Integrating common and rare genetic variation in diverse human populations

The International HapMap 3 Consortium

Despite great progress in identifying genetic variants that influence human disease, most inherited risk remains unexplained. A more complete understanding requires genome-wide studies that fully examine less common alleles in populations with a wide range of ancestry. To inform the design and interpretation of such studies, we genotyped 1.6 million common single nucleotide polymorphisms (SNPs) in 1,184 reference individuals from 11 global populations, and sequenced ten 100-kilobase regions in 692 of these individuals. This integrated data set of common and rare alleles, called ‘HapMap 3’, includes both SNPs and copy number polymorphisms (CNPs). We characterized population-specific differences among low-frequency variants, measured the improvement in imputation accuracy afforded by the larger reference panel, especially in imputing SNPs with a minor allele frequency of ≤5%, and demonstrated the feasibility of imputing newly discovered CNPs and SNPs. This expanded public resource of genome variants in global populations supports deeper interrogation of genomic variation and its role in human disease, and serves as a step towards a high-resolution map of the landscape of human genetic variation.

The Human Genome Project, the SNP Consortium and the International HapMap Project collectively identified ~10 million common DNA variants, primarily SNPs, in a limited set of DNA samples. Knowledge of these SNPs and their linkage-disequilibrium patterns enabled genome-wide association studies, which have successfully identified hundreds of novel genomic loci that influence human diseases.

Nonetheless, our knowledge of human genetic variation remains limited with respect to variant type, frequency and population diversity. Only common DNA variants (minor allele frequency (MAF) ≥5%) have yet been well studied, even though low MAF variants no doubt contribute to a substantial fraction of hereditary risk for common diseases. Only recently have systematic studies of other types of variants, in particular copy number variation, begun to guide our knowledge of their frequency spectra, population distributions and patterns of linkage disequilibrium.

To inform efforts aimed at rectifying this, we expanded the public HapMap Phase I and II resource by performing genome-wide SNP genotyping and CNP detection, as well as polymerase chain reaction (PCR) resequencing in selected genomic regions. We collected and studied an extended set of 1,184 samples from 11 populations (Supplementary Information). These included all HapMap Phase I and II samples, along with further samples from the same four populations: individuals from the Centre d’Etude du Polymorphisme Humain collected in Utah, USA, with ancestry from northern and western Europe (CEU); Han Chinese in Beijing, China (CHB); Japanese in Tokyo, Japan (JPT); and Yoruba in Ibadan, Nigeria (YRI). Samples from seven additional populations were also included: African ancestry in the southwestern USA (ASW); Chinese in metropolitan Denver, Colorado, USA (CHD); Gujarati Indians in Houston, Texas, USA (GIH); Luhya in Webuye, Kenya (LWK); Maasai in Kinyawa, Kenya (MKA); Mexican ancestry in Los Angeles, California, USA (MXL); and Tuscanans in Italy (Toscani in Italia, TSI). These populations were included to provide further variation data from each of the three continental regions represented in HapMap Phase I and II, as well as data from some more admixed populations residing in the US. The specific populations and localities were chosen based on contacts with researchers who worked in those regions and had established trusting relationships with local communities. (See Supplementary Table 1 and the Supplementary Information for more details.)

SNP genotyping

Genotype data were obtained with the Affymetrix Human SNP array 6.0 (interrogating 1,832,600 genomic sites) and the Illumina Human1M single beadchip (1,199,187 genomic sites), initially applied to 1,486 and 1,284 samples, respectively. Following genotype calling and initial filtering of low-quality and incomplete data, 909,622 variant SNPs from 1,326 samples (Affymetrix) and 1,055,111 sites from 1,121 samples (Illumina) remained. Data from the two platforms were merged; genotype concordance was 99.5% (across 335,014 overlapping SNPs) at a call rate of 99.8%. Further filters were applied to this merged data set on the basis of population-specific call rates, deviation from Hardy–Weinberg equilibrium and the expected Mendelian inheritance patterns (Supplementary Methods). The consensus genotype set contains 1,440,616 SNPs that are polymorphic in 1,184 individuals from 11 populations. Analysis shows a small but statistically significant bias against rare (MAF = 0.05–0.5%) allele calls (observed in both platforms), consistent with previous reports (Supplementary Information). The data were then phased (Supplementary Information).

