

# UC San Diego

## UC San Diego Previously Published Works

### Title

A Cost-effective, High-throughput, Highly Accurate Genotyping Method for Outbred Populations.

### Permalink

<https://escholarship.org/uc/item/19x4g2gc>

### Authors

Chen, Denghui

Chitre, Apurva

Nguyen, Khai-Minh

et al.

### Publication Date




2024-12-13

### DOI

10.1093/g3journal/jkae291

Peer reviewed

# A cost-effective, high-throughput, highly accurate genotyping method for outbred populations

Denghui Chen <sup>1</sup>, Apurva S. Chitre <sup>1</sup>, Khai-Minh H. Nguyen,<sup>2</sup> Katerina A. Cohen,<sup>2</sup> Beverly F. Peng,<sup>2</sup> Kendra S. Ziegler,<sup>2</sup> Faith Okamoto,<sup>2</sup> Bonnie Lin,<sup>2</sup> Benjamin B. Johnson,<sup>2</sup> Thiago M. Sanches,<sup>2</sup> Riyan Cheng,<sup>2</sup> Oksana Polesskaya,<sup>2</sup> Abraham A. Palmer <sup>2,3,\*</sup>

<sup>1</sup>Bioinformatics and System Biology Program, University of California San Diego, La Jolla, CA 92093, USA

<sup>2</sup>Department of Psychiatry, University of California San Diego, La Jolla, CA 92093, USA

<sup>3</sup>Institute for Genomic Medicine, University of California San Diego, La Jolla, CA 92093, USA

\*Corresponding author: Department of Psychiatry, University of California San Diego, La Jolla, CA 92093, USA. Email: aap@ucsd.edu

Affordable sequencing and genotyping methods are essential for large-scale genome-wide association studies. While genotyping microarrays and reference panels for imputation are available for human subjects, nonhuman model systems often lack such options. Our lab previously demonstrated an efficient and cost-effective method to genotype heterogeneous stock rats using double-digest genotyping by sequencing. However, low-coverage whole-genome sequencing offers an alternative method that has several advantages. Here, we describe a cost-effective, high-throughput, high-accuracy genotyping method for N/NIH heterogeneous stock rats that can use a combination of sequencing data previously generated by double-digest genotyping by sequencing and more recently generated by low-coverage whole-genome sequencing data. Using double-digest genotyping-by-sequencing data from 5,745 heterogeneous stock rats (mean 0.21× coverage) and low-coverage whole-genome sequencing data from 8,760 heterogeneous stock rats (mean 0.27× coverage), we can impute 7.32 million biallelic single-nucleotide polymorphisms with a concordance rate > 99.76% compared to high-coverage (mean 33.26× coverage) whole-genome sequencing data for a subset of the same individuals. Our results demonstrate the feasibility of using sequencing data from double-digest genotyping by sequencing or low-coverage whole-genome sequencing for accurate genotyping and demonstrate techniques that may also be useful for other genetic studies in nonhuman subjects.

**Keywords:** low-coverage; whole-genome sequencing; genotyping; heterogeneous stock rat

## Introduction

In both humans and model organisms, genome-wide association studies (GWAS) are valuable for identifying genetic variants associated with diseases and other complex traits. GWAS results facilitate the discovery of novel biological pathways and potential therapeutic targets (Palmer et al. 2021; Uffelmann et al. 2021; Alliance of Genome Resources Consortium 2022; Abdellaoui et al. 2023). The success of large-scale population and quantitative genetics studies depends on the availability of dense and high-quality genotype data (Welter et al. 2014). Single-nucleotide polymorphism (SNP) arrays, paired with reference panels (e.g. HapMap or the 1000 Genomes Project), are commonly used to infer genotypes and perform genetic studies in humans (Frazer et al. 2007; Marchini and Howie 2010; McVean et al. 2012; Uffelmann et al. 2021; Aganezov et al. 2022). However, SNP arrays often perform poorly when applied to populations other than the one used for array design, leading to a need for costly development of population-specific SNP arrays (Didion et al. 2012). This issue is even more critical in model organisms, where population structure is often very pronounced (Gileta et al. 2022). An alternative to genotyping microarrays is to use next-generation sequencing. Because sequencing at sufficient depth to make calls directly remains expensive, low-coverage sequencing paired with imputation from reference panels provides a more economical

solution (Davies et al. 2016; Petter et al. 2020; Li et al. 2021, 2024; Wasik et al. 2021).

Our lab has performed GWAS using various mouse and rat populations (Chitre et al. 2020, 2023; Zhou et al. 2020; Gileta et al. 2022; Gunturkun et al. 2022; Parker et al. 2022; Fowler et al. 2023). In particular, we have now phenotyped and genotyped almost 20,000 N/NIH heterogeneous stock (HS) rats. HS rats were created in 1984 by intercrossing 8 inbred rat strains (ACI/N, BN/SsN, BUF/N, F344/N, M520/N, MR/N, WKY/N, and WN/N). To genotype outbred mice and rats, we have used genotyping by sequencing (GBS) (Elshire et al. 2011; Parker et al. 2016; Gonzales et al. 2018) and subsequently double-digest GBS (ddGBS) protocols, followed by imputation (Gileta et al. 2020). More recently, we have reported on our use of commercial whole-genome sequencing (WGS) library preparation kits to generate low-coverage WGS (lcWGS) data, followed by imputation using outbred mice (Davies et al. 2016; Nicod et al. 2016; Zou et al. 2022). However, we have not previously reported on our methods for genotyping rats using lcWGS followed by imputation, nor have we reported a method for jointly calling genotypes using a combination of ddGBS and lcWGS data.

In this paper, we present a cost-effective, high-throughput, and highly accurate genotyping method for HS rats that utilizes both previously generated ddGBS data and more recently generated lcWGS data. This method allowed us to impute 7.32 million

biallelic SNPs with a concordance rate of >99.76% compared to genotypes obtained from 33.26× coverage WGS without imputation for a subset of the same individuals.

## Materials and methods

### Animals

As reviewed elsewhere, the N/NIH HS rat population was created by interbreeding 8 inbred rat strains (ACI/N, BN/SsN, BUF/N, F344/N, M520/N, MR/N, WKY/N, and WN/N) in the mid-1980s (Solberg Woods and Palmer 2019). Since then, HS rats have been maintained as an outbred population for more than 100 generations. Because they have been maintained as an outbred population for such a long time, HS rats possess short haplotypes that are derived from the 8 inbred founders, making them ideal for high-resolution genetic mapping (Johannesson et al. 2009; Baud et al. 2013; Woods and Mott 2017; Solberg Woods and Palmer 2019). In this study, we used sequence data from a total of 15,552 HS rats (7,797 males and 7,755 females) from generation 81 to 97 that were bred at the Medical College of Wisconsin (RRID: RGD\_2314009), Wake Forest University (RRID: RGD\_13673907), the University of Tennessee Health Sciences Center, or Oregon Health and Sciences University. The colony at the Medical College of Wisconsin moved to Wake Forest University in 2016, which resulted in 2 sites that existed sequentially. The University of Tennessee Health Sciences Center and Oregon Health and Sciences University bred rats from Wake Forest University for a single generation to produce offspring locally; therefore, RRIDs have not been issued for these 2 sites. Detailed composition of rats by sex and site is outlined in Supplementary Table 1. All procedures that occurred prior to tissue collection were approved by the relevant Institutional Animal Care and Use Committees. As described in the following sections, of the 15,552 HS rats, 477 were sequenced with both ddGBS and lcWGS. Eighty-eight of those 477 were also whole-genome sequenced at an average depth of 33.26×; we refer to those 88 rats as the “truth set.”

