Predicting recombination frequency from map distance

Mikko Kivikoski[a,1], Pasi Rastas[b], Ari Löytynoja[b], Juha Merilä[a,c]

[a] Ecological Genetics Research Unit, Organismal and Evolutionary Biology Research Programme, Faculty of Biological and Environmental Sciences, FI-00014 University of Helsinki, Finland

[b] Institute of Biotechnology, HiLIFE, FI-00014 University of Helsinki, Finland

[c] Area of Ecology and Biodiversity, School of Biological Sciences, Kadoorie Biological Science Building, Pokfulam Road, The University of Hong Kong, Hong Kong SAR

[1] corresponding author: Mikko Kivikoski, E-mail: mikko.kivikoski@helsinki.fi

Author ORCID iDs:

M.K. **0000-0001-7105-651X**, P.R. **0000-0003-2768-1339**, A.L. **0000-0001-5389-6611**, J.M. **0000-0001-9614-0072**

Keywords: Crossover interference, *Gasterosteus aculeatus*, *Homo sapiens*, Mapping function, *Pungitius pungitius*, Stickleback

Running title: Predicting recombination frequency

Abstract

Map distance is one of the key measures in genetics and indicates the expected number of crossovers between two loci. Map distance is estimated from the observed recombination frequency using mapping functions, the most widely used of those, Haldane and Kosambi, being developed at the time when the number of markers was low and unobserved crossovers had a substantial effect on the recombination fractions. In contemporary high-density marker data, the probability of multiple crossovers between adjacent loci is negligible and different mapping functions yield the same result, that is, the recombination frequency between adjacent loci is equal to the map distance in Morgans. However, high-density linkage maps contain an interpretation problem: the map distance over a long interval is additive and its association with recombination frequency is not defined. Here, we demonstrate with high-density linkage maps from humans and stickleback fishes that the inverse of Haldane or Kosambi mapping functions fail to predict the recombination frequency from map distance, and show that this is because the expected number of crossovers is not sufficient to predict recombination frequency. We formulate a piecewise function to calculate the probability of no crossovers between the markers that yields more accurate predictions of recombination frequency from map distance. Our results demonstrate that the association between map distance and

recombination frequency is context-dependent and no universal solution exists. We anticipate that our study will motivate further research on this subject to yield a more accurate mathematical description of map distance in the context of modern data.

Introduction

Crossovers in meiosis break the physical linkage among loci and allow formation of recombinant chromosomes and ensure chromosome segregation in meiosis I (Koehler et al. 1996, Hassold et al. 2021). Although crossovers and the resultant recombinations have been studied for more than a hundred years (Sturtevant 1913), many related questions remain unanswered. Due to their profound importance in sexual reproduction, substantial research efforts have focused on better understanding the among-organism and individual variation of crossover rate (e.g., Stapley et al. 2017, Haenel et al. 2018), and on the other hand, recent technologies have been utilized for detecting recombinations at the gamete level (Dréau 2019, Bell et al. 2020, Yang et al. 2022). Crossovers also have implications for statistical properties essential in population genetics, such as the variance of genetic relatedness (Veller et al. 2020). However, one aspect that has gained little attention in the era of high-throughput sequencing is the interpretation of genetic map distances.

Recombinant gametes or offspring can be utilized to build linkage maps that quantify the physical order and map distance (i.e., expected number of crossovers) between loci. Map distances are estimated with mapping functions that attempt to account for the non-additivity of the recombination

46 frequencies due to multiple crossovers between adjacent loci. The two most
47 widely used and referred to mapping functions are probably those of
48 Haldane (1919) and Kosambi (1944, e.g. Lynch and Walsh 1998, Hill and
49 Weir 2011, Otto and Payseur 2019). However, modern sequencing methods
50 and increasing marker density have reduced the utility of these functions in
51 linkage map reconstruction; the probability of multiple crossovers between
52 closely positioned adjacent loci is negligible and all mapping functions yield
53 essentially the same result, $r = d$, where r is recombination frequency and d
54 is map distance in Morgans (e.g., Purcell et al. 2007). As the map distances
55 are estimated only for adjacent loci, they are additive over longer intervals.
56 Consequently, recombination frequency over intermediate or long map
57 distances (e.g., 50 cM) does not follow any simple association.

58

59 Although the importance of mapping functions in linkage map construction
60 has decreased over time with increasing access to dense marker data,
61 inverse mapping functions have recently been utilized to predict genetic
62 shuffling in meiosis (Veller et al. 2019) and from that the variance in
63 genetic relatedness (Veller et al. 2020). It is technically trivial to translate
64 map distance to recombination frequency with an inverse of mapping
65 function, but as all mapping functions effectively yield the same map for
66 high-density data, it is not clear which inverse mapping function, if any,

would yield the true recombination frequencies. Veller et al. (2020) observed that empirical variance in genetic relatedness among human (*Homo sapiens*) subjects did not match those predicted with the inverse of Kosambi function, indicating that the inverse of Kosambi function is invalid for translating map distances into recombination frequencies.

Here, we show with empirical data from humans and from nine-spined (*Pungitius pungitius*) and three-spined (*Gasterosteus aculeatus*) sticklebacks, that the inverse of Kosambi, Haldane, or linear mapping functions do not translate additive map distances correctly to recombination frequencies. To that end, we propose a new approach to translate map distances to recombination frequencies using a piecewise function based on the probability of no crossovers between the markers. We motivate this approach with empirical data from the nine-spined stickleback and demonstrate its performance with empirical data from humans and nine- and three-spined sticklebacks.

Materials and methods

Number of crossovers and recombination frequency

An odd number of gametic crossovers between two loci cause recombination, and the recombination frequency of two loci is equal to the probability of an odd number of gametic crossovers between them. Assuming that there is no chromatid interference, any positive number of crossovers in the bivalent leads to equal proportions of recombinant and non-recombinant gametes, while the absence of crossovers between two loci always leads to non-recombinant gametes. From this follows that recombination frequency r, can be expressed as a function of probability of no crossovers between two loci in the bivalent p_0, so that $r = ½(1- p_0)$ (Mather 1938, Weeks et al. 2009). To translate map distance into recombination frequency, the association between map distance and p_0 is necessary.

