J. Math. Biol. (2012) 64:87–108
DOI 10.1007/s00285-011-0406-8

Mathematical Biology

Refining the relationship between homozygosity and the frequency of the most frequent allele

Shashir B. Reddy · Noah A. Rosenberg

Received: 15 July 2010 / Revised: 7 December 2010 / Published online: 9 February 2011

Abstract Recent work has established that for an arbitrary genetic locus with its number of alleles unspecified, the homozygosity of the locus confines the frequency of the most frequent allele within a narrow range, and vice versa. Here we extend beyond this limiting case by investigating the relationship between homozygosity and the frequency of the most frequent allele when the number of alleles at the locus is treated as known. Given the homozygosity of a locus with at most K alleles, we find that by taking into account the value of K, the width of the allowed range for the frequency of the most frequent allele decreases from $2/3 - \pi^2/18 \approx 0.1184$ to $1/3 - 1/(3K) - \{K/[3(K-1)]\} \sum_{k=2}^{K} 1/k^2$. We further show that properties of the relationship between homozygosity and the frequency of the most frequent allele in the unspecified-K case can be obtained from the specified-K case by taking limits as $K \to \infty$. The results contribute to a greater understanding of the mathematical properties of fundamental statistics employed in population-genetic analysis.

Keywords Allele frequency · Homozygosity · Population genetics

Mathematics Subject Classification (2000) 92D10

S. B. Reddy
Center for Computational Medicine and Bioinformatics,
University of Michigan, 100 Washtenaw Ave, Ann Arbor, MI 48109, USA

N. A. Rosenberg (✉)
Department of Human Genetics, Center for Computational Medicine
and Bioinformatics, and the Life Sciences Institute, University of Michigan,
100 Washtenaw Ave, Ann Arbor, MI 48109, USA
e-mail: rnoah@umich.edu

1 Introduction

For a variable genetic locus in a diploid population, homozygosity is the fraction of individuals in the population expected to have two identical copies at the locus under the assumption of Hardy-Weinberg proportions (Weir 1996). Consider a polymorphic locus with at most K alleles, whose allele frequencies are represented by the sequence $(p_1, \ldots, p_K)$. The sequence is arranged in descending order such that if $i < j$, then $p_i \geq p_j$. The p_i can be viewed as probabilities; for all i, $p_i \in [0, 1)$, and $\sum_{i=1}^{K} p_i = 1$. The homozygosity H of the locus is defined as the sum of the squares of allele frequencies at the locus,

$$H = \sum_{i=1}^{K} p_i^2. \tag{1}$$

Homozygosity depends primarily on the frequencies of high-frequency alleles, so that most individuals homozygous for some allele are homozygous for one of the alleles of highest frequency. Because empirical studies sometimes report limited information about individual loci, precisely determining the relationship between homozygosity H and the frequency p_1 of the most frequent allele would provide a basis for approximating one of these two quantities when only the other quantity is reported. Clarifying this relationship would also assist in understanding the properties of statistics based on H or p_1 in scenarios in which a population-genetic phenomenon influences one of the two quantities directly and only indirectly influences the other. For example, positive selection favoring a specific allele can directly inflate p_1 while indirectly inflating H; a bottleneck event in a population can lead to a loss of diversity, inflating H directly while indirectly inflating p_1. In both contexts, because of the close relationship of H and p_1, statistics based on either quantity can be suitable in measuring the phenomenon of interest.

For a locus whose number of alleles was treated as indeterminate, Rosenberg and Jakobsson (2008) examined the relationship between H and p_1, showing that given either H or p_1, the other can be determined to within an interval of mean size $2/3 - \pi^2/18$. Here, we seek to refine this relationship in the case that an upper bound K is specified for the number of alleles at the locus.

If $K = 2$, then $H = p_1^2 + p_2^2 = 2p_1^2 - 2p_1 + 1$, and from the value of H or p_1, the other quantity is uniquely determined. For $K > 2$, however, given either H or p_1, the other is only localized to a particular interval dependent on K. We show that the mean size of this interval is smaller than $2/3 - \pi^2/18$, so that if K is given, then H and p_1 localize each other more precisely than in the unspecified-K case of Rosenberg and Jakobsson (2008). We consider a variety of properties of the dependence of the relationship of H and p_1 on K, showing that for $K \to \infty$, our results agree with those of Rosenberg and Jakobsson (2008) for the case of K unspecified. We illustrate the relationships among H, p_1, and K using allele frequency data from human populations.

2 Bounds on H and M

As in Rosenberg and Jakobsson (2008), we refer to p_1 by M, and we label the half-open interval $[1/k, 1/(k-1))$ by I_k. Much of our analysis parallels that of Rosenberg and Jakobsson (2008), except with K, the maximal number of alleles, specified rather than unspecified. By taking limits of various quantities as $K \to \infty$, we can compare our results to those of Rosenberg and Jakobsson (2008) and we can verify that results for the case of K specified converge on results obtained in the unspecified-K case.

Given a fixed maximal number of alleles $K \geq 2$, our main results provide upper and lower bounds on M in terms of H (Theorem 1), and upper and lower bounds on H in terms of M (Theorem 2). These results are analogous to Theorems 1 and 2 of Rosenberg and Jakobsson (2008), respectively.

Theorem 1 *Consider a sequence of the allele frequencies at a locus, $(p_1, \ldots, p_K)$, with $K \geq 2$ fixed, such that $p_i \in [0, 1)$, $\sum_{i=1}^{K} p_i = 1$, $H = \sum_{i=1}^{K} p_i^2$, $M = p_1$, and $i < j$ implies $p_i \geq p_j$. Then given $H \in [1/K, 1)$,*

$$\frac{1}{\lceil H^{-1} \rceil}\left(1 + \sqrt{\frac{\lceil H^{-1} \rceil H - 1}{\lceil H^{-1} \rceil - 1}}\right) \leq M \leq \frac{1}{K}\left(1 + \sqrt{(KH-1)(K-1)}\right).$$

Equality of M with its lower bound occurs if and only if $p_i = M$ for $1 \leq i \leq K' - 1$, $p_{K'} = 1-(K'-1)M$, and $p_i = 0$ for $i > K'$, where $K' = \lceil H^{-1} \rceil = \lceil M^{-1} \rceil$. Equality of M with its upper bound occurs if and only if $p_1 = M$ and $p_i = (1-M)/(K-1)$ for $2 \leq i \leq K$.

Theorem 2 *Consider a sequence of the allele frequencies at a locus, $(p_1, \ldots, p_K)$, with $K \geq 2$ fixed, such that $p_i \in [0, 1)$, $\sum_{i=1}^{K} p_i = 1$, $H = \sum_{i=1}^{K} p_i^2$, $M = p_1$, and $i < j$ implies $p_i \geq p_j$. Then given $M \in [1/K, 1)$,*

$$\frac{KM^2 - 2M + 1}{K-1} \leq H \leq 1 - M(\lceil M^{-1} \rceil - 1)(2 - \lceil M^{-1} \rceil M).$$

Equality of H with its lower bound occurs if and only if $p_1 = M$ and $p_i = (1-M)/(K-1)$ for $2 \leq i \leq K$. Equality of H with its upper bound occurs if and only if $p_i = M$ for $1 \leq i \leq K'-1$, $p_{K'} = 1-(K'-1)M$, and $p_i = 0$ for $i > K'$, where $K' = \lceil H^{-1} \rceil = \lceil M^{-1} \rceil$.

2.1 A geometric argument

Before proving the theorems, we introduce a geometric perspective that can assist in understanding them. In each of the three panels of Fig. 1, representing three different loci, the area of a red box represents the square of an allele frequency, and therefore, the total area of all boxes represents homozygosity. The x-axis, which is divided into sections corresponding to separate allele frequencies, represents the constraint that the sum of allele frequencies is equal to 1. The dotted line prespecifies the maximal value

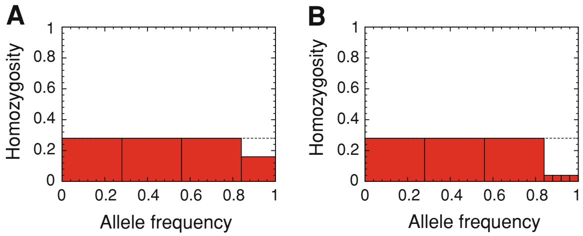

Fig. 1 A geometric argument for obtaining the upper and lower bounds on the homozygosity H, as functions of the frequency M of the most frequent allele. In each panel, vertical lines partition the x-axis into a set of allele frequencies with sum equal to 1. *Red boxes* represent the squares of allele frequencies, and each box indicates the contribution to homozygosity of an individual allele. Homozygosity is represented by the total area in red. The panels depict the case in which $M = 0.28$ and the number of alleles is $K = 7$. The *dashed line* indicates the maximal box height. (**a**) The maximal H of 0.2608, produced when $(p_1, p_2, p_3, p_4, p_5, p_6, p_7) = (0.28, 0.28, 0.28, 0.16, 0, 0, 0)$. (**b**) An intermediate H of 0.2416, produced when $(p_1, p_2, p_3, p_4, p_5, p_6, p_7) = (0.28, 0.28, 0.28, 0.04, 0.04, 0.04, 0.04)$. (**c**) The minimal H of 0.1648, produced when $(p_1, p_2, p_3, p_4, p_5, p_6, p_7) = (0.28, 0.12, 0.12, 0.12, 0.12, 0.12, 0.12)$.

of the frequency M of the most frequent allele, so that none of the boxes can exceed its height.