### ddGBS sequencing

Of the 15,552 HS rats used in this study, 6,379 individuals (3,219 males and 3,160 females) were sequenced using a ddGBS library preparation protocol described by (Gileta et al. 2020). Briefly, DNA was extracted from spleen tissues using Agencourt DNAdvance Kit (Beckman Coulter Life Sciences, Indianapolis, IN, USA) and digested using the restriction enzymes PstI and NlaIII. After adapter ligation, DNA purification, and library pooling, sample DNA was sequenced as 48 samples per library on Illumina HiSeq 4000 with 100-bp single-end reads at the University of California San Diego Institute for Genomic Medicine Genomics Center (UCSD IGM).

### lcWGS sequencing

In addition, 9,173 (4,578 males and 4,595 females) of 15,552 HS rats underwent lcWGS sequencing. DNA was extracted from spleen tissues using the Agencourt DNAdvance Kit, and the Twist 96-Plex Library Prep Kit (Twist Bioscience, South San Francisco, CA, USA) was used for library preparation following the manufacturer's protocol. In each library, 96 samples were barcoded separately. Then, the samples' DNA was sequenced on Illumina NovaSeq 4000 or 6000 with 150-bp paired-end reads at UCSD IGM. DNA extraction, normalization, randomization, and library preparation were all performed on the EPmotion 5075 (Eppendorf, Hamburg, Germany) liquid-handling robot. Detailed

lcWGS protocols for many of these steps can be found in the Center for GWAS in Outbred Rats Database protocol repository on protocols.io ([https://www.protocols.io/workspaces/cgord\\_spleen\\_cutting](https://www.protocols.io/workspaces/cgord_spleen_cutting): [http://dx.doi.org/10.17504/protocols.io.36wgg7nr\\_yvk5/v1](http://dx.doi.org/10.17504/protocols.io.36wgg7nr_yvk5/v1), DNA extraction: <http://dx.doi.org/10.17504/protocols.io.8epv59reng1b/v1>, normalization and randomization: <http://dx.doi.org/10.17504/protocols.io.261genw5dg47/v1>, library preparation: <http://dx.doi.org/10.17504/protocols.io.j8nlkkm8515r/v1>, pooling and sequencing: <http://dx.doi.org/10.17504/protocols.io.yxmvmnw29g3p/v1>).

### Reference panel preparation

To obtain the best possible imputation reference panel for outbred HS rats, we used consensus biallelic homozygous SNP calls from 3 different inbred HS rat founder data sets. The first data set was produced from publicly available 30.34× coverage WGS sequences (NCBI SRA: PRJNA487943) using the Genome Analysis Toolkit (GATK) joint calling workflow (Supplementary Method 2 and Supplementary Fig. 1) (Ramdas et al. 2019; Van der Auwera and O'Connor 2020). In that data set, BN/SsN and MR/N are female, and other rats are male. The second data set was produced using the same GATK joint calling workflow using an independent data set with an average of 41.81× coverage WGS sequences (NCBI SRA: PRJNA1048943) generated with high-coverage WGS sequencing procedures (Supplementary Methods 1 and 2 and Supplementary Fig. 1). Details of this data set have not been previously published. In this data set, all 8 HS founders were male. The third data set was produced using the same 41.81× coverage WGS sequences, but using the DeepVariant multisample calling workflow (Supplementary Method 3 and Supplementary Fig. 2). Filters applied after variant calling processes were described in the corresponding supplementary method sections. For autosomal chromosomes, chromosome X, and mitochondria, 7,406,667, 184,934, and 117 SNPs respectively that had consensus homozygous genotypes across all 3 call sets were retained; however, because BN/SsN and MR/N in the first data set are female, we dropped them from the consensus check process for chromosome Y, resulting in 5,220 consensus homozygous SNPs for chromosome Y. In total, 7,596,938 SNPs were retained for the reference panel.

### Biallelic SNP positions preparation

We employed STITCH for the imputation process. STITCH was designed for imputing biallelic SNPs in lcWGS reads by constructing haplotypes (Davies et al. 2016). STITCH accepts a position file for the biallelic SNPs to be imputed. In order to capture the common variants derived from the HS founders, as well as new SNPs observed in recent generations of the outbred HS population, we compiled the SNP position file using biallelic SNPs discovered in the founder data sets mentioned above and in 88 HS rats (44 males and 44 females). Variants in the subset of 88 HS rats were called on 33.26× coverage WGS sequences (NCBI SRA: PRJNA1076141) using the GATK joint calling workflow (Supplementary Methods 1 and 2 and Supplementary Fig. 1). The resulting SNPs position file contained 10,684,883 SNPs with 10,227,209 on autosomal chromosomes, 331,389 on chromosome X, 126,141 on chromosome Y, and 144 on mitochondria.

### Truth set preparation

To assess the quality of imputed genotypes, we sequenced the aforementioned 88 HS outbred rats using 3 methods: ddGBS, lcWGS, and high-coverage WGS (33.26×). The biallelic SNPs imputed from the ddGBS and lcWGS genotyping pipeline were

compared with the variants discovered on high-coverage WGS GATK joint calling pipeline (Supplementary Methods 1 and 2 and Supplementary Fig. 1). Variants filtering process was described in the supplementary method as well. We treated the genotypes called by high-coverage WGS as our truth set and used them to check the concordance of the genotypes imputed with the other 2 methods.

## Genotyping

Our full bioinformatic pipeline is outlined in Fig. 1. The pipeline inputs each sample's raw ddGBS or lcWGS sequences, maps them to *Rattus norvegicus* reference genome mRatBN7.2 (NCBI Genome Assembly Accession: GCF\_015227675.2) in parallel, and then jointly imputes biallelic SNPs. The complete source code for the pipeline can be found in the Palmer Lab GitHub repository (<https://github.com/Palmer-Lab-UCSD/HS-Rats-Genotyping-Pipeline>, DOI:<https://doi.org/10.5281/zenodo.10002191>).

ddGBS sequences were demultiplexed using `fastx_toolkit` v0.0.14 (Hannon Lab 2010). Barcode, adapter, and quality trimming were subsequently performed using `Cutadapt` v4.1 (Martin 2011) with 25 bp as the minimum length per read and 20 as the minimum base quality. `BWA-mem` v0.7.17 (Li 2013) was used to align ddGBS sequences with a constraint of an alignment score greater than 20, and the aligned BAM files were sorted and indexed by coordinates using `SAMtools` v1.14 (Danecek et al. 2021) for fast random access.

lcWGS sequences were demultiplexed using `fgbio` v1.3.0 (Tim and Nils 2023). `BBDuk` v38.94 (Bushnell 2024) (`ktrim = r`, `k = 23`, `mink = 11`, `hdist = 1`, `trimpolyg = 50`, `tpe`, `tbo`) was used to trim adapters, and `Cutadapt` v4.1 (Martin 2011) was used to trim sequences with Phred base quality < 5 and length shorter than 70 bp. Alignment of the lcWGS sequences was carried out using `BWA-mem` v0.7.17 (Li 2013). Duplicated reads were marked using `Picard` v2.25.7 (Broad Institute 2019) and indexed by coordinates using `SAMtools` v1.14 (Danecek et al. 2021) for fast random access.

Aligned sequences were used to jointly impute biallelic SNPs at given positions with `STITCH` v1.6.6 (Davies et al. 2016) (`niterations = 2`, `k = 8`, `nGen = 100`). At the first iteration of `STITCH`'s EM algorithm, the reference haplotypes are used to initialize the ancestral haplotype. After the first iteration, `STITCH` uses information from the samples' reads to update the ancestral haplotypes. In our genotyping pipeline, we set `niterations` parameter to 2 to enable `STITCH` to capture variants not present in the provided reference panel. Since the HS rat population was derived from 8 inbred founders, we set the `STITCH` `k` parameter to 8 to specify the number of founder haplotypes to use. `STITCH` also requires an `nGen` parameter for the number of population generations. For the results presented in this paper, we used 100. We experimented with other values and found that this parameter had virtually no impact on our results. During the imputation step, a reference panel based on the genotypes of the 8 inbred founder strains and the SNP position file mentioned above were provided to `STITCH` to construct haplotypes for imputation. To increase computational efficiency, imputation was performed parallelly on chromosome chunks with a 1-Mb buffer on each end. Each chunk had a length of at least 7 Mb and contained at least 1,000 SNPs. Then, we used `BCFtools` v1.14 (Danecek et al. 2021) to concatenate the chunks back to individual chromosomes.

