Title

Adapting Genotyping-by-Sequencing and Variant Calling for Heterogeneous Stock Rats

Journal

G3: Genes, Genomes, Genetics, 10(7)

ISSN

2160-1836

Authors

Gileta, Alexander F

Gao, Jianjun

Chitre, Apurva S

et al.

Publication Date

2020-07-01

DOI

10.1534/g3.120.401325

Peer reviewed

Adapting genotyping-by-sequencing and variant calling for heterogeneous stock rats

Alexander F. Gileta[*] [†⁋], Jianjun Gao[*⁋], Apurva S. Chitre[*], Hannah V. Bimschleger[*], Celine L. St. Pierre[*], Shyam Gopalakrishnan[#], Abraham A. Palmer[*‡]

[*] Department of Psychiatry, University of California San Diego, La Jolla, California, 92093

[†] Department of Human Genetics, University of Chicago, Chicago, Illinois, 60637

[#] Natural History Museum of Denmark, University of Copenhagen, 2200 København N, Denmark

[‡] Institute for Genomic Medicine, University of California San Diego, La Jolla, California, 92093

[⁋] These authors contributed equally to this work.

Running title: GBS and variant calling in HS rats

Key words: genotyping-by-sequencing, heterogeneous stock, rat, imputation

Corresponding author: Abraham A. Palmer

Mailing address: 9500 Gilman Drive #0667, La Jolla, CA, 92093

Phone number: 858-534-2093

Email: aapalmer@ucsd.edu

ABSTRACT

The heterogeneous stock (**HS**) is an outbred rat population derived from eight inbred rat strains. HS rats are ideally suited for genome wide association studies; however, only a few genotyping microarrays have ever been designed for rats and none of them are currently in production. To address the need for an efficient and cost effective method of genotyping HS rats, we have adapted genotype-by-sequencing (**GBS**) to obtain genotype information at large numbers of single nucleotide polymorphisms (**SNPs**). In this paper, we have outlined the laboratory and computational steps we took to optimize double digest genotype-by-sequencing (**ddGBS**) for use in rats. We also evaluate multiple existing computational tools and explain the workflow we have used to call and impute over 3.7 million SNPs. We also compared various rat genetic maps, which are necessary for imputation, including a recently developed map specific to the HS. Using our approach, we obtained concordance rates of 99% with data obtained using data from a genotyping array. The principles and computational pipeline that we describe could easily be adapted for use in other species for which reliable reference genome sets are available.

INTRODUCTION

Advances in next-generation sequencing technology over the past decade have enabled the discovery of high-density, genome-wide single nucleotide polymorphisms (**SNPs**) in model systems. Comprehensive assays of the

standing genetic variation in these organisms has allowed for the identification of quantitative trait loci (**QTL**) and the application of numerous population genetic and phylogenetic methods. However, due to the high degree of linkage disequilibrium (**LD**) in QTL mapping populations, sequencing whole genomes is not necessary. Many populations are the result of numerous generations of interbreeding inbred strains, allowing for recombination to produce an admixed population with known founder haplotypes. Due to the relatively slow rate of accumulation of recombination events, these populations contain large chunks of the genome derived from the same founder haplotype. Nearby SNPs are therefore often in strong LD with physically adjacent loci, effectively 'tagging' nearby variation and thereby reducing the number of sites that need to be directly genotyped. Several reduced-representation sequencing approaches that take advantage of LD structure have been previously described (Miller *et al.* 2007; van Orsouw *et al.* 2007; Van Tassell *et al.* 2008; Baird *et al.* 2008; Huang *et al.* 2009; Andolfatto *et al.* 2011; Elshire *et al.* 2011; Davey *et al.* 2011; Poland and Rife 2012; Peterson *et al.* 2012; Sun *et al.* 2013; Scheben *et al.* 2017). Using these methods, thousands of SNPs can be identified in large numbers of samples for a fraction of the price of whole-genome sequencing (Chen *et al.* 2013; He *et al.* 2014). The advantages of these methods are especially attractive when applied to less commonly utilized species or strains for which genotyping microarrays are not available.

Of the existing reduced-representation protocols, the genotyping-by-sequencing (**GBS**) approach developed by Elshire et al. (Elshire *et al.* 2011) has been frequently modified to accommodate other species: soybean (Sonah *et al.* 2013), rice (Furuta *et al.* 2017), oat (Fu and Yang 2017), chicken (Pértille *et al.* 2016; Wang *et al.* 2017), mouse (Parker *et al.* 2016), fox (Johnson *et al.* 2015), and cattle (De Donato *et al.* 2013), among others. The greatly varying genomic composition among organisms necessitates a diverse and customized set of approaches for obtaining high-quality genotypes. As such, both the GBS protocol and computational pipeline require modifications when applied to a new species. Recent work from our group showed that GBS can be effectively applied to outbred mice (Parker *et al.* 2016; Zhou *et al.* 2018; Gonzales *et al.* 2018) and rats (Fitzpatrick *et al.* 2013). However, those publications used protocols that had not been optimized, leaving significant room for improvement in genotype quality and marker density. Additionally, although several tools and workflows for the analysis of GBS data have been described, including Stacks (Catchen *et al.* 2013), IGST-GBS (Sonah *et al.* 2013), TASSEL-GBS (Glaubitz *et al.* 2014), Fast-GBS (Torkamaneh *et al.* 2017), and GB-eaSy (Wickland *et al.* 2017), the majority were developed and optimized for use in plant species. Given the lack of well-developed genomic resources in these species, they do not leverage the wealth of genomic data available for model organisms such as rats. Here we describe the customized computational and laboratory protocols for applying GBS to HS rats.

The HS is an outbred rat population created in 1984 using eight inbred strains and has been maintained since then with the goal of minimizing inbreeding and maximizing the genetic diversity of the colony (Johannesson *et al.* 2008; Woods and Mott 2017). After more than 80 generations of accumulated recombination events, their genome has become a fine-scale mosaic of the inbred founders' haplotypes. The breeding scheme and the number of accumulated generations has made the HS colony attractive for genetic studies. Additionally, extensive deep sequencing data exists for many inbred rat strains, including the eight progenitor strains (Rat Genome Sequencing and Mapping Consortium *et al.* 2013; Hermsen *et al.* 2015; Ramdas *et al.* 2019), allowing for accurate imputation to millions of additional SNPs following direct genotyping of only a subset.

Detailed here are the steps we have taken to optimize a rat GBS protocol and computational pipeline. Drawing on existing protocols (Elshire *et al.* 2011; Poland *et al.* 2012; Peterson *et al.* 2012; Parker *et al.* 2016) as templates, we redesigned our previous GBS approach (Parker *et al.* 2016; Gonzales *et al.* 2018) and have developed a novel, reference-based, high-throughput workflow to accurately and cost-effectively call and impute variants from low-coverage double digest GBS (**ddGBS**) data in HS rats. This publication is intended as a resource for others who might wish to perform GBS in rats and should provide a roadmap for adapting GBS for use in new species. We demonstrate that with a suitable reference panel, applying

reduced representation approaches and imputation in model systems can provide high-confidence genotypes on millions of genome-wide markers.

MATERIALS AND METHODS

Tissue samples and DNA extraction

Samples for this study originated from three sources: an in house advanced intercross line (**AIL**) derived from LG/J and SM/J mice (Gonzales *et al.* 2018), Sprague Dawley (**SD**) rats from Charles River Laboratories and Harlan Sprague Dawley, Inc. (Gileta *et al.* 2018), and an HS rat colony (Woods and Mott 2017; Chitre *et al.* 2018). Early stages of ddGBS optimization utilized AIL genomic DNA extracted from spleen by a standard salting-out protocol. Later optimization steps were performed using genomic DNA from SD rats extracted from tail tissue using the PureLink Genomic DNA Mini Kit (Thermo Fisher Scientific, Waltham, MA). HS rat DNA was extracted from spleen tissue using the Agencourt DNAdvance Kit (Beckman Coulter Life Sciences, Indianapolis, IN). All genomic DNA quality and purity was assessed by NanoDrop 8000 (Thermo Fisher Scientific, Waltham, MA). Interestingly, we observed that rat genomic DNA derived from either spleen or tail tissue appears to degrade faster than mouse genomic DNA following extraction by either of the above protocols; therefore, we recommend storing rat genomic DNA at -20° and using it within weeks of extraction whenever possible.

In silico digest of rat genome

We used *in silico* digests to aid in the selection of restriction enzymes, with the goal of maximizing the proportion of the genome captured at sufficient depth to make confident genotype calls (Kent *et al.* 2002). We used the *restrict* function in EMBOSS (version 6.6.0) (Rice *et al.* 2000) in conjunction with the REBASE database published by New England BioLabs (NEB; version 808) (Roberts and Macelis 1999) to perform *in silico* digest of the current release of the Norway brown rat reference genome, designated rn6. For the primary restriction enzyme, we chose PstI, which had been successfully used in numerous project (Fitzpatrick *et al.* 2013; Parker *et al.* 2016; Gonzales *et al.* 2018). We performed the digest with PstI alone and then with PstI paired with each of 7 secondary enzymes: AluI, BfaI, DpnI, HaeIII, MluCI, MspI, and NlaIII. We only considered fragments with one PstI cut site and one cut site from the secondary enzyme because the adapter and primer sets are designed to only allow these fragments to be amplified.

Restriction enzyme selection

Initial criteria for selecting a secondary restriction enzyme were a 4bp recognition sequence, no ambiguity in the recognition sequence (i.e. N's), compatibility with the NEB CutSmart Buffer, and an incubation temperature of 37ºC. The list of enzymes meeting these criteria at the time included AluI, BfaI, DpnI, HaeIII, MluCI, MspI, and NlaIII. Using the *in silico* digest data, we looked to maximize the portion of the genome contained within a fragment size range of 125-275bp (250-400bp with annealed adapters and primers) (Figure 2; Table 1). We excluded enzymes that produced blunt ends, both

because it would be more difficult to anneal adapters to blunt ended fragments and because our adapters would then also anneal to blunt ends produced by DNA shearing. We also excluded methylation-sensitive enzymes, as we did not want to limit the breadth of our sequencing efforts, accepting the possibility of read pileup in repetitive regions. Based on these criteria, as well as maximizing the percent of the genome captured, NlaIII, BfaI, and MluCI were selected for further testing. The final choice of enzyme (NlaIII) was determined empirically and is detailed in the Results.

ddGBS library preparation and sequencing

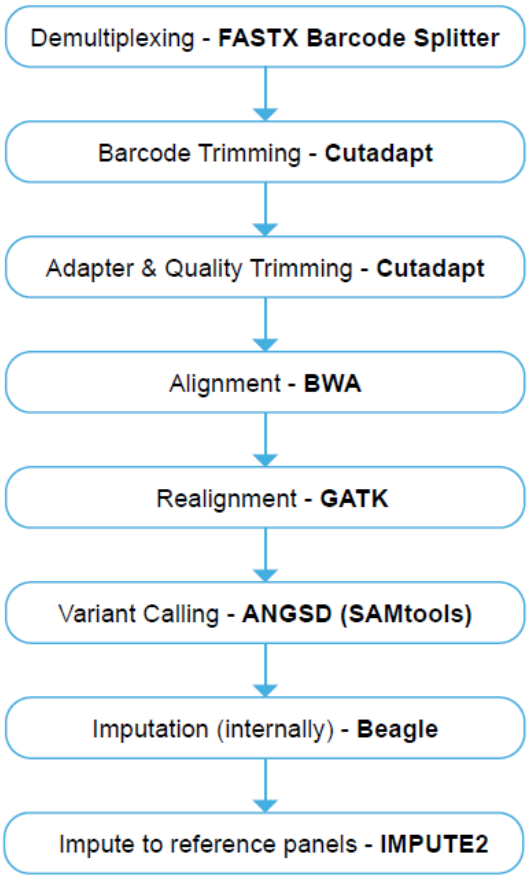
The full ddGBS protocol is available in File S1. In brief, approximately 1µg of DNA was used per sample. Sample DNA, PstI barcoded adapters, and NlaIII Y-adapter were combined in a 96-well plate and allowed to evaporate at 37°C overnight. The PstI adapter barcode is “in-line” such that each sequencing read from a given sample contains both the PstI overhang sequence (4bps) and a unique adapter sequence (4-8bps) prior to the beginning of the insert sequence. Sample DNA and adapters were re-eluted on day two with a PstI/NlaIII digestion mix and incubated at 37°C for two hours to allow for complete digestion. Ligation reagents were then added and incubated at 16°C for one hour to anneal the adapters to the DNA fragments, followed by a 30-minute incubation at 80°C to inactivate the restriction enzymes. Sample libraries were purified using a plate from a MinElute 96 UF PCR Purification

Kit (QIAGEN Inc., Hilden, Germany), vacuum manifold, and ddH_2O. Once re-eluted, libraries were quantified in duplicate with Quant-IT PicoGreen (Thermo Fisher Scientific, Waltham, MA) and pooled to the desired level of multiplexing (i.e. 12, 24, or 48 samples per library). Each pooled library was then concentrated by splitting the pooled volume across 2-3 wells of the MinElute vacuum plate and resuspending the library at desired volume for use in the Pippin Prep. The concentrated pool was quantified to ensure the gel cassette was not overloaded with DNA (>5µg). The pool was then loaded into the Pippin Prep for size selection (300-450bps) using a 2% agarose gel cassette on a Pippin Prep (Sage Science, Beverly, MA). Size-selected libraries were then PCR amplified for 12 cycles to add the Illumina sequencing primers and increase the quantity of DNA, concentrated again, and checked for quality on an Agilent 2100 Bioanalyzer with a DNA 1000 Series II chip (Agilent Technologies, Santa Clara, CA)., Bioanalyzer results were used to assure sufficient DNA concentration and to identify excessive primer dimer peaks.

