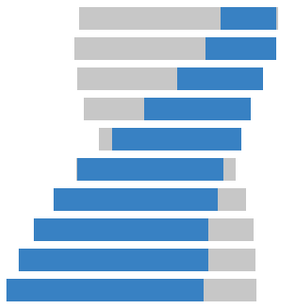


NGI-RNAseq

Processing RNA-seq data at the National Genomics Infrastructure

SciLifeLab

NGI stockholm

Phil Ewels
phil.ewels@scilifelab.se

NBIS RNA-seq tutorial
2017-11-09

SciLifeLab NGI

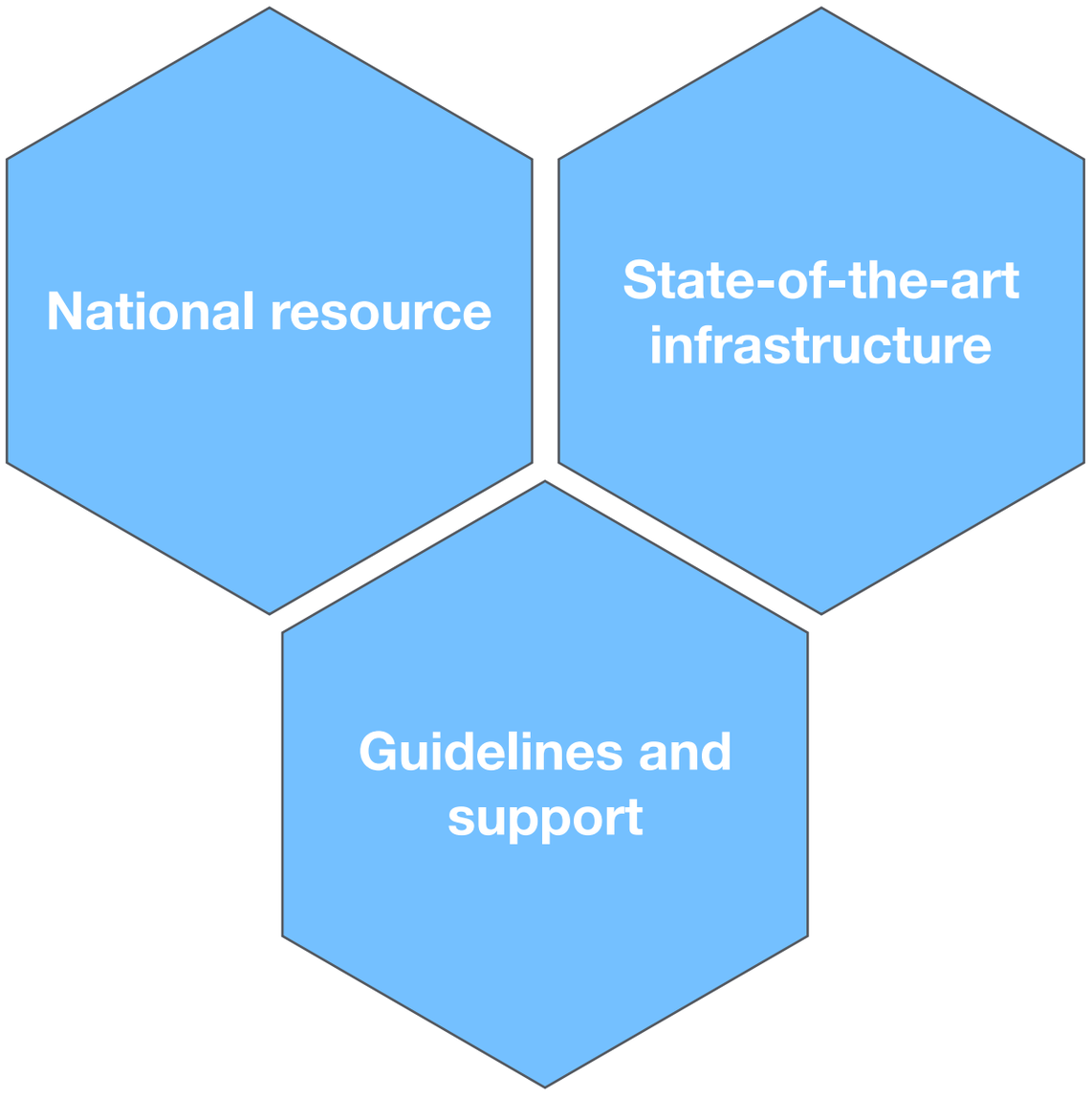
GANATIONALCTAC
ATCAGENOMICSGT
INFRASTRUCTURE

