nature genetics

Common deletion polymorphisms in the human genome

Steven A McCarroll[1–3], Tracy N Hadnott[1], George H Perry[4], Pardis C Sabeti[3], Michael C Zody[3], Jeffrey C Barrett[3], Stephanie Dallaire[4], Stacey B Gabriel[3], Charles Lee[4,5], Mark J Daly[2,3,5,6] & David M Altshuler[1–3,5,6], for the International HapMap Consortium

The locations and properties of common deletion variants in the human genome are largely unknown. We describe a systematic method for using dense SNP genotype data to discover deletions and its application to data from the International HapMap Consortium to characterize and catalogue segregating deletion variants across the human genome. We identified 541 deletion variants (94% novel) ranging from 1 kb to 745 kb in size; 278 of these variants were observed in multiple, unrelated individuals, 120 in the homozygous state. The coding exons of ten expressed genes were found to be commonly deleted, including multiple genes with roles in sex steroid metabolism, olfaction and drug response. These common deletion polymorphisms typically represent ancestral mutations that are in linkage disequilibrium with nearby SNPs, meaning that their association to disease can often be evaluated in the course of SNP-based whole-genome association studies.

Recently, comparative hybridization of genomic DNA from multiple individuals revealed extensive copy number variation across the human genome[1–3]. These methods generally detected changes >100 kb in size; it is not yet known which of those regions involve losses of genetic information and which involve extra copies (**Supplementary Note** online). Comparison of fosmid end sequences from a single individual with the finished human genome sequence identified 102 regions containing intermediate-sized (>8 kb) deletion variants[4]. It is not known which of these are common or what additional common deletions are present in the human population beyond the single individual studied. As dense SNP genotype data (greater than 1 per 5 kb) is already available in 269 people from the International HapMap Project[5] and will soon be generated in many thousands of people from clinical cohorts, it seemed valuable to develop a method for discovering deletions directly from SNP genotype data.

We developed an approach for discovering deletions from SNP genotypes, based on the observation that a segregating deletion can leave 'footprints' in SNP genotype data, including apparent deviations

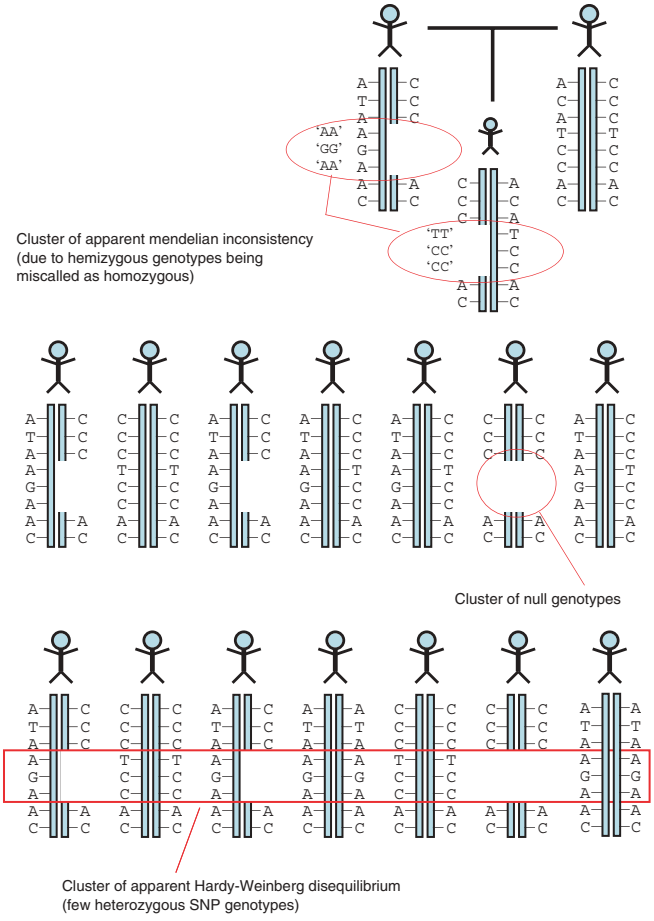

Figure 1 Using SNP genotypes to discover segregating deletion variants. Segregating deletions leave a 'footprint' in SNP genotype data by causing physically clustered patterns of null genotypes, apparent mendelian inconsistencies and apparent Hardy-Weinberg disequilibrium.

[1]Department of Molecular Biology and [2]Center for Human Genetic Research, Massachusetts General Hospital, 55 Fruit Street, Boston, Massachusetts 02114, USA. [3]Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02141, USA. [4]Department of Pathology, Brigham and Women's Hospital, 75 Francis Street, Boston, Massachusetts 02115, USA. [5]Harvard Medical School, Boston, Massachusetts 02115, USA. [6]Department of Medicine, Massachusetts General Hospital, Simches Research Center, 185 Cambridge St., Boston, Massachusetts 02114, USA. Correspondence and requests for materials should be addressed to D.M.A. (altshuler@molbio.mgh.harvard.edu).

Received 12 October; accepted 31 October; published online 4 December 2005; doi:10.1038/ng1696

