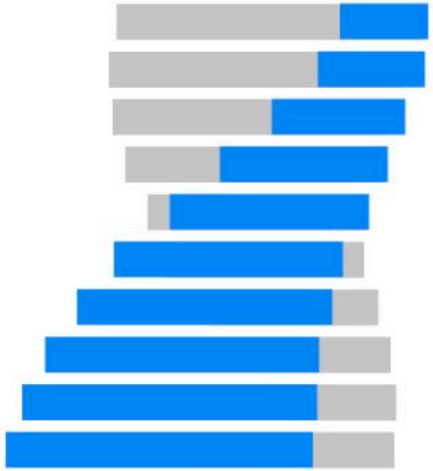
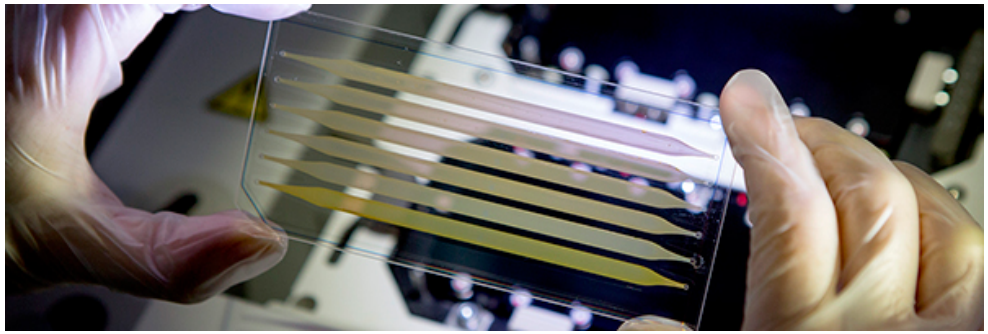
Data Generation
–where, what and how much



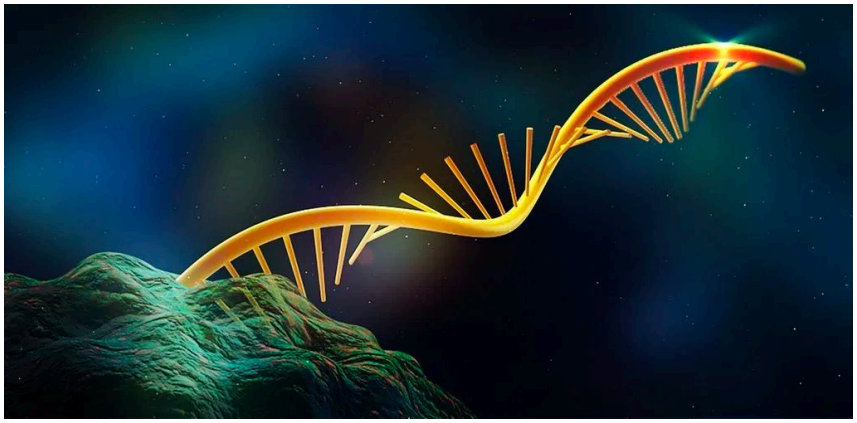


Carl-Johan Rubin
Head of Applications Development,
National Genmics Infrastructure (NGI)
SciLifeLab, Stockholm

Talk outline

- **Where can I get sequencing data**
  - NGI
    - Organization
    - Technologies by node
    - User projects flow
  - Sequencing service providing companies
  - Data repositories

- **Data formats**
    - Fastq: compression
    - Fastq/sam/bam/cram
      - Typical space requirements
    - Typical examples per unit of WGS/RNA
    - Best practise analyses



Genomics data
– *data types and repositories*

NATIONAL GENOMICS INFRASTRUCTURE

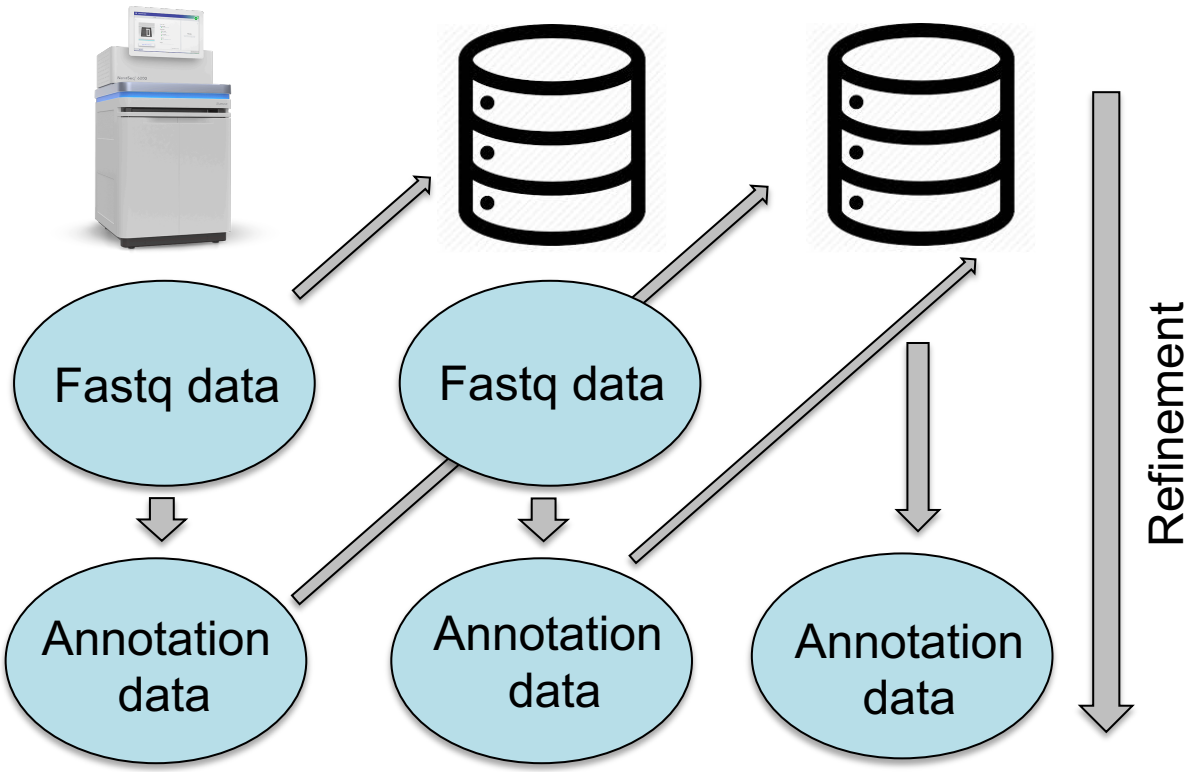
SciLifeLab

Sequencing data

- Raw data (images, bcl, signal data)
- Basecalled data
  - Typically .fastq format
  - Sam/bam

Annotation data

- Standardized formats
  - Bed, Wig, bedgraph, BigWig, BigBed, gff, gtf3, fasta, gfa, vcf, gvcf etc.



ENA
European Nucleotide Archive

SRA (NCBI)

https://www.ensembl.org/biomart
Encode
NCBI/EMBL-EBI
db-SNP
Etc.

Data repositories

- Sequence Read Archive (SRA)
  - https://www.ncbi.nlm.nih.gov/sra
  - SRA-toolkit
  - Link download + API

- European Nucleotide Archive (ENA)
  - https://www.ebi.ac.uk/ena
  - raw sequencing data, sequence assembly info functional annotation
  - Link download + API

Sequence data from repositories

European nucleotide archive (ENA)

EMBL-EBI

Services | Research | Training | About u

ENA
European Nucleotide Archive

Search

Examples: BN000065, histone

Advanced
Sequence

Home | Search & Browse | Submit & Update | Software | About ENA | Support

ENA > Search & Browse > Downloading ENA data

Downloading ENA data

The main tool for downloading ENA data is the ENA Browser. The ENA Browser can be used both interactively and programmatically through REST URLs. All ENA data including assembled and annotated sequences is available for download through the ENA Browser.

