Type: Letter/Methods

selscan 2.0: scanning for sweeps in unphased data

Zachary A. Szpiech[1,2,*]

[1] *Department of Biology, Pennsylvania State University, University Park, PA 16802, USA*
[2] *Institute for Computational and Data Sciences, Pennsylvania State University, University Park, PA 16802, USA*
[*] *Correspondence: szpiech@psu.edu*

Abstract

Haplotype-based scans to identify recent and ongoing positive selection have become commonplace in evolutionary genomics studies of numerous species across the tree of life. However, the most widely adopted approaches require phased haplotypes to compute the key statistics. Here we release a major update to the selscan software that re-defines popular haplotype-based statistics for use with unphased "multi-locus genotype" data. We provide unphased implementations of iHS, nSL, XP-EHH, and XP-nSL and evaluate their performance across a range of important parameters in a generic demographic history. Source code and executables are available at https://www.github.com/szpiech/selscan.

1 Introduction

Haplotype-based summary statistics—such as iHS (Voight, et al. 2006), nSL (Ferrer-Admetlla, et al. 2014), XP-EHH (Sabeti, et al. 2007), and XP-nSL (Szpiech, et al. 2021)—have become commonplace in evolutionary genomics studies to identify recent and ongoing positive selection in populations (e.g.,Colonna, et al. 2014; Zoledziewska, et al. 2015; Nedelec, et al. 2016; Crawford, et al. 2017; Meier, et al. 2018; Lu, et al. 2019; Zhang, et al. 2020; Salmon, et al. 2021). When an adaptive allele sweeps through a population, it leaves a characteristic pattern of long high-frequency haplotypes and low genetic diversity in the vicinity of the allele. These statistics aim to capture these signals by summarizing the decay of haplotype homozygosity as a function of distance from a putatively selected region, either within a single population (iHS and nSL) or between two populations (XP-EHH and XP-nSL). However, each of these statistics presumes that haplotype phase is known.

Recent work has shown that converting haplotype data into multi-locus genotype data is an effective approach for using haplotype-based selection statistics such as G12, LASSI, and saltiLASSI (Harris, et al. 2018; Harris and DeGiorgio 2020; DeGiorgio and Szpiech 2021) in unphased data. Recognizing this, we have reformulated the iHS, nSL, XP-EHH, and XP-nSL statistics to use multi-locus genotypes and provided an easy-to-use implementation in selscan 2.0 (Szpiech and Hernandez 2014). We also evaluate the performance of these unphased statistics under various generic demographic models.

2 New Approaches

When the --unphased flag is set in selscan v2.0+, biallelic genotype data is collapsed into multi-locus genotype data by representing the genotype as either 0, 1, or 2—the number of derived alleles observed. In this case, selscan v2.0+ will then compute iHS, nSL, XP-EHH, and XP-nSL as described below. We follow the notation conventions of Szpiech and Hernandez (2014).

2.1 Extended Haplotype Homozygosity

In a sample of n diploid individuals, let C denote the set of all possible genotypes at locus x_0. For multi-locus genotypes, $C \coloneqq \{0,1,2\}$, representing the total counts of a derived allele. Let $C(x_i)$ be the set of all unique haplotypes extending from site x_0 to site x_i either upstream or downstream of x_0. If x_1 is a site immediately adjacent to x_0, then $C(x_1) \coloneqq \{00,01,02,10,11,12,20,21,22\}$, representing all possible two-site multi-locus genotypes. We can then compute the extended haplotype homozygosity (EHH) of a set of multi-locus genotypes as

$$EHH(x_i) \sum_{h \in C(x_i)} \frac{\binom{n_h}{2}}{\binom{n}{2}},$$

where n_h is the number of observed haplotypes of type h.

If we wish to compute the EHH of a subset of observed haplotypes that all contain the same 'core' multi-locus genotype, let $\mathcal{H}_c(x_i)$ be the partition of $\mathcal{C}(x_i)$ containing genotype $c \in \mathcal{C}$ at x_0. For example, choosing a homozygous derived genotype ($c = 2$) as the core, $\mathcal{H}_2 \coloneqq \{20,21,22\}$. Thus, we can compute the EHH of all individuals carrying a given genotype at site x_0 extending out to site x_i as

$$EHH_c(x_i) = \sum_{h \in \mathcal{H}_c(x_i)} \frac{\binom{n_h}{2}}{\binom{n_c}{2}},$$

where n_h is the number of observed haplotypes of type h and n_c is the number of observed multi-locus genotypes with core genotype of c. Finally, we can compute the complement EHH of a sample of multi-locus genotypes as

$$cEHH_c(x_i) = \sum_{h \in \mathcal{C}(x_i) \setminus \mathcal{H}_c(x_i)} \frac{\binom{n_h}{2}}{\binom{n_{c'}}{2}},$$

where $n_{c'}$ is the number of observed multi-locus genotypes with a core genotype of not c.

2.2 iHS and nSL

Unphased iHS and nSL are calculated using the equations above. First, we compute the integrated haplotype homozygosity (iHH) for the homozygous ancestral ($c = 0$) and derived ($c = 2$) core genotypes as

$$iHH_c = \sum_{i=1}^{|\mathcal{D}|} \frac{1}{2}\big(EHH_c(x_{i-1}) + EHH_c(x_i)\big)\mathrm{d}(x_{i-1}, x_i) + \sum_{i=1}^{|\mathcal{U}|} \frac{1}{2}\big(EHH_c(x_{i-1}) + EHH_c(x_i)\big)\mathrm{d}(x_{i-1}, x_i),$$

where $\mathcal{D}$ is the set of downstream sites from the core locus and $\mathcal{U}$ is the set of upstream sites. $d(x_{i-1}, x_i)$ is a measure of genomic distance between to markers and is the genetic distance in centimorgans or physical distance in basepairs for iHS (Voight, et al. 2006) or the number of sites observed for nSL (Ferrer-Admetlla, et al. 2014). We similarly compute the complement integrated haplotype homozygosity (ciHH) for both homozygous core genotypes as

$$ci HH_c = \sum_{i=1}^{|\mathcal{D}|} \frac{1}{2}\left(\mathrm{c}EHH_c(x_{i-1}) + \mathrm{c}EHH_c(x_i)\right)\mathrm{d}(x_{i-1}, x_i)$$

$$+ \sum_{i=1}^{|\mathcal{U}|} \frac{1}{2}\left(\mathrm{c}EHH_c(x_{i-1}) + \mathrm{c}EHH_c(x_i)\right)\mathrm{d}(x_{i-1}, x_i).$$

The (unstandardized) unphased iHS is then calculated as

$$iHS = \begin{cases} \mathrm{iHS}_2, & \text{if } \mathrm{iHS}_2 > \mathrm{iHS}_0 \\ \text{-}\mathrm{iHS}_0, & \text{otherwise'} \end{cases}$$

where $iHS_2 = \log_{10}\left(\frac{iHH_2}{\mathrm{c}iHH_2}\right)$ and $iHS_0 = \log_{10}\left(\frac{iHH_0}{ciHH_0}\right)$. Unstandardized iHS scores are then normalized in frequency bins, as previously described (Voight, et al. 2006; Ferrer-Admetlla, et al. 2014). Unstandardized unphased nSL is computed similarly with the appropriate distance measure. Large positive scores indicate long high-frequency haplotypes with a homozygous derived core genotype, and large negative scores indicate long high-frequency haplotypes with a homozygous ancestral core genotype. Clusters of extreme scores in both directions indicate evidence for a sweep.

2.3 XP-EHH and XP-nSL

Unphased XP-EHH and XP-nSL are calculated by comparing the iHH between populations A and B, using the entire sample in each population. iHH in a population P is computed as

$$iHH_{\mathrm{P}} = \sum_{i=1}^{|\mathcal{D}|} \frac{1}{2}\left(\mathrm{EHH}(x_{i-1}) + \mathrm{EHH}(x_i)\right)\mathrm{d}(x_{i-1}, x_i) + \sum_{i=1}^{|\mathcal{U}|} \frac{1}{2}\left(\mathrm{EHH}(x_{i-1}) + \mathrm{EHH}(x_i)\right)\mathrm{d}(x_{i-1}, x_i),$$

where the distance measure is given as centimorgans or basepairs for XP-EHH (Sabeti, et al. 2007) and number of sites observed for XP-nSL (Szpiech, et al. 2021). The XP statistics between population A and B are then computed as $XP = \log_{10}\left(\frac{iHH_A}{iHH_B}\right)$ and are normalized genome wide in a single bin. Large positive scores indicate long high-frequency haplotypes in population A, and large negative scores indicate long high-frequency haplotypes in population B. Clusters of extreme scores in one direction indicate evidence for a sweep in that population.

3 Methods

3.1 Simulations

We evaluate the performance of the unphased versions of iHS, nSL, XP-EHH, and XP-nSL under a generic two-population divergence model using the coalescent simulation program discoal (Kern and Schrider 2016). We explore five versions of this generic model and name them Demo 1 through Demo 5 (Table 1). Let N_0 and N_1 be the effective population sizes of Population 0 and Population 1 after the split from their ancestral population (of size N_A). For Demo 1, we keep a constant population size post-split and let $N_0 = N_1 = 10{,}000$. For Demo 2, we keep a constant population size post-split and let $N_0 = 2N_1 = 10{,}000$. For Demo 3, we keep a constant population size post-split and let $2N_0 = N_1 = 10{,}000$. For Demo 4, we initially set $N_0 = N_1 = 10{,}000$ and let N_0 grow stepwise exponentially every 50 generations starting at 2,000 generations ago until $N_0 = 5N_1 = 50{,}000$. For Demo 5, we initially set $N_0 = N_1 = 10{,}000$ and let N_1 grow stepwise exponentially every 50 generations starting at 2,000 generations ago until $5N_0 = N_1 = 50{,}000$.

For each demographic history we vary the population divergence time $t_d \in \{2000,\ 4000,\ 8000\}$ generations ago. For non-neutral simulations, we simulate a sweep in Population 0 in the middle of the simulated region across a range of selection coefficients $s \in \{0.005,\ 0.01,\ 0.02\}$. We vary the frequency at which the adaptive allele starts sweeping as $e \in \{0,\ 0.01,\ 0.02,\ 0.05,\ 0.10\}$, where $e = 0$ indicates a hard sweep and $e > 0$ indicates a soft sweep, and we also vary the frequency of the selected allele at time of sampling $f \in \{0.7,\ 0.8,\ 0.9,\ 1.0\}$ as well as $g \in \{50,\ 100\}$ representing fixation of the sweeping allele g generations ago. For all simulations we set the genome length to be $L = 500{,}000$ basepairs, the ancestral effective population size to be $N_A\ =\ 10{,}000$, the per site per generation mutation rate at $\mu = 2.35 \times 10^{-8}$, and the per site per generation recombination rate at $r = 1.2 \times 10^{-8}$. For neutral simulations, we simulate 1,000 replicates for each parameter set, and for non-neutral simulations we simulate 100 replicates for each parameter set. As iHS and nSL are single population statistics, we only analyze Demo 1, Demo 3, and Demo 4 with these statistics, as Demo 2 and Demo 5 have a constant size history identical to Demo 1 for Population 0, where the sweeps are simulated.

For all simulations, we compute the relevant statistics (--ihs, --nsl, --xpehh, or --xpnsl) with selscan v2.0, using the --unphased and --trunc-ok flags. For iHS and XP-EHH, we also use the --pmap flag in order to use physical distance instead of a recombination map.

3.2 Power and False Positive Rate

To compute power for iHS and nSL, we follow the approach of Voight et al. (2006). For these statistics, each non-neutral replicate is individually normalized jointly with all matching neutral replicates in 1% allele frequency bins. Because extreme values of the statistic are likely to be clustered along the genome (Voight, et al. 2006), we then compute the proportion of extreme scores ($|iHS| > 2$ or $|nSL| > 2$) within 100kbp non-overlapping windows. We then bin these windows into 10 quantile bins based on the number of scores observed in each window and call the top 1% of these windows as putatively under selection. We calculate the proportion of non-neutral replicates that fall in this top 1% as the power. To compute the false positive rate, we compute the proportion of neutral simulations that fall within the top 1%.

To compute power for XP-EHH and XP-nSL, we follow the approach of Szpiech et al. (2021). For these statistics, each non-neutral replicate is individually normalized jointly with all matching neutral replicates. Because extreme values of the statistic are likely to be clustered along the genome (Szpiech, et al. 2021), we then compute the proportion of extreme scores (XP-EHH > 2 or XP-nSL > 2) within 100kbp non-overlapping windows. We then bin these windows into 10 quantile bins based on the number of scores observed in each window and call the top 1% of these windows as putatively under selection. We calculate the proportion of non-neutral replicates that fall in this top 1% as the power. To compute the false positive rate, we compute the proportion of neutral simulations that fall within the top 1%.

4 Results

We find that the unphased versions of iHS and nSL have good power (Figures 1, S1-S4, S13-16, and S21-24) to detect selection prior to fixation of the allele, with nSL generally outperforming iHS. In smaller populations (Figure 1C and 1D), power does suffer relative to larger populations (Figure 1A, 1B, 1E, 1F). Each of these statistics also have low false positive rates hovering around 1% (Table S1).

Similarly, we find that the unphased versions of XP-EHH and XP-nSL have good power as well (Figures 2, 3, S5-S12, S17-S20, and S25-S32). When the sweep takes place in the smaller of the two populations (Figure 2C and 2D), we see a similar decrease in power. When one population is undergoing exponential growth (Figure 3) performance is generally quite good, likely the result of a larger effective selection coefficient in large populations. These two-population statistics generally outperform their single-population counterparts, especially for sweeps that have reached fixation recently. Each of these statistics also have low false positive rates hovering around 1% (Table S1).

5 Discussion

We introduce multi-locus genotype versions of four popular haplotype-based selection statistics—iHS (Voight, et al. 2006), nSL (Ferrer-Admetlla, et al. 2014), XP-EHH (Sabeti, et al. 2007), and XP-nSL (Szpiech, et al. 2021)—that can be used when the phase of genotypes is unknown. We implement these updates in the latest v2.0 update of the program selscan (Szpiech and Hernandez 2014), with source code and pre-compiled binaries available at https://www.github.com/szpiech/selscan.

6 Acknowledgements

This work was supported by start-up funds from the Pennsylvania State University's Department of Biology. Computations for this research were performed using the Pennsylvania State University's Institute for Computational Data Sciences' Roar supercomputer.

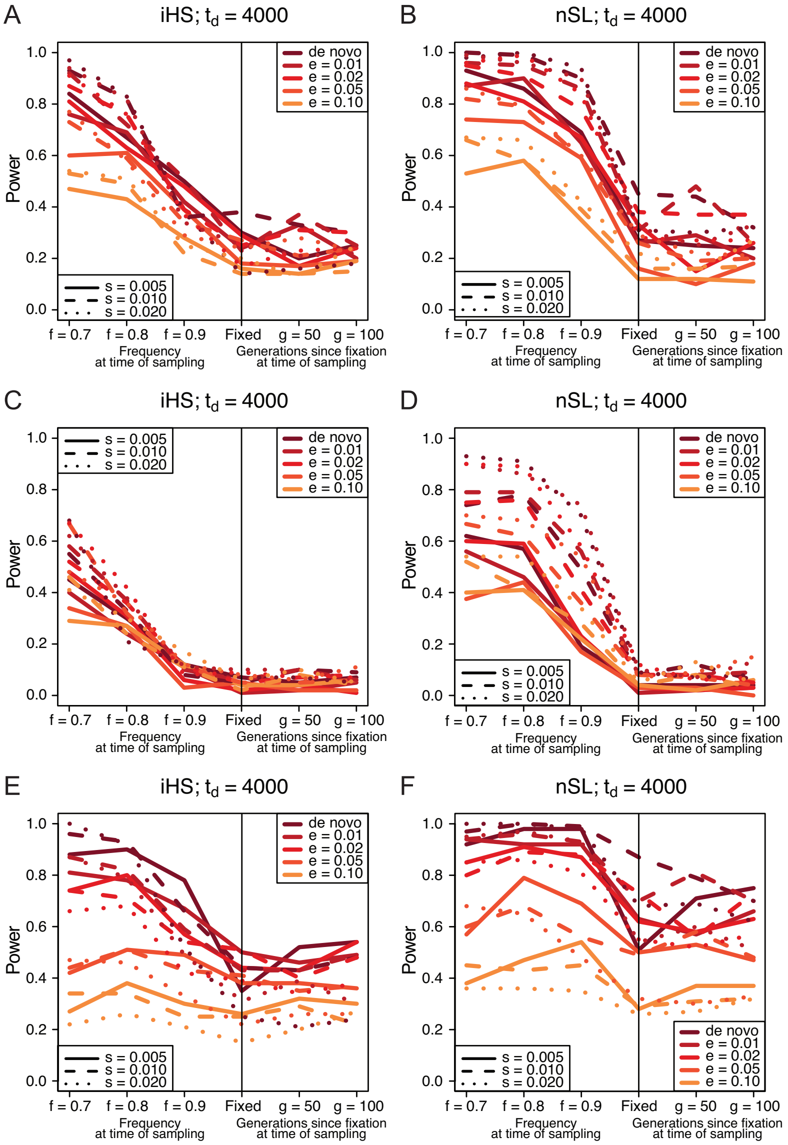


Figure 1. Power curves for unphased implementations of iHS (A, C, and E) and nSL (B, D, and F) under demographic histories Demo 1 (A and B), Demo 3 (C and D), and Demo 4 (E and F). s is the selection coefficient, f is the frequency of the adaptive allele at time of sampling, g is the number of generations at time of sampling since fixation, e is the frequency at which selection began, and t_d is the time in generations since the two populations diverged.

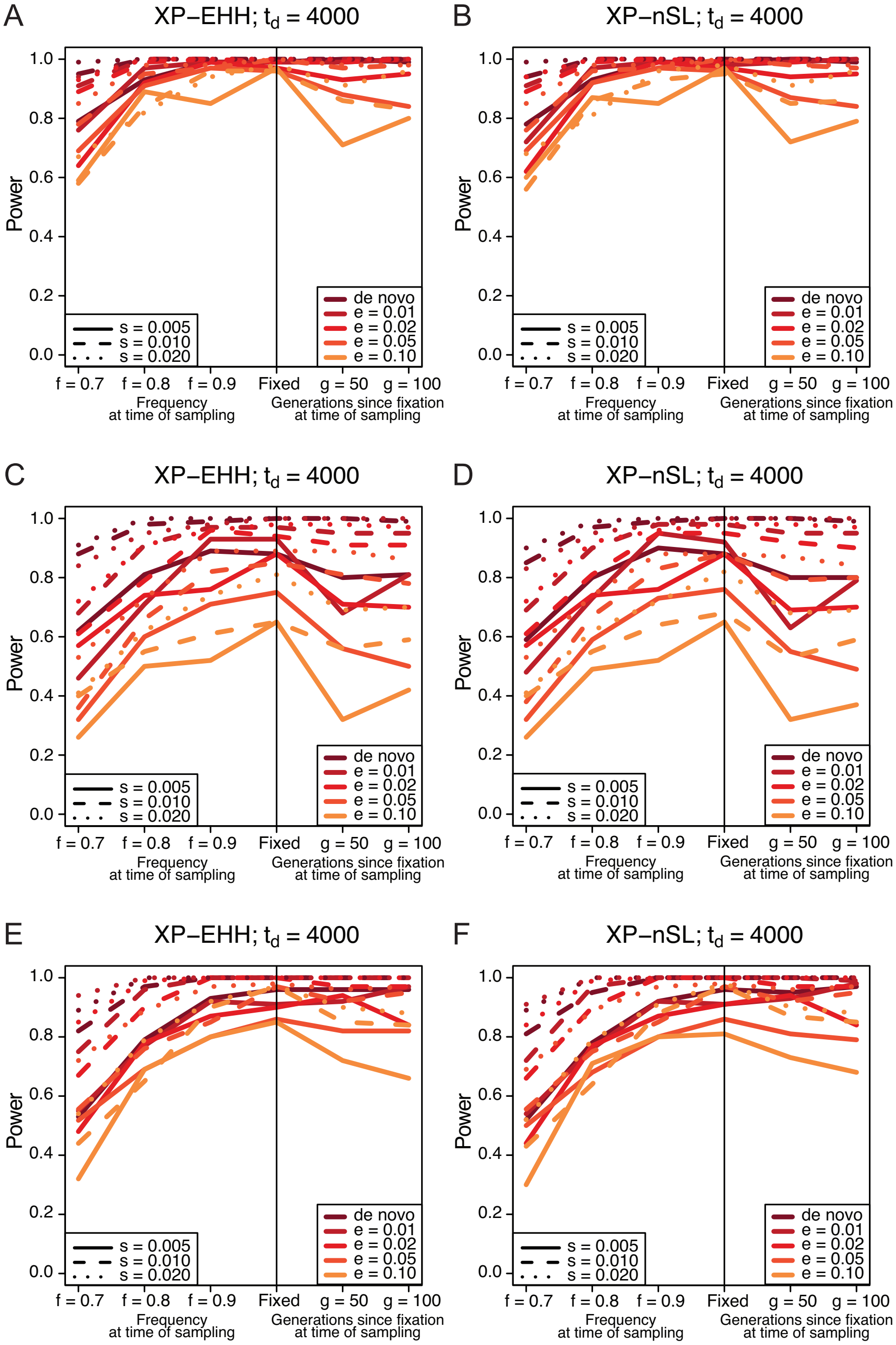


Figure 2. Power curves for unphased implementations of XP-EHH (A, C, and E) and XP-nSL (B, D, and F) under demographic histories Demo 1 (A and B), Demo 2 (C and D), and Demo 3 (E and F). s is the selection coefficient, f is the frequency of the adaptive allele at time of sampling, g is the number of generations at time of sampling since fixation, e is the frequency at which selection began, and t_d is the time in generations since the two populations diverged.

