Supplementary Methods: Comparative-Grade Sequence Finishing

Blakesley RW, Hansen NF, Mullikin JC, Thomas PJ, McDowell JC, Maskeri B, Young AC, Benjamin B, Brooks SY, Coleman BI, Gupta J, Ho S-L, Karlins EM, Maduro QL, Stantripop S, Tsurgeon C, Vogt JL, Walker MA, Masiello CA, Guan X, NISC Comparative Sequencing Program, Bouffard GG, and Green ED: An intermediate grade of finished genomic sequence suitable for comparative analyses. *Genome Research* 14:2235-2244, 2004.

Introduction

We developed an experimental and computational approach for sequence finishing that is tailored for comparative analyses— we call the product of this process 'comparative-grade finished sequence.' Following assembly of highly redundant shotgun reads, the sequence is refined to substantially higher quality with a small amount of additional effort. The primary goal is to establish order and orientation of all significant (>2 kb) contigs in the assembly, then verify the results by separate analyses. Virtually all refinement and verification steps are performed *in silico*, avoiding the generation of additional sequence reads. Available auxiliary data are utilized when possible. This finishing process has been developed specifically for BAC-based shotgun sequence assemblies.

To achieve a consistent quality of product, all comparative-grade finished sequences must meet a set of specifications (Table I). The sequence is matured to these specifications by subjecting the initial assembled sequence to a series of steps. Many of these steps should be familiar to someone accustomed to sequence finishing, and they can be completed by application of now-standard computational methods.

Table I: Specifications for Comparative-Grade Finished Sequence

• Assembly is based on a minimum of 7-8X redundant coverage of Q20 bases
• All contigs >2 kb are ordered and oriented
• No significant misassemblies remain
• Both BAC ends are accounted for
• Questionable contig joins are removed
• Major sequencing artifacts and anomalies are corrected or removed
• Low quality/chimeric data is removed from contig ends • Contig order and orientation is established mostly by read-pair information • Contig order and orientation is verified by separate analyses

Procedure

Sequence reads are generated from both insert ends of randomly selected plasmid (shotgun) subclones derived from a BAC. These reads are assembled using Phrap. The initial assembly is subjected to quality control processing whereby: (a) contaminating reads are eliminated; (b) identity of the BAC is confirmed through PIPMaker-based comparisons to a reference sequence; and (c) redundancy level is determined to be a minimum of seven- to eight-fold in Q20 bases. Once these criteria are satisfied, the resulting full-shotgun draft sequence is submitted to GenBank and becomes the starting point for sequence finishing.

In contrast to the specifications established for finishing the human genome sequence ('human-grade finished sequence'), there is no minimum sequence quality (error rate) required for comparative-grade finished sequence (i.e., the average error rate can exceed 1 per 10,000 bases). Quality scores are submitted to GenBank together with the comparative-grade finished sequence so any user may understand the underlying quality and error rates of that sequence. To keep the quality scores as accurate as possible, sequence editing is minimized in order to avoid changing a consensus base to quality 98 or 99. Another difference with comparative-grade finished sequence (compared to human-grade finished sequence) is that the BAC sequence is not required to be represented by a single contig. Indeed, the sequence may be in many pieces, as long as the order and orientation of all contigs >2 kb is established.

For each BAC, the full-shotgun draft sequence is subject to the entire comparative-grade sequence-finishing procedure, including BAC assemblies with only one or two contigs. Any contig containing a BAC insert-end is included, regardless of its length (i.e., even those <2 kb are included). The consensus sequence is generated only from the shotgun sequence reads generated from the BAC. It does not include reads or consensus sequence (i.e., 'stolen reads or consensus') obtained from other sources, such as an overlapping BAC or a whole-genome shotgun assembly. However, sequence from a neighboring BAC is often used in the verification of contig order and orientation, while consensus sequence from a whole-genome shotgun assembly can be used as a guide to untangle misassembled sequence. A flowchart of the major procedures used to generate comparative-grade finished sequence is provided in Figure 1. There are three major stages of this process usually performed by a single finishing technician. A final stage of quality assessment and database submission is then performed.

