nature genetics

Integrated detection and population-genetic analysis of SNPs and copy number variation

Steven A McCarroll[1–4,10], Finny G Kuruvilla[1–4,10], Joshua M Korn[1–6], Simon Cawley[7], James Nemesh[1], Alec Wysoker[1], Michael H Shapero[7], Paul I W de Bakker[1,4,8], Julian B Maller[3], Andrew Kirby[3], Amanda L Elliott[1], Melissa Parkin[1], Earl Hubbell[7], Teresa Webster[7], Rui Mei[7], James Veitch[7], Patrick J Collins[7], Robert Handsaker[1], Steve Lincoln[7], Marcia Nizzari[1], John Blume[7], Keith W Jones[7], Rich Rava[7], Mark J Daly[1,3,4,9], Stacey B Gabriel[1] & David Altshuler[1–4,9]

Dissecting the genetic basis of disease risk requires measuring all forms of genetic variation, including SNPs and copy number variants (CNVs), and is enabled by accurate maps of their locations, frequencies and population-genetic properties. We designed a hybrid genotyping array (Affymetrix SNP 6.0) to simultaneously measure 906,600 SNPs and copy number at 1.8 million genomic locations. By characterizing 270 HapMap samples, we developed a map of human CNV (at 2-kb breakpoint resolution) informed by integer genotypes for 1,320 copy number polymorphisms (CNPs) that segregate at an allele frequency >1%. More than 80% of the sequence in previously reported CNV regions fell outside our estimated CNV boundaries, indicating that large (>100 kb) CNVs affect much less of the genome than initially reported. Approximately 80% of observed copy number differences between pairs of individuals were due to common CNPs with an allele frequency >5%, and more than 99% derived from inheritance rather than new mutation. Most common, diallelic CNPs were in strong linkage disequilibrium with SNPs, and most low-frequency CNVs segregated on specific SNP haplotypes.

Genome-wide association studies were made possible by accurate, detailed maps of human sequence variation, and by highly accurate methods for typing SNPs. Over the same time that first-generation genome-wide association studies have been conducted, the human genome has been found to show extensive copy number variation[1–9]. Progress has been made in identifying large genomic regions that seem to harbor CNVs[9] and in finer-scale descriptions of many CNVs in specific individuals[10,11]. However, the ability to assess copy number variation in disease has been limited by the lack of techniques for accurately measuring the copy number level of each CNV in each individual, and the lack of enabling basic knowledge about the precise locations and allele frequencies of most of the copy number polymorphisms (CNPs) that segregate in the human population[12]. We sought to develop hybrid oligonucleotide microarrays to accurately analyze SNPs and copy number variation simultaneously; to use these arrays to map the genomic locations, allele frequencies and population-genetic properties of human CNPs; and to apply this knowledge to advance strategies for querying CNV in genome-wide association studies.

RESULTS

Development of hybrid SNP-CNV genotyping arrays

The expansion of content on genotyping arrays was enabled by the empirical observation that genotype information might be captured more efficiently. An earlier genotyping array (Affymetrix 500K) interrogated each SNP with 24–40 different 25-mer probes, designed to query both strands at multiple offsets with respect to each SNP. We evaluated the information content of each probe with respect to the correct genotype (from HapMap[13,14]) and found that for each SNP, some probes were more informative than others (**Fig. 1a**). Using only the best A/B probe pair for each SNP supported genotyping performance only slightly diminished compared to using all 24–40 probes (**Fig. 1b**).

This result, supported by independent studies[15,16], suggested an alternative design involving multiple replicates of the most informative probes rather than a single copy each of many different probe sequences. Simulation of such a design achieved improved performance with markedly fewer probes (**Fig. 1c**). A prototype microarray (Affymetrix SNP 5.0) was manufactured using the best A/B probe pair

[1]Program in Medical and Population Genetics and Genetic Analysis Platform, The Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA. [2]Department of Molecular Biology, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. [3]Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. [4]Harvard Medical School, Boston, Massachusetts 02115, USA. [5]Harvard-MIT Division of Health Sciences and Technology, Cambridge, Massachusetts 02139, USA. [6]Graduate Program in Biophysics, Harvard University, Cambridge, Massachusetts 02138, USA. [7]Affymetrix Inc., Santa Clara, California 95051, USA. [8]Division of Genetics, Brigham and Women's Hospital, and Harvard Medical School-Partners HealthCare Systems Center for Genetics and Genomics, Boston, Massachusetts 02115, USA. [9]Department of Medicine, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. [10]These authors contributed equally to this work. Correspondence should be addressed to S.M.(smccarro@broad.mit.edu) or D.A.(altshuler@molbio.mgh.harvard.edu).

Received 20 February; accepted 18 August; published online 7 September 2008; doi:10.1038/ng.238

