ASSEMBLING AND ANALYSING A BACTERIAL GENOME USING A COMBINATION OF SANGER CAPILLARY DATA AND 454 SHOTGUN DATA

Mandy Sanders, Stephen Bentley, Ian Goodhead, Pathogen Sequencing Unit
The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA. U.K.

Introduction

Streptococcus suis SC84 is the first bacterial genome at the Wellcome Trust Sanger Institute to be co-assembled using Sanger capillary shotgun data and 454 shotgun data. We have previously finished the sequence of *Streptococcus suis* P1/7 using fluorescent di-deoxy DNA sequencing and subsequently used *Streptococcus suis* P1/7 as a model to test new sequencing technologies. Streptococcal genomes are a favoured organism to test new technologies on due to a severe cloning bias in conventional fluorescent di-deoxy DNA sequencing. The shotgun of *Streptococcus suis* SC84 has been achieved using 19 x 454 GS20 shotgun coverage combined with 5 x shotgun Sanger capillary data. After repeat tagging and contig ordering, using *Streptococcus suis* P1/7 as a reference strain, two different assemblies have been compared using ACT (Artemis Comparison Tool).

Comparison of the two assemblies against *S.suis P1/7* in ACT

Assembly using phrapped capillary data as the backbone with de novo 454 data assembled on top = **SC84_454**

Finished ***S.suis P1/7*** capillary sequence

Assembly using mapped 454 data (mapped against *S.suis P1/7*) as the backbone with capillary data assembled on top = **SC84_454_MAP**

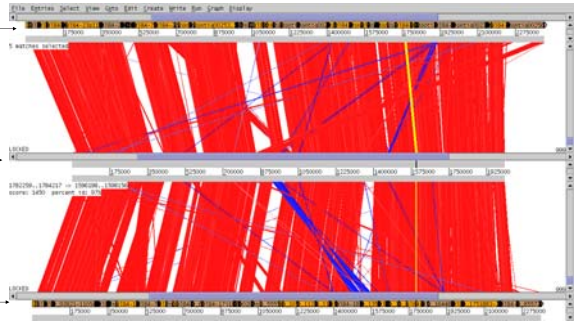

FIG.1

Differences between the two assemblies

On comparing the two merged assemblies in ACT with the reference strain *S.suis P1/7* (FIG.1) very few differences were found. There were gaps:

1)Over repetitive regions
2)Where there was bad quality capillary sequence
3)Where the 454 data stops
4)Where a chimeric capillary read joins onto 454 data

plus some misassembled repeats. A large proportion of the gaps over repetitive regions occur at the same position in both assemblies but the other gaps tended to vary in position but appear to be random. Because the two assemblies contained so few differences we proceeded with the capillary and de novo 454 data assembly (SC84_454) so that no biases were carried over from the *S.suis P1/7* assembly.

Assembly Information for *S.suis SC84*

Sanger ABI3730 capillary assembly = 218 contigs > 2kb
SC84_454 assembly = 39 contigs > 2kb
SC84_454_MAP = 39 contigs > 2kb

View of 454 data in Gap4

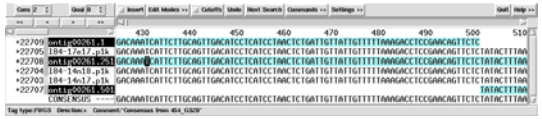

FIG.2

Closing the Gaps

454 data can be viewed in Gap4 in the form of a consensus read shredded into 500 base pieces which then overlap at 250 base intervals (FIG.2) The bases are given a nominal quality value of 30.
Work has been done to close gaps by cutting back bad quality data and chimeric reads and this resulted in an assembly of 15 contigs >2kb with no finishing reactions having been carried out (FIG.3).
PCRs have been done to close the remaining gaps which are mainly repeats.
Internal low quality capillary data has been cut back.

Comparison of *S.suis SC84* to *S.suis P1/7* in ACT

S.Suis SC84 sequence before any finishing reactions have been carried out

Finished ***S.suis P1/7*** capillary sequence

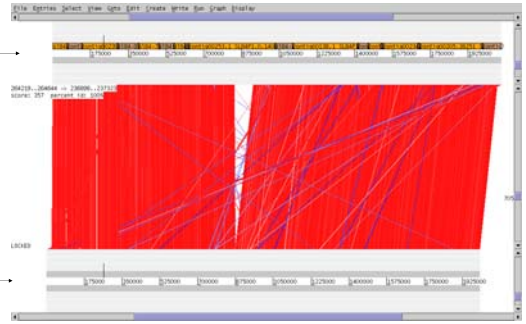

FIG.3

Analysing the Data

Once *S.suis SC84* is contiguous it will be compared in ACT to *S.suis P1/7* and frameshifts and SNP's will be identified in gene coding regions.
PCR's will be carried out to confirm if these differences are real in 454 only regions.
454 regions will be tagged.
The resulting data will be presented as "a contiguous draft".