1

2

3

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

5 Alexander F. Gileta[*†⁋], Jianjun Gao[*⁋], Apurva S. Chitre*, Hannah V. Bimschleger[*], Celine L. St.

6 Pierre[*], Shyam Gopalakrishnan[#], Abraham A. Palmer[*‡]

7

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

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

10 [#] Department of Biology, University of Copenhagen, 2200 København N, Denmark

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

12

13

14 [⁋] 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. 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 the 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; X. Huang et al. 2009; Andolfatto et al. 2011; 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 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; 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

70 been described, including Stacks (Catchen et al. 2013), IGST-GBS (Sonah et al. 2013), TASSEL-
71 GBS (Glaubitz et al. 2014), Fast-GBS (Torkamaneh et al. 2017), and GB-eaSy (Wickland et al.
72 2017), the majority were developed and optimized for use in plant species and given the lack of
73 well-developed genomic resources in these species, do not leverage the wealth of genomic data
74 available for model organisms such as rats. Here we describe the customized computational and
75 laboratory protocols for applying GBS to HS rats.

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

85 Detailed here are the steps we have taken to optimize a rat GBS protocol and computational
86 pipeline. Drawing on existing protocols (Elshire et al. 2011; Peterson et al. 2012; Poland et al.
87 2012; Parker et al. 2016) as templates, we redesigned our previous GBS approach (Parker et al.
88 2016; Gonzales et al. 2018) and have developed a novel, reference-based, high-throughput
89 workflow to accurately and cost-effectively call and impute variants from low-coverage double
90 digest GBS (**ddGBS**) data in HS rats. This publication is intended as a resource for others who
91 might wish to perform GBS in rats and should provide a roadmap for adapting GBS for use in new
92 species. We demonstrate that with a suitable reference panel, applying reduced representation

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

95 MATERIALS AND METHODS

96 **Tissue samples and DNA extraction**

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

110 ***In silico* digest of rat genome**

111 We used *in silico* digests to aid in the selection of restriction enzymes, with the goal of maximizing
112 the proportion of the genome captured at sufficient depth to make confident genotype calls (Kent
113 et al. 2002). We used the *restrict* function in EMBOSS (version 6.6.0) (Rice, Longden, and
114 Bleasby 2000) in conjunction with the REBASE database published by New England BioLabs

115 (NEB; version 808) (Roberts and Macelis 1999) to perform *in silico* digest of the current release
116 of the Norway brown rat reference genome, designated rn6. For the primary restriction enzyme,
117 we chose PstI, which had been successfully used in numerous project (Fitzpatrick et al. 2013;
118 Parker et al. 2016; Gonzales et al. 2018). We performed the digest with PstI alone and then with
119 PstI paired with each of 7 secondary enzymes: AluI, BfaI, DpnI, HaeIII, MluCI, MspI, and NlaIII.
120 We only considered fragments with one PstI cut site and one cut site from the secondary enzyme
121 because the adapter and primer sets are designed to only allow these fragments to be amplified.

122 **Restriction enzyme selection**

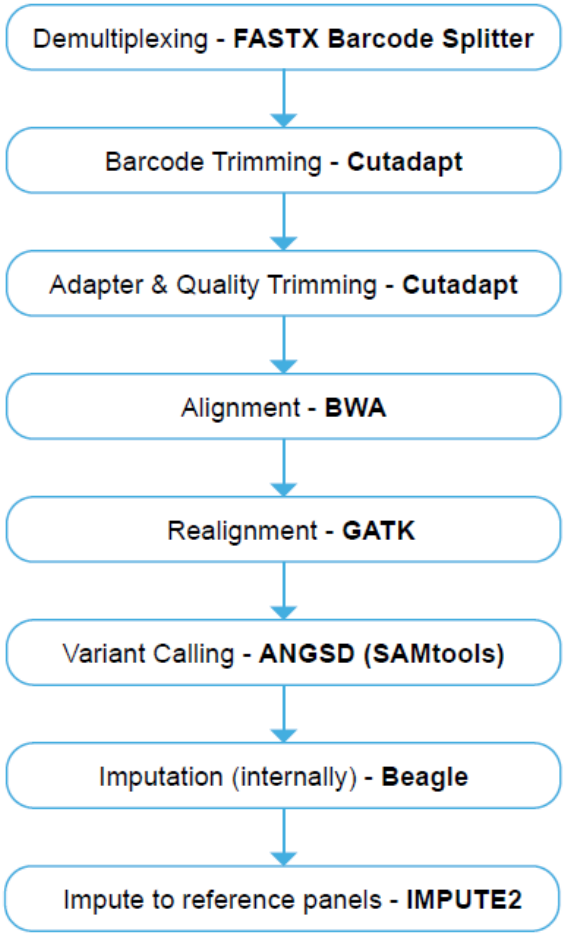
123 Initial criteria for selecting a secondary restriction enzyme were a 4bp recognition sequence, no
124 ambiguity in the recognition sequence (i.e. N's), compatibility with the NEB CutSmart Buffer, and
125 an incubation temperature of 37ºC. The list of enzymes meeting these criteria at the time included
126 AluI, BfaI, DpnI, HaeIII, MluCI, MspI, and NlaIII. Using the *in silico* digest data, we looked to
127 maximize the portion of the genome contained within a fragment size range of 125-275bp (250-
128 400bp with annealed adapters and primers) (Figure 1; Table 1). We excluded enzymes that
129 produced blunt ends, both because it would be more difficult to anneal adapters to blunt ended
130 fragments and because our adapters would then also anneal to blunt ends produced by DNA
131 shearing. We also excluded methylation-sensitive enzymes, as we did not want to limit the breadth
132 of our sequencing efforts, accepting the possibility of read pileup in repetitive regions. Based on
133 these criteria, as well as maximizing the percent of the genome captured, NlaIII, BfaI, and MluCI
134 were selected for further testing. The final choice of enzyme (NlaIII) was determined empirically
135 and is detailed in the Results.

136 **ddGBS library preparation and sequencing**

137 The full ddGBS protocol is available in File S1. In brief, approximately 1µg of DNA was used per
138 sample. Sample DNA, PstI barcoded adapters, and NlaIII Y-adapter were combined in a 96-well
139 plate and allowed to evaporate at 37ºC overnight. The PstI adapter barcode is "in-line" such that
140 each sequencing read from a given sample with contain both the PstI overhang sequence (4bps)
141 and unique adapter sequence (4-8bps) prior to the beginning of the insert sequence. Sample DNA
142 and adapters were re-eluted on day two with a PstI/NlaIII digestion mix and incubated at 37ºC for
143 two hours to allow for complete digestion. Ligation reagents were then added and incubated at
144 16ºC for one hour to anneal the adapters to the DNA fragments, followed by a 30-minute
145 incubation at 80ºC to inactivate the restriction enzymes. Sample libraries were purified using a
146 plate from a MinElute 96 UF PCR Purification Kit (QIAGEN Inc., Hilden, Germany), vacuum
147 manifold, and ddH_2O. Once re-eluted, libraries were quantified in duplicate with Quant-IT
148 PicoGreen (Thermo Fisher Scientific, Waltham, MA) and pooled to the desired level of
149 multiplexing (i.e. 12, 24, or 48 samples per library). Each pooled library was then concentrated by
150 splitting the pooled volume across 2-3 wells of the MinElute vacuum plate and resuspending the
151 library at desired volume for use in the Pippin Prep. The concentrated pool was quantified to ensure
152 the gel cassette will not be overloaded with DNA (>5µg). The pool was then loaded into the Pippin
153 Prep for size selection (300-450bps) using a 2% agarose gel cassette on a Pippin Prep (Sage
154 Science, Beverly, MA). Size-selected libraries were then PCR amplified for 12 cycles to add the
155 Illumina sequencing primers and increase the quantity of DNA, concentrated again, and checked
156 for quality on an Agilent 2100 Bioanalyzer with a DNA 1000 Series II chip (Agilent Technologies,
157 Santa Clara, CA)., Bioanalyzer results were used to assure sufficient DNA concentration and to
158 identify excessive primer dimer peaks.

159 As a pilot, an initial 96 HS samples were sequenced, 12 samples per library, at Beckman
160 Coulter Genomics (now GENEWIZ) on an Illumina HiSeq 2500 with v4 chemistry and 125bp
161 single-end reads. Subsequently, we began using a set of 48 unique barcoded adapters (File S2) to
162 multiplex 48 HS samples per ddGBS library. Each library thereafter was run on a single flow cell
163 lane on an Illumina HiSeq 4000 with 100bp single-end reads at the IGM Genomics Center
164 (University of California San Diego, La Jolla, CA). We also obtained ddGBS data on the HiSeq
165 4000 for a select set of 96 samples that had been previously genotyped on a custom Affymetrix
166 Axiom MiRat 625k microarray (Part#: 550572), providing us with a truth set.

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



169
170

171 **Evaluation of ddGBS pipeline performance**

172 We present the steps required to call and impute genotypes from raw ddGBS sequencing data in
173 Figure 2. During optimization of the pipeline, performance was assessed by two primary metrics:
174 (1) the number of variants called and (2) genotype concordance rates for calls made in 96 HS rats
175 that had both ddGBS genotypes and array genotypes from a custom Affymetrix Axiom MiRat
176 625k microarray. There were two checkpoints in the GBS pipeline where genotype quality
177 (measured by concordance rate) was assessed. The first was after "internal" imputation with
178 Beagle (Browning and Browning 2009; 2016), whereby we leverage information from samples
179 that had sufficient read depth to make a confident genotype call at a given locus in order to impute
180 the genotype of other samples that had lower read depths at that locus. The second checkpoint was
181 after "external" imputation, meaning imputation to our reference panel with IMPUTE2 to obtain
182 genotype calls at loci we did not directly capture by our GBS method. (B. N. Howie, Donnelly,
183 and Marchini 2009; B. Howie et al. 2012). A third, additional metric we checked was the transition
184 to transversion ratio ($T_S T_V$), which is expected to be ~2 for intergenic regions. The steps as outlined
185 in the following sections reflect the final version of the pipeline. Variant calling and imputation
186 steps utilized all available samples run on the HiSeq 4000 (3,000+ rats), though genotype
187 concordance rates could only be calculated for the set of 96 HS samples for which we had array
188 genotype calls.

189 **Demultiplexing**

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

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

199 **Barcode, adapter, and quality trimming**

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

205 **Read alignment and indel realignment**

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

212 **Variant calling**

213 Variants were called, and genotype likelihoods were computed at variant sites using ANGSD
214 v0.911, under the SAMtools model for genotype likelihoods (ANGSD-SAMtools) (Korneliussen,
215 Albrechtsen, and Nielsen 2014; Durvasula et al. 2016). Further, using ANGSD-SAMtools, we

216 inferred the major and minor alleles (*-domajorminor* 1) from the genotype likelihoods, retaining
217 only high confidence polymorphic sites (*-snp_pval* 1e-6), and estimated the allele frequencies
218 based on the inferred alleles *(-domaf* 1). We discarded sites missing read data in more than 4% of
219 samples (*–minInd*). Additionally, we tested multiple thresholds for minimum base (*-minQ*) and
220 mapping (*-minMapQ*) qualities.

