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.

17

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

19

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

21

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

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

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

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

26

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. A few genotyping microarrays have been created for rats; however, they were expensive and are no longer in production. 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. 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. We include a detailed protocol to perform ddGBS in rats.

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 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

50 sequencing approaches that take advantage of LD structure have been previously described
51 (Miller et al. 2007; van Orsouw et al. 2007; Van Tassell et al. 2008; Baird et al. 2008; X. Huang
52 et al. 2009; Davey et al. 2011; Elshire et al. 2011; Poland et al. 2012; Peterson et al. 2012; Sun et
53 al. 2013; Scheben, Batley, and Edwards 2017). Thousands of SNPs can be identified in large
54 numbers of samples for a fraction of the price of other genotyping methods (Chen et al. 2013; He
55 et al. 2014). The advantages of these methods are especially attractive when applied to less
56 commonly utilized species or strains for which genotyping microarrays are not available.

57 Of the existing reduced-representation protocols, the genotyping-by-sequencing (**GBS**)
58 approach developed by Elshire et al. (Elshire et al. 2011) has been frequently modified to
59 accommodate non-model species, such as: soybean (Sonah et al. 2013), rice (Furuta et al. 2017),
60 oat (Fu and Yang 2017), chicken (Pértille et al. 2016; Wang et al. 2017), mouse (Parker et al.
61 2016), fox (Johnson et al. 2015), and cattle (De Donato et al. 2013), among others. The greatly
62 varying genomic composition among organisms necessitates a diverse and customized set of
63 approaches for obtaining high-quality genotypes. As such, both the GBS protocol and
64 computational pipeline require modifications when being applied to a new species. Recent work
65 from our group showed that GBS can be effectively applied to outbred mice (Parker et al. 2016;
66 Gonzales et al. 2017; Zhou et al. 2018) and rats (Fitzpatrick et al. 2013). However, those
67 publications used protocols that had not been optimized, leaving significant room for
68 improvement in genotype quality and marker density. Additionally, although several tools and
69 workflows for the analysis of GBS data have been described, including Stacks (Catchen et al.
70 2013), IGST-GBS (Sonah et al. 2013), TASSEL-GBS (Glaubitz et al. 2014), Fast-GBS
71 (Torkamaneh et al. 2017), and GB-eaSy (Wickland et al. 2017), the majority were developed and
72 optimized for use in plant species and given the lack of well-developed genomic resources in

these species, do not include an imputation step. 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 models, 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

94 Samples for this study originated from three sources: an inhouse advanced intercross line (**AIL**)
95 derived from LG/J and SM/J mice (Gonzales et al. 2018), Sprague Dawley (**SD**) rats from
96 Charles River Laboratories and Harlan Sprague Dawley, Inc. (Gileta et al. 2018), and an HS rat
97 colony (Woods and Mott 2017; Chitre et al. 2018). Early ddGBS optimization utilized AIL
98 genomic DNA extracted from spleen by a standard salting-out protocol. Later optimization steps
99 were performed using genomic DNA from SD rats extracted from tail tissue using the PureLink
100 Genomic DNA Mini Kit (Thermo Fisher Scientific, Waltham, MA). Optimization of the ddGBS
101 sequencing data analysis pipeline was performed on various subsets of our final sample of 4,973
102 HS rats. HS rat DNA was extracted from spleen tissue using the Agencourt DNAdvance Kit
103 (Beckman Coulter Life Sciences, Indianapolis, IN). All genomic DNA quality and purity was
104 assessed by NanoDrop 8000 (Thermo Fisher Scientific, Waltham, MA). Interestingly, we have
105 observed that rat genomic DNA appears to degrade faster than mouse genomic DNA following
106 extraction; therefore, we recommend storing rat genomic DNA at -20° and using it within
107 months of extraction whenever possible.

108 ***In silico* digest of rat genome**

109 We used *in silico* digests to aid in the selection of restriction enzymes, with the goal of
110 maximizing the portion of the genome captured at sufficient depth to make confident genotype
111 calls. We used the *restrict* function within EMBOSS (version 6.6.0) (Rice, Longden, and
112 Bleasby 2000) in conjunction with the REBASE database published by New England BioLabs
113 (NEB; version 808) (Roberts and Macelis 1999) to perform *in silico* digest of rn6, the current
114 release of the Norway brown rat reference genome. For the primary restriction enzyme, we chose
115 PstI. We performed the digest with PstI alone and then with PstI paired with each of 7 secondary
116 enzymes: AluI, BfaI, DpnI, HaeIII, MluCI, MspI, and NlaIII. We only considered fragments with

117 one PstI cut site and one cut site from the secondary enzyme because the adapter and primer sets
118 are designed to only allow these fragments to amplify. The final choice of enzyme (NlaIII) was
119 determined empirically and is detailed in the Results.

120 **Restriction enzyme selection**

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

132 **ddGBS library preparation and sequencing**

133 The full ddGBS protocol is available in File S1. In brief, approximately 1µg of DNA is used per
134 sample. Sample DNA, PstI barcoded adapters, and NlaIII Y-adapter are combined in a 96-well
135 plate and allowed to evaporate at 37ºC overnight. Sample DNA and adapters are re-eluted on day
136 two with a PstI/NlaIII digestion mix and incubated at 37ºC for two hours to allow for complete
137 digestion. Ligation reagents are then added and incubated at 16ºC for one hour to anneal the
138 adapters to the DNA fragments, followed by a 30-minute incubation at 80ºC to inactivate the

139 restriction enzymes. Sample libraries are purified using a plate from a MinElute 96 UF PCR
140 Purification Kit (QIAGEN Inc., Hilden, Germany), vacuum manifold, and ddH_2O. Once re-
141 eluted, libraries are quantified in duplicate with Quanit-iT PicoGreen (Thermo Fisher Scientific,
142 Waltham, MA) and pooled to the desired level of multiplexing (i.e. 12, 24, or 48 samples per
143 library). Pooled libraries are concentrated to obtain the desired volume for use in the Pippin Prep.
144 The concentrated pool is quantified to ensure the gel cassette will not be overloaded with DNA
145 (>5µg). The pool is then loaded into the Pippin Prep for size selection (300-450bps) using a 2%
146 agarose gel cassette on a Pippin Prep (Sage Science, Beverly, MA). Size-selected libraries were
147 then PCR amplified for 12 cycles to increase the quantity of DNA, concentrated, and checked for
148 quality on an Agilent 2100 Bioanalyzer with a DNA 1000 Series II chip (Agilent Technologies,
149 Santa Clara, CA)., Bioanalyzer results were used to assure sufficient DNA concentration and to
150 identify excessive primer dimer peaks.

151 An initial 96 HS samples were sequenced, 12 samples per library, at Beckman Coulter
152 Genomics (now GENEWIZ) on an Illumina HiSeq 2500 with v4 chemistry and 125bp single-end
153 reads. Subsequently, we began using a set of 48 unique barcoded adapters (File S2) to multiplex
154 48 HS samples per ddGBS library. Each library was run on a single flow cell lane on an Illumina
155 HiSeq 4000 with 100bp single-end reads at the IGM Genomics Center (University of California
156 San Diego, La Jolla, CA).

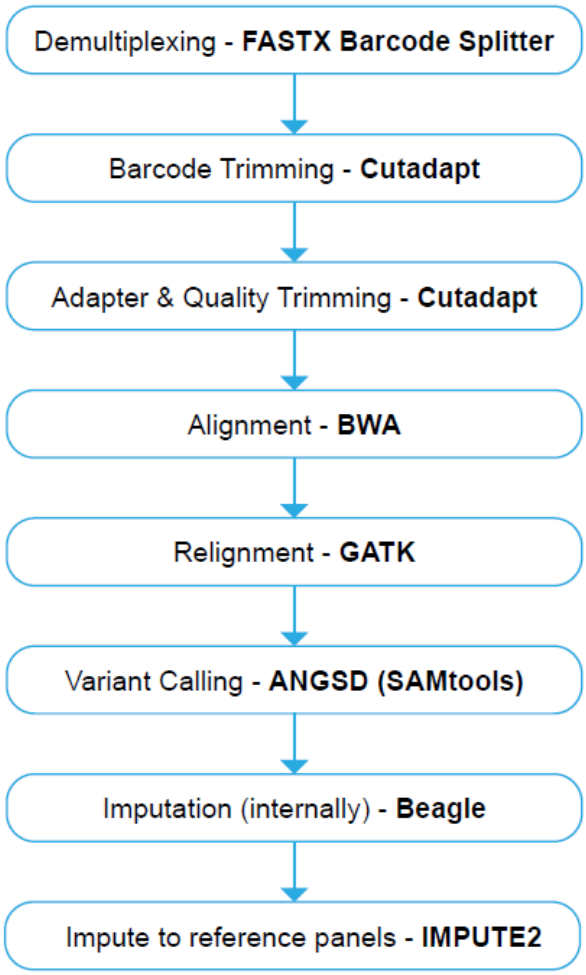
157

158

159

160

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



163
164

165 **Evaluation of ddGBS pipeline performance**

166 We present the steps required to call and impute genotypes from ddGBS data in Figure 2. During
167 optimization of the pipeline, performance was assessed by two primary metrics: (1) the number
168 of variants called and (2) genotype concordance rates for calls made in 96 HS rats that had both
169 ddGBS genotypes and array genotypes from the Axiom MiRat microarray. After each step of the
170 pipeline, the variant count and concordance rates were computed again. There were two
171 checkpoints in the GBS pipeline where genotype quality (measured by concordance rate) was
172 assessed: after internal imputation within Beagle (Browning and Browning 2009, 2016) and

173 again after imputation to the reference panel with IMPUTE2 (B. N. Howie, Donnelly, and
174 Marchini 2009; B. Howie et al. 2012). A third, additional metric we checked was the transition to
175 transversion ratio ($T_S T_V$), which is expected to be ~2.

