European Journal of Human Genetics (2004) 12, 527–534
© 2004 Nature Publishing Group All rights reserved 1018-4813/03 $30.00
www.nature.com/ejhg

ARTICLE

Linkage disequilibrium in young genetically isolated Dutch population

Yurii S Aulchenko*[,1,2], Peter Heutink[1], Ian Mackay[3], Aida M Bertoli-Avella[1], Jan Pullen[3], Norbert Vaessen[1], Tessa AM Rademaker[1], Lodewijk A Sandkuijl[4,6], Lon Cardon[5], Ben Oostra[1] and Cornelia M van Duijn[1]

[1]*Department of Epidemiology and Biostatistics and Department of Clinical Genetics, Erasmus Medical Center Rotterdam, The Netherlands;* [2]*Institute of Cytology and Genetics SD RAS, Novosibirsk, Russia;* [3]*Oxagen limited, UK;* [4]*Leiden University Medical Centre, The Netherlands;* [5]*Wellcome Trust Center for Human Genetics, UK*

The design and feasibility of genetic studies of complex diseases are critically dependent on the extent and distribution of linkage disequilibrium (LD) across the genome and between different populations. We have examined genomewide and region-specific LD in a young genetically isolated population identified in the Netherlands by genotyping approximately 800 Short Tandem Repeat markers distributed genomewide across 58 individuals. Several regions were analyzed further using a denser marker map. The permutation-corrected measure of LD was used for analysis. A significant ($P<0.0004$) relation between LD and genetic distance on a genomewide scale was found. Distance explained 4% of the total LD variation. For fine-mapping data, distance accounted for a larger proportion of LD variation (up to 39%). A notable similarity in the genomewide distribution of LD was revealed between this population and other young genetically isolated populations from Micronesia and Costa Rica. Our study population and experiment was simulated *in silico* to confirm our knowledge of the history of the population. High agreement was observed between results of analysis of simulated and empirical data. We conclude that our population shows a high level of LD similar to that demonstrated previously in other young genetic isolates. In Europe, there may be a large number of young genetically isolated populations that are similar in history to ours. In these populations, a similar degree of LD is expected and thus they may be effectively used for linkage or LD mapping.

European Journal of Human Genetics (2004) **12**, 527–534. doi:10.1038/sj.ejhg.5201188
Published online 31 March 2004

Keywords: population; Europe; genetic isolation; linkage disequilibrium; simulation

Introduction

There is an increasing interest in linkage disequilibrium (LD) mapping. LD mapping has a potential for the precise location of genes involved in common disease, but may also be used to identify novel genes in genomewide scans in population-based studies. Classical linkage analysis in families will typically resolve the position of a novel gene to 10–20 cM, with further precise location obtained by using LD mapping within this region.[1,2] Yet under certain conditions for complex diseases, genomewide LD studies may have more power than linkage studies.[3] The power of these mapping techniques depends strongly on disease allele frequencies and on the extent of disequilibrium between marker and disease alleles.[4] The latter may depend for a large part on the age of mutations involved and on the history of the size and structure of the population studied.

Throughout Europe, there are various genetically isolated populations, founded in the 18th century with

*Correspondence: Dr YS Aulchenko, Genetic Epidemiology Unit, Department of Epidemiology and Biostatistics, Erasmus Medical Center Rotterdam, PO Box 1738, Rotterdam, 3000 DR, The Netherlands.
Tel: +31 104087362;
Fax: +31 104089406; E-mail: i.aoultchenko@erasmusmc.nl

[6]In memoriam.

Received 21 October 2003; revised 22 January 2004; accepted 11 February 2004

subsequent exponential growth. These populations are a valuable resource for mapping genes for complex disease because large segments of DNA are expected to be shared identical-by-descent between carriers of a disease allele. In young isolates, the boundary between linkage and LD mapping becomes obscured. They may provide a researcher with the advantage of extensive pedigree information, which may be utilized by recently developed statistical methods.[5,6] At the same time, the connections between people may be so remote that it makes possible effective fine-mapping. Moreover, smaller isolates show an increased degree of inbreeding that can also be exploited for the purposes of gene mapping.[7]

Empirical studies have demonstrated that the decay of LD with distance does not always follow the pattern expected under standard population genetics models. Compared to expectations, there are examples of too little LD over a few kb and too much at greater distances.[8] Also, other studies have shown that the pattern of LD varies between populations and that its distribution is irregular across the genome.[9,10]

For future LD-mapping projects, it is important to know the expected magnitude and genomewide pattern of LD and how these may vary in different populations. LD should therefore be described in and compared between different populations. One issue, frequently overlooked, is that the comparison of LD between different populations comprises a methodological problem. Two widely used measures of LD (D' and P-values coming from the test of significance of LD) are not suitable for comparison purposes: while D' is biased upwards with decreasing sample size and increasing number of alleles,[8,11–13] the power to detect significant LD increases with sample size. Thus, any studies reporting D' or P-values alone cannot be compared unless similar sample sizes and sets of markers have been used. Recently, a method that makes D' less sensitive to sample size and extreme marker allele frequencies was suggested and implemented in a study of LD in the population of Palau, Micronesia.[11] We have adopted this approach and thus our results should be comparable with these obtained in Palau. By using exact P-values from the test for LD, our study could also be compared with other studies that use a similar sample size.

