The Hitchhiking Effect of a Strongly Selected Substitution in Male Germline on Neutral Polymorphism in a Monogamy Population

Junrui Li[1,2], Kristan A. Schneider[3,4], Haipeng Li[1]*

1 Laboratory of Evolutionary Genomics, CAS Key Laboratory of Computational Biology, CAS-MPG Parter Institute for Computational Biology, Chinese Academy of Sciences, Shanghai, China, **2** Graduate School of the Chinese Academy of Sciences, Beijing, China, **3** University of Applied Sciences, Mittweida, Germany, **4** Department of Mathematics, University of Vienna, Vienna, Austria

Abstract

Comparative genomic studies suggest that a huge number of genes that show the strongest evidence for positive selection in human are testis- or sperm-specific genes, which are possibly due to germline selection. We propose a novel selection model in which the germlines of heterozygous males in a monogamous population are under natural selection. Under this model, we study the dynamics of a strongly selected substitution in the male germline and its hitch-hiking effect on the preexisting linked neutral polymorphism. We show that the expected heterozygosity at the neural locus is reduced by $\frac{8c}{s}(\frac{\alpha}{4})^{-8c/s}\Gamma(-\frac{8c}{s},\frac{4}{\alpha})$, where c is the recombination rate between selected and neutral locus, s is selective coefficient of advantageous allele, $\alpha=2Ns$ and N is diploid effective population size.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: lihaipeng@picb.ac.cn

Introduction

The hitchhiking effect [1] is commonly referred to a phenomenon that a selectively favored allele will change the frequencies of polymorphisms at linked loci on its way to fixation. This effect has been extensively studied in the last several decades [1–3] when the beneficial allele is codominant. Recently, Teshima and Przeworski examined the hitchhiking effect when the dominance coefficients of advantageous mutations are unknown [4]. To date, most models developed to characterize the hitchhiking effect assume that positive selection acts at the individual level through differential viabilities (Figure 1A). However, positive selection may also act on the male germline (Figure 1B) which might be common. In a comparative genomic study, based on the ratio of nonsynonymous substitutions per nonsynonymous site to synonymous substitutions per synonymous site (d_N/d_S), Nielsen *et al*. found a surprising large number of testis- or sperm-specific genes positively selected in the humans and chimpanzees [5]. A possible explanation for this is sperm competition [5].

Males produce tens of thousands of sperms, but only one of them can fertilize an egg. This severe competition implies potential positive selection that exert pressure on sperms activities in the spermatogenesis and fertilization process, for example, motility, acrosome reaction, penetration, apoptosis during spermatogenesis etc. Indeed, evidence for selection on the male germline has emerged taking the advantage of highly improved techniques in molecular biology. Several independent studies show genotype-dependent chances to fertilize eggs [6–8] and some reproductive

proteins are identified as targets of positive selection [9]. As data on germline selection accumulates, it is critical to develop a model to characterize the dynamics of positive directional selection due to sperm competition, and its effects on linked neutral polymorphism.

This article considers the consequences of strongly-selected substitutions in the male germline on preexisting linked neutral ploymorphism. For simplicity, we make the following assumptions: (1) absence of viability selection, i.e., individuals with different genotypes have the same viability, and (2) monogamy, i.e., competition is only among different sperm haplotypes from the same heterozygous male, and there is no competition between sperms from different males.

