Use of Multi-BAC Assemblies to Improve the Dog Genome

Abouelleil, A., Aftuck, L., Arachchi, H., Berlin, A., Brown, A., Fitzgerald, M., Gearin, G., Johnson, J., Lui, A., Macdonald, P., Pirun, M., Priest, M., Shea, T., Zimmer, A.

Broad Institute of MIT and Harvard, Cambridge, MA 02142

Abstract

The dog genome is important for comparative analysis of mammalian genome biology and evolution. CanFam 2.0, the current dog draft assembly, covers approximately 99% of the euchromatic portion of the genome. Through the use of targeted finishing, the accuracy and integrity of this draft assembly can be significantly improved in an efficient and cost effective manner.

The use of multi-BAC contigs is central to the aforementioned improvement process. BACs and genome chunks (sections of WGA pulled for direct finishing work) targeting large gaps, ENCODE regions, and uncertified regions (areas of questionable integrity) are finished. Overlapping BACs and chunks are then identified as work regions, which are in turn integrated into multi-BAC assemblies. This important step minimizes the draft finished sequence interfaces during finished data integration into the genome assembly. These assemblies undergo a quality control step and are then patched into the genome assembly. Finally, the chromosome that is patched undergoes another QC analysis. Through this process, a near-finished quality genome can be produced without the expense of generating a shotgun BAC library for the entire genome. We will describe our process and results from generating these contigs.

Introduction

A near finished dog genome would serve as the third mammalian reference sequence, and the first belonging to Clade Laurasiatheria. Furthermore, the dog genome retains many of the features of the "ancestral" eutherian genome, making it ideal for comparative studies between mammals.
Dog is also important both as a model for human diseases such as cancer and as a commonly used organism in pharmaceutical clinical trials[1]. Hence, the importance of producing a near-finished quality dog genome can not be overstated.

DNA from Tasha, a female boxer, was selected for sequencing.

The current dog assembly, CanFam 2.0, is a high quality draft assembly with ~98% of bases at or above Q40, and 99.3% considered to be "certified." Additional regions for targeted finishing include "uncertified"* regions greater than 50kb in size and any fragment gaps greater than 35kb[1]. Finally, supercontig gaps and chromosome ends are also selected for finishing.

In order to achieve a near-finished quality genome, the six standards[1] that must be met have been specified below (table 1):

TABLE 1: Standards for Near-Finished Quality Dog Genome

1	No global misassembly, with 99.9% of the assembly certified
2	Greater than 99.5% coverage for the euchromatic portion of the genome
3	More than 99.5% of bases must be Q40 or greater
4	All genes with 1-1 orthologs with human fully and accurately represented with most gene family members resolved
5	Less than 10,000 gaps in the genome, and fewer than 4 gaps/Mb
6	Finished quality for ENCODE[2]** regions

A key step in reaching these goals is the integration of the finished sequence back into the whole genome assembly. Doing so for each BAC individually poses two problems. Neighboring BACs may represent different haplotypes, and it increases the number of exchange points in the AGP and therefore problems due to shifting coordinates. A multi-BAC assembly approach addresses these concerns by "normalizing" haplotype differences and minimizing the number of exchanges.

Methods

Targeted Finishing

Selection of regions for targeted finishing is performed by extracting coordinates from CanFam 2.0. These regions comprise approximately 80MB of the dog genome and have the following characteristics:

1. ENCODE regions
2. Uncertified regions
3. Internal gaps >10kb
4. Supercontig gaps and chromosome ends
5. Clusters of gaps 1-10kb in size

