Recent Developments in Finishing Strategies for BAC-based and Medical Sequencing

J. Gupta[1], B. Barnabas[1], S.Y. Brooks[1], A. Young[1], N.F. Hansen[2], NISC Comparative Sequencing Program[1,2], G.G. Bouffard[1,2], E.D. Green[1,2], and R.W. Blakesley[1,2]

[1]NIH Intramural Sequencing Center (NISC) and [2]Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA

Abstract

We continue to refine our sequence-finishing procedures in order to improve efficiency and expand capabilities. First, we investigated potential modifications of sequence-assembly programs to reduce the amount of manual correction required with initial Phrap assemblies of BAC sequences. In an examination of 102 BAC sequences with misassemblies, alternative assembly-program routines completely resolved the problems in 29 BACs, while another 38 BAC sequences showed a reduction in the severity of the misassembly. Investigation of alternative assembly routines is now being applied to any Phrap-based assembly showing misassembled sequences. Second, we have found that the use of a plasmid copy-control strain of *E. coli* improves the uniformity of sequence-read distribution across assembled BAC sequences. Furthermore, the frequency of uncaptured sequencing gaps due to 'unclonable' DNA fragments is dramatically reduced when such a strain is used. This has proven especially helpful for sequencing BACs that yield assemblies with higher numbers of gaps, such as those derived from platypus, owl monkey and hedgehog. Finally, as our scientific portfolio expands to include major initiatives in medical sequencing, we are refining our capabilities in sequence finishing. Towards that end, we are adapting our traditional tools to improve the sequence data being generated in this context.

NISC website : http://www.nisc.nih.gov

Alternative Sequence Assemblers

Approximately 15% of our BAC-based shotgun sequencing projects appear 'misassembled' after initial read assembly by Phrap. The misplaced sequence reads result from one or more of the following: high quality discrepant joins, low quality misjoins, collapsed repeats, satellite repeats, or >2-copy repeats. Such misassemblies cause us to make time-consuming and labor-intensive manual corrections. In order to reduce finisher effort, we examined changes to our standard Phrap assembly routines. Here are results from two alternative assembly routines using Autosort and rPhrap (ver. SPS-4.24.0).

Characteristics of Alternative Sequence Assemblers

rPhrap
- "Mate-pair aware" version of Phrap.
- Allows merging based on percent identity (-M1 option) or length of match (default).
- Handles reads from fosmid libraries.
- Developed and distributed at Geospiza by Todd Smith.

Autosort
- "Mate-pair aware" wrapper for Phrap.
- Determines incorrect read matches and prevents Phrap from using them to join contigs.
- Developed by Nancy Hansen (NHGRI), but not distributed.

Improvement of Sequence Assemblies

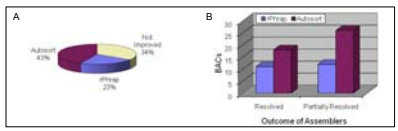


Figure 1. Sequence assembly improvement using alternative assemblers. A) Sample of 102 misassembled BACs were reassembled with rPhrap and Autosort. The resulting assemblies were then analyzed and the most improved selected for continuation of finishing. The altered Phrap assemblies often provide a better starting point for finishing. B) The 66% of BACs showing fewer or less-complex misassemblies were scored as either reduced (Partially Resolved) or completely absent of misassembly (Resolved). These re-assemblies provided a better starting point for manual sorting.

Example of Improvement Using Autosort

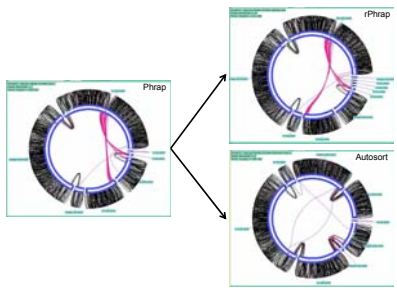


Figure 2. Graphical display (Orchid) of original and alternative BAC assemblies, where cluster of red lines indicate incorrect orientation/spacing of read pairs.

Example of Improvement Using rPhrap

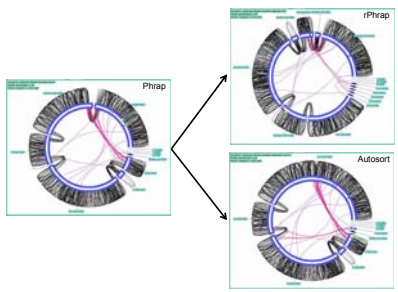


Figure 3. Graphical display (Orchid) of original and alternative BAC assemblies, where cluster of red lines indicate incorrect orientation/spacing of read pairs.

Summary

- Alternative assemblers often provide a better starting point for misassembled clones.
- Time taken to manually sort misassemblies is reduced.
- Either alternative assembly routine is useful as a secondary assembly option.

Alternative Competent Cells for Subcloning

Phrap assemblies of BACs derived from either certain species or particular genomic regions had a higher than average number of contigs. Up to half of the sequence gaps per BAC were not captured by spanning subclones, making it more difficult to order and orient all contigs of these BACs. Plasmids containing missing sequences were recovered by lowering their copy number during growth for DNA purification.

Characteristics of Copy-Control Competent Cells*

- CopyCutter EPI400 *E.coli* cell line was derived by manipulating the host *pcnB* gene to control plasmid copy number.
- Copy control is cell-based, not plasmid-based, maximizing applicability.
- Able to clone AT- and GC-rich sequences, secondary structure sequences, and inserts that inhibit cell growth.
- Adaptable to a wide variety of high copy-number vectors.

