Published in final edited form as:

Proteomics. 2015 August ; 15(15): 2557–2559. doi:10.1002/pmic.201500253.

Integration of RNA-seq and Proteomics Data with Genomics for Improved Genome Annotation in Apicomplexan Parasites

Natalie C. Silmon de Monerri² and Louis M. Weiss^{1,2,*}

¹Department of Medicine, Albert Einstein College of Medicine, 1300 Morris Park Avenue, Room 504 Forchheimer, Bronx, NY, 10461

²Department of Pathology, Albert Einstein College of Medicine, 1300 Morris Park Avenue, Room 504 Forchheimer, Bronx, NY, 10461

Abstract

While high quality genomic sequence data is available for many pathogenic organisms, the corresponding gene annotations are often plagued with inaccuracies that can hinder research that utilizes such genomic data. Experimental validation of gene models is clearly crucial in improving such gene annotations; the field of proteogenomics is an emerging area of research wherein proteomic data is applied to testing and improving genetic models. Krishna *et al* [Proteomics 2015, 00, 1–11] investigated whether incorporation of RNA-seq data into proteogenomics analyses can contribute significantly to validation studies of genome annotation, in two important parasitic organisms *Toxoplasma gondii* and *Neospora caninum*. They applied a systematic approach to combine new and previously published proteomics data from *T. gondii* and *N. caninum* with transcriptomics data, leading to substantially improved gene models for these organisms. This study illustrates the importance of incorporating experimental data from both proteomics and RNA-seq studies into routine genome annotation protocols.

Keywords

Proteogenomics; Protozoa; Toxoplasma; Neospora; Proteomics; Apicomplexa

In this issue, Krishna *et al* [1] present an integrated analysis of RNA-seq and mass spectrometry data to improve genome annotation, in the closely related protozoan parasites *Toxoplasma gondii* and *Neospora caninum*. They build new RNA-seq-derived genetic models and provide the first global proteomic dataset for *N. caninum*. The resulting genome annotations cover 35% and 17% of the *T. gondii* and *N. caninum* predicted proteomes, respectively. Furthermore, this analysis led to the identification of a significant number of novel protein-coding genes, which are absent from current annotations.

The genomes of many important human pathogens have now been sequenced, providing an essential resource for research. For most of these organisms, however, there is a very limited understanding of their encoded transcripts and proteins. Accurate genome annotations are

*corresponding author: louis.weiss@einstein.yu.edu.

essential for many approaches, particularly genetic studies or for constructing databases for proteomics. Two *ab initio* prediction programs for eukaryotic genes are widely used to annotate genomic sequence: TigrScan and GlimmerHMM [2]. While tools such as these are essential for prediction of gene models, inaccuracies in these models are common and create significant issues for global proteomic studies of many organisms. Gene models can have an incorrect or missing start site, wrong intron or exon boundaries, or a novel gene may not even be predicted by such approaches. Furthermore, information on alternative splicing is lacking from most annotations. How accurate genome annotations are is unclear, and varies from organism to organism. Proteogenomics, the integration of proteomic, transcriptomic and genomics data, can be a powerful approach to improving genome annotation and identifying novel genes.

The first efforts to sequence the *T. gondii* genome were performed by Shotgun sequencing and EST assembly [3]. Strains representing the main lineages of *T. gondii* have been sequenced providing critically important data for understanding the biology of this ubiquitous pathogen. The most recent annotations of the *T. gondii* and *N. caninum* genome [4] are maintained by ToxoDB.org as part of the Eukaryotic Pathogen Database Resource Center (EuPathDB) [5], an important resource for the Apicomplexa community. Over 8000 genes are currently annotated in the draft *T. gondii* genome, which were originally annotated using conventional computational algorithms (including TigrScan, Twinscan and GlimmerHMM) [3,6]. While such tools have been useful for predicting *T. gondii* genes, the algorithms on which they are based result in the prediction of different gene models, which has led to uncertainty about the accuracy of these predictions [7]. By comparing *T. gondii* gene annotations generated from TigrScan and GlimmerHMM with proteomics and EST data, Dybas *et al* calculated a false negative rate of these gene models of up to 41% [8], illustrating the problems inherent in gene annotation based on the analytical programs available at that time of publication of this paper.