In Theorem 2, given M and K, we seek to partition the x-axis of Fig. 1 into at most K components so that the resulting boxes have maximal or minimal area. Comparing Figs. 1a and 1b, the two figures differ only in the partition of the interval [0.84, 1]. In Fig. 1a, this interval contains a single allele, whereas in Fig. 1b, it contains four alleles of equal frequency. The scenario in Fig. 1a has greater area, illustrating the principle that because the square of the sum $(p_1 + p_2)^2$ exceeds the sum of squares $p_1^2 + p_2^2$, greater area is produced when a larger allele is carved from a fixed interval than when the interval is divided into smaller alleles. This principle that a single box is larger than two boxes that occupy the same total length on the x-axis can be used to show that the maximal area is obtained by proceeding greedily along the x-axis, sequentially separating alleles with frequency M until the remaining length along the axis is less than M, and then choosing a single allele to fill the remaining space. Similarly, the minimum is obtained in the opposite manner, as can be seen in Fig. 1c: after choosing one allele with frequency equal to the prespecified maximum M, the minimal sum of squares is produced by subdividing the remaining part of the unit interval into as many small boxes as allowed by the prespecified number of alleles K.

2.2 Proofs of Theorems 1 and 2

To prove the theorems, we use a corollary of the Cauchy-Schwarz inequality.

Lemma 3 *Consider a sequence of length* K, $(a_1, \ldots, a_K)$ *with* $a_i \geq 0$, *such that* $\sum_{i=1}^K a_i = A$ *for some* $A \geq 0$. *Then* $\sum_{i=1}^K a_i^2 \geq A^2/K$, *with equality if and only if* $a_1 = a_2 = \cdots = a_K = A/K$.

Proof By the Cauchy-Schwarz inequality, $(\sum_{i=1}^K a_i^2)(\sum_{i=1}^K 1) \geq (\sum_{i=1}^K a_i)^2$, with equality if and only if $a_i = \lambda$ for all i and some constant λ. Because $\sum_{i=1}^K a_i = A$, equality holds if and only if $a_i = A/K$ for all i. □

In examining the bounds on M in terms of H and on H in terms of M, it is important to take note of the allowable values of H and M. We now show that given the number of alleles K, the homozygosity H and the frequency M of the most frequent allele lie in the interval $[1/K, 1)$.

Lemma 4 *At a locus with at most $K \geq 2$ alleles, $H, M \in [1/K, 1)$.*

Proof By construction, $M \in (0, 1)$. Because $M \geq p_i$ for all i, $M = \sum_{i=1}^{K} M/K \geq \sum_{i=1}^{K} p_i/K = 1/K$. As $p_i \in [0, 1)$, $p_i^2 \leq p_i$. Summing from $i = 1$ to K, and noting that $M^2 < M$ because $M \in (0, 1)$, we obtain $H < 1$. $H \geq 1/K$ follows from application of Lemma 3 to $(p_1, \ldots, p_K)$. □

We now prove Theorems 1 and 2. We define the upper and lower bound functions on M in terms of H as $UM_K, LM_K : [1/K, 1) \to [1/K, 1)$, respectively, and we define the upper and lower bound functions on H in terms of M as $UH_K, LH_K : [1/K, 1) \to [1/K, 1)$. We aim to determine these functions and to show that they match the formulas in Theorems 1 and 2. For convenience, we denote the corresponding functions in the unspecified-K case, as derived by Rosenberg and Jakobsson (2008) and previously denoted F, f, G, and g, respectively, by UM_∞, LM_∞, UH_∞, $LH_\infty : (0, 1) \to (0, 1)$.

Proof of Theorem 2 We wish to confirm that for $K \geq 2$, $UH_K(M) = 1 - M(\lceil M^{-1} \rceil - 1)(2 - \lceil M^{-1} \rceil M)$ and $LH_K(M) = (KM^2 - 2M + 1)/(K - 1)$. By Theorem 2ii of Rosenberg and Jakobsson (2008), on the interval $[1/K, 1)$, even if the number of alleles is permitted to exceed K, the upper bound on H is achieved when it is at most K. Thus, $UH_K(M) = UH_\infty(M)$ on $[1/K, 1)$. From eq. A3 of Rosenberg and Jakobsson (2008), $UH_\infty(M) = 1 - M(\lceil M^{-1} \rceil - 1)(2 - \lceil M^{-1} \rceil M)$, with the appropriate condition for $UH_\infty(M) = H$.

By definition, $LH_K(M)$ is the minimum of $M^2 + \sum_{i=2}^{K} p_i^2$ over sequences $(p_1, \ldots, p_K)$. Because $\sum_{i=2}^{K} p_i = 1 - M$, by Lemma 3, $LH_K(M) = M^2 + (1 - M)^2/(K - 1)$, with $LH_K(M) = H$ if and only if $p_i = (1 - M)/(K - 1)$ for each i with $2 \leq i \leq K$. □

The upper bound $UH_K(M)$ for H is equivalent on $[1/K, 1)$ to the corresponding function in the unspecified-K case, $UH_\infty(M)$. In the unspecified-K case, the upper bound is achieved by choosing $\lceil M^{-1} \rceil - 1$ alleles to have the highest allele frequency M, and assigning all remaining frequency to one "leftover" allele (see Lemma 3 of Rosenberg and Jakobsson (2008)). Because $\lceil M^{-1} \rceil \leq K$ on $[1/K, 1)$, in the unspecified-K case, the configuration that achieves the upper bound has K or fewer alleles. The lower bound $LH_K(M)$ is attained simply when the remaining frequency after excluding the most frequent allele is distributed equally among the $K - 1$ remaining alleles. It is easily shown that $\lim_{K\to\infty} LH_K(M) = LH_\infty(M) = M^2$.

Figure 2 superimposes $UH_K(M)$ and $LH_K(M)$ for K equal to 3, 5, and 15, and in the limiting case of unspecified K. The figure depicts the different domains for different K, and the piecewise structure for UH_K. It also illustrates that the bounds are monotonically increasing, continuous, and bijective.

Lemma 5 $UH_K, LH_K : [1/K, 1) \to [1/K, 1)$ *are monotonically increasing, continuous, and bijective.*

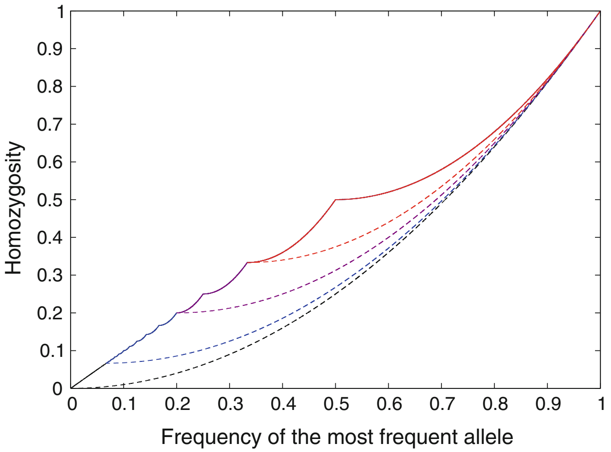

Fig. 2 Upper bounds UH_K (*solid*) and lower bounds LH_K (*dashed*) on the homozygosity H, as functions of the frequency of the most frequent allele. The bounds for loci with $K = 3$, $K = 5$, and $K = 15$ alleles are plotted in red, purple, and blue, respectively, each with domain $[1/K, 1)$. The bounds for the case of unspecified K are plotted in black, with domain $(0, 1)$. The various curves for the upper bound overlap.

Proof For UH_K, the result follows from Lemma 4i of Rosenberg and Jakobsson (2008). The function LH_K is a (continuous) quadratic function with positive leading coefficient $K/(K-1)$. The vertex of LH_K occurs at the endpoint $M = 1/K$; thus, the function LH_K is monotonically increasing over $[1/K, 1)$. Because $LH_K(1/K) = 1/K$ and $LH_K(1) = 1$, the domain and range of the function are identical. □

As a consequence of Lemma 5, UH_K and LH_K have well-defined inverse functions over $[1/K, 1)$, such that $UH_K^{-1}, LH_K^{-1} : [1/K, 1) \to [1/K, 1)$ are monotonically increasing, continuous, and bijective. To establish the upper and lower bounds of M in terms of H, we use the invertibility of UH_K and LH_K over $[1/K, 1)$.

Corollary 6 *For $K \geq 2$, the inverse functions of UH_K and LH_K are*

$$UH_K^{-1}(H) = \frac{1}{\lceil H^{-1} \rceil}\left(1 + \sqrt{\frac{\lceil H^{-1} \rceil H - 1}{\lceil H^{-1} \rceil - 1}}\right) \tag{2}$$

$$LH_K^{-1}(H) = \frac{1}{K}\left(1 + \sqrt{(KH-1)(K-1)}\right). \tag{3}$$

Proof It is easy to check that the function in eq. 3 satisfies $LH_K^{-1}(LH_K(M)) = LH_K(LH_K^{-1}(M)) = M$. For each $k \in [2, K]$, for M in I_k, the function in eq. 2 satisfies $UH_K^{-1}(UH_K(M)) = UH_K(UH_K^{-1}(M)) = M$. □

Proof of Theorem 1 It suffices to confirm that on $[1/K, 1)$, $UM_K(H)$ and $LM_K(H)$, defined as the upper and lower bounds on M in terms of H, satisfy $UM_K(H) = LH_K^{-1}(H)$ and $LM_K(H) = UH_K^{-1}(H)$.