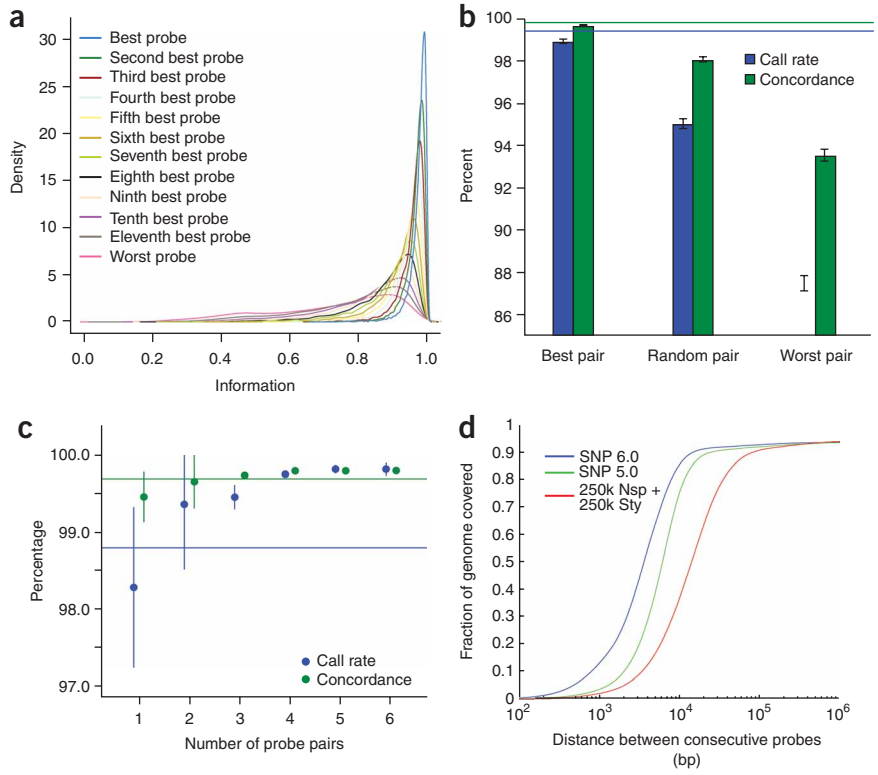

Figure 1 Design of new microarrays. (**a**) Density plots of information content at a probe level for 10,000 randomly selected SNPs from the Affymetrix 500K platform. The *x* axis of information is proportional to the *C* statistic of logistic regression; this statistic measures how well a given covariate or set of covariates can predict an outcome (in this case, HapMap genotypes) in standard logistic regression. (**b**) Using the information metric described above, we could define the best and worst probe pair (in a joint sense). A random probe pair was also selected. Using these selections, we ran the Bayesian robust linear model using Mahalanobis distance (BRLMM) algorithm on the 270 HapMap samples (Sty fraction) but only used two probes ('blinded' performance). The comparison of the normal operation of BRLMM (all pairs) to the 'blinded' performance of the best pair, random pair and worst pair is shown in terms of call rates and concordances. Error bars designate the 90% confidence intervals (*t*-test, $n = 270$). (**c**) A single sample was hybridized to 21 Affymetrix 500K arrays. Using these empirical data, we simulated virtual chips in which each probe was tiled up to 21 times. For each simulation of the number of probes, the experiment was conducted in triplicate, that is, three random draws from the 21 replicates were done. Shown are the means and 90% confidence intervals of call rates and concordances (*t*-test, $n = 3$). Horizontal lines represent empirical performance of the Affymetrix 500K array (Nsp fraction). (**d**) Physical coverage of the genome for copy number analysis. Cumulative fraction of the nucleotides in the genome that lie within probe-to-probe intervals of the specified size or smaller.

each in four copies, requiring only four million probes (less than one array) to interrogate the same 500K SNPs (**Supplementary Table 1** online). Cluster separation between SNP genotype classes was improved (**Supplementary Fig. 1** online). To design a next-generation array (Affymetrix SNP 6.0), we screened more than two million additional SNPs (chosen from HapMap and dbSNP) on a prototype array, selecting 936,000 SNPs to optimize coverage of common patterns of variation in the three HapMap analysis panels (**Supplementary Table 1**).

We empirically selected 940,000 'copy number probes' to directly interrogate copy number variation, unrestricted by the locations and sequence properties of SNPs. To provide uniform coverage across the genome, we selected 800,000 probes, from a screening array of 13 million candidate probes[17], on the basis of genomic spacing and performance in a titration experiment (**Supplementary Methods** online). To maximize detection in regions of previously reported CNV, 140,000 additional probes were targeted at high density in 3,700 regions previously reported to contain CNVs[1–9].

Initial performance evaluation

Using a novel SNP genotyping algorithm[18] applied across a variety of samples and laboratories, we observed genotyping completeness greater than 99%; concordance with HapMap genotypes exceeded 99.5% (ref. 18). Coverage of common SNPs in HapMap via linkage disequilibrium was substantially improved (**Supplementary Table 1**).

We evaluated the performance of the array probes for copy number analysis by using a signal-to-noise (SNR) metric to evaluate the ability of X-chromosome probes to distinguish DNA from males and females. Individual copy number probes on the SNP 6.0 array showed greater SNR (median = 2.0) than SNP probe sets on the 500K or SNP 5.0 arrays (median = 1.4). An individual BAC probe (spanning 150 kb, 6,000 times larger) provides still-higher SNR (for BAC probes from ref. 9, median SNR = 6.9). The SNR of the typical BAC probe was achieved with about 10 probes from the SNP 6.0 array, spanning a mean of 22 kb (**Fig. 2a**). In practice, however, both techniques seem to be able to identify CNVs much smaller than 150 kb or 22 kb, as we describe below.

Physical extent of human copy number variation

We used this array to develop a high-resolution map of copy number variation in the 270 HapMap samples. Our goal was to construct a map that was precise and accurate in two dimensions: (i) the boundaries of the genomic regions affected by CNV and (ii) the measurement of an accurate integer copy number level for each segment in each individual.

We used two computational approaches to identify CNVs: the hidden Markov model Birdseye[18], and an approach based on correlation between nearby probes across a population sample (Methods). To maximize the quality of reported findings, we ran duplicate experiments in independent labs, and report the CNVs that were observed in both experiments, in the same samples and at essentially identical genomic locations. Using these stringent criteria, we identified 3,048 CNV regions among the HapMap samples, of which 60% overlapped with (and 40% fell outside) the regions of previously reported CNV in which we had increased the density of copy number probes during array design.

To estimate the sensitivity of our approach for identifying CNVs and its precision for estimating their locations, we compared these results to a set of CNVs that had been identified in eight of the same HapMap individuals by fosmid end-sequence-pair (ESP) analysis