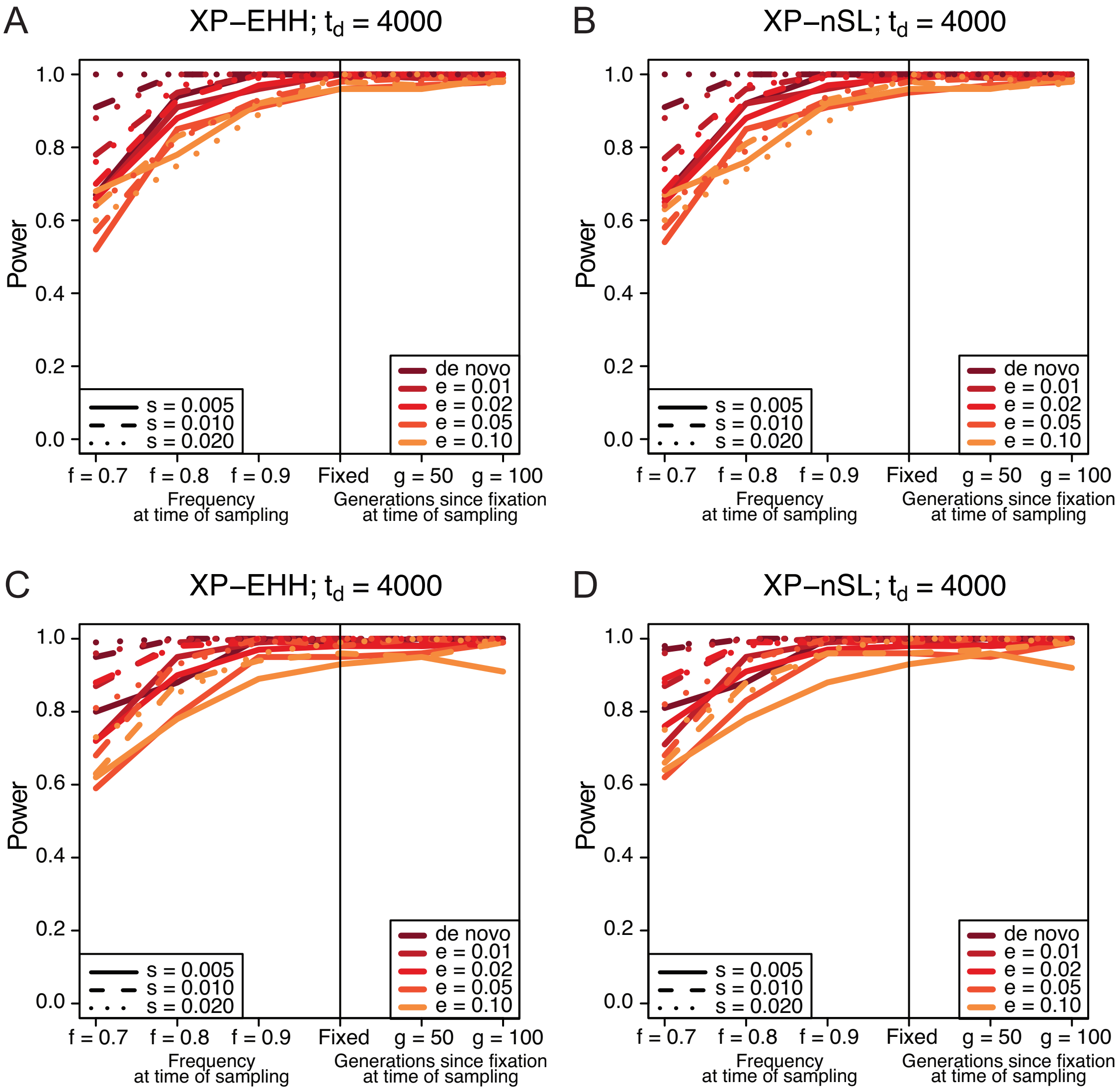


Figure 3. Power curves for unphased implementations of XP-EHH (A and C) and XP-nSL (B and D) under demographic histories Demo 4 (A and B), and Demo 5 (C and D). s is the selection coefficient, f is the frequency of the adaptive allele at time of sampling, g is the number of generations at time of sampling since fixation, e is the frequency at which selection began, and t_d is the time in generations since the two populations diverged.

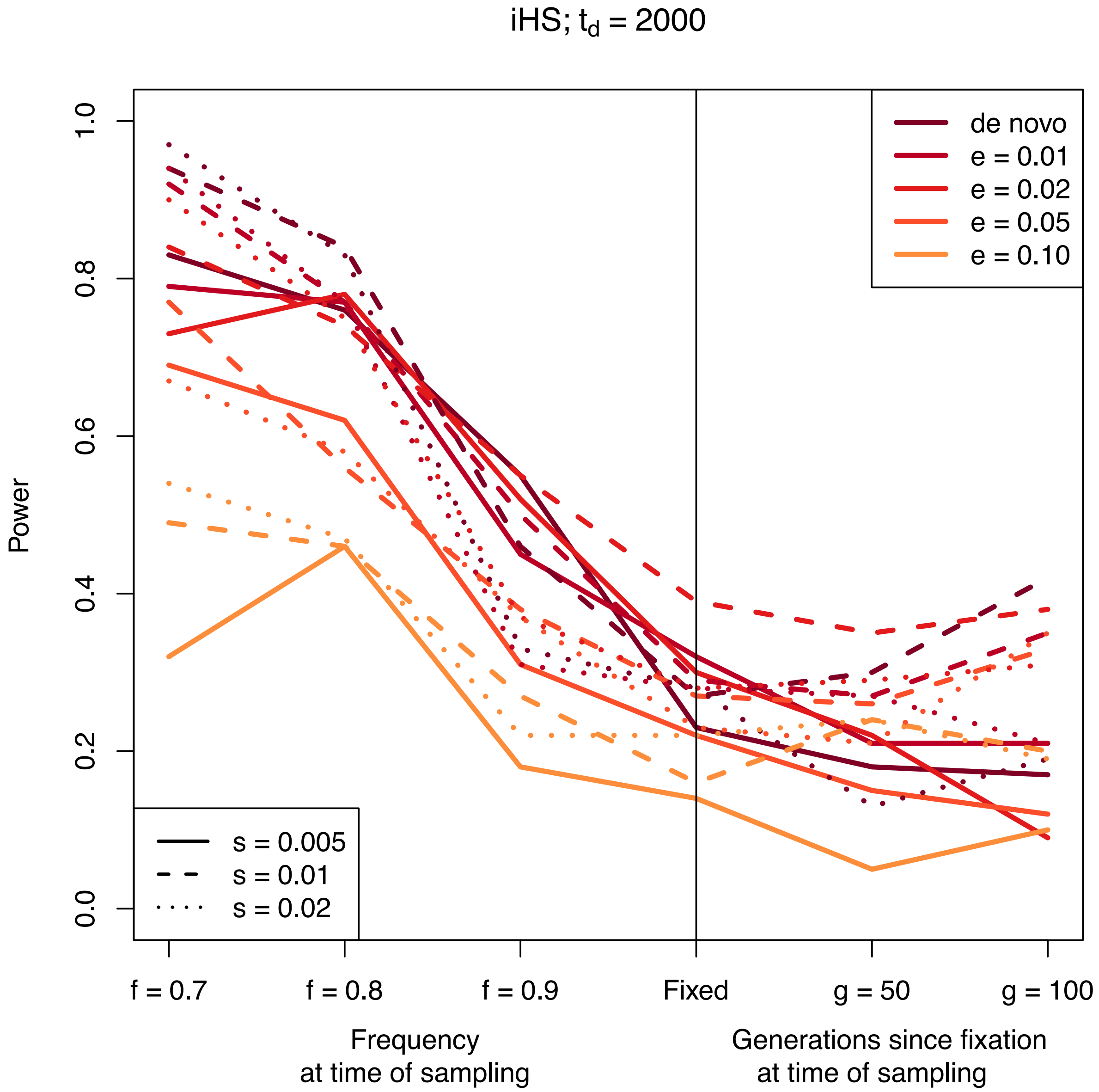


Figure S1. Demo 1 iHS $t_d = 2000$ power curves. s is the selection coefficient, f is the frequency of the adaptive allele at time of sampling, g is the number of generations at time of sampling since fixation, e is the frequency at which selection began, and t_d is the time in generations since the two populations diverged.

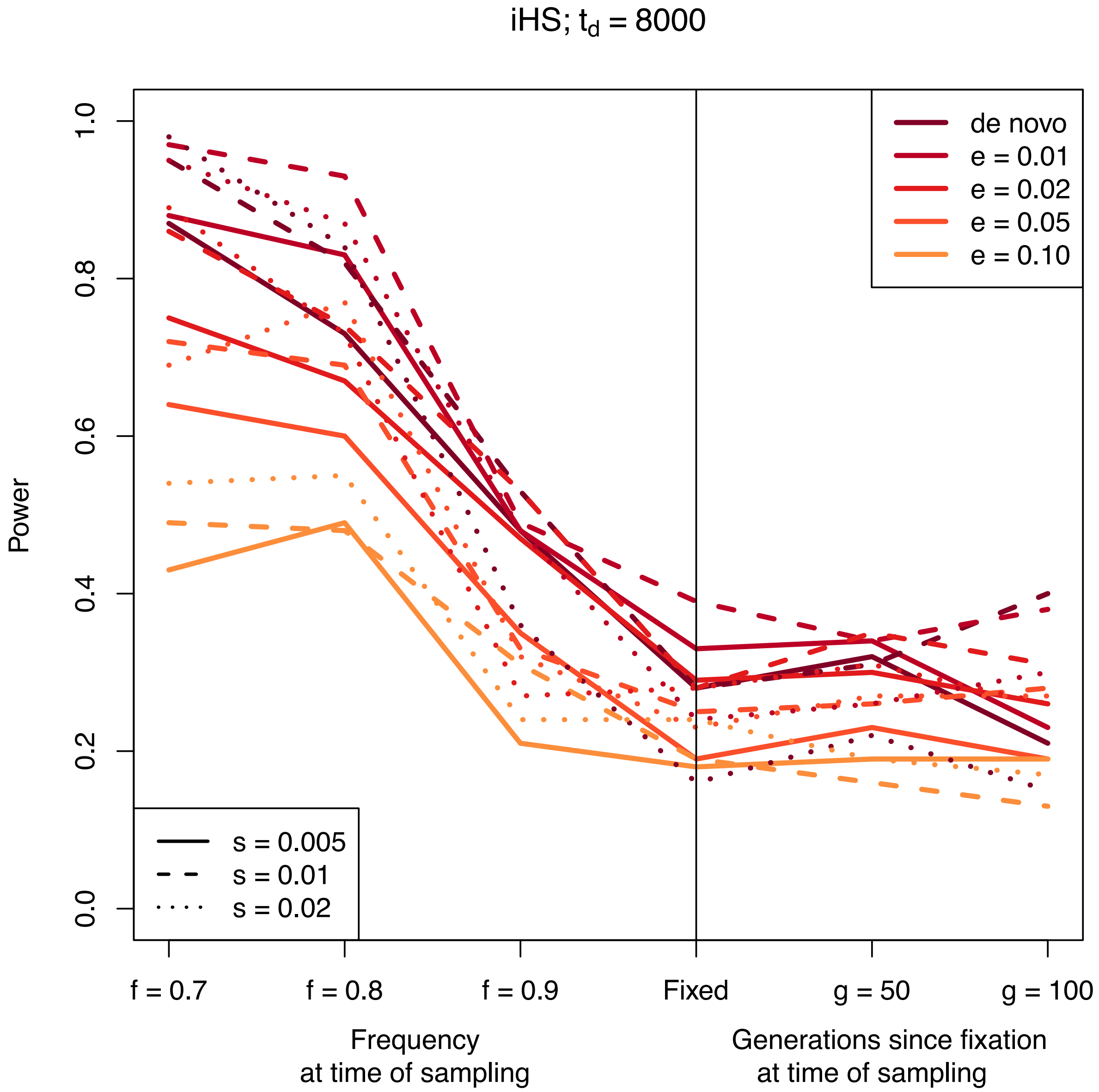
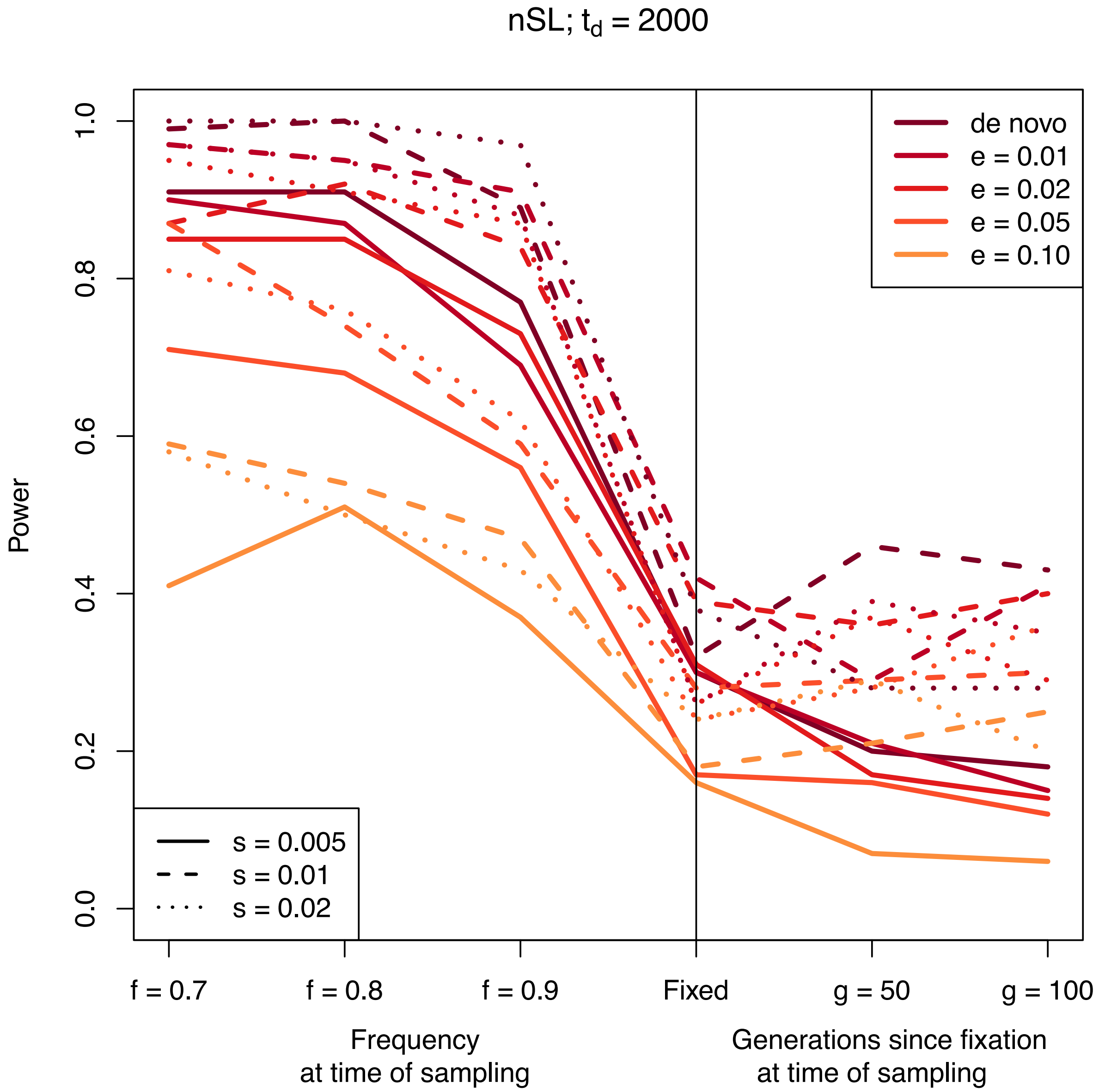


Figure S2. Demo 1 iHS $t_d = 8000$ power curves. s is the selection coefficient, f is the frequency of the adaptive allele at time of sampling, g is the number of generations at time of sampling since fixation, e is the frequency at which selection began, and t_d is the time in generations since the two populations diverged.



206

207 **Figure S3**. Demo 1 nSL $t_d = 2000$ power curves. s is the selection coefficient, f is the
208 frequency of the adaptive allele at time of sampling, g is the number of generations at time of
209 sampling since fixation, e is the frequency at which selection began, and t_d is the time in
210 generations since the two populations diverged.

