nature genetics

Evaluating potential for whole-genome studies in Kosrae, an isolated population in Micronesia

Penelope E Bonnen[1,9], Itsik Pe'er[2,3,9], Robert M Plenge[3,4], Jackie Salit[1], Jennifer K Lowe[1], Michael H Shapero[5], Richard P Lifton[6,7], Jan L Breslow[1], Mark J Daly[2,3], David E Reich[8], Keith W Jones[5], Markus Stoffel[1], David Altshuler[3,4,8] & Jeffrey M Friedman[1,7]

Whole-genome association studies are predicted to be especially powerful in isolated populations owing to increased linkage disequilibrium (LD) and decreased allelic diversity, but this possibility has not been empirically tested[1–3]. We compared genome-wide data on 113,240 SNPs typed on 30 trios from the Pacific island of Kosrae to the same markers typed in the 270 samples from the International HapMap Project[4,5]. The extent of LD is longer and haplotype diversity is lower in Kosrae than in the HapMap populations. More than 98% of Kosraen haplotypes are present in HapMap populations, indicating that HapMap will be useful for genetic studies on Kosrae. The long-range LD around common alleles and limited diversity result in improved efficiency in genetic studies in this population and augments the power to detect association of 'hidden SNPs'.

The use of LD-based mapping strategies makes it practical to perform whole-genome association studies without typing every common variant in the human genome[6,7]. It has been suggested that the power of such studies is increased in isolated populations, and this has been demonstrated for rare alleles[8]. However, it remains unclear whether LD among common alleles in isolated populations extends significantly farther than in other populations—far enough to have impact for studies that seek to identify alleles contributing to common complex traits[9–11]. Thus, although LD has been observed around rare alleles (such as disease genes) in isolated populations, facilitating disease gene mapping and cloning, this finding has not been extended to the common alleles and haplotypes that are the focus of the HapMap project. Even among isolated populations, variation in demographic history has powerful consequences for LD structure around rare alleles[3,12,13], necessitating an evaluation of each population to assess the power of an LD-based mapping strategy. Thus, with the intense worldwide effort to generate a haplotype map for the general population, it is important to establish the utility of marker sets developed using samples from the HapMap project for use in other populations. We addressed these issues using samples from the island of Kosrae, Federated States of Micronesia[14], which was settled by a small number of Micronesian founders ~2,000 years ago[15].

In the current study, we generated a genome-wide high-density dataset of ~110,000 SNPs to assess the effects of Kosraen population history on genomic variation in 30 Kosraen (KOS) trios. These same SNPs have been typed in the samples used in the HapMap project (see Methods). The 30 Kosraen trios were chosen such that all trios would be five or more generations separated from each other, which represents the most distant branches of the Kosrae pedigree. Thus, the individuals in these trios allow a fuller sampling of the range of diversity on Kosrae than if we had analyzed a random set of individuals. Comparing data on the same set of markers typed in samples included in the HapMap project permits examination of the relative extent of LD and haplotype diversity in Kosrae compared with these other populations.

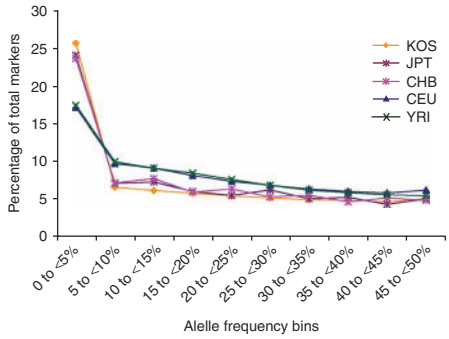

Figure 1 Allele frequency distribution. Allele frequency distribution for ~110,000 SNPs typed in five populations.

[1]Rockefeller University, 1230 York Avenue, New York, New York 10021, USA. [2]Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, Massachusetts 02142, USA. [3]Broad Institute of Harvard and the Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [4]Department of Medicine, Harvard Medical School, Boston, Massachusetts 02115, USA. [5]Affymetrix Inc., 3380 Central Expressway, Santa Clara, California 95051, USA. [6]Departments of Medicine and Genetics, Yale University School of Medicine, New Haven, Connecticut 06510, USA. [7]Howard Hughes Medical Institute. [8]Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA. [9]These authors contributed equally to this work. Correspondence should be addressed to J.F. (friedj@rockefeller.edu).