## SNP quality control

Following the imputation process, we implemented a quality control procedure to filter out SNPs with low genotype quality. A total of 10,684,883 biallelic SNPs were imputed using our genotyping

pipeline. Among them, we removed 2,737,742 SNPs with an imputation info score < 0.9 using `BCFtools` v1.14 (Danecek et al. 2021). Furthermore, we filtered out 623,881 SNPs that have low concordance with the ground truth data set described above. As a result, we retained 7,323,260 SNPs. The genotypes after quality control can be found in UC San Diego Library Digital Collections (<https://doi.org/10.6075/J0445MPC>).

## Sample quality control

A sample quality control step was also performed to ensure sample quality. In total, 15,552 samples, representing 14,629 unique outbred HS rats, were used in this study. We excluded 66 samples whose ratio of mapped reads on chromosomes X and Y were incompatible with their reported sex (Supplementary Fig. 3). We also filtered out samples with high genotype missing rate and samples with possible contamination based on their genotype heterozygosity rate. Specifically, we excluded 153 samples that either had a genotype missing rate exceeding 0.1 or a genotype heterozygosity rate falling outside the range of  $\pm 4$  SD (Supplementary Fig. 4). Because of the differences between ddGBS and lcWGS data, we conducted these 2 sample quality control criteria for different sequencing methods separately. Additionally, in the cases where we had multiple sequencing runs for the same samples, we kept only the one with the highest number of sequence reads. This quality control process resulted in the retention of 14,505 distinct HS rats (7,283 males and 7,222 females) with 5,745 individuals from ddGBS (2,903 males and 2,842 females) and 8,760 individuals from lcWGS (4,380 males and 4,380 females).

## Results

### Sequence statistics

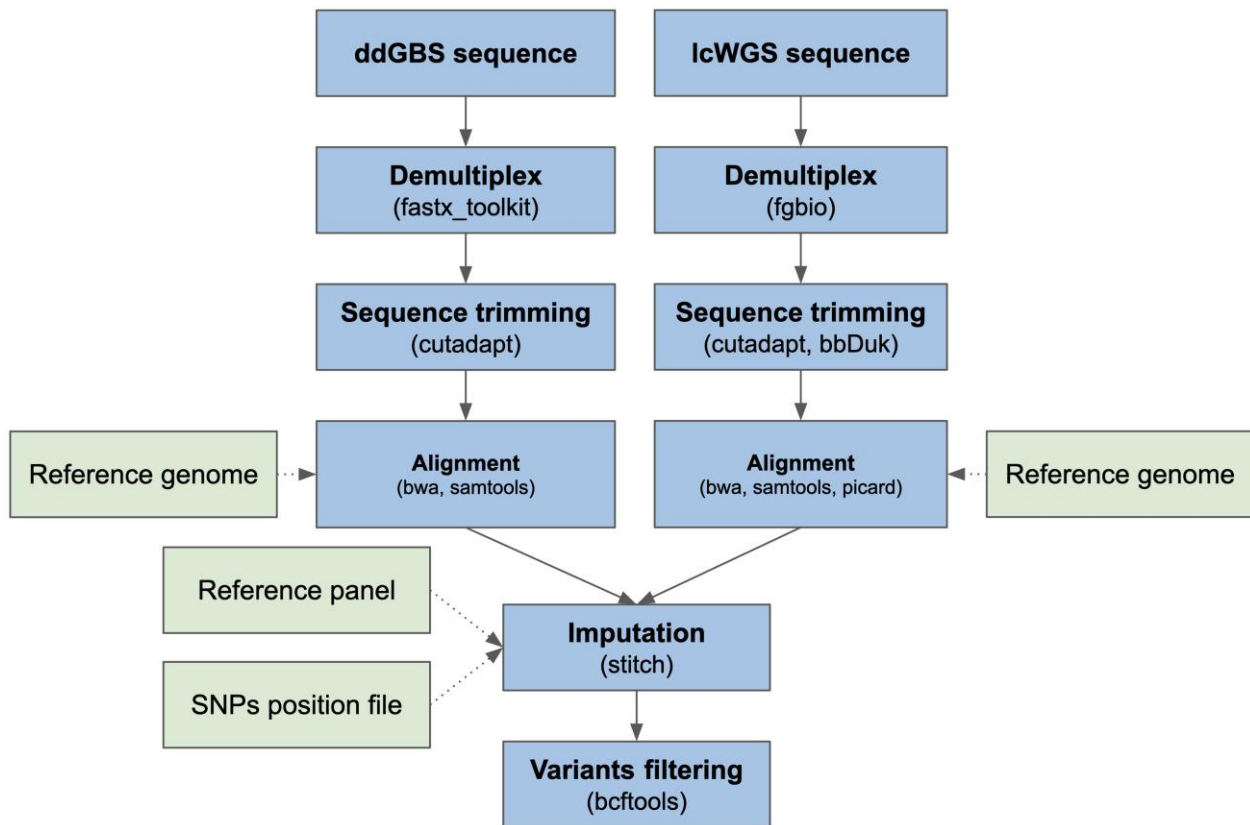
Our genotyping pipeline was applied to 15,552 samples, representing 14,629 unique outbred HS rats. A total of 14,505 distinct samples were retained after the quality control steps described in *Materials and Methods* section, 5,745 of which were sequenced using ddGBS and 8,760 using lcWGS.

After demultiplexing and aligning to reference genome mRatBN7.2 (NCBI Genome Assembly Accession: GCF\_015227675.2), a mean of 8.44 million 100-bp reads per sample was mapped to the reference genome in the case of ddGBS (Fig. 2a). Because of the double restriction enzyme digestion employed in ddGBS, only the chromosomal regions near the enzyme cut sites were sequenced. This led to ddGBS sequences covering 4.97% of the genome per sample, with a mean coverage of 4.22 $\times$  at each captured site (Fig. 2b and c). Consequently, this approach resulted in an average mapped coverage of 0.21 $\times$  per sample across the entire genome although that coverage was highly nonuniform, by design (Fig. 2d).

For lcWGS, a mean of 16.03 million 150-bp reads were mapped for each sample (Fig. 2a). Due to the random priming process of lcWGS, a more diverse set of DNA fragments was sequenced. This enabled lcWGS sequences to cover a wider range of the genome at 18.28% per sample on average, but with a lower mean coverage of 1.39 $\times$  at each capture site (Fig. 2b and c). This resulted in a mean mapped coverage of 0.27 $\times$  per sample genome wide (Fig. 2d).

