Selective Whole Genome Amplification for Resequencing Target Microbial Species from Complex Natural Samples

Aaron R. Leichty and Dustin Brisson[1]
Department of Biology, University of Pennsylvania, Philadelphia, Pennsylvania 19104-6018

ABSTRACT Population genomic analyses have demonstrated power to address major questions in evolutionary and molecular microbiology. Collecting populations of genomes is hindered in many microbial species by the absence of a cost effective and practical method to collect ample quantities of sufficiently pure genomic DNA for next-generation sequencing. Here we present a simple method to amplify genomes of a target microbial species present in a complex, natural sample. The selective whole genome amplification (SWGA) technique amplifies target genomes using nucleotide sequence motifs that are common in the target microbe genome, but rare in the background genomes, to prime the highly processive phi29 polymerase. SWGA thus selectively amplifies the target genome from samples in which it originally represented a minor fraction of the total DNA. The post-SWGA samples are enriched in target genomic DNA, which are ideal for population resequencing. We demonstrate the efficacy of SWGA using both laboratory-prepared mixtures of cultured microbes as well as a natural host–microbe association. Targeted amplification of *Borrelia burgdorferi* mixed with *Escherichia coli* at genome ratios of 1:2000 resulted in $>10^5$-fold amplification of the target genomes with <6.7-fold amplification of the background. SWGA-treated genomic extracts from *Wolbachia pipientis*-infected *Drosophila melanogaster* resulted in up to 70% of high-throughput resequencing reads mapping to the *W. pipientis* genome. By contrast, 2–9% of sequencing reads were derived from *W. pipientis* without prior amplification. The SWGA technique results in high sequencing coverage at a fraction of the sequencing effort, thus allowing population genomic studies at affordable costs.

CLASSICAL population genetics, coupled with advances in coalescent modeling, has been foundational to studies of the evolutionary histories and ecological forces that shape natural populations (Rosenberg and Nordborg 2002; Hume *et al.* 2003; Wakeley 2004). However, detecting fine scale processes using population genetics and coalescent analyses is limited by the amount of available sequence data per sample. Datasets with substantially greater genetic information per sample, such as genomic data from population-level sampling, would be optimal to study biological processes at all relevant scales. The promise of population genomics for many microbial species is tempered, however, by the difficulty of isolating and preparing microbial genomes for next-generation sequencing. Currently, sequencing microbial genomes requires laboratory culture to isolate them from other organisms with which they are naturally associated to obtain the appropriate samples for sequencing—sufficient numbers of the target genome with limited contaminating DNA (Mardis 2008).

Methodological issues in obtaining populations of genomes from microbial species occur both because the target microbial genomes often constitutes only a miniscule fraction of the DNA in complex, field-derived samples and because many important microbial species are difficult to isolate and culture. The shotgun approach of next-generation sequencing provides very limited sequence coverage of the rare microbial genomes from these samples, thus requiring laboratory culture prior to sequencing. However, the overwhelming majority of microbes cannot be cultured in the laboratory. Thus, a primary hindrance to collecting population genomic data from microbes is the need for a cost-effective, practical, and unbiased method to collect sufficient amounts of microbial genomic DNA from the

doi: 10.1534/genetics.114.165498
Manuscript received April 21, 2014; accepted for publication July 22, 2014; published Early Online August 5, 2014.
Supporting information is available online at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.165498/-/DC1.
Sequence data have been deposited in the European Nucleotide Archive (study accession no. PRJEB6142).
[1]Corresponding author: Department of Biology, University of Pennsylvania, Leidy Laboratories, 209, 433 S. University Ave., Philadelphia, PA 19104-6018.
E-mail: dbrisson@sas.upenn.edu

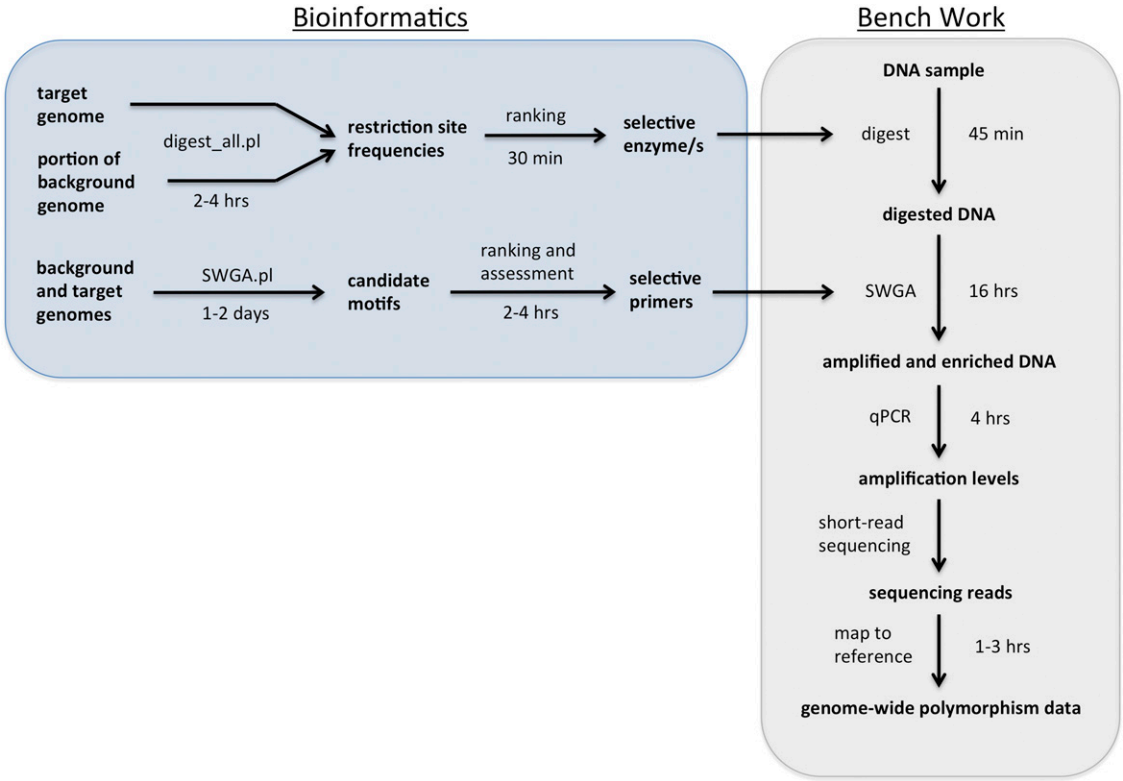

Figure 1 Schematic overview of the bioinformatic and laboratory steps involved in selective whole genome amplification. Target and background genomes were investigated using PERL scripts (SWGA.pl) to select primers that are underrepresented in the background genome and overrepresented in the target genome. Additionally, restriction enzymes with cut sites that are overrepresented in the background (digest_all.pl) are used to pretreat genomic DNA extracts. Digested samples are then amplified using selective primers and phi29 polymerase to enrich the sample in target DNA. Amplification success is evaluated using qPCR of multiple regions in the target and background genomes. Samples can then be used for high-throughput genome sequencing and mapped to a reference genome to obtain genome-wide polymorphism data. The entire SWGA pipeline, from initial bioinformatics to submitting samples for sequencing can be completed in less than 1 week and multiple samples can be run simultaneously.

target species while limiting the amount of contaminating DNA from organisms with which the target microbe naturally associates.

Here we present a culture-free technology applied to genomic preparations directly from complex environmental samples that results in high concentrations of a target microbial genome with limited contaminating DNA. The selective whole genome amplification (SWGA) technique was developed to amplify only a specified target genome from total genomic extracts derived from an environmental sample (Figure 1). Thus, SWGA is akin to PCR in that a specific portion of the DNA in a sample is enriched. Amplifying an entire genome by SWGA differs technically from amplifying a single gene region by PCR both in how primers are chosen and the amplification technology used. The primers used for SWGA bind to DNA sequence motifs that are common in the target genome but rare in the genomes of other species present in the environmental sample. The SWGA procedure takes advantage of the inherent differences in the frequencies of sequence motifs among species to design primers specific to a target species. These primers are then used to selectively amplify the target microbial genomes using phi29 multidisplacement amplification technology (Dean *et al.* 2001, 2002). The phi29 polymerase is strand displacing and amplifies DNA from primers with high processivity (up to 70-kbp fragments) and is 100 times less error prone than Taq (Fuller *et al.* 2009), making it ideal for genome amplification prior to sequencing (Rodrigue *et al.* 2009; Blainey 2013; McLean *et al.* 2013). We present evidence of the potential to selectively amplify the genome of a target species from a complex sample using both experimental mixtures of two bacterial species as well as selectively amplifying and sequencing *Wolbachia pipientis*, a natural endosymbiont of *Drosophila melanogaster*, from a complex, natural genomic DNA preparation.

