ViGWAS: automation and visualization for population-scale association studies

Louise Cui (Louise.Cui@csiro.au)

Summary

To conduct Genome-Wide association studies (GWAS) with careful quality control, we developed a Jupyter notebook style automation of quality control (QC) and GWAS visualizations: ViGWAS. Utilizing Hail, it is capable to massively distribute the analysis of variant association from whole genome sequencing data, thereby enabling population-scale cohorts to be processed. Automatically joining sample annotations located on cloud storage, it generates QC parameters, conduct association analysis as well as visualize distributions of analysis results. We considered a comprehensive range of plots summarized in an output report for GWAS QC and extended interactive features for plots for improved readability of graphs with massive data points. ViGWAS greatly simplified the process to conduct a quality GWAS for scientists with little programming knowledge.

Availability

ViGWAS, including a Jupyter notebook and a couple of packages required, can be downloaded and installed from scratch, as a virtual machine or on cloud platform AWS and is freely available at https://github.com/aehrc/VIGWAS.

Introduction

Genome-Wide Association Studies (GWAS) has been applied to reveal linkages between specific genetic variants and human diseases. The big data GWAS involves has brought on the challenges in processing and analysing the data. There have been solutions developed for GWAS to process, manage and conduct statistical analysis on such large-scale data (Zhang et al. 2015) (Privé et al. 2017);(Zhang et al. 2015; Zheng et al. 2012). Hail 0.2 is one of the tools, which enables the parallel processing of massive genomic data, visualizations of the data distributions and generations of quality control (QC) statistics of data on cloud platforms. However, for a successful GWAS, a non-noisy and unbiased dataset input is essential. Efficient quality control process prior to GWAS is yet been explored adequately. Meanwhile, current tools performing GWAS still requires a good knowledge of programming from parsing datasets to well-presented results.

Therefore, our project aims to develop an automated visualization tool that takes in genomic data and sample annotations as input, perform and visualize QC and GWAS, and finally output a constructive report including visualizations for QC from multiple aspects and GWAS

with multiple methods. The project is built as a Jupyter notebook on top of Hail 0.2, automating and extending visualizations of QC and GWAS.

Features

Input parsing

Variant SNP data (in VCF: https://samtools.github.io/hts-specs/VCFv4.2.pdf or PLINK: https://www.cog-genomics.org/plink2/formats format) and sample annotations (in CSV or TSV format) can be read in and joined as matrix table (https://hail.is/docs/0.2/hail.MatrixTable.html) in Hail 0.2. Variant data can be read in as multiple vertically or horizontally continuing files.
Some compulsory header fields and value types are required for sample annotations and it must include all samples in the variant data.

Quality Control (QC)

1. Visualized QC on Sampling:

Quality Control (QC) on samples is conducted by visualizing the distribution of samples by different attributes. With user's input of interesting attributes, bar graphs exhibiting sample distributions by each attribute are plotted on the result page. Furthermore, the notebook supports complex visualization of sample distributions, allowing users to analyse sample distributions of an attribute in cohorts.

2. Visualized QC on samples and variants:

Using Hail 0.2 built-in functions, QC statistics of variant/sample tables are generated, annotated and visualized as histograms on the result page. The data to be plotted can be QC statistics for variant tables includes allele count, allele frequency, call rate and homo/heterozygote count etc. Sample QC statistics includes call rate, homo/heterozygous calls, insertion/deletion/transition alternate allele count and more. A full list of QC stats can be found from Hail 0.2 documentation (https://hail.is/docs/0.2/methods/index.html).

Genome-Wide Association Studies (GWAS)

1. PCA

Principal Component Analysis (PCA) is conducted with a selected number of factors by the user. The result of PCA is plotted as a scatter graph by case/control with all combinations of components by default. Visualizations of PCA by other attributes are available with the input of an attribute list. A scatter plot of the first two components is generated for each attribute input.

2. Logistic Regression

Components calculated from the last step are included together with gender to perform logistic regression. Three methods: wald, score and LRT are applied (Ma et al. 2013) and the results are plotted as Manhattan plots and quantile-quantile plots for each method respectively.

3. Variant Spark

Annotated variant data is exported and run on VariantSpark (O'Brien et al. 2015). The output of Variant-Spark is formatted and imported after the run. Association analysis by Variant-Spark is plotted as the Manhattan plot and presented on the result page.

Output Report

All the visualizations are summarized and presented on a series of linked local HTML pages. All plots are interactive and can be zoomed in and zoomed out. Hovering fields are enabled for particular types of plots to assist users for quick information reference of significant candidate variants. Downsample option for histograms are also available for quicker plotting and display.
Exportation of annotated datasets is available in matrix table, VCF or plink file formats. Visualized plots are stored under a directory grouped by analysis as individual HTML files. Saving plots directly from the result page is also available.

Implementations

The tool was developed as a Jupyter notebook, implemented mainly using Hail 0.2 but also extended with bokeh (http://www.bokeh.pydata.org) for plotting and other several python packages for data processing. Results are provided in HTML and presented as local web pages. The notebook can be run on the cloud, virtual machine or installed locally depending on user requirements and data size.

Conclusion

ViGWAS presents a comprehensive series of QC visualizations on large-scale genomic data, which enables users to identify potential biases in a more complete and efficient way. Meanwhile, it extends plot types and data types to be visualized by Hail, thus allowing a more readable result presentation for GWAS and QC. Furthermore, multiple methods for GWAS included and automated in ViGWAS provide comparable GWAS results, avoiding

systematic errors due to methodology, therefore, producing more accurate conclusions. Overall, ViGWAS only requires users to input variant and sample annotation files, automates all analysis and visualizations and provides a summarized and easy-to-navigate report to users. It bridges the gap between scientists and computer science, enabling their usage of cloud platforms to solve problems with large-scale genomic data.

Acknowledgements

This work was supported by CSIRO.

References

Ma, C., Blackwell, T., Boehnke, M., Scott, L.J. and GoT2D investigators 2013. Recommended joint and meta-analysis strategies for case-control association testing of single low-count variants. *Genetic Epidemiology* 37(6), pp. 539–550.

O'Brien, A.R., Saunders, N.F.W., Guo, Y., Buske, F.A., Scott, R.J. and Bauer, D.C. 2015. VariantSpark: population scale clustering of genotype information. *BMC Genomics* 16, p. 1052.

Privé, F., Aschard, H., Ziyatdinov, A. and Blum, M.G.B. 2017. Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr. *Bioinformatics* 34(16), pp. 2781–2787.

Zhang, Y., Blanton, M. and Almashaqbeh, G. 2015. Secure distributed genome analysis for GWAS and sequence comparison computation. *BMC Medical Informatics and Decision Making* 15 Suppl 5, p. S4.

Zheng, X., Levine, D., Shen, J., Gogarten, S.M., Laurie, C. and Weir, B.S. 2012. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* 28(24), pp. 3326–3328.