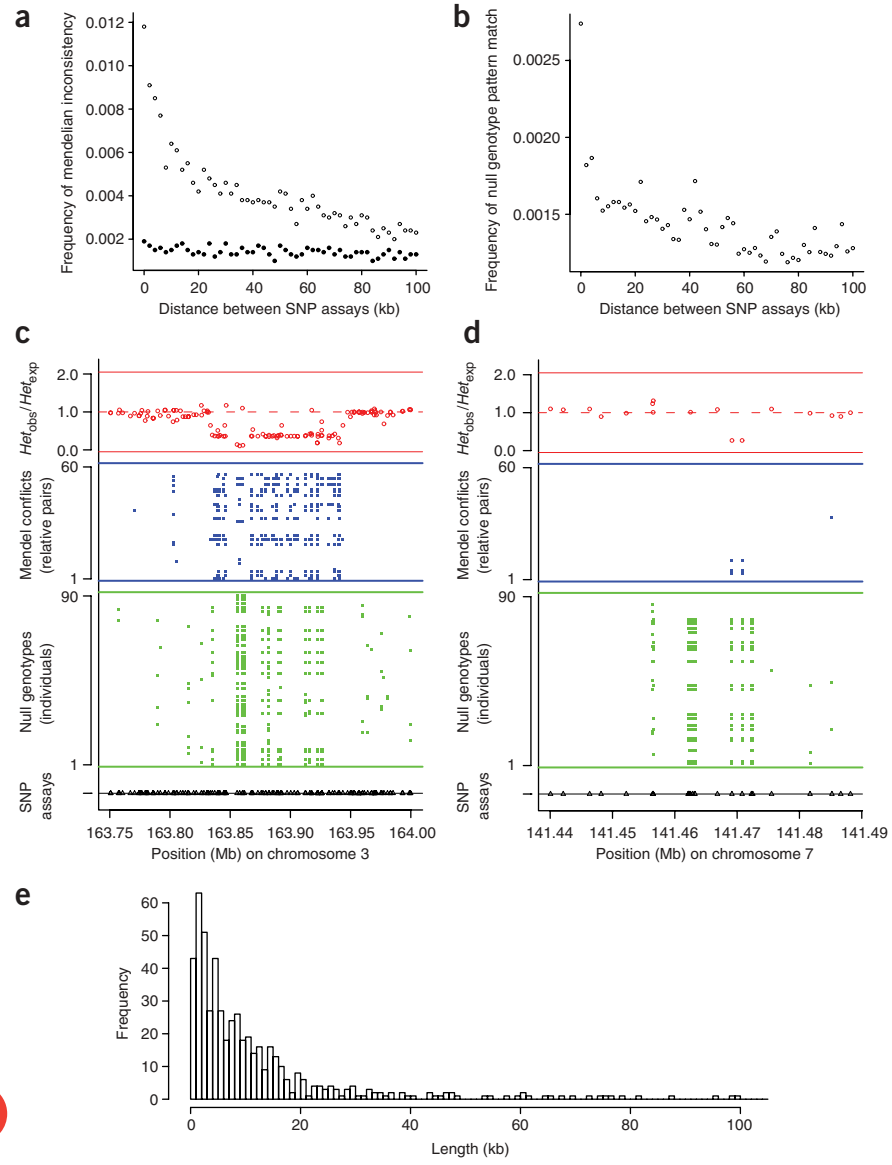

Figure 2 Spatially patterned aberrations in SNP genotypes. (**a**) Clustering of mendelian inconsistencies. Mendelian-inconsistent SNP genotypes ('Mendel failures') appear more frequently at genomic locations close to other Mendel failures when those earlier failures are observed in the same individuals (open circles) but not when they are observed in other individuals (filled circles). The figure is based on pairs of HapMap SNP assays that were typed using different genotyping technologies. (**b**) Clustering of population patterns of null genotypes. (**c**,**d**) Spatially patterned failure of SNP genotype assays at the sites of segregating deletions. Tracks show, for each SNP assay (triangles), the pattern of null genotypes across 90 individuals (green), the pattern of Mendel failure across 60 pairs of relatives (blue) and the ratio of observed to expected heterozygosity (red). Physically clustered sets of similarly aberrant genotypes identify a common, 85-kb segregating deletion on chromosome 3q and a common, 10-kb segregating deletion on chromosome 7q, both in a sample of 30 trios with European ancestry. Additional figures showing these deletion variants in all HapMap population samples are in **Supplementary Figure 2**. (**e**) Size distribution of deletion variants identified from regional patterns of aberrant SNP genotypes. A few deletions larger than 100 kb (up to 745 kb) were also observed.

from mendelian inheritance[6,7], apparent deviations from Hardy-Weinberg equilibrium and null genotypes (**Fig. 1**). Using these clues to discover true variants is challenging, however, because the vast majority of such observations represent technical artifacts and genotyping errors[5]. In fact, because such deviations are thought usually to be errors, 'failed' assays of this sort are routinely discarded from medical genetic studies.

To determine whether a subset of 'failed' SNP genotyping assays in the HapMap data might reflect structural variation, we asked whether such failures are physically clustered in a manner that is specific to individuals. Consistent with this hypothesis, the rate of mendelian-inconsistent genotypes was elevated near other mendelian-inconsistent genotypes in the same individual (regardless of whether the same genotyping platform was used for both assays) but was unrelated to mendelian inconsistencies in other individuals (**Fig. 2a**). A similar relationship was observed for null genotypes (**Fig. 2b** and **Supplementary Fig. 1** online). Thus, such clustering is a property of individual variation in local sequence, rather than the local sequence *per se*.

Encouraged, we developed a systematic approach for identifying deletions from patterns of Mendel failures, null genotypes and Hardy-Weinberg disequilibrium in dense genotype data (**Fig. 1** and Methods). To distinguish aberrant genotypes due to structural variants from sporadic errors, we looked for regions of the genome in which the same failure profile appeared repeatedly at nearby markers (**Fig. 2c**,**d** and **Supplementary Fig. 2** online) in a manner that was statistically unexpected based on chance. A set of statistical thresholds was tailored to each mode of failure, genotyping center and genotyping platform used in the project. Specifically, we calculated the binomial probability of observing each such pattern n times in m markers (based on the empirically observed rate of each error type in each center and platform; see Methods and **Supplementary Fig. 1**). Although these specific thresholds are tailored to the HapMap genotyping centers, the same procedure can readily be applied to dense SNP data from any platform or study.

Applying this method to data from Phase I of the International HapMap Project[5]—1.3 million SNPs, typed in 269 individuals of European, Yoruba, and Chinese and Japanese ancestry—resulted in a map of 541 candidate deletion variants ranging from 1–745 kb (median, 7.0 kb) in size (**Fig. 2e** and **Supplementary Table 1** online). Some 278 of these loci (1–230 kb, median 5.5 kb) were identified in multiple unrelated individuals, suggesting that they segregate at an appreciable frequency in the human population; 120 were observed as homozygous deletions. (Observations within the immunoglobulin κ, λ and heavy chain loci were attributed to somatic rearrangements in the lymphoblastoid cell lines used to derive DNA for analysis and were not included in subsequent analyses.)

