NGI-ChIPseq

Processing ChIP-seq data at the National Genomics Infrastructure

SciLifeLab

NGI stockholm

Phil Ewels
phil.ewels@scilifelab.se
NBIS ChIP-seq tutorial
2018-11-08

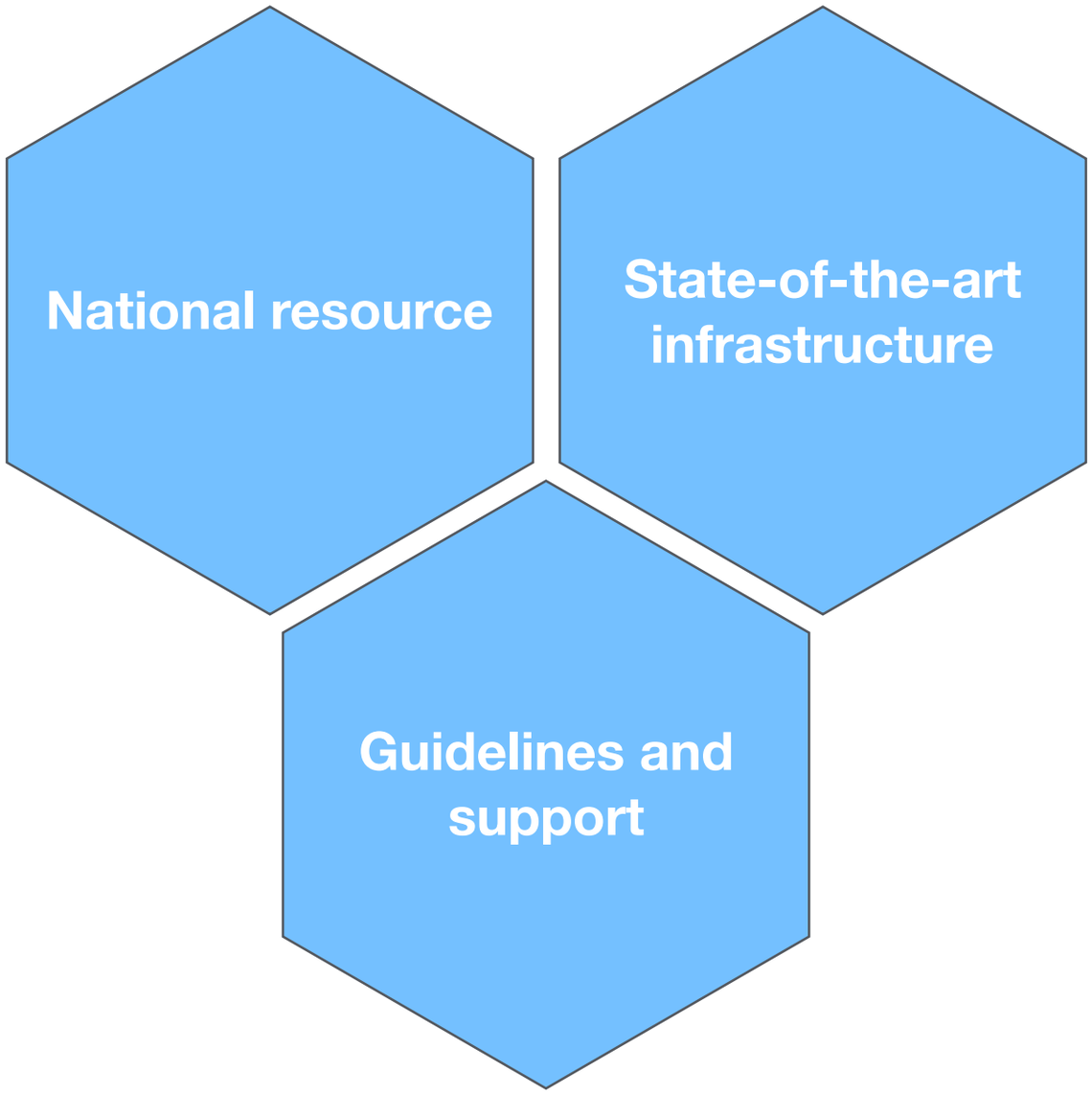
SciLifeLab NGI

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

