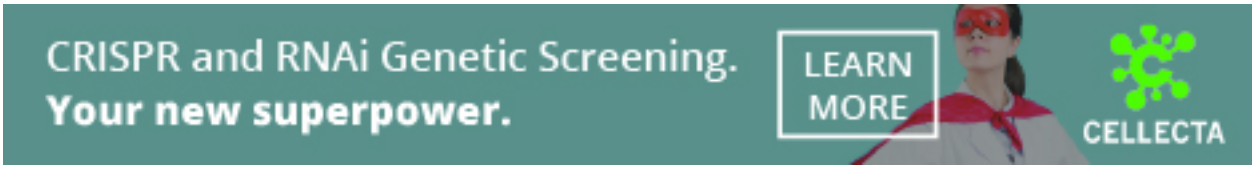
A new strategy for genome assembly using short sequence reads and reduced representation libraries

Andrew L. Young, Hatice Ozel Abaan, Daniel Zerbino, et al.

Genome Res. 2010 20: 249-256
Access the most recent version at doi:10.1101/gr.097956.109



A new strategy for genome assembly using short sequence reads and reduced representation libraries

Andrew L. Young,[1] Hatice Ozel Abaan,[1] Daniel Zerbino,[2] James C. Mullikin,[1] Ewan Birney,[2] and Elliott H. Margulies[1,3]

[1]*Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA;* [2]*European Bioinformatics Institute, Hinxton, Cambridge CB10 1SD, United Kingdom*

We have developed a novel approach for using massively parallel short-read sequencing to generate fast and inexpensive de novo genomic assemblies comparable to those generated by capillary-based methods. The ultrashort (<100 base) sequences generated by this technology pose specific biological and computational challenges for de novo assembly of large genomes. To account for this, we devised a method for experimentally partitioning the genome using reduced representation (RR) libraries prior to assembly. We use two restriction enzymes independently to create a series of overlapping fragment libraries, each containing a tractable subset of the genome. Together, these libraries allow us to reassemble the entire genome without the need of a reference sequence. As proof of concept, we applied this approach to sequence and assembled the majority of the 125-Mb *Drosophila melanogaster* genome. We subsequently demonstrate the accuracy of our assembly method with meaningful comparisons against the current available *D. melanogaster* reference genome (dm3). The ease of assembly and accuracy for comparative genomics suggest that our approach will scale to future mammalian genome-sequencing efforts, saving both time and money without sacrificing quality.

[Supplemental material is available online at http://www.genome.org. The sequence data from this study have been submitted to NCBI Short Read Archive (http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi) under accession no. SRA010040.]

Genomes are the fundamental unit by which a species can be defined and form the foundation for deciphering how an organism develops, lives, dies, and is affected by disease. In addition, comparisons of genomes from related species have become a powerful method for finding functional sequences (Eddy 2005; Xie et al. 2005, 2007; Pennacchio et al. 2006; The ENCODE Project Consortium 2007). However, the high cost and effort needed to produce draft genomes with capillary-based sequencing technologies have limited genome-based biological exploration and evolutionary sequence comparisons to dozens of species (Margulies et al. 2007; Stark et al. 2007). This limitation is particularly true for mammalian-sized genomes, which are gigabases in size.

With the advent of massively parallel short-read sequencing technologies (Bentley et al. 2008), the cost of sequencing DNA has been reduced by orders of magnitude, now making it possible to sequence hundreds or thousands of genomes. However, the reduced length of the sequence reads, compared with capillary-based approaches, poses new challenges in genome assembly. Here, we sought to address those experimental and bioinformatics hurdles by combining classical biochemical methodologies with new algorithms specifically tailored to handle massive quantities of short-read sequences.

To date, the whole-genome shotgun (WGS) approach using massively parallel short-read sequencing has shown significant promise in silico (Butler et al. 2008) and has been applied to de novo sequencing and assembly of small genomes that do not contain an overabundance of low-complexity repetitive sequence (Hernandez et al. 2008). This presents a challenge when scaled to larger more complex genomes, where the information contained in a single short read cannot unambiguously place that read in the genome. Additionally, de novo assembly from WGS short-read sequencing currently requires large computational resources, on the order of hundreds of gigabytes of RAM, when scaled to larger genomes. As a compromise, current mammalian genomic analyses utilizing short-read sequencing technology either use alignments of individual reads against a reference genome (Ley et al. 2008; Wang et al. 2008) or require elaborate parallelization schemes across a large compute farm (Simpson et al. 2009) for assembly. Regardless of computational improvements, effectively handling repetitive sequences in a whole-genome assembly still remains a challenge.

Our goal was to establish wet-lab and bioinformatics methods to rapidly sequence and assemble mammalian-sized genomes in a de novo fashion. Importantly, we wanted an approach that (1) did not rely on existing reference assemblies, (2) could be accomplished using commodity computational hardware, and (3) would yield functional assemblies useful for comparative sequence analyses at a fraction of the time and cost of existing capillary-based methods. To accomplish this, we propose a generic genome partitioning approach to solve both the biological and computational challenges of short-read assembly.

Traditionally, partitioning of genomic libraries was accomplished through the use of clonal libraries of bacterial artificial chromosomes (BACs) (or yeast artificial chromosomes). This method accurately partitions genomes into more manageable subregions for sequencing and assembly. However, the high financial cost and overhead associated with creating and maintaining these libraries make this method unattractive for scaling to hundreds of genomes. In addition, a single BAC clone, which contains ~200 kb of sequence, is not large enough to leverage the amount of sequence obtained from a single lane of Illumina data

[3]Corresponding author.
E-mail elliott@nhgri.nih.gov.
Article is online at http://www.genome.org/cgi/doi/10.1101/gr.097956.109.

(currently ~2.5 Gb of sequence), requiring the need for various pooling or indexing strategies (Meyer et al. 2007). Furthermore, virtually all BAC libraries exhibit some degree of variable cloning bias, with some regions overrepresented and others not at all. In silico studies have investigated the cost-saving potential of using a randomized BAC clone library with short-read sequencing (Sundquist et al. 2007); however, even with this shortcut, the clonal library concept does not lend itself to fast and cheap whole-genome partitioning.

