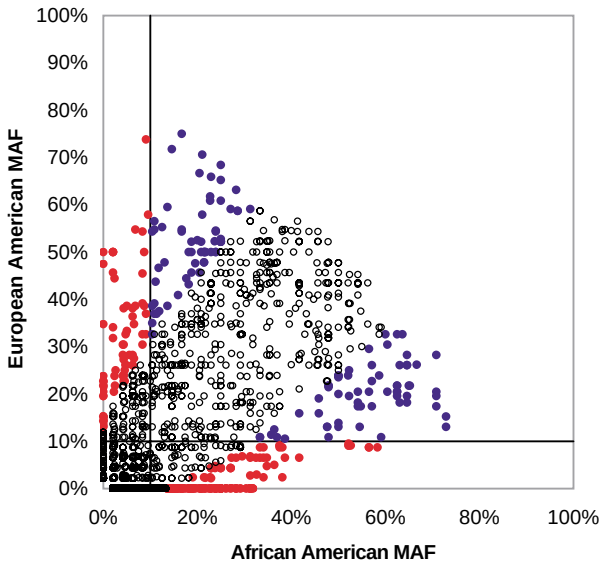
Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans

Christopher S. Carlson[1], Michael A. Eberle[2], Mark J. Rieder[1], Joshua D. Smith[1], Leonid Kruglyak[2,3] & Deborah A. Nickerson[1]

Published online 24 March 2003; doi:10.1038/ng1128

More than 5 million single-nucleotide polymorphisms (SNPs) with minor-allele frequency greater than 10% are expected to exist in the human genome[1]. Some of these SNPs may be associated with risk of developing common diseases[2–4]. To assess the power of currently available SNPs to detect such associations, we resequenced 50 genes in two ethnic samples and measured patterns of linkage disequilibrium between the subset of SNPs reported in dbSNP and the complete set of common SNPs. Our results suggest that using all 2.7 million SNPs currently in the database would detect nearly 80% of all common SNPs in European populations but only 50% of those common in the African American population and that efficient selection of a minimal subset of SNPs for use in association studies requires measurement of allele frequency and linkage disequilibrium relationships for all SNPs in dbSNP.

Testing whether common SNPs are associated with modestly higher risk of developing common diseases is an important challenge in human genetics. It has been suggested that a map of over 300,000 SNPs will be required for such genome-wide association studies[5,6], but it is not yet clear whether the currently available public set of 2.7 million uniquely mapped SNPs is adequate for assembling such a map.



We have determined the patterns of common variation in a set of candidate genes related to the inflammatory process by comprehensively resequencing the complete genomic region of each gene in 47 human samples. We selected and sequenced 50 genes distributed across 17 autosomes and spanning 564 kb. We analyzed samples from two ethnic populations, African Americans (24 individuals) and European Americans (23 individuals). Defining a SNP as a biallelic variant, we identified 2,729 SNPs in the 50 genes; defining a common SNP as one with minor-allele frequency greater than 10% in one or both populations, 1,081 of 2,729 SNPs were common (888 in African Americans and 761 in European Americans). The observed frequency of common SNPs (one per 506 bp scanned) suggests that roughly 6 million common SNPs exist in the genome, consistent with previous estimates[1].

We note that only 52% of common SNPs were common in both populations (561 of 1,081; Fig. 1). Defining private polymorphisms as those observed in only one population, 22% of common SNPs in African Americans were private (199 of 888), as were 5% of common SNPs in European Americans (40 of 761). Furthermore, 36% of common SNPs had significantly different frequencies between populations (384 of 1,081 at $P < 0.01$). Of these, 127 were common in both populations, 185 were common in African Americans but not European Americans and 72 were common in European Americans but not African Americans. Thus, an appreciable fraction of all common variation is either private or common in only a single population, and therefore SNP discovery in a single population is probably inadequate for assembling a catalog of common SNPs that could be used for association studies in all human populations.

