

1 **Exponential increase in QTL detection with increased sample size**

2 **Authors:** Apurva S. Chitre¹, Oksana Polesskaya¹, Daniel Munro^{1,13}, Riyan Cheng¹, Pejman
3 Mohammadi^{13,14}, Katie Holl², Jianjun Gao¹, , Hannah Bimschleger¹, Angel Garcia Martinez³,
4 Anthony George⁶, Alexander F. Gileta^{1,10}, Aidan Horvath⁴, Alesa Hughson⁴, Keita Ishiwari⁶,
5 Christopher P. King⁵, Alexander Lamparelli⁵, Cassandra L. Versaggi⁵, Connor Martin⁶, Celine L.
6 St. Pierre¹¹, Jordan A. Tripi⁵, Jerry B. Richards⁶, Tengfei Wang³, Hao Chen³, Shelly B. Flagel¹²,
7 Paul Meyer⁵, Terry E. Robinson⁷, Leah C. Solberg Woods⁹, Abraham A. Palmer^{1,8}

8 **Affiliations:**

9 1 Department of Psychiatry, University of California San Diego

10 2 Human and Molecular Genetic Center, Medical College of Wisconsin

11 3 Department of Pharmacology, University of Tennessee Health Science Center

12 4 Department of Psychiatry, University of Michigan

13 5 Department of Psychology, University at Buffalo

14 6 Clinical and Research Institute on Addictions, University at Buffalo

15 7 Department of Psychology, University of Michigan

16 8 Institute for Genomic Medicine, University of California San Diego

17 9 Department of Internal Medicine, Wake Forest School of Medicine

18 10 Department of Human Genetics, University of Chicago

19 11 Department of Genetics, Washington University in St. Louis

20 12 Molecular and Behavioral Neuroscience Institute, University of Michigan

21 13 Department of Integrative Structural and Computational Biology, Scripps Research

22 14 Scripps Research Translational Institute, Scripps Research

23

24 **Abstract**

25 Power analyses are often used to determine the number of animals required for a
26 genome wide association analysis (GWAS). These analyses are typically intended to estimate
27 the sample size needed for at least one locus to exceed a genome-wide significance threshold.
28 A related question that is less commonly considered is the number of significant loci that will be
29 discovered with a given sample size. We used simulations based on a real dataset that
30 consisted of 3,173 male and female adult N/NIH heterogeneous stock (HS) rats to explore the
31 relationship between sample size and the number of significant loci discovered. Our simulations
32 examined the number of loci identified in sub-samples of the full dataset. The sub-sampling
33 analysis was conducted for four traits with low (0.15 ± 0.03), medium (0.31 ± 0.03 and $0.36 \pm$
34 0.03) and high (0.46 ± 0.03) SNP-based heritabilities. For each trait, we sub-sampled the data
35 100 times at different sample sizes (500, 1,000, 1,500, 2,000, and 2,500). We observed an
36 exponential increase in the number of significant loci with larger sample sizes. Our results are
37 consistent with similar observations in human GWAS and imply that future rodent GWAS should
38 use sample sizes that are significantly larger than those needed to obtain a single significant
39 result.

40 **Introduction**

41 Genome wide association studies (GWAS) in both humans and rodents have been
42 extremely successful in understanding the genetics of quantitative traits. Outbred rodent
43 populations such as Heterogeneous stock (HS) rats, Diversity Outbred (DO) mice, and
44 Advanced Intercross Lines (AIL) have proven to be an invaluable resource for genetic mapping
45 studies. The success of these outbred rodent strains can be attributed to the ability to provide
46 high resolution QTL mapping (Solberg Woods and Palmer 2019). With each generation of
47 recombination, the number of markers and independent tests increases, which in turn increases
48 the threshold for statistical significance. In comparison to an F_2 cross, outbred rodent
49 populations offer better resolution for mapping QTLs (Solberg Woods 2014; Gonzales and
50 Palmer 2014). Inbred rodent strains such as the Hybrid Rat Diversity panels (HRDP), Hybrid
51 Mouse Diversity Panels (HMDP) and Recombinant Inbred (RI) strains (such as the BXD and CC
52 panels) have also been successfully employed for mapping studies (Williams and Williams
53 2017). However, the sample size involving these panels is usually limited by the number of
54 strains available in the population. QTL mapping studies are not limited to rodent populations.
55 These genetic studies are also conducted in zebrafish (Kwon et al. 2019), fruit flies (Wangler et
56 al. 2017) and plants such as *Arabidopsis thaliana* (Togninalli et al. 2020).

