2006

Physical map-assisted whole-genome shotgun sequence assemblies

David Messina
Washington University School of Medicine in St. Louis

Shiaw-Pyng Yang
Washington University School of Medicine in St. Louis

Wesley C. Warren
Washington University School of Medicine in St. Louis

John W. Wallis
Washington University School of Medicine in St. Louis

LaDeana W. Hillier
Washington University School of Medicine in St. Louis

See next page for additional authors

Authors

David Messina, Shiaw-Pyng Yang, Wesley C. Warren, John W. Wallis, LaDeana W. Hillier, Asif T. Chinwalla, and Richard K. Wilson

Physical map-assisted whole-genome shotgun sequence assemblies

René L. Warren, Dmitry Varabei, Darren Platt, et al.

Genome Res. 2006 16: 768-775
Access the most recent version at doi:10.1101/gr.5090606

Physical map-assisted whole-genome shotgun sequence assemblies

René L. Warren,[1,6] Dmitry Varabei,[1,6] Darren Platt,[4] Xiaoqiu Huang,[3] David Messina,[2] Shiaw-Pyng Yang,[2] James W. Kronstad,[5] Martin Krzywinski,[1] Wesley C. Warren,[2] John W. Wallis,[2] LaDeana W. Hillier,[2] Asif T. Chinwalla,[2] Jacqueline E. Schein,[1] Asim S. Siddiqui,[1] Marco A. Marra,[1] Richard K. Wilson,[2] and Steven J.M. Jones[1,7]

[1]British Columbia Cancer Agency, Genome Sciences Centre, Vancouver, British Columbia V5Z 4S6, Canada; [2]Washington University School of Medicine, Genome Sequencing Center, St. Louis, Missouri 63108, USA; [3]Department of Computer Science, Iowa State University, Ames, Iowa 50011-1040, USA; [4]U.S. Department of Energy, Joint Genome Institute, Walnut Creek, California 94598, USA; [5]The Michael Smith Laboratories, Department of Microbiology and Immunology, The University of British Columbia, Vancouver, British Columbia V6T 2Z4, Canada

We describe a targeted approach to improve the contiguity of whole-genome shotgun sequence (WGS) assemblies at run-time, using information from Bacterial Artificial Chromosome (BAC)-based physical maps. Clone sizes and overlaps derived from clone fingerprints are used for the calculation of length constraints between any two BAC neighbors sharing 40% of their size. These constraints are used to promote the linkage and guide the arrangement of sequence contigs within a sequence scaffold at the layout phase of WGS assemblies. This process is facilitated by FASSI, a stand-alone application that calculates BAC end and BAC overlap length constraints from clone fingerprint map contigs created by the FPC package. FASSI is designed to work with the assembly tool PCAP, but its output can be formatted to work with other WGS assembly algorithms able to use length constraints for individual clones. The FASSI method is simple to implement, potentially cost-effective, and has resulted in the increase of scaffold contiguity for both the *Drosophila melanogaster* and *Cryptococcus gattii* genomes when compared to a control assembly without map-derived constraints. A 6.5-fold coverage draft DNA sequence of the *Pan troglodytes* (chimpanzee) genome was assembled using map-derived constraints and resulted in a 26.1% increase in scaffold contiguity.

The technical ease with which whole-genome shotgun sequencing (WGS) approaches can now be implemented has resulted in most mammalian genome projects including a substantial WGS component (Adams et al. 2000; Venter et al. 2001; Holt et al. 2002; Waterston et al. 2002; Gibbs et al. 2004; Margulies et al. 2005; Mikkelsen et al. 2005). This is largely because of the development of assembly algorithms capable of using paired-end sequence read information in the form of clone length constraints (Sutton et al. 1995; Huang and Madan 1999; Huang et al. 2003, 2006; Myers et al. 2000; Batzoglou et al. 2002; Jaffe et al. 2003; Mullikin and Ning 2003). Clone length constraints are supplied to WGS assembly programs as a set of permissible distances between the forward–reverse read pair of a single clone. Their use is crucial for resolving repeated sequences at run-time and permitting the construction of scaffolds by linking, ordering, and orienting sequence contigs, thus increasing the longer-range contiguity of the resulting assemblies. Within a given genomic library, clone inserts follow approximate Gaussian length distributions that correlate with both the quality of the library and the size of its clones. In an attempt to account for this, length constraints are often represented as a relatively wide range, but this potentially introduces incorrect joins and poor repeat resolution. Although ideal, it would be impractical to accurately size every single clone sequenced using current laboratory protocols.

[6]These authors contributed equally to this work.
[7]Corresponding author.
E-mail sjones@bcgsc.bc.ca; fax (604) 877-6085.
Article is online at http://www.genome.org/cgi/doi/10.1101/gr.5090606.

Construction of genome physical maps, however, generates bacterial artificial chromosome (BAC) insert sizes as part of the process. Physical maps are built by clustering together BACs or fosmids sharing portions of a DNA "fingerprint," a pattern of fragments of various sizes generated by restriction enzyme digestion of individual clones (Marra et al. 1997). A typical 10-fold coverage physical map of a 3-Gb mammalian genome is comprised of ~200,000 BAC clones, each with a known insert size approximated from individual clone fingerprints. Aside from providing the starting point for the clone-by-clone (CBC) genome sequencing approach (*C. elegans* Sequencing Consortium 1998; *Arabidopsis* Genome Initiative 2000; Lander et al. 2001), the availability of a physical map also provides an orthogonal tool for assembly assessment, both at the level of comparisons of individual clones to their fingerprints and for comparison of clone order and overlap (L.W. Hillier, pers. comm.). A physical map is also invaluable for validating and improving the contig layout of existing whole-genome shotgun sequence assemblies (WGA) (Warren et al. 2005). But the usefulness of the map for genome sequencing extends well beyond post-assembly validation. The information provided by physical maps has been used extensively to improve CBC assemblies in a variety of combined strategies (Kent and Haussler 2001; Choi and Farach-Colton 2003; Havlak et al. 2004; Hillier et al. 2004), but still remains unused within WGA algorithms.

