

ARTICLE

# The effective size of the Icelandic population and the prospects for LD mapping: inference from unphased microsatellite markers

Thomas Bataillon<sup>\*1,2</sup>, Thomas Mailund<sup>1</sup>, Steinunn Thorlacius<sup>3</sup>, Eirikur Steingrímsson<sup>3,4</sup>, Thorunn Rafnar<sup>3</sup>, Magnus M Halldórsson<sup>3,5</sup>, Violeta Calian<sup>3,6</sup> and Mikkel H Schierup<sup>1</sup>

<sup>1</sup>Bioinformatics Research Center, University of Aarhus, Høegh-Guldbergs Gade 10, DK-8000 Aarhus C, Denmark; <sup>2</sup>INRA – UMR 1097 ‘Diversité & Génomes des Plantes Cultivées’, Domaine de Melgueil, Mauguio, France; <sup>3</sup>Iceland Genomics Corporation, Snorrabraut, Reykjavik, Iceland; <sup>4</sup>Faculty of Medicine, University of Iceland, Reykjavik, Iceland; <sup>5</sup>Department of Computer Science, Faculty of Engineering, University of Iceland, Reykjavik, Iceland; <sup>6</sup>The Science Institute, University of Iceland, Reykjavik, Iceland

Characterizing the extent of linkage disequilibrium (LD) in the genome is a pre-requisite for association mapping studies. Patterns of LD also contain information about the past demography of populations. In this study, we focus on the Icelandic population where LD was investigated in 12 regions of ~15 cM using regularly spaced microsatellite loci displaying high heterozygosity. A total of 1753 individuals were genotyped for 179 markers. LD was estimated using a composite disequilibrium measure based on unphased data. LD decreases with distance in all 12 regions and more LD than expected by chance can be detected over approximately 4 cM in our sample. Differences in the patterns of decrease of LD with distance among genomic regions were mostly due to two regions exhibiting, respectively, higher and lower proportions of pairs in LD than average within the first 4 cM. We pooled data from all regions, except these two and summarized patterns of LD by computing the proportion of pairs of loci exhibiting significant LD (at the 5% level) as a function of distance. We compared observed patterns of LD with simulated data sets obtained under scenarios with varying demography and intensity of recombination. We show that unphased data allow to make inferences on scaled recombination rates from patterns of LD. Patterns of LD in Iceland suggest a genome-wide scaled recombination rate of  $\rho^* = 200$  (130–330) per cM (or an effective size of roughly 5000), in the low range of estimates recently reported in three populations from the HapMap project.

*European Journal of Human Genetics* advance online publication, 31 May 2006; doi:10.1038/sj.ejhg.5201669

**Keywords:** linkage disequilibrium mapping; coalescence; recombination rate; effective population size; Iceland; microsatellites

\*Correspondence: Dr T Bataillon, Department of Genetics and Ecology, Bioinformatics Research Center, University of Aarhus, Høegh-Guldbergs Gade 10, Building 1090, Aarhus 8000 C, Denmark.  
Tel: +45 8942 3359; Fax: +45 8942 3077;  
E-mail: tbata@daimi.au.dk  
Received 9 September 2005; revised 27 April 2006; accepted 28 April 2006

## Introduction

The isolated population of Iceland has a well-characterized history starting 1100 years ago when the Icelandic population was founded by emigrants from Scandinavia, probably mostly from Norway but with some significant contributions from the British Isles.<sup>1</sup> Its modest population size of 300 000 individuals, combined with very good historical records, makes Iceland a population of choice for the location of disease-causing genes using either linkage

mapping or association mapping, the latter also termed linkage disequilibrium (LD) mapping.<sup>2</sup> LD is defined as the statistical association between alleles at different loci in the genome. Such associations can be generated by some forms of epistatic selection or chance alone through the interplay of mutation and drift.<sup>3</sup> Amounts of LD within an isolated population are modulated mainly by the intensity of genetic drift, through the effective size of the population, and the per generation recombination rate between loci.<sup>4,5</sup> Patterns of LD in a sample of linked loci can thus be used to estimate the absolute effective size of a population provided independent information on the recombination rates between loci is available.<sup>6–8</sup> Effective population size, hereafter  $N_e$ , is a central parameter in population genetics as it governs the magnitude of drift and thus, among other things, LD patterns. In the context of association mapping, knowledge of  $N_e$  is needed to predict genome-wide patterns of LD, the density of single-nucleotide polymorphisms (SNPs) needed to cover a genomic region, and to model the power of LD-based association mapping methods to detect polymorphism-causing phenotypic variation for traits/disease of interest.<sup>5</sup>

Here we survey LD patterns in 12 genomic regions, spanning ~15 cM each, using sets of linked microsatellites genotyped in a random sample of the contemporary Icelandic population. We characterize LD within each region and find that the loci exhibiting significant LD can be found as distant as 4 cM suggesting a reduced effective population size. We then use extensive simulations to compare observed LD patterns with LD patterns simulated under various demographic scenarios with varying population size and recombination rates. This allows us to obtain point estimates of (1) the scaled recombination rate and (2) the long-term  $N_e$  in the Icelandic population.

## Materials and methods

### Sampling of individuals and microsatellite genotyping

The study population consisted of 1753 individuals (962 males and 791 females) who had participated in the Icelandic Cancer Project as controls.<sup>2</sup> Twelve genomic regions were selected based on the criteria that they were gene-rich, approximately 15 cM in size and of high sequence quality, that is, >90% of the sequence was in the finished state and contained no gaps. For convenience, the regions were named B, C, E, F, G, H, I, K, L, N, P and Q. Regions C, E and H were genotyped in a total of 1753 individuals, whereas the remaining regions were genotyped in a subset of 1400 individuals. For each of the genomic regions, 14–16 markers were selected, spaced at an interval as close to 1 cM as possible. The genetic markers were selected for high heterozygosity of dinucleotide repeats. When possible, available dinucleotide Marshfield

markers of high heterozygosity were used. The heterozygosity detected in our samples ranged from 0.44 to 0.96, and was on average 0.77. A complete list of genetic regions, marker locations and marker heterozygosity is presented in Supplementary Table 1.

