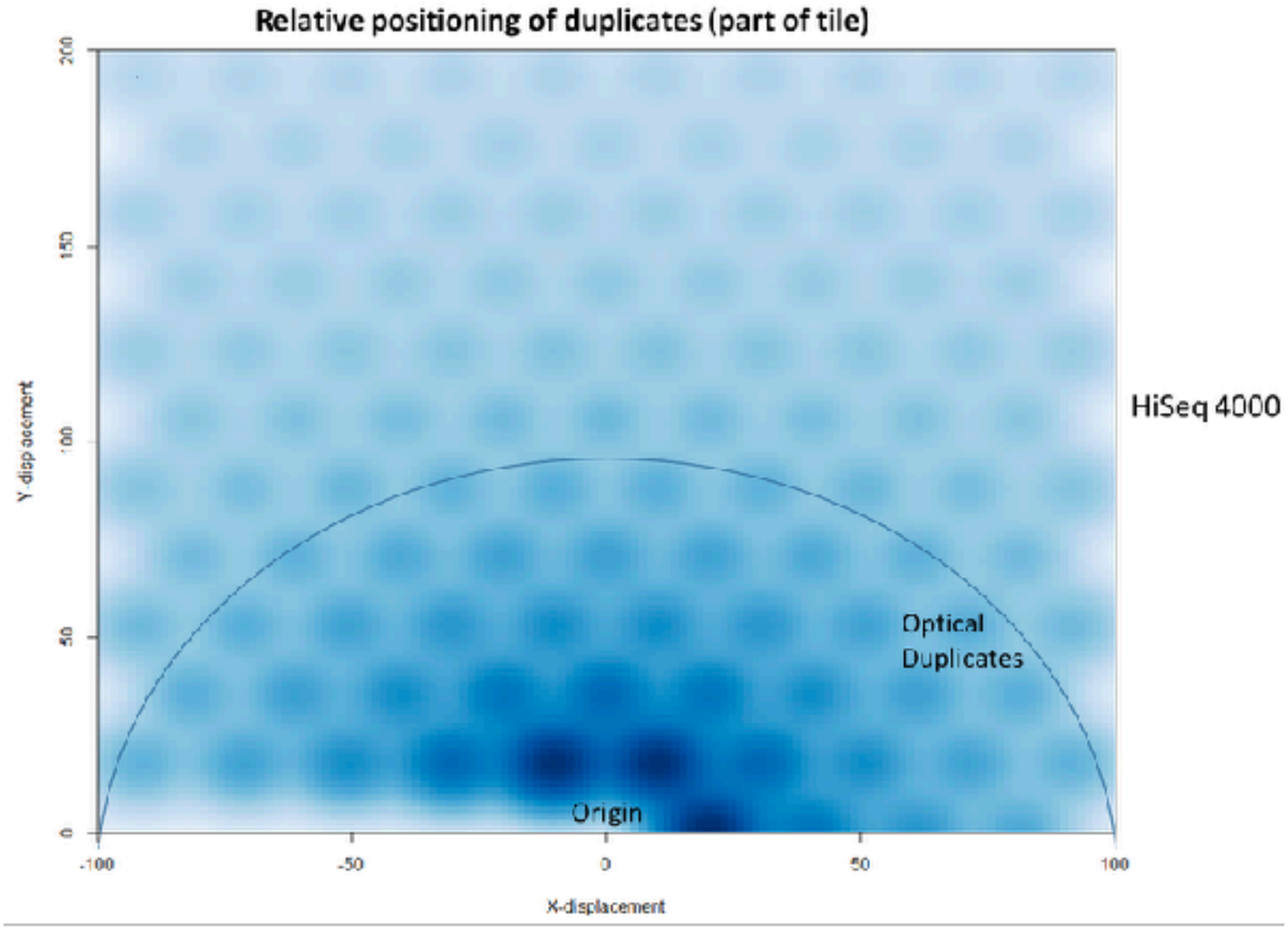


Tile A



Tile A

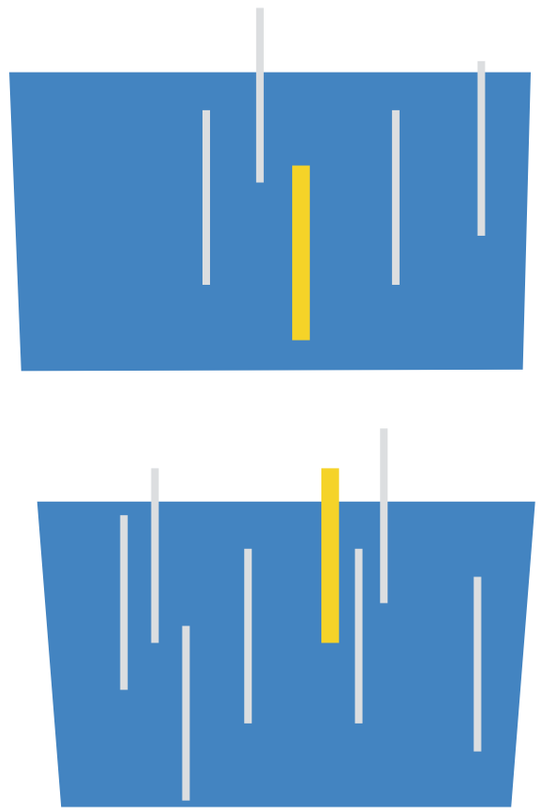
Patterned duplicates



HiSeq 4000

Patterned
flow cell

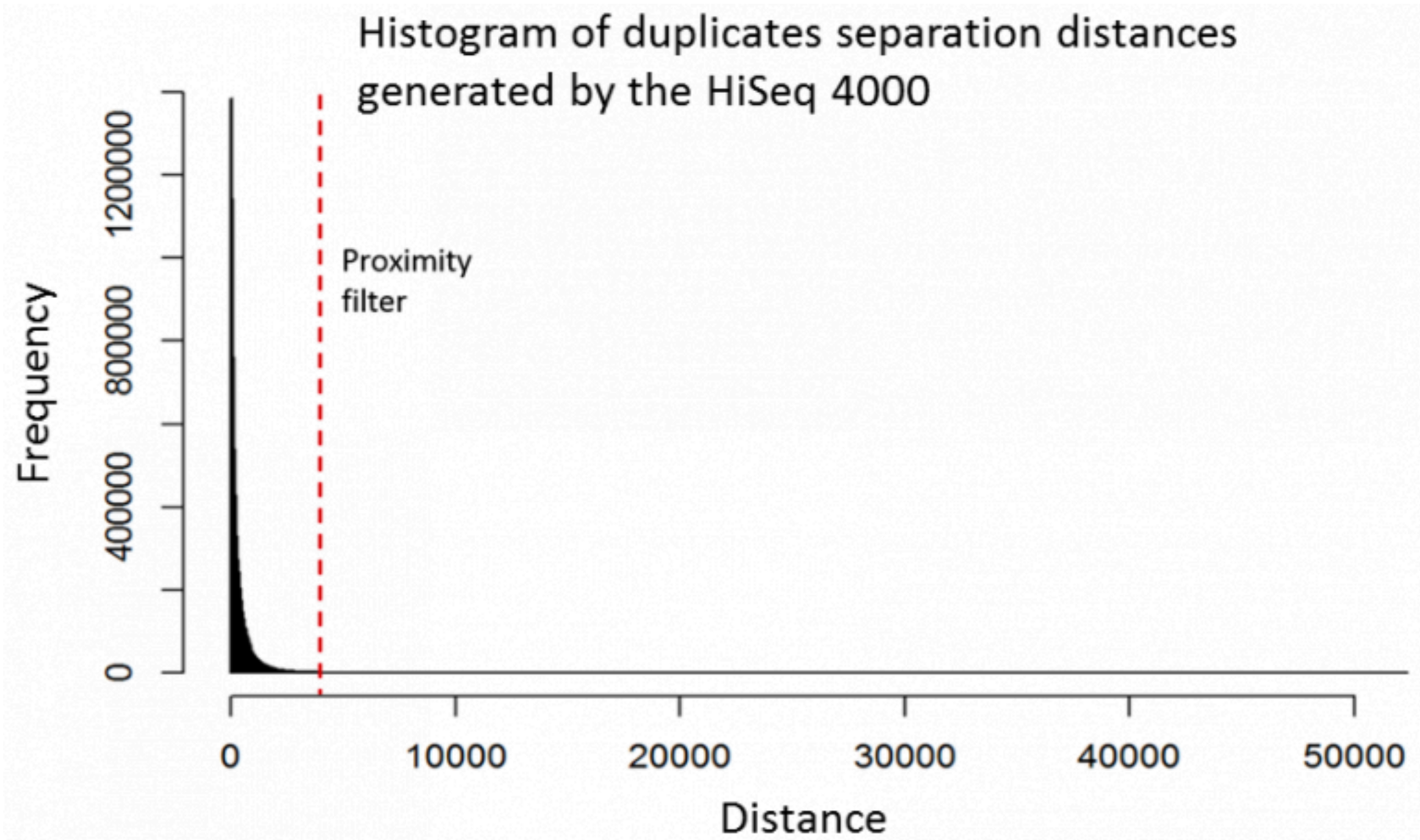
Duplicates on
the same tile



Tile A

Tile A

Patterned duplicates



Patterned duplicates

- Regular duplicate removal works fine
  - Sequence alignment positions should be identical
- Can use Picard MarkDuplicate optical duplicate settings
  - May need to increase the default pixel threshold
- Specialised tools such as `EdinburghGenomics/well_duplicates` work directly with `.bcl` files
  - https://github.com/EdinburghGenomics/well_duplicates