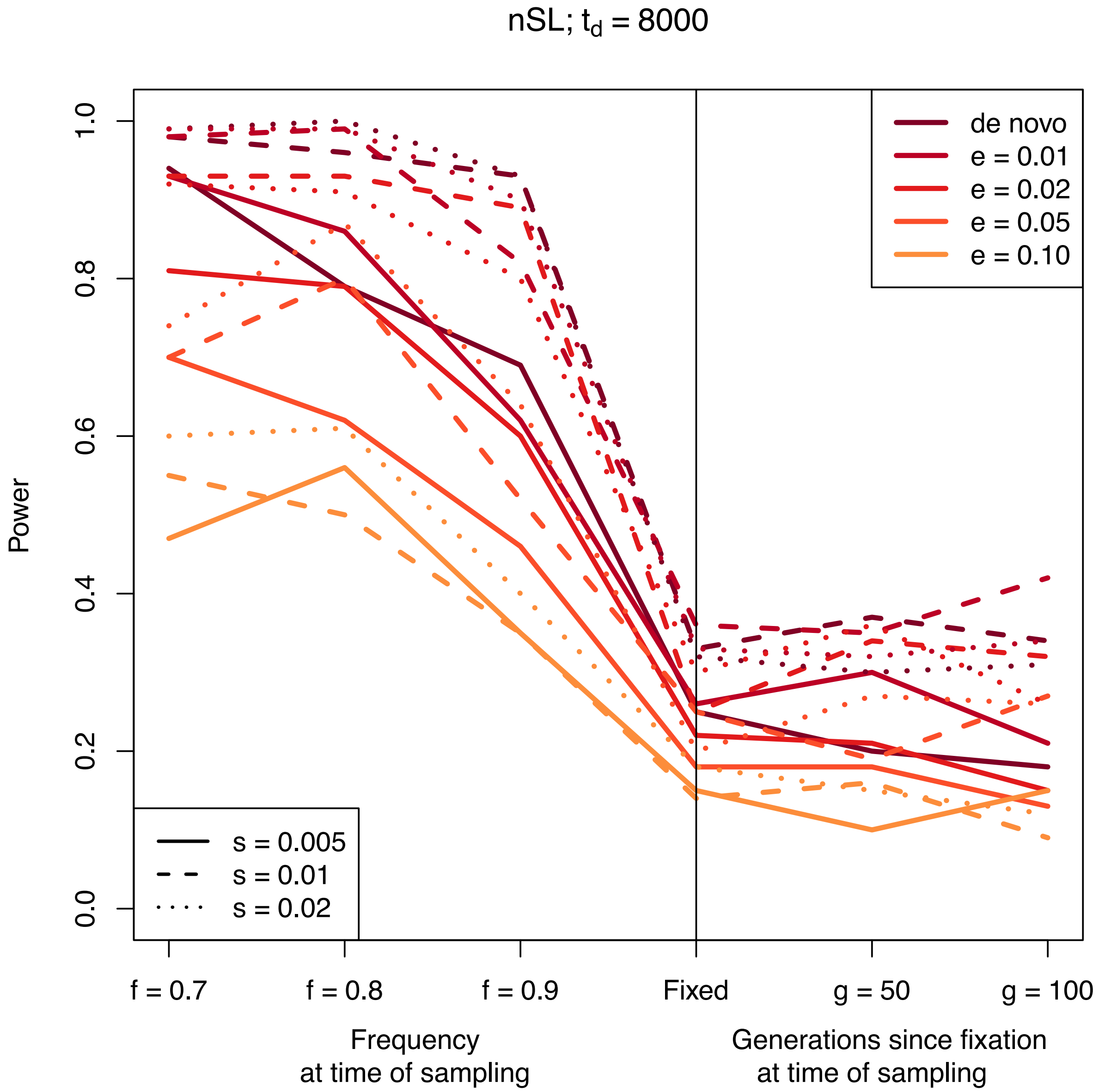
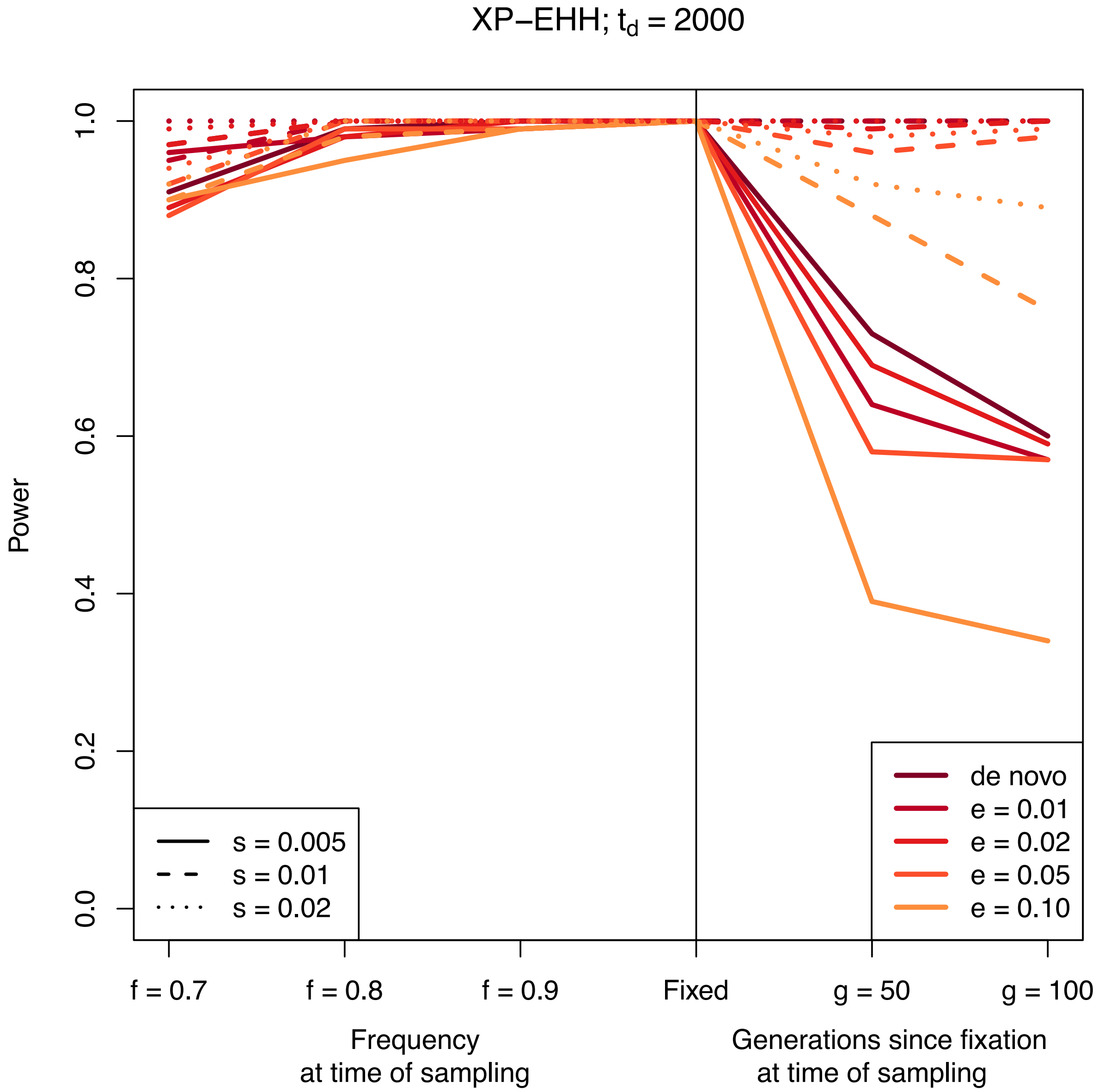


Figure S4. Demo 1 nSL $t_d = 8000$ power curves. s is the selection coefficient, f is the frequency of the adaptive allele at time of sampling, g is the number of generations at time of sampling since fixation, e is the frequency at which selection began, and t_d is the time in generations since the two populations diverged.



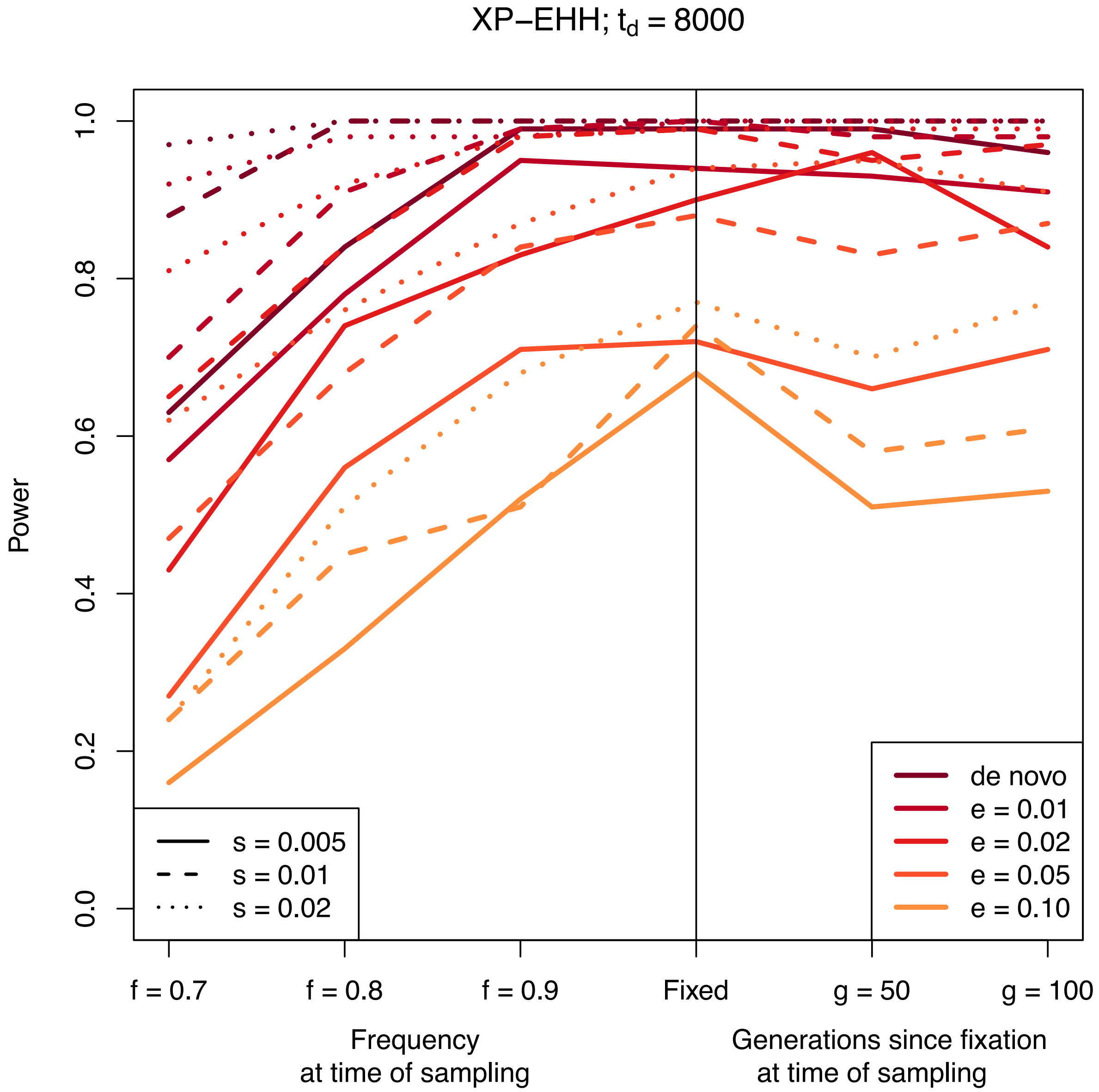
216

217 **Figure S5**. Demo 1 XP-EHH $t_d = 2000$ power curves. s is the selection coefficient, f is the

218 frequency of the adaptive allele at time of sampling, g is the number of generations at time of

219 sampling since fixation, e is the frequency at which selection began, and t_d is the time in

220 generations since the two populations diverged.



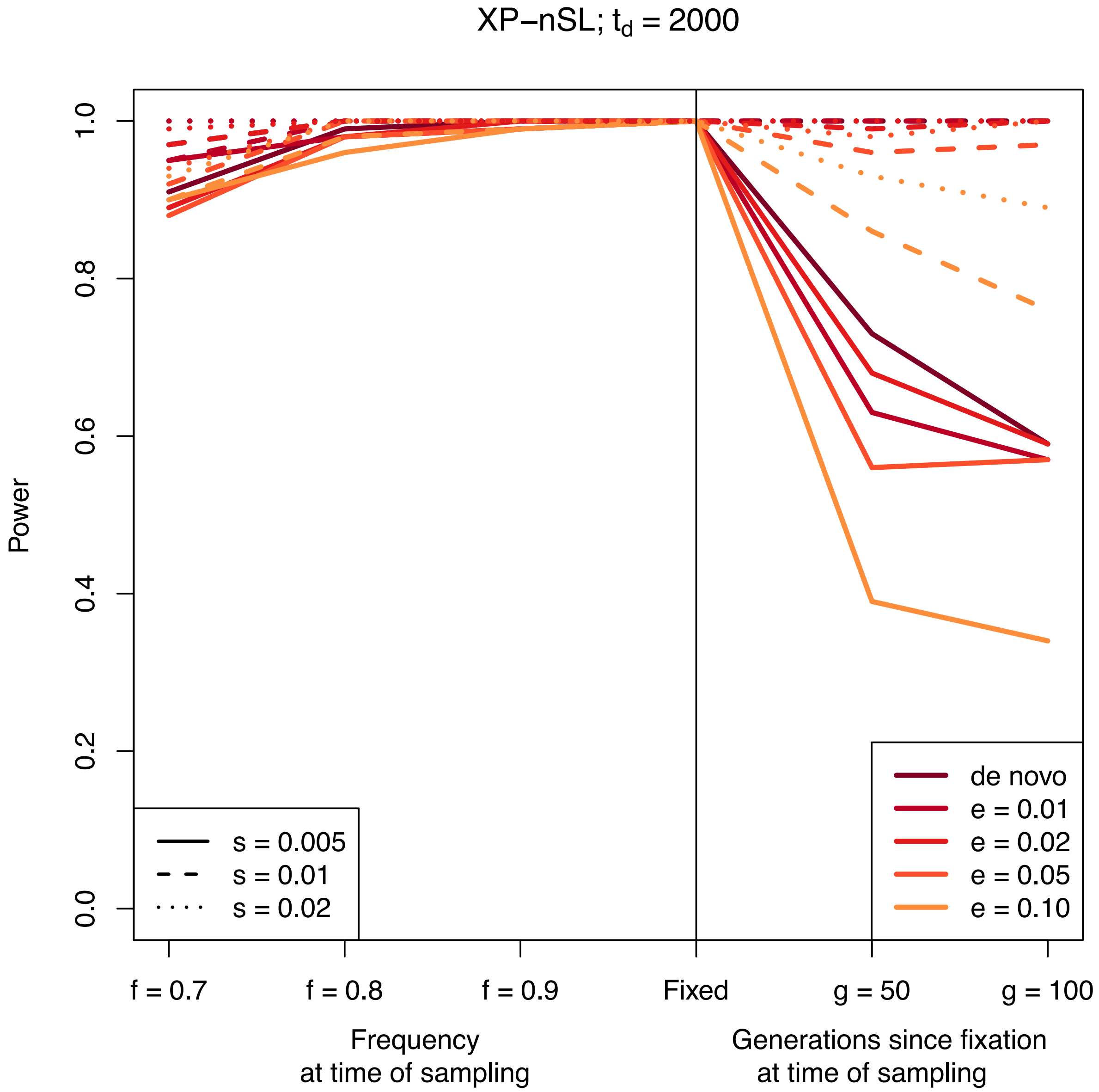
221

222 **Figure S6**. Demo 1 XP-EHH $t_d = 8000$ power curves. s is the selection coefficient, f is the

223 frequency of the adaptive allele at time of sampling, g is the number of generations at time of

224 sampling since fixation, e is the frequency at which selection began, and t_d is the time in

225 generations since the two populations diverged.



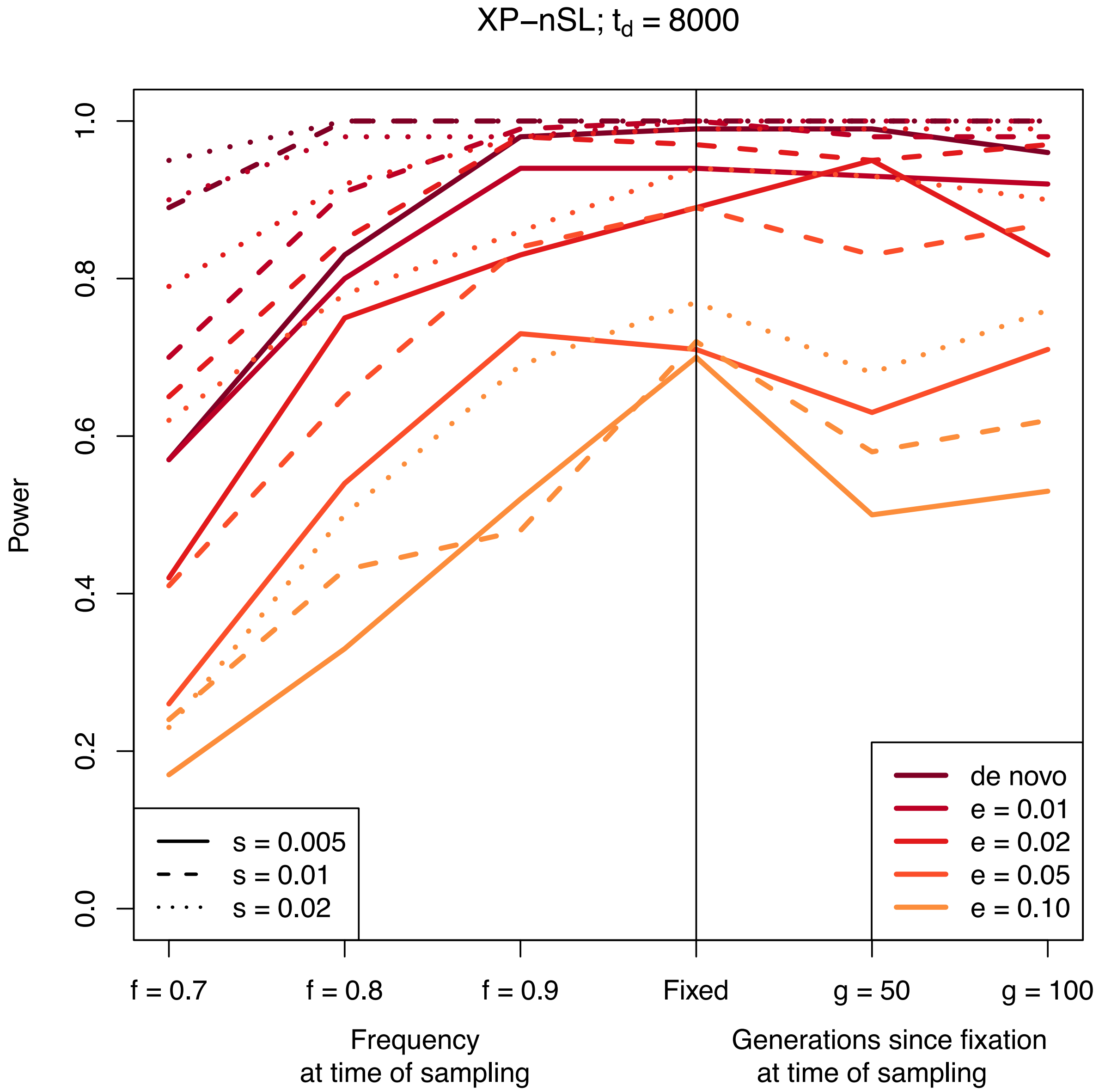
226

227 **Figure S7**. Demo 1 XP-nSL $t_d = 2000$ power curves. s is the selection coefficient, f is the

228 frequency of the adaptive allele at time of sampling, g is the number of generations at time of

229 sampling since fixation, e is the frequency at which selection began, and t_d is the time in

230 generations since the two populations diverged.



231

232 **Figure S8**. Demo 1 XP-nSL $t_d = 8000$ power curves. s is the selection coefficient, f is the
233 frequency of the adaptive allele at time of sampling, g is the number of generations at time of
234 sampling since fixation, e is the frequency at which selection began, and t_d is the time in
235 generations since the two populations diverged.

