H3AGWAS : A portable workflow for Genome Wide Association Studies

Jean-Tristan Brandenburg[1], Lindsay Clark[2], Gerrit Botha[3], Sumir Panji[3], Shakuntala Baichoo[4], Christopher Fields[2] and Scott Hazelhurst[1,5]

May 2, 2022

Abstract

Background Genome-wide association studies (GWAS) are a powerful method to detect associations between variants and phenotypes. A GWAS requires several complex computations with large datasets, and many steps may need to be repeated with varying parameter. Manual running of these analyses can be tedious, error-prone and hard to reproduce. **Results** The H3AGWAS workflow from the Pan-African Bioinformatics Network for H3Africa is a powerful, scalable and portable workflow implementing pre-association analysis, implementation of various association testing methods and post-association analysis of results. **Conclusions** The workflow is scalable — laptop to cluster to cloud (e.g., SLURM, AWS Batch, Azure). All required software is containerised and can run under Docker on Singularity.

Background

Genome-wide association studies (GWAS) are a powerful method to detect associations between variants and phenotypes; from initial raw genotype data until detection of putative causal variant requires numerous steps, software and approaches to extract and understand results [1]. Common steps after genotyping include:

1. Preparing data into standard formats
2. Quality control (QC) of genotypes to remove uncertain positions and individuals – e.g., discrepancy between genotyped sex and known sex, and bias due to high relatedness between individuals. These are important steps to reduce noise and false positive discovery rate [2–4].

[1]Sydney Brenner Institute for Molecular Bioscience, University of the Witwatersrand, Johannesburg, South Africa

[2]HPCBio, Roy J. Carver Biotechnology Center, University of Illinois at Urbana-Champaign, Urbana, Illinois, United States

[3]Computational Biology Division, Institute of Infectious Disease and Molecular Medicine, University of Cape Town, cnySouth Africa

[4]Department of Digital Technologies, Faculty of Information, Communication and Digital Technologies, University of Mauritius, Mauritius

[5]School of Electrical & Information Engineering, University of the Witwatersrand, Johannesburg, South Africa

3. Associating genetic variation with phenotype. This step is very expensive, with millions of positions and sample sizes ranging from several thousands to several hundred thousand. These methods take account of relatedness between individuals with mixed models and different algorithms to improve detection and/or approximations for a very large sample size.

4. post-association analysis, which may include highly complex methods such as meta analysis considering different GWAS summary statistics, fine-mapping to define causal variants, heritability of phenotype, replication and transferabilty of previous results, annotation, integration of eQTL, and/or calculation of polygenic risk score.

Motivation

The phases of GWAS are all complex, and typically require multiple executions, sometimes on different platforms by different collaborators and replicability of analyses is crucial. The Pan-African Bioinformatics Network of the Human Heredity and Health in Africa Consortium [5] (H3ABioNet) has as one of its goals the task of supporting the work of H3Africa, and African scientists more broadly by developing workflow for commonly performed analyses. Baichoo et al. [6] provide an overview of workflow development within H3ABioNet, including an introduction to a much earlier version of this workflow. The goal of the H3AGWAS workflow is to support scientists undertaking GWAS taking into account access to heterogeneous computing environments.

In summary, the goal of the H3AGWAS workflow is to provide a flexible, powerful and portable workflow for genome-wide association studies. The use of a workflow reduces the manual intervention required by human analysts, thereby reducing the overall time for a project to complete. Some phases of a GWAS are exploratory and analyses may need to be re-run as QC proceeds, and different parameters and analytic techniques tried after assessing initial results. The workflow needs to support reproducible analyses and be portable and scalable across many different computational environments (laptop to cluster to cloud), reflecting the heterogeneous environments across Africa. Using Nextflow and containerisation promotes scalability and portability.

Implementation

The workflow has been developed in Nextflow [7], with Python, bash and R scripts and uses well-known bioinformatics tools. It can easily be ported to different execution environments (e.g., standalone, job scheduling, cloud) and using containers to package software and dependencies assures replicability and simple installation. Figure 1 gives an overview.

Rather than producing one workflow which operates end-to-end, the H3AGWAS workflow is split into several independent sub workflows mapping to separate phases of work. Independent workflows allow users to execute parts that are only relevant to them at those different phases. For example, our experience has shown that the QC step requires multiple iterations over several weeks to find the best QC parameters and resolve problems with data. Once the QC is complete, the analysis moves to the next phase, which in turn may take weeks.

Sample runs and extensive documentation for the different phases can be found at `http://github.com/h3abionet/h3agwas/`.