Multiplex PCR was performed on the samples using unlabelled, marker-specific primers with a common oligonucleotide tail sequence, followed by a second PCR with fluorescent-labelled primers that anneal to the tails. Each multiplex PCR reaction contained 2–4 markers. Three such reactions were pooled and run on a MegaBace 1000 automated DNA sequencer (Amersham Biosciences). MegaBace Genetic Profiler software was used for fragment sizing and genotype calling. Two researchers checked all genotyping scores individually and a third person scored discrepancies between the two. Median success per marker was 98% (84.7–99.6); 2.8% of genotypes were unreadable.

### LD analysis

LD was analyzed using the composite measure of genotypic disequilibria devised by Weir<sup>9</sup> (see Schaid<sup>10</sup> for a recent account). This measure, hereafter denoted by  $\Delta$ , is defined for a pair of loci with arbitrary number of alleles, is insensitive to deviations from Hardy–Weinberg equilibrium (hereafter HWE) and can be used with unphased data.  $\Delta$  was calculated after discarding, at each locus, alleles with frequency lower than 5%. For each pair of loci, A and B, each with, respectively,  $n_A$  and  $n_B$  alleles with frequency greater than 5%,  $\Delta$  was calculated using the counting scheme suggested by Schaid.<sup>10</sup> Significance of the composite LD was calculated for each pair by computing the  $\chi^2$  statistic from the counts of the  $n_A \times n_B$  contingency table and assuming that, under the null hypothesis of absence of LD,  $\chi^2$  follows a  $\chi^2$  distribution with  $(n_A - 1)(n_B - 1)$  degrees of freedom. This approximation is very accurate provided that all cells in the contingency table have five observations or more, an assumption that was met in virtually all cases because of the large sample size (1753 individuals) and the 5% threshold imposed on allele frequencies. The counting scheme and significance testing were first implemented in *Mathematica* and then as a C routine (see below for the analysis of simulated data). The significance testing procedure was checked by analyzing the contingency tables obtained for ~20 pairs of loci from all regions both in *Mathematica* and using the statistical software *R*. For these loci, we checked (1) that the *P*-values obtained in *R* under the  $\chi^2$  approximation, using the *chisq.test* function, were identical to the ones computed in *Mathematica* and (2) that these *P*-values were accurate by comparing them with exact *P*-values computed under the null hypothesis of no-linkage disequilibrium using the function *fisher.test* implemented in *R*.

The individual *P*-values obtained for all 12 regions were pooled ( $n = 1235$  pairs) and the distribution of *P*-values was

**Table 1** Number of pairs of loci exhibiting significant LD within each region

Reg	Loc	$n_{min}$	Pairs	1 cM	2 cM	3 cM	4 cM	5 cM	> 5cM	> 10cM	Total
B	13q	1347	S	6	6	3	2	1	3	0	21
			NS	4	9	12	13	9	32	5	84
C	17q21	1534	S	6	5	3	3	1	1	0	19
			NS	0	5	7	8	8	35	23	86
E	16q21	1632	S	2	8	4	3	0	0	0	17
			NS	0	2	4	4	9	31	24	74
F	20q11	1329	S	4	7	5	2	0	0	0	18
			NS	2	6	6	10	10	35	18	87
G	22q13	1289	S	2	4	1	0	0	0	0	7
			NS	3	9	10	9	11	39	17	98
H	6p21	1634	S	13	9	4	0	0	0	0	26
			NS	2	8	12	11	12	49	16	110
I	10q24	1340	S	3	7	0	1	1	0	0	12
			NS	1	6	9	12	13	35	17	93
K	15q	1292	S	2	4	0	2	0	0	0	8
			NS	2	9	9	9	8	37	23	97
L	9q34	1330	S	4	3	1	0	0	0	0	8
			NS	0	3	10	8	7	30	25	83
N	11q13	1285	S	2	6	2	1	0	0	0	11
			NS	2	6	8	9	8	28	19	80
P	7q22	1331	S	3	3	3	1	0	0	0	10
			NS	5	8	9	9	12	37	15	95
Q	5q31	1343	S	4	6	4	0	0	0	0	14
			NS	2	5	6	8	11	30	15	77

Notes: Reg: region name; Loc: chromosomal location of the region;  $n_{min}$ : minimum number of individuals effectively used to assess LD, calculated as the total number of individual genotyped times the maximum proportion of missing data per locus in the region (rounded to the nearest integer); S: number of pairs of loci in significant LD (at the 5% level) within a region (binned by class of genetic distance) and NS: number of pairs of loci in nonsignificant LD (at the 5% level) within a region (binned by class of genetic distance).

used to estimate the so-called  $q$ -value, the frequency of false positives among the pairs declared as 'significant' at a given individual threshold, using the method of Storey and Tibshirani,<sup>11</sup> as implemented in the *R* routine ' $q$ -value'. Examination of the graph depicting the relation between individual  $P$ -values and corresponding  $q$ -values suggests that retaining pairs of loci with associated  $P$ -values of 1% will yields less than 10% of false positives over all pairs (data not shown).

A logistic regression was used to model the decrease in the proportion of pairs exhibiting significant LD with genetic distance within each region. We used the genetic distance (in cM) between microsatellites from the recombination map reported by Kong *et al.*<sup>12</sup> We binned our data in seven classes of distance between loci: from 0 to 1, 1–2, 2–3, 3–4, 4–5 cM, from 5 to 10 cM and distances larger than 10 cM. For convenience, these seven categories are hereafter denoted in tables and figures as, respectively,

'1 cM', '2 cM', '3 cM', '4 cM', '5 cM', '> 5 cM' and '> 10cM'. The proportion of significant pairs (at the 5% level),  $p_{ij}$ , found within region  $i$  at distance  $D_j$  was modeled as  $Z_{ij} = \mu + R_i + D_j$ , where  $Z_{ij} = \log(p_{ij}/(1-p_{ij}))$ ,  $\mu$  is the intercept of the regression,  $R_i$  is the effect associated with region  $i$  and  $D_j$  the effect of distance between pairs of loci (in cM). The analysis was performed in *R* using the binomial family and a logit link function.

