Scalable Statistical Introgression Mapping Using Approximate Coalescent-Based Inference

Qiqige Wuyun
Department of Computer Science and
Engineering
Michigan State University
East Lansing, Michigan
wuyunqiq@msu.edu

Nicholas W. VanKuren
Department of Ecology and Evolution
University of Chicago
Chicago, Illinois
nvankuren@uchicago.edu

Marcus Kronforst
Department of Ecology and Evolution
University of Chicago
Chicago, Illinois
mkronforst@uchicago.edu

Sean P. Mullen
Department of Biology & Center for
Ecology and Conservation Biology
Boston University
Boston, Massachusetts
smullen@bu.edu

Kevin J. Liu
Department of Computer Science and
Engineering
Michigan State University
East Lansing, Michigan
kjl@msu.edu

ABSTRACT

Recent advances in biomolecular sequencing have revealed the important role that interspecific gene flow has played in genome evolution throughout the Tree of Life. Current and future genomic studies will bring large amounts of genomic sequence to bear upon this topic, and scalable computational methodologies are needed to detect and analyze genomic signatures of interspecific introgression in large-scale datasets.

To address the methodological gap, we introduce a new computational framework known as PHiMM (or “fast PhyloNet + Hidden Markov Model”). PHiMM combines inference and learning under a combined model of genetic drift, substitutions, recombination, and gene flow with a coalescent-based approximation technique. We compare the performance of PHiMM against the state of the art using synthetic and empirical genomic sequence data. We find that PHiMM offers better computational runtime and main memory usage by multiple orders of magnitude, while returning comparable inference accuracy.

An open-source software implementation of the PHiMM framework and open data are publicly available at <https://gitlab.msu.edu/liulab/phimm-dataset>.

CCS CONCEPTS

• **Applied computing** → *Computational genomics; Computational biology; Molecular sequence analysis; Molecular evolution; Computational genomics; Bioinformatics; Population genetics;*

KEYWORDS

introgression, coalescent, phylogenetic network, gene flow, hidden Markov model, non-parametric resampling, mouse, butterfly

ACM Reference Format:

Qiqige Wuyun, Nicholas W. VanKuren, Marcus Kronforst, Sean P. Mullen, and Kevin J. Liu. 2019. Scalable Statistical Introgression Mapping Using Approximate Coalescent-Based Inference. In *ACM-BCB '19: ACM-BCB '19: ACM Conference on Bioinformatics, Computational Biology, and Health Informatics, September 07–10, 2019, Niagara Falls, NY*. ACM, New York, NY, USA, ?? pages. <https://doi.org/10.1145/3307339.3343352>

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/10.1145/3307339.3343352>