Bioinformatics, 36(3), 2020, 930–933
doi: 10.1093/bioinformatics/btz633
Advance Access Publication Date: 8 August 2019
Applications Note

Genome analysis

RICOPILI: Rapid Imputation for COnsortias PIpeLIne

Max Lam [1,2,3,4,5,†], Swapnil Awasthi,[1,6,†], Hunna J. Watson[7,8,9], Jackie Goldstein[1,2], Georgia Panagiotaropoulou[1,6], Vassily Trubetskoy[1,6], Robert Karlsson [10], Oleksander Frei[11], Chun-Chieh Fan[11], Ward De Witte[12], Nina R. Mota [12,13], Niamh Mullins[14], Kim Brügger[15], S. Hong Lee[16], Naomi R. Wray[17,18], Nora Skarabis[6], Hailiang Huang[1,2], Benjamin Neale[1,2], Mark J. Daly[1,2], Manuel Mattheisen[19,20,21], Raymond Walters[1,2] and Stephan Ripke[1,2,6,*]

[1]Stanley Center for Psychiatric Research, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA, [2]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Cambridge, MA 02114, USA, [3]Division of Psychiatry Research, The Zucker Hillside Hospital, Northwell Health, Glen Oaks, NY 11004, USA, [4]Research Division, Institute of Mental Health Singapore, Singapore 539747, Singapore, [5]Human Genetics, Genome Institute of Singapore, Singapore 138672, Singapore, [6]Department of Psychiatry and Psychotherapy, Charité - Universitätsmedizin, 10117 Berlin, Germany, [7]Department of Psychiatry, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA, [8]Division of Paediatrics, School of Medicine, The University of Western Australia, Perth, WA 6009, Australia, [9]School of Psychology, Curtin University, Perth, WA 6102, Australia, [10]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, 171 77 Stockholm, Sweden, [11]NORMENT, KG Jebsen Centre for Psychosis Research, Institute of Clinical Medicine, University of Oslo, Ullevål Hospital, N-0424 Oslo, Norway, [12]Department of Human Genetics, Donders Institute for Brain, Cognition and Behaviour, [13]Department of Psychiatry, Donders Institute for Brain, Cognition and Behaviour, Radboud University Medical Center, 6525 GA Nijmegen, The Netherlands, [14]Pamela Sklar Division of Psychiatric Genomics, Department of Genetics and Genomic Science, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA, [15]Department of Informatics, University of Bergen, 5020 Bergen, Norway, [16]Australian Centre for Precision Health, University of South Australia Cancer Research Institute, University of South Australia, Adelaide, South Australia 5000, Australia, [17]Institute for Molecular Bioscience, The University of Queensland, [18]Queensland Brain Institute, The University of Queensland, Brisbane, QLD 4072, Australia, [19]Department of Psychiatry, Psychosomatics and Psychotherapy, Center of Mental Health, University Hospital Würzburg, Würzburg, Germany, [20]Department of Clinical Neuroscience, Centre for Psychiatric Research, Karolinska Institutet, Stockholm, Sweden and [21]Department of Biomedicine, Aarhus University, Aarhus, Denmark

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first two authors contributed equally.
Associate Editor: Russell Schwartz

Received on April 18, 2019; revised on July 31, 2019; editorial decision on August 2, 2019; accepted on August 7, 2019

Abstract

Summary: Genome-wide association study (GWAS) analyses, at sufficient sample sizes and power, have successfully revealed biological insights for several complex traits. RICOPILI, an open-sourced Perl-based pipeline was developed to address the challenges of rapidly processing large-scale multi-cohort GWAS studies including quality control (QC), imputation and downstream analyses. The pipeline is computationally efficient with portability to a wide range of high-performance computing environments. RICOPILI was created as the Psychiatric Genomics Consortium pipeline for GWAS and adopted by other users. The pipeline features (i) technical and genomic QC in case-control and trio cohorts, (ii) genome-wide phasing and imputation, (iv) association analysis, (v) meta-analysis, (vi) polygenic risk scoring and (vii) replication analysis. Notably, a major differentiator from other GWAS pipelines, RICOPILI leverages on automated parallelization and cluster job management approaches for rapid production of imputed genome-wide data. A comprehensive meta-analysis of simulated GWAS data has been incorporated demonstrating each step of the pipeline. This includes all the associated visualization plots, to allow ease of data interpretation and manuscript preparation. Simulated GWAS datasets are also packaged with the pipeline for user training tutorials and developer work.

