Genet. Res., Camb. (1999), **74**, *pp.* 81–85. Printed in the United Kingdom © 1999 Cambridge University Press

SHORT PAPER

Speed congenics: accelerated genome recovery using genetic markers

P. M. VISSCHER*

Institute of Ecology and Resource Management, University of Edinburgh, Edinburgh EH9 3JG, Scotland, UK

(*Received 30 October 1998 and in revised form 16 January 1999*)

Summary

Genetic markers throughout the genome can be used to speed up 'recovery' of the recipient genome in the backcrossing phase of the construction of a congenic strain. The prediction of the genomic proportion during backcrossing depends on the assumptions regarding the distribution of chromosome segments, the population structure, the marker spacing and the selection strategy. In this study simulation was used to investigate the rate of recovery of the recipient genome for a mouse, *Drosophila* and *Arabidopsis* genome. It was shown that an incorrect assumption of a binomial distribution of chromosome segments, and failing to take account of a reduction in variance in genomic proportion due to selection, can lead to a downward bias of up to two generations in the estimation of the number of generations required for the formation of a congenic strain.

1. Introduction

The time taken to construct a congenic mouse strain has been shown to be greatly accelerated when genetic markers are used during the backcross phase to speed up the 'recovery' of the recipient genome (Markel *et al.*, 1997; Wakeland *et al.*, 1997). Markel *et al.* (1997) showed both theoretical and empirical results to demonstrate that 'speed congenics', i.e. the accelerated creation of congenic strains using multiple markers throughout the genome, works. However, the theoretical section includes a number of questionable assumptions which may give a biased view of what can be achieved in practice. In particular: (i) the authors assume that the distribution of homozygous versus heterozygous segments in the genome follows a binomial distribution, and that the chromosome segments segregate independently, (ii) the authors incorrectly use the truncation point of the normal distribution to calculate response to selection, rather than the mean of the selected group, (iii) the author's derivations are based on a large sample of *selected parents*, yet the tabulated example is based on a single selected male and (iv) the authors ignore the distinction between a marker-based estimate of the proportion of recipient genome and its true value. The first assumption was also made by Hillel *et al.* (1990), who derived that the recipient genome could be recovered in as few as two generations of backcrossing in poultry populations.

In this short note I will take these issues in turn, and show how they affect the variation in genomic proportion during a backcross breeding programme. In addition, a more realistic prediction of the possible achievements of a marker-aided backcross introgression programme using mouse lines is obtained through simulation. It is shown that in practice it would take about two generations of backcrossing more than the previous predictions by Markel *et al.* (1997) suggest before the recipient genome is sufficiently ($> 99{\cdot}5\,\%$) recovered. Finally, simulation results for two species, *Drosophila* and *Arabidopsis*, are presented, to demonstrate how the genome recovery depends on the genome size and number of chromosomes.

2. Theory and simulation results

(i) *Distribution of genomic proportion*

In the absence of selection, the theory underlying the distribution of homozygous and heterozygous

* Tel: +44 (0)131 535 4052. Fax: +44 (0)131 667 2601. e-mail: peter.visscher@ed.ac.uk

Table 1. *Mean and standard deviation* ($\times 100$) *of the proportion of the genome originating from the recipient line during backcrossing, for a mouse genome*

		Theory[a]			Simulation	
Generation	Mean	SD	m[b]	SD[c]	SD[d]	SD[e]
F1	50·00	0	0	0	0	0
N2	75·00	4·59	30	3·54	4·57	4·58
N3	87·50	3·76	22	2·50	2·96	3·76
N4	93·75	2·72	21	1·77	1·92	2·71
N5	96·88	1·87	22	1·25	1·24	1·87
N6	98·44	1·27	24	0·88	0·79	1·26
N7	99·22	0·85	27	0·63	0·47	0·85
N8	99·61	0·57	30	0·44	0·27	0·57
N9	99·81	0·39	33	0·32	0·15	0·38
N10	99·90	0·26	36	0·22	0·08	0·25

[a] Using theory from Hill (1993).
[b] Equivalent number of independent segments giving the same variance in genomic proportion.
[c] Using the assumption of Markel *et al.* (1997).
[d] Randomly selecting 1 male out of 40. SD based on 1000 simulations.
[e] Randomly selecting 100 males out of 400. SD based on 1000 simulations.

chromosome segments in a backcross breeding programme is well understood (e.g. Stam & Zeven, 1981; Hospital *et al.*, 1992; Hill, 1993; Visscher *et al.*, 1996), as is the use of genetic markers to estimate the genomic proportion in an individual (Visscher, 1996). Genomic proportion is the relative proportion of the genome of an individual that originates from either the donor or the recipient line. Recently, good agreement between empirical results and simulation studies in the reduction of the proportion of the donor genome in a mouse introgression programme was found (Wakeland *et al.*, 1997).