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

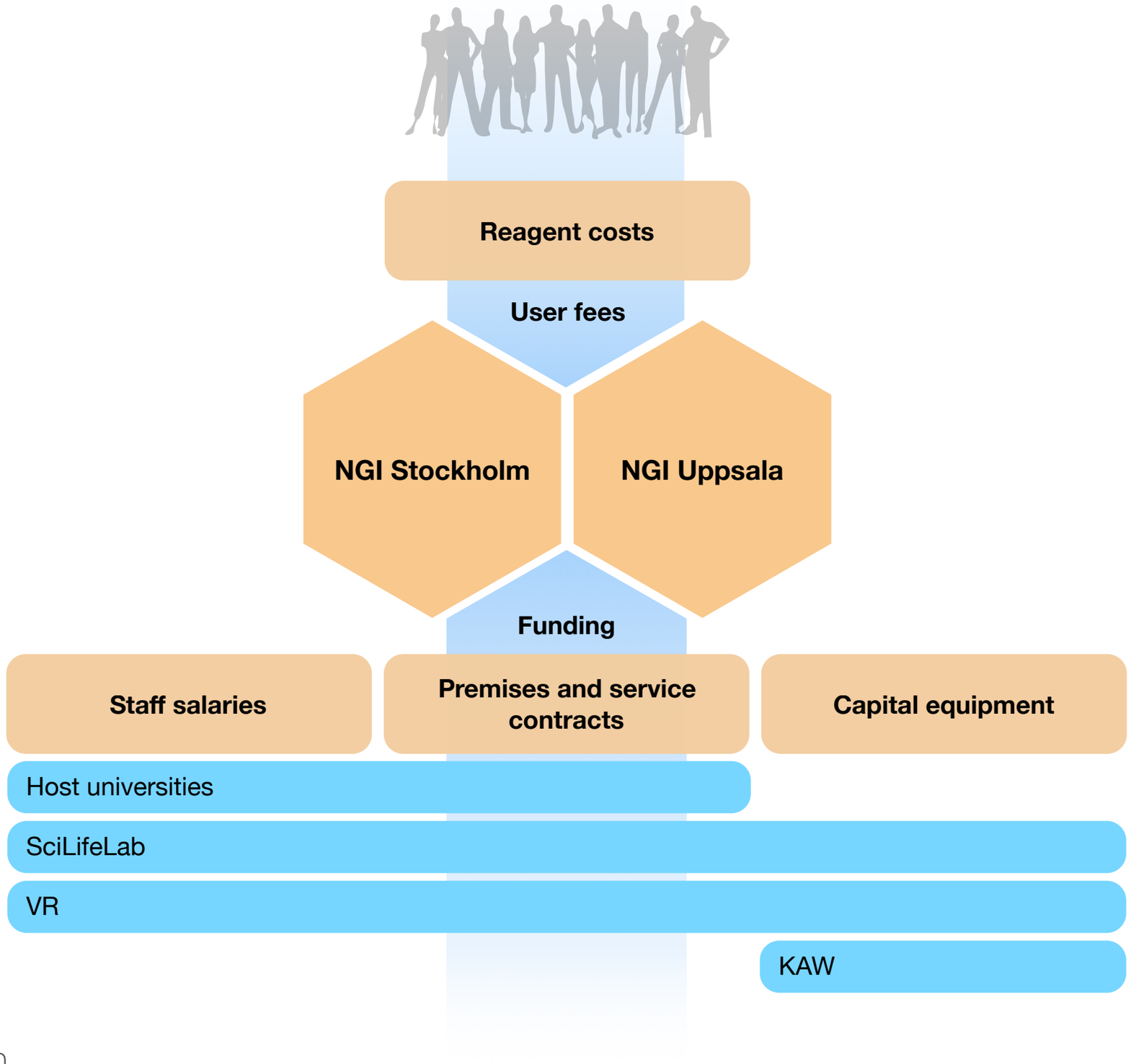
SciLifeLab
NGI stockholm

NGI Organisation

SciLifeLab

NGI stockholm

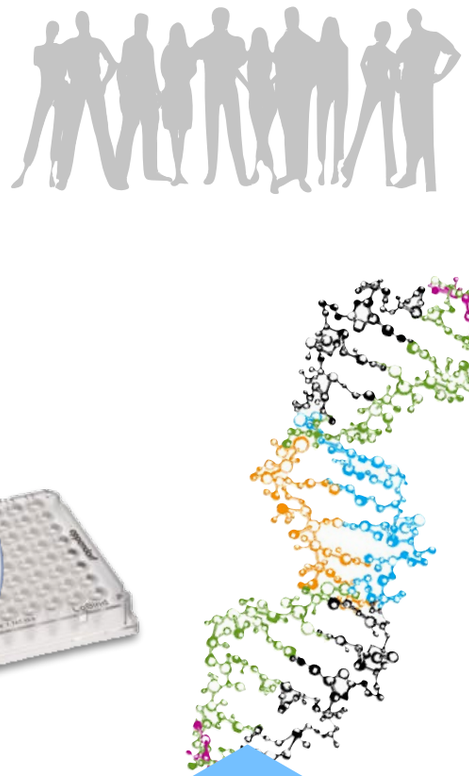
NGI Organisation

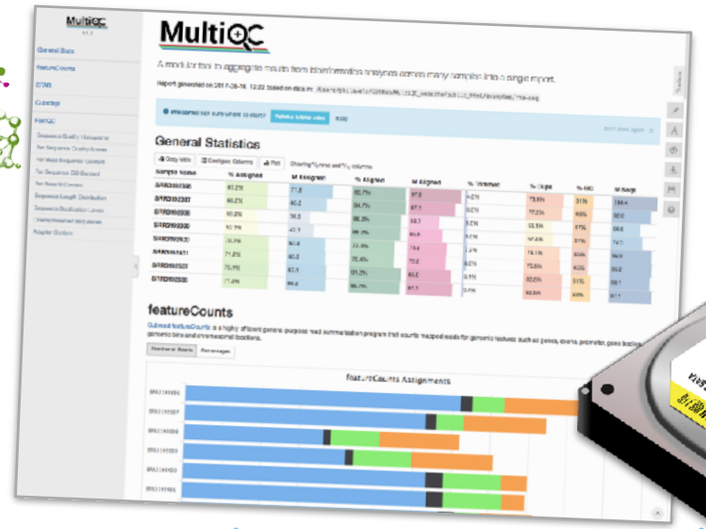


SciLifeLab

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