Here, we examine the amount and decay of LD with genetic distance in a young genetically isolated Dutch population using approximately 800 polymorphic markers distributed throughout the genome. In four autosomal regions, LD is investigated in more detail using a denser marker map in order to investigate the potential for fine-mapping in this population. We compare the amount of LD observed in our study with that in previous studies of LD in young[11,14] and older[15,16] genetic isolates. To assess whether the amount and decay of LD with genetic distance observed in our study population could be explained based on our knowledge of the history of the population, we performed a simulation study and compared the results to our empirical findings.

Materials and methods

Subjects

The subjects were derived from an isolated village in the Southwest of the Netherlands (the GRIP population). The village was founded by approximately 150 people in the middle of the 18th century, and until the last few decades descendants of these founders have lived in social isolation with minimal immigration (less than 5%). From the year 1848, the population has expanded from 700 up to 20 000 inhabitants.

Two (partly overlapping) panels of subjects were studied. To evaluate genomewide LD and LD in specific regions of chromosome 18 and 3, data from an ongoing study of the genetics of Type 2 diabetes were used. Data from 58 spouses of probands were included in the analysis. To evaluate LD at the telomeric region of chromosome 10, we studied 88 subjects, who were healthy controls in ongoing studies of Type 2 diabetes, Parkinson's and Alzheimer's disease. All of the subjects had genotypes available from first-degree relatives, thus allowing haplotype estimation. The study was approved by the medical ethics committee of the Erasmus Medical Center, Rotterdam, and written consent was obtained from all subjects.

Markers and maps

We examined 734 autosomal and 47 X-linked Short tandem repeat (STR) markers. Four genomic regions were subjected to further analysis using a more dense map of STR markers: an 11.9 Mb long telomeric region on chromosome 18p11 (15 markers), a 4.2 Mb telomeric region on chromosome 10q26 (12 markers), a 1.6 Mb centromeric region on chromosome 3p12 (8 markers) and a 12 Mb middle-arm region on chromosome 3p13 (16 markers).

For the whole genome scan, the sex-average Marshfield genetic map was used to define the order of markers and intermarker distances.

For more densely typed regions, none of the genetic maps currently available allowed for the establishment of marker order and intermarker distances accurately. Therefore, for chromosomes 18 and 10, marker order and distances were obtained using the Celera physical map.[17] For the two regions on chromosome 3, the NCBI STS physical map was used. We estimated region-specific genetic to physical map ratios by using genetic and physical distances between the markers flanking a region. For the regions 3p12, 3p13, 10q26 and 18p11, we estimated the genetic to physical map ratio as 0.34 (deCode map), 1.76, 3.63 and 3.76 (Marshfield map) cM/Mb, respectively. Using these estimates and assuming constant cM/Mb ratio across the fine-mapping regions, it is possible to convert distance from the physical to the genetic scale.

Models and statistical methods

For each subject used in the analysis, the haplotypes were estimated using GeneHunter v. 2.1_r3.[18] For estimating X-linked haplotypes, X-GeneHunter-Plus[19] was used. For a few loci, marker genotypes were missing for a large proportion of pedigree members. To minimize the influence of these loci, we dropped from the analysis any pair of loci with fewer than 70 and 50 inferred two-locus haplotypes for autosomes and X-linked markers, respectively.

Haplotype data were subject to an analysis of pairwise linkage disequilibrium. For all pairs of loci on the same chromosome the multiallelic version of the D' statistic was calculated, namely, $D' = \Sigma_{ij} p_i q_j |D'_{ij}|$, where D'_{ij} is Lewontin's standard measure of LD.[20] Permutation analysis was used to correct the bias occurring due to finite sample size.[8,11–13] Alleles were permutated at each locus independently of alleles at other loci. Then, D'_{sim} was calculated as the average of D' over 1000 simulations. Taking the difference between observed and mean simulated values yielded permutation-corrected linkage disequilibrium (D'_{cp}).[11–13] It is interesting to note that the bias uncovered by the correction was large: averaged over loci, the D'_{sim} was 0.317 for the autosomes and 0.324 for the X-chromosome. For chromosomal regions 18p11, 3p12, 3p13, and 10q26, the average bias was equal to 0.295, 0.268, 0.227 and 0.189, respectively.