Received 29 September 2005; accepted 17 November 2005; published online 22 January 2006; doi:10.1038/ng1712

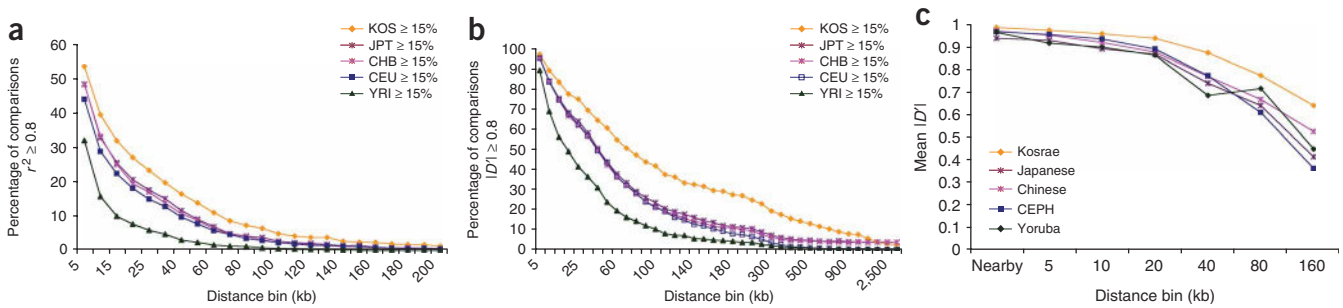

Figure 2 Decay of linkage disequilibrium over distance. LD was measured by pairwise comparison between markers that had minor allele frequency ≥15% and that fell into the same intermarker distance bin. (**a**,**b**) Decay of LD over distance is represented here by the percentage of pairwise comparison scores equal to 0.8 for each distance bin, using the statistic r^2 (in **a**) and the statistic |*D*'| (in **b**). (**c**) LD was measured using an independently ascertained set of SNPs on a separate set of samples.

The allele frequency distribution of the genotyped SNPs in Kosrae was similar to that in other populations for markers with a minor allele frequency (MAF) ≥15% (**Fig. 1**). This confirms that SNP ascertainment strategies did not substantially alter the common part of the frequency spectrum. However, not unexpectedly, a higher proportion of alleles was nonpolymorphic in the samples from KOS, JPT and CHB (~20% of all markers) versus CEU and YRI samples (11% and 9%, respectively). We next assessed the extent of LD among common alleles (MAF >15%) (**Fig. 2**). The half-life of LD decay with genomic distance was substantially longer in Kosrae than in CEU, CHB, JPT and YRI. Measured as the fraction of SNP pairs with highly correlated minor alleles ($r^2 > 0.8$), we observed this fraction to be 1.25- to 1.75-fold higher in Kosrae than in reference populations for the entire distance range above 10 kb (**Fig. 2a**). Moreover, for longer genomic distances (>100 kb), LD on Kosrae was twice that of the reference populations, when measured as the fraction of such SNP pairs showing little or no recombination (|*D*'| > 0.8) (**Fig. 2b**). Both LD decay plots show a clear correlation between the extent of LD and population demography since the migration out of Africa. YRI displays the least LD, and CEU, CHB and JPT cluster together with respect to extent of LD. The gap between the curves for the CEU, JPT and CHB populations and curve for the Kosrae population suggests that the additional LD in the Kosrae population is the result of one or more recent demographic events during the history of this population.

We found that the diversity of long-range haplotypes was significantly lower in the Kosrae population than in other populations. As a measure of haplotype diversity, we first considered consecutive sets of markers that spanned approximately 1 Mb (haplotype diversity over shorter distances was also evaluated using a different panel of markers; see below). For these 1-Mb intervals, we counted the total number of haplotypes observed in each region and averaged this number across regions (see Methods). We asked how many haplotypes were needed to account for 95% of chromosomes. On Kosrae, a genome-wide average of 35 haplotypes/Mb accounted for 95% of the chromosomes, whereas 35 haplotypes accounted for only 50% of chromosomes among CEU and 40% of the chromosomes among YRI (**Fig. 3a**). In the CHB and JPT populations, these 35 haplotypes accounted for 75% of the total. In the Asian cohort, it is likely that we are underestimating haplotype diversity even for these specific markers, as fewer chromosomes were analyzed in the Asian groups (88 JPT, 90 CHB) than in KOS, YRI and CEU (120 each).

