ABACAS: Algorithm Based Automatic Contiguation of Assembled Sequences

Sammy Assefa[1,*], Thomas D. Otto[1], Thomas M. Keane[1], Chris Newbold[1,2] and Matthew Berriman[1]

[1] Pathogen Genomics, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SA, UK

[2] Institute of Molecular Medicine, University of Oxford, John Radcliffe Hospital, Oxford, OX3 9DS, UK

[*] sa4@sanger.ac.uk

1. Introduction

There is increasing interest in sequencing the genomes of strains that are closely related to existing finished reference genomes. Recently a number of *de novo* and mapping based assemblers have been developed to produce high quality draft genomes from second generation sequencing technology reads. However, a significant amount of manual effort is still required to go from the point of having a draft assembly to obtaining a fully contiguated genome. ABACAS is intended as a tool to rapidly extend contigs into larger scaffolds by, first, aligning, ordering, and orientating contigs against a reference sequence and, second, designing primers to close gaps. The input to ABACAS is a set of contigs which will be aligned to the reference genome, ordered and orientated, visualized in the ACT comparative browser, and optimal primer sequences are automatically generated. We present its application to the wider research community on a number of eukaryote and prokaryote genome projects. In particular we have tested it on the following eukaryotic genome projects: *Babesia bigemina* (4 chromosomes), *Trypanosoma vivax* (11chromosomes) and *Plasmodium berghei* (14 chromosomes). Taking the results of the *P. berghei* chromosome 2 finishing effort as a quantitative example, the number of contigs was reduced from 60 to 36 with 9 potential joins via ABACAS. Forty-six PCR products were generated to close gaps and 38 of these were successful in closing gaps in the assembly.

2. Methods

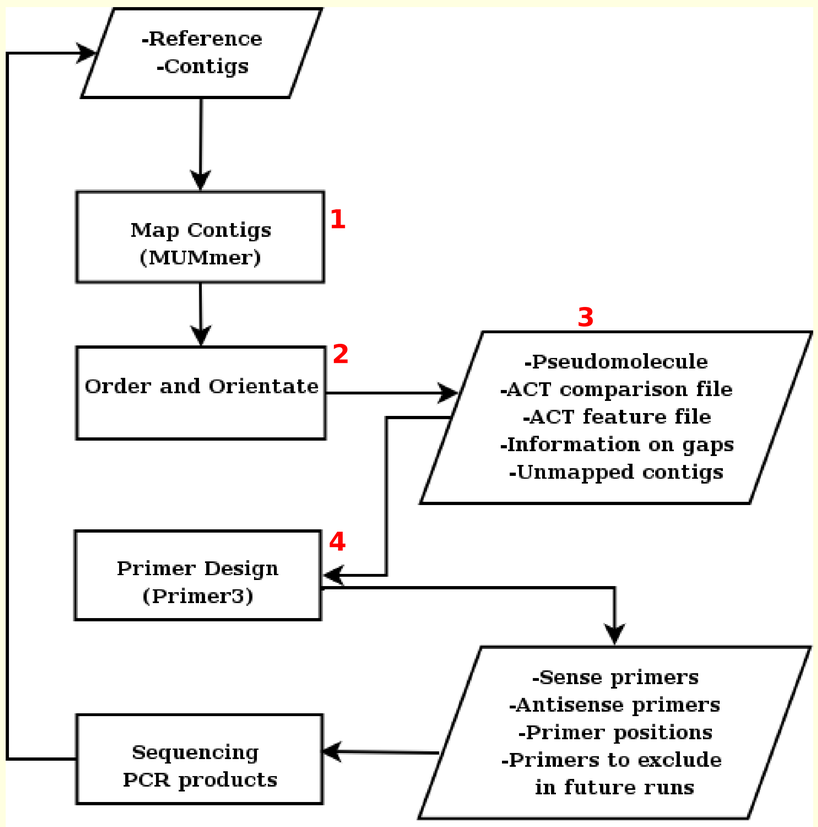

Fig. 1: A flow-chart describing the pipeline implemented in ABACAS

1. It uses MUMmer [1] to find alignment positions and identify areas of synteny.
2. Overlapping contigs and gaps taken into account while generating a pseudomolecule
3. Default output files.
4. ABACAS automatically extracts gaps on the pseudomolecule and generates primer oligos for gap closure using Primer3 [2] considering base quality (if provided).