The significance of LD was tested using the program MLD, which performs a shuffling version of the exact conditional tests for different combinations of allelic and genotypic disequilibrium on haploid and diploid data, or their combination.[21] A total of 5000 permutations were used to assess the *P*-values. D' and D'_{cp} were computed using our own software, miLD 2.0.[13]

A simple model, similar to that of Abecasis *et al*,[10] was used to study the decay of pairwise linkage disequilibrium with time and distance:

$$E(D'_{\mathrm{T}}) = L + (H - L)\exp\{-\theta T\} \qquad (1)$$

Here, θ is recombination fraction between two loci, and T is the number of generations since founding. To allow for LD between unlinked loci and for incomplete LD between tightly linked markers, two parameters are introduced into the model: L, the minimum expected LD between markers, and H, the maximum D' between closely linked markers.

Model (1) is equivalent to the Malecot model[9]. The model's parameters are estimated by minimizing the sum of squares $\mathrm{SSQ} = \Sigma_{i>j}\,(D'_{ij} - E[D'_{ij}])^2$, where the sum is taken over all N pairs of marker loci studied, and $E[D'_{ij}]$ is the expectation of LD between i and j defined by expression (1).

The most general model (H_2) is described by the set of three parameters: $\{H, L, T\}$. Restricting L to 0 results in the nested hypothesis H_1, which assumes that LD between unlinked markers is 0. Note, when the model is applied to D' corrected by permutation, L should be 0 unless a large amount of LD is generated by genetic drift or there is population admixture. Imposing the further restriction, $T = 0$, leads to the null hypothesis H_0 of independence of LD and distance. The above hypotheses are nested, thus the F-test can be used for comparison. It may be argued that the F-ratio test is not appropriate because the sampling distribution of D' is not normal with small sample sizes and/or a small number of different alleles at the loci tested.[22] Under these conditions, resampling techniques may be preferred for hypothesis testing. Therefore, *P*-values and 95% confidence intervals were also obtained using 2500 bootstrap samples, as described in Aulchenko *et al*.[13]

Results

Genomewide LD

In the GRIP population, the mean corrected LD for all pairs of autosomal markers was $D'_{\mathrm{cp}} = 0.\ 0054 \pm 0.0004$. Only pairs of markers belonging to the same linkage group (syntenic markers) were considered. We did not observe extreme values of corrected LD: only for two pairs of markers was D'_{cp} over 0.30. Overall, 7.57% of the disequlibrium values were significant at $\alpha = 0.05$. If we partition the sample according to recombination distance between pairs of loci, we find that a steadily declining fraction is significant for more distant pairs of loci (Table 1, GRIP Autosomes row). Interestingly, while the variance of D's in our sample was 0.00574, the variance of D'_{cp} was only 0.00208. Thus, about 64% of the total variation of D' could be explained by the fixed factors such as distribution of allelic frequencies and sample size.

Under the unrestricted model (H_2), we obtained the maximal corrected LD of 0.057, while LD for unlinked markers was virtually zero (−0.0002). Indeed, model H_1, restricting L to 0, did not differ significantly from H_2 (both asymptotic and empirical $P > 0.6$, Table 2) suggesting that admixture and drift are not generating a detectable LD between unlinked loci in our study population. The test of LD decay with distance (H_1 *versus* H_0) was highly significant (both $P < 0.0004$). However, distance alone explains only 4.4% of total variance in our data set.

As a large proportion of pairs of markers have one marker in common, the data are correlated. To assess whether this departure from independence may affect our results significantly, we repeated the analysis of LD using a sample of independent marker pairs. In all, 104 D'_{cp} values, used in this analysis, were derived from pairs of adjacent markers, with the requirement that these pairs were separated by at least 20 cM. Each marker was involved in only one pair. The results obtained using this sample demonstrated high

Table 1 Number of marker pairs, mean corrected LD ± SE and percent of LD values significant (lower line) for recombination intervals between pairs of loci

Population	Recombination Interval							
	<0.02	*0.02–<0.05*	*0.05–<0.1*	*<0.1*	*0.1–<0.2*	*0.2–<0.3*	*0.3–<0.4*	*>0.4*
GRIP								
Autosomes	65	393	775	1233	1705	2124	2720	3520
	0.05±0.008	0.037±0.003	0.024±0.002	0.030±0.001	0.010±0.001	0.003±0.001	0.000±0.001	0.001±0.001
	35.38±5.98	24.68±2.18	17.68±1.37	20.84±1.16	8.97±0.69	6.40±0.53	4.82±0.41	5.06±0.37
X-chromosomes	31	58	110	199	247	268	202	6
	0.054±0.014	0.035±0.009	0.021±0.006	0.030±0.005	0.004±0.004	0.005±0.004	0.012±0.004	−0.030±0.016
	25.80±7.99	20.69±5.37	9.09±2.75	15.08±2.54	6.48±1.57	7.46±1.61	7.92±1.90	0
Palau								
Autosomes	—	—	—	—	—	—	—	—
	—	—	—	0.031	0.019	0.017	0.012	0.009
	—	—	—	16.2	11.6	11.6	7.1	4.4
*X-chromo*somes	—	—	—	—	—	—	—	—
	—	—	—	0.123	0.041	0.041	0.020	0.026
	—	—	—	44.0	0	13.6	13.6	21.4

Data from GRIP study and Palau[11] are shown.