Due to linkage and recombination, chromosome segments do not segregate independently, and the distribution of genomic proportion is not binomial (Stam & Zeven, 1981; Hill, 1993; Visscher, 1996). In Table 1 I show the predicted standard deviation in genomic proportion during backcrossing, assuming a 'mouse genome' (19 autosomes with a total length of 1493 cM, from the 1998 Jackson laboratory linkage map) and the Haldane mapping function, using theoretical methods (Hill, 1993). The predicted variation is substantially larger than that predicted by Markel *et al.* (1997), whose predictions are also shown in Table 1. Note that the standard deviation of the proportion of heterozygosity (which was used by Markel *et al.*) is twice the standard deviation of genomic proportion. Table 1 also shows the equivalent number of independent segments that gives the same variation in genomic proportion, using the binomial distribution as in Markel *et al.* (1997). It can be seen that the number of independent segments is closer to 30 than the 50 that was assumed by Markel *et al.*

(ii) *Response to selection*

Markel *et al.* (1997) used a selected fraction of 5% (1 in 20 males selected), and assume that the best male is 1·65 standard deviations (SD) above (or below, depending on the selection criterion) the mean. However, the mean of a selected group that is at least 1·65 SD above the mean, i.e. the standardized selection intensity, is 2·06 SD (e.g. Falconer & Mackay, 1996). Therefore, the calculations of Markel *et al.* are really based on a selected fraction of about 12·5% (giving a selection intensity of 1·65 SD). The above selection intensities are based on selection from a very large population. If only a single parent is selected from a small group of candidates, the selection intensity is reduced relative to selecting the same proportion from a large population (e.g. Falconer & Mackay, 1996). In that case a superiority of a single selected parent of 1·65 SD corresponds to a selection of about 1 in 13 males, and the selection intensity of selecting 1 out of 20 males is 1·867 SD (Falconer & Mackay, 1996). Hence, table 3 of Markel *et al.* is incorrect, and the potential reduction in the proportion of donor line segments in their table should be larger.

(iii) *Selecting a single parent*

A mouse genome (19 autosomes, using map lengths as described previously) was simulated, assuming Haldane's mapping function to generate crossovers. A single male from an inbred line was mated to 10 females from a different inbred line, each producing 8 progeny (4 males and 4 females). Choice of a single male parent out of the 40 candidates was done at random, and 1000 replicates population were simulated for 5 generations of backcrossing. Genomic proportion was calculated by tracing back the origin of each genomic location to one of the two founder populations. Hence, the figures in Table 1 are for observed genomic proportions, and not for estimates of the genomic proportion based on a small set of markers. For further results of the simulation process, see Visscher *et al.* (1996). The results in Table 1 show that relative to the predicted variation in genomic proportion, the observed variation is much reduced. This is due to the selection of a small number of parents (a single parent in these simulations), analogous to the phenomenon of genetic drift. Simulations using more parents (100 males selected) showed that the observed variation was as expected from theory (Table 1). The impact of a small number of parents can be explained by considering a single (marker) locus only. If the F1 generation from a cross between inbred lines is *Mm* at this locus, then a randomly

Table 2. *Mean and standard deviation* ($\times 100$) *of genomic proportion during backcrossing when a single selected male is used each generation. Results from 1000 replicates of simulation, and from predictions of Markel et al. (1997)*

	Proportion ($\times 100$) of genome from recipient population					
	Simulation			Prediction from Markel *et al.* (1997)[a]		
Generation	Mean	SD	Best male[b]	Mean	SD	Best male
N2	75·03	4·58	82·01	75·00	3·54	81·60
N3	90·99	2·45	94·35	90·80	2·14	94·80
N4	97·19	1·19	98·31	97·40	1·14	99·53
N5	99·15	0·54	99·30	99·77	0·34	100
N6	99·65	0·29	99·53	100	0	100
N7	99·77	0·21	99·61	100	0	100
N8	99·80	0·18	99·65	100	0	100
N9	99·82	0·16	99·69	100	0	100
N10	99·84	0·15	99·72	100	0	100

[a] Assuming a selection intensity of 1·867 standard deviations, corresponding to selecting 1 out of 20 individuals. Markel *et al.* (1997) used a selection intensity of 1·65 SD in their calculations.