57 In GWAS studies power is defined as the likelihood of detecting a single significant QTL
58 of a certain effect size. Power analyses are often performed for GWAS studies so that an
59 appropriate sample size can be selected. In general, larger sample sizes increase the power to
60 detect significant loci in humans (Spencer et al. 2009), rodents (Li et al. 2006; Keele et al.
61 2019), livestock (Wittenburg et al. 2020) and crops (Wang and Xu 2019). Software to perform
62 power analyses has also typically focused on power to detect a single locus given its effect size
63 (Sen et al. 2007; Delongchamp et al. 2018).

64 In this study, we sought to examine a related question, namely the relationship between
65 sample size and the number of significant loci discovered. We used simulations based on a real
66 dataset that consisted of 3,173 male and female adult N/NIH heterogeneous stock (HS) rats.
67 This dataset is part of our recent publication on the GWAS of obesity related traits in HS rats,
68 which is among the largest rodent GWAS ever performed (Chitre et al. 2020). The dataset in
69 Chitre et al. was collected as part of a large multi-site project focused on genetic analyses of
70 behavioral phenotypes related to drug abuse in HS rats (www.ratgenes.org). We repeatedly
71 subsampled this dataset to determine the number of significant loci that could be identified with
72 various sample sizes.

73

74 **Results**

75 The number of significant loci discovered increased exponentially as sample size
76 increased. **Figure 1** shows the average number of significant loci detected for each trait at each
77 sample size. When we ran the analysis with the maximum number of individuals, we detected
78 28 loci for body weight ($h^2 = 0.46 \pm 0.03$), 16 loci for body length_Tail ($h^2 = 0.36 \pm 0.03$), 5 loci
79 for BMI_Tail ($h^2 = 0.31 \pm 0.03$) and 3 for fasting glucose ($h^2 = 0.15 \pm 0.03$). As expected, fewer
80 QTLs were discovered with smaller sample sizes. We note the largest increase in the number of
81 QTL detected for body weight, the trait with the highest heritability, with more than a ten-fold
82 increase in detected QTL when the sample size is increased from 500 to 2500. Similar trends
83 are seen for both BMI and fasting glucose.

84 To determine whether the increase in the number of significant loci was more consistent
85 with a linear or an exponential (log-linear) function, we fitted both models on the data to identify
86 least squares parameters. The two models we defined as

87

88 Linear: $y = b_0 x + b_1 + e$,

89 and

90 Exponential: $y = \exp(b_0 x + b_1) + e$

91

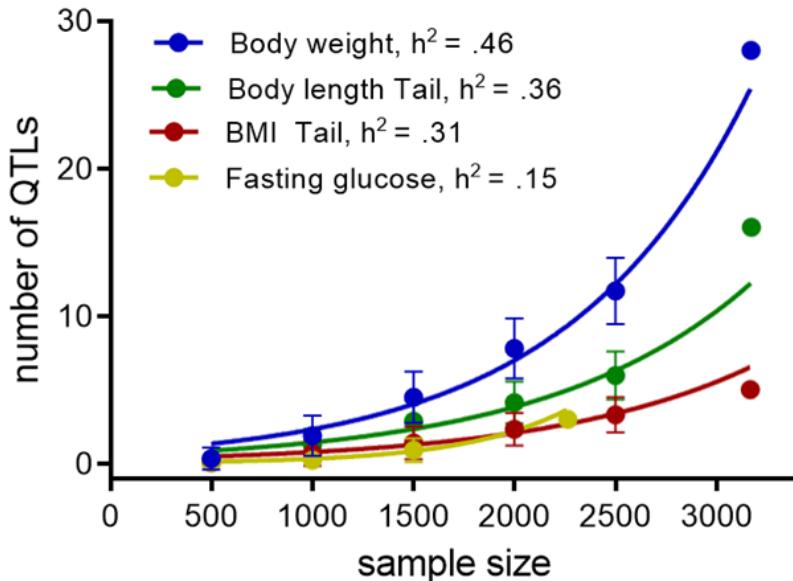
92 where, b_0 and b_1 are the model parameters, x is the sample size, y is the average number of
93 QTLs and e is the error term.

