contains all genotype submissions, 'redundant filtered' contains all submissions that pass QC, and 'non-redundant filtered' contains a single QC+ submission for each SNP in each analysis panel.

The QC filters remove SNPs showing gross errors. However, it is also important to understand the magnitude and structure of more subtle genotyping errors among SNPs that pass QC. We therefore carried out a series of analyses to assess the influence of the long-range PCR amplicon structure on genotyping error, the concordance rates between genotype calls from different genotyping platforms and between those platforms and re-sequencing assays, as well as the rates of false monomorphism and mis-mapping of SNPs (see Supplementary Text 2, Supplementary Figs 1–3 and Supplementary Tables 1–4). We estimate that the average per genotype accuracy is at least 99.5%. However, there are higher rates of missing data and genotype discrepancies at non-reference alleles, with some clustering of errors resulting from the amplicon design and a few incorrectly mapped SNPs.

Table 1 shows the numbers of SNPs attempted and converted to QC+ SNPs in each analysis panel (Supplementary Table 5 shows a breakdown by each major submission). Haplotypes and missing data were estimated for each analysis panel separately using both trio information and statistical methods based on the coalescent model (see Methods). To enable cross-population comparisons, a consensus data set was created consisting of 3,107,620 SNPs that were QC+ in all analysis panels and polymorphic in at least one analysis panel. The equivalent figure from Phase I was 931,340 SNPs. Unless stated otherwise, all analyses have been carried out on the consensus data set. An additional set of haplotypes was created for those SNPs in the consensus where a putative ancestral state could be assigned by comparison of the human alleles to the orthologous position in the chimpanzee and rhesus macaque genomes.

The variation in SNP density within the Phase II HapMap is shown in Fig. 1. On average there are 1.14 genotyped polymorphic SNPs per kilobase (average spacing is 875 base pairs (bp)) and 98.6% of the assembled genome is within 5 kb of the nearest polymorphic SNP. Still, there is heterogeneity in genotyped SNP density at both broad (Fig. 1a) and fine (Fig. 1b) scales. Furthermore, there are systematic changes in genotyped SNP density around genomic features including genes (Fig. 1c).

The Phase II HapMap differs from the Phase I HapMap not only in SNP spacing, but also in minor allele frequency distribution and patterns of linkage disequilibrium (Supplementary Fig. 4). Because the criteria for choosing additional SNPs did not include consideration of SNP spacing or preferential selection for high MAF, the SNPs added in Phase II are, on average, more clustered and have lower MAF than the Phase I SNPs. Because MAF predictably influences the distribution of linkage disequilibrium statistics, the average r^2 at a given physical distance is typically lower in Phase II than in Phase I; conversely, the $|D'|$ statistic is typically higher (data not shown). One notable consequence is that the Phase II HapMap includes a better representation of rare variation than the Phase I HapMap.

The increased resolution provided by Phase II of the project is illustrated in Fig. 2. Broadly, an additional SNP added to a region shows one of three patterns. First, it may be very similar in distribution to SNPs present in Phase I. Second, it may provide detailed resolution of haplotype structure (for example, a group of chromosomes with identical local haplotypes in Phase I can be shown in Phase II to carry

Table 1 | Summary of Phase II HapMap data (release 21)

Phase	SNP categories	Analysis panel		
		YRI	CEU	CHB+JPT
I	Assays submitted	1,304,199	1,344,616	1,306,125
	Passed QC	1,177,312 (90%)	1,217,902 (91%)	1,187,800 (91%)
	Did not pass QC	126,887 (10%)	126,714 (9%)	118,325 (9%)
	>20% missing	82,463 (65%)	95,684 (76%)	78,323 (66%)
	>1 duplicate inconsistent	6,049 (5%)	5,126 (4%)	9,242 (8%)
	>1 mendelian error	18,916 (15%)	11,310 (9%)	N/A
	<0.001 Hardy–Weinberg *P* -value	10,265 (8%)	8,922 (7%)	13,722 (12%)
	Other failures	19,345 (15%)	13,858 (11%)	20,674 (17%)
II	Assays submitted	5,044,989	5,044,996	5,043,775
	Passed QC	3,150,433 (62%)	3,204,709 (64%)	3,244,897 (64%)
	Did not pass QC	1,894,556 (38%)	1,840,287 (36%)	1,798,878 (36%)
	>20% missing	1,419,000 (75%)	1,398,166 (76%)	1,403,543 (78%)
	>1 duplicate inconsistent	0 (0%)	0 (0%)	6,617 (0%)
	>1 mendelian error	172,339 (9%)	127,923 (7%)	N/A
	<0.001 Hardy–Weinberg *P* -value	96,231 (5%)	82,268 (4%)	108,880 (6%)
	Other failures	334,511 (18%)	337,906 (18%)	340,370 (19%)
Overall	Assays submitted	6,349,188	6,389,612	6,349,900
	Passed QC	4,327,745 (68%)	4,422,611 (69%)	4,432,697 (70%)
	Did not pass QC	2,021,443 (32%)	1,967,001 (31%)	1,917,203 (30%)
	>20% missing	1,501,463 (74%)	1,493,850 (76%)	1,481,866 (77%)
	>1 duplicate inconsistent	6,049 (0%)	5,126 (0%)	15,859 (1%)
	>1 mendelian error	191,255 (9%)	139,233 (7%)	N/A
	<0.001 Hardy–Weinberg *P* -value	106,496 (5%)	91,190 (5%)	122,602 (6%)
	Other failures	353,856 (18%)	351,764 (18%)	361,044 (19%)
	Non-redundant (unique) SNPs	3,796,934	3,868,157	3,890,416
	Monomorphic	861,299 (23%)	1,246,183 (32%)	1,410,152 (36%)
	Polymorphic	2,935,635 (77%)	2,621,974 (68%)	2,480,264 (64%)

SNP categories	All analysis panels
Unique QC-passed SNPs	4,000,107
Passed in one analysis panel	88,140 (2%)
Passed in two analysis panels	268,534 (7%)
Passed in three analysis panels (QC+3)	3,643,433 (91%)
QC+3 and monomorphic across three analysis panels	535,813
QC+3 and polymorphic in at least one analysis panel	3,107,620
QC+3 and polymorphic in all three analysis panels	2,006,352
QC+3 and MAF ≥ 0.05 in at least one of three analysis panels	2,819,322

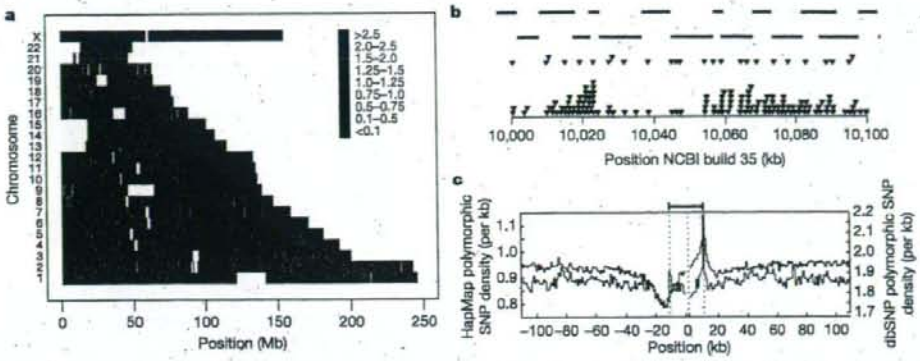

Figure 1 | SNP density in the Phase II HapMap. a, SNP density across the genome. Colours indicate the number of polymorphic SNPs per kb in the consensus data set. Gaps in the assembly are shown as white. **b**, Example of the fine-scale structure of SNP density for a 100-kb region on chromosome 17 showing Perlegen amplicons (black bars), polymorphic Phase I SNPs in the consensus data set (red triangles) and polymorphic Phase II SNPs in the consensus data set (blue triangles). Note the relatively even spacing of Phase I SNPs. **c**, The distribution of polymorphic SNPs in the consensus Phase II HapMap data (blue line and left-hand axis) around coding regions. Also shown is the density of SNPs in dbSNP release 125 around genes (red line and right-hand axis). Values were calculated separately 5′ from the coding start site (the left dotted line) and 3′ from the coding end site (right dotted line) and were joined at the median midpoint position of the coding unit (central dotted line).

multiple related haplotypes). Third, the novel SNP (or group of added SNPs) may reveal previously missed recombinant haplotypes. The extent to which each type of event occurs varies among populations and chromosomal regions. The greatest gains in resolution, in terms of identifying new recombinant haplotypes and haplotype groupings, occur in YRI. Consequently, the Phase II HapMap provides increased resolution in the estimated fine-scale genetic map and improved power to detect and localize recombination hotspots (Fig. 2b).

The use of the Phase II HapMap in association studies

The increased SNP density of the Phase II HapMap has already been extensively exploited in genome-wide studies of disease association. In this section, we quantify the gain in resolution and outline how the HapMap data can be used to improve the power of association studies.

Improved coverage of common variation. We previously predicted that the vast majority of common SNPs would be correlated to Phase II HapMap SNPs by extrapolation from the ten HapMap ENCODE regions[3]. Using the actual Phase II marker spacing and frequency distributions (Table 2), we repeated the simulations and estimate that Phase II HapMap marker sets capture the overwhelming majority of all common variants at high r^2. For common variants (MAF ≥ 0.05) the mean maximum r^2 of any SNP to a typed one is 0.90 in YRI, 0.96 in CEU and 0.95 in CHB+JPT. The impact of the

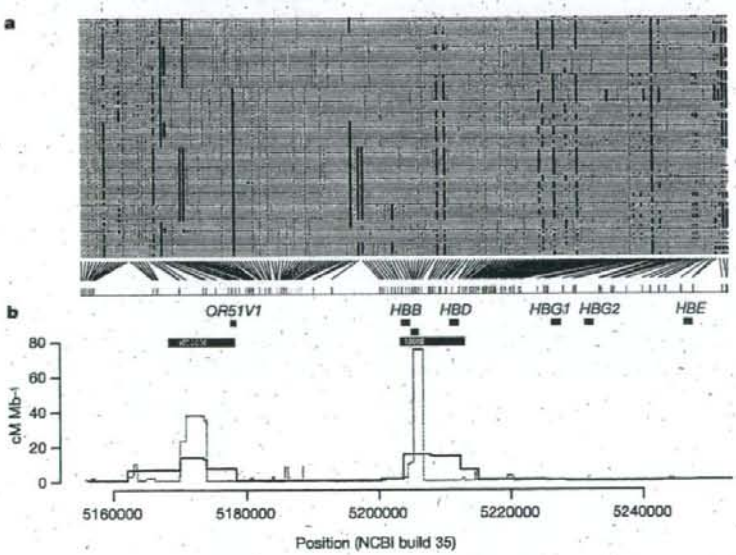

Figure 2 | Haplotype structure and recombination rate estimates from the Phase II HapMap. a, Haplotypes from YRI in a 100 kb region around the β-globin (*HBB*) gene. SNPs typed in Phase I are shown in dark blue. Additional SNPs in the Phase II HapMap are shown in light blue. Only SNPs for which the derived allele can be unambiguously identified by parsimony (by comparison with an outgroup sequence) are shown (89% of SNPs in the region); the derived allele is shown in colour. **b**, Recombination rates (lines) and the location of hotspots (horizontal blue bars) estimated for the same region from the Phase I (dark blue) and Phase II HapMap (light blue) data. Also shown are the location of genes within the region (grey bars) and the location of the experimentally verified recombination hotspot[37,38] at the 5′ end of the *HBB* gene (black bar).

Table 2 | Estimated coverage of the Phase II HapMap in the ten HapMap ENCODE regions

Panel	MAF bin	Phase I HapMap[a]		Phase II HapMap			
				Pairwise linkage disequilibrium		Additional 2-SNP tests	
		$r^2 \geq 0.8$ (%)	Mean maximum r^2	$r^2 \geq 0.8$ (%)	Mean maximum r^2	$r^2 \geq 0.8$ (%)	Mean maximum r^2
YRI	≥0.05	45	0.67	82	0.90	87	0.93
	<0.05			61	0.76	62	0.78
	0.05–0.10			81	0.89	81	0.89
	0.10–0.25			90	0.94	90	0.95
	0.25–0.50			87	0.93	92	0.96
CEU	≥0.05	74	0.85	93	0.96	95	0.97
	<0.05			70	0.79	72	0.81
	0.05–0.10			87	0.92	88	0.93
	0.10–0.25			94	0.96	95	0.97
	0.25–0.50			95	0.97	97	0.98
CHB+JPT	≥0.05	72	0.83	92	0.95	95	0.97
	<0.05			65	0.74	65	0.74
	0.05–0.10			81	0.89	82	0.89
	0.10–0.25			90	0.94	90	0.95
	0.25–0.50			94	0.96	97	0.98

2-SNP tests, linkage disequilibrium to haplotypes formed from two nearby SNPs.