Methods offered at NGI

Accredited methods

SWEDAC ACKREDITERING
Ackred. nr 1850
Provning
ISO/IEC 17025

Whole Genome seq

RNA-seq

de novo

Just Sequencing

Data analysis included for **FREE**

Metagenomics

Nanopore sequencing

Exome sequencing

RAD-seq

Bisulphite sequencing

ChIP-seq

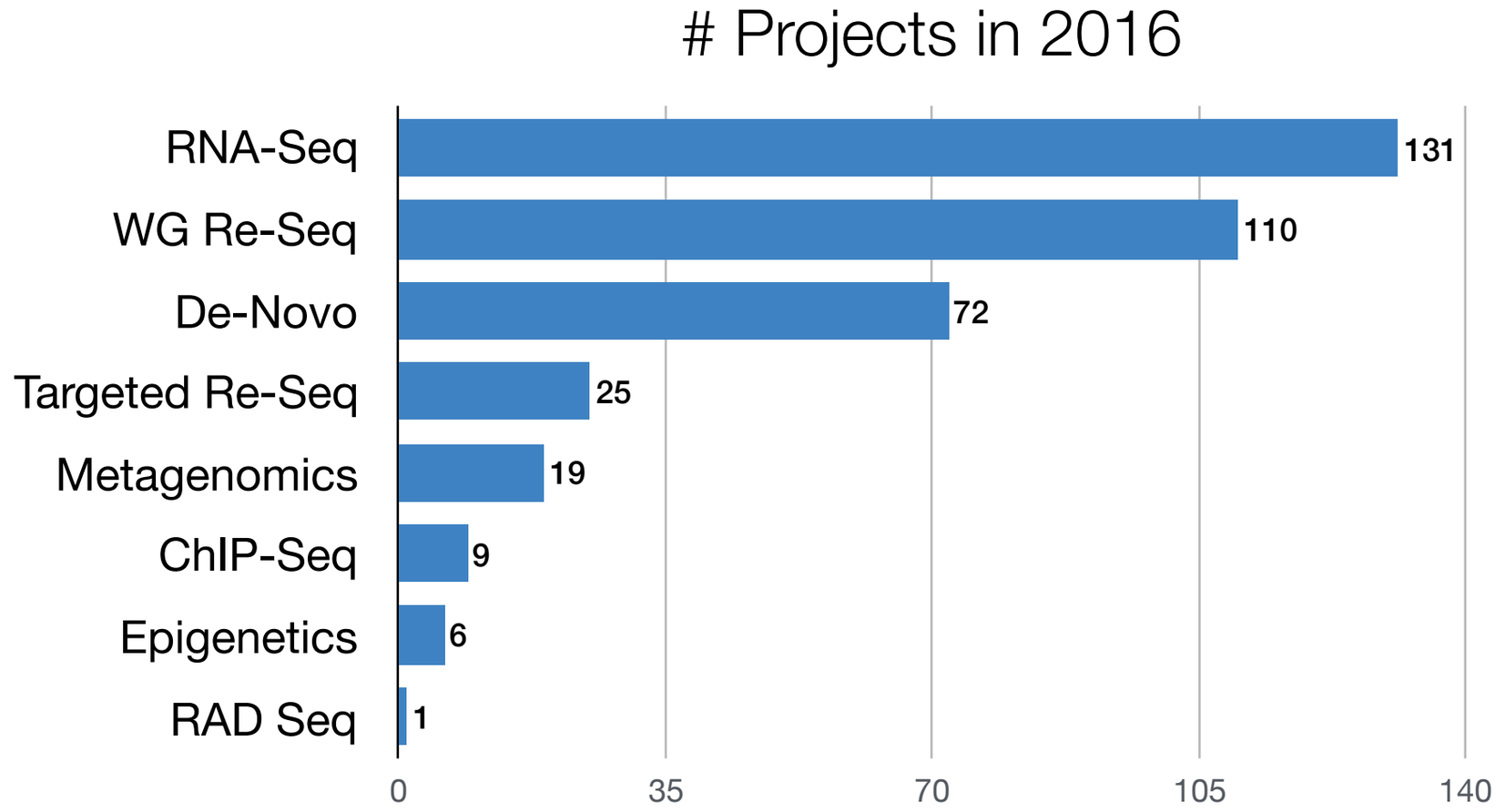
ATAC-seq