We used our data set to estimate the power of the variants previously reported in dbSNP to detect all existing high frequency variants. At least two SNPs were reported in dbSNP for each gene (denoted throughout as dbSNPs), for a total of 837 dbSNPs. The

Fig. 1 Allele frequency comparison between African American and European American populations. The minor allele was set as the less frequent allele in the combined population, and minor-allele frequency (MAF) was calculated in each population for all 2,729 SNPs analyzed. A linear regression of European American minor-allele frequency on African American minor-allele frequency had R^2 of only 0.37, illustrating the marked differences in minor-allele frequency between populations at many sites. Sites where the population allele frequencies were not significantly different ($\chi^2 < 6.635$, $P > 0.01$) are shown as open circles. Sites with significant allele frequency differences ($\chi^2 \geq 6.635$, $P < 0.01$) that were common in both populations are shown in blue (127 sites), and sites with significant allele frequency differences ($\chi^2 \geq 6.635$, $P < 0.01$) that were only common in one population are shown in red (185 SNPs in African Americans, 72 SNPs in European Americans).

[1]*Department of Genome Sciences, University of Washington, 1705 NE Pacific, Seattle, Washington 98195-7730, USA.* [2]*Division of Human Biology, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue N, Seattle, Washington 98109.* [3]*Howard Hughes Medical Institute, Chevy Chase, Maryland, USA. Correspondence should be addressed to C.S.C. (e-mail: csc47@u.washington.edu) or D.A.N. (e-mail: debnick@u.washington.edu).*

Table 1 • Mapping of SNPs in 50 candidate genes using r^2

	African American (888[a])			European American (761[b])		
r^2 threshold[c]	Unmapped[d]	Mapped correctly[e]	Mapped incorrectly[f]	Unmapped[d]	Mapped correctly[e]	Mapped incorrectly[f]
≥0.1	0 (0%)	777 (87.5%)	111 (12.5%)	0 (0%)	729 (95.8%)	32 (4.2%)
≥0.2	0 (0%)	777 (87.5%)	111 (12.5%)	1 (0.1%)	729 (95.9%)	31 (4.1%)
≥0.3	6 (0.7%)	776 (88%)	106 (12%)	5 (0.7%)	729 (96.4%)	27 (3.6%)
≥0.4	31 (3.5%)	767 (89.5%)	90 (10.5%)	18 (2.4%)	722 (97.2%)	21 (2.8%)
≥0.5	106 (11.9%)	733 (93.7%)	49 (6.3%)	37 (4.9%)	714 (98.6%)	10 (1.4%)
≥0.6	178 (20%)	689 (97%)	21 (3%)	59 (7.8%)	700 (99.7%)	2 (0.3%)
≥0.7	250 (28.2%)	628 (98.4%)	10 (1.6%)	79 (10.4%)	681 (99.9%)	1 (0.1%)
≥0.8	310 (34.9%)	575 (99.5%)	3 (0.5%)	107 (14.1%)	653 (99.8%)	1 (0.2%)
≥0.9	405 (45.6%)	483 (100%)	0 (0%)	158 (20.8%)	603 (100%)	0 (0%)
≥1	447 (50.3%)	441 (100%)	0 (0%)	192 (25.2%)	569 (100%)	0 (0%)

[a]Total number of common SNPs in the African American population. [b]Total number of common SNPs in the European American population. [c]Threshold r^2 for pairwise comparisons. [d]Total number of common SNPs that do not exceed threshold r^2 with any other SNP in the data set, within or across loci (percentage of all common SNPs that are unmapped). [e]Total number of common SNPs for which the maximum r^2 within locus exceeds the maximum r^2 across loci (percentage of all mapped that are correctly mapped). [f]Total number of common SNPs for which the maximum r^2 within locus is less than or equal to the maximum r^2 across loci (percentage of all mapped that are incorrectly mapped). [e,f]These numbers include only SNPs for which r^2 exceeds threshold with another SNP.

average dbSNP density in these genes (1 per 654 bp) was considerably higher than in the genome overall (roughly 1 dbSNP per 1,100 bp), reflecting the fact that these genes are all candidate genes for inflammatory disease processes and therefore have been the targets of multiple directed SNP-discovery efforts[7–9].