Map distance of two loci is the expected number of gametic crossovers between them. However, the distance does not tell the variation around the expectation or the likelihood for no bivalent crossovers (p_0) in that particular interval. Depending on the number of crossovers in the bivalent and their localization, the same map distance can be associated with different values

105 of p_0 and recombination frequency (Fig. 1). To address this ambiguity, we
106 formulate a piecewise function $p_0(k)$ that gives the probability of no
107 crossovers between the two loci in the bivalent with k crossovers.
108
109
110 *Derivation of $p_0(k)$*
111
112 The impetus for the model comes from chromosome 8 (LG8) of the nine-
113 spined stickleback (*Pungitius pungitius*). LG8 is a metacentric chromosome
114 (Kivikoski et al. 2021), where maternal map lengths of the two chromosome
115 arms are 52 and 60 cM, corresponding on average to 1.04 and 1.20
116 crossovers per bivalent, respectively (Table S1). Out of 934 gametes, 265
117 have two maternal crossovers and 94% of those gametes have one crossover
118 in both chromosome arms. Because these are gametic crossovers, the
119 number of crossovers in the bivalent could have been two or more.
120 However, the inferred likelihoods for two, three, or four crossovers in the
121 bivalent are 0.85, 0.12, and 0.03, respectively, and likelihoods for other
122 crossover counts are zero (see the section 'Inference of bivalent crossover
123 rate' below). Assuming no chromatid interference, the proportion of the
124 gametes with two crossovers descending from meiosis with the same
125 number of crossovers is 0.79 (see supplementary methods). Given this high
126 proportion, those crossovers are, for the sake of example, interpreted to

largely represent bivalent crossovers in meiosis with two crossovers. The map positions of the two crossovers are uncorrelated (r=0.029, p=0.64) and they are distributed approximately uniformly along the map distance (Fig. 2).

Putting this observation in a general context, we assume that when there are k crossovers in the bivalent they occur in k distinct regions, and within those regions the exact localization is independent of other crossovers in the bivalent. Let d be the map length of the whole chromosome. If there are $k(k > 0)$ crossovers in the bivalent, they are assumed to occur in k non-overlapping regions of equal size, $\frac{d}{k}$, such that every crossover occurs in a different chromosomal region. Within each region, their localizations are uniformly distributed and independent across the regions. For a given marker pair, $p_0 = \sum_{k=0}^{n} P(\text{no crossovers}|\ k)\ P(k\ \text{crossovers}) = \sum_{k=0}^{n} p_k\, p_0(k)$, where n is the highest possible number of crossovers in the bivalent, p_k is the probability of k crossovers in the bivalent and $p_0(k)$ is the probability of no crossovers between the markers when there are k crossovers in the bivalent. For markers at map positions m_i and m_j ($m_j \geq m_i$), $p_0(k)$ is defined as:

148 $$p_0(k) = \begin{cases} 1,\ if\ k = 0 \\ 1 - \frac{|m_j - m_i|}{\frac{d}{k}},\ if\ \left\lceil \frac{m_j}{\frac{d}{k}} \right\rceil = \left\lceil \frac{m_i}{\frac{d}{k}} \right\rceil and\ k > 0 \\ \left(1 - \frac{|b_i - m_i|}{\frac{d}{k}}\right) \frac{|b_j - m_j|}{\frac{d}{k}}, if\ \left\lceil \frac{m_j}{\frac{d}{k}} \right\rceil - \left\lceil \frac{m_i}{\frac{d}{k}} \right\rceil = 1\ and\ k > 0 \\ 0, if\ \left\lceil \frac{m_j}{\frac{d}{k}} \right\rceil - \left\lceil \frac{m_i}{\frac{d}{k}} \right\rceil > 1\ and\ k > 0 \end{cases} \text{(Eq. 1)},$$

149 where $\left\lceil \frac{m_i}{\frac{d}{k}} \right\rceil$ and $\left\lceil \frac{m_j}{\frac{d}{k}} \right\rceil$ are the crossover regions of the markers m_i and m_j,

150 respectively, and b_i and b_j are the upper boundaries for these regions,

151 respectively, so that $b_i = \left\lceil \frac{m_i}{\frac{d}{k}} \right\rceil \frac{d}{k}$ and $b_j = \left\lceil \frac{m_j}{\frac{d}{k}} \right\rceil \frac{d}{k}$. Notation $\lceil \quad \rceil$ refers to

152 ceiling function, and $\left\lceil \frac{m_j}{\frac{d}{k}} \right\rceil$ is the least integer greater than, or equal to $\frac{m_j}{\frac{d}{k}}$.

153

154 This function is also applicable for multiple loci. In the case of three

155 markers (m_i, m_j, m_k), the recombination frequencies derived with this

156 function meet the criteria $r_{ij} + r_{jk} \geq r_{ik}$, where r_{ij}, r_{jk}, and r_{ik} are the

157 recombination frequencies between markers m_i and m_j, m_j and m_k, and m_i

158 and m_k, respectively (Karlin and Liberman 1978, Weeks 1994; see also

159 Supplementary methods).

160

161 Our approach is built on three premises: (1) Absence of chromatid

162 interference; (2) even spacing of crossovers due to crossover interference;

163 and (3) the fact that the number of bivalent crossovers in meiosis I varies

164 across individuals, chromosomes, sexes, and species. The second and the
165 third premises have been demonstrated several times in the literature and are
166 merely facts. Crossover interference causes even distribution of crossovers
167 by their physical distances (micrometers) (Zhang et al. 2014, Zickler and
168 Kleckner 2015) and approximately by base pairs, as the chromatin packing
169 ratio per nucleus is roughly constant (see Fig. S2 in Veller et al. 2019).
170 However, the spatial distribution per map distance cannot be concluded
171 directly. The number of crossovers per chromosome is not constant but
172 varies between chromosomes, nuclei, individuals, sexes, and species (e.g.,
173 Stapley et al. 2017) and hence affects the spatial distribution of crossovers
174 (Charles 1938, Zhang et al. 2014). While the absence of chromatid
175 interference is a common assumption in the literature (e.g., Weinstein 1936,
176 Zhao and Speed 1996, Sandor et al. 2012), this has been empirically tested
177 and confirmed only in a few organisms, including humans and yeast (Zhao
178 et al. 1995, Mancera et al. 2008, Hou et al. 2014, Wang et al. 2019).
179
180 In contrast to existing mapping functions, the proposed model is not based
181 on the renewal process (Zhao and Speed 1996) and it does not model
182 crossover interference parametrically. Instead, crossover interference is
183 implemented structurally by assuming that crossovers occur in close
184 proximity less often than would be expected by chance in the absence of
185 crossover interference.

We applied the function for sex-specific recombination data in the 21 chromosomes of nine- and three-spined sticklebacks and in the 22 human autosomes. The total map lengths were derived from the linkage maps and the likelihoods for different numbers of crossovers in the bivalent were inferred from the observed number of crossovers (see 'Inference of bivalent crossover frequency' below). Implementation of the $p_0(k)$ function in R is provided in the 'Data archiving' section.