Two-Locus Model

We study a two-locus model that describes one selected and one linked neutral locus. At the selected locus, the ancestral allele is called a and the mutated beneficial allele A. According to our assumption, the viability of individuals with genotype $AA/Aa/aa$ are the same and can be assigned as 1, which is different from the traditional dominant additive fitness model, a special case of arbitrary dominance model that is well studied by Teshima *et al*. [4], but sperms with allele A have a selective advantage over sperms with allele a and their fitness are assigned as $1+s$, and 1, respectively. However, sperms from a homozygous male (AA or aa) have no advantage over other sperms from the same individual since they are identical. Thus their fitnesses are set to 1. The alleles at the neutral locus are denoted by B and b. We assume that a beneficial mutation A arise at time $t=0$ and replaces allele a
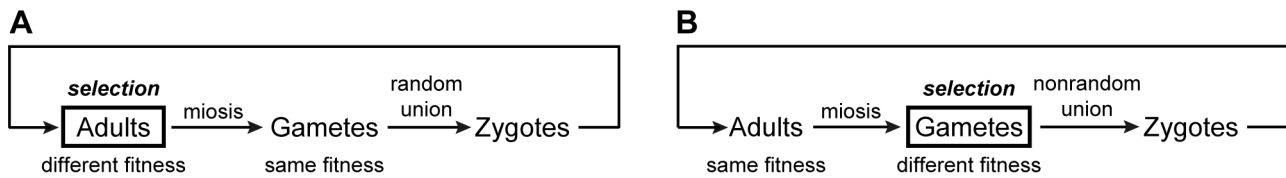
Figure 1. Model of positive natural selection in germline is different from that in individuals. (A) Individuals with different fitness are under selection but gametes have same fitness. (B) Individuals have same fitness but gametes with different fitness are under selection.
doi:10.1371/journal.pone.0071497.g001

subsequently. The fixation process of alleles A may alter heterozygosity levels at the linked neutral locus.

To analyze this model, we follow Ohta and Kimura's treatment that separated the dynamics of the selected locus and that of the neutral locus [2]. Afterwards, we consider only the frequency trajectory of the beneficial allele, which is acceptable when selection is so strong that the dynamics can be treated deterministically [10]. Furthermore, we adopt a modified moment-analysis method [3] to study the effect of the beneficial allele on the neutral locus.

Results

Frequency trajectory of beneficial allele

Assume the frequency of individuals with genotype AA, Aa and aa in the current generation is P, $2Q$, and R, respectively, where $P+2Q+R=1$. In terms of genotype frequencies, the allele frequencies p of A and q of a are as follows:

$$p=(2\times P+2Q)/2=P+Q, \qquad (1a)$$

$$q=(2\times R+2Q)/2=R+Q. \qquad (1b)$$

Note that $p+q=P+2Q+R=1$.

According to the sperm competition model we used, we can calculate the frequencies of allele A and a in the next generation. In total, there are nine possible mating pairs (Table 1). Since competition only occurs among sperms, there will be no selection at the individual level. So mating will be random

and these matings take place in proportion to the genotype frequencies. For example, the proportion of Aa and AA matings is $2Q\times P$. The frequencies of all the nine mating pairs are given in Table 1.

Recall that, competitions only happen in the gametes produced by the heterozygous male. Thus, a mating of male Aa with female AA produces proportionally $\frac{1+s}{2+s}AA$ and $\frac{1}{2+s}Aa$ zygotes, while the mating of male AA with female Aa produce $\frac{1}{2}AA$ and $\frac{1}{2}Aa$ zygotes. The frequency of zygotes produced by all possible mating pairs is listed in Table 1.

The genotype frequencies of AA, Aa, and aa zygotes at the next generation are denoted as P', $2Q'$, and R', respectively. These frequencies can be calculated as follows:

$$P'=P^2+PQ\left(\frac{4+3s}{2+s}\right)+2Q^2\left(\frac{1+s}{2+s}\right), \qquad (2a)$$

$$2Q'=PQ\left(\frac{4+s}{2+s}\right)+2PR+QR\frac{4+3s}{2+s}+2Q^2, \qquad (2b)$$

$$R'=R^2+RQ\left(\frac{4+s}{2+s}\right)+2Q^2\left(\frac{1}{2+s}\right). \qquad (2c)$$

Then, the frequencies of A and a, which are denoted as p' and q', respectively, can be calculated from genotype frequencies:

Table 1. Genotype frequency in next generation.

male×female	freq. of mating pair	freq. of AA zygotes	freq. of Aa zygotes	freq. of aa zygotes
$AA\times AA$	P^2	1	0	0
$AA\times Aa$	$2PQ$	$\frac{1}{2}$	$\frac{1}{2}$	0
$AA\times aa$	PR	0	1	0
$Aa\times AA$	$2PQ$	$\frac{1+s}{2+s}$	$\frac{1}{2+s}$	0
$Aa\times Aa$	$4Q^2$	$\frac{1+s}{2(2+s)}$	$\frac{1}{2}$	$\frac{1}{2(2+s)}$
$Aa\times aa$	$2QR$	0	$\frac{1+s}{2+s}$	$\frac{1}{2+s}$
$aa\times AA$	PR	0	1	0
$aa\times Aa$	$2QR$	0	$\frac{1}{2}$	$\frac{1}{2}$
$aa\times aa$	R^2	0	0	1