221 **Internal imputation**

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

229 **Quality Control for genotypes before imputation using an external reference panel**

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

238 we formed curves for the distributions of the expected number of Mendelian errors for both SNPs
239 and samples and chose the inflection points as thresholds for the number of Mendelian errors
240 allowed.

241 **Data preparation for phasing with external reference panel**

242 First, in our study sample of 96 samples, we only retained variants previously identified in
243 the 8 HS founder strains because we expected the polymorphisms in our samples to be limited to
244 the variation present in the founders (Hermsen et al. 2015; Ramdas et al. 2019). Further, to improve
245 imputation efficiency, we employed a pre-phasing step with IMPUTE2 (*prephase_g*) (B. Howie
246 et al. 2012) prior to imputation. Pre-phasing only needs to be performed once, allowing us to reuse
247 the estimated haplotypes from out dataset for imputation with multiple different reference panels.
248 A flowchart outlining the pre-phasing protocol is presented in Figure S2.

249 **Genetic maps**

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

257 **Imputation to reference panel**

258 We used a combination of existing sequencing and array genotyping data from the HS rat founder
259 and other inbred laboratory rat strains (Hermsen et al. 2015) as reference panel for imputation.

260 Genotype data underwent QC and were phased by Beagle into single chromosome haplotype files.
261 Haplotype files were then created using the workflow detailed in Figure S2. Imputation by
262 IMPUTE2 was performed in 5Mb windows using the aforementioned reference panels and genetic
263 maps.

264 **Data availability**

265 Genotype data will be available at https://dx.doi.org/10.6084/m9.figshare.8243222. The code
266 necessary to run the steps outlined in the publication are provided at
267 https://dx.doi.org/10.6084/m9.figshare.8243156. Supplementary Files are available at
268 https://dx.doi.org/10.6084/m9.figshare.8243129. Remaining files necessary for imputation
269 (genetic maps, reference data, etc.) can be found with the following links:
270 https://dx.doi.org/10.6084/m9.figshare.11919615 ,
271 https://dx.doi.org/10.6084/m9.figshare.11919573 ,
272 https://dx.doi.org/10.6084/m9.figshare.11919597.

273 RESULTS

274 **ddGBS optimization**

275 Previous projects utilizing GBS in mice and rats (Fitzpatrick et al. 2013; Parker et al. 2016;
276 Gonzales et al. 2018) often encountered an issue where certain regions of the genome experienced
277 high pileups of reads per sample (>100x), while other regions were covered by just 1-2 reads. This
278 read distribution imbalance can be caused in part by PCR amplification bias, where a subset of
279 fragments are preferentially amplified until they dominate the final library (Kanagawa 2003; Aird
280 et al. 2011). These previous protocols utilized 18 cycles of amplification. We tested reducing this
281 to 8, 10, 12, or 14 cycles and found that below 12 cycles, there was insufficient PCR product to

282 accurately quantify and pool for sequencing. The reduction in the number of PCR cycles was
283 expected to reduce PCR bias, though this was not explicitly tested.

284 Another concern regarding previous sequencing results was an excess of long fragments
285 (>700bps as determined by *in silico* digest), which do not provide sufficient reads to make
286 confident genotype calls (< 5 reads per sample) and are therefore wasteful. We tested three
287 methods of combating this issue, including increasing the digestion time or enzyme concentration,
288 performing size selection on the libraries, and using a two-enzyme restriction digest.

289 We considered the possibility that the restriction enzyme digests might not be running to
290 completion. To address this possibility, we increased the duration of the digestion from 2 hours to
291 3 or 4 hours. We also tried increasing the number of units of PstI enzyme added, to ensure complete
292 digest. Neither of these modifications impacted the final fragment length distribution of the library,
293 indicating that the digest was reaching completion after 2 hours using the original concentration
294 of PstI (File S3 – wells 1-6).

295 Our previous GBS protocol did not have an explicit library fragment size selection step.
296 The final library was purified using a MinElute PCR Purification Kit (QIAGEN Inc., Hilden,
297 Germany), which isolates PCR products 70bp-4kb in length, leaving a wide range of fragment
298 sizes in the final library, under the assumption that only shorter fragments would bridge amplify
299 on the flow cell. This method was imprecise and had low reproducibility, negatively impacting our
300 ability obtain reads at consistent sites across libraries. Rather than attempt size selection by gel
301 extraction, we chose to utilize a Pippin Prep, which automates the elution of DNA libraries of
302 desired fragment size ranges. By using this automated size selection, we reduced the proportion of
303 the genome targeted for sequencing, Additionally, since restriction enzymes make predominantly
304 consistent cuts across samples (barring the presence of polymorphisms in RE recognition sites), it

305 is ensured that highly similar sets of genomic fragments will be sequenced across sample libraries.
306 Since the clustering process involves a bridge amplification step that preferentially amplifies
307 library fragments with shorter insert sizes (Illumina, Inc. 2014), we kept the size selection window
308 narrow (250-400bps) to avoid introducing a bias in which fragments were sequenced. A
309 comparison of the fragment size distributions for the protocols before and after introduction of the
310 Pippin Prep is shown in File S4.

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

325 The number of fragments with one of each of the cut sites were summed for all observed
326 lengths and the results summarized in Figure 1 and Table 1. BfaI, MluCI, and NlaIII were chosen
327 for further testing due to their compatibility with PstI digestion reagents and temperatures, sticky

328 ends, and the proportion of the genome falling in the size selection window. We ruled out BfaI
329 because it only had a 2bp overhang after cleavage, which led to a high concentration of adapter
330 dimer in the sequencing libraries (S5 File). NlaIII was chosen because it contained the greatest
331 portion of the genome in the size selection window.

332 **Table 1. Restriction enzyme options for double digest.**

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

335

336

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%

337 * Restriction enzyme is methylation sensitive.

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

339

340

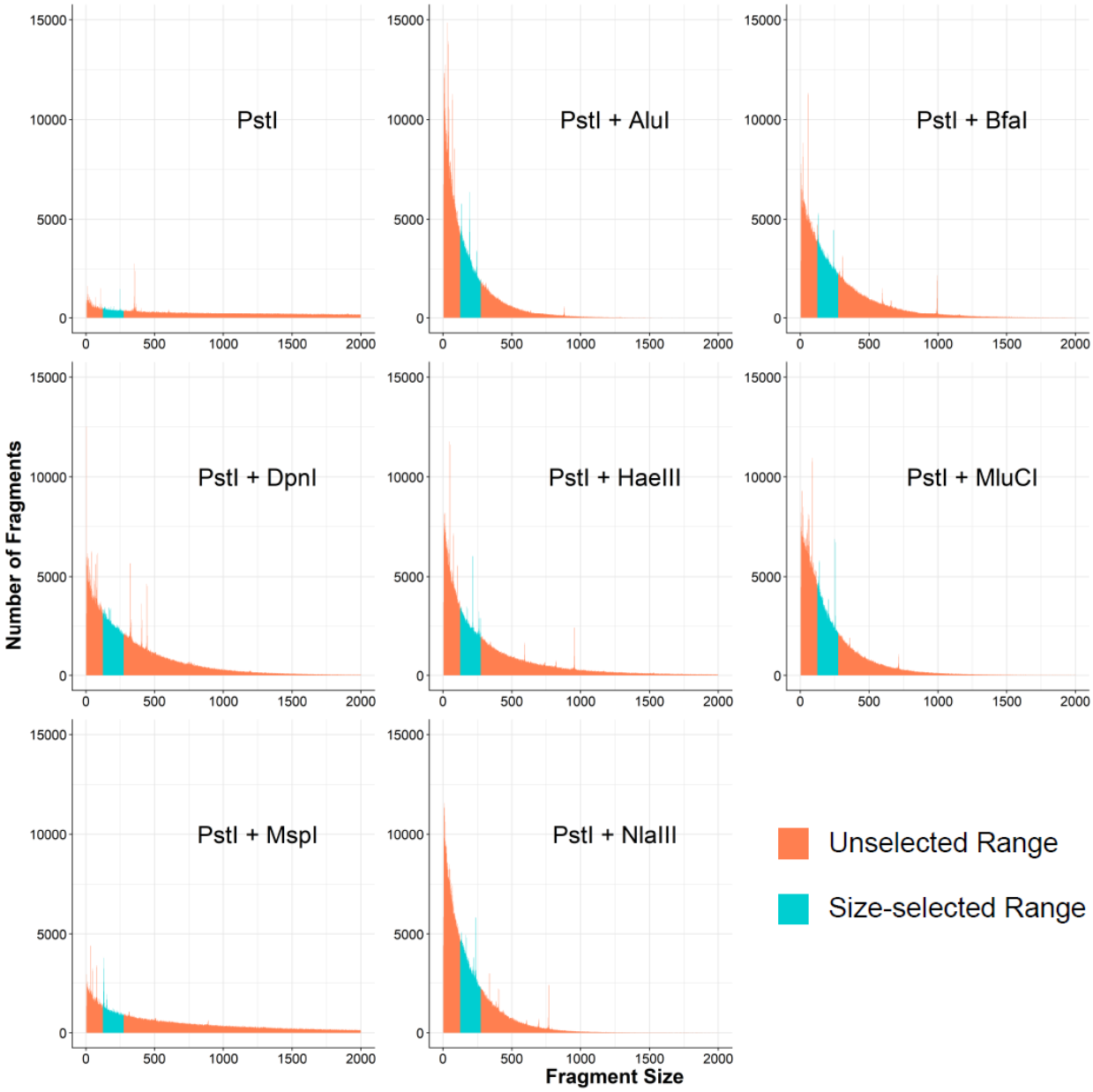
341

342

343

344

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.

355 In our previous GBS protocol, all fragments were cut on both ends by PstI. By using a
356 substantially lower concentration of the barcoded PstI adapter than the common PstI adapter, we
357 ensured the barcoded adapter would be the limiting reagent and the majority of fragments with an
358 annealed barcoded adapter would have a common adapter on the other end. This is crucial, as
359 having one of each of the adapters is required for proper amplification of the fragments on the flow
360 cell. However, when using both PstI and NlaIII, the library is predominantly composed of
361 fragments cut on both sides by NlaIII (File S6), which will amplify during PCR with a common
362 adapter, but not on the flow cell. Therefore, we employed a Y-adapter (Poland et al. 2012) to
363 control the direction of the first round of PCR and prevent two-sided NlaIII fragments from
364 dominating the final sequencing library (File S2).

