Measuring the extent of linkage disequilibrium in commercial pig populations

N. Harmegnies*, F. Farnir*, F. Davin*, N. Buys†, M. Georges* and W. Coppiters*

*Unit of Animal Genomics, Faculty of Veterinary Medicine, University of Liège (B43), 20 Bd de Colonster, 4000-Liège, Belgium. †Gentec N.V., Brusselsesteenweg 168, 9280 Lebbeke, Belgium.

Summary
To evaluate the extent of linkage disequilibrium in domestic pigs, we genotyped 33 and 44 unrelated individuals from two commercial populations for 29 and five microsatellite markers located on chromosomes 15 and 2 respectively. A high proportion of marker pairs up to 40 cM apart exhibited significant linkage disequilibrium in both populations. Pairwise $r^2$ values averaged between 0.15 and 0.50 (depending on chromosome and population) for markers <1 cM apart and declined to values of 0.05 for more distant syntenic markers. Our results suggest that both populations underwent a bottleneck approximately 20 generations ago, which reduced the effective population size from thousands to <200 animals.

Keywords effective population size, linkage disequilibrium, pigs.

Introduction
Gametic association (GA) refers to the population-wide, non-random association of alleles at distinct loci, which accumulates as a result of random drift, migration, mutation and selection but erodes with time as a result of recombination. Steady-state levels of GA are, therefore, typically higher for tightly linked loci, justifying the common usage of the term linkage disequilibrium (LD).

There are multiple measurements of LD between two biallelic loci. Several of these are based on $|D|$, the difference between the observed and expected haplotype frequencies at linkage equilibrium. $D'$ is a rescaled version of $|D|$, ranging from 0 to 1 and accounting for differences in allelic frequencies that make comparisons of $|D|$ between distinct pairs of loci meaningless. A $D'$ value of 1 is referred to as a complete LD and indicates that one of the alleles at the first locus is associated with only one allele at the second locus, without having necessarily having identical frequencies. The $r^2$ is the squared correlation between the alleles at two biallelic loci, corresponding to $D^2$ divided by the product of all four allelic frequencies, and it ranges from 0 to 1. An $r^2$ of 1 is referred to as a perfect LD and corresponds to complete LD with identical allelic frequencies at both loci. One of the advantages of $r^2$ is its known distribution under the null hypothesis of linkage equilibrium (e.g. Pritchard & Przeworski 2001). Estimates of $r^2$ are also less sensitive to sample size than are estimates of $D'$ (e.g. Weiss & Clark 2002). Linkage disequilibrium between two multi-allelic loci can be quantified by averaging $D'$ or $r^2$ over all possible allelic pairs weighted by the frequency of the corresponding haplotypes.

In a finite, randomly mated population and in the absence of selection, the expectation of the sample estimate of $r^2$ is roughly $1/(1 + 4N_e\theta + 1/n)$ (e.g. Hudson 2003). In this formula, $N_e$ is the effective population size, $\theta$ is the recombination rate between the two loci, and $n$ is the sample size. If $\theta$ is known, $r^2$ thus provides an estimate of the effective population size. Estimates of $r^2$ for markers at decreasing distance provide estimates of $N_e$ in the more distant past (e.g. Hayes et al. 2003).

Although highly variable between chromosome regions, LD only extends over sub-centiMorgan levels in most natural populations including humans. In Caucasian populations, for example, useful levels of LD ($r^2 \geq 0.3$) typically do not extend for more than 3–80 kb (e.g. Kruglyak 1999, Reich et al. 2001, Altschuler et al. 2005). Unexpectedly, it was found that significant LD extends...
over much longer distances in domestic animal species including the cow, sheep, pig and dog (e.g. Farnir et al. 2000, McRae et al. 2002, Nsengimana et al. 2004, Sutter et al. 2004, Jungerius et al. 2005). This is most likely due to random drift because of the unusual population structure imposed by breeding designs causing dramatic reductions in \(N_e\).

