

# SUPERCONTIGS: A Contig Scaffolding Tool

Daniela Puiu, Ping Xu, J. M. Alves, Myrna Serrano and Gregory A. Buck  
Center for the Study of Biological Complexity, Virginia Commonwealth University  
 [{dpuiu,pxu,jmalves,mgsserrano,gabuck}@vcu.edu](mailto:{dpuiu,pxu,jmalves,mgsserrano,gabuck}@vcu.edu)

## Abstract

*SUPERCONTIGS is a genome-finishing tool that orders, orients and groups contigs based on clone pair information and an alignment with a related genome. The program can be used in the genome finishing and gap closing process. SUPERCONTIGS is a Perl script that runs on Unix-like systems. It is available at <http://www.vcu.edu/csbc/bccl/research-downloads.htm> as a copy-left public domain program.*

## 1. Introduction

Assembling together all the sequences based on their similarity is one of the major steps in any genome project. The basic assembly principle is that two sequences with overlapping ends probably come from the same region of a genome and can be assembled together in a contig. A number of programs that have been developed to solve this NP-complete problem: Phrap [2], Cap [4] and Arachne [3]; these programs have similar approaches: they search for overlapping sequences, merge them into contigs and generate a consensus. Clone pair and insert size information can be employed to further assemble the contigs into supercontigs.

Genome alignment is used to identify conserved genome regions, study gene homology and evolution. It can be done either at the DNA level or at the protein level and can be classified into local/global alignment.

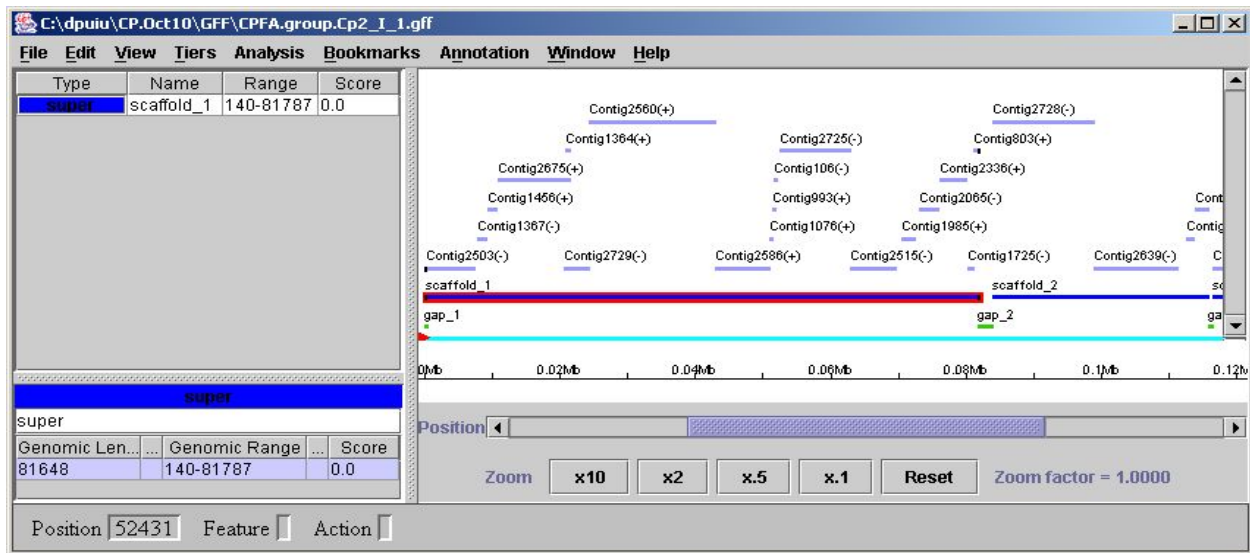
## 2. Overview

The main idea developed in SUPERCONTIGS is the combination of assembly and alignment results towards scaffold generation. This is done by locating clone pairs assembled in different contigs and based on their position in the contigs and the contig alignments to the related genome, compute their *absolute* alignment coordinates. If the clones are located within a maximum clone library insert size and have opposite

orientations, it is most likely that the contigs they belong to are adjacent and can be clustered together. The contigs are hierarchically clustered into scaffolds based on transitivity relation. The order and size of the scaffolds is computed, followed by an estimation of the gap types and sizes. The results are formatted and saved for analysis and visualization.