176 **Demultiplexing**

177 The PstI adapter barcodes were used to demultiplex FASTQ files into individual sample files.
178 Three demultiplexing software packages were tested: FASTX Barcode Splitter (Hannon Lab
179 2010), GBSX (Herten et al. 2015), and an in-house Python script (Parker et al. 2016). Reads that
180 could not be matched with any barcode (maximum of 1 mismatch allowed), or that lacked the
181 appropriate enzyme cut site, were discarded. Samples with less than two million reads after
182 demultiplexing were discarded. Data concerning demultiplexing are shown in Table S1 are from
183 a single HS rat sequenced in a 12-sample library on one lane after demultiplexing and
184 adapter/quality trimming.

185 **Barcode, adapter, and quality trimming**

186 Read quality was assessed using FastQC v0.11.6 (Andrews 2017). We compared the efficacy of
187 two rapid, lightweight software options for trimming barcodes, adapters, and low-quality bases
188 from the NGS reads: Cutadapt (Martin 2011) and the FASTX Clipper/Trimmer/Quality Trimmer
189 tools (Hannon Lab 2010) (Table S2). A base quality threshold of 20 was used and reads trimmed
190 to below 25bps were discarded.

191 **Read alignment and indel realignment**

192 Rattus norvegicus genome assembly rn6 was used as the reference genome for read alignment
193 with the Burrows-Wheeler Aligner (BWA) (H. Li and Durbin 2009) using the *mem* algorithm.
194 We then used GATK IndelRealginer (McKenna et al. 2010) to improve alignment quality by

195 locally realigning reads around a reference set of known indels in 42 whole-genome sequenced
196 inbred rat strains, including the eight HS progenitor strains (Hermsen et al. 2015).

197 **Variant calling and imputation**

198 Variants were called using ANGSD v0.911, for which we employed the SAMtools model
199 (Korneliussen, Albrechtsen, and Nielsen 2014; Durvasula et al. n.d.). Beagle v4.1 (Browning and
200 Browning 2009, 2016) was used to impute missing genotypes at sites that had calls for only a
201 subset of individuals (no external reference panel was used). Before settling on the combination
202 of ANGSD and Beagle for genotype calling, we also tried GATK's UnifiedGenotyper and
203 HaplotypeCaller (McKenna et al. 2010) with various parameter settings, but we did not obtain
204 satisfactory performance. Using ANGSD, we inferred the major and minor alleles (-
205 *domajorminor* 1) from the genotype likelihoods, kept only high confidence polymorphic sites (-
206 *snp_pval* 1e-6), and estimated the allele frequencies based on the inferred alleles *(-domaf* 1). We
207 discarded sites missing read data in more than 4% of samples (*–minInd*). Additionally, we tested
208 altering thresholds for minimum base quality score (-*minQ*) and mapping quality (-*minMapQ*).

209 **QC and pre-phasing for imputation using an external reference panel**

210 Prior to imputing SNPs from the 42 inbred strain reference panel (Hermsen et al. 2015), we
211 checked concordance rates for the 96 HS animals with array genotypes, examined the $T_S T_V$ ratio,
212 and assessed whether the sex as recorded in the pedigree records agreed with the sex empirically
213 determined by the proportion of reads on the X-chromosome out of the total number of reads
214 (Figure S1). We also identified Mendelian errors using the --mendel option in *plink* and known
215 pedigree information for 1,136 trios from 214 families within the HS sample. Using the fraction
216 of the trios that were informative for a given SNP and the formula $1-(1-2p(1-p))^3$, where p

217 represents the minor allele frequency of the allele, we formed curves for the distributions of the
218 expected number of Mendelian errors for both SNPs and samples and chose the inflection points
219 as thresholds for the number of Mendelian errors allowed. Lastly, we only retained variants
220 previously identified in the 8 HS founders (Ramdas et al. 2018) for use in imputation to the 42
221 genomes (Hermsen et al. 2015).

222 To improve imputation accuracy and computational efficiency, we employed a pre-
223 phasing step prior to reference imputation. A flowchart outlining the pre-phasing protocol is
224 presented in Figure S2. We tested three methods of converting genotype data to phased .hap
225 files: VCFtools (*--impute*) (Danecek et al. 2011), SHAPEIT (Delaneau, Marchini, and Zagury
226 2011), and IMPUTE2 (*-prephase_g*) (B. N. Howie, Donnelly, and Marchini 2009).

227 **Genetic maps**

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

235 **Imputation to reference panel**

236 We used existing sequencing and array data from the HS rat founder and other inbred laboratory
237 rat strains (Hermsen et al. 2015) as reference panel for imputation. Genotype data underwent QC
238 and were phased by Beagle into single chromosome haplotype files. Haplotype files were then

239 created using the workflow detailed in Figure S2. Imputation by IMPUTE2 was performed in
240 5Mb using the aforementioned reference panels and genetic maps.

241 **Data availability**

242 Data Genotype data will be available at https://doi.org/10.6084/m9.figshare.7597697 and the
243 code necessary to run the steps outline in the publication provided here
244 https://doi.org/10.6084/m9.figshare.7597694. Additional data is available upon request.

245 RESULTS

246 **ddGBS optimization**

247 Previous projects utilizing GBS in mice and rats (Fitzpatrick et al. 2013; Parker et al. 2016;
248 Gonzales et al. 2018) often encountered an issue where certain regions of the genome
249 experienced high pileups of reads per sample (>100x), while other regions were covered by just
250 1-2 reads. This read distribution imbalance can be caused in part by PCR amplification bias,
251 where a subset of fragments are preferentially amplified until they dominate the final library
252 (Kanagawa 2003; Aird et al. 2011). These previous protocols utilized 18 cycles of amplification.
253 We tested reducing this to 8, 10, 12, or 14 cycles and found that only 12 cycles were necessary to
254 yield sufficient PCR product for sequencing; the reduction in the number of PCR cycles was
255 expected to reduce PCR bias.

256 Another concern regarding previous sequencing results was an excess of long fragments
257 (>700bps as determined by *in silico* digest), which do not provide sufficient reads to make
258 confident genotype calls (< 5 reads per sample) and are therefore wasteful. We tested multiple
259 methods of combatting this issue, including: reducing the PCR extension time, increasing the

260 digestion length or enzyme concentration, increasing the selectivity of the primers, performing
261 size selection on the libraries, or using a two-enzyme restriction digest.

262 We began by testing shorter extension times (15s/20s vs. 30s) to try and reduce the
263 average fragment size of our sequencing libraries by preventing fragments with larger inserts
264 from completing amplification. However, we did not see noticeable improvement in the final
265 library composition at shorter extension times and determined the original length of 30 seconds
266 was appropriate.

267 We also considered the possibility that the restriction enzyme digests might not be
268 running to completion. To address this possibility, we increased the length of the digestion from
269 2 hours to 3 or 4 hours. We also tried increasing the number of units of PstI enzyme added, to
270 ensure complete digest. Neither of these modifications impacted the final fragment length
271 distribution of the library, indicating that the digest was reaching completion after 2 hours using
272 the original concentration of PstI.

273 We also explored reducing the complexity of our library as a means of obtaining more
274 reads at fewer sites. We accomplished this using the methods described by Sonah et al. 2013
275 (Sonah et al. 2013). This method involved adding an additional 1-2 nucleotides to the sequencing
276 PCR primers (A or AG). The addition of these nucleotides increases the selectivity of the primers
277 and reduces the number of amplifiable fragments, sacrificing sequencing breadth for read depth.
278 However, this did not ameliorate the issue of reads being wasted on long fragments but did
279 appear to increase the PCR bias (data not shown). Therefore, we chose to retain our original
280 primer design.

281 Our previous GBS protocol did not have an explicit library fragment size selection step.
282 The final library was purified using a MinElute PCR Purification Kit (QIAGEN Inc., Hilden,
283 Germany), which isolates PCR products 70bp-4kb in length, leaving a wide range of fragment
284 sizes in the final library, under the assumption that only shorter fragments would bridge amplify
285 on the flow cell. This method was imprecise and had low reproducibility, negatively impacting
286 our ability obtain reads at consistent sites across libraries. Rather than attempt size selection by
287 gel extraction, we chose to utilize a Pippin Prep, which automates the elution of DNA libraries of
288 desired fragment size ranges. By using this automated size selection, we reduced the proportion
289 of the genome targeted for sequencing, and because restriction enzymes make the consistent cuts
290 across samples, ensure the same fragments are sequenced in the majority of libraries. Since the
291 clustering process involves a bridge amplification step that preferentially amplifies library
292 fragments with shorter insert sizes (Illumina, Inc. 2014), we kept the size selection window
293 narrow (250-400bps) to avoid introducing a bias in which fragments were sequenced.