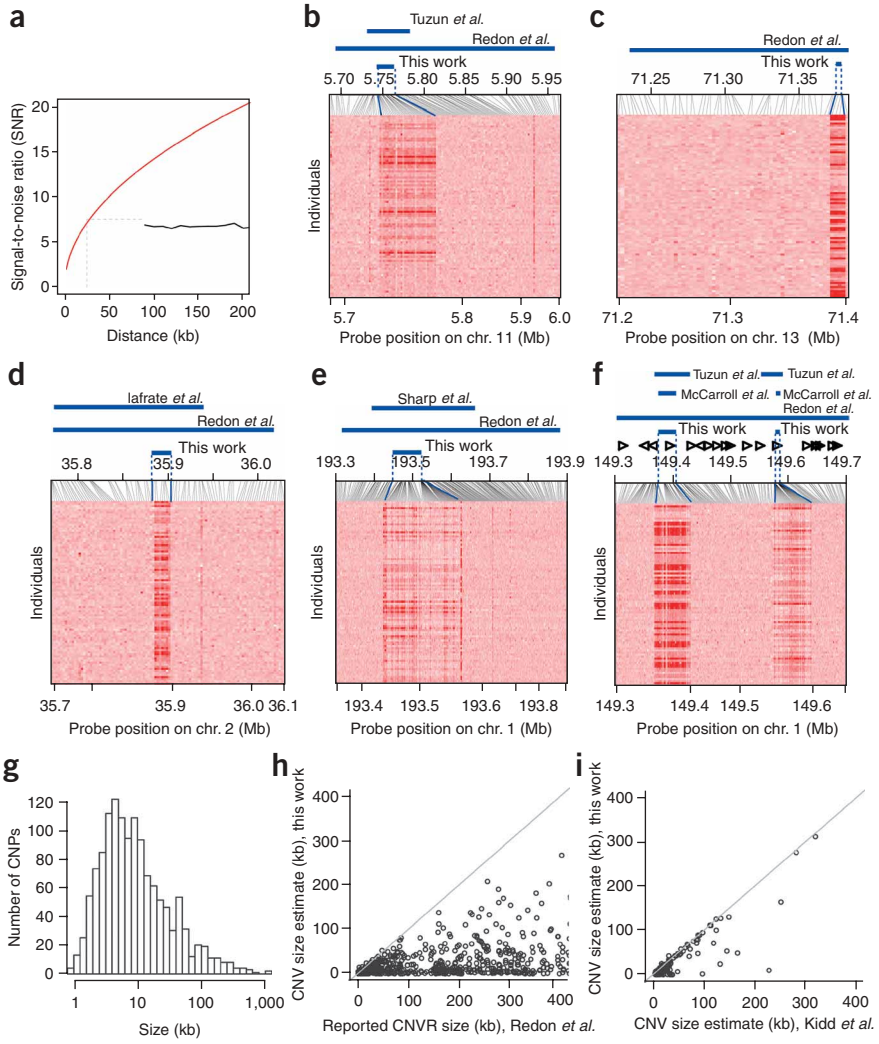

Figure 2 Discovery and sizes of CNV regions. (**a**) Signal-to-noise ratio (SNR) as a function of physical resolution. Black curve indicates the SNR of WGTP BAC probes from an earlier study used to make an initial map of copy number variation across the HapMap samples[9]. Red curve indicates the SNR of sets of probes on the SNP 6.0 array, as a function of the distance spanned by those probes. Dotted lines indicate that the SNR achieved by a typical BAC (150 kb) is achieved at approximately 22-kb physical resolution by probes on the SNP 6.0 array. (**b**–**f**) Revision of CNV regions. Heat-map representation of copy number measurements from the SNP 6.0 array in 90 individuals (HapMap CEU). At each probe, intensity measurements (relative to median sample) are represented by shades of orange, with red corresponding to reduced intensity and yellow to increased intensity. Note that at sites of common CNPs, the median individual can be heterozygous for a CNP allele and therefore have an intermediate copy number. CNV definitions from earlier studies are indicated by blue horizontal lines. In panel **f**, triangles indicate genes in the late cornified envelope (LCE) gene family. (**g**) Length distribution of CNPs observed to segregate at a MAF greater than 1% in one or more of the HapMap populations. (**h**,**i**) Comparison of the estimated sizes of CNPs identified in this study to corresponding size estimates of the same CNPs from studies by Redon *et al.* (**h**) and Kidd *et al.*

and localized by complete fosmid resequencing or targeted 200-bp-resolution oligo array CGH[11]. Of the independently identified CNVs from these eight individuals, our data identified 76% of the CNVs larger than 10 kb and 64% of the CNVs 5–10 kb. We identified only 27% of the ESP-identified CNVs smaller than 5 kb (and believe the true sensitivity of our approach for such small CNVs to be much less than 27%, as both approaches have sharply diminishing sensitivity for CNVs <4 kb). Our estimates of the boundaries of CNV regions differed from the sequencing-established breakpoints by a median of 1.6 kb.

We compared this map to two other datasets: a catalog of CNV regions from the same 270 HapMap samples identified by Redon *et al.* using BAC array CGH and the 500K array[9], and the full set of structural variants identified in eight of these same individuals by analysis of fosmid ESPs[11]. We identified CNVs within 82% of the regions identified by Redon *et al.*, but our CNVs were generally far (5–15 times) smaller than the size of the reported CNV regions (**Fig. 2b–h**). By contrast, our size estimates showed good concordance (**Fig. 2i**) with estimates based on the apparent size discrepancy of fosmid ESPs.

The agreement of our data with the results of a sequence-based method confirms that the physical scale of CNVs is far smaller than the CNV regions initially described in the HapMap samples. This challenges the prevailing interpretation of the earlier data, which has assumed that such events are large (>100 kb) and are a leading indicator of a far-larger number of intermediate-size (10–100 kb) CNVs yet to be discovered[19]. Published analyses of the functional content of CNVs—the genes, ultraconserved elements and segmental duplications they contain—based on a literal interpretation of the reported coordinates of CNV regions will need to be reevaluated in light of the 80–90% downsizing of most of these regions. For example, a finding that CNVs disproportionately affect structural proteins is based on a 400-kb CNV region that contains a family of 20 structural protein–encoding genes in the late cornified envelope family[9,20]; our results indicate that only 45 kb of this 400-kb region is affected (by 2 distinct, common CNVs), and that only 2 (rather than 20) of these genes are copy number variant (**Fig. 2f**).

Moreover, estimates that 12% of the genome is involved in large-scale copy number variation[9] and that 18% of the genome is involved in copy number variation at all scales (Database of Genomic Variants) were not supported by our analyses. Even after correcting for sensitivity with respect to the sequencing-discovered CNVs, we estimate that large-scale (>50 kb) CNVs affect less than 5% of the genome in these 270 individuals, and cause a still-smaller fraction of the genome (less than 0.5%) to differ in copy number between any two individuals.

Common copy number polymorphisms

Almost half (1,320) of the CNV regions were observed in multiple unrelated individuals, generally across genomic segments that were indistinguishable at the resolution of the array (**Fig. 2b–f** and

