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

Jianjun Gao[*⁋], Alexander F. Gileta[* †⁋], 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

[#] Department of Biology, 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.

15

16 **Running title:** GBS and variant calling in HS rats

17

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

19

20 **Corresponding author:** Abraham A. Palmer

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

22 **Phone number:** 858-534-2093

23 **Email:** aapalmer@ucsd.edu

ABSTRACT

The heterogeneous stock (**HS**) is an outbred rat population derived from eight inbred rat strains. The population is maintained with the goal of minimizing inbreeding and maximizing the genetic diversity of the stock. To effectively utilize this rat strain for fine-scale genetic mapping, genotype data is necessary for large numbers of animals. A few genotyping microarrays have been created for rats; however, they were expensive and are no longer in production. Thus, to obtain high-density genome-wide marker data for genetic mapping, we have adapted genotype-by-sequencing (**GBS**) for use in rats. Here, we outline the laboratory and computational steps we took to design and optimize an efficient double digest genotype-by-sequencing (**ddGBS**) protocol for rats. We include a detailed protocol to perform ddGBS in rats. To analyze the ddGBS sequencing data, we evaluated multiple existing computational tools and designed a workflow that allowed us to call and impute over 3.7 million SNPs genome-wide in the HS. We also compared various rat genetic maps for use in imputation, including a recently developed map specific to the HS. Using the pipeline, we obtained concordance rates of 99% with data from a rat genotyping array. The computational pipeline that we have developed can be easily adapted for use in other species.

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 many structured breeding populations, sequencing whole genomes is not necessary. SNPs are frequently in strong LD with adjacent loci, effectively 'tagging' nearby variation, and thereby

reducing the number of sites that need to be 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; X. Huang et al. 2009; Davey et al. 2011; Elshire et al. 2011; Poland et al. 2012; Peterson et al. 2012; Sun et al. 2013; Scheben, Batley, and Edwards 2017). Thousands of SNPs can be identified in large numbers of samples for a fraction of the price of whole-genome sequencing methods (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 non-model species, such as: 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; Gonzales et al. 2017; Zhou 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 and given the lack of well-developed genomic resources in these species, 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 the eight progenitor strains, allowing for accurate imputation from sites directly captured by GBS to millions of additional SNPs.

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; Peterson et al. 2012; Poland et al. 2012; Parker et al. 2016) as templates, we redesigned our GBS approach 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 inhouse 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 appears to degrade faster than mouse genomic DNA following extraction; 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 We used the *restrict* function in EMBOSS (version 6.6.0) (Rice, Longden, and Bleasby 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. The final choice of enzyme (NlaIII) was determined empirically and is detailed in the Results.

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 1; 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, NlaIII, BfaI, and MluCI were selected for further testing.

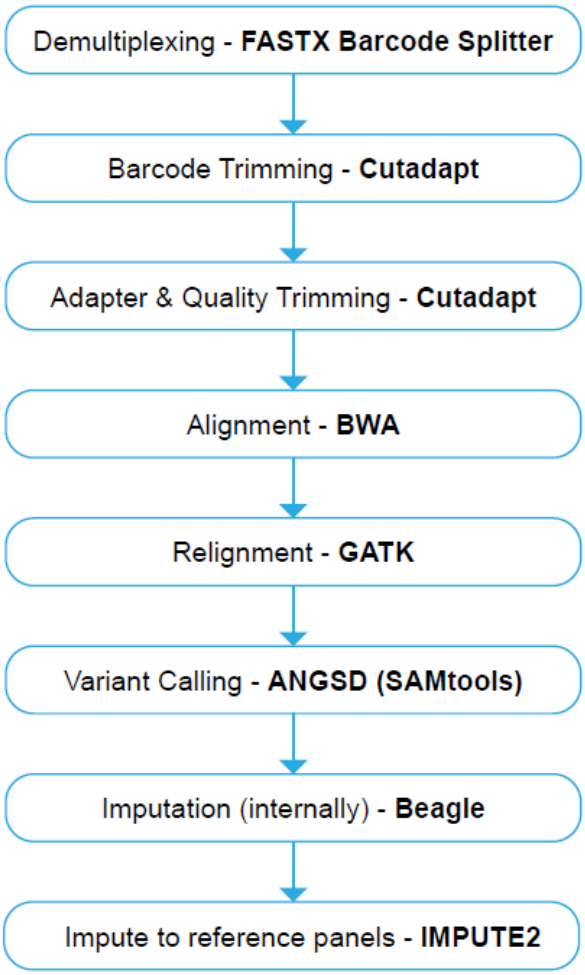
ddGBS library preparation and sequencing

The full ddGBS protocol is available in File S1. In brief, approximately 1µg of DNA is used per sample. Sample DNA, PstI barcoded adapters, and NlaIII Y-adapter are combined in a 96-well plate and allowed to evaporate at 37ºC overnight. Sample DNA and adapters are re-eluted on day two with a PstI/NlaIII digestion mix and incubated at 37ºC for two hours to allow for complete digestion. Ligation reagents are then added and incubated at 16ºC for one hour to anneal the adapters to the DNA fragments, followed by a 30-minute incubation at 80ºC to inactivate the restriction enzymes. Sample libraries are 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 are quantified in duplicate with Quanit-iT PicoGreen (Thermo Fisher Scientific, Waltham, MA) and pooled to the desired level of multiplexing (i.e. 12, 24, or 48 samples per library). Pooled libraries are concentrated to obtain the desired volume for use in the Pippin Prep. The concentrated pool is quantified to ensure the gel cassette will not be overloaded with DNA (>5µg). The pool is 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 increase the quantity of DNA, concentrated, 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.

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 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).

158 **Figure 2. ddGBS sequencing data analysis workflow.** Each step of the workflow is described
159 in the text.



160
161

162 **Evaluation of ddGBS pipeline performance**

163 We present the steps required to call and impute genotypes from raw ddGBS sequencing data in
164 Figure 2. During optimization of the pipeline, performance was assessed by two primary metrics:
165 (1) the number of variants called and (2) genotype concordance rates for calls made in 96 HS rats
166 that had both ddGBS genotypes and array genotypes from a custom Affymetrix Axiom MiRat
167 625k microarray (Part#: 550572). There were two checkpoints in the GBS pipeline where genotype
168 quality (measured by concordance rate) was assessed: after internal imputation within Beagle
169 (Browning and Browning 2009, 2016) and again after imputation to the reference panel with

IMPUTE2 (B. N. Howie, Donnelly, and Marchini 2009; B. Howie et al. 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.

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. Data concerning demultiplexing are shown in Table S1 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] (H. 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 (Korneliussen, Albrechtsen, and Nielsen 2014; Durvasula et al. n.d.). Further, using ANGSD, 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. . It must be noted that before settling on the combination of ANGSD and Beagle for genotype calling and internal imputation, we also experimented with GATK's UnifiedGenotyper and HaplotypeCaller (McKenna et al. 2010) with various parameter settings, but with unsatisfactory results.

Quality Control for genotypes before imputation using and 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. 2018). Further, to improve imputation accuracy and computational efficiency, we employed a pre-phasing step with IMPUTE2 (*prephase_g*) (B. Howie et al. 2012) prior to imputation. 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. the Medical College of Wisconsin (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

Genotype data will be available at https://figshare.com/articles/Heterogeneous_Stock_Genotype_Data/8243222 and the code necessary to run the steps outline in the publication are provided at https://figshare.com/articles/ddGBS_Pipeline_Commands/8243156. Supplementary Files are available at https://figshare.com/articles/Supplementary_Files/8243129. Additional data is available upon request.

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), which do not provide sufficient reads to make confident genotype calls (< 5 reads per sample) and are 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, and because restriction enzymes make the consistent cuts across samples, ensure the same fragments are sequenced in the majority of 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 1; 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 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 incomplete capture of the genome (Flanagan and Jones 2018).

The number of fragments with one of each of the cut sites were summed for all observed lengths and the results summarized in Figure 1 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. We ruled out BfaI because it only had a 2bp overhang after cleavage, which led to a high concentration of adapter dimer in the sequencing libraries (S5 File). NlaIII was chosen because it contained the greatest portion of the genome in the size selection window.

Table 1. Restriction enzyme options for double digest.