SciLifeLab

NGI stockholm

QCFAIL.com

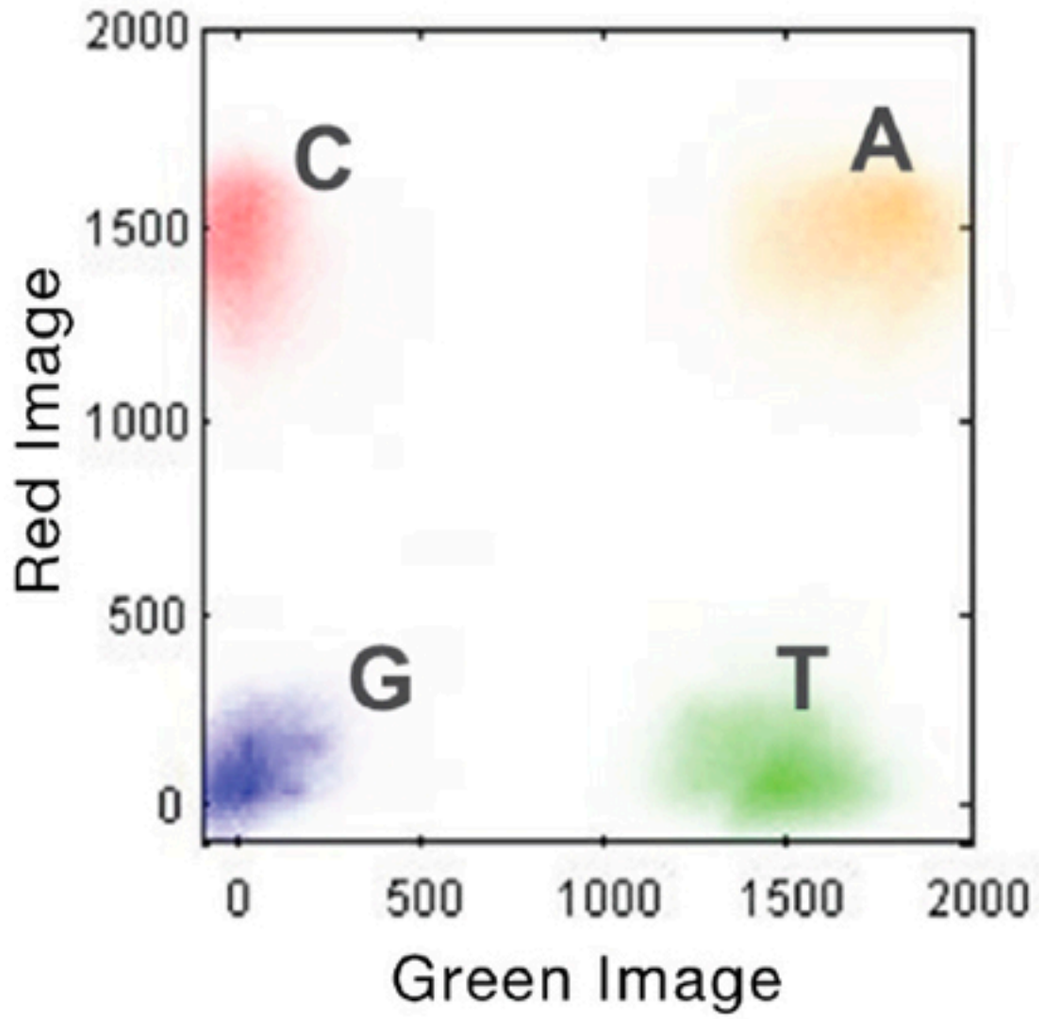
Illumina 2 colour chemistry can overcall high confidence G bases

Simon Andrews

https://sequencing.qcfail.com/articles/illumina-2-colour-chemistry-can-overcall-high-confidence-g-bases/

Colour chemistry

- Older SBS used four different fluorophores
  - One for each nucleotide
- New machines use two
  - Faster and cheaper
  - NextSeq, NovaSeq, iSeq



SciLifeLab

NGI stockholm

Colour chemistry

4-colour chemistry

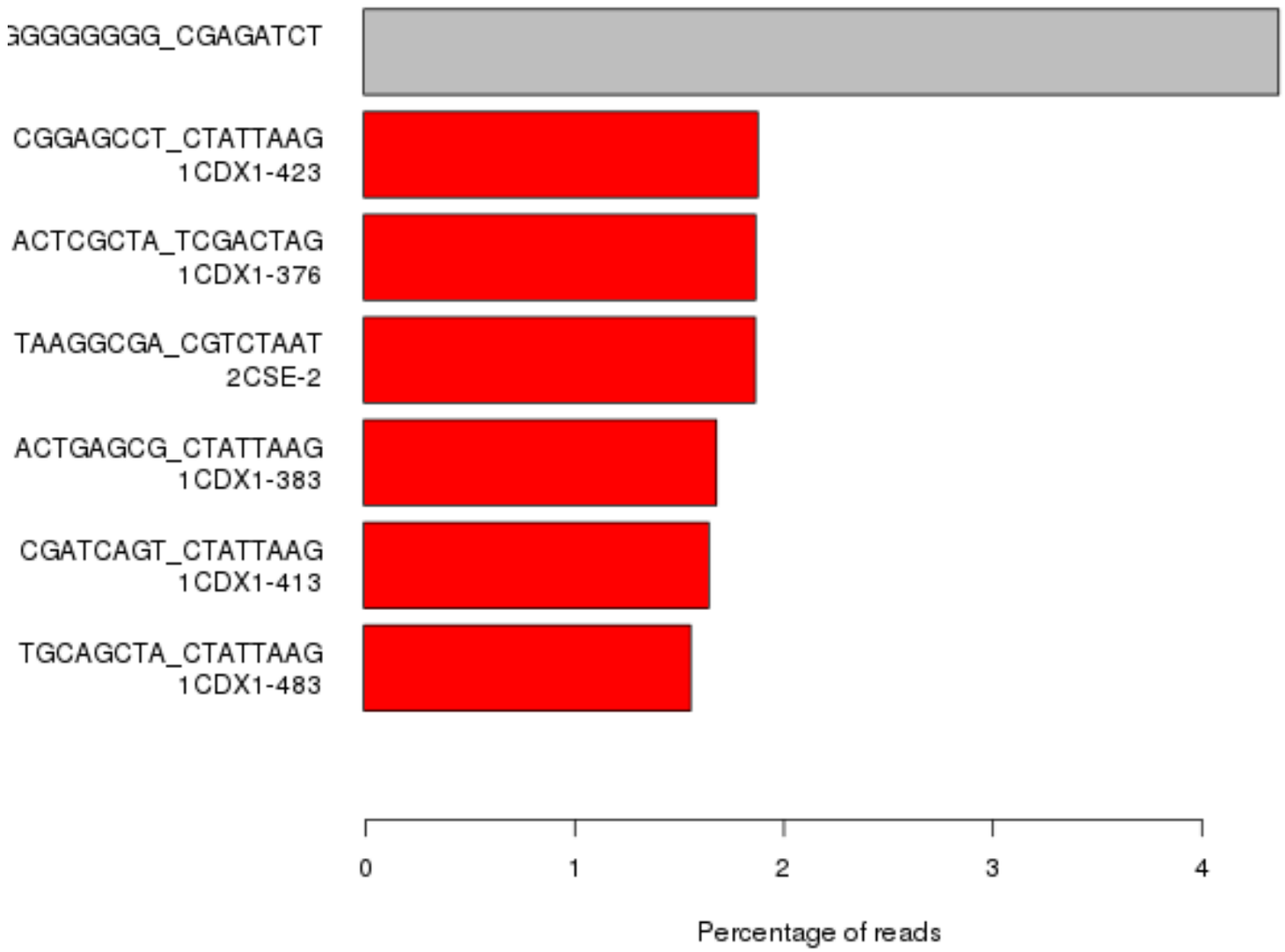
Base	G Filter	A Filter	T Filter	C Filter
G	✅	❌	❌	❌
A	❌	✅	❌	❌
T	❌	❌	✅	❌
C	❌	❌	❌	✅
N	❌	❌	❌	❌

Colour chemistry

2-colour chemistry

Base	Green Filter	Red Filter
G	❌	❌
A	✅	✅
T	✅	❌
C	❌	✅
N	❌	❌

Colour chemistry



Index	Count	Total %
AGCGATAG+ATTGTTGC	3 408 780	3.22%
GGGGGGGG+GGGGGGGG	3 051 380	2.89%
TCCGCGAA+ATTGTTGC	2 844 020	2.69%
GAGATTCC+ATTGTTGC	2 602 140	2.46%
TAATGCGC+ATTGTTGC	2 578 560	2.44%
TCTCGCGC+ATTGTTGC	2 405 340	2.28%
ATTACTCG+ATTGTTGC	2 281 500	2.16%
CGCTCATT+ATTGTTGC	2 279 700	2.16%
ATTCAGAA+ATTGTTGC	2 252 180	2.13%
CTGAAGCT+ATTGTTGC	2 249 940	2.13%
TCCGGAGA+CGTTTACT	2 236 880	2.12%
GGGGGGGG+TGTTTCCC	2 160 120	2.04%
CGGCTATG+ATTGTTGC	2 132 900	2.02%
GAATTCGT+ATTGTTGC	2 080 700	1.97%
TCCGGAGA+ATTGTTGC	2 070 820	1.96%
ATTACTCG+CGTTTACT	2 017 160	1.91%
TGCGATTG+TTTGTGGC	1 730 080	1.64%
ATTCAGAA+CGTTTACT	1 640 140	1.55%
GGGGGGGG+TTTTGCCT	1 475 160	1.4%
CGCTCATT+CGTTTACT	1 408 140	1.33%