The extent of LD has some important practical implications. It determines the marker density that is required to locate genes underlying phenotypes of interest by means of whole-genome association studies. The extensive genome-wide LD observed in cattle has already allowed mapping of several disease genes by association using the available medium-density microsatellite maps (e.g. syndactyly in Holstein-Friesian, Charlier et al. 1996). It determines the marker density needed for and limits of LD-based fine-mapping strategies. Linkage disequilibrium has been used to refine the map positions of quantitative trait loci (QTL) to centiMorgan levels (e.g. Grisart et al. 2002, 2004, Meuwissen et al. 2002, Blott et al. 2003, Nezer et al. 2003, Olsen et al. 2005). It is also of importance when computing matching probabilities and exclusion power when using multiple markers for identification or pedigree verification purposes.

In this study, we measured the level of LD in two commercial pig populations to evaluate the feasibility of mapping genes underlying inherited defects by means of whole-genome association analysis.

## Materials and methods

### Population sample

We sampled 33 and 40 individuals from two commercial pig populations referred to as A and B respectively. Line A is a synthetic population derived from the intercrossing of Large White and Landrace animals approximately 20 generations ago, with the occasional introgression of new blood. At any given point in time, the population consists of approximately 10 boars and 250 sows. Line B is a purebred Large White line founded approximately 30 generations ago, and it contains approximately 10 boars and 200 sows at any given point in time. Line B has remained essentially closed since its foundation. In both lines, individuals without common parents were sampled.

### Marker selection and genotyping

To test for LD between markers covering a broad range of genetic distances, we selected 29 microsatellite markers covering 62 cM (Kosambi) on chromosome SSC15: SW1989, SW1401, SW1673, S0284, SW423, KS108, SW2129, SW1263, KS161, SWR1533, SW1316, SWR1002, SW120, SW1683, SW2053, S1000, S1002, S1001, S1003, S1004, S1005, S1006, S1008, SW936, SW906, SW2608, KS135, SW1262 and SW1119. The map positions of these markers are shown in Table 1, with the distance between adjacent markers ranging from <0.1 cM (<100 kb) to 10.5 cM (Milan et al. 2000, Rohrer et al. 1996). Note that the RN* mutation, causing high glycogen content in skeletal muscle and mapping to SSC15, is not present in the A or B lines.

To test GA between non-syntenic loci, we genotyped the same individuals for five microsatellite markers located near IGF2 on distal SSC2p: KVLQT1, PULGE3, PULGE1, SWC9 and SWR2516 (Table 1). These markers cover an estimated 1 cM, with adjacent markers from <0.1 cM (<63 kb) to 0.6 cM (Amarger et al. 2002, Nezer et al. 2003). This SSC2p region is likely to have undergone selection in lines A and B, which are both segregating for the IGF2 intron3 3072G>A quantitative trait nucleotide (Van Laere et al. 2003).

Multiplex microsatellite genotyping was performed using standard procedures and automatic ABI3700 capillary sequencers. Electropherograms were read independently by two experienced persons.

### Hardy–Weinberg (HW) equilibrium

Departure from HW proportions was tested using a Monte-Carlo approximation of Fisher’s exact test. Following Weir (1996), we computed the probability, Pr, of the observed genotype numbers \((n_{uv})\), conditional on observed allele numbers \((n_u)\) as:

\[
Pr(n_{uv}||n_u) = \frac{n!}{(2n)!} \prod_u (n_u)! \prod_v (2n_{uv})! 
\]

We then randomly simulated 16 000 genotype numbers from the observed allele numbers, computed Pr for the corresponding tables using equation (1) and sorted the tables with Pr-value smaller or equal to the Pr-value of the observed genotype numbers. To test the actual significance, \(P_c\), of the departure from HW for individual markers, the nominal significance, \(P_n\), was corrected for the testing of multiple markers by applying the Bonferroni correction \(P_c = 1-(1-P_n)^{14}\) because 34 microsatellites were examined.

Given the limited power of the previous test, we compared the distribution of \(P_n\) values across markers with that expected assuming that all markers were in HW equilibrium. The latter was obtained from the distribution of \(P\)-values computed from the Pr-values of the 34 \(\times\) 16 000 simulations in a manner analogous to the one utilized for the observed tables. Note that this distribution was not uniform (i.e. 10% of the \(P\)-values between 0 and 0.1, 10% between 0.1 and 0.2, etc.), as intuitively expected. Comparisons between observed and simulated distributions...
were done by deciles using a chi-squared goodness-of-fit test.