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

371 We sequenced a trial flow cell with 8 pooled ddGBS libraries of 12 SD rat samples each
372 (96 total) on a HiSeq 2500 (Illumina, San Diego, CA) with 125bp reads and v3 chemistry,
373 obtaining an average of 15.3 million reads per sample. Given the NlaIII *in silico* digest results
374 suggested we were capturing ~3.4% of the genome and that we were using 125bp reads, this
375 corresponded to approximately 20x coverage of captured sites. We subsequently increased the
376 number of samples to 48 per library for the HS rats because we hypothesized 5x would be sufficient
377 coverage per sample when utilizing imputation to a reference panel. We also discovered that a

378 portion of the reads contained sequence fragments of the NlaIII adapter sequence, indicating there
379 were fragments with insert sizes smaller than 125bps in the final library. To avoid this, we
380 increased the fragment size range to 300-450bps (Table 1), which corresponds to a 175-325bp
381 insert size once the adapters and primers are accounted for. We noted however that the library size
382 distribution obtained from the Pippin Prep was uniformly shifted towards higher fragment lengths
383 (Figure S3). This is a result of the high concentrations of our libraries after pooling and loading
384 the gel cartridge near the upper limit of the recommend number of micrograms of DNA, which
385 can cause slower migration of the DNA across the gel matrix.

386 The final ddGBS protocol can be found in File S1 and the necessary primer and adapter
387 sequences in File S2. This protocol was used for the sequencing of all HS rats included in the
388 following computational optimization steps. These steps were optimized using

389 **Demultiplexing**

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

396 **Adapter and quality trimming**

397 Read quality was substantially improved after trimming the barcode and adapter sequences and
398 low-quality base pairs at the ends of reads (Figure S5). Overall read counts were only marginally
399 reduced by quality trimming (Table S1). We observed that the number of called variant sites and

400 the genotyping rate were both greater when using reads initially processed by cutadapt (Martin,
401 2011) than reads processed by the FASTX_Toolkit (Table S2). Importantly, a large portion of the
402 additional identified variants were known variant sites from the 42 inbred strains reference set
403 (Figure S6), indicating the elevated call rate was at least in part due to capturing more true variant
404 sites. We viewed this as sufficient support for proceeding with cutadapt for adapter removal and
405 quality trimming.

406 **Mapping quality**

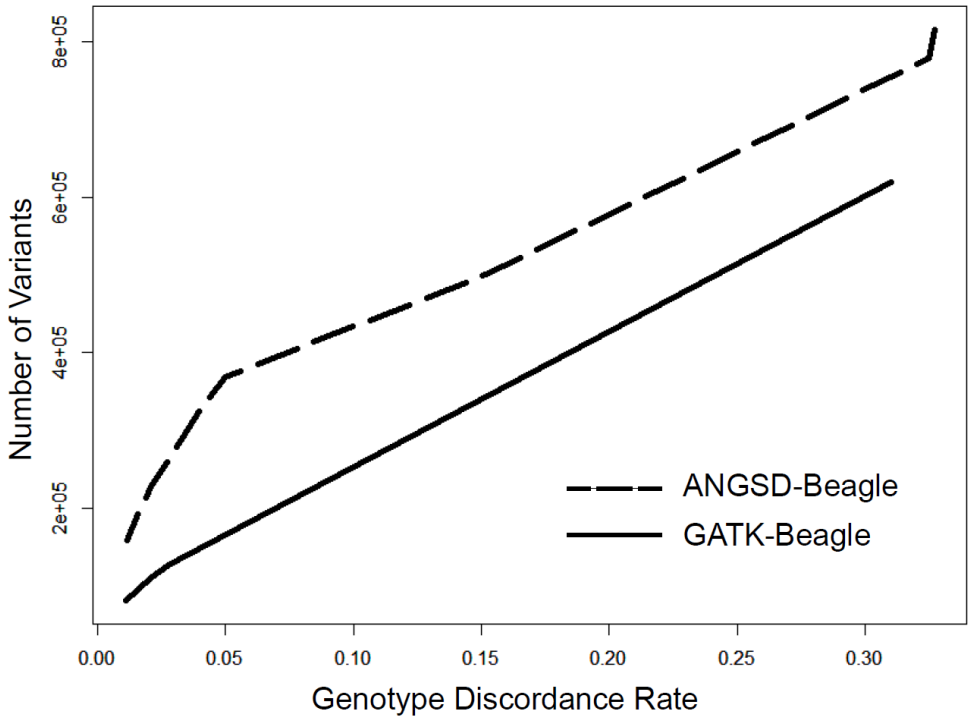
407 The number of called variants and genotype call rates were identical at read mapping quality
408 (mapQ) thresholds of either 20 or 30 (Table S3) within ANGSD. As the ANGSD mapQ threshold
409 was raised to 45, there was a small reduction in the number of called variants, and then much
410 greater losses at thresholds of 60 or 90. Fortunately, discordance rates between ddGBS and array
411 genotypes were stable at both low and high mapQ thresholds, despite the putatively higher quality
412 of the alignments (Figure S7). This permitted us to select a lower mapQ threshold (mapQ = 20),
413 maximizing the number of variants called without sacrificing genotyping accuracy.

414 **Variant calling**

415 Figure 3 shows that across all levels of genotype discordance rates (with the array genotyping
416 data), the combination of the ANGSD-SAMtools with BEAGLE produced more SNPs, at various
417 discordance thresholds, than GATK's HaplotypeCaller (McKenna et al. 2010; DePristo et al.
418 2011). This observation held when variants were limited only to biallelic sites and SNPs with an
419 MAF > 0.05 (Figure S8). We speculate that the poorer performance of HaplotypeCaller may be in
420 part to the sparsity and non-uniform distribution of GBS genotype data across the genome and the
421 high level of genotype call missingness across samples prior to imputation.

422

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



425

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

431

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

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

449 **Imputation to reference panel**

450 Imputation is used in two complimentary ways in our protocol. As described earlier, after ddGBS,
451 not all samples will have sufficient sequencing coverage at captured polymorphic loci to make a
452 confident genotype call. Therefore, we first use imputation from other well-covered samples to
453 "fill in the blanks" and assign genotypes to SNP loci in the subset of individuals that lacked
454 confident calls at these sites. imputation in this section to call genotypes at sites where GBS that
455 were inaccessible to ddGBS sequencing. After these missing genotypes have been imputed in all
456 samples, we then use the genotype information we have for the SNPs captured by ddGBS along
457 with the reference panel data on the original 8 HS founders (Hermsen et al. 2015; Ramdas et al.
458 2019) to impute genotype calls at sites that were inaccessible to ddGBS sequencing. Thus, our
459 second application of imputation is similar to the human genetics application in which imputation
460 using 1000 Genomes (1000 Genomes Project Consortium et al. 2010) increases the number of

461 SNPs beyond those included on a given microarray platform. IMPUTE2 was selected over Beagle
462 for this application because it is able to handle multiple reference panels, allowing us to provide
463 data from all 8 HS founder strains.

464 Before starting this imputation step, we observed an inflated transition/transversion ratio
465 (Table S4) in our ANGSD-SAMtools/Beagle SNPs. This issue was ameliorated when the SNP set
466 was filtered for only "known" variants that were previously identified in either the 42 inbred strains
467 (Hermsen et al. 2015) or the 8 deep-sequenced HS founders (Ramdas et al. 2019). For imputation,
468 we therefore only provided IMPUTE2 with previously identified variant sites from our ANGSD-
469 SAMtools/Beagle output. Prior to running IMPUTE2, we also filtered the variants for biallelic
470 sites with a genotype call in at least two individuals. Using pedigree data for the HS rats, we further
471 removed samples showing an order of magnitude higher level of Mendelian error than the sample
472 mean. We further removed SNPs that had an error rate surpassing a threshold of ~0.005 (Figure
473 S10; inflection point). There were 4 samples and 4,179 SNPs removed from subsequent analyses.
474 Lastly, we removed any samples where the X chromosome read ratio was incompatible with their
475 reported sex using hard threshold of 3% of total reads, above indicating female and below
476 indicating male (Figure S1).

477 There were three major genomic reference datasets available for the HS rats. The first
478 reference set was obtained from Baud et al. (Rat Genome Sequencing and Mapping Consortium
479 et al. 2013) and contained sequence data and genotype calls for the 8 founders of the HS. The
480 second came from Hermsen et al. (Hermsen et al. 2015) and sequence and genotype data on 42
481 distinct laboratory rats strains and substrains, 8 of which were the founders of the HS from Baud
482 et al., but analyzed alongside a new set of strains. The third reference set came from Ramdas et al.
483 (Ramdas et al. 2019), who independently performed whole-genome sequencing and made

484 genotype calls on the 8 HS founder strains. It was unclear which set of genotypes would provide
485 the best reference for imputation from out ddGBS data, so we tested five different possible subsets
486 of available data (Table 2). From Hermsen et al., we used (1) all 42 inbred strains, (2) only the 34
487 strains that were not the HS founders, and (3) only the 8 HS founder strains. Then from Baud et
488 al. and Ramdas et al., we tested the 8 HS founders only from each study. The most accurate
489 imputation was observed for the reference set containing only the 8 deep-sequenced HS founder
490 strains (Ramdas et al. 2019); however, imputation to this set had the lowest genotyping rate of all
491 panels. In contrast, using the 42 rat inbred strains displayed a balance of high accuracy and low
492 missingness, leading us to choose this as our reference set. To better understand the role of the 8
493 founder strains, which were part of the 42 strain reference panel, we created a reference panel that
494 included only the 34 non-HS founder strains. As expected, discordance rates were much higher
495 when only considering non-founders. However, the genotype missingness was lower for the 34
496 than the 8 founders alone, suggesting a combination of the two was the optimal set.

497

498

499

500

501

502

503

504

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 map performed similarly (Table S5). Surprisingly, the linear genetic maps performed just as well as the HS-specific map (Littrell et al.

517 2018). Thus, for simplicity, we chose to use the chromosome-specific values initially published by
518 Jensen-Seaman (Jensen-Seaman 2004).

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

522 DISCUSSION

523 The use of microarrays and WGS for genotyping large samples in model organisms remains cost-
524 prohibitive. There is therefore an urgent and wide-spread need for high-performance and
525 economical methods of obtaining genome-wide genotype data. While reduced-representation
526 approaches have been utilized in numerous species of plants and animals, including rodents
527 (Peterson et al. 2012; Fitzpatrick et al. 2013; Parker et al. 2016; Gonzales et al. 2017; Zhou et al.
528 2018), there has yet to be a published protocol optimized specifically for rats. Prior to sequencing
529 thousands of HS samples with GBS for our mapping efforts, we wanted to ensure we were
530 capturing the greatest possible number of high-quality variants at the lowest possible cost. The
531 protocol we present here is the culmination of careful testing and optimization of each step of the
532 GBS protocol for rats. We have now applied the approach to 4,973 HS rats, as well as 4,608
533 Sprague Dawley rats (Gileta et al. 2018).