We found that fewer dbSNPs were polymorphic in our samples than in previous reports[6,10,11]. Only 496 of the 837 previously reported SNPs were polymorphic, with 413 of these common in either African Americans or European Americans (see Supplementary Table 1 online). We confirmed variants that were independently reported by multiple groups at a considerably higher frequency (183 of 214, 85.5%) than SNPs that were uniquely reported by a single group (313 of 623, 50.2%) and observed considerable variation in confirmation rates among submitting groups (see Supplementary Table 1 online). We confirmed approximately equal numbers of SNPs in each population (438 in African Americans, 431 in European Americans). Given that the African American population has higher nucleotide diversity and considerable European admixture, this suggests a bias toward the European population as the source of dbSNPs. Extrapolating from 413 common SNPs in 837 dbSNPs, we estimate that roughly 50% of the SNPs in dbSNP are common (1.35 million of 2.7 million), representing 20–25% of the estimated 6 million common SNPs in the genome, although other ethnicities may have substantial numbers of common, population-specific SNPs.

In an association study, risk variants can be detected either by direct assay or by indirect assay of an associated marker in linkage disequilibrium (LD) with the risk variant[3]. Assessing the power of a collection of SNPs to detect risk variants indirectly requires specification of the strength of LD between each unassayed marker and the set of assayed markers. We chose the LD statistic r^2 for this analysis, because power to detect a risk variant indirectly in n samples is equivalent to power to detect it directly in nr^2 samples[12]. We calculated the observed r^2 for all pairs of common SNPs in each population and determined the fraction of all common SNPs ascertained across a range of r^2 values using several subsets of dbSNP (Fig. 2). Each common SNP was categorized as ascertained if it either belonged to the subset (directly assayed) or exceeded a threshold level of observed r^2 with a SNP from the same gene that was in the subset (indirectly assayed). Although low r^2 thresholds allow assay of fewer variants, they also require much larger samples to retain power. We applied a stringent threshold correlation between assayed and unassayed variants ($r^2 > 0.8$) because we observed few false positive associations (<1%) in the data set at this threshold (Table 1). If all 2.7 million dbSNPs were developed into assays, at $r^2 > 0.8$ roughly 50% of all common SNPs in the genome would be ascertained in African Americans and 77% in European Americans (Fig. 2). Results are similar for other r^2 thresholds between 0.5 and 0.9.

To better approximate dbSNPs in anonymous regions, we also examined the subset of dbSNP described using random discovery

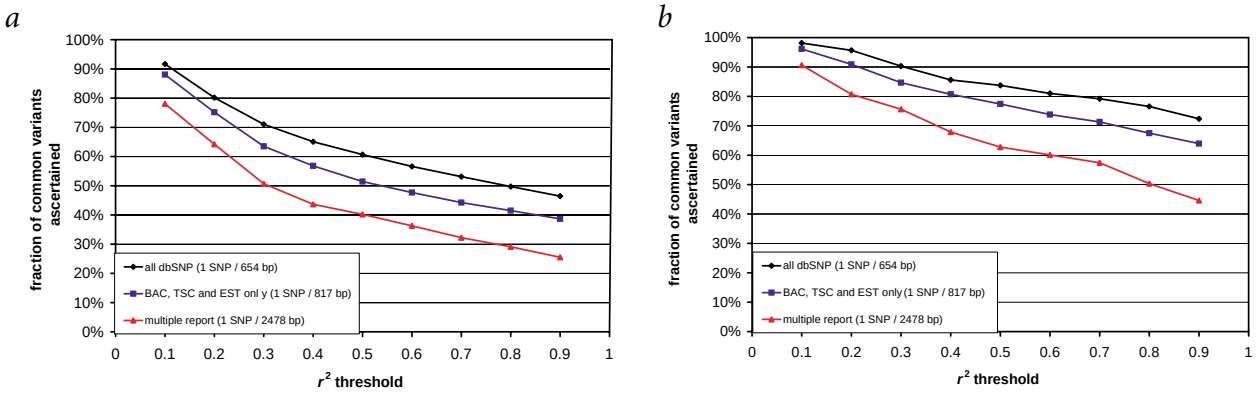


Fig. 2 Detection of common SNPs by linkage disequilibrium using subsets of dbSNP. Detection of common SNPs (minor-allele frequency >10%) in African American (***a***) and European American (***b***) samples was plotted against threshold r^2 value. Detection was defined for each common SNP as either direct assay or r^2 above threshold with an assayed SNP. Results are shown for detection of all identified common SNPs using several subsets of dbSNP: all of dbSNP, only dbSNPs identified by random discovery strategies (BAC, TSC or EST) and dbSNPs independently reported by multiple submitters.