### Simulated data sets

Data were simulated under scenarios with variable amounts of recombination and exponential population growth in the population history using a coalescence approach. Samples drawn from populations were simulated using the ancestral recombination graph approach implemented in the software *CoaSim* version 4.0,<sup>13</sup> an extensible coalescence simulation tool (<http://www.birc.dk/Software/CoaSim/>).

The following parameters were varied in the coalescent simulations:

$\rho = 4N_e rL$ , the scaled recombination rate for a chromosomal region of length  $L$  nucleotides recombining at a rate  $r$  per nucleotide and per generation, varied in {500, 1000, 2000, 3000, 4000, 4500, 5000, 10000}.

$\theta = 4N_e \mu$ , the scaled mutation rate of the microsatellite markers simulated on the chromosome region, was set to either 1.5 or 3. All microsatellites were assigned the same scaled mutation rate within a given simulation. To check the influence of the model of mutation on LD patterns, simulations were run under two alternative mutation models commonly used for modeling microsatellite: the  $K$ -alleles model and the stepwise mutation model. A  $K$ -alleles mutation model with  $K = 50$  different alleles was assumed throughout the study. The  $K$ -allele mutation model is directly implemented in *CoaSim*, but the stepwise mutation model was implemented using the scheme-based extension language of *CoaSim*.

Population growth was simulated through the parameter  $\beta = 2N_e b$ , where  $b$  is the (exponential) growth rate per generation and  $N_e$  is the effective population size.  $\beta$  was set to 0 (stable population) 10, and 100. For  $\beta > 0$ , the value of the parameter  $\theta$  was adjusted so that the expected number of alleles per locus would be the same as under  $\beta = 0$ . This was carried out by first obtaining, through repeated simulation ( $10^4$  independent simulations), the mean length of the whole genealogy of a sample drawn from a population with growth rate  $\beta$  relative to that of a population with growth rate  $\beta = 0$  (the latter being known analytically<sup>14</sup>).

For each combination of parameters ( $\rho$ ,  $\theta$ ,  $\beta$ ) and each mutation model, 100 independent data sets consisting of 10 regions of length  $\rho$  with 15 microsatellite loci per region were simulated. For all simulations, a sample size of 2600 haploid chromosomes (pooled at random to generate 1300 diploid individuals) was used. This sample size corresponds to the minimum of genotypic data available across all regions (Table 1). The rationale for choosing this sample size was to run stochastic simulations with the biggest sample size still compatible with the data observed in all regions genotyped.

### Inference of demography and evolutionary parameters from LD patterns

LD patterns in simulated data were summarized by computing the proportion of pairs of loci exhibiting significant LD at seven different classes of genetic distance (expressed as fractions of  $\rho$ ) using the same binning scheme as the one adopted for the observed data. The distance between a single simulated data set and the observed data was calculated as

$$d = \sum_{i=1}^7 2 \frac{(p_{\text{obs}}^i - p_{\text{sim}}^i)^2}{(p_{\text{obs}}^i + p_{\text{sim}}^i)} \quad (1)$$

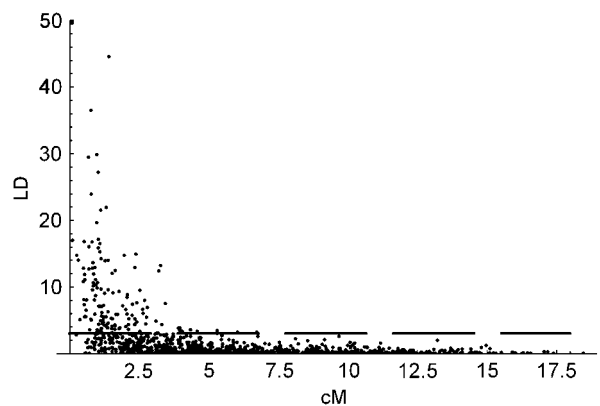
where  $p_{\text{obs}}^i$  (and  $p_{\text{sim}}^i$ ) is the observed (and simulated) proportion of pairs of loci in significant LD within the class

of genetic distance  $i$ . For each combination of parameters ( $\rho$ ,  $\theta$ ,  $\beta$ ), a mean distance was calculated by averaging over a hundred simulated data sets. The combination ( $\rho^*$ ,  $\theta^*$ ,  $\beta^*$ ) yielding the lowest mean distance was retained as fitting the data best.

## Results

### Overall patterns of LD in the data

A total of 1753 individuals were genotyped at 179 loci. The mean amount of missing data per locus ranged from 1.7 (Region B) to 3.6% (Region C), yielding a minimum of 1285 individuals effectively genotyped for assessing LD within each region (Table 1). LD patterns between pairs of loci in each region are summarized in Table 1. Hereafter, we consider a pair of loci exhibits significant LD when the  $\chi^2$  test statistic has an associated  $P$ -value that is lower than 5%. The proportion of pairs of loci exhibiting significant LD decreased with genetic distance, although there was considerable variation within a given distance (Figure 1). A logistic regression analysis confirmed that the genetic distance between loci (in cM) explained a significant amount of the variation in the proportion of pairs exhibiting significant LD at the 5% level ( $P < 10^{-16}$ ). Regions E and G displayed significant heterogeneity in their patterns of LD relative to the remaining regions (Supplementary Figure 1). Region E (respectively G) had a higher (respectively lower) average proportion of pairs of loci exhibiting significant LD within the first 4 cM. When these two regions were discarded from the logistic regression, the decrease in the proportion of pairs exhibiting LD with genetic distance did not exhibit further heterogeneity between regions ( $P > 0.05$ ), whereas the effect of genetic distance was still highly significant ( $P < 10^{-15}$ ).