As a pilot, an initial 96 HS samples were sequenced, 12 samples per library, at Beckman Coulter Genomics (now GENEWIZ) on an Illumina HiSeq 2500 with v4 chemistry and 125bp single-end reads. Subsequently, we began using a set of 48 unique barcoded adapters (File S2) to multiplex 48 HS samples per ddGBS library. Each library thereafter was run on a single flow cell lane on an Illumina HiSeq 4000 with 100bp single-end reads at the IGM Genomics Center (University of California San Diego, La Jolla, CA). We

also obtained ddGBS data on the HiSeq 4000 for a select set of 96 samples that had been previously genotyped on a custom Affymetrix Axiom MiRat 625k microarray (Part#: 550572), providing us with a “gold standard” array-based dataset with which to compare to our ddGBS data.

Figure 1. ddGBS sequencing data analysis workflow. Each step of the workflow is described in the text.



Evaluation of ddGBS pipeline performance

We present the steps required to call and impute genotypes from raw ddGBS sequencing data in Figure 1. During optimization of the pipeline, performance was assessed by two primary metrics: (1) the number of variants called and (2) genotype concordance rates for calls made in 96 HS

rats that had both ddGBS genotypes and array genotypes from a custom Affymetrix Axiom MiRat 625k microarray. There were two checkpoints in the GBS pipeline where genotype quality (measured by concordance rate) was assessed. The first was after “internal” imputation with Beagle (Browning and Browning 2009, 2016), whereby we leverage information from samples that had sufficient read depth to make a confident genotype call at a given locus in order to impute the genotype of other samples that had lower read depths at that locus. The second checkpoint was after “external” imputation, meaning imputation to our reference panel with IMPUTE2 to obtain genotype calls at loci we did not directly capture by our GBS method. (Howie *et al.* 2009, 2012). A third, additional metric we checked was the transition to transversion ratio ($T_S T_V$), which is expected to be ~2 for intergenic regions. The steps as outlined in the following sections reflect the final version of the pipeline. Variant calling and imputation steps utilized all available samples run on the HiSeq 4000 (3,000+ rats), though genotype concordance rates could only be calculated for the set of 96 HS samples for which we had array genotype calls.

Demultiplexing

The PstI adapter barcodes were used to demultiplex FASTQ files into individual sample files. Three demultiplexing software packages were tested: FASTX Barcode Splitter v0.0.13 [RRID: SCR_005534] (Hannon Lab 2010), GBSX v1.3 (Herten *et al.* 2015), and an in-house Python script (Parker *et al.* 2016). Reads that could not be matched with any barcode (maximum of 1

mismatch allowed), or that lacked the appropriate enzyme cut site, were discarded. Samples with less than two million reads after demultiplexing were discarded as these appeared to be outliers (Figure S4) and were observed to have high rates of missingness in their genotype calls. Data concerning demultiplexing are shown in Table S1 and are from a single HS rat sequenced in a 12-sample library on one lane after demultiplexing and adapter/quality trimming.

Barcode, adapter, and quality trimming

Read quality was assessed using FastQC v0.11.6 (Andrews 2017). We compared the efficacy of two rapid, lightweight software options for trimming barcodes, adapters, and low-quality bases from the NGS reads: Cutadapt v1.9.1 (Martin 2011) and the FASTX Clipper/Trimmer/Quality Trimmer tools v0.0.13 (Hannon Lab 2010) (Table S2). A base quality threshold of 20 was used and reads shorter than 25bp were discarded.

Read alignment and indel realignment

Rattus norvegicus genome assembly rn6 was used as the reference genome for read alignment with the Burrows-Wheeler Aligner v0.7.5a (BWA) [RRID: SCR_010910] (Li and Durbin 2009) using the *mem* algorithm. We then used GATK IndelRealginer v3.5 [RRID: SCR001876] (McKenna *et al.* 2010) to improve alignment quality by locally realigning reads around a reference set of known indels in 42 whole-genome sequenced inbred rat strains, including the eight HS progenitor strains (Hermsen *et al.* 2015).

Variant calling

Variants were called, and genotype likelihoods were computed at variant sites using ANGSD v0.911, under the SAMtools model for genotype likelihoods (ANGSD-SAMtools) (Korneliussen *et al.* 2014; Durvasula *et al.* 2016). Further, using ANGSD-SAMtools, we inferred the major and minor alleles (*-domajorminor* 1) from the genotype likelihoods, retaining only high confidence polymorphic sites (*-snp_pval* 1e-6), and estimated the allele frequencies based on the inferred alleles *(-domaf* 1). We discarded sites missing read data in more than 4% of samples (*–minInd*). Additionally, we tested multiple thresholds for minimum base (*-minQ*) and mapping (*-minMapQ*) qualities.

Internal imputation

Beagle v4.1 (Browning and Browning 2009, 2016) was used to improve the genotyping within the samples without the use of an external reference panel. Missing and low quality genotypes were imputed by borrowing information from other individuals in the dataset with high quality information at these same variant sites. Before settling on the combination of ANGSD-SAMtools and Beagle for genotype calling and internal imputation, we also experimented with GATK's HaplotypeCaller (McKenna *et al.* 2010) with various parameter settings, but with unsatisfactory results (Figure 3).

Quality Control for genotypes before imputation using an external reference panel

To verify the quality of the “internally” imputed genotypes prior to imputing SNPs from the 42 inbred strain reference panel (Hermsen *et al.* 2015), we checked concordance rates for the 96 HS animals with array genotypes, examined the $T_S T_V$ ratio, and assessed whether the sex as recorded in the pedigree records agreed with the sex empirically determined by the proportion of reads on the X-chromosome out of the total number of reads (Figure S1). We also identified Mendelian errors using the --mendel option in *plink* and known pedigree information for 1,136 trios from 214 families within the HS sample. Using the fraction of the trios that were informative for a given SNP and the formula $1-(1-2p(1-p))^3$, where p represents the minor allele frequency of the allele, we formed curves for the distributions of the expected number of Mendelian errors for both SNPs and samples and chose the inflection points as thresholds for the number of Mendelian errors allowed.

Data preparation for phasing with external reference panel

First, in our study sample of 96 samples, we only retained variants previously identified in the 8 HS founder strains because we expected the polymorphisms in our samples to be limited to the variation present in the founders (Hermsen *et al.* 2015; Ramdas *et al.* 2019). Further, to improve imputation efficiency, we employed a pre-phasing step with IMPUTE2 (*prephase_g*) (Howie *et al.* 2012) prior to imputation. Pre-phasing only needs to be performed once, allowing us to reuse the estimated haplotypes from

our dataset for imputation with multiple different reference panels. A flowchart outlining the pre-phasing protocol is presented in Figure S2.

Genetic maps

Genetic maps are required for phasing and imputation with IMPUTE2. When we began this project, no strain-specific recombination map was available for HS rats. Thus, we considered a sparse genetic map for SHRSPxBN (Steen *et al.* 1999). We also tested two types of linearly interpolated genetic maps, with recombination rates set at either 1cM/Mb or the chromosome specific averages for rats, as reported by Jensen-Seaman et al. (Jensen-Seaman 2004). Lastly, late in the course of this project, we experimented with an HS-specific genetic map developed by Littrell et al. (Littrell *et al.* 2018).

Imputation to reference panel

We used a combination of existing sequencing and array genotyping data from the HS rat founder and other inbred laboratory rat strains (Hermsen *et al.* 2015) as reference panel for imputation. Genotype data underwent QC and were phased by Beagle into single chromosome haplotype files. Haplotype files were then created using the workflow detailed in Figure S2. Imputation by IMPUTE2 was performed in 5Mb windows using the aforementioned reference panels and genetic maps.

Data availability

The ddGBS protocol and adapter sequences used to generate the data presented in this paper are available at https://doi.org/10.6084/m9.figshare.12284432.v1. All supplementary figures are available at https://doi.org/10.6084/m9.figshare.12280814.v1. Supplementary tables can be found at https://doi.org/10.6084/m9.figshare.12284444.v1. Genotype data will be available at https://dx.doi.org/10.6084/m9.figshare.8243222. The code necessary to run the steps of the computational pipeline outlined in this publication is available at https://dx.doi.org/10.6084/m9.figshare.8243156. Supplementary Files are available at https://dx.doi.org/10.6084/m9.figshare.8243129. Remaining files necessary for imputation (genetic maps, reference data, etc.) can be found with the following links: https://dx.doi.org/10.6084/m9.figshare.11919615, https://dx.doi.org/10.6084/m9.figshare.11919573, https://dx.doi.org/10.6084/m9.figshare.11919597.

RESULTS

ddGBS optimization

Previous projects utilizing GBS in mice and rats (Fitzpatrick *et al.* 2013; Parker *et al.* 2016; Gonzales *et al.* 2018) often encountered an issue where certain regions of the genome experienced high pileups of reads per sample (>100x), while other regions were covered by just 1-2 reads. This read distribution imbalance can be caused in part by PCR amplification bias,

where a subset of fragments are preferentially amplified until they dominate the final library (Kanagawa 2003; Aird *et al.* 2011). These previous protocols utilized 18 cycles of amplification. We tested reducing this to 8, 10, 12, or 14 cycles and found that below 12 cycles, there was insufficient PCR product to accurately quantify and pool for sequencing. The reduction in the number of PCR cycles was expected to reduce PCR bias, though this was not explicitly tested.

Another concern regarding previous sequencing results was an excess of long fragments (>700bps as determined by *in silico* digest). We observed that longer sequencing fragments often do not provide sufficient reads to make confident genotype calls (< 5 reads per sample), putatively due to inefficient bridge amplification and clustering on Illumina flow cells. Sequencing these long fragments is therefore wasteful. We tested three methods of combating this issue, including increasing the digestion time or enzyme concentration, performing size selection on the libraries, and using a two-enzyme restriction digest.

We considered the possibility that the restriction enzyme digests might not be running to completion. To address this possibility, we increased the duration of the digestion from 2 hours to 3 or 4 hours. We also tried increasing the number of units of PstI enzyme added, to ensure complete digest. Neither of these modifications impacted the final fragment length distribution of the library, indicating that the digest was reaching completion after 2 hours using the original concentration of PstI (File S3 – wells 1-6).

Our previous GBS protocol did not have an explicit library fragment size selection step. The final library was purified using a MinElute PCR Purification Kit (QIAGEN Inc., Hilden, Germany), which isolates PCR products 70bp-4kb in length, leaving a wide range of fragment sizes in the final library, under the assumption that only shorter fragments would bridge amplify on the flow cell. This method was imprecise and had low reproducibility, negatively impacting our ability obtain reads at consistent sites across libraries. Rather than attempt size selection by gel extraction, we chose to utilize a Pippin Prep, which automates the elution of DNA libraries of desired fragment size ranges. By using this automated size selection, we reduced the proportion of the genome targeted for sequencing, Additionally, since restriction enzymes make predominantly consistent cuts across samples (barring the presence of polymorphisms in RE recognition sites), it is ensured that highly similar sets of genomic fragments will be sequenced across sample libraries. Since the clustering process involves a bridge amplification step that preferentially amplifies library fragments with shorter insert sizes (Illumina, Inc. 2014), we kept the size selection window narrow (250-400bps) to avoid introducing a bias in which fragments were sequenced. A comparison of the fragment size distributions for the protocols before and after introduction of the Pippin Prep is shown in File S4.

To increase the proportion of the genome captured within the fragment size window, we pursued a double digest of the genome using a secondary enzyme with a more frequently occurring recognition sequence. When used

alone, *in silico* digest of the rn6 reference genome by PstI (Figure 2; Table 1) showed that only ~0.5% of the genome would have fallen within a 150bp fragment size window selected on the Pippin Prep. Previously, we performed GBS in CFW mice using the single-enzyme approach and observed that large regions of the genome that were not covered by sequencing reads (Parker *et al.* 2016). Therefore, we sought to increase the fraction of the genome that was accessible to GBS, so that there would be sufficient SNPs to tag the majority of the variation in the rat genome. Additionally, we were concerned about potential biases in coverage, heterozygosity, and the minor allele frequency (**MAF**) spectrum that may be introduced by a less complete capture of the genome. Flanagan and Jones have performed an empirical study comparing single- to double- digest RAD-seq and found that double-digest RAD-seq had lower rates of allelic dropout, decreased variance in between-sample per SNP coverage, less allele frequency inflation due to PCR bias, and reduced batch effects (Flanagan and Jones 2018).

The number of fragments with one of each of the cut sites was summed for all observed lengths and the results summarized in Figure 2 and Table 1. BfaI, MluCI, and NlaIII were chosen for further testing due to their compatibility with PstI digestion reagents and temperatures, sticky ends, and the proportion of the genome falling in the size selection window in the in silico analysis. We ruled out BfaI because it only had a 2bp overhang after cleavage, which we empirically showed leads to a high concentration of adapter dimer in the sequencing libraries (S5 File). NlaIII was chosen over

MluCI because it contained the greatest portion of the genome in the desired size selection window.

Table 1. Restriction enzyme options for double digest.

