

Title

Detection of secondary structure-based regulatory nucleotide elements

Background

The trypanosomatids are a monophyletic sub-group of excavate parasites [1,2] which collectively exhibit an extraordinarily broad host-range which extends from crocodiles [3] to coconut palms [4]. Several members of the trypanosomatid group cause globally important parasitic diseases of humans including sleeping sickness, Chagas disease and leishmaniases which together kill and debilitate hundreds of thousands of people worldwide each year (<http://www.who.int>). The study of gene-expression in trypanosomatids has been highly productive and has led to the discovery of several fundamental aspects of transcription, some of which have subsequently been found to be utilised in many other eukaryotes. These discoveries include RNA editing [5], trans-splicing [6,7,8] and RNA polymerase I transcription of protein coding genes [9,10]. Despite these breakthroughs in understanding gene expression in trypanosomatids, we know little about the identity and role of the many regulatory elements which must govern this system. This is particularly pertinent as most protein coding genes in trypanosomatids are expressed in polycistronic transcription units where promoters appear to be absent from the 5' end of genes. Moreover, almost nothing is known about regulatory elements which govern this system, and only a small number of cis-elements and transacting protein factors have been identified [11,12,13,14,15].

In trypanosomatids, transcription of all protein coding genes by RNA polymerase II proceeds polycistronically [16,17,18,19]. The average polycistron contains ~70 genes [20] and each gene must be processed to monocistronic RNA units before nuclear export and translation can occur. This obligatory transition to monocistronic RNA molecules is mediated by trans-splicing, a process in which a 39 nucleotide capped RNA molecule is added upstream of the start codon of each coding sequence [7,8,21]. This profound difference in mechanism from how transcription proceeds in well studied model organisms such as Humans and yeast is mirrored by an absence of transcription regulatory proteins from the trypanosomatid genomes.

Several methods have been developed to identify regulatory motifs in DNA sequences. The methods are diverse and can range from information-rich methods (for example [22,23]), which utilise groups of co-regulated or functionally related genes to search for regulatory sequences in upstream regions, to information-poor methods (for example [24,25]), which look for over-represented sequence motifs in a collection of sequences. Unsurprisingly, these and other methods have failed to shed light on any recurring regulatory motifs in trypanosomatid DNA sequences. Hence there is a pressing need to develop novel *ab initio* approaches for detecting regulatory information in these genomes.

Given that complex patterns of gene expression regulation exist and that cis-acting regulatory motifs appear to be rare this suggests that RNA secondary structure may play an important role in modulating gene expression regulation.

This project will utilise the 8 available trypanosomatid genomes to identify secondary

structure elements in the non-translated regions surrounding coding sequences. Putative secondary structures are easy to predict. Individual structures can be clustered according to several properties including nucleotide composition, free energy at physiologically relevant temperatures, topology, position relative to known processing sites and conservation across different trypanosomatid genomes. This data will then be used as a grammar to develop an *ab initio* prediction algorithm for identification of putative regulatory structural elements.

Timeline

Week 1: Read key papers from the literature list and make preliminary contents of the report

Week 2: Analyze genome data to identify stable secondary structure elements up to 200bp in length located in intergenic regions. At temperatures relevant to the different life cycle stages of the parasite eg @37°C for bloodstream form <=28°C for procyclic form.

Week 3-4: Cluster structural elements based on intrinsic properties, temperature sensitivity and location relative to coding sequences.

Week 5-6: Correlate expression data from RNAseq experiments with presence of identified structural elements.

Trypanosomatid genomes are small, gene dense and lack introns so are easy to work with computationally. The Ideal candidate would have some expertise in a scripting or programming language and would have some experience in data clustering. Though this is not essential.