Table 2 Modeling the decay of disequilibrium with distance, estimated for autosomes, X-chromosome and four different genomic regions

	No.	*Model parameters (H_1)*		*Variance explained*	*H_1 vs H_2*		*H_0 vs H_1*	
Region	*D'_{cp}*	*H*	*T*	*(%)*	*P_A*	*P_B*	*P_A*	*P_B*
Autosomes	11302	0.057 (0.05–0.067)	12 (10.5–13.9)	4.36	0.76	0.63	**<0.0001**[c]	**<1/2500**
Autosomes[a]	104	0.069 (0.044–0.101)	15.8 (4.4–28.6)	5.72	0.9	0.56	**0.014**	**0.004**
X-chromosomes[b]	922	0.053 (0.028–0.090)	12.8 (5.1–28.1)	2.58	0.063	**0.026**[b]	**<0.0001**	**0.001**
3p12	28	0.299 (0.160– 0.462)	200.8 (80.7–335)	39.3	0.73	0.38	**0.0003**	**<1/2500**
3p13	120	0.145 (0.043– 0.350)	63.8 (3.7–171.2)	8.16	0.23	0.11	**0.002**	**0.015**
10q26	66	0.241 (0.009– 1)	1015 (19–2292.3)	15.7	0.42	0.17	**0.001**	**0.025**
18p11	105	0.124 (0.032– 0.179)	289.5 (14.8–571.3)	7.07	0.12	0.1	**0.006**	**0.001**

The number of D'_{cp} values used is given in brackets. Parameter estimates (95% bootstrap confidence interval) and percent of variance explained for the accepted hypotheses and P-value coming from F-ratio test (P_{A}) and bootstrap (P_{B}) are shown. P-values less than 0.05 are in bold.

[a] Only adjacent marker pairs separated by at least 20 cM used.

[b] Parameters of H_2 model are ($H=0.06$ (0.038–0.126), $L=0.006$ (0.000–0.011), $T=19.4$ (11.6–367.7)).

similarity to that obtained using all pairs: the H_1 hypothesis is accepted, while H_0 is rejected. Further, the estimates obtained are very similar to those obtained using all pairs, despite the fact that the sample size was over 100 times smaller (Table 2). These results indicate that the departure from independence is not crucial in our analysis.

To evaluate whether the pattern of disequilibrium differed with chromosome, a separate analysis was carried out for every chromosome. No autosome showed a significant deviation of L from 0 and each chromosome showed significant evidence for decay of LD with distance (all $P \leq 0.002$), except for chromosome 21 and 22 ($P=0.14$ and 0.17). Given the number of typed markers (11 and 13, for chromosome 21 and 22, respectively), it is likely that in these cases we did not have power to reject the null hypothesis. Although most chromosomes gave a consistent estimate of H (between 0.03 and 0.1) and T (between 6 and 23), for two chromosomes a large deviation was observed. For chromosome 2 and 13, H was estimated as 1.0, that is, perfect LD is predicted at very short distances. For chromosome 2, these results were mainly determined by a single D'_{cp} value ($D'_{\mathrm{cp}}=0.36$, $\theta=0.005$, Monte-Carlo $P<0.0002$). Excluding this data point from analysis led to more consistent estimates of $H=0.04$ and $T=10.5$. For chromosome 13, H was also estimated as unity. We did not find a single value determining the result; rather it was determined by a set of closely linked marker pairs (at $\theta \sim 0.03$–0.04) demonstrating relatively high LD.

The mean-corrected LD between 922 pairs of X-linked markers was 0.0114 (±0.002). None of the markers demonstrated corrected LD of more than 0.3. Overall, 8.89% of the disequlibrium values were significant. For the X chromosome, the H_1 hypothesis of no LD between distant markers was rejected based on the empirical estimate of $P = 0.026 \pm 0.003$.

LD in four genomic regions, using a denser map

The results from the analysis of the four genomic regions (chromosomes 3p12, 3p13, 10q26 and 18p11) using a denser map are shown in Tables 2 and 3. If we partition the sample according to physical distance, we find a steady decline of LD (Table 3). As is the case with the whole-genome scan, LD between distant markers is effectively zero thus suggesting that admixture and drift are not generating a detectable LD between unlinked loci in our study population.

The model restricting L to 0 does not differ (all $P > 0.1$) from the model allowing for LD between unlinked loci. At the same time, exclusion of distance from the model (H_0) significantly decreases the fit to the data and H_0 is rejected (all $P < 0.01$) for all four regions.

Although the same model H_1 is accepted for all four genomic regions, the extent and distribution of LD differs (Figure 1). The largest proportion of variance explained by distance is 39.3% for the centromeric region 3p12. The next largest is 15.7% for the telomeric region 10q26, then 8.2% obtained for the middle-arm region 3p13 and 7.1% for the telomeric region 18p11. The estimate of LD at small distances (H) ranges from 0.3 (3p12) to 0.12 (18p11); the T parameter ranges from 64 (3p13) to 1015 (10q26).