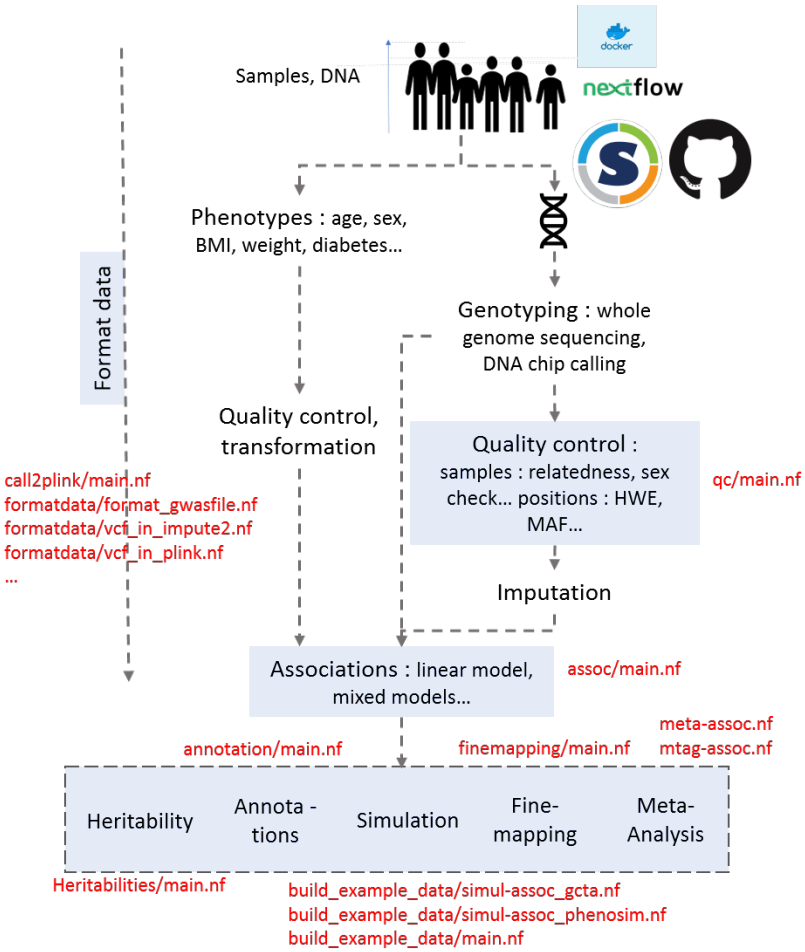

Figure 1: Overview of GWA studies from DNA sample until post-association analysis, box in blue corresponding to part of GWA studie present in H3AGWAS workflow and text in red corresponding to scripts in H3AGWAS workflow

Pre-association Workflows

Producing PLINK data The `call2plink` workflow converts Illumina genotyping reports into PLINK format.

Quality control The `qc` workflow performs quality control on a set of input PLINK file. The workflow considers per-sample and per-Single nucleotide polymorphism (SNP) missingness, minor allele frequency, levels of heterozygosity, highly related samples, possible duplicates, and sex mismatches, and also examines possible batch effects (for example, between cases and controls, for samples collected from different sites, or genotyped in different runs). A detailed report is produced which helps the user understand the data and which can be used in the methods section of a paper. All QC and workflow parameters (including software versions) and the MD5 checksums of input and output data are recorded in order to promote replicability and reduce the risk of version skew.

Association testing

The `assoc` workflow performs association on PLINK formatted files, including adjustment for multiple testing in PLINK. In addition to the basic association tests, the workflow currently supports Cochran-Mantel-Haenszel (CMH), linear and logistic regression, permutation and mixed-model association testing. This workflow provides user-selectable choices of software for association testing. PLINK is the work-horse for basic linear models, including support of covariates and adjusting for population structure. Exact linear mixed models with relatedness matrix have been included (Fast-LMM [8] and GEMMA [9]). For larger datasets, BOLT-LMM [10] and fastGWA or SAIGE [11] which use approximation of relatedness can be selected (and the workflow can compute the SNP-derived genetic relationship matrix (GRM) from genotype data using GCTA [12]). Each software uses as PLINK-formatted input files, but SAIGE also needs a VCF file containing genotype information. The workflow provides an option for BOLT-LMM's use of imputation probability in BGEN format – BGEN format can be extracted from VCF files after imputation using `formatdata/vcf_in_bgen.nf`

Many common complex traits are believed to be a result of the combined effect of genes, environmental factors and their interactions. Gene–environment interaction (G×E) can be analysed to detect loci where genotype-phenotype association may depend on the environment: G×E options from GEMMA and PLINK are implemented in the workflow (see Figure 2 and Table 1).

The PLINK input files are also used to perform a principal component analysis (PCA) and a PCA plot is generated that can be used to identify any possible population structure in the dataset.

Phase	Software/Resource	Workflows / Task	References
All	PLINK (1.9)	All	[13]
	python (3.0)	All	[14]
	R (3.6)	Plot and extraction data	[15]
Association testing	GEMMA	Association testing / Heritability	[9]
	Fast-LMM (binary version)	Association testing	[8]
	BOLT-LMM	Association testing / Heritability	[10]
	SAIGE	Association testing	[11]
	GCTA : fastGWA	Association testing	[12, 16]
Post-association analysis	GCTA : COJO, simulation, GREML	Fine-mapping/heritability/Simulation	[17]
	MetaSoft	Meta-analyses	[18]
	GWAMA	Meta-analyses	[19]
	METAL	Meta-analyses	[20]
	MTAG	Multi-trait association	[21]
	LDSC	Heritability	[22]
	PhenoSim	Simulation	[23]
	LocusZoom	Annotations	[24]
	Annovar	Annotations	[25]
Format data	BCFtools	Convert vcf	[26]
	QCTOOL (v2)	Convert vcf to bgen/bimbam	[27]
	CrossMap	Convert positions between genome builds	[28]

Table 1: List of softwares and resources used in H3AGWAS workflow, softwares are classify by phase of GWAS and task.

Post-association analysis

The post-association analysis workflows use genotype data and results of association testing in order to (1) find putative causal variants; (2) perform a meta-analysis or multi-trait genome-wide association study using summary statistics; (3) estimate global heritability; and (4) annotate positions (see Tables 1 and 2).

