POST-NGS: INTERPRETATION AND ANALYSIS OF NEXT GENERATION SEQUENCING DATA FOR BASIC AND TRANSLATIONAL SCIENCE

GURKAN BEBEK
Case Center for Proteomics and Bioinformatics,
Case Western Reserve University 10900 Euclid Ave.
Cleveland, OH 44106-4988, USA
Email: gurkan@case.edu

MEHMET KOYUTÜRK
Department of Electrical Engineering and Computer Science,
Case Western Reserve University 10900 Euclid Ave.
Cleveland, OH 44106-4988, USA
Email: koyuturk@eecs.case.edu

THOMAS LAFRAMBOISE
Department of Genetics,
Case Western Reserve University 10900 Euclid Ave.
Cleveland, OH 44106-4955, USA
Email: thomas.laframboise@case.edu

BENJAMIN J. RAPHAEL
Department of Computer Science,
Brown University, 115 Waterman St.
Providence, RI 20912, USA
Email: braphael@cs.brown.edu

MARK R. CHANCE
Case Center for Proteomics and Bioinformatics,
Case Western Reserve University 10900 Euclid Ave.
Cleveland, OH 44106-4988, USA
Email: mark.chance@case.edu

1. Introduction

Next generation sequencing has dramatically changed our view of what is achievable in genomics. In recent years, research has focused on using next generation sequencing data to characterize genomic content and many methods have been developed for de novo sequence assembly, identification of genomic variants, detection of splice variants etc. Now that the scientific community is equipped with efficient and reliable methods to characterize genomic content, it is natural to expect that the vast amount of information generated by these methods will be further analyzed to seek answers to fundamental biological and medical questions within the context of biological systems. Such questions range from the relationship between genotype and phenotype to regulatory mechanisms of development and principles of evolution.

Recently, large-scale projects such as the Cancer Genome Atlas (TCGA) Project or the 1000 Genomes Project have utilized these technologies extensively, driving remarkable conclusions. However, the methods that have been utilized therein have been specific to these applications and limited in number. Further downstream analyses of NGS data require development of new computational techniques to derive biological knowledge from this vast pool of information. Importantly, there is great need for new methods for integrating these large datasets within the current and emerging research paradigms. From basic science to clinical applications, the –omes that are identified can steer research efforts in transformative directions.

This session will provide a forum for methods and algorithms developed for analysis of finalized next-generation sequencing data. Motivated by the flourishing availability of genome sequences and related data, novel computational methods that interpret these data for research and clinical applications are included in this session. Developing innovative new methodologies and tools for analyzing post-NGS datasets will stimulate more basic and clinical investigation. The aim of this session is to raise awareness of these challenges in the biocomputing community and provide a forum for discussing and disseminating a broad range of computational methods that aim to construct this leap.

2. Session Summary

This session includes an invited talk, three reviewed papers contributed as oral presentations, two contributed papers as posters and a tutorial prepared by the session chairs. The studies presented in this session focus on the development of computational methods to utilize next generation sequencing data to further study diverse biological and translational science problems.

2.1. *Accepted Session Papers*

The following talks will be presented at the Post-NGS session:

- "ChIPModule: Systematic discovery of transcription factors and their cofactors from ChIP-seq data" by Jun Ding, Xiaohui Cai, Ying Wang, Haiyan Hu, and Xiaoman Li.
- "SHPlace: Fast phylogenetic placement using locality-sensitive hashing" by Daniel G. Brown and Jakub Truszkowski.
- "Detecting highly differentiated copy-number variants from pooled population sequencing" by Daniel R. Schrider, David J. Begun, and Matthew W. Hahn.

The following papers will be presented as posters at the symposium:

- "MetaSeq: Privacy preserving meta-analysis of sequencing-based association studies" by Angad Pal Singh, Samreen Zafer, and Itsik Pe'er.
- "Using BioBin to explore rare variant population stratification" by Carrie B. Moore, John R. Wallace, Alex T. Frase, Sarah A. Pendergrass, and Marylyn D. Ritchie

The breadth of research presented in the Post-NGS Session excites us, and we are hopeful that our session will help bring together researchers from various fields and lead to fruitful discussions.

3. Acknowledgments

We would like to thank all the anonymous reviewers who contributed with their invaluable time, and expertise to evaluate submitted papers, and the authors who submitted their work to the Post-NGS Session. This publication was made possible by the Clinical and Translational Science Collaborative of Cleveland, UL1TR000439 from the National Center for Advancing Translational Sciences (NCATS) component of the National Institutes of Health (NIH) and NIH roadmap for Medical Research. MRC, GB and MK would like to mention support from NIH Grant P30-CA043703. MK and MRC are supported in part by NIH grant R01LM011247. TL is supported by NIH Grant R01- CA131341 and American Cancer Society Grant RSG-12-159-01-DMC. MK and TL are supported in part by National Science Foundation (NSF) Award IIS-0916102. BJR is supported by the National Institutes of Health Grant R01-HG5690, a CAREER Award from NSF (CCF-1053753), and Burroughs Wellcome Career Award at the Scientific Interface.