After converting distance from the physical to the genetic scale, the estimates of T became 584.1, 36.3, 279.6 and 77 for regions 3p12, 3p13, 10q26 and 18p11, respectively.

LD in simulated data

We simulated our study by modeling a population founded 12 generations ago by 75 spouse pairs. We chose 12 generations not by estimation from this genetic study (which also suggested 12 generations), but rather because from historical records it is known that GRIP was founded approximately 250 years ago, corresponding to 10–14 generations. The number of founders was chosen based on available historic information. The distribution of the number of offspring was set as Poisson with an average of three, which roughly approximates the known growth curve for the GRIP population. The lifespan of an individual was set to two generations. For the simulations we have used the same marker map as in the empirical study. Initial allelic frequencies were set to the values found in our sample. The mutation frequency was set to 0.001. From a resulting population, we sampled randomly 88 chromosomes. All simulations were conducted by the GENOOM program.[23] The simulations were repeated 10 times. Each sample underwent analysis in a manner replicating that for the GRIP sample. The average estimate of parameters were $\{H = 0.095 \pm 0.002,\ L = 0.001 \pm 0.0002,\ T = 13.8 \pm 0.33\}$ with an average proportion of the variance explained equal to $8.9 \pm 0.3\%$. Thus, the estimates of L and T resulting from simulated data did not differ significantly from the estimates obtained in the empirical study (Z-test, $P > 0.05$). However, H (LD at very short distances) was significantly ($P < 0.001$) higher in simulated data than that in the empirical study.

Comparison between GRIP and other populations

We compared LD in the GRIP population with LD in the young genetically isolated populations of Palau, Micronesia[11] and the Central Valley of Costa Rica.[14]

Table 3 Number of marker pairs, mean corrected LD ± SE, percent of LD values significant ($P < 0.05$) for distances between pairs of loci in fine-mapping regions

	Interval (Mb)					
Region	*0–0.2*	*0.2–0.5*	*>0.5–1.0*	*>1.0–2.0*	*>2.0–5.0*	*>5.0*
3p12	3	9	10	6	0	0
	0.228±0.080	0.158±0.030	0.057±0.030	0.049±0.027	—	—
	100	88.9±11.1	30.0±15.3	33.3±21.1	—	—
3p13	5	3	6	19	42	45
	0.165±0.079	0.048±0.079	0.120±0.046	0.049±0.018	0.013±0.010	0.026±0.018
	60±24.5	66.7±33.3	50.0±22.4	42.1±11.6	19.0±6.1	8.9±4.3
10q26	7	6	11	14	28	0
	0.066±0.038	−0.006±0.015	−0.002±0.010	0.012±0.007	0.006±0.008	—
	42.9±20.2	0	0	0	3.6±3.6	—
18p11	4	5	6	21	39	30
	0.098±0.015	0.037±0.009	0.018±0.019	0.014±0.010	0.012±0.011	0.002±0.010
	50±28.9	20±20	16.7±16.7	4.8±4.8	10.3±4.9	10.0±5.6

