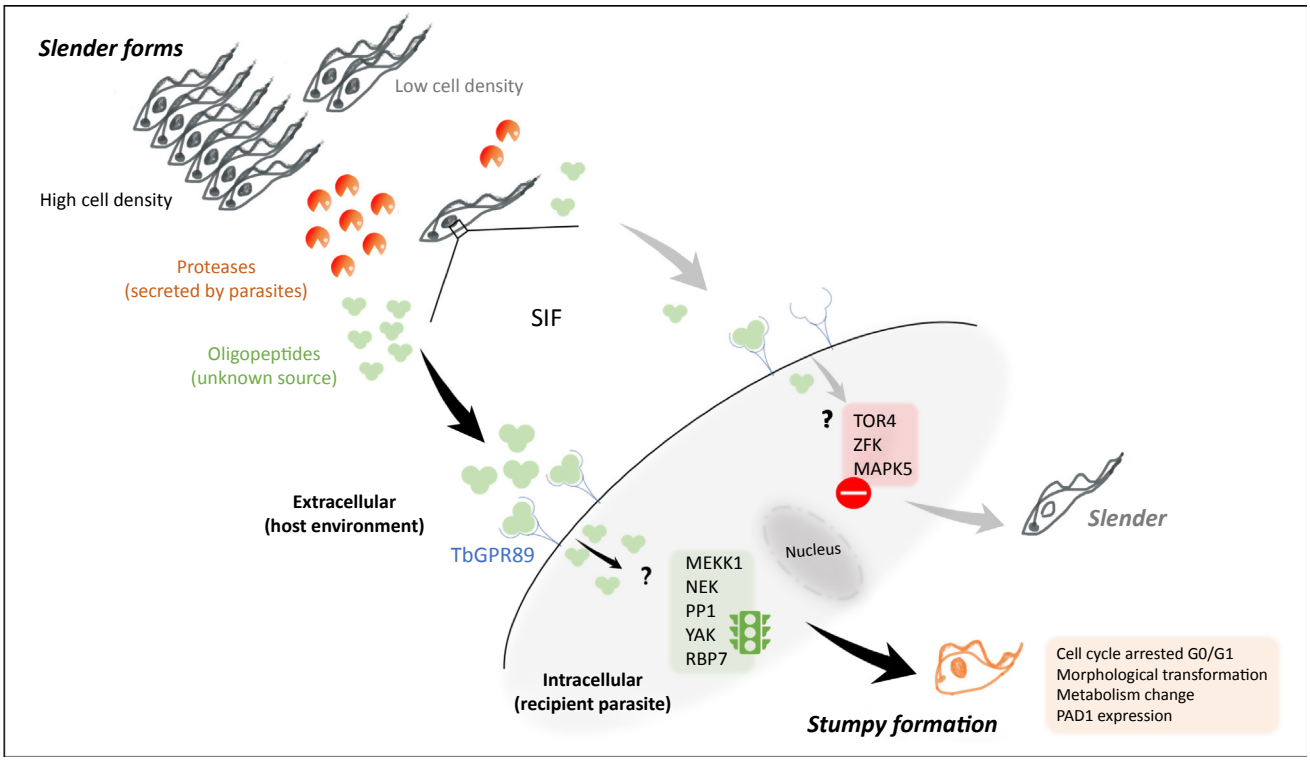

Figure 1. A Putative Model of Stumpy Formation Based on Recent Findings. Parasite density correlates with secreted protease activity, and hence the levels of oligopeptides in the extracellular environment of the parasite. These oligopeptides, the putative stumpy induction factor (SIF), are sensed and internalized by recipient parasites through the putative SIF receptor, TbGPR89. At high parasitaemia, accumulation of oligopeptides activates signalling pathways (e.g., MEEK1, NEK1, PP1, YAK, RBP7 . . .) that lead to stumpy formation. At lower cell density, fewer oligopeptides are produced and internalized, thus activating an alternative pathway (e.g., TOR4, ZFK, MAPK5 . . .) that represses stumpy formation and keeps the cell in a replicative state, the slender form.

development, as TbGRP89 is essential for both parasite survival and stumpy formation and a TbGRP89-targeting drug could potentially reduce parasite burden and at the same time block transmission.