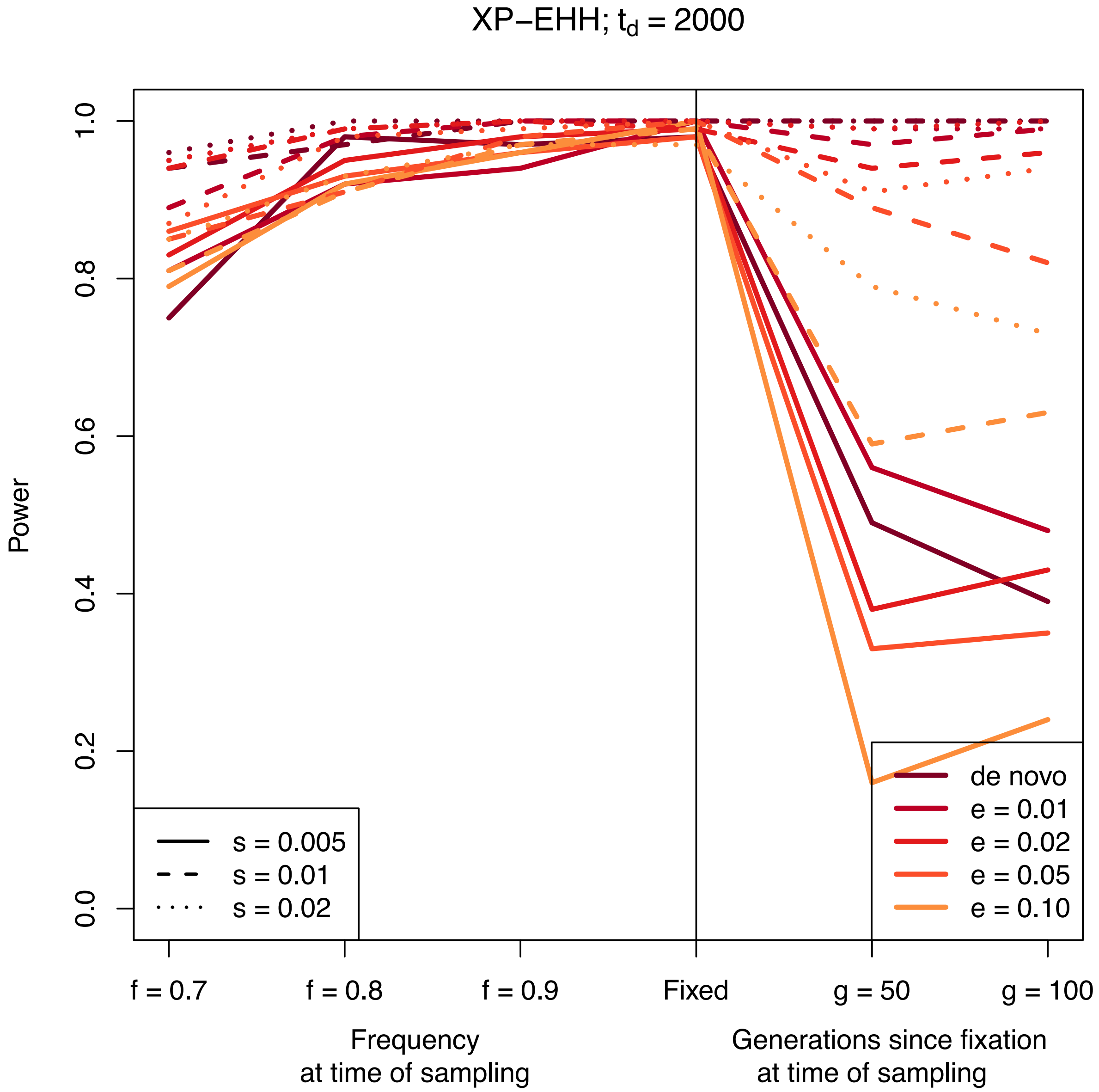
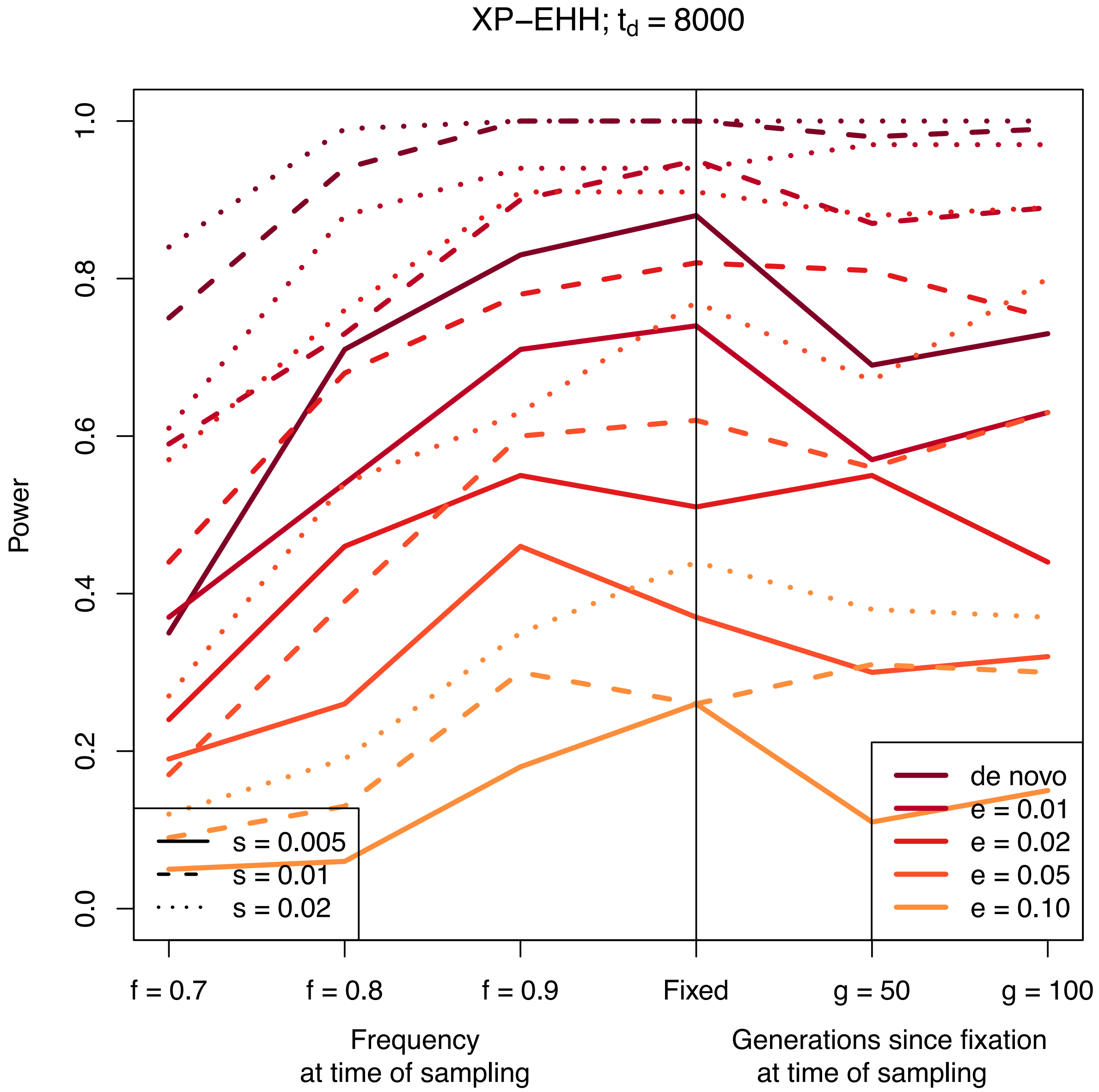
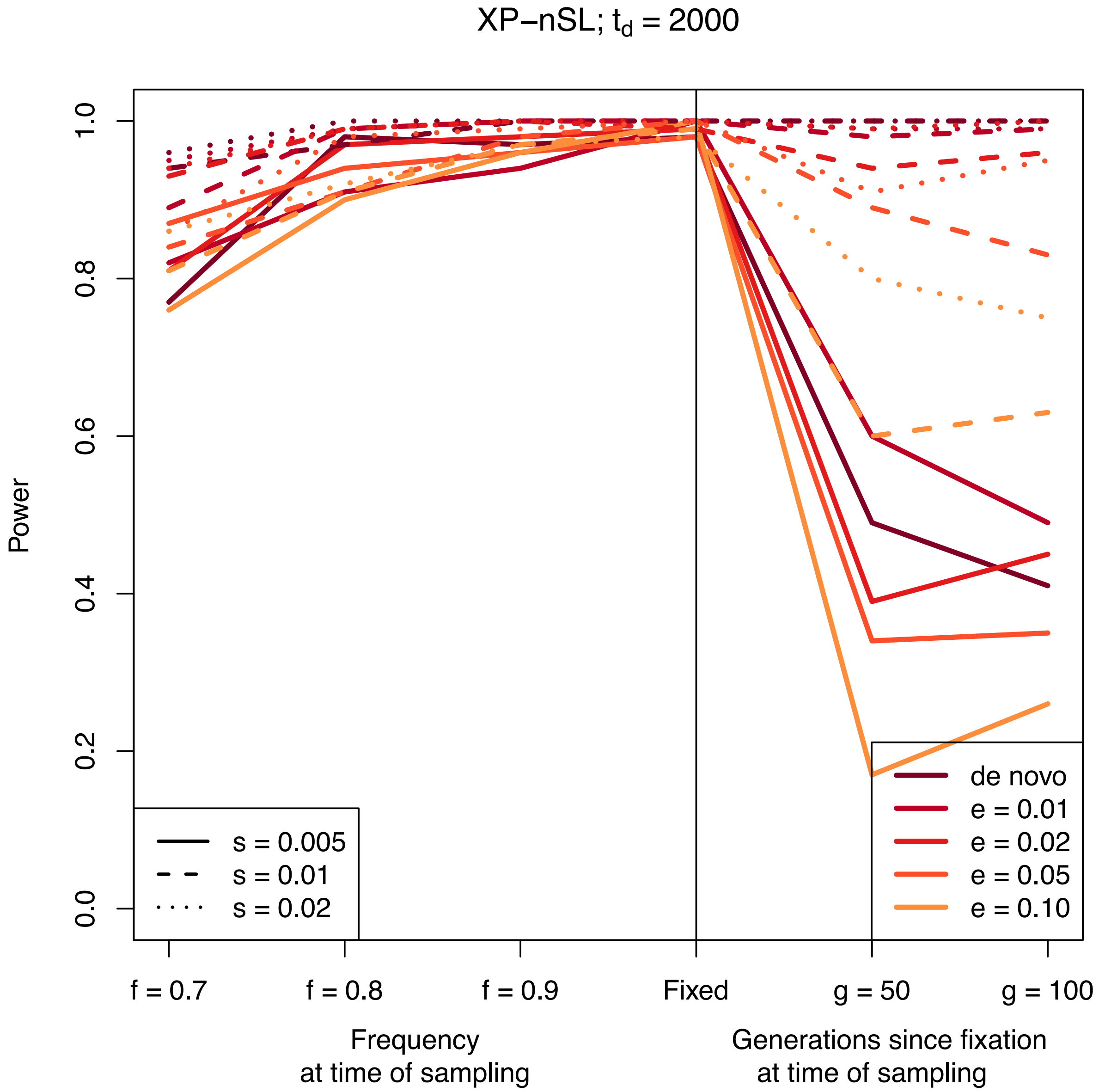


Figure S9. Demo 2 XP-EHH $t_d = 2000$ power curves. s is the selection coefficient, f is the frequency of the adaptive allele at time of sampling, g is the number of generations at time of sampling since fixation, e is the frequency at which selection began, and t_d is the time in generations since the two populations diverged.



241

242 **Figure S10**. Demo 2 XP-EHH $t_d = 8000$ power curves. s is the selection coefficient, f is the
243 frequency of the adaptive allele at time of sampling, g is the number of generations at time of
244 sampling since fixation, e is the frequency at which selection began, and t_d is the time in
245 generations since the two populations diverged.



246

247 **Figure S11**. Demo 2 XP-nSL $t_d = 2000$ power curves. s is the selection coefficient, f is the
248 frequency of the adaptive allele at time of sampling, g is the number of generations at time of
249 sampling since fixation, e is the frequency at which selection began, and t_d is the time in
250 generations since the two populations diverged.

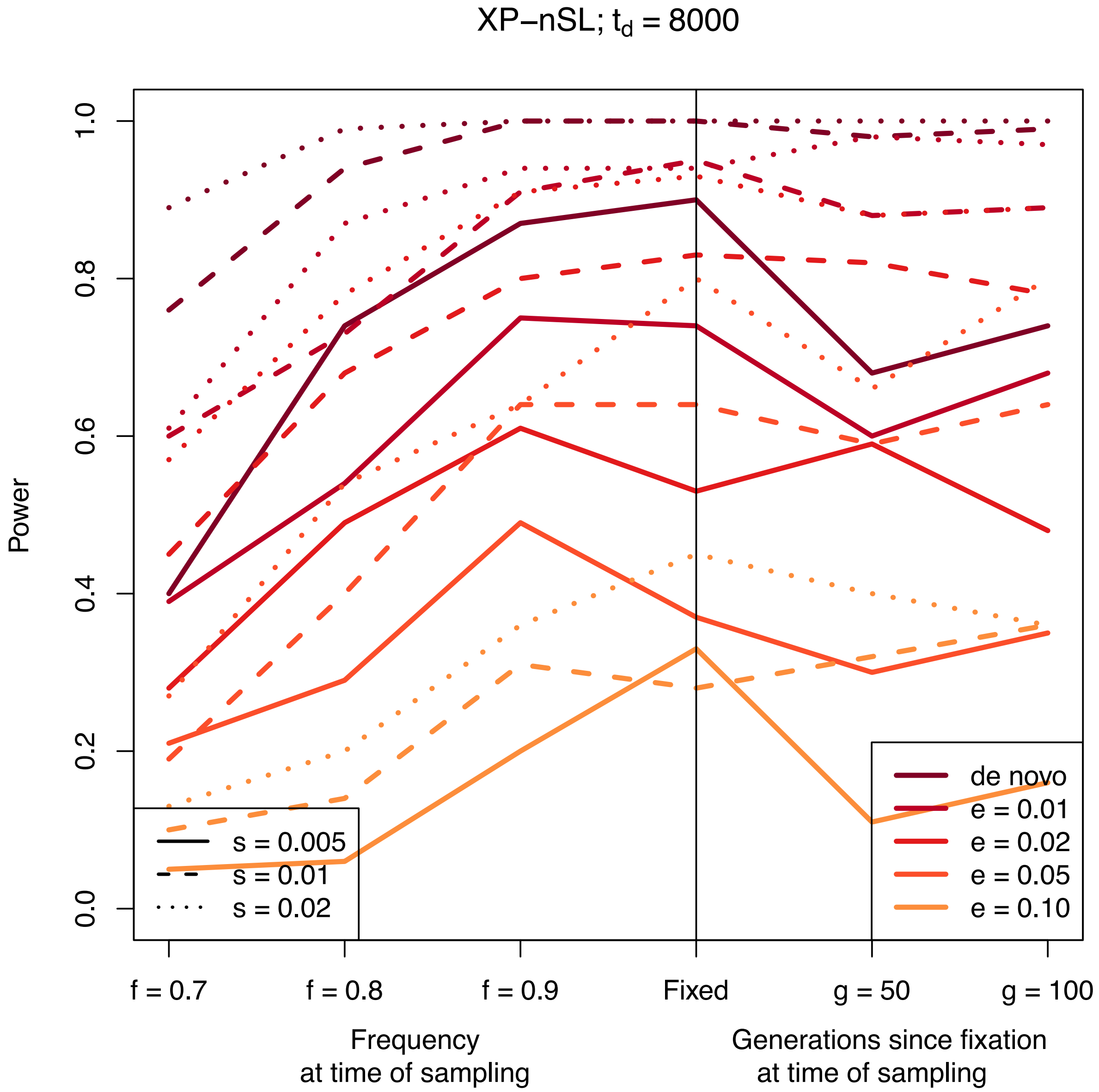
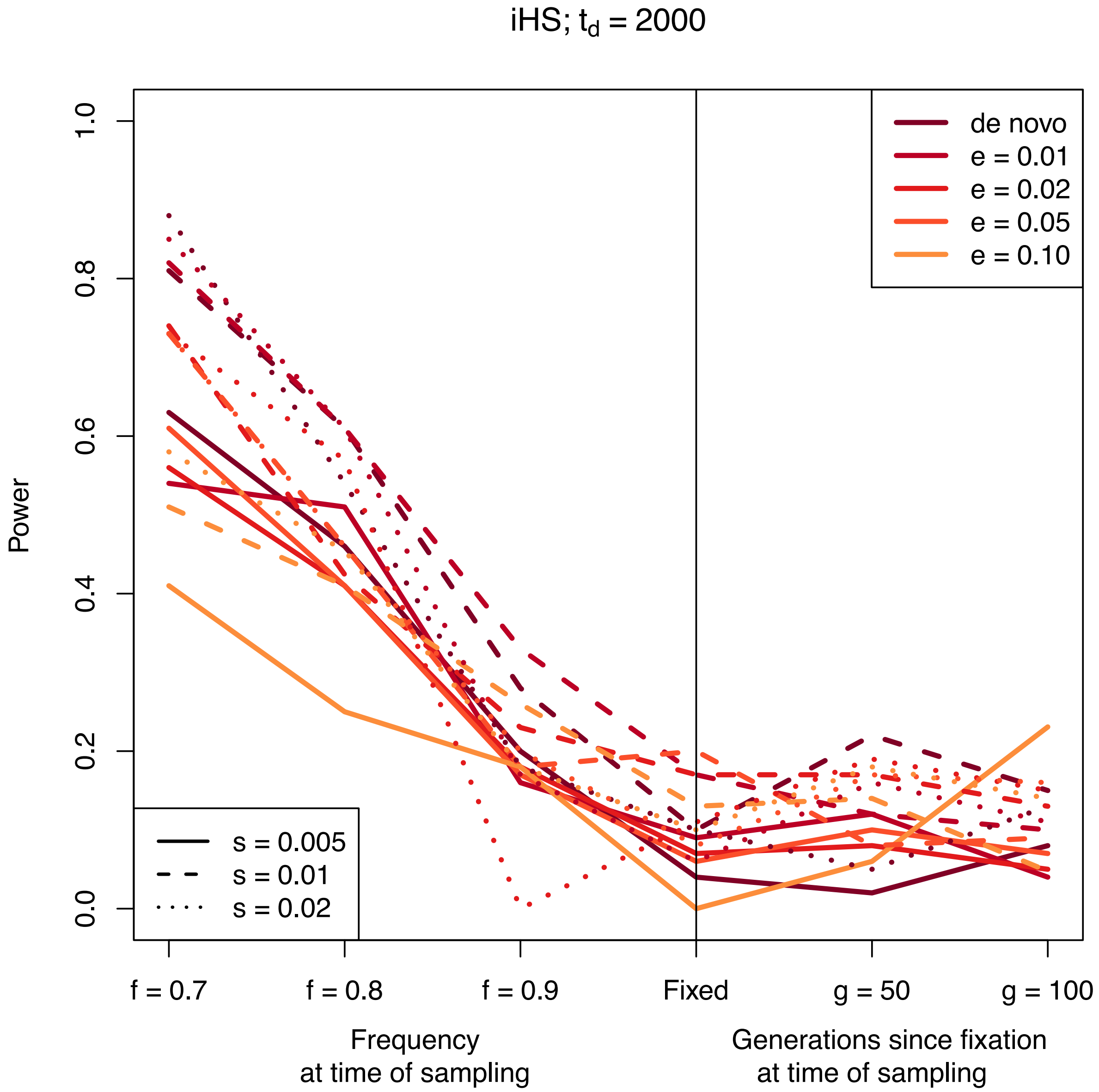


Figure S12. Demo 2 XP-nSL $t_d = 8000$ power curves. s is the selection coefficient, f is the frequency of the adaptive allele at time of sampling, g is the number of generations at time of sampling since fixation, e is the frequency at which selection began, and t_d is the time in generations since the two populations diverged.