Table 3 | Number of tag SNPs required to capture common (MAF ≥ 0.05) Phase II SNPs

Threshold	YRI	CEU	CHB+JPT
$r^2 \geq 0.5$	627,458	290,969	277,831
$r^2 \geq 0.8$	1,093,422	552,853	520,111
$r^2 = 1.0$	1,616,739	1,024,665	1,078,959

increased density of the Phase II HapMap is most notable in YRI (in the Phase I HapMap the mean maximum r^2 was 0.67). Similar results are found if a threshold of $r^2 \geq 0.8$ is used to determine whether an SNP is captured (Table 2). As expected, very common SNPs with MAF > 0.25 are captured extremely well (mean maximum r^2 of 0.93 in YRI to 0.97 in CEU), whereas rarer SNPs with MAF < 0.05 are less well covered (mean maximum r^2 of 0.74 in CHB+JPT to 0.76 in YRI). The latter figure is probably an overestimate because it is based on lower frequency SNPs discovered via re-sequencing 48 HapMap individuals, and does not include a much larger number of very rare SNPs. We also assessed the increase in coverage provided by using two-SNP haplotypes as proxies for SNPs that are poorly captured by single SNPs[16] (Table 2). These two-SNP haplotypes lead to a modest increase in mean maximum r^2 of 0.01 to 0.03 across all allele frequencies. However, in some regions, particularly where marker density is low, gains from multi-marker and imputation approaches in practical situations can be substantial (see below).

Currently, the Phase II HapMap provides the most complete available resource for selecting tag SNPs genome-wide. Using a simple pairwise tagging approach, we find that 1.09 million SNPs are required to capture all common Phase II SNPs with $r^2 \geq 0.8$ in YRI, with slightly more than 500,000 required in CEU and CHB+JPT (Table 3). These numbers are approximately twice those required to capture SNPs in the Phase I HapMap (which has one-third as many SNPs). The number of SNPs required to achieve perfect tagging ($r^2 = 1.0$) in each analysis panel is almost double that required to achieve the $r^2 \geq 0.8$ threshold. It becomes increasingly expensive to improve the coverage afforded by tags from the Phase I and, now, the Phase II HapMap, because additional tag SNPs are unlikely to capture large groups of additional SNPs.

Phase II HapMap and genome-wide association studies. Although the efficient choice of tag SNPs is one use of the Phase II HapMap, for most disease studies the tag SNPs genotyped will be primarily determined by the choice of a commercial platform for the experiment[17,18]. Using Phase II data, we estimated the coverage of several available products on which genome-wide association studies are already underway (Table 4). Similar to earlier estimates[17,18], these products typically perform well in CEU and CHB+JPT, and some also perform well in YRI. For example, arrays of approximately 500,000 SNPs capture 68–88% (depending on selection method) of all HapMap Phase II variation with $r^2 \geq 0.8$ in CEU. SNPs that are not included in the Phase II HapMap will be covered more poorly because most genotyping products were designed using HapMap data.

HapMap data have several additional roles in the analysis of disease-association studies using fixed marker sets. For example, the high-quality haplotype information within the Phase II HapMap can be used to aid the phasing of genotype data from new samples because additional haplotypes are likely to be locally very similar to at least one haplotype in the Phase II data. By a similar argument, missing genotypes can potentially be inferred through comparison to the Phase II haplotypes. Genotypes may be missing either because of genotyping failure or because the SNP was not assayed within the experiment. Therefore, the HapMap haplotypes provide a way of *in silico* genotyping Phase II SNPs that were not included in the experiment.

Although there is no clear consensus yet about the role of SNP imputation in the analysis of genome-wide association studies, high imputation accuracy can be achieved using model-based methods[19–23] and can lead to an increase in power[23,24]. To illustrate the possibilities, in the 500-kb HapMap ENCODE region on 8q24.11 (Supplementary Fig. 5) we evaluated imputation of Phase II SNPs from the Affymetrix GeneChip 500K array. To do this, we used a

Table 4 | Estimated coverage of commercially available fixed marker arrays

Platform*	YRI		CEU		CHB+JPT	
	$r^2 \geq 0.8$ (%)	Mean maximum r^2	$r^2 \geq 0.8$ (%)	Mean maximum r^2	$r^2 \geq 0.8$ (%)	Mean maximum r^2
Affymetrix GeneChip 500K	46	0.66	68	0.81	67	0.80
Affymetrix SNP Array 6.0	66	0.80	82	0.90	81	0.89
Illumina HumanHap300	33	0.56	77	0.86	63	0.78
Illumina HumanHap550	55	0.73	88	0.92	83	0.89
Illumina HumanHap650Y	66	0.80	89	0.93	84	0.90
Perlegen 600K	47	0.68	92	0.94	84	0.90

* Assuming all SNPs on the product are informative and pass QC; in practice these numbers are overestimates.

leave-one-out procedure to assess the accuracy of genotype prediction in the YRI. For SNPs with MAF ≥ 0.2, the average maximum r^2 to a typed SNP in the region is 0.59 compared to an average genotype prediction r^2 of 0.86. Furthermore, whereas 44% of such SNPs in the region have no single-marker proxy with $r^2 \geq 0.5$, fewer than 6% of the SNPs have a genotype imputation accuracy of $r^2 < 0.5$, establishing that accurate imputation can be achieved even in the population where linkage disequilibrium is the weakest.

New insights into linkage disequilibrium structure

The paradigm underlying association studies is that linkage disequilibrium can be used to capture associations between markers and nearby untyped SNPs. However, the Phase II HapMap has revealed several properties of linkage disequilibrium that illustrate the full complexity of empirical patterns of genetic variation. Two striking features are the long-range similarity among haplotypes, and SNPs that show almost no linkage disequilibrium with any other SNP.

The extent of recent common ancestry and segmental sharing. A simplified view of linkage disequilibrium is that genetic variation is organized in relatively short stretches of strong linkage disequilibrium (haplotype blocks), each containing only a few common haplotypes and separated by recombination hotspots across which little association remains[25]. Although this view has heuristic value, if chromosomes share a recent common ancestor then similarity between chromosomes can extend over considerable genetic distance and span multiple recombination hotspots[26]. The extent of such recent ancestry in the four populations surveyed here has not been characterized previously. Therefore we identified stretches of identity between pairs of chromosomes, both within and across individuals, reflecting autozygosity and identity-by-descent (IBD) (Fig. 3a). After first checking for stratification within each analysis panel (see Supplementary Text 3; none was found for YRI, CEU and JPT, and only small stratification was found for CHB), we calculated genome-wide probabilities of sharing 0, 1 or 2 chromosomes identical by descent for each pair of individuals (see Supplementary Text 4). In addition to identifying a few close relationships (as reported in HapMap Phase I[3]), we estimate that, on average, any two individuals from the same population share approximately 0.5% of their genome through recent IBD (Table 5). Using a hidden Markov model approach[27] (see Supplementary Text 5), we searched for such shared segments over 1-megabase (Mb) long and containing at least 50 SNPs, after first pruning the list of SNPs to remove local linkage disequilibrium. We find that 10–30% of pairs in each analysis panel share regions of extended identity resulting from sharing a common ancestor within 10–100 generations. These regions typically span hundreds of SNPs and can extend over tens of megabases (Table 5).

Similarly, extended stretches of homozygosity are indicative of recent inbreeding within populations[28,29]. Although short runs of homozygosity are commonplace, covering up to one-third of the genome and showing population differences reflective of ancient linkage disequilibrium patterns (Table 5 and Fig. 3b), very long homozygous runs exist that are clearly distinct from this process. Including two JPT individuals who have unusually high levels of homozygosity (NA18987 and NA18992) and one CEU individual (NA12874), we identified 79 homozygous regions over 3 Mb in 51 individuals, with many segments extending over 10 Mb (Supplementary Tables 7 and 8). Segments intersecting with suspected deletions were first removed from the analysis (Supplementary Text 6).

In studies of rare mendelian diseases, the extended haplotype sharing surrounding recent mutations, usually with a frequency of much less than 1%, has been exploited to great advantage through homozygosity mapping[30,31] and haplotype sharing[32] methods. In studies of common disease, extended haplotype sharing among patients potentially offers a route for identifying rare variants (MAF in the range of 1–5%) of high penetrance[33,34], which tend to be poorly captured through single-marker association with genome-wide arrays. To illustrate the idea, we identified SNPs where only two copies of the minor allele are present (referred to as '2-SNPs'), which have minor allele frequencies of 1–2%. We find that these are enriched approximately sevenfold (Table 5) among regions of IBD identified by the hidden Markov model approach. Notably, identification of IBD regions can be performed with the same genome-wide SNP data being

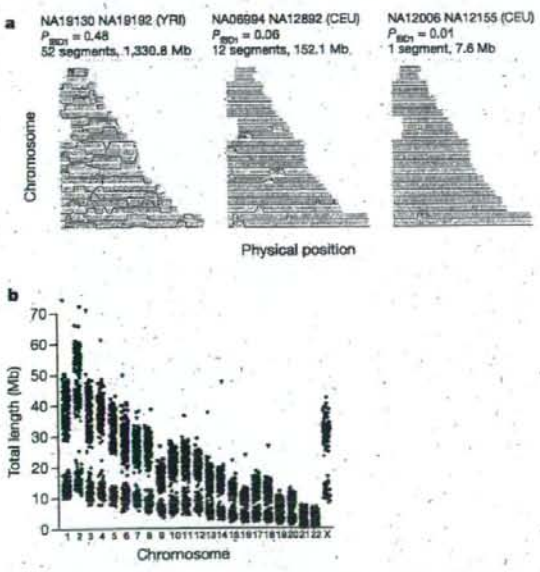

Figure 3 | The extent of recent co-ancestry among HapMap individuals. **a**, Three pairs of individuals with varying levels of identity-by-descent (IBD) sharing illustrate the continuum between very close and very distant relatedness and its relation to segmental sharing. The three pairs are: high sharing (NA19130 and NA19192 from YRI; previously identified as second-degree relatives[3]), moderate sharing (NA06994 and NA12892 from CEU) and low sharing (NA12006 and NA12155 from CEU). Along each chromosome, the probability of sharing at least one chromosome IBD is plotted, based on the HMM method described in Supplementary Text 5. Red sections indicate regions called as segments: in general, the proportion of the genome in segments is similar to each pair's estimated global relatedness. **b**, The extent of homozygosity on each chromosome for each individual in each analysis panel. Excludes segments <106 kb and chromosome X in males. Asterisk, NA12874, length = 107 Mb. YRI, green; CEU, orange; CHB, blue; JPT, magenta.

Table 5 | Relatedness, extended segmental sharing and homozygosity

Property	YRI	CEU	CHB	JPT
Number of pairs included	1,767	1,708	990	861
Mean identity by state (IBS) (%)	81.9	83.7	85.0	85.1
Mean identity by descent (IBD) (%)	0.04	0.34	0.36	0.42
Number of pairs with >1% IBD (%)	8.8	20.4	21.1	29.7
Number of pairs with one or more segment (%)	195 (11.0)	350 (20.5)	135 (13.6)	216 (25.1)
Total number of segments	250	427	146	273
Total distance spanned (Mb)	1,416	2,336	704	1,301
Mean segment length (Mb)	5.7	5.5	4.8	4.8
Maximum segment length (Mb)	51.7	56.2	15.0	25.3
Maximum segment length (Mb) (including close relatives)	141.4	128.5	N/A	N/A
Total number of 2-SNPs	6,219	9,220	8,174	8,750
Number of 2-SNPs in segments	109	162	116	132
2-SNP fold increase	6.7	7.3	7.6	7.0
Number of homozygous segments ($\times 10^3$)*	0.9	2.2	2.6	2.6
SNPs in homozygous segments ($\times 10^5$)	1.6	4.2	5.3	5.4
Total length of homozygous segments (Mb)	160	410	510	520

2-SNP, SNPs where only two copies of the minor allele are present.
* Homozygous segments >106 kb.

collected in large-scale association studies, making haplotype-sharing approaches an attractive and complementary analysis to standard SNP association tests, with the potential to identify rare variants associated with complex disease.

The distribution and causes of untaggable SNPs. Despite the SNP density of the Phase II HapMap, there are high-frequency SNPs for which no tag can be identified. Among high-frequency SNPs (MAF ≥ 0.2), we marked as untaggable SNPs to which no other SNP within 100 kb has an r^2 value of at least 0.2. In Phase II, approximately 0.5–1.0% of all high-frequency SNPs are untaggable and the proportion in YRI is approximately twice as high as in the other panels. Similar proportions are observed across the ten HapMap ENCODE regions.

To identify factors influencing the location of untaggable SNPs we considered their distribution relative to segmental duplications, repeat sequence, CpG dinucleotide density, regions of low SNP density, unusual allele frequency distribution, linkage disequilibrium patterns and recombination hotspots. We find no evidence for an enrichment of untaggable SNPs in segmental duplications or repeat sequence, as would be expected from mis-mapping of SNPs (2% and 35% of common SNPs lie in segmental duplications and repeat sequence, respectively, compared to 1.8% and 29%, respectively, of untaggable SNPs). Untaggable SNPs are slightly enriched in CpG islands (0.37% of common SNPs are in CpG islands compared to 1.4% of untaggable SNPs) and have slightly reduced MAF (Fig. 4). Most notably, untaggable SNPs are strongly enriched in regions of low linkage disequilibrium, particularly in recombination hotspots. To test whether these untaggable SNPs are themselves responsible for the identification of recombination hotspots, we eliminated them from 100 randomly chosen recombination hotspots and reassessed the evidence for a local peak in recombination. In all cases we still find evidence for a considerable increase in local recombination rate.