techniques (BAC, RRS or EST) and found that it was only modestly less powerful than all of dbSNP. Two factors may bias our ascertainment estimates. First, some unascertained SNPs in our reference sequence may be in LD with dbSNPs in flanking regions that we did not sequence, which would lead to an underestimate of ascertainment. Second, the higher dbSNP density in our regions relative to the genome as a whole would lead to an overestimate of ascertainment. Simulations showed that the magnitude of bias from each factor was similar (data not shown), so that the two factors offset, leaving little overall bias in our estimates of ascertainment.

Our analysis suggests that most but not all of the SNPs required to assemble a comprehensive map useful for the European American population have already been discovered but that considerable additional SNP discovery is needed to assemble a map useful for the African American population. Similar studies in other populations will be required before conclusions can be drawn as to the adequacy of dbSNP for each population.

Given that the set of all dbSNPs can directly or indirectly assay nearly 80% of all common SNPs in the European American population, how can we select a maximally informative subset of dbSNP without designing assays for all 2.7 million unique dbSNPs? If allele frequency information were available for all dbSNPs in each population of interest, it would be straightforward to design assays only for common variants. In our data set, this would reduce the number of assay designs from 837 to 413, translating to roughly 1.35 million assays genome-wide. Further efficiency could be achieved by eliminating redundant markers, which requires determination of pairwise LD values for all dbSNPs in each population. We optimized the list of common dbSNPs by retaining only one variant from each pair with $r^2 > 0.8$, which yielded 258 non-redundant assay designs in European Americans and 341 in African Americans, translating to a set of 800,000 to 1.1 million SNPs for the entire genome. At an ascertainment threshold of $r^2 > 0.8$, the set of common, non-redundant dbSNPs allowed essentially the same ascertainment of all common SNPs as did the complete set of all dbSNPs.

Further reductions in the final map density may be possible if some SNPs are strongly associated with haplotypes, by defining haplotypes across each gene, subdividing the genes into haplotype 'blocks' showing little evidence for recombination and asking for the fraction of haplotypes 'tagged' by various subsets of dbSNP[6,13]. Such an analysis, however, involves a number of inferences and assumptions that complicate its interpretation, including computational inference of haplotypes from genotype data, the definition of haplotype blocks and the choice of a measure for fraction of haplotypes captured. We focused our analysis on pairwise LD because interpretation of the results is straightforward.

Unfortunately, the LD data or allele frequency data necessary to identify a minimal set of dbSNP with reasonable power to detect associations are currently not available for most dbSNPs[14]. As a possible alternative, we examined ascertainment using only dbSNPs reported by multiple groups, as these are much more likely to be common (see Supplementary Table 1 online). This strategy yielded a set of 214 variants, 162 of which were common, but ascertainment with the multiply reported set was markedly lower than with the complete set (50% versus 77% for European Americans and 29% versus 50% in African Americans at $r^2 > 0.8$; Fig. 2) because some of the multiply reported dbSNPs are rare and some are strongly associated with one another. We examined other subsets of dbSNP, but none had better ascertainment than multiply reported SNPs with the same number of assays developed. Even the insufficiently powerful multiply reported subset would require development of 700,000 assays across the genome, and a randomly selected subset of dbSNPs would require many more markers to achieve reasonable power.