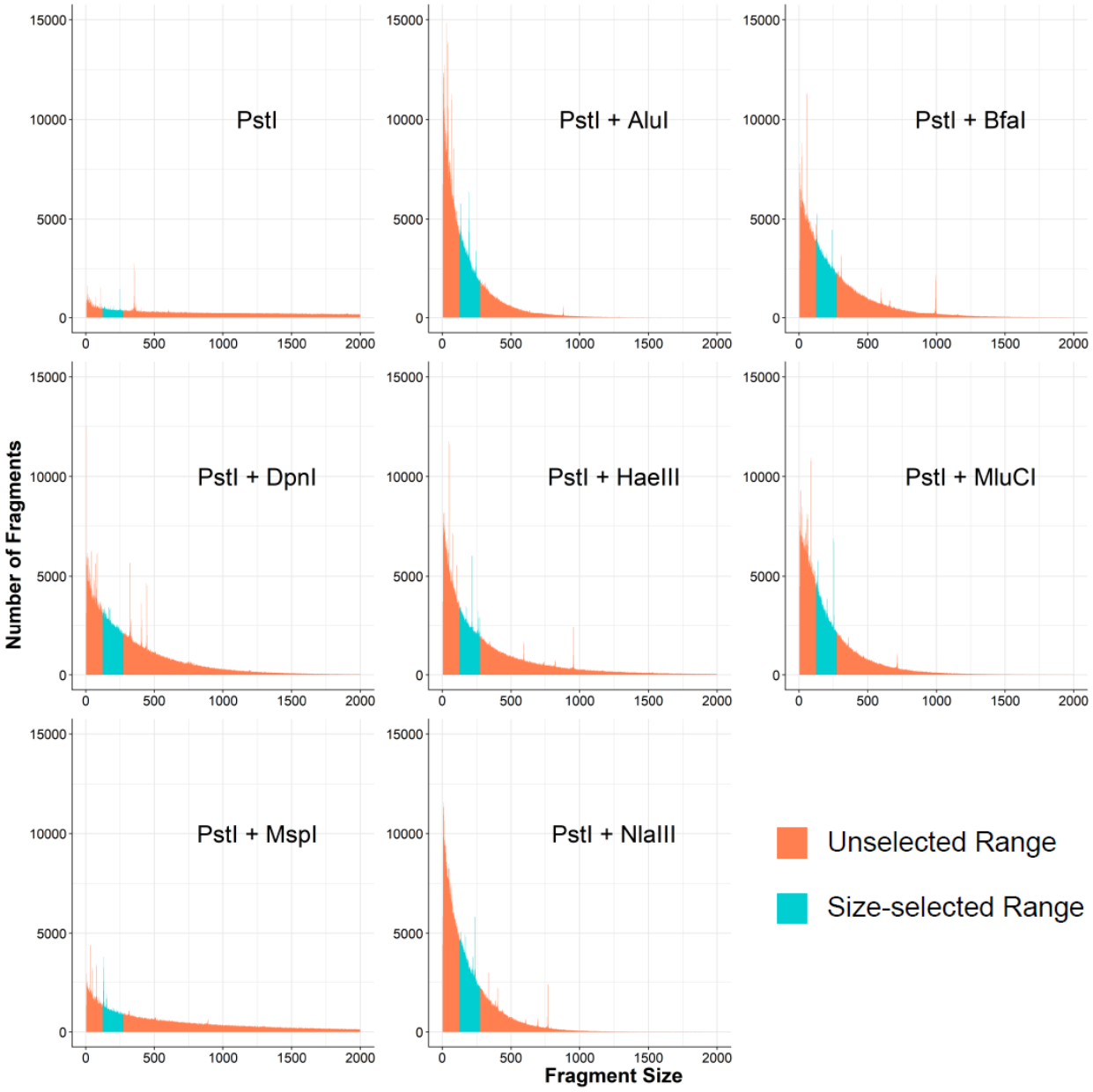
Restriction Enzyme(s)	Recognition sequence	Length of Overhang (bp)	% Genome in 250-400bp Region[+]	% Genome in 300-450bp Region[+]
PstI	CTGCA^G	4	0.48%	0.56%
PstI + AluI	AG^CT	0	3.06%	2.88%
PstI + BfaI	C^TAG	2	3.10%	3.25%
PstI + DpnI*	GA^TC	0	2.69%	3.00%
PstI + HaeIII	GG^CC	0	2.71%	2.79%
PstI + MluCI	^AATT	4	3.32%	3.21%
PstI + MspI	C^CGG	2	1.16%	1.24%
PstI + NlaIII	CATG^	4	3.45%	3.31%

The percent genome in region columns indicate the percentage of the genome that falls within the provided fragment size ranges and can therefore be captured by GBS.

* Restriction enzyme is methylation sensitive.

[+] Calculated using rn6 genome length of 2,870,182,909bps.

Figure 2. *In silico* digest fragment distributions for PstI and potential secondary restriction enzymes.



Each panel represents an independent digest of rn6 with the listed enzyme(s). Regions highlighted in blue are fragments that would be selected by the Pippin Prep (125-275bp) after annealing adapters and primers. These regions are quantified in Table 1 by multiplying the length of the fragments by the number of fragments to estimate the portion of the genome captured.

In our previous GBS protocol, all fragments were cut on both ends by PstI. By using a substantially lower concentration of the barcoded PstI

adapter than the common PstI adapter, we ensured the barcoded adapter would be the limiting reagent and the majority of fragments with an annealed barcoded adapter would have a common adapter on the other end. This is crucial, as having one of each of the adapters is required for proper amplification of the fragments on the flow cell. However, when using both PstI and NlaIII, the library is predominantly composed of fragments cut on both sides by NlaIII (File S6), which will amplify during PCR with a common adapter, but not on the flow cell. Therefore, we employed a Y-adapter (Poland *et al.* 2012) to control the direction of the first round of PCR and prevent two-sided NlaIII fragments from dominating the final sequencing library (File S2).

We tested numerous quantities of PstI and NlaIII adapters in an attempt minimize the amount used and avoid adapter dimers in the final libraries. For the barcoded PstI adapters, we tested 120pmol, 60pmol, 20pmol, 4.0pmol, 2.67pmol, 1.60pmol, 0.53pmol, and 0.20pmol; for the NlaIII Y-adapter, 30pmol, 10pmol, 5.0pmol, 4.0pmol, and 1.0pmol (Files S7 & S8). We found that 0.20pmol of PstI adapter and 4pmol of NlaIII Y-adapter yielded sufficient library and minimized the presence of adapter dimers.

We sequenced a trial flow cell with 8 pooled ddGBS libraries of 12 SD rat samples each (96 total) on a HiSeq 2500 (Illumina, San Diego, CA) with 125bp reads and v3 chemistry, obtaining an average of 15.3 million reads per sample. Given the NlaIII *in silico* digest results suggested we were capturing ~3.4% of the genome and that we were using 125bp reads, this

corresponded to approximately 20x coverage of captured sites. We subsequently increased the number of samples to 48 per library for the HS rats because we hypothesized 5x would be sufficient coverage per sample when utilizing imputation to a reference panel. We also discovered that a portion of the reads contained sequence fragments of the NlaIII adapter sequence, indicating there were fragments with insert sizes smaller than 125bps in the final library. To avoid this, we increased the fragment size range to 300-450bps (Table 1), which corresponds to a 175-325bp insert size once the adapters and primers are accounted for. We noted however that the library size distribution obtained from the Pippin Prep was uniformly shifted towards higher fragment lengths (Figure S3). This is a result of the high concentrations of our libraries after pooling and loading the gel cartridge near the upper limit of the recommend number of micrograms of DNA, which can cause slower migration of the DNA across the gel matrix.

The final ddGBS protocol can be found in File S1 and the necessary primer and adapter sequences in File S2. This protocol was used for the sequencing of all HS rats included in subsequent computational optimization steps.

Demultiplexing

The number of base pairs of sequencing data retained after demultiplexing was fairly consistent across demultiplexing software (Table S1). We ultimately decided to use FASTX Barcode Splitter because it yielded the

greatest number of reads after quality/adapter trimming and had faster run times. An average of 330 million 100bp reads were obtained per library, resulting in ~7 million reads per sample. Figure S4 shows the distribution of reads counts for all samples after demultiplexing.

Adapter and quality trimming

Read quality was substantially improved after trimming the barcode and adapter sequences and low-quality base pairs at the ends of reads (Figure S5). Overall read counts were only marginally reduced by quality trimming (Table S1). We observed that the number of called variant sites and the genotyping rate were both greater when using reads initially processed by cutadapt (Martin 2011) than reads processed by the FASTX_Toolkit (Table S2). Importantly, a large portion of the additional identified variants were known variant sites from the 42 inbred strains reference set (Figure S6), indicating the elevated call rate was at least in part due to capturing more true variant sites. We viewed this as sufficient support for proceeding with cutadapt for adapter removal and quality trimming.

Mapping quality

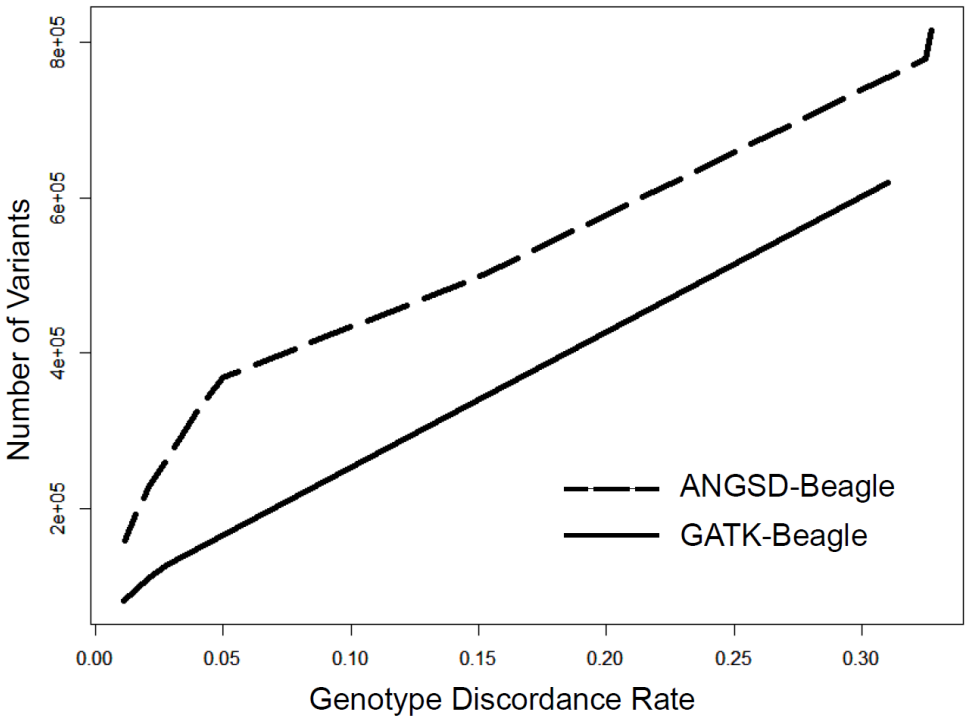
The number of called variants and genotype call rates were identical at read mapping quality (mapQ) thresholds of either 20 or 30 (Table S3) within ANGSD. As the ANGSD mapQ threshold was raised to 45, there was a small reduction in the number of called variants, and then much greater losses at thresholds of 60 or 90. Fortunately, discordance rates between ddGBS and

array genotypes were stable at both low and high mapQ thresholds, despite the putatively higher quality of the alignments (Figure S7). This permitted us to select a lower mapQ threshold (mapQ = 20), maximizing the number of variants called without sacrificing genotyping accuracy.

Variant calling

Figure 3 shows that across all levels of genotype discordance rates (comparing ddGBS with the array genotyping data), the combination of the ANGSD-SAMtools with BEAGLE produced more SNPs at various discordance thresholds than GATK's HaplotypeCaller (McKenna *et al.* 2010; DePristo *et al.* 2011). This observation held when variants were limited only to biallelic sites and SNPs with an MAF > 0.05 (Figure S8). We speculate that the poorer performance of HaplotypeCaller may be due in part to the sparsity and non-uniform distribution of GBS genotype data across the genome and the high level of genotype call missingness across samples prior to imputation.

Figure 3. Genotype discordance rates between array data and variants called by GATK/Beagle or ANGSD-SAMtools/Beagle.



The figure compares the number of variants called by combination of ANGSD-SAMtools and Beagle or GATK HaplotypeCaller and Beagle at various thresholds of genotype discordance with array data. Calls were made using the 96 HS rats with array data. The x-axis represents the genotype discordance rate thresholds and the y-axis is the number of variants that surpass that threshold for each genotype calling method.

ANGSD supports four different models for estimating genotype likelihoods: SAMtools, GATK, SOAPsnp and SYK. We compared the methods to determine which produced the most SNPs with the lowest error rates. The SOAPsnp model demonstrated an advantage in genotype accuracy and number of variants called post-imputation with Beagle (Figure S9). However, SOAPsnp requires considerably more time (1.7x for 96 samples) and memory and scales poorly with sample size. With greater than 2,000 samples, we were unable to allocate sufficient memory for the SOAPsnp model to

successfully run, even after dividing the chromosomes into several, smaller chunks. The marginal benefits of SOAPsnp in accuracy and number of variants were far outweighed by its limitations when applied to a large sample set. The GATK model showed a greater number of variants for more lenient genotype discordance rate threshold. This is in contrast with what was observed in Figures 3 and Figure S8 because ANGSD utilizes the direct genotype likelihood method from the first implementation of GATK's Unified Genotyper, whereas we had previously tested GATK's HaplotypeCaller. Interestingly though, as the stringency for discordance rate increased, the number of variants converged across the SAMtools, GATK, and SOAPsnp models. We proceeded with the SAMtools model for genotype likelihood estimation due to its previous support in the GBS literature (Torkamaneh *et al.* 2017), accepting a nominal decrease in highly concordant variants (Figure S9) for a large reduction in run time and memory usage.

