SNPQT: FLEXIBLE, REPRODUCIBLE, AND COMPREHENSIVE
QUALITY CONTROL AND IMPUTATION OF GENOMIC DATA

A PREPRINT

Christina Vasilopoulou
Northern Ireland Centre for Stratified Medicine
Altnagelvin Hospital Campus
Biomedical Sciences Research Institute
Ulster University
Derry/Londonderry, UK, BT47 6SB
vasilopoulou-c@ulster.ac.uk

Benjamin Wingfield
Centre for Personalised Medicine
Altnagelvin Hospital Campus
Biomedical Sciences Research Institute
Ulster University
Derry/Londonderry, UK, BT47 6SB
b.wingfield@ulster.ac.uk

Andrew P. Morris
Centre for Genetics and Genomics Versus Arthritis
Centre for Musculoskeletal Research
Manchester Academic Health Science Centre
University of Manchester
Manchester, M13 9PT, UK
andrew.morris-5@manchester.ac.uk

William Duddy
Northern Ireland Centre for Stratified Medicine
Altnagelvin Hospital Campus
Biomedical Sciences Research Institute
Ulster University
Derry/Londonderry, UK, BT47 6SB
w.duddy@ulster.ac.uk

May 6, 2021

ABSTRACT

Motivation: Quality control of genomic data is an essential but complicated multi-step procedure, often requiring separate installation and expert familiarity with a combination of disparate bioinformatics tools.

Results: To provide an automated solution that retains comprehensive quality checks and flexible workflow architecture, we have developed snpQT, a scalable, stand-alone software pipeline, offering some 36 discrete quality filters or correction steps, with plots before-and-after user-modifiable thresholding. This includes build conversion, population stratification against 1,000 Genomes data, population outlier removal, and built-in imputation with its own pre- and post-quality controls. Common input formats are used and users need not be superusers nor have any prior coding experience. A comprehensive online tutorial and installation guide is provided through to GWAS (https://snpqt.readthedocs.io/en/latest/), introducing snpQT using a synthetic demonstration dataset and a real-world Amyotrophic Lateral Sclerosis SNP-array dataset.

Availability: snpQT is open source and freely available at https://github.com/nebfield/snpQT.
Contact: Vasilopoulou-C@ulster.ac.uk, w.duddy@ulster.ac.uk

Keywords GWAS · Quality Control · GWAS pipeline · Nextflow · Imputation · SNPs · Genomic Variants

1 Introduction

Assuring high quality of genomic data is necessarily a complex multi-step procedure, but it is critical to generate reproducible and reliable results in genome-wide association studies (GWAS). Multiple challenges are encountered in carrying out QC ([1]). Although there are well-established steps and good practices ([2,3]), there is no standardised and universally followed workflow, contributing to low reproducibility of results.
Existing approaches, including semi-automated tools [4], can involve a time-consuming "trial and error" approach, requiring the analyst to check the distributions of parameters in plots produced over many rounds of adjustments, and to manually enter commands in a long list of QC steps one-by-one or in a series of shell scripts. The analyst may encounter incompatibility problems and installation difficulties. Software architecture tools such as nextflow and BioContainers can address these issues and have been proposed as automated solutions [5], but limitations exist in terms of limited and relatively rigid QC analysis, lacking such steps as imputation, limited variety of threshold choice and plot outputs, and the requirement for users to have extensive knowledge of the software in order to tailor their analysis.

2 Methods

snpQT was developed as a set of nine core workflow components implemented with the nextflow workflow management system [6]. Each workflow component consists of independent containerised modules, using BioContainers curated by the bioinformatics community wherever possible [7]. Nextflow allows snpQT to be easily scaled from a laptop to a high-performance computing (HPC) or cloud environment, and enables caching at continuous checkpoints, so users can alter thresholds without needing to rerun earlier parts of the analysis.