We describe here an approach to improve the contiguity of PCAP (Huang et al. 2003) WGS assemblies at run-time, using BAC end length constraints and clone overlap information derived from a physical map. BAC clone sizes and overlaps are derived

16:768–775 ©2006 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/06; www.genome.org

from clone fingerprints and are used for the calculation of these constraints. We hypothesize that more accurate and overlapping length constraints between BAC end sequences (BES) can be used to promote linkage and guide the arrangement of sequence contigs within a scaffold at the layout phase of WGA. This process is facilitated by the Fingerprint and ASSembly Incorporation software (FASSI), a stand-alone application that calculates BAC end length constraints from clone fingerprint map contigs created by the FPC package (Soderlund et al. 1997, 2000; Ness et al. 2002). The output of FASSI is fully compatible with the PCAP assembly program and could be easily formatted to work with other assembly algorithms able to accept length constraints for individual clones as input.

Using FASSI, six assembly protocols were designed, each representing a different set of BAC end and BAC overlap length constraints, and supplied to PCAP during WGA of *Cryptococcus gattii* strain WM276, a fungal pathogen (Sorrell 2001). The two best assembly approaches resulting from this analysis, named FASSI-1 and FASSI-6, are presented here. The *Drosophila melanogaster* genome sequence (Adams et al. 2000) was assembled at variable depth of read coverage with and without the FASSI-1 and FASSI-6 BAC length constraints, and the resulting assemblies were compared to the finished genome sequence. The analysis shows that chromosomes are covered by fewer and larger scaffolds. Lower read-depth FASSI assemblies appear equivalent in contiguity to higher-depth WGAs constructed without map-derived constraints, and thus may result in a more efficient use of sequencing reads, making the approach potentially more cost-effective. A 6.5× coverage draft sequence of the *Pan troglodytes* (chimpanzee) genome was assembled at the Genome Sequencing Center, Washington University in St. Louis, using FASSI length constraints in conjunction with PCAP.REP (Huang et al. 2006). When compared to a control chimpanzee WGA without map-derived constraints, the use of FASSI length constraints led to a 26.1% increase in the scaffold N50 length (the length such that 50% of the assembled genome lies in scaffolds of that size or longer).

Results

C. gattii WGS assemblies

C. gattii strain WM276 was sequenced and assembled to approximately fivefold coverage at the BC Cancer Agency Genome Sciences Center (BCCAGSC; http://www.bcgsc.ca) in collaboration with James W. Kronstad at the University of British Columbia. Using a BAC-based physical map assembled and hand-edited on location at the BCCAGSC, we constructed two experimental WGS assemblies, each using a different set of BAC length constraints, named FASSI-1 and FASSI-6 (Fig. 1). FASSI-1 consists of four different sets of length constraints between any two overlapping BAC neighbors sharing >40% of their length (based on total length of shared restriction fragments). In addition to sizes derived from the fingerprints of each individual BAC clone, the FASSI-1 length constraints also include two pairs of length constraints between the forward and reverse sequence reads of two different, but overlapping BAC neighbors (Fig. 1A, wide overlap constraints). Constraints in FASSI-6 include the length constraints from the first approach, plus short overlap constraints between any two proximal BES from BAC sharing 40% of their length. Reverse-complemented BES are generated in order to produce these logical short overlap constraints between proximal BES (proximal BAC end sequences are reoriented to face inward, thus creating a logical pair) (Fig. 1B). The FASSI-1 and FASSI-6 sets include additional BAC length constraints for singletons (BACs that do not assemble with any other clones in the physical map), as well as buried constraints. Buried constraints are generated whenever a BAC clone is a subset of another (buried clone) (Coulson et al. 1986).

At 5× read depth, the control PCAP assembly yielded 610 scaffolds, 13 between 100 kb and 1 Mb and seven >1 Mb in size, gaps excluded. With FASSI-1, the addition of map-derived constraints had a modest effect, yielding 607 scaffolds, nine between 100 kb and 1 Mb and eight >1 Mb (data not shown). FASSI-6 length constraints produced the most contiguous assembly, with 540 scaffolds, seven between 100 kb and 1 Mb and nine >1 Mb in size. Interestingly, throughout all FASSI experiments the FASSI contig N50 length increased by 2%–6% compared to the control assembly, while the scaffold N50 length remained unchanged or increased only slightly (1% increase at the most) (data not shown). Although length constraints are supplied at the layout stage of the assembly, additional contig merges are made by PCAP to accommodate additional forward–reverse pair restrictions (Huang et al. 2003). The low scaffold N50 increase for FASSI assemblies can be explained by merges that occur between large scaffolds already comprised of more than half of all bases in the assembly. It is worth noting that the control assembly is very contiguous at this depth, leaving little room for further improvement in scaffold length. In this respect, any improvement based on the physical map that brings the number of large scaffolds closer to the number of map contigs should not be taken lightly.

