NGI-ChIPseq

Processing ChIP-seq data at the National Genomics Infrastructure

SciLifeLab

NGI stockholm

Phil Ewels
phil.ewels@scilifelab.se
NBIS ChIP-seq tutorial
2017-11-29

SciLifeLab NGI

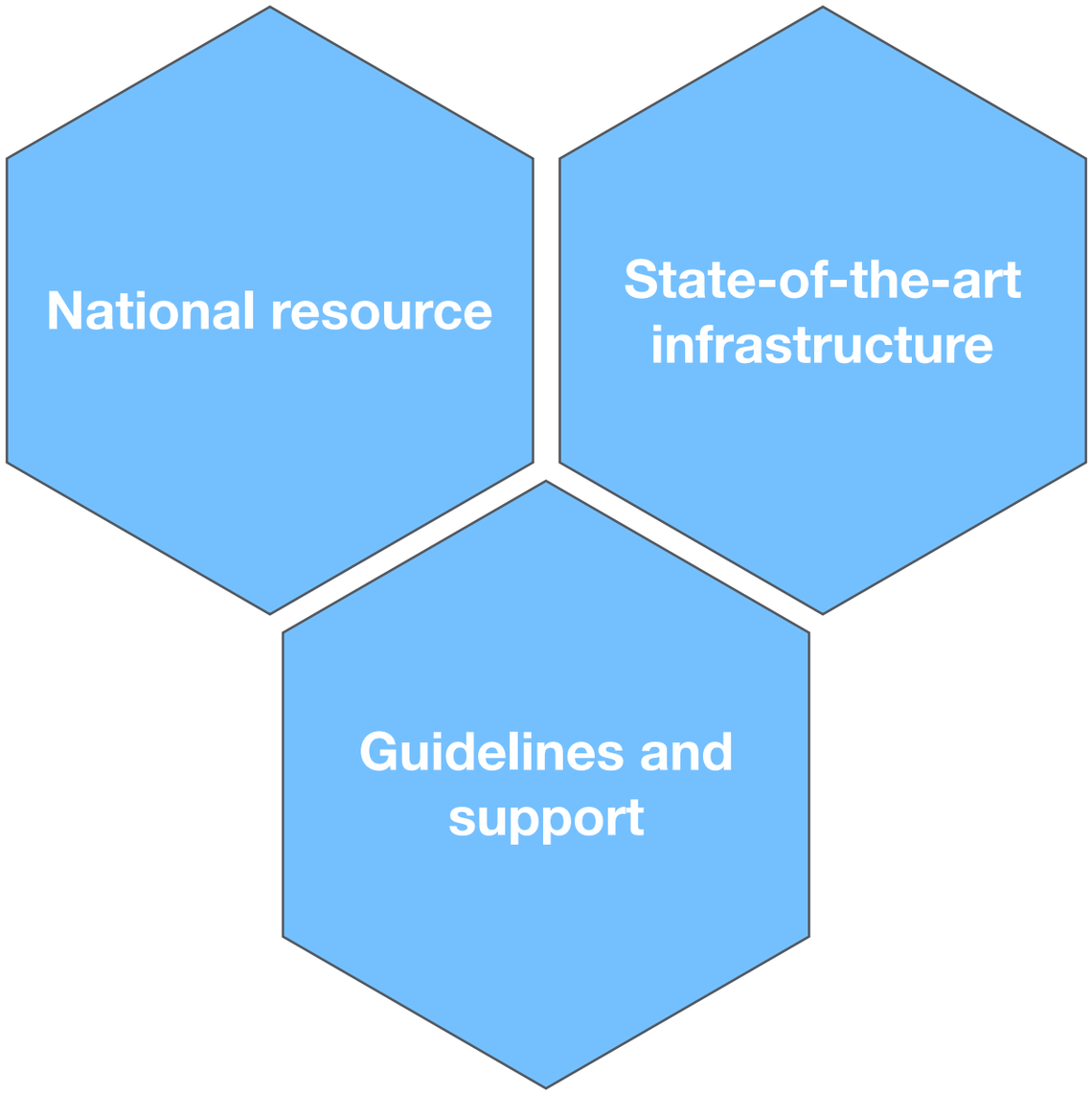
NATIONAL GENOMICS INFRASTRUCTURE

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



NATIONAL GENOMICS INFRASTRUCTURE

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

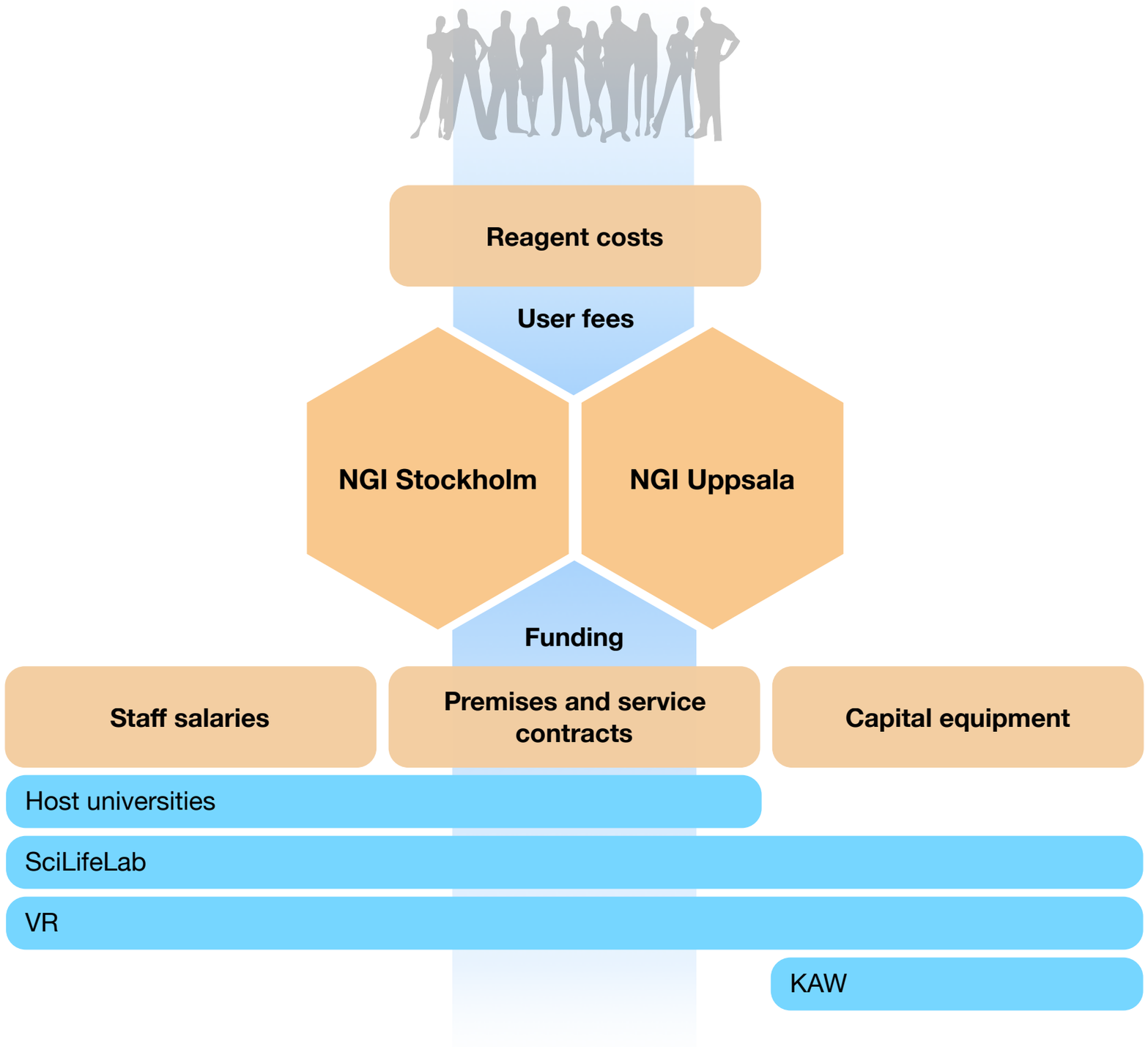