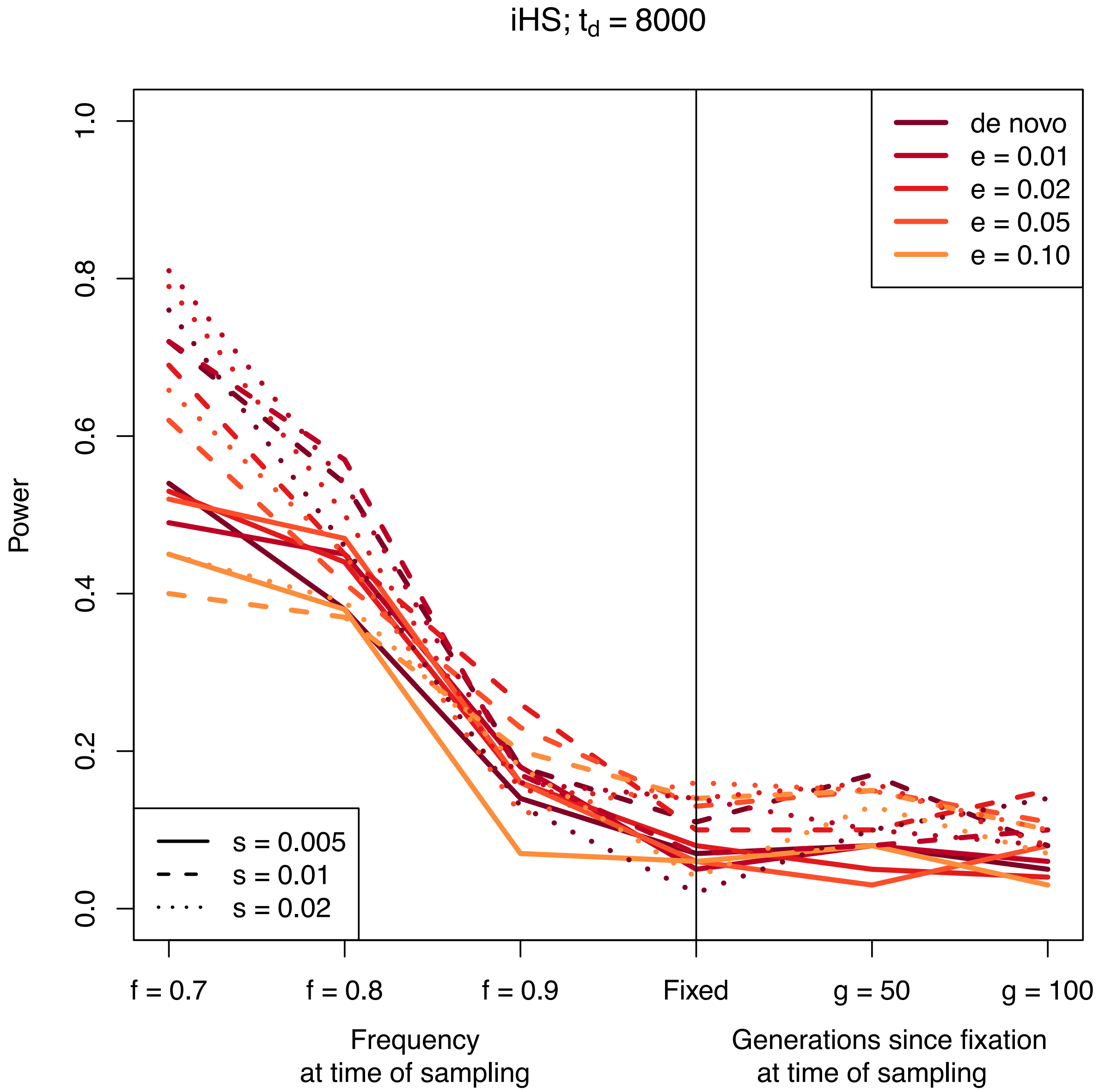
256

257 **Figure S13**. Demo 3 iHS $t_d = 2000$ power curves. s is the selection coefficient, f is the

258 frequency of the adaptive allele at time of sampling, g is the number of generations at time of

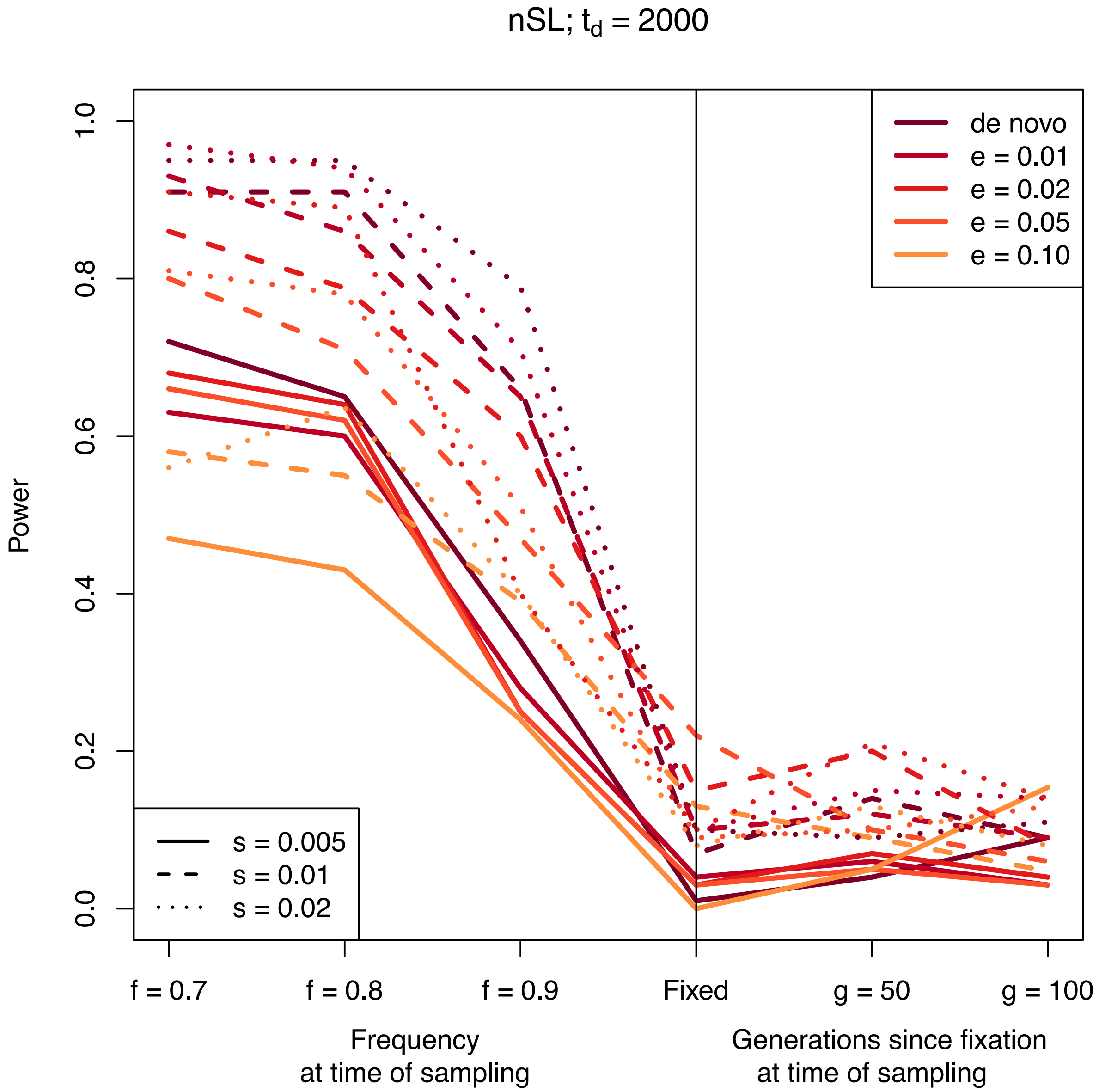
259 sampling since fixation, e is the frequency at which selection began, and t_d is the time in

260 generations since the two populations diverged.



261

262 **Figure S14**. Demo 3 iHS $t_d = 8000$ power curves. s is the selection coefficient, f is the
263 frequency of the adaptive allele at time of sampling, g is the number of generations at time of
264 sampling since fixation, e is the frequency at which selection began, and t_d is the time in
265 generations since the two populations diverged.



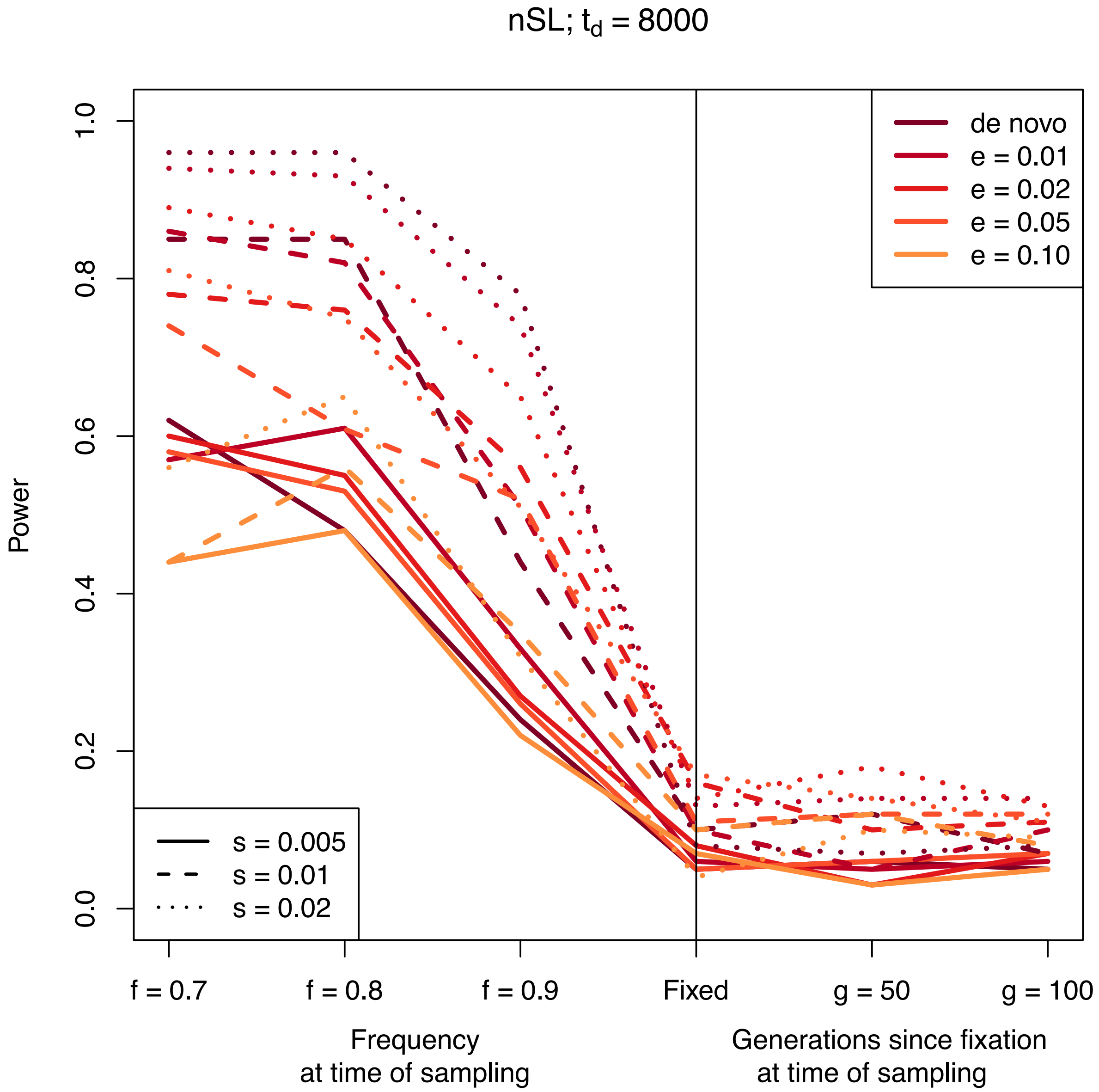
266

267 **Figure S15**. Demo 3 nSL $t_d = 2000$ power curves. s is the selection coefficient, f is the

268 frequency of the adaptive allele at time of sampling, g is the number of generations at time of

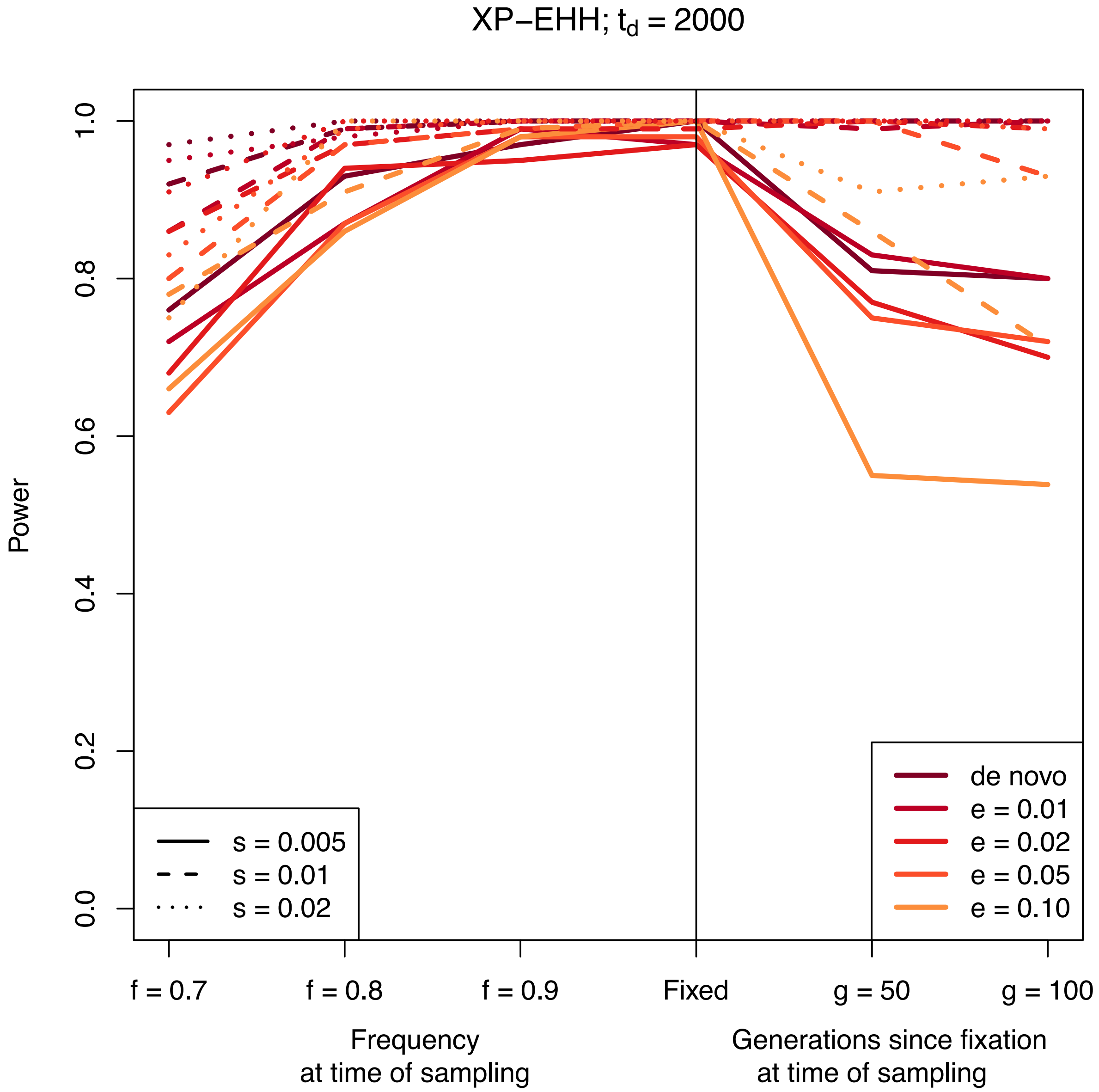
269 sampling since fixation, e is the frequency at which selection began, and t_d is the time in

270 generations since the two populations diverged.



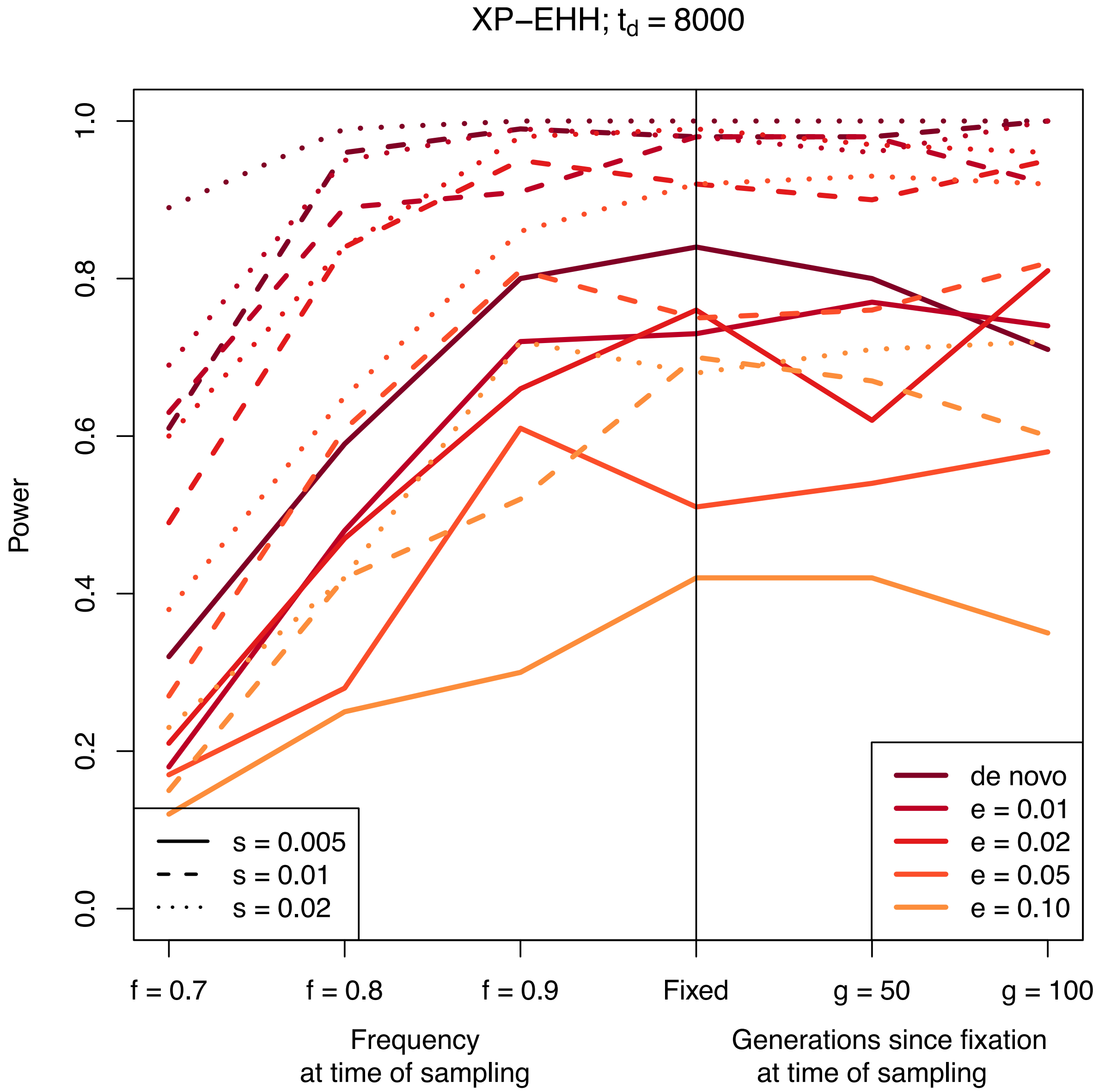
271

272 **Figure S16**. Demo 3 nSL $t_d = 8000$ power curves. s is the selection coefficient, f is the
273 frequency of the adaptive allele at time of sampling, g is the number of generations at time of
274 sampling since fixation, e is the frequency at which selection began, and t_d is the time in
275 generations since the two populations diverged.



276

277 **Figure S17**. Demo 3 XP-EHH $t_d = 2000$ power curves. s is the selection coefficient, f is the
278 frequency of the adaptive allele at time of sampling, g is the number of generations at time of
279 sampling since fixation, e is the frequency at which selection began, and t_d is the time in
280 generations since the two populations diverged.