Table 1 The functions defining the bounds on homozygosity H and the frequency of the most frequent allele M, given the maximal number of alleles K (*left*), and in the limiting case of unspecified K (*right*)

Function	Definition	Limiting function	Definition
UH_K	$1 - M(\lceil M^{-1}\rceil - 1)(2 - \lceil M^{-1}\rceil M)$	UH_∞	Same as UH_K
LH_K	$\frac{KM^2-2M+1}{K-1}$	LH_∞	M^2
UM_K	$\frac{1}{K}\left(1+\sqrt{(KH-1)(K-1)}\right)$	UM_∞	$\sqrt{H}$
LM_K	$\frac{1}{\lceil H^{-1}\rceil}\left(1+\sqrt{\frac{\lceil H^{-1}\rceil H-1}{\lceil H^{-1}\rceil-1}}\right)$	LM_∞	Same as LM_K

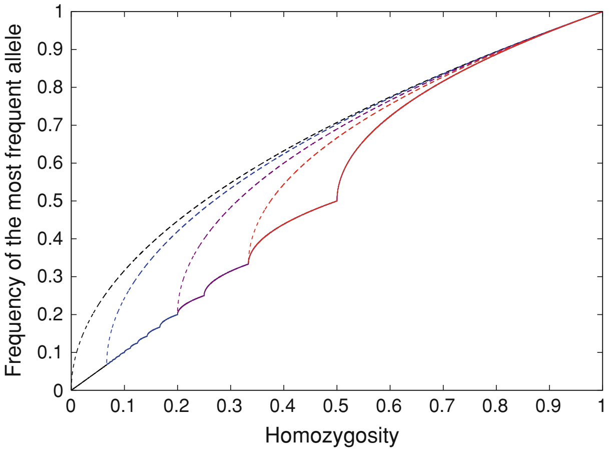

Fig. 3 Upper bounds UM_K (*dashed*) and lower bounds LM_K (*solid*) on the frequency M of the most frequent allele, as functions of homozygosity. The bounds for loci with $K = 3$, $K = 5$, and $K = 15$ alleles are plotted in red, purple, and blue, respectively, each with domain $[1/K, 1)$. The bounds for the case of unspecified K are plotted in black, with domain $(0, 1)$. The various curves for the lower bound overlap. The upper and lower bounds in this figure are inverse functions of the lower and upper bounds in Figure 2.

That $LM_K(H) = UH_K^{-1}(H)$ follows from Lemma 5 of Rosenberg and Jakobsson (2008). To obtain $UM_K(H)$, consider $H \in [1/K, 1)$. Rewriting eq. 1 as $H - \sum_{i=2}^{K} p_i^2 = M^2$, M is maximized when $\sum_{i=2}^{K} p_i^2$ is minimized. By Lemma 3, at the minimum, $H = M^2 + (1 - M)^2/(K - 1)$. Solving for M, we take the larger root so that M will be the frequency of the most frequent allele. □

The various functions appear in Table 1. Just as Theorem 2 finds that for $M \in [1/K, 1)$, $UH_K(M) = UH_\infty(M)$, Theorem 1 finds that for $H \in [1/K, 1)$, $LM_K(H) = LM_\infty(H)$. Similarly, just as $\lim_{K\to\infty} LH_K(M) = LH_\infty(M) = M^2$, it is easily shown that $\lim_{K\to\infty} UM_K(H) = UM_\infty(H) = \sqrt{H}$.

Figure 3 plots the upper and lower bounds on M in terms of H, with K equal to 3, 5, and 15, and in the limiting unspecified-K case. Comparing Figs. 2 and 3, it is visually apparent that the lower bound of M and the upper bound of H are inverse functions, as are the upper bound of M and the lower bound of H.

3 Features of the bounds on *M* in terms of *H*

We next highlight some of the features of the upper and lower bounds identified in Theorems 1 and 2. For many of our results, when applying a limit as $K \to \infty$, we obtain corresponding results from the unspecified-K case of Rosenberg and Jakobsson (2008). Although the input value $H = 1$ or $M = 1$ does not represent a polymorphic genetic locus, it is convenient to view UM_K and LM_K as producing output $M = 1$ at $H = 1$ and UH_K and LH_K as producing $H = 1$ at $M = 1$. First, we consider the upper and lower bounds on the frequency of the most frequent allele M in terms of homozygosity H, as obtained in Theorem 1.

3.1 Mean values of the bounds

Proposition 7 *For $K \geq 2$, averaging across values of $H \in [1/K, 1)$, (i) the mean of $UM_K(H)$ is $2/3 + 1/(3K)$; (ii) the mean of $LM_K(H)$ is $1/3 + 2/(3K) + \frac{K}{3(K-1)} \sum_{k=2}^{K} 1/k^2$; (iii) the mean of $UM_K(H) - LM_K(H)$ is $1/3 - 1/(3K) - \frac{K}{3(K-1)} \sum_{k=2}^{K} 1/k^2$.*

Proof i.

$$\frac{1}{1-1/K} \int\limits_{\frac{1}{K}}^{1} UM_K(H)dH = \frac{K}{K-1} \int\limits_{\frac{1}{K}}^{1} \frac{1}{K} + \frac{1}{K}\left(\sqrt{(K-1)}\sqrt{(KH-1)}\right)dH.$$
$$= \frac{2}{3} + \frac{1}{3K}.$$

ii.

$$\frac{1}{1-1/K} \int\limits_{\frac{1}{K}}^{1} LM_K(H)dH = \frac{K}{K-1} \sum_{k=2}^{K} \int\limits_{1/k}^{1/(k-1)} \frac{1}{k}\left(1 + \frac{\sqrt{kH-1}}{\sqrt{k-1}}\right)dH$$
$$= \frac{K}{3(K-1)} \sum_{k=2}^{K} \left[\frac{1}{k} - \frac{1}{k-1} + \frac{2}{(k-1)^2} - \frac{1}{k^2}\right].$$

$\sum_{k=2}^{K} 1/k - 1/(k-1)$ and $\sum_{k=2}^{K} 2/(k-1)^2 - 1/k^2$ simplify to $-1 + 1/K$ and $2 - 2/K^2 + \sum_{k=2}^{K} 1/k^2$, respectively.

iii. This result follows by taking the difference between the results in parts i and ii. □

Proposition 8 *For $K \geq 2$, averaging across values of $H \in [1/K, 1)$, (i) the mean of $UM_K(H) - H$ is $1/6 - 1/(6K)$; (ii) the mean of $LM_K(H) - H$ is $-1/6 + 1/(6K) + \frac{K}{3(K-1)} \sum_{k=2}^{K} 1/k^2$.*

Proof By Lemma 7 of Rosenberg and Jakobsson (2008), on the interval $[1/K, 1)$, $LM_K(H) \geq H$. Averaging over $H \in [1/K, 1)$, the mean of H is $1/2 + 1/(2K)$. We subtract this expression from the results in parts i and ii of Proposition 7 to acquire the means of $UM_K(H) - H$ and $LM_K(H) - H$, respectively. □

By noting that $\sum_{k=2}^{\infty} = \pi^2/6 - 1$, we can observe that as $K \to \infty$, the limiting values in the three parts of Proposition 7 approach the corresponding values in Proposition 6 of Rosenberg and Jakobsson (2008): $2/3$, $\pi^2/18$, and $2/3 - \pi^2/18$, respectively. Similarly, the limits in Proposition 8 approach the quantities in Proposition 8 of Rosenberg and Jakobsson (2008): $1/6$ and $\pi^2/18 - 1/2$.

It is also noteworthy that the mean difference $UM_K(H) - LM_K(H)$ for a fixed K is always smaller than the large-K limiting mean difference. Thus, when incorporating the value of K, the interval in which M is confined by its upper and lower bounds has a narrower range than in the case of K unspecified. We can measure the mean improvement that specification of K provides in ascertaining M given H, by evaluating the difference between the mean difference of the upper and lower bounds of M with K unspecified and the reduced mean difference of the upper and lower bounds for fixed finite K.

Proposition 9 *For $K \geq 2$, averaging across values of $H \in [1/K, 1)$, (i) the mean of $UM_\infty(H) - LM_\infty(H)$ is $[K\sqrt{K} + K - 2]/[3K(\sqrt{K} + 1)] - \frac{K}{3(K-1)} \sum_{k=2}^{K} 1/k^2$; (ii) the mean difference between $UM_\infty(H) - LM_\infty(H)$ and $UM_K(H) - LM_K(H)$ is $(\sqrt{K} - 1)/[3K(\sqrt{K} + 1)]$.*

Proof i. Using Proposition 7ii to obtain the mean of $LM_K(H)$, which equals $LM_\infty(H)$ on $[1/K, 1)$, the result follows by simplifying

$$\frac{1}{1 - 1/K} \int_{\frac{1}{K}}^{1} (UM_\infty(H) - LM_K(H))dH = \frac{K}{K-1} \left(\int_{\frac{1}{K}}^{1} \sqrt{H} dH \right) - \left(\frac{1}{3} + \frac{2}{3K} + \frac{K}{3(K-1)} \sum_{k=2}^{K} \frac{1}{k^2} \right).$$

ii. The result follows by subtracting the result in Proposition 7iii from the result in i. □

As $K \to \infty$, the limiting values in Proposition 9i and ii approach $2/3 - \pi^2/18$ and 0, respectively. These results are sensible: as $K \to \infty$, the set of allowed values of H approaches $(0, 1)$, so that the mean of $UM_\infty(H) - LM_\infty(H)$ over $[1/K, 1)$ approaches its mean over the entire interval $(0, 1)$. $UM_K(H) - LM_K(H)$ approaches $UM_\infty(H) - LM_\infty(H)$, so the mean of $UM_K(H) - LM_K(H)$ approaches the mean of $UM_\infty(H) - LM_\infty(H)$.