SciLifeLab
NGI stockholm

NGI Organisation

SciLifeLab

NGI stockholm

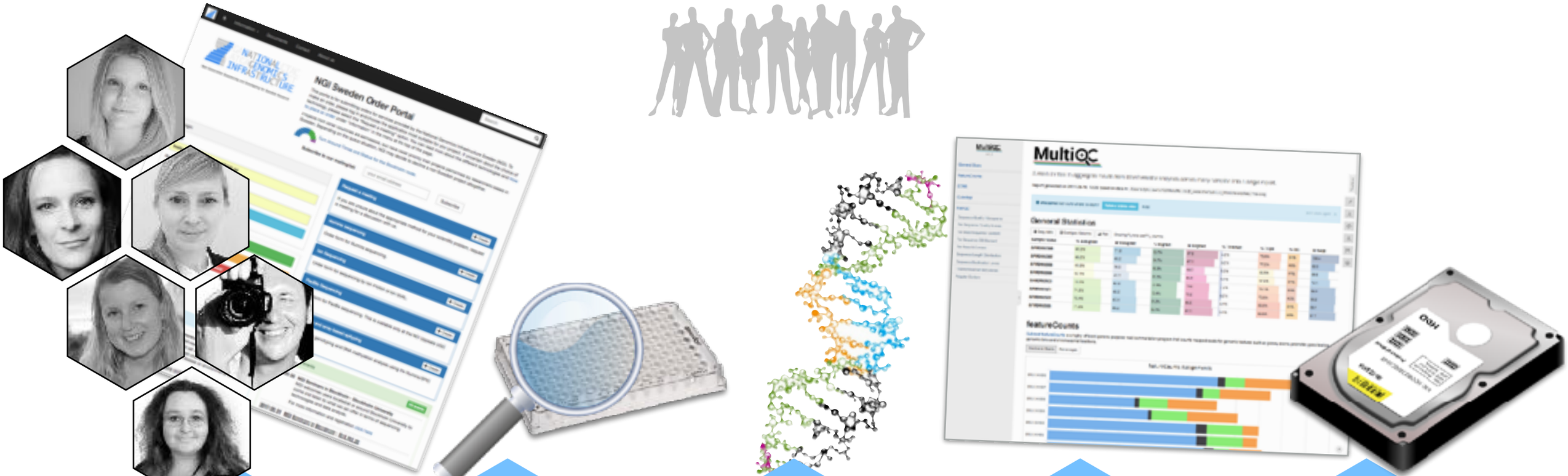
NGI Organisation



SciLifeLab

NGI stockholm

Project timeline



Scientific support and project consultation

Sample QC

Library preparation, Sequencing, Genotyping

Data processing and primary analysis

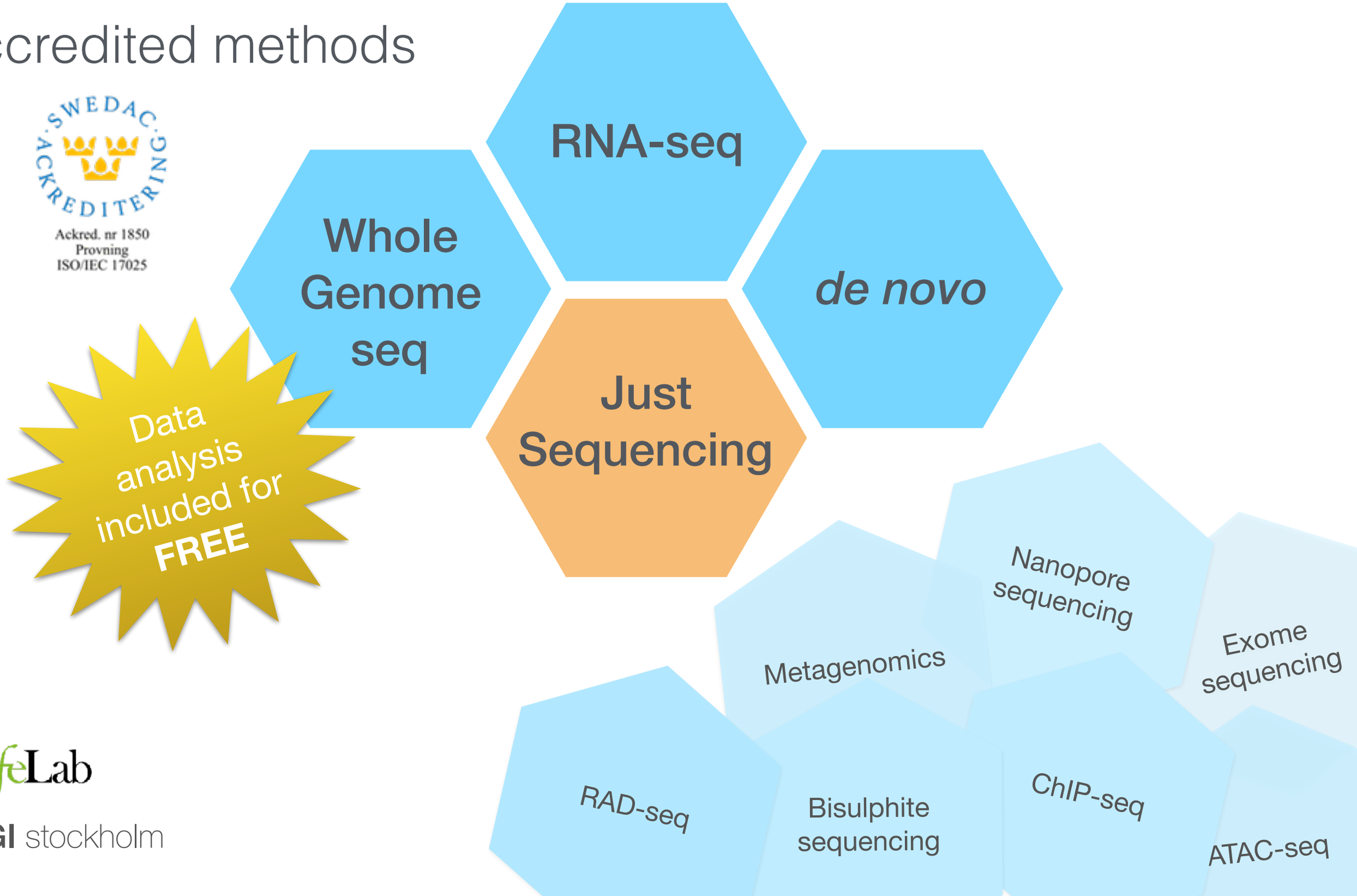
Data delivery

SciLifeLab

NGI stockholm

Methods offered at NGI

Accredited methods



SciLifeLab

NGI stockholm

ChIP-seq: NGI Stockholm

- You do the ChIP, we do the seq
- Rubicon ThruPlex DNA (NGI Production)
  - Min 1 ng input
  - Min 10 µl
  - 0.2-10 ng/µl
  - Ins. size 200-800 bp
  - 963 kr / prep

SciLifeLab

NGI stockholm

ChIP-seq: NGI Stockholm

- You do the ChIP, we do the seq
- Rubicon ThruPlex DNA (NGI Production)
- Typically run SE 50bp
  - Illumina HiSeq High Output mode v4, SR 1x50bp
  - 1226 kr / sample (40M reads)