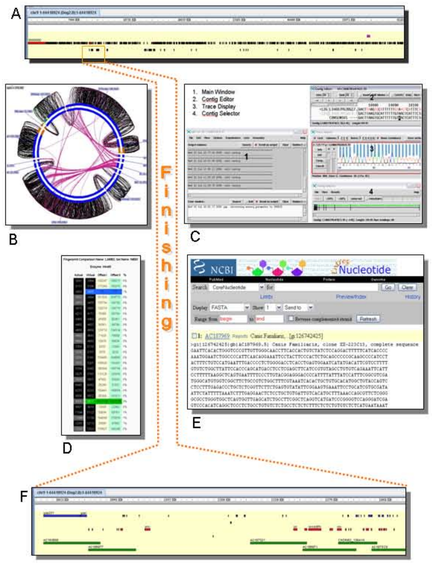

Fig. 1: Finishing workflow from target selection to sequence reintegration. A: CanFam 2.0 is scanned for regions exhibiting characteristics targeted for finishing (orange box) B. Circhid display indicating good 6-bad links in a misassembled project C: The four primary interfaces of the genome assembly progress (GAP4) D: Fingerprint fragment analysis comparison of a virtual gel of the finished BAC sequence with the actual BAC. E: FASTA display of finished dog sequence. F: Finished BACs covering targeted regions of the genome assembly.

BACs covering the selected regions enter the finishing process (fig. 1). In some instances, rather than finish a BAC, the genome assembly is extracted between coordinates and finished using fosmid-based strategies, such as fosmid primer walks. These regions undergo the same QC as BACs except for fingerprint analysis, and are not submitted to GenBank as phase 3 sequence.

Multi-BAC Assembly

Regions covered by several overlapping BACs and/or finished genome chunks are identified as work regions and processed for multi-BAC assemblies. These BACs and chunks are assigned to a project group to provide an identifier for the work region (Fig.3) and the consensus fasta files are used to generate representative reads for each member of the project group. These reads are then assembled into representative BACs. A complete depiction of the multi-BAC assembly process is illustrated in figure 4.

Multi-BAC Assembly QC

Completed Multi-BAC assemblies are reviewed by comparing a dot plot against the CanFam 2.0 assembly (see fig. 2). Coordinates and orientation resulting from the plot are compared with expected values. If the assembly orientation is not positive or the coordinates are discrepant, the multi-BAC assembly is returned to the finisher for subsequent analysis.

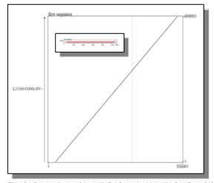

Fig. 2: Comparison of a multi-BAC assembly with CanFam 2.0. A: A dot plot of the assembly against chromosome 30 between 13.16 Mb and 13.17 Mb created using PIPmaker. B: The same sequence represented in a pip plot form.

Multi-BAC Assembly Integration

A patcher program is used to integrate the completed multi-BAC assemblies into the genome assembly and produce a single coherent consensus. The sequence is inserted into the WGA using the start and stop coordinates of the multi-BAC assembly. Note that the patched sequence does not replace the original WGA so as to preserve the underlying structure of the original WGA.

Fig 3: Sample of a project group listing for a multi-BAC assembly consisting of 4 constituent neighboring BACs.

Fig. 4: Flowchart depicting the multi-BAC assembly process. A: Four overlapping BACs are selected as a work region. B: The finished BACs consensus sequence is extracted to a fasta file, broken down into 3kb overlapping representative reads and assembled together into representative BACs. These representative BACs are combined into a multi-BAC assembly. B, insert: Gap4 contig editor window with 2 constituents of the multi-BAC assembly displayed. L35001 is the 'dominant' clone as it provides more sequence across the entire overlapping region. Therefore, it's sequence is represented at bases in which there are haplotype differences. C: The completed multi-BAC assembly resolves several uncertified regions in the original dog assembly with a minimum number of exchanges between the genome and finished sequence.

Chromosome QC

Once work regions are integrated, a new fasta is generated and aligned against the original CanFam 2.0 assembly. Dot plots and pips (such as those represented in fig.2) are used to verify coordinates or sequence shifts. Any discrepancies are corrected and the sequence is patched into the assembly again. The flexibility of the patcher software allows for iterative patching and correction of localized regions without the need to reintegrate the entirety of the updated sequence.