Materials and Methods

Samples

Genomic extracts from cultures of *Borrelia burgdorferi* (strain B31) and *Escherichia coli* (strain BL21) were prepared using the Qiagen DNeasy kit. Total DNA from each genomic preparation was quantified using UV absorbance (NanoDrop, Thermo Scientific). Laboratory-generated complex samples were prepared with 40 ng of *E. coli* DNA

Table 1 qPCR data from *B. burgdorferi-E. coli* SWGA

Expected *B. burgdorferi*:*E. coli* genome ratios	Average fold amplification ($2^{-\Delta CT}$)	Quantity of *B. burgdorferi* relative to *E. coli*	
	B. burgdorferi (*E. coli*)	Pre-SWGA	Post-SWGA
1:2	466.7 (1.7)	1.215	326.034
1:20	5074.6 (3.0)	0.065	110.337
1:200	18991.1 (4.9)	0.009	34.924
1:2000	115777.1 (6.7)	0.001	13.117

($\sim 10^{6.9}$ genome copies) combined with *B. burgdorferi* DNA at 1:2, 1:20, 1:200, and 1:2000 genome copy-number ratios to evaluate the SWGA method (Table 1). Individual *D. melanogaster* females of isofemale lines from K18 Vienna and MD9 Cameroon were extracted for genomic DNA using a Qiagen Gentra Puregene kit according to the manufacturer's protocol. Sample quality was assessed using gel electrophoresis and UV absorbance. In anticipation of overamplification of *D. melanogaster* mtDNA due to rapid rolling-circle amplification with phi29, all restriction enzymes in REBASE (Roberts *et al.* 2010) were ranked by ratio of cut sites in the *W. pipientis* genome and in the mitochondrial genome and chromosome 4 of *D. melanogaster* (where lower is better) (digest_all.pl; Supporting Information, File S1). The restriction enzyme *Nar*I was chosen for empirical evaluation. The SWGA procedure was performed on both *Nar*I-digested and -undigested extracts from each fly (Table 2). Digested samples (40 ng) were digested with 1 unit of *Nar*I in 1× NEB4 reaction buffer at 37° for 30 min, followed by heat inactivation at 65° for 15 min.

Selection of amplification primers

We created a PERL script (SWGA.pl; File S2) to identify primers for selective genome amplification. The PERL script quantifies the number of times all motifs of user-defined lengths occur in the target species genome and nontarget species genomes (background). Highly selective primers are those that bind motifs that are common in target genome and rare in background genomes. The list of highly selective primers are then filtered to remove those with predicted melting temperatures, $T_m = 4(N_G + N_C) + 2(N_A + N_T)$ °, above a user defined value. *B. burgdorferi*-selective primers were identified using the *B. burgdorferi* strain B31 genome (GenBank AE000783.1) (Fraser *et al.* 1997) and the *E. coli* strain BL21 genome (GenBank AM946981.2) using 12-bp motifs with melting temperatures <30° (the optimal temperature for phi29 amplification). Primer pairs with runs of greater than three pairing nucleotides at their ends were removed to prevent potential primer dimerization. *W. pipientis*-selective primers were identified using the *W. pipientis* (GenBank NC_002978.6) (Wu *et al.* 2004) and *D. melanogaster* (FlyBase v5.9, all linear chromosomes, mitochondria, and unplaced scaffolds) (Marygold *et al.* 2013) genomes with a motif length of 8–12 bp, and T_m cutoff of 30°. Motifs were then ranked on their ratio of occurrence between the *W. pipientis* and *D. melanogaster* genome (where higher is better), frequency of occurrence in the *D. melanogaster* mitochondrial genome (where lower is better), and their predicted melting temperature (where lower is better). The most selective primers were aligned to the mitochondrial genome and only motifs with one or more mismatches in the 3′ end were selected to reduce mitochondrial sequence amplification. This restriction was imposed on mitochondrial DNA because even small amounts of nonspecific priming can lead to rapid "rolling-circle" amplification of this small, circular element resulting in exorbitant mitochondrial amplification (data not shown). In cases where a motif was wholly or partially contained within a longer motif, the longer motif was selected.

Using these computational approaches, one primer set consisting of 20 primers was selected to selectively amplify *B. burgdorferi* and two primer sets consisting of 10 and 2 primers were selected to selectively amplify *W. pipientis* (Table S1). The primers in the smaller *W. pipientis*-specific set were the two primers from the larger primer set with the least permissive melting temperatures (22.1° and 22.2°). Primers were ordered from Integrated DNA Technologies with phosphorothioate bonds between the two most 3′ nucleotides to prevent primer degradation by phi29. There was no expectation that *B. burgdorferi*-specific primers would be useful for selective amplification of *W. pipientis*, given the inherent differences in motif frequencies among species. Consistent with this expectation, there was no overlap among the primers in the sets chosen to amplify *B. burgdorferi* and those chosen to amplify *W. pipientis*.

Selective whole genome amplification

All samples were equilibrated at 35° for 5 min and then combined with 30 units of phi29 polymerase (New England Biolabs), reaction buffer and BSA to 1×, dNTPs to 1 mM, and each amplification primer to 2.5 μM in a final volume of 50 μl. Samples were then run with a "stepdown" protocol consisting of 35° for 5 min, 34° for 10 min, 33° for 15 min, 32° for 20 min, 31° for 30 min, 30° for 16 hr, and 65° for 15 min. Success of amplification was assessed by qPCR using qPCR primer sets arrayed across *B. burgdorferi*, *E. coli*, and the *W. pipientis* genome (five sets per species) and one qPCR primer set for each arm of each *D. melanogaster* chromosome and mitochondria (seven sets; Table S2). Standard curves were used to ensure that all qPCR primer sets had measured efficiencies between 90 and 110%. SWGA reactions were diluted 100× in water and used for qPCR with Power SYBR Green master mix (Life Technologies) on an ABI StepOne Plus. Fold amplification for samples was calculated as $2^{-\Delta CT}$, where threshold values (CT) of SWG amplified samples were

Table 2 Summary of SWGA sequencing

Sample	Treatment conditions	Bases sequenced (after filtering)	*D. melanogaster* mapping reads (%)	*W. pipientis* mapping reads (%)	Average predicted coverage[a]
Fly 1 (MD9 Cameroon)	*Nar*I, SR2	159e + 6	58.7	26.6	22.7
	Control	359e + 6	81.0	2.4	1.9
Fly 2 (K18 Vienna)	*Nar*I, SR2	88e + 6	27.0	62.2	53.3
	Control	451e + 6	80.8	7.6	6.1
Fly 3 (K18 Vienna)	*Nar*I, SR2	161e + 6	19.8	70.4	59.8
	Control	415e + 6	79.9	8.8	7.1

[a] Actual coverage normalized to throughput from a single lane on a HiSeq 2000 with 96 multiplex samples and 50-bp SE reads (104,166,667 bp per sample).

compared to threshold values from nonamplified controls (no phi29 added) of the same samples.

Library preparation and sequencing

W. pipientis-amplified and unamplified sample libraries were prepared for high-throughput sequencing using the Nextera DNA Sample Preparation kit following the manufacturer's protocol (Illumina). Briefly, samples were column purified and concentrated (Zymo Research) and quantified using a quBIT fluorometer (Invitrogen). A total of 50 ng of sample was then subjected to a 5-min tagmentation reaction, followed by column purification (Zymo Research). Adapters were added using PCR and products were size selected and purified using Agencourt AMPure beads with a single binding reaction at 0.6×. The resulting libraries were quantified using a Qubit fluorometer (Invitrogen) and size distributions were analyzed using a Bioanalyzer (Agilent). The resulting libraries were multiplexed on an Illumina MiSeq for 150 paired-end sequencing.

Data analysis and read mapping

Reads were first trimmed of low-quality bases from their ends using Prinseq-lite (v0.20.3) (Schmieder and Edwards 2011) with the following parameters: -trim_qual_left 20, -trim_qual_right 20, -trim_qual_rule lt, -trim_qual_window 1, -trim_qual_step 1, -trim_qual_type min, -ns_max_p 10 -min_len 20 -min_qual_mean 20. These parameters first eliminated bases sequentially from either end of a read until a base had a quality score ≥20 (Phred+33) and then eliminated reads with >10% ambiguous bases, lengths <20 bases, or an average quality score <20. These reads were then mapped as paired reads to the *D. melanogaster* genome (r5.9) using Bowtie2 (v2.1.0) (Langmead and Salzberg 2012) with the following parameters: -I 0 -X –fr –score-min L,0,-0.1466667. These parameters return the best scoring alignment with a minimum alignment score of −44 for 300-bp reads (2 × 150) (*e.g.*, five mismatches and a 3-bp gap), fragments can map up to 800 bp between outer ends and are allowed to overlap and contain each other (dovetails not allowed). To reduce the likelihood of falsely identifying a read as coming from *W. pipientis*, *D. melanogaster*-mapped reads were filtered out and the remaining paired-end reads were mapped to the *W. pipientis* genome (NC_002978.6) using identical mapping parameters.

GenomeCoverageBed (v2.14.2) (Quinlan and Hall 2010) was used to generate single-base resolution coverage maps for both *W. pipientis* and *D. melanogaster*. The resulting coverage values were divided by the number of total bases sequenced and multiplied by 1 million to obtain a normalized measure of coverage for each sample (reads per million bases sequenced, RPMS). Estimates of predicted coverage were calculated by multiplying per-site RPMS by 104.1667 (200,000,000 reads × 50 bp read length/96 samples/1,000,000), a factor that reflects the expected throughput of an Illumina HiSeq run multiplexing 96 samples with 50-bp single-end sequencing. This metric denotes the expected output from a HiSeq run while normalizing samples to allow direct comparisons among samples. Estimates of amount of sequencing needed to obtain 10× coverage over 50 and 90% of a genome for a given sample were obtained from nonlinear models fit using *pcrfit* in the qpcR package (Ritz and Spiess 2008) of R (v2.15.1) (R Development Core Team 2012). These models describe the relationship between the $\log_{10}$ transformed number of bases sequenced and number of sites in the genome with >10× coverage.

SAMtools (Li *et al.* 2009) was used to call bases and extract whole genomes from alignments using the mpileup | bcftools | vcfutils.pl pipeline with a minimum coverage cutoff of 8× to analyze variant calls between samples (http://samtools.sourceforge.net/mpileup.shtml). The rate of chimera formation induced by SWG amplification, a potential problem caused by phi29 amplification (Lasken and Stockwell 2007), was investigated by examining the alignments of reads that did not initially map to *D. melanogaster* nor *W. pipientis*. Using an approach similar to Lasken and Stockwell (2007), nonmapping reads were aligned against the *D. melanogaster* and *W. pipientis* genomes using BLAST (Altschul *et al.* 1997), and chimeric sequences were defined as reads that mapped to two distinct locations where the total alignment length was less than the read length + 10 bp.