Over 50% of all untaggable SNPs lie within 1 kb of the centre of a detected recombination hotspot and over 90% are within 5 kb. Because only 3–4% of all SNPs lie within 1 kb from the centre of a detected recombination hotspot (16% are within 5 kb), this constitutes a marked enrichment and implies that at least 10% of all SNPs within 1 kb of hotspots are untaggable. The implication for association mapping is that when a region of interest contains a known hotspot it may be prudent to perform additional sequencing within the hotspot. Many of the variants identified in this manner will be untaggable SNPs that should be genotyped directly in association studies. From a biological perspective, the proximity of untaggable SNPs to the centre of hotspots suggests that they may lie within gene conversion tracts associated with the repair of double-strand breaks. Double-strand breaks are thought to resolve as crossover events only 5–25% of the time[35]. Consequently, SNPs lying near the centre of a hotspot are liable to be included within gene conversion tracts and will experience much higher effective recombination rates than predicted from crossover rates alone.

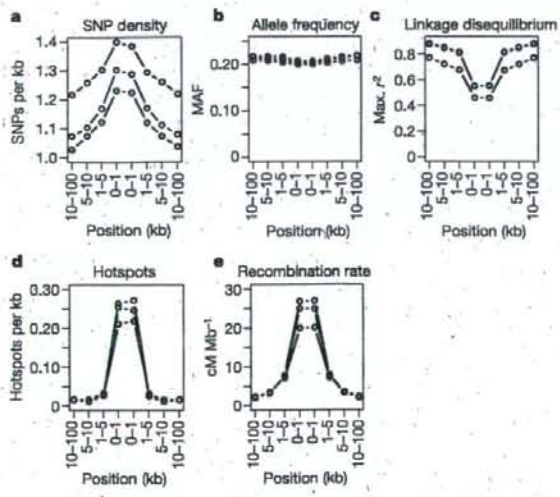

Figure 4 | Properties of untaggable SNPs. a–e, Properties of the genomic regions surrounding untaggable SNPs in terms of: **a**, the density of polymorphic SNPs within the consensus data set; **b**, mean minor allele frequency of polymorphic SNPs; **c**, maximum r^2 of SNPs to any others in the Phase II data; **d**, the density of estimated recombination hotspots (defined from hotspot centres); and **e**, the estimated mean recombination rate. YRI, green; CEU, orange; CHB+JPT, purple.

The distribution of recombination

In the Phase II HapMap we identified 32,996 recombination hotspots[3,6,36] (an increase of over 50% from Phase I) of which 68% localized to a region of ≤5 kb. The median map distance induced by a hotspot is 0.043 cM (or one crossover per 2,300 meioses) and the hottest identified, on chromosome 20, is 1.2 cM (one crossover per 80 meioses). Hotspots account for approximately 60% of recombination in the human genome and about 6% of sequence (Supplementary Fig. 6). We do not find marked differences among chromosomes in the concentration of recombination in hotspots, which implies that obligate differences in recombination among chromosomes of different size result from differences in hotspot density and intensity[6].

The increased number of well-defined hotspots allows us to understand better the influence of genomic features on the distribution of recombination. Previous work identified specific DNA motifs that influence hotspot location[6,37] as well as additional influences of local sequence context including the location of genes[6] and base composition[38]. The Phase II HapMap provides the resolution to separate these influences. Figure 5a shows the distribution of recombination, hotspot motifs and base composition around genes. Within the transcribed region of genes there is a marked decrease in the estimated recombination rate. However, 5′ of the transcription start site is a peak in recombination rate with a corresponding local increase in the density of hotspot motifs. This region also shows a marked increase in G+C content, reflecting the presence of CpG islands in promoter regions. There is also an asymmetry in recombination rate across genes, with recombination rates 3′ of transcribed regions being elevated (as are motif density and G+C content) compared to regions 5′ of genes. Studies in yeast have previously suggested an association between promoter regions and recombination hotspots[39]. Our results suggest a significant, although weak, relationship between promoters and recombination in humans. Nevertheless, the vast majority of hotspots in the human genome are not in gene promoters. The association may reflect a general association between regions of accessible chromatin and crossover activity.

Systematic differences in recombination rate by gene class. Previous work has demonstrated differences in the magnitude of linkage disequilibrium, as measured at a megabase scale, among genes associated with different functions[3,40]. Using the fine-scale genetic map estimated from the Phase II HapMap data we can quantify local increases in recombination rate associated with genes of different function using the Panther gene ontology annotation[41]. Average recombination rates vary more than sixfold among such gene classes (Fig. 5b), with defence and immunity genes showing the highest rates (1.9 cM Mb^{-1}) and chaperones showing the lowest rates (0.3 cM Mb^{-1}). Gene functions associated with cell surfaces and external functions tend to show higher recombination rates (immunity, cell adhesion, extracellular matrix, ion channels, signalling) whereas those with lower recombination rates are typically internal to cells (chaperones, ligase, isomerase, synthase). Controlling for systematic differences between gene classes in base composition and gene clustering, the differences between groups remain significant.

We also find that the density of hotspot-associated DNA motifs varies systematically among gene classes and that variation in motif density explains over 50% of the variance in recombination rate among gene functions (Supplementary Fig. 7).

These results pose interesting evolutionary questions. Because recombination involves DNA damage through double-strand breaks, hotspots may be selected against in some highly conserved parts of the genome. In regions exposed to recurrent selection (for example, from changes in environment or pathogen pressure) it is plausible that recombination may be selected for. However, because the fine-scale structure of recombination seems to evolve rapidly[42,43] it will be important to learn whether patterns of recombination rate heterogeneity among molecular functions are conserved between species.

Natural selection

The Phase I HapMap data have been used to identify genomic regions that show evidence for the influence of adaptive evolution[3,9], primarily through extended haplotype structure indicative of recent positive selection. Using two established approaches[9,44], we identified approximately 200 regions with evidence of recent positive selection from the Phase II HapMap (Supplementary Table 9). These regions include many established cases of selection, such as the genes *HBB* and *LCT*, the HLA region, and an inversion on chromosome 17. Many other regions have been previously identified in HapMap Phase I including *LARGE*, *SYT1* and *SULT1C2* (previously called *SULT1C1*). A detailed description of the findings from the Phase II HapMap is published elsewhere[45].

The Phase II HapMap also provides new insights into the forces acting on SNPs in coding regions. Effort was made to genotype as many known or putative non-synonymous SNPs as possible. Of the 56,789 non-synonymous SNPs identified in dbSNP release 125, attempts were made to genotype 36,777, which resulted in 17,427 that are QC+ in all three analysis panels and polymorphic. We selected only those SNPs for which ancestral allele information was available (approximately 90%). For comparison, we used patterns of variation at synonymous SNPs. As previously reported[46,47], non-synonymous SNPs show an increase in frequency of rare variants and a slight decrease of common variants compared to synonymous SNPs, compatible with widespread purifying selection against non-synonymous mutations (Fig. 6a). In contrast, we find no excess of high-frequency derived non-synonymous mutations, as might be expected if positive selection were widespread.

Natural selection also influences the extent to which allele frequencies differ between populations, not only through local selective pressures that drive alleles to different frequencies[48,49], but also through local variation in the strength of purifying selection. We compared the distribution of population differentiation (as measured by F_{ST}, the proportion of total variation in allele frequency that is due to differences between populations) at non-synonymous SNPs and synonymous SNPs matched for allele frequency (Fig. 6b). We find a systematic bias for non-synonymous SNPs to show stronger differentiation than synonymous SNPs. Among SNPs showing high levels of differentiation there is a strong tendency for the derived allele to be at higher frequency in non-YRI populations. Among SNPs with $F_{ST} > 0.5$ between CEU and YRI, in 79% and 75% of non-synonymous and synonymous variants, respectively, the derived allele is more common in CEU. Although this difference between non-synonymous and synonymous SNPs is not significant, among the eight exonic SNPs with $F_{ST} > 0.95$, all are non-synonymous. We see no such bias towards increased MAF in CEU at high-differentiation SNPs, indicating that SNP ascertainment is unlikely to explain the difference. Rather, this effect can largely be explained by more genetic drift in the non-African populations, as confirmed by simulations (data not shown). In addition, reduced selection against deleterious mutations and local adaptation within non-African populations will both act to increase the frequency of derived variants in non-African populations.

To assess the evidence for widespread local adaptation influencing non-synonymous mutations we considered the distribution of integrated extended haplotype homozygosity (iEHH) statistics[9,44] (Fig. 6c). We find no evidence for systematic differences between non-synonymous and synonymous SNPs, suggesting that local adaptation does not explain their higher differentiation. Although hitch-hiking effects will tend to obscure differences between selected

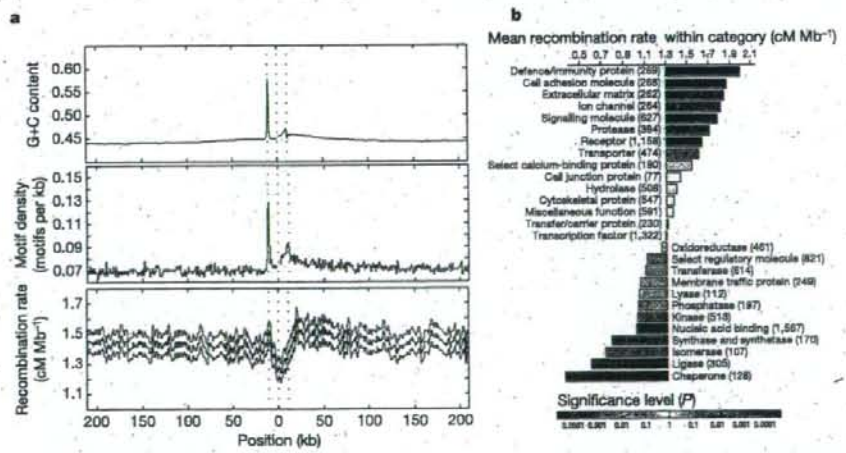

Figure 5 | Recombination rates around genes. a, The recombination rate, density of recombination-hotspot-associated motifs (all motifs with up to 1 bp different from the consensus CCTCCCTNNCCAC) and G+C content around genes. The blue line indicates the mean. For the recombination rate, grey lines indicate the quartiles of the distribution. Values were calculated separately 5′ from the transcription start site (the first dotted line) and 3′ from the transcription end site (third dotted line) and were joined at the median midpoint position of the transcription unit (central dotted line). Note the sharp drop in recombination rate within the transcription unit, the local increase around the transcription start site and the broad decrease away from the 3′ end of genes. These patterns only partly reflect the distribution of G+C content and the hotspot-associated motif, suggesting that additional factors influence recombination rates around genes. **b**, Recombination rates within genes of different molecular function[41]. The chart shows the increase or decrease for each category compared to the genome average. *P* values were estimated by permutation of category; numbers of genes are shown in parentheses.

and neutral SNPs, these results are consistent with a scenario in which the higher differentiation of non-synonymous SNPs is primarily driven by a reduction in the strength or efficacy of purifying selection in non-African populations.

Discussion and prospects

The International HapMap Project has been instrumental in making well-powered, large-scale, genome-wide association studies a reality. It is now clear that the HapMap can be a useful resource for the design and analysis of disease association studies in populations across the world[50–53]. Furthermore, the decreasing costs and increasing SNP density of standard genotyping panels mean that the focus of attention in disease association studies is shifting from candidate gene approaches towards genome-wide analyses. Alongside developments in technology, new statistical methodologies aimed at improving aspects of analysis, such as genotype calling[21,54], the identification of and correction for population stratification and relatedness[55,56], and imputation of untyped variants[21–23], are increasing the accuracy and reliability of genome-wide association studies.

Within this context, it is important to consider the future of the HapMap Project. Currently, additional samples from the populations used to develop the initial HapMap, as well as samples from seven additional populations (Luhya in Webuye, Kenya; Maasai in Kinyawa, Kenya; Tuscans in Italy; Gujarati Indian in Houston, Texas, USA; Denver (Colorado) metropolitan Chinese community; people of Mexican origin in Los Angeles, California, USA; and people with African ancestry in the southwestern United States; http://ccr.coriell.org/Sections/Collections/NHGRI/?SsId=11) will be sequenced and genotyped extensively to extend the HapMap, providing information on rarer variants and helping to enable genome-wide association studies in additional populations. There are also ongoing efforts by many groups to characterize additional forms of genetic variation, such as structural variation, and molecular phenotypes in the HapMap samples. Finally, in the future, whole-genome sequencing will provide a natural convergence of technologies to type both SNP and structural variation. Nevertheless, until that point, and even after, the HapMap Project data will provide an invaluable resource for understanding the structure of human genetic variation and its link to phenotype.