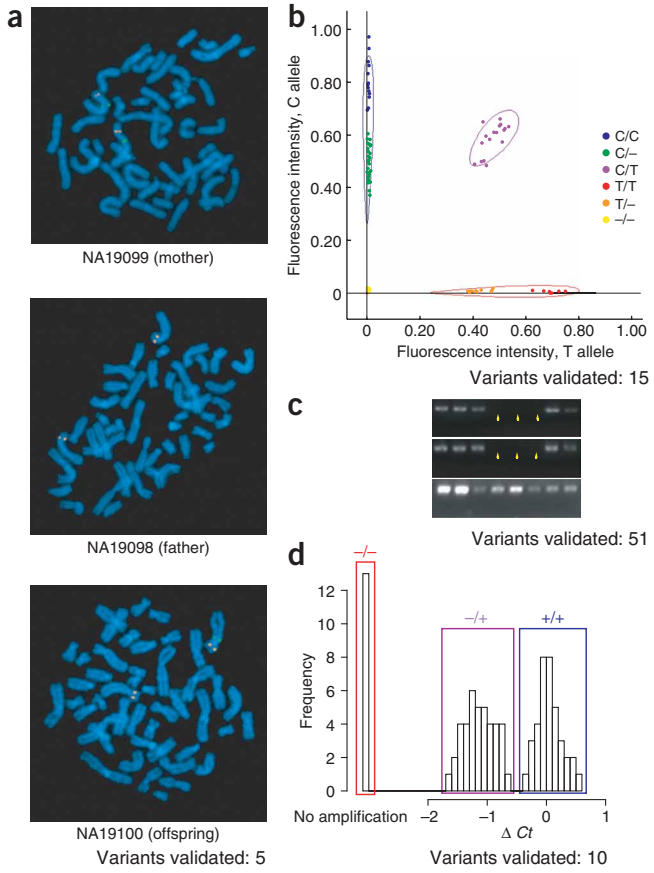

Figure 3 Experimental validation of segregating deletions. (**a**) Fluorescence *in situ* hybridization confirmed the presence and mendelian inheritance of an 85-kb deletion on chromosome 4q13.2 at 70.4 Mb. FISH results for four additional loci are in **Supplementary Figure 4**. (**b**) Two-color, allele-specific fluorescence intensity measurements (Illumina BeadArray) for 90 individuals at a site that shows both single-nucleotide and deletion polymorphism. Intensity measurements fall into six genotype clusters (versus the two to three clusters typical of SNPs). The extra clusters were subsequently shown (by quantitative PCR) to correspond to individuals with hemizygous and homozygous deletions of the locus (green, orange and yellow circles). (**c**) Confirmation by PCR of a predicted population pattern of homozygous deletion of sequence on chromosome 8p23.3 at 2.4 Mb. Yellow arrows indicate the individuals predicted (from having multiple null genotypes at the locus) to carry homozygous deletions. Bottom panel shows results for a control locus outside of the deletion. (**d**) Measurements of copy number obtained by quantitative PCR (shown here for a deletion on chromosome 4 at 70.5 Mb) fall into three discrete clusters, allowing accurate inference of the deletion genotype in each individual.

Nearly all of these deletion variants were novel (507/541, or 94%). This could be because in many cases, available SNP genotype data provides finer resolution than previous assays, and most deletion variants we identified were small (**Fig. 2e**): 55% were less than 8 kb in size. Moreover, only 11 deletions mapped to the locations of large 'copy number polymorphisms', and only 28 of 541 had been identified in a previous analysis of 589,275 fosmid end reads[4] in a single person (**Supplementary Note**).

Of course, a high rate of new discoveries (94% novel) could be simply explained if the method has a very high rate of false positives. This is a valid concern, given the origin of our discoveries in data that fail typical quality control standards and the statistical nature of the inference. Thus, we set out to validate a subset of the candidates using four methods: fluorescent *in situ* hybridization (FISH), two-color fluorescence intensity measurements, PCR amplification and quantitative PCR.

We performed fluorescent *in situ* hybridization (FISH) for five candidate deletions large enough to span available FISH probes. In all five cases, FISH assays confirmed the deletions in the predicted individuals (**Fig. 3a** and **Supplementary Fig. 3** online).

We examined two-color allele-specific fluorescence data from SNP genotyping assays from a subset of data available at the Broad Institute of Harvard and the Massachusetts Institute of Technology by looking for a reduction in fluorescence intensity in individuals predicted to carry a deletion. At most SNPs in the genome, fluorescence intensity measurements cluster into two or three discrete groups corresponding to homozygous and heterozygous genotypes. At 15 of 17 candidate deletion loci (for which allele-specific intensity values were available for three or more SNPs), fluorescence intensity data for one or more SNPs clustered into additional groups that corresponded to the predicted deletion genotypes (**Fig. 3b**).

We used PCR amplification to query 60 loci for which the pattern of genotypes suggested multiple individuals with homozygous deletions. Variants were considered confirmed if the pattern of amplification success and failure matched prediction across a set of 12–24 individuals (including at least two individuals with each predicted result). We confirmed 51 of 60 candidate variants by this criterion (**Fig. 3c**).

We performed quantitative PCR in all 269 HapMap DNA samples for 11 candidate deletions that overlapped the coding exons of genes (described below) and that were discovered in many individuals: at 10/11 loci, we observed three discrete clusters, identifying individuals with zero, one and two gene copies (**Fig. 3d**).