[1]Wellcome Centre for Molecular Parasitology, Institute of Infection, Immunity and Inflammation, University of Glasgow, Glasgow, UK
[2]Department of Immunology and Infectious Disease, Harvard T.H. Chan School of Public Health, Boston, MA, USA

*Correspondence:
Matthias.Marti@glasgow.ac.uk (M. Marti).

https://doi.org/10.1016/j.pt.2018.11.009

References

1. Reuner, B. *et al.* (1997) Cell density triggers slender to stumpy differentiation of *Trypanosoma brucei* bloodstream forms in culture. *Mol. Biochem. Parasitol.* 90, 269–280
2. Vassella, E. *et al.* (1997) Differentiation of African trypanosomes is controlled by a density sensing mechanism which signals cell cycle arrest via the cAMP pathway. *J. Cell Sci.* 110, 2661–2671
3. Mony, B.M. *et al.* (2013) Genome-wide dissection of the quorum sensing signalling pathway in *Trypanosoma brucei*. *Nature* 505, 681
4. McDonald, L. *et al.* (2018) Non-linear hierarchy of the quorum sensing signalling pathway in bloodstream form African trypanosomes. *PLoS Pathog.* 14, e1007145
5. Rojas, F. *et al.* (2019) Oligopeptide signaling through *Tb*GPR89 drives trypanosome quorum sensing. *Cell* 176, 1–12
6. Tetaert, D. *et al.* (2018) Unusual cleavage of peptidic hormones generated by trypanosome enzymes released in infested rat serum. *Int. J. Pept. Protein Res.* 41, 147–152
7. Geiger, A. *et al.* (2010) Exocytosis and protein secretion in *Trypanosoma*. *BMC Microbiol.* 10, 20
8. Bossard, G. *et al.* (2013) Secreted proteases of *Trypanosoma brucei gambiense*: possible targets for sleeping sickness control? *BioFactors* 39, 407–414
9. Capewell, P. *et al.* (2016) The skin is a significant but overlooked anatomical reservoir for vector-borne African trypanosomes. *eLife* 5, e17716
10. Trindade, S. *et al.* (2016) *Trypanosoma brucei* parasites occupy and functionally adapt to the adipose tissue in mice. *Cell Host Microbe* 19, 837–848
11. Brancucci, N.M.B. *et al.* (2017) Lysophosphatidylcholine regulates sexual stage differentiation in the human malaria parasite *Plasmodium falciparum*. *Cell* 171, 1532–1544.e15

Forum

Gene Function Discovery for Kinetoplastid Pathogens

Reza Salavati[1,2,3,*] and Vahid H. Gazestani[4]

We propose to integrate the existing and new experimental data

with computational tools to model interaction networks for the most prominent kinetoplastid pathogens. These interaction networks will vastly expand the functional annotation of the kinetoplastid genomes, which in turn are critical for identifying new routes of disease intervention.

The Kinetoplastid Pathogens and Annotation of Their Genomes

Parasitic kinetoplastids consist of the medically important trypanosomatids which affect large populations globally and cause numerous deaths [1]. They include the causal agents of human African sleeping sickness, Chagas' disease, and leishmaniasis (*Trypanosoma brucei* group, *T. cruzi*, and *Leishmania* species, respectively).

The advent of recent technologies and the availability of genome sequences for trypanosomatid parasites has led to the functional characterization of trypanosomatid genes and their expression patterns at different life stages [2]. However, the functional roles of many trypanosomatid genes still remain poorly understood (Table 1). For instance, the genome of *T. brucei* encodes 11 203 genes, of which only 6912 have annotated Gene Ontology (GO) – biological process annotation (622 distinct terms). Furthermore, only 1221 genes are assigned to 93 different KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways. As a further complication, most of these annotations are computationally predicted using sequence homology-based approaches. For example, *T. brucei* is the only organism in Table 1 for which the functional role of more than 1000 of its genes is supported by experimental evidence in terms of GO biological process annotations. Owing to high evolutionary divergence and low sequence similarity with model organisms [3], the reliability of such computational predictions remains to be established. These issues present a major bottleneck in understanding the biology of these parasites and finding new therapeutic options.