All nine workflows are illustrated in Figure 1. Workflow A runs only once, performing a local database set up, downloading and preparing reference files [8, 9] and setting up specific versions of tools using conda or docker. snpQT processes data in human genome build 37, but Workflow B has been created for the user to convert from build 38 to 37 or vice versa. Workflow C performs sample QC, including checks for missing call rate, sex discrepancies, heterozygosity, cryptic relatedness, and missing phenotypes. Workflow D performs population stratification for the automatic removal of samples that are predicted as ethnic outliers (using EIGENSOFT, [10]). Workflow E performs the main Variant QC, checking missing call rate, Hardy-Weinberg equilibrium deviation, minor allele frequency, missingness in case/control status, and generates covariates for GWAS, based on a user-modifiable number of Principal Components (or users may...
provide a covariates file). Workflow F is for pre-imputation quality control, while workflow G performs local phasing and imputation using shapeit4 ([11]) and impute5 ([12]), and workflow H performs post-imputation QC. The workflows structure also allows for users to upload their data to an external imputation server, or use a different reference panel. Workflow I performs GWAS, outputting summary statistics, along with a Manhattan plot and a QQ-plot. Detailed summary logs and graphs are provided throughout, depicting the total number of samples and variants in each step, and prompting users towards the locations of intermediate files and logs.

snpQT is implemented in nextflow, R and bash. As well as those already listed, the following tools are used: picard (https://broadinstitute.github.io/picard/), PLINK ([13]), PLINK2.0 ([14]), samtools ([15]), and snpflip (https://github.com/biocore-ntnu/snpflip). The latest release of 1,000 human genome data ([8]) is used as a reference panel in both VCF and processed PLINK2 formats ([9]). A part of population stratification implementation was inspired by the work of [3].

We demonstrate snpQT using a synthetic dataset which is available with the tool, and an Amyotrophic Lateral Sclerosis SNP-array dataset of 2,000 samples (1,000 cases and 1,000 controls) taken from a restricted-access dbGaP project ([16]), at https://snpqt.readthedocs.io/en/latest/.

3 Conclusion

snpQT offers robust QC combined with scalability, reproducibility, flexibility and user-friendly design which can appeal to a broad spectrum of users. It is stand-alone software that needs neither additional coding nor manual installation/download of any data or other program apart from nextflow and conda or docker. The input is a VCF file and/or binary plink files, formats which are widely used. For users who have limited experience with QC analysis a thorough "how-to" guide and step-by-step tutorials are provided, using the demonstration dataset that is available with the tool.

Acknowledgements

We would like to thank Dr. Priyank Shukla for helpful discussion, Dr. Apostolos Malatras for his assistance, and Peter Timlett for designing the snpQT logo.

Funding

This work was financed by the EU Regional Development Fund EU Sustainable Competitiveness Programme for N. Ireland, NI Public Health Agency (HSC R&D) & Ulster University. C.V. was the recipient of a DfE international scholarship from Ulster University.

References

[1] Christina Vasilopoulou, Andrew P. Morris, George Giannakopoulos, Stephanie Duguez, and William Duddy. What Can Machine Learning Approaches in Genomics Tell Us about the Molecular Basis of Amyotrophic Lateral Sclerosis? *Journal of Personalized Medicine*, 10(4):247, 11 2020.

[2] Carl A Anderson, Fredrik H Pettersson, Geraldine M Clarke, Lon R Cardon, P Morris, and Krina T Zondervan. Data quality control in genetic case-control association studies. *Nature Protocols*, 5(9):1564–1573, 2011.

[3] Andries T. Marees, Hilde de Kluiver, Sven Stringer, Florence Vorspan, Emmanuel Curis, Cynthia Marie-Claire, and Eske M. Derks. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *International Journal of Methods in Psychiatric Research*, 27(2):1–10, 2018.

[4] Ryan J. Eller, Sarath C. Janga, and Susan Walsh. Odyssey: A semi-automated pipeline for phasing, imputation, and analysis of genome-wide genetic data. *BMC Bioinformatics*, 20(1):364, 6 2019.

[5] Zeyuan Song, Anastasia Gurinovich, Anthony Federico, Stefano Monti, and Paola Sebastiani. nf-gwas-pipeline: A Nextflow Genome-Wide Association Study Pipeline. *Journal of Open Source Software*, 6(59):2957, 3 2021.