Table 1 Common gene deletion polymorphisms

Gene	Function	Frequency of deletion allele			Tagging SNP r^2		
		CEU (%)	JCH (%)	YRI (%)	CEU	JCH	YRI
UGT2B17	Sex steroid hormone metabolism	30	84	22	1.00	0.96	0.63
UGT2B28	Sex steroid hormone metabolism	13	15	35	1.00	1.00	0.90
TRY6	Proteolysis	41	74	12	1.00	1.00	1.00
LCE3C	Epidermal cornified envelope	56	69	30	0.93	0.93	1.00
GSTM1	Detoxification, drug metabolism	77	70	48	0.76	1.00	0.97
GSTT1	Detoxification, drug metabolism	39	65	57	1.00	1.00	0.48
CYP2A6	Detoxification, drug metabolism	0	17	2	–	0.80	–
PRB1	Secreted salivary proteoglycan	8	3	16	–	–	0.28
OR51A2	Olfactory receptor	49	40	28	0.38	0.47	0.20
OR4F5	Olfactory receptor	3	0	17	–	–	–

Table shows confirmed common human deletion polymorphisms that remove the coding exons of genes. Each gene deletion appears to be a null. All 269 individuals from the three HapMap population samples were typed for each gene deletion variant using quantitative PCR. To assess linkage disequilibrium, the distribution of each variant across each population sample was compared to the distribution of all SNP alleles within 200 kb of the deletion. The value shown is the correlation to the highest-scoring tagging SNP having LOD >3. (**Fig. 5a** and **Supplementary Fig. 5** show the broader patterns of linkage disequilibrium throughout each of these regions. The gene deletion genotypes for each HapMap individual are available in **Supplementary Table 3**.) Populations: CEU (European ancestry), JCH (Japanese and Chinese ancestry), YRI (Yoruba ancestry).

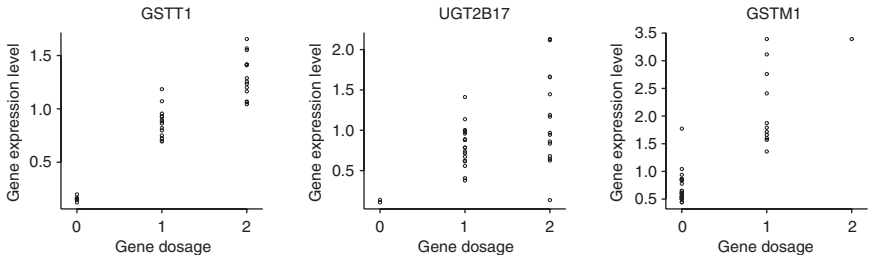

Figure 4 Variation in gene expression due to gene copy number variation. The measured expression level of each gene[13] in lymphoblastoid cell lines from individuals whom we determined by quantitative PCR to have zero, one and two gene copies.

In total, 90 predicted deletion variants were validated by one or more of these approaches or comparison with earlier studies (**Supplementary Table 2** online), showing that a high false positive rate is not the explanation for the large number of new discoveries in the analysis. There were ample SNP assays at the sites of most common, large–copy number variants identified in two earlier, hybridization-based studies[1,2]; thus, the observed lack of overlap with these studies indicates that most of the common, large (>100 kb) copy number variants identified by those studies involve extra copies, rather than losses, of genetic material. (The fosmid-based study examined only one individual, and we queried 269, which is likely to explain the lack of overlap between these two methods.)

The high-resolution mapping of this collection of deletion variants allowed us to ask whether they are flanked by low-copy repeats in the same orientation, potentially extending an earlier finding that low-copy repeats are enriched within BAC-sized regions that contain copy number variants[3]. Supporting the model that non-allelic recombination between homologous sequences is a frequent cause of structural mutation[8], we found that 24 of the deleted sequences (versus 3.4 expected by chance) were flanked by homologous pairs of low-copy repeats in the same orientation (**Supplementary Fig. 4** online).

We also asked whether any of the deletions removed the coding exons of genes. Some 266 genes were affected by deletion variants in at least one individual. Notably, ten genes were validated as being deleted at an appreciable frequency and were observed as homozygous nulls (**Table 1**). Two of these commonly deleted genes are involved in the metabolism of sex steroid hormones (*UGT2B28* and *UGT2B17*). Common deletions also removed two genes encoding olfactory receptors (*OR51A2* and *OR4F5*) and three genes (*CYP2A6*, *GSTT1* and *GSTM1*) with roles in detoxification and drug metabolism. Three of these common gene deletions were previously known[9–12].

We used quantitative PCR assays (**Fig. 3d**) to accurately genotype each individual as carrying zero, one or two gene copies (**Table 1** and **Supplementary Table 3**). We observed mendelian inheritance for the ten common gene deletions in all 60 trios, Hardy-Weinberg equilibrium in all four populations surveyed, and transmission rates close to 50%, suggesting that each behaves as a stable, heritable genetic polymorphism (**Supplementary Table 3**). We observed the gene deletion polymorphisms in individuals of European, Yoruba, Chinese and Japanese ancestry, although the frequency of each deletion polymorphism varied from population to population (**Table 1**), as is commonly seen with SNPs.

Exonic deletions should affect the expression of encoded genes. Three genes observed as common deletions (*GSTT1*, *UGT2B17* and *GSTM1*) are expressed at appreciable levels in lymphoblastoid cell lines previously used to survey gene expression variation[13,14]. We compared published expression measurements from these cell lines[13] with deletion genotypes obtained from the same individuals. Variation in gene dosage explained 88%, 26% and 75% of the observed variation in expression of the three genes, respectively (**Fig. 4**); individuals with one copy showed 30%, 35% and 38% less expression, respectively, than individuals with two gene copies.

Assessing the medical consequences of common deletion polymorphisms requires determining their frequencies in patient cohorts. Common SNPs are being tested for a role in disease using genome-wide association studies, which rely on correlations among nearby variants (linkage disequilibrium, LD) to test variants both known and as-yet undiscovered[5,15]. Thus far, it has not been evaluated whether such LD-based approaches will also suffice to detect the influence of common deletion polymorphisms or whether dedicated technology (such as comparative genomic hybridization) will be required. The answer depends on the population genetic histories of common deletion polymorphisms: if common deletions are typically due to recurrent mutation, then deletions will exist on many haplotypes, and LD will be of little value[3]. To the extent that deletions result from unique ancestral mutational events, they will often be in LD with nearby SNPs, and ancestral SNP haplotypes can serve as proxy in disease studies.