Linkage disequilibrium

Pair-wise LD was measured for all possible marker pairs. We tested LD using the likelihood ratio test \( A = \ln(L_{LE}/L_{LD}) \), where \( L_{LE} \) corresponds to the likelihood of the data assuming linkage equilibrium and \( L_{LD} \) corresponds to the likelihood of the data allowing for departure from linkage equilibrium. \( L_{LE} \) was computed as:

\[
L_{LE} = \prod_{i=1}^{N} p_i^2 \prod_{j=1}^{M} 2p_i q_j \prod_{k=1}^{O} q_k^2 \prod_{l=1}^{P} 2q_k q_l
\]

where \( N \) and \( M \) are the number of individuals that are homozygous and heterozygous, respectively, for the first marker, \( O \) and \( P \) are the equivalent numbers for the second
marker, $p_i$ (and $q_i$) are the frequencies of the alleles for which individuals $i$ ($k$) are homozygous, $p_{i1}$ and $p_{i2}$ ($q_{i1}$ and $q_{i2}$) are the frequencies of the two alleles of individual $j$ ($l$). Allele frequencies were determined simply from the data by counting.

$L_{LD}$ was computed as:

$$L_{LD} = \frac{N^2 + M - \sum_{i=1}^{N} h_i}{2T}$$

where $N$ is the number of double homozygous individuals, $M$ is the number of individuals that are homozygous for one of the two markers, $O$ is the number of double heterozygous individuals, $h_i$ is the frequency of the haplotype for which individual $i$ is homozygous, $h_{i1}$ and $h_{i2}$ are the frequencies of the two haplotypes of individual $j$, $h_{k1A}$ and $h_{k2A}$ are the frequencies of the two haplotypes of individual $k$ assuming phase $A$, and $h_{k1B}$ and $h_{k2B}$ are the frequencies of the two haplotypes of individual $k$ assuming phase $B$. Values of $h_i$ that maximized $L_{LD}$ were determined using an expectation–maximization (EM) approach (e.g. Weir 1996). In this, the values of $h_{x,t+1}$ at EM-cycle ($t + 1$) are estimated as:

$$h_{x,t+1} = \frac{2N + M - \sum_{i=1}^{N} p_i}{2T}$$

where $N$ is the number of individuals that are homozygous for haplotype $x$, $M$ is the number of individuals that are unambiguously heterozygous for haplotype $x$ as they are homozygous at one of the two markers, $O$ is the number of double heterozygous individuals that have a probability $p_i \neq 0$ to have haplotype $x$, and $T$ is the total number of individuals in the sample. $p_i$ is computed as:

$$p_i = \frac{2h_{x,t}h_{k2A}}{2h_{x,t}h_{k2A} + 2h_{x,t}h_{k2B}}$$

where $h_{x,t}$ is the frequency of haplotype $x$ at EM-cycle $t$, $h_{k2A,t}$ is the $t$th estimate of the frequency of the second haplotype (the first one being $x$) of individual $l$ assuming phase $A$, $h_{k1B,t}$ and $h_{k2B,t}$ are the $t$th estimates of the frequencies of the two haplotypes of individual $l$ assuming phase $B$. All values of $h_x$ are updated at each cycle until convergence of all $h_{x,t+1}$ and $h_{x,t}$.

The nominal statistical significance ($P_e$-values) of observed $A$ were estimated from the predicted distribution of $A$ under the null hypothesis of LE. The latter was obtained separately for each marker pair from the $A$ values obtained from 1000 population samples simulated under the assumption of LE. The allele numbers of the simulated samples were kept identical to the real data. The $P_e$-value of the observed $A$ was computed as the proportion of simulations with $A$ larger or equal to the observed $A$.

As a result of this multiple testing, the power to detect significant deviation from LE for individual marker pairs was extremely low. As for HW equilibrium (see above), we compared the distribution of observed $P_e$-values with that expected assuming that all markers are in linkage equilibrium. This was done after grouping the marker pairs by distance: 0–5, 5–10, 10–20, 20–40, 40–60 cM, and non-syntenic. The distribution of $P_e$-values assuming linkage equilibrium was obtained by computing the $P_e$-value for all simulated data sets as was done for the real data and pooling over all markers. The resulting distribution was shown to be uniform (i.e. 10% of the simulations being significant at the 10% level, 20% at the 20% level, etc.). As a result, the observed distribution of $P_e$-values was compared by deciles with the expected uniform distribution using a chi-squared goodness-of-fit test.

