Computational requirements for the H3ABioNet GWAS workflows

Scott Hazelhurst

http://www.bioinf.wits.ac.za/gwas/gwas-comp-handout.pdf

H3ABioNet

Pan African Bioinformatics Network for H3Africa

SDIMB

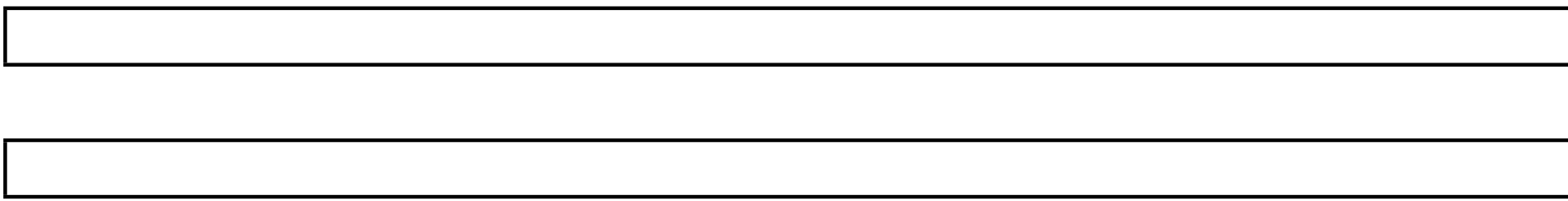
SYDNEY BRENNER INSTITUTE FOR MOLECULAR BIOSCIENCE

School of Electrical and Information Engineering

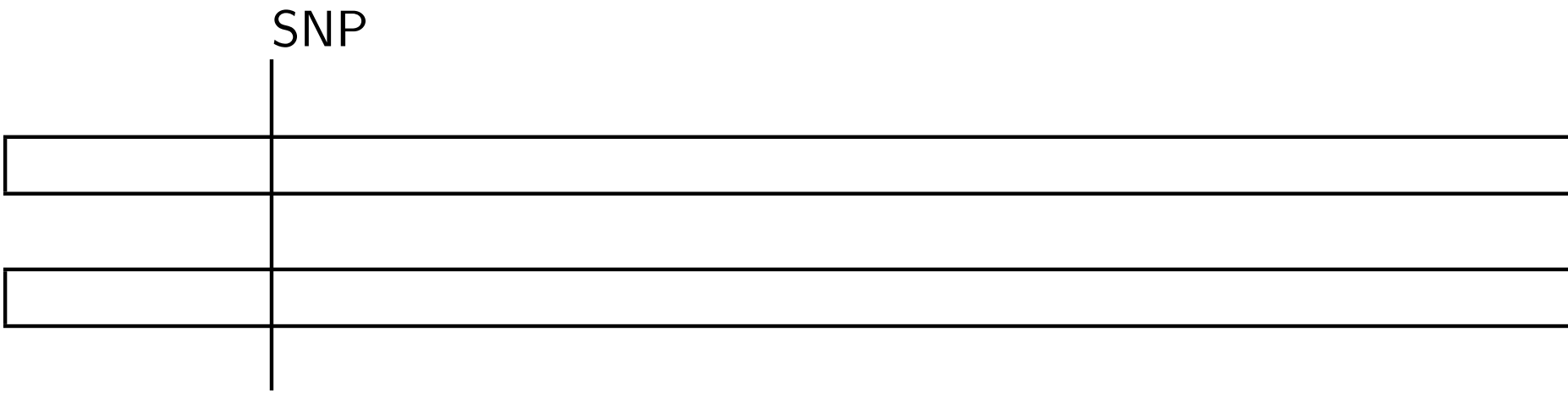
AWI-Gen

Africa Wits-INDEPTH Partnership for Genomic Reseach

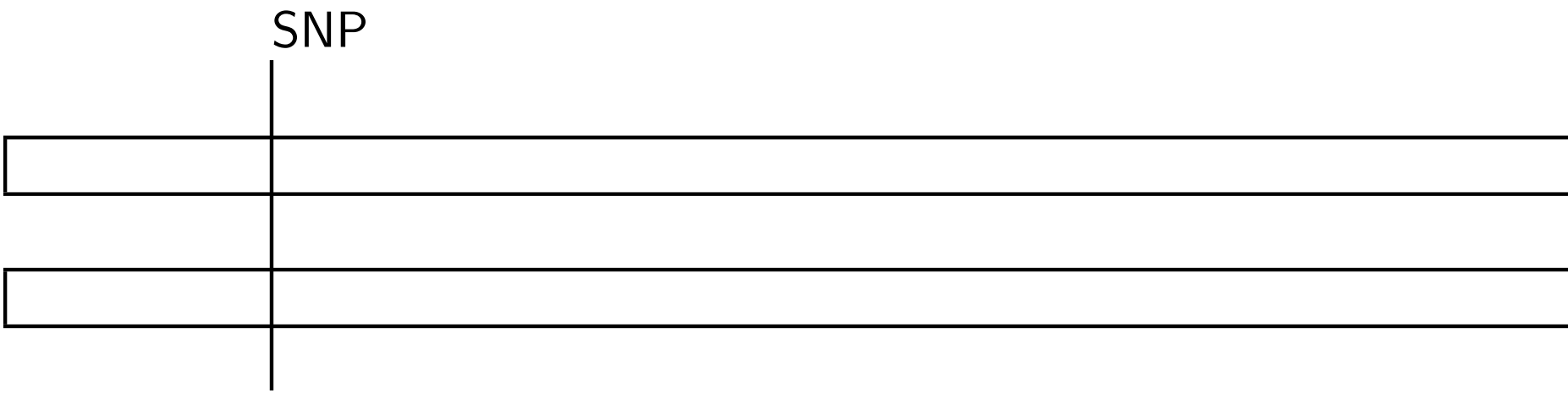
Recap of GWAS



Recap of GWAS

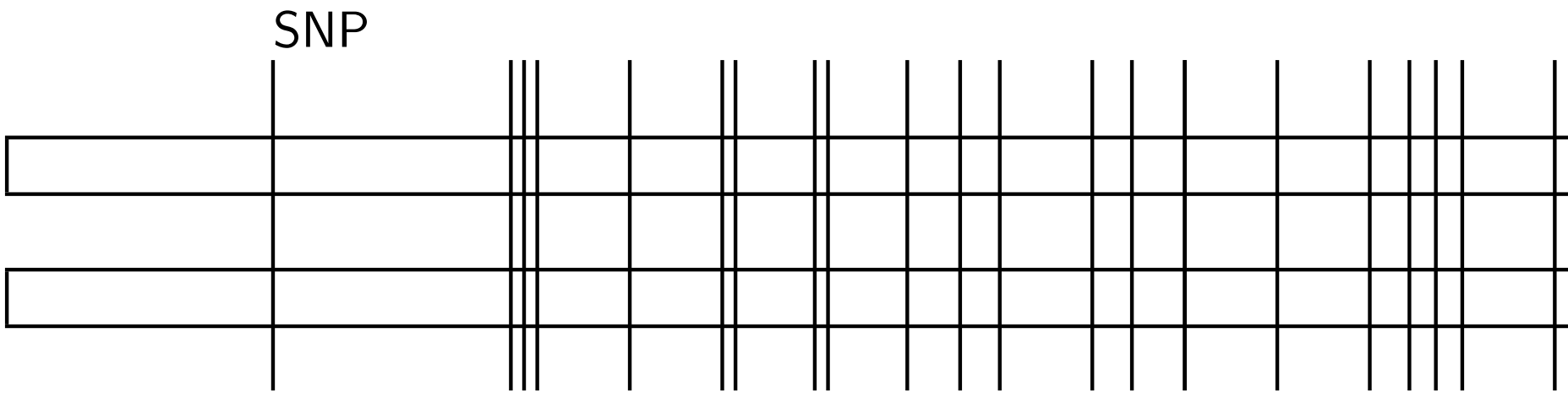


Recap of GWAS



A/A, A/G, G/G

Recap of GWAS



Alleles

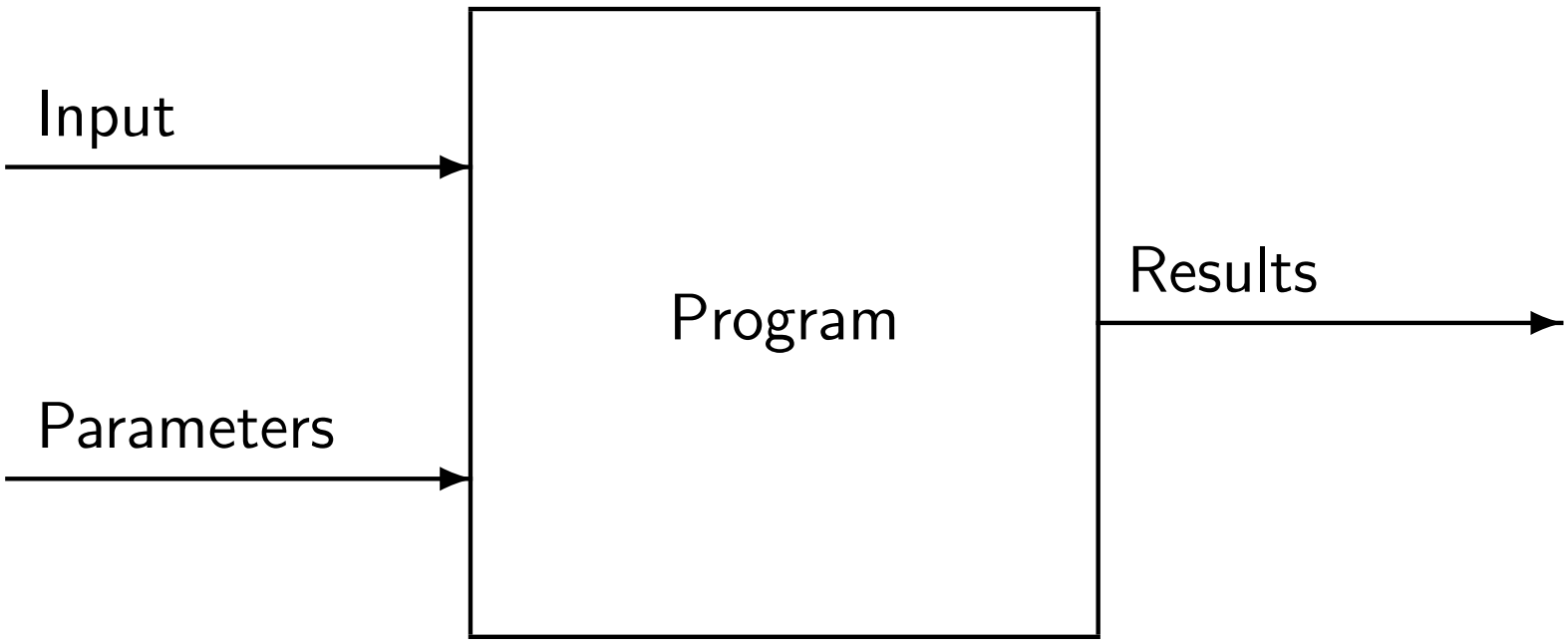
A/A, A/G, G/G

Introduction

Genome Wide-Association Studies important application area – complex computing requirements

- hardware requirements – could probably manage with quad-core, 8GB RAM
  - e.g., 10k samples, 2m chip, individual steps would take 1-2 days
  - modest-size cluster *very* helpful
- Software requirements complex and heterogeneous

Ideal world



Need to build pipeline for GWAS

GWAS

- complex – several programs
- multiple steps, software dependencies
- have multiple parameters

Constraints of good scientific practice – needs to be

- re-rerun often to understand data
- reproducible by others
- portable

Workflow/Pipeline

Packaging of steps in a complex analysis

- user runs the package
- automate the steps
- not black-box – user needs to understand

Use appropriate software technology to support this

- Nextflow
- Containerisation (Docker/Singularity)

Nextflow – workflow language and system

Nextflow developed at the Comparative Genomics Group at the Centre for Genomic Regulation in Barcelona