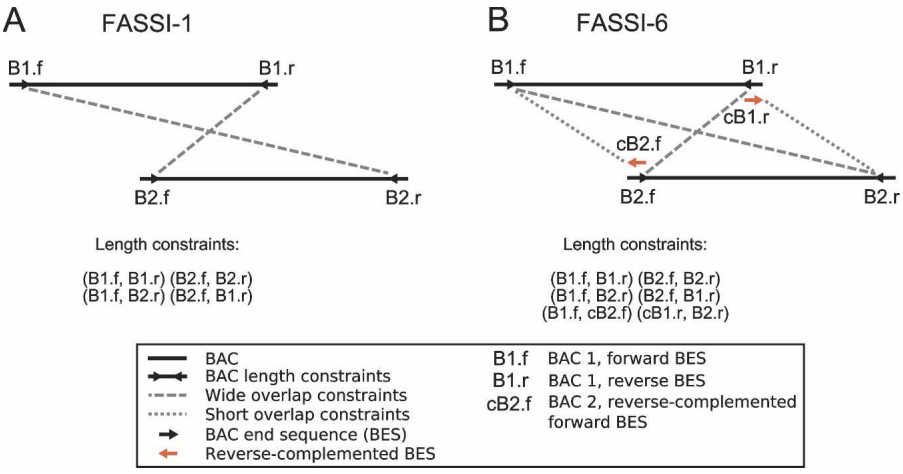

Figure 1. BAC length constraints derived from physical maps. (*A*) FASSI-1 consists of four different sets of length constraints between any two overlapping BAC neighbors (40% overlap in our implementation). In addition to introducing accurate length constraints between mate pairs of a given BAC, two additional, wide overlap constraints are introduced between neighbors. (*B*) Constraints in FASSI-6 include the length constraints from FASSI-1, plus short overlap constraints between any two proximal BAC end sequences from BAC sharing 40% of their length. The FASSI parameters are detailed in Methods.

WGS assembly alignments to the finished *C. gattii* genome using wuBLAST (http://blast.wustl.edu) corroborate that large scaffolds (>100 kb) are effectively merging as a result of adding map-derived constraints (Table 1). With FASSI-1, 93% of the *C. gattii* genome is covered by 17 large scaffolds. With FASSI-6, a genome coverage of 92% is achieved by 16 scaffolds >100 kb. This is one less scaffold than the number of map contigs and two more than the electrophoretic karyotype for this genome (G. Hu and J. Kronstad, unpubl.).

To further quantify the effect of map-derived constraints on sequence assemblies, we define the contiguity score, *C*, as the number of aligned bases divided by the number of scaffolds aligning in the 100 kb–1 Mb and >1 Mb size range. Large *C* values represent a clear improvement in contiguity, with more bases being covered by larger and fewer scaffolds. The contiguity score also gives a fair indication of the quality of the assembly at the scaffold and contig level. This is because contigs that do not align in the context of their scaffolds are misplaced, align to the reference in the wrong direction, or interrupt the contiguity of another scaffold, all creating breaks in contiguity. For any given scaffold, two or more consecutive contig alignments contribute to the overall scaffold contiguity, until a break is encountered. For the FASSI-1 and FASSI-6 WGAs, the contiguity score for large aligning scaffolds exceeds 1 Mb, a 17.5% and 23.6% increase relative to the control assembly, respectively (Table 1).

Table 1. PCAP assemblies of *C. gattii* at 5× read depth using map-derived BAC length constraints

	Control	FASSI-1	FASSI-6
Scaffolds (>100 kb)	20	17	16
Aligned bases (Mb)	17.12	17.10	16.94
Contiguity score (×1000)	856	1006	1059
Genome coverage[a]	93%	93%	92%

Number of major scaffolds (>100 kb) aligning to the finished genome, aligned bases, contiguity score, and genome coverage.
[a]Based on an 18.36-Mb genome.

D. melanogaster WGS assemblies

FASSI assemblies compared to a control

Similarly to *C. gattii*, the biggest change observed is in scaffold rearrangements (Fig. 2). While the total number of assembled bases remains fairly constant between WGAs at any given shotgun depth, there are obvious shifts in bases from smaller to larger scaffolds. In Figure 2, this base shift can be observed clearly starting at 2× coverage. The difference between the FASSI WGAs and the control assembly is most predominant at 4× coverage and remains fairly constant as the read depth increases. At this depth, the control assembly yielded 12 scaffolds in the 1-Mb range, totaling 20 million bases. The FASSI-1 and FASSI-6 WGAs yielded 11 and 19 scaffolds in that range, respectively. These scaffolds totaled 21 and 34 Mb, in that order (data not shown). FASSI-1 and FASSI-6 length constraints have the potential to double and triple (Fig. 1) the number of supportive linking pairs between distant contigs, respectively. Supported map-derived length constraints can have a considerable effect on the long-range contiguity of the resulting assembly, especially at low read depth when less information from small clones is available to link scaffolds. Consistently, at 4× coverage WGA, the scaffold N50 length is increased by 57% and 236% with the use of FASSI-1 and FASSI-6 length constraints compared to the control WGA, respectively (Fig. 3B). The FASSI-1 contig N50 length is increased by an average of 2% across all read depths. The observed increase in FASSI contig N50 length caused by the incorporation of overlap FASSI constraints confirms that map-derived constraints also improve the assembly contiguity locally (Fig. 3A).

Whole–genome alignments

To assess the assembly accuracy, WGAs were aligned to the release 4 of the *D. melanogaster* genome (http://www.fruitfly.org) using wuBLAST, and contiguity scores were calculated as described previously.

At 4× coverage, 59.2% of the *D. melanogaster* genome is covered by well-aligned FASSI-6 scaffolds >100 kb, compared to 46.1% and 49.1% for the control and FASSI-1 WGAs, respectively (Table 2). At low shotgun depth, namely, 2× to 4×, FASSI-6 constraints consistently outperform FASSI-1 constraints on the basis of contiguity and coverage by large scaffolds. For *D. melanogaster*, maximum genome coverage by scaffolds >100 kb is attained at fivefold coverage for the FASSI WGAs and at sixfold coverage for the control assembly without map-derived constraints. Beyond this depth, additional shotgun reads are more beneficial to increasing the contiguity of sequence contigs, as Figure 3A suggests. However, it is worth noting that even at higher depths, FASSI constraints still lead to larger, more well-assembled scaffolds compared to the control assembly (Table 2, 10× read depth).