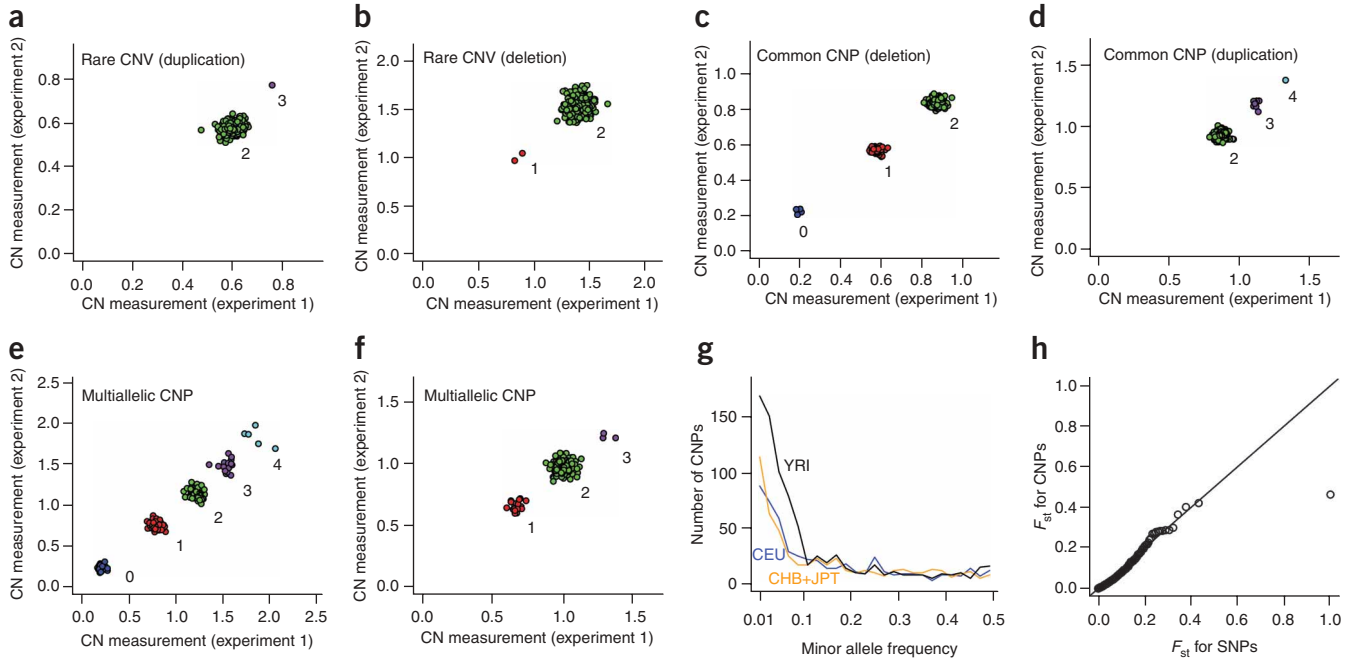

Figure 3 Classes of copy number variants and reproducibility and discrete quality of copy-number measurements. (**a**–**f**) Reproducible population distributions of discrete-valued copy number measurements. Each HapMap sample was analyzed in two different labs; probes across each CNP were used to derive a summarized measurement for each sample for each CNP, in each experiment. Observed categories of variation include rare CNVs (**a**,**b**), in which an altered copy number level is observed in only a single individual, or in an individual and her offspring; diallelic CNPs (**c**,**d**), for which two copy number alleles each seem to segregate at an appreciable frequency; and multiallelic CNPs (**e**,**f**) that are not explained by a simple, two-allele system. (**g**) Frequency distribution for CNPs in different HapMap populations (blue, CEU; orange, JPT + CHB; black, YRI). (**h**) Quantile-quantile plot comparing F_{st} (for HapMap CEU and YRI populations) for CNP and SNP genotypes.

Supplementary Table 2 online). We refer to these as copy number polymorphisms (CNPs), because they appear (at 2-kb resolution) to involve the same affected genomic sequence in each individual (an inference that was not possible in BAC-resolution CNV studies or in CNV catalogs made from isolated individuals) and are therefore consistent with a model of a polymorphism that segregates in the population at an allele frequency greater than 1%.

Because CNPs segregate at an appreciable frequency, they can be heterozygous or homozygous in an individual's genome, giving rise to three or more potential copy-number levels in a population. An individual's status for a CNP is therefore not well described by terms such as 'gain' or 'loss' relative to a 'normal' reference. To assess the role of CNPs in disease and population genetics, the integer copy-number level of each CNP locus must be accurately measured in each individual[12,21]. To type these 1,320 CNPs, we summarized the intensity measurements of the probes corresponding to each CNP into a single measurement for each sample (Methods); these measurements were then clustered into discrete classes corresponding to successive integer copy number levels (**Fig. 3a–f** and **Supplementary Fig. 2** online). A set of heuristics, informed by the population-wide distribution of copy number intensity measurements (Methods), was then used to assign a specific integer copy number level to each class (**Fig. 3a–f** and **Supplementary Table 3** online).

We used this data to delineate diallelic and multiallelic CNPs. For most autosomal CNPs (1,154 of 1,292), we observed two or three diploid copy number levels, distributed across populations and within families in a manner consistent with the mendelian segregation of two underlying copy number alleles in Hardy-Weinberg equilibrium (**Fig. 3c**,**d**). A minority of autosomal CNPs (138 of 1,292, approximately 10%) could not be completely explained by a two-allele system, because the population data was distributed into four or more diploid copy number levels (**Fig. 3e**) or into three levels for which the intermediate level was overwhelmingly the most common (**Fig. 3f**). (For an additional 180 regions, measurements did not cluster into well-separated classes; these could in principle represent more-complex CNVs, or events for which our array lacked sufficient measurement precision.)

To evaluate the accuracy of these copy number assignments, we used several approaches. Quantitative PCR evaluation of 810 CNP genotypes (27 common CNPs in 30 individuals) yielded concordance of 99.3% with our determinations of integer copy number. To globally evaluate genotypes for diallelic CNPs, we applied quality control criteria analogous to those used to assess SNP genotypes. Diallelic CNPs showed a low rate of deviation from mendelian inheritance (0.1% per trio per CNP), consistent with a low rate of genotype error (comparable to that of high-quality filtered SNP genotypes in HapMap[13,14]). Genotypes for common diallelic CNPs also conformed to Hardy-Weinberg equilibrium (failing at $P < 0.01$ with a rate of 0.02). To evaluate genotypes for multiallelic CNPs (**Fig. 3e**,**f**), we used inheritance (Fisher's h) to measure the correlation between parental and offspring copy number levels. Across 53 multiallelic CNPs that segregated at high frequency among the CEU and YRI trios (with at least 10% of individuals showing a nonmodal copy number), Fisher's h was distributed with a mean of 0.98 (statistically equivalent to the level of 1.0 expected for perfectly heritable traits).

We then assessed the nature of human copy number variation in a manner informed by the frequency of each copy number class in each population. For the CNV loci identified here, two unrelated

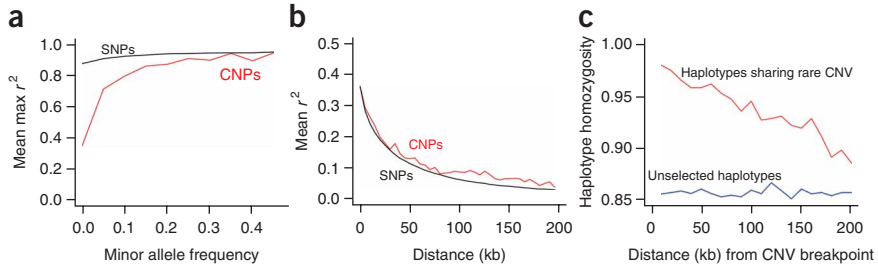

Figure 4 Linkage disequilibrium properties of CNPs. (**a**) Taggability of common CNPs (red) and common SNPs (black), expressed as the maximum correlation (r^2) to an individual SNP (data shown for HapMap CEU). (**b**) Average strength of linkage disequilibrium (mean r^2) as a function of distance from a polymorphism, for common (MAF > 5%) CNPs (red) and common SNPs (black) in HapMap CEU. (**c**) Haplotype homozygosity around low-frequency CNVs, expressed as the homozygosity of SNPs on haplotypes sharing a low-frequency CNV (red), compared to the homozygosity of the other haplotypes in the same population (blue).

individuals from the same population differed in copy number at an average of 175 (in CEU, JPT and CHB) or 230 (in YRI) autosomal loci, spanning 5.9 Mb of the genome (6.3 Mb in YRI) and overlapping the transcribed regions of 100 genes (120 in YRI). Approximately 92% of the copy number differences in this comparison resulted from CNPs (versus only 8% from rare CNVs observed in a single individual or family), of which 85% (80% in YRI) resulted from common CNPs with minor allele frequency (MAF) of at least 5%. This suggests that a limited set of common CNPs captures most human copy number differences, a pattern of variation similar to that observed for SNPs.