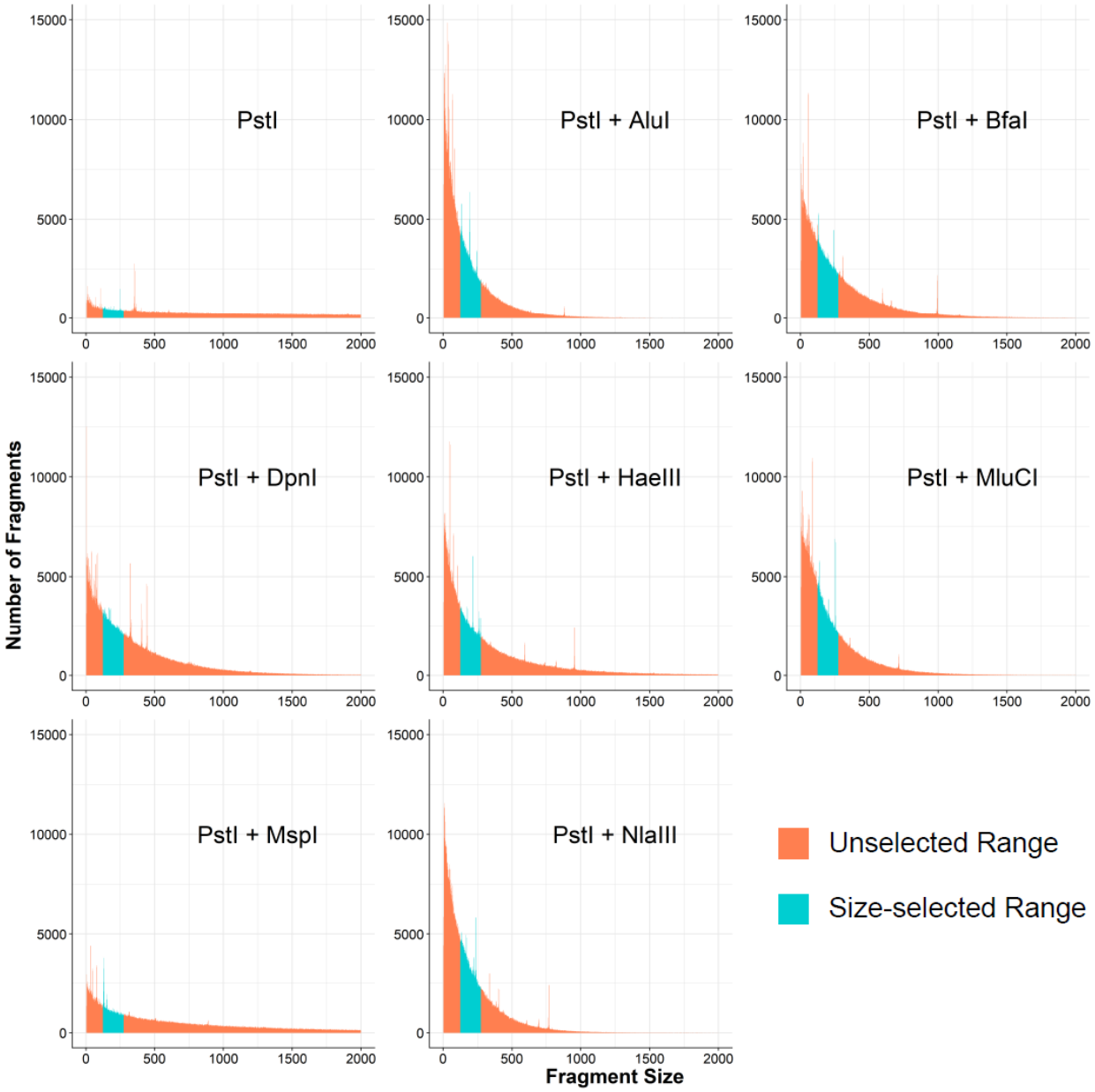
Restriction Enzyme(s)	Recognition sequence	Length of Overhang (bp)	% Genome in 250-400bp Region[+]	% Genome in 300-450bp Region[+]
PstI	CTGCA^G	4	0.48%	0.56%
PstI + AluI	AG^CT	0	3.06%	2.88%
PstI + BfaI	C^TAG	2	3.10%	3.25%
PstI + DpnI*	GA^TC	0	2.69%	3.00%
PstI + HaeIII	GG^CC	0	2.71%	2.79%
PstI + MluCI	^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 1. *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

351 portion of the reads contained sequence fragments of the NlaIII adapter sequence, indicating there
352 were fragments with insert sizes smaller than 125bps in the final library. To avoid this, we
353 increased the fragment size range to 300-450bps (Table 1), which corresponds to a 175-325bp
354 insert size once the adapters and primers are accounted for. Due to the high concentrations of our
355 libraries after pooling, the library size distribution obtained from the Pippin Prep was uniformly
356 shifted towards higher fragment lengths (Figure S3).

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

360 **Demultiplexing**

361 The number of base pairs of sequencing data retained after demultiplexing was fairly consistent
362 across demultiplexing software (Table S1). We ultimately decided to use FASTX Barcode Splitter
363 because it yielded the greatest number of reads after quality/adapter trimming and had faster run
364 times. An average of 330 million 100bp reads were obtained per library, resulting in ~7 million
365 reads per sample. Figure S4 shows the distribution of reads counts for all samples after
366 demultiplexing.

367 **Adapter and quality trimming**

368 Read quality was substantially improved after trimming the barcode and adapter sequences and
369 low-quality base pairs at the ends of reads (Figure S5). Overall read counts were only marginally
370 reduced by quality trimming (Table S1). We observed that the number of called variant sites and
371 the genotyping rate were both greater when using reads initially processed by cutadapt (Martin,
372 2011) than reads processed by the FASTX_Toolkit (Table S2). Importantly, a large portion of the

373 additional identified variants were known variant sites from the 42 inbred strains reference set
374 (Figure S6), indicating the elevated call rate was at least in part due to capturing more true variant
375 sites. We viewed this as sufficient support for proceeding with cutadapt for adapter removal and
376 quality trimming.

377 **Mapping quality**

378 The number of called variants and genotype call rates were identical at read mapping quality
379 (mapQ) thresholds of either 20 or 30 (Table S3) within ANGSD. As the ANGSD mapQ threshold
380 was raised to 45, there was a small reduction in the number of called variants, and then much
381 greater losses at thresholds of 60 or 90. Fortunately, genotype concordance rates at both low and
382 high mapQ thresholds were stable, despite the putatively higher quality of the alignments (Figure
383 S7). This permitted us to select a lower mapQ threshold (mapQ = 20), maximizing the number of
384 variants called without sacrificing genotyping accuracy.

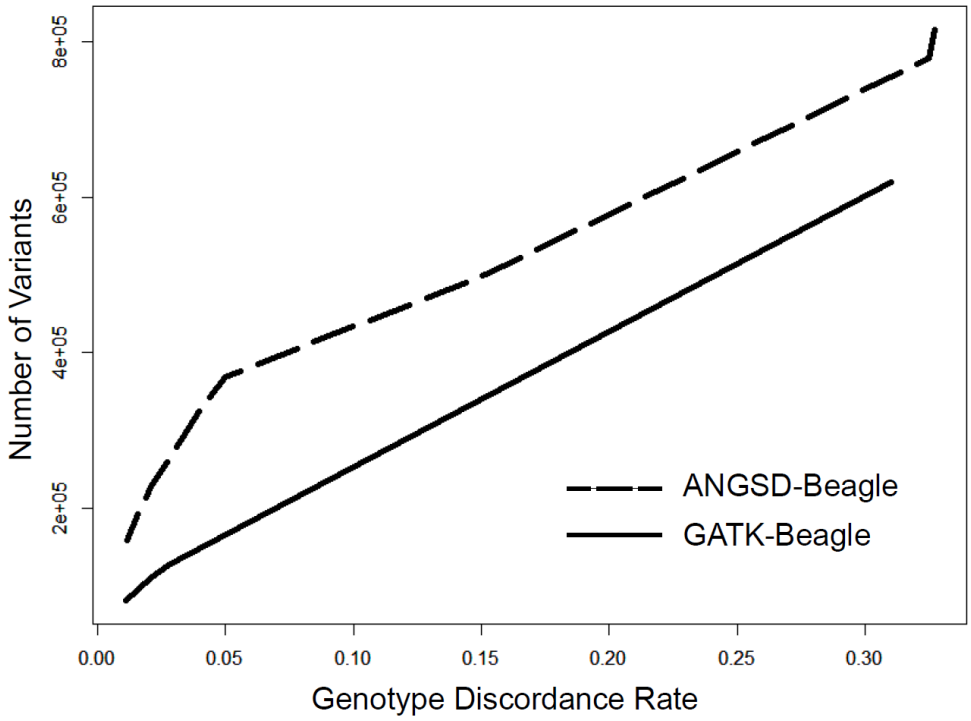
385 **Variant calling**

386 Figure 3 shows that across all levels of genotype discordance rates (with the array genotyping
387 data), the combination of the ANGSD (*samtools* model) with BEAGLE produced more SNPs, at
388 various genotyping concordance thresholds, than GATK's HaplotypeCaller (McKenna et al. 2010;
389 DePristo et al. 2011). This observation held when variants were limited only to biallelic sites and
390 SNPs with an MAF > 0.05 (Figure S8).

391

392

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



The figure compares the number variants called by combination of ANGSD 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. (A) 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. (B) Additional filters were applied to the original SNP sets and the plot zooms in on a smaller range of acceptable discordance rates. 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.

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

411 with sample size. With greater than 2,000 samples, we were unable to allocate sufficient memory
412 for the SOAPsnp model to successfully run, even after dividing the chromosomes into several,
413 smaller chunks. The marginal benefits of SOAPsnp in accuracy and number of variants were far
414 outweighed by its limitations when applied to a large sample set. The GATK model showed a
415 greater number of variants for more lenient genotype discordance rate threshold, but as stringency
416 increased, the number of variants converged across the remaining 3 models. We proceeded with
417 the SAMtools model for genotype likelihood estimation due to its previous support in the GBS
418 literature (Torkamaneh et al. 2017), accepting a nominal decrease in highly concordant variants
419 (Figure S9) for a large reduction in run time and memory usage.

420 **Imputation to reference panel**

421 Imputation is use in two ways in our protocol. As described above, we use imputation to assign
422 missing genotypes at SNPs called in only a subset of individuals. In addition, we use imputation
423 in this section to call genotypes at sites where GBS that were inaccessible to ddGBS sequencing.
424 Thus, our second application (described here) is similar to the human genetics application in which
425 imputation using 1000 Genomes increases the number of SNPs beyond those included on a given
426 microarray platform.

427 Before starting this imputation step, we observed an inflated transition/transversion ratio
428 (Table S4) in our ANGSD/Beagle SNPs. This issue was ameliorated when the SNP set was filtered
429 for only "known" variants that were previously identified in either the 42 inbred strains (Hermsen
430 et al. 2015) or the 8 deep-sequenced HS founders (Ramdas et al. 2018). For imputation, we
431 therefore only provided IMPUTE2 with previously identified variant sites from our
432 ANGSD/Beagle output. Prior to running IMPUTE2, we also filtered the variants for biallelic sites
433 with a genotype call in at least two individuals. Using pedigree data for the HS rats, we further

434 removed samples showing an order of magnitude higher level of Mendelian error than the sample
435 mean. We further removed SNPs that had an error rate surpassing a threshold of ~0.005 (Figure
436 S10; inflection point). There were 4 samples and 4,179 SNPs removed from subsequent analyses.
437 Lastly, we removed any samples where the sex chromosome read ratio was incompatible with their
438 reported sex (Figure S1).

439 To determine which reference set to use for imputation, we tested six different possible
440 combinations of available reference data (Table 2). The most accurate imputation was observed
441 for the reference set containing only the 8 deep-sequenced HS founder strains (Ramdas et al. 2018);
442 however, imputation to this set had the lowest genotyping rate of all panels. In contrast, using the
443 42 rat inbred strains displayed a balance of high accuracy and low missingness, leading us to
444 choose this as our reference set. To better understand the role of the 8 founder strains, which were
445 part of the 42 strains reference panel, we created a reference panel that included only the 34 non-
446 HS founder strains. As expected, discordance rates were much higher when only considering non-
447 founders. However, the genotype missingness was lower for the 34 than the 8 founders alone,
448 suggesting a combination of the two was the optimal set.