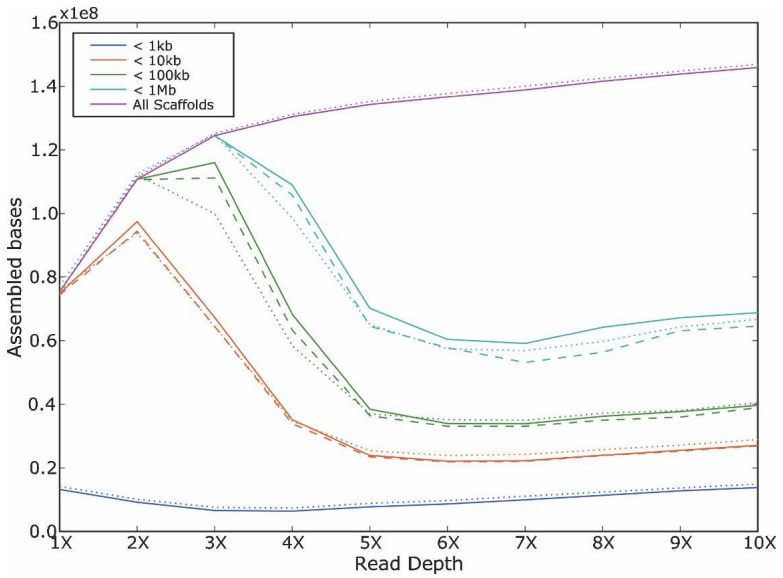

Figure 2. Effect of FASSI length constraints on *D. melanogaster* WGA scaffold sizes. Bases contained in scaffolds smaller than 1 kb, 10 kb, 100 kb, 1 Mb, and in all scaffolds are depicted by different colors on the stacked line graph. For every shotgun depth ranging from 1× to 10×, a base distribution for the control without map-derived constraints (solid line), FASSI-1 (dashed line), and FASSI-6 (dotted line) is shown.

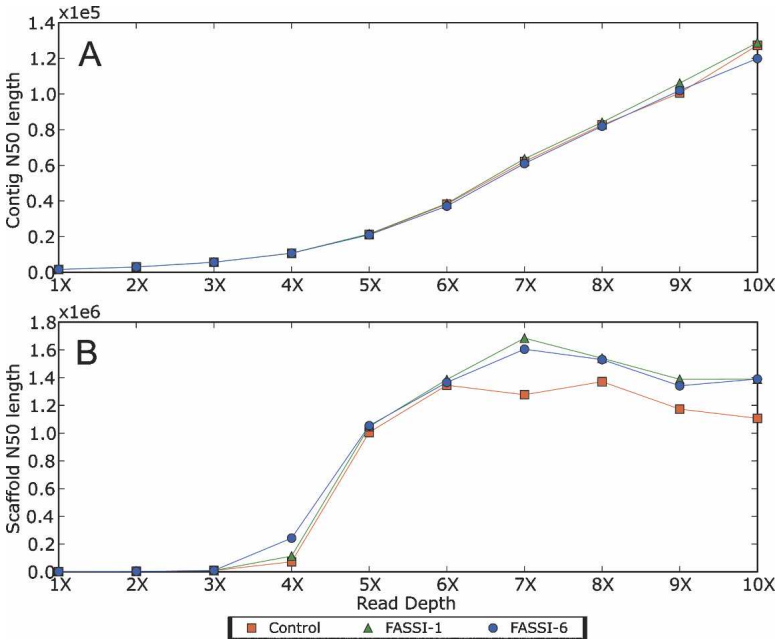

Figure 3. Effect of FASSI on the assembly contiguity of *D. melanogaster*. (*A*) Contig and (*B*) scaffold N50 length are calculated for every assembly at shotgun depths ranging from 1× to 10×.

To investigate further the effect of map-derived length constraints on the shotgun assemblies of *D. melanogaster*, we plotted all contig alignments in the context of their scaffold and WGA (Fig. 4). The right arm of chromosome 3 (arm 3R) was selected for graphical representation of 4× coverage WGA alignments, because of the effect observed on scaffold contiguity at this coverage. By comparing FASSI alignments to the control, we observe that the longest scaffolds to align to the reference are from the FASSI WGAs (displayed on the figure as a reduction in alternating colors). As seen on the first track of Figure 4, a higher concentration of repeated sequences, such as near coordinate 6 M and close to the telomere between coordinate 0–1 M, still poses a challenge for shotgun assemblies, FASSI or not. However, it is clear that map-derived BAC constraints improve scaffold contiguity regardless of most repeats on arm 3R.

We investigated shotgun assembly mistakes at the scaffold level. The following results are based on contig alignments between WGAs and their respective finished genome. We took into account all contigs >500 bases in size and aligned to the genome with >70% sequence identity over the contig length and >90% over the wuBLAST alignment hit length. We minimized the effect of repeats by counting contigs as misplaced only if they did not align in the right order and orientation relative to the other contigs of the parent scaffold. Using this scheme, we find that misplaced FASSI-1 contigs account for fewer misassembled bases than the control assembly, for both the *C. gattii* and *D. melanogaster* genomes (Table 3) and across all shotgun depths for *D. melanogaster* (data not shown). FASSI-6 length constraints, however, lead to WGAs having additional contigs that do not align in the right order for both genomes. For *D. melanogaster*, misaligned bases calculated from aligning the 4× coverage FASSI-6 WGA account for merely 0.25% of all contiguously aligned bases and represent a 4.1% increase relative to the control assembly. Error rates that account for the longer-range contiguity of the FASSI assemblies are 0.3% and 0.1% lower for FASSI-1 and FASSI-6 compared to the control assembly, respectively (data not shown). For both *Drosophila* and *Cryptococcus*, we did not find a single case of a global misassembly introduced by the physical map. Obviously, wrong joins between BACs in the physical map would be reflected in the sequence assembly, unless sufficient length constraints from plasmid and fosmid clones contradicted the map-derived BAC length and overlap constraints we introduced. With PCAP, if the number of correct fosmid/plasmid read pairs is greater than the number of incorrect map-derived read pairs by at least two for a region, then no misassembly will ensue (X. Huang, pers. comm.).