We observed strong LD between SNPs from HapMap and validated deletions. For example, nine of the ten gene deletions (for which we had designed accurate quantitative PCR genotyping assays) showed significant LD with nearby SNPs, and six of the ten had a perfect SNP proxy ($r^2 = 1$) in one or more populations (**Table 1**, **Fig. 5a** and **Supplementary Fig. 5** online). In each case, the deletion was associated to the same SNP allele(s) in each population (**Fig. 5b** and **Supplementary Table 4** online), indicating an ancestral mutation that occurred before humans migrated from Africa to Europe and Asia. In the larger collection of 51 deletion variants validated by PCR, we found elevated homozygosity at SNPs flanking the homozygous deletions (relative to randomly selected individuals at the same loci), indicating that the deletion alleles travel on specific SNP haplotypes (**Fig. 5c**). On average, the rate of decay of haplotype homozygosity around deletion alleles was similar to that observed for a frequency- and population-matched set of SNP alleles (**Fig. 5c**).

In summary, we developed a method to systematically screen for deletion variants using SNP genotype data and, by applying it to data from the HapMap Project, generated an initial high-resolution map of deletion variants in the human genome. A related method and deletion map have been independently developed by Conrad and colleagues (reported in this issue)[16]. These methods are not limited to the HapMap data: tailored to the error properties of other data sets, such an approach can be used to screen for ancestral and *de novo* deletion variants in any dense set of SNP genotypes. This should be of increasing value given that such an approach has no cost, and that whole-genome SNP association studies are being undertaken in many clinical cohorts. Existing SNP genotype data from medical genetic studies could also be reanalyzed by such an approach.

Of more general importance, the finding of LD between SNPs and deletions indicates the utility of a single database that integrates data on SNP genotypes and structural polymorphisms. These results for deletion polymorphisms are likely to be generalizable, as Hinds *et al.*, in an independent study in this issue, have also reached the conclusion that LD between SNPs and deletions is similar to LD between SNPs

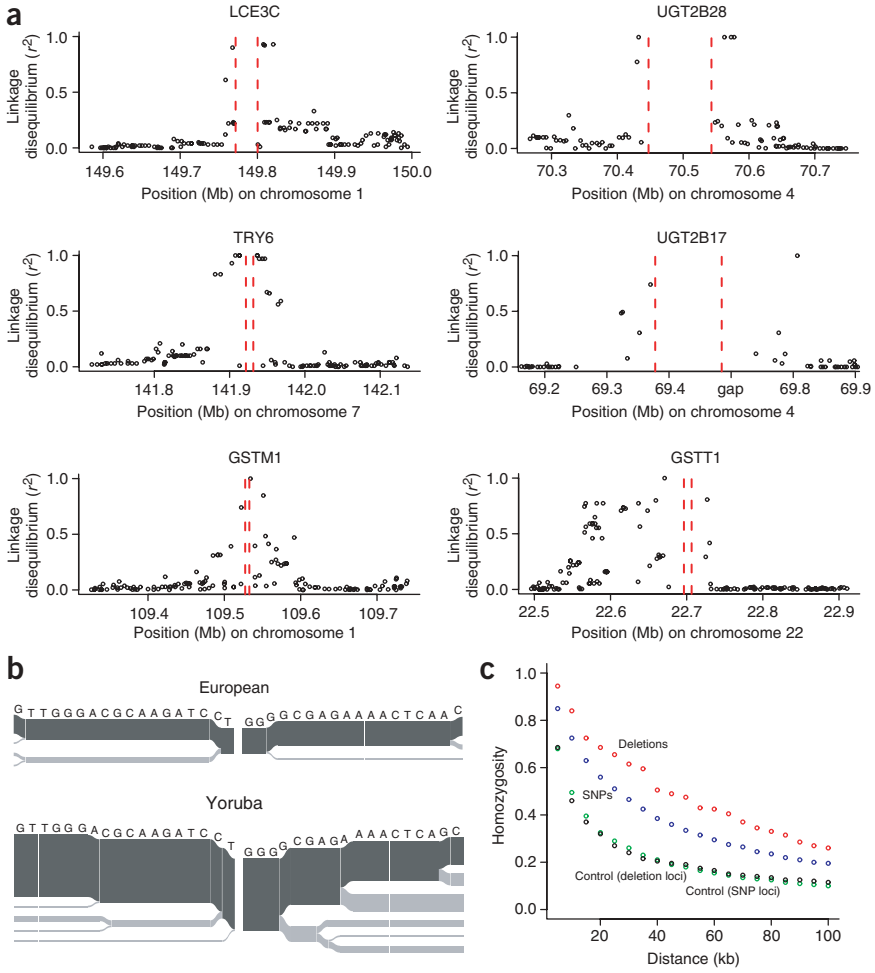

Figure 5 LD of deletion polymorphisms with SNPs. (**a**) Linkage disequilibrium (r^2) of gene deletion polymorphisms with SNPs. For each gene deletion, strong linkage disequilibrium is observed with SNPs to the left and right of the deletion breakpoints (red dashed lines). Separate LD plots for these gene deletion polymorphisms in European, Yoruba, and Chinese and Japanese population samples are in **Supplementary Figure 5**. (**b**) Residence of the UGT2B28 deletion allele on the same core haplotype in European and Yoruba populations. Letters indicate the consensus haplotype in each population. Image was generated using the Bifurcator program[19]. (**c**) Haplotype homozygosity across flanking SNPs in individuals homozygous for 51 experimentally validated deletions (red), in randomly selected control individuals at the same deletion loci (black), in individuals homozygous for a frequency- and population-matched set of SNP variants (blue) and in randomly selected control individuals at these SNP loci (green).

and other SNPs[17]. In addition to illuminating the origin and evolutionary history of structural variants, the identification of informative 'tag' SNPs for structural variants will make it possible (for at least a subset of structural variants) to query both SNP and deletion variants using LD-based whole-genome association studies. It remains an important and open question whether these conclusions about deletion variants can be extended to gains of genetic material and to highly duplicated regions of the human genome in which SNP genotype data are harder to collect.

METHODS

Genotype data. We used the genotypes from HapMap release 16 provided without quality control filtering. These consisted of separate genotype files for four population samples: 90 CEPH individuals (30 trios) of European ancestry; 90 individuals (30 trios) of Yoruba ancestry sampled in Ibadan, Nigeria; 45 unrelated individuals of Han Chinese ancestry sampled in Beijing and 45 unrelated individuals of Japanese ancestry sampled in Tokyo. The population samples are described in detail elsewhere[5]. We combined the data from the Chinese and Japanese population samples and thereafter treated the data set as three analysis panels of 90 individuals each.