Paolo Di Tommaso, Nextflow Lead

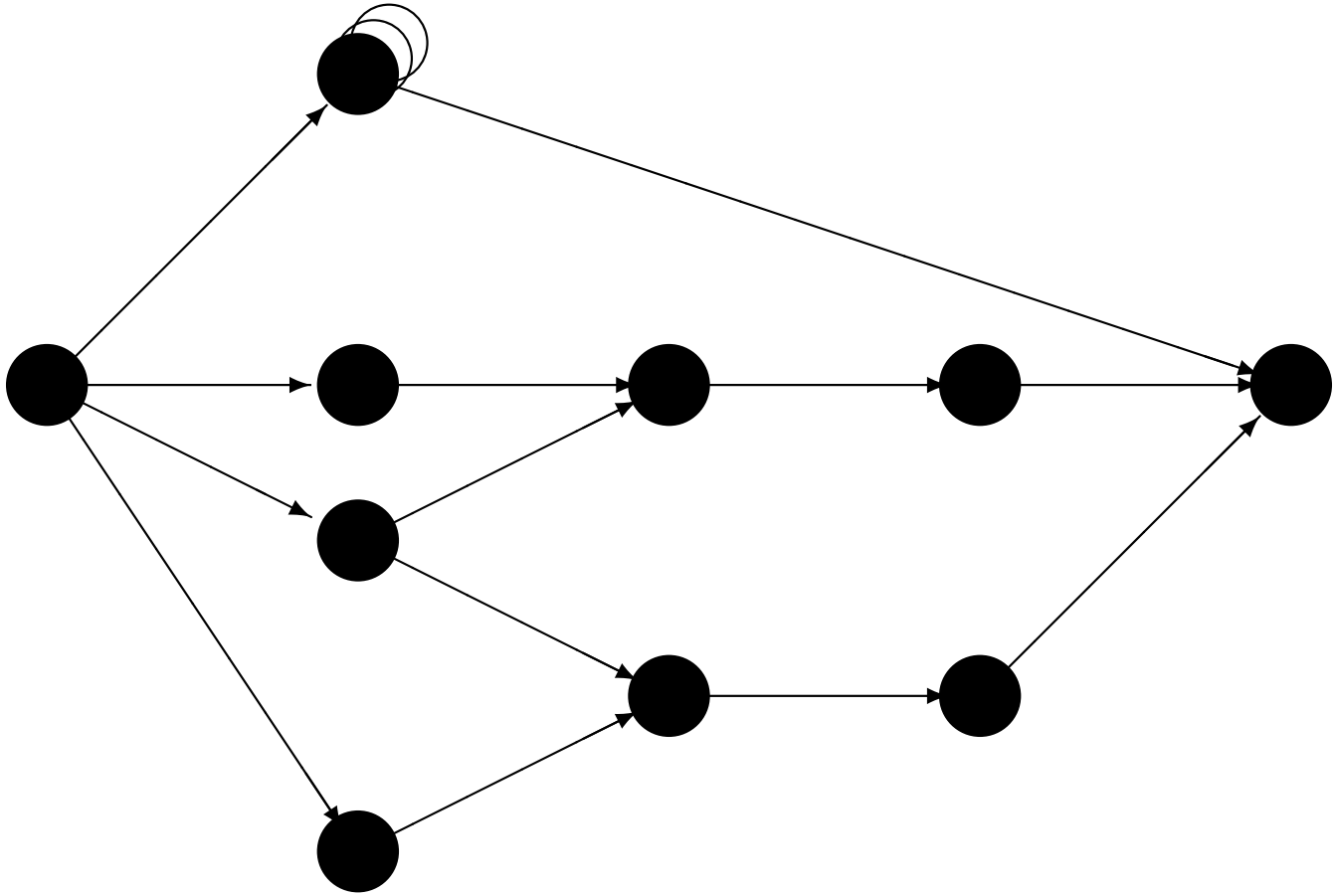
Evan Floden, Bioinformatician

Emilio Palumbo, Bioinformatics Engineer

Cedric Notredame, Principal Investigator

Nextflow is a language/system to coordinate individual steps of workflow

- Special purpose language with high level support for coordination of work
- Individual steps : written in more general purpose language, or call tools to be used



Features

- Detects dependencies, parallelism in workflow
- Schedules tasks when ready — maps to computational resources available
- Supports partial resumption
- Execute locally, on head node of cluster, cloud computing
- Supports Docker and Singularity

Installing Nextflow

Requires

- Java 8
- Nextflow

Detailed instructions at

- `http://www.bioinf.wits.ac.za/gwas/gwas-comp-handout.pdf` and
- video at `http://www.bioinf.wits.ac.za/gwas/videos/10-nextflow.mp4`

Software dependencies

GWAS requires many different pieces of software

- Install them yourself
  Requires work but gives you flexibility
- Use Docker or Singularity containers
  Packages all dependencies for you
  Requires Docker or Singularity to be on your system.

Detailed instructions and videos at
`http://www.bioinf.wits.ac.za/gwas/gwas-comp-handout.pdf`

Containerisation

Light-weight support for virtual machines

- a software *container* image is a package of an operating system, libraries, tools needed for an application

Containerisation

Light-weight support for virtual machines