Chimpanzee genome assembly

The *P. troglodytes* (chimpanzee) genome was sequenced to 6.5-fold coverage, assembled and mapped at the Genome Sequencing Center in St. Louis, Missouri. FASSI-1 was used to create BAC end and BAC overlap length constraints from a ninefold coverage draft chimpanzee physical map (W. Warren, unpubl.) and incorporated into the genome assembly process using PCAP.REP (Huang et al. 2006). Although the effect of adding map-derived constraints is not as sizeable as it is for *D. melanogaster*, scaffold contiguity is increased as the scaffold N50 length and base distribution for scaffolds suggest (data not shown). Overall, 32.6 million bases from scaffolds shorter than 1 Mb shifted into larger scaffolds (>1 Mb) as a result of using accurate and redundant BAC length constraints derived from the chimpanzee physical map. Compared to a control assembly, this is a 1.26% increase. Furthermore, the FASSI scaffold N50 length reaches 10.9 Mb, a 26.1% increase compared to the control. With twice the number of BAC constraints, supportive FASSI-1 pairing is responsible for bringing into the assembly >197,000 additional assembled bases and reducing the total number of scaffolds by 238 (data not shown).

Discussion

The versatility of physical maps and low cost compared to sequencing has made genome mapping a natural component of large genome sequencing endeavors (Gregory et al. 2002; Wallis et al. 2004). Mapping has found an important niche in sequencing applications by providing large clone resources and a framework for genome finishing and validation. Physical maps are routinely used by large sequencing centers as a tool to increase the scaffold contiguity of WGA and identify incorrect joins post-assembly. In this paper, we report the use of physical maps during the assembly process in order to guide and help assembly algorithm make the right decision regarding the validity of a join at the contig and scaffold levels. To this end, we have developed FASSI, a program that calculates BAC end and BAC overlap length constraints between overlapping BACs from physical maps stored in FPC (Soderlund et al. 1997; Ness et al. 2002).

From our observations, BAC length constraints derived from individual fingerprints are used to compute more contiguous assemblies, with less assembly errors for some, but not all, FASSI experiments. Overlap length constraints add redundancy to the read pairs set, substantiating weak joins both at the contig and

Table 2. Improving *D. melanogaster* genome contiguity and coverage by scaffolds >100 kb in length

WGA	Number of scaffolds >100 kb			Genome coverage by scaffolds >100 kb[a]			Contiguity score (×1000)		
Shotgun depth	Control	FASSI-1	FASSI-6	Control	FASSI-1	FASSI-6	Control	FASSI-1	FASSI-6
1×	—	—	—	0.0%	0.0%	0.0%	—	—	—
2×	—	—	1	0.0%	0.0%	0.1%	—	—	139
3×	35	51	74	4.8%	7.4%	14.2%	162	171	226
4×	131	139	117	46.1%	49.1%	59.2%	415	417	597
5×	100	94	109	74.0%	75.0%	79.8%	873	942	864
6×	92	83	87	77.0%	77.4%	77.7%	987	1101	1053
7×	104	85	93	77.2%	77.8%	77.8%	876	1081	988
8×	100	82	85	74.5%	75.4%	76.0%	879	1085	1054
9×	107	95	95	72.9%	74.5%	74.0%	804	925	920
10×	110	96	94	73.8%	73.2%	74.0%	791	899	929

[a]Based on 118.4-Mb *D. melanogaster* euchromatin.

scaffold level and improving assembly contiguity in the process. The efficiency of this technique was evaluated at shotgun depths ranging from 1× to 10× by randomly selecting and assembling shotgun reads from the fruit fly data set. The improvement in scaffold contiguity is seen at all depths, but is optimum at fourfold to fivefold coverage. At low shotgun depths, the pinnacle in assembly contiguity is achieved by FASSI-6. Provided that the accuracy of the resulting assembly is preferred over its contiguity, FASSI-1 might be a better choice. In conjunction with the data set used in the present study, the magnitude of the assembly errors detected in the FASSI-6 assemblies is low, and error rates that take into account improvements in scaffold contiguity are lower for the FASSI WGAs when compared to a control without map-derived constraints.

For a mammalian-sized genome, the cost of a 10-fold coverage physical map is negligible even when compared to the sequencing cost required to cover that genome only once. With this in mind, the ability of FASSI to achieve the scaffold contiguity of higher-depth WGAs with 1× less sequence reads has tremendous potential in reducing the cost of large genome sequencing projects. The simplicity and generic nature of the FASSI approach also preclude the need to alter existing assembly algorithms, unless these programs cannot accept length constraints for individual clones. Since the mapping information is used at run-time during the sequence assembly process, the output of a sequence assembler can be used directly for genome finishing if need be. Given that FASSI length constraints are derived from physical maps, it is clear that map depth, contiguity, and accuracy are instrumental to the success of this approach.

As resources necessary to sequence, assemble, and finish whole genomes with Sanger-based sequencing methods become increasingly limiting, it is imperative to find novel ways to reduce costs and develop more efficient ways to process the low-coverage shotgun data at hand while improving the quality and contiguity of the resulting assemblies. With the advent of short read sequencing and its challenges for de novo sequencing of genomes, the value of a physical map in guiding the placement and orientation of contigs otherwise devoid of pairing information is also indisputable. Physical map-assisted genome sequence assemblies represent an initial step in addressing these points as well as providing a key experimental resource for subsequent genome finishing and completion.

Software availability

FASSI is implemented in Perl and run on Linux. It is distributed under the same terms as Perl and is available from Canada's Michael Smith Genome Science Center (Vancouver, BC) at http://www.bcgsc.ca/bioinfo/software/FASSI. FASSI is free for academic and noncommercial use.