Imputation to reference panel

Imputation is used in two complimentary ways in our protocol. As described earlier, after ddGBS, not all samples will have sufficient sequencing coverage at captured polymorphic loci to make a confident genotype call. Therefore, we first use imputation from other well-covered samples to "fill in the blanks" and assign genotypes to SNP loci in the subset of individuals that lacked confident calls at these sites. After these missing genotypes have been imputed in all samples, we then use the genotype information we have for the SNPs captured by ddGBS along with the reference panel data on the

original 8 HS founders (Hermsen *et al.* 2015; Ramdas *et al.* 2019) to impute genotype calls at sites that were inaccessible to ddGBS sequencing. Thus, our second application of imputation is similar to the human genetics application in which imputation using 1000 Genomes (1000 Genomes Project Consortium *et al.* 2010) increases the number of SNPs beyond those included on a given microarray platform. IMPUTE2 was selected over Beagle for this application because it has been shown to perform better with smaller reference panels from populations with substantial LD (Frischknecht *et al.* 2014; Friedenberg and Meurs 2016)

Before starting this imputation step, we observed an inflated transition/ transversion ratio (Table S4) in our ANGSD-SAMtools/Beagle SNPs. This issue was ameliorated when the SNP set was filtered for only “known” variants that were previously identified in either the 42 inbred strains (Hermsen *et al.* 2015) or the 8 deep-sequenced HS founders (Ramdas *et al.* 2019). For imputation, we therefore only provided IMPUTE2 with previously identified variant sites from our ANGSD-SAMtools/Beagle output. Prior to running IMPUTE2, we also filtered the variants for biallelic sites with a genotype call in at least two individuals. Using pedigree data for the HS rats, we further removed samples showing an order of magnitude higher level of Mendelian error than the sample mean. We further removed SNPs that had an error rate surpassing a threshold of ~0.005 (Figure S10; inflection point). There were 4 samples and 4,179 SNPs removed from subsequent analyses. Lastly, we removed any samples where the X chromosome read ratio (reads mapped to

593 the X chromosome divided by total reads) was incompatible with their
594 reported sex. We used hard threshold of 3% of total reads (empirically
595 determined), where individuals with more than 3% X-mapped reads were
596 determined to be female and below 3%, male (Figure S1).

597 There were three major genomic reference datasets available for the
598 HS rats. The first reference set was obtained from Baud et al. (Rat Genome
599 Sequencing and Mapping Consortium *et al.* 2013) and contained sequence
600 data and genotype calls for the 8 founders of the HS. The second came from
601 Hermsen et al. (Hermsen *et al.* 2015) which contains sequence and genotype
602 data on 42 distinct laboratory rats strains and substrains, 8 of which were
603 the founders of the HS from Baud et al., but analyzed alongside a new set of
604 strains. The third reference set came from Ramdas et al. (Ramdas *et al.*
605 2019), who independently performed whole-genome sequencing and made
606 genotype calls on the 8 HS founder strains. It was unclear which set of
607 genotypes would provide the best reference for imputation from our ddGBS
608 data, so we tested five different possible subsets of available data (Table 2).
609 From Hermsen et al., we used (1) all 42 inbred strains, (2) only the 34 strains
610 that were not the HS founders, and (3) only the 8 HS founder strains. Then
611 from Baud et al. and Ramdas et al., we tested the 8 HS founders only from
612 each study. The most accurate imputation was observed for the reference
613 set containing only the 8 deep-sequenced HS founder strains (Ramdas *et al.*
614 2019); however, imputation to this set had the lowest genotyping rate of all
615 panels. In contrast, using the 42 rat inbred strains displayed a balance of

high accuracy and low missingness, leading us to choose this as our reference set. To better understand the role of the 8 founder strains, which were part of the 42 strain reference panel, we created a reference panel that included only the 34 non-HS founder strains. As expected, discordance rates were much higher when only considering non-founders. However, the genotype missingness was lower for the 34 than the 8 founders alone, suggesting a combination of the two was the optimal set.

Table 2. Imputation accuracy based on different variant reference panels for IMPUTE2.

The table includes five different possible reference panels for imputation. The 42 inbred strains, 34 non-founder inbred strains, and 8 HS founders from the 42 inbred strains all were derived from Hermsen et al. 2015 (Hermsen et al. 2015). The UMich 8 HS founders were obtained from Ramdas et al. 2019 (Ramdas et al. 2019). The final set of 8 HS founder was taken from Baud et al. 2013 (Rat Genome Sequencing and Mapping Consortium et al. 2013).

		Chr1	Chr2
42 inbred strains	**Discordance rate**	0.011	0.010
	# Variants	790,659	882,993

	Genotyping Rate	0.85	0.81
All 34 non-founder inbred strains	**Discordance rate**	0.035	0.030
	# Variants	812,550	912,749
	Genotyping Rate	0.84	0.80
8 HS founders only from the 42 inbred strains	**Discordance rate**	0.012	0.011
	# Variants	805,424	902,061
	Genotyping Rate	0.57	0.53
UMich 8 HS founders only	**Discordance rate**	0.0059	0.008
	# Variants	865,514	898,621
	Genotyping Rate	0.42	0.41
Baud et. al 2013 8 HS founders only	**Discordance rate**	0.0095	0.0096
	# Variants	507,909	540,844
	Genotyping Rate	0.43	0.40

IMPUTE2 requires a genetic map. As described in the methods section, we considered four different genetic maps, two that were empirically derived and two that were linear extrapolations based on the physical map (Figure S11). All genetic maps performed similarly (Table S5). Surprisingly, the linear genetic maps performed just as well as the HS-specific map (Littrell *et al.* 2018). Thus, for simplicity, we chose to use the chromosome-specific values initially published by Jensen-Seaman (Jensen-Seaman 2004).

To obtain our final set of ~3.7 million variants, a final round of variant filtering was performed after imputation to the 42 strain reference panel. We removed SNPs with MAF < 0.005, a post-imputation genotyping rate < 90%, and SNPs that violated HWE with $p<1x10^{-10}$.

DISCUSSION

The use of microarrays and WGS for genotyping large samples in model organisms remains cost-prohibitive. There is therefore an urgent and widespread need for high-performance and economical methods of obtaining genome-wide genotype data. While reduced-representation approaches have been utilized in numerous species of plants and animals, including rodents (Peterson *et al.* 2012; Fitzpatrick *et al.* 2013; Parker *et al.* 2016; Zhou *et al.* 2018; Gonzales *et al.* 2018), there has yet to be a published protocol optimized specifically for rats. Prior to sequencing thousands of HS samples with GBS for our mapping efforts, we wanted to ensure we were capturing the greatest possible number of high-quality variants at the lowest possible cost. The protocol we present here is the culmination of careful testing and optimization of each step of the GBS protocol for rats. We have now applied the approach to 4,973 HS rats, as well as 4,608 Sprague Dawley rats (Gileta *et al.* 2018).

Our previous GBS protocol (Parker *et al.* 2016), which was designed for use with CFW mice, was unsuitable for our current genotyping efforts in HS rats, due to the much higher levels of genetic diversity in the HS population. There are multiple reasons we chose to develop our own computational pipeline for GBS rather than using existing workflows. Foremost, the prominent GBS analysis pipelines were developed and optimized for use in crop species (Sonah *et al.* 2013; Catchen *et al.* 2013; Glaubitz *et al.* 2014; Torkamaneh *et al.* 2017; Wickland *et al.* 2017), some of which are polyploid

673 and have differing levels of variation and LD than outbred rodent
674 populations. Additionally, there were elements of each pipeline that did not
675 meet our needs or lacked customizability. For instance, TASSEL-GBS v2
676 (Glaubitz *et al.* 2014) trims all reads to 92 base pairs; however, other
677 projects underway in our lab utilized up to 125bp reads, leading to a ~20%
678 reduction in data. TASSEL-GBS also ignores read base quality scores, which
679 are informative in probabilistic frameworks for estimating uncertainty in
680 alignments and variant calls (Li *et al.* 2008; DePristo *et al.* 2011; Nielsen *et*
681 *al.* 2011), and uses a naïve binomial likelihood ratio method for calling SNPs.
682 Stacks has previously shown poor performance in demultiplexing (Herten *et*
683 *al.* 2015; Torkamaneh *et al.* 2017) and does not make use of the reference
684 genome for priors when calling SNPs (Catchen *et al.* 2013). Fast-GBS relies
685 on Platypus (Rimmer *et al.* 2014) for variant calling (WGS500 Consortium *et*
686 *al.* 2014; Torkamaneh *et al.* 2017), which employs a Bayesian method of
687 constructing candidate haplotypes that works poorly with low-pass
688 sequencing data and does not scale well to large sample sizes (Li *et al.*
689 2018). Lastly, none of these pipelines included an imputation step, which is
690 crucial for filling in missing genotypes in GBS data and can provide millions
691 of additional SNPs given an appropriate composite reference panel (Howie *et*
692 *al.* 2011; Huang and Tseng 2014).

693 Though we have not explicitly tested each alternate GBS pipeline for
694 the purposes of this publication, this has been recently done by Wickland et
695 al. (Wickland *et al.* 2017). Their pipeline GB-eaSy, which ours most closely

resembles, was found to be superior by a number of metrics to Stacks, TASSEL-GBS, IGST, and Fast-GBS. Similar to GB-eaSy, our pipeline utilizes a double-digest GBS protocol, aligns reads to the reference genome with *bwa mem*, and uses the SAMtools genotype likelihood model for calling SNPs (Li 2011). The combination of bwa mem and SAMtools algorithm was independently shown to have the best performance for calling SNPs from Illumina data (Hwang *et al.* 2015), further supporting our choice of these programs for read alignment and variant calling. Additionally, using the ANGSD wrapper provided us with the ability to convert the posterior genotype probabilities into genotype dosages for mapping studies (Korneliussen *et al.* 2014).

A minor difference between GB-eaSy and our pipeline is the use of cutadapt (Martin 2011) rather than GBSX (Herten *et al.* 2015) for demultiplexing, though both performed equally well (Table S1). The primary improvement is our extension of the pipeline with the implementation of effective internal and reference-based imputation steps using the 42 inbred rat genomes (Hermsen *et al.* 2015) and 8 deep-sequenced HS founders from UMich (Ramdas *et al.* 2019). There are two stages of imputation in our pipeline. The first one is accomplished by Beagle and aims to fill in missing genotypes at called variants using information from other samples. This raises the genotype call rate to 100%, but it may also introduce errors due to insufficient information, emphasizing the need for careful filtering steps. The second stage of imputation made use of IMPUTE2 and an external reference

719 panels of variants called from WGS data on the 8 inbred HS founders, as well
720 as 34 additional inbred rat strains. We decided to include the 34 additional
721 strains because of the elevated genotyping rate we observed upon their
722 inclusion in the IMPUTE2 reference panel. We attribute this to the presence
723 of haplotypes that exist in both the 8 the HS founder strains and a subset of
724 the 34 additional strains in this panel. The benefits of using a composite
725 reference panel have been previously noted (Zhang *et al.* 2013; Huang and
726 Tseng 2014); there is increased accuracy and decreased missingness in the
727 imputed genotype data.

728 In summary, we have adapted a GBS protocol and genotyping and
729 imputation pipeline to obtain dense genotypes on genome-wide markers in
730 highly-multiplexed HS rats. After quality filtering on the level of SNP and
731 sample, over 3.7 million SNPs were called with a concordance rate of 99%.
732 The ddGBS protocol and bioinformatic methods used to produce this data are
733 publicly available, easy to handle, and cost-effective. The presented
734 workflow could be feasibly followed with marginal modifications for
735 application in other species. The steps taken toward optimizing the wet lab
736 protocols are easily applied to novel organisms, as is the computational
737 pipeline so long as there are reliable reference genome sets available for use
738 in alignment and imputation.

739

740

ACKNOWLEDGEMENTS:

This work was supported by P50 DA037844, R21 DA036672, T32 GM07197, and F31 DA039638.

LITERATURE CITED

1000 Genomes Project Consortium, G. R. Abecasis, D. Altshuler, A. Auton, L. D. Brooks *et al.*, 2010 A map of human genome variation from population-scale sequencing. Nature 467: 1061–1073.

Aird, D., M. G. Ross, W.-S. Chen, M. Danielsson, T. Fennell *et al.*, 2011 Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biology 12: R18.

Andolfatto, P., D. Davison, D. Erezyilmaz, T. T. Hu, J. Mast *et al.*, 2011 Multiplexed shotgun genotyping for rapid and efficient genetic mapping. Genome Res. 21: 610–617.

Andrews, S., 2017 *FastQC*.

Baird, N. A., P. D. Etter, T. S. Atwood, M. C. Currey, A. L. Shiver *et al.*, 2008 Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers (J. C. Fay, Ed.). PLoS ONE 3: e3376.

Browning, B. L., and S. R. Browning, 2009 A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals. The American Journal of Human Genetics 84: 210–223.

Browning, B. L., and S. R. Browning, 2016 Genotype Imputation with Millions of Reference Samples. The American Journal of Human Genetics 98: 116–126.

Catchen, J., P. A. Hohenlohe, S. Bassham, A. Amores, and W. A. Cresko, 2013 Stacks: an analysis tool set for population genomics. Molecular Ecology 22: 3124–3140.

