Reproducing Population Genomics Analyses

John McCallum[1,2] , Rebekah Frampton[2], Eric Bergueno[2], Tejas Sevak[2], Roy Storey[2]

[1] Biochemistry Dept, University of Otago

[2] Plant and Food Research

Funding

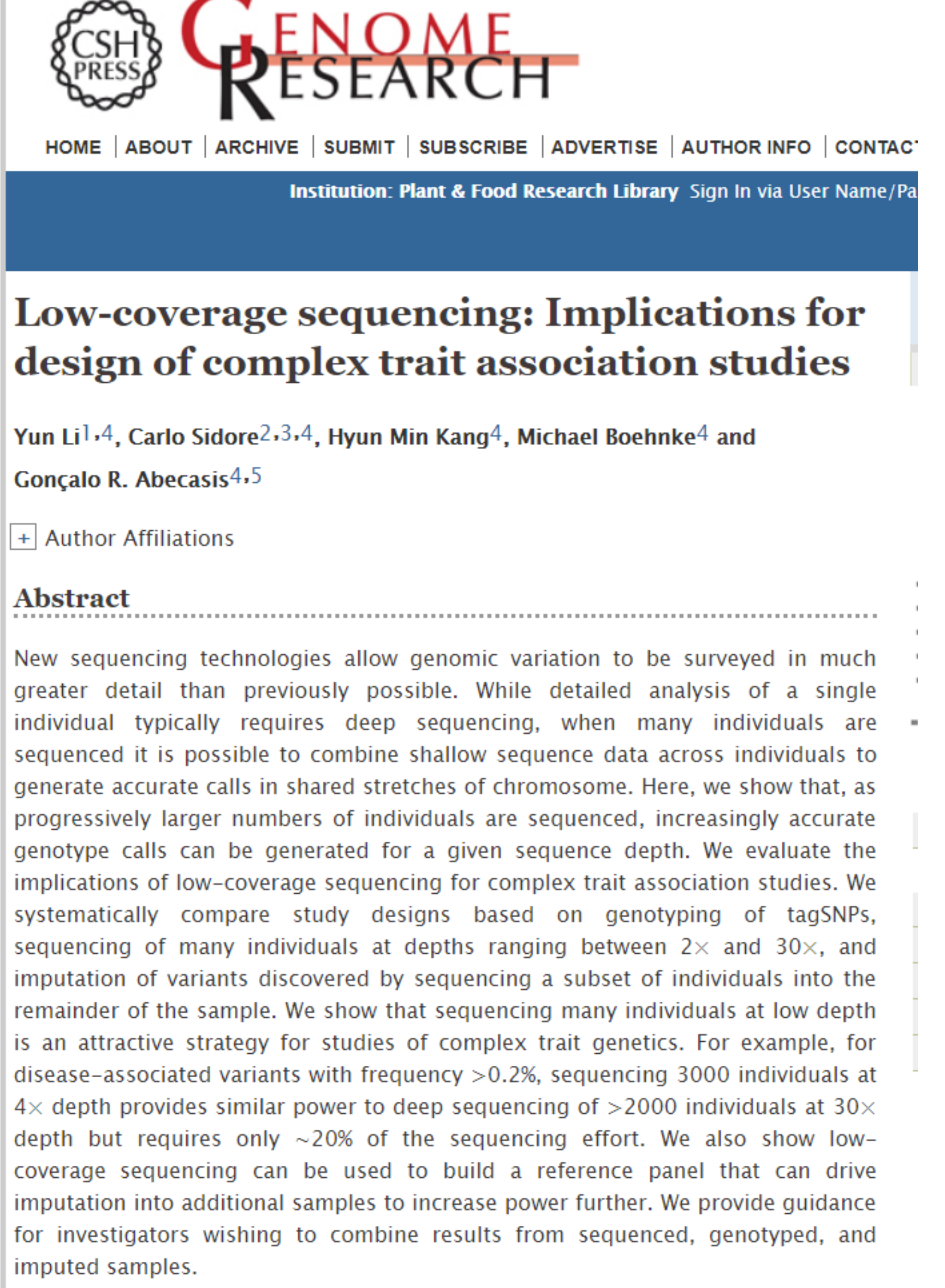
- MBIE Catalyst Fund
- Genomics Aotearoa

REPO

https://github.com/MapNetNZ/Pop-Genomics-Workshop2019

Burning Issues…

- How do we allocate sequencing resources?
- How do we manage sample meta-data?
- How do we carry out QC and analysis?
- How do we minimise software configuration time-wasting?