Availability and implementation: RICOPILI has a flexible architecture to allow for ongoing development and incorporation of newer available algorithms and is adaptable to various HPC environments (QSUB, BSUB, SLURM and others). Specific links for genomic resources are either directly provided in this paper or via tutorials and external links. The central location hosting scripts and tutorials is found at this URL: https://sites.google.com/a/broadinstitute.org/RICOPILI/home
Contact: sripke@broadinstitute.org
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Genome-wide association studies (GWASs) have enabled the discovery of genetic variants underlying a plethora of complex traits (https://www.ebi.ac.uk/gwas/diagram). GWASs have highlighted previously unknown biological mechanisms associated with complex diseases and traits (Breen *et al.*, 2016). The Psychiatric Genomics Consortium (PGC) (http://www.med.unc.edu/pgc) the largest umbrella organization for psychiatric genetics (Sullivan *et al.*, 2018)—have made possible to advance the objectives of (i) revealing biological insights of psychiatric illness, (ii) informing clinical practice and (iii) presenting new therapeutic targets through sheer number of cohorts for GWASs across various psychiatric traits (Breen *et al.*, 2016; Sullivan *et al.*, 2012). The exponential availability of cohorts requires efficient, consistent and standardized approaches for various aspects of GWAS data management and analysis. Here, we introduce RICOPILI, the pipeline that automates rapid GWAS analysis workflow across various PGC workgroups. The pipeline is state-of-art, constantly incorporating latest available GWAS computational techniques and methods. With open-sourced simulated GWAS datasets and training tutorials packaged with the pipeline, RICOPILI is ideal for those contributing to large-scale genetic studies.

1.1 Comparison with other GWAS quality control and imputation pipelines

To our understanding, RICOPILI is the only open-sourced GWAS pipeline allowing secure data management, efficient data processing and downstream analysis scalable on both desktop and cluster environment. First, RICOPILI features an integrated quality control (QC), imputation and association analysis within its framework. Second, RICOPILI allows more than one imputation approach and reference panel to be utilized within its framework. Furthermore, computer intensive imputation can be processed locally within a closed cluster system. Third, the RICOPILI framework allows scalable processing of GWAS data, from a single CPU, to a cluster set up, or even within the cloud-based systems.

We compare RIOCPILI to existing available GWAS processing pipelines in Supplementary Table S1. All other tools focus on specific stages of GWAS analysis and do not provide the comprehensive features of RICOPILI. In the ensuing sections we will further highlight and discuss the features and functions of RICOPILI.

2 Design and implementation

2.1 Pipeline description

RICOPILI automates and integrates standard GWAS analysis methods, allowing for automated cluster submission and parallelization. The pipeline unifies standard software for its functions and implements best data analysis practices, provides sensible default settings while permitting the user to flexibly customize filters, thresholds and job resources as required. The optimization of cluster resources allows computations and visualizations to be completed quickly without significant user intervention. Written predominantly in Perl and R, the pipeline is organized according to analysis modules. Each module runs in its entirety via a single command line. The main module functions include:

- Pre-imputation technical QC;
- Principal components analysis (PCA) and relatedness estimation;
- Genome-wide imputation of genotype probabilities and generation of best guess genotypes in PLINK format (Purcell *et al.*, 2007);
- Downstream analyses, including GWAS, meta-analysis and polygenic risk scoring;
- Harmonizing large imputation reference panels (such as 1000 Genomes and the Haplotype Reference Consortium) to fit the architecture of RICOPILI.