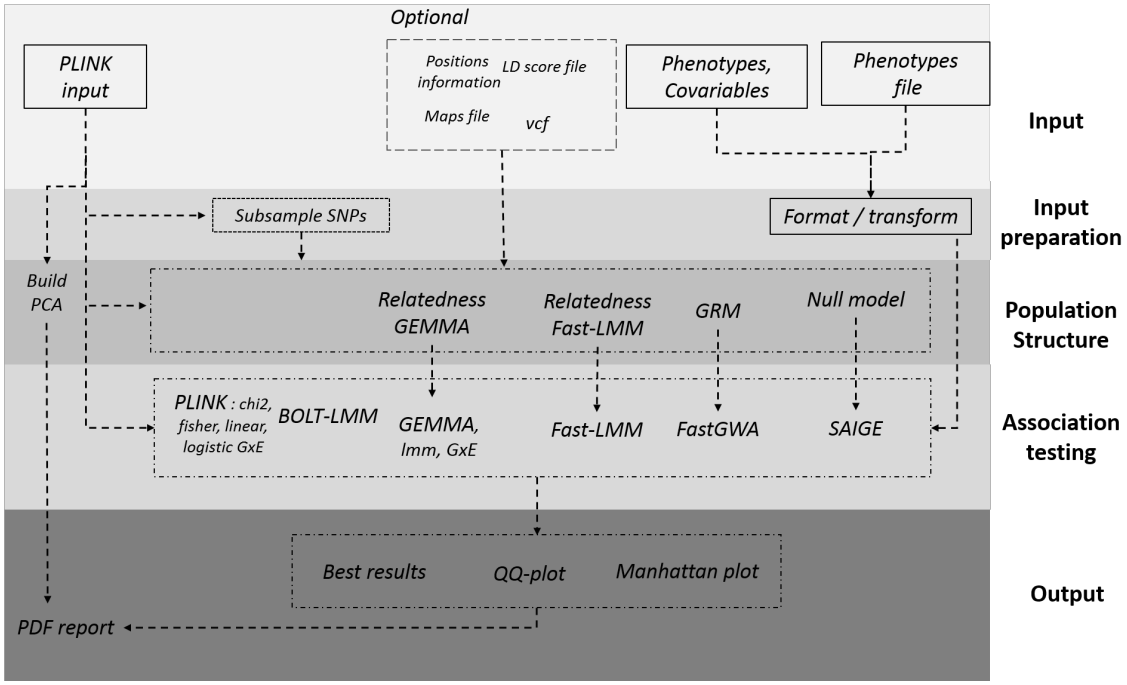

Figure 2: Workflow of association testing each background color represent different steps from input to output with preparation of input data, generated relatdness matrix or GRM to take account population structure and association testing

Genetic heritability and co-heritability of phenotypes We implemented scripts to compare heritability and co-heritability of phenotypes using different strategies. Two approaches are implemented. The first uses relatedness and phenotypes, based on REML or variance components analysis with BOLT-LMM [10, 29], GEMMA [9] and GCTA [16, 30]. The second uses summary statistics and methods implemented in GEMMA [31] and LDSC [22]. Furthermore, the workflow can compute the co-variability and co-heritability between phenotypes using relatedness and phenotypes using BOLT-LMMand GCTA or with summary statistics with LDSC.

Annotations We implemented a script to extract genotypes of each individual and compared to phenotype, annotated lead snps using Annovar [25] and plot region using LocusZoom software [24].

Simulations of phenotype We implemented a script to randomly build phenotype in order to estimate true and false positive detection in GWA. `simul-assoc_phenosim.nf` builds phenotypes using PhenoSim software [23] and genetics data where loci are randomly selected, followed by a GWA on the simulated data using BOLT-LMM and GEMMA.

In addition, the `build_example_data/main.nf` script builds phenotypes of individuals using initial genotype and allele effect. By default, the workflow uses 1000 Genomes data [32] and extracts effect of positions from the GWAS Catalog. The steps are: (1) extract and format 1000 Genomes data; (2) download GWAS catalog positions and results; (3) simulate phenotype in 1000 Genome individuals using effect of position, using GCTA [13].

Causal variants Two workflows have been implemented to detect causal variants: `assoc/cojo-assoc.nf` uses LD and conditional and joint analysis method to define lead SNPs [17]; and `finemapping/main.nf` uses genotypes, summary statistics and a region of interest to extract putative causal variants under a Bayesian framework. Various software can be : FINEMAP [19], CaviarBF [33] and PAINTOR [34]. Output includes results of all steps and plots of regions of interest with p-value and post probabilities obtained by fine-mapping to compare results.

Meta-analysis and multi-trait genome-wide association Script `mtag-assoc` performs a multi-trait genome-wide association using the `mtag` software [21] for joint analysis of summary statistics from GWASs of different traits. The `meta-assoc.nf` workflow performs meta-analysis with different software and statistical approaches, to account for variability between datasets, genomic inflation or overlap between samples with METAL [20] or GWAMA [19] and Metasoft [18, 35].

Format file

Many GWAS tools use different formats and being able to convert easily between them is useful. We provide various scripts to support this. For instance the `formatdata/plk_in_vcf_imp.nf` script prepares data for imputation. There are scripts that transform VCF data imputed in various formats to PLINK, bimbam or impute2 format. `formatdata/convert_posversiongenome.nf` converts genomic coordinates between different assemblies, for example between GRCh38 and hg19, using a combination of conversion with CrossMap [28].

Installation and support

The H3AGWAS workflow requires Java 8 or later and Nextflow, and can either be cloned from GitHub explicitly or run directly using Nextflow.

In addition, the workflow relies on a number of state-of-the art bioinformatics tools (Table 2, 1). We recommend that users install either Singularity or Docker and then run H3AGWAS workflow workflow using the appropriate profile – we provide containers with all tools bundled. These containers will automatically be installed on the first execution of the workflow. However, for those users who are not able to use Singularity or Docker or who would like control over which versions of the tools are used, the Docker files can be used to guide someone with basic system administration skills to install the necessary dependencies.