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

539 Catchen et al. 2013; Glaubitz et al. 2014; Torkamaneh et al. 2017; Wickland et al. 2017), some of
540 which are polyploid and have differing levels of variation and LD than outbred rodent populations.
541 Additionally, there were elements of each pipeline that did not meet our needs or lacked
542 customizability. For instance, TASSEL-GBS v2 (Glaubitz et al. 2014) trims all reads to 92 base
543 pairs; however, other projects underway in our lab utilized up to 125bp reads, leading to a ~20%
544 reduction in data. TASSEL-GBS also ignores read base quality scores, which are informative in
545 probabilistic frameworks for estimating uncertainty in alignments and variant calls (H. Li, Ruan,
546 and Durbin 2008; DePristo et al. 2011; Nielsen et al. 2011), and uses a naïve binomial likelihood
547 ratio method for calling SNPs. Stacks has previously shown poor performance in demultiplexing
548 (Herten et al. 2015; Torkamaneh et al. 2017) and does not make use of the reference genome for
549 priors when calling SNPs (Catchen et al. 2013). Fast-GBS relies on Platypus (Rimmer et al. 2014)
550 for variant calling (WGS500 Consortium et al. 2014; Torkamaneh et al. 2017), which employs a
551 Bayesian method of constructing candidate haplotypes that works poorly with low-pass
552 sequencing data and does not scale well to large sample sizes (Z. Li, Wang, and Wang 2018).
553 Lastly, none of these pipelines included an imputation step, which is crucial for filling in missing
554 genotypes in GBS data and can provide millions of additional SNPs given an appropriate
555 composite reference panel (B. Howie, Marchini, and Stephens 2011; G.-H. Huang and Tseng
556 2014).

557 Though we have not explicitly tested each alternate GBS pipeline for the purposes of this
558 publication, this has been recently done by Wickland et al. (Wickland et al. 2017). Their pipeline
559 GB-eaSy, which ours most closely resembles, was found to be superior by a number of metrics to
560 Stacks, TASSEL-GBS, IGST, and Fast-GBS. Similar to GB-eaSy, our pipeline utilizes a double-
561 digest GBS protocol, aligns reads to the reference genome with *bwa mem*, and uses the SAMtools

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

568 A minor difference between GB-eaSy and our pipeline is the use of cutadapt (Martin 2011)
569 rather than GBSX (Herten et al. 2015) for demultiplexing, though both performed equally well
570 (Table S1). The primary improvement is our extension of the pipeline with the implementation of
571 effective internal and reference-based imputation steps using the 42 inbred rat genomes (Hermsen
572 et al. 2015) and 8 deep-sequenced HS founders from UMich (Ramdas et al. 2019). There are two
573 stages of imputation in our pipeline. The first one is accomplished by Beagle and aims to fill in
574 missing genotypes at called variants using information from other samples. This raises the
575 genotype call rate to 100%, but it may also introduce errors due to insufficient information,
576 emphasizing the need for careful filtering steps. The second stage of imputation made use of
577 IMPTUE2 and an external reference panels of variants called from WGS data on the 8 inbred HS
578 founders, as well as 34 additional inbred rat strains. We decided to include the 34 additional strains
579 because of the elevated genotyping rate we observed upon their inclusion in the IMPUTE2
580 reference panel. We attribute this to the presence of haplotypes that exist in both the 8 the HS
581 founder strains and a subset of the 34 additional strains in this panel. The benefits of using a
582 composite reference panel have been previously noted (Zhang et al. 2013; G.-H. Huang and Tseng
583 2014); there is increased accuracy and decreased missingness in the imputed genotype data.

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

592

593

594

595

596

597

598

599

600

601

602

603

ACKNOWLEDGEMENTS:

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

607 LITERATURE CITED

608 1000 Genomes Project Consortium, Gonçalo R. Abecasis, David Altshuler, Adam Auton, Lisa D. Brooks,
609 Richard M. Durbin, Richard A. Gibbs, Matt E. Hurles, and Gil A. McVean. 2010. "A Map of Human
610 Genome Variation from Population-Scale Sequencing." *Nature* 467 (7319): 1061–73.
611 https://doi.org/10.1038/nature09534.
612 Aird, Daniel, Michael G Ross, Wei-Sheng Chen, Maxwell Danielsson, Timothy Fennell, Carsten Russ,
613 David B Jaffe, Chad Nusbaum, and Andreas Gnirke. 2011. "Analyzing and Minimizing PCR
614 Amplification Bias in Illumina Sequencing Libraries." *Genome Biology* 12 (2): R18.
615 https://doi.org/10.1186/gb-2011-12-2-r18.
616 Andolfatto, Peter, Dan Davison, Deniz Erezyilmaz, Tina T. Hu, Joshua Mast, Tomoko Sunayama-Morita,
617 and David L. Stern. 2011. "Multiplexed Shotgun Genotyping for Rapid and Efficient Genetic
618 Mapping." *Genome Research* 21 (4): 610–17. https://doi.org/10.1101/gr.115402.110.
619 Andrews, Simon. 2017. *FastQC* (version 0.11.6).
620 http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
621 Baird, Nathan A., Paul D. Etter, Tressa S. Atwood, Mark C. Currey, Anthony L. Shiver, Zachary A. Lewis,
622 Eric U. Selker, William A. Cresko, and Eric A. Johnson. 2008. "Rapid SNP Discovery and Genetic
623 Mapping Using Sequenced RAD Markers." Edited by Justin C. Fay. *PLoS ONE* 3 (10): e3376.
624 https://doi.org/10.1371/journal.pone.0003376.
625 Browning, Brian L., and Sharon R. Browning. 2009. "A Unified Approach to Genotype Imputation and
626 Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals." *The*
627 *American Journal of Human Genetics* 84 (2): 210–23.
628 https://doi.org/10.1016/j.ajhg.2009.01.005.
629 Browning, Brian L., and Sharon R. Browning. 2016. "Genotype Imputation with Millions of Reference
630 Samples." *The American Journal of Human Genetics* 98 (1): 116–26.
631 https://doi.org/10.1016/j.ajhg.2015.11.020.
632 Catchen, Julian, Paul A. Hohenlohe, Susan Bassham, Angel Amores, and William A. Cresko. 2013. "Stacks:
633 An Analysis Tool Set for Population Genomics." *Molecular Ecology* 22 (11): 3124–40.
634 https://doi.org/10.1111/mec.12354.
635 Chen, Qiang, Yufang Ma, Yumei Yang, Zhenliang Chen, Rongrong Liao, Xiaoxian Xie, Zhen Wang, et al.
636 2013. "Genotyping by Genome Reducing and Sequencing for Outbred Animals." Edited by
637 Shuhong Zhao. *PLoS ONE* 8 (7): e67500. https://doi.org/10.1371/journal.pone.0067500.
638 Chitre, Apurva S, Oksana Polesskaya, Katie Holl, Jianjun Gao, Riyan Cheng, Angel Martinez, Tony George,
639 et al. 2018. "Genome Wide Association Study of Body Weight, Body Mass Index, Adiposity, and
640 Fasting Glucose in 3,173 Outbred Rats," September. https://doi.org/10.1101/422428.
641 Davey, John W., Paul A. Hohenlohe, Paul D. Etter, Jason Q. Boone, Julian M. Catchen, and Mark L.
642 Blaxter. 2011. "Genome-Wide Genetic Marker Discovery and Genotyping Using next-Generation
643 Sequencing." *Nature Reviews Genetics* 12 (7): 499–510. https://doi.org/10.1038/nrg3012.
644 De Donato, Marcos, Sunday O. Peters, Sharon E. Mitchell, Tanveer Hussain, and Ikhide G. Imumorin.
645 2013. "Genotyping-by-Sequencing (GBS): A Novel, Efficient and Cost-Effective Genotyping
646 Method for Cattle Using Next-Generation Sequencing." Edited by James C. Nelson. *PLoS ONE* 8
647 (5): e62137. https://doi.org/10.1371/journal.pone.0062137.
648 DePristo, Mark A, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl,
649 Anthony A Philippakis, et al. 2011. "A Framework for Variation Discovery and Genotyping Using
650 Next-Generation DNA Sequencing Data." *Nature Genetics* 43 (5): 491–98.
651 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. 2016. "ANGSD -Wrapper: Utilities for Analysing next-Generation Sequencing Data." *Molecular Ecology Resources* 16 (6): 1449–54. https://doi.org/10.1111/1755-0998.12578.

Elshire, Robert J., Jeffrey C. Glaubitz, Qi Sun, Jesse A. Poland, Ken Kawamoto, Edward S. Buckler, and Sharon E. Mitchell. 2011. "A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species." Edited by Laszlo Orban. *PLoS ONE* 6 (5): e19379. https://doi.org/10.1371/journal.pone.0019379.

Fitzpatrick, Christopher J., Shyam Gopalakrishnan, Elizabeth S. Cogan, Lindsay M. Yager, Paul J. Meyer, Vedran Lovic, Benjamin T. Saunders, 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." Edited by Patrizia Campolongo. *PLoS ONE* 8 (10): e75042. https://doi.org/10.1371/journal.pone.0075042.

Flanagan, Sarah P., and Adam G. Jones. 2018. "Substantial Differences in Bias between Single-Digest and Double-Digest RAD-Seq Libraries: A Case Study." *Molecular Ecology Resources* 18 (2): 264–80. https://doi.org/10.1111/1755-0998.12734.

Fu, Yong-Bi, and Mo-Hua Yang. 2017. "Genotyping-by-Sequencing and Its Application to Oat Genomic Research." In *Oat*, edited by Sebastian Gasparis, 1536:169–87. New York, NY: Springer New York. https://doi.org/10.1007/978-1-4939-6682-0_13.

Furuta, Tomoyuki, Motoyuki Ashikari, Kshirod K. Jena, Kazuyuki Doi, and Stefan Reuscher. 2017. "Adapting Genotyping-by-Sequencing for Rice F2 Populations." *G3: Genes|Genomes|Genetics* 7 (3): 881–93. https://doi.org/10.1534/g3.116.038190.

Gileta, Alexander F., Christopher J. Fitzpatrick, Apurva S. Chitre, Celine L. St. Pierre, Elizabeth V. Joyce, Rachael J. Maguire, Africa M. McLeod, et al. 2018. "Genetic Characterization of Outbred Sprague Dawley Rats and Utility for Genome-Wide Association Studies," September. https://doi.org/10.1101/412924.