Computational and Experimental Approaches to Assign Genes to Pathways and Biological Functions

In the past several years, we and others have pioneered an array of computational and experimental approaches that make it possible to move away from homology-based annotation of genes, and to integrate diverse sources of organism-specific evidence to assign genes to pathways and biological functions [4–6]. These approaches establish a strong foundation for the development of innovative, integrated frameworks for functional annotation of genes that do not have a closely related characterized homolog.

Function Prediction Based on Computational Analysis

One of the most promising approaches for the annotation problem is to derive machine-learning-based models that incorporate a wide range of variables to functionally classify genes (Figure 1A). The premise behind these approaches is that each dataset has only a limited predictive potential on its own, but the combination of the information from orthogonal resources can lead to substantially improved predictions [5,7]. As an illustration, machine-learning techniques have been successfully used to infer protein networks by combining different datasets based on positive (i.e., proteins known to interact with one another) and negative (i.e., proteins that do not interact with one another) gold standards [8]. Next, functions of proteins with unknown functions can be inferred based on the detection of protein complexes [9] or the propagation of biological information through the constructed networks [6]. The extent to which information is propagated from one protein to another is proportional to their interaction weight that is learned by the employed machine-learning approach. The interaction weight summarizes the similarity that the two proteins exhibit in terms of different biological measures such as patterns of transcription and gene essentiality, as well as similarities in cor-

Table 1. Annotation State of the Kinetoplastid Pathogens[a]

Organism	No. identified genes	No. annotated genes	No. genes annotated based on experimental evidence	No. genes annotated based on physical interaction evidence
Trypanosoma brucei TREU927	11 203	6912	4122	368
Leishmania major Friedlin	8519	4796	197	32
Trypanosoma cruzi CLBrener Esmeraldo-like	10 338	5218	692	3
Leishmania infantum JPCM5	8237	4704	132	3
Trypanosoma vivax Y486	11 394	4680	16	0

[a]The data are based on the annotations present in the TriTrypDB release 37 (April 25, 2018). GO annotations with the experimental evidence codes were considered as experimentally verified.