For K equal to 3, 5, and 15, Fig. 4 plots the areas that are lost from the unspecified-K case in refining the upper bound on M given specified values of K. For K equal to 2, 3, 4, 5, and 15, evaluating the quantity in Proposition 9ii, the reductions in the mean difference between the upper and lower bounds on M are approximately 0.0286, 0.0298,

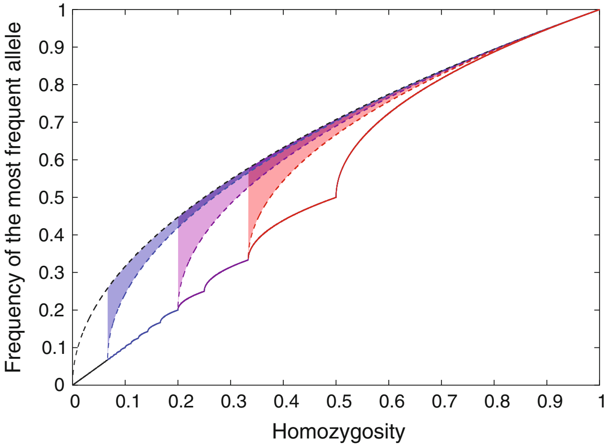

Fig. 4 The refinement in the range for the frequency of the most frequent allele as a function of homozygosity, when K is specified compared to unspecified. The red, purple, and blue regions correspond to the reductions in range for $K = 3$, $K = 5$, and $K = 15$, respectively.

0.0278, 0.0255, and 0.0131, respectively. These values are not insignificant fractions of 0.1184, the mean difference between the upper and lower bounds on M over the whole unit interval in the case of K unspecified. The maximal improvement occurs at $K = 3$, accounting for ∼25% of the total area between $UM_\infty(H)$ and $LM_\infty(H)$.

3.2 Maximal and minimal differences between the bounds

Figure 5 plots the pairwise differences between the upper bound of M, the lower bound of M, and H itself. We now prove a variety of results about $UM_K(H)$, $LM_K(H)$, and H, based on patterns visible in the figure.

In Fig. 5a, we notice that the difference between the upper bound of M and H has a single maximal value within $[1/K, 1)$. The following proposition identifies the location of this point.

Proposition 10 *On* $[1/K, 1)$*, the maximal value of* $UM_K(H) - H$ *is* $(K-1)/(4K)$*, and it is achieved at* $H = (K+3)/(4K)$.

Proof The derivative of $UM_K(H) - H$ with respect to H, $(\sqrt{K-1})/(2\sqrt{KH-1})-1$, has a single critical point at $((K+3)/(4K), (K-1)/(4K))$. The second derivative of $UM_K(H) - H$ is negative over the entire domain, so that $UM_K(H) - H$ achieves a global maximum at the critical point. □

As $K \to \infty$, the maximum of $UM_K(H) - H$ approaches $(1/4, 1/4)$, the location of the maximum of $UM_\infty(H) - H$, as derived in Corollary 13 of Rosenberg and Jakobsson (2008) for the case of unspecified K.

Figure 5b plots the difference $LM_K(H) - H$. As shown in Proposition 9 of Rosenberg and Jakobsson (2008), because $LM_K(H) = LM_\infty(H)$ on I_k for $k \leq K$, the local maximum of $LM_K(H) - H$ in I_k occurs at $((4k-3)/[4k(k-1)], 1/[4k(k-1)])$.

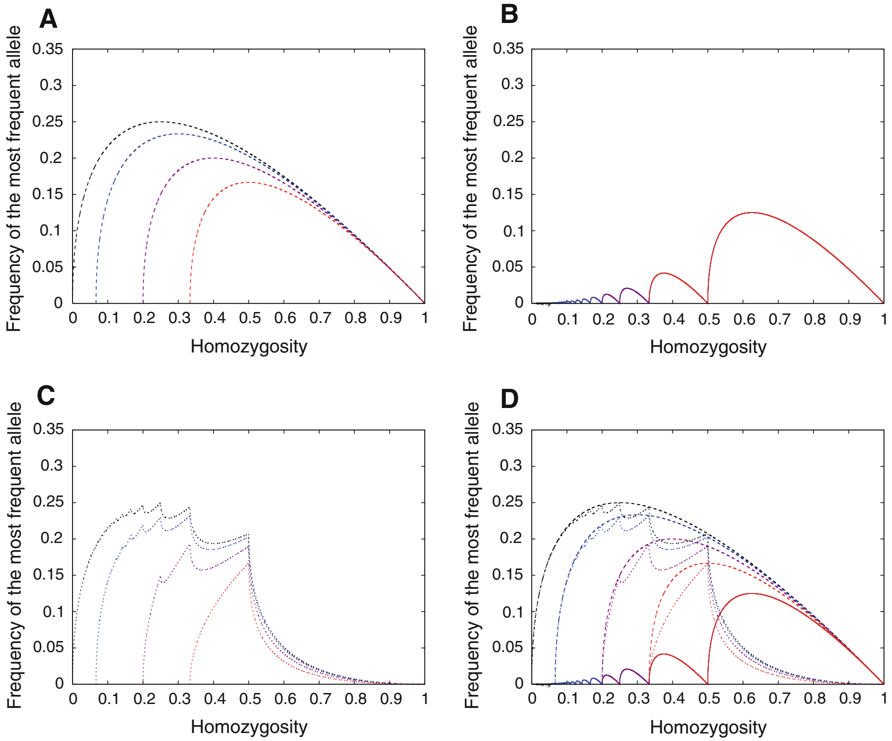

Fig. 5 The difference between the upper and lower bounds on the frequency of the most frequent allele, for a given homozygosity, and the difference between the bounds and homozygosity itself. (**a**) $UM_K(H) - H$. (**b**) $LM_K(H) - H$. (**c**) $UM_K(H) - LM_K(H)$. (**d**) Superposition of parts (**a**), (**b**), and (**c**). Plots for loci with $K = 3$, $K = 5$, and $K = 15$ alleles appear in red, purple, and blue, respectively, each with domain $[1/K, 1)$. The case of unspecified K is plotted in black, with domain $(0, 1)$.

As shown in Corollary 10 of Rosenberg and Jakobsson (2008), considering all k with $2 \le k \le K$, the highest of the local maxima is achieved at $(5/8, 1/8)$.

In Fig. 5c, we observe that the difference between the upper and lower bounds of M has a series of local minima, with maxima occurring at reciprocals of integers. We now derive the locations of the minima.

Proposition 11 *Suppose $K \ge 3$. On $[1/k, 1/(k-1)]$, for $k \in [2, K]$, the minimal value of $UM_K(H) - LM_K(H)$ occurs at $H = [(k-1)(K-1)-1]/[k(k-1)(K-1)-K]$, and is*

$$\frac{\sqrt{K-k}\left[\sqrt{k(k-1)(K-1)-K}-\sqrt{(k-1)(K-k)}\right]}{Kk\sqrt{k-1}}.$$

Proof If $k = K = 2$, then $UM_K(H) = LM_K(H)$. Otherwise, $\frac{d}{dH}(UM_K(H) - LM_K(H)) = (1/2)[\sqrt{K-1}/\sqrt{KH-1} - 1/\sqrt{(k-1)(kH-1)}]$. The only critical value occurs at $H = [(k-1)(K-1)-1]/[k(k-1)(K-1)-K]$, which we denote by $\phi(K, k)$. This value lies in the interior of $[1/k, 1/(k-1)]$, except if $k = K$, for which it lies at the left endpoint, and for $k = 2$, for which it lies at the right endpoint.

To verify that this critical value is a minimum, we evaluate the second derivative of $UM_K(H) - LM_K(H)$:

$$\frac{d^2}{dH^2}(UM_K(H) - LM_K(H)) = \frac{1}{4}\left(\frac{k}{\sqrt{k-1}(kH-1)^{3/2}} - \frac{K\sqrt{K-1}}{(KH-1)^{3/2}}\right).$$

The second derivative is positive at H if $(K/k)\sqrt{k-1}\sqrt{K-1} < [(KH-1)/(kH-1)]^{3/2}$. At $H = \phi(K,k)$, $(KH-1)/(kH-1) = (K-1)(k-1)$, from which it follows that the second derivative is positive at $H = \phi(K,k)$ if $K/(K-1) < k(k-1)$. For $K \geq k \geq 2$ and $K \geq 3$, $K/(K-1) < 2$ and $k(k-1) \geq 2$. Thus, $H = \phi(K,k)$ is a minimum. □

Taking the limit of $\phi(K,k)$ as $K \to \infty$ yields $(k-1)/(k^2-k-1)$, the location of the minimum of $UM_\infty(H) - LM_\infty(H)$ on $H \in [1/k, 1/(k-1)]$ according to Proposition 11ii of Rosenberg and Jakobsson (2008). The value of $UM_K(H) - LM_K(H)$ at the minimum approaches $(\sqrt{k^2-k-1}/\sqrt{k-1} - 1)/k$, the minimum of $UM_\infty(H) - LM_\infty(H)$ on $[1/k, 1/(k-1)]$.

We now identify the location of the highest of the local minima of $UM_K(H) - LM_K(H)$.

