Journal of
Data Mining in Genomics & Proteomics

Editorial Open Access

Bioinformatics for High Throughput Sequencing

Heinz-Ulli G Weier*

Department of Cancer & DNA Damage Responses, Life Sciences Division, University of California-LBNL, Berkeley, CA, USA

Over 12 years have passed since the publication of the first rough draft of the sequence of the human genome, and almost 10 years since a complete DNA sequence of the euchromatic part of the human genome was published by a large international consortium [1,2]. While the assembly of the initial maps was costly and labor-intensive, it has been a very important step towards understanding biological functions and interactions, phenotypes, diversity, disease and interactions of an organism with its environment.

In this last decade, ground-breaking technologies such as pyrosequencing, next generation (nexgen) sequencing and Pacific Biosciences' '3rd generation' sequencing [3,4] have not only helped to greatly accelerate the rate of DNA/RNA sequence generation and scientific discoveries, they have also allowed to significantly cut the cost of sequencing and the time to deeply sequence entire genomes or transcriptomes. While these achievements are certainly laudable, they have also created new challenges: with contemporary data sets extending into the range of tens to hundreds of gigabytes per run, data storage, management and interpretation began to face new problems. Other issues relate to quality control (QC) in high throughput sequencing studies.

The good news is that there exists an emerging new generation of researchers with broad training in computer science, informatics, wet lab sciences and computational biology that will allow them to tackle the challenges and solve many of the problems, once the appropriate, and definitely not inexpensive, infrastructures have been put in place [5].

The potential pay-off of these endeavors, investments and efforts could become enormous. To name a few examples, a deeper understanding of the interaction of microbial communities requires multidisciplinary teams of well trained researchers. To understand how microbial communities inhabit and interact with the termite gut environment and how cellulose degrading enzymes are produced and compartmentalized starts with the isolation of cellulose-degrading microorganisms, a metagenomic analysis of enzymes from gut inhabiting microbes and a high resolution analysis of the termite gut environment [6-8]. While microbial communities had been studied just a few years ago using shot-gun sequencing approaches [9], nexgen and 3rd generation sequencing approaches are expected to generate much more information in less time and for a fraction of the price. Results from this type of research may be shifting paradigms in biotechnology/-processing in just a few years.

Human health, susceptibility and acute disease have always been mentioned as drivers of biotechnology-/sequencing-based diagnostic technology developments. Very encouraging results from RNA sequencing, for example, showing recurrent gene fusions and cancer-associated expression of long non-coding RNAs as well as atypical gene splicing in prostate cancer may one day allow to predict the course of the disease [10]. This is exciting new research in the discovery of biomarkers for tumor aggressiveness, metastasis or response to therapy that might direct therapeutic interventions in prostate cancer patients. Other non-coding RNAs including micro-RNAs might also become prognostic markers for disease progression or disease-free survival [11].

This list of promising applications of high throughput sequencing technology could go on for many pages, including the sequencing of plant genome [12] or disease causing viruses, microbes or agents. We prefer to keep this editorial concise and introduce the reader to a collection of cutting-edge articles that describe innovative solutions to today's problems in bioinformatic analysis of high throughput sequencing data.

In 2013, the publishers of the Journal of Data Mining in Genomics and Proteomics (JDMGP) and myself issued a call to the scientific community to consider publishing high-quality, peer-reviewed articles in a Special Issue of JDMGP entitled 'Bioinformatics for High Throughput Sequencing' for streamlined review by their peers and open access publishing. The Open Access publishing model makes these articles available shortly after acceptance, and world-wide readers will not have to pay fees or order a copy through libraries.

The current issue of JDMGP contains an exciting collection of nine research articles from labs that work at the cutting edge of high throughput sequencing.