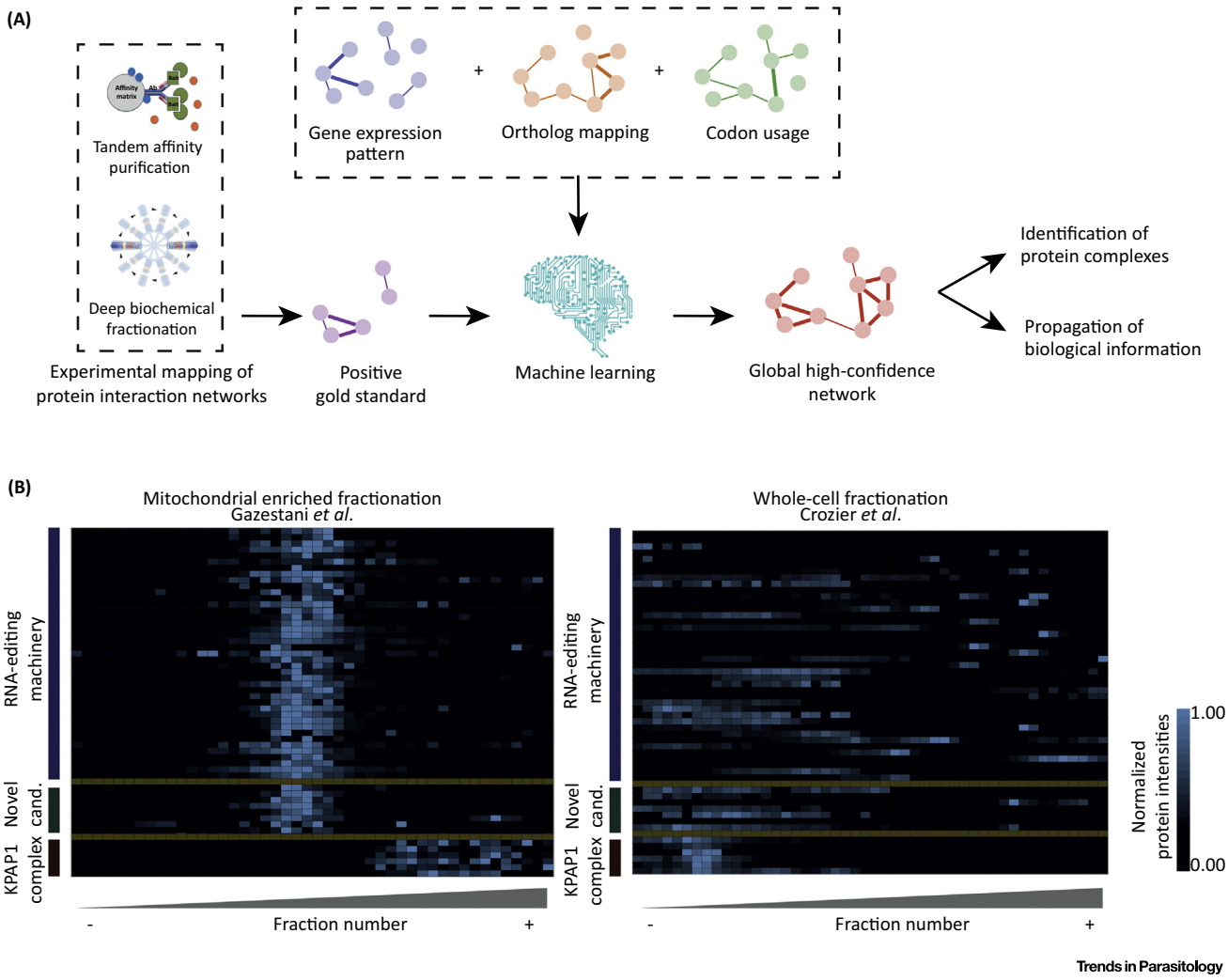

Figure 1. Annotation of Gene Function Based on Computational and Experimental Approaches. (A) Schematic representation of the general framework for deciphering gene functions by machine-learning-based integration of a wide range of resources, including gene expression, evolutionary history, and codon usage. To properly train these classifiers, large and representative positive and negative gold standards of trypanosomatid proteomes are needed. As elaborated in the text, we propose that such gold standards could be experimentally generated by combining results from different large-scale experiments. (B) Organelle enrichment can significantly enhance resolution of deep biochemical fractionation experiments. The figure compares the cofractionation patterns of two distinct mitochondrial complexes (RNA-editing machinery and poly(A) polymerase1 (KPAP1) complex) between enriched mitochondrial fractionation (Gazestani *et al*. [9,12]) and whole-cell fractionation (Crozier *et al*. [11]) experiments. While RNA-editing machinery (including core editosome and accessory elements) is involved in the editing of some of the mitochondrial RNAs, KPAP1 complex regulates the translation of mitochondrial RNAs. Each row represents a member of these two complexes, and each column demonstrates a fraction in an increasing order. The color intensities represent the relative intensity of a protein at a specific fraction. As illustrated, the members of each complex are highly cofractionated with each other while well separated from the other complex in mitochondrial-enriched fractionation patterns from Gazestani *et al*. The high resolution of the fractionation patterns allowed to identify novel candidates involved in RNA-editing machinery which was verified by follow-up experiments. However, as shown, the resolution of fractionation patterns is diminished in whole-cell experiments. For illustration purposes, we have randomly chosen one experiment from several whole-cell experiments conducted by Crozier *et al*. (the fractionation resolutions were similar across whole-cell experiments).

egulation mechanisms, protein domains, and genetic interactions [6,8].

Although promising, machine-learning approaches are not widely adopted to nonmodel organisms, such as trypanosomatids, primarily due to the lack of large-scale positive gold standard data for model training. As we argue below, this gap can be addressed by systematic efforts to experimentally infer the functional role of trypanosomatid proteins at large scale. Such information can be exploited to develop classifiers that predict the functional and/or physical

interactions for the remaining genes in the genome.