Manuals and examples can be found at https://github.com/h3abionet/. Problems or help with the workflow itself is provided through GitHub issues. The H3ABioNet supports general queries from African researchers about the use of the workflow or GWAS in general through its help desk [36] (https://helpdesk.h3abionet.org).

FAIR

The workflow was developed to be "Findable, Accessible, Interoperable and Reusable" according to guidelines on the FAIR https://fair-software.eu/ website. The H3AGWAS workflow has been registered in bio.tools (https://bio.tools/h3agwas), uses a MIT Licence, contain citation metadata files, and uses a software quality checklist via a Core Infrastructure Initiative (CII) Best Practices badge (https://bestpractices.coreinfrastructure.org/en).

Phase	script names	Description
Pre-association	qc/qc.nf	Quality control of genetics data
	call2plink/main.nf	Converting from Illumina genotyping reports in TOP/BOTTOM or Forward
Association testing	assoc/main.nf	Run association and GxE using genetics on PLINK file and phenotype(s)
Post-association analysis	finemapping/cojo-assoc.nf	Conditional and joint analysis using GCTA
	finemapping/finemap_region.nf	fine-mapping on specific region using FineMap,
	finemapping/main.nf	Extract lead SNPs and perform a fine-mapping on each region using FineMap,
	heritabilities/main.nf	Estimated heritability using GCTA, BOLT-LMM, LDSC or GEMMA
	replication/gwascat/main.nf	Extraction of replication using GWAS catalog by positions and linkage disequilibrium.
	meta/meta-assoc.nf	Meta analysis using PLINK, Meta-soft, GWAMA and Metal
	meta/mtag-assoc.nf	Multi-trait genome-wide association using mtag software
	utils/annotation/annot-assoc.nf	Locus zoom, annotation using , distribution of phenotype by genotype
	utils/build_example_data/main.nf	Simulated phenotype with gcta using 1000Genome and GWAS catalog result
	utils/build_example_data/simul-assoc.nf	Simulated phenotype using phenosim
Format data	formatdata/ convert_posversiongenome.nf	Convert position between reference
	formatdata/ format_gwasfile.nf	Format GWAS file
	formatdata/ plk_in_vcf_imp.nf	Format PLINK in VCF for imputation
	formatdata/ vcf_in_bimbam.nf	Use VCF output of imputation to format in bimbam
	formatdata/ vcf_in_impute2.nf	Use VCF output of imputation to format in impute2 format
	formatdata/ vcf_in_plink.nf	Use VCF output of imputation to format in PLINK format
	formatdata/ vcf_in_bgen_merge.nf	Use VCF output of imputation to format in bgen format

Table 2: List and description of Nextflow scripts by phase of GWAS

Results and discussion

Each workflow was tested on the Wits University Core Research Cluster (CentOS 7, SLURM) and Singularity images [37], and on Amazon AWS and Microsoft Azure. It has also been used in production on other environments. Since it uses Nextflow and containers, it can run on any environment that Nextflow supports such as PBS/Torque.

We illustrate the use of the workflow with a real dataset from the H3Africa AWI-Gen Collaborative Centre [38]. The data comes from a cross-sectional study that investigated populations from six sub-Saharan African sites – ≈12,000 black African men and women from two urban settings (Nairobi and Soweto) and four rural settings (Agincourt, Dikgale, Nanoro and Navrongo), aged 40 to 80 years. DNA from these individuals was genotyped on the H3Africa Custom Array (https://chipinfo.h3abionet.org), designed as an African common variant enriched GWAS array with ≈2.3 million SNPs. QC was run on the array dataset resulting in $\approx 10,600$ individuals and ≈1.733m SNPs. Imputation was performed on the cleaned dataset using the Sanger Imputation Server and the African Genome Resources as a reference panel. We selected EAGLE2 [39] for pre-phasing and the default PBWT algorithm was used for imputation. The resulting data was used for the following phases.

Testing of different workflow

- QC : Quality control of genotype data was tested using AWI-Gen dataset with 12,000 individuals before imputation.
- Association testing: For association testing, we used four residuals of lipid phenotype: LDL, cholesterol, HDL and triglycerides normalised using sex and age followed by an inverse normal transformation. We ran simultaneous linear associations with PLINK [13], GEMMA using Univariate Linear Mixed Model [9], BOLT-LMM using mixed model analysis [10], fastGWA from GCTA using mixed linear model and SAIGE [11] [12, 16].
- Meta-analysis : The meta-analysis workflow was tested using GEMMA summary statistics of cholesterol from each region of AWI-Gen dataset: South Africa, east Africa and west Africa.
- Other scripts : Testing of other scripts is summarized in Table 5. The `finemapping` script was tested using cholesterol result of GEMMA. Conversion of PLINK to VCF was tested using genotypes processed by the QC workflow. Conversion of VCF to PLINK, bimbam, impute2 was tested using data after imputation.

Association testing

The association workflow was tested using 10,700 individuals, four phenotypes and 14 million imputed positions with PLINK, GEMMA, BOLT-LMM, fastGWA and SAIGE. We excluded Fast-LMM from testing given that it required over 100 GB of memory for a single chromosome. Using the Wits Core cluster[1], the workflow ran with an elapsed time of 14h 37min. Among the four programs used for association, GEMMA used most computing time and jobs, followed by SAIGE, fastGWA, BOLT-LMM and PLINK. Other processes took less than 8% of

[1]This is a production cluster and while the cluster was lightly loaded at this time there were other jobs running

CPU time Table 3. The largest maximum memory (resident set size) used by any job was 6.9 GB.