- Uniqueness of candidate primer sets is checked.
- Comparison file can be used to visualize ordered and oriented contigs in ACT, the Artemis Comparison Tool [3].
- Contigs that were not mapped can be included to the pseudomolecule.
- Repetitive regions in the reference can also be identified and visualized in ACT alongside quality of the contigs.
- If contigs are not mapped, there is an option to run tblastx [4] on contigs that are not included in the pseudomolecule using sequences from the reference that correspond to the gaps.
- Additional contigs to the pseudomolecule can be dragged and dropped to the correct position using ACT.

3. Results and Discussion

- ABACAS has already been used on a number of eukaryote and prokaryote genome projects at the Wellcome Trust Sanger Institute.

Table 1. Using ABACAS on a number of genome projects

Project(Species)	**#Contigs**		**PCR**	
	Total	**Placed**	**Total**	**Successful**
Ecoli K88	925	574	-	-
Ecoli K99	397	297	-	-
Yersinia enterocolitica – 5603	2766	2402	-	-
Yersinia enterocolitica – 1203	394	293	-	-
Yersinia enterocolitica – 14902	500	394	-	-
Yersinia enterocolitica – 21202	489	323	-	-
Yersinia enterocolitica – 5303	364	269	-	-
C. difficle cdbi1	37 reduced to 10		-	-
C. difficle cd196	6 joined to 1		-	-
			-	-
P.Berghei Chromosome 2	60	36 (9 joins)	46	38
Pf IT Finishing (>2k contigs)	2611	2496	-	-
Pf 3D7 contigs	23356	21795	-	-

- At the Sanger institute, ABACAS has been included in the production pipeline to automatically finish draft assemblies.
- Ongoing work includes improving mapping of contigs in highly divergent species.

4. Examples

4.1. *P. berghei* contig ordering

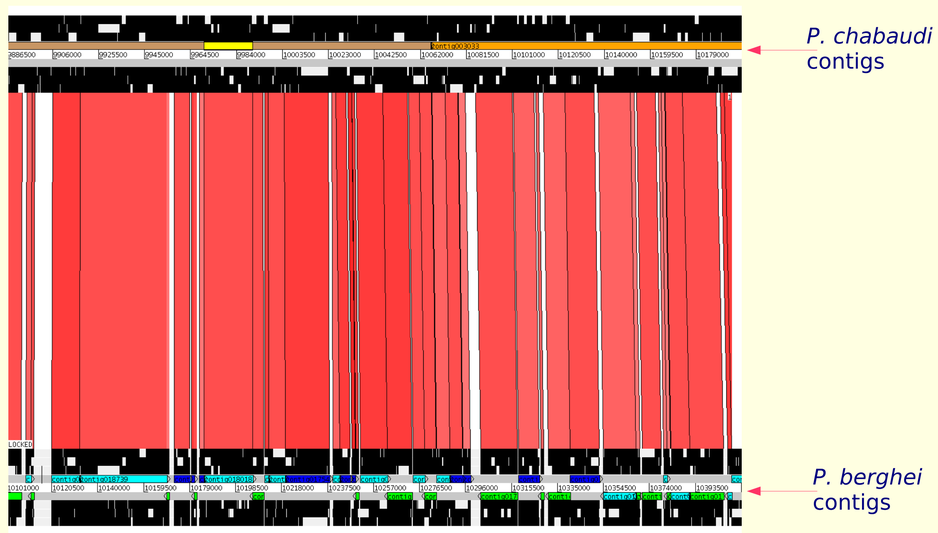

Figure 2. ACT view of *P.berghei* ANKA contigs ordered against *P.chabaudi* AS

- Stop codons are shown in *black ticks*.
- White boxes represent open reading frames.
- Contigs mapped to the forward and reverse strands are shown in blue and green respectively.
- Cyan boxes represent overlapping contigs.