### Genotype statistics

In our genotyping pipeline, we imputed a total of 10,684,883 biallelic SNPs. Following the quality control procedures outlined in *Materials and Methods* section, 7,323,260 SNPs were retained. Out of these retained SNPs, 7,148,654 were located on autosomal



**Fig. 1.** Genotyping pipeline flow chart.

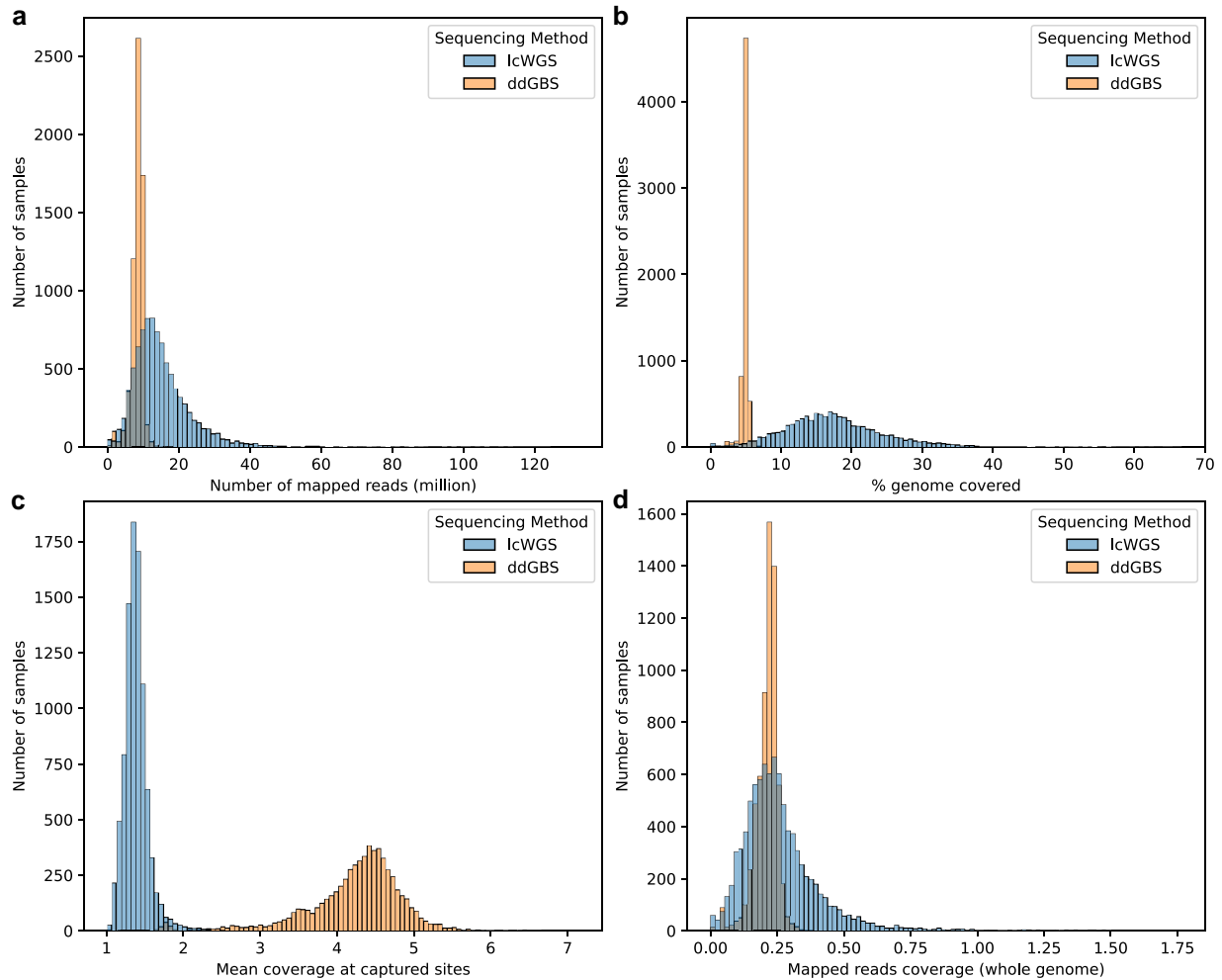
chromosomes, 174,374 were on chromosome X, 118 were on chromosome Y, and 114 were on mitochondria (Fig. 3).

Among the 7,148,654 SNPs on autosomes, 1,602,374 were found to be monomorphic with a minor allele frequency (MAF) of 0. We assume that these SNPs, which were polymorphic in the HS founders, became monomorphic in the outbred HS population due to genetic drift, the proportion of SNPs showing this pattern is consistent with simulations we have performed in the past (Munro et al. 2022). Additionally, new mutations may have arisen since the creation of the HS population, most of which are expected to have low MAF. The primary objective of our SNP genotyping is to identify variants useful for GWAS; however, low MAF SNPs are not well powered to detect associations. Therefore, we applied a MAF threshold of 0.005. A total of 183,621 SNPs fell below the  $MAF \leq 0.005$  threshold whereas 5,362,659 were above the threshold (Supplementary Fig. 5a). Further examination revealed that 143,402 of the SNPs with a  $MAF \leq 0.005$  had an allele count lower than 10, and 136,542 had an allele count lower than 5 (Supplementary Fig. 8a), suggesting that many of the low MAF SNPs were genotyping errors. In this study, all the HS rats used were from the same population; however, familial relationships within the colony could lead to deviations from Hardy–Weinberg equilibrium (HWE) that would not justify excluding the SNPs. Therefore, we applied a lenient HWE threshold of  $-\log_{10}(P\text{-value}) < 10$ . Out of the 7,148,654 SNPs, 39,606 violated HWE with a  $-\log_{10}(P\text{-value}) \geq 10$  (Supplementary Fig. 5b), and 36,664 had a genotype missing rate higher than 0.1 (Supplementary Fig. 5c). Consequently, a total of 5,292,916 autosomal SNPs had a  $MAF > 0.005$ ,  $HWE -\log_{10}(P\text{-value}) < 10$  and missing rate  $\leq 0.1$ .

## Sex chromosomes

Due to the different inheritance patterns on sex chromosomes in males and females, we investigated the SNPs on chromosomes X and Y separately in each sex. Among the 7,222 female samples included in this study, we observed that out of the 174,374 SNPs on chromosome X, 47,882 were monomorphic, 1,375 had a  $MAF \leq 0.005$ , and 125,117 had a  $MAF > 0.005$  (Supplementary Fig. 6a). A total of 627 SNPs violated HWE (Supplementary Fig. 6b), and 582 SNPs had a missing rate higher than 0.1 (Supplementary Fig. 6c). This led to a total of 123,997 chromosome X SNPs for females with a  $MAF > 0.005$ ,  $HWE -\log_{10}(P\text{-value}) < 10$  and missing rate  $\leq 0.1$ . Chromosome Y SNPs were discarded for female samples. In the 7,283 male samples used in this study, among the 174,374 SNPs on chromosome X, 46,319 were monomorphic, 3,227 had a  $MAF \leq 0.005$ , and 124,828 had a  $MAF > 0.005$  (Supplementary Fig. 6d). Because males have only one copy of the X chromosome, we did not test them for HWE, but we found 2,223 chromosome X SNPs had a missing rate higher than 0.1 (Supplementary Fig. 6e). This resulted in a total of 122,693 chromosome X SNPs for males with a  $MAF > 0.005$  and missing rate  $\leq 0.1$ . The 118 SNPs on chromosome Y for male samples had a missing rate  $\leq 0.1$ , but they were all monomorphic SNPs with a MAF of 0.

Out of the 114 SNPs on the mitochondrial chromosome, 30 were found to be monomorphic with a MAF of 0, and the remaining 74 were SNPs with  $MAF > 0.005$  (Supplementary Fig. 7a). HWE was also not tested for mitochondrial SNPs, but all of them had a genotype missing rate lower than 0.1 (Supplementary Fig. 7b). Consequently, a total of 74 mitochondria SNPs had a  $MAF > 0.005$  and missing rate  $\leq 0.1$ .



**Fig. 2.** Aligned sequence statistics. a) Number of reads mapped to reference genome (million). ddGBS mean: 8.44, SD: 1.65; lcWGS mean: 16.03, SD: 10.32. b) Percentage of genome covered by mapped reads in width (%). ddGBS mean: 4.97, SD: 0.54; lcWGS mean: 18.28, SD: 8.31. c) Mean coverage at captured sites. ddGBS mean: 4.22x, SD: 0.67x; lcWGS mean: 1.39x, SD: 0.16x. d) Mapped reads coverage genome wide. ddGBS mean: 0.21x, SD: 0.04x; lcWGS mean: 0.27x, SD: 0.16x.