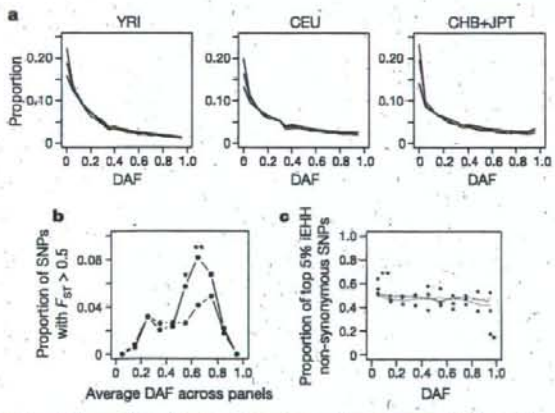

Figure 6 | Properties of non-synonymous and synonymous SNPs. a, The derived allele frequency (DAF) spectrum in each analysis panel for all SNPs (black), synonymous SNPs (green) and non-synonymous SNPs (red). Note the excess of rare variants for coding sequence SNPs but no excess of high-frequency derived variants. **b,** Enrichment of non-synonymous SNPs among genic SNPs showing high differentiation. For each of ten classes of derived allele frequency (averaged across analysis panels) the fraction of non-synonymous (red) and synonymous (green) variants in that class that show $F_{ST} > 0.5$ is shown. Note the strong enrichment of non-synonymous SNPs among SNPs of moderate to high derived-allele frequency (asterisk, $P < 0.05$; double asterisk, $P < 0.01$). **c,** Lack of enrichment of non-synonymous SNPs among those showing long-range haplotype structure. The integrated extended haplotype homozygosity (iEHH) statistic[9] was calculated for non-synonymous and synonymous SNPs in each analysis panel (YRI, green; CEU, orange; CHB+JPT, purple). For each of ten derived allele frequency classes, the proportion of non-synonymous SNPs among those showing the 5% most extreme statistics (within the allele frequency class) is shown (points). Also shown is the proportion of non-synonymous SNPs among SNPs in the coding sequence for each frequency class (dotted lines). Differences between synonymous and non-synonymous SNPs are tested for using a contingency table test.

METHODS SUMMARY

Of approximately 6.9 million SNPs in dbSNP release 122 approximately 4.7 million were selected for genotyping by Perlegen. 2.5 million SNPs were excluded because no assay could be designed and a further 350,000 were excluded for other reasons (see Methods). Perlegen performed genotyping using custom high-density oligonucleotide arrays as previously described[13]. Additional genotype submissions are described in the text. QC filters were applied as previously described[3]. Where multiple submissions met the QC criteria the submission with the lowest missing data rate was chosen for inclusion in the non-redundant filtered data set. Haplotypes were estimated from genotype data as described previously[3]. Ancestral states at SNPs were inferred by parsimony by comparison to orthologous bases in the chimpanzee (panTro2) and rhesus macaque (rheMac2) assemblies. Recombination rates and the location of recombination hotspots were estimated as described previously[3]. Additional details can be found in the Methods section and the Supplementary Information. The data described in this paper are in release 21 of the International HapMap Project.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 12 April; accepted 18 September 2007.

1. The International HapMap Consortium. Integrating ethics and science in the International HapMap Project. *Nature Rev. Genet.* 5, 467–475 (2004).
2. The International HapMap Consortium. The International HapMap Project. *Nature* 426, 789–796 (2003).
3. The International HapMap Consortium. A haplotype map of the human genome. *Nature* 437, 1299–1320 (2005).
4. Bowcock, A. M. Genomics: guilt by association. *Nature* 447, 645–646 (2007).
5. Altshuler, D. & Daly, M. Guilt beyond a reasonable doubt. *Nature Genet.* 39, 813–815 (2007).
6. Myers, S., Bottolo, L., Freeman, C., McVean, G. & Donnelly, P. A fine-scale map of recombination rates and hotspots across the human genome. *Science* 310, 321–324 (2005).
7. McCarroll, S. A. *et al.* Common deletion polymorphisms in the human genome. *Nature Genet.* 38, 86–92 (2006).
8. Conrad, D. F., Andrews, T. D., Carter, N. P., Hurles, M. E. & Pritchard, J. K. A high-resolution survey of deletion polymorphism in the human genome. *Nature Genet.* 38, 75–81 (2006).
9. Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. A map of recent positive selection in the human genome. *PLoS Biol.* 4, e72 (2006).
10. Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* 444, 444–454 (2006).
11. de Bakker, P. I. *et al.* A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. *Nature Genet.* 38, 1166–1172 (2006).
12. Pastinen, T. *et al.* Mapping common regulatory variants to human haplotypes. *Hum. Mol. Genet.* 14, 3963–3971 (2005).
13. Stranger, B. E. *et al.* Genome-wide associations of gene expression variation in humans. *PLoS Genet.* 1, e78 (2005).
14. Cheung, V. G. *et al.* Mapping determinants of human gene expression by regional and genome-wide association. *Nature* 437, 1365–1369 (2005).
15. Hinds, D. A. *et al.* Whole-genome patterns of common DNA variation in three human populations. *Science* 307, 1072–1079 (2005).
16. de Bakker, P. I. *et al.* Efficiency and power in genetic association studies. *Nature Genet.* 37, 1217–1223 (2005).
17. Pe'er, I. *et al.* Evaluating and improving power in whole-genome association studies using fixed marker sets. *Nature Genet.* 38, 663–667 (2006).
18. Barrett, J. C. & Cardon, L. R. Evaluating coverage of genome-wide association studies. *Nature Genet.* 38, 659–662 (2006).
19. Burdick, J. T., Chen, W. M., Abecasis, G. R. & Cheung, V. G. *In silico* method for inferring genotypes in pedigrees. *Nature Genet.* 38, 1002–1004 (2006).
20. Servin, B. R. & Stephens, M. Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genet.* 3, e114 (2007).

21. The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–668 (2007).
22. Scott, L. J. *et al.* A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* **316**, 1341–1345 (2007).
23. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies via imputation of genotypes. *Nature Genet.* **39**, 906–913 (2007).
24. Chapman, J. M., Cooper, J. D., Todd, J. A. & Clayton, D. G. Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Hum. Hered.* **56**, 18–31 (2003).
25. Paabo, S. The mosaic that is our genome. *Nature* **421**, 409–412 (2003).
26. McVean, G., Spencer, C. C. & Chaix, R. Perspectives on human genetic variation from the HapMap Project. *PLoS Genet.* **1**, e54 (2005).
27. Purcell, S. *et al.* PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
28. Broman, K. W. & Weber, J. L. Long homozygous chromosomal segments in reference families from the centre d'Etude du polymorphisme humain. *Am. J. Hum. Genet.* **65**, 1493–1500 (1999).
29. Gibson, J., Morton, N. E. & Collins, A. Extended tracts of homozygosity in outbred human populations. *Hum. Mol. Genet.* **15**, 789–795 (2006).
30. Lander, E. S. & Botstein, D. Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. *Science* **236**, 1567–1570 (1987).
31. Leutenegger, A. L. *et al.* Using genomic inbreeding coefficient estimates for homozygosity mapping of rare recessive traits: application to Taybi-Linder syndrome. *Am. J. Hum. Genet.* **79**, 62–66 (2006).
32. Te Meerman, G. J., Van der Meulen, M. A. & Sandkuijl, L. A. Perspectives of identity by descent (IBD) mapping in founder populations. *Clin. Exp. Allergy* **25** (Suppl 2), 97–102 (1995).
33. Houwen, R. H. *et al.* Genome screening by searching for shared segments: mapping a gene for benign recurrent intrahepatic cholestasis. *Nature Genet.* **8**, 380–386 (1994).
34. Durham, L. K. & Feingold, E. Genome scanning for segments shared identical by descent among distant relatives in isolated populations. *Am. J. Hum. Genet.* **61**, 830–842 (1997).
35. Jeffreys, A. J. & May, C. A. Intense and highly localized gene conversion activity in human meiotic crossover hot spots. *Nature Genet.* **36**, 151–156 (2004).
36. McVean, G. A. *et al.* The fine-scale structure of recombination rate variation in the human genome. *Science* **304**, 581–584 (2004).
37. Myers, S. *et al.* The distribution and causes of meiotic recombination in the human genome. *Biochem. Soc. Trans.* **34**, 526–530 (2006).
38. Spencer, C. C. *et al.* The influence of recombination on human genetic diversity. *PLoS Genet.* **2**, e148 (2006).
39. Petes, T. D. Meiotic recombination hot spots and cold spots. *Nature Rev. Genet.* **2**, 360–369 (2001).
40. Smith, A. V., Thomas, D. J., Munro, H. M. & Abecasis, G. R. Sequence features in regions of weak and strong linkage disequilibrium. *Genome Res.* **15**, 1519–1534 (2005).
41. Thomas, P. D. *et al.* PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res.* **13**, 2129–2141 (2003).
42. Winckler, W. *et al.* Comparison of fine-scale recombination rates in humans and chimpanzees. *Science* **308**, 107–111 (2005).
43. Ptak, S. E. *et al.* Fine-scale recombination patterns differ between chimpanzees and humans. *Nature Genet.* **37**, 429–434 (2005).
44. Sabeti, P. C. *et al.* Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**, 832–837 (2002).
45. Sabeti, P. C. *et al.* Genome-wide detection and characterization of positive selection in human populations. *Nature* doi:10.1038/nature06250 (this issue).
46. Bustamante, C. D. *et al.* Natural selection on protein-coding genes in the human genome. *Nature* **437**, 1153–1157 (2005).
47. Cargill, M. *et al.* Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genet.* **22**, 231–238 (1999).
48. Akey, J. M., Zhang, G., Zhang, K., Jin, L. & Shriver, M. D. Interrogating a high-density SNP map for signatures of natural selection. *Genome Res.* **12**, 1805–1814 (2002).
49. Sabeti, P. C. *et al.* Positive natural selection in the human lineage. *Science* **312**, 1614–1620 (2006).
50. de Bakker, P. I. *et al.* Transferability of tag SNPs in genetic association studies in multiple populations. *Nature Genet.* **38**, 1298–1303 (2006).
51. Conrad, D. F. *et al.* A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nature Genet.* **38**, 1251–1260 (2006).
52. Service, S., Sabatti, C. & Freimer, N. Tag SNPs chosen from HapMap perform well in several population isolates. *Genet. Epidemiol.* **31**, 189–194 (2007).
53. Lim, J. *et al.* Comparative study of the linkage disequilibrium of an ENCODE region, chromosome 7p15, in Korean, Japanese, and Han Chinese samples. *Genomics* **87**, 392–398 (2006).
54. Rabbee, N. & Speed, T. P. A genotype calling algorithm for affymetrix SNP arrays. *Bioinformatics* **22**, 7–12 (2006).
55. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
56. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genet.* **38**, 904–909 (2006).
57. Smith, R. A., Ho, P. J., Clegg, J. B., Kidd, J. R. & Thein, S. L. Recombination breakpoints in the human β-globin gene cluster. *Blood* **92**, 4415–4421 (1998).
58. Holloway, K., Lawson, V. E. & Jeffreys, A. J. Allelic recombination and *de novo* deletions in sperm in the human β-globin gene region. *Hum. Mol. Genet.* **15**, 1099–1111 (2006).
59. Weir, B. S. & Cockerham, C. C. Estimating F-statistics for the analysis of population structure. *Evolution* **38**, 1358–1370 (1984).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank many people who contributed to this project: all members of the genotyping laboratory and the sample, primer, bioinformatics, data quality and IT groups at Perlegen Sciences for technical and infrastructural support; J. Beck, C. Beiswanger, D. Coppock, A. Leach, J. Mintzer and L. Toji for transforming the Yoruba, Japanese and Han Chinese samples, distributing the DNA and cell lines, storing the samples for use in future research, and producing the community newsletters and reports; J. Greenberg and R. Anderson for providing funding and support for cell line transformation and storage in the NIGMS Human Genetic Cell Repository at the Coriell Institute; T. Dibling, T. Ishikura, S. Kanazawa, S. Mizusawa and S. Saito for help with genotyping; C. Hind and A. Moghadam for technical support in genotyping and all members of the subcloning and sequencing teams at the Wellcome Trust Sanger Institute; X. Ke for help with data analysis; Oxford E-Science Centre for provision of high-performance computing resources; H. Chen, W. Chen, L. Deng, Y. Dong, C. Fu, L. Gao, H. Geng, J. Geng, M. He, H. Li, H. Li, S. Li, X. Li, B. Liu, Z. Liu, F. Lu, F. Lu, G. Lu, C. Luo, X. Wang, Z. Wang, C. Ye and X. Yu for help with genotyping and sample collection; X. Feng, Y. Li, J. Ren and X. Zhou for help with sample collection; J. Fan, W. Gu, W. Guan, S. Hu, H. Jiang, R. Lei, Y. Lin, Z. Niu, B. Wang, L. Yang, W. Yang, Y. Wang, Z. Wang, S. Xu, W. Yan, H. Yang, W. Yuan, C. Zhang, J. Zhang, K. Zhang and G. Zhao for help with genotyping; P. Fong, C. Lai, C. Lau, T. Leung, L. Luk and W. Tong for help with genotyping; C. Pang for help with genotyping; K. Ding, B. Qiang, J. Zhang, X. Zhang and K. Zhou for help with genotyping; Q. Fu, S. Ghose, X. Lu, D. Nelson, A. Perez, S. Poole, R. Vega and H. Yonath for help with genotyping; C. Bruckner, T. Brundage, S. Chow, O. Iartchouk, M. Jain, M. Moorhead and K. Tran for help with genotyping; N. Addleman, J. Atilano, T. Chan, C. Chu, C. Ha, T. Nguyen, M. Minton and A. Phong for help with genotyping, and D. Lind for help with quality control and experimental design; R. Donaldson and S. Duan for help with genotyping, and J. Rice and N. Saccone for help with experimental design; J. Wigginton for help with implementing and testing QA/QC software; A. Clark, B. Keats, R. Myers, D. Nickerson and A. Williamson for providing advice to NIH; C. Juenger, C. Bennet, C. Bird, J. Melone, P. Nailer, M. Weiss, J. Witonsky and E. DeHaut-Combs for help with project management; M. Gray for organizing phone calls and meetings; D. Leja for help with figures; the Yoruba people of Ibadan, Nigeria, the people of Tokyo, Japan, and the community at Beijing Normal University, who participated in public consultations and community engagements; the people in these communities who donated their blood samples; and the people in the Utah CEPH community who allowed the samples they donated earlier to be used for the Project. This work was supported by the Japanese Ministry of Education, Culture, Sports, Science and Technology, the Wellcome Trust, Nuffield Trust, Wolfson Foundation, UK EPSRC, Genome Canada, Génome Québec, the Chinese Academy of Sciences, the Ministry of Science and Technology of the People's Republic of China, the National Natural Science Foundation of China, the Hong Kong Innovation and Technology Commission, the University Grants Committee of Hong Kong, the SNP Consortium, the US National Institutes of Health (FIC, NCI, NCRR, NEI, NHGRI, NIA, NIAAA, NIAID, NIAMS, NIBIB, NIDA, NIDCD, NIDCR, NIDDK, NIEHS, NIGMS, NIMH, NINDS, NLM, OD), the W.M. Keck Foundation, and the Delores Doré Eccles Foundation. All SNPs genotyped within the HapMap Project are available from dbSNP (http://www.ncbi.nlm.nih.gov/SNP); all genotype information is available from dbSNP and the HapMap website (http://www.hapmap.org).