Chen, Q., Y. Ma, Y. Yang, Z. Chen, R. Liao *et al.*, 2013 Genotyping by Genome Reducing and Sequencing for Outbred Animals (S. Zhao, Ed.). PLoS ONE 8: e67500.

Chitre, A. S., O. Polesskaya, K. Holl, J. Gao, R. Cheng *et al.*, 2018 Genome wide association study of body weight, body mass index, adiposity, and fasting glucose in 3,173 outbred rats.

Davey, J. W., P. A. Hohenlohe, P. D. Etter, J. Q. Boone, J. M. Catchen *et al.*, 2011 Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nature Reviews Genetics 12: 499–510.

De Donato, M., S. O. Peters, S. E. Mitchell, T. Hussain, and I. G. Imumorin, 2013 Genotyping-by-Sequencing (GBS): A Novel, Efficient and Cost-Effective Genotyping Method for Cattle Using Next-Generation Sequencing (J. C. Nelson, Ed.). PLoS ONE 8: e62137.

DePristo, M. A., E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire *et al.*, 2011 A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics 43: 491–498.

Durvasula, A., P. J. Hoffman, T. V. Kent, C. Liu, T. J. Y. Kono *et al.*, 2016 ANGSD - wrapper: utilities for analysing next-generation sequencing data. Mol Ecol Resour 16: 1449–1454.

Elshire, R. J., J. C. Glaubitz, Q. Sun, J. A. Poland, K. Kawamoto *et al.*, 2011 A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species (L. Orban, Ed.). PLoS ONE 6: e19379.

Fitzpatrick, C. J., S. Gopalakrishnan, E. S. Cogan, L. M. Yager, P. J. Meyer *et al.*, 2013 Variation in the Form of Pavlovian Conditioned Approach Behavior among Outbred Male Sprague-Dawley Rats from Different Vendors and Colonies: Sign-Tracking vs. Goal-Tracking (P. Campolongo, Ed.). PLoS ONE 8: e75042.

803 Flanagan, S. P., and A. G. Jones, 2018 Substantial differences in bias between
804 single-digest and double-digest RAD-seq libraries: A case study. Molecular
805 Ecology Resources 18: 264–280.
806 Friedenberg, S. G., and K. M. Meurs, 2016 Genotype imputation in the domestic dog.
807 Mamm. Genome 27: 485–494.
808 Frischknecht, M., M. Neuditschko, V. Jagannathan, C. Drögemüller, J. Tetens *et al.*,
809 2014 Imputation of sequence level genotypes in the Franches-Montagnes
810 horse breed. Genet. Sel. Evol. 46: 63.
811 Fu, Y.-B., and M.-H. Yang, 2017 Genotyping-by-Sequencing and Its Application to
812 Oat Genomic Research, pp. 169–187 in *Oat*, edited by S. Gasparis. Springer
813 New York, New York, NY.
814 Furuta, T., M. Ashikari, K. K. Jena, K. Doi, and S. Reuscher, 2017 Adapting
815 Genotyping-by-Sequencing for Rice F2 Populations. G3: Genes|
816 Genomes|Genetics 7: 881–893.
817 Gileta, A. F., C. J. Fitzpatrick, A. S. Chitre, C. L. St. Pierre, E. V. Joyce *et al.*, 2018
818 Genetic characterization of outbred Sprague Dawley rats and utility for
819 genome-wide association studies.
820 Glaubitz, J. C., T. M. Casstevens, F. Lu, J. Harriman, R. J. Elshire *et al.*, 2014 TASSEL-
821 GBS: A High Capacity Genotyping by Sequencing Analysis Pipeline (N. A.
822 Tinker, Ed.). PLoS ONE 9: e90346.
823 Gonzales, N. M., J. Seo, A. I. Hernandez-Cordero, C. L. St. Pierre, J. S. Gregory *et al.*,
824 2018 Genome wide association analysis in a mouse advanced intercross line.
825 Hannon Lab, 2010 *FASTX-Toolkit*.

He, J., X. Zhao, A. Laroche, Z.-X. Lu, H. Liu *et al.*, 2014 Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. Frontiers in Plant Science 5:.

Hermsen, R., J. de Ligt, W. Spee, F. Blokzijl, S. Schäfer *et al.*, 2015 Genomic landscape of rat strain and substrain variation. BMC Genomics 16:.

Herten, K., M. S. Hestand, J. R. Vermeesch, and J. K. Van Houdt, 2015 GBSX: a toolkit for experimental design and demultiplexing genotyping by sequencing experiments. BMC Bioinformatics 16:.

Howie, B. N., P. Donnelly, and J. Marchini, 2009 A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies (N. J. Schork, Ed.). PLoS Genetics 5: e1000529.

Howie, B., C. Fuchsberger, M. Stephens, J. Marchini, and G. R. Abecasis, 2012 Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nature Genetics 44: 955.

Howie, B., J. Marchini, and M. Stephens, 2011 Genotype Imputation with Thousands of Genomes. G3: Genes|Genomes|Genetics 1: 457–470.

Huang, X., Q. Feng, Q. Qian, Q. Zhao, L. Wang *et al.*, 2009 High-throughput genotyping by whole-genome resequencing. Genome Research 19: 1068–1076.

Huang, G.-H., and Y.-C. Tseng, 2014 Genotype imputation accuracy with different reference panels in admixed populations. BMC Proceedings 8: S64.

Hwang, S., E. Kim, I. Lee, and E. M. Marcotte, 2015 Systematic comparison of variant calling pipelines using gold standard personal exome variants. Scientific Reports 5: 17875.

Illumina, Inc., 2014 Nextera(R) Library Validation and Cluster Density Optimization: Guidelines for generating high-quality data with Nextera library prepartion kits.

Jensen-Seaman, M. I., 2004 Comparative Recombination Rates in the Rat, Mouse, and Human Genomes. Genome Research 14: 528–538.

Johannesson, M., R. Lopez-Aumatell, P. Stridh, M. Diez, J. Tuncel *et al.*, 2008 A resource for the simultaneous high-resolution mapping of multiple quantitative trait loci in rats: The NIH heterogeneous stock. Genome Research 19: 150–158.

Johnson, J. L., H. Wittgenstein, S. E. Mitchell, K. E. Hyma, S. V. Temnykh *et al.*, 2015 Genotyping-By-Sequencing (GBS) Detects Genetic Structure and Confirms Behavioral QTL in Tame and Aggressive Foxes (Vulpes vulpes) (W. J. Murphy, Ed.). PLOS ONE 10: e0127013.

Kanagawa, T., 2003 Bias and artifacts in multitemplate polymerase chain reactions (PCR). Journal of Bioscience and Bioengineering 96: 317–323.

Kent, W. J., C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle *et al.*, 2002 The Human Genome Browser at UCSC. Genome Research 12: 996–1006.

Korneliussen, T. S., A. Albrechtsen, and R. Nielsen, 2014 ANGSD: Analysis of Next Generation Sequencing Data. BMC Bioinformatics 15:.

Li, H., 2011 A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27: 2987–2993.

Li, H., and R. Durbin, 2009 Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754–1760.

Li, H., J. Ruan, and R. Durbin, 2008 Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Research 18: 1851–1858.

Li, Z., Y. Wang, and F. Wang, 2018 A study on fast calling variants from next-generation sequencing data using decision tree. BMC Bioinformatics 19:.

Littrell, J., S.-W. Tsaih, A. Baud, P. Rastas, L. Solberg-Woods *et al.*, 2018 A High-Resolution Genetic Map for the Laboratory Rat. G3: Genes| Genomes|Genetics g3.200187.2018.

Martin, M., 2011 Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17: 10.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis *et al.*, 2010 The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research 20: 1297–1303.

Miller, M. R., J. P. Dunham, A. Amores, W. A. Cresko, and E. A. Johnson, 2007 Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. Genome Research 17: 240–248.

Nielsen, R., J. S. Paul, A. Albrechtsen, and Y. S. Song, 2011 Genotype and SNP calling from next-generation sequencing data. Nature Reviews Genetics 12: 443–451.

van Orsouw, N. J., R. C. J. Hogers, A. Janssen, F. Yalcin, S. Snoeijers *et al.*, 2007 Complexity Reduction of Polymorphic Sequences (CRoPS™): A Novel Approach for Large-Scale Polymorphism Discovery in Complex Genomes (I. Baxter, Ed.). PLoS ONE 2: e1172.

Parker, C. C., S. Gopalakrishnan, P. Carbonetto, N. M. Gonzales, E. Leung *et al.*, 2016 Genome-wide association study of behavioral, physiological and gene expression traits in outbred CFW mice. Nature Genetics 48: 919–926.

Pértille, F., C. Guerrero-Bosagna, V. H. da Silva, C. Boschiero, J. de R. da S. Nunes *et al.*, 2016 High-throughput and Cost-effective Chicken Genotyping Using Next-Generation Sequencing. Scientific Reports 6: 26929.

Peterson, B. K., J. N. Weber, E. H. Kay, H. S. Fisher, and H. E. Hoekstra, 2012 Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species (L. Orlando, Ed.). PLoS ONE 7: e37135.

Poland, J. A., P. J. Brown, M. E. Sorrells, and J.-L. Jannink, 2012 Development of High-Density Genetic Maps for Barley and Wheat Using a Novel Two-Enzyme Genotyping-by-Sequencing Approach (T. Yin, Ed.). PLoS ONE 7: e32253.

Poland, J. A., and T. W. Rife, 2012 Genotyping-by-Sequencing for Plant Breeding and Genetics. The Plant Genome Journal 5: 92.

Ramdas, S., A. B. Ozel, M. K. Treutelaar, K. Holl, M. Mandel *et al.*, 2019 Extended regions of suspected mis-assembly in the rat reference genome. Sci Data 6: 39.

Rat Genome Sequencing and Mapping Consortium, A. Baud, R. Hermsen, V. Guryev, P. Stridh *et al.*, 2013 Combined sequence-based and genetic mapping analysis of complex traits in outbred rats. Nature Genetics 45: 767–775.

Rice, P., I. Longden, and A. Bleasby, 2000 EMBOSS: The European Molecular Biology Open Software Suite. Trends in Genetics 16: 276–277.

Rimmer, A., H. Phan, I. Mathieson, Z. Iqbal, S. R. F. Twigg *et al.*, 2014 Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. Nat. Genet. 46: 912–918.

Roberts, R. J., and D. Macelis, 1999 REBASE--restriction enzymes and methylases. Nucleic Acids Research 27: 312–313.

Scheben, A., J. Batley, and D. Edwards, 2017 Genotyping-by-sequencing approaches to characterize crop genomes: choosing the right tool for the right application. Plant Biotechnology Journal 15: 149–161.

Sonah, H., M. Bastien, E. Iquira, A. Tardivel, G. Légaré *et al.*, 2013 An Improved Genotyping by Sequencing (GBS) Approach Offering Increased Versatility and Efficiency of SNP Discovery and Genotyping (Z. Liu, Ed.). PLoS ONE 8: e54603.

Steen, R. G., A. E. Kwitek-Black, C. Glenn, J. Gullings-Handley, W. Van Etten *et al.*, 1999 A high-density integrated genetic linkage and radiation hybrid map of the laboratory rat. Genome Res. 9: AP1-8, insert.

Sun, X., D. Liu, X. Zhang, W. Li, H. Liu *et al.*, 2013 SLAF-seq: An Efficient Method of Large-Scale De Novo SNP Discovery and Genotyping Using High-Throughput Sequencing (J. Aerts, Ed.). PLoS ONE 8: e58700.

Torkamaneh, D., J. Laroche, M. Bastien, A. Abed, and F. Belzile, 2017 Fast-GBS: a new pipeline for the efficient and highly accurate calling of SNPs from genotyping-by-sequencing data. BMC Bioinformatics 18:.

Van Tassell, C. P., T. P. L. Smith, L. K. Matukumalli, J. F. Taylor, R. D. Schnabel *et al.*, 2008 SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nature Methods 5: 247–252.

Wang, Y., X. Cao, Y. Zhao, J. Fei, X. Hu *et al.*, 2017 Optimized double-digest genotyping by sequencing (ddGBS) method with high-density SNP markers and high genotyping accuracy for chickens (P. Xu, Ed.). PLOS ONE 12: e0179073.

WGS500 Consortium, A. Rimmer, H. Phan, I. Mathieson, Z. Iqbal *et al.*, 2014 Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. Nature Genetics 46: 912–918.

Wickland, D. P., G. Battu, K. A. Hudson, B. W. Diers, and M. E. Hudson, 2017 A comparison of genotyping-by-sequencing analysis methods on low-coverage crop datasets shows advantages of a new workflow, GB-eaSy. BMC Bioinformatics 18:.

Woods, L. C. S., and R. Mott, 2017 Heterogeneous Stock Populations for Analysis of Complex Traits, pp. 31–44 in *Systems Genetics*, edited by K. Schughart and R. W. Williams. Springer New York, New York, NY.

Zhang, P., X. Zhan, N. A. Rosenberg, and S. Zöllner, 2013 Genotype imputation reference panel selection using maximal phylogenetic diversity. Genetics 195: 319–330.

Zhou, X., C. L. St. Pierre, N. M. Gonzales, R. Cheng, A. S. Chitre *et al.*, 2018 Genome-wide association study, replication, and mega-analysis using a dense marker panel in a multi-generational mouse advanced intercross line.

Figure S1. Ratio of reads on X-chromosome to total sequencing reads.

The color of the points indicates the pedigree-recorded sex of the samples. Females are expected to have approximately twice as many reads for the X-chromosome. Samples that did not cluster with their pedigree-recorded sex were removed from the study for possible sample mix-up.

