nature genetics

A high-resolution survey of deletion polymorphism in the human genome

Donald F Conrad[1], T Daniel Andrews[2], Nigel P Carter[2], Matthew E Hurles[2] & Jonathan K Pritchard[1]

Recent work has shown that copy number polymorphism is an important class of genetic variation in human genomes[1–4]. Here we report a new method that uses SNP genotype data from parent-offspring trios to identify polymorphic deletions. We applied this method to data from the International HapMap Project[5] to produce the first high-resolution population surveys of deletion polymorphism. Approximately 100 of these deletions have been experimentally validated using comparative genome hybridization on tiling-resolution oligonucleotide microarrays. Our analysis identifies a total of 586 distinct regions that harbor deletion polymorphisms in one or more of the families. Notably, we estimate that typical individuals are hemizygous for roughly 30–50 deletions larger than 5 kb, totaling around 550–750 kb of euchromatic sequence across their genomes. The detected deletions span a total of 267 known and predicted genes. Overall, however, the deleted regions are relatively gene-poor, consistent with the action of purifying selection against deletions. Deletion polymorphisms may well have an important role in the genetics of complex traits; however, they are not directly observed in most current gene mapping studies. Our new method will permit the identification of deletion polymorphisms in high-density SNP surveys of trio or other family data.

It has long been known that chromosomal deletions can lead to a variety of serious developmental and malformation disorders, such as DiGeorge and Prader-Willi syndromes[6–9]. While deletions that cause severe diseases are rare in the population, recent work has indicated that more benign deletions are widespread in the human genome, in many cases deleting genes[1–4]. These observations suggest that it is plausible that deletions may have a significant role both in the genetics of complex traits, as previously proposed for autism[10], and in genome evolution[11,12].

In this study we set out to obtain a detailed picture of the extent and distribution of deletion variation in the human genome. Our analysis made use of transmission patterns of SNP genotypes within parent-offspring trios from the International HapMap Project[5]. The samples consisted of two sets of thirty parent-offspring trios: a European-derived 'CEPH' sample (denoted 'CEU') and an African sample of Yoruba individuals from Ibadan, Nigeria ('YRI'). Our detection scheme identifies deletions that, in a given trio, are transmitted to the child; hence, our sample sizes are, in effect, 30 individuals from each population.

Deletions normally go undetected by current SNP genotyping methods. Instead, SNPs in regions that are hemizygous for a deletion are generally miscalled as homozygous for the allele that is present[13]. Hence, when a deletion is transmitted from parent to child, the genotypes at SNPs within the deletion region will often appear to violate the rules of mendelian transmission (**Fig. 1** (configuration A)). In what follows, we will distinguish between two classes of mendelian incompatibilities that we call 'Type I' (consistent with a deletion) and 'Type II' (inconsistent with a deletion; **Fig. 1** (configuration C)). Based on this logic, we developed a simple algorithm for scanning the HapMap trio data for unusual runs of consecutive SNPs that, in a single family, are either Type I mendelian incompatibilities or have other, less informative, genotype configurations consistent with the presence of a deletion (**Fig. 1** (configurations A, E); Methods).

Overall, the data quality of the HapMap is very high (Methods); the occurrence of even two Type I mendelian incompatibilities within a single run of deletion-compatible SNPs is very unlikely in the absence of a deletion. Thus, we scanned the HapMap data, labeling any such runs with more than two Type I mendelian incompatibilities as potential deletions. The outside Type I mendelian incompatibilities define the minimal deleted interval, with the maximal extent of the deletion bounded by genotypes incompatible with a heterozygous deletion.

Using the unfiltered version of HapMap build 16c.1, we identified a total of 453 and 680 regions in the CEU and YRI samples that met our initial criteria as candidate deletion regions. After removing deletions that appear to be cell line–specific, likely somatic deletions and obvious artifacts of experimental error (Methods), we were left with 345 candidate deletions in the CEU samples and 590 in the YRI (**Supplementary Tables 1** and **2** online). Two different analytical methods, using very different assumptions, suggest that the false positive rate for the predicted deletions is low (the two methods yield overall false positive rate estimates of 5% and 14% respectively; Methods).

In order to experimentally validate our deletion detection method, we first tested 12 predicted deletions using quantitative PCR. For

[1]Department of Human Genetics, The University of Chicago, 920 East 58th Street, Chicago, Illinois 60637, USA. [2]Genome Dynamics and Evolution Group, The Wellcome Trust, Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. Correspondence should be addressed to J.K.P. (pritch@uchicago.edu).

Received 13 June; accepted 6 October; published online 4 December 2005; doi:10.1038/ng1697

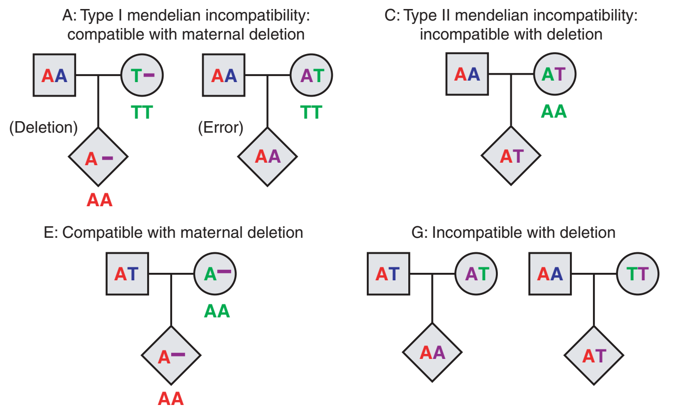
A: Type I mendelian incompatibility: compatible with maternal deletion
AA
T–
TT
(Deletion)
A–
AA
AA
AT
TT
(Error)
AA
C: Type II mendelian incompatibility: incompatible with deletion
AA
AT
AA
AT
E: Compatible with maternal deletion
AT
A–
AA
A–
AA
G: Incompatible with deletion
AT
AT
AA
AA
TT
AT

