

Fine-Scale Recombination Mapping of High-Throughput Sequence Data

Catherine E. Welsh
Department of Computer Science
University of North Carolina
Chapel Hill, NC 27599
cwelsh@cs.unc.edu

Fernando Pardo-Manuel de Villena
Department of Genetics
Lineberger Comprehensive Cancer Center
University of North Carolina
Chapel Hill, NC 27599
fernando@med.unc.edu

Chen-Ping Fu
Department of Computer Science
University of North Carolina
Chapel Hill, NC 27599
ping@cs.unc.edu

Leonard McMillan^{*}
Department of Computer Science
University of North Carolina
Chapel Hill, NC 27599
mcmillan@cs.unc.edu

ABSTRACT

In this paper, we contrast the resolution and accuracy of determining recombination boundaries using genotyping arrays compared to high-throughput sequencing. In addition, we consider the impacts of sequence coverage and genetic diversity on localizing recombination boundaries. We developed a hidden Markov model for estimating recombination breakpoints based on variant observations seen in the read coverage spanning uniformly sized genomic windows. Our model includes 36 states representing all combinations of 8 genomes, and estimates a founder mosaic that is consistent with the variants observed in the aligned sequences. At HMM transition locations we consider the most likely founder-pair and refine the recombination breakpoints down to an interval spanning two informative variants. We compare this solution to alternate solutions based on microarrays that we have estimated. At 30x coverage the recombination mapping accuracy far exceeds the resolution attainable by any microarray. Even at coverages of 1x and below we are generally able to estimate recombination breakpoints with comparable accuracy.

Categories and Subject Descriptors

J.3 [Computer Applications]: LIFE and MEDICAL SCIENCES –Biology and genetics

Keywords

high-throughput sequencing, hidden Markov model, haplotype reconstruction

^{*}To whom correspondence should be addressed.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

BCB '13, September 22 - 25, 2013, Washington, DC, USA
Copyright 2013 ACM 978-1-4503-2434-2/13/09 ...\$15.00.

1. INTRODUCTION

High-Throughput Sequencing (HTS) of short reads is rapidly becoming cost competitive with full-genome genotyping using microarrays. A key difference between HTS and microarray genotyping is that microarrays sample specific genomic locations, whereas HTS samples the genome randomly. Categorizing genetic differences in HTS data requires a database of known sequence variants, while microarray-based genotyping is based on a set of reliable variants that were selected previously as part of the array's design. A common application of full-genome genotyping is to determine the ancestral origin of genomic segments arising from recombination. In this paper, we contrast the resolution and accuracy of determining recombination boundaries using genotyping microarrays with HTS. In addition, we consider the impacts of sequence coverage and genetic diversity on localizing recombination boundaries.

We have been monitoring the genomes of a multi-parental Recombinant Inbred Line (RIL) panel, called the Collaborative Cross (CC)[5] throughout its development. This is being done to ascertain the level of heterozygosity in various developing RILs as well as to decrease the number of generations of inbreeding required to achieve fully inbred animals[17]. We have monitored these genomes using three different genotyping arrays, two of which were designed specifically to be informative for the CC[20, 5]. For each of these genotyping platforms, algorithms have been designed to assign founders and estimate recombination breakpoints[10, 6]. Versions of these founder assignment algorithms have been demonstrated to work on a number of different mouse resources, including the Diversity Outcross (DO) [16] and other outbred populations.

Recently others have considered using HTS technologies to determine ancestral origins[13] and have also used sparse sequence data for this same analysis[15]. Sequencing data from four pooled samples were used to establish that the genetic variants and haplotypes of commercial outbred mouse stocks are largely shared with common laboratory strains[19]. We perform a similar analysis with eight-founder CC RILs, which leverages high-throughput sequencing data for three CC lines (OR867m532, OR1237m224, and OR3067m352)

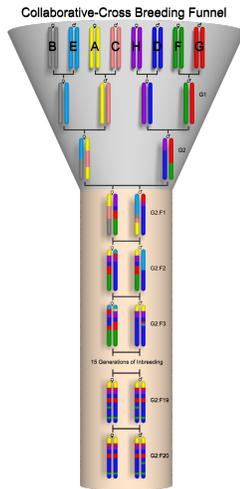


Figure 1: Collaborative Cross breeding scheme. Each independent CC strain begins with a funnel breeding stage that mixes eight founders, which are crossed for two generations, G1 and G2. The lines are then inbred for at least 20 generations to obtain recombinant inbred lines. CC lines are regularly genotyped after their 6th generation of inbreeding to monitor their residual heterozygosity, detect breeding errors, and to accelerate the inbreeding of selected lines.

that have also been previously genotyped on two of our genotyping platforms.

Beissinger et al.[2] addresses determining the necessary read coverage needed to genotype-by-sequencing in order to perform Quantitative Trait Loci (QTL) mapping. In our analysis we perform a similar determination of the read coverage necessary to map recombination breakpoints and compare this resolution to that obtained using genotyping arrays. We do this using the same 3 CC lines mentioned previously and sampling the reads to simulate various coverage levels.

2. MATERIALS AND METHODS

2.1 CC Strains

The Collaborative Cross is a multi-parental recombinant inbred panel derived from a set of eight genetically diverse inbred laboratory mouse strains. The set of founders consists of five classical inbred strains (A/J, C57BL/6J, 129S1/SvImJ, NOD/ShiLtJ, NZO/H1LtJ) and three wild-derived inbred strains (CAST/EiJ, PWK/PhJ, WSB/EiJ). They were chosen to capture a high level of genetic diversity, representing on average 90% of known genetic variation in laboratory stocks across all 1-Mb intervals[14]. A single CC strain is derived from the eight founders through a funnel breeding scheme that consists of two generations of mixing crosses, followed by 20 or more generations of inbreeding (Figure 1). Throughout the development of the CC, we have genotyped samples of strains at various stages of development using several different genotyping platforms. These genotyping platforms were used to track the remaining heterozygosity as well as the founder contributions at various generations.