These data suggest that the Kosraen population shows two substantial differences compared with the other populations: increased LD and reduced haplotype diversity. To rule out the possibility that these conclusions were an artifact of SNP selection process, we performed two validation experiments. We typed an independently ascertained set of SNPs on a different cohort of 14 Kosraen trios and compared the genotype data with that from 14 trios from the CEPH, Han Chinese and Beni (from Nigeria) populations. These 473 SNPs were distributed across 17 unlinked genomic regions of ~150 kb (see ref. 16), with additional SNP discovery by deep resequencing in a diverse sample. Consistent with results from the ~110,000-SNP dataset, these SNPs showed significantly higher LD in the Kosraen samples than in the other samples (**Fig. 2c**). There was also reduced haplotype diversity, with ten haplotypes accounting for 94% of Kosraen chromosomes, 80% of CEPH and Han Chinese and 70% of Beni, using substantially shorter, more densely typed intervals than were tested using the ~110,000-SNP dataset (**Fig. 3b**).

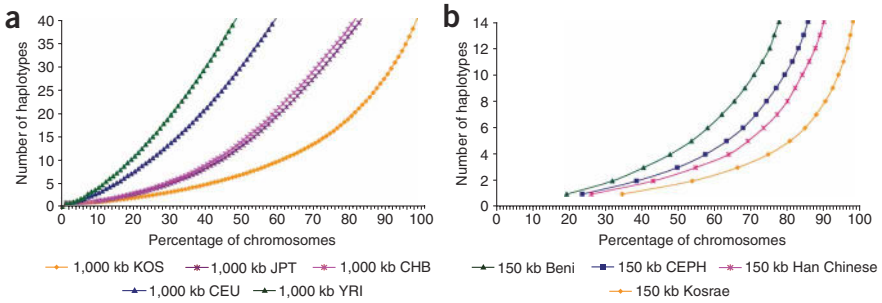

Figure 3 Haplotype diversity. (**a**) The genome was divided into 1-Mb regions, with each region having roughly the same number of markers (~35). Haplotypes were derived and frequencies counted for each region. The number of haplotypes and their frequencies per 1-Mb region was averaged over all regions and plotted against the percentage of total chromosomes the haplotypes accounted for. (**b**) Seventeen unlinked genomic regions spanning ~150 kb were selected, and ~20 SNPs spanning these regions were genotyped. The number of haplotypes in each region and the percentage of chromosomes they account for was averaged over the 17 regions.

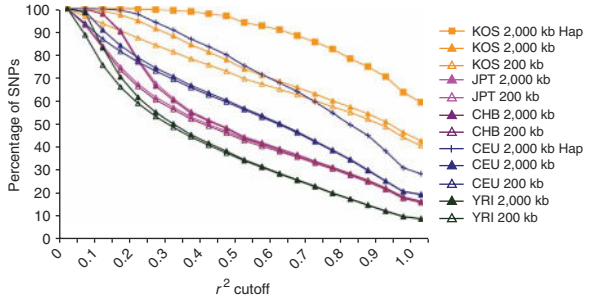

Figure 4 Association study power comparison. We evaluated the effectiveness of these ~110,000 SNPs to detect a potentially causative allele by hiding one SNP and evaluating its best correlate on the array using r^2 as the metric for correlation. This was done using single SNPs within 200 kb of the hidden SNP, single SNPs within 2,000 kb of the hidden SNP and for every possible three-SNP haplotype (KOS and CEU).