Data in ENA can be searched via the search box in the header of all our pages. The search results are presented through the ENA Browser.

Please refer to the following sections for information on how to bulk download ENA data.

Sequences

Assembled and annotated sequences are available for bulk download. Information on how to do this can be found here.

Read data

Read data is available for bulk download. Information on how to do this can be found here

Taxonomy data

Taxonomy data is available for bulk download. Information on how to do this can be found here.

Search & Browse

- Data formats
  - Genome assemblies
- Marker portal
- Taxon portal
- Programmatic access
  - Data retrieval
  - Taxon portal
  - Marker portal
  - Search
  - File reports
  - XREF service
- Genome assembly database
- Taxonomy Service
  - Translation tables

Sequence data from repositories

NCBI Sequence Read Archive (SRA)

SciLifeLab

NCBI Resources How To Sign in to NCBI

SRA SRA Search

Advanced

Documentation | Downloading | Submitting | Browsing | SRA Toolkit | SRA-BLAST Use Cases | Archives

Download SRA sequences from Entrez search results

- Obtain search results
- Obtain run accessions
- Download sequence data files using SRA Toolkit
- Download metadata associated with SRA data
- Download sequence data from the Run Browser
- Download SRA sequence data using Amazon Web Services (AWS)
- Contact SRA

Obtain search results

Go to:

Task: find RNA-Seq records for lymph node tissue in BALB/c mice in SRA Entrez

To learn how to use Advanced Search Builder please refer to Search in SRA

- In the Entrez search bar enter the query: ((("mus musculus"[Organism]) AND BALB/c*) AND "lymph*") AND "rna seq"[Strategy].
- To limit your search to only aligned data add to the above query **AND aligned data"[Properties]**.
- Click the checkboxes next to records (experiments) to select data of interest. Leave all checkboxes unchecked to select all records (experiments) from your search.

Obtain run accessions

Go to:

Run accessions are used to download SRA data. To download a list of Run accessions selected from your Entrez search:

- Click **Send to** on the top of the page, check the radiobutton **File** , select **Accession List**.
- Save this file in the location from which you are running the SRA Toolkit.

Download sequence data files using SRA Toolkit

Go to:

! Please make sure you are running the most recent release of the toolkit as older versions may not be compatible with more recently loaded data or the most current network protocols.

Documentation

- SRA Overview
- SRA Fact Sheet (.pdf)
- SRA database growth
- File Format Guide
- Search in SRA

Downloading SRA data

- Download Guide
- dbGaP Download Guide
- dbGAP Cloud Access

Submitting Data to SRA

General

- Quick Start
- BioProject & BioSample
- SRA Metadata Overview
- SRA File Upload
- Frequently Asked Questions

SRA Submission Portal

- Submitting to SRA
- Troubleshooting submission

Submitting for dbGaP & GEO

- Submitting for dbGaP
- Submitting for GEO

Updating SRA Data

Sequence data from repositories

SRA Explorer

https://ewels.github.io/sra-explorer/

SRA-Explorer

24 saved datasets

SRA Explorer

This tool aims to make datasets within the Sequence Read Archive more accessible.

Search for: SRP043510[All Fields]

Max Results 100 **Start At Record** 0

Need inspiration? Try GSE30567, SRP043510, PRJEB8073, ERP009109 or human liver miRNA.

24 Saved Datasets

Remove all from collection and send to search results

FastQ Downloads | SRA Downloads | Full Metadata

To download FastQ files directly, sra-explorer queries the ENA for each SRA run accession number.

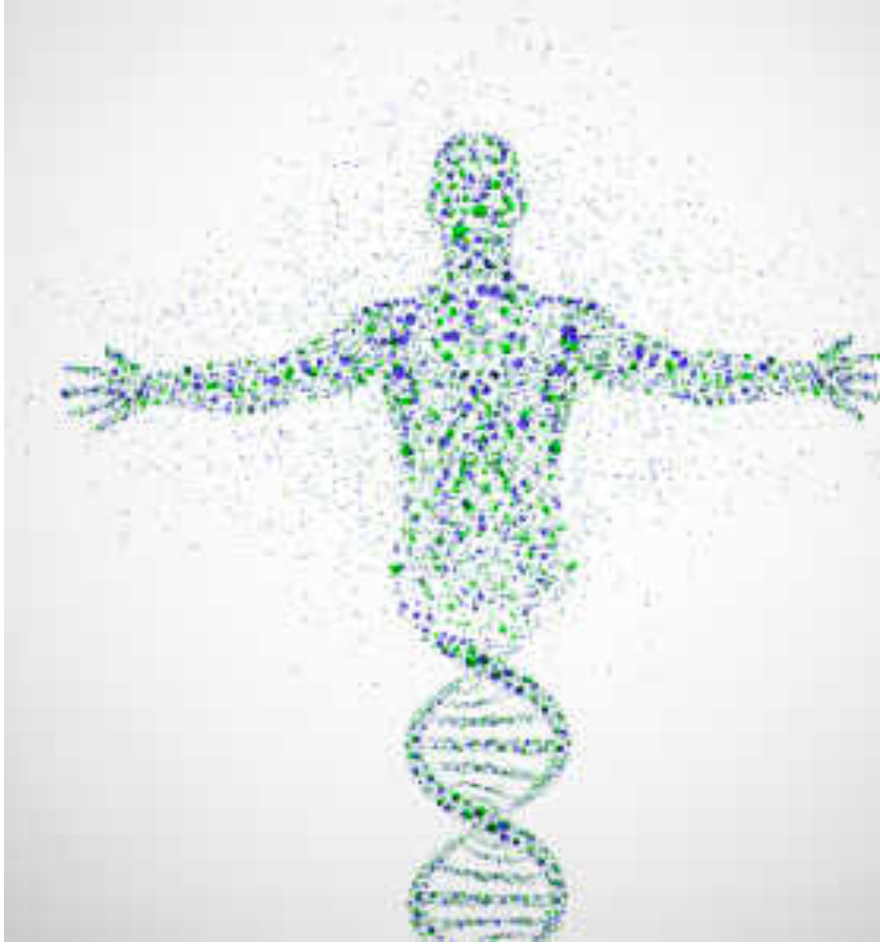
- Raw FastQ Download URLs
- Bash script for downloading FastQ files
- Aspera commands for downloading FastQ files
- Cluster Flow FastQ download file (nice filenames)
- bcbio project file for FastQ downloads (nice filenames)

Where can you get seq. data?