Methods

FASSI length constraints

FASSI is written in Perl and runs on Linux. The software produces three types of length constraints, supplied to PCAP (Huang et al. 2003) as a permissible lower and upper distance limit between any two BES having a relationship. Relationships include BES from the same clone, logical intrinsic, and logical reverse-complemented BES between overlapping clones. The relationships are diagrammed in Figure 1. Logical BES refers to end sequences properly oriented relative to each other (end sequences with opposite orientation, facing each other). We are generating three types of length constraints based on the physical map:

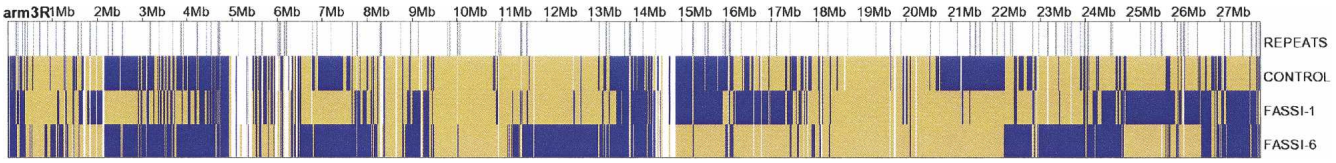

Figure 4. Sequence assembly alignments to *D. melanogaster* arm 3R show longer well-assembled scaffolds for the FASSI-1 and FASSI-6 WGAs compared to a control without map-derived constraints at 4× coverage. Scaffolds of all sizes containing at least two contigs, each aligning to the reference genome with 70% sequence identity over the contig length and 90% over the wuBLAST hit length, are displayed. The blue and yellow lines represent contigs of a specific scaffold. The color switch signifies a change in the identity of aligned scaffolds. White spaces indicate regions without suitable contig alignment. The first track shows repeated segments (vertical gray lines) at specified positions on the chromosome. Repeats >500 bp in size and sharing >95% sequence identity are shown on the first track. Tracks 2, 3, and 4 show scaffold alignments between the finished genome and the control assembly, FASSI-1 and FASSI-6, respectively.

Table 3. Assembly errors (per megabase of sequence) based on WGA alignment to reference genomes

WGA	Contigs in wrong orientation	Misplaced contigs	Misaligned bases
A. *D. melanogaster* genome at 4× read depth			
Control	0.12	0.46	2212
FASSI-1	0.10	0.52	2060
FASSI-6	0.08	0.57	2306
B. *C. gattii* WM276 genome at 5× read depth			
Control	—	0.03	55
FASSI-1	—	0.02	32
FASSI-6	0.04	0.03	94

1. BAC length constraints, derived from the fingerprint estimation of the BAC insert size, plus or minus 40% for the upper and lower size limits, respectively. These limits account for possible errors in the calculation of BAC insert sizes from fingerprints, which are rarely more than 10%–20% of a calculated insert size (M. Krzywinski, pers. comm.).
2. Buried length constraints, generated when a BAC is a subset of another BAC, using the distance between proximal or distal BES of the two clones.
3. Overlap length constraints, generated when BAC neighbors in the map share a portion (40% in our implementation) of their length. The shared length is calculated as the sum of shared restriction fragment sizes.

Six FASSI assembly strategies were designed, two of which are presented here (Fig. 1). FASSI-1 consists of four different sets of length constraints between any two overlapping BAC neighbors sharing >40% of their length. FASSI-1 also includes constraints for buried BACs and singletons. In addition to length constraints delimiting the allowable distance between the mate pairs of each BAC, two additional wide overlap constraints are introduced between overlapping and buried neighbors as a distance between the left end read of one BAC and the right end sequence read of the other, for both possible left–right pairs (Fig. 1A). Constraints in FASSI-6 include the length constraints from the first approach, plus short overlap constraints between any two proximal BES from BAC sharing 40% of their length. Reverse-complemented BES are generated in order to produce these logical short overlap constraints between the proximal BAC end sequences (Fig. 1B). FASSI-6 also includes length constraints for singletons and overlap length constraints between unburied and buried BACs.

FASSI takes four files as input, three of which are compulsory for the generation of BAC length constraints for each clone fingerprinted. The fourth is optional, but essential to produce overlap and buried length constraints between overlapping BACs of a physical map. A fingerprint map in FPC file format (Soderlund et al. 1997, 2000) is first used to get the order and relative position of BACs within map contigs. This information is used in conjunction with BAC fingerprints to calculate BAC length constraints. BAC fingerprint data are generated by BandLeader (Fuhrmann et al. 2003) and supplied to FASSI in the Image file format (Sulston et al. 1988) (http://www.sanger.ac.uk/Software/Image). To account for possible differences in nomenclature between map and sequencing clone names, a third file lists the name of every BES as well as the BAC it is associated with in the physical map. Calculation of overlap and buried length constraints necessitate knowledge about the relative orientation of every BAC in the fingerprint map. This information can only be generated by bootstrapping the sequence assembly process and is optional to FASSI. However, it is essential to the generation of overlap and buried length constraints between neighboring BACs.

Shotgun data

A total of 4,087,972 *D. melanogaster* traces were downloaded from the NCBI trace archive repository (http://www.ncbi.nlm.nih.gov/Traces/trace.cgi). The base calling was done using phred (Ewing and Green 1998; Ewing et al. 1998) and length constraints were approximated for small (1.8 kb) and medium (9.8–11.5 kb) size plasmid clones and optimized for PCAP assemblies, as per the recommendations of Xiaoqiu Huang (pers. comm.). Shotgun reads were quality-trimmed using trim2 (X. Huang, pers. comm.) and vector-clipped using cross_match (http://www.phrap.org). 28,840 BES were downloaded from the Genoscope (the French National Sequencing Center) Web site (http://www.genoscope.cns.fr/externe/English/Outils), and quality scores were set to phred 20. As directed (X. Huang, pers. comm.), BAC length constraints used for the control assembly of *C. gattii* and *D. melanogaster* were set to 60 kb and 300 kb for the lower and upper boundaries, respectively. For *P. troglodytes*, the lower and upper boundaries were set to 110 and 180 kb, respectively. *C. gattii* strain WM276 plasmid (2.8 kb, 15 kb), fosmid (35 kb), and BAC libraries, prepared by our group, Cletus D'Souza, and James W. Kronstad at the University of British Columbia, were sequenced at Canada's Michael Smith Genome Sciences Center in Vancouver. After quality and vector clipping, 163,936 processed shotgun reads were assembled, generating fivefold coverage of the *C. gattii* genome. The *P. troglodytes* genome was sequenced to ~6.5-fold coverage by both the Washington University Genome Sequencing Center (WUGSC) and the Broad Institute (Boston, MA). A total of 35 million sequence paired-end reads were generated from libraries ranging in size from 4 kb to 180 kb.