Originally, preCC mice[1] were genotyped using the Mouse Diversity Array[20], which has approximately 500K SNPs at a cost of more than \$500 per sample. To aid the inbreeding process of the CC, we designed the Mouse Universal Genotyping Array (MUGA), a 7,854-marker array based on the Illumina Infinium platform [5], which costs \$100/sample. Finally, since the cost of genotyping arrays has continued to decrease, we designed a second generation genotyping array called MegaMUGA which contains 77K SNPs, including those on MUGA, and costs \$90/sample. We have genotyped over 2,800 CC samples using the three genotyping arrays, and for each of these arrays, we have inferred the genomic mosaic of the original eight founder genomes[10, 6].

2.2 Sequence Data

Whole-genome sequencing for three extant CC lines was completed by the Washington University School of Medicine Genome Sequencing and Analysis Center using Illumina sequencing technology with 30x haploid coverage. DNA was extracted from the spleen of a single male sample from each of the three extant CC strains. The resulting 100 base pair paired-end sequence fragments were aligned to a consensus reference genome using Bowtie (v 2.0.5)[9, 8]. The consensus genome was created by inserting the majority allele of the 8 CC founders at all known variant positions into the NCBI37/mm9 mouse genome[4]. The genetic variants were provided by the Wellcome Trust/Sanger Institute’s Mouse Genome’s Project[7]. We applied our techniques to these three extant lines since MUGA and MegaMUGA genotypes and sequencing data existed for all three samples.

3. APPROACH

3.1 Sequence Similarity Maps

In order to separate the degree to which resolving recombination boundaries depends on sequencing depth versus sequence similarity between the two sequences on either side of the recombination event, we developed a pairwise sequence similarity map. Sequence similarity varies throughout the genome and serves as a fundamental limit to our ability to resolve recombination boundaries. No amount of additional read coverage can improve the localization of a recombination breakpoint beyond the resolution determined by a sequence similarity map. In order to measure the accuracy of a given recombination boundary estimate, it is necessary to factor in the extent to which genomic variations exist near the region in question. A sequence similarity map provides such a gauge. It can also be used to normalize accuracy measures of recombination breakpoint positions.

3.2 HMM Algorithm

We use a Hidden Markov Model (HMM) algorithm to determine the founder mosaic for our sequenced animals. Since CC animals have eight founders and each loci can be heterozygous or homozygous, our HMM has 36 possible states (8 inbred and 28 founder-pair combinations). To help alleviate some of the noise inherent in sequencing data, we binned the genome into uniform sized genomic windows, so that each bin would contain sufficient evidence to discriminate between 36 possibilities using primarily biallelic variants. We then used a standard Viterbi algorithm to solve for the most likely founder mosaic represented in the HMM as described below.

3.2.1 Variants

A database of 65 million variants in 17 laboratory mouse strains has recently been produced by the Wellcome Trust/Sanger Institute[7]. They included the eight Collaborative Cross founder strains. Of these 65M SNPs, 31M high-confidence SNPs are informative among the eight CC founders. The majority allele at each of these 31M SNPs was used to construct the consensus genome used for alignment. We further filtered these down to a subset 29M SNPs such that there are no unknown genotypes among all eight founders, eliminating any need for imputation.

3.2.2 Emission Probabilities

The aligned reads were then examined at each of these 29M SNP positions and binned using uniform-sized non-overlapping bins. The bin size is a user specified parameter which should be set based on the amount of genetic diversity between the founders. Unless otherwise specified, it was set to 1000bp in this paper. Within each bin, emission probabilities are computed for each of the 28 heterozygous founder-pair combinations and the 8 inbred founders by counting the number of variants consistent with each of the 36 possible states. Counts for each of the 36 states were converted to a likelihood score based on the number of reads supporting each genotype call, and subsequently adjusted to compensate for the likelihood of the same counts occurring by chance as modeled by a binomial distribution. A noise model of 1 sequencing error per 100 sequenced bases was assumed, so that the binomial distribution of a homozygous call is 0.99, while the assumed split for a heterozygous call is 0.495. Three possible probabilities (homozygous for each allele and heterozygous) are calculated at each SNP based on the number of reads that supports each allele, and applied appropriately to each bin. The probabilities of all SNPs in a bin were combined, and then the values are normalized so that the sum of all probabilities in the 36 states sum to 1. When there are no SNPs or no reads present in a bin, the emission probabilities are assumed to be equal for all 36 states.

We also reweighted the informativeness of each bin based on the average number of reads and the total number of SNPs within each bin modeled as a Poisson distributed random variable, as shown in the formulas below, where R_{avg} is the average number of reads in all bins, R_{std} is the standard deviation of reads in all bins, and h_R is the number of reads in the current bin. Similarly, N_{avg} is the average, N_{std} is the standard deviation, and h_N is the current bin count of SNPs.

$$\alpha_1 = e^{-\frac{(h_R - R_{avg})^2}{R_{std}}} \quad (1)$$

$$\alpha_2 = e^{-\frac{(h_N - N_{avg})^2}{N_{std}}} \quad (2)$$

$$\alpha = \min(\alpha_1, \alpha_2) \quad (3)$$

$$P_{s'} = P_s * \alpha + \frac{1}{36} * (1 - \alpha) \quad (4)$$

This was done so that bins with a large number of reads (typical of highly repetitive regions of the genome) and bins with a small number of SNPs would not overly influence our solution. Parameters R_{avg} , R_{dev} , N_{ave} and N_{dev} are based on the reads and SNPs per bin for each given data set.

3.2.3 Transition Probabilities

The transition probabilities for the HMM are estimated based on observed recombinations seen in previous MUGA haplotype reconstructions for 350 unique, emerging CC lines[18]. There are four classes of transitions that can occur between states, as shown in Figure 2. The most likely class of transition is that the state remains the same between two adjacent bins. This is because over a genome of about 2,470 Mb, we observed an average of 100 recombinations among our 350 genotyped CC samples when founders were assigned using the intensity-based algorithm described by Fu et al.[6]. A similar number of recombinations were found using the Liu et. al[10] algorithm as reported by Fu et al.[6]. Another class of transitions occurs when a recombination on one chromosome generates a heterozygous state, or when a recombination on a single chromosome causes a transition from one heterozygous state to another. The homozygous to heterozygous transitions appear in two versions: either the homozygous founder is included in the heterozygous state (more likely) or the transition from a homozygous to a heterozygous state involves a simultaneous transitions on both chromosomes. The heterozygous to heterozygous states have two variants as well, such that either 1 or 0 of the founders remain the same between the two states. Based on the observed recombinations in the CC lines, we calculated the expected transition probabilities at a specified bin size. We assumed that 100 bins on average should contain a transition, and the rest should maintain the same state between consecutive bins. Therefore, the probability of remaining the same is (total bins - 100) / total bins. Of the 100 transitions, we observed that 41.85% of them are between a homozygous state and a heterozygous state that contained the homozygous state's founder, 37.14% were between two different homozygous states, 17.92% were between heterozygous states that share a founder, and the remaining 2.89% was between a homozygous state and a heterozygous state with no shared founder.