294 To increase the proportion of the genome captured within the fragment size window, we pursued
295 a double digest of the genome using a secondary enzyme with a more frequently occurring
296 recognition sequence. When used alone, *in silico* digest of the rn6 reference genome by PstI
297 (Figure 1; Table 1) showed that only ~0.5% of the genome would have fallen within a 150bp
298 fragment size window selected on the Pippin Prep. Given the LD structure of other outbred
299 rodent lines such as DO and CFW mice (Parker et al. 2016), we speculated this small fraction
300 would not yield sufficient SNPs to tag the majority of variation in the genome. Additionally, we
301 were concerned about potential biases in coverage, heterozygosity, and the minor allele
302 frequency (**MAF**) spectrum that may be introduced by incomplete capture of the genome
303 (Flanagan and Jones 2018).

304 The number of fragments with one of each of the cut sites were summed for all observed
305 lengths and the results summarized in Figure 1 and Table 1. BfaI, MluCI, and NlaIII were
306 chosen for further testing due to their compatibility with PstI digestion reagents and
307 temperatures, sticky ends, and the proportion of the genome falling in the size selection window.
308 We ruled out BfaI because it only had a 2bp overhang after cleavage, which led to a high
309 concentration of adapter dimer in the sequencing libraries. NlaIII was chosen because it
310 contained the greatest portion of the genome in the size selection window.

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%

311

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

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

315

316 * Restriction enzyme is methylation sensitive.

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

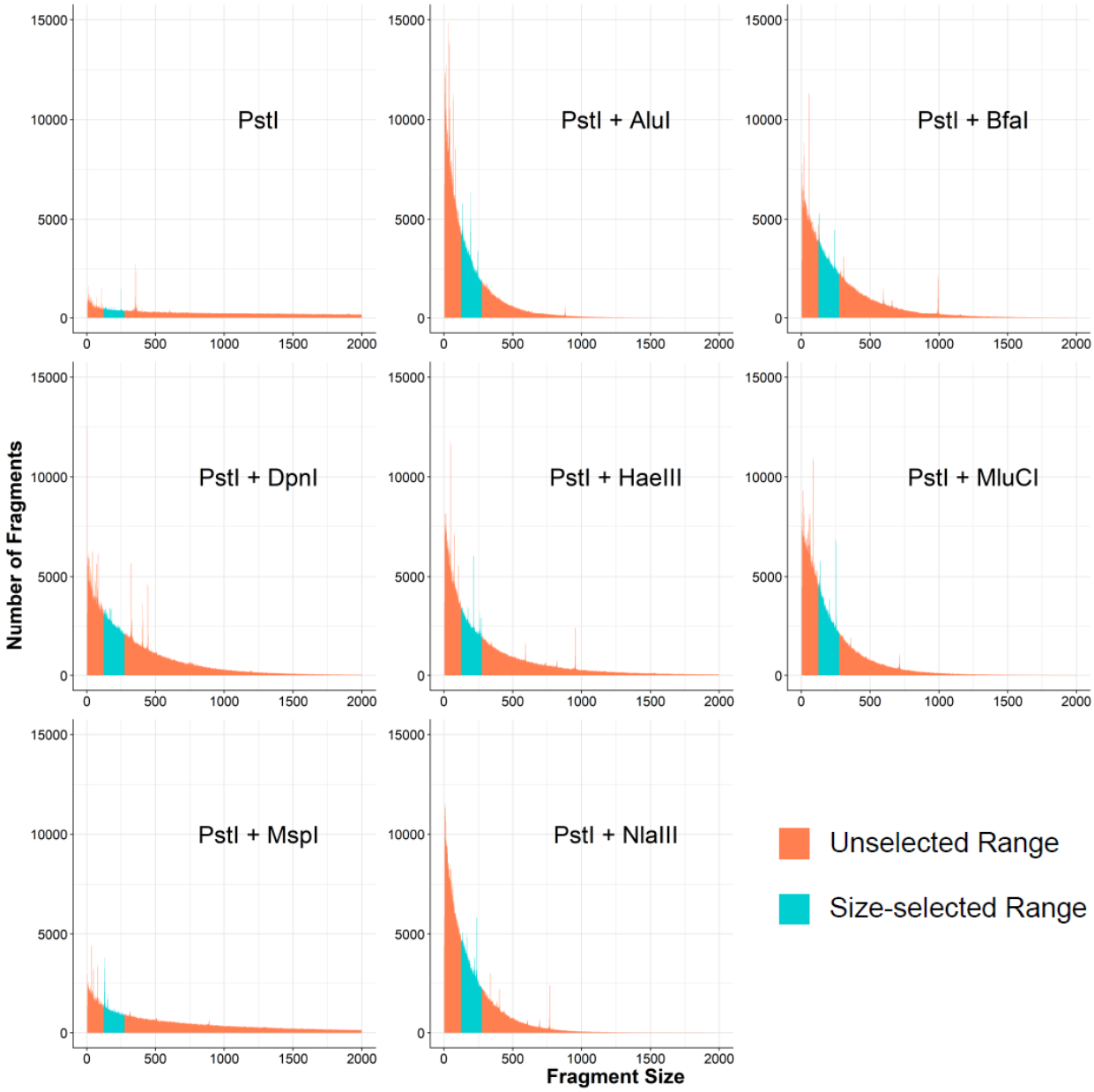
318

319

320

321

322 **Figure 1. *In silico* digest fragment distributions for PstI and potential secondary restriction**
323 **enzymes.**



324

325

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

331

332

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

343 We tested numerous quantities of PstI and NlaIII adapters in an attempt minimize the
344 amount used and avoid adapter dimers in the final libraries. For the barcoded PstI adapters, we
345 tested 120pmol, 40pmol, 20pmol, 4pmol, 3pmol, 1.5pmol, 0.6pmol, and 0.2pmol and for the
346 NlaIII Y-adapter, 16pmol, 8pmol, 4pmol, 2pmol, 1pmol, and 0.5pmol. We found that 0.2pmol of
347 PstI adapter and 4pmol of NlaIII Y-adapter yielded sufficient library and avoided adapter dimers.

348 We sequenced a trial flow cell with 8 pooled ddGBS libraries of 12 SD rat samples each
349 (96 total) on a HiSeq 2500 (Illumina, San Diego, CA) with 125bp reads and v3 chemistry,
350 obtaining an average of 15.3 million reads per sample. Given the NlaIII *in silico* digest results
351 suggested we were capturing ~3.4% of the genome and that we were using 125bp reads, this
352 corresponded to approximately 20x coverage of captured sites. We subsequently increased the
353 number of samples to 48 per library for the HS rats because we hypothesized 5x would be
354 sufficient coverage per sample when utilizing imputation to a reference panel. We also
355 discovered that a portion of the reads contained sequence fragments of the NlaIII adapter

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

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

364 **Demultiplexing**

365 The number of base pairs of sequencing data retained after demultiplexing was fairly consistent
366 across demultiplexing software (Table S1). We ultimately decided to use FASTX Barcode
367 Splitter because it yielded the greatest number of reads after quality/adapter trimming and had
368 faster run times, though this may be predominantly attributable to outputting uncompressed
369 FASTQ files. An average of 330 million 100bp reads were obtained per library, resulting in ~7
370 million reads per sample. Figure S4 shows the distribution of reads counts for all samples after
371 demultiplexing, but prior to sample filtering based on read count.

372 **Adapter and quality trimming**

373 Read quality was substantially elevated after trimming the barcode and adapter sequences and
374 low-quality base pairs at the ends of reads (Figure S5). Overall read counts were only marginally
375 reduced by quality trimming (Table S1). We observed that the number of called variant sites and
376 the genotyping rate were both greater when using reads initially processed by Cutadapt (Martin,
377 2011) than reads processed by the FASTX_Toolkit (Table S2). Importantly, a large portion of

378 the additional identified variants were known variant sites from the 42 inbred strains reference
379 set (Figure S6), indicating the elevated call rate was at least in part due to capturing more true
380 variant sites. We viewed this as sufficient support for proceeding with Cutadapt for quality
381 trimming.

382 **Read alignment quality**

383 The number of called variants and genotype call rates were identical at read mapping quality
384 (mapQ) thresholds of either 20 or 30 (Table S3) within ANGSD. As the ANGSD mapQ
385 threshold was raised to 45, there was a small reduction in the number of called variants, and then
386 much greater losses at thresholds of 60 or 90. Fortunately, genotype concordance rates at both
387 low and high mapQ thresholds were stable, despite the putatively higher quality of the
388 alignments (data not shown). This permitted us to select a lower mapQ threshold (mapQ = 20),
389 maximizing the number of variants called without sacrificing genotype quality.