449

450

451

452

453

454

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

The table includes six 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. 2018 (Ramdas et al. 2018). 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 map 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 wide-spread 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; Gonzales et al. 2017; Zhou 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), which are polyploid and have differing levels of variation and LD than outbred rodent populations. Additionally, there were elements of each pipeline that did not meet our needs or lacked customizability. For instance, TASSEL-GBS v2 (Glaubitz et al. 2014) trims all reads to 92 base pairs; however, other projects underway in our lab utilized up to 125bp reads, leading to a ~20% reduction in data. TASSEL-GBS also ignores read base quality scores, which are informative in probabilistic frameworks for estimating uncertainty in alignments and variant calls (H. Li, Ruan, and Durbin 2008; DePristo et al. 2011; Nielsen et al. 2011), and uses a naïve binomial likelihood ratio method for calling SNPs. Stacks has previously shown poor performance in demultiplexing (Herten et al. 2015; Torkamaneh et al. 2017) and does not make use of the reference genome for priors when calling SNPs (Catchen et al. 2013). Fast-GBS relies on Platypus (Rimmer et al. 2014) for variant calling (WGS500 Consortium et al. 2014; Torkamaneh et al. 2017), which employs a Bayesian method of constructing candidate haplotypes that works poorly with low-pass sequencing data and does not scale well to large sample sizes (Z. Li, Wang, and Wang 2018). Lastly, none of these pipelines included an imputation step, which is crucial for filling in missing genotypes in GBS data and can provide millions of additional SNPs given an appropriate composite reference panel (B. Howie, Marchini, and Stephens 2011; G.-H. Huang and Tseng 2014).

Though we have not explicitly tested each alternate GBS pipeline for the purposes of this publication, this has been recently done by Wickland et 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

513 genotype likelihood model for calling SNPs (H. Li 2011). The combination of bwa mem and
514 SAMtools algorithm was independently shown to have the best performance for calling SNPs from
515 Illumina data (Hwang et al. 2015), further supporting our choice of these programs for read
516 alignment and variant calling. Additionally, using the ANGSD wrapper provided us with the
517 ability to convert the posterior genotype probabilities into genotype dosages for mapping studies
518 (Korneliussen, Albrechtsen, and Nielsen 2014).

519 A minor difference between GB-eaSy and our pipeline is the use of cutadapt (Martin 2011)
520 rather than GBSX (Herten et al. 2015) for demultiplexing, though both performed equally well
521 (Table S1). The primary improvement is our extension of the pipeline with the implementation of
522 effective internal and reference-based imputation steps using the 42 inbred rat genomes (Hermsen
523 et al. 2015) and 8 deep-sequenced HS founders from UMich (Ramdas et al. 2018). There are two
524 stages of imputation in our pipeline: the first one is accomplished by Beagle and aims to fill in
525 missing genotypes at called variants using information from other samples; this raising the
526 genotype call rate to 100%, but it may also introduce errors due to insufficient information,
527 emphasizing the need for careful filtering steps. The second stage of imputation made use of
528 IMPTUE2 and an external reference panels of variants called from WGS data on the 8 inbred HS
529 founders, as well as 34 additional inbred rat strains. We decided to include the 34 additional strains
530 because of the elevated genotyping rate we observed upon their inclusion in the IMPUTE2
531 reference panel. We attribute this to the presence of haplotypes that exist in both the 8 the HS
532 founder strains and a subset of the 34 additional strains in this panel. The benefits of using a
533 composite reference panel have been previously noted (Zhang et al. 2013; G.-H. Huang and Tseng
534 2014); there is increased accuracy and decreased missingness in the imputed genotype data.

In summary, we have adapted a GBS protocol and genotyping and imputation pipeline to obtain dense genotypes on genome-wide markers in highly-multiplexed HS rats. After quality filtering on the level of SNP and sample, over 3.7 million were called with a concordance rate of 99%. The ddGBS protocol and bioinformatic methods used to produce this data are publicly available, easy to handle, and cost-effective. The presented workflow could be feasibly followed with marginal modifications for application in other species.

554 ACKNOWLEDGEMENTS:

555 This work was supported by P50 DA037844, R21 DA036672, T32 GM07197, and

556 F31 DA039638.

LITERATURE CITED

Aird, Daniel, Michael G Ross, Wei-Sheng Chen, Maxwell Danielsson, Timothy Fennell, Carsten Russ, David B Jaffe, Chad Nusbaum, and Andreas Gnirke. 2011. "Analyzing and Minimizing PCR Amplification Bias in Illumina Sequencing Libraries." *Genome Biology* 12 (2): R18. https://doi.org/10.1186/gb-2011-12-2-r18.

Andrews, Simon. 2017. *FastQC* (version 0.11.6). http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Baird, Nathan A., Paul D. Etter, Tressa S. Atwood, Mark C. Currey, Anthony L. Shiver, Zachary A. Lewis, Eric U. Selker, William A. Cresko, and Eric A. Johnson. 2008. "Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers." Edited by Justin C. Fay. *PLoS ONE* 3 (10): e3376. https://doi.org/10.1371/journal.pone.0003376.

Browning, Brian L., and Sharon 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 (2): 210–23. https://doi.org/10.1016/j.ajhg.2009.01.005.

Browning, Brian L., and Sharon R. Browning. 2016. "Genotype Imputation with Millions of Reference Samples." *The American Journal of Human Genetics* 98 (1): 116–26. https://doi.org/10.1016/j.ajhg.2015.11.020.

Catchen, Julian, Paul A. Hohenlohe, Susan Bassham, Angel Amores, and William A. Cresko. 2013. "Stacks: An Analysis Tool Set for Population Genomics." *Molecular Ecology* 22 (11): 3124–40. https://doi.org/10.1111/mec.12354.

Chen, Qiang, Yufang Ma, Yumei Yang, Zhenliang Chen, Rongrong Liao, Xiaoxian Xie, Zhen Wang, et al. 2013. "Genotyping by Genome Reducing and Sequencing for Outbred Animals." Edited by Shuhong Zhao. *PLoS ONE* 8 (7): e67500. https://doi.org/10.1371/journal.pone.0067500.

Chitre, Apurva S, Oksana Polesskaya, Katie Holl, Jianjun Gao, Riyan Cheng, Angel Martinez, Tony George, et al. 2018. "Genome Wide Association Study of Body Weight, Body Mass Index, Adiposity, and Fasting Glucose in 3,173 Outbred Rats," September. https://doi.org/10.1101/422428.

Davey, John W., Paul A. Hohenlohe, Paul D. Etter, Jason Q. Boone, Julian M. Catchen, and Mark L. Blaxter. 2011. "Genome-Wide Genetic Marker Discovery and Genotyping Using next-Generation Sequencing." *Nature Reviews Genetics* 12 (7): 499–510. https://doi.org/10.1038/nrg3012.

De Donato, Marcos, Sunday O. Peters, Sharon E. Mitchell, Tanveer Hussain, and Ikhide G. Imumorin. 2013. "Genotyping-by-Sequencing (GBS): A Novel, Efficient and Cost-Effective Genotyping Method for Cattle Using Next-Generation Sequencing." Edited by James C. Nelson. *PLoS ONE* 8 (5): e62137. https://doi.org/10.1371/journal.pone.0062137.

DePristo, Mark A, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony A Philippakis, et al. 2011. "A Framework for Variation Discovery and Genotyping Using Next-Generation DNA Sequencing Data." *Nature Genetics* 43 (5): 491–98. https://doi.org/10.1038/ng.806.

Durvasula, Arun, Paul J Hoffman, Tyler V Kent, Chaochih Liu, Thomas J Y Kono, Peter L Morrell, and Jeffrey Ross-Ibarra. n.d. "ANGSD-Wrapper: Utilities for Analyzing next

601 Generation Sequencing Data." Accessed September 5, 2018.
602 https://doi.org/10.7287/peerj.preprints.1472v2.
603 Elshire, Robert J., Jeffrey C. Glaubitz, Qi Sun, Jesse A. Poland, Ken Kawamoto, Edward S.
604 Buckler, and Sharon E. Mitchell. 2011. "A Robust, Simple Genotyping-by-Sequencing
605 (GBS) Approach for High Diversity Species." Edited by Laszlo Orban. *PLoS ONE* 6 (5):
606 e19379. https://doi.org/10.1371/journal.pone.0019379.
607 Fitzpatrick, Christopher J., Shyam Gopalakrishnan, Elizabeth S. Cogan, Lindsay M. Yager, Paul
608 J. Meyer, Vedran Lovic, Benjamin T. Saunders, et al. 2013. "Variation in the Form of
609 Pavlovian Conditioned Approach Behavior among Outbred Male Sprague-Dawley Rats
610 from Different Vendors and Colonies: Sign-Tracking vs. Goal-Tracking." Edited by
611 Patrizia Campolongo. *PLoS ONE* 8 (10): e75042.
612 https://doi.org/10.1371/journal.pone.0075042.
613 Flanagan, Sarah P., and Adam G. Jones. 2018. "Substantial Differences in Bias between Single-
614 Digest and Double-Digest RAD-Seq Libraries: A Case Study." *Molecular Ecology*
615 *Resources* 18 (2): 264–80. https://doi.org/10.1111/1755-0998.12734.
616 Fu, Yong-Bi, and Mo-Hua Yang. 2017. "Genotyping-by-Sequencing and Its Application to Oat
617 Genomic Research." In *Oat*, edited by Sebastian Gasparis, 1536:169–87. New York, NY:
618 Springer New York. https://doi.org/10.1007/978-1-4939-6682-0_13.
619 Furuta, Tomoyuki, Motoyuki Ashikari, Kshirod K. Jena, Kazuyuki Doi, and Stefan Reuscher.
620 2017. "Adapting Genotyping-by-Sequencing for Rice F2 Populations." *G3:*
621 *Genes|Genomes|Genetics* 7 (3): 881–93. https://doi.org/10.1534/g3.116.038190.
622 Gileta, Alexander F., Christopher J. Fitzpatrick, Apurva S. Chitre, Celine L. St. Pierre, Elizabeth
623 V. Joyce, Rachael J. Maguire, Africa M. McLeod, et al. 2018. "Genetic Characterization
624 of Outbred Sprague Dawley Rats and Utility for Genome-Wide Association Studies,"
625 September. https://doi.org/10.1101/412924.
626 Glaubitz, Jeffrey C., Terry M. Casstevens, Fei Lu, James Harriman, Robert J. Elshire, Qi Sun,
627 and Edward S. Buckler. 2014. "TASSEL-GBS: A High Capacity Genotyping by
628 Sequencing Analysis Pipeline." Edited by Nicholas A. Tinker. *PLoS ONE* 9 (2): e90346.
629 https://doi.org/10.1371/journal.pone.0090346.
630 Gonzales, Natalia M., Jungkyun Seo, Ana Isabel Hernandez-Cordero, Celine L. St. Pierre,
631 Jennifer S. Gregory, Margaret G. Distler, Mark Abney, Stefan Canzar, Arimantas
632 Lionikas, and Abraham A. Palmer. 2017. "Genome Wide Association Study of
633 Behavioral, Physiological and Gene Expression Traits in a Multigenerational Mouse
634 Intercross," December. https://doi.org/10.1101/230920.
635 ———. 2018. "Genome Wide Association Analysis in a Mouse Advanced Intercross Line,"
636 September. https://doi.org/10.1101/230920.
637 Hannon Lab. 2010. *FASTX-Toolkit* (version 0.0.13).
638 http://hannonlab.cshl.edu/fastx_toolkit/index.html.
639 He, Jiangfeng, Xiaoqing Zhao, AndrÃ© Laroche, Zhen-Xiang Lu, HongKui Liu, and Ziqin Li.
640 2014. "Genotyping-by-Sequencing (GBS), an Ultimate Marker-Assisted Selection
641 (MAS) Tool to Accelerate Plant Breeding." *Frontiers in Plant Science* 5 (September).
642 https://doi.org/10.3389/fpls.2014.00484.
643 Hermsen, Roel, Joep de Ligt, Wim Spee, Francis Blokzijl, Sebastian Schäfer, Eleonora Adami,
644 Sander Boymans, et al. 2015. "Genomic Landscape of Rat Strain and Substrain
645 Variation." *BMC Genomics* 16 (1). https://doi.org/10.1186/s12864-015-1594-1.