Results

To date, 27 multi-BAC assemblies have been finished, composing 13% of the total sequence targeted by the dog genome improvement project. These regions range from 200KB to 1.1MB in size (see fig. 6) for a total of approximately 12MB of total sequence. Analysis of haplotype differences in regions with overlapping BACs suggests a bi-modal polymorphism pattern (see fig 5).

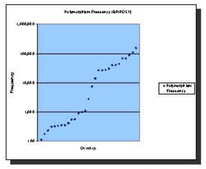

Fig 5: Polymorphism frequency (POLYBAP) amongst overlapping BACs in 27 finished multi-BAC assemblies.

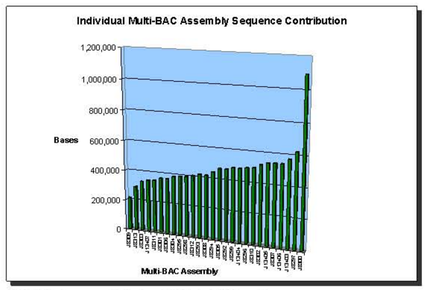

Fig 6: Individual sequence contribution (bases) of finished multi-BAC assemblies.

Table 2: Current Status of the Dog Genome Improvement Project

Goal	CanFam 2.0	Improvement Target	Current Status	% Goal Complete
ENCODE	H.Q. Draft	Finished	Finishing	Finished
euchromatic (% cov.)	99.2	>99.5	99.3	99.5
Certified Regions (% cov.)	99.5	>99.8	~99.7	99.86
Base Quality >40	~98%	99.50%	~98.6	99.56
Contiguity (gaps/MB)	12	<4	11.7	11
Gene Coverage	~5000	<100	TBD	TBD

Table 2, above, displays the current status of each targeted goal in the Dog Genome Improvement Project, and indicates that the majority of the goals are nearly complete. Table 3, below, demonstrates the contributions of the multi-BAC assemblies as compared to regular BACs.

Table 3: Multi-BAC Assembly Contributions***

	MBA (current)	MBA (projected)	Non-MBA	Total
Sequence	12MB	51MB	39MB	90MB
% of Total	13%	57%	43%	100%
Exchanges	27	98	201	299
Exchange Efficiency (KB/exchange)	440	520	190	

Conclusion

The use of the Multi-BAC assembly method to patch in large segments of finished sequence has made a significant contribution towards the Dog Genome Improvement Project at the Broad Institute. It currently accounts for 13% of the targeted sequence, and is projected to cover approximately 57% of the targeted region. Furthermore, this method reduced the number of exchanges into the WGA by 70%, thereby saving significant human time, effort and computer processing power.

Literature cited

1. Lindblad-Toh, K., Lander, E. Improvement of the Dog Genome to Near Finished Quality. <http://www.genome.gov/Pages/Research/Sequencing/SeqProposals/DogImprovementSeq.pdf>: 1-4.
2. The Encode Project: ENCyclopedia of DNA Elements, May 14, 2007, National Human Genome Research Institute, June 12, 2007. <http://www.genome.gov/10005107>.

Footnotes

* Uncertified regions are those areas of CanFam 2.0 that contain clustered read pair discrepancies and/or >2 haplotypes.
** Encyclopedia of DNA Elements is an NHGRI project launched to identify all functional elements in the human genome[2].
*** Numbers in this table are approximations, as the ratio of work regions to BACs is constantly adjusted. Also note, the 90MB total number includes 10MB of untargeted sequence that is incidentally captured by BACs and work regions.

Acknowledgments

Special thanks to the following people who provided data, images and general assistance with this poster:

Mike Fitzgerald
Jeremy Johnson
Andrew Zimmer

For further information

Amr Abouelleil
320 Charles Street
Cambridge, MA 02141-2023

email: amr@broad.mit.edu
tel #: 617-324-2396
fax #: 617-258-0901

Poster # FF0028