3.2.4 Viterbi Solution

We initially assume that all states were equally likely and set our priors to reflect that. The Viterbi algorithm then proceeds to find the path maximizing the sum of log-likelihoods, thus computing the most probable sequence of founder assignments. This process is repeated for each chromosome independently.

3.3 Refining Recombination Breakpoints

The HMM solution at best determines a recombination boundary to the resolution of a bin (typically 1Kb). In a post process, we utilize all informative SNPs between the most likely two founders identified on each side of the recombination by the HMM solution to refine the recombination breakpoints down to the distance between two consecutive informative SNPs. This becomes complicated in regions of high sequence similarity, leading to regions where the resolution of the recombination boundary depends on the pair of founders on each side of the event. However, in most cases, we are able to determine the two informative SNPs between which the recombination occurs and these SNP positions are then used to bound the recombination event. Where these informative SNPs are far apart, areas of uncertainty are drawn and we assume that the founders are Identical-By-State (IBS) within the determined interval.

All Possible HMM State Transitions

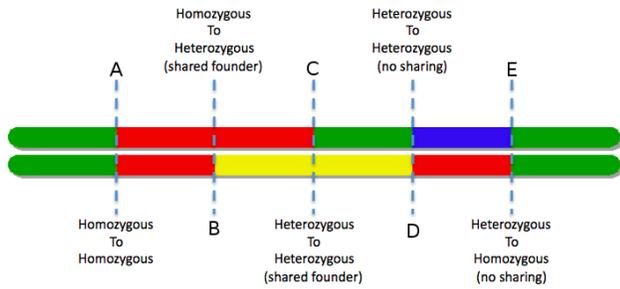


Figure 2: There are four classes of transitions that can occur between HMM states. The most likely transition is to remain the same founder state between two adjacent bins. In inbred animals, a shared recombination on both chromosomes generates the typical homozygous to homozygous transition (A). Another class of transitions occurs when a recombination on one chromosome transitions a homozygous state to heterozygous state (B), or causes a transition from one heterozygous state to another (C). Rare heterozygous to heterozygous transitions occur as a result of recombinations on both chromosomes (D) and are usually due to a recombination hotspot. Likewise, rare transitions from heterozygous to homozygous states can result from two aligned but separate recombinations (E).

3.4 Determining Necessary Read Coverage

The most significant variable influencing cost in HTS is the read coverage. In order to use HTS as a cost-effective alternative to genotyping arrays in the future, one needs to determine the necessary read coverage to compute haplotype reconstructions that are, at a minimum, equivalent in resolution to algorithms based on a fixed marker set. The resolution of array-based methods is a function of marker density, genetic state, and the informativeness of each marker. We designed MUGA to be able to resolve recombinations to within 1Mb on average when the sample was nearly inbred. MegaMUGA, was designed to resolve recombinations to within 160Kb for samples that are highly heterozygous. To determine the necessary read coverage, we sampled the reads at various coverage levels, such that if we wanted 2x coverage, we use about 1/15th of the available reads. In this way, we sampled the genome at 0.25x, 0.5x, 1x, 4x, and 16x. Since we randomly decided which reads to keep, each experiment was run 10 times with a different random seed and the resulting solutions are used in this analysis. For coverage levels of 1x and above, we used the same size bins (1Kb) as the full coverage solution. However, in order to maintain a similar level of evidence per bin at the lower coverage levels, we used 2Kb bins for the 0.5x coverage and 4Kb bins for the 0.25x coverage.

4. RESULTS

4.1 Founder-pair Resolution

Sequence similarity maps were constructed between all 28 founder pairs. In Figure 3 we show visualizations of these sequence similarity maps. These images depict the

number of 1000bp bins that have at least one informative SNP within each 100Kb bin. The sequence similarity map indicates where in the genome there are few or no informative SNPs distinguishing a particular founder-pair. The frequency of informative SNPs in a genomic region places a fundamental limit on the resolution for which a recombination breakpoint can be mapped using sequence information, regardless of read coverage. The density of informative SNPs varies significantly between founder pairs from the CC. The three wild-derived CC founder strains (CAST/EiJ, PWK/PhJ, and WSB/EiJ), include many variants and result in very few regions of ambiguity (areas with few or no informative SNPs) when they participate in a founder-pair. However, there is considerable sequence similarity among the five classical CC founder strains (A/J, C57BL/6J, 129S1/SvImJ, NOD/ShiLtJ, and NZO/H1LtJ) and therefore there are many regions of ambiguity in founder-pairs involving two classical CC strains. Figure 3 depicts three sequence-similarity maps, one between two classical strains, a second between a wild-derived strain and a classical strain and a third between two wild-derived strains. Also shown is the distribution of distances between informative SNPs genome-wide for all 28 founder-pairs, which peaks at 512 base pairs, thus justifying our choice of bin size (i.e. most 1000 base pair bins are likely to include a informative SNP variant between most founder pairs). The sequence similarity maps also depict regions of the genome where there are few annotated variants due to lack of sequence complexity, such as the large gaps on chromosomes 7, 12, and X. In these regions we would also expect to be limited in our capabilities to resolve recombination breakpoints. These sequence similarity maps are used to assess the possible localization accuracy of a specific recombination event as determined by experiments with variable read coverage.

4.2 Breakpoint Mapping of HTS data

Initially, we will consider the recombination breakpoint mapping accuracy attainable from the full 30x coverage sequence data. Accuracy depends both on sampling density and the genetic diversity between the founders surrounding each breakpoint. Our HMM solution pools evidence within regions of a user specified window size (1000 bases for 30X coverage) to infer the most likely source of the genome within a window. HMM transitions, which are suggestive of a recombination breakpoint, occur between window boundaries.