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details accompany the full-text HTML version of the paper at www.nature.com/nature. Correspondence and requests for materials should be addressed to G.M. (mcvean@stats.ox.ac.uk) or M.D. (mjdaly@chgr.mgh.harvard.edu).

The International HapMap Consortium (Participants are arranged by institution and then alphabetically within institutions except for Principal Investigators and Project Leaders, as indicated.)

Genotyping centres: Perlegen Sciences Kelly A. Frazer (Principal Investigator)[1], Dennis G. Ballinger[2], David R. Cox[2], David A. Hinds[2], Laura L. Stuve[2]; **Baylor College of Medicine and ParAllele BioScience** Richard A. Gibbs (Principal Investigator)[3], John W. Belmont[3], Andrew Boudreau[4], Paul Hardenbol[5], Suzanne M. Leal[3], Shiran Pasternak[6], David A. Wheeler[3], Thomas D. Willis[4], Fuli Yu[7]; **Beijing Genomics Institute** Huanming Yang (Principal Investigator)[8], Changqing Zeng (Principal Investigator)[8], Yang Gao[8], Haoran Hu[8], Weitao Hu[8], Chaohua Li[8], Wei Lin[8], Siqi Liu[8], Hao Pan[8], Xiaoli Tang[8], Jian Wang[8], Wei Wang[8], Jun Yu[8], Bo Zhang[8], Qingrun Zhang[8], Hongbin Zhao[8], Hui Zhao[8], Jun Zhou[8]; **Broad Institute of Harvard and Massachusetts Institute of Technology**

Stacey B. Gabriel (Project Leader)[7], Rachel Barry[7], Brendan Blumenstiel[7], Amy Camargo[7], Matthew Defelice[7], Maura Faggart[7], Mary Goyette[7], Supriya Gupta[7], Jamie Moore[7], Huy Nguyen[7], Robert C. Onofrio[7], Melissa Parkin[7], Jessica Roy[7], Erich Stahl[7], Ellen Winchester[7], Lluda Ziaugra[7], David Altshuler (Principal Investigator)[7,9]; **Chinese National Human Genome Center at Beijing** Yan Shen (Principal Investigator)[10], Zhijian Yao[10]; **Chinese National Human Genome Center at Shanghai** Wei Huang (Principal Investigator)[11], Xun Chu[11], Yungang He[11], Li Jin[12], Yangfan Liu[11], Yayun Shen[11], Weiwei Sun[11], Haifeng Wang[11], Yi Wang[11], Ying Wang[11], Xiaoyan Xiong[11], Liang Xu[11]; **Chinese University of Hong Kong** Mary M. Y. Waye (Principal Investigator)[13], Stephen K. W. Tsui[13]; **Hong Kong University of Science and Technology** Hong Xue (Principal Investigator)[14], J. Tze-Fei Wong[14]; **Illumina** Luana M. Galver (Project Leader)[15], Jian-Bing Fan[15], Kevin Gunderson[15], Sarah S. Murray[1], Arnold R. Oliphant[16], Mark S. Chee (Principal Investigator)[17]; **McGill University and Génome Québec Innovation Centre** Alexandre Montpetit (Project Leader)[18], Fanny Chagnon[18], Vincent Ferretti[18], Martin Leboeuf[18], Jean-François Olivier[4], Michael S. Phillips[19], Stéphanie Roumy[19], Clémentine Sallée[19], Andrei Verner[18], Thomas J. Hudson (Principal Investigator)[20]; **University of California at San Francisco and Washington University** Pui-Yan Kwok (Principal Investigator)[21], Dongmei Cai[21], Daniel C. Koboldt[22], Raymond D. Miller[22], Ludmila Pawlikowska[21], Patricia Taillon-Miller[22], Ming Xiao[21]; **University of Hong Kong** Lap-Chee Tsui (Principal Investigator)[23], William Mak[23], You Qiang Song[23], Paul K. H. Tam[23]; **University of Tokyo and RIKEN** Yusuke Nakamura (Principal Investigator)[24,25], Takahisa Kawaguchi[25], Takuya Kitamoto[25], Takashi Morizono[25], Atsushi Nagashima[25], Yozo Ohnishi[25], Akihiro Sekine[25], Toshihiro Tanaka[25], Tatsuhiko Tsunoda[25]; **Wellcome Trust Sanger Institute** Panos Deloukas (Project Leader)[26], Christine P. Bird[26], Marcos Delgado[26], Emmanouil T. Dermitzakis[26], Rhian Gwilliam[26], Sarah Hunt[26], Jonathan Morrison[27], Don Powell[26], Barbara E. Stranger[26], Pamela Whittaker[26], David R. Bentley (Principal Investigator)[28]

Analysis groups: Broad Institute Mark J. Daly (Project Leader)[7,9], Paul I. W. de Bakker[7,9], Jeff Barrett[7,9], Yves R. Chretien[7], Julian Maller[7,9], Steve McCarroll[7,9], Nick Patterson[7], Itsik Pe'er[29], Alkes Price[7], Shaun Purcell[9], Daniel J. Richter[7], Pardis Sabeti[7], Richa Saxena[7,9], Stephen F. Schaffner[7], Pak C. Sham[29], Patrick Varilly[7], David Altshuler (Principal Investigator)[7,9]; **Cold Spring Harbor Laboratory** Lincoln D. Stein (Principal Investigator)[6], Lalitha Krishnan[6], Albert Vernon Smith[6], Marcela K. Tello-Ruiz[6], Gudmundur A. Thorisson[30]; **Johns Hopkins University School of Medicine** Aravinda Chakravarti (Principal Investigator)[31], Peter E. Chen[31], David J. Cutler[31], Carl S. Kashuk[31], Shin Lin[31]; **University of Michigan** Gonçalo R. Abecasis (Principal Investigator)[32], Weihua Guan[32], Yun Li[32], Heather M. Munro[33], Zhaohui Steve Qin[32], Daryl J. Thomas[34]; **University of Oxford** Gilean McVean (Project Leader)[35], Adam Auton[35], Leonardo Bottolo[35], Niall Cardin[35], Susana Eyheramendy[35], Colin Freeman[35], Jonathan Marchini[35], Simon Myers[35], Chris Spencer[7], Matthew Stephens[36], Peter Donnelly (Principal Investigator)[35]; **University of Oxford, Wellcome Trust Centre for Human Genetics** Lon R. Cardon (Principal Investigator)[37], Geraldine Clarke[38], David M. Evans[38], Andrew P. Morris[38], Bruce S. Weir[39]; **RIKEN** Tatsuhiko Tsunoda (Principal Investigator)[25], Todd A. Johnson[25]; **US National Institutes of Health** James C. Mullikin[40]; **US National Institutes of Health National Center for Biotechnology Information** Stephen T. Sherry[41], Michael Feolo[41], Andrew Skol[42]

Community engagement/public consultation and sample collection groups: Beijing Normal University and Beijing Genomics Institute Houcan Zhang[43], Changqing Zeng[8], Hui Zhao[8]; **Health Sciences University of Hokkaido, Eubios Ethics Institute, and Shinshu University** Ichiro Matsuda (Principal Investigator)[44], Yoshimitsu Fukushima[45], Darryl R. Macer[46], Eiko Suda[47]; **Howard University and University of Ibadan** Charles N. Rotimi (Principal Investigator)[48], Clement A. Adebamowo[49], Ike Ajayi[49], Toyin Aniagwu[49], Patricia A. Marshall[50], Chibuzor Nkwodimmah[49], Charmaine D. M. Royal[48]; **University of Utah** Mark F. Leppert (Principal Investigator)[51], Missy Dixon[51], Andy Peiffer[51]

Ethical, legal and social issues: Chinese Academy of Social Sciences Renzong Qiu[52]; **Genetic Interest Group** Alastair Kent[53]; **Kyoto University** Kazuto Kato[54]; **Nagasaki University** Norio Niikawa[55]; **University of Ibadan School of Medicine** Isaac F. Adewole[49]; **University of Montréal** Bartha M. Knoppers[19]; **University of Oklahoma** Morris W. Foster[56]; **Vanderbilt University** Ellen Wright Clayton[57]; **Wellcome Trust** Jessica Watkin[58]

SNP discovery: Baylor College of Medicine Richard A. Gibbs (Principal Investigator)[3], John W. Belmont[3], Donna Muzny[3], Lynne Nazareth[3], Erica Sodergren[3], George M. Weinstock[3], David A. Wheeler[3], Imtaz Yakub[3]; **Broad Institute of Harvard and Massachusetts Institute of Technology** Stacey B. Gabriel (Project Leader)[7], Robert C. Onofrio[7], Daniel J. Richter[7], Lluda Ziaugra[7], Bruce W. Birren[7], Mark J. Daly[7,9], David Altshuler (Principal Investigator)[7,9]; **Washington University** Richard K. Wilson (Principal Investigator)[59], Lucinda L. Fulton[59]; **Wellcome Trust Sanger Institute** Jane Rogers (Principal Investigator)[26], John Burton[26], Nigel P. Carter[26], Christopher M. Clee[26], Mark Griffiths[26], Matthew C. Jones[26], Kirsten McLay[26], Robert W. Plumb[26], Mark T. Ross[26], Sarah K. Sims[26], David L. Willey[26]

Scientific management: Chinese Academy of Sciences Zhu Chen[60], Hua Han[60], Le Kang[60]; **Genome Canada** Martin Godbout[61], John C. Wallenburg[62]; **Génome Québec** Paul L'Archevêque[63], Guy Bellemare[63]; **Japanese Ministry of Education, Culture, Sports, Science and Technology** Koji Saeki[64]; **Ministry of Science and Technology of the People's Republic of China** Hongguang Wang[65], Daochang An[65], Hongbo Fu[65], Qing Li[65], Zhen Wang[65]; **The Human Genetic Resource Administration of China** Renwu Wang[66]; **The SNP Consortium** Arthur L. Holden[15]; **US National Institutes of Health** Lisa D. Brooks[67], Jean E. McEwen[67], Mark S. Guyer[67], Vivian Ota Wang[67,68], Jane L. Peterson[67], Michael Shi[69], Jack Spiegel[70], Lawrence M. Sung[71], Lynn F. Zacharia[67], Francis S. Collins[72]; **Wellcome Trust** Karen Kennedy[61], Ruth Jamieson[58], John Stewart[58]