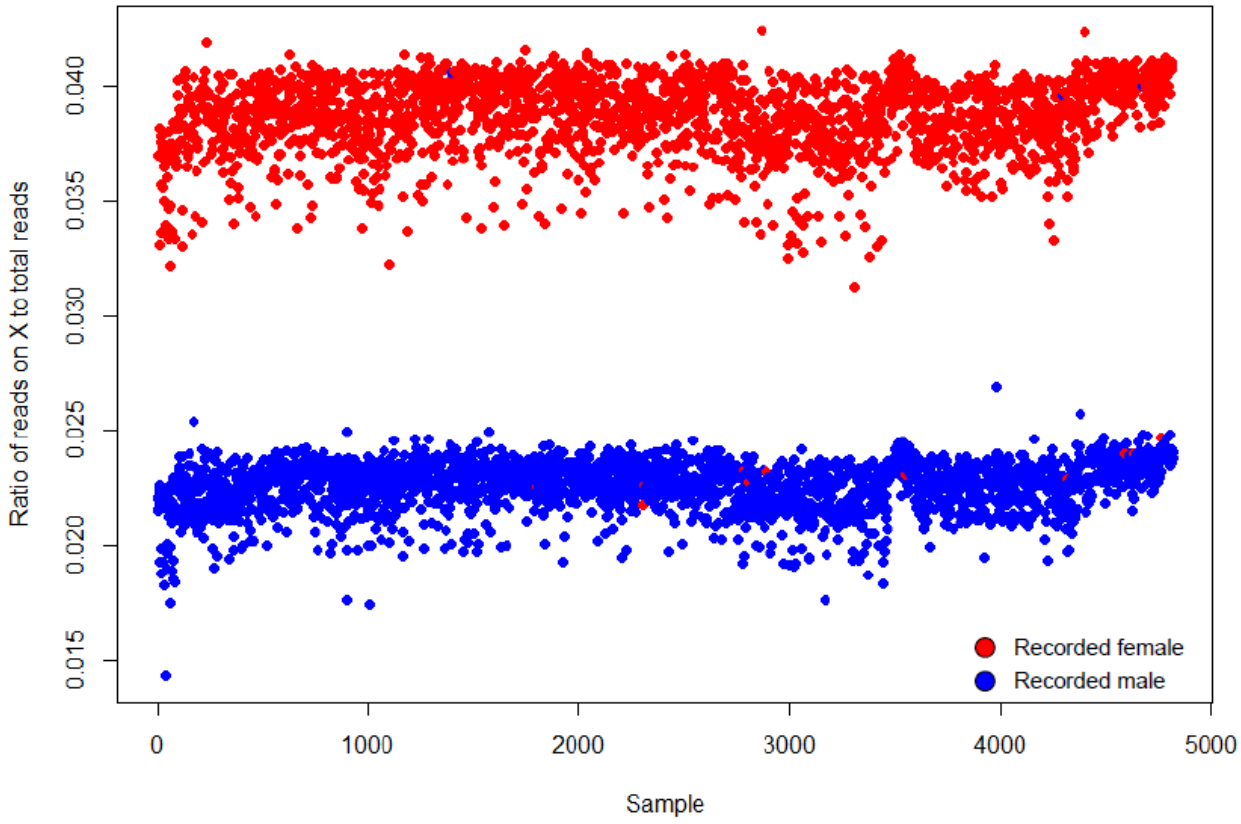

Ratio of reads on X to total reads
0.015
0.020
0.025
0.030
0.035
0.040
0
1000
2000
3000
4000
5000
Sample
Recorded female
Recorded male

Figure S2. Data preparation workflow for imputation with IMPUTE2.

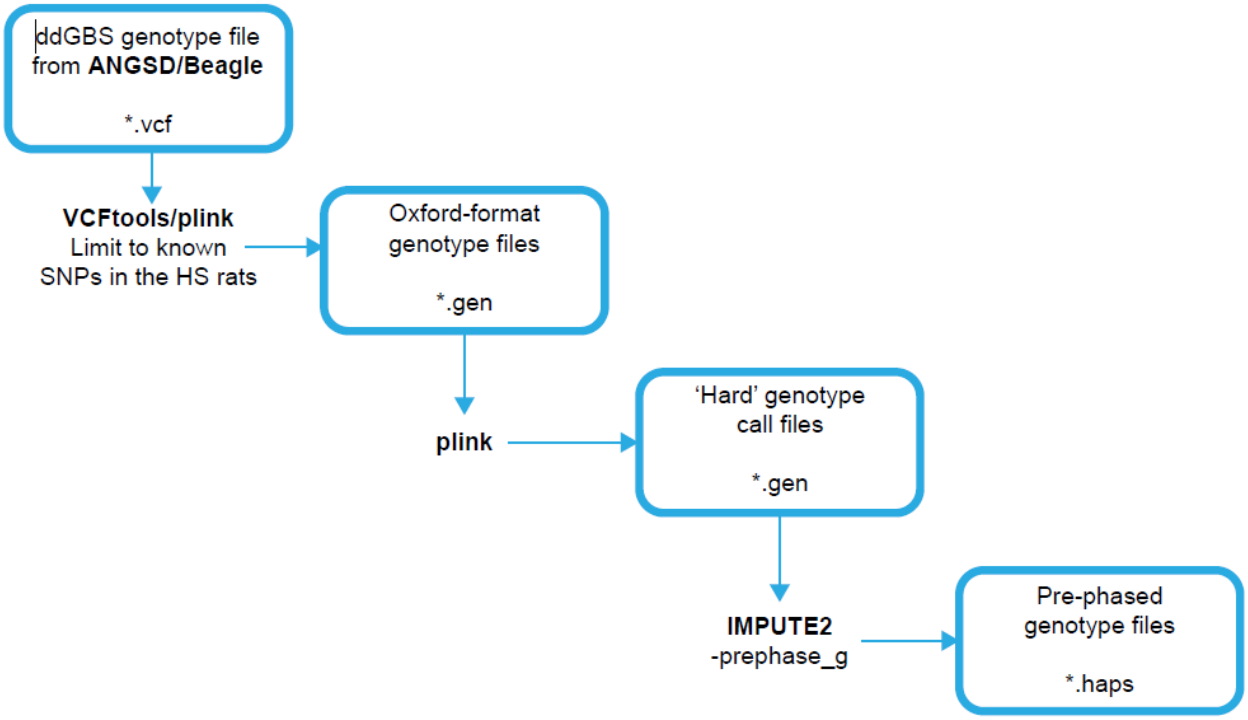


Figure S3. Programmed vs. empirical Pippin Prep fragment size range.

This plot comes from the Bioanalyzer output for a pooled HS library. The x-axis shows the library fragment sizes in base pairs, and the y-axis is in fluorescent units, which represent the quantity of the fragments on the gel chip. There is approximately a 50-75bp shift in the empirical library distribution compared to expectation due to the high quantity of fragments loaded into the Pippin Prep gel cassette.

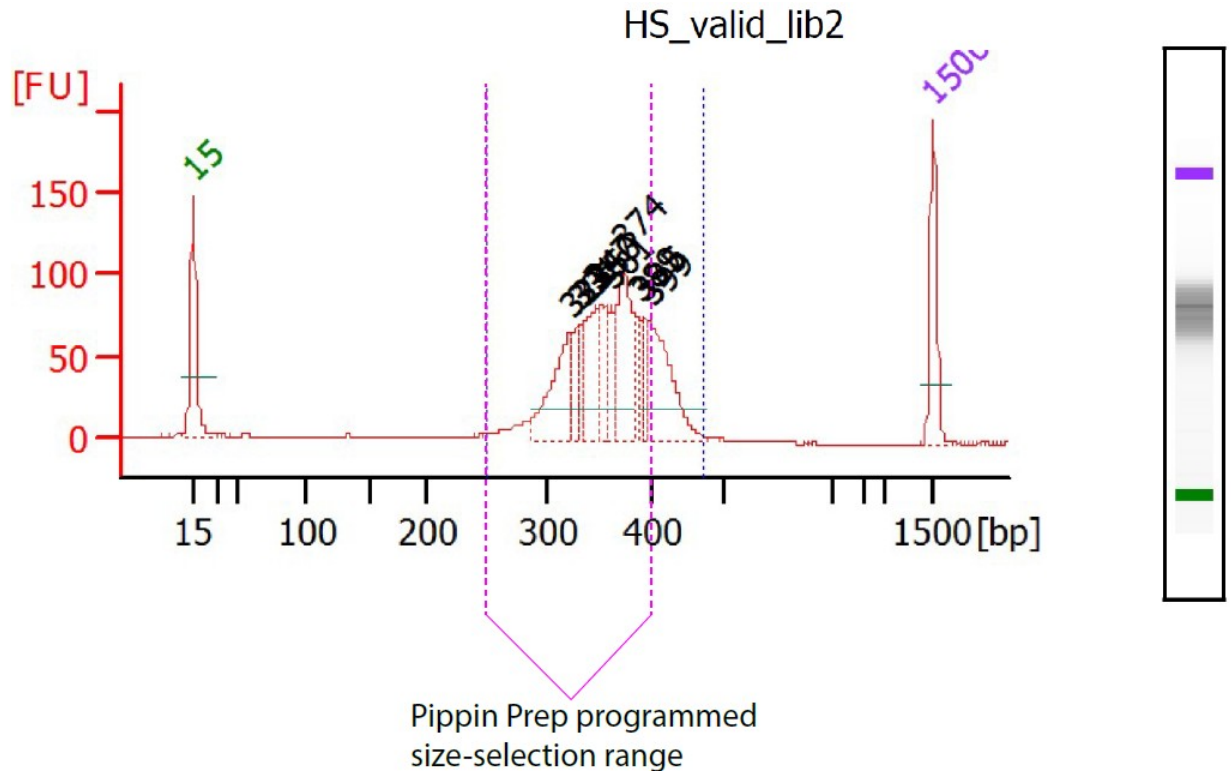


Figure S4. Raw read counts grouped by shipment batch.

Raw read counts are on a per-sample basis after demultiplexing FASTQ files with FASTX Barcode Splitter. Each batch represents a set of samples from a given shipment.

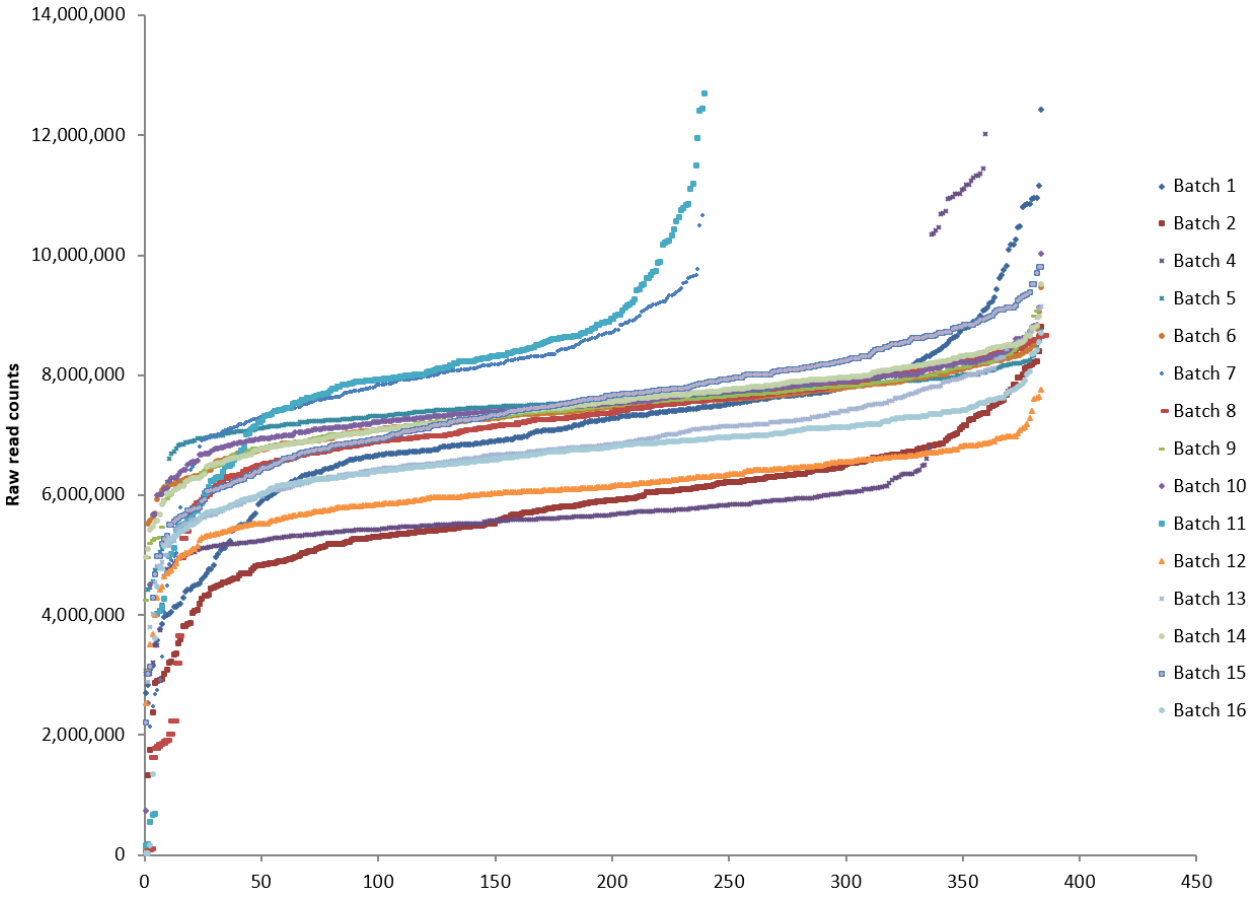


Figure S5. FASTQC results pre- and post-filtering with Cutadapt.