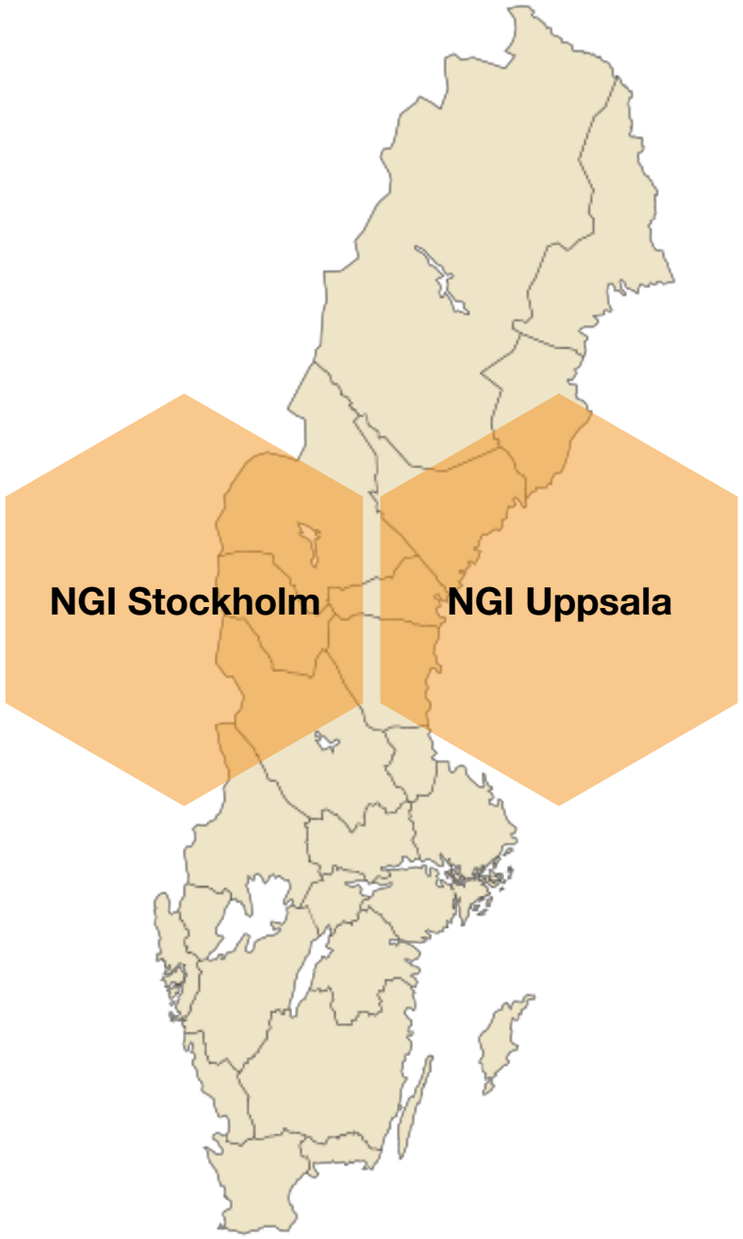


GANATIONALCTAC
ATCAGENOMICSGT
INFRASTRUCTURE

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

SciLifeLab
NGI stockholm

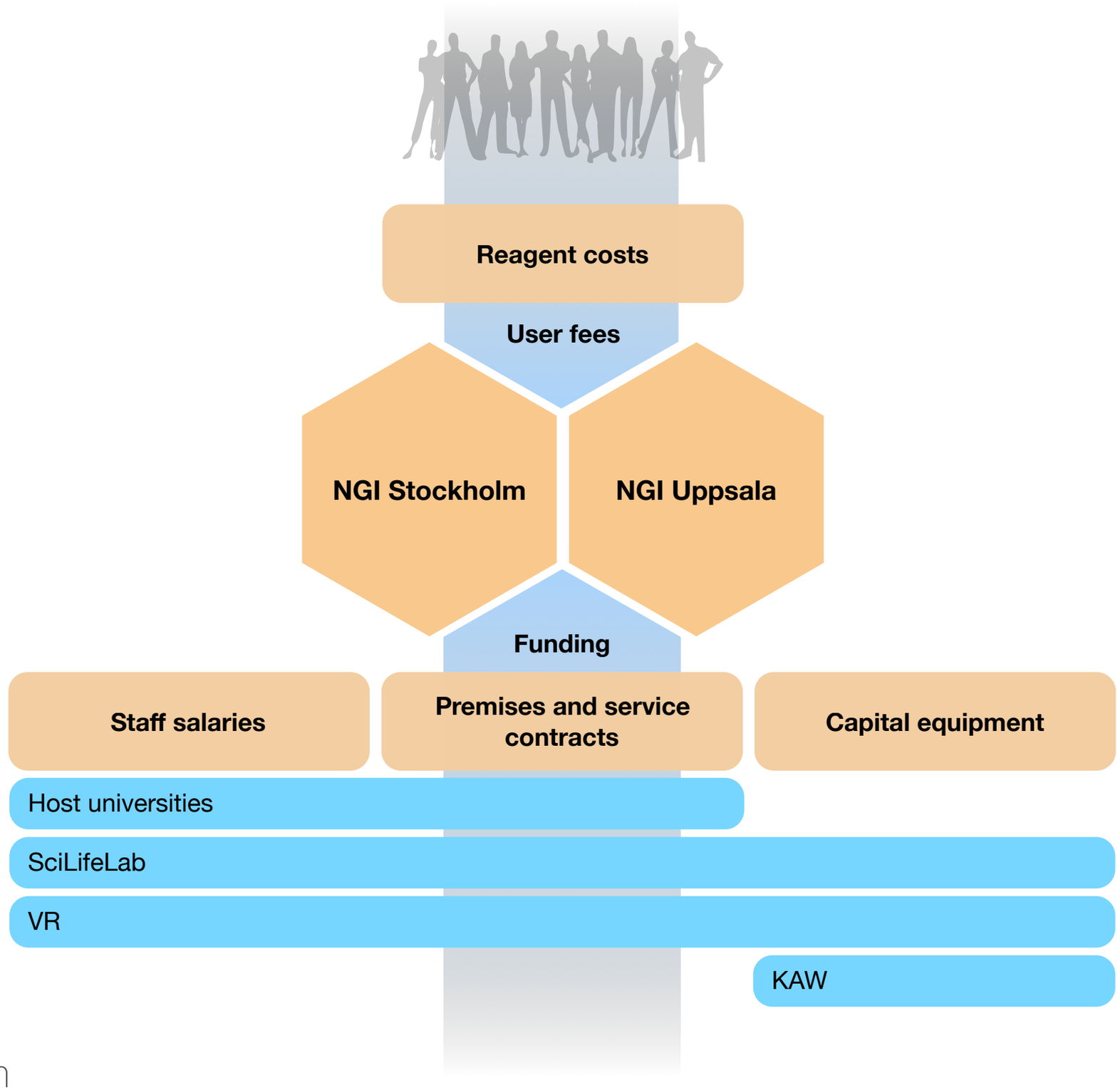
NGI Organisation



SciLifeLab

NGI stockholm