390 **Variant calling**

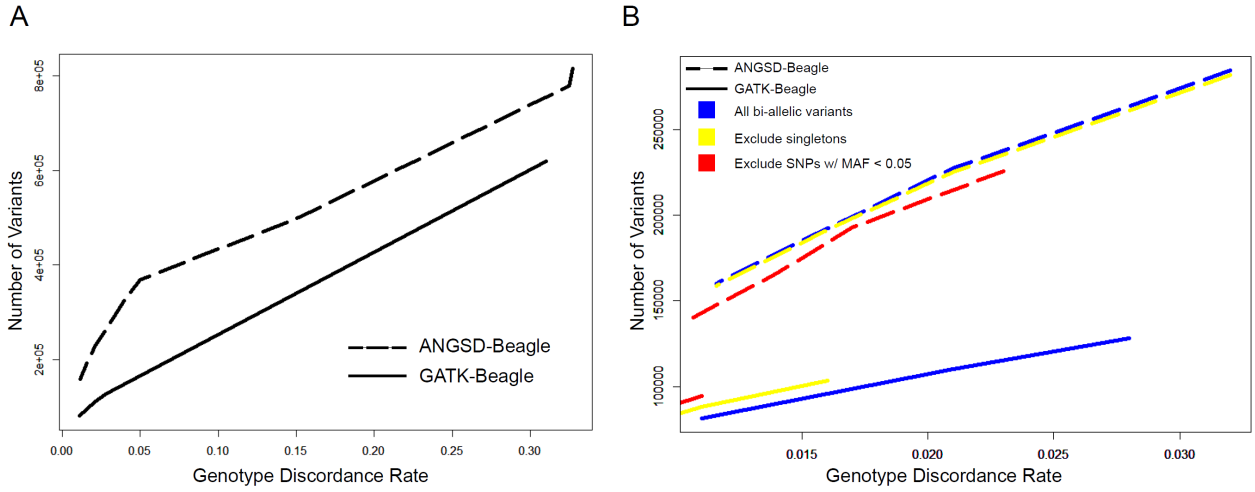
391 Neither GATK's UnifiedGenotyper nor HaplotypeCaller (McKenna et al. 2010; DePristo et al.
392 2011) called an adequate number of SNPs for fine-mapping in the HS. Panel A of Figure 3
393 shows that the combination of HaplotypeCaller and Beagle had lower rates of genotype
394 concordance with array data than calls made by ANGSD and Beagle. We tried several
395 adjustments to the GATK default thresholds for variant quality, heterozygosity, graph pruning,
396 dangling branch length, and number of reads sharing an alignment start to declare an active
397 region; however, none of these changes significantly improved the results (data not shown). As
398 shown in Panel B of Figure 3, ANGSD/Beagle yielded a substantially greater number of SNPs at

399 all genotyping error rate thresholds. This observation held when variants were limited only to

400 biallelic sites and SNPs with an MAF > 0.05.

401

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



404

405 The two panels compare the number variants called by combination of ANGSD and Beagle or
406 GATK HaplotypeCaller and Beagle at various thresholds of genotype discordance with array
407 data. Calls were made using the 96 HS rats with array data. (A) The x-axis represents the
408 genotype discordance rate thresholds and the y-axis is the number of variants that surpass that
409 threshold for each genotype calling method. (B) Additional filters were applied to the original
410 SNP sets and the plot zooms in on a smaller range of acceptable discordance rates. Blue lines
411 represent the unfiltered SNP set. Yellow lines have been filtered for singletons. Red lines have
412 further excluded SNPs with an MAF < 0.05. Each line contains the same number of points.

413

414 ANGSD supports four different models for estimating genotype likelihoods: SAMtools,

415 GATK, SOAPsnp and SYK. We compared the methods to determine which produced the most

416 SNPs with the lowest error rates. The SOAPsnp model demonstrated an advantage in genotype

417 accuracy and number of variants called post-imputation with Beagle (Figure S7). However,

418 SOAPsnp requires considerably more computational resources, outweighing its marginal benefits

419 when applied to a large sample set (>2,000). The GATK model showed a greater number of

420 variants for more lenient discordance rate threshold, but as stringency increased, the number of
421 variants converged across the remaining 3 models. We proceeded with the SAMtools model for
422 genotype likelihood estimation due to its previous support in the GBS literature (Torkamaneh et
423 al. 2017), accepting a nominal decrease in highly concordant variants for a large gain in
424 efficiency.

425 **Imputation to reference panel**

426 Imputation is use in two ways in our protocol. As described above, we use imputation to assign
427 missing genotypes at SNPs called in some but not all individuals. In addition, we use imputation
428 in this section to call genotypes at site where GBS did not provide a call for any individuals.
429 Thus, our second application (described here) is similar to the human genetics application in
430 which imputation using 1000 Genomes increases the number of SNPs beyond those included on
431 a given microarray platform.

432 Before starting this imputation step, we observed an inflated transition/transversion ratio
433 (Table S4) in our ANGSD/Beagle SNPs. This issue was ameliorated when filtered the SNP set
434 for only "known" variants that were previously identified in either the 42 inbred strains
435 (Hermsen et al. 2015) or the 8 deep-sequenced HS founders (Ramdas et al. 2018). For
436 imputation, we therefore only provided IMPUTE2 with previously identified variant sites from
437 our ANGSD/Beagle output. Prior to running IMPUTE2, we also filtered the variants for biallelic
438 sites with a genotype call in more than one individual. Using pedigree data for the HS rats, we
439 further removed samples and SNPs showing uncharacteristically high levels of Mendelian errors.
440 There were 4 samples and 4,179 SNPs removed from subsequent analyses. Lastly, we removed
441 any samples where the sex chromosome read ratio was contrary to their reported sex (Figure S1).

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

462
463 **Table 2. Imputation accuracy based on different variant reference panels for IMPUTE2.**

464 The table includes six different possible reference panels for imputation. The 42 inbred strains,
465 34 non-founder inbred strains, and 8 HS founders from the 42 inbred strains all were derived
466 from Hermsen et al. 2015 (Hermsen et al. 2015). The UMich 8 HS founders were obtained from
467 Ramdas et al. 2018 (Ramdas et al. 2018). The final set of 8 HS founder was taken from Baud et
468 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
UMich 8 founders + 42 inbred strains	**Discordance rate**	0.007	0.011
	# Variants	864,670	898,621
	Genotyping Rate	0.63	0.64
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

469

470

471 IMPUTE2 requires a genetic map. As described in the methods section, we considered
472 four different genetic maps, two that were empirically derived and two that were linear
473 extrapolations based on the physical map. All genetic map performed similarly (Table S5).
474 Surprisingly, the linear genetic maps performed just as well as the HS-specific map (Littrell et al.
475 2018). Thus, for simplicity, we chose to use the chromosome-specific values initially published
476 by Jensen-Seaman (Jensen-Seaman 2004).

477 Pre-phasing with VCFtools and IMPUTE2 performed similarly, therefore we used
478 IMPUTE2. SHAPEIT was not used because it showed a higher rate of genotype discordance
479 compared to VCFtools and IMPUTE2 (data not shown).

480 To obtain our final set of ~3.7 million variants, a final round of variant filtering was
481 performed after imputation to the 42 strain reference panel. We removed SNPs with MAF <
482 0.005, a post-imputation genotyping rate < 90%, and SNPs that violated HWE with $p<1x10^{-10}$.

483 DISCUSSION

484 The use of microarrays and WGS for genotyping large samples in model organisms remains
485 cost-prohibitive. There is therefore an urgent and wide-spread need for a high-performance and
486 economical methods of obtaining genome-wide genotype data. While reduced-representation
487 approaches have been utilized in numerous species of plants and animals, including rodents
488 (Peterson et al. 2012; Fitzpatrick et al. 2013; Parker et al. 2016; Gonzales et al. 2017; Zhou et al.
489 2018), there has yet to be a published protocol optimized specifically for rats. Prior to
490 sequencing thousands of HS samples with GBS for our mapping efforts, we wanted to ensure we
491 were capturing the greatest possible number of high-quality variants at the lowest possible cost.
492 The protocol we present here is the culmination of careful testing and optimization of each step

493 of the GBS protocol for rats. We have now applied the approach to 4,973 HS rats, as well as
494 4,608 Sprague Dawley rats (Gileta et al. 2018).

495 Our previous GBS protocol (Parker *et al*, 2016), which was designed for use with CFW
496 mice, was unsuitable for our current genotyping efforts in HS rats, due to the much higher levels
497 of genetic diversity in the HS population. There are multiple reasons we chose to develop our
498 own computational pipeline for GBS rather than using existing workflows. Foremost, the
499 prominent GBS analysis pipelines were developed and optimized for use in crop species (Sonah
500 et al. 2013; Catchen et al. 2013; Glaubitz et al. 2014; Torkamaneh et al. 2017; Wickland et al.
501 2017), which are polyploid and have differing levels of variation and LD than outbred rodent
502 populations. Additionally, there were elements of each pipeline that did not meet our needs or
503 lacked customizability. For instance, TASSEL-GBS v2 (Glaubitz et al. 2014) trims all reads to
504 92 base pairs; however, other projects underway in our lab utilized up to 125bp reads, leading to
505 a ~20% reduction in data. TASSEL-GBS also ignores read base quality scores, which are
506 informative in probabilistic frameworks for estimating uncertainty in alignments and variant
507 calls (H. Li, Ruan, and Durbin 2008; DePristo et al. 2011; Nielsen et al. 2011), and uses a naïve
508 binomial likelihood ratio method for calling SNPs. Stacks has previously shown poor
509 demultiplexing (Herten et al. 2015; Torkamaneh et al. 2017) and does not make use of the
510 reference genome for priors when calling SNPs (Catchen et al. 2013). Fast-GBS relies on
511 Platypus for variant calling (WGS500 Consortium et al. 2014; Torkamaneh et al. 2017), which
512 employs a Bayesian method of constructing candidate haplotypes that works poorly with low-
513 pass sequencing data and does not scale well (Z. Li, Wang, and Wang 2018). Lastly, none of
514 these pipelines included an imputation step, which is crucial for filling in missing genotypes in

515 GBS data and can provide millions of additional SNPs given an appropriate composite reference
516 panel (B. Howie, Marchini, and Stephens 2011; G.-H. Huang and Tseng 2014).