4.2.1 Comparison with Refined Breakpoint Solution

We refine our HMM estimates by expanding the region surrounding each transition, and then consider only the informative markers between the two founders identified on each side of the breakpoint. Generally, there is a clear transition where every marker distal to a boundary marker is consistent with one founder and every marker proximal to a second boundary marker is consistent with the other founder. For most recombinations, we were able to find two consecutive informative SNPs that were obviously on different sides of the recombination breakpoint. The actual breakpoint is most likely to have occurred between these two SNPs. We then tracked the distance between each of these informative SNPs and the HMM solution at each recombination for each of our three samples. Figure 4 depicts a histogram of these distances for each recombination.

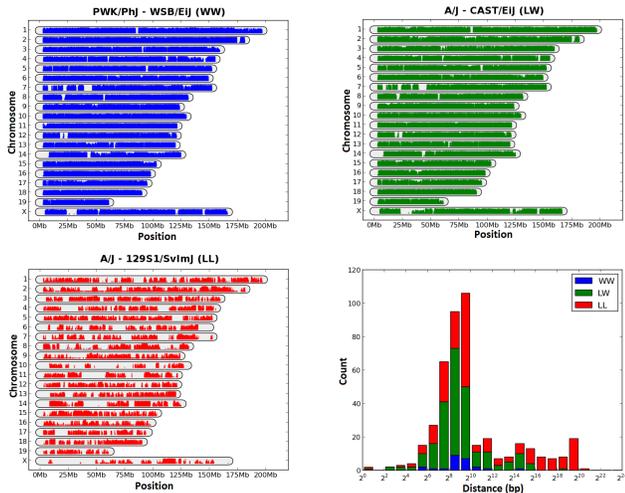


Figure 3: Sequence similarity maps for three founder-pairs and histogram of spacing between informative SNPs for all founder-pairs. The red, green, and blue subplots illustrate the percentage of 1000 base pair bins within a 100 kilobase window for which there is at least 1 informative SNP distinguishing the founder pair. Because CC founders fall into two categories, classical lab strains and wild-derived lab strains, we have three possible categories of founder-pair combinations. WW, shown in blue, occurs when both founders are wild-derived. These founder-pairs typically have low sequence similarity and many informative variants as seen by the relative density of the blue plot. LW, shown in green, occurs when one founder is wild-derived and the other is a classical lab strain. These founder-pairs also typically have many informative SNPs, but less than WW pairs. LL, shown in red, is an example where both founders are classical lab strains. These combinations typically have significantly more sequence similarity. As shown in the red plot there are many 1000 bp bins with no informative SNPs. Regions shown in white on these three sequence similarity maps indicate areas of the genome that will be difficult to detect recombinations between these founder-pairs. The stacked histogram plot shows the distance between all informative SNPs genome wide. It is divided into the 3 founder-pair categories to illustrate the larger distances between informative SNPs in red founder-pairs, as compared to the distances between informative SNPs for the blue and green founder-pairs.

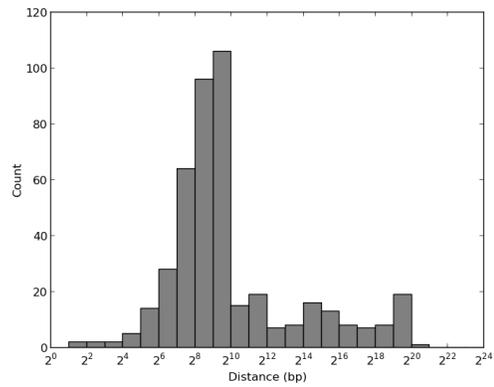


Figure 4: Histogram of distance between HMM solution and refined recombination breakpoint. The HMM solution can at best determine a recombination breakpoint to the resolution of a bin, which in this case is 1Kb. In a post process, we further refine these breakpoints by searching for informative SNPs within the region of the transitions and determining between which two consecutive SNPs the breakpoint actually occurs. We calculate the distance between each of these SNPs and the HMM solution and plot a histogram of the frequency at which each distance occurs. The high peak at 1Kb and the large number of distances < 1Kb indicate that the HMM solution is typically within the range of the informative SNPs.

Next we analyzed the mapping accuracy of the HMM solution relative to the refined informative-marker solution. For the three samples given, the HMM transition occurred at a median distance of 527 base pairs from the midpoint of the surrounding informative markers, with the first quartile falling 284 base pairs from the median, and the third quartile falling 899 base pairs from the median. A summary histogram of the distance of our HMM solution from the refined solution is shown in Figure 4. This histogram shows that the majority of our breakpoints were actually in the bin that the HMM transitioned, but there were some instances where there were no informative SNPs and the breakpoint estimation could not be narrowed down to within 2Mb of the HMM solution. In 61.8% of the recombinations, the HMM solution fell between the informative SNPs, while 18.6% transitioned before the informative SNP pair and 19.5% transitioned after. Transitions that occurred before the informative SNP pair tended to occur within a median distance of 546 base pairs, while transitions that occurred after, were a median distance of 233 base pairs.

We estimated the precision of the recombination-breakpoint localization using the gap spacing between the two informative markers of the refined solution. Over the 220 detected recombinations, we were able to localize each to a median region of 1,022 bases with the first quartile falling within 749 bases of the median, and the third quartile falling within 26,412bp of the median. The closest that a recombination breakpoint was determined was 5bp between strains NZO/H1LtJ and PWK/PhJ on Chromosome 1 around 29Mb in sample OR867m532. The poorest precision that we could assign an observed breakpoint was to 1,623,010 bases be-

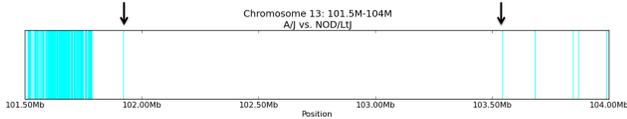


Figure 5: Depiction of all informative SNPs between A/J and NOD/ShiLtJ on Chromosome 13 from 101.5Mb to 104.0Mb. Informative SNPs are shown in cyan and arrows are used to depict the beginning and end of the ambiguous region between these two strains, since there are no informative markers between 101.9Mb and 103.5Mb.