First two columns are all nine possible mating pairs and their frequencies. Last three columns are frequencies of zygotes produced by different mating pairs under germline selection.
doi:10.1371/journal.pone.0071497.t001

$$p' = (2 \times P' + 2Q')/2 = P + Q + \frac{Qs}{2(2+s)} = p + \frac{Qs}{2(2+s)}, \quad (3a)$$

$$q' = (2 \times R' + 2Q')/2 = R + Q - \frac{Qs}{2(2+s)} = q - \frac{Qs}{2(2+s)}, \quad (3b)$$

where p and q are the allele frequencies given in Equation (1).

From Equation (3a), we obtain the frequency change of the A allele after one generation of germline selection,

$$\Delta p = p' - p = \frac{Qs}{2(2+s)}. \quad (4)$$

In the traditional positive selection models, random union of gametes guaranties Hardy-Weinberg equilibrium, which means

$$P = p^2, \quad 2Q = 2p(1-p), \quad R = (1-p)^2. \quad (5)$$

But in the germline selection model, gametes from the same heterozygous male have different fitness, which leads to non-random union of gametes. Thus genotype frequencies deviate from the expectation of Hardy-Weinberg principle. However, we can use Hardy-Weinberg proportions as a good approximation to calculate frequency change of allele A:

$$\Delta p_A = p'_A - p = \frac{p(1-p)s}{2(2+s)}. \quad (6)$$

The difference between equation (4) and (6) arises from frequency differences of heterozygotes (Aa) after one generation of germline selection. Here, without loss of generality, we use frequencies of heterozygote Aa in (i+1)-th generation as an example to show that differences between frequencies of Aa heterozygotes calculated from equation (4) and (6) are a first order infinitesimal item of s, i.e., $O(s)$.

Equation (6) yields the frequency of allele A after one generation of germline selection as

$$p'_A = p + \frac{p(1-p)s}{2(2+s)}. \quad (7)$$

Since, $p = P + Q$, we can derive the deviation between Q' and $p'_A(1-p'_A)$ to be

$$Q' - p'_A(1-p'_A) = \frac{s}{4(s+2)^2}(sf(P,Q) + g(P,Q))$$
$$= \frac{sg(P,Q)}{16} + O(s^2) \approx \frac{sg(P,Q)}{16}, \quad (8)$$

a first order infinitesimal item of s, where $f(P,Q)$ and $g(P,Q)$ are polynomials of P and Q.

Then, equation (4) can be written as

$$\Delta p = \frac{Qs}{2(2+s)} = \frac{p(1-p)s + (Q - p(1-p))s}{2(2+s)}$$
$$= \frac{p(1-p)s}{2(2+s)} + O(s^2) \approx \frac{p(1-p)s}{2(2+s)}, \quad (9)$$

which means that Hardy-Weinberg frequencies are a good approximation. The deviation between frequency trajectories that are calculated iteratively by Equation (4) and Equation (9) is demonstrated in Figure 2. Given selective strength s, the absolute value of this deviation is positively correlated with the frequency of heterozygotes Aa (Figure S1). That means the more heterozygotes, the larger the deviation, and the deviation attains its maximum when the frequency of heterozygotes is maximized (results not shown). Of note, the maximum deviation increases with increasing selective coefficients s. However, the deviation is still very small even when s is as large as 0.1 (Figure 2). Moreover, difference in fixation times of these two frequency trajectories are also negligible (results not shown). Therefore, Equation (9) is a good approximation for the frequency trajectory of allele A.

If the selective advantage of allele A is large, the frequency change of allele A can be treated deterministically as long as its frequency is not very close to either 0 or 1 [11]. Given $0 < \varepsilon \ll 1$, the frequency of allele A at time t can be approximated by $x(t)$ that satisfies the differential equation:

$$\frac{dx(t)}{dt} = \frac{x(t)(1-x(t))s}{2(2+s)} \approx \frac{s}{4}x(t)(1-x(t)), \quad x(t_\varepsilon) = \varepsilon, \quad (10)$$

where we make a further approximation using the Taylor expansion of $\frac{1}{2(2+s)}$ and ignoring terms of order s^2 and higher.