SciLifeLab

NGI stockholm

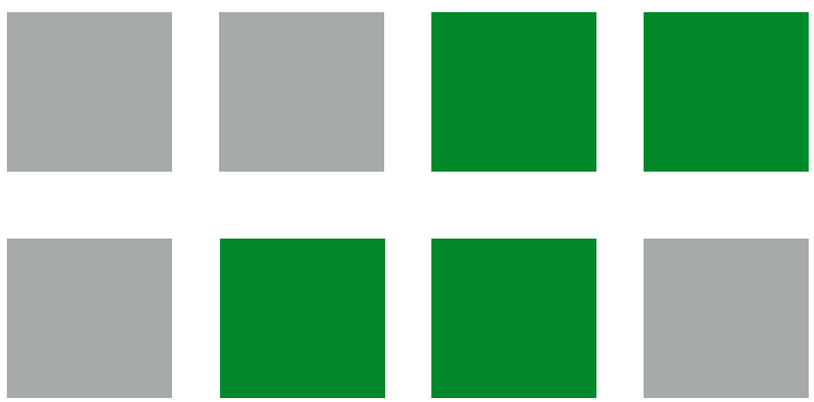
Colour chemistry

Base	Green Filter	Red Filter
G	❌	❌
A	✅	✅
T	✅	❌
C	❌	✅
N	❌	❌

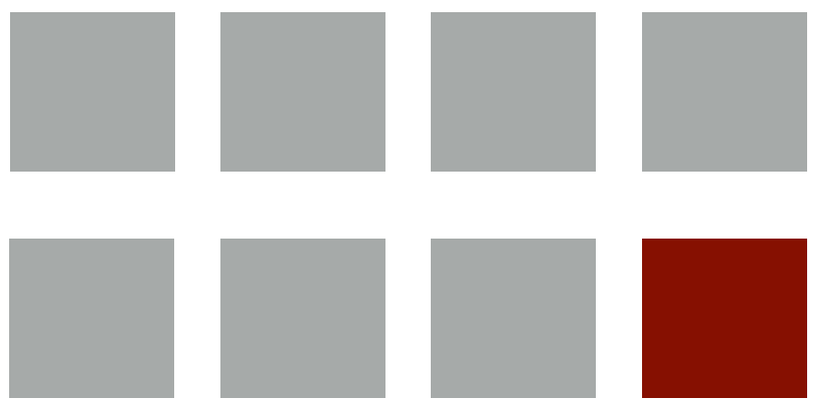
Sample 1 GGTT

Sample 2 GTTC

Green



Red



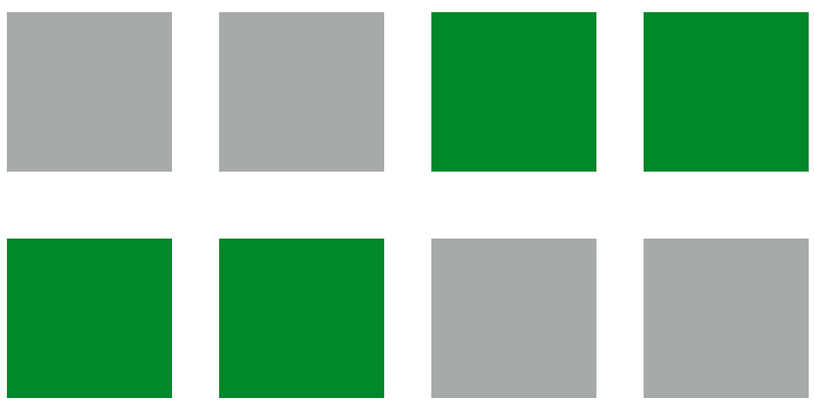
Colour chemistry

Base	Green Filter	Red Filter
G	❌	❌
A	✅	✅
T	✅	❌
C	❌	✅
N	❌	❌

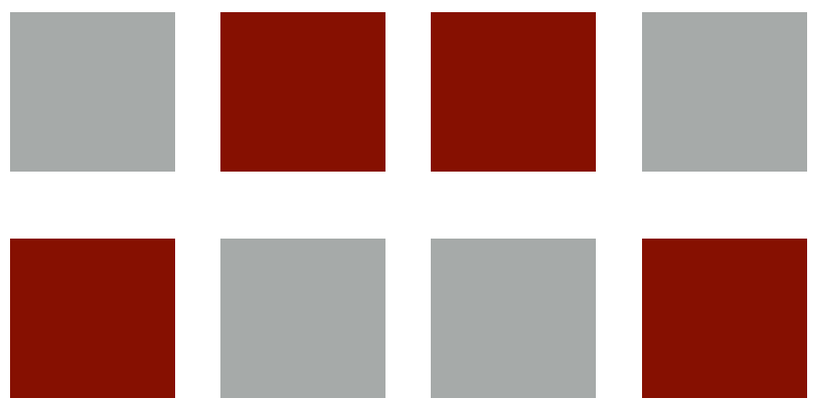
Sample 1 GCAT

Sample 2 ATGC

Green



Red



Colour chemistry

- Poor quality reads may show up as G instead of N
  - For example, missing bases from short insert sizes
- Trimming tools such as cutadapt now updated to handle this
- Careful colour balancing of indexes can avoid problems with deduplication
  - This isn't new - it's just more sensitive than before
- Check the illumina recommendations:
  - http://emea.support.illumina.com/downloads/index-adapters-pooling-guide-1000000041074.html?langsel=/se/

Balanced pooling

- New NovaSeqs make the S4 the best option
- Proper sample concentration normalisation more important than ever
  - Big (expensive) flow cells = high stakes!
- Our plans: always improving library quantitation and normalisation
  - Constant benchmarking of quant tools
  - More accurate automation

SciLifeLab

NGI stockholm

illumina®

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)

SciLifeLab

NGI stockholm

PacBio

https://youtu.be/NHCJ8PtYCFc

SciLifeLab

NGI stockholm

PacBio RS II

SciLifeLab

NGI stockholm

PacBio Sequel

SciLifeLab

NGI stockholm

PacBio Sequencing

- Long reads are excellent for *de-novo* genome assembly, haplotype phasing 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