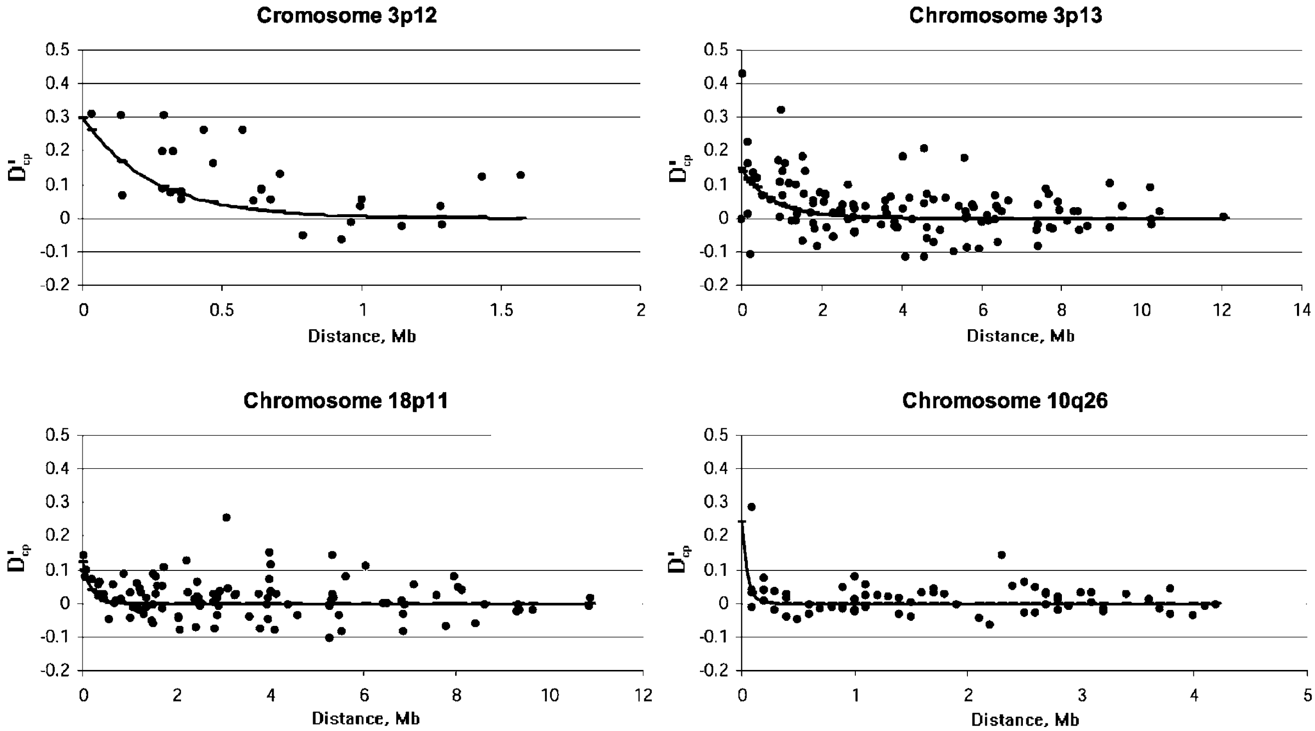

Figure 1 D'_{cp} *versus* physical distance in four genomic regions. The solid lines correspond to the expected LD under the model of decay explained by distance.

In. the Palau study, 84 individuals were used to study LD in autosomes and 60 males were investigated to study the X-chromosome. The relation between corrected LD and the recombination fraction followed a linear regression model. Adding a quadratic term into the regression did not improve the fit.[11] In contrast, in our data we found that adding a quadratic term improved the model significantly ($P<0.0001$), while the exponential model explained the largest proportion of variance (% of variance explained by the linear, quadratic and exponential regression were 2.79, 4.2 and 4.4, respectively). At shorter distances between loci ($\theta<0.1$) LD in GRIP was very close to that in Palau (Table 1). At larger distances ($\theta>0.1$), LD starts decaying more strongly in GRIP. As the density of our marker set was nearly twice the density used in Palau, we conclude that LD is likely to be higher in Palau than in GRIP, especially at longer distances ($\theta>0.1$).

On the X-chromosome, LD in GRIP was much smaller than that in Palau (see Table 1). Again, Devlin *et al*[11] found that adding the quadratic term in the regression model did not improve the fit to the data, while in GRIP the quadratic term was significant ($P<0.0001$), and the exponential model gave the best fit to the data (% of variance explained by linear and quadratic models were 1.04 and 2.52, respectively, while H_2 explained 2.9%). We found that the distribution of LD at the X chromosome is similar to the distribution found for the autosomes. In contrast, Devlin *et al*[11] found LD on the X-chromosome (mean corrected D' of 0.12 for $\theta<0.1$) to be four times larger than LD for the autosomes. This has also been noted by the authors and remains to be explained.

We also compared our results with results from the previous genomewide evaluation of LD in a young genetically isolated population from the Central Valley of Costa-Rica (CVCR).[14] In the CVCR study, 157 chromosomes, nontransmitted to individuals with bipolar disorder, were studied. Although this sample is slightly larger, the power may be approximately comparable with that of our study (116 chromosomes). From Figure 2, it can be seen that the extent and distribution of LD is similar in GRIP and CVCR. The significance of LD tends to be higher in CVCR at smaller distances, which can be probably explained by greater sample size. However, the decline tends to be slower in GRIP, suggesting that the GRIP effective population size is smaller.

The results of the genomewide evaluation of the percentage of significant LD coefficients at intervals 0–0.02θ and 0.02–0.05θ in GRIP (Table 1) indicate a significant increase of LD at these distances. This is also true for selected regions in older populations which were subject to genetic drift (Saami, Gavoi; Table 4).[15,16] In contrast, evaluation of these regions in older isolates, which underwent exponential expansion (Sardinia, Finland), and in the general UK population reveals much lower levels of LD.[15,16] Additionally, LD declines very fast in these populations (only for pairs of markers separated by less than 0.02θ are significant results found, Table 4).

Thus, at small distances (<10 cM) there is much similarity in LD between young genetically isolated populations (GRIP, Palau and CVCR): the percent of significant *P*-values is similar between GRIP, Palau and CVCR, and mean permutation-corrected D' is similar between GRIP and Palau. The drop of LD with distance is steadier in young isolates compared to older expanding isolates.