Process	Tot hours	% times	% cpu number used (Mean)	Max mem (MB)	NF processes
computePCA	2.3	0.42	300.40	1100.0	1
drawPCA	0.0	0.00	104.30	87.9	1
extractPheno	0.0	0.00	76.20	39.1	1
buildindex	7.8	1.39	98.24	7.4	22
getBoltPhenosCovar	0.0	0.00	83.70	54.3	1
select_rs_format	0.1	0.01	100.00	609.7	1
format_genetic_ldscore	0.6	0.10	36.3	2200.00	1
FastGWADoGRM	7.5	1.33	492.67	2200.0	100
MergFastGWADoGRM	0.0	0.01	1.30	148.0	1
computeTest	8.5	1.52	95.35	824.1	4
doBoltmm	17.2	3.05	623.23	5600.0	4
getListeChroGem	0.0	0.00	78.40	5.0	1
getGemmaRel	1.4	0.25	832.00	5100.0	1
doGemmaChro	286.6	50.93	958.16	5400.0	88
doMergeGemma	0.1	0.02	8.43	3.8	4
FastGWARun	41.3	7.35	972.45	1900.0	4
ShowManhattan	0.6	0.11	96.34	8200.0	16
checkidd_saige	0.1	0.01	12.50	135.6	1
getSaigePheno	0.0	0.00	72.60	43.9	1
subplink_heritability_saige	0.30	0.05	7.60	929.2	1
saige_computed_variance	20.6	3.66	919.58	3200.0	4
doSaige	167.3	29.73	99.90	293.2	88
doMergeSaige	0.2	0.04	12.18	3.8	4
drawPlinkResults	0.1	0.02	71.58	3700.0	4
showPhenoDistrib	0.0	0.00	80.10	121.2	1
doReport	0.00	0.00	46.50	26.1	1

Table 3: Cost of association running on cluster, using 4 phenotypes and 10,000 individuals. The elapsed time for running the entire workflow was 14h 37min, with a high degree of parallelisation. *Process* is the Nextflow process name; *Tot hours* – total CPU hours used by instances of this NF process; *% times* – percentage of times used by process compared to other process ; *% cpu number used (Mean)* – mean % cpu number used by instances of the process and so is a measure of achievable parallelism for instances of that process; *Max mem (MB)* is maximum of memory (resident set size) used by one process; *NF processes* – number of Nextflow process used for the steps, and a measure of parallelism at very coarse level.

Meta-analysis workflow

As an illustration, we performed meta-analysis (`meta/meta-assoc.nf`) using 3 summary statistics, each with 14 million SNPs. The script ran in 34 minutes in total, with METAL using the smallest amount of processing time (1.8 minutes) and GWAMA using the greatest amount of processing time. The greatest amount of memory (10 GB) was also used by GWAMA, whereas PLINK used the least (2 GB; Table 4).

Process	Tot hours	% times	% cpu number used (Mean)	Max mem (MB)	NF processes
GetRsFile	1.1	1.18	99.50	8.5	1
ChangeFormatFile	29.90	30.69	65.23	4100.0	3
doGWAMA	18.0	18.45	67.40	10400.0	1
doMetal	2.9	2.99	99.60	3500.0	1
doMetaSoft	12.3	12.65	103.10	7700.0	1
doMRMEGA	15.8	16.23	68.10	8100.0	1
doPlinkMeta	8.9	9.15	99.80	2000.0	1
showGWAMA	1.4	1.49	100.20	5800.0	1
showMetal	1.8	1.80	97.70	3800.0	1
showMetasoft	1.9	1.92	100.10	7500.0	1
showMRMEGA	1.6	1.62	100.20	6200.0	1
showPlink	1.2	1.25	100.00	4800.0	1
doReport	0.6	0.58	46.10	24.3	1

Table 4: Cost of running meta-analysis worflow using Wits cluster, using 3 summary statistics and 14 millions of positions by summary statistics. Column labels as in Table 3

Others tests

Each script has been tested using the AWI-Gen dataset, as summarised in Table 5. The supplementary materials give more detail showing the costs of each step being run on a Linux cluster with SLURM and using Singularity images.

Cloud computing

The QC and association workflows have been tested on Amazon Web Services (AWS) as well as Microsoft Azure using batch processing through Nextflow. All workflows have configuration files that include profiles for use on AWS and Azure, and instructions are provided in the README for the workflow. Using a large simulated dataset with 22k individuals across 2.2m SNPs, the QC script took 8.6 hours to run on AWS and 20 hours to run on Azure, with cost between US$5-US$10 using spot pricing.

Contribution and related work

The H3AGWAS workflow provides a comprehensive suite of workflows for GWAS that are portable and scaleable. Few existing workflows integrate so many steps of GWAS, from QC to post-association analysis.

