

Effective migration rates

Arthur Zwaenepoel

Here is a solid and fairly transparent derivation of the formulae in Zwaenepoel, Sachdeva, and Fraïsse (2024), without the haplodiplontic complications.

Assume a source population A with known allele frequencies q_i^* $i = 1 \dots L$ for variants that are locally deleterious in a sink population B . We write q_i for the corresponding allele frequencies in B . Migration occurs from A to B at rate m . The mean fitness among resident individuals is

$$\overline{W}_B = \prod_{i=1}^L (p_i^2 + 2p_i q_i (1 - s_i h_i) + q_i^2 (1 - s_i)) \quad (1)$$

$$= \prod_{i=1}^L (1 - 2s_i h_i p_i q_i - s_i q_i^2) \quad (2)$$

$$\approx \exp \left(- \sum_i^L 2s_i h_i p_i q_i + s_i q_i^2 \right) \quad (3)$$

The relative fitness of a migrant individual in the resident background is:

$$W_M \approx \frac{1}{\overline{W}_B} \exp \left(- \sum_i^L 2s_i h_i p_i q_i^* + s_i q_i^{*2} \right) \quad (4)$$

write (to avoid double exponents, LaTeX...) q_0 for q^* , we get factors for each locus of the form

$$sq^2 + 2shpq - sq_0^2 - 2shp_0q_0 = sq^2 + 2shq - 2shq^2 - sq_0^2 - 2shq_0 + 2shq_0^2 \quad (5)$$

$$= s(q^2(1 - 2h) - q_0^2(1 - 2h) + 2h(q - q_0)) \quad (6)$$

$$= s \left[(1 - 2h)(q^2 - q_0^2) + 2h(q - q_0) \right] \quad (7)$$

$$= s \left[(1 - 2h)(q - q_0)(q + q_0) + 2h(q - q_0) \right] \quad (8)$$

$$= s(q - q_0) \left[(1 - 2h)(q + q_0) + 2h \right] \quad (9)$$

$$= -s(q_0 - q) \left[(1 - 2h)(q + q_0) + 2h \right] \quad (10)$$

so we get

$$W_M \approx \exp \left(- \sum_i^L s(q_i^* - q_i) \left[(1 - 2h_i)(q_i + q_i^*) + 2h \right] \right)$$

This is also the expression in Himani's notes on heterosis.

Now consider the first generation (F1). The mean fitness among F1's is

$$\tilde{W}_0 \approx \exp \left(- \sum_i^L s_i h_i (p_i q_i^* + p_i^* q_i) + s_i q_i^* q_i \right)$$

relative to the resident, we get factors of the form

$$sq^2 + 2shpq - sh(pq_0 + p_0q) + sq_0q = -s(q_0 - q)(h + q(1 - 2h))$$

So

$$W_0 \approx \exp \left(- \sum_i^L s_i (q_i^* - q_i)(h_i + q_i(1 - 2h_i)) \right) \quad (11)$$

$$= \exp \left(- \sum_i^L s_i (q_i^* - q_i)(h_i(p_i - q_i) + q_i) \right) \quad (12)$$

Under the assumption that all migrants and their descendants cross with residents, and that the effects of selection on allele frequencies within these F1s, BC1s, *etc.* is negligible, than we get in the k th BC generation at a generic locus

$$q^{(k)} = q + \frac{1}{2^{k+1}}(q^* - q) \quad (13)$$

The proportion of heterozygotes is hence

$$pq^{(k)} + p^{(k)}q = pq^{(k)} + q - q^{(k)}q \quad (14)$$

$$= q^{(k)}(p - q) + q \quad (15)$$

$$= \left(q + \frac{1}{2^{k+1}}(q^* - q) \right) (p - q) + q \quad (16)$$

$$= 2pq + \frac{1}{2^{k+1}}(q^* - q)(p - q) \quad (17)$$

The proportion of homozygotes is

$$q^{(k)}q = q^2 + \frac{1}{2^{k+1}}(q^* - q)q \quad (18)$$

In general for the k th generation backcross, we obtain

$$\tilde{W}_k \approx \exp \left(- \sum_i^L s_i h_i \underbrace{\left(2p_i q_i + \frac{1}{2^k}(p_i - q_i)(q_i^* - q_i) \right)}_{\text{heterozygotes}} + s_i \underbrace{\left(q_i^2 + \frac{1}{2^k}q_i(q_i^* - q_i) \right)}_{\text{homozygotes}} \right)$$

so, dividing by the mean resident fitness, we get factors of the form

$$2shpq + sq^2 - 2shpq - sh(p - q)(q_0 - q)/2^k - sq^2 - sq(q_0 - q)/2^k \quad (19)$$

$$= -sh(p - q)(q_0 - q)/2^k - sq(q_0 - q)/2^k \quad (20)$$

$$= - \frac{s(q_0 - q)(h(p - q) + q)}{2^k} \quad (21)$$

for the heterozygotes' part, and of the form

$$sq^2 - sq^2 - q(q_0 - q)/2^k = q(q_0 - q)/2^k \quad (22)$$

for the part coming from homozygotes. Putting everything together, this yields:

$$W_k = \exp \left(-\frac{1}{2^k} \sum_i^L s_i(q_i^* - q_i)(h_i(p_i - q_i) + q_i) \right) \quad (23)$$

