1 **Title:** Towards absolute abundance for conservation applications: estimating the number of

2 contributors via microhaplotype genotyping of mixed-DNA samples

3 **Running title**: Microhaplotype genotyping of DNA mixtures

4 **Authors:** Yue Shi[1,2,*], Cory M. Dick[3], Kirby Karpan[4], Diana Baetscher[4],

5

6 **Contact information**:

7 [1] College of Fisheries and Ocean Sciences, University of Alaska Fairbanks, 17101 Point Lena

8 Loop Road, Juneau, AK 99801, USA.

9 [2] Wisconsin Cooperative Fishery Research Unit, College of Natural Resources, University of

10 Wisconsin-Stevens Point, 800 Reserve Street, Stevens Point, WI 54481, USA.

11 [3] California Cooperative Fish and Wildlife Research Unit, Department of Fisheries Biology,

12 Humboldt State University. 1 Harpst Street, Arcata, CA 95521, USA.

13 [4] National Oceanic and Atmospheric Administration, National Marine Fisheries Service, Alaska

14 Fisheries Science Center, Auke Bay Laboratories, 17109 Point Lena Loop Road, Juneau, AK

15 99801, USA.

16 [5] U.S. Geological Survey, California Cooperative Fish and Wildlife Research Unit, Department

17 of Fisheries Biology, Humboldt State University. 1 Harpst Street, Arcata, CA 95521, USA.

18 [6] U.S. Geological Survey, New York Cooperative Fish and Wildlife Research Unit, Cornell

19 University, Ithaca, NY 14853, USA

20 [7] U.S. Geological Survey, Wisconsin Cooperative Fishery Research Unit, College of Natural
21 Resources, University of Wisconsin-Stevens Point, 800 Reserve Street, Stevens Point, WI 54481,
22 USA.

23 *Corresponding author. yshi8@alaska.edu.

24

25
26
27
28
29

Abstract

Molecular methods including metabarcoding and qPCR have shown promises for estimating species abundance by quantifying the concentration of genetic material in field samples. However, the relationship between specimen abundance and detectable concentrations of genetic material is often variable in practice. DNA mixture analysis represents an alternative approach to quantify specimen abundance based on the identity of unique alleles in a sample. The DNA mixture approach provides novel opportunities to inform ecology and conservation by estimating the absolute abundance of target taxa through molecular methods; yet, challenges with genotyping many highly variable markers in mixed-DNA samples have prevented its widespread use. To advance molecular approaches for abundance estimation we explored the utility of microhaplotypes for DNA mixture analysis by applying a 125-marker panel to 1,179 Chinook salmon (*Oncorhynchus tshawytscha*) smolts from the Sacramento-San Joaquin Delta. We assessed the accuracy of DNA mixture analysis through a combination of mock mixtures containing DNA from up to 20 smolts and a trophic ecological application enumerating smolts in predator diets. Mock DNA mixtures of up to 10 smolts could reliably be resolved using microhaplotypes and increasing the panel size would likely facilitate identification of more individuals. However, while analysis of predator gastrointestinal tract contents indicated DNA mixture analysis could discern the presence of multiple prey items, poor DNA quality prevented accurate genotyping and abundance estimation. Our results indicate that DNA mixture analysis can perform well with high-quality DNA, but methodological improvements in genotyping degraded DNA are necessary before this approach can be used on marginal quality samples.

51 **Keywords:** microhaplotype, DNA mixtures, number of contributors, absolute abundance, GT-
52 seq, diet analysis

53 **Introduction**

54 Molecular tools can provide important insights on species abundance, which is critical for many
55 ecological and conservation applications, such as understanding population dynamics
56 (Bravington, Skaug, & Anderson, 2016; Roy et al., 2014), assessing dietary profiles (Shi,
57 Hoareau, Reese, & Wasser, 2021), investigating community composition (Gehri, Larson,
58 Gruenthal, Sard, & Shi, 2021), and biomonitoring (Darling & Blum, 2007). One popular
59 ecological application is molecular diet analysis of fecal samples or stomach content samples to
60 estimate the composition of prey species consumed by predators (King, Read, Traugott, &
61 Symondson, 2008). The two primary methods used to conduct molecular diet analysis are
62 quantitative polymerase chain reaction (qPCR) and metabarcoding (Deiner et al., 2017; Harper et
63 al., 2018; Pompanon et al., 2012). qPCR is conducted by designing species-specific primers and
64 tracking DNA amplification across PCR cycles, and metabarcoding is conducted by amplifying
65 primers targeting certain taxa (e.g., vertebrates, fishes) and sequencing the targeted amplicons on
66 a high-throughput sequencer (e.g., Illumina MiSeq).

67 While the majority of molecular diet studies have focused on detecting presence/absence, there
68 has been substantial interests in using these tools to estimate species abundance for conservation
69 and management (Rourke et al., 2022). Past studies have demonstrated that both qPCR and
70 metabarcoding can provide information on the amount of input DNA in a sample that can
71 theoretically be used to estimate species abundance (Hänfling et al., 2016; Shelton et al., 2019).
72 While there is often a positive correlation between qPCR or metabarcoding results and

73 abundance in laboratory setting, the correlation is often variable and can be weak in natural
74 environments (Fonseca, 2018; Kelly, Shelton, & Gallego, 2019; Yates, Fraser, & Derry, 2019).
75 qPCR can directly estimate the amount of species-specific target DNA present in a sample and
76 therefore is not influenced by the mixture of species in a sample (Nathan, Simmons, Wegleitner,
77 Jerde, & Mahon, 2014). Metabarcoding, on the other hand, produces read counts for each species
78 present in each sample. Relative read abundance produced from a metabarcoding study is
79 effective for determining the major taxa in an environment, but there is only a weak quantitative
80 relationship between RRA and input DNA amount due to technical bias across species in the
81 processes of sampling, library prep and sequencing (Harrison, Calder, Shuman, & Buerkle, 2021;
82 Lamb et al., 2019). That being said, relative read abundance could still be informative if no other
83 data about community composition exists (Deagle et al., 2018) and the incorporation of internal
84 standards or "spike-in" into DNA pools can help ameliorate bias with this approach (Harrison et
85 al., 2021; Thomas, Deagle, Eveson, Harsch, & Trites, 2015).

86 Estimates of abundance using molecular tools are improving and will likely continue to improve
87 in the future. However, even if researchers were able to calculate the input DNA amount from a
88 given species in a given sample without bias, input DNA amount may not accurately reflect
89 organismal abundance due to variation in animal size, shedding rates, digestion rates, and myriad
90 other environmental factors (Barnes et al., 2014; Carreon-Martinez, Johnson, Ludsin, & Heath,
91 2011; Levi et al., 2019; Stoeckle et al., 2017). An alternative approach is to estimate absolute
92 abundance in a sample by leveraging within-species genetic variation to quantify the number of
93 unique contributors in a mixed-DNA sample (Curran, Triggs, Buckleton, & Weir, 1999; Weir,
94 Triggs, Stowell, Walsh, & Buckleton, 1997). A major advantage of this DNA mixtures approach
95 over qPCR and metabarcoding is that abundance estimates are decoupled from DNA quantity. In

96 other words, as long as sufficient DNA from each individual is present in the sample and as long
97 as individuals can be distinguished genetically, the absolute abundance estimate is insensitive to
98 differences in the amount of DNA contributed by individual specimens (Sethi, Larson,
99 Turnquist, & Isermann, 2019). This means that factors that influence input DNA quantities such
100 as organism size, sloughing rate, and digestion rate do not influence estimates to the same degree
101 as with qPCR and metabarcoding. The ability to count the number of organisms present in a
102 mixed DNA sample opens up promising opportunities for count-based ecological inferences,
103 including but not limited to estimating the number of individuals on an invasive species front
104 from eDNA, estimating the absolute abundance of a low-population species of conservation
105 concern from eDNA, or estimating the number of individuals of an endangered prey species
106 consumed by invasive predators using diet samples (Sethi et al., 2019).

107 The DNA mixtures approach, which was first utilized for criminal forensics, relies on identifying
108 unique genetic variation at the individual level to infer the number of contributors to a mixed
109 DNA sample (Bieber, Buckleton, Budowle, Butler, & Coble, 2016; Haned, Pène, Lobry, Dufour,
110 & Pontier, 2011; Weir et al., 1997). An early ecological application of DNA mixture analysis
111 involved genotyping five microsatellites and using a heuristic 'allele counting' approach to
112 estimate the number of larval yellow perch (*Perca flavescens)* consumed by predators in river
113 plumes (Carreon-Martinez, Wellband, Johnson, Ludsin, & Heath, 2014). While allele counting is
114 conceptionally simple, this approach is not reliable beyond 3-individual mixtures (Dembinski,
115 Sobieralski, & Picard, 2018). In comparison, the maximum likelihood-based approach can make
116 explicit use of observed alleles present in a sample and their associated population allele
117 frequencies to substantially improve accuracy of estimates (Haned, 2011; Haned et al., 2011;
118 Perez, Mitchell, Ducasse, Tamariz, & Caragine, 2011).

119 Recently, Sethi et al. (2019) explored the utility of the likelihood approach for ecological and
120 conservation applications and found that it was possible to accurately estimate the number of
121 contributors in mixture samples containing up to 10 individuals using simulated data from SNP
122 and microsatellite panels. Sethi et al. (2019) also constructed mock mixtures containing extracted
123 DNA from 1-5 yellow perch and genotyped these mixtures with 14 microsatellite markers.
124 Results from this analysis suggested that relatively accurate results could be obtained with a
125 small panel but with a downward bias of ~1 individual on average. In the same study, Sethi et al.
126 (2019) analyzed stomach samples from predators collected in the field, which likely to contain
127 yellow perch, and demonstrated that multiple individuals could be identified; however, this
128 experiment was a proof of concept demonstration given that the number of perch consumed was
129 unknown.

130 Andres, Sethi, Lodge, & Andrés (2021) further explored the utility and performance of the DNA
131 mixtures approach by using it to estimate the number of contributors in eDNA samples. Using a
132 panel of 28 microsatellites, Andres et al. (2021) were able to accurately estimate the number of
133 contributors in mixtures of up to 10 individuals constructed with both tissue and eDNA samples
134 (i.e., filtered water) in mesocosm experiments. However, as the mixtures approached 10
135 individuals, bias of estimate increased to approximately 2-3 individuals and varied from positive
136 to negative depending on the effect of allele frequency cutoffs on individual alleles. These results
137 emphasize both the importance of rare alleles for accurately estimating the number of
138 contributors and the difficulty associated with accurately detecting these rare alleles without
139 introducing false positive alleles, which can upwardly bias estimates. While microsatellite
140 markers contain many rare alleles, making them a logical choice for DNA mixtures applications,
141 idiosyncrasies associated with microsatellite genotyping, such as PCR stutter and allelic dropout,

142 makes it challenging to accurately call rare alleles and differentiate them from artefacts (Andres
143 et al., 2021). Single-nucleotide polymorphisms (SNPs) are much easier to genotype, but because
144 they are generally biallelic, obtaining the genetic variation necessary to accurately identify
145 mixtures containing large numbers of individuals is difficult. One potential solution is to employ
146 a panel of microhaplotype markers as suggested by Andres et al. (2021). These markers leverage
147 the inherent phase information in short-read DNA sequence data to derive multi-allelic
148 microhaplotypes from multiple, proximate SNPs on the same read (Baetscher, Clemento, Ng,
149 Anderson, & Garza, 2018). Importantly, they can be genotyped accurately without the issues
150 associated with microsatellites and microhaplotypes contain low frequency alleles that are
151 important for the accuracy of the DNA mixtures approach.

152 Here, we build on the previous work by Andres et al. (2021) and Sethi et al. (2019), and apply
153 the likelihood-based DNA mixture genotyping using a 125-locus microhaplotype panel. Our
154 study was motivated by the need to inform conservation of imperiled salmon populations in the
155 Sacramento-San Joaquin Delta (hereafter referred to as the Delta). Habitat changes and the
156 introduction of non-native fish species has fundamentally altered the Delta ecosystem, and many
157 native fishes including Chinook salmon (*Oncorhynchus tshawytscha)* have experienced
158 significant declines for decades (Carlson & Satterthwaite, 2011; Munsch, Greene, Mantua, &
159 Satterthwaite, 2022). One contributing factor to Chinook salmon declines may be predation by
160 non-native fish species (Grossman, 2016). Previous studies using visual and genetic techniques
161 have shown that non-native piscivores such as striped bass (*Morone saxatilis*), largemouth bass
162 (*Micropterus salmoides*), and channel catfish (*Ictalurus punctatu*s) consume Chinook salmon in
163 the Delta (Brandl, Schreier, Conrad, May, & Baerwald, 2021; Michel, Smith, Demetras, Huff, &
164 Hayes, 2018; Sabal, Hayes, Merz, & Setka, 2019). However, the impacts of this predation on

165 Chinook salmon populations have been difficult to quantify using conventional visual and
166 molecular assessment of diet contents that generally lack information on the number of Chinook
167 salmon consumed and the rate at which various prey items are digested. Our study aims to
168 address these limitations by experimentally assessing the feasibility of the DNA mixtures
169 approach for counting Chinook salmon smolts in predator diets.