FASTQC results are from a single sample from the original set of 96 HS samples prepared in 12-plex and sequenced on the Illumina HiSeq 2500 with 125bp reads.

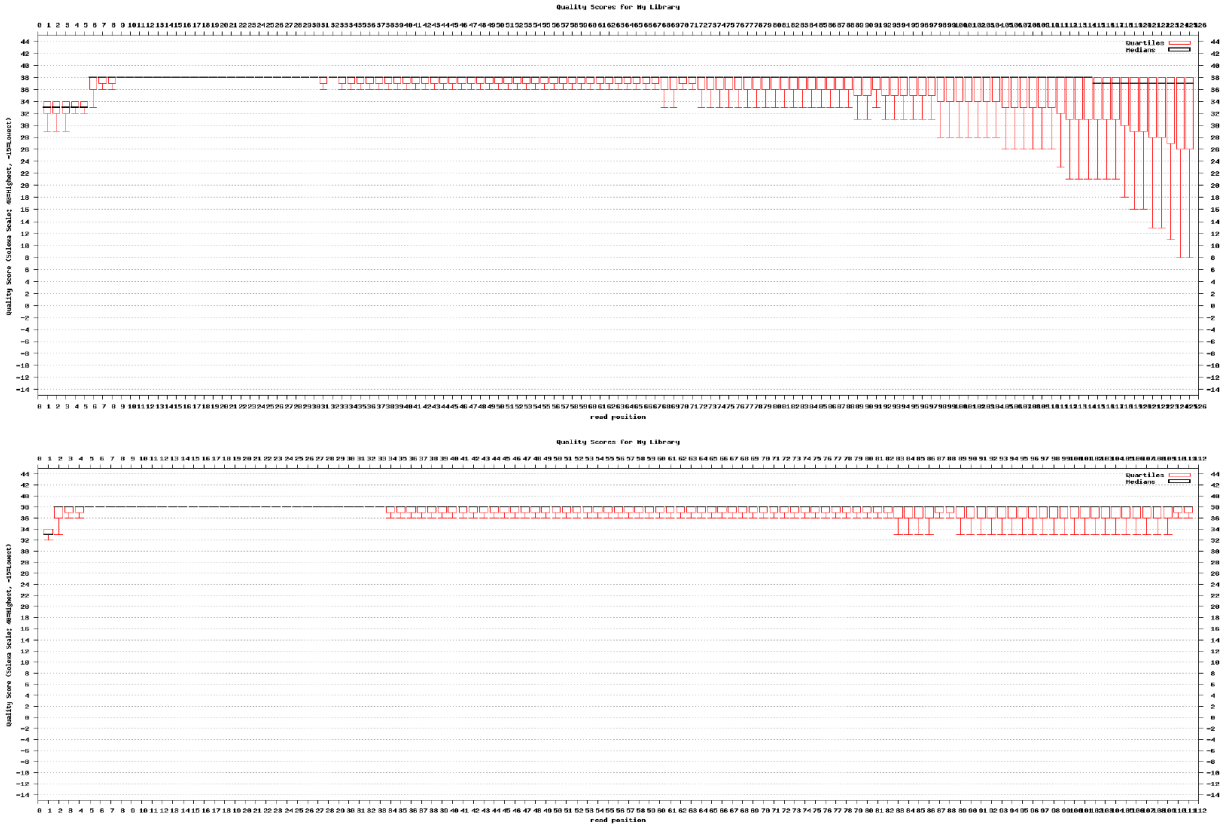


Figure S6. Overlap of called SNPs with known variants after read trimming with FASTX or Cutadapt.

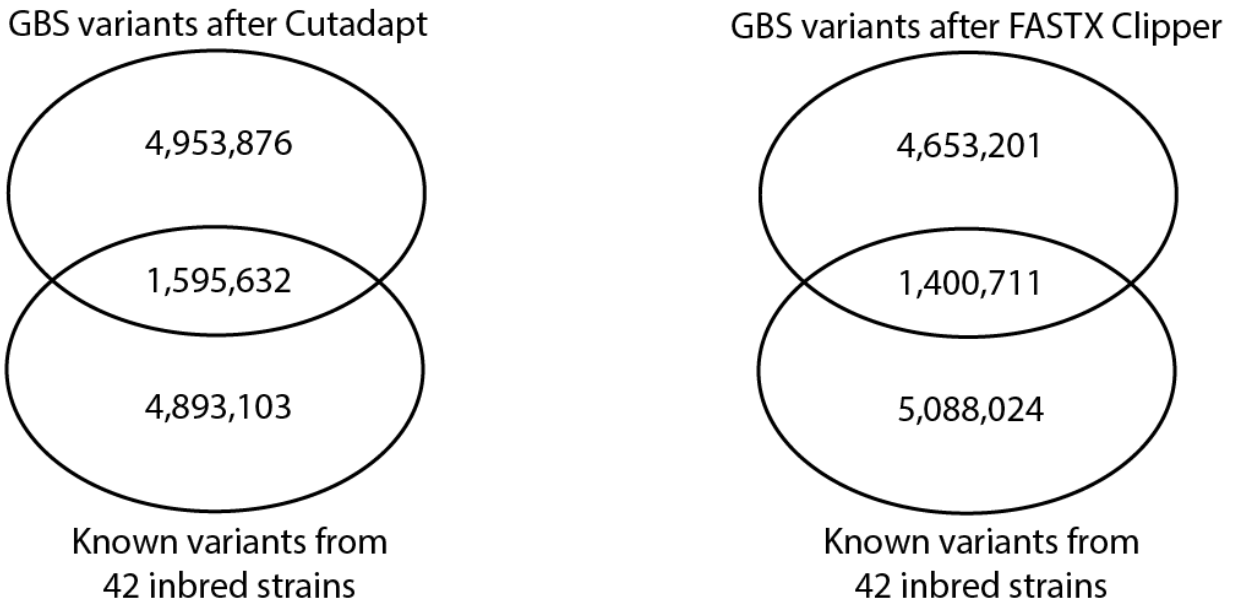


Figure S7. Mapping quality thresholds.
Genotyping error rate and number of variants by mean depth per sample per variant site for mapping quality thresholds of 20, 30, and 60.

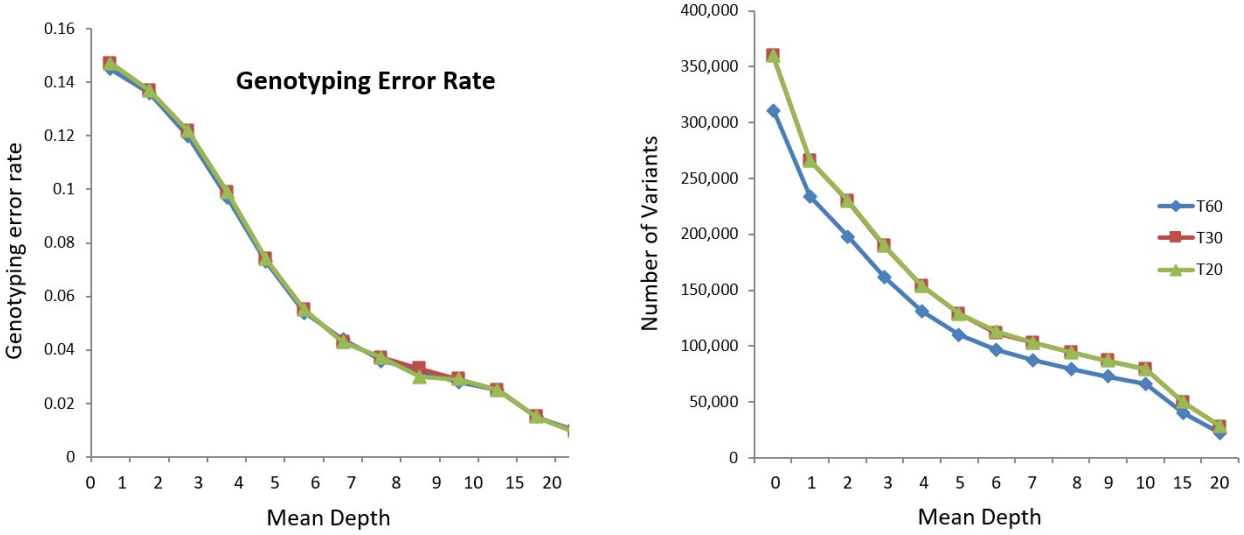


Figure S8. ANGSD-SAMtools vs GATK HaplotypeCaller, filtered calls.

The panel compares the number variants called by combination of ANGSD-SAMtools and Beagle or GATK HaplotypeCaller and Beagle at various thresholds of genotype discordance with array data. Calls were made using the 96 HS rats with array data. The x-axis represents the genotype discordance rate thresholds and the y-axis is the number of variants that surpass that threshold for each genotype calling method. Additional filters were applied to the original SNP sets and the plot zooms in on a smaller range of acceptable discordance rates compared to Figure 3. Blue lines represent the unfiltered SNP set. Yellow lines have been filtered for singletons. Red lines have further excluded SNPs with an MAF < 0.05. Each line contains the same number of points.

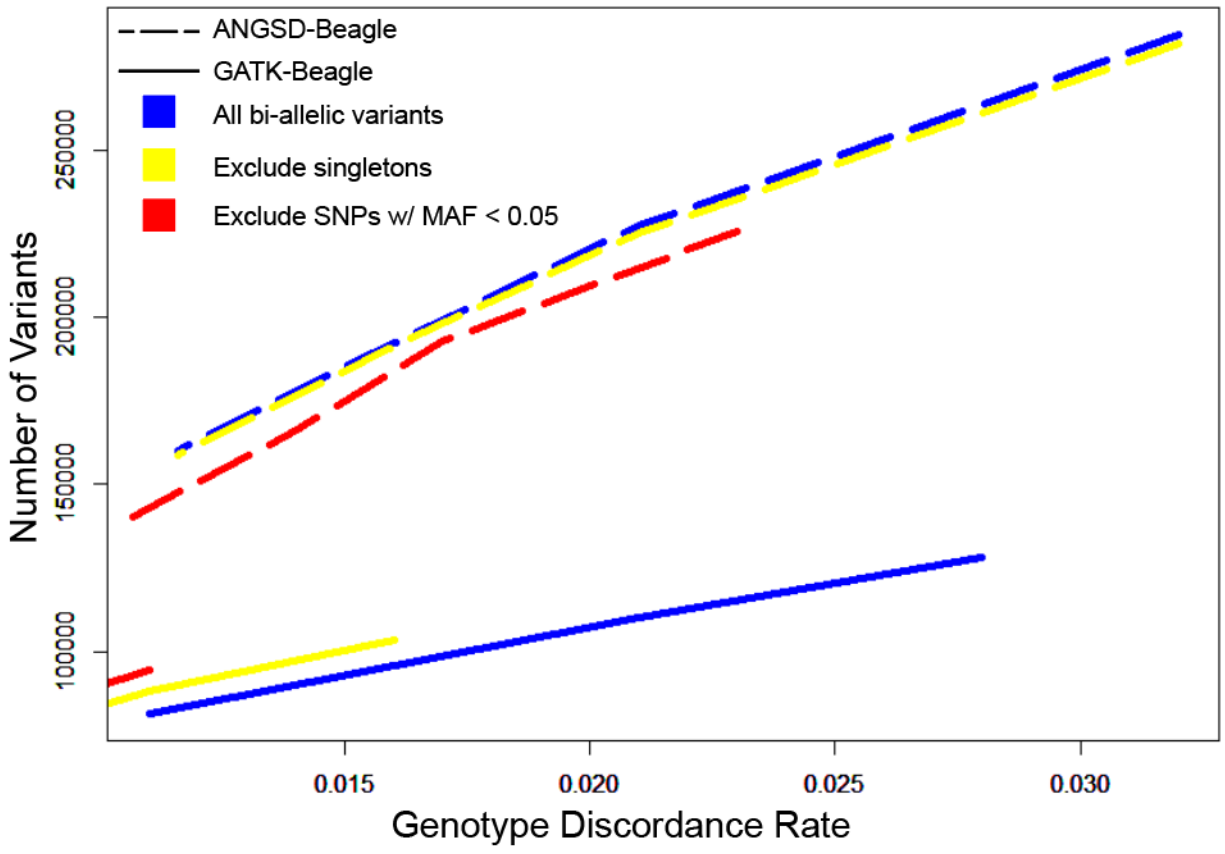


Figure S9. Number of variants by genotype discordance rates for 4 ANGSD genotype likelihood models.