Proposition 12 *Suppose* $K \geq 4$. *On* $[1/K, 1]$, *the highest local minimum of* $UM_K(H) - LM_K(H)$ *occurs at* $H = \phi(K,3)$ *for* $K \in \{4, 5\}$, *at both* $H = \phi(K,3)$ *and* $H = \phi(K,4)$ *for* $K = 6$, *at* $H = \phi(K,4)$ *for* $K \in [7, 17]$, *and at* $H = \phi(K,5)$ *for* $K \geq 18$.

Proof By Proposition 11, we aim to find the maximal value among local minima that occur at $(\phi(K,k), UM_K(\phi(K,k)) - LM_K(\phi(K,k)))$, for integers $k \in [3, K-1]$. The interval $[1/k, 1/(k-1)]$ has no interior local minimum for $k = 2$ or $k = K$. Thus, for $K = 2$ and $K = 3$ there are no nonzero minima, and for $K = 4$ the only nonzero minimum occurs at $\phi(4, 3)$.

Consider $K \geq 5$. For $k \neq 2$ and $k \neq K$, the proof of Proposition 11 shows that each interval $[1/k, 1/(k-1)]$ has a single interior critical point for $UM_K(H) - LM_K(H)$, and that this point is a minimum. As a result, the maximum of the function must occur at one of the endpoints of the interval. Thus, given K, if there exist values of k and J such that $UM_K(\phi(K,k)) - LM_K(\phi(K,k))$ exceeds $UM_K(1/j) - LM_K(1/j)$ for all $j \geq J$, then $UM_K(\phi(K,k)) - LM_K(\phi(K,k))$ exceeds the local maxima—and consequently, the local minima—on all intervals $[1/j, 1/(j-1)]$ for all $j > J$. It can be shown that $UM_K(\phi(K,5)) - LM_K(\phi(K,5))$ exceeds $UM_K(1/7) - LM_K(1/7)$ for all $K \geq 7$.

Denoting $\alpha_K(m,n) = [UM_K(\phi(K,m)) - LM_K(\phi(K,m))] - [UM_K(\phi(K,n)) - LM_K(\phi(K,n))]$, it can be shown that $\alpha_K(5,6) > 0$ for all $K \geq 6$, and $\alpha_K(5,7) > 0$ for all $K \geq 7$. Thus, for all $K \geq 4$, the desired maximum among the local minima occurs at $\phi(K,k)$ for some $k \leq 5$.

It can be shown that $\alpha_K(3,4) > 0$ for $K = 5$, $\alpha_K(3,4) = 0$ for $K = 6$, $\alpha_K(4,5) > 0$ for $K \in [5, 17]$, $\alpha_K(4,3) > 0$ for $K \geq 7$, and $\alpha_K(5,4) > 0$ for $K \geq 18$. □

For large K, the highest local minimum of $UM_K(H) - LM_K(H)$ occurs at $\phi(K, 5) = 4/19$, the location of the highest local minimum of $UM_\infty(H) - LM_\infty(H)$ by Proposition 12 of Rosenberg and Jakobsson (2008).

We next consider the maxima of the difference $UM_K(H) - LM_K(H)$. First we find the global maximum. We then examine the locations of the local maxima in intervals $[1/k, 1/(k-1)]$.

Proposition 13 *Suppose* $K \geq 3$. *On* $[1/K, 1]$, *the maximal value of* $UM_K(H) - LM_K(H)$ *occurs at* $H = 1/2$ *for* $K \in \{3, 4\}$, *at* $H = 1/3$ *for* $K \in [5, 18]$, *and at* $H = 1/4$ *for* $K \geq 19$.

Proof For $K = 2$, $UM_K(H) - LM_K(H) = 0$. For $K \geq 3$, local maxima of $UM_K(H) - LM_K(H)$ occur at integers $1/k$, where $UM_K(H) - LM_K(H)$ coincides with $UM_K(H) - H$. By Proposition 10, the global maximum of $UM_K(H) - H$, which has a single critical value, occurs at $(K+3)/(4K)$. Consequently, the highest local maximum of $UM_K(H) - LM_K(H)$ occurs at one of the endpoints of the interval for which $1/k \leq (K+3)/(4K) \leq 1/(k-1)$, that is, the interval on whose interior $k = \lceil 4K/(K+3) \rceil$.

If $K = 3$, $(K+3)/(4K)$ is exactly equal to $1/2$, and $UM_K(H) - LM_K(H)$ is therefore maximized at $1/2$. Similarly, if $K = 9$, $(K+3)/(4K) = 1/3$, and $UM_K(H) - LM_K(H)$ is maximized at $1/3$. For all other K, the global maximum of $UM_K(H) - H$ at $(K+3)/(4K)$ occurs interior to some interval $[1/k, 1/(k-1)]$: $[1/2, 1/3]$ for $K \in [4, 8]$ and $[1/3, 1/4]$ for $K \geq 10$.

For $K \neq 3$ and $K \neq 9$, to determine whether $UM_K(1/k) - LM_K(1/k)$ is maximized at $k = 2$ or $k = 3$, we define $\psi(K, k) = [UM_K(1/k) - 1/k] - [UM_K(1/(k-1)) - 1/(k-1)]$. $\psi(4, 3) < 0$, but $\psi(K, 3) > 0$ for $K \in \{5, 6, 7, 8\}$. Therefore, the global maximum of $UM_K(H) - LM_K(H)$ for $K = 4$ occurs at $H = 1/2$, and for $K \in [5, 8]$, it occurs at $H = 1/3$.

For $K \geq 10$, it can be shown that the function $\psi(K, 4)$ is negative until its single root at $K \approx 18.83$, and positive thereafter. Consequently, the global maximum of $UM_K(H) - LM_K(H)$ for $K \in [10, 18]$ occurs at $H = 1/3$, and for $K \geq 19$, it occurs at $H = 1/4$. □

For large K, the location of the maximal difference between the upper and lower bounds of M is at $H = 1/4$, as in the unspecified-K case in Corollary 14 of Rosenberg and Jakobsson (2008).

Proposition 14 *Suppose* $K \geq 3$. *On* $[1/k, 1/(k-1)]$, *for* $k \in [2, K]$, *the maximal value of* $UM_K(H) - LM_K(H)$ *occurs (i) at* $H = 1/k$ *if* $k < \lceil 4K/(K+3) \rceil$, *(ii) at* $H = 1/(k-1)$ *if* $k > \lceil 4K/(K+3) \rceil$, *and (iii) at the location in Table 2 if* $k = \lceil 4K/(K+3) \rceil$.

Proof By Lemma 7 of Rosenberg and Jakobsson (2008), $UM_K(H) - LM_K(H) \leq UM_K(H) - H$, with equality if and only if $H = 1/k$ for integers $k \in [2, K]$. By Proposition 10, $UM_K(H) - H$ is increasing over $[1/K, (K+3)/(4K)]$ and decreasing over $[(K+3)/(4K), 1]$. Proposition 11 shows that $UM_K(H) - LM_K(H)$ has at most one critical value on $[1/k, 1/(k-1)]$, a local minimum. Thus, the maximal value of

Table 2 The location of the maximum of $UH_K(M) - LH_K(M)$ on $[1/k, 1/(k-1)]$, when $k = \lceil 4K/(K+3) \rceil$. The table is part of the statement of Proposition 14

K	$k = \lceil 4K/(K+3) \rceil$	Location of maximum of $UH_K(M) - LH_K(M)$ on $[1/k, 1/(k-1)]$	
		Interval endpoint	Value
3	2	$1/k$	$1/2$
4	3	$1/(k-1)$	$1/2$
$5, 6, 7, 8, 9$	3	$1/k$	$1/3$
$10, 11, \dots, 18$	4	$1/(k-1)$	$1/3$
≥ 19	4	$1/k$	$1/4$

$UM_K(H)$–$LM_K(H)$ on the interval occurs at one of the endpoints, where $UM_K(H) - LM_K(H)$ has the same values as $UM_K(H) - H$. For $k < \lceil 4K/(K+3) \rceil$, the maximum occurs at $1/k$, and for $k > \lceil 4K/(K+3) \rceil$, it occurs at $1/(k-1)$. If $k = \lceil 4K/(K+3) \rceil$, then the maximum occurs at the global maximum of $UM_K(H) - LM_K(H)$ over $[1/K, 1]$, whose location, derived in Proposition 13, is given in Table 2. □

As $K \to \infty$, the maximum on $[1/k, 1/(k-1)]$ occurs at $1/(k-1)$ for $k > 4$ and at $1/k$ for $k \leq 4$. This limit agrees with the result in Proposition 11i of Rosenberg and Jakobsson (2008) for $UM_\infty(H) - LM_\infty(H)$.

In Fig. 5d, we notice that the difference between the upper and lower bounds of M intersects the difference between the lower bound of M and H at a coordinate within the interval $[1/2, 1)$. In the following proposition, we analyze the behavior of $UM_K(H) - LM_K(H)$ and $LM_K(H) - H$.

Proposition 15 *Suppose* $K \geq 3$. *Define* $R(K)$ *to be the location of the single root of* $(1-K+KH-K\sqrt{2H-1}+\sqrt{K-1}\sqrt{KH-1})/K$ *that lies in* $(1/2, 1)$. *Then the difference* $UM_K(H)-LM_K(H)$ *is (i) greater than* $LM_K(H)-H$ *if* $1/K < H < R(K)$; *(ii) equal to* $LM_K(H) - H$ *if* $H \in \{1/K, R(K), 1\}$; *(iii) less than* $LM_K(H) - H$ *if* $R(K) < H < 1$.