Regional sequencing

We selected ten 100-kb regions for direct PCR-Sanger capillary sequencing analysis. These regions included the central 100 kb from five previously sequenced HapMap-ENCODE 500-kb regions and five ENCODE regions not previously subject to sequencing in the HapMap Project (Supplementary Table 4). A total of 692 unrelated samples chosen from the ten then-available genotyped population samples (ASW, CEU, CHB, CHD, GIH, JPT, LWK, MXL, TSI and YRI) were interrogated and passed quality control metrics (Supplementary Table 1). SNPs were discovered from the raw sequence data using SNP Detector 3.0 software. Subsequent genotyping showed an overall genotype concordance rate of 99.2% and an 86.8% genotype concordance rate for genotypes with minor alleles (Supplementary Table 5a). Also, a 93.6% genotype concordance rate...
was found for singleton genotypes with minor alleles and 88% for two to six copies of the minor allele. The higher genotype concordance rate in singletons reflects the higher stringency applied in making singleton calls. (See Supplementary Information and Supplementary Table 5 for details.) Unlike SNPs present on microarray platforms, which are intentionally biased towards high frequency by the discovery and selection process, the SNPs discovered by sequencing provide a direct estimate of the underlying allele frequency spectrum in each population. As in previous surveys, common (MAF ≥5%) and low-frequency (MAF = 0.5–5%) variants account for the vast majority of the heterozygosity in each sample, but we also observed a large number of rare (MAF = 0.05–0.5%) and private (singletons and MAF <0.05%) variants (see Supplementary Table 2 for definitions of variant frequency classes). Each population had 42–66% of sites with a MAF <5%, compared to 10–13% in the genotyping data; 37% of SNPs with a MAF <0.5% were observed in only one population. In total, 77% of the discovered SNPs were new (that is, not in the SNP database) compared to 10–13% in the genotyping data; 37% of SNPs with a MAF <0.5% were observed in only one population. In total, 77% of the discovered CNPs were new (that is, not in the SNP database (dbSNP) build 129) and 99% of those had a MAF <5%.

Copy number variation

To assess copy number variation we merged and analysed the probe-intensity data from both the Affymetrix and Illumina arrays, identifying 1,610 genomic segments that probably varied in copy number (CNPs) with an estimated MAF of at least 1% of the cohort (see Methods). Further quality control steps yielded a set of reference CNPs for 856 CNPs with a 99.0% mean call rate and 0.3% Mendelian inconsistency—very high accuracy, but still less than that observed from SNP genotyping (Mendelian inconsistency <0.14% in this data set; Supplementary Information). We estimate that the resolution of this analysis to detect CNVs is at a multi-kilobase scale, but not smaller (Fig. 1a).

The overall allele frequency spectrum of CNPs resembled that of the SNPs ascertained by resequencing; most variants were at low frequency (Fig. 1b), but most heterozygosity was due to a limited set of common variants. This extends an observation previously made in the original HapMap population samples1 to additional populations. The allele frequency spectrum of common CNPs (MAF >10%) was similar across populations, but differed markedly at lower frequencies. African-ancestry and admixed populations showed by far the greatest number of variants with MAF <5%, and had a higher average number of CNPs differing in copy number between two individuals (160–171) than non-admixed populations without African ancestry (127–142) (Fig. 1b).

At 95% of the CNPs the variation observed was explained by a simple biallelic model obeying Mendelian inheritance and Hardy–Weinberg equilibrium. The remaining 5% of loci showed multi-allelic patterns, somewhat lower than the 15% reported in a recent study,2 which may reflect improved resolution of the assays and analyses used in this study. Among the biallelic loci, 92% were deletions (diploid copy numbers ≤2) and 8% were duplications (diploid copy numbers ≥2); the disparity reflects our higher power to detect small deletions than small insertions. The median size of CNPs genotyped in this study was 7.2 kb (Fig. 1a), with biallelic deletions significantly smaller on average than biallelic duplications because of this difference in power.

The 856 genotyped CNPs represent an average of 3.5 megabases of sequence in each individual; this is 0.1% of the human genome, and similar to the overall rate of SNP variation. One-third (33.5%) of the genotyped CNPs overlap ReSeq genes, with duplications more likely than deletions to overlap genes (after correcting for the greater average length of duplications (P = 0.006)), which probably reflects greater purifying selection acting on deletions of genes.