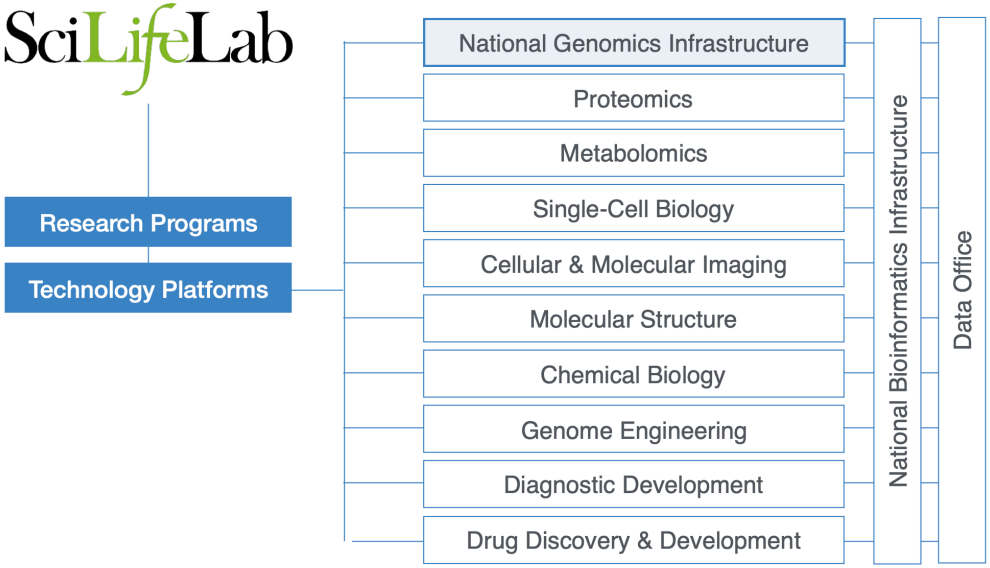
- **SciLifeLab platforms**
  - **National Genomics Infrastructure (NGI)**
  - Eukaryotic Single Cell Genomics (ESCG)
  - Ancient DNA
  - Microbial Single Cell Genomics
  - Diagnostics delevopment (Clinical Genomics

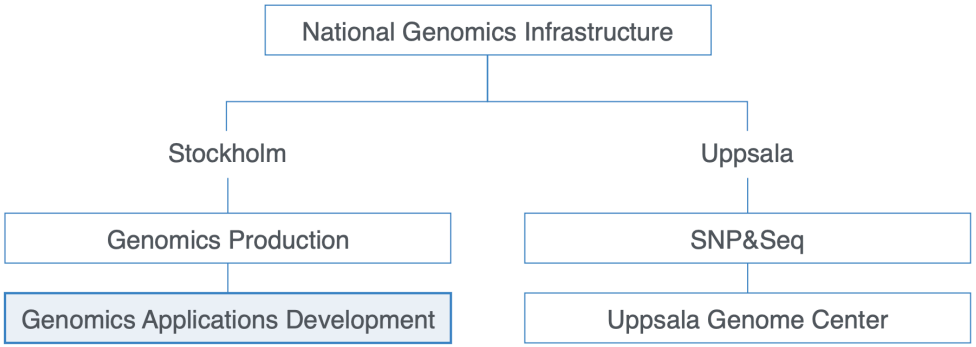
- **Companies**
  - Eurofins
  - TATAA
  - Etc.

- **Data repositories**
  - European Nucleotide Archive (ENA)
  - Sequence Read Archive (SRA)
  - NCBI / EMBL-EBI



NGI organization





SciLifeLab NGI mission

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

NGI methods/tech. by node

NATIONAL GENOMICS INFRASTRUCTURE

SciLifeLab

- **Stockholm**
- Bulk DNA-seq
- Bulk RNA-seq
- HiC + Omni-C
- 10X-chromium
- Nanopore
- ATAC-seq
- Low input RNA/DNA
- etc.

- **Uppsala (SNP&Seq)**
  - Bulk DNA-seq
  - Bulk RNA-seq
  - 10x single cell
  - Genotyping
  - ChIP-seq
  - WGBS + RRBS

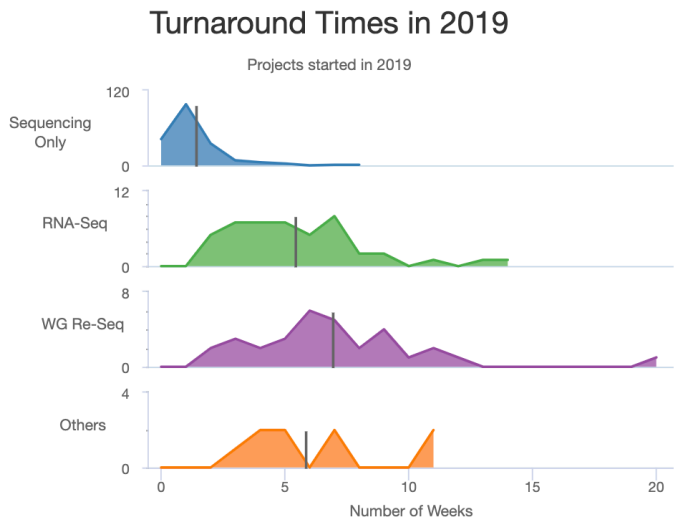
- **Uppsala (UGC)**
  - PacBio
  - Oxford nanopore
  - Ion Torrent
  - Assembly