The HapMap data (and all physical coordinates in this paper) use the hg16 (build 34) assembly of the Human Genome Project.

HapMap genotyping platforms. The HapMap genotypes were generated at ten genotyping centers, using seven different genotyping technologies. The background rates of Mendel failure, null genotypes and Hardy-Weinberg disequilibrium differed from technology to technology and even for the same technology when used by different centers. Furthermore, we observed sample-by-batch interactions, in which particular samples show elevated rates of null genotypes or Mendel failures in particular experimental batches. Thus, we needed to calibrate all statistical thresholds used for each of the 12 platforms (combinations of center and technology) used on the project, as discussed below.

We also observed that deletions left different signatures in SNP genotypes depending on the genotyping technology used. The most common signature (**Fig. 1**) was that (i) homozygous deletions resulted in null SNP genotypes and (ii) hemizygosity resulted in SNP genotypes being miscalled as homozygous, often resulting in mendelian inconsistency. On some genotyping platforms (particularly the Third Wave platform used by the RIKEN center), hemizygosity instead resulted in null genotypes (perhaps because samples with aberrant intensity values were not assigned a genotype or because mendelian inconsistencies were excluded during curation). On other platforms (particularly the Illumina platform used at the Sanger Center), samples with homozygous deletions were often given an (incorrect) SNP genotype call that resulted in mendelian inconsistency with relatives' genotypes.

Because we sought to identify independent but accordant multiassay patterns in genotype data, genotyping assays that were based on a shared amplification, labeling or restriction site were not considered independent. Thus we excluded all the Perlegen assays, because the use of 10-kb amplicons on that platform potentially caused long-range patterns of aberrant genotypes wherever an undiscovered SNP altered either primer binding site. We also excluded data from any experiments whose batch structure corresponded to physical regions of the genome, because this design potentially allowed batch-specific experimental artifacts to appear as regional patterns in the data.

Definitions of 'failure profiles' for Mendel conflicts and null genotypes. We separately assayed for clustering of aberrant genotype patterns in each of the three population samples.

Null genotypes. For each genotype assay and population sample, we defined the 'null genotype pattern' of that assay as the binary vector (length 90) of null genotype calls across the 90 individuals in that population sample. For each such pattern that was observed, we considered each pattern together with its close neighbors ($r^2 > 0.8$) in pattern space; this fuzzy clustering was necessary because genotype assays do not always obtain 100% complete calls, even in diploid samples.

Mendel failures. For each genotype assay and population sample (CEPH and Yoruba samples only), we defined the 'Mendel failure pattern' of that assay as the binary vector (length 60) of Mendel consistency status across the 60 parent-offspring pairs in that population sample. For each such pattern that was observed, we considered each pattern together with its close neighbors ($r^2 > 0.8$) in pattern space; this fuzzy clustering was used because the same deletion segregating in a population can give rise to nonidentical patterns of Mendel failure at different SNPs, owing to the fact that the other (nondeletion) SNP haplotypes (whose conflicts result in the mendelian inconsistencies) may not disagree at all SNPs.

Null genotypes plus Mendel failures. For each genotype assay and population sample (CEPH and Yoruba samples only), we defined the 'combined failure pattern' as the binary vector (length 90), with one element for each individual in a population, equal to 1 if that individual either had a null genotype or was involved in a Mendel failure (and equal to 0 otherwise). We treated these vectors as above. Although this third type of 'failure profile' was highly redundant with the first two and resulted in only two additional discoveries, it was useful for discovering deletions that crossed SNP assays performed on distinct genotyping technologies, as we occasionally observed that the same deletion polymorphism gave rise to Mendel failures on one platform and null genotypes on another (see note on genotyping platforms, above).

Assessment of clustering of 'failure profiles'. For both Mendel failure profiles and null genotype profiles, we observed that highly similar ($r^2 > 0.8$) profiles tended to be physically clustered in the genome—more specifically, that the probability of observing a 'match' to any particular profile was a decreasing function of physical distance from that profile, even when we considered only pairs of SNP assays that were typed using different technology platforms (**Supplementary Figure 1**).

The Phase I HapMap data was produced by ten different genotyping centers, with each chromosome arm primarily genotyped by one particular center[5]. Approximately 120,000 SNP assays were performed by centers outside of their primary regions or on genome-wide platforms such as Affymetrix 100K SNP arrays, allowing cross-platform analyses like those in **Figure 2a** and **Supplementary Figure 1**. However, because the overwhelming majority of assays in any particular region were performed at a single genotyping center, any effort to identify local multimarker features in the HapMap data must of necessity compare many SNP assays that were produced by the same center and genotyping technology. It was therefore critical to control for center- and platform-specific patterns in the data.

We therefore analyzed the data from each genotyping center separately. For each genotyping center, we first ordered all of the SNP assays from that center by genomic position. For each pattern (clustered set of highly similar profiles) that was observed multiple times, we determined that pattern's background frequency at that center. We then analyzed the physical distribution of all observations of that pattern relative to all of the SNP assays from that center (ordered by genomic position). A list of 'candidate clusters' was determined by considering every consecutive pair and consecutive trio of observations of that pattern, together with any other intervening SNP assays from that center. To assess the tightness of each such candidate cluster, a 'clustering *P* value' was calculated to assess the probability of observing a cluster at least as tight (in consecutive-assay space) as that cluster, given (i) the background frequency of the pattern, (ii) the number of SNP assays spanned by the cluster and (iii) the total number of SNP assays performed by that center. The distribution of these *P* values is shown in **Supplementary Figure 1**: it shows a generally uniform distribution of *P* values from 0 to 1, but with an excess of very low *P* values. The region of excess low *P* values can be thought to identify a set of candidate clusters in which the alternate hypothesis (nonrandom degree of clustering) is likely to be true; this region is separated by a 'knee' from the rest of the distribution, which is organized as a generally uniform distribution (**Supplementary Fig. 1**). We chose a significance threshold for promoting potential clusters, with the goal of capturing as many true discoveries as possible while maintaining a false discovery rate of no greater than 10% of all discoveries. This required selecting a significance threshold somewhat to the left of the 'knee' in the distribution.