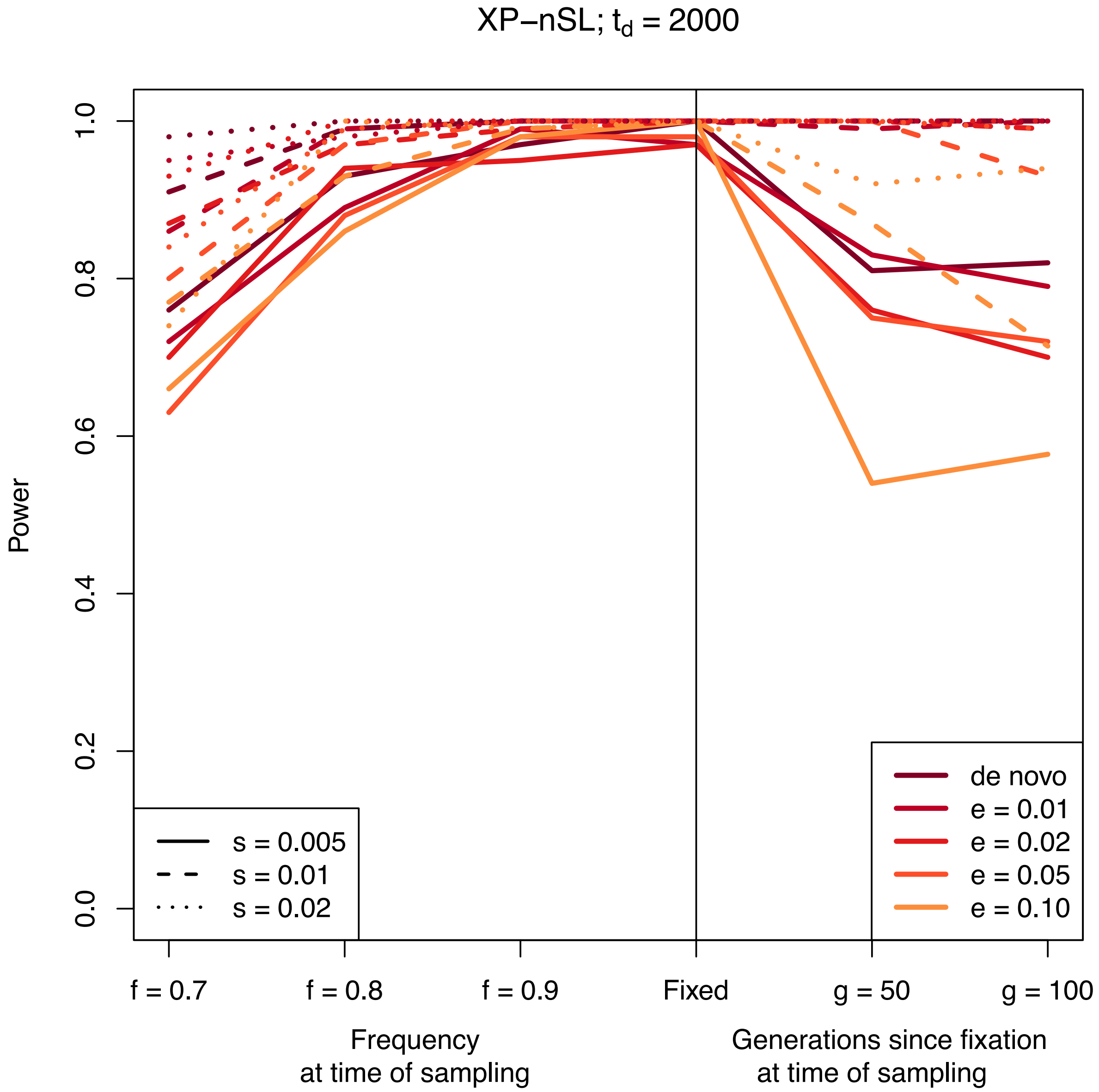
281

282 **Figure S18**. Demo 3 XP-EHH $t_d = 8000$ power curves. s is the selection coefficient, f is the

283 frequency of the adaptive allele at time of sampling, g is the number of generations at time of

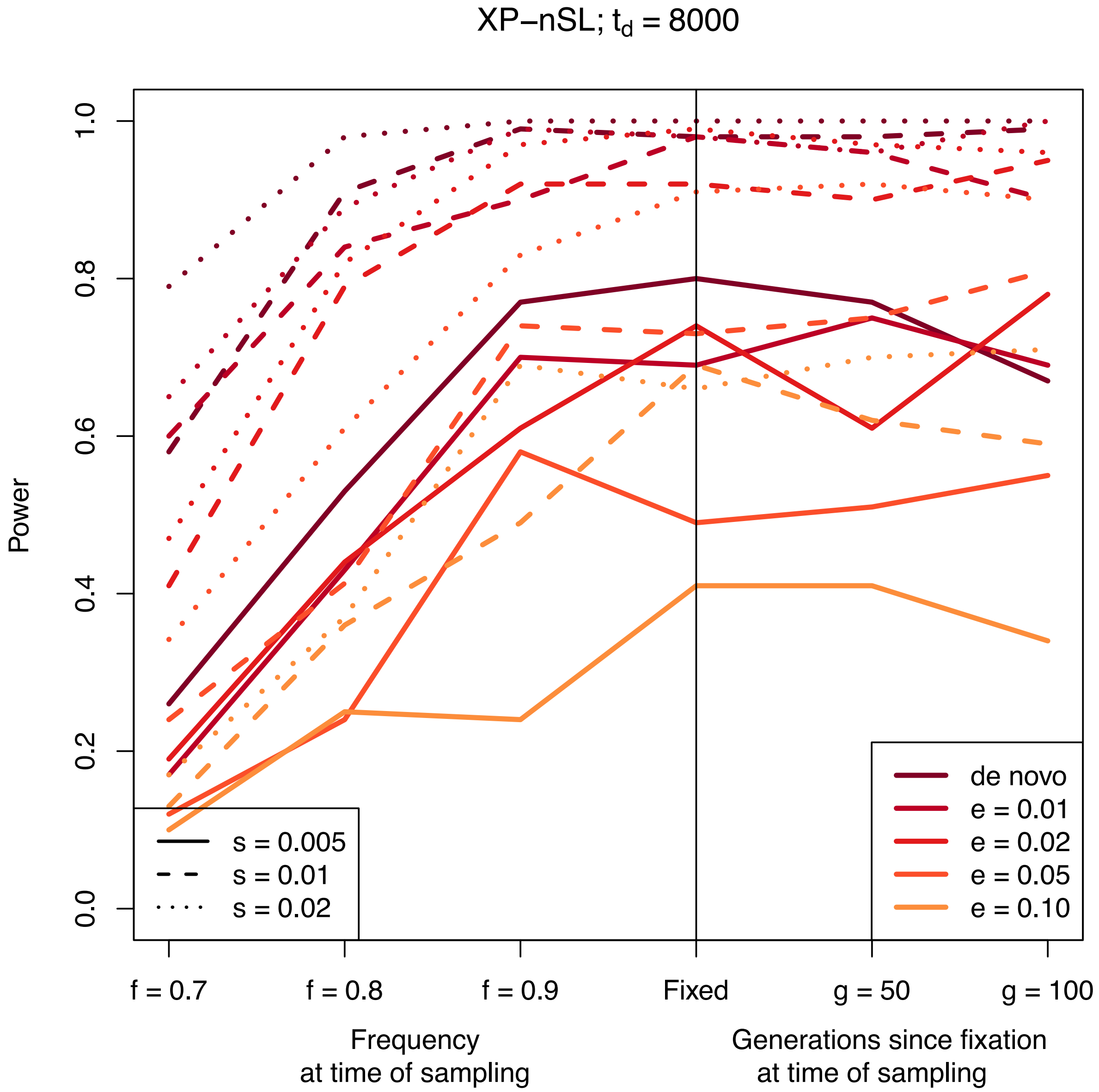
284 sampling since fixation, e is the frequency at which selection began, and t_d is the time in

285 generations since the two populations diverged.



286

287 **Figure S19**. Demo 3 XP-nSL $t_d = 2000$ power curves. s is the selection coefficient, f is the
288 frequency of the adaptive allele at time of sampling, g is the number of generations at time of
289 sampling since fixation, e is the frequency at which selection began, and t_d is the time in
290 generations since the two populations diverged.



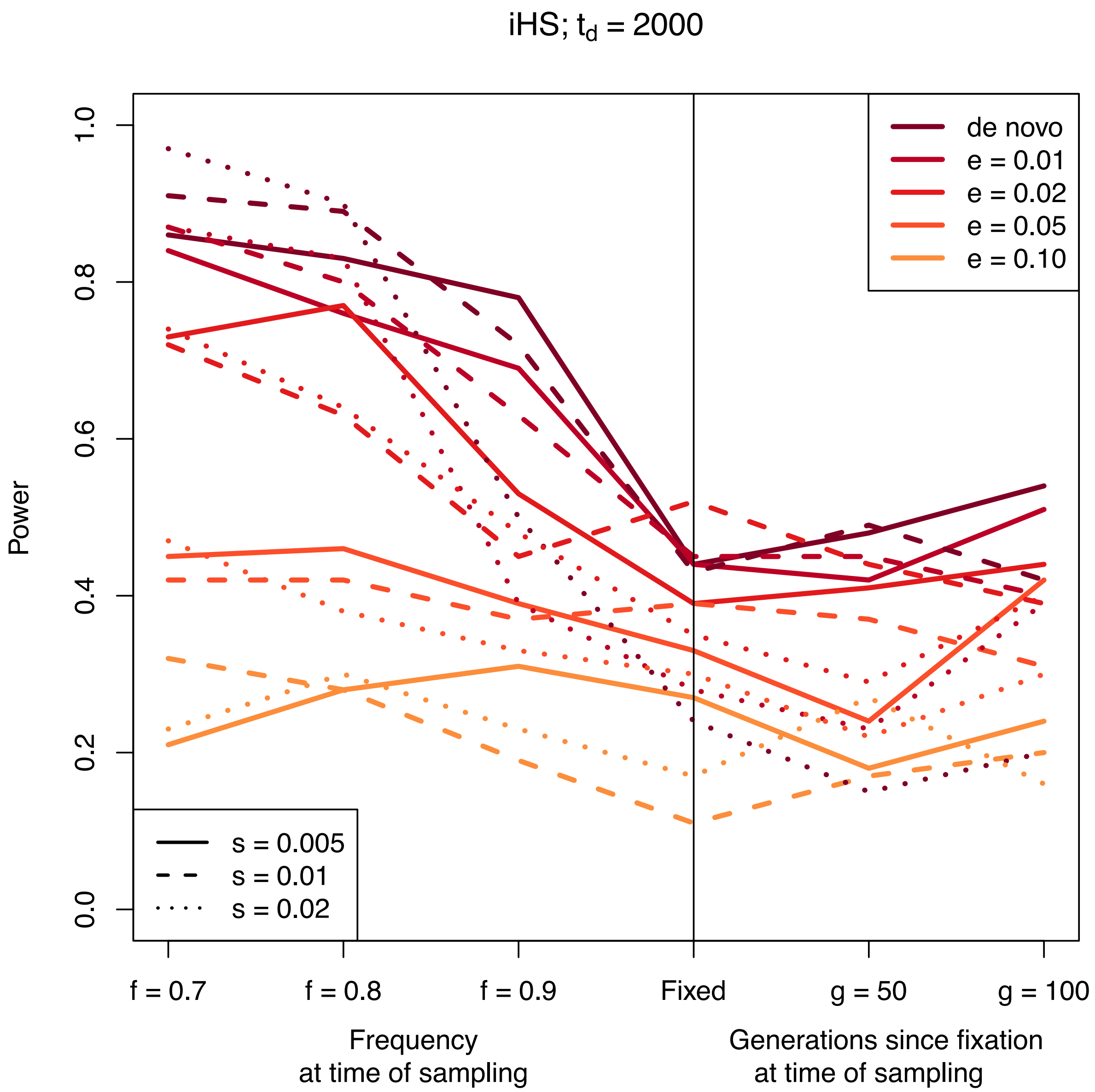
291

292 **Figure S20**. Demo 3 XP-nSL $t_d = 8000$ power curves. s is the selection coefficient, f is the

293 frequency of the adaptive allele at time of sampling, g is the number of generations at time of

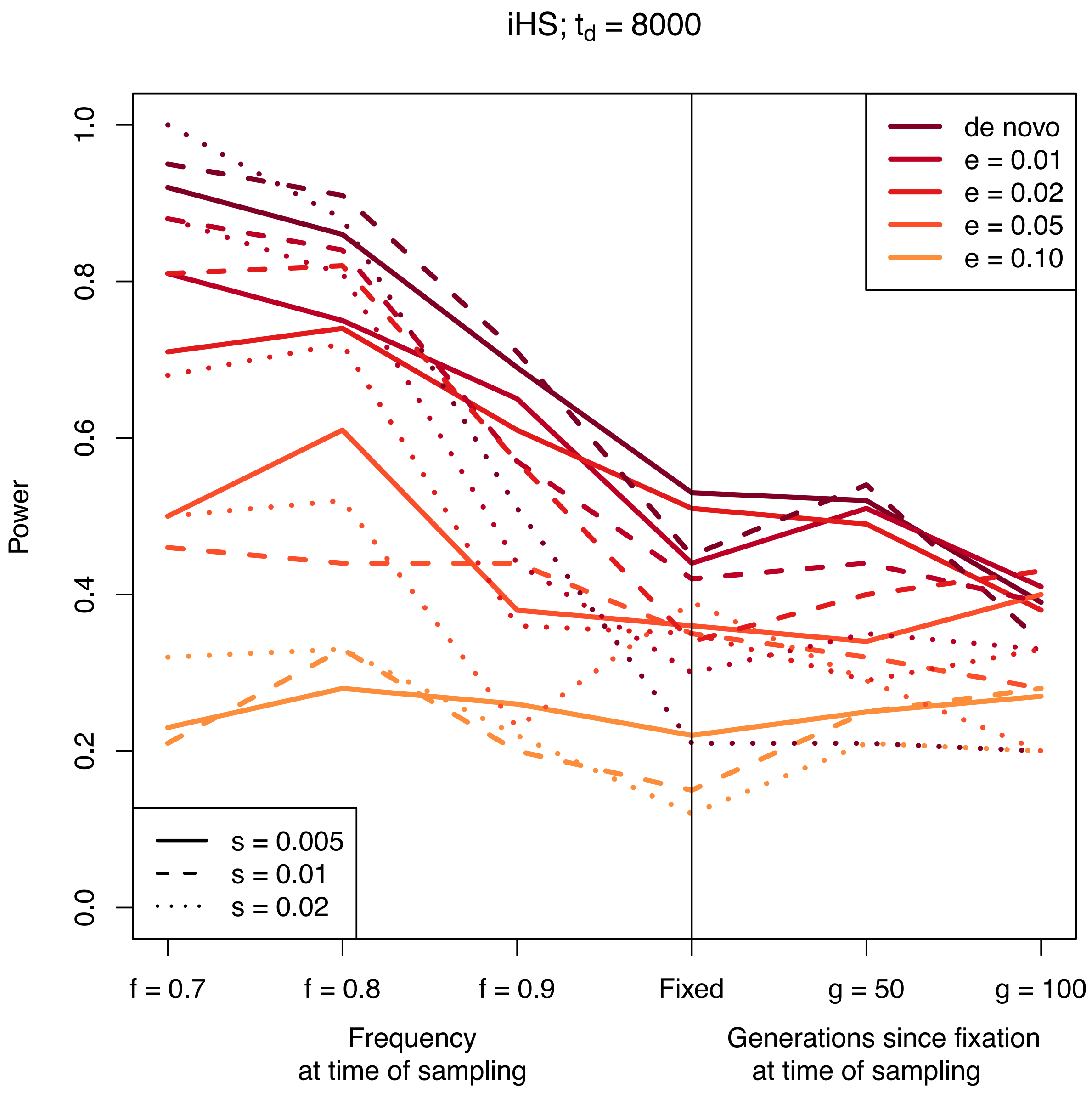
294 sampling since fixation, e is the frequency at which selection began, and t_d is the time in

295 generations since the two populations diverged.



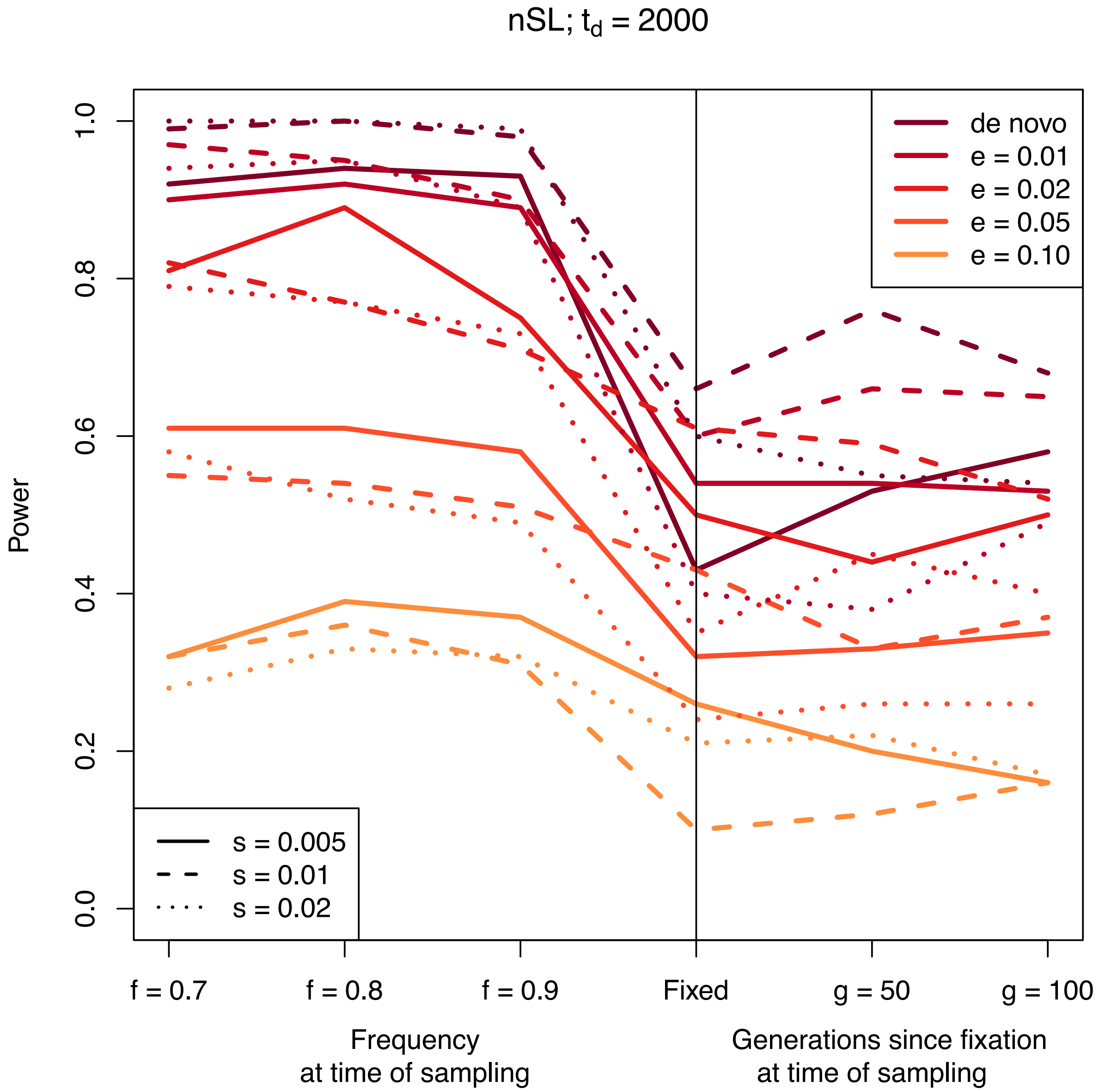
296

297 **Figure S21**. Demo 4 iHS $t_d = 2000$ power curves. s is the selection coefficient, f is the 298 frequency of the adaptive allele at time of sampling, g is the number of generations at time of 299 sampling since fixation, e is the frequency at which selection began, and t_d is the time in 300 generations since the two populations diverged.



301

302 **Figure S22**. Demo 4 iHS $t_d = 8000$ power curves. s is the selection coefficient, f is the
303 frequency of the adaptive allele at time of sampling, g is the number of generations at time of
304 sampling since fixation, e is the frequency at which selection began, and t_d is the time in
305 generations since the two populations diverged.