Briefly, the first article be Roy et al. [13] describes the analysis of small RNA libraries for the discovery of novel Citrus Leprosus Virus cytoplasmic type 2 by nexgen sequencing using small RNA libraries and bioinformatics analysis [14]. The following article entitled 'Computational Approach for MicroRNA Identification in Plants: Combining Genome-Based Predictions with RNA-Seq Data' by J.S. Oliveira and colleagues describes the test and validation of a single genome microRNA finding tool for the analysis of Eucalyptus spp., for which no microRNAs had been described before.

The paper 'Bioinformatics in High Throughput Sequencing: Application in Evolving Genetic Diseases' by M.M.S. Al-Haggar et al. [15] provides a broad overview over previous and ongoing large scale sequencing projects and describes specific examples of successful translation of sequencing results in the treatment of human diseases.

The following two papers focus on algorithm development. A contribution by G. Natsoulis and colleagues describes a novel two-step algorithm for the 'identification of insertion deletion mutations from deep targeted resequencing' [16], while I.Y. Zhbannikov and coauthors present 'SlopMap: A Software Application Tool for Quick and Flexible Identification of Similar Sequences Using Exact K-Mer Matching' in Roche 454- and Illumina-generated data sets [17].

***Corresponding author:** Heinz-Ulli G Weier, Department of Cancer & DNA Damage Responses, Life Sciences Division, E.O. Lawrence Berkeley National Lab, 1 Cyclotron Road, MS 977, Berkeley, CA 94720, USA; Tel: 001-510-486-5363; Fax: 001-510-486-5343; E-mail: ugweier@lbl.gov

Received October 04, 2013; **Accepted** October 07, 2013; **Published** October 07, 2013

Citation: Weier HUG (2013) Bioinformatics for High Throughput Sequencing. J Data Mining Genomics Proteomics 4: e108. doi:10.4172/2153-0602.1000e108

Copyright: © 2013 Weier HUG. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Alignment of nucleic acid and protein sequences is the subject of the next two papers. Davit Bzhalava and J. Dillner discuss 'Bioinformatics for Viral Metagenomics' [18], and the paper 'mBLAST: Keeping up with the Sequencing Explosion for (Meta) Genome Analysis' by Davis et al. [19] introduces a novel search algorithm for large datasets based on the Basic Local Alignment Search Tool (BLAST).

With 3rd generation sequencing technology becoming available to the research community, the article by X. Jiao et al. entitled 'A Benchmark Study on Error Assessment and Quality Control of CCS Reads Derived from the PacBio RS' [20] is considered very timely. Finally, the fastest sequencing tools will be underutilized, if bottlenecks continue to exist in the pipeline of template generation and processing. E. Avsar-Ban and colleagues present a 'High-Throughput Injection System for Zebrafish Fertilized Eggs' intended of overcome problems in the use of zebrafish as a vertebrate model system [21].

The present nine articles describe mostly the research focus of their teams of authors. Present efforts at the publishing house are underway to publish a further volume with additional contributions on 'Bioinformatics for High Throughput Sequencing' before the end of the year. Please see the JDMGP's Special Issue web site for the timeline and further information.

Acknowledgement

The skillful assistance of staff of the Weier laboratory, LBNL, is gratefully acknowledged. This work was supported in parts by NIH grants CA136685 and CA168345 carried out at the Earnest Orlando Lawrence Berkeley National Laboratory under contract DE-AC02-05CH11231.

Disclaimer

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof, or The Regents of the University of California.

Conflict of Interest

The author declares no conflict of interest.