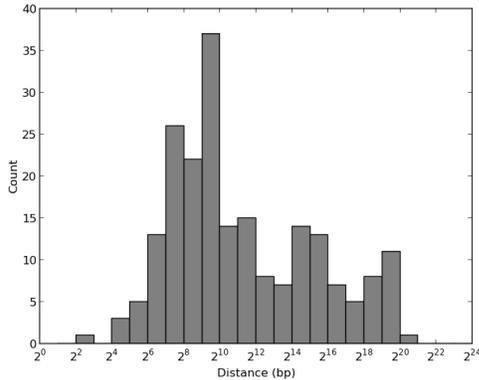


Figure 6: Histogram of distance between informative SNPs in the refined breakpoint solution. Starting from each HMM transition, we found the two consecutive SNPs informative for the different founders on each side of the breakpoint as determined by maximum likelihood founder-pair. The separation between these informative SNPs were used to compute the histogram shown. We consider the true recombination breakpoint to have occurred somewhere between these two SNPs. This indicates that the precision of recombination breakpoint mapping varies from a few bases to over a megabase in dense read coverage.

Sample	Number of Intervals			Concordance with HTS	
	HTSMUGA	MegaMUGA	MUGA	MegaMUGA	MUGA
OR867m532	117	108	118	95.56	98.09
OR1237m224	116	102	116	95.97	98.37
OR3067m352	112	102	113	96.76	98.95

Table 1: Comparison of HTS to Genotype Solutions, showing both the number of intervals found using each algorithm as well as the concordance between the HTS solution and the genotyping solutions. The concordance is measured such that at every base pair in the genome, we find the total number of base pairs where the genotyping solution is the same as the HTS solution divided by the total number of base pairs genome wide.

tween A/J and NOD/ShiLtJ on Chromosome 13 between 101.9Mb and 103.5Mb. This poor mapping is consistent with the sequence similarity map for A/J and NOD/ShiLtJ [21], as is shown in Figure 5, where all informative SNPs between A/J and NOD/ShiLtJ are shown in cyan and arrows depict the start and end of this ambiguous region. A summary histogram of distance between recombination breakpoint informative SNPs is shown in Figure 6.

Throughout the rest of our analysis we use the full-coverage HTS solution with refined breakpoints as the standard with which to evaluate alternative genotyping approaches and lower-coverage solutions.

4.2.2 Comparison to Genotype Solutions

Next we compared the recombination breakpoints determined from the whole-genome sequence data to the breakpoints estimated from the 7K MUGA and 77K MegaMUGA genotyping platforms. Given the relatively low sampling density of microarray based genotyping when compared to whole-genome sequencing, it is possible that some small genomic intervals (regions between two recombination breakpoints attributable to a single founder) can be missed entirely. The size of the minimum detectable genomic intervals was a design consideration for both MUGA and MegaMUGA. MUGA was designed to detect haploid founder intervals larger than 1Mb on average, whereas MegaMUGA was designed to detect both homozygous haploid or heterozygous diploid intervals larger than 160Kb on average. For our three samples, OR867m532, OR1237m224, and OR3067m352, MegaMUGA missed 1, 2, and 0 small genomic regions respectively. On OR867m532, MegaMUGA missed a 106Kb heterozygous region on Chromosome 8 from 19.68Kb - 19.79Kb, while on OR1237m224, it missed a 102Kb heterozygous region on Chromosome 8 from 19.68Kb-19.79Kb, and a 394Kb heterozygous region on Chromosome 11 from 97.50Kb-97.89Kb. On OR3067m352, there were no missing regions. The two missing heterozygous regions on Chromosome 8 of OR867m532 and OR1237m224 are in the same range, and examination of the sequence similarity maps shows that this region is adjacent to an area of very few informative SNPs for all founder-pairs (see Figure 3). MUGA solutions for our three samples tended to miss 7-11 intervals ranging in size from 102Kb - 3.1Mb.

A second aspect of recombination breakpoint accuracy is whether the two sequences on either side of the recombination breakpoint are consistent with the HTS predic-

tions. MegaMUGA chose a different founder-pair in 2-3 intervals per sample and also had about 1-2 false positives (extra recombinations) per solution. MUGA results had only 1 false positive total among our three samples and it occurred at the beginning of a chromosome. MUGA also only chose a different founder-pair in two instances total for our three samples. We attribute this to the relative maturity of the MUGA founder assignment solutions compared to MegaMUGA. MUGA has been available for almost three years and the HMM parameters have been tuned to match the marker performance. MegaMUGA on the other hand has been available for less than a year and fewer samples have been run on it. Moreover there is an order-of-magnitude more markers to characterize and tune parameters for. We expect MegaMUGA's founder assignment performance to improve over time both as the number of samples run on MegaMUGA increases and as we continue to refine the MegaMUGA founder assignment algorithm.

The final aspect of comparison is the breakpoint accuracy, which applies only to genomic intervals that are both detected and whose genomic intervals have founders consistent with the whole-genome sequence solution on both sides. On average, MegaMUGA localized the recombination breakpoint to within 161Kb-320Kb while MUGA's breakpoints were within 820Kb-870Kb. We would expect MUGA to be able to refine breakpoints to within 1Mb of the actual location and MegaMUGA to be within 160Kb on average. MUGA performed slightly better than anticipated on average, while MegaMUGA is not quite as good as expected at this point, but it is still 3 to 5 times more accurate than the MUGA platform it replaced. A comparison of the founder solutions for each of our three CC samples is shown in Figure 7 and Table 1. In Table 1, the concordance between the HTS solution and the genotype solution is measured such that for every base pair in the genome, we find the total number of base pairs where the genotyping solution is the same as the HTS solution divided by the total number of base pairs genome wide. Where one solution is found to be inbred and the other is heterozygous but includes the inbred solution, we consider this to be half right, and count it accordingly.