[b] A single male was mated to 10 females, each producing 8 progeny. Of the male progeny that were heterozygous for the allele to be introgressed, a single male was selected based on the estimate of the proportion of its genome from the recipient population. The latter selection step corresponds to a selected proportion of approximately 1 in 20.

selected parent in the N2 generation is either *Mm* or *mm* (assuming that the recipient line is *mm*), each with a probability of 1/2. Genotype classes *Mm* and *mm* correspond to a proportion of the recipient 'genome' of 0·5 and 1·0, respectively. The variation in genomic proportion among the N3 progeny of this parent is either 1/16 (parent = *Mm*) or 0·00 (parent = *mm*). Hence, the average observed variance in genomic proportion in the N3 generation is only $(1/2)(1/16) = 1/32 = 2/64$. With a large number of parents in the N2 generation, both *Mm* and *mm* genotypes are represented, and the variance in the N3 generation among their progeny is $(1/4)(1/4)(1-1/4) = 3/64$ (e.g. Hospital *et al.*, 1992; Visscher *et al.*, 1996). This same principle applies to many linked loci, although the difference in variation in genomic proportion between the case of a single and many selected parents becomes smaller.

(iv) *Variation in genomic proportion when introgressing a gene and using markers*

Table 2 shows simulation results (1000 replicated populations) using the same population structure as before, i.e. 1 male mated to 10 females, producing 80 progeny, but now simultaneously introgressing a gene. An identified allele at position 30 cM on chromosome 1 was introgressed from an inbred donor population, while actively selecting against the rest of the donor genome using the estimated proportion of the genome which originated from the recipient line as selection criterion. Eight evenly spaced markers per chromosome were used to identify donor and recipient segments, corresponding to a marker spacing of about 10 cM, and these markers were used to estimate the proportion of the genome which originated from the recurrent line (Visscher *et al.*, 1996). Clearly the reduction in variance in genomic proportion has a large impact on how quickly a congenic strain can be formed. The simulation results from Table 2 are close to the results from empirical data shown by Markel *et al.* (1997). For example, our simulations show that the best male in generation N5 has, on average, a proportion of genes from the recipient line of 99·30%. For different crosses of inbred lines, Markel *et al.* empirically obtained 99·11, 99·41, 99·70, 95·88, 99·38 and 99·73%, respectively, whereas their prediction was 100%.

Table 2 also shows the results of the prediction of Markel *et al.* (1997) if the correct selection intensity (1·867) is used. Until generation N3, their predictions are quite similar to the simulation results, because the effect of a selection intensity which is too large is compensated by the downward-biased prediction of the standard deviation. However, after 3 generations the simulations suggest that it would take at least until generation N6 before a congenic strain can be created, whereas Markel *et al.* predict this could take place

https://doi.org/10.1017/S0016672399003857 Published online by Cambridge University Press

after generation N4 (the best male in N4 is predicted to have a proportion of 99·53% of the recipient genome).

Selection based on markers which are unique to the donor and recipient lines ignores double recombinants within marker brackets. The probability of double recombinants is low with markers spaced every 10 cM, but not insignificant. For example, at generation N6 the simulations resulted in all individuals being fixed for the recipient line markers (results not shown), whereas the underlying proportion of the genome which was from the recipient line was on average 99·65% (Table 2). Contamination from the donor genome after fixation of all recipient genome marker was pointed out previously (Wakeland *et al*., 1997). A denser marker map would explain more of the underlying variance in genomic proportion (Visscher, 1996).

(v) *Other species*

Hill (1993) pointed out the variation in genomic proportion was dependent more on the total genome length than on the distribution of chromosome lengths for a given total genome length; the longer the genome, the smaller the variance in genomic proportion. Therefore, simulations were performed with a very short genome length, using *Drosophila* parameters, and an intermediate genome length, using parameters from the *Arabidopsis* genome. For each set of parameters, 1000 replicate populations were simulated.

For the *Drosophila* genome, three chromosomes of length 66, 105 and 99 cM were used in the simulation study, with 11 fully informative markers per chromosome. A single male from the recurrent population was repeatedly backcrossed to a single female from the crossbred population. One hundred progeny were simulated, and the best female (out of 50 females) that carried one copy of the allele which was introgressed was selection based on the marker-based estimate of the proportion of her genome from the recipient population. On average only half the female candidates have inherited one copy of the allele to be introgressed, so that approximately 1 in 25 females were selected on the genomic proportion criterion. Ten generations of backcross populations were simulated. The allele which was introgressed was at location 30 cM on the first (X) chromosome. Results are in Table 3. The average genomic proportion reaches 99·5% faster than in the mouse genome, because of a larger selection intensity assumed (1 of 50 females for the flies versus 1 of 40 males in the mice) and because of a shorter genome length. After 5 generations of backcrossing the increase in mean proportion of the genome from the recipient population is small. This is due to linkage drag around the introgressed allele (Stam & Zeven, 1981), and to contamination from the donor genome after fixation of all recipient genome marker (Wakeland *et al*., 1997).