Functional Characterization Based on Experimental Approaches

Protein–protein interaction maps can provide evidence for the functional relatedness of proteins as the functionally related proteins usually interact with one another [10]. Although highly useful for the functional characterization of proteins, they have not been extensively used to annotate the genome of trypanosomatids (Table 1). We recently pioneered cost-effective experimental approaches for proteome-wide mapping of complexes by coupling deep biochemical fractionation of proteins with semiquantitative mass spectrometry analysis [9]. By this approach, we systematically interrogated the co-complex interactions of 40% of genes that are expressed in the insect-form life stage of *T. brucei*. The accuracy of deciphered protein interactions was confirmed by assessing the reproducibility of cofractionation patterns, and showing that the interacting genes are involved in similar biological processes and molecular pathways, coexpressed across the parasite life stages, and exhibit cellular colocalization. One salient observation was that the accuracy of inferred interactions increased from 34% to 68% when considering interactions that were supported based on two different fractionation approaches. This increase in accuracy was not limited to the results from fractionation methods and was even more pronounced for interactions that were supported by more distant techniques such as those interactions supported by both fractionation and tandem affinity purification. For example, there are ~360 proteins that are reported to copurify by trypanosomatid RNA-editing machinery based on affinity purification methods. Of these, only 50 proteins were strongly cofractionated with each other in mitochondrial enriched fractionation experiments (Figure 1B). Most of these 50 proteins were already known to be involved in RNA-editing machinery, and follow-up experiments confirmed the accuracy of the new predictions [9]. Similar results were observed for the *T. brucei* aminoacyl-tRNA synthetase proteins, where merging a list of 262 copurified proteins with fractionation patterns led to the identification of proteins involved in tRNA synthesis. The other important observation from fractionation-based approaches was that organelle enrichment prior to fractionation significantly enhances the resolution of observed fractionation patterns and consequently leads to more accurate predictions. This point is illustrated clearly by comparing the fractionation patterns of RNA-editing machinery from mitochondrial-enriched extracts reported by Gazestani *et al*. [9] with those of whole-cell extract from Crozier *et al*. [11] (Figure 1B). Moreover, comparison of the cofractionation pattern of complexes demonstrated that different fractionation methods provide a complementary view of the interactome landscape. For example, while glycerol gradient fractionation preserves less stable interactions, ion-exchange chromatography favors more stable interactions potentially due to the application of a salt gradient to the mobile phase. By leveraging these observations, a high-confidence core network with 6636 interactions among 665 *T. brucei* proteins was constructed [9]. This single study not only provided experimental evidence on biological roles of 472 proteins with no available experimentally derived GO annotations, but also resulted in the inference of functional roles for 131 proteins annotated as hypothetical.

Protein interaction maps in combination with gene expression across life stages, gene essentiality data, and gene regulatory networks, provide a comprehensive toolbox for functional annotation of proteins. TrypsNetDB (http://trypsnetdb.org/) [12] tries to address this gap by combining large- and small-scale datasets across trypanosomatid parasites. The database currently covers the physical protein interactions extracted from 97 different studies, and is accessible through widely used TriTrypDB (http://tritrypdb.org) [13]. Moreover, the TrypsNetDB automatically propagates information from various resources (e.g., gene expression, RNA decay, gene essentiality, protein fractionation patterns, etc.) across trypanosomatid parasites to provide information on queried proteins in the species of interest. Currently, the database includes 101,187 total physical interactions, providing network context to 13,395 proteins across 16 trypanosomatid parasites. We anticipate that this database will lead to an initiative for pipelining computational tools that we and others have developed for functional annotation of genes and proteins. The proposed resource will vastly expand the power and ease of use of the proposed analyses, making them available to every researcher with access to limited computational facilities. Integration of such a pipeline with TriTrypDB will significantly enhance the annotation of parasite genomes and will lead to a better understanding of the biology of these organisms and, potentially, new approaches for intervention in parasite survival and infection.