4.3 Read Coverage Analysis

We sampled the HTS reads at various coverage levels (16x, 4x, 1x, 0.5x, and 0.25x) to ascertain the level of accuracy of our haplotype reconstructions and the recombination breakpoints at each level. Since reads were chosen randomly, we repeated each coverage level 10 times. We compared each of the 10 solutions to our full coverage solution to determine the number of times recombinations were found, missed or when new recombinations not in the full coverage solution were created (false positives). For all true recombinations, we calculated the average distance from the recombination breakpoints of our low coverage solutions to the full coverage solution, and also noted the maximum distance between the full coverage recombination location and the low coverage solutions. A synopsis of these comparisons can be seen in Table 2. For comparison, similar statistics for the genotyping solutions are also shown in Table 2. In addition, Figure 8 shows the 3 full-coverage solutions compared to one of their 4x and 0.25x coverage solutions. At 4x coverage, most solutions were very similar to the 30x baseline and were between 99.8% and 99.9% concordant with the full coverage

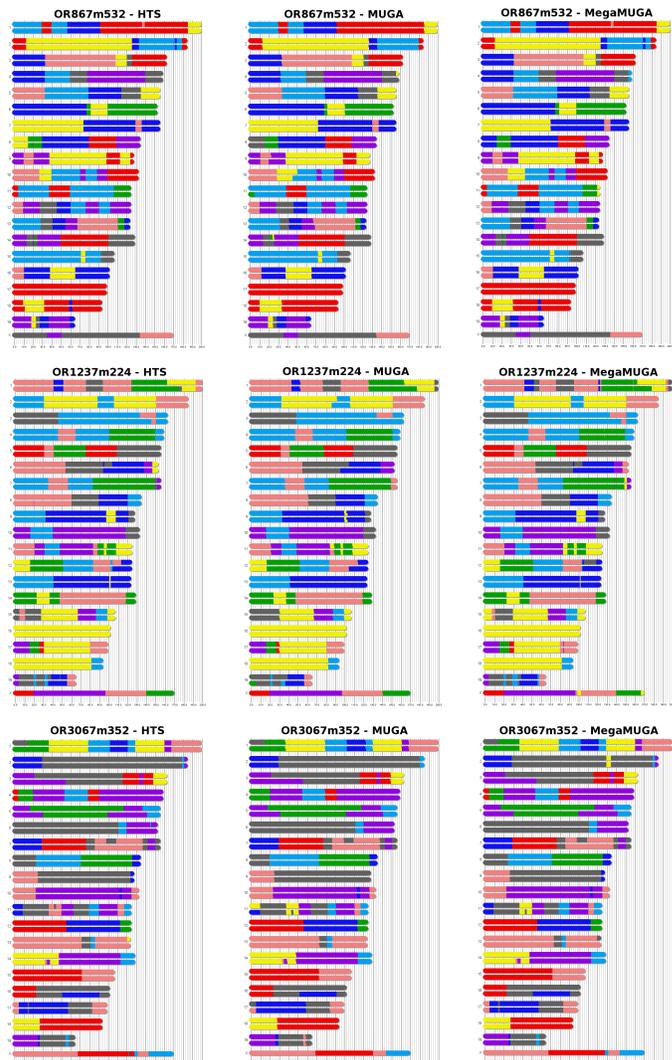


Figure 7: Comparison of HTS full coverage solutions with MUGA and MegaMUGA solutions. We show here the founder mosaic solutions for each of our three CC samples. The HTS solution is shown first, followed by the MUGA solutions and then MegaMUGA solution for each sample. There are a number of small differences between the MUGA and MegaMUGA solutions and the HTS solution. An example of one is that both MegaMUGA and MUGA mislabeled the pink founder (129S1/SvImJ) as yellow (A/J) on chromosome 11 of OR3067m352.

Sample	Avg. # Recombs	Avg # FP	Avg. # Missing Recombs	Avg. Distance to HTS	Max. Distance to HTS
OR867m532	95	-	-	-	-
16.0x	93	0	2	14471.51	824000
4.0x	93	0.2	2.2	19088.47	824000
1.0x	92.6	1.8	4.2	34485.99	869000
0.5x	79.6	9.6	25	46993.45	855000
0.25x	66	18	47	66466.53	998000
MUGA	86	1	8	820717.70	3832590
MegaMUGA	93	1	1	161699.68	1748837
OR1237m224	95	-	-	-	-
16.0x	92.4	0.4	3	7623.21	558000
4.0x	92.6	0.8	3.2	14482.41	558000
1.0x	91.3	3.1	6.8	35263.74	815000
0.5x	78.5	12.1	28.6	49397.94	982000
0.25x	77.3	10.3	28	51197.37	993000
MUGA	81	0	11	827496.10	3204258
MegaMUGA	114	2	2	252832.99	2263917
OR3067m352	90	-	-	-	-
16.0x	88	0.3	2.3	152.93	19000
4.0x	88.2	0.8	2.6	2765.18	753000
1.0x	87.6	4.1	6.5	26921.33	919000
0.5x	77.4	8.5	21.1	45009.11	933000
0.25x	76.6	9.9	23.3	48702.63	990000
MUGA	81	0	7	870420.44	3575568
MegaMUGA	92	1	0x	320968.16	3562834

Table 2: Statistics for various coverage levels of sequencing and genotyping data for our three CC samples.

solution. At 0.25x coverage though, the solutions varied more dramatically depending on whether or not the randomly selected reads fell over enough informative SNPs for a founder pair in a particular region. These solutions ranged from 49.1% to 99.4% concordant with the full coverage solution. In Figure 8 we have shown a 99.1% concordant 0.25x solution for sample OR867m532, a 95.6% concordant 0.25x solution for sample OR1237m224, and a 76.8% concordant 0.25x solution for OR3067m352. Note that the majority of the discordant solutions include a heterozygous state rather than the expected homozygous state chosen by the full coverage HMM. Calls of heterozygous states with only a single observation tend to be 50% correct, in that they always have one founder that matches the correct homozygous state solution. This issue could be addressed in low coverage cases by considering the degree of inbreeding when establishing the emission probabilities. We have shown that it is possible to accurately reconstruct founder mosaics using HTS data at relatively low coverage levels. In order to maintain the ability to distinguish between homozygous and heterozygous founder-pair states, we found that 1x coverage was sufficient. Below this level of coverage, our results were highly variable depending if we had reads at the informative SNPs within our recombination breakpoint areas. Solutions at 16x were very consistent, with the majority of the solutions choosing the exact same bins at which to transition for 87.2% of found recombinations. As we lowered our coverage level, the 10 solutions at each coverage level became more inconsistent, as shown in Figure 9, although they still maintained relatively concordant solutions with the full coverage HTS solution. As shown in Table 2, less recombinations are found and of the recombinations found, the distance from the HTS solution grows. The largest difference in coverage

levels comes between 1x and 0.5x, where we start to lose informative SNPs since only about half of the SNPs will have reads, and the bin size is also doubled in order to maintain a similar level of evidence.