Rare and *de novo* CNVs

We carried out a carefully curated analysis to identify *de novo* CNVs among the offspring in 60 HapMap father-mother-offspring trios (**Supplementary Methods** and **Supplementary Table 4** online). All but ten of the rare CNVs observed in trio offspring were also observed in at least one parent (some of these ten could be somatic mutations or cell-line mutations that are not present in the individual's germ line.) Thus, even when ascertained in a cultured lymphoblastoid cell line, the contribution of *de novo* CNV formation to an individual's genome seems to be at least 100 times smaller than the contribution of inheritance. Combined with the observed low rate of mendelian inconsistencies for accurately typed common CNPs, these results suggest that the copy number differences observed among normal individuals are overwhelmingly inherited from parents.

The fact that we did not detect CNVs in thousands of the CNV regions described in current databases[22] (despite having dozens of probes within most of those regions) suggests that many such reported CNVs are rare variants (or false discoveries), consistent with the allelic spectrum we observe (in which fully half of the CNV regions we observe are singleton CNVs) and therefore partly explaining the previously observed lack of concordance among CNV datasets ascertained in different individuals[23].

Population-level properties of copy number variation

Many studies have identified a larger number of CNVs in population samples with African ancestry than in similarly sized population samples without African ancestry. We found that this effect is entirely explained by CNVs with allele frequency less than 10% (**Fig. 3g**). As does a similar relationship for SNPs, this indicates that many low-frequency alleles were lost in the population bottlenecks associated with human migration out of Africa.

Some CNPs have been observed to segregate at different frequencies in different populations, potentially owing to the action of recent selection on CNP alleles[24,25]. The hypothesis that CNPs are particularly likely to have functional effects that would cause their allele frequencies to be shaped by recent selection has not been tested on a genome scale. We assessed the distribution of F_{st}, a measure of population differentiation in allele frequencies, for SNP and CNP genotypes. The distributions of F_{st} for SNPs and diallelic CNPs were indistinguishable (**Fig. 3h**), suggesting that common CNPs are not more influenced by recent selection than are common SNPs, and that most population differentiation in CNP allele frequencies is explained by simple genetic drift.

Linkage disequilibrium properties of CNPs

Previous estimates of linkage disequilibrium between SNPs and CNPs have been constrained by the limited number of CNPs for which accurate genotypes could be obtained, by insufficient knowledge of the true locations of CNPs detected by large clones, and by diminished density of effectively typed SNPs (that could serve as potential tags) in the repeat-rich regions in which CNVs are enriched.

We assessed linkage disequilibrium in two ways: for common CNPs, by correlation to other genetic markers; for rare CNVs, by the haplotype diversity of chromosomes that carry the event. Most common, diallelic CNPs (with MAF greater than 5%) were perfectly captured ($r^2 = 1.0$) by at least one SNP tag from HapMap Phase II (**Fig. 4a**). Common CNPs showed a modest shift in the distribution of r^2 compared to frequency-matched SNPs (**Fig. 4a**). This taggability

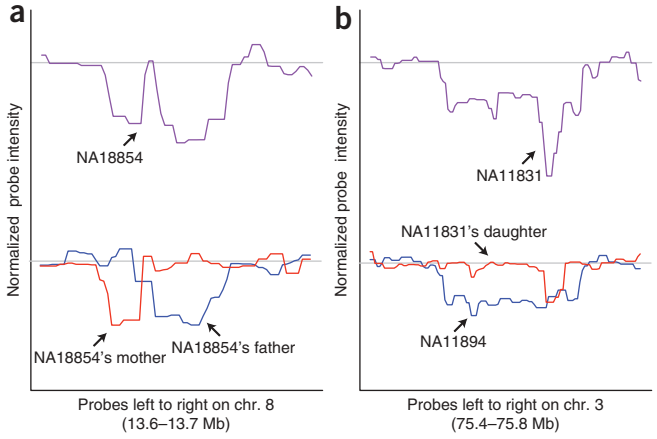

Figure 5 Dissection of complex CNVs. (**a**,**b**) Two CNV regions that are shaped by multiple mutational events. In **a**, the appearance of a long, complex deletion in individual NA18854 (purple) results from two nearby, segregating deletion polymorphisms that NA18854 inherited from different parents (red and blue). In **b**, the appearance of an architecturally complex CNV in individual NA11831 (purple) is explained by the combination of two simpler deletion polymorphisms (red and blue) that are segregating at the same locus but appear to reside on different haplotypes.