Glaubitz, Jeffrey C., Terry M. Casstevens, Fei Lu, James Harriman, Robert J. Elshire, Qi Sun, and Edward S. Buckler. 2014. "TASSEL-GBS: A High Capacity Genotyping by Sequencing Analysis Pipeline." Edited by Nicholas A. Tinker. *PLoS ONE* 9 (2): e90346. https://doi.org/10.1371/journal.pone.0090346.

Gonzales, Natalia M., Jungkyun Seo, Ana Isabel Hernandez-Cordero, Celine L. St. Pierre, Jennifer S. Gregory, Margaret G. Distler, Mark Abney, Stefan Canzar, Arimantas Lionikas, and Abraham A. Palmer. 2017. "Genome Wide Association Study of Behavioral, Physiological and Gene Expression Traits in a Multigenerational Mouse Intercross," December. https://doi.org/10.1101/230920.

———. 2018. "Genome Wide Association Analysis in a Mouse Advanced Intercross Line," September. https://doi.org/10.1101/230920.

Hannon Lab. 2010. *FASTX-Toolkit* (version 0.0.13). http://hannonlab.cshl.edu/fastx_toolkit/index.html.

He, Jiangfeng, Xiaoqing Zhao, AndrÃ© Laroche, Zhen-Xiang Lu, HongKui Liu, and Ziqin Li. 2014. "Genotyping-by-Sequencing (GBS), an Ultimate Marker-Assisted Selection (MAS) Tool to Accelerate Plant Breeding." *Frontiers in Plant Science* 5 (September). https://doi.org/10.3389/fpls.2014.00484.

Hermsen, Roel, Joep de Ligt, Wim Spee, Francis Blokzijl, Sebastian Schäfer, Eleonora Adami, Sander Boymans, et al. 2015. "Genomic Landscape of Rat Strain and Substrain 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.

Li, H. 2011. “A Statistical Framework for SNP Calling, Mutation Discovery, Association Mapping and Population Genetical Parameter Estimation from Sequencing Data.” *Bioinformatics* 27 (21): 2987–93. https://doi.org/10.1093/bioinformatics/btr509.

Li, H., and R. Durbin. 2009. “Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform.” *Bioinformatics* 25 (14): 1754–60. https://doi.org/10.1093/bioinformatics/btp324.

Li, H., J. Ruan, and R. Durbin. 2008. “Mapping Short DNA Sequencing Reads and Calling Variants Using Mapping Quality Scores.” *Genome Research* 18 (11): 1851–58. https://doi.org/10.1101/gr.078212.108.

Li, Zhentang, Yi Wang, and Fei Wang. 2018. “A Study on Fast Calling Variants from Next-Generation Sequencing Data Using Decision Tree.” *BMC Bioinformatics* 19 (1). https://doi.org/10.1186/s12859-018-2147-9.

Littrell, John, Shirng-Wern Tsaih, Amelie Baud, Pasi Rastas, Leah Solberg-Woods, and Michael J. Flister. 2018. “A High-Resolution Genetic Map for the Laboratory Rat.” *G3: Genes|Genomes|Genetics*, May, g3.200187.2018. https://doi.org/10.1534/g3.118.200187.

Martin, Marcel. 2011. “Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads.” *EMBnet.Journal* 17 (1): 10. https://doi.org/10.14806/ej.17.1.200.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, et al. 2010. “The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data.” *Genome Research* 20 (9): 1297–1303. https://doi.org/10.1101/gr.107524.110.

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 (2): 240–48. https://doi.org/10.1101/gr.5681207.

Nielsen, Rasmus, Joshua S. Paul, Anders Albrechtsen, and Yun S. Song. 2011. “Genotype and SNP Calling from Next-Generation Sequencing Data.” *Nature Reviews Genetics* 12 (6): 443–51. https://doi.org/10.1038/nrg2986.

Orsouw, Nathalie J. van, René C. J. Hogers, Antoine Janssen, Feyruz Yalcin, Sandor Snoeijers, Esther Verstege, Harrie Schneiders, et al. 2007. “Complexity Reduction of Polymorphic Sequences (CRoPS^{TM}): A Novel Approach for Large-Scale Polymorphism Discovery in Complex Genomes.” Edited by Ivan Baxter. *PLoS ONE* 2 (11): e1172. https://doi.org/10.1371/journal.pone.0001172.

Parker, Clarissa C, Shyam Gopalakrishnan, Peter Carbonetto, Natalia M Gonzales, Emily Leung, Yeonhee J Park, Emmanuel Aryee, et al. 2016. “Genome-Wide Association Study of Behavioral, Physiological and Gene Expression Traits in Outbred CFW Mice.” *Nature Genetics* 48 (8): 919–26. https://doi.org/10.1038/ng.3609.

Pértille, Fábio, Carlos Guerrero-Bosagna, Vinicius Henrique da Silva, Clarissa Boschiero, José de Ribamar da Silva Nunes, Mônica Corrêa Ledur, Per Jensen, and Luiz Lehmann Coutinho. 2016. “High-Throughput and Cost-Effective Chicken Genotyping Using Next-Generation Sequencing.” *Scientific Reports* 6 (May): 26929.

Peterson, Brant K., Jesse N. Weber, Emily H. Kay, Heidi S. Fisher, and Hopi E. Hoekstra. 2012. “Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species.” Edited by Ludovic Orlando. *PLoS ONE* 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, Mary K. Treutelaar, Katie Holl, Myrna Mandel, Leah C. Solberg Woods, and Jun Z. Li. 2019. “Extended Regions of Suspected Mis-Assembly in the Rat Reference Genome.” *Scientific Data* 6 (1): 39. https://doi.org/10.1038/s41597-019-0041-6.

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.

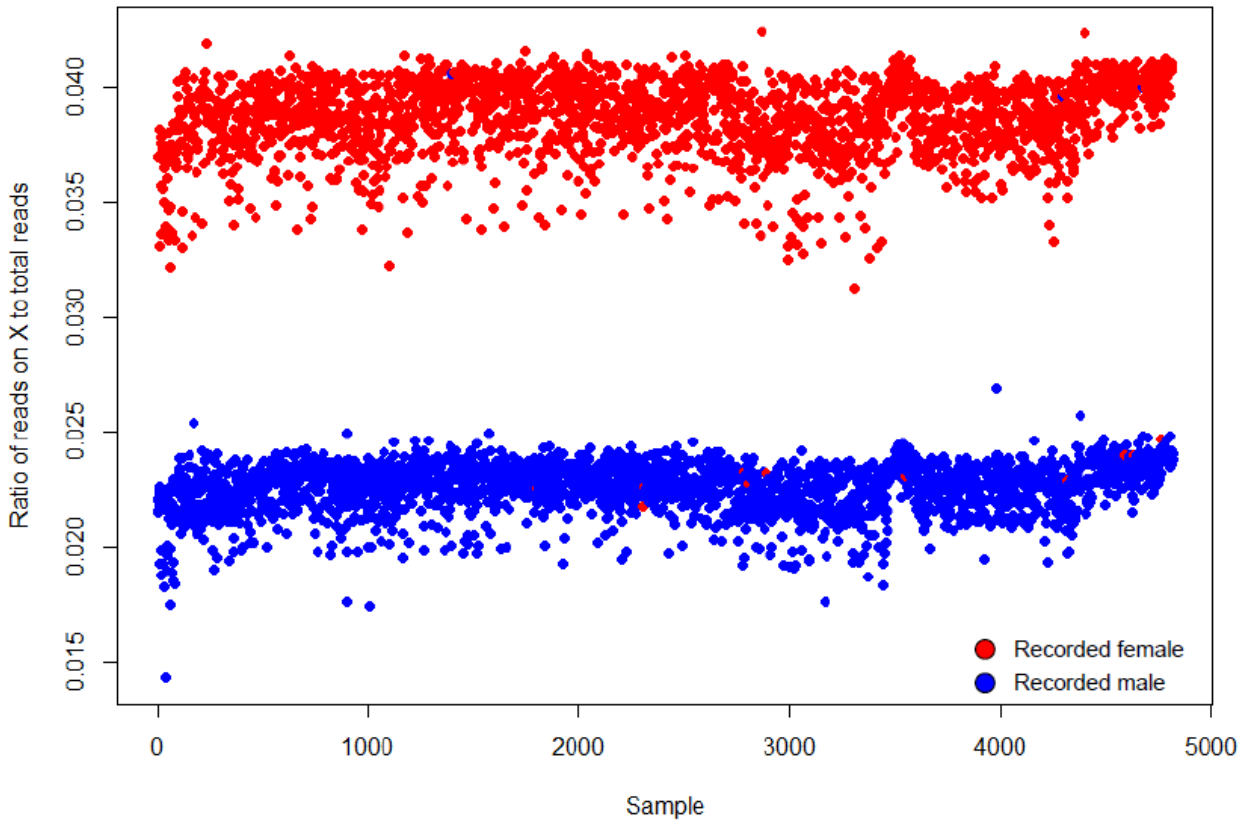
Zhang, Peng, Xiaowei Zhan, Noah A. Rosenberg, and Sebastian Zöllner. 2013. "Genotype Imputation Reference Panel Selection Using Maximal Phylogenetic Diversity." *Genetics* 195 (2): 319–30. https://doi.org/10.1534/genetics.113.154591.

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.