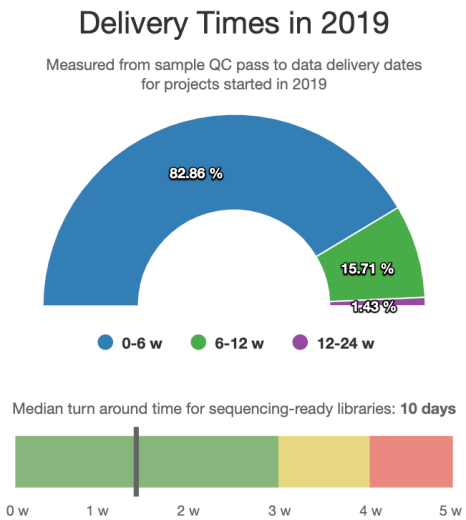
HiSeq X decommissioned

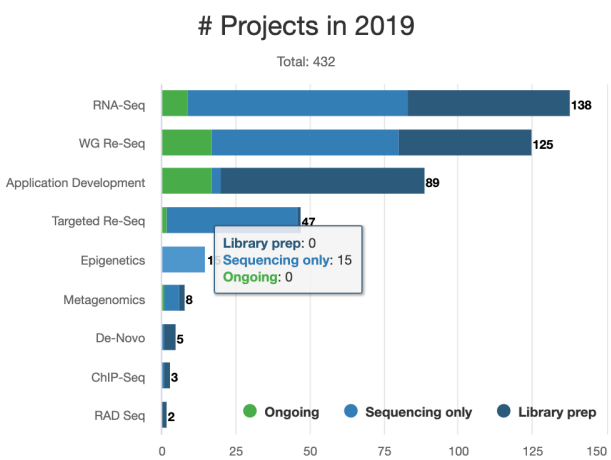
NGI technologies by node

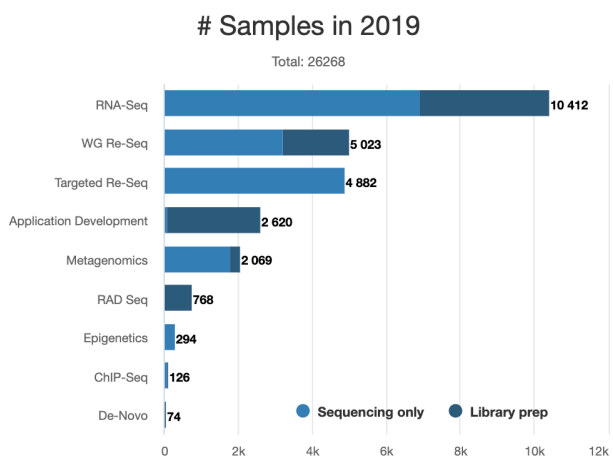
SciLifeLab NGI – throughput

NGI Stockholm 2019 stats

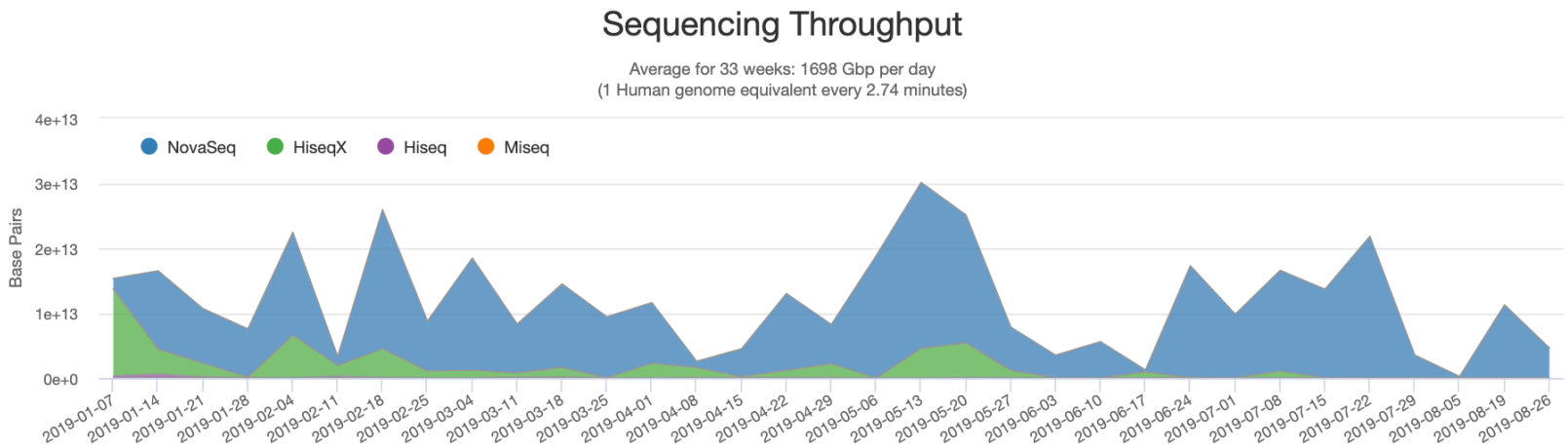


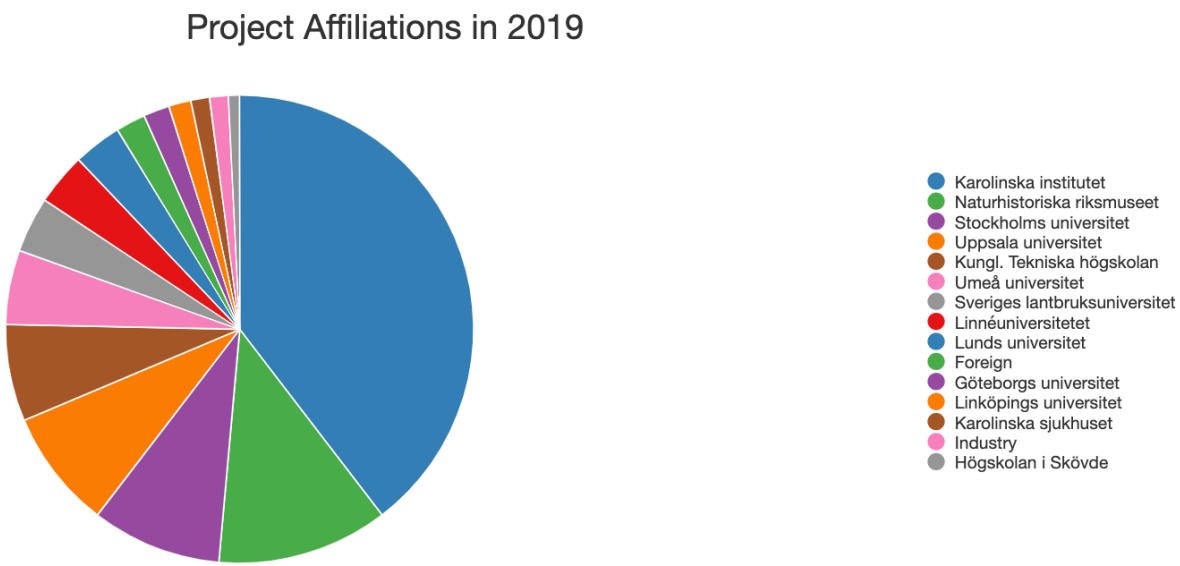




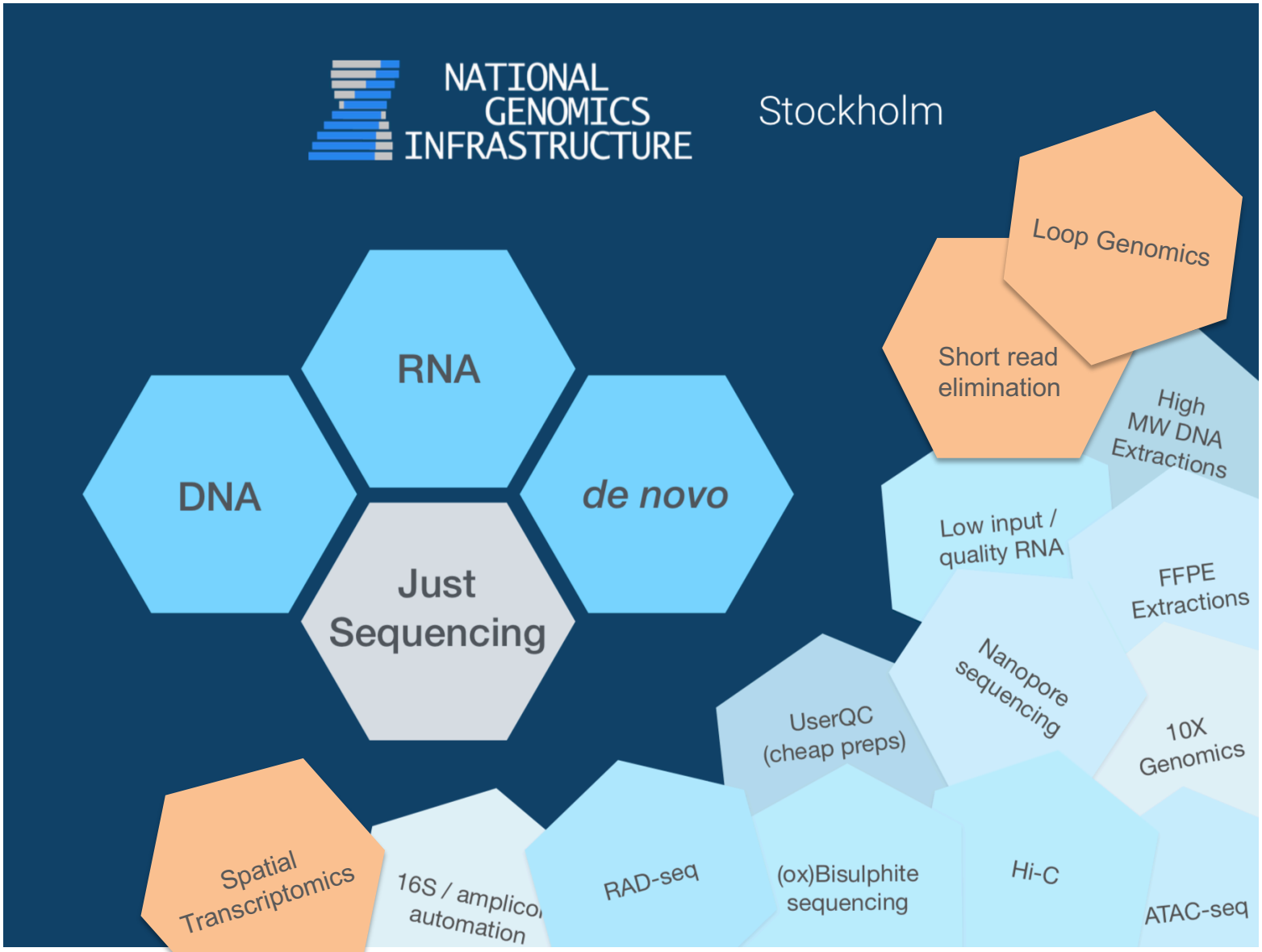


NGI Stockholm 2019 stats





NGI Stockholm methods by facility



NovaSeq Flowcell throughput and costs

					Output per flowcell (M clusters)			Cost per flowcell	Output per lane (based on Illumina spec max)		Cost per lane (including XP kit for NovaSeq)		
					Spec		Typical						
Instrument	Flowcell	Read setup (base pairs)	Total cycles	Lanes / flowcell	Min	Max	Approx.	SEK	M clusters / lane	Gbp / lane	SEK	SEK / M bp	SEK / M clusters
NovaSeq	S Prime	2x50	100	2	650	800	?	24,330	400	40	16,435	0.41	41.09
NovaSeq	S Prime	2x150	300	2	650	800	?	41,631	400	120	25,085	0.21	62.71
NovaSeq	S Prime	2x250	500	2	650	800	?	59,473	400	200	34,006	0.17	85.02
NovaSeq	S1	2x50	100	2	1300	1600	1800	44,335	800	80	26,437	0.33	33.05
NovaSeq	S1	2x100	200	2	1300	1600	1800	59,473	800	160	34,006	0.21	42.51
NovaSeq	S1	2x150	300	2	1300	1600	1800	71,368	800	240	39,954	0.17	49.94
NovaSeq	S2	2x50	100	2	3300	4100	4000	102,726	2050	205	55,633	0.27	27.14
NovaSeq	S2	2x100	200	2	3300	4100	4000	140,573	2050	410	74,556	0.18	36.37
NovaSeq	S2	2x150	300	2	3300	4100	4000	164,903	2050	615	86,721	0.14	42.30
NovaSeq	S4	2x100	200	4	8000	10000	10000	194,715	2500	500	51,325	0.10	20.53