The closest competing workflow is BIGwas [40] which provides both QC and association testing. Kässens *et al.* compared BIGwas to an earlier version of H3AGWAS workflow. With respect to QC they found that the two were roughly equivalent in functionality but that BIGwas was much faster. However, we have been unable to replicate their findings and our experimentation shows that the QC and association testing H3AGWAS workflow execution on default parameters is much faster (see supplementary material). However, although the workflow engineering is important to performance, the computational cost primarily depends on underlying tools rather than the virtues of the workflow. With respect to association testing and pre- and post-analysis, they found their workflow superior. Whatever arguable

Script	Test descriptive
qc/main.nf	QC of genotype 12,000 individuals from AWI-Gen project
finemapping/main.nf	Extraction of lead SNPs of cholesterol result from GEMMA
heritabilities/main.nf	Estimation of heritabilities of 4 phenotypes lipid using genotype and phenotype and/or summary statistics
formatdata/plk_in_vcf_imp	Transformation of PLINK after qc in vcf to prepared data for imputation
formatdata/vcf_in_plink.nf	Transformation of data imputed in PLINK format
meta/mtag-assoc.nf	mtag performed using 4 lipid phenotypes
utils/build_example_data/main.nf	Build an example dataset using diabetes phenotype from lead snps of GWAS catalog and 1000 genomes project genotype [32]

Table 5: List of evaluation of additional workflow implemented in H3AGWAS workflow. using AWI-Gen dataset or 1000 genome project

shortcomings the H3AGWAS workflow may have had in October 2020, in March 2022, in the H3AGWAS workflow it has significantly more extensive in functionality. In addition, the H3AGWAS workflow has two significant advantages: (1) it supports cloud computing directly through the use of AWS and Azure batch; and (2) relatively lightweight Singularity/Docker containers allow deployment in HPC environments where *setuid* for Singularity is often disabled (see the supplementary for an explanation).

Other tools that are available are summarised in Table 6.

Conclusion

The H3AGWAS workflow provides a suite of workflows from quality control of genomic data to post-association analysis of result. Using Nextflow and containers allows the workflow to be easily installed and to be portable and scalable – from laptop to server to cloud (AWS and Azure). The multiple workflow scripts map to natural GWAS workflow phases.

Pre-association scripts focus on quality control, with imputation performed by a separate workflow. We plan to add calling of array data to the workflow in the future. Association studies, including G×E analysis, can be performed in our workflow using six different techniques provided by state-of-the-art tools. Post-analysis of GWAS supports meta analysis, heritability

computation, identifying causal SNPs, co-localisation and fine-mapping.

We support multiple tools, providing users with opportunities to compare results (e.g., different approaches for fine-mapping and association testing). Furthermore, different Nextflow scripts for each step allows the user to run analyses with different parameters and customise the analysis to their needs. Each script is associated with a Docker image to simplify installation, and returns a PDF report to the researcher to help to interpret the results.

The workflows are available on GitHub and we strive to comply with FAIR principles.

Future development

Several additional features are in development. In pre-association, calling genotypes from raw array data is challenging, and we are currently working on a workflow to perform this step. New features to be added are supporting replication and transferability of previous result using GWAS Catalog result [51] or full summary statistics. We plan to port to DSL2 and make nf-core compatible.

Availability and requirements

Project name:	H3AGWAS workflow
Project home page:	`https://github.com/h3abionet/h3agwas`
Operating system:	Linux (or MacOS and Windows with Docker)
Program language:	Nextflow, Python, R, bash
Other requirements:	Java 8 or later, Nextflow Docker/Singularity (or software dependencies listed in Dockerfile)
Licence:	MIT Licence
Restrictions on non-academic use:	None

Docker images are available from `https://quay.io/organization/h3abionet_org/` and `https://github.com/h3abionet/h3agwas-docker`.

Ethics

We have used the AWI-Gen data set as our main example as a real data set. The AWI-Gen study received approval from the Human Research Ethics Committee (Medical), University of the Witwatersrand, South Africa (M121029, M1706110). The AWI-Gen data set is available from the European Genome-Phenome Archive on application to the independent H3Afica Data Access Committee (EGAD00001006425 and EGAD00010001996). We hope to make our synthetic data sets publicly available.

Funding

The work is supported by National Human Genome Research Institute/National Institutes of Health: JTB is supported by the AWI-Gen Collaborative Centre (U54HG006938) and all other

authors and the Wits Core Cluster are supported by the Pan-African Bioinformatics Network for H3Africa (U24HG006941). The views expressed are solely those of the authors and not that of the NIH.

Conflict interest and consents

Each authors contribute and corrected papers and Consent for publication. There is no conflict of interest.

Authors contribution

JTB developed the workflow and led the writing of this paper. LC contributed to worflow development. SB, CJF and SP led the H3ABioNet workflows project and provided scientific input and direction. SH led the H3AGWAS workflow project, and wrote parts of the workflow, and co-led the writing of the paper. All authors contributed to the writing of this paper.

Software	Analysis	link
GWASTools (Bioconductor)	QC	https://www.bioconductor.org/packages/release/bioc/html/GWASTools.html, [41]
BigGWAS	QC, Association testing	https://github.com/ikmb/gwas-assoc, [40]
plinkQC	QC	https://meyer-lab-cshl.github.io/plinkQC/articles/plinkQC.html, [42]
GWAS-ellingson	QC	https://github.com/sallyrose0425/GWAS, [43]
TASSEL	Association testing with general linear model and mixed linear model approaches, some post analysis analysis	[44]
CAUSALdb-finemapping	Fine-mapping using different software	https://github.com/mulinlab/CAUSALdb-finemapping-pip, [45, 46]
FINNGEN	Fine-mapping	https://github.com/FINNGEN/finemapping-pipeline
FUMA	post-association analysis using web interface	https://fuma.ctglab.nl/, [47, 48]
postgap	post-association analysis with annotation through cis-regulatory datasets using python	https://github.com/Ensembl/postgap, [49]
nf-gwas-pipeline	QC, Association testing integrating R packages SNPRelate/GENESIS/GMMAT and ANNOVAR using Nextflow	https://github.com/montilab/nf-gwas-pipeline [50]
gwasglue	post-association analysis with colocalisation, fine-mapping Mendelian randomisation using R	https://mrcieu.github.io/gwasglue

Table 6: Non-exhaustive list of workflows that perform QC, association testing and/or post-association analysis of GWAS

Acknowledgements

The authors thank all AWI-Gen collaborators for use of their dataset and more particularly Michèle Ramsay Principal Investigator and all participants of AWI-Gen. Our SBIMB colleagues have made many useful and generous contributions: Carl Wenlong Chen, Palwende R. Boua and Shaun Aron in particular. Many people contributed to the workflow and we especially thank Lerato Magosi, Rob Clucas and Eugene de Beste whose effort at the start of the project was so important. We thank Professor Nicola Mulder from the University of Cape Town whose leadership of H3ABioNet made the work possible.