SciLifeLab

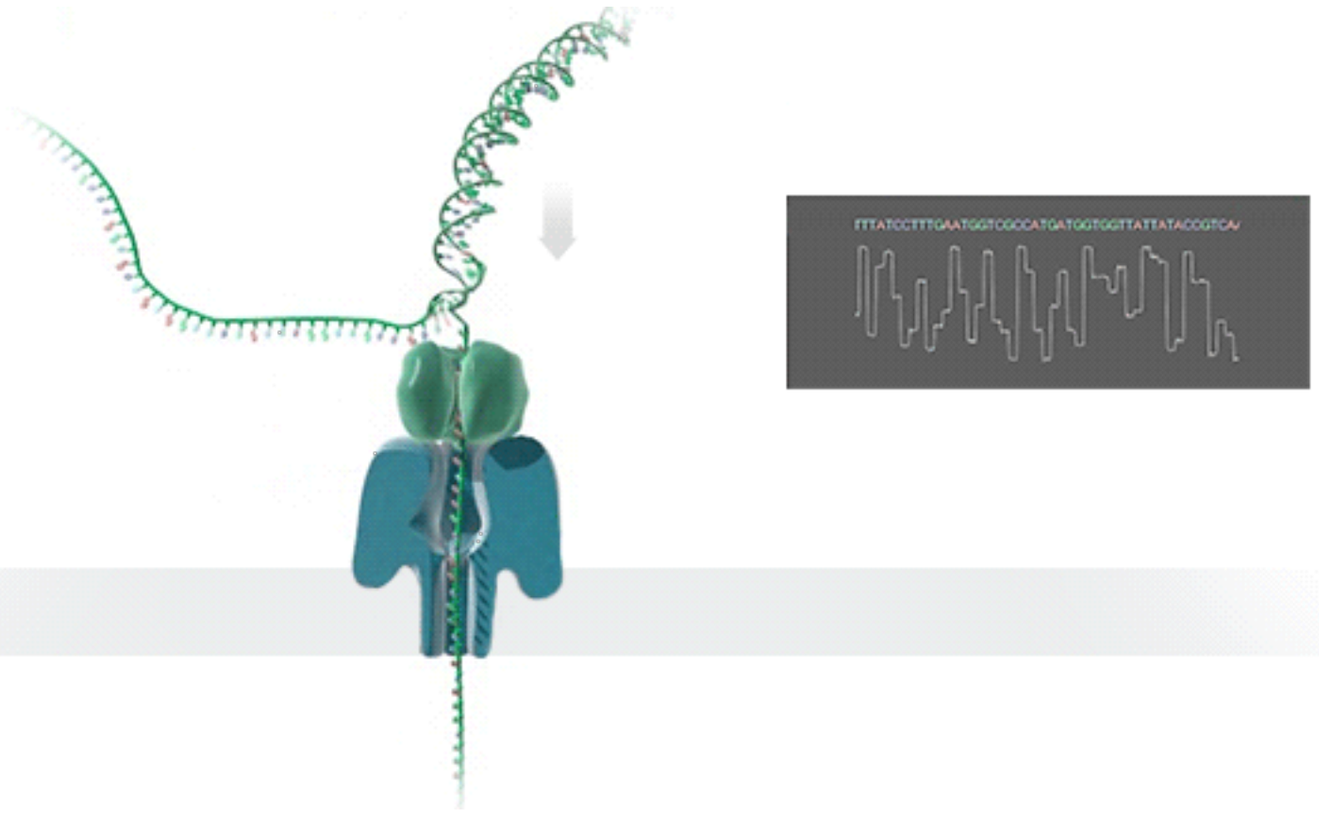
NGI stockholm


Oxford
NANOPORE
Technologies

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



SciLifeLab

NGI stockholm

MinION

SciLifeLab

NGI stockholm

MinION

SciLifeLab

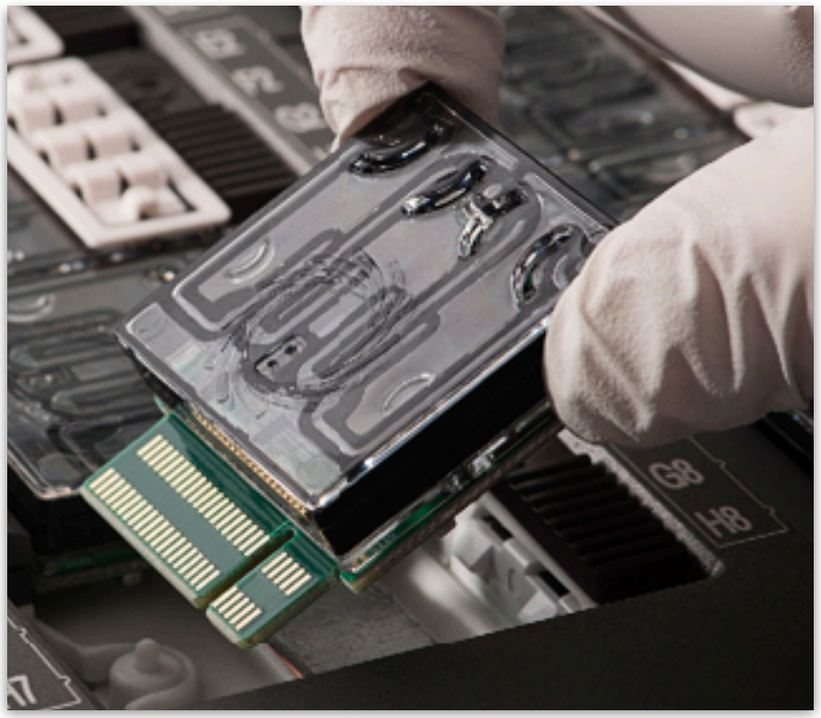
NGI stockholm

GridION

SciLifeLab

NGI stockholm

PromethION



SciLifeLab

NGI stockholm

SmidgION

(not yet released)

SciLifeLab

NGI stockholm

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
  - Need to balance yield with read length
- Price dropping rapidly, but still expensive compared to illumina
- NGI has 2x MinIONs and a PromethION