We have recently published a separate paper that uses the same genotypes described here to examine Y and mitochondrial chromosome haplogroups (Okamoto et al. 2024).

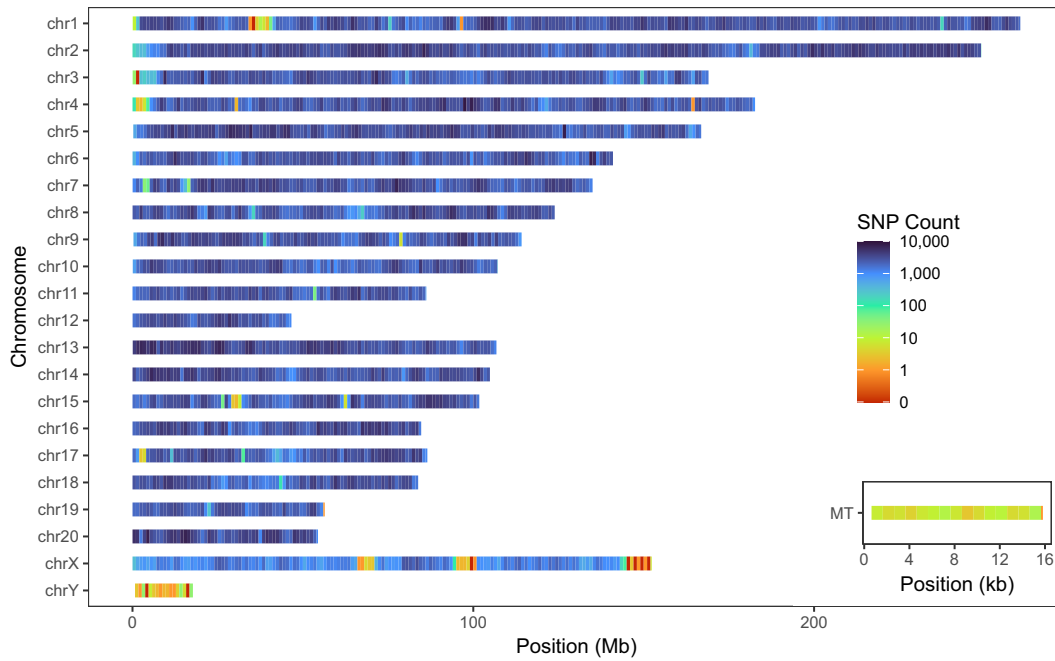
## Genotyping accuracy

As described in *Materials and Methods* section, in the 15,552 outbred HS rats, we genotyped, there were 88 outbred HS rats that had been sequenced with ddGBS, lcWGS, and 33.26x high-coverage WGS. We tested our genotyping pipeline's accuracy by comparing genotypes imputed from ddGBS and lcWGS with SNPs called using high-coverage WGS without any imputation, which we refer to as the "truth set." Specifically, for each sample, we looked at the concordance rate of overlap and nonmissing SNPs between the imputed genotypes and the truth set. Concordance rate calculations were based on SNPs that passed quality control filters:  $MAF > 0.005$ ,  $HWE -\log_{10}(P\text{-value}) < 10$ , and missing rate  $\leq 0.1$ . On average, 5,417,913 polymorphic SNPs were shared between imputation from ddGBS sequences and variant calling from 33.26x high-coverage WGS, with a mean concordance rate of 99.76% (Fig. 4). Similarly, we observed that 5,429,453 SNPs were shared between lcWGS and 33.26x high-coverage WGS, with a mean concordance rate of 99.78% (Fig. 4). Additionally, we examined the concordance across different MAFs. SNPs at different MAFs were

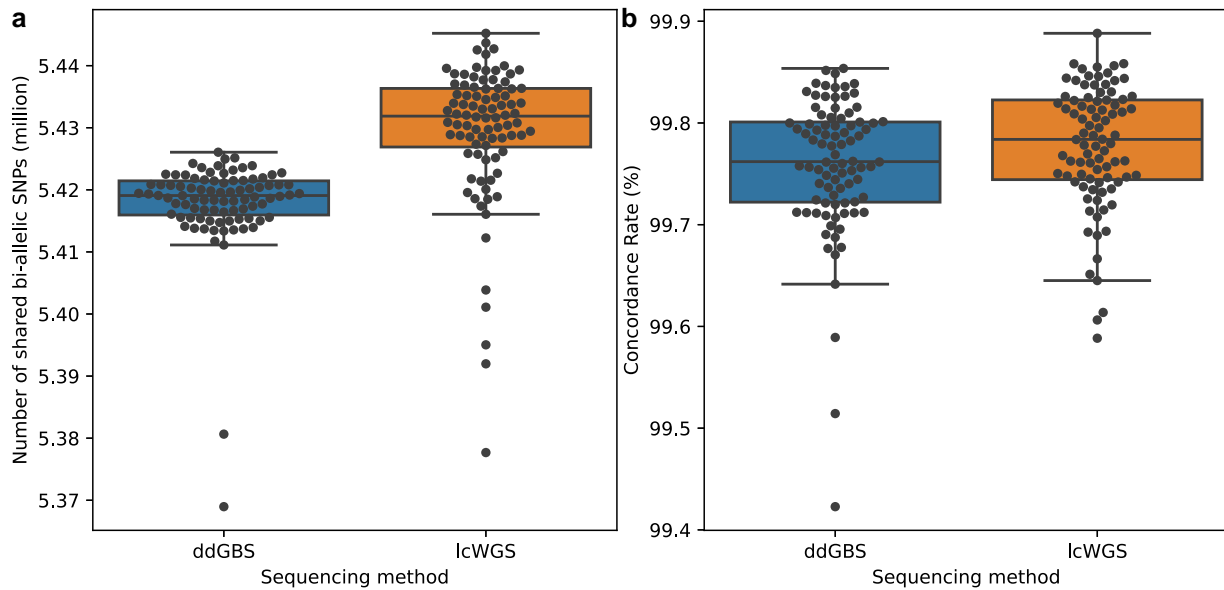
relatively uniformly distributed. The genotype concordance rate started at around 99.98% and decreased slightly as MAF increased such that accuracy dropped to about 99.6% as MAF approached 0.5. Overall these results indicate a high concordance across all allele frequencies (Supplementary Fig. 9). ddGBS sequences a smaller portion of the genome at higher depth, while lcWGS covers a larger portion at lower depth. These differences lead to more regions without reads in ddGBS compared to lcWGS, resulting in a slightly higher number of discordant calls in those regions (Supplementary Fig. 10). For the imputed genotypes on chromosome X, the concordance rates from 2 sequencing approaches are comparable (Supplementary Fig. 11).

## Batch effects on ddGBS and lcWGS genotypes

To investigate potential batch effects of different sequencing methods, we performed a principal component (PC) analysis on the autosomal genotypes of the 88 HS outbred rats sequenced with both ddGBS and lcWGS (Fig. 5). Overlapping first and second PC values without apparent clustering between the 2 methods indicate that both methods capture equivalent information from the genome, meaning there are no obvious method-specific batch effects introduced by the pipeline. Additionally, we did not observe any batch effects in any other PCs that explained more



**Fig. 3.** Imputed biallelic SNPs density distribution heatmap on each chromosome with 1-Mb windows and mitochondrial chromosome with 1-kb windows.