- a software *container* image is a package of an operating system, libraries, tools needed for an application
- can run *containers* from the image
  - each container has its own isolated set of resources
  - own file system
  - can run different OS to the host operating system

Containerisation

Light-weight support for virtual machines

- a software *container* image is a package of an operating system, libraries, tools needed for an application
- can run *containers* from the image
  - each container has its own isolated set of resources
  - own file system
  - can run different OS to the host operating system
- We provide Docker images and Singularity support for our workflows
- You don't need to be an expert in Docker and/or Singularity – Nextflow manages it for you

Docker versus Singularity

Docker

- better known, better support
- not intended for multi-user computers – security issue – so probably won't find on your local university cluster
- Linux, recent MacOS, Windows 10 with MS Hyper-V

Singularity:

- Better security – can run on shared computer
- Linux or
- MacOS or Windows with extra requirements

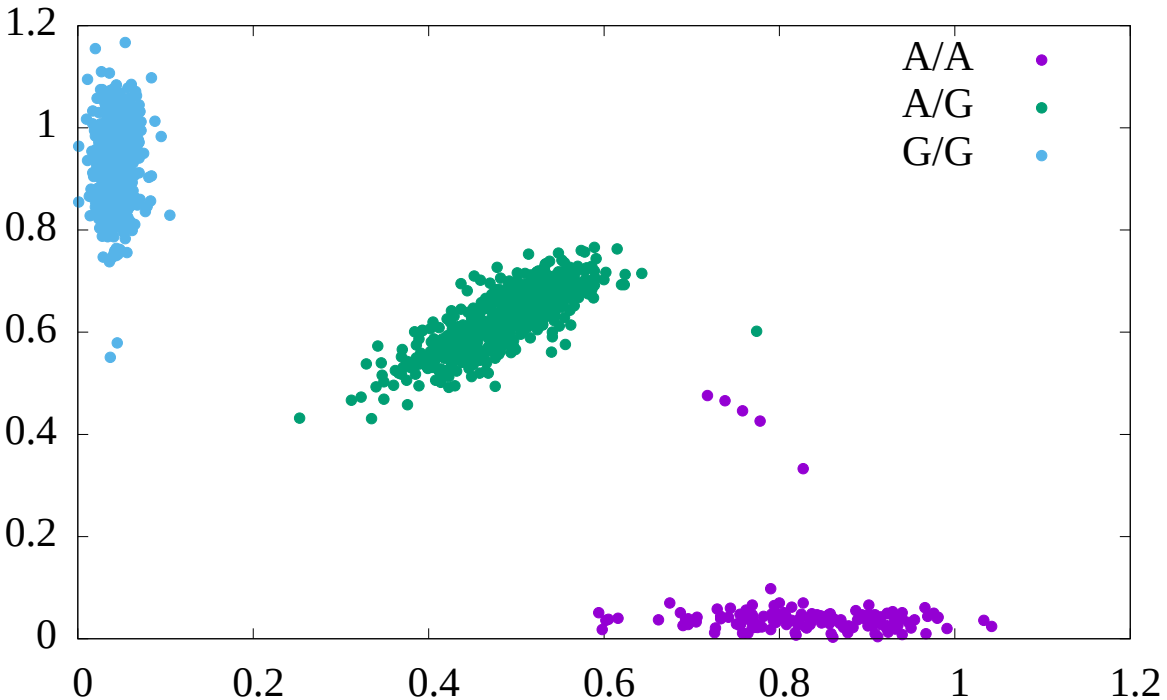
H3A GWAS pipeline

https://github.com/h3abionet/h3agwas

Possible data types

1. Image file (we don't support – yet)
2. Called data
3. Raw PLINK data
4. QC PLINK data
5. GWAS results

Image data



Called data

Programs like GenomeStudio analyse the image files and *call* the data.