RICOPILI takes dataset with unfiltered genotype calls, through trait association analysis, multi-cohort meta-analysis, linkage disequilibrium (LD) score regression (Bulik-Sullivan *et al.*, 2015), conditional analysis, replication analysis and polygenic risk scoring (Supplementary Fig. S1). Little intermediate interaction is required, allowing for efficient standardized analysis of genome-wide data and results. Standardized file naming conventions are designed to optimize overview and analysis record tracking within large-scale genetic projects. Publication-ready data visualizations and reports (in PDF and Excel format) permits easy evaluation of the results. Simulated datasets are also available with the pipeline for training and development purposes. In the ensuing sections, we describe the main components of the pipeline.

2.2 Pre-imputation/QC

The pre-imputation/QC module (Supplementary Section S1) consists of the following general steps (Supplementary Fig. S2):

- Inferring the genotyping chip;
- Standardizing file names and sample identifiers, incorporating chip information and ensuring that sample IDs across distinct cohorts are unique while keeping original sample IDs intact;
- Carrying out technical sample and variant QC procedures: RICOPILI will assign red, yellow and green flags to various QC parameters to help with the decision if a cohort needs further work before going into the following modules (Supplementary Fig. S1.1).

Detailed sample and variant filtering reports provide diagnostics to identify possible QC issues and solutions. Quality controlled datasets are saved separately for downstream analysis.

2.3 Principal components analysis

The PCA module (Supplementary Section S2) fulfils two objectives (Supplementary Fig. S3):

- Identify and remove duplicated or related samples for case-control and trio cohorts;
- Assess ancestral outliers and population stratification with EIGENSTRAT (Price *et al.*, 2006);
- Principal component scores are computed and could be utilized for visualization or as covariates to adjust for population structure in downstream post-imputation GWAS.

2.4 Imputation

RICOPILI automates computationally costly genotype imputation with an optimized routine for high-performance computing (HPC) environments (Supplementary Section S3 and Fig. S4). This module aligns genotype data to the imputation reference, pre-phases haplotypes and executes imputation. Users have the option to:

- Impute genotypes to the 1000 Genomes (1000 Genomes Project Consortium et al., 2015) or Haplotype Reference Consortium panel (McCarthy *et al.*, 2016);
- Perform pre-phasing with Eagle (Loh *et al.*, 2016) or SHAPEIT (Delaneau et al., 2011);
- Perform imputation with IMPUTE (Bycroft *et al.*, 2018; Howie *et al.*, 2009) or Minimac (Das *et al.*, 2016; Howie *et al.*, 2012).

RICOPILI allows for automated data preparation, alignment and sharing with public imputation servers (https://docs.google.com/document/d/18dupvU4kw11slREc1TUfwQwhO_eI0n_MeKVpwi4HLNA/) [e.g. Michigan (https://imputationserver.sph.umich.edu/index.html#!pages/home), Sanger (https://imputation.sanger.ac.uk/)], and reintegration of the results back into the RICIOPILI data structure. This is especially beneficial if an HPC environment is not accessible, and imputation by third party services has been approved by the user's local Institutional Review Board (IRB). More importantly with larger reference panels, such as the HRC and TopMed imputation panels becoming available but not directly accessible, RICOPILI allows such resources to be utilized.

The imputation output files are a set of genotype probabilities for all markers and 'best-guess' genotype hardcall files filtered on imputation quality and minor allele frequency. Hard call genotypes are available in three levels (hardcall with genotype probability >0.8, otherwise missing): (i) no further filter, (ii) lightly filtered (missingness <0.02) and (iii) filtered with strict criteria (missingness <0.01; MAF >5%).

RICOPILI allows the creation of case-pseudo-controls to handle imputation and association procedures for trios.