Physical maps

For both *D. melanogaster* and *C. gattii*, BAC clone fingerprints were processed by Bandleader (Fuhrmann et al. 2003), assembled using FPC (Soderlund et al. 1997; Ness et al. 2002) and hand-edited in-house. The BAC-based physical map for *D. melanogaster* is being built in collaboration with the Berkeley *Drosophila* Genome Project (BDGP) (J.E. Schein, R. Hoskins, J. Carlson, S. Celniker, and G. Rubin, unpubl.). The physical map for *P. troglodytes* is being constructed at the Genome Sequencing Center in St. Louis, who kindly provided the FPC files. The chimpanzee map had not been hand-edited at the time of experimentation. The clone depth for the physical maps of *C. gattii*, *D. melanogaster*, and *P. troglodytes* used in this study was 16×, 10×, and 9×, respectively.

Whole-genome shotgun sequence assemblies

D. melanogaster sequence reads were selected at random to simulate shotgun depths ranging from 1× to 10×. BES were not subjected to random selection, and corresponding sets of length constraints were used in their entirety across all depths for the FASSI experiments and the control assembly. PCAP was used as the assembly engine for *D. melanogaster* and *C. gattii* (Huang et al. 2003), and PCAP.REP (Huang et al. 2006) was used to assemble the chimpanzee genome. For *Drosophila*, 20 parallel PCAP jobs ran on a cluster of AMD opteron computers (AMD) with 2 Gb Random Access Memory (RAM) and dual processor running SUSE 9.0. PCAP programs to calculate the layout (bcontig) and consensus (bconsen) subsequently ran on a Sun E2900 with 96 Gb RAM and 12 dual-core UltraSPARC IV processors. For the chimpanzee genome assembly, 400 parallel PCAP jobs ran on a blade

center with Intel Xeon or AMD opteron with 2–8 Gb RAM. The bcontig and bconsen program of the PCAP suite subsequently ran on Intel Itanium machines with 96 Gb RAM.

Assembly figures

Whole-genome assemblies were compared to each other using various metrics. These include rate of pairing and satisfied constraints, contig and scaffold N50 length calculations, number and size of gaps, contig, scaffold, and gap size distribution for size ranges of 1 bp–1 kb, 1 kb–10 kb, 10 kb–100 kb, 100 kb–1 Mb, and >1 Mb. The N50 length is the length that marks 50% of genome content, and is a measure of contiguity for both contigs and scaffolds.

WGA alignments and assessment of assembly accuracy

For genome alignment purposes, the release 4 of the *D. melanogaster* genome was downloaded from the BDGP Web site (http://www.fruitfly.org). For *C. gatti*, WGAs were compared to the fully finished 18.36-Mb genome completed at the BCCAGSC earlier this year (J.W. Kronstad, C. D'Souza, G. Taylor, R.L. Warren, J. Schein, M. Marra, S. Jones, B.F.F. Ouellette, and R. Holt, unpubl.). Repetitive sequences within contigs were masked using RepeatMasker (http://repeatmasker.org) and aligned to their reference genomes using wuBLAST (http://blast.wustl.edu). wuBLAST alignments were parsed and ordered numerically based on the hit coordinates. Broken consecutive alignments were resolved prior to the analysis, when applicable. For the analysis, we took into account all alignments >500 bases from contigs sharing at least 70% sequence identity with the chromosome over the entire contig length and 90% sequence identity over the hit length. Aligned bases were counted exclusively from contigs aligning contiguously to the genome, in the correct order and orientation relative to the whole scaffold. Visual alignments between whole-genome assemblies and a reference genome were generated using the same logic, and repeats were identified using cross_match (http://www.phrap.org). Misplaced contigs, scaffold interruptions caused by high-quality alignments, and incorrect contig orientation based on sequence alignment all cause breaks in contiguity affecting the number of contiguously aligned bases. Contiguously aligned bases are segregated by scaffold alignment size ranges. The contiguity score is calculated as the number of contiguous and correctly assembled bases per scaffold aligning in a specific size range. Alignment plots were generated by Python scripts using the Python Imaging Library (PIL; http://www.pythonware.com/products/pil).

Acknowledgments

This project was supported by a grant from the National Human Genome Research Institute (NHGRI U54 HG00307). M.A.M and S.J.M.J. are Michael Smith Foundation for Health Research Scholars. Funding for sequencing *C. gattii* strain WM276 was provided by Genome Canada.

References

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoslins, R.A., Galle, R.F., et al. 2000. The genome sequence of *Drosophila melanogaster*. *Science* **287:** 2185–2195.

Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408:** 796–815.

Batzoglou, S., Jaffe, D.B., Stanley, K., Butler, J., Gnerre, S., Mauceli, E., Berger, B., Mesirov, J.P., and Lander, E.S. 2002. ARACHNE: A whole-genome shotgun assembler. *Genome Res.* **12:** 177–189.

C. elegans Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science* **282:** 2012–2018.

Choi, V. and Farach-Colton, M. 2003. Barnacle: An assembly algorithm for clone-based sequences of whole genomes. *Gene* **320:** 165–176.

Coulson, A., Sulston, J., Brenner, S., and Karn, J. 1986. Toward a physical map of the genome of the nematode *Caenorhabditis elegans*. *Proc. Natl. Acad. Sci.* **83:** 7821–7825.

Ewing, B. and Green, P. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8:** 186–194.

Ewing, B., Hillier, L., Wendl, M.C., and Green, P. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8:** 175–185.