We also resequenced Kosrae DNAs to verify that alleles ascertained in reference populations were representative of the common variation in the Kosraen genome and that we were not missing a class of common alleles on Kosrae by using SNPs generated in studies of other populations. We resequenced a total of 80 kb across the same 17 regions referred to above in 15 samples of unrelated Kosraens. We then compared the SNP discovery results with the same regions resequenced in 15 CEPH samples, 15 Han Chinese samples and 15 Beni samples. A total of 441 polymorphic sites were identified from the 120 chromosomes that were sequenced. SNPs that were observed in three or more chromosomes in Kosrae were nearly always seen in CEPH or Han Chinese samples (97/105); allele frequency distributions in these three populations were statistically indistinguishable. Overall, these independent results verify that the ~110,000-SNP set provides excellent representation of the common variants in Kosrae.

To evaluate the effect of increased LD and reduced haplotype diversity on the ability of ~110,000 genotypes to detect a potentially causative allele, we performed an analysis in which we 'hid' one SNP from the set and evaluated its best correlate in the dataset. We observed that the Kosraen samples showed significantly higher levels of correlation between 'hidden' SNPs and the remaining markers as compared with HapMap populations (**Fig. 4**). In the Kosraen samples, 57% of 'hidden' SNPs have a highly correlated[17] ($r^2 \geq 0.8$) proxy within 200 kb on the array, compared with 34% for CEU, 28% for JPT and CHB, and 17% for YRI. On Kosrae, even more SNPs were captured when allowing long-range (2-Mb) correlations, whereas in other populations, we did not observe this additional advantage (**Fig. 4**). This suggests that the full benefits for genetic studies in this isolated population may be maximized by using approaches that incorporate long-range LD.

Haplotype testing can have more power in association studies than testing individual SNPs alone[18]. We repeated the hidden SNP analysis described above, this time implementing a haplotype-based analysis (see Methods) instead of simply testing genotypes of individual SNPs. Using this haplotype approach resulted in 78% of SNPs having $r^2 > 0.8$ in Kosrae versus 49% of SNPs having $r^2 > 0.8$ in the CEU population (**Fig. 4**).

Finally, we compared the Kosraen haplotypes with those in HapMap populations. Data was phased, and five nonoverlapping SNP haplotypes spanning 100 kb were compared across populations. Critically, over 98% of the haplotypes present in Kosrae were represented in the CEU, JPT and CHB populations. Ninety-one percent of haplotypes were shared in both the European and Asian populations. An additional 5% of haplotypes were observed only in the Asian populations, and another 3% of haplotypes were present in the European population only. These data indicate that the HapMap dataset contains nearly all haplotypes observed in Kosrae. Thus, the HapMap data is highly informative in this Oceanic population and should permit us to perform haplotype-based tests that increase the power to detect potentially causal SNPs that have not been genotyped.

The detection of some haplotypes in Kosrae that are observed only in the European HapMap population suggests the possibility of some European admixture on Kosrae. Thus in rare instances (3%), we observed Caucasian haplotypes that extend over several megabases (data not shown), so we considered the possibility that admixture might be falsely elevating our estimates of LD. We used a modification of the Hidden Markov Model[19] to demarcate chromosomal segments that are highly likely to be of recent European origin. The same degree of long-range LD on Kosrae was observed when these segments were deleted from the analysis and measures of LD were recalculated (data not shown).

In summary, this study reports the first high-density, genome-wide SNP haplotype map for an isolated population. We estimate that with current technology, we can cover 78% of the SNPs in the genome with high ($r^2 \geq 0.8$) efficiency. Moreover, we show that resources developed from the HapMap project are useful for the Kosraen population. Over the last decade we (J.M.F., M.S., J.L.B.) and collaborators in Kosrae have ascertained 3,150 subjects on Kosrae (total population is estimated at 8,000) and compiled a pedigree for the entire island[14]. A total of 4,854 sibling pairs in 750 nuclear families with 885 trios are available for analysis of clinical data for numerous traits relevant for metabolic syndrome, including BMI, glucose tolerance tests, blood pressure and plasma lipid levels. Genotyping the entire Kosraen cohort of 3,150 samples using the Affymetrix GeneChipTM100K Mapping Arrays is underway and will allow us to use a variety of analytical strategies that are capable of detecting association to disease variants in a population with the genetic structure present on Kosrae[20–26].