NovaSeq	S4	2x150	300	4	8000	10000	10000	224,475	2500	750	58,765	0.08	23.51
MiSeq	v2	1x50	50	1	10	10	10	8,982	10	0.5	8,982	17.96	898.17
MiSeq	v2	2x150	300	1	10	10	10	11,551	10	3	11,551	3.85	1,155.08
MiSeq	v2	2x250	500	1	10	10	10	12,938	10	5	12,938	2.59	1,293.81
MiSeq	v3	2x75	150	1	18	18	18	9,948	18	2.7	9,948	3.68	552.65
MiSeq	v3	2x300	600	1	18	18	18	17,367	18	10.8	17,367	1.61	964.85
MiSeq	Nano v2	2x150	300	1	1	1	1	3,309.03	1	0.3	3,309.03	11.03	3309.03

Prices last confirmed: 2019-02-15

Most cost efficient NovaSeq flow cell
7-8 mammalian genomes at 30X / lane
4 lanes → 28-32 genomes per flow cell

More flexible sequencing

- NovaSeq throughput is high
  - Users want more flexibility
  - Not just full S4 lanes
- Now also ¼ lanes.
- --> Lower sequencing costs
- Development efforts at NGI
  - Pooling balance
  - New indexes

			Cost per lane (including XP kit for NovaSeq)		
Instrument	Flowcell	Read setup (base pairs)	SEK	SEK / M bp	SEK / M clusters
NovaSeq	S Prime	2x50	16,435	0.41	41.09
NovaSeq	S Prime	2x150	25,085	0.21	62.71
NovaSeq	S Prime	2x250	34,006	0.17	85.02
NovaSeq	S1	2x50	26,437	0.33	33.05
NovaSeq	S1	2x100	34,006	0.21	42.51
NovaSeq	S1	2x150	39,954	0.17	49.94
NovaSeq	S2	2x50	55,633	0.27	27.14
NovaSeq	S2	2x100	74,556	0.18	36.37
NovaSeq	S2	2x150	86,721	0.14	42.30
NovaSeq	S4	2x100	51,325	0.10	20.53
NovaSeq	S4	2x150	58,765	0.08	23.51
MiSeq	v2	1x50	8,982	17.96	898.17
MiSeq	v2	2x150	11,551	3.85	1,155.08
MiSeq	v2	2x250	12,938	2.59	1,293.81
MiSeq	v3	2x75	9,948	3.68	552.65
MiSeq	v3	2x300	17,367	1.61	964.85
MiSeq	Nano v2	2x150	3,309.03	11.03	3309.03