Proof We show that $UM_K(H) - LM_K(H) \geq LM_K(H) - H$ for $H \in I_k$ and $k \in [3, K]$, with equality if and only if $H = 1/K$. In interval I_K,

$$(UM_K(H) - LM_K(H)) - (LM_K(H) - H) = \frac{1}{K}\sqrt{KH-1}\Big(\sqrt{KH-1} + \sqrt{K-1} - 2/\sqrt{K-1}\Big). \quad (4)$$

Eq. 4 is 0 for $H = 1/K$. Elsewhere on I_K, $\sqrt{KH-1}$ is positive, and $\sqrt{K-1} - 2/\sqrt{K-1} \geq 0$ for $K \geq 3$.

For $H \in [1/K, 1)$, if $K' > K$, then $UM_{K'}(H) > UM_K(H)$ and $LM_{K'}(H) = LM_K(H)$. Consider an interval I_k, for $k \in [3, K']$. For $k = K'$, $UM_{K'}(H) - LM_{K'}(H) \geq LM_{K'}(H) - H$ on I_k, with equality if and only if $H = 1/K'$, based on the argument in the previous paragraph with K' in place of K. For $k \in [3, K'-1]$,

$(UM_{K'}(H)-LM_{K'}(H))-(LM_{K'}(H)-H) > (UM_k(H)-LM_k(H))-(LM_k(H)-H)$ on the interval I_k. But $(UM_k(H) - LM_k(H)) - (LM_k(H) - H) > 0$ based on the argument in the previous paragraph with k in place of K. Thus, we can conclude that for all $K \geq 3$, $UM_K(H) - LM_K(H) \geq LM_K(H) - H$ for $H \in [1/K, 1/2)$, with equality if and only if $H = 1/K$.

Now consider $k = 2$, for which $H \in [1/2, 1]$. On this interval,

$$(UM_K(H) - LM_K(H)) - (LM_K(H) - H) = \big[1 - K + KH - K\sqrt{2H-1} + \sqrt{K-1}\sqrt{KH-1}\big]/K. \tag{5}$$

For each $K \geq 3$, this function is positive at $H = 1/2$, negative at $H = 1 - 1/K$, and zero at $H = 1$. It is straightforward to show that its second derivative is positive on $[1/2, 1]$. We can therefore conclude that eq. 5 has a single zero interior to $(1/2, 1)$. The location of this zero does not have a convenient formula, and we simply label it $R(K)$. Because $(UM_K(H) - LM_K(H)) - (LM_K(H) - H)$ is continuous and is positive at $H = 1/2$, it switches from positive to negative at $H = R(K)$ before reaching zero again at $H = 1$. □

As $K \to \infty$, $R(K)$ becomes the single root of $-1 + H - \sqrt{2H-1} + \sqrt{H}$ that lies in $(1/2, 1)$, or $4 - 2\sqrt{3}$. The result then accords with the unspecified-K case in Proposition 15 of Rosenberg and Jakobsson (2008).

4 Features of the bounds on *H* in terms of *M*

We now turn our attention to the various properties of the upper and lower bounds of H, UH_K and LH_K, in terms of M. As in the case of the bounds on M in terms of H in Sect. 3, we begin by examining the means of the upper and lower bounds over the interval $[1/K, 1)$.

4.1 Mean values of the bounds

Proposition 16 *For* $K \geq 2$, *averaging across values of* $M \in [1/K, 1)$, *(i) the mean of* $UH_K(M)$ *is* $2/3 + 1/(3K) - \frac{K}{3(K-1)} \sum_{k=2}^{K} 1/k^2$; *(ii) the mean of* $LH_K(M)$ *is* $1/3 + 2/(3K)$; *(iii) the mean of* $UH_K(M) - LH_K(M)$ *is* $1/3 - 1/(3K) - \frac{K}{3(K-1)} \sum_{k=2}^{K} 1/k^2$.

Proof iii. Using the inverse relationships between $UH_K(M)$ and $LM_K(H)$ and between $LH_K(M)$ and $UM_K(H)$, the area between $UH_K(M)$ and $LH_K(M)$ is identical to the area between $UM_K(H)$ and $LM_K(H)$. The result follows from Proposition 7iii.

ii. $\frac{1}{1-1/K} \int_{1/K}^{1} LH_K(M)dM = \frac{K}{K-1} \int_{1/K}^{1} \frac{KM^2-2M+1}{K-1} dM = 1/3 + 2/(3K)$.

i. The result follows by summing the results in parts ii and iii. □

Proposition 17 *For* $K \geq 2$, *averaging across values of* $M \in [1/K, 1)$, *(i) the mean of* $M - UH_K(M)$ *is* $-1/6 + 1/(6K) + \frac{K}{3(K-1)} \sum_{k=2}^{K} 1/k^2$; *(ii) the mean of* $M - LH_K(M)$ *is* $1/6 - 1/(6K)$.

Proof i. By Lemma 18 of Rosenberg and Jakobsson (2008), $M \geq UH_K(M)$. Subtracting the result of Proposition 16i from $\frac{1}{1-1/K}\int_{1/K}^{1} MdM = 1/2 + 1/(2K)$, the result follows.
ii. The result follows by summing the results in part i and Proposition 16iii. □

Taking the limits as $K \to \infty$ for the results in Proposition 16, we obtain $1-\pi^2/18$, $1/3$, and $2/3-\pi^2/18$, respectively, as in Proposition 17 of Rosenberg and Jakobsson (2008). Similarly, the limits of $\pi^2/18 - 1/2$ and $1/6$ in Proposition 17i and ii agree with Proposition 19 of Rosenberg and Jakobsson (2008).

As in the case of the bounds on M, we can compute the reduction in mean difference between the upper and lower bounds on H yielded by specifying K.

Proposition 18 *For $K \geq 2$, averaging across values of $M \in [1/K, 1)$, (i) the mean of $UH_\infty(M) - LH_\infty(M)$ is $1/3 - 1/(3K^2) - \frac{K}{3(K-1)}\sum_{k=2}^{K} 1/k^2$; (ii) the mean difference between $UH_\infty(M) - LH_\infty(M)$ and $UH_K(M) - LH_K(M)$ is $(K-1)/(3K^2)$.*

Proof ii. Noting that $UH_\infty(M) = UH_K(M)$, the desired mean reduces to $\frac{1}{1-1/K}\int_{1/K}^{1} LH_K(M) - LH_\infty(M)dM$. Subtracting $\frac{1}{1-1/K}\int_{1/K}^{1} LH_\infty(M)dM = \frac{K}{K-1}\int_{1/K}^{1} M^2dM = 1/3 + 1/(3K) + 1/(3K^2)$ from the quantity in Proposition 16ii, the result follows.
i. This result follows by summing the results in part ii and Proposition 16iii. □

As $K \to \infty$, the limiting values in Proposition 18i and ii approach $2/3 - \pi^2/18$ and 0, respectively. These results are sensible, as the region represented in Proposition 18i approaches the region between UH_∞ and LH_∞ over the whole unit interval, a region with area $2/3 - \pi^2/18$. The region in Proposition 18ii becomes progressively smaller as the region between UH_K and LH_K approaches the region between UH_∞ and LH_∞.

For K equal to 3, 5, and 15, Fig. 6 plots the areas lost in refining the upper bound of H given K compared to the case of K unspecified. For K equal to 2, 3, 4, 5, and 15, using Proposition 18ii, the reductions in the mean difference between the upper and lower bounds on M are approximately 0.0833, 0.0741, 0.0625, 0.0533, and 0.0207, respectively. Especially for small K, these values provide substantial reductions compared to 0.1184, the mean difference between the upper and lower bounds on H over $(0, 1)$ when K is unspecified. The largest reduction occurs for $K = 2$.

4.2 Maximal and minimal differences between the bounds

Figure 7 plots the differences among the upper bound of H, the lower bound of H, and M itself. We now examine the properties of the local maxima and minima visible in the figure.

The difference between M and the upper bound of H is the same for specified K as for K unspecified, except with a different domain. Thus, in Fig. 7a, which plots the difference between M and the upper bound of H, a local maximum occurs in each interval $[1/k, 1/(k-1)]$ for $2 \leq k \leq K$, as in Proposition 20 of Rosenberg and Jakobsson (2008). This maximum occurs at $((2k-1)/[2k(k-1)], 1/[4k(k-1)])$.

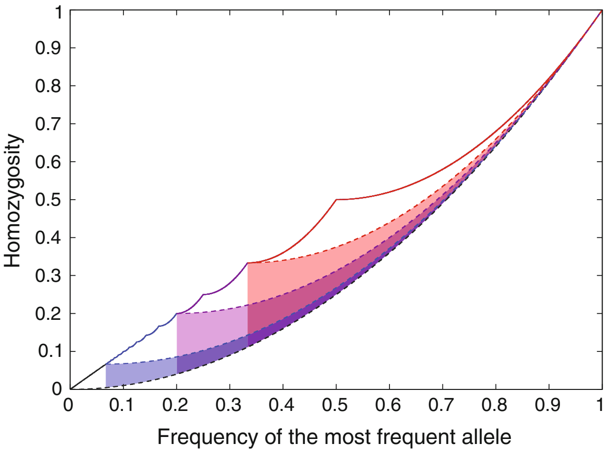

Fig. 6 The refinement in the range for homozygosity as a function of the frequency of the most frequent allele, when K is specified compared to unspecified. The red, purple, and blue regions correspond to the reductions in range for $K = 3$, $K = 5$, and $K = 15$, respectively.