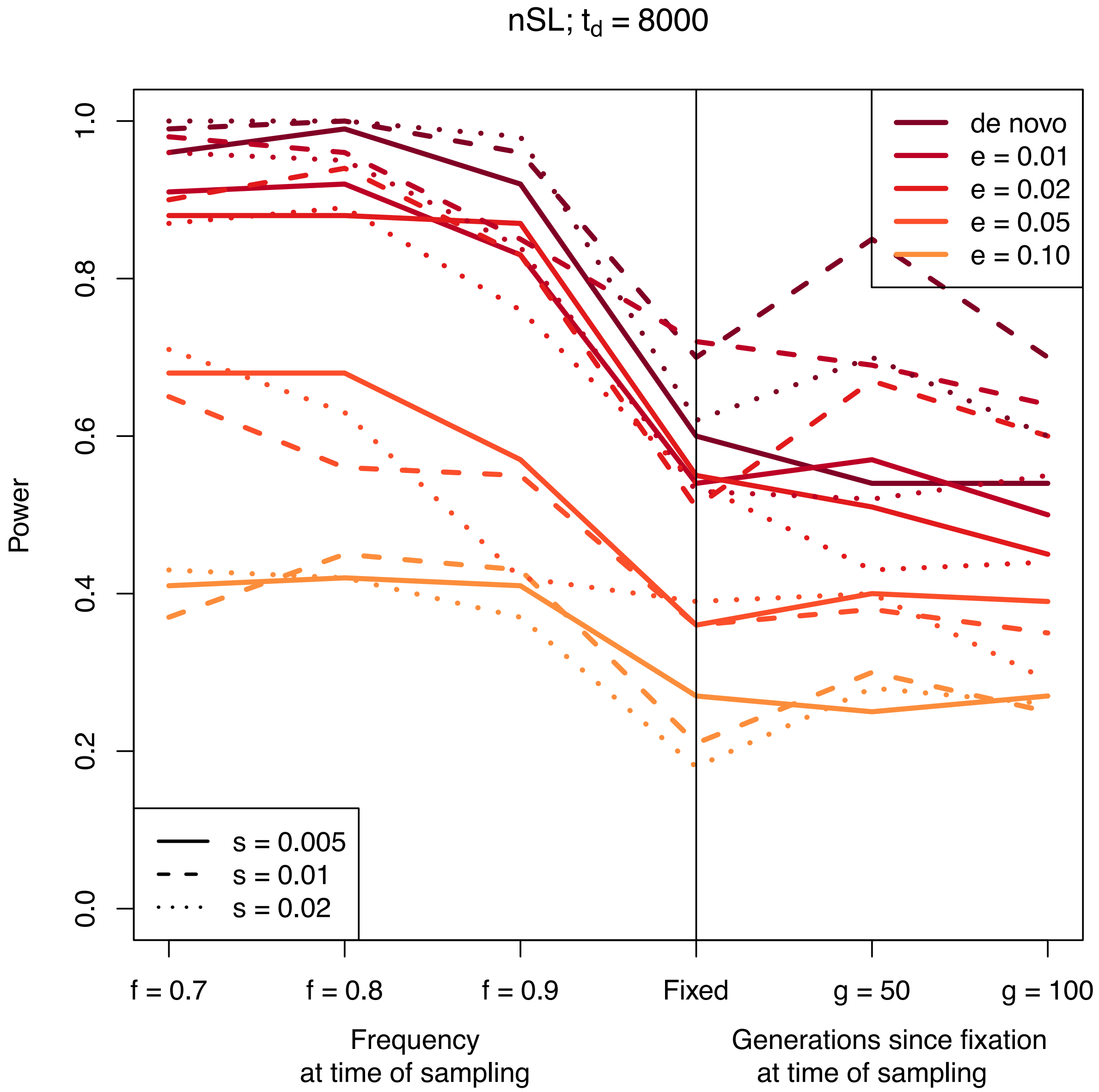
306

307 **Figure S23**. Demo 4 nSL $t_d = 2000$ power curves. s is the selection coefficient, f is the

308 frequency of the adaptive allele at time of sampling, g is the number of generations at time of

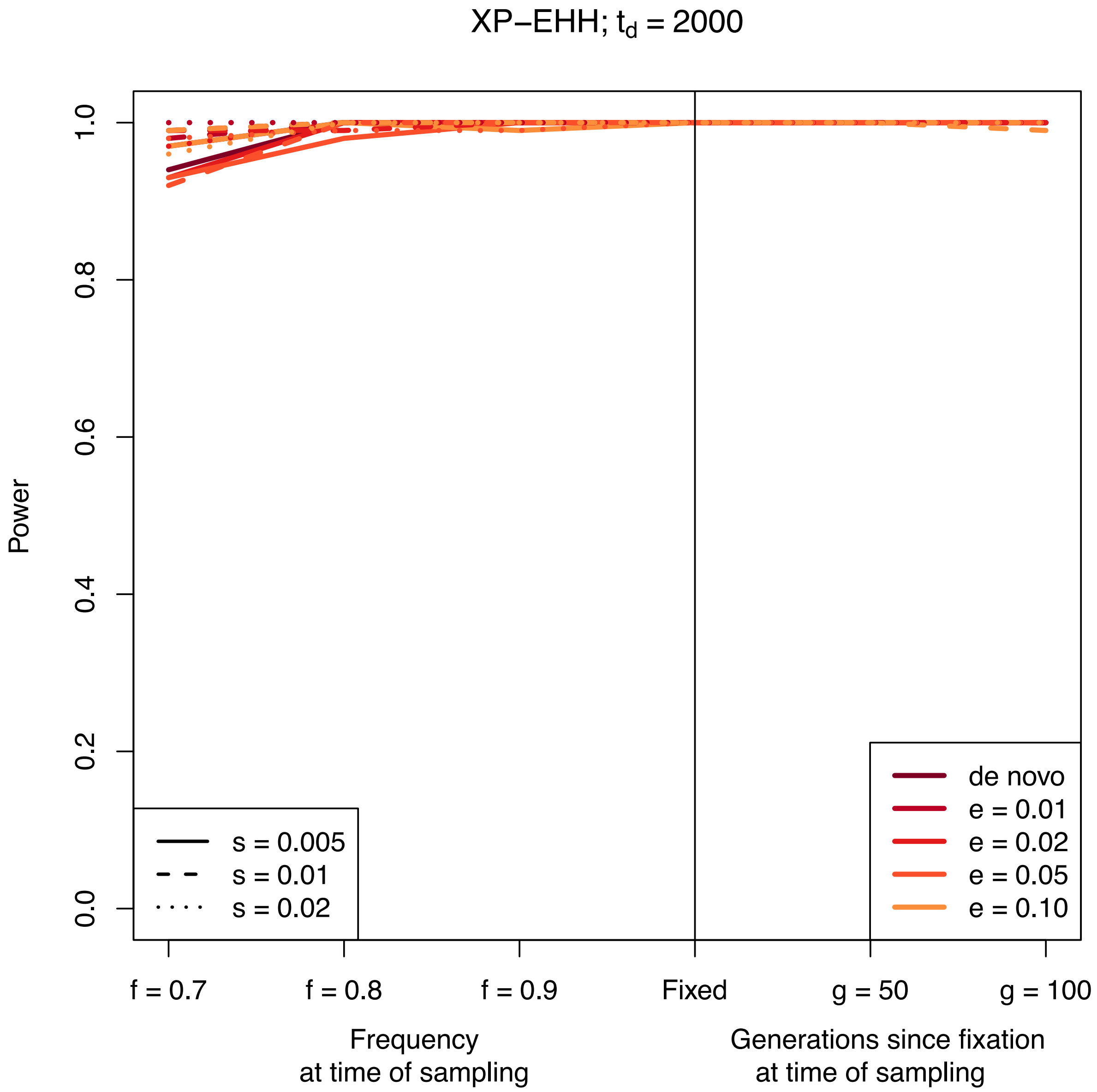
309 sampling since fixation, e is the frequency at which selection began, and t_d is the time in

310 generations since the two populations diverged.



311

312 **Figure S24**. Demo 4 nSL $t_d = 8000$ power curves. s is the selection coefficient, f is the
313 frequency of the adaptive allele at time of sampling, g is the number of generations at time of
314 sampling since fixation, e is the frequency at which selection began, and t_d is the time in
315 generations since the two populations diverged.



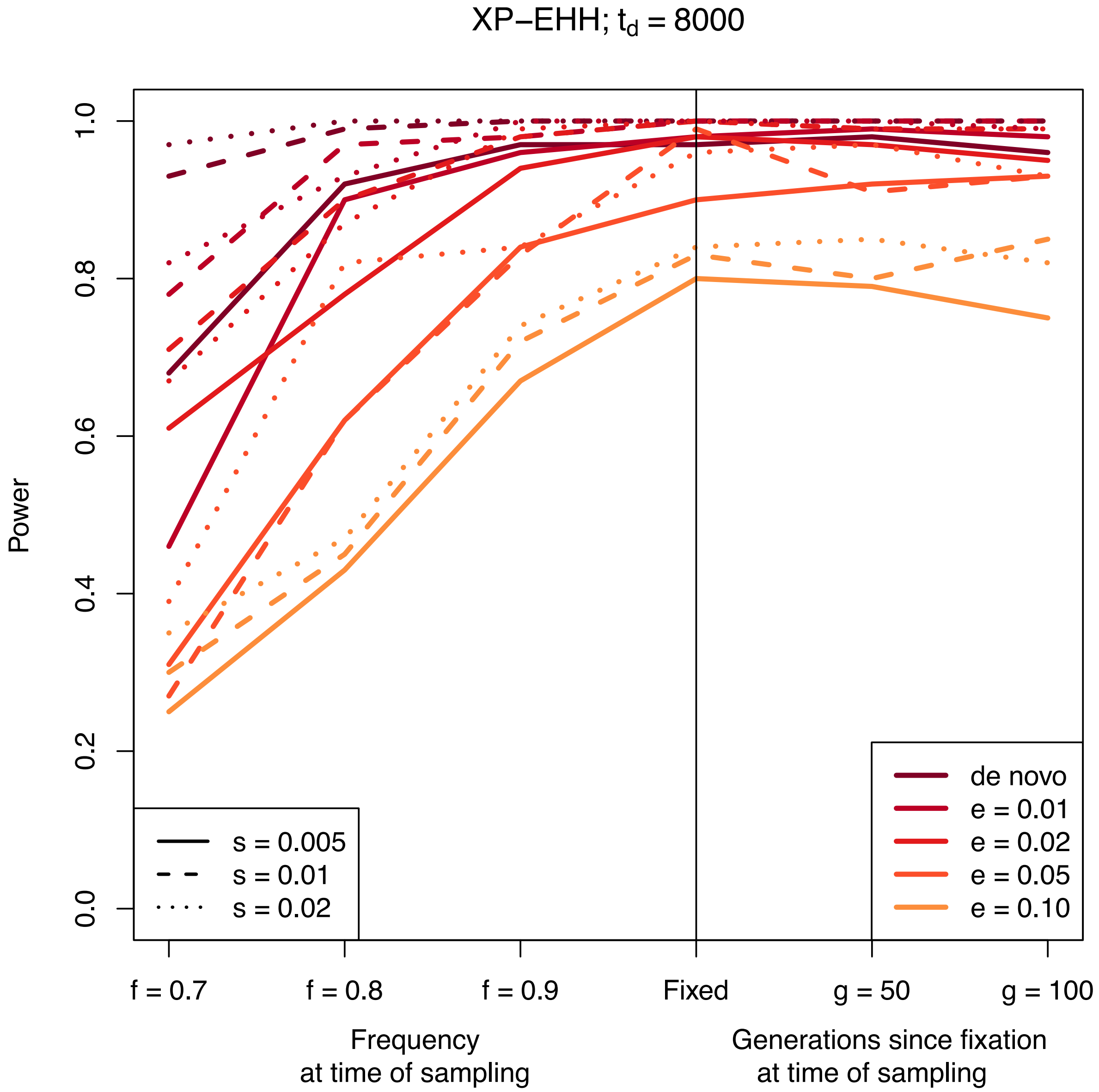
316

317 **Figure S25**. Demo 4 XP-EHH $t_d = 2000$ power curves. s is the selection coefficient, f is the

318 frequency of the adaptive allele at time of sampling, g is the number of generations at time of

319 sampling since fixation, e is the frequency at which selection began, and t_d is the time in

320 generations since the two populations diverged.



321

322 **Figure S26**. Demo 4 XP-EHH $t_d = 8000$ power curves. s is the selection coefficient, f is the

323 frequency of the adaptive allele at time of sampling, g is the number of generations at time of

324 sampling since fixation, e is the frequency at which selection began, and t_d is the time in

325 generations since the two populations diverged.

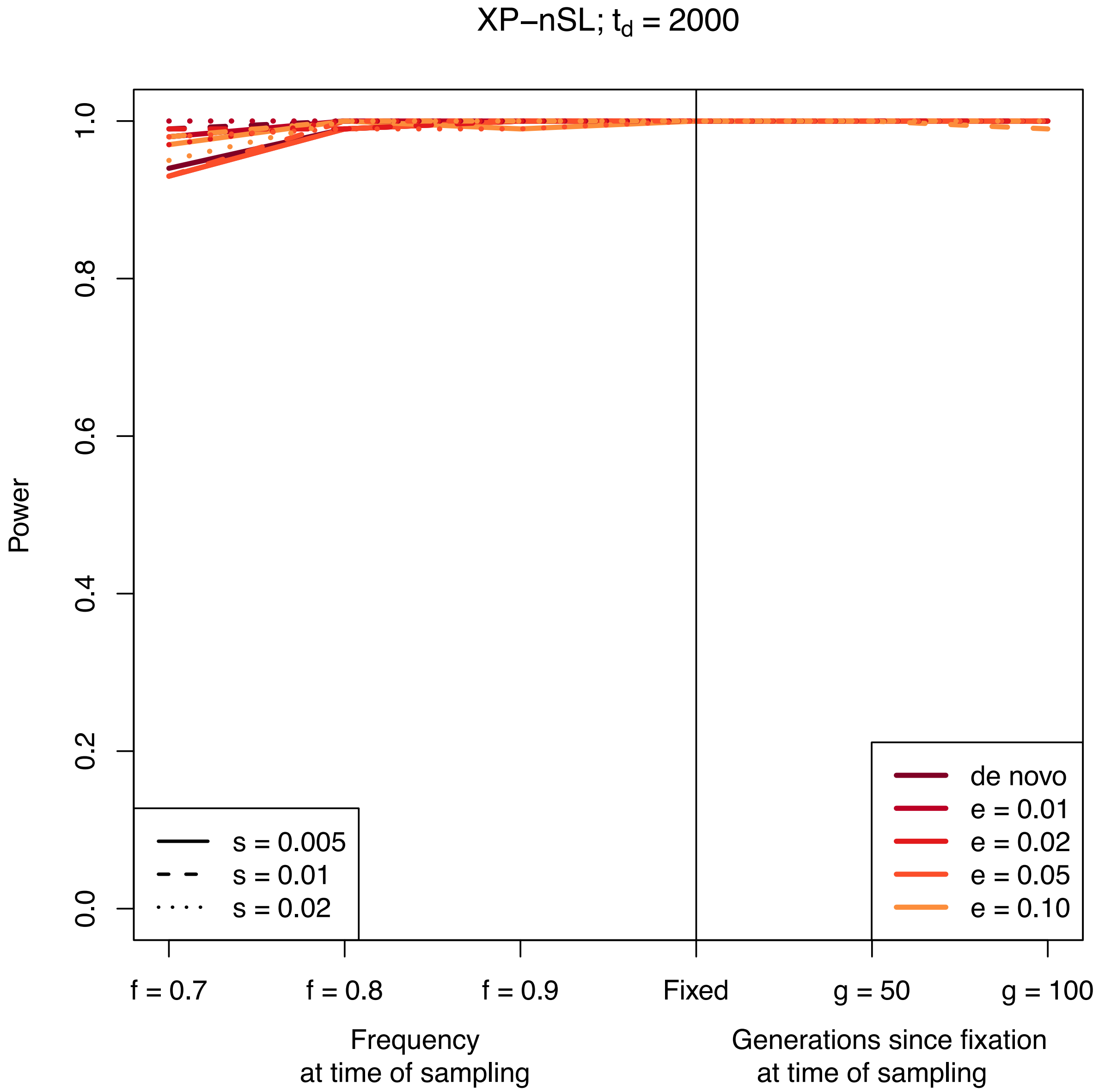
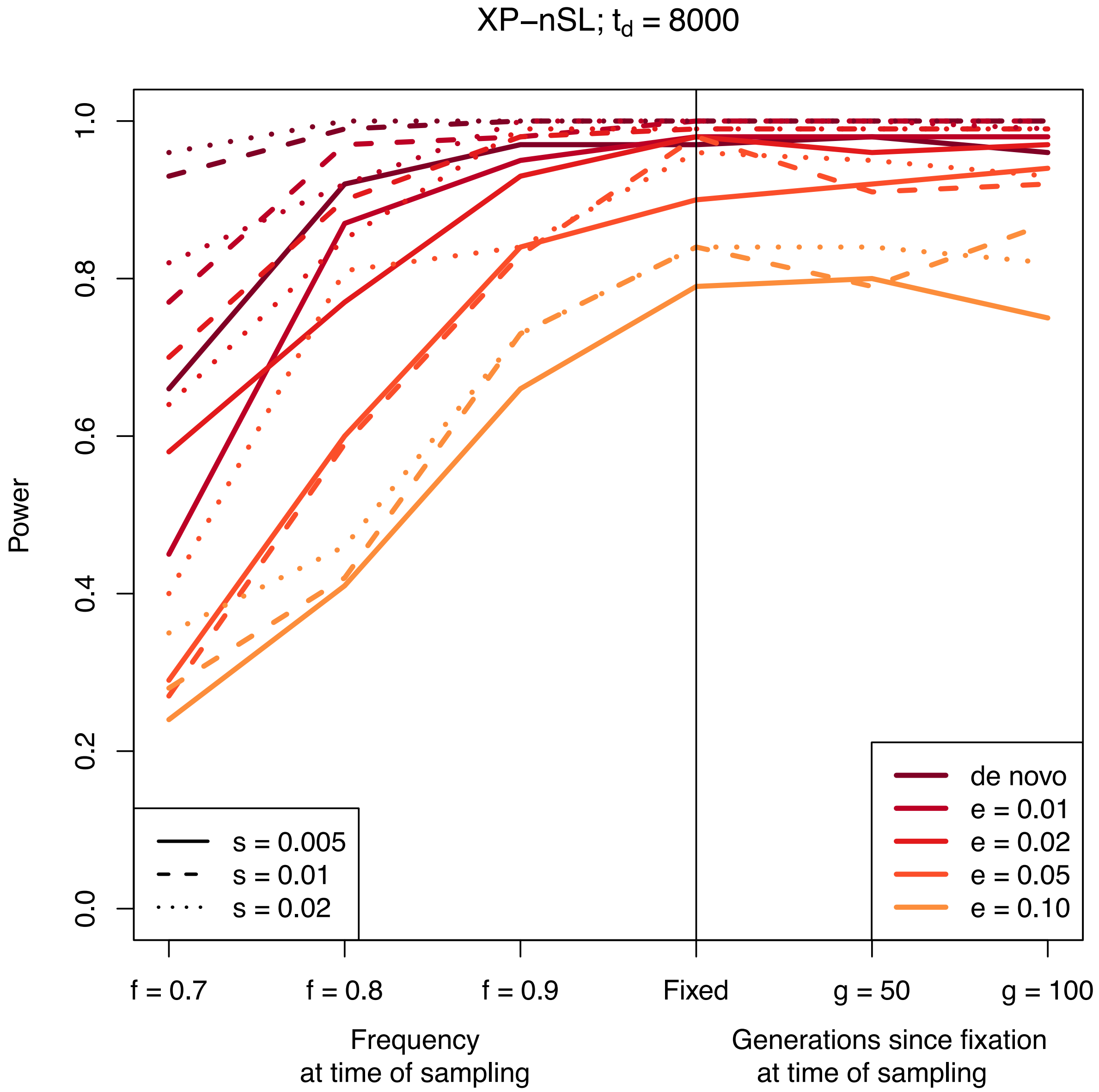
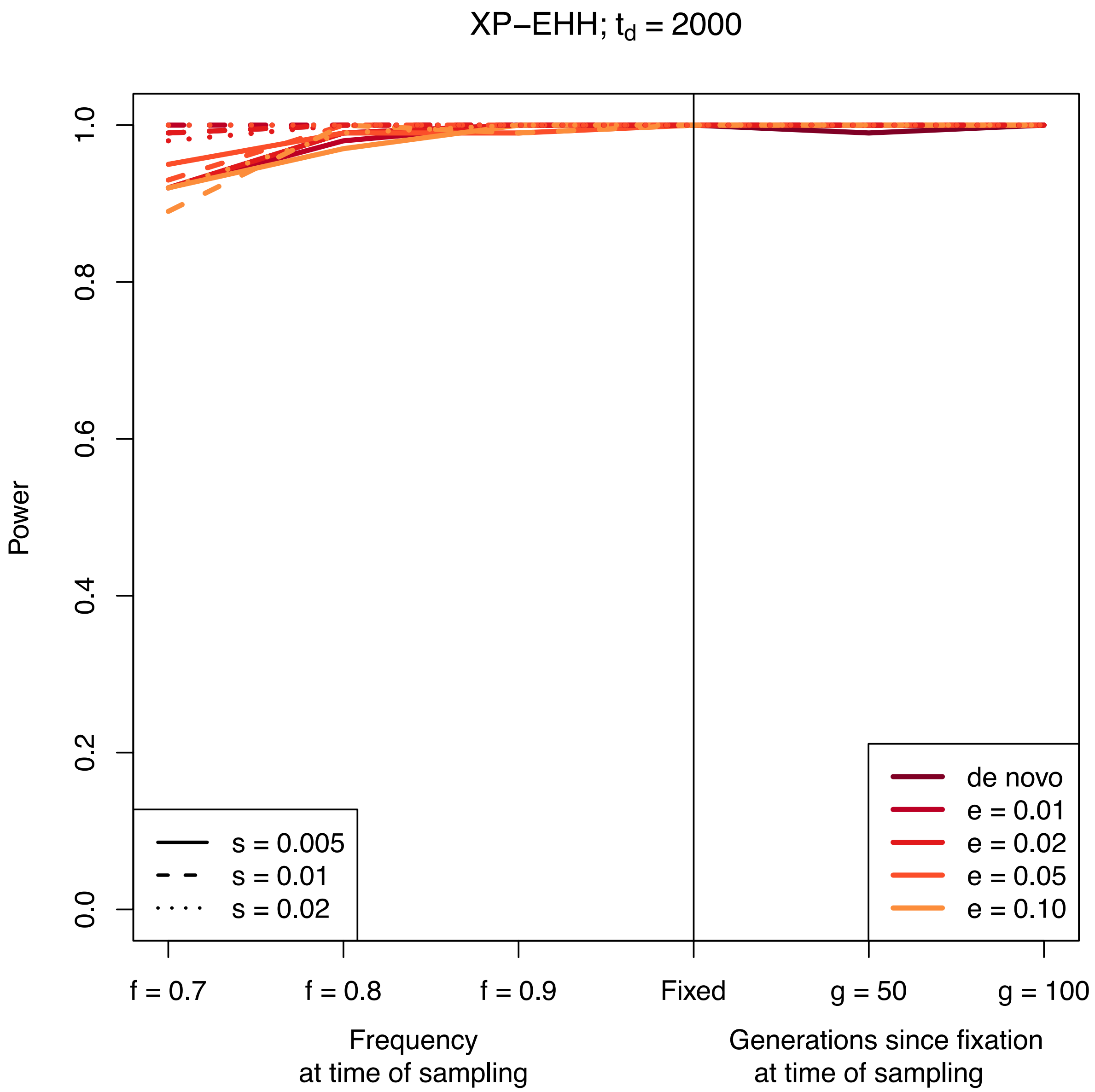


Figure S27. Demo 4 XP-nSL $t_d = 2000$ power curves. s is the selection coefficient, f is the frequency of the adaptive allele at time of sampling, g is the number of generations at time of sampling since fixation, e is the frequency at which selection began, and t_d is the time in generations since the two populations diverged.