170 In this study, we first obtained haplotype frequencies by genotyping 1,179 Chinook fin-clip
171 samples using a panel of 125 microhaplotype loci. Secondly, to test the microhaplotype panel on
172 real amalgamations of DNA in a controlled setting, we estimated the number of contributors
173 (hereafter referred to as NOC) in mock mixtures containing DNA extracts from 2-20 Chinook
174 individuals. Lastly, to apply the optimized estimator in a more realistic setting, we explored the
175 utility of DNA mixture analysis for diet analysis in a large controlled-feeding experiment by
176 estimating the number of Chinook individuals found in the gastrointestinal (GI) tracts of two
177 non-native predators in the Delta, largemouth bass (LMB) and channel catfish (CCF). Our results
178 illustrate the utility of microhaplotype panels for DNA mixture analysis but also illuminate some
179 challenges associated with applying this approach to degraded DNA samples - which are
180 typically what researchers encountered in diet samples.

181 Materials and Methods

182 *Curating a Catalog of Haplotypes and Estimating Their Frequencies*

183 Accurate haplotype frequency estimation is crucial for precise estimation of NOC using the
184 DNA mixture approach (Andres et al., 2021; Sethi et al., 2019). Therefore, we aimed to genotype
185 a large number of Chinook smolts from Mokelumne hatcher in CA (N=1,179) using Genotyping-

186 in-Thousands by Sequencing (GT-seq; Campbell, Harmon, & Narum, 2015) and a panel of 125
187 microhaplotype markers. This microhaplotype panel was developed for Chinook salmon in the
188 Klamath and Sacramento river basins and has product sizes of 90 -143 bp (Thompson et al.,
189 2020). Because cross-amplification of predator DNA might interfere with NOC estimation of the
190 prey, we also genotyped fin clip samples of LMB and channel CCF specimens collected for this
191 study to examine the level of cross-amplification between predators and Chinook salmon. DNA
192 was isolated from dried fin clip samples either with Qiagen DNeasy Blood and Tissue Kits or
193 10% Chelex 100 solution containing 1% of Triton-X 100 and 1% Tween 20. One negative
194 control was included on each 96-well extraction plate. GT-seq was conducted following the
195 methods of Campbell et al. (2015) with modifications detailed in Bootsma et al. (2020) except
196 that we used the original post-normalization double-sided SPRI bead size-selection protocol of
197 0.5x to 1.2x (Campbell et al., 2015). Libraries were sequenced on the Illumina MiSeq platform
198 using a single v2 300 cycle kit (2 x 150 bp paired end). An initial GT-seq test run on 377
199 Chinook samples was conducted to evaluate the 125 microhaplotype markers and we removed
200 any loci with over or under amplification from the panel.

201 Demultiplexed reads (forward reads only) were processed with *trimmomatic* v0.39 (Bolger,
202 Lohse, & Usadel, 2014) to remove adapter sequence using the following parameters:
203 ILLUMINACLIP:2:30:10 SLIDINGWINDOW:4:15 MINLEN:50 and the adaptor sequences
204 *fasta* file provided by *trimmomatic*, TruSeq-3-PE-2.fa. After trimming, forward reads were
205 mapped to the reference file of consensus sequences of the 125 microhaplotype markers using
206 bwa-mem v 0.7.17 with default settings (Li, 2013). On-target rate was calculated for each sample
207 as the proportion of reads that aligned to amplicons in the microhaplotype panel. To assemble
208 microhaplotypes and obtain their read depths in each individual, we used the R package

MICROHAPLOT (https://github.com/ngthomas/microhaplot). MICROHAPLOT uses the reference VCF file to obtain SNP positions for each locus and assemble SNPs into microhaplotypes, and then extracts microhaplotypes from SAM files (Baetscher et al., 2018).

To obtain a reliable catalog of microhaplotypes and their frequencies from Chinook salmon tissue samples, we conducted the following filtering steps modified from (Baetscher et al., 2018): (1) we removed incomplete haplotypes, i.e., haplotypes with N or X, (2) we removed haplotypes with fewer than 20 reads at a locus and a read depth ratio of < 0.2 within an individual (read depth ratio is defined as the ratio between read depth of a haplotype at a locus and the read depth of the haplotype with the highest read depth), (3) we removed monomorphic loci, i.e. loci with only one haplotype present across all chinook samples, and (4) we removed loci with more than two haplotypes in any individuals. Genotypes were called from the remaining loci. An individual was called as a heterozygote if two haplotypes remained and a homozygote if only one haplotype remained. Finally, we used an iterative filtering approach to remove samples genotyped in < 80% of loci and remove loci genotyped in < 70% of samples. After the aforementioned filtering steps, haplotype frequency was calculated as the number of copies of a haplotype at a given locus, divided by the total number of copies present at that locus in the dataset. The large number of sampled specimens and associated haplotype frequencies are believed to be representative of population-level frequencies in the Delta.

To check cross-amplification of the microhaplotype panel in the predator fish species, we extracted DNA from LMB (N=190) and CCF (N=94) fin-clip samples and genotyped these samples using GT-seq as described above. We used the same filtering criteria on predator samples, i.e., we removed haplotypes with fewer than 20 reads at a locus and a read depth ratio

231 of less than 0.2 within an individual and assessed the read coverage at the loci/haplotypes shared
232 with Chinook salmon. To check for contamination, we conducted the same analysis on negative
233 controls samples.

234 *Estimating NOC in Mock DNA Mixtures*

235 We constructed 285 mock DNA mixtures containing DNA from 2 to 20 Chinook (Table 1)
236 smolts to assess the ability of the optimized microhaplotype panel (from above) to accurately
237 estimate NOC across variable numbers of contributors. These mock DNA mixtures were made
238 by pooling 2 µl of extracted DNA per individual and prepared in three replicates (Table 1). No
239 two pools contained the same set of individuals. Genotyping was determined using GT-seq as
240 described above, and we implemented the likelihood-based model described in (Andres et al.,
241 2021; Sethi et al., 2019) to estimate NOC in mock DNA mixtures.

242 Distinguishing true alleles from technical artefacts is relatively simple for single-source diploid
243 individuals, but this problem becomes more difficult in DNA mixture samples, which contain
244 multiple individuals and thus many alleles appear at low frequencies (Andres et al., 2021). The
245 parameter that needs to be tuned to ensure accurate detection of alleles present in DNA mixture
246 samples is the read depth ratio. For individual tissue samples, we used a read depth ratio of 0.2.
247 However, this value needs to be greatly reduced for DNA mixture analysis. We used a read depth
248 ratio of 0.02 because mock DNA mixture samples included multiple individuals, which
249 corresponds to multiple haplotypes per sample with variable read depths. Bias for mock DNA
250 mixture samples was calculated as the estimated NOC - true NOC.

251 Genetic samples collected for some ecological applications, such as molecular diet analyses, tend
252 to have lower DNA quantity and quality than tissue samples, which often results in locus
253 dropout. Therefore, we attempted to assess how variation in locus dropout rate affects the NOC
254 estimates by randomly subsampling 10% - 90% of 74 loci in the final panel for a total of nine
255 levels with each level increasing 10%. At each level of locus dropout, we also assessed how
256 NOC estimates were affected by different read depth ratio cutoffs (0.002, 0.02, 0.2). We
257 conducted 50 trials for each combination of locus dropout rate and read depth ratio.

258 *Estimating NOC in the Feeding Trial*

259 We conducted a large feeding trial to understand how temperature and predator species influence
260 digestion rates using diet analysis results obtained from visual identification, qPCR,
261 metabarcoding, and the DNA mixtures approach. Additional details on the feeding trial and
262 results from the visual, metabarcoding, and qPCR analyses are available in Dick et al. (in
263 review). Briefly, two non-native predators in the Delta, LMB and CCF, were acclimated for two
264 weeks at 15.5˚C or 18.5˚C prior to the initiation of the feeding trial. After the acclimation period,
265 individual predators were force fed three fall run Chinook salmon smolts (average 6.4g per
266 smolt). At regular intervals post-ingestion, a subset of 5-10 predators from each species by
267 temperature treatment were euthanized. GI tracts were removed and preserved in 100% non-
268 denatured ethanol. The dissections began 6 hours post-ingestion (t=6) and continued every 12
269 hours until t=96 hours, and then a final sample occurred at t=120 hours (5 days) resulting in a
270 total of ten time points (Table 2).

271 Stomach contents and stool were collected from the preserved GI tract samples. We then
272 combined stool and small pieces of each visible diet item into a 1.5 ml tube, and excess ethanol

273 was removed by centrifugation and pipetting followed by evaporation. DNA was extracted using
274 a Macherey-Nagel Nucleospin 96 DNA Stool kit with three modifications: (1) we replaced bead-
275 induced lysis with enzymatic lysis, (2) we used a per-sample volume of 25 µL of proteinase-k
276 and 850 µL of lysis buffer ST1, and (3) we incubated overnight at 56°C. See Dick et al. (in
277 review) for detailed dissection methods. One negative control was included on each 96-well
278 extraction plate. GT-seq genotyping and DNA mixture analysis was conducted in the same way
279 as described above.

280 In total, we dissected 277 GI tract content samples, including 173 samples from LMB and 104
281 samples from CCF (Table 2). These samples were genotyped using the microhaplotype panel to
282 determine the ability of the DNA mixtures method to accurately recover NOC as prey items were
283 digested. We chose a read depth ratio of 0.02 for genotyping the stomach samples, which was
284 informed by the results of the mock DNA mixture subsampling experiment and the relatively
285 small number of Chinook smolts fed to each predator. We fit an exponential decay model, $y =$
286 $a(1-r)^x$, with x=hour post-ingestion and y=mean NOC estimate to examine the loss rate of the
287 number of detected Chinook salmon smolts in predator GI tract content samples over time across
288 species and temperature. We estimated the initial amount (parameter 'a') and rate of decay
289 (parameter 'r') in each decay model.

290 **Results**

291 *Curating a Catalog of Haplotypes and Estimating Their Frequencies*

292 Initial GT-seq testing using 377 Chinook fin-clip samples showed that 11 out of 125
293 microhaplotype loci had either over or under amplification based on the total number of on-target

294 reads across individuals (Figure S1). After removing these 11 loci, the final panel consisted of
295 114 microhaplotype loci (Table S1). Primer sequence information of the 114 loci can be found in
296 Data S5 in Thompson et al. (2020). GT-seq data using the final panel of 114 microhaplotype loci
297 on 1,179 Chinook fin-clips yielded an average of 8,517 forward reads per sample (range = 1 -
298 26,336 reads) and an average of 6,509 on-target reads per sample (range = 0 - 16,656). The
299 median on-target rate of sequencing data was 80.21% (range = 0 - 98.5%), with only 11 samples
300 having on-target rate less than 40% (Figure 1).

301 After quality filtering on the extracted haplotypes, a total of 74 loci and 565 samples remained.
302 Details of the number of loci and samples remaining after each filtering step can be found in
303 Table S2. These 74 loci contained 252 unique haplotypes with a median of 3 haplotypes and a
304 range of 2-7 haplotypes per locus (Figure S2). The curated catalog of 252 haplotypes had a wide
305 range of haplotype frequencies, ranging from 0.001 to 0.997 with a median of 0.190 (Figure 2a).
306 The majority of the 74 loci contained low-frequency haplotypes. Specifically, 54 loci (73%) had
307 haplotypes with a frequency of less than 0.1, and 46 loci (62%) had haplotypes with a frequency
308 less than 0.05.

309 We applied the same haplotype filtering on 190 LMB and 94 CCF fin-clip samples. Four
310 LMB/CCF samples were outliers in terms of total number of on-target reads (1,166 - 11,544
311 reads; Figure S3). After filtering, these four samples still had nonnegligible amount of on-target
312 reads remained (51 - 3,084 reads) whereas the rest of the LMB/CCF samples had zero or close to
313 zero on-target reads after filtering. These four outlier samples shared 80 haplotypes across 58
314 loci with Chinook. Our results suggest that these four samples were likely contaminated with
315 Chinook salmon DNA, and thus we conclude that overall, there was no evidence of cross-

316 amplification of our microhaplotype panel between two predator fish species (LMB and CCF)
317 and Chinook salmon. In addition, there was no systematic contamination in our dataset as the
318 number of on-target reads after filtering was zero across all negative control samples.

319 *Estimating NOC in Mock DNA Mixtures*

320 The 285 mock DNA mixture samples yielded an average of 9,001 forward reads per sample
321 (range = 1 - 13,038 reads) and an average of 6,807 on-target reads per sample (range = 0 -
322 9,603). Three samples were dropped due to failed library prep and sequencing run. For the
323 remaining 282 samples, the median on-target rate was 76.24% (range = 64.52 - 82.51%; Figure
324 1). All 74 microhaplotype loci were successfully genotyped in all 282 mock DNA mixture
325 samples (Figure 3a). Using the curated catalog of haplotypes across the 74 microhaplotype loci
326 described above along with their frequencies and a read depth ratio of 0.02, NOC estimates
327 generally fell within ± 2 from the true NOC in mock DNA mixtures of up to 10 individuals with
328 a mean bias of 0.2 ± 1.1 (Figure 2b). However, apparent negative bias emerged when true NOC
329 was greater than 10, with a mean bias = -2 ± 2.1 for NOC=15 and mean bias = -5.7 ± 1.6 for
330 NOC=20 (Figure 2b).