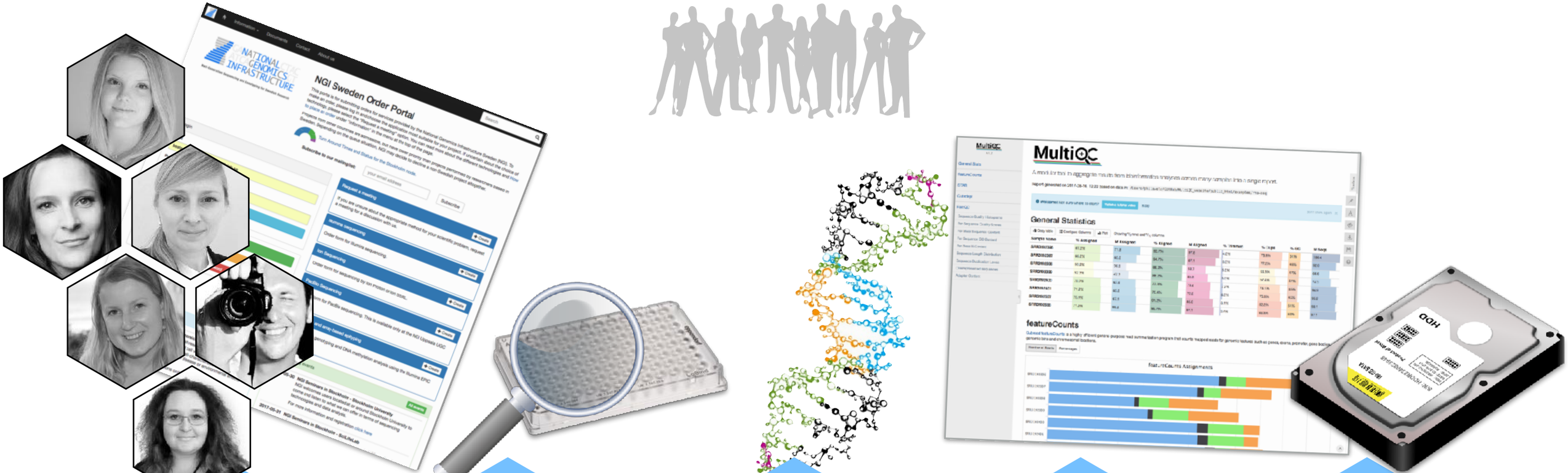
NGI Organisation



SciLifeLab

NGI stockholm

Project timeline



Scientific support and project consultation

Sample QC

Library preparation, Sequencing, Genotyping

Data processing and primary analysis