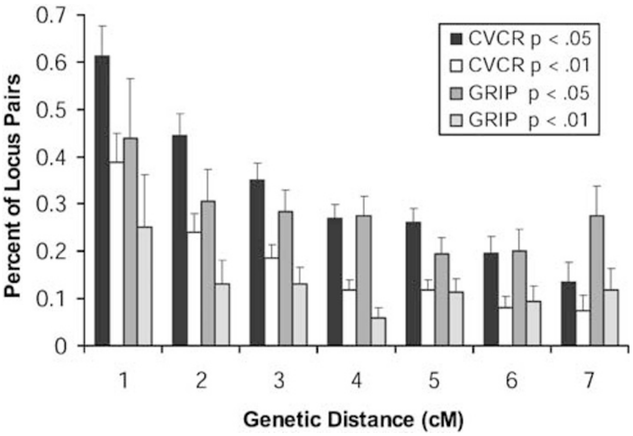

Figure 2 Percentage of *P*-values <0.05 and <0.01 between 630 and 1012 pairs of adjacent markers in the GRIP and CVCR populations, respectively. Distance is given as right boundary of 1 cM – binning interval (1: all marker pairs at <1 cM, 2: all marker pairs at <2 and ≥1 cM, etc.).

Discussion

We examined genomewide LD in a young genetically isolated Dutch population and characterized in detail four genomic regions using a dense marker map. As expected, we found a significant ($P<0.0004$) relation between LD and genetic distance. More importantly, LD was still detectable at large distances up to 20 cM. We did not detect LD between unlinked autosomal loci, suggesting that admixture and drift are not generating a detectable LD between unlinked loci in our study population.

The pattern of LD in GRIP was studied using the most likely haplotypes for each individual as input data. These were estimated from pedigree data using the Lander–Green algorithm, as implemented in GeneHunter.[18] Since this method assumes absence of LD between markers, concerns have been expressed that it may be inaccurate under some circumstances.[24] Fallin and Schork[25] demonstrated that although the EM algorithm gives good accuracy when estimating LD between SNPs using samples of greater than 100 people, accuracy decreases with increased heterozygosity and reduced sample size. Given the nature of our data (a sample of 58 people, highly polymorphic STR markers), the EM algorithm is not a suitable alternative method in our case. However, given the density of the map used and the fact that genotype data also exist for spouse and children for most subjects in the study, pedigree-based methods will assure good accuracy.[26]

The results obtained in our simulation study were close to those obtained in our empirical study. Although the estimates of L and T resulting from simulated data were within the 95% confidence interval for the estimates obtained in the empirical study, H (LD at very short distances) was not. This indicates that LD in GRIP is less than expected under the simple model we used for our simulations. There are a few possible explanations for this discordance. First, the modeled effective population size might be less than the actual one. That is, either the number of founders in the GRIP population was more than

Table 4 Number of marker pairs and percent of LD values significant (lower line) for recombination intervals between pairs of loci on X-chromosome

Population	*No. of chromosomes*	*Recombination interval*		
		<0.02	*0.02–<0.05*	*0.05–<0.1*
Saami	54	17	4	—
		82.35±9.53	75±25	—
Gavoi	73	17	4	—
		94.12±5.88	75±25	—
Sardinia	73	17	4	—
		11.76±8.05	0	—
Finland	80	26	15	8
		7.69±5.33	13.33±9.09	0
UK	73	17	4	—
		11.76±8.05	0	—

Data from Saami, Gavoi, Sardinia, UK[15] and Finland[16] are shown.

150, or there was higher immigration. Also possible heterogeneity of the population's growth parameters across time that was not accounted for in our simulation study may change the effective population size.

It appears from our simulation study (9% of total variance explained by genetic distance) that on a genome-wide scale one should not expect a large proportion of the variance to be explained by genetic distance, given the marker map used and the history of the population. Thus, in our study a very large proportion of variance of LD is a consequence of the highly stochastic nature of genetic processes in natural populations.

The distribution of LD is highly irregular across the genome.[9,10,27] The choice of the density of a marker map to 'catch' a risk factor would have to take the regional variation in LD into account as suggested by our results for chromosome 10q26, where we see that LD is dropping very fast compared to the other fine-mapping regions we studied.

We also compared LD in GRIP with LD in other young genetically isolated populations in Palau, Micronesia.[11] and the Central Valley of Costa Rica.[14] At smaller distances (<10 cM) there is much similarity in LD between young genetically isolated populations. In contrast, the drop of LD with distance was much faster in older isolates, which underwent exponential growth. This implies that for a young isolate the fact of recent isolation/fast growth is far more important than the geographical position and the ethnic background of a population. In Europe, there are many young genetically isolated populations that are very similar in history to the GRIP population. In these populations, a similar degree of LD is expected and thus they may be effectively used for mapping genes underlying complex diseases.

Acknowledgements

We thank Dr Tatiana I Axenovich, Institute of Cytology and Genetics, Novosibirsk, Russia, for useful discussion and for the METHGI subroutine. We are also grateful to Susan Service (University of California at Los Angeles, USA) for valuable discussion and data on LD in the population of Central Valley of Costa Rica. This work was supported by the Dutch Diabetes Foundation (DFN) and the Netherlands Organization for Scientific Research (NWO). The financial support from Oxagen Ltd is acknowledged.