SciLifeLab

NGI stockholm

ion torrent

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

SciLifeLab

NGI stockholm

Ion Torrent PGM

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

SciLifeLab

NGI stockholm

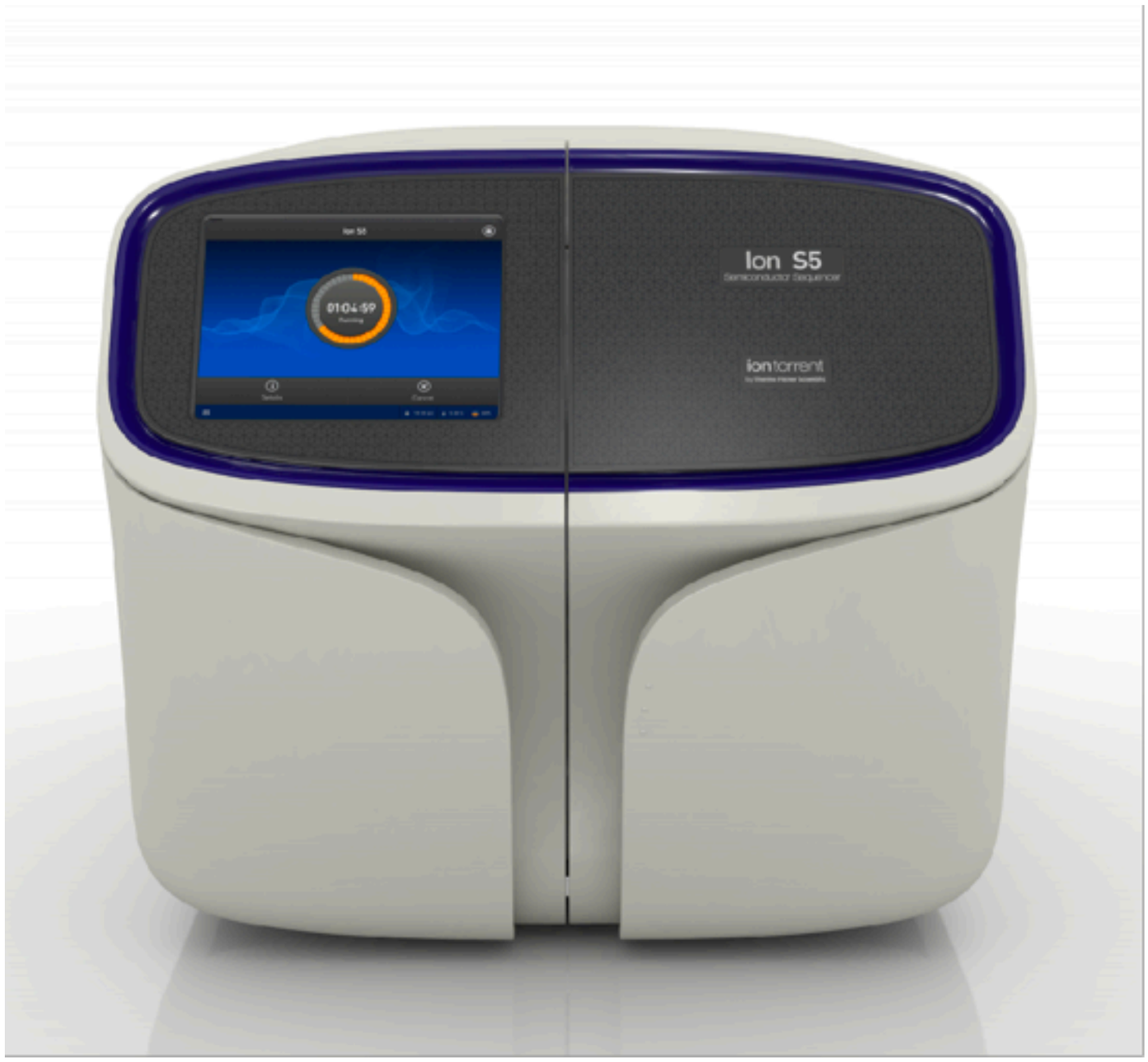
Ion Torrent Proton

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

SciLifeLab

NGI stockholm

Ion Torrent S5 XL



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

SciLifeLab

NGI stockholm

Sequencing Type

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

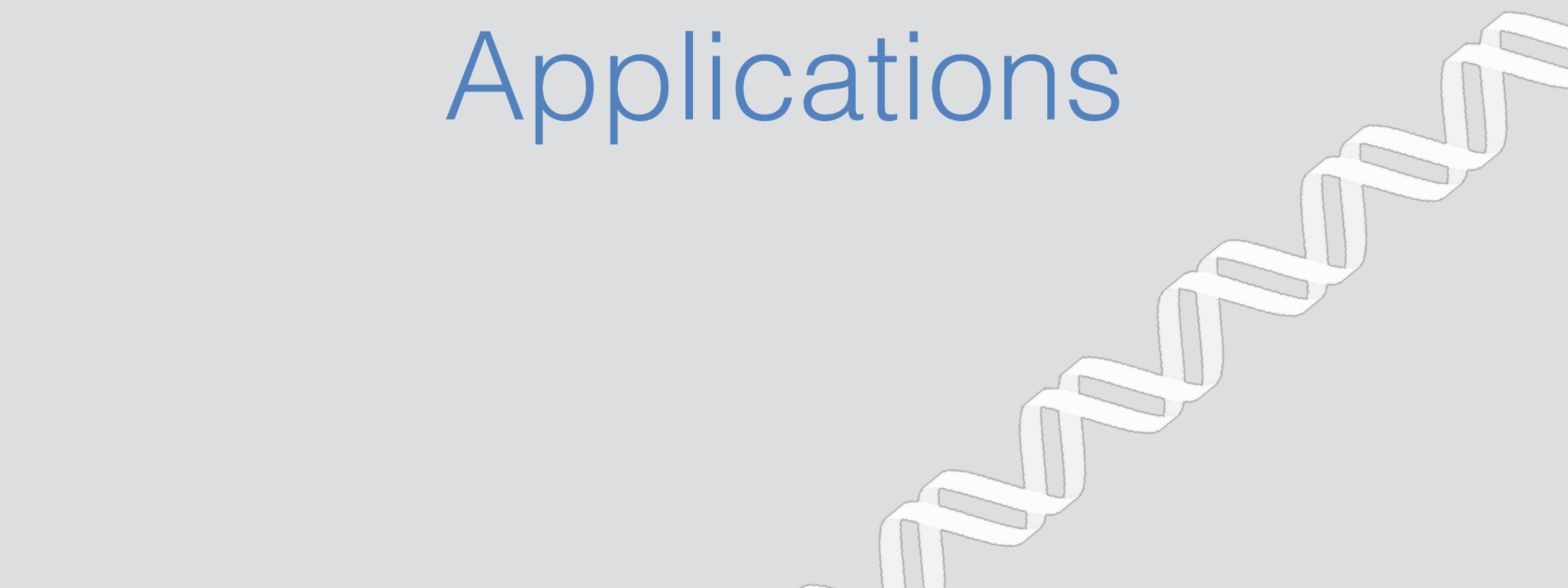
support@ngisweden.se

https://ngisweden.scilifelab.se/

SciLifeLab

NGI stockholm

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

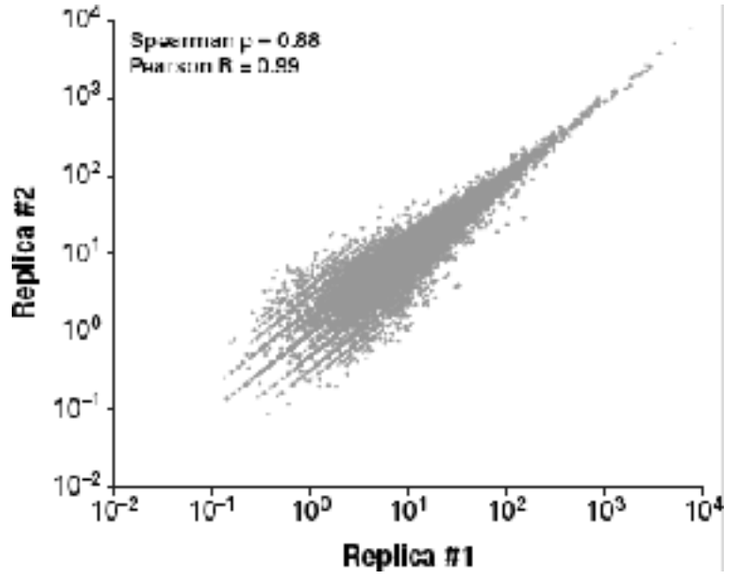
SciLifeLab

NGI stockholm

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



SciLifeLab

NGI stockholm

RNA Sequencing

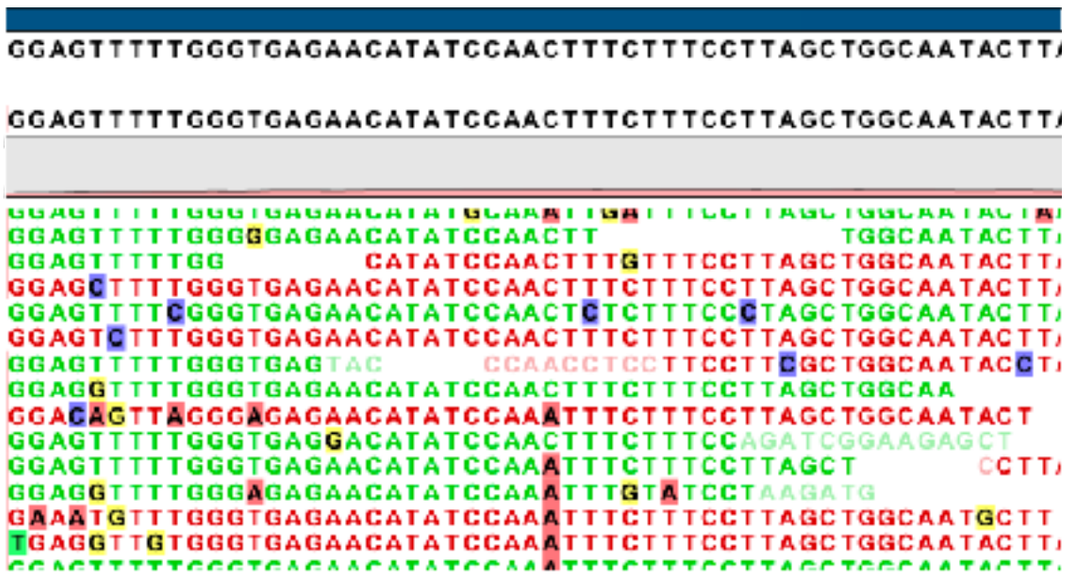
- Illumina sequencing RNA library prep kits
  - Illumina TruSeq RNA — *Protein-coding poly-A*
  - Illumina RiboZero — *rRNA depletion*
  - Illumina TruSeq RNA Exome — *FFPE / low quality*
  - Clontech SMARTER Pico — *low input*
  - Illumina TruSeq Small RNA — *small RNA*
- Oxford Nanopore, PacBio, IonTorrent

SciLifeLab

NGI stockholm

DNA Sequencing

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



SciLifeLab

NGI stockholm

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

SciLifeLab

NGI stockholm

10X Genomics

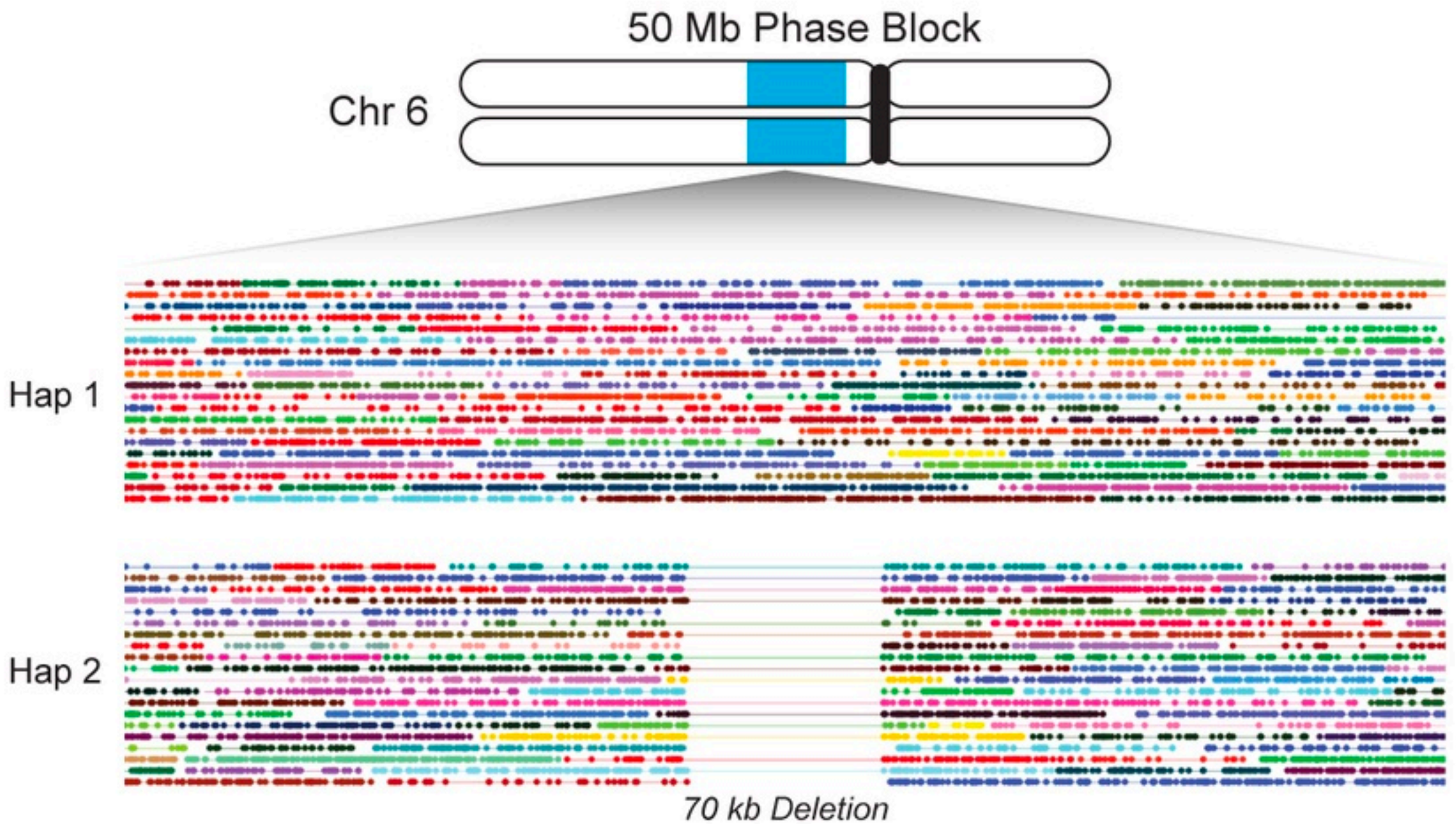
- Chromium instrument uses droplet emulsion technology for nanoliter reaction volumes