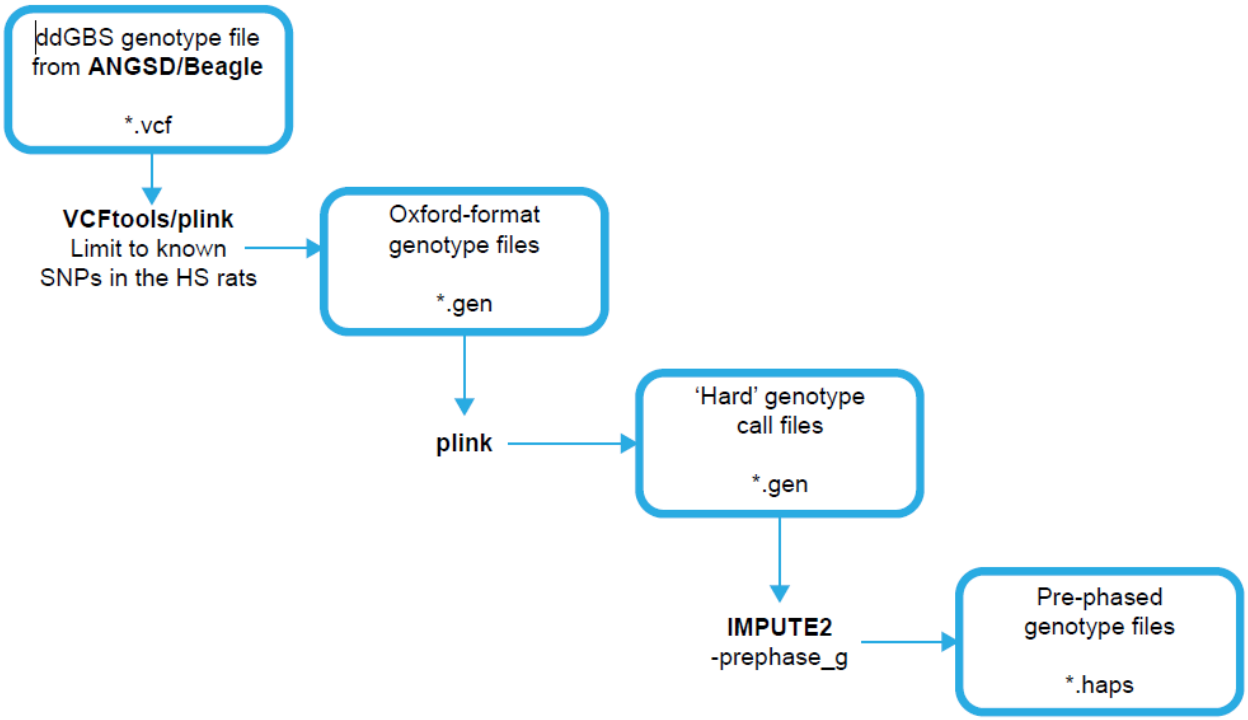
Figure S1. Ratio of reads on X-chromosome to total sequencing reads.

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.



859 **Figure S2. Data preparation workflow for imputation with IMPUTE2.**

860



861

862

863

864

865

866

867

868

869

870

871

872

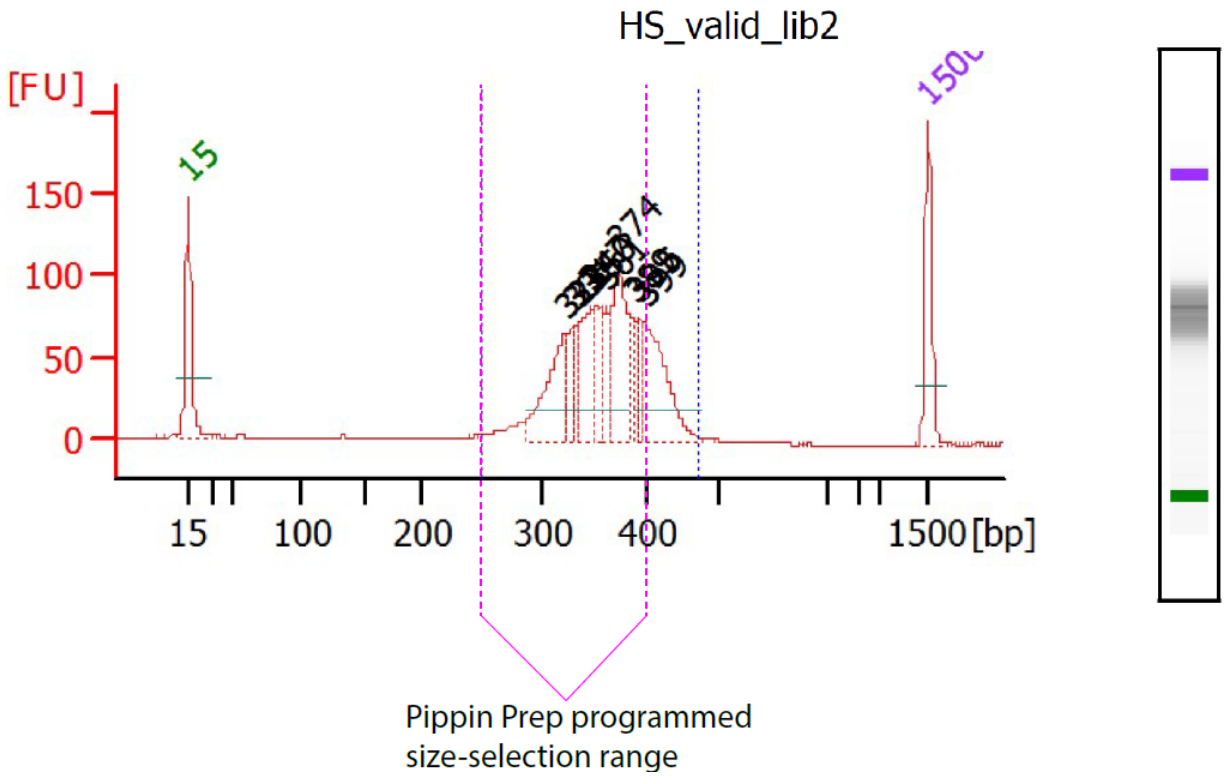
873

874

875

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.