Common and low-frequency variation across populations

We used the ENCODE data to assess how well each sample set could serve as a SNP discovery resource for other populations. This is an important practical matter, because it determines the effectiveness of scanning multiple populations for variation discovery as compared to sampling more deeply in a single population. To estimate how informative SNPs discovered in population A were for those present in population B, we counted the fraction of variants found in a sample of 30 A individuals that were also seen in a sample of 30 B individuals. Our measure of informativeness was the ratio of this fraction to that observed for a second, non-overlapping sample of 30 A individuals (Fig. 2a).

As judged by this measure, informativeness varied greatly for different population pairs. Consistent with the observation that non-African diversity is largely a subset of African diversity14, African samples provided a more complete discovery resource for variant sites in non-African samples than the converse (Fig. 2a). Focusing only on low-frequency variants in the original sample of 30 A individuals (one or two copies, corresponding to allele frequencies of 3.3% or less), even African samples were highly incomplete for diversity outside of Africa, with informativeness ratios dropping to 40–60% in LWK and YRI (Fig. 2b). In general, for low-frequency variants only closely related populations did an adequate job of capturing variation (Fig. 2b), probably reflecting the recent origins of low-frequency variants. Two populations, LWK and GIH, stand out as being poorly captured by any of our other populations, the result of admixture with an ancestral population not closely related to any in our regional sequencing data (Supplementary Methods). (Although the MKK captures similar East African ancestry to that of LWK (Supplementary Fig. 2), it had not been included in the regional sequencing.)

In all cases, FST, a measure of the degree of population differentiation (Supplementary Table 6) correctly predicted the most informative population, despite the FST estimates being based on genotyping array data with SNP ascertainment biases15. However, FST was not a perfect predictor: the correlation coefficient between FST and ascertainment informativeness was highly variable, ranging across populations between −0.67 and −0.99 for all SNPs and between −0.51 and −0.97 for low-frequency SNPs. Furthermore, FST is symmetrical between a pair of populations, whereas informativeness is not. For example, the most informative population for low-frequency GIH SNPs was TSI, with informativeness being only 55% of that of an independent GIH sample (because TSI captures only one of GIH’s ancestral populations; Fig. 2b). Conversely, the informativeness of GIH on low-frequency TSI SNPs was 71% (Fig. 2b).

Within a single population, increasing the sequenced sample size yields diminishing returns of new SNPs. Figure 3 quantifies the number of SNPs discovered by resequencing as a function of sample size; it demonstrates the expected partitioning between populations with genetic proximity to Africa, and therefore higher diversity, and the rest of the populations. The new SNPs are mostly of lower frequency, and account for the majority of the discovered variant sites as the number of interrogated samples is increased (Supplementary Fig. 5).
haplotypes for array SNPs are expected because of SNP ascertainment, frequency alleles (whether or not discovered by sequencing) had using the consensus genotype data around each low-frequency SNP. Type homozygosity (that is, perfect concordance between haplotypes) 272 SNPs were examined in YRI and 106 in CEU. For comparison, a ing (Supplementary Methods) to ensure highly accurate genotypes, the ENCODE data a set of SNPs observed two to six times in YRI or in populations with genetic proximity to Africa (LWK, ASW and YRI), compared to populations of non-African ancestry, which was biased towards SNPs shared across populations and therefore towards older SNPs with shorter-range linkage disequilibrium. Among the ENCODE SNPs, there was little difference in haplotype sharing between alleles seen twice and those seen four to six times in the sample, indicating that these minor differences in frequency are not good predictors of the age and haplotype sharing of alleles (presumably due to drift and sampling error in the frequency estimate). Haplotype sharing was also greater for derived than for ancestral alleles, although the effect was modest (Supplementary Fig. 6).

We next characterized the extent to which alleles share haplotype backgrounds as a function of frequency, a question related to the imputation of variants not directly observed in each clinical sample. Population genetic models predict that lower-frequency variants should on average be younger than more common variants, and thus have a longer physical extent of haplotype sharing. We selected from the ENCODE data a set of SNPs observed two to six times in YRI or in CEU; we estimated haplotype phase with high confidence using parent–offspring trio data. After validation using Sequenom genotyping (Supplementary Methods) to ensure highly accurate genotypes, 272 SNPs were examined in YRI and 106 in CEU. For comparison, a set of SNPs from the genotyping arrays with the same frequencies were analysed. Haplotype sharing was measured by calculating the haplotype homozygosity (that is, perfect concordance between haplotypes) using the consensus genotype data around each low-frequency SNP.