It is not surprising that additional SNP discovery is required for association studies in the African American population. Analyses similar to those presented are necessary to determine how much common variation is private or population-specific in other large ethnic populations. In combination with further SNP discovery in high-diversity populations, such as those of recent African descent, such studies will help ensure that a linkage disequilibrium map is adequately powerful in all ethnic populations, particularly those in which a substantial fraction of common variation is population-specific. Even when sufficient SNPs have been discovered, however, there is no simple way to develop an optimal subset without knowledge of SNP allele frequencies and the patterns of LD between SNPs in each population.

Methods

Samples. We analyzed 24 African Americans from the Coriell HD50AA panel (NA17101–NA17116, NA17133–NA17140) and 23 individuals of European descent from the CEPH families (NA06990, NA07019, NA07348, NA07349, NA10830, NA10831, NA10842–NA10845, NA10848, NA10850–NA10854, NA10857, NA10858, NA10860, NA10861, NA12547, NA12548 and NA12560).

Sequencing. The SeattleSNPs Program for Genomic Applications resequences candidate genes involved in inflammatory processes in humans. For all genes analyzed, we resequenced the complete genomic region of the transcript, including introns, 2.5 kb 5′ of the gene and 1.5 kb 3′ of the gene, using standard dye primer chemistry on an ABI 3700. For each gene, sequence analysts assembled the sequence data into a contig using Phred and Phrap, edited the contig in Consed to ensure that the assembly was accurate and identified polymorphisms using the PolyPhred program, version 4.0. At insertion–deletion polymorphisms, the sequence analysts manually genotyped each sample and designed primers from the other strand to sequence beyond the insertion–deletion. Analysts reviewed every polymorphic site flagged by PolyPhred to remove a few false positives associated with biochemical artifacts, such as GC compressions, unincorporated dye terminators and heterozygous insertion–deletion polymorphisms.

We assessed data quality in a number of ways. We trimmed each chromatogram to remove low-quality sequence (Phred score below 25), resulting in analyzed reads averaging >450 bp with an average quality of Phred 40. We obtained second-strand confirmation from a different sequencing primer at 66% of all polymorphic sites and third strand confirmation at 33% of all polymorphic sites. We observed all three possible genotypes (heterozygotes and homozygotes with respect to each allele) for approximately 38% of common polymorphic sites with an average Phred quality greater than 45 (1:50,000 probability of being incorrectly assigned). The average flanking-sequence quality associated with polymorphic sites (± 5 bp on each side of the polymorphic site) was greater than 40. Eighty percent of all common sites were significantly associated with at least one other site in the same gene ($\chi^2 > 10.828$, $P \leq 0.02$ corrected for multiple tests in each gene). We independently genotyped 59 of the identified common sites by Taqman allelic discrimination on an ABI 7900 (ref. 15) and observed only 8 discrepancies in 2,773 genotypes compared between technology platforms, suggesting an error rate well below 1% for genotype calls.

Loci analyzed. We have resequenced over 90 genes to date, and details of all SNPs identified have been submitted to dbSNP. This analysis was limited to autosomal genes with complete resequencing data and assigned refSNP numbers. We identified 50 genes meeting these criteria, spanning a total of 565 kb, or an average of 11 kb per gene. GenBank accession numbers for the reference sequence of each gene position in the corresponding genomic contig are shown in Supplementary Table 2 online. We scanned 547 kb (96.8%) of this set; the remainder fell in regions that were difficult to amplify or yielded low-quality sequence data. Thus, the data set comprised more than half a megabase of genomic sequence with nearly complete SNP ascertainment in two ethnic groups across more than 46 chromosomes for each group. We identified 2,729 biallelic polymorphisms: 2,577 single-nucleotide substitutions and 152 biallelic insertion–deletion variants. We also identified multiallelic markers, but these were not included in the analysis. Only 4.4% of all genotypes could not be determined.