SciLifeLab

NGI stockholm

RNA-Seq: NGI Stockholm

- RNA-seq is the most common project type

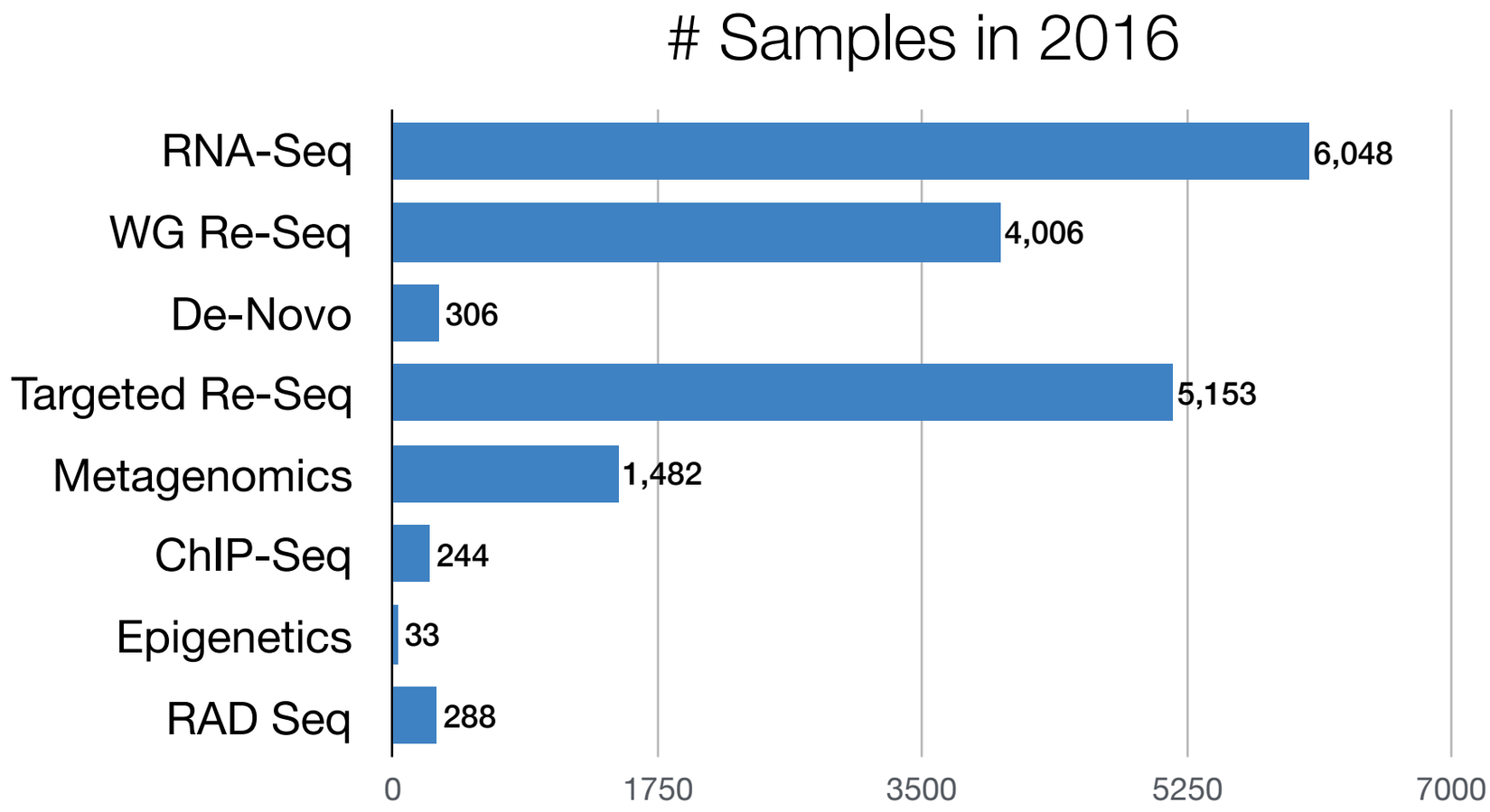


SciLifeLab

NGI stockholm

RNA-Seq: NGI Stockholm

- RNA-seq is the most common project type
- Production protocols:
  - TruSeq (poly-A)
  - RiboZero
- In development:
  - SMARTer Pico
  - RNA Access