[1]The Scripps Research Institute, 10550 North Torrey Pines Road MEM275, La Jolla, California 92037, USA. [2]Perlegen Sciences, Inc., 2021 Stierlin Court, Mountain View, California 94043, USA. [3]Baylor College of Medicine, Human Genome Sequencing Center, Department of Molecular and Human Genetics, 1 Baylor Plaza, Houston, Texas 77030, USA. [4]Affymetrix, Inc., 3420 Central Expressway, Santa Clara, California 95051, USA. [5]Pacific Biosciences, 1505 Adams Drive, Menlo Park, California 94025, USA. [6]Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, New York 11724, USA. [7]The Broad Institute of Harvard and Massachusetts Institute of Technology, 1 Kendall Square, Cambridge, Massachusetts 02139, USA. [8]Beijing Genomics Institute, Chinese Academy of Sciences, Beijing 100300, China. [9]Massachusetts General Hospital and Harvard Medical School, Simches Research Center, 185 Cambridge Street, Boston, Massachusetts 02114, USA. [10]Chinese National Human Genome Center at Beijing, 3-707 N. Yongchang Road, Beijing Economic-Technological Development Area, Beijing 100176, China. [11]Chinese National Human Genome Center at Shanghai, 250 Bi Bo Road, Shanghai 201203, China. [12]Fudan University and CAS-MPG Partner Institute for Computational Biology, School of Life Sciences, SIBS, CAS, Shanghai 201203, China. [13]The Chinese University of Hong Kong, Department of Biochemistry, The Croucher Laboratory for Human Genetics, 6/F Mong Man Wai Building, Shatin, Hong Kong. [14]Hong Kong University of Science and Technology, Department of Biochemistry and Applied Genomics Center, Clear Water Bay, Knowloon, Hong Kong. [15]Illumina, 9885 Towne Centre Drive, San Diego, California 92121, USA. [16]Complete Genomics, Inc., 658 North Pastoria Avenue, Sunnyvale, California 94085, USA. [17]Prognosys Biosciences, Inc., 4215 Sorrento Valley Boulevard, Suite 105, San Diego, California 92121, USA. [18]McGill University and Génome Québec Innovation Centre, 740 Dr. Penfield Avenue, Montréal, Québec H3A 1A4, Canada. [19]University of Montréal, The Public Law Research Centre (CRDP), PO Box 6128, Downtown Station, Montréal, Québec H3C 3J7, Canada. [20]Ontario Institute for Cancer Research, MaRS Centre, South Tower, 101 College Street, Suite 500, Toronto, Ontario M5G 1L7, Canada. [21]University of California, San Francisco, Cardiovascular Research Institute, 513 Parnassus Avenue, Box 0793, San Francisco, California 94143, USA. [22]Washington University School of Medicine, Department of Genetics, 660 South Euclid Avenue, Box 8232, St Louis, Missouri 63110, USA. [23]University of Hong Kong, Genome Research Centre, 6/F, Laboratory Block, 21 Sassoon Road, Pokfulam, Hong Kong. [24]University of Tokyo, Institute of Medical Science, 4-6-1 Sirokanedai, Minato-ku, Tokyo 108-8639, Japan. [25]RIKEN SNP Research Center, 1-7-22 Suehiro-cho, Tsurumi-ku Yokohama, Kanagawa 230-0045, Japan. [26]Wellcome Trust Sanger Institute, The Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. [27]University of Cambridge, Department of Oncology, Cambridge CB1 8RN, UK. [28]Solexa Ltd, Chesterford Research Park, Little Chesterford, Nr Saffron Walden, Essex CB10 1XL, UK. [29]Columbia University, 500 West 120th Street, New York, New York 10027, USA. [30]University of Leicester, Department of Genetics, Leicester LE1 7RH, UK. [31]Johns Hopkins University School of Medicine, McKusick-Nathans Institute of Genetic Medicine, Broadway Research Building, Suite 579, 733 North Broadway, Baltimore, Maryland 21205, USA. [32]University of Michigan, Center for Statistical Genetics, Department of Biostatistics, 1420 Washington Heights, Ann Arbor, Michigan 48109, USA. [33]International Epidemiology Institute, 1455 Research Boulevard, Suite 550, Rockville, Maryland 20850, USA. [34]Center for Biomolecular Science and Engineering, Engineering 2, Suite 501, Mail Stop CBSE/ITI, UC Santa Cruz, Santa Cruz, California 95064, USA. [35]University of Oxford, Department of Statistics, 1 South Parks Road, Oxford OX1 3TG, UK. [36]University of Chicago, Department of Statistics, 5734 South University Avenue, Eckhart Hall, Room 126, Chicago, Illinois 60637, USA. [37]Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, Seattle, Washington 98109, USA. [38]University of Oxford/Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, UK. [39]University of Washington Department of Biostatistics, Box 357232, Seattle, Washington 98195, USA. [40]US National Institutes of Health, National Human Genome Research Institute, 50 South Drive, Bethesda, Maryland 20892, USA. [41]US National Institutes of Health, National Library of Medicine, National Center for Biotechnology Information, 8600 Rockville Pike, Bethesda, Maryland 20894, USA. [42]University of Chicago, Department of Medicine, Section of Genetic Medicine, 5801 South Ellis, Chicago, Illinois 60637, USA. [43]Beijing Normal University, 19 Xinjiekouwai Street, Beijing 100875, China. [44]Health Sciences University of Hokkaido, Ishikari Tobetsu Machi 1757, Hokkaido 061-0293, Japan. [45]Shinshu University School of Medicine, Department of Medical Genetics, Matsumoto 390-8621, Japan. [46]United Nations Educational, Scientific and Cultural Organization (UNESCO Bangkok), 920 Sukhumwit Road, Prakanong, Bangkok 10110, Thailand. [47]University of Tsukuba, Eubios Ethics Institute, PO Box 125, Tsukuba Science City 305-8691, Japan. [48]Howard University, National Human Genome Center, 2216 6th Street, NW, Washington DC 20059, USA. [49]University of Ibadan College of Medicine, Ibadan, Oyo State, Nigeria. [50]Case Western Reserve University School of Medicine, Department of Bioethics, 10900 Euclid Avenue, Cleveland, Ohio 44106, USA. [51]University of Utah, Eccles Institute of Human Genetics, Department of Human Genetics, 15 North 2030 East, Salt Lake City, Utah 84112, USA. [52]Chinese Academy of Social Sciences, Institute of Philosophy/Center for Applied Ethics, 2121, Building 9, Caoqiao Xinyuan 3 Qu, Beijing 100067, China. [53]Genetic Interest Group, 4D Leroy House, 436 Essex Road, London N130P, UK. [54]Kyoto University, Institute for Research in Humanities and Graduate School of Biostudies, Ushinomiya-cho, Sakyo-ku, Kyoto 606-8501, Japan. [55]Nagasaki University Graduate

School of Biomedical Sciences, Department of Human Genetics, Sakamoto 1-12-4, Nagasaki 852-8523, Japan. [56]University of Oklahoma, Department of Anthropology, 455 West Lindsey Street, Norman, Oklahoma 73019, USA. [57]Vanderbilt University, Center for Genetics and Health Policy, 507 Light Hall, Nashville, Tennessee 37232, USA. [58]Wellcome Trust, 215 Euston Road, London NW1 2BE, UK. [59]Washington University School of Medicine, Genome Sequencing Center, Box 8501, 4444 Forest Park Avenue, St Louis, Missouri 63108, USA. [60]Chinese Academy of Sciences, 52 Sanlihe Road, Beijing 100864, China. [61]Genome Canada, 150 Metcalfe Street, Suite 2100, Ottawa, Ontario K2P 1P1, Canada. [62]McGill University, Office of Technology Transfer, 3550 University Street, Montréal, Québec H3A 2A7, Canada. [63]Génome Québec, 630, boulevard René-Lévesque Ouest, Montréal, Québec H3B 1S6, Canada. [64]Ministry of Education, Culture, Sports, Science, and Technology, 3-2-2 Kasumigaseki, Chiyodaku, Tokyo 100-8959, Japan. [65]Ministry of Science and Technology of the People's Republic of China, 15 B. Fuxing Road, Beijing 100862, China. [66]The Human Genetic Resource Administration of China, b7, Zaojunmiao, Haidian District, Beijing 100081, China. [67]US National Institutes of Health, National Human Genome Research Institute, 5635 Fishers Lane, Bethesda, Maryland 20892, USA. [68]US National Institutes of Health, Office of Behavioral and Social Science Research, 31 Center Drive, Bethesda, Maryland 20892, USA. [69]Novartis Pharmaceuticals Corporation, Biomarker Development, One Health Plaza, East Hanover, New Jersey 07936, USA. [70]US National Institutes of Health, Office of Technology Transfer, 6011 Executive Boulevard, Rockville, Maryland 20852, USA. [71]University of Maryland School of Law, 500 West Baltimore Street, Baltimore, Maryland 21201, USA. [72]US National Institutes of Health, National Human Genome Research Institute, 31 Center Drive, Bethesda, Maryland 20892, USA.

METHODS

SNP selection and genotyping. All SNPs in dbSNP release 122 were considered for genotyping by Perlegen. Among these the following were excluded: SNPs for which no assay could be designed (primarily through location in repeat-rich regions; approximately 2.5 million); SNPs shown previously in samples from related populations[15] to be most probably in perfect association ($r^2 = 1$) with a Phase I SNP (approximately 122,000); all but one of SNPs shown previously[15] to be most probably in perfect association ($r^2 = 1$) with each other but not with a Phase I SNP (approximately 62,000); and SNPs shown previously[15] to have MAF < 0.05 (approximately 119,000). In addition, a few SNPs were excluded for efficiency (for example, if an amplicon contained a single SNP). Approximately 30,000 SNPs that had been typed in Phase I were deliberately retyped in Phase II to allow detailed comparisons of data quality, and an additional 15,000 SNPs that showed discrepancies between multiple genotyping attempts in Phase I were re-typed in Phase II. A further 2,000 SNPs identified by the Mammalian Gene Collection were also typed.

Perlegen performed genotyping using custom high-density oligonucleotide arrays as previously described[15]. Initially, a pilot phase was carried out on chromosome 2p to optimize experimental workflow and data handling. Details of amplicons used in the experiment and PCR primers can be found at http://genome.perlegen.com/pcr/ and also on the HapMap website. The arrays were tiled with sets of 25-bp probes for each SNP, with either 40 or 24 probes per SNP. These consisted of four sets of features, corresponding to forward and reverse strand tilings of sequences complementary to each of the two SNP alleles. Within a feature set, the position of the SNP within the oligonucleotide varied from position 11 to position 15. Mismatch probes were used to measure background, and by comparison with the perfect match probes, to detect the presence or absence of a specific PCR product. The 40-feature and 24-feature tilings both provided 10 perfect-match features for each SNP allele and differed only in the number of mismatch probes.

Genotypes were scored by clustering intensity measurements as previously described[15]. In addition, quality scores similar to Phred scores were computed for each genotype call, based on a combination of experimental metrics correlated to data quality. Assays with overall call rates less than 80% or with poor average quality scores were flagged as failed. About 38% of the tiled assays failed these basic criteria, and the remainder were processed using the more rigorous HapMap Project data quality control filters. For analysis of the whole genome, probes for 4,373,926 distinct SNPs were tiled onto 32 chip designs, with 32 SNPs tiled in replicate onto each chip design for quality control (QC). Perlegen did not type the samples by plates as had been done for the Phase I genotyping, instead typing large numbers of SNPs one sample at a time. Consequently, blank wells on each plate were not included as a component of QC for this genotyping. In the Phase I HapMap a single JPT sample had been excluded because of technical problems. Perlegen typed a replacement sample (from the original JPT collection) for all new SNPs. This sample was not specifically genotyped on the Phase I SNPs, although a substantial fraction of these was typed in Phase II.

Additional genotype submissions came from the Affymetrix GeneChip Human Mapping 500K array called with the BRLMM algorithm. In release 21a additional genotype submissions were incorporated from the MHC haplotype consortium[11], the Illumina HumanHap300 BeadChip, the Illumina Human-1 Genotyping BeadChip and the 10K non-synonymous SNP set from Affymetrix (ParAllele).

Details of primer design, DNA amplification, DNA labelling and hybridization and signal detection for the Perlegen platform can be found in Supplementary Text 7.

QC analyses. Genotype submissions were assessed for mendelian errors (where possible), missing data rates and Hardy–Weinberg proportions. QC filters were applied as previously described[3]; to achieve QC+ status a SNP had to have fewer than two mendelian errors, less than 20% missing data and $P > 0.001$ for Hardy–Weinberg analysis. The consensus data set consists only of SNPs for which QC+ submissions were available from all analysis panels. Where multiple submissions met the QC criteria the submission with the lowest missing data rate was chosen for inclusion in the non-redundant filtered data set. Comparison of the Phase II HapMap with the Affymetrix 500K genotypes has shown approximately 20 SNPs where the reported minor allele is discrepant (referred to as 'allele-flipping'). Over the entire data set, we expect that 500–2,000 SNPs have this problem and the vast majority will occur in SNPs from Phase I of the project. The Data Coordination Center (DCC) is working to resolve as many of these as possible.

Analyses of data quality. See Supplementary Text 2.

Analyses of population stratification, relatedness and homozygosity. See Supplementary Texts 3–6.