Figure 1 Examples of four of the seven types of trio genotype configurations used in this analysis. The true genetic state of each individual is depicted within his or her pedigree symbol. The called genotype, when it differs from the true genotype, is shown outside the pedigree symbol. The three upper configurations (A and C) all result in mendelian incompatibilities. We define 'Type I mendelian incompatibilities' as those that are compatible with a deletion transmitted from parent to child and 'Type II mendelian incompatibilities' as those that are incompatible with the deletion model. Key to figure: A: mendelian incompatibility, genotypes compatible with a deletion transmitted from the mother; C: mendelian incompatibility, genotypes incompatible with a transmitted deletion; E: no mendelian incompatibility, genotypes compatible with a deletion transmitted from the mother (but not the father); G: no mendelian incompatibility, genotypes incompatible with a transmitted deletion. Candidate deletion regions are runs of consecutive SNPs with at least two Type I mendelian incompatibilities and other SNPs that are compatible with a deletion; all the SNPs must suggest transmission from the same parent. See further details in Methods.

all 12 deletions, we observed DNA concentrations consistent with transmission of a deletion from parent to child (**Supplementary Methods** online).

Next, to provide more extensive validation by comparative genome hybridization (CGH), we designed a custom oligonucleotide microarray (see Methods) that comprises 380,000 probes that tile across all candidate deletions identified in nine HapMap offspring (eight YRI and one CEU). The results of this CGH analysis indicate that the great majority of candidate deletions detected by our method are real. After correcting for multiple comparisons, we found evidence for deletions in all but 13 out of 93 scorable deletions, representing an empirical false positive rate of 14% (**Table 1** and **Supplementary Table 3** online). Beyond providing validation, the CGH data underline just how widespread copy number polymorphism (CNP) is in the human genome: there are numerous examples where a validated deletion is present in more individuals than were predicted from HapMap genotypes, including occasionally in deletion homozygotes, as well as additional CNPs near to candidate deletions. The CGH data also show that the deletion boundaries generally fall within the minimal and maximal extents predicted from the SNP data, with a few notable

