**Bacterial Genome Synchronizer**

I am interested in making a computational tool developed by Matt become publicly accessible via the Innovation Centre computational infrastructure. The basis of the tool is a recognition that bacterial genomes are circular, and independent genome sequencing projects (which generate linear contigs) will often result in an arbitrary start and end of the sequence. By providing a defined start point, the tool will reconfigure input genomes into an output of genomes reconfigured for a common start point.

**Tool description:**

Clients would submit two pieces of input.

1. An NCBI bacterial genome sequence accession number (preferable) or a fasta file of a bacterial genome sequence. The genomic sequence would generally range in length from a 2 million to 6 million nucleotide text file. *Value added: clients could submit a list (limits?) or multiple files.*
2. A 50-100 nucleotide text file indicating the client’s desired start point for the genome sequence(s) provided. Per job, only a single start point sequence would be allowed. Sequences shorter than 50 nt would be discontinued with an appropriate error message. Sequences longer than 100 nt would be truncated at 100 nt, with an appropriate warning message.

Computational protocol.

1. A reverse complement of the 50-100 nt start point sequence would be calculated.
2. The start point sequences (original and revC) would be text matched to the genome sequence(s).

If neither of the start point sequences do not exactly match any loci in the genome sequence, an error message “start not found” is displayed, and the processing of that genome is halted.

If the original and/or revC exactly match more than one locus, an error message (non-unique start point” is displayed, and the processing of that genome is halted.

1. When the original start point sequence exactly matches at a single genomic locus (and the revC has zero matches), the procedure progresses to step 5.
2. When the revC start point sequence exactly matches at a single genomic locus (and the original sequence has zero matches), the reverse complement of the genome sequence is calculated, and the revC\_genome is used in step 5.
3. At this point, the genomes should have a single locus exactly matching the original start point sequence. The tool would then “break” the sequence immediately ahead of the first nucleotide of the start sequence. The entirety of sequence preceding the new start nucleotide would then be appended at the end of the genome sequence. (Visually, if S is the desired start point, the sequence A-B-C-D-E-S-G-H-I-J-K would be transformed to S-G-H-I-J-K-A-B-C-D-E.)
4. The tool would do a QC/QA to ensure that the reconfigured genome sequence exactly matches the length of the input genome sequence.
5. The QC/QA approved synchronized genome sequence(s) would be made available to the client for downloading.

Considerations:

1. Scale: what is a reasonable number of jobs, or number of genomes/job?
2. Time: for public uptake, the entire process (especially for a single genome) should be nearly instantaneous.