CSH PRESS
Genome Research
HOME | ABOUT | ARCHIVE | SUBMIT | SUBSCRIBE | ADVERTISE | AUTHOR INFO | CONTAC
Institution: Plant & Food Research Library Sign In via User Name/Pa

Low-coverage sequencing: Implications for design of complex trait association studies

Yun Li[1,4], Carlo Sidore[2,3,4], Hyun Min Kang[4], Michael Boehnke[4] and Gonçalo R. Abecasis[4,5]

+ Author Affiliations

Abstract

New sequencing technologies allow genomic variation to be surveyed in much greater detail than previously possible. While detailed analysis of a single individual typically requires deep sequencing, when many individuals are sequenced it is possible to combine shallow sequence data across individuals to generate accurate calls in shared stretches of chromosome. Here, we show that, as progressively larger numbers of individuals are sequenced, increasingly accurate genotype calls can be generated for a given sequence depth. We evaluate the implications of low-coverage sequencing for complex trait association studies. We systematically compare study designs based on genotyping of tagSNPs, sequencing of many individuals at depths ranging between 2× and 30×, and imputation of variants discovered by sequencing a subset of individuals into the remainder of the sample. We show that sequencing many individuals at low depth is an attractive strategy for studies of complex trait genetics. For example, for disease-associated variants with frequency >0.2%, sequencing 3000 individuals at 4× depth provides similar power to deep sequencing of >2000 individuals at 30× depth but requires only ~20% of the sequencing effort. We also show low-coverage sequencing can be used to build a reference panel that can drive imputation into additional samples to increase power further. We provide guidance for investigators wishing to combine results from sequenced, genotyped, and imputed samples.

Blockers

- HPC requirements
- Linux/HPC know-how
- Limited documentation
- Configuration issues
- Finite time
- **Unicorns** with HPC + stats + genetics + genomics are rare..

Today

- Essentials for Prototyping/documenting/reproducing Linux workflows
  - Git/Gist
  - Jupyter notebook + bash kernel
- Large Data publishing & exchange
  - Zenodo
  - OSF
- Running tools as a non-privileged user on a Linux box or cluster
  - Singularity
  - Conda
- **HOW CAN WE DEVELOP COMMUNITY TRAINING RESOURCES IN THIS SPACE?**

Linux Problem , Jupyter Solution

- Encourages tidy, visible, documented Linux shell usage
- Can set up metadata, processing & exploration in one document
- https://github.com/takluyver/bash_kernel
- Encourages **Explore, Share and Copy Culture**
- https://github.com/MapNetNZ/Pop-Genomics-Workshop2019/blob/master/Using_MAPGD_via_Singularity.ipynb

GISTs: ‘Git-Lite’ Notebook + Data Sharing

- https://gist.github.com/
- Can contain multiple files
- Simple linear versioning
- Can be secret (unguessable hash)
- Can create using a gister application e.g. https://pypi.org/project/gister/

```
 $ cat ~/.gister
# My gister config file
[gister]
public_oauth = 773a5dccf509ad3922c202e48fb674d1XXXXXXXX
```

HOW Do I Install MAPGD?

- Download and Compile https://github.com/LynchLab/MAPGD#-quick-start-
- But… need various dependencies ..far from simple https://github.com/LynchLab/MAPGD/issues/34
- SOLUTION : Singularity https://www.sylabs.io/singularity/
  - Build trusted containerised applications for HPC from recipe files
  - No privilege escalation – can download and run without admin rights!

  - MAPGD recipe https://github.com/powerPlant/mapgd-srf
  - Builds on SingularityHub https://www.singularity-hub.org/collections/2319
  - HOWTO https://github.com/MapNetNZ/Pop-Genomics-Workshop2019/blob/master/Using_MAPGD_via_Singularity.ipynb

HOW do I Install ANGSD (+ Samtools etc etc)?

- http://www.popgen.dk/angsd/index.php/ANGSD
- Conda https://conda.io/en/latest/index.html is a
  - Package management system
  - Environment management system
- Can build reproducible environments without admin rights https://github.com/MapNetNZ/Pop-Genomics-Workshop2019/blob/master/environment.yml
  - e.g. `conda env create -f environment.yml`
- Easy way to install ANGSD from https://anaconda.org/bioconda/angsd

```
conda install -c bioconda angsd
```

- More ANGSD tools & Docs: https://github.com/mfumagalli/ngsTools

HOW Do I share Larger Sets of Data for Publication,Collaboration, Training?

- Zenodo CERN
  - lots of space 50 Gb size limit pre record!
  - REST API for Upload https://gist.github.com/f811ed1793b455a278e35d4e539a049b
  - Download with helper https://github.com/MapNetNZ/Pop-Genomics-Workshop2019/blob/master/DownloadFromZenodo.ipynb
  - Nice author tracking
  - DOI issuing
- OSF https://osf.io/u9qd8/
  - Nice activity tracking
  - Git integration API is a work in progress.. A bit broken
  - Looking good as a pre-publication option
    - https://osf.io/7d968/