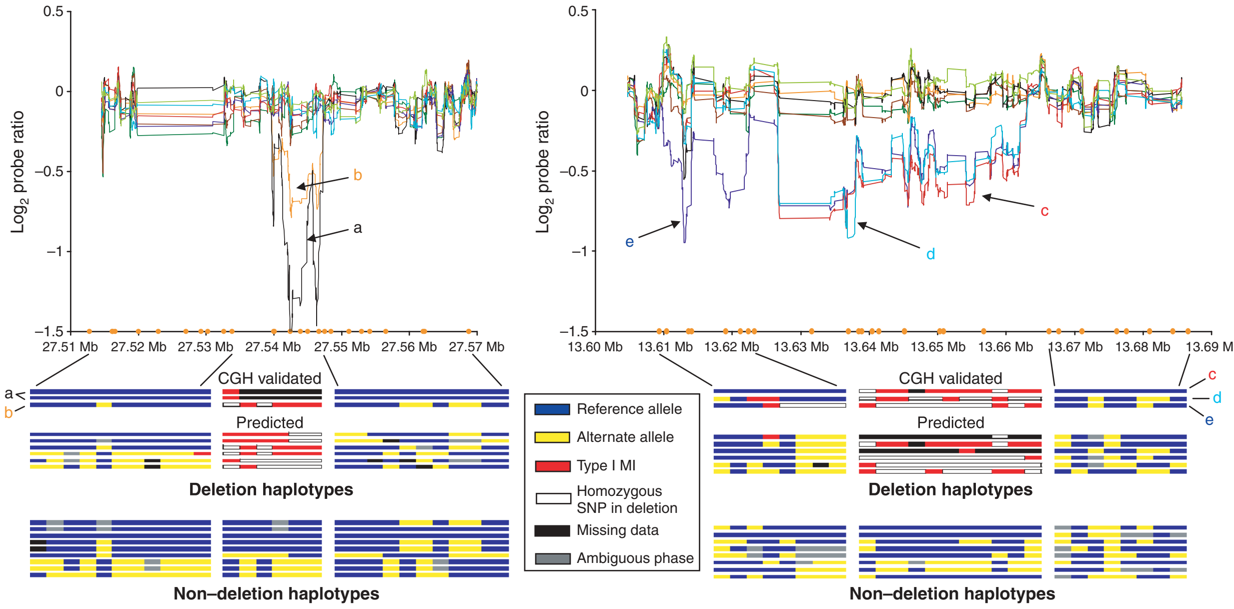

Figure 2 Comparison of high-resolution CGH results (top) with haplotype patterns (bottom) for two regions containing deletions. The colored lines in the top panels show moving averages of the relative probe intensities for each of eight YRI individuals relative to a single CEU reference individual. Values are plotted on a $\log_2$ scale so that values around 0 indicate that both individuals have the same copy number, and significantly lower values indicate deletions in the YRI individual. In region A (12p11.23), individuals a and b are homozygous and heterozygous for deletions, respectively; in region B (8p22), individuals c, d, and e are all heterozygous for deletions, although the deletion in individual e has a different boundary on the left side. The lower panels plot haplotypes in the same regions for (i) the deletion-bearing YRI chromosomes confirmed by CGH; (ii) additional YRI chromosomes that have unusual patterns of mendelian incompatibilities and missing data in the deletion region suggesting that they, too, carry deletions; and (iii) some of the YRI chromosomes shown by CGH not to be carrying deletions. Haplotype phase was established using the trio information. SNPs in the haplotypes are colored as shown in the key; the choice of which allele is blue at each SNP is arbitrary. SNP positions are marked in orange along the *x*-axes, with black lines extending down to the haplotypes to show the correspondence between the SNPs and key haplotype positions. At both loci, the deletion-bearing haplotypes suggest recurrent deletion events, although there are no intrachromosomal segmental duplications of at least 1 kb in length and >90% identity in either region.

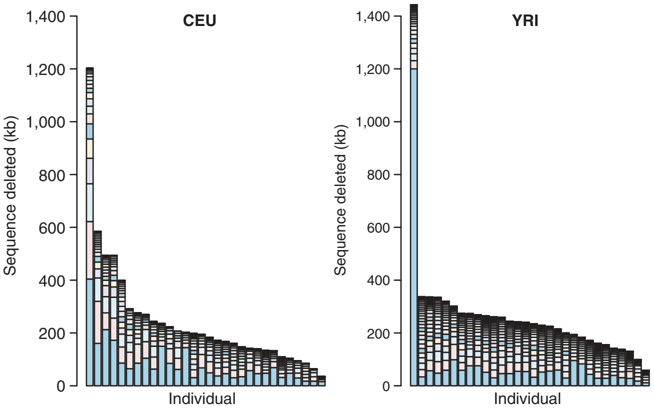

Figure 3 Cumulative length of hemizygous sequence detected within the 30 children in each sample. The deletions were sorted into bins corresponding to the individual in which they were detected and then were sorted within bins by length. The same color deletion in two different individuals does not indicate a shared variant; rather, the rank of length is the same for each color. The CEU individual (left) has about 1.1 Mb of deleted material in 8p23, a well-known region for rearrangements[14]. Although our method splits this into four distinct deletions, it may be parsimonious to believe that this is actually a single event.

exceptions in which genotyping errors split a single long deletion into shorter deletions (**Supplementary Methods**).

The CGH data also show that even within small regions, deletion polymorphism was often complex, with multiple deletion breakpoints (**Fig. 2**, **Supplementary Note** online). Moreover, by superimposing the SNP data onto the CGH results, it was apparent that deletions in a single region often sat on more than one haplotype, suggesting that the deletion events have occurred multiple times (**Fig. 2**), although formal quantification is difficult.

In total, the lengths of the deletions identified from the SNP data followed an L-shaped distribution, with many small deletions and few large ones (**Figs. 3** and **4**; refs. 14,15). The 345 predicted CEU deletion regions had a median length of 10.6 kb (range: 0.3–404 kb). There were ~70% more deletions detected in the YRI sample, but these were smaller, with a median length of 8.5 kb and with fewer long deletions (range: 0.5–1,200 kb). This difference in deletion prevalence cannot be explained by the slightly higher rate of false positives in the YRI sample; it is more likely to reflect the greater diversity commonly found in African-derived populations[15,16].

Most of the deletions that we detected were at low frequencies: 39% of deletions did not overlap with mendelian incompatibilities in any other trios; the most common was a deletion detected in 15/30 YRI children (see Methods). The 590 YRI deletions occurred at 396 distinct genomic locations; the 345 CEU deletions were at 228 locations. Thirty-seven deletion locations were shared between the populations. We found that 15.7% of our deletions (but 47.3% of deletions larger than 60 kb) contained or overlapped segmental duplications. This is less than in a previous study[3], perhaps owing to underrepresentation of segmental duplications by HapMap SNPs. Just 11% of our predicted deletions matched CNPs identified by previous studies[1–3]. Low overlap between studies has been noted previously[3,17]. This may be explained in part by the low frequencies of most of these variants and, in the present case, by the fact that we have greater power to detect small deletions.

Although we have detected many candidate deletion regions, there may be many more in these trios that have gone undetected. The power to find any given deletion depends on both the number of SNPs within the deleted region and the SNP allele frequencies (which determine the probability that the deletion will generate mendelian incompatibilities). In order to estimate how many deletions we are missing, we developed a simulation to assess the power to detect deletions in HapMap Build 16c.1 (**Fig. 5**). On most chromosomes, we achieved 50% power to detect deletions larger than about 25 kb. The average power curves were similar in the YRI and CEU samples, suggesting that the much larger number of deletions observed in YRI was not an artifact of having greater power in that sample.