Herten, Koen, Matthew S Hestand, Joris R Vermeesch, and Jeroen KJ Van Houdt. 2015. "GBSX: A Toolkit for Experimental Design and Demultiplexing Genotyping by Sequencing Experiments." *BMC Bioinformatics* 16 (1). https://doi.org/10.1186/s12859-015-0514-3.

Howie, Bryan, Christian Fuchsberger, Matthew Stephens, Jonathan Marchini, and Gonçalo R Abecasis. 2012. "Fast and Accurate Genotype Imputation in Genome-Wide Association Studies through Pre-Phasing." *Nature Genetics* 44 (July): 955.

Howie, Bryan, Jonathan Marchini, and Matthew Stephens. 2011. "Genotype Imputation with Thousands of Genomes." *G3: Genes|Genomes|Genetics* 1 (6): 457–70. https://doi.org/10.1534/g3.111.001198.

Howie, Bryan N., Peter Donnelly, and Jonathan Marchini. 2009. "A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies." Edited by Nicholas J. Schork. *PLoS Genetics* 5 (6): e1000529. https://doi.org/10.1371/journal.pgen.1000529.

Huang, Guan-Hua, and Yi-Chi Tseng. 2014. "Genotype Imputation Accuracy with Different Reference Panels in Admixed Populations." *BMC Proceedings* 8 (Suppl 1): S64. https://doi.org/10.1186/1753-6561-8-S1-S64.

Huang, X., Q. Feng, Q. Qian, Q. Zhao, L. Wang, A. Wang, J. Guan, et al. 2009. "High-Throughput Genotyping by Whole-Genome Resequencing." *Genome Research* 19 (6): 1068–76. https://doi.org/10.1101/gr.089516.108.

Hwang, Sohyun, Eiru Kim, Insuk Lee, and Edward M. Marcotte. 2015. "Systematic Comparison of Variant Calling Pipelines Using Gold Standard Personal Exome Variants." *Scientific Reports* 5 (December): 17875.

Illumina, Inc. 2014. "Nextera(R) Library Validation and Cluster Density Optimization: Guidelines for Generating High-Quality Data with Nextera Library Prepartion Kits." https://www.illumina.com/documents/products/technotes/technote_nextera_library_validation.pdf.

Jensen-Seaman, M. I. 2004. "Comparative Recombination Rates in the Rat, Mouse, and Human Genomes." *Genome Research* 14 (4): 528–38. https://doi.org/10.1101/gr.1970304.

Johannesson, M., R. Lopez-Aumatell, P. Stridh, M. Diez, J. Tuncel, G. Blazquez, E. Martinez-Membrives, 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 (1): 150–58. https://doi.org/10.1101/gr.081497.108.

Johnson, Jennifer L., Helena Wittgenstein, Sharon E. Mitchell, Katie E. Hyma, Svetlana V. Temnykh, Anastasiya V. Kharlamova, Rimma G. Gulevich, et al. 2015. "Genotyping-By-Sequencing (GBS) Detects Genetic Structure and Confirms Behavioral QTL in Tame and Aggressive Foxes (Vulpes Vulpes)." Edited by William J. Murphy. *PLOS ONE* 10 (6): e0127013. https://doi.org/10.1371/journal.pone.0127013.

Kanagawa, Takahiro. 2003. "Bias and Artifacts in Multitemplate Polymerase Chain Reactions (PCR)." *Journal of Bioscience and Bioengineering* 96 (4): 317–23. https://doi.org/10.1016/S1389-1723(03)90130-7.

Korneliussen, Thorfinn Sand, Anders Albrechtsen, and Rasmus Nielsen. 2014. "ANGSD: Analysis of Next Generation Sequencing Data." *BMC Bioinformatics* 15 (1). https://doi.org/10.1186/s12859-014-0356-4.

690 Li, H. 2011. "A Statistical Framework for SNP Calling, Mutation Discovery, Association
691 Mapping and Population Genetical Parameter Estimation from Sequencing Data."
692 *Bioinformatics* 27 (21): 2987–93. https://doi.org/10.1093/bioinformatics/btr509.
693 Li, H., and R. Durbin. 2009. "Fast and Accurate Short Read Alignment with Burrows-Wheeler
694 Transform." *Bioinformatics* 25 (14): 1754–60.
695 https://doi.org/10.1093/bioinformatics/btp324.
696 Li, H., J. Ruan, and R. Durbin. 2008. "Mapping Short DNA Sequencing Reads and Calling
697 Variants Using Mapping Quality Scores." *Genome Research* 18 (11): 1851–58.
698 https://doi.org/10.1101/gr.078212.108.
699 Li, Zhentang, Yi Wang, and Fei Wang. 2018. "A Study on Fast Calling Variants from Next-
700 Generation Sequencing Data Using Decision Tree." *BMC Bioinformatics* 19 (1).
701 https://doi.org/10.1186/s12859-018-2147-9.
702 Littrell, John, Shirng-Wern Tsaih, Amelie Baud, Pasi Rastas, Leah Solberg-Woods, and Michael
703 J. Flister. 2018. "A High-Resolution Genetic Map for the Laboratory Rat."
704 *G3: Genes|Genomes|Genetics*, May, g3.200187.2018.
705 https://doi.org/10.1534/g3.118.200187.
706 Martin, Marcel. 2011. "Cutadapt Removes Adapter Sequences from High-Throughput
707 Sequencing Reads." *EMBnet.Journal* 17 (1): 10. https://doi.org/10.14806/ej.17.1.200.
708 McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, et
709 al. 2010. "The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-
710 Generation DNA Sequencing Data." *Genome Research* 20 (9): 1297–1303.
711 https://doi.org/10.1101/gr.107524.110.
712 Miller, M. R., J. P. Dunham, A. Amores, W. A. Cresko, and E. A. Johnson. 2007. "Rapid and
713 Cost-Effective Polymorphism Identification and Genotyping Using Restriction Site
714 Associated DNA (RAD) Markers." *Genome Research* 17 (2): 240–48.
715 https://doi.org/10.1101/gr.5681207.
716 Nielsen, Rasmus, Joshua S. Paul, Anders Albrechtsen, and Yun S. Song. 2011. "Genotype and
717 SNP Calling from Next-Generation Sequencing Data." *Nature Reviews Genetics* 12 (6):
718 443–51. https://doi.org/10.1038/nrg2986.
719 Orsouw, Nathalie J. van, René C. J. Hogers, Antoine Janssen, Feyruz Yalcin, Sandor Snoeijers,
720 Esther Verstege, Harrie Schneiders, et al. 2007. "Complexity Reduction of Polymorphic
721 Sequences (CRoPS^{TM}): A Novel Approach for Large-Scale Polymorphism Discovery in
722 Complex Genomes." Edited by Ivan Baxter. *PLoS ONE* 2 (11): e1172.
723 https://doi.org/10.1371/journal.pone.0001172.
724 Parker, Clarissa C, Shyam Gopalakrishnan, Peter Carbonetto, Natalia M Gonzales, Emily Leung,
725 Yeonhee J Park, Emmanuel Aryee, et al. 2016. "Genome-Wide Association Study of
726 Behavioral, Physiological and Gene Expression Traits in Outbred CFW Mice." *Nature*
727 *Genetics* 48 (8): 919–26. https://doi.org/10.1038/ng.3609.
728 Pértille, Fábio, Carlos Guerrero-Bosagna, Vinicius Henrique da Silva, Clarissa Boschiero, José
729 de Ribamar da Silva Nunes, Mônica Corrêa Ledur, Per Jensen, and Luiz Lehmann
730 Coutinho. 2016. "High-Throughput and Cost-Effective Chicken Genotyping Using Next-
731 Generation Sequencing." *Scientific Reports* 6 (May): 26929.
732 Peterson, Brant K., Jesse N. Weber, Emily H. Kay, Heidi S. Fisher, and Hopi E. Hoekstra. 2012.
733 "Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and
734 Genotyping in Model and Non-Model Species." Edited by Ludovic Orlando. *PLoS ONE*
735 7 (5): e37135. https://doi.org/10.1371/journal.pone.0037135.