We clustered all overlapping genomic segments that were identified by this analysis into 702 genomic loci.

We were concerned that multiplexed batches of SNP assays that were performed together could also give rise to potential patterns in the data, which (if distributed non-randomly in genomic space with respect to that center's other SNP assays) could give rise to potential batch artifacts. We therefore excluded those clusters that consisted entirely of SNP assays from the same experimental batch. (Batch information was obtained from the International HapMap Consortium.) This resulted in a set of 541 predictions.

Use of Hardy-Weinberg disequilibrium. We observed that a deletion tended to reduce the ratio of observed heterozygosity to expected heterozygosity (het_{obs}/het_{exp}) by a uniform amount (**Fig. 2c**), this amount being determined by the population frequency of the deletion haplotype. We thus looked for genomic regions in which het_{obs}/het_{exp} consistently fell below some cutoff (we used cutoffs of 0.7 and 0.4). We included only those assays with a minor allele frequency greater than 10%. For each genotyping platform, we determined the background frequency of assays for which het_{obs}/het_{exp} was less than the cutoff, and we used this frequency to determine statistical thresholds for clustering as described above.

Wherever the resulting genomic segments overlapped with clusters of Mendel failure or null genotypes as discovered above, we clustered those segments together. Because heterozygosity can show regional correlations owing to haplotype structure, selection and potentially duplicated sequence, we did not promote loci based on het_{obs}/het_{exp} alone unless confirmed by one of the other lines of evidence. However, the het_{obs}/het_{exp} deviations were useful for extending clusters discovered by Mendel failures, because the Mendel failures themselves may not be observed at every SNP in the deleted region.

Fluorescent *in situ* hybridization (FISH). Fosmid clones with end sequences mapped to locations within predicted deletion intervals were obtained from the BAC/PAC resource, and DNA was isolated from each fosmid with the Maxi DNA plasmid kit (Qiagen). Fosmid DNAs were labeled by nick translation with Vysis Spectrum Green 11 dUTP (G248P89259F2 and G248P87989C3 on chromosome 4) or Spectrum Orange 11 dUTP (G248P87609A7 on chromosome 8 and G248P81036F4 on chromosome 18). We hybridized the test probes with appropriate positive control probes: Spectrum Orange 11 dUTP–labeled BAC clone RP11-363G1 (BAC/PAC; chromosome 4p15.1), and biotin-16 dUTP–labeled chromosome 8 and 18 paint probes (Roche). FISH experiments were performed using standard hybridization conditions on metaphase chromosome preparations derived from lymphoblastoid cell lines obtained from the Coriell Institute for Medical Research. Cy5-labeled streptavidin was used for detection of the biotin-labeled chromosome 8 and 18 paint probes. Images were captured on an Olympus AX70 fluorescent microscope equipped with a charge-coupled device (CCD) camera (Photometrics KAF 1400) with appropriate fluorescent filters and were analyzed with Applied Imaging's Genus software.

The chromosome 4 fosmids used for FISH validation (G248P89259F2 and G248P87989C3) are mapped to segmental duplication-containing regions[8]. Sequences with >94% nucleotide similarity are located <1 Mb (on chromosome 4) from each fosmid. We considered the possibility that these probes could hybridize to the paralogous sequence and yield a positive FISH signal, even if the target sequence were deleted. To investigate this, we repeated these experiments six times under various hybridization conditions, including once with an extended hybridization of 48 h. In four out of these six experiments for a given probe and in a minimum of 25 metaphase spreads examined per individual, we consistently observed zero fluorescent probe signals (for example, for fosmid probe G248P89259F2: NA19098), one signal (NA19100, NA19200, NA19202) or two fluorescent probe signals (NA19099, NA19201) per individual. Furthermore, in these experiments, we included parent-offspring trios, and FISH results were consistent with mendelian inheritance of deletions. We examined a minimum of 25 metaphase spreads per individual and found, in two experiments (including the 48-h hybridization protocol), that individuals believed to homozygous for the deletion had two faint signals (for example, fosmid probe G248P89259F2: NA19098); those heterozygous for the deletion had one faint and one strong signal (NA19100, NA19200, NA19202) and those homozygous for the nondeletion allele had two strong signals (NA19099, NA19201). **Supplementary Figure 1** shows such a signal

intensity difference in an individual heterozygous for the chromosome 4 deletion containing fosmid G248P87989C3.

Illumina (allele-specific fluorescence) validation of deletion variants. Seventeen candidate deletion variants covered at least three SNPs that had been assayed on the Illumina BeadArray platform at the Broad Institute. The BeadArray platform generates a quantitative allele-specific intensity measurement for each SNP allele in each individual in a population. The normalized allele-specific intensity measurements are comparable across individuals and generally fall into two or three discrete clusters, corresponding to individuals homozygous for allele 1, individuals homozygous for allele 2 and individuals heterozygous for alleles 1 and 2. For SNPs covered by predicted deletion variants, we observed additional genotype classes corresponding to individuals hemizygous for allele 1, individuals hemizygous for allele 2 and individuals homozygous for the deletion allele. We considered a deletion variant validated if (i) we observed one or more of these additional, well-separated genotype clusters and (ii) all of the individuals predicted (from multimarker genotype patterns) to be hemizygous or homozygous deleted in fact fell into the appropriate additional cluster.

PCR validation of homozygous deletions. To validate predicted homozygous deletions by PCR, we selected 60 candidate deletion loci for which the pattern of genotypes predicted the existence of at least two individuals with homozygous deletions in at least one population. The criterion for validation was confirmation of a precise predicted pattern of amplification success and amplification failure across at least 12 samples that included at least two predicted examples of each result. Any deviation from that pattern was classified as a confirmation failure. The predictions (about which individuals harbored homozygous deletions) were derived from the SNP genotypes—the individuals in whom multiple null genotypes had given rise to the prediction of a deletion (**Supplementary Table 1**) were predicted to be homozygous null; all other individuals were predicted to have genetic material at that locus. Importantly, we chose PCR amplification sites that were distinct from any of the sequences used in the SNP genotyping assays so that this would be an independent confirmation of a predicted result.