dbSNP comparisons. To make comparisons with the dbSNP database (build 104), we identified the reference sequence of the region scanned for each gene using BLAST (see Supplementary Table 2 online) and retrieved refSNP numbers for all variations mapped to the reference sequence by the National Center for Biotechnology Information. Although we submitted 2,729 common sites to dbSNP, we retrieved only 2,486 using this method, evidently reflecting the difficulty of mapping SNPs uniquely to the genome based on 100 bp of flanking sequence. For each variation reported in the reference sequence, we established which submitter(s) reported the SNP. We manually inspected sequence traces at all sites not initially confirmed in the data set and categorized as unconfirmed those dbSNPs that had valid sequence coverage but were not observed to be polymorphic in our populations. We grouped submitters according to their discovery strategy: BAC-overlap discovery (BAC) submitter KWOK[16] or SC_JCM[17], random-clone overlap or reduced representation sequencing (TSC) submitter TSC-CSHL[10,18], EST overlap (EST) submitters LEE[19] and CGAP-GAI[20], pooled PCR discovery (PCR) submitter YUSUKE and all other submitters. Confirmation rates by SNP discovery strategy are given in Supplementary Table 1 online.

Our confirmation rate for TSC-reported SNPs (64.8%) was markedly lower than that found in a previous report[6]. To determine whether this might reflect SNPs specific to other ethnicities, we analyzed TSC confirmation in 50 genes sequenced in the Environmental Genome Project using 90 individuals (24 European Americans, 24 African Americans, 24 Asian Americans, 12 Hispanic Americans and 6 Native Americans) from the polymorphism-discovery resource[21]. Although the confirmation rate from the Environmental Genome Project is higher than that from the Program for Genomic Applications (123 of 171 TSC-reported dbSNPs; 71.9%), it is still below that from the previous report, suggesting that SNPs specific to Asian and Hispanic populations do not entirely explain the low TSC SNP confirmation rates.

Linkage disequilibrium. Given two biallelic sites with minor-allele frequencies p_{1+} and p_{+1}, the major-allele frequencies are p_{2+} $(= 1 - p_{1+})$ and p_{+2} $(= 1 - p_{+1})$, and there are four possible haplotypes with frequencies p_{11}, p_{12}, p_{21} and p_{22}. We estimated haplotype frequencies for every pair of SNPs in each gene from the observed genotype frequencies according to the method of Hill[22]. We inferred r^2 from the estimated two-site haplotype frequencies[23] using the equation $r^2 = (p_{11}p_{22} - p_{12}p_{21})^2 / (p_{1+}p_{2+}p_{+1}p_{+2})$. Simulations show that bias in r^2 is relatively small in samples of 23 individuals (see Supplementary Table 3 online). Simulations under a standard neutral model[24] suggest that in this sample size roughly 80% of all site pairs with an observed r^2 above a given threshold represent true r^2 above threshold (see Supplementary Fig. 1 and Supplementary Table 4 online).

SNP ascertainment. Using various subsets of dbSNP, we calculated the fraction of all common variants ascertained as the fraction of all common variants previously reported (directly assayed) plus the fraction of all common variants indirectly ascertained by association at r^2 greater than threshold with a previously reported SNP (indirect assay). Using all of dbSNP, 336 of 888 common sites in the African American population were already in dbSNP (38%) and 105 other common sites were associated with these sites at $r^2 > 0.8$, for a total of 441 sites ascertained at this threshold. Similarly, 359 of 761 common sites in the European American population were already in dbSNP (47%) and 226 common sites not previously in dbSNP were associated with these sites at $r^2 > 0.8$, for a total of 585 sites ascertained. In the absence of reported SNPs in a gene, unreported SNPs would not be ascertained. Thus, for each subset of dbSNP, we considered only genes with at least one dbSNP in the subset (49 genes using only BAC, TSC or EST SNPs and 48 genes using multiply reported SNPs) and adjusted the potential number of dbSNPs ascertained accordingly (see Supplementary Table 1 online).

URLs. dbSNP, http://www.ncbi.nlm.nih.gov/SNP/index.html; Phred, Phrap and Consed, http://www.phrap.org; Polyphred version 4.0, http://droog.mbt.washington.edu/PolyPhred.html; GenBank, http://www.ncbi.nlm.nih.gov/Genbank/index.html. Genotype files for all data reported are available at http://pga.gs.washington.edu. Data for the EGP project are available at http://egp.gs.washington.edu.

Note: Supplementary information is available on the Nature Genetics website.