Table 3. *Mean and standard deviation* ($\times 100$) *of genomic proportion during backcrossing when a single female, selected from 50 candidates, is used each generation. Results from 1000 replicates of simulation, using the* Drosophila *and* Arabidopsis *genome*

	Proportion ($\times 100$) of genome from recipient population			
	Drosophila genome		*Arabidopsis* genome	
Generation	Mean	SD	Mean	SD
N2	74·98	11·19	75·04	8·57
N3	94·41	3·97	93·18	3·78
N4	98·25	1·51	98·13	1·47
N5	99·22	0·73	99·26	0·66
N6	99·47	0·51	99·56	0·41
N7	99·56	0·43	99·64	0·34
N8	99·61	0·38	99·67	0·32
N9	99·64	0·35	99·68	0·31
N10	99·66	0·33	99·69	0·30
N11	99·68	0·31	99·70	0·29

For the *Arabidopsis* genome (five chromosomes, of lengths 122, 77, 96, 76 and 98 cM), 100 progeny were simulated from a single plant, and a single plant was selected that was heterozygous for the allele which was introgressed and had the largest proportion of the genome from the recipient population. The allele which was introgressed was at a position 30 cM on chromosome I. Each chromosome contained 11 equidistant markers. Hence, apart from the total genome length and number of chromosomes, the same parameters were used as in the *Drosophila* simulation. Results are presented in Table 3, and are very similar to those for the *Drosophila* genome. The largest difference is at generation N3 (94·41% vs 93·18%), because of a larger variation in genomic proportion in the previous generation for the smaller (*Drosophila*) genome.

3. Conclusions

I have shown that the theoretical results of Hillel *et al*. (1990) and Markel *et al*. (1997) give a biased prediction of the introgression process in the chicken and mouse, and that better, more realistic predictions can be achieved using either theoretical results (Stam & Zeven, 1981; Hill, 1993) (for the case of random selection) or simulations (Hospital *et al*., 1992; Visscher *et al*., 1996; Wakeland *et al*., 1997). The biases in the prediction were based on an incorrect assumption regarding the distribution of genomic

https://doi.org/10.1017/S0016672399003857 Published online by Cambridge University Press

proportion and a failure to account for the reduction in variation in genomic proportion when only a few parents from the recurrent population are used each generation.

The only method to create speed congenics in this study was the use of genetic markers throughout the genome. There are other methods to create speed genetic programmes, for example in mammals by employing reproductive techniques such as super-ovulation of prepubertal females, *in vitro* maturation of oocytes and embryo transfer (Behringer, 1998), and the use of gamete harvesting and nuclear transfer in recurrent selection schemes (e.g. Haley & Visscher, 1998).

I thank Chris Haley and Bill Hill for encouragement and helpful comments.

References

Behringer, R. (1998). Supersonic congenics? *Nature Genetics* **18**, 108.

Falconer, D. S. & Mackay, T. F. C. (1996). *Introduction to Quantitative Genetics*, 4th edn. Harlow, UK: Longman.

Haley, C. S. & Visscher, P. M. (1998). Strategies to utilise marker-QTL information. *Journal of Dairy Science* **81**(Suppl. 2), 85–97.

Hill, W. G. (1993). Variation in genetic composition in backcrossing programs. *Journal of Heredity* **84**, 212–213.

Hillel, J., Schaap, T., Haberfield, A., Jeffreys, A. J., Plotzky, Y., *et al*. (1990). DNA fingerprints applied to gene introgression in breeding programs. *Genetics* **124**, 783–789.

Hospital, F., Chevalet, C. & Mulsant, P. (1992). Using markers in gene introgression breeding programs. *Genetics* **132**, 1199–1210.

Markel, P., Shu, P., Ebeling, C., Carlson, G. A., Nagle, D. L., *et al*. (1997). Theoretical and empirical issues for marker-assisted breeding of congenic mouse strains. *Nature Genetics* **17**, 280–284.

Stam, P. & Zeven, A. C. (1981). The theoretical proportion of the donor genome in near-isogenic lines of self-fertilizers bred by backcrossing. *Euphytica* **30**, 227–238.

Visscher, P. M. (1996). Proportion of the variation in genetic composition in backcrossing programs explained by genetic markers. *Journal of Heredity* **87**, 136–138.

Visscher, P. M., Haley, C. S. & Thompson, R. (1996). Marker assisted introgression in backcross breeding programs. *Genetics* **144**, 1923–1932.

Wakeland, E., Morel, L., Achey, K., Yui, M. & Longmate, J. (1997). Speed congenics: a classic technique in the fast lane (relatively speaking). *Immunology Today* **18**, 472–477.