331 To assess the effects of different levels of locus dropout on the NOC estimate, we subsampled 10
332 - 90% of the above 74 microhaplotype loci, which corresponded to 7 to 67 loci for a total of nine
333 levels with each level increasing 10%. Fewer number of loci genotyped corresponded to the
334 higher levels of locus dropout. At each level of locus dropout, we compared three different read
335 depth ratio cutoffs (0.002, 0.02, 0.2; Figure 4). At the read depth ratio of 0.02, higher locus
336 dropout rate resulted in larger variance in NOC estimates with the largest variances observed
337 when genotyping coverage was 20% of loci or less (≤15 loci retained) across all NOC scenarios

338 tested, although the effect was minimal when true NOC was 2 or 3 (Figure S4). In addition, with
339 higher locus dropout rate, the estimate bias moved in the positive direction, especially when true
340 NOC was less than 15. Interestingly, locus dropout rate only had a marginal effect on the mean
341 estimate bias, which was within ± 2 from true NOC up to 15 individuals, suggesting moderate
342 robustness in DNA mixture analysis to locus dropout type errors (Figure S4). Patterns of
343 variance and mean estimate bias at the lowest read depth ratio (0.002) were similar to what was
344 observed at the ratio of 0.02 (Figure 4), indicating such a low threshold of 0.002 was likely
345 below the read ratio of all haplotypes within samples. In contrast, the highest read ratio of 0.2
346 likely exceeded the read ratio of all but the dominant haplotypes (i.e. haplotypes with highest
347 read depth) within samples, resulting in negative bias in NOC estimates (Figure 4 & Figure S4).

348 *Estimating NOC in the Feeding Trial*

349 In the feeding trial, we force fed LMB and CCF with three Delta Chinook salmon smolts at two
350 different temperature conditions (15.5˚C and 18.5˚C). A total of 277 GI tract content samples
351 (173 LMB samples and 104 CCF samples) yielded an average of 11,453 forward reads per
352 sample (range = 0 - 111,528 reads) and an average of 2,230 on-target reads per sample (range =
353 0 - 46,580). Compared to mock DNA mixture samples, these GI tract samples demonstrated a
354 wide range of on-target rates across samples (0 - 87.82%) with the median of 14.66% (Figure 1).
355 The wide range of on-target rates was associated with time post-ingestion (Figure 5a).
356 Specifically, the on-target rate decreased over time in both species and dropped significantly
357 after 72 hours at 15.5 ˚C and after 48 hours at 18.5 ˚C (Figure 5a). Six samples were removed
358 due to extremely low on-target reads (≤ 2 reads). For the remaining 271 samples, number of
359 successfully genotyped loci increased with the number of on-target reads (Figure 3b). When the

360 total on-target reads reached 429 reads or above (N=96), at least 90% of 74 loci (67 loci) were
361 genotyped (Figure 3b). We further removed 89 samples with fewer than 20% of 74 loci
362 genotyped (15 loci) because too few genotyped loci led to large variance in NOC estimate based
363 on the subsampling experiment (Figure 4). Notably, these removed samples included all or most
364 samples at 84 - 120 hours post-ingestion in CCF (Table S3). In general, we observed a
365 downward trend in estimated NOC over time in both CCF and LMB (Figure 5b), and mean
366 estimates of NOC were larger than one for up to 48-72 hours though with high variance (Table
367 S3). Both LMB and CCF showed exponential decay patterns in NOC through time, presumably
368 as Chinook salmon DNA was digested or evacuated from predator guts (Figure S5). The two
369 predator species showed different patterns of decay (Figure S5 and Table S4). The initial amount
370 at six hours (parameter 'a') was slightly greater in CCF compared to LMB at either temperature.
371 In addition, the rate of decay (parameter 'r') was slightly faster for CCF compared to LMB at
372 either temperature. Notably, higher temperature was associated with higher rate of decay in NOC
373 estimates of Chinook salmon smolts over time post-ingestion for CCF whereas the temperature
374 effect for LMB was weaker (Figure S5; Table S4).

375 **Discussion**

376 Our study provided strong evidence that the likelihood-based DNA mixture analysis paired with
377 a sufficiently variable microhaplotype panel can be used to accurately quantify the number of
378 contributors to mixed DNA samples containing up to ten individuals and possibly more.
379 However, we faced substantial methodological challenges associated with highly degraded DNA
380 when applying this method to GI tract content samples from piscivorous fish predators in a
381 feeding trial. Our results reveal promises, but also potential pitfalls associated with the DNA

382 mixtures approach. Below we discuss the methodological advances achieved in this study, some
383 important considerations and limitations of the study, and how to potentially address them in the
384 future.

385 *The DNA mixture analysis paired with a microhaplotype panel: a promising approach for future*
386 *studies*

387 The most significant advancement of our study is demonstrating the benefit of microhaplotype
388 markers for DNA mixture analysis. The most recent study to conduct similar analyses by Andres
389 et al. (2021) used microsatellite markers genotyped with high-throughput sequencing and faced
390 significant difficulty calling low-frequency alleles. They recommended that future studies test
391 microhaplotype markers to specifically address this issue with low-frequency alleles. We
392 followed their advice and have confirmed that microhaplotype markers are well-suited for DNA
393 mixture analysis. The most substantial reason for this high performance is their ability to reliably
394 genotype low frequency alleles, which is critical for achieving accurate estimates of NOC.

395 It is important to note that the panel we used was developed for genetic stock identification of
396 West Coast Chinook salmon (Thompson et al., 2020) and was not designed to maximize the
397 number of haplotypes at each locus within the Delta population, which would be the goal for
398 optimizing the DNA mixtures analysis. In contrast, panels developed for parentage analysis or
399 other applications often enrich for loci with a high number of alleles, and loci containing over ten
400 alleles/haplotypes are common (Baetscher et al., 2018), compared to a maximum of seven
401 haplotypes in our study. Our results demonstrated that an existing microhaplotype panel not
402 necessarily designed for NOC estimation can still be effective for DNA mixture analysis.
403 Fortunately, designing new panels specifically for DNA mixture applications is not overly

404 onerous, and the workflow for constructing these panels has been thoroughly described in
405 previous papers (Baetscher et al., 2018; Bootsma et al., 2020).

406 Designing larger panels containing a high number of loci with more haplotypes would likely
407 facilitate accurate NOC estimates for mixtures containing more than the 10 individuals that we
408 could reliably resolve with our current panel. Previous investigations into DNA mixtures suggest
409 that the maximum number of individuals that can be resolved is a function of the number of low
410 frequency alleles present in a dataset and the ability to accurately identify them. Andres et al.
411 (2021) demonstrated that a microsatellite panel containing 28 loci and 253 total alleles could
412 accurately estimate NOC in samples of up to 58 individuals in silico, but in practice this panel
413 was limited to resolving mixtures of ~10 individuals due to issues with differentiating true rare
414 alleles from artefacts. Identification of rare alleles was more straightforward with our
415 microhaplotype panel, but we were still potentially limited by (1) the number of loci and the
416 number of total alleles and (2) sequencing coverage. Future studies could explore increasing
417 panel size and sequencing coverage to increase the number of rare alleles and the ability to
418 reliably detect them. One related option would be to design multiple small panels that are easier
419 to optimize and genotype then combine data from those panels to increase both the number of
420 loci and coverage.

421 Our resampling analysis of known mixtures suggested that the read ratio cutoff should be set as
422 low as possible to facilitate identification of rare alleles without mischaracterizing true alleles as
423 artefacts. Setting this value is a balance between biasing estimates upwards because artefact
424 alleles are retained and biasing estimates downwards because true alleles are not detected, as
425 discussed in Andres et al. (2021). Increasing sequencing coverage could allow better detection of

426 true alleles and facilitate the use of smaller read ratio cutoffs, but the utility of this approach
427 should be tested on known mixtures due to diminishing returns associated with increasing
428 sequencing coverage of finite PCR products (Rochette et al., 2022). One potential solution to this
429 issue could be to conduct multiple PCR replicates for each sample and combine the products to
430 reduce the stochastic effects of PCR (Miller, Joyce, & Waits, 2002), which could cause certain
431 alleles to amplify more readily.

432 Interestingly, our analysis of various levels of locus dropout rates revealed an unexpected
433 relationship between the number of loci genotyped and the direction of bias in NOC estimates.
434 As fewer loci were genotyped, the bias increased in the positive direction. Especially when the
435 true NOC was fewer than 15, the NOC was overestimated. In simulated data, this trend occurs in
436 the opposite direct (Sethi et al., 2019), indicating that the positive bias that we observed may be
437 due to artefacts. Specifically, we hypothesize that the upward bias due to artefact alleles is
438 reduced when additional loci are genotyped. The locus dropout subsampling results indicate that
439 accurate mean estimates of NOC can be obtained with relatively few loci, but as the number of
440 loci genotyped decreases, the variance in NOC estimates increases, and it becomes more
441 important to ensure that rare alleles are called correctly and distinguished from artefacts. Our
442 empirical data from the feeding trial suggests that, for degraded DNA samples, the percentage of
443 loci that can be successfully genotyped is positively correlated with the number of on-target
444 reads, meaning that if few loci are genotyped, the sequencing coverage for each locus is likely
445 low, potentially leading to inaccurate identification of rare alleles. This is characteristic of poor-
446 quality input DNA, such as that obtained from diets and some environmental samples. We
447 therefore urge caution when estimating NOC using genotype data when a large number of loci
448 failed to genotype. However, our resampling simulations do suggest that a relatively small

449 number of loci can be effective for estimating NOC in mixtures with ≤10 contributors if rare
450 alleles are accurately identified.

451 One aspect of NOC estimation that we anticipated could be a problem was cross-amplification of
452 microhaplotype loci in predator species. Cross-amplification could inflate the number of alleles
453 at a given locus and upwardly bias NOC estimates. Therefore, a best practice is to verify that no
454 cross-amplification between species occurs. Luckily, we found no evidence that loci included in
455 our panel amplified in LMB and CCF suggesting that cross-amplification is likely to be minimal
456 in distantly related taxa. However, cross amplification could become a problem if more closely
457 related species are analyzed, such as in systems where multiple congeners are found. Certain
458 microsatellite loci have been shown to amplify in a large number of salmonid species (Scribner,
459 Gust, & Fields, 1996; Williamson, Cordes, & May, 2002), and microhaplotype loci developed
460 for kelp rockfish (*Sebastes atrovirens)* amplify in many other *Sebastes* species (Baetscher,
461 Nuetzel, & Garza, 2022). Fortunately, when loci cross-amplify, alleles are often species-specific
462 and can be dealt with in downstream analyses. If alleles overlap among species, loci containing
463 these alleles should be removed prior to analysis. While it is important to address cross-
464 amplification in DNA mixture studies where multiple species contribute to DNA samples, our
465 study suggests that this issue should be relatively easy to resolve in most instances.

466 *The utility of the DNA mixtures approach is hindered by low quality DNA: some potential*
467 *solutions and future research directions*

468 Our mock DNA mixtures demonstrated the feasibility of accurately resolving NOC from mixed-
469 DNA samples when DNA quality is high. However, resolving NOC in more degraded samples
470 from the feeding trial proved difficult. Mean estimates of NOC were 2-3 (true NOC = 3) for up

471 to 48 - 72 hours post-ingestion. However, the variance in estimates, even in the early part of the
472 trial, was generally high. These results indicate that the DNA mixtures approach we used can
473 identify whether more than one individual was consumed by a predator. However, the accuracy
474 of individual NOC estimates is likely to be low, limiting the practical resolution of the current
475 approach. One potential way to increase accuracy could be to conduct multiple DNA extractions
476 and/or PCR replicates and use the mean of the replicates as the NOC estimate (Alberdi et al.,
477 2019; Mata et al., 2019). Our subsampling experiment also showed that the mean NOC estimate
478 among replicates tended to be accurate regardless of the number of loci genotyped. However,
479 this does not address the fundamental problem of reduced performance of the microhaplotype
480 panel on degraded samples.

481 The percentage of on-target reads was already ~30% lower six hours post-ingestion than in tissue
482 samples (~80% on-target in tissue samples vs ~50% on-target 6 hours into feeding trial). This
483 value continued to descend over time, reaching ~10% at 72 hours and functionally zero after that.
484 Interestingly, the trend in proportion of on-target reads across the feeding trial was very similar
485 to the number of mtDNA metabarcoding reads across the same timespan (Dick et al. in review).
486 Our data strongly suggest that DNA degradation as diet items become more digested is leading
487 to lower proportions of on-target reads, which prevents accurate microhaplotype genotyping.
488 One major advantage of the DNA mixtures approach, in theory, is that it should be robust to
489 variation in DNA quantity. However, our data indicates that poor performance of the nuclear
490 microhaplotype panel in degraded samples largely negated the advantage of the DNA mixtures
491 approach compared to methods amplifying more abundant mtDNA such as qPCR and
492 metabarcoding.