SciLifeLab

NGI stockholm

RNA-Seq: NGI Stockholm

- RNA-seq is the most common project type
- Production protocols:
  - TruSeq (poly-A)
  - RiboZero
- In development:
  - SMARTer Pico
  - RNA Access

SciLifeLab

NGI stockholm

RNA-Seq Pipeline

- Takes raw FastQ sequencing data as input
- Provides range of results
  - Alignments (BAM)
  - Gene counts (Counts, FPKM)
  - Quality Control
- First RNA Pipeline running since 2012
- Second RNA Pipeline in use since April 2017

SciLifeLab

NGI stockholm

NGI-RNAseq

RNA-Seq Pipeline

NGI-RNAseq

FastQC	*Sequence QC*
TrimGalore!	*Read trimming*
STAR	*Alignment*
dupRadar	*Duplication QC*
featureCounts	*Gene counts*
StringTie	*Normalised FPKM*
RSeQC	*Alignments QC*
Preseq	*Library complexity*
edgeR	*Heatmap, clustering*
MultiQC	*Reporting*

SciLifeLab

NGI stockholm

RNA-Seq Pipeline

NGI-RNAseq

FastQ

BAM

TSV

HTML

FastQC	*Sequence QC*
TrimGalore!	*Read trimming*
STAR	*Alignment*
dupRadar	*Duplication QC*
featureCounts	*Gene counts*
StringTie	*Normalised FPKM*
RSeQC	*Alignments QC*
Preseq	*Library complexity*
edgeR	*Heatmap, clustering*
MultiQC	*Reporting*

SciLifeLab

NGI stockholm

Nextflow

- Tool to manage computational pipelines
- Handles interaction with compute infrastructure
- Easy to learn how to run, minimal oversight required

SciLifeLab

NGI stockholm

Nextflow

nextflow

```
#!/usr/bin/env nextflow

cheers=Channel.from "Bonjour","Ciao","Hello","Hola"

process sayHello {
  input:
  val x from cheers

  """
  echo $x world!
  """
}
```

SciLifeLab

NGI stockholm

https://www.nextflow.io/

Nextflow


```
#!/usr/bin/env nextflow

input = Channel.fromFilePairs( params.reads )
process fastqc {
  input:
  file reads from input

  output:
  file "*_fastqc.{zip,html}" into results

  script:
  """
  fastqc -q $reads
  """
}
```

SciLifeLab

NGI stockholm

Nextflow

```
#!/usr/bin/env nextflow

input = Channel.fromFilePairs( params.reads )
process fastqc {
  input:
  file reads from input

  output:
  file "*_fastqc.{zip,html}" into results

  script:
  """
  fastqc -q $reads
  """
}
```

Default: Run locally, assume software is installed

```
process {

  executor = 'slurm'
  clusterOptions = { "-A b2017123" }

  cpus = 1
  memory = 8.GB
  time = 2.h

  $fastqc {
    module = ['bioinfo-tools', ‘FastQC']
  }

}
```

Submit jobs to SLURM queue
Use environment modules

UPPNEX

SciLifeLab

NGI stockholm

Nextflow

```
#!/usr/bin/env nextflow

input = Channel.fromFilePairs( params.reads )
process fastqc {
  input:
  file reads from input

  output:
  file "*_fastqc.{zip,html}" into results

  script:
  """
  fastqc -q $reads
  """
}
```

```
process {

  executor = 'slurm'
  clusterOptions = { "

  cpus = 1
  memory = 8.GB
  time = 2.h

  $fastqc {
    module = ['bioinfo-tools', ‘FastQC']
  }

}
```

```
docker {
 enabled = true
}

process {
  container = 'biocontainers/fastqc'

  cpus = 1
  memory = 8.GB
  time = 2.h
}
```

Run locally, use docker container for all software dependencies

SciLifeLab

NGI stockholm

NGI-RNAseq

This repository Search Pull requests Issues Marketplace Explore

SciLifeLab / NGI-RNAseq
forked from ewels/NGI-RNAseq

Unwatch 13 Unstar 23 Fork 19

Code Issues 13 Pull requests 1 Insights Settings

Nextflow RNA-Seq Best Practice analysis pipeline, used at the SciLifeLab National Genomics Infrastructure.
https://ngisweden.scilifelab.se/