Analysis of recombination rate and gene ontology. We used the Panther Database[41] to obtain details of the gene molecular function and biological process. Genes are grouped into 28 top-level molecular function groups and 30 top-level biological process groups, with each gene allowed to exist in more than one group. We identified 14,979 non-overlapping autosomal genes from the Panther RefSeq Annotation for which we could obtain recombination rates. Of these, 9,735 had at least one assigned molecular function and 9,432 had at least one assigned biological process. Genes without a molecular function or biological process were removed from the corresponding analysis. To control for gene size, we estimated the mean recombination rate over a 20-kb region centred on the mid-point of each gene transcription region.

Genes were grouped based on molecular function and biological process. A mean recombination rate was calculated for each group. The significance of the result from each group was calculated via a permutation test involving 10^5 random groupings of genes. No correction was made for multiple testing. To account for the effect of G+C content on recombination, we performed a linear regression between the G+C content and recombination rate of all genes in each sample. Using the estimated regression parameters, the proportion of recombination explained by G+C content was subtracted from each gene.

Identification of non-synonymous SNPs and tests for natural selection. Using annotations from dbSNP release 125 we identified 17,427 polymorphic non-synonymous SNPs in release 21 and 15,976 polymorphic synonymous SNPs. Of these, 15,583 non-synonymous and 14,324 synonymous SNPs were autosomal and could have ancestral allele status unambiguously assigned by parsimony through comparison to the chimpanzee and macaque genomes. We used the phased haplotypes for analysis in which missing data had been imputed. F_{ST} was calculated using the method of Weir and Cockerham[39].

To detect recent partial selective sweeps we used the long-range haplotype (LRH) test[44,49] and the integrated haplotype score (iHS) test[5]. On simulated data[45], we found that the tests have similar power to detect recent selection but the iHS test has slightly lower power at low haplotype frequency and the LRH test has slightly lower power at high frequency. This can be seen in applications to HapMap Phase I data[3,5], where the iHS test misses the well-known cases of *HBB* and *CD36* and the LRH test misses the *SULT1C2* region. Although both tests are based on the concept of EHH[44], we observed that the false positives produced by the two tests tend not to overlap and thus that signals detected by both tests have a very low false-positive rate.

GSP online Genomics, Society and Policy
2007, Vol.3, No.1, pp.44-47

HUGO Statement on Pharmacogenomics (PGx): Solidarity, Equity and Governance

HUMAN GENOME ORGANISATION ETHICS COMMITTEE

Reason for Statement

The HUGO Ethics Committee,

- Recognising that there have been significant discussions of the ethical issues arising in the application of genetic knowledge;[1]
- Recognising that PGx has the potential to maximize therapeutic outcomes and minimize adverse reactions to therapy, and that it is consistent with the traditional goals of public health and medical care to relieve human suffering and save lives;
- Recognising that the ethical, research and policy issues that need to be considered in PGx include:
    - the ways in which PGx requires a novel focus on families, populations and communities;
    - the impact of 'personalised' medicine on a society;
    - the implications for populations in developing countries, including access to therapies for neglected diseases;
    - the impact of PGx on health care costs and policies worldwide;
    - the significance of PGx for research priorities;
    - the importance of re-evaluation of prior clinical trial data for efficacy among particular genotypes (as in resuscitation of abandoned drugs);
    - the significance of PGx for existing, as opposed to new, drugs; and
    - the fear that PGx could reinforce genetic determinism and lead to discrimination against, and stigmatization of, individuals and groups;

hereby identifies a pressing need to reach consensus on the most important ethical principles that are applicable and for workable guidelines in clinical and public health settings.

Definitions and scope

There has been extensive discussion of the relative benefits of using the terms pharmacogenomics (PGx) and pharmacogenetics, and various definitions have been suggested. To some extent the terms have been used interchangeably. Although both 'genetics' and 'genomics' are popularly used in different ways in different cultures, and have different meanings in a variety of languages, in this statement we use the term pharmaco*genomics* (PGx) and understand 'PGx' to mean the total sum of genetic variation that affects response to therapeutic agents.

The Committee recognises that principles developed in this Statement on PGx also apply to other therapeutic modalities such as ionising radiation and biologicals, and to variation in response in contexts such as nutrition, environmental exposures and toxicology.

Genomics, Society and Policy, Vol.3, No.1 (2007) ISSN: 1746-5354

GSP online Genomics, Society and Policy
2007, Vol.3, No.1, pp.44-47

Principles

Past HUGO Ethics Committee Statements have reflected a commitment to the view that the highest ethical priority in implementing genomic knowledge is that of saving life and reducing suffering, but the Committee considers it urgent that the ethical principles of solidarity and equity be given increased attention

- Solidarity: Because of shared vulnerabilities, people have common interests and moral responsibilities to each other. Willingness to share information and to participate in research is a praiseworthy contribution to society.
- Equity: To reduce health inequalities between different populations, and to work towards equal access to care is an important prerequisite for implementing genomic knowledge for the benefit of society.

while also reaffirming the following long-accepted ethical considerations:

- Respect for human rights
- Protection of confidentiality and privacy
- Avoiding harm
- Beneficence: there is an obligation to do good, and to maximise the possible benefits of genomics, which should be regarded as a global public good
- Autonomy: the freedom of persons to make decisions regarding their medical care is fundamental to the modern practice of medicine
- The authority of communities to participate in decisions that affect them
- The fundamental relationship between the quality of scientific research and its ethical acceptability

Recommendations

1. **Research priorities**

1.1. There needs to be a careful consideration of research priorities in PGx and translational research for each society and these should not merely be led by economic priorities determined by market forces.

1.2. The continuing creation of a sound scientific basis for PGx should be pursued, such as:

identification of genetic factors, including genes, haplotypes, SNPs and copy number variants (CNVs) with significant pharmacogenomic effects; identification of interactions with other genetic, environmental and social factors among drugs currently in common use in different countries; meta-analyses for consistency of data; ongoing development of methods for bio-marker analyses that are suitable for large scale studies.

1.3. The continuing creation of a sound corpus of research on ethical and legal issues, appropriate governance, social science and policy research, and methods of community and public dialogue and participation.

Genomics, Society and Policy, Vol.3, No.1 (2007) ISSN: 1746-5354

2. Governance of research

2.1. There needs to be appropriate governance at national and international levels of the collection, storage, utilization, sharing, and protection of data and biological specimens.
2.2. There is an urgent need for institutions and scientists to apply the principle of open access and sharing of data, consistent with the protection of personal privacy of the persons contributing to the database, in order to maximise benefit.
2.3. The establishment of necessary infrastructure, including international initiatives for sharing data between biobanks, should be supported.
2.4. There should be support and encouragement by research funding agencies to require and support the sharing of data.
2.5. HUGO should consider establishing mechanisms for on-going international co-ordination and evaluation of developments in PGx, with special reference to harmonising standards for reliability and replicability of PGx association studies.
2.6. While many institutions have developed governance mechanisms, such as ethics committees, attention should be paid to preparing them to assess appropriately the benefits and risks of PGx protocols.

3. Maximising the benefits of research

3.1. It should be recognised that PGx can be of benefit to communities as well as to individuals, even in the absence of optimal infrastructure/resources
3.2. In order to reduce health inequalities, there is a need both to develop new drugs for people with certain genetic variants, especially in the case of neglected and orphan diseases, and to consider the possibility of resuscitation of abandoned drugs for particular population groups.

4. Participation in research and social responsibility

4.1. All stakeholders in PGx research should exercise their ethical responsibilities in a spirit of equity and solidarity.
4.2. Voluntary participation of members of a community in PGx research provides an opportunity to actualise the principle of solidarity. Researchers have an obligation to engage the community while maintaining the highest standard of research conduct to earn the trust of the community.
4.3. The participation of all stakeholders, including the wider community, in such research requires public and professional dialogue and education in the science and ethics of PGx.

5. Clinical implementation

5.1. The principle of equity implies that therapies should be made equally available to those with equal needs. If this were not the case, the translation of genomic knowledge into clinical practice would aggravate disparities among people.

Genomics, Society and Policy, Vol.3, No.1 (2007) ISSN: 1746-5354

5.2. Barriers to translation of genomic knowledge into practice should be identified and addressed, e.g., by training of appropriate personnel; by working towards more equitable health care systems; etc.
5.3. An individual's serious side effect, or absence of response, may have important implications for drug treatment of blood relatives. The unit of care with respect to drug treatment may include the family as well as the individual, e.g., physicians should be alert to the implications for the relatives of a patient who suffers a serious adverse drug reaction, and should initiate genetic counselling.
5.4. Voluntary sharing of PGx information within families should be encouraged.

6. **Monitoring and quality control**

6.1. There needs to be agreement upon a standard set of data to be collected and a common format to facilitate international data sharing.
6.2. Systematic recording of clinical drug reactions to obtain data useful for PGx research should be encouraged.
6.3. Robust mechanisms of quality control need to be in place to minimise the possibility of error in individual PGx testing and maintenance of patient records.
6.4. Regulatory agencies should address the implications of PGx practice on a wide scale, including the possibilities of off-label use and liability issues.

7. **Education, training and awareness**

7.1. In the light of the rapid growth in PGx knowledge there is an urgent need to increase the level of awareness, education and training in the above issues for all stakeholders, including researchers, clinicians, policy makers, social scientists, patients and publics.

The HUGO Ethics Committee

Professor Kåre Berg (Vice Chair), Norway; **Professor Ruth F. Chadwick** (Chair), UK; **Professor Jose Maria Cantu**, Mexico; **Professor Abdallah S. Daar**, Oman and Canada; **Professor Kazuto Kato**, Japan; **Professor Darryl R. J. Macer**, New Zealand and Thailand; **Professor John J. Mulvihill**, USA; **Professor Thomas H. Murray**, USA; **Professor Carlos M. Romeo-Casabona**, Spain; **Professor Ishwar Verma** (Vice Chair), India; **Professor Zhai Xiaomei**, China

Montreal, May 2007

[1] The Universal Declaration on Bioethics and Human Rights was unanimously adopted on 19 October 2005 by all member countries of UNESCO. It follows two specific Declarations on ethics of applications of genomics that all member countries of UNESCO have agreed to; the Universal Declaration on the Human Genome and Human Rights (1997), also adopted unanimously by all UN members in 1998, and the International Declaration on Human Genetic Data (2003). In addition there have been many declarations and scholarly articles, and proclamations by agencies and committees of national governments, which provide insight into the theoretical issues and practical applications of ethics to issues raised by genomics, that we draw upon herein.

Genomics, Society and Policy, Vol.3, No.1 (2007) ISSN: 1746-5354

ゲノム医療の発展に向けた研究体制と市民との対話に関する考察
——全ゲノム関連解析とデータ共有を例にして

Importance of GWAS data sharing and public dialogue in the human genome research

高橋貴哲(写真)　加藤和人
Kitetsu Takahashi[1] and Kazuto Kato[1,2]
京都大学生命科学研究科生命文化学分野[1]，同人文科学研究所文化研究創成部門[2]

◎近年，疾患関連遺伝子を探索するための研究手法として全ゲノム関連解析(GWAS)が注目されているが，GWAS を行うには大規模なサンプル収集が必要になる．そこで英米の研究機関は各研究者の研究データを統合したデータベースを構築し，審査に合格した特定の研究者間でのみデータを共有する体制を整備することで，研究参加者のプライバシーに配慮しつつ，効率的に GWAS を推進する体制を整備している．その有効性から，日本でもデータ共有を行うことが望ましいと考えられるが，さまざまな社会的・倫理的問題により現段階では慎重に行わざるをえない状況がある．将来のゲノム医療の発展のためにはまだまだ基礎研究が必要であるため，これらの問題を解決し，GWAS に限らず，ヒトゲノム研究全体を強力に推進する体制を整備していく必要があるであろう．

Key word：全ゲノム関連解析(GWAS)，データ共有，dbGaP，WTCCC

2007 年，英米を中心に糖尿病や高血圧といった“ありふれた病気”に関連する遺伝子が多数報告され，ヒトゲノム研究は大きく前進した．その重要性はアメリカ『*Science*』誌が同年の科学的進歩トップ 10 の第 1 位として“ヒトゲノム多様性”研究の発展をあげたことからもうかがえるであろう．研究を一段と加速させるため，世界中の国々はヒトゲノム研究にいっそう力を入れている．

わが国でもこれまで以上にヒトゲノム研究を推進していく必要があるが，日本のヒトゲノム研究推進体制は研究の発展に合わせて柔軟に整備されてこなかったため，最新の研究手法に対応できない場合がある．したがって，世界の研究動向を調査し，日本の状況と比較検討することで，研究体制を刷新していく必要があるであろう．

本稿では世界の重要な研究動向の一例として，英米で先進的に行われている全ゲノム関連解析(genome wide association studies：GWAS)のデータ共有に関する話題を取り上げる．両国は研究参加者の多型解析データや表現型データなどをデータベース化して研究者間で共有しており(以下，GWAS データ共有)，効率的に GWAS を推進する体制を整備している．上記の成果もこの GWAS データ共有によるものが大きいが，日本では両国の GWAS データ共有体制についてあまり紹介されてこなかった．そこで，まず英米における GWAS の推進体制について簡単に紹介した後に，日本で GWAS データ共有を行ううえで課題となる事項について考察したい．