493 Some potential ways to improve in the future include (1) additional replication such as extraction
494 and PCR replicates, which was discussed above, (2) additional sequencing coverage, which
495 could potentially improve genotyping accuracy even when the percentage of on-target reads is
496 low, and (3) laboratory protocols that enhance the performance of microhaplotype panels with
497 degraded samples. Increasing sequencing coverage could improve results, especially in terms of
498 confidently identifying rare alleles. Previous research has shown that increasing sequence depth
499 increases the number of taxa recovered for eDNA samples (Shirazi, Meyer, & Shapiro, 2021),
500 which is similar in concept to identifying rare alleles. However, while increased depth could
501 improve results, it is likely that this is not a problem that researchers can "sequence their way out
502 of" given the extremely poor performance of highly degraded samples. Instead, we suggest that
503 future studies focus efforts on improving laboratory protocols for extracting and amplifying
504 degraded DNA and incorporate best practices (Deagle, Eveson, & Jarman, 2006; Rohland,
505 Glocke, Aximu-Petri, & Meyer, 2018)

506 Two previous studies that have used the GT-seq approach to genotype DNA from fecal samples,
507 although both studies targeted the predator not the prey (Burgess, Irvine, & Russello, 2022;
508 Eriksson, Ruprecht, & Levi, 2020). Burgess et al. (2022) used a similar protocol to ours with two
509 modifications: (1) primer pools were divided to reduce the number of primers in each multiplex
510 and (2) DNA was quantified and normalized after PCR1 rather than normalized with SequalPrep
511 plates after barcoding (PCR2). Eriksson et al. (2020) conducted an additional bead cleanup after
512 PCR1 that both our study and Burgess et al. (2022) did not. Eriksson et al. (2020) also quantified
513 and normalized after PCR1 rather than using normalization plates. It is important to note that the
514 protocol we used was designed for high-throughput analysis of thousands of fish tissue samples
515 for genetic stock identification and was not optimized for degraded DNA.

516 At minimum, we suggest that future studies conduct an additional bead cleanup and quantify and
517 normalize after PCR1 rather than use normalization plates after PCR2, which are designed to
518 reduce high concentrations of DNA to a uniform concentration but are not effective if DNA
519 concentrations are already below the expected input threshold (250 ng/well). Additionally, we
520 suggest that future studies use quantification results to pool samples of similar quantity and
521 therefore likely similar quality, together to reduce high variation in read counts across samples.
522 Finally, we suggest conducting iterative rounds of library preparation and sequencing to obtain
523 usable data from as many samples as possible. In our experience, samples perform better in
524 smaller batches, and this time-consuming iterative approach of analyzing small batches of poor-
525 quality samples may be the most feasible way to improve results barring sequencing for
526 sequencing degraded samples. Our suggestions focus on analysis of degraded but high-quantity
527 DNA samples from diet studies, but they may also be useful for eDNA studies with water
528 samples, where DNA quantity is potentially more of an issue than quality (Harrison, Sunday, &
529 Rogers, 2019). Quantifying the performance of different amplicon sequencing approaches with
530 highly degraded and low-quantity DNA using controlled dilution and DNA shearing experiments
531 would help advance the application of the DNA mixtures method for both eDNA and molecular
532 diet studies.

533 *Conclusion*

534 As discussed at length in Andres et al. (2021) and Sethi et al. (2019), the DNA mixtures method
535 could be leveraged to address a multitude of important topics related to conservation,
536 management, and ecology of wild populations. However, developing methods to reliably
537 estimate NOC in mixed-DNA samples with variable qualities and quantities has been difficult.

538 Our study demonstrated that accurate NOC estimates for samples containing up to 10 individuals
539 can be obtained using a panel of ~100 microhaplotype loci genotyped with GT-seq chemistry,
540 and that this approach is more effective for accurately identifying rare alleles compared to
541 microsatellites. However, analysis of highly degraded samples from a feeding trial produced
542 relatively poor results due to a low percentage of on-target reads. We suggest that future studies
543 focus on improving laboratory protocols for GT-seq analysis with highly degraded and low-
544 quality samples. Substantial methodological improvements have made it feasible to implement
545 the DNA mixtures method to for non-model organisms in ecological studies. However, some
546 final technical barriers still exist. We expect that future studies will successfully address these
547 barriers, facilitating the widespread use of the DNA mixtures method to address important
548 questions in conservation and ecology.

549 **Acknowledgements**

550 This study was supported by a grant from California Metropolitan Water District (project
551 number L04719). All animal experiments were conducted under Humboldt State University
552 IACUC # 2021F5A. We thank Katie D'Amelio from NOAA Alaska Fisheries Science Center for
553 assistance with laboratory work. We thank Fred Feyrer and Justin Clause from the U.S.
554 Geological Survey for collection of wild fish and Nann Fangue, Dennis Cochran, and Sarah
555 Baird from UC Davis for assistance with animal husbandry. We thank Anthony Clemento from
556 NOAA Southwest Fisheries Science Center for providing primer pool aliquots and reference files
557 of the microhaplotype panel and help with MICROHAPLOT. Any use of trade, firm or product
558 names is for descriptive purposes only and does not imply endorsement by the US Government.

559 **References**

560 Alberdi, A., Aizpurua, O., Bohmann, K., Gopalakrishnan, S., Lynggaard, C., Nielsen, M., &
561 Gilbert, M. T. P. (2019). Promises and pitfalls of using high-throughput sequencing for diet
562 analysis. *Molecular Ecology Resources*, *19*(2), 327–348. http://doi.org/10.1111/1755-
563 0998.12960
564 Andres, K. J., Sethi, S. A., Lodge, D. M., & Andrés, J. (2021). Nuclear eDNA estimates
565 population allele frequencies and abundance in experimental mesocosms and field samples.
566 *Molecular Ecology*, *30*, 685–697. http://doi.org/10.1111/mec.15765
567 Baetscher, D. S., Clemento, A. J., Ng, T. C., Anderson, E. C., & Garza, J. C. (2018).
568 Microhaplotypes provide increased power from short-read DNA sequences for relationship
569 inference. *Molecular Ecology Resources*, *18*(2), 296–305. http://doi.org/10.1111/1755-
570 0998.12737
571 Baetscher, D. S., Nuetzel, H., & Garza, J. C. (2022). Highly accurate species identication of
572 eastern Pacic rockshes (Sebastes spp.) with high- throughput DNA sequencing. *Research*
573 *Square*, 1–20. http://doi.org/10.21203/rs.3.rs-1895338/v1
574 Barnes, M. A., Turner, C. R., Jerde, C. L., Renshaw, M. A., Chadderton, W. L., & Lodge, D. M.
575 (2014). Environmental Conditions Influence eDNA Persistence in Aquatic Systems.
576 *Environmental Science Technology*, *48*, 1819–1827. http://doi.org/10.1021/es404734p
577 Bieber, F. R., Buckleton, J. S., Budowle, B., Butler, J. M., & Coble, M. D. (2016). Evaluation of
578 forensic DNA mixture evidence: protocol for evaluation, interpretation, and statistical
579 calculations using the combined probability of inclusion. *BMC Genetics*, *17*, 1–15.
580 http://doi.org/10.1186/s12863-016-0429-7
581 Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimomatic: a flexible trimmer for Illumina
582 sequence data. *Bioinformatics*, *30*, 2114–2120. http://doi.org/10.1093/bioinformatics/btu170/
583 Bootsma, M. L., Gruenthal, K. M., McKinney, G. J., Simmons, L., Miller, L., Sass, G. G., &
584 Larson, W. A. (2020). A GT-seq panel for walleye (*Sander vitreus*) provides important
585 insights for efficient development and implementation of amplicon panels in non-model
586 organisms. *Molecular Ecology Resources*, *20*(6), 1706–1722. http://doi.org/10.1111/1755-
587 0998.13226
588 Brandl, S., Schreier, B., Conrad, J. L., May, B., & Baerwald, M. (2021). Enumerating Predation
589 on Chinook Salmon, Delta Smelt, and Other San Francisco Estuary Fishes Using Genetics.
590 *North American Journal of Fisheries Management*, *41*, 1053–1065.
591 http://doi.org/10.1002/nafm.10582
592 Bravington, M. V., Skaug, H. J., & Anderson, E. C. (2016). Close-Kin Mark-Recapture.
593 *Statistical Science*, *31*, 259–274. http://doi.org/10.1214/16-STS552
594 Burgess, B. T., Irvine, R. L., & Russello, M. A. (2022). A genotyping-in-thousands by
595 sequencing panel to inform invasive deer management using noninvasive fecal and hair
596 samples. *Ecology and Evolution*, *12*(6), e8993. http://doi.org/10.1002/ece3.8993
597 Campbell, N. R., Harmon, S. A., & Narum, S. R. (2015). Genotyping-in-Thousands by
598 sequencing (GT-seq): A cost effective SNP genotyping method based on custom amplicon
599 sequencing. *Molecular Ecology Resources*, 855–867. http://doi.org/10.1111/1755-
600 0998.12357
601 Carlson, S. M., & Satterthwaite, W. H. (2011). Weakened portfolio effect in a collapsed salmon
602 population complex. *Canadian Journal of Fisheries and Aquatic Sciences*, *68*, 1–11.
603 http://doi.org/10.1139/F2011-084

604 Carreon-Martinez, L. B., Wellband, K. W., Johnson, T. B., Ludsin, S. A., & Heath, D. D. (2014).
605 Novel molecular approach demonstrates that turbid river plumes reduce predation mortality
606 on larval fish. *Molecular Ecology*, *23*(21), 5366–5377. http://doi.org/10.1111/mec.12927
607 Carreon-Martinez, L., Johnson, T. B., Ludsin, S. A., & Heath, D. D. (2011). Utilization of
608 stomach content DNA to determine diet diversity in piscivorous fishes. *Journal of Fish*
609 *Biology*, *78*, 1170–1182.
610 Curran, J. M., Triggs, C. M., Buckleton, J., & Weir, B. S. (1999). Interpreting DNA mixtures in
611 structured populations. *Journal of Forensic Sciences*, *44*(5), 987–995.
612 Darling, J. A., & Blum, M. J. (2007). DNA-based methods for monitoring invasive species: a
613 review and prospectus. *Biological Invasions*, *9*, 751–765. http://doi.org/10.1007/s10530-006-
614 9079-4
615 Deagle, B. E., Eveson, J. P., & Jarman, S. N. (2006). Quantification of damage in DNA
616 recovered from highly degraded samples--a case study on DNA in faeces. *Frontiers in*
617 *Zoology*, *3*(1), 11–10. http://doi.org/10.1186/1742-9994-3-11
618 Deagle, B. E., Thomas, A. C., McInnes, J. C., Clarke, L. J., Vesterinen, E. J., Clare, E. L., et al.
619 (2018). Counting with DNAin metabarcoding studies: How should we convert sequence
620 reads to dietary data? *Molecular Ecology*, *28*(2), 391–406. http://doi.org/10.1111/mec.14734
621 Deiner, K., Bik, H. M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., et al.
622 (2017). Environmental DNA metabarcoding: Transforming how we survey animal and plant
623 communities, *26*(21), 5872–5895. http://doi.org/10.1111/mec.14350
624 Dembinski, G. M., Sobieralski, C., & Picard, C. J. (2018). Estimation of the number of
625 contributors of theoretical mixture profiles based on allele counting: Does increasing the
626 number of loci increase success rate of estimates? *Forensic Science International. Genetics*,
627 *33*, 24–32. http://doi.org/10.1016/j.fsigen.2017.11.007
628 Eriksson, C. E., Ruprecht, J., & Levi, T. (2020). More affordable and effective noninvasive
629 single nucleotide polymorphism genotyping using high-throughput amplicon sequencing.
630 *Molecular Ecology Resources*, *20*(6), 1505–1516. http://doi.org/10.1111/1755-0998.13208
631 Fonseca, V. G. (2018). Pitfalls in relative abundance estimation using eDNA metabarcoding.
632 *Molecular Ecology Resources*, *18*, 923–926. http://doi.org/10.1111/1755-0998.12902
633 Gehri, R. R., Larson, W. A., Gruenthal, K., Sard, N. M., & Shi, Y. (2021). eDNA metabarcoding
634 outperforms traditional fisheries sampling and reveals fine-scale heterogeneity in a temperate
635 freshwater lake. *Environmental DNA*, *3*(5), 912–919. http://doi.org/10.1002/edn3.197
636 Grossman, G. D. (2016). Predation on fishes in the Sacramento-San Joaquin Delta: Current
637 knowledge and future directions. *San Francisco Estuary and Watershed Science*, *14*(2), 1–
638 26. http://doi.org/10.15447/sfews.2016v14iss2art8
639 Haned, H. (2011). Forensim: An open-source initiative for the evaluation of statistical methods
640 in forensic genetics, *5*(4), 265–268. http://doi.org/10.1016/j.fsigen.2010.03.017
641 Haned, H., Pène, L., Lobry, J. R., Dufour, A. B., & Pontier, D. (2011). Estimating the number of
642 contributors to forensic DNA mixtures: Does maximum likelihood perform better than
643 maximum allele count ? *Journal of Forensic Sciences*, *56*(1), 23–28.
644 Harper, L. R., Handley, L. L., Hahn, C., Boonham, N., Rees, H. C., Gough, K. C., et al. (2018).
645 Needle in a haystack? A comparison of eDNA metabarcoding and targeted qPCR for
646 detection of the great crested newt (Triturus cristatus). *Ecology and Evolution*, *8*, 6330–
647 6341. http://doi.org/10.1002/ece3.4013