We next fitted a simple model to the observed data, making use of the number and size distribution of the detected deletions, along with the power to detect deletions of a given size, to estimate the true number and size distribution of all deletions in this sample (Methods). This analysis confirms that most deletions in the genome are small—in fact, the skew towards small deletions was even stronger than suggested by the size distribution of detected deletions (**Fig. 4**). Furthermore, we were able estimate the total number of deletions of a given size. For the summaries below, we report values for a minimum deletion size of 5 kb, as the data provide little direct information about deletions smaller than this. In total, the CEU children are estimated to carry about 900 deletions of >5 kb, with an average size of 18.5 kb. Hence a typical CEU individual has around 30 deletions larger than 5 kb and is hemizygous for a total of about 555 kb of sequence in deletions >5 kb. The YRI sample is estimated to have 1,525 deletions of >5 kb, with a smaller average size (14.8 kb). An average YRI individual is estimated to have about 50 deletions >5 kb and to be hemizygous for an average of 740 kb. These estimates correspond to the deletion load in the easy-to-genotype, 'HapMappable' genome.

A priori, one might expect that deletions may often be subject to purifying selection. To further study the impact of selection on these variants, we determined the proportion of SNPs within deletions that were either in coding sequence or within introns (**Supplementary**

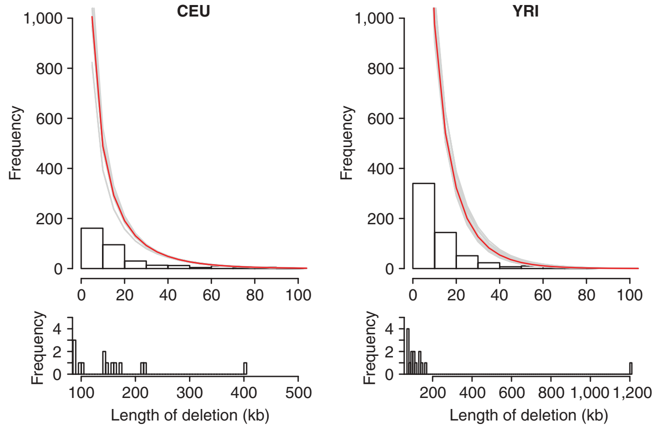

Figure 4 Observed and predicted distributions of deletion sizes in the CEU and YRI samples. The size distributions of the observed deletion regions are plotted as histograms. For each sample, the upper panel shows a histogram for deletions ≤100 kb in size; the lower panel, deletions >100 kb. The red lines show our estimates of the true numbers of deletions of each size in the two-population samples, taking account of the fact that there is incomplete power to detect small deletions. The gray lines show a sample of 100 random draws from the posterior distribution to give an indication of the degree of uncertainty in the estimates. The estimated distributions are not shown on the lower panels, as they are essentially zero.

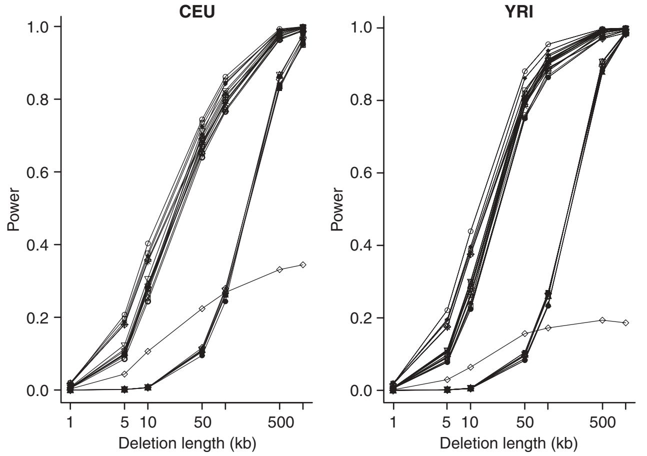

Figure 5 Power to detect a single deletion transmitted from parent to child, in HapMap Release 16c.1, as a function of deletion size (see Methods). A separate curve is drawn for each chromosome. The power to detect deletions on chromosomes 5, 11, 14, 15, 16, 17 and 19 is greatly reduced owing to filtering of mendelian incompatibilities by one genotyping center before data release. The power to detect deletions on the X chromosome (open diamonds) is also reduced, as we can detect only those deletions passed from mother to daughter and because there are only 16 and 7 daughters in the CEU and YRI samples, respectively.

Table 1 CGH validation results for predicted deletions

	≥10 probes			≥30 probes		
MIs	True	False+	Complex	True	False+	Complex
2 MIs	42	13	4	32	10	1
3 MIs	20	0	1	19	0	1
>3 MIs	18	0	0	18	0	0
Total	80	13	5	69	10	2

CGH arrays were used to test predicted deletions in one CEU and eight YRI individuals. Of 134 predicted deletions in those individuals, we were able to target 98 and 81 deletions with at least 10 and 30 unique probes, respectively. Predicted deletions were classified as true, false positives, or 'complex'. Deletions classified as complex were deemed unscorable due to the presence of extensive segmental duplication and the appearance of complex copy number variation that was not easily reconciled with a simple deletion event. The classifications are more reliable in the ≥30 probe set, but at the cost of reduced sample size (**Supplementary Methods**). MI: mendelian incompatibility.

Methods). We found that genic SNPs were strongly underrepresented in deletion regions compared with the HapMap average (23.6% of 6,470 YRI SNPs and 18.6% of 4,312 CEU SNPs in deletion regions versus 33.8% in the entire Phase I HapMap data; $P < 10^{-15}$), consistent with the hypothesis that genic deletions tend to be deleterious. There was also a modest deficit of X-chromosome deletions, perhaps reflecting stronger selection against X-chromosome deletions in males (**Supplementary Methods**).

