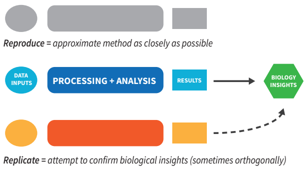
Create a reproducible paper with FireCloud

T. Miller[1], K. Noblett[1], I. Rosenberg[1], M. J. Miossec[2], G. A. Van der Auwera[1]

[1] Data Sciences Platform (DSP), Broad Institute, Cambridge, MA, USA

[2] Universidad Andrés Bello, Center for Bioinformatics and Integrative Biology, Santiago, Chile

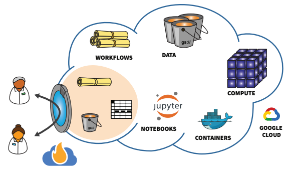


***Reproduce** = approximate method as closely as possible*

***Replicate** = attempt to confirm biological insights (sometimes orthogonally)*

Abstract

The lack of portability and reproducibility of analysis methods limits the effectiveness with which biomedical researchers can benefit from the democratization of genomic analysis. FireCloud is an open-source, freely accessible cloud-based analysis platform developed at the Broad Institute that empowers developers and consumers of analysis methods to overcome these challenges. It bundles data storage, workflow management and interactive Jupyter Notebooks into secure yet readily shareable workspaces. We demonstrated how this system can be used to reproduce someone else's analysis and make your own research more reproducible.



FireCloud empowers researchers to store data and run analyses as workflows and in Jupyter notebooks, then share all materials and results with others to make their research fully reproducible.

Reproducing someone else's research: A case study

Objective

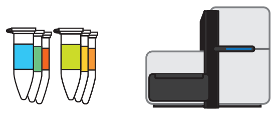
We set out to reproduce the work described by Matthieu Miossec and collaborators in a biorXiv preprint titled "Deleterious genetic variants in *NOTCH1* are a major contributor to the incidence of non-syndromic Tetralogy of Fallot" (ToF). The authors analyzed high-throughput exome sequence data from 867 cases and 1252 controls, identifying 49 deleterious variants within the *NOTCH1* gene that appeared associated with this congenital heart disease. Others had previously identified *NOTCH1* in families with congenital heart defects, including ToF; however the work by Miossec *et al.* is the first to scale variant analysis of ToF to nearly a thousand case samples and show that *NOTCH1* is a significant contributor to ToF risk.

Preprint URL: https://www.biorxiv.org/content/early/2018/04/13/300905

Overall approach

We used the information provided in the preprint and its Supplemental Materials to reconstruct the main phases of the work, distinguishing **Data Input**, **Processing** and **Analysis** as recommended by Kitzes *et al.* (see section below). For the **Processing** phase, we created a synthetic dataset to get around the lack of appropriate public data to use as input, and applied a variant discovery workflow that we judged equivalent. For the **Analysis** phase, we obtained the original scripts and commands from Dr. Miossec and with his assistance, reimplemented them in two parts: the prediction of variant effects as a workflow in WDL (Workflow Description Language) and the clustering analysis as R code in a Jupyter notebook. We did all the work in the Broad Institute's open-source analysis platform, **FireCloud**.

DATA INPUTS



Exome sequencing

- 829 ToF patients (excl. carriers of known deletion)
- 1252 healthy controls
- Agilent SureSelectXT v4
- Illumina HiSeq2000

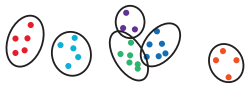
PROCESSING

Mapping & Variant Discovery

- MUGQIC GenPipes DNAseq including:
- Trimmomatic
- BWA 0.6.2 (b37/hg19)
- GATK 3.2 HaplotypeCaller
- QS (QUAL) > 100

Generated synthetic data based on public 1000G Project data
Created case samples by spiking in the *NOTCH1* variants
Joint variant discovery on cases + controls with GATK4

ANALYSIS

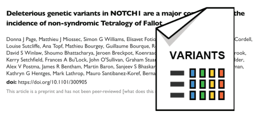


Effect Prediction & Clustering Analysis

- SnpEff + Gemini
- OMIM, GERP, 1000G, ExAC
- MAF ≤ 0.001 in ExAC
- CADD >= 20
- W_d statistic and test

Translated scripts to WDL & R
Same analysis commands
with same data resources

SHARING



Deleterious genetic variants in NOTCH1 are a major co... the incidence of non-syndromic Tetralogy of Fallot

Donna J Page, Matthieu J Miossec, Simon G Williams, Elisavet Fotio... Cordell, Louise Sutcliffe, Ana Topf, Mathieu Bourgey, Guillaume Bourque, R... David S Winlaw, Shoumo Bhattacharya, Jeroen Breckpot, Koenraad... rook, Kerry Setchfield, Frances A Bu'Lock, John O'Sullivan, Graham Stua... lder, Alex V Postma, James R Bentham, Martin Baron, Sanjeev S Bhaskar... man, Kathryn G Hentges, Mark Lathrop, Mauro Santibanez-Koref, Berna...

doi: https://doi.org/10.1101/300905

This article is a preprint and has not been peer-reviewed [what does this

Preprint in biorXiv

- Summary of methods
- Table of 49 NOTCH1 variants
- Pers. communication with author to translate bash and Perl scripts

Public FireCloud workspace
Bundles all data, workflows,
Jupyter notebook and results

Conclusions

We were able to reproduce the original analysis with some compromises: we used synthetic data instead of the original data, and we adapted the processing. In addition, we are still working to address scaling problems that currently prevent us from running the analysis on the full synthetic cohort. Note that this reproduction would not have been possible without direct interaction with one of the authors.

Perspectives

The FireCloud workspace that bundles all materials required for reproducing this work is freely available to all; see https://broad.io/ASHG2018 for access. Given the challenges involved in synthetic data creation step, we plan to develop a more comprehensive dataset and accompanying tooling that will serve as a community resource to empower others to do similar work. We welcome collaborations!

Making your own research reproducible: best practices and their implementation in FireCloud

- Document data sharing policy and who to contact for access
- Create a synthetic dataset or describe how to obtain equivalent data
- Organize raw data inputs in a cloud-based container

- Automate processing as workflows
- Use portable / interoperable tooling
- Bundle original software and dependencies in containers
- Prefer open-access or fee-for-service based hardware over local clusters

- Document analysis design decisions
- Provide scripts for computations and visualizations
- Make manuscripts version-controlled and readily distributable

- Enable easy recomputation with open-source services for sharing data, methods, and papers
- Archive with open-access services that use Digital Object Identifier (DOI)

Best practices recommendations from the book: "The Practice of Reproducible Research" by Justin Kitzes, Daniel Turek and Fatma Deniz

Implementation in FireCloud: An open-source, freely accessible cloud-based analysis platform developed at the Broad Institute (https://www.firecloud.org)

- Workspaces backed by Google storage
- Documentation in the Workspace description and/or Readme.txt in the Google bucket

- Workflows written in WDL
- Connects to container repositories such as DockerHub and Dockstore
- Fee-for-service execution on Google Cloud; can specify exact configuration

- Jupyter notebooks embedded in workspaces for continuity of analysis
- Notebooks can hold both the code and the explanation of what/how/why
- Collaborative manuscript writing

- Sharing the workspace provides all materials necessary to reproduction
- Google buckets have persistent URLs enabling re-use and flexible sharing of resources across projects