94

95 Since both models have the same number of parameters we compared them in terms of
96 residual sum of squares (RSS) and used bootstrapping to ascertain statistical boundaries of the
97 estimates. We found that an exponential curve fits better; the estimated 95% confidence interval
98 is (0.344, 7.67) for the exponential fit, and (9.57, 217.522) for the linear fit.

99

100



101

102 **Figure 1. Number of detected QTLs increases with the increase of sample size.** Each dot is
103 an average number of QTLs obtained in 100 GWAS, each performed on a randomly selected
104 subset of the actual dataset. Error bars indicate standard deviation. The final point (at ~3100
105 animals for body weight, body length_Tail, BMI_Tail and at 2,246 for fasting glucose) used the
106 full dataset and therefore does not include error bars. This simulation was performed on four
107 traits with different heritability: body weight ($h^2 = 0.46 \pm 0.03$), body length_Tail ($h^2 = 0.36 \pm$
108 0.03), BMI_Tail ($h^2 = 0.31 \pm 0.03$) and fasting glucose ($h^2 = 0.15 \pm 0.03$).

109

110 We also conducted linkage analysis using haplotypes to confirm that our findings were in
111 agreement with those obtained from GWAS analysis that utilized SNPs. We found that, similar
112 to the GWAS analysis, an exponential increase in the number of QTL identified with increasing
113 sample size for linkage analysis using haplotypes (**Supplementary Figure 1**). We performed
114 this analysis for BMI with tail ($SNP\ h^2 = 0.31 \pm 0.03$) using R/qltl2 (Broman et al. 2019) using a
115 permutation derived threshold of 18.2 LOD at alpha = 0.05. We used the residual sum of
116 squares to compare the linear and exponential models. The RSS values for exponential fit
117 (0.069) are smaller than for linear fit (1.217), suggesting that an exponential curve fits better
118 than a linear curve.

119

120 **Discussion**

121 In this study, we used a real dataset to explore the effect of sample size on the number
122 of significant loci identified. This represents a conceptually different approach compared to
123 conventional power analyses, which focus on estimating power to detect at least one genome-
124 wide significant locus. We found an exponential increase in the number of QTL identified with

125 increasing sample size, particularly for body weight, a trait with relatively high heritability. Our
126 results suggest (but do not prove) that our findings would generalize to other similar laboratory
127 populations (HS/Npt, HS-CC, DO, etc.). The results from the haplotype-based linkage mapping
128 analysis also support an exponential increase in the number of QTL identified with increasing
129 sample size for the trait BMI with tail ($SNP\ h^2 = 0.31 \pm 0.03$).

130 Similar observations in human genetics (Visscher et al. 2012; Sullivan et al. 2018)
131 suggest an initial exponential growth in the number of loci, which is what we have observed,
132 followed by a linear phase when increasing sample size produces a linear increase in the
133 number of significant loci. In the current study, we did not find strong evidence of this linear
134 phase. This could reflect the fact that our sample size, which is still small by the standards of
135 human genetics, was not large enough to get beyond the initial exponential phase. As is the
136 case in human GWAS, the effect size of loci that require larger sample sizes will tend to be
137 smaller than those identified with larger sample sizes, assuming a constant allele frequency.
138 There are several reasons that this dataset was able to identify multiple significant loci despite
139 having a sample size that is smaller than those typically used for human GWAS. First, the effect
140 sizes of alleles discovered in model systems are often much larger than alleles found in
141 humans. The reasons for this are unknown but might include relaxed selection in captive
142 breeding populations, which allows alleles that would have been selected against in a natural
143 population to rise to high frequency. A second reason that smaller sample sizes are sufficient in
144 model systems is that the linkage disequilibrium among SNPs is greater, meaning that fewer
145 tests are performed, thus reducing the multiple testing burden and correspondingly the
146 threshold for significance. The greater LD might also mean that multiple smaller alleles are
147 inherited in blocks that have greater effect sizes. A third advantage of model systems is that
148 they are often created by crossing a small number of inbred strains, meaning that allele
149 frequencies are higher; greater power is always available when alleles are more common.
150 Despite these differences, our observation of exponential growth in the number of significant loci
151 with increasing sample size is very similar to observations in human genetics.

