Reviewer's report

Title: SICTIN: Rapid footprinting of massively parallel sequencing data.

Version: 1 **Date:** 3 April 2010

Reviewer: Richard Green

Reviewer's report:

Enroth et al. present a computational tool, SICTIN, for summarizing and querying a generic signal over a genome. They demonstrate the utility of the method by converting some existing nucleosome position data into the indexed, binary format they describe. Then, the authors use the associated tools to interrogate and display some nucleosome positioning trends recently discovered elsewhere.

While, there is no doubt that SICTIN is a useful tool for handling and querying specific kinds of data, I have several reservations about the work and the manuscript. These are detailed below, in order of severity.

1. It is not immediately obvious that the authors have chosen the optimal data representation for the kinds of data SICTIN is designed to handle. Since each data index requires several Gb of space over a typical genome-scale data set, this representation may not scale well to allow simultaneous interrogation of dozens or hundreds of genome-scale datasets. A brief but thoughtful exposition of the merits of an independent, disk-based index over each dataset would be useful. I would not necessarily advocate a relational database solution for this project, but it is an obvious candidate.

2. The manuscript is poorly written. It would greatly benefit from at least one thorough, outside reading. The main shortcomings are issues of clarity. Below I detail many of these:

a. The “Background” section of the abstract is a muddled progression of semi-related sentences. The second sentence begins “Examples include…” but it is not clear if these are examples of “Massively parallel sequencing” or “genome-wide hypothesis-free investigation”. I think the meaning to be conveyed by the third sentence is better represented by “Although nucleotide-resolution, detailed information can be easily generated, biological insight often requires a more general view of these data.” The “it” referred to in the last sentence is not clear.

b. The “Methods” section of the abstract uses the undefined term “footprint generation”. I think I understand the concept of a footprint in this context, but it required reading the manuscript. The term footprint has a more common connotation – a region of DNA protected by a DNA-biding protein during enzymatic digestion.

c. The “Conclusions” section of the abstract states that “…the main emphasis has been on transcriptional start sites…”. It is not at all whose or what’s main

emphasis is being described.

d. Page 3 & 4 contains an oddly detailed description of various C/C++ data types (unsigned short, float, etc.). More helpful to the reader would be description of the disk-index rational at this level of description.

e. Top of p. 5 – "…has been described to be phased with respect to the TSS's." It is left undefined and thus unclear what "phased" means in this context.

f. Page 7 – Figure 2 legend. The term "ultra-conserved" has been adopted in a different context to mean inexplicable and high conservation over hundreds of base-pairs across a large evolutionary distance, e.g., human to mouse. The authors should not mis-use this term to describe the conservation of splice donor sites. Simply describing splice-donor sites as "conserved" would be succinct and less confusing.

Level of interest: An article of limited interest

Quality of written English: Needs some language corrections before being published

Statistical review: No, the manuscript does not need to be seen by a statistician.

Declaration of competing interests:

I declare that I have no competing interests.