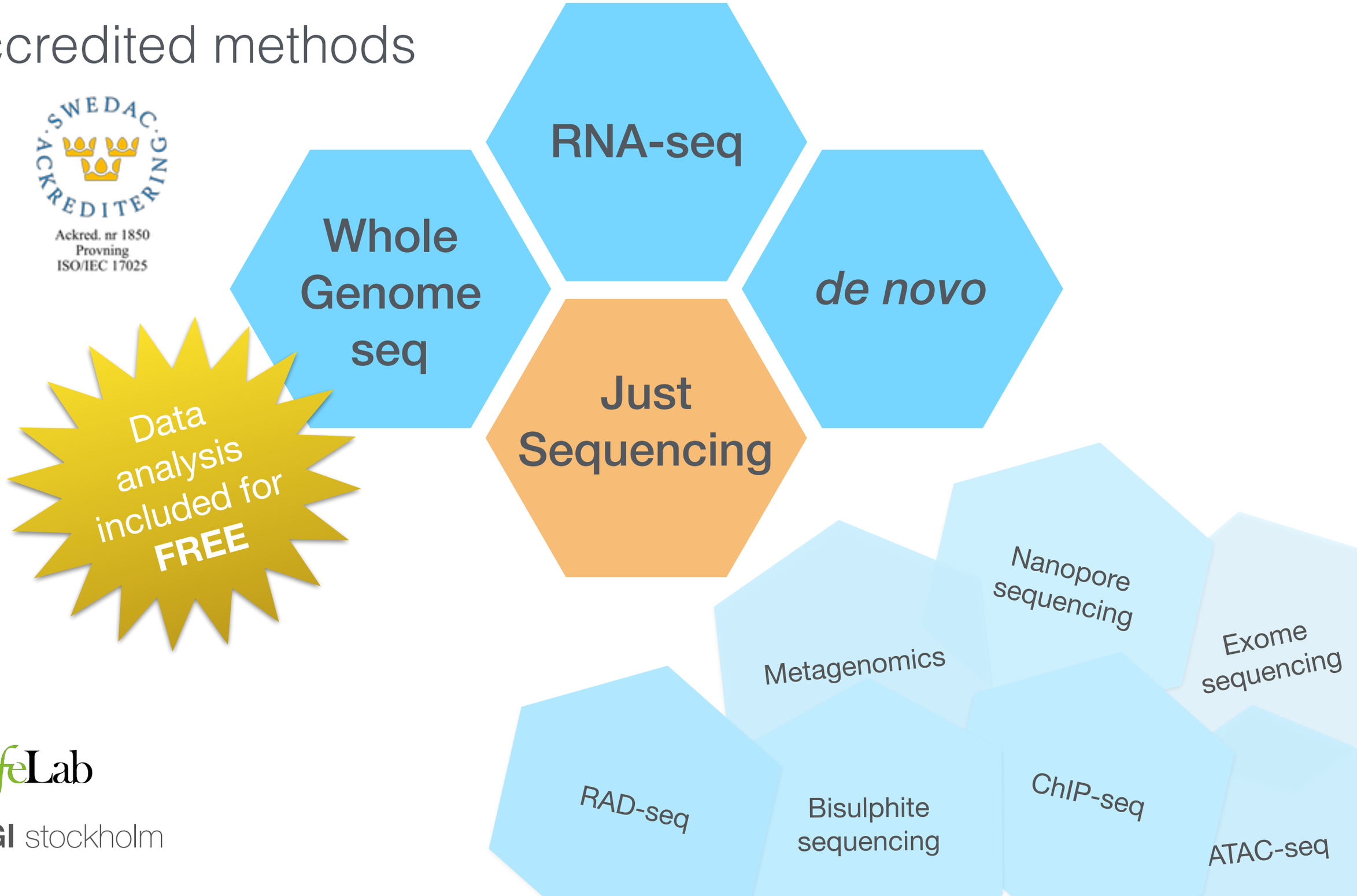
Data delivery

SciLifeLab

NGI stockholm

Methods offered at NGI

Accredited methods



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
  - Approx 1000 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
  - ~1300 kr / sample (40M reads)