We then computed $r^2$ for all marker pairs as:

$$r^2 = \sum_{ij} h_{ij} \frac{(h_{ij} - p_{ij})^2}{p_i(1-p_i)q_j(1-q_j)}$$

where $p_i$ is the frequency of allele $i$ at the first marker, $q_j$ the frequency of allele $j$ at the second marker, and $h_{ij}$ is the frequency of haplotype $ij$ estimated using the previously described EM algorithm.

Effective population size

We estimated the effective population size, $N_e$, from pairwise $r^2$ assuming that:

$$E(r^2) = \frac{1}{1 + 4N_e \theta} + \frac{1}{n}$$

in which $\theta$ is the recombination rate between the markers and $n$ is the diploid sample size. This equation shows that $E(r^2)$ comprises two parts. The first component $1/(1 + 4N_e \theta)$ reflects the past finite population history. Its adequacy in the case of multi-allelic markers was demonstrated by Hayes et al. (2003). The second component is due to the limited sample size (e.g. Hill 1981).

From the point-wise estimates of $N_e$ sorted by marker distance we computed the average $N_e$ for ‘moving’ windows, each spanning 5 cM. For marker pairs ($d$ recombination units apart) yielding negative estimates of $N_e$ (whenever $r^2 < 1/n$), the value of $N_e$ was replaced by the average $N_e$ estimates over all marker pairs separated by a distance of $d \pm 2.5\%$.

Results

Hardy–Weinberg equilibrium

Table 1 reports the number of alleles, the heterozygosity and the nominal and Bonferroni-corrected $P$-values of the departure from HW for the 34 microsatellites in lines A and B. The average number of alleles per microsatellite was 4.48 and 4.03 in lines A and B while the heterozygosity averaged 0.61 and 0.59 respectively. The slightly higher genetic variation observed for line A when compared with line B might be due to its synthetic origin. Two markers had
nominal $P_n$-values $<0.05$ in line A (S1000 and SW1119, both on SSC15), and one in line B (KS161 on SSC15), suggesting possible departure from HW. However, after correcting for multiple testing, none of the $P_c$-values remained significant. The distribution of $P_n$-values did not differ significantly from that expected assuming HW equilibrium ($P = 0.72$ in line A, $P = 0.82$ in line B; data not shown).

**Linkage disequilibrium**

We measured LD for all marker pairs, first by estimating the probability of the marker genotypes under the null hypothesis of linkage equilibrium. The cumulative frequency distributions of obtained $P_n$-values for all marker pairs sorted by distance are shown in Fig. 1a and b. Significant LD was detected in both lines for markers up to 40 cM, and even up to 60 cM in line B. No LD was found between non-syntenic markers.

We then used the haplotype frequencies corresponding to the maximum likelihood solutions to measure LD using $r^2$ (Fig. 2a and b). The $r^2$ profiles obtained between SSC15 markers were relatively similar for lines A and B, averaging 0.15 and 0.19 for markers $<1$ cM apart, 0.10 and 0.12 for markers between 1 and 5 cM apart, and 0.07 and 0.10 for markers between 5 and 10 cM apart respectively. For larger distances, the average $r^2$ values were around 0.05 in both lines. For the SSC2 markers, which were all within sub-centiMorgan distance, $r^2$ values were slightly higher, averaging 0.48 in line A and 0.35 in line B.

**Estimating effective population size**

The $r^2$ values were used to estimate $N_e$ according to Hayes et al. (2003). We only utilized the SSC15 values, as the SSC2 region that was explored underwent selection (see above), thus violating the theoretical assumptions underlying equation (7). The moving averages obtained using a sliding 5-cM window are shown in Fig. 2a and b. The profiles were again similar in both populations, showing a sharp decline from values in the order of 1500 with markers $<6–7$ cM apart to values averaging 100 for more distant markers ($>20$ cM).

![Figure 1](image1.png)  
**Figure 1** Cumulative frequency distribution of $P_n$-values of the likelihood ratio test statistic $\lambda$ for linkage equilibrium for marker pairs sorted by inter-marker distance for lines A (graph a) and B (graph b). Statistical significance ($P$-values) of the deviation from the expected uniform distribution assuming that all markers are in linkage equilibrium (black diagonal line) estimated by means of a chi-squared goodness-of-fit test is given under the corresponding legend.