4.2. *Pf 3D7* new technology contigs

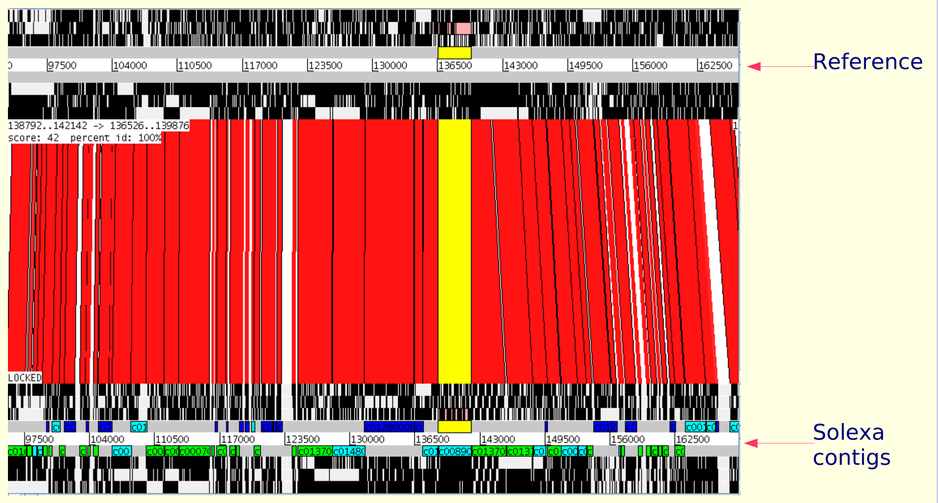

Figure 3: Artemis view of 3D7 FuzzyPaths Solexa contigs ordered against 3D7

- Of 23356 contigs 21795 could be mapped.
- It can be seen that the marked contig (yellow) is overlapping with the next contig.

4.3. *Pf IT* genome finishing

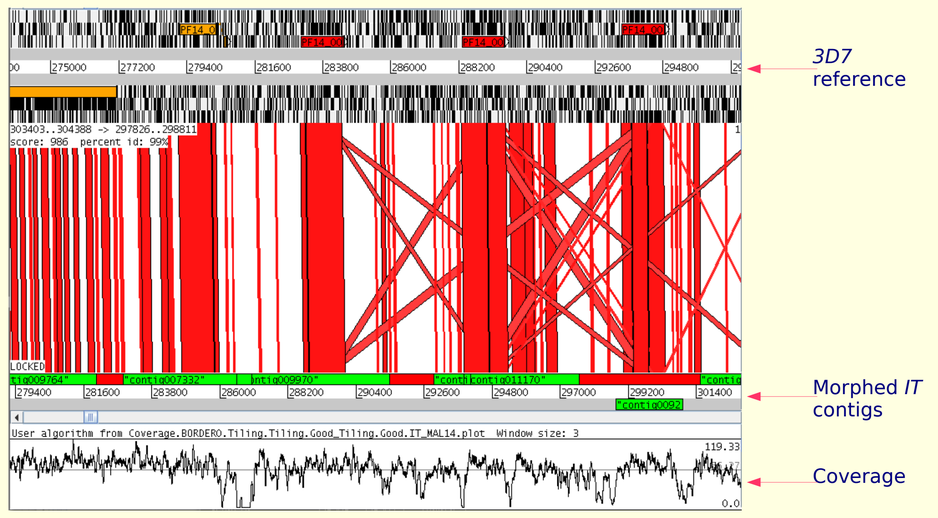

Figure 4: Chromosome 14 of morphed *Pf IT* with capillary contigs compared to *Pf 3D7*

- The morphed *IT* was obtained by CARMA (Otto, unpublished). By iteratively mapping IT Solexa reads on 3D7, and correcting SNP/indels, 3D7 was transformed (or morphed) toward IT. Where sufficient IT Solexa reads coverage (>=5x) exits, the sequence is now IT.
- The capillary contigs were included include in the morphed IT using ABACAS.
- This new sequence is compared to *3D7* with ABACAS.
- The lower graph shows the coverage of the mapping Solexa reads of *IT*.

5. Implementation and Availability

- ABACAS is implemented in Perl and is freely available for download from **http://abacas.sourceforge.net**
- It requires MUMmer and (optionally) blast for mapping, and Primer3 for primer design.
- Can be used in iterative process of contiguating a genome sequence.
- The output files produced after each run can be fed back into the program as input.

6. References

[1] Kurtz S et al. (2004) Versatile and open software for comparing large genomes. *Genome Biology*, **5**:R12.
[2] Triinu Koressaar , and Maido Remm (2007) Enhancements and modifications of primer design program Primer3, *Bioinformatics*, **23**:1289-1291.
[3] Carver T, et al. (2008) Artemis and ACT: viewing, annotation and comparing sequences stored in relational database. *Bioinformatics*, **24**(23): 2672-2676.
[4] Altschul, S.F. Et al. (1990) Basic local alignment search tool. *J. Molecular Biology*, **215**:403-410.

7. Acknowledgements

We would like to thank Andrew Berry, Mandy Sanders, and Danielle Walker of the Pathogen Genomics group at the Wellcome Trust Sanger Institute who provided feedback on earlier versions of the program.
Funding: This work was supported by the Wellcome Trust [grant number WT085775/Z/08/Z]; and European Union 6th Framework Program grant to the BioMalPar Consortium [grant number LSHP-LT-2004-503578].