METHODS

Human subjects. Genotyping was carried out on 30 trios from the island of Kosrae, Federated States of Micronesia. The parents in these trios are all unrelated to each other by at least five generations. Kosraean samples were collected with approval from The Rockefeller University Institutional Review Board (IRB), and informed consent was obtained for all Kosraen participants in the study. DNA was extracted from blood. Samples typed by the International HapMap Project include 30 CEPH trios (CEU) from Utah, USA with ancestry from northern and western Europe; 30 Yoruban (YRI) trios from Ibadan, Nigeria; 44 unrelated Japanese from Tokyo, Japan (JPT) and 45 unrelated Han Chinese from Beijing, China (CHB).

Genomic DNA samples that were genotyped for the independent haplotype diversity study included 14 Kosraen trios, 14 CEPH trios, 14 Han Chinese trios and 14 African trios. CEPH samples were obtained from CORIELL. African samples were collected from unrelated civil servants in Benin City, Nigeria. Sample collection was organized by Robert Ferrell at the University of Pittsburgh and was approved by the University of Pittsburgh IRB. Han Chinese samples were from the University of Southern California (USC), with IRB approval for studies of medical genetics by the USC IRB.

Genotyping. Genotyping of the ~100,000 SNP collection was carried out using the Affymetrix GeneChip Mapping 100K Arrays. The average call rate per array was ~98%, and an assay was repeated if the call rate was <94%. To be included in data analyses, each individual SNP was required to meet the following quality metrics: missing data across samples <25%, Mendel error ≤1 and Hardy-Weinberg Equilibrium p value cutoff = 0.001.

Regions selected for the independent validation study were previously described in ref. 16. SNPs genotyped included all polymorphic sites discovered by targeted resequencing of these regions in 141 multiethnic samples (60 African, 54 European, 14 East Asian, five nonhuman primates, four American, two Pacific, two other Asian).

Statistical methods. Evaluation of LD measures was performed using Haploview[27] and special-purpose computer code. Haplotypes were inferred by Mendel inheritance combined with an EM-type phasing algorithm. Haplotype diversity was measured using two independent data sets. We used the Affymetrix GeneChip 100K data to study longer-range (1 Mb) haplotypes and a separate dataset with higher intermarker density to study shorter-range haplotypes (150 kb). The average number of markers in 1 Mb on the Affymetrix chips is 35, yielding an intermarker distance of 1 SNP every 29 kb. There were an average of 28 SNPs spanning each of the 17 150-kb regions studied. The more comprehensive coverage of the 100K dataset as well as the lower intermarker density made it more suitable for assessing haplotypes over longer intervals such as the 1-Mb span that we used. The higher density dataset provided a means of testing haplotype diversity over shorter distances. Haplotype diversity was inferred by comparing the average number of haplotypes across these regions in each population. We assessed the presence of Kosraen haplotypes in HapMap populations by comparing the identity of haplotypes composed of five SNPs spanning nonoverlapping 100-kb regions.

Efficiency of association studies was estimated using a 'hidden SNP' approach. In this strategy, a single genotyped SNP is 'hidden' and the maximum correlation coefficient (r^2) of the 'hidden' SNP to any tested variant within 200 kb is determined. This process was repeated using a 2-Mb distance cutoff. This was done for each SNP in the ~110,000-SNP dataset. The haplotype-based analysis considered in turn all combinations of up to three markers within 2 Mb of the hidden SNP. The occurrence of each of the combinations (up to eight) of alleles of these markers was treated as a potential proxy for the hidden SNP, and the correlation between the potential proxy and the hidden SNP was evaluated. The most correlated proxy was registered. Owing to computational constraints, results for haplotype testing are reported for the 3,906 SNPs on chromosome 14.

ACKNOWLEDGMENTS
We thank M. Sullivan for technical assistance and R. Ferrell for the Beni samples. We thank the Government and Department of Health of Kosrae for their partnership and the people of Kosrae for making this study possible. The authors thank The Starr Foundation for their support.

COMPETING INTERESTS STATEMENT
The authors declare that they have no competing financial interests.