**Figure 1** Decrease of LD with genetic distance ( $n = 1235$  pairs, five pairs with  $P$ -values lower than  $10^{-20}$  are not shown in the graph but were used in the fitting procedure). X-axis: genetic distance between loci of a pair in centiMorgans (cM). Y-axis: LD measured for each pair as minus the natural logarithm of  $P$ -values. The horizontal dotted line denotes the 5% threshold for individual  $P$ -values ( $-\text{Log}(0.05) = 3$ ).

**Table 2** Properties of the studied regions

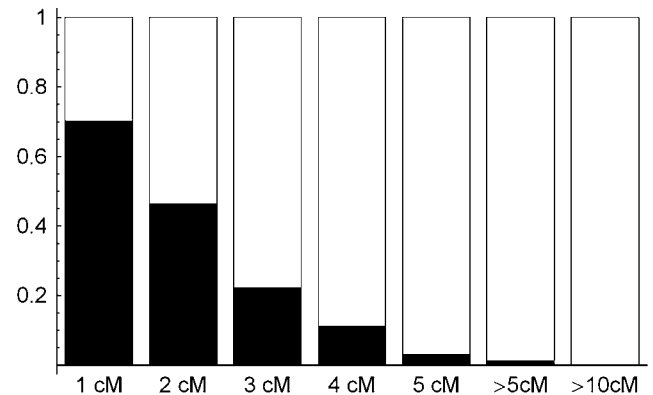
Region	Location	Mb	$cM_{\text{Decode}}$	$cM_{\text{Marshfield}}$	$cM_{\text{Decode per Mb}}$	$cM_{\text{HapMap per Mb}}$
B	13q	9	13	13	1.44	1.38
C	17q21	17	17	15	1.00	1.30
E	16q21	19	19	17	1.00	1.13
F	20q11	14	15	16	1.07	1.17
G	22q13	12	16	20	1.33	1.65
H	6p21	18	15	20	0.83	0.96
I	10q24	19	16	16	0.84	0.96
K	15q	20	18	13	0.90	1.01
L	9q34	13	20	17	1.54	2.13
N	11q13	14	15	15	1.07	1.06
P	7q22	17	15	17	0.88	0.99
Q	5q31	17	15	14	0.88	0.96

Notes: Mb: estimated physical length of the region in Megabases;  $cM_{\text{Decode}}$ : estimated length of the region (in cM) based on the deCODE genetic map;  $cM_{\text{Marshfield}}$ : estimated length of the region (in cM) based on the Marshfield genetic map and  $cM_{\text{HapMap per Mb}}$ : average local recombination rates estimated using SNPs data from the HapMap project (release 16a). Data and details on the estimation method used to infer these scaled recombination rates are available at <http://www.hapmap.org/downloads/recombination/latest/>.

Region G exhibited the second highest ratio of genetic (cM) to physical (Mb) among the 12 regions studied (Table 2). However, the differences among regions revealed by the logistic regression cannot be explained by difference in sample size between regions E and G or by differences in the average cM per Mb among regions (not shown). Note that regions genotyped were precisely chosen to be roughly 1 cM per Mb (0.83–1.54 cM per Mb). Inspection of the fine scale variation in local scaled recombination rates estimated in these regions using the HapMap data recently released<sup>15</sup> confirms that Region G has one the highest average cM per Mb among the regions we studied (Table 2) and may harbor a couple of hot spot in the middle of the region that may explain the faster decay of LD among pairs (Supplementary Figure S2). Inspection of the sequence properties of regions E and G in the genome browser available at Ensembl (<http://www.ensembl.org/>) did not reveal striking features for either GC content, gene density or SNP density (data not shown).

### Estimation of demographic and evolutionary parameters

We discarded data from Regions E and G and used all data from the 10 remaining genomic regions to estimate population parameters ( $\rho$ ,  $\beta$ ,  $\theta$ ). The rationale for doing so was to only pool data from regions where patterns of LD were as similar as possible to yield a meaningful average. Observed data from the remaining 10 genomic regions were pooled and patterns of LD were summarized by computing the proportion of pairs of loci exhibiting significant LD as a function of genetic distance (Figure 2). This summary of LD patterns was subsequently used, using equation (1), to compute the distance  $d$  between the observed data and a simulated data set obtained under a given ( $\rho$ ,  $\beta$ ,  $\theta$ ) combination. Given that we discarded a region of low LD and a region of high LD, doing so does

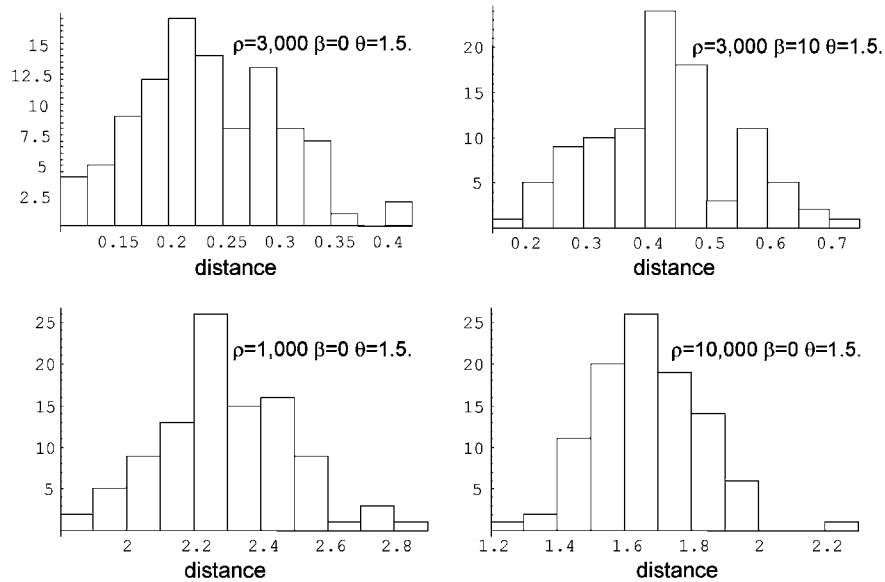


**Figure 2** Proportion of pairs of loci exhibiting significant LD. Data from all regions (but excluding Regions E and G) were pooled and binned in seven classes of genetic distance ( $n = 1039$  pairs).