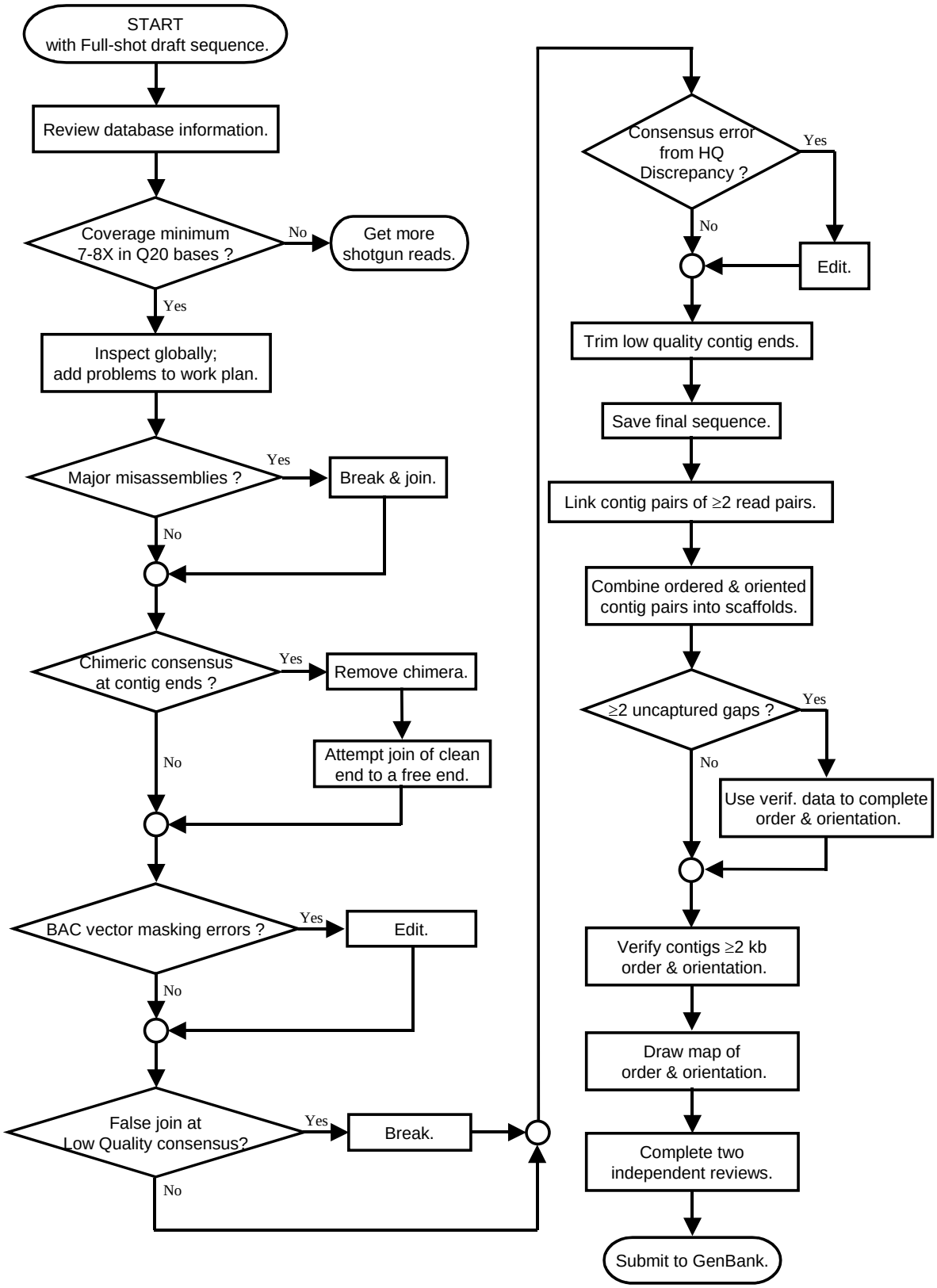

Figure 1: Flowchart Summarizing the Comparative-Grade Sequence Finishing Process.