In both populations, ENCODE variant alleles had longer shared haplotypes than array-based SNPs of the same frequency, and all low-frequency alleles (whether or not discovered by sequencing) had longer haplotypes than did higher-frequency SNPs (Fig. 4). Shorter haplotypes for array SNPs are expected because of SNP ascertainment,
large collections of samples, genotyping arrays, in concert with statistical imputation of untyped alleles, offer a complementary approach to increase power for previously observed alleles. We therefore evaluated the effect on imputation afforded by the larger HapMap 3 resource and also studied how well imputation performs when applied to lower frequency variants and to CNPs.

One use of imputation is to combine data for genome-wide association studies performed using different array platforms. Therefore, we first measured the change in performance of imputation for common (array-based) SNPs using a HapMap 3 panel of 410 phased European-ancestry chromosomes (CEU+TSI) in comparison with a HapMap Phase II panel of 120 CEU chromosomes (HMII-CEU). Each panel was used to impute array SNPs in 1,393 Europeans of the 1958 British birth cohort (58BBC), which had previously been genotyped using earlier versions of the Affymetrix and Illumina chips. 

We next investigated imputation across populations. We compared imputation of CEU or TSI using the CEU reference panel, CHD or CHB+JPT using the CHB+JPT reference panel, and YRI or LWK using the YRI reference panel. Imputation into closely related populations worked well for common but not for low-frequency alleles (Supplementary Table 11).

Imputation in admixed populations was examined by comparing reference panels based on either one population, or on mixtures of other populations; one mixture (COSMO1) combined chromosomes from the original three HapMap population panels, whereas the other (COSMO2) included seven populations (CEU, CHB, GIH, JPT, MKK, MXL, and YRI, see Methods for details). For ASW, the best reference panel was YRI+CEU, which yielded mean $r^2$ = 0.87 and mean $r^2$ = 0.72 for common and low-frequency SNPs, respectively. For the other admixed populations, the best reference panels were the same-population panel (when available) followed by the diverse reference panel of seven populations (COSMO2) (Supplementary Table 10).