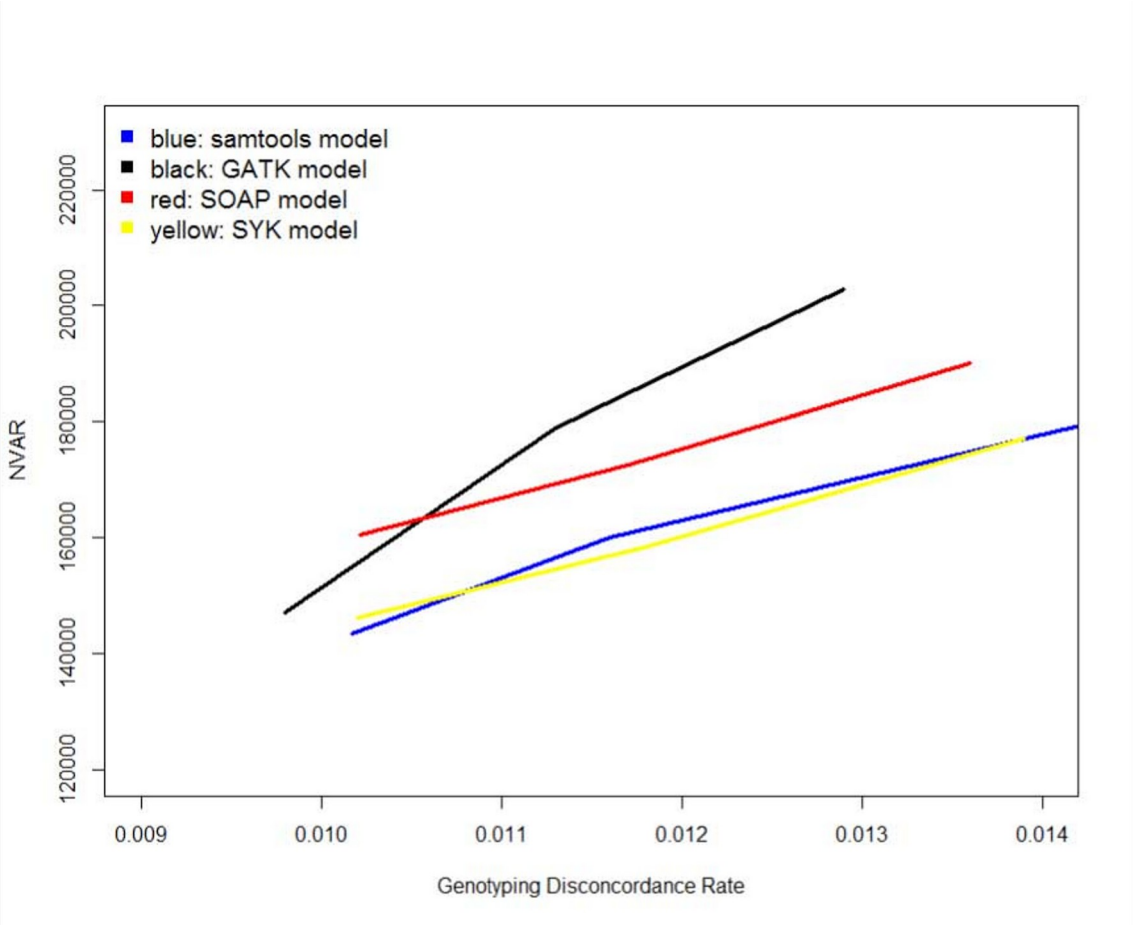


Figure S10. Mendelian error rates.

The plot shows the Mendelian error rate for all SNPs. A threshold was set at the inflection point of the curve (~0.005) and all SNPs above that threshold were removed from the data set.

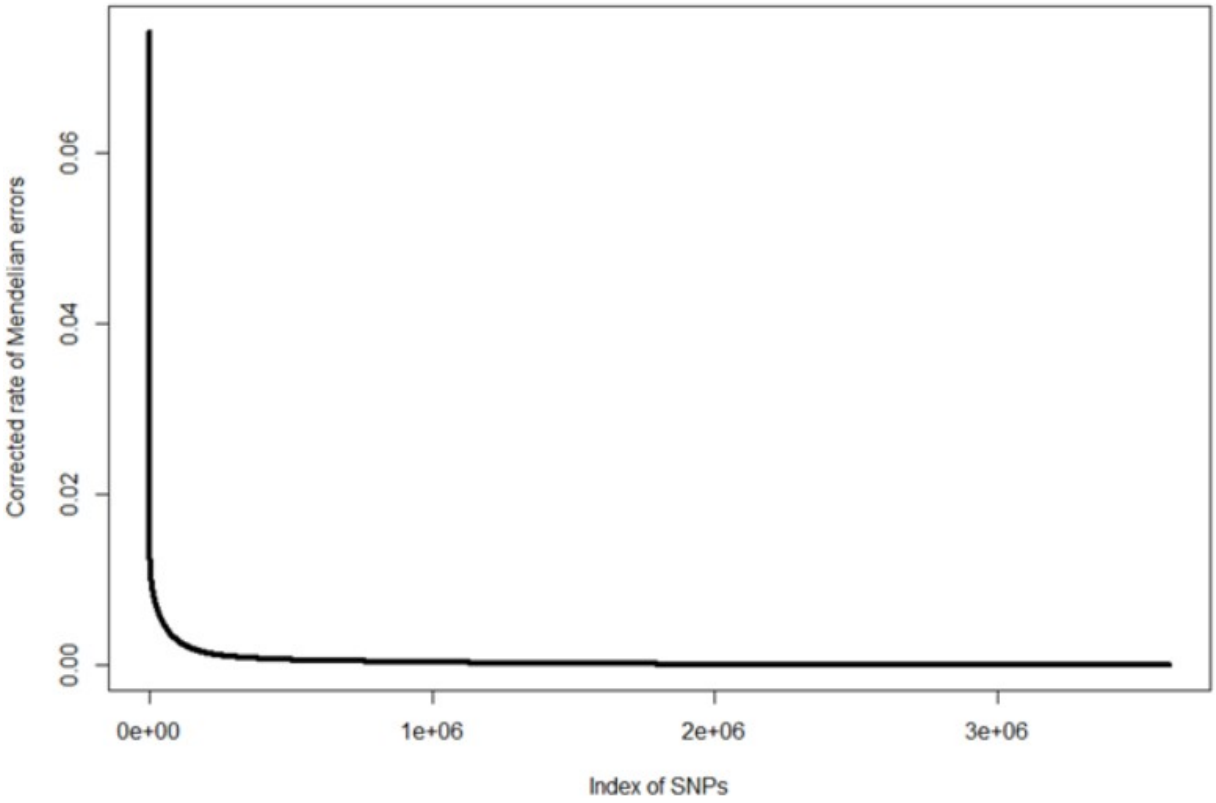


Figure S11. Available rat genetic maps.

Plotted physical and genetic distances are for chromosome 12.

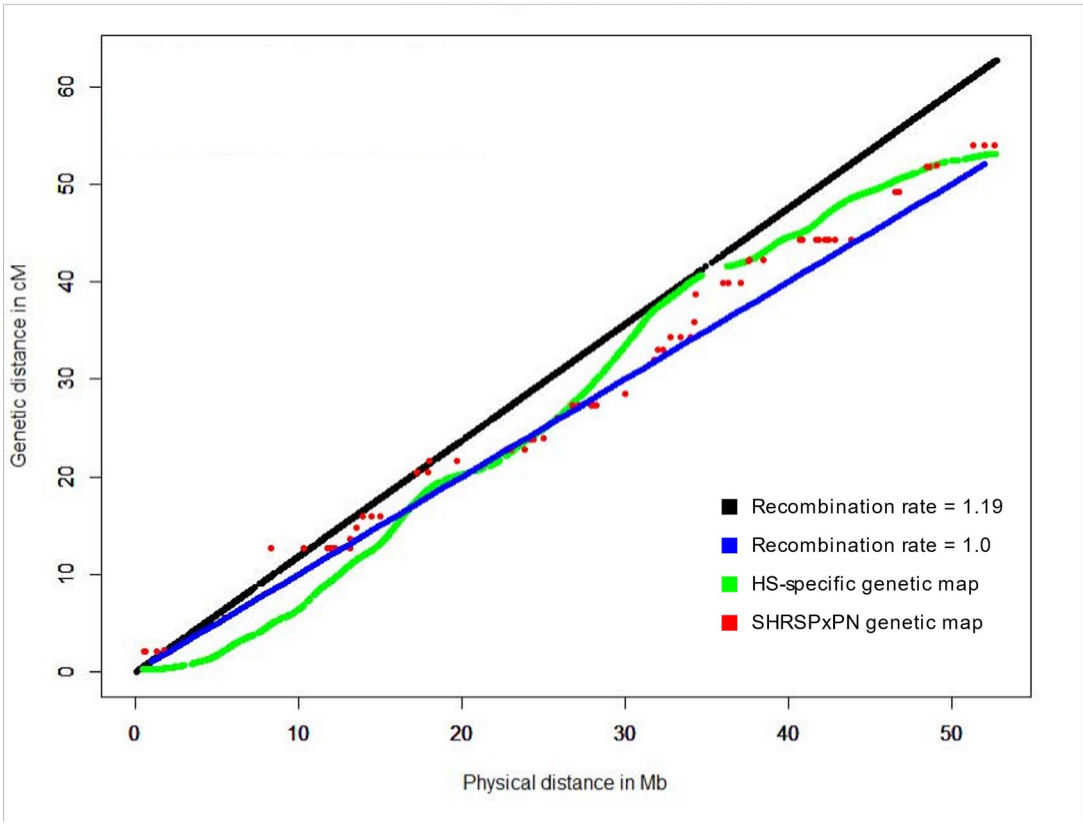


Table S1. Demultiplexing performance.

All methods began with the same number of reads from the original FASTQ. Final read and base pair counts are from after the reads have been trimmed of adapter, barcode, and restriction site sequences, as well as low-quality base pairs (< Q20).

	In-house Python Script	GBSX	FASTX Barcode Splitter
Reads with NlaIII adapter sequence	545,177 (3.07%)	475,581 (2.67%)	547,697 (3.07%)
Total bps processed	2,061,523,464	2,116,436,361	2,227,542,500
Total bps written to file	2,059,714,312	2,114,841,934	2,225,724,833
Proportion of bps retained	99.91%	99.92%	99.92%
Reads post-processing	17,771,754	17,786,280	17,820,340

Table S2. Comparison of variants calls after filtering with FASTX vs Cutadapt.

Data shown comes from the original set of 96 HS samples prepared in 12-plex and sequenced on the Illumina HiSeq 2500. At this early step of pipeline optimization, variants were called utilizing GATK UnifiedGenotyper.

	FASTX Clipper	Cutadapt
Number of variants	6,075,821	6,581,115
Genotyping call rate	0.17	0.19
Mean minor allele count	3.96	4.25
Mean minor allele frequency	0.15	0.15
Number of singletons	433,960	548,975
Number monomorphic sites	807,453	773,074
Transition/ transversion ratio	2.32	2.40
$T_I T_V$ ratio for singletons	3.23	3.40
Mean variant read depth	109.56	126.35
Mean quality score	601.79	715.56

Table S3. Variant metrics resulting from reads filtered at different mapping quality thresholds.

Data shown comes from the original set of 96 HS samples prepared in 12-plex and sequenced on the Illumina HiSeq 2500. Variants were called utilizing the SAMtools model and the -minMapQ filter in ANGSD. Calls were unfiltered.

	MAPQ = 20	**MAPQ = 30**	**MAPQ = 45**	**MAPQ = 60**	**MAPQ = 90**
Number of variants	372,860	372,330	363,790	316,949	233,322
Genotyping call rate	0.64	0.64	0.64	0.61	0.75
Mean minor allele count	5.96	5.96	6.06	5.86	7.36
Mean minor allele frequency	0.18	0.18	0.18	0.18	0.19
Number of singletons	16,781 (4.50%)	16,732 (4.49%)	16,550 (4.55%)	17,352 (5.47%)	11,773 (5.05%)
Number of monomorphic sites	122,478 (32.85%)	122,188 (32.82%)	116,738 (32.09%)	100,074 (31.57%)	56,179 (24.08%)
Transition/ transversion ratio	1.23	1.24	1.26	1.31	1.41
T_iT_v ratio for singletons	1.27	1.28	1.28	1.31	1.38
Mean variant read depth	157.78	157.73	159.25	152.48	188.80
Mean quality score	2,547	2,548	2,556	2,461	2,954

Table S4. Transition/transversion ratio before and after known sites filtering.

The presented data comes from ANGSD-SAMtools/Beagle variant calls for 3,601 HS samples, prior to imputation with IMPUTE2. Known SNPs came from both the 42 inbred genomes from Hermsen et. al 2015 (Hermsen et al. 2015) and the 8 inbred HS founder strains sequenced by the University of Michigan (Ramdas et al. 2019).

	Unfiltered SNPs	**Filtered for known SNPs**
AC	15,157	9,166
AG	888,657	42,275
AT	15,432	7,610
CG	18,043	8,061
CT	893,653	41,938
GT	15,118	9,177
T_s	1,782,310	84,213
T_v	63,750	34,014
T_sT_v	27.96	2.48
Total # SNPs	1,846,060	118,227

Table S5. Imputation accuracy for chromosome 12 across different genetic maps.

The number of variants used for the concordance check is dependent on the overlap of the imputed variants with array data for the 96 HS rats with array genotypes. The MAF filter only removes monomorphic sites within the 96 HS rat sample used for the concordance check.

	cM/Mb = 1.00	cM/Mb = 1.16	SHRSPxPN	HS-specific
Number of variants before QC	158,452	158,452	158,452	158,452
Genotyping rate before QC	0.94	0.92	0.92	0.92
Variant removed for missingness > 10%	22,217	28,959	28,356	28,858
Variants removed for MAF < 0.005	50,380	61,270	61,592	59,812
Variants removed for HWE < $1x10^{-10}$	53	56	57	56
Number of variants after QC	85,802	68,167	68,447	69,726
Genotyping rate after QC	0.93	0.91	0.92	0.91
Number of variants in concordance check	5,912	5,590	5,594	5,646
Discordance rate	0.095	0.011	0.011	0.010