Integrating transcriptomic and proteomic data for accurate assembly and annotation of genomes

Manish Kumar[1,2,14], T. S. Keshava Prasad[1,3,4,14,*], Ajeet Kumar Mohanty[5,6,14], Sreelakshmi K. Sreenivasamurthy[1,2,14], Gourav Dey[1,2,14], Raja Sekhar Nirujogi[1,7], Sneha M. Pinto[1,2,4], Anil K. Madugundu[1,7], Arun H. Patil[1,8], Nirbhay Kumar[9], Photini Sinnis[10], Igor V Sharakhov[11], Charles Wang[12], Harsha Gowda[1,4], Zhijian Tu[11], Ashwani Kumar[5] and Akhilesh Pandey[13]

[1]Institute of Bioinformatics, International Technology Park, Bangalore, Karnataka 560066, India
[2]Manipal University, Madhav Nagar, Manipal, Karnataka 576104, India
[3]Proteomics and Bioinformatics Laboratory, Neurobiology Research Centre, National Institute of Mental Health and Neuro Sciences, Bangalore, Karnataka 560 029, India
[4]YU-IOB Center for Systems Biology and Molecular Medicine, Yenepoya University, Mangalore 575018, India
[5]National Institute of Malaria Research, Field Station, Goa 403001, India
[6]Department of Zoology, Goa University, Taleigao Plateau, Goa 403206, India
[7]Centre for Bioinformatics, Pondicherry University, Puducherry 605014, India
[8]School of Biotechnology, KIIT University, Bhubaneswar, Odisha 751024, India
[9]Department of Tropical Medicine, Tulane University School of Public Health and Tropical Medicine, New Orleans, LA 70112, USA
[10]Malaria Research Institute, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205, USA
[11]Department of Biochemistry, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA
[12] Center for Genomics and Department of Basic Sciences, School of Medicine, Loma Linda University, CA 92350, USA
[13]Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA
[14]These authors contributed equally to this work

Abstract:

The primary goal of whole genome sequencing efforts in any new organism is to provide accurate assembly and annotation of all protein-coding genes in the genome. Complementing genome sequence with deep transcriptome and proteome data could enable more accurate assembly and annotation of newly sequenced genomes. To this end, we provide a proof-of-concept for an analysis pipeline integrating experimental and computational workflows along with expert curation to greatly enhance genome assembly and annotation. We demonstrate a number of advantages of the proposed approach by analyzing the genome of *Anopheles stephensi*, which is one of the most important vectors of the malaria parasite. To achieve broad coverage of genes, we carried out transcriptome sequencing and deep proteome profiling of multiple anatomically distinct sites. Based on transcriptomic data alone, we identified and corrected 535 events of incomplete genome assembly involving 1,196 scaffolds and 868 protein-coding gene models. This approach enabled us to add 401 genes that were missed during genome annotation and identify 1,148 gene correction events through discovery of 173 novel exons, 357 protein extensions, 280 exon extensions, 207 novel protein start sites, 20 novel translational frames, 29 events of joining of exons and 82 events of joining of adjacent genes as a single gene. Incorporating proteomic evidence allowed us to change the designation of over 87 predicted ‘non-coding RNAs’ to conventional mRNAs coded by protein-coding genes. Importantly, extension of the newly corrected genome assemblies and gene models to 15 other newly assembled Anopheline genomes led to the discovery of a large number of apparent discrepancies in assembly and annotation of these genomes. Our data provide a framework for how future genome sequencing efforts should incorporate transcriptomic and proteomic analysis in combination with simultaneous manual curation to achieve near complete assembly and accurate annotation of genomes.

Biography:

Manish Kumar is a Ph.D. student at the Institute of Bioinformatics, Bangalore, India. He has obtained his M.Sc. degree in Biotechnology from Guru Jambheshwar University, India.