Poland, Jesse A., Patrick J. Brown, Mark E. Sorrells, and Jean-Luc Jannink. 2012. "Development of High-Density Genetic Maps for Barley and Wheat Using a Novel Two-Enzyme Genotyping-by-Sequencing Approach." Edited by Tongming Yin. *PLoS ONE* 7 (2): e32253. https://doi.org/10.1371/journal.pone.0032253.

Ramdas, Shweta, Ayse Bilge Ozel, Katie Holl, Myrna Mandel, Leah Solberg Woods, and Jun Z Li. 2018. "Extended Regions of Suspected Mis-Assembly in the Rat Reference Genome," September. https://doi.org/10.1101/332932.

Rat Genome Sequencing and Mapping Consortium, Amelie Baud, Roel Hermsen, Victor Guryev, Pernilla Stridh, Delyth Graham, Martin W McBride, et al. 2013. "Combined Sequence-Based and Genetic Mapping Analysis of Complex Traits in Outbred Rats." *Nature Genetics* 45 (7): 767–75. https://doi.org/10.1038/ng.2644.

Rice, Peter, Ian Longden, and Alan Bleasby. 2000. "EMBOSS: The European Molecular Biology Open Software Suite." *Trends in Genetics* 16 (6): 276–77. https://doi.org/10.1016/S0168-9525(00)02024-2.

Rimmer, Andy, Hang Phan, Iain Mathieson, Zamin Iqbal, Stephen R. F. Twigg, WGS500 Consortium, Andrew O. M. Wilkie, Gil McVean, and Gerton Lunter. 2014. "Integrating Mapping-, Assembly- and Haplotype-Based Approaches for Calling Variants in Clinical Sequencing Applications." *Nature Genetics* 46 (8): 912–18. https://doi.org/10.1038/ng.3036.

Roberts, R. J., and D. Macelis. 1999. "REBASE--Restriction Enzymes and Methylases." *Nucleic Acids Research* 27 (1): 312–13. https://doi.org/10.1093/nar/27.1.312.

Scheben, Armin, Jacqueline Batley, and David Edwards. 2017. "Genotyping-by-Sequencing Approaches to Characterize Crop Genomes: Choosing the Right Tool for the Right Application." *Plant Biotechnology Journal* 15 (2): 149–61. https://doi.org/10.1111/pbi.12645.

Sonah, Humira, Maxime Bastien, Elmer Iquira, Aurélie Tardivel, Gaétan Légaré, Brian Boyle, Éric Normandeau, et al. 2013. "An Improved Genotyping by Sequencing (GBS) Approach Offering Increased Versatility and Efficiency of SNP Discovery and Genotyping." Edited by Zhanjiang Liu. *PLoS ONE* 8 (1): e54603. https://doi.org/10.1371/journal.pone.0054603.

Steen, R. G., A. E. Kwitek-Black, C. Glenn, J. Gullings-Handley, W. Van Etten, O. S. Atkinson, D. Appel, et al. 1999. "A High-Density Integrated Genetic Linkage and Radiation Hybrid Map of the Laboratory Rat." *Genome Research* 9 (6): AP1-8, insert.

Sun, Xiaowen, Dongyuan Liu, Xiaofeng Zhang, Wenbin Li, Hui Liu, Weiguo Hong, Chuanbei Jiang, et al. 2013. "SLAF-Seq: An Efficient Method of Large-Scale De Novo SNP Discovery and Genotyping Using High-Throughput Sequencing." Edited by Jan Aerts. *PLoS ONE* 8 (3): e58700. https://doi.org/10.1371/journal.pone.0058700.

Torkamaneh, Davoud, Jérôme Laroche, Maxime Bastien, Amina Abed, and François Belzile. 2017. "Fast-GBS: A New Pipeline for the Efficient and Highly Accurate Calling of SNPs from Genotyping-by-Sequencing Data." *BMC Bioinformatics* 18 (1). https://doi.org/10.1186/s12859-016-1431-9.

Van Tassell, Curtis P, Timothy P L Smith, Lakshmi K Matukumalli, Jeremy F Taylor, Robert D Schnabel, Cynthia Taylor Lawley, Christian D Haudenschild, Stephen S Moore, Wesley C Warren, and Tad S Sonstegard. 2008. "SNP Discovery and Allele Frequency Estimation by Deep Sequencing of Reduced Representation Libraries." *Nature Methods* 5 (3): 247–52. https://doi.org/10.1038/nmeth.1185.

Wang, Yuzhe, Xuemin Cao, Yiqiang Zhao, Jing Fei, Xiaoxiang Hu, and Ning Li. 2017. "Optimized Double-Digest Genotyping by Sequencing (DdGBS) Method with High-Density SNP Markers and High Genotyping Accuracy for Chickens." Edited by Peng Xu. *PLOS ONE* 12 (6): e0179073. https://doi.org/10.1371/journal.pone.0179073.

WGS500 Consortium, Andy Rimmer, Hang Phan, Iain Mathieson, Zamin Iqbal, Stephen R F Twigg, Andrew O M Wilkie, Gil McVean, and Gerton Lunter. 2014. "Integrating Mapping-, Assembly- and Haplotype-Based Approaches for Calling Variants in Clinical Sequencing Applications." *Nature Genetics* 46 (8): 912–18. https://doi.org/10.1038/ng.3036.

Wickland, Daniel P., Gopal Battu, Karen A. Hudson, Brian W. Diers, and Matthew 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 (1). https://doi.org/10.1186/s12859-017-2000-6.

Woods, Leah C. Solberg, and Richard Mott. 2017. "Heterogeneous Stock Populations for Analysis of Complex Traits." In *Systems Genetics*, edited by Klaus Schughart and Robert W. Williams, 1488:31–44. New York, NY: Springer New York. https://doi.org/10.1007/978-1-4939-6427-7_2.

Zhang, Peng, Xiaowei Zhan, Noah A. Rosenberg, and Sebastian Zöllner. 2013. "Genotype Imputation Reference Panel Selection Using Maximal Phylogenetic Diversity." *Genetics* 195 (2): 319–30. https://doi.org/10.1534/genetics.113.154591.

Zhou, Xinzhu, Celine L. St. Pierre, Natalia M. Gonzales, Riyan Cheng, Apurva S. Chitre, Greta Sokoloff, and Abraham A. Palmer. 2018. "Genome-Wide Association Study, Replication, and Mega-Analysis Using a Dense Marker Panel in a Multi-Generational Mouse Advanced Intercross Line," August. https://doi.org/10.1101/387613.

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.

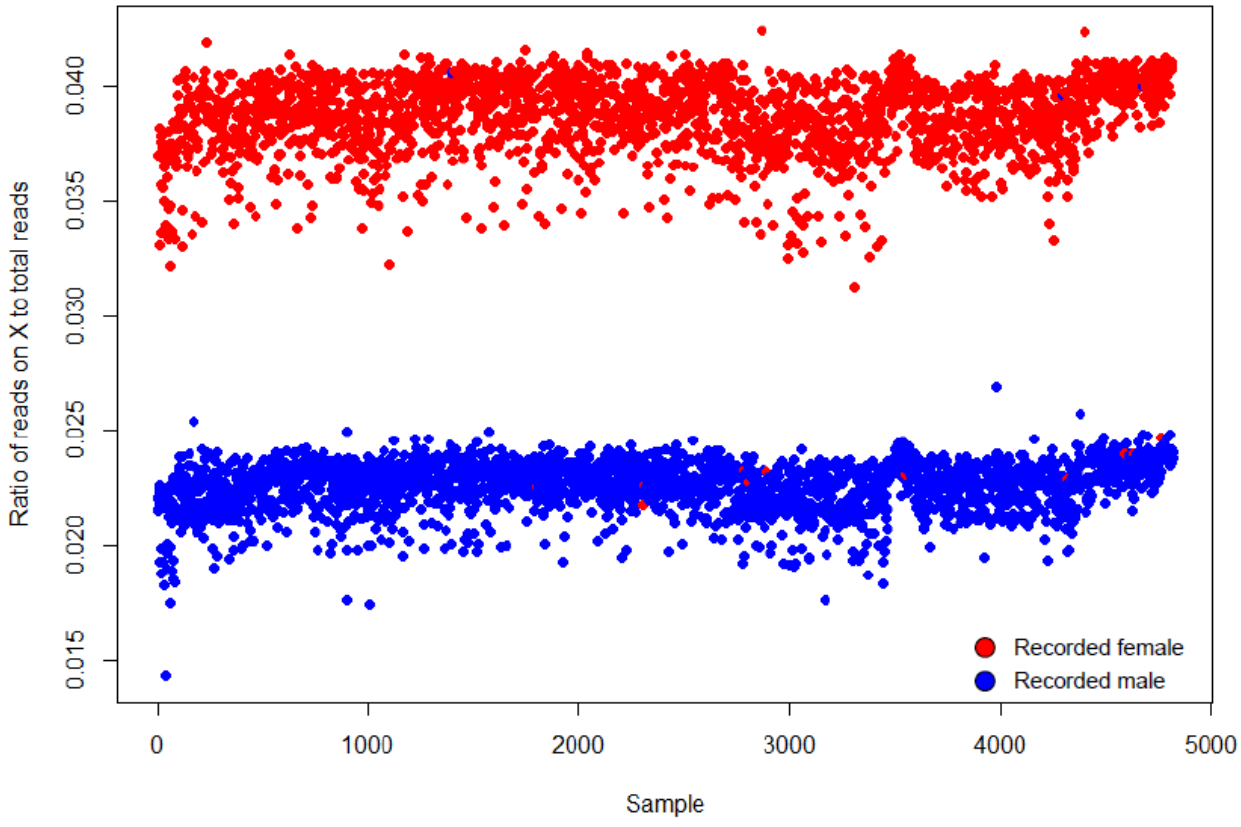


Figure S2. Data preparation workflow for imputation with IMPUTE2.