Nonetheless, a large number of genes were affected by the predicted deletions. The deletions spanned at least part of 267 known and predicted genes (**Supplementary Table 4** online). Coding sequence was deleted in 201 of these genes; 92 genes of these were entirely deleted. The observation that genic deletions are frequent suggests that deletion polymorphism may well be an important contributor to complex disease risk. We therefore further investigated the types of genes involved in deletions.

In total, 23 of the genes containing deletions had disease-associated OMIM entries. At least four deletions occurred within exons or introns of autosomal recessive mendelian disease genes, including *STRC* (hereditary deafness), *FSHB* (follicle-stimulating hormone deficiency), *GCNT2* (i blood group, congenital cataract) and *NEB* (nemaline myopathy). In addition, numerous cancer-related genes were identified as partly or totally deleted in this study (such as *DLEU7*, *TUSC3*, *BCAS1*, *HIC2*, *LOH12CR1*). These may represent genes which are much more likely to experience loss of heterozygosity due to natural deletion polymorphisms or might represent fragile areas of the genome that are more likely to experience rearrangement in cancerous cells.

Using information from the PANTHER database, we assigned each of the 267 affected genes to a molecular function and biological process and compared the resulting distribution of annotations with the functional distribution of all the genes in the genome (**Supplementary Methods**). Genes involved in immunity and defense, sensory perception, cell adhesion and signal transduction seem to be especially prone to deletion (see also refs. 1,3,18). Genes that encode nucleic acid binding proteins or proteins that are involved in nucleic acid metabolism seem to be underrepresented in our set of genes when compared with the genome average. Various multigene families were represented in our set of deletions, including the MHC, assorted olfactory receptor families, a pregnancy-specific glycoprotein family, the salivary protein complex, killer cell immunoglobulin-like receptors, the defensins, the B melanoma antigen family and the UGT2B family of drug detoxification enzymes.

These deletion-enriched functional classes[1,3] overlapped strikingly with those identified in studies of genes in segmental duplications[18], genes under positive selection[19] and lineage-specific gene family expansions and contractions[20]. This relationship may reflect the presence of these gene families in regions of structural dynamism (for example, in regions of segmental duplication) that lend themselves to (i) deletion, (ii) lineage-specific expansions and (iii) positive selection after gene duplication.

In summary, we report here on a new approach to studying deletion variation on a genomic scale. We find that deletions of 5 kb and larger are extremely widespread in the human genome; indeed, our estimates of the abundance of deletions are most likely underestimated (i) because our estimates are biased downward on chromosomes where there is low power (data not shown) and (ii) owing to the low SNP density in segmental duplication regions. In view of the large number of genes affected by the deletions that we detected and the evidence for purifying selection against deletions, it seems highly plausible that deletions may be important in the genetic basis of complex traits. Therefore, it will be important to continue the development of high-resolution techniques for studying deletion variation, including in genetic association studies.

Note added in proof: see related papers by Frazer and colleagues[21] and McCarroll et al.[22] in this issue for related analyses of deletion variants.

METHODS

Data processing. To create our working data set, we downloaded the 'unfiltered' version of the Phase I (Build 16c.1) International HapMap Consortium (IHMC) HapMap for the autosomes and X chromosome. These uncleaned data include 1,302,761 SNP genotyping records in 30 CEPH trios from Utah ('CEU') and 1,273,629 SNP genotyping records in 30 Yoruba trios from Ibadan, Nigeria ('YRI'). Each record corresponds to 90 genotype calls at a single SNP (30 trios × three people per trio). Five of the ninety individuals in each sample were genotyped twice at every locus ('plate duplicates'). Except when assessing error rates (as noted in the text), we used the first recorded genotype for each of the duplicated individuals in all our analyses. Many SNPs in the HapMap have been genotyped multiple times (for quality control); when we encountered multiple genotype records with the same reference SNP (rs) number, we selected the record with the least amount of missing data. To remove SNPs with potential genotyping problems, we filtered the data as follows. Records were discarded if they failed any of the following checks: (i) if a χ^2 test for Hardy-Weinberg equilibrium produced a P-value ≤ 0.001, (ii) if the record was missing $\geq$20% data, (iii) if there was more than one discrepancy between plate duplicates, or (iv) if the record was flagged by the submitter. These criteria match the filtering process used for the HapMap's 'filtered' data releases, except that we did not remove SNPs with mendelian incompatibilities, as such SNPs may indicate deletions. We added another layer of filtering by removing records for loci that were typed more than once and were discordant for more than one genotype call across experiments. After filtering and removing duplicate records, we obtained 1,108,950 and 1,085,823 unique SNPs in the CEU and YRI samples, respectively.

Deletion detection. After experimenting with several procedures for identifying deletions, including a Hidden Markov Model approach, we settled on the following fairly simple method. Our algorithm aims to detect deletions that are transmitted from a hemizygous parent to a child. For each trio, every SNP was coded into one of seven categories (**Fig. 1**): (A) Type I mendelian incompatibility (that is, consistent with deletion) involving mother; (B) Type I mendelian incompatibility involving father; (C) Type II mendelian incompatibility (that is, inconsistent with deletion); (D) child homozygous or missing data, both parents homozygous or missing data; (E) child homozygous or missing data, father heterozygous, mother homozygous or missing data; (F) child homozygous or missing data, mother heterozygous, father homozygous or missing data; (G) child heterozygous or both parents heterozygous (see **Supplementary Methods** for further details). SNPs were assigned to states D–G only if they did not contain mendelian incompatibilities. A run of consecutive SNPs in a particular trio was considered to be consistent with a maternal transmitted deletion if all SNPs were in states A, D or E, or with a paternal deletion if all SNPs were in states B, D or F.