The solution of this ordinary differential equation (ODE) is

$$x(t) = \frac{\varepsilon}{\varepsilon + (1-\varepsilon)e^{-s(t-t_\varepsilon)/4}}. \quad (11)$$

For convenience, by the substitution $\tau = t - t_\varepsilon$, the time it takes for allele A to reach quasi-fixation, i.e., the time needed to increase from frequency ε to $1-\varepsilon$, can be calculated to be

$$\hat\tau = -8\ln(\varepsilon)/s. \quad (12)$$

Hitchhiking effect on heterozygosity

In order to study the effect of the selected locus on the neutral one, we adopt the method of Stephan et al. [3], which is modified from Ohta and Kimura's moment-analysis method [2]. Ohta and Kimura divide the population into two parts: one part contains chromosomes carrying the advantageous allele A and the other part carrying the disadvantageous allele a [2]. Let p_1 be the frequency of allele B among chromosomes carrying A, and p_2 be the frequency of B among chromosomes carrying a. Then the frequency of allele B can be expressed by p_1 and p_2 as

$$p = p_1 x(\tau) + p_2(1 - x(\tau)). \quad (13)$$

Following Stephan et al.'s method [3], which distinguishes the situations that the beneficial mutation occurs on a B- or b-carrying chromosomes, we calculate a weighted expectation for an arbitrary polynomial function f of p_1 and p_2 as

$$\overline{E}(f) = p_{20}E(f|p_{10}=1) + (1-p_{20})E(f|p_{10}=0), \quad (14)$$
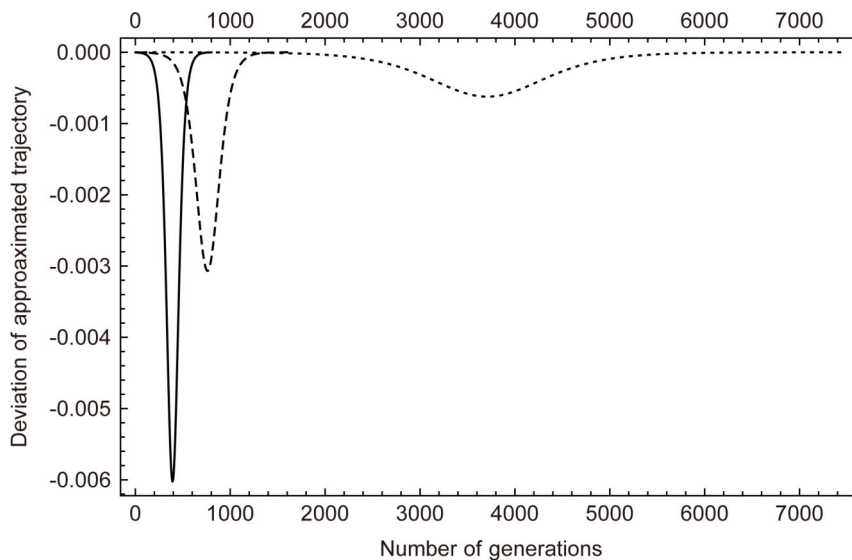
Figure 2. Deviation between HW approximated and real frequency trajectories of allele *A*. Solid line, dashed line, and dotted line are for selective coefficients $s = 0.1$, 0.05, 0.01, respectively.
doi:10.1371/journal.pone.0071497.g002

where p_{10} is the frequency p_1 at time $t = 0$, and p_{20} is the frequency p_2 at time $t = 0$. The former equals either one or zero, depending on whether the beneficial mutation occurred on a *B* or *b* background.

The expected heterozygosity H at *B/b* locus is defined by

$$H = \overline{E}(2p(1-p)), \tag{15}$$

where p is the frequency of allele *B* (Equation (17)).