2.5 Post-imputation

The post-imputation module (Supplementary Section S4 and Fig. S5) performs association analysis using imputed dosage files, meta-analysis via METAL (Willer *et al.*, 2010), conditional analysis, polygenic risk scoring, LD score regression (Bulik-Sullivan *et al.*, 2015) and replication analysis. Covariates (e.g. age, sex, principal components from PCA) and alternative phenotypes, including quantitative traits may be incorporated within the post-imputation module. Automated 'clumping' of genome-wide significant single nucleotide polymorphisms to facilitate identification of independently associated genetic loci. Publication-ready reports and visualizations such as Manhattan plots, QQ-plots, forest plots, annotated region plots and polygenic risk distributions are generated by the module as well. It is notable that genome-wide summary statistics as well as input statistics for various Manhattan and QQ-plots, as well as clumped summary statistics are automatically made available in the distribution/folder as part of the pipeline. These could then be utilized for downstream and follow-on analysis (https://docs.google.com/document/d/1jiD25BYjPAO-TLRAPkYSspiovn8wiQ29ZmZv9Pe2I2U/) (e.g. GCTA; Yang *et al.*, 2011, Spredixcan; Barbeira *et al.*, 2018 and FUSION; Gusev *et al.*, 2016) for the GWAS results.

2.6 Additional utility modules

RICOPILI allows for additional features and modules (see Supplementary Information). Including, (i) reference builder: builds reference data for genotype imputation from publicly accessible reference panels (Supplementary Fig. S6), (ii) replication of GWAS: using external summary data or those generated by RICOPILI and (iii) polygenic leave-one-out analysis: where each input dataset is used as a hold out and polygenic risk prediction is done iteratively across hold out data. All helper scripts and modules are saved in a *centralized location* specified by the user within a folder called rp_bin/and logging files with *_info suffix are also available.

2.7 Availability of simulated GWAS data

To allow new users to familiarize themselves with RICOPILI and experienced users to develop new functionality for the pipeline, we simulated freely available GWAS data using HAPGEN (Su *et al.*, 2011) (Supplementary Section S6). The dataset comprises 6200 'individuals' across ~600 000 markers based on the Illumina OmniExpress, a widely used genotyping platform. For training and development purposes, population stratification, cross-sample relatedness and technical errors were introduced to the simulated data. The sample is separated into five datasets 'HapGen5' packaged with RICOPILI (https://docs.google.com/document/d/1ux_FbwnvSzaiBVEwgS7eWJoYlnc_o0YHFb07SPQsYjI/). Data description and results are described in further detail in Extended Data Analysis and User Guide.

2.8 Cluster portability

RICOPILI is portable (https://docs.google.com/document/d/14aaoeT5hF541I8hHsDAL_42oyvlHRC5FWR7gir4xco/) to various LINUX-based HPC environments {e.g. BSUB (https://docs.google.com/document/d/1fNFnC3-rBZkmtH47Je_yUfGatB9qhDGi9HtMSA3_MPw/), QSUB (https://docs.google.com/document/d/1oY5IA4a6yG_pmbvWJC8A6MTzjYoGzVlqQ_aXUwWCl8I/), SLURM, GCP [Google Cloud Platform (https://docs.google.com/document/d/115NAaH6c8_C6Gn7D5JTldfW0CMGwUOMhxqd1Sthku-E/)]} (Supplementary Section S7). Support for Docker (https://hub.docker.com/r/bruggerk/ricopili; https://github.com/vtrubets/ricopili_docker) implementation of RICOPILI is also underway. In the absence of an HPC environment, RICOPILI can use the full potential of multi-core machines with parallel optimization. Regular updates and maintenance of the pipeline are carried out to incorporate the latest advances in genetic association methods. Ongoing support includes an active user forum (https://groups.google.com/forum/#!forum/ricopili-user-group), support website (https://sites.google.com/a/broadinstitute.org/ricopili/home) and detailed tutorials written by current RICOPILI analysts (consult footnotes).

2.9 RICOPILI web app

RICOPILI is now usable via browser on a cluster backed by Google Cloud: http://34.74.48.153. Here the user does not need any UNIX knowledge. Naturally the user needs to make sure that IRB allows for uploading genotype data to third party computer environments.