In addition to the validation experiments described in the main text, we also tested an additional 56 loci that were not among our core predictions but that met a more relaxed set of statistical thresholds; the confirmation rate among these other candidate variants was only 20%, suggesting that although additional true variants could be discovered, relaxation of the statistical thresholds would result in an unacceptable false positive rate.

Quantitative PCR. Individuals' deletion genotypes cannot be unambiguously inferred from SNP genotypes data. (That is, the detection of a deletion based on either Mendel errors or null genotypes is dependent on the genotype on the other (nondeletion) chromosome.) For this reason, it was necessary to develop assays for accurately typing the deletion variants. We performed two-color TaqMan assays, using a FAM-labeled probe for the test gene and a VIC-labeled probe (Applied Biosystems) for PMP22, a diploid control gene (assay primer sequences are described in **Supplementary Table 5**). Small (60–90 nt) amplicons from the test and control loci were simultaneously amplified in each tube, in 96-well plates (one plate per population, including five replicate samples and one blank sample) on a Bio-Rad iCycler. The threshold cycle (C_t) was calculated for each fluorophore separately, and the difference between the threshold cycles for the two fluorophores (ΔC_t) was used as a measurement of relative copy number that could be compared from sample to sample on the same plate. For each assay, the ΔC_t measurements clustered into three discrete groups (with one group typically showing no amplification of the test locus at all). For some assays, these groups were initially incompletely separated; in these cases, averaging of the ΔC_t measurements across three to five replicates resulted in discrete, well-separated clusters of average measurements. For each assay, we treated these three clusters as '+/+', '+/−' and '−/−' genotypes. In each case, the resulting genotype calls for replicate samples agreed completely, and the resulting genotypes showed mendelian inheritance and Hardy-Weinberg equilibrium.

Assessment of LD. To assess LD between gene deletion polymorphisms and SNPs, we used the gene-deletion genotypes ('+/+', '+/–', '–/–') obtained by quantitative PCR above. We obtained all HapMap SNP genotypes 200 kb to either side of the deletion locus, removed SNPs that were covered by the deletion and replaced them with the quantitative PCR–derived deletion genotypes from the same individuals. We used the Haploview program[18] to determine the phase of all SNP and deletion alleles and to calculate linkage disequilibrium (r^2) between deletions and SNPs.

URLs. HapMap release 16, before quality control filtering: http://hapmap.org/genotypes/2005-03_16a_phaseI/full/redundant-unfiltered/. Physical coordinates of candidate deletion polymorphisms have been deposited in the UCSC Genome Browser (http://genome.ucsc.edu).

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

ACKNOWLEDGMENTS

The authors wish to thank J. Moore and L. Ziaugra for contributing their expertise on the behavior of genotyping platforms and C. Patil, J. Melo and E. Lander for commenting on manuscript drafts. We thank G. Thorisson and A. Vernon-Smith for extensive help with data coordination, and D. Conrad, J. Pritchard and K. Frazer for exchanging manuscripts before publication.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

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

1. Sebat, J. *et al.* Large-scale copy number polymorphism in the human genome. *Science* **305**, 525–528 (2004).
2. Iafrate, A.J. *et al.* Detection of large-scale variation in the human genome. *Nat. Genet.* **36**, 949–951 (2004).
3. Sharp, A.J. *et al.* Segmental duplications and copy-number variation in the human genome. *Am. J. Hum. Genet.* **77**, 78–88 (2005).
4. Tuzun, E. *et al.* Fine-scale structural variation of the human genome. *Nat. Genet.* **37**, 727–732 (2005).
5. International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
6. Daiger, S.P. & Chakravarti, A. Deletion mapping of polymorphic loci by apparent parental exclusion. *Am. J. Med. Genet.* **14**, 43–48 (1983).
7. Chance, P.F. *et al.* DNA deletion associated with hereditary neuropathy with liability to pressure palsies. *Cell* **72**, 143–151 (1993).
8. Bailey, J.A. *et al.* Recent segmental duplications in the human genome. *Science* **297**, 1003–1007 (2002).
9. Seidegard, J., Vorachek, W.R., Pero, R.W. & Pearson, W.R. Hereditary differences in the expression of the human glutathione transferase active on trans-stilbene oxide are due to a gene deletion. *Proc. Natl. Acad. Sci. USA* **85**, 7293–7297 (1988).
10. Nunoya, K. *et al.* A new deleted allele in the human cytochrome P450 2A6 (CYP2A6) gene found in individuals showing poor metabolic capacity to coumarin and (+)-cis-3,5-dimethyl-2-(3-pyridyl)thiazolidin-4-one hydrochloride (SM-12502). *Pharmacogenetics* **8**, 239–249 (1998).
11. Pemble, S. *et al.* Human glutathione S-transferase theta (GSTT1): cDNA cloning and the characterization of a genetic polymorphism. *Biochem. J.* **300**, 271–276 (1994).
12. Wilson, W. *et al.* Characterization of a common deletion polymorphism of the UGT2B17 gene linked to UGT2B15. *Genomics* **84**, 707–714 (2005).
13. Monks, S.A. *et al.* Genetic inheritance of gene expression in human cell lines. *Am. J. Hum. Genet.* **75**, 1094–1105 (2004).
14. Morley, M. *et al.* Genetic analysis of genome-wide variation in human gene expression. *Nature* **430**, 743–747 (2004).
15. de Bakker, P.W. *et al.* Efficiency and power in genetic association studies. *Nat. Genet.* (in the press).
16. 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. *Nat. Genet.*, advance online publication 4 December 2005 (doi:10.1038/ng1697).
17. Hinds, D.A., Kloek, A.P. & Frazer, K.A. Common deletions and SNPs are in linkage disequilibrium in the human genome. *Nat. Genet.*, advance online publication 4 December 2005 (doi:10.1038/ng1695).
18. Barrett, J.C. *et al.* Haploview: Analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263–265 (2005).
19. Fry, B. *Computational Information Design*. Thesis, Massachusetts Institute of Technology (2005).