The gff is then

$$g = W_M \prod_k^\infty W_k \quad (24)$$

$$= \exp \left(-\sum_i^L s_i(q_i^* - q_i) [(1 - 2h_i)(q_i + q_i^*) + 2h] \right) \exp \left(-2 \sum_i^L s_i(q_i^* - q_i)(h_i(p_i - q_i) + q_i) \right) \quad (25)$$

$$= \exp \left(-\sum_i^L s_i(q_i^* - q_i) [(1 - 2h_i)(q_i + q_i^*) + 2h] + 2s_i(q_i^* - q_i)(h_i(p_i - q_i) + q_i) \right) \quad (26)$$

$$= \exp \left(-\sum_i^L s_i(q_i^* - q_i) [(1 - 2h_i)(q_i + q_i^*) + 2h + 2(h_i(p_i - q_i) + q_i)] \right) \quad (27)$$

$$= \exp \left(-\sum_i^L 2s_i(q_i^* - q_i) \left[2h + \frac{1}{2}(3q_i - q_i^*)(1 - 2h) \right] \right) \quad (28)$$

in the absence of dominance, this becomes

$$g = \exp \left(-\sum_i^L s(q_i^* - q_i) \right) \exp \left(-2 \sum_i^L \frac{s_i}{2}(q_i^* - q_i) \right) \quad (29)$$

$$= \exp \left(-2 \sum_i^L s(q_i^* - q_i) \right) \quad (30)$$

which is the same as the result for a haploid model.

When migration happens after selection, the factor W_M should be dropped. In general it appears more transparent to write the gff as in eq. 25.

When allele frequencies fluctuate, one can take $\mathbb{E}[g]$ as an approximation, dropping terms that are $O(s^2)$, assuming LE, and assuming the source and island allele frequencies are independent. The gff will then be a function of the first two moments of the allele frequency

distribution (only the first moment in the absence of dominance).

$$\mathbb{E}[g] \approx \mathbb{E}[W_M] \prod_k^{\infty} \mathbb{E}[W_k] \quad (31)$$

$$\approx \exp \left(- \sum_i^L \mathbb{E} [s_i (q_i^* - q_i) [(1 - 2h_i)(q_i + q_i^*) + 2h]] \right) \quad (32)$$

$$\times \exp \left(-2 \sum_i^L \mathbb{E} [s_i (q_i^* - q_i) (h_i(p_i - q_i) + q_i)] \right) \quad (33)$$

$$\approx \exp \left(- \sum_i^L s_i ((q_i^* - \mathbb{E}[q_i]) - (1 - 2h_i) [pq_i^* - \mathbb{E}[pq_i]]) \right) \quad (34)$$

$$\times \exp \left(-2 \sum_i^L s_i (h_i(q_i^* - \mathbb{E}[q_i]) - (1 - 2h_i) [p_i^* \mathbb{E}[q_i] - \mathbb{E}[pq_i]]) \right) \quad (35)$$

This is what's in equation 10 of our genetics paper.

Question: how do random allele frequencies affect the assumptions regarding the allele frequencies in the BC generations? Are we safe in using the $\mathbb{E}[g]$ where g is obtained conditional on known allele frequencies?

Two-locus theory suggests that the effective migration rate at a neutral locus linked to a selected one in a diploid is given by

$$m_e = m \left(1 - \frac{s(q - q^*)(h + (1 - 2h)q)}{r + m + s(h + (1 - 2h)q)(p - q)} \right) \quad (36)$$

$$= m \left(1 - \frac{s(q - q^*)(h(p - q) + q)}{r + m + s(h(p - q) + q)(p - q)} \right) \quad (37)$$

$$= m \exp \left(- \frac{s(q - q^*)(h(p - q) + q)}{r + m + s(h(p - q) + q)(p - q)} \right) \quad (38)$$

For the case where $r \approx 0.5 \gg m, s$ this yields the same prediction as above (barring the contribution from the diploid migrant).