The algorithm is designed to detect deletions that are hemizygous in one parent only and passed to the child. Hence, it would miss homozygous deletions, as these would likely result in missing data rather than mendelian incompatibilities. However, missing data are more common than mendelian incompatibilities in the HapMap, and hence it is harder to identify deletions using this kind of pattern. One might also wish to identify duplications from SNP genotype data; however, the likely genotype patterns are less predictable, making this harder. For example, an AAT individual might plausibly be recorded as AT, or AA or ambiguous (and hence missing data). Our algorithm may also detect cell line–specific deletions. If these occur in the child, they should mimic our deletion signal; when they occur in a parent only, they should be recognizable as producing a mix of both Type I and Type II mendelian incompatibilities.

In addition to the SNP filtering described above, we also performed additional filtering to remove candidate deletions that may not be transmitted deletions. That is, we discarded deletion regions that (i) contained one or more Type II mendelian incompatibilities, as these can indicate a somatic deletion in a parent; (ii) occurred in unusual clusters of deletions in a trio that was an outlier in terms of total number of deletions; or (iii) occurred in known regions of somatic recombination. We believe that this filtering is conservative, in the sense of incorrectly removing some germline deletions, though it does not guarantee that all cell line–specific deletions are removed. The largest single anomaly is a cell-line deletion of 104 Mb on chromosome 1q of one individual. Further details are provided in the **Supplementary Methods**.

In one analysis, we report estimates of the frequencies of deletions. In that analysis and in **Figure 2** only, as we have modest power to detect small deletions, we have also considered trios in which a single mendelian incompatibility overlaps an identified deletion in another family to represent likely additional copies of the original deletion (**Supplementary Methods**). Simple calculations indicate that this approach should have an acceptably low false positive rate.

Computational assessment of false positive rate. In our scheme, the bulk of the signal comes from observing unusual clusters of Type I mendelian incompatibilities. Mendelian incompatibilities can arise for a variety of reasons besides true deletions: these include random genotyping errors, errors due to null alleles in the primer binding sequence for an assay and errors due to duplicated SNPs or paralogous sequence variants[23]. Therefore, it is essential to assess the overall rate of mendelian incompatibilities in the HapMap data. Overall, the genotyping accuracy in the dataset is extremely high. After data cleaning, the discordance rates for repeated genotype experiments on a single genotyping plate were 9.1×10^{-4} per genotype and 1.0×10^{-3} across independent experiments at different centers or on different platforms. The overall frequencies of Type I mendelian incompatibilities in the data are 5.3×10^{-4} (CEU) and 1.1×10^{-3} (YRI) per SNP trio. The frequencies of Type II mendelian incompatibilities are 3.9×10^{-4} (CEU) and 5.0×10^{-4} (YRI), respectively (**Supplementary Methods**).

In order to assess the significance of potential deletions, we used the following simulation study to determine the frequency of deletion candidate regions under the null hypothesis of 'no deletions'. We tabulated the frequency of each state for each combination of genotyping center and platform, separately for the CEU and YRI samples, yielding a matrix of 7 states × 11 platforms/centers × 2 samples. We then simulated 5,000 HapMap datasets, sampling the state of each SNP independently from the appropriate frequencies, given the platform/center that was originally used to type that SNP. These simulations suggested that requiring $\geq$2 Type I mendelian incompatibilities in a deletion-compatible region is a moderately stringent criterion, as we observed an average of 11 such runs per simulated CEU HapMap and 34 per simulated YRI HapMap. These results suggest that the false positive rates are ~11/345 (or 3%) and ~34/590 (or 6%) in the CEU and YRI, respectively.

The latter simulation approach assumes that mendelian incompatibilities occur independently across neighboring SNPs, an assumption that may not hold in practice. Therefore, we also developed an empirical method for estimating the false positive rate that naturally accommodates correlation in genotyping errors across neighboring SNPs. Consider four possible arrangements of two Type I mendelian incompatibilities in which the two mendelian incompatibilities are either contiguous or separated by uninformative trio configurations compatible with the first mendelian inconsistency. Using the nomenclature described above, these are (i) A-[D,E]-A; (ii) B-[D,F]-B; (iii) A-[D,E]-B; (iv) B-[D,F]-A. Only the first two arrangements, which represent pairs of mendelian incompatibilities involving the same parent, are called as deletions. The latter two configurations are inconsistent with a single transmitted deletion and are presumably the result of genotyping errors. The number of Type iii and iv arrangements should be a good estimate of the number of Type i and ii arrangements that are due to genotyping error. In the CEU data, there are 345 Type i and ii arrangements and 41 Type iii and iv arrangements; this suggests that the false positive rate in our deletion calls is 41/345 or 11.9%. Likewise, in the YRI, there are 590 Type i and ii arrangements and 88 Type iii and iv arrangements, suggesting a false positive rate of 14.9%. This method is conservative in that some Type iii and iv arrangements may represent true, adjacent deletions that originate in different parents.