References

1. International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. Nature409: 860-921.
2. Linton, Birren B, Nusbaum C, Michael CZ, Jennifer B et al., (2004) International Human Genome Sequencing Consortium (IHGSC) Finishing the euchromatic sequence of the human genome. Nature431: 931-945.
3. Fernandes F, da Fonseca PG, Russo LM, Oliveira AL, Freitas AT (2011) Efficient alignment of pyrosequencing reads for re-sequencing applications. BMC Bioinformatics 12: 163.
4. Okoniewski MJ, Meienberg J, Patrignani A, Szabelska A, Matyas G, et al. (2013) Precise Breakpoint localization of large genomic deletions using PacBio and Illumina next-generation sequencers. Biotechniques 54: 98-100.
5. Webb S (2011) A decade after the genome, bioinformatics comes of age. BioTechniques 51: 157-161.
6. McDonald JE, Rooks DJ, McCarthy AJ (2012) Methods for the isolation of cellulose-degrading microorganisms. Methods Enzymol 510: 349-374.
7. Nimchua T, Thongaram T, Uengwetwanit T, Pongpattanakitshote S, Eurwilaichitr L (2012) Metagenomic analysis of novel lignocellulose-degrading enzymes from higher termite guts inhabiting microbes. J Microbiol Biotechnol 22: 462-469.
8. Köhler T, Dietrich C, Scheffrahn RH, Brune A (2012) High-resolution analysis of gut environment and bacterial microbiota reveals functional compartmentation of the gut in wood-feeding higher termites (Nasutitermes spp.). Appl Environ Microbiol 78: 4691-4701.
9. Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, et al., (2005) Comparative metagenomics of microbial communities. Science 308: 554-557.
10. Ren S, Peng Z, Mao JH, Yu Y, Yin C, et al. (2012) RNA-seq analysis of prostate cancer in the Chinese population identifies recurrent gene fusions, cancer-associated long noncoding RNAs and aberrant alternative splicings. Cell Res 22: 806-821.
11. Weier HUG (2012) Rags to Riches: Re-viewing the Tumor-specific Expression of Long Noncoding DNA and Satellite DNA Repeats in Humans in the Era of Next Generation Sequencing. J Data Mining in Genomics Proteomics 3: e101.
12. He N, Zhang C, Qi X, Zhao S, Tao Y, et al. (2013) Draft genome sequence of the mulberry tree Morus notabilis. Nat Commun 4: 2445.
13. Roy A, Shao J, Hartung JS, Schneider W, Brlansky RH (2013) A Case Study on Discovery of Novel Citrus Leprosis Virus Cytoplasmic Type 2 Utilizing Small RNA Libraries by Next Generation Sequencing and Bioinformatic Analyses. J Data Mining Genomics Proteomics 4: 129.
14. Oliveira JS, Mendes ND, Carocha V, Graça C, Paiva JA, et al. (2013) A Computational Approach for MicroRNA Identification in Plants: Combining Genome-Based Predictions with RNA-Seq Data. J Data Mining Genomics Proteomics 4: 130.
15. Al-Haggar MMS, Khair-Allaha BA, Islam MM, Mohamed ASA (2013) Bioinformatics in High Throughput Sequencing: Application in Evolving Genetic Diseases. J Data Mining Genomics Proteomics 4: 131.
16. Natsoulis G, Zhang N, Welch K, Bell J, Ji HP (2013) Identification of Insertion Deletion Mutations from Deep Targeted Resequencing. J Data Mining Genomics Proteomics 4: 132.
17. Zhbannikov IY, Hunter SS, Settles ML, Foster JA (2013) SlopMap: A Software Application Tool for Quick and Flexible Identification of Similar Sequences Using Exact K-Mer Matching. J Data Mining Genomics Proteomics 4:133.
18. Bzhalava D, Dillner J (2013) Bioinformatics for Viral Metagenomics. J Data Mining Genomics Proteomics 4: 134.
19. Davis C, Kota K, Baldhandapani V, Gong W, Abubucker S, et al. (2013) Mblast: Keeping up with the Sequencing Explosion for (Meta) Genome Analysis. J Data Mining Genomics Proteomics 4: 135.
20. Jiao X, Zheng X, Ma L, Kutty G, Gogineni E, et al. (2013) A Benchmark Study on Error Assessment and Quality Control of CCS Reads Derived from the PacBio RS. J Data Mining Genomics Proteomics 4: 136.
21. Avsar-Ban E, Miyake H, Obata M, Hashimoto M, Tamaru Y (2013) High-Throughput Injection System for Zebrafish Fertilized Eggs. J Data Mining Genomics Proteomics 4: 137.