Bioinformatics computing project

The problem:

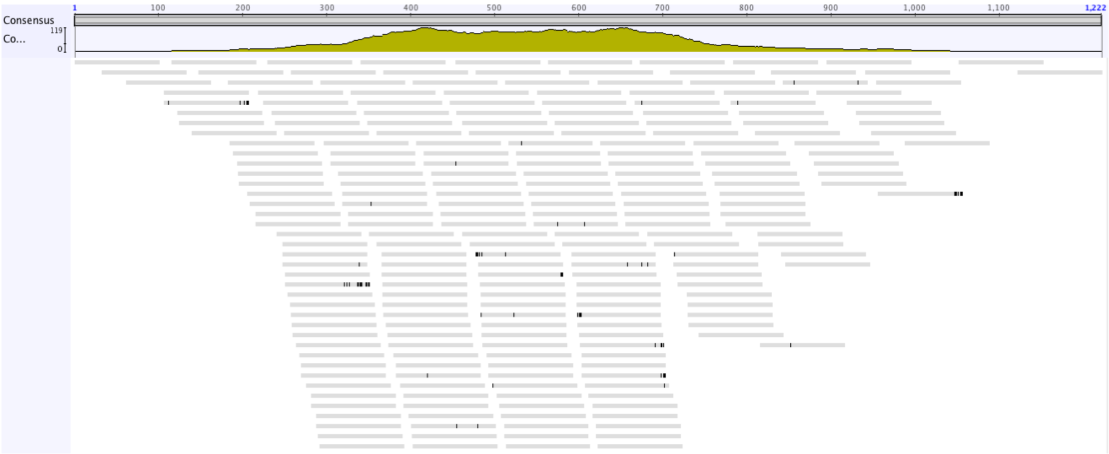
Obtaining the spa type of a *Staphylococcus aureus* bacteria using whole genome sequence data.

Background:

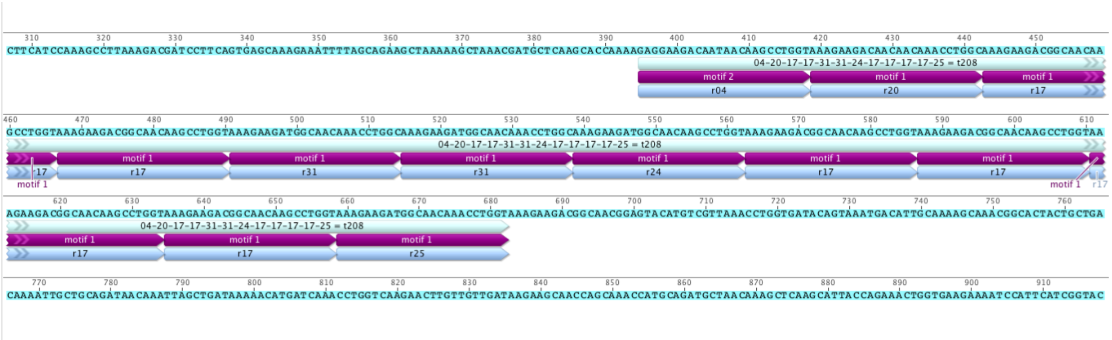
The spa type is obtained by matching a pattern of spa repeats (e.g. the spa type t843 = r04-r82-r17-r25-r17-r25-r25-r16-r17). Each repeat is a specific DNA sequence of 21,24, or 27 DNA base pairs (e.g. r17 = AAAGAAGACGGCAACAAGCCTGGT).

If we started from a complete DNA sequence (as a single string) you would just search the string for each of the repeats in the library of repeats and establish the sequence of repeats and then look that up in the library of spa types.

The problem is that whole genome sequencing only provides reads (string lengths) of around 100 base pairs and that these reads contain occasional errors. There are many computer programs designed to reconstruct the genetic sequence from these short reads but they are optimized for reconstruction of the whole genome (about 3.5K base pairs) from the large number of short reads that are produced (about 3,000 strings of 100 base pairs which cover the whole genome). Most algorithms sacrifice accuracy for processor time. They often fail to reconstruct areas of the genome where there are repetitions of short sequences (e.g. the spa gene) properly.



(Diagrammatic illustration of short reads aligned against a reference sequence)



```
310       320       330       340       350       360       370       380       390       400       410       420       430       440       450
CTTCATCCAAAGCCTTAAAGACGATCCTTCAGTGAGCAAAGAAATTTTAGCAGAAGCTAAAAAGCTAAACGATGCTCAAGCACCAAAAGAGGAAGACAATAACAAGCCTGGTAAAGAAGACAACAACAAACCTGGCAAAGAAGACGGCAACAA
                                                                                          04-20-17-17-31-31-24-17-17-17-17-25 = t208

460       470       480       490       500       510       520       530       540       550       560       570       580       590       600       610
GCCTGGTAAAGAAGACGGCAACAAGCCTGGTAAAGAAGATGGCAACAAACCTGGCAAAGAAGATGGCAACAAACCTGGCAAAGAAGATGGCAACAAGCCTGGTAAAGAAGACGGCAACAAGCCTGGTAAAGAAGACGGCAACAAGCCTGGTAA
04-20-17-17-31-31-24-17-17-17-17-25 = t208

620       630       640       650       660       670       680       690       700       710       720       730       740       750       760
AGAAGACGGCAACAAGCCTGGTAAAGAAGACGGCAACAAGCCTGGTAAAGAAGATGGCAACAAACCTGGTAAAGAAGACGGCAACGGAGTACATGTCGTTAAACCTGGTGATACAGTAAATGACATTGCAAAAGCAAACGGCACTACTGCTGA
04-20-17-17-31-31-24-17-17-17-17-25 = t208

770       780       790       800       810       820       830       840       850       860       870       880       890       900       910
CAAAATTGCTGCAGATAACAAATTAGCTGATAAAAACATGATCAAACCTGGTCAAGAACTTGTTGTTGATAAGAAGCAACCAGCAAACCATGCAGATGCTAACAAAGCTCAAGCATTACCAGAAACTGGTGAAGAAAATCCATTCATCGGTAC
```

(*example of output from a program that identifies the spa type from a complete sequence of the spa gene*)

The information that you have to work from is:

A list of the possible spa types
A list of the repeat sequences
A list of strings identifying the start and stop sequences either side of the spa repeats in the genome
The short reads from the whole genome sequencing (in FASTQ format). This includes a string indicating the quality of the read for each base pair. The reads are also 'paired'. Paired reads are obtained from sequencing the same short fragment of the original genome from both ends. These fragments are approximately 300 base pairs long but this length is not consistent and will vary among the paired reads. The coverage of the genome (i.e. the number of reads that include a particular position in the genome) will average 40-60 x.

This project originated from Dr Mark Holmes at the Vet School in Cambridge arising from collaborative work performed at the Wellcome Trust Sanger Centre. It is a real problem for which a solution is required. It is reasonable for the program to fail to identify a spa type for a genome (at a low rate) but wrongly identifying the spa type is to be avoided. This requires some sort of the estimate of the confidence of each solution.

Dr Holmes is available for help/advice/supervision and would be happy to answer any questions (mah1@cam.ac.uk). We have a collection of approximately 200 genomes for which the spa type is known.