Validation experiments using CGH. Nine trio offspring (eight YRI (NA18500, NA18503, NA18506, NA18515, NA18521, NA18854, NA18857 and NA18860) and one CEU (NA10851)) were selected for extensive validation of predicted deletions by array CGH. We designed custom long oligonucleotide arrays that tile at high density (8 bp median spacing) across the nonrepeat masked portion of genomic intervals that encompass 134 predicted deletions in these nine individuals. These intervals were designed such that a substantial proportion of the oligonucleotide probes lay in undeleted flanking sequence outside the maximal extent of the predicted deletion. Genomic DNA for the nine individuals was obtained from Coriell Cell Repositories. Eight array CGH experiments were performed in which each YRI individual in turn was labeled with Cy3 and hybridized along with Cy5-labeled NA10851 (as a common reference DNA) to an array. Isothermal oligo design, array fabrication, DNA labeling, CGH experiments, data normalization and $\log_2$(Cy3/Cy5) ratio calculations were performed by NimbleGen[24]. Only intensity data from oligonucleotide probes with unique matches in the genome were used in downstream analyses (representing 288,629 informative probes per array).

For each individual, we applied Mann-Whitney U tests to compare the distribution of $\log_2$ ratios within each minimal deleted region defined by mendelian incompatibility SNPs with $\log_2$ ratios in the union of the undeleted flanking sequences of these deletions. This test is less sensitive to cryptic (invisible in HapMap genotypes) copy number variation than comparisons between a deletion and its neighboring flanking sequences. We observed that cryptic copy number variation within some flanking sequences arose both because some deletions were longer than predicted (due to erroneous genotypes within HapMap data) and through independent deletion or duplication events. The resultant P values were used to estimate the number of probes required within a minimal deleted interval in order to minimize false negative calls. The proportion of deletions called as false positives (using a P value threshold of 0.05, with or without Bonferroni correction) decreases quickly as the minimal number of 'deletion' probes is increased to 10 and asymptotes at around 30 (**Supplementary Methods**). Out of the original 134 candidate deletions, 21 had no probes with unique matches in the genome. Ninety-eight predicted deletions contained ten or more probes within the minimal deleted interval and were carried forward for detailed analysis. Five of these predicted deletions showed evidence of complex copy number variation within highly segmentally duplicated intervals and were removed from further analysis. Of the remaining ninety-three predicted deletions, seventy-two had Mann-Whitney U test P values of less than 0.00054 (0.05 Bonferroni corrected for 93 tests).

Statistical tests for divergent $\log_2$ ratios are not sufficient to accurately assign predicted deletions as either true deletions or false positives. If the reference DNA shares the same deletion as the test individual, then no significant variation in $\log_2$ ratios between deleted and undeleted flanking sequence will be observed (see example plots in **Supplementary Note** and **Supplementary Methods**). However, deletions in the reference can be detected as elevated $\log_2$ ratios in other test individuals who do not share the same deletion. Similarly, $\log_2$ ratios may show significant ($P < 0.05$ with or without Bonferroni correction) discrepancies in the absence of copy number variation as a result of regional biases in incorporation of Cy3 and Cy5 dyes. However, such regional biases should be common to all eight array CGH experiments. Therefore, plots of the moving average over 30 probes of $\log_2$ ratios in each of the eight experiments were overlaid for each predicted deletion. The overlaid plot for each predicted deletion was inspected, and the calls of true deletions and false positives manually curated. As a result, ten predicted deletions in which the Mann-Whitney U Test P value exceeded 0.00054 (0.05 Bonferroni corrected for 93 tests) were reclassified as true deletions, and two predicted deletions in which the P value was less than 0.00054 were reclassified as false positives.

In principle, a true deletion may be called as a false positive if all nine individuals share the same heterozygous deletion. However, for all 13 of the predicted deletions called as false positives there were two or more other test individuals with heterozygous SNP genotypes within the minimal deleted interval, indicating the absence of a heterozygous deletion.

Another possible concern is that the length of a given deletion may have an effect on our ability to validate it and in turn bias our false positive estimate. However, when we compared the length of the minimal deletion interval of the false positive two–mendelian incompatibility deletions to the true deletions with two mendelian incompatibilities, the failed deletions were of slightly longer mean length in both the >10 and >30 probe datasets.

Estimation of power. The power to detect a deletion in the HapMap is a function of the number of SNPs that are typed within the deletion and the allele frequencies at those SNPs. We estimated power as a function of deletion length, for a range of lengths between 1 kb and 2 Mb, by simulating deletions in the existing (filtered) HapMap data. The following procedure was repeated 10,000 times separately for each chromosome and for both populations. For each simulated deletion, a parent and a deletion region of specified size on the appropriate chromosome were chosen at random. This parent's transmitted gamete was then deleted in both the parent and offspring in the selected region, and the parent and offspring genotypes were reconstructed to be homozygous for the nondeleted nucleotides. If this process created at least two mendelian incompatibilities, the deletion was considered to be 'detected'.

Model fitting. We created a formal hierarchical model for the distribution of deletion sizes in the genome in order to improve the accuracy of the length estimates of the observed deletions and to estimate the total distribution of deletions, both observed and unobserved, in the genome.