References

[1] Uffelmann E, Huang QQ, Munung NS, de Vries J, Okada Y, Martin AR, et al. Genome-wide association studies. Nature Reviews Methods Primers. 2021 Aug;1(1):1–21. Available from: https://www.nature.com/articles/s43586-021-00056-9.

[2] Marees AT, de Kluiver H, Stringer S, Vorspan F, Curis E, Marie-Claire C, et al. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. International Journal of Methods in Psychiatric Research. 2018 Feb;27(2). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6001694/.

[3] Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT. Data quality control in genetic case-control association studies. Nature Protocols. 2010 Sep;5(9):1564–1573. Available from: https://www.nature.com/articles/nprot.2010.116.

[4] Laurie CC, Doheny KF, Mirel DB, Pugh EW, Bierut LJ, Bhangale T, et al. Quality control and quality assurance in genotypic data for genome-wide association studies. Genetic epidemiology. 2010 Sep;34(6):591–602. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3061487/.

[5] Mulder NJ, Adebiyi E, Alami R, Benkahla A, Brandful J, Doumbia S, et al. H3ABioNet, a sustainable pan-African bioinformatics network for human heredity and health in Africa. Genome Research. 2016;26(2).

[6] Baichoo S, Souilmi Y, Panji S, Botha G, Meintjes A, Bendou H, et al. Developing reproducible bioinformatics analysis workflows for heterogenous computing environments to support African genomics. BMC Bioinformatics. 2018;19(457):1–9.

[7] Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nature Biotechnology. 2017 Apr;35(4):316–319. Available from: https://www.nature.com/articles/nbt.3820.

[8] Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D. FaST linear mixed models for genome-wide association studies. Nature Methods. 2011 Oct;8(10):833–835. Available from: https://www.nature.com/articles/nmeth.1681.

[9] Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nature Genetics. 2012 Jul;44(7):821–824. Available from: https://www.nature.com/articles/ng.2310.

[10] Loh PR, Kichaev G, Gazal S, Schoech AP, Price AL. Mixed-model association for biobank-scale datasets. Nature Genetics. 2018 Jul;50(7):906–908.

[11] Zhou W, Nielsen JB, Fritsche LG, Dey R, Gabrielsen ME, Wolford BN, et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nature Genetics. 2018 Sep;50(9):1335–1341.

[12] Jiang L, Zheng Z, Qi T, Kemper KE, Wray NR, Visscher PM, et al. A resource-efficient tool for mixed model association analysis of large-scale data. Nature Genetics. 2019 Dec;51(12):1749–1755. Available from: https://www.nature.com/articles/s41588-019-0530-8.

[13] Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience. 2015;4(1):1–16. Available from: http://www.gigasciencejournal.com/content/4/1/7.

[14] Van Rossum G, Drake FL. Python 3 Reference Manual. Scotts Valley, CA: CreateSpace; 2009.

[15] R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2020. Available from: https://www.R-project.org/.

[16] Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: A Tool for Genome-wide Complex Trait Analysis. American Journal of Human Genetics. 2011 Jan;88(1):76–82. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3014363/.

[17] Yang J, Ferreira T, Morris AP, Medland SE, Genetic Investigation of ANthropometric Traits (GIANT) Consortium, DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nature Genetics. 2012 Mar;44(4):369–375, S1–3.

[18] Han B, Eskin E. Random-Effects Model Aimed at Discovering Associations in Meta-Analysis of Genome-wide Association Studies. The American Journal of Human Genetics. 2011 May;88(5):586–598. Available from: https://www.cell.com/ajhg/abstract/S0002-9297(11)00155-8.

[19] Mägi R, Morris AP. GWAMA: software for genome-wide association meta-analysis. BMC bioinformatics. 2010 May;11:288.

[20] Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics (Oxford, England). 2010 Sep;26(17):2190–2191.

[21] Turley P, Walters RK, Maghzian O, Okbay A, Lee JJ, Fontana MA, et al. Multi-trait analysis of genome-wide association summary statistics using MTAG. Nature genetics. 2018 Feb;50(2):229–237. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5805593/.

[22] Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, Loh PR, et al. An atlas of genetic correlations across human diseases and traits. Nature Genetics. 2015 Nov;47(11):1236–1241. Available from: https://www.nature.com/articles/ng.3406.

[23] Günther T, Gawenda I, Schmid KJ. phenosim–A software to simulate phenotypes for testing in genome-wide association studies. BMC bioinformatics. 2011 Jun;12:265.

[24] Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, et al. LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics (Oxford, England). 2010 Sep;26(18):2336–2337.

[25] Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Research. 2010 Sep;38(16):e164–e164. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkq603.

[26] Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. GigaScience. 2021 Feb;10(2):giab008.

[27] Wigginton JE, Cutler DJ, Abecasis GR. A Note on Exact Tests of Hardy-Weinberg Equilibrium. The American Journal of Human Genetics. 2005 May;76(5):887–893. Available from: https://www.cell.com/ajhg/abstract/S0002-9297(07)60735-6.