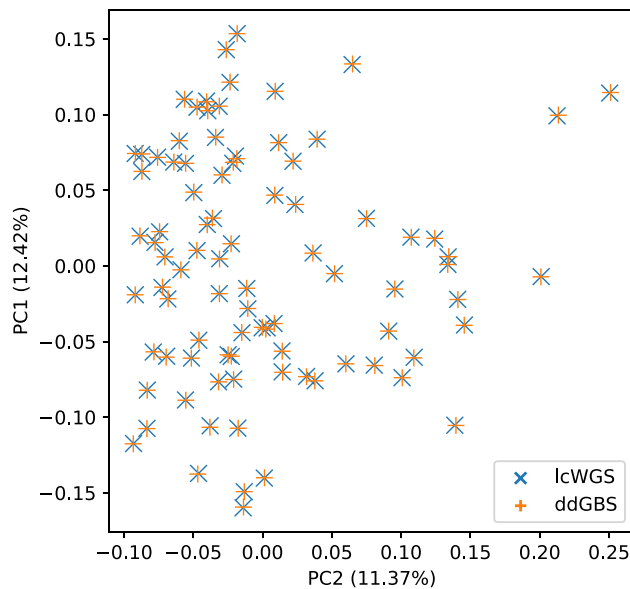
**Fig. 4.** Imputed genotypes demonstrate high concordance with 33.26x high-coverage WGS for millions of biallelic SNPs. a) Number of biallelic SNPs compared (million). ddGBS mean: 5.42, SD: 0.01; lcWGS mean: 5.43, SD: 0.01. b) Concordance rate with 33.26x high-coverage WGS (%). ddGBS mean: 99.76, SD: 0.07; lcWGS mean: 99.78, SD: 0.06.

than 10% of the variance (Supplementary Fig. 12). The MAF and HWE distributions were also comparable across different sequencing batches with minor indication of genetic drift on MAF (Supplementary Fig. 13).

## Discussion

While large-scale genetic studies in humans often use genotyping microarrays and imputation, similar resources are not available for most other species. Although there are examples where

human genetic studies use lcWGS and imputation for genotyping, they typically require higher coverage because of more diverse and smaller haplotype blocks (Cai et al. 2015; Petter et al. 2020; Li et al. 2021, 2024; Wasik et al. 2021). Our genotyping method takes advantage of the unique HS population structure caused by interbreeding 8 inbred founders. Because the founders are fully sequenced, we are able to construct a high-quality reference panel for HS rats, which enables us to achieve highly accurate imputation for their genotypes even with low read coverage (0.21x mean ddGBS and 0.27x mean lcWGS).



**Fig. 5.** Overlapping first and second PCs on genotypes shows no batch effects between different sequencing methods.

Others have reported a similar genotyping strategy of using GBS or lcWGS alone and imputation for AIL and CFW mice (Nicod *et al.* 2016; Parker *et al.* 2016; Gonzales *et al.* 2018). Nicod *et al.* (2016) used lcWGS sequence data with STITCH imputation on CFW mice. Parker *et al.* (2016) used GBS sequence data with IMPUTE2 on CFW mice. Gonzales *et al.* (2018) used GBS sequence data with BEAGLE on AIL mice. Their estimated genotype concordance rates were 98.1, 97.0, and 96.96% respectively compared to the MegaMUGA array. Our previous work of using ddGBS and imputation (2 rounds of imputation: BEAGLE and IMPUTE2) to genotype HS rats was able to produce over 3.7 million SNPs with a concordance rate of 99.0% compared to a custom Affymetrix Axiom MiRat 625k microarray (Gileta *et al.* 2020). These 4 studies also included a variant calling step to identify candidate variants using either ANGSD or GATK before imputation. Our genotyping method described here does not require such a variant calling step, which reduces computation. Our method combines ddGBS and lcWGS sequence data and uses STITCH in conjunction with a fully sequenced founder reference panel to achieve genotype imputation. As a result, we achieve a high genotype concordance rate (>99.76%) compared to high-coverage (33.26x coverage) WGS.

Our genotyping method provides a robust method for genotyping HS rats by effectively imputing SNP genotypes from 2 different sequencing protocols without significant batch effects. Even with low read coverage, our method produced highly accurate genotypes. Our method is cost-effective due to the previously developed affordable ddGBS technique and low-cost lcWGS, which use commercially available library preparation kits and liquid-handling robots, improving throughput. Additionally, our method combines ddGBS and lcWGS sequences for genotype imputation, enabling old ddGBS genotyped rats to be analyzed in tandem with more recently genotyped HS rats.

The differences we observed in aligned sequence statistics between ddGBS and lcWGS (Fig. 2) reflect the different nature of the DNA sequences captured by 2 sequencing methods. Double restriction enzyme digestion limits ddGBS to only capture the DNA fragments near the enzyme cut sites, while random priming helps

lcWGS capture DNA fragments across the genome randomly. Despite the differences in captured DNA fragments, the genotype concordances of imputed SNPs for both ddGBS and lcWGS are remarkably high, at 99.76 and 99.78%, respectively. This concordance demonstrates the strength of our pipeline in producing high-accuracy genotypes in HS rats, which provides a strong foundation for genetic studies in this population.

The GBS sequencing method was originally developed by Elshire *et al.* (2011) and modified to accommodate other species such as soybean (Sonah *et al.* 2013), rice (Furuta *et al.* 2017), oat (Fu 2018), chicken (Pétille *et al.* 2016; Wang *et al.* 2017), fox (Johnson *et al.* 2015), cattle (Donato *et al.* 2013), and mouse (Parker *et al.* 2016; Gonzales *et al.* 2018). Our lab modified GBS for use in HS rats (Gileta *et al.* 2020). In this study, we further improved our genotyping methods by harmonizing the previously produced ddGBS sequences and newly sequenced lcWGS sequences with commercial WGS technique in support of large-scale genetic studies. The principles of our genotyping method can be easily adapted for use in other populations, especially for those in which the founders are fully sequenced.

In summary, we developed a genotyping method for HS rats that is both cost-effective and high-throughput, yielding highly accurate genotypes. Our method can be readily applied to other species with minimal adjustments, forming a basis for conducting extensive genetic research in nonhuman species.

## Data availability

HS rats are available at <https://ratgenes.org/cores/core-b/>. Wet lab procedures are documented in protocols.io <https://www.protocols.io/workspaces/cgord> (spleen cutting: <http://dx.doi.org/10.17504/protocols.io.36wgq7nryvk5/v1>, DNA extraction: <http://dx.doi.org/10.17504/protocols.io.8epv59reng1b/v1>, normalization and randomization: <http://dx.doi.org/10.17504/protocols.io.261genw5dg47/v1>, library preparation: <http://dx.doi.org/10.17504/protocols.io.j8nlkkm85l5r/v1>, pooling and sequencing: <http://dx.doi.org/10.17504/protocols.io.yxmvnmw29g3p/v1>). Raw sequencing reads for ddGBS and lcWGS are available in NCBI SRA: PRJNA1022514. Eight HS inbred founder WGS raw reads are available in NCBI SRA: PRJNA487943 and PRJNA1048943. Eighty-eight selected HS rat WGS raw reads are available in NCBI SRA: PRJNA1076141. High-coverage WGS GATK genotyping pipeline code is available in Zenodo (<https://doi.org/10.5281/zenodo.6584834>) and GitHub (<https://github.com/Palmer-Lab-UCSD/High-Coverage-WGS-GATK-Genotyping-Pipeline>). High-coverage WGS DeepVariant genotyping pipeline code is available in Zenodo (<https://doi.org/10.5281/zenodo.10027133>) and GitHub (<https://github.com/Palmer-Lab-UCSD/High-Coverage-WGS-DeepVariant-Genotyping-Pipeline>). Genotyping pipeline and analysis code is available in Zenodo (<https://doi.org/10.5281/zenodo.10002191>) and GitHub (<https://github.com/Palmer-Lab-UCSD/HS-Rats-Genotyping-Pipeline>). Genotype data after quality control are available in UC San Diego Library Digital Collections (<https://doi.org/10.6075/J0445MPC>).