517 Though we have not explicitly tested each alternate GBS pipeline for the purposes of this
518 publication, this has been recently done by Wickland et al. (Wickland et al. 2017). Their pipeline
519 GB-eaSy, which ours most closely resembles, was found to be superior by a number of metrics
520 to Stacks, TASSEL-GBS, IGST, and Fast-GBS. Similar to GB-eaSy, our pipeline utilizes a
521 double-digest GBS protocol, aligns reads to the reference genome with *bwa-mem*, and uses the
522 SAMtools genotype likelihood model for calling SNPs (H. Li 2011). The combination of bwa-
523 mem and SAMtools algorithm was independently shown to have the best performance for calling
524 SNPs from Illumina data (Hwang et al. 2015), further supporting our choice of these programs
525 for read alignment and variant calling. Additionally, using the ANGSD wrapper provided us with
526 the ability to convert the posterior genotype probabilities into genotype dosages for mapping
527 studies (Korneliussen, Albrechtsen, and Nielsen 2014).

528 A minor difference between GB-eaSy and our pipeline is the use of Cutadapt (Martin
529 2011) rather than GBSX (Herten et al. 2015) for demultiplexing, though both performed equally
530 well (Table S1). The primary improvement is our extension of the pipeline with the
531 implementation of effective internal and reference-based imputation steps using the 42 inbred rat
532 genomes (Hermsen et al. 2015) and 8 deep-sequenced HS founders from UMich (Ramdas et al.
533 2018). There are two stages of imputation in our pipeline: the first one is accomplished by
534 Beagle, and aims to fill in missing genotypes at called variants using information from other
535 samples; this raises the genotype call rate to 100% but may also introduce errors if there is low
536 confidence in the existing genotype calls, emphasizing the need for careful filtering steps. The
537 second stage of imputation made use of IMPTUE2 and an external reference panels of variants

538 called from WGS data on the 8 inbred HS founders, as well as 34 additional inbred rat strains.
539 We decided to include the 34 additional strains because of the elevated genotyping rate we
540 observed upon their inclusion in the IMPUTE2 reference panel. We attribute this to the presence
541 of haplotypes that exist in both the 8 the HS founder strains and a subset of the 34 additional
542 strains in this panel.

543 In summary, we have adapted a GBS protocol and genotyping and imputation pipeline to
544 obtain dense genotypes on genome-wide markers in highly-multiplexed HS rats. After quality
545 filtering on the level of SNP and sample, over 3.7 million were called with a concordance rate of
546 99%. The ddGBS protocol and bioinformatic methods used to produce this data are publicly
547 available, easy to handle, and cost-effective. The presented workflow could be feasibly followed
548 with marginal modifications for application in other species.

549

550 ACKNOWLEDGEMENTS:

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

553 LITERATURE CITED

554 Aird, Daniel, Michael G Ross, Wei-Sheng Chen, Maxwell Danielsson, Timothy Fennell, Carsten Russ,
555 David B Jaffe, Chad Nusbaum, and Andreas Gnirke. 2011. "Analyzing and Minimizing PCR
556 Amplification Bias in Illumina Sequencing Libraries." *Genome Biology* 12 (2): R18.
557 https://doi.org/10.1186/gb-2011-12-2-r18.
558 Andrews, Simon. 2017. *FastQC* (version 0.11.6).
559 http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
560 Baird, Nathan A., Paul D. Etter, Tressa S. Atwood, Mark C. Currey, Anthony L. Shiver, Zachary A. Lewis,
561 Eric U. Selker, William A. Cresko, and Eric A. Johnson. 2008. "Rapid SNP Discovery and Genetic
562 Mapping Using Sequenced RAD Markers." Edited by Justin C. Fay. *PLoS ONE* 3 (10): e3376.
563 https://doi.org/10.1371/journal.pone.0003376.
564 Browning, Brian L., and Sharon R. Browning. 2009. "A Unified Approach to Genotype Imputation and
565 Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals." *The*
566 *American Journal of Human Genetics* 84 (2): 210–23.
567 https://doi.org/10.1016/j.ajhg.2009.01.005.
568 Browning, Brian L., and Sharon R. Browning. 2016. "Genotype Imputation with Millions of Reference
569 Samples." *The American Journal of Human Genetics* 98 (1): 116–26.
570 https://doi.org/10.1016/j.ajhg.2015.11.020.
571 Catchen, Julian, Paul A. Hohenlohe, Susan Bassham, Angel Amores, and William A. Cresko. 2013. "Stacks:
572 An Analysis Tool Set for Population Genomics." *Molecular Ecology* 22 (11): 3124–40.
573 https://doi.org/10.1111/mec.12354.
574 Chen, Qiang, Yufang Ma, Yumei Yang, Zhenliang Chen, Rongrong Liao, Xiaoxian Xie, Zhen Wang, et al.
575 2013. "Genotyping by Genome Reducing and Sequencing for Outbred Animals." Edited by
576 Shuhong Zhao. *PLoS ONE* 8 (7): e67500. https://doi.org/10.1371/journal.pone.0067500.
577 Chitre, Apurva S, Oksana Polesskaya, Katie Holl, Jianjun Gao, Riyan Cheng, Angel Martinez, Tony George,
578 et al. 2018. "Genome Wide Association Study of Body Weight, Body Mass Index, Adiposity, and
579 Fasting Glucose in 3,173 Outbred Rats," September. https://doi.org/10.1101/422428.
580 Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, et al. 2011.
581 "The Variant Call Format and VCFtools." *Bioinformatics* 27 (15): 2156–58.
582 https://doi.org/10.1093/bioinformatics/btr330.
583 Davey, John W., Paul A. Hohenlohe, Paul D. Etter, Jason Q. Boone, Julian M. Catchen, and Mark L.
584 Blaxter. 2011. "Genome-Wide Genetic Marker Discovery and Genotyping Using next-Generation
585 Sequencing." *Nature Reviews Genetics* 12 (7): 499–510. https://doi.org/10.1038/nrg3012.
586 De Donato, Marcos, Sunday O. Peters, Sharon E. Mitchell, Tanveer Hussain, and Ikhide G. Imumorin.
587 2013. "Genotyping-by-Sequencing (GBS): A Novel, Efficient and Cost-Effective Genotyping
588 Method for Cattle Using Next-Generation Sequencing." Edited by James C. Nelson. *PLoS ONE* 8
589 (5): e62137. https://doi.org/10.1371/journal.pone.0062137.
590 Delaneau, Olivier, Jonathan Marchini, and Jean-François Zagury. 2011. "A Linear Complexity Phasing
591 Method for Thousands of Genomes." *Nature Methods* 9 (December): 179.
592 DePristo, Mark A, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl,
593 Anthony A Philippakis, et al. 2011. "A Framework for Variation Discovery and Genotyping Using
594 Next-Generation DNA Sequencing Data." *Nature Genetics* 43 (5): 491–98.
595 https://doi.org/10.1038/ng.806.
596 Durvasula, Arun, Paul J Hoffman, Tyler V Kent, Chaochih Liu, Thomas J Y Kono, Peter L Morrell, and
597 Jeffrey Ross-Ibarra. n.d. "ANGSD-Wrapper: Utilities for Analyzing next Generation Sequencing
598 Data." Accessed September 5, 2018. https://doi.org/10.7287/peerj.preprints.1472v2.

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.

646 Howie, Bryan, Jonathan Marchini, and Matthew Stephens. 2011. “Genotype Imputation with Thousands
647 of Genomes.” *G3: Genes|Genomes|Genetics* 1 (6): 457–70.
648 https://doi.org/10.1534/g3.111.001198.
649 Howie, Bryan N., Peter Donnelly, and Jonathan Marchini. 2009. “A Flexible and Accurate Genotype
650 Imputation Method for the Next Generation of Genome-Wide Association Studies.” Edited by
651 Nicholas J. Schork. *PLoS Genetics* 5 (6): e1000529.
652 https://doi.org/10.1371/journal.pgen.1000529.
653 Huang, Guan-Hua, and Yi-Chi Tseng. 2014. “Genotype Imputation Accuracy with Different Reference
654 Panels in Admixed Populations.” *BMC Proceedings* 8 (Suppl 1): S64.
655 https://doi.org/10.1186/1753-6561-8-S1-S64.
656 Huang, X., Q. Feng, Q. Qian, Q. Zhao, L. Wang, A. Wang, J. Guan, et al. 2009. “High-Throughput
657 Genotyping by Whole-Genome Resequencing.” *Genome Research* 19 (6): 1068–76.
658 https://doi.org/10.1101/gr.089516.108.
659 Hwang, Sohyun, Eiru Kim, Insuk Lee, and Edward M. Marcotte. 2015. “Systematic Comparison of Variant
660 Calling Pipelines Using Gold Standard Personal Exome Variants.” *Scientific Reports* 5 (December):
661 17875.
662 Illumina, Inc. 2014. “Nextera(R) Library Validation and Cluster Density Optimization: Guidelines for
663 Generating High-Quality Data with Nextera Library Prepartion Kits.”
664 https://www.illumina.com/documents/products/technotes/technote_nextera_library_validatio
665 n.pdf.
666 Jensen-Seaman, M. I. 2004. “Comparative Recombination Rates in the Rat, Mouse, and Human
667 Genomes.” *Genome Research* 14 (4): 528–38. https://doi.org/10.1101/gr.1970304.
668 Johannesson, M., R. Lopez-Aumatell, P. Stridh, M. Diez, J. Tuncel, G. Blazquez, E. Martinez-Membrives,
669 et al. 2008. “A Resource for the Simultaneous High-Resolution Mapping of Multiple Quantitative
670 Trait Loci in Rats: The NIH Heterogeneous Stock.” *Genome Research* 19 (1): 150–58.
671 https://doi.org/10.1101/gr.081497.108.
672 Johnson, Jennifer L., Helena Wittgenstein, Sharon E. Mitchell, Katie E. Hyma, Svetlana V. Temnykh,
673 Anastasiya V. Kharlamova, Rimma G. Gulevich, et al. 2015. “Genotyping-By-Sequencing (GBS)
674 Detects Genetic Structure and Confirms Behavioral QTL in Tame and Aggressive Foxes (Vulpes
675 Vulpes).” Edited by William J. Murphy. *PLOS ONE* 10 (6): e0127013.
676 https://doi.org/10.1371/journal.pone.0127013.
677 Kanagawa, Takahiro. 2003. “Bias and Artifacts in Multitemplate Polymerase Chain Reactions (PCR).”
678 *Journal of Bioscience and Bioengineering* 96 (4): 317–23. https://doi.org/10.1016/S1389-
679 1723(03)90130-7.
680 Korneliussen, Thorfinn Sand, Anders Albrechtsen, and Rasmus Nielsen. 2014. “ANGSD: Analysis of Next
681 Generation Sequencing Data.” *BMC Bioinformatics* 15 (1). https://doi.org/10.1186/s12859-014-
682 0356-4.
683 Li, H. 2011. “A Statistical Framework for SNP Calling, Mutation Discovery, Association Mapping and
684 Population Genetical Parameter Estimation from Sequencing Data.” *Bioinformatics* 27 (21):
685 2987–93. https://doi.org/10.1093/bioinformatics/btr509.
686 Li, H., and R. Durbin. 2009. “Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform.”
687 *Bioinformatics* 25 (14): 1754–60. https://doi.org/10.1093/bioinformatics/btp324.
688 Li, H., J. Ruan, and R. Durbin. 2008. “Mapping Short DNA Sequencing Reads and Calling Variants Using
689 Mapping Quality Scores.” *Genome Research* 18 (11): 1851–58.
690 https://doi.org/10.1101/gr.078212.108.
691 Li, Zhentang, Yi Wang, and Fei Wang. 2018. “A Study on Fast Calling Variants from Next-Generation
692 Sequencing Data Using Decision Tree.” *BMC Bioinformatics* 19 (1).
693 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™): 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, 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.