Fuhrmann, D.R., Krzywinski, M.I., Chiu, R., Saeedi, P., Schein, J.E., Bosdet, I.E., Chinwalla, A., Hillier, L.W., Waterston, R.H., McPherson, J.D., et al. 2003. Software for automated analysis of DNA fingerprinting gels. *Genome Res.* **13:** 940–953.

Gibbs, R.A., Weinstock, G.M., Metzker, M.L., Muzny, D.M., Sodergren, E.J., Scherer, S., Scott, G., Steffen, D., Worley, K.C., and Burch, P.E. 2004. Genome sequence of the brown Norway rat yields insights into mammalian evolution. *Nature* **428:** 475–476.

Gregory, S.G., Sekhon, M., Schein, J., Zhao, S., Osoegawa, K., Scott, C.E., Evans, R.S., Burridge, P.W., Cox, T.V., Fox, C.A., et al. 2002. A physical map of the mouse genome. *Nature* **418:** 743–750.

Havlak, P., Chen, R., Durbin, K.J., Egan, A., Ren, Y., Song, X.-Z., Weinstock, G.M., and Gibbs, R.A. 2004. The atlas genome assembly system. *Genome Res.* **14:** 721–732.

Hillier, L.W., Miller, W., Birney, E., Warren, W., Hardison, R.C., Ponting, C.P., Bork, P., Burt, D.W., Groenen, M.A., Delany, M.E., et al. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **433:** 695–716.

Holt, R.A., Subramanian, G.M., Halpern, A., Sutton, G.G., Charlab, R., Nusskern, D.R., Wincker, P., Clark, A.G., Ribeiro, J.M., Wides, R., et al. 2002. The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science* **298:** 129–149.

Huang, X. and Madan, A. 1999. CAP3: A DNA sequence assembly program. *Genome Res.* **9:** 868–877.

Huang, X., Wang, J., Aluru, S., Yang, S.-P., and Hillier, L. 2003. PCAP: A whole-genome assembly program. *Genome Res.* **13:** 2164–2170.

Huang, X., Yang, S.-P., Chinwalla, A.T., Hillier, L.W., Minx, P., Mardis, E.R., and Wilson, R.K. 2006. Application of a superword array in genome assembly. *Nucleic Acids Res.* **34:** 201–205.

Jaffe, D.B., Butler, J., Gnerre, S., Mauceli, E., Lindblad-Toh, K., Mesirov, J.P., Zody, M.C., and Lander, E.S. 2003. Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res.* **13:** 91–96.

Kent, W.J. and Haussler, D. 2001. Assembly of the working draft of the human genome with GigAssembler. *Genome Res.* **11:** 1541–1548.

Lander, E.S., Linton, L.M., Bireen, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.

Margulies, E.H., Vinson, J.P., NISC Comparative Sequencing Program, Miller, W., Jaffe, D.B., Lindblad-Toh, K., Chang, J.L., Green, E.D., Lander, E.S., Mullikin, J.C., et al. 2005. An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing. *Proc. Natl. Acad. Sci.* **102:** 4795–4800.

Marra, M.A., Kucaba, T.A., Dietrich, N.L., Green, E.D., Brownstein, B., Wilson, R.K., McDonald, K.M., Hillier, L.W., McPherson, J.D., and Waterson, R.H. 1997. High throughput fingerprint analysis of large-insert clones. *Genome Res.* **7:** 1072–1084.

Mikkelsen, T.S., Hillier, L.W., Eichler, E.E., Zody, M.C., Jaffe, D.B., Yang, S.-P., Enard, W., Hellmann, I., Lindblad-Toh, K., Altheide, T.K., et al. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437:** 69–87.

Mullikin, J.C. and Ning, Z. 2003. The phusion assembler. *Genome Res.* **13:** 81–90.

Myers, E.W., Sutton, G.G., Delcher, A.L., Dew, I.M., Fasulo, D.P., Flanigan, M.J., Kravitz, S.A., Mobarry, C.M., Reinert, K.H., Remington, K.A., et al. 2000. A whole-genome assembly of *Drosophila*. *Science* **287:** 2196–2204.

Ness, S.R., Terpstra, W., Krzywinski, M., Marra, M.A., and Jones, S.J. 2002. Assembly of fingerprint contigs: Parallelized FPC. *Bioinformatics* **18:** 484–485.

Soderlund, C., Longden, I., and Mott, R. 1997. FPC: A system for building contigs from restriction fingerprinted clones. *Comput. Appl. Biosci.* **13:** 523–535.

Soderlund, C., Humphray, S., Dunham, A., and French, L. 2000. Contigs

built with fingerprints, markers, and FPC V4.7. *Genome Res.* **10:** 1772–1787.
Sorrell, T.C. 2001. *Cryptococcus neoformans* variety *gattii*. *Med. Mycol.* **39:** 155–168.
Sulston, J., Mallet, F., Staden, R., Durbin, R., Horsnell, T., and Coulson, A. 1988. Software for genome mapping by fingerprinting techniques. *Comput. Appl. Biosci.* **4:** 125–132.
Sutton, G.G., White, O., Adams, M.D., and Kerlavage, A.R. 1995. TIGR assembler: A new tool for assembling large shotgun sequencing projects. *Genome Sci. Technol.* **1:** 9–19.
Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. 2001. The sequence of the human genome. *Science* **291:** 1304–1351.
Wallis, J.W., Aerts, J., Groenen, M.A., Crooijmans, R.P., Layman, D., Graves, T.A., Scheer, D.E., Kremitzki, C., Fedele, M.J., Mudd, N.K., et al. 2004. A physical map of the chicken genome. *Nature* **432:** 761–764.
Warren, R.L., Butterfield, Y.S., Morin, R.D., Siddiqui, A.S., Marra, M.A., and Jones, S.J.M. 2005. Management and visualization of whole genome shotgun assemblies using SAM. *Biotechniques* **38:** 715–720.
Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.

Received December 28, 2005; accepted in revised form April 11, 2006.