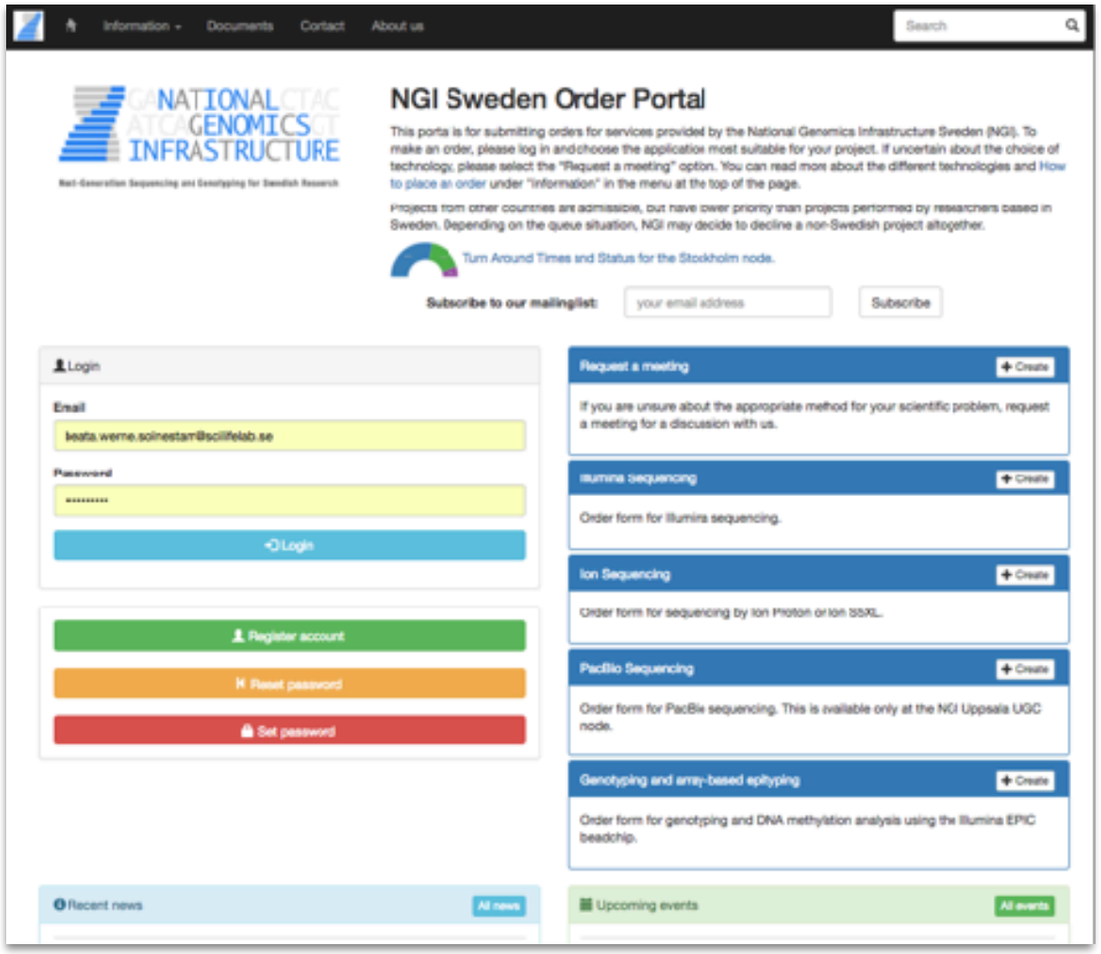
SciLifeLab

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

- 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

nf-core/chipseq

ChIP-seq Pipeline

nf-core/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

nf-core

CRG Centre for Genomic Regulation

GANATIONALCTAC ATCAGENOMICSGT INFRASTRUCTURE

CANCER RESEARCH UK | BEATSON INSTITUTE

QBiC

SciLifeLab

A*STAR Genome Institute of Singapore

SLU

lifebit

THE FRANCIS CRICK INSTITUTE

wellcome sanger institute

International Agency for Research on Cancer

World Health Organization

https://nf-co.re/

Home Pipelines Usage Developers Tools About

nf-core

A community effort to collect a curated set of analysis pipelines built using Nextflow.

VIEW PIPELINES

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.

Nextflow is an incredibly powerful and flexible workflow language.

nf-core pipelines adhere to strict guidelines - if one works, they all will.

Nextflow is an incredibly powerful and flexible workflow language.

nf-core pipelines adhere to strict guidelines - if one works, they all will.

Documentation

Extensive documentation covering installation, usage and description of output files ensures that you won't be left in the dark.

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.

Travis CI

Stable Releases

nf-core pipelines use GitHub releases to tag stable versions of the code and software, making pipeline runs totally reproducable.

Docker

Software dependencies are always available in a bundled docker container, which Nextflow can automatically download from dockerhub.

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.

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.

Get started in minutes

Nextflow lets you run nf-core pipelines on virtually any computing environment.

nf-core pipelines come with built-in support for AWS iGenomes with common species.