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.

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.

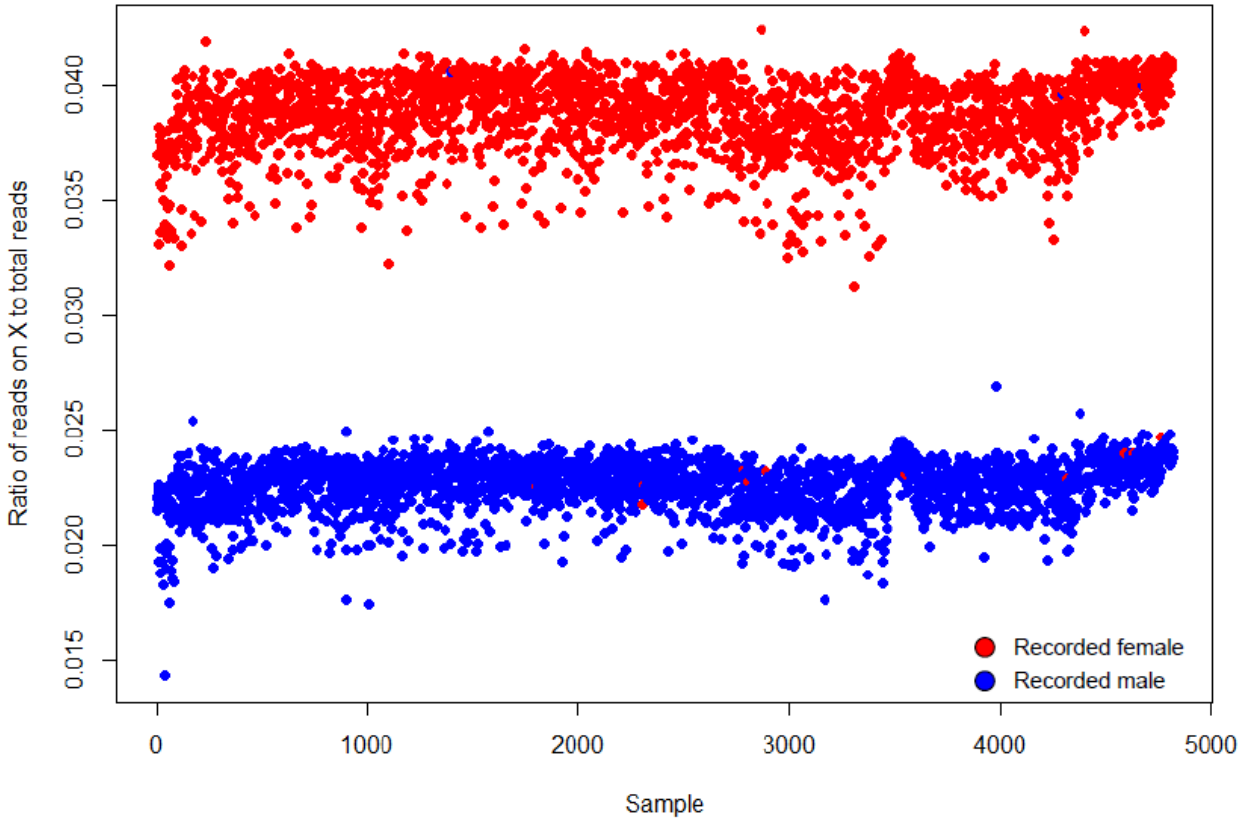


Figure S2. Data preparation workflow for imputation with IMPUTE2.

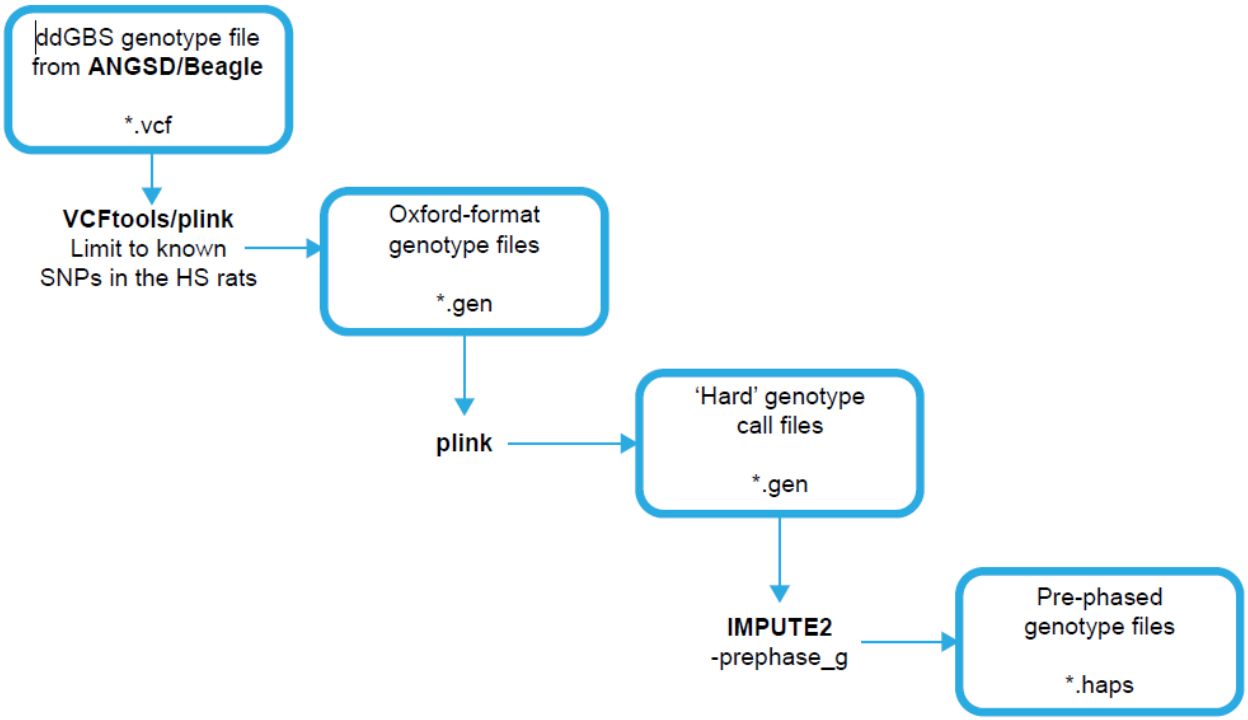


Figure S3. Programmed vs. empirical Pippin Prep fragment size range.

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.