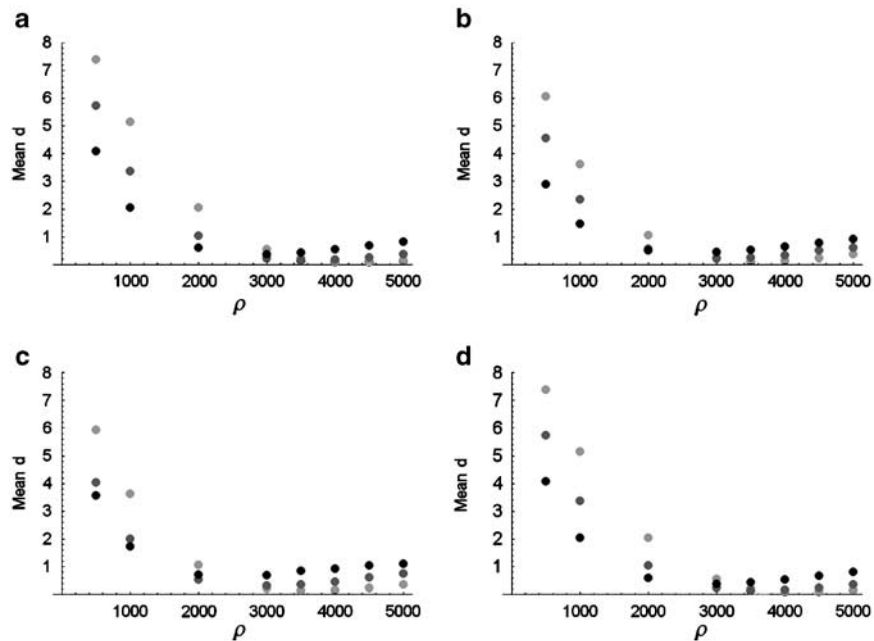
not affect the overall patterns of decrease of LD as summarized by the proportion of pairs exhibiting significant LD within each class of distance. In fact, keeping these two regions in the analysis would result in virtually identical proportions of pairs with significant LD in the various classes of distance. For each parameter ( $\rho$ ,  $\beta$ ,  $\theta$ ) combination, the distance between the observed data and each of the 100 simulated data set was computed. Examples of the distribution of distances between observed and data simulated under four different scenarios are given Figure 3. In all cases examined, the distributions were roughly normal (Figure 3) and could be adequately characterized by their mean and variance (not shown). The mean distance between data simulated under a given scenario and the observed data was used as a measure of fit (Figure 4). Data sets simulated under scenarios involving sharp exponential growth ( $\beta = 1000$ ) yielded data sets with very few polymorphic pairs of loci (at the 5% threshold),

making it difficult to compare the simulation to the observed data. Data sets simulated under the condition  $\beta=1000$  had a large distance to the observed data (not shown). These are not examined further here. For a given

$(\theta, \beta)$  value, the mean distance between simulation and data had a single minimum when varying  $\rho$ , the lowest mean distance being achieved for  $\beta=0$  (Figure 4). The value of  $\rho^*$  minimizing the distance between data and



**Figure 3** Distribution of distance  $d$  between the observed data and simulated data sets obtained under four alternative scenarios  $(\rho, \theta, \beta)$ . All simulations were run under the  $K$ -alleles model ( $K=50$ ). For each simulated data set, the distance  $d$  to the observed data was calculated using equation (1). Each empirical distribution is based on 100 simulated data sets and therefore comprises 100 counts.



**Figure 4** Mean distance between observed and simulated data as a function of  $\rho$ . Plots of mean distance are given for data simulated under scenarios with various degrees of exponential growth:  $\beta=0$  (open circles),  $\beta=10$  (gray circles) and  $\beta=100$  (black circles). Each dot represent the distance averaged over 100 simulated data sets. Top panels: simulations assuming a stepwise mutation model with  $\theta=1.5$  (a) or  $\theta=3$  (b). Bottom panels: simulations run under the  $K$ -alleles model ( $K=50$ ) with  $\theta=1.5$  (c) or  $\theta=3$  (d).

simulation was somewhat sensitive to  $\theta$  and the mutation model chosen (compare panels of Figure 4). However, for a given  $\theta$ , the value of  $\rho$  minimizing distance was nearly insensitive to  $\beta$ . Simulation of data under a constant effective size or moderate exponential growth ( $\beta$  varying from 0 to 10) suggests simulations obtained under the assumption that a region of 15 cM corresponds to a  $\rho$  of 2500–3500 best fit the data (Figure 4). A point estimate of  $\rho^* = 3000$ , together with the assumption of  $r = 10^{-8}$  (1 cM equals 1 Mb), yields an effective population size of the Icelandic population of  $N_e = 3000 / (4 * 0.15) = 5000$  individuals.