Cross-population imputation can be less effective for low-frequency alleles both because the sets of alleles in the two samples do not overlap perfectly (see earlier), and because haplotype patterns differ between populations. To isolate the effect of differing haplotype patterns, imputation within a population (CEU or YRI) was
1,393 subjects of the 1958 British birth cohort. Scatter plots show Affymetrix 500K SNPs on chromosome 20 imputed for (CEU/TSI copy. Imputation accuracy into a closely related European population the reference panel, and a mean populations was more substantial, with mean panel, with a mean worked well for low-frequency alleles when using the correct reference both target and reference panels (Fig. 6a). Notably, the imputation LWK), but restricting the analysis to SNPs that were polymorphic in compared with imputation into a closely related population (TSI or LWK), but restricting the analysis to SNPs that were polymorphic in both target and reference panels (Fig. 6a). Notably, the imputation worked well for low-frequency alleles when using the correct reference panel, with a mean \( r^2 > 0.7 \) with only two copies of the minor allele in the reference panel, and a mean \( r^2 > 0.6 \) when imputing from a single copy. Imputation accuracy into a closely related European population (CEU/TSI \( F_{ST} = 0.004 \)) was almost indistinguishable from the accuracy within a single population. For the two African populations, where low-frequency diversity is greater and the populations more diverged (\( F_{ST} = 0.008 \)), the difference between reference and target populations was more substantial, with mean \( r^2 \) only rising above 0.7 when five copies of the minor allele were in the reference panel. In both cases, however, the cross-population accuracy was much better than that seen in Supplementary Table 10, indicating that cross-population loss of accuracy largely results from the incomplete sharing of low-frequency alleles between reference and target samples, rather than from differences in haplotype backgrounds. Using the same approach, we also checked the dependence of imputation accuracy on pedigree information, as trios improve the accuracy of haplotype phasing and therefore imputation. We compared the within-CEU results described earlier to imputation done purely within the TSI sample, with the sample size held fixed. The two populations are closely related, but the CEU samples were genotyped in trios and the TSI samples as individuals. The results were virtually identical (data not shown), indicating that poor phasing was not a problem for our unrelated samples, at least for array SNPs. (Note that pedigree information was used indirectly in our TSI phasing, with phased CEU chromosomes used as a reference panel for phasing TSL.) In a second set of analyses, we assessed imputation of newly discovered variants using as our test sets SNPs found in CEU and YRI by the complete ENCODE sequencing and CNPs. We created a reference panel of phased haplotypes that incorporated the new variant and the surrounding consensus genotype data, and used it to impute genotypes in additional samples. This models (for example) the imputation into an existing genome-wide association study of new SNPs and CNPs discovered by the 1000 Genomes Project or an exome sequencing project. We assessed imputation accuracy by masking each individual in the sample in turn and imputing its genotype from the rest of the sample, thus preserving the largest reference panel possible. For comparison, we also repeated the analysis by masking randomly selected, frequency-matched array SNPs, rather than newly discovered variants. The imputation accuracy was quite similar for the SNPs and the CNPs (Fig. 6b), given their similar haplotype properties. Accuracy depended on high SNP density; reducing the set of tag SNPs from the full HapMap 3 set to the subset found on an earlier generation of array (approximately a threefold reduction in density) reduced \( r^2 \) by roughly a factor of two for low-frequency SNPs (Supplementary Fig. 8). Somewhat unexpectedly, the accuracy was consistently higher for YRI than for CEU for both classes of variant, despite the former’s greater haplotype diversity and the identical panel sizes and SNP frequencies.
One possible explanation is that for less common variants, the relationship between frequency and age has been partly obscured by population bottlenecks in the history of European populations, so that minor allele frequency is less effective as a predictor of allele age than in samples from Africa.

Overall, we observed that imputation works well for the newly discovered SNPs, although not as well as for frequency-matched SNPs on the available genotyping arrays, even though newly discovered SNPs show greater haplotype sharing. This difference may be due to an ascertainment bias in the discovery and choice of SNPs on the arrays—most SNPs in HapMap and on arrays were originally detected by sequencing a few individuals, representing a fraction of haplotypes in the population; these haplotypes are better represented on arrays (which focused on SNPs that served as good proxies) than are newly discovered SNPs. This difference is markedly seen in a comparison of nearby, frequency-matched SNPs from within either the array or ENCODE: looking only at SNPs with two copies of the minor allele, 5% of the time, two frequency-matched ENCODE SNPs are perfect proxies for each other, whereas the fraction is 70–80% for a pair of frequency-matched array SNPs (Supplementary Fig. 9). This highlights the need for caution in extrapolating from low-frequency array SNPs to low-frequency sequencing SNPs.

Natural selection

We searched the larger and more diverse HapMap 3 genotype data for genomic regions showing signals of positive natural selection using a recently published method, the composite of multiple signals (CMS)19. In the three original HapMap populations, CEU, CHB+JPT and YRI, comparing the regions identified in HapMap 3 with published results from HapMap Phase II (Supplementary Methods), we replicated 83% (147 out of 178) of the previous HapMap Phase II candidate regions (Supplementary Fig. 10a–d). Of the 17% of regions that did not replicate, most had lower SNP density in HapMap 3 than in HapMap Phase II; in 20 regions, none of the high-scoring HapMap Phase II SNPs was genotyped in HapMap 3.

Next we sought to identify candidate selection loci in the new HapMap 3 populations TSI, LWK and MKK (that is, all populations except those likely to be recently admixed). First we identified 54 broad candidate regions for selection using long haplotype tests. Applying CMS to these regions, we localized signals to new and intriguing candidates (Supplementary Table 12). In TSI, pigmentation genes were again identified, including KITLG and MLPH1 (Supplementary Fig. 10a, f). We found other signals, like LAMA3, a gene involved in wound healing, and an olfactory receptor cluster. In the Kenyan populations we identified several immune-related genes, such as CD2264, ITGAE13 and DPP7 (Supplementary Fig. 10g–i). A novel signal identified in MKK localized to the gene ANKH; ANKH has a role in bone growth and susceptibility to arthritis, and has previously been identified as being under positive selection in horses12 (Supplementary Fig. 10). The complete set of new candidates (Supplementary Table 12) may suggest hypotheses regarding natural selection in these populations.