Edit

bioinformatics rna-seq rnaseq rna-sequencing pipeline nextflow Manage topics

598 commits 3 branches 12 releases 8 contributors MIT

Branch: master New pull request Create new file Upload files Find file Clone or download

This branch is even with ewels:master. Pull request Compare

ewels Merge pull request #162 from na399/na399-patch-1 ... Latest commit 6dbb58a 5 days ago

assets	Got software version stuff working.	4 months ago
bin	Dynamically generate base64 images for pipeline report. New 'assets' ...	6 months ago
conf	Merge branch 'master' of github.com:SciLifeLab/NGI-RNAseq	26 days ago

SciLifeLab

NGI stockholm

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

NGI-RNAseq

README.md

NGI-RNAseq Documentation

The NGI-RNAseq documentation is split into a few different files:

- `installation.md`
  - Pipeline installation and configuration instructions
- `usage.md`
  - Instructions on how to run the NGI-RNAseq pipeline
- `output.md`
  - Document describing all of the results produced by the pipeline, and how to interpret them.
- `amazon_web_services.md`
  - Docs about running the pipeline in the cloud with AWS.

SciLifeLab

NGI stockholm

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

Running NGI-RNAseq

Step 1: Install Nextflow

- Uppmax - load the Nextflow module
  `module load nextflow`
- Anywhere (including Uppmax) - install Nextflow

```
curl -s https://get.nextflow.io | bash
```

Step 2: Try running NGI-RNAseq pipeline

```
nextflow run SciLifeLab/NGI-RNAseq --help
```

SciLifeLab

NGI stockholm

Running NGI-RNAseq

Step 3: Choose your reference

- Common organism - use iGenomes

  `--genome GRCh37`

- Custom genome - Fasta + GTF (minimum)

  `--fasta genome.fa --gtf genes.gtf`

Step 4: Organise your data

- One (if single-end) or two (if paired-end) FastQ per sample
- Everything in one directory, simple filenames help!

SciLifeLab

NGI stockholm

Running NGI-RNAseq

Step 5: Run the pipeline on your data

- Remember to run detached from your terminal
  `screen / tmux / nohup`

Step 6: Check your results

- Read the Nextflow log and check the MultiQC report

Step 7: Delete temporary files

- Delete the `./work` directory, which holds all intermediates

SciLifeLab

NGI stockholm

Typical pipeline output

SciLifeLab

NGI stockholm

Using UPPMAX

```
nextflow run SciLifeLab/NGI-RNAseq
    --project b2017123
    --genome GRCh37
    --reads "data/*_R{1,2}.fastq.gz"
```

- Default config is for UPPMAX
  - Knows about central iGenomes references
  - Uses centrally installed software

SciLifeLab

NGI stockholm

Using other clusters

```
nextflow run SciLifeLab/NGI-RNAseq
    -profile hebbe
    --fasta genome.fa --gtf genes.gtf
    --reads "data/*_R{1,2}.fastq.gz"
```

C3SE

BIOCONDA

- Can run just about anywhere
  - Supports local, SGE, LSF, SLURM, PBS/Torque, HTCondor, DRMAA, DNAnexus, Ignite, Kubernetes

SciLifeLab

NGI stockholm

Using Docker

```
nextflow run SciLifeLab/NGI-RNAseq
    -profile docker
    --fasta genome.fa --gtf genes.gtf
    --reads "data/*_R{1,2}.fastq.gz"
```

- Can run anywhere with Docker
  - Downloads required software and runs in a container
  - Portable and reproducible.

SciLifeLab

NGI stockholm

Using AWS

```
nextflow run SciLifeLab/NGI-RNAseq
    -profile aws
    --genome GRCh37
    --reads "s3://my-bucket/*_{1,2}.fq.gz"
    --outdir "s3://my-bucket/results/"
```

- Runs on the AWS cloud with Docker
  - Pay-as-you go, flexible computing
  - Can launch from anywhere with minimal configuration

SciLifeLab

NGI stockholm

Input data

```
ERROR ~ Cannot find any reads matching: XXXX
NB: Path needs to be enclosed in quotes!
NB: Path requires at least one * wildcard!
If this is single-end data, please specify
--singleEnd on the command line.
```

```
--reads '*_R{1,2}.fastq.gz'
--reads '*.fastq.gz' --singleEnd
```

```
--reads sample.fastq.gz
--reads *_R{1,2}.fastq.gz
--reads '*.fastq.gz'
```

SciLifeLab

NGI stockholm

Read trimming

- Pipeline runs TrimGalore! to remove adapter contamination and low quality bases automatically

- Some library preps also include additional adapters
  - Will get poor alignment rates without additional trimming

```
--clip_r1 [int]
--clip_r2 [int]
--three_prime_clip_r1 [int]
--three_prime_clip_r2 [int]
```