The nf-core companion tool makes it easy to list all available nf-core pipelines and shows which are available locally. Local versions are checked against the latest available release.

```
# Install nextflow
curl -s https://get.nextflow.io | bash
mv nextflow ~/bin

# Launch the RNAseq pipeline
nextflow run nf-core/RNAseq \
    -profile standard,docker \
    --genome GRCh37 \
    --reads "data/*_{R1,R2}.fastq.gz"

# Install nf-core tools
pip install nf-core

# List all nf-core pipelines and show available updates
nf-core list
```

Pipelines

Browse the **16** pipelines that are currently available as part of nf-core.

Available Pipelines

Can you think of another pipeline that would fit in well? Let us know!

Search keywords | Filter: Released 4 | Under development 12 | Sort: Last Release | Alphabetical | Status | Stars

nf-core/eager

8

adna ancientdna pathogen-genomics population-genetics

A fully reproducible and state of the art ancient DNA analysis pipeline.

Version **2.0.2** Published 3 days ago

nf-core/rnaseq

48

rna rna-seq

RNA sequencing analysis pipeline using STAR or HISAT2, with gene counts and quality control

Version **1.1** Published 1 month ago

nf-core/hlatyping

dna hla hla-typing immunology optitype personalized-medicine rna

Precision HLA typing from next-generation sequencing data

Version **1.1.1** Published 3 months ago

nf-core/methylseq

15

bisulfite-sequencing dna-methylation methyl-seq

Methylation (Bisulfite-Sequencing) analysis pipeline using Bismark or bwa-meth + MethylDackel

Version **1.1** Published 3 months ago

nf-core/rnafusion

nf-core/rrna-ampliseq

9

nf-core/chipseq

Search or jump to... / Pull requests Issues Marketplace Explore

nf-core / chipseq
forked from SciLifeLab/NGI-ChIPseq

Unwatch 23 | Star 14 | Fork 34

Code | Issues 16 | Pull requests 2 | Insights | Settings

Chromatin immunoprecipitation (ChIP-seq) analysis using BWA and MACS2 with QC steps. http://nf-co.re