Conclusions and implications

With improvements in sequencing technology, low-frequency variation is becoming increasingly accessible. This greater resolution will no doubt expand our ability to identify genes and variants associated with disease and other human traits. This study integrates CNPs and lower-frequency SNPs with common SNPs in a more diverse set of human populations than was previously available. The results underscore the need to characterize population-genetic parameters in each population, and for each stratum of allele frequency, as it is not possible to extrapolate from past experience with common alleles. As expected, lower-frequency variation is less shared across populations, even closely related ones, highlighting the importance of sampling widely to achieve a comprehensive understanding of human variation.

We find that variants discovered through large-scale sequencing have longer haplotypes than more common variants, and that imputation can perform well for both CNPs and low-frequency SNPs. Success was partial (as compared to common variants), and required a number of conditions: large reference panels, dense and accurate genotyping and good phasing. Moreover, some variants were not well imputed, although it is unclear if this is fundamental or due to a need for improved methods of imputation of lower frequency variants.

Informed by preliminary analyses of these data, the 1000 Genomes Project is studying the collection of samples from five populations within each continental region. Our data indicate that a strategy of identifying polymorphic SNPs and CNPs followed by imputation in densely genotyped samples can provide information even for lower-frequency alleles. Necessary components of such a reference panel include accurate genotyping and characterization of the haplotype background for the alleles (which included here the use of pedigree information to inform phasing), and a broad range of reference populations to capture geographically local variants. The ultimate utility of such a strategy (as compared to a more complete approach using exome or whole genome sequencing) will depend on the as yet poorly characterized deleteriousness of allelic variants across populations as compared to non-coding regions, and the relative cost and accuracy of sequencing as compared to genotyping followed by imputation. The development of a robust reference panel will be a necessary step in the evaluation of these different strategies across a wide variety of diseases.

METHODS SUMMARY

Genotyping and genotype data quality control. Genotyping was done using Affymetrix 6.0 and Illumina 1.0 Million SNP mass arrays. Data quality control filters were applied as detailed in the text and Supplementary Information.

CNP analysis. For CNP discovery, we combined the genotype data from the Affymetrix and Illumina arrays and applied two algorithms, QuantiSNP27 and Birdseye2. First, approximately 60,000 CNP calls were made by each algorithm (~50 per sample), generally supported by data from both platforms. Shared genomics segments of common CNPs were identified and refined by an algorithm that used cross-sample correlations between nearby probes (Supplementary Information).

For CNP genotyping, we used two algorithms for summarizing the data from the probe sets into a single measurement, followed by clustering the resulting measurements into discrete copy-number classes (Supplementary Information). Although the two approaches agreed on the majority of calls (genotype concordance ≥99% for 96% of common CNPs), wherever they disagreed the approach that yielded the best separated clusters for that particular CNP was preferred. The joint use of the two platforms considerably improved the separation of genotype classes (Supplementary Fig. 1).

Sequence SNPs. Ten ENCODE regions were chosen on the basis of their overlap with previously sequenced ENCODE regions15, 16. PCR primers and conventional fluorescent DNA sequencing were used, and the SNPs were identified and filtered as described in Supplementary Information.

Imputation. Imputation was performed using the MACH program28 (http://www.sph.umich.edu/csg/abecasis/MACH/download/). In all analyses, the set of samples whose genotypes were imputed did not overlap the set of samples used to construct reference panels. For the 1958 British birth cohort analysis, we imputed all available SNPs on chromosome 20. The 1958 British birth cohort samples had been previously genotyped on the Affymetrix 500K and Illumina 550K chips, so we used the 1958 British birth cohort Illumina 550K genotypes in tandem with either reference panel (HMIJ-CEU or CEU+TSI) to impute the known (but masked) Affymetrix 500K SNPs (Supplementary Information).