Stickleback linkage maps

The crossover frequencies and locations were estimated from the linkage maps described in Kivikoski et al. (2021). For the nine-spined stickleback, high-density linkage maps (22,468 markers informative to conclude crossover) were reconstructed with Lep-MAP3 software (Rastas 2017) from a data set of 133 parents and 938 F_1 offspring. The parental fish, 46 females and 87 males, were wild-caught individuals from the Baltic Sea coast of Finland (Helsinki, 60°13'N, 25°11'E) that were artificially crossed in laboratory to produce the aforementioned F_1 offspring (Kivikoski et al. 2021, see also Rastas et al. 2016). Five females were each crossed with a

207 different male, forming five full-sib families, and the other 41 females were
208 each crossed with two different males, which formed 41 half-sib families.
209
210 Identification of the single nucleotide polymorphisms (SNPs) of the parental
211 and the F_1 fish were based on whole-genome sequencing of the parents (5–
212 10X coverage; Illumina Hiseq platforms, BGI Hong Kong) and DarTseq
213 (Diversity Arrays Technology, Pty Ltd) genotyping of the F_1 fish (Kivikoski
214 et al. 2021). The read mapping was conducted with BWA-mem (ver. 0.7.15,
215 Li 2013) and the variants were called with SAMtools mpileup (ver 1.9, Li et
216 al. 2009) following the Lep-MAP3 software pipeline (Rastas 2017). The
217 linkage maps were built with Lep-MAP3, and the number of paternal and
218 maternal crossovers were inferred from the observed changes in the
219 haplotype phase of the F_1 offspring. Crossovers could not be inferred for
220 four crosses with only single offspring, and the final dataset included 934
221 offspring in total.
222
223 The three-spined stickleback linkage maps were based on previously
224 published sequencing data (Pritchard et al. 2017). In short, ninety wild-
225 caught parental fish (Baltic Sea, Helsinki, Finland, 60°13'N, 25°11'E) were
226 crossed such that the males (n=30) were each crossed with two different
227 females (n=60). This yielded 60 and 30 full-sib and half-sib families,
228 respectively, with 517 F_1 offspring. Genotyping of the parental and the F_1

229 fish were based on genotype-by-sequencing, according to the Restriction-
230 site Associated DNA (RAD) sequencing protocol of Elshire et al. (2011).
231 The crossing, rearing protocols and sequencing data are explained in more
232 detail in Leder et al. (2015) and Pritchard et al. (2017). For this study, the
233 RAD reads from Pritchard et al. (2017) were mapped to three-spined
234 stickleback reference genome (v4, Peichel et al. 2017) and the variants were
235 called following the Lep-MAP3 pipeline as for the nine-spined sticklebacks.
236 This yielded 28,187 informative markers to build linkage maps with Lep-
237 MAP3.
238
239 For all linkage maps (maternal and paternal maps of the nine- and three-
240 spined stickleback), the genetic distances between the adjacent markers
241 were calculated from recombination frequency with the Haldane mapping
242 function. The distances are additive for non-adjacent markers. There were,
243 on average, 1,070 and 1,342 markers per chromosome in the nine- and
244 three-spined stickleback maps, respectively. Hence, the inter-marker
245 distances were short: on average 19,571 bp corresponding to 0.054 cM and
246 0.106 cM in the paternal and maternal maps of the nine-spined stickleback,
247 respectively, and 15,461 bp corresponding to 0.043 cM and 0.075 cM in the
248 paternal and maternal maps of the three-spined stickleback, respectively. As
249 all conventional functions yield very similar results for small recombination

250 frequencies, the choice of the mapping function has a minor impact on the
251 map distances.
252
253
254 *Analysis of human recombination data*
255
256 To evaluate the general applicability of the new function, we analyzed
257 human data from Halldorsson et al. (2019). This consisted of sex-specific
258 linkage maps (their supplementary Data S1 and S2) and the crossover data
259 (their supplementary Data S4). Crossover locations and counts were
260 obtained from the column 'medsnp' of the sex-specific linkage maps for the
261 41,092 probands with both paternal and maternal crossover information. All
262 crossovers were used irrespective of their status regarding the gene
263 conversions (complex, non-complex, or not assessed; see Halldorsson et al.
264 2019). Moreover, no probands were discarded based on the total number of
265 crossovers in them; the highest number of crossovers per proband per
266 chromosome was 17 maternal crossovers in chromosome 13. The number of
267 markers in the linkage maps ranged from 17,894 to 90,036 depending on the
268 chromosome. We used R (ver. 4.1.1 R Core Team 2018) with seed value
269 2021 to sample 1.5% of the markers of every chromosome, which yielded
270 268–1,351 markers (i.e., 35,778–911,925 marker pairs) per chromosome.
271 For every marker pair, we estimated the sex-specific recombination

frequency by calculating the proportion of the studied probands (n=41,092) with an odd number of crossovers between the markers.

Inference of crossover frequency

Assuming there is no chromatid interference, the probability of observing k crossovers in a randomly sampled meiotic product depends on the number of crossovers in the bivalent so that $P(k) = \binom{n}{k}\left(\frac{1}{2}\right)^n$, where n is the number of crossovers in the bivalent and $0 \leq k \leq n$ (Weinstein 1936). The number of crossovers in the bivalent varies not only between the sexes and chromosomes but also between individuals and individual meioses (see Broman and Weber 2000 for an example). Therefore, the observed crossovers in the gamete pool are a sample of crossovers from meioses with a different number of crossovers in the bivalent. As the sampling function above is known, the multinomial distribution of the number of crossovers in the bivalent can be estimated. We used the expectation-maximization (EM) algorithm of Yu and Feingold (2001) to estimate the multinomial distribution for different numbers of crossovers in bivalent. The algorithm approximates the multinomial distribution that maximizes the likelihood of the data consisting of the numbers of meiotic products with $0 \ldots N$ observed crossovers in each chromosome. According to Yu and Feingold (2001), for

294 data where the highest number of observed crossovers in a single meiotic
295 product is N, it is sufficient to estimate the multinomial distribution between
296 0 and $2N - 1$. The algorithm was applied separately for maternal and
297 paternal crossovers and for each chromosome by pooling all meiotic
298 products (n=934, n=517, n=41,092 for nine- and three-spined stickleback
299 and human data, respectively, Tables S1–S6). The estimated multinomial
300 distribution includes the maximum-likelihood estimate for no crossovers in
301 the bivalent. We applied the bootstrapping test of Yu and Feingold (2001) to
302 estimate if, in case of an estimate above zero, the deviation from that is
303 statistically significant ($p < 0.05$). For the chromosomes with non-
304 significant p-values, we used a restricted multinomial distribution that
305 restricts the likelihood of no crossovers to zero for further analyses. This
306 choice was made to assume the obligate crossover as a null hypothesis.
307
308 Chromosomes with p-value below 0.05 were chr15 of the three-spined
309 stickleback (paternal meioses) and in humans chr21 and chr22 (both
310 maternal and paternal meioses) and chr3 (paternal meioses). For humans,
311 meioses with no crossovers in chromosomes 21 and 22 have been
312 previously reported in cytological studies (e.g. Wang et al. 2017, Hassold et
313 al. 2021), but we are not aware of such findings for the chr3 and further
314 verification is needed. For the sticklebacks, this is the first study testing the
315 obligate crossover hypothesis.