Supplemental material available at G3 online.

## Acknowledgments

This publication includes data generated at the UC San Diego IGM Genomics Center: San Diego Supercomputer Center (2022) (Triton Shared Computing Cluster); University of California, San Diego Service (<https://doi.org/10.57873/T34W2R>).



## Funding

This work was funded by National Institute on Drug Abuse P50DA037844.

## Conflicts of interest

The authors declare no conflicts of interest.

## Literature cited

- Abdellaoui A, Yengo L, Verweij KJH, Visscher PM. 2023. 15 years of GWAS discovery: realizing the promise. *Am J Hum Genet.* 110(2):179–194. doi:10.1016/j.ajhg.2022.12.011.
- Aganezov S, Yan SM, Soto DC, Kirsche M, Zarate S, Avdeyev P, Taylor DJ, Shafin K, Shumate A, Xiao C, et al. 2022. A complete reference genome improves analysis of human genetic variation. *Science.* 376(6588):eabl3533. doi:10.1126/science.abl3533.
- Alliance of Genome Resources Consortium. 2022. Harmonizing model organism data in the Alliance of Genome Resources. *Genetics.* 220(4):iyac022. doi:10.1093/genetics/iyac022.
- Baud A, Hermsen R, Guryev V, Stridh P, Graham D, McBride MW, Foroud T, Calderari S, Diez M, Ockinger J, et al. 2013. Combined sequence-based and genetic mapping analysis of complex traits in outbred rats. *Nat Genet.* 45(7):767–775. doi:10.1038/ng.2644.
- Broad Institute. 2019. Picard Toolkit. [accessed 2023 Apr 10]. <https://broadinstitute.github.io/picard/>.
- Bushnell B. BBTools. <http://sourceforge.net/projects/bbmap/2014>.
- Cai N, Bigdeli TB, Kretschmar W, Li Y, Liang J, Song L, Hu J, Li Q, Jin W, Hu Z, et al. 2015. Sparse whole-genome sequencing identifies two loci for major depressive disorder. *Nature.* 523(7562):588–591. doi:10.1038/nature14659.
- Chitre AS, Hebda-Bauer EK, Blandino P, Bimschleger H, Nguyen K-M, Maras P, Li F, Ozel AB, Pan Y, Polesskaya O, et al. 2023. Genome-wide association study in a rat model of temperament identifies multiple loci for exploratory locomotion and anxiety-like traits. *Front Genet.* 13:1003074. doi:10.3389/fgene.2022.1003074.
- Chitre AS, Polesskaya O, Holl K, Gao J, Cheng R, Bimschleger H, Garcia Martinez A, George T, Gileta AF, Han W, et al. 2020. Genome-wide association study in 3,173 outbred rats identifies multiple loci for body weight, adiposity, and fasting glucose. *Obesity.* 28(10):1964–1973. doi:10.1002/oby.22927.
- Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, et al. 2021. Twelve years of SAMtools and BCFtools. *GigaScience.* 10(2):giab008. doi:10.1093/gigascience/giab008.
- Davies RW, Flint J, Myers S, Mott R. 2016. Rapid genotype imputation from sequence without reference panels. *Nat Genet.* 48(8):965–969. doi:10.1038/ng.3594.
- Didion JP, Yang H, Sheppard K, Fu C-P, McMillan L, de Villena FP-M, Churchill GA. 2012. Discovery of novel variants in genotyping arrays improves genotype retention and reduces ascertainment bias. *BMC Genomics.* 13(1):34. doi:10.1186/1471-2164-13-34.
- Donato MD, Peters SO, Mitchell SE, Hussain T, Imumorin IG. 2013. Genotyping-by-sequencing (GBS): a novel, efficient and cost-effective genotyping method for cattle using next-generation sequencing. *PLoS One.* 8(5):e62137. doi:10.1371/journal.pone.0062137.
- Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity Species. *PLoS One.* 6(5):e19379. doi:10.1371/journal.pone.0019379.
- Fowler S, Wang T, Munro D, Kumar A, Chitre AS, Hollingsworth TJ, Garcia Martinez A, St. Pierre CL, Bimschleger H, Gao J, et al. 2023. Genome-wide association study finds multiple loci associated with intraocular pressure in HS rats. *Front Genet.* 13:1029058. doi:10.3389/fgene.2022.1029058.
- Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, et al. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature.* 449(7164):851–861. doi:10.1038/nature06258.
- Fu Y-B. 2018. Oat evolution revealed in the maternal lineages of 25 *Avena* species. *Sci Rep.* 8(1):4252. doi:10.1038/s41598-018-22478-4.
- Furuta T, Ashikari M, Jena KK, Doi K, Reuscher S. 2017. Adapting genotyping-by-sequencing for rice F2 populations. *G3 (Bethesda).* 7(3):881–893. doi:10.1534/g3.116.038190.
- Gileta AF, Fitzpatrick CJ, Chitre AS, Pierre CLS, Joyce EV, Maguire RJ, McLeod AM, Gonzales NM, Williams AE, Morrow JD, et al. 2022. Genetic characterization of outbred Sprague Dawley rats and utility for genome-wide association studies. *PLOS Genet.* 18(5):e1010234. doi:10.1371/journal.pgen.1010234.
- Gileta AF, Gao J, Chitre AS, Bimschleger HV, St. Pierre CL, Gopalakrishnan S, Palmer AA. 2020. Adapting genotyping-by-sequencing and variant calling for heterogeneous stock rats. *G3 (Bethesda).* 10(7):2195–2205. doi:10.1534/g3.120.401325.
- Gonzales NM, Seo J, Hernandez Cordero AI, St. Pierre CL, Gregory JS, Distler MG, Abney M, Canzar S, Lionikas A, Palmer AA. 2018. Genome wide association analysis in a mouse advanced intercross line. *Nat Commun.* 9(1):5162. doi:10.1038/s41467-018-07642-8.
- Gunturkun MH, Wang T, Chitre AS, Garcia Martinez A, Holl K, St. Pierre C, Bimschleger H, Gao J, Cheng R, Polesskaya O, et al. 2022. Genome-wide association study on three behaviors tested in an open field in heterogeneous stock rats identifies multiple loci implicated in psychiatric disorders. *Front Psychiatry.* 13:790566. doi:10.3389/fpsy.2022.790566.
- Hannon Lab. FASTX-Toolkit. [accessed 2023 Apr 10]. <https://www.hannonlab.org/resources/2010>.
- Johannesson M, Lopez-Aumatell R, Stridh P, Diez M, Tuncel J, Blázquez G, Martinez-Membrives E, Cañete T, Vicens-Costa E, Graham D, et al. 2009. A resource for the simultaneous high-resolution mapping of multiple quantitative trait loci in rats: the NIH heterogeneous stock. *Genome Res.* 19(1):150–158. doi:10.1101/gr.081497.108.
- Johnson JL, Wittgenstein H, Mitchell SE, Hyma KE, Temnykh SV, Kharlamova AV, Gulevich RG, Vladimirova AV, Fong HWF, Acland GM, et al. 2015. Genotyping-by-sequencing (GBS) detects genetic structure and confirms behavioral QTL in tame and aggressive foxes (*Vulpes vulpes*). *PLoS One.* 10(6):e0127013. doi:10.1371/journal.pone.0127013.
- Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM [preprint]. arXiv:1303.3997v2 [q-bio.GN]. <https://doi.org/10.48550/arXiv.1303.3997>.
- Li JH, Findley K, Pickrell JK, Bleasdale K, Zhao J, Kruglyak S. 2024. Low-pass sequencing plus imputation using avidity sequencing displays comparable imputation accuracy to sequencing by synthesis while reducing duplicates. *G3 (Bethesda).* 14(2):jkad276. doi:10.1093/g3journal/jkad276.
- Li JH, Mazur CA, Berisa T, Pickrell JK. 2021. Low-pass sequencing increases the power of GWAS and decreases measurement error of polygenic risk scores compared to genotyping arrays. *Genome Res.* 31(4):529–537. doi:10.1101/gr.266486.120.
- Marchini J, Howie B. 2010. Genotype imputation for genome-wide association studies. *Nat Rev Genet.* 11(7):499–511. doi:10.1038/nrg2796.

- Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* 17(1):10–12. doi:[10.14806/ej.17.1.200](https://doi.org/10.14806/ej.17.1.200).
- McVean GA, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, Donnelly P, Eichler EE, Flicek P, et al. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature.* 491(7422):56–65. doi:[10.1038/nature11632](https://doi.org/10.1038/nature11632).
- Munro D, Wang T, Chitre AS, Polesskaya O, Ehsan N, Gao J, Gusev A, Woods LCS, Saba LM, Chen H, et al. 2022. The regulatory landscape of multiple brain regions in outbred heterogeneous stock rats. *Nucleic Acids Res.* 50(19):10882–10895. doi:[10.1093/nar/gkac912](https://doi.org/10.1093/nar/gkac912).
- Nicod J, Davies RW, Cai N, Hassett C, Goodstadt L, Cosgrove C, Yee BK, Lionikaite V, McIntyre RE, Remme CA, et al. 2016. Genome-wide association of multiple complex traits in outbred mice by ultra-low-coverage sequencing. *Nat Genet.* 48(8):912–918. doi:[10.1038/ng.3595](https://doi.org/10.1038/ng.3595).
- Okamoto F, Chitre AS, Missfeldt Sanches T, Chen D, Munro D, Aron AT, Beeson A, Bimschleger HV, Eid M, Garcia Martinez AG, et al. 2024. Y and mitochondrial chromosomes in the heterogeneous stock rat population. *G3-Genes Genom Genet.* 14(11). doi:[10.1093/g3journal/jkae213](https://doi.org/10.1093/g3journal/jkae213).
- Palmer RHC, Johnson EC, Won H, Polimanti R, Kapoor M, Chitre A, Bogue MA, Benca-Bachman CE, Parker CC, Verma A, et al. 2021. Integration of evidence across human and model organism studies: a meeting report. *Genes Brain Behav.* 20(6):e12738. doi:[10.1111/gbb.12738](https://doi.org/10.1111/gbb.12738).
- Parker CC, Gopalakrishnan S, Carbonetto P, Gonzales NM, Leung E, Park YJ, Aryee E, Davis J, Blizard DA, Ackert-Bicknell CL, et al. 2016. Genome-wide association study of behavioral, physiological and gene expression traits in outbred CFW mice. *Nat Genet.* 48(8):919–926. doi:[10.1038/ng.3609](https://doi.org/10.1038/ng.3609).
- Parker CC, Philip VM, Gatti DM, Kasperek S, Kreuzman AM, Kuffler L, Mansky B, Masneuf S, Sharif K, Sluys E, et al. 2022. Genome-wide association mapping of ethanol sensitivity in the diversity outbred mouse population. *Alcohol Clin Exp Res.* 46(6):941–960. doi:[10.1111/acer.14825](https://doi.org/10.1111/acer.14825).
- Pértille F, Guerrero-Bosagna C, Silva Vd, Boschiero C, Nunes JdRdS, Ledur MC, Jensen P, Coutinho LL. 2016. High-throughput and cost-effective chicken genotyping using next-generation sequencing. *Sci Rep.* 6(1):26929. doi:[10.1038/srep26929](https://doi.org/10.1038/srep26929).
- Petter E, Schweiger R, Shahino B, Shor T, Aker M, Almog L, Weissglas-Volkov D, Naveh Y, Navon O, Carmi S, et al. 2020. Relative matching using low coverage sequencing. *bioRxiv* 289322. <https://doi.org/10.1101/2020.09.09.289322>, preprint: not peer reviewed.
- Ramdas S, Ozel AB, Treutelaar MK, Holl K, Mandel M, Woods LCS, Li JZ. 2019. Extended regions of suspected mis-assembly in the rat reference genome. *Sci Data.* 6(1):39. doi:[10.1038/s41597-019-0041-6](https://doi.org/10.1038/s41597-019-0041-6).
- Solberg Woods LC, Palmer AA. 2019. Using heterogeneous stocks for fine-mapping genetically complex traits. In: Hayman GT, Smith JR, Dwinell MR, Shimoyama M, editors. *Rat Genomics*. New York (NY): Springer. p. 233–247.
- Sonah H, Bastien M, Iquira E, Tardivel A, Légaré G, Boyle B, Normandeau É, Laroche J, Larose S, Jean M, et al. 2013. An improved genotyping by sequencing (GBS) approach offering increased versatility and efficiency of SNP discovery and genotyping. *PLoS One.* 8(1):e54603. doi:[10.1371/journal.pone.0054603](https://doi.org/10.1371/journal.pone.0054603).
- Tim F, Nils H. 2023. *fgbio*. [accessed 2023 Apr 10]. <https://github.com/fulcrumgenomics/fgbio>.
- Uffelmann E, Huang QQ, Munung NS, de Vries J, Okada Y, Martin AR, Martin HC, Lappalainen T, Posthuma D. 2021. Genome-wide association studies. *Nat Rev Methods Primer.* 1(1):1–21. doi:[10.1038/s43586-021-00056-9](https://doi.org/10.1038/s43586-021-00056-9).
- Van der Auwera GA, O'Connor BD. 2020. *Genomics in the cloud: using Docker, GATK, and WDL in Terra*. 1st ed. Sebastopol, CA: O'Reilly Media.
- Wang Y, Cao X, Zhao Y, Fei J, Hu X, Li N. 2017. Optimized double-digest genotyping by sequencing (ddGBS) method with high-density SNP markers and high genotyping accuracy for chickens. *PLoS One.* 12(6):e0179073. doi:[10.1371/journal.pone.0179073](https://doi.org/10.1371/journal.pone.0179073).
- Wasik K, Berisa T, Pickrell JK, Li JH, Fraser DJ, King K, Cox C. 2021. Comparing low-pass sequencing and genotyping for trait mapping in pharmacogenetics. *BMC Genomics.* 22(1):197. doi:[10.1186/s12864-021-07508-2](https://doi.org/10.1186/s12864-021-07508-2).
- Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L, et al. 2014. The NHGRI GWAS catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* 42(D1):D1001–D1006. doi:[10.1093/nar/gkt1229](https://doi.org/10.1093/nar/gkt1229).
- Woods LCS, Mott R. 2017. Heterogeneous stock populations for analysis of complex traits. In: Schughart K, Williams RW, editors. *Systems Genetics: Methods and Protocols (Methods in Molecular Biology)*. New York (NY): Springer. p. 31–44.
- Zhou X, St. Pierre CL, Gonzales NM, Zou J, Cheng R, Chitre AS, Sokoloff G, Palmer AA. 2020. Genome-wide association study in two cohorts from a multi-generational mouse advanced intercross line highlights the difficulty of replication due to study-specific heterogeneity. *G3 (Bethesda).* 10(3):951–965. doi:[10.1534/g3.119.400763](https://doi.org/10.1534/g3.119.400763).
- Zou J, Gopalakrishnan S, Parker CC, Nicod J, Mott R, Cai N, Lionikas A, Davies RW, Palmer AA, Flint J. 2022. Analysis of independent cohorts of outbred CFW mice reveals novel loci for behavioral and physiological traits and identifies factors determining reproducibility. *G3 (Bethesda).* 12(1):jkab394. doi:[10.1093/g3journal/jkab394](https://doi.org/10.1093/g3journal/jkab394).