[6] Paolo DI Tommaso, Maria Chatzou, Evan W. Floden, Pablo Prieto Barja, Emilio Palumbo, and Cedric Notredame. Nextflow enables reproducible computational workflows, 4 2017.

[7] Felipe da Veiga Leprevost, Björn A. Grünning, Saulo Alves Affitos, Hannes L. Röst, Julian Uszkoreit, Harald Barsnes, Marc Vaudel, Pablo Moreno, Laurent Gatto, Jonas Weber, Mingze Bai, Rafael C. Jimenez, Timo Sachsenberg, Julianus Pfeuffer, Roberto Vera Alvarez, Johannes Griss, Alexey I. Nesvizhskii, and Yasset Perez-Riverol.
Fiona Cunningham, Ian Dunham, Kasper Lage, Jakob Berg Jespersen, Heiko Horn, Donghoon Kim, Rob Desalle, Apurva Narechania, Melissa A. Wilson Sayres, Fernando L. Mendez, G. David Poznik, Peter A. Underhill, David Mittelman, Ruby Banerjee, Maria Cerezio, Thomas W. FitzGerald, Sandra Louzada, Andrea Massaia, Fengtang Yang, Divya Kalra, Walker Hale, Xu Dan, Kathleen C. Barnes, Christine Beiswanger, Hongyu Cai, Hongzhi Cao, Brenna Henn, Danielle Jones, Jane S. Kaye, Alastair Kent, Angeliki Kerasidou, Raiska Mathias, Pilar N. Ossorio, Michael Parker, Charles N. Rotimi, Charmaine D. Royal, Karla Sandoval, Yeyang Su, Zhongming Tian, Sarah Tishkoff, Marc Via, Yuhong Wang, Huanming Yang, Ling Yang, Jiayong Zhu, Walter Bodmer, Gabriel Bedoya, Zhiming Cai, Yang Gao, Jayou Chen, Leena Peltonen, Andres Garcia-Montero, Alberto Orfao, Julie Dutil, Juan C. Martinez-Cruzado, Raiska A. Mathias, Anselm Hennis, Harold Watson, Colin McKenzie, Firdausi Qadri, Regina LaRocque, Xiaoyan Deng, Danny Asogun, Onikeke Folarin, Christian Happi, Omonwummi Omoniwa, Matt Stremliad, Ridhi Timariy, Minummatua Sisay Joof, Tomani Corrah, Kirk Rockett, Dominic Kwiatkowski, Jaspal Kooner, Tran Tinh Hien, Sarah J. Dunstan, Nguyen Thuy-Hang, Richard Fonnie, Robert Garry, Lansen S. Kanne, Linda Moses, John Schieffelin, Donald S. Grant, Carla Gallo, Giovanni Poletti, Danish Saleeheen, Asif Rasheed, Lisa D. Brooks, Adam L. Felsenfeld, Jean E. McEwen, Yukaterina Vaydylevich, Audrey Duncanson, Michael Dunn, and Jeffery A. Schloss. A global reference for human genetic variation, 9 2015.

[9] Chang CC. 1000 Genomes phase 3, phased and annotated data for use in plink2.0 worked examples. GigaScience Database, 2018.

[10] Alkes L. Price, Nick J. Patterson, Robert M. Plenge, Michael E. Weinblatt, Nancy A. Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics, 38(8):904–909, 8 2006.

[11] Olivier Delaneau, Jean François Zagury, Matthew R. Robinson, Jonathan L. Marchini, and Emmanouil T. Dermitzakis. Accurate, scalable and integrative haplotype estimation. Nature Communications, 10(1):1–10, 12 2019.

[12] Simone Rubinacci, Olivier Delaneau, and Jonathan Marchini. Genotype imputation using the Positional Burrows Wheeler Transform. PLOS Genetics, 16(11):e1009049, 11 2020.

[13] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A.R. Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I.W. De Bakker, Mark J. Daly, and Pak C. Sham. PLINK: A tool set for whole-genome association and population-based linkage analyses. American Journal of Human Genetics, 81(3):559–575, 2007.