Long read sequencing

- ONT

- PacBio

Long read sequencing

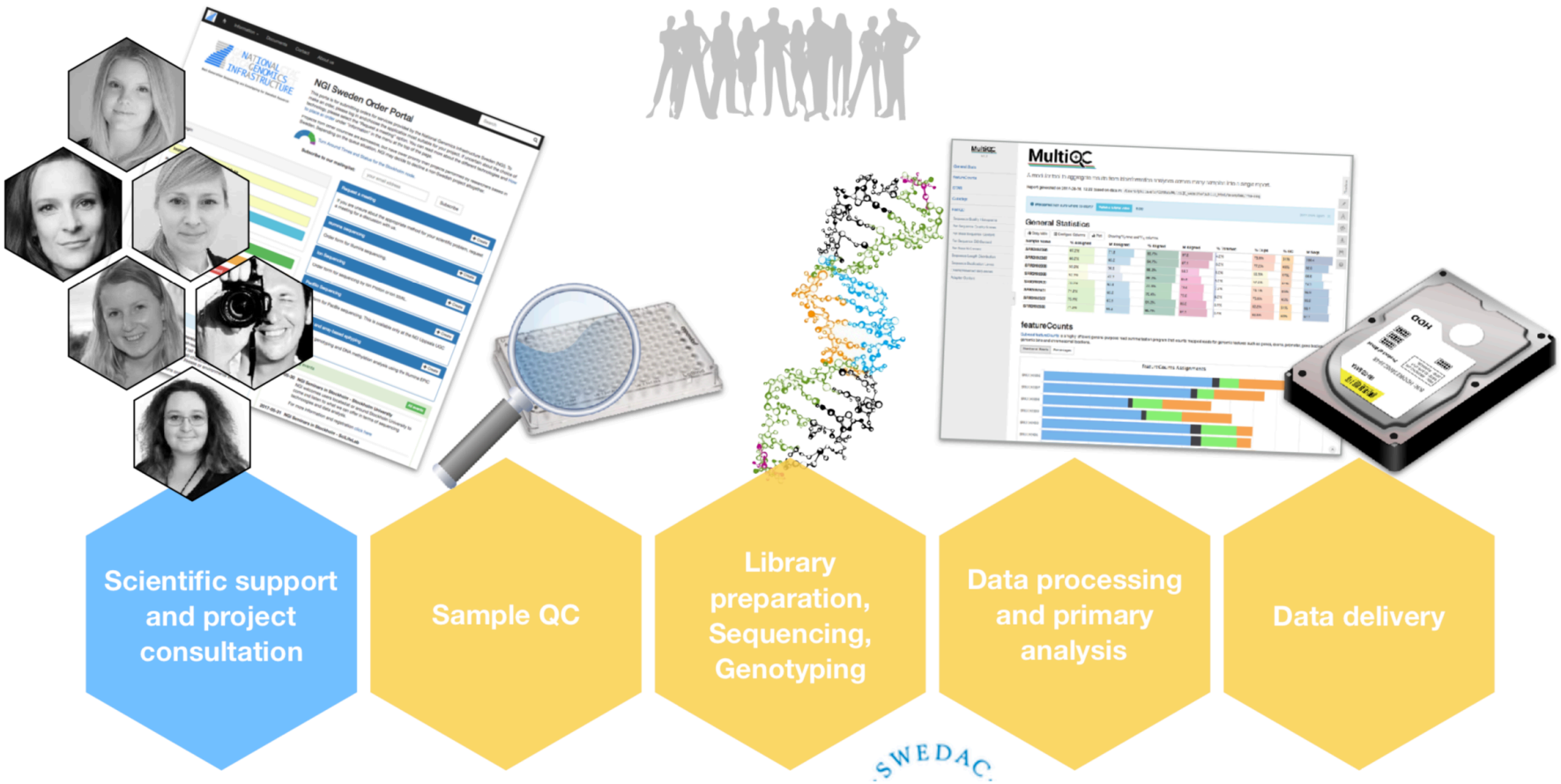
– *current stats*

Platform	Throughput (flowcell	SEK/ Gb	Read lengths	Quality (Phred)
ONT (P)	80-150Gb	200-400	Max 2 Mb	R9 read: 12 R9 consensus: 30 R10 consensus: 40
PacBio Sequel	10 Gb	1000-1500	Max 170 kb	Read: 8-9 HiFi: 20-50
PacBio Sequel II	100-150Gb	200-300	Max 170 kb	Read: 8-9 HiFi: 20-50

ONT signal level data 5-10x size of basecalled data

NGI projects pipeline

from consultation to data



NGI – orders and information

https://ngisweden.scilifelab.se/

Information Documents Contact About us

NATIONAL GENOMICS INFRASTRUCTURE

Next-Generation Sequencing and Genotyping for Swedish Research

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

email address Subscribe

Login

Email

Email address of account

Password

Login

Register account

Reset password

Illumina Sequencing + Create order

Order form for Illumina sequencing.

Request a meeting + Create order

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

Ion Sequencing + Create order

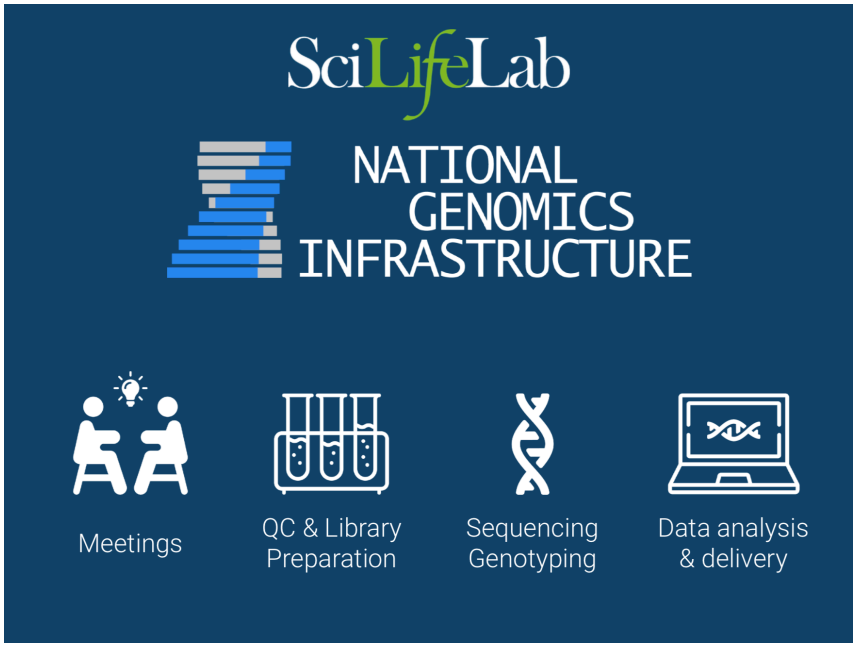
Order form for sequencing by Ion Proton or Ion S5XL.

PacBio Sequencing + Create order

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

NGI – orders and information
https://ngisweden.scilifelab.se/

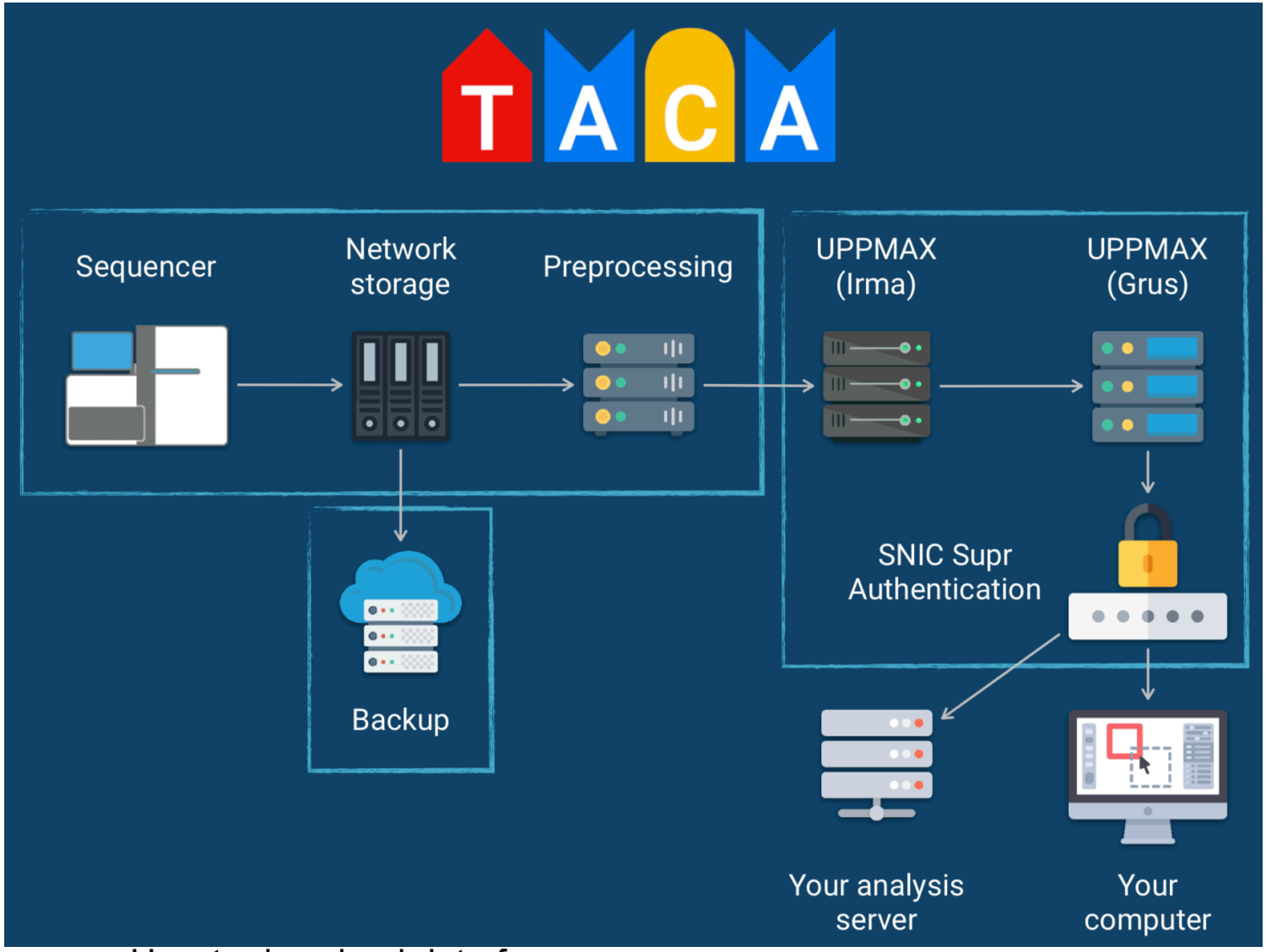
- Meeting with NGI project coordinators
  - *Feasibility discussion*
  - *Limitations (samples/amounts/etc.)*
  - *Capabilities*
  - *Pilot projects*
- Submit order
  - *Project information*
  - *Sample sheets*
  - *Plates sent out*
- Lab + data management
- Deliveries
  - *Usually 6-12 weeks **
  - *Secure server deliveries (2-factor authentication) (SNIC/SUPR)*