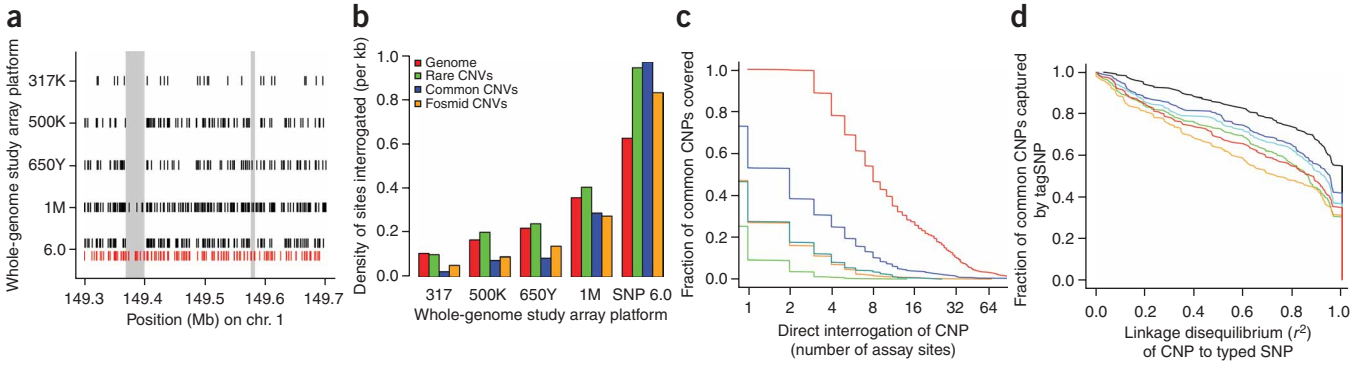

Figure 6 Capture of CNPs in genome-wide association studies via direct interrogation and linkage disequilibrium. (**a**) Locations of sites interrogated on five SNP array platforms, at a representative locus containing two common CNPs (gray regions). Black tick marks indicate coverage by SNP assays; red tick marks indicate coverage by copy-number probes. The horizontal span of the figure corresponds to the region described as copy number variant in the Database of Genomic Variants. (**b**) Density of interrogated sites (number of sites per kb) in the genome as a whole (red bars), in low-frequency CNPs (MAF < 5%, green), in common CNPs (MAF > 5%, blue) and in a set of CNVs identified without regard to frequency by an independent, sequence-based method (orange). (**c**) Capture of the common CNPs (MAF > 5%) identified in this study by direct interrogation on genome-wide association platforms. For each number of probe locations, the height of the curve indicates the fraction of common CNPs containing at least that number of probes on that array platform (orange, Affymetrix 500K; green, Illumina 317; light blue, Illumina 650Y; blue, Illumina 1M; red, Affymetrix SNP 6.0). Note that because these CNPs were discovered on the SNP 6.0 array, estimates of the coverage of short CNPs (covering only a few probes) on the SNP 6.0 array are biased upward relative to what one would presumably observe for independently ascertained, independently genotyped CNPs; the analysis is intended for comparison to efficacy of linkage disequilibrium–based strategies. (**d**) Capture of common (MAF > 5%) CNPs by linkage disequilibrium to SNPs typed on genome-wide SNP array platforms (black, SNPs in HapMap; other colors, same as in panel **c**; data are for HapMap CEU analysis panel).

gap, noted in earlier studies, could in principle be due to recurrent rearrangements at CNV sites[4,9] or to a paucity of SNPs in repeat-rich regions to serve as potential tags[8,9]. To distinguish between these models, we examined the relationship between r^2 and physical distance from the estimated CNP breakpoint: recurrent CNV formation would reduce correlations to flanking markers, but a paucity of SNP tags would reduce the number of potential tags without affecting the relationship between r^2 and physical distance. Mean r^2 as a function of distance was indistinguishable for SNPs and diallelic CNPs (**Fig. 4b**), suggesting that common, diallelic CNPs are overwhelmingly ancestral mutations. (We note that multiallelic CNPs, and CNPs not genotyped by our platform, may have different properties.)

The linkage disequlibrium properties of rare CNVs have not been previously investigated. A CNV site observed in multiple individuals may arise from shared ancestry at a locus or, alternatively, from independent structural mutations at the site. To evaluate the extent to which shared rare CNVs in normal individuals arise from shared ancestry, we identified 162 CNV segments that were observed in exactly two unrelated YRI individuals in HapMap. We used inheritance to phase these CNV alleles onto the SNP haplotypes defined by HapMap SNP genotypes in the same trios. Rare CNVs that were observed in two unrelated individuals were almost always present on the same SNP haplotype in both individuals (**Fig. 4c**). This haplotype sharing was robustly detectable at distances greater than 200 kb (**Fig. 4c**), suggesting that sharing of rare CNVs can imply sharing of much-larger genomic regions.

Dissecting complex CNVs

Studies of CNV have noted its potential complexity; several CNVs have been identified that cannot be understood in terms of a single mutational event. We explored a number of these and found that they could be explained not by one, but by two or three simple mutations. In some cases, the apparent complexity of a CNV resulted from interrogating multiple chromosomes together (**Fig. 5**). For example, a deletion region that was previously reported to be complex and longer in one individual than in others[6] seems in fact to comprise two nearby, nonoverlapping deletion polymorphisms (both common) that the individual inherited from different parents (**Fig. 5a**); both deletions segregate on specific SNP haplotypes and therefore seem to be unique, ancestral events. Another CNV that initially seemed to show architectural complexity was readily explained as the combination of two simpler deletion polymorphisms that were segregating separately (**Fig. 5b**) on specific SNP haplotypes. Many apparently complex CNV regions can thus be incorporated into association studies as their molecular and population-genetic nature is elucidated.

Copy number analysis of earlier whole-genome scans

We had hypothesized that the content of earlier SNP arrays was systematically biased against genomic regions affected by common CNPs, because common CNPs cause SNP data to fail the quality control checks that were used to qualify SNPs for inclusion on commercial arrays. Our map of CNPs and their allele frequencies confirmed this hypothesis: common (but not rare) CNPs generally corresponded to bald spots in the physical coverage of earlier SNP arrays, a bias that is now largely ameliorated in new platforms (from both Affymetrix and Illumina) with CNV-targeted content (**Fig. 6a**,**b**). Of 423 CNPs that we observed to be common (MAF > 5%) in HapMap, fewer than half (44%) were interrogated by even a single SNP assay on the Affymetrix 500K or Illumina 650Y arrays, and less than 20% were interrogated by three or more SNPs (**Fig. 6c**). By contrast, low-frequency CNPs (as well as rare CNVs) showed little coverage bias and were much better captured (**Fig. 6b**). The differences between earlier and CNV-targeted arrays persisted when we analyzed coverage of a completely independent set of 100 CNV regions discovered by fosmid ESP analysis and refined by complete sequencing[11]; these sequencing-defined CNV regions showed a coverage bias similar to (though slightly less extreme than) that of common CNPs, reflecting that they are a mixture of common and rare variants (**Fig. 6b**).

The extent to which earlier SNP arrays are blind to common CNPs has not been previously appreciated, because most estimates of CNV coverage utilize the Database of Genomic Variants[22], which seems to include extensive non-CNV genome in CNV definitions (**Figs. 2b–f,h** and **6a**) and does not distinguish between common and rare CNVs. Although there is extensive literature on mining copy number information from SNP array data, our results suggest that efforts to extract copy number information from earlier SNP arrays will miss most of the CNPs that are common in the populations used to screen SNP assays for earlier commercial arrays (populations with European, African and East Asian ancestry). Thus, a recent report (based on an earlier SNP array) that human CNV consisted almost entirely of rare variants, and that CNVs were more common in populations from Oceania and the Americas[26], is likely to reflect an inability to ascertain most of the CNPs that are common in European, African and East Asian populations.

Even if most common CNPs cannot be observed directly on earlier SNP arrays, the disease effects of some common CNPs could potentially be captured through linkage disequilibrium to SNPs that are typed. We assessed the capture of common CNPs by linkage disequilibrium to SNPs typed on five commercial array platforms widely used for genome-wide association studies (**Fig. 6d**). Slightly under half (40–50%) of common CNPs were captured ($r^2 > 0.8$) by markers on first-generation SNP arrays. This suggests that a partial picture of the contribution of common CNPs to disease might be obtained by analyzing the disease association of CNP-tagging SNPs. In fact, we found a common deletion polymorphism that is in perfect linkage disequilibrium with Crohn's disease–associated SNPs at *IRGM* and is a strong candidate to explain the Crohn's association there[27].