648 Harrison, J. B., Sunday, J. M., & Rogers, S. M. (2019). Predicting the fate of eDNA in the
649 environment and implications for studying biodiversity. *Proceedings of the Royal Society B:
650 Biological Sciences*, *286*(1915), 20191409. http://doi.org/10.1098/rspb.2019.1409
651 Harrison, J. G., Calder, W. J., Shuman, B., & Buerkle, C. A. (2021). The quest for absolute
652 abundance: The use of internal standards for DNA-based community ecology. *Molecular
653 Ecology*, *21*(1), 30–43. http://doi.org/10.1111/1755-0998.13247
654 Hänfling, B., Handley, L. L., Read, D. S., Hahn, C., Li, J., Nichols, P., et al. (2016).
655 Environemtnal DNA metabarcoding of lake fish communities reflects long-term data from
656 established survey methods, *25*(13), 3101–3119. http://doi.org/10.1111/mec.13660
657 Kelly, R. P., Shelton, A. O., & Gallego, R. (2019). Understanding PCR processes to draw
658 meaningful conclusions from environmental DNA studies. *Scientific Reports*, *9*.
659 http://doi.org/10.1038/s41598-019-48546-x
660 King, R. A., Read, D. S., Traugott, M., & Symondson, W. O. C. (2008). Molecular analysis of
661 predation: a review of best practice for DNA-based approaches. *Molecular Ecology*, *17*(4),
662 947–963. http://doi.org/10.1111/j.1365-294X.2007.03613.x
663 Lamb, P. D., Hunter, E., Pinnegar, J. K., Creer, S., Davies, R. G., & Taylor, M. I. (2019). How
664 quantitative is metabarcoding: A meta-analytical approach. *Molecular Ecology*, *28*, 420–
665 430. http://doi.org/10.1111/mec.14920
666 Levi, T., Allen, J. M., Bell, D., Joyce, J., Russell, J. R., Tallmon, D. A., et al. (2019).
667 Environmental DNA for the enumeration and management of Pacific salmon. *Molecular
668 Ecology Resources*, *19*(3), 597–608. http://doi.org/10.1111/1755-0998.12987
669 Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.
670 *arXiv*, *1303.3887v2*, 1–3.
671 Mata, V. A., Rebelo, H., Amorim, F., McCracken, G. F., Jarman, S., & Beja, P. (2019). How
672 much is enough? Effects of technical and biological replication on metabarcoding dietary
673 analysis. *Molecular Ecology*, *28*(2), 165–175. http://doi.org/10.1111/mec.14779
674 Michel, C., Smith, J., Demetras, N., Huff, D., & Hayes, S. (2018). Non-Native Fish Predator
675 Density and Molecular-Based Diet Estimates Suggest Differing Impacts of Predator Species
676 on Juvenile Salmon in the San Joaquin River, California. *San Francisco Estuary and
677 Watershed Science*, *16*(4), 1–19. http://doi.org/10.15447/sfews.2018v16iss4art3
678 Miller, C. R., Joyce, P., & Waits, L. P. (2002). Assessing allelic dropout and genotype reliability
679 using maximum likelihood. *Genetics*, *160*, 357–366.
680 http://doi.org/10.1093/genetics/160.1.357
681 Munsch, S. H., Greene, C. M., Mantua, N. J., & Satterthwaite, W. H. (2022). One hundred-
682 seventy years of stressors erode salmon fishery climate resilience in California's warming
683 landscape. *Global Change Biology*, *28*(7), 2183–2201. http://doi.org/10.1111/gcb.16029
684 Nathan, L. M., Simmons, M., Wegleitner, B. J., Jerde, C. L., & Mahon, A. R. (2014).
685 Quantifying Environmental DNA Signals for Aquatic Invasive Species Across Multiple
686 Detection Platforms. *Environmental Science Technology*, *48*(21), 12800–12806.
687 http://doi.org/10.1021/es5034052
688 Perez, J., Mitchell, A. A., Ducasse, N., Tamariz, J., & Caragine, T. (2011). Estimating the
689 number of contributors to two-, three-, and four-person mixtures containing DNA in high
690 template and low template amounts. *Croatian Medical Journal*, *52*(3), 314–326.
691 http://doi.org/10.3325/cmj.2011.52.314

692 Pompanon, F., Deagle, B. E., Symondson, W. O. C., Brown, D. S., Jarman, S. N., & Taberlet, P.
693 (2012). Who is eating what: diet assessment using next generation sequencing. *Molecular*
694 *Ecology*, *21*, 1931–1950. http://doi.org/10.1111/j.1365-294X.2011.05403.x
695 Rochette, N. C., Rivera-Colón, A. G., Walsh, J., Sanger, T. J., Campbell-Staton, S. C., &
696 Catchen, J. M. (2022). *On the causes, consequences, and avoidance of PCR duplicates:*
697 *towards a theory of library complexity*. *bioRxiv* (pp. 1–42).
698 Rohland, N., Glocke, I., Aximu-Petri, A., & Meyer, M. (2018). Extraction of highly degraded
699 DNA from ancient bones, teeth and sediments for high-throughput sequencing. *Nature*
700 *Protocols*, *13*(11), 2447–2461. http://doi.org/10.1038/s41596-018-0050-5
701 Rourke, M. L., Fowler, A. M., Hughes, J. M., Broadhurst, M. K., DiBattista, J. D., Fielder, S., et
702 al. (2022). Environmental DNA (eDNA) as a tool for assessing fish biomass: A review of
703 approaches and future considerations for resource surveys. *Environmental DNA*, *4*, 9–33.
704 http://doi.org/10.1002/edn3.185
705 Roy, J., Vigilant, L., Gray, M., Wright, E., Kato, R., Kabano, P., et al. (2014). Challenges in the
706 use of genetic mark-recapture to estimate the population size of Bwindi mountain gorillas
707 (Gorilla beringei beringei). *Biological Conservation*, *180*(C), 249–261.
708 http://doi.org/10.1016/j.biocon.2014.10.011
709 Sabal, M., Hayes, S., Merz, J., & Setka, J. (2019). Habitat Alterations and a Nonnative Predator,
710 the Striped Bass, Increase Native Chinook Salmon Mortality in the Central Valley,
711 California. *North American Journal of Fisheries Management*, *36*(2), 309–320.
712 http://doi.org/10.1080/02755947.2015.1121938
713 Scribner, K. T., Gust, J. R., & Fields, R. L. (1996). Isolation and characterization of novel
714 salmon microsatellite loci: cross-species amplification and population genetic applications.
715 *Canadian Journal of Fisheries and Aquatic Sciences*, *53*, 833–541.
716 http://doi.org/10.1139/cjfas-53-4-833
717 Sethi, S. A., Larson, W. A., Turnquist, K., & Isermann, D. (2019). Estimating the number of
718 contributors to DNA mixtures provides a novel tool for ecology. *Methods in Ecology and*
719 *Evolution*, *10*, 109–119. http://doi.org/10.1111/2041-210X.13079
720 Shelton, A. O., Kelly, R. P., O Donnell, J. L., Park, L., Schwenke, P., Greene, C., et al. (2019).
721 Environmental DNA provides quantitative estimates of a threatened salmon species.
722 *Biological Conservation*, *237*, 383–391. http://doi.org/10.1016/j.biocon.2019.07.003
723 Shi, Y., Hoareau, Y., Reese, E. M., & Wasser, S. K. (2021). Prey partitioning between sympatric
724 wild carnivores revealed by DNA metabarcoding: a case study on wolf (Canis lupus) and
725 coyote (Canis latrans) in northeastern Washington. *Conservation Genetics*, *22*(2), 293–305.
726 http://doi.org/10.1007/s10592-021-01337-2
727 Shirazi, S., Meyer, R. S., & Shapiro, B. (2021). Revisiting the effect of PCR replication and
728 sequencing depth on biodiversity metrics in environmental DNA metabarcoding. *Ecology*
729 *and Evolution*, *11*(22), 15766–15779. http://doi.org/10.1002/ece3.8239
730 Stoeckle, B. C., Beggel, S., Cerwenka, A. F., Motivans, E., Kuehn, R., & Geist, J. (2017). A
731 systematic approach to evaluate the influence of environmental conditions on eDNA
732 detection success in aquatic ecosystems. *PLoS ONE*, *12*, e0189119.
733 http://doi.org/10.1371/journal.pone.0189119
734 Thomas, A. C., Deagle, B. E., Eveson, J. P., Harsch, C. H., & Trites, A. W. (2015). Quantitative
735 DNA metabarcoding: improved estimates of species proportional biomass using correction
736 factors derived from control material. *Molecular Ecology Resources*, *16*(3), 714–726.
737 http://doi.org/10.1111/1755-0998.12490

738 Thompson, N. F., Anderson, E. C., Clemento, A. J., Campbell, M. R., Pearse, D. E., Hearsey, J.
739 W., et al. (2020). A complex phenotype in salmon controlled by a simple change in
740 migratory timing. *Science*, *370*, 609–613. http://doi.org/10.1126/science.aba9059
741 Weir, B. S., Triggs, C. M., Stowell, L. I., Walsh, K. A. J., & Buckleton, J. (1997). Interpreting
742 DNA Mixtures. *Journal of Forensic Sciences*, *42*, 213–222.
743 Williamson, K. S., Cordes, J. F., & May, B. (2002). Characterization of microsatellite loci in
744 chinook salmon (Oncorhynchus tshawytscha) and cross-species amplification in other
745 salmonids. *Molecular Ecology Notes*, *2*(1), 17–19. http://doi.org/10.1046/j.1471-
746 8286.2002.00129.x
747 Yates, M. C., Fraser, D. J., & Derry, A. M. (2019). Meta-analysis supports further refinement of
748 eDNA for monitoring aquatic species-specific abundance in nature. *Environmental DNA*,
749 *1*(1), 5–13. http://doi.org/10.1002/edn3.7

750 **Data Accessibility Statement**

751 Demultiplexed GT-seq data used in this study are archived in the NCBI Sequence Read Archive
752 with a BioProject ID (TBD). Consensus sequences (*.fasta*) and SNP info (*.vcf*) of the
753 microhaplotype panel with 125 markers used in the study to run MICROHAPLOT are archived
754 on DRYAD (TBD). Other intermediate data files and all bioinformatic scripts supporting this
755 article are available on the Github repository (https://github.com/melodysyue/DNAmixture).
756 Questions pertaining to data generated for this project should be directed toward the
757 corresponding author.

758 **Author Contributions**

759 Y.S., C.M.D., M.J.H., and W.A.L designed the study. C.M.D. conducted the feeding trials. K.K.
760 and D.B. carried out the molecular laboratory work. Y.S. conducted the data analyses. Y.S. and
761 W.A.L. drafted the manuscript. M.V.M. supervised the project. All authors commented on the
762 manuscript and gave final approval for publication.

763 **Figures**

764 **Figure 1** Comparison of on-target rates among Delta Chinook salmon smolt tissue samples
765 (N=1,179), mock DNA mixture samples made up of Delta Chinook salmon smolts (N=283) and
766 GI tract content samples from the feeding trial (N=277). On-target rate was calculated as the total
767 number of on-target reads divided by the total number of reads. Median on-target rates are
768 indicated with red vertical lines.

769 **Figure 2** (a) Broad haplotype frequency distribution of the curated catalog of 252 unique
770 haplotypes across 74 loci and 565 Delta Chinook salmon smolt samples after stringent filtering.
771 (b) Bias in the estimated number of contributors using genotypes of the above curated catalog
772 from various DNA mock mixture samples made up of Delta Chinook salmon smolts. Light gray
773 points are individual mock DNA mixtures, and red points and lines are mean bias ±1 *SD.* A read
774 depth ratio of 0.02 was used, below which haplotypes were removed.

775 **Figure 3** Effects of total on-target reads on the number of loci successfully genotyped in (a)
776 mock DNA mixture samples (N=282) and (b) GI tract samples from the feeding trial (N=271). A
777 total of 89 GI tract samples (gray points) with fewer than 20% of 74 loci genotyped (15 loci)
778 were removed from the downstream analyses.

779 **Figure 4** Bias in the estimated number of individuals contributing to mock DNA mixtures made
780 up from Delta Chinook salmon smolt samples (range: 2- 20 individuals per mixture) with
781 varying simulated genotyping rates (10 - 90% of 74 loci) and three read depth ratios (0.002, 0.02,
782 0.2), below which haplotype sequence reads were removed. Lower genotyping rates
783 corresponded to higher locus dropout rates.

784 **Figure 5** Changes in on-target rate (a) over time post-ingestion (up to 120 hours) in the GI tract
785 content samples (N=277) of largemouth bass (LMB) and channel catfish (CCF) at two different
786 feeding trial water temperatures (15.5˚C and 18.5˚C). We removed six samples due to their
787 extremely low on-target reads (≤ 2 reads) and additional 89 samples due to fewer than 15 loci
788 genotyped in these samples. We estimated number of contributors (NOC) in each remaining
789 sample (N=182; b). In (b), light gray points are individual samples (N=184), and red points and
790 red lines are mean estimate ±1 *SD.* A read depth ratio of 0.02 was used, below which haplotypes
791 were removed.

Supplementary Figures

Figure S1 Panel optimization based on the sum of on-target reads across 377 Chinook salmon smolt samples in the initial GT-seq testing. Microhaplotype loci were ordered by the sum of on-target reads. A total of 11 microhaplotype loci were removed (gray points) due to either over amplification (>59,000 on-target reads) or under amplification (<7,000 on-target reads). Cutoffs were chosen based on the breakpoints of the distribution. The final GT-seq panel consisted of 114 microhaplotype loci (red points; Table S1).