Stage 1. Editing Sequence

1. Review database information for the BAC and become familiar with details of the project.

   Information noted by the finishing technician should include species, genomic region, insert size based on fingerprint analysis, neighboring BAC(s), and any unusual characteristics or problems encountered during the processing of that BAC through the sequence-production pipeline. Such facts can be helpful, for example, to resolve apparent inconsistencies in contig order and orientation or to reveal repetitive regions collapsed on themselves. Also, a significant problem with the assembly may be suspected from large discrepancies with the estimated insert size calculated from the sum of assembled sequence contigs and that based on the sum of estimated sizes of restriction-digest fragments. Finally, the database information is useful for prioritizing a weekly schedule, where one dispenses with the easiest BAC projects first as predicted from prior experience with a species or genomic region.

2. Note whether or not a minimum of 7-8X sequence redundancy in Q20 bases has been generated.

   A minimum of 7-8X redundancy is required for comparative-grade finished sequence. Projects with lower initial redundancy are returned for generation of more shotgun sequence reads.

3. Inspect BAC assembly globally and develop a work list for the project.

   Inspect globally the sequence assembly using Consed Assembly View, Orchid, and/or Phrapview. Note areas of possible sequence misassembly or low quality/incorrect joins, and note the extent and uniformity of read coverage and the number of uncaptured gaps. Add these features to the work list for finishing this project.

4. Correct major misassemblies in the sequence assembly.

   Fixing misassembled sequences is the first sequence-editing step performed by the finishing technician. The initial assembly is inspected for major misassemblies using software that displays in overview format read-pair relationships for the entire BAC sequence— specifically, Consed Assembly View, Orchid, or Phrapview. Major misassemblies typically appear as a cluster of red lines, indicating read pairs whose separation is significantly different than the average subclone insert size or whose relative orientations are not pointing toward each other. Among many reasons, these errors can be caused by alignment of low-quality sequences or by chimeric subclones/reads incorrectly bridging two non-adjacent segments of sequence.

   Another type of major misassembly is seen as a region of unusual depth of sequence reads (multiples of the average fold redundancy), where very similar or repetitive sequences have inappropriately coalesced. Navigating in Consed by high-quality discrepancies is useful for identifying such misassemblies, followed by sorting the different versions of a repetitive sequence assembled on top of one another.

   In all of these cases, the finishing technician uses Consed to correct the misassembled contigs. Within the program, appropriate breaks of incorrect joins are made, chimeric reads inappropriately linking two regions are removed, or reads from very similar but distinct copies of repetitive sequences are separated into discrete groups. Joining of sequence fragments from separated contigs is permitted only when read-pair data fully support the new relationship; if

uncertainty exists, such contigs are not merged.

Minor misassemblies are not edited during comparative-grade sequence finishing. Identified by occasional red lines, these misassembled reads are typically scattered across the project rather than clustered together. Minor misassemblies frequently contain read(s) of low quality incorrectly assembled with other high-quality reads. The finishing technician should also ignore a chimeric read that piggybacked on correct sequence without causing a false join. Any misplaced or misassembled read(s) are considered minor if the consensus sequence is unaffected by their presence.

5. Inspect the first 1 kb of sequence at both the 5′ and 3′ ends of each contig within the Consed Aligned Reads window. If high-quality consensus sequence resulting from a chimeric read is found, then remove that read to its own contig and add it to file read_list.txt. Search the project for a contig end to join with the newly edited contig.

   A chimeric read, often the result of a subcloning problem or a sequencing reaction artifact, assembled into a contig may or may not require editing. A chimeric read assembled into a group of good reads but not affecting the consensus sequence is ignored. On the other hand, corrective action is taken when the consensus sequence is based on the secondary sequence of the chimeric read. A frequently encountered example is high-quality sequence based on a single chimeric read extending from the end of a contig. Removal of this offending read from the end corrects the consensus sequence. The presence of the chimeric sequence might have blocked the contig from merging with another contig. Following removal, manual joining of the 'cleaned' contig to another contig is attempted, thereby reducing the number of contigs in the assembly. If in an earlier step, a misassembled contig was broken, the resulting ends should be inspected for any chimera-driven consensus sequence. When present, each of these assembly artifacts can usually be seen by inspection of the first 1000 bases of each contig end.