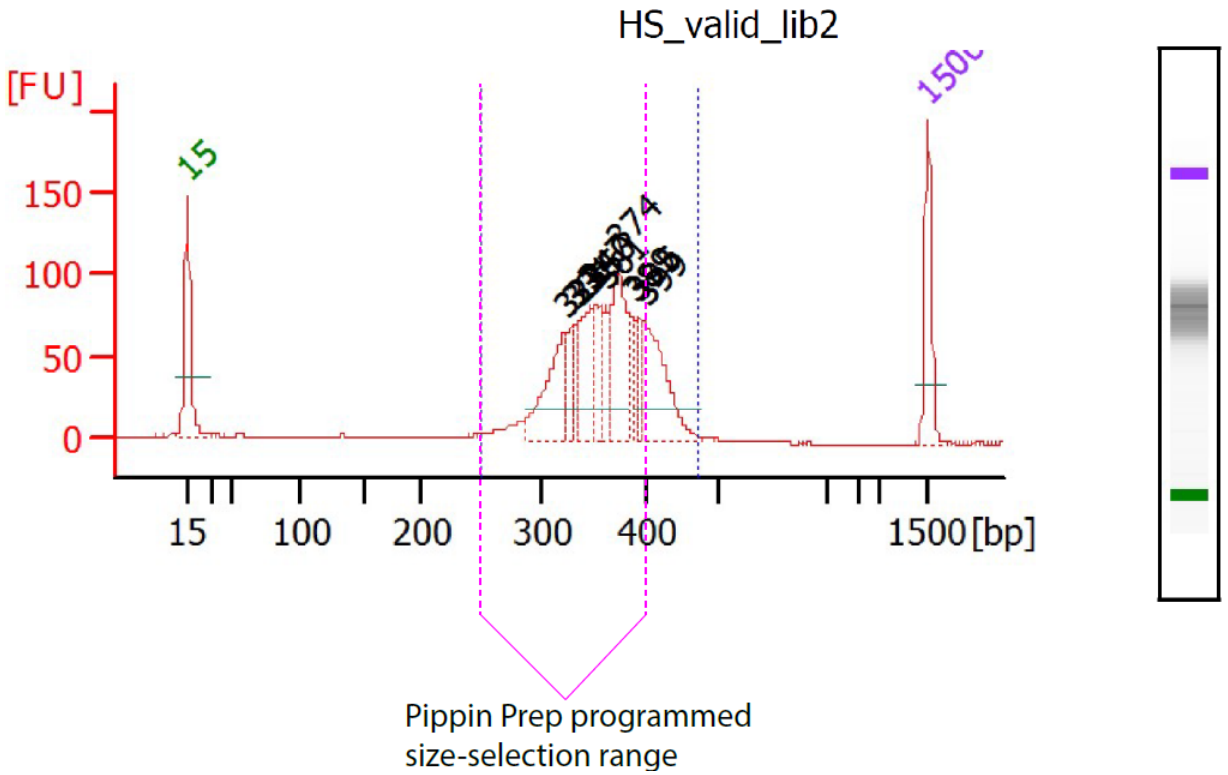


Figure S4. Raw read counts grouped by shipment batch.

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

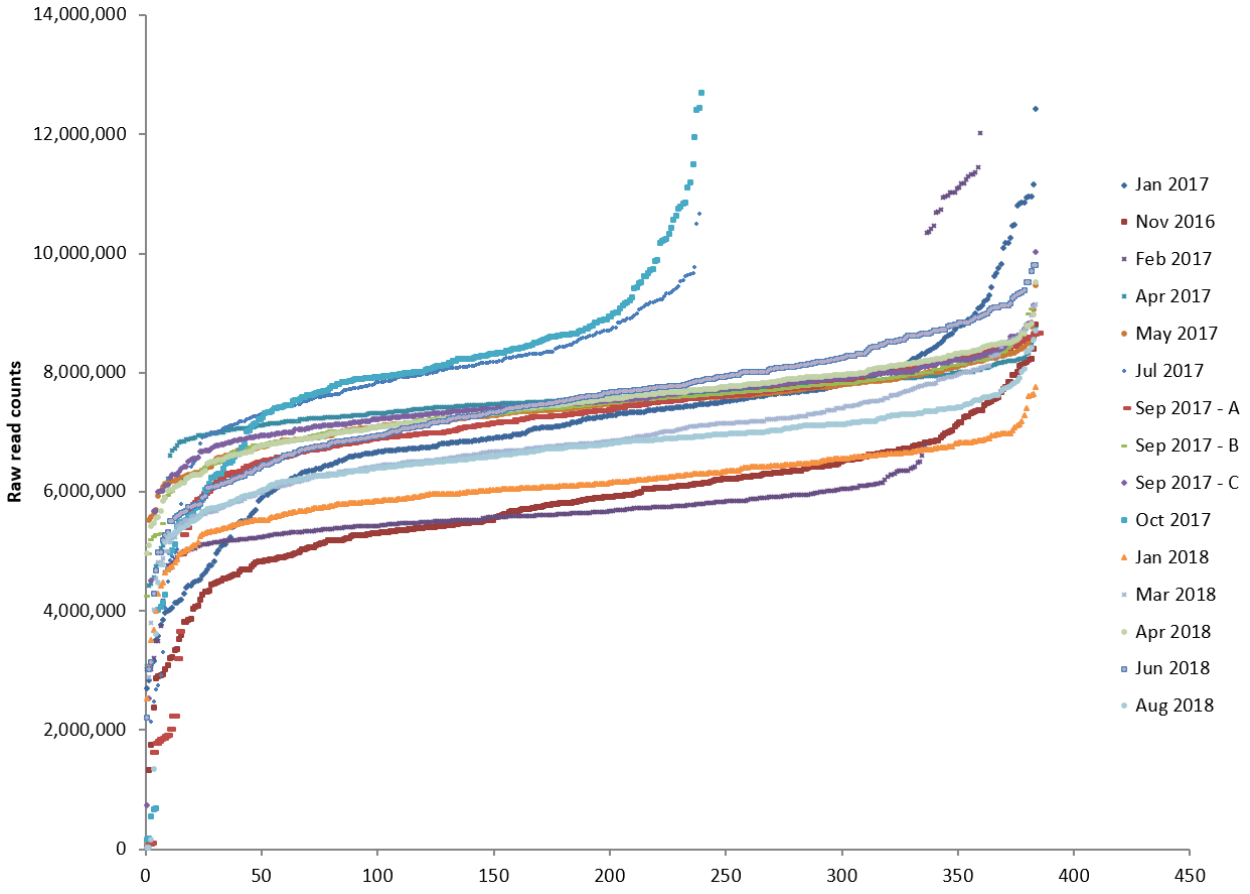


Figure S5. FASTQC results pre- and post-filtering with Cutadapt.

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.