- 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

SciLifeLab

NGI stockholm

10X Genomics

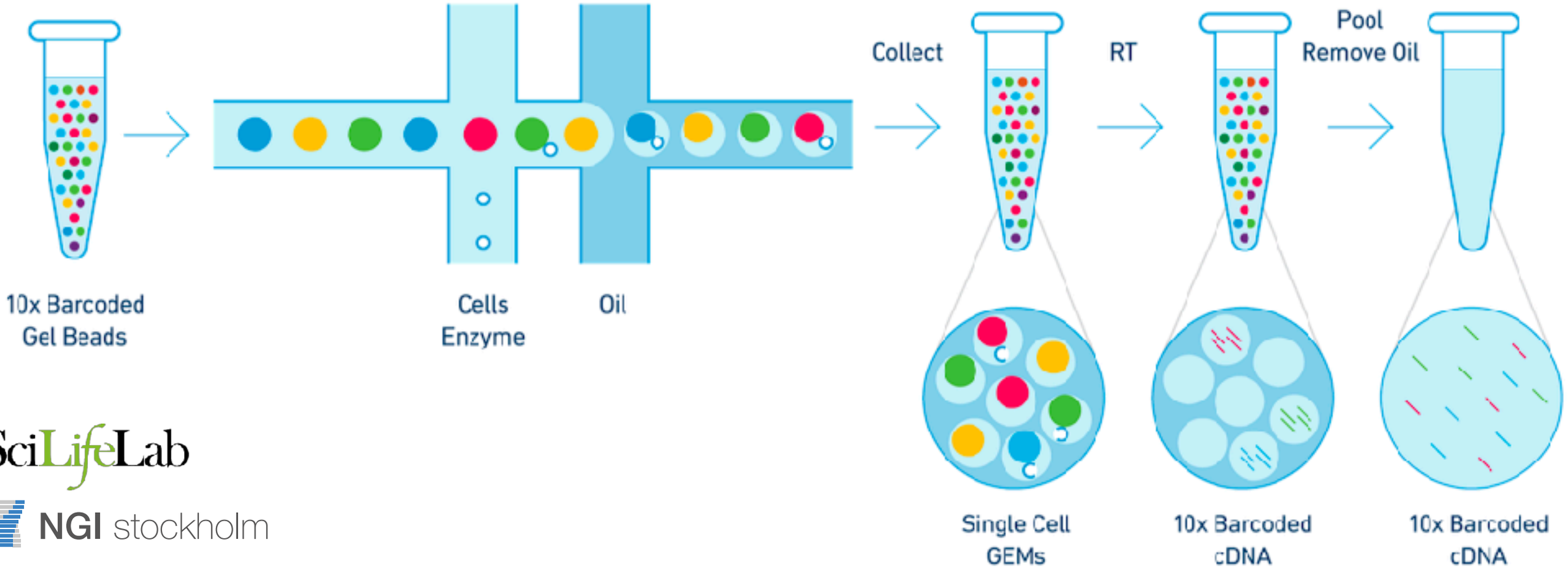


SciLifeLab

NGI stockholm

10X Genomics

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

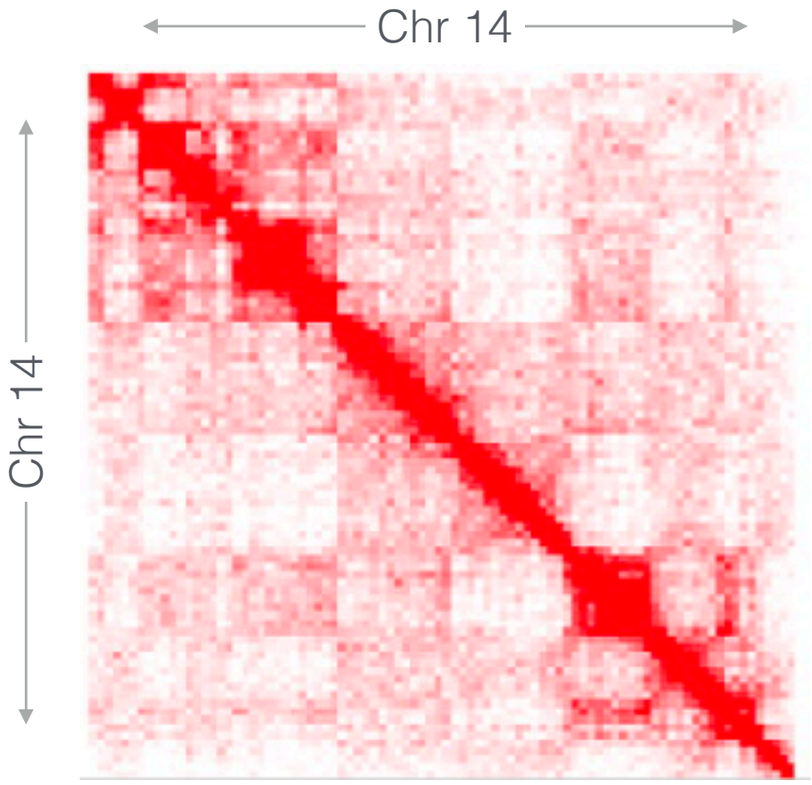


SciLifeLab

NGI stockholm

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



SciLifeLab
NGI stockholm
ARIMA
Dovetail GENOMICS

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

SciLifeLab

NGI stockholm

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

SciLifeLab

NGI stockholm

Amplicon Sequencing

- 16S / 18S / Custom amplicons
- High sample throughput
  - Plates of 96 samples processed using liquid handling automation
  - Large numbers of index combinations allow large pools
- Cheap and convenient for metagenomics and metabarcode sequencing projects
  - Contact us to talk about a pilot project

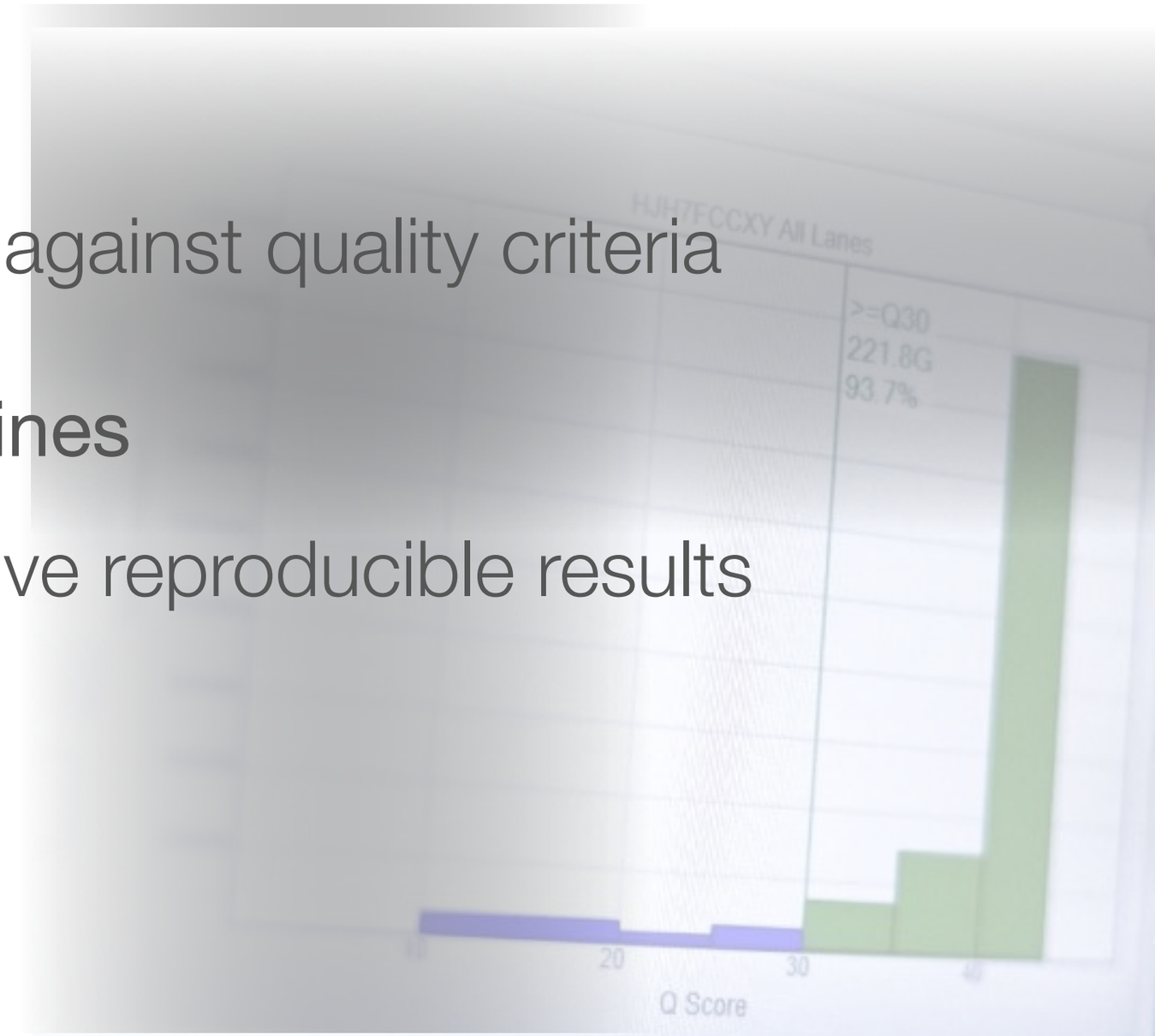
SciLifeLab

NGI stockholm

Bioinformatics at the NGI

Bioinformatics at NGI

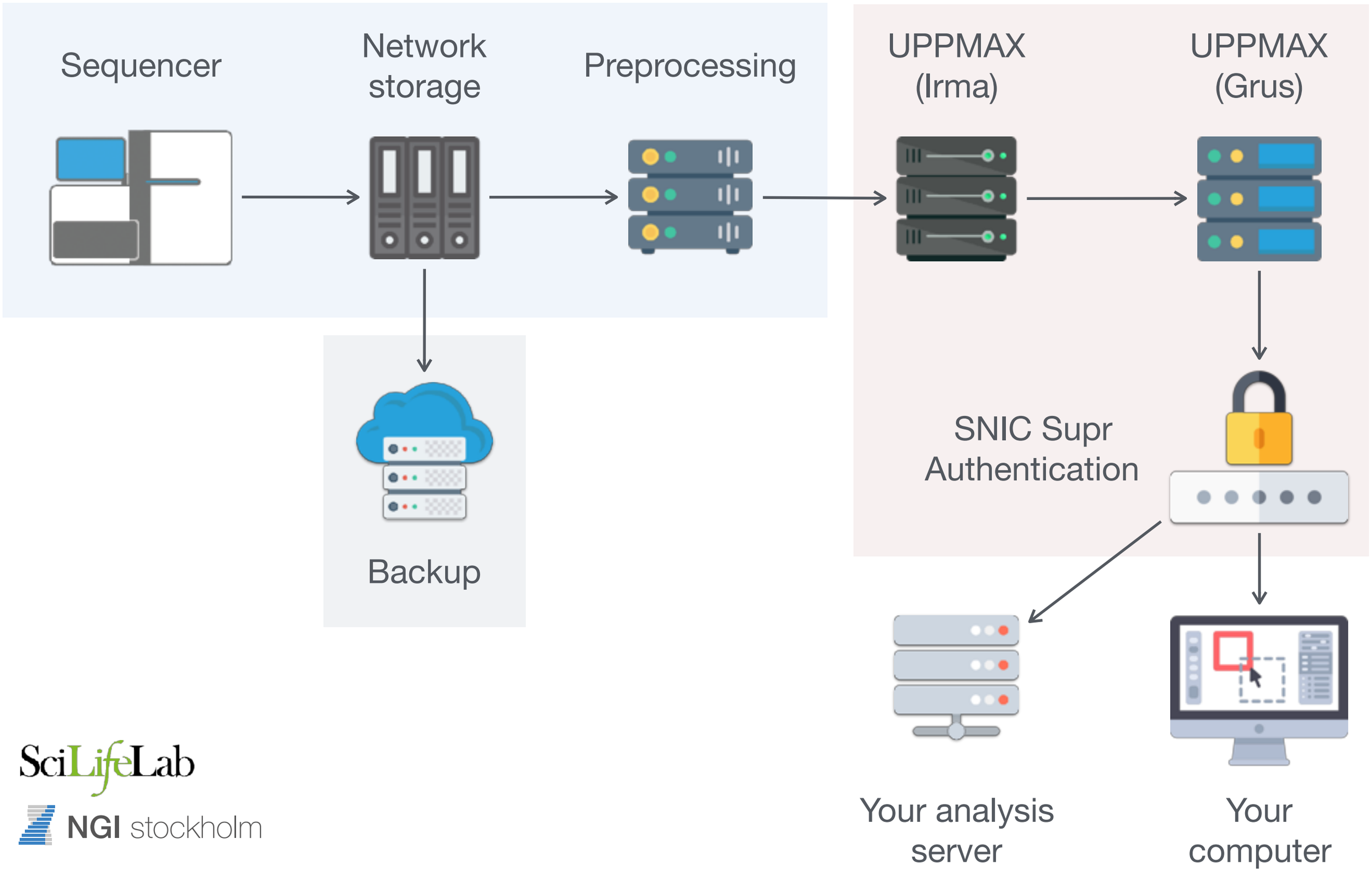
- Raw sequencing data management
  - Demultiplexing, data transfers, backups, delivery
- Quality control
  - Every project is checked against quality criteria
- Automated analysis pipelines
  - Standardised pipelines give reproducible results
- Software development