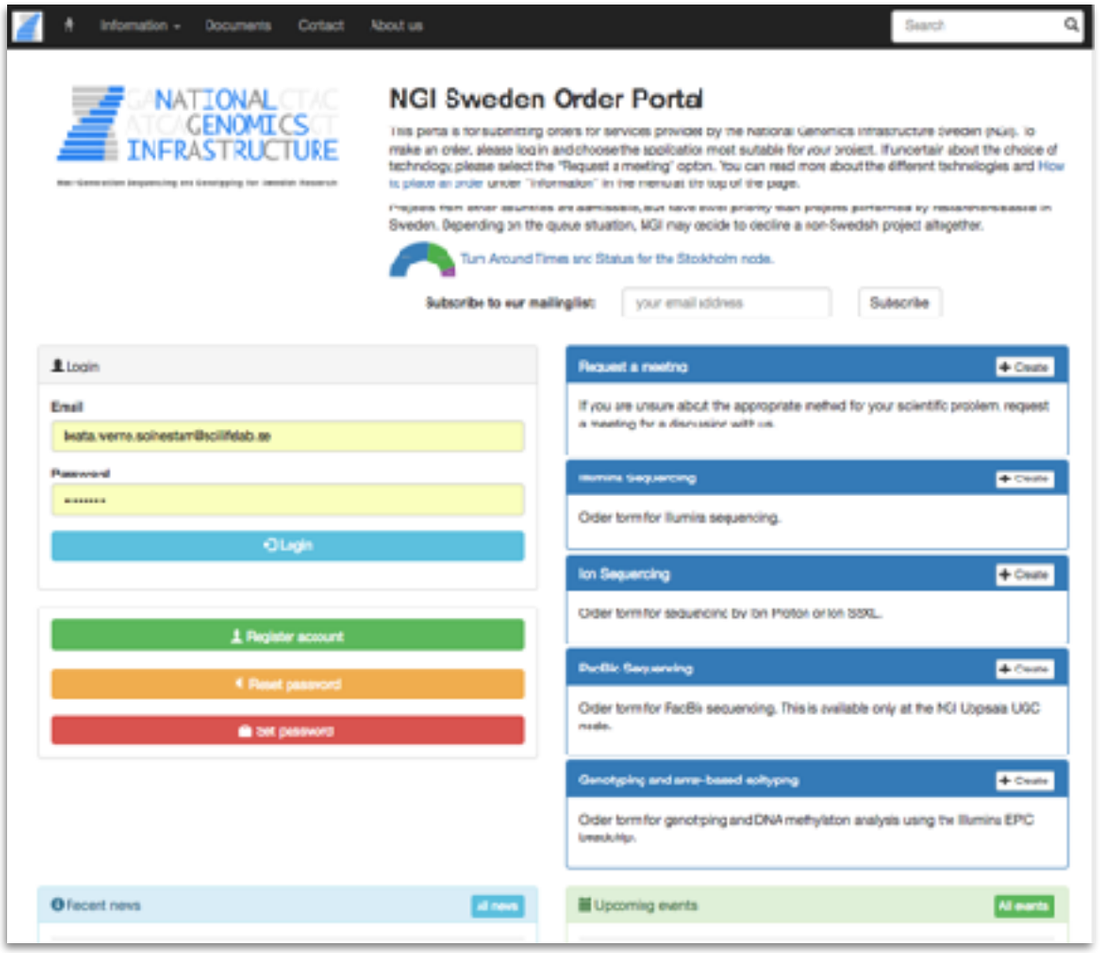
SciLifeLab

NGI stockholm

ChIP-seq: NGI Stockholm

- You do the ChIP, we do the seq
- Rubicon ThruPlex DNA (NGI Production)
- Typically run SE 50bp
- Start by organising a planning meeting

https://ngisweden.scilifelab.se



SciLifeLab

NGI stockholm

ChIP-seq Pipeline

- Takes raw FastQ sequencing data as input
- Provides range of results
  - Alignments (BAM)
  - Peaks (optionally filtered)
  - Quality Control
- Pipeline in use since early 2017 (on request)

SciLifeLab

NGI stockholm

NGI-ChIPseq

ChIP-seq Pipeline

NGI-ChIPseq

FastQ	FastQC	*Sequence QC*
	TrimGalore!	*Read trimming*
BAM	BWA	*Alignment*
	Samtools, Picard	*Sort, index, mark duplicates*
	Phantompeakqualtools	*Strand cross-correlation QC*
	deepTools	*Fingerprint, sample correlation*
	NGSPlot	*TSS / Gene profile plots*
BED	MACS2	*Peak calling*
	Bedtools	*Filtering blacklisted regions*
HTML	MultiQC	*Reporting*

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

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

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
  """
}
```

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

UPPNEX

SciLifeLab

NGI stockholm

NGI-ChIPseq

This repository Search Pull requests Issues Marketplace Explore

SciLifeLab / NGI-ChIPseq
forked from chuan-wang/NGI-ChIPseq

Unwatch 7 Star 2 Fork 9

Code Issues 5 Pull requests 0 Projects 0 Wiki Insights Settings

Nextflow ChIP-seq data analysis pipeline, National Genomics Infrastructure, Science for Life Laboratory in Stockholm Edit

nextflow bioinformatics bioinformatics-pipeline chip-seq pipeline Manage topics

178 commits 1 branch 0 releases 2 contributors