152 Our result indicates that many previous studies, which have performed power analyses
153 designed to assure that they find a single significant locus, are likely underpowered to find
154 multiple loci that have diminishing effect sizes. Our recommendation is that future studies of
155 complex traits in outbred rodents should use significantly larger sample sizes since they are
156 likely to provide a larger number of findings; this recommendation assumes that the cost of
157 increasing sample size is linear, however in some cases there might be efficiencies of scale that

158 would make the addition of each additional subject less expensive. Fewer studies with larger
159 sample sizes, rather than larger numbers of studies with modest sample sizes might be
160 preferable. Alternatively, multiple traits from separate studies that are genetically correlated
161 might be jointly analyzed, since this can provide some of the advantages of larger sample sizes
162 assuming that certain loci are important for more than one of the traits under study.

163

164 **Materials and Methods**

165 The data used in this study are thoroughly described in our recent publication (Chitre et
166 al. 2020). Briefly, phenotypic data on body weight and length (which permit calculation of BMI),
167 fat pad weight, and fasting glucose levels of adiposity traits were collected at three different
168 research sites at multiple ages. We used data from all three sites. Prior to combining data from
169 the three sites, we regressed out the effects of covariates and then performed quantile
170 normalizations within each site and within each sex after which data from all sites and sexes
171 were combined and jointly analyzed to explore the relationship between sample size and the
172 number of significant loci identified.

173 HS rats used in this study were obtained from the NMci:HS colony which was initiated
174 by the NIH in 1984 by interbreeding eight inbred founder strains and were subsequently
175 maintained as an outbred population, making them ideal for fine mapping of genetic loci
176 (Hansen and Spuhler 1984; Solberg Woods and Palmer 2019). Rats were genotyped at 3.4
177 million autosomal SNPs, however, because there was extensive LD among these SNPs and to
178 reduce computational burden, we used LD pruning ($r^2 < 0.95$) which yielded a set of 128,477
179 SNPs that were used for all analyses described in this paper.

180 To determine the number of QTLs detected by different samples sizes, we subsampled
181 data from four phenotypes chosen to have low (0.15 ± 0.03 ; fasting glucose), medium ($0.36 \pm$
182 0.030 ; body length_Tail and $.31 \pm 0.03$; BMI) and high (0.46 ± 0.03 ; body weight) chip
183 heritabilities (calculated using GCTA). For each dataset, we performed 100 random
184 subsamples in which we retained 500, 1,000, 1,500, 2,000, or 2,500 individuals (for fasting
185 glucose we could not include 2,000 and 2,500 because the total sample size was smaller than
186 2,000). Thus, we produced 1,300 total subsamples for the three phenotypes. We then
187 performed a GWAS for each subsampled dataset using an automated pipeline based on the
188 LMM software package GEMMA (Zhou and Stephens 2012); we implemented the leave one
189 chromosome out (LOCO) method (Cheng et al. 2013). We have previously shown that an LMM

190 in conjunction with the LOCO methods effectively controls type I error rate (Gonzales et al.
191 2018; Gileta et al. 2022), meaning that our observations in this study are unlikely to be due to
192 type I errors that can be caused by population structure.