5. DISCUSSION

By developing a method for computing founder mosaics from HTS data, we enable users of the CC, DO and other mouse populations to easily transition from genotyping arrays to HTS. This means that in the near future when HTS becomes price competitive with genotyping, our current pipelines for marker-assisted inbreeding[17], detection of residual heterozygosity and other tools for the CC lines[18] will be able to transition from using genotyping arrays to HTS. We have shown that even at relatively low coverage levels of 1x, our founder mosaics are just as reliable, if not more accurate than our current genotyping platform algorithms. This is caused by our ability to see almost all informative SNPs for each founder-pair genome-wide, rather than a pre-selected subset of SNPs. Designing informative genotyping arrays is also a very tedious process,[5] and with HTS, we can avoid this step entirely.

The accuracy with which one can resolve recombination breakpoints in HTS data depends on both the density of reads and the genetic diversity of the genomes on either side of the breakpoint. We have attempted to address both of these factors by combining an HMM data driven model with a refine process that is based entirely on the known genetic differences between a given founder pair. In this setting the HMM is responsible for finding a rough estimate of the breakpoint location, but more importantly it is responsible for identifying the founders on either side of the breakpoint. We can then refine the location of the breakpoint using in-

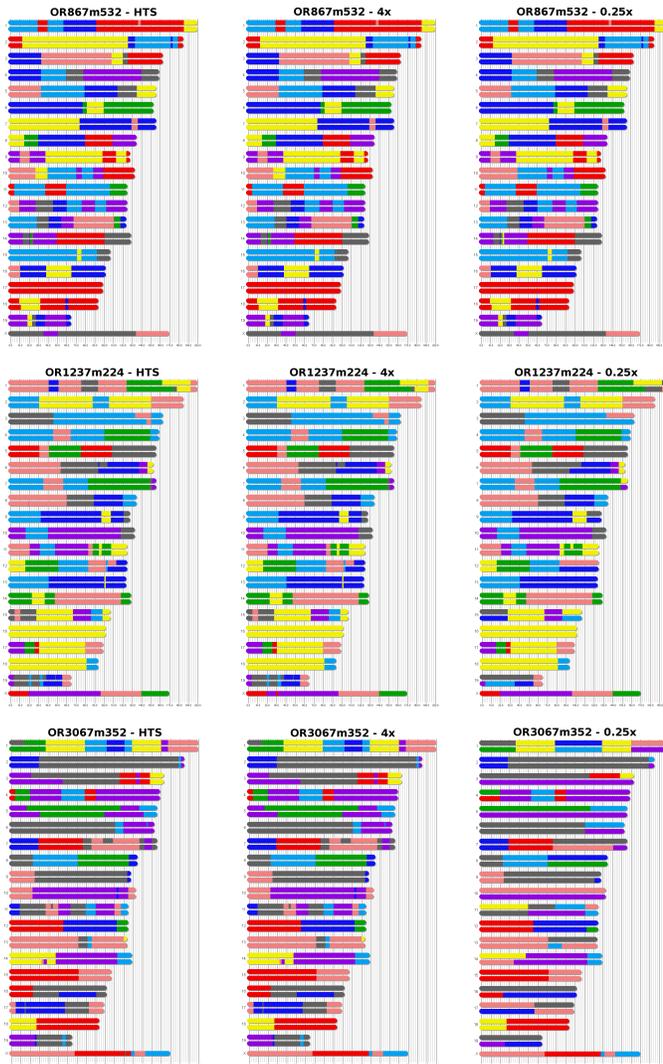


Figure 8: Comparison of HTS full coverage solutions with 4x and 0.25x coverage solutions. At 4x coverage, most solutions are very similar and were between 99.8% and 99.9% concordant with the full coverage solution. At 0.25x coverage, the solution were more variable, depending on which reads were selected. The 0.25x solutions shown here are 99.1%, 95.6%, and 76.8% concordant with the full coverage OR867m532, OR1237m224, and OR3067m352 solutions respectively. In the OR3067m352 0.25x solution shown, it can be seen that the majority of the discordance is from a heterozygous state being selected rather than the correct homozygous state. In each of these cases, the 0.25x solution is 50% correct, in that the heterozygous state selected includes the correct homozygous founder.

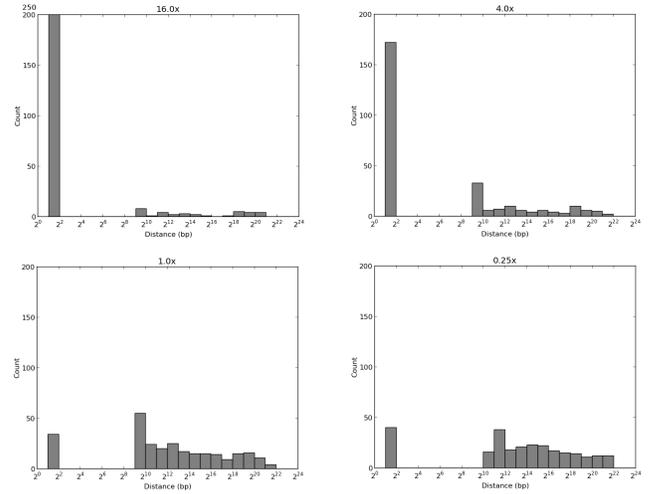


Figure 9: Histograms of the delta between the maximum position and minimum position found at each recombination among the 10 runs at each coverage level. Coverage levels of 16x, 4x, 1x, and 0.25x are shown. We can see that at 16x coverage, almost all 10 solutions were identical, while at 1x coverage, our solutions tended to be more divergent, although the majority still found transitions within 1-4 bins of each other.

formative SNPs down to the limits of the sequence diversity.

It has also been suggested that HTS may enable us to detect gene conversions that were previously undetectable. Gene conversions appear as two nearby recombinations, as if they were a tiny double recombination. Finding gene conversions is very difficult when pooling read data as they tend to be very small (100bp-3000bp) and the rate at which they occur is currently unknown. Using our full coverage HMM solutions, we plan to explore each bin for evidence of informative SNPs for some founder pair similar to how we refine recombination breakpoints. The primary difference being that when refining breakpoints the founder pair is given by the HMM solution. In the case of gene conversion all combinations would have to be explored while controlling for noise. We plan to take advantage of the observation that gene conversions tend to fall near recombinations, and in particular are found primarily in recombination hotspots. By looking at both the recombination regions in our HTS solutions as well as those regions of the genome known to be hotspots in mouse[11, 12, 3], we can test at what coverage level HTS allows for the discovery of gene conversions among multi-parental crosses.