Gene models can be significantly improved by combining experimental data with existing annotations. *T. gondii* genetic models are continuously reassessed by semi-automated reannotation using experimental data, or manual curation [3,6]. Proteomics has played an important role in shaping the current *T. gondii* genome annotations, amounting to at least 68% coverage of the predicted proteome [9]. Proteomic data can be used to validate gene annotations, and is also resource for new open reading frames and novel proteins. A global proteomic study of *T. gondii* tachyzoites, performed by Xia *et al* [10], provided coverage of 27% of the predicted proteome, and was the first study to use mass spectrometry data to validate genetic models in *T. gondii*. Integration of this data with EST information led to the validation of 91% of the proteins in the proteome, arguing that transcriptomic data can be used to validate proteomic datasets. A subsequent study by Che *et al* [11] that employed three proteomic strategies (LC-MS/MS, TLSGE MudPIT, and BDAP LC-MSMS) identified 2241 *T. gondii* proteins that were classified into 841 protein clusters. For analysis, they employed a hypothetical *T. gondii* proteome based on a combination of computationally predicted proteins from TigrScan, TwinScan, GlimmerHMM, Release 6.0 ToxoDB and the available experimental *T. gondii* sequences from the NCBI nonredundant protein database,

confirming that the experimental proteomic data identified valid predictions that were unique to each computational model.

Next generation sequencing has exponentially improved the quality of transcriptional information that can be obtained, and can provide information on alternative splicing, intron/exon boundaries and lead to identification of novel transcripts. In their study, Krishna *et al* queried mass spectrometry data against RNA-seq derived gene models for *T. gondii* and *N. caninum* [1], leading to the identification of loci that were not present in current genome annotations, indicating that RNA-seq is a valuable tool for validation of genome annotation models. Furthermore, Krishna *et al* introduced an RNA-seq compliant version of CRAIG, a tool for gene model generation, which is an alternative to TigrScan and other widely used genome annotation algorithms [12]. Studies based on sequencing, such as these, are beginning to play an important role in genome annotation in *T. gondii*. RNA-seq has been used to perform *de novo* assembly in ME49 strain [13], which led to the identification of over 2000 transcripts that did not correspond to any previously annotated gene. In addition, this provided information on alternative splicing, which was previously uncharacterised in *T. gondii*. TSS-seq has also been used to profile the 5'UTR regions of genes and determine transcriptional start site locations [14]. In addition, a recent study used strand-specific RNA-seq to explore untranslated regions in *T. gondii* and *N. caninum* [15], which led to the identification of putative antisense transcripts and long noncoding RNAs. With RNA-seq data on different *T. gondii* life cycle stages and in various strains now available [16, 17], it may be possible to further increase coverage of genome annotations and mine these datasets for novel transcripts that are stage-and/or strain-specific.

Considering the value proteomics and next generation sequencing information has had in gene annotations in *T. gondii*, Krishna *et al* have combined both types of data for a more powerful proteogenomics analyses [1]. This resulted in a significant improvement in the genome annotation, as demonstrated by this study, providing the greatest coverage of the predicted proteome compared to previous studies on these organisms that used other models. The genomes of many Apicomplexan parasites and other pathogens have been sequenced, providing an important resource for researches; however, the annotations are far from complete. Combinatorial approaches such as those used by Krishna *et al* can provide important validation of computational predictions of annotations. Incorporation of proteogenomics into genome annotation pipelines as standard practice is likely to be hugely useful in the generation of accurate gene models in the future. To this end, the current version of ToxoDB.org is actively curated, employing heuristic gene prediction methods incorporating experimental data sets such as proteomics and transcriptomic data to improve gene annotations. This represents a significant model for gene annotation and illustrates the importance of maintaining active curation efforts to improve and maintain the utility of these critical scientific community resources.

Acknowledgements

Supported by NIH/NIAID R01AI93220 (Weiss).