193 Our pipeline used an algorithm to automatically record the number of significant QTLs in
194 each subsampled dataset. First, we scanned each chromosome to determine if there was at
195 least one SNP that exceeded the threshold of $-\log_{10}(p) > 5.6$, which is the threshold used in
196 Chitre et al. 2020. To avoid situations where only a single, presumably anomalous, SNP
197 showed a significant association, we required that at least one other SNP within 0.5 Mb have a
198 p-value that was within $2 - \log_{10}(p)$ of the index SNP. If we found a second supporting SNP, we
199 recorded the identification of a QTL for that dataset. Some chromosomes were expected to
200 contain more than one independent QTL, but we were also concerned that we might count a
201 single locus twice. To avoid counting the same locus twice, we excluded all SNPs with $r^2 > 0.4$
202 relative to the just identified index SNP. We then rescanned the chromosome to see if any
203 additional SNPs on this chromosome exceeded the threshold of $-\log_{10}(p) > 5.6$. If they did and
204 they were supported by a second SNP within 0.5 Mb that had a p-value that was within $2 -$
205 $\log_{10}(p)$ of the index SNP, we recorded an additional QTL for that dataset. We then repeated
206 these steps as often as needed until no further significant QTLs could be identified on a given
207 chromosome. We then continued this process for all subsequent chromosomes. After scanning
208 the last chromosome, we tabulated the number of QTLs detected for that dataset. We repeated
209 this procedure for each of the 1,300 subsampled datasets. In this way, we determined the
210 number of significant QTLs in 100 possible sub-samplings of each of four traits when using
211 500, 1,000, 1,500, 2,000, and 2,500 individuals, and in the maximal number of individuals
212 (~3100 for all traits except fasting glucose).

213 We performed linkage mapping with haplotypes using R/qltl2 (Broman et al. 2019). We
214 estimated founder haplotypes using the calc_genoprob_fst function with the cohort and founder
215 strain genotypes. We used the scan1perm function to perform 1,000 permutations for
216 establishing the significance threshold. The kinship matrices were derived using the “leave one
217 chromosome out” method with the calc_kinship function. For each sub-sampled dataset for the
218 trait BMI with tail, we performed a genome scan using a linear mixed model with the scan1
219 function. We used the function find_peaks to identify LOD peaks that exceeded the permutation
220 derived threshold of 18.2.

221

222 **Data availability**

223 The data presented in the study are deposited in the UC San Diego Library Digital Collections
224 repository at <https://library.ucsd.edu/dc/object/bb9156620z> (DOI
225 <https://doi.org/10.6075/J0Q240F0>).

226

227 **Acknowledgements**

228 This work was supported by the National Institute on Drug Abuse (P50 DA037844) and the
229 National Institute of Diabetes and Digestive and Kidney Diseases (R01 DK106386)

230

231 **References**

232 Broman, K. W., Gatti, D. M., Simecek, P., Furlotte, N. A., Prins, P., Sen, Š., Yandell, B. S., &
233 Churchill, G. A. (2019). R/qtL2: Software for Mapping Quantitative Trait Loci with High-
234 Dimensional Data and Multiparent Populations. *Genetics*, 211(2), 495–502.
235 <https://doi.org/10.1534/genetics.118.301595>

236 Cheng, R., Parker, C. C., Abney, M., & Palmer, A. A. (2013). Practical Considerations
237 Regarding the Use of Genotype and Pedigree Data to Model Relatedness in the Context of
238 Genome-Wide Association Studies. *G3 Genes|Genomes|Genetics*, 3(10), 1861–1867.
239 <https://doi.org/10.1534/g3.113.007948>

240 Chitre AS, Polesskaya O, Holl K, Gao J, Cheng R, Bimschleger H, Garcia Martinez A, George
241 T, Gileta AF, Han W, et al. 2020. Genome-Wide Association Study in 3,173 Outbred Rats
242 Identifies Multiple Loci for Body Weight, Adiposity, and Fasting Glucose. *Obesity*. 28(10):1964–
243 1973.

244 Delongchamp R, Faramawi MF, Feingold E, Chung D, Abouelenein S. 2018. The Association
245 between SNPs and a Quantitative Trait: Power Calculation. *European Journal of Environment
246 and Public Health* [Internet]. [accessed 2020 Dec 28] 2(2). [http://www.ejeph.com/article/the-
association-between-snps-and-a-quantitative-trait-power-calculation-3925](http://www.ejeph.com/article/the-
247 association-between-snps-and-a-quantitative-trait-power-calculation-3925)

248 Gileta, A. F., Fitzpatrick, C. J., Chitre, A. S., St. Pierre, C. L., Joyce, E. V., Maguire, R. J.,
249 McLeod, A. M., Gonzales, N. M., Williams, A. E., Morrow, J. D., Robinson, T. E., Flagel, S. B., &
250 Palmer, A. A. (2022). Genetic characterization of outbred Sprague Dawley rats and utility for
251 genome-wide association studies. *PLOS Genetics*, 18(5), e1010234.
252 <https://doi.org/10.1371/journal.pgen.1010234>

253 Gonzales NM, Palmer AA. 2014. Fine-mapping QTLs in advanced intercross lines and other
254 outbred populations. *Mamm Genome*. 25(7–8):271–292.