SciLifeLab

NGI stockholm

NGI Data Handling



SciLifeLab

NGI stockholm

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

SciLifeLab

NGI stockholm

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

SciLifeLab

NGI stockholm

SWEDAC. ACKREDITERING

Ackred. nr 1850
Provning
ISO/IEC 17025

Analysis Requirements

Automated

Reliable

nextflow

Easy for others to run

Reproducible results

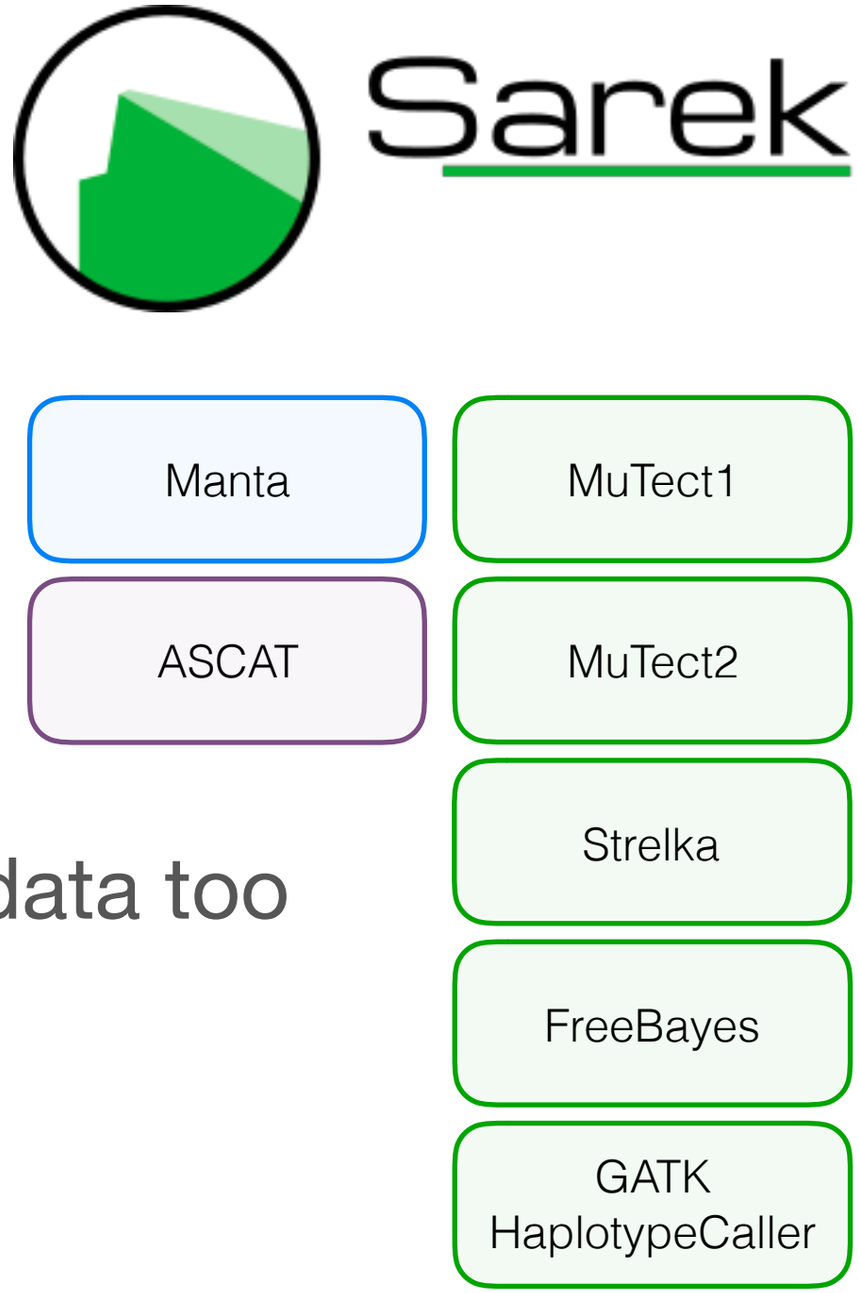
SciLifeLab

NGI stockholm

Sarek

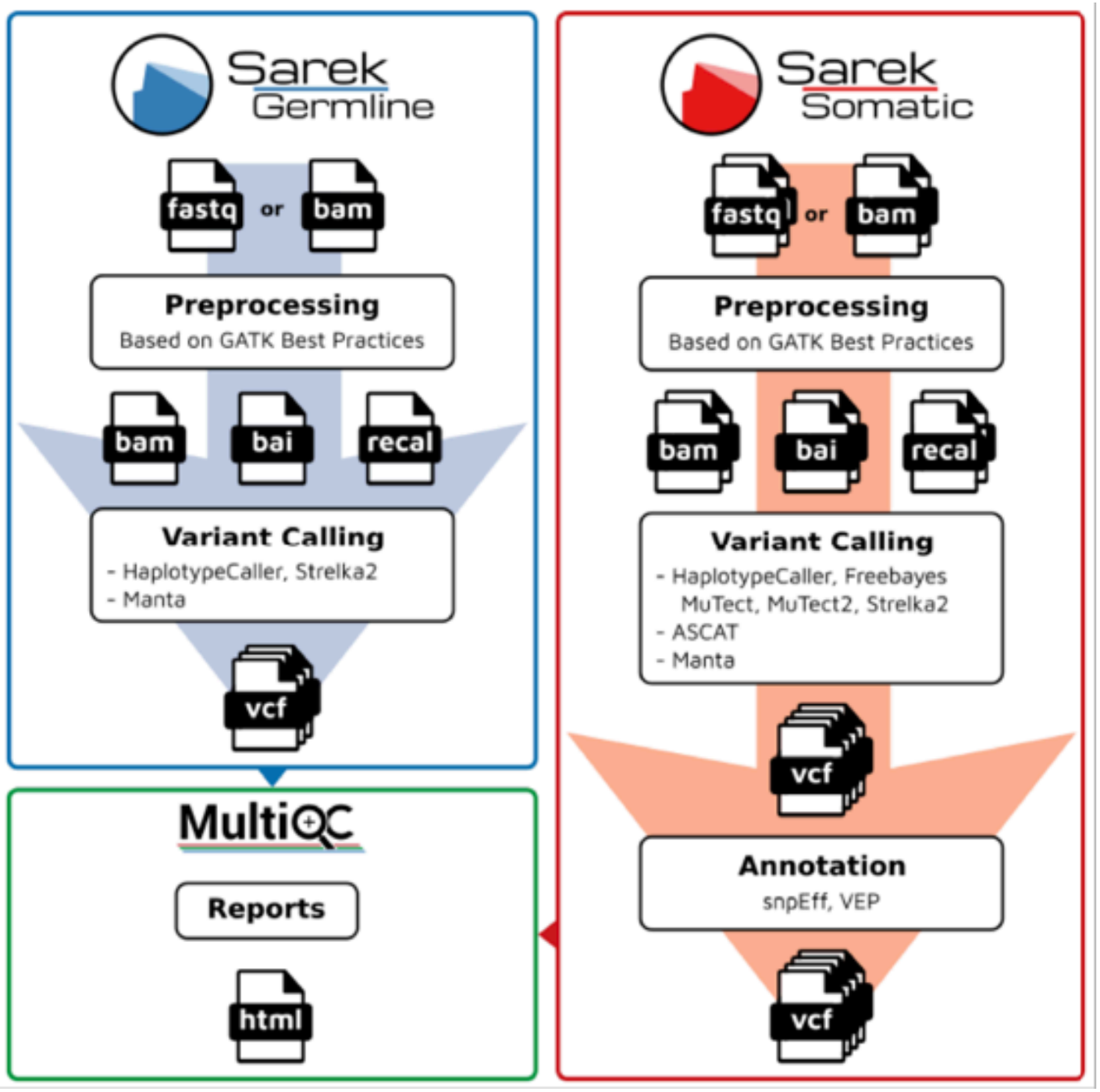
GitHub https://github.com/SciLifeLab/Sarek