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

This branch is 2 commits ahead of chuan-wang:master. Pull request Compare

ewels Merge pull request #49 from chuan-wang/master ··· Latest commit 51b7c2d 6 days ago

assets	Fix errors in scripts for reporting software versions; Add scripts fo...	7 days ago
bin	Fixing bugs	6 days ago
blacklists	v1.4 Updates: Adding post peak calling filtering and annotations; Gen...	14 days ago

SciLifeLab

NGI stockholm

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

NGI-ChIPseq

README.md

NGI-ChIPseq Results

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

- `installation.md`
  - Pipeline installation and configuration instructions
- `usage.md`
  - Instructions on how to run the NGI-ChIPseq pipeline
- `output.md`
  - Document describing all of the results produced by the pipeline, and how to interpret them.

SciLifeLab NATIONAL GENOMICS

SciLifeLab
NGI stockholm

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

Running NGI-ChIPseq

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-ChIPseq pipeline

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

SciLifeLab

NGI stockholm

Running NGI-ChIPseq

Step 3: Choose your reference

- Common organism - use iGenomes

```
--genome GRCh37
```

- MACS peak calling config file

```
--macsconfig config.csv
```

Step 4: Organise your data

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

Running NGI-ChIPseq

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

Using UPPMAX

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

UPPNEX

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

SciLifeLab

NGI stockholm

Using other clusters