Edit

nf-core nextflow workflow chip-seq chromatin-immunoprecipitation peak-calling Manage topics

342 commits | 2 branches | 1 release | 7 contributors | MIT

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

This branch is 107 commits ahead, 3 commits behind SciLifeLab:master. #55 Compare

ewels Merge pull request #42 from Rotholandus/master ... Latest commit 5f67d82 on 10 Aug

assets	nf-core/chipseq, not ChIPseq	5 months ago
bin	Merge pull request #16 from nf-core/bioconda	5 months ago

SciLifeLab

NGI stockholm

https://github.com/nf-core/chipseq

nf-core/chipseq

README.md

nf-core/chipseq Results

The nf-core/chipseq documentation is split into a few different files:

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

SciLifeLab

NGI stockholm

https://github.com/nf-core/chipseq

Running nextflow

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 nf-core/chipseq pipeline

`nextflow run nf-core/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

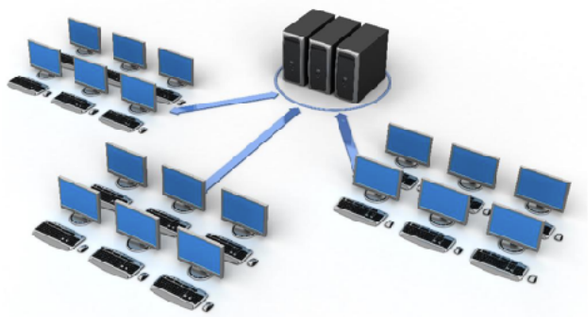
- Read the Nextflow log and check the MultiQC report

Step 7: Delete temporary files

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

Using UPPMAX

```
nextflow run nf-core/chipseq
    -profile uppmax
    --project b2017123
    --genome GRCh37 --macsconfig p.txt
    --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 nf-core/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 nf-core/chipseq
    -profile standard,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 nf-core/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

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
```

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

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"

}
```

SciLifeLab

NGI stockholm

nf-core/chipseq config

```
N E X T F L O W  ~  version 0.30.1
Launching `/home/travis/build/nf-core/chipseq/main.nf` [determined_ekeblad] - revision:
b11db350eb
=========================================================
                                          ,--./,-.
          ___     __   __   __   ___     /,-._.--~'
    |\ | |__  __ /  ` /  \ |__) |__         }  {
    | \| |       \__, \__/ |  \ |___     \`-._,-`-,
                                          `._,._,'
 nf-core/chipseq : ChIP-Seq Best Practice v1.0dev
=========================================================
Run Name                  : determined_ekeblad
Reads                     : data/*{1,2}*.fastq.gz
Data Type                 : Paired-End
Genome                    : false
Fasta Ref                 : https://github.com/nf-core/test-datasets/raw/chipseq/reference/genome.fa
MACS Config               : https://github.com/nf-core/test-datasets/raw/chipseq/macsconfig.txt
Saturation analysis       : false
MACS broad peaks          : false
Blacklist filtering       : false
Extend Reads              : 100 bp
Container                 : nfcore/chipseq:latest
Output dir                : ./results
Script dir                : /home/travis/build/nf-core/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            : test,docker
Email                     : phil.ewels@scilifelab.se
```

Version control

Releases | Tags

Pre-release

v1.3

9d8b6b5

NGI-

ewels

Version

Down

Sourc

Sourc

PUBLIC | AUTOMATED BUILD

scilifelab/ngi-chipseq

Last pushed: a day ago

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

Status	Actions	Tag	Created	Last Updated
Building	Cancel	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 nf-core/chipseq -r 1.0
```

SciLifeLab

NGI stockholm

Conclusion

- Use nf-core/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

nf-core/
chipseq

Conclusion

Phil Ewels

phil.ewels@scilifelab.se

ewels

tallphil

Acknowledgements

Chuan Wang

Jakub Westholm

Rickard Hammarén

Max Käller

Denis Moreno

NGI Stockholm Genomics
Applications Development Group

https://nf-co.re

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

SciLifeLab

NGI stockholm