References

1. Hamilton PB, Gibson WC, Stevens JR (2007) Patterns of co-evolution between trypanosomes and their hosts deduced from ribosomal RNA and protein-coding gene phylogenies. *Mol Phylogenet Evol* 44: 15-25.
2. Simpson AG, Gill EE, Callahan HA, Litaker RW, Roger AJ (2004) Early evolution within kinetoplastids (euglenozoa), and the late emergence of trypanosomatids. *Protist* 155: 407-422.
3. Hoare C, A (1929) Studies on *Trypanosoma grayi*. *Transactions of the Royal Society of Tropical Medicine and Hygiene* 23: 18.
4. Donovan C (1909) Kala-azar in Madras, especially with regard to its connexion with the dog and the bug (*Conorrhinus*). *Lancet* 177: 2.
5. Benne R, Van den Burg J, Brakenhoff JP, Sloof P, Van Boom JH, et al. (1986) Major transcript of the frameshifted coxII gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA. *Cell* 46: 819-826.
6. Boothroyd JC (1985) Antigenic variation in African trypanosomes. *Annu Rev Microbiol* 39: 475-502.
7. Borst P (1986) Discontinuous transcription and antigenic variation in trypanosomes. *Annu Rev Biochem* 55: 701-732.
8. Van der Ploeg LH (1986) Discontinuous transcription and splicing in trypanosomes. *Cell* 47: 479-480.
9. Rudenko G, Bishop D, Gottesdiener K, Van der Ploeg LH (1989) Alpha-amanitin resistant transcription of protein coding genes in insect and bloodstream form *Trypanosoma brucei*. *EMBO J* 8: 4259-4263.
10. Navarro M, Gull K (2001) A pol I transcriptional body associated with VSG mono-

- allelic expression in *Trypanosoma brucei*. *Nature* 414: 759-763.
11. Engstler M, Boshart M (2004) Cold shock and regulation of surface protein trafficking convey sensitization to inducers of stage differentiation in *Trypanosoma brucei*. *Genes Dev* 18: 2798-2811.
 12. Estevez AM (2008) The RNA-binding protein TbDRBD3 regulates the stability of a specific subset of mRNAs in trypanosomes. *Nucleic Acids Res* 36: 4573-4586.
 13. Walrad P, Paterou A, Acosta-Serrano A, Matthews KR (2009) Differential trypanosome surface coat regulation by a CCCH protein that co-associates with procyclin mRNA cis-elements. *PLoS Pathog* 5: e1000317.
 14. Furger A, Schurch N, Kurath U, Roditi I (1997) Elements in the 3' untranslated region of procyclin mRNA regulate expression in insect forms of *Trypanosoma brucei* by modulating RNA stability and translation. *Mol Cell Biol* 17: 4372-4380.
 15. Schurch N, Furger A, Kurath U, Roditi I (1997) Contributions of the procyclin 3' untranslated region and coding region to the regulation of expression in bloodstream forms of *Trypanosoma brucei*. *Mol Biochem Parasitol* 89: 109-121.
 16. McDonagh PD, Myler PJ, Stuart K (2000) The unusual gene organization of *Leishmania major* chromosome 1 may reflect novel transcription processes. *Nucleic Acids Res* 28: 2800-2803.
 17. Martinez-Calvillo S, Yan S, Nguyen D, Fox M, Stuart K, et al. (2003) Transcription of *Leishmania major* Friedlin chromosome 1 initiates in both directions within a single region. *Mol Cell* 11: 1291-1299.
 18. Wright JR, Siegel TN, Cross GA Histone H3 trimethylated at lysine 4 is enriched at probable transcription start sites in *Trypanosoma brucei*. *Mol Biochem Parasitol* 172: 141-144.
 19. Siegel TN, Hekstra DR, Kemp LE, Figueiredo LM, Lowell JE, et al. (2009) Four histone variants mark the boundaries of polycistronic transcription units in *Trypanosoma brucei*. *Genes Dev* 23: 1063-1076.
 20. Ivens AC, Peacock CS, Worthey EA, Murphy L, Aggarwal G, et al. (2005) The genome of the kinetoplastid parasite, *Leishmania major*. *Science* 309: 436-442.
 21. Clayton CE (2002) Life without transcriptional control? From fly to man and back again. *Embo J* 21: 1881-1888.
 22. Cora D, Di Cunto F, Provero P, Silengo L, Caselle M (2004) Computational identification of transcription factor binding sites by functional analysis of sets of genes sharing overrepresented upstream motifs. *BMC Bioinformatics* 5: 57.
 23. Gupta M, Liu JS (2005) De novo cis-regulatory module elicitation for eukaryotic genomes. *Proc Natl Acad Sci U S A* 102: 7079-7084.
 24. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, et al. (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* 37: W202-208.
 25. Thompson W, Rouchka EC, Lawrence CE (2003) Gibbs Recursive Sampler: finding transcription factor binding sites. *Nucleic Acids Res* 31: 3580-3585.