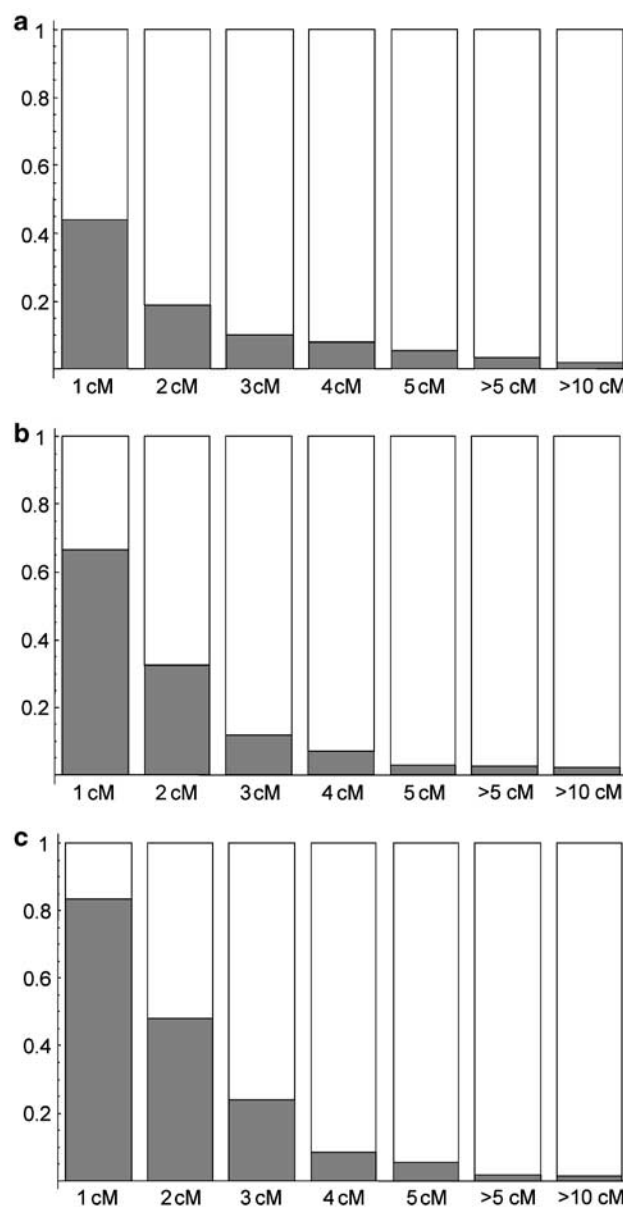
### Expected extent of LD between SNP markers

Assuming that a region of 15 cM corresponds to  $\rho = 3000$ , we simulated a set of 100 evenly spaced SNP markers with minor allele frequencies of 5% and investigated the LD between these. Clearly, SNP markers have lower power to detect significant LD. However, with the large sample sizes used here (1300 diploid individuals simulated), even SNPs may exhibit more LD than expected by chance alone at a range of 3–4 cM in the Icelandic population (Figure 5). The range of detectable LD (as measured by the range of distance where the proportion of pairs of SNPs with significant LD as higher than expected by chance alone) in our simulations extends further than previous simulation studies.<sup>16</sup> This is easily explained by the lower effective size we simulated, the fact that other studies have used a different criteria such as the half-life of LD decay<sup>16</sup> to characterize the range of LD, as well as the large sample size (1300 individuals) we used. Given the large sample size simulated here, many pairs of loci will display significant LD along a region albeit with values of  $r^2$  probably much lower than what is needed to ‘tag’ efficiently these regions.<sup>17</sup>

### Discussion

#### Patterns of observed LD and prospects for association mapping

We have studied a large data set consisting solely of unphased data. It is experimentally difficult and very time consuming to obtain phased data, especially for markers that are separated by more than 10 kb.<sup>18</sup> Here, we show that, instead of inferring haplotypes and then analyzing these as if they were true haplotypes, the direct use of unphased measures of LD still allows us to extract substantial information about the patterns of LD and its decrease with distance. Our study validates recent predictions obtained through a Monte Carlo simulation study.<sup>19</sup> Furthermore, the use of unphased measures of LD, such as Weir’s  $\Delta$ , allows making inferences that are robust to deviations from HWE within the population studied. The hypothesis of HWE is often needed for inferring haplotypes and can be violated even in seemingly homogeneous



**Figure 5** Decrease of the proportion of pairs in significant LD with distance: comparing simulated microsatellites and SNP data. (a) Data comprises 10 genomic regions of length  $\rho$  and 100 SNPs per region ( $n = 49\,500$  pairs in total). Simulations were based on the following parameters  $\rho = 4000$ ,  $\beta = 0$ . (b) Data comprises 10 genomic regions of length  $\rho$  with 15 microsatellites per region ( $n = 12,000$  pairs). Simulations were based on the following parameters:  $K$ -alleles mutation model ( $K = 50$ ),  $\rho = 4000$ ,  $\beta = 0$ ,  $\theta = 1.5$ . (c) Microsatellite simulated data set. Simulations were based on the following parameters:  $K$ -alleles mutation model ( $K = 50$ ),  $\rho = 4000$ ,  $\beta = 0$  and  $\theta = 3$ .

populations such as the Icelandic one that still may exhibit mild population substructure.<sup>20</sup> Altogether, this *a priori* bodes well of the use of large quantity of unphased data to achieve genome-wide coverage.

Our study shows that the decay of LD with distance in Iceland can be approximated by a simple model of a panmictic population of constant size. Microsatellites reveal more significant LD than expected by chance alone over distances of 4 cM (on average 3 Mb), where 35% of pairs (164/471) exhibit significant LD. There is no apparent sign of a strong influence of population structure in the population sample used here, as we see virtually no cases (four among 1235 pairs examined, Table 1) of significant LD at ranges larger than 5 cM despite numerous comparisons at this range. A question that remains is whether the mere existence of significant LD is sufficient for successful LD mapping. Ultimately, the power to detect an unknown susceptibility variant through a marker is dependent on the phenotypic effect of the true variant times the amount of LD between the marker and the causative variant measured through the  $r^2$  measure of LD.<sup>14</sup> As data on SNPs accumulate rapidly in the Icelandic population, a natural extension of this work will be to (1) contrast predicted patterns of LD between SNPs based on our very simple population model calibrated using our microsatellite data sets and (2) investigate which additional features need to be added to this model (demography, variation in recombination rates) to account for the features of the patterns of LD observed using SNPs. Calibrating a simulation model that remain parsimonious in the number of parameters, while capturing the crucial features of LD patterns, will be an important tool to conduct realistic simulations to predict the power of alternative association mapping strategies (see Schaffner *et al*<sup>5</sup> for a recent example using the data from the Phase 1 of the HapMap project).