Figure S2 Distribution of number of Delta Chinook salmon haplotypes per locus across 74 microhaplotype loci after filtering.

Figure S3 Distribution of total number of on-target reads across 190 largemouth bass samples and 94 channel catfish samples using the microhaplotype panel designed for Delta Chinook salmon. Four samples were outliers with 1,166 - 11,544 on-target reads and were on the right side of the red vertical line (x intercept = 1,100).

Figure S4 Comparison of variance (top panel) and mean bias (bottom panel) of estimated number (NOC) of Chinook salmon smolts contributing to mock DNA mixtures (rang: 2- 20 individuals) across varying simulated genotyping rates (10% to 90% of 74 loci) and three read depth ratios (0.002, 0.02, 0.2). Lower genotyping rates corresponded to higher locus dropout rates.

Figure S5 Number of Chinook salmon smolts detected over time post-ingestion (up to 120 hours) in the GI tracts of channel catfish (CCF) and largemouth bass (LMB) at two different feeding trial water temperatures ( ). Predators were each fed three smolt specimens. We fit with an exponential decay model ($y = a(1-r)^x$) with x=hour post-ingestion and y=mean NOC estimate calculated for each treatment group (species * temperature * time point). Mean NOC estimates can be found in Table S3.

817 **Table 1.** Number of samples used in the mock DNA mixture experiment
818

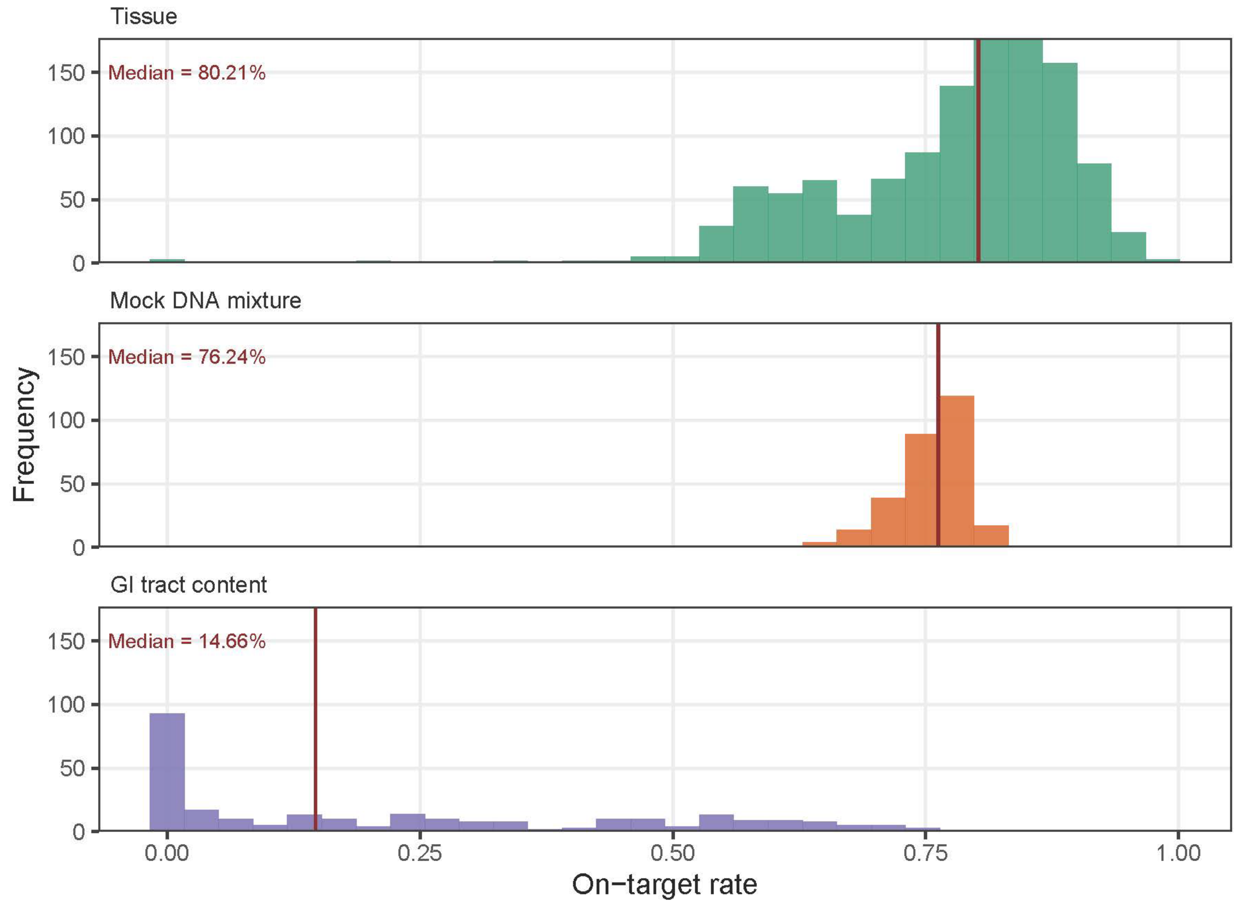
# Inds	Replicate 1	Replicate 2	Replicate 3
2	8	8	8
3	8	8	8
5	16	16	16
7	16	16	16
9	16	16	16
10	16	16	16
15	8	8	8
20	7	7	7

819
820

821 **Table 2.** Number of samples used in the feeding trial experiment
822

Species		CCF		LMB	
Temperature		15.5	18.5	15.5	18.5
# Smolt		1	3	1	3
Hours post-ingestion	6	6	5	9	9
	12	5	5	9	9
	24	5	5	9	9
	36	5	5	8	9
	48	6	6	8	10
	60	5	5	9	9
	72	6	5	8	10
	84	5	5	9	9
	96	6	5	8	8
	120	4	5	5	9