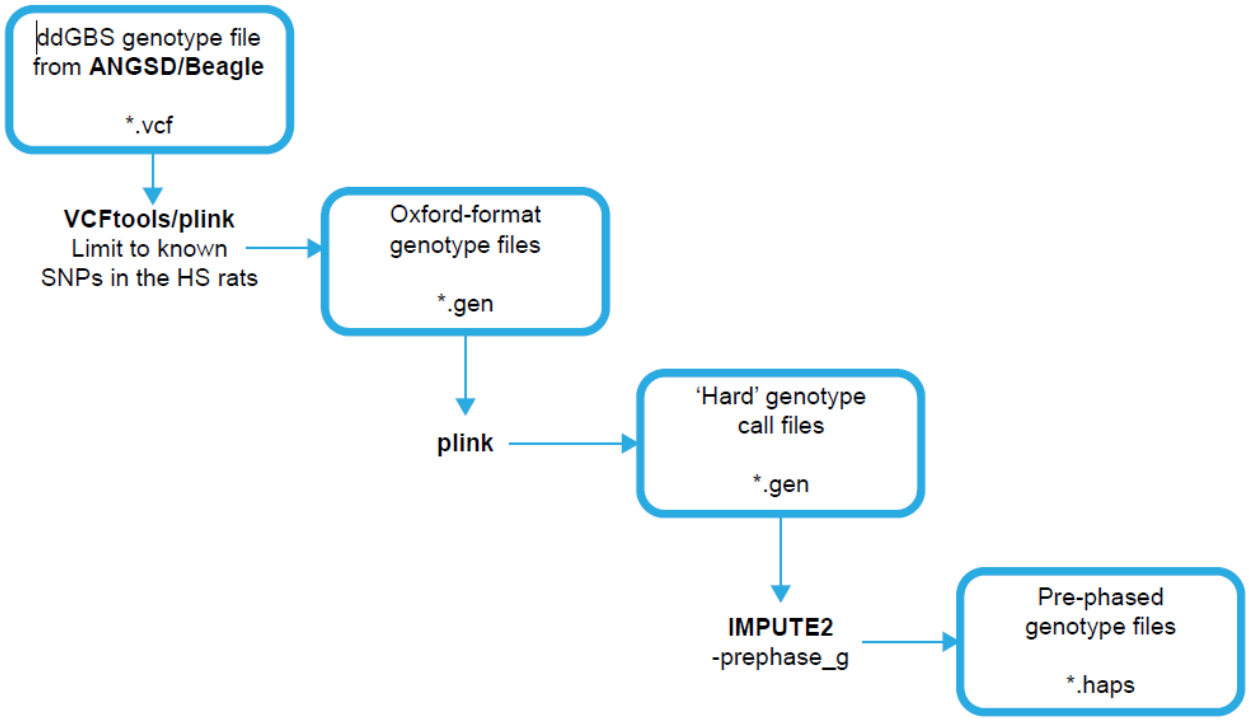


Figure S3. Programmed vs. empirical Pippin Prep fragment size range.

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.

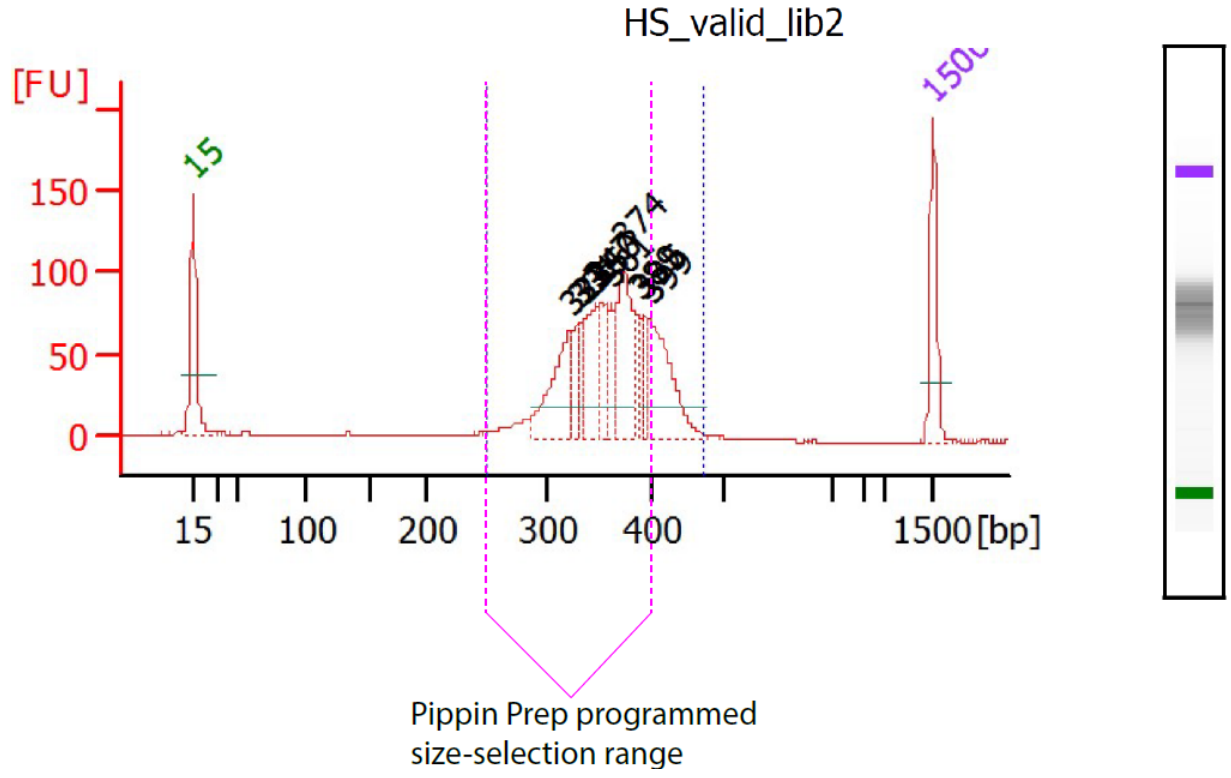


Figure S4. Raw read counts grouped by shipment batch.

Raw read counts are on a per-sample basis after demultiplexing FASTQ files with FASTX Barcode Splitter.

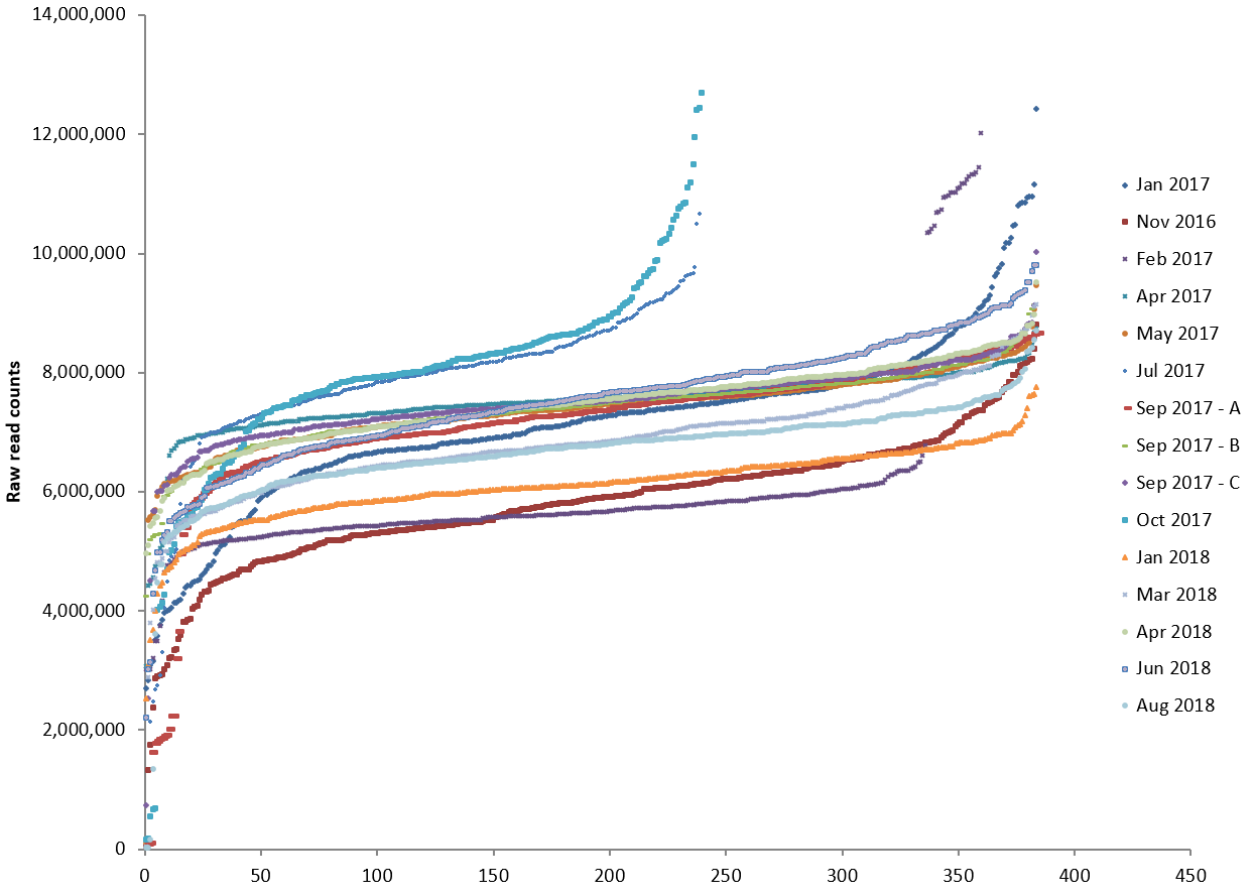


Figure S5. FASTQC results pre- and post-filtering with Cutadapt.

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.