316

317

318 *Performance assessment of functions*

319

320 The performance of $p_0(k)$ and the three inverse mapping functions in

321 predicting recombination frequency from map distance were assessed by

322 calculating the mean absolute error of the predictions and the intra-

323 chromosomal component of genetic shuffling, $\bar{r}$ (Veller et al. 2019). The

324 equations for the inverse mapping functions were: $r = \frac{1}{2}(1 - e^{-2d})$

325 (Haldane 1919), $r = \frac{1}{2} tanh(2d)$ (Kosambi 1944), and

326 $\begin{cases} r = d, \text{ when } d \leq 0.5 \\ r = \frac{1}{2}, \text{when } d > 0.5 \end{cases}$ (linear). In all functions, r is the recombination

327 frequency and d is the map distance in Morgans.

328

329 The mean absolute error of the predicted recombination frequency was

330 calculated as $\frac{1}{n}\sum |r_e - r_p|$, where r_e and r_p are empirical and predicted

331 recombination frequencies per marker pair, respectively, and n is the total

332 number of marker pairs. The intra-chromosomal component of $\bar{r}$ is the part

333 of genetic shuffling due to crossover rate and localization. A higher number

334 of crossovers, and their even distribution increase $\bar{r}$, while a low crossover

335 rate and terminal or aggregated localization decrease it. Here, we calculated

336 the intra-chromosomal component of $\bar{r}$, by first predicting the

recombination frequency from map distance and then converting it to shuffling according to Eq. 10 of Veller et al. (2019) $\sum_{i<j} r_{ij} / \binom{\Lambda}{2}$, where r_{ij} is the rate of shuffling, i.e. recombination frequency of locus pair (*i, j*), Λ is the number of loci, and $\binom{\Lambda}{2}$ is the number of locus pairs.

342 # Results
343

344

345 *Map distance and recombination frequency do not have a fixed*

346 *association*

347

348 Mapping functions formulate a fixed association between recombination
349 frequency and the expected number of crossovers in a gamete, the map
350 distance. However, the same additive map distance can be associated with
351 different recombination frequencies depending on the crossover positions
352 and the context of the focal loci (Fig. 1). This demonstrates that inverse
353 mapping functions have limitations in predicting recombination frequency
354 from map distance. Maternal crossovers in chromosome 8 of the nine-spined
355 stickleback have a distribution similar to the idealized example (Fig. 2),
356 suggesting that a structural model of crossover localization may be
357 applicable in predicting recombination frequency from map distance.
358 Another limitation of the mapping functions (and their inverses) is that they
359 do not account for the variation in the number of bivalent crossovers.
360 Inferred probabilities for bivalent crossover counts show that the number of
361 crossovers varies between sexes, among chromosomes, and in meioses in all
362 three studied species (Fig. 3, Tables S1–S6). This further indicates that the

inverse of the Kosambi or Haldane mapping function may not be suitable for different species, sexes, or chromosomes.

Inverse Kosambi and Haldane mapping functions underestimate recombination frequency

For the three studied species, empirical recombination frequencies lie between the predictions of the inverse Kosambi and linear function (Fig. 4, Fig. S1–S3). This shows that inverse Kosambi and Haldane mapping functions systematically underestimate the recombination frequencies and the linear function overestimates them. The poor performance of the inverse Haldane function is not surprising as it assumes no crossover interference. However, the fact that the inverse Kosambi function, which does implement crossover interference, systematically underestimates recombination frequencies implies that its crossover interference model does not reflect the underlying biology.

The new $p_0(k)$ function outperforms the existing functions

384 We evaluated the new function against existing functions by calculating the
385 mean absolute error of predictions and the intra-chromosomal component of
386 genetic shuffling for the two sexes of three different species. The
387 predictions made with the new $p_0(k)$ functions depict the pattern of
388 empirical data (Fig. 4, Fig. S4–S6), and the mean absolute error of those
389 predictions are clearly lowest in five out of the six cases. In nine-spined
390 stickleback males, most meioses have one crossover and the linear function
391 gives a slightly lower error, than $p_0(k)$ (Table 1). The excellent overall
392 performance (Table S7–S9) demonstrates that the new function works for
393 different species and on different types of chromosomes. Consistent with the
394 mean absolute error, the new function is superior when assessed on the
395 intra-chromosomal components of $\bar{r}$ estimated from the empirical and
396 predicted recombination frequencies. For each species and for both sexes,
397 the predictions by the $p_0(k)$ function gave results closest to the empirical
398 results (Table 2), demonstrating its potential for the application.
399
400 Altogether, our analyses show that the inverses of the Kosambi and the
401 Haldane mapping functions systematically underestimate the recombination
402 frequencies. Although the linear function works for chromosomes with an
403 overall crossover rate close to 1 (map length ca. 50 cM), the $p_0(k)$ function
404 does not underestimate the recombination frequency to the same extent as

the other two functions. Overall, the new $p_0(k)$ function gives qualitatively and quantitatively the best results.

Discussion

Intrinsic limitations of mapping functions

Map distance tells the expected number of crossovers between two loci and can be used as a proxy of recombination; the longer the map distance the higher the recombination frequency. In principle, translating a map distance to a recombination frequency is trivial, one only needs an inverse of a mapping function such as Kosambi (1944) or Haldane (1919). However, the problem with this approach is that in modern high-density linkage maps, map distances of non-adjacent loci are additive and an inverse mapping function is not guaranteed to give the correct recombination frequency.

The intrinsic limitations of mapping functions, namely that they describe the interference only at a general level and do not account for the variation in the strength of crossover interference or the number of crossovers, have been previously recognized (Crow 1990, Zhao and Speed 1996, Otto and Payseur 2019). However, empirical tests on the performance of inverse mapping functions with modern data are scarce.