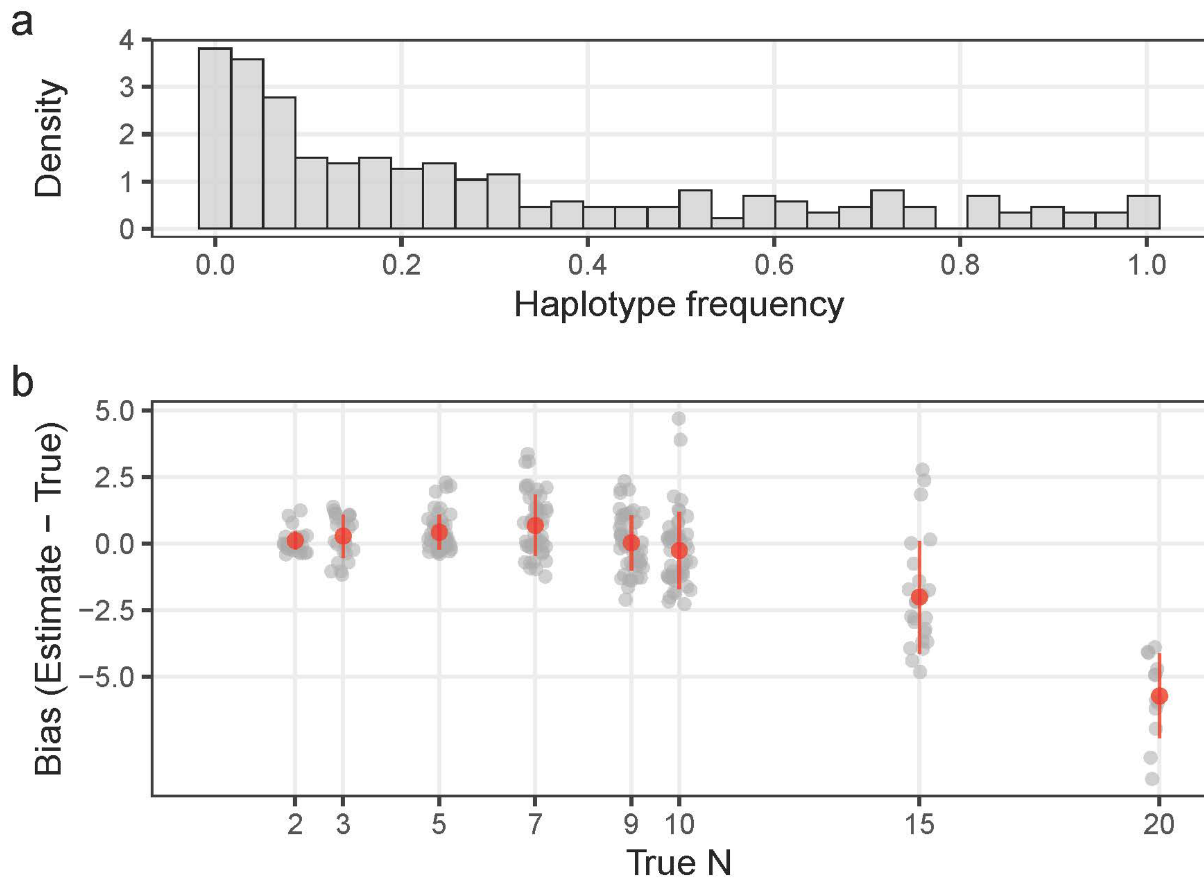
823



824

825 **Figure 1.**

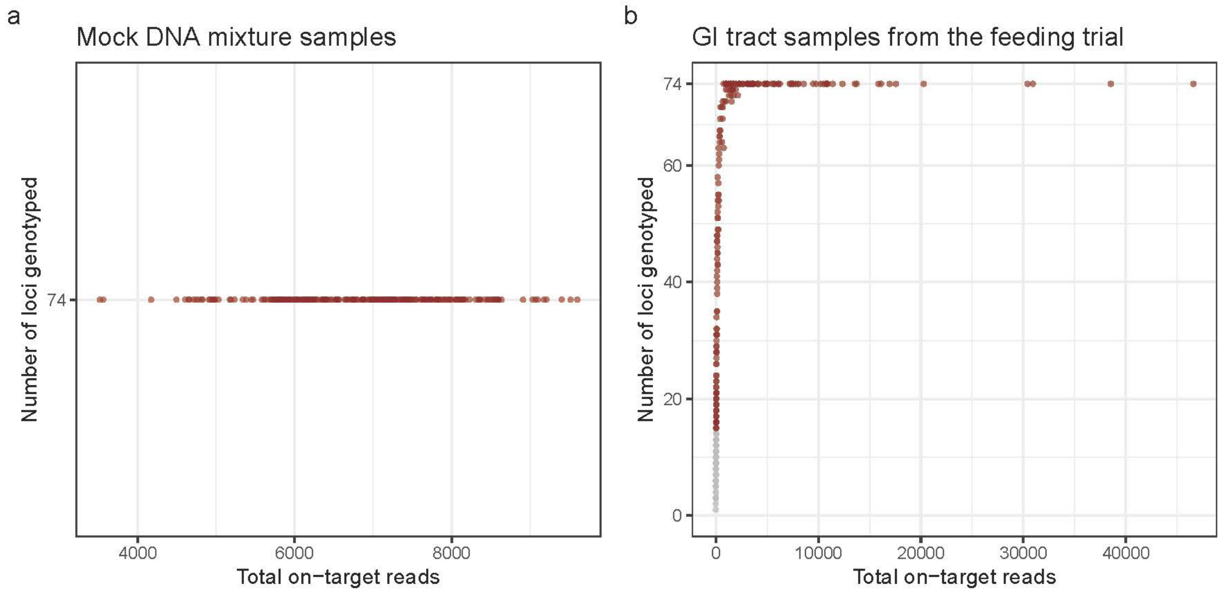
826



827

828 **Figure 2.**

829

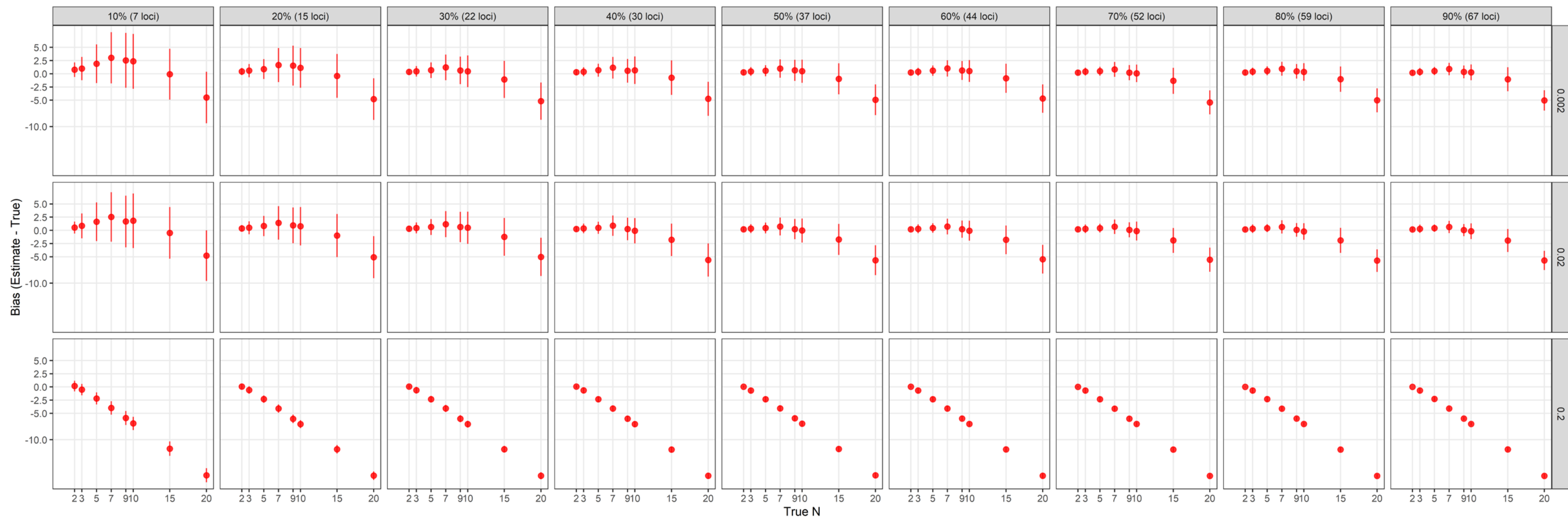


830

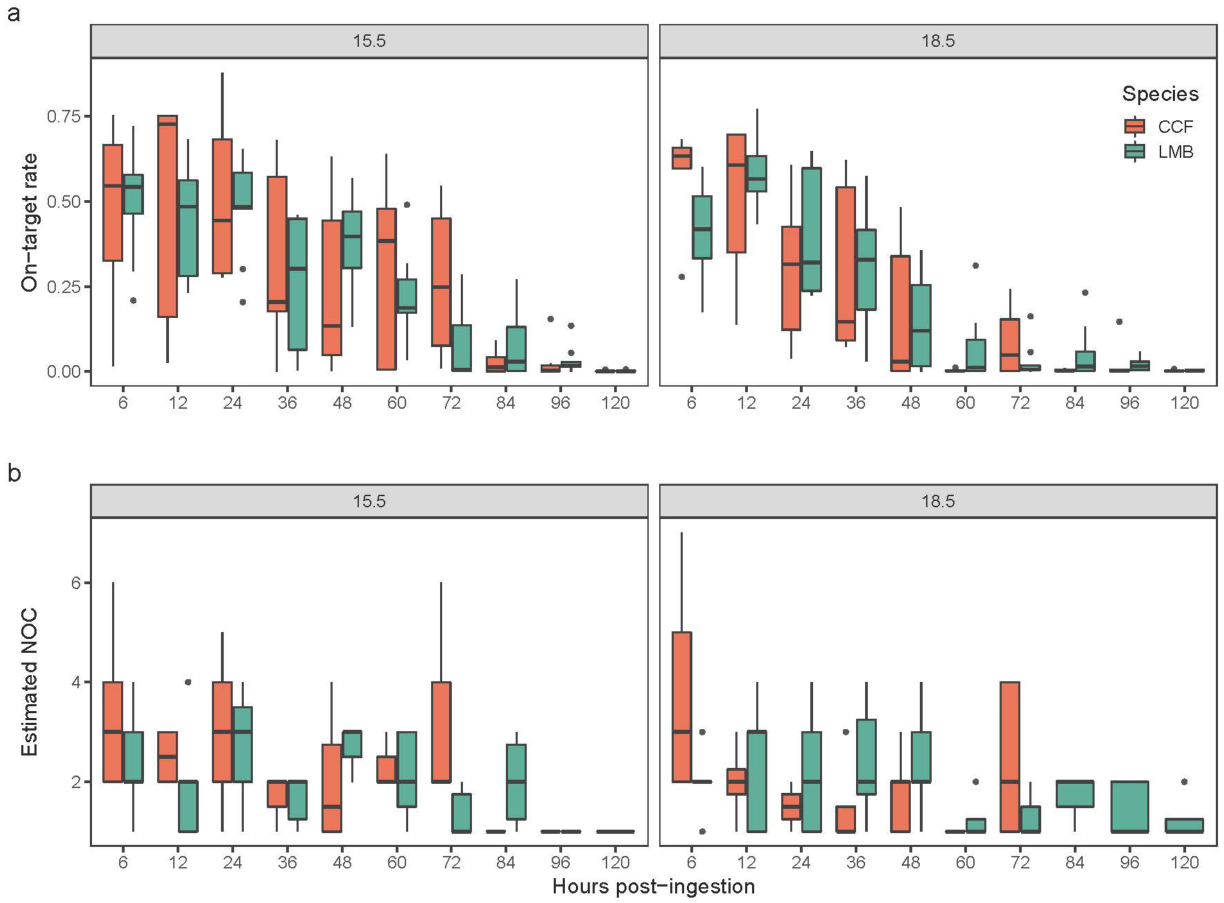
831 **Figure 3.**

832

833



834 **Figure 4.**



835

836 **Figure 5.**

837

838 **Supplemental Table 1.** List of microhaplotype loci (N=114) used in the final GT-seq panel

839

tag_id_2_206
tag_id_716
tag_id_2_40
tag_id_2_1348
tag_id_1733
tag_id_1276
tag_id_2_939
tag_id_1554
tag_id_2_1268
tag_id_2_188
tag_id_5617
tag_id_278
tag_id_4969
tag_id_1363
tag_id_2_661
tag_id_2_1114
tag_id_235
tag_id_2_2222
tag_id_2_1029
tag_id_2_978
tag_id_481
tag_id_5686
tag_id_2_911
tag_id_2_859
tag_id_744
tag_id_3920
tag_id_2_487
tag_id_2_935
tag_id_425
tag_id_423
tag_id_669
tag_id_2_502
tag_id_2_58
tag_id_2_694
tag_id_2_284
tag_id_1243
tag_id_757

tag_id_2_98
tag_id_1872
tag_id_3221
tag_id_2_136
tag_id_427
tag_id_2_1382
tag_id_773
tag_id_2_1693
tag_id_2_953
tag_id_2_2973
tag_id_2_234
tag_id_2_1586
tag_id_819
tag_id_2_1539
tag_id_1281
tag_id_2_9
tag_id_430
tag_id_650
tag_id_275
tag_id_2_1158
tag_id_1413
tag_id_2_1579
tag_id_32
tag_id_1227
tag_id_2_419
tag_id_282
tag_id_2_855
tag_id_2_1016
tag_id_5385
tag_id_2_700
tag_id_1629
tag_id_2_2787
tag_id_542
tag_id_2_414
tag_id_968
tag_id_70
tag_id_1692
tag_id_2_705
tag_id_251
tag_id_2_3577
tag_id_999

tag_id_787
tag_id_2_3026
tag_id_945
tag_id_1551
tag_id_2_3094
tag_id_2_20
tag_id_1126
tag_id_1144
tag_id_491
tag_id_1079
tag_id_2_251
tag_id_1428
tag_id_826
tag_id_381
tag_id_684
tag_id_5720
tag_id_2_311
tag_id_2_123
tag_id_2_749
tag_id_2_633
tag_id_1470
tag_id_664
tag_id_1191
tag_id_2_332
tag_id_2_3471
tag_id_186
tag_id_784
tag_id_554
tag_id_120
tag_id_1030
tag_id_2_321
tag_id_2_3452
tag_id_2_786
tag_id_2_113
tag_id_603
tag_id_384

840

841

842 **Supplemental Table 2.** Quality filtering on microhaplotype loci and samples.
843

Filtering steps	**Number of microhaplotype loci**	**Number of samples**
Prior to filtering	114	1179
Call haplotypes	114	1176
Remove haplotypes with N or X	114	1176
Remove haplotypes < 20 reads at a locus and a read depth ratio < 0.2	114	1152
Remove monomorphic loci	110	1152
Remove loci with more than 2 haplotypes in any individuals	96	1148
Remove samples genotyped in fewer than 80 of remaining loci and remove loci genotyped in fewer than 70% of samples	74	565

844
845

846 **Supplemental Table 3.** Mean NOC estimate across replicates calculated for each treatment group (species*temperature*time point).
847

Species	Temperature	Hours post-ingestion									
		6	12	24	36	48	60	72	84	96	120
CCF	15.5	3.33	2.5	3	1.67	2	2.33	3.2	1	1	NA
	18.5	3.8	2	1.5	1.5	1.8	1	2.4	NA	NA	NA
LMB	15.5	2.33	1.88	2.71	1.67	2.71	2.14	1.33	2	1	1
	18.5	2	2.44	2.2	2.38	2.43	1.25	1.29	1.67	1.43	1.25

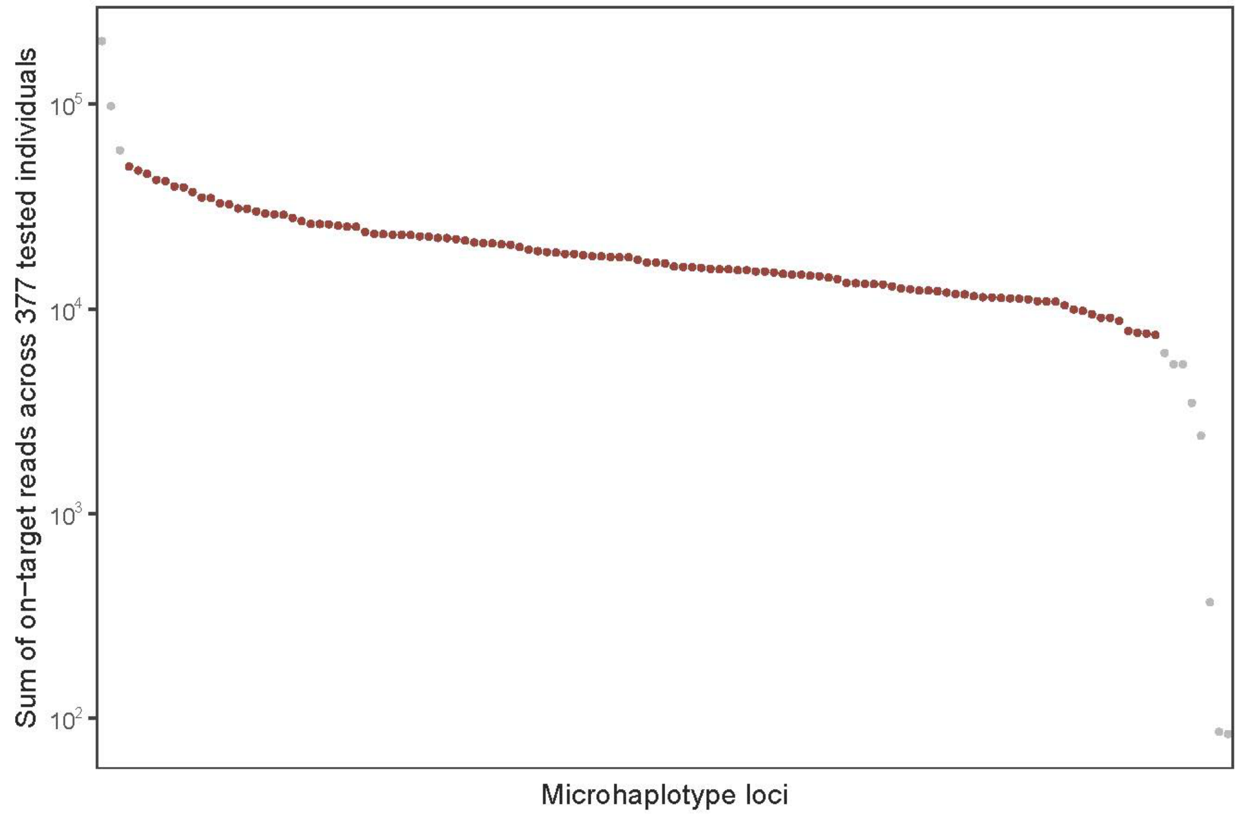
848
849 Note: 89 samples got removed because fewer than 20% of 74 loci (15 loci) were genotyped in these samples. These removed samples
850 included all or most samples at 84 - 120 hours post-ingestion in CCF.

