Comparative sequencing of eukaryote genomes: experiences with three Leishmania genomes

David Harris, Kathy Seeger, Lee Murphy, Chris Peacock and the Pathogen Sequencing Unit, The Wellcome Trust Sanger Institute, The Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK

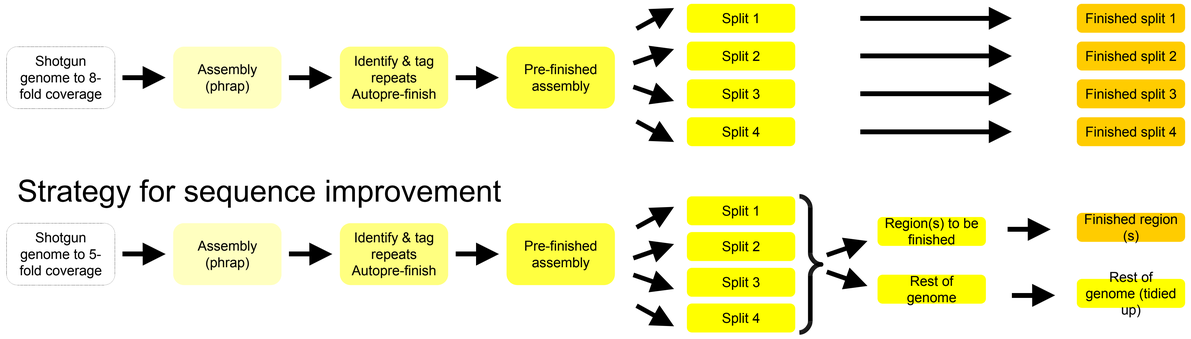
Objective

To produce genome sequences from two Leishmania species in a cost effective way for comparison with the high quality finished sequence of *L. major.*

Introduction

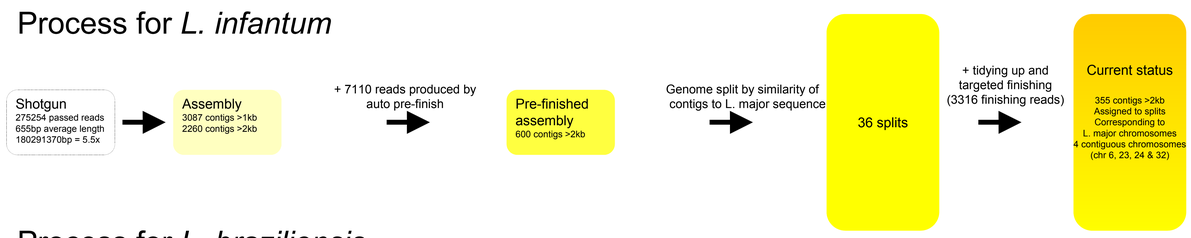
The genome of *L. major* consists of 33Mb in 36 chromosomes. Between 1998 and 2005 the WTSI as part of an international collaboration, finished 27.8Mb from 30 chromosomes. This work involved two main strategies; clone by clone and shotguns of libraries enriched for chromosomes by preparative PFGE. The completed genome sequence (published in *Science*, July 2005) has no gaps and is of finished quality. Two comparative Leishmania genome projects have been performed with modified wgs strategies. These projects were more challenging than expected but provided valuable experience for similar projects planned for the future.

Typical strategy for finishing small eukaryote wgs projects

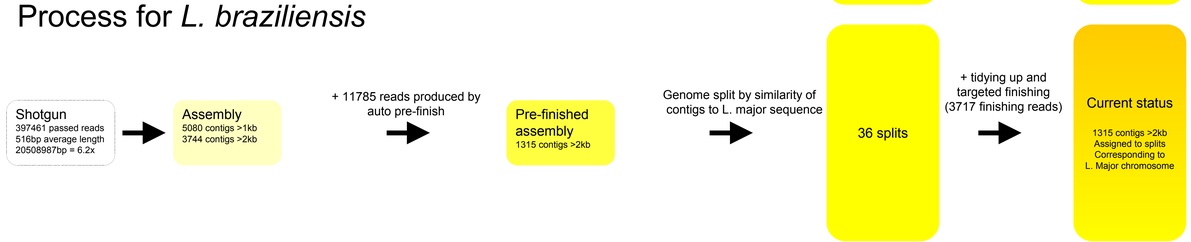


Strategy for sequence improvement

Process for *L. infantum*



Process for *L. braziliensis*



Tidying up = looking for joins that the assembler missed, breaking mis-assemblies, moving contigs between splits when they were mis-assigned BUT no more reactions

Targeted finishing = confirming the sequence of "interesting" genes, checking possible synteny breaks

Problems

Not possible to define the final product at the start of the project

- much less clearly defined than "finished"

Projects needed much more management than typical finishing projects

- decisions on the project objectives needed frequent discussions between annotator (whose job was to define the objectives of the project) and finisher (whose job was to explain the limits of the data)
- possible to draw incorrect conclusions from inadequate data

What software should we use?

- much of the software (phrap) was optimized for high redundancy shotguns
- assembly viewer (gap4) was not compatible with annotation tool (Artemis), this led to difficulties when the annotator and finisher needed to share data

How do we describe the final product?

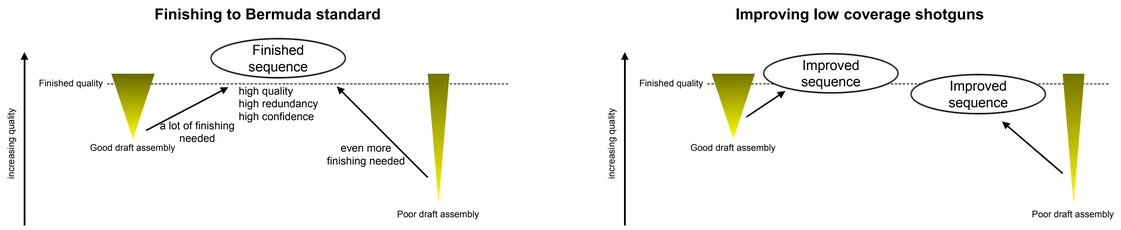
- no clear criteria for improved quality and so it is difficult to give a measure for how reliable the conclusions drawn from the data are
- best seems to be by comparisons (e.g. final contig numbers) or as shown below:

	<10	<20	<30	<40	<50	<60
L. major	1 in 26780	1 in 5013	1 in 2450	1 in 274	1 in 104	1 in 75
L. infantum	1 in 1341	1 in 483	1 in 218	1 in 108	1 in 43	1 in 19
L. braziliensis	1 in 243	1 in 144	1 in 104	1 in 71	1 in 39	1 in 21

Table 1. Frequency of consensus bases with gap4 quality scores less than values shown. For example, our final sequence fo

An important observation

The extent of possible improvement is very dependant on shotgun data quality



Suggestions for future projects

3. Make sure that shotgun quality is very high (must include very high fosmid coverage to scaffold contigs)
4. Put a lot of effort into optimising the assembly
5. Put effort into identifying repeats
6. Perform auto pre-finish to close gaps
7. Aim to produce chromosome scaffolds (using fosmid scaffolds and similarity to sequenced genomes)
8. Put a lot of effort into identifying interesting regions
9. Finish interesting regions to Sanger standard