[28] Zhao H, Sun Z, Wang J, Huang H, Kocher JP, Wang L. CrossMap: a versatile tool for coordinate conversion between genome assemblies. Bioinformatics (Oxford, England). 2014 Apr;30(7):1006–1007.

[29] Loh PR, Bhatia G, Gusev A, Finucane HK, Bulik-Sullivan BK, Pollack SJ, et al. Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. Nature Genetics. 2015 Dec;47(12):1385–1392.

[30] Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. Nature Genetics. 2010 Jul;42(7):565–569. Available from: https://www.nature.com/articles/ng.608.

[31] Zhou X. A UNIFIED FRAMEWORK FOR VARIANCE COMPONENT ESTIMATION WITH SUMMARY STATISTICS IN GENOME-WIDE ASSOCIATION STUDIES. The Annals of Applied Statistics. 2017 Dec;11(4):2027–2051.

[32] Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, et al. A global reference for human genetic variation. Nature. 2015 Oct;526(7571):68–74. Available from: https://www.nature.com/articles/nature15393.

[33] Chen W, Larrabee BR, Ovsyannikova IG, Kennedy RB, Haralambieva IH, Poland GA, et al. Fine Mapping Causal Variants with an Approximate Bayesian Method Using Marginal Test Statistics. Genetics. 2015 Jul;200(3):719–736.

[34] Gong Y, Greenbaum J, Deng HW. A statistical approach to fine-mapping for the identification of potential causal variants related to human intelligence. Journal of human genetics. 2019 Aug;64(8):781–787. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6712985/.

[35] Han B, Eskin E. Interpreting Meta-Analyses of Genome-Wide Association Studies. PLOS Genetics. 2012 Mar;8(3):e1002555. Available from: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1002555.

[36] Kumuthini J, Zass L, Panji S, Salifu SP, Kayondo JK, Nembaware V, et al. The H3ABioNet helpdesk: An online bioinformatics resource, enhancing Africa's capacity for genomics research. BMC Bioinformatics. 2019;20(1):1–7.

[37] Kurtzer GM, Sochat V, Bauer MW. Singularity: Scientific containers for mobility of compute. PLOS ONE. 2017 May;12(5):e0177459. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0177459.

[38] Ramsay M, Crowther N, Tambo E, Agongo G, Baloyi V, Dikotope S, et al. H3Africa AWI-Gen Collaborative Centre: a resource to study the interplay between genomic and environmental risk factors for cardiometabolic diseases in four sub-Saharan African countries. Global Health, Epidemiology and Genomics. 2016 Nov;1:e20. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5732578/.

[39] Loh PR, Danecek P, Palamara PF, Fuchsberger C, A Reshef Y, K Finucane H, et al. Reference-based phasing using the Haplotype Reference Consortium panel. Nature Genetics. 2016 Nov;48(11):1443–1448. Available from: https://www.nature.com/articles/ng.3679.

[40] Kässens JC, Wienbrandt L, Ellinghaus D. BIGwas: Single-command quality control and association testing for multi-cohort and biobank-scale GWAS/PheWAS data. GigaScience. 2021 06;10(6). Giab047. Available from: https://doi.org/10.1093/gigascience/giab047.

[41] Gogarten SM, Bhangale T, Conomos MP, Laurie CA, McHugh CP, Painter I, et al. GWASTools: an R/Bioconductor package for quality control and analysis of genome-wide association studies. Bioinformatics (Oxford, England). 2012 Dec;28(24):3329–3331.

[42] Meyer HV. HannahVMeyer/plinkQC: plinkQC version 0.2.3. Zenodo; 2019. Available from: https://zenodo.org/record/3373798.

[43] Ellingson SR, Fardo DW. Automated quality control for genome wide association studies. F1000Research. 2016 Jul;5. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5007749/.

[44] Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics. 2007 06;23(19):2633–2635. Available from: https://doi.org/10.1093/bioinformatics/btm308.

[45] Wang J, Huang D, Zhou Y, Yao H, Liu H, Zhai S, et al. CAUSALdb: a database for disease/trait causal variants identified using summary statistics of genome-wide association studies. Nucleic Acids Research. 2019 11;48(D1):D807–D816. Available from: https://doi.org/10.1093/nar/gkz1026.

[46] Schaid DJ, Chen W, Larson NB. From genome-wide associations to candidate causal variants by statistical fine-mapping. Nature Reviews Genetics. 2018 Aug;19(8):491–504. Available from: https://www.nature.com/articles/s41576-018-0016-z.

[47] Watanabe K, Taskesen E, van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. Nature Communications. 2017 Nov;8(1):1826. Available from: https://www.nature.com/articles/s41467-017-01261-5.

[48] Watanabe K, Umićević Mirkov M, de Leeuw CA, van den Heuvel MP, Posthuma D. Genetic mapping of cell type specificity for complex traits. Nature Communications. 2019 Jul;10(1):3222. Available from: https://www.nature.com/articles/s41467-019-11181-1.

[49] Peat G, Jones W, Nuhn M, Marugán JC, Newell W, Dunham I, et al. The open targets post-GWAS analysis pipeline. Bioinformatics. 2020 May;36(9):2936–2937. Available from: https://doi.org/10.1093/bioinformatics/btaa020.

[50] Song Z, Gurinovich A, Federico A, Monti S, Sebastiani P. nf-gwas-pipeline: A Nextflow Genome-Wide Association Study Pipeline. Journal of Open Source Software. 2021 Mar;6(59):2957. Available from: https://joss.theoj.org/papers/10.21105/joss.02957.

[51] Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Research. 2019 Jan;47(D1):D1005–D1012.