英米におけるGWASの推進

GWAS とは，アメリカ国立保健衛生研究所(National Institute of Health：NIH)の定義を参照すれば，特定の疾患と関連する遺伝的多様性を発見するために，多数の人びとのゲノム上のマーカーを解析する研究手法のことである[1]．GWAS 自体は図 1 に記したように，2005 年あたりからじわじわと成果を出している[2]．しかし，2007 年にイギリスウェルカム・トラストが運営する Wellcome Trust Case Control Consortium(WTCCC)な

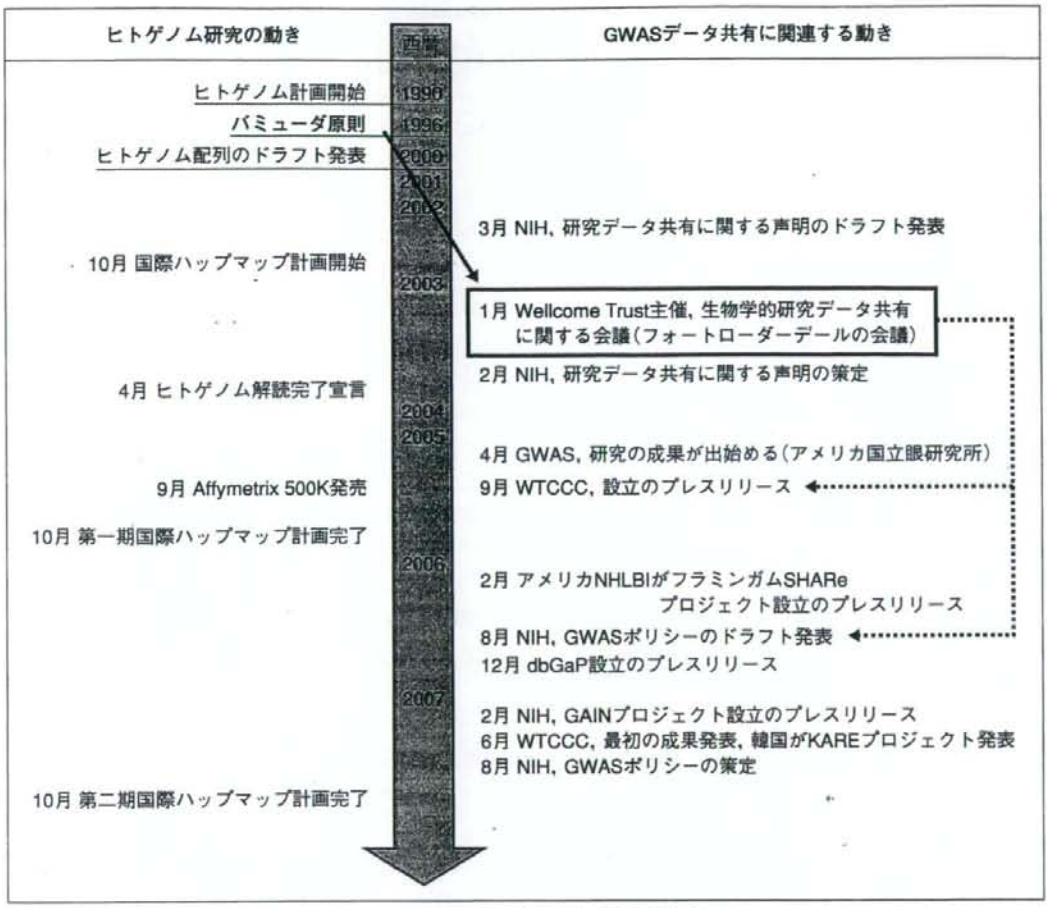

図 1　GWASデータ共有体制の整備過程

2005 年から GWAS 研究の成果が上がりはじめ，WTCCC や dbGaP，そして韓国の Korean Association REsource(KARE)プロジェクトといった GWAS データ共有を行う計画がつぎつぎと発表された．しかし，GWAS データ共有のように成果発表前の研究データを共有するという考え方の出発点は，1996 年のバミューダ原則にある．バミューダ原則は成果として発表する前のゲノム解読結果を研究者間で広く共有するための原則であり，その共有対象を生物学的データ全般に拡大するため，2002 年から 2003 年にかけ，NIH とウェルカム・トラスト主導による議論が行われた．これらの議論が 2005 年の WTCCC 設立や，2006 年に NIH が発表した GWAS データ共有の具体的方針について定めた GWAS ポリシーに影響を与えていると考えられる．

どの研究機関が行った大規模サンプル解析により，大量の疾患関連遺伝子が報告されたことによって，その注目度は飛躍的に上昇した．

現在，NIH と WTCCC は GWAS を推進するうえで強力なツールとなるデータ共有を行い，大規模サンプルを利用した疾患関連遺伝子探索や，さまざまな解析ツールの開発などを行っている．以下では両機関による GWAS データ共有システムの概観や，その共有体制が整備されていった過程について紹介する．

英米のGWASデータベース

アメリカには NIH の一部門である国立バイオテクノロジー情報センター(National Center for Biotechnology Information：NCBI)が運営する database of Genotype and Phenotype(dbGaP)[3]というデータベースがあり，ここには NIH の出資で行われた GWAS プロジェクト(かならずしもすべてではない)のデータが収められている．

一方，イギリスには WTCCC[4]が管理するデータベースがあり，約 20,000 人分の多型データや表現型データなどが収録されている．これはイギリス

の遺伝学者たちがもっていた研究データと古くからのコホート研究のデータを統合して構築されたものである．

GWASデータ共有システム

両データベースにはオープンアクセスデータとアクセス制限データが存在する．前者には多型頻度やデータの性質(サンプルの数や性別など)といった，個人の特定につながらない情報が収録され，後者には多型データや表現型のデータなどの，個人の特定につながりうるデータが収録されている．

後者にアクセスするための仕組みは両データベースで類似しており，アクセスを希望するすべての研究者はデータを管理する各機関に存在するデータアクセス委員会に自らの研究計画を申請し，審査を受ける必要がある．また，研究者のデータ運用状況を監視する委員会も存在し，年次報告書の提出などを通じて研究者の活動を監視している．それぞれのデータにはインフォームドコンセントの内容に対応したデータ使用基準が存在し，研究者はこれに準じてデータを運用しなければならない．

両データベースでは以上のような仕組みで研究参加者のプライバシーに配慮しつつ，研究者間でのデータ共有を実現している．

NIH，WTCCCにおけるGWASプロジェクトの特徴

NIHでは現在10を超えるGWASプロジェクトが運営されている．そのなかには Genetic Association Information Network(GAIN)[5]のように Pfizer や Abott などの企業と連携して行われるプロジェクトや，古くから行われてきたコホート研究とゲノム研究を組み合わせて環境因子と遺伝子の関係を探索するフラミンガム SHARe[6]プロジェクトなどが存在し，数万人規模の膨大なデータが dbGaP に収録されている．

これらのプロジェクトに参画する研究者はタイピングが完了した後，その成果を発表するより前に，速やかに dbGaP にデータを提出する(GAIN では7カ月以内が推奨されている)ことが求められている点が特徴的である．また，データ提供者の貢献に配慮するため，データ公開から最低9カ月間(プロジェクトによって異なる)は，データ提供者以外の研究者がそのデータを用いた成果発表(論文発表や学会発表など)を行ってはならないことになっている．

また，WTCCC も個別グループが解析した研究データを WTCCC の名義で論文発表した[4]ことから，NIH と同様に成果発表前の研究データを共有する姿勢をもっていると考えられる．

GWASデータ共有に至る背景

GWAS データ共有体制に特徴的な成果発表前のデータを共有する姿勢が形成された過程を追うと，その出発点は1996年のバミューダ原則にあると考えられる(図1)．同原則は解読したゲノム配列データを速やかに国際的データベースに登録して共有するように定めた原則であり，ヒトゲノム計画の推進に多大な貢献をした．この原則をもとに，共有の対象をゲノム配列に限らず，生物学的研究データ全般へと拡大する提案がなされ，2002年から2003年にかけてさまざまな議論がおもに英米主導で行われた．とくに2003年のフォートローダーデールの会議は世界中から識者が招待され，基盤的な生物学的研究データを成果発表前に共有することの有効性についてコンセンサスが形成された点で重要である．

一連の活動で発行された文書には個人の多型データや表現型データの共有という言葉自体は現れないが，WTCCC の発足が2005年の第一期国際ハップマップ計画完了，および解析に使用している Affymetrix 500K の発売と同時期にプレスリリースされていることから，これらの議論の過程で GWAS データの共有のような可能性を想定していたことがうかがえる．

また，成果発表前の生物学的研究データを共有することの利点が広く認識されたことにより，企業のゲノム配列の所有権に対する認識をも変化させた点は重要であった．バミューダ原則の前年には，ミリアド社が *BRCA1*，*BRCA2* 遺伝子に関する検査に対して特許を出願したが，現在では GAIN などの公的データベースを整備するプログ

ラムに種々の企業が出資しているという事実が，それを証明しているであろう．基礎研究に投資をし，研究全体の質が向上することが，結局は自らのビジネスを拡大することにつながるという考え方は，日本の企業も見習うべき姿勢であろう．

日本のGWASデータ共有推進における課題

2007年は疾患に関連する遺伝子が多数報告され，ヒトゲノム研究の歴史に刻まれた年であった．その波紋は世界中に広がり，すでに韓国はGWASデータ共有を行うプロジェクトとしてKorean Association REsource(KARE)プロジェクトを2007年9月に発表している(図1)．日本においても限定的に特定領域研究“ゲノム4領域”内で行うことが検討されているが，将来的には日本全体でGWASデータ共有を行うことが望まれる．ここでは日本でGWASデータ共有を推進するために必要となる事項について考察したい．

まず，日本のヒトゲノム研究倫理指針である，「ヒトゲノム・遺伝子解析研究に関する倫理指針」をみると，GWASで用いられる多型データや表現型データなどの共有に関する記述がなく，この指針をGWASデータ共有の規範として用いることはできないことがわかる．これらのデータは単体で個人を特定することができないので，個人情報とは区別して扱う必要があり，広くGWASデータ共有を行うためには，この点に配慮した指針を整備する必要があるであろう．

このように，指針が新しい研究手法に対応できない状況を招いている原因として，指針にさまざまな研究者の意見を反映できていないことがあげられる．たとえば，NIHでは種々のポリシーを策定する際にドラフトを最初に発表してパブリックコメントを広く募集し，それらの意見をもとに正式なポリシーを策定している．日本においても指針策定の初期段階からさまざまな研究者に広く意見を求め，その意見を柔軟に反映していくような体制を整備する必要があるであろう．

また，GWASデータ共有のように参加者のプライバシーを脅かす可能性がある研究は，社会との信頼関係なくして行うことはできない．したがって，GWASデータ共有を本格的に実施するためには，それと並行して社会との信頼関係を醸成するための活動を行うことも重要である．

たとえば，当研究室で運営している“ゲノムひろば”というイベントではゲノム研究に携わる研究者が市民からの疑問や質問に答え，研究者と直接交流する場を提供している[7]．多くの人が最先端のゲノム研究に触れることを目的に来場しており，このような場でGWASデータ共有に関する議論を行うことは，限定的ではあるが，市民のGWASデータ共有に関する関心を高め，社会との信頼関係を構築する助けになる可能性がある．

欧米諸国に比べ日本ではこのようなコミュニケーション活動の歴史が浅く，われわれもいまだ試行錯誤の段階である．しかし，社会との信頼関係が科学の発展を左右することはイギリスのBSE問題からも明らかであり，GWASデータ共有が原因となって日本の科学全体の信頼が損なわれるような事態を避けるためには，研究者間での議論や研究者と社会間での持続的なコミュニケーション活動などが重要になるであろう．

以上，GWASデータ共有を推進するために必要な事項として，指針整備の必要性とコミュニケーション活動の活性化という2つの点を指摘したが，まだまだ課題は山積しており，さらなる考察が必要であろう．ヒトゲノム研究はたしかに大きく飛躍したが，ゲノム医療の実現にはさらなる基礎研究が必要である．したがって，日本のヒトゲノム研究全体の水準を上げる必要があり，そのためにはさまざまな社会的・倫理的課題を，研究者と社会が一丸となって解決していくことができる体制を国としてつくっていくことが重要ではなかろうか．

文献/URL

1) http://www.genome.gov/20019523
2) Couzin, J. and Kaiser, J.：*Science*, **316**：820-822, 2007.
3) Mailman, M. D. et al.：*Nat. Genet.*, **39**：1181-1186, 2007.
4) Wellcome Trust Case Control Consortium：*Nature*, **447**：661-678, 2007.
5) Manolio, T. A. et al.：*Nat. Genet.*, **39**：1045-1051, 2007.
6) http:/ /www.ncbi.nlm.NIH.gov/projects/gap/cgi-bin/study.cgi?study_id＝phs000007.v1.p1
7) 白井哲哉，加藤和人：蛋白質・核酸・酵素，**53**：274-280，2008.