References

1 Boehnke M: Limits of resolution of genetic linkage studies: implications for the positional cloning of human disease genes. *Am J Hum Genet* 1994; **55**: 379–390.

2 Lander ES: The new genomics: global views of biology. *Science* 1996; **274**: 536–539.

3 Risch N, Merikangas K: The future of genetic studies of complex human diseases. *Science* 1996; **273**: 1516–1517.

4 Muller-Myhsok B, Abel L: Genetic analysis of complex diseases. *Science* 1997; **275**: 1328–1329, ; author reply 1329–1330.

5 Sobel E, Lange K: Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. *Am J Hum Genet* 1996; **58**: 1323–1337.

6 Abney M, Ober C, McPeek MS: Quantitative-trait homozygosity and association mapping and empirical genomewide significance in large, complex pedigrees: fasting serum-insulin level in the Hutterites. *Am J Hum Genet* 2002; **70**: 920–934.

7 Wright A, Charlesworth B, Rudan I, Carothers A, Campbell H: A polygenic basis for late-onset disease. *Trends Genet* 2003; **19**: 97–106.

8 Weiss KM, Clark AG: Linkage disequilibrium and the mapping of complex human traits. *Trends Genet* 2002; **18**: 19–24.

9 Collins A, Lonjou C, Morton NE: Genetic epidemiology of single-nucleotide polymorphisms. *Proc Natl Acad Sci USA* 1999; **96**: 15173–15177.

10 Abecasis GR, Noguchi E, Heinzmann A *et al.*: Extent and distribution of linkage disequilibrium in three genomic regions. *Am J Hum Genet* 2001; **68**: 191–197.

11 Devlin B, Roeder K, Otto C, Tiobech S, Byerley W: genomewide distribution of linkage disequilibrium in the population of Palau and its implications for gene flow in Remote Oceania. *Hum Genet* 2001; **108**: 521–528.

12 Teare MD, Dunning AM, Durocher F, Rennart G, Easton DF: Sampling distribution of summary linkage disequilibrium measures. *Ann Hum Genet* 2002; **66**: 223–233.

13 Aulchenko YS, Axenovich TI, Mackay I, van Duijn CM: miLD and booLD programs for calculation and analysis of corrected linkage disequilibrium. *Ann Hum Genet* 2003; **67**: 372–375.

14 Service SK, Ophoff RA, Freimer NB: The genomewide distribution of background linkage disequilibrium in a population isolate. *Hum Mol Genet* 2001; **10**: 545–551.

15 Zavattari P, Deidda E, Whalen M *et al.*: Major factors influencing linkage disequilibrium by analysis of different chromosome regions in distinct populations: demography, chromosome recombination frequency and selection. *Hum Mol Genet* 2000; **9**: 2947–2957.

16 Varilo T, Laan M, Hovatta I, Wiebe V, Terwilliger JD, Peltonen L: Linkage disequilibrium in isolated populations: Finland and a young sub-population of Kuusamo. *Eur J Hum Genet* 2000; **8**: 604–612.

17 Venter JC, Adams MD, Myers EW *et al.*: The sequence of the human genome. *Science* 2001; **291**: 1304–1351.

18 Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES: Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am J Hum Genet* 1996; **58**: 1347–1363.

19 Kong A, Cox NJ: Allele-sharing models: LOD scores and accurate linkage tests. *Am J Hum Genet* 1997; **61**: 1179–1188.

20 Lewontin RC: The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics* 1964; **49**: 49–67.

21 Zaykin D, Zhivotovsky L, Weir BS: Exact tests for association between alleles at arbitrary numbers of loci. *Genetica* 1995; **96**: 169–178.

22 Zapata C, Carollo C, Rodriguez S: Sampling variance and distribution of the D' measure of overall gametic disequilibrium between multiallelic loci. *Ann Hum Genet* 2001; **65**: 395–406.

23 Quesneville H, Anxolabehere D: GENOOM: a simulation package for GENetic Object Oriented Modelling. *Ann Hum Genet* 1997; **61**: 543.

24 Schaid DJ, McDonnell SK, Wang L, Cunningham JM, Thibodeau SN: Caution on pedigree haplotype inference with software that assumes linkage equilibrium. *Am J Hum Genet* 2002; **71**: 992–995.

25 Fallin D, Schork NJ: Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data. *Am J Hum Genet* 2000; **67**: 947–959.

26 Schaid DJ: Relative efficiency of ambiguous *vs* directly measured haplotype frequencies. *Genet Epidemiol* 2002; **23**: 426–443.

27 Reich DE, Schaffner SF, Daly MJ *et al.*: Human genome sequence variation and the influence of gene history, mutation and recombination. *Nat Genet* 2002; **32**: 135–142.