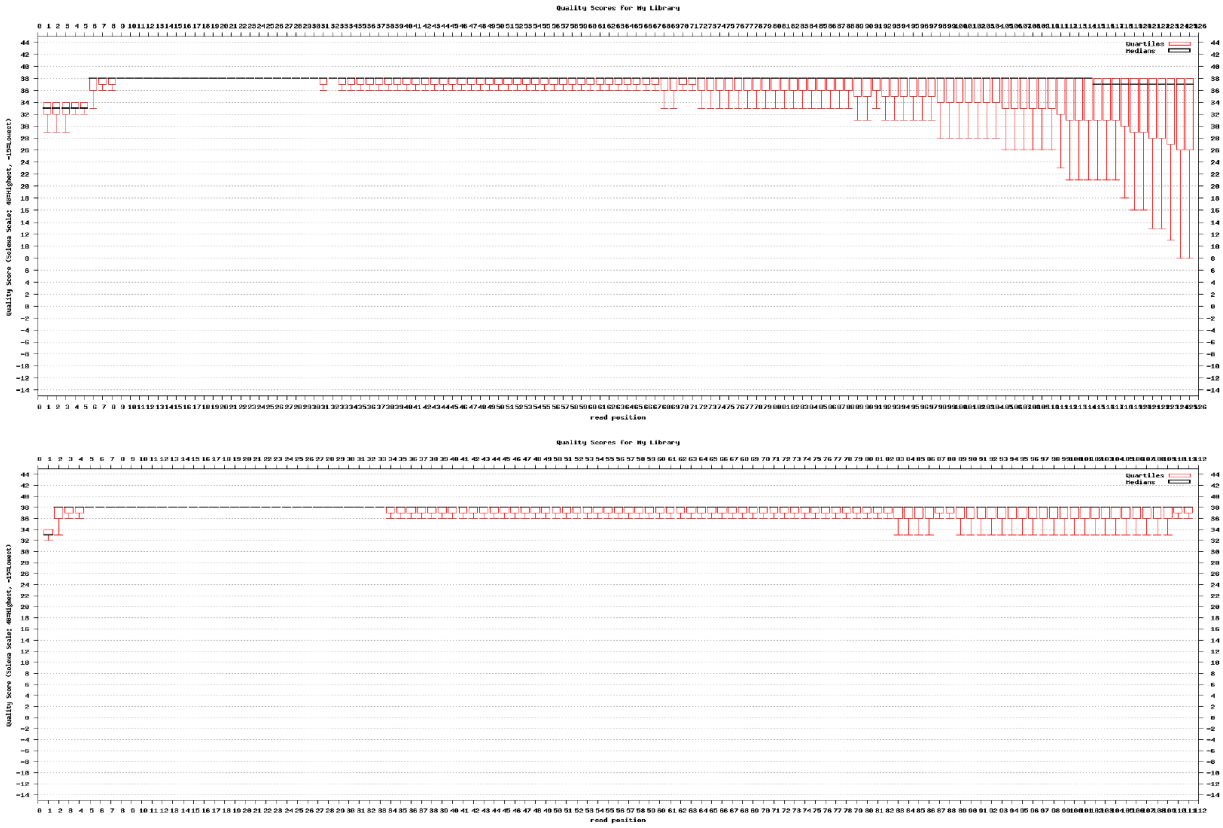
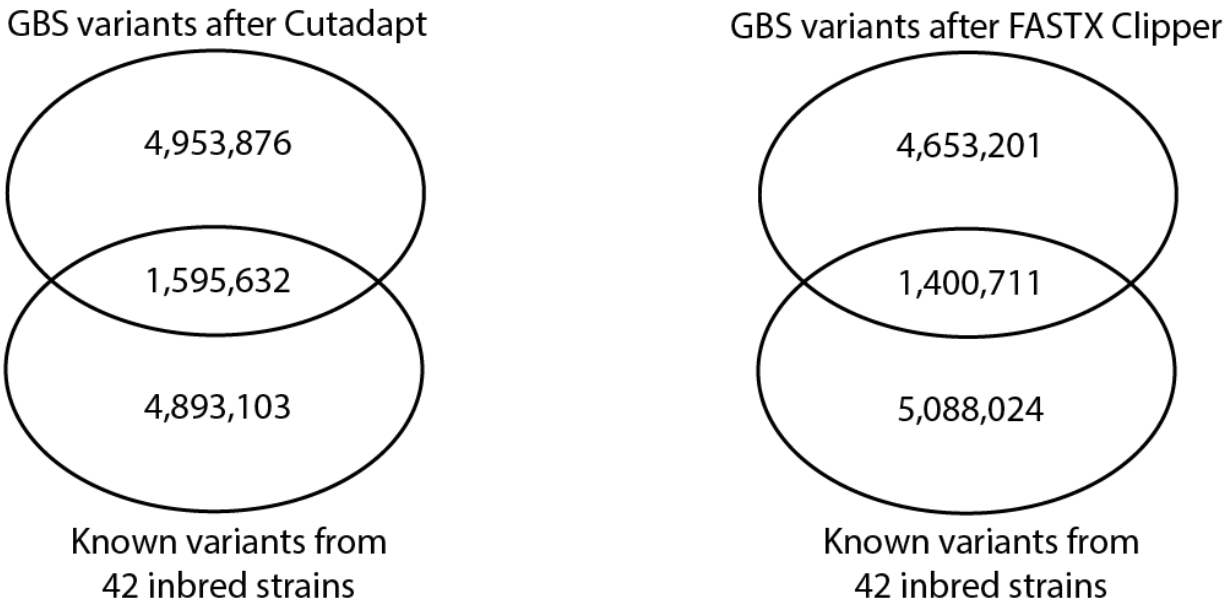
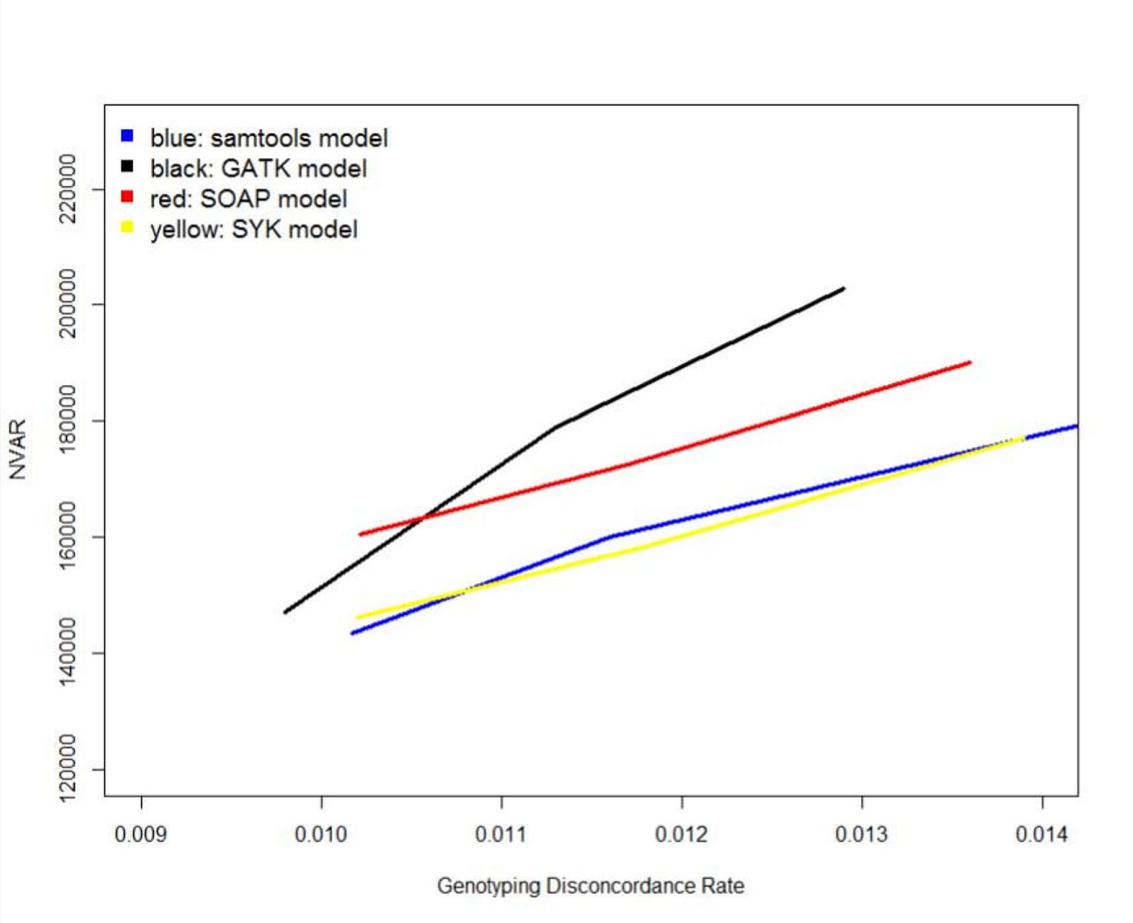


Figure S6. Overlap of called SNPs with known variants after read trimming with FASTX or Cutadapt.



887

888 **Figure S7. Number of variants by genotype discordance rates for 4 ANGSD genotype**
889 **likelihood models.**



890

891

892

893

894

895

896

897

898

899

900

901

Figure S8. Available rat genetic maps.

Plotted physical and genetic distances are for chromosome 12.

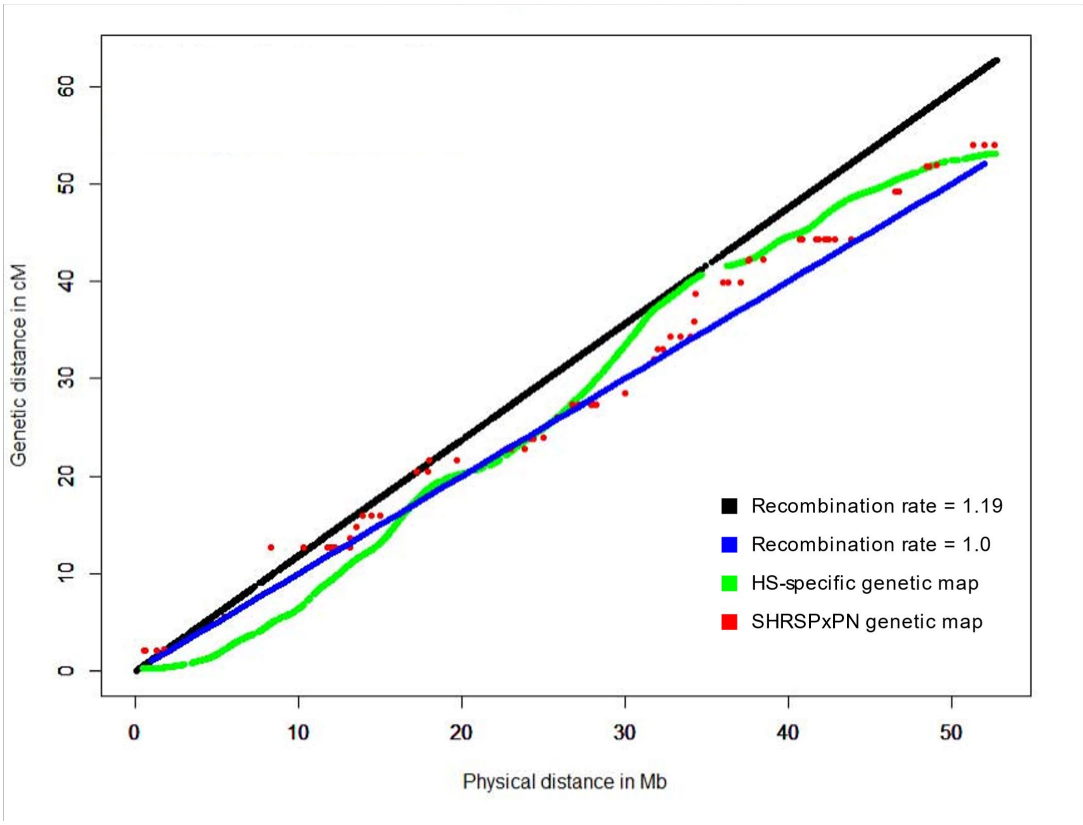


Table S1. Demultiplexing performance.

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

929

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

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

934

	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

935

936

937

938

939

940

941

942

943

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

946 Data shown comes from the original set of 96 HS samples prepared in 12-plex and sequenced on
947 the Illumina HiSeq 2500. Variants were called utilizing the SAMtools model and the -minMapQ
948 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

949

950
951

952

953

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

955 The presented data comes from ANGSD/Beagle variant calls for 3,601 HS samples, prior to
956 imputation with IMPUTE2. Known SNPs came from both the 42 inbred genomes from Hermsen
957 et. al 2015 (Hermsen et al. 2015) and the 8 inbred HS founder strains sequenced by the
958 University of Michigan (Ramdas et al. 2018).

959

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

960

961
962

963

964

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

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

969

	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

970

971

972

973