3 Discussion

RICOPILI has supported the analytical capability of the PGC, encompassing over 800 investigators internationally. The consortium is a testament to collaborative science that has unified much of the field and collated data collections, and enabled rapid progress in uncovering the genetic and biological basis of psychiatric disorders. RICOPILI addresses the need for a rapid computational pipeline for GWAS that integrates leading bioinformatics resources and produces publication-ready outputs. The PGC has reported GWAS studies in high-impact publications, most of which featured RICOPILI as the main analysis pipeline—including the seminal report identifying 108 GWAS loci for schizophrenia (Ripke et al., 2014). The pipeline has been adapted across various consortia, with 112 analysts performing rapid computation for GWAS to date. For this reason, we introduce RICOPILI to an audience of principal investigators, academics, analysts and all personnel tasked with determining the common variation underlying complex, heritable diseases and traits.

Acknowledgements

Computing and network infrastructure was provided by various sources, including SURFsara (Genetic Cluster Computer; LISA cluster) and the Stanley Center for Psychiatric Research at the Broad Institute of MIT and Harvard.

Funding

The PGC has received major funding from the US National Institutes of Health (NIH) [U01 MH109528 and U01 MH1095320]; M.L. received funding support from National Medical Research Council, Singapore [NMRC/TCR/003; MH095: 003/008-1014; NMRC/CG/004/2013; NMRC/SEEDFD/019/2017]; S.R. received funding support from NIH/NIMH [U01MH109528

01 (Ripke, PI) 2016-2021 'Psychiatric Genomics Consortium: Finding Actionable Variation'], Broad Institute of MIT and Harvard [691099; 'Psychiatric Genomics Consortium'] and the German Research Foundation [DFG, RI2846/1-1].

Conflict of Interest: none declared.

References

Barbeira,A.N. *et al*. (2018) Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nat. Commun*., **9**, 1825.

Breen,G. *et al*. (2016) Translating genome-wide association findings into new therapeutics for psychiatry. *Nat. Neurosci*., **19**, 1392–1396.

Bulik-Sullivan,B.K. *et al*. (2015) LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet*., **47**, 291–295.

Bycroft,C. *et al*. (2018) The UK Biobank resource with deep phenotyping and genomic data. *Nature*, **562**, 203–209.

Das,S. *et al*. (2016) Next-generation genotype imputation service and methods. *Nat. Genet*., **48**, 1284–1287.

Delaneau,O. *et al*. (2011) A linear complexity phasing method for thousands of genomes. *Nat. Methods*, **9**, 179–181.

Gusev,A. *et al*. (2016) Integrative approaches for large-scale transcriptome--wide association studies. *Nat. Genet*., **48**, 245–252.

Howie,B. *et al*. (2012) Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet*., **44**, 955–959.

Howie,B.N. *et al*. (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet*., **5**, e1000529.

Loh,P.R. *et al*. (2016) Fast and accurate long-range phasing in a UK Biobank cohort. *Nat. Genet*., **48**, 811–816.

McCarthy,S. *et al*. (2016) A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet*., **48**, 1279–1283.

Price,A.L. *et al*. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet*., **38**, 904–909.

Purcell,S. *et al*. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet*., **81**, 559–575.

Ripke,S. *et al*. (2014) Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, **511**, 421–427.

Su,Z. *et al*. (2011) HAPGEN2: simulation of multiple disease SNPs. *Bioinformatics*, **27**, 2304–2305.

Sullivan,P.F. *et al*. (2012) Genetic architectures of psychiatric disorders: the emerging picture and its implications. *Nat. Rev. Genet*., **13**, 537–551.

Sullivan,P.F. *et al*. (2018) Psychiatric genomics: an update and an agenda. *Am. J. Psychiatry*, **175**, 15–27.

Willer,C.J. *et al*. (2010) METAL: fast and efficient meta-analysis of genome-wide association scans. *Bioinformatics*, **26**, 2190–2191.

Yang,J. *et al*. (2011) GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet*., **88**, 76–82.

1000 Genomes Project Consortium *et al*. (2015) A global reference for human genetic variation. *Nature*, **536**, 68–74.