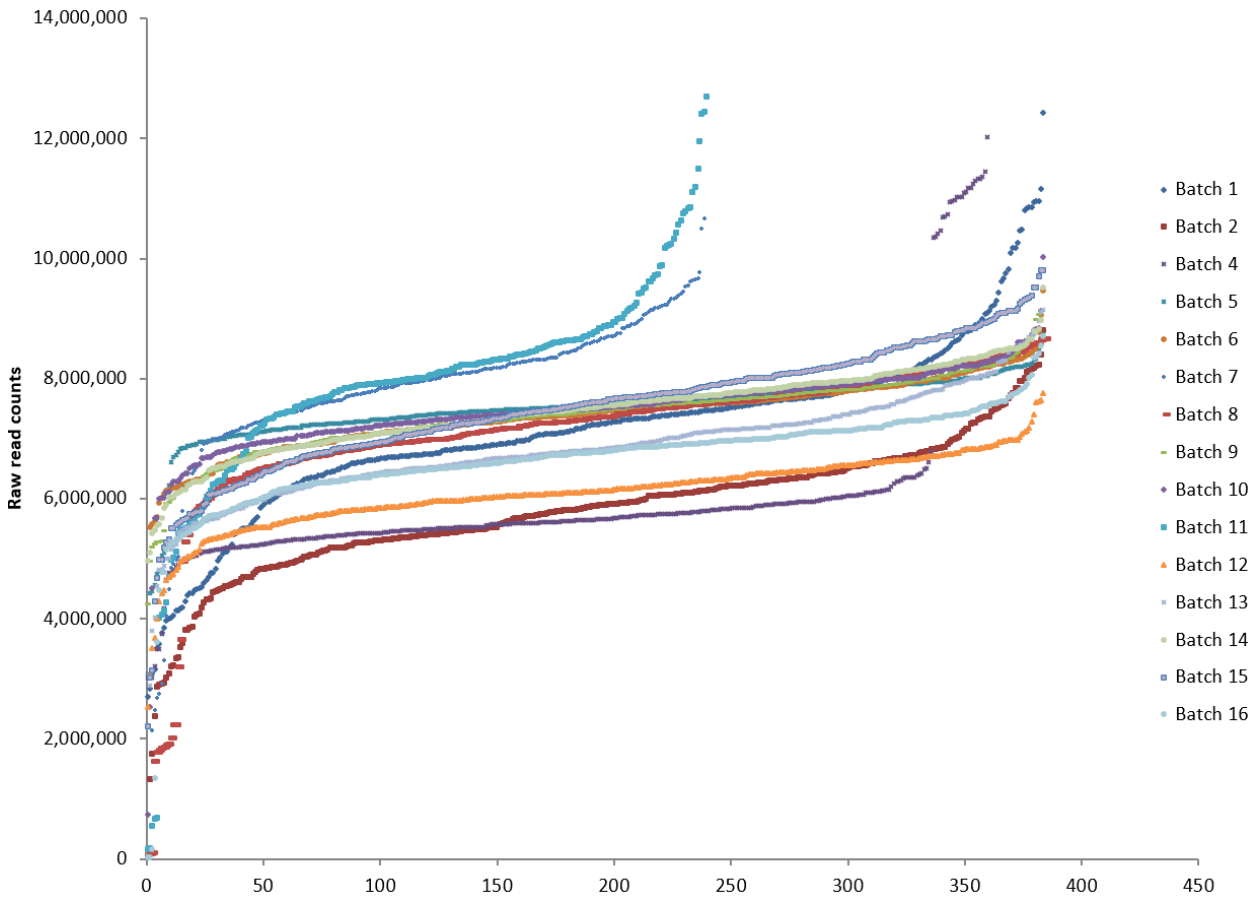


894

895 **Figure S4. Raw read counts grouped by shipment batch.**

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

898



899

900

901

902

903

904

905

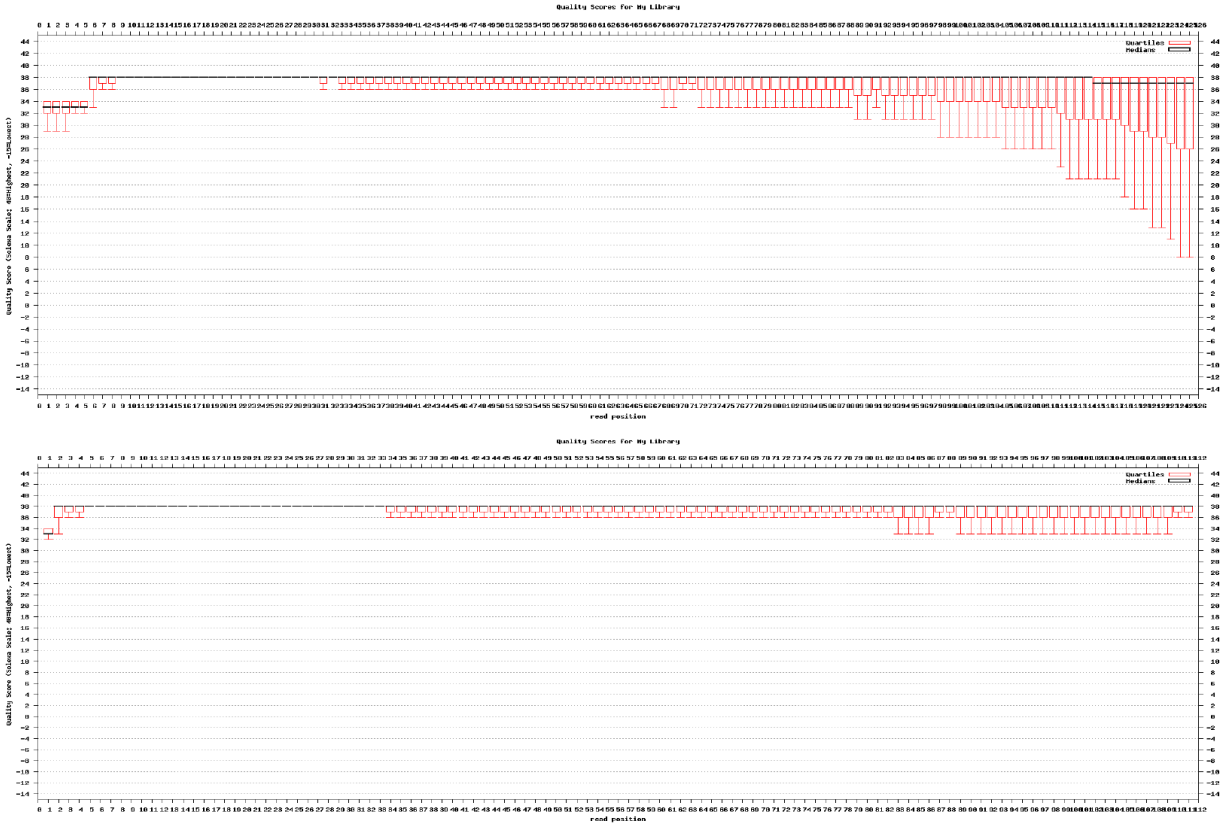
906

907

908

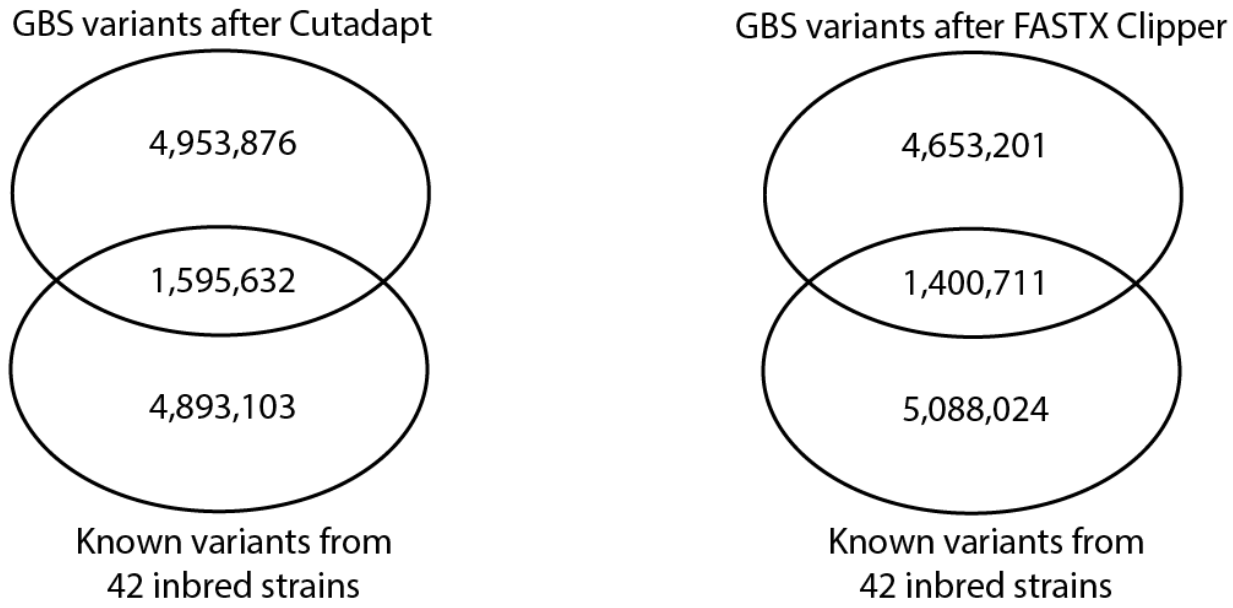
Figure S5. FASTQC results pre- and post-filtering with Cutadapt.

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.



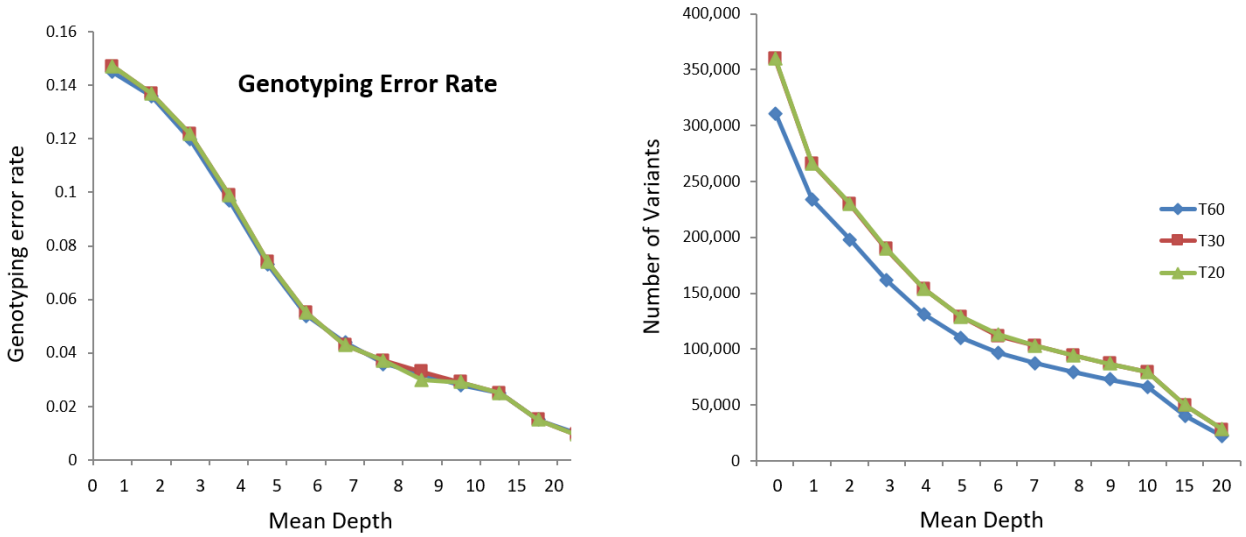
927
928 **Figure S6. Overlap of called SNPs with known variants after read trimming with FASTX or**
929 **Cutadapt.**

930



931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946

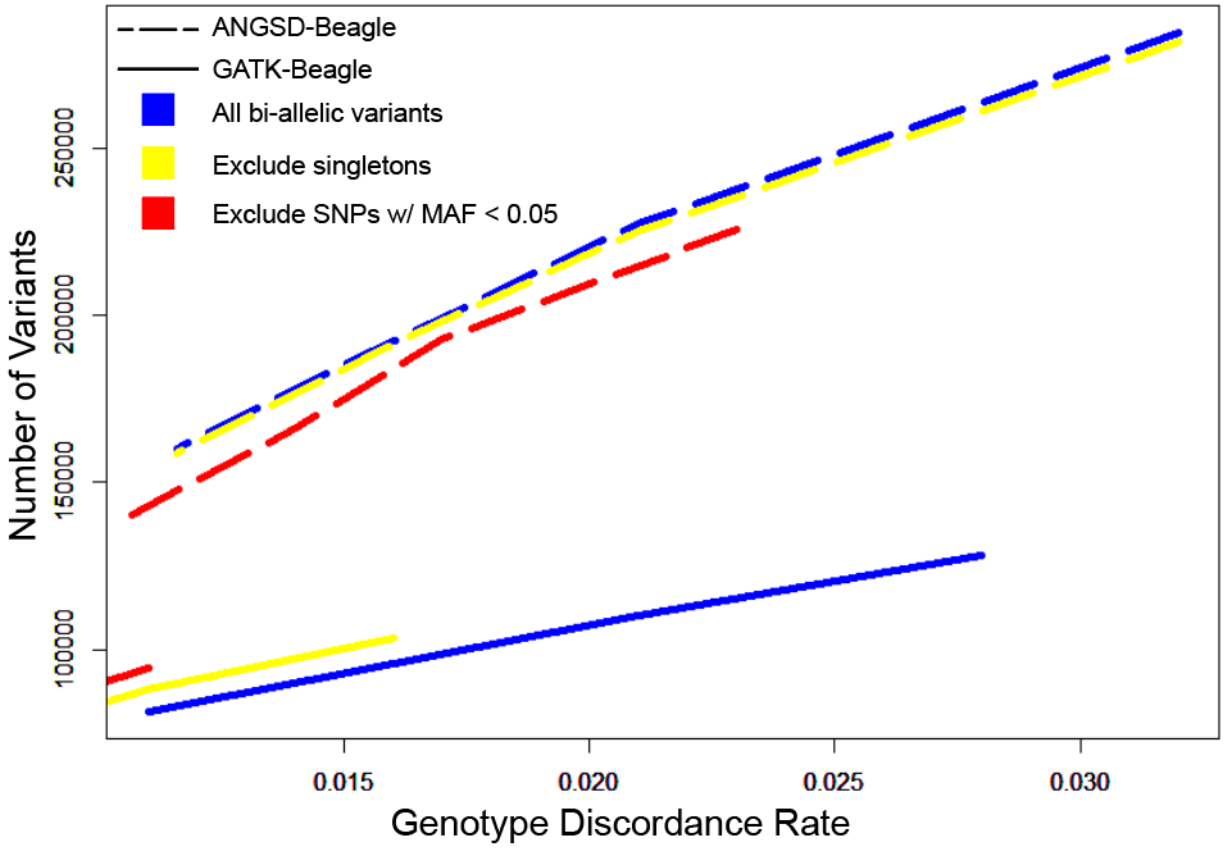
Figure S7. Mapping quality thresholds.
Genotyping error rate and number of variants by mean depth per sample per variant site for mapping quality thresholds of 20, 30, and 60.