- For each SNP: which cluster does each individual fall in

Different strand alignment formats possible

- TOP/BOTTOM
- Forward/Reverse

Very verbose format

PLINK format

Binary PLINK format data is found in three files

- FAM file: who the people are
- BIM file: describes each SNP (e.g., chromosome, position, possible alleles)
- BED file: for each person, for each SNP what the SNP is

We have three workflow

- `topbottom.nf`
  Conversion from Illumina Top/Bottom or Forward/Reverse format
- `plink-qc.nf`
  Quality control
- `plink-gwas.nf`
  Basic association study

Installing the workflow

1. Use Nextflow to manage – will download all the code for you
2. Use git to manage
   More advanced, needed if you want to modify the workflow

Using Nextflow to manage

```
nextflow pull h3abionet/h3agwas
```

Running Nextflow

If you use Nextflow to manage the workflows you'll run Nextflow as follows

- `nextflow run h3abionet/h3agwas/topbottom.nf ....`
- `nextflow run h3abionet/h3agwas/plink-qc.nf .....`
- `nextflow run h3abionet/h3agwas/plink-gwas.nf ...`

Updating the workflow

Each time we update Nextflow, the next time you run the script you'll get a message like

> *NOTE: Your local project version looks outdated - a different revision is available in the remote repository*

Update by saying: `nextflow pull h3abionet/h3agwas`

Configuration files

Nextflow runs are controlled by configuration files -

- Can have several
- Recommended – use two
  - default nextflow.config file
  - a smaller config file that redefines just those things that you need

Default config file is `https://github.com/h3abionet/h3agwas/blob/master/nextflow.config`

```
nextflow run h3abionet/h3agwas/plink-qc.nf -c b.config
```

Will use :

- default nextflow.config file plus
- the specified config file

Overview of QC steps

- Remove duplicate SNPs
- Remove SNPs, individuals with high missingness, HWE, MAF
- Remove outliers on sample heterozygosity
- Remove relatedness
- Tests differential missingness
- Produce reports

Config file : Main components

- Input directory and file
- Output directory and file
- Batch analysis: strongly recommended
  Case-control: binary – compulsory
  By phenotype: e.g., *site*, strongly recommended
- QC cut-offs

Need phenotype file(s) with headers.

Running a QC

Controlled by a config file

```
params.input_dir  = "sample"
params.input_pat  = "sampleA"
params.output     = "test-qc"
params.output_dir = "output"
params.case_control = "sample/sample.phe"
params.case_control_col = "PHE"
params.batch = "sample/sample-batch-site.phe"
params.batch_col = "batch"
params.phenotype = "sample/sample-batch-site.phe"
params.pheno_col = "site"
params.sexinfo_available = true
params.pi_hat = 0.18
params.cut_maf= 0.05
```

```
nextflow run h3abionet/h3agwas -c sc.config plink-qc.nf
```

```
N E X T F L O W  ~  version 0.30.1
Launching `../plink-qc.nf` [pedantic_kare] - revision: ac5217ecef
Sexinfo available command
[warm up] executor > local
[fe/c0582e] Submitted process > inMD5 (1)
[d4/1e3bbb] Submitted process > getDuplicateMarkers (1)
[f4/ed17d8] Submitted process > removeDuplicateSNPs (1)
[ad/6d480c] Submitted process > getInitMAF (1)
[8b/40df13] Submitted process > getX (1)
[57/ef097e] Submitted process > identifyIndivDiscSexinfo (1)
[91/59ea3c] Submitted process > generateIndivMissingnessPlot (1)
[ab/e1643d] Submitted process > generateSnpMissingnessPlot (1)
[d8/9d32e8] Submitted process > removeQCPhase1 (1)
..
..
[e5/2817fb] Submitted process > generateMafPlot (1)
[d4/e6d01c] Submitted process > produceReports (1)
The output report is called output/test-qc.pdf
```