Received 10 November 2009; accepted 21 June 2010.
17. The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.
18. The Cleveland Clinic, Beohriner, W. E. and Durbin, R. for critical reading of the manuscript. We thank the Wellcome Trust Case Control Consortium. genomewide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 437, 1299–1302 (2005).
19. The Broad Institute, Cheng, Y. Li, Y. & Abecasis, G. R. Mach 1.0: rapid haplotype reconstruction and missing data.
20. The Institute of Medical Genetics, University Central Hospital, Helsinki, Finland. We thank the Wellcome Trust Case Control Consortium (http://www.wtcc.org.uk), which was funded by Wellcome Trust award 070113.
21. Broad Institute, 7 Cambridge Center, Cambridge, Massachusetts 02138, USA. "Bayor College of Medicine, Human Genome Sequencing Center, Department of Molecular and Human Genetics, One Baylor Plaza, Houston, Texas 77030, USA. "University of Geneva, Medical School, Department of Genetic Medicine and Development, Faculty of Medicine, Geneva 1211, Switzerland. "Division of Genetics, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts 02115, USA. "Wellcome Trust Sanger Institute, Department of Human Genetics, Wellcome Trust Genome Campus, Cambridge CB10 1HH, UK. "Institute for Medical Genetics, University of Helsinki, Haartmaninkatu 8, 00290 Helsinki, Finland. "Department of Medical Genetics, University of Helsinki and University Central Hospital, Haartmaninkatu 8, 00290 Helsinki, Finland. "Harvard Medical School, Brigham and Women's Hospital, Department of Pathology, Boston, Massachusetts 02215, USA. "Cornell University, Department of Biological Statistics and Computational Biology, I02A Weill Hall, Ithaca, New York 14853, USA. "Harvard School of Public Health, Departments of Biostatistics and Biostatistical Consulting, 677 Huntington Avenue, Boston, Massachusetts 02115, USA. "Massachusetts General Hospital, Center for Human Genetic Research, Simchon Research Center, 185 Cambridge Street, Boston, Massachusetts 02114, USA. "University of Oxford, Department of Oncology, 1 South Parks Road, Oxford OX1 3TG, UK. "Harvard University, Department of Organismic and Evolutionary Biology, Center for Systems Biology, 52 Oxford Street, Room 469, Cambridge, Massachusetts 02115, USA. "University of Maryland School of Medicine, Department of Epidemiology and Preventive Medicine, N106 Institute of Human Virology, 725 West Lombard Street, Baltimore, Maryland 21201, USA. "University of Oklahoma, Department of Anthropology, 455 West Lindsey Room 505C, Norman, Oklahoma 73019, USA. "University of California, San Francisco, Department of Anthropology, History and Social Medicine, 333 California Street Suite 485, San Francisco, California 94143-0850, USA. "The Australian National University, John Curtin School of Medical Research, Garran Road, Building 131, Canberra, ACT 2601, Australia. "Institute for Oncological Study and Prevention, Florence 50139, Italy. "Case Western Reserve University, Department of Bioethics, School of Medicine TA200, 10900 Euclid Avenue, Cleveland, OH 44106, USA. "National Institutes of Health, University of Hawaii, 1757 Kanawai, Tsubetsu-cho, Ishikari-gun, Hokkaido 061-0293, Japan. "University of Minnesota, Department of Population and Family Health, PO Box 4606, Elrod 3010, Kenya. "National Human Genome Research Institute, National Institutes of Health, 5635 Fishers Lane Suite 3039, Bethesda, Maryland 20892, USA. "University of Houston at Clear Lake, Department of Anthropology, 2700 Bay Area Boulevard, PO Box 295, Houston, Texas 77058-1098, USA. "National Genome Research Institute, Center for Research on Genomics and Global Health, 12 South Drive, MSC 5635, Building 12A, Room 4047, Bethesda, Maryland 20892-5635, USA. "Duke University, Institute for Genome Sciences and Policy, 450 Research Drive, PO Box 91009, LSCR B-B Wing, Room 3208, Durham, North Carolina 27708, USA. "The Cleveland Clinic, Department of Bioethics, 9500 Euclid Avenue 16J0, Cleveland, Ohio 44124, USA. "Institute of Medical Genetics, Chinese Academy of Science, Beijing Airport Industrial Zone B-6, Beijing 101300, China. "National Human Genome Research Institute, National Institutes of Health, 5635 Fischers Lane Suite 4076, Bethesda, Maryland 20892, USA. ©2010 Macmillan Publishers Limited. All rights reserved.