Results

The SWGA procedure effectively enriched samples in target genomes—*B. burgdorferi* or *W. pipientis*—in all tested genomic extracts. *B. burgdorferi* mixed with *E. coli* at genome copy number ratios of 1:2, 1:20, 1:200, and 1:2000 was used to evaluate the SWGA technique with primers that bind

commonly in *B. burgdorferi* but rarely in *E. coli*. SWG amplification of both *B. burgdorferi* and *E. coli* was quantified by comparing qPCR values of five *B. burgdorferi* and five *E. coli* loci for SWGA amplified and nonamplified control (no phi29 included) samples. In all cases, amplification of *B. burgdorferi* was orders of magnitude greater than that of *E. coli* resulting in a dramatic increase in the relative proportion of *B. burgdorferi* DNA (Table 1). Despite being rare in the original samples, *B. burgdorferi* genomes dominated the post-SWG amplification samples making up the overwhelming majority of all DNA in all cases (Table 1). *B. burgdorferi* loci were near the limit of qPCR detection in the 1:2000 *B. burgdorferi*:*E. coli* preamplification mixture such that the reported estimates of fold amplification for these samples is very conservative.

Amplification of *W. pipientis* genomes was also much greater than amplification of *D. melanogaster* genomes in all samples regardless of the primer set employed. Amplification was quantified by comparing qPCR values at five *W. pipientis* and seven *D. melanogaster* loci for each sample before and after amplification. SWGA using the 10 primers in set SR1 resulted in substantial amplification of the *W. pipientis* DNA with limited amplification of *D. melanogaster* DNA (Figure 2). Cutting the sample with a restriction enzyme prior to SWGA dramatically increased the amplification of *W. pipientis* genomes and reduced total *D. melanogaster* amplification. The two primers from this set with the least permissive melting temperatures (set SR2) further decreased amplification from *D. melanogaster*-derived DNA (Figure 2). Restriction digest of samples prior to SWG amplification further increased the amplification of *W. pipientis* DNA, likely due to lower polymerase time directed toward nontarget amplification. Combining the least permissive primer set (SR2) with restriction digestion prior to SWG amplification resulted in nearly 140 times greater amplification of *W. pipientis* than chromosomal *D. melanogaster* DNA (1264-fold *vs.* 9-fold). The pre-SWGA genomic extracts from all flies, as well as the post-SWGA samples using the SR2 primer set derived from those flies, were prepared for whole genome sequencing on an Illumina MiSeq platform.

Sequencing of *W. pipientis* genomes directly from post-SWGA samples is substantially more efficient than sequencing directly from total genomic extracts of *D. melanogaster*. Sequencing from total genomic extracts of three female *D. melanogaster* flies on an Illumina MiSeq platform resulted in only 2.4–8.8% of the reads deriving from *W. pipientis*, while ~80% were derived from *D. melanogaster* (Table 2). The remaining reads mapped to neither genome due in part to the stringent mapping parameters used during analysis. These conservative mapping parameters allowed for unequivocal mapping to the correct source genome, which is essential to effectively evaluate this methodology. To achieve at least 10× coverage across 90% of the *W. pipientis* genome by deep sequencing of a *D. melanogaster* genomic extract would require sequencing 1.3–8.7 billion bp per sample, equivalent to 13–87% of one Illumina HiSeq lane (Figure 3).

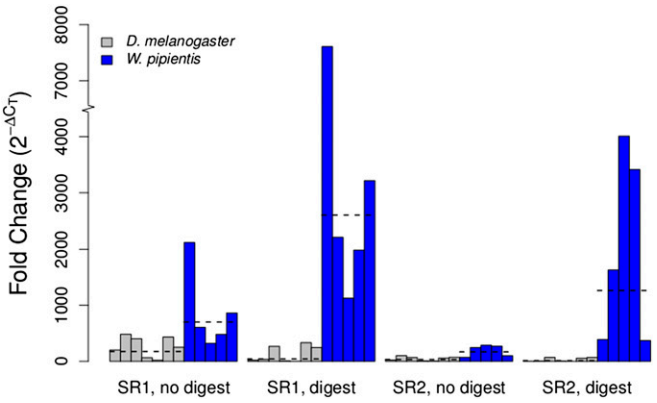

Figure 2 Selective amplification of *W. pipientis* DNA from genomic extracts of infected *D. melanogaster* using different primer sets. The relative concentration of *W. pipientis* DNA, as measure by qPCR, increased as much as 7500-fold after selective genome amplification while *D. melanogaster* DNA had limited amplification. Fold amplification after SWGA of each sample was calculated relative to qPCR levels prior to SWG amplification at seven *D. melanogaster* sites (one per chromosome arm and one on the mitochondria) and five sites around the *W. pipientis* genome, averaged across three biological replicates. The degree of amplification differed among regions of the *W. pipientis* genome, depending on the primer set used (SR1 and SR2). Restriction digestion of the sample prior to amplification resulted in substantial improvements in selectively amplifying the target *W. pipientis* genome for all primer sets used. Horizontal bars represent means across all loci within a species.

Sequencing post-SWGA samples from the same three female *D. melanogaster* flies resulted in significantly more reads mapping to *W. pipientis* than from sequencing *W. pipientis* directly from *D. melanogaster* genomic preparations. Similarly, a much smaller proportion of contaminating *D. melanogaster* was sequenced from the postamplified samples (Table 2). Although variation in sequence coverage across the *W. pipientis* genome remained in the postamplification samples (Figure S1), only 1–6% of the genome had very low coverage (<2×). The majority (56–91%) of the *W. pipientis* genome from each sample had deep coverage (>10×). Interestingly, the areas of the *W. pipientis* genome with high coverage were consistent across the three samples investigated (Figure S2). To achieve at least 10× coverage across 90% of the *W. pipientis* genome would require 0.6–2.2 billion bp sequenced per sample (Figure 3), corresponding to 6–22% of one Illumina HiSeq lane. Interestingly, to obtain 10× coverage across only 50% of the genome would only require 2–5% of a HiSeq lane per SWGA sample, whereas non-SWGA samples would require as much as 36% (Figure 3).

The selective genome amplification technique showed no evidence of introducing point mutations or indels due to amplification prior to sequencing. In our datasets, the total number of bases that differed between the reference genome (wMel) and the post-SWGA sequenced samples was similar to the total number of bases that differed between the reference genome and the preamplification samples sequenced from the same fly. Further, the sites that

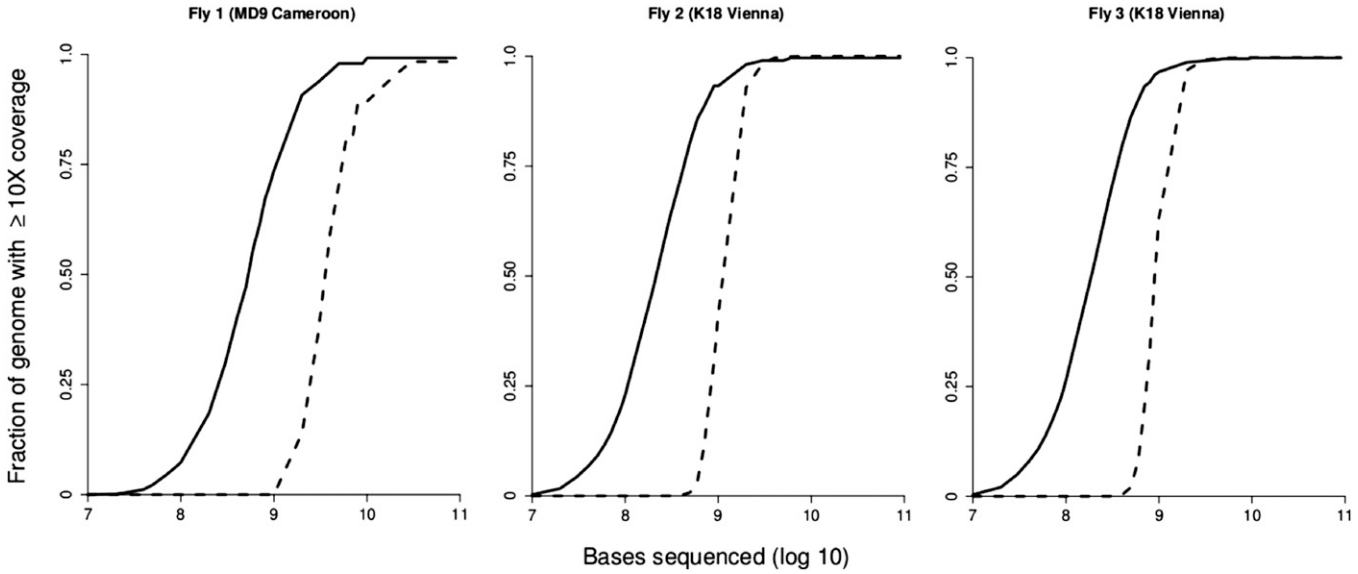

Figure 3 Sequence coverage improves across the *W. pipientis* genome due to selective whole genome amplification (SWGA) in all three *D. melanogaster–W. pipientis* samples tested. Selective amplification decreases the amount of sequencing necessary to achieve ≥10× coverage across the genome. For example, to achieve ≥10× coverage in 50% of the genome in fly 1 would require ~0.4 billion bp with SWGA (solid curve), but >3.6 billion bp without SWGA (dashed curve).

differed between the reference genome and the post-SWGA sequenced samples were congruent with the bases that differed between the reference genome and the preamplification samples, suggesting these variant bases are real differences between our strains and the reference genome (Table 3). Additionally, the rate of chimera formation due to phi29 amplification was very low (0.26–0.9% of reads, Table S4). No chimeric reads were included in the coverage or mutation analyses as all were filtered during the original mapping analysis. The high level of fidelity in the phi29 enzyme, which is currently used in single-cell genome sequencing (Pinard *et al.* 2006), makes it ideal for presequencing amplification.