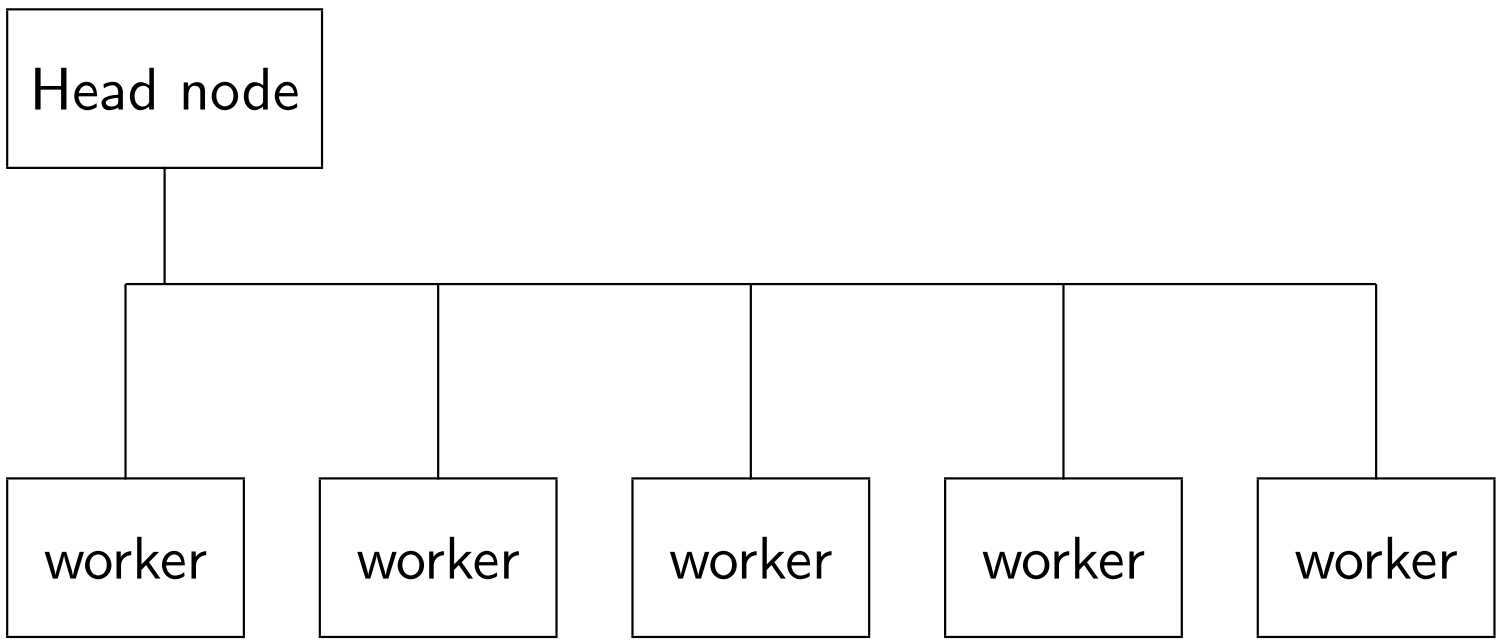
Running with Docker

If tools have not all been installed on the system, run with Docker

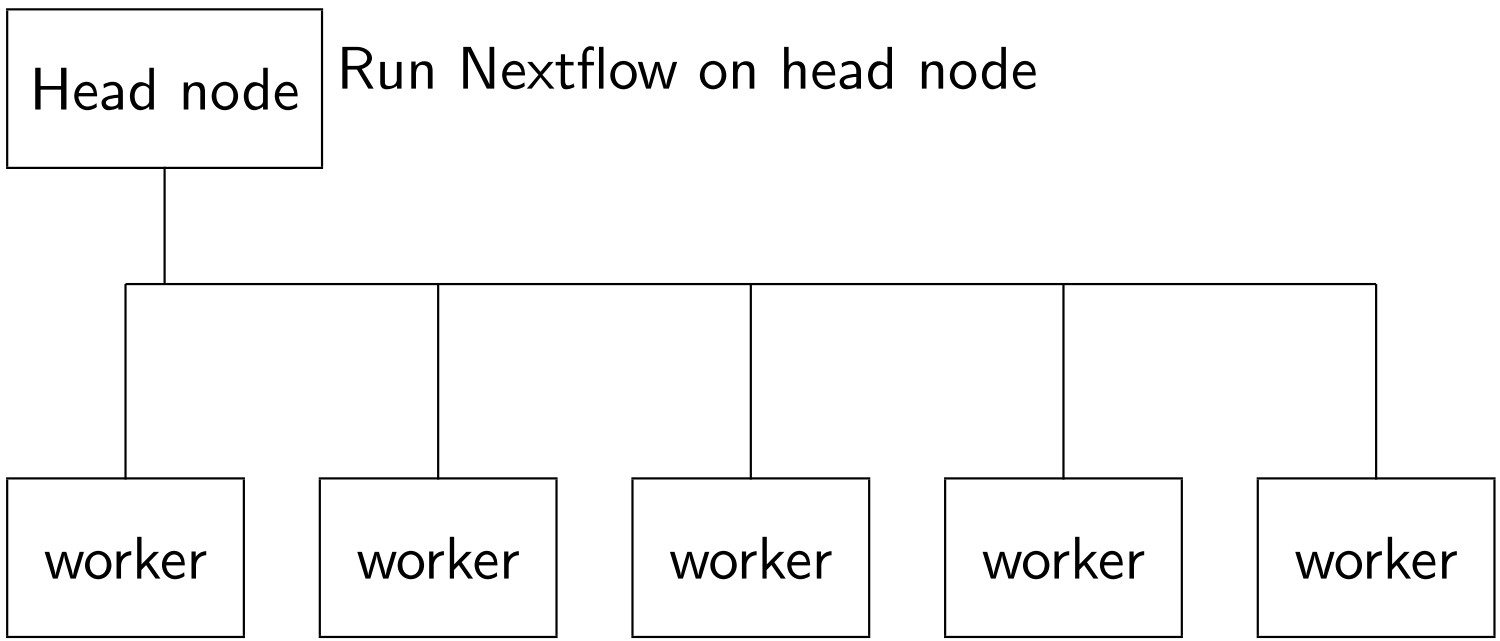
```
nextflow run h3abionet/h3agwas -c sc.config plink-qc.nf\
             -profile docker
```

The first time you do this, there may be very long delays as the Docker images are fetched from their repositories.