*From D. Haskins, EPICENTRE

Example of Improvement Using Control Copy Competent Cells

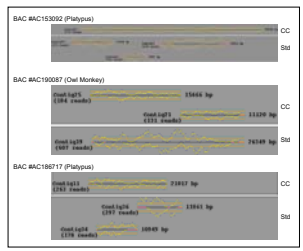


Figure 4. Variation in depth of reads derived from subclone libraries transformed in standard and copy control *E. coli* strains. Subclone ligation reactions from sheared BAC DNAs AC153092, AC190087 and AC186717 each were used to transform two strains of *E. coli*, standard (Std) and plasmid copy control (CC). Reads from shotgun sequencing to an average 8X coverage were separately assembled with Phrap for the six libraries. Each panel displays read depths for a portion of an assembly. For each BAC, identical sequences from each strain are aligned vertically. Yellow lines indicate depth of reads on the upper and lower strands for each contig (black line). A depth value of 10 is indicated by the orange line.

Table I. Plasmids from an *E. coli* copy control strain show improved shotgun read assemblies.

BAC Accession Number	Species	Read Depth Variation [a]		No. Contigs >2 kb		No. Uncaptured Gaps	
		Standard	Copy Control	Standard	Copy Control	Standard	Copy Control
AC190087	Cat	++++	-	18	13	6	2
AC153092	Platypus	++++	-	15	8	13	3
AC188899	Owl Monkey	++++	-	14	9	6	3
AC190076	Owl Monkey	++	-	22	15	9	1
AC182792	Shrew	++	-	22	11	11	0
AC175233	Shrew	++	+	21	22	9	10
AC186717	Platypus	++	-	20	10	7	3
AC188506	Rabbit	++	+	11	6	5	0
AC190002	Owl Monkey	++	++	10	6	5	4
AC187194	Platypus	++	+	9	12	7	3
AC172296	Hedgehog	+	-	21	13	9	0
AC190000 [b]	Owl Monkey	+	-	10	6	2	0
AC188356 [b]	Dusky titi	+	-	8	13	0	2

Note: [a] The variation in read depth over all contigs was assessed qualitatively as low (-), medium (+), high (++) or very high (++++). [b] BAC representing a typical assembly of average depth and number of contigs.

Summary

- Improved sequence representation in BAC subclone library.
- Increased efficiency in ordering and orienting contigs
- Cost was comparable to DH10B cells, but cells yielded fewer colonies and less DNA for sequencing
- Using copy-controlled fosmids also increased recovery of larger, difficult-to-clone DNA inserts.

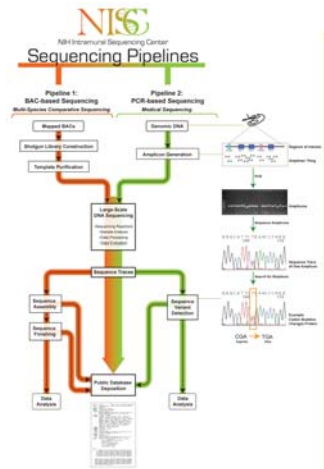
Recent Experiences with New Species

Table II. Recently sequenced BACs that presented finishing challenges.

Species		Observation
Small Brown Bat		High frequency of uncaptured gaps; dinucleotide repeats throughout clone; high GC content in gaps.
Tenrec		High frequency of single-strand read coverage; average gap size ≥1 kb.
Shrew		High frequency of uncaptured gaps; average gap size ≥2 kb; CT-repeat rich.
Echidna		High frequency of uncaptured gaps; average gap size ≥2 kb.
Xenopus		Extra effort to close gaps; structural problems with hard stops at end of gaps.

Finishing in Medical Sequencing

We have developed a second DNA input branch to our high-throughput sequencing pipeline in support of PCR-based medical sequencing; see flow chart below. In many instances, it is important to develop as complete a data set as possible, that is, to generate sequence from every amplicon from all patient samples. We are adapting our finishing techniques to recover missing or low quality sequence data.



Examples of Sequence Finishing Challenges

Figure 5. Example of patient sequence with a deletion relative to reference sequence. Extension of this deletion through the primer region may be the cause of failure to recover sequence in other patient DNAs.

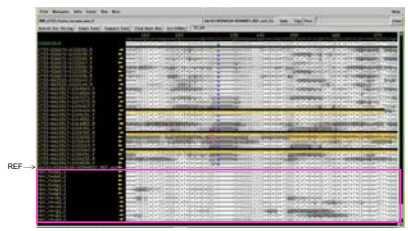


Figure 6. Example of improved quality of finishing reads compared to initial production sequence reads. Finishing reads are outlined by pink box. "REF" is reference sequence.

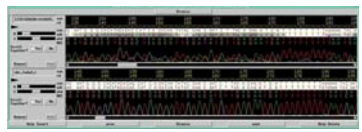


Figure 7. Traces from above figure demonstrating improved quality of finishing read (lower trace) over initial production sequence read (upper trace). Finishing reaction used a) DNA from a PCR reaction catalyzed by an alternative polymerase (ThermoStart – ABgene), and 10X DNA concentration.

Summary

- Establishing a pre-finishing protocol to address routine problems.
- Surveying a variety of enzymes and conditions to optimize PCR and sequencing.

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research