6. Inspect the consensus sequence of each BAC end in Consed Aligned Reads window. Edit the consensus sequence that is a result of vector-masking errors.

   Sometimes the vector sequence within reads spanning a BAC vector-insert junction is incompletely masked. An erroneous consensus sequence occurs when the unmasked vector portion of the read is used. Determine which read generated the vector sequence in the consensus, and then edit (change to "X") each problematic vector base. Occasionally, vector masking continues beyond the junction and into actual insert sequence. In this case, change the "X" bases of the insert to the correct sequence using the highest quality read.

7. Navigate in the Consed Aligned Reads window by 'low-consensus quality' through the entire length of each contig. Identify and break the contig at any low-quality areas representing a false join. Consensus sequence of ten or more consecutive Phred quality zero (Q0) bases is assumed to represent a false join and should be broken.

   In general, low-quality consensus sequence is not edited, even when the finishing technician believes the data may be stronger than the quality score indicates. As long as each read resembles or confirms the consensus sequence, then no action is taken. On the other hand, a region is broken if composed of just low-quality reads without sequence agreement, and not linking higher-quality flanking sequences in the same reads.

8. Navigate in the Consed Aligned Reads window by 'high quality discrepancies' through the entire length of each contig. Where appropriate, correct consensus sequence base derived from a high-quality discrepant base.

   Though many high-quality discrepancies may be found, most are not edited. The general rule is do not edit a read, alter a chimeric read, 'N' out bases, and so forth. The finishing technician edits only where convincing discrepancies indicate an incorrect consensus base or region of sequence; consensus sequence is left unchanged in areas with ambiguous reads. The consensus sequence is not edited directly, as it will change the associated quality score. Rather, the offending read is removed from the assembly. When the consensus is based on a low-quality read, remove and re-insert a high-quality read to change the quality in the consensus to that of the read and not an artificial "99". Editing is allowed when consensus is derived from the sequence of a chimeric read, a simple repeat/homopolymer, or an improperly masked vector sequence. In the case of simple repeats, the longest version of the repeat unit is usually used, but only if it is from a high-quality read. The majority should rule in cases of high-quality discrepancies from multiple reads at a particular consensus base. A series of multiple high-quality discrepancies across a region of sequence may indicate an unresolved, problematic repetitive sequence, and the finishing technician should go back to step 4.

9. Trim low-quality sequence from both ends of every contig >2 kb.

   NISC uses a script based on the program mkTrimTable, which trims the consensus sequence from a contig end through bases of consistently low quality or a continuous series of "N's". The finishing technician inspects the consensus sequence at each end of each contig in Consed Aligned Reads window to verify that few low-quality bases remain. Errors in the sequence are subject to manual editing and re-trimming. Contigs that become <2 kb after trimming are not considered further.

10. Save the edited sequence in Consed as a new ace file. This sequence will form the basis for ordering and orienting the contigs in the next stage.

    NISC convention is to make the name of the final sequence file include " ace.0". No further editing or manipulation of the sequence is expected. If re-work is necessary, then the project is saved again as an ace.0 file to indicate the final sequence.

Stage 2. Establishing Order and Orientation of Contigs

1. For each captured gap, identify the relative orientation of the associated contig pair. Develop scaffolds joining contig pairs into an ordered and oriented map of the sequence contigs. The scaffolds can be viewed in the Consed Assembly View window or in Orchid. Only accept a contig relationship with two or more subclone read-pairs spanning the gap.

   Contigs are ordered and oriented in comparative-grade finished sequence based on the pairing of end reads from individual subclones. To establish a relationship between two contigs, at least two read pairs (with each member of each pair pointing towards the other) must connect adjacent contigs. These read pairs can be readily seen in Consed Assembly View or in Orchid.

2. If two or more uncaptured gaps remain, use data from a verification method in Stage 3 to establish contig order and orientation.

Often, sequence contigs cannot be completely ordered and oriented by the use of subclone read-pair information. Additional data generated in Stage 3 can be used in such cases. The overall contig map can be developed from a mixture of different data types, as long as there are no conflicts among the data and one consistent map can be derived. In addition, two different types of data are required for comparative-grade finished sequence— one for establishment of contig relationships and a second for verification.