```
nextflow run SciLifeLab/NGI-ChIPseq
    -profile hebbe
    --bwaindex ./ref --macsconfig p.txt
    --reads "data/*_R{1,2}.fastq.gz"
```

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

Using Docker

```
nextflow run SciLifeLab/NGI-ChIPseq
    -profile docker
    --fasta genome.fa --macsconfig p.txt
    --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-ChIPseq
    -profile aws
    --genome GRCh37 --macsconfig p.txt
    --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
  - Use `--notrim` to disable this

- Some library preps also include additional adapters

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

Blacklist filtering

- Some parts of the reference genome collect incorrectly mapped reads
  - Good practice to remove these peaks
- Pipeline has ENCODE regions for Human & Mouse
- Can pass own BED file of custom regions

```
--blacklist_filtering
--blacklist regions.bed
```

Broad Peaks

- Some chromatin profiles don't have narrow, sharp peaks
  - For example, H3K9me3 & H3K27me3
- MACS2 can call peaks in "broad peak" mode
  - Pipeline uses default qvalue cutoff of 0.1

```
--broad
```

Extending Read Length

- When using single-end data, sequenced read length is shorter than the sequence fragment length
- For DeepTools, need to "extend" the read length
  - Set to 100bp by default. Use this parameter to customise this value.
  - Expected fragment length - sequence read length

```
--extendReadsLen [int]
```

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
`--saturation`	Run saturation analysis, subsampling reads from 10% - 100%
`--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-ChIPseq config

```
N E X T F L O W  ~  version 0.26.1
Launching `SciLifeLab/NGI-ChIPseq` [deadly_bose] - revision: 28e24c2a2a
=========================================
 NGI-ChIPseq: ChIP-Seq Best Practice v1.4
=========================================
Run Name             : deadly_bose
Reads                : data/*fastq.gz
Data Type            : Single-End
Genome               : GRCh37
BWA Index            : /sw/data/uppnex/igenomes//Homo_sapiens/Ensembl/GRCh37/Sequence/BWAIndex/
MACS Config          : data/macsconfig.txt
Saturation analysis  : false
MACS broad peaks     : false
Blacklist filtering  : false
Extend Reads         : 100 bp
Current home         : /home/phil
Current user         : phil
Current path         : /home/phil/demo_data/ChIP/Human/test
Working dir          : /home/phil/demo_data/ChIP/Human/test/work
Output dir           : ./results
R libraries          : /home/phil/R/nxtflow_libs/
Script dir           : /home/phil/GitHub/NGI-ChIPseq
Save Reference       : false
Save Trimmed         : false
Save Intermeds       : false
Trim R1              : 0
Trim R2              : 0
Trim 3' R1           : 0
Trim 3' R2           : 0
Config Profile       : UPPMAX
UPPMAX Project       : b2017001
E-mail Address       : phil.ewels@scilifelab.se
=====================================
```

Version control

Releases | Tags

Pre-release

v1.3

9d8b6b5

NGI-

ewels

Version v

```
$fastqc.
$trim_ga
$bw.modu
$samtool
$picard.
$phantom
$deepToo
$ngsplot
$macs.mo
$saturat
$saturat
```

PUBLIC | AUTOMATED BUILD

scilifelab/ngi-chipseq

Last pushed: a day ago

Repo Info | Tags | Dockerfile | Build Details | Build Settings | Collaborators | Webhooks

Status	Actions	Tag	Created	Last Updated
Building	Cance	v1.3	2 minutes ago	a minute ago
Canceled		v1.4	a day ago	a day ago
Success		latest	a day ago	a day ago

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-ChIPseq -r v1.3
```

SciLifeLab

NGI stockholm

Conclusion

- Use NGI-ChIPseq to prepare your data if you want:
  - To not have to remember every parameter for every tool
  - Extreme reproducibility
  - Ability to run on virtually any environment
- Now running for all ChIPseq projects at NGI-Stockholm

NGI-ChIPseq

Conclusion

GitHub https://github.com/

Open Source Initiative

NGI-ChIPseq SciLifeLab/NGI-ChIPseq

NGI-RNAseq SciLifeLab/NGI-RNAseq

NGI-smRNAseq SciLifeLab/NGI-smRNAseq

NGI-MethylSeq SciLifeLab/NGI-MethylSeq

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

Chuan Wang

Jakub Westholm

Rickard Hammarén

Max Käller

Denis Moreno

NGI Stockholm Genomics Applications Development Group

support@ngisweden.se

http://opensource.scilifelab.se

SciLifeLab

NGI stockholm