We propose a novel partitioning approach using restriction enzymes to create a series of reduced representation (RR) libraries by size fractionation. This method was originally described for single nucleotide polymorphism (SNP) discovery using Sanger-based sequencing methods on AB377 machines (Altshuler et al. 2000) and was subsequently used with massively parallel short-read sequencing (Van Tassell et al. 2008). Importantly, this method allows for the selection of a smaller reproducible subset of the genome for assembly. We extended this concept to create a series of distinct RR libraries consisting of similarly sized restriction fragments. Individually, these libraries represent a tractable subset of the genome for sequencing and assembly; when taken together, they represent virtually the entire genome. Using two separate restriction enzymes generates overlapping libraries, which allow for assembly of the genome without using a reference sequence.

As proof of concept, we present here a de novo *Drosophila melanogaster* genomic assembly, equivalent in utility to a comparative grade assembly (Blakesley et al. 2004). Two enzymes were used to create a total of eight libraries. Short reads (~36 bp) from each library were sequenced on the Illumina Genome Analyzer and assembled using the short-read assembler Velvet (Zerbino and Birney 2008). Contigs assembled from each library were merged into a single nonoverlapping meta assembly using the lightweight assembly program Minimus (Sommer et al. 2007). Furthermore, we sequenced genomic paired-end libraries with short and long inserts to order and orient the contigs into larger genomic scaffolds. When compared with a whole-genome shotgun assembly of the same data, we produce a higher quality assembly more rapidly by reducing the biological complexity and computational cost to assemble each library. Finally, we compare this assembly to the dm3 fly reference to highlight the accuracy of our assembly and utility for comparative sequence analyses. Our results demonstrate that this method is a rapid and cost-effective means to generate high-quality de novo assemblies of large genomes.

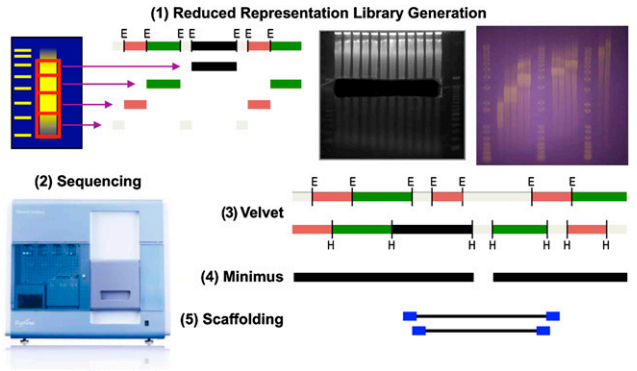


Figure 1. A schematic overview of the sequencing and assembly methods used to generate our de novo fly assembly. (*1*) Reduced representation libraries were created by digestion of genomic DNA with two restriction enzymes separately. Shown are a single library's gel slice and a subsequent second purification step to ensure library fragment fidelity. Resolution on an agarose gel allowed for libraries to be selected between 1–4 kb, 4–7 kb, 7–9 kb, and 9–30 kb in size for each enzyme independently. (*2*) Each library was then sequenced independently on the Illumina Genome Analyzer. (*3*) The short-read libraries were then assembled using Velvet. (*4*) Overlapping contigs from all eight libraries were merged using the lightweight assembler Minimus. (*5*) Finally, genomic paired-end short sequence reads were incorporated into the assembly process to order and orient the contigs generated in previous steps.

Results

Reduced representation library generation

An overview of our method is presented in Figure 1. Two sets of reduced representation (RR) libraries were prepared using genomic DNA from the *Sxl-EGFP-3* laboratory strain of *D. melanogaster* used for sexing embryos (Hempel et al. 2008). For each RR set, genomic DNA was digested to completion with EcoRI or HindIII and resolved on agarose gels to purify four distinct size-ranges: 1–4 kb, 4–7 kb, 7–9 kb, and 9–30 kb (see Methods). These fragment size boundaries and enzymes were established based on in silico analyses of the dm3 *D. melanogaster* reference genome (Adams et al. 2000; Celniker et al. 2002), which we expected (and subsequently showed) to be similar to the genomic sequence of our *D. melanogaster* strain.

We created fragment libraries from each purified RR aliquot and sequenced to roughly 30× base coverage on an Illumina Genome Analyzer (see Methods). Table 1 summarizes the amount of sequencing completed for each RR library. To determine the specificity of each library, we aligned the reads back to the dm3 reference genome and examined the overlap between the reads from each RR library and the predicted RR fragments from the dm3 genome. In general, we found the majority of reads aligned to the expected regions of the dm3 genome representing restriction fragment sizes commensurate with the originating RR library (Fig. 2). In addition, we quantitatively analyzed the proportion of reads from each RR library that matched a range of restriction fragment sizes from the dm3 genome (Fig. 3).

We note that alignment to the dm3 reference genome cannot entirely verify the accuracy of the partitioning scheme because a portion of EcoRI and HindIII sites are polymorphic between our individual fly's genome and the dm3 genome. Therefore, some fragments in a library expected to have high coverage, do not, because the set of predicted restriction fragments is not completely concordant between the two genomes. This leads to a bimodal distribution of the in silico predicted fragment coverage as summarized

Table 1. Summary of sequencing reads for each RR library, the theoretical size of each library, and the coverage based on alignment to the theoretical reduced representation (RR) library

Library	Theoretical size (Mb)	Reads sequenced (millions)	Bases sequenced (Gb)	Reads aligned in target RR library (%)
EcoRI1k4k	35.6	101.1	3.5	45.7
EcoRI4k7k	34.1	75.1	2.6	42.3
EcoRI7k9k	16.9	80.0	3.1	32.3
EcoRI9k30k	36.6	153.1	5.3	49.1
HindIII1k4k	35.2	73.0	2.7	44.4
HindIII4k7k	33.1	72.4	2.7	40.7
HindIII7k9k	16.1	167.5	6.5	15.4
HindIII9k30k	40.0	191.7	7.1	39.4
Total		913.5	33.6	37.2

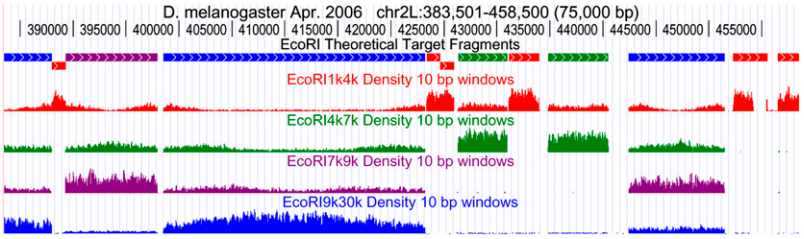


Figure 2. UCSC Genome Browser screen shot highlighting read coverage for each sequenced library. The *top* track represents theoretical fragments generated by in silico EcoRI restriction enzyme digestion. Read density tracks are color-coded by library as red (1–4 kb), green (4–7 kb), purple (7–9 kb), or blue (9–30 kb). The following four tracks are reads from each individual library aligned back to the dm3 reference using Illumina's short read aligner, ELAND, with standard parameters.

for the EcoRI1k4k library in Figure 4. The EcoRI1k4k results from this library accurately represent the findings for the other libraries (data not shown).

Assembly

Each of the eight RR short-read libraries were assembled separately using Velvet (Zerbino and Birney 2008) (see Methods) with the results summarized in Table 2. The Velvet assemblies were qualitatively analyzed based on alignment results to the dm3 reference genome (see Methods). Figure 5, top, describes the comparison of our assembled contigs with the theoretical fragments generated by in silico restriction enzyme digestion.

At this point we had created two distinct but overlapping libraries of nominally sized contigs aligning to 117.7 Mb of the 125-Mb dm3 reference genome. Next, we merged the overlapping contigs between the EcoRI and HindIII libraries using the assembler Minimus (Sommer et al. 2007). This resulted in a 121.4-Mb meta-assembly aligning to 116.2 Mb of the dm3 reference genome. The N50 contig length was 4243 bases. These results are summarized in Table 2 with alignments to the reference shown in Figure 5.

A separate WGS assembly was also performed for comparison with the RR approach. The eight RR libraries were pooled and assembled by presorting and hashing the data to allow the assembly to run at capacity on a 256 G-memory machine. The resulting assembly contained 113.4 Mb of sequence in 156,904 contigs with an N50 of 1444 bases. Table 2 summarizes comparison with the RR assembly. Figure 6 shows the contribution of each contig to the total assembly size for both assembly approaches. While the WGS assembly covered roughly the same amount of the genome, it was broken up into twice as many contigs with an N50 one-third less compared with the RR-generated assembly, highlighting a key benefit to our RR approach.

Finally, we created scaffolds of the Minimus contigs using genomic (i.e., non-RR) paired-end (PE) reads. PE libraries with different insert sizes were generated—a standard mate-pair library with an average insert size of 365 bp and a large-insert "jumping" library (Collins et al. 1987) with an average insert size of 2363 bp (see Methods). Components of the Phusion assembler (Mullikin and Ning 2003) were modified and used to order and orient the Minimus contigs based on the additional information contained in the PE reads (see Methods). The resulting assembly comprised 4718 scaffolds containing 118.5 Mb of sequence spanning 121.4 Mb (including gaps). The N50 scaffold length was 88,605 bases. In total, 95.5% of this assembly aligned to some portion of the dm3 reference genome. Specifically, there were 2766 scaffolds containing 115.4 Mb of our assembly that aligned to the euchromatic portion of dm3 (120.4 Mb). The remaining alignable scaffolds resided in heterochromatic regions (879 scaffolds representing 2.9 Mb of our assembly) or chromosome U (1.1 Mb in 484 scaffolds). Chromosome U contains *D. melanogaster* scaffolds that could not be unambigusouly placed when the dm3 assembly was created. It is this final assembly (summarized in Table 3) that we used for subsequent validation, quality assessment, and comparative analysis.

Analysis and quality assessment

Our assembly contained 589 scaffolds that did not align to the *D. melanogaster* reference genome. A total of 284 scaffolds had robust matches with entries in GenBank (see Methods). Three hits were specific to *D. melanogaster*, but not aligned by BLAT. The remaining hits were bacterial in origin and overlapped heavily with the fly microbiome (Corby-Harris et al. 2007). This analysis is summarized in Figure 7, with each resulting alignment organized by genus. The 284 hits to GenBank were spread over 38 species from 32 genera. Five of the top six represented genera, comprising 245 of the scaffolds, were from the Acetobacteraceae family. This family consists of proteobacteria that colonize rotting fruit, the primary food source of the fruit fly. These findings suggest that we sequenced genetic material from bacteria cohabitating with the flies sequenced and did not originate from a foreign source of contamination.

Next, we examined single-nucleotide differences between dm3 and our genome assembly. We used the SSAHA package (Ning et al. 2001) to determine the variation rate patterns across a variety of genomic features (Fig. 8). Low variation between strains was observed in first and second codon positions, regions independently annotated as conserved, and was lowest in splice-site junctions (0.05%), representing the upper bound of base-wise accuracy in our assembly. Conversely, the highest variation rates were seen in the third codon position. Variation was higher on the autosomes than the X chromosome, even in the randomized controls (data not shown). This pattern of variation is consistent

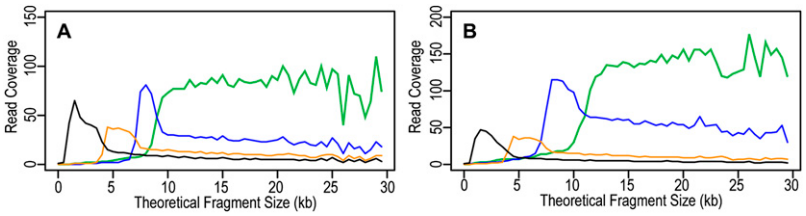


Figure 3. Read coverage for the four EcoRI (*A*) and four HindIII (*B*) RR libraries sequenced. The reads from each library were aligned to the theoretical fragments generated by restriction enzyme digestion. The short reads from each library 1–4 kb (black), 4–7 kb (orange), 7–9 kb (blue), and 9–30 kb (green) aligned to fragments corresponding to their expected size.

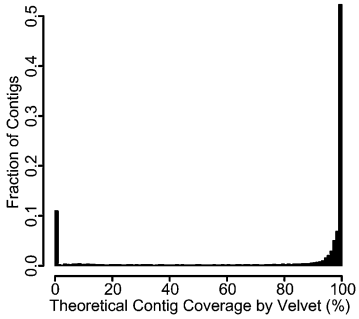


Figure 4. The coverage of the theoretical contigs in the EcoRI1k4k library by the EcoRI1k4k sequence reads exhibits a bimodal distribution. A majority of contigs are covered completely. However, one-tenth of the theoretical contigs are not covered at all. This is also observed in the other libraries assembled.

with expectations and provides strong support that the base-wise accuracy of our genome assembly is high.

In addition to single nucleotide variation between assemblies, larger insertion and deletion (indel) events were frequently discovered by alignment to the fly reference genome (see next paragraph). We note that this type of structural variation can only be detected at high resolution with a de novo assembly. Numerous LTR-bearing elements in the kilobase size range predicted by the dm3 reference were not present in our assembly. However, the scaffolds surrounding those regions aligned to the reference genome with high fidelity. Additionally, genomic paired-end reads independently validate our assembled scaffolds based on the average insert size from alignments back to the reference genome. For the LTR to exist in those genomic PE reads, the insert sizes would have to be several standard deviations above the mean. An example of this is marked by an asterisk in the middle panel of Figure 5.

We quantified the indel rate by examining the BLAT alignments of the assembled scaffolds against the dm3 reference sequence. Overall, we found 117,261 insertions in our fly's genome, averaging 35.9 bases in size. We also found 85,663 insertions in the dm3 reference fly's genome, averaging 88.9 bases in size. Indel events were observed less frequently in coding and conserved regions of the genome compared with other regions in the genome, providing support that the identified indel events are real and not assembly artifacts. These results are summarized in Figure 9.

In order to identify potential misjoins in our assembly, we examined scaffolds that unequivocally aligned to two distinct regions of the dm3 reference. Of the 4129 scaffolds that aligned using BLAT (see Methods), 52 scaffolds contained two separate regions (>20 kb) that aligned to different locations in the reference genome. One of these matched with chr U, and 31 had one or both ends overlapping repetitive regions, which are all known sources of error as well as hot spots for sources of rearrangement. These results provide good evidence for the high structural accuracy of our assembly.

One known difference between the dm3 reference genome and our assembled genome was the presence of a GFP construct attached to the promoter for the *Sex lethal* (*Sxl*) gene (Hempel et al. 2008). We located the construct in a single 147-kb scaffold. The first 3 kb of the scaffold aligned to chr X upstream of the *Sxl* gene. The terminal 143 kb aligned with 99.5% identity to chr 3L of the dm3 reference. The 1-kb intervening sequence aligned via BLAST with 98% identity to a GFP construct used in *D. melanogaster* (Swanson et al. 2008). This discovery supported that we sequenced the correct fly strain and can accurately resolve modestly sized features in the genome.

Cost

The Illumina sequencing for this project was completed across several flow cells over several months with constant improvements to cluster density, read length, and sequence quality. Additionally, each library was sequenced in excess to determine the optimal read coverage necessary for assembly. Based on current capacity (circa August 2009) the same project could be completed on a single flow cell with a 100-base sequencing run using paired-end reads from the outset for an approximate reagent cost of $11,500. In addition, we would gain additional improvements from utilizing paired-end reads from the outset—methodology that was not available initially. For comparison, whole-genome shotgun sequencing for a draft 8–10× assembly using capillary-based technology for the same-sized genome would cost roughly $600,000 and take 2–3 yr of run time on a single AB 3730xl, assuming 700-base read lengths and a reagent cost of 40 cents per read.

Discussion

We have created a de novo draft assembly of a *D. melanogaster* genome using a massively parallel short-read sequencing platform and reduced representation libraries. The reduced representation method simplifies the size and complexity of the assembly problem by breaking the genome into smaller portions, allowing us to assemble larger contigs than with the WGS approach. Additionally, the WGS assembly was conducted at the upper limit of current computational capabilities, requiring 256 G of memory compared with 32–64 G required for the RR approach. Thus, the RR method is scalable by dividing the assembly into smaller more computationally reasonable libraries.

The combination of algorithmic improvements and sequencing advancements will ease the future assembly of larger more complex genomes with this method. The availability of paired-end reads with ever increasing read lengths will facilitate the sequencing of RR libraries containing more fragments with fewer reads. This should allow for partitioning of mammalian-sized genomes into a relatively small number of fragment pools.

Table 2. Summary of each phase of the assembly process

	Bases (Mb)	No. of contigs	Contig N50	Aligned (Mb)
EcoRI libraries				
1–4 k	38.4	54,794	1332	38.0
4–7 k	38.3	61,377	1209	38.0
7–9 k	29.9	70,829	675	29.6
9–30 k	35.4	41,586	1854	34.6
HindIII libraries				
1–4 k	34.2	49,159	1319	33.8
4–7 k	43.5	61,289	1423	42.9
7–9 k	52.0	73,112	1505	50.7
9–30 k	33.7	47,870	1624	32.0
Minimus assembly	121.4	78,649	4243	116.2
WGS Velvet assembly	113.4	156,904	1441	115.8

Each library was assembled independently using Velvet. The eight RR libraries were merged using Minimus into a single assembly. For comparison with the Minimus assembly, the whole-genome shotgun (WGS) Velvet assembly is included. N50 contig length is the minimum contig length, such that all contigs larger than that size represent 50% of the assembly.

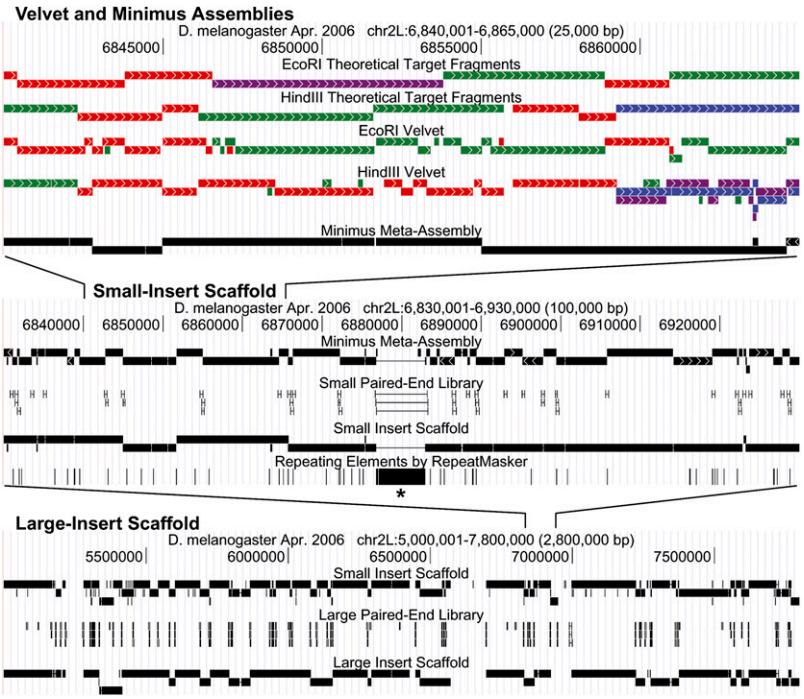


Figure 5. This is a screenshot from the UCSC Genome Browser exhibiting the alignment of each RR library to the reference genome. The *top four* tracks summarize the Velvet assembly steps. Contigs are color-coded by library as red (1–4 kb), green (4–7 kb), purple (7–9 kb), or blue (9–30 kb) bars. The *top two* tracks are the theoretical in silico restriction enzyme digested contigs. The actual EcoRI and HindIII contigs aligned back to the dm3 reference are show in the *next two* tracks. There is relatively little overlap between contigs in different libraries generated from the same restriction enzyme. The *next* track shows the RR meta-assembled contigs, resulting from merging the eight libraries with Minimus. Short and large genomic paired-end libraries facilitate the scaffolding of the contigs into the final assembly, shown in the *bottom* two panels. An LTR element deletion suggested by our assembly, marked by an asterisk (*) in the *middle* panel, was verified by alignment of the genomic paired-end reads flanking the deletion. If those read pairs contained the LTR element, their insert size would be several deviations larger than the mean for that library. The alignments depicted here are from the same region of chromosome 2 with the window zoomed out for each successive panel.

Similarly, algorithmic improvements will decrease the computational barriers to sequence assembly allowing for the assembly of larger RR libraries on commodity hardware. This will also benefit the WGS method of genome assembly. However, the biological complexity issue will persist, making our RR method useful for future large genome assemblies.

We selected EcoRI and HindIII for RR library generation based on typical laboratory availability, robustness, and in silico analyses of the dm3 reference genome. We initially simulated fragmentation with a variety of different type II restriction enzymes, which yielded different distributions of fragment sizes and theoretical coverage of the dm3 genome. However, multiple pairwise combinations of enzymes resulted in fragments covering >95% of the dm3 genome. Additional in silico analysis of larger, more complex genomes (hg18, mm9) revealed similar results with pairwise combinations of enzymes producing modestly sized fragments (1–30 kb) covering >95% of the genome. Based on these findings, we do not anticipate enzyme selection to be a barrier for future de novo genome assembly utilizing the RR approach. Indeed, where little is known about the genome content, gel electrophoresis analysis of genomic digestions using a variety of enzymes will likely yield robust results.

Our RR assemblies were made solely from fragment reads, using paired-end reads exclusively for scaffolding; the ability to generate accurate paired-end data was not available to us when we initiated this project. Utilizing paired-end reads for RR library assembly will improve the size and quality of future, more complex assemblies. Indeed, we have subsequently resequenced the EcoRI9k30k library with 50-base paired-end reads. We utilized 25.9 million read-pairs to produce a 33.5-Mb Velvet assembly (using *k*mer = 31 and expected coverage of 20) in 19,449 contigs with an N50 of 4514 bases (creating 9765 scaffolds with an N50 of 10.3 Kb). The original single read EcoRI9k30k assembly was 35.4 Mb in 41,586 contigs with an N50 of 1854 bases. While the paired-end assembly is relatively the same size, it is composed of much larger contigs, highlighting the advantage of using paired-end reads for the initial RR assembly step.

Variation for *Drosophila* populations has been reported between 0.4% in coding regions and 2% in noncoding regions, with autosomes varying more than sex chromosomes and *D. melanogaster* varying less than other species in the same genus (Moriyama and Powell 1996). Our findings support these conclusions. Our fly varies less than 2% in noncoding regions. This may result from our strain originating from a lineage closely related to the dm3 reference strain. However, the variation patterns observed across different annotated regions supports the accuracy of our assembly.

Confirming the presence of the GFP construct in our fly provided additional verification of our assembly's accuracy. The GFP construct containing the *Sxl* promoter region was inserted into chromosome 3. This discovery also demonstrates how our assembly method deals with segmental duplications. The *Sxl* promoter region was only represented once in our assembly. The endogenous promoter

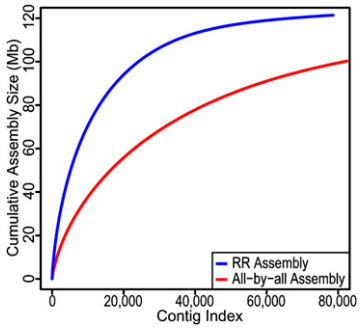


Figure 6. Contigs created by the RR approach were compared with those generated by the all-by-all WGS approach. The contigs were ordered by decreasing size. The figure depicts the increase in cumulative assembly size by adding successive contigs from each sorted list. The assembly size increases with fewer contigs for the RR approach than the all-by-all approach.

Table 3. The contigs generated with Velvet and Minimus were scaffolded using paired-end genomic reads (see Methods)

	Final assembly
Size (Mb)	121.4
No. of scaffolds	4718
Scaffold N50	88,605
Aligned to reference (Mb)	115.9
Base accuracy	99.95%
Structural accuracy	98.74%

Base accuracy was determined by examining variation in the invariable splice site junctions surrounding coding exons. Structural accuracy was estimated identifying scaffolds that unequivocally align to two locations (see Results). These accuracy measures represent an upper-level estimate on potential errors in our assembly, since it is expected that these two strains might differ slightly.

sequence from the X chromosome was merged with the transgenic version and assembled with the version inserted in chromosome 3. Additionally, most of the non-fly-aligning scaffolds represent inhabitants of the *D. melanogaster* microbiome. These findings demonstrate the ability of our assembly methods to accurately segregate sequence reads from different species into separate scaffolds, highlighting its potential for combined genomic/microbiome analyses.

Our analysis examined insertion and deletion events in our fly's genome relative to the dm3 reference genome. Interestingly, we found a distinct difference between deletion events occurring inside of regions annotated as coding DNA compared with noncoding DNA. We observed enrichment for deletions in coding sequence that were multiples of three bases in size (Fig. 9B). These findings are expected, given the low tolerance of protein synthesis to frameshift mutations and the triplet nature of the genetic code. This further suggests the utility of using de novo assembly by RR libraries to discover both small and large genomic variations with high precision.

The robustness of the RR method was demonstrated by the ability of this approach to tolerate poor quality in one of the eight libraries. We found low diversity in the reads from the HindIII7k9k library (i.e., we sequenced a large proportion of duplicate molecules). The diversity issue resulted from 16 cycles of PCR biasing the RR library fragments. This was corrected when only six cycles of PCR were used for additional sequencing in that library. However, the final overall assembly only improved slightly once the complexity issue in the HindIII7k9k library was corrected, or if the library was removed (data not shown). This suggests that the overlapping libraries created by two restriction enzymes accommodated moderate deteriorations in quality without a significant adverse effect to the resulting assembly. However, as larger and more complex genomes are assembled by this method, the tolerance to a poor library may decrease, as each library will represent a larger amount of sequence from the genome.

Our findings indicate that libraries of sufficient quality and complexity can be generated without the high cost and time requirements of a BAC clone library. However, there are shortcomings with using restriction enzymes for partitioning. Specifically, bias is introduced by the nonrandom location of restriction sites, and the library sizes will always vary to some degree, especially when applied to a previously unsequenced genome. This limitation is minimized by using two different restriction enzymes and is also offset by the relatively low cost and short time needed to implement the RR approach. Additionally, due to insert-size limitations for the paired-end scaffolding reads at the onset of this study, we could not provide chromosome scale scaffolding.

Our assembly appears more fragmented at both the contig and scaffold level when compared with Sanger-based *Drosophila* assemblies completed by the *Drosophila* 12 Genomes Consortium (2007). This is expected given our use of short (~36 bp) single-end reads for assembly instead of longer paired-end Sanger reads that can span modestly sized repeats. In addition, we used 2-kb insert paired-end reads for scaffolding instead of larger insert fosmids. However, from a comparative genomics standpoint, this is less important. The utility of our assembly lies in our ability to discern meaningful inferences about our fly's genome from comparison with the dm3 reference.

Our study shows the utility of using short-read sequencing to rapidly generate a de novo genomic assembly adequate for comparative genomic analyses. Additionally, the use of RR libraries is a scalable method that can be applied to assemble larger, more complicated genomes. Our assembly method combined with continual advancements in sequencing technology and computing power brings the promise of fast and cost-effective mammalian genome sequencing assembly closer to a reality.

Methods

Data availability

Sequence reads used for assembly are available at the NCBI Short Read Archive (accession no. SRA010040). All assemblies generated and analyzed in this study are available at ftp://kronos.nhgri.nih.gov/pub/outgoing/elliott/fly/.

Reduced representation libraries creation and sequencing

Genomic DNA preparations were digested to completion using EcoRI or HindIII and separated on an agarose gel. Libraries were recovered from gel slices 1–4 kb, 4–7 kb, 7–9 kb, and 9–30 kb in size. Individual libraries were sheared with a nebulizer using standard protocols. Illumina-specific adapters were ligated to the sheared fragments. Gel electrophoresis allowed for size selection of fragments around 250 bp. PCR amplification was used to select for fragments containing both Illumina-specific adapter sequences. Library complexity was reduced with the prescribed 16-cycle PCR amplification step (most notably in the HindIII7k9k library). PCR cycle count analysis was subsequently used to limit the cycles of PCR for each library. We found that six cycles of PCR improved the complexity of the sequenced libraries. A subsequent bead purification step eliminated unbound adapter sequences from the preparation. Single-end sequencing reads, 30–48 bases in length, for each of the eight libraries, were sequenced on the Illumina

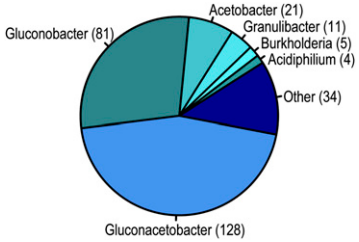


Figure 7. Distribution of GenBank alignment hits for nonfly sequence scaffolds. Of the 284 contigs discontinuous MEGABLAST aligned, 250 fell in only six genera. These top hits are all proteobacteria with all but Burkholderia residing in the Acetobacteraceae family.

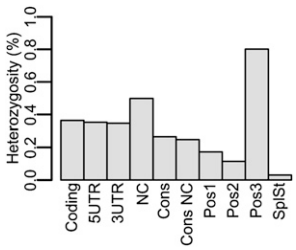


Figure 8. Our fly was compared with the dm3 reference across various annotated regions using SSAHA-SNP. Within annotated regions the variation rate matches previous *Drosophila* variation findings. The most conserved regions were the first (Pos1) and second (Pos2) codon positions, conserved sequence (Cons), and splice-site junctions (SplSt). The most variable region was the third codon position (Pos3). For comparison, annotated protein-coding regions (Coding), 5′ untranslated regions (5UTR), 3′ untranslated regions (3UTR), noncoding sequence (NC), and conserved noncoding sequence (Cons NC) were included.

Genome Analyzer II platform. Reads passing Illumina's chastity filtering parameters (0.6) were selected for assembly.

Assembly

The RR libraries were sequenced individually using Velvet. The parameters for Velvet were *k*mer = 23, cov_cutoff = 6, and min_contig_lgth = 100. Due to poor read coverage of the EcoRI7k9k library, a cov_cutoff = 10 was used to salvage contigs that could be assembled. Assembled contigs were aligned back to the reference genome using BLAT with the parameter max_Intron = 100, allowing for alignment of contigs containing small deletion events. Multiple degenerate hits were discounted from the reported alignment.

Reduced representation meta-assembly

We used a hierarchical RR meta-assembly scheme, first merging EcoRI and HindIII libraries separately. This specifically merged fragment assemblies that were represented partially in two or more libraries from the same enzyme. The two independent libraries were then merged together to create a single nonoverlapping RR meta-assembly.

The parameters for Minimus were overlap = 30, conserr = 0.1, minid = 95, and maxtrim = 20. Subsequent alignments were completed using BLAT with the max_Intron = 100 parameter.

Paired-end scaffold generation

We used the standard Illumina paired-end library preparation and sequencing methods to generate the short insert paired-end library. The large-insert library had a median 2363 bases between each read pair. Parts of the SOLiD mate-pair kit were used to generate this library. Briefly, this involved ligation of the EcoP15 restriction site containing adapter sequences, circularization with a biotinylated adapter sequence, digestion with the type III restriction enzyme EcoP15, and recovery of the fragments. Both libraries were sequenced using the previously mentioned pipeline.

Scaffold organization

The high-quality genomic paired-end reads were aligned to de novo assembled contigs with Illumina's short read aligner ELAND. Read pairs aligning without mismatches to the assembly were retained for scaffolding. Read pairs aligning within contigs were used to determine the average insert size and distribution for read pairs in the library, statistics necessary for the Phusion assembler to determine spacing between joined contigs. The first Phusion step used the pairing information to condense contigs that overlapped with each other but could not be joined by Minimus, because the overlap was too short or in low-complexity sequence. Subsequent iterations of Phusion joined contigs and scaffolds separated by 50, 100, 150, 200, 250, and 300 bases using the short-insert paired end libraries. At this point only scaffolds or contigs >300 bp were retained.

The final iterations of Phusion joined contigs and scaffolds separated by 300, 500, 750, 1000, 1500, 2000, 2500, and 3000 bases using the large-insert genomic paired-end library. Each step built upon the previous scaffolds to progressively join contigs and scaffolds further and further apart. There were a few modifications made to the software to reduce the rate of misassembly. At the read level, an aligned paired-end read was rejected if the alignment indicated that the insert size was five standard deviations away from the mean insert size. A join was prevented if the number of reads joining the two contigs together was not 30% of the total number of reads aligned to that end of the contig. Alignments to the reference genome were made using BLAT with the parameter maxIntron = 10,000 to allow for alignment when scaffolds contained large gaps. Only the top BLAT hit was counted from the alignment.

Variation analysis

FlyBase coordinates (Drysdale and Crosby 2005) for specific annotated regions were retrieved from the UCSC Table Browser (Kuhn et al. 2009). SNPs were examined between the two assemblies in annotated regions using SSAHA-SNP (Ning et al. 2001) without significant modification.

Microbiome determination

Scaffolds that did not align to the dm3 reference using BLAT were separated from the assembly. These scaffolds were aligned to

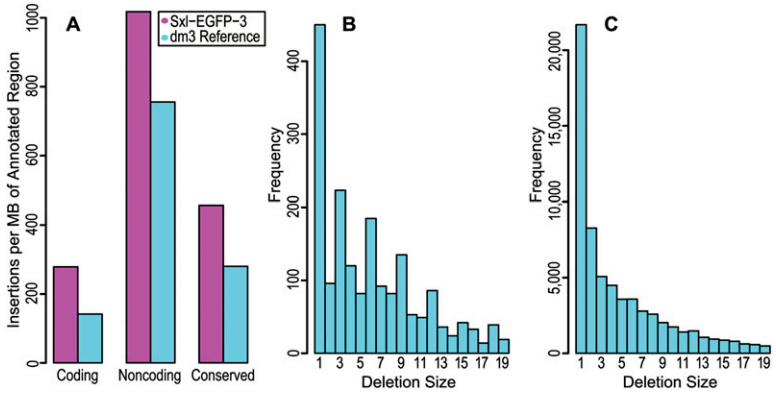


Figure 9. Depicted is a summary of indel events from comparison of the dm3 reference genome to the genome of our *D. melanogaster* individual. FlyBase annotations of the dm3 reference were used to determine the location of each indel event. (*A*) The number of insertions (magenta) and deletions (cyan) with respect to our fly are shown. The results are normalized by the megabases of sequence in each annotated track. (*B*) The distribution of deletion sizes is shown for regions annotated as coding sequence. An enrichment of deletion sizes that are multiples of three is visible, in addition to the underlying exponential decay visible in the deletion size distribution for regions annotated as noncoding (*C*).

GenBank using Nucleotide Discontiguous MegaBLAST. Only alignments that were over 100 bases in length and had an *E*-score < 10×10^{-10} were selected to improve the accuracy of the results.

Acknowledgments

We thank members of S. Salzberg's lab for assistance with implementing the Minimus assembler; J. Becker for computational assistance; L. Brody, E. Green, R. Blakesley, and an anonymous reviewer for helpful comments; Alice Young and the NIH Intramural Sequencing Center for sequencing support; and B. Oliver for providing the strain of *Drosophila* we sequenced and for helpful feedback. This research is supported in part by the Intramural research program of the National Human Genome Research Institute.

References

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, et al. 2000. The genome sequence of *Drosophila melanogaster*. *Science* **287:** 2185–2195.

Altshuler D, Pollara VJ, Cowles CR, Van Etten WJ, Baldwin J, Linton L, Lander ES. 2000. An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* **407:** 513–516.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456:** 53–59.

Blakesley RW, Hansen NF, Mullikin JC, Thomas PJ, McDowell JC, Maskeri B, Young AC, Benjamin B, Brooks SY, Coleman BI, et al. 2004. An intermediate grade of finished genomic sequence suitable for comparative analyses. *Genome Res* **14:** 2235–2244.

Butler J, MacCallum I, Kleber M, Shlyakhter IA, Belmonte MK, Lander ES, Nusbaum C, Jaffe DB. 2008. ALLPATHS: De novo assembly of whole-genome shotgun microreads. *Genome Res* **18:** 810–820.

Celniker SE, Wheeler DA, Kronmiller B, Carlson JW, Halpern A, Patel S, Adams M, Champe M, Dugan SP, Frise E. et al. 2002. Finishing a whole-genome shotgun: Release 3 of the *Drosophila melanogaster* euchromatic genome sequence. *Genome Biol* **3:** RESEARCH0079. doi: 10.1186/gb-2002-3-12-research0079.

Collins FS, Drumm ML, Cole JL, Lockwood WK, Vande Woude GF, Iannuzzi MC. 1987. Construction of a general human chromosome jumping library, with application to cystic fibrosis. *Science* **235:** 1046–1049.

Corby-Harris V, Pontaroli AC, Shimkets LJ, Bennetzen JL, Habel KE, Promislow DE. 2007. Geographical distribution and diversity of bacteria associated with natural populations of *Drosophila melanogaster*. *Appl Environ Microbiol* **73:** 3470–3479.

Drosophila 12 Genomes Consortium. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* **450:** 203–218.

Drysdale RA, Crosby MA. 2005. FlyBase: Genes and gene models. *Nucleic Acids Res* **33:** D390–D395.

Eddy SR. 2005. A model of the statistical power of comparative genome sequence analysis. *PLoS Biol* **3:** e10. doi: 10.1371/journal.pbio.0030010.

The ENCODE Project Consortium. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447:** 799–816.

Hempel LU, Kalamegham R, Smith JE III, Oliver B. 2008. *Drosophila* germline sex determination: Integration of germline autonomous cues and somatic signals. *Curr Top Dev Biol* **83:** 109–150.

Hernandez D, Francois P, Farinelli L, Osteras M, Schrenzel J. 2008. De novo bacterial genome sequencing: Millions of very short reads assembled on a desktop computer. *Genome Res* **18:** 802–809.

Kuhn RM, Karolchik D, Zweig AS, Wang T, Smith KE, Rosenbloom KR, Rhead B, Raney BJ, Pohl A, Pheasant M, et al. 2009. The UCSC Genome Browser Database: Update 2009. *Nucleic Acids Res* **37:** D755–D761.

Ley TJ, Mardis ER, Ding L, Fulton B, McLellan MD, Chen K, Dooling D, Dunford-Shore BH, McGrath S, Hickenbotham M, et al. 2008. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456:** 66–72.

Margulies EH, Cooper GM, Asimenos G, Thomas DJ, Dewey CN, Siepel A, Birney E, Keefe D, Schwartz AS, Hou M, et al. 2007. Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. *Genome Res* **17:** 760–774.

Meyer M, Stenzel U, Myles S, Prufer K, Hofreiter M. 2007. Targeted high-throughput sequencing of tagged nucleic acid samples. *Nucleic Acids Res* **35:** e97. doi: 10.1093/nar/gkm566.

Moriyama EN, Powell JR. 1996. Intraspecific nuclear DNA variation in *Drosophila*. *Mol Biol Evol* **13:** 261–277.

Mullikin JC, Ning Z. 2003. The Phusion assembler. *Genome Res* **13:** 81–90.

Ning Z, Cox AJ, Mullikin JC. 2001. SSAHA: A fast search method for large DNA databases. *Genome Res* **11:** 1725–1729.

Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, Minovitsky S, Dubchak I, Holt A, Lewis KD, et al. 2006. In vivo enhancer analysis of human conserved non-coding sequences. *Nature* **444:** 499–502.

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I. 2009. ABySS: A parallel assembler for short read sequence data. *Genome Res* **19:** 1117–1123.

Sommer DD, Delcher AL, Salzberg SL, Pop M. 2007. Minimus: A fast, lightweight genome assembler. *BMC Bioinformatics* **8:** 64. doi: 10.1186/1471-2105-8-64.

Stark A, Lin MF, Kheradpour P, Pedersen JS, Parts L, Carlson JW, Crosby MA, Rasmussen MD, Roy S, Deoras AN, et al. 2007. Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. *Nature* **450:** 219–232.

Sundquist A, Ronaghi M, Tang H, Pevzner P, Batzoglou S. 2007. Whole-genome sequencing and assembly with high-throughput, short-read technologies. *PLoS One* **2:** e484. doi: 10.1371/journal.pone.0000484.

Swanson CI, Hinrichs T, Johnson LA, Zhao Y, Barolo S. 2008. A directional recombination cloning system for restriction- and ligation-free construction of GFP, DsRed, and lacZ transgenic *Drosophila* reporters. *Gene* **408:** 180–186.

Van Tassell CP, Smith TP, Matukumalli LK, Taylor JF, Schnabel RD, Lawley CT, Haudenschild CD, Moore SS, Warren WC, Sonstegard TS. 2008. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nat Methods* **5:** 247–252.

Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, Fan W, Zhang J, Li J, Guo Y, et al. 2008. The diploid genome sequence of an Asian individual. *Nature* **456:** 60–65.

Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M. 2005. Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature* **434:** 338–345.

Xie X, Mikkelsen TS, Gnirke A, Lindblad-Toh K, Kellis M, Lander ES. 2007. Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. *Proc Natl Acad Sci* **104:** 7145–7150.

Zerbino DR, Birney E. 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18:** 821–829.

Received July 1, 2009; accepted in revised form November 11, 2009.