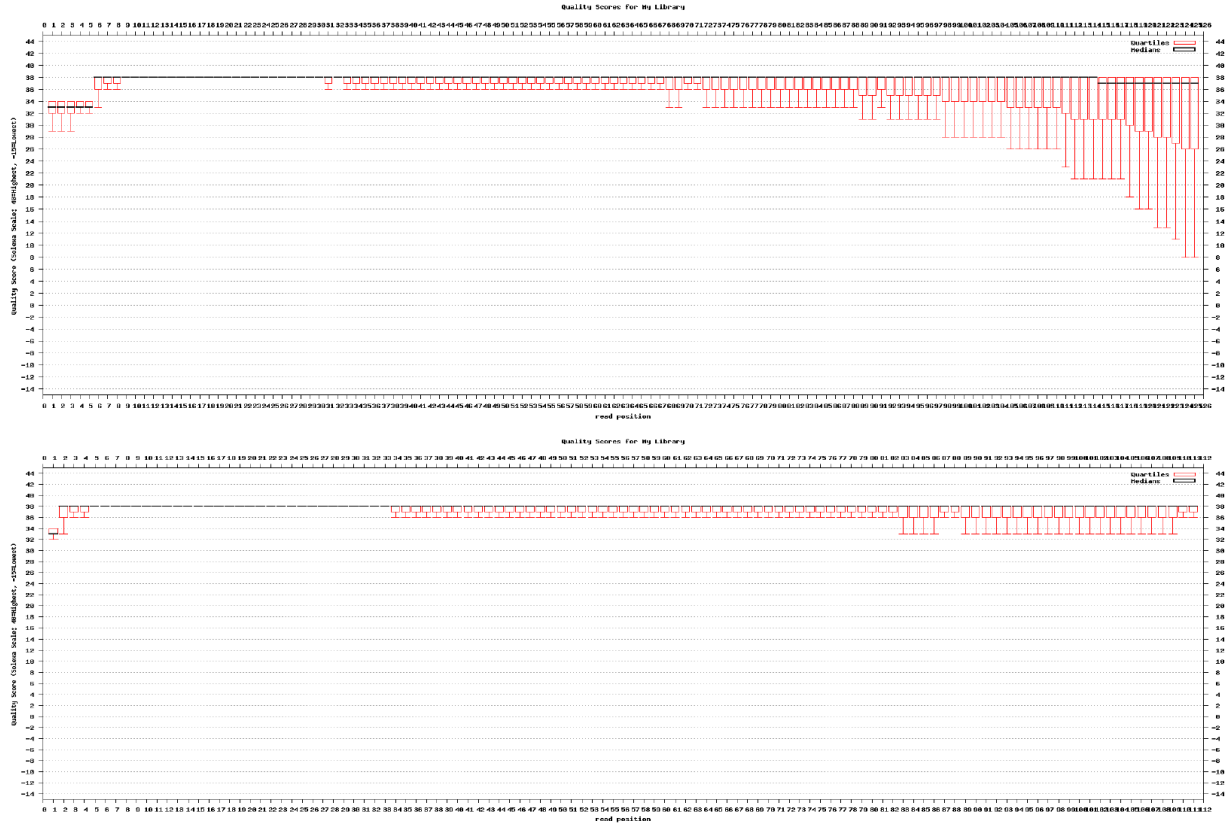


Figure S6. Overlap of called SNPs with known variants after read trimming with FASTX or Cutadapt.

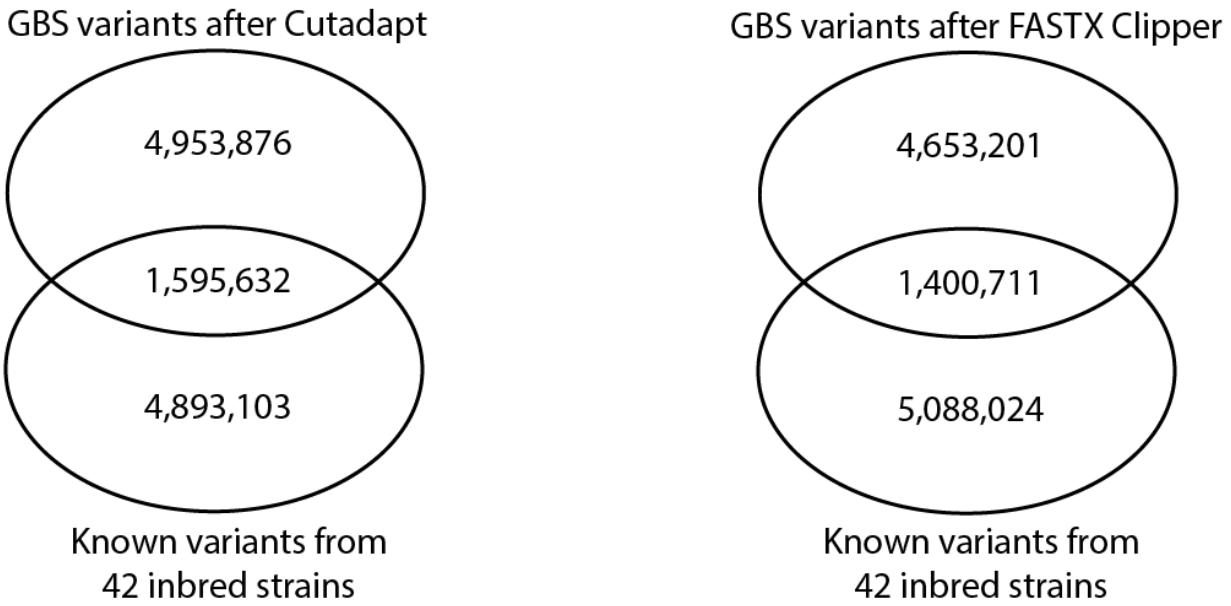


Figure S7. Mapping quality thresholds.
Genotyping error rate and number of variants by men depth per sample per variant site for mapping quality thresholds of 20, 30, and 60.

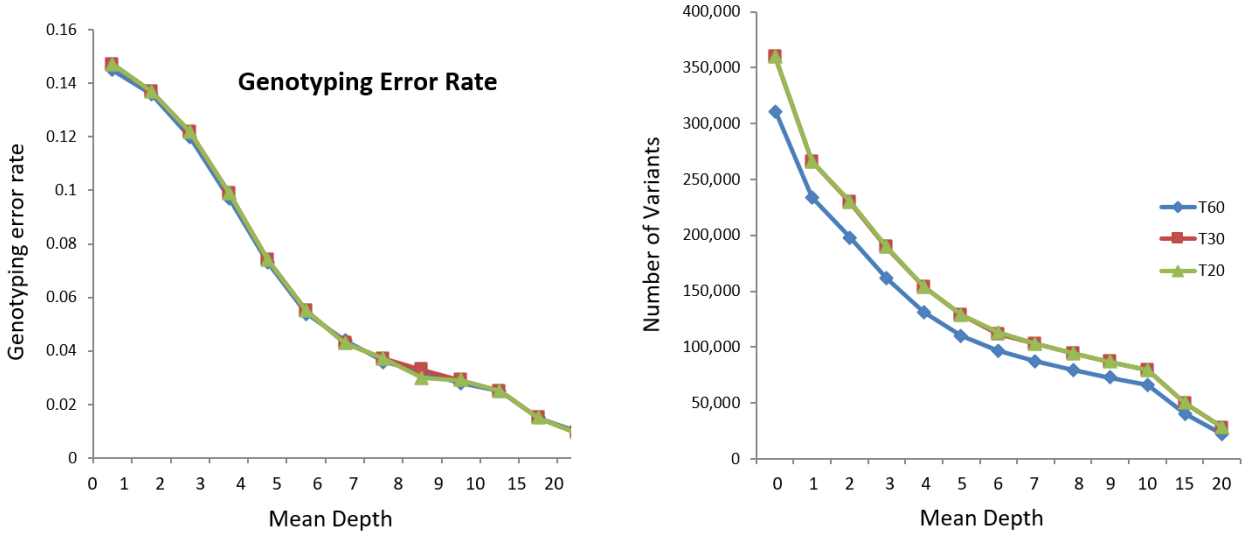


Figure S8. ANGSD vs GATK HaplotypeCaller, filtered calls.

The panel compares the number variants called by combination of ANGSD 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.