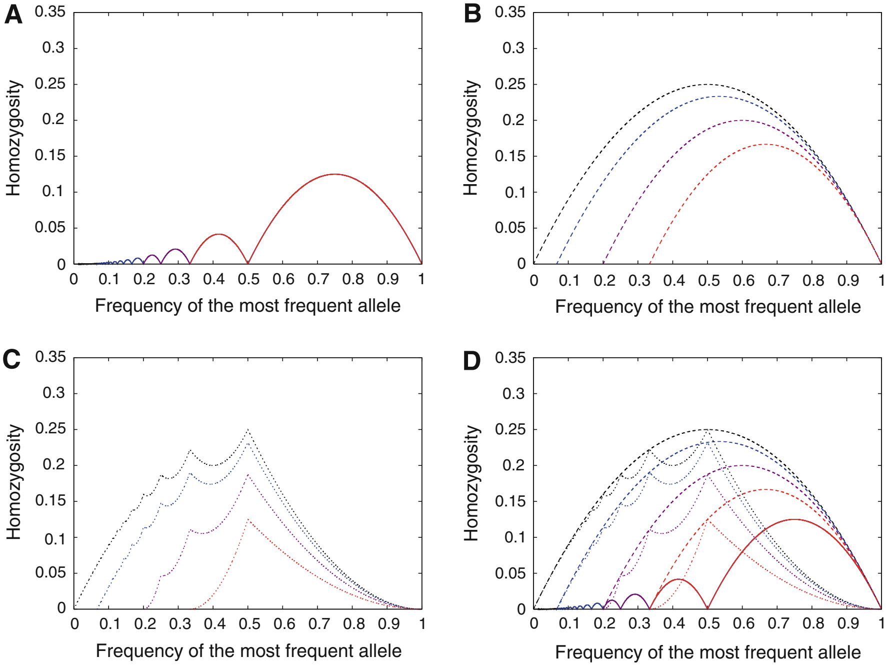

Fig. 7 The difference between the upper and lower bounds on homozygosity, for a given frequency of the most frequent allele, and the difference between the frequency of the most frequent allele and the bounds. (**a**) $M - UH_K(M)$. (**b**) $M - LH_K(M)$. (**c**) $UH_K(M) - LH_K(M)$. (**d**) Superposition of parts (**a**), (**b**), and (**c**). Plots for loci with $K = 3$, $K = 5$, and $K = 15$ alleles appear in red, purple, and blue, respectively, each with domain $[1/K, 1)$. The case of unspecified K is plotted in black, with domain $(0, 1)$.

As in Corollary 21 of Rosenberg and Jakobsson (2008), the highest maximum occurs at (3/4, 1/8), when $k = 2$.

In Fig. 7b, which plots the difference between M and the lower bound of H, we notice that a single maximum value occurs in the domain $[1/K, 1)$. We now establish the location of this maximum.

Proposition 19 *On* $[1/K, 1)$*, the maximal value of* $M - LH_K(M)$ *is* $1/4 - 1/(4K)$*, and it is achieved at* $M = 1/2 + 1/(2K)$.

Proof $M - LH_K(M) = [-KM^2 + (K+1)M - 1]/(K-1)$. The vertex of this parabola is at $M = (K+1)/(2K)$, which necessarily lies in the interior of $[1/K, 1)$. □

The location of the maximum is at the midpoint of the interval $[1/K, 1)$, and $M - LH_K(M)$ is zero at the endpoints $1/K$ and 1. As $K \to \infty$, the location of the maximum approaches (1/2, 1/4), the location of the maximum of $M - LH_\infty(M)$ in Corollary 24 of Rosenberg and Jakobsson (2008).

Figure 7c plots the difference between the upper and lower bounds of H. For $K \geq 3$ and $3 \leq k \leq K - 1$, this curve has an interior local minimum in the interval $[1/k, 1/(k-1)]$.

Proposition 20 *Suppose* $K \geq 3$. *On* $[1/k, 1/(k-1)]$*, for* $k \in [2, K]$*, the minimal value of* $UH_K(M) - LH_K(M)$ *occurs at* $M = [k(K-1) - K]/[k(k-1)(K-1) - K]$*, and is* $[(k-2)(K-k)]/[k(k-1)(K-1) - K]$.

Proof On $[1/k, 1/(k-1)]$, $UH_K(M) - LH_K(M) = (k^2 - k - \frac{K}{K-1})M^2 + (\frac{2}{K-1} - 2k + 2)M + (1 - \frac{1}{K-1})$. Excluding the case of $k = K = 2$, this function is a parabola with positive leading coefficient, whose vertex is at $M = [k(K-1) - K]/[k(k-1)(K-1) - K]$. Except if $k = 2$ or $k = K$, this point lies in the interior of the interval $[1/k, 1/(k-1)]$; if $k = K$, it lies at the left endpoint, and if $k = 2$, it lies at the right endpoint. At the minimum, the value of $UH_K(M) - LH_K(M)$ is $(k-2)(K-k)/[k(k-1)(K-1) - K]$. This minimum is zero in the leftmost interval, with $k = K$, and in the rightmost interval, with $k = 2$. □

As $K \to \infty$, the locations of the local minima approach $((k-1)/(k^2 - k - 1), (k-2)/(k^2 - k - 1))$. These locations match those found for $UH_\infty(M) - LH_\infty(M)$ in Proposition 22ii of Rosenberg and Jakobsson (2008).

We now identify the location of the greatest local minimum of $UH_K(M) - LH_K(M)$.

Proposition 21 *Suppose* $K \geq 4$. *On* $[1/K, 1]$*, the highest local minimum of* $UH_K(M) - LH_K(M)$ *is* $(K-3)/(5K-6)$*, and it occurs at* $M = (2K-3)/(5K-6)$.

Proof For $2 \leq k \leq K$, each interval $[1/k, 1/(k-1)]$ has an interior local minimum, except if $k = 2$ or $k = K$. Thus, no interior local minima occur for $K = 2$ or $K = 3$, and we only consider $K \geq 4$. Using Proposition 20, define $\omega_K(k)$ as the minimum value of $UH_K(M) - LH_K(M)$ over the interval $[1/k, 1/(k-1)]$. We show that among integers k with $3 \leq k \leq K - 1$, $\omega_K(k)$ is maximal at $k = 3$:

$$\omega_K(k) - \omega_K(k+1) = \frac{k(k-3)K^2 + (K^2 - k^2) + k(2K-1) + K}{[k(k-1)(K-1) - K][(k+1)k(K-1) - K]}.$$

For $K \geq 4$ and $k \geq 3$, the denominator is positive. For $k \geq 3$ and $K \geq k$, the numerator is positive. Thus, as $\omega_K(k) > \omega_K(k+1)$, $k = 3$ produces the maximal value of $\omega_K(k)$, at the location in Proposition 20. □

As $K \to \infty$, the limiting location of the highest of the local minima is at $(2/5, 1/5)$, the location of the highest local minimum of $UH_\infty(M) - LH_\infty(M)$ in Proposition 23 of Rosenberg and Jakobsson (2008).

Our next result examines local maxima of $UH_K(M) - LH_K(M)$.

Proposition 22 *Suppose $K \geq 3$. On $[1/k, 1/(k-1)]$, for $k \in [3, K]$, the maximal value of $UH_K(M) - LH_K(M)$ is $(k-2)(K-k+1)/[(k-1)^2(K-1)]$, and it is achieved at $M = 1/(k-1)$; if $k = 2$, the maximal value of $UH_K(M) - LH_K(M)$ is $(K-2)/[4(K-1)]$, and it is achieved at $M = 1/2$.*

Proof The maximum must occur at one of the two endpoints of the interval. For integers k with $2 \leq k \leq K$, $UH_K(1/k) = 1/k$. At $M = 1/k$, $UH_K(M) - LH_K(M) = (k-1)(K-k)/[k^2(K-1)]$, and at $M = 1/(k-1)$, $UH_K(M) - LH_K(M) = (k-2)(K-k+1)/[(k-1)^2(K-1)]$. It is straightforward to show that the latter value is greater than the former for $k \geq 3$ and $K \geq k$, and that the reverse is true for $k = 2$. □

Corollary 23 *Suppose $K \geq 3$. On $[1/K, 1]$, the maximal value of $UH_K(M) - LH_K(M)$ is $(K-2)/[4(K-1)]$, and it is achieved at $M = 1/2$.*

Proof By Proposition 22, the local maxima of $UH_K(M) - LH_K(M)$ occur at points $1/k$ for $2 \leq k < K$. At such points, $UH_K(M) = M$. From Proposition 19, $M - LH_K(M)$ is increasing on $[1/K, 1/2 + 1/(2K)]$. Thus, considering points $1/k$ where k is an integer with $2 \leq k < K$, the maximum of $UH_K(M) - LH_K(M)$ occurs at $M = 1/2$. □

As $K \to \infty$, local maxima occur at $(1/(k-1), (k-2)/(k-1)^2)$ for $k > 2$, and at $(1/2, 1/4)$ for $k = 2$. These locations accord with Proposition 22i of Rosenberg and Jakobsson (2008). The global maximum approaches $(1/2, 1/4)$, the location in Corollary 25 of Rosenberg and Jakobsson (2008).

In Fig. 7d, we notice that the curves in Figures 7b and 7c coincide at reciprocals of integers, while the curves in Figs. 7a and 7c coincide only in the two intervals $(1/K, 1/(K-1))$ and $(1/2, 1)$ for $K \geq 3$. We now determine the precise locations where $M - UH_K(M)$ and $UH_K(M) - LH_K(M)$ intersect.