428 *Inverse Kosambi and Haldane mapping functions err with high-*
429 *density data*
430
431 Here, we demonstrated that additive map distance can yield an array of
432 different recombination frequencies (Fig. 1) and showed with empirical data
433 from humans and nine- and the three-spined sticklebacks that the inverse of
434 Kosambi and Haldane mapping functions underestimate the recombination
435 frequencies (Fig. 4). We also formulated a new function that outperforms
436 those mapping functions in this task and yields lower error (Table 1) and a
437 more accurate estimate of genetic shuffling (Table 2). The fact that mapping
438 functions fail to predict recombination frequencies implies that those models
439 of crossover interference does not predict crossover localization correctly.
440
441 Similar findings were reported by Veller et al. (2020), who showed that the
442 per-chromosome variances of genetic relatedness estimated with the inverse
443 of Kosambi function were higher than those from cytological data that
444 should approximate "true" variance. Our results regarding both sexes of all
445 three studied species were concordant: inverse Kosambi and Haldane
446 functions underestimate recombination frequencies and genetic shuffling
447 ($\overline{r}$), causing overestimation of variance in genetic relatedness, which
448 decreases as a function of $\overline{r}$ (Veller et al. 2020). Underestimation of
449 recombination frequencies means that the Kosambi map function predicts

450 too many crossovers (excessively long map distance) for observed
451 recombination frequency. The Haldane map function yields even longer
452 map distance, as it assumes no crossover interference.
453
454 Stickleback males have primarily one crossover per bivalent (Fig. 3, Table
455 S2, S4) and in those chromosomes the linear function had the best
456 performance (Table S7–S9). The linear function predicts recombination
457 frequency equal to map distance, whereas the $p_0(k)$ gives the map distance
458 per the total map length of the chromosome, which yields an underestimate
459 in certain chromosomes. This indicates that the linear function is adequate
460 and should be preferred for chromosomes that mainly have one bivalent
461 crossover, especially in species where this is a norm, such as in Lepidoptera
462 (Davey et al. 2017).
463
464 Accurate estimates of recombination frequency are needed in predicting
465 genetic shuffling and from that the variance in genetic relatedness (Visscher
466 et al. 2006). On the other hand, map distance per base pair (cM/Mbp) is
467 used as a measure of recombination in comparative genomics, especially in
468 non-model organisms (Stapley et al. 2017, Martin et al. 2019). This metric
469 is easy to obtain from linkage maps, but as shown here, meaningful
470 interpretation of map distance requires knowledge of the recombination

process and interpretation with an inverse mapping function can lead to biased estimates.

Limitations of the approach

In contrast to existing methods, the approach presented here builds on the fact that the crossover rate varies among organisms and meioses, which leads to differences in spatial distributions of crossover sites, and this variation should be implemented in the model. However, this approach requires knowledge about the probability of different numbers of crossovers in the bivalent, which is not needed by other methods and that cannot be concluded directly from the total map length. These can be inferred from gametic crossovers obtained from linkage maps (as done here) or analysis of haploid offspring (Liu et al. 2015). Alternatively, cytological methods can be used to obtain the bivalent crossover counts directly (Froenicke et al. 2002, Wang et al. 2017). Especially with gametic crossovers that contain the sampling variance caused by the fact that only two chromatids are involved in one crossover event, the sample size must be sufficient to obtain reasonably accurate probability distributions. Implicitly, gametic crossovers were, as a null expectation, assumed to present an unbiased sample of those in the bivalent and that meiotic drive or selection for crossover count has not

occurred. However, effects of natural selection could also be involved, for example if the offspring are studied in later life-stages instead of direct investigation of gametes.

In the $p_0(k)$ function itself, the most important assumption is the uniform distribution of crossovers per map distance and the arbitrarily defined breakpoints of the "segments". These assumptions are simplifications that allow mathematically tractable formula. The distribution of crossovers per map distance can be studied from the gametic crossovers only to some extent because they present a subset of those in the bivalent. However, the likelihood for gametes with a certain number of crossovers that show all bivalent crossovers can be calculated and the inherent uncertainty can be estimated (Supplementary methods). Based on the distribution of gametic crossovers per map distance, the assumptions of the model approximate the data in many chromosomes (Fig. S7–S9). However, the non-random distribution of error per map distance shows that all assessed functions have a systematic bias (Fig. S10–S12).

The motivation for the approach presented here came from wild individuals of the nine-spined stickleback, which is not a canonical organism to study crossovers. In contrast to humans for example, the number of crossovers in meiosis varies very little, especially in males. Despite the overall differences

515 in crossover rates (map lengths) of human and sticklebacks, the presented

516 method was demonstrated to work for human data as well.

Conclusions

The most salient finding of this study is that the inverse of the Kosambi and the Haldane mapping functions systematically underestimate recombination frequencies. Another caveat of using inverse mapping functions to translate additive map distances to recombination frequencies is that they yield one prediction per map distance, which does not match empirical findings. These findings demonstrate that (intermediate) map distances must be interpreted with care and context-specifically. We also formulated a piecewise function that allows the association between map distance and recombination frequency to be ambiguous. The fact that this function outperforms existing mapping functions in this task indicates that its implementation of crossover interference is more concordant with the data when compared with functions devised earlier. However, the presented function does not replace mapping functions in building linkage maps; it only replaces their inverses in predicting recombination from map distance.

Acknowledgements

We thank the Biodata Analytics Unit of the University of Helsinki (in particular Jukka Siren) for advice in statistics. We wish to acknowledge

CSC – IT Center for Science, Finland, for access to computational resources. Our research was supported by Academy of Finland (grants: 129662, 134728 and 218343 to JM; 322681 to AL; 343656 to PR), the Doctoral Programme in Wildlife Biology Research (LUOVA; funding to MK), the Alfred Kordelin Foundation (grant 210190 to MK), National Natural Science Foundation of China/Research Grants Council (RGC) Joint Research Scheme 2021/2022 ('N_HKU763/21' to JM), and support from the Helsinki Institute for Life Sciences (HiLife; grant to JM).

Author contribution statement

MK formulated the mathematical function introduced here, executed the data analyses, and led the manuscript writing. PR made the stickleback linkage maps and was involved in the formulation of the mathematical function. AL and JM supervised the study and participated in writing the manuscript.

Conflict of Interest

The authors declare no competing interests.

Data archiving

The stickleback linkage maps and computer code for replication of the analyses of this study are available in Github https://github.com/mikkokivikoski/recombinationStudies. In the repository, the p0kFunction.r file provides an implementation of $p_0(k)$ function in R. To facilitate reproducibility of the analyses, analytical steps following the linkage map reconstruction step are integrated into an Anduril2 workflow (Cervera et al. 2019).

References

Bell, A. D., Mello, C. J., Nemesh, J., Brumbaugh, S. A., Wysoker, A., & McCarroll, S. A. (2020). Insights into variation in meiosis from 31,228 human sperm genomes. *Nature, 583*: 259–264.