A series of differential equations are derived and solved approximately following Stephan *et al.*'s method [3], and then we obtained the reduction in expected heterozygosity at the end of selection:

$$\frac{H_{1-\varepsilon}}{2p_{2\varepsilon}(1-p_{2\varepsilon})} = \frac{8c}{s}(\alpha/4)^{-8c/s}[\Gamma(-\frac{8c}{s}, \frac{4}{\alpha}) - \Gamma(-\frac{8c}{s}, \frac{4}{\alpha\varepsilon})], \tag{16}$$

where $\alpha = 2Ns$, c is recombination rate between selected locus and neutral locus, $H_{1-\varepsilon}$ and $2p_{2\varepsilon}(1-p_{2\varepsilon})$ are the expected heterozygosities at time when the frequencies of the allele *A* are ε and $1-\varepsilon$, respectively. Note, that this is exactly the same form as the one derived by Stephan *et al.* [3] with s replaced by $s/4$. Intuitively, this is not too surprising, because selection occurs only in heterozygote males. Since males account only for half of the population, and at most half of the males are heterozygous, selection is roughly four times less efficient than in a corresponding viability selection model. The reduction in expected heterozygosity of these two models is demonstrated in Figure 3.

If $\varepsilon \leq 1/\alpha$, then the Equation (16) can be approximated as

$$\frac{H_{1-\varepsilon}}{2p_{2\varepsilon}(1-p_{2\varepsilon})} = \frac{8c}{s}(\alpha/4)^{-8c/s}\Gamma(-\frac{8c}{s}, \frac{4}{\alpha}), \tag{17}$$

which imply that reduction in expected heterozygosity is only weakly dependent on ε. Figure 3 shows that this is a very accurate approximation for various choices of c.

Discussion

Exploring alternatives to adaptive evolution driven by differential viabilities, we proposed a model in which sperms from a heterozygous male are under natural selection. This model is different from the classic models since the male germline rather than individuals are selected. Notably, this model is also distinct from fertility selection [12–14] where different mating pairs have different fertility, as well as sexual selection [15,16] in which selection functions through asymmetrical mating preferences. A similar model which is mentioned as "meiotic drive" has been study by Chevin and Hospital [17], however, their model sticks to hitchhiking effect results from non-random segregation of chromosomes during meiosis rather than germline competition after meiosis. Thus, their results introduce an additional factor that is related to recombination rate to describe hitchhiking effect, while our results related to selective strength. Compared to other sperm competition models [18–20] where ployandrous populations are considered, this model mainly focuses on sperm competition in heterozygous males, and it is more suitable to describe the evolutionary dynamics of germline selection in monogamous population, for example the testis- or sperm-specific genes in human population [5]. Interestingly, recently evolved new genes are often testis-specific as documented by Vinckenbosch *et al.* [21], and Zhang *et al.* [22]. Moreover, these genes are often associated with strong adaptive signals [22]. As argued by Meiklejohn and Tao, such new genes may originate under the pressure of meiotic drive [23]. However, as the model discussed in this manuscript, such genes may adaptively emerge under sperm competition.