331

332 **Figure S28**. Demo 4 XP-nSL $t_d = 8000$ power curves. s is the selection coefficient, f is the
333 frequency of the adaptive allele at time of sampling, g is the number of generations at time of
334 sampling since fixation, e is the frequency at which selection began, and t_d is the time in
335 generations since the two populations diverged.



336

337 **Figure S29**. Demo 5 XP-EHH $t_d = 2000$ power curves. s is the selection coefficient, f is the
338 frequency of the adaptive allele at time of sampling, g is the number of generations at time of
339 sampling since fixation, e is the frequency at which selection began, and t_d is the time in
340 generations since the two populations diverged.

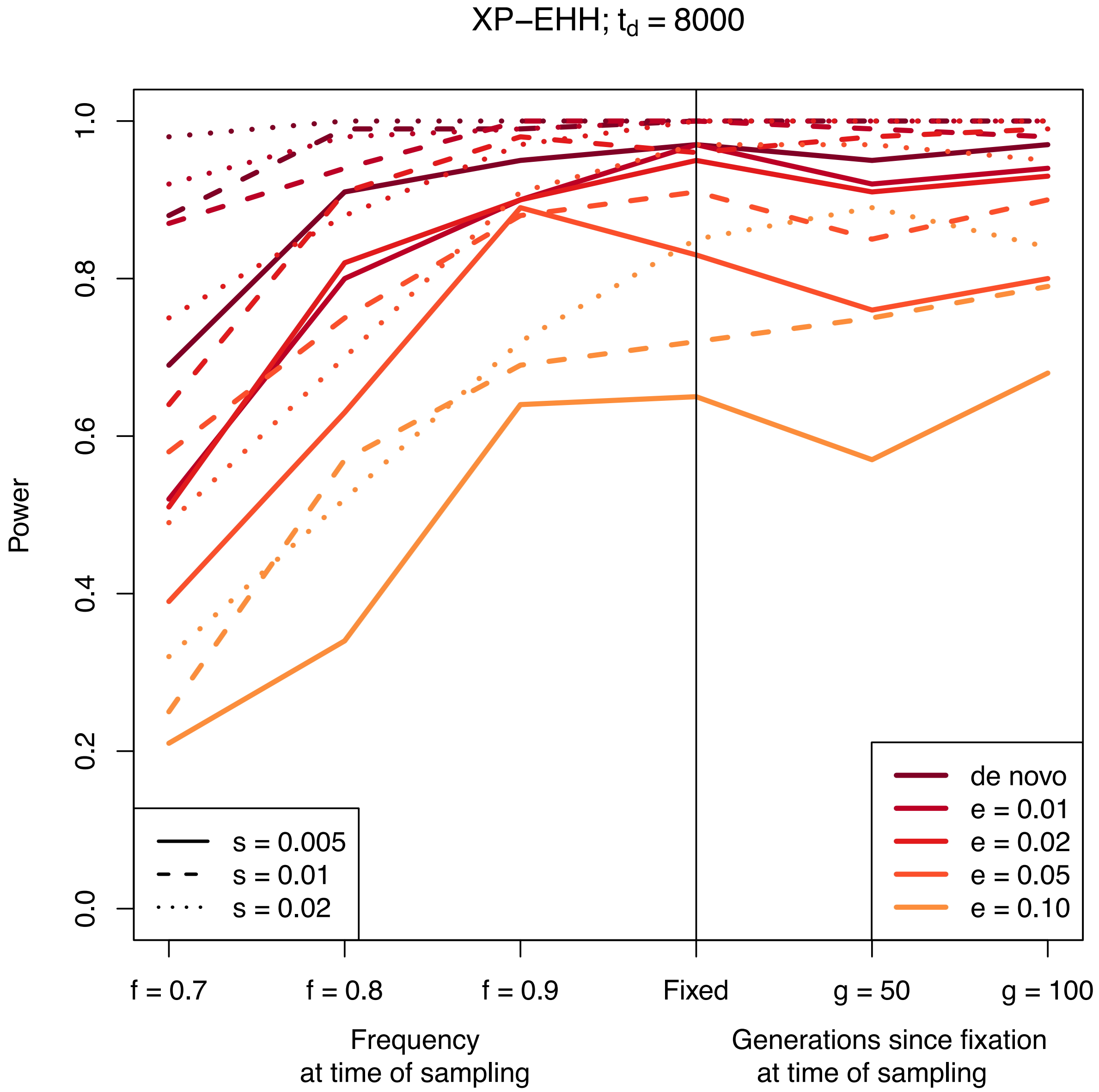
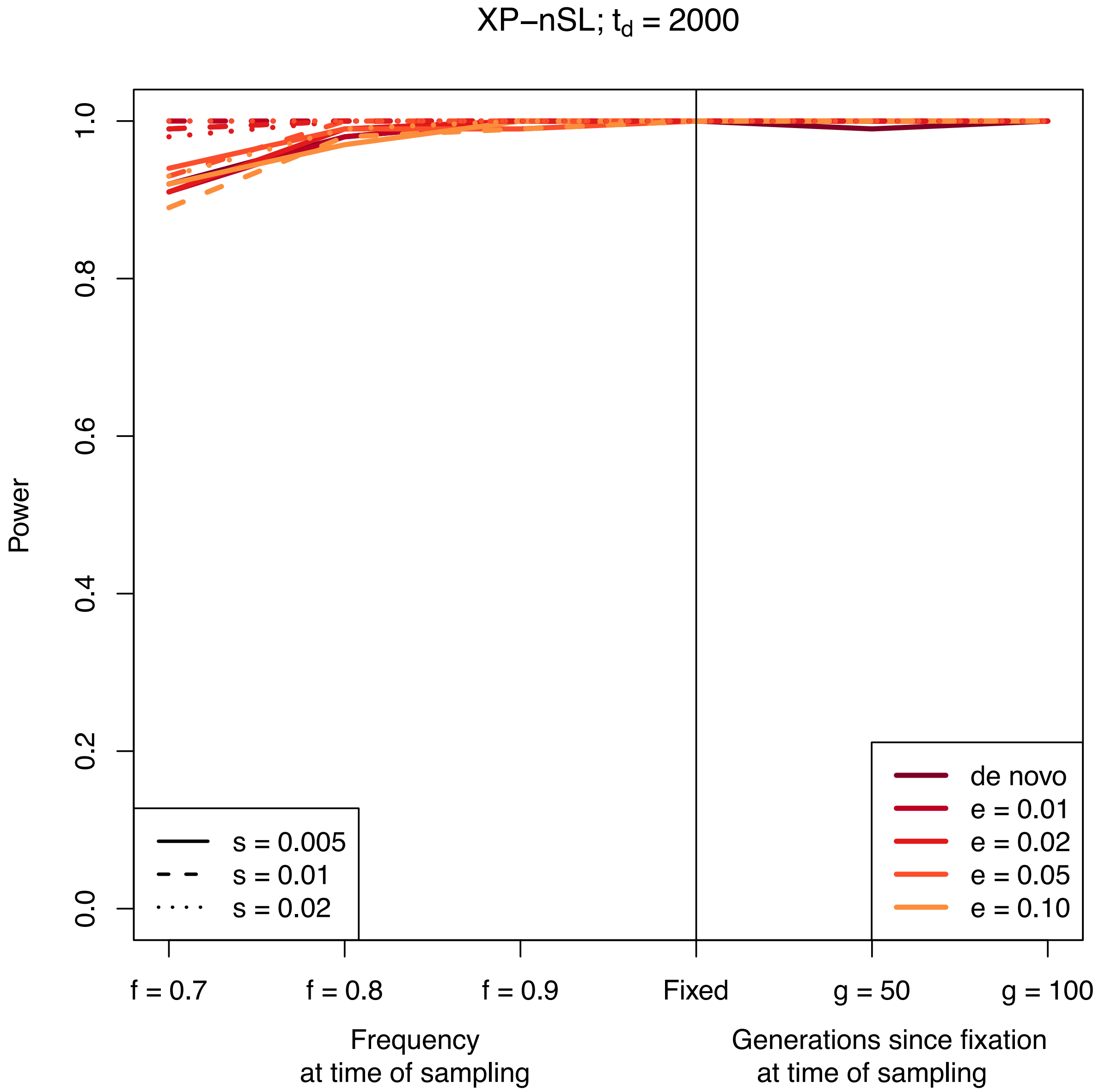


Figure S30. Demo 5 XP-EHH $t_d = 8000$ power curves. s is the selection coefficient, f is the frequency of the adaptive allele at time of sampling, g is the number of generations at time of sampling since fixation, e is the frequency at which selection began, and t_d is the time in generations since the two populations diverged.



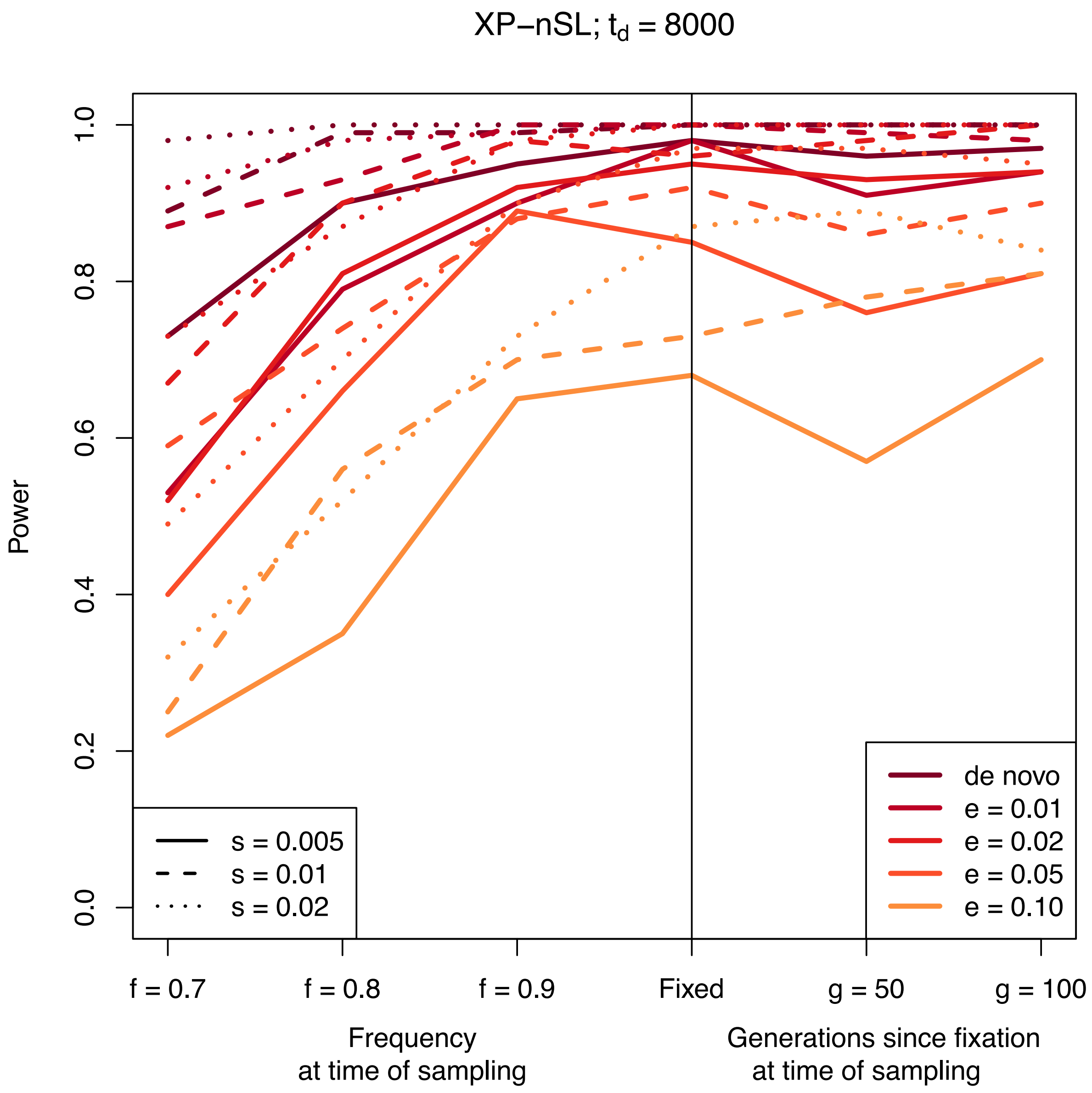
346

347 **Figure S31**. Demo 5 XP-nSL $t_d = 2000$ power curves. s is the selection coefficient, f is the

348 frequency of the adaptive allele at time of sampling, g is the number of generations at time of

349 sampling since fixation, e is the frequency at which selection began, and t_d is the time in

350 generations since the two populations diverged.



351

352 **Figure S32**. Demo 5 XP-nSL $t_d = 8000$ power curves. s is the selection coefficient, f is the
353 frequency of the adaptive allele at time of sampling, g is the number of generations at time of
354 sampling since fixation, e is the frequency at which selection began, and t_d is the time in
355 generations since the two populations diverged.

356
357
358
359
360

Table 1. Demographic history parameters for simulations.

	N_A	N_0 at split	N_0 at present	N_1 at split	N_1 at present	t_d
Demo 1	10,000	10,000	10,000	10,000	10,000	2,000/4,000/8,000
Demo 2	10,000	10,000	10,000	5,000	5,000	2,000/4,000/8,000
Demo 3	10,000	5,000	5,000	10,000	10,000	2,000/4,000/8,000
Demo 4	10,000	10,000	50,000†	10,000	10,000	2,000/4,000/8,000
Demo 5	10,000	10,000	10,000	10,000	50,000†	2,000/4,000/8,000

†The reached via exponential growth starting 2,000 generations ago.

Table S1. False positive rate computed from neutral simulations for varying t_d and demographic history.

		$t_d = 2000$	$t_d = 4000$	$t_d = 8000$
iHS	Demo 1	0.013	0.1	0.009
	Demo 3	0.007	0.013	0.007
	Demo 4	0.015	0.018	0.008
nSL	Demo 1	0.01	0.015	0.008
	Demo 3	0.008	0.011	0.007
	Demo 4	0.014	0.021	0.014
XP-EHH	Demo 1	0.013	0.013	0.016
	Demo 2	0.017	0.009	0.015
	Demo 3	0.01	0.011	0.012
	Demo 4	0.012	0.014	0.014
	Demo 5	0.011	0.012	0.013
XP-nSL	Demo 1	0.014	0.011	0.013
	Demo 2	0.019	0.011	0.012
	Demo 3	0.011	0.011	0.012
	Demo 4	0.012	0.012	0.014
	Demo 5	0.011	0.012	0.014

References

Colonna V, Ayub Q, Chen Y, Pagani L, Luisi P, Pybus M, Garrison E, Xue Y, Tyler-Smith C, Genomes Project C, et al. 2014. Human genomic regions with exceptionally high levels of population differentiation identified from 911 whole-genome sequences. Genome Biol 15:R88.

Crawford NG, Kelly DE, Hansen MEB, Beltrame MH, Fan S, Bowman SL, Jewett E, Ranciaro A, Thompson S, Lo Y, et al. 2017. Loci associated with skin pigmentation identified in African populations. Science 358.

DeGiorgio M, Szpiech ZA. 2021. A spatially aware likelihood test to detect sweeps from haplotype distributions. bioRxiv:2021.2005.2012.443825.

Ferrer-Admetlla A, Liang M, Korneliussen T, Nielsen R. 2014. On detecting incomplete soft or hard selective sweeps using haplotype structure. Mol Biol Evol 31:1275-1291.

Harris AM, DeGiorgio M. 2020. A likelihood approach for uncovering selective sweep signatures from haplotype data. Mol Biol Evol.

Harris AM, Garud NR, DeGiorgio M. 2018. Detection and Classification of Hard and Soft Sweeps from Unphased Genotypes by Multilocus Genotype Identity. Genetics 210:1429-1452.

Kern AD, Schrider DR. 2016. Discoal: flexible coalescent simulations with selection. Bioinformatics 32:3839-3841.

Lu K, Wei L, Li X, Wang Y, Wu J, Liu M, Zhang C, Chen Z, Xiao Z, Jian H, et al. 2019. Whole-genome resequencing reveals Brassica napus origin and genetic loci involved in its improvement. Nat Commun 10:1154.

Meier JI, Marques DA, Wagner CE, Excoffier L, Seehausen O. 2018. Genomics of Parallel Ecological Speciation in Lake Victoria Cichlids. Mol Biol Evol 35:1489-1506.

Nedelec Y, Sanz J, Baharian G, Szpiech ZA, Pacis A, Dumaine A, Grenier JC, Freiman A, Sams AJ, Hebert S, et al. 2016. Genetic Ancestry and Natural Selection Drive Population Differences in Immune Responses to Pathogens. Cell 167:657-669 e621.

Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll SA, Gaudet R, et al. 2007. Genome-wide detection and characterization of positive selection in human populations. Nature 449:913-918.

Salmon P, Jacobs A, Ahren D, Biard C, Dingemanse NJ, Dominoni DM, Helm B, Lundberg M, Senar JC, Sprau P, et al. 2021. Continent-wide genomic signatures of adaptation to urbanisation in a songbird across Europe. Nat Commun 12:2983.

Szpiech ZA, Hernandez RD. 2014. selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. Mol Biol Evol 31:2824-2827.

Szpiech ZA, Novak TE, Bailey NP, Stevison LS. 2021. Application of a novel haplotype-based scan for local adaptation to study high-altitude adaptation in rhesus macaques. Evol Lett 5:408-421.

Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. Plos Biology 4:e72.

Zhang SJ, Wang GD, Ma P, Zhang LL, Yin TT, Liu YH, Otecko NO, Wang M, Ma YP, Wang L, et al. 2020. Genomic regions under selection in the feralization of the dingoes. Nat Commun 11:671.

Zoledziewska M, Sidore C, Chiang CWK, Sanna S, Mulas A, Steri M, Busonero F, Marcus JH, Marongiu M, Maschio A, et al. 2015. Height-reducing variants and selection for short stature in Sardinia. Nat Genet 47:1352-1356.