Broman, K. W., & Weber, J. L. (2000). Characterization of human crossover interference. *American Journal of Human Genetics, 66*(6): 1911–1926.

Cervera, A., Rantanen, V., Ovaska, K., Laakso, M., Nuñez-Fontarnau, J., Alkodsi, A., Casado, J., Facciotto, C., Häkkinen, A., Louhimo, R., Karinen, S., Zhang, K., Lavikka, K., Lyly, L., Pal Singh, M., & Hautaniemi, S. (2019). Anduril 2: upgraded large-scale data integration framework. *Bioinformatics, 35*(19): 3815–3817.

Charles, D. R. (1938). The spatial distribution of cross-overs in X - chromosome tetrads of *Drosophila melanogaster*. *Journal of Genetics, 36*(1): 103–126.

591 Crow, J. F. (1990). Mapping functions. *Genetics, 125*(4): 669–67.

592

593 Davey, J. W., Barker, S. L., Rastas, P. M., Pinharanda, A., Martin, S. H.,

594 Durbin, R., McMillan, W. O., Merrill, R. M., & Jiggins, C. D. (2017). No

595 evidence for maintenance of a sympatric *Heliconius* species barrier by

596 chromosomal inversions. *Evolution Letters*, *1*(3), 138–154.

597

598 Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K.,

599 Buckler, E. S., & Mitchell, S. E. (2011) A robust, simple genotyping-by-

600 sequencing (GBS) approach for high diversity species. *PLoS One* 6: e19379.

601

602 Froenicke, L., Anderson, L. K., Wienberg, J., & Ashley, T. (2002). Male

603 mouse recombination maps for each autosome identified by chromosome

604 painting. *American Journal of Human Genetics, 71(6)*: 1353–1368.

605

606 Haenel, Q., Laurentino, T. G., Roesti, M., & Berner, D. (2018). Meta-

607 analysis of chromosome-scale crossover rate variation in eukaryotes and its

608 significance to evolutionary genomics. *Molecular Ecology, 27*(11): 2477–

609 2497.

610

Haldane, J. B. S. (1919). The combination of linkage values and the calculation of distances between the loci of linked factors. *Journal of Genetic*s, 8(4): 299–309.

Halldorsson, B. V., Palsson, G., Stefansson, O. A., Jonsson, H., Hardarson, M. T., Eggertsson, H. P., Gunnarsson, B., Oddsson, A., Halldorsson, G. H., Zink, F., Gudjonsson, S. A., Frigge, M. L., Thorleifsson, G., Sigurdsson, A., Stacey, S. N., Sulem, P., Masson, G., Helgason, A., Gudbjartsson, D. F., Thorsteinsdottir, U., & Stefansson, K. (2019). Characterizing mutagenic effects of recombination through a sequence-level genetic map. *Science, 363*(6425): eaau1043.

Hassold, T., Maylor-Hagen, H., Wood, A., Gruhn, J., Hoffmann, E., Broman, K. W., & Hunt, P. (2021). Failure to recombine is a common feature of human oogenesis. *American Journal of Human Genetics, 108*(1): 16–24.

Hill, W., & Weir, B. (2011). Variation in actual relationship as a consequence of Mendelian sampling and linkage. *Genetics Research, 93*(1), 47–64.

632 Karlin, S., & Liberman, U. (1978). Classifications and comparisons of
633 multilocus recombination distributions. *Proceedings of the National*
634 *Academy of Sciences U.S.A., 75*(12): 6332–6336.
635
636 Kivikoski, M., Rastas, P., Löytynoja, A., & Merilä, J. (2021). Automated
637 improvement of stickleback reference genome assemblies with Lep-Anchor
638 software. *Molecular Ecology Resources, 21*(6): 2166–2176.
639
640 Koehler, K. E., Hawley, R. S., Sherman, S., & Hassold, T. (1996).
641 Recombination and nondisjunction in humans and flies. *Human Molecular*
642 *Genetics 5*(1): 1495–1504.
643
644 Kosambi, D. D. (1944). The estimation of map distances from
645 recombination values. *Annals of Eugenics, 12*(1): 172–175.
646
647 Leder, E. H., McCairns, R. J. S., Leinonen, T., Cano, J. M., Viitaniemi, H.
648 M., Nikinmaa, M., Primmer, C. R., & Merilä, J. (2015) The evolution and
649 adaptive potential of transcriptional variation in sticklebacks – signatures of
650 selection and widespread heritability. *Molecular Biology and Evolution 32*:
651 674–689.
652

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics, 25*: 2078–2079.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at http://arxiv.org/abs/1303.3997. 1– 3.

Liu, H., Zhang, X., Huang, J., Chen, J. Q., Tian, D., Hurst, L. D., & Yang, S. (2015). Causes and consequences of crossing-over evidenced via a high-resolution recombinational landscape of the honey bee. *Genome Biology,* 16: 15.

Lynch, M., & Walsh, B. (1998). Genetics and Analysis of Quantitative Traits. Sinauer Associates, Sunderland, MA, pp 393–398.

Martin, S. H., Davey, J. W., Salazar, C., & Jiggins, C. D. (2019). Recombination rate variation shapes barriers to introgression across butterfly genomes. *PLoS Biology, 17*(2): e2006288.

Mather, K. (1938). Crossing-over. *Biological Reviews, 13*(3): 252–292.

674 Otto, S. P., & Payseur, B. A. (2019) Crossover interference: Shedding light
675 on the evolution of recombination. *Annual Reviews of Genetics,* 53: 19–44.
676
677 Peichel, C. L., Sullivan, S. T., Liachko, I., & White, M. A. (2017)
678 Improvement of the threespine stickleback genome using a Hi-C-based
679 proximity-guided assembly. *Journal of Heredity, 108*: 693–700.
680
681 Pritchard, V, L., Viitaniemi, H. M., McCairns, R. J. S., Merilä, J., Nikinmaa,
682 M., Primmer, C. R., & Leder, E. H. (2017). Regulatory architecture of gene
683 expression variation in the threespine stickleback *Gasterosteus aculeatus.*
684 *G3: Genes, Genomes, Genetics, 7*(1): 165–178.
685
686 Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R.,
687 Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M., & Sham, P.
688 C. (2007). PLINK: A tool set for whole-genome association and population-
689 based linkage analyses. *American Journal of Human Genetics, 81*(3): 559–
690 575.
691
692 Rastas, P., Calboli, F. C. F., Guo, B., Shikano, T., & Merilä, J. (2016).
693 Construction of ultradense linkage maps with Lep-MAP2: Stickleback F2
694 recombinant crosses as an example. *Genome Biology and Evolution, 8*(1):

78–93.

Rastas, P. (2017). Lep-MAP3: robust linkage mapping even for low-coverage whole genome sequencing data. *Bioinformatics, 33*(23): 3726–3732.