Acknowledgments

The authors would like to thank Q. Yi, T. Armel, E. Calhoun, D. Carrington, M. Chung, P. Keyes, P. Lee, C. Poel and E. Toth for producing sequence variation data for the SeattleSNPs Program for Genomic Applications and M. Lundberg and S. Banks-Schlegel for their advice and encouragement. This work was supported by a Program for Genomic Applications grant from the National Heart Lung and Blood Institute (to D.A.N., M.J.R. and L.K.) with additional support from the National Institute of Mental Health (to L.K.). L.K. is a James S. McDonnell Centennial Fellow.

Competing interests statement

The authors declare competing financial interests. Details accompany the paper on the Nature Genetics website (http://www.nature.com/naturegenetics).

Received 18 November 2002; accepted 20 February 2003.

1. Kruglyak, L. & Nickerson, D.A. Variation is the spice of life. *Nat. Genet.* **27**, 234–236 (2001).
2. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517 (1996).
3. Collins, F.S., Guyer, M.S. & Chakravarti, A. Variations on a theme: cataloging human DNA sequence variation. *Science* **278**, 1580–1581 (1997).
4. Lander, E.S. The new genomics: global views of biology. *Science* **274**, 536–539 (1996).
5. Kruglyak, L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat. Genet.* **22**, 139–144 (1999).
6. Gabriel, S.B. *et al.* The structure of haplotype blocks in the human genome. *Science* **296**, 2225–2229 (2002).
7. Cargill, M. *et al.* Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat. Genet.* **22**, 231–238 (1999).
8. Cambien, F. *et al.* Sequence diversity in 36 candidate genes for cardiovascular disorders. *Am. J. Hum. Genet.* **65**, 183–191 (1999).
9. Halushka, M.K. *et al.* Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nat. Genet.* **22**, 239–247 (1999).
10. Sachidanandam, R. *et al.* A map of human genome sequence variation containing 1.42 million single-nucleotide polymorphisms *Nature* **409**, 928–933 (2001).
11. Reich, D.E., Gabriel, S.B. & Altshuler, D. Quality and completeness of SNP databases. *Nat. Genet.* **33**; advance online publication 24 March 2003; doi:10.1038/ng1133.
12. Pritchard, J.K. & Przeworski, M. Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.* **69**, 1–14 (2001).
13. Johnson, G.C. *et al.* Haplotype tagging for the identification of common disease genes. *Nat. Genet.* **29**, 233–237 (2001).
14. Marth, G. *et al.* Single-nucleotide polymorphisms in the public domain: how useful are they? *Nat. Genet.* **27**, 371–372 (2001).
15. Livak, K.J. Allelic discrimination using fluorogenic probes and the 5′ nuclease assay. *Genet. Anal.* **14**, 143–149 (1999).
16. Marth, G.T. *et al.* A general approach to single-nucleotide polymorphism discovery. *Nat. Genet.* **23**, 452–456 (1999).
17. Ning, Z., Cox, A.J. & Mullikin, J.C. SSAHA: a fast search method for large DNA databases. *Genome Res.* **11**, 1725–1729 (2001).
18. Altshuler, D. *et al.* An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* **407**, 513–516 (2000).
19. Irizarry, K. *et al.* Genome-wide analysis of single-nucleotide polymorphisms in human expressed sequences. *Nat. Genet.* **26**, 233–236 (2000).
20. Buetow, K.H., Edmonson, M.N. & Cassidy, A.B. Reliable identification of large numbers of candidate SNPs from public EST data. *Nat. Genet.* **21**, 323–325 (1999).
21. Collins, F.S., Brooks, L.D. & Chakravarti, A. A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res.* **8**, 1229–1231 (1998).
22. Hill, W.G. Estimation of linkage disequilibrium in randomly mating populations. *Heredity* **33**, 229–239 (1974).
23. Devlin, B., Risch, N. & Roeder, K. Disequilibrium mapping: composite likelihood for pairwise disequilibrium. *Genomics* **36**, 1–16 (1996).
24. Hudson, R.R. Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics* **18**, 337–338 (2002).