### Estimation of scaled recombination rates and effective size in the Icelandic population

We have studied patterns of LD in a large sample of Icelandic individuals and used patterns of LD to estimate an effective size for the Icelandic population. We found that simulation of the patterns of LD under a simple model assuming a constant population size through time could fit the observed patterns of LD in our sample fairly well. Based on (1) the  $\rho^*$ -values (2500–3500) that yielded the best fit to the data, and (2) assuming a rate of recombination of 1 cM per Mb – a reasonable assumption for the regions surveyed here when comparing physical and genetic maps (Table 2) – and therefore setting  $r$  to  $10^{-8}$ , we estimated an effective size of approximately 5000 individuals. Using our minimum distance scheme, we only obtain a point estimate of  $\rho$ . By inspecting the distribution of distances, we can bracket a set of parametric values of  $\rho$  that yields distances that are not significantly greater than the minimum observed. This leads to a rough confidence interval around  $\rho^*$  of (2000–5000). Note that this confidence interval is somewhat sensitive to values assumed for  $\theta$  and the underlying mutation model of microsatellites.

A number of studies using microsatellites for studying the patterns of LD in other isolated European populations, such as the Faroe Islands<sup>21</sup> and various isolates within the Finnish<sup>22</sup> and Swedish populations,<sup>23</sup> are potentially available for comparison. Unfortunately, no estimates of the scaled recombination rate,  $\rho$ , were reported in those studies. Comparison across studies is then very difficult as the differences in the range of LD detected can be due to differences in the scaled recombination rate (which includes differences in true recombination rates as well as differences in the realized effective size of these isolates) but also to vast differences in the sample size used in these studies and ours.

Mean estimates of  $\rho$  averaged on 1 Mb windows, a scale comparable to our study, obtained using a composite likelihood estimator of  $\rho$ ,  $\rho_{CL}$  (Hudson, 2001) and the SNP data available for three populations involved in the Hap Map project, were recently reported.<sup>24,25</sup> Our point estimate of  $\rho^* = 200$  (130–330) per Mb yields an estimate of  $\rho^* = 0.0002$  per bp (0.00013–0.00033). Our estimate falls directly in the range of mean estimates for  $\rho$  reported in the European–Americans (Utah residents with ancestry from northern and Western Europe,  $\rho_{CL} = 0.000192$  per bp), African–Americans ( $\rho_{CL} = 0.000405$ ) and Han Chinese (from the Los Angeles area,  $\rho_{CL} = 0.000200$ ). Interestingly, these estimates were based on studies using different markers, different estimation procedure and different strategies of genotyping (150 microsatellite markers on 1300 individuals in our case *versus* data from only 23 individuals but genotyped for 1 586 383 SNPs).

A number of cautionary remarks are needed before discussing the implications of this result. We have simulated data under simple demographic scenarios to investigate whether such simple models could match observed LD patterns. This by no means implies that the complex demographic history of Iceland should be viewed as constant through time or even adequately approximated by the simple exponential growth scenario we also used. The mere existence of an effective size adequately capturing the rate of coalescence of genes in a real population that has undergone fluctuating sizes in the past can be questioned.<sup>26</sup> Austerlitz and Heyer,<sup>27</sup> studying the isolated population of Saguenay-Lac Saint Jean (Québec), also emphasized that beyond the effects of variable population size and overlapping generations, the transmission of reproductive success from parents to offspring could dramatically lower effective size relative to the census size. In that regard, the Wright–Fisher model, upon which the whole concept of effective size is based, may not be adequately modeling genetic drift in the Icelandic population. However, we feel that fitting observed patterns of LD using a simple Wright–Fisher model allows us to obtain an effective size estimate that can in turn be used to make some testable predictions about patterns of LD in other genomic regions or different markers such as SNPs.



Previous estimates of effective size estimates in human population relied on different methods and were mostly based on levels of polymorphism. Levels of polymorphism at a single locus can be used to estimate the product of effective population size and mutation rate (at that locus). If one assumes a mutation rate for the sequence or the marker used, one can obtain an estimate of the (long-term) effective population size. Harpending *et al*<sup>28</sup> reviewed these methods and, with the data available at that time, they suggested a small effective population size of the entire human population, on the order of 10 000 breeding individuals, due to reduced demographic population size in the Pleistocene period. Sherry *et al*<sup>29</sup> used the observed distribution of sample frequencies of 13 dimorphic *Alu* elements. They used coalescence theory to compute the expected total genealogies branch lengths for monomorphic and dimorphic elements, leading to an estimate of human effective population size of ~18 000 during the last one to two million years. In that respect, the estimate we obtain here is very large given that Iceland represents a rather small fraction of the contemporary human genetic diversity. One possibility is that estimates of long-term effective size based on LD are based on relatively recent coalescence time. Hayes *et al*<sup>6</sup> obtained similar results using simulation with phased data and argued that one could actually obtain estimates of  $N_e$  at different times in the past by using pairs of loci with different amounts of recombination (the tighter the linkage between markers used, the older the time back in time). As a consequence, our estimation method yields estimates of  $N_e$  that are less affected by ancient demographic events and thus much higher than estimates based on levels of polymorphism alone.

In conclusion, we have shown that using LD measures requiring only unphased genotypic data from linked loci still convey substantial information about effective recombination rates and past  $N_e$ . We have illustrated this method by analyzing a large sample of individuals and estimating the scaled recombination rate in the Icelandic population. We found that the data were broadly consistent with an estimated scaled recombination rate of  $\rho \sim 200$  per cM. This estimate will be of importance for predicting, through simulations, the power of future association mapping studies using the Icelandic population. Last, our approach for the estimation of  $\rho$  from LD patterns can be used in populations where deviations from HWE proportions occur due to either nonrandom mating or hidden population structure. As such, it should be widely applicable not only to analyze LD in a variety of human populations but also in other species, such as numerous plant species that reproduce through partial self fertilization, where the HWE assumption is untenable and phased data is still difficult and time consuming to obtain.