Here, we studied the dynamics of a strongly selected substitution in male germline and its effect on the preexisting linked neutral polymorphism. Due to the selection on male germline, random union of sperms and eggs are disrupted. Thus, genotype frequency and allele frequency are no longer in Hardy-Weinberg equilibrium. However, the deviation of genotype frequencies from Hardy-Weinberg proportions are negligible (Figure 2) and we found that Hardy-Weinberg frequencies calculated from genotype frequencies can still produce a good approximation for allele frequencies.
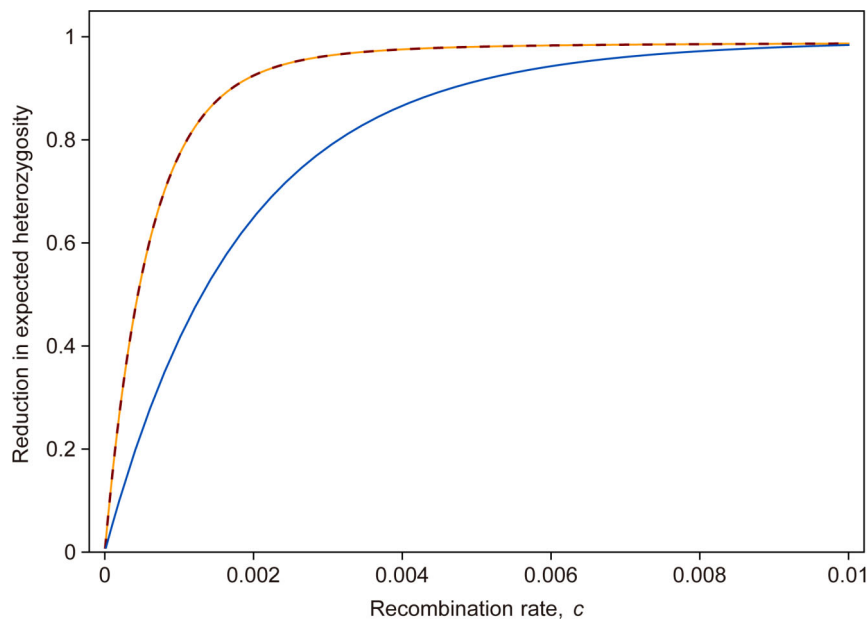
Figure 3. Reduction in expected heterozygosity at the end of selection process. Solid orange line shows the expected heterozygosity calculated from Equation (16). Dashed line is the approximated result (Equation (17)). Solid blue line demonstrates the reduction in heterozygosity according to result of Stephan *et al.* (1992). We assumed $s = 0.02$.
doi:10.1371/journal.pone.0071497.g003

The dynamics of the selected allele and its effect on linked neutral polymorphism are similar to the results derived by Stephan *et al.* [3], except that the selective coefficient in our formulas is $s/4$ rather than s, which means that germline selection is weaker than individual selection, given the same selection coefficient. This is reasonable because selection only occurs in heterozygous males. However, the selective strength in germlines may usually be much larger than that in individuals, which results in faster evolution of reproductive genes [24–26].

Notably, the hitchhiking effect of a beneficial mutation selected in the male germline is approximately identical to that of a beneficial mutation under viability selection in the absence of dominance, however, having just a quarter of the selective advantage. Hence, in genome-wide scans for traces of selection, positively selected targets in the male germline might be incorrectly inferred as candidates under moderate viability selection. Even further, in such situations between-population comparison studies designed to verify adaptive evolution might not provide evidence for viability selection and hitchhiking patters might be incorrectly explained by demographic effects. Hence, in populations in which germeline selection could potentially occur, experimental setups need to be adjusted accordingly to correctly infer the mechanisms of selection.

As a basic model for germline selection, the model can be extended to some more complex models. For example, fitness of individuals can also be considered in the model. Several recent studies about a substitution in fibroblast growth factor receptor 2 (*FGFR2*) in the male germline showed that a selective germline advantage leads to unexpected high mutant prevalence, although this substitution causes defects to the descendants [27,28]. It indicates that the substitution is positively selected in germlines, but negatively selected in individuals, which could

lead to overall balancing selection. Hence, models combining these two aspects may give a better prediction for diseases and worth further investigations. Notably our model assumes a monogamous population. However, it can also be extended to polyandry populations where females mate with different males in a short time period, which certainly leads to more severe sperm competition. Such phenomenon can be frequently found in social insects, for example bees [29]. In such cases the reduction of linked neutral polymorphism may accordingly be more severe.

Supporting Information

Figure S1 Correlation between the deviation and HW frequency of heterozygotes. Horizontal axis is the number of iteration. HW frequency of heterozygotes is demonstrated by the solid line and left vertical axis. The deviation between HW approximated and the real frequency trajectories of allele *A* is demonstrated by the dashed line and right vertical axis. We assumed $s = 0.1$.
(TIF)

Acknowledgments

Author Contributions

Conceived and designed the experiments: JL HL. Performed the experiments: JL KAS HL. Analyzed the data: JL KAS HL. Contributed reagents/materials/analysis tools: JL KAS HL. Wrote the paper: JL KAS HL.