DISCUSSION

We have developed new experimental tools for simultaneously interrogating SNPs and CNVs across the genome, and used them to develop a map of segregating CNPs in the HapMap cohorts.

Our results document that large-scale (>100-kb) CNV affects far less of the human genome than reported in initial studies, mainly because most CNVs are far smaller than reported CNV regions (**Fig. 2**). The scale of human CNV seems to have been overestimated in large part because of the interpretation of data from arrays of large-insert clones; the assumption that large clones detected similarly large CNVs may have seemed to be reasonable, and was the basis for many analyses, but was largely incorrect (**Fig. 2**). Our approach has its own limitations: it is limited to sequences that are already present in the finished human genome sequence, and it misses many regions of high-multiplicity duplication. Even with allowances for these factors, however, the fraction of human genetic variation attributable to large-scale (>100-kb) CNV seems unlikely ever to approach the estimates of earlier studies.

More than 90% of the observed copy number differences between any two individuals seem to be due to CNPs that segregate in the population at a MAF greater than 1%. Common CNPs seem to show patterns of allele frequency, linkage disequilibrium and population differentiation that mirror the properties of SNPs. Cataloging the genomic locations, haplotypes and sequence properties of these alternative structural alleles will therefore be an important direction for completing databases of common patterns of genetic variation in the human population[13]. Because common CNPs do not need to be discovered *ab initio* in each study, we propose that they could be analyzed much more accurately (than in current practice) by using focused algorithms analogous to those used to genotype SNPs[12,18]; the resulting genotypes could then be analyzed for association to disease. As human copy number variation seems to derive overwhelmingly from inheritance and shared ancestry, linkage disequilibrium–based analyses will provide an important context for discovering and interpreting putative associations of both SNPs and CNVs with phenotypes.

Rare CNVs represent a leading edge of a next frontier of human genetics research, which involves studying collections of rare variants[28,29] for roles in disease. An important direction will be to develop rigorous analytical approaches for collecting rare variants into biologically meaningful classes that can be tested for enrichment in affected individuals. The simultaneous study of SNPs and structural variants, both common and rare, will be needed to understand the relative contribution of each form of variation to disease in human populations.

METHODS

CNV discovery. We developed two approaches for CNV discovery. The first was a hidden Markov model, Birdseye[18], which identifies CNVs in one sample at a time. We also developed a conceptually different method to identify common CNPs by searching for genomic regions across which copy number probes showed cross-sample patterns of intensity that were highly correlated. All sets of 2, 4 and 8 consecutive probes were analyzed. (Simulation indicated that CNVs spanning intermediate and larger numbers of probes would be captured with high sensitivity from the 2-, 4- and 8-probe windows that they overlapped; the selection of 2-, 4- and 8-probe windows seemed to optimally balance between the goals of covering a range of potential CNV sizes and minimizing redundant hypothesis testing.) The 'correlation score' for any set of 2, 4 or 8 probes was the median pairwise correlation of their measurements across the 270 samples. To create an empirical null distribution with which we could determine the probability of observing particular correlation scores by chance, we randomly permuted the locations of all probes and repeated the analysis. To set thresholds (separate for each window size) for determining correlation scores to be significant, we compared the empirical distribution to the null distribution, identified the correlation score corresponding to a likelihood ratio of 100 (the correlation at which the density of the empirical distribution was 100 times greater then the density of the null distribution), and used this as a threshold. This identified groups of highly correlated neighboring probes across the genome. We agglomeratively clustered overlapping groups of probes. The result was the identification of a series of genomic segments over which particular population copy number patterns prevailed.

To further refine the boundaries of these genomic segments, we used data from the independent experiments in which the same 270 HapMap samples were independently prepared and hybridized at Affymetrix (experiment 1) and the Broad Institute (experiment 2). For each CNV segment identified as described above from the data in experiment 1, we summarized the probe-level intensity measurements into a single summarized measurement per sample per CNV segment (using median polish as described below), then identified (from the data at the same genomic locus in experiment 2) the individual copy number probes and SNP probe sets for which intensity measurements were significantly ($P < 10^{-4}$) correlated with the (CNV-summarized) measurement from experiment 1. In this way, we defined a potential breakpoint using information across the entire sample set; the approach also implicitly required that there be evidence for CNV in both independent experiments (a criterion which filtered out approximately 10% of the CNVs identified in either experiment on its own).

We excluded a class of extremely short candidate CNV regions for which the contributing probes all came from one or two consecutive Nsp or Sty genomic restriction fragments, as these could be artifacts of the sample preparation process (which involved digestion of genomic DNA with Nsp and Sty).

CNV regions and sizes were defined by the span of the genomic sites interrogated by all probes contributing evidence to a CNV (which tends to slightly underestimate CNV size).

Summarization of CNV probe sets. For each CNV, we generated a set of summarized intensity measurements (one summarized measurement per

sample) using the Tukey median polish of the log intensities of the contributing probes. Briefly, our application of median polish involved creating a probes-by-samples matrix of log-intensity measurements for all probes spanned by the CNV and all samples in the same experimental plate, then solving an additively fit model (for this matrix data) of the form row effect + column effect + overall median; the resulting column effects (corresponding to sample effects, with one measurement per sample) were then transformed back out of log space and used for genotyping (described below).

Genotyping CNPs. A typical use of genotyping arrays involves hybridizing each sample to a single array and performing automated computational analyses of the data; we describe in ref. 18 a suite of algorithms for such an application. To create a reference dataset of particular crispness and completeness, we used here the independent replicate experiments performed on the same samples in labs at Broad and Affymetrix, and further undertook a process of manual curation of the genotypes for each CNP. For each CNP probe set, we generated a scatter plot comparing the summarized intensity measurement of each HapMap sample in experiment 1 versus the summarized intensity measurement for the same samples in experiment 2; we then clustered this data into the discrete copy number classes present in the population (**Fig. 3a–f** and **Supplementary Fig. 1**). For CNVs that had been identified only by Birdseye, we used the copy number call from Birdseye. For each CNV, we generated cluster assignments only for those samples for which the cluster assignment was clear in both experiments and identical between experiments, resulting in 98.1% data completeness (**Supplementary Fig. 1** and **Supplementary Table 1**).

Determination of integer copy number. To assign an integer copy number to each genotype cluster (**Fig. 3a–f**), we used an observation and an assumption: (i) the observation that hybridization intensity increased less than linearly with the number of DNA copies and (ii) the assumption that the series of copy number classes present in the population represent consecutive integer copy number levels. This led to a simple heuristic, in which the ratios of the average intensities for consecutive copy number classes were used to inform the choice of consecutive integer copy number levels. For example, for a CNP for which two copy number classes were observed (**Fig. 3c,d**), the observation of an intensity ratio less than 2:1 but greater than 3:2 between the two classes evidenced that the CNP was a deletion CNP (producing copy number classes of 2 and 1) rather than a duplication CNP (producing copy number classes of 3 and 2) (**Fig. 3a,b**). The integer copy number determinations made with this heuristic were tested by quantitative PCR and determined to be correct for 27/27 loci tested.