6. ACKNOWLEDGMENTS

We would like to thank The Genome Institute at Washington University in St. Louis for generating the whole-genome sequence data used in this analysis. We would also like to thank Darla Miller for her help in obtaining the DNA and genotypes necessary for this analysis.

This project was partially supported by National Institutes of Health (NIH) grants MH-090338 and HG-006582 (F. P.-M. de Villena).

7. REFERENCES

- [1] D. L. Aylor, W. Valdar, W. Foulds-Mathes, and et. al. Genetic analysis of complex traits in the emerging Collaborative Cross. *Genome research*, 21(8):1213–22, Aug. 2011.
- [2] T. M. Beissinger, C. N. Hirsch, R. S. Sekhon, J. M. Foerster, J. M. Johnson, G. Muttoni, B. Vaillancourt, C. R. Buell, S. M. Kaeppler, and N. de Leon. Marker Density and Read-Depth for Genotyping Populations Using Genotyping-by-Sequencing. *Genetics*, 193(April):1073–1081, Feb. 2013.
- [3] H. Brunshwig, L. Levi, E. Ben-David, R. W. Williams, B. Yakir, and S. Shifman. Fine-Scale Maps of Recombination Rates and Hotspots in the Mouse Genome. *Genetics*, 191(3):757–764, July 2012.
- [4] D. M. Church, L. Goodstadt, L. W. Hillier, and et. al. Lineage-specific biology revealed by a finished genome assembly of the mouse. *PLoS biology*, 7(5):e1000112, May 2009.
- [5] C. C. Consortium. The Genome Architecture of the Collaborative Cross Mouse Genetic Reference Population. *Genetics*, 190(2):389–401, Feb. 2012.
- [6] C.-P. Fu, C. E. Welsh, F. P.-M. de Villena, and L. McMillan. Inferring ancestry in admixed populations using microarray probe intensities. In *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine*, BCB '12, pages 105–112, New York, NY, USA, 2012. ACM.
- [7] T. M. Keane, L. Goodstadt, P. Danecek, and et. al. Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature*, 477(7364):289–94, Sept. 2011.
- [8] B. Langmead and S. L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature methods*, 9(4):357–9, Apr. 2012.
- [9] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*, 10(3):R25, Jan. 2009.
- [10] E. Y. Liu, Q. Zhang, L. McMillan, F. P.-M. de Villena, and W. Wang. Efficient genome ancestry inference in complex pedigrees with inbreeding. *Bioinformatics*, 26(12):i199–i207, June 2010.
- [11] K. Paigen, J. P. Szatkiewicz, K. Sawyer, N. Leahy, E. D. Parvanov, S. H. S. Ng, J. H. Graber, K. W. Broman, and P. M. Petkov. The Recombinational Anatomy of a Mouse Chromosome. *PLoS Genet*, 4(7):e1000119, July 2008.
- [12] E. Parvanov, P. Petkov, and K. Paigen. Prdm9 controls activation of mammalian recombination hotspots. *Science*, 327(5967):10–13, 2010.
- [13] J. E. Pool, I. Hellmann, J. D. Jensen, and R. Nielsen. Population genetic inference from genomic sequence variation. *Genome research*, 20(3):291–300, Mar. 2010.
- [14] A. Roberts, F. Pardo-Manuel de Villena, W. Wang, L. McMillan, and D. W. Threadgill. The polymorphism architecture of mouse genetic resources elucidated using genome-wide resequencing data: implications for QTL discovery and systems genetics. *Mammalian genome : official journal of the International Mammalian Genome Society*, 18(6-7):473–81, July 2007.
- [15] T. B. Sackton, R. J. Kulathinal, C. M. Bergman, A. R. Quinlan, E. B. Dopman, M. Carneiro, G. T. Marth, D. L. Hartl, and A. G. Clark. Population genomic inferences from sparse high-throughput sequencing of two populations of *Drosophila melanogaster*. *Genome biology and evolution*, 1:449–65, Jan. 2009.
- [16] K. L. Svenson, D. M. Gatti, W. Valdar, C. E. Welsh, R. Cheng, E. J. Chesler, A. A. Palmer, L. McMillan, and G. A. Churchill. High-Resolution Genetic Mapping Using the Mouse Diversity Outbred Population. *Genetics*, 190(2):437–447, Feb. 2012.
- [17] C. E. Welsh and L. McMillan. Accelerating the inbreeding of multi-parental recombinant inbred lines generated by sibling matings. *G3: Genes| Genomes| Genetics*, 2(2):191–198, 2012.
- [18] C. E. Welsh, D. R. Miller, K. F. Manly, J. Wang, L. McMillan, G. Morahan, R. Mott, F. A. Iraqi, D. W. Threadgill, and F. P. M. de Villena. Status and access to the Collaborative Cross population. *Mammalian Genome*, pages 1–7, 2012.
- [19] B. Yalcin, J. Nicod, A. Bhomra, S. Davidson, J. Cleak, L. Farinelli, M. Ø sterås, A. Whitley, W. Yuan, X. Gan, M. Goodson, P. Klenerman, A. Satpathy, D. Mathis, C. Benoist, D. J. Adams, R. Mott, and J. Flint. Commercially available outbred mice for genome-wide association studies. *PLoS genetics*, 6(9), Sept. 2010.
- [20] H. Yang, Y. Ding, L. N. Hutchins, J. Szatkiewicz, T. a. Bell, B. J. Paigen, J. H. Graber, F. P.-M. de Villena, and G. a. Churchill. A customized and versatile high-density genotyping array for the mouse. *Nature methods*, 6(9):663–6, Sept. 2009.
- [21] H. Yang, J. R. Wang, J. P. Didion, R. J. Buus, T. a. Bell, C. E. Welsh, F. Bonhomme, A. H.-T. Yu, M. W. Nachman, J. Pialek, P. Tucker, P. Boursot, L. McMillan, G. a. Churchill, and F. P.-M. de Villena. Subspecific origin and haplotype diversity in the laboratory mouse. *Nature genetics*, 43(7):648–55, July 2011.