R Core Team (2018) R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing, Vienna, Austria).

Sandor, C., Li, W., Coppieters, W., Druet, T., Charlier, C., & Georges, M. (2012) Genetic Variants in *REC8*, *RNF212*, and *PRDM9* influence male recombination in cattle. *PLoS Genetics, 8*(7): e1002854.

Stapley, J., Feulner, P. G. D., Johnston, S. E., Santure, A. W., & Smadja, C. M. (2017). Variation in recombination frequency and distribution across eukaryotes: patterns and processes. *Philosophical Transactions of the Royal Society B* 372: 20160455.

Sturtevant, A. H. (1915). The behavior of chromosomes as studied through linkage. *Z. Abstam. Vererbung,* 13: 234–287.

Veller, C., Kleckner, N., & Nowak, M. A. (2019). A rigorous measure of genome-wide genetic shuffling that takes into account crossover positions and Mendel's second law. *Proceedings of National Academy of Science U.S.A. 116* (5) 1659–1668.

Veller, C., Edelman, N. B., Muralidhar, P., & Nowak, M. A. (2020). Variation in genetic relatedness is determined by the aggregate recombination process. *Genetics, 216*(4): 985–994.

Veller, C., Wang, S., Zickler, D., Zhang, L. & Kleckner, N. (2022). Limitations of gamete sequencing for crossover analysis. *Nature* 606: E1–E3.

Visscher, P. M., Medland, S. E., Ferreira, M. A. R., Morley, K. I., Zhu, G., Cornes, B. K., Montgomery, G. W., & Martin, N. G. (2006). Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS Genetics* 2(3): e41.

Wang, S., Hassold, T., Hunt, P., White, M. A., Zickler, D., Kleckner, N., & Zhang, L. (2017). Inefficient crossover maturation underlies elevated aneuploidy in human female meiosis. *Cell, 168*(6): 977–989.

Weeks, D. E. (1994) Invalidity of the Rao map function for three loci. *Human Heredity,* 44: 178–180.

Weinstein, A. (1936). The theory of multiple-strand crossing over. *Genetics, 21*(3): 155–199.

Yang, L., Gao, Y., Li, M. Park, K.-E., Liu, S., Kang, X., Liu, M., Oswalt, A., Fang, L., Telugu, B. P., Sattler, C. G., Li, C., Cole, J. B., Seroussi, E., Xu, L., Yang, L., Zhou, Y., Li, L., Zhang, H., Rosen, B. D., Van Tassell, C. P., Ma, L., & Liu, G. E. (2022). Genome-wide recombination map construction from single sperm sequencing in cattle. *BMC Genomics 23*: 181.

Yu, K., & Feingold, E. (2001). Estimating the frequency distribution of crossovers during meiosis from recombination data. *Biometrics, 57*(2), 427–434.

Zhang, L., Liang, Z., Hutchinson, J., & Kleckner, N. (2014) Crossover patterning by the Beam-Film model: Analysis and implications. *PLoS Genetics, 10*(1): e1004042.

Zhao, H., Speed, T. P. (1996). On genetic map functions. *Genetics, 142*(4): 1369–1377.

Zickler, D., & Kleckner, N. (2015). Recombination, pairing, and synapsis of homologs during meiosis. *Cold Spring Harbor perspectives in biology*, *7*(6): a016626.

Figures

A

B

	A-B, B-C, C-D, D-E	A-C, C-E	A-D, B-E	A-E	B-D
p0	0.5	0	0	0	0.25
p1	0.5	1	0.5	0	0.5
p2	0	0	0.5	1	0.25
Expected no. crossovers	0.5	1	1.5	2	1
Recombination frequency	0.25	0.5	0.5	0.5	0.375
Additive map distance (cM)	25	50	75	100	50
Haldane map distance (cM)	34.7	Inf	Inf	Inf	69.3
Kosambi map distance (cM)	27.5	Inf	Inf	Inf	48.7

C

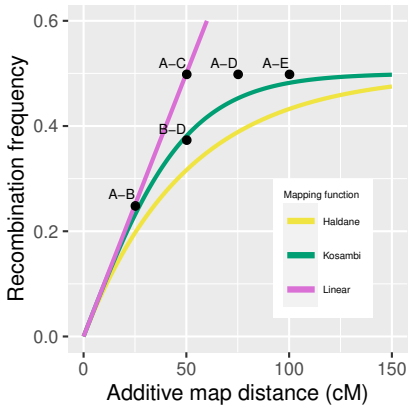

Fig. 1 Number and location of crossovers affect the recombination frequency. (A) An example of a chromosome with always two crossovers *in the bivalent* so that one occurs in the yellow and the other one in the blue area. The locations of the crossovers within their distinct regions are independent. (B) The table shows the probability for 0, 1, or 2 crossovers between the markers (p_0, p_1, and p_2, respectively) and the expected number

773 of crossovers *in the bivalent*, the resulting recombination frequency of the
774 marker pair (*in the gamete*) and the estimated map distance for every marker
775 pair using different mapping functions. (C) Graph showing the relationship
776 between the recombination frequency and map distance with different
777 mapping functions. Notably, intervals A-C, A-D, and A-E have the same
778 recombination frequencies but different map distances, whereas marker
779 pairs A-C and B-D have the same map distances but different recombination
780 frequencies.
781
782

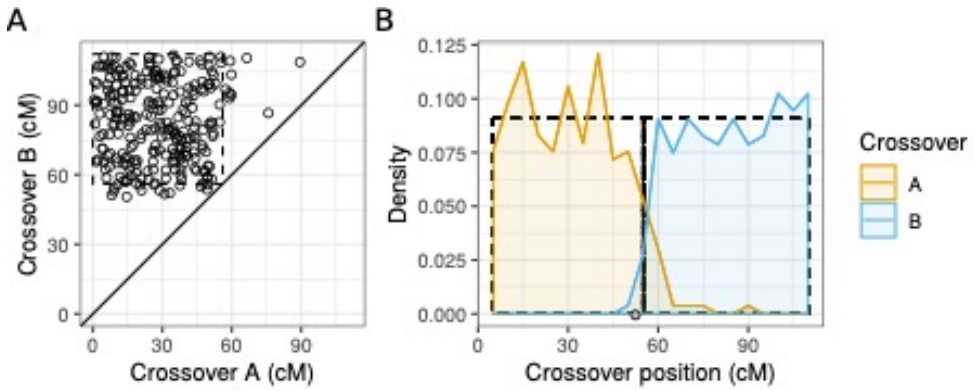

Fig. 2 Observed spatial distribution of maternal crossovers in gametes with two maternal crossovers. (A) Each dot corresponds to one offspring (n=265) and coordinates show the map positions of the two crossovers, 'A' referring to the crossover closer to chromosome start and 'B' to the one closer to chromosome end (r=0.029, p=0.64). (B) Density plots show the distribution of the two crossovers. Densities are calculated in 5 cM windows. In both panels, the dashed-line rectangles show the expected distribution of crossovers expected under the model presented here.