851 **Supplemental Table 4.** Exponential decay function parameters (± standard errors) for different
852 predator species at different temperatures.
853

Species	Temperature	a (initial amount)	r (rate of decay)
CCF	**15.5**	3.19 ± 0.549	0.008 ± 0.00382
	18.5	2.9 ± 0.747	0.011 ± 0.00766
LMB	**15.5**	2.55 ± 0.345	0.006 ± 0.00253
	18.5	2.47 ± 0.243	0.006 ± 0.00183

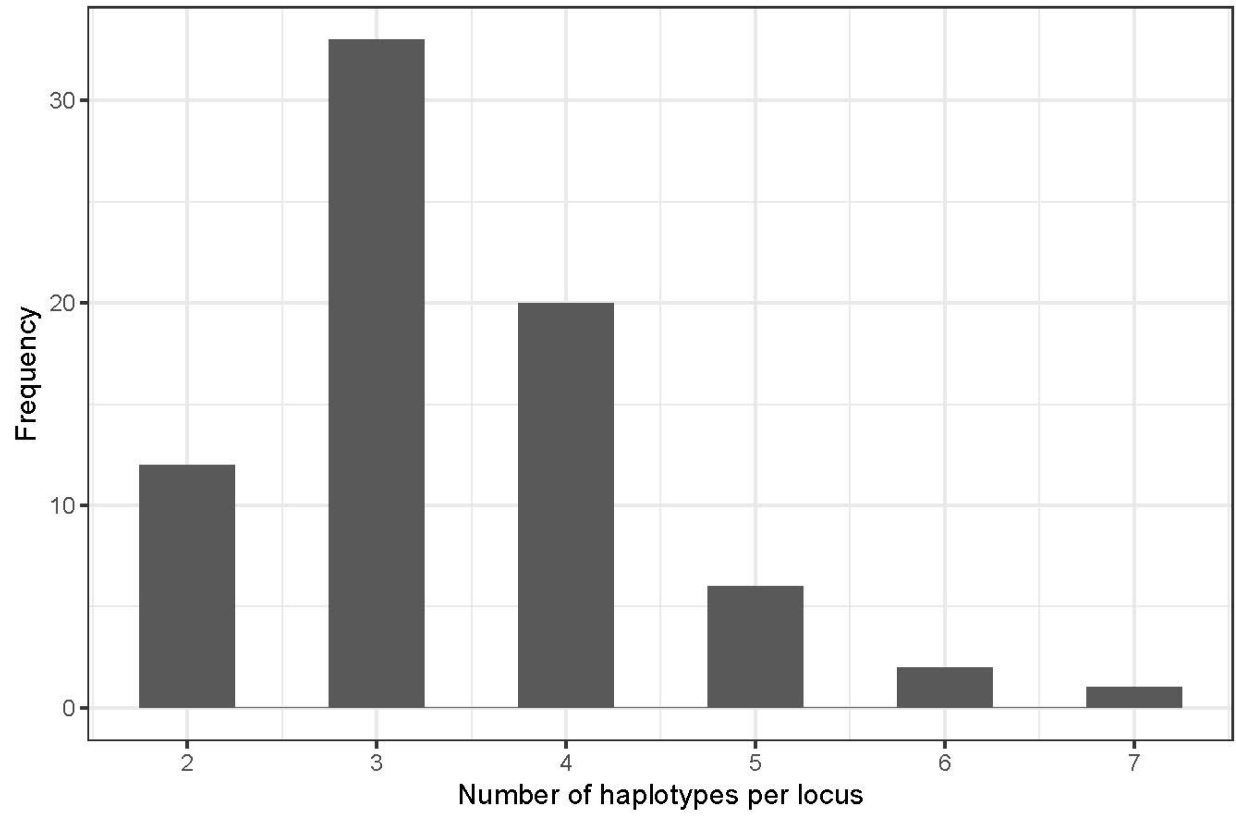
854

855



856 **Supplemental Figure 1.**

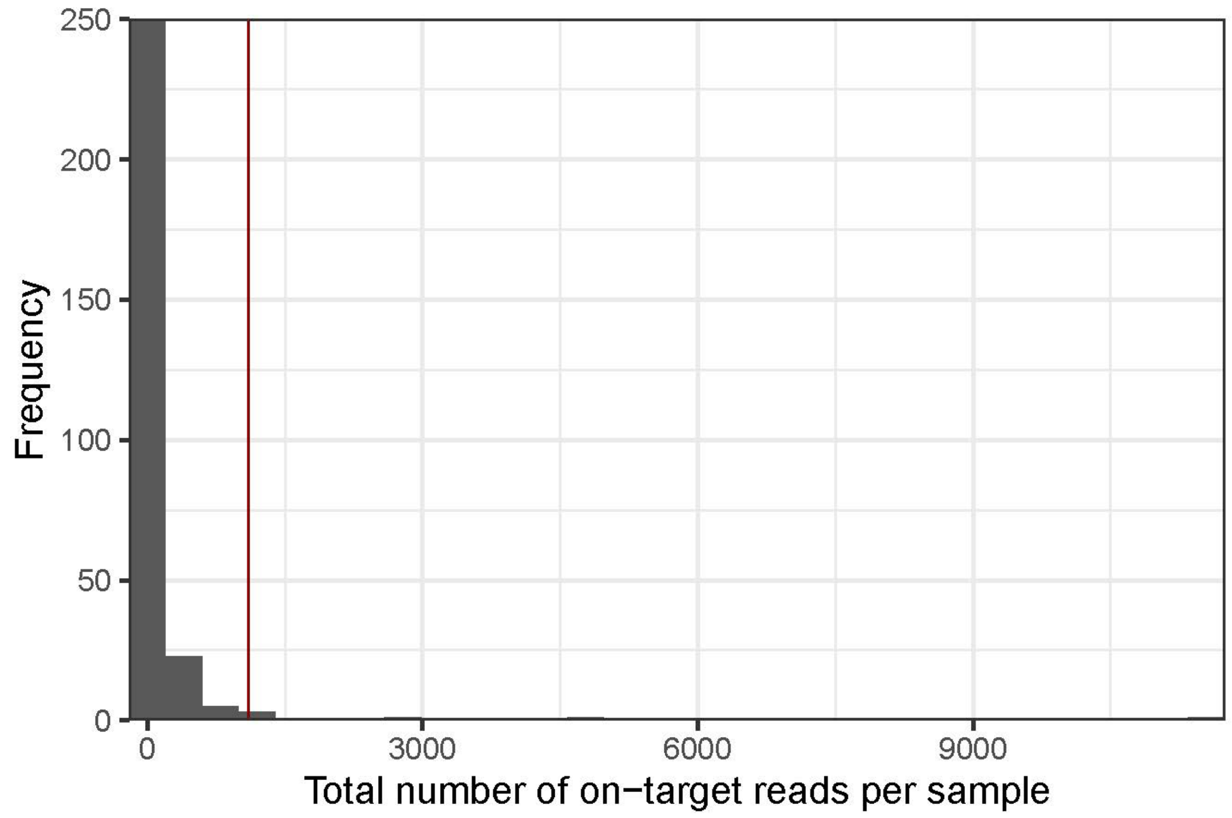
857



858

859 **Supplemental Figure 2.**

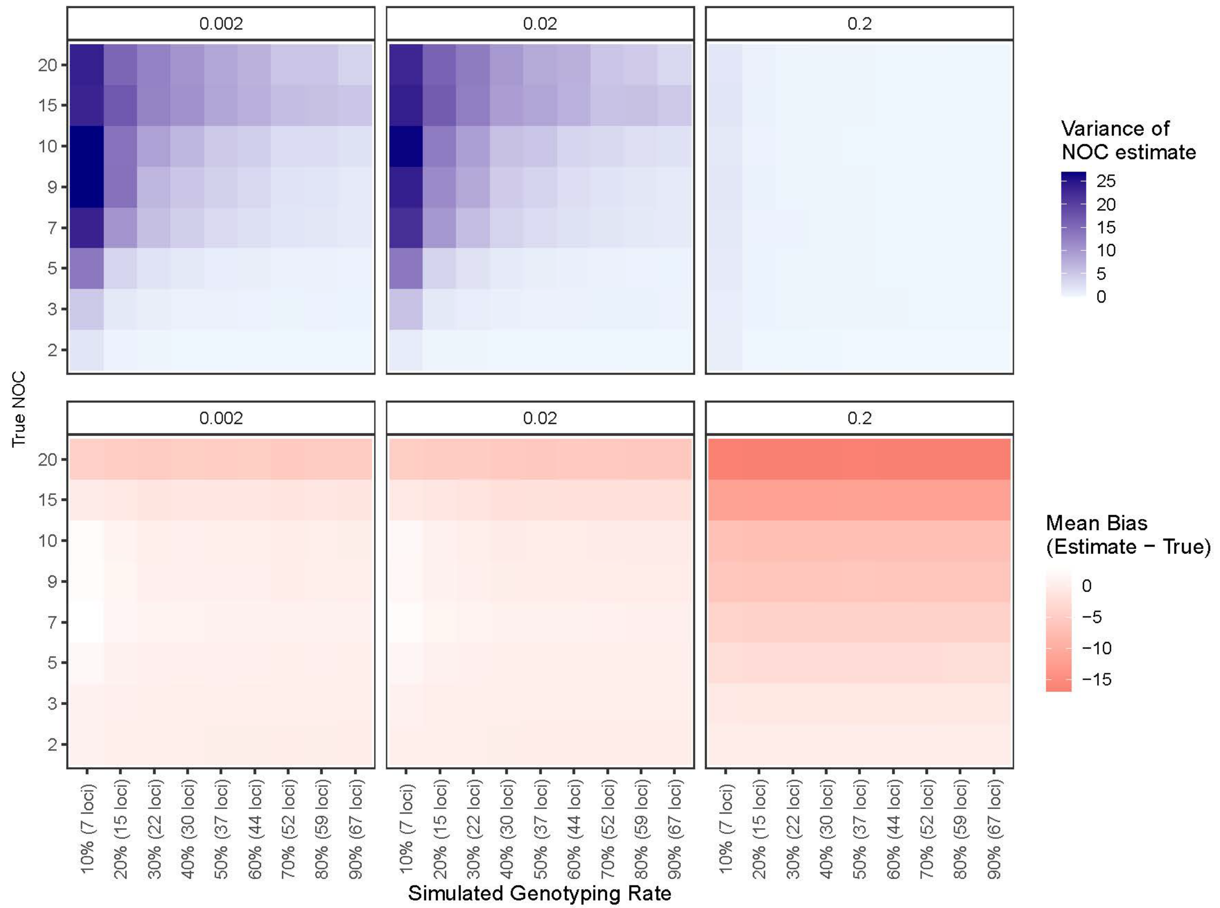
860



861

862 **Supplemental Figure 3.**

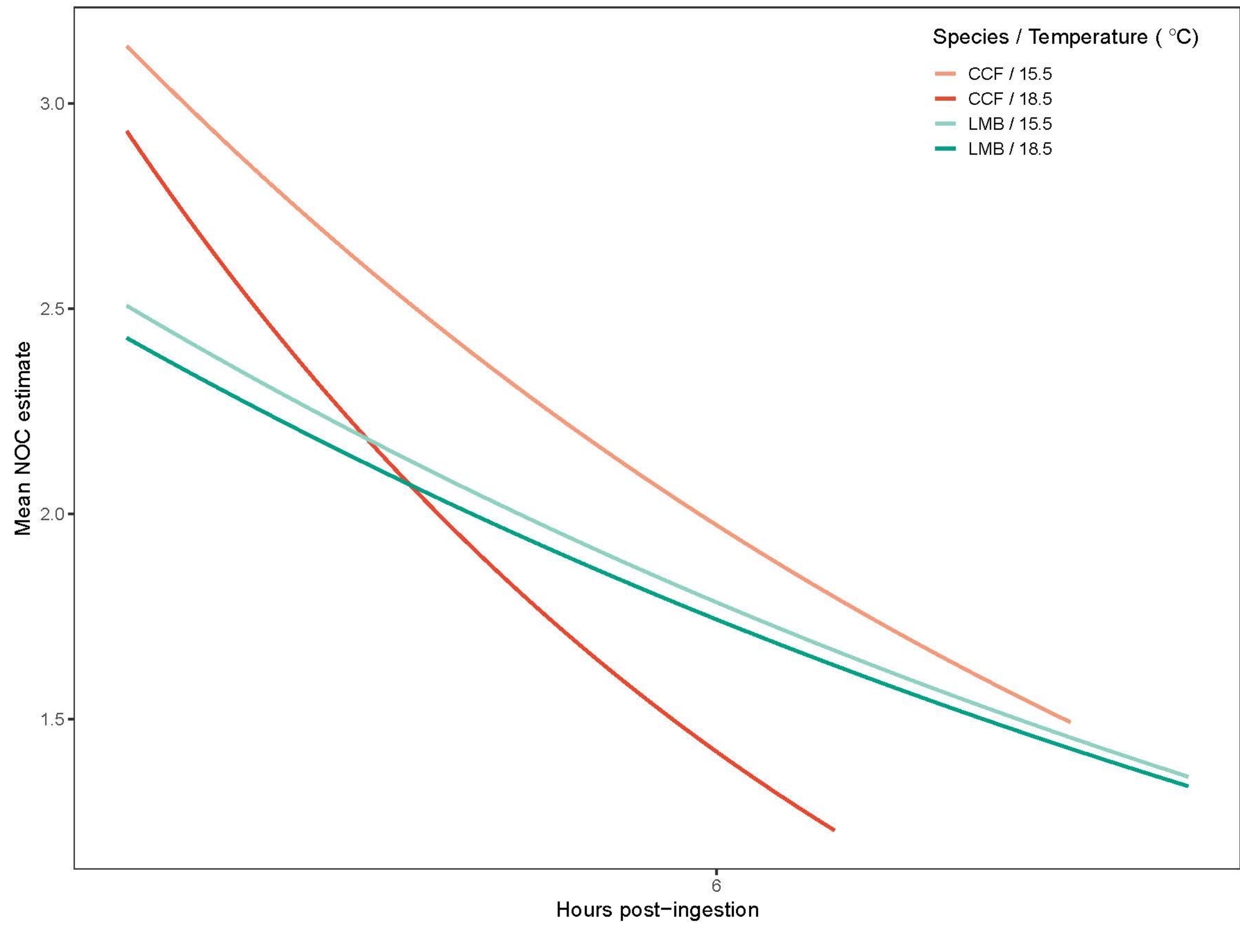
863



864

865 **Supplemental Figure 4.**

866



867

868 **Supplemental Figure 5.**