255 Gonzales, N. M., Seo, J., Hernandez Cordero, A. I., St. Pierre, C. L., Gregory, J. S., Distler, M.
256 G., Abney, M., Canzar, S., Lionikas, A., & Palmer, A. A. (2018). Genome wide association

257 analysis in a mouse advanced intercross line. *Nature Communications*, 9(1), 5162.
258 <https://doi.org/10.1038/s41467-018-07642-8>

259 Hansen C, Spuhler K. 1984. Development of the National Institutes of Health genetically
260 heterogeneous rat stock. *Alcohol Clin Exp Res*. 8(5):477–479.

261 Keele GR, Crouse WL, Kelada SNP, Valdar W. 2019. Determinants of QTL Mapping Power in
262 the Realized Collaborative Cross. *G3*. 9(5):1707–1727.

263 Kwon RY, Watson CJ, Karasik D. 2019. Using zebrafish to study skeletal genomics. *Bone*.
264 126:37–50.

265 Li X, Quigg RJ, Zhou J, Xu S, Masinde G, Mohan S, Baylink DJ. 2006. A critical evaluation of
266 the effect of population size and phenotypic measurement on QTL detection and localization
267 using a large F2 murine mapping population. *Genet Mol Biol*. 29(1):166–173.

268 Sen S, Satagopan JM, Broman KW, Churchill GA. 2007. R/qtlDesign: inbred line cross
269 experimental design. *Mamm Genome*. 18(2):87–93.

270 Solberg Woods LC. 2014. QTL mapping in outbred populations: successes and challenges.
271 *Physiol Genomics*. 46(3):81–90.

272 Solberg Woods LC, Palmer AA. 2019. Using Heterogeneous Stocks for Fine-Mapping
273 Genetically Complex Traits. *Methods Mol Biol*. 2018:233–247.

274 Spencer CCA, Su Z, Donnelly P, Marchini J. 2009. Designing Genome-Wide Association
275 Studies: Sample Size, Power, Imputation, and the Choice of Genotyping Chip. Storey JD, editor.
276 *PLoS Genet*. 5(5):e1000477.

277 Sullivan PF, Agrawal A, Bulik CM, Andreassen OA, Børglum AD, Breen G, Cichon S, Edenberg
278 HJ, Faraone SV, Gelernter J, et al. 2018. Psychiatric Genomics: An Update and an Agenda. *Am*
279 *J Psychiatry*. 175(1):15–27.

280 Togninalli M, Seren Ü, Freudenthal JA, Monroe JG, Meng D, Nordborg M, Weigel D, Borgwardt
281 K, Korte A, Grimm DG. 2020. AraPheno and the AraGWAS Catalog 2020: a major database
282 update including RNA-Seq and knockout mutation data for *Arabidopsis thaliana*. *Nucleic Acids*
283 *Res*. 48(D1):D1063–D1068.

284 Visscher PM, Brown MA, McCarthy MI, Yang J. 2012. Five years of GWAS discovery. *Am J*
285 *Hum Genet*. 90(1):7–24.

286 Wang M, Xu S. 2019. Statistical power in genome-wide association studies and quantitative trait
287 locus mapping. *Heredity*. 123(3):287–306.

288 Wangler MF, Hu Y, Shulman JM. 2017. *Drosophila* and genome-wide association studies: a
289 review and resource for the functional dissection of human complex traits. *Dis Model Mech*.
290 10(2):77–88.

291 Williams RW, Williams EG. 2017. Resources for Systems Genetics. In: Schugart K, Williams
292 RW, editors. *Systems Genetics* [Internet]. Vol. 1488. New York, NY: Springer New York;
293 [accessed 2020 Dec 16]; p. 3–29. http://link.springer.com/10.1007/978-1-4939-6427-7_1

294 Zhou X, Stephens M. 2012. Genome-wide efficient mixed-model analysis for association
295 studies. *Nat Genet.* 44(7):821–824.