Poster: Pangenome-Wide Association Studies with Frequented Regions

Buwani Manuweera
Gianforte School of Computing
Montana State University
Bozeman, MT, USA
buwani.manuweera@student.
montana.edu

Joann Mudge
National Center for Genome
Resources
Santa Fe, NM, USA
jm@ncgr.org

Indika Kahanda
Gianforte School of Computing
Montana State University
indika.kahanda@montana.edu

Brendan Mumey
Gianforte School of Computing
Montana State University
brendan.mumey@montana.edu

Thiruvarangan Ramaraj
National Center for Genome
Resources
tr@ncgr.org

Alan Cleary
National Center for Genome
Resources
acleary@ncgr.org

ABSTRACT

Connecting genetic variation (genotype) to trait variation (phenotype) is a critical but often difficult step in genetic research. A genome-wide association study (GWAS) is a common approach to connect underlying genetic variation to complex phenotypic traits, allowing for phenotypic prediction. GWAS is important in many disciplines, including identifying genetic risk factors for common, complex diseases, identifying genes underlying important traits and predicting phenotypes from genotypes. GWAS is limited, though, in that the types of variations typically studied are single nucleotide polymorphisms (SNPs) identified relative to a single reference genome. These limitations lead to bias and preclude GWAS from studies across related species. The advent of next-generation sequencing has brought an exponential growth in DNA sequence data. This has led to the more comprehensive pangenomics approach, where the entire sequence content and variation of a population are succinctly represented independent of a reference. In prior work, we developed a method for identifying genomic regions that characterize complex variations within pangenomic data and showed that these regions provide a more general way to study genetic variation than existing approaches. This poster describes our initial results to develop new methods for a new branch of genomic analysis called pangenome-wide association studies (PWAS) that generalizes GWAS to pangenomic datasets both within and across species. We make use of recently developed algorithms for fast compressed De Bruijn graph construction and identifying frequented regions in these graphs that can be used as machine-learning features to identify pangenomic regions, overlaid with gene annotations, that relate to complex phenotypic traits. Initial results on a pangenome composed of 100 yeast indicate

that frequented region features provide better machine-learning regression models than SNPs for predicting phenotypic traits.

KEYWORDS

pangenomics; regression; machine-learning

ACKNOWLEDGMENTS

This work was supported by NSF grant DBI-1759522.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the Owner/Author.
ACM-BCB '19, September 7–10, 2019, Niagara Falls, NY, USA
© 2019 Copyright is held by the owner/author(s).
ACM ISBN 978-1-4503-6666-3/19/09
<https://doi.org/10.1145/3307339.3343244>