- Tumour/Normal pair WGS analysis based on GATK best practices
  - SNPs, SNVs and indels
  - Structural variants
  - Heterogeneity, ploidy and CNVs
- Works with regular WGS and Exome data too



Sarek

- Tool split into sub-workflows
- Preprint available on bioRxiv
  - https://www.biorxiv.org/content/early/2018/05/09/316976
- Will soon be main DNA pipeline at NGI



NATIONAL GENOMICS INFRASTRUCTURE
NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN
Barntumörbanken

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

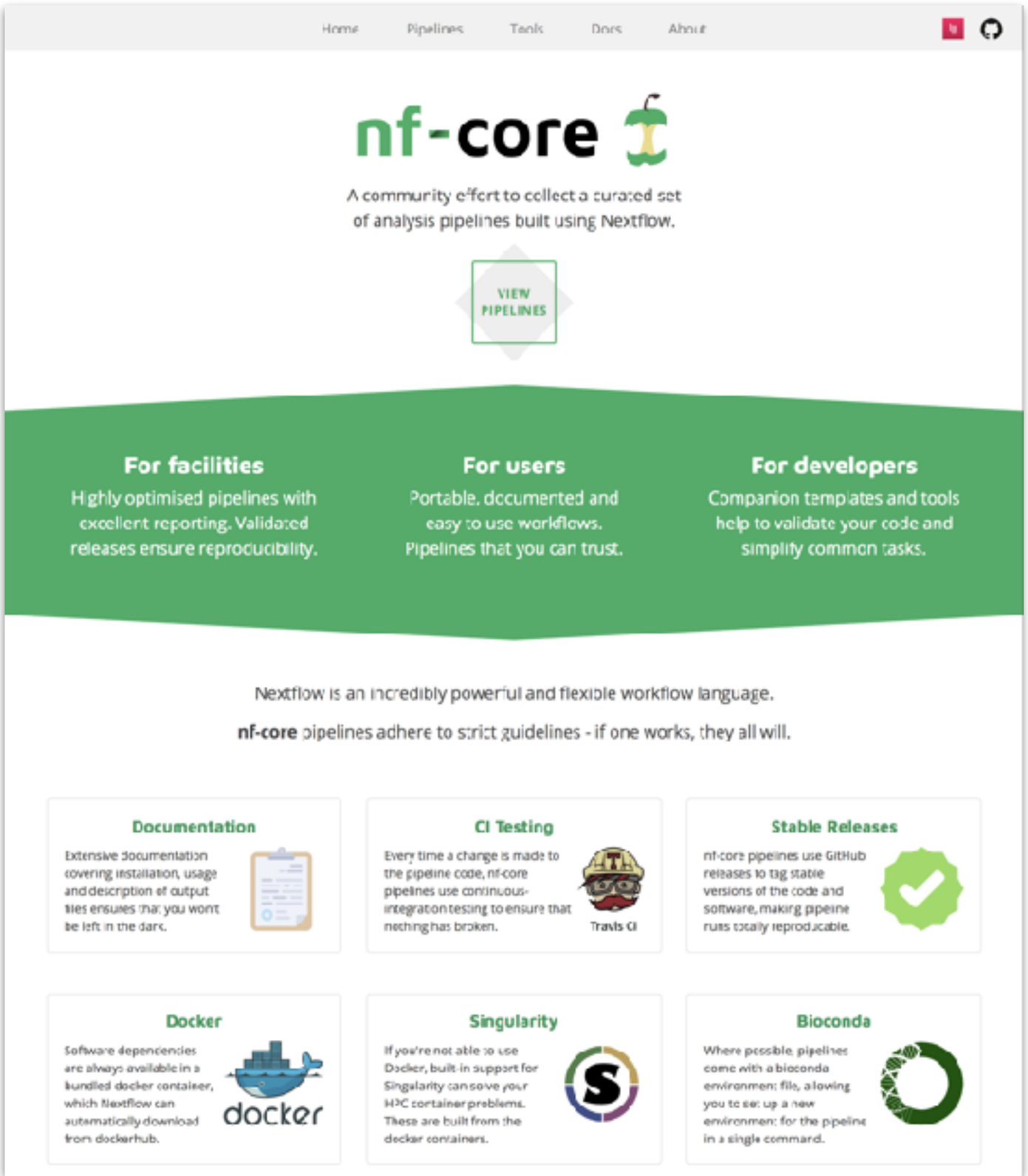
SciLifeLab

NGI stockholm

nf-core

https://nf-co.re

nf-core



nf-core

https://nf-co.re

- 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

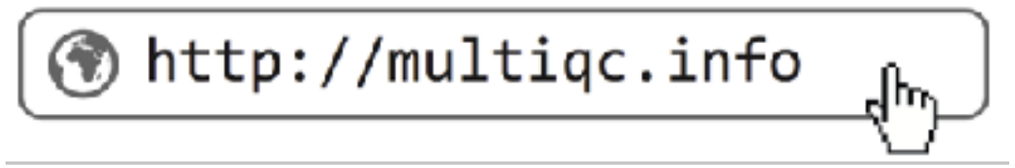
SciLifeLab

NGI stockholm

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 **67** supported tools
- Works with anything from tens → thousands of samples
- Highly customisable



MultiQC

MultiQC

NATIONAL GENOMICS INFRASTRUCTURE

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 6/9 columns.

Sample Name	% Aligned	M Aligned	% Trimmed	% Dups	% GC	M Seqs
P1234_1001	68.2%	22.8	10.3%	71.3%	49%	33.7
P1234_1002	67.9%	20.9	10.7%	70.1%	50%	31.1
P1234_1003	64.7%	21.7	11.0%	72.3%	50%	33.7
P1234_1004	55.2%	17.0	13.2%	73.4%	51%	31.2
P1234_1005	53.0%	17.7	15.9%	75.8%	52%	33.8
P1234_1006	52.7%	16.1	14.1%	73.8%	52%	30.8
P1234_1007	33.0%	7.0	32.0%	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%	46%	20.5

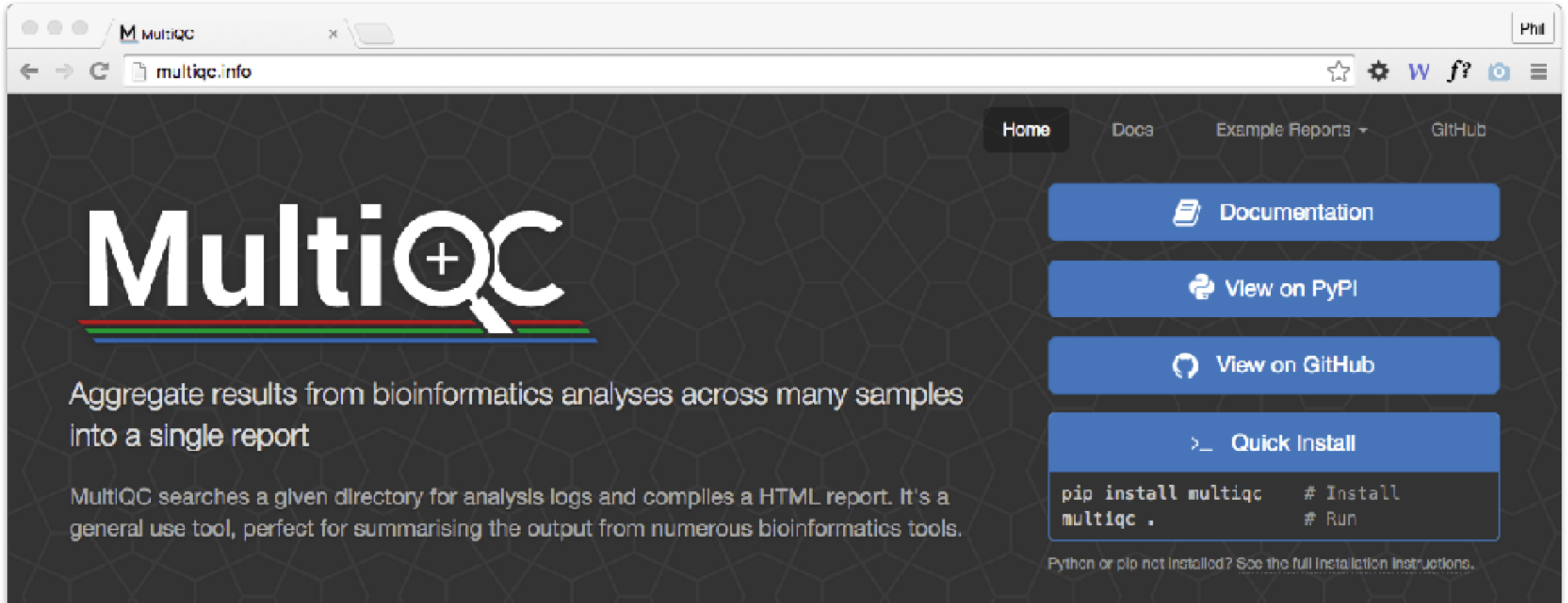
Getting MultiQC

ewels / MultiQC

BIOCONDA

http://multiqc.info

PyPI



Conclusions

If you have a project

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

https://ngisweden.scilifelab.se

support@ngisweden.se

SciLifeLab

NGI stockholm

Find our tools

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

http://opensource.scilifelab.se

SciLifeLab

NGI stockholm

Acknowledgements

Phil Ewels

phil.ewels@scilifelab.se

ewels

tallphil

Thanks to:

Max Käller

Olga Vinnere Pettersson

NGI Sweden

ngisweden

support@ngisweden.se

http://ngisweden.scilifelab.se

http://opensource.scilifelab.se

SciLifeLab

NGI stockholm