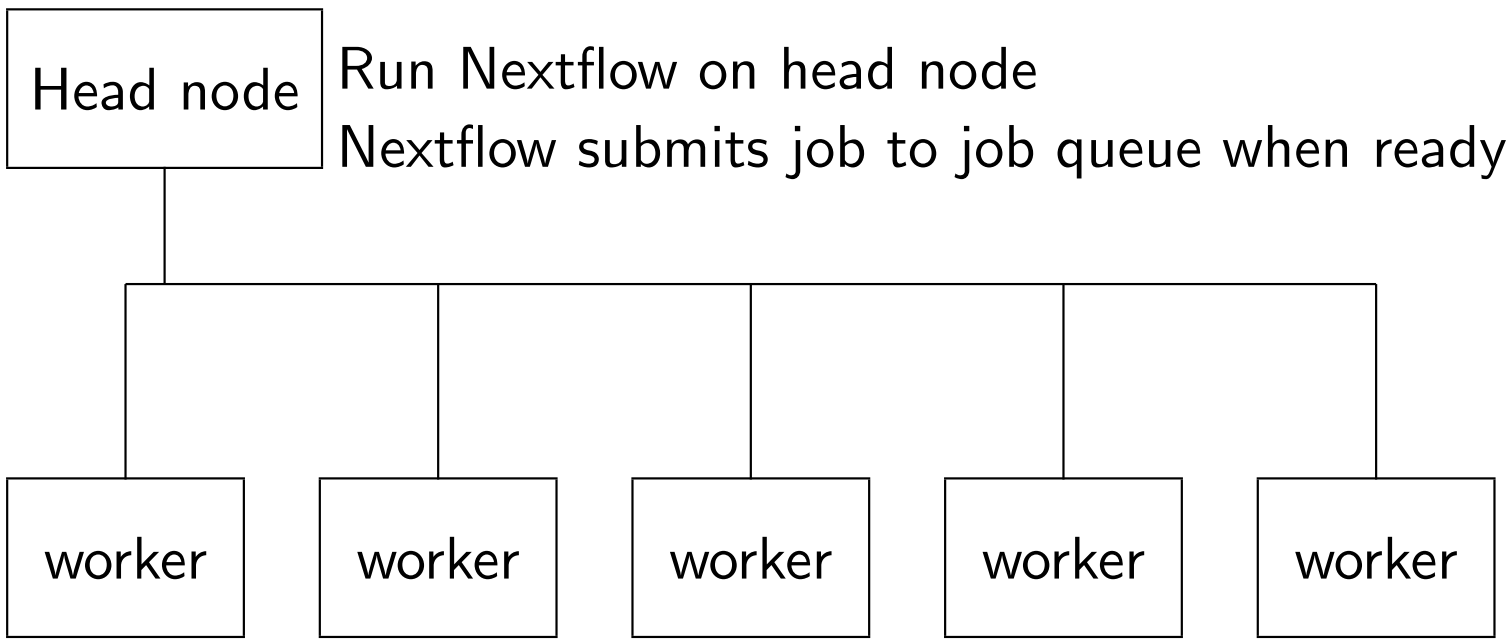
Running on a cluster



Running on a cluster



Running on a cluster



Running with PBS

Run on the head node of the cluster

```
nextflow run h3abionet/h3agwas -c sc.config plink-qc.nf\
             -profile pbs
```

```
nextflow run h3abionet/h3agwas -c sc.config plink-qc.nf\
             -profile pbsDocker
```

Basic association study

plink-assoc.nf

Association workflow very experiment dependant

- data, population structure, co-variate, question
- basic workflow implemented for initial study
- can be extended

Example config file

```
params {
    input_dir = "/data/scott/assoc/agt"
    input_pat = "t25"
    output  =  "allgemma"
    output_dir = "assocresults"
    data = "/data/scott/assoc/data.csv"

    covariates = "age,sex"
    pheno="bmi_c/np.log,wst_hip_r_c,standing_height_mm"
    gemma_num_cores = 8
    gemma             = 1
    linear             = 1
}
```

Running the workflow

```
nextflow run h3abionet/h3agwas/plink-assoc.nf \
         -c assoc.config
```