Stage 3. Verifying Order and Orientation of Contigs

1. Confirm the co-linearity of the established BAC sequence with independently derived, overlapping sequence(s) from a neighboring BAC(s) [Overlap].

   Use the program Pal to compare the BAC sequence to that of any overlapping BAC(s). Both the graphic and text outputs should be examined for the relative order and orientation of contigs that have sequence in common with overlapping BACs. For verification purposes, the overlapping BAC sequence(s) must be either finished to a comparative-grade stage, reside in a single contig, or have read pairs that span any common gaps. It is also permissible to use an unfinished overlapping sequence when the presence of BAC ends in both BACs defines only one possible orientation of the sequence(s) in common.

2. Compare the generated BAC sequence to a suitable orthologous reference sequence, verifying the order and orientation of contigs [Alignment].

   Align the generated BAC sequence to an established reference sequence (e.g., human genome sequence from UCSC build) using the sequence-alignment program PIPMaker. Verify the long-range contig map using the output that displays order and orientation of the BAC contigs relative to the reference sequence.

   Comparisons to orthologous sequences have proven quite effective for verifying contig order and orientation. Comparing sequences from closely related species usually demonstrates co-linearity throughout most of the BAC sequence. Sometimes, there may be a short stretch of sequence deleted in one sequence, or there may be clusters of short repetitive sequences that obscure sequence alignment. The overall data in these instances are typically sufficient to determine contig order and orientation. Sequence comparisons are typically not useful for BAC sequences from species that are distantly related to the reference sequence; in such cases, little alignment is encountered (e.g., comparing chicken sequence to human sequence). In some instances, whole-genome sequences are available for that species or a more closely related species, and this can then be used to verify contig order and orientation.

3. Compare the pattern (i.e., set of sizes) of restriction fragments generated from the BAC in the laboratory to that established *in silico* using Consed Display Digests [Fingerprint]. All fragments >1 kb in size must be consistent with the generated BAC sequence, as based on the proposed contig map.

   Open the Consed Display Digests window, which shows the restriction fragment pattern calculated from the BAC sequence side-by-side with the data imported from a lab-generated gel image. All bands predicted from the sequence that are internally contained in a contig should align side-by-side with the corresponding bands in the gel image. Sort the text list of fragments

by position, focusing on the remaining unaccounted-for bands, which should be those extending from the end of one contig into its adjacent neighbor. The text shows the predicted positions of the ends in two potentially adjacent contigs and the distance between them (fragment size). If one adds ~500 bases (for an average sized gap or more to account for small contigs within the gap) to the calculated fragment size, then a band in the gel should match within ±10%. Occasionally, an actual restriction site may fall within a sequence gap, so two smaller bands within the gel will add up to the predicted spanning fragment. For restriction digest data to verify contig order and orientation, all fragments should be accounted for (i.e., gap-spanning as well as internal fragments). If there is an ambiguous result, then the restriction digest data is not used.

4. If any contigs remain unordered or unoriented, then attempt to generate amplicons between candidate adjacent contig ends by PCR using purified BAC DNA as template [PCR]. Multiple PCR assays are typically performed with primers paired up in a combinatorial fashion, with the resulting amplicons then sized by gel electrophoresis (using suitable size markers). Only assays yielding single, clear amplicon bands whose sizes agree with those predicted from the sequence should be accepted as evidence for verification.

   PCR is associated with many limitations (e.g., often yielding multiple amplicons, false-positive fragments, weakly amplified fragments, or no amplicon); as such, PCR-generated data is considered the least effective at establishing or verifying contig order and orientation. Nonetheless, it is often the only remaining option. In some cases, different PCR enzymes or reaction mixtures are used if standard enzymes and conditions give ambiguous results. Several attempts are usually made to obtain acceptable results. Multiple unspanned gaps within a BAC sequence can necessitate the use of a PCR assay for each possible combination of primers. NISC uses data from combinatorial PCR as evidence for verification if only one possible order and orientation of contigs is supported by the resulting data.

5. Finalize the order and orientation of all contigs and indicate that result (on the submission form) as a linear map from one BAC end to the other. Summarize data used to establish that contig map. Attach all data collected and generated during the finishing process.