NGI laboratory and data flow

Lab	Clarity LIMS™ Intuitive Lab Management™
Data Management	TACA
Analysis	nextflow
Quality Control	MultiQC

NGI – data processing



How to download data from grus:
https://www.youtube.com/watch?v=-6ZufZRuwyU

NGI – data deliveries

MultiQC

Multiple tools

Multiple samples

One report

Standardised output

MultiQC

1. Install MultiQC — `pip install multiqc`
2. Run MultiQC — `multiqc .`
3. Read the report

MultiQC v1.0

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

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 6/8 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%	80.5%	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

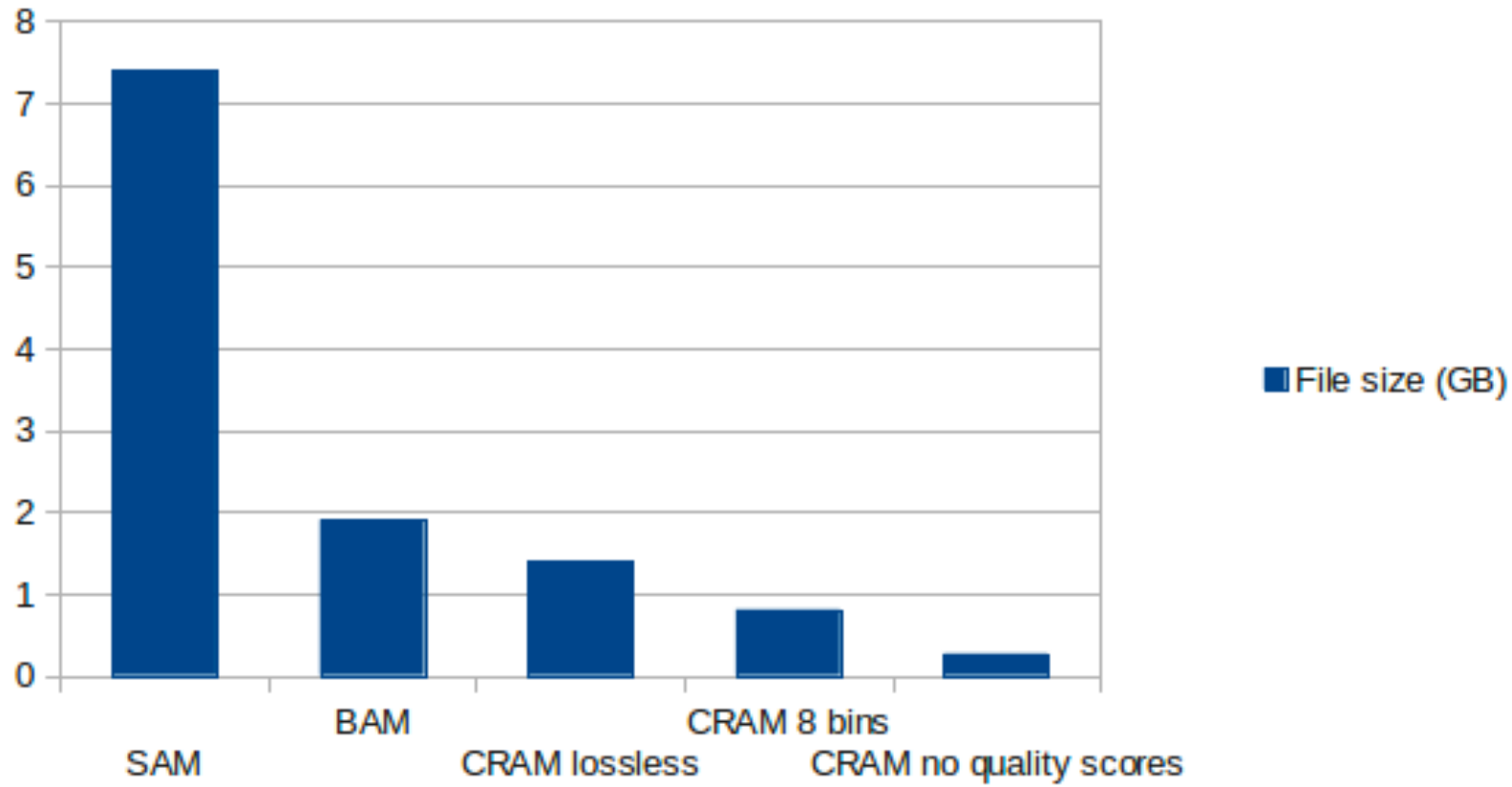
Best practise pipelines

- **WGS human:**
  - Delivers: fastq, bams, vcf-files from different tools
  - Preprocessing: *bwa, GATK*
  - Germline/somatic variant calling
  - Annotation
  - Reporting

- RNA-seq
  - The workflow processes raw data from FastQ
  - FastQC,
  - Trim Galore
  - STAR/HiSAT2),
  - generates gene counts (featureCounts, StringTie)
  - quality-control RSeQC, dupRadar, Preseq, edgeR, Multi QC).

SAM – BAM - CRAM

SAM: Sequence Alignment Map format (raw tex)
BAM: binary SAM (factor 3-4 compression)
CRAM: more efficiently compressed bam (lossless to lossy)



Further compression possible

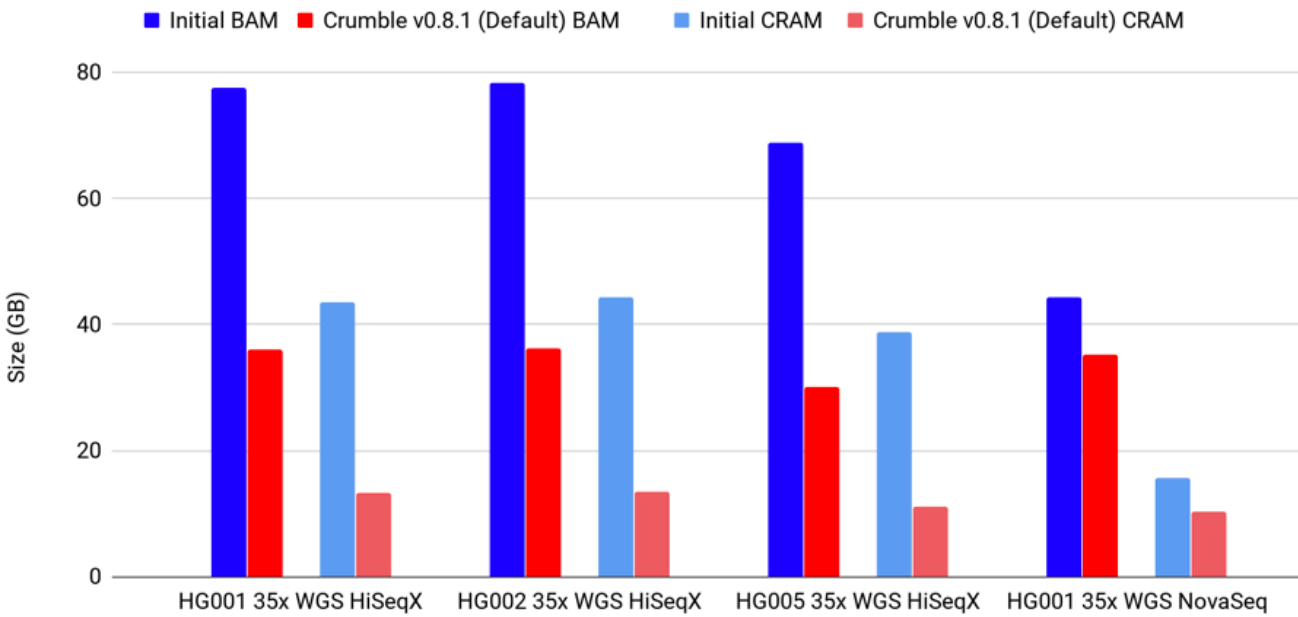
Crumble: reference free lossy compression of sequence quality values