Supplementary methods. Additional methods on array hybridization, population-genetic analysis and analysis of *de novo* CNVs are described in **Supplementary Methods**.

Genome coordinates. All physical positions described in the figures utilize the hg17 (b35) build of the human genome sequence.

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

ACKNOWLEDGMENTS

We thank J. Kidd, G. Cooper and E. Eichler for sharing data on high-resolution breakpoints of select CNVs prior to its publication; E. Lander, J. Hirschhorn and S. Kathiresan for thoughtful readings of the manuscript; the Affymetrix team of the Broad Institute Genetic Analysis Platform: W. Brodeur, N. Chia, M. DaSilva, J. Gibbons, N. Houde, M. McConnell, R. Barry, K. Nguyen, J. Camarata, M. Fava and T. Nyinjee under the supervision of C. Gates, B. Blumenstiel, D. Gage and M. Parkin; members of the Affymetrix informatics team: X. Di, H. Gorrell, G. Liu, M. Mittmann, M. Shen, C. Sugnet, A. Willams and G. Yang; members of the Affymetrix arrays and assays team: T. Berntsen, M. Chadha, J. Law, H. Matsuzaki, B. Nguyen, K. Travers, N. Vissa and S. Walsh. S.A.M. was supported by a Lilly Life Sciences Research Fellowship.

AUTHOR CONTRIBUTIONS

F.G.K. conceived a strategy for empirical probe reduction of SNP probe sets. S.A.M. conceived of hybrid arrays consisting of polymorphic (SNP) and nonpolymorphic (copy number) probes. F.G.K., S.A.M. and D.A. proposed to Affymetrix a specific redesign of the 500K SNP array based on these concepts. The idea was further developed with input from R.R., J.B., S.C., S.L., K.W.J., S.B.G. and M.J.D., and a pilot initiated. For the pilot (which became the SNP 5.0 array), F.G.K. and J.B.M. selected SNP probe sets, and S.A.M. and J.M.K. selected copy number probes. For the development of the SNP 6.0 array, R.M. directed laboratory SNP screening experiments which were analyzed by S.C., E.H. and T.W. P.I.W.d.B., J.B.M. and S.C. selected SNPs from those which passed the screening effort, using a linkage-disequilibrium tagging strategy. S.A.M. and M.H.S. designed and M.H.S. directed laboratory work for the titration experiment that guided empirical selection of copy number probes; on the basis of these results, together with informatic analyses which A.K. performed, S.A.M. and J.M.K. selected copy number probes. Laboratory experiments at Broad Institute were led by M.P. and S.B.G. A.W., J.N., R.H. and E.H. developed supporting software. S.A.M., J.M.K. and J.N. analyzed the data to identify CNVs. S.A.M., J.N., F.G.K. and J.M.K. developed CNP genotyping analysis. P.J.C. conducted and J.V. analyzed experiments to validate CNP genotypes experimentally. S.A.M. analyzed the population-genetic and linkage-disequilibrium properties of CNVs. J.M.K. analyzed the data for evidence of *de novo* CNVs. A.L.E. analyzed platforms' coverage of CNVs. S.A.M., F.G.K., J.M.K., M.J.D. and D.A. wrote the manuscript. Discussions among all authors informed the array design, the development of algorithms for analysis and the interpretation of results.

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturegenetics/.

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. Hinds, D.A., Kloek, A.P., Jen, M., Chen, X. & Frazer, K.A. Common deletions and SNPs are in linkage disequilibrium in the human genome. *Nat. Genet.* **38**, 82–85 (2006).
6. 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.* **38**, 75–81 (2006).
7. McCarroll, S.A. *et al.* Common deletion polymorphisms in the human genome. *Nat. Genet.* **38**, 86–92 (2006).
8. Locke, D.P. *et al.* Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. *Am. J. Hum. Genet.* **79**, 275–290 (2006).
9. Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444**, 444–454 (2006).
10. Korbel, J.O. *et al.* Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**, 420–426 (2007).
11. Kidd, J.M. *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56–64 (2008).
12. McCarroll, S.A. & Altshuler, D.M. Copy-number variation and association studies of human disease. *Nat. Genet.* **39**, S37–S42 (2007).
13. The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
14. Frazer, K.A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).
15. Smemo, S. & Borevitz, J.O. Redundancy in genotyping arrays. *PLoS ONE* **2**, e287 (2007).
16. Antipova, A.A., Tamayo, P. & Golub, T.R. A strategy for oligonucleotide microarray probe reduction. *Genome Biol* **3**, RESEARCH0073 (2002).
17. Shen, F. *et al.* Improved detection of global copy number variation using high density, non-polymorphic oligonucleotide probes. *BMC Genet.* **9**, 27 (2008).
18. Korn, J.M. *et al.* Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat. Genet.* advance online publication, doi:10.1038/ng.237 (7 September 2008).
19. Stranger, B.E. *et al.* Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315**, 848–853 (2007).
20. Cooper, G.M., Nickerson, D.A. & Eichler, E.E. Mutational and selective effects on copy-number variants in the human genome. *Nat. Genet.* **39**, S22–S29 (2007).
21. McCarroll, S.A. Copy-number analysis goes more than skin deep. *Nat. Genet.* **40**, 5–6 (2008).
22. Zhang, J., Feuk, L., Duggan, G.E., Khaja, R. & Scherer, S.W. Development of bioinformatics resources for display and analysis of copy number and other structural variants in the human genome. *Cytogenet. Genome Res.* **115**, 205–214 (2006).

23. Scherer, S.W. *et al.* Challenges and standards in integrating surveys of structural variation. *Nat. Genet.* **39**, S7–S15 (2007).
24. Kidd, J.M., Newman, T.L., Tuzun, E., Kaul, R. & Eichler, E.E. Population stratification of a common APOBEC gene deletion polymorphism. *PLoS Genet.* **3**, e63 (2007).
25. Perry, G.H. *et al.* Diet and the evolution of human amylase gene copy number variation. *Nat. Genet.* **39**, 1256–1260 (2007).
26. Jakobsson, M. *et al.* Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* **451**, 998–1003 (2008).
27. McCarroll, S.A. *et al.* Deletion polymorphism upstream of *IRGM* associated with altered *IRGM* expression and Crohn's disease. *Nat. Genet.* advance online publication, doi:10.1038/ng.215 (24 August 2008).
28. Cohen, J.C., Boerwinkle, E., Mosley, T.H. Jr & Hobbs, H.H. Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. *N. Engl. J. Med.* **354**, 1264–1272 (2006).
29. Cohen, J.C. *et al.* Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science* **305**, 869–872 (2004).