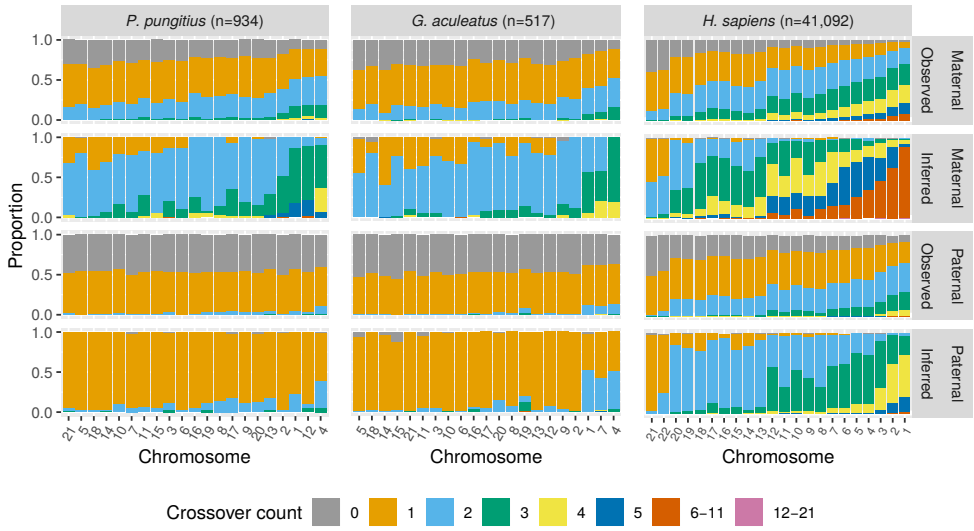

Fig. 3 Observed (gametic) and inferred (bivalent) crossover frequency distributions in the nine-spined stickleback (*P. pungitius*), three-spined stickleback (*G. aculeatus*), and human (*H. sapiens*). Each bar shows the proportion of offspring with a certain number of crossovers (Observed) and the inferred proportions of meiosis with a certain number of crossovers in the bivalent in maternal and paternal meioses (Inferred). The inference is based on the expectation-maximization algorithm by Yu and Feingold (2001). Chromosomes are ordered from shortest to longest by length in base pairs. Crossover counts 6–11 and 12–21 are grouped for readability.

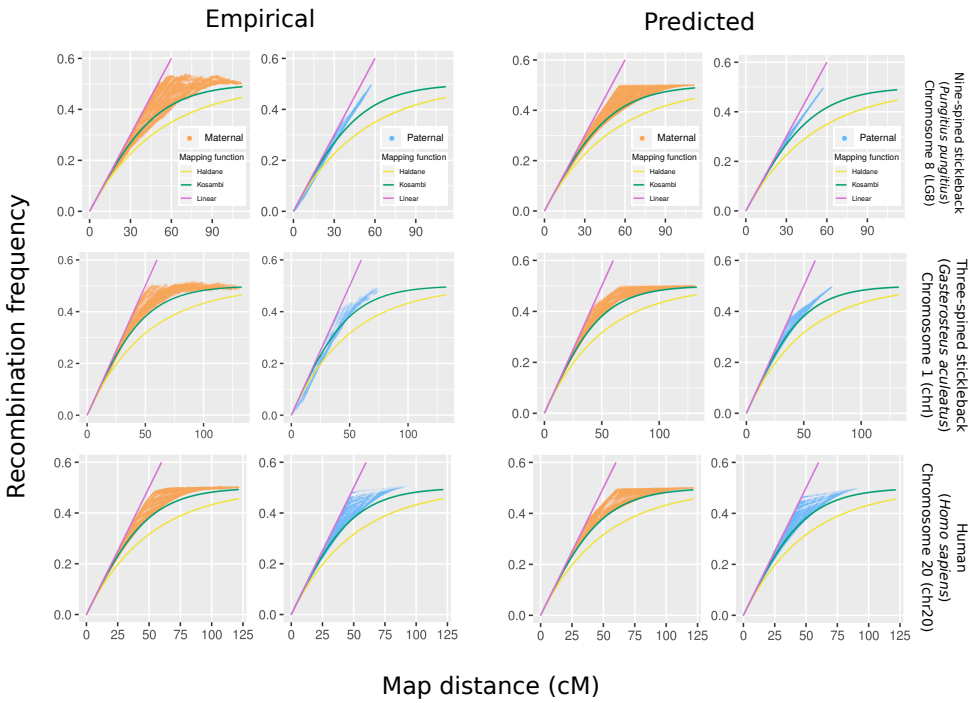

Fig. 4 Examples of empirical recombination frequencies (left panels) and predictions by the new function (right) in the nine-spined stickleback (top), three-spined stickleback (middle), and human (bottom). Solid lines show the three inverse mapping functions. Each orange and blue dot is a marker pair in maternal and paternal data, respectively.

Tables

Table 1. Means of a per-marker pair absolute error in the predicted recombination frequencies for both sexes of the three study species**.**

Species	Sex	Haldane	Kosambi	Linear	$p_0(k)$
Human	Paternal	0.0282	0.00936	0.00853	0.00375
	Maternal	0.0257	0.00709	0.00823	0.00346
Nine-spined stickleback	Paternal	0.0104	0.00442	0.00267	0.00302
	Maternal	0.0417	0.0152	0.0154	0.00910
Three-spined stickleback	Paternal	0.00867	0.00399	0.00311	0.00232
	Maternal	0.0438	0.0167	0.0139	0.00981

Table 2. Intra-chromosomal components of genetic shuffling ($\bar{r}$) for autosomes. The five columns show $\bar{r}$ calculated from the empirical recombination frequencies and those predicted by the four functions.

Species	Sex	Haldane	Kosambi	Linear	$p_0(k)$	Empirical
Human	Paternal	0.012799	0.014872	0.016796	0.015558	0.015882
	Maternal	0.018091	0.020133	0.021721	0.020672	0.020831
Nine-spined	Paternal	0.007945	0.009261	0.01041	0.009627	0.009819
stickleback	Maternal	0.01243	0.01449	0.01646	0.01509	0.01538
Three-spined	Paternal	0.007979	0.009292	0.010406	0.009733	0.009648
stickleback	Maternal	0.011867	0.013869	0.015800	0.014449	0.014815