![Figure 2](image2.png)  
**Figure 2** $r^2$ (left Y-axis) values computed for SSC15 (○) and SSC2 (▲) marker pairs sorted by distance (expressed in cM) for lines A (graph a) and B (graph b). Average $r^2$ values were computed for markers between 0 and 1, 1 and 5, 5 and 10 and 15, 10 and 20, 20 and 30, 30 and 40, 40 and 50 and 50 and 60 cM apart (SSC15, □; SSC2, ▪). Average estimates of the effective population size, $N_e$ (right Y-axis), for 5-cM moving windows (●), which were computed from the SSC15 data points as a function of $r^2$, $\theta$ and the sample size, $n$, according to Hill (1981).
**Discussion**

Herein, we demonstrate the occurrence of statistically significant, long-range (>10 cM) LD between syntenic markers in two commercial pig populations. This situation is to some extent similar to what has been previously reported in livestock (Farnir et al. 2000, McRae et al. 2002, Nsengimana et al. 2004). Contrary to what was found previously in cattle, however, we found no evidence in this work for GA between non-syntenic markers.

We assume that, as demonstrated for cattle, this long-range LD is for a large degree because of random drift caused by reduced \( N_e \) resulting from the population structure imposed by modern breeding designs. This hypothesis is supported by estimates of \( N_e \) derived from \( r^2 \), which point to a drastic reduction in \( N_e \) from thousands of individuals to hundreds or less (Fig. 2a and b). Assuming that \( r^2 \) estimates \( N_e \) \( 1/(20) \) generations ago (Hayes et al. 2003), this event occurred an estimated ~20 generations ago, which corresponds well with the known history of these populations. The \( r^2 \)-based estimates of present \( N_e \) also agree reasonably well with estimates based on the known numbers of males and females in these populations using \( N_e = 4N_mN_h/N \) \((N_m + N_h)\), corresponding approximately to 40 in both lines (e.g. Crow & Kimura 1970).

The fact that \( r^2 \) was higher for markers at sub-centi-Morgan distances from the IGF2 locus when compared with distant markers at the RN locus suggests that selection might contribute to LD as well. Because of its effect on muscle mass (Van Laere et al. 2003), the IGF2 locus has been under selective pressure in both the studied populations. As none of the lines are related to the Hampshire breed, they are devoid of the RN- allele causing an increase in muscle glycogen content (Milan et al. 2000), which might otherwise have been under negative selective pressure.

Although statistically significant, the \( r^2 \) values obtained even for closely linked markers are low: in the order of \( 10^{-4} \), does not affect the level of ‘long-range’ LD observed in livestock populations because the average number of generations to coalescence is expected to be much lower (in the order of tens to hundreds) than, for instance, in humans.

These results cast some doubt on the relative power of whole-genome association studies to map genes underlying phenotypes of interest with the available medium-density marker maps. Because of the direct relationship between \( r^2 \) and the required sample sizes (e.g. Pritchard & Przeworski 2001), values of 0.10 indicate the need for a sample size 10 times as large for association studies as for experimental designs, which would allow direct tracing of the causal mutation. Note that in designs based on conventional linkage, the causal mutation is in essence traced directly in large, informative (‘Qq’ heterozygous sire) paternal half-sib pedigrees, a pedigree structure commonly available in livestock. However, information is only extracted from the paternal chromosome, contrary to association studies that extract information from both chromosomes. Also, when performed in outbred populations, only a fraction of the studied half-sib pedigrees will segregate for the QTTL and, therefore, provide mapping information. The relative merit of both approaches needs to be evaluated on a case-by-case basis. It is likely that so as to be effective, association studies will impose much larger genotyping requirements than linkage-based designs. A considerable reduction in genotyping costs is, therefore, essential to make association studies a viable approach in livestock. The development of high-density SNP maps and progress in SNP genotyping technology may render this accessible in the not-too-distant future. In the meantime, hybrid strategies using linkage-based approaches first, followed by the exploitation of LD for fine-mapping purposes, seems to be a more realistic and effective modus vivendi.

**References**


Charlier C., Farnir F., Bezzz P., Vannmanshoven P., Brouwers B. & Georges M. (1996) BBD mapping of recessive traits in livestock,
application to map the bovine syndactyly locus to chromosome 15. Genome Research 6, 580–9.