SciLifeLab

NGI stockholm

Library strandedness

- Most RNA-seq data is strand-specific now
  - Can be "forward-stranded" (same as transcript) or "reverse-stranded" (opposite to transcript)

- UPPMAX config runs as reverse stranded by default

- If wrong, QC will say most reads don't fall within genes

```
--forward_stranded
--reverse_stranded
--unstranded
```

SciLifeLab

NGI stockholm

Lib-prep presets

- There are some presets for common kits
- Clontech SMARTer PICO
  - Forward stranded, needs R1 5' 3bp and R2 3' 3bp trimming

```
--pico
```

- Please suggest others!

SciLifeLab

NGI stockholm

Saving intermediates

- By default, the pipeline doesn't save some intermediate files to your final results directory
  - Reference genome indices that have been built
  - FastQ files from TrimGalore!
  - BAM files from STAR (we have BAMs from Picard)

```
--saveReference
--saveTrimmed
--saveAlignedIntermediates
```

SciLifeLab

NGI stockholm

Resuming pipelines

- If something goes wrong, you can resume a stopped pipeline
  - Will use cached versions of completed processes
  - NB: Only one hyphen!

```
-resume
```

- Can resume specific past runs
  - Use `nextflow log` to find job names

```
-resume job_name
```

SciLifeLab

NGI stockholm

Customising output

`-name`	Give a name to your run. Used in logs and reports
`--outdir`	Specify the directory for saved results
`--aligner hisat2`	Use HiSAT2 instead of STAR for alignment
`--email`	Get e-mailed a summary report when the pipeline finishes

SciLifeLab

NGI stockholm

Nextflow config files

- Can save a config file with defaults
  - Anything with two hyphens is a params

```
./nextflow.config

~/.nextflow/config

-c /path/to/my.config
```

```
params {

  email = 'phil.ewels@scilifelab.se'
  project = "b2017123"

}

process.$multiqc.module = []
```