The output will be found in the *assocresults* configuration file

- because that's specified in the *assoc.config* file

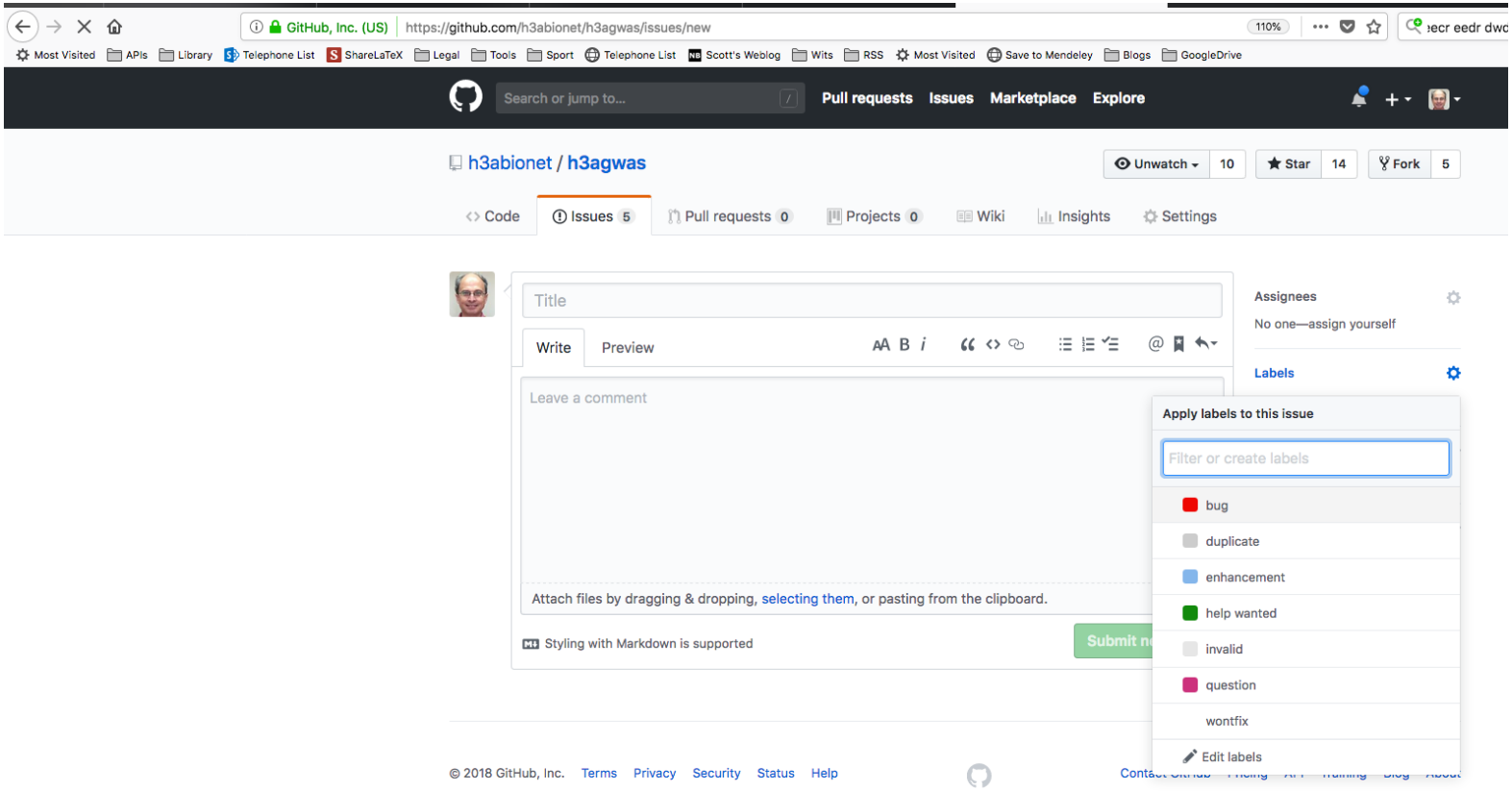
Problems

If something goes wrong, may be difficult to understand why

- workflow
- data

Asking for help

- H3ABioNet Help desk
  `https://www.h3abionet.org/support`
- On GitHub – need a GitHub account
  `https://github.com/h3abionet/h3agwas/issues`


GitHub, Inc. (US) https://github.com/h3abionet/h3agwas/issues/new
Search or jump to...
Pull requests
Issues
Marketplace
Explore
h3abionet / h3agwas
Unwatch 10
Star 14
Fork 5
Code
Issues 5
Pull requests 0
Projects 0
Wiki
Insights
Settings
Title
Write
Preview
Leave a comment
Attach files by dragging & dropping, selecting them, or pasting from the clipboard.
Styling with Markdown is supported
Assignees
No one—assign yourself
Labels
Apply labels to this issue
Filter or create labels
bug
duplicate
enhancement
help wanted
invalid
question
wontfix
Edit labels
© 2018 GitHub, Inc. Terms Privacy Security Status Help

Funded by NIH NHGRI grants U41HG006941, HG006938. Work at different institutions and individuals.

Eugene de Beste
Lerato Magosi
Phelelani Mpangase
Rob Clucas
Jean-Tristan
Brandenberg Harry
Noyce

Ayton Meintjes
Don Armstrong
Fourie Joubert
Gerrit Botha
Sumir Panji
Nicky