Future Direction

Highly accurate annotation of trypanosomatid genomes can be obtained by conducting similar deep fractionation-based experiments and establishing experimentally derived interacting networks in other related trypanosomatid parasites (e.g., *T. cruzi* and *Leishmania* spp.). The primary benefits of this approach are to experimentally examine and revise current computational predictions, and provide high-confidence, experimentally derived data to infer the function of other

uncharacterized proteins using machine-learning algorithms. These approaches are also able to identify the parasite-specific subunits of evolutionarily conserved complexes, such as proteasome, ribosome, and RNA polymerases. As these complexes are often crucial for the survival of the parasite, the parasite-specific subunits of these complexes are potentially ideal targets for drug development. Moreover, by performing deep fractionation experiments across different life stages or environmental conditions, one could gain insight on the molecular reorganizations that are causal to parasite adaptation and survival. Dynamic interaction of functionally related proteins in cell lines with different genetic backgrounds or different environmental conditions can be valuable for creating more reliable interaction maps. These environmental conditions can include growth in the presence of different chemicals and drugs in current use. The dynamic interaction models can predict new interactions, provide a platform for exploring context-specific networks, rediscover new interactions, and reveal how pathways respond to environmental inputs or drugs, which are critical for improved target-based discovery and interpretation of results from phenotypic screening of chemical libraries.

Acknowledgments

The research in the Salavati laboratory is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) grant # RGPIN 328186.

[1]Institute of Parasitology, McGill University, 21111 Lakeshore Road, Ste. Anne de Bellevue, Montreal, Quebec H9X3V9, Canada
[2]McGill Centre for Bioinformatics, McGill University, 3649 Promenade Sir William Osler, Montreal, Quebec H3G0B1, Canada
[3]Department of Biochemistry, McGill University, McIntyre Medical Building, 3655 Promenade Sir William Osler, Montreal, Quebec H3G1Y6, Canada
[4]Department of Pediatrics, University of California, San Diego, School of Medicine, La Jolla, CA 92093, USA

*Correspondence: reza.salavati@mcgill.ca (R. Salavati).

https://doi.org/10.1016/j.pt.2018.09.003

References

1. Field, M.C. *et al.* (2017) Anti-trypanosomatid drug discovery: an ongoing challenge and a continuing need. *Nat. Rev. Microbiol.* 15, 217
2. Glover, L. *et al.* (2015) Genome-scale RNAi screens for high-throughput phenotyping in bloodstream-form African trypanosomes. *Nat. Protocols* 10, 106
3. El-Sayed, N.M. *et al.* (2005) Comparative genomics of trypanosomatid parasitic protozoa. *Science* 309, 404–409
4. Salavati, R. and Najafabadi, H.S. (2010) Sequence-based functional annotation: what if most of the genes are unique to a genome? *Trends Parasitol.* 26, 225–229
5. Najafabadi, H.S. and Salavati, R. (2008) Sequence-based prediction of protein-protein interactions by means of codon usage. *Genome Biol.* 9, R87
6. Mostafavi, S. *et al.* (2008) GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biol.* 9, S4
7. Jansen, R. *et al.* (2003) A Bayesian networks approach for predicting protein–protein interactions from genomic data. *Science* 302, 449–453
8. Greene, C.S. *et al.* (2015) Understanding multicellular function and disease with human tissue-specific networks. *Nat. Genet.* 47, 569
9. Gazestani, V. *et al.* (2016) A protein complex map of *Trypanosoma brucei*. *PLoS Negl. Trop. Dis.* 10, e0004533
10. Oliver, S. (2000) Proteomics: guilt-by-association goes global. *Nature* 403, 601
11. Crozier, T.W. *et al.* (2017) Prediction of protein complexes in *Trypanosoma brucei* by protein correlation profiling mass spectrometry and machine learning. *Mol. Cell. Proteom.* 16, 2254–2267
12. Gazestani, V.H. *et al.* (2017) An integrated framework for the functional characterization of trypanosomatid proteins. *PLoS Negl. Trop. Dis.* 11, e0005368
13. Aslett, M. *et al.* (2009) TriTrypDB: a functional genomic resource for the Trypanosomatidae. *Nucleic Acids Res.* 38, D457–D462