The following model was evaluated separately for the CEU and YRI samples. Let c be the rate of deletions (of all sizes) per individual per base pair. (Multiple copies of the same deletion are counted separately.) Then on chromosome i, which is of length b_i base pairs, the total number of deletions, t_i, in the 30 offspring in the trios is taken as Poisson distributed with parameter $30b_ic$. In the experiment, we detect n_i deletions among the 30 trios, where $n_i \leq t_i$. Each of the t_i deletions is assumed to draw its size from a Gamma distribution with parameters α and β. The probability that a given deletion is detected depends on the power to find a deletion of that size on chromosome i (p_il, estimated as above). Then, conditional on the data, we perform joint inference for the size distribution (specified by α and β), the true rate of deletions per individual per base pair (c), and the sizes of each of the deletions detected. A Markov chain Monte Carlo algorithm was implemented to sample from the posterior distribution of this model. Throughout the article, length estimates reported for detected deletions are based on the posterior median lengths obtained from this method. The CGH data indicate a modest tendency for the largest deletions to be broken into multiple smaller events. However, this occurs at a low rate and should have only a modest effect on the estimates.

Accession codes. The full CGH data reported here are available in the Gene Expression Omnibus (GEO) as accession number GSE3474.

URLs. The database of segmental duplications used in this paper is available at http://humanparalogy.gs.washington.edu. PANTHER database: http://panther.appliedbiosystems.com.

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

ACKNOWLEDGMENTS
We thank G. Coop, D. Cutler, A. DiRienzo, H. Fiegler, M. Przeworski, G. Raca, C. Tyler-Smith and D. Vetrie for comments and discussions; R. Redon for managing the genomic DNA collection; B. Voight for extracting the SNP classifications; S. Das, A. DiRienzo and C. Ober for lab space and equipment and the members of the International HapMap Consortium for their work in creating this dataset. This work was supported by a grant to J.K.P. from the Packard Foundation. D.F.C. was supported in part by US National Institutes of Health/National Institute of General Medical Sciences Genetics and Regulation Training Grant GM07197. Additional funding was provided by the Wellcome Trust.

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. Tuzun, E. *et al.* Fine-scale structural variation of the human genome. *Nat. Genet.* **37**, 727–732 (2005).
4. Sharp, A.J. *et al.* Segmental duplications and copy-number variation in the human genome. *Am. J. Hum. Genet.* **77**, 78–88 (2005).
5. The International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–796 (2003).
6. Schmickel, R.D. Contiguous gene syndromes: a component of recognizable syndromes. *J. Pediatr.* **109**, 231–241 (1986).
7. Chen, K.S. *et al.* Homologous recombination of a flanking repeat gene cluster is a mechanism for a common contiguous gene deletion syndrome. *Nat. Genet.* **17**, 154–163 (1997).
8. Flint, J. *et al.* The detection of subtelomeric chromosomal rearrangements in idiopathic mental retardation. *Nat. Genet.* **9**, 132–140 (1995).
9. Gardner, R.J. & Sutherland, G.R. *Chromosomes Abnormalities and Genetic Counseling* (Oxford Univ. Press, Oxford, 2004).
10. Yu, C.E. *et al.* Presence of large deletions in kindreds with autism. *Am. J. Hum. Genet.* **71**, 100–115 (2002).
11. Petrov, D.A. Mutational equilibrium model of genome size evolution. *Theor. Popul. Biol.* **61**, 531–544 (2002).
12. Olson, M.V. When less is more: gene loss as an engine of evolutionary change. *Am. J. Hum. Genet.* **64**, 18–23 (1999).
13. Amos, C.I., Shete, S., Chen, J. & Yu, R.K. Positional identification of microdeletions with genetic markers. *Hum. Hered.* **56**, 107–118 (2003).

14. Giglio, S. *et al.* Olfactory receptor-gene clusters, genomic-inversion polymorphisms, and common chromosome rearrangements. *Am. J. Hum. Genet.* **68**, 874–883 (2001).
15. Weber, J.L. *et al.* Human diallelic insertion/deletion polymorphisms. *Am. J. Hum. Genet.* **71**, 854–862 (2002).
16. Bhangale, T.R., Rieder, M.J., Livingston, R.J. & Nickerson, D.A. Comprehensive identification and characterization of diallelic insertion-deletion polymorphisms in 330 human candidate genes. *Hum. Mol. Genet.* **14**, 59–69 (2005).
17. Carter, N.P. As normal as normal can be? *Nat. Genet.* **36**, 931–932 (2004).
18. Bailey, J.A. *et al.* Recent segmental duplications in the human genome. *Science* **297**, 1003–1007 (2002).
19. Clark, A.G. *et al.* Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science* **302**, 1960–1963 (2003).
20. Emes, R.D., Goodstadt, L., Winter, E.E. & Ponting, C.P. Comparison of the genomes of human and mouse lays the foundation of genome zoology. *Hum. Mol. Genet.* **12**, 701–709 (2003).
21. 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).
22. McCarroll, S.A. *et al.* Common deletion polymorphisms in the human genome. *Nat. Genet.* Advance online publication, 4 December 2005 (doi:10.1038/ng1696).
23. Fredman, D. *et al.* Complex SNP-related sequence variation in segmental genome duplications. *Nat. Genet.* **36**, 861–866 (2004).
24. Selzer, R.R. *et al.* Analysis of chromosome breakpoints in neuroblastoma at sub-kilobase resolution using fine-tiling oligonucleotide array CGH. *Genes Chromosom. Cancer* **44**, 305–319 (2005).