If we take the exponent in the last line, and expand in powers of s , we find

$$- \frac{s}{m + r} (2hq^2 - 2hqq_0 - hq + hq_0 - q^2 + qq_0) + O(s^2)$$

Taking expectations and rearranging, we obtain to first order in s

$$- \frac{s}{m + r} (h(q^* - \mathbb{E}[q]) - (1 - 2h) [p^* \mathbb{E}[q] - \mathbb{E}[pq]])$$

This suggests a gff approximation of the form

$$\mathbb{E}[g(x)] \approx \mathbb{E}[W_M] \prod_k^\infty \mathbb{E}[W_k] \quad (39)$$

$$\approx \exp \left(- \sum_i^L s_i ((q_i^* - \mathbb{E}[q_i]) - (1 - 2h_i) [pq_i^* - \mathbb{E}[pq_i]]) \right) \quad (40)$$

$$\times \exp \left(- \sum_i^L \frac{s_i (h_i (q_i^* - \mathbb{E}[q_i]) - (1 - 2h_i) [p_i^* \mathbb{E}[q_i] - \mathbb{E}[pq_i]])}{m + r(x, x_i)} \right) \quad (41)$$

However, this ignores the fact that we often will have $m = O(s)$ and $r = O(s)$ as $s \rightarrow 0$. Assuming $m = O(s)$ and $r = O(s)$, we get

$$-s \left(\frac{qq_0 - q^2 + hq_0 - hq - 2hqq_0 + 2hq^2}{m + r + s(4hq^2 - 4hq + h - 2q^2 + q)} \right) + O(s^2)$$

Taking expectations is now more tricky because of the q 's in the denominator. Proceeding naively with the numerator and denominator separately, we get

$$- \frac{s(h(q^* - \mathbb{E}[q]) - (1 - 2h)[p^* \mathbb{E}[q] - \mathbb{E}[pq]])}{m + r + s(h - \mathbb{E}[q] + (1 - 2h)2\mathbb{E}[pq])}$$

This suggests a gff approximation of the form

$$\mathbb{E}[g(x)] \approx \mathbb{E}[W_M] \prod_k^\infty \mathbb{E}[W_k] \quad (42)$$

$$\approx \exp \left(- \sum_i^L s_i ((q_i^* - \mathbb{E}[q_i]) - (1 - 2h_i) [pq_i^* - \mathbb{E}[pq_i]]) \right) \quad (43)$$

$$\times \exp \left(- \sum_i^L \frac{s_i (h_i (q_i^* - \mathbb{E}[q_i]) - (1 - 2h_i) [p_i^* \mathbb{E}[q_i] - \mathbb{E}[pq_i]])}{m + r(x, x_i) + s_i (h_i - \mathbb{E}[q_i] + (1 - 2h_i)2\mathbb{E}[pq_i])} \right) \quad (44)$$

which is equation 5 in the Genetics paper.

Consider $h = 1/2$ to obtain a haploid/no dominance model:

$$\mathbb{E}[g(x)] \approx \mathbb{E}[W_M] \prod_k^\infty \mathbb{E}[W_k] \quad (45)$$

$$\approx \exp \left(- \sum_i^L s_i (q_i^* - \mathbb{E}[q_i]) \right) \exp \left(- \sum_i^L \frac{s_i h_i (q_i^* - \mathbb{E}[q_i])}{m + r(x, x_i) + s_i h_i (1 - 2\mathbb{E}[q_i])} \right) \quad (46)$$

$$\text{where } h_i = 1/2 \quad (47)$$

What does this entail for a single selected haploid locus at equilibrium frequency m/s , fixed

in the mainland?

$$\mathbb{E}[g(x)] \approx \exp(-s(1 - m/s)) \exp\left(-\frac{s(1 - m/s)}{m + r + s(1 - 2m/s)}\right) \quad (48)$$

$$\approx \exp(-(s - m)) \exp\left(-\frac{s - m}{r + s - m}\right) \quad (49)$$

$$\approx \exp\left(-(s - m) \left(1 + \frac{1}{r + s - m}\right)\right) \quad (50)$$

$$\approx \exp\left(\frac{(m - s)(1 - m + r + s)}{r + s - m}\right) \quad (51)$$

$$\text{where } m < s \quad (52)$$

as $m \rightarrow 0$, this becomes

$$\mathbb{E}[g] = \exp\left(\frac{-s(1 - r - s)}{r + s}\right) \approx \exp\left(\frac{-s}{r + s}\right) \approx 1 - \frac{s}{r + s} = \frac{r}{r + s} \quad (53)$$

Which is the result of Petry (1983).

Two island coalescent with unidirectional migration

References

- Petry, Doug. 1983. “The Effect on Neutral Gene Flow of Selection at a Linked Locus.” *Theoretical Population Biology* 23 (3): 300–313.
- Zwaenepoel, Arthur, Himani Sachdeva, and Christelle Fraïsse. 2024. “The Genetic Architecture of Polygenic Local Adaptation and Its Role in Shaping Barriers to Gene Flow.” *Genetics*, iyae140.