## Acknowledgements

We thank EB Knudsen for improving the English of this manuscript, BK Ehlers for discussion, and T Damm Als and three anonymous reviewers for their comments.

## References

- 1 Helgason A, Sigurethardottir S, Gulcher JR, Ward R, Stefansson K: mtDNA and the origin of the Icelanders: deciphering signals of recent population history. *Am J Hum Genet* 2000; **66**: 999–1016.
- 2 Rafnar T, Thorlacius S, Steingrimsdottir E *et al*: The Icelandic cancer project – a population-wide approach to studying cancer. *Nat Rev Cancer* 2004; **4**: 488–492.
- 3 Ohta T, Kimura M: Linkage disequilibrium due to random genetic drift. *Genet Res* 1969; **13**: 47–55.
- 4 Pritchard JK, Przeworski M: Linkage disequilibrium in humans: models and data. *Am J Hum Genet* 2001; **69**: 1–14.
- 5 Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D: Calibrating a coalescent simulation of human genome sequence variation. *Genome Res* 2005; **15**: 1576–1583.
- 6 Hayes BJ, Visscher PM, McPartlan HC, Goddard ME: Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Res* 2003; **13**: 635–643.
- 7 Hill WG: Linkage disequilibrium among multiple neutral alleles produced by mutation in finite population. *Theor Popul Biol* 1975; **8**: 117–126.
- 8 Hill WG: Estimation of effective population size from data on linkage disequilibrium. *Genet Res* 1981; **38**: 209–216.
- 9 Weir BS: *Genetic Data Analysis*. Sunderland, MA: Sinauer, 1996.
- 10 Schaid DJ: Linkage disequilibrium testing when linkage phase is unknown. *Genetics* 2004; **166**: 505–512.
- 11 Storey JD, Tibshirani R: Statistical significance for genomewide studies. *Proc Natl Acad Sci USA* 2003; **100**: 9440–9445.
- 12 Kong A, Gudbjartsson DF, Sainz J *et al*: A high-resolution recombination map of the human genome. *Nat Genet* 2002; **31**: 241–247.
- 13 Maillund T, Schierup M, Pedersen C, Mechlenborg P, Madsen J, Schausser L: CoaSim: a flexible environment for simulating genetic data under coalescent models. *BMC Bioinform* 2005; **6**: 252.
- 14 Hein J, Schierup MS, Wiuf C: *Gene Genealogies, Variation And Evolution: A Primer in Coalescent Theory*. Oxford: Oxford University Press, 2004.
- 15 Altshuler D, Brooks LD, Chakravarti A, Collins FS, Daly MJ, Donnelly P: A haplotype map of the human genome. *Nature* 2005; **437**: 1299–1320.
- 16 Reich DE, Cargill M, Bolk S *et al*: Linkage disequilibrium in the human genome. *Nature* 2001; **411**: 199–204.
- 17 de Bakker PI, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, Altshuler D: Efficiency and power in genetic association studies. *Nat Genet* 2005; **37**: 1217–1223.
- 18 Raymond CK, Subramanian S, Paddock M *et al*: Targeted, haplotype-resolved resequencing of long segments of the human genome. *Genomics* 2005; **86**: 759–766.
- 19 Morris AP, Whittaker JC, Balding DJ: Little loss of information due to unknown phase for fine-scale linkage-disequilibrium mapping with single-nucleotide-polymorphism genotype data. *Am J Hum Genet* 2004; **74**: 945–953.
- 20 Helgason A, Yngvadottir B, Hrafnkelsson B, Gulcher J, Stefansson K: An Icelandic example of the impact of population structure on association studies. *Nat Genet* 2005; **37**: 90–95.
- 21 Jorgensen TH, Degn B, Wang AG *et al*: Linkage disequilibrium and demographic history of the isolated population of the Faroe Islands. *Eur J Hum Genet* 2002; **10**: 381–387.
- 22 Varilo T, Paunio T, Parker A *et al*: The interval of linkage disequilibrium (LD) detected with microsatellite and SNP markers

- in chromosomes of Finnish populations with different histories. *Hum Mol Genet* 2003; **12**: 51–59.
- 23 Johansson A, Vavruch-Nilsson V, Edin-Liljegren A, Sjolander P, Gyllensten U: Linkage disequilibrium between microsatellite markers in the Swedish Sami relative to a worldwide selection of populations. *Hum Genet* 2005; **116**: 105–113.
- 24 Hudson RR: Two-locus sampling distributions and their application. *Genetics* 2001; **159**: 1805–1817.
- 25 Serre D, Nadon R, Hudson TJ: Large-scale recombination rate patterns are conserved among human populations. *Genome Res* 2005; **15**: 1547–1552.
- 26 Sjodin P, Kaj I, Krone S, Lascoux M, Nordborg M: On the meaning and existence of an effective population size. *Genetics* 2005; **169**: 1061–1070.
- 27 Austerlitz F, Heyer E: Social transmission of reproductive behavior increases frequency of inherited disorders in a young-expanding population. *Proc Natl Acad Sci USA* 1998; **95**: 15140–15144.
- 28 Harpending HC, Batzer MA, Gurven M, Jorde LB, Rogers AR, Sherry ST: Genetic traces of ancient demography. *Proc Natl Acad Sci USA* 1998; **95**: 1961–1967.
- 29 Sherry ST, Harpending HC, Batzer MA, Stoneking M: Alu evolution in human populations: using the coalescent to estimate effective population size. *Genetics* 1997; **147**: 1977–1982.

Supplementary Information accompanies the paper on European Journal of Human Genetics website (<http://www.nature.com/ejhg>)