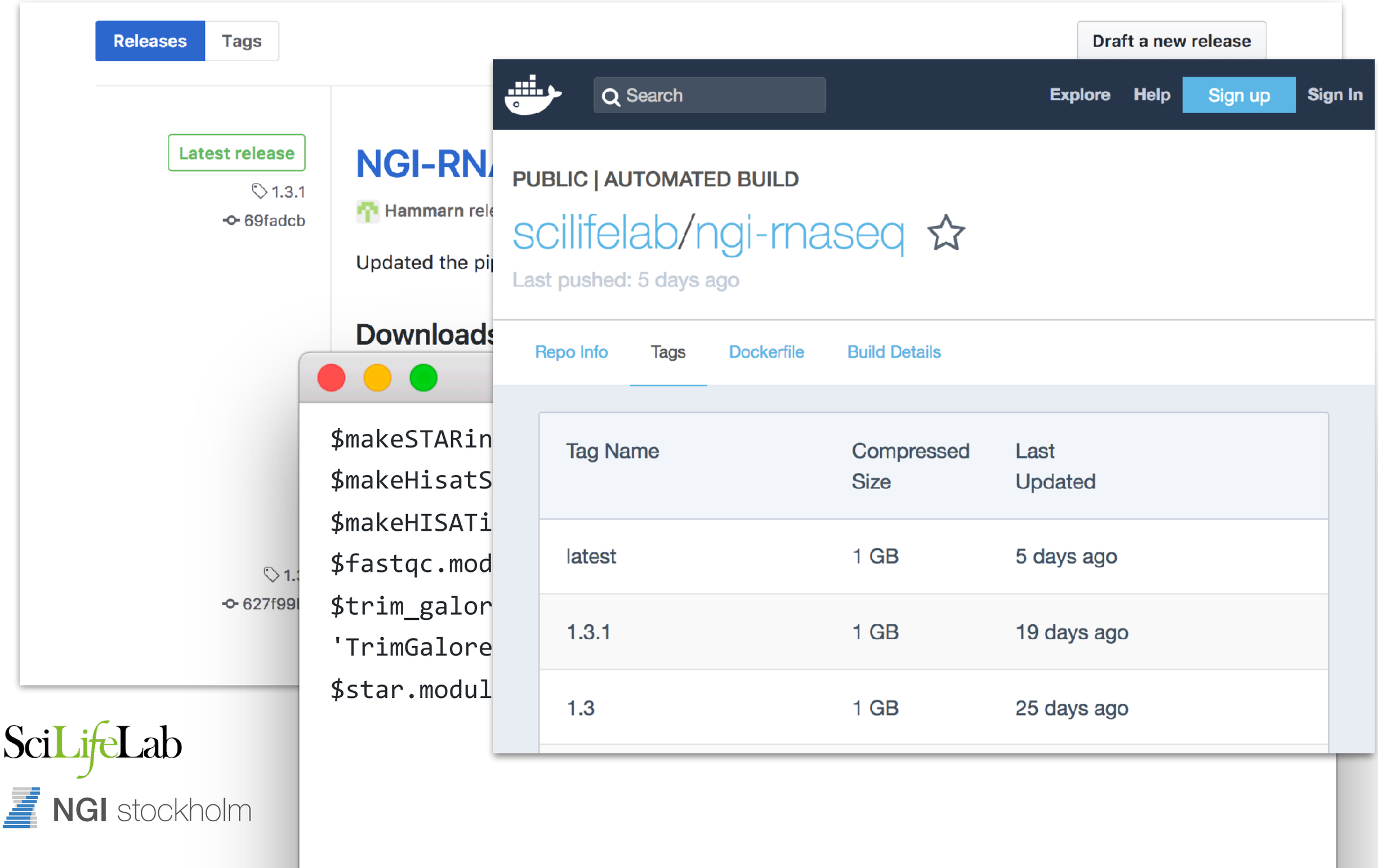
SciLifeLab

NGI stockholm

NGI-RNAseq config

```
N E X T F L O W  ~  version 0.25.5
Launching `/home/phil/GitHub/NGI-RNAseq/main.nf` [amazing_laplace] - revision: 8b9f416d01
=========================================
 NGI-RNAseq : RNA-Seq Best Practice v1.3.1
=========================================
Run Name        : amazing_laplace
Reads           : data/7_111116_AD0341ACXX_137_*_{1,2}.fastq.gz
Data Type       : Paired-End
Genome          : GRCh37
Strandedness    : Reverse
Trim R1         : 0
Trim R2         : 0
Trim 3' R1      : 0
Trim 3' R2      : 0
Aligner         : STAR
STAR Index      : /sw/data/uppnex/igenomes//Homo_sapiens/Ensembl/GRCh37/Sequence/STARIndex/
GTF Annotation : /sw/data/uppnex/igenomes//Homo_sapiens/Ensembl/GRCh37/Annotation/Genes/genes.gtf
BED Annotation : /sw/data/uppnex/igenomes//Homo_sapiens/Ensembl/GRCh37/Annotation/Genes/genes.bed
Save Reference : Yes
Save Trimmed    : No
Save Intermeds : No
Output dir      : ./results
Working dir     : /pica/h1/phil/nbis_rnaseq/work
Current home    : /home/phil
Current user    : phil
Current path    : /home/phil/nbis_rnaseq
R libraries     : /home/phil/R/nxtflow_libs/
Script dir      : /home/phil/GitHub/NGI-RNAseq
Config Profile : UPPMAX
UPPMAX Project : b2017123
E-mail Address : phil.ewels@scilifelab.se
=========================================
```

Version control



SciLifeLab

NGI stockholm

Version control

- Pipeline is always released under a stable version tag
- Software versions and code reproducible
- For full reproducibility, specify version revision when running the pipeline

```
nextflow run SciLifeLab/NGI-RNAseq -r v1.3.1
```

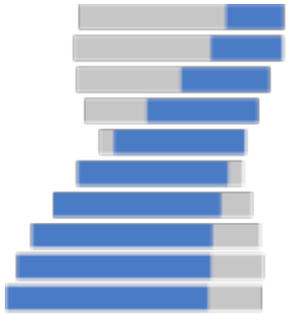
SciLifeLab

NGI stockholm

Conclusion

- **Use NGI-RNAseq to prepare your data if you want:**
  - To not have to remember every parameter for STAR
  - Extreme reproducibility
  - Ability to run on virtually any environment

- Now running for all RNA projects at NGI-Stockholm



NGI-RNAseq

Conclusion

GitHub https://github.com/

Open Source Initiative

NGI-RNAseq SciLifeLab/NGI-RNAseq

NGI-smRNAseq SciLifeLab/NGI-smRNAseq

NGI-MethylSeq SciLifeLab/NGI-MethylSeq

NGI-ChIPseq SciLifeLab/NGI-ChIPseq

MIT Licence

SciLifeLab

NGI stockholm

Conclusion

GitHub

https://github.com/

NGI-RNAseq

SciLifeLab/NGI-RNAseq

NGI-smRNAseq

SciLifeLab/NGI-smRNAseq

NGI-MethylSeq

SciLifeLab/NGI-MethylSeq

NGI-ChIPseq

SciLifeLab/NGI-ChIPseq

Acknowledgements

Phil Ewels

Rickard Hammarén

Anders Jemt

Max Käller

Denis Moreno

Chuan Wang

NGI Stockholm Genomics Applications Development Group

support@ngisweden.se

http://opensource.scilifelab.se

SciLifeLab

NGI stockholm