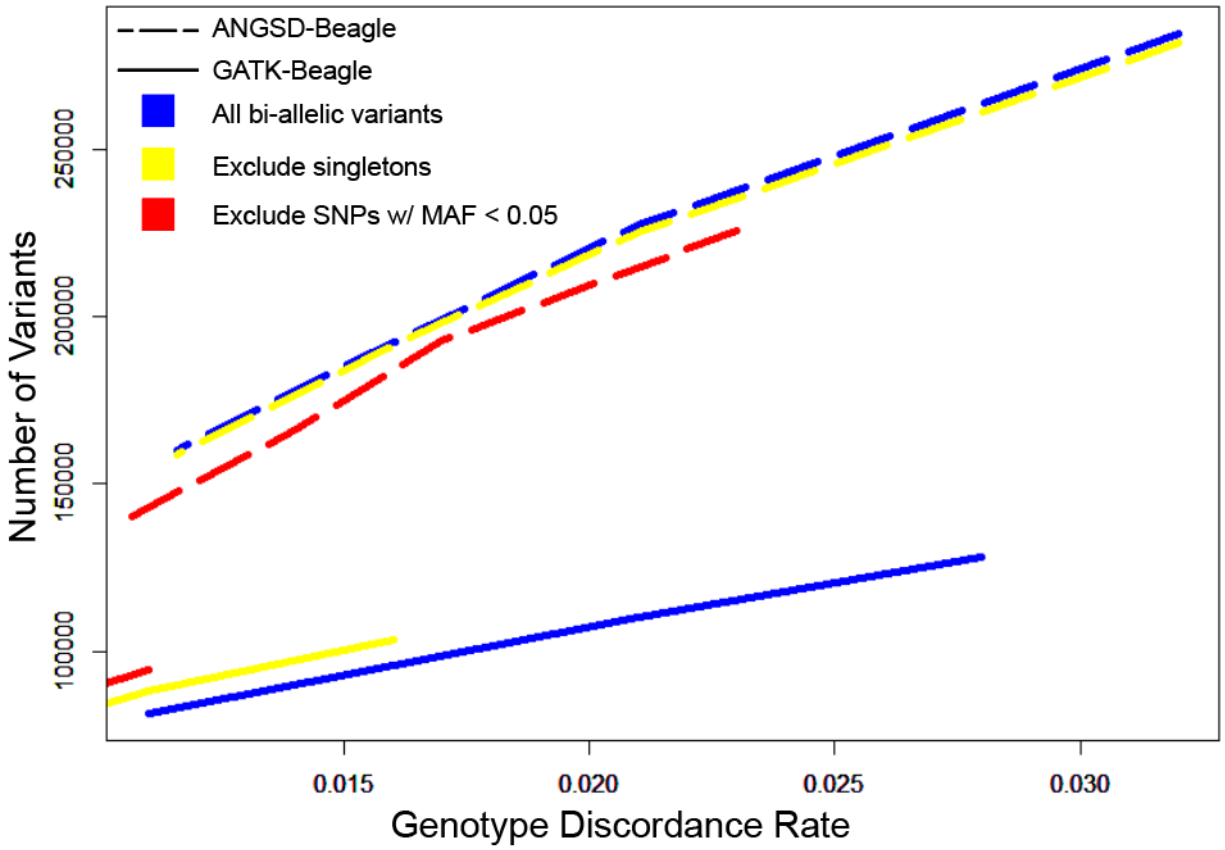
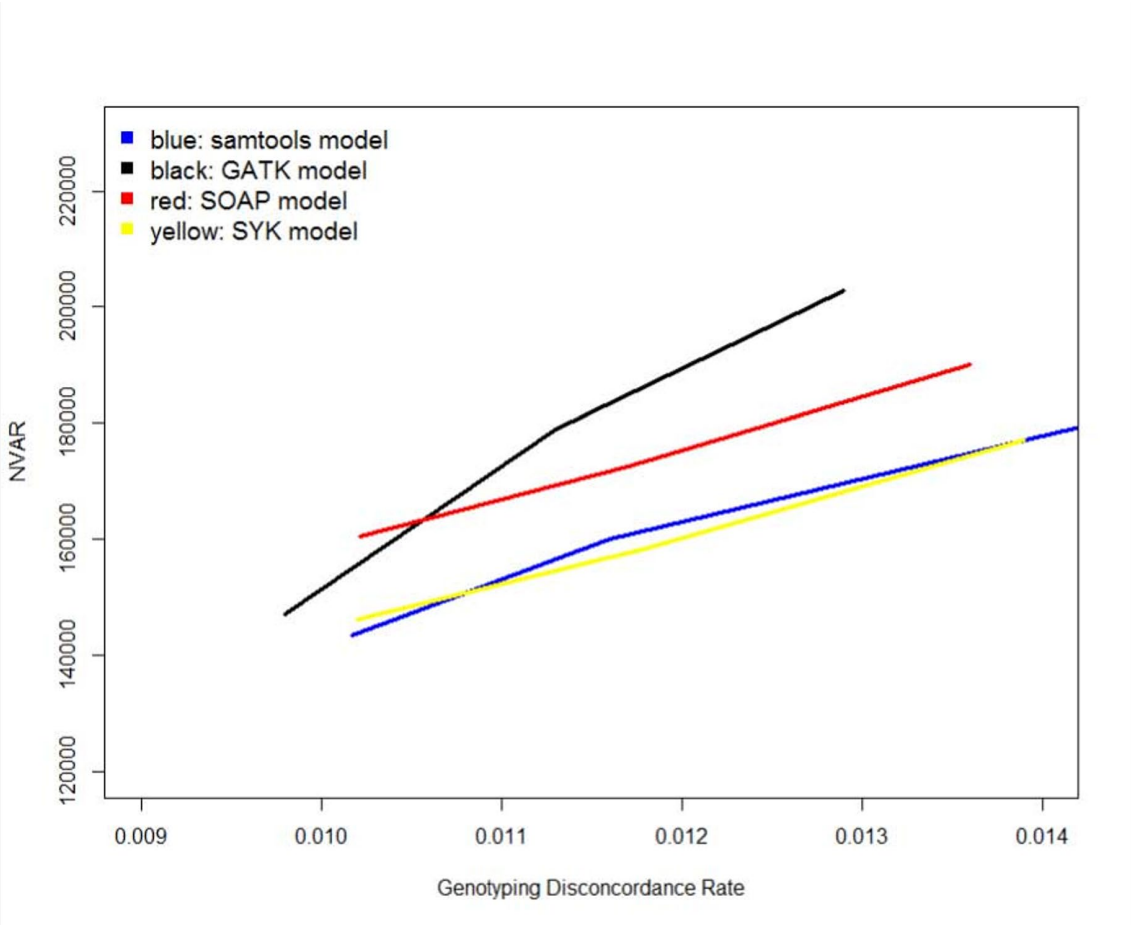
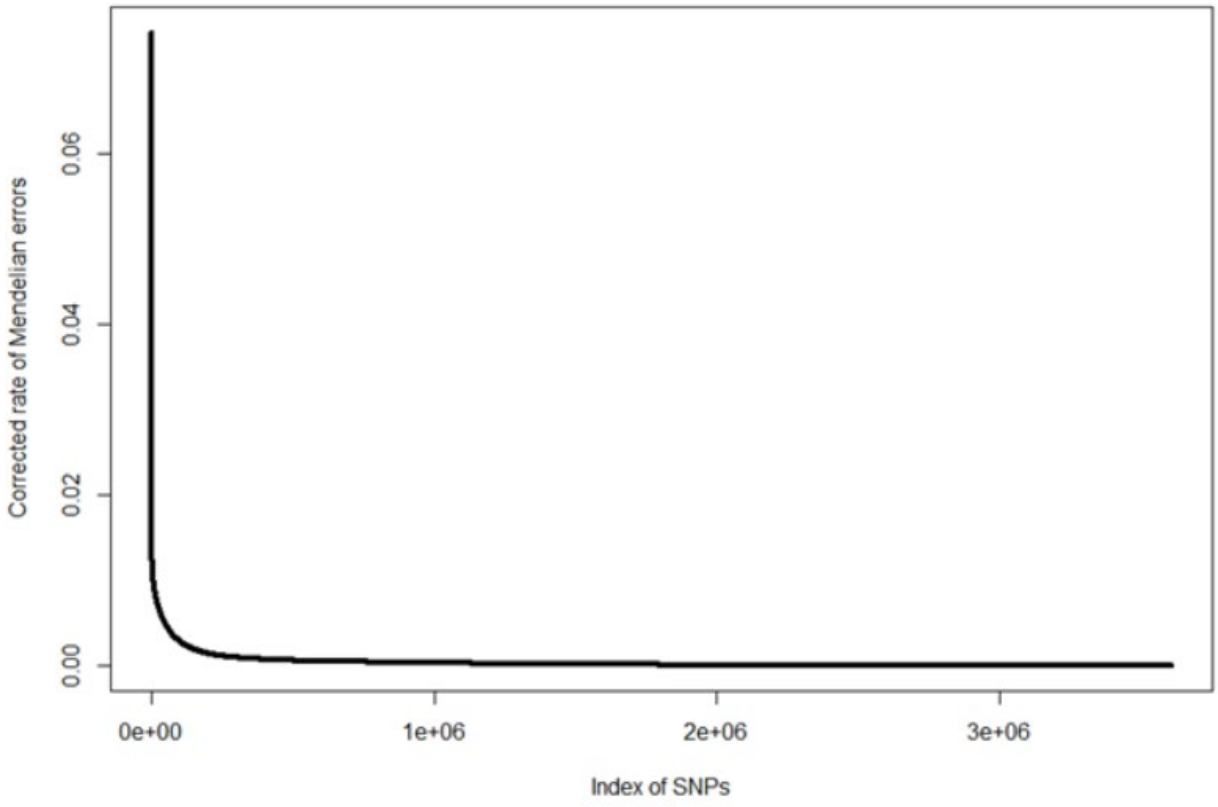


Figure S9. Number of variants by genotype discordance rates for 4 ANGSD genotype likelihood models.



963
964 **Figure S10. Mendelian error rates**

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



967
968
969
970
971
972
973
974
975
976
977
978

Figure S11. Available rat genetic maps.

Plotted physical and genetic distances are for chromosome 12.

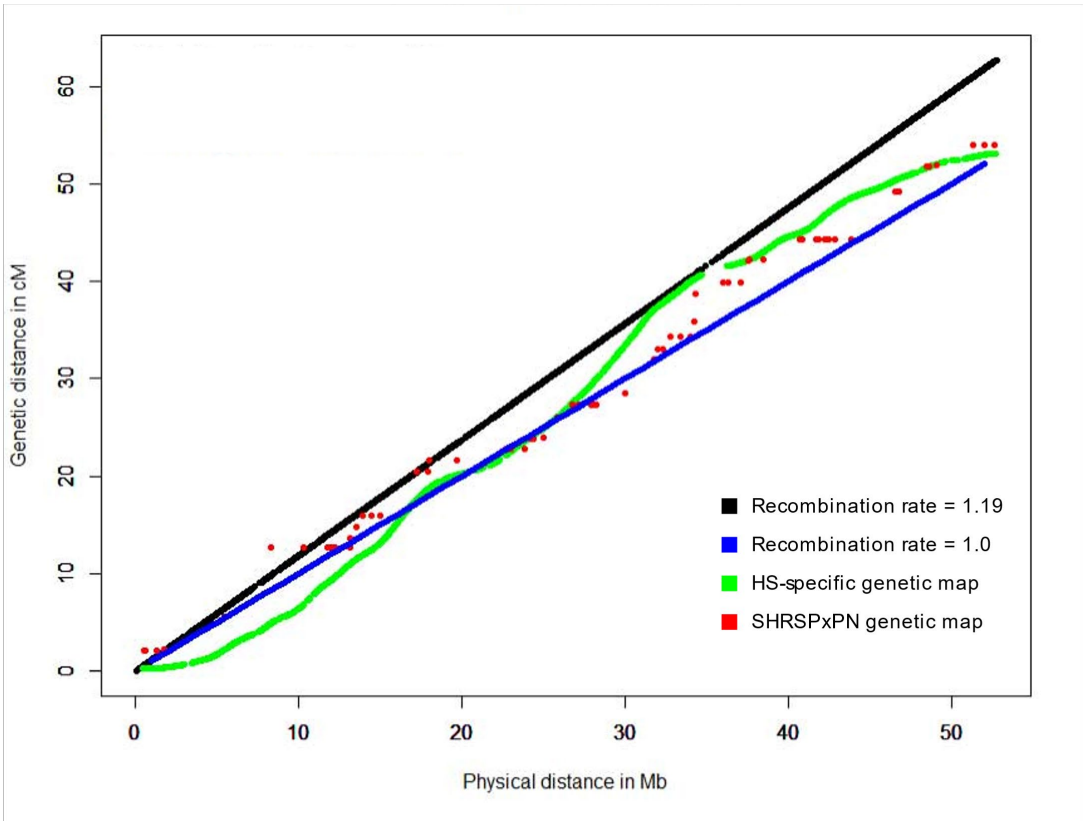


Table S1. Demultiplexing performance.

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 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
$T_I T_V$ 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/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. 2018).

	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
$T_S T_V$	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