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

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

978



979

980

981

982

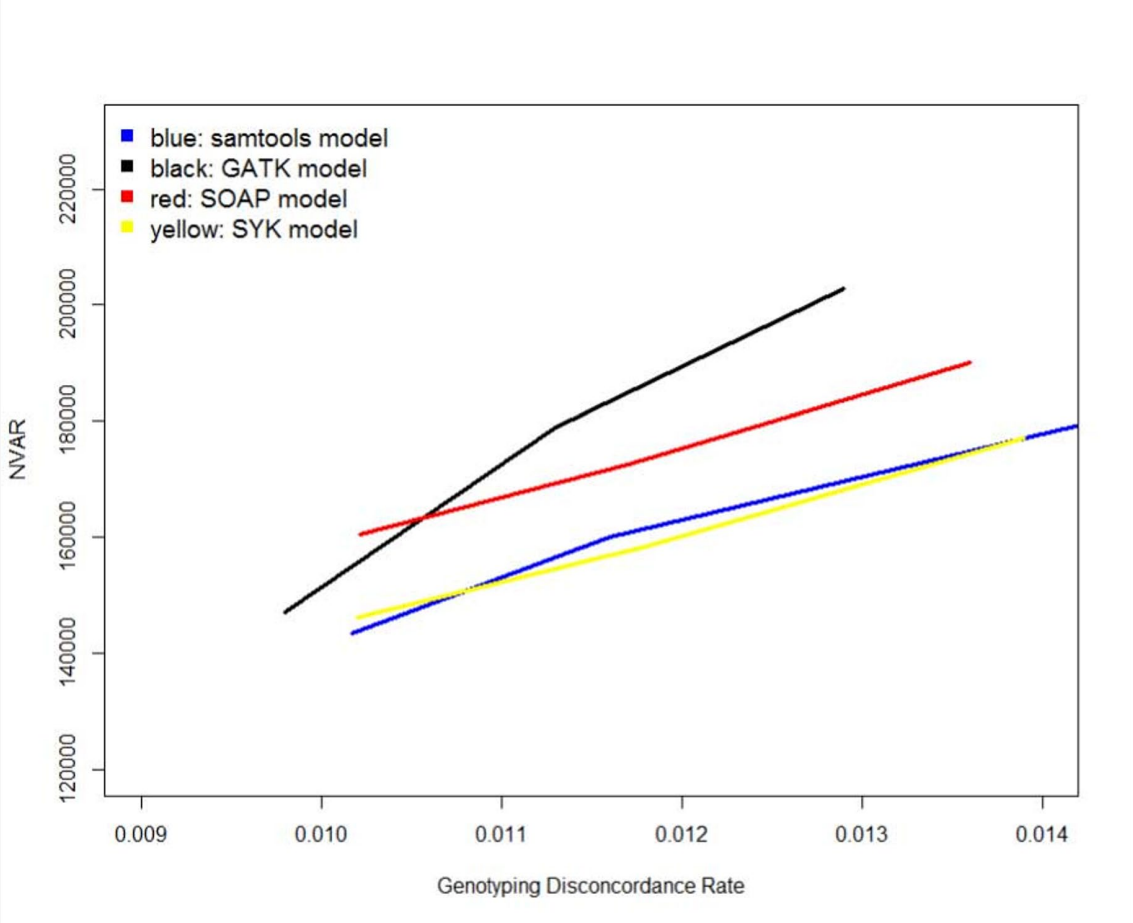
983

984

985

986

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



989

990

991

992

993

994

995

996

997

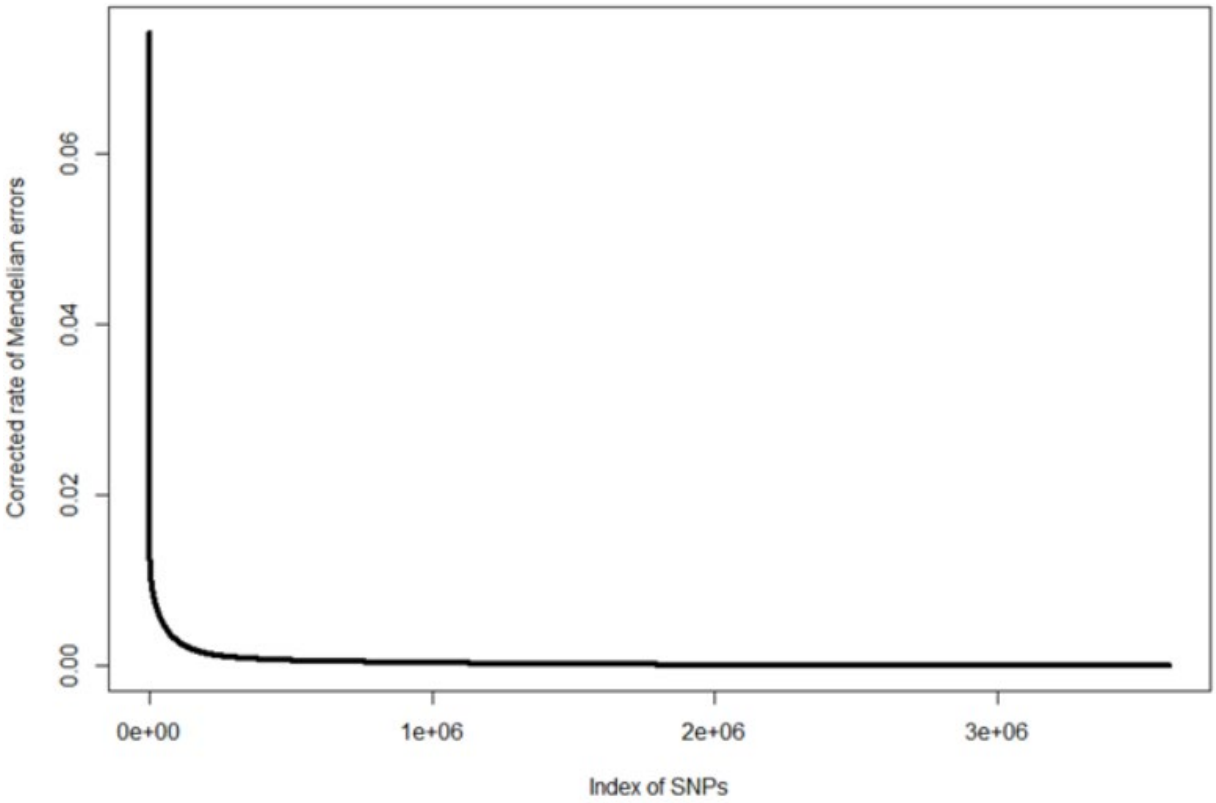
998

999

1000

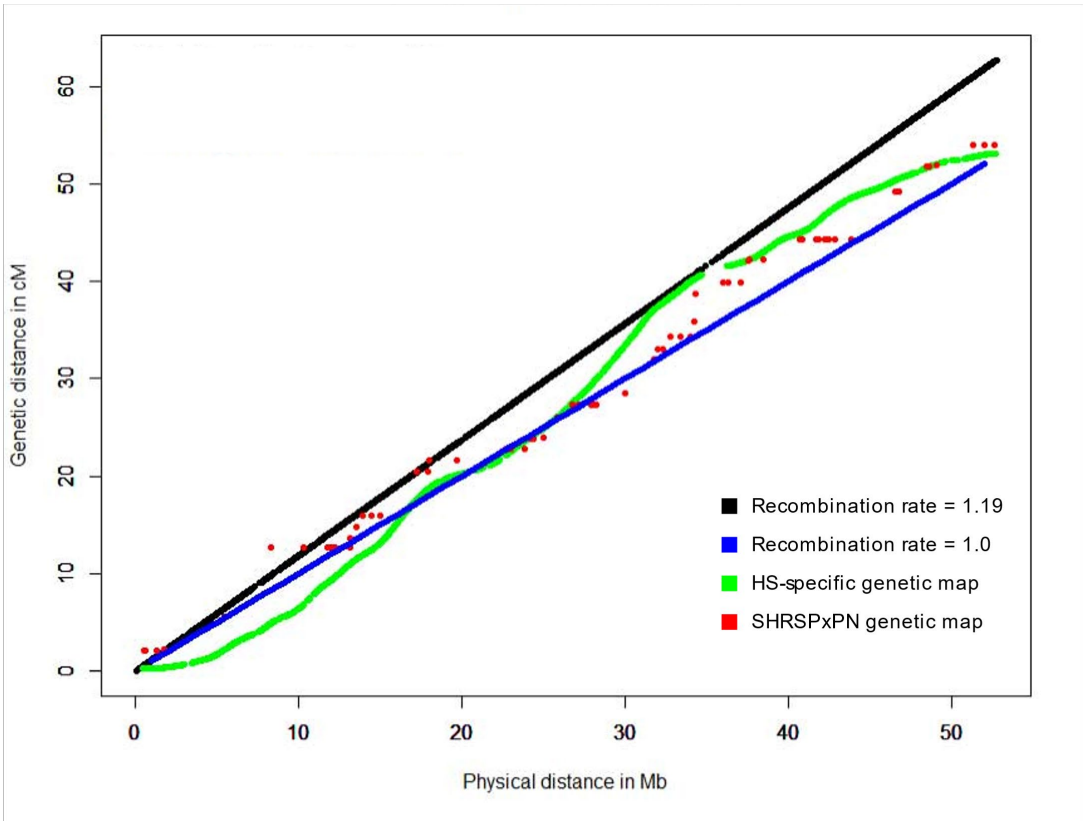
Figure S10. Mendelian error rates

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.



1017
1018 **Figure S11. Available rat genetic maps.**

1019 Plotted physical and genetic distances are for chromosome 12.



1020

1021

1022

1023

1024

1025

1026

1027

1028

1029

1030

1031 **Table S1. Demultiplexing performance.**

1032 All methods began with the same number of reads from the original FASTQ. Final read and base
1033 pair counts are from after the reads have been trimmed of adapter, barcode, and restriction site
1034 sequences, as well as low-quality base pairs (< Q20).

1035

	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

1036

1037

1038
1039

1040

1041 **Table S2. Comparison of variants calls after filtering with FASTX vs Cutadapt.**

1042 Data shown comes from the original set of 96 HS samples prepared in 12-plex and sequenced on
1043 the Illumina HiSeq 2500. At this early step of pipeline optimization, variants were called utilizing
1044 GATK UnifiedGenotyper.

1045

	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

1046

1047

1048

1049

1050

1051
1052

1053

1054 **Table S3. Variant metrics resulting from reads filtered at different mapping quality**
1055 **thresholds.**

1056 Data shown comes from the original set of 96 HS samples prepared in 12-plex and sequenced on
1057 the Illumina HiSeq 2500. Variants were called utilizing the SAMtools model and the -minMapQ
1058 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

1059

1060
1061

1062

1063 **Table S4. Transition/transversion ratio before and after known sites filtering.**

1064 The presented data comes from ANGSD-SAMtools/Beagle variant calls for 3,601 HS samples,
1065 prior to imputation with IMPUTE2. Known SNPs came from both the 42 inbred genomes from
1066 Hermsen et. al 2015 (Hermsen et al. 2015) and the 8 inbred HS founder strains sequenced by the
1067 University of Michigan (Ramdas et al. 2019).

1068

	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
$\mathbf{T_S}$	1,782,310	84,213
$\mathbf{T_V}$	63,750	34,014
$\mathbf{T_S T_V}$	27.96	2.48
Total # SNPs	1,846,060	118,227

1069

1070
1071

1072

1073 **Table S5. Imputation accuracy for chromosome 12 across different genetic maps.**

1074 The number of variants used for the concordance check is dependent on the overlap of the imputed
1075 variants with array data for the 96 HS rats with array genotypes. The MAF filter only removes
1076 monomorphic sites within the 96 HS rat sample used for the concordance check.

1077

	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

1078

1079

1080