References

1. Maynard Smith J, Haigh J (1974) The hitch-hiking effect of a favourable gene. Genetical Research 23: 23–35.
2. Ohta T, Kimura M (1975) The effect of selected linked locus on heterozygosity of neutral alleles (hitch-hiking effect). Genetical Research 25: 313–326.
3. Stephan W, Wiehe THE, Lenz MW (1992) The effect of strongly selected substitutions on neutral polymorphism - analytical results based on diffusion-theory. Theoretical Population Biology 41: 237–254.
4. Teshima KM, Przeworski M (2006) Directional positive selection on an allele of arbitrary dominance. Genetics 172: 713–718.
5. Nielsen R, Bustamante C, Clark AG, Glanowski S, Sackton TB, et al. (2005) A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol 3: e170.
6. Clark AG, Begun DJ, Prout T (1999) Female×male interactions in drosophila sperm competition. Science 283: 217–220.
7. Palumbi SR (1999) All males are not created equal: Fertility differences depend on gamete recognition polymorphisms in sea urchins. Proc Natl Acad Sci U S A 96: 12632–12637.
8. Chow CY,Wolfner MF, Clark AG (2010) The genetic basis for male×female interactions underlying variation in reproductive phenotypes of drosophila. Genetics 186: 1355–1365.
9. Pujolar JM, Pogson GH (2011) Positive darwinian selection in gamete recognition proteins of strongylocentrotus sea urchins. Mol Ecol 20: 4968–4982.
10. Kaplan NL, Hudson RR, Langley CH (1989) The "hitchhiking effect" revisited. Genetics 123: 887–899.
11. Kurtz TG (1970) Solutions of ordinary differential equations as limits of pure jump markov processes. Journal of Applied Probability 7: 49–58.
12. Hadeler K, Liberman U (1975) Selection models with fertility differences. Journal of Mathematical Biology 2: 19–32.
13. Feldman MW, Christiansen FB, Liberman U (1983) On some models of fertility selection. Genetics 105: 1003–1010.
14. Nagylaki T (1987) Evolution under fertility and viability selection. Genetics 115: 367–375.
15. Kirkpatrick M (1982) Sexual selection and the evolution of female choice. Evolution: 1–12.
16. Heisler IL, Curtsinger JW (1990) Dynamics of sexual selection in diploid populations. Evolution: 1164–1176.
17. Chevin LM, Hospital F (2006) The hitchhiking effect of an autosomal meiotic drive gene. Genetics 173: 1829–1832.
18. Birkhead TR, Møller AP (1998) Sperm competition and sexual selection. Academic Press.
19. Parker GA (1970) Sperm competition and its evolutionary consequences in the insects. Biological Reviews 45: 525–567.
20. Parker GA (1998) Sperm Competition and Sexual Selection. Academic Press, 3–54 pp.
21. Vinckenbosch N, Dupanloup I, Kaessmann H (2006) Evolutionary fate of retroposed gene copies in the human genome. Proc Natl Acad Sci U S A 103: 3220–3225.
22. Zhang YE, Vibranovski MD, Krinsky BH, Long M (2010) Age-dependent chromosomal distribution of male-biased genes in drosophila. Genome Research 20: 1526–1533.
23. Meiklejohn CD, Tao Y (2010) Genetic conict and sex chromosome evolution. Trends in Ecology and Evolution 25: 215–223.
24. Metz EC, Robles-Sikisaka R, Vacquier VD (1998) Nonsynonymous substitution in abalone sperm fertilization genes exceeds substitution in introns and mitochondrial dna. Proc Natl Acad Sci U S A 95: 10676–10681.
25. Wyckoff GJ, Wang W, Wu CI (2000) Rapid evolution of male reproductive genes in the descent of man. Nature 403: 304–309.
26. Swanson WJ, Vacquier VD (2002) The rapid evolution of reproductive proteins. Nat Rev Genet 3: 137–144.
27. Goriely A, McVean GA, Rojmyr M, Ingemarsson B, Wilkie AO (2003) Evidence for selective advantage of pathogenic FGFR2 mutations in the male germ line. Science 301: 643–646.
28. Choi SK, Yoon SR, Calabrese P, Arnheim N (2008) A germ-line-selective advantage rather than an increased mutation rate can explain some unexpectedly common human disease mutations. Proc Natl Acad Sci U S A 105: 10143–10148.
29. Franck P, Coussy H, Le Conte Y, Solignac M, Garnery L, et al. (1999) Microsatellite analysis of sperm admixture in honeybee. Insect Mol Biol 8: 419–421.