[14] Christopher C Chang, Carson C Chow, Laurent CAM Tellier, Shashaank Vattikuti, Shaun M Purcell, and James J Lee. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience, 4(1):7, 12 2015.

[15] Petr Danecek, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whittham, Thomas Keane, Shane A McCarthy, Robert M Davies, and Heng Li. Twelve years of SAMtools and BCFtools. GigaScience, 10(2):1–4, 2 2021.

[16] Aude Nicolas, Kevin Kenna, Alan E. Renton, Nicola Ticozzi, Faraz Faghi, Ruth Chia, Janice A. Dominov, Brendan J. Kenna, Mike A. Nalls, Pamela Keagle, Alberto M. Rivera, Wouter van Rheenen, Natalie A. Murphy, Joke J.F.A. van Vught, Joshua T. Geiger, Rick van der Spek, Hannah A. Pliner, Shankaracharya, Bradley N. Smith, Giuseppe Marangio, Simon D. Topp, Yevgeniya Abramzon, Athina Sorgia Gkazi, John D. Eicher, Aoife Kenna, Francesco O. Logullo, Isabella Simone, Giancarlo Logroscino, Fabrizio Salvi, Ilaria Bartolomei, Giuseppe Borghero, Maria Rita Murru, Emanuela Costantino, Carla Pani, Roberta Puorro, Carla Ceredda, Valeria Piras, Stefania Tranquilli, Stefania Cuccu, Daniela Corongiu, Maurizio Melis, Antonio Milia, Francesco Marrosu, Maria Giovanni Marrosu, Gianluca Floris, Antonino Cannas, Margherita Capasso, Claudia Caponnetto, Gianluigi Mancardi, Paola Origone, Paola Mandich, Francesca L. Conforti, Sebastiano Cavallaro, Gabriele Mora, Kalliopi Marinou, Riccardo Sideri, Silvana Penco, Lorena Mosca, Christian Lunetta, Giuseppe Lauria Pinter, Massimo Corbo, Nilo Riva, Paola Carrera, Paolo Volanti, Jessica Mandrioli, Nicola Fini, Antonio Fasano, Lucio Tremolizzo, Alessandro Arosio, Carlo Ferrari, Francesca Trojisi, Gioachino Tedeschi, Maria Rosaria Monsurrò, Giovanni Piccirillo, Cinzia Mescino, Anna Tetta, Enzo Ortu, Vincenzo La Bella, Rossella Spataro, Tiziana Colletti, Mario Sabatelli, Marcella Zollino, Amelia Conte, Marco Luigetti, Serena Lattante, Marialuisa Santarelli, Antonio Petrucci, Maura Pugliatti, Angelo Pirisi, Leslie D. Parish, Patrizia Occhineri, Fabio Giannini, Stefania Battistini, Claudia Ricci, Michele Benigni, Tea B. Cau, Daniela Loi, Andrea Calvo, Cristina Moglia, Maura Brunetti, Marco Barberis, Gabriella Restagno, Federico Casale, Giuseppe Marrali, Giuseppe Fuda, Irene Ossola, Stefania Cammarosano, Antonio Canosa, Antonino Iard, Umberto Manera, Maurizio Grassano, Raffaella Tanel, Fabrizio Pisano, Letizia Mazzini, Sonia Messina, Sandra D’Alfonso, Lucia Corrado, Luigi Ferrucci, Matthew B. Harms, David B. Goldstein, Neil A. Shneider, Stephen Gottman, Zachary Simmons, Timothy M. Miller, Siddharthan Chandran, Suvankar Pal, George Manousakis, Stanley Appel, Ericka Simpson, Leo Wang, Robert H. Baloh, Summer Gibson, Richard S. Bedlack, David Lacomis, Dhruv Sareen, Alexander Sherman, Lucie Bruijn, Michelle
Genome-wide Analyses Identify KIF5A as a Novel ALS Gene. Neuron, 97(6):1268–1283, 3 2018.