James K Bonfield, Shane A McCarthy, Richard Durbin

Bioinformatics, Volume 35, Issue 2, 15 January 2019, Pages 337–339,
https://doi.org/10.1093/bioinformatics/bty608
Published: 10 July 2018 **Article history**

jkbonfield / **crumble**

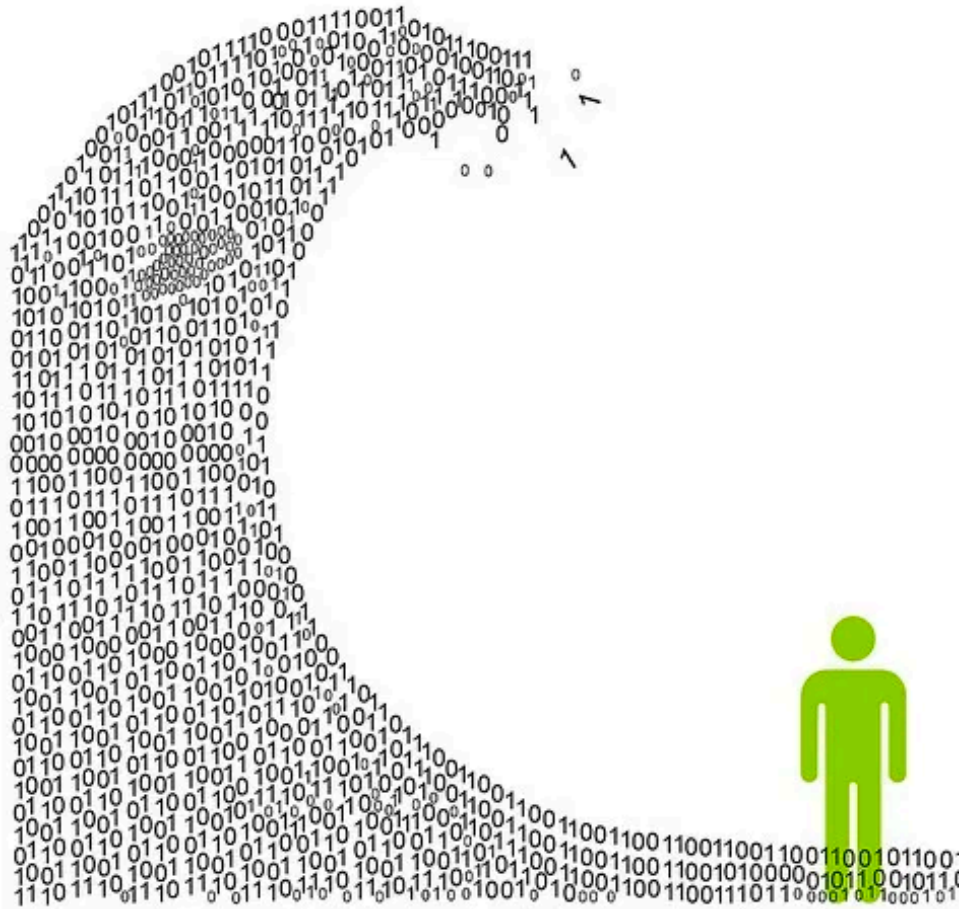
Figure 2. File Size Before and After Crumble Compression on 35X WGS



https://blog.dnanexus.com/2018-07-23-breaking-down-crumble/

How much data will I get

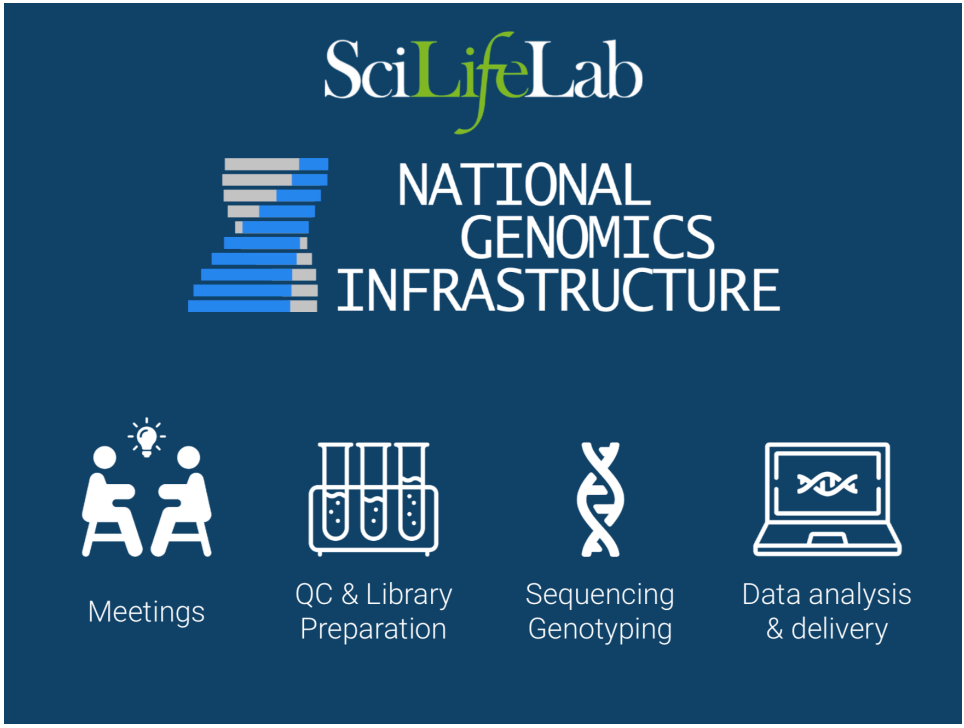
	WGS 30x	RNA-seq (20M)
Fastq.gz	50 Gb	4-5 Gb
bam	80 Gb	6-8 Gb
100 samples	**13 Tb**	**1-1.3 Tb**



Integrating information from NGI in data management plan

Integrating NGI orders into Data Management Plan

- What data to get
  - Contact NGI / NGI website
  - http://ngisweden.scilifelab.se
- Deliveries:
  - GRUS
  - Hard drive (not recommended)
- No long term storage obligation
  - Plan for storage accordingly
  - SNIC / ENA / other backup



Thanks for your attention

Carl-Johan Rubin
@callerubin
carl.rubin@scilifelab.se

support@ngisweden.se
http://ngisweden.scilifelab.se
http://opensource.scilifelab.se