Proposition 24 *Suppose $K \geq 3$. Define $S(K) = (2K-3)/(2K^2-4K+1)$ and $T(K) = (2K-3)/(3K-4)$. Then the difference $UH_K(M) - LH_K(M)$ is (i) greater than $M - UH_K(M)$ if $S(K) < M < T(K)$; (ii) equal to $M - UH_K(M)$ if $M \in \{1/K, S(K), T(K), 1\}$; (iii) less than $M - UH_K(M)$ if $1/K < M < S(K)$ or $T(K) < M < 1$. The intersection points of $UH_K(M) - LH_K(M)$ and $M - UH_K(M)$ interior to $[1/K, 1]$ occur at $(S(K), (K-1)(K-2)/(2K^2-4K+1)^2)$ and $(T(K), (K-1)(K-2)/(3K-4)^2)$.*

Proof In each interval $[1/k, 1/(k-1)]$ for $k > 2$, by Proposition 20, the minimal value of $UH_K(M) - LH_K(M)$ is $(k-2)(K-k)/[k(k-1)(K-1)-K]$. By Proposition 20 of Rosenberg and Jakobsson (2008), the maximal value of $M - UH_K(M)$ is $1/[4k(k-1)]$.

We show that except if $k = K$ or $k = 2$, the minimum of $UH_K(M) - LH_K(M)$ exceeds the maximum of $M - UH_K(M)$. Writing $\zeta(K, k) = (k-2)(K-k)/[k(k-1)(K-1)-K] - 1/[4k(k-1)]$, substituting $c = K-k$ where $1 \leq c \leq K-3$, and simplifying, it follows that $\zeta(K, k) > 0$ if $\eta(c, k) = (4c-1)k^3 + (2-13c)k^2 + 9ck + c > 0$. For $k = 3$, $\eta(c, k) = 19c - 9$, which exceeds 0 because $c \geq 1$.

The difference $\eta(c, k+1) - \eta(c, k)$ is equal to $(12c-3)k^2 + (1-14c)k + 1$. Because $c \geq 1$, $12c-3 \geq 9c$. Consequently, $\eta(c, k+1) - \eta(c, k) \geq (9k-14)ck + k + 1$, which exceeds 0 for $k \geq 3$. Thus, because $\eta(c, 3) > 0$ and $\eta(c, k+1) > \eta(c, k)$ for $k \geq 3$, it follows that $UH_K(M) - LH_K(M)$ exceeds $M - UH_K(M)$ over $[1/k, 1/(k-1)]$ for $3 \leq k \leq K-1$.

Setting $k = K$, $UH_K(M) - LH_K(M) < M - UH_K(M)$ if $K(2K^2 - 4K + 1)M^2 - (4K^2 - 7K + 1)M + (2K-3) < 0$, that is, if M lies in $(1/K, S(K))$. The intersection points for $UH_K(M) - LH_K(M)$ and $M - UH_K(M)$ on $[1/K, 1/(K-1)]$ occur at $(1/K, 0)$ and $(S(K), (K-1)(K-2)/(2K^2-4K+1)^2)$.

Setting $k = 2$, $UH_K(M) - LH_K(M) < M - UH_K(M)$ if $(3K-4)M^2 + (7-5K)M + (2K-3) < 0$, that is, if M lies in $(T(K), 1)$. The intersection points for $UH_K(M) - LH_K(M)$ and $M - UH_K(M)$ on $[1/2, 1]$ occur at $(T(K), (K-1)(K-2)/(3K-4)^2)$ and $(1, 0)$. □

As $K \to \infty$, the intersection points at $S(K)$ and $T(K)$ approach $(0, 0)$ and $(2/3, 1/9)$, respectively. Thus, the limiting result agrees with Proposition 26 of Rosenberg and Jakobsson (2008).

5 Application to data

Our mathematical results are informative for examining homozygosity and the frequency of the most frequent allele in multiallelic population-genetic data. We considered the values of H, M, and K for 783 microsatellite loci in 1,048 individuals from worldwide human populations (Rosenberg *et al.* 2005), treating allele frequency estimates as parametric allele frequencies. To illustrate the effect of K on the bounds on H in terms of M and M in terms of H, we show results for two distinct values of K, $K = 7$ and $K = 15$.

Superimposing the graphs of $UM_{15}(H)$ and $UM_7(H)$ along with graphs of $UM_\infty(H)$ and $LM_\infty(H)$, we can see in Fig. 8 that for the 27 loci with $K = 7$ alleles, M and H tend to be greater than for the 33 loci with $K = 15$ alleles. In some cases, loci with $K = 15$ alleles have values of M and H that do not lie in the allowed region for loci with $K = 7$ alleles. Both for $K = 7$ and for $K = 15$, the region between $UM_K(H)$ and $LM_K(H)$ circumscribes the points plotted for loci with the given value of K more precisely than does the region between $UM_\infty(H)$ and $LM_\infty(H)$.

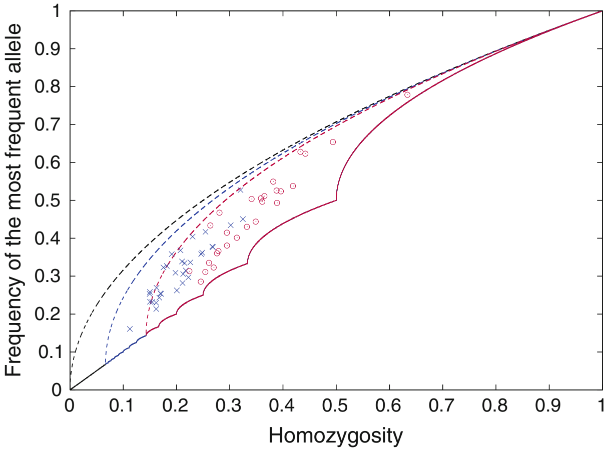

Fig. 8 Frequency of the most frequent allele and homozygosity for 27 loci with $K = 7$ alleles (purple circles) and 33 loci with $K = 15$ alleles (blue crosses). The plot shows the upper and lower bounds on M given H for $K = 7$ (purple) and $K = 15$ (blue) alleles, and in the case of unspecified K (black).

6 Discussion

By considering the value of the number of alleles at a locus in evaluating the relationship between homozygosity and the frequency of the most frequent allele, we have refined the range in which one of the two quantities must lie when the other is given. Our analysis extends the work of Rosenberg and Jakobsson (2008) on the case of unspecified K, and indeed, it can be used to obtain many of the earlier results by taking limits as $K \to \infty$.

Rosenberg and Jakobsson (2008) identified several examples in which the relationship between H and M could be used to understand the behavior of haplotype frequency statistics relying on H or M in tests of neutrality using haplotype frequencies in a population. Our refinements in range for H given M and M given H when K is specified provide an improved basis for interpreting population-genetic statistics based on H and M. These refinements can be particularly helpful in comparing inferences based on loci with different numbers of alleles, for which the precise relationship between H and M will differ. Additionally, they can also be useful in cases in which H, M, and K are measured from sample frequency distributions, as in our human microsatellite example. In this context, treating sample frequencies as parametric frequencies, the number of observations at a locus is an upper bound on K. As the range reduction for H given M or M given H owing to specification of K is maximal when K is small (Propositions 9 and 18), when H, M, and K are obtained from small samples, the bounds in this article can potentially provide a noticeable improvement compared to earlier work.

Our results trace largely to the fact that the allele frequencies at a locus constitute a set of nonnegative numbers with sum equal to 1. This property of allele frequencies has a variety of mathematical consequences not only for homozygosity and heterozygosity, but also for measures of differentiation across populations (Long and Kittles

2003; Hedrick 2005; Jost 2008) and measures of association among loci (Rosenberg and Calabrese 2004; Wray 2005; VanLiere and Rosenberg 2008). In each of these various situations, mathematical understanding of the underlying properties of population-genetic statistics provides insights into the ways in which the statistics behave when applied in data analysis. Thus, our work on homozygosity and the frequency of the most frequent allele, beyond providing new results on their specific relationship, can be seen in a broader context as a new addition to the mathematical theory of the fundamental statistics of population genetics.

Acknowledgments The authors gratefully acknowledge two reviewers for detailed comments, T. Pemberton for technical assistance, and the Burroughs Wellcome Fund and NIH grant GM081441 for financial support.

References

Hedrick PW (2005) A standardized genetic differentiation measure. Evolution 59:1633–1638

Jost L (2008) G_{ST} and its relatives do not measure differentiation. Mol Ecol 17:4015–4026

Long JC, Kittles RA (2003) Human genetic diversity and the nonexistence of biological races. Hum Biol 75:449–471

Rosenberg NA, Calabrese PP (2004) Polyploid and multilocus extensions of the Wahlund inequality. Theor Pop Biol 66:381–391

Rosenberg NA, Jakobsson M (2008) The relationship between homozygosity and the frequency of the most frequent allele. Genetics 179:2027–2036

Rosenberg NA, Mahajan S, Ramachandran S, Zhao C, Pritchard JK, Feldman MW (2005) Clines, clusters, and the effect of study design on the inference of human population structure. PLoS Genet 1:660–671

VanLiere JM, Rosenberg NA (2008) Mathematical properties of the r^2 measure of linkage disequilibrium. Theor Pop Biol 74:130–137

Weir BS (1996) Genetic data analysis II. Sinauer, Sunderland

Wray NR (2005) Allele frequencies and the r^2 measure of linkage disequilibrium: impact on design and interpretation of association studies. Twin Res Hum Genet 8:87–94