References

1. Krishna R, Xia D, Sanderson S, Shanmugasundram A, Vermont S, Bernal A, Daniel-Naguib G, Ghali F, Brunk BP, Roos DS, Wastling JM, Jones AR. A large-scale proteogenomics study of apicomplexan pathogens-Toxoplasma gondii and Neospora caninum. *Proteomics*. 2015
2. Majoros WH, Pertea M, Salzberg SL. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics*. 2004; 20:2878–2879. [PubMed: 15145805]
3. Kissinger JC, Gajria B, Li L, Paulsen IT, Roos DS. ToxoDB: accessing the Toxoplasma gondii genome. *Nucleic Acids Res*. 2003; 31:234–236. [PubMed: 12519989]
4. Reid AJ, Vermont SJ, Cotton JA, Harris D, Hill-Cawthorne GA, Konen-Waisman S, Latham SM, Mourier T, Norton R, Quail MA, Sanders M, Shanmugam D, Sohal A, Wasmuth JD, Brunk B, Grigg ME, Howard JC, Parkinson J, Roos DS, Trees AJ, Berriman M, Pain A, Wastling JM. Comparative genomics of the apicomplexan parasites Toxoplasma gondii and Neospora caninum: Coccidia differing in host range and transmission strategy. *PLoS Pathog*. 2012; 8:e1002567. [PubMed: 22457617]
5. Aurrecochea C, Barreto A, Brestelli J, Brunk BP, Cade S, Doherty R, Fischer S, Gajria B, Gao X, Gingle A, Grant G, Harb OS, Heiges M, Hu S, Iodice J, Kissinger JC, Kraemer ET, Li W, Pinney DF, Pitts B, Roos DS, Srinivasamoorthy G, Stoeckert CJ Jr, Wang H, Warrenfeltz S. EuPathDB: the eukaryotic pathogen database. *Nucleic Acids Res*. 2013; 41:D684–D691. [PubMed: 23175615]
6. Gajria B, Bahl A, Brestelli J, Dommer J, Fischer S, Gao X, Heiges M, Iodice J, Kissinger JC, Mackey AJ, Pinney DF, Roos DS, Stoeckert CJ Jr, Wang H, Brunk BP. ToxoDB: an integrated Toxoplasma gondii database resource. *Nucleic Acids Res*. 2008; 36:D553–D556. [PubMed: 18003657]
7. Wakaguri H, Suzuki Y, Sasaki M, Sugano S, Watanabe J. Inconsistencies of genome annotations in apicomplexan parasites revealed by 5'-end-one-pass and fulllength sequences of oligo-capped cDNAs. *BMC Genomics*. 2009; 10:312. [PubMed: 19602295]
8. Dybas JM, Madrid-Aliste CJ, Che FY, Nieves E, Rykunov D, Angeletti RH, Weiss LM, Kim K, Fiser A. Computational analysis and experimental validation of gene predictions in Toxoplasma gondii. *PLoS One*. 2008; 3:e3899. [PubMed: 19065262]
9. Wastling JM, Armstrong SD, Krishna R, Xia D. Parasites, proteomes and systems: has Descartes' clock run out of time? *Parasitology*. 2012; 139:1103–1118. [PubMed: 22828391]
10. Xia D, Sanderson SJ, Jones AR, Prieto JH, Yates JR, Bromley E, Tomley FM, Lal K, Sinden RE, Brunk BP, Roos DS, Wastling JM. The proteome of Toxoplasma gondii: integration with the genome provides novel insights into gene expression and annotation. *Genome Biol*. 2008; 9:R116. [PubMed: 18644147]
11. Che FY, Madrid-Aliste C, Burd B, Zhang H, Nieves E, Kim K, Fiser A, Angeletti RH, Weiss LM. Comprehensive proteomic analysis of membrane proteins in *Toxoplasma gondii*. *Mol Cell Proteomics*. 2010; 10 M110.000745.
12. Bernal A, Crammer K, Hatzigeorgiou A, Pereira F. Global discriminative learning for higher-accuracy computational gene prediction. *PLoS Comput Biol*. 2007; 3:e54. [PubMed: 17367206]
13. Hassan MA, Melo MB, Haas B, Jensen KD, Saeij JP. De novo reconstruction of the Toxoplasma gondii transcriptome improves on the current genome annotation and reveals alternatively spliced transcripts and putative long non-coding RNAs. *BMC Genomics*. 2012; 13:696. [PubMed: 23231500]
14. Yamagishi J, Watanabe J, Goo YK, Masatani T, Suzuki Y, Xuan X. Characterization of Toxoplasma gondii 5' UTR with encyclopedic TSS information. *J Parasitol*. 2012; 98:445–447. [PubMed: 22010783]
15. Ramaprasad A, Mourier T, Naeem R, Malas TB, Moussa E, Panigrahi A, Vermont SJ, Otto TD, Wastling J, Pain A. Comprehensive Evaluation of Toxoplasma gondii VEG and Neospora caninum LIV Genomes with Tachyzoite Stage Transcriptome and Proteome Defines Novel Transcript Features. *PLoS One*. 2015; 10:e0124473. [PubMed: 25875305]
16. Croken MM, Ma Y, Markillie LM, Taylor RC, Orr G, Weiss LM, Kim K. Distinct strains of Toxoplasma gondii feature divergent transcriptomes regardless of developmental stage. *PLoS One*. 2014; 9:e111297. [PubMed: 25393307]

17. Hehl AB, Basso WU, Lippuner C, Ramakrishnan C, Okoniewski M, Walker RA, Grigg ME, Smith NC, Deplazes P. Asexual expansion of *Toxoplasma gondii* merozoites is distinct from tachyzoites and entails expression of non-overlapping gene families to attach, invade, and replicate within feline enterocytes. *BMC Genomics*. 2015; 16:66. [PubMed: 25757795]