Published online at http://www.nature.com/naturegenetics
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. de la Chapelle, A. & Wright, F.A. Linkage disequilibrium mapping in isolated populations: the example of Finland revisited. *Proc. Natl. Acad. Sci. USA* **95**, 12416–12423 (1998).
2. Shifman, S. & Darvasi, A. The value of isolated populations. *Nat. Genet.* **28**, 309–310 (2001).
3. Wright, A.F., Carothers, A.D. & Pirastu, M. Population choice in mapping genes for complex diseases. *Nat. Genet.* **23**, 397–404 (1999).
4. the International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–796 (2003).
5. The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
6. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517 (1996).
7. Jorde, L.B. Linkage disequilibrium as a gene-mapping tool. *Am. J. Hum. Genet.* **56**, 11–14 (1995).
8. Lee, N. *et al.* A genomewide linkage-disequilibrium scan localizes the Saguenay-Lac-Saint-Jean cytochrome oxidase deficiency to 2p16. *Am. J. Hum. Genet.* **68**, 397–409 (2001).
9. Eaves, I.A. *et al.* The genetically isolated populations of Finland and sardinia may not be a panacea for linkage disequilibrium mapping of common disease genes. *Nat. Genet.* **25**, 320–323 (2000).
10. Shifman, S., Kuypers, J., Kokoris, M., Yakir, B. & Darvasi, A. Linkage disequilibrium patterns of the human genome across populations. *Hum. Mol. Genet.* **12**, 771–776 (2003).
11. Kruglyak, L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat. Genet.* **22**, 139–144 (1999).
12. Laan, M. & Paabo, S. Demographic history and linkage disequilibrium in human populations. *Nat. Genet.* **17**, 435–438 (1997).
13. Angius, A. *et al.* Not all isolates are equal: linkage disequilibrium analysis on Xq13.3 reveals different patterns in Sardinian sub-populations. *Hum. Genet.* **111**, 9–15 (2002).
14. Shmulewitz, D. *et al.* Linkage analysis of quantitative traits related to obesity, type II diabetes, hypertentsion and dyslipidemia (metabolic syndrome) on the island of Kosrae, Federated States of Micronesia. *Proc. Natl. Acad. Sci. USA* (accepted).
15. Segal, H.G. *Kosrae: The Sleeping Lady Awakens* (Kosrae State Tourist Division, Department of Conservation and Development, Kosrae State Government, Federated States of Micronesia, 1995).
16. Reich, D.E. *et al.* Linkage disequilibrium in the human genome. *Nature* **411**, 199–204 (2001).
17. Carlson, C.S., Eberle, M.A., Kruglyak, L. & Nickerson, D.A. Mapping complex disease loci in whole-genome association studies. *Nature* **429**, 446–452 (2004).
18. Lin, S., Chakravarti, A. & Cutler, D.J. Exhaustive allelic transmission disequilibrium tests as a new approach to genome-wide association studies. *Nat. Genet.* **36**, 1181–1188 (2004).
19. Patterson, N. *et al.* Methods for high-density admixture mapping of disease genes. *Am. J. Hum. Genet.* **74**, 979–1000 (2004).
20. Spielman, R.S., McGinnis, R.E. & Ewens, W.J. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am. J. Hum. Genet.* **52**, 506–516 (1993).
21. Churchill, G.A. & Doerge, R.W. Empirical threshold values for quantitative trait mapping. *Genetics* **138**, 963–971 (1994).
22. Doerge, R.W. & Churchill, G.A. Permutation tests for multiple loci affecting a quantitative character. *Genetics* **142**, 285–294 (1996).
23. Van Steen, K. *et al.* Genomic screening and replication using the same data set in family-based association testing. *Nat. Genet.* **37**, 683–691 (2005).
24. Wilk, J.B. *et al.* Family-based association tests for qualitative and quantitative traits using single-nucleotide polymorphism and microsatellite data. *Genet. Epidemiol.* **21** (Suppl. 1), S364–S369 (2001).
25. Zhang, J., Schneider, D., Ober, C. & McPeek, M.S. Multilocus linkage disequilibrium mapping by the decay of haplotype sharing with samples of related individuals. *Genet. Epidemiol.* **29**, 128–140 (2005).
26. Ober, C., Abney, M. & McPeek, M.S. The genetic dissection of complex traits in a founder population. *Am. J. Hum. Genet.* **69**, 1068–1079 (2001).
27. Barrett, J.C., Fry, B., Maller, J. & Daly, M.J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263–265 (2005).