The SUPERCONTIGS tool is a Perl script that executes a number of steps. Each of these steps is an independent program. The intermediate files generated are saved for further examination. Described below are the main steps executed by the SUPERCONTIGS tool:

1. parse the alignment file and extract the start/stop coordinates for each alignment
2. parse the assembler output file and associate each read name with its corresponding contig
3. pair clones based on their base name
4. filter clone pairs, keeping only those pairs that are assembled in different contigs, within a maximum insert size from either contig end and are in the correct orientation
5. align the clones to the related genome
6. filter clone pairs, keeping only the ones that align within a certain maximum insert size
7. pair the contigs based on the filtered clone pairs
8. cluster all the contigs based on the transitivity relation and generate scaffolds
9. order and orient the contigs within scaffolds
10. order and orient each scaffold with respect to the related genome
11. estimate the physical gap sizes
12. estimate the sequence gap sizes based on the alignment and the size of the insert library
13. compute scaffold/gap statistics
14. generate GFF files containing the corresponding scaffold/gap information; these files can be visualized using the Apollo Genome Annotation Tool [5]. Figure 1 illustrates a sample output from the SUPERCONTIGS program.



**Figure 1. Two of the chromosome I scaffolds visualized from Apollo. The dark blue lines represent the scaffolds; the light blue lines represent the contigs, while the green lines stand for the physical gaps. scaffold\_1, the currently selected scaffold is highlighted in red. Details about the selected scaffold are given in the left pane of the window.**

The tool was initially developed for the *Cryptosporidium hominis* [9] genome assembly. *C. hominis* is an 8 chromosome intestinal parasite whose closest relative genome, *Cryptosporidium parvum* [8] is almost complete and was used for alignment.

### 3. Results

The SUPERCONTIGS approach and performance has been compared to the one of Bambus [6] developed at TIGR [7]. The programs are similar in the way they group contigs together based on clone pair information. However SUPERCONTIGS requires that the clone pairs align to the related chromosome and considers evidence contigs with multiple alignments.

**Table 1. SUPERCONTIGS and Bambus results**

Program	Supercontigs	Bambus
Number of scaffolds	305	658
Min number of contigs in the scaffolds	1	1
Max number of contigs in scaffolds	57	31
Min scaffold span	83	251
Max scaffold span	1,276,985	316,934
Physical gap no.	226	-
Min physical gap size	4	-
Max physical gap size	7,770	-

As input were used the 1,573 Phrap-generated *C. hominis* contigs and their nucmer [1] alignment to the *C. parvum* genome. SUPERCONTIGS has generated less than half the number of scaffolds Bambus has generated; the scaffolds contain more contigs and their span is larger.

### 4. References

- [1] Delcher, A.L. *et al.*, "Fast Algorithms for Large-scale Genome Alignment and Comparison", *Nucleic Acids Research*, 30, 2002, pp. 2478-2483.
- [2] Green, P., Phrap <http://www.phrap.org>, 1999.
- [3] Jaffe, D. B., "Whole-Genome Sequence Assembly for Mammalian Genomes: ARACHNE 2", *Genome Research* 13, 1999, pp. 91-96.
- [4] Huang, X. and Madan, A., "CAP3: A DNA Sequence Assembly Program", *Genome Research* 9, 1999, pp. 868-877
- [5] Lewis, S.E. and all, "Apollo: a sequence annotation editor", *Genome Biology* 3, 2002.
- [6] Pop, M. and all, "A hierarchical approach to building contig scaffolds", Second Annual RECOMB Satellite Meeting on DNA Sequencing and Characterization. Stanford University, 2002.
- [7] TIGR Web Site, <http://www.tigr.org/software>
- [8] UMN C Parvum project, <http://www.cbc.umn.edu/ResearchProjects/AGAC/Cp>
- [9] VCU C Hominis project, <http://www.parvum.mic.vcu.edu>