Quality Control and Submission

1. A second, highly experienced finishing technician thoroughly reviews all of the data for each BAC project, and determines whether or not contig order and orientation is firmly established. If not, the project is returned to the originating finishing technician for correction.

   The reviewer utilizes all of the software tools and available data to quickly determine the validity of the BAC sequence and its associated contig map; in a simple project, this could take as little as 10 minutes. Experience aids in determining whether certain subjective decisions made by the finishing technician were correct (e.g., distinguishing a major versus minor misassembly, proper editing of a consensus base, and breaking a contig at a low-quality region). First, the overall assembly is viewed to get a sense of any unresolved problematic areas and to detect any new problems created by the finishing process. For example, in an effort to correct a misassembled sequence, broken contigs may have subsequently become incorrectly joined or sequence may have become unnecessarily fragmented into too many contigs. Next, the sequence is subjected to closer inspection for errors, such as those emanating from the incomplete masking of BAC vector at the insert-vector junctions, the incorrect interpretation of alignment(s) with the

reference sequence, the misinterpretation of an overlap with a neighboring BAC sequence, the improper trimming of contig ends, or the failure to remove a chimeric read used for the consensus. Additional finishing technician actions, such as correction of a misassembled sequence or separation of repetitive sequences into discrete units, are scrutinized for each of the several steps (ace files) performed.

In reviewing how contig order and orientation was established, the data is weighted in the following order (see Table II): (a) read pairs from two or more spanning subclones, (b) co-linearity with sequence from an overlapping BAC, (c) alignment with a reference sequence, (d) agreement between sequence-predicted restriction digest pattern and the lab-produced BAC fingerprint, and (e) the unambiguous agreement of the presence and sizes of PCR amplicons relative to that expected. When some verification data (Stage 3) is used for establishing contig order and orientation in cases of multiple unspanned gaps, the strength of the data is examined carefully. For example, in the case of distantly related species, the alignment of orthologous sequences may be suspicious. Further, a convincing set of PCR data is often difficult to generate.

Verification data is reviewed for confirmation of the proposed order and orientation of contigs. The strength of the data is weighted in the same order as that stated above. In the case of uncaptured gaps, the preferred data combinations for establishment and then verification are as follows (in order): (a) overlap with neighboring BAC(s) combined with (in order): reference sequence alignment, fingerprint comparisons, and PCR; and (b) reference sequence alignment combined with (in order) fingerprint comparisons and PCR.

Table II: Preferential Order of Data for Establishing the Order and Orientation of Contigs.

Establish Map	Establish & Verify Map with ≥2 Uncaptured Gaps
1. Two or more spanning subclones	1. Overlap & Alignment
2. Overlap with neighboring BAC sequence [Overlap]	2. Overlap & Fingerprint
3. Alignment with reference sequence [Alignment]	3. Overlap & PCR
4. Comparison of restriction digests [Fingerprint]	4. Alignment & Fingerprint
5. PCR	5. Alignment & PCR

2. After passing the first review, the BAC project is passed to a member of the Bioinformatics group for final review and submission to GenBank.

The GenBank record includes the following text in the comment section:

> ″The sequence data in this record represents an ‘enhanced’ version of a Phase 2 submission. Specifically, the indicated order and orientation of each sequence contig has been established using one or more of the following: read-pair data from individual subclones, overlaps with neighboring clones, alignment with available reference sequence (e.g., human), and/or confirmation by PCR testing. In addition, the sequence assembly is based on at least 8X average coverage in Q20 bases and has been reviewed to rule out gross misassemblies, the low-quality ends of sequence contigs have been trimmed away, and each base is associated with a Phrap-derived quality score.″

The second review includes a final check for completeness of the supporting data and confirmation that the information about contig order and orientation is clear. The sequence is then submitted to GenBank as Phase 2. Since the full-shotgun draft sequence was thoroughly

examined and edited en route to determining contig order and orientation, and then verified by additional separate analyses, NISC regards comparative-grade finished sequence as an enhanced version of HTGS_Phase 2 sequence. Finally, comparative-grade finished sequence represents a very good starting point for further refinement, such as to near perfect or human-grade finished sequence.