Discussion

Many major outstanding questions in microbiology can be addressed through analyses of populations of genomes. Obtaining genomic sequence data at population levels is hindered by the absence of a cost-effective and practical method to collect sufficient amounts of microbial genomic DNA while limiting the amount of contaminating DNA necessary for efficient high-throughput sequencing. The SWGA technique is a simple, rapid, and cost-effective methodology to overcome this impediment. SWGA transforms a complex sample with nearly all DNA originating from nontarget species to a sample enriched for the target microbial genomic DNA. The SWGA procedure primes phi29 amplification from computationally selected primers that are frequent in the target genome but rare in the nontarget DNA. The resulting sample is enriched in target genomic DNA and is thus ideal for high-throughput sequencing. In all samples tested, the target genome made up only a small fraction of the total DNA prior to SWG amplification but became in many samples the overwhelmingly dominant fraction after SWG amplification. Further, as much as 70% of all sequencing reads were derived from the target genome, *W. pipientis*, after selective amplification of total genomic extracts from whole *D. melanogaster*. Selective amplification resulted in ~10-fold increase in sequence coverage of the target genome compared to unamplified samples in all tested samples. Thus, equivalent sequence coverage can be accomplished using less than one-tenth the number of sequencing reads, allowing studies of populations of genomes at research feasible costs.

The primers chosen to selectively amplify *W. pipientis* resulted in as much as 70% of the sequencing reads derived from the target genome while only 20% were derived from *D. melanogaster*. Thus, there is little room for significant improvement in sequencing efficiency by improved primer design. Improving primer selection criteria to increase sequencing efficiency will require experimentation in systems with exceedingly rare target DNA. However, there is room for improvement in the evenness of sequencing coverage of *W. pipientis* (Figure S2) although the current data are insufficient to identify all of the factors that affect the variation in amplification across the target genome. Proximity to restriction cut sites had a small, negative effect on sequence coverage, suggesting that digesting samples may result in a tradeoff between better overall amplification of the target genome and evenness in sequencing coverage (Figure S3 and Figure S4). This tradeoff may be circumvented by mixing multiple independent amplifications each treated with a different restriction enzyme prior to sequencing. Primer

Table 3 High level of shared SNPs between amplified and nonamplified samples

	SNPs with reference genome			Total number of called bases	
Sample	SWGA	No SWGA	Shared	SWGA	No SWGA
Fly 1	36	36	30	966,334	795,941
Fly 2	22	21	15	1,072,087	1,196,998
Fly 3	20	22	16	1,208,554	1,210,532

density had a minor correlation with sequencing coverage across the genome (Figure S3), but in regions surrounding predicted priming sites, sequencing coverage was positively correlated with distance from the priming site (Figure S5). However, there is considerable variation in sequencing coverage, suggesting other factors are important to amplification in the SWGA process. Using different primer sets for SWGA also resulted in different patterns of sequence coverage across the genome, suggesting that improvements in primer choice may affect evenness in coverage (Figure 2). Including nonspecific priming in the analysis did little to improve the correlation between sequence coverage and distance from a primer (Figure S5B). Similarly, there was little correlation between GC content and sequence coverage, although there was not sufficient variation in GC content across the *W. pipientis* genome to effectively assess this factor (Figure S6). It is important to note that next-generation sequencing technologies are inherently biased and uneven sequence coverage is common (Lam *et al.* 2012; Quail *et al.* 2012; Ross *et al.* 2013), similar to the data from our pre- and postamplification samples (Figure S2).

Despite the success in amplifying both of the investigated target species from complex samples, there remain several aspects of the technology that can be improved by further empirical research. In particular, studies of the mechanism of phi29 priming is essential to codify criteria for primer design. Currently, primer design requires substantial sequence information to identify the motifs that are common in the target and rare in the background. However, complete genomes are not essential for primer design. Using only a randomly selected fraction (0.1× in unassembled 100-bp windows) of both the *W. pipientis* and *D. melanogaster* genomes resulted in rankings of primers for *W. pipientis* selective amplification that were highly similar to those estimated from whole genome data (Figure S7). Importantly, the most selective primers were identical using either genome dataset. Not including sequence information from the microflora colonizing *D. melanogaster* in the computational analyses to design *W. pipientis*-specific primers also did not affect *W. pipientis* amplification.

Laboratory culture, the current standard to isolate microbes for genome sequencing, is possible for only a very small fraction of microbial species and is both time consuming and expensive (Wilson 2012). Further, laboratory culture may introduce sampling biases if different strains of a species are cultured with different efficiencies (Snyder *et al.* 2004; Gorski 2012), a problematic confounder for population genomic analyses (Beerli 2004; Simmons *et al.* 2008). The selective genome amplification technique amplified all of the *W. pipientis* strains tested without noticeable bias despite the samples originating from different continents (Figure S2). Further, the SWG amplification showed no evidence of introducing point mutations or indels (Table 3), suggesting utility for resequencing populations of microbial genomes.

The SWGA technology has many advantages over previously employed methodologies used to sequence genomes of unculturable microorganisms. Many methodologies physically separated the *W. pipientis* DNA from *D. melanogaster* DNA using differential centrifugation and pulse-field gel separation, which requires ~1000 live adult flies to obtain sufficient quantities of *W. pipientis* DNA for genomic sequencing (Sun *et al.* 2001; Wu *et al.* 2004), thus eliminating the possibility of acquiring *W. pipientis* genomes from individual flies. Very deep sequencing of flies (Richardson *et al.* 2012) or physical isolation of target DNA are becoming more feasible with technological advances (Richardson *et al.* 2012; Ellegaard *et al.* 2013), although they are still inefficient in both cost and labor and are thus prohibitive for population-level studies. The SWGA method maximizes the amount of target microbial DNA sequenced to create an efficient and cost-effective method to pursue population genomic studies.

The SWGA technology has the potential to amplify the genomic DNA of nearly any target species in a complex sample due to the intrinsic differences in the frequencies of nucleotide sequence motifs between species. These data will be useful for many applications, including fine-scale mapping of the location and timing of epidemiological outbreaks; identification of horizontal gene transfer among microbes or between hosts and microbes; identification of genomic regions that have experienced natural selection due to environmental changes, such as host species switches or migration to novel habitats; and identification of genetic loci that are associated with a particular trait or process. Additionally, the SWGA technology can be used for in-depth evolutionary or functional association analyses of one or several target species in microbiome samples. While we focused narrowly on microbial genomics, other researchers may repurpose the foundations from the SWGA technology for other applications such as amplifying large fragments of metazoan chromosomes.

Classical population genetics, coupled with advances in coalescent modeling, has been foundational to studies of the evolutionary histories and ecological forces that shape natural populations (Rosenberg and Nordborg 2002; Hume *et al.* 2003; Wakeley 2004). These analytical frameworks have identified genes under selection, characterized population structure and migration routes, characterized population dynamic and evolutionary processes, and identified mutations leading to epidemics and pandemics in pathogens (Grenfell *et al.* 2004; Deng *et al.* 2008; Holmes and Grenfell

2009; Humphrey *et al.* 2010; Hofinger *et al.* 2011; Castro-Nallar *et al.* 2012). However, precision of estimates from conventional population genetic methods are limited by the amount of available sequence data. Population genomic analyses offer unprecedented capabilities to investigate precise evolutionary, ecological, and epidemiological processes on both coarse and very fine scales. The proposed selective whole genome amplification technology allows the population genomic analyses necessary to address major outstanding questions about the microbiota in nature.

Acknowledgments

We are grateful to John Jaenike and James Fry for *Drosophila* samples and consultation and to Scott Poethig and Beatrice Hahn for useful and encouraging discussion. This work was supported in part by the National Institutes of Health (T32GM008216, AI076342, and AI097137) and the Burroughs Wellcome Fund.

Literature Cited

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang *et al.*, 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25: 3389–3402.

Beerli, P., 2004 Effect of unsampled populations on the estimation of population sizes and migration rates between sampled populations. Mol. Ecol. 13: 827–836.

Blainey, P. C., 2013 The future is now: single-cell genomics of bacteria and archaea. FEMS Microbiol. Rev. 37: 407–427.

Castro-Nallar, E., M. Perez-Losada, G. F. Burton, and K. A. Crandall, 2012 The evolution of HIV: inferences using phylogenetics. Mol. Phylogenet. Evol. 62: 777–792.

Dean, F. B., J. R. Nelson, T. L. Giesler, and R. S. Lasken, 2001 Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification. Genome Res. 11: 1095–1099.

Dean, F. B., S. Hosono, L. Fang, X. Wu, A. F. Faruqi *et al.*, 2002 Comprehensive human genome amplification using multiple displacement amplification. Proc. Natl. Acad. Sci. USA 99: 5261–5266.

Deng, X., H. Liu, Y. Shao, S. Rayner, and R. Yang, 2008 The epidemic origin and molecular properties of B′: a founder strain of the HIV-1 transmission in Asia. AIDS 22: 1851–1858.

Ellegaard, K. M., L. Klasson, K. Naslund, K. Bourtzis, and S. G. Andersson, 2013 Comparative genomics of Wolbachia and the bacterial species concept. PLoS Genet. 9: e1003381.

Fraser, C. M., S. Casjens, W. M. Huang, G. G. Sutton, R. A. Clayton *et al.*, 1997 Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. Nature 390: 580–586.

Fuller, C. W., L. R. Middendorf, S. A. Benner, G. M. Church, T. Harris *et al.*, 2009 The challenges of sequencing by synthesis. Nat. Biotechnol. 27: 1013–1023.

Gorski, L., 2012 Selective enrichment media bias the types of *Salmonella enterica* strains isolated from mixed strain cultures and complex enrichment broths. PLoS ONE 7: e34722.

Grenfell, B. T., O. G. Pybus, J. R. Gog, J. L. Wood, J. M. Daly *et al.*, 2004 Unifying the epidemiological and evolutionary dynamics of pathogens. Science 303: 327–332.

Hofinger, B. J., J. R. Russell, C. G. Bass, T. Baldwin, M. dos Reis *et al.*, 2011 An exceptionally high nucleotide and haplotype diversity and a signature of positive selection for the eIF4E resistance gene in barley are revealed by allele mining and phylogenetic analyses of natural populations. Mol. Ecol. 20: 3653–3668.

Holmes, E. C., and B. T. Grenfell, 2009 Discovering the phylodynamics of RNA viruses. PLOS Comput. Biol. 5: e1000505.

Hume, J. C., E. J. Lyons, and K. P. Day, 2003 Human migration, mosquitoes and the evolution of *Plasmodium falciparum*. Trends Parasitol. 19: 144–149.

Humphrey, P. T., D. A. Caporale, and D. Brisson, 2010 Uncoordinated phylogeography of *Borrelia burgdorferi* and its tick vector, *Ixodes scapularis*. Evolution 64: 2653–2663.

Lam, H. Y., M. J. Clark, R. Chen, R. Chen, G. Natsoulis *et al.*, 2012 Performance comparsion of whole-genome sequencing platforms. Nat. Biotechnol. 30: 78–82.

Langmead, B., and S. L. Salzberg, 2012 Fast gapped-read alignment with Bowtie 2. Nat. Methods 9: 357–359.

Lasken, R. S., and T. B. Stockwell, 2007 Mechanism of chimera formation during the Multiple Displacement Amplification reaction. BMC Biotechnol. 7: 19.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan *et al.*, 2009 The sequence alignment/map format and SAMtools. Bioinformatics 25: 2078–2079.

Mardis, E. R., 2008 Next-generation DNA sequencing methods. Annu. Rev. Genomics Hum. Genet. 9: 387–402.

Marygold, S. J., P. C. Leyland, R. L. Seal, J. L. Goodman, J. Thurmond *et al.*, 2013 FlyBase: improvements to the bibliography. Nucleic Acids Res. 41: D751–D757.

McLean, J. S., M. J. Lombardo, M. G. Ziegler, M. Novotny, J. Yee-Greenbaum *et al.*, 2013 Genome of the pathogen *Porphyromonas gingivalis* recovered from a biofilm in a hospital sink using a high-throughput single-cell genomics platform. Genome Res. 23: 867–877.

Pinard, R., A. de Winter, G. J. Sarkis, M. B. Gerstein, K. R. Tartaro *et al.*, 2006 Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. BMC Genomics 7: 216.

Quail, M. A., M. Smith, P. Coupland, T. D. Otto, S. R. Harris *et al.*, 2012 A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics 13: 341.

Quinlan, A. R., and I. M. Hall, 2010 BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841–842.

R Development Core Team, 2012 *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna.

Richardson, M. F., L. A. Weinert, J. J. Welch, R. S. Linheiro, M. M. Magwire *et al.*, 2012 Population genomics of the Wolbachia endosymbiont in *Drosophila melanogaster*. PLoS Genet. 8: e1003129.

Ritz, C., and A. Spiess, 2008 qpcR: an R package for sigmoidal model selection in quantitative real-time polymerase chain reaction analysis. Bioinformatics 24: 1549–1551.

Roberts, R. J., T. Vincze, J. Posfai, and D. Macelis,2010 REBASE—a database for DNA restriction and modification: enzymes, genes and genomes. Nucleic Acids Res. 38: D234–D236.

Rodrigue, S., R. R. Malmstrom, A. M. Berlin, B. W. Birren, M. R. Henn *et al.*, 2009 Whole genome amplification and de novo assembly of single bacterial cells. PLoS ONE 4: e6864.

Rosenberg, N. A., and M. Nordborg, 2002 Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. Nat. Rev. Genet. 3: 380–390.

Ross, M. G., C. Russ, M. Costello, A. Hollinger, N. J. Lennon *et al.*, 2013 Characterizing and measuring bias in sequence data. Genome Biol. 14: R51.

Schmieder, R., and R. Edwards, 2011 Quality control and preprocessing of metagenomic datasets. Bioinformatics 27: 863–864.

Simmons, S., G. Dibartolo, V. Denef, D. Goltsman, M. Thelen *et al.*, 2008 Population genomic analysis of strain variation in Leptospirillum group II bacteria involved in acid mine drainage formation. PLoS Biol. 6: e177.

Snyder, J. C., J. Spuhler, B. Wiedenheft, F. F. Roberto, T. Douglas *et al.*, 2004 Effects of culturing on the population structure of a hyperthermophilic virus. Microb. Ecol. 48: 561–566.

Sun, L. V., J. M. Foster, G. Tzertzinis, M. Ono, C. Bandi *et al.*, 2001 Determination of Wolbachia genome size by pulsed-field gel electrophoresis. J. Bacteriol. 183: 2219–2225.

Wakeley, J., 2004 Recent trends in population genetics: More data! More math! Simple models? J. Hered. 95: 397–405.

Wilson, D. J., 2012 Insights from genomics into bacterial pathogen populations. PLoS Pathog. 8: e1002874.

Wu, M., L. V. Sun, J. Vamathevan, M. Riegler, R. Deboy *et al.*, 2004 Phylogenomics of the reproductive parasite *Wolbachia pipientis* wMel: a streamlined genome overrun by mobile genetic elements. PLoS Biol. 2: E69.

Communicating editor: J. Miller

GENETICS

Supporting Information
http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.165498/-/DC1

Selective Whole Genome Amplification for Resequencing Target Microbial Species from Complex Natural Samples

Aaron R. Leichty and Dustin Brisson

DOI: 10.1534/genetics.114.165498

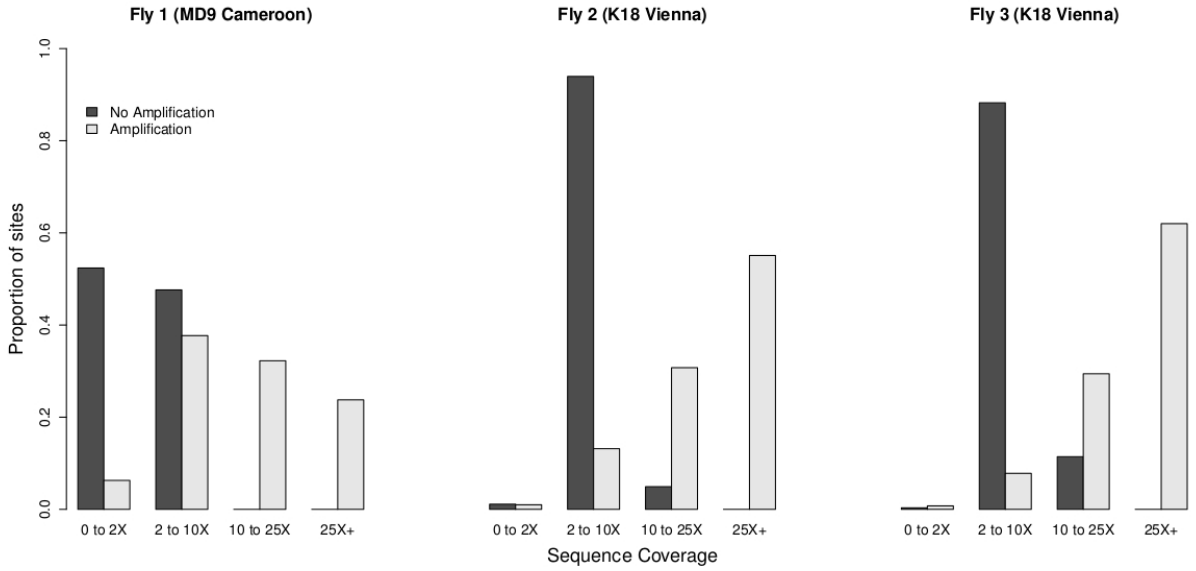

Figure S1 Sequence coverage improves across the *W. pipientis* genome due to selective whole genome amplification (SWGA). In all three *D. melanogaster*-*W. pipientis* samples tested, SWGA resulted in nearly 10 fold elevation in sequence coverage across the *W. pipientis* genomes. For example, while none of the sites in the Fly 1 genome had greater than 10X coverage when sequenced directly from the fly (dark bars), the majority of sites from the same fly extract had greater than 10X coverage after SWG amplification. The sequence coverage estimates for each sample were standardized to the equivalent of $1/96^{th}$ of an Illumina HiSeq lane (200 million 50 bp SE reads).

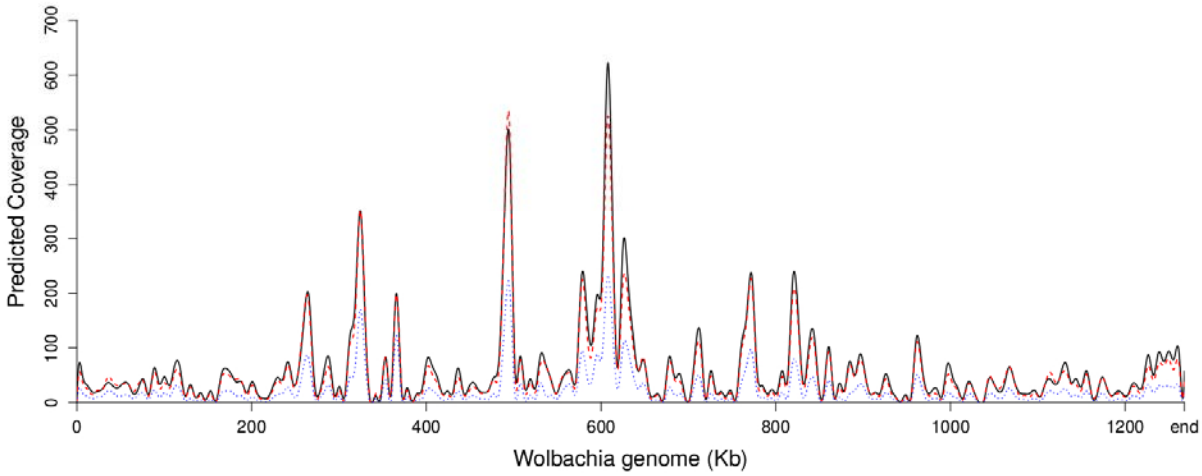

Figure S2 Genome-wide patterns of coverage are similar among samples. Curves represent sequencing coverage for post-SWGA samples from three flies across the *W. pipientis* genome. Curves were smoothed for visualization.

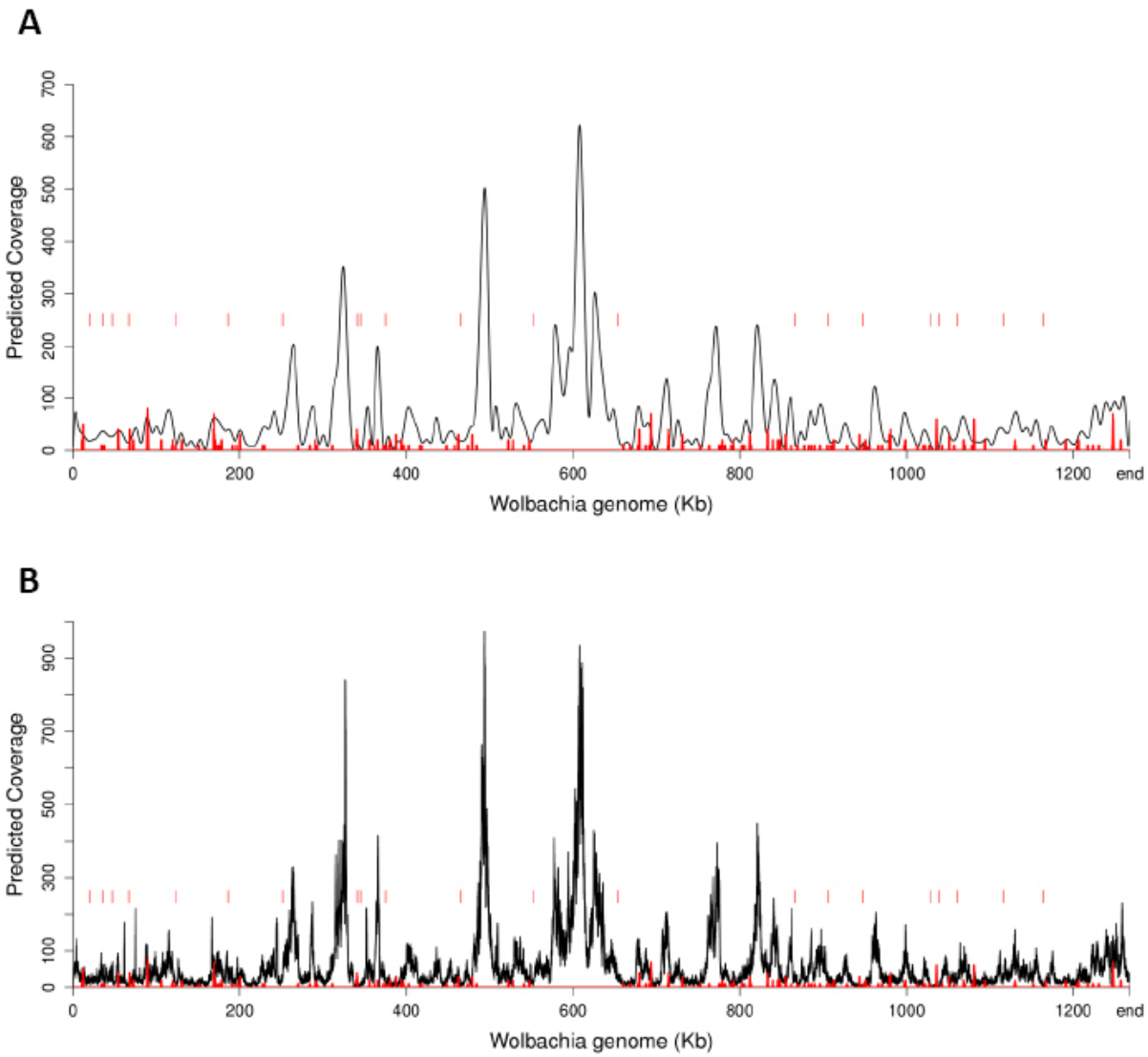

Figure S3 Primer density (red curve; number of motif occurrences in a 1,000 bp sliding window) is weakly correlated with sequencing coverage (black curve). Red vertical bars are locations of NarI restriction sites in the reference genome. Primer density was enlarged by a factor of 10 for visualization. Coverage curve was smoothed for visualization in panel A, raw coverage is shown in panel B. The origin of replication is at ~988 Kb.

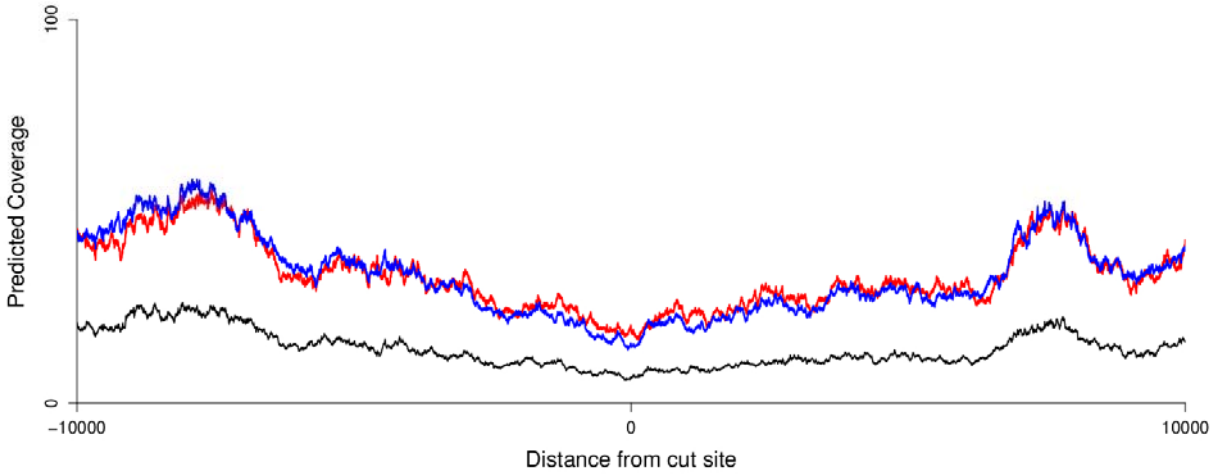

Figure S4 Sequencing coverage is positively correlated with distance from a restriction cut site. Each curve represents the average coverage across all NarI restriction sites in the *W. pipientis* genome. Curves are for individual flies (black = Fly 1, red = Fly 2, blue = Fly3).

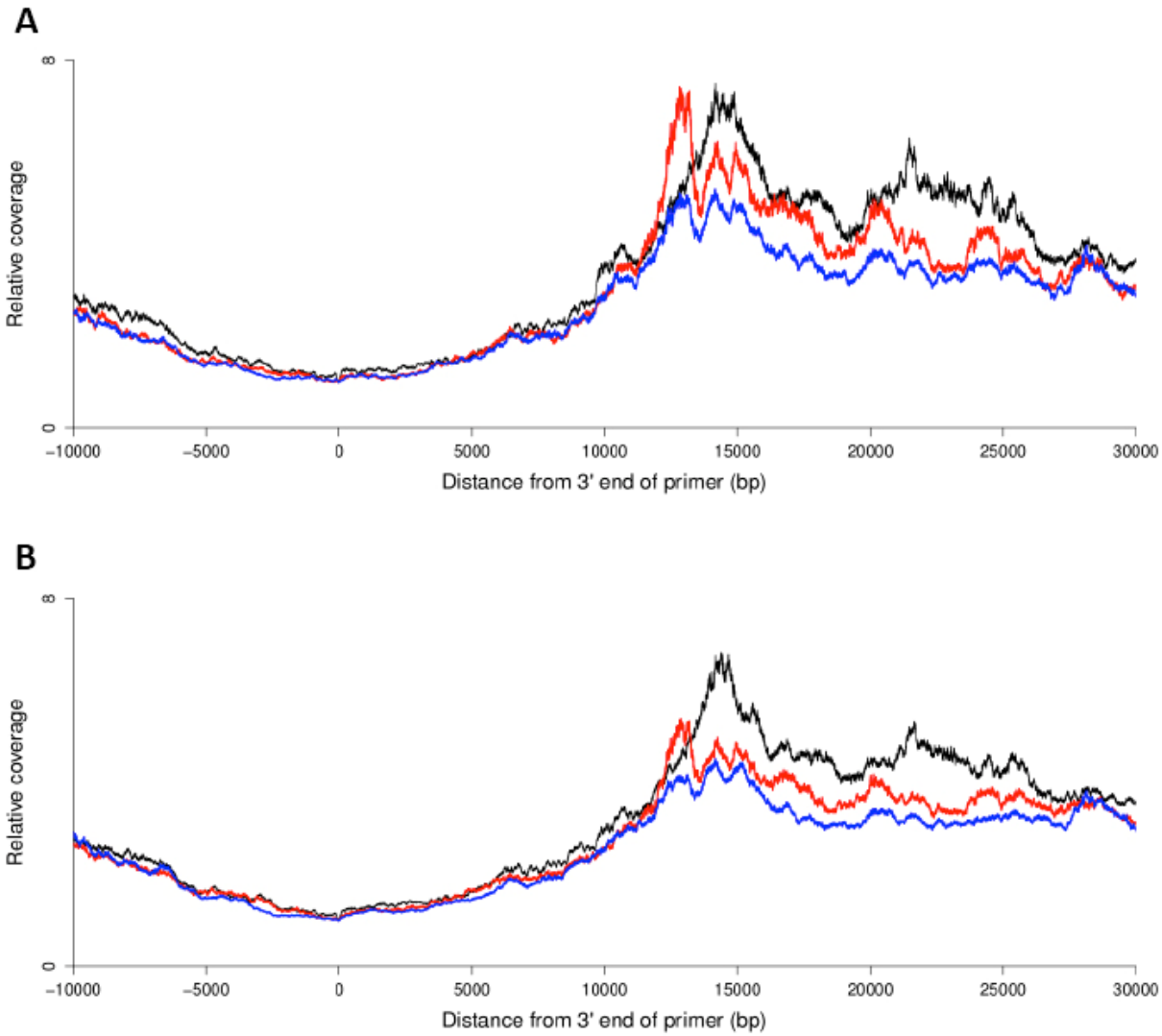

Figure S5 Sequencing coverage is positively correlated with distance from a predicted priming site. Panel A represents the average sequencing coverage on either side of each predicted priming site across the *W. pipientis* genome relative to the coverage at the 3' end of each primer. Panel B represents the average sequencing coverage on either side of each predicted priming site, including sites with 1 mismatching bp at any position excluding the 5 most 3' nucleotides of a primer, relative to the coverage at the 3' end of each primer. Each curves represents an individual fly (black = Fly 1, red = Fly 2, blue = Fly 3).

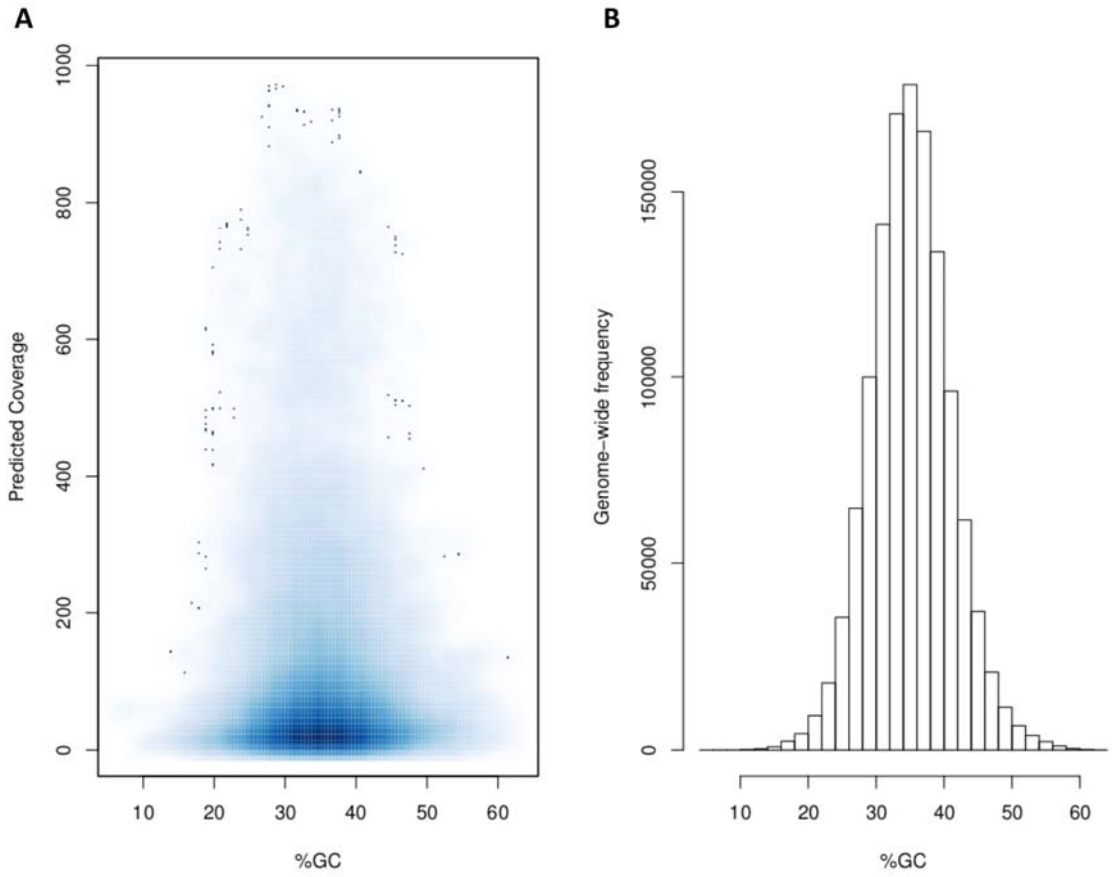

Figure S6 Sequencing coverage shows no correlation with GC content when controlling for GC frequencies. (A) The highest levels of coverage occur in the areas of the genome with the most frequent levels of GC content. Darker shading corresponds to a higher density of points. (B) Histogram of GC content across the *W. pipientis* genome. GC content was calculated for each position as the average across a 101 bp window.

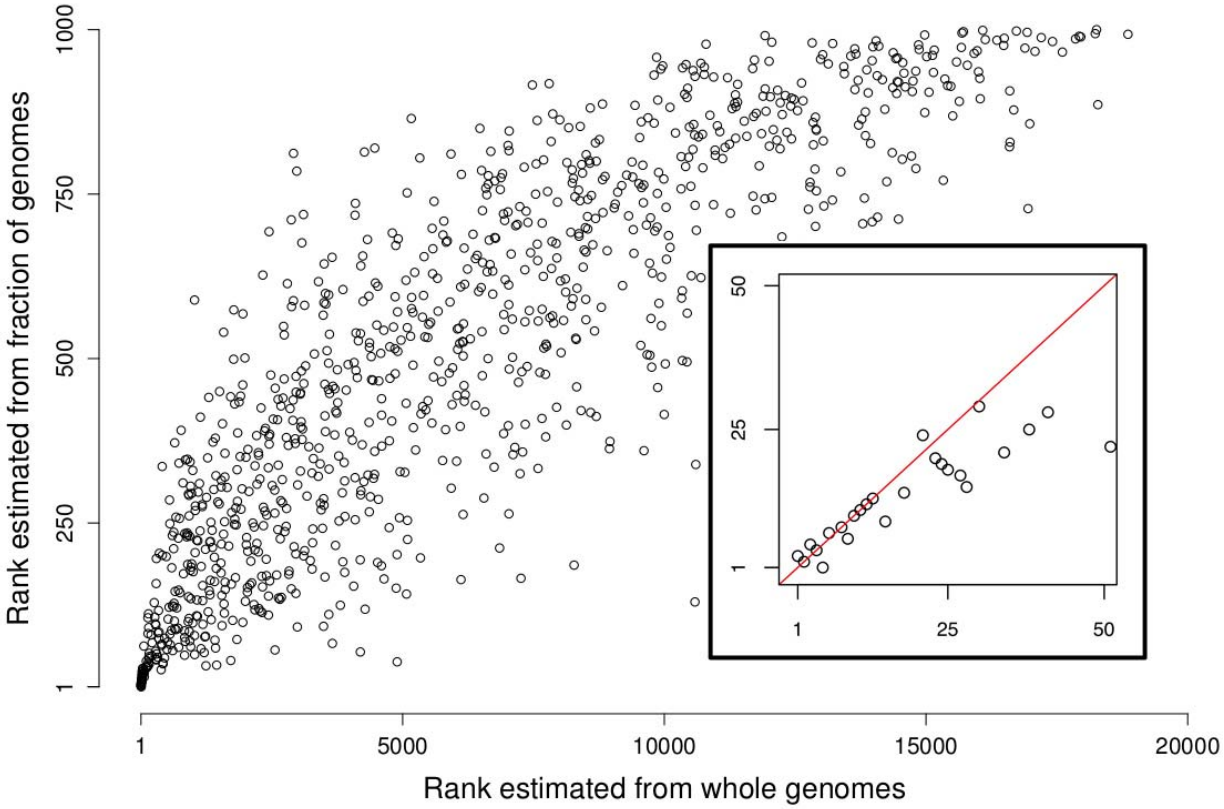

Figure S7 Primers selected for SWGA using a fraction of both the *W. pipientis* and *D. melanogaster* genomes are highly correlated ($R^2 = 0.72$) with those selected using the whole genomes of both species. 100 bp fragments were sampled with replacement to a depth of ~0.1X coverage for each genome (*D. melanogaster* chromosomes were sampled evenly), left unassembled, and used with SWGA.pl to estimate the ratio of frequencies for the top 1,000 most frequent motifs in the *W. pipientis* dataset. These ranks were compared with the ranks of the same motifs calculated from the entire assembled genomes of both species. The subwindow at right of figure shows the 1:1 correlation of motif ranks for the top 50 estimated motif ranks. Note that many motifs are equivalently ranked between datasets, but the 1:1 correlation lessens after rank 30.

File S1

PERL script digest_all.pl

File S1 is available for download as a .pl file at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.165498/-/DC1

File S2

PERL script SWGA.pl

File S2 is available for download as a .pl file at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.165498/-/DC1

Table S1 Primer sequences used for SWGA

Primer ID	Sequence	Primer Set	Tm (°C)
B31_1	ATTTTTTTATTT	B31-BL21	17.4
B31_2	TTTATTATTTTT	B31-BL21	16.0
B31_3	TTTTAAAATTTT	B31-BL21	17.9
B31_4	TTTAATATTTTT	B31-BL21	16.0
B31_5	TTTTTAAATTTT	B31-BL21	17.9
B31_6	TTTTTAATTTTT	B31-BL21	17.9
B31_7	TATTTTTTTATT	B31-BL21	16.0
B31_8	TTTTATTATTTT	B31-BL21	16.0
B31_9	TAATATTTTTTT	B31-BL21	16.0
B31_10	TTTTAATATTTT	B31-BL21	16.0
B31_11	TTTAAAATTTTT	B31-BL21	17.9
B31_12	AATATTTTTTTT	B31-BL21	17.4
B31_13	TTTTTAATTTTA	B31-BL21	16.5
B31_14	TTTTTTAAAATT	B31-BL21	17.9
B31_15	TTTTTTGATTTT	B31-BL21	22.3
B31_16	TTAATATTTTTT	B31-BL21	16.0
B31_17	ATTTTTTTTATT	B31-BL21	17.4
B31_18	TTTTTATTATTT	B31-BL21	16.0
B31_19	AAATTTTTTATT	B31-BL21	17.4
B31_20	TTTTTTTAAATT	B31-BL21	17.9
WMel_34	TCATACCGC	SR1	26.7
WMel_35	ATCCCGCTA	SR1	27.5
WMel_36	CCGCTAACA	SR1	27.2
WMel_38	CGTCATACC	SR1	23.5
WMel_41	TGTCATTCCA	SR1	26.2

WMel_43	CGGTATGAC	SR1	23.5
WMel_44	AACCGTCAT	SR1	24.1
WMel_46	CGATTCATTC	SR1	23.3
WMel_40	ATCCAAGTAG	SR1, SR2	22.1
WMel_45	CGGTATCTC	SR1, SR2	22.2

Table S2 Primer sequences used for qPCR

Primer ID	Sequence	Coordinates
	E. coli BL21	
bl21_qPCR_F1	GAACACGCTGAAAGCGGCTAACAT	468370
bl21_qPCR_R1	ATAGCAGCACGAGCGCCTTTAGTA	468498
bl21_qPCR_F2	GAGCGTATCAACAAAGCGCTGGAT	1344057
bl21_qPCR_R2	TCGCCCGTCTCGATATTGACGAAT	1344193
bl21_qPCR_F3	TCTCAATCGGATGGCAGAACGTGA	2448710
bl21_qPCR_R3	AGCGCCGTTATCAACCAAACGAAC	2448857
bl21_qPCR_F4	GCCAATCTCCGGTCGCTAATCTTT	3521473
bl21_qPCR_R4	GTTTGAACAGGGTTTCGCTCAGGT	3521327
bl21_qPCR_F5	CACGCCTGAGTTCGGCTAATTTGT	4314234
bl21_qPCR_R5	TGGCAGATGAAGATCTGAGCCGTT	4314099
	B. burgdorferi B31	
b31_qPCR_F1	AACCCAACCCAATATTTCCGCCAG	102184
b31_qPCR_R1	AAACCGATGCTGCAATCAACAGGG	102303
b31_qPCR_F2	AGTTTAGGGCCTCAGTGCGCTATT	298677
b31_qPCR_R2	TCCCGCTAAATCCTTCATAGGCCA	298545
b31_qPCR_F3	GCGGCACACTTAACACGTTAGCTT	445245
b31_qPCR_R3	AAGGCGAACTTCTGGGTCAAGACT	445393
b31_qPCR_F4	TGGCTTGCCTTAAACCGCTATCAC	592244
b31_qPCR_R4	TGATGCTATCAGGCAGTTGTGGGA	592120
b31_qPCR_F5	ACCTCTTGTGACGACTGTTGCGTA	811489
b31_qPCR_R5	AAGTCTTGAGGCAATCTCAGGCAC	811364
	D. melanogaster assembly 5	
Dmel_qPCR_F1	AGTGCGACTTCTCAGCCCAATACT	X: 11257135
Dmel_qPCR_R1	GCAGGTTGCCATCAAGATGCTGAA	X: 11257262
Dmel_qPCR_F2	TGTAACACTCCACGGCGATTTGAC	2L: 12399366
Dmel_qPCR_R2	TGATTGCTCACTACCCTGCTCACT	2L: 12399486
Dmel_qPCR_F3	TCGAGCAGTTGACGGTGTCATTCT	2R: 11144264
Dmel_qPCR_R3	AATCGTAGGAGGCCTGCATCTTCA	2R: 11144377
Dmel_qPCR_F4	TAATGGAGCTGATGCTGTGGGTGA	3L: 12415999
Dmel_qPCR_R4	CAACAGTCAACCGTGCAACACCAT	3L: 12416134
Dmel_qPCR_F5	GGCCCTTGCAATTGCTAACATCCA	3R: 14012047
Dmel_qPCR_R5	ATCTAACACGGGAATGACGTGGGT	3R: 14012157
Dmel_qPCR_F6	ACCTTAACCAACACAAGCGCATCC	4: 687216
Dmel_qPCR_R6	ACCAGTGTTGTAGGCACTTGAGGA	4: 687327
Dmel_qPCR_F7	TGCTCCTGATATAGCATTCCCACG	MT: 1731
Dmel_qPCR_R7	ACAGTTCATCCTGTTCCAGCTCCA	MT: 1850
	Wolbachia pipientis wMel	
wol_qPCR_F1	GTGCACAATGAATAACCGGAGGCA	193574
wol_qPCR_R1	GACATAGCATCGTCTGTTGTGCCA	193683
wol_qPCR_F2	CAAGCTGCTTCCTTAGGCTTTGCT	404642
wol_qPCR_R2	TCAAGAGATTGAGCGCAGGCTGAT	404533
wol_qPCR_F3	TTGCCCACATTGCTGCTGCTTTAG	597811
wol_qPCR_R3	TAAGGGCGTTGTGGGAAATAGGGT	597686
wol_qPCR_F4	TGCCACTGCTGTTTGAATCCTTCC	817294
wol_qPCR_R4	ACGCACGTTCGTACAACAAATGGG	817419

wol_qPCR_F5	CAGTCATTGCAACGCACCCAACTT	1009497
wol_qPCR_R5	AGGAAGCGGAGTATTGAGCGGATT	1009639

Table S3 qPCR results for each SWGA reaction (geometric mean ± standard deviation)

	qPCR for primer sets ($2^{-\Delta CT}$)			
Chromosome/locus	SR1, no digest	SR1, NarI	SR2, no digest	SR2, NarI
D. melanogaster X	203 ± 27	20 ± 5	23 ± 8	3 ± 0
D. melanogaster 2L	483 ± 356	36 ± 24	102 ± 38	6 ± 5
D. melanogaster 2R	403 ± 218	268 ± 90	68 ± 35	70 ± 6
D. melanogaster 3L	65 ± 15	9 ± 2	9 ± 2	6 ± 1
D. melanogaster 3R	19 ± 3	2 ± 0	6 ± 1	1 ± 0
D. melanogaster 4	433 ± 47	333 ± 40	59 ± 27	56 ± 33
D. melanogaster mtDNA	252 ± 142	247 ± 63	74 ± 22	69 ± 3
W. pipientis 1	2120 ± 1756	7608 ± 4510	72 ± 27	390 ± 195
W. pipientis 2	606 ± 376	2213 ± 1009	247 ± 34	1630 ± 577
W. pipientis 3	320 ± 269	1128 ± 594	287 ± 112	4010 ± 1782
W. pipientis 4	481 ± 220	1986 ± 129	270 ± 48	3413 ± 1054
W. pipientis 5	863 ± 406	3217 ± 416	100 ± 6	371 ± 114
mean *D. melanogaster*	176 ± 182	43 ± 145	33 ± 37	12 ± 33
mean *W. pipientis*	702 ± 722	2611 ± 2558	169 ± 101	1264 ± 1689

Table S4 Analysis of reads that do not map to *D. melanogaster* or *W. pipientis* using initial mapping parameters

Sample	Conditions	Chimera rate (% of total reads)	Top BLAST hit for non-mapping reads			
			Drosophila species	*Wolbachia* species	Other notable hits	No BLAST hit
Fly 1	Control	0.44	88.30	0.02	0.01 (Human)	11.14
Fly 1	NarI, SR2	0.90	90.49	1.62	0.18 (Human)	7.30
Fly 2	Control	0.28	81.13	0.12	1.02 (Acetobacter)	16.94
Fly 2	NarI, SR2	0.65	87.06	9.12	0.031 (Acetobacter)	3.39
Fly 3	Control	0.25	81.58	0.12	0.50 (Acetobacter)	17.11
Fly 3	NarI, SR2	0.65	82.55	13.78	0.025 (Human)	3.15