

# Practical Evolutionary Genomics from Short-Read Sequence Experiments

Patrick J. Biggs<sup>1,2</sup>

[P.Biggs@massey.ac.nz](mailto:P.Biggs@massey.ac.nz)

Lesley J. Collins<sup>1,2</sup>

[L.J.Collins@massey.ac.nz](mailto:L.J.Collins@massey.ac.nz)

Sylvia Xiwei Chen<sup>1,2</sup>

[sylvia.x.chen@gmail.com](mailto:sylvia.x.chen@gmail.com)

Nigel French<sup>3</sup>

[N.P.French@massey.ac.nz](mailto:N.P.French@massey.ac.nz)

1. Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, Palmerston North, New Zealand.
2. Institute for Molecular BioSciences, Massey University, Palmerston North, New Zealand.
3. Institute of Veterinary, Animal and Biomedical Sciences, Massey University, Palmerston North, New Zealand.

## Abstract

Short-read sequences (< 36nt) as obtained by some Next-Generation Sequencing platforms are often considered too short for evolutionary analysis. From experience we have found that this is not necessarily the case. Using information from two projects (bacterial genomic analysis and protist small RNA analysis) we show that evolutionary information can be recovered using very short-reads. Genomic sequencing gives us SNP (single nucleotide polymorphisms) as an obvious indicator of genomic differences, but we can also readily find areas showing increased numbers of mutations. Small RNA analysis not only allows identification of new ncRNAs and evidence of expression but allow the investigation the evolution of some ncRNA groups such as snoRNAs. Although our work at this stage is only preliminary, we show that high-throughput sequence data even when 'short' and from a non-model organism, is indeed very useful in the area of evolutionary genomics.

**Keywords:** short-read sequencing, Solexa, evolutionary genomics

## 1 Introduction

Although relatively new, Short-read sequencing technology has revolutionized how we plan genomic scale projects. Despite the fact that many Next-Generation Sequencing platforms are moving to longer sequence reads there remains a perception that data from shorter reads (< 35nt) are of insufficient length to contribute to evolutionary analysis. We present here two examples showing that such reads are useful for evolutionary analysis but also that with small RNA work, our example here being small RNAs from *Giardia lamblia*, even shorter sequences (17nt) can be used.

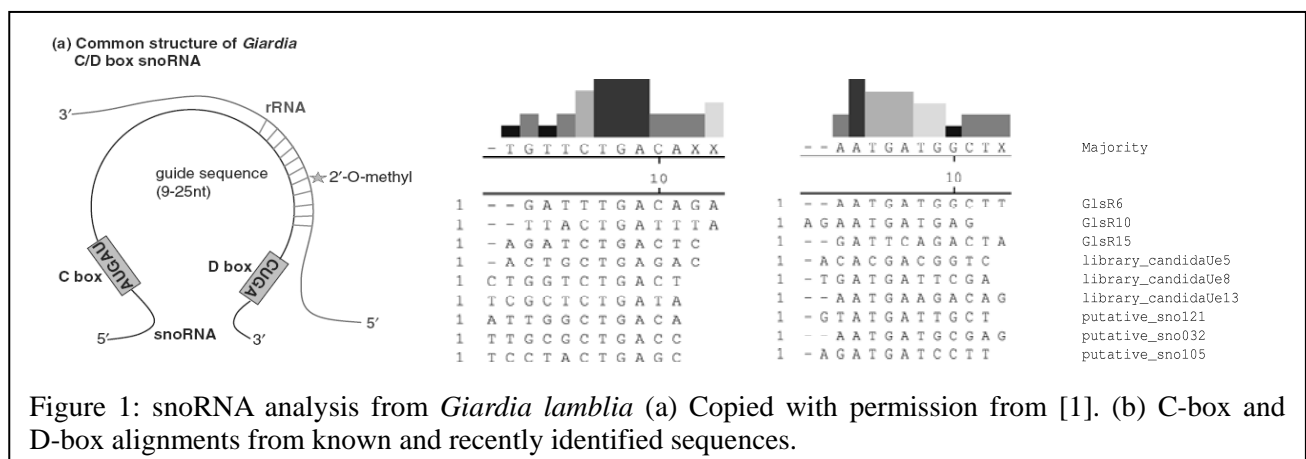


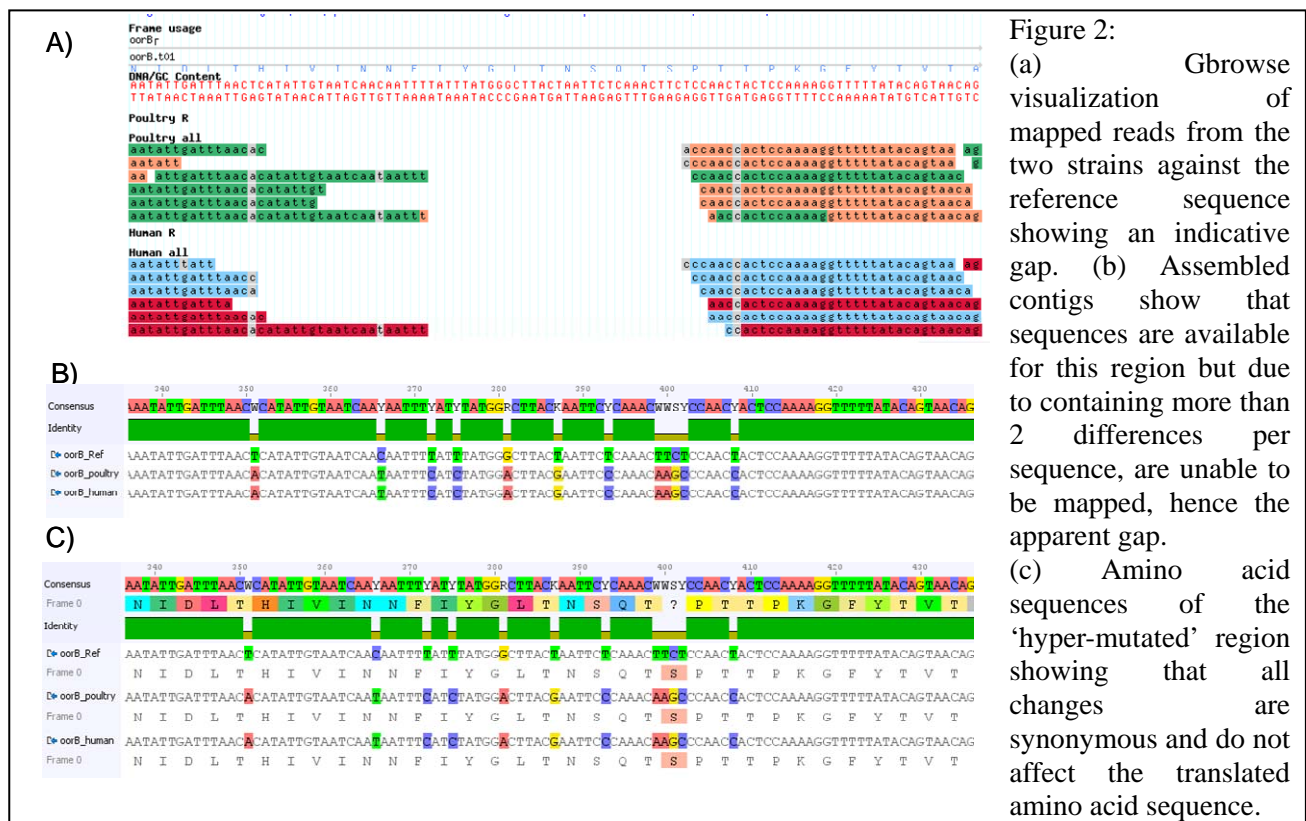
Figure 1: snoRNA analysis from *Giardia lamblia* (a) Copied with permission from [1]. (b) C-box and D-box alignments from known and recently identified sequences.

## 2 Method and Results

Example1: High-throughput sequencing of *Giardia lamblia* small RNAs (band size 20-200nt) allowed

complete coverage of some smaller ncRNAs (e.g. tRNAs and snoRNAs) but incomplete coverage of other known ncRNAs[1]. As well as confirming the expression of some ncRNA candidates, the sequences that we obtained allowed investigation of differences between snoRNAs. Figure 2 shows an alignment of snoRNAs for which hits were obtained and sequences verified. Evolutionary analyses of these sequences are now underway. Because many ncRNAs are smaller than 36nt (e.g. miRNAs < 22nt) mappings used 17nt as well as the full 36bp lengths of the sequences. Results could be recovered with both lengths.

Example 2: Short-reads of 36nt length from a high-throughput Solexa re-sequencing project of two *Campylobacter jejuni* strains were mapped to the reference genome (AL111168). Sequences were also assembled *de novo* into contigs using Velvet [2] and analyzed by specific Perl scripts to find possible indicators of ‘hyper-mutated’ regions. An example of results is shown in Figure 2. Single nucleotide gaps were also found in some genes (not shown), as well as other larger deletions, some of which may be possible indicators of chromosomal rearrangements.



### 3 Discussion

The use of high-throughput short-read sequencing is only in its infancy. We conclude that even the shortest reads produced by this technology contain valuable evolutionary indicators. As the technology improves, reads will become longer and coverage more intense. Since expression tag analysis often uses sequences of 17-18nt it raises the issue that results from one type of experiment may hold much more information than can be initially appreciated, especially for evolutionary analysis.

### References

[1] Chen, X.S., Rozhdestvensky, T.S., Collins, L.J., Schmitz, J. and Penny, D. Combined experimental and computational approach to identify non-protein-coding RNAs in the deep-branching eukaryote *Giardia intestinalis*. *Nucleic